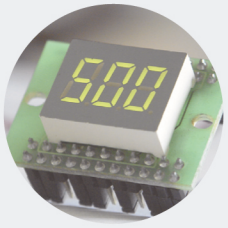
INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA ELECTROTÉCNICA E DE COMPUTADORES






Data Analysis in the BIG DATA scope in Basketball

DIOGO FILIPE PINTO ALVES
setembro de 2022

POLITÉCNICO
DO PORTO

Politécnico do Porto
Instituto Superior de Engenharia do Porto

Data Analysis in the BIG DATA scope in Basketball

Diogo Filipe Pinto Alves

Master in Electrical and Computer Engineering
Specialization Area of Systems and Industrial Planning

Departamento de Engenharia Eletrotécnica
Instituto Superior de Engenharia do Porto

July, 2022

This dissertation partially satisfies the requirements of the Thesis/Dissertation course of the program Master in Electrical and Computer Engineering, Specialization Area of Systems and Industrial Planning.

Candidate: Diogo Filipe Pinto Alves, No. 1081693, 1081693@isep.ipp.pt

Scientific Guidance: Carlos José Campos, crc@isep.ipp.pt

Scientific Co-Guidance: Veríssimo Santos, vms@isep.ipp.pt

Departamento de Engenharia Eletrotécnica
Instituto Superior de Engenharia do Porto
Rua Dr. António Bernardino de Almeida, 431, 4200–072 Porto

July, 2022

All the involved people did part of this research by developing content to understand the concept of large data volume, applying numerical and computer mathematical methods with their support and knowledge.

A big thank you to Doutor Carlos José Campos and Doutor Veríssimo Santos, that provided the opportunity to work on a new subject - Big Data, that will help me, in the future, to understand and analyse data, identify and apply the numerical and computation methods.

A special thanks to Doutor Tenrreiro Machado for guiding and helping me along this path, which he started, and also for being always available to support and help student in too many scopes.

Also and not forgotten, a thank you to ISEP. It's been a pleasure to be inside of a great university, always with the spirit to learn and teach the best practices to their students, making them betters professionals in their work environment.

Abstract

Nowadays we are witnessing a great growth in the generation, storage and treatment of large amounts of data. These data are generated by different sources, such as the Web, IoT devices, computers software, smartphone apps, sports and so on.

Big Data is a term used for large, varied and complex sets of data, with difficulties in storage, analysis and visualization for later processes or results.

The process of searching large amounts of data to reveal hidden patterns and correlations is called Big Data data analysis. This information is useful for companies or organizations and with the help of numerical and computational methods, results can be obtained in a short space of time.

For this reason, data implementations in Big Data need to be analysed and executed as accurately as possible. With the amount of data generated, data storage is being crucial. The huge increase in data does not stop and data analysis and visualization are adding to the Big Data era with the amount of data generated by computers, social networks, mobile devices, data collection in sports, etc. This research presents an overview of the content, scope, samples, methods, advantages, challenges and concerns of data analysis in Big Data.

Basketball is one of the examples, where we will work with the data, applying types of analysis and data visualization, to understand them and in the end, show their results.

Using Clustering algorithms and, with criteria defined in the being of the problems, we will have the same information spread over two or more clusters. Important steps, such as the analysis of each of the indicators and, the objective, we determine rule settings for the expected result.

In the results demonstration, we verified that the applied clustering algorithm, K-Means, obtained good results comparing with other data.

With the completion of this work, we can better understand the scope of Big Data and apply mathematical clustering methods to extract useful information from large amounts of data.

Keyword:

Big Data, Data analysis, K-Means, computational and mathematical methods, unstructured data.

Resumo

Atualmente assistimos a um grande crescimento na geração, armazenamento e tratamento de grandes quantidades de dados. Esses dados, gerados por diferentes fontes, como a Web, dispositivos IoT, aplicações computacionais, aplicativos de smartphones, desporto, entre outros.

Big Data é um termo usado para um grande conjuntos de dados, variados e complexos, com dificuldades de armazenamento, análise e visualização para processos ou resultados posteriores.

O processo de pesquisa em grandes quantidades de dados para revelar padrões ocultos e correlações é chamado de análise de dados em Big Data. Essas informações são úteis para empresas ou organizações e com a ajuda de métodos numéricos e computacionais, pode-se obter resultados num espaço curto de tempo.

Por esse motivo, as implementações de dados em Big Data precisam ser analisadas e executadas com a maior precisão possível. Com a quantidade de dados gerados, o armazenamento de dados está sendo crucial. O aumento enorme dos dados não para e a análise e visualização de dados estão agregando a era do Big Data com a quantidade de dados gerados por computadores, redes sociais, dispositivos móveis, coleta de dados em desporto, etc. Esta pesquisa apresenta uma visão geral do conteúdo, âmbito, amostras, métodos, vantagens, desafios e preocupações da analise de dados em Big Data.

O Basketball é um dos exemplos, onde trabalharemos os dados, aplicando tipos de análise e visualização de dados, para entendê-los e no final, mostrar os seus resultados.

Utilizando os algoritmos de Clustering e, com critérios definidos no ser dos problemas, teremos a mesma informações espalhadas por dois ou mais clusters. Etapas importantes, como a análise de cada um dos indicadores e, o objetivo, determinamos configurações de regras para o resultado esperado.

Na demonstração de resultados, verificamos que o algoritmo de clustering aplicado, K-Means, obteve bom resultados comparando com outros dados.

métodos matemáticos de clustering para extrair informações úteis de grandes quantidades de dados.

Palavras-chave:

Big Data, Análise de dados, K-Means, métodos computacionais e matemáticos, dados não estruturados

Contents

List of Figures

List of Tables

List of Acronyms

AHC	*Agglomerative Hierarchical Clustering*
AST	*Assist Made*
BLK	*Total Block Made*
CRM	*Customer Relationship Management*
DB	*DataBase*
DBMS	*Database Management System*
DREB	*Defensive Rebound*
ERP	*Enterprise Resource planning*
ETL	*Extract, Transform and Load*
FG3%	*Field Goal 3 Point Percentage*
FG3A	*Field Goal 3 Point Attempt*
FG3M	*Field Goal 3 Point Made*
FG%	*Field Goal Percentage*
FGA	*Field Goal Attempt*
FGM	*Field Goal Made*
FT%	*Free Throw Percentage*
FTA	*Free Throw Attempt*
FTM	*Free Throw Made*
IBM	*International Business Machines Corporation*
IoT	*Internet of Things*
IT	*Information Technology*
JSON	*JavaScript Object Notation*

KPI	*Key Performance Indicator*
MDS	*Multidimensional Scaling*
MP	*Minute Played*
NBA	*National Basketball Association*
NBL	*National Basketball League*
NCAA	*National Collegiate Athletic Association*
NoSQL	*Not only SQL*
NSA	*National Security Agency*
OREB	*Offensive Rebound*
PC	*Personal Computer*
PCA	*Principal Component Analysis*
PTS	*Point Made*
SQL	*Structured Query Language*
STL	*Total Steal*
SVD	*Singular Value Decomposition*
TO	*Turnover Made*
USA	*United States of America*
XML	*Extensible Markup Language*
YMCA	*Young Men's Christian Association*

Chapter 1

Introduction

Big Data analysis is becoming one of the most important methods trends that have the potential for dramatically changing the way organizations use the information to enhance the customer experience and transform their business models.

In this document, applied to basketball sport, we will provide an understanding of how data analyses are changing the way that unstructured data can be now studied and used in new ways in order to improve future companies actions.

Big Data is not a single market, rather, it is a combination of data management technologies and methodologies that have evolved over time. Big Data enables organizations to store, manage, and manipulate wide amounts of data at the highest speed and at the right time to gain the right insights.

Data has to be managed to meet the business requirements for designing a solution. Most organizations are at an early stage with their Big Data journey. Many organizations are experimenting with techniques that allow them to collect massive amounts of data to determine whether hidden patterns exist within that data that might be an early indication of an important change. Some data may indicate that customer buying patterns are changing or that new elements are in the business that needs to be addressed before it is too late.

As organizations begin to evaluate new types of Big Data solutions, many new opportunities will unfold. The new types of technologies such as Industry 4.0. The *Internet of Things* (IoT) has been connecting elements for many years but, the values extracted through a database has taken to new levels. IoT will be the next focus but, we already faced some change behaviours in sociality that made some

organizations have evolved in their data modelling. Facebook is an example. With the exponential growth, the organizations adapt their data model to be fast and capable to read a large volume of data.

Implementing a Big Data solution requires that the infrastructure is in place to support the scalability, distribution, and management of that data. Therefore, it is important to put both a business and technological strategy in place to make use of this important technology trend.

It is important to understand Big Data technologies and know the way organizations are using emerging technologies and new *DataBase* (DB) engines to transform the values of their data.

Storing amount of data, in Big Data is important but, its also relevant to mention the usage of methods to store and work the data, for visualization. Cluster algorithms have been used often. With clustering, and applying filters on raw data, we can obtain results to be studied in the future. Performance of player, team, and so on, can be used the cluster. With this approach, we can spread the raw data in small groups to be analysed further. The outcomes can be used in a different way and the analysis can be applied using the identification on each cluster point, to select and view other type of KPI.

There are multiple ways to work and show the data. Some of these procedures will be explain in this study.

1.1 Scope Definition

Today, Big Data is influencing the *Information Technology* (IT) industry like few technologies have done before.

The massive data generated from sensor-enabled machines, mobile devices, cloud computing, social media, and satellites helps different organizations improve their decision making and take their business to another level.

Also in sports, Big Data is being used to enrich organizations to grow. The data acquired is stored, in a large volume, to be analysed further. In basketball, performance analysis is very used. Big Data absolutely has the potential to change the way organizations, and academic institutions conduct their business and make discoveries, and its likely to change how everyone lives their day-to-day.

Data has been the most important information in organizations since *Personal Computer* (PC) was invented. Every day, data is generated in such a rapid manner that, traditional DB and other data storing system will gradually give up on gathering, retrieving, and finding relationships among data. As we know, in basketball, data is collected by organizations. Leagues, teams and players are being monitored. Gather information of point scored, fouls, 2-points made, 2-point attempts, rebound, steal and other attributes makes the enrichment of data. At the end of each season

or even during a season, the team can analyse its opponents in order to verify where the flaws are. With this amount of data, its also possible to verify the performance of a player, to be hired.

Now, the scope is to find a way to process and visualize the vast amount of data that a system has in place. In this process, numerical computer methods shall be used to work the data with outcomes views for better understanding.

1.2 Problem Setting

Each user or application, at the end of each year, consume and generate lots of data. Validate, aggregate and enrich the data with a high level of automation are the focus now. Data is recorded for business or individual matters and can also trigger action for real-time decision, such as controlling the performance of a player, gathering information by a team, collect data from other teams to make further analysis. In basketball, organizations are collecting the data to have a better view of league, players, team and so on.

Many researchers are being evolved in the Big Data area and we will explain, briefly, the concept of Big Data, the types, solution, scandal due to wrong data protection, data processing, etc.

NBA Basketball, was the sport chosen to make the study. Teams are using and recording data, to be used as aggregated values for forthcoming decisions. Studying offensive and defensive attributes, collecting the data and clustering it in different types of classification, will help to segregate de amount of data in a small population, to have better precision in the analysis.

1.3 Goals

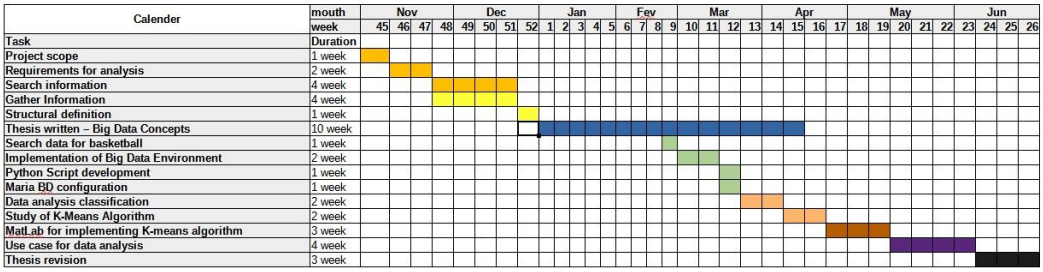
The aim of this study contains the learning of mathematics and computational algorithms to be applied in the Big Data environment, to extract and analyse crucial information in the Basketball scope. Big Data is often being used and in the new era, will be more used to collect, aggregate, store and display the amount of data that is being generated by many organizations.

Basketball is a sport well-known and, like in football, we will study the performance of the players, separating the data into clusters, to be more easy and versatile to be analysed. Clustering, has been a method used for a long time, and is often being used due to the amount of source data generated. On basketball, attributes are gathered and, with classification, we will generate clusters to spread the information.

Once information is spread and classified by clusters, we will use this data to enrich our study in different aspects such as: offensive, defensive, and percentage of successful field goals.

1.4 Calendar

The project plan, can be observed in figure 1.1, where is visible all the tasks, planned for the study - Data Analysis in the BIG Data Scope in Basketball, with its duration. In total, 41 weeks were spend to present all the content available in this study.



Calender	mouth	Nov				Dec				Jan					Fev				Mar					Apr					May					Jun			
	week	45	46	47	48	49	50	51	52	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Task	Duration																																		
Project scope	1 week																																		
Requirements for analysis	2 week																																		
Search information	4 week																																		
Gather Information	4 week																																		
Structural definition	1 week																																		
Thesis written – Big Data Concepts	10 week																																		
Search data for basketball	1 week																																		
Implementation of Big Data Environment	2 week																																		
Python Script development	1 week																																		
Maria BD configuration	1 week																																		
Data analysis classification	2 week																																		
Study of K-Means Algorithm	2 week																																		
MatLab for implementing K-means algorithm	3 week																																		
Use case for data analysis	4 week																																		
Thesis revision	3 week																																		

Figure 1.1: Project Calendar

1.5 Report Organization

In chapter 1, is presented the aim of this study - Data Analysis in the BIG DATA scope in Basketball. The definition scope, problem setting, work aim defines problems that we are facing nowadays.

In chapter 2, we will study the behaviour of BIG DATA, the types of data used, collection data and the DataBase. Data can be sent from various formats and its important to understand, learn and use methodologies to store the amount of data generated to be processed, analysed and visualized, for a better understanding.

In chapter 3, we present an overview of Basketball. The rules, players, season and impact that are being used in the game that generate millions of millions of dollars, in the USA.

In chapter 4 we will use mathematical and computer algorithms, in a Big Data simulated environment, using Maria BD DataBase to store the information, *Extract, Transform and Load* (ETL) mechanism to pre-process the data before it is stored and for last, the K-Means implementation (Clustering algorithm), using MatLab, to demonstrate the results of this study.

And lastly, in chapter 5, the conclusion of this study and the further work to enrich it, even more, are presented.

Chapter 2

Big Data

Big Data in a way just means "all data" and there is quite some data nowadays. The volume of data is impressive and, looking at the growth rates of the digital data universe, data volume will not stop increasing.

Originally, Big Data mainly was used as a term to refer the size and complexity of datasets, as well as to the different forms of processing, analysing and so forth that were needed to deal with those larger and more complex datasets and unlock them in value [1].

2.1 History

The term Big Data was introduced on Digital World in 2015 by Roger Mougalas. However, the application of Big Data and the search to understand the available data is something that exist for a long time ago.

In 1663, John Graunt recorded and analysed information on the rate of mortality in London. John Graunt provided the world with the first statistical analysis of data ever recorded. John Graunt studied the causes of death in seventeenth-century England. Due to his work, John Graunt is widely regarded as the pioneer in the field of statistics.

After the work of John Graunt, accounting principles continue to improve and develop but nothing extraordinary was shown until the 20th Century when the information era began. The starting point of modern data begins in 1889 when

a computing system was invented by Herman Hollerith in an attempt to organize census data.

Herman Hollerith did a great achievement in introducing a computer system and at the time being, in the year 1937, Franklin D. Roosevelt President of the *United States of America* (USA) requested the tracking of USA population.

International Business Machines Corporation (IBM), contracted by the government of the USA and the first company to introduce the DB, was ordered to develop a punch card-reading system, a data processing machine that would be applied in this extensive data project to record data.

The first data-processing machine was called "Colossus" and was developed by the British in order to decipher Nazi codes in World War II, 1943. This machine worked by searching for any patterns that would appear regularly in the intercepted messages. This machine worked at a record rate of five thousand characters per second, reducing the work that would take weeks to a few hours.

In 1952, *National Security Agency* (NSA) was created in USA and employees started to develop advanced machines that could independently and automatically collect and process information to decrypted messages during Cold War.

The first Data Center was built by USA government in 1965 with the purpose of storing million of tax returns and fingerprints set. This was achieved by transferring every record onto magnetic tapes that were to be stored systematically in a central location. This project, however, did not persist due to fear of sabotage or acquisition. However, its widely accepted that this initiative was the starting point of electronic big storage.

Tim Berners-Lee, a British computer scientist invented the World Wide Web in 1989 with the intention of sharing information. This invention had a huge impact on the world and in 1990, the creation of data grew at an extremely high rate as more devices gained capacity to access the Internet.

The first super computer was built in 1995, with a capacity to handle work that hat would take a single person thousands of years in a matter of seconds [2].

In 2005, the term Big Data was introduced by Roger Mougalas and at the same time, Yahoo created a mechanism with intention of indexing the entire World Wide Web with open source code called "Hadoop". Hadoop is a software platform in JAVA, with the ability to process a large volume of data. During this period, social networks were rapidly increasing and a large amount of data was being created daily. In figure 2.1 we can see the evolution of each type of data.

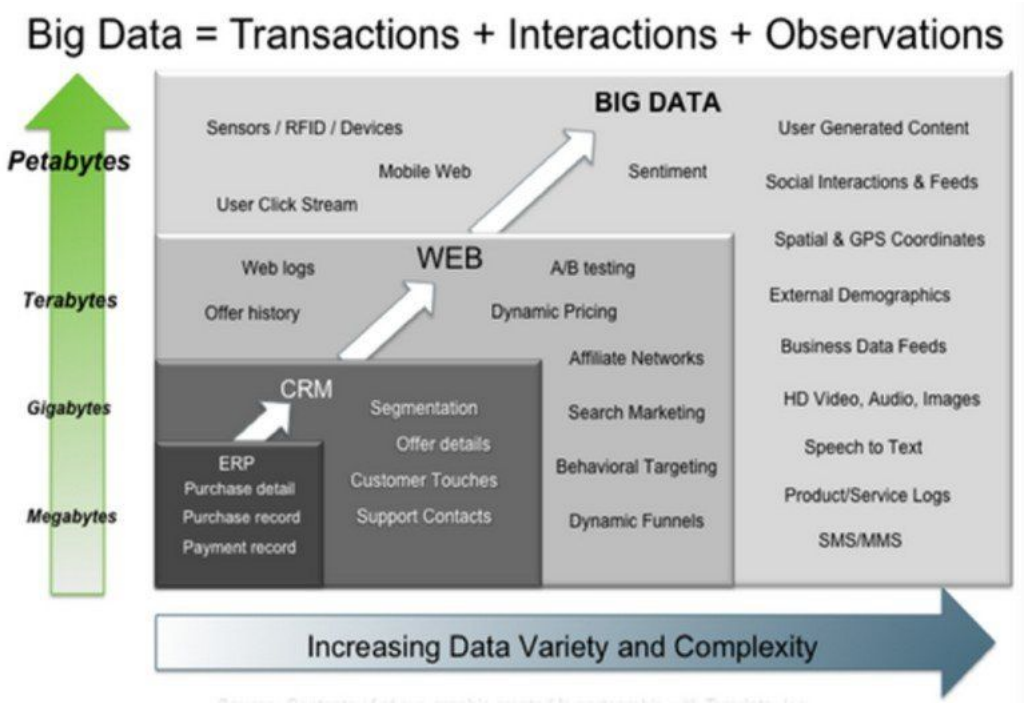


Figure 2.1: Big Data market size revenue forecast

Businesses and Governments began to establish Big Data projects and since then, Big Data is growing very fast [1], as shown in figure 2.2.

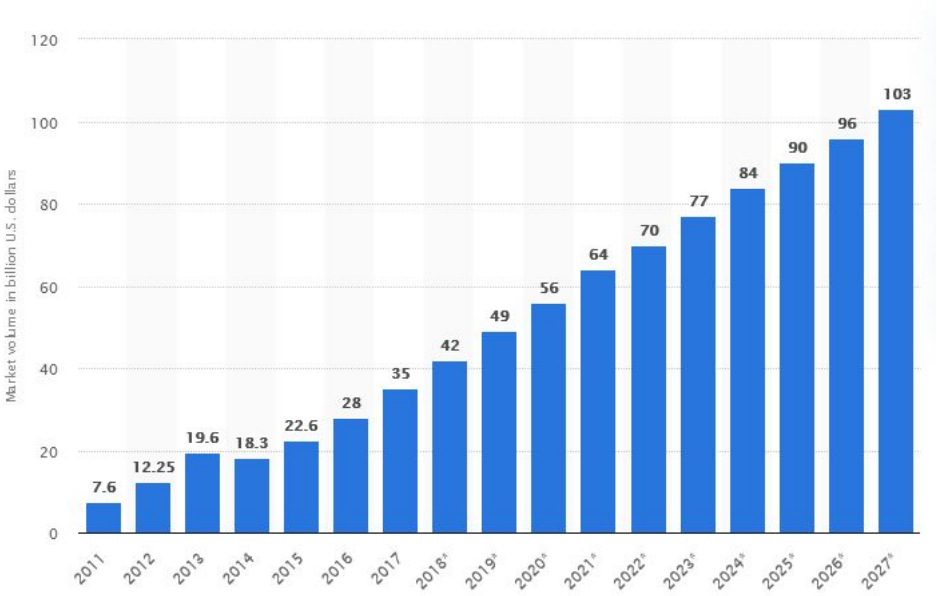


Figure 2.2: Big Data market revenue forecast

2.2 What is

Big Data is an ambiguous and relative term. Big Data it's data that doesn't fit well into a familiar analytic paradigm and it's been developing for many years. It has the ability to change the nature of a business. In fact, there are many firms whose sole existence is based upon their capability to generate knowledge that only Big Data can deliver.

In simple words, Big Data won't fit into rows and columns of an Excel spreadsheet, even compared to the traditional DB. It can't be analysed with conventional multiple regression and, it probably won't fit on a normal computer's drive.

Big Data is a dedicated field for analysis, processing and storing large collections of data that frequently originate from scattered sources. News solutions are required when comparing with traditional data analysis, processing, storage technologies and techniques. Big Data address distinct requirements, such as the combination of multiples unrelated datasets (collection of data), processing large amount of data of unstructured data and harvesting hidden information in a time-sensitive manner.

In addition to the traditional analytic approach based on statistics, Big Data became with new techniques that leverage computational resources and approaches to execute analytics algorithms. This step is important as datasets continue to become larger, more diverse, more complex and streaming-centric. While the statistics approach has been used to approximate measurements via sampling, on Big Data, the processing of an entire dataset is allowed.

The analysis of Big Data datasets is an interdisciplinary endeavour that cover mathematics, statistics, computer science and subject matter expertise. With this new approach, Big Data takes into account the impact of data characteristics on the design of a solution.

Data within Big Data is being amassed and collected via applications, sensors and external sources. Data processed by Big Data solution can be used by enterprise applications or can be stored to enrich existing data. As a result, through the processing, Big Data can lead to a wide range of acknowledgement and benefits such as:

- Operational optimization
- Actionable intelligence
- Identification of new markets
- Accurate predictions
- Fault and fraud detection
- More detailed records

- Improved decision-making
- Scientific discoveries

Evidently, the application and potential benefits of Big Data are deep but, as with any kind of technology, many issues need to be considered when adopting Big Data approaches. These issues need to be understood and weighed against anticipated benefits.

The most important characteristic of a Big Data, is the dataset to be considered. It must have one or more characteristics that require accommodation in the solution design and architecture of the analytic environment. The final goal is to conduct the analysis of the data in the best way to achieve high-quality results to provide enterprise values.

To fulfil the requirements, there are five Big Data characteristics that can be used to help differentiate data categorized as "Big" from the other forms of data. The five Big Data attributes shown in Figure 2.3 are commonly referred to as the Five Vs: Volume, Velocity, Variety, Veracity and Value.

Figure 2.3: The Five Vs of Big Data

Volume is the base of Big Data that is processed. High data volumes impose distinct data storage and processing demands, as well as additional data preparation, curation and management process.

One of the most important attributes of the 5Vs is velocity. Data can arrive at fast speeds and an enormous dataset can accumulate within a very short period of time. The velocity of data translates into the amount of time that it takes for the data to be processed and requires a high elastic design available data processing

solution corresponding with the data storage capabilities. Depending on the source, velocity may not be high. Data velocity is put in perspective when considering that the flowing data can be easily generated in a given minute.

Variety refers to the multiple formats and types of data that need to be supported by Big Data Solutions. The amount of data type, brings challenges in terms of data integration, transformation, processing and storage.

Veracity is the data that enters the Big Data environments and needs to be assessed for quality, which can lead to data processing activities to resolve invalid data and remove noise. Noise, is data that cannot be converted into information and thus has no values, whereas signals have value and lead to meaningful information. Data with a high signal-to-noise ratio has more veracity than data with a lower ratio. Uncontrolled sources of data usually contain more noise instead data from controlled sources. Thus signal-to-noise ratio is dependent upon the source of data and its type.

Values are defined as the usefulness of data. The value characteristics are intuitively related to veracity. With higher data fidelity, more value it holds on a system. Values are also dependent on how long data processing takes because analytics results have a shelf-life. i.e, a 20 minute of delay has no values instead a 20 milliseconds of delay. Time processing on Big Data is a strong key.

Nowadays, from small to large organizations, data is with a huge increment. It is important to analyse and predict all the Five V's available [3].

2.3 Types of data

The data processed by Big Data can be generated by humans or machines. Humans data generated results with an iteration such as online and digital service. Machine-generated data is provided by software or hardware devices in response to real-world events. Table 2.1 presents some examples for each data type.

Table 2.1: Examples of Human and Machine-generated data on Big Data

Human data	**Machine data**
Textual data	Web logs
Videos	Sensor data
Musics	Telemetry data
Pictures	Smart meters
Structured Data	GPS

Human and Machine-generated data can be provided from a variety of sources and be represented in various formats and types. The processed data by a Big Data solution can be categorized by types:

- structured data
- unstructured data
- semi-structured data

This data types categorization is referring to the internal organization of the data that can be called data formats.

Structured data is data that follows a model or schema and is often stored in a tabular form. As an example, we can see a traditional DB that is structured to capture relationships between entities to store data. This type of data is frequently generated by enterprise applications and information systems like *Enterprise Resource planning* (ERP) and *Customer Relationship Management* (CRM) systems. Examples of this type of data includes banking transactions, invoices and customer records.

On the other hand, we have unstructured data. Data that does not fit in a data model or schema is known as unstructured data. Unstructured data has a faster growth rate than structured data nowadays due to the numerous applications that has been inserted into society. A Smartphone is an example. We can make videos and record audio via an application and store it in a cloud. It is estimated that unstructured data makes up 80 percent of the data within any given enterprise. This form of data is either textual or binary and is often conveyed via files that are self-contained and non-relational. Technically, both text and binary files have a structure defined by the format itself but, this aspect is disregarded and the notion of being unstructured is in relation to the format of data contained in the file. As an example, we have video, audio records and pictures that are types of unstructured data.

To find a balance between structured and unstructured data, in Big Data, we can have data semi-structured. This method has a defined level of structure and consistency that is not relative in nature. This data is hierarchical or graph base stored in files that contain text. *Extensible Markup Language* (XML) and *JavaScript Object Notation* (JSON) are common forms of semi-structured data used in Software languages. In that kind of files, the data is structured and can be easily processed by unstructured data. Semi-structured data has special pre-processing and storage requirements that can be validated to ensure that it conforms to its schema definition [3].

Figure 2.4 presents all types of data on Big Data with their format types.

Figure 2.4: Type of data in Big Data

2.4 Data Collection

Data Collection is the first step for any analytics application. Before data can be analysed, the data must be collected and ingested into a Big Data stack. The tools and frameworks for data collection depends on the source and type of data being ingested. For that, various types of connectors can be used as DB connectors, custom connectors, publish-subscribe messaging frameworks, messaging queues, etc. Data collection systems allow collecting, aggregating and moving data from various sources into a centralized data store.

Data preparation steps involves various tasks such as data cleaning, data wrangling or munging, de-duplication, normalization, sampling and filtering [4].

Table 2.2 enumerates the aspects to be considered.

Table 2.2: Data Preparation in Big Data

Steps	**Description**
Data Cleaning	Detect and fix issues for corrupt records, line with missing values and bad formatting.
Wrangling/Munging	Process to transforming and mapping data from raw data into another format to meet certain rules.
De-duplication	Detect duplicated information.
Normalization	Required when data is from different sources using various units/scales or have a diverse abbreviation for the same thing.
Sampling	To process data that meet certain rules.
Filtering	Can be useful to reject bad records with incorrect or out-of-range values.

2.4.1 Data and Database

Data can be stored in multiple formats and sizes. In that way, the system has to have the ability and capacity to learn and interact. System can only stored the data with an unrecognised format however, data will be stored. Phases such as automation

and consolidation have to be in place to build a methodology and process model of data that shall be in the business model.

Analysis needs to be requested or specify the various operations on the data. This transaction involve many routine activities explained in the subsection.

Figure 2.5 illustrates the data process to be considered in a DB.

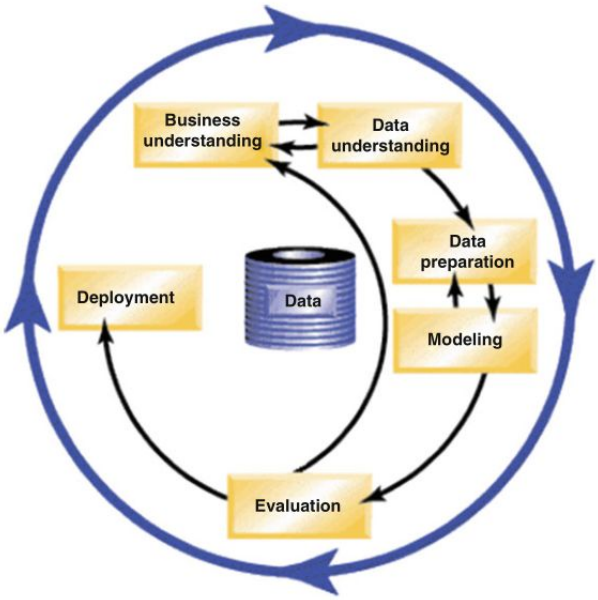


Figure 2.5: Data process on Big Data

During all these stages, and if data is according to the business rules, the data will be stored in the DB according to the system recommendation. That will take into account the location and types of categorization defined by the business [5].

2.4.2 Aggregation

Data aggregation is rapidly emerging. An efficient solution to deal with the voluminous data by combining similar data, eliminating the redundant data problem and reducing resource consumption. Data is aggregated at multiples resolutions and provide a trade-off between efficiency and accuracy. It represents one of the processing challenges of Big Data that is being at the centre of big companies [6].

The structure is built once, updated incrementally, and serves as a common data input for multiple mining and learning algorithms. Data mining algorithms are modified to accept the aggregated data as input. Hierarchical data aggregation serves as a paradigm under which novel data representations and algorithms work together for the analysis and mining of Big Data. Data aggregation is a technique that is used to reduce the number of instances in which similar instances are combined into one stance to eliminate redundancy.

More advanced methods for aggregation include a multidimensional data cube, which holds aggregated data in subspaces to support advanced analysis and decision. The multilevel structure forms a different granularity of data aggregation, where multiple regions at a lower level are grouped to form one region in the next higher level. We assume aggregation of each level to half of the size of its immediate lower level. This method provides a hierarchy concept for continuous data.

Figure 2.6 shows an example of a multidimensional data cube generated from three dimensions, each data has four levels of aggregation.

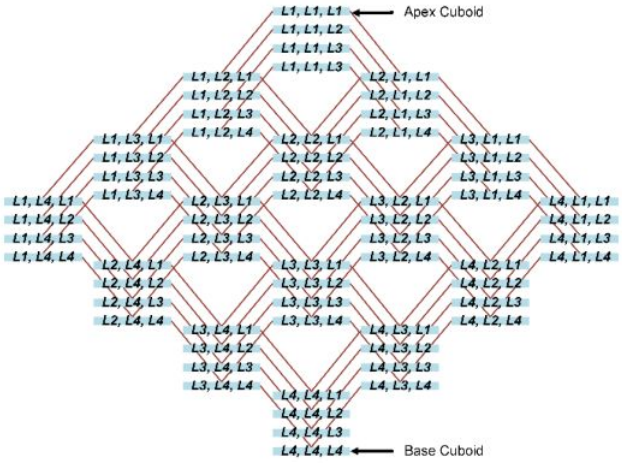


Figure 2.6: Data Aggregation type on Big Data

Each node in the lattice is a cuboid that represents a single combination of aggregation levels (one level from each dimension). Storing aggregated data in a tree structure makes valuable tree operations such as search operations available. Searching the tree structure is faster and more efficient than searching the massive raw dataset.

The lowest layer of data cube aggregation (L4, L4, L4) is stored in the tree leaves, and as we move toward the tree root the level of aggregation increases. The number of tree levels (denoted by h) is determined by the domain expert.

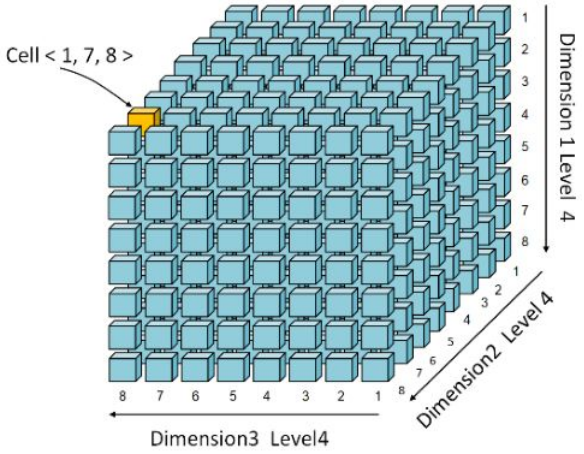


Figure 2.7: Data Aggregation on Big Data, the base cuboid

Each node except the root in the tree represents a non-empty cell from the corresponding cuboid. The typical contents of an intermediate node include statistical measures of aggregation data instances in the node. These measures should be updated incrementally when new data instances join the dataset map to its parent and children nodes [7].

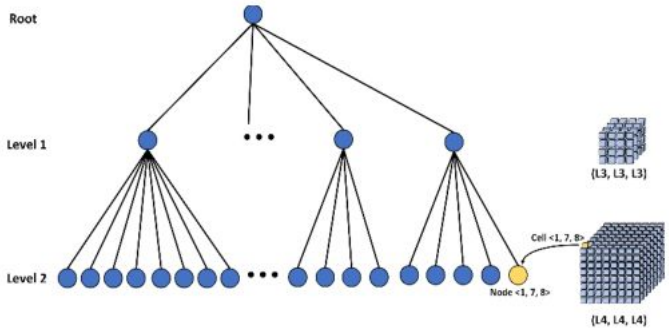


Figure 2.8: Data Aggregation on Big Data, the tree node

2.4.3 Challenges

Since we are in a new area, a wide range will be needed for data processing due to the enormous quantity of devices to cover. We see four issues requiring high-performance processing of large data volumes and appropriated Big Data approaches.

1. A large and continuously growing amount of data: The Big Data integration generated to access and process information as well as the life cycle of entities.

2. Knowledge processing: Processing large and complex amount of ancient data for analytics, mining and prognosis to enhance decision making and for results optimization.

3. Random access to data: individual situations and events in production that are triggered to build and ad-hoc networks for decision making.

4. Real-time access and processing: real-time control and decision making. The decision process has to be consider in the overall system.

The issues presented before form two general Big Data use cases. The reading access and processing patterns, underlying data. Processing data and underlying data is a complex work.

- Data Mining: Big Data solution that supports batch processing and distributed computing.

- Entity Access: requires Big Data solution support real-time queries and random access.

To make up these gaps, infrastructure and processing methodologies shall be supported [8].

2.5 Infrastructure, Platforms and Software

In Big Data, Infrastructure and Platforms aren't mentioned a lot but, are the fundamental structure of it. The size and process model, in a DB, has to be dimensioned. A DB is a collection of related data that is organized to ease and speed up access to data in a standardized format. The interface between the DB, users, and other applications is integrated into a software system known as *Database Management System* (DBMS).

Most of the data management functionalities are used to provide only relational DBMS (*Structured Query Language* (SQL)). SQL is used to communicate with a DB as a standard language for relational DBMS. SQL statements are used to perform tasks such as updating data on a DB, or retrieving data from a DB.

In the last decades, new applications emerged and new requirement were raised. With the amount of data and due to the storage data being relational, the designer of DB started to question the fundamentals and properties of relational DB. Big Data bring challenges and solutions to mitigate the principal requirements - Volume, Velocity, Variety, Veracity and Value.

This new approach resulted in the appearance of *Not only SQL* (NoSQL) DB. NoSQL DB provides a mechanism for storage and retrieval of data that is modelled in means other than the tabular relations used in relational DB. The motivations for this approach include simplicity of design, simpler scaling to clusters of machines (which is a problem for relational DB), finer control over availability and limiting the object-relational impedance mismatch.

The data structures used by NoSQL DB (e.g. key–value pair, wide column, graph, or document) are different from those used by default in relational DB, making some operations faster in NoSQL. The particular suitability of a given NoSQL DB depends on the problem it must solve. Sometimes the data structures used by NoSQL DB are also viewed as more flexible than relational DB tables.

In the past, information was stored on DB on a single server. This is no longer possible today, as a single server is no longer sufficient to cope with the colossal volume of data. In that way, new infrastructure were also developed to ensure all the requirements expected of Big Data. Because of the enormous quantities of data, it must incorporate a robust infrastructure for storage, processing and networking, in addition to an analytics software (Platforms).

Infrastructure is based on hardware and must be chosen according to scalability, parallel processing, low-latency resources, and data optimization. Many products from various sources appear and one of the most rated is the IBM products. IBM infrastructure meets six critical business needs. Examples to acquire, grow, and retain customers, to transform financial management processes, optimize operations, reduce fraud, manage risk, improve IT economics, to create new business models. For each entry point, certain infrastructure design points (speed, access, availability) are more important than others. Certain design points are better enabled by specific hardware and software(platforms) infrastructure capabilities and architectures [9].

In a Big Data infrastructure can be implemented one or more platforms (based on cloud servers). Big Data platforms are a type of IT solution that combines the features and capabilities of several Big Data applications and utilities within a single solution. Generally consists of storage, servers, DB, management, business intelligence and other Big Data management utilities. It also supports custom development, querying and integration with other systems. The primary benefit behind the Big Data platform is to reduce the complexity of multiple vendors/solutions into one cohesive solution. Big Data platform is also delivered through the cloud where the provider provides an all-inclusive Big Data solutions and services [10].

Figure 2.9: Infrastructure and Platforms

In figure 2.9 is visible the cloud infrastructure that contains the functional parts of the system. On Big Data platform, where is set the specification of Big Data and on top the of it, the business data. Business data cover the compliance of the data spread in four types: Governance, Archive, Privacy and Protection.

2.5.1 Infrastructure

Infrastructure is the cornerstone of Big Data architecture. Processing the rights tools for storing, processing and analysing the data is crucial in any Big Data system. In the last years, Big Data infrastructure is being changed [11].

Any organization can benefit from the insights offered by Big Data however, the ability to assimilate massive quantities of data from any resource and turn it into intelligible, actionable insight can be hard to implemented. Across the projection of a Big Data infrastructure, the 5Vs need to be taken into account. With the scope of it, the chosen type of processing and storage data will have an impact on that architecture. Where Big Data infrastructure support comes in is in supporting storage for data volume, maintaining throughput for data velocity, and providing speciality programming to accommodate data variety.

In the beginning, the bare metal, was a sustainable solution, with physical hardware where the owner bought all the required equipment and built the solution with the limited capacity of processing and storage however, with the time being, the native could infrastructure is being adopted (the virtual computer machines).

Cloud computing provides the elastic data storage and massively parallel process demanded by Big Data. The data can be hosted by third parties. Clients only pay a rent and with a short time, can increase the power processing, and storage capacity in a short period.

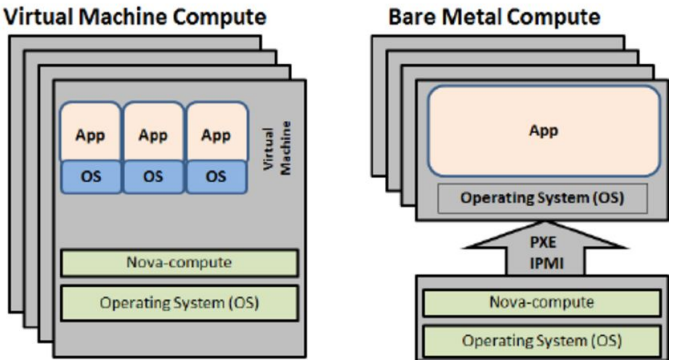


Figure 2.10: Infrastructure types: Bare Metal vs Cloud Computing

The demand for more Big Data infrastructure support is growing, making the Hardware, Software and Services demands increase too [12].

2.5.2 Platforms

A Big Data platform consists of many components where we have many alternatives for the same job. A chosen platform in an organization shall fulfil all the requirements and a platform should adapt, accept, and evolve due to new expectations. The challenge is to design a simple platform with an efficient cost in terms of development, maintenance, deployment, and the actual running expense.

Big Data platform is an integrated computing solution that collect, store, process and discover. The value of storing and processing data to get actionable insights have been a competitive advantage across the world.

Organizations use Big Data platforms for business intelligence, data analytics and data science, among others, because they identify, extract, predict and forecast information based on the collected data, thus aiding companies to make informed decisions, improve their strategies, and evaluate parts of their business. The more the data recorded in different aspects of business, the better understanding of the data can exist. The solutions for Big Data processing vary based on the company strategy however, all solutions shall follow the main three aims which are: Scalability, Availability, Performance and Security [13].

There are a lot of available Big Data platforms in the market, freeware and enterprise. Some of the most rated platforms are:

- Windows Server Data Center
- Red Hat Enterprise Linux
- Ubuntu Server
- SQL Server Enterprise
- Reserved VM Instances
- Web App For containers

Both Microsoft Windows and Linux operating system platforms are providing flexibility for developers and organizations. Evolution has been made year by year to provide the market with new solutions.

2.5.3 Software

The increase in data generation and data consumption has led to challenges and opportunities while massive data increase. The ingestion of the data, the analysis and the storage of data in a fast/slow moving had some challenges and Google was the pioneer in trying to solve them.

The computing and storage require large server machines (in a cloud base or traditional way with a vertically scaled of high number of CPU and RAM) as well as network storage with a very costly and very slow process of data growth.

Google introduced in the Big Data market the "Hadoop Framework" to address these problems, utilizing cheaper commodity server machines (which provides easier and more cost-effective).

Organizations could deploy and manage such environments, or they could leverage a cloud service that provides easier and more cost-effective options for organizations with the possibilities of cloud computing to capture, store, and analyse large volumes of data without having to worry about building such environments or managing the complexity of such solutions.

Organizations leveraging such technologies can rest assured that open-source or compatible software is utilized, which provides choice and portability. Figure 2.11 illustrates some of the vast open source ecosystems of Big Data [14].

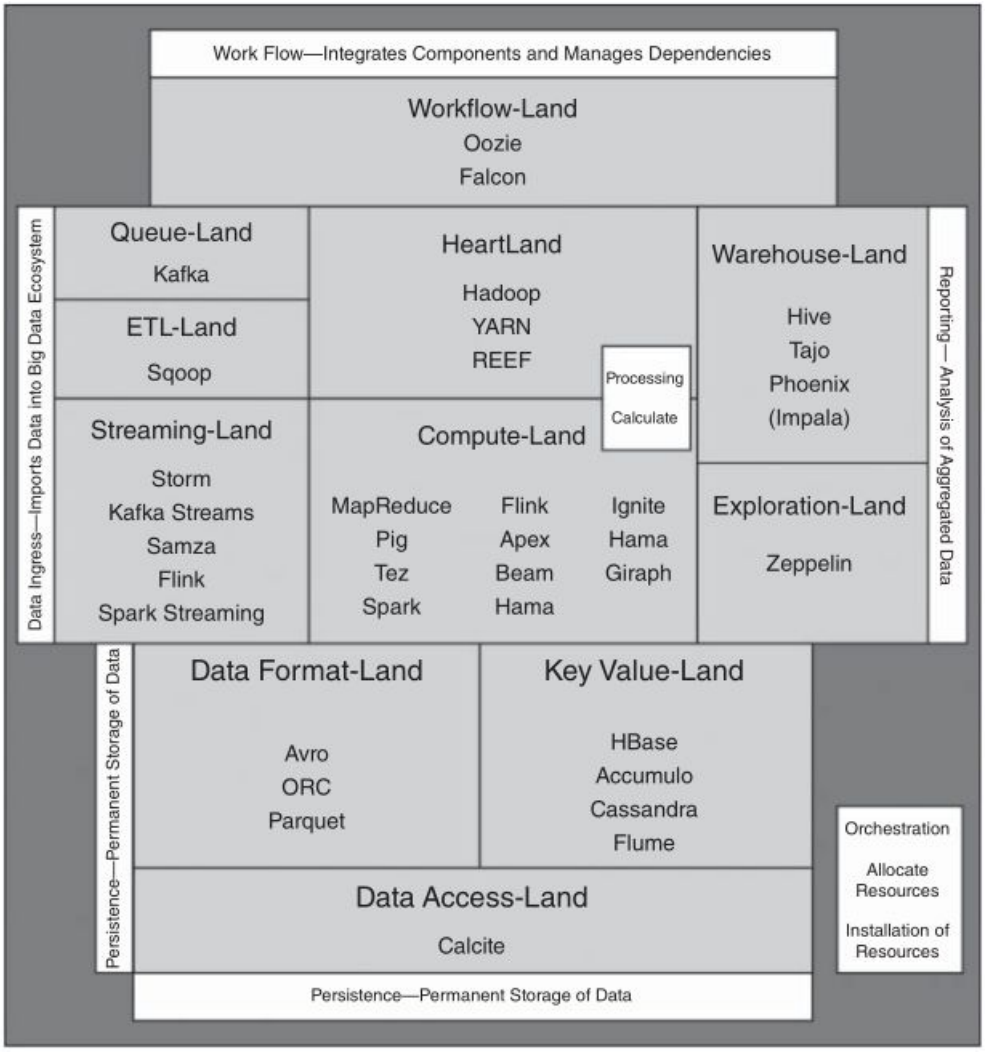


Figure 2.11: Software Open Source Ecosystem

In figure 2.11 we can visualize the type of open source software that we can use to normalize the data, storage, work flows for reporting. The reporting is very important where, the user can visualize the data but, the other function behind of the report are even more important. Having good normalization and storage, is crucial for the performance of a Big Data environment.

2.5.4 Advantages and Disadvantages

The adoption of such a platform may be difficult in an organization's current environment. This is because data aggregation technology is complex and requires the correct technical knowledge for implementation. Another area of concern is the quality of the imported data: if it is of low quality, accurate, etc.

There are six major advantages to using a data management platform: gathering data in one place, using third-party data to discover new markets, gaining audience insights, creating a full view of customers, targeting your audience, and effectively budgeting your expenditures on marketing.

As a disadvantage, we have the data management platforms rely heavily on "Cookie Technology" to identify behaviours. Recent moves from Apple and now Google are moving towards blocking third-party advertising "Cookies" which places the DBMS platform value proposition at risk [13].

Big Data is growing exponentially. With Big Data, a better understanding of Science, Politics, Healthcare, Education, Social Media, trend Markets, etc. making flows the worlds economics, having a tremendous impact. Big Data is recognized well by the market companies and is bringing high values [15].

Organization are:

- Adopting the methods required to generate, collect and store these new forms of data
- Using advanced data processing techniques
- Applying this data knowledge to business decisions and activities

2.6 Opportunities

The Big Data marketplace is growing on a scale of 26 percent per year, faster than any other industry, impacting business and academic sectors, and delivering significant value in job creation, creativity, innovation and productivity.

Organizations already have a great trade of Data and already struggle to realise the full value of this data. Data is the heart of prediction and accurate decisions in

the market, based on good quality data. Automation and sourcing data from what is already out is becoming more important in an organization, helping the enrichment of it.

Some organizations like Facebook, Twitter, Google, etc. already understand that data helps to find out what costumer and markets are saying about the organization and its reputation.

Technically, there are significant opportunities for advances with platforms, data mining, analytics and forecasting application, data management, business intelligent and other aspects. There are a lot of available Big Data technologies where organization can use to understand and improve, for example their revenues, with Big Data (with open source or enterprise systems).

The success of using Big Data is that each organization shall make their homework and have the right advisers to have a good system to enrich their knowledge about the marker that they are inserted.

The challenge with Big Data is not only about technology. The goal of each system is to have the right people with the right skills to understand the data.

The skills that have been most used to understand the Big Data analytics techniques are:

- **Data mining models:** which help to answer questions such as "Which products are costumer likely to buy?" or "Which workers are likely to leave the company?"
- **Text models:** which address questions such as "What are people saying about my products and services?" or "Can I detect emerging issues from customer feedback or service claims?"
- **Forecasting models:** which tackle questions such as "How many products will be sold this year or next year?" or "How does this break down by each product over the next three or six months?"
- **Operation research:** which looks at questions such as "What is the optimal inventory and stock to be held of each of the products to minimise overall holding costs?" or "What is the least cost route for transporting goods from warehouses to final destinations?"

Despite all these challenges, high-performance analytics will drive high-impact results in many businesses.

Big Data is part of a wider innovation revolution in our domestic, social and business worlds, enabling a better understanding of our history and place in space. We have a big opportunity to improve our use of Big Data to interpret better our future environment and possibilities on earth [16].

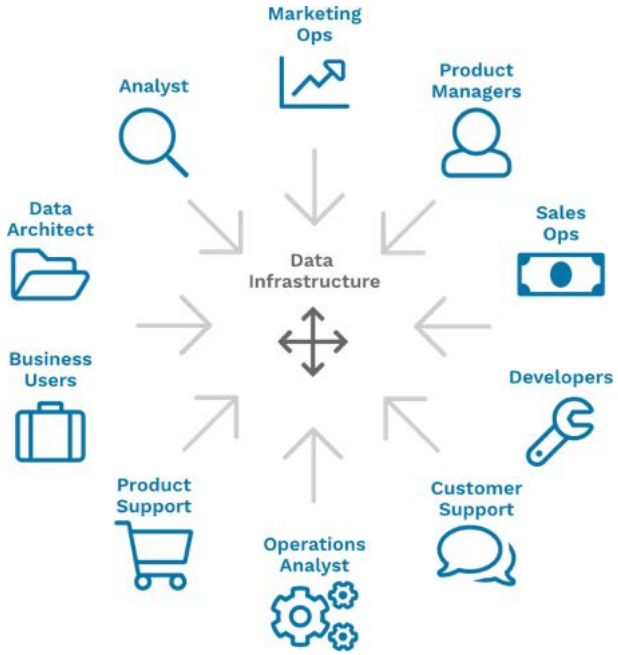


Figure 2.12: Big Data - Opportunities in the marketplace

Sales, marketing, product support, analysis and other work types is the most used areas in the Big Data, as figure 2.12 is shown. With the time being, other areas will have the possibility to be integrated as an opportunity in Big Data.

2.7 Data Analysis

The analysis of data, in Big Data, can be made in various options. It can be implemented types of analysis like basic statistics, regression, clustering, time series, etc. It will depend on what is the purpose of the analysis and its aim. These skills include analysis methods called "Big Analytics", used to benefit from the data.

Big Analytics helps organizations harness their data and use it to identify new opportunities, changing the way the world does business. This means that organizations can improve their customer retention, develop better products, and gain a competitive advantage by taking rapid action to respond to market changes, indications of critical customer shifts, and other metrics that impact business.

- Descriptive Analytics tells what happened in the past and helps a business understand how it is performing by providing context to help stakeholders interpret information.
- Diagnostic Analytics takes descriptive data a step further and helps you understand why something happened in the past.
- Predictive Analytics predicts what is most likely to happen in the future and provides organizations with actionable insights based on the information.

- Prescriptive Analytics provides recommendations regarding actions that will take advantage of the predictions and guide the possible actions toward a solution.

Within the analysis data, we can determine the analysis mode, which can be either batch, real-time or interactive.

The choice of the mode depends on the requirements. If demand results to be updated after short intervals of time (every few seconds), then real-time analysis mode is chosen. The results are to be generated and updated on larger timescales (daily or monthly), then batch mode can be used or if your demands flexibility to query data on demand, then the interactive mode is useful.

Basic statistical data are data collected on a regular basis, using the most simple mathematics operation. Simple operations are often used in Big Data to get information like "Counts", "Max/Min/Mean", "Top-N", "Distinct" data and "Correlation" data.

Classification is achieved by algorithms that belong to supervised machine learning. Supervised machine learning involves inferring a model from a set of input data and known responses to the data and then uses this model to predict responses to new data.

Some models like Decision tree, Random Forest, and Naive Bayes algorithms are used to classify the data.

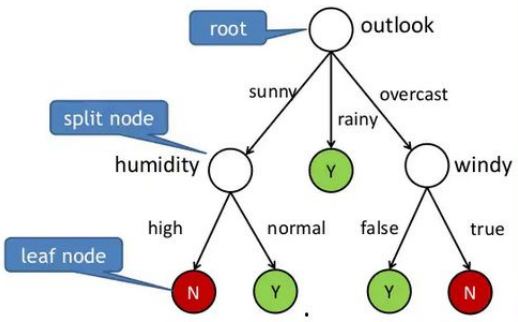


Figure 2.13: Data Analysis - Classification, Decision Tree

While in classification, figure 2.13, the response variable is categorical and not ordered in a Regression, the response variable takes continuous values, regression involves modelling the relationship between a dependent variable and one or more independent variables.

Figure 2.14 shows some Regression models like Linear Least Squares, Generalized models and Isotonic Regression are used to predict "events".

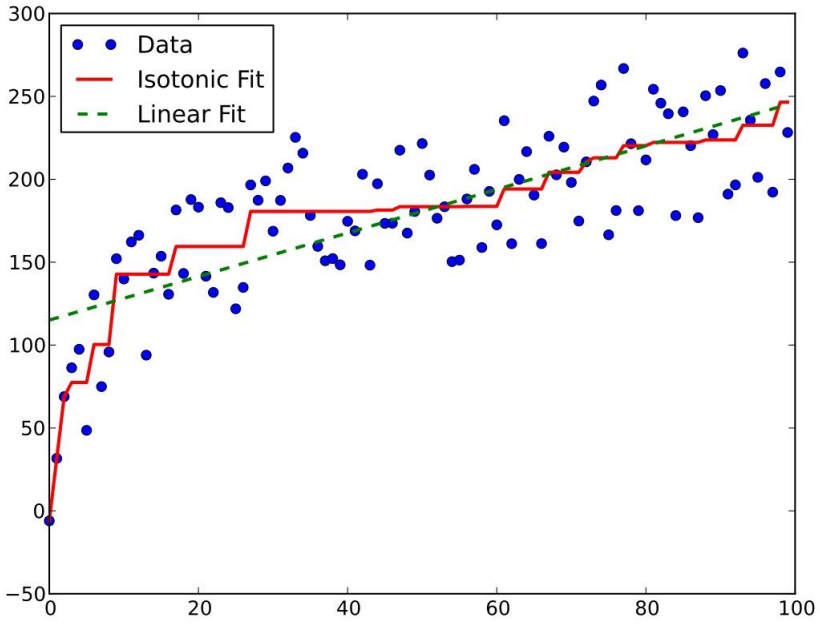


Figure 2.14: Data Analysis - Linear Regression, Isotonic Regression

The process of grouping data items is a method of identify similar data items that are closer to each other. Clustering Big Data happens in applications such as:

- Clustering Social Network data to find a group of similar users
- Clustering Electronic health record (EHR) data to find similar patients
- Clustering sensor data to group similar or related faults in a machine
- Clustering market research data to group similar customers
- Clustering clickstream data to group similar users

Clustering is achieved by cluster algorithms that belong to a broad category of algorithms called unsupervised machine learning. Unsupervised machine learning algorithms find the patterns and hidden structures in data for which no training data is available. Some algorithms used are K-means, Gaussian Mixture and Power Interaction Clustering.

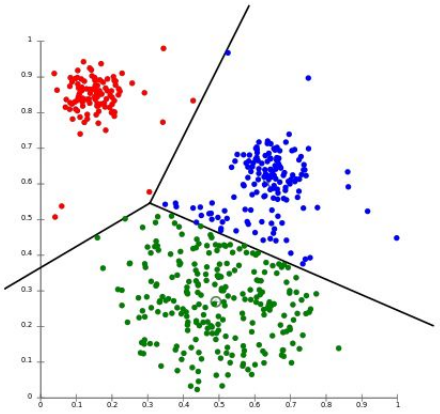


Figure 2.15: Data Analysis - Clustering, K-Means

Time series analysis is a specific way of analysing a sequence of data point collected over an interval of time. In the time series analysis, analytics data points at consistent intervals over a set period than just recording the data points intermittently or randomly.

However, this type of analysis is not merely the act of collecting data over time. What sets time series data apart from other data is that the analysis can show how variables change over time. In other words, time is a crucial variable because it shows how the data adjusts over the course of the data points as well as the final results.

There are a lot of algorithms to be used with Time Series analysis and, for example, we have the "Hidden Markov Model", Kalman Filters, Time-Frequency Models, Outliers, etc.

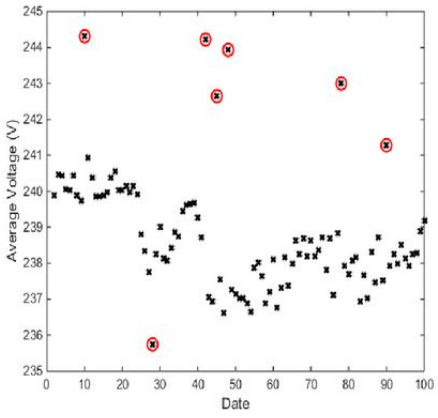


Figure 2.16: Data Analysis - Time Series Analysis, Outlier Detection

Dimension Reduction refers to the process of reducing the number or dimensions (input variable) in a dataset. It is commonly used during the analysis of high-dimensional data.

High-dimensionality statistics and dimensionality reduction techniques are often used for data visualization. Nevertheless, these techniques can be used to simplify a classification or regression dataset in order to better fit a predictive model.

- Dimensionality reduction is a general field of study concerned with reducing the number of input features.
- Dimensionality reduction methods include feature selection, linear algebra methods, projection methods, and auto-encoders.

Principal Component Analysis (PCA), *Singular Value Decomposition* (SVD) and *Multidimensional Scaling* (MDS) are the most frequently used algorithms for Dimension Reduction [4].

2.8 Visualization

Data analysis, also called data analytics and visualization is joining the Big Data. It's essential and complex to visualize and interpret the large scale of data since they require significant data processing and storage capacity. Nowadays, we are having huge amount of data in a system due to the fast growth of the amount of data generated by computers, social media, mobile devices, industry 4.0, IoT, sports and so on.

It is crucial to have a process to acquire the data and to visualise it, in a traditional/advanced chart, tables, etc. The normal process to do it is shown in figure 2.17.

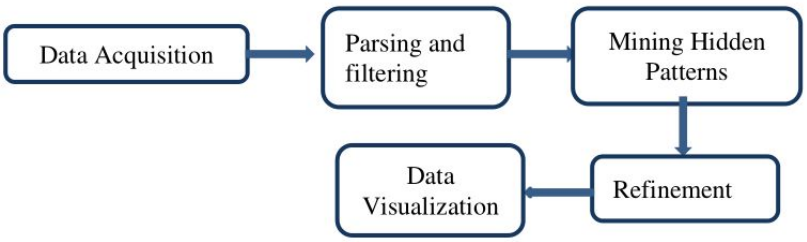


Figure 2.17: Data Visualization Process

The first step in the method of visualization is the retrieval of data from multiple sources. There could be unstructured/semi-structured data obtained from heterogeneous sources, so it needs to be parsed in a structured format. For visualization, all the data might not be necessary. The next move is to strip out the unimportant data. In the form of diagrams and charts, useful patterns are then derived and represented. Useful patterns are then extracted and depicted in charts and graphs in order to expose the users, and understand the secret knowledge. It allows data scientists to find secret data patterns and how they are stored. Business Analysts

may also use techniques of data visualization to define areas that require change or enhancement, concentrate on variables that affect consumer behaviour, and forecast revenue volumes.

Big Data visualization is complicated based on the number, variety, and speed of data. The biggest problem when dealing with Big Data is how to manage huge data volumes and efficiently show the practical and usable outcomes of data visualization and analysis [17].

Visualization is the process of displaying data in charts, graphs, maps, and other visual forms. Its used to help people easily to understand and interpret their data at a glance, and to clearly show trends and patterns that arise from this data.

Big Data visualization not only makes understanding and interpretation of data faster and easier, but its also a way of identifying and highlighting observations that might not be as noticeable when viewing a list of numbers and values.

Furthermore, raw data often comes in a variety of formats, so creating data visualizations is a vital part of an intensive process of gathering, managing, and transforming data into a format that's most usable and meaningful. The best data visualizations are those that clearly communicate an idea and simplify complex data, and they can be used by as many people as possible.

Line Chart, shown on figure 2.18, is the most used type of visualization that we are currently using for data analysis that displays information as a series of data points called 'markers' connected by straight line segments. A line chart is often used to visualize a trend in data over intervals of time.

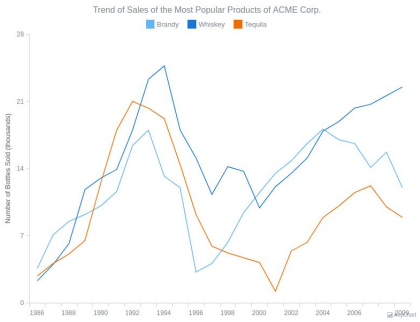


Figure 2.18: Data Visualization - Line chart

A Scatter plot, shown in figure 2.19, is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. The data is displayed as a collection of points, each having the values of one variable determining the position of the horizontal and vertical axis.

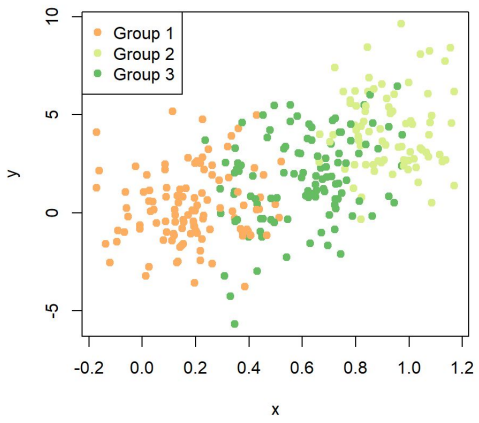


Figure 2.19: Data Visualization - Scatter plot

A Bar chart, shown in figure 2.20, presents categorical data with rectangular bars with heights/lengths proportional to the values. The bars can be plotted vertically or horizontally.

A Bar chart shows comparisons among discrete categories to show, as an example, the ranking of a set of data.

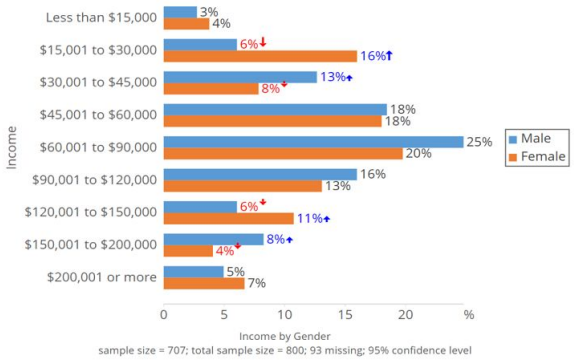


Figure 2.20: Data Visualization - Bar chart

Force-direct drawing algorithms are a class of algorithms for drawing graphs in an aesthetically-pleasing way. Their purpose is to position the nodes of a graph in two-dimensional or three-dimensional space so that all the edges are of more or less equal length and there are as few crossing edges as possible, by assigning forces among the set of edges and the set of nodes, based on their relative positions, and then using these forces either to simulate the motion of the edges and nodes or to minimize their energy.

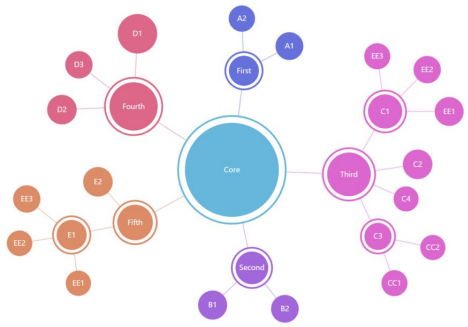


Figure 2.21: Data Visualization - Force-Direct Graph

A residual plot, shown in figure 2.22, is a graphical technique that attempts to show the relationship between a given independent variable and the response variable given that other independent variables are also in the model.

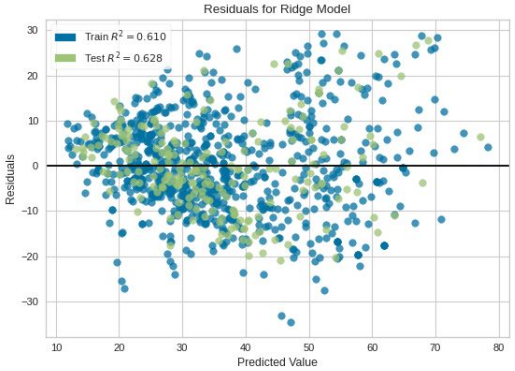


Figure 2.22: Data Visualization - Residual plot

Big Data is a potential research area, receiving considerable attention from academia and IT communication. The amount of data generated and stored expands within a short period.

In this "new digital world", with the evolution of mobile devices, digital sensors, IoT, communication, computing and storage have provided means to collect data. Data collected, has been a fundamental key for entrepreneur and organizations to enrich their knowledge in different areas.

Big Data is a collection of large amounts of complex data that cannot be manage efficiently by the state-of-the-art data processing technologies.

The data utilized to store and analysed a large scale of data cannot be operated by human being manually (amount of year will be needed to be provided with a complex solution of a certain problem by "hand"). Only advanced data mining and storage techniques can make the storage, management and analysis of enormous

data collected. The critical challenge for research is the exponential growth rate of data, which overloads the current ability of humans to design, analyse and manage the large amounts of data.

Challenges have been overpassed however, the storage architecture, computing systems, storage techniques, analytics techniques and user experience is being the centre of the evolution of Big Data.

Nowadays, technologies are the key to everything and Big Data has been increasing efficiency and as result, increase successfully the "system" analysis for any area [18].

Chapter 3

Understanding Basketball

Basketball is one of the most American sports played in the world. The game was invented on a cold day, in 1891, by a teacher of Physical Education at the International *Young Men's Christian Association* (YMCA) training school, in Springfield-Massachusetts, by James Naismith.

Naismith was a 31 year-old graduate student when he created the indoor sport to keep athletes indoors during the winters. The usual winter athletic activities were marching, callisthenics, and apparatus work but they were not nearly as thrilling as football or lacrosse which were played during the warmer seasons, that would be simple to understand but complex enough to be interesting.

Figure 3.1: James Naismith - Founder of Basketball

3.1 History

Naismith approached the school janitor, hoping he could find two square boxes to use for goals. When the janitor came back from his search, he had two peach baskets instead. Naismith nailed the peach basket to the lower rail of the gymnasium balcony, one on each side. The height of that lower balcony rail happened to be 10 feet (around 3 meters). The students would play on teams to try to get the ball into their teams basket. A person was stationed at each end of the balcony to retrieve the ball from the basket and put it back into play. The first game ever played between students was a complete brawl.

The humble beginnings of the only professional sport to originate in the United States laid the foundation for today's multi-billion-dollar business. The current *National Collegiate Athletic Association* (NCAA) March Madness college basketball tournament includes the best 68 of more than 1,000 college teams, stadiums that seat tens of thousands of spectators and lucrative television contracts.

Naismith didn't create all of the rules at once. For the time being, he was adapting and modifying the rules until got the 13 original rules.

The first public game of basketball was played in a YMCA gymnasium and was recorded by the Springfield Republican on March 12th, 1892. The instructors played against the students. Around 200 spectators attended to discover this new sport they had never heard of or seen before. In the story published by the Republican, the teachers were credited with "agility" but the student's "science" is what led them to defeat the teachers by 5-1.

Within weeks the sport's popularity grew rapidly. Students attending other schools introduced the game at their YMCAs. The original rules were printed in a college magazine, which was mailed to YMCAs across the country. With growth of the sport in popularity, it gained notice from the International Olympic Committee and was introduced at the 1904 Olympic Games in St. Louis as a demonstration event. High schools began to introduce the new game, and by 1905, basketball was officially recognized as a permanent winter sport.

As the sport continued its rapid spread, professional leagues began to form across the United States. Basketball fans cheered on their new home town teams. The first professional league was the *National Basketball League* (NBL) formed in 1898, and comprised six teams in the north-east. The league only lasted about five years. After it dissolved in 1904, the league would be reintroduced 33 years later in 1937 with an entirely new support system, with Goodyear, Firestone, and General Electric corporations as the league owners, and 13 teams.

While professional sports leagues gained nationwide attention, college basketball was also a major fixture. The first NCAA tournament, which included eight teams, was held in 1939 at North-western University. The first collegiate basketball

national champion was the University of Oregon. The team defeated Ohio State University[19].

3.2 Rules and Players

Across the years, the necessity of adjust the rules compared with the first 13 were crucial for basketball. There are 8 official rules, and each rule has a subsection for a better understanding. For details refer to FIBA website (www.fiba.basketball).

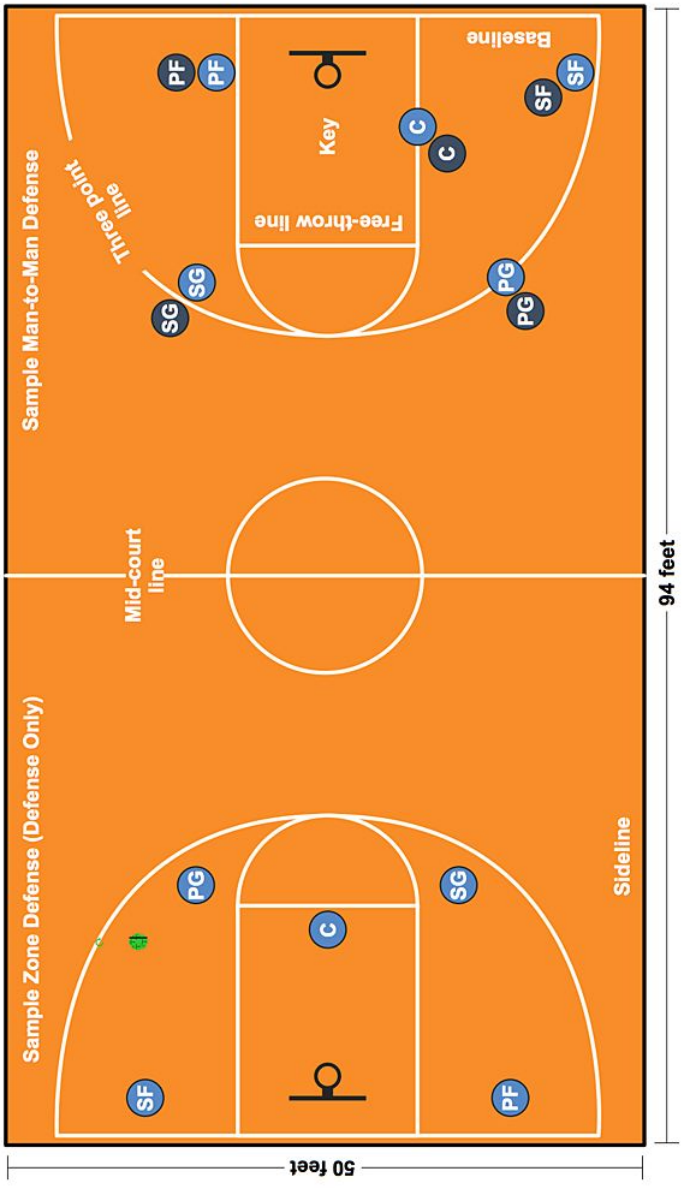


Figure 3.2: Basketball field diagram

A basketball team consist of at least 12 members entitled to play, including the captain, head coach and a maximum of 8 accompanying delegation member

(including 2 assistance coaches).

The team, in an official game, has five players. The tallest player is usually the "Center" (C), the second-tallest and strongest is the "Power Forward" (PF), a slightly shorter but more agile player is the "Small Forward" (SF) and the shortest player or best ball handler are the "Shooting Guard" (SG) and the "Point Guard" (PG) shown on figure 3.2.

Players advance the ball by bouncing it while walking or running (dribbling) or by passing it to a teammate. The aim of each team is to score the maximum number of points in the opposing team's basket. There are three ways to score point. The first, is to get 2 point when a player spear to the basket inside the "Three-point line". The seconds opposite of the first and get 3 point when a player is outside the "Three-point line". The third, is when a player gets 1 point through a "Free Throw". The "Free throw" occurs when a foul is made during the game and the player has two opportunities to get 2 points by a single point. To reach the 3-point, player has to spear to the basket outside of the "Three-point line" (shown in figure 3.2).

Each team can substitute their player during a time-out or an interval of play (a substitute request the substitution to the timer) or the referee beckons the substitute to enter the playing court. if in a game, a player has an injury, the referees may stop the game, only if the game is stopped otherwise and if only necessary, in case of an emergency, referees shall stop the game immediately and the injured player cannot continue to play.

The game shall consist of 4 quarters of 10 minutes each. During each quarters, a timer is present to track only the minutes that the game is running (when the ball becomes "live"). If the ball becomes "dead" (referee stops the game, there is a foul, etc.) the clock is stopped. The interval of the pay is 2 minutes between the first and the second quarter (first half). The second half is between the third and the fourth quarter with an interval of 15 minutes.

A foul happens when a player commits a personal foul, flagrant foul, technical foul, etc. Depending on the severity of the foul, the player can be excluded from the game after 5 personal fouls and the opposite team has the three-point loop, or the free loop by the corner line sides. For a better understanding, please visit the FIBA site.

At the end of the fourth quarter, the team with more points, win the game. In Basketball, the two teams cannot have the same score. If in a game this happens, an extra 5 minutes is played to have a winner.

At the end of the game, is attributed to the winner team 2 points on the score table and for the loser team, 1 point. In basketball, compared to a soccer game, the "home" and "visited" teams always gain points [20].

3.3 Season

A season, in all federated sports, is a period where the teams are playing with each other. This time is defined by the organization that is controlling the sport, *National Basketball Association* (NBA).

In basketball, the winter sport, season started at the beginning of November and ends in the middle of June as a typical period. The season of 2019/2020 in America (NBA) ends early due to Corona Virus (atypical season due to pandemic world).

Each team plays twice with the same opposite team. One game in the home team and another in the visitor team.

During a season, the team should gain the most of the games to win the championships.

3.4 Impacts

Every federated sport, has an impact on society. Some of them have more impact than others. In the case of Basketball, analysing the NBA League (placed in USA), has a huge impact at social, economic and even political levels.

This sport, in USA, has a lot of spectators. The total population in 2019 was 328.3 million people and the NBA league has 21.96 million spectators, being rated in the fourth position with most spectators.

With this amount of people, the sport is a hobby for some, and a business for others. Being the basketball games, there is an industry beginning for controller and management.

There is a tremendous economic impact that generates hundreds of millions of dollars in television viewership, advertisement revenue, in-arena purchases such as team jerseys and merchandises and the games ticket sales.

Everyone, or most people, has their favourite sport as well as their favourite team. People have a preference for their team to make an impact on society. When a team win the championship, there is a festival of it and sometimes, makes disturbs society.

There is a lot of impact on each sport and the NBA is one of the most that make it happens.

Chapter 4

Study of the behaviour of the USA NBA league (2018-2019 season)

Basketball is a sport well-known by Portuguese people but, is not the favourite one. In this chapter, we will study the behaviour of USA NBA league, classifying the data by clusters and in the end, show the performance of each cluster.

The basketball game has changed completely from a decade ago. Nowadays, is much more difficult to be a top basketball player than before. Many basketball players started playing in grade school or earlier. Players should look to practice as often as possible, challenging themselves to be stronger, taller, faster and better opponents.

The old game played was focused on the rotation and movement of the player synchronously, while in today's game, the player needs to run faster and be agile. The rotation is still important but at different level of movement speed. Today, the game is more agile and versatile.

The offensive and defensive strategies, as in other sports, are used often to get good results in a game. From a game we can get some player´s data of is behaviour, which are important to analyse later. Understanding the data behaviour as well as the performance of the player is important to the team owner, to make some decisions and reorganize his team. For example, and very common nowadays, this data can

be used to determine the best player, ranking classification of player, decision-maker to hire a player, and so on.

4.1 Data storage

Data for the USA NBA league is available at [21]. For each USA national league, the match result is catalogued for the given season. Also to be mentioned, the player performance and the statistics of each *Key Performance Indicator* (KPI) will be given. Therefore, for a specific match, we find the names of the home and the visitor teams, points, KPIs and dates.

To study a Big Data DB, we started to collect data from several seasons. The data, stored in an excel file, contain all seasons from 2015 until 2019. Due to the world Pandemic, the season of 2020 until 2021 was not taken into account in this study.

Maria Data Bases is one of the most popular relational DB. Was made by the original developers of MySQL and guarantee values of performance, stability and openness. MariaDB server was the DB selected to store all the information of these seasons.

During the process, to import the data into the DB, was used ETL, figure 4.1. This mechanism is used to extract the data from an original data source, transform it into pre-defined values, defined across in the DB, and load it in the DB. For this process, it was used python scripts.

It is also important to mention that it was required to build the DB structure. In the next subsection, we will detailedly explain its structure.

Figure 4.1: Data Preparation - ETL Mechanism

Raw data can be sent from various sources. Each source has its way of storing the data and sending the data in different formats (as an example, we have XML data, JSON data, File Transfer to be stored in a location, etc.)

The job of an ETL is to map the data from the source to the destination DB. Here, we can have an auxiliary process to re-build the data sent (that is not optimized

for analytics). With Python, we can easily re-build the information, but in Big Data, we already have applications to map the data and store it in the DB. ETL-Land and Storm are well used in this process. Using an exciting application available on the market, we can configure the type of data from different destinations to be stored. It is also important to note that the data type must be defined for further processing. Strings and numbers are the most common data types used in relational databases, but other formats, such as file storage (music, pictures, documents, and so on), are also possible and important.

In this study, we will use the data to cluster in different groups of players. Data will be extracted from spreadsheets, transformed, and loaded into the DB. The format of each attribute and the data type were taken into account for further analysis (in the future, to not make data convert in a proper format). Below, we can see a piece of code used to work the data. One of the principal attributes of a DB is to set the null values especially, when the columns are defined as numbers to be easy to perform the mathematical operation in the DB. Parts of the code made in Python, will be presented and explained below.

```
  Python code:

#!/usr/bin/python
import time
import csv
import pandas as pd
import numpy as np

t=time.asctime( time.localtime(time.time()) )
print (t)

filename = '.\games_details_python.csv'

df = pd.read_csv(filename, delimiter=",")

...
```

In Python code, it's important to know what we want to do. Once we are using CSV files, we used some of the libraries available. Pandas, is a library that can read easily and manipulate data in the CSV format. Numpy library is used for matrix and large-scale data and the CSV library is to read CSV files. It was also added, the time library to have the possibility to record the time expend of running the script. Data was loaded in the "df" variable which contains all the information of a CSV file. In this particular stage, the data to be imported will be the game details. This data contains all the KPI of each game, per player. After the import, we can define and manipulate the data in the right format to be imported in the DB.

```
  Python code:

  addstr = "'"

  df["PLAYER_NAME"] = df["PLAYER_NAME"].str.replace("'","''")
  df["COMMENT"] = df["COMMENT"].str.replace("'","''")
  df["TEAM_ABBREVIATION"] = df["TEAM_ABBREVIATION"].str.replace("'","''")
...
  #df["START_POSITION"].fillna('null', inplace=True)
  #df["COMMENT"].fillna("null", inplace=True)
  #df["MIN"].fillna("null", inplace=True)
...
  df['TEAM_ABBREVIATION'] = addstr + df['TEAM_ABBREVIATION'].astype(str) + addstr
  df['TEAM_CITY'] = addstr + df['TEAM_CITY'].astype(str) + addstr
  df['PLAYER_NAME'] = addstr + df['PLAYER_NAME'].astype(str) + addstr

  t1=time.asctime( time.localtime(time.time()) )
```

In the below code presented, the main changes are to replace the blank spaces of all attributes at null, to be in line with DB rules and to include "'" to be inserted in a DB.

```
  Python code:

  ...

  writePath = './save.txt'
  #with open(writePath, 'a') as f:
  #    f.write(
  df.to_csv(writePath, header = False, index = False, sep=',', mode='a')
  #   )

  with open('./output.txt', 'w') as out_file:
  with open('./save.txt', 'r') as in_file:
  for line in in_file:
  out_file.write(line.rstrip('\n') + "),"+'\n')

  t2=time.asctime( time.localtime(time.time()) )
  print (t2)
  print ("done")
```

After all the changes were made, we saved the file in a text file and worked again the data, in a SQL format, to be inserted in the DB. It was very helpful to use a script to normalize all data to be imported into the DB. Imagine making this step by hand. It will require a lot of effort without any benefit.

4.2 Database

Database technology has rapidly evolved in storage data for three decades and now, the Big Data concept, is growing even faster. Relational DB has been implemented as a set of software designs, as a software tool and, many of than offer interactive modelling capabilities, using a simplified data modelling approach.

The structure of a basic data relationship and its definition in a particular DB shall have a logical design. In logical design, relationships, entities and attributes are concepts that need to be structured in the DB. The three classes of objects (entities, relationships and attributes) are the key aspects of the data modelling.

- **Entities** are the principal data object about which information is to be collected. Usually, denoted a person, place, things, etc. For example, team, division, league and location are all examples of entities.
- **Relationship** represents the association among one or more entities. Are described in terms of degree.
- **Attributes** are characteristics of entities, that can provide details of an attribute. There are two types of attributes: identifiers and descriptors. An identifier (or key) is used to uniquely determine an instance of an entity. A descriptor is used to specify a non-unique characteristic of a particular entity instance. For example, an identifier or key of a player is the player-id, and a descriptor of is player-name or position.

The logic structure, implemented for the DB of the USA NBA league, is following some of the important concepts described above. As an entity, we select the "nickname" and in the attributes, we selected the "league id", "team id", "abbreviation", etc. for the table denominated as "team". The dataset, also includes different tables like "game", "Player", "Ranking", "Game" and "Game details". All this table has a relationship between "Team", "Player", "Ranking" and "Game details". To not have too much repetitive information, in this dataset, we have a table for description information like "teams" and "player" and, in the other table, we use the entities to link the information between tables [22].

One of the important features that the MariaDB server has, is the view of SQL statements (all relational DB vendors have included this feature). To reduce the number of SQL queries and to facilitate its usage of it (when we are querying the

DB often), created, by season, a view that contains all the data required to proceed with this study. This view includes data from "Games" and "Games details" table.

```
  SQL Query for a view:

  select `aa`.`season` AS `season`,
  `aa`.`game_date_est` AS `game_date_est`,
  `aa`.`game_id` AS `game_id`,
  `bb`.`team_id` AS `team_id`,
...
  round(cast(substring_index(`bb`.`min_`,':',1) as float) +
      cast(substring_index(`bb`.`min_`,':',-1) as float) /
      60,3) AS `min_played`,
...
  from (`games` `aa` join `games_details` `bb` on(`aa`.`
      game_id` = `bb`.`game_id`)) where `aa`.`season` = 2018
```

One of the aspects that is required to build a query, is the type of DB. Once we are using MariaDB, the default language is MySQL. We need to build our query with the right rules of MySQL. The first step to initiate the query is to include the "select". Select is used to start the query and then, we can filter the information, by attributes. Inserting the attributes is a process of knowledge. Its important to know which attributes are allocated in the table because we can have the same attribute in many tables. On the code, we are using the table alias and the attribute name and then, we give the name of this attribute (not mandatory). On each table, we are using an alias to be easier to build the query. As an example, we have the "`aa`.`game_id` AS `game_id`," selection, where the "'aa'" is the alias of a table, "game_id" is the attribute of the table "'aa'" and the "AS game_id" is the alias of the attribute. Notice, that on each query made, we cannot have duplicated alias in the same query. Now, in the "from" we can select the tables with their alias and apply some rules. MySQL has many rules but, for this project, we only use the "join" and the "where". Join function allow us to have multiple tables, sharing the same attributes and where, we can make the filter.

In figure 4.2, is shown the logical structure of the USA NBA league DB.

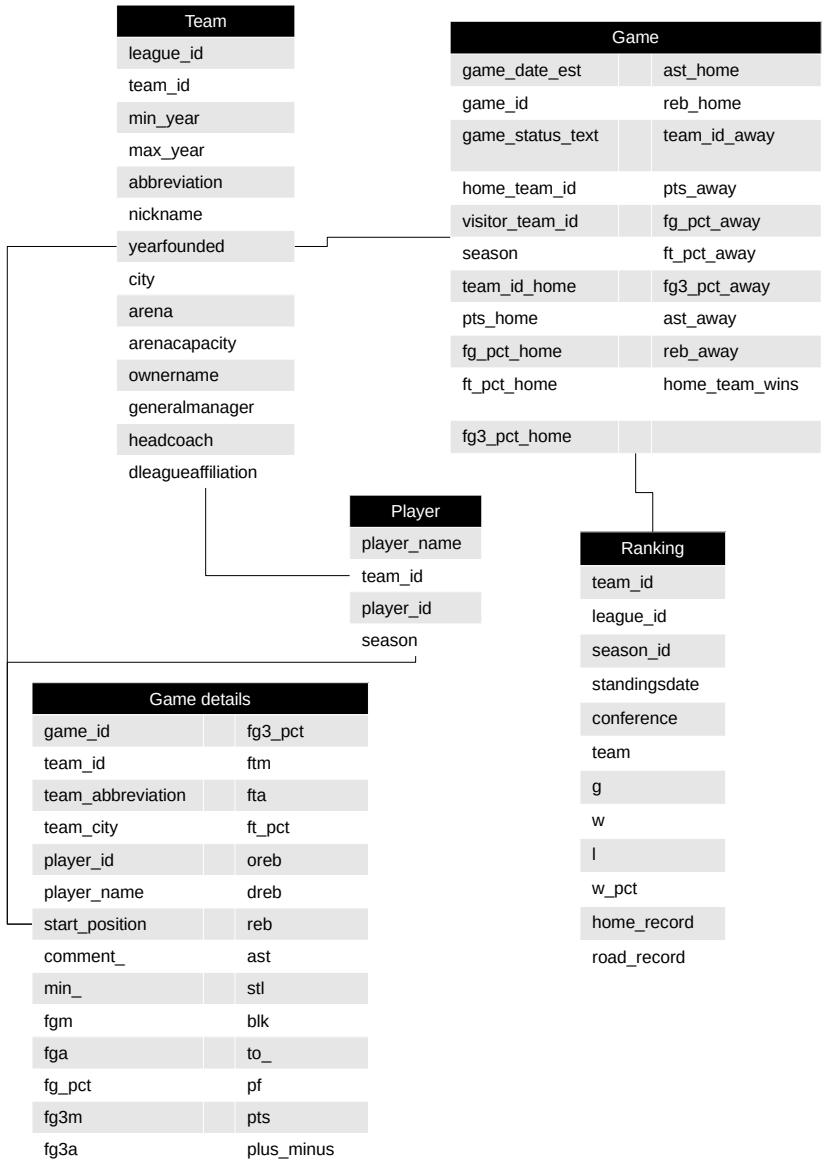


Figure 4.2: Database Structure - NBA data

Different tables were created to spread the information to different "locations". Organization in a DB is very important. Aspects like performance are needed, to have the information selected by a user at the proper time. The time consumed when a user is making a query to a DB has an impact on the defined structure. In this way, we created different tables for each piece of information collected.

On "Team", we have each team throughout the whole season, with relevant information like nickname, city, arena, owner, head-coach, etc. Table "Team" is the

main information for the other tables. "Team" has a relationship with the following tables: "Games", "Players", "Ranking" and "Game Details".

"Games Details" contains information by player, for each game played. *Minute Played* (MP), *Field Goal Made* (FGM), *Field Goal Attempt* (FGA), *Total Block Made* (BLK) and other types of information are stored to record the performance of each player.

In "Game", we have storage information by team. The season, the total points scored by the visitor and home team, and other factors are stored for further analysis. As example of it, we can easily check how many points the home or visitor team scored during a season. On "player", we gather information like the player, team and season. With this information, we can follow each player's progress throughout his career. And last, the "Ranking" table where is stored information by team, season, home and road record, and the point made.

As mentioned in subsection 2.2.3 - Aggregation, in the "Ranking" Table we have an aggregation mechanism. For a new entry on the "Game" table and with defined rules, the "Ranking" table will automatically update the information by team, league, season, home and road record, using mathematical operation, for example on the points, updated with the latest information.

The "Game details" table will be the information most used. It has the most important information to cluster the information into different group types.

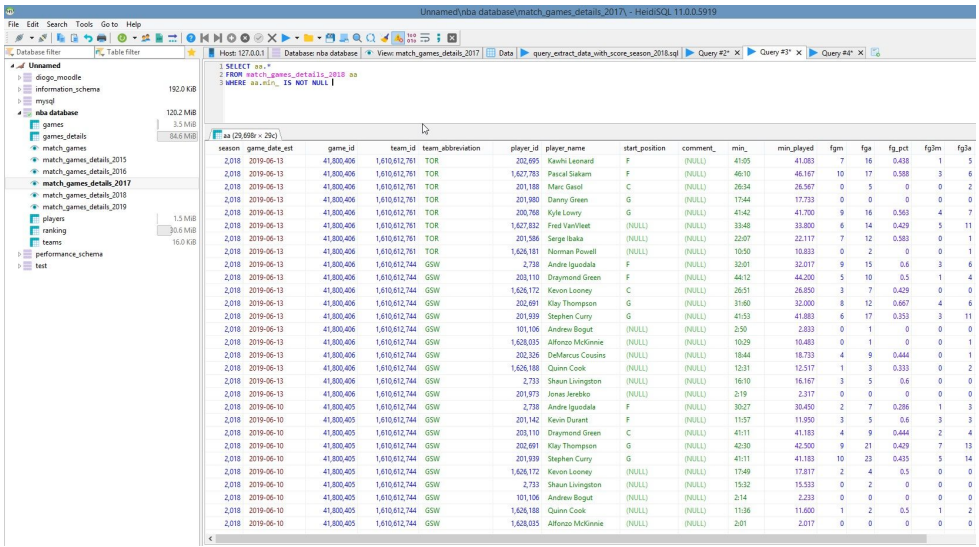


Figure 4.3: Database - MariaDB

To list all the data generated in figure 4.3, we used a simple query under the view "match_games_details_2018". The selection of the data was made, excluding the null values in the attribute minutes, having now the possibility to export the data, to be used for analysis.

The DataBase used in this study, contains a total of 120.2MB of data. Using the view, many people can work with it (in fact, is like a virtual table), and not have to create all the time complex query, like using join table, making operations in the DB, filtering the data, and so on. It's easy to create a view and brings benefits to simplify the query and a standardized way of accessing it.

4.3 Clustering algorithms

Cluster analysis provides insight into the data by dividing the objects into groups (clusters). These objects that are inside of a cluster, are more related and have a similar trend to each other, compared with data from another cluster. As the objects are not using other external information, such as class labels, the cluster analysis is called unsupervised learning in some traditional fields, such as, machine learning and pattern recognition.

The purpose of using cluster analysis is to understand the data and its utility of it. Understanding the cluster is a way to use the cluster analysis for automatic findings, in a group of objects that share common behaviour/characteristics, helping specialists to analyse, describe and utilize the valuable information hidden in a group. These techniques are being used often, having an important role in a wide variety of application domains such as business intelligence, psychology and social science, pattern classification, etc. Clustering analysis is always valuable for the exploration of unknown data emerging from real life, like basketball.

Cluster analysis has been used when, Karl Pearson used the moment matching method to determine the mixture parameters of two single variable components. Since then, tremendous research efforts have been made to design new clustering algorithms for cluster analysis. During the time being, the researchers have found some aspects that difficult the cluster analysis:

1. Clustering is essentially an inexhaustible combinational problem;
2. There exist no widely accepted theories for clustering;
3. The definition of a cluster seems to be a bit "arbitrary", which is determined by the data characteristics and the understanding of users.

These three points illustrated before, explain why there are so many clustering algorithms and why it is valuable to formulate the clustering problems as optimization problems, which can be solved by some heuristics, categorizing them into different categories.

- **Prototype-Based Algorithms:** Learn a prototype for each cluster, and forms a cluster by data objects around the prototypes. Some algorithms such

as K-Means and Fuzzy c-Means, the prototype of a cluster is a centroid and the cluster tends to be globular.

- **Density-Based Algorithms:** Takes a cluster as a dense region of data objects that are surrounded by region of low densities. They are often used when the cluster is irregular or when noise and outliers are present. DBSCAN and DENCLUE are two representative density-based algorithms. DBSCAN divides data objects into core points, border points and noise, respectively, based on the Euclidean density, and then finds the clusters naturally. DENCLUE defines a probability density function based on the kernel function of each data object, and then finds the clusters by detecting the variance of densities.

- **Graph-Based Algorithms:** Data can be representative in a graph when, data objects as nodes are used. The distance between two objects as the weight of the edge connecting the two nodes. The well-known *Agglomerative Hierarchical Clustering* (AHC), which merges the nearest two nodes/groups in one round until all nodes are connected.

- **Hybrid Algorithms:** Use two or more clustering algorithms in combination, in order to overcome the shortcomings of single clustering. Chameleon, is a typical hybrid algorithm, which first uses a graph-based algorithm to separate data into many small components and then, employs a special AHC to get the final cluster.

- **Algorithm-Independent Methods:** Called also as clustering aggregation or cluster ensemble, runs on the clustering results, and basic clustering algorithms rather than the original data. Given a set of basic partitioning of data, consensus clustering aims to find a single partitioning that matches every basic partitioning as closely as possible. It has been recognized that consensus clustering has merits in generating better clusterings, finding bizarre clusters, handling noise and outliers, and integrating partitioning of distributed or even inconsistent data.

Evaluation of a cluster, is a necessary step but also, a challenging task in cluster analysis. It is an evaluation of the cluster results. A validation measurement is a great need to tell us how the cluster is, classifying the information. At this moment, after evaluation and classification of the given cluster, the user can decide for each cluster [23].

In this use case, Study of the behaviours of the USA NBA league, we will use a clustering algorithm categorized as a prototype-base algorithm called K-Means.

4.4 K-Means implementation

K-Means is a prototype-based, simple partitional clustering algorithm that attempts to find K non-overlapping clusters. These clusters are represented by their centroids (a cluster centroid is typically the mean of the points in that cluster).

First, K initial centroids are selected, where K is specified by the user and indicates the desired number of clusters. Every point in the data is then assigned to the closest centroid, and each collection of points assigned to a centroid forms a cluster. The centroid of each cluster is then updated based on the points assigned to that cluster. This process is repeated until no point changes in clusters.

Centroid defines the central point of each cluster and will be updated, during the cycle of the algorithm, to address each point in a cluster.

Suppose D = {x1, · · · , xn} is the data set to be clustered. K-Means can be expressed by an objective function that depends on the proximities of the data points to the cluster centroids as follows:

$$min_{\{m_k\},1\leq k\leq K}\sum_{k=1}^{K}\sum_{x\in C_k}\pi_x dist(x,m_k), \tag{4.1}$$

where π_x is the weight of x, n_k is the number of data objects assigned to cluster

$$C_k, m_k = \sum_{x\in C_k}\frac{\pi_x x}{n_k} \tag{4.2}$$

is the centroid of cluster C_k, K is the number of clusters set by the user, and the function *dist* computes the distance between object x and centroid $m_k, 1 \leq k \leq K$. While the selection of distance function is optional,the squared Euclidean distance, i.e. $\|x - m\|^2$, has been the most widely used [23].

Once the clustering algorithm is explained, we can look for the data of the USA NBA league to try to identify the most appropriate variables to be used in the cluster.

The dataset contains a total of 29 columns of attributes (using the view created and explained in the subsection - DataBase) however, not all of it would be analysed in the K-Means algorithm. Manually, was chosen some attributes that we think are the most representative. Before implementing K-means clustering, a pre-processing of the dataset was made, mentioned in the subsection - Database, where was worked the data and pre-defined in a DataBases view. In the View, all the attributes are mentioned below, excluding the null data for the season 2018-2019. These attributes, in general, explain the offensive and defensive skills of basketball players.

The offensive attributes are *Point Made* (PTS), FGM, FGA, *Field Goal Percentage* (FG%), *Field Goal 3 Point Made* (FG3M), *Field Goal 3 Point Attempt* (FG3A),

Field Goal 3 Point Percentage (FG3%), *Free Throw Made* (FTM), *Free Throw Attempt* (FTA), *Free Throw Percentage* (FT%), *Assist Made* (AST) and *Offensive Rebound* (OREB).

While the defensive attributes are *Defensive Rebound* (DREB), *Total Steal* (STL), BLK and *Turnover Made* (TO). Both, offensive and defensive attributes, are also considered the MP. Depending on both offensive and defensive attributes, we will identify the most important variables, to be inserted into the K-Means algorithm, to spread the whole information in some clusters, for the dataset for the season 2018-2019.

A scatter and histogram plots, of each offensive attributes in the dataset are represented in figure 4.4.

Similar scatter and histogram plots, for each defensive attributes in the dataset are represented in figure 4.5.

In both histograms (figure 4.4 and figure 4.5), we can see that the data is not having a trend. Some of them have a Gaussian distribution and others, we can simply classify as data series. The time was considered in this analysis but, was not taken into account the period of each season, excluding the possibility of a time series analysis. Time series are analysed for a date or time, during a period. In this study, and after further analysis of each variable, defensive and offensive, where we can decide the most important variable to apply the K-Means algorithm, we will consider the MP and PTS, during each game played, under season 2018-2019. We consider the time played as a reference, to differentiate the attributes that need to be considered in this analysis.

```
MatLab code:

(...) loaded dataset from BD, with offensives and defensive
    attributes
(...) histogram plot for offensive and defensive attributes

X = [min_played, pts];

scatter(X(:,1), X(:,2));
title('Raw data - input for clustering')
xlabel('MP')
ylabel('PTS')
```

"X" variable, will be the matrix with data for the MP and PTS and the starting point to build the code for implementing the K-Means algorithm.

In figure 4.6, the ideal MP and PTS for outstanding players are between 30-40 MP and 20-30 PTS scored, considering that each game has 4 quarters (with 10

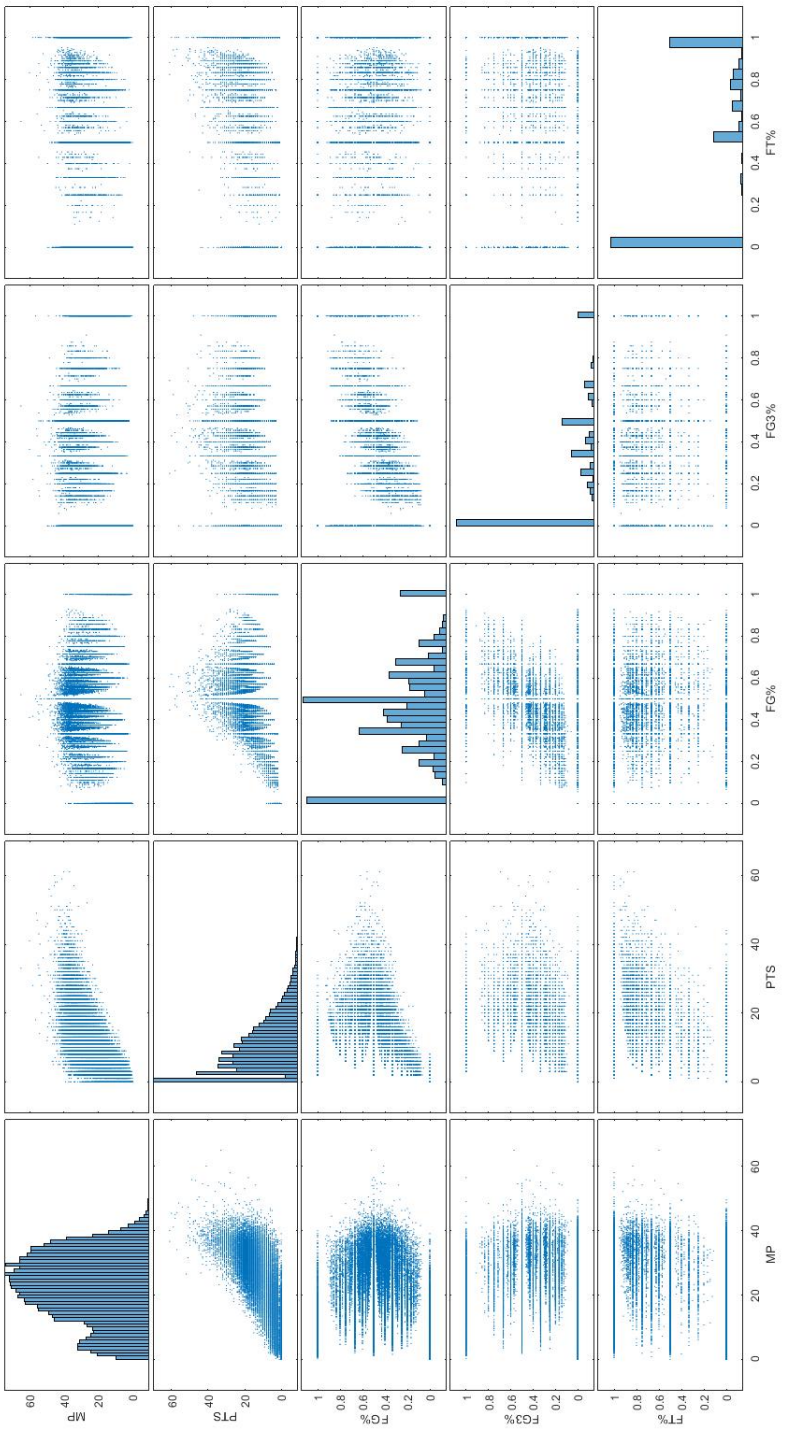


Figure 4.4: The Distribution of Offensive Attributes

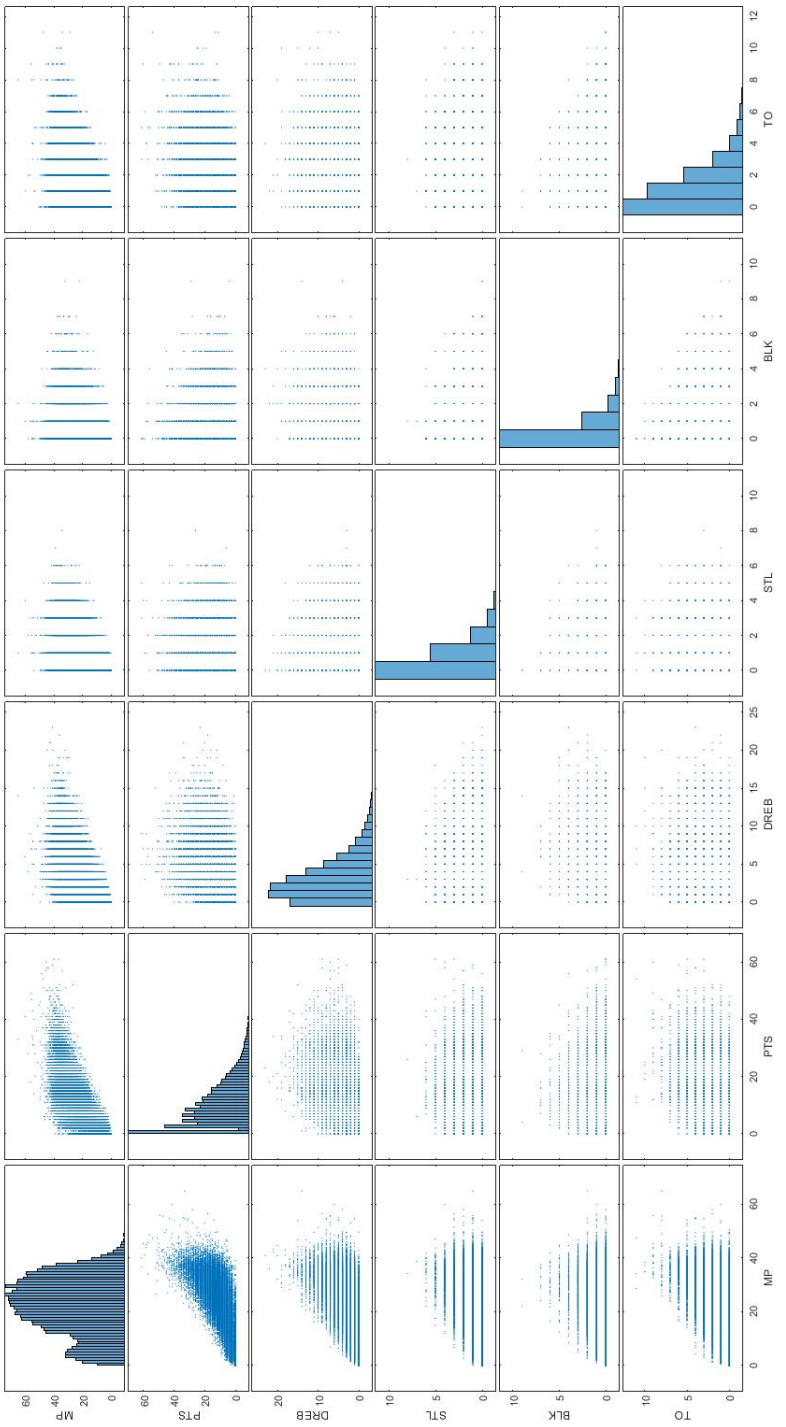


Figure 4.5: The Distribution of Defensive Attributes

minutes of duration each).

In Basketball, there is always a winner. Unlike football, there are no draws. When a game, at the end of 4 quarters ends in a tie, additional 5 minutes to be played is added. If at the end of the extra 5 minutes, there is a winning team, the game end's. Otherwise, there are more additional 5 extra minutes to be played.

MP is the aggregation times of 4 quarters that each player played during a full game. On Basketball, a player can be replayed for other teammates and can back to the game, multiple times.

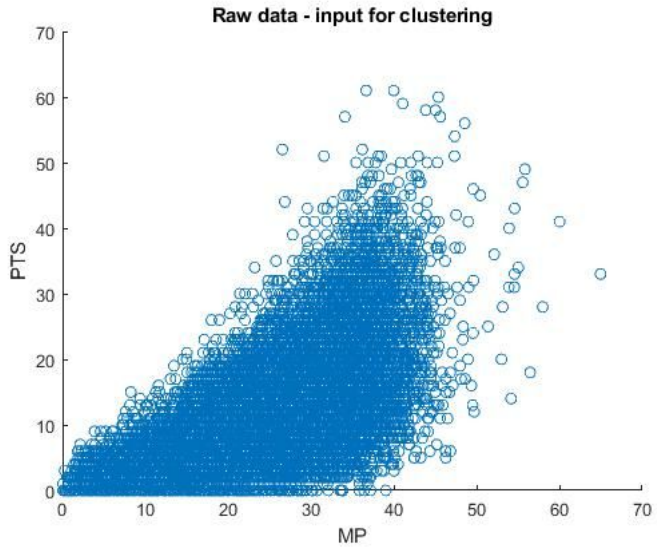


Figure 4.6: Raw Data input for Clustering

The input variable for clustering is defined and now, we also need to identify the separation of data into clusters.

NBA games are very versatile. Nowadays, players have good attributes in the offensive and defensive structure. The number of clusters, three, was chosen to determine the bad, intermediate or good and the outstanding players. With less than three clusters, it will be hard to segment the data and with only two groups, we can only have the bad and outstanding players. The intermediate or good will not be visible when a team doesn't have a good budget to buy the best players (as an example), data analysis will be more difficult because, this player will be on the edge of each cluster. More than three, does not allow us to reach a better understanding. It would be very segmented.

K (equal to three), will be our variable with the number of clusters to segment the dataset. For better accuracy in the clustering, and after some experiences, we also define the number of iterations to be five.

With the repetition of the K-Means algorithm, the centroid ("C_k"), initially selected as random values in the code, will be updated on each iteration, and the

algorithm will relocate the data from one cluster to another, obtaining the ideal division of data into three clusters.

The steps and the MatLab code will be described for a better understanding.

```
MatLab code:

...

K=3;
max_iters = 5;

rand_idx= randperm(size(X,1));
initial_centroids= X(rand_idx(1:K), :);
centroids = initial_centroids;
previous_centroids = centroids;
```

Three observations are randomly selected from the dataset as initial centroids. Note that the centroids should be unique. "K" defines the number of centroids and "max_iters" as the number of cycles to run the algorithm to have more precision in the outcomes. As mentioned, after some experience, we find an optimal cycle running the algorithm five times.

```
MatLab code:

...

for i=1:max_iters
[m,n]= size(X); % m and n are number of rows and columns
    of the dataset X ;
% Set K
K = size(centroids, 1);
dist= zeros(size(X, 1), K);
c=0;
for i= 1:m
for k=1:K
for j= 1:n
% the distance from each observation to a centroid
c= c+ (X(i,j)-centroids(k,j)).^2 ;
end
dist(i,k)= c; %distance of each observation from each
    centroid
c=0 ; % restart c to zero for next iteration
end
end
[M I]= min(dist , [], 2);
```

```
I; % a vector containing centroids closest to each
    observation
```

In the above code, the assignment of each centroid in the dataset to the closest centroids is made, returning a vector containing the closest centroid for each observation,m assigned to each cluster.

```
MatLab code:

...

  figure(4)
scatter(X(:,1), X(:,2), 15, 'w');
hold on

\%Data wrok by cluster
XX = [I,X];
ind1=XX(:,1)==1;
XX1=XX(ind1,:);
scatter(XX1(:,2), XX1(:,3),15,color_1,'p');
hold on

ind2=XX(:,1)==2;
XX2=XX(ind2,:);
scatter(XX2(:,2), XX2(:,3),15,color_2,'+');
hold on

ind3=XX(:,1)==3;
XX3=XX(ind3,:);
scatter(XX3(:,2), XX3(:,3),15,color_3,'d');

plot(centroids(:,1), centroids(:,2), 'x', 'MarkerEdgeColor
    ','k', 'MarkerSize', 10, 'LineWidth', 3);
for j=1:size(centroids,1)
plot([centroids(j, 1), previous_centroids(j, 1)], [
    centroids(j, 2), previous_centroids(j, 2)],'k');
str = strcat('\leftarrow',' ',num2str(centroids(j, 1)), ',
    ',num2str(centroids(j, 2)));
text(centroids(j, 1),centroids(j, 2),str);

end
hold off
title('K-Means, Centroid evolution');
legend('','Cluster 1','Cluster 2','Cluster 3', 'Centroid
    tracker');
```

Now, we will work the data and split the data by cluster to be shown in a

plot where, its also visible the path of each centroid cluster during the cycle of five iterations (visible in figure 4.8).

Considering the code implemented, in figure 4.7 we can see the first division of our population X (containing the MP and PTS scored by a player) into three clusters population and their respective centroids.

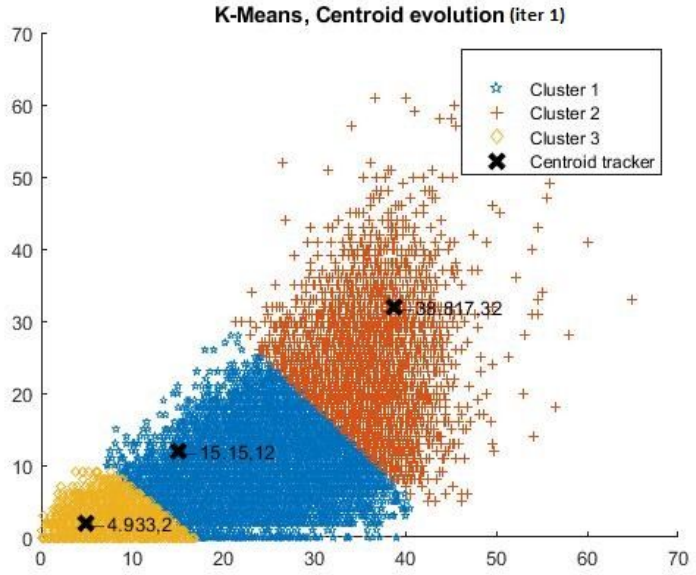


Figure 4.7: K-Means clustering - the first iteration

Clusters 1 and 2, is having more population than cluster 3 but, in this step, its hard to classify the type of player. Without an external data analysis, recalling the statistical methods, we cannot have this classification right now. Ending the implementation of the K-Means algorithm code and, running the 5 iterations, to update the centroid values to have more precision in each cluster population. For each iteration, the mean of each cluster will be made to update the centroid with a new value. This value, will be taken into account in the re-selection of each data player in the population cluster selection.

```
MatLab code:

...

clstr=0;
for k=1:K
clstr= find(I==k); % to find row index where idx==k
centroids(k,:)= mean(X(clstr,:));
clstr=0;
end
```

```
11    end
```

In figure 4.8 we can see the evolution of the K-Means algorithm.

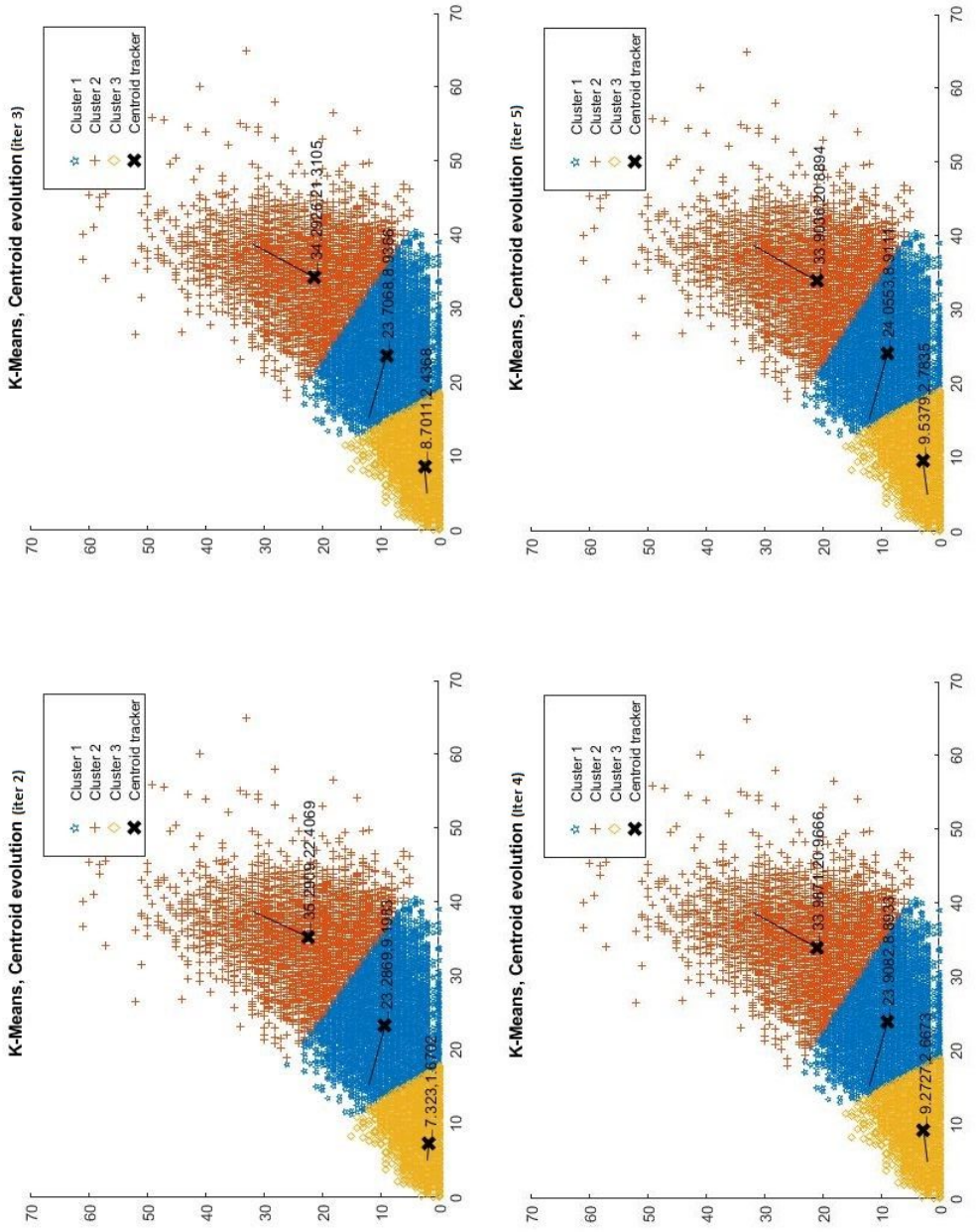


Figure 4.8: K-Means clustering - iteration evolution

Over time, and after the 1st iteration, the K-Means is adjusting the centroid values. In each adjustment, the raw data of each cluster are moving from one cluster to another, in other to obtain the optimal distance centre, the centroid point assigned to each cluster. From the first iteration to the second, we can see a big

adjustment in terms of the centroid (also tracked by a line) on clusters 2 and 3. The initial centroids were randomly selected by the system, causing a big adjustment.

As we only used 5 iterations (after some experience and analysing of the centroid values, we conclude that 5 iteration are a good value to achieve the right accuracy) to find the optima distance centroid in each cluster, and after the second iteration, small adjustments in the centroid was made. In these 4 steps, more precision is given in the clustering. On figure 4.9 we can see the final population for each cluster.

```
MatLab code:

...

figure(5)
hold on
scatter(X(:,1), X(:,2),'w');
hold on
XX = [I,X];
ind1=XX(:,1)==1;
XX1=XX(ind1,:);
scatter(XX1(:,2), XX1(:,3),15,color_1,'p');
hold on
ind2=XX(:,1)==2;
XX2=XX(ind2,:);
scatter(XX2(:,2), XX2(:,3),15,color_2,'+');
hold on
ind3=XX(:,1)==3;
XX3=XX(ind3,:);
scatter(XX3(:,2), XX3(:,3),15,color_3,'d');
% Plot the centroids as black x's

for j=1:size(centroids,1)
plot(centroids(j,1), centroids(j,2), 'x', 'MarkerEdgeColor
    ','k', 'MarkerSize', 10, 'LineWidth', 3);
end
hold off
% Title
title(sprintf('K-Means, grouping cluster = %d', K))
xlabel('MP')
ylabel('PTS')
legend('','Cluster 1','Cluster 2','Cluster 3', 'Centroid
    per cluster');
```

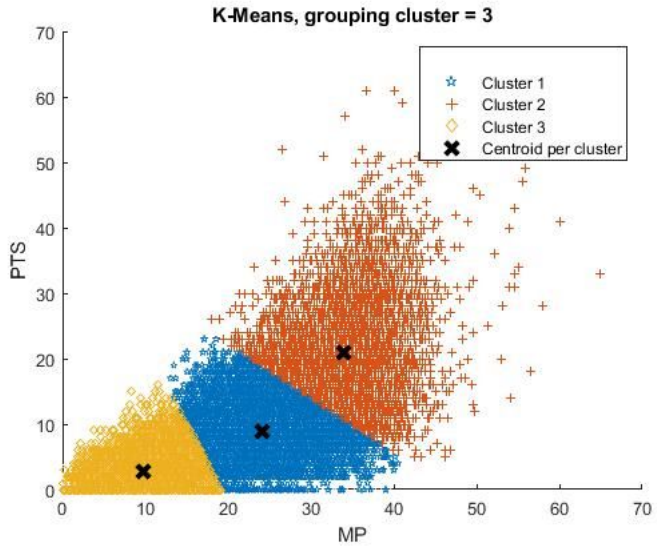


Figure 4.9: K-Means clustering - Final cluster

Raw data now are assigned to each cluster, meaning that all data players were clustered. Using the offensive and defensive attributes, and classifying them to a specific cluster, we can now make an analysis. "X1", is the variable that is stored, in a matrix, all the attributes. "I", is the matrix from the previous code, where is located the cluster number of each raw data of the given "X" variable (containing the MP and PTS values). Once "X" was extracted by the "X1" matrix, we can guarantee that each position of "X" will be the same in "X1", classifying each data value in "X1" with the right cluster number of "I".

```
MatLab code:

...

data = [I, X1];

data_cluster1_temp = ismember(data(:,1),1);
data_cluster1 = data(data_cluster1_temp,:);

data_cluster2_temp = ismember(data(:,1),2);
data_cluster2 = data(data_cluster2_temp,:);

data_cluster3_temp = ismember(data(:,1),3);
data_cluster3 = data(data_cluster3_temp,:);
```

The data now is classified by groups and using the reference cluster of each point, by filtering the offensive and defensive attributes. The mathematical methods and matrix methods were used often to generate and process the data to obtain the

following charts. On matrix, we used the transport to have more than one cluster in an abscissa axis and the average to obtain the mean of each cluster, by each attribute.

For the whole data of each cluster, we applied the average for each attribute. MP and PTS, will be presented in the offensive and defensive analysis. They are the Key to clustering. The results are displayed in figure 4.10 and 4.11.

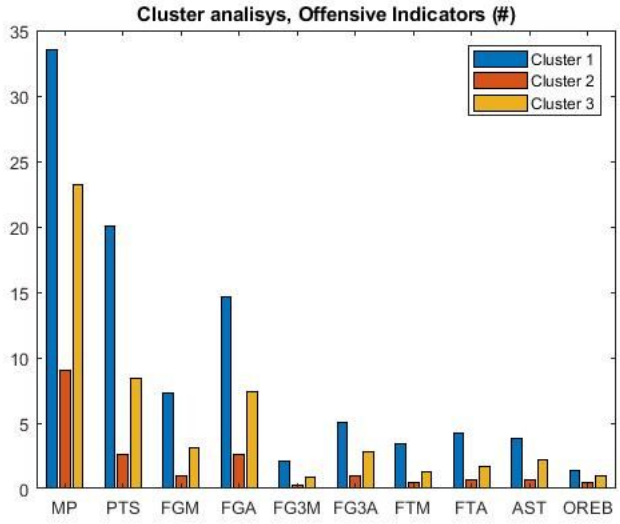


Figure 4.10: Cluster Analysis - Offensive Indicators

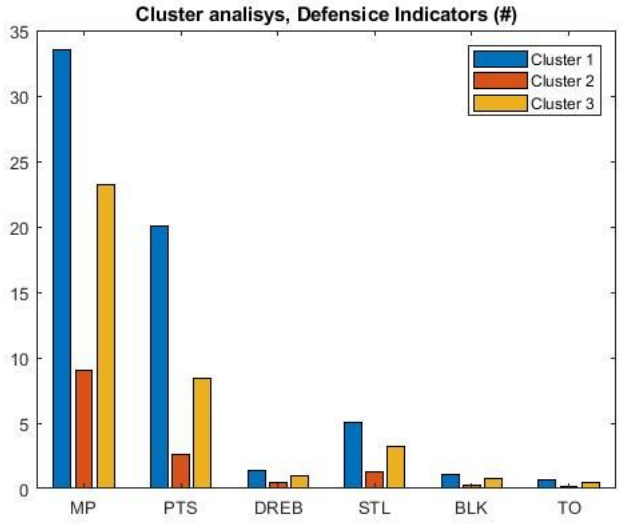


Figure 4.11: Cluster Analysis - Defensive Indicators

The results are quite similar for each attribute studied, offensive or defensive. Filtering the additional data, the clustering was made between the MP and PTS, we can conclude that cluster 1 is in the first ranking, saying that the outstanding

player was located in cluster 1. Next, we can classify cluster 3 as second and the last cluster, as third.

Focusing on cluster 1, MP, PTS and FGA is having a high average compared with the other attributes. FG3A, FTM, FTA, AST and STL are also having quite good results. The player of cluster 3, had good skill scoring a basket but, also had a good sense of defence. Overall, this kinds of players are the players that are good in offence and defence. Analysing of all clusters, by field goals success, presented in figure 4.12, the difference is not too much. On FG(%) and FG3(%), we can see a difference of 10 percent between clusters but, in this analysis is also important to mention, that in clusters 2 and 3, players have less MP, PTS and FGA when compared to the ones in cluster 1.

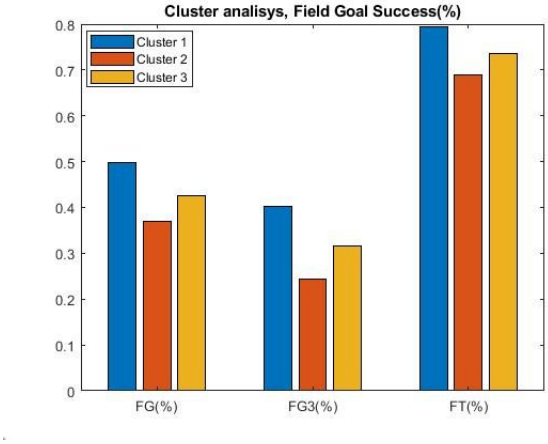


Figure 4.12: Cluster Analysis - Field Goal Success

Let's see some players and their statistics for the cluster with better performance and making a top 10 rank, by PTS only, to list the player.

Table 4.1: Top 10 Player in Cluster 1

Player name	**Point made**
James Harden	3232
Damian Lillard	2494
Stephen Curry	2483
Paul George	2340
Kevin Durant	2332
Giannis Antetokounmpo	2284
Kawhi Leonard	2276
Kemba Walker	2070
Bradley Beal	2070
Klay Thompson	2038

Analysing the result, and according with the site [24], we can conclude that 50% of players are within our results.

Table 4.2: Top 10 Players in Cluster 1

Player name	**Ranking**
LeBron James	1
Kevin Durant	2
Stephen Curry	3
Anthony Davis	4
James Harden	5
Russell Westbrook	6
Kawhi Leonard	7
Giannis Antetokounmpo	8
Chris Paul	9
Joel Embiid	10

LeBron James, in the season, was considered the best player and is not included in table 4.2. This means that this analysis, is not only taking into account the sum of PTS for the top 10 rank. In the list, we can see James Kevin Durant, Stephen Curry, James Harden, Kawhi Leonard and Giannis Antetokounmpo, classified as the top 10 best players in this season [24].

Chapter 5

Conclusion

The purpose of this study was to offer a literature overview on the topic of Big Data and data analysis in Basketball, applying numerical and computer methods. This began with the presentation of a general background of Big Data concepts, including Big Data definitions and characteristics.

Big Data is a significant area which offers many potential benefits and innovations. It is a remarkable domain with a promising future, if approached correctly. The difficulty with Big Data comes mainly from its size, which requires proper storage, management, integration, cleansing, processing, and analysis. The 5V's: Volume, Velocity, Variety, Veracity and Value of data, increase the difficulty of dealing with it in terms of traditional data management. This creates a need to study and explore new analysis methods which might help in overcoming such difficulties to promote the positive role of Big Data analysis to as many sectors as possible. Future research could thus usefully focus on Big Data analysis challenges.

The method used in this study is the K-Means algorithm and was explained in enough detail to understand the aim of this work. First, we covered the principle mechanics with the mathematical background of the method, explained it in detail and then, moved to the implementation of the code, in Matlab. MatLab is a powerful tool used, where we can apply mathematical and numerical methods as well as, obtaining visualizations, using the available charts to show the result of this study.

The results obtained in this study are good. Differentiating initially the bad, good and the outstanding player, using the appropriate variables (MP and PTS) was the right choice. Once set the input raw data in three clusters - bad, good

and outstanding players, we obtain the data into the corresponding clusters, helping in the selection of the players to be studied. Once we were looking for the players with outstanding performance, applying statistical methods and visualizing the data to find the cluster with outstanding player performance. To enhance the study, and using the information of each data point identified as belonging to a cluster, using the offensive and defensive attributes, we could find the cluster with the best performance, for our case study.

The cluster selected with the best performance, had half of the players classified as top 10 in the season 2018-2019 by the site "TSJ101 sports!". Summing the scored points obtained in the cluster 1 and comparing with data from the site and, taking into account some percentage error, we can conclude for the top 10 ranking, that maybe other attributes are used to classify the players.

5.1 Further Work

The implementation of mathematical and numerical methods in Big Data was an important approach in the scope of this work. We used the K-Means algorithm to cluster the information into some groups. The goal of this study was to identify what players had an outstanding performance, using the defensive and offensive attributes. With the DB implemented and with the historical data to be analysed, in the future, we would like to implement prediction methods in the cluster with the best performance, to try to find out the top 10 players for a season and in the end, compare it with reliable data. Another approach could be to use a different clustering implementation that allows to study other game aspects or, for example, allows to automatically determine the best number of clusters.

The other work that could be done is the implementation of a machine learning algorithm such neural networks. Neural Networks is a powerful algorithm, designed to recognize patterns without being explicitly programmed (supervised and unsupervised learning).

Also important to mention is the utilization of data in table 4.2. With the top 10 ranking players, we can create a dataset to use an supervised approach that will help us to identify high performance standards, improving the performance of the model implemented.

References

[1] CleverISM. Brief history of big data. https://www.cleverism.com/brief-history-big-data/. Accessed: 13-01-2022. [Cited on pages 5 and 7]

[2] Datafloq Data and Technology Insights. A short history of big data. https://datafloq.com/read/big-data-history/. Accessed: 15-01-2022. [Cited on page 6]

[3] Paul Buhler Thomas Erl, Wajid Khattak. *Big Data Fundamentals Concepts, Drivers and Techniques*. Prentice Hall, 2015. [Cited on pages 10 and 11]

[4] Bernard Marr. *Big Data Science and Analytics - A Hands-On Approach*. Arshdeep Bahga and Vijay Madisetti, 2019. [Cited on pages 12 and 27]

[5] Xiaotong Shen Wolfgang Karl Hardle, Henry Horng-Shing Lu. *Handbook of Big Data Analytics*. Springer, 2018. [Cited on page 13]

[6] IEEE. Big data challenges and data aggregation strategies in wireless sensor networks. https://ieeexplore.ieee.org/document/8353765. Accessed: 31-01-2022. [Cited on page 13]

[7] wmich.edu. Multi-resolution hierarchical structure for efficient data aggregation and mining of big data. https://amity.edu/icactm/Proceeding/Paper. Accessed: 12-02-2022. [Cited on page 15]

[8] Nasir Raheem. *Big Data, A Tutorial-Based Approach*. CRC Press, 2019. [Cited on page 16]

[9] Amir Vahid Dastjerdi Rajkumar Buyya, Rodrigo N. Calheiros. *Big Data Principles and Paradigms*. ELSEVIER, 2016. [Cited on page 17]

[10] techopedia. Big data platform. https://www.techopedia.com/definition/29951/big-data-platform. Accessed: 16-02-2022. [Cited on page 17]

[11] Dataconomy. Understanding big data - infrastructure. https://dataconomy.com/2014/06/understanding-big-data-infrastructure/. Accessed: 20-02-2022. [Cited on page 18]

[12] Research and Data Alliance. Big data security - issues, challenges, tech and concerns. https://www.rd-alliance.

org/group/big-data-ig-data-security-and-trust-wg/wiki/big-data-security-issues-challenges-tech-concerns. Accessed: 05-03-2022. [Cited on page 18]

[13] Yusuf Aytas. *Designing Big Data Platforms, How to Use, Deploy, and Maintain Big Data Systems.* Wiley, 2021. [Cited on pages 19 and 21]

[14] Vinit Sharma. *The Cloud-Based Demand-Driven Supply Chain.* Wiley, 2019. [Cited on page 20]

[15] International Journal of Computer Science and Information Technologies. No science no humans, no new technologies no changes. [Cited on page 21]

[16] John Morton. *Big Data Opportunities and Challenges.* bcs The Chartered Institute fot IT, 2015. [Cited on page 22]

[17] Subhi R. M. Zeebaree Zhwan M. Khalid. *Big Data Analysis for Data Visualization - A Review.* International Journal of Sicence and Business, 2021. [Cited on page 28]

[18] Abdullah Gani Salimah Mokhtar Ejaz Ahmed Nor Badrul Anuar Athanasios V. Vasilako International Journal of Information Management by Ibrar Yaqoob, Ibrahim Abaker Targio Hashem. Big data: From beginning to future. https://www.researchgate.net/publication/305736330_Big_Data_From_Beginning_to_Future. Accessed: 26-03-2022. [Cited on page 31]

[19] National Geographic. Here's the history of basketball—from peach baskets in springfield to global phenomenon. https://www.nationalgeographic.com/history/article/basketball-only-major-sport-invented-united-states-how-it-was-created. Accessed: 09-02-2022. [Cited on page 35]

[20] FIBA Central Board. 2020 official basketball rules, basketball rules and basketball equipment. https://www.fiba.basketball/documents#tab=efb3a7a8-15d1-494b-8070-f55bd809304c. Accessed: 17-02-2022. [Cited on page 36]

[21] Fixture. Download basketball fixtures, schedules and results. https://fixturedownload.com/sport/basketball. Accessed: 07-03-2022. [Cited on page 40]

[22] Tom Nadeau H.V. Jagadish Toby J. Teorey, Sam S. Lightstone. *Database Modeling and Design, Logical Design.* MORGAN KAUFMANN publishers, fourth edition edition, 2005. [Cited on page 43]

[23] Junjie Wu. *Advances in K-means Clustering - A Data Mining Thinking.* Springer, 2021. [Cited on pages 48 and 49]

[24] TSJ101 Sport. 2018-19 season top 10 nba players. `https://tsj101sports.com/2018/09/26/2018-19-season-top-10-nba-players/`. Accessed: 23-06-2022. [Cited on page 62]