

MAA

Mestrado em Métodos Analíticos Avançados

Master Program in Advanced Analytics

Data Analytics in the Football Industry:

The Portuguese Football Federation Case

Sofia Alexandra Grifo Cunha

Internship report presented as partial requirement for
obtaining the Master's degree in Advanced Analytics

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

DATA ANALYTICS IN THE FOOTBALL INDUSTRY

The Portuguese Football Federation Case

by

Sofia Alexandra Grifo Cunha

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

Advisor: Associate Professor Pedro Cabral

November 2021

To my Parents and my Sister.

ACKNOWLEDGEMENTS

First of all, I would like to thank the Portuguese Football Federation for the opportunity to carry out the curricular internship that allowed me to complete my master's degree. Just as I also wanted to thank all the team I was inserted in during this period, the Intelligence and Customer Service Department, but especially the Analytics team and the members of the Portugal Football Observatory.

I want to thank my advisor, Professor Pedro Cabral, for his guidance throughout the preparation of this internship report, and for his constant willingness to help me at all times.

I thank the College, NOVA Information Management School (NOVA IMS) for the training during these two years of master's degree, for all the knowledge I have acquired along this route, and also for the opportunities I have been given.

I would like to thank all my friends and companions since early age, from school, ISCSP and IMS who have always accompanied me along this journey, who have always helped me and motivated me to be better.

Finally, and the most special thanks of all is to my parents and to my sister, for constantly motivating me, supporting, and guiding me, both in my academic and personal career. Thank you for always being there. Thank you for believing in me. Without the three of you, I never would have gotten where I am.

ABSTRACT

In the football industry, an innovation that has emerged and is increasingly taking on a more important role is Data Analysis. With the attempt to seek a competitive advantage, Big Data was introduced as a way to obtain value and insights from information that at first seemed useless. Which leads to a new era in football and sports organizations where decisions are made based on data. Thus, the main objective of this internship report is to present a description of the implementation of the various data analysis and intelligence methodologies in the sport sector. Namely in a Football Federation, where the main sports under study are Football, Futsal and Beach Football, comprising all educational and professional sectors. This internship took place at the FPF (Portuguese Football Federation), in two different spaces. These being the Portugal Football Observatory and the Department of Intelligence and Service to the Fan, thus supporting the various studies and analyzes developed. The main result expected from the completion of this internship was the development of research methodologies, analysis and visualization of data related to football, and their application through data analysis technological tools.

KEYWORDS

Football; Sports; Data Analysis; Segmentation; Business Intelligence; Decision Making.

INDEX

1. Introduction.....	1
1.1Institution Context.....	1
1.1.1History.....	1
1.1.2The Organization.....	2
1.1.3Vision	2
1.1.4Mission.....	2
1.1.5Values.....	2
1.2Project Overview	3
1.2.1Portugal Football Observatory	3
1.2.2Intelligence and Customer Service Department:	4
2. Theoretical Framework	5
Business Intelligence and Analytics	5
Data Preprocessing and Data Preparation	6
Sports Analytics	7
Sports Big Data	10
3. Methodology	11
3.1 Tools and Technology.....	11
3.1.1Microsoft SQL Server Management Studio (SSMS).....	11
3.1.2Microsoft Office Excel.....	11
3.1.3Microsoft Office PowerPoint	11
3.1.4Microsoft Power BI	11
3.2 Data.....	12
3.2.1Internal Data Sources	12
3.2.2. External Data Sources	13
3.3Internship Description	14
3.3.1Training Period.....	14
3.3.2Portugal Football Observatory	15
3.3.3Intelligence and Customer Service Department	17
4. Results.....	21
4.1Internship experiences	21
4.1.1Training Period.....	21
4.2Portugal Football Observatory	23
4.2.1Women’s Football.....	23

4.2.2Futsal.....	27
4.2.3Relative Age Effect.....	29
4.2.4Replacement Minute Impact	30
4.3Intelligence and Customer Service Department	30
4.3.1Covid-19 Impact.....	30
4.3.2Databases Analysis.....	31
4.3.3Analysis of the growth of federated athletes	31
4.3.4Listing of Football Associations for financial support.....	32
4.3.5Audiences LIGA NOS and SuperLiga	33
4.3.6Portugal Store Stocks Management	33
5. Conclusion	38
Connection to Master's Program	38
Internship Evaluation and Results	38
Bibliography.....	41

LIST OF FIGURES

Figure 1. Portuguese Football Federation Logo	1
Figure 2. Portugal Football Observatory Logo.....	3
Figure 3. Forms of Data Preprocessing (J.Han et al, 2012).....	7
Figure 4: Sports Analysis Framework (Morgulev et al., 2018).....	8
Figure 5. Model of value creation from Business Intelligence & Analytics in competitive (Caya & Bourdon,2016)).....	9
Figure 6: Five Keys Features of Sports Big Data (Adapted from Bai & Bai, 2021).....	10
Figure 7. Training Period Methodology Diagram	15
Figure 8. Women’s Football Study Methodology Diagram	16
Figure 9. Futsal Study Methodology Diagram	16
Figure 10. Relative Age Effect Study Methodology Diagram	17
Figure 11. Replacement Minute Impact Methodology Diagram.....	17
Figure 12. Covid-19 impact Methodology Diagram	18
Figure 13. Databases Analysis Methodology Diagram	18
Figure 14. Analysis of the growth of federated athletes Methodology Diagram.....	18
Figure 15. Football Associations Financial Support Methodology Diagram.....	19
Figure 16. Audiences Liga NOS and SuperLiga Methodology Diagram	19
Figure 17. Portugal Store Stocks Management Methodology Diagram	20
Figure 18. Supporter Value Matrix	21
Figure 19: Distribution of the number of players of each type of school by birth quarter.....	30
Figure 20. Evolution of federated football players in Portugal over the last 9 seasons.....	32
Figure 21. Dashboard Dimensional Model	35
Figure 22. Provisory Dashboard Design	37

LIST OF TABLES

Table 1. List of the Dashboard Visualizations36

LIST OF ABBREVIATIONS AND ACRONYMS

AF	Associações de Futebol (Football Associations)
BI	Business Intelligence
CBFF	Centros Básicos para a Formação de Futebol e Futsal (Basic Football and Futsal Training Centers)
CTT	Correiros de Portugal (Portuguese Post Office)
DSS	Decision Support System
ETL	Extract, Transform and Load
FA	The Football Association
FFF	La Fédération Française de Football (The French Football Federation)
FIFA	Fédération Internationale de Football Association (International Federation of Association Football)
FPF	Federação Portuguesa de Futebol (Portuguese Football Federation)
INE	Instituto Nacional de Estatística (Statistics National Institute)
KDD	Knowledge Discovery from Data
KPI	Key Performance Indicator
PFO	Portugal Football Observatory
PFS	Portugal Football School
PVP	Preço de Venda ao Público (Sale price to the public)
RAE	Relative Age Effect
RFEF	Real Federación Española de Fútbol (Royal Spanish Football Federation)
SQL	Structured Query Language
SSMS	SQL Server Management Studio
UEFA	Union of European of Football Associations

1. INTRODUCTION

Based on the learnings throughout the 1st year of the Master's Program, this internship report is the application of the knowledges acquired. The internship was for 8 months (approximately 700h) in a Football Federation, the Portuguese Football Federation.

The internship aims to develop a set of studies and fieldwork, within the scope of the Portugal Football Observatory project, as well as in the Intelligence and Customer Service Department, supporting the studies and analyzes that have been developed over the course of the project internship, contributing to the completion of the master's thesis.

The purpose of the internship was to allow a privileged contact with the professional reality, aiming to complete and perfect the acquired skills, as a way of continuing the school path already accomplished. The main expected outcome with the internship was the development of research methodologies, analysis and visualization of data related to football and other themes related with the work developed by a Football Federation, as well as its application through technological tools.

Throughout the report it is possible to verify the division into several chapters, specifically: An introduction to the internship, the institution and the several projects carried out (Chapter 1); In this part, the Literature review/Theoretical Framework is made a contextualization about the theme (Chapter 2); The third part of the report is where the internship will be describe, namely the projects, data and tools (Chapter 3); The outcomes in terms of Acquired Skills are described next (Chapter 4); The conclusion on the internship and work done (Chapter 5).

1.1 INSTITUTION CONTEXT

1.1.1 History

The predecessor to the current Portuguese Football Federation (Figure 1) was the Portuguese Football Union (UPF) that was created by the regional associations of Lisbon, Portalegre and Porto on March 31, 1914. Its purpose was to officially create the National Team, this because FIFA it only allowed the participation of teams when they were part of a National Federation, made official on August 26, 1914 (FPF, 2021).



Figure 1. Portuguese Football Federation Logo.

The FPF as an institution comprises 22 district and regional associations, professional club leagues, sports agents' associations, sports clubs or societies, players, coaches, and referees.

In the modalities (Football, Futsal and Beach Football) governed by the supervision of the FPF there are 23 National Teams, divided by the various categories of formation and senior, as well as by the gender of the athletes. Currently, Portugal has 221,137 Football players, 56,558 Futsal players and

1,029 Beach Football players. Which makes a total of 270,724 practitioners in Portugal of the modalities governed by the FPF.

The main focus of FPF is the representation at national and international level of Portuguese Football, its competitive management and organization of competitions. It also aims to develop the sport in Portugal according to the values that govern the institution. These objectives are achieved through the training of sports agents.

1.1.2The Organization

The FPF is organized in such a way that it is possible to enhance and realize its vision, mission, and values. This structure is subdivided into two essential parts, the Social Entities, and the different Departments.

The Social Entities are: Chairman, General Meeting, Board of Directors, Fiscal Council, Justice Council, Disciplinary Council, Arbitration Council.

The existing Departments are: Commercial, Football City, Competitions and Events, Financial, Legal, Information, Marketing and New Business, People and Media, Portugal Football School, Intelligence and Customer Service, Human Resources and Administration, Records and Transfers, Responsibility Social and Public Relations, Service to Selections, Technology, Certification Unit, Health and Performance Unit.

1.1.3Vision

FPF's vision is to reinforce its role as the main reference of Football in Portugal, as well as to promote its development. This vision has as objective the recognition of the project with high relevance in Europe and in the World.

It works as a dynamic organization for sports, social, economic, educational and leisure activities in Football. This, considering the constant growth in the importance of training at younger categories, in the economy and in the Portugal international footprint.

1.1.4Mission

FPF's mission is to promote, in a way, the coordination, dynamism, development and organization of football as a whole. This mission aims to ensure the continued growth of the sport in Portugal, consequently increasing competitiveness in the international scenario.

For the fulfillment of this mission, there is international and intranational cooperation, as well as public-private partnerships. These cooperation's work to achieve the mission as expected, this in a vision of independence and mutual benefit.

1.1.5Values

To support FPF's vision and mission, there are several values that work to improve conditions, as well as to achieve the objectives. The values go through:

1. For FPF, success only exists when all its partners achieve it too, as well as always helping them reach their goals.

2. The principles that govern the FPF have the function of sustainable development, corresponding to the ethical, social, and environmental obligation standards its objectives.
3. FPF works based on a democratic structure and through dialogue with partners.
4. FPF encourages inclusion, knowledge acquisition, as well as a healthy lifestyle.
5. Zero tolerance for racism, discrimination, violence, and the lack of clarity of the games and results is an essential point of the FPF.
6. FPF intends to rationally and with attention to the economic part of Football, ensure that there is always transparency, integrity, loyalty, and honesty at all times.

1.2PROJECT OVERVIEW

During the internship I was working in two important parts of the FPF, namely the recent project Portugal Football Observatory, as well as in the Intelligence and Customer Service Department, both describe bellow.

1.2.1Portugal Football Observatory

Created in March 2021, under Portugal Football School (Figure 2) responsibility, the research and development unit of the Portuguese Football Federation. The creation of the Portugal Football Observatory (PFO) was made with a collaboration between three essential departments of the Portuguese Football Federation (FPF), namely the Intelligence and Customer Service Department, the Health and Performance Unit and the Portugal Football School (Portugal Football Observatory, 2021).



Figure 2. Portugal Football Observatory Logo.

The main objective of the PFO project is to collect and analyze information from several sources. This research makes possible to obtain conclusions on Football, a term that includes all the existing variants (Football, Futsal, Beach Football and eFootball). All the analysis has into account the gender of the athletes, the category they are inserted and the competition levels. And also, the season along with the evolution during them.

Other points where the PFO is focused is in the support for research, creation of a capability to disseminate knowledge to improve the sports policy and practices of the modalities studied. As well as the establishment of partnerships with other institutions for the advancement of research in the area.

Its main objective is the construction of a stage where it is possible to acquire knowledge, debate and discuss. This in order to make possible a better and greater development of Football in Portugal.

1.2.2 Intelligence and Customer Service Department:

Intelligence/Analytics

The Intelligence and Analytics part is related to everything that concerns knowledge and raising it, in order to bring as many insights as possible. The initial focus of this sub-department was to obtain data through collection campaigns, the integration of data that already existed in external sources, communication actions with supporters.

The knowledge part works a lot with the development of studies and analyzes. Namely regarding the satisfaction of FPF supporters, the production of analyzes to reflect and adjust strategies, interactions on social networks, support for the development of business plans, production of studies on various topics.

Customer Service

This part of the department comprises a few different tasks. These are based on Customer Care, that is, everything related to customer service, ticket office, CRM (Customer Relationship Management), among other points. The attendance of the fan and all this monitoring of his experience made the approach of the fans something real, as well as a digital and face-to-face management in moments of the game.

Other part of a great focus is Email-Marketing, which works as a way of sending messages, in this case emails, relevant and personalized according to the target audience of each message. This submission is made to subscribers of the newsletters, fan page, ticket office customers, among other data in the FPF databases.

As for the Portugal Store, which is partly a cooperation between this department and that of Marketing and New Business. It focuses on marketing campaigns, stock management, sales, etc.

Another very essential point is Portugal +, the official fan club of the national teams. It was launched in 2018 to bring fans and teams closer together, with the aim of transforming platform followers into acquaintances (members of databases).

Social Media

They manage the social networks of the different pages, and of the various types of digital platforms. The pages managed by the sub-department of social networks are Portugal, Liga Revelação, Liga BPI, Liga Placard, Canal 11. This separation of supporters by different pages serves as a way to create a segmentation by tastes and interests, and thus achieve a greater and more specific target audience.

Responsible for image editing, content creation and production for publication on the various pages.

2. THEORETICAL FRAMEWORK

There is a growing connection between sport and information technologies, which makes it a much more used tool in sports organizations. With the attempt to grow more and more, organizations have been trying to understand how to do it. In this way and taking into account the amount of data that can be collected during sports practices and that can promote important conclusions for this growth and in order to improve sports practices.

With the growth of Data Analysis, the Football Industry saw as an added value the use of Data Intelligence, what corresponds to use Data Analysis and Business Intelligence in the daily analysis, namely in a way to support the decision making for games, growth, and to improve the value. In this way, the usage of Football data has to face a set of challenges during the process of utilization. The first one, is related to the data collection and organization into databases, because it wasn't a normal procedure to capture data during the practices, or even to collect data from supporters to make decisions. The other challenge is related to the language of the data, by this what it is meant to be is that the data needs to be translated in a way that the information and conclusions acquired by the data are accessible to everybody, that is, that the data need to answer to the proposed question de forma simples, e above all, the analyses must be with a daily language that they can easily comprehend to be accessible to all in the Football Organization. (Jan Van Haaren, 2021).

BUSINESS INTELLIGENCE AND ANALYTICS

Progressively throughout the evolution of the world and technologies, the combination of Business Intelligence and data analysis has become a topic of greater interest for researchers and information systems workers. In this way, the real importance of the data was perceived, but above all the information that can be obtained from them and the advantages they provide in the short and long term. What made that several industries, such as insurance, sales, telecommunications, entertainment, among others, have increasingly bet on these technologies.

Business Intelligence and Analytics is described by Chen et al., (2012) as "the techniques, technologies, systems, practices, methodologies, and applications that analyze critical business data to help in the enterprise better understand its business and market and make timely decisions". Therefore, the use of these tools by organizations has as its main focus the attempt to take the greatest possible advantage of numerous data and information sources, this to somehow facilitate the decision-making processes.

So, **Business Intelligence** works as a tool/software whose objective is to transform data and information into insights that are accessible to everyone in order to help in decision making and establishing strategies for the organization. The tools used make it possible to present data in the form of reports, Dashboard, as well as in different types of views, which allows all users to understand the information that results from them about the organization. This is because Business Intelligence also refers to the fact that the tools result in quick and objective insights into the current state of the organization, as well as the past due to the fact that the data is historical (Pratt and Fruhlinger, 2019). Its main objective is to better support decision making, working as a DSS (Decision Support System). In this way, it is a system that allows access in the form of visualizations to historical and current data and, at the same time, to forecasts.

Taking value out of data in order to promote value, to create answers to questions that until then were rhetorical. In this way, **Data Analysis** is the process where statistical and logic tools are applied to perform the description, evaluation, and illustration of the collected data. What can also be understood as the set of analytical methods where it is possible to “draw inductive inferences from data and distinguish the signal (the phenomenon of interest) from the noise (statistical fluctuations) present in the data (Shamoo and Resnik, 2009)”.

There are several applications of Business Intelligence and Analytics tools, including Dashboard, Data Mining and Data Warehouse, ETL process (Extract, Transform and Load), KPI (Key Performance Indicators), among others.

Having into account the applications listed above, Data Mining, which corresponds to the extraction of knowledge from data via Data Science techniques. Data Mining is, as well, an Unsupervised Learning method, essentially it uses clustering and segmentation to find patterns and classes, so it works as a descriptive method. Also known as Knowledge discovery from Data (KDD), which the main focus is to discover patterns and trends from large amounts of Data using several techniques. Because it transforms large amounts of data, it works like finding a small piece of knowledge in the middle of something that was apparently just information (Chapman P. et al., 2000). For this, there is a process composed of seven steps for the discovery of knowledge, namely:

1. Data Cleaning, where the data is treated as the noise is eliminated and the consistency of the data verified.
2. Data Integration, in which data are combined so that later conclusions can be drawn.
3. Data Selection, which is the moment when the most relevant data are selected from the base data.
4. Data Transformation where data is processed and consolidated so that it is ready for aggregation, summarization, and analysis operations.
5. Data Mining is the essential process in which Intelligence and Analytics methods are used to find patterns.
6. Pattern Evaluation is the moment when the patterns and trends that represent the knowledge obtained through the data are identified.
7. Knowledge Presentation is where the reporting of the conclusions obtained is carried out, mainly via visualizations.

DATA PREPROCESSING AND DATA PREPARATION

Of the seven steps for discovery of knowledge mentioned in the paragraph above, the ones that belong to this part of the process are: Data Cleaning, Data Integration, Data Selection and Data Transformation (Figure 3). These four steps correspond to process of transforming raw data into data that can be used in analysis and also in the modelling process. Having this phase as the objective to answer questions such as *“How can the data be preprocessed in order to help improve the quality of the data and, consequently, of the mining results? How can the data be preprocessed so as to improve the efficiency and ease of the mining process? (J.Han et al., 2012)”*.

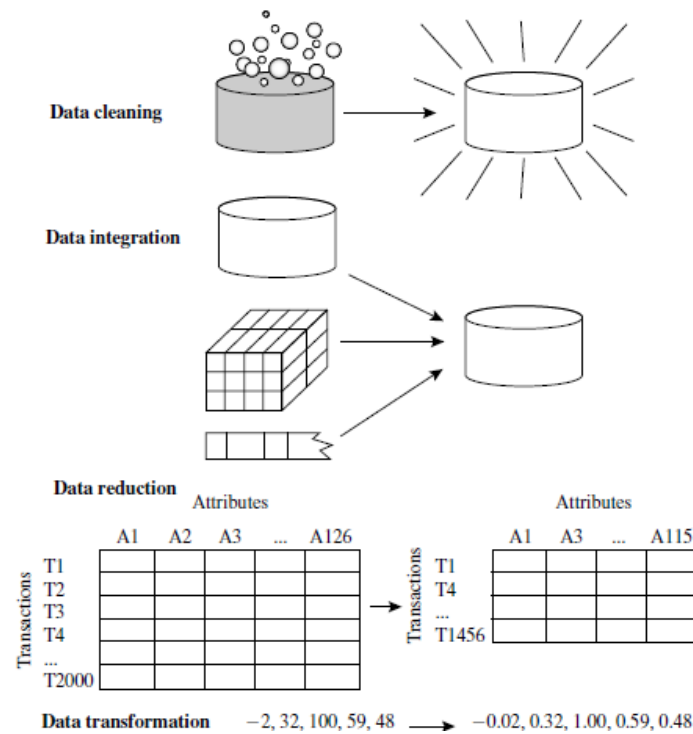


Figure 3. Forms of Data Preprocessing (J.Han et al., 2012).

In this way, it is verified that the fact that these preprocessing techniques are applied before exploring the data, there is an improvement in the quality of the analyses, as well as the fact that the time needed for their elaboration is also smaller, due to the efficiency and effectiveness allowed (J.Han et al., 2012)". This is because for an analysis to work well and correctly, the existence of quality in the data is necessary, which is possible to guarantee through these techniques, allowing the Accuracy, Consistency and Completeness factors to exist. This data quality is permissible due to the fact that this steps, the main objective is to transform dataset into a more comprehensible set of information. Basically, these steps allow that the quality of the analysis improve, this because if good data exists, good analysis will also exist, since it's a prerequisite.

The main focus of the data pre-processing and data preparation is to eliminate the signals and noise that exists on the data. Basically, it corresponds to eliminate or treat the outliers (values that differ from the rest of data), treat the missing values, treat duplicate values, and treat some other inconsistencies.

SPORTS ANALYTICS

The importance of data science and analytics in the world is increasingly noted, namely Big Data, Artificial Intelligence, Business Intelligence, Machine Learning and Data Analysis, so 'Sports industry seems to be one of the most promising applications of these modern technologies '(Patel et al, 2020). In this way, we can characterize sport analysis as an application of sports historical data or related to sports organizations (example of sports fans and agents), descriptive and predictive models in order to understand the present and future of the industry, as well as the use of information systems for more informed decision making (Patel et al., 2020) These data, after collected, are treated in order to understand the information obtained, namely, they are

standardized, and analyzed using a variable of metrics. Therefore, the data influence the decision-making process, as when analyzed, it is possible to see trends and patterns in the data, which in a way makes the chances of success greater, both at the sporting and the organization level. Functioned as a Sports Analysis Framework (Figure 4), whose knowledge makes decision-making more sustainable in organizations, as well as improves their decision system (Morgulev et al., 2018).

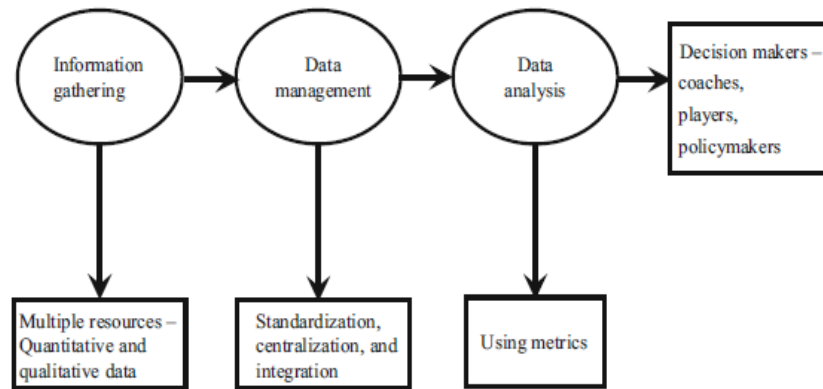


Figure 4: Sports Analysis Framework (Morgulev et al., 2018).

This Sports Analysis Framework (Figure 4) shows that there is a workflow that begins with obtaining information from different sources and in different ways. When having contact with this information, it is then possible to carry out a management, the so-called data management phase, where the previously mentioned step of data preprocessing and data preparation comes into action. With these two parts of the flow carried out, it is then possible to carry out the proper analysis, using different methodologies and metrics to facilitate the final objective, which is decision-making based on the conclusions obtained during the analysis process.

While this framework provides a high-level representation of how and where this type of project creates value in competitive sports, it drives more accurate examples of value creation at each level of analysis (institutional, organizational, and individual levels). Within each "sub-model", specific details about the nature of the Business Intelligence and Analytics initiatives, along with a specific conversion, contingencies and value measures created from those same initiatives.

A structure was also developed (Caya & Bourdon, 2016) in order to realize the value that investments in Business Intelligence and Analytics can promote for projects in sports organizations in where they will be inserted (Figure 5). Thus, there are 4 categories in the organization, which are the following ones:

- Institutional, which relates to the institutions and sports federations that manage the sport, as they are the main structure, hence this level is considered the main category.
- Organizational, it is the intermediate category that corresponds to the teams and their staff. For example, the entire administrative and management part of the club, coaches, and analytical support team. This category also includes information related to club costs and income, as well as ticket sales, salaries, among others. Individual relates to athletes.

- External Actors, which are the group of people who are not directly concerned with the transformation and the creation of value, but are the main consumers of sports data, as they are fans, data analysis companies, sports bloggers, ect. These actors are one of the main people responsible for making it essential to invest in Business Intelligence and Analytics.

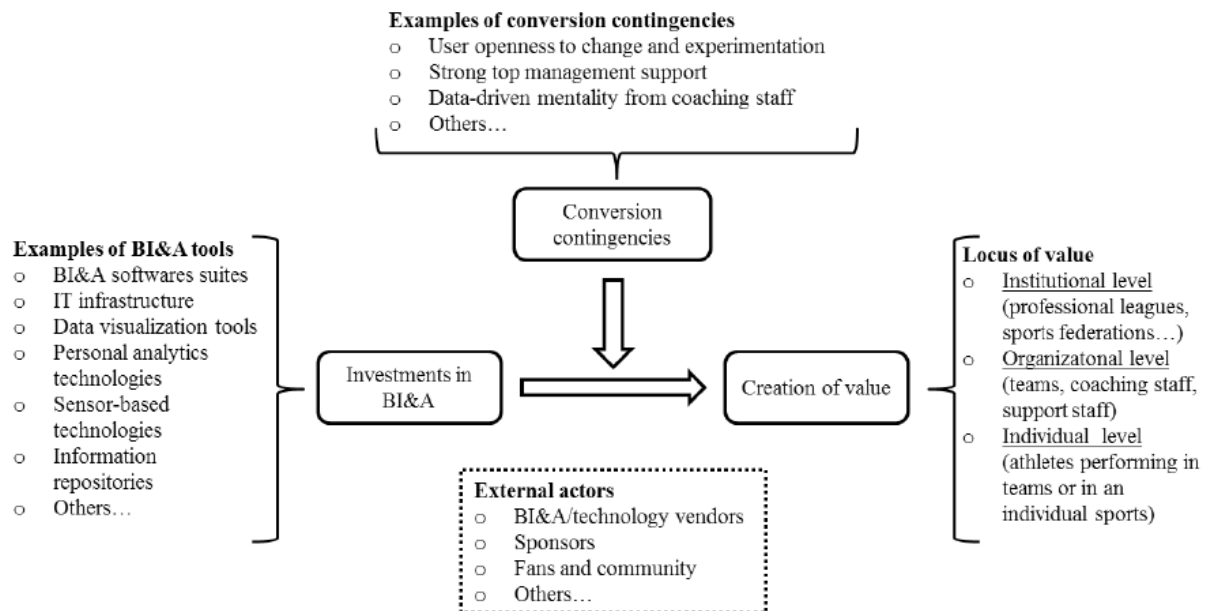


Figure 5. Model of value creation from Business Intelligence & Analytics in competitive sports (Caya & Bourdon, 2016).

This model (Figure 5) shows which investments need to be made in BI and Analytics, in which tools to create value and what possible outcomes may arise, which also exist for each level of analysis. Thus, this model demonstrates the types of investments that must be made, such as in software, visualizations, information repositories, among others. As well as saying examples of contingencies of conversion. Finally, reaching the end of the model and the final objective, which is the creation of value for each of the levels through the use of BI and Analytics tools.

Meanwhile, Caya and Bourdon also provide examples of how the application of these technologies is important at other levels beyond the competition itself in order to create even more value from existing data, in this case at three levels:

- At the Institutional level there is an investment in the creation of iterative views, as well as in social media and advertising, this because they intend to have a strategic position. Thus, the creation of aims to realize the satisfaction of fans and the entire sports community, as well as promoting an increase in the institution's revenues.
- At the Organizational level there is an investment in specialized software to analyze data, as well as in algorithms and statistical tools. Tools that help in data management, opening up to digital transformation. These investments are aimed at creating value by improving the preparation of games and athletes, historically due to historical analysis, and also so that there is an increase in revenue.

- At the Individual level there is an investment in analysis technologies, as well as in reporting and data visualization tools. Thus, the main objective of this level is the improvement of the athletes' performance, as well as the prevention of injuries and any possible unhealthy issues.

SPORTS BIG DATA

As a form of growth for the sport, both institutionally and in terms of the sports practice itself, Big Data has played a very important role in the treatment and analysis of data. In this way, and due to the growth, that sports data has had over time, as well as the profit that can be built. In this way, Big Data for sport is defined as determined, storage, management and analysis of the data (Bai & Bai, 2021).

For this Sports Big Data analysis, 5 key points are taken into account in the analysis process (Figure 6).

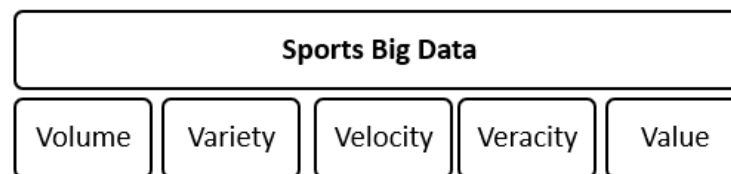


Figure 6: Five Keys Features of Sports Big Data (Adapted from Bai & Bai, 2021).

- Volume - Each time there is a greater amount of data for transport.
- Variety - Data come from different sources of information, which ends up leading to a giant variety.
- Velocity - With the increase in competitive advantages and the need for quick recovery, data processing has to be faster and faster.
- Veracity - The data used must have a true base, so you always have to check the source of provenance.
- Value - This is the most important point of Big Data, as the main objective of data analysis is to obtain the value of information that in part seemed useless or unimportant.

The whole purpose of using Big Data in sports, in this case in the football industry, is to obtain relevant knowledge for decision making, as well as for improving the performance of athletes and to improve the sport itself. Basically, it focuses on trying to solve problems based on data, problems that are solved with data mining techniques, statistics, among other information technologies. In this way, all valuable information and data are then presented through clear and objective views, so that conclusions are as noticeable as possible, so that everyone understands them.

3. METHODOLOGY

3.1 TOOLS AND TECHNOLOGY

The support of tools and technologies is essential to perform data analysis tasks. The projects and analysis carried out during the internship were based on tools learned throughout the curricular year of the Master's degree, and also during the internship itself, mostly in the training month. The principal tools and technologies used during the internship to the development of projects and every type of analysis are presented below.

3.1.1 Microsoft SQL Server Management Studio (SSMS)

Microsoft SQL Server Management Studio is an integrated environment for managing any type of SQL infrastructures, where the language used is SQL (Structured Query Language), that is a language with the purpose of store, manipulate and retrieve data in a database. It presents a simplified graphical interface in terms of writing, editing, and maintaining databases. SSMS provides tools to configure, monitor, and administer instances of SQL Server and databases. The biggest and most important feature of SSMS is to explore. It is used to query, design and manage the databases and data warehouses (Microsoft, 2021).

The most used function used in SQL queries were: SELECT, SELECT, SELECT TOP, JOIN, WHERE, GROUP BY, ORDER BY, SUM, COUNT, MIN and MAX, UNION ALL.

3.1.2 Microsoft Office Excel

Microsoft Office Excel is a spreadsheet application with the purpose for create, edit, view, and share files in a quick and easy way. It is a program whose objective is data analysis and documentation. It uses a grid of cells form of store the data, where the rows are numbered, and the columns are letter-named. Compost by a diverse variety of functions to help in the analysis of the data, concerning to statistics, segmentation, targeting, concatenation, and others. It also displays of a set of visualizations, types of storage, and even pivot tables (GooglePlay, 2021).

The most used function used in Excel were: COUNT.IF, SUM, VLOOKUP, CONCAT, 'Remove Duplicates'.

3.1.3 Microsoft Office PowerPoint

Microsoft office PowerPoint is a program used to create presentations, where the main objective is to present about a certain topic or report, which can be used for the same: images, sounds, texts and videos that can be animated in different ways (Microsoft, 2021).

3.1.4 Microsoft Power BI

Microsoft Power BI is a set of applications, software services and connectors working together used to create interactive and immersive visualizations and insights. Basically, it turns unrelated sources of information into coherent data. It is also very important to create dashboards which assists in decision making. This due to the possibility of having current actualization of the data (Microsoft, 2021).

3.2 DATA

3.2.1 Internal Data Sources

There are several information sources in the FPF Databases, where the id (identification of the row) of each one is the email. The Databases were accessible in SQL. During the internship, I use four of them, which are:

SCORE

The Database SCORE is related to data relating to all sports agents associated with the FPF, currently comprising 5 129 138 records, divided into 4 fields: Athletes (1 113 364), Coaches (33 035), Referees (17 705), Directors (139 552) and Staff (181 056).

In general, there is a large discrepancy when it comes to gender distribution. So, there are 95% male records and only 5% female. This is equivalent to 1 376 879 men and 73 724 women.

When dividing by different sports agents:

- Athletes: 96% (1 065 145) male records and only 6% (48 220) female.
- Coaches: 98% (27 897) male records and only 2% (626) female.
- Referees: 91% (16 005) male records and only 9% (1 491) female.
- Managers: 93% (120 709) male records and only 7% (8 852) female.
- Staff: 91% (147 122) male records and only 9% (14 535) female.

In terms of age-related distribution of sports agents, there is a greater concentration of middle age groups, with the following distribution: 13-17: 9% (113954); 18-24: 15% (188313); 25-34: 21% (275 633); 35-44: 22% (278 228); 45-54: 18% (233 378); 55-64: 11% (147,132); +65: 4% (52 411)

The distribution by different sports agents does not show statistically relevant changes when compared to the general distribution.

As for the nationality of sports agents, they are mostly in Portugal, corresponding to 97% (1 282 203). Only in coaches there is a difference from the pattern where 73% (24, 187) are nationals.

Portugal +

The Database do Portugal + corresponds to the data related to the members of the Fan Club of the Selections, and there are currently 238 975 records.

When it comes to gender distribution, there are 66% (157 637) male records and 34% (81340) female, not presenting a discrepancy as high as in other FPF databases.

In terms of age-related distribution of Portugal + members, there is a greater concentration of middle age groups, with the following distribution: 13-17: 2% (4,997); 18-24: 17% (41,082); 25-34: 22% (51,519); 35-44: 23% (53,396); 45-54: 20% (47,155); 55-64: 10% (24,556); +65:6% (14,419).

As for nationality, around 71% (170,325) are nationals, while 29% (68,654) are foreigners.

Ticket Office

The Ticket Database corresponds to the purchasers of tickets for official matches of the national teams, with the number of users being 117,350.

Taking into account the gender of ticket users, 73% (57,619) of the registrations are male, and 27% (21,538) are female.

As for the age group distribution, there is a greater concentration in the Middle Ages, with the following distribution: 13-17: 0% (311); 18-24:13% (12,352); 25-34: 32% (31,419); 35-44: 28% (27,143); 45-54:19% (18,923); 55-64: 6% (5,572); +65: 3% (2,918)

Focusing on Nationality, the vast majority of users are nationals of Portugal corresponding to 91% (107 280). Since international are about 9% (10,213) of the registrations.

Portugal Store

The Portugal Store Database corresponds to buyers of articles in the official FPF store, so the number of users so far is 39,084.

Taking into account the gender of Portugal Store users, 68% (7,999) of the records are male, and 32% (3,689) are female.

As for the age group distribution, there is a greater concentration in the Middle Ages, with the following distribution: 13-17: 1% (131); 18-24: 22% (2189); 25-34: 29% (2932); 35-44: 26% (2,669); 45-54: 16% (1645); 55-64: 4% (433); +65: 1% (118.)

As for nationality, around 68% (8 231) are nationals, while 32% (3 847) are foreigners.

3.2.2. External Data Sources

In a way to complement the already existing database, it is necessary to access to external data.

ZEROZERO

Founded in 2003 it works as a way to offer information about Football, both about Portugal and about the sport internationally. Follows football competitions at different levels, starting with the under-8. As well as the various variants of Football, such as 7-a-side football and beach football. As for the following competitions, these are divided into 4 types: Portuguese Competitions-14; Foreign National Competitions-6; International Club Competitions-9; Selection Competitions-9 (ZEROZERO,2021).

INE (Instituto Nacional de Estatística)

The National Institute of Statistics, founded in 1935, is the body responsible for creating and disseminating statistical information related to national issues and data. Also responsible for the demographic studies of Portugal and the Communities (INE, 2021).

PORDATA

Project created by the Francisco Manuel dos Santos Foundation in 2009 and works as the 'Contemporary Portugal Database'. The Foundation's work involves collecting, organizing, systematizing and sharing information about Portugal. The data refer to Portugal as a whole, municipalities and also European countries. The main focus is to make the information accessible and clear to the general population (PORDATA, 2021).

OPTA

It is a sports data company founded in 1996 after the merger between two existing companies, OPTA Sports Data and SportingStatz. Its purpose is the analysis, storage and distribution of sports data (OPTA, 2021).

Liga NOS

In the 2020/2021 season, NOS was the official sponsor of the competition, so the league's website was responsible for information regarding the games, their calendars, and information about the games (day, time, teams, among others) (Liga NOS, 2021).

Transfermarkt

It is a German website launched in 2000 dedicated to football and its modalities, where it is possible to collect information related to the game of the various international and national competitions in each country (mainly from the main European leagues), as well as national team competitions. Statistics are also available related to three themes: Players, Coaches and Clubs (Transfermarkt, 2021).

3.3 INTERNSHIP DESCRIPTION

As part of the internship, I was able to develop and be part of the following projects, in the project Portugal Football Observatory and in the Intelligence and Customer Service Department. But, before starting the internship as a Data Analyst, a training period was also carried out.

3.3.1 Training Period

During the first month of internship as a data analyst, I trained in order to learn about the work I would carry out over the next 8 months. As was the period when small tasks were carried out with the future work. Throughout this period, three themes were carried out essentially, being them about a targeting for all the supporters of Portugal+ and of the Portugal Store, a targeting of all the database SCORE members, as well as a presentation about the EURO2020. The main point of the work done during this period and the method to do it is synthesized in the follow diagram (Figure 7).

Theme	Targeting supporters of Portugal+ and Portugal Store for email marketing campaigns	Targeting SCORE Members	Presentation Creation
Work	Segmentation of supporters according to their engagement	Collection of data relating to athlete profiles; Segmentation of athletes and sports agents according to data available in the database; Answers to questions relating to National Teams	Creating a presentation on the EURO2020 and supporters' journeys to the games
Methodology	Data collection via SQL, namely the email	Data collection via SQL	Synthesis of data to be presented in Power Point

Figure 7. Training Period Methodology Diagram.

3.3.2 Portugal Football Observatory

Within the project of the Portugal Football Observatory (PFO), several studies were developed in order to understand how the growth of Football and the two modalities in Portugal is. As well as assess the impacts that certain also cause in situations related to football as a whole.

Thus, four studies were developed within the PFO project, which were:

1. Women's Football: 'How to recruit more and retain better in Women's Football?'
2. Futsal: 'The evolution of Futsal in Portugal and the victory in the 2018 European Championship as an accelerator for growth?'
3. Relative Age Effect: 'Does the time of the year in which one is born condition the opportunity to be a grassroots athlete?'
4. Replacement Minute Impact.

Women's Football

In this part of the internship, and within the PFO project, it is intended to answer the question 'How to recruit more and retain better in Women's Football?'. In this way, a number of tasks have been developed, including a benchmark related to the role and potential growth of women's football in Europe. As well as being to the main point of this study were carried out analyses of surveys made to girls who like and practice or do not play Football, parents who are the great motivators of the practice, and also to football clubs as a way of perceiving in their view what are the main factors that lead to growth and what are the limitations of the sport in Portugal. Therefore, the following diagram (Figure 8) is the summary of the tasks performed in this study.

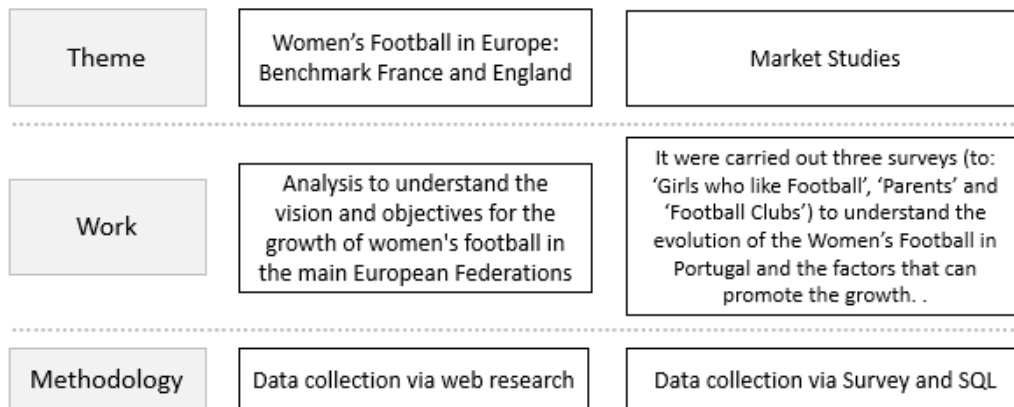


Figure 8. Women's Football Study Methodology Diagram.

Futsal

In this particular part of the internship, and within the PFO project, it is intended to answer the question 'The evolution of Futsal in Portugal and the victory in the 2018 European Championship as an accelerator for growth?'. Thus, we analyzed the surveys conducted to athletes who practice the sport, to former athletes, those who only like futsal. Therefore, and analyzing the main source of data it was possible to perceive the impact that the victory in the European had in the modality, as well as the growth of the practice of futsal in Portugal. Therefore, the following diagram (Figure 9) is the square of the market study realized for this study.

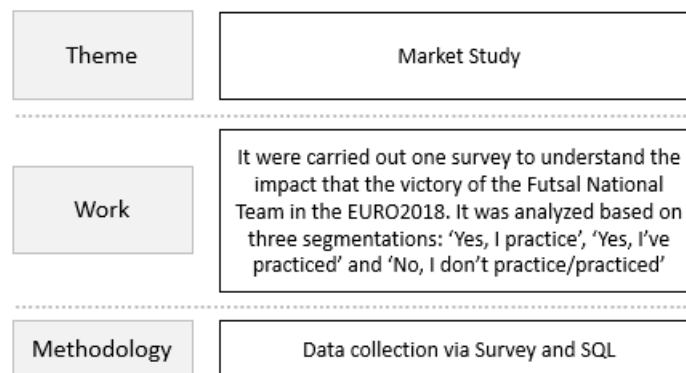


Figure 9. Futsal Study Methodology Diagram.

Relative Age Effect

The main objective of this study is to respond to this particular question 'Does the time of the year in which one is born condition the opportunity to be a grassroots athlete?'. So, this is all intended to analyze the impact that the month, whether the fact of being born at the beginning or at the end of the year is a decisive factor to be a good athlete or not in the future. Thus, data were collected to be analyzed in order to perceive the existing distribution by quarter of births. For this study it was also necessary to create a table with data related to certified clubs for other analyses performed by other PFO members. Therefore, the following diagram (Figure 10) is the resume of the work performed in this study.

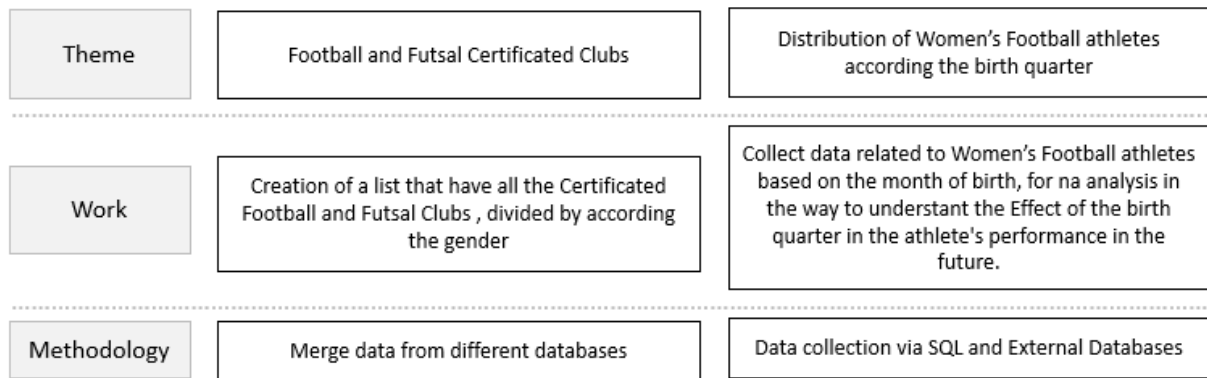


Figure 10. Relative Age Effect Study Methodology Diagram.

Replacement Minute Impact

This study is about the impact of altering the number of players substitutions in the Portuguese league. In this way it was necessary to collect data on substitutions made in the matches of the main European football leagues. Therefore, it was necessary to collect data through the external database, OPTA. The following diagram is the summary of the task concerned (Figure 11).

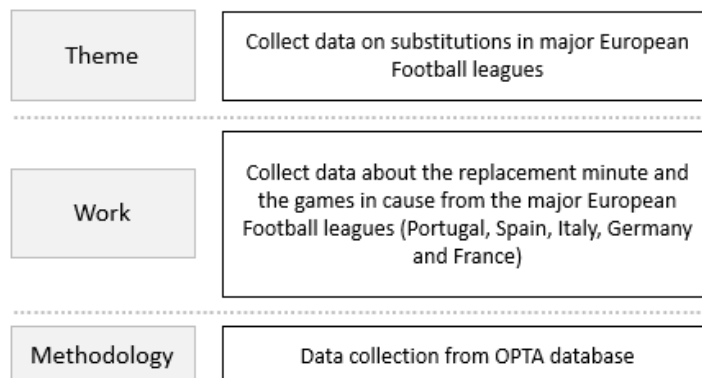


Figure 11. Replacement Minute Impact Methodology Diagram.

3.3.3Intelligence and Customer Service Department

Within the project of the Intelligence and Customer Service Department several ad hoc analyses were developed in order to help in the decision making, and also to evaluate the current image of the different modalities inside Football, and also the image from the FPF as an organization with a diverse purpose. In the way it was developed Analysis, being them:

1. Covid-19 Impact.
2. Databases Analysis.
3. Analysis of the growth of federated athletes.
4. Listing of Football Associations for financial support.
5. Portugal Store Stocks Management.

Covid-19 Impact

During this ad hoc analysis, the main objective was to understand the impact that the pandemic situation caused in the Football Clubs in terms of income and associated costs, which is summarized in the diagram (Figure 12).

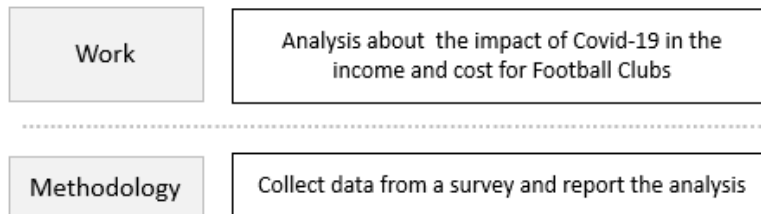


Figure 12. Covid-19 impact Methodology Diagram.

Databases Analysis

With the purpose of knowing how the image of the databases is, this analysis was carried out. To better understand it was performed in general, since the creation of each of the databases. As well as only since 2018/2019 to have a more real image of the growth of recent times. The resume of the analysis is the following diagram (Figure 13).

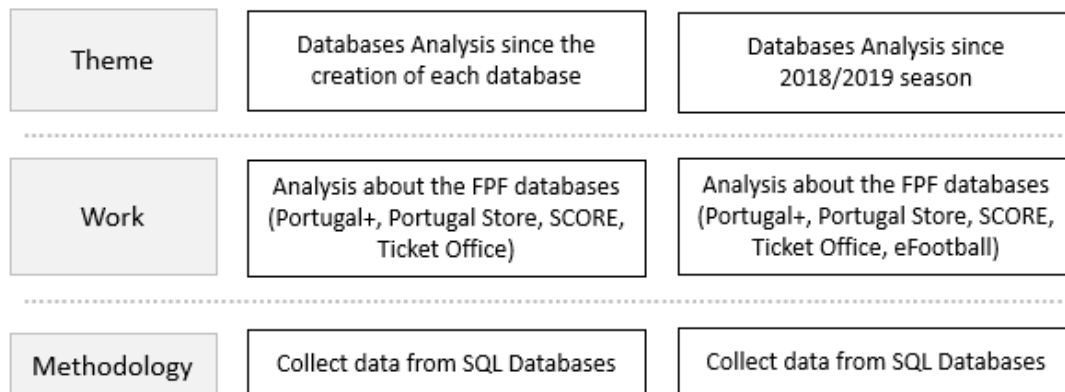


Figure 13. Databases Analysis Methodology Diagram.

Analysis of the growth of federated athletes

This analysis is just to understand and visualize the evolution of the growth rate of federated Football players over the last 9 seasons (Figure 14).

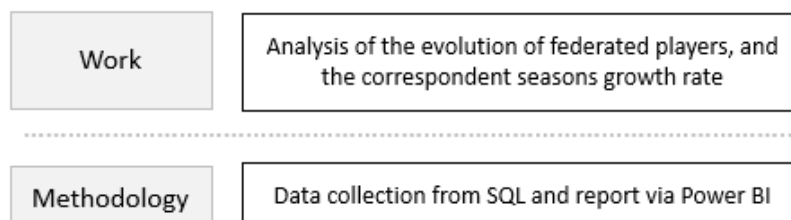


Figure 14. Analysis of the growth of federated athletes Methodology Diagram

Listing of Football Associations for financial support

A fund has been set up for the allocation of financial support by the FPF and the Portuguese State, due to the situation and the lack of revenue largely because of the stopping of competitions. Thus, it was necessary to analyze the data sent by the Football Associations about the clubs of each association and buy them with those of the databases. This process is summarized in the following diagram (Figure 15).

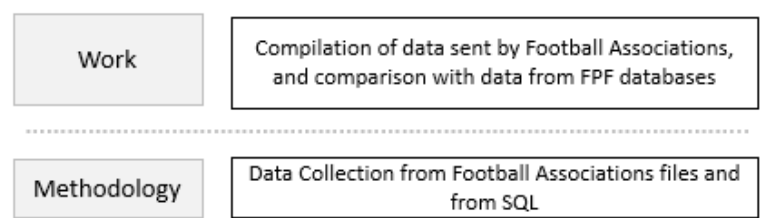


Figure 15. Football Associations Financial Support Methodology Diagram.

Audiences Liga NOS and SuperLiga

An analysis was carried out on the audiences of the main Portuguese Football League, the Liga NOS, and SuperLigue, a league that would be created by the main European Football Clubs. For this analysis it was necessary to collect data about the football games played by the clubs from different data sources. The plan of work is the following diagram (Figure 16).

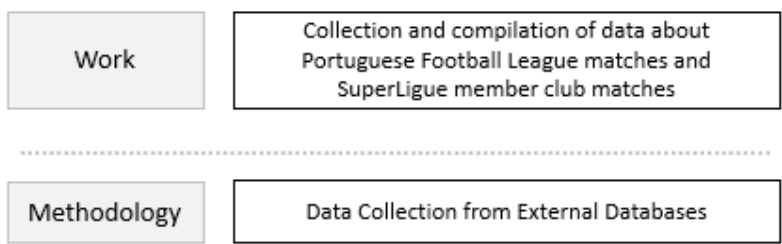


Figure 16. Audiences Liga NOS and SuperLiga Methodology Diagram.

Portugal Store Stocks Management

In order to understand and visualize what’s happening in the Portugal Store in terms of products stocks it was necessary to create a file with all the information about that, namely about the amount of each product that exists in stock, for how long the products have been in storage, what are the production and sale price, and finally how much is the profit margin of each one of them. Another part of work, and after this one, was the construction of a Dashboard with all the information present in the file created, in order that the data is accessible to all members of the organization. Therefore, the following diagram (Figure 17) is the summary of the tasks performed in this analysis.

Theme	Portugal Store Products Stocks and Profit Margin	Portugal Store Stocks Dashboard
Work	Creation of a table with the information about all the store products, and the calculate of the profit margin	Creation of a Dashboard with all the information present in the Excel spreadsheet previously created
Methodology	Collect data from SQL databases and compile into an Excel spreadsheet	Creation of a Dashboard in Power BI

Figure 17. Portugal Store Stocks Management Methodology Diagram.

4. RESULTS

4.1 INTERNSHIP EXPERIENCES

As explained in chapter 3, in the point on the Internship Description, it is possible to verify that it was composed of three different phases:

- Training Period.
- The projects that are part of the Portugal Football Observatory.
- Ad hoc analysis for the Intelligence and Customer Service Department.

4.1.1 Training Period

During this period, I had my first contact with the FPF databases, previously described in section 3.1.5 of this report, which are composed of data from the Portugal Store, Ticket Office, SCORE and Portugal +. In order to be able to work with these data, particularly with regard to their collection, to later be worked on and conclusions drawn from them, it was planned to learn how to work with SQL.

For learning and better consolidation of SQL knowledge, several introductory exercises were performed, which worked as hypothetical situations of data that in the future would have to be removed, or even related to how to know and have knowledge of existing data in each of the databases, as well as the different tables that compose them. As well as to have an atmosphere with the work, presentations of the analyzes were also carried out to learn how they would have to be carried out in the future in official work and for decision-making.

In this way, the initiation exercises and the small analyzes were related to:

1.Targeting supporters of Portugal + and Portugal Store for email marketing campaigns. Therefore, to create this segmentation it was necessary to take into account the Supporter Value Matrix (Figure 18). In this matrix, each supporter is classified according to the commitment he holds, such as visits to the website, open emails, and visualization of the content through the click. And also taking into account the income from it, such as purchases at the official store, purchases at the ticket office. Another point to be considered is the supporter's seniority in the FPF database.

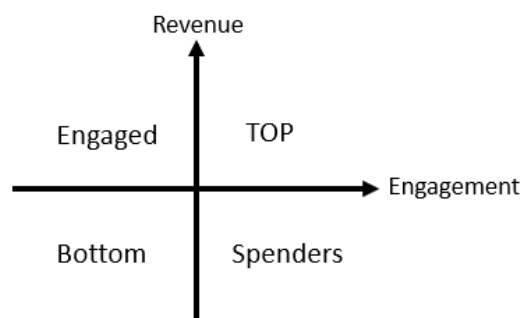


Figure 18. Supporter Value Matrix

To carry out segmentations also related to email marketing campaigns, it was also necessary to take into account the so-called Supporter Life Cycle, which corresponds to the classification of supporters according to

1. Database not activated: the supporter does not authorize contact via email.
2. Pre-active: supporters who have not yet been contacted but are eligible for the same.
3. New: supporters who have 1 to 3 contacts, however they have not yet created Engagement, as they have not made purchases or opened their emails.
4. Active: supporters who open emails and make purchases, creating Engagement, but who at the same time are never classified as Suspended.
5. Unknown: supporters who had more than 3 contacts, however they never opened any, who do not show interest in the information sent.
6. Non-active: supporters who have more than 10 contacts, but who have never opened any emails, or created any type of Engagement.
7. Suspended: supporters who were active, however did not open the last 10 emails sent.
8. Reactive: are those who were suspended or not active, but who made a purchase or opened an email, that is, they returned to creating Engagement after some time stopped.

Basically, and considering these ratings for fans, the exercises performed were related to the collection of emails via SQL from fans, being filtered according to the commitment and the revenue they promote for the FPF, namely for the Portugal Store, Portugal + and for the Ticket Office, but also according to how active they are for campaigns, for contact via email, among other ways to create commitment.

Thus, email collections were carried out via SQL of questions such as 'What is the email address of Portugal+ members who have already been to games?', 'What is the email address of Active/Reactive members who have been to the National Team games (dividing between which ones are TOP and which ones are Engaged)?'. These two questions are a small sample of the segmentations and targeting that were necessary as a way of setting these databases.

2.Targeting SCORE members. As a way to know the composition of the SCORE database, which tables there are (the joins between tables that were necessary to know later on to carry out the necessary analysis and which could be done), the columns and data that can be removed in order to create knowledge, and to answer questions for future works and projects, namely Ad Hoc and PFO analyses. This entire exercise was performed using queries and SQL.

Therefore, and as a first exercise, it was the answer to questions in order to draw the profile of a sports agent, in this case the profile of an athlete with only the agent's license number. The collected data then corresponded to name, age, nationality, email, telephone number, gender, sport, level, division, type of football, in how many seasons he played, and in how many teams he played. Therefore, it was possible to fully characterize, taking into account the existing data in the SCORE database, the profile of an athlete.

Another exercise carried out based on the SCORE database was related to the number of entries in the 2020/2021 season (the time in which the analysis was carried out), as well as for previous seasons. Therefore. Questions were answered regarding the number of entries by level, by Regional Sports Association, by Nationality, by age. As well as, regarding the Regional Sports Associations and the member clubs of each one of them, the conclusions being regarding the total number of teams, levels, the total number of athletes per club.

With data related to the National Teams, both in Football and Futsal, another exercise was carried out. Thus, questions were answered regarding the clubs represented by the selection, the number of players called up per season to play for the national teams, the number of players who played in certain events and which ones. To complement this analysis, data and tables relating to the games were also used, the so-called Game Sheets, where it is possible to access information about the event, namely the games that took place and when, how many games took place per competition, how many teams played, what warnings there were. Therefore, it is necessary to access the game sheets to be able to say which players were part of the national team in the games.

3. Presentation Creation. In this task, the intention was to be familiarized with performing data presentations after the analyses. After receiving data from a survey already carried out on the visit of supporters to the cities where the 2020 European Football Championship would be held, namely Budapest and Munich, they were summarized using graphical views and tables for the presentation of sample values, particularly about how many supporters would travel to see these games, whether they would tour the cities in question, how they would travel to them, on what days would they travel, where they would leave (Lisbon or Porto). These data also served to collect information for the possibility of holding meeting points in the cities in question, or whether it would be important to create a space to celebrate only with Portuguese people. In this way, this exercise served to consolidate and learn how the FPF worked in terms of synthesizing data and presenting them, however, in Power Point format.

4.2 PORTUGAL FOOTBALL OBSERVATORY

4.2.1 Women's Football

"How to recruit more and retain better in Women's Football?" (Portugal Football Observatory, March 2021)

Within the scope of the PFO, a study on women's football was developed in which the main point of investigation and the big question was 'How to retain and raise more in Women's Football?'. This study aimed to understand which factors were preventing or leveraging the growth of the sport, to women's, in Portugal. As the main motivations for the growth of the sport, there are certain points that try to promote the increase of female football practitioners, teams and professionals, as well as fight and understand the reasons that lead to the retirement. Other points in view to increase the value of the sport in Portugal are faced with the attempt to increase the representation of women in the world of football, by increasing the commercial value, gaining and improving the already existing image of the sport.

Women's Football in Europe: Benchmark France and England

One of the works done during this project, and in an attempt to understand the vision and objectives for the growth of women's football in the main European federations, it was carried a benchmark where the main focus were the French (FFF - La Fédération Française de Football) and the English (FA - The Football Association). To carry out this analysis, proposals from Spain (RFEF – Real Federación

Española de Fútbol) and UEFA were also analyzed, namely the plans and projects for the future of women's football, in order to compare and understand the European vision for the growth of the sport in the female aspect. In this way, it was possible to verify that the main objective is to promote growth through its promotion, as well as through the guarantee of better conditions and opportunities. Another point that became known was a growing attempt to increase the representation of women in football, as athletes, directors, coaches, referees, presidents, and representatives of Football Federations. The analysis was made throughout research of the programs and goals proposed by the federations for the next years, available on the Federations websites.

Market Studies

For the market study, data whose time interval was 2010/2011 to 2019/2020 seasons were used. The current season (having into account the season were the study and the internship was carried out), 2020/2021, was not counted as it was still ongoing during the study, which would bring inconsistencies when comparing with the other seasons, and also because it is the season of the pandemic, what in a certain way doesn't retract the reality of a normal season.

For this part of the project, several surveys were carried out, namely to three large groups: Girls who like football over 13 years old, practitioners or not practitioners, federated or informal football practitioners; Parents of young individuals aged between 5 and 19 years; Football clubs with women's football teams at different training levels.

The surveys were released in 3 different ways: Facebook with the intention of obtaining a more comprehensive and diversified sample, even though there was made a targeting based on Facebook interest; via email to the federated athletes listed in the FPF database; and via direct contact with the Football clubs. All these ways of dissemination were intended to ensure that the sample for analysis was sufficient to make segmentations possible.

"Girls who like Football"

This survey, 'Girls who like Football' was carried out to girls older than 13 years old, in a way to understand the behavior and how engaged they are with the modality. For that, the survey was divided into two different analyses, creating at this way a segmentation between the athletes that are federated (formal football) and the ones who play football as a form of leisure (informal football). This in a way of understanding the existing patterns inside these two groups, the reasons why they play informal or formal, what are the biggest difficulties, what would make them continue or retired from practice football.

Within formal football, there was a segmentation between the athletes of the BPI League (It is the Women's National Football Championship, the main level of the female football league system in Portugal) and the athletes of the other leagues. This segmentation served to understand the real difficulties of women's football in the training levels and in the lower leagues without benefits. As well as a way to understand the main differences between women's football and men's football. This because the lower leagues have less sponsorship, conditions, basically have less opportunities than men's football.

The main survey questions and answers analyzed in this study, and which were later published on the PFO website, were related to three different groups, namely: about playing football, about

dropouts from football, and about informal football. As well as to see if the questionnaire's answers had a real sample compared with the data from the databases, personal questions related to gender, age group, location, among others, were also asked.

Thus, and referring to the first group 'About Playing Football', where the main focus was to understand what the motivations are for playing football, in which the most important questions, and the ones used in the study related to this point were:

1. To the question 'Why did you start playing football (Top3)?', the most answered responses remained that they practice football because they like, because they are good at practicing football, and lastly because they like to compete.
2. To the question 'How did you become Federated?', the answers were that they were recruited by a club, they search for a club that already compete, and because of some influence of the people closest to them.
3. To the question 'How did you start playing football (Top5)?', the most chosen answers were that they start playing football with friends, in the school recess, due to of family influences, in the physical education classes, and because they signed up for school sports.

Even with the positive evolution of the sport's growth in Portugal, it was also necessary to enter the factors that lead to the athletes' withdrawal, as well as the main barriers to the continuity of the practice in the future. Thus, the following questions were asked for the group 'About Dropouts from Football'.

1. To the question 'Why did you stop playing football (Top3)?', the most chosen responses were that they dropout due to the incompatibility of schedules, because they are going to continue their studies, and also because of the difficulty of finding a team.
2. To the question 'What is the main reason for you to abandon the practice of football in the future (Top3)?', the answers were due to the fact that the plans to be a football player as a career are not part of the future, because they are going to continue the studies, and also due to the incompatibility of schedules.
3. To the question 'What time does the train starts?', the three hypotheses of responses were 6pm/20:30pm, 8:30pm/22pm or 10pm/more.
4. To the question 'Why are studies an impediment (Top3)?', the most answered responses were that it is not part of the plans to be a football player as career, why will you change residence to study, and because they want to keep the current club.

Related to the practice of informal football, which corresponds to non-federated football players, and in this way to understand what factors exist to make it difficult, or so that they are not playing in a club, or even just playing. Thus, to understand these points, the following two questions were asked:

1. To the question 'Why don't you play in a club (Top4)?', the most chosen answers were due to the difficulty of finding a team, because they do not feel prepared to practice in a formal way, it was necessary to spent too much time on the practice, and due to schedules incompatibility.

2. To the question 'Why don't you play football (Top4)?', the most chosen answers were the difficulty of finding a team, because they appreciate to watch games but not to play, because they are not good at practicing, and because they practice another sport.

Parents

This survey was carried out to Parents whose children are between 5 and 18 years old, that were asked about how Football practice is viewed, in order to understand the motivation given the parents regarding the practice of the sport, or why they don't motivate the children to practice.

When analyzing the data collected, it was necessary to segment them according to their gender and age, dividing this into two intervals each group, if the children are between 6 and 12 years old, or if they are between 12 and 18 years old. By analyzing the results obtained, it was possible to see that the sample for each of the clusters were not significant in terms of the number of responses, as well as that there were no major differences in terms of age intervals. A second segmentation was then performed based only on the gender of the children, in a way to improve and guarantee that the sample have a significant size to represent the reality. The main conclusion obtain with this analysis was that parents tend to motivate more the boys than the girls to practice football, so the girls only practice if they want to do it, and that doesn't append to boys when they are young.

Having the previous analysis, the following question was used and published for the study: 'What are the main reasons to encourage (TOP3)?'. Following the question, the three most answered responses, respectively, were that practice team sports, like football facilitates socialization, it is a healthy practice, and lastly the children enjoy and have fun practicing football.

Football Clubs Case Study

39 clubs that have or do not have a women's football training team were questioned so that in this it is possible to understand from the ones who are responsible for training the athletes what the main difficulties are, what factors they think can promote the growth and evolution of women's football in Portugal. They were asked about how the project in the various clubs emerged, how the process of recruiting athletes is carried out, what reasons in their eyes are the reason for the success in the growth of the sport, how they overcome the difficulties that are often related to infrastructure and training schedule, and what are the main differences between women's and men's football in their view. The great conclusion reached out with this small study was that most clubs have as main objective to promote the growth of the modality by trying to create more training levels, as well as finishing earlier in the age of athletes with mixed practice (teams of boys and girls). And to create and promotes, as much as possible, equality between women's and men's football, in terms of sponsorship, conditions, professionalization, among other points.

From this study, a football club was chosen to work as an example case study, in order to represent the other clubs that are also an example of a significant growth in the formation of women's football. In this case, the selected club was Grupo Desportivo da Ilha, due to the gross growth it had from having only 6 athletes in 2018/2019 and 26 athletes in 2019/2020, to 45 athletes in 2020/2021. This is a great example that the measures and project apply to promotes the growth of women in the football world.

Methodology

Much of the exploratory analysis of the market studies was carried out in Excel, namely the Girls who like football and the Case study of the clubs. Being that of the parents, the views were created in Power BI. Another tool used throughout this task was the SQL, for the collection of complementary data to those obtained through the questionnaires. The functions used are the ones described previously in the point 3.1 'Tools and Technology' of the report.

With regard to the questionnaires, one of the tasks that also needed to be carried out was their validation. In other words, it was necessary to check before its publication so that the public could respond correctly, if they were according to the proposed, if all the paths followed according to the plan. Considering that, due to some choices, the path changed, as well as if the multiple choices worked in the right form, if they were all the questions and if they were according to the type of answer intended (select, open field, slide stream, etc.). The surveys were published in the questionnaires were published through the company SurveyMonkey.

4.2.2Futsal

The evolution of Futsal in Portugal. And the victory in the 2018 European Championship as an accelerator for growth?" (Portugal Football Observatory, April 2021)

Within the scope of the PFO, the second study carried out analyzes 'The evolution of Futsal in Portugal and the victory in Euro 2018 as a growth accelerator'. The main focus of this study was to analyze the evolution of the growth of futsal in Portugal from the 2010/2011 season to the 2019/2020 season. The study focused largely on the existing perception that the fact that the National Futsal Team won the title of European champion in 2018 is one of the main levers for the existence of an above-average growth.

Market Studies

For this part of the study, three different surveys were carried out, two of which had the same structure and were only sent to different targeting groups. This survey essentially served to understand what kind of athletes they were, whether formal or informal, what reasons they could give up, why they started practicing, and also what impact the European title had. One via Facebook in order to capture a broader and more diverse audience. The other was sent directly to athletes listed as federated in the FPF database. The third survey had a different structure, as it sought to understand the reasons for the withdrawal by the former athletes that counted in the database, for this purpose, a targeting of athletes who were no longer registered as federated was carried out.

All the data collected throughout the three surveys was merge into one single table, and then was created a targeting of responses that had as main objective to ensure that there was a sufficient sample for each of the clusters that were formed after a segmentation of the collected data. The three analysis groups/clusters were: 'Yes, I practice', 'Yes, I've practiced' and 'I don't practice/practiced'.

A question and limitation that arose in this study was the fact that the database created through the responses to the surveys did not correspond to the reality existing in the FPF database. This would promote a fallacy when analyzing the results, as it did not correspond to reality mainly in terms of

age and gender distribution. So, the solution was to carry out ponderations (3.1 and 3.2), which corresponds “re-balancing the data in order to more accurately reflect the population (DataStar, 2013)” in order to assign different weights to each answer depending on gender and level (education or senior).

$$\text{Ponderations} = \text{Database Distribution} / \text{Survey Distribution} \quad (3.1)$$

$$\text{Adjusted } N = \text{Survey Distribution} \times \text{Ponderation} \quad (3.2)$$

These questionnaires, as well as the different target audiences that were reached, have as main objective to complete the current photography of the sport in numerical terms, in order to understand the representativeness of School Sport before federated athletes. And also to understand the main reasons for giving up for the future of current athletes or what would be needed to guarantee their permanence. Another point of study based on the questionnaires was the impact of the Euro 2018 victory, in order to understand the motivation promoted by the victory and how it changed expectations of having a career as a professional futsal athlete.

As in the study related to women's football, questionnaires were carried out, however only a few questions were published in the Futsal study in the PFO. Thus, the following questions were asked and published in order to better understand how the perception is, and even the image of futsal in the eyes of those who practice and enjoy the sport. The questions are as follows:

1. To the phrase/question ‘Representativeness of federated athletes in School Sports’, it was made a distinction between whether the athletes play or not at the same time in school sports and in a club, or if they only play in one of them. This distribution was also divided into four levels of training: Under15, Under17, Under19 and Senior.
2. To the question ‘What would you do not quit futsal within 2/3 years (Top5)?’, the most answered responses were the competition, playing with friends, the healthy lifestyle it promotes, the National Futsal Team, and also if they have a career perspective.
3. To the question ‘What would you do not to leave futsal next year (Top5)?’, the most considered answers were if they have better conditions, if the professionalization was easier, if they can reconcile schedules, competition, and if they find a club.
4. To the phrase/question ‘European Title and Motivation’, it was necessary to choose between 1(Do not agree at all) to 5(Agree entirely) to the sentence. It was divided by gender, and if they were registered if they were registered in the modality before or after the title.
5. To the phrase/question ‘European Title and Expectations’, it was necessary to choose between 1(Do not agree at all) to 5(Agree entirely) to the sentence, and if they were registered if they were registered in the modality before or after the title.

The exploratory analysis of the market studies was carried out in Excel. Another tool used throughout this task was the SQL, for the collection of complementary data to those obtained through the questionnaires. The functions used are the ones described previously in the point 3.1 ‘Tools and Technology’ of the report.

Considering the surveys, one task that was needed, was their validation, as well as in the Women's Football Study. In other words, it was necessary to check before its publication so that the public could respond correctly, if they were according to the proposed, if all the paths followed according to the plan. The surveys were published in the questionnaires were published through the company SurveyMonkey.

4.2.3 Relative Age Effect

“Does the time of the year in which one is born condition the opportunity to be a grassroots athlete?” (Portugal Football Observatory, May 2021)

In the PFO, a study was carried out that answers the question 'Does the time of year in which you are born affect the opportunity to be an athlete by training?'. The main objective of this study is to understand the effect of relative age (RAE), which corresponds to the phenomenon that the month of birth can influence the athlete's ability, since 'Those born early in the selection year (e.g. January) are relatively older—by as much as 12 months minus 1 day—than those born later in the selection year (e.g. December) (Wattie et al., 2008). Which in a way causes players to be categorized according to the month of Birth (Barnsley et al., 1985).

Project

During the project, as a way of initiation, it was necessary to extract and collect data from several databases.

Initially, a list of existing clubs was carried out, making a segmentation between Football and Futsal, as well as according to gender (Women and Men). This list took into account the certification of the clubs, that is, it works as the attribution of a grade to the training schools of the sport, in order to guarantee excellence. At this stage of the work, a merge was carried out between two tables, one of the data from the FPF Database, in which the extraction was via SQL and the other table of clubs depending on the certification. For this task, the VLOOKUP function of Excel Microsoft Office was used. There were some exceptions where the merge was not done directly, having to be done manually. Exceptions that were due to abbreviations of club names in one of the sources.

The other part that was carried out at this point of relative age was the collection of data related to women's Football athletes through SQL queries, based on the month of birth of each athlete. This research and data extraction was related to four points, namely: 1. General distribution of births from 2011 to 2019 over the four quarters; 2. Types of schools taking into account the certification (could be Training Entities, Football Schools, CBFF-Basic Football and Futsal Training Centers); 3. Training level; 4. District. At all these points, the distribution of births was evaluated according to the question under analysis. That is, and as a demonstrative example (Figure 19).

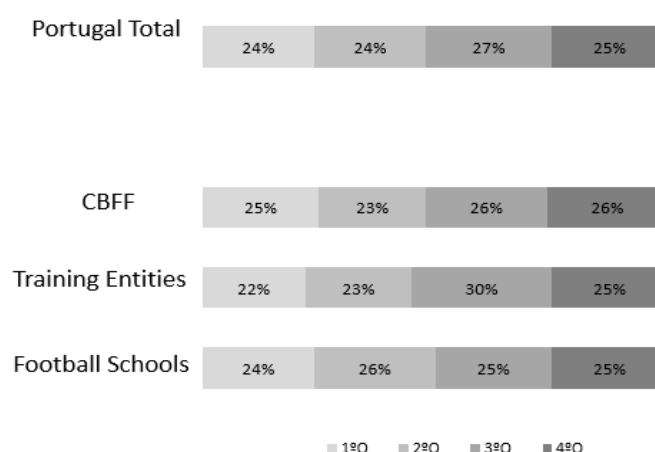


Figure 19: Distribution of the number of players of each type of school by birth quarter.

4.2.4 Replacement Minute Impact

This study/analysis is about the impact of changing the number of players substitutions in the Portuguese league, so the change during the game went from 3 to 5 substitutions for each game, however the game can only be interrupted 3 times. This change was due to the impact of Covid-19 on players' performance, which somehow caused excessive fatigue. And by being able to increase trades throughout the game, it can make performance better.

In order to carry out this study, it was necessary, in a first analysis, to collect data on substitutions in the main European leagues, as a way of comparing the different championships. The Leagues studied were: Liga NOS- Portugal; La Liga-Spain; Bundesliga-Germany; Serie A-Italy; Ligue 1-France. The data collected corresponded to three seasons: 2020/2021; 2019/2020; 2018/2019 (in this last season there was a division between the pre and post pandemic championship). The data was collected in the external Database OPTA, previously described.

The collected data were organized in an Excel table in which each line corresponded to a substitution, while the columns were the game in question, the team, the player substituted, date, minutes played, goals scored and assists. As well as several sheets were created corresponding to the league in question and at the time of the data. The functions used in the Excel are part of the ones described previously in the point 3.1 'Tools and Technology' of the report.

4.3 INTELLIGENCE AND CUSTOMER SERVICE DEPARTMENT

4.3.1 Covid-19 Impact

This analysis appears as a way to understand the impact that the Covid-19 pandemic situation had on Training Entities. For this analysis there was a sample of 96 Training Entities, in order to carry out a comparative analysis on some key topics. The main points go through two large groups:

1. Income: This point is related to the continuation, or not, of obtaining monetary income through the athletes' monthly fees, club and local government contributions, sponsorships and other sources of income.

2. Costs: This point is related to the costs that the Training Entities had during the period of the pandemic. These costs have to do with Club personnel, transport, food, infrastructure, factors related to Covid-19, and other operating costs.

The analysis had as main considerations: the number of practitioners, evaluating their growth or not; the monthly fee, evaluating if there was a reduction in its cost, or even if the charge was maintained; the income and income obtained by the Training Entity; and the costs borne by the Training Entity.

The presentation of results was divided according to the type of training entity, considering its certification. Thus, there are three groups: Training Entity (holders of 5, 4 or 3 stars); Training Schools (2 or 1 star holders); CBFF. According to this division, the sample for each of them was 42 Training Entities, 9 Training Schools and 7 CBFF.

For the analysis, the tool used was Microsoft Office Excel, and to report the conclusions the tool used was Microsoft Office PowerPoint.

4.3.2 Databases Analysis

In a way to understand how the figure of the FPF databases is, and to visualize the engagement which several points of gains news database members, as well as buyers, fans, players, and another. And in a way to evaluate the existing growth.

For that, the analysis of the databases works by taking into account the databases previously described in this report in section 3.2- Data, namely in the subsection 'Internal Data Sources'.

In the first part, the task consists of an analysis of the users of all existing FPF databases (Portugal +; Portugal Store; SCORE; Ticket). In the other part, a segmentation of database users was performed, taking into account only the database users since the 2018/2019 season. In this segmentation, data related to eFootball were also analyzed.

To perform this analysis, data were extracted using SQL queries. And later treated in Excel, as well as the visualizations of them. These views were intended to show the value of the data, namely the following variables: Existing data; Number of users; Nationality; National or International; Genre; Age range (13-17, 18-24, 25-34, 35-44, 45-54, 55-64, 65+).

4.3.3 Analysis of the growth of federated athletes

In this task, the main point was to assess the evolution of federated football players in Portugal over the seasons. Having been counted 9 times: 2011/2012, 2012/2013, 2013/2014, 2014/2015, 2015/2016, 2016/2017, 2017/2018, 2018/2019, 2019/2020.

In this analysis, based on the total number of practitioners in each season, the growth rate (3.3) in each season was calculated. This rate was calculated using the formula:

$$\text{Growth Rate} = \frac{\text{Old Year} - \text{New Year}}{\text{New Year}} \quad (3.3)$$

For the demonstration of the results, these were presented in the form of a bar graph (Figure 20), with the indication of the number of athletes in each one of the seasons and the respective growth rate, between each one of the bars that corresponds to a season in question.

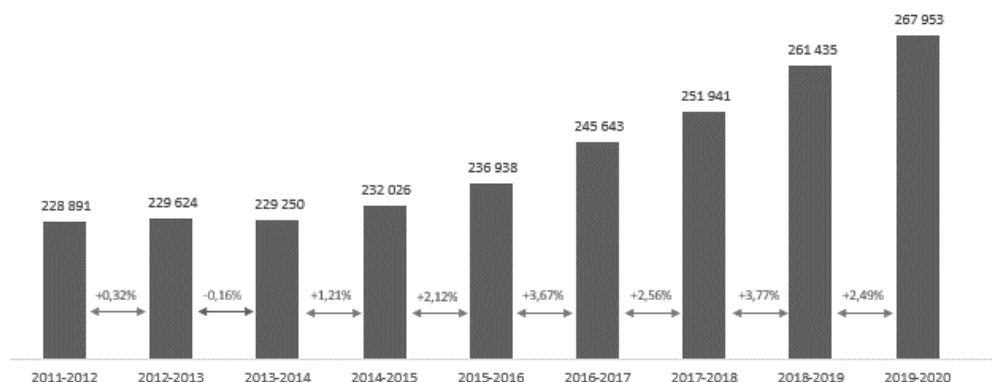


Figure 20. Evolution of federated football players in Portugal over the last 9 seasons.

Data were extracted from the FPF database through SQL queries, namely from the SCORE table (previously described in point 3.2 of this report). Its treatment was carried out through Power BI.

4.3.4 Listing of Football Associations for financial support

During the pandemic situation, Football Associations suffered in terms of support and monetary income. In this way, the FPF, together with the Portuguese State, decided to create a fund of financial support for the sports, divisions and levels that suffered most from the stoppages in competitions. The amount of support to each club and team varied according to three different groups, two of them for the Football modality (1st Groups - 1st District Division, Honor Division and Pro-National; 2nd Group - 2nd and 3rd District Divisions) and another for Futsal.

There are 18 Football Associations in Portugal, from the following Regions: Algarve, Aveiro, Beja, Braga, Bragança, Castelo Branco, Coimbra, Évora, Guarda, Leiria, Lisboa, Portalegre, Porto, Santarém, Setúbal, Viana do Castelo, Vila Real and Viseu. About 1047 teams were considered for financial support, of which 167 were training and 880 were senior teams.

The task of this analysis consisted of compiling and organizing all the data sent by the Football Associations. As well as performing the comparison of the data sent with the data previously collected through the FPF Database through SQL. This comparison appears as a way to understand the differences, and in a way to verify whether the data sent by the Football Associations were close to reality, as cases were verified where the teams were not registered at the time most affected by the pandemic, or they did not participate in competitions, therefore they were not eligible for financial support.

In this way a single list was created where the lines were the teams, and the columns were the groups and subgroups previously described. The columns were made binary variables, which later

facilitated the analysis and counting of how many teams existed in each of the groups. This list was created in Excel, as well as the analysis and summary accounting of each variable.

4.3.5 Audiences LIGA NOS and SuperLiga

An analysis was carried out as a way of understanding, and in a way trying to prove the theory that the games between the main teams of the European leagues had a larger audience compared to the other games. This analysis then appears as a way to understand the reality behind this theory, and in this way the data collected on the Superliga were compared with the Liga NOS.

My work in this analysis consisted of collecting data related to the games played, date and time of the game. And in the case of data related to the Superliga, information regarding the competition in which the game was played was also collected. Data extraction was divided into two distinct parts, one relating to each of the leagues under study for the 2020/2021 season.

The first data collected were from the Superliga, which is a project in which the main European clubs join: AC Milan, Arsenal, Atletico Madrid, Chelsea, Barcelona, Inter Milan, Juventus, Liverpool, Manchester City, Manchester United, Real Madrid and Tottenham. These data were extracted from the Transfermarkt.pt website. For this part of the task form collected information about 55 games between the teams above.

The second data are related to the main Portuguese league, Liga NOS (league name at the time of the analysis), during 28 rounds out of 34 (days played at the time of analysis). The clubs whose data relate to, are: Sporting CP, FC Porto, Benfica, SC Braga, FC Paços de Ferreira, Vitória SC, Santa Clara, Moreirense FC, CD Tondela, Portimonense, Belenenses SAD, Gil Vicente, FC Famalicão, Rio Ave, Boavista FC, Marítimo M, SC Farense and Nacional. These data were extracted from the website liganos.pt, the official website of the competition. For this part of the task, information was collected on 261 matches between the teams.

All data were just compiled in an Excel file, as my task was only to collect data from different external databases.

4.3.6 Portugal Store Stocks Management

Analysis

The Portugal Store was created in 2018, however its products are stored (on pallets or shelves) in the operational logistics centers of CTT (Correios de Portugal – Portuguese Post Office), the same center that is responsible for distributing the products. In this way, all the analysis carried out aims to understand the existing costs for the FPF in relation to the storage of products, as well as, in a final part, the calculation of the profit margin of each product was carried out. This analysis also seeks to understand if the revenue obtained compensates for the storage costs.

In the first part, a compilation of two tables from different sources is carried out, one table provided by CTT and the other that corresponds to the current stock of the store's products, where there is also a reference to the sales made (monthly file), as well as the difference in the stock between the beginning and end of the month. The compilation between these two tables is performed by product reference, however all the analysis is performed considering the variable ID_SKU (product identification number). In this initial part of the analysis, it was possible to verify the existing stock,

how many sales were made and where these products are stored. The data were processed in Excel, and to carry out this compilation the VLOOKUP function was used for the direct compilation between the data from the two files. In this way, and based on the data collected and compiled, it was possible to calculate the storage costs per day, month, and year for all products in the store. These costs varied depending on the storage location.

When analyzing the storage costs for all products in the store, a segmentation was made to carry out a more specific analysis for certain items. The first segmentation performed was a cost analysis for the oldest products in the store, but with a high number of old products in stock and few sales, so 30 items were selected. The objective of this segmentation was to understand which products have been in stock for the longest time, in order to increase their sales, or even to donate them, since they only increase the store's costs when they are only stored.

The second one was an evolution of all scarf sales per month since the creation of the store (March 2018) until now and was just to comprehend how the sales of scarfs were doing during the years. This analysis had as main reason the fact that scarfs were the product with the biggest amount in stock and less sales, so in this way it is possible to understand the sales of this products.

The last segmentation concerned a division of the analysis by the different existing special collections, namely: POR Collection, NIKE Products, Legends Collection and Christmas Collections. This segmentation Was just to understand how the sales and associated costs of the store's main collections were doing.

In a second part, all the initial analysis carried out, as well as the subsequent segmentations, were complemented with the calculation of the profit margin (3.4 and 3.5). The formulas used to calculate the profit margin of the store's products were:

$$Profit = Income - Production Cost \quad (3.4)$$

$$Profit Margin = \frac{Profit}{Income} \times 100 \quad (3.5)$$

Thus, the variables taken into account to perform the final analysis are the following: ID_SKU, Reference, Product Description, Sales, Stock, Production Cost per product unit, Production Cost per day, Production Cost per month, Production Cost by Total Sales, Storage Costs by Product Unit, Storage Costs per day for the total quantity in stock, Storage Costs per month for the total quantity in stock, Storage Costs by total sales, Costs of Media Support, Total Costs (Production Cost per Total Sales+ Storage Costs per Total Sales+ Media Support Costs), PVP excluding VAT (Sales price excluding tax), Revenue, Profit and Profit Margin.

A temporal analysis of costs and sales was carried out, working as an evolution from 2018 until the date of the analysis. In this case, a table was created in which the lines are the years (2018, 2019, 2020 and 2021) and the columns correspond to the Production Cost, Storage Cost per day, Media Support Cost, PVP without VAT, Total Costs, Revenue, Profit and Profit Marge. As a result of this

analysis, the totals for the 4 years for Total Costs, Revenue and Profit were calculated, as well as the total Profit Margin since the beginning of the Portugal Store was calculated.

Dashboard

In an attempt to make the data easy to access and understand for the entire organization, a Dashboard was created with the data previously collected and processed. This because a dashboard works as a page where the information is presented objectively throughout visualizations with the purpose of telling the story of the data, containing only the most important information.

In this way, several metrics and views were created for them. And for the creation of the dashboard, one of the first points to consider was the perception of where the data came from, as it was signed in order to achieve direct automation when there is new data in the database. Just as it was necessary to merge some data in order to be able to have the maximum amount of information essential to the analysis and from these data.

The Dimensional Model created for this dashboard was a snowflake schema, composed by five tables (Figure 21), were the Fact Table 'Vendas_Detalhe' represents the sales of the store's products in detail, mostly information related to the quantity sold per purchase. Then, there is an existence of four Dimensional Tables. The 'Produtos_Stock' Table corresponds to the number of products in stock per ID_SKU (identification number of the product) and the respective description. The 'Vendas_Portugal' corresponds to sales details in terms of dates relative to the time period in which the purchase was made. Thus, and so that it was possible to perform a temporal filter, by semester, quarters, month, among other filters, it was necessary to attach the table 'Ano_Mes_Tempo'. The last Dimensional Table used was the 'PVP_Todos', which is related to the prices of sale to the public of the products, where it is possible to find values with or without taxes.

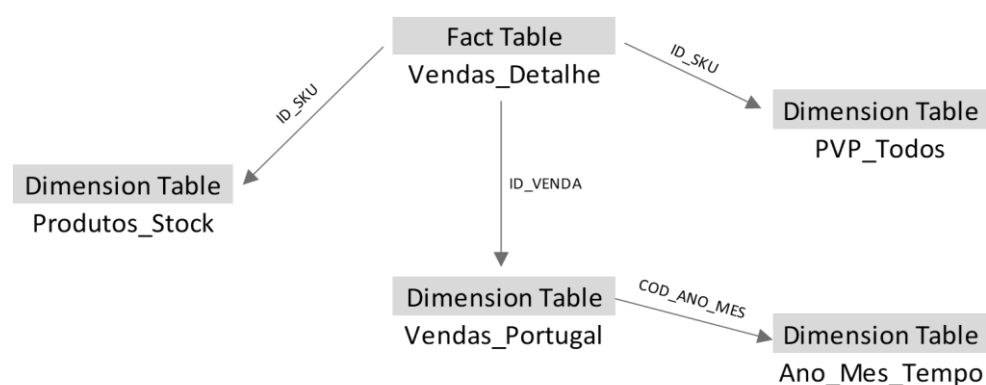


Figure 21. Dashboard Dimensional Model.

Since the main goal of creating a Dashboard was to make the data easily understandable, the visualizations were also simple and straightforward. Thus, it was necessary to take some points into account, namely the fact of the importance of choosing the right type of views and also the fact that they should be simple and direct at the end, because the entire dashboard must focus on the user and the target public, therefore in a dashboard you must consider the fact that the data must be

accessible immediately. Another point to consider is that the visualizations are intended to give a detail and knowledge only in the title, so that the understanding itself is clear to the user (Magalhães, M.S., 2017).

Accordingly, to the information to considered above, in the table below (Table 1), there are a list of the type of visualization used and what data it represents.

Visualization Type	Data and Metrics
CARDS	ID_SKU
	Amount of Products sold
	Production Cost
	Income (no Taxes)
	Profit (Production Cost + Income)
	Storage Costs per day
	Storage Costs per month
Filters	Year
	Month
	Semester
	ID_SKU
Charts	Evolution of the amount of sales per year
	Evolution of costs, revenues and profit per month
Table (Data)	ID_SKU
	Production Cost
	Price (no Taxes)
	Profit (Production Cost + Income)
	Amount of Products sold

Table 1. List of the Dashboard Visualizations.

However, and until the time of completion of the internship, the design of the Dashboard was the following (Figure 22):

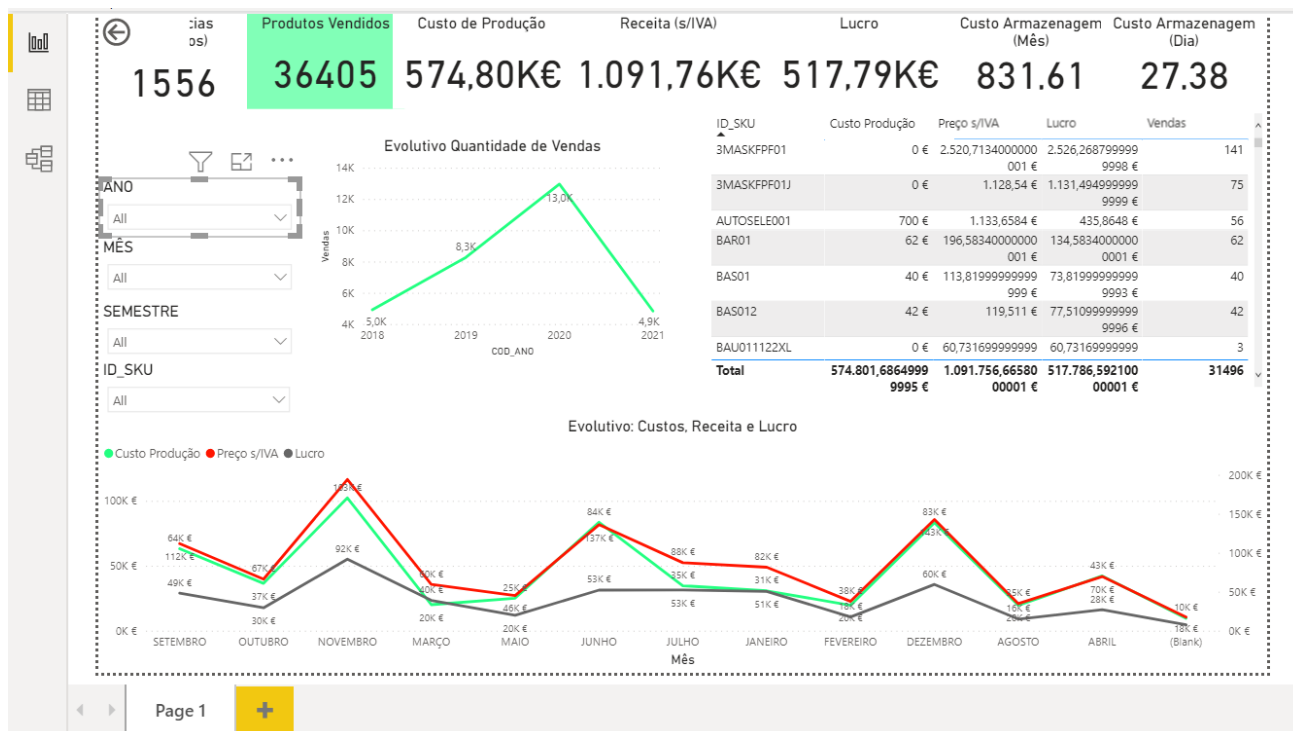


Figure 22. Provisory Dashboard Design.

It was not possible to complete this task in full, as the internship period ended. But for future work it was left to do: Automate the data so that Dashboard updates are fast and automatic; Placing the FPF template on the Dashboard.

5. CONCLUSION

This report had as its main objective to understand the importance of data analysis in the industry of sport, essentially in what concerns to a sport organization as a Football Federation, in this case the Portuguese Football Federation (FPF) where the internship was developed during approximately 8 months.

In this way, the internship was part of two projects, at the Portugal Football Observatory and at the Intelligence and Customer Service Department, for which several analyses were developed. As for the PFO project, 3 published studies were developed, and a study that, during the internship, was in the creation process. Being the studies about the growth in Women's Football, about the impact of the futsal national team to have been European Champion in 2018, the impact of the Relative age effect in training level, and the last study was about the replacement minute impact during the Football games. For the Intelligence and Customer Service Department several ad hoc analyses were developed, namely about the Covid-19 impact in the Football practice, an analysis concerning the FPF Databases, about the growth of the federated athletes during several seasons, regarding financial support for Football Associations, about TV audiences of the biggest Football leagues, and the last performed analysis was about the manage the stock of the Portugal Store and also about studying the profit margin of the products.

CONNECTION TO MASTER'S PROGRAM

In the Master's program of studies in Data Science and Advanced Analytics there is a high focus regarding the treatment of large amounts of data, in the strategic and decision-making process, statistics, Data Mining, Machine Learning, as well as creation processes and maintenance of predictive and descriptive models. All of these processes have the main objective of transforming data into information that is of value to the organization.

In the case of my specialization, Business Analytics, there is an orientation towards a more strategic and management aspect of companies and organizations. This will allow you to be able to analyze data and make decisions in order to optimize business processes, as well as obtain more information and better and structured data outputs.

As a way of consolidating the knowledge learned throughout the master's degree, as well as completing it with new visions and learning, the internship was a crucial element. Just as the connection between the internship and the master's degree is important in the sense that it gives an idea of what the application reality of knowledge is as a data analyst in an institutional context.

Throughout the internship, the curricular units that were most important for the internship were: Business Intelligence; Digital Transformation; Business cases; Data Mining; Data Visualization.

INTERNSHIP EVALUATION AND RESULTS

Positive Points

The internship functioned as a great learning port, both personally and professionally. Insofar as it was the first contact with the data analysis area outside the faculty, it was possible to understand

and put into practice many of the knowledge acquired during the Master's, as well as to learn even more.

Having acquired skills divided between hard and soft skill, described in the following points.

Soft Skills

During the internship, and with the development of the various projects, new capabilities were acquired, or even improve some of them. Thus, and because it was my first professional experience in the area, and because they were not just college projects, but with real importance for the organization, I ended up gaining a sense of responsibility that I didn't have until then, since small mistakes can compromise all the work of the organization. Which in a way leads to another point, which was better organizational skills since most of the work was done in a team.

Other points I developed and acquired during the internship was gaining agility in learning, this because even with the knowledge from the master's there was always point, subjects and programs that I had to learn from scratch, as well as I had to learn in different ways, either from self-taught manner or training of my advisors. I can also say that my communication skills have increased considerably, due in large part to the teamwork developed.

In this way I can emphasize that I acquire essential skills and abilities for my future, both personally and professionally.

Hard Skills

In a more technical and practical aspect of the internship, there were three factors that I had to learn to manage during the internship, which were: Time Management, Data Management and Project Management.

As for data management and what it encompasses, the more specific ways in which I applied this ability that I was acquiring was through the collection of information and data from different sources and forms, through: Questionnaires, where I realized that the simplification of questions, like this how placing choice options instead of open fields makes future work in terms of analysis immensely easier; Both internal and external databases; Excel files; Literature reviews, including documents from other football federations or institutions.

In a way, time and project management are intertwined, as for the project to be well managed it was necessary to have excellent time management and organization.

In terms of more technical skills and working with data and information, I deepened my knowledge in data collection and analysis programs, namely in SQL, Excel, and Power BI.

Negative Points

The internship was carried out during a pandemic and confinement period, which allowed it to be fully teleworked. The fact that it was carried out under these circumstances ends up somehow compromising direct contact with my colleagues, ending up making the whole process more impersonal. As well as all interactions, meetings, and explanations were all carried out via Microsoft Teams, which somehow ends up hindering the fluidity of the work, or even teamwork.

BIBLIOGRAPHY

- Bai, Z., & Bai, X. (2021). Sports Big Data: Management, Analysis, Applications, and Challenges. In *Complexity* (Vol. 2021). Hindawi Limited. <https://doi.org/10.1155/2021/6676297>
- Barnsley, R., Thompson, A., and Barnsley, P. (1985). Hockey success and birthdate: the RAE. *Can. Assoc. Health Phys. Educ. Recreat.* 51, 23–28.
- Caya, O., & Bourdon, A. (2016). A framework of value creation from business intelligence and analytics in competitive sports. *Proceedings of the Annual Hawaii International Conference on System Sciences, 2016-March*, 1061–1071. <https://doi.org/10.1109/HICSS.2016.136>
- Chapman P., Clinton, J., Kerber R., Khabaza T., Reinartz T., Shearer C., & Wirth R. (2000). *Step-by-step data mining guide*. DaimlerChrysler.
- Chen, H., Chiang, R. H. L., Storey, V. C., Lindner, C. H., & Robinson, J. M. (2012). *Business Intelligence and Analytics: From Big Data to Big Impact Quarterly-Business Intelligence and Analytics: From Big Data to Big Impact* (Vol. 36, Issue 4).
- DataStar (2021). Weight, Weight... Please Tell Me! (Blog Page) Retrieved 6 October, 2021, from: <https://www.surveystar.com/startips/apr2013.pdf>
- Dobraja, I., & Kraak, M. (2020) Principles of dashboard adaptability to get insights into origin-destination data, *Journal of Location Based Services*, 28-48. <https://doi.org/10.1080/17489725.2020.1738577>
- FPF: Federação Portuguesa de Futebol. (Web Page) Retrieved 27 May, 2021, from: <https://www.fpf.pt/pt>
- GooglePlay, Microsoft Excel: Ver e Editar Folhas de Cálculo (2021, 24 November) Retrieved 8 July, 2021, from: https://play.google.com/store/apps/details?id=com.microsoft.office.excel&hl=pt_PT&gl=US
- Haaren, J. (2021, 17 March) State of the Football Analytics Industry in 2021 (Web page) Retrieved 18 November, 2021, from: <https://www.scisports.com/state-of-the-football-analytics-industry-in-2021/>
- Han J. Pei J. & Kamber M. (2011). *Data Mining Concepts and Techniques*. Elsevier
- Instituto Nacional de Estatística/Statistics Portugal (Web Page) Retrived 20 July, 2021, from: https://ine.pt/xportal/xmain?xpgid=ine_main&xpid=INE
- Liga NOS (Web Page) Retrieved 20 July, 2021, from: <https://liganos.pt/>
- Magalhães, M.S., (2017, 21 February) Princípios de Desing para a criação de dashboards (Web Page) Retrieved 3 November, 2021, from: <https://www.linkedin.com/pulse/princ%C3%ADpios-de-desing-para-cria%C3%A7%C3%A3o-dashboards-marcel-magalh%C3%A3es/?originalSubdomain=pt>

- Microsoft, Download SQL Server Management Studio (SSMS) (Web Page) Retrieved 8 July, 2021, from: <https://docs.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-ver15>
- Microsoft, What is Power Bi? (Web Page) Retrieved 8 July, 2021, from: <https://docs.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview>
- Morgulev, E., Azar, O. H., & Lidor, R. (2018). Sports analytics and the big-data era. *International Journal of Data Science and Analytics*, 5(4), 213–222. <https://doi.org/10.1007/s41060-017-0093-7>
- OPTA (Web Page) Retrieved 20 July, 2021, from: <https://www.statsperform.com/opta/>
- Patel, D., Shah, D., & Shah, M. (2020). The Intertwine of Brain and Body: A Quantitative Analysis on How Big Data Influences the System of Sports. *Annals of Data Science*, 7(1). <https://doi.org/10.1007/s40745-019-00239-y>
- PORDATA: Base de Dados Portugal Contemporâneo (Web Page) Retrieved 20 July, 2021, from: <https://www.pordata.pt/>
- Portugal Football Observatory (April 2021). The evolution of Futsal in Portugal and the victory in the 2018 European Championship as an accelerator for growth. (Web Page) Retrieved from: <https://indd.adobe.com/view/7342e002-a699-48ab-95c9-8b6c5c73d4d2>
- Portugal Football Observatory (March 2021). How to recruit more and retain better in Woman's Football? (Web Page) Retrieved from: <https://indd.adobe.com/view/b42a8285-e762-44f1-bb07-bbd3d6643e40>
- Portugal Football Observatory (May 2021). Does the time of the year in which one is born condition the opportunity to be a grassroots athlete? (Web Page) Retrieved from: <https://indd.adobe.com/view/246f47ca-e1d6-4669-9a70-6807b1638d84>
- Portugal Football Observatory (Web Page) Retrieved from: <https://portugalfootballobservatory.fpf.pt/Default.aspx>
- Pratt M. & Fruhlinger J., (2019). What is business intelligence? Transforming data into business insights (Web Page) Retrieved 10 November, 2021, from: <https://www.cio.com/article/2439504/business-intelligence-definition-and-solutions.html>
- Realmonkey Blog, The 10 Principles of Dashboards Design 2021 (Blog Page) Retrieved 3 November, 2021, from: <https://realmonkey.co/blog/principles-of-dashboards-design/>
- Shamoo, A.E., & Resnik, B.R. (2009). Responsible Conduct of Research. *Oxford University Press*. DOI:10.1093/acprof:oso/9780195368246.001.0001
- Transfermarkt (Web Page) Retrieved 20 July, 2021, from: <https://www.transfermarkt.pt/>
- Wattie, N., Cobley, S., and Baker, J. (2008). Towards a unified understanding of relative age effects. *J. Sports Sci.* 26, 1403–1409. doi: 10.1080/02640410802233034

ZEROZERO (Web Page) Retrieved 20 July, 2021, from: <https://www.zerozero.pt/>

