NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

A holistic approach to information mining

Internship experience at Oticon A/S

Vlad Robu

Internship report presented as partial requirement for obtaining the Master's degree in Statistics and Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

2019

A holistic approach to information mining

Internship experience at Oticon A/S

Vlad Robu

MGI

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

A HOLISTIC APPROACH TO INFORMATION MINING

by

Vlad Robu

Internship report presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence

Advisor: Prof. Vitor Duarte dos Santos

September 2019

DEDICATION

This paper is dedicated to all those who supported, encouraged and inspired me during this interesting and challenging journey. It is dedicated to those who contributed making this a beautiful experience of personal enrichment, growth and discovery.

It is dedicated to my Market Intelligence Team: Kasper, Jonas and Alessandro. Who welcomed me to this adventure and contributed to my professional, and most important to my internal growth.

To all my friends, who allways encouraged, patiently waited and understood my struggle.

To Elisa, who put joy in my life.

To my family, and especially to my beloved mother, without whom, all these achievements and successes would have been impossible.

ACKNOWLEDGMENTS

I would like to thank Prof. Vitor Duarte dos Santos, who accepted this challenge and encouraged me to pursue the achievement of the full potential of this internship report.

Additionally, I would like to thank my Oticon colleagues, for being amazing collaborators, for hosting me during these ten months of intership and making this experience more than just work. Thank you for transforming it into fun, for transforming it into friendships.

In conclusion, I would like to acknowledge NOVA IMS for making this experience possible and for contributing to make it being partially sponsored by the European Erasmus + Traneeship program.

ABSTRACT

In our days, companies have the huge possibility to understand their customers like never before. They have access to numerous sources of customer related data. Insights that can't be retrieved from secondary sources, can be gathered through market research. Data is the key element to create customer centric value through data-driven decisions. Making the right decisions based on empirical analysis can become a competitive advantage and a critical success factor for big industry players like Oticon A/S - one of the largest in the world manufacturer of hearing aid devices.
This internship report will describe the author's ten months internship experience in the Market Intelligence Team of Oticon. The description of the projects carried throughout the internship and especially their diversity in terms of tools and methodologies, aims to represent a holistic approach to data mining, showing how the latter should be performed considering it as a part of a larger ecosystem of actors and processes. Quantitative research, development of reporting solutions, database management, and application of frequent pattern mining algorithms, are all used to transform data into actionable knowledge.

KEYWORDS

Frequent pattern mining; Apriori; Market Research, PowerBI Reporting Solution

TABLE OF CONTENTS

1. Introduction 1
   1.1. Internship company overview 2
   1.2. Internship overview – the team and activities 3
   1.3. Internship motivation 4
   1.4. Internship definition and goals 4
2. Theoretical framework 5
   2.1. Market Research 5
      2.1.1. Market Research Steps 7
      2.1.2. Data collection methods 9
      2.1.3. Scaling 14
      2.1.4. Data types 17
   2.2. Data Mining 19
   2.3. Big data 20
   2.4. Knowledge discovery in large datasets 20
   2.5. Machine Learning 24
   2.6. Association Rules 25
      2.6.1. Association rules evaluation methods 26
      2.6.2. Association rule algorithms 29
3. Tools and technology 32
   3.1. PowerBI 32
   3.2. IBM SPSS STATISTICS 33
   3.3. R 34
   3.4. Qualtrics 36
4. Projects 38
   4.1. Internship timeline 39
   4.2. Market research Projects 41
      4.2.1. Project A – OM 43
      4.2.2. Project B – SCE 46
      4.2.3. Project C – N2 52
      4.2.4. Project D – Adk 59
   4.3. Reporting solution. Project E - DP 62
   4.4. Project Genie 71
      4.4.1. Business understanding 71
      4.4.2. Problem identification and definition 73

4.4.3. Data understanding and retrieval 75
4.4.4. Data preparation 79
4.4.5. Modelling 83
4.4.6. Model performance 85
4.4.7. Rule evaluation and selection 86
4.4.8. Limitations and Future steps 89
5. Conclusions 90
5.1. Overall evaluation of the internship 90
5.2. Lessons learned 90

LIST OF FIGURES

Figure 1. William Demant Holding – brand umbrella. Source: www.demant.com 1
Figure 2. Oticon logo. Source: www.oticon.com 2
Figure 3. Oticon range of hearing care solutions. Source: www.oticon.com 3
Figure 4. Example of Likert-type "Satisfaction" scale 16
Figure 5. Process of Knowledge Discovery in Databases. Source: Fayyad, Piatetsky-Shapiro, & Smyth, 1996. 21
Figure 6. Universe of Data Mining and Knowledge Discovery methodologies. Source: Mariscal, Marbàn, & Fernàndez, 2010. 22
Figure 7. CRISP-DM process. Source: Chapman P., et al., 2000. 23
Figure 8. The "Magic Quadrant for Analytics and Business Intelligence Platforms". Source: Gartner, 2019. 33
Figure 9. IBM SPSS Statistics - "Analyse" tab. 34
Figure 10. Qualtrics question type. 36
Figure 11. SCE project – Data view screen. Source: IBM SPSS, Project B. 48
Figure 12. SCE project – SPSS variable View screen. Source: IBM SPSS, Project B. 49
Figure 13. Project SCE – Satisfaction Scale. Source: IBM SPSS, Project B. 49
Figure 14. Project SCE – IBM SPSS "Frequencies" tab. Source: IBM SPSS, Project B. 50
Figure 15. Project SCE – IBM SPSS Statistics Viewer. Source: IBM SPSS, Project B. 50
Figure 16. Project SCE – Data visualization. Source: MS Power Point, Project B. 51
Figure 17. Project N2 – IBM SPSS visualization. Relevance Likert scale. Source: IBM SPSS, Project C. 54
Figure 18. Project N2 – IBM SPSS visualization. Quantity scale. Source: IBM SPSS, Project C. 54
Figure 19. Project N2 - IBM SPSS visualization. Assumed reason for non-usage. Source: IBM SPSS, Project C. 55
Figure 20. Project N2 – IBM SPSS. Define Variable Sets option. Source: IBM SPSS, Project C. 57
Figure 21. Project N2 – IBM SPSS Statistics Syntax Editor visualization. Source: IBM SPSS, Project C. 57
Figure 22. Project N2 – Data visualization examples. Source: MS Power Point, Project C. 58
Figure 23. The original data analysis process setup. 62
Figure 24. New reporting solution phases. 63
Figure 25. Adapted modified waterfall methodology. 64
Figure 26. Power Query editor. Source dataset. Source: Power BI, Project E. 66
Figure 27. Power BI editor. Interaction between attributes. Source: Power BI, Project E. 66
Figure 28. Relational model. Many to One relationship. Source: Power BI, Project E. 67

Figure 29. Example of survey data visualization with Power BI. Source: Power BI, Project E. 68
Figure 30. Example of histogram and table used as data visualization in Power BI. Source: Power BI, Project E. 68
Figure 31. Card chart Power BI visualization. Source: Power BI, Project E. 69
Figure 32. Power BI workspace. End-user online report visualization. Source: Power BI, Project E. 69
Figure 33. The journey of the HA from the manufacturer to the end user. Source: own adaptation. 72
Figure 34. Genie software interface. Source: www.google.com; search key: "genie 2 software". 73
Figure 35. MS SQL Data Query - first 1000 rows. Source: SSMS, Project Genie. 76
Figure 36. Simplified conceptual representation of the dataset structure after JSON data partitioning. 77
Figure 37. MS SQL Query - SELECT DISTINCT to identify the levels of the geographical attribute. Source: SSMS, Project Genie. 77
Figure 38. MS SQL Query - Selection of the data to be used for the data mining project. Source. SSMS, Project Genie. 78
Figure 39. Simplified conceptual representation of the dataset after dimensionality reduction. 79
Figure 40. Simplified conceptual representation of the binary transactional dataset. 81
Figure 41. Simplified conceptual representation of the tool dimension creation process. 81
Figure 42. Simplified conceptual representation of the Inner Join performed to select complete and valid sessions. 82
Figure 43. Simplified conceptual representation of the LeftOuterJoin performed to select digital fitting tools used during a digital fitting session. 82
Figure 44. Association Rules by default parameters. Source R: software; Project Genie. 84
Figure 45. Association rules generated with minimum support threshold 0.001. Source R: software; Project Genie. 87
Figure 46. First 20 association rules based on Lift. Source R: software; Project Genie. 88
Figure 47. First 20 association rules based on Support. Source R: software; Project Genie. ... 88

LIST OF TABLES

Table 1. A comparison of Frequent Pattern Mining Algorithms. 31
Table 2. Matrix of tools and projects. 37
Table 3. Gantt chart diagram of the projects during the internship (part 1). 39
Table 4. Gantt chart diagram of the projects during the internship (part 2). 40
Table 5. Gantt chart of Project A. 43
Table 6. Gantt chart of Project B - SCE. 46
Table 7. Gantt chart of Project C - N2. 52
Table 8. Gantt chart of Project D -Adk. 59
Table 9. Gantt chart of Project E - DP. 62
Table 10. Gantt chart of Project Genie. 71
Table 11. Comparison of Apriori and ECLAT algorithm performance. 85

LIST OF ABBREVIATIONS AND ACRONYMS

AI - Artificial Intelligence

AR - Association Rules

BI - Business Intelligence

CRISP-DM - Cross-Industry Standard Process for Data Mining

ECLAT - Equivalent Class Transformation

ETL – Extract, Transform, Load

HA - Hearing Aid

HCP - Hearing Care Professional

JSON - Java Script Object Notation

KDD - Knowledge Discovery in Datasets

ML - Machine Learning

R&D - Research and Development

SaaS - Software as a Service

SEMMA - Sample, Explore, Modify, Model, Asses

1. INTRODUCTION

In 2018 the World Health Organization has estimated that by 2050, over 900 million people (1 in 10 people) will have disabling hearing loss. Hearing loss may result from genetic causes, complications at birth, certain infectious diseases, chronic ear infections, the use of particular drugs, exposure to excessive noise, and ageing. Hearing loss, might have a big, but underestimated impact on one's life: (i) functional impact, by impacting on the individual's ability to communicate with others; (ii) social and emotional impact, caused by isolation and frustration; (iii) economic impact - hearing loss poses an annual global cost of US $ 750 billion (health sectors costs, educational support, loss of productivity, etc.).

As a countermeasure, people with hearing loss can benefit from the use of hearing devices, such as hearing aids, cochlear implants and other assistive devices (World Health Organization, 2018). These trends provided a fertile soil for a sector in growth. A market analysis made by *Orbis research*, shows that the "Global Hearing Aid Devices Market" is estimated to witness a CAGR of 5,1% during the forecast period of 2017-2023 (OrbisResearch, 2017) with one of the main players being the William Demant Holding A/S.

Demant, is the company behind the commercial successes of world-renowned brands like Oticon, Bernafon, Sonic, Oticon Medical, Maico, Interacoustics, Grason-Stadler and Sennheiser Communications (figure 1). The Group operates in a global market with companies in more than 30 countries, has a total staff exceeding 13,000 and generates annual revenues of more than 1.75 billion Euros. Event though the group leads a multi-brand strategy, its main subsidiary is Oticon A/S - one of the market's biggest hearing care solutions manufacturers - which revenues 87% of the business' activities (Oticon, 2018) (William Demant, 2016).

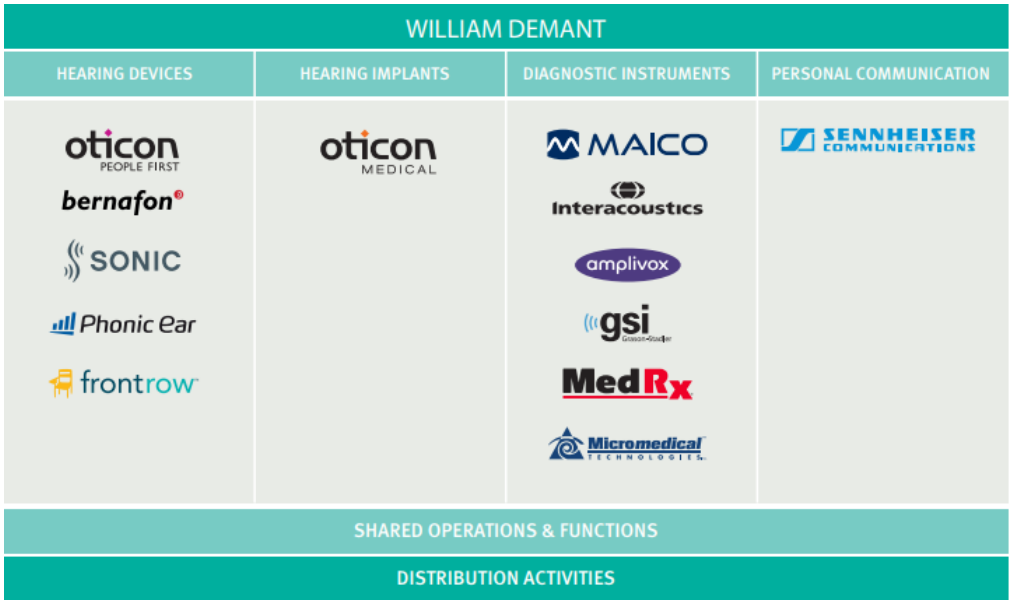

Figure 1. William Demant Holding – brand umbrella. Source: www.demant.com

This paper will describe the author's ten months internship experience - from August 2018 to May 2019 - in the Market Intelligence team of Oticon A/S.

The internship report is presented as a as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence.

During the internship, the author had the possibility to work on a variety of projects requiring the use of different tools and methodologies in order to reach the established goals. Based on specific tasks, and in order to create meaningful insights from data, market research, reporting solution design and data mining techniques were used. In this way aiming to demonstrate that tools and knowledge domains should not be treated as separate silos but combined in order to create knowledge and lead to data-driven decisions.

This internship report will start by giving an overview of the company, the team and the motivations behind the internship choice. Afterwards, the theoretical framework needed during the projects will be presented. In conclusion the report will describe the projects done by the author by dividing them in three categories: (i) Market Research projects; (ii) Reporting solution development project; (iii) Data Mining project. After the projects' description, a brief conclusion will be made together with the lessons learned during the internship.

1.1. INTERNSHIP COMPANY OVERVIEW

With more than 600 patents registered, Oticon A/S (Patent Justia, 2018), strives to mix the most sophisticated technology and audiology, to meet the needs of the hearing-impaired people. It is one of the largest HA manufacturers in the world, based in Denmark.
Since 1904, the company has seen its international growth. At the moment, Oticon A/S (figure 2), contributes to William Demant Group revenue by 87 % of its total. Oticon focuses on manufacturing hearing care solutions, enabling people to be fully reintegrated into their social lives

Figure 2. Oticon logo. Source: www.oticon.com

The company is the home of many industry firsts, including the first fully automatic hearing instrument, the first digital hearing instrument, the first hearing instrument with artificial intelligence, one of the first instruments to use wireless binaural technology and the first design-focused hearing instrument (Oticon, 2018). Oticon's products are represented by a range of products commonly known on the market, designed for different types of hearing loss, personal needs and preferences. The hearing aids are differentiated by different functions, power of the sound, dimensions and technological generations. The range of HA instruments offered by the company is

represented by: (i) BTE, Beside The Ear, (ii) miniRITE-T, mini Receiver In The Ear - Telecoil, (iii) miniRITE, (iv) ITE - FS, Inside The Ear - Full Shell, (v) ITE HS - Inside The Ear, Half Shell, (vi) ITC, Inside The Canal, (vii) CIC, Completely In the Canal, (viii) IIC, Invisible In the Canal (figure 3).

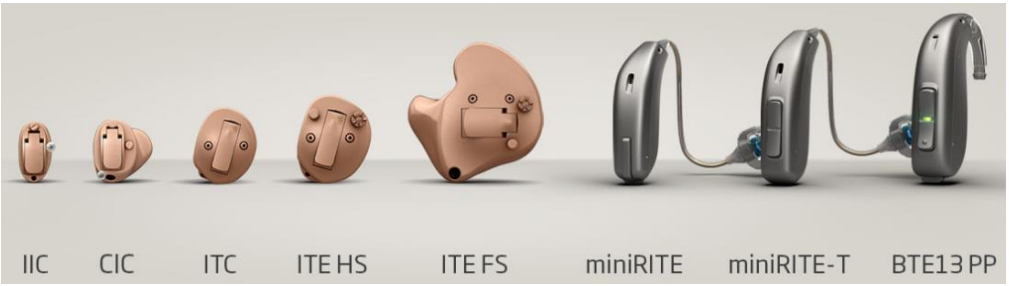

Figure 3. Oticon range of hearing care solutions. Source: www.oticon.com

In order to develop HAs in a way to effectively help the end users it is essential to understand their needs and problems. Market research allows to quantify the end users' behavior, needs, problems and create insights that can be transformed in customer-centric business decisions. Furthermore, market research can be used in order to understand how provide better and more efficient services that run behind the selling of a HA. Even more data related to product usage and user behavior is available now that the HAs can be connected to apps and fitting processes done through software connected to internet. The availability of even more data gives the possibility to get limitless insights on dimensions such as usage of the HIs, app usage, understanding of the use of the fitting software and involved processes, and much more. All insights leading to a better understanding of the people for whom the company strives to create value.

Different type of data is analyzed internally by different teams, based on the type of data, type of insights to be generated and types of decisions to be taken based on the latter. The goal of the Market Intelligence team is that of leveraging the data, either primary or secondary, to deliver customer centric insights.

1.2. INTERNSHIP OVERVIEW – THE TEAM AND ACTIVITIES

The internship experience took place in the Market Intelligence Team of Oticon - a team part of the InMarket and Customer Experience department. The team is formed by four analysts, one of which being the intern. The team's main role is that of providing the company with data driven insights about existing products, potential products, concepts, prototypes, marketing processes, user behaviour and other customer-centric business needs. The insights are provided most of the times through quantitative, qualitative or mixed research. The team works usually with decision driven research problems, where the researcher doesn't define himself the research problem and hypothesis, but the latter are defined by another stakeholder. Furthermore, the team may analyse, when available, secondary data and integrate the results with the findings from the research. The secondary data is retrieved from company's databases which store information from a variety of sources (apps, software usage, product usage, sales, retail data, etc.). Based on the specific problem

the appropriate research method is chosen by the team and applied in order to answer the problem (most of the times the research being exploratory).

The researchers are making use of methodologies such as online surveys and interviews to explore and answer the problem at hand. Statistical software such as IBM SPSS is used for the statistical analysis of the collected data, while tools such as PowerBI, PowerQuery, R and MS SQL Management are used to retrieve, visualize and analyse secondary data.

During the internship, the student had therefore the opportunity to work on different types of projects, having the chance to find the best methodology and tool to reach the project goals. The projects ranged from doing descriptive research using online survey, to the development of a reporting solution, to the application of association rule mining algorithms to large datasets.

Due to the sensitive nature of the data, in this internship report no sensitive data will be disclosed – data such as sensitive variable names, specific data sources and information that could reveal business confidential information. Nevertheless, each project will be introduced with an own background that will allow the reader to have a holistic overview of the project, project methodology and project results.

1.3. INTERNSHIP MOTIVATION

From a practical point of view, the internship was meant to be an opportunity for the student to apply his acquired theoretical knowledge during the Knowledge Management and Business Intelligence Master studies at NOVA University of Lisbon - faculty of Information Management Business school.

Additionally, it was of great interest the possibility to become part of global team operating in the health care industry and a team that uses a mix of tools and methodologies to extract knowledge out of data and transform it into insights – destroying the silos usually existing between statistical analysis, business intelligence and data mining. Nevertheless, from the point of view of value creation, the company's mission of helping the hearing-impaired people in having a normal live was an essential motivation to purse the internship at Oticon.

1.4. INTERNSHIP DEFINITION AND GOALS

The main goal of the internship was that of forming an analyst able of using a mix of tools and techniques in order to provide relevant customer centric insights. Through designing and developing online questionnaires, scoping and managing research projects, querying large datasets for data mining applications and building insightful data visualizations; the analyst intern would have gained confidence in performing market research and data analysis in various business environments, in order to answer different business and research problems.

2. THEORETICAL FRAMEWORK

2.1. MARKET RESEARCH

Market research is a process of gathering, analyzing and interpreting information about a market, product, service or customer (Al-Shatanawi, Osman, & Ab Halim, 2014). It is done to gather insights about customers' characteristics, habits, needs, feelings, market and whole industry. Webb (2003), Kothari (2004), (McCusker & Gunaydin, 2014) and Kumar (2014) mention different types of market research that can be performed. It therefore can be classified as follows, sometimes putting the different market research types at two opposite sides of a spectrum:

- **Descriptive research** *vs* **Analytical research:**
  - Descriptive research: used as well with the term *Ex post facto,* is applied usually when the goal of the research is that of obtaining a descriptive picture of the reality. One of the main features of this type, is the researcher having no control over the variables to be observed. Descriptive researches can include attempts of the researchers to understand the cause of the identify things. The most common methods used for this type of research are survey methods.
  - Analytical research: the researcher makes use of the already available facts or information.
- **Applied research** *vs* **Fundamental research**:
  - Applied research has at its core the finding of a solution for an immediate problem.
  - Fundamental research strives to formulate a general theory over a matter.
- **Quantitative research** *vs* **Qualitative research**:
  - Quantitative research is a type of research based on the measurements of *quantities*. It aims to quantify a certain problem by generating numerical data that can be used afterwards to calculate statistics. It is used to quantify perceptions, behaviors, opinions, and consequently generalize the results for a larger population.
  - Qualitative research is used to gather non-numerical data in order to explore and answer the "why" and "how" of research problems. This type of research is used to produce explanations of the particular problem under study, while a generalization of the study is to be considered a tentative proposition.
- **Conceptual research** *vs* **Empirical research**:
  - Conceptual research is a research related to abstract ideas or theory. Usually used to develop new concepts or to reinterpret existing ones.
  - Empirical (or experimental research) research is a type of research that relies exclusively on experience or observation. It is a data-driven research. This type of research is characterized by the researcher's control over the variables.
- **One-time research** *vs* **longitudinal research:**

- One-time research is a research confined to a single time period.
- Longitudinal research is research repeated more times over time.

❖ **Conclusion-oriented research** *vs* **Decision oriented research:**

- Conclusion-oriented research, the researcher is free to pick up a problem, design the enquiry as he proceeds and is prepared to conceptualize as he wishes.
- Decision-oriented research is always done for the need of a decision maker and the researcher is not free to manage the research process based on his own inclination.

When performing market research, one should consider the research approach to be adopted as well. For instance, McCusker and Gunaydin (2014) and Kothari (2004), mention three research approches:

1. **Quantitative approach**, which generates quantitative data that can be subject to quantitative analysis. This approach can be consequently sub-divided in:
   - ❖ Inferential approach through which data is collected in order to infer characteristics or relationships of the population. Usually implies survey research.
   - ❖ Experimental approach which implies a greater control over the research environment and allows for manipulation of some variables in order to observe the effect on other variables.
   - ❖ Simulation approach, which involves the setting of an artificial environment where artificial data and information can be generated.
2. **Qualitative approach,** which usually considers the subjective assessment of attitudes, opinions and behavior. The qualitative research can be performed as a function of researcher's insights and impressions. Normally the generated data can't be considered for rigorous quantitative analysis. Techniques such as focus group interviews, projective techniques and depth interviews are used.
3. **Mixed methods approach,** McCusker and Gunaydin (2014) which makes use of the both above mentioned approaches. Usually qualitative research is to be done prior to quantitative research, hence the first comes to enforce the second.

Even before starting research project the researcher has to have a clear distinction between two fundamental concepts which at the same time are both different and complementary. These two key concepts are those of *Research Methodology* and *Research method.*

❖ **Research methodology**: is the way of systematically answering the research problem. It defines how the research shall be done scientifically. The research methodology describes the main steps followed by the researcher in studying the research problem. By defining the research methodology, the researcher has to understand and explicit the assumptions udnerlying various techniques, it requires to be aware of the criteria by wich a certain technique is to be applied so to answer a specific research problem. In other words, when defining the methodology, the research

decisions have to be described, specifying the *what,* the *why,* and the consequences of each decision. The methodology comprehends the hypothesis and the logic behind each step done.(Webb, 2003),(Kothari, 2004), (Kumar, 2014).

- **Research method**, can be seen as the choice over the methods and techniques used for conducting research. Examples can be analysis of historical data, non-participant direct observation, questionnaire, personal interviews, focused interviews, group interview, etc. (Webb, 2003),(Kothari, 2004), (Kumar, 2014).

  Additionally, Kothari (2004) classifies the research methods in three groups:

  1. Methods aiming to *data collections*. They are used when the already available data is not enough to arrive at the required solution the data already available are not enough to arrive at the required solution.
  2. Methods using *statistical techniques* to identify relationships within the data and between the data and the research problem.
  3. Methods used to evaluate the accuracy of the obtained results.

2.1.1. Market Research Steps

The literature identifies different market research project steps and their variants, nevertheless the most common and mentioned steps are the eight ones described below (Al-Shatanawi, Osman, & Ab Halim, 2014), (Kumar, 2014), (Verschuren & Doorewaard, 2010), (Kothari, 2004).

1**) Definition of research problem.** Consists in the identification of the area of interest that has to be studied and the statement of the objective of the research. This step will consequently influence the definition of all the other incoming steps.

2) **Review of the literature**. This step has to be done in order to define the *status quo* of the literature on similar research problems and projects. It helps the researcher to frame the problem in a broader picture and get an overview of how similar problems were approached by fellow researchers. It might help identify potential limitations, difficulties and solutions of the research project. It might help to see the problem under different point of views, enriching and targeting better the research project.

3) **Hypothesis formulation**. During this step a tentative assumption is formulated in order to draw out and test its logical or empirical consequences. Hypothesis formulation provide the focal point for the research, i.e. defining the research hypothesis helps to target the research project, identify the right tools, methods and techniques to test the hypothesis and answer the research quest.

4) **Research design**. It is about defining the conceptual structure within the research that would be conducted. The research design highly depends on the purpose of the research that might be for example explorative, descriptive, diagnostical, experimental. Kothari (2004) advices to consider the following aspects when defining the research design:

i. The means of obtaining information;

ii. The skills possessed by the researcher and team;

iii. Explanation and definition of the means selected for obtaining information;

iv. The time available for the research;

v. The cost factor relating to research;

5) **Definition of the sample design**. Since in many researches it is quite impossible to run a research project on the whole population, the researcher has to identify an appropriate sample from the population that would allow to extract enough accurate insights and represent the population. During the sample design it must be considered taking into account the nature of the research problem and other related factors. Kothari (2004), classifies the sampling techniques into two major categories: *probability* and *non-probability* sampling.

Examples of *probability sampling* are the following techniques:

i. Simple random sampling: each item in the population has an equal chance of inclusion in the sample.

ii. Systematic sampling: each item in the population has a certain probability of being selected and a selection pattern is defined in order to select the sample items.

iii. Stratified sampling. when the population does not represent a homogeneous group, stratified sampling is required in order to obtain a representative sample

iv. Cluster/area sampling involves grouping the population and then consequently selecting specific groups or clusters rather than individuals.

Examples of *non-probability* sampling are considered to be:

i. Deliberate sampling, which involves deliberate selection of specific units from the population.

ii. Convenience sampling is a sampling technique with which deliberate sampling is done by selecting the sampling items based on their ease of access.

iii. Judgement sampling is a method that implies the use of researcher's judgement to select the sample in such a way to make it representative of the population.

iv. Quota sampling techniques require quotas to be given and to be filled in from different strata. The size of quotas usually is proportionate to the size of the stratum in the population

Other sampling methods could be multi-stage sampling and sequential sampling.

6) **Data collection**. Data is required to answer the research problem. The quality of the data will greatly influence the research results and the insights generated from them. When defining a data collection method, it is essential to consider elements such as

nature, scope and objective of the research problem, available budget for the research project, time and other resources available.

7) **Data analysis**; is the phase of the project during which the collected data gets structured, cleansed and categorized, coded and tabulated to draw statistical inferences. This phase will help to transform simple data into knowledge. Consequently, depending on the goal of the research, the appropriate data analysis techniques and statistics have to be used.

8) **Interpretation and report**. This last phase consists into structuring the research results and presenting them in such a way to facilitate the insights generation process in the audience. Here the initial research problem and hypothesis are answered.

2.1.2. Data collection methods

A review of the market research literature allowed to identify few key methods of data collection. As following below, the key data collection methods across multiple authors and researchers resulted to be the observation method, interview, questionnaire, schedules, consumer pannels, projective techniques and depth interviews. (Kothari, 2004) (Al-Shatanawi, Osman, & Ab Halim, 2014) (Kumar, 2014), (Creswell, 2014), (Webb, 2014), (McCusker, 2014) Each of this methods comes with a variety of techniques that can implemented to answer the research problem. For the purose of this paper only the most used methods were described, despite many more existing and being adopted for market reseach goals.

❖ **Observation method**

With this method the information is collected through researcher's direct observation of the events, where the researcher does not interact with the subject being observed. The main advantages of this type of data collection method is it being independent from the respondent's willingness to respond, additionally it allows to eliminate respondent's subjective bias. Furthermore, it tracks the present moment and the information is not influenced by past behavior or future intentions or attitudes. On the other hand, the limitations of this method are represented by the fact that it is expensive to run, it might provide little information, and unforeseen factors can interfere with the observational task.

Depending of the conditions under which the observation occurs, it can be either *controlled observation* or *uncontrolled observation*, either *structured* or *unstructured*. Depending on whether the researcher is living the experience personally or not, it can be either *participant observation* or *non-participant observation.*

❖ **Interview method.**

This data collection method involves a presentation of some kind of verbal stimuli by the researcher to the respondent and a verbal response from the latter one. The literature identifies two main types of interviews: *personal* and *telephone interviews*.

- ❖ **Personal interviews.** Implies an interviewer asking questions to anoother person, or persons (generally face-to-face). During the interview the interviewer records and collects the information given by the respondent. When the interview involves the use of a set of predetermined questions and a standardized way of recording, it can be classified as a *structured interview.* Contrary, when the elements for a structured interview are missing, it can be considered as an *unstructured interview.* The latter type of interviewing gives more freedom to the interviewer while at the same time diminish the later possibility of result comparison.

  Focused interviews. This type of interviews is a variant of *personal interviews* and implies the attention of the interviewer to be focused on the given experience of the respondent and its effects. The interviewer is free to decide the manner and sequence in which the questions are presented.

  When using interviews as a data collection method, some advantages can be considered the more information and greater depth that can be obtained, the method gives the interviewer the possibility to use his/hers skills to overcome the respondent's resistance to answer certain questions, it offers a greater flexibility in presenting the questions, Usually non-response rates are very low due to a higher control of the process, the emotions and non-verbal language can be tracked, additional information about the respondent can be collected.

  On the other hand, disadvantages are represented by the method being expensive (both in time and finances) especially when having a geographically spread sample, possibility of the bias of the interviewer, the presence of the interviewer can overstimulate the respondent.

- ❖ **Telephone interviews.** A method which consists in running interviews via phone. It lost its popularity with the more frequent use of online questionnaires. Nevertheless, some of the advantages of this method are that it results to be cheaper than face-to-face interviews, helps to reach respondents that are spread geographically, respondents might feel more secure to release more sensitive information via phone. On the contrary, some disadvantages of the method are that this type of interviewing gives little time to run the interviews, it might be difficult to elaborate on more complex questions, probes are difficult to handle.

With the increasing trend of using computers to assist the process of data collections through interviews, four types of computer-assisted interviewing processes emerge - the so-called: CAPI (Computer-Assisted Personal Interviewing), CASI (Computer-Assisted Self-Interviewing), CATI (Computer-Assisted Telephone Interviewing), and CAWI (Computer-Assisted Web Interviewing). These methodologies reflect the above described approach, with the difference of researcher or respondent being assisted by a computer device.

❖ **Questionnaires.**

It is so far the most widely adopted method of primary data collection. It used by a multitude of industries to gain insights on specific research problem. Questionnaires are represented by a number of questions that are typed in a certain order on a form. Consequently, the questionnaire is shared with the respondent who autonomously answer the questions written in the questionnaire.

The main advantages claimed on behalf of the questionnaires can be listed as follows: 1) It is a low cost method to gather primary data; 2) Allows to reach a large sample widely spread geographically; 3) Lowers to the minimum the bias of the interviewer, since the respondent answer with its own words; 4) Respondents can take all the time they require to answer the questions; 5) It can ensure a high degree of answer confidentiality, therefore the respondents may be more propense to give information on sensitive topics; 6) Large samples are easier to manage.

Nevertheless, the questionnaires may present the following drawbacks: 1) Depending on the sample and research problem, there might be present a medium-to high risk of low response rate; 2) It is a method that relies completely on the cooperation of respondents, despite requiring them to be educated on the tool and topic; 3) Once the questionnaire has been sent out, there is nothing that the researcher can do in case of any detected errors in the questionnaire structure or questions; 4) In case of open-ended questions there is a risk of ambiguous replies; 5) In case of omitted questions, the interpretation of the omission may result difficult.

Considering the drawbacks of the questionnaires, it is a common use to run pilot studies when large samples are involved. A pilot study is a replica of the data collection method done on a smaller representative sample in order to test details such as response rate, errors, omission tendency etc. Most of the times a pilot study would allow to target better the process and leverage the tool as much as possible.

Some main aspects to be considered when designing a questionnaire can be listed as follows:

1. **Form** of the questionnaire.
    - *Structured questionnaire*. Is a questionnaire form where all the questions are definite, concrete and pre-determined. The questions and their sequence are the same for all respondents. This type of questionnaire is simple to manage and allows to gather data from very large samples. The data collected through this method is easier to analyze statistically. However, structured questionnaires are not advisable when a problem is being first explored or when the working hypothesis are sought from the research.
    - *Unstructured questionnaire.* Is a questionnaire form that is provided to the respondent by the interviewer. The interviewer knows the data that has to be collected and has the freedom to formulate that questions in the appropriate way, based on each respondent.
2. **Question sequence**.
    An appropriate question sequence is required in order to ensure the highest quality of the replies. It may as well influence the drop out rate of the respondents, since a good sequence is required to keep the respondents engaged during the answering

process. A proper sequence of the questions expects the questions to follow from the least difficult to the most difficult. Additionally, questions researching behavior should be listed before questions researching attitude this will help to avoid data about behavior biased by attitudes.

3. **Question formulation.**
   The questions should be formulated in a clear, simple and unambiguous way. When formulating the questions, the researcher has to have in mind the audience of the questionnaire and use appropriate understandable terms. To avoid ambiguity the questions should evaluate one item at a time and avoid evaluation on multiple items or problems.

4. **Question type**.

   - *Close-ended questions.*
     This type of questions offers a pre-defined range of answers to the respondent allowing for either single or multiple-choice selection. Close-ended questions are easy to handle and inexpensive to analyze statistically. However, it is not advisable to use this type of questions when complex matters are to be explored. Additionally, they assume to cover all the possible range of answers. When creating closed-ended questions, researcher may make use of scales to analyze degrees of agreement, perceptions, attitudes, etc.

   - *Open-ended questions*.
     Open-ended questions allow the respondent to answer using his or her own words. Despite the use of the latter type of questions being preferable for explorative research, it might bring data that is difficult to analyze and interpret. Furthermore, the interpretation makes space for bias caused by researcher's interpretation. It can be as well hard to compare data collected through open-ended questions. In conclusion, the researcher should aim for the right balance between these two types of questions.

5. **Best practices.**
   Questionnaires should be kept short and simple, using a language easily understandable by respondents. It should contain the right balance between closed-ended and open-ended questions. To ensure the quality of the information and a representative sample it is essential to make use of screening and control questions. The latter two type of questions will help to filter out the respondents that are not representative for the study.

❖ **Schedules**

It is a data collection method similar to questionnaires. The difference stands in the fact that the schedules are being filled in by enumerators. The appointed enumerators face the respondents and record the replies in the space meant for them in the proforma.

❖ **Consumer panels.**

Is a data collection method considering a set of consumers that track and regularly record data on their consumption behavior. These consumers are consequently interviewed by the investigator.

❖ **Projective techniques.**

Is a data collection techniques developed by psychologists, where the latter make use of projections for respondents to infer about underlying motives, urges, or intentions. This type of techniques is performed when the information to be obtained sees the respondents reluctant to give it or simply sees them lacking the knowledge to express the underlying attitudes and/or behavior. In other words, the respondent is supplying attitude and behavior information unconsciously. This technique is usually used in motivational research. It requires though highly specialized expert skills and training. Examples of these techniques are word association tests, sentence completion tests, story completion tests, verbal projection tests.

❖ **Depth interviews.**

Are interviews that aim to bring to surface underlying motives, needs, feelings, desires and attitudes of the respondents. Like the projective techniques, it is a methodology used in motivational research. It aims to uncover unconscious type of knowledge related to personality dynamics and motivations. Running depth interviews requires skilled interviewer and enough time. The difference with the projective interviews is that depth interviews may not necessarily be projective in nature.

Examples of ***other data collection methods*** are content-analysis, pantry audits, use of mechanical devices (data collected through indirect means such as eye camera, pupillometric camera, Psychogalvanometer, etc.).

When using a certain data collection method, the researcher has to be aware of potential sources of errors occurring during the data collection activity that might influence the output of the research. By being aware of the possible sources of error, one can act in order to prevent poor quality data and consequently enhance the quality and the output of the research (Kothari, 2004). Five main sources of errors and poor data can be the respondent himself/herself, the specific situation, the measurer, the instrument used to collect the data, and poor sampling.

1. **Respondent caused errors**. Based on the topic of the research or type of information asked, the respondent may be reluctant to give a truthful answer. Additionally, it may happen that they have little knowledge on the topic and could not admit this lack of knowledge. Other factors like fatigue, boredom, anxiety may limit the ability of the respondent to answer the questions accurately and fully.

2. **Situation caused errors**. Any conditions which strains the interview can have effect on the rapport between the interviewer and respondent. For example, if the

respondent may perceive that the anonymity is not granted, he or she may be reluctant to express certain ideas, feelings or beliefs.

3. **Measurer caused errors**: The errors done by the interviewer during the data collection period may distort the findings. The way in which interviewer makes use of the questions and relates to the respondent might as well influence the collected answers. Errors from incorrect coding, tabulations and statistical calculations can as well distort the findings.

4. **Instrument caused errors**. Errors may arise because of defective measurement instrument. Use of complex logic to design the questionnaires, use of complex words and formulation in the questions, ambiguous meanings, inadequate space for replies, response choice omissions are examples of things that can lead to measurement errors.

5. **Poor sampling caused errors.**

2.1.3. Scaling

When using questionnaires with closed-ended questions an important branch of the latter type of questions is represented by questions using scales. Scaling describes the procedures of assigning number to various degrees of opinion, attitude and other concepts. Questions using scaling acquire importance since they allow the closed-ended questions to evaluate more abstract concepts and perceptions. Scales are used to measure attitudes and opinions; they help to measure abstract concepts more accurately. Commonly in the literature the scale is considered to be a continuum, consisting of the highest point and the lowest point, with several intermediate points between these two extreme points. Usually, a scale-point placed after another one, will indicate a higher degree in terms of a given characteristic (Kothari, 2004). Generally, to types of scaling are cited in the literature: *rating scales* and *ranking scales.*

- **Rating scales.**

  They involve qualitative description of a limited number of aspects of a thing or traits of an object. When using rating scales, an object gets evaluated in absolute terms against some specific criteria. There is no specific rule on the number of points that have to be used for a scale (e.g. two, three, five, seven, nine points). Examples of scale classification with more categories can be "*always-often-occasionally-rarely-never*", "*not at all satisfied- somewhat dissatisfied - neutral - somewhat satisfied - very satisfied*", "*poor - below average - average - good - excellent*", etc.

  Some advantages of the rating scales are the fact that they have a wide range of applications, they may be used with a large number of properties or variables, require less time and are interesting to use. Nevertheless, they work under the assumption that the respondent is capable of making good judgements, otherwise errors in data my occur. Collected data might be incorrect or can represent a

distorted picture of reality. Three type of common errors are known: the *error of leniency*, the *error of central tendency*, the *error of hallo effect*. The first type of error, leniency, comes out when the respondents are either easy or hard raters. The second type of error, the error of central tendency, occurs when respondents are reluctant to give extreme judgements. The hallo effect error is represented by a bias occurring when the respondent carries over a generalized impression of the subject from one rating to another.

- **Ranking scales.**
  The so-called comparative scales are used to make relative judgements against other similar objects. Usually the respondent compares two or more objects and make choices among them.

Based on the goal of the research and the most appropriate data to be gathered, it is important to consider the different types of approaches to formulate the scales and the respective scale types resulting. Therefore, there are several approaches to scale creation mentioned by Kothari (2004).

- **Arbitrary approach**.
  This approach is related to scales that are developed *ad hoc* by the researcher based on his or her own experience, competence and subjective selection of items. This technique fore scale development results to be resource efficient, since it can be easily and quickly implemented. The drawbacks of using this technique is that there is no objective evidence that such a scale will measure the concept for which it has been developed.

- **Consensus approach** (*Differential or* ***Thurstone-type scales***).
  By adopting this approach, the selection of items is made by a panel of judges who evaluate the items in terms of whether they are relevant to the topic area and unambiguous in implication. Differential sales have been widely used to measure attitude towards issues like wars, religion and other major society topics.

- **Item analysis approach** (*Summated scales or* ***Likert-type Scales****).*
  Summated or Likert-type scales consist of a number of statements which express either a favorable or unfavorable attitude towards a given object to which the respondent is asked to react. Using the specific instrument adopted by the researcher, the respondent uses a liker-scale to show a certain agreement or disagreement with each statement. However, other items rather than agreement can be evaluated on a Likert-type scale, examples can be *satisfaction, ease of use, frequency of a certain behavior, helpfulness,* etc. Based on the item considered by the scale, the degrees on the scale change appropriately. Each response is coded with a numerical score, indicating the favorableness or unfavorableness. The respondent has consequently to respond to each of the statements in terms of several degrees. Indeed, Likert-type scales can be created based on three, five, seven or nine-point scale. An example of evaluation of the satisfaction of the respondent with a certain item can be done by using the Likert-type scale in figure 4.

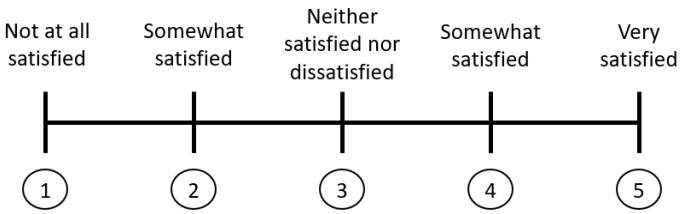

Figure 4. Example of Likert-type "Satisfaction" scale

The score values are normally not printed with the scale. However, when a respondent indicates an evaluation degree, it will be assigned a score. In the end the scores will be used to run descriptive statistics and search for patterns in the answers of the sample. Some of the advantages of using this type of scales are: 1) Likert-type scales are considered more reliable since the respondents answer each statement included in the instrument. This leads to gathering more data and information; 2) They are easy to construct and implement; 3) Are easy to use in respondent-centred and stimulus-centred studies; 4) Require less time to be constructed. On the other hand, limitations are represented by the fact that the points on a Likert-type scale might not be equally spaced in the belief of the respondent (interval between "not at all satisfied" and "somewhat satisfied" might not be the same as the one between "somewhat satisfied" and "very satisfied"). Despite its coding into numerical values, it is often ground for disputes whether it should be considered as an ordinal or interval scale. Additionally, there might the possibility that people may answer according to what they think they should feel rather than how they actually feel.

- **Cumulative approach** (*Cumulative scales such as **Guttman's Scalogram***).
  The cumulative approach consists of a series of statements to which a respondent expresses agreement or disagreement. The peculiarity of this scale is that they are set in such a way that they are related to one another. For instance, the individual who replied unfavorably to item 3, for example, most likely had replied unfavorably to item 2 and item 1. In order to count the respondent's score, the number of points concerning the number favorably answered statements has to be counted. Therefore, once a respondent has answered to a certain number of questions, a pattern can be identified for his future answers.

- **Factor analysis approach** (*Factor scales such as Osgood's Semantic Differential, Multi-dimensional scaling, etc.*).
  This approach is developed through factor analysis or on the basis of intercorrelations of items which indicate that a common factor accounts for a relationship between items (Kothari, 2004). Making use of factor scales turns particularly useful in uncovering latent attitude dimensions and approach scaling through the concept of multiple-dimension attribute space (Emory, 1985).

2.1.4. Data types

In order to answer a research problem, specific mathematical properties can be sought by the researcher. Therefore, it is important to consider the existing types of data scales based on their mathematical properties. So far, the types of data scales can be classified in *nominal, ordinal, interval* and *ratio.*

- **Nominal scale.**
  This type of scale represents data of nominal type. Here the numbers are used to map the nominal values. The ordering of the data does not carry any meaning. The numbering is therefore only used to keep track of the nominal data points. This type of scale has the least powerful level of measurement, since its points don't represent any distance between each other and there is no arithmetic origin. The only statistics that can be run on this type of scale and data are count and mode, the latter representing the most appropriate measure of central tendency. Any other type of statistical analysis would be meaningless. Nevertheless, the use this scale happens frequently when the goal is that of identifying and classifying data in sub-groups (Kothari, 2004).

- **Ordinal scale.**
  The ordinal scales are the ones that assign an importance to the order of the items. Most commonly, this type of scale is used for ranking purposes. Hence it implies statements of order "less than" or "greater than". Even though providing more information that the nominal scales, the distance between the adjacent datapoints or ranks, does not always have an arithmetic meaning. The most appropriate central tendency measure for an ordinal scale is the median. Additionally, percentiles can be used to measure dispersion. This type of scale can be considered when performing correlation statistics and are restricted to non-parametric methods (Kothari, 2004).

- **Interval scale.**
  When using interval scales, the intervals between data points are defined to be equal based on a specified rule. The rule is true when the assumptions holding the rule are accepted. The first difference from the first two type of scales, is that interval scales hold an arbitrary zero (hence they don't have an absolute zero). This implies that an interval scale cannot measure the complete absence of a trait or characteristic. An example of an interval scale is the temperature scale. Once can state that 60° are double than 30° but cannot state for instance that at 60° the temperature is twice as warmer than at 30°. The ratio of the two temperatures cares no meaning. Nevertheless, interval scales carry more arithmetical concepts since it incorporates the concept of equality of interval. Hence, more powerful statistical measures can be considered for analysis. Mean is usually used as a central tendency measure and data dispersion is measured by standard deviation. Additionally, correlations and statistical significance tests such as "t-test" and "F-test" can be run on this type of data (Kothari, 2004).

- **Ratio scale.**

  Ratio scale and ratio data type represent the highest level of measurement and carry the highest amount of arithmetical information. First of all, compared to the above cited type of data and scales, it has an absolute zero; therefore, it can measure the complete absence of a trait. The complete absence of the trait carries a meaning (for example a speed equal to zero represents the absence of movement). At the same time, a ratio scale englobes the concept of equal interval between data points. These properties make it possible both to calculate meaningful distance between data points as well as ratios. Generally, all statistical analysis techniques can be applied to ratio scales. It creates the possibility to use additional measures of central tendency, like geometric and harmonic mean, and more coefficients of variations may as well be calculated (Kothari, 2004).

2.2. DATA MINING

"In God we trust, all others bring data"

(Edwin R. Fisher, 1978)

In our days, Data Mining become a buzz word and it is acquiring more and more importance due to the immense universe of data standing behind it. This chapter aims to give an overview of the main data mining concepts and techniques through a thorough literature review. Since data mining and analysis of data *per se* without a specific goal don't bring any value, this chapter will represent as well a bridge that will lead from pure data mining concepts to their application to the specific projects described in this paper.

Before data mining, there was statistics. It was a statistics' task that of understanding the reality, analysing data and trying to obtain valuable knowledge out of it. The need of placing knowledge on a systematic evidence base drove the use of this science. In the past the main issue regarding data was not certainly having too much of it but having too little. Lately, new type of problems, resources to analyse them, amount of data and expansion of frontiers of machine learning, artificial intelligence and database management, switched the attention from statistics to data mining. The goal was switched from extracting knowledge from each *rare datum*, to making sense, instead, of quantities of data so large that it is beyond the human ability to comprehend the data in its raw format (Grover & Mehra, 2008). Because it sits at the frontier between computer science, statistics, data visualization, database management, artificial intelligence; the definition of data mining changes with the background and view of the one who defines it. These are just some definitions cited by Friedman in one of his papers (Friedman, 1997):

- *"Data mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data".* - Fayyad.
- *"Data mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions"*. - Zekulin.
- *"Data Mining is a set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data".* - Ferruzza.
- "*Data mining is the process of discovering advantageous patterns in data*". - John
- Data mining is a decision support process where we look in large data bases for unknown and unexpected patterns of information. - Parsaye

Even though data mining embeds some of the statistical tools and techniques for data analysis and many data mining tools can be considered nothing else than multivariate statistical data analysis methods (Kuonen, 2004), there are substantial differences that separate the two:

- Statistics deals mainly with primary data while data mining, with secondary data, which is already collected and stored in data bases.

- One of the main concerns in statistics is how to make statements about a population when one has observed only a sample. On the other hand, a data mining problem can have the entire population at hand and its goal would be to find interesting patterns in it.
- Classical statistics deals with numerical data, while nowadays database have numerous different formats of data (text data, audio and video data, geographical data, etc.
- Statistics usually deals with inferring from small samples, while data mining makes use of large datasets containing millions or billions of rows and hundreds or thousands of dimensions.

(Grover & Mehra, 2008), (Kuonen, 2004), (Hand, 1998).

2.3. BIG DATA

Since the quantity of available data can be considered one of the main features differentiating between statistics, it would be nice to specify the concept of big data and what does it comprehend. When speaking about big data, in simplistic terms, we speak about the 5 "V"'s that characterize big data: Volume, Velocity, Variety, Variability and Value (Jain, 2016). These 5 terms lead us to find new approaches in analyzing and treating data in order to convert information into knowledge.

Different approaches and processes were defined and reviewed in time to establish or, at least suggest, a path that would lead to the transformation of the information into knowledge (Matheus, Chan, & Piatetsky-Shapiro, 1993) (Brachman & Anand, 1996) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) (Chapman, et al., 2000) (Mariscal & Marbàn, 2010).

2.4. KNOWLEDGE DISCOVERY IN LARGE DATASETS

Considering that the main goal of an analyst is that of potentially discovering useful knowledge for the company's business from a dataset containing business related data; it could be of great value to ununderstand better how this process is treated. For the latter mentioned reason, a review o the literature was done in order to create a rapresentative picture of how the process of knowledge discovery happens and what are its main methodologies.

The term describing the process of discovering knowledge in datasets, called "K*nowledge Discovery in Databases (KDD)"*, was firstly coined in 1989 (Piatetsky-Shapiro, 1991) to highlight that knowledge is the end products of a data-driven discovery. KDD is defined as a nontrival process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). It comes to be described as a process, because it is composed from several

steps (figure 4), that start with the data stored in a database and end with it being transformed in knowledge. It is meant to be seen both as an interactive and iterative process, allowing the user to partecipate in many decision during the process development.

One of the key steps, through which information is transformed in knowledge, is the *Data Mining* step. Because of its importance and value in the process, the scientific world started to associate the *Data Mining* step to the process of KDD itself, making the two terms synonyms and leading to a confusionary definition of the term (Han, Cai, & Cercone, 1992), (Clifton, 2017), (Azavedo & Santos, 2008). Lately, the term *data mining and knowledge discovery* has been proposed as the most adequate name for the overall process of KDD (Mariscal, Marbàn, & Fernàndez, 2010)

If being considered one of the steps of the KDD process, then Data Mining consists in applying different machine learning algorithms, pattern recognition and statistiscs to a dataset in order to extract patterns from data which, if consequently interpered, can become actionable knowledge.

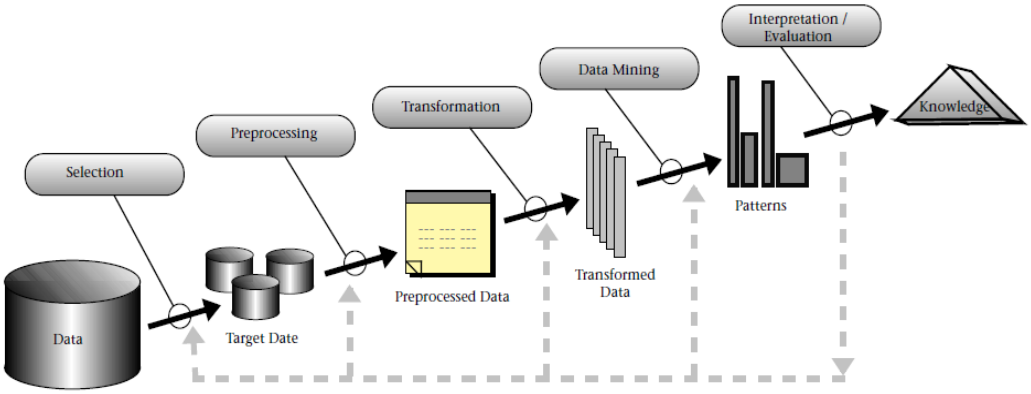

Figure 5. Process of Knowledge Discovery in Databases. Source: Fayyad, Piatetsky-Shapiro, & Smyth, 1996.

The definition of the process of KDD, led it in time to become a benchmark and ispiration for other methods used for knowledge discovery (figure 5) (Mariscal, Marbàn, & Fernàndez, 2010). The process itself, can be seen as an initial approach. Subsequently, as researched by Mariscal et al., the two other main methodolgies which evelved from the original one and dominate nowadays the market, result to be SEMMA and CRISP-DM.

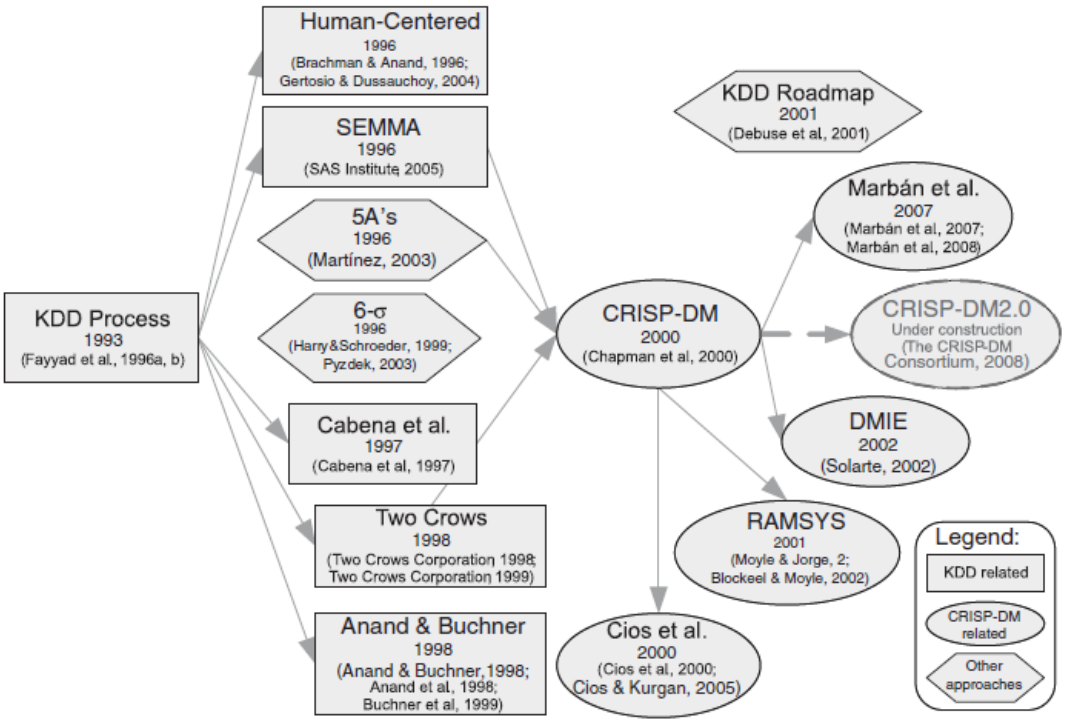

Figure 6. Universe of Data Mining and Knowledge Discovery methodologies. Source: Mariscal, Marbàn, & Fernàndez, 2010.

SAS Institute defines SEMMA as an acronym to describe the SAS data mining process. For instance, SEMMA stands for Sample, Explore, Modify, Model and Assess. It proposes an arrangement of tools provided by the software that allows the user to run through a knowledge discovery process using SAS Enterprise Miner (SAS Institute, 2014). The main difference between SEMMA and KDD process, is that SEMMA is linked to the tool present in SAS Enterprise Miner and it is unlikely to use SEMMA methodolgy out of it. While flexibility of KDD, allows its application in different business or organizational environments and its application on different types of projects.

In response to common needs in data mining projects, a group of organizations involved in data mining projects (Teradata, SPSS, Daimler-Chrysler and OHRA), proposed a guide to manage the data mining projects : CRISP -DM - standing for CRoss Industry Standard Process for Data Mining (Mariscal, Marbàn, & Fernàndez, 2010). AN important factor of CRISP-DM's success, is the fact that it is industry-tool-and application-neutral (Chapman P. , et al., 2000).

CRISP-DM, consists of six phases (figure 6):

1. Business understanding;
2. Data understanding;
3. Data preparation;
4. Modelling;
5. Evaluation;
6. Deployment.

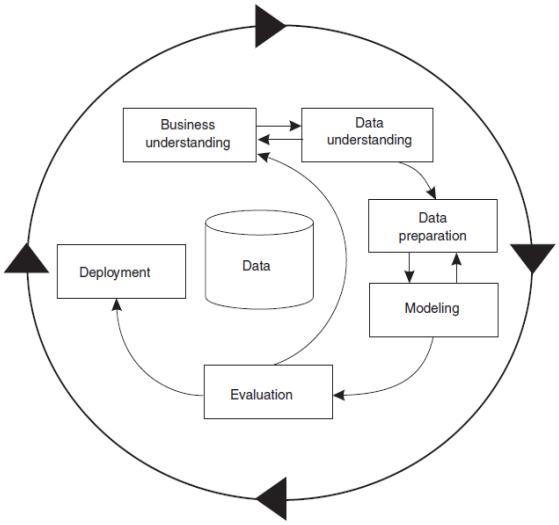

Figure 7. CRISP-DM process. Source: Chapman P., et al., 2000.

The latter methodology describes and provides an overview of the life-cycle of a data mining project. It contains the corresponding phases of a project, their respecitve tasks, and relationships between these tasks. Since this methodology abeard and put roots in the market and industries, these same industries and their business needs continued to change and new ones to emerge: new types of data, integration and deplyment of results with operational systems, more demanding requirements for scalability, need to mine large scale databases *in situ,* etc. These new trends and needs forced CRISP-DM to adapt, therefore new versions of it emerged - CRISP-DM 2.0. It has to be mentioned that despite its newness, it has not been fully accepted by the industries yet.

Compared to KDD, CRISP-DM has a fewer number of steps, nine of the KDD compared to the six of CRISP-DM, but this factor on the other hand gives the latter one a higher flexibility and adaptability to new envirnoments and a very broad range of applicability on different types of projects. Because of its popularity and adoption, it is considered *de facto* the standard method for developing data mining and knowledge discovery projects.

Despite being a broadly adopted methodology, the methodology come with some weaknesses as well. Some of which being the following:

- Lack of project management processes (such as choosing a lify cycle type for the project);
- Lack of integral processes (which may assure eventually a project's completeness and quality);
- Lack of organizational processes (that eventually can help to achieve a more effective organization).

(Mariscal, Marbàn, & Fernàndez, 2010).

Despite the flows of the methodology, CRISP-DM was chosen to be used for the data mining project described in this internship report. To compensate for the flows of the methodology, a thorough quality check and process management will be done internally by the managers of the company.

2.5. MACHINE LEARNING

Machine Learnings is a highly interdisciplinary field of studies involving concepts from statistics, computer science, engineering, cognitive science, mathematics, optimisation theory and many other fields (Ghahramani, 2003). It is the learning and building of algorithms that can learn from and make predictions on datasets.(Simon, Deo, Venkatesan, & Ramesh, 2015).

> "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, imporves with experience E" – Tom Mitchell, Carnegie Mellon University.

The machine learning field can be divided in three main underlying types of learning: (1) Reinforcement learning; (2) Supervised learning and, (3) Unsupervised learning.

- In **Reinforcement Learning**, a machine interacts with its environment by producing actions. Based on the effect produced by the actions on the environment, the machine receives rewards. Its future actions will aim to maximize the future rewards. In a different variant of reinforcement learning combines the game theory, by letting the machine to interact in a dynamic environment with other machines.

- In **Supervised Learning**, the machine learns by having a given sequence of desired outputs. There is a specified set of classes, and training data is labeled with the appropriate class. In this way the machine can learn on the errors made when being trained. The main goal is that of producing the correct output given a new input. Numerous unsupervised learning techniques have been developed over time based on Artificial Intelligence and Statistics (Kotsiantis S., 2007). The most common techniques are Decision Trees based, Perceptron-Based techniques (ANN), Bayesian Networks, Instance-based learning (k-nn), Support Vector Machines.

- The fird type of learning is **Unsupervised Learning**. In this type of learning the machine receives inputs, but obtains neither supervised target outputs, nor rewards from its environment. The machine learns to organize information without providing an error signal to evaluate the potential solution (Sathya & Abraham, 2013). The goal in this type of learning is that of creating a representation of the input that can be used for decision making, predicting future inputs, efficiently communicating with other machines. Through unsupervised learning the machine tries to find patterns in data that otherwise would be considered pure unstructured noise. Often, the goal is to decide which data should be grouped together by creating similar groups. The better the developed algorithm the more efficiently one can compress and communicate new data. (Ghahramani, 2003). Some of the most common unsupervised learning techniques are: Clustering (k-means, Gaussian mixtures as Soft k-means, K-medoids, Hierarchical Clustering, Self-Organizing Maps, Principal Components, Spectral Clustering, Kernel and Sparce Principal Components, Multidimensional Scaling), Association Rules or Frequent Pattern Mining.

2.6. ASSOCIATION RULES

In data mining, associacion rules are a well known set of machine learining algorithms that allow for discovery of positive relationships between variables in big datasets. It aims to extract interesting correlations, frequent patterns, casual structures or associations among sets of items (Sethi & Mahajan, 2012) (Arora, Bhalla, & Rao, 2013) (Bathla & Kathuria, 2015).

Association rules are used to find relationships between objects that are frequently used together (Kumbhare & Chobe, 2014).

Association rule mining can be seen as a the process of discovering interesting and unexpected rules from large data sets. Tipically association rules are an implication of IF-THEN-rule, supported by data, trying to identifiy how a subset of items influences the presence of another subset (Ghosh, Biswas, Sarkar, & Sarkar, 2012).

Generally, the association rule can be expressed as:

$$A \Rightarrow B$$

Where A and B are sets of items, A being the antecedent and B being the consequent. The intuitive meaning of the rule is that transactions of the database which contain A, tend to contain B. A real world example would be:

$$\{ \text{Bread} \} \Rightarrow \{ \text{Milk} \}$$

The above rule can be interpreted as "*customers who buy bread, tend as well to buy milk".* Considering that datasets can potentially contain milions or billions of rows, there is the risk that the algorithm will end up finding a quantity of rules to large to be analyzed or too complext to be interpreted by the researcher. Since the goal is that of finding meaningful patterns and interesting rules, there is need to set some parameters that would help through the research of the most interesting and insightful rules. Some basic measures of interestingness are: support and confidence. Only patterns that respect the minimum thresholds of the support and confidence are considered for the analysis.

Let $I = \{ I_1, I_2, \ldots, I_m \}$ be an itemset. Let *D*, be a set of database transactions where each transaction *T* is a nonempty itemset such that $T \subseteq I$. Each transaction is associated with an identifier, called TID. Let *A* be a set of itms. A trasaction *T* is said to contain *A,* if $A \subseteq T$. An association rule is an implication of the form A => B, where $A \subset I$, $B \subset I$, $A \neq \emptyset$, $B \neq \emptyset$, and $A \cap B = \Psi$.
The rule A => B holds in the transaction set *D* with support *s*, where *s* is the percentage of transactions in *D* that contain *A* U *B* (i.e. both A and B). This is take to be the probability P(A U B). The rule *A => B* has confidence *c* in the transaction set *D,* where *c* is the percentage of transactions in *D* containing *A* that also contain *B.* This is taken to be the conditional porbability, P(A | B).

$$\text{Support}(A \Rightarrow B) = P(A \cup B)$$
$$\text{Confidence}(A \Rightarrow B) = P(A \mid B)$$

Rules that satisfy both a minimum support threshold and minimum confidence threshold are called strong.

A set of itmes is referred to as an itemset. An itemset that contains *k* items is a *k*-itemset (e.g. {milk, bread} is a 2-itemset). The occurrence frequency (i.e. frequency, support count, count) of an itemset is the number of transactions that contain the itemset. If the support of an itemset *I* satisfies the corresponding minimum support threshold, then *I* is a frequent itemset.

Generally, the association rule mining can be summarized into two main steps:

1. Finding all frequent itemsets. By definition, the ones that satisfy the minimum support threshold.

2. Generate strong association rules from the frequent itemsets. By definition, these rules must satisfiy both minimum support and minimum confidence threshold (Han, Kamber, & Pei, 2012).

2.6.1. Association rules evaluation methods

Even though setting minimum threshold for support and confidence, there still may emerge a huge number of patterns to be analysed. From the immense ocean of rules and patterns that can emerge after using association rule mining, there might be few spaceleft for uderstanding, interpretation and knowledge extraction, since the rules can become very complex and contain hundreds of items. Furthermore, it becomes highly difficult to spot those rules that might be interesting for the researcher and could bring value for the company.

Whether a rule results to be interesting can be assessed either subjectively or objectively. The former one, depends on the experience and domain knowledge of the researcher. For the latter method, the literature propose some additional objective evaluation methods that can help the researcher in finding interesting and relevant patterns.

Han, Kamber, & Pei (2012) mention correlation to be used additionally to support and confidence. From different correlation measures they propose:

1. Lift.

This measure is based on the claim that the occurrence of itemset *A* is indipendent of the occurrence of itemset *B* if P(A U B) = P(A)(B); otherwise, itemsets A and B are dependent and correlated as events.

$$lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)}.$$

If the lift is less then 1, then the occurrence of *A* is negatively correlated with the occurrence of *B.* If the resulting value of the lift is more than 1, then *A* and *B* are positively correlated, meaning that the occurrence of the one implies the occurrence of the other. Finally if the lift is 1, then *A* and *B* are independent and there is no correlation between them

2. Chi squared.

Is a technique that allows to gauge the degree of dependence between the variables *A* and *B.* The Chi squared method is based on the comparison of observed frequencies and the corresponding expected frequencies. It is used to test the significance of the derivation from expected values.

$$\chi^2 = \Sigma \frac{(observed - expected)^2}{expected}$$

The test can be used to calculate the p-value by comparing the value of statistics to a chi-square distribution to determine the significance level of the rule. If the p-value is higher than 0,05 (when the Chi squared value is less than 3,84), it can be considered that A and B are significantly indipendent, and hence the rule A => B, can be pruned form the results (Jaiswal & Agarwal, 2012)

3. Max confidence.

Is the measure that expresses the maximum confidence of the two association rules “A =>B” and “B=> A”.

$$max\_conf(A, B) = max\{P(A \mid B), P(B \mid A)\}.$$

4. Kulczynski (Kulc).

Is a measure of interestingness of an association rule that can be viewed as an average of two confidence measures. It is calculated as the average of two conditional probabilities: the probability of itemset *B* given itemset *A,* and the probability of itemset *A* given itemset *B.*

$$Kulc(A, B) = \frac{1}{2}(P(A|B) + P(B|A)).$$

5. Cosine measure.

The *cosine* measure can be viewed as a *harmonized lift.* For the cosine measure, the square root is taken on the product of the probabilites of A and B. The particularity and importance of this measure is given by the fact that the square root makes so that the cosine measure is influenced only by the support of A, B and A U B, and not by the total number of transactions. Closer to 1 the more transactions containing X also contain Y. Closer to 0, vice versa (Manimaran & Velmurugan, 2015)

$$cosine(A, B) = \frac{P(A \cup B)}{\sqrt{P(A) \times P(B)}} = \frac{sup(A \cup B)}{\sqrt{sup(A) \times sup(B)}} = \sqrt{P(A|B) \times P(B|A)}.$$

When considering the evaluation measueres of the rules, one should pay specific attention to the *null-invariance* properity of the measure. A measure is *null-invariant* if its value is free from the influence of null-transactions. This is specially important in large databases where many transactions do not contain the specified itemsets, therefore null-transaction.

A good interestingness measure should not be affected by transactions that do not contain the itemsets of interest, otherwise it would generate unstable results – for different ammount of data, the measures can generate completely different results.

Han, Kamber, & Pei (2012) suggest to use the Kulc measure in conjunction with the imbalance ratio measure:

$$IR(A, B) = \frac{|sup(A) - sup(B)|}{sup(A) + sup(B) - sup(A \cup B)},$$

It will give an idea of the imbalance of the two itemsets and about whether the directional implications between A and B are the same.

The literature presents many other interestingness measures like *All_confidence* and *Bond* (*Omiecinski, 2003*), *Hyper-lift, Hyper-confidence, Conviction, Coverage, Leverage* (Maninmaran & Velmurugan, 2015), but for the sake of simplicity this paper will focus only on the above described ones.

2.6.2. Association rule algorithms

The concept of Association Rules (AR) was initially popularised because of Agrawal, Imielinski and Swami's paper where they proposed a unifying framework for problems involving classification, association, and sequences, arguing that these problems could be seen under the process of rule discovery. They proposed a simple rule mining model, that could be applied for the discovery of one item consequent asociation rules. Indeed the first algorithm is called based on its authors, the AIS algorithm (Kotsiantis & Kanellopoulos, 2006). Since then many other have contributed to developing and proposing new models that could improve the association rule mining (Agrawal, Imielinski, & Swami, 1993) (Kumbhare & Chobe, 2014). Most of the efforts went in two main directions: improving the algorithm performance and reducing the output set, by allowing the possibility to express constraints on the desired result (Garcia, Romero, Ventura, & Calders, 2008).

The most popular algorithms for association rule mining mentioned in the literature are listed in the table below. The following table (table 1) will give an overview of the background research made on association rules, together with the characteristics and drawbacks of each algorithm and the respective paper where the algorithm is cited.

Algorithm	Papers and authors citing the algorithm	Description	Drawbacks
Apriori	Association Rule Mining: A review - Sethi & Mahajan, 2012 A review on Association Rule Mining Algorithms - Arora, Bhalla & Rao 2013 "An Overview of Association Rule Mining Algorithms" - Kumbhare & Chobe, 2014 "Mining of Association Rule: A review paper" - Harne & Deshpande, 2015 "Mining of Association Rule: A review paper"- Harne & Deshpande, 2015	Apriori was created to overcome the limitations of AIS and SETM. Apriori is also called "level-wise" algorithm. Developed for Boolean association rules, it uses prior knowledge of frequent itemset properties. The algorithm uses two functions at every iteration: candidate generation and pruning. The known k-item-sets are used to explore the (k+1)-item-sets. It generates the candidate item-sets by joining the large item-sets of the previous pass and deleting those subsets which are small in the previous pass without considering the transaction in the database. The frequent individual items are extended as long as those items appear sufficiently often in the database.	The algorithm tends to generate a large number of associations. Thus, when having large number of frequent patterns may suffer from: high costs in handling big number of candidate sets; tedious to repeatedly scan the database and check a large set of candidates by pattern matching. This imply a very high cost in terms of time, space and memory. Additionally, a computational problem emerges when having to consider the size of the frequent item-sets. For example, if k is the number of items in the itemset, and k=20, around one million subsets need to be considered. For finding long frequent item-sets, Apriori has to be abandoned.
Apriori TID	"Association Rule Mining: A review" - Sethi & Mahajan, 2012 "A review on Association Rule Mining Algorithms" - Arora, Bhalla & Rao, 2013 "Mining of Association Rule: A review paper" - Harne & Deshpande, 2015	The database is not used at all for counting the support of candidate item-sets after the first pass. Inspired from Apriori, it uses the Apriori function to generate the candidate item-sets. The difference from Apriori is that another set C' is generated of which each member has the TID (Transaction ID) of each transaction and the large item-sets present in this transaction. This set is used to count the support of each candidate itemset	Its performance depends on the number of candidates sets to be generated. The performance of Apriori-TID decreases when there is large number of candidate item-sets and when they don't have a skewed distribution with a long tail.
Apriori Hybrid	"Association Rule Mining: A review - Sethi & Mahajan, 2012 "An Overview of Association Rule Mining Algorithms" - Kumbhare & Chobe, 2014 "A review on Association Rule Mining Algorithms" - Arora, Bhalla & Rao, 2014 "Mining of Association Rule: A review paper- Harne & Deshpande, 2015	Since Apriori-TID outperforms Apriori when the itemsets distribution has a long tail (high distribution at the early stage) and Apriori outperforms Apriori-TID for larger datasets, these two algorithms were merged, based on the concept that it is not necessary to use the same algorithm in all passes over data. Apriori Hybrid uses Apriori in the initial passes and switches to Apriori-TID when it is expected that the candidate itemset at the end of the pass will fit in memory.	Because of switching from Apriori-TID to Apriori, it requires an estimation of the candidate itemset at the end of each pass. It requires a high computational cost in in determining the transaction point from Apriori-TID to Apriori.

Algorithm	Papers and authors citing the algorithm	Description	Drawbacks
FP-Growth	"Association Rule Mining Algorithms and Genetic Algorithms: A comparative study" - Ghosh & Sarkar,2012 "Association Rules Mining: A Recent Overview - Kotsiantis & Kanellopoulos, 2006 "An Overview of Association Rule Mining Algorithms" - Kumbhare & Chobe, 2014 "Mining of Association Rule: A review paper- Harne & Deshpande, 2015	Overcomes the bottlenecks of the Apriori algorithm by compressing the large dataset, by adopting a pattern-fragment growth method to avoid the costly generation of large number of candidates sets and by using a portioning-based method to decompose the mining task, also known as divide and conquer strategy. The algorithm requires two scans: in the first it creates a frequency list of the items while during the second it performs the mining of the FP-tree. Later on, it performs the FP-tree recursively. In the end it generates frequent patterns form the FP-tree.	Data becomes very inter dependent which makes it more difficult than Apriori based methods to be parallelizable on a shared memory. It is difficult to use the algorithm in an interactive mining process, for instance changing the support threshold may require the repetition of the whole mining process. Additionally, the FP-growth is not suitable for incremental mining. When new data is added to the database, it may as well require a repetition of the whole mining process.
AIS	"Database Mining: A performance Perspective" - Agrawal, Imielinski & Swami, 1993 "Association Rules Mining: A Recent Overview - Kotsiantis & Kanellopoulos, 2006 "An Overview of Association Rule Mining Algorithms" - Kumbhare & Chobe, 2014 Mining of Association Rule: A review paper- Harne & Deshpande, 2015	Considered to be the first algorithm proposed for mining association rules. It generates only one-item association rules. As a consequence, the consequent of the rule has only one item. The frequent item sets are generated by scanning the database several times. At each iteration, after checking the candidate item-sets against the threshold, the (k+1) item-sets are generated.	Too many frequent candidate item-sets that turn out to be short are generated. It requires considerable computational power to generate longer candidate item-sets.
SETM	"An Overview of Association Rule Mining Algorithms" - Kumbhare & Chobe, 2014 "Mining of Association Rule: A review paper" - Harne & Deshpande, 2015	The new candidate item-sets are generated as in AIS, but it makes use of the transaction identifier TID of the generating transaction, which is saved with the candidate itemset in a sequential structure. The algorithm separates the candidate generation process from frequency counting. At the end of the pass, the support count of candidate item-sets is determined by aggregating the sequential structure	Too many frequent candidate item-sets that turn out to be short are generated. It requires considerable computational power to generate longer candidate item-sets. Furthermore, for each candidate itemset, there are as many entries as its support.
ECLAT	"Performance Comparison of Apriori, ECLAT and FP-Growth Algorithm for Association Rule Learning" (Gayathri, 2017) "ECLAT Algorithm for Frequent Itemset Generation" (Kaur & Grag, 2014)	Finds elements from bottom like depth first search. Compared to other algorithms, it requires to scan the dataset only once reducing the computational effort needed and increasing the time efficiency (having even better performance than FP-Growth).	The algorithm has to be used on vertical datasets. Because of the way in which the algorithm runs, it does not support other interestingness measures such as confidence

Table 1. A comparison of Frequent Pattern Mining Algorithms.

3. TOOLS AND TECHNOLOGY

The following chapter will give an overview of the tools that were used throughout the projects during the internship experience in the Market Intelligence Team.

3.1. POWERBI

Microsoft Power BI (Business Intelligence), is a BI tool and analytics platform developed by Microsoft in 2013, and first released to the general public in 2015. It is a cloud-based data analysis and reporting tool. It consists of applications and services designed to provide coherent visual, and interactive insights into data, being a suite of business analytics tools and services that work together to access data sources, shape, analyze, visualize data, and share insights (Powell, 2017). The components of Power BI which make it a competitive tool against other BI tools are: (1) Power Query (Self-service ETL, Excel add-in); (2) Power Pivot (in-memory data modeling component); (3) Power View (interactive visualization tool providing a drag-and-drop interface); (4) Power Map (three dimensional data visualization tool, Excel add-in); (5) Power Q&A (recognizes the words a user write and help it to find the answer); (6) Power BI Desktop (a canvas which allows to build drag-and-drop, single unified and interactive visualizations) (Gowthami & Pavan Kumar, 2017). In order to better manipulate and make a better use of data, Power BI allows the use of DAX (Data Analysis Expression), which is a collection of functions, operators, and constants that can be used in a formula to calculate and return one or more values. DAX is a functional language, containing the full executed code inside the function. Aditionally, Power BI online version – Power BI Service – a SaaS option of Power BI runing in the Azure cloud allows users to share datasets, reports and dashboard by publishing them online. The great advantage of PowerBI desktop is the possibility to analyze data from different sources such as different file formats, databases, azure infrastructure, online services and other; facilitating this way the integration of a universe of sources into one tool.

On the other hand, it gets to be a limitation when one wants to export the insights into a format which is not a PowerBI format. One of the limitations of Power BI is its data storage capacity. Regarding the workspace, with a Pro licence each workspace is limited up to 10 GB of data storage. Regarding the datasets imported into Power BI, the tools is limited to 1 GB of data storage (the space downgrades to 250 MB when a user chooses to keep the Excel experience). Other limitations are given by the fact that Microsoft does not give customers flexibility to choose a cloud infrastructure as a service (IaaS) offering, instead running only in Azure.

Based on Power BI possibilities, its accessibility, flexibility, cost for the Pro version, upgrading pace and community, ease of use, this self-service intelligence tool was chosen to be used in several projects for ETL purposes, analysis, data exploration and reporting. When comparing with other tools, the 2019 report of Gartner places Microsoft in a leading position in its Magic Quadrant for Analytics and Business Intelligence Platforms (Gartner, 2019) (figure 7).

Figure 8. The "Magic Quadrant for Analytics and Business Intelligence Platforms". Source: Gartner, 2019.

For the purpose of this internship paper, PowerBI was used in order to develop a reporting tool that would have allowed to explore and visualize data extracted from an ongoing monthly questionnaire. By updating the dataset, the report would have been automatically updated in this way sharing valuable and updated insight every month.

3.2. IBM SPSS STATISTICS

IBM SPSS Statistics (Statistical Package for the Social Sciences), is a statistical interactive software designed to solve business and research problems by means of ad hoc analysis, statistical testing and predictive analytics. Some examples of the statistics included in the base software are (*figure 8*):

- Descriptive statistics: Cross tabulation, Frequencies, P-P plots and Q-Q plots
- Bivariate statistics: means, t-test, ANOVA, Correlation, Nonparametric tests, etc.
- Prediction for numerical outcomes: Linear Regression, Curve estimation, Partial Lest Squares, etc.

- ❖ Prediction for identifying groups: K-nearest neighbor, Hierarchical clustering, K-means.
- ❖ Forecasting: Autocorrelations, Cross-correlation, Sequence charts.

Besides the statistics, the software offers options as well for data visualization, using its “Graph” tab.

The many options are available via pull-down menus or can be programmed with a proprietary command syntax language.

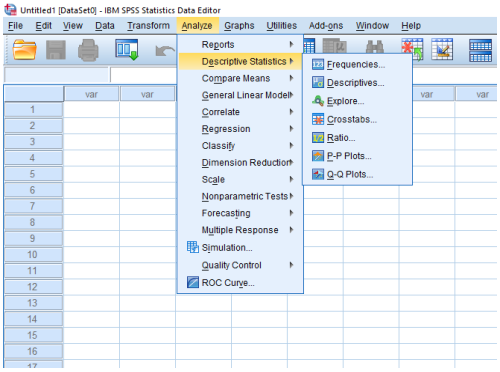

Figure 9. IBM SPSS Statistics - “Analyse” tab.

The SPSS datasets have a two-dimensional table structure, where columns represent measurement variables and rows represent cases. The graphical user interface is presented by two windows, data view and variable view. The first one contains the data table while in the second, the researcher can create a metadata dictionary, indicating the name of the variables, the type of variables (numeric, date, string, etc.), the type of the measures (ordinal or nominal), the role, the width, the type of missing values, and indicate a label.

The software supports a variety of source formats, which allows for flexibility when it comes to read data from different file formats, while its output has a proprietary file format (*.spv). Despite this, one can export the data to other file formats like *.csv, *.xlsx, *.sas7bdat, *.dta, suitable for MS Excel, SAS, Stata and other software. Additionally, when working with the software there are more windows in which the researcher can operate, for instance a script window and an output window. Using the first window, the researcher can run analysis by using strings of code while the output window will serve to visualize the output of the analysis and commands done in SPSS.

Internally the team works with IBM SPSS version 21. The software is used primarily to analyze datasets containing data from online questionnaires through ad hoc statistical analysis.

3.3. R

R is a programming language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and it is considered to be highly

extensible. R is available as a Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS. Furthermore, R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes:

- An effective data handling and storage facility,
- A suite of operators for calculations on arrays, in particular matrices,
- A large, coherent, integrated collection of intermediate tools for data analysis,
- Graphical facilities for data analysis and display either on-screen or on hardcopy, and
- A well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software. R can be extended (easily) via packages. There are about eight packages supplied with the R distribution and many more are available through the CRAN family of Internet sites covering a very wide range of modern statistics (The R Foundation, 2019). With more than 15.000 additional packages available at the Comprehensive R Archive Network (CRAN), R comes to be a language suitable for an immense variety of statistical and data mining tasks.

Despite of other advanced statistical data analysis and data mining tools, R was chosen to be used because of its flexibility, support community, ease of use and integration with other programming languages and software. Furthermore, the decision was done having in mind the scope of using it:

- Statistical analysis,
- Data visualization and exploration,
- Frequent pattern mining algorithm implementation and hence using available packages in R such as *arules and arulesViz.*

If there are several competitors on the market allowing to implement with ease machine learning algorithms, R seems to be unbeatable with regard to the first two points.

In conclusion, an extra factor that contributed to the choice of R as a tool was the already existing team's familiarity with the tool.

Within this internship, the programming language was mainly used to run the data mining project with the implementation of the frequent pattern mining algorithms and the identification of meaningful rules in the datasets.

3.4. QUALTRICS

Qualtrics is an online based survey tool, used to conduct survey research, evaluation and other data collection activities. It provides the researcher with a large array of question types like standard questions (multiple choice, matrix, text entry, etc.), specialty questions (graphic slider, heat map, drill down, etc.) and more advanced option (figure 9). It allows highly customizable survey appearance (using brands, pictures, videos and integration of other media), advanced conditional logic tools allowing for complex experimental designs and user-tailored survey paths, built-in email distribution capabilities (facilitating reminders and follow-ups) and the ability to export data as an SPSS data file, coma-delimited file (*.csv), text file (*.txt), HTML or XML. These latter features and many more of the online tool, make it an advantageous choice for survey online programming and distribution.

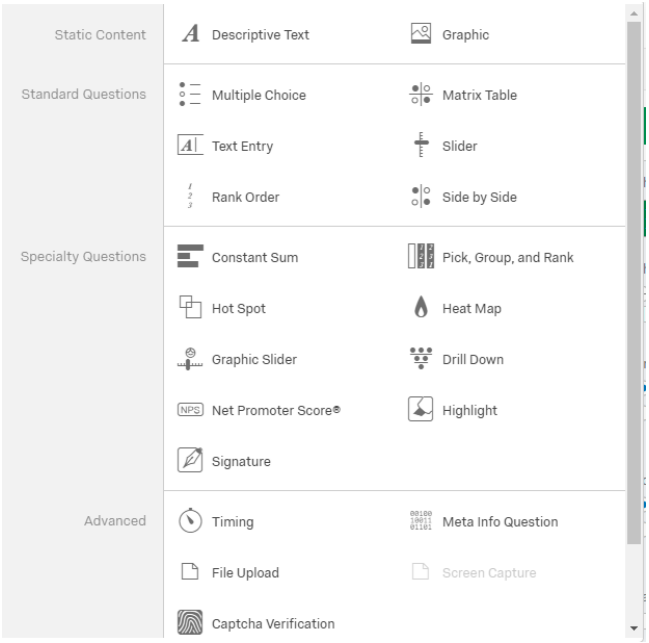

Figure 10. Qualtrics question type.

Qualtrics is being used by the Market Intelligence team to program highly personalized online questionnaires that are afterwards distributed to the chosen sample. The possibility to extract the answers in SPSS format, facilitates consequently the statistical data analysis performed in IBM SPSS.

The following matrix (table 2) will resume the use of the tools for each project. Intuitively, Qualtrics and IBM SPSS were used for market research projects. The former being used for online survey design, launch and data collection; while the latter for data analysis and visualization. MS Power Point was used to present the output of each project, nevertheless it was decided not to insert the tool in the matrix since its use was secondary to each project. Similarly, MS Excel was used as well in some of the projects as a support tool to store data and run quick analysis. During Project A, its spreadsheets were used to analyze and categorize the string variables, while during Project E and G, it was used as a way to store data and consecutively to use it as a source for the reports. Power BI and its Power Query feature was used respectively to build a reporting solution; while R was used to implement the frequent pattern mining algorithms.

Projects & Tools	Qualtrics	IBM SPSS	PowerBI	PowerQuery	R
Project A - OM	x	x			
Project B - SCE	x	x			
Project C - N2	x	x			
Project D - Adk	x	x			
Project E - DP			x	x	
Project G - Genie				x	x

Table 2. Matrix of tools and projects.

4. PROJECTS

The Market Intelligence Team's role is that of providing the company with customer and product centric insights based on data. Depending on the projects, the team is able to perform exploratory, descriptive or causal research, doing both quantitative and qualitative analysis. In order to get the necessary insights to explore and understand a specific business problem, the team may use different methods and sources of data, both primary and secondary. The team is involved in multi-disciplinary projects, demanded usually by internal stakeholders from different corporate areas, like marketing, R&D, software and digital solutions, etc.

Based on the goals of the projects, business problem and type of data required, the projects in which the student was involved during the internship can be grouped in three main areas:

- Market Research: Project OM, Project SCE, Project Adk and Project N2;
- BI reporting solutions: Project DSP;
- Data Mining: Project Genie.

This chapter aims therefore to provide a description of each group of projects and of each project individually. The chapter will be divided in 4 sections:

- The 1st section will put the projects into a time frame using a Gantt chart. It will be possible to see the duration of each project and understand where and how the projects overlapped. The Gantt chart will as well represent the key phases of each project in which the student has taken part in.
- The 2nd section will group the projects related to market research. It will try to give both a description of the overall methodology uniting these projects and a detailed picture of each market research project individually.
- The 3rd section will talk about the project involving the development of an internal reporting solution. It will start with describing the methodology used for this project, the project management approach and will conclude by presenting the final results.
- The 4th and last section will present a data mining related project and the journey throughout it, involving the choices that have been done, the adopted methodology, the encountered difficulties, solutions and results.

Each section will put the projects into a bigger picture, highlighting the reason for developing them, the vision and the benefits that were expected from each project.

4.1. INTERNSHIP TIMELINE

The following table (table 3 and 4) shows in detail each project with the main phases of the projects. The timeline gives an overview of the year, months and number of the week during which the project has been taking place.

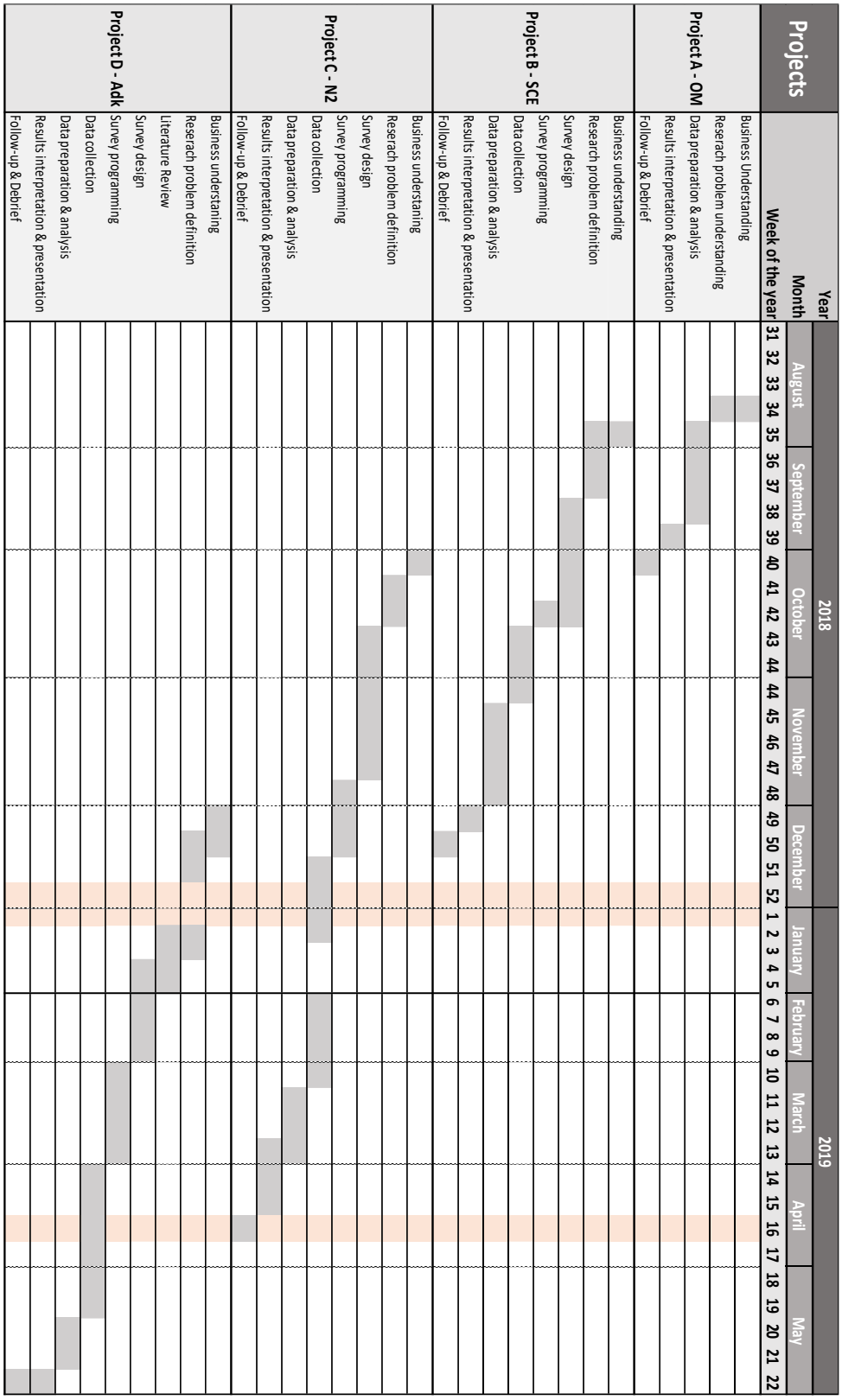

Table 3. Gantt chart diagram of the projects during the internship (part 1).

Projects	Year	2018																							2019																					
	Month	August					September				October					November					December				January					February				March				April				May				
	Week of the year	31	32	33	34	35	36	37	38	39	40	41	42	43	44	44	45	46	47	48	49	50	51	52	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
	Problem definition & Feasibility study																																													
	Requrirement specificaiton																																													
	System design																																													
Project E - DGS P	Implementation																																													
	Integration & Testing (Verification)																																													
	Deployment																																													
	Maintenance																																													
	Business & case understanding																																													
	Data retrieval																																													
	Data exploration & understanding																																													
Project G - Genie	Data preparation																																													
	Modelling																																													
	Evaluation																																													
	Deployment																																													

Table 4. Gantt chart diagram of the projects during the internship (part 2).

4.2. MARKET RESEARCH PROJECTS

Because the researchers in the team emphasize the research problem, exploring all available approaches (leading most of the time to adopt a mixed method research) in order to answer it, the market intelligence team's research philosophy could be classified in the pragmatic branch (Creswell, 2014). This type of philosophy allows the researchers to have the freedom of choice regarding the methods, techniques and procedures of research that best meet their needs and purposes. Typically, considering the research problem at hand presented to the team, the study design often contemplates the use of online questionnaires with both closed and open-ended questions, consequently generating both quantitative and qualitative data. (Creswell, 2014) (Kumar, 2011) (Mooi & Sarstedt, 2011).

Despite engaging in both qualitative and qualitative analyses, during the internship the intern analyst was mostly engaged in the latter type - having only occasionally the task to analyze qualitative data by performing categorization. The methodological approach applied to each developed market research project was not only inspired by the methodology commonly used by the team, but was as well deeply grounded in the market research literature (*see "Theoretical framework" chapter)*.

Inspired by a thorough review of the literature and considering the problem at hand, it was decided to make use of the market research steps suggested by authors such as Kothari, Kumar and Webb, and adapt them to the specific projects at hand. As a consequence of the adaptation, the market research projects described in this chapter have at their core the following steps:

1. Identification and formulation of the research problem;
2. Literature review;
3. Definition of the research design;
4. Definition of the sample;
5. Definition of the data collection method;
6. Collection and analysis of the data;
7. Interpretation and result presentation of the data;
8. Follow-up

It is worth mentioning the domain specific literature review was done when new knowledge was needed for a project covering an unknown to the team business area or research topic.

The above listed research steps applied to the business context, can be described as follows:

1. The business problem is given on request by one of company's departments/teams, which respectively become the stakeholder of the project. The stakeholder shares with share with the team the research idea or business need. After several iterations of meetings, the business problem and scope of the research project are defined. This allows to formulate the research problem.

2. In case the research area or problem results to be a new one, or a new research methodology might be more appropriate, the team will explore the existing academic literature to create a consistent knowledge frame. On the contrary, in case of the research area or problem resulting to be similar to one encountered in the past, the team will adapt one of the already existing and approve protocols.

3. The research design of the projects under the intern analyst during the traineeship, had its boundaries defined by the business environment and type of the projects. Consequently, the research design was limited by the only use of online questionnaires as a research tool for the research problem exploration and data collection. Nevertheless, the questionnaires varied among the projects by having different categories of questions, Likert scales, display logics, question flows, etc.

4. The "sample definition" step was as well confined by the business environment and the way the team was operating. Hence, for the projects described in this paper the sample was uniquely defined by the stakeholders.

5. Similarly to the previous point, this step was highly influenced by factors such as: business environment, stakeholder's needs, tools used for the project. The "data collection" method was in all the research project influenced by the requirements and needs of the stakeholder. Together with the sample definition, the stakeholder was in charge of distributing the online survey. Not being in control of these steps, the intern analyst was not able to influence the response rates and the significance of the collected data.

6. After the ending of the data collection timeframe, the data analysis could be initiated. The intern analyst was in charge of cleansing, structuring and analyzing the gathered data. For analysis purposes, the datasets were downloaded from Qualtrics in a SPSS file format to be analyzed in the IBM SPSS software. The data cleansing was done by eliminating the inconsistent, irrelevant, empty and duplicate data. Afterwards the data was explored and analyzed in order to describe and answer the initial research problem. The following chapters will cover in depth the data analysis methodology and implications.

7. The "data exploration" allowed to make sense of the answers collected from the sample, identifying key patterns and insights. After the conclusion of each research project the insights were presented to the stakeholders in order to allow them to act towards the business need and take data driven decisions.

8. After concluding each project, the researcher was available to provide angle of analysis not covered in the insights presentation as well as to provide additional ad hoc analysis. Additionally, each project was followed by internal team debriefs with the objective to share the knowledge acquired during each project, encountered difficulties and adopted solutions.

4.2.1. Project A – OM

Projects	Year	2018																																		2019																							
	Month	August					September					October					November					December					January					February				March					April					May											
	Week of the year	31	32	33	34	35	36	37	38	39	40	41	42	43	44	44	45	46	47	48	49	50	51	52	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22													
Project A - OM	Business Understanding																																																											
	Reserach problem understanding																																																											
	Data preparation & analysis																																																										
	Results interpretation & presentation																																																										
	Follow-up & Debrief																																																										

Table 5. Gantt chart of Project A.

Methodology

The Project A – OM, can be classified as a market research project, and it was the first project of the internship. As can be observed from table 5, the trainee has been involved in it for a total duration of 7 weeks, successfully finalizing it after this period. When the trainee entered Project A, it was already initiated and in its "data analysis" step. The trainee was assigned with the task to explore, analyze, generate and present insights from the already gathered data. To run the task and achieve the project's goals, it was essential to firstly understand the research and business problem, together with understanding the business activity of the stakeholder demanding the research. The project was made as a request of one of the companies belonging to Demant's corporate umbrella which is in charge of the production and distribution of a specific audiological product of the product portfolio. Because of the newness of the business area (in continuous growth) and the product itself, the research project had as its major goal that of understanding and creating a reality picture around the product in different markets, focusing on parameters such as: market competitors, health policies in different markets, retail and marketing strategies implemented by resellers, management of provided accessories and services.

The sample, already defined by the project stakeholder, was represented by professionals from different countries dealing with the product. The data collection was done through an online survey distributed via email; resulting in a response rate of 54% (from 52 contacts). Considering the different backgrounds of the respondents and the multitude of possible answers, it was important to design a survey that would have gathered as much as possible information, would have been as short and clear as possible to avoid the risk of drop out and would have been as tailored as possible to each respondent's experience to keep the engagement high. To achieve the latter goals, it was made use of the strengths of using an online questionnaire: using conditional question display flow, conditional personalization of the content of the questions.

Because of the "semi-exploratory" nature of the research, a mixed research method was used when designing the questionnaire; implying the use of both closed and open-ended questions – generating both quantitative and qualitative data. The close-ended questions were used for the questions where the range of answers was known and finite. Because of research problem's exploratory nature, not all the questions had predefined answers; without a concrete prior knowledge of certain topics it was fundamental to give the respondents the opportunity to express themselves, therefore open-ended questions were used. A mix of the two types of questions was used as well, for example inserting a question with pre-coded answers and leaving one option with the possibility to contain text (extending this way the range of answer options).

Data analysis

After understanding the research design and the initial project scope, the analyst was able to focus the analysis and search for insights that would have answered the research problem. To proceed with the data analysis step, the dataset was downloaded in an SPSS file format. The dataset was containing both variables representing the answers given to each question as well as automatically generated variables such as geographical coordinates, IP addresses, time spent on the questionnaire, etc. – variables which were deleted afterwards. Hence, each variable was representing a specific answer to a specific question (single choice and multiple-choice answers were stored as individual variables), while the rows were representing the record of a single respondent. To help the data interpretation, the "Variable View" window contained the metadata dictionary specifying each variable's name, label, type, width, measure, and role.

Before starting the data analysis process, the dataset was checked for data related problems, such as inconsistency, irrelevancy, missing or incomplete values and outliers in answers. This step was necessary to guarantee a high quality of the data and unbiased results; making in this way the insights more trustful and reliant for future managerial decisions. For this purpose, the records with incomplete or empty responses were deleted, maintaining only the respondents that have completed the questionnaire. The inconsistencies in the data were recoded manually, this was possible because of the relatively small sample size. An example of inconsistent data can be present from the variable "Country" (with the following recoding form): dk -> Denmark, denmark -> Denmark, Netherland/the Netherlands/Holland -> The Netherlands, etc. An implication of having respondents from different countries was their use of different currencies when indicating prices. It was necessary to transform the values into a single and interpretable currency, using the official currency rate of the national banks. A new variable listing the currency rate for each country was created and used for data transformation. The transformation generating new variables while keeping the old ones as well (excluded at this point from the analysis).

The dataset contained two different types of variables (the only two allowed by SPSS): ordinal (represented by numerical values) and nominal (represented either by the open-ended option of the answers or by the open-ended questions). The most common statistical measures such as mean, median, standard deviation were used to describe the ordinal data. Histograms and box plots were used to visualize the data distribution and identify the outliers in data. Because of the different nature of the nominal data, only descriptive statistics such as count, frequency tables, relative frequency and mode, were possible to calculate. The Open-ended questions had to be analyzed separately; since they were containing long strings of text it was inappropriate to run the usual statistics. For this purpose, categorization was done. Categorization is a technique used for text analytics, the researcher identifies each string main themes and the creates a frequency of the themes in order to visualize the insights. This allowed to quantify the qualitative information and facilitated its interpretation. Because not requested by the stakeholder, no hypothesis was formulated and therefore no model explaining the reality was developed.

Crosstabs were used to dice the data and analyze it from different dimensions, allowing for an in-depth exploration of hidden patterns in data. Analysis based on groups like "Country" or "Work were performed to explore whether there were any significand differences among groups of respondents. So far, no different patterns were identified.

Results

This exploratory market research allowed to answer the initially identified business needs, providing an in-depth picture of the business reality in different markets, comparing among countries, products and competitors. This allowed the stakeholder to gain a general picture and plan actions or identify specific point of interests based on data generated insights.

To facilitate the presentation of insights, the results were grouped in topics such as finance, market policies, accessories, service related, etc., allowing to present the insights into a coherent flow and give a holistic picture of the reality. The categorization of the string variables was integrated with other statistics, amplifying the horizon of knowledge on the single areas of analysis. The analysis was presented using MS PowerPoint and consequently shared with the stakeholder.

4.2.2. Project B – SCE

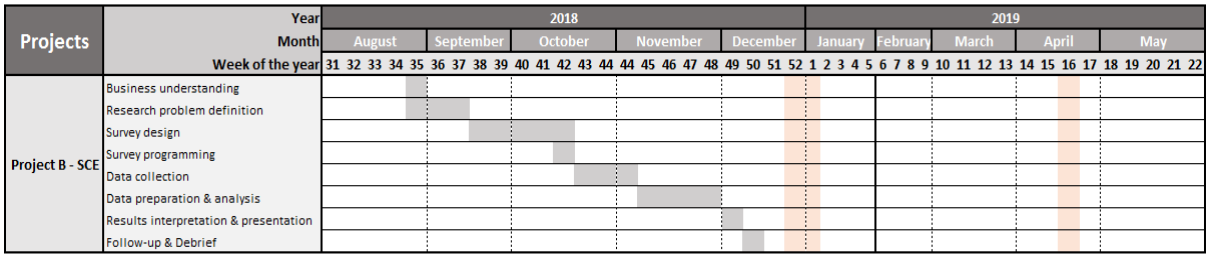

Table 6. Gantt chart of Project B - SCE.

The Project B - SCE, is the second market research project carried on by the analyst intern in the team. As shown in table 6, it had a total duration of 17 weeks. The main stakeholders of the Project SCE were the Market Intelligence team and the Marketing department. Contrary to the previous project, in this one the trainee was involved from the first phases of the project till the project's closing. The analyst intern was in charge of all the phases listed in table 6, receiving guidance and feedback from management. The project has been developed in accordance to the main guidelines described in the theoretical framework related to market research.

Identification of the business and research problem

One of the core parts of the project was the scoping of the research project in order to identify what was the business need and how to formulate the research problem. The scoping phase allowed to better understand the intent of the research that would have allowed to define the research design. From a business point of view, the research was aiming to understand how well were perceived and how helpful were the marketing materials released for the markets for a specific marketing campaign. The gained knowledge would have helped to align better the future releases of materials to the expectations of the end users.

The goal of the research was that of describing the reality of the facts, having at hand an *ex post facto research,* also known as descriptive research. It is worth mentioning that the main practical limitations were given by the time and resources allocated to the research. Consequently, it was prioritized to have a descriptive AS IS picture of the reality and postpone further in time a causal research. In the end, the scope of the research was that of measuring and describing the satisfaction of the end users of the released marketing materials, as well as to explore the reasons behind certain satisfaction levels.

Definition of the research design

The research design of the market research project was done after consulting the academic literature and taking into consideration factors such as team's expertise, the research problem itself and the resources allocated to the project. Based on the latter considerations it was decided to use a questionnaire as a data collection method to answer the research problem.

To measure and explore the satisfaction levels, both closed-ended and open-ended questions were used. In order to create the structure of the online survey the stakeholder provided the analyst with

appropriate terminology and content. This created a constraint in terms of length and complexity of the survey. Since not being in control of the latter factors, the analyst was aware that a bigger length and higher complexity of the survey could have cause a higher dropout rate. Display logic was used to personalize the experience of each respondent and to engage them. Using the available team's tool, the online questionnaire was programmed in Qualtrics and distributed to the selected sample via mail by the project stakeholder. Therefore, similarly to the content definition, the sample definition was not under the responsibility of the analyst or any other team member. By not participating in the sample and distribution method definition, the analyst was aware that this could lead to low response rates and thus not representative answers. The total data collection period was limited to three and a half weeks (17 working days). After closing the survey, the data was cleansed, explored and analyzed using SPSS, and visualized using PowerPoint.
Because of the confidentiality of the survey content and results, the questions will not be listed completely. Nevertheless, an overview of the type of questions, scales and structure of the survey will be given.

Structure of the data collection method

Aware of the limitation of an online questionnaire, the focus was set on the advantages of using such a tool – resulting to be the most optimal tool to achieve the goal of the research project. The online questionnaire was programmed using Qualtrics. Based on the business requirements, the questions were targeted to understand the satisfaction with specific marketing materials and specific categories of materials used by the respondents during the marketing campaigns.
To gather the data needed to answer the research problem, three types of questions were used:

1. Background open-ended questions, allowing to understand the name, company role and country of the respondent. This would have allowed to follow-up at a later point in time, and additionally to identify specific patterns in the data, if any.
2. Closed-ended questions related to the specific marketing materials and categories. The closed-ended questions were used to assess the satisfaction and ease of use or implementation of the different content under evaluation. For this purpose, 4-point Likert scale was used[1].
3. Open-ended questions were used to explore the reasons of a certain evaluation. The question invited the respondents to comment on their rating. Despite many variants found in the literature, following one was used for the open-ended questions:

   "Do you have any comments in relation to the rating above?".

   This formulation was preferred over the others because of its simplicity and low bias risk.

Before being launched online, the survey went through multiple validation processes in order to ensure as much as possible the maximum clarity, comprehensiveness and ease of interpretability of the questions, with respect to the main pillars of the survey design.

[1] The 4-point satisfaction Likert scale was inspired from the commonly known 5-point Likert scale. Based on previous experience with similar market research projects, the "Neutral" point of the scale was excluded to allow data generalization and insight creation by forcing the respondents towards a positive or negative evaluation. The 5-point Likert scale was preferred to the 7-point Likert scale because of the small sample size, in order to avoid results that were too spread for pattern identification.

Collection and analysis of the data.

Once the collection period ended, the data was collected in an SPSS file, to be further on analyzed in the statistical software (figure 10).

The dataset contained a total of 64 variables and 21 input records. 17 of the variables were automatically generated by Qualtrics with information regarding the time and date a response has been recorded, IP addresses, the overall duration of the survey, etc. The other 47 variables represented the questions and the respective items that had to be evaluated within each question. For instance, if a question in the survey had three items to be evaluated this generated three new variables in the dataset, corresponding to the three evaluated items.

The variables representing to the survey options were defined either as a string variable (17 string variables) or as a numeric (30 numeric variables) variable (figure 11). The numeric variables were storing the Likert scale evaluation of each respondent. Despite being considered as a categorical variable in the literature, the numerated Likert are stored usually as integers (e.g. from 1 to 4, based on the length of the scale) which are labeled with the appropriate scale level (e.g. 1= Very dissatisfied, 2= Dissatisfied, 3 = Satisfied, 4= Very Satisfied) (figure 12). Defining the Likert scale as numeric gives the researcher the possibility to run descriptive statistics such as *mean, standard deviation* and *median* when necessary, which would not be possible with categorical data.

Before proceeding with the data analysis, the dataset was assessed for inconsistencies, duplicates, missing values and other data specific errors.

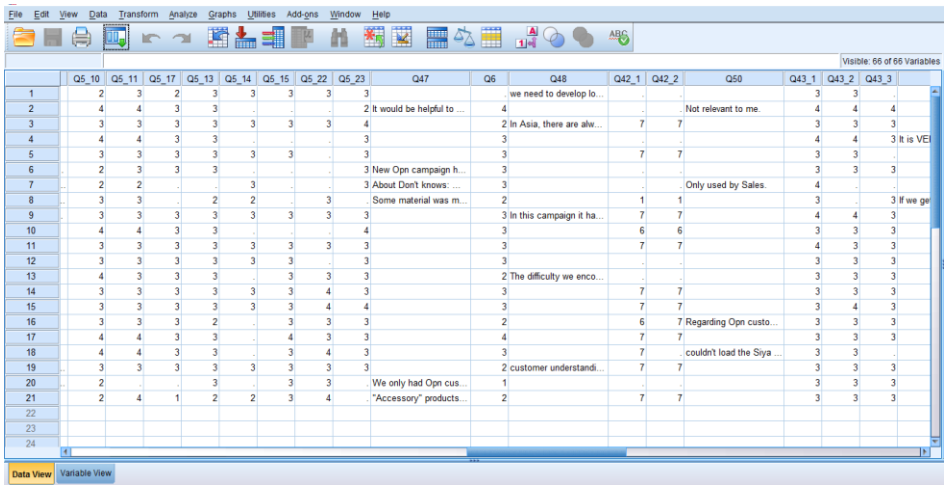

Figure 11. SCE project – Data view screen. Source: IBM SPSS, Project B.

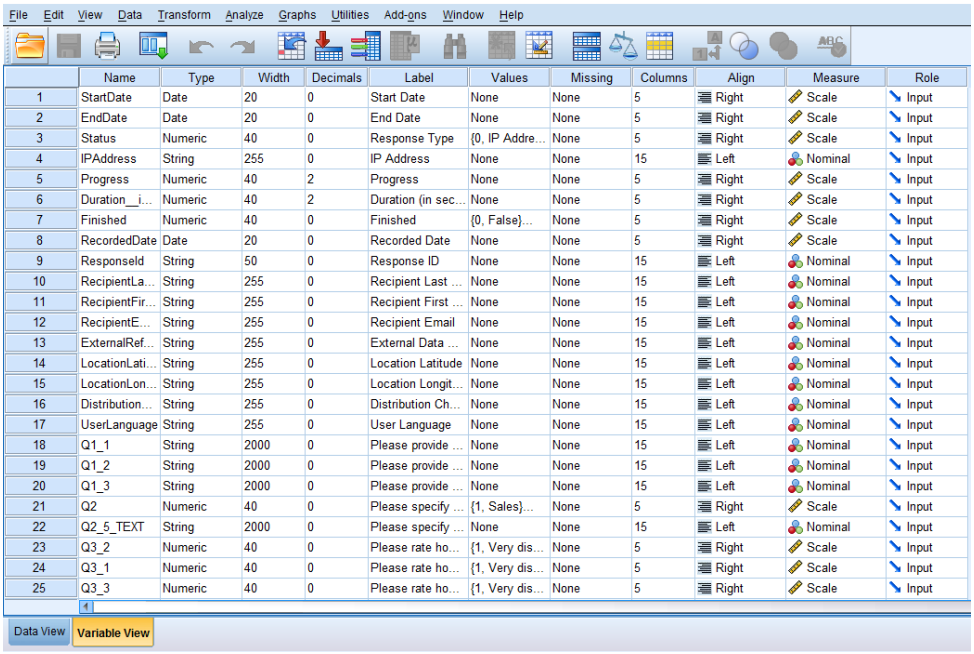

Figure 12. SCE project – SPSS variable View screen. Source: IBM SPSS, Project B.

Figure 13. Project SCE – Satisfaction Scale. Source: IBM SPSS, Project B.

So far the following data manipulations were done: (i) The country names were normalized by renaming the values corresponding to the same country but typed in different ways (e.g. Poland, PL and Poland -> Poland, etc.), (ii) incomplete survey answers were deleted from the dataset (iii) Analysis of the missing values: the missing values were identified, but no changes were done towards them since in this dataset missing values were carrying information themselves – showing for example that a specific questions was not representative for the respondent.

The overall descriptive analysis was done by generating descriptive statistics of the numeric variables using the Frequencies tab option in IBM SPSS (Figure 13). Additionally, crosstabs were used to search for pattern and differences among different groups of respondents.

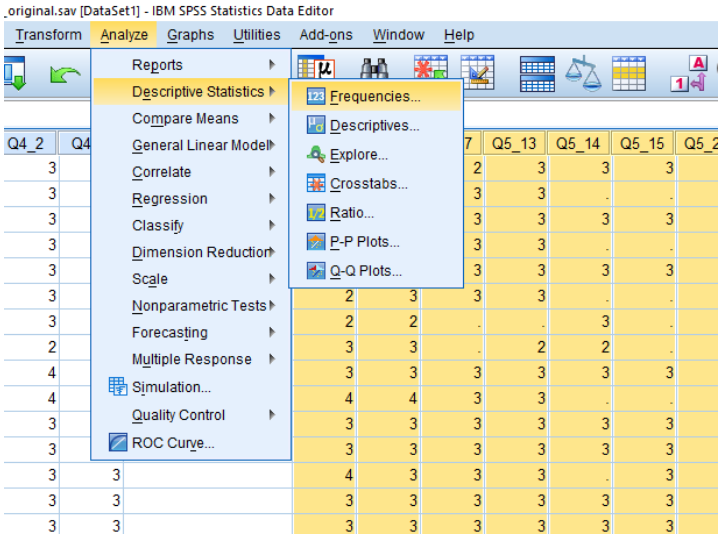

Figure 14. Project SCE – IBM SPSS "Frequencies" tab. Source: IBM SPSS, Project B.

The output of the Frequencies was generated in the IBM SPSS statistics viewer window, figure 14 shows in one example.

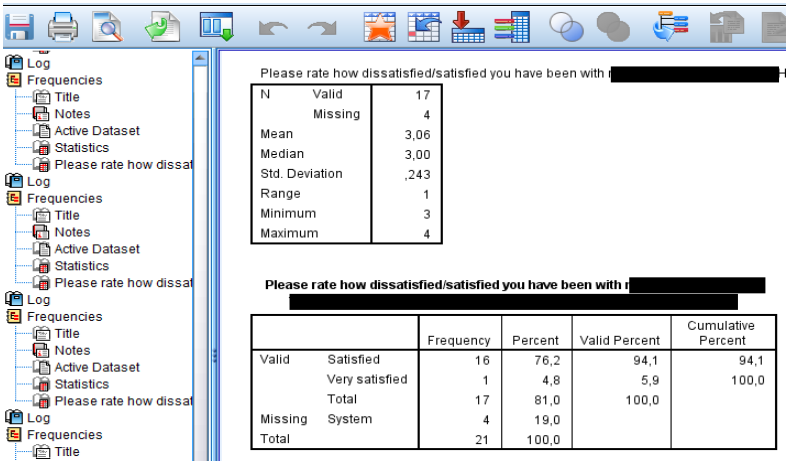

Figure 15. Project SCE – IBM SPSS Statistics Viewer. Source: IBM SPSS, Project B.

From the example above, it can be observed that the output consists both in descriptive statistics, (giving an overview of the std. deviation within the answer) and the frequency table of the Likert Scale. Since the sample size is only composed by 21 observation, it is common that few missing values represent a high percentage (it can be seen in the example that 4 missing values represent almost 20% of data).

The frequency tables were used secondarily for creating the data visualization in MS PowerPoint.

The string variables containing the comments of the respondents were analyzed and delivered separately to the stakeholders. The was no categorization done on the text because of the relatively small sample size, the nature of the comments (in depth comments on certain business issues) and the specificity of each observation (every comment was country specific and there was no possibility of generalization of the insights).

Interpretation & result presentation

To deliver key insights and make them actionable, it was essential to visualize the findings and knowledge in an easy interpretable way. For this purpose, stacked bar charts were used to visualize and compare the distribution of the satisfaction among the items relative to a specific question, while bar charts were used instead for the questions with only one item to be evaluated (figure 15).

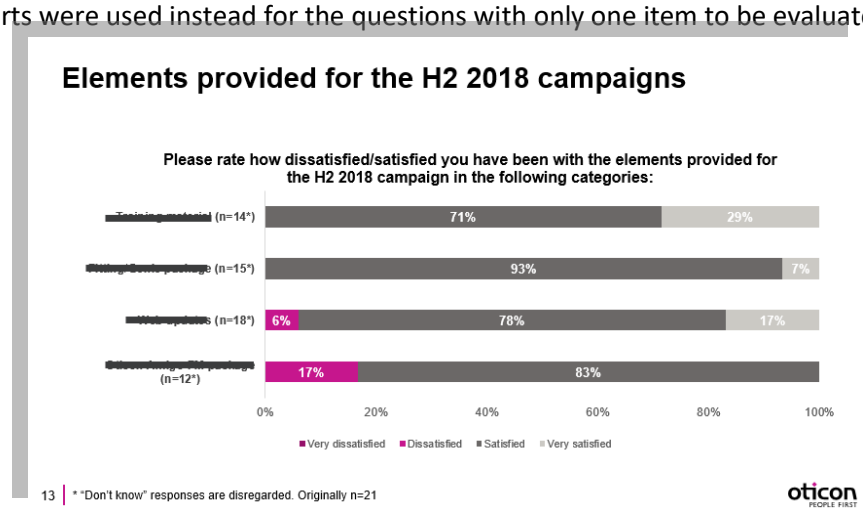

Figure 16. Project SCE – Data visualization. Source: MS Power Point, Project B.

The data analysis and visualization allowed the stakeholder to have a data driven overview of how the developed marketing materials were perceived on different markets. This helped to target better the future releases and marketing actions in order to improve customer satisfaction and develop more efficient processes.

4.2.3. Project C – N2

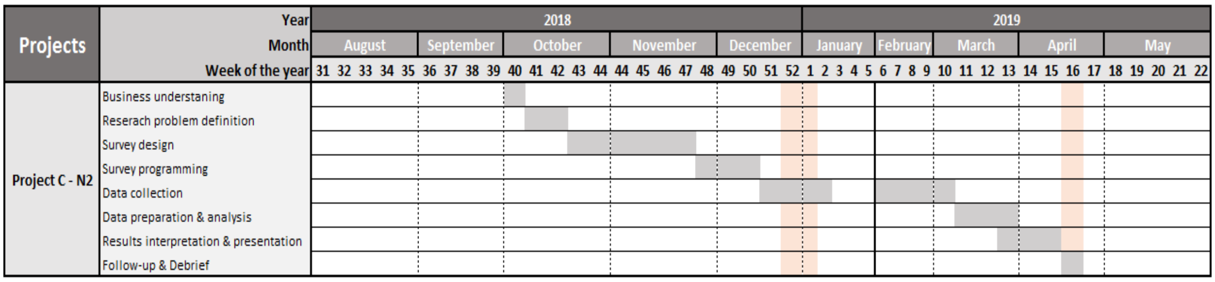

Table 7. Gantt chart of Project C - N2.

Identification and formulation of the research problem

Similarly to the other projects, Project C or N2, emerged from the request of an internal stakeholder. With the launch of a new product on the market, the marketing team of Oticon had to produce materials that could endorse the product and assist the retailers, at the same time it developed and implemented communication activities in order to ensure a successful launch and support for the involved stakeholders in the different markets. This project has been developed and research decisions have been taken in accordance to the main guidelines described in the theoretical framework related to market research. The main guidelines related to the research steps, data collection methods and scaling were considered together with the best practices of questionnaire design and development.

During this project the student was responsible to carry on all the phases delegated to the Marketing Intelligence Team, having to report to his management and get feedback. When starting the research project, the stakeholder engaged the team starting from a broad problem to research about, therefore it was fundamental to identify and specify the business and research problems in order to proceed. The research was needed in order to allow the department to have a data driven approach when defining future campaign developments, potentially reduce campaign costs and waste, while increasing the overall satisfaction with the materials and the marketing efforts when launching a new product globally. Additionally, the research would have allowed to better target the content distribution and marketing strategies among the markets.

Once the business needs were identified, it was possible to transform the business concerns into a research problem. After several iterations it was defined that the research had to be done to understand four key points: (i) How relevant were the implemented communication activities; (ii) How appropriate was the number of the materials developed and delivered to the different countries; (iii) Which materials were used by the different markets when creating their campaigns; (iv) Why the ones not being used were not used.

Definition of the research design

The case at hand was treated as a decision-oriented descriptive research type. Considering the available resources for the project, time frame and sample size, it was decided to adopt a quantitative approach to the problem. The short time frame for the project didn't allow to run focus groups or in-depth interviews and collect qualitative data. At the beginning, a close partnership with the stakeholder was necessary to understand better the elements that had to be put under research, such as: (i) Which where the commercial activities to be analyzed for their relevancy; (ii) Which materials were to be analyzed for their quantity; (iii) Which elements were to be analyzed whether they were used or not and why. The stakeholder had to provide the trainee with an overview of the business processes, so the trainee had a better chance to correctly formulate the questions so to extract relevant data for the research problem. The researcher had freedom to find the most appropriate way to obtain the data that would have answered the research problem, to implement the identified method, and analyze the data to mine and deliver insights.

The elements of the case such as the available resources for the project, the time frame, a sample of respondents distributed all over the world, and type of research problem, influenced the tools and methodology to be used. Consequently, it was decided to make use of an online questionnaire as the only tool satisfying the parameters of the project and being able to collect the necessary data to answer the research problem. Furthermore, the capabilities of the online tool, such as creating highly personalized experience during the survey, would have allowed to explore more in depth and in a more targeted way the above stated problems.

Once designed and programmed in Qualtrics, the online survey was launched in two phases: a soft launch, to test during a period of two weeks for bugs in the survey flow and identify possible problems with the survey experience. Once the soft launch phase was completed and the survey validated, it was distributed via email to the rest of the sample (for a total data collection period of five weeks). During the second phase, a series of follow-ups were done in order to increase the response rate. After ending the data collection period, the researcher was in charge of analyzing the data, extracting the insights and presenting them.

It must be mentioned that sample definition, sample size and the online survey distribution were under the responsibility of the stakeholder. This implies that the intern was only partially responsible for factors such as response rate, since the definition of the sample and survey distribution method highly influence the number of answers. A wrongly targeted sample, an inefficient way of communicating the survey, a too small sample size, could lead to a low response rate and irrelevant data. On the other hand, to increase the chances of a low drop out the questions were formulated in simple, unambiguous and unbiased way by using user friendly wording, evaluating only on a single parameter or element at a time and by avoiding words or element that could have allowed to influence the answers of the respondents.

Definition of the data collection method

Based on the initial scoping, the online survey had to cover three main areas of interest: (i) Relevance; (ii) Quantity; (iii) Usage of materials. The latter factor together with the content provided by the stakeholders made the respondents to provide information related to a total of 47 elements (9 elements related to the first area, 2 elements related to the second, 36 related to the third). This made the survey design a problem of finding the right balance between the length of the survey, the goal of obtaining high quality and relevant data, and the goal of obtaining a higher as possible response rate. As stated for the previous surveys, literature regarding the survey design indicates that the longer and the more complex are the questions in a survey, the higher the possibility of a question's misunderstanding and of the dropout rate. The risk of a high dropout rate becomes higher when there are no incentives to keep the respondents engaged in the survey process.

Consequently, considering the above stated risk the online survey was designed appropriately. To collect as much insights as possible, a mixed method was adopted for the survey – using both closed ended and open-ended questions. The closed-ended questions gave the possibility to present predefined answers (since the range of the answers was already known), thus facilitating the task of answering the questions and saving answering time. The closed-ended questions where used when formulating the following categories of questions: (i) Single choice predefined question used for Likert scale ratings (when rating the relevancy of certain materials) (figure 16 and 17);

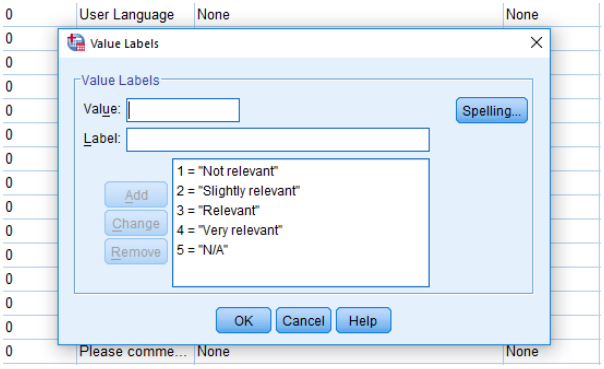

Figure 17. Project N2 – IBM SPSS visualization. Relevance Likert scale. Source: IBM SPSS, Project C.

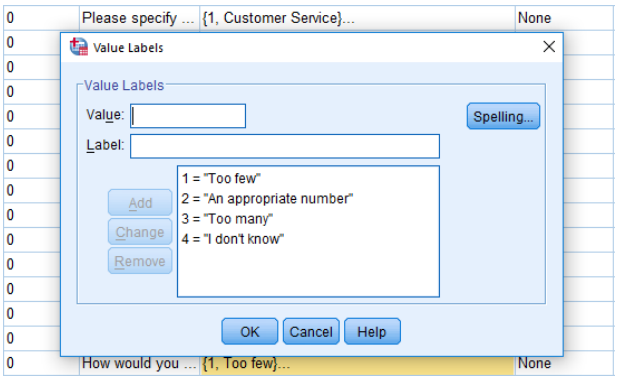

Figure 18. Project N2 – IBM SPSS visualization. Quantity scale. Source: IBM SPSS, Project C.

(ii) Multiple choice predefined questions for the selection of materials that have been used; (iii) Single choice pre-coded questions when selecting the reason why the materials were not used. For this option, the business expertise of the stakeholder helped to define a hypothesis about the range of possible reasons of not using the materials (figure 18).

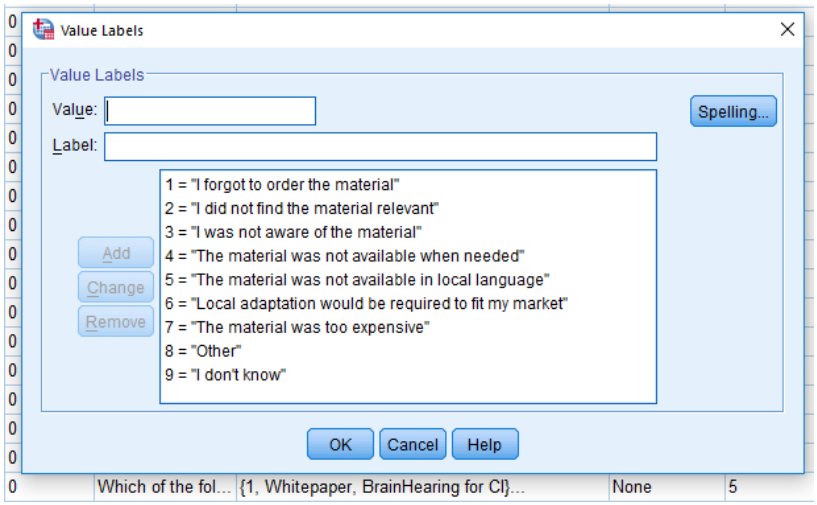

Figure 19. Project N2 - IBM SPSS visualization. Assumed reason for non-usage. Source: IBM SPSS, Project C.

This choice of predefining the reasons of not using the materials was done to be able to summarize the results and reduce the overall time needed to answer the survey. Because of a large number of items to be evaluated, it was expected that by asking the respondents to specify with an open-ended question the reason of not using a specific item for their marketing campaigns, would have led them to abandon the survey (due to a considerable increase in the time needed to complete it) - considering that the more items the respondents were not using, the more items they would have had to evaluate consequently. To get more in-depth insights and still leave the respondents with the possibility to give their input, open-ended questions were used after each category of materials. When related to ratings the following expression was used (inspired from the approaches cited in the literature and similar cases faced by the market intelligence team):

"Please comment on the ratings above"

While to invite the respondents to comment on the usage of materials, the below formulation was used:

"Please elaborate on the above selected materials:
(*i.e. Any of the XXXX Materials that did not live up to your expectations, market requirements or customer needs? Anything you would like us to do more of/less of?)"*

To reduce the overall time needed to answer the survey, make the answering experience as much as personalized and targeted as possible (thus increasing the chances of engaging the respondents), display logic has been used when answering on the usage of materials. For example, if a certain material is not select then the following question will display only the materials that have not been selected, so the respondents could give their input only on the items of interest. Additionally, background questions were introduced in order to track the work title and country of the

respondents. This information would have helped to identify possible patterns in data and check for significant differences in the answers between groups of respondents.

After having completed the online programming phase, the overall time expected for the completion of the survey was estimated to be between 15 and 20 minutes.

Collection and analysis of the data

After closing the data collection period, the data was stored in an SPSS file in order to be consequently analyzed in the statistical software. The survey generated a dataset with a total of 127 variables (17 of which generated automatically by the online survey tool – such as start and end date, time, IP address, completion, time required to complete the survey, etc.) and 43 inputs was created (corresponding to 43 responses).

The variable type was automatically defined by the IBM SPSS software considering the type of input. As a consequence, out of the 110 variables 96 were defined as numeric and 14 were defined as a string (containing the text with the comments of the respondents). Despite being automatically defined as numeric, the variables not representing a Likert scale could not be used for descriptive statistics calculating metrics such as mean, sum, standard deviation, etc. These variables had to be considered as categorial and consequently use the appropriate descriptive statistics, such as count and frequency tables. The only numerical variables on which descriptive statistics could have been run were the variables containing representing Likert scales.

Referring to the concept "garbage in, garbage out"[2], in order to obtain relevant and trustful insights, the dataset was cleansed as follows: (i) incomplete answers and test answer were eliminated because considered not relevant. Incomplete surveys[3] can create noise and their validity could be questioned, since the respondents have abandoned the survey; (ii) Duplicate records were eliminated, by keeping only the most recent record; (iii) Inconsistent values, like different typing of the same country were normalized applying a coherent naming.

Considering the data type, the analysis was done by generating frequency tables of the variables representing the pre-coded answers. In order to analyze the multiple response questions[4], variables had to be defined using the "Define Variable Sets" option in IBM SPSS. By using this option, it was possible to combine the information carried by different variables in one single analysis by creating frequency tables summarizing the answers of the respondents by the different answer options per single question.

[2] Garbage in, garbage out (GIGO), is a common computer science concept stating that regarding the accuracy of the method to analyze the data, poor inputs will always deliver poor outputs.

[3] Incomplete surveys referred to respondents that have not completed all the survey, going through all the questions.

[4] The multiple response questions are questions for which the respondent has the possibility to select multiple answers from a predefined range. When storing the choices, the tool creates a new variable for each option selected by the respondent, therefore it's important to combine the information carried by the multiple variables in one single analysis. IBM SPSS offers such a possibility with the "Define Variable Sets" tab.

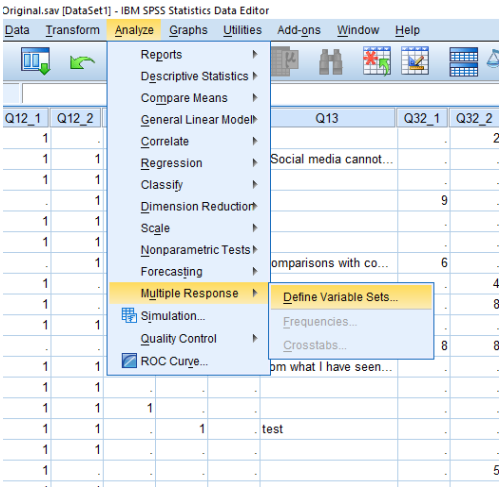

Figure 20. Project N2 – IBM SPSS. Define Variable Sets option. Source: IBM SPSS, Project C.

IBM SPSS Statistics Syntax Editor was used to write and store the script of the analysis (figure 20). The code was used to run the summary and descriptive statistics on the data. An additional importance of the script was given by the fact that it allowed to track the steps in the data cleansing and analysis as well as to identify and correct potential mistakes.

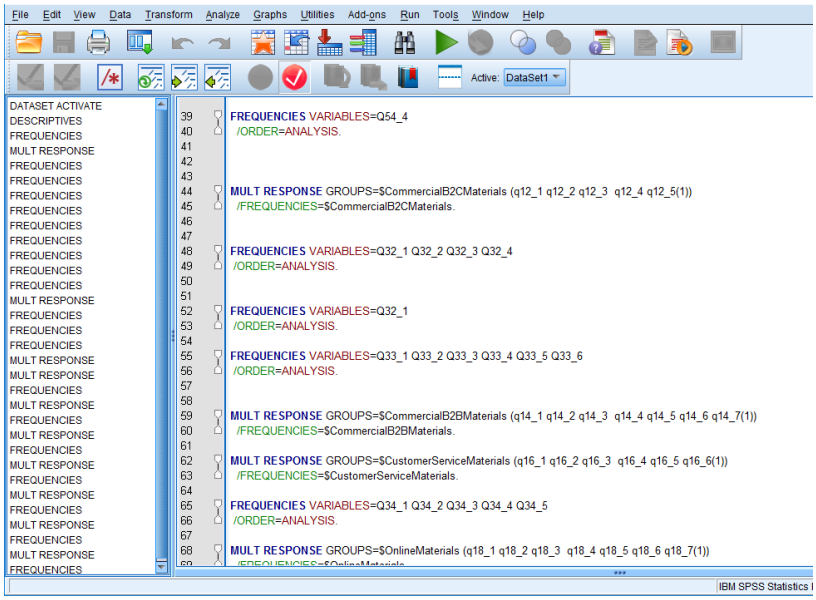

Figure 21. Project N2 – IBM SPSS Statistics Syntax Editor visualization. Source: IBM SPSS, Project C.

IBM SPSS Statistics Viewer was used to visualize the output of the analysis and identify easier possible patterns, anomalies in the data and generate insights from data.

By cross-tabulating the data with the country and work title dimension it was possible to check for patterns, understanding whether these two dimensions could influence the answers. So far, no patterns were identified since the answers were distributed homogenously across the countries and work titles. Additionally, it's worth mentioning that the single categories of materials were different from each other and hence not comparable.

The string variables containing text related to the comments were not categorized and hence, left to be analyzed by the stakeholder because of the insights being deeply business and market specific - having therefore an individual value. Additionally, it emerged that the assumption of the stakeholder on the reasons why the items were not used, was mistaken since the most frequent answer was "other".

Interpretation and result presentation

MS PowerPoint was used to visualize the data. The visualization was organized in order to answer the three initial business needs, creating three sections in which the issues were addressed.
Examples of data visualization can be seen in the figure 21.

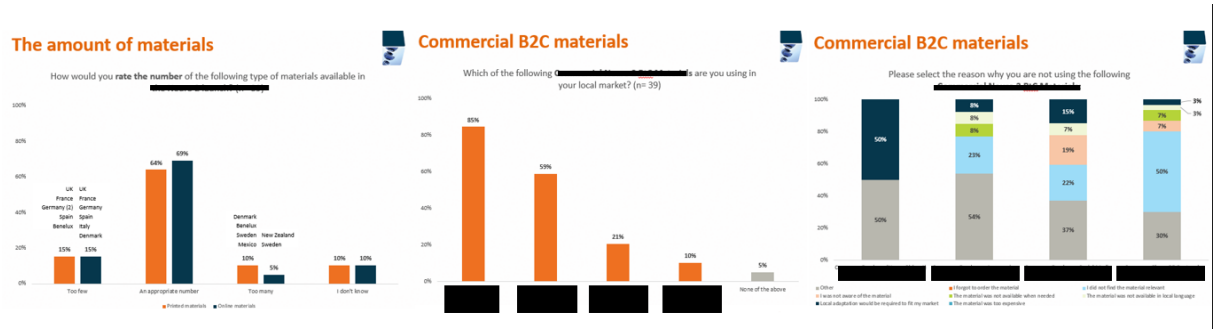

Figure 22. Project N2 – Data visualization examples. Source: MS Power Point, Project C.

Visualizing the data allowed the stakeholder to understand the reality present among its network. The survey allowed to quantify the campaign efforts and identify adjustments to be done for the future.

Limitations and lessons learned

One of the limitations of this research project is represented by the initially defined goals, analysis level requirements and allocated resources for the projects (especially in terms of time). Additional analyses could have been done by performing non-parametric tests between groups defined on the "country" or "work title" variable. Further data aggregation could have allowed to search for hidden patterns and insights.

Despite performing a descriptive quantitative research, the end user of the project was satisfied with the output since it helped to derive valuable knowledge and to assess the present situation on the markets. This project gave the student the opportunity to grasp the knowledge on how to manage the phases of a market research projects, starting from the scoping till the presentation of insights.

4.2.4. Project D – Adk

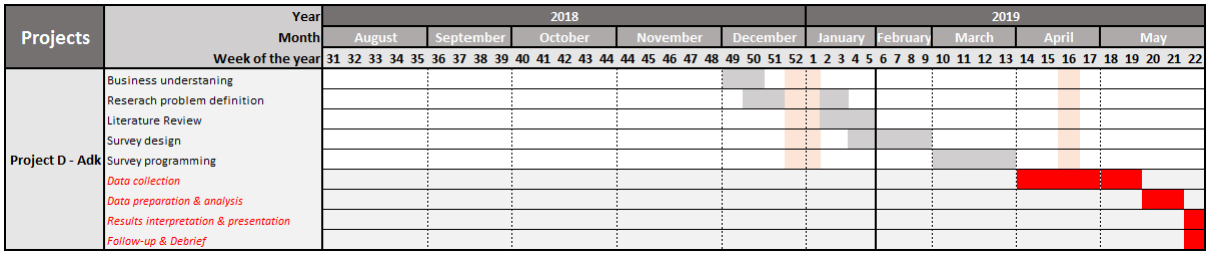

Table 8. Gantt chart of Project D -Adk.

Oticon which focused on developing highly technological health devices, has at its core constant technological innovation. The new generations of hearing aids are capable of connecting with the users' smartphone, bringing the user experience to a new level through the usage of *ad hoc* developed apps. A new app was developed for the market to help the hearing aid users to track their experience with the hearing aids and successively optimise their hearing aid fitting sessions.

Identification and formulation of the research problem

Project D (Adk) was initiated in December 2018 and was supposed to have a total duration of 26 weeks. The Marketing Intelligence team was asked to understand whether the app had the potential to become a success on the market (and be popular among HA users). It was important to understand how the app was perceived by the test users and whether there were any issues in using it. Aditionally, it was required to understand whether the app was meeting its predefined purpose and if it was bringing any additional value to the users.

The analyst intern had the responsibility to manage all the phases of the project and deliver the final insights after the conclusion of the market research project. He was partially assisted by the management during the scoping phase and throught the project by getting feedback on the project development.

Literature review

After defining the business problem, the literature was analyzed in order to identify similar cases and understand how to approach the research project. A review of the literature was essential in order to create the knowledge frame and identify possible parameters on which to evaluate the app and collect relevant data that could answer the research problem (Nikov, Vassilieva, Anguelova, Stoeva, & Tzvetanova, 2003) (Folmer & Bosch, 2004)(Skiba, 2017). From other similar research projects, relevant parameters for app evaluation were identified, such as: frequency of use, ease of use, technical issue identification, net promoter score evaluation. The literature review helped additionally to decide on the most appropriate method for data collection, this being an online questionnaire submitted to test users in which display logic, graphical elements and links to web pages could have been used. The method definition was influenced as well by the desire to make the

evaluation scalable, therefore run it potentially at a global level and gather quantified and comparable data.

Definition of the research design

Considering the literature examples, it was opted (similarly to the previously described projects) to use an online survey as a tool for data collection method. Therefore, the research could be classified as a decision-oriented descriptive research, done through a quantitative methodology. Descriptive, because there was no initial hypothesis prior to the research and the stakeholder wanted only to understand and grasp a picture of the reality. The online survey was planned to be programmed with Qualtrics, and to be distributed via mail to a sample of app users pre-defined by the stakeholder.

After gathering the business context and information related to the app it was possible to proceed with the survey protocol[5]. The business and technical overview of the app under evaluation was needed in order to target as much as possible the questions and understand what an appropriate formulation of the questions would be to avoid misinterpretation and bias (consequently generating poor quality results).

A series of screening, background and evaluation questions would have allowed to ensure both the quality of the sample and the quality of the gathered data. Consequently, after concluding the data collection period (initially planned for a duration of 6 weeks), the data had to be analyzed in IBM SPSS using the main descriptive statistics to find the main patterns and insights to be presented to the stakeholder.

Definition of the data collection method

As previously anticipated, a number of screening questions to ensure the quality of the sample, in other words, to ensure that only users that have used the app were participating in the survey. Consequently demographic and behavioral questions were used to collect data on how the app was used and identify at a later point in time whether there was any correlation between the behavioral data and attitudinal data[6] The evaluation questions aiming to collect data answering the research problem were formulated with the use of 5-point Likert scales measuring the ease of use[7], helpfulness[8] and satisfaction[9]. To analyze and measure the popularity and perceived value generated by the app, Net Promoter Score was used by asking them to evaluate the likelihood of recommending the app to a person in a similar situation.

[5] Survey protocol refers to the formulation of the questions to be used in the survey together with the order and logic of the questions to be displayed.

[6] Attitudinal data referred to data representing the attitude of the respondents toward the parameters under evaluation.

[7] Ease of use 5-point Likert scale: very easy, somewhat easy, neither easy nor difficult, somewhat difficult, very difficult.

[8] Helpfulness 5-point Likert scale: Not at all helpful, slightly helpful, moderately helpful, very helpful, extremely helpful.

[9] Satisfaction 5-point Likert scale: Very unsatisfied, somewhat unsatisfied, neither satisfied nor unsatisfied, somewhat satisfied, very satisfied.

The above described group of questions were programmed using a display logic. Therefore, during the screening question, an inappropriate answer would have ended the survey experience. With the same logic and with the aim to extract as much insights as possible, open-ended questions were displayed in case of a negative evaluation on the Likert scales. The open-ended questions were inviting the respondents to comment on their ratings.

Before launching the survey, a series of iterations were needed in order to align the content with the expectations of the stakeholder. Nevertheless, to test the quality of the survey it was planned to do first a soft launch of the survey and afterwards a full launch. In this way avoiding mistakes and inconsistencies in the survey structure.

Collection and analysis of the data

After the online survey approval, it was launched but the data collection phase had to be extended due to few responses that were not significantly representative.

Limitations and lessons learned

One of the limitation of the research project is the type of research being decision-oriented research, where the researcher doesn't have the possibility to define and test hypothesis or to develop a own research problem, since the research problem is stricly linked to the business problem. Fruthermore, in a research project the sample definition and distribution method have an impact on the response rate and data quality, this leaves the research with limited space in terms of delivering highly insightful information. In the future, datasets containing detailed background questions regarding the behaviour of the users could be used to test research hypothesis and develop more complex models

4.3. REPORTING SOLUTION. PROJECT E - DP

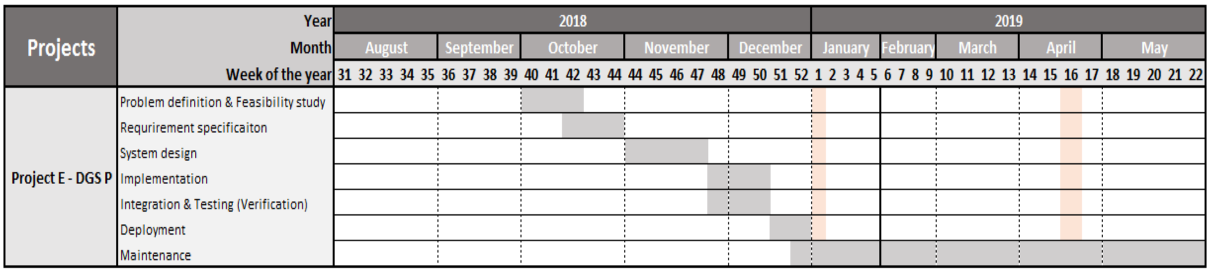

Table 9. Gantt chart of Project E - DP.

The Project DP was inspired by the willingness to optimize and automatize an internal process re-occurring monthly, in order to save team's working hours and create a more efficient tool of providing insights to the stakeholders. The original reporting process consisted in analysing data collected through a monthly survey evaluating internal stakeholder satisfaction on specific services and activities provided internally. The monthly analysis of the satisfactions was done in order to keep track of it over time and use the insights to provide better services. Hence, Project E was centred on developing a reporting tool that would have allowed to automatize the data analysis process and sharing of the insights of the end users. The project had a total duration of 13 weeks (table 9), including the national holidays during which project related activities were interrupted). The analyst intern was in charge of developing and managing all the phases of the project, including stakeholder management, feedback implementation and testing.

The framework of the original data analysis process can be described with the following five steps (figure 22):

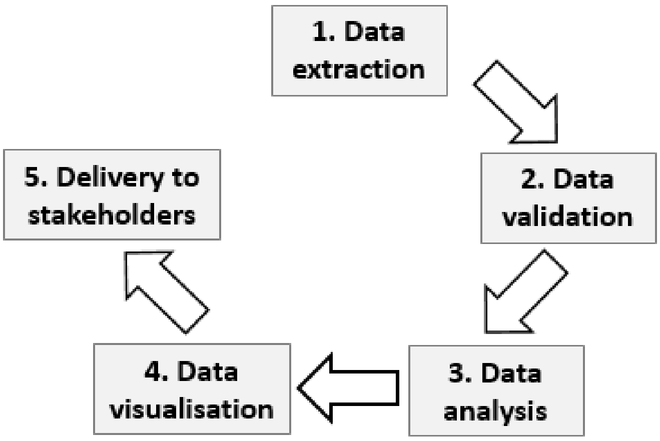

Figure 23. The original data analysis process setup.

In more detail, the steps can be described as follows:

1. Data was extracted from Qualtrics into an SPSS format.

2. The validation step consisted obtaining stakeholder's feedback on whether the respondents have answered the right group of questions[10]. After the validation, the dataset was checked for irregularities and anomalies in the answers. After the data cleansing, it was possible to proceed with the analysis.
3. The dataset was analysed using IBM SPSS through frequency tables generations and cross-tabulation to summarize the outputs for each variable. The outputs of the analysis were used to update the graphical visualization of the data using PowerPoint. Update, since the content and consequently the data analysis process were not changing over time.
4. After being updated monthly the PowerPoint report was delivered to the stakeholder.

Because of the process being handled manually, it was time consuming and inefficient with space for errors in the output (the higher risk of errors being during the update phase of the report). Additionally, considering the static delivery format, there was no space for interactivity with the data and it was limiting the generation of more insights. The inefficiency aroused when the stakeholder had to ask for more data analysis to understand specific trends. A side effect of this methodology was using employee working hours (up to one working day) to update the report instead of being involved in more challenging and engaging projects.

Having at hand a repetitive process, it was decided to automatize it. For this purpose, considering the available tools on the market (see chapter 3), it was decided to use PowerBI. The new reporting process would have had the flow as represented in figure 23.

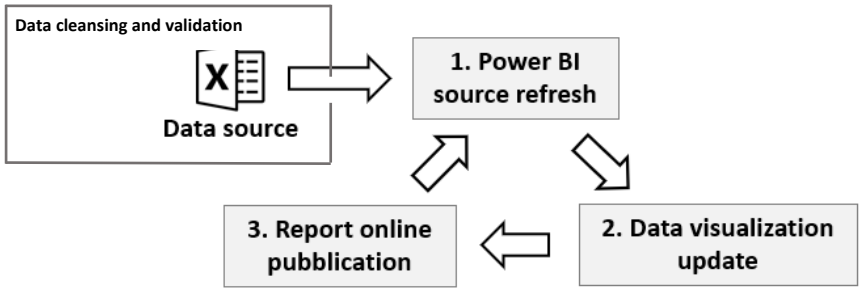

Figure 24. New reporting solution phases.

The new reporting process could be described as follows:

1. After validating the data with the stakeholder and cleansing the dataset, the data would have been stored in an excel file. The monthly updated dataset would have worked as a source for the developed PowerBI report. The report would been refreshed to upload the new data.
2. The new uploaded data would refresh all the available visualization in the report. It is important to mention that the visualizations developed in PowerBI were resembling the

[10] There was a known amount of survey respondents, since the stakeholder had a list of contacts to which the survey had to be distributed. Additionally, their role was known too, therefore it was possible to validate whether a certain respondent has answered the right group of questions targeting a specific role.

ones in the previous reporting solution in order to keep the same template and visualization of data.

3. After updating the visualizations in the PowerBI report, it would have been published online, using the PowerBI cloud infrastructure. The end users of the report would have been able to access the updated version by simply accessing a link to the online report. PowerBI capabilities would have allowed the end users to filter and interact with data to get more insights and a broader picture of the reality.

The project followed a Modified Waterfall methodology[11] (Sridhar, 2015) which was better fitting the reality of the project and was allowing for more flexibility during the project development steps. Despite a multitude of system and software development methodologies (as well as their variations) commonly adopted on the market, the modified waterfall methodology sufficiently fulfilled the requirements to achieve the project scope. The key elements that contributed to the choice of this methodology were the linearity of the project, the initially clear objectives and requirements, as well as the medium complexity of the project. Additionally, the presence of only few stakeholders involved in the development of the solution did not require a more complex methodology, since a more complex methodology would have simply increased the complexity of the solution development processes.

Following the Modified Waterfall Methodology, Project DP, was developed through the following steps (figure 24):

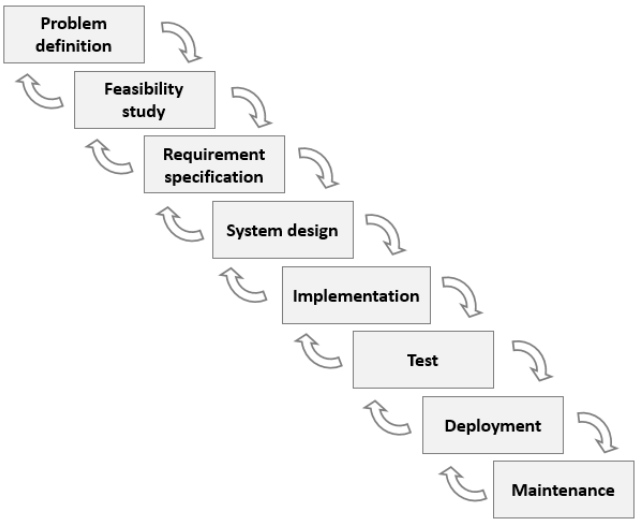

Figure 25. Adapted modified waterfall methodology.

As stated at the beginning of this chapter, the defined problem was that of the inefficiency of the old reporting methodology (see above for the problem description). Consequently, the goal of the project was to develop a new reporting solution that would allow to automatize the whole reporting process; in this way optimizing the activities needed to generate the monthly report and adding a

[11] The Modified Waterfall Methodology is a system development methodology that is inspired to the Waterfall methodology, but which fits better the reality. The main difference between the "traditional" waterfall methodology and the modified one, is the fact that the latter is more flexible since it allows to go back to the previous steps and make correction, while the former is more linear without the possibility to turn back on owns steps

higher flexibility in terms of insights generation. The “Problem definition” phase was critical since it allowed to engage the stakeholders into the project and get their permission to start the development of a new solution.

The feasibility study involved taking into consideration the whole process and documenting all the steps of the reporting in order to conceptualize the architecture. It was important to take into consideration the structure of the dataset, variable types, possible occurring errors, possible ways of visualizing the data in order to understand whether it would be possible to automatize the whole process and to understand which tool to use for this purpose. The data source initially being an SPSS format could be transformed in an Excel file, which made it readable for tool such as Power BI. The dataset represented a table having its rows (tuples) corresponding to the entries of each survey respondent and respectively the columns (attributes) being the variables representing each question and the respective answer to it. The intersection between each row and column representing respectively the answer given by a respondent to a specific question. The structure of the dataset made it suitable to become a source to the Power BI solution and hence to be analysed and visualized. Since the survey providing pre-coded answers, the data errors were reduced to a minimum risk. Additionally, Power BI offering a wide range of data visualizations, allowed to replicate the visualization of old reporting solution. Following the above considerations, the project was defined as feasible and consequently the analyst could proceed to its development.

During the “requirements specification” phase, the key stakeholders of the reporting solution were involved in order to align on the project scope, define the success metrics and define the requirements. Alignment and requirement definition were focused on defining the new template of the report, data to visualize, and visualizations to be used for this purpose. Some examples of the mandatory requirements can be listed: (i) Reporting solution to be updated on monthly basis; (ii) Visualization or description the survey demographics; (iii) The core part of the dataset to be visualized using the same graphs as in the old solution: 100% stacked bar charts, representing the satisfaction related to the specific questions, and bar charts to represent distributions; (iv) Display the open-ended questions; (iv) Creation of summaries for categories of questions, using 100% stacked bar charts.

The new reporting solution was developed through the following steps:

1. PowerBI report was sourced by an excel file containing the survey dataset. The whole visualization universe of the tool could be used now to explore and visualize the data. In order to avoid future errors during the data refresh, the PowerBI file was linked to the table containing the dataset in the spreadsheet. By maintaining the link to a defined table, it would be possible to refresh the data without the need to update every time the references. Additionally, using the data table as a source, it would help to allow the load of empty rows or data present on the excel spreadsheet.
2. Power Query feature was used to query the data and define the appropriate data types (figure 25). Despite the fact most of the attributes store numbers (ranging from 1 to 6, corresponding to the satisfaction scale), all of the data types were set to a text format. This was done because there was no need to perform operations such as sum or average. Because defining the variables correspondent to the questions as text, the only summary statistics possible to be calculated were count, count distinct and calculation of percentage of grand total. The table format kept its original structure

and was not unpivoted. The latter choice would have had restricted the interaction between different attributes representing the answers of a respondent (figure 26). By filtering on one attribute it was essential to be able to visualize all the other attributes belonging to a specific record. In other words, to perform a filter across all the variables. An important feature that would have allowed to automatize this process was the fact that PowerQuery registers all the steps that have been done in order to clean and structure the dataset. The applied steps for instance can be viewed in the right-side panel called "Applied steps". Once a new data source would have been applied, the data would have been processed through the registered steps making not necessary any other modifications to the data. By excluding the human factor from future data manipulation, it was possible to guarantee coherence over time in data formatting and changes.

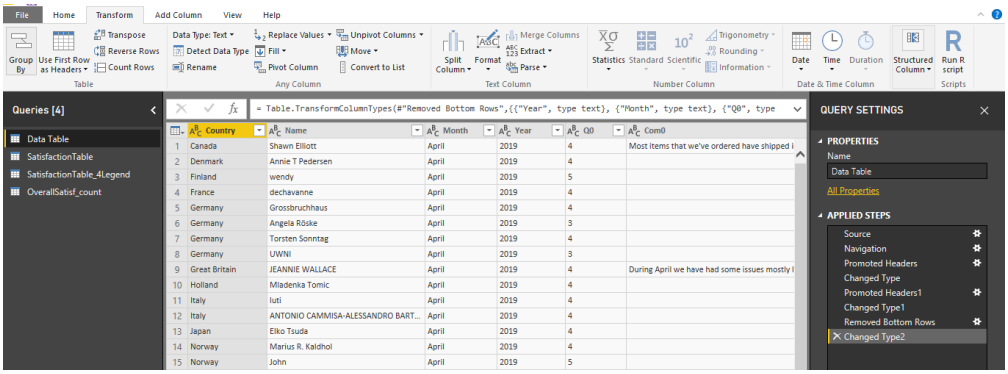

Figure 26. Power Query editor. Source dataset. Source: Power BI, Project E.

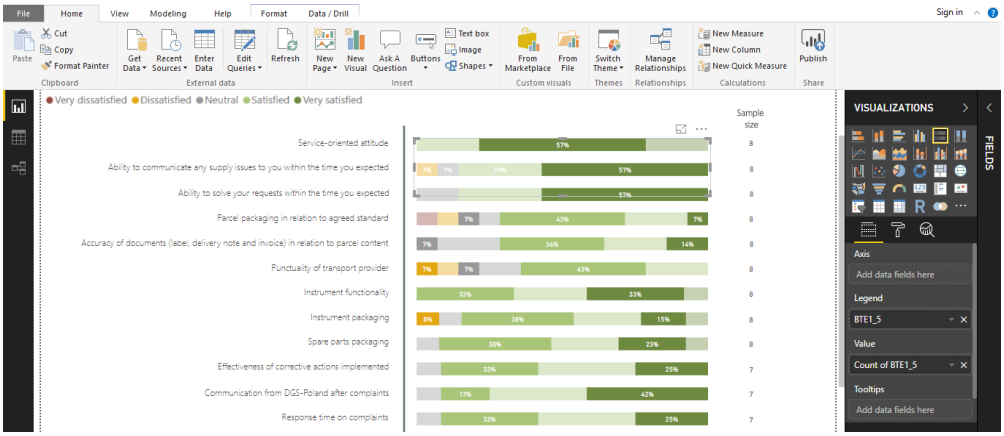

Figure 27. Power BI editor. Interaction between attributes. Source: Power BI, Project E.

3. The successive step consisted in creating a relational data model. Taking into consideration that the attributes' values were unlabelled, it was necessary create labels for a better interpretability of the charts[12]. To label the values of the attributes, new data table containing the dictionary were created (containing one column with

[12] The labelling was necessary to make the charts of an easy interpretation. For example a chart representing the distribution of values from 1 to 6 would carry no insights to the end users without knowing what exactly those values mean.

the unique value id and another with the label[13]). The link between the values and labels was done by establishing a *Many to One* relationship (figure 27) between the dataset and the table storing the labels (1, 2, 3, 5 and 6 acting as Primary Key to establish the relationship).

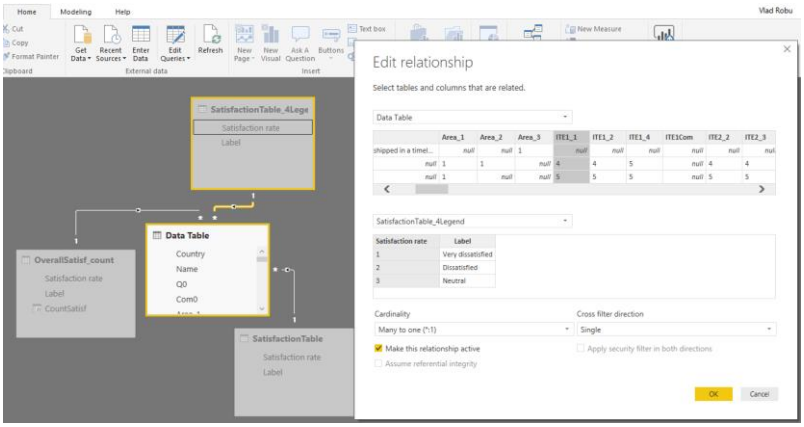

Figure 28. Relational model. Many to One relationship. Source: Power BI, Project E.

4. After creating all the necessary tables to map the data, it was possible to proceed to the development of data visualizations, following the previously defined requirements. The 100% stacked bar charts (figure 28) were showing the distribution of satisfaction per each attribute as a percentage of the grand total. The visualization design allowed the end user to easily compare among different attributes belonging to the same group. The interactivity of the charts allowed to apply visual filters and inspect the way the results for one attribute were impacting the results in another. Card visual were applied to summarize the sample size of each question, while tables were used to visualize the data from open-ended questions (figure 14). By interacting with the open-ended questions, it was possible for the end user to get an instant idea of the satisfaction level of that specific respondent (an operation not possible with the old setup).

[13] The Likert satisfaction scale was labelled as follows: 1 = Very dissatisfied, 2 = Dissatisfied, 3 = Neutral, 4 = Satisfied, 5 = Very satisfied, 6 = Not applicable / Not relevant

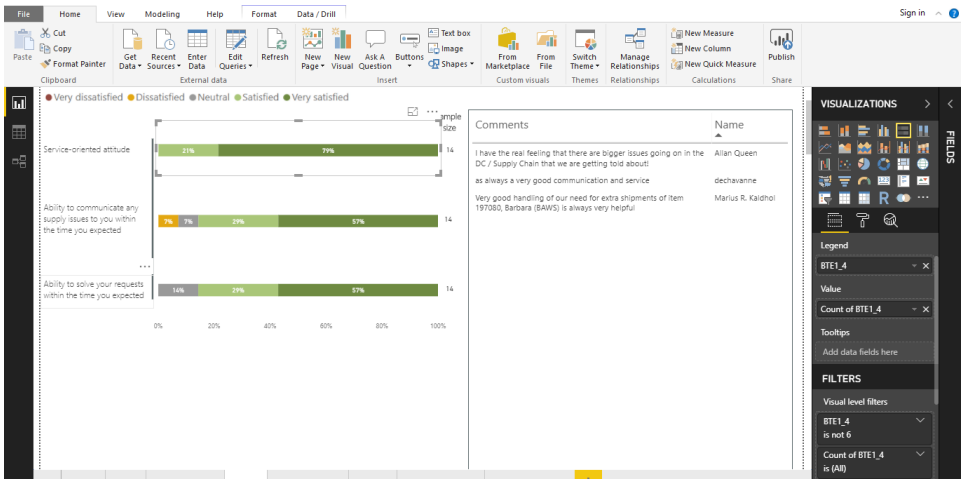

Figure 29. Example of survey data visualization with Power BI. Source: Power BI, Project E.

Other types of charts were used in order to create the reporting solution and to make it dynamic (with elements updating at each refresh of data). For example, histograms were used to visualize the overall distribution of the satisfaction as well as to show the countries participating in the survey (for a more immediate interpretation it was decided not to use the map chart), while a table was used to summarized key information about respondents and spot anomalies (figure 29). Card charts were used to integrate the text present in the report sheets, changing every time a change in data would occur. Cards were used for describing elements such as month, year, count elements, etc (figure 30).

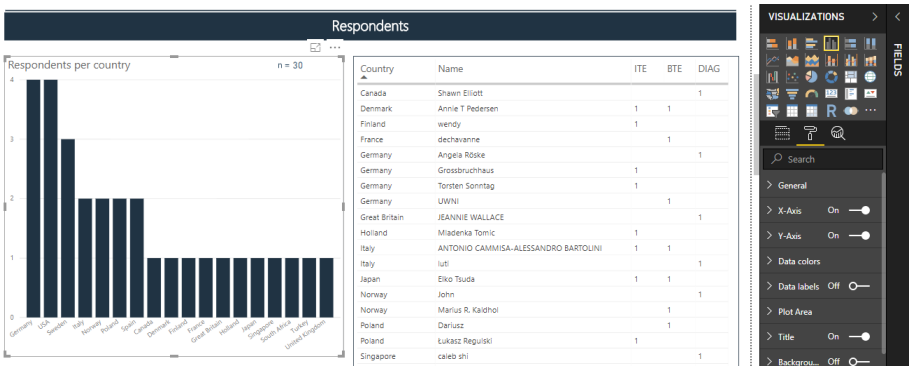

Figure 30. Example of histogram and table used as data visualization in Power BI. Source: Power BI, Project E.

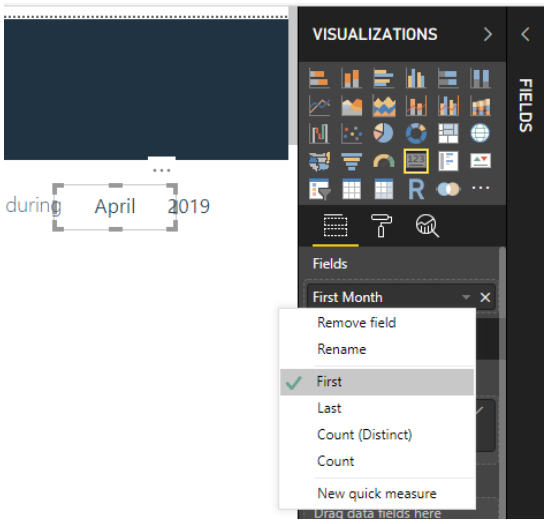

Figure 31. Card chart Power BI visualization. Source: Power BI, Project E.

Once the report was designed, a new reporting solution app was generated in the PowerBI online workspace. This operation was necessary in order to publish online the report (figure 31). Having the reporting solution online would have allowed the end-users to access it by simply using link generated through the app publication. In case of new users wanting to access the report, a dedicated access had to be created. Besides a better management of the end-users, having an online report would have enabled the users to always have the access to the latest version of the report, avoiding endless sharing via mail. This was a milestone in creating a unique point of truth and simplifying its delivery.

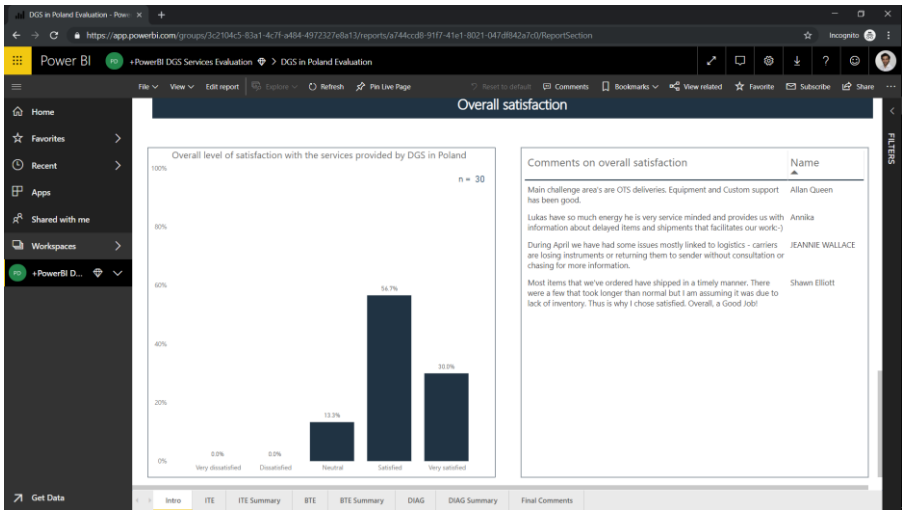

Figure 32. Power BI workspace. End-user online report visualization. Source: Power BI, Project E.

5. After completing the reporting solution design, several tests were done to verify whether it was working as expecting at each data refresh and there were no bits of information lost. Before deploying, a last confrontation was done was done with the

stakeholder to certify that the solution was meeting the expectations and requirements.

6. With the management and stakeholder's approval it was possible to deploy the new solution globally. After being deployed, the update of the report was requiring only a few minutes, significantly decreasing human engagement in this process.
7. Finally, the analyst was responsible to maintain and update monthly the report as well as to grant access to new users. He had the additional responsibility to answer and solve any technical questions solution related.

Overall the project had a high success, by significantly facilitating the reporting and insight generation processes across departments and globally. Meanwhile, the analyst intern had the possibility to apply business intelligence concepts, enriching the domain knowledge, in order to optimize data analysis processes and meet business needs.

4.4. PROJECT GENIE

Projects	Year	2018																																2019																							
	Month	August					September					October					November					December					January					February				March					April					May										
	Week of the year	31	32	33	34	35	36	37	38	39	40	41	42	43	44	44	45	46	47	48	49	50	51	52	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22											
Project G - Genie	Business & case understanding																																																									
	Data retrieval																																																									
	Data exploration & understanding																																																									
	Data preparation																																																									
	Modelling																																																									
	Evaluation																																																									
	Deployment																																																									

Table 10. Gantt chart of Project Genie.

Project Genie represents the largest and most challenging project developed during the internship. It started as project aiming to simply explore a database[14] which collected data from the company's hearing aid fitting software. After achieving the explorative goal, the project scope expanded to become a data mining project, aiming to identify hidden patterns that could become foundation for future managerial decisions. Considering the theoretical framework developed in chapter 2, the project was managed using CRISP-DM methodology[15]. Throughout the project, which had a duration of almost eight months (table 10), many side tasks emerged as per request of management involving *ad hoc* data exploration and analysis; nevertheless, this chapter will focus only on the data mining aspect of the project. It is worth mentioning that the data mining approach to the project was a trainee's initiative, and therefore the project was completely run and developed by the trainee himself. Concluding, this project led to the writing of a research paper authored by Vlad Robu and co-authored by Vitor Duarte dos Santos, which was submitted to the 6th International Conference on Systems and Informatics (ICSAI 2019) - Lingang, Shanghai, China (*see the research paper as Appendix 7*).

***Disclaimer note**: Because of the confidentiality of the information contained, variable names, data sample, software code containing variables' names used to perform data manipulation, rule output, will not be presented in this internship report. Nevertheless, if necessary, requests can be evaluated, and details can be presented under special terms and conditions.*

4.4.1. Business understanding

The journey of a hearing aid involves two important stakeholders of the hearing care industry and a fundamental step. The two stakeholders are the hearing care professional (the audiologist) and the hearing-impaired person (figure 32). Before a hearing-impaired person can fully benefit from its newly bought HAs, they have to be fitted to the client's specific hearing loss and sound preferences (e.g. in a given environment some people may prefer a sharper sound while others may prefer a softer one). During a fitting session, the HCP uses the company's fitting software and the available

[14] The initial goal was that of simply exploring the database after certifying that it could store valuable data related to HA users and digital HA fitting session. After a reporting solution created in the past that was not anymore in use, the data stopped to be leveraged to extract data driven insights for business decisions.

[15] CRISP-DM methodology includes the following steps: 1) Business understanding; 2) Data retrieval; 3) Data exploration and understanding; 4) Data preparation; 5) Modelling; 6) Evaluation; 7) Deployment.

features in the software to program in a personalized way the hearing instrument[16]. The fitting process represents a crucial step in one's experience with the HAs: (i) Without an adequate fitting the user won't be able to benefit from the full sound experience of the hearing instrument. A bad fitting can become a frustrating experience with the product, which can later cause a switch to another brand, or even worse, the hearing-impaired person will be discouraged from trying a solution ever again. Therefore, a competent audiologist and a well-functioning software are essential; (ii) A software that covers a wide range of tools and functionalities allows the audiologist to personalize as much as possible the fitting for each user. This translates both into a user that will be satisfied with its HAs and into an HCPs that can perform its professional activity without the frustration of using a poorly performing or limited software (avoiding as a consequence both unsatisfied user and HCP).

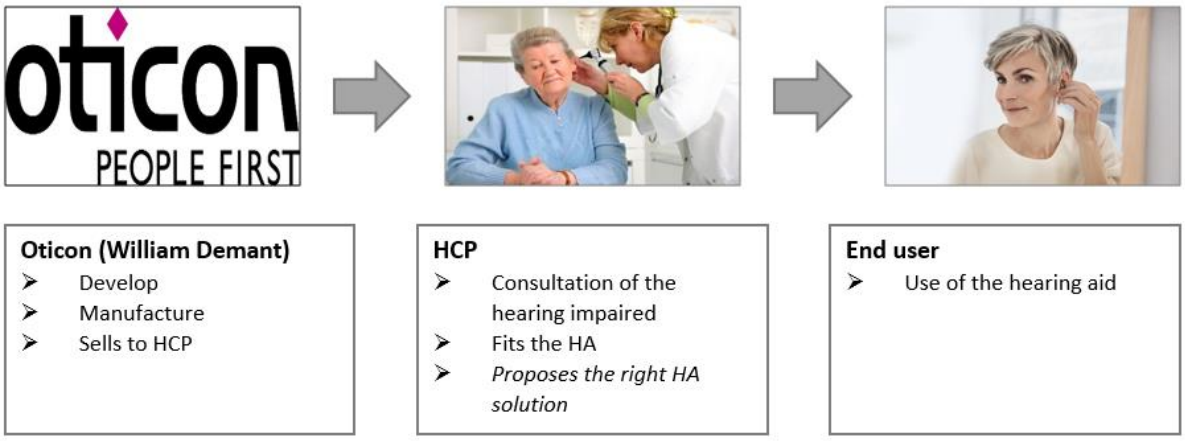

Figure 33. The journey of the HA from the manufacturer to the end user. Source: own adaptation.

An additional, less obvious, reason of the importance of a fitting session is the type and amount of data it generates. Considering that the software used for the fitting records almost all the steps done from the beginning till the end of a session, it generates data on key elements of the business: (i) Data about the client (demographics, medical condition, etc.); (ii) Data about the HA (type used, programs, features, etc.); (iii) Data about how the software was used, i.e. what were the main actions of an HCP within the software to perform the session. It is obvious how this data represents a golden mine for business insights, knowledge generation, and business strategies based on data. Considering that the data is recorded worldwide and the sample of observations being the whole population (which can be millions of fitting sessions), the insights achieved through different analysis come to be very robust and representative (compared to insights based on a sample of few observations).

The fitting software developed by Oticon is called Genie. An example of how it looks like is shown in figure 33. On the top part of the screen a ribbon with the main phases of the fitting sessions is present, while for each phase a set of tools and options are available on the left side of the screen. By selecting the tools and adequately setting the parameters, the HCP can fits the HAs. The central area of the window visualizes the data representing the output of the fitting or other details regarding the

[16] The fitting of the HAs takes into consideration factors such as hearing loss severity, HAs style, sound preferences of the user and in some cases lifestyle.

hearing aid and sound preferences. Therefore, the software logs the events (clicks done by the HCP) happening during a digital fitting process, besides registering other valuable information (as mentioned above).

The logged events become this way a precious pool of data that can be extracted, transformed and queried for analysis related to end users, markets, products and their fitting, software usage, etc. The insights gained from the analysis of this data can be combined with business trends, market development and become actionable and integrated into business strategies. The possibilities are limitless.

Figure 34. Genie software interface. Source: www.google.com; search key: "genie 2 software".

4.4.2. Problem identification and definition

An HCP spends a considerable amount of time fitting its clients. Every session requires the HCP to go through all the steps of the fitting software and program the HAs based on the user's preferences. Among the sessions, most of the cases tend to be similar: similar hearing-loss severity, similar hearing aids choice, similar sound preferences, similar choice of extra features. Meanwhile, the HCP still has to go through all the steps and individually select these preferences. This results to be time consuming and inefficient most of the times.

What if it would be possible to analyze the most common tools used by the HCP in the software? Based on this, what if it would be possible to personalize the experience of the HCP by already suggesting the most preferred fitting tools[17] throughout each fitting step, this way saving time and simplifying the process itself?

Identifying the most common steps done by the HCP in the software could open the possibly of offering a personalized set of steps to be performed, tools to be used and features to be set - avoiding therefore unnecessary clicks and waste of time. Suggesting only the most frequent

[17] A tool in the fitting software represents a button or an operation that allows to perform a specific group of settings on hearing aid.

behavioral patterns could optimize and make a session more pleasant and more efficient process. Additionally, identifying the most common patterns in the fitting software usage could help to identify which of the implemented software tools are used the most or on the contrary, are not used at all.

The initiative of mining the most frequent patterns within the software tool would transcend multiple stakeholders, and potentially bring benefits such as: (i) gaining knowledge of the most common used combination of tools; (ii) optimizing the flow of a fitting session; (iii) potentially reducing the time needed to run a session; (iv) creating data-driven awareness on which patterns are most present and which less, allowing to explore consequently the reasons; (v) targeting better the efforts of product owners, developers and business analysists when planning future developments of the fitting software; (vi) etc.

Based on the above considerations, the purpose of this data mining project is that of identifying the most frequent patterns within the software tools usage by the HCPs.

For this purpose, the Apriori algorithm was used. The choice of the algorithm was based on the careful review of the literature treating similar problems. Despite the presence of many frequent pattern algorithms, there are several factors that weighted in favor of Apriori:

1. Despite being one of the oldest algorithms developed for such problems, it is still widely adopted by researchers and companies to mine frequent patterns and develop Market Basket Analysis[18].
2. It benefits from a large literature and research communities treating about the algorithm, its applications, its problems and solutions
3. Many frequent pattern algorithms have been developed to improve the Apriori approach and overcome its limitations. While most comparisons are done based on the computational efficiency, rarely something is stated about the quality of the identified rules. When disregarding the computational efficiency, Apriori delivers the same output as other more advanced algorithms.

   Indeed, during the project it was possible to observe that the computational effort increased when significantly lowering the minimum support threshold (min_sup = 0,0001), i.e. making the algorithm to run on a higher number of item-sets. The effort may increase as well when working on larger datasets and/or working with low support and/or confidence thresholds. When ECLAT[19] was run on the same dataset, keeping the same parameters, it had a better performance in terms of time required for the output. On the other hand, when comparing the quality of the outputs of the two algorithms - Apriori and ECLAT, the two delivered the same rules when comparing the measures[20].
4. Because of the way Apriori works and the simple calculations to mine the association rules, it allowed the definition of a multitude of interestingness measures compared

[18] Although the data mining problem at hand is not centred on transactions involving goods, it still resembles a Market Basket Analysis by aiming to identify the most frequent patterns among the transactions.

[19] The test has been run by using a support threshold of 0,1, both for Apriori and ECLAT.

[20] Even though Eclat overcomes the computational limitations of Apriori, it doesn't offer a range of measures to evaluate the rules as wide as Apriori.

to the fewer ones of other algorithms. These interestingness measures both help to overcome the limitations of Apriori and to enhance the evaluation of the rules as well as the generation of insights.

5. In conclusion, because of not having an extremely large dataset and not aiming to use extremely low support or/and confidence threshold, it was opted for the adoption of a simpler solution rather than a more sophisticated and complex one. On one hand, a simpler solution allowed to program in a simple and clear way the algorithm. Leaving this way space for future scalability and increase of complexity based on the needed solution (and hence better control the development of the project).

4.4.3. Data understanding and retrieval

The data logged from the software is stored in a data lake in the Azure infrastructure. From the data lake, the data is extracted, transformed and loaded into the Azure SQL database where it is stored under the form of a transactional database. Having the data structured in a database[21], made the data easier to query, analyze and visualize. When creating the database, the data engineers were able to eliminate most of the data anomalies through the ETL (Extract, Transform, Load) process Because of having a "reader" access type to the data, the only operations that were possible, were the data query operations - no additional table could be created nor other sources could be used to integrate the query with additional attributes.

The first exploration of the data was done in MS SQL Management Studio (figure 34). An additional support in understanding the variables and their meaning was given by the available technical documentation, containing information about the meaning of the variables, meanings of the IDs, and structure of the variables[22]. At first, it was important to understand how a digital fitting session is mapped in the dataset and which steps of the session are logged. Once this would have been done, it would have been possible to understand the dynamics of a fitting session.

[21] For this project the team has been granted a "reader" role to access the transactional database. The transactional database represented the final selection of a set of attributes from an underlying relational model (to which the access was not available).

[22] Despite having technical documentation at hand, it was related to one of the first software versions, hence it was unsure whether the newer software versions were having the same specifications.

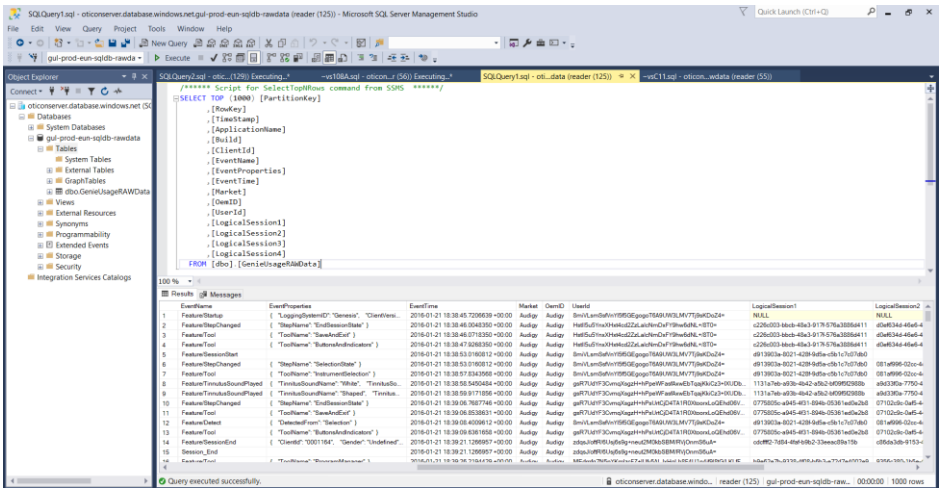

Figure 35. MS SQL Data Query - first 1000 rows. Source: SSMS, Project Genie.

For the first exploration, the SQL code was used to explore the first 1000 rows. It allowed to have a glance at the variables present in the database and understand their structure, hence it was possible to create expectations on what would be possible to achieve and would need further exploration.

Initially, the database has 16 variables. The types of variables could be distinguished in string and "date" variables. Among the variables, many of them were ID variables, storing IDs relative to the fitting session, HCP, client and fitting step. These ID variables were stored as string variables (nvarchar). Each row of the database represented a specific step of a fitting session, for example selection of a tool, change of the step, start and end of a session, occurrence of an invalid step, etc. Therefore, a session will be represented by a multitude of rows. The attributes that were not ID variables, were storing information regarding the occurring step. This information was stored in a JSON[23] format.

Based on the first data exploration it resulted possible to identify when the hearing care professional started or ended a session, when did it use a specific tools and which tools were used. Additionally, the variables compressed in a JSON format were storing information related to the details of the steps, to the hearing aid, to the patient's medical conditions, and much more[24]. In order to use the data compressed in JSON, it had to be parsed. For specific steps of the fitting session, the information stored in a JSON format could store many dozens of variables representing specific details of the step. Because of this, a partition of the format would have expanded the dimensionality of the dataset to hundreds of variables[25] and give the dataset a structure similar to the example in figure 35. Considering that the dataset had millions of rows, it would have been computationally expensive

[23] JSON (JavaScript Object Notation) is a lightweight data-interchange format. An open-standard file format that uses human-readable text to transmit data objects consisting of attribute-value pairs and array data types. The format is built on two structures: (i) A collection of name/value pairs; (ii) An ordered list of values. (Crockford, 2019)

[24] Due to the confidentiality of the information, no further details regarding the data stored in the JSON format will be given.

[25] The dimensionality would increase due to data partitioning. By only partitioning three variables each containing up to 20 variables stored in a JSON format, would increase the dimensionality of the dataset by 60 new attributes.

to parse and analyze the JSON format. For this reason, the partition was done in a second moment, after the data was filtered and only the necessary records were selected.

Session	Step	Var3	Var4	Var5	Var6	Var7
1	1	x				x
1	2		x			x
1	3		x			x
2	1			x		x
2	2				x	x
...	...	...	...	...	...	...

Partioned JSON data structure

Figure 36. Simplified conceptual representation of the dataset structure after JSON data partitioning.

Because of being part of a business framework, it was decided to run the project and test the data mining model only on the most relevant for the business data. The relevant data had to be filtered and extracted out of the dataset based on four key attributes: geographical zone, software version, session step name and date. To apply the filters, it was first needed to explore the levels of each of the attributes but date (since the levels of it would be each single second present in the database). For this purpose, the SELECT DISTINCT query was run on each of the first three attributes, an example of the script can be seen in figure 36. Following this example, a similar query was run to identify the distinct level of the "software version" attribute and of the "session step name" attribute. The immense size of the database can be deducted from the fact that almost three hours were needed to run a fairly simple query.

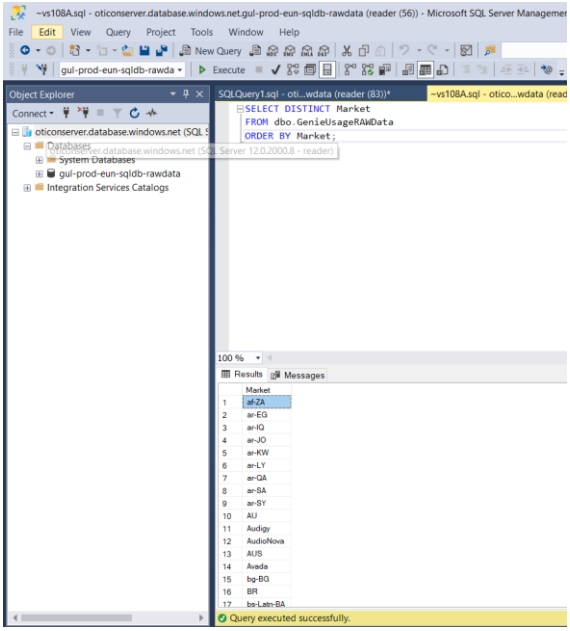

Figure 37. MS SQL Query - SELECT DISTINCT to identify the levels of the geographical attribute. Source: SSMS, Project Genie.

After identifying the levels of the key attributes, it was possible to select the ones of interest when filtering the database as well as to define only the necessary variables to be kept during the data extraction process (figure 37).

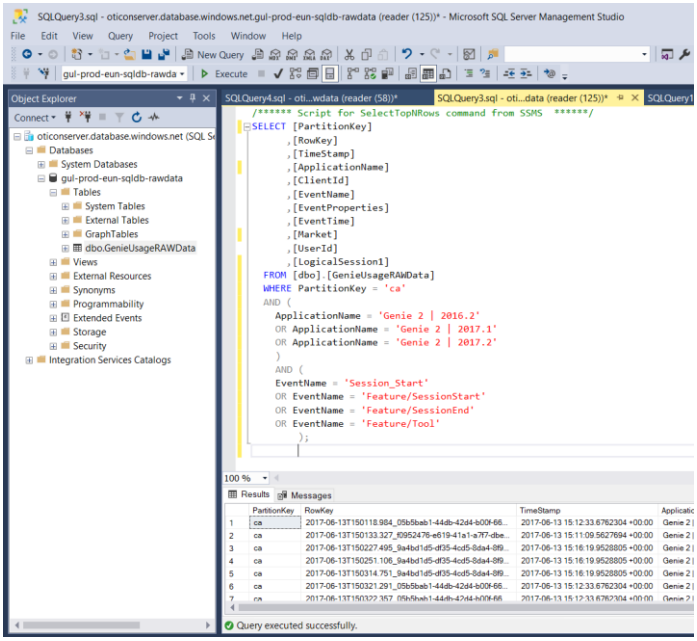

Figure 38. MS SQL Query - Selection of the data to be used for the data mining project. Source. SSMS, Project Genie.

Considering inputs from the management, the focus was set only on specific geographical areas of interest as well as on certain software versions; additionally, to reduce the size of the dataset only specific steps done during the session were selected. Having in mind the goal of the data mining project, only the relevant steps of the fitting session were extracted, specifically:

1. To develop a data model on quality data, it was important to ensure that the frequent pattern mining model would have been run on both complete and valid sessions. From the technical documentation available on the database it was identified that a session was valid only if complete. Consequently, it was decided to extract the steps of a session corresponding to the session's start and end[26]. Based on the session ID, it would have been possible later on to match the two steps[27]. Therefore, a complete session would have had both a correspondent Start and End step, while an incomplete one would have had only one of the two steps. Implicitly, the sessions without an id, would be excluded from the analysis.
2. Only the session step identifying the tool used by the HCP during the fitting was extracted. Based on the session id, it would have been possible to identify all the tools used during a specific session by a specific HCP. Once identifying the tools used during a session, it would have been possible to mine the frequent patterns in the tool's usage.

[26] There are two steps indicating the ending of a session. The reason is that one is recorded once the session is ended, the other one stores valuable information about the session in a compacted JSON format.

[27] The process of joining the steps of a sessions based on their unique Id is part of creating a dataset suitable for the application of frequent pattern mining algorithms.

The query was done only on the above four mentioned session steps. After running the query, a dataset of a total of 9.548.479 rows was extracted, which was consequently exported to a CSV file format. Storing the data in a CSV format would have allowed to store such quantity of data occupying a relatively small computational memory space. Afterwards, the file would have been used as a data source for data manipulations to be performed with Power Query and R (together with the algorithm implementation). Querying the data directly from MS SQL server would have significantly impacted the performance of analysis because of the computational limitations. It is important to notice that despite the possibility to run all the data manipulation steps in R, the choice of using Power Query for data structuring was due to the requirements of using an already available tool from the company's tool box, consequently allowing for scalability of the project and access to non-users of R.

4.4.4. Data preparation

Base on the GIGO principle, the data preparation step was necessary to cleanse the dataset and increase the possibility of a high-quality data output from the project. As mentioned previously, the first steps of data preparation were done in Power Query. The data manipulation was performed through the following steps:

1. The initial dataset extracted from MS SQL in a CSV format was explored to identify whether it was appropriate for the goal of the project. After visually exploring the content of the variables using PowerBI and the information carried by each of the them, it was decided to reduce the dimensionality of the dataset by removing the variables that would not contribute to the future data model. Only the variables identifying the session id, session step name and the variable storing the session step related information in a JSON format were kept. After reducing the dimensionality, only three variables of the initial dataset were left. Due to confidentiality figure 38 shows only an example of the dataset structure.

Session ID	Session Step Name	JSON
1	SessionStart	SessionStart
1	Tool	Tool 1
1	Tool	Tool 2
1	Tool	Tool 3
1	SessionEnd	SessionEnd
2	SessionStart	SessionStart
2	Tool	Tool 2
2	Tool	Tool 3
2	SessionEnd	SessionEnd
3	SessionStart	SessionStart
3	Tool	Tool 3
3	SessionEnd	SessionEnd
4	SessionStart	SessionStart
4	SessionStart	SessionStart
....	....	....

Figure 39. Simplified conceptual representation of the dataset after dimensionality reduction.

During the exploration of the variables, no inconsistencies were detected. This can be explained by the fact that the software logged automatically each step with no human interaction, reducing the error margin to minimum. Additionally, the reduced dataset had no missing values. Variables with missing values due to not logged information were removed during the dimensionality reduction (because not relevant to the model). It is worth mentioning that the data quality was checked for all the variables but for the one storing the data in JSON.

2. The next step consisted in partitioning the data from the JSON format. Nevertheless, the format was storing HA, medical and patient related information[28] - only the values corresponding to the tools were left for the analysis (the name of the tool corresponding to the "tool" step). After performing the partition, the dataset kept the same dimensionality.

3. A second data exploration was done to ensure the quality of the data. At this point the dataset structured was as follows: (i) A variable storing the session id. Each id representing a unique session. Nevertheless, the unique id of a session was present on multiple records, corresponding to the session steps; (ii) A categorical variable containing the name of the session steps; (iii) A categorical variable containing the name of the tools.

   Again, a search for missing values and inconsistencies was done. Some values were found to be recorded in the wrong variables, risking causing inconsistencies in the final output. From previous data explorations, the errors were identified to belong to invalid sessions, test sessions or incomplete sessions. When eliminating from the analysis the latter type of sessions, these errors in data would be removed too.

4. The last step of the data manipulation consisted in structuring the dataset for the application of the Apriori algorithm.

   i. Ensuring that the unsupervised learning would have run only on complete and valid sessions.

   ii. Ensuring the appropriate structure of the data for the algorithm. Even though being a transactional dataset, it didn't have an adequate structure for the applications of Apriori.

   The Apriori algorithm requires to be applied on a transactional dataset. The transactional dataset shall have each record representing a transaction, while the items in the transaction are represented through binary variables (where 0 = absence of the item in the transaction, 1= presence of the item in the transaction). Consequently, the Genie dataset had to be structured in such a way that it would store each session as a single record, and each tool used during the session will have to be represented by the binary variables. The dataset acquiring a structure as in figure 39[29].

[28] The additional information stored in JSON format but not essential for the project, was used by the company to perform ad hoc analysis on markets, HCPs and HA users.

[29] Due to data confidentiality only the concept of the actual dataset is shown.

Session	tool 1	tool 2	tool 3	tool 4	...
1	0	0	1	0	...
2	1	0	1	0	...
3	0	0	1	1	...
4	1	1	0	1	...
...	...	...	...	...	...

Figure 40. Simplified conceptual representation of the binary transactional dataset.

To structure and at the same time cleanse the dataset. The following operations were performed on the original dataset using Power Query.

1. Given the data structure, it was needed to create a single dimension corresponding to each step type in the dataset. This was possible by using the "session id" as a unique id, that was used later for table join operations. This data manipulation step was done in order to overcome the limitation of having all the tools in a single variable. An example of the concept can be seen in figure 40.

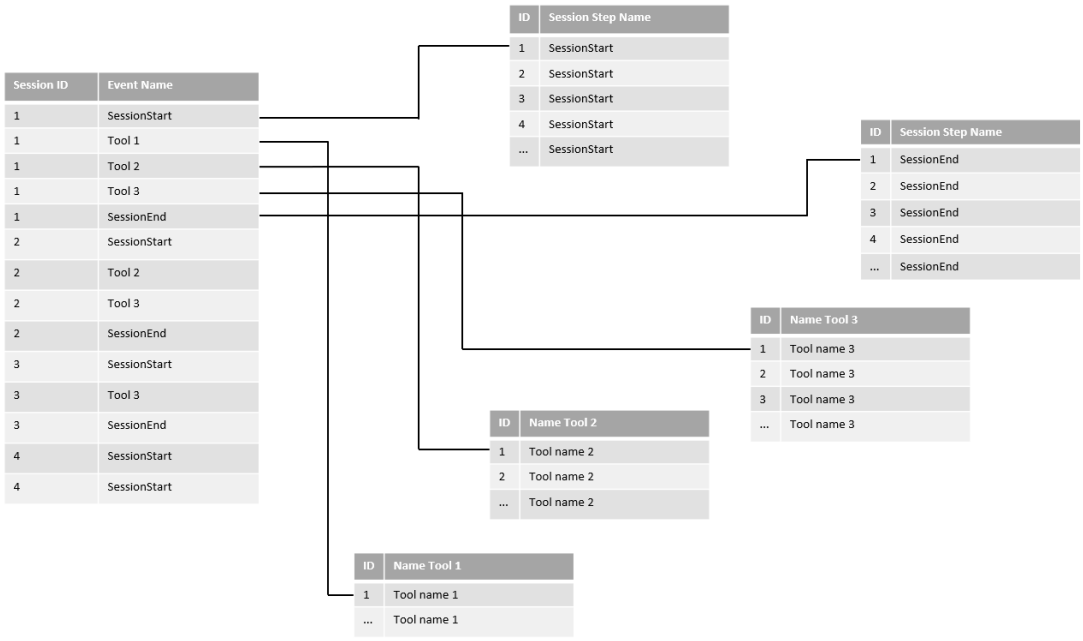

Figure 41. Simplified conceptual representation of the tool dimension creation process.

2. Afterwards, a first table join was performed using an "Inner Join" type based on the "session id" variable (figure 41). The "InnerJoin" was done between the "SessionStart" and "SessionEnd" variables and allowed to select only sessions having both a "Session Start" and a "Session End".

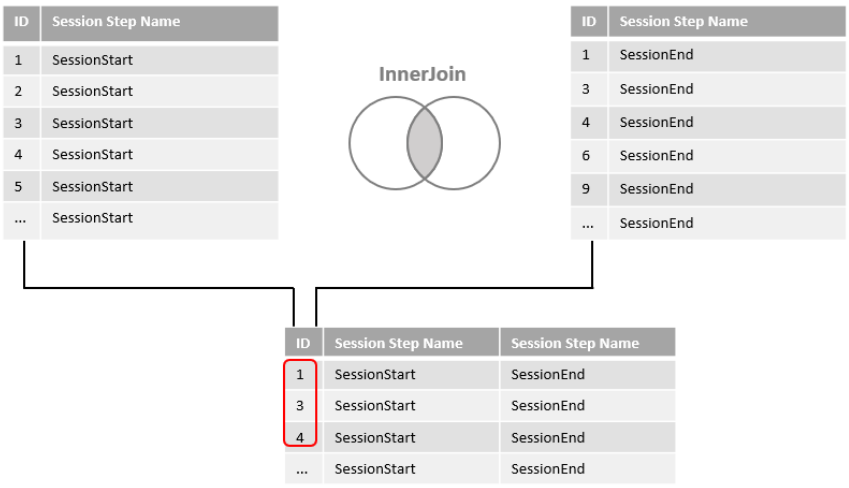

Figure 42. Simplified conceptual representation of the Inner Join performed to select complete and valid sessions.

3. Finally, after selecting the complete and valid sessions it was possible to create the required data structure for the application of the Apriori algorithm. The structure was created by using a "LeftOuterJoin" on the "Session ID" between the sessions and each tool dataset. A simplified example of this step is showed in figure 42. By running this operation, it was possible to identify which digital tools were used or not used during the specific digital fitting sessions

Figure 43. Simplified conceptual representation of the LeftOuterJoin performed to select digital fitting tools used during a digital fitting session.

After the merging phase, two separate datasets were obtained with digital fitting sessions recorded in two different key markets. The one dataset contained a total of 25.239 transactions while the other a total of 27.699 transactions - in total 52.938 transactions.

4.4.5. Modelling

RStudio was used to model the association rule using the Apriori algorithm, which was selected based on the reasons listed in paragraph 4.4.2. It was used as well to make comparisons between the identified rules using Apriori and Eclat.

The modelling process occurred as follows:

1. First, the source was set by using the previously created two transactional datasets. Consequently, the two datasets were appended to each other, creating a unique transactional dataset of 52.938 rows (see appendix 1 for the code used in RStudio).

2. Before applying the association rules to the data, a few data manipulation had to be done:

    a. Because the variables were automatically identified by R to be of a "character" type, they had to be converted to a "factor" type - required for the application of the algorithm. For this purpose the "factor()" function was used (an example for a single variable can be seen in appendix 2).

    b. The variables not being part of the transactions (e.g. Market and TID) were removed from the dataset together with the variables with no values (appendix 3). Since Apriori runs based on minimum support and confidence threshold, empty variables would have never figured in the data model, hence the decision to remove them. Out of the initial 31, only 26 variables were kept.

3. In order to apply the algorithm, the necessary software packages had to be installed in R - Arules and ArulesViz packages, and the dataset had to be defined of a transactional type using the "transactions" function. After which it was possible to apply Apriori using the following code:

```
install.packages("arules")
install.packages("arulesViz")
library(arules)
library(arulesViz)
library(dplyr)

tr <- as(cl.genie2, "transactions")
ar.genie <- apriori(tr)
```

    By running the algorithm with the default parameters[30] first output consisted of a set of 312 rules. Nevertheless, before exploring the set of rules, redundancy had to be eliminated[31]. The redundant rules were first identified and then removed from the analysis the code below. Despite using confidence as a measure to eliminate redundancy, other measures such as support and lift can be used.

[30] Apriori parameters default threshold are: minimum support = 0.1 and minimum confidence = 0.8

[31] Apriori is subject to redundancy, consequently it can identify rules that contain represent identical information, therefore is good practice to eliminate the redundancy from the output.

```
genie.redundant <- is.redundant(ar.genie, measure="confidence")
clean.arules <- ar.genie[!is.redundant(ar.genie)]
```

The cleansing process left only 85 association rules. By using the "summary", "inspect" and "plot" functions it was possible both to view the identified rules and to visualize them.

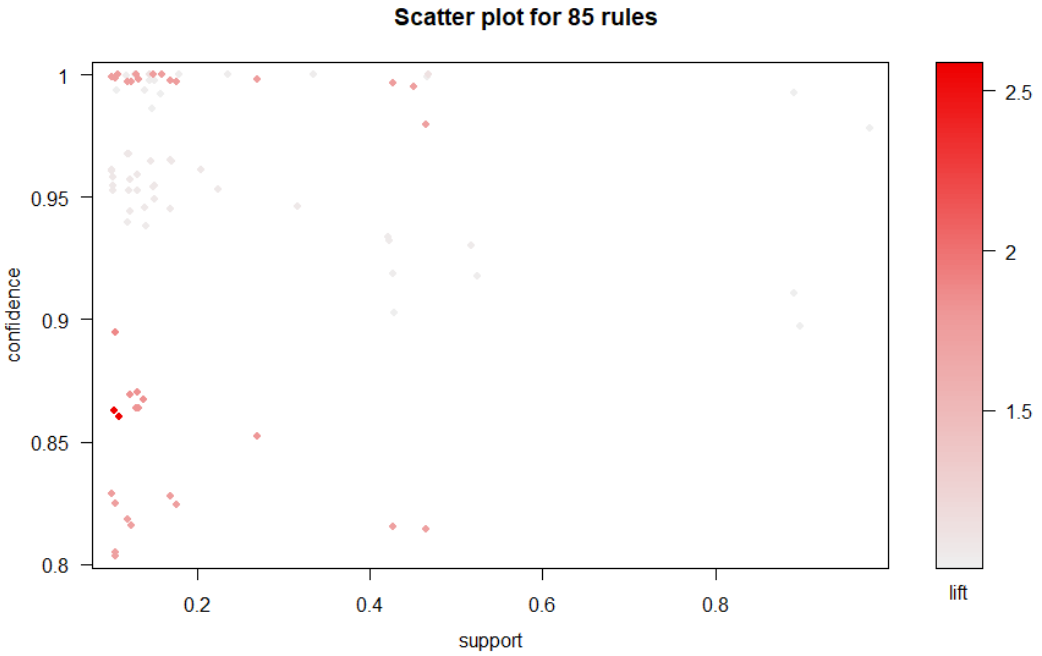

Figure 44. Association Rules by default parameters. Source R: software; Project Genie.

To be able later to evaluate the rules from different perspective, additional interestingness measures were added (see paragraph 2.5.1.): ChiSquared, Kulczynski, Cosine and hyperLift (used for very low support thresholds).

```
quality(clean.arules) <- cbind(quality(clean.arules), kulczynski =
interestMeasure(clean.arules, measure = "kulczynski", transactions =
tr))

quality(clean.arules) <- cbind(quality(clean.arules), cosine =
interestMeasure(clean.arules, measure = "cosine", transactions = tr))

quality(clean.arules) <- cbind(quality(clean.arules), hyperLift =
interestMeasure(clean.arules, measure = "hyperLift", transactions =
tr))

quality(clean.arules) <- cbind(quality(clean.arules), chiSquared =
interestMeasure(clean.arules, measure = "chiSquared", transactions =
tr))
```

Considering the goal of the project being that of identifying interesting and frequent patterns in the dataset (i.e. identifying which are the most frequent and interesting patterns in the fitting tool usage), it was decided to lower the min support in order to allow the search of less

frequent but interesting patterns - evaluating a larger amount of rules. Therefore, the min support was lowered to a minimum threshold of 0,001 (rules that occur in 0,1% of the transactions), while the min confidence threshold was left at 0.8 (80% of the transactions containing a would have contained B)[32]. Additionally, in order to understand how Apriori performed with different minimum thresholds, it was tested for the following values: 0.001; 0.005; 0.01; 0.05; 0.1; 0.5. To have a comparison against another algorithm, ECLAT was tested for the same minimum support thresholds.

For ECLAT, the following code was applied - respectively changing the minimum threshold at each iteration. The "minlen" parameter was set to 2 in order to obtain rules containing at least 2 items.

```
ar.genie.eclat <- eclat(tr, parameter = list(support=0.1, minlen=2))
```

The following paragraphs will explore the algorithm performance and rule evaluation and selection.

4.4.6. Model performance

Note: *A more detailed model performance analysis and comparison can be viewed in Appendix 7 which contains the research paper written by the author of this internship report - Vlad Robu and co-authored by Vitor Duarte dos Santos. For the sake of the length, in this chapter only the core findings from the algorithm comparison and performance were highlighted.*

After running Apriori and ECLAT with the different support thresholds (specified above), the performance of the algorithms is summarized in table 11.

Min Support Threshold	Apriori			ECLAT	
	Nr of total rules	Nr of not redundant	Time (s)	Nr of total rules	Time (s)
0.001	686981	37442	1.35	154242	0.45
0.005	90280	7993	0.31	26735	0.08
0.01	26347	2616	0.17	9341	0.03
0.05	1213	199	0.06	633	0.02
0.1	312	85	0.04	187	0.02
0.5	8	6	0.01	4	0.01

Table 11. Comparison of Apriori and ECLAT algorithm performance.

Confirming the knowledge emerging from the literature review, Apriori struggled at a lower threshold values while ECLAT was more efficient and required less computational power and time mine the most frequent patterns. Additionally, Apriori is less parsimonious than ECLAT, being propense to find a significantly higher number of rules since it is subject to redundancy. As a

[32] The confidence was not lowered because it would have meant to significantly increase the number of the identified rules by taking into consideration less important rules.

consequence of eliminating the redundant rules (see paragraph above), there is a higher possibility to store only more meaningful and interesting rules. It is interesting to notice how after eliminating from analysis the redundant rules, there are much fewer rules left then ECLAT. Indeed, when considering the rule length distribution[33], Apriori generates a normal distribution while ECLAT a more skewed one (appendix 4). This means that by keeping the not redundant rules for Apriori, it will be optimized and deliver rules with a more optimal number of items. Because of how ECLAT is calculated there is no formula eliminating redundancy, therefore it is up to the researcher to identify the redundant rules, if any. Despite the fact that with higher minimum support thresholds Apriori increased its efficiency, ECLAT was more efficient in identifying the rules.

As stated in the previous paragraphs, it is important mentioning that regardless the performance of the algorithms in different situations, after exploring the quality of the rules with the highest support and the same amount of items, the same set of rules was identified. To explore the first 30 rules, the following functions were used (see the appendix 4 for a more detailed performance of the algorithms):

```
sorted.arules.001 <- clean.arules.001 %>% sort(by='support')
inspect(sorted.arules.001[1:30])


sorted.eclat.001 <- ar.genie.eclat.001 %>% sort(by='support')
inspect(sorted.eclat.001[1:30])
```

While the first quartile defined on support values might not change too much, there is a slight difference for the 4th quartile of rules. Apriori tending to distribute rules with higher support in the last quartile.

If there is a lower efficiency in Apriori compared to ECLAT, it is balanced by the whole set of additional interestingness measures that can help the researcher to evaluate the rules. Having different evaluation parameters rather than “Support”, makes Apriori to offer a more complete picture of the interestingness of measures and help during the rule selection process.

4.4.7. Rule evaluation and selection

In order to identify interesting and frequent pattern in the dataset, Apriori rules evaluation was enriched with additional interestingness measures. In the end, the analyst intern relied on the following measures[34] to evaluate the rules in an objective way: (i) Support; (ii) Confidence; (iii) Lift; (iv) Count; (v) Kulczynski; (vi) Cosine; (vii) HyperLift; (viii) ChiSquared. Among these only Kulczynski results to have the null-invariant property, while the other parameters may result affected by the number of rules (i.e. may tend to skew). It is a common rule that the selection of the rules doesn’t depend solely on objective measures, but their selection and interpretations depend as well on the

[33] For this project, both for Apriori and ECLAT the max length of the rule was kept to 10 items, which is the default parameter.

[34] The calculation and interpretation of the mentioned interestingness measures can be viewed in the “Theoretical Framework” chapter.

business knowledge of the researcher. Hence, the selection of the rules was done based both on objectivity and subjectivity.

Since it was of interest to identify interesting and frequent rules among a larger universe, the support was left at a threshold of 0.001, delivering as an output a total of 686981 rules, of which not redundant 37442 rules (figure 44). By plotting the rules, it can be observed how most of the rules are grouped at a lower level of support, while confidence is not a good discriminator (since for lower support levels there are both rules with high confidence and low confidence). The graph shows as well how the highest values of lift are present at the low support. This being caused by the fact that rule that occur rarely, are more likely to appear more frequently together in the data than expected - hence under the assumption of conditional independence causing high lift values.

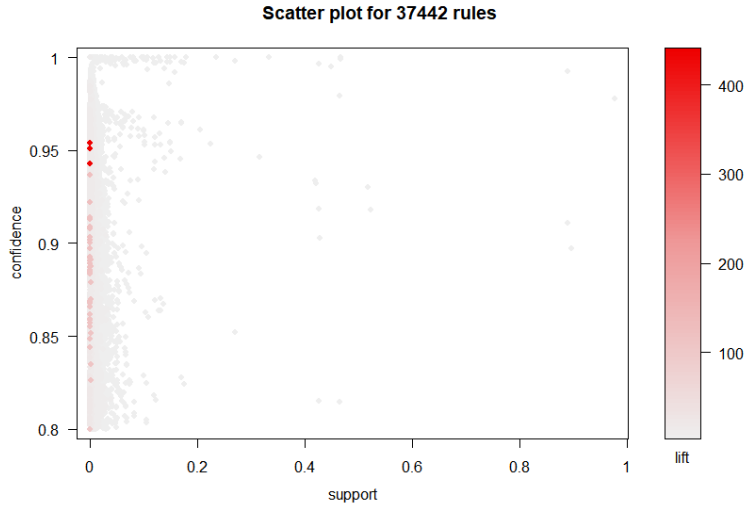

Figure 45. Association rules generated with minimum support threshold 0.001. Source R: software; Project Genie.

Based on the above considerations, a double strategy was used for the identification of interesting rules:

1. A first selection of rules based on their likelihood to appear together and being positively interdependent, using for this purpose the "Lift parameter". The additional interestingness measures were used for the interpretation of the rules (see appendix 5). It was decided to proceed with the first 20 rules (figure 45), leaving space for further exploration in future analysis. The generated rules had a max number of 4 items. The following code being used.

```
sorted.arules.001 <- clean.arules.001 %>% sort(by='Lift')

inspect(sorted.arules.001[1:20])
```

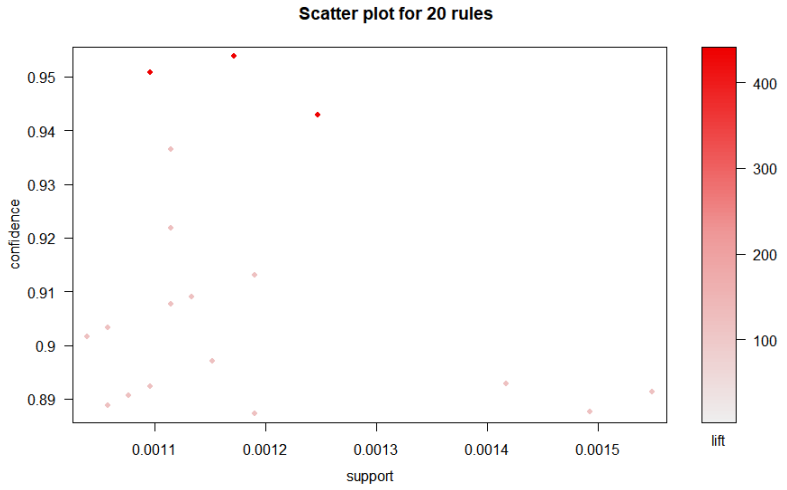

Figure 46. First 20 association rules based on Lift. Source R: software; Project Genie.

2. A second selection of rules was done in order to identify the most frequent patterns occurring in the dataset. Therefore, the first 20 rules with the highest support (figure 46) were selected and consequently evaluated based on additional measures such as their Lift, ChiSquare values, Cosine and Confidence (see appendix 6). The following code being used:

```
sorted.arules.001 <- clean.arules.001 %>% sort(by='support')

inspect(sorted.arules.001[1:20])
```

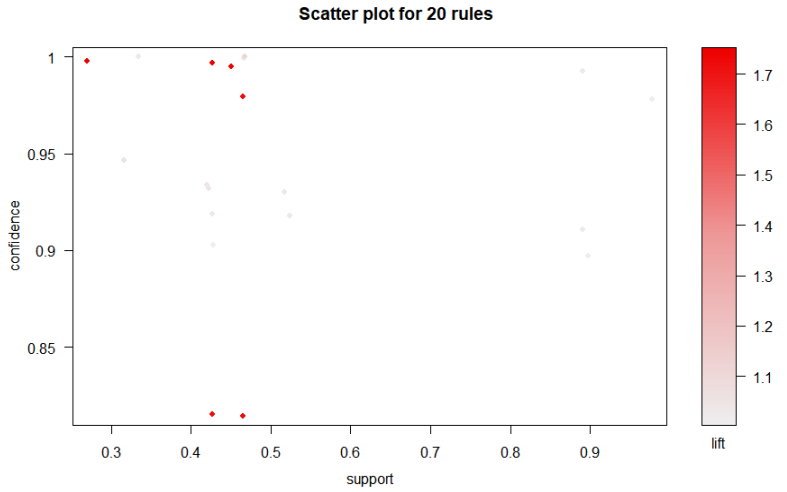

Figure 47. First 20 association rules based on Support. Source R: software; Project Genie.

Since Lift, Support and Confidence might be influence by large number of rules and imbalanced presence of items, the additional interestingness measures helped to interpret the rules. While there are rules with a high lift in the first selection, they occur rarely at the same time and measures such as Cosine and Kulczynski help to highlight that these rules might not be interesting, since they tend to be significantly lower after the 10th rule. Contrary, the second selection identified rules that result to be more interesting. Despite a lift value close to 1 or slightly above one, high values of Confidence, Cosine and Kulczynsky point to their interestingness.

Concluding, in addition to the picture of interestingness created by the objective evaluation parameters, the business acumen of the analyst was useful to make the final decision on rules[35] and identify their business values.

4.4.8. Limitations and Future steps

The frequent pattern mining problems with search for interesting rules in data represent an immense universe and would deserve a separate focus and thesis project. Because of this, the limitations of the research are represented mainly by the choices done in terms of tools to be used, algorithms to be applied, datasets and goals of the research. Despite Apriori represented a good solution for the problem at hand, with larger and continuously growing datasets it can become computationally expensive, hence there is space for other algorithms to be tested and implemented together with an exploration for other tools and broader project scopes.

In the future it could become interesting to explore in depth how applying different parameters influence the identification of patterns. Additionally, rules generated from use of lower minimum thresholds can be analysed more in depth for their interestingness to explore their business value. It can be interesting to apply the algorithm on real-time data in order to deliver updated insights on patterns, for more targeted managerial and strategic decisions. The insights can be both used to track the online behaviour of the users and to understand how to optimize their experience together with the resources invested in the solution. Other future applications of Apriori or similar algorithms, can be the enforcement of clustering or prediction algorithms or projects - helping to better group, understand and target the users and the markets, leading towards data-driven decisions.

[35] Due to the confidentiality of the rules it was decided not to publish them in this report. Nevertheless, requests can be evaluated under specific terms and conditions.

5. CONCLUSIONS

5.1. OVERALL EVALUATION OF THE INTERNSHIP

During the internship, the student had the possibility to work in a global company with a highly international environment and at the same time in a small and dedicated team - experiencing both an environment defined by rules, procedures and hierarchy and at the same time a very dynamic and flexible environment, ready to adapt and absorb quickly new skills and knowledge. Being part of such a team allowed the analyst intern to participate in a variety of different categories of projects, requiring the ability to destroy the knowledge silos and use tools and knowledge from different areas of competency in order to see the holistic picture of the knowledge and to achieve the projects' goal. Becoming part of the Market Intelligence Team of Oticon meant to embark on a steep learning curve, enriching the theoretical knowledge with practice. The diversity of the projects allowed for the application of a multitude of knowledge domains, ranging from statistics, to business intelligence to data mining - picturing how these should not be viewed individually, but should be combined to leverage the data in order to create knowledge and value.

The analyst intern had the opportunity to enrich his academic knowledge baggage by acquiring new competencies and tools - for example SPSS for statistical data analysis; Qualtrics for programming online questionnaires; knowledge in designing and implementing quantitative questionnaires; analysis categorical data through categorization.

Furthermore, additionally to the hard skills (practical knowledge), each project contributed to the creation of important soft skills: organizational skills and project management, capability of requirement definition and project scoping, story-telling for data analysis, business oriented and customer centric vision, stakeholder management, problem solving, attitudinal approach towards encountered difficulties and problems, and team spirit.

5.2. LESSONS LEARNED

The continuous ambition of growth generates continuously new products, new solutions (either digital or not), and new data. With a fast-paced business and technological development, in a global company it becomes hard to keep track of all the knowledge that is either generated or acquired - hence retain the knowledge and manage it. Indeed, one of the biggest obstacles to overcome during some major projects, was the internal knowledge retrieval regarding the developed solutions. Having to retrieve all the puzzle pieces of metadata and knowledge necessary to understand a data sample or business problem, can become a frustrating and time-consuming process, leading to a planning which becomes approximate and uncertain results (considering that bad data deliver bad results). Being part of an analytical team allowed to understand from a practical perspective how data by itself is not enough to create relevant insights, it should be representative, correct and meaningful. Considering the latter aspects, customer and product related data become a key asset and a powerful weapon in the battle for the competitive advantage and future market leadership.

Appendix

Appendix 1

```
install.packages('tidyverse')
library(tidyverse)
library(dplyr)
library(readxl)
library(magrittr)

#Importing & merging the files
ca <- read_excel("C:/Users/[redacted]/Desktop/Genie data/[redacted]/Genie data/[redacted]
[redacted]")

de <- read_excel("C:/Users/[redacted]/Desktop/Genie data/[redacted]/Genie
data/[redacted]")

#merging the two different datasets
genie <- rbind(ca, de)

#creating a copy of the datasets
genie2 <- genie

#exploring the dataset
str(genie2)
names(genie2)
```

Appendix 2

```
#Declaring genie2 variables as factors for the data mining algorithm
genie2$Market <- factor(genie2$Market)
```

Appendix 3

```
#removing the columns that will not be part of the algorithm
genie3$Market <- NULL
genie3$TID <- NULL
genie3$FittingAudiometric[redacted] <- NULL
genie3$FittingAudiometric[redacted] <- NULL
genie3$InstrumentProgram[redacted] <- NULL
```

Appendix 4

Below, the summary statistics of Apriori not redundant rules with a min support threshold of 0.001

```
> summary(clean.arules.001)
set of 37442 rules

rule length distribution (lhs + rhs):sizes
    1     2     3     4     5     6     7     8     9    10
    2    69   412  2105  6601 11002 10002  5289  1657   303

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   6.000   6.000   6.385   7.000  10.000

summary of quality measures:
    support           confidence          lift             count          kulczynski        cosine          hyperLift        chiSquared
 Min.   :0.001001   Min.   :0.8000   Min.   :  1.000   Min.   :   53.0   Min.   :0.4009   Min.   :0.03408   Min.   : 0.945   Min.   :    0.07
 1st Qu.:0.001681   1st Qu.:0.8500   1st Qu.:  2.922   1st Qu.:   89.0   1st Qu.:0.4374   1st Qu.:0.09066   1st Qu.: 2.429   1st Qu.:  295.68
 Median :0.002815   Median :0.8907   Median :  4.680   Median :  149.0   Median :0.4565   Median :0.11701   Median : 3.259   Median :  543.35
 Mean   :0.004772   Mean   :0.8910   Mean   :  5.431   Mean   :  252.6   Mean   :0.4566   Mean   :0.12786   Mean   : 3.590   Mean   :  766.73
 3rd Qu.:0.004534   3rd Qu.:0.9310   3rd Qu.:  7.396   3rd Qu.:  240.0   3rd Qu.:0.4750   3rd Qu.:0.15287   3rd Qu.: 4.833   3rd Qu.:  959.29
 Max.   :0.977899   Max.   :1.0000   Max.   :439.084   Max.   :51768.0   Max.   :0.9889   Max.   :0.98889   Max.   :62.000   Max.   :32614.86
                                                                                                                               NA's   :2

mining info:
 data ntransactions support confidence
   tr          52938   0.001        0.8
```

Below, the summary statistics of ECLAT rules with a min support threshold of 0.001

```
> summary(sorted.eclat.001)
set of 154242 itemsets

most frequent items:
                  BasicFitting=BasicFitting
                                      73627
FittingProgramManager=FittingProgramManager
                                      69809

element (itemset/transaction) length distribution:sizes
    2     3     4     5     6     7     8     9    10
  265  1455  5116 12661 23032 31570 33363 27920 18860

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   6.000   8.000   7.486   9.000  10.000

summary of quality measures:
    support             count
 Min.   :0.001001   Min.   :   53.0
 1st Qu.:0.001436   1st Qu.:   76.0
 Median :0.002191   Median :  116.0
 Mean   :0.003978   Mean   :  210.6
 3rd Qu.:0.003910   3rd Qu.:  207.0
 Max.   :0.890551   Max.   :47144.0

includes transaction ID lists: FALSE

mining info:
 data ntransactions support
   tr          52938   0.001
>
```

Appendix 5

Below the selection of the first 20 rules based on Lift

support	confidence	lift	count	kulczynski	cosine	hyperLift	chiSquared
0.001171181	0.9538462	439.08441	62	0.7464883	0.7171105	62.00000	27191.760
0.001095621	0.9508197	437.69123	58	0.7275837	0.6924910	58.00000	25354.455
0.001246741	0.9428571	434.02584	66	0.7583851	0.7356072	33.00000	28613.770
0.001643432	0.8787879	114.02273	87	0.5460116	0.4328840	29.00000	9840.847
0.001643432	0.8700000	112.88250	87	0.5416176	0.4307142	29.00000	9740.881
0.001624542	0.8514851	110.48020	86	0.5311347	0.4236505	28.66667	9420.515
0.001548982	0.8913043	115.64674	82	0.5461424	0.4232431	27.33333	9408.479
0.001530092	0.8350515	108.34794	81	0.5167905	0.4071637	27.00000	8697.780
0.001511202	0.8791209	114.06593	80	0.5375997	0.4151827	26.66667	9051.174
0.001492312	0.8876404	115.17135	79	0.5406340	0.4145739	26.33333	9025.843
0.001416752	0.8928571	115.84821	75	0.5383403	0.4051273	25.00000	8619.265
0.001378972	0.8588235	111.43235	73	0.5188725	0.3919975	24.33333	8064.217
0.001794552	0.8260870	107.18478	95	0.5294650	0.4385757	23.75000	10092.985
0.001284522	0.8000000	99.88302	68	0.4801887	0.3581925	22.66667	6721.265
0.001246741	0.8571429	111.21429	66	0.5094538	0.3723647	22.00000	7275.309
0.001227851	0.8441558	109.52922	65	0.5017348	0.3667228	21.66667	7054.545
0.001190071	0.9130435	118.46739	63	0.5337276	0.3754792	21.00000	7404.623
0.001190071	0.8873239	115.13028	63	0.5208679	0.3701530	21.00000	7192.763
0.001171181	0.8857143	114.92143	62	0.5188375	0.3668703	20.66667	7065.393
0.001171181	0.8857143	114.92143	62	0.5188375	0.3668703	20.66667	7065.393

Appendix 6

Below the selection of the first 20 rules based on Support

support	confidence	lift	count	kulczynski	cosine	hyperLift	chiSquared
0.9778987	0.9778987	1.000000	51768	0.9889493	0.9888876	1.000000	NA
0.8971249	0.8971249	1.000000	47492	0.9485625	0.9471668	1.000000	NA
0.8905512	0.9926725	1.015108	47144	0.9516754	0.9507920	1.014569	4662.09579
0.8905512	0.9106784	1.015108	47144	0.9516754	0.9507920	1.014569	4662.09579
0.5238959	0.9177973	1.023043	27734	0.7508847	0.7320983	1.020007	326.01960
0.5173410	0.9300122	1.036659	27387	0.7533389	0.7323292	1.033511	777.73807
0.4671880	1.0000000	1.114672	24732	0.7603807	0.7216380	1.110602	5322.82813
0.4666969	0.9989487	1.021526	24706	0.7380967	0.6904657	1.019856	951.65516
0.4649401	0.9795439	1.716033	24613	0.8970292	0.8932260	1.700380	32614.85554
0.4649401	0.8145145	1.716033	24613	0.8970292	0.8932260	1.700380	32614.85554
0.4503948	0.9950338	1.743170	23843	0.8920334	0.8860669	1.726503	32157.73430
0.4285013	0.9027739	1.006297	22684	0.6902061	0.6566578	1.002696	16.53739
0.4271223	0.9186609	1.024005	22611	0.6973811	0.6613438	1.020260	231.16733
0.4271223	0.9967819	1.746232	22611	0.8725222	0.8636288	1.728670	29397.43917
0.4271223	0.8152809	1.717648	22611	0.8575748	0.8565312	1.700459	27105.50290
0.4219275	0.9321426	1.039033	22336	0.7012267	0.6621153	1.035128	581.65087
0.4205675	0.9337751	1.040853	22264	0.7012849	0.6616259	1.036934	631.39204
0.3343912	1.0000000	1.022601	17702	0.6709743	0.5847638	1.020406	601.07372
0.3164268	0.9462773	1.054789	16751	0.6494946	0.5777226	1.049759	696.19073
0.2699951	0.9979055	1.748200	14293	0.7354508	0.6870266	1.723502	14619.70942

Appendix 7

Research paper authored by Vlad Robu and co-authored by Vitor Duarte dos Santos, which was submitted to the 6th International Conference on Systems and Informatics (ICSAI 2019) - Lingang, Shanghai, China.

Mining frequent patterns in data using Apriori and Eclat

A comparison of the algorithm performance and association rule generation

Vlad Robu
NOVA Information Management School
NOVA University of Lisbon
Lisbon, Portugal

Vitor Duarte dos Santos
NOVA Information Management School
NOVA University of Lisbon
Lisbon, Portugal

Abstract—**This paper aims to compare Apriori and Eclat algorithms for association rules mining by applying the them on a real-world dataset. In addition to considering performance efficiency of the algorithms, the research takes into consideration the distribution of the support, as well as the number of rules generated by Apriori and Eclat.**

Apriori; ECLAT; Association rules; Algorithm comparison

I. Introduction

In a world where the data is continuously generated by a multitude of sources and in extremely large quantities, it becomes crucial to transform it into valuable and actionable knowledge. The process of transforming raw data into knowledge becomes especially important for large companies that strive to understand as much as possible the behavior of their customers and build a strong competitive advantage by offering a better targeted and a more personalized experience. This is often transformed into a problem of finding the most frequent patterns in large datasets and create a generalizable and interpretable picture of reality. A branch of this problem area is represented by those problems that imply transactional datasets and require the identification of the most frequent item-sets, the so-called Market Basket Analysis type problems or in other words the Association Rule Learning problems. The latter type of problems aims to identify which items tend to occur together most frequently in the transactions of a dataset. It strives to find the most frequent relationship between the respective items. There are several algorithms designed to serve the purpose of the pattern mining nevertheless, this paper will focus only on the two most known in the industry algorithms: Apriori and Elcat. In order to understand which one is better to be applied in a certain situation, many past researches focus on comparing Apriori and Eclat efficiency, interpreted as the time needed to mine the rules; this being done by making the algorithms perform under certain controlled circumstances and often being applied simulated data [1][2][3][4]. Inspired by past researches, this paper adds new dimensions to the comparison, considering additional elements related to association rule generation. Since the two algorithms under analysis were run on a real-world dataset, the comparison is done with a more business driven approach, analyzing other factors that might result important in a business environment, such as the quantity of generated rules, interestingness of the rules, drawbacks of the algorithms to be considered, and finally the time efficiency needed to run the two.

II. Frequent Pattern Mining Algorithms

Association rules mining is commonly used for market basket analysis problems with the goal to identify combination of items more likely to appear together in the transaction of a dataset. It helps to identify interesting rules, frequent patterns, correlations or just casual data structures in transactional datasets [4]. An association rule can be seen as an expression of type $X \Rightarrow Y$, where X and Y are sets of items. In a real-world scenario such a rule can be that, for example, customers that have purchased X also purchased Y. The interestingness of such a rule is represented by its *support* and *confidence*. Where *support* of a rule $X \Rightarrow Y$ is the percentage of transaction that contain both X and Y. While the *confidence* of the rule represents the percentage of transactions of X that also contain Y [5].

A. Apriori algorithm

Apriori is an association rule mining algorithm proposed in 1994 by Srikant and Agrawal for Boolean association rules [6]. It uses a level-wise algorithm, using bottom-up research and moving upward level-wise in the lattice [7]. In order to identify the association rules, the algorithm runs through two phases: *Candidate Generation* and *Pruning*. In the *Candidate Generation* phase, the Apriori algorithm uses a prior knowledge of frequent itemset properties and k item-sets are used to explore k+1 item-sets. Intuitively, if an item-set X has minimum support, so do all subsets of X. After generating all the k+1 candidates, a new scan of the transaction is done and the support these new candidates is determined. Afterwards, during the *pruning* step, the k-1 which are found not reaching the minimum support threshold are eliminated [8][9].

B. ECLAT algorithm

Eclat or Equivalence Class Transformation Algorithm is simple algorithms that mines efficiently frequent patterns by

Identify applicable sponsor/s here. If no sponsors, delete this text box. (sponsors)

performing a bottom like depth first search or in other words, a bottom up Lattice traversal. In order to mine frequent patterns, it requires to be applied on a vertical database [10]. For this methodology, all the transactions that contain a certain itemset are grouped into the same record. After intersecting the frequent k-itemsets, the frequent k+1 itemsets are generated. This process occurs until no more frequent itemsets can be found. What makes Eclat advantageous is that it does not need to scan the database multiple times in order to identify the k+1 itemsets. Indeed, after scanning the database once, the k+1 itemsets are discovered by intersecting the k-itemsets with one another [11]. The support for each transaction is calculated, if it is equal or greater than the minimum support threshold set by the researcher, then it is considered for the analysis otherwise it gets discarded. One of the main advantages of Eclat is that it reduces the access time while a disadvantage can be represented by the fact that it does not consider *confidence* as an interestingness measure but only the *support* of the rules [12].

III. METHODOLOGY

This research emerged from a real-world business scenario, where the researchers had to identify the most frequent patterns in a transactional dataset generated by a software logging the choices of the users. After cleansing the data and structuring it into a transactional binary dataset, where the presence of a certain item was identified by a binary variable, a dataset of 52.938 transactions and 26 emerged.

Apriori and Eclat were run using the R opensource software. For this purpose, the "arules" and "arulesViz" software packages had to be installed. After declaring the variables as *factors* (1) and the dataset type as transactional (2), Apriori and Eclat were applied by using the formulas (3) and (4), as showed in the example.

factor() (1)

as(x, transactions) (2)

apriori() (3)

eclat() (4)

In order to check the performance of the two algorithms, and identify both often and rarely occurring rules, the two algorithms were run with different minimum support thresholds. Therefore, the minimum support threshold used for this research was set at the following levels: 0.01%, 0.1%, 0.5%, 1%, 5%, 10%, 20%, 30%, 50%. Setting very low minimum thresholds, would have allowed both to stress the algorithms and allow for the possibility to find both rare but interesting rules. The minimum confidence threshold for the Apriori algorithm was set at a level of 80%.

In conclusion, the performance of the algorithms was tracked and compared considering the time needed to generate the rules, the number of generated rules at each iteration and the distribution of the interestingness of the rules.

IV. COMPARATIVE ANALYSIS

The following Fig.1, Fig 2 and table 1, compare the performance of Apriori and Eclat under three aspects: time in seconds needed for the association rule generation, distribution of the support of the rules and the number of rules generated at different support thresholds.

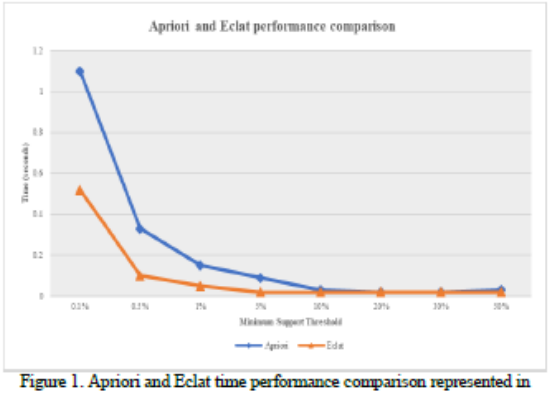

Figure 1. Apriori and Eclat time performance comparison represented in seconds

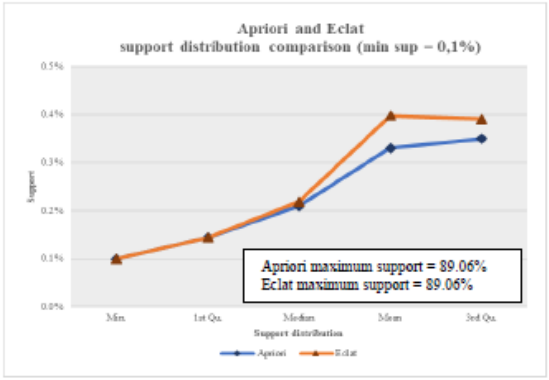

Figure 2. Apriori and Eclat rules' support distribution comparison

TABLE I: Apriori and Eclat comparison on the number of generated rules

Comparison of number of rules generated by Apriori and Eclat at different support thresholds		
Min support threshold	**Apriori**	**Eclat**
0.01%	7,465,565	1,443,983
0.1%	686,979	154,242
0.5%	90,278	26,735
1%	26,345	9,341
5%	1,211	633
10%	312	187
20%	62	40
30%	34	21
50%	6	4

For the sake of visualization, in Fig.2, the maximum support reached by the rules generated by the two algorithms, is written on the graph instead of being visualized.

In line with previous researches [1][2], Apriori shows to perform slower than Eclat, resulting in more time required to generate the association rules. The performance of Apriori is slower and requires more computational power compared to that of Eclat, especially when lowering the minimum support threshold at which the algorithms are applied. Considering that in a real-world scenario the researcher aims to find not only the most frequent but as well the most interesting patterns, even though rare, it might result crucial to operate with low support (and confidence) thresholds and with larger datasets. In the latter situation, one may opt to avoid the use of Apriori in favor of faster association rule mining algorithms. A lower time required by Eclat can be explained partially by the fact that at the same time it generates less rules than Apriori. For lower support thresholds, the gap between the output in terms of number of rules, of the two algorithms under comparison increases significantly. Meanwhile, the algorithms reach the same maximum support value for the generated association rules and the rules with the highest support tend to be in the 4rh quartile of data for both Apriori and Eclat. Nevertheless, Eclat tends to generate less rules than Apriori, they tend to have a slightly higher support, which means they can be potentially more interesting for the researcher.

V. Conclusions

In order to mine the most interesting and frequent patterns, the association rule mining algorithms come to be perfectly suitable for the task. By comparing Apriori and Eclat in this paper, it was possible to highlight how it becomes more computationally expensive for Apriori rather than for Eclat to mine frequent patterns, especially when rare but interesting patterns have to be mined at very low thresholds. Furthermore, Apriori tends to generate more rules than Eclat, which in a real-world scenario can only add complexity to the task of identifying interesting association rules. In addition to the analysis carried in this paper, other future researches can be done focusing on the difference between the quality of the generated rules in addition to the efficiency of the algorithms. To decrease the quantity of generated rules, it can be accounted for the redundancy generated by the algorithms [13]. Despite being less efficient than Eclat, in a real-world scenario Apriori has the advantage of considering other interestingness measure rather than *support*, such as *lift* and *confidence*, which can contribute to the identification of interesting and unique patterns in large datasets of rules overweighting this way its lower efficiency.

References

[1] Chee, CH., Jaafar, J., Aziz, I.A. et al. Artif Intell Rev (2018). https://doi.org/10.1007/s10462-018-9629-z

[2] Gayathri, G. (2017). Performance comparison of Apriori, Eclat and FP-Growth algorithm for association rules learning. *International Journal of Computer Science and Mobile Computing*, 81-89.

[3] Vani, K. (2015). Comparative Analysis of Association Rule Mining Algorithms Based on Performance Survey. *International Journal of Computer Science and Information Technologies*, 3980-3985.

[4] Garg, K., & Kumar, D. (2013). Comparing the Performance of Frequent Pattern Mining Algorithms. *International Journal of Comoputer Applications*, 29-32.

[5] Sethi, A., & Mahajan, P. (2012). Association Rule Mining: A review. The International Journal Of Computer Science And Applications.

[6] R.Agrawal and R.Srikant, "*Fast Algorithms for Mining Asscociation Rules*," In Proc. of VLDB '94, pp. 487-499, Santiago, Chile, Sept. 1994.

[7] Ghosh, S., Biswas, S., Sarkar, D., & Sarkar, P. (2012). 2012 Third International Conference on Emerging Appplications of Information Technology. International Conference on Emerging Applications of Information Technology, (pp. 202-205). Kolkata.

[8] Arora, J., Bhalla, N., & Rao, S. (2013). A review on association rule mining algorithms. *International Journal of Innovative Researc in Computer and Communication Engineering*, 1246-1251

[9] Han, J., Kamber, M., & Pei, J. (2012). Apriori Algorithm: Finding Frequent Itemsets by Confined Candidate Generation. In J. Han, M. Kamber, & J. Pei, *Data Mining Concepts and Techniques - Third Edition* (pp. 243-246). Amsterdam: Elsevier.

[10] Kaur, M., Garg, U., & Kaur, S. (2015). Advanced Eclat Algorithm for Frequent Itemsets Generation. *International Journal of Applied Engineering Research*, 23263-23279.

[11] Chee, CH., Jaafar, J., Aziz, I.A. et al. Artif Intell Rev (2018). https://doi.org/10.1007/s10462-018-9629-z

[12] Gayathri, G. (2017). Performance comparison of Apriori, Eclat and FP-Growth algorithm for association rules learning. *International Journal of Computer Science and Mobile Computing*, 81-89.

[13] Ashrafi M.Z., Taniar D., Smith K. (2004) A New Approach of Eliminating Redundant Association Rules. In: Galindo F., Takizawa M., Traunmüller R. (eds) Database and Expert Systems Applications. DEXA 2004. Lecture Notes in Computer Science, vol 3180. Springer, Berlin, Heidelberg

Bibliography

Agrawal, R., Imielinski, T., & Swami, A. (1993). Database Mining: A performance perspective. *IEEE Transactions of Knowledge and Data Engineering*, 914-925.

Al-Shatanawi, H. A., Osman, A., & Ab Halim, M. S. (2014). The Importance of Market Research in Implementing Marketing Programs. *International Journal of Academic Research in Economics and Management Sciences*, 150-159

Arora, J., Bhalla, N., & Rao, S. (2013). A review on Association Rule mining algorithms. *International Journal of Innovative Research in Computer and Communication Engineering*, 1246 - 1251.

Azavedo, A., & Santos, M. (2008). KDD, SEMMA and CRISP-DM: A parallel overview. *IADIS Conference on Data Mining*, (pp. 182 - 185). Amsterdam, The Netherlands.

Bathla, H., & Kathuria, K. (2015). Association Rule Mining: Algorithms Used. *International Journal of Computer Science and MObile Computing*, 271-277.

Brachman, R., & Anand, T. (1996). The process of knowledge discovery in databases. Advances in Knowledge Discovery. *AAAI*, 35 - 37.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0.* SPSS.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-Step Data Mining Guide.*

Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A., & Rong, X. (2015). Data Mining for the Internet of Things: Literature review and challenges. *International Journal of Distributed Sensor Networks*, 1-14.

Clifton, C. (2017, September 26). *Data Mining Computer Science*. Retrieved from Encyclopedia britannica: https://www.britannica.com/technology/data-mining

Creswell, J. (2014). *Research Design: Qualitative, Quantitative and Mixed Methods Approaches - 4th edition.* California: SAGE Publications, Inc.

Crockford, D. (2019, March 02). *Introducing JSON*. Retrieved from JSON org: https://www.json.org/

Emory, C., Williams, *Business Research Methods,* p. 264 - 265

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to Knowldge Discovery in Databases. *AAAI*, 37 - 54.

Folmer, E., & Bosch, J. (2004). Architecting for usability: a survey. *The Journal of Systems and Software*, 61 - 78.

Friedman, J. H. (1997). Data Mining and Statistics: What's the Connection? *Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics*, (pp. 1-6). Huston, United States.

Gayathri, G. (2017). Performance comparison of Apriori, Eclat and FP-Growth algorithm for association rules learning. *International Journal of Computer Science and Mobile Computing*, 81-89.

Garcia, E., Romero, C., Ventura, S., & Calders, T. (2008). Drawbacks and solutions of applying association rule mining in learning managemetn systems. *International Workshop on Applying Data Mining in e-Learning* (pp. 1-10). Aachen: CEUR Workshop Proceedings.

Gartner. (2019). *Magic Quadrant for Analytics and Business Intelligence Platforms.* Gartner.

Ghahramani, Z. (2003). Unsupervised Learning. In O. Bousquet, U. Von Luxburg, & G. Ratsch, *Advanced Lectures on Machine Learning* (pp. 72-112). Berlin: Springer.

Ghosh, S., Biswas, S., Sarkar, D., & Sarkar, P. (2012). Association Rule Mining Algorithms and Genetic Algorithms: A Comparative Study. *Third International Conference on Emerging Applications of Information Technology*, (pp. 202-205). Kolkata, India.

Gioia, C., Ben-Akiva, M., Jorgensen, O., Kasper, J., & Don, S. (2015). Case Factors Affecting Hearing Aid Recommendations by Hearing Care Professionals. *Journal of the American Academy of Audiology*, 229-246.

Gowthami, K., & Pavan Kumar, M. (2017). Study on Business Intelligence Tools for Enterprise Dashboard Development. *International Research Journal of Engineering and Technology*, 29877 - 2992.

Grover, L., & Mehra, R. (2008). The Lure of Statistics in Data Mining. *Journal of Statistics Education*.

Han, J., Cai, Y., & Cercone, N. (1992). Knowleedge Discovery in Databases: An Attribute-Oriented Approach. *VLDB*, 547 - 559.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques.* Waltham: Elsevier.

Hand, D. (1998). Data Mining: Statistics and More? *The American Statistician*, 112-118.

Haugaard, S., Egger, C., & Abrams, H. (2011). Hearing Aids Improves Hearing - and a lot more. *The Hearing Review*, May.

Jain, A. (2016, September 17). *The 5 Vs of Big Data*. Retrieved from IBM: https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/

Jaiswal, V., & Agarwal, J. (2012). The Evolution of the Association Rules. *International Journal of Modeling and Optimization*, 726-729.

Jenstad, L., & Moon, J. (2011). Systematic Review of Barriers and Facilitators to Hearing Aid Uptake in Older Adults. *Audiology Research*, March.

Jenstad, L., Oberg, M., Nielsen, C., Naylor, G., & Kramer, S. (2010). Factors Influencing Help Seeking, Hearing Aid Uptake, Hearing Aid Use and Satisfaction With Hearing Aids: A review of the Literature. *Trends in amplification*, 91-96.

Kaur, M., & Grag, U. (2014). ECLAT Algorithm for Frequent Itemsets Generation. *International Journal of Computer Systems*, 82-84.

Kothari, C. R. (2004). Research Methodology: An Introduction. In C. R. Kothari, *Research Methodology: Methods & Techniques* (pp. 1-21). New Delhi: New Age International Ltd.

Kothari, C. R. (2004). Measurement and Scaling Techniques. In C. R. Kothari, *Research Methodology, Methods and Techniques* (pp. 69-82). New Delhi: New Age International Ltd.

Kothari, C. R. (2004). Methods of Data Collection. In C. R. Kothari, *Research Methodology, Methods and Techniques* (pp. 95-120). New Delhi: New Age International Ltd.

Kothari, C. R. (2004). Research Methodology: An Introduction. In C. R. Kothari, *Research Methodology: Methods & Techniques* (pp. 1-21). New Delhi: New Age International Ltd.

Kothari, C. R. (2004). Sampling Design. In C. R. Kothari, *Research Methodology, Methods and Techniques* (pp. 55-67). New Delhi: New Age International Ltd.

Kotsiantis, S. (2007). Supervised Machine Learning: A Review of Classification Techniques. In I. Maglogiannis, K. Karpouzis, B. Wallace, & J. Soldatos, *Frontiers in Artificial Intelligence and Applications* (pp. 3-15). Amsterdam: IOS Press.

Kotsiantis, S., & Kanellopoulos, D. (2006). Association Rules Mining: A recent overview. *International Transactions on Computer Science and Engineering*, 71-82.

Kumar, R. (2014). *Research Methodology, a step-by-step guide for beginners - 3rd edition.* Los Angeles: Sage.

Kumbhare, T., & Chobe, S. (2014). An overview of Association Rule Mining Algorithms. *International Journal of Computer Science and Information Technologies*, 927-930.

Kuonen, D. (2004). Data Mining and Statistics: What is the Connection? *The Data Administration Newsletter*, 1-6.

Maninmaran, J., & Velmurugan, T. (2015). Analysing the equality of Association Rules by Computing an Interestingness Measures. *Indian Journal of Science and Technology*, 1-12.

Mariscal, G., Marbàn, O., & Fernàndez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The knowledge Engineering Review*, 137 - 166.

Matheus, C., Chan, P., & Piatetsky-Shapiro, G. (1993). Systems for Knowledge Disvoery in Databases. *IEEE Transactions on knowldge and data engineering*, 903 - 913.

McCusker, K., & Gunaydin, S. (2014). Research using qualitative, quantitative or mixed methods and choice based on the research. *Perfusioin*, 1-6.

Mooi, E., & Sarstedt, M. (2011). *A concise guide to market research. The process, data, and methods using IBM SPSS Statistics.* Berlin: Springer.

Nikov, A., Vassilieva, S., Anguelova, S., Stoeva, S., & Tzvetanova, S. (2003). Webuse: An approach for web usability evaluation. *3rd Symposium on Production Research*, (pp. 511 - 519). Istanbul.

Omiecinski, E. (2003). Alternative Interest Measures for Mining Associations in Databases. *IEEE Transactions on Knowledge and Data Engineering*, 57-69.

OrbisResearch. (2017, June 20). *2017-2023 Hearing Aid Devices Market to Grow at a CAGR of 5,1% during the Forecasted Period*. Retrieved from Reuters: https://www.reuters.com/brandfeatures/venture-capital/article?id=11672

Oticon. (2018). *Oticon Hearing Aid Solutions*. Retrieved from Oticon: https://www.oticon.com/

Patent Justia. (2018). *Patents assigned to Oticon A/S*. Retrieved from pATENTS jUSTIA: https://patents.justia.com/assignee/oticon-a-s

Piatetsky-Shapiro, G. (1991). Knowledge Discovery in Real Databases: A report on the IJCAI-89 Workshop. *Al Magazine*, 68 - 70.

Powell, B. (2017). *Microsoft Power BI Cookbook.* Birmingham, UK: Packt Publishing Ltd.

SAS Institute. (2014, October 1). *Data Mining Using SAS Enterprise Mining: A case study approach, Third Edition*. Retrieved from SAS: http://support.sas.com/documentation/cdl/en/emcs/66392/HTML/default/viewer.htm#n0pejm83csbja4n1xueveo2uoujy.htm

Sathya, R., & Abraham, A. (2013). Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. *International Journal of Advanced Research in Artificial Intelligence*, 34-38.

Sethi, A., & Mahajan, P. (2012). Association Rule Mining: A Review. *The International Journal of Computer Science & Applications*, 72-83.

Simon, A., Deo, S. D., Venkatesan, S., & Ramesh, B. (2015). An Overview of Machine Learning and its Applications. *International Journal of Electrical Sciences and Engineering*, 22-24.

Skiba, D. (2017). Evaluation Tools to Appraise Social Media and Mobile Applications. *Informatics*, 32-40.

Sridhar, M. (2015). Model Driven Software Engineering in the Mobile Era with an Emphasis on Security. *International Journal of Electronics Communication and Computer Engineering*, 619-623.

The R Foundation. (2019, April 22). *What is R? Introduction to R*. Retrieved from R-project: https://www.r-project.org/about.html

U.S. Government Printing Office. (1978). *Effect of Smoking on Nonsmokers. Hearing before the Subcommittee on Tobacco of the Committee on Agriculture House of Representatives. Ninety-Fifth Congress.* Washington D.C.: U.S. Government Printing Office.

Verschuren, P., & Doorewaard, H. (2010). *Designing a Research Project.* The Hague: Eleven International Publishing.

Webb, J. (2003). Marketing Research. In M. J. Baker, *The Marketing Book, 5th edition* (pp. 171-197). Oxford: Butterworth Heinemann.

William Demant. (2016). *William Demant*. Retrieved from Demant: https://www.demant.com/

World Health Organization. (2018, March 15). *Deafness and hearing loss*. Retrieved from World Health Organization: http://www.who.int/en/news-room/fact-sheets/detail/deafness-and-hearing-loss