NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Stream Smarter, Not Harder

A Data-Driven Solution based on Recommendation Systems for Streaming Platforms through Power BI.

Adriana Gamboa Campos Calheiros de Brito

Project Work

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

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

Stream Smarter, Not Harder

A Data-Driven Solution based on Recommendation Systems for Streaming Platforms through Power BI.

by

Adriana Gamboa Campos Calheiros de Brito

Project Work presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics.

Supervised by

Professor Vitor Duarte dos Santos, Integrated Researcher at Information Management Research Center (MagIC), NOVA Information Management School.

November, 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 1 November 2023

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my dear family, Anabela, and Daniel, who have always supported me in every challenging decision. Both gave me the confidence and encouragement to achieve all my goals and dreams. I would like to extend my sincere appreciation to Alan, who has been relentless in encouraging and motivating me during every stage of this project.

Finally, I am also very thankful to my supervisor, Professor Vitor Duarte dos Santos, for all the knowledge, expertise, guidance, availability, and patience he shared with me during this period.

ABSTRACT

Streaming video platforms have gained remarkable popularity in recent years, offering users affordable access to a vast variety of entertainment content. Platforms such as Netflix and Spotify have fundamentally transformed the consumption of digital content, from films and TV programmes to music and podcasts, allowing instant access without the need for lengthy downloads, often through a simple subscription fee or freemium model. This revolution has reshaped the entertainment industry, establishing a dynamic global distribution channel that quickly connected creators to a massive audience. Nevertheless, the abundance of competitors and a large volume of content competing for viewers' attention has forced industry leaders to re-evaluate their content recommendation strategies. To battle for their market share in this competitive environment, these industry giants must leverage the power of data-driven insights through Artificial Intelligence and Machine Learning in order to select appealing content recommendations and ultimately foster deeper engagement. Therefore, the goal of this project was to create a complete Business Intelligence solution using the data of four streaming platforms and deliver an effective content-based recommendation system to boost audience level. Therefore, aimed of achieving this objective, a recommendation algorithm was applied to the different datasets and eight dashboards were created to display powerful insights of the four most popular streaming platforms worldwide.

KEYWORDS

Business Intelligence; Machine Learning; Recommendation Systems; Data Visualization; Word2vec; Natural Language Processing

Sustainable Development Goals (SDG):

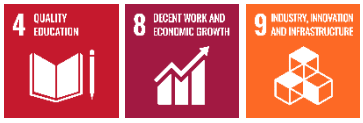

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

3NF Third Normal Form

BI Business Intelligence

BIA Business Intelligence and Analytics

CBF Content-Based Filtering

CBOW Continuous Bag of Words

CF Collaborative Filtering

CSV Comma-Separated Values

CTR Click Through Rate

DAX Data Analysis Expressions

DV Data Visualization

DW Data Warehouse

ETL Extract Transform Load

FK Foreign Key

HS Hybrid System

KPI Key Performance Indicator

ML Machine Learning

NLP Natural Language Processing

NLTK Natural Language Toolkit

NPS Net Promoter Score

OKR Objectives and Key Results

OLAP Online Analytical Processing

OLTP Online Transaction Processing

PK Primary Key

TV Television

RS Recommendation Systems

SG Skip Gram

SSBI Self-Service Business Intelligence

VRR Visit after Recommendation Rate

1. Introduction

1.1. CONTEXT

Streaming video platforms have become increasingly popular in recent years, as they offer convenient and affordable entertainment video content to their users. Over the last few years, online streaming applications have increasingly grown to thousands of subscribers, in fact, Disney+ achieved more than 160 million subscribers and Netflix reached 231 million users worldwide in the last quarter of 2022 (Statista, 2023).

Video streaming services allowed users to access and watch digital content, such as movies, TV shows, music, and podcasts, without waiting for the full download in exchange of an affordable monthly subscription fee or a freemium account. These applications have transformed the entertainment industry by providing a new distribution channel for leisure content and for reaching a global audience in a very short amount of time.

The considerable number of players along with the vast information available competing for users' attention led the market leaders to reinvent the way they recommend their content. To accomplish that, market leaders must fully understand their own data and insights in order to make meaningful data-driven decisions, recommend successful items and promote their audience engagement.

1.2. MOTIVATION

The increased number of web-based services that use recommender systems created a significant impact on several industries in the last couple of years. The rise of new suggested products' advertisement and the revenue growth due to video advertising's monetisation due to recommended videos are only few examples of profitable recommendation outcomes. Recommender systems are an irreplaceable tool in daily web searches and sales worldwide.

In 2006, Netflix decided to further improve its recommendation system on the "What to watch next" content. In order to achieve this goal, Netflix announced a public competition, named "The Netflix Prize". The goal of this competition was to create an improved recommendation system, in which the accuracy score would be at least 10% better than Cinematch, a sophisticated recommendation system based on collaborative filtering algorithm. The platform offered a $1M prize to the winner.

Although Netflix never put into practice the work developed by the awarded team, the public competition had a significant impact on the field of data science and machine learning. Some of the reasons why Netflix did not use the algorithms created were the operational cost associated with the scale of the solution and a lawsuit against Netflix due to the use of personal users' data on this competition. Nevertheless, the creation of events such as this one

contributed to develop and boost ML algorithms capable of improving personalized predictions and effective recommender systems.

In addition to that, information has proved its value over the last decades. In fact, data represents nowadays a new class of economic asset, similar to currency or gold. However, with the large amount of data, information and knowledge management has become a real challenge for most of the organizations. Its importance has been especially recognised by the development of decision-making tools based on computer analysis and structured data architecture.

The motivation to elaborate this master thesis was based on the importance of analysing data frequently in order to support the decision-making process of managers and leaders. In this manner, a complete Business Intelligence solution was developed, which included a deep analysis of the platforms' data and the creation of a Recommendation System. This one was able to generate a recommended list of similar video content, based on content description and category. This BI solution shall contribute to create an economic impact in these organizations, but also benefit socially its subscribers and stakeholders.

On the users' behalf, this project contributed to provide customized content recommendations. This solution was able to suggest new video content based on the other video content's attributes. This way, the model will provide better options and alleviates the user with the burden of choosing from a large number of movies and tv shows. This process leads to a fast unconscious decision-making process and thus, to an increase of individual amusement and spontaneous contentment. Besides, with a successful recommendation system, the users won't feel the need to have multiple streaming accounts to watch the same type of content, which ultimately reduces financial stress on their households.

On the other side, companies in the streaming industry who use BI tools can benefit from a range of advantages, such as access to a better understanding of user behaviour insights and preferences, allowing managers to make data-driven decisions and, ultimately, improve user experience and retention time. Additionally, it can result in cost savings by identifying areas where operations can be optimized, and costs reduced. For example, reducing the investment in one movie category at the expense of a more economically attractive one. Lastly, leveraging BI capabilities can lead to a competitive advantage in this popular industry by providing an efficient and tailored-made user experience and interface, leading to higher user retention, increased number of users and larger revenues.

In short, a Business Intelligence solution such as the one elaborated in this project can become an essential component for the success of streaming platforms and its users, in order to enhance performance, market share, improve user-experience and drive value growth.

1.3. OBJECTIVES

The goal of this project was to create a complete BI solution for four of the most relevant streaming platforms worldwide and to suggest an effective content-based recommendation system applied to the US market.

In order to reach the research goal, the following intermediate objectives were defined:

- Perform a comprehensive literature review on the topic of recommendation systems applied to streaming platforms, including several filtering methods, areas of use and methods of evaluation.
- Study and implement the most efficient recommendation system algorithm using Python in a Jupyter Notebook.
- Develop a deep understanding and analysis regarding the main concepts of Business Intelligence, including data warehouses, architecture approaches, effective data visualization techniques such as KPIs, OKRs and data charts, and more.
- Integrate all sources and develop a complete BI solution using Power BI from Microsoft to enable data-driven decision making. One type of report displays the main data insights and the other provides information regarding the recommendation system. In total, two reports per streaming platform were created.

These four milestones provide a path to guide the fulfilment of the main goal.

1.4. EXPECTED OUTCOME

The creation of an efficient recommendation system combined with a complete Business Intelligence solution was expected. This solution was planned and computed in order to be extremely beneficial not only for the mentioned streaming platforms, but for any organization which is competing for their market share in this popular industry. As mentioned before, a holistic perspective of their data allows companies to make more accurate informed decisions, enhance organizational growth and improve performance. In Figure 1, a representative report of the general statistics of each platform was provided.

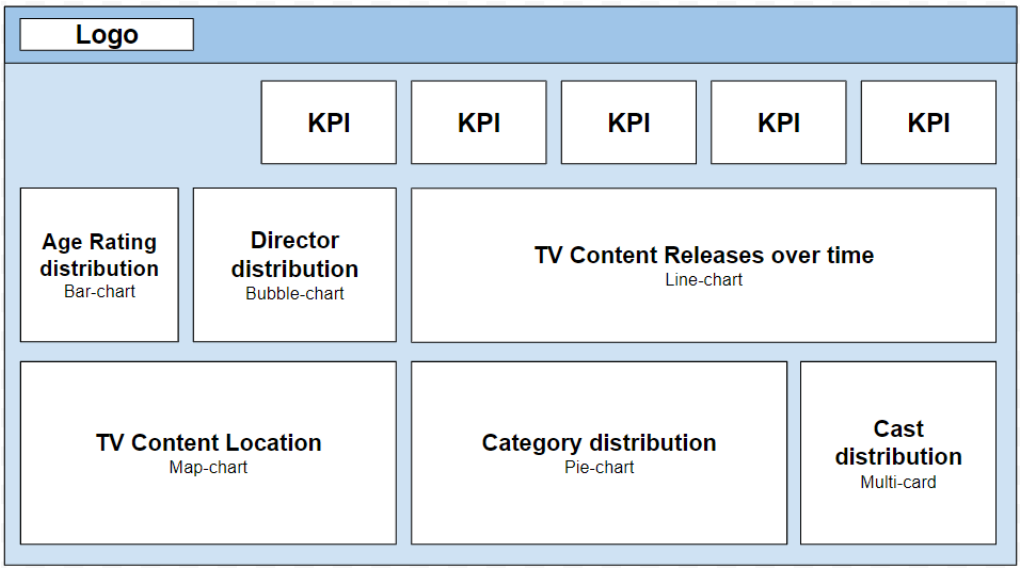

Figure 1 – Report representation of platforms' main dataset on Power BI.
Source: Illustration prepared by the author.

In addition, Figure 2 represented a suggestion of the second view of this BI solution, presenting the expected solution of the recommendations generated for each of the platforms.

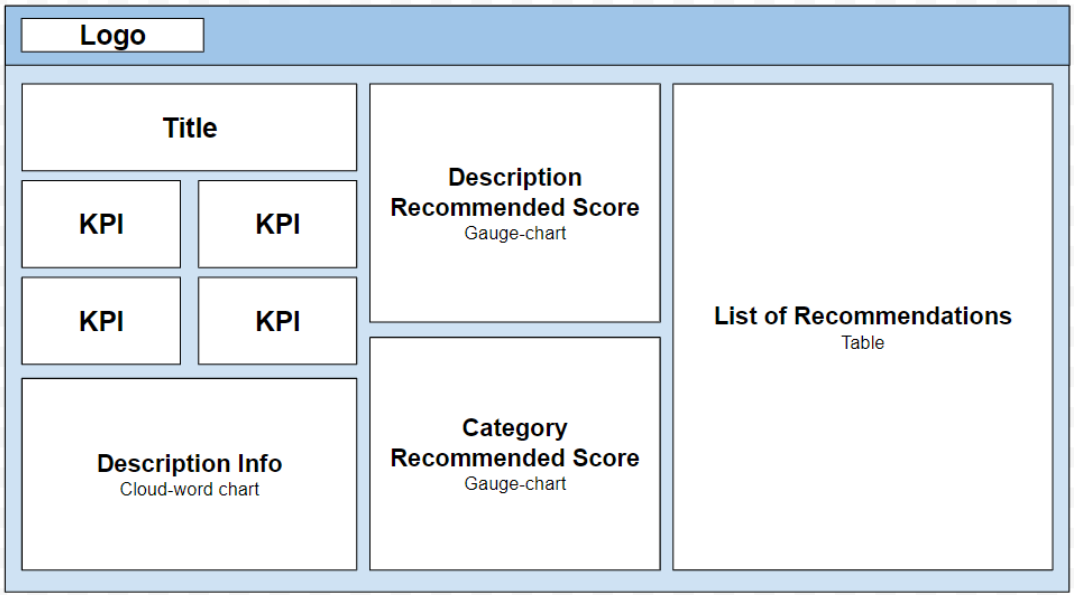

Figure 2 - Report representation of each platforms' recommendation on Power BI.
Source: Illustration prepared by the author.

2. Work Programme

2.1. PROJECT PHASES

The development of this project was divided in the following planned stages:

- **Identify, Analyse and Define the Necessity** – In the initial phase, the economic and social need by the streaming platforms for such a project was identified, followed by an extensive definition and analysis of such need.
- **Project Planning** – In the second stage, the necessary information to perform the work was gathered and scheduled in order to organize the execution of the project.
- **Theoretical Framework** – A complete study and research of the Recommendation System and Business Intelligence was performed to assess the best algorithms and techniques.
- **Project Development** - In this phase, the practical part of the project was developed, such as the implementation of the algorithm and the creation of the Business Intelligence solution in Microsoft Power BI.
- **Final Considerations** - Lastly, a complete description and analysis of the final result of this project was provided, along with future work suggestions and improvements in order to achieve the best possible result.

2.2. TOOLS

As mentioned before, the main goal of this project was to create a BI solution based on the data of the top four streaming platforms worldwide and suggest an effective content-based recommendation system. In order to accomplish this goal, a set of tools were required, such as the web-based interactive computing platform named Jupyter Notebook.

To display the output of the BI solution, Power BI tool developed by Microsoft was used to build the final dashboards applying DAX and M query languages. Lastly, this project also required other applications to be applied, such as Microsoft Excel and Google Workspace apps.

2.2.1. Python

The first part of this project was developed using the programming language Python to compute an efficient and accurate recommendation system algorithm. According to PYPL - PopularitY of Programming Language Index in 2022, Python is one of the most popular

computer languages in the world, with more than 8 million programmers (Cutting & Stephen, 2021).

This programming language holds a wide range of benefits, such as versatility, simplicity, ease of use, large developers-support community, and popular libraries. Python can be used for a wide range of tasks, from simple scripts to complex web applications and data analysis, according to official webpage of the programming language Python.org.

In detail, Python is a high-level, object-oriented, and structured interpreted programming language that was designed by Guido van Rossum and released in 1991. It is widely used for multiple industries and purposes such as Data Analysis, Artificial Intelligence, Machine Learning and much more. One of the main reasons for its popularity and adoption is due to its wide range of libraries and frameworks including NumPy, Pandas, Sklearn, TensorFlow, Seaborn and Matplotlib.

2.2.2. Microsoft Power BI

In the second part of this project, a complete BI solution was developed using the desktop application of Power BI, a Business Intelligence and Data Visualization tool developed by Microsoft.

According to its owner organization Microsoft, Power BI is a unified, scalable platform for self-service and enterprise Business Intelligence applications. It allows users to connect to a wide variety of data sources, including excel spreadsheets, csv files, databases, cloud-based and on-premises data sources, and perform dynamic visualizations, KPIs, OKRs, reports, and dashboards with any kind of data.

A major benefit of Power BI comparable with other applications is the fact that it offers a user-friendly interface and a powerful set of tools for data modelling, data preparation, and data analysis in its own programming language. It provides a few Machine Learning techniques such as data clustering and data prediction. Besides, Python and R queries are also possible to computed in this application.

2.2.3. M and DAX

In Microsoft Power BI, two programming languages are provided, which are M and DAX. As the opposite of many BI competitor tools, Power BI has its own Extract Transform Load tool named Power Query which allows to perform data engineering and preprocessing on the application directly. In this section of the platform, it is possible to explore and process complex sets of data using M query language.

After performing the necessary cleaning and preprocessing steps in the Power Query section applying M, it is also possible to create additional columns and measures in the Data Visualization section of the platform. In this section, Data Analysis Expressions programming language is available to support analysts developing KPIs, OKRs, reports, and charts (Towards Data Science, 2020).

2.3. CHRONOGRAM

In Figure 3, an execution plan chronogram with the timeline associated with each section of this project was provided. The schedule was divided in seven phases, four milestones (M) and two deliverables (D).

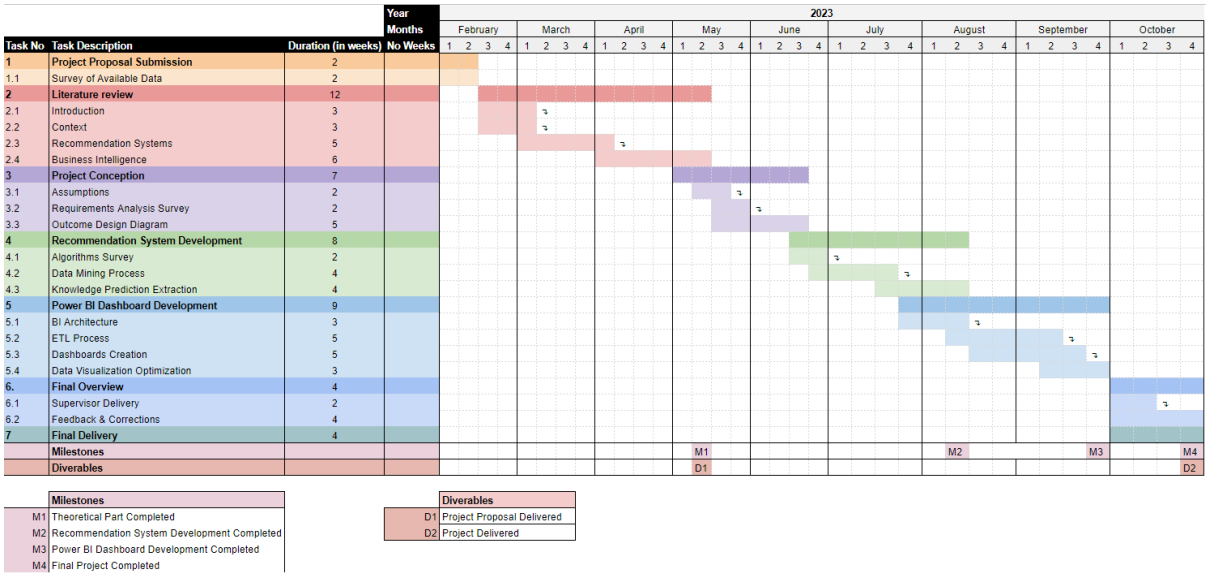

Figure 3 - Chronogram with the execution plan of the project.
Source: Illustration prepared by the author.

3. Theoretical Framework

3.1. OVERVIEW

To achieve the objectives mentioned previously and deliver the overall project goal, extensive research was performed, and knowledge collected from Machine Learning and Business Intelligence experts of the scientific community. The selection of the documentation for this project was carried out through the following electronic sources: ResearchGate, Elsevier and Google Scholar. These information sources were selected based on the number of citations and by full-text documents.

Therefore, in this section, a vast and deep research on Recommendation Systems was conducted respecting its concept, areas of use, and specifically how to apply it on streaming services in the entertainment industry.

Furthermore, an additional comprehensive study regarding the topic of Business Intelligence was executed, which included its concept, architecture, data warehouse models, advantages, and disadvantages of its implementation.

3.2. RECOMMENDATIONS SYSTEMS

3.2.1. Concept

Recommendation systems are an extremely supportive technique that mitigate the issue of overloading users with a vast amount of information from online services and providers. This technology predicts the grade of items to be recommended, creates a list of recommendation ranking for each item and makes it possible to recommend related items to the same user or to a similar one (Ko et al., 2022).

Since it is impossible to capture and track users' actions in precise real time (Paul & Kundu, 2019), the goal of the information filtering technique is to create an individualized list of items that match user's preferences (Ko et al., 2022).

3.2.2. Filtering Techniques

In accordance with the predict tailored-made list of recommendation items which satisfy users' taste and preferences based on its relationship with the product or service, a specific system to filter a large amount of information needs to be computed. In particular, it is possible to divide this particular knowledge-based filtering system into three major categories: Content-Based Filtering, Collaborative Filtering and finally, Hybrid System (Shahbazi & Byun, 2020; Reddy et al., 2018).

In Figure 4, the author aimed to display a general overview of the broad processes of a recommender system.

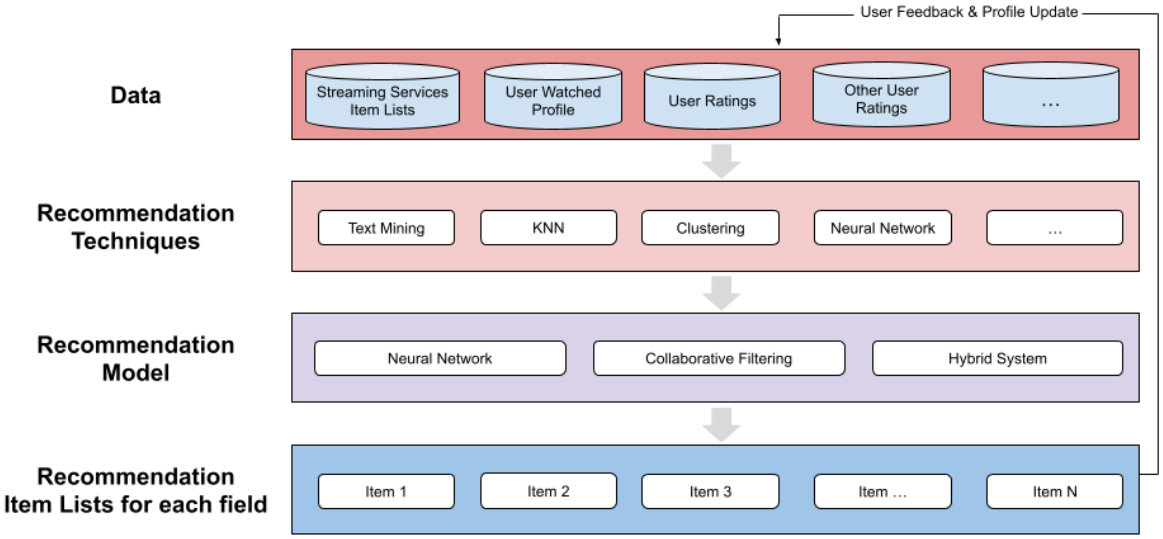

Figure 4 - Recommendation System flow process, including algorithms and techniques. Source: (Ko et al., 2022)

To have a better overview regarding the differences among the three types of recommendation system techniques, an elaborate description of the three was provided below.

3.2.2.1. Content-Based Filtering

Content-Based Filtering technique analyses the past behaviour of the user and according to the user's parameters and measures, identifies patterns and recommends other similar items (Reddy et al., 2018).

In this technique, the items are recommended based on the comparison done by the model between the features captured on the content of the item and the user profile (Paul et al., 2019).

The data available of each item is represented through a set of attributes or features. For example, if a user has rated high for a certain movie, other films with similar genres are recommended by this system (Reddy et al., 2018).

In this technique, the most popular algorithms to compute an accurate recommendation based on the similarity component are XGBoost, Random Forest, SVM (Shahbazi et al., 2020), K-means Clustering, and Expectation-Maximization with Monte Carlo sampling (Paul et al., 2019).

Researchers are continually developing new and innovative algorithms to enhance the performance and effectiveness of content-based models. These emerging algorithms aim to deliver more accurate and relevant outcomes, keeping up with the evolving needs of users and content. However, several challenges are associated with this approach such as the learning algorithm used in order to capture and learn the observed items' features in order to make accurate recommendations, the correctness of the item model, and finally it faces a glass-ceiling effect, which implies that this technique cannot discriminate major differences between similar goods (Paul et al., 2019).

3.2.2.2. Collaborative Filtering

On the contrary, the Collaborative Filtering method analyses the user's behaviour and predicts the similarity score between this particular user and other users in order to provide its recommendations (Shahbazi et al., 2020; Reddy et al., 2018). To successfully suggest the items, this technique uses a rating technique for its recommendation, which can be collected explicitly such as the NPS rating tools or implicitly extracted through user behaviour, such as the number of times a music was played by the same user (Paul et al., 2019).

In Collaborative Filtering, the widely used method for calculating a precise recommendation system is the K-Nearest Neighbour algorithm (Paul et al., 2019).

Additionally, this filtering method is based on the assumption that the users continue to rate the items similarly in the future as they did in the past, which means that users won't suffer a change of taste or preferences. Despite the fact of that, the major downside of this technique is its poor performance in the early stage of the recommendation known as the Cold-Start problem described later in detail (Paul et al., 2019; Wang et al., 2018).

3.2.2.3. Hybrid System

In this last filtering technique and as the name suggests, the Hybrid System is a combination of both Content-Based Filtering and Collaborative Filtering approach. By integrating the strengths of both systems, this model attempts to address the limitations and leverage the benefits of each approach (Reddy et al., 2018).

As mentioned earlier, both filtering models have limitations because the CBF model relies largely on metadata of the item, and CF depends mainly on the user item's classification data. In order to overcome the constraints of both filtering methods and to boost the performance of the recommendation, a Hybrid System has been developed (Ko et al., 2022).

This technique can be used to overcome several problems associated with recommendation systems, such as the narrowness and metadata dependency from CBF, and cold-start problem

and sparsity from CF. Resulting in a filtering system more robust and efficient (Pandya et al., 2016; Bhatt et al., 2014).

In Figure 5, an adapted example was added which represents an overview of the different recommendation filtering techniques.

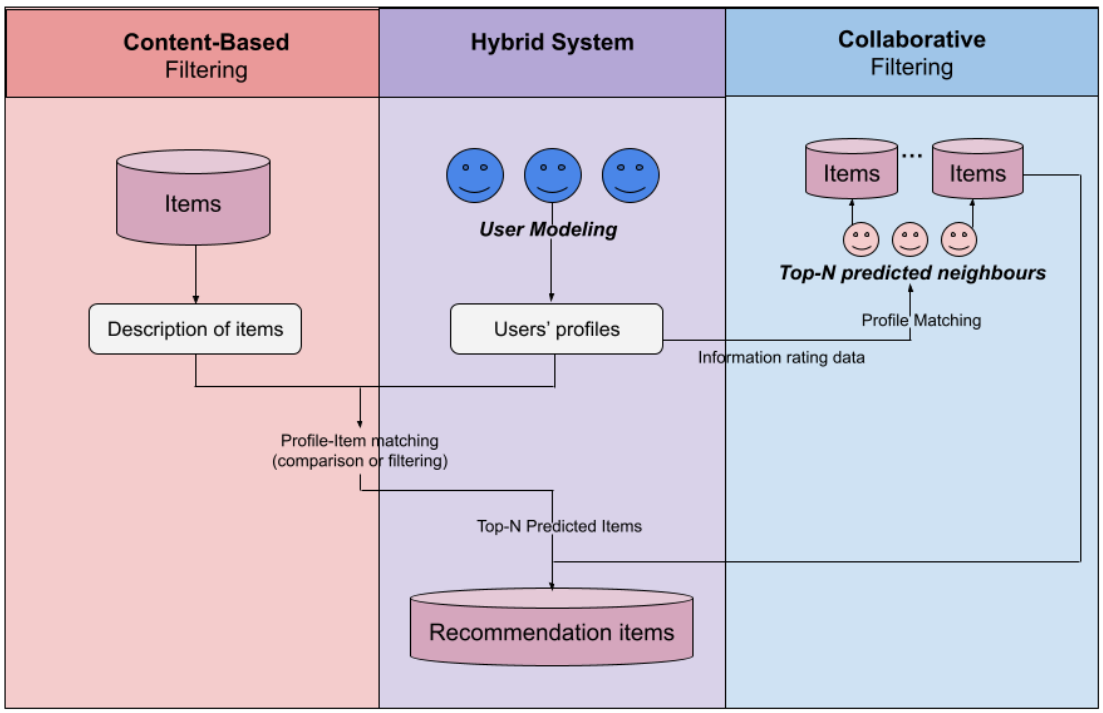

Figure 5 - Overall process of the three recommendation system models.
Source: (Ko et al., 2022)

3.2.3. Cosine Similarity

To extract the similarity value and quantify semantic similarities between two words, a well-known metric was used. Cosine similarity, vector similarity or cosine coefficient is a calculated formula to measure the similarity between two vectors in a multi-dimensional space. It computes the relationship or degree of similarity between two vectors, words or paragraphs as shown in Figure 6 (Bhatt et al., 2014; Zhao et al., 2020).

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Figure 6 - Mathematical formula of Cosine Similarity models.
Source: (Bhatt et al., 2014; Zhao et al., 2020)

Technically, this metric measures the cosine of the angle between two vectors. In fact, it calculates cosϴ between the two data points, ranging from -1 to 1 (Bhatt et al., 2014).

3.2.4. Areas of Use

Recommendation systems have developed under the main domain of Machine Learning and Artificial Intelligence. These filtering models are not a recent discovery, in fact, data scientists have been using them for the past 29 years in order to provide an accurate list of products according to customers' preferences (Ko et al., 2022).

There are several use-cases and attributes to boost engage using methods of recommendation, such as music, movies, social media, news, and advertising in general. Nowadays, major large corporations implement recommendation systems for fulfilling customer requirements and preferences. For example, LinkedIn recommends relevant connections, Amazon suggests interesting products, and Netflix recommends movies and series that the users might like to watch (Reddy et al., 2018).

In this project, a Recommendation System was developed and applied to the datasets of four different platforms. Therefore, the main focus of the research in the next section was to study in depth the usage of Recommendation System applied to video streaming platforms.

3.2.4.1. Online Movie Streaming Services

In the past, movies and TV shows were mostly consumed and watched by users through movie theatres or TV. However, in the early 2000's, that habit started to change with the entrance of streaming platforms in the market, such as Netflix and Amazon Prime (Ko et al., 2022).

To clarify, a streaming platform is an on-demand broadcast service which provides multiple entertainment content, such as TV shows, movies, series, online concerts, and other streaming media in real-time (McAuley, 2021).

In fact, companies such as Netflix led the innovative study of the data science field in this industry. By adapting, combining, and correcting available models, Netflix is constantly creating newer and more efficient algorithms in order to meet organization's needs and increase revenue. And this is why Netflix is using several ML models, such as Linear Regression, Logistic Regression, Elastic Nets, Singular Value Decompression, Gradient Boosted Decision Trees, Random Forest, Clustering Techniques, among many more to make each individual profile as personalized as possible (Fouladirad et al., 2015).

In this case, the recommendation system predicts which video content should be recommended to the user based on movie characteristics presented and experienced by the user. In order to create an accurate list of recommendations, the model analyses a vast amount of metadata content, such as movie's genres, film's cast, show's director, but also user activity data and user similarity to other users (Ko et al., 2022).

Due to ML techniques, Netflix is able to analyse practically everything, including forecast how many more movies they will produce, test emotional responses to films and consequently, adapting the movies' covers for individual personalization purposes (Fouladirad et al., 2015).

Such algorithms are beneficial for competitive enterprises, whose main driver is to deliver the best suggestions in order to retain their users on the platform as long as time allows (Reddy et al., 2018).

Overall, the usage of high-quality recommendation systems in the streaming field has increased user satisfaction (Ko et al. 2022), client experience and created a positive impact on enterprises' revenue (Fayyaz et al., 2020).

3.2.5. Challenges and Opportunities

Recommendation system applied to a particular streaming service is a comprehensive and complex task that involves multiple decisions, opportunities, and a variety of challenges.

These models are no longer new when it comes to e-commerce or streaming services, in fact these techniques are a serious and important tool to help users' decision-making process and to alleviate them from a large amount of information (Fayyaz et al., 2020).

The diversity of recommendation system usage can be seen as an opportunity, since it is able to provide recommendations and content to different areas in e-commerce targeting various attributes, such as location-based ad recommendations, agriculture-items to farmers and educational assistance to healthcare workers (Fayyaz et al., 2020).

However, recommendation models face several challenges. One of the most mentioned problems during the research is the Cold-Start problem, which occurs in the early stage of the learning process when it has not yet gathered sufficient information (Ko et al. 2022), and it is not feasible to provide credible recommendations due to lack of initial data. It is possible to break this problem down into three categories: new community, new item and new user (Bobadilla et al., 2012).

The first refers to the difficulty in obtaining sufficient amount of data which enables reliable recommendations to be initially made, the second arises due to the fact that the newly added movies or products do not usually have initial votes, and consequently they are not likely to be recommended by the algorithm. Lastly, the new user problem happens when the new

subscriber has no votes and, therefore, the model cannot generate any personalized recommendations to the user (Bobadilla et al., 2012).

The researchers face another issue in their investigation for the most accurate recommendation model which is the ignorance of the users' sentiment. Usually, users tend to choose the movies that most people prefer and skip the content that most other consumers dislike (Wang et al., 2018).

Lastly, another problem is the scalability of an effective recommendation system (Ko et al., 2022). Provide new recommendations to users has become a serious problem due to the massive amount of data, and the key challenge is to build an efficient algorithm that can handle such a large scale of information (Wang et al., 2018; Xin, 2015).

3.2.6. Methods of Evaluation

When the process of selecting the appropriate algorithm starts, the first step is to decide which specific attributes should be considered in order to feed the algorithm and, ultimately, provide the most efficient recommendations (Shani, & Gunawardana, 2010).

Over the years, academic researchers have paid more attention to traditional evaluation scenarios based on historical data, whereas industry practitioners tend to value the outcome of online evaluation on live systems, such as A/B testing, questionnaires or more user involved frameworks (Peška, & Vojtáš, 2020).

While offline evaluation is typically easier to assess, it is frequently argued that it might not accurately reflect the true effectiveness of recommendation systems, as demonstrated in online experiments (Peška et al., 2020).

Therefore, to assess the performance of the live recommendation system, two other online metrics were suggested: Click-Through Rate and Visit after Recommendation Rate. The first metric represents the ratio between the number of items clicked and the number of items recommended displayed, while VRR counts whether users visited the recommended items after those were presented (Peška et al., 2020).

3.3. BUSINESS INTELLIGENCE

3.3.1. Concept

Per definition, Business Intelligence is considered to be a decision-making process which supports the analysis of large amounts of data. It also provides actionable insights which allows informed business decisions (Romero et al., 2021).

BI combines operational organization's data with advanced analytical and data visualization tools to display complex information to decision makers (Negash, 2004).

The main purpose of a Business Intelligence solution is to improve the quality and responsiveness of critical business decisions through actionable data visualization. Therefore, it supports executives, managers, and other stakeholders to understand their available capabilities, but also perceive trends, key performance indicators, technologies to be used, up-coming directions in the markets and the current actions facing competition (Negash, 2004).

As the amount of data has exploded over the last few decades, BI systems are performing an increasingly critical role for companies. Due to those, organizations are able to understand, process and present insights about internal data within a short period of time. This information has been proven to be one of the most valuable assets of the business (Romero et al., 2021).

Therefore, large enterprises are heavily dependent on BI reports in order to provide fast and efficient business decisions and create competitive advantage in the market (Romero et al., 2021).

Business Intelligence solution serves also as a preprocessing tool that converts both structured and semi-structured data into valuable and meaningful information. Combined with an efficient analysis from a decision-maker, the use of BI systems can foster a powerful impact on the organization (Negash, 2004).

According to Negash (2004), a few examples of Business Intelligence assignments are:

- Creation of forecasts based on previous data.
- What-if analysis about the impacts of changes and alternative scenarios.
- Use the data to answer specific corporate questions.
- Creation of strategies based on the insights captured with the current data and performance.

Finally, BI applications have been proven to be essential for the strategy and competitiveness of the companies worldwide. These systems play not only an ultimate role in the survival of businesses, but also in counterintelligence between governments (Romero et al., 2021).

3.3.2. Business Intelligence Architecture

A successful BI solution implementation starts with a structured and consolidated BI architecture. This one articulates the technology and methodology to manage data in order to support the company's BI efforts (Ong et al., 2011).

To guarantee high data quality and smooth information flow within a BI system, five layers of BI Architecture are suggested which include data sources, ETL process, data warehouse, end-user, and metadata layers (Ong et al., 2011).

3.3.2.1. Data Warehouse

One of the most important components of BI is a structured and well-defined data warehouse. There are several definitions of DW. According to Kimball, a DW is "A copy of transactional data specifically structured for query and analysis" whose main purpose is "to provide information to support decision-making in a company". Whereas Inmon argues that “A warehouse is subject-oriented, integrated, time variant and non-volatile collection of data in support of management's decision-making process" (Yessad & Labiod, 2016).

In short, a data warehouse is a centralized repository that stores operational data in a particular manner to make it available and usable for analysis (Yessad et al., 2016). In other words, it is a process of combining multiple data sources in one single place (Linstedt, 2002).

As mentioned, the key objective of a DW is to store data provided from the core business, in a certain format, from multiple sources. Ultimately, to perform analysis and generate insights to support data-driven decisions (Yessad et al., 2016).

Aiming to build a data warehouse, the BI researchers must decide which data modelling approach should the solution be built upon. The approach should consider several aspects, such as data modelling, project management, risk management, deployment, and many other essential aspects for the organization (Yessad et al., 2016).

3.3.2.2. Modelling Types of Data Warehouse

To create the best conditions for analysing data, the organization should first choose a specific method of shaping their data. Data warehouse modelling represents the “process that produces abstract data models for one or more database components of the data warehouse” (Ballard et al., 1999).

The data warehouse modelling process consists of all tasks related to requirements gathering, analysis, validation, and modelling (Ballard et al., 1999).

In accordance with Yessad et al. (2016), there are three methods that dominate the BI market:

- Subject-modelling approach of Inmon.
- Dimensional modelling approach of Kimball.
- Lastly, the flexible and scalable "Data Vault" approach.

INMON APPROACH

During the 90's, a popular data modelling approach was founded by Bill Inmon. This method was the first being created and rapidly became known for its top-down development approach of an enterprise data warehouse, to meet the requirements of the organizations and to support the development of the decisional systems (Yessad et al., 2016; Breslin, 2004).

Inmon's method is built on the premises of the Entity-relationship diagrams of operational systems (Yessad et al., 2016).

The architecture of the data warehouse developed by Inmon include all data available from the company's information system in the database prior to knowing the user requirements. In his approach, Inmon proposed to load all the company's information systems into the data warehouse and believed that this one should be physically detached from the data mart (Yessad et al., 2016).

In Figure 7, according to Inmon's approach, the company's database is divided into four levels, which are operational, atomic (DW), departmental (data marts) and individual levels. The first one is related to day-to-day business transactions, the data processed into the second level of the database, the atomic. This one represents the data warehouse section which includes the ETL process. Typically, the data warehouse has its own physical existence, and it is planned towards storage, traceability, and scalability, avoiding data redundancy as much as possible, which results in clear responses to new business requirements. Moreover, the data is then split into physical existent data marts which respond to users' requirements and needs. Finally, the data is redirected to specific data access tools to be manipulated and analysed (Yessad et al., 2016).

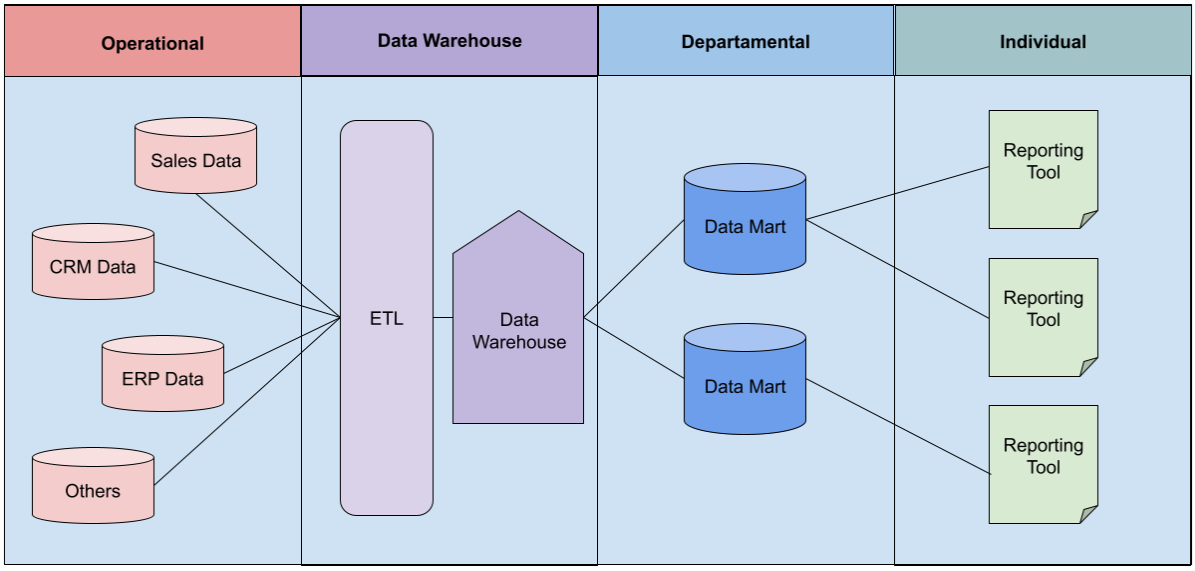

Figure 7 - Inmon's data warehouse architecture.
Source: (Yessad et al., 2016)

Inmon's data warehouse architecture approach has its advantages and disadvantages as represented in Table 1.

Table 1 – Inmon's data warehouse architecture advantages and disadvantages.

Advantages	Disadvantages
The essence of "one source of truth" is met, since all the data in the data warehouse is integrated (Yessad et al., 2016).	Long installation and delivery process, which can be a great issue to meet the business deadlines (Breslin, 2004).
Minimal redundancy since it stores cleaned and normalized data. Eases the ETL process, avoids duplicates and makes it less likely to fail (Inmon, 2008).	Low query performance over time, as it involves more computational power to join more tables respecting 3NF data structure (Yessad et al., 2016).
Addresses a large company-wide reporting needs, since it integrates the complete organizational data, even noncritical business metrics (Breslin, 2004).	End-users have mostly passive roles in the development of the data warehouse (Breslin, 2004; Inmon, 2008).
Very flexible to adapt when business requirements or new data sources are added (Yessad et al., 2016).	High ETL work overload since it is harder to develop data marts from the data warehouse and relies heavily on IT professionals (Breslin, 2004).
	Due to its initial complexity and size solution, it requires higher start-up costs with lower further development costs (Yessad et al., 2016; Breslin, 2004).

KIMBALL APPROACH

In the same period of time, another data modelling method was developed and presented, this time by Ralph Kimball. This approach became known as a bottom-up approach of a dimensional data warehouse. Rapidly, it was accepted and adopted due to its new data architecture design, vision, and an innovative modelling strategy of a data warehouse. As in the previous method and as shown in Figure 8, in this approach, the database is also divided into four levels, which are operational, departmental, Data Warehouse (data marts, included) and individual levels. The major difference between these two methods is that Kimball's involves the end-users in the initial stage of the process, and this is why it is also known as the "user requirements-driven" approach. Moreover, Kimball also defends that a data warehouse

can be envisioned and built as a collection of coherent data marts that rely on shared and standardized conformed dimensions (Yessad et al., 2016).

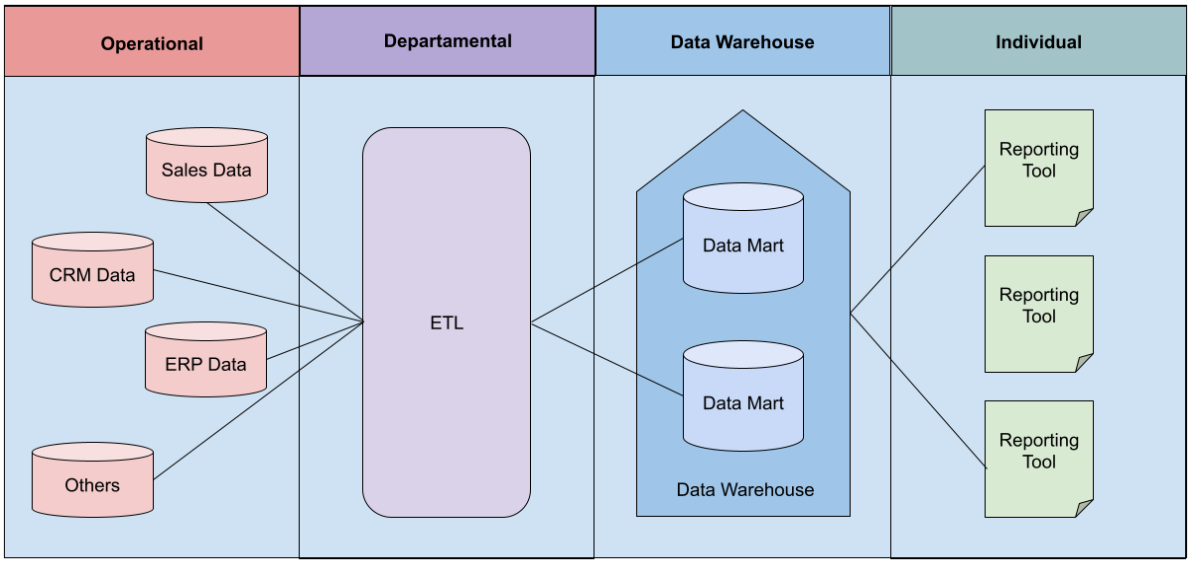

Figure 8 – Kimball's data warehouse architecture.
Source: (Yessad et al., 2016)

Kimball's approach is based on the concept of dimensional modelling which highlights the subject and the different analysis perspectives. Kimball also suggested developing one database per major business process, such as sales, marketing, or any other relevant dimension (Breslin, 2004).

Moreover, it presents several unique elements such as fact table, dimensions, conformed dimensions, and bus matrix (Yessad et al., 2016). In detail, the first two elements represent the main concepts of this method. According to Kimball, "*a fact table is the primary table in a dimensional model where the numerical performance measurements of the business are stored*" (Kimball & Ross, 1996).

As with the previous method, this method also has advantages and disadvantages that should be considered. In Table 2, they are described in detail.

Table 2 – Kimball's data warehouse architecture advantages and disadvantages.

Advantages	Disadvantages
Faster approach, quickly data mart delivery and initial set-up phase of the data warehouse (Breslin, 2004).	Complex ETL process, as mentioned before not all raw data is integrated, which requires an initial complex process of data cleaning (Yessad et al., 2016; Kimball et al., 1996).
User-friendly and fast query performance due to the dimensional methodology, since it needs less computational power to conduct the needed joins (Breslin, 2004).	Ongoing incremental development with the addition of new metrics and attributes, a constant DW work is needed (Yessad et al., 2016).
Since it only uses business data, it avoids loading other organizational data. Therefore, uses less storage space then other approaches (Kimball et al., 1996).	Does not address enterprise reporting needs, since it is a tactical approach oriented (Breslin, M., 2004).
Addresses better smaller teams of developers, since it integrates mostly the core business data (Breslin, 2004).	Low flexibility to adapt when business requirements change (Yessad et al., 2016).
Great tracking approach for a department-wise since it suggests one data mart per business-unit (Breslin, 2004).	
Due to its flexibility and simple installation phase, it also provides lower start-up costs and each further phase costs as much (Yessad et al., 2016; Breslin, 2004).	

"DATA VAULT" APPROACH

In the early 2000, Dan Linstedt introduced an emerging data warehouse architecture approach. According to the author, Data Vault's approach is "a detail oriented, historical tracking and uniquely linked set of normalized tables that support one or more functional areas of business" (Linstedt, 2002).

The main goal of this approach is to meet the requirements of the enterprise data warehouses. This approach is also known for its flexible, scalable, consistent, and adaptable design in order to meet the needs of the organization. Data Vault's method is considered to be a hybrid approach merging the best of the above methods, between the 3rd normal form (3NF) and the star schema (Linstedt, 2002).

To address the other approaches' weaknesses, Linstedt suggests a few changes and adaptations in processes and data structures (Yessad et al., 2016).

As shown in Figure 9, at the very beginning, the structural and descriptive information are separated so that the principle of flexibility and adaptability prevails. This approach allows loading of data from multiple sources of data in parallel. After loading, the data is never processed, filtered, or changed in order to be available "all the data, all of the time" as stated by Dan Linstedt (Yessad et al., 2016).

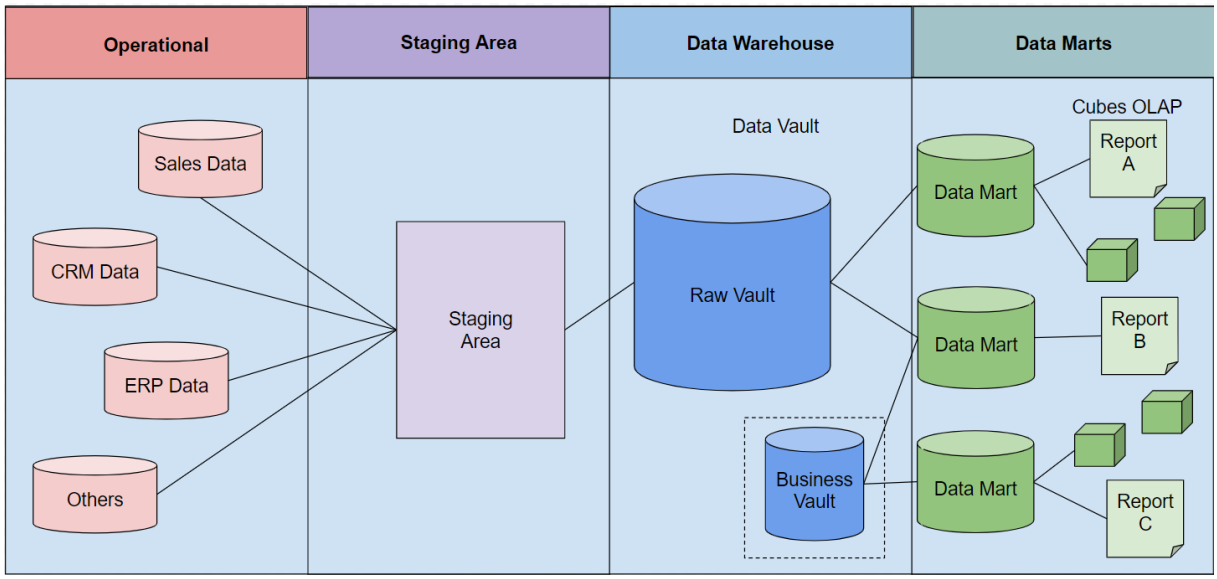

Figure 9 – Data Vault's data warehouse architecture.
Source: (Yessad et al., 2016)

Finally, Data Vault's approach is based on three tiers architecture, which are the preparation phase which includes Operational and Staging area, the Data Vault in data warehouse, and the data marts (Yessad et al., 2016).

Even though Data Vault's approach tries to address many of the other DW methods' weaknesses, besides its advantages, it also has a few disadvantages that should be mentioned. In Table 3, these are described in detail.

Table 3 – Data Vault's data warehouse architecture advantages and disadvantages based on the academic paper from Yessad et al. (2016).

Advantages	Disadvantages
Fast installation and build the initial phase of the data warehouse.	Very slow query performance, due to high data standardization. Multiple dimensional data marts are required to create the reports.
Data Vault's method requires less resources and therefore is less costly, when compared with the previous two approaches.	Not effective in a business environment with a few data sources with limited changes of their structures.
Low complexity in the ETL, since the rules are simple to apply and load hubs, links, and satellites.	

3.3.3. Visualization

Besides data architecture, Data Visualization represents a crucial component in BI, since it illustrates the data through graphical images and shares its knowledge (Chen et al., 2022).

The memorable proverb "A picture is worth a thousand words" fits perfectly into the concept of Data Visualization. DV extracts information from complex data and provides a clear and meaningful outcome. In fact, it is able to identify structure, patterns, trends, anomalies, and relationships in data that otherwise would be significantly harder (Negash, 2004).

The key objective of Data Visualization is to enable the researchers and general players to explore, interpret, and present the results in order to make data-driven decisions (Chen et al., 2022). Ultimately, DV is used to create advanced reporting dashboards with large amounts of data in a clearer, structured, and dynamic way presented on a single screen (Negash, 2004).

However, creating meaningful Data Visualization is not an easy task, especially when developed by web-lab scientists instead of data designers (Chen et al., 2022).

In the current fast-paced business environment, professional roles of DV designers are increasing in popularity. More than ever before, the relationship between academic researchers and professional practice communities is gaining interest in an attempt to create a single model design framework which would approximate closely towards a more scientific method. With the intention of better understanding this relationship, 87 DV practitioners and 20 professional practitioners were interviewed in order to understand what the best practices are when creating a Data Visualization dashboard (Parsons, 2022).

In Figure 10, the results of techniques and principles was displayed.

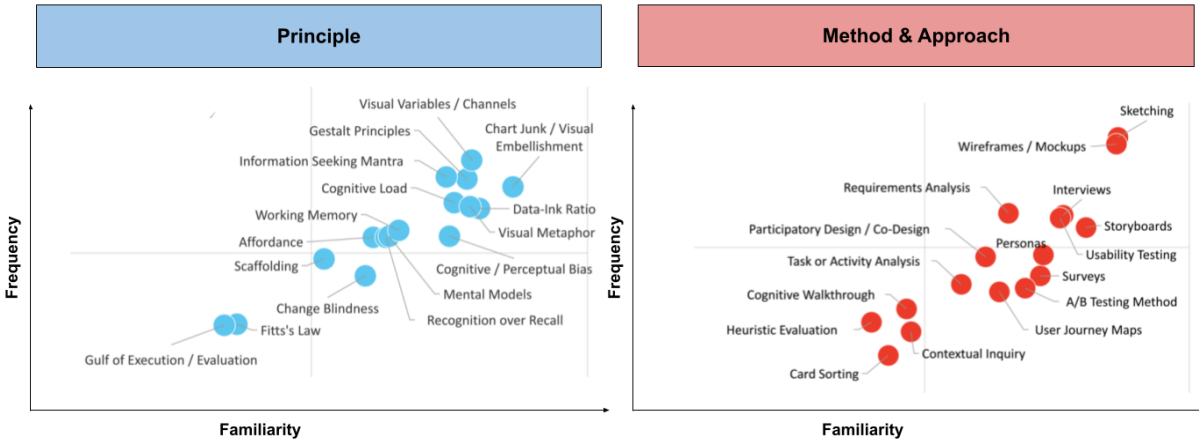

Figure 10 – Results of the survey regarding the use for design methods and principles. Source: (Parsons, 2022)

The first group stated to be most familiar with the principle "*Chartjunk*", a term coined by DV expert Edward Tufte to refer to elements of a chart that do not add value to the presented data, but rather serve to distract or confuse the viewer. Other familiar fundamentals used by the DV practitioners included visual variables, Gestalt principles, and data-ink ratio. On the other hand, other methods that they confessed are less used due to time or resource consumption are A/B testing and surveys (Parsons, 2022).

The other group reported not to rely on any logical methodology, such as distributed cognition, activity, or dual process theory. Indeed, the 20 professional practitioners confessed using some form of "intuition" or "gut feeling" that led them through their creative DV process. To start, the group described to perform some methods, such as data cleaning and wrangling, brainstorming, talking to clients, sketching, and prototyping, or user testing. After that and to guide them to create, they not only use sketching and precedent visualization ideas, but also use Pinterest boards, art, books, and museums to inspire them. Nevertheless, they recognize the importance of scientific knowledge, especially when choosing chart elements, such as lines instead of angles (Parsons, 2022).

3.3.4. Tools

In current business environments, companies are expecting that their decision makers are able to create their own BI dashboards in order to explore, analyse and ultimately make critical business decisions to hopefully gain competitive advantages (Romero et al., 2021).

To implement an effective Business Intelligence solution, an impactful Self-Service BI software has to be identified and chosen by the company. Through SSBI, the DV designers and business users have the ability to create, select and analyse their own reports and dashboards without requiring direct assistance from technical resources, including IT departments (Schuff et al., 2016).

Moreover, the design study methodology developed by DV researchers resulted in the development of useful software (Parsons, 2022). In specific, SSBI tools have numerous benefits, such as increased flexibility, release resources for other activities, and reduce the workload on the IT department (Romero et al., 2021).

According to Romero et al. (2021), a few examples of cutting-edge software are:

- Microsoft Power BI.
- Tableau from Salesforce.
- Jasper Reports.
- Pentaho.
- SpagoBI.
- Palo / Jedox.
- and Qlik.

3.3.5. Usage of Business Intelligence on Streaming Platforms

As mentioned before, BI systems are playing an increasingly important role for enterprises. The use of these systems is not exclusive to a single industry, in fact more industries are adopting these mechanisms, such as healthcare, sustainability, education, automotive, finance, gaming, environmental surveillance, and many more (Romero et al., 2021).

It has been proven that Business Intelligence and Analytics technologies are principles of development for companies, since they support the decision-making process, forecasting, and corporate economy (Romero et al., 2021). The development of advanced models has a big impact on how companies make better data-based decisions and become more competitive in the market (Fouladirad et al., 2015), and the Online Streaming industry is no exception.

Finally, through a large amount of historical data, the goal of BIA is to provide relevant and reliable information to the right people in a meaningful way to support the decision-making process in real-time (Sudhakaran, 2021).

3.3.6. Challenges and Opportunities

The implementation of a new technology brings challenges and opportunities that decision-makers should be aware of and address them.

To start, implementing a Business Intelligence solution is not an easy task and requires a great number of resources. One of the biggest challenges is the implementation costs associated with this system, such as hardware, software, and personnel costs, including data warehouse purchase, BI packages, training, and BI jobs (Negash, 2004).

One of the most difficult challenges to any leadership also caused by this technology is the cultural and behaviour change within the organization that needs to be made. Lastly, the challenge of misinterpretation by the ones who read a BI reporting, due to lack of communication or different perceptions between the author and the users (Romero et al., 2021).

On the other hand, studies confirm that multiple benefits can originate from the implementation of BI tools, such as improved performance, efficiency, productivity, business growth, resource planning, supplier–buyer relationship, among others. All of those prove that implementing a BI system can lead the organization to a competitive advantage (Romero et al., 2021).

Indeed, SSBI solutions such as Power BI reporting, can certainly increase flexibility, release resources which can fund other critical activities, and reduce the workload on technical departments (Romero et al., 2021).

Finally, Business Intelligence and Analytics solutions can greatly benefit enterprises by supporting them in their strategic decisions, predicting future opportunities and leading towards the growth of the organizations (Sudhakaran, 2021).

4. Project Development

Based upon the theoretical research compiled and analysed in the previous chapter, this section outlines the comprehensive five steps taken to develop the ultimate objective of this project. This chapter provides a detailed description of all the relevant procedures involved.

4.1. Defining Project Goals, Needs, and Assumptions

In the initial phase of this project, it was extremely important to establish the main key topics regarding the project, such as goals, needs and assumptions taken throughout the work developed. The combination of these key topics formed the foundations of the project, since it effectively communicated the thought process and identified the specific topics that the author aimed to address.

NEEDS

In a highly competitive entertainment market, it is essential to understand the users' behaviour and acknowledge patterns of preference. Therefore, two major needs were identified: economic and social. These needs addressed the requirements of financial sustainability and user satisfaction in this market.

Starting with the first one, through a Business Intelligence solution with an efficient recommendation system, the streaming platforms can observe a substantial improvement of user engagement and increase of revenue due to monthly subscriptions, monetization of personalized ads, or content purchases according to users' preferences. Moreover, recommendation systems applied in a BI solution play a crucial role in retaining users by offering personalized content suggestions, enhancing user satisfaction, boosting loyalty, reducing user churn rates and, therefore, raising a stable subscriber base number for each streaming platform.

Concerning the social needs, the author identified and highlighted the fact that users deal daily with the challenge of content overload. Therefore, the need of simplify content portfolio, narrow the number of suggestions displayed, and reduce the time and effort required to the users to find the desired movie was found. Moreover, a well-designed recommendation system could also promote cultural exchange by suggesting content from different regions, languages and perspectives which share similar traits such as movie category, rating, or some elements from the cast.

GOALS

As mentioned previously, the main goal of this project was to develop a comprehensive BI solution for the four leading global streaming platforms, Netflix, Amazon Prime, Disney+ and Hulu. The primary focus was to design and implement an efficient content-based recommendation system, in order to support data-driven decisions of the business analytics team. Ultimately, the aiming outcome of this project was to improve user growth, enhance user engagement, and increase company's revenue.

ASSUMPTIONS

Considering the risks involved in the extraction process of the data from open sources on the internet, it was extremely important to state the assumptions made during the development and analysis of the project. By clearly documenting these, the work developed can maintain transparency and communicate potential limitations associated with the data.

To start, it was assumed that the five principles of data integrity were present and reflected an accurate reflection of the content consumed by their users, such as completeness, consistency, accuracy, validity, and timeliness of the samples.

In particular, it was assumed the Data Accuracy, meaning that the data was accurate and reflected a reliable representation of the content consumptions of the platforms. Moreover, it was assumed the Data Relevancy, which means that the data obtained from multiple streaming platforms held all relevant attributes and features for a meaningful analysis and comparisons. And finally, it was also assumed that Data Bias was respected, that the sample collected from an open source was free from significant inherent biases that may jeopardize the results and insights from this analysis.

4.2. TECHNICAL OPTIONS

According to project specificalities, including goals, needs and assumptions, and as analyzed in the theoretical framework section, particular techniques were chosen. In this section, a detailed explanation was provided.

4.2.1. Word Embedding and Content-Based Filtering Techniques

Due to the availability of data for the development of this project, two types of techniques were used in order to improve the quality and performance of the final recommendation system.

To capture the real semantic context and similarity between words present in the content attributes, the first technique used was Word Embedding. In recent times, word embedding techniques have increased in popularity, primarily due to their remarkable performance across several NLP tasks (Zhao et al., 2020).

Word embedding algorithms create a low-dimensional vector space representation of words, enhancing machines to comprehend and process textual data as a human would do. This methodology aims to capture the semantic essence of words and their contextual connections, enabling extraction of meaningful data, establishing a profound semantic and relational data understanding and, ultimately, quantifying semantic similarities between words or documents (Zhao et al., 2020).

Word2Vec, Doc2Vec, Glove, and SciBERT represent a few of the most notorious word embedding algorithms employed for content representation (Zhao et al., 2020).

While the second technique used was Content-Based Filtering, this approach was considered to be suitable for the project's objectives and requirements, since it compared several attributes from each video content and based on those, it recommended other identical items to the user (Paul et al., 2019).

Several algorithms were considered in order to create the most adequate model and respective recommender outcome. A comprehensive description of the chosen algorithm is presented later in this development chapter.

4.2.2. Kimball’s Approach

In today's fast-paced and demanding business environment, it is crucial for end-users to have permanent access to data to effectively carry out their queries and deliver insightful business intelligence reports. With the aim to reduce IT department dependency and reliance, there is a growing emphasis on adopting the use of self-service and user-friendly models in designing data warehouses (Moody, & Kortink, 2000).

With the goal of creating an efficient data warehouse design using an accessible model for this project, the author has decided to choose Kimball's data modelling method as the second technical choice presented in the chapter. This decision was based on the low level of architecture’s complexity in this technique, its high and fast query performance, clear understanding by its final-users and, lastly, due to its well-known business requirements (Yessad et al., 2016).

In Figure 11, the process of data collection was presented according to Kimball’s approach. First, it can start from two possible type of sources, the operational system (1) or other external data sources (2). The first one, represents the internal system which stores the information from the business and the second, includes other additional information from

external sources to support the data analysis. After all the data collected, the extract process (3) starts which includes preprocessing and standardization of the data in order to consolidate all the information in a consistent and structured form. Then, a central Data Warehouse (4) is created, where it becomes a central data source of decision support. Once it is created, the loading process initiates (5), in this step the data is distributed and allocated to the respective data marts (6). Usually, these represent real or virtual "data outlets" which provide usable information to the final consumers. Data marts are usually created to fit the needs of a specific group of users or decision-making agents (7). Finally, those users are the ones who perform the queries and analysis using self-service tools (Moody et al., 2000).

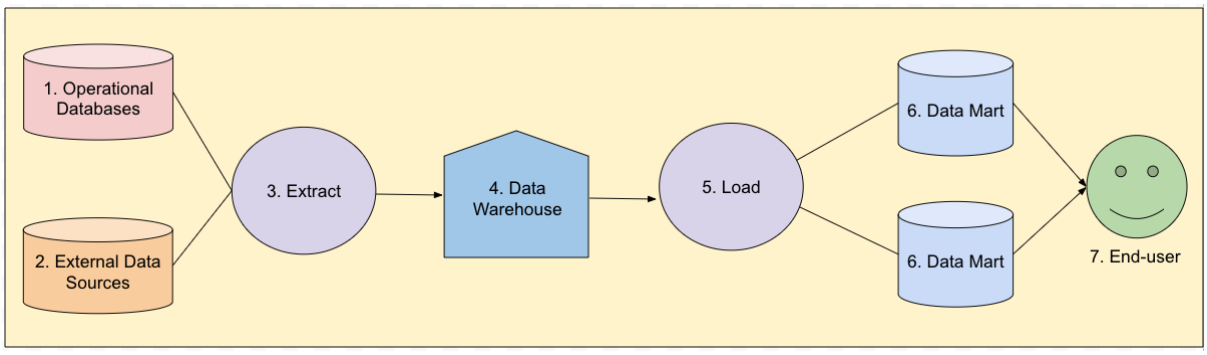

Figure 11 – Data Warehouse architecture.
Source: (Moody et al., 2000)

According to Kimball, the data warehousing (OLAP) environment is deeply different from the operational (OLTP) environment. Therefore, a more suitable technique should be used in order to design effective data warehouses. For this reason, a new clearer and accessible modelling approach for data modelling was created, named dimensional modelling (Moody et al., 2000).

Albert Einstein understood the basis of driving dimensional design when he said, "*Make everything as simple as possible, but not simpler*." This technique has become widely accepted as the preferred technique for presenting analytic data due to its simplicity, since it addresses two simultaneously important requirements. First, by creating an optimized database structure that facilitates access to the end-users in order for them to comprehend the data and write their queries; and secondly, to maximize the efficiency of query performance by reducing the number of tables in the model, simplifying the relationships between the attributes and minimize the need of joins required to perform the queries (Kimball et al., 1996; Moody et al., 2000).

Dimensional modelling is a longstanding technique for making databases simple by reducing its level of complexity (Kimball et al., 1996).

Implement structures of dimensional models in relational database management systems are known as star schemas due to their resemblance to a star-like structure. They consist of one

large fact table and various smaller dimension tables connected through primary and foreign keys relationships which radiate around the central table (Moody et al., 2000).

In contrast, Online Analytical Processing (OLAP) cubes are dimensional structures found in multidimensional databases. As shown in Figure 12, both star schemas and OLAP cubes share a common design with identifiable dimension tables, but the process of their physical implementation is different. OLAP cubes store and index data using formats and techniques specifically designed for dimensional data using more analytical complex languages than SQL, such as XMLA and MDX. Those include precalculated summary tables and optimizations for superior query performance. Kimballs' recommended loading detailed atomic information into a star schema and then, populated OLAP cubes from it (Kimball et al., 1996).

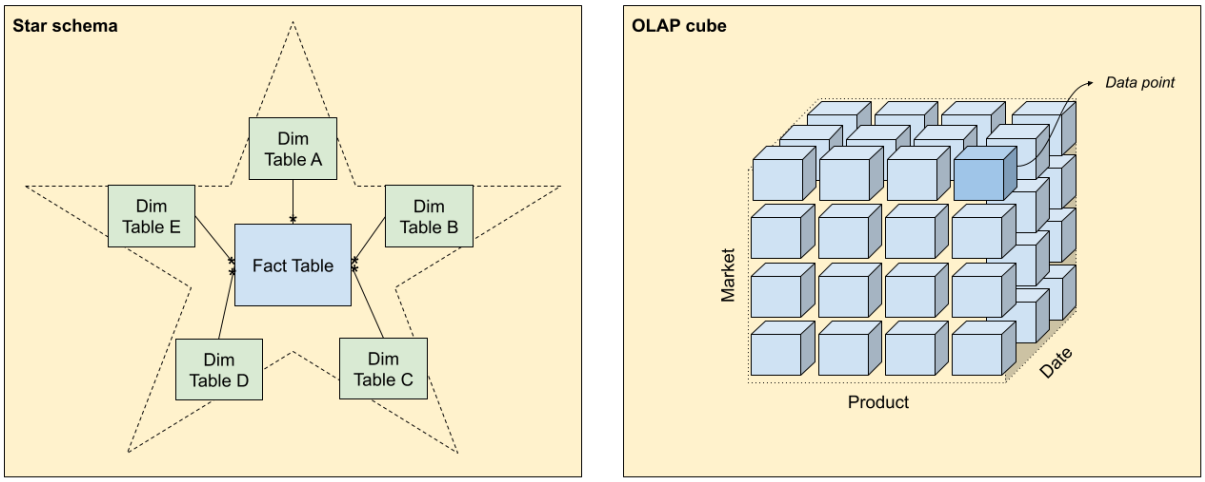

Figure 12 – Star Schema and OLAP cube design according to Kimball's dimensional.
Source: (Kimball et al., 1996)

The use of a star schema provides significant advantage to end-users by reducing the complexity of the database and improving the query performance. This efficiency is achieved by minimizing the number of tables and relationships that need to be considered and joined.

According to Kimball et al. (1996), approximately 80% of queries performed by end-users using star schemas in data warehouse design resulted in accessing just a single table (Moody et al., 2000).

Moreover, Kimball recommended building the data modelling based on a first principles approach. This means that it starts by identifying the relevant facts and dimensional attributes that need to be aggregated, based on the analysis of the business and user requirements (Moody et al., 2000).

4.2.2.1. Fact and Dimensional Tables

As mentioned before, the dimensional model technique includes one central fact table and various smaller dimensional tables. Starting with the first one, the fact table contains the performance measurements resulting from the business operations which can be aggregated in various ways, such as the price, the quantity, currency, and collection of products sold (Kimball et al., 1996; Moody et al., 2000).

Usually, fact tables represent 90% of the total consumed space of the dimensional model. They tend to be deep, i.e., have a large number of rows, but narrow in terms of the number of columns. That's the reason why redundant information should be avoided in fact tables. Concerning the data type of the facts, the most beneficial ones in the fact table are the continuously valued ones, such as the numeric and additive. In theory, it is also possible to add textual facts, however this condition is very hard to achieve. The textual data is hardly unique or error-prove, which makes it nearly impossible to analyse. Unless the data is unique for every row, this information belongs in the dimension table (Kimball et al., 1996).

The fact table typically has two or more foreign keys that connect to the dimension tables' primary keys through a one-to-many relationship. The primary key, i.e., the unique identifier assigned to each row in a table, plays a crucial role in maintaining data integrity and enabling the join operations necessary for querying and analysing data in the model. When all keys in the fact table correctly match the respective primary keys in the dimension tables, these satisfy the referential integrity condition. The primary key of the fact table is the combination of the primary keys of all the related dimension tables, which is often called a composite key (Kimball et al., 1996; Moody et al., 2000).

On the other hand, the dimension tables represent the information associated with business process measurement events, usually providing text information such as the "who, what, where, when, how, and why" of each event (Kimball et al., 1996).

Dimension tables provide the foundation for aggregating and calculating the measurements present on the fact table. In fact, dimension tables tend to have fewer rows, but typically a wider range of columns or text attributes (Moody et al., 2000).

Each dimension is defined by a single identifier primary key, establishing referential integrity when joined with a fact table as presented in Figure 13. Dimension attributes play a crucial role in query constraints, groupings, and report labels. That's the reason why dimension attributes should consist of meaningful and relevant business terminology rather than codes or abbreviations (Kimball et al., 1996).

Deciding whether a numeric element is a fact or dimension attribute can occasionally be ambiguous. In order to correctly make the decision, the DW/BI analyst should consider if the element represents a measurement that takes on numerous values and participates in calculations, such as a fact, or if it is a discretely valued description that remains relatively

constant and contributes to constraints and row labels, such as a dimensional attribute (Kimball et al., 1996).

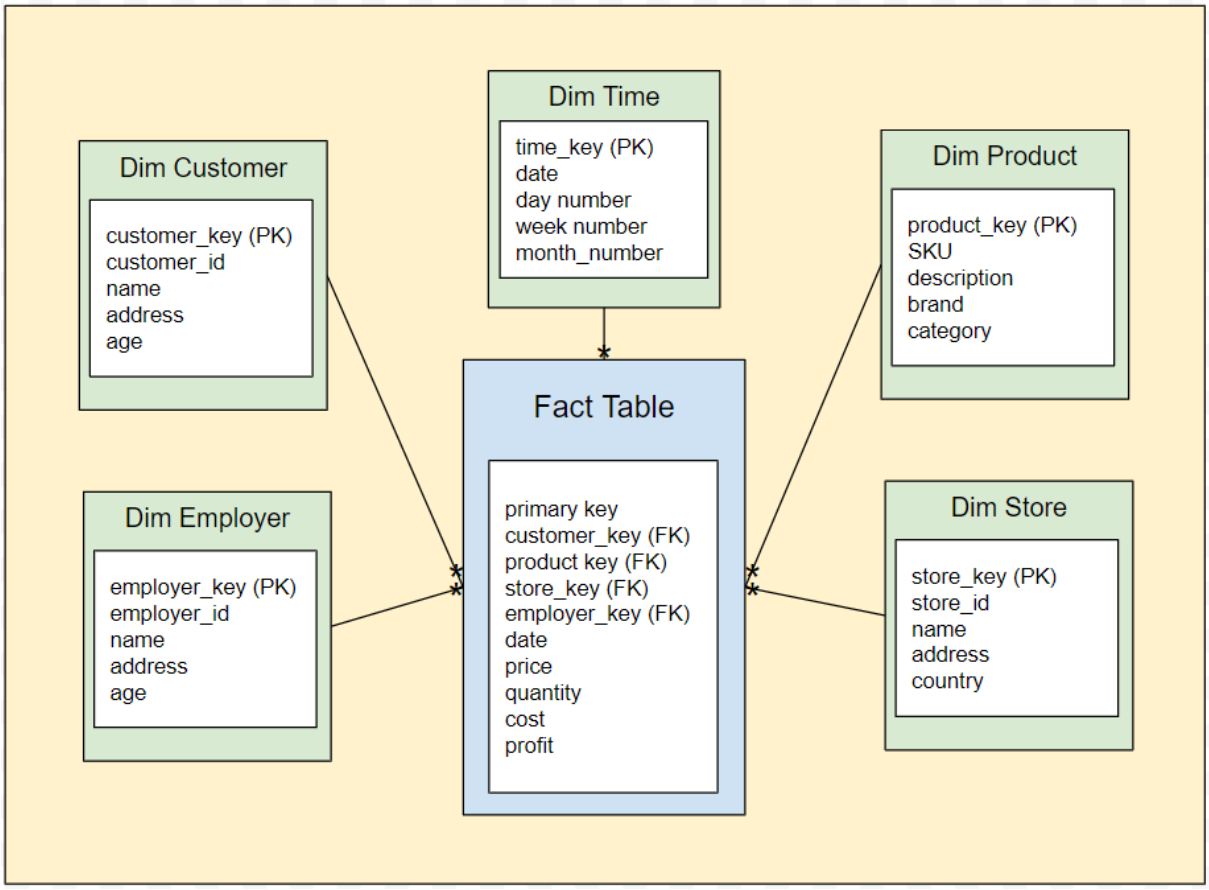

Figure 13 – Example of a fact table and dimensional tables in a dimensional model. Source: (Kimball et al., 1996)

The dimensional model stands out due to its simplicity, symmetry, and performance advantages. The final business users benefit greatly from its user-friendly navigation and readability. The simple structure of the schema makes data easier to understand and query it, reducing the probability of making mistakes or assumptions. The reduced number of tables and joins combined with the meaningful business descriptors enhance further the simplicity and user performance (Kimball et al., 1996).

Moreover, dimensional models are outstanding in accepting change and extension. The consistent and predictable framework allows the integration of unexpected shifts in user behaviour, since every dimension holds equal importance, serving as symmetrical entry points into the fact table. Also, new dimensions can be seamlessly added to the existing schema as long as each existing fact row has a defined value for the new dimension. Similarly, new facts can be incorporated into the fact table, provided they maintain a consistent level of detail with the existing data. The dimensional model is not designed based on prefixed questions or

preferences, and therefore no inherent bias towards specific query patterns or preferences (Kimball et al., 1996).

4.2.3. DV Tool: Microsoft Power BI

According to the owner entity Microsoft, Power BI is a unified, scalable platform for self-service and enterprise business intelligence. In fact, Power BI consists of several elements such as software services, applications, and connectors that all together can connect unrelated data and transform it into coherent, visually immersive, and dynamic reporting dashboards.

As illustrated in Figure 14, Power BI offers three different ways of collaborative work, such as a desktop application called Power BI Desktop, an online software as a service named Power BI service, and finally, Power BI Mobile apps for Windows, iOS, and Android devices. Moreover, it is able to connect different data sources, such as excel spreadsheets, cloud-based data storage or on-premises data warehouses all in one dashboard sheet.

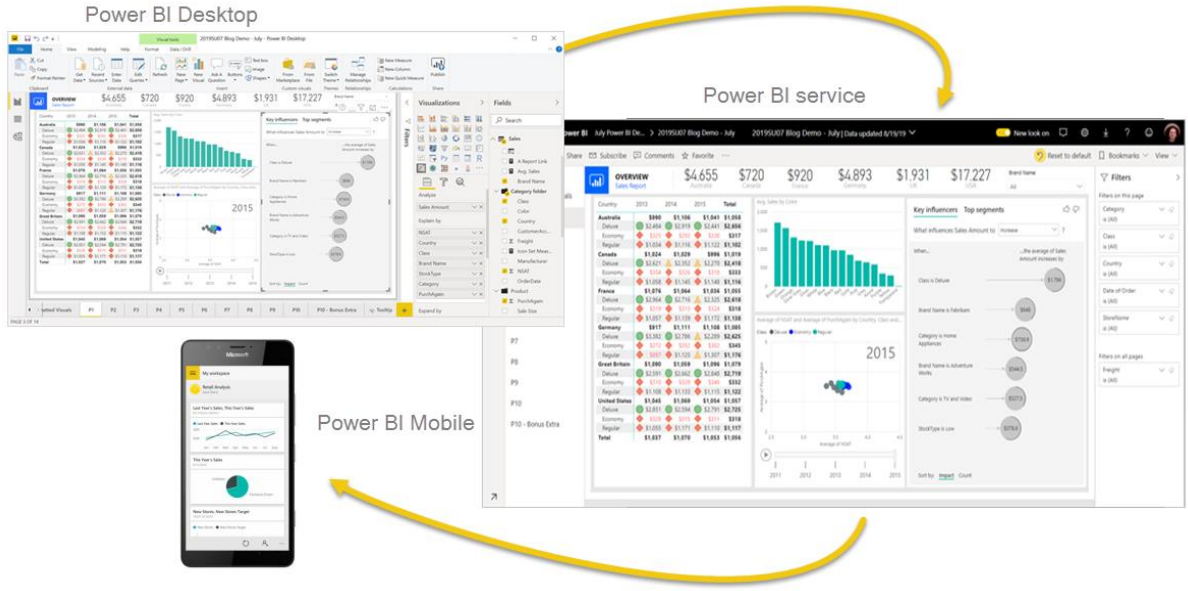

Figure 14 – Summary of all Power BI elements: Power BI Desktop, Service and Mobile app. Source: (Microsoft Learn, 2023)

Furthermore, the Power BI tool offers two other very important features which represent a real competitive advantage compared with other BI tools. The first element is the Report Builder, the area dedicated to creating paginated reporting dashboards to share in the Power BI service. The second feature is the Power BI Report Server, an on-premises report server where it is possible to publish and edit Power BI reports, after creating them in the Desktop version.

Besides the features and benefits named previously, Microsoft Power BI also built a strong and reputable name across the market. Gartner, Inc. (2023), has recognized Microsoft Power BI as the top leader of Analytics and Business Intelligence Platforms in 2023 as displayed in Figure 15 for the 5th consecutive year.

Figure 15 – Magic Quadrant for Analytics and Business Intelligence Platforms.
Source: (Gartner, Inc. 2023)

Therefore, due to all the reasons and benefits mentioned before and with the aim of using a convenient tool to develop this Business Intelligence solution, the author has decided to choose Microsoft Power BI tool as the third and final technical choice presented in the chapter.

4.3. DATA EXPLORATION AND UNDERSTANDING

With the primary objective of developing a Business Intelligent solution which resembled real-world scenarios, the data used in this project was obtained from an open data source from the internet provided from the four world's largest online streaming services, namely Netflix,

Amazon Prime, Disney+, and Hulu. The data was extracted in 2021 in the geographical location of the United States of America.

NETFLIX

To start, Netflix is a popular streaming service that allows subscribers to watch TV shows, movies, documentaries and more on demand. The company was founded in 1997 by Reed Hastings and Marc Randolph, with the aim of providing people with a convenient way to rent movies and TV shows without having to leave home. Currently, Netflix has become one of the world's most popular streaming services, with more than 200 million subscribers by 2021 (Statista, 2023).

It started by operating as a DVD rental service, but in 2007, the company launched its streaming service, which allowed subscribers to watch content instantly on their computers. In the following years, Netflix scaled up its service to include streaming on a wide range of devices, including smartphones, tablets, and smart TVs.

Netflix's business goal is to provide a wide range of entertainment options to subscribers around the world. The company creates and licenses content from a variety of sources, including television networks, movie studios and independent producers, and makes it available to subscribers upon a monthly subscription. Netflix's most popular original content included series such as Stranger Things, The Crown and Orange is the New Black, as well as films such as Bird Box and The Irishman.

AMAZON PRIME

Slightly different, Amazon Prime is also a subscription-based service offered by Amazon. It provides members with a variety of benefits, including free shipping, access to streaming video and music content, and other exclusive deals and discounts. The service was launched in 2005, initially offering just the free shipping benefit, but it didn't take long to expand to a wide range of other features. Nowadays, Amazon Prime has over 200 million members worldwide, making it one of the largest subscription-based services in the world (Statista, 2023).

The business purpose of Amazon Prime is to provide an all-in-one subscription service that makes it easier and more convenient for customers to shop on the web provider Amazon.com and access digital content. By offering free shipping and other benefits, Amazon aims to encourage customers to make more purchases on its site, while also increasing engagement and loyalty regarding other services such as online streaming.

Amazon Prime was created by Jeff Bezos, the founder and CEO of Amazon.com, and has since become one of the company's most important services. In addition to the benefits listed

above, Amazon Prime also offers members access to exclusive content, such as original TV shows and movies, as well as a selection of e-books, magazines, and other digital content.

DISNEY+

Thirdly, Disney+ is also a streaming service based on a monthly subscription owned and operated by The Walt Disney Company. The service was launched in November 2019 and offers a wide range of TV shows, movies, and other content from Disney, Pixar, Marvel, Star Wars, and National Geographic.

Disney+ was created by a team led by Chairman and CEO Bob Iger. The business purpose of this additional service is to provide a dedicated platform for Disney to showcase its extensive library of content, as well as to create and distribute its own original programming. The service is focused on family-friendly programming, features a mix of animated and live-action content, as well as documentaries and other educational programming.

In addition to its library of classic movies and TV shows, Disney+ has also produced several original series and movies, including The Mandalorian, WandaVision, and Soul. At the moment, Disney+ has over 100 million subscribers worldwide, making it one of the fastest-growing streaming services in the world (Statista, 2023).

HULU

Finally, Hulu is a subscription-based streaming service that offers a wide range of TV shows, movies, and other content to subscribers.

In 2008, the streaming service was launched by a team of executives from several major media companies, including Fox Broadcasting Company President, Peter Chernin, and NBCUniversal CEO, Jeff Zucker. The service has been acquired by The Walt Disney Company, which now owns a majority stake in the company. Globally, Hulu has over 41 million subscribers in the United States and Japan, making it one of the largest streaming services in the world.

The business aim of Hulu is to provide a streaming platform for its content partners to reach audiences online, as well as to create and distribute its own original content. The service is focused on providing a broad selection of TV shows, including current and classic series from major networks like ABC, NBC, and Fox.

In addition to its extensive library of TV shows, Hulu also offers a selection of movies, documentaries, original content such as The Handmaid's Tale, Little Fires Everywhere, and Pen15, among other programming content.

In Table 4, an overview of the four streaming platforms main information was provided.

Table 4 – Streaming Movie Platforms Overview until 2022.

Platform	Founder	Year	Users (in million)	Revenue (in billion US$)
Netflix	Reed Hastings & Marc Randolph	1997	230.7	31.6
Amazon Prime	Jeff Bezos	2005	157.4	25.2
Disney+	The Walt Disney Company	2019	164.2	23.5
Hulu	Jason Kilar & Beth Comstock	2007	47.2	10.7

In the following table, a detailed summary of the volume of the data collected through open sources was added, as well as a description of each attribute.

Table 5 – Streaming Movie Platforms datasets in detail.

Company	Netflix	Amazon Prime	Disney+	Hulu	Description
Show_id	8807	9668	1450	3073	Represented the unique identifier assigned to each TV show or movie within a streaming service's database.
Type	8807	9668	1450	3073	Indicated whether the content was a TV show or a movie.
Title	8807	9668	1450	3073	The title of the content.
Director	6173	7586	1450	3073	The name of the director(s) of the TV show or movie.
Cast	7982	8435	977	3	The names of the actors and actresses who performed in the content.

Country	7976	672	1260	0	The country where the content was produced or filmed.
Date_added	8797	155	1231	1620	The date when the TV show or movie was added to the streaming service's library.
Release_year	8807	9668	1447	3045	The year when the content was first released to the public.
Rating	8803	9331	1450	3073	The appropriate age rating for the content.
Duration	8804	9668	1447	2553	The length of time, in minutes or hours, of the content.
Listed_in	8807	9668	1450	2594	The category of the TV show or movie.
Description	8807	9668	1450	3073	Provided a synopsis of the TV show or movie.
Platform	8807	9668	1450	3073	Indicated the specific streaming service where the content is available.

4.4. LITERATURE FRAMEWORK FOR SUPPORT RATIONAL DECISION-MAKING

During this phase of the project, a compilation of scientific papers was gathered and analysed within the scope of Recommendation Systems and Business Intelligence. The research included several concepts and techniques in order to establish a robust theoretical framework to support practical decisions made throughout the project's development.

The literature collected during this step was introduced and explained in detail previously in the Theoretical Framework section of this project. The theoretical knowledge gathered represented the foundation for the entire practical work in the development section.

4.5. ELABORATION OF THE RECOMMENDATION SYSTEM ALGORITHM

At this stage of the project, the four streaming platforms were successfully extracted and organized in four different comma-separated values files. Then, a machine learning project was developed aiming to assess and identify the most suitable and effective recommendation system model. As mentioned previously, the programming language used in this project was Python.

In the following section, a detailed explanation of all the techniques and procedures made were provided. It's important to note that the same procedure and technique were applied to all four streaming platform datasets. To avoid repetition, the Netflix dataset was used as the main illustrative example of the solution's development in this chapter. By presenting a comprehensive clarification of the implemented techniques, the author aimed to provide a clear understanding of the approach and steps taken.

DATA CLEANING, EDA AND FEATURE SELECTION

The first phase of this solution had its focus on an initial analysis of the extracted csv files in order to access any discrepancies, such as missing values, outliers, incorrect data types, or any other inconsistencies as shown in Figure 16. To do so, well-known Python libraries were used, such as Pandas, and NumPy. These libraries provided powerful tools for data exploration and analysis, enabling to effectively identify and address any possible issues with the data.

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	Platform
total_missing	0.0	0.0	0.0	2634.00	825.00	831.00	10.00	0.0	4.00	3.00	0.0	0.0	0.0
%_missing	0.0	0.0	0.0	29.91	9.37	9.44	0.11	0.0	0.05	0.03	0.0	0.0	0.0

Figure 16 – Example of missing values analysis on Netflix dataset.
Source: Prepared by the author.

Moreover, an exploratory analysis was conducted in order to gain a comprehensive overview of the data distribution, identify potential trends, and explore relationships or correlations between the data attributes such as patterns, or outliers as presented in Figure 17. For that purpose, popular Python visualization libraries were used, such as Seaborn and Matplotlib.

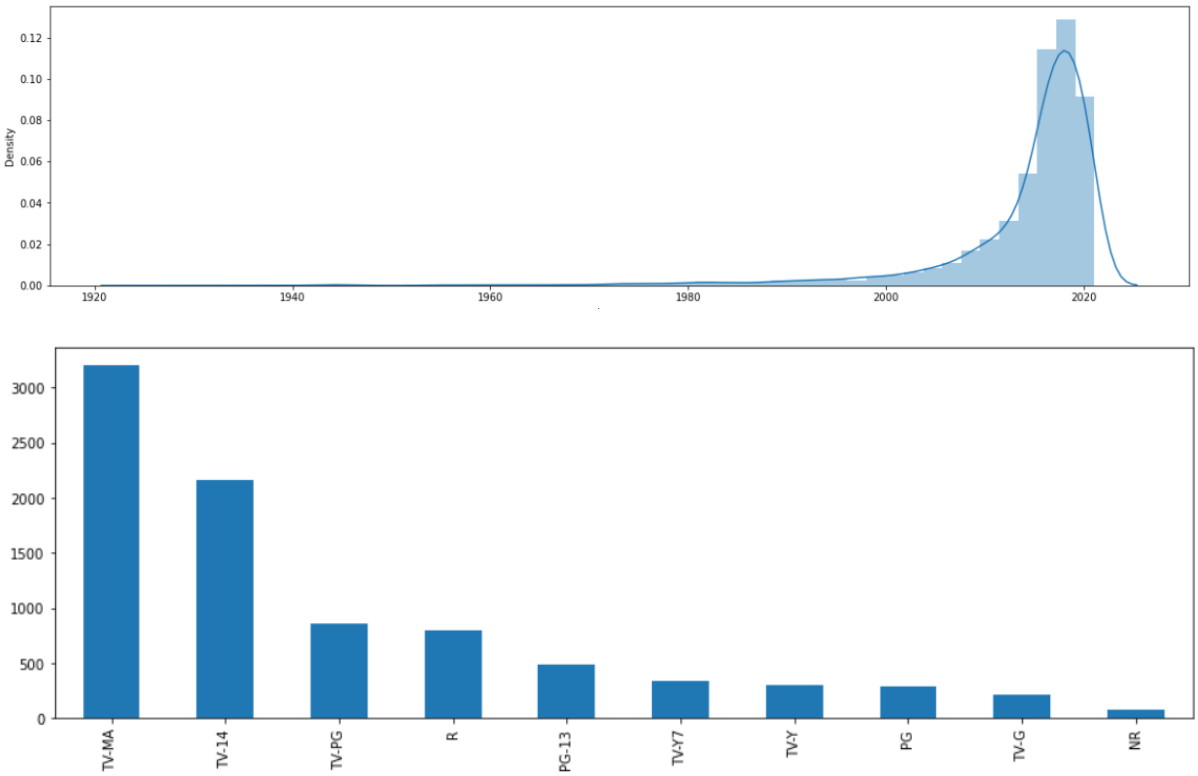

Figure 17 – Example of movie date release and rating distribution on Netflix dataset.
Source: Prepared by the author.

After this analysis, a preprocessing and feature selection was added in order to optimize the data to fit the machine learning model. In detail, a data type transformation was performed to change any attribute into object data type across all the datasets used in this project.

Moreover, a feature selection process was carried out to remove any attribute that was not relevant for the recommendation model used, which it will be explained later. For this matter and as explained in the Theoretical Framework section, all the attributes that do not fit the purpose for Word Embedding model were not considered, such as Director, Cast, Country, Release Year, Rating, Duration, Added Year, and Content Type.

Lastly, to not feed unnecessary words from the Description parameter to the model, those were identified and removed from the main attribute column in order to fit the algorithm with relevant keywords only. After all the preprocessing tasks, the four datasets were ready to feed the recommendation model with the following attributes: Show ID, Title, Movie Category, Description, and Platform.

PREDICTIVE MODEL: GOOGLE'S WORD2VEC

According to the filtering techniques explained during the Theoretical Framework section, several content-based filtering algorithms were analysed and tested in order to generate a tailored-made recommendation list.

To perform an adequate recommendation of video content based on similar movies, it was important to take into consideration textual information besides the category of the movie or TV series. As mentioned before in the literature review, Nature Language Process is an extremely important field in Artificial Intelligence, which enables computing machines to understand and process human language in a similar way as humans do. This field of Machine Learning includes statistical analysis and tasks, such as speech recognition, language translation, sentiment analysis, text classification, named entity recognition, text generation, and much more.

Word Embedding techniques have revolutionized NLP tasks, enabling models to understand and process natural language by representing text analysis in the form of real-valued vectors. In fact, this technique can solve multiple problems and issues since it considers the relationships and similarities between words unlike other more traditional NLP techniques. It is believed to be one of the most significant advances in Deep Learning for addressing complex natural language processing problems. Word Embedding technique is used to transform words into dense vector representations which capture the semantic and taxonomy meaning significance and contextual information of each word. Through a Neural Network approach, Word Embedding can learn throughout the time and enable to measure the similarity or dissimilarity between words.

Therefore, a well-known word embedding model was used in this project, the Word2Vec algorithm developed by Google. This algorithm is highly effective due its ability to understand similarities between words based on their occurrences in the dataset. For this reason, Word2Vec allows it to effectively predict the meaning of each word and establish associations with other words in the corpus of the text. There are two versions at the root of Word2Vec's model success: the Skip Gram model and the Continuous Bag of Words.

Gensim is a popular open-source Python library dedicated to NLP tasks which includes the pre-trained Word2Vec algorithm, while both models are 1-hidden-layer neural network, SG uses pairs of words generated by sliding a window across text data and trains the network to predict a probability distribution of nearby words given an input word. To normalize it, it uses a one-hot encoding to preprocess categorical data projected into a hidden layer, resulting in 300-dimensional word embeddings for a network with 300 neurons in the hidden layer. On the other hand, CBOW uses as input the context words within a fixed window and tries to predict the centre word. The projection weights that transform one-hot words into averageable vectors serve as the word embeddings.

As exemplified in Figure 18, when considered words such as "King" and "Queen", it is possible to observe a high significant similarity. The Word2Vec algorithm has the ability to perform algebraic operations on word embeddings. Using a famous example, subtracting the embedding vector of "male" (4, 2) from the embedding vector of "king" (7, 3) and adding the embedding vector of "female" (5, 1), results in a vector that closely approximates the embedding vector of "queen" (9, 2) as the final outcome.

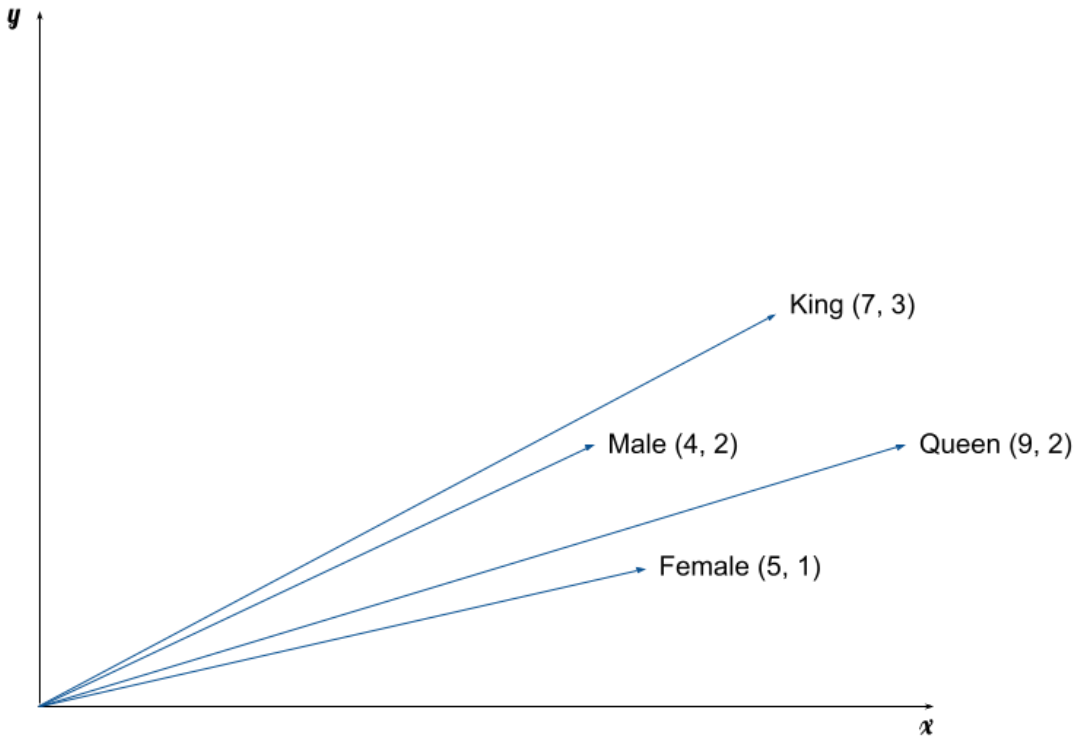

Figure 18 - Example of Google's Word2Vec algorithm performance, using the embedding vectors.
Source: Prepared by the author.

PREDICTIVE MODEL: IMPLEMENTATION

Concerning the implementation of the algorithm in this project, a few processing steps were added in order to transform the data to fit the model.

To start, the standardization process was carried out including multiple steps. The first one aimed to convert every word to lowercase all the text from the data attributes. To do so, from the Pandas library, the .str function and .lower() were used. The next step included breaking down every sequence or string of text into individual words or tokens, using a well-known NLP Python library named Natural Language Toolkit (NLTK) with the function .word_tokenize(). The further step aimed to convert all the text information from multiple attributes into a more suitable range of words for further analysis and to fit the pre-trained Word2Vec algorithm.

Therefore, the author created a list of common stopwords and punctuation in English, such as “the”, “and”, “I”, and other articles and prepositions. Those words are very frequent in common texts, however, do not add any relevant insight or significant meaning to the model. In this step, the author has removed from the dataset all the stopwords and punctuation included in the list. Once again NLTK Python library was used along with function .corpus and the module .stopwords(), which contains a collection of commonly used stopwords in English.

Last step of this process involved removing all the duplicate words of the analysed columns. For that, the Pandas function .set() was used in order to convert the list of words into a set of unique elements, automatically removing any duplicate words. The final relevant title list was presented in Figure 19.

	show_id	title	listed_in	description	Platform	title_list
0	s1	Dick Johnson Is Dead	[documentaries]	[father, death, ways, face, inevitable, life, ...	Netflix	[dick, johnson, dead]
1	s2	Blood & Water	[international, dramas, tv, mysteries, shows]	[sister, cape, sets, prove, swimming, star, pa...	Netflix	[water, blood]
2	s3	Ganglands	[international, crime, action, tv, shows, adve...	[family, expert, mehdi, team, robbers, lord, v...	Netflix	[ganglands]
3	s4	Jailbirds New Orleans	[tv, docuseries, reality]	[talk, toilet, gritty, justice, flirtations, s...	Netflix	[jailbirds, orleans, new]
4	s5	Kota Factory	[comedies, international, tv, romantic, shows]	[campus, train, unexceptional, india, centers,...	Netflix	[factory, kota]
5	s6	Midnight Mass	[mysteries, horror, dramas, tv]	[young, believe, arrival, ominous, mysteries, ...	Netflix	[mass, midnight]
6	s7	My Little Pony: A New Generation	[family, children, movies]	[hero, prove, equestria, earth, unicorns, pals...	Netflix	[little, pony, new, generation]
7	s8	Sankofa	[independent, international, dramas, movies]	[ancestral, model, ghana, agony, past, witness...	Netflix	[sankofa]
8	s9	The Great British Baking Show	[british, reality, shows, tv]	[whipping, uk, face, 10week, bakers, dishes, b...	Netflix	[show, british, baking, great]
9	s10	The Starling	[comedies, dramas]	[feisty, adjusting, woman, find, way, —, life,...	Netflix	[starling]
10	s11	Vendetta: Truth, Lies and The Mafia	[international, crime, tv, docuseries, shows]	[antimafia, organized, crime, sicily, trying, ...	Netflix	[vendetta, mafia, truth, lies]
11	s12	Bangkok Breaking	[international, crime, action, tv, shows, adve...	[earn, man, joins, rescue, living, conspiracy,...	Netflix	[breaking, bangkok]
12	s13	Je Suis Karl	[international, dramas, movies]	[family, young, bombing, woman, unknowingly, j...	Netflix	[suis, karl, je]
13	s14	Confessions of an Invisible Girl	[family, children, comedies, movies]	[joins, ideas, clever, queen, fit, bee, anythi...	Netflix	[invisible, confessions, girl]
14	s15	Crime Stories: India Detectives	[british, crime, tv, docuseries, shows]	[complex, four, challenging, investigations, i...	Netflix	[crime, india, detectives, stories]

Figure 19 - Example of the output of the Netflix dataset after the preprocessing process. Source: Prepared by the author.

PREDICTIVE MODEL: RECOMMENDATION OUTCOME

Once each dataset was converted into a set of words, a pre-trained collection of vectors trained on Google News dataset of around 100 billion words was loaded. Therefore, “api.load('word2vec-google-news-300')” module and function from the popular NLP library named Gensim was used, using Skip Gram model as default version.

By adding a model which contained 300-dimensional vectors for 3 million words and phrases already trained, the project benefited and leveraged its knowledge to the highest level. Both semantic and syntactic relationships between words captured and learned through this

model. Ultimately, these vectors learned meaningful representations of words based on their co-occurrence patterns and distances.

As the objective was to retrieve a list of recommendations based on the text information of each movie or TV series attributes in the datasets, a smaller amount of word trained vectors was needed in order to ensure that only matching valid words were retained for further modelling.

Therefore, in the next step, the author has filtered out from the collection of the trained vectors all the words that were not presented in the platforms' datasets. This filtering step helped to improve the quality and accuracy of the recommendation by focusing on relevant words that had a known vector representation in the pre-trained word vectors. This was achieved by iterating over each element in the DataFrame and keeping only the words that were present in the pre-trained word2vec model's vocabulary, using wv.key_to_index() function and method.

Moving forward, this next phase aimed to generate a function which iterated through the entire streaming platform dataset resulting in a recommendation list of 10 movies or TV shows for each title. This recommendation was based on the similarity to the other content, according on their description and category.

The function initialized with an empty list called recommendations to store the recommendation results. It iterated over each title in the dataset using the .tqdm() function. Within the loop, the function created a list called title_vocab to store the rows that have the same title as the current iteration. It iterated over the subset of the DataFrame where the title matches the current iteration and filters out words that were not present in the pre-trained vocabulary of the word2vec model using once again the .wv.key_to_index.() function. The filtered lines were then added to the title_vocab list. After, if the title_vocab list was empty, indicating that there are no lines with the same title, the cycle would move to the next iteration.

Afterwards, the function initialized an empty list called similarity_scores which stored the similarity results between the searched title and other video contents in the streaming platform data frames. It iterated over each row in the original dataset and calculated the similarity scores between each two sets of words represented by their word vectors using the .wv.n_similarity() function from the pre-trained word2vec model. The similarity scores were appended to the initial list if the category score was above the defined threshold, which in this case was defined to be 0,85.

Subsequently, from the similarity_scores list calculated previously, another data frame was created named similarity_df with columns as Title, Recommended title, Category score, and Description score. Then, the data frame was sorted in descending order based on the description score, category score and only the top 10 recommendations were selected.

Next, the similarity_df was appended to the empty list of recommendations. Once all iterations were completed, the recommendations were concatenated into a final data frame called df_netflix_recommendations and sorted by the title, category score, and description score. Finally, the function returned the data frame df_netflix_recommendations as the final result.

Finally, the outcome provided recommendations for each title in the original streaming platform dataset based on the similarity of the associated words. In Figure 20, the detailed code of the recommendation function described previously was provided.

```
def recommendation(df_netflix):
    recommendations = []
    for title in tqdm(df_netflix['title'].unique()):
        title_vocab = []
        for row in df_netflix[df_netflix['title'] == title].to_numpy():
            row[2] = [word for word in row[2] if word in wv.key_to_index]
            row[3] = [word for word in row[3] if word in wv.key_to_index]
            row[4] = [word for word in row[4] if word in wv.key_to_index]
            title_vocab.append(row)

        if not title_vocab:
            continue

        similarity_scores = []
        for row in df_netflix.to_numpy():
            if row[1] != title:
                category_score = wv.n_similarity(row[2], title_vocab[0][2])
                description_score = wv.n_similarity(row[3], title_vocab[0][3])
                if category_score > 0.85:
                    similarity_scores.append([title, row[1], category_score, description_score])

        similarity_df = pd.DataFrame(similarity_scores, columns=['Title', 'Recommendation', 'score_category', 'score_description'
        similarity_df = similarity_df.sort_values(by=['score_description', 'score_category'], ascending=False).head(10)

        recommendations.append(similarity_df)

    df_netflix_recommendations = pd.concat(recommendations, ignore_index=True)
    df_netflix_recommendations.sort_values(by=['Title',
                                               'score_category',
                                               'score_description'],ascending=[True, False, False], inplace=True)
    return df_netflix_recommendations
```

Figure 20 - Recommendation function code applied to Netflix dataset.
Source: Prepared by the author.

Performing an evaluation of the effectiveness of Word2Vec algorithm presented a challenge to any recommendation system, since it is an unsupervised learning model. An assessing process for this model depended on the specific use-case of the application. Therefore, an online evaluation process was planned and explained later in this research.

4.6. CREATION OF DATA VISUALIZATION & REPORTING SOLUTION

To build a successful and dynamic Business Intelligence solution, a powerful visualization tool was used. As mentioned before in the theoretical framework and initial part of the development section, Microsoft Power BI gathers all the conditions to hold the Business Intelligence report planned in this research. This specific tool displays great and attractive

domestic charts and graphics which can be used to effectively visualize data in such a way that boosts insights comprehension and knowledge retention.

The design and graphics used in a report represented a fundamental part of this Business Intelligence solution since it allowed the users to connect with the data and interact with the information provided in an effective and dynamic manner.

Once again, a detailed description of all steps and procedures was presented in this section. Considering the similar attributes inherent to the four datasets, identical methodologies and logic were applied consistently. For simplification, only the Netflix dataset was used as an illustrative example in the present section.

After computing the recommendation content for the four streaming platforms, the first step to create a specific BI solution was to connect all the data files into Microsoft Power BI. Both four recommendation lists and four main datasets were stored in comma-separated values file, therefore the domestic Text/CSV connector was used in order to connect the eight csv files.

Subsequently, based on the user needs as described in the literature review section, a second ETL process was conducted to the main datasets and to its recommendation files in order to prepare the data model.

Several steps were applied to the main datasets as shown in Figure 21:

- Several data types were changed.
- The Date_added column was removed due to its irrelevance to this solution.
- The repetitive standard Show_id code was replaced to a more suitable and logical one. To establish a logical connection to the dataset, the first letter of the database name was added to the index number. In this Netflix example, the Show_id key was changed from s100 to n100.
- The Show_id column was renamed to Primary key column.
- The duration column was split in two columns Duration (min) and Seasons.
- The Age Rating column was added in order to have a better overview of maturity audience target content according to Netflix Help Centre.
- Several names from title column were cleaned, such as the ones which contained special characters and irrelevant information as (Sub) or (Dub).
- Followed the logic that the first element mentioned in the columns Category, Country, Director, and Cast was the most relevant, additional columns with the unique element were generated named “Main + (Previous column name)”.

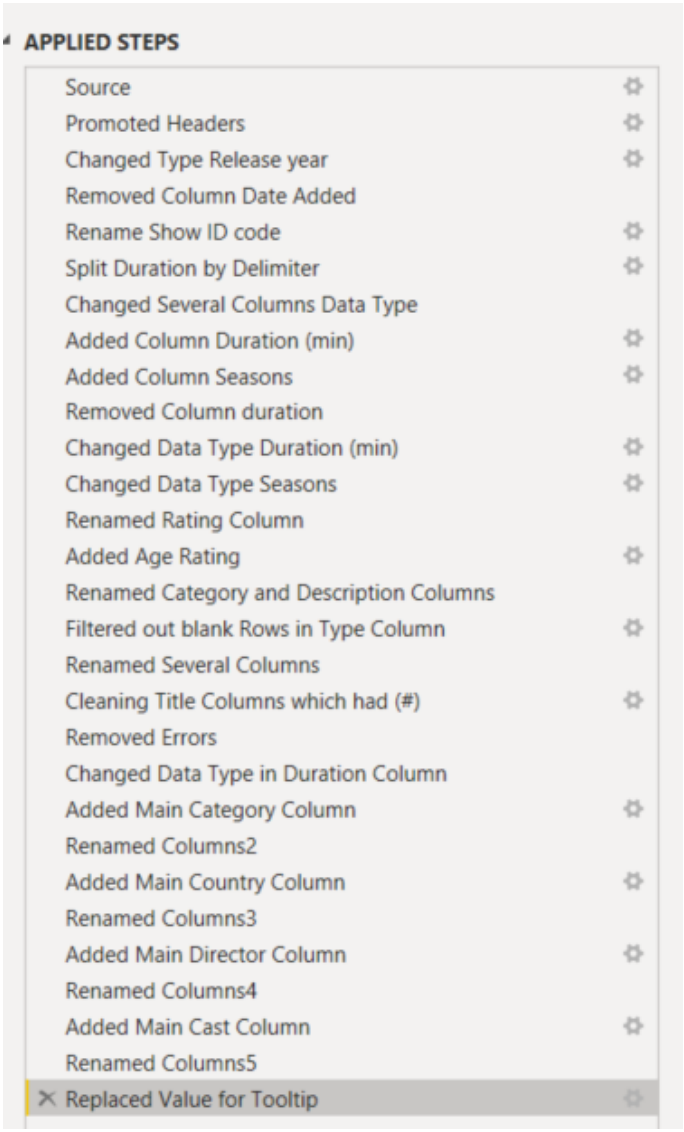

Figure 21 - Power Query Editor steps for data processing applied to Netflix dataset.
Source: Prepared by the author.

As presented in Figure 22, the steps applied to the recommendation tables were:

- Several data types were changed, including to Score_category and Score_description.
- Names which contained special characters from Title column ware cleaned.
- The columns Show ID, Release Year, Platform, Duration (min), Seasons which have numeric data types from main Netflix dataset were merged to the recommendation table, based on the Title column.
- The Title column was removed.

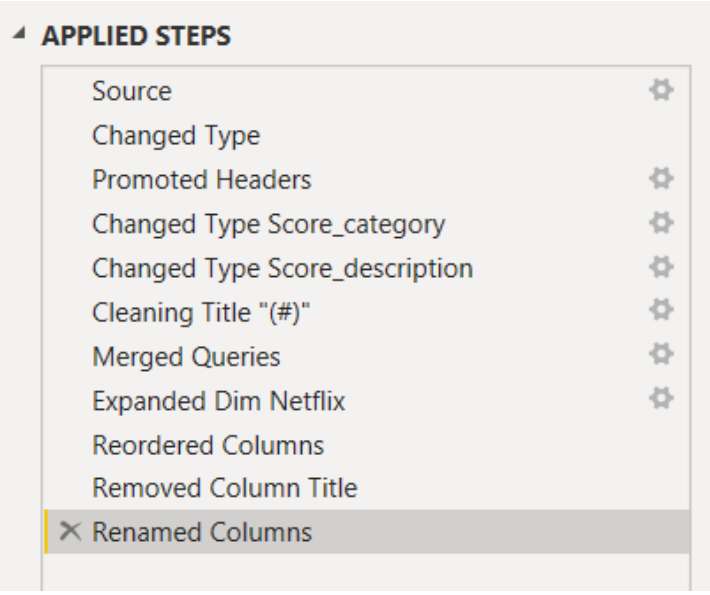

Figure 22 - Power Query Editor steps for pre-processing Recommendation Netflix file.
Source: Prepared by the author.

4.6.1. Data Modelling – Star Schema

To create a self-service and user-friendly model suggested by Ralph Kimball and described in the theoretical framework section, relevant changes were performed to the data in order to create the well-known star schema data model.

As explained previously, this type of schema consists of one large fact table and various smaller dimension tables. Those are connected through primary and foreign keys relationships which radiate around the central table. In order to build one in this BI solution, several changes to the data took place.

To start and as mentioned in the steps applied, all numeric columns from the main datasets were merged to the recommendation tables in order to isolate all the categorical attributes in the datasets table. After this step, the author confirmed that the dataset tables contained only unique foreign keys represented in the Show_id column. This assured the “one” part of the one-to-many relationship between Dimensional tables and future Fact table, through Show_id as unique foreign key. Afterwards, all four dataset tables were renamed to Dim + Streaming Platform Name, such as Dim Netflix, Dim Amazon, Dim Disney and Dim Hulu.

On the other hand, by merging all numeric columns into the recommendation table, the author aimed to join all the performance measurements into one single table. After applying the mentioned steps to the four platforms recommendation files, additional steps were taken in order to create the central table of this BI solution, also known as Fact table, as shown in Figure 23.

Detail description of all the steps involved to create the Fact table:

- New empty table created.
- Manually entered one single column which included the platform names.
- Merged this column with the four recommendation tables on the Platform column.
- The following columns Show ID, Recommend Titles, Score_category, Score_description, Release Year, Duration (min), and Seasons were merged and extracted from the four recommendation tables. The outcome was a deep table with all the data of the four streaming platforms.
- Removed the Platform column.
- Renamed the Show ID column to Primary Key column.
- Finally, renamed the column name to Fact Platform.

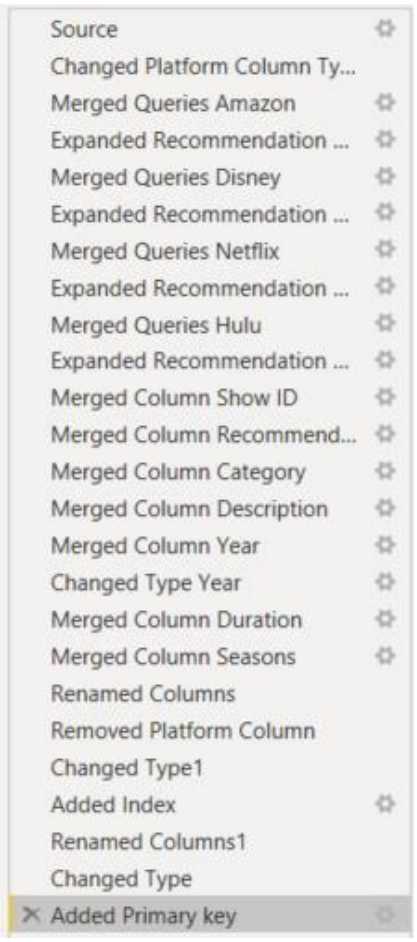

Figure 23 - Power Query Editor steps for data processing applied to Fact Platform table.
Source: Prepared by the author.

After the described steps were made to guarantee the efficiency and optimization of this BI solution, the initial four recommendation tables were disabled to load.

Further, a Calendar Table was created. Typically, this kind of table is fundamental in a star schema, since it allows the user to explore the report with more flexibility and if need, to be used for time intelligence functions. This one was created in DAX using the function CALENDARAUTO(), which automatically returns a table with a single column named Date with a range of dates based on existent date information in the dataset.

Subsequently, to ensure data integrity and consistency, a one-to-many relationship was settled between the fact table and each one of the remaining tables through the common attribute Show ID, later renamed to Foreign Key and Primary Key.

Lastly, the final data model consists of four Dimensional tables, Dim Netflix, Dim Amazon, Dim Disney, Dim Hulu, one Calendar table and one Fact table, Fact Platform, as shown in Figure 24.

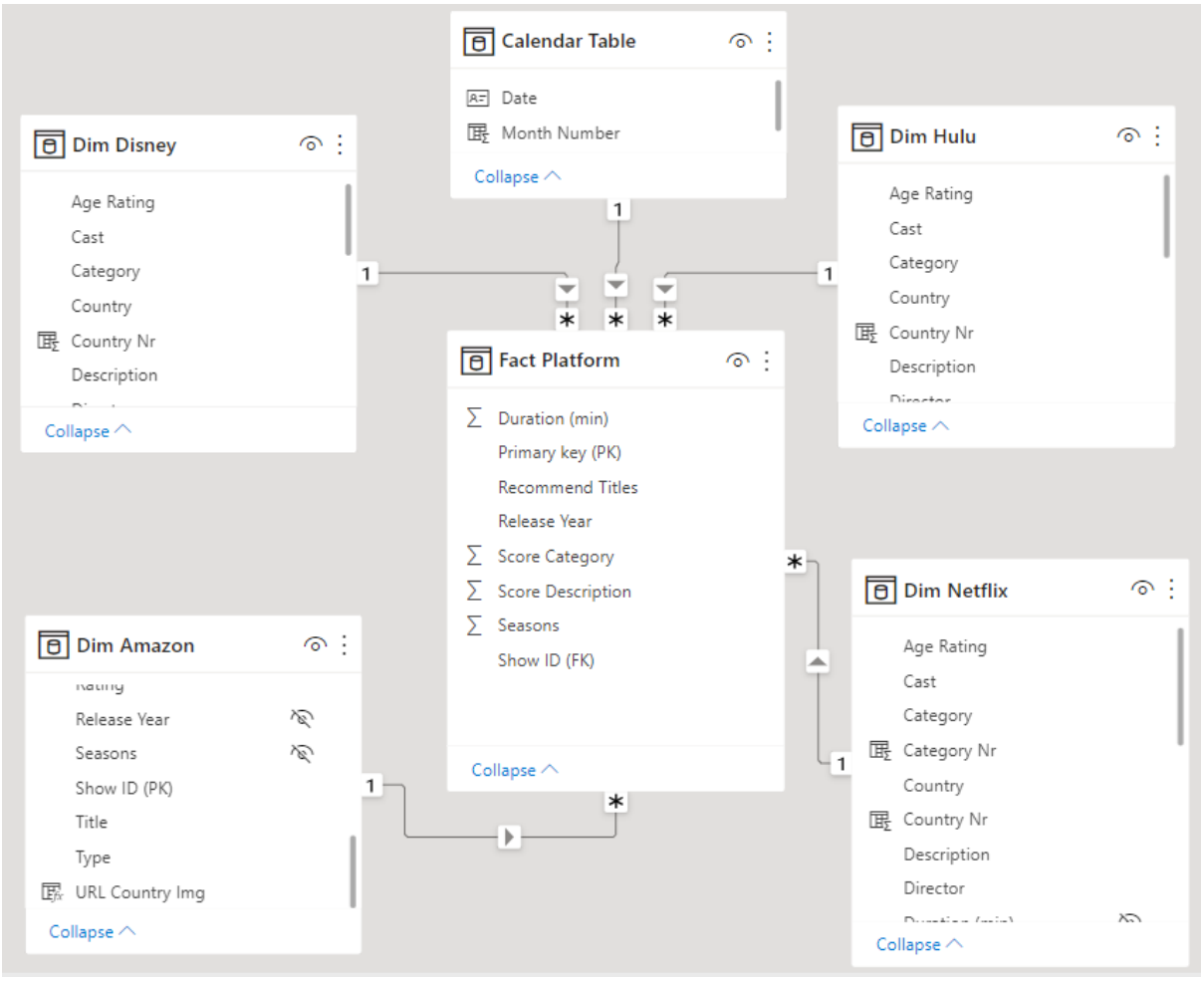

Figure 24 - Star Schema Model from the fours Streaming Platforms on Power BI.
Source: Prepared by the author.

4.6.2. Data Visualization

Data visualization is a powerful and meaningful tool to comprehend and communicate insights in a clear and effective way. In this Business Intelligence solution, the author aimed to display each platform's complex data in dynamic and understandable reports.

To successfully present this dashboard, as presented in the introduction section, an initial exploration of the data storytelling from each platform's report was visualized and sketched. Subsequently, extensive research into various data chart types was carried out in order to comprehend proper circumstances and methods to employ each distinct chart.

In the following section, a detailed description of all used charts was described in detail and structured in Platform report and Recommendation report.

4.6.3. Platform Report

This BI solution started with an initial page which represents the Home Page of this report with four navigation bookmarks that would direct the user to each one platform report, as illustrated in Figure 25.

Figure 25 – Home page of the BI solution on Microsoft Power Bi.
Source: Prepared by the author.

Once clicked in one of the platform's icons, the platform report showcasing a significant portion of the platform's principal data was loaded. This report provided a comprehensive

overview and valuable insights regarding the content and diversity of information available on each platform. In Table 6, a comprehensive description of each chart and graph was provided.

Table 6 – Description of each visualization on Platforms' report.

Technique / Chart	Description
Canvas Background	To optimize the BI solution and avoid the loading of additional assets, a background image was designed in Microsoft PowerPoint to fill the canvas scenario and fit to the page size.
Slice filters	In order to filter dynamically this dashboard, four dropdown slicers were added. These slicers allow the dynamic and flexible use of this dashboard. Each filter allowed single selection only and permitted filtering by Country, Category and Type of content.
Navigation Buttons	As mentioned before, a background image was added to the BI solution, in which the icons of the three other streaming platforms were already incorporated in the asset. In order to transform these into navigation buttons, three transparent text boxes were added on top of the icons with the option page navigation action activated.
Single Card	Five KPIs cards were displayed to provide better and meaningful insights regarding the data from each platform. Specifically, the sum of total video content available in each platform data was added, along with the sum of main categories and countries. The average duration in minutes and average number of seasons of each movie or TV show were also added.
Bar-chart	To perform a metric comparison across different subgroups in Age Rating attribute, the popular bar-chart visualization was used in order to provide clear information on how each subgroup compare against the others.
Bubble Chart	To present the directors who took the leading role in the production of the platforms' content, a Packed Bubble chart was employed. The AppSource custom visual created by xViz LLC was downloaded to highlight the overall number of Movies and TV Shows primarily directed by these individuals.
Line-chart	Moreover, a time-series analysis was also provided using a line-chart visualization with the Type of content and Release date as parameters. Through this graph, it was possible to conclude that Movies have always been more prominent and superior then TV Shows, in terms of the quantity of content released. This fact was confirmed across the four platforms, however a slight change in this trend was registered after 2010, this category of content suffered a decrease.
Map Bubble Chart	The four streaming platforms offered a wide range of film content across the globe, including diverse geolocations. Considering the first country mentioned in each content as their prime location, a bubble

	map-chart was used to visualize the amount of content distributed by country.
Tooltip	In order to include additional information to the map chart, the powerful and customized tooltip technique was used. Specifically, by hovering on top of each location, additional data regarding the total number of contents, total average of categories, flag and name of each country were displayed. In particular, the flag images were used based on the name and URL provided through an online connection, as shown in Figure 26.
Pie-chart	Furthermore, as mentioned before, to compute the prime location of the Country attribute, the same logic was applied to retrieve the main Category of each content. Afterwards, a pie-chart was considered in order to visualize the 10 most relevant category's types of each platform.
Multicard	The primary objective of this visual was to showcase a list of the 10 most frequently mentioned actors and/or actresses on each platform. To achieve this, a multi-row card was used, exhibiting not only the persons' names but also the total number of films participated.
Bookmark	Finally, two bookmark assets were also prepared in this BI solution. One was added on top of the dashboard, a Recommendation button was provided with the important role of displaying the second part of this report. This part provides more insights and a list generated through the Recommendation System. The second was added through home button in order to provide the possibility to return to the Home page.

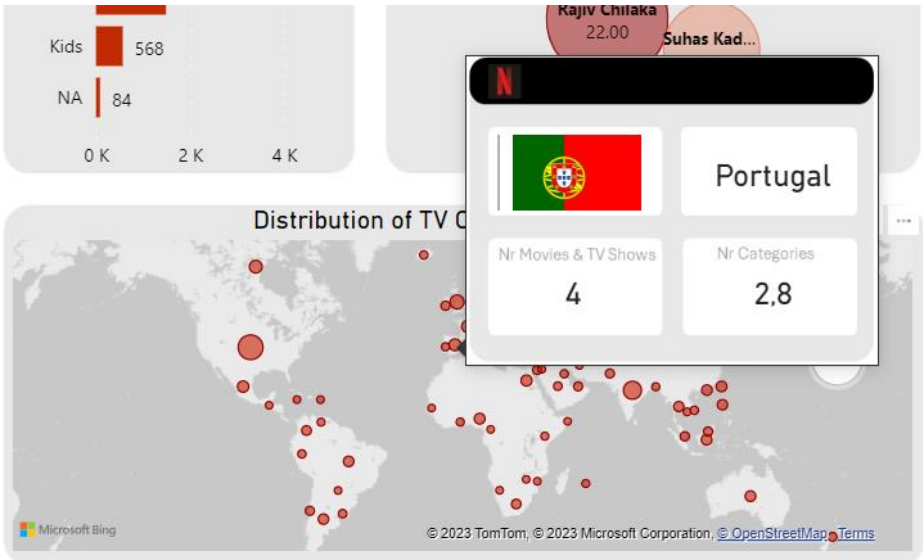

Figure 26 – Tooltip function on Map chart from Netflix's report on Microsoft Power Bi.
Source: Prepared by the author.

The final outcome of the Business Intelligence solution regarding general statistics from Netflix dataset was presented in Figure 27 and the remaining tabs of this solution were displayed in the Appendix A section.

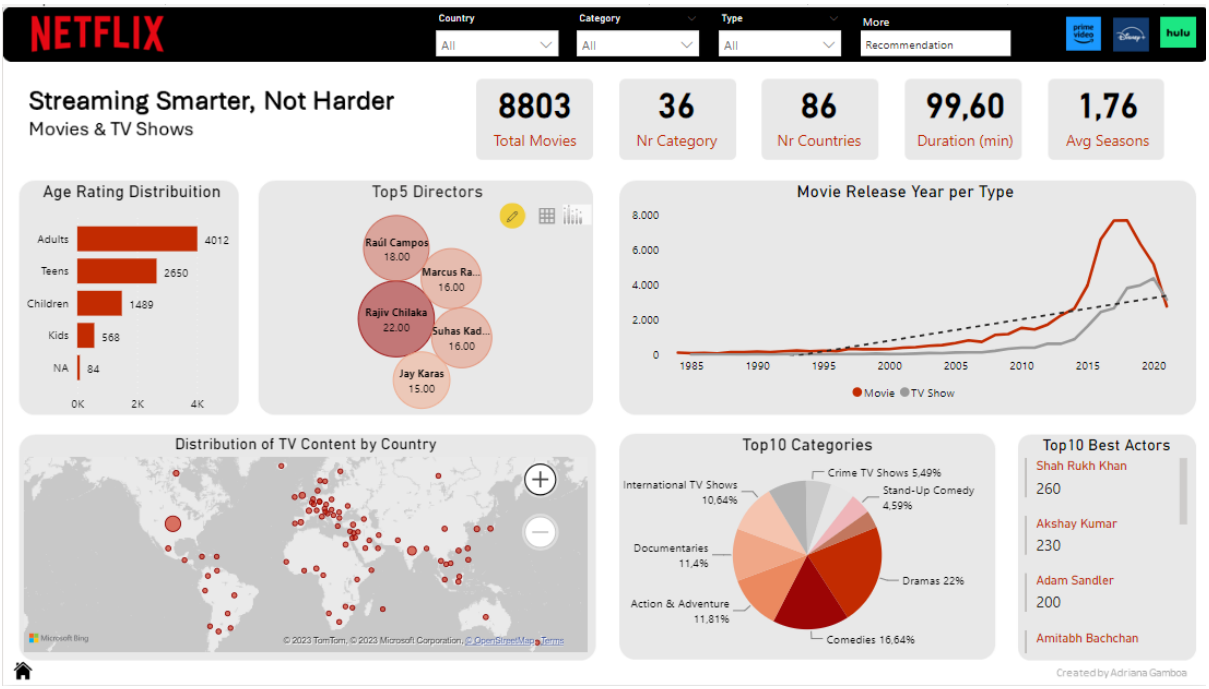

Figure 27 – Final outcome from Netflix's dashboard on Microsoft Power Bi.
Source: Prepared by the author.

4.6.4. Recommendation Report

As described, a second part of the platform's dashboard included a comprehend analysis of the Recommendation System outcome. In Table 7, a description of all the charts and techniques performed in this section of the report was provided.

Table 7 – Description of each visualization on Recommendations' report.

Technique / Chart	Description
Slice filters	On the contrary of the previous section, the Recommendation report focused exclusively on the analysis of the recommendations per Movie or TV Show title. Therefore, only one relevant slice was added to this view, the Title filter.
Single Card	To have a full comprehension of the recommendation list, four KPIs cards were presented to provide more information regarding the Movie or TV Show title chosen. In particular, the main category of the film content, the age rating provided, the main country where it was produced, and finally the main director.

Word Cloud chart	Following the same framework, a very interesting AppSource custom chart was added to highlight important keywords of the Description of the title chosen.
Table	As described before, the Recommendation System created a list of recommendations for each item and made it possible to suggest related items to the same user or to a similar one, based on certain characteristics. The outcome of RS was finally presented through a table with the recommended Movies or TV Shows, ranked whether by the description or category score.
Gauge-chart	For last, to present the only two metrics that were able to rank the list of recommended titles, two Gauge-charts were added to this report. This chart provided a very intuitive comprehension and quick access to the current state of the metric against the target, which in this particular case was 1.
Bookmark	Once again, a bookmark asset was used as a return button in order to allow a flexible and dynamic navigation between dataset and recommendation reports.

Lastly, the recommendation system dashboard from Netflix was displayed in Figure 28.

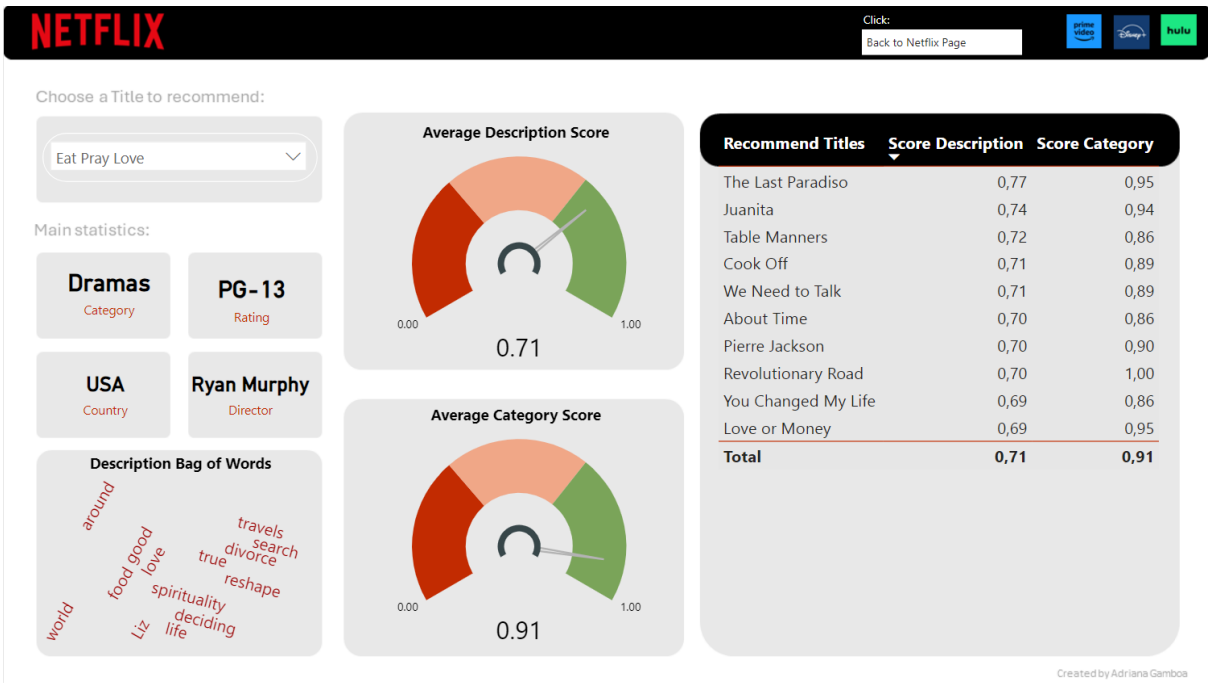

Figure 28 – Netflix Recommendation System's final report on Microsoft Power Bi.
Source: Prepared by the author.

5. Results and Discussion

Following the development of the comprehensive Business Intelligence solution involving recommendation systems of four global streaming platforms, evaluation of the results is an essential aspect to measure the performance and effectiveness of the recommendation.

On the contrary of other machine learning projects, recommendation systems do not split the data into test and training datasets in order to perform evaluation metrics to evaluate the predicted result using metrics such as F1-score, Precision or Recall. Nevertheless, other effective evaluation approaches are possible.

To start, a successful evaluation metric to assess the recommendation systems' efficiency could be done through User Engagement Metrics. These allow BI scientists to have access to interesting insights regarding how engaged the audience is with the digital content. Two great measures are CTR, which estimates the percentage of the clicked or interactions on the recommendation content by the user; and Conversion Rate, which calculates the percentage of recommendation content actual watch by the users or had any other relevant behaviour considered as a conversion action.

Another interesting evaluation metric is to compare the performance of different recommendations or models though A/B Testing approach. This technique involves dividing the audience in two random groups and recommend content based on different aspects. For example, recommend content to group A based on Description Score, while to group B, a recommendation would be based on Category Score.

Lastly and as the evaluation method applied to this project, a User Survey was made in order to collect valuable user feedback and assess interesting insights. This technique was extremely helpful to gather qualitative insights from the users and to understand the level of satisfaction with the recommendations. Therefore, a survey was created using Google Forms which collected a total of 42 answers as shown in Appendix B.

In this survey, four different questions were enquired to the users in order to test the recommendations from each platform. Specifically, the author asked the likelihood to the user of watching the recommendation content after watching one particular movie from 0 (not likely) to 10 (very likely) as possible responses. In Figure 29, a pie-chart visualization of the results was presented per platform, in which displayed first the vote of the user, and secondly the vote rating percentage.

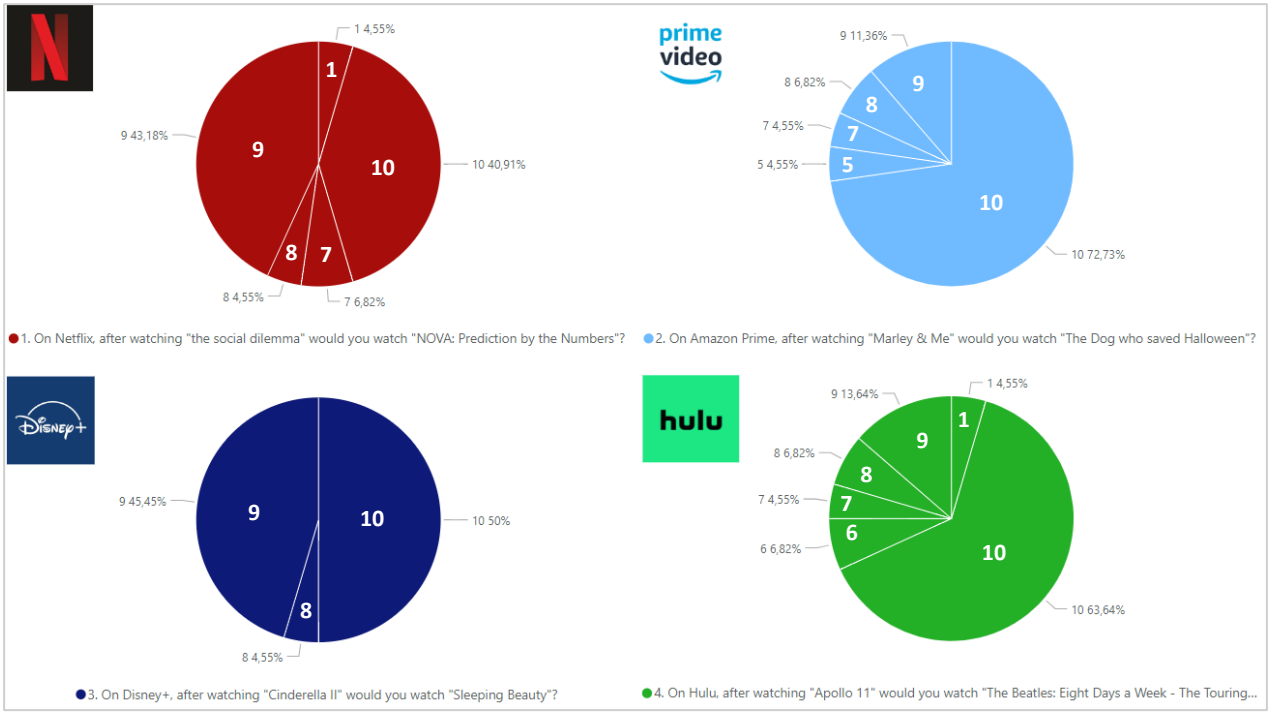

Figure 29 – Collected answers from Stream Smarter, Not Harder Survey.
Source: Prepared by the author.

As result, a Net Promoter Score (NPS) rating which assesses the customer satisfaction and engagement was computed. Out of these responses, two major outcomes were computed.

The first represented the division of the audience in three major cluster groups:

- **Detractors** - the individuals who voted between [0; 6]
- **Passives** - the users who voted in the middle of the range, between [7; 8]
- **Promoters** - the passionate users who evaluate between [9; 10] as shown in Figure 30.

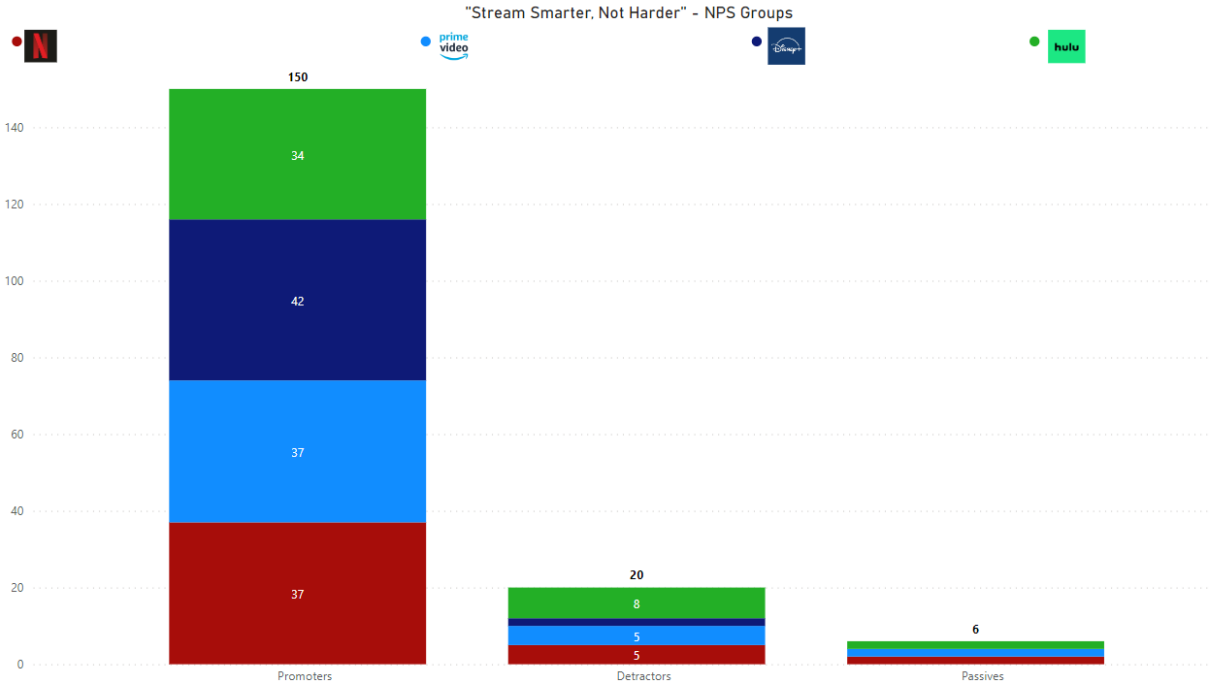

Figure 30 – Stream Smarter, Not Harder Survey - NPS Groups.
Source: Prepared by the author.

The second and most significant outcome was computed in single KPI for benchmarking within [-100; 100] interval. This benchmark served as an essential indicator of the user sentiment and can generate helpful insights to improve the recommendation system. According to the creators of the NPS metric, Bain & Company, which stated that a NPS score above 20 is “great” and above 50 is “amazing”. As displayed in Figure 31, the NPS score of this projected achieved an outstanding result of 73.86.

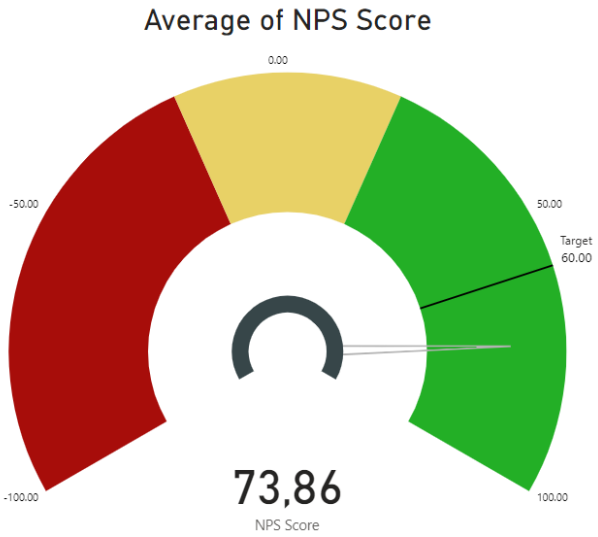

Figure 31 – Stream Smarter, Not Harder Survey - NPS Result.
Source: Prepared by the author.

Overall, the results obtained from the recommendation based on the Description and Category score were extremely satisfactory. On average, the recommendations generated through the model resulted in a cosine similarity above 0.7 on description score, 0.8 on category score and a NPS score above 70. These results proved the high probability of the user to choose any of the recommended content and ultimately, increase watching hours and user retention on each of the platforms.

6. Conclusion, Limitations and Future Work

6.1. Summary of the Developed Project

Following the conclusion of this comprehensive BI solution involving the main data and recommendation system for the four global streaming platforms, this section served the purpose of providing an understanding summary of the numerous phases of this project.

The primary objective of this project was to design and implement a sophisticated Business Intelligence solution which included a deep analysis of the general statistics of four streaming platforms. Moreover, a recommendation system was generated and added to the BI report. Overall, the project featured a diverse range of tools and programming languages, including Power BI, Python, M, and DAX.

The foundation of this study was based on the clear definition of objectives and strategies to successfully complete this dissertation. Subsequently, an extensive theoretical framework regarding Recommendation Systems and Business Intelligence was done. In this section, a deep investigation concerning the details of these two major topics occurred, including an exploration into several recommendation techniques, specifically Content-Based Filtering and areas of use. Additionally, deep research about Business Intelligence also took place, from data warehousing to its application in streaming platforms, allowing a comprehensive understanding of the several architectural models, including Kimball's main approach. Regarding both subjects, challenges and opportunities were discussed in order to get all the requirements and knowledge to conduct the success of this project.

Furthermore, the project's development was divided into a five-step approach. To start the practical phase of this study, a clear definition of the project goals, needs and assumptions was provided to align with industry's demands and expectations. Then, a comprehensive exploration and understanding of the platforms' data was conducted. Through this step, it was possible to identify key insights and trends. As mentioned before, a complete literature review was carried out in order to perform practical decisions based on scientific-proven knowledge involved in the project development. In the further step, the transformation from raw data into generated recommendations was computed. The author applied a Recommendation System algorithm, allowing each platform to enhance the user experience and engagement by improving the content suggested. Finally, it was possible to visualize meaningful insights from the raw data when a user-friendly interface solution was created in Power BI. This solution involved an extensive data architecture procedure to ensure the connection between four different platforms' data into a reachable six table star schema model, Dim Netflix, Dim Amazon, Dim Disney, Dim Hulu, one Calendar table and one Fact Platform.

To conclude, this particular industry will face continuous improvements and adaptations to answer to the fast-changing user preferences. Therefore, the insights and BI solution provided

through this study shall serve as proof of the potential and innovation of Business Intelligence and Recommendation Systems to improve industry's competitiveness.

6.2. CONTRIBUTIONS OF THE SOLUTION

This project contributed to analyse and provide a deep comprehension regarding the business needs in the mentioned streaming platforms. This Business Intelligence and Recommendation System solution was able to address the requirements and provide meaningful results.

Consequently, the final result of this project did not only achieve the goals proposed, but it will also contribute to improve the efficiency and effectiveness in an innovative entertainment industry.

6.3. LIMITATIONS

This project faced two notorious challenges which affected the development of sophisticated techniques and final outcome of this solution.

One barrier encountered was the lack of a large variety of numerical data. This important barrier made it impossible to develop sophisticated intelligent numerical metrics or forecasts, which could have provided more granular information of each streaming platform.

Moreover, it was important to acknowledge that there is no perfect recommendation technique in the field of recommender systems, as stated by Paul et al. (2019). Therefore, this study struggled along with the limitations and challenges inherent to the particular recommendation model used and its outcome.

6.4. FUTURE WORK

This project has the potential to evolve and increase its scoop with further additional future iterations. Firstly, it is highly probable that other data sources could be integrated into this recommendation system, such as additional numeric attributes which might complete the current datasets or additional data sources. As a result, this expansion requires further effort and development to adapt the existing data model architecture, ensuring its compatibility and effectiveness in facilitating the new incoming streams of data.

Future work should focus on a systematic analysis of the final recommendations in relation to actual user interactions and outcomes. It should use a range of statistical measures such as accuracy metrics in order to assess its performance over time.

Lastly, the search for a more efficient recommendation model persists. Continuous research efforts should prioritise methods that guarantee high-quality recommendations rapidly. Big Data technology has become a powerful ally in this matter, offering ways to explore scalable algorithms and efficient data processing techniques (Wang et al., 2018).

References

Ballard, C., Herreman, D., Schau, D., Bell, R. E., Kim, E., & Valencic, A. (1999). Data Modeling Techniques for Data Warehousing. International Technical Support Organization.

Bhatt, B., Patel, P., Gaudani, H. (2014). A review paper on Machine Learning based Recommendation System. International Journal of Engineering Development and Research. 2(4):3955-3961.

Bobadilla, J., Ortega, F., Hernando, A., & Bernal, J. (2012). A collaborative filtering approach to mitigate the new user cold start problem. Knowledge Based Systems, 26, 225–238.

Breslin, M. (2004). Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models. Business Intelligence Journal.

Chen, T., Liu, Y., & Huang, L. (2022). ImageGP: An easy-to-use data visualization web server for scientific researchers. iMeta, 1(1).

Cutting, Vineesh & Stephen, Nehemiah. (2021). A Review on using Python as a Preferred Programming Language for Beginners. 8. 4258-4263.

Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., & Kashef, R. (2020). Recommendation systems: algorithms, challenges, metrics, and business opportunities. Applied Sciences, 10(21), 7748.

Fouladirad, M., Neal, J., Vilaplana Ituarte, J., Alexander, J., & Ghareeb, A. (2015). Entertaining Data: Business Analytics and Netflix. International Journal of Data Analysis and Information Systems, 10(1).

Gartner (2023) Magic Quadrant for Analytics and Business Intelligence Platforms

Inmon, B. (2008). Building a Data Warehouse. Wiley, 4, 576.

Kimball R, Ross M. (1996). The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling.

Ko H, Lee SY, Park Y, Choi A. (2022). A survey of recommendation systems: recommendation models, techniques, and application fields. Electronics. 11(1):141. doi:10.3390/electronics11010141

Linstedt, D. (2002). Data Vault Series 1 – Data Vault Overview. The Data Administration Newsletter, LLC.

Maturity ratings for TV shows and movies on Netflix. (n.d.). Netflix Help Centre. Retrieved August 1, 2023, from https://help.netflix.com/en/node/2064

McAuley, J. (2021). Recommendation on Live-Streaming Platforms: Dynamic Availability and Repeat Consumption. Proceedings of the 15th ACM Conference on Recommender Systems.

Moody, D. L., & Kortink, M. a. R. (2000). From enterprise models to dimensional models: a methodology for data warehouse and data mart design. Design and Management of Data Warehouses, 5.

Negash, S. (2004). Business Intelligence. Communications of the Association for Information Systems, 13.

Ong, I. L., Siew, P. H., & Wong, S. F. (2011). A Five-Layered business intelligence architecture. Communications of the IBIMA, 1–11.

Pandya, S., Shah, J., Joshi, N., Ghayvat, H., Mukhopadhyay, S., Yap, M. (2016). A novel hybrid based recommendation system based on clustering and association mining. 10th International Conference on Sensing Technology (ICST), 1-6.

Parsons, P. (2022). Understanding data visualization design practice. IEEE Transactions on Visualization and Computer Graphics, 28(1), 665–675.

Paul D, Kundu S. (2019). A Survey of Music Recommendation Systems with a Proposed Music Recommendation System. Advances in Intelligent Systems and Computing. 279-285.

Peška, L., & Vojtáš, P. (2020). Off-line vs. On-line Evaluation of Recommender Systems in Small E-commerce. Faculty of Mathematics and Physics, Charles University.

Reddy, S., Nalluri, S., Kunisetti, S., Sharmila, A., Venkatesh, B. (2018) Content-Based Movie Recommendation System using genre correlation. In: Smart Innovation, Systems and Technologies, 2018:391-397.

Romero, C., Ortiz, J., Khalaf, O., & Prado, A. (2021). Business Intelligence: Business Evolution after Industry 4.0. Sustainability, 13(18), 10026.

Schuff, D., Corral, K., St Louis, R. D., & Schymik, G. (2016). Enabling self-service BI: A methodology and a case study for a model management warehouse. Information Systems Frontiers, 20(2), 275–288.

Shahbazi, Z., Byun, Y. (2020). Product Recommendation Based on Content-based Filtering Using XGBoost Classifier. International Journal of Advanced Science and Technology, 29(04), 6979 –6988.

Shani, G., & Gunawardana, A. (2010). Evaluating recommendation systems. In: Springer EBooks. 257-297.

Statistica (2023) Number of Netflix paid subscribers worldwide from 1st quarter 2013 to 2nd quarter 2023. Retrieved in July, 2023, from https://www.statista.com/statistics/250934/quarterly-number-of-netflix-streaming-subscribers-worldwide/

Statistica (2023) Number of Disney Plus subscribers worldwide from 1st quarter 2020 to 3rd quarter 2023. Retrieved in August, 2023, from https://www.statista.com/statistics/1095372/disney-plus-number-of-subscribers-us/

Sudhakaran, P. (2021). Research of Business Intelligence and Analytics Platforms.

Towards Data Science (2020) Power BI: M vs. DAX and Measures vs. Calculated Columns. Retrieved in August, 2023, from https://towardsdatascience.com/power-bi-m-vs-dax-vs-measures-4c77ae270790

Wang, Y., Wang, M., Xu, W. (2018). A Sentiment-Enhanced hybrid recommender system for movie recommendation: a big data analytics framework. Wireless Communications and Mobile Computing. 1-9.

What is Power BI. (n.d.). Microsoft Learn. Retrieved in August 1, 2023, from https://learn.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview

Xin, Y. (2015). Challenges in Recommender Systems: Scalability, Privacy, and Structured Recommendations. Massachusetts Institute of Technology.

Yessad, L., & Labiod, A. (2016). Comparative study of data warehouses modeling approaches: Inmon, Kimball and Data Vault. International Conference on System Reliability and Science (ICSRS).

Zhao, X., Kang, H., Feng, T., Meng, C., & Nie, Z. (2020). A Hybrid Model Based on LFM and BiGRU Toward Research Paper Recommendation. IEEE Access, 8, 188628-188640.

Appendix A

The following figures represented the additional dashboards regarding the main data and recommendation system from the remaining platforms in this project, including Amazon Prime, Disney+ and Hulu.

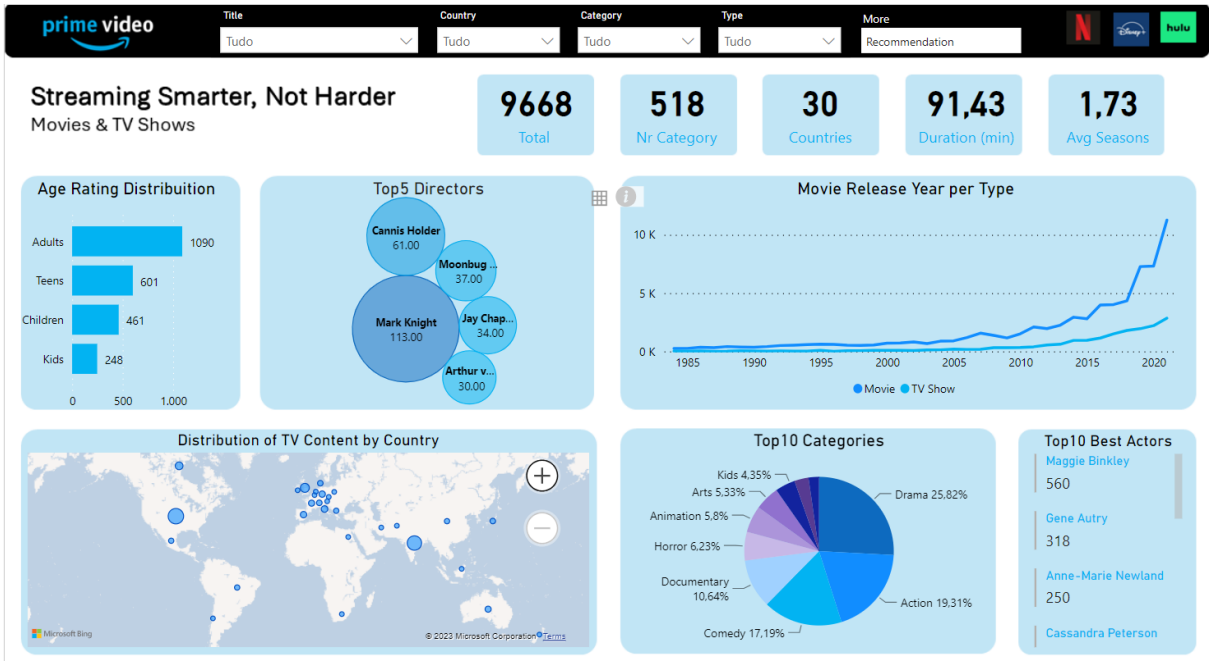

Figure a.1 – Final outcome from Amazon Prime dashboard on Microsoft Power Bi.

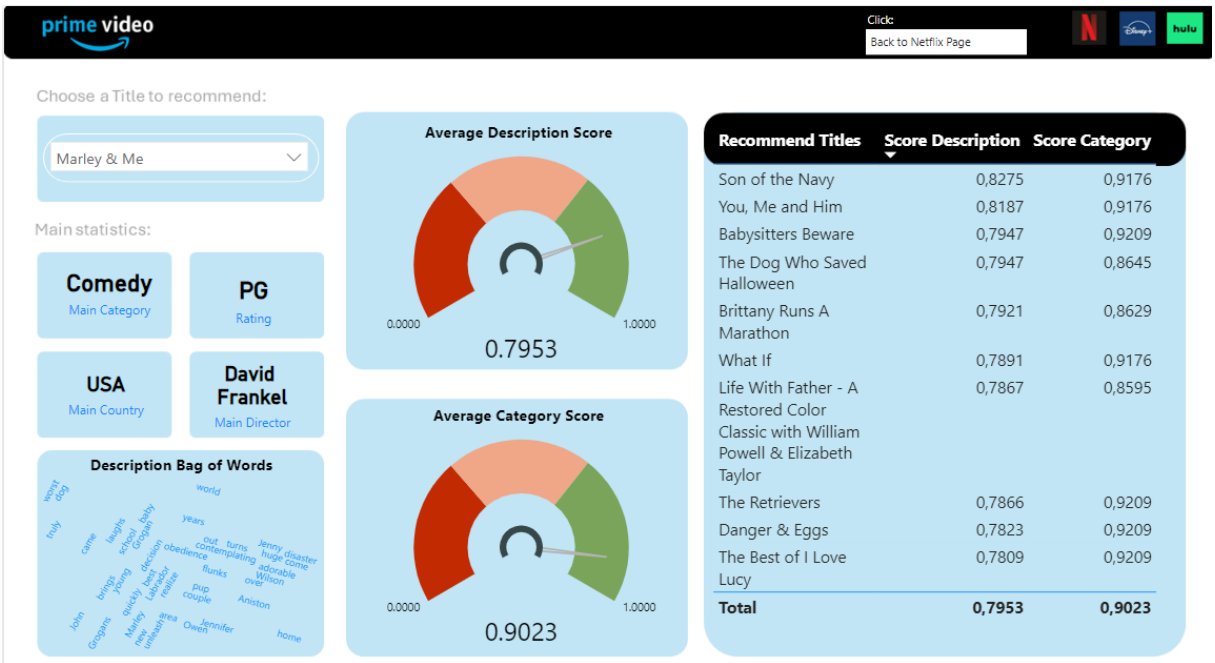

Figure a.2 – Amazon Prime Recommendation System's final report on Microsoft Power Bi.

Figure a.3 – Final outcome from Disney+ dashboard on Microsoft Power Bi.

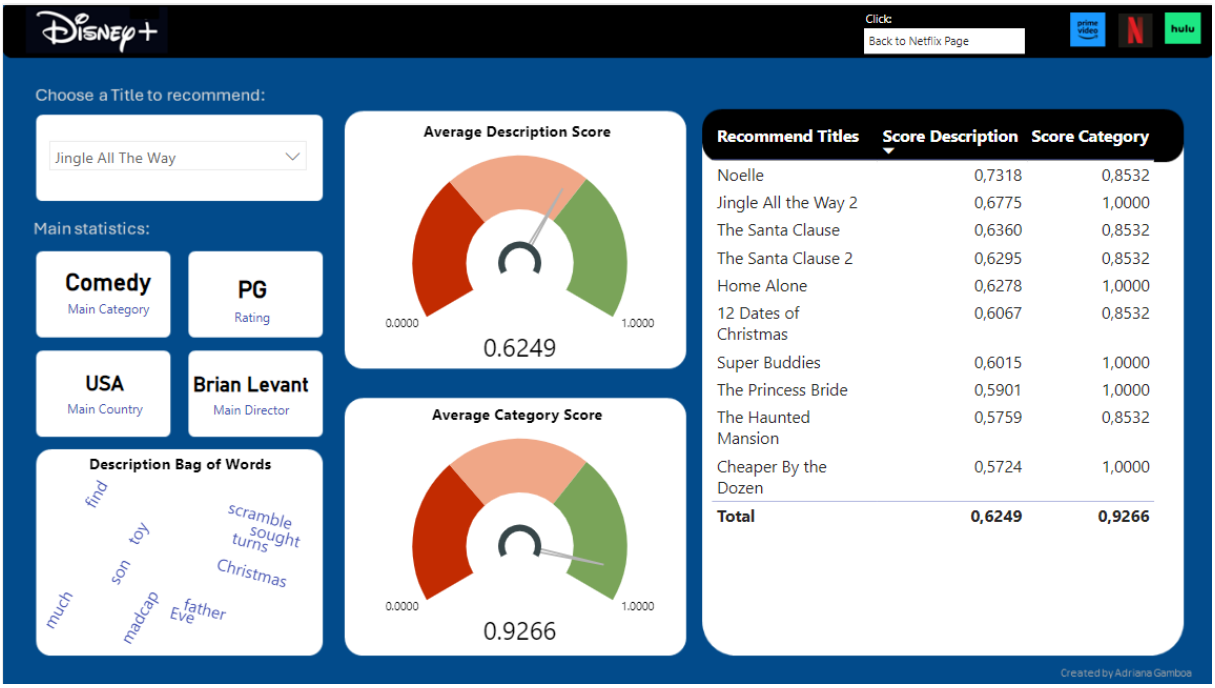

Figure a.4 – Disney+ Recommendation System's final report on Microsoft Power Bi.

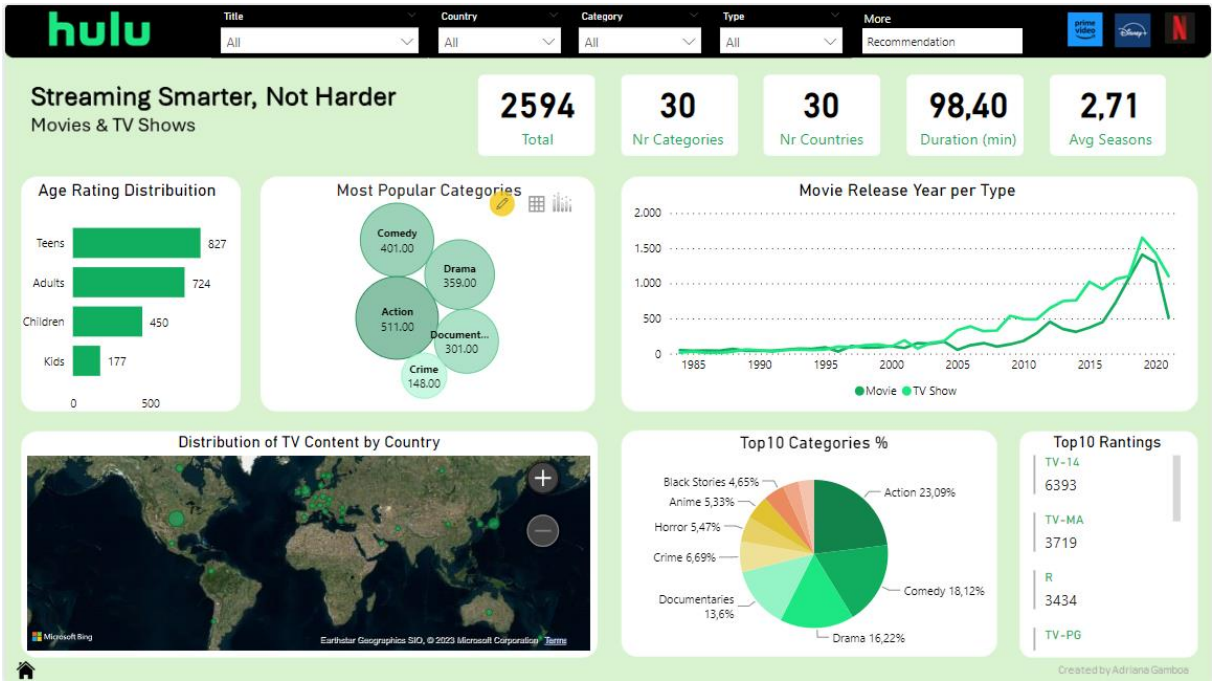

Figure a.5 – Final outcome from Hulu's dashboard on Microsoft Power Bi.

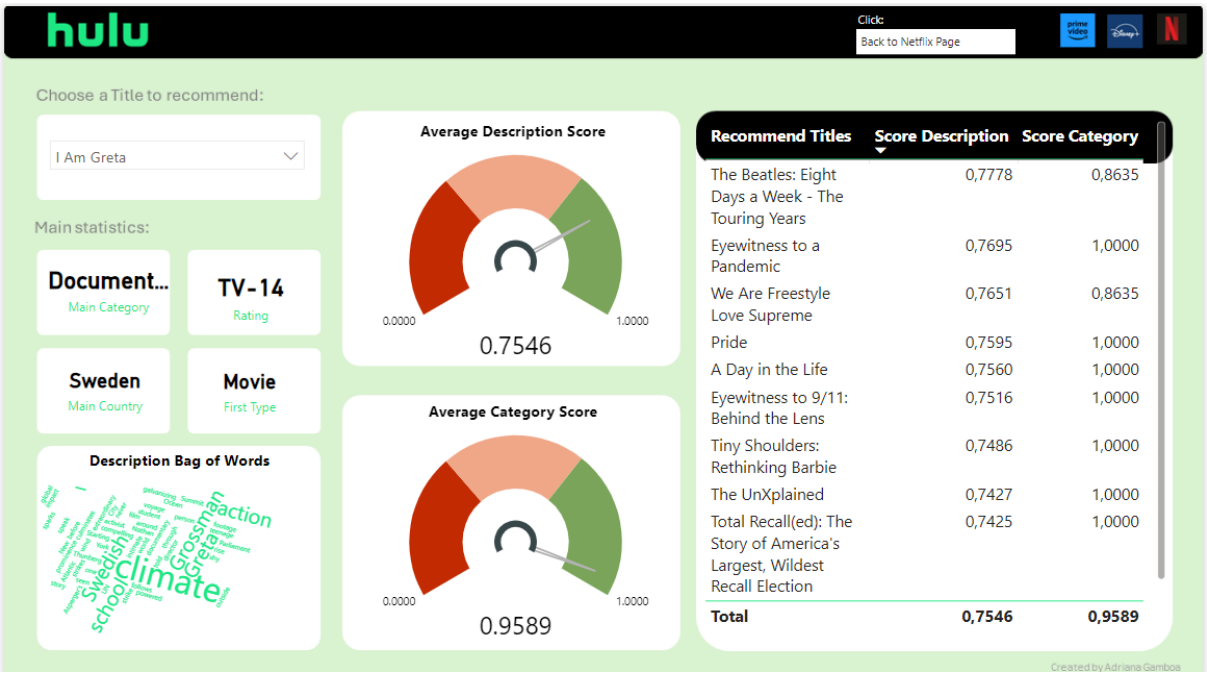

Figure a.6 – Hulu Recommendation System's final report on Microsoft Power Bi.

Appendix B

The following figures represented the User Survey made to collect valuable user feedback and assess interesting insights.

Stream Smarter, Not Harder

Dear Participants,

This invitations aims to motivate you to take part in this crucial survey designed to evaluate the Recommendation System developed as a part of the "Stream Smarter, Not Harder" Master's Thesis project. Your insights and opinions are invaluable in supporting the authour understand the effectiveness of this system and make improvements if necessary.

How?
You will be asked to rate your likelihood of the recommendation content based on what the recommendation system generated, on a scale of 0 to 10, with 0 being "Not at all likely" and 10 being "Extremely likely".

Thank you!

For being part of this important research initiative. Your contribution is invaluable in making streaming smarter for all. Your Feedback Matters!

Sincerely,

Adriana Gamboa

adriana.gamboa.brito@gmail.com Mudar de conta

Não partilhado

* Indica uma pergunta obrigatória

On Netflix, after watching "the social dilemma" would you watch "NOVA: Prediction by the Numbers"? *

	1	2	3	4	5	6	7	8	9	10	
Not likely.	○	○	○	○	○	○	○	○	○	○	Extremely likely.

On Amazon Prime, after watching "Marley & Me" would you watch "The Dog who saved Halloween"? *

	1	2	3	4	5	6	7	8	9	10	
Not likely.	○	○	○	○	○	○	○	○	○	○	Extremely likely.

On Disney+, after watching "Cinderella II" would you watch "Sleeping Beauty"? *

	1	2	3	4	5	6	7	8	9	10	
Not likely.	○	○	○	○	○	○	○	○	○	○	Extremely likely.

On Hulu, after watching "Apollo 11" would you watch "The Beatles: Eight Days a Week - The Touring Years"? *

	1	2	3	4	5	6	7	8	9	10	
Not likely.	○	○	○	○	○	○	○	○	○	○	Extremely likely.

Figure b.1 – Stream Smarter, Not Harder Survey.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa