Similarity Analysis to aid decision making on NBA Draft

Miguel Alejandro Houghton López

Bachelor's Degree Final Project in Computer Science and Engineering

Facultad de Informática

Universidad Complutense de Madrid

2019 - 2020

Director: Dra. Yolanda García Ruiz

Análisis de la similaridad para la toma de decisiones en el Draft de la NBA

Miguel Alejandro Houghton López

Trabajo de Fin de Grado en Grado de Ingeniería Informática

Facultad de Informática

Universidad Complutense de Madrid

2019 - 2020

Director: Dra. Yolanda García Ruiz

Contents

List of figures

List of tables

Abstract

This work is based on the different statistical studies published by Mock Draft Websites and on webs that store the official statistics of the NBA players. The data associated with NBA players and teams are currently very precious since their correct exploitation can materialize in great economic benefits. The objective of this work is to show how data mining can be useful to help the scouts in this real problem. Scouts participating in the Draft could use the information provided by the models to make a better decision that complements their personal experience. This would save time and money since by simply analyzing the results of the models, teams would not have to travel around the world to find players who could be discarded for the choice. In this work, unsupervised grouping techniques are studied to analyze the similarity between players. Databases with statistics of both current and past players are used. Besides, three different clustering techniques are implemented that allow the results to be compared, adding value to the information and facilitating decision- making. The most relevant result is shown at the moment in which the shooting in the NCAA is analyzed using grouping techniques.

Keywords: Draft NBA, Basketball, Clustering analysis, Chernoff faces, R programming.

Resumen

Este trabajo se basa en los diferentes estudios estadísticos publicados en páginas web de predicción de Drafts de la NBA y en webs que almacenan las estadísticas oficiales de los jugadores de la NBA. Los datos asociados a los jugadores y equipos de la NBA son actualmente muy valiosos ya que su correcta explotación puede materializarse en grandes beneficios económicos. El objetivo de este trabajo es mostrar cómo la minería de datos puede ser útil para ayudar a los entrenadores y directivos de los equipos en la elección de nuevas incorporaciones. Los ojeadores de los equipos que participan en el Draft podrían utilizar la información proporcionada por los modelos para tomar una mejor decisión que la que tomarían valorando su experiencia personal. Esto ahorraría mucho tiempo y dinero ya que, simplemente analizando los resultados de los modelos, los equipos no tendrían que viajar por el mundo para encontrar jugadores que pudieran ser descartados para la elección. En este trabajo se estudian técnicas de agrupación para analizar la similitud entre jugadores. Se utilizan bases de datos con estadísticas de jugadores tanto actuals como del pasado. Además, se implementan tres técnicas de clustering diferentes que permiten comparar los resultados, agregando valor a la información y facilitando la toma de decisiones. El resultado más relevante se muestra en el momento en el que se analiza el tiro en liga universitaria mediante técnicas de agrupación.

Palabras clave: Draft NBA, Baloncesto, Análisis Clúster, Caras de Chernoff, Programación en R.

Chapter 1. Introduction

1.1 Motivation

Technology have brought us new ways of transmitting digital information. Social media, websites, smartphones and smartwatches are some tools that allow us to manage huge amounts of data. All these data are useful for any business and are a very valuable asset. It's a great way to improve and develop new ways of earning economic power and innovate in any area of a company. Descriptive data analysis and predictive data analysis have become two of the best ways of studying new patrons and behaviors in the information and helped the different business to innovate. But the data exploitation is not only done in the business world. In recent years, data has taken a center stage in the sports sector and is currently a very important part of any sports activity. Athletes are great generators of data, and without a doubt, the world of sports is undergoing a transformation. In any sports event, such as a football or basketball game, around 8 million events and data are generated (field condition, goals, jumps, distances traveled, sanctions, players' fitness, passes, dribbles, etc.). The treatment of this information offers the opportunity to make the most appropriate decisions, such as choosing a strategy during a match, improving performance, substitute a player to prevent injuries, transfer management, etc. For this, it will be necessary to apply the most modern statistical and computational methods.

In recent years, sports analysis has become a very attractive field for researchers. In [1] theycollect psychometric data on athletes during their competition season to study how anxiety and mood influence the performance of players. Image recognition is used to monitor the position and movement of players in real time. This information allows inferring other data like the player's speed, distances between players, etc., very valuable information for coaches. Applying modern statistical and computational techniques to the data obtained from previous matches, it is possible to detect patterns in the competitors. Consequently, coaches can predict the decisions of the opposing team and modify their own tendency. The use of sensors in athletes (see [2]) allows a large amount of data to be collected during training (level of fatigue, heart rate, etc.). The processing of this data can be used to design personalized training routines. The digital trail that athletes leave

on social networks is also valuable information for Clubs when making signings, since with the analysis of behavior it is possible to predict whether an athlete will give the Club a good or bad image in the future. Signing a star player can be expensive and a great risk if he is injured or has a bad image.

Basketball is one of the biggest sports in the world. Nowadays, is the sport that has the most impact in the economy worldwide. Thanks to great basketball leagues such as NBA, Euroleague, ACB or CBA, basketball moves billions of dollars every year. For example, one of the most valuable teams in the NBA, the New York Knicks, has an average value of 4 billion dollars. One of the biggest issues about every general manager is concerned, is the optimal composition of their teams. Building the best team possible according to the budget of any team is a complex process to accomplish. To do this, the team management must take into account several factors such as everything related to the players (physical, age, statistics), as well as everything related to the economic power of the team (renewal of contracts and salary limits for instance). In addition, in the NBA, to give dynamism to the league and make it as even as possible, Draft night is held every year. This event consists of the selection of the best university or other league players to become part of an NBA team. Teams place a high value on Draft, given that on numerous occasions their primary choice on Draft night can shape the future of an entire franchise.

The night of the Draft consists of two rounds, of 30 players each. The first elections in the first round are divided by lottery among the 16 teams not classified for the NBA playoffs. The rest of the elections of the round go to the teams classified in the playoffs of the previous season. The draw to know which position each team chooses in the Draft is called the 'NBA Draft Lottery' and the chances of choosing higher positions depend on the number of losses in the regular season of the previous year. Given the importance of this choice, management teams invest a lot of time and money in getting the best possible choice for the position that has been obtained in the Draft. This choice is conditioned by the needs of each team at all times. For example, if a team has a solid player in the Center or Power-Forward position, the wisest course would be to 'draft' a smaller player to cover the shortcomings of the team.

The objective of this work is the design and analysis of different statistical models based on real statistics. The teams participating in the Draft could use the information provided by the models to make a better choice and not only based on personal intuition. This would save time and money, since by simply analyzing the results of the models, teams would not have to travel around the world to find players who could be discarded for the choice.

1.2 Objectives

The general objective of this work is to show how data mining can be useful in a real-life problem as it is the Draft problem. More specifically, this work aims to carry out a follow-up and a complete study of each Draft class in order to help teams to choose the best player, based on real statistics and visualizations of data mining analytics as a complement to the expertise of the basketball scouts.

1.3 Methodology and work organization

During the development of this study, I've noticed the great influence of the knowledge acquired on my studies of Computer Science, especially thanks to subjects like 'Statistics', 'Data Mining and Big Data Paradigm' and 'Data Bases'. In general terms, the critical vision that gives you the development of the different areas of the degree has been essential for this essay. Thanks also to some subjects of the degree related to programming, I've been able to install and develop all the code necessary to carry out this project. The methodology used in this Final Degree Project is to carry out the following tasks:

1. Search of information and state of the art.
2. Location and download of required datasets in the study.
3. Cleaning and integration of datasets.
4. Documentation of data.
5. Installation of the necessary software tools and libraries.
6. Study of the applications and contact with the data.
7. Analysis of the data using the available libraries and own code.
8. Visualization of results.
9. Writing this report.

This work is structured in the following way. Chapter 2 presents the state of the art providing a critical vision of the apps and other work available. Besides it includes a general description of the clustering techniques that we choose for the study. Chapter 3 describes the materials, datasets and software tools that have been used throughout the project. Chapter 4 explains the statistical methods and measurement techniques used from a theoretical point of view. Chapter 5 shows the results obtained after applying these techniques. Finally, the conclusions and future work are proposed in Chapter 6. The bibliography necessary to complement the information is also included, which is also a reference to the sources of information used in this work.

Chapter 2. State of the Art

This study is based on the different statistical studies made by Mock Draft Websites and on webs that store the official statistics of the NBA players. The data associated with NBA players and teams are currently very precious, since their correct exploitation can materialize in great economic benefits. Along the lines of our work, there are several companies that perform sophisticated analytics as a result of complex game and team studies, and which are subsequently published on their websites for consumer consumption. In general, these companies use a wide variety of data; data related to the physical conditions of the players (height, weight, injury history, etc.) and data collected during games. Many of these companies predict player selection in the NBA Draft.

As we can suspect, this statistical information is of great value in the betting sector since the fans guide their betting decisions in each Draft based on the information cooked by these companies. These websites have a great influence on the betting industry because every year the NBA Draft bets move millions of dollars. You can also look for Databases that store lots of stats of active and inactive players. The fact that you can use the data from retired players can make the study way more complete, because thanks to this, we are going to be able to compare the prospects with any player who has ever played.

This chapter is divided into two sections. The first is dedicated to existing applications whose objective is related to that of the present work. Although the objective of this work is not to implement an application of this type, its analysis is helpful to understand all the elements of the study. The second section is dedicated to grouping techniques. In addition to an introduction to these techniques, the techniques used at work are explained in more detail and the results of which are presented in the corresponding chapter.

2.1. Related work

In this section the most known Draft prediction and NBA stats websites are described. Every item contains details of the most important features.

- **NBADRAFT.NET**

The company Sports Phenoms.Inc [3] publishes on its website (see Figure 1) a complete analysis of players who will participate in the next NBA Draft. For each player, not only

physical characteristics such as height and weight are included, but also different characteristics related to game statistics, the team to which he belongs, information on nationality and many others are presented. This website allows fans to create an account and make their own drafts according to their own criteria and based on the information contained on the page. Player rankings based on the different categories to which they belong are also shown, such as High School players, NCCA players and International players among others. The way in which this website extracts data is not too clear. Each website has different prediction models and, although they are probably similar, they hide their way of obtaining data from the public to prevent competition from benefiting.

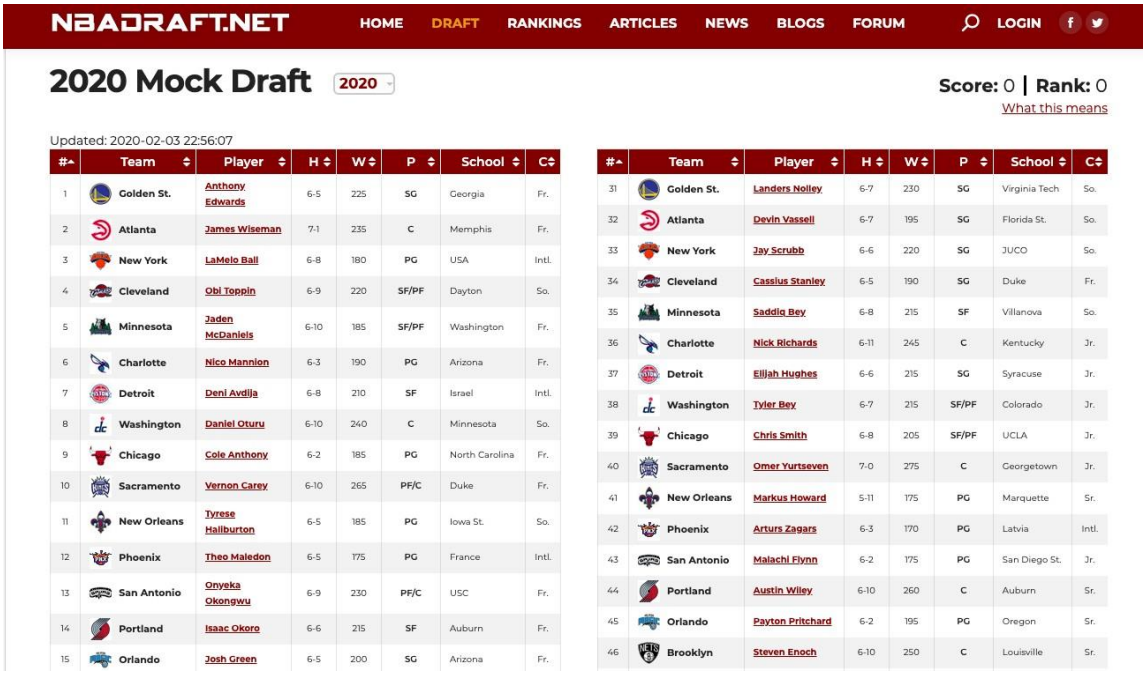

Figure 1. Nbadraft.net 2020 web site

- **TANKATHON.COM**

The company Tankathon [4] offer a web site (see Figure 2) even more complete: one can make predictions for any player but for the right amount of minutes played that you want. It is updated every 2 days, and it's clear and simple interface makes an easy experience for the user. It also makes predictions of the Draft Lottery, without taking into account the players yet. In addition, it allows you to perform a large number of actions, such as comparing players and consulting previous editions of Draft, with complete statistics both per game and per approximations for 36 minutes of play.

In addition, it is not limited only to the NBA, but rather goes further and offers these same services for the rest of the major American leagues, such as the NHL (ice hockey), the NFL (American football) and the MLB (baseball), where a player selection process is also carried out using Draft.

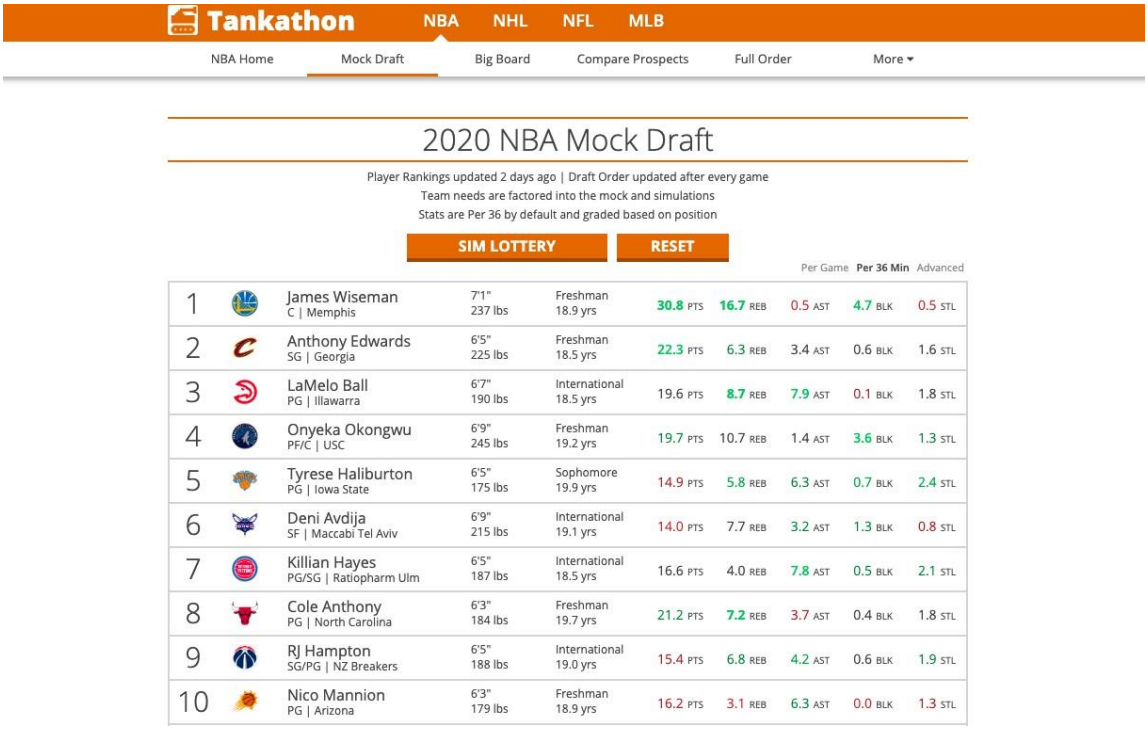

Figure 2. Tankathon.net 2020 web site

The ways to obtain the models on this page are more accessible to the public, such that in the 'Power Rankings' section where you can see the rankings generated by the Web model, explain how it is to obtain the ranking including complete explanations and visuals to which users have access. For example, Aaron Barzilai, who was a player on the varsity basketball team at MIT before earning his Ph.D. in Mechanical Engineering at Stanford University, wrote this article [5] where he concludes that:

> *"The right to draft with a given pick in the NBA draft is an important but often overlooked topic in basketball analytics. .... The position of a team in a future draft is not known until the end of the season before that draft, and many trades involve picks multiple seasons into the future. Additionally, the strength of a future draft class is particularly hard to quantify. Finally, when valuing a pick a team*

might include an estimation of their skill or their trading partner's skill in drafting players."

- **STATS.NBA.COM:**

This is the official stats website [6] of the NBA. It shows a great amount of different advanced stats (see Figure 3), allowing you to compare how two different players are playing together or how they are playing separated. This website will be used to compare the ideal stats of the player that we would want for any team with any player of the actual Draft Class that we are studying. This website uses a SAP data collection model created specifically for sports statistics, which is not accessible to the public. Is called SAP Global Sports One. SAP is at the forefront of transforming the sport and entertainment industries by helping athletes, performers, teams, leagues, and organizations run at their best.

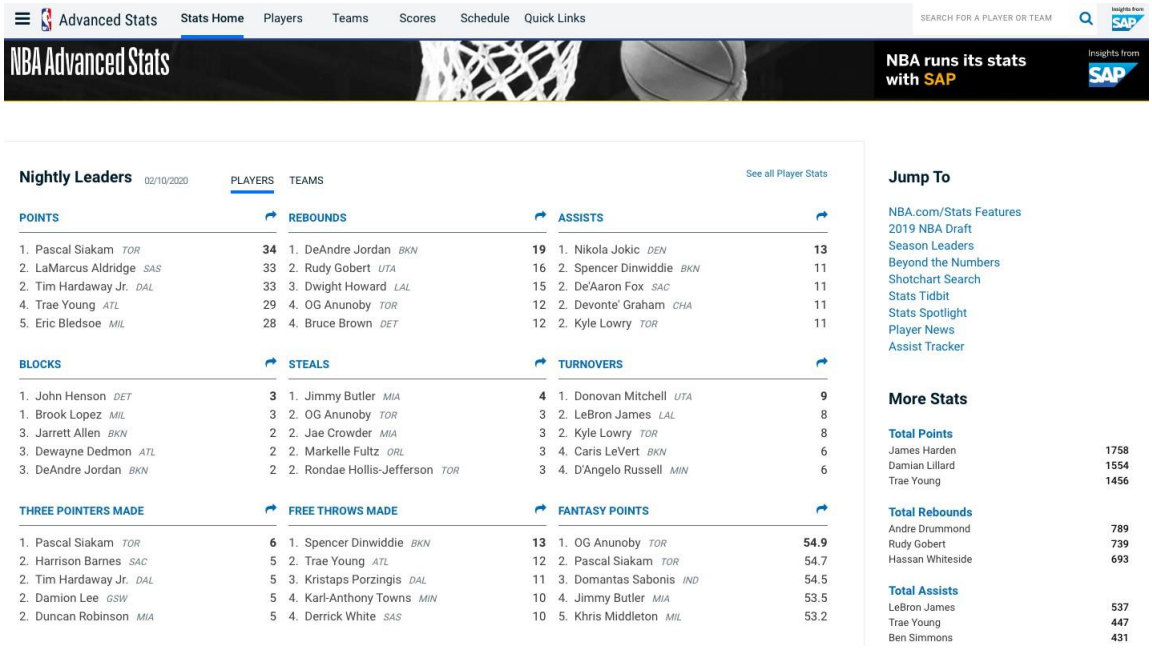

Figure 3. NBA official 2020 web site

- **BASKETBALL-REFERENCE.COM**

This web [7] stores every kind of stats for any player who has ever played in the NBA. BasketBall-Reference (Figure 4) will be used in a similar way compared to stats.nba.com, but adding the advantage of having advanced stats of retired players, which could make the study more advanced and complete. The data sets on this page are open to the public and offer the download of all the tables that make up the web page in different formats to make it as accessible as possible to users. It is the best-known website for NBA statistics

(it has variants dedicated to other sports and university sports). It offers complete statistics on Playoffs, Draft, candidate approaches to MVP, ROY (rookie of the year), etc.

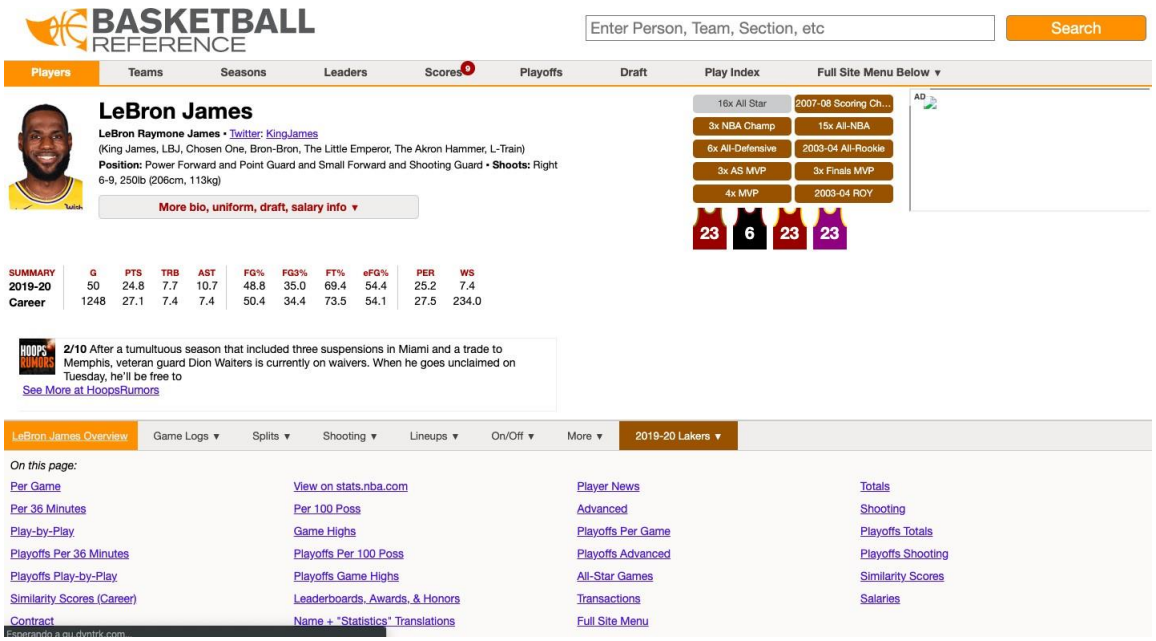

Figure 4. Basketball Reference 2020 web site

In addition to the information collected on these applications and websites, studies carried out at universities have also been analyzed, such as the aforementioned work by [5] and more recently the work by Adhiraj Watave at the University of California, Berkeley entitled Relative Value of Draft Position in the NBA [8].

2.2 Clustering techniques

Clustering algorithms consist of grouping the observations in sets trying to maximize the similarity between the elements of the groups. For this, the distance between the different variables of the observations is measured. Consequently, members of the same group have common characteristics. These common characteristics also serve to characterize the group. To obtain good results, it is necessary that the number of variables associated with the observations are sufficiently descriptive. On the one hand, the results of the grouping depend on the number and quality of the variables. The existence of dependent variables or with little variability, adds noise to the system. On the other hand, the object of grouping must also be taken into account when choosing variables. An objective may be to detect inherent groups, but we may also be interested in finding representatives of homogeneous groups or in detecting outliers. Hierarchical grouping methods based on

distances have been used in this work. These methods do not require prior information about the groups, called clusters, and are therefore called unsupervised methods. In this work we use the Chernoff faces [9-13], Singlelinkage [14] (with dendograms) and Kmeans methods [15, 16]. The idea of using these methods has been motivated by Professor Krijnen's work on similarity in gene observations [17]. Krijnen uses the database of Golub99 [18] to show how to find similarities by making groupings between observations from people with different types of leukemia.

The results of hierarchical grouping techniques are visualized using dendrograms. These objects are trees that serve to represent the similarity between the different observations in the study. Observations are leaves of the tree. If the leaves are close, it means that these observations are very similar. In contrast, two distant leaves represent very different observations.

The exploration of all possible trees is computationally intractable, so when the number of observations is large it cannot be carried out in polynomial time. Approximate algorithms or heuristics are used to solve the problem. There are two approaches based on spanning trees and that correspond to the ideas of Prim's algorithm on the one hand and Kruskal's algorithm on the other. The first case, called the divisive method, starts from a single cluster that is split into pieces until the appropriate number of clusters is reached. The second case, called the agglomerative method, begins by considering that each observation is a single cluster and at each stage, it joins one or more clusters until the appropriate number of groups is reached. This last method is the one we use in this work through the Single Linkage or Minimum Spanning Tree and KMeans algorithms. The fundamental difference between these two proposals is that the first only requires the set of observations as input while the second also requires knowing in advance the desired number of clusters, known as the number K. Thanks to its good results, Kmeans is the most popular algorithm. To solve the problem of choosing a suitable K number, in this work we have chosen to carry out a tracker code of a range of options to find the value that optimizes the results. All details about these methods and techniques are shown in the chapter 4 of the present work.

Chapter 3. Software tools

The different sections of this chapter show the details of the software tools that have been needed to carry out the project. These tools have been chosen due to their usability and suitability for the proposed objectives. They are all widely used tools today and their learning has contributed to my training.

3.1 R, RStudio, and libraries

R [19] is a programming language developed by Ross Ihaka and Robert Gentleman in 1993. R possesses an extensive catalog of statistical and graphical methods. It includes machine learning algorithms, linear regression, time series, and statistical inference to name a few. Most of the R libraries are written in R, but for heavy computational tasks, C, C++ and Fortran codes are preferred. R is not only entrusted by academic, but many large companies also use R programming language, including Uber, Google, Airbnb, Facebook and so on. Data analysis with R is done in a series of steps; programming, transforming, discovering, modeling and communicate the results.

RStudio [20] is an integrated development environment (IDE) for the R programming language, dedicated to statistical computing and graphics. It includes a console, syntax editor that supports code execution, as well as tools for plotting, debugging, and workspace management. RStudio is available for Windows, Mac and Linux or for browsers connected to RStudio Server or RStudio Server Pro (Debian / Ubuntu, RedHat / CentOS, and SUSE Linux) .3 RStudio has the mission of providing the statistical computing environment R. It allows analysis and development so that anyone can analyze the data with R.

There are many libraries and packages for R. In addition to the Base library, the most important libraries for carrying out this work are listed below according to their functionality and their utility in this work.

- **Tidyverse, dplyr, ggplot2 and ggdendro.** The tidyverse is a set of open source R packages provided by RStudio. The aim of this set is preprocessing data and it

provides functions to make data transformations as filtering, selecting, renaming, grouping, and so on. The most useful libraries in tidyverse are dplyr ang ggplot2. In this work, dplyr has been used to prepare the data and ggplot2 to make graphics and visualizations. The library ggdendro is useful to make dendrogram clustering and works in combination with ggplot2.

- **Alpk and TeachingDemos.** These libraries have been used for the classification and display of Chernoff's faces, although they provide many other functions.
- **Rmarkdown.** This library is very useful to make reports about the code generated. It helps also to relate the R code, comments, the input and the output and to remember easily the steps of the procedures.

3.2 Microsoft Power BI

It is a tool [21] dedicated to the visualization of large data sets used by many companies, mainly for the Business Intelligence area. It has a great capacity to manipulate large data sets efficiently, in such a way that the processes of cleaning data and relationships between tables are done quickly and easily. Through the creation of reports and interactive panels, different graphs related to statistics and relationships of the work developed will be displayed. Besides, the work will be made of the different possibilities that Power BI offers, such as connecting to the most popular statistics pages, in such a way that updated data can be obtained in real-time. The environment in which these reports and dashboards will be developed is Power BI Desktop, which requires a license provided by the company in which the author has carried out his practices and is currently working.

3.3 Github

At a higher level, GitHub [22] is a website and a cloud service that helps developers store and manage their code, as well as keep track and control of any changes to this code. To understand exactly what GitHub is, first you need to know the two principles that connect it: Control version and Git. A Control Version helps developers keep track and manage any changes to the software project code. As this project grows, the control version becomes essential. Git is an open source version specific control system created by Linus Torvalds in 2005. Specifically, Git is a distributed version control system, which means that the entire code base and its history are available on every developer's computer, allowing easy access to forks and merges. According to the survey among Stack Overflow

developers, over 87% of developers use Git. So, GitHub is a non-profit company that offers a cloud-based repository hosting service. Essentially, it makes it easier for individuals and teams to use Git as the collaborative and control version.

In this final year project, we will use Github to have a shared repository with the development and implementation of the NBA prediction model. The code will be implemented in R.

The github repository used for this project is located here:

https://github.com/houghton97/Similarity-Analysis-to-aid-decision-making-on-NBA-Draft

3.4 Others

For the correct manipulation of the new datasets extracted from the stats web pages, these text editors have been used to format and import them into Rstudio. Sublime Text and VS Code have been the two text editors chosen to carry out these tasks of organizing datasets and formatting tables. The large number of functionalities that VS Code offers thanks to its extensions, and the simplicity and compatibility of Sublime Text are the reasons that make them stand out from other editors.

Microsoft Excel has been used basically to organize the data obtained from the statistics websites in a similar way to text editors. The difference is that with this tool it has been possible to transform and manipulate the data obtained in spreadsheet format instead of in .csv format.

Chapter 4. Data sources, Methods and Techniques

This chapter shows the sources of information used, as well as their characteristics. From the data offered, the most interesting ones for this study have been chosen and added, and after the cleaning and data treatment operations they have been prepared as input for the clustering algorithms that are explained at the end of this chapter. Details of Chernoff's methods, hierarchical clustering and Kmeans algorithms are also provided in the chapter to complete the brief introduction given within the second chapter.

First, the problem we want to solve can be formulated as follows: Given a dataset S of players and given a specific player P_0 the problem is to find which are the most important features that a player needs to perform at a high level. Each player is characterizedby a list of features (variables structured into a data frame) in the way that Figure 5 shows as an example. In it, the 2013 draft dataset has 18 features of 62 players (observations).

Player	Draft pick	Height (No Shoes)	Wingspan	Standing reach	Vertical (Max)	Vertical (Max Reach)	Vertical (No Step)
Adonis Thomas	NA	76.75	85.00	99.0	40.5	139.5	34.5
Allen Crabbe	31	77.25	83.25	103.5	36.0	139.5	30.5
Andre Roberson	26	78.25	83.00	104.5	36.5	141.0	30.0
Archie Goodwin	29	75.75	81.50	102.0	36.0	138.0	30.0
B.J. Young	NA	74.25	80.25	99.0	NA	NA	NA
Ben McLemore	7	75.50	79.75	100.5	42.0	142.5	32.5
Brandon Davies	NA	80.50	85.50	108.5	31.5	140.0	26.0

Figure 5. 2013 draft dataset (7 first rows, 7 first variables)

The 18 features (or variables) are physical characteristics of the player. Players with a vector of features in small distance may have similar functions and may be potentially interesting for further research. In order to discover which observations (players) form a group there are several methods developed in the well-known cluster analysis. Most of these methods are based on a distance function and an algorithm to join data points to clusters.

The numerical variables of the observations are interpreted as points in the multidimensional space and the distance between them is calculated. A very common option is to use the Euclidean distance.

However, the results depend on a good selection of variables. Information sources tend to contain many redundant, aggregated, dependent and sometimes uninformative variables. There are also variables that will be essential to recognize the value of a player for a team and other variables whose information will not be of any interest. The problem of selecting the right variables is not a trivial problem. Scouts perform this task intuitively and based on their previous experience, so in a future automatic variable selection system without any prior information, they must use brute force algorithms or heuristics that relax the computation. In the case of this job, I have used my previous experience as a basketball player and as a coach for several years.

For this reason, it is important to carry out a preliminary data preparation phase. This chapter has been divided into three parts: the first is dedicated to the sources of information and the collection of data; the second part is dedicated to the preparation of the data and its treatment and finally, the third section is dedicated to the description of the classification methods used on the data.

4.1 Data collecting and integration

The datasets that have been used for this project are exposed in the Table 1. The table shows the data source, the web site where each dataset is available and a description.

Table 1. Data Sources downloaded for use in the project

Data Source	**WebSite**	**Description**
Data.World	https://data.world/achou/nba-draft-combine-measurements	Draft Combine Stats 2009-2016
Kaggle	https://www.kaggle.com/drgilermo/nba-players-stats	Shooting Stats NCAA + NBA of many players.
Basketball Reference (Draft)	https://www.basketball-reference.com/draft/	Stats of every season in the NBA from any Draft Class
Github	https://github.com/sshleifer/nbaDraft/	Complete NCAA stats dataset

The variables of interest corresponding to each table are shown in Table 2. As there are many, the most notable of the datasets are chosen. Furthermore, the variable that

refers to the name of the players is not included, as the extreme importance of this variable is taken for granted.

Table 2. Variables of interest per dataset

Datasets	**Description**
Draft_Combine_All_Years	Draft pick, Height, Weights, Wingspan, Vertical, Standing Reach.
Players_Shooting	NCAA_ppp, NBA_ppp, NBA_3papg, NCAA_3papg, NBA_g_played
Seasons_Stats	Stats of every season in the NBA from any player.
Lot_of_players_info	2P%, PPG, FG%, 3Pt, GP

Next, we explain the composition of the different datasets and the possible uses that will be given during the project.

- **Draft Combine Stats - *Draft_Combine_All_Years***

This Dataset provides a wealth of information and statistics about the physical tests that players perform each year before presenting to the Draft. The Dataset has information on each class of the Draft from 2009 to 2016. Among the variables stand out height, span, vertical jump, weight or position of the Draft of each player. It will allow us to extract information about the physical characteristics of players in their preparation stage for the NBA Draft and thus be able to compare future Draft classes with characteristics of model players in the early stages of their professional career. Figure 6 shows a fragment of this dataset to illustrate this information.

Player	Year	Draft pick	Height (No Shoes)	Height (With Shoes)	Wingspan	Standing reach	Vertical (Max)	Vertical (Max Reach)	Vertical (No Step)
Blake Griffin	2009	1	80.50	82.00	83.25	105.00	35.5	140.50	32.0
Terrence Williams	2009	11	77.00	78.25	81.00	103.50	37.0	140.50	30.5
Gerald Henderson	2009	12	76.00	77.00	82.25	102.50	35.0	137.50	31.5
Tyler Hansbrough	2009	13	80.25	81.50	83.50	106.00	34.0	140.00	27.5
Earl Clark	2009	14	80.50	82.25	86.50	109.50	33.0	142.50	28.5
Austin Daye	2009	15	81.75	82.75	86.75	110.00	28.0	138.00	25.0
James Johnson	2009	16	79.00	79.75	84.75	105.50	35.0	140.50	30.5
Jrue Holiday	2009	17	75.25	76.25	79.00	100.50	34.0	134.50	28.5
Ty Lawson	2009	18	71.25	72.50	72.75	94.50	36.5	131.00	29.0
Jeff Teague	2009	19	72.25	73.50	79.50	98.50	36.5	135.00	30.5

Figure 6: Fragment of Draft_Combine_All_Years dataset

- **Shooting Stats NCAA NBA - *Players_Shooting***

It is made up of information about shooting percentages for numerous players from 1950 to the present. It is made up of shooting variables grouped in such a way that shooting statistics can be observed in both the university league and the NBA. In addition, it includes an organization of statistics both by game and by quarter, as well as variables that allow distinguishing converted shots from attempted shots. This dataset has passed an important data cleaning phase, given that the database included players who developed their careers in an era prior to the existence of the three-point shot. It will allow us to observe the influence of the change of league in the shot of the players, and thus see which players are more affected by the change and which less. Figure 7 shows a fragment of this dataset to illustrate this information.

NBA_fta_p_g	NBA_g_played	NBA_ppg	NCAA__3ptapg	NCAA__3ptpct	NCAA__3ptpg
1.4	151	5.6	1.9	0.369	0.7
2.0	557	7.5	0.7	0.356	0.2
2.7	387	8.4	0.1	0.400	0.0
0.9	113	6.1	0.1	0.250	0.0
0.5	55	3.3	0.1	0.250	0.0
1.2	135	3.4	0.1	0.250	0.0
2.4	552	9.2	0.1	0.000	0.0
1.6	518	3.7	0.0	1.000	0.0
1.8	375	5.3	0.0	0.333	0.0
1.0	74	5.1	0.0	0.000	0.0
0.8	490	2.2	0.0	0.000	0.0
2.0	541	5.5	0.0	0.000	0.0

Figure 7: Fragment of Players_Shooting dataset

- **NBA Season Stats – Seasons_Stats**

The original dataset contains complete statistics for a large group of players for each year. Summarize the statistics that each player performs in the year you want. In this way, if we filter by the first year of each player, the statistics of the first season of these are obtained. This Dataset will be very useful when evaluating the performance of players belonging to a common Draft class in their first year as professionals. Figure 8 shows the appearance of this dataset.

Year	Player	Pos	Age	Tm	G	MP	PER	TS%	FTr	FG	FGA	FG%	2P
1981	Kareem Abdul-Jabbar*	C	33	LAL	80	2976	25.5	0.616	0.379	836	1457	0.574	836
1981	Tom Abernethy	SF	26	TOT	39	298	8.0	0.459	0.373	25	59	0.424	25
1981	Tom Abernethy	SF	26	GSW	10	39	3.2	0.463	1.000	1	3	0.333	1
1981	Tom Abernethy	SF	26	IND	29	259	8.7	0.458	0.339	24	56	0.429	24
1981	Alvan Adams	C	26	PHO	75	2054	20.3	0.567	0.298	458	870	0.526	458
1981	Darrell Allums	PF	22	DAL	22	276	5.3	0.385	0.328	23	67	0.343	23
1981	Tiny Archibald*	PG	32	BOS	80	2820	14.3	0.582	0.547	382	766	0.499	382
1981	Dennis Awtrey	C	32	SEA	47	607	6.7	0.501	0.215	44	93	0.473	44
1981	James Bailey	PF	23	SEA	82	2539	14.5	0.546	0.406	444	889	0.499	443
1981	Greg Ballard	SF	26	WSB	82	2610	16.7	0.500	0.165	549	1186	0.463	542

Figure 8: Fragment of Seasons_Stats dataset

- **Complete NCAA Stats - *Lot_of_players_info***

With both physical and game statistics and personal characteristics (age, university, Draft election, position), this dataset gives a lot of information to the project, and allows combining characteristics of the Draft_Combine_All_Years datasets (with characteristics physical as wingspan, weight ...) with historic NCAA stats. In addition, it is made up of a large number of players and former players (with debuts from 1950 to 2017). In Figure 9 part of this dataset is shown.

Name	Team	pick	age	GP	Min	Pts	FG	FGA	FG%	2Pt	2PtA	2P%	3Pt
Dajuan Wagner	Memphis	6	19	36	31.8	25.0	8.7	21.2	41.0	6.5	14.4	45.3	2.2
Carmelo Anthony	Syracuse	3	19	35	36.4	23.2	8.3	18.2	45.3	6.6	13.3	49.6	1.7
Luol Deng	Duke	7	19	37	31.1	18.8	7.1	14.9	47.5	5.7	11.1	51.4	1.3
Kris Humphries	Minnesota	14	19	29	34.1	25.0	8.8	19.8	44.4	8.1	17.8	45.5	0.7
Trevor Ariza	UCLA	43	19	25	31.6	14.8	5.5	12.9	42.6	4.6	9.0	50.9	0.9
Kevin Durant	Texas	2	19	35	35.9	27.0	9.2	19.4	47.3	6.7	13.3	50.5	2.5
Thaddeus Young	Georgia Tech	20	19	31	29.6	18.4	7.3	15.3	47.8	5.7	11.5	49.8	1.6
Michael Beasley	Kansas State	2	19	33	31.5	30.8	10.9	20.5	53.2	9.6	17.1	56.2	1.3
Anthony Randolph	LSU	13	19	31	32.8	18.7	7.1	15.2	46.4	7.0	14.5	48.3	0.1
Jrue Holiday	UCLA	8	19	35	27.1	12.7	4.8	10.6	45.0	3.6	6.9	52.8	1.2
Xavier Henry	Kansas	15	19	34	27.7	19.1	6.4	13.9	45.9	3.6	7.3	49.2	2.8

Figure 9: Fragment of Lot_of_Players_Info dataset

These are the datasets imported from the data storage web pages, but also, datasets created from combinations and leaks of those extracted directly from the internet have been used. After the cleaning and data preparation phase, these datasets have been created that facilitate the implementation and development of the visualizations of this project. Both for Power BI dashboards and Rstudio renderings.

4.2 Data Cleaning and pre-processing

The data cleaning phase has had a considerable duration, due to the variety and quantity of datasets found related to the matter, and also because what was found by the network, despite having similar characteristics, does not use specific datasets that have the same objectives. and the same methodology as required. Due to this, each of the datasets found has had to be manipulated and modified, as well as creating new ones and also correcting the errors found in those of third parties.

- ***Players_Shooting:*** In order to manipulate the data from Players_shooting, the data disposition was corrected first, since several cells were filled in erroneously due to specific player situations such as: having played in two different universities, not having played in NCAA Before the NBA, having played in times where the three-point shot did not exist ... Once corrected errors, we proceeded to create new datasets made up of subsets of data obtained from the original dataset. These subsets of data were distinguished according to the position of the players to be studied. Stages in data cleaning can be differentiated:

    ○ Deletion of columns that do not bring real interest, such as birthday or institute name.

    ○ Eliminate columns with a large number of NAs, as well as replacing existing NAs with valid numerical values.

    ○ Filtered by players who have had consistent NBA careers (50+ games)

Draft_Combine_All_Years: We have proceeded to create datasets made up of data sets from each year of the Draft class. In this way you can compare entire classes of the Draft. In addition, since they only have physical characteristics, several statistics have had to be corrected for players who will perform physical tests other than the most common ones (generally, before the Draft, complete physical tests are carried out on the players, including vertical jump, long jump and agility among others). It is usual that with international players only characteristics such as height, wingspan and weight are considered. In addition, this table had incomplete characteristics such as the size of the hands, which are included in other datasets.

- ***NBA_Stats_Draft***: In this table, the main errors that have been corrected have been caused by the way of exposing the data on their original websites. The layout of the columns and rows was incorrect, in addition to adding all the categories as if they were cell data instead of as column titles. Player name fixes and removal of advanced non-job stats, as well as addition of new tables with subsets of data by years following the same process as with all other original datasets.

- **Seasons_Stats**: The content of the dataset did not have notable formatting errors, but as it contained information referring to very old players (since 1950), the long distance shooting statistics and various percentages were included as NAs. The corrections have been based on eliminating variables that included very advanced and far-fetched statistics (% rebounds per Game), as well as filtering the table in such a way that only statistics from 1981 are shown.

- ***Lot_of_players_info***: Despite being a good quality dataset (good organization, complete data for each player), changes have had to be made in several old players whose exact physical characteristics were not registered, and data had to be added to categories such as body fat and dimension of hands.

In addition to the specific changes made to the different datasets that make up the study, general guidelines have been taken into account to make the interpretation and handling of the data as simple as possible. General guidelines applied to all datasets such as:

1. It has been established as the common name of the variable that refers to the players' first name as' Player '. This variable will be composed of a name and a surname.
2. As common names for the variables that refer to game statistics, the format of the most popular sports statistics websites has been followed, such as: Basketball Reference or NBA stats. This format consists of distinguishing the statistics by initial or by value of the basket in this case, as well as in terms of percentages, add a '%' at the end of the initials.
3. The positions of the players have been limited to the 5 main basketball players, with their corresponding acronym (C, G, F, PF, SG). In this way, secondary player positions are not taken into account.

4. All the individual values marked 'NA' in the tables have been modified. They have been replaced by values such as 0, or by values that indicate the NA reason (in columns such as 'college', in case of having reached the NBA as an international or from high school, a common value has been created for the variable).
5. Added players who might be missing due to different data sources. In some tables some, players referring to high positions in the Draft simply did not appear, and have been added with their respective characteristics.
6. The final datasets have been renamed in such a way that their name ends with the ending ‘final’. This gives us security when it comes to using the correct datasets in PowerBI.

In addition to all the data cleansing carried out on the datasets, the process of preparing data in the Power BI environment must be considered. Different actions have been carried out in this tool to carry out the work in the most efficient and correct way possible:

1. The datasets, once manipulated in RStudio, when converting them to CSV and importing them into Power BI, several columns lose formatting and numbering. For example: physical statistics such as weight or wingspan lose their decimal places and have to be recalculated again.
2. To obtain more complete graphs, formulas have been calculated with the columns from the datasets. Formulas such as the season or race averages for each statistic.
3. Connectors have been created to the reference pages of this job to automate the updating of data in Power BI charts.

4.3 Clustering techniques: Analysis and Visualization

In many real-life problems, we need to analyze data in search of solutions to problems that are not clear-cut. This frequently occurs when the hypotheses are not clear or when, as in our case, we do not know exactly what information is necessary to solve the problem. In these cases, it is usually common to use data visualization. Using graphical representations, the analyst can formulate hypotheses and conclude. In our case, the

visualization before the analysis has helped us to discover the variables of interest and reduce their number. Visualization helps as well to understand the analytics. The analysis techniques also inform us of the importance of the variables in the creation of the groups and this information can be useful in future applications.

This section is structured as follows. First, we introduce the Single Linkage method and KMeans from their theoretical point of view and its application on a generic Draft class. Then we explain how Chernoff's faces works for the clustering and how interactive charts and dashboards with Power BI work. The application of this techniques on specific Draft class and its results are shown in Chapter 5.

- **Single linkage**

The Single Linkage cluster analysis is important because it is intuitively appealing and often applied in very similar works [16]. By single linkage several clusters of players can be discovered without specifying the number of clusters on beforehand. This is an advantage because at first instance we do not have information about the composition of the dataset. That is why this is called unsupervised method. The analysis produces a tree, which represents similar players as close leaves and dissimilar ones on different edges.

The Single Linkage is based on clusters. A cluster C is a set of observations C = {p1,…pN}. In our case, each observation corresponds with a player. In Single Linkage cluster analysis the distance between clusters C1 and C2 is defined by the distance formula:

$$dist(C1, C2) = \min_{i,j}\{dist(pi, pj): pi \in C1, pj \in C2\} \qquad \text{(Eq1)}$$

That means, the smallest distance over all pairs of points of the two clusters. The method is also known as nearest neighbors because the distance between two clusters is the same as that of the nearest neighbors. The algorithm starts with creating as many clusters as players (observations). Next, the distance between each pair of clusters is calculated (by Eq1) and nearest two are merged into one cluster. The process continuous until we get desirable number of clusters or even until all points belong to one cluster. Figure 10 shows the flow chart of the algorithm. In this algorithm, the process 'merge' is indicated. Every loop can make one or more merges in case of a tie. In Figure 11 an

example is shown. The code with *hclust*, method *single* and distance *Euclidean* is on top of the figure. Below the code, the calculations and activity of the algorithm are shown in the same figure. As a result, a dendrogram is drawn. Some examples of these dendrograms come next chapter.

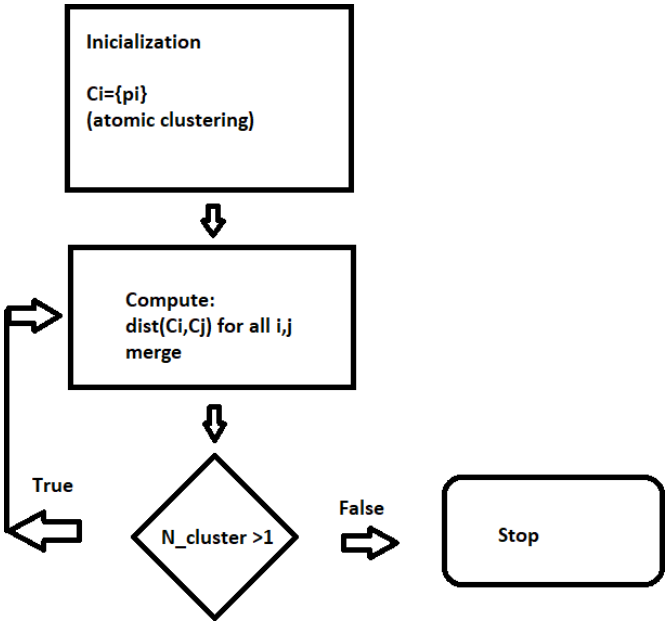

Figure 10: Flow chart of the Single Linkage algorithm (own elaboration)

```
61  ###Single Linkage Custering with dendrograms, hclustering applyed to the
62  # group of players, with a different selection of variables, here 3:10
63  ######
64
65  playersHC<-hclust(dist(active_from2015[,4:10],
66                         method="euclidian"), method="single")
67  plot(players)
```

playersHC	list [7] (S3: hclust)	List of length 7
merge	integer [17 x 2]	-12 -13 -8 -3 -1 1 -16 -14 -18 3 4 5 ...
height	double [17]	0.540 0.601 0.633 0.638 0.655 0.776 ...
order	integer [18]	5 15 17 13 14 12 ...
labels	NULL	Pairlist of length 0
method	character [1]	'single'
call	language	hclust(d = dist(active_from2015[, 4:10], method = "eu..
[[1]]	symbol	`hclust`
d	language	dist(active_from2015[, 4:10], method = "euclidian")
method	character [1]	'single'
dist.method	character [1]	'euclidean'

Figure 11: Example of execution of the Single Linkage algorithm.

- **KMeans**

In clustering analysis, K-means is a well-known method. The goal is to solve the same problem than in Single Linkage but that here it is re-formulated taking into account the number of clusters that plays a special role because it will be part of the input. Given the data observations (or players, in this work)

$$Players = \{p1, \cdots, pn\}$$

The method seeks to minimize the function:

$$\sum_{i=1}^{K} \sum_{j \in C_i}^{n_i} dist^2(p_j, q_i) \qquad (Eq.2)$$

Over all possible observations $q_1, \cdots, q_K$.

The method (see [22]) is based on building clusters trying to minimize the variance within the observations of the same group. That is, minimizing the within-cluster sum of squares over K clusters (Eq.2) is what the algorithm aims. However, this goal is an NP problem at it is said before and therefore, the algorithm described in [22] is just a heuristic. According to [23], the algorithm has strong consistency for clustering. This is the reason why it is very popular despite not guaranteeing the optimization goal or its convergence. The steps of the algorithm are as follows.

- Step 1: Initialization. The algorithm begins by partitioning the observation set into K random initial clusters. Also, this initialization can be done by using some heuristic device which improves the speed of the execution and quality of results especially when some information is known in advance.
- Step 2: The algorithm computes the cluster means.
- Step 3: The algorithm constructs a new partition by associating each point with the closest cluster mean. The latter yields new clusters.
- Step 4: Clustering comparison. If there are no changes, the process stops. If one or more observations change clusters, the process jumps to Step 2.

Therefore, the algorithm works by making a new partition every step by associating each observation with the closest cluster mean. These steps are repeated until convergence. The number of clusters has no changes. Only the observations changes (or not) the cluster where they belong. The stop occurs when the observations no longer change clusters, the process converges.

The algorithm works better when the variables are normalized, therefore, in the first place, we have used the 'scale' function of R that allows this process to be carried out easily on the variables chosen for the study. Next, we have made a loop to calculate the values of the objective function (Eq.2 wich is equivalent to $tot.withinss in this model), and thus determine the number K of clusters that we want to calculate. We have also compared these values with those of the intra-group variance, given by the variable $betweens. However, it is important to consider the variance within each group individually, this is $withinss, which allows determining the homogeneity of each group. To illustrate this, Figure 12 shows the output of an execution of Kmeans R function over the same dataset example of players. In the top of the figure it is the code executed. The line 113 of the code is a combination of the following actions: selection of the variables (sequence 3:14 of the datosStudy dataset), normalization of the data with scale and structure the result as data.frame mode. Line 115 instances the Kmeans algorithm for k=5 clusters (parameter ‘centers’). Line 115 produce the output below. In it, the structure of the new object ‘datos.kmeans’ is shown. There are a total of 9 fields including $centers (final centroid of each cluster), $withinss (list of the 5 within-cluster sum of squares), $tot.withinss (the sum of the latter), $betweenss (intra-cluster sum of squares) , $size (list of the 5 sizes of the clusters) and $iter (number of iterations of the algorithm). This information is essential to analyze the quality of the results. On the right it is the list of player’s names and the cluster they below.

```
113 dataStudy.scale<-as.data.frame(scale(datosStudy[,3:14]))
114 ## First try with 5 centers
115 datos.kmeans<-kmeans(dataStudy.scale, centers = 5) #k=centers
116 ## checking the results
117 View(datos.kmeans)
```

datos.kmeans	list [9] (S3: kmeans)	List of length 9
cluster	integer [37]	3 3 5 3 3 1 ...
centers	double [5 x 13]	2.02e+03 2.01e+03 2.01e+03 2.(
totss	double [1]	973.2432
withinss	double [5]	0.0 35.5 241.7 35.0 33.5
tot.withinss	double [1]	345.723
betweenss	double [1]	627.5202
size	integer [5]	1 6 16 6 8
iter	integer [1]	2
ifault	integer [1]	0

datos.kmeans	List of length 9
cluster	3 3 5 3 3 1 ...
Arron Afflalo	3
Tony Allen	3
Kyle Anderson	5
Trevor Ariza	3
Luke Babbitt	3
Lonzo Ball	1
Tarik Black	4
Bobby Brown	3
DeMarre Carroll	3

Figure 12. Data computed by the Kmeans algorithm

To make a good choice of the number k, some authors recommend analyzing the evolution of the variables $betweens and $tot.withinss. The first use to increase when the number of clusters grows but from a certain value of K, the growth is negligible or just decrease.

On the contrary, the variable tot.withinss decreases normally as the number of clusters increases but it also can increase at a certain point or slow its decrease. The appropriate K value is one where this variable starts to grow again or decreases very slowly. Figure 13 is an example of NBA 2009 dataset. The figure shows the evolution of these variables in the execution of 11 instances of Kmeans algorithm. It can be seen how 7 can be an adequate number of clusters according to these criteria.

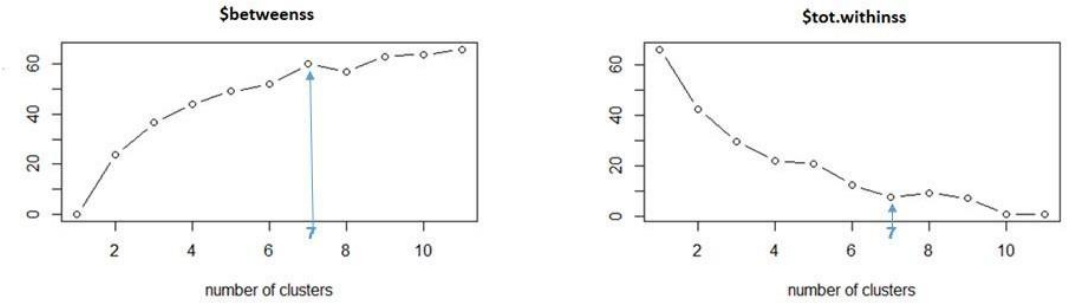

Figure 13: Selection of an adequate K number of clusters for the NBA 2009 dataset

Once the clustering is done, it is easy to plot the results for a better visualization in reports and helping with decision making. Next chapter contains some of these plots over the specific datasets in the Draft study.

- **Visualization with Chernoff faces**

Chernoff faces is a method invented by Herman Chernoff [9] for easy classification of multivariate data. The process consists of displaying multivariate data in the shape of a human face. Each individual item within a face (eyes, ears, mouth, nose, chin and even a hat of other elements if necessary) represents a value by different shapes and sizes. This idea comes because face recognition is a natural way that humans can easily apply and notice small changes without difficulty. The main use of face graphs was to enhance "the user's ability to detect and comprehend important phenomena" and to serve "as a mnemonic device for remembering major conclusions" [9]. To draw the output faces, the system handles each variable differently and can be fitted to facilitate the final classification decision by choosing the features that are going to be mapped. In [11], for example, the authors demonstrate that eye size and eyebrow-slant have been found to carry significant weight.

Recent applications of Chernoff faces include classification of drinking water samples [10], graphical representation of multivariate data [11], experimental analysis [12] and last year (2019) in cartography classification [13].

TeachingDemos and AplPack libraries provide both the function 'faces()' that makes the calculations of the distance matrix and the visualization of the resultant faces. The AplPack function is more complete because also provide the characterization matrix per each observation. Figure 14 shows an example of code where a set of players from 2015 is selected, and variables 3:14 are taken as input. The result output of distance is shown below the code. For example, eye right (eyer) and eye left (eyel) of the player Kyle Anderson have a list of numbers symmetric because they come from the same feature of the player. Every item within the face is related in this way with a feature to draw it later, as we show in the next chapter.

```
47 ## Chernoff faces :#
48 ####
49 # in this example, with all players active from 2015 or later,
50 # and still active in 2018, the Chernoff faces is shown
51 library(TeachingDemos)
52 active_from2015<-filter(datosStudy, active_from>=2015)
53 basketfacesTD<-TeachingDemos::faces(active_from2015[,3:14])
54
55 library(aplpack)
56 active_from2015<-filter(datosStudy, active_from>=2015)
57 basketfacesAP<-faces(active_from2015[,3:14], labels = active_from2015$name, face.type =5)
58
```

basketfaces	list [3] (S3: faces)	List of length 3
faces	list [18]	List of length 18
Kyle Anderson	list [11]	List of length 11
eyer	double [7 x 2]	14.40 17.49 22.34 25.43 22.34 17.49 0.00 4.66 4.66 0.00 -4.66 -4.66 ...
eyel	double [7 x 2]	-14.40 -17.49 -22.34 -25.43 -22.34 -17.49 0.00 4.66 4.66 0.00 -4.66 -4 ...
irisr	double [5 x 2]	17.93 19.70 21.90 19.70 17.93 0.00 2.33 0.00 -2.91 0.00 ...
irisl	double [5 x 2]	-17.93 -19.70 -21.90 -19.70 -17.93 0.00 2.33 0.00 -2.91 0.00 ...
lipso	double [9 x 2]	0.00 6.87 14.49 7.52 0.00 -7.52 -53.53 -54.47 -50.62 -59.63 -60.57 -59 ...

Figure 14: Example of Chernoff faces execution and the characterization matrix.

- **Visualization with Power BI**

The use of the Microsoft Power BI tool will offer us an overview of the different relationships between both physical and game statistics of the players, which will allow us to obtain conclusions about the behavior of these statistics. The main objective of creating interactive visualizations is to be able to locate trends, and thus have a graphic representation of the possible predictions resulting from this work. The first actions to be carried out have to do with the arrangement and organization of the data. Making decisions about what data to study and relating to obtain the best-performing graphs is important. An example: it makes sense to relate two physical characteristics such as weight and wingspan, but it does not make so much sense to relate age to a purely physical characteristic. The main relationships that it has been decided to study are:

● Physical characteristics with position: players who can stand out in terms of physical characteristics over the rest of players of the same position are more likely to excel.

● Physical characteristics with shooting: In the historical records, a clear trend is noted that indicates that larger players have worse shooting percentages.

● Age with shooting percentages and number of shots: The evolution of the shot over time of the players will be studied.

● Shooting in NBA with shooting in NCAA: all statistics related to shooting will be studied to obtain the trends of change between leagues (professional and

university)

● Draft classes with physical characteristics: This relationship allows us to know the general election trends over the years exclusively by studying physical characteristics.

Chapter 5. Results

This chapter shows the results of the application of the clustering techniques developed in this work, as well as the visualizations of the most remarkable results implemented in RStudio and Power BI. To obtain the most representative results possible, a real case that occurred in the NBA about a wrong choice in the order of the Draft players is exposed. In the 2009 draft, the choices of each team had special relevance for their franchises, given that they are considered one of the best generations of the Draft in recent years. Stars like James Harden, Stephen Curry or Blake Griffin, were selected in this class of the Draft. In this Draft class, bad predictions were given that marked the future of the teams, by letting players leave who turned out to be better than the choice made. Stephen Curry, Demar Derozan and Jrue Holiday were drafty underrated players by teams, but they would end up being stars.

During the study, the behavior of the characteristics of these players, especially from Stephen Curry, has been of vital importance.

The shooting of the players is an essential factor for the game of basketball, every attack of a player revolves largely around his shot, and the study carried out has been based on statistics and general shooting percentages. In this case, a study has been carried out around the guards and guards belonging to the first round of said Draft. To access the data more easily and efficiently, the results have been obtained by manipulating a dataset merged between Seasons_Stats and Shooting_Stats. To complete the work, the results of the application of the same methods are exposed to the players belonging to the most recent Draft classes.

5.1 Clustering analytics results

This section shows the results of the clustering methods applied. We will present results from both Single Linkage, as well as K-Means on selected sets of variables according to the previous experience. The section includes also a critical comparison and explanation of the charts.

- **Single Linkage results**

In the first place, the variables referring to percentages of field goals and free throws have been selected. Besides, a distinction is made of the same type of variables in 3 different

stages of the players: NBA career, First year in NBA, NCAA career. By applying hierarchical clustering with the single method, specific dendrograms are created, referring to the different percentages of each player, which indicate similarities between them. The results are shown in Figures 15 and 16.

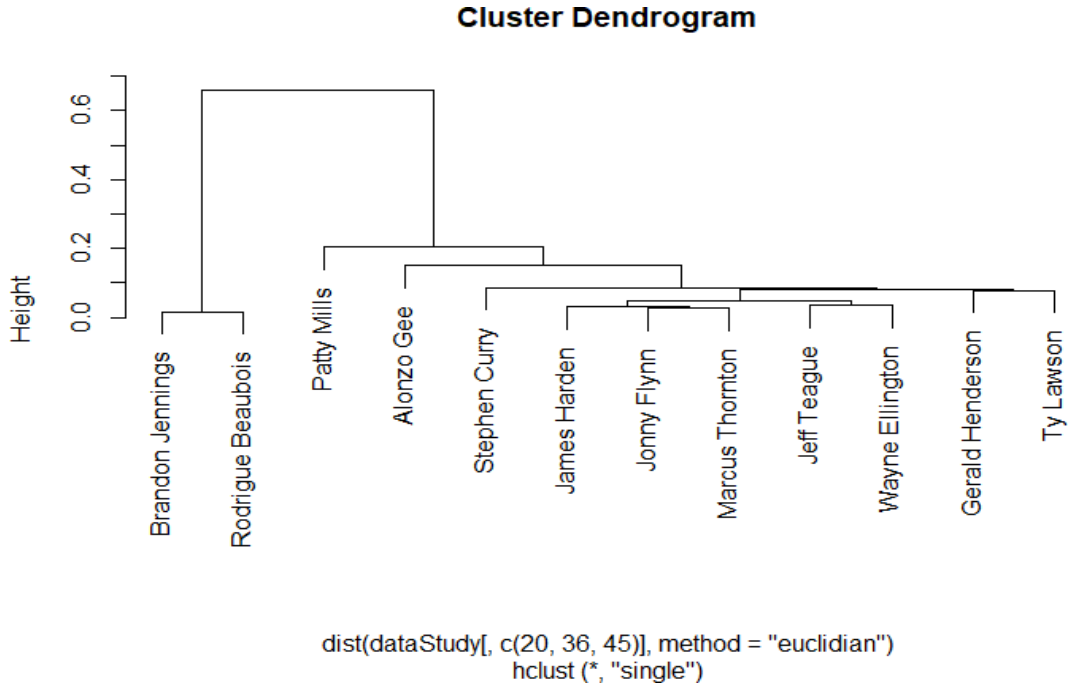

Figure 15: Dendrogram FT% on 3 features: Free Throw percentage in first season in NBA, NBA career, NCAA career.

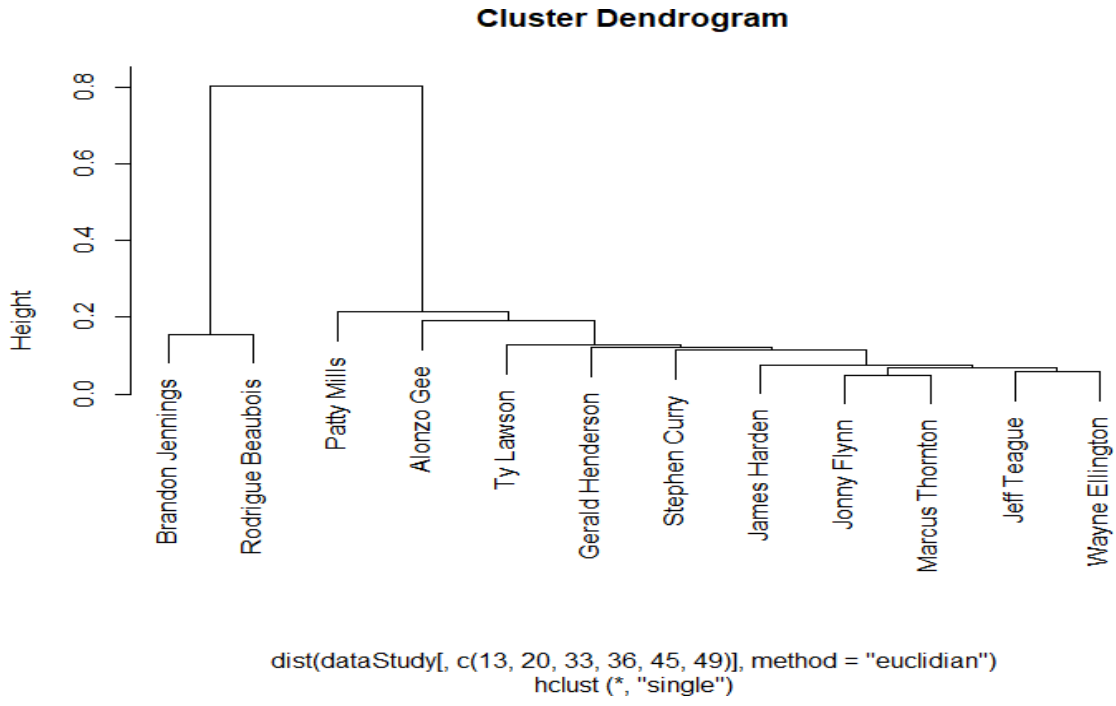

Figure 16: Dendrogram FG% on 6 features: Field goals attempted and Field goals made in first season of NBA, NBA career and NCAA career.

We can see that the evolution of field kicks and free kicks is similar in most players (it does not mean that they all shoot equally well, but rather that their evolution in the shooting has followed similar patterns). Players Brandon Jennings and Roudrigue Beaubois are outliers, having not played in the NCAA. What can be highlighted from the above charts is the consistent shooting similarity between Jonny Flynn and Marcus Thorton (both would end up with mediocre runs).

In the development of the work, the decision was made to compare the results of shooting statistics in both competitions, also, adding the latest players with data collected and available on the Internet. In such a way, when putting the classes of 2009 and 2017 together, interesting results are obtained as shown in Figure 17.

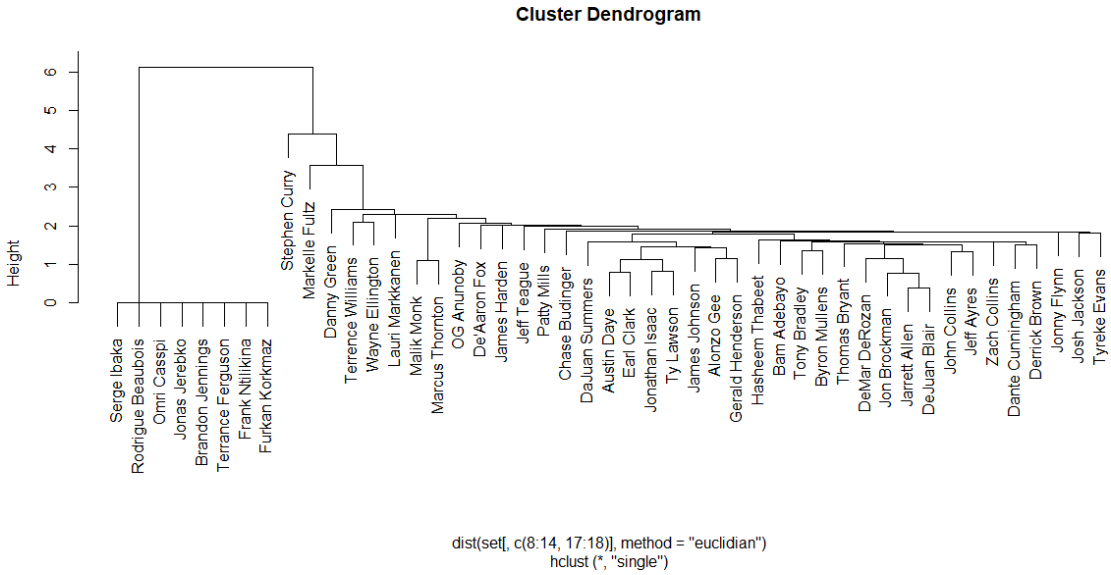

Figure 17: NCAA Overall Shooting 2009 & 2018

It can be seen how there is a common point regarding most of players, but Stephen Curry and Markell Fultz follow a very unusual pattern concerning the rest. Markelle Fultz was the pick No. 1 in the 2017 Draft due to his great shooting ability (as well as excellent physical qualities), a factor that shows that teams, do not let that players as special as Stephen Curry, have a hard time in their Draft night. This result already tells us that it is essential to highlight the importance of shooting over many other qualities in basketball.

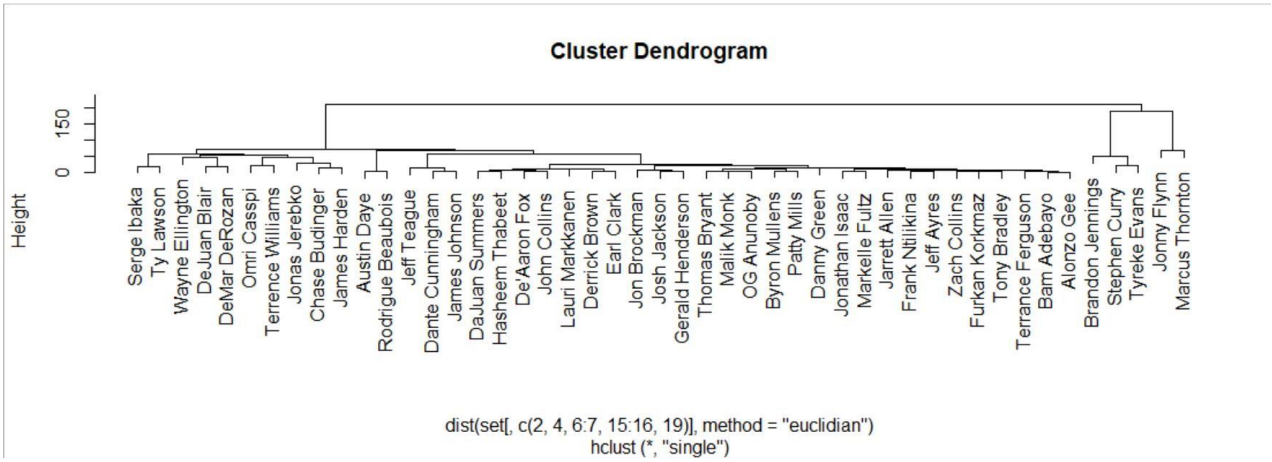

Figure 18: NBA Overall Shooting 2009 & 2018

In Figure 18 we can see how the players during their NBA Career are located similarly in the dendrogram, except for 5 players: Jennings, Curry, Evans, Flynn and Thorton. The explanation is simple: in this representation, the number of shots tried is considered, and of these players, all except Stephen Curry, have ended up having mediocre runs in long terms. These 4 players managed to be established players during their early years. But the low percentages hurt them over time, and if they don't improve much in terms of efficiency and scoring, the team loses confidence in them and they try to replace him with other younger players. This tells us that having a good shooting percentage in the NCAA, together with a large shooting volume, makes the player more likely to have good results in the NBA. Therefore, the importance of the percentage of successes before in the annotation per game is concluded. High registers of both cases are essential.

- **K Means**

This section shows the results of the datasets 2009 and merge 2009+2018 analyzing the results obtained by kmeans and comparing them with the dendrograms applied to the same data. First, we must do normalization of the data since kmeans is very sensitive to outliers. Figure 19 shows the 2009 dataset after scaling. The scaling is done by dividing the (centered) columns by their standard deviations.

	FG%	FT%	NBA_fg%	NBA_ft%	NCAA_ft	NCAA_FG
Alonzo Gee	0.7964587	-1.6719588	0.04371681	-1.6404842	-0.03939307	0.3480265
Brandon Jennings	-1.2115827	0.3890722	-1.81256614	-0.6247544	-2.09978665	-2.1205328
Gerald Henderson	-1.5012041	-0.3680413	0.28584067	-0.6973065	0.20185554	0.3572315
James Harden	-0.5937238	0.2944330	0.40690260	0.7779200	0.36160125	0.6335902
Jeff Teague	-0.7288804	0.5993815	0.56831851	0.5360796	0.54742788	0.4155554
Jonny Flynn	-0.3234105	0.4837114	-1.32831842	-0.3103619	0.44962439	0.3867726
Marcus Thornton	0.3330646	0.3575258	-0.31946899	0.1975030	0.41376311	0.3655340
Patty Mills	-0.3234105	-2.1977320	-0.03699115	0.6328158	0.52786718	0.1622378
Rodrigue Beaubois	1.6267067	0.2944330	0.24548669	-0.2861778	-2.09978665	-2.1205328
Stephen Curry	0.5454536	1.1041238	1.77893783	1.9629381	0.75607533	0.4428018
Ty Lawson	1.5687824	-0.2418557	1.09292022	-1.2535395	0.44310416	0.7070335
Wayne Ellington	-0.1882539	0.9569073	-0.92477865	0.7053679	0.53764753	0.4222825

Figure 19: The dataset 2009 after scaling

Figure 19 shows how the data is organized after applying the scale to the different stats selected. We can see that the values surround the number 0 depending on how good or bad they are. For example, on column ‘NBA_fg%’, we can see that Stephen Curry is above the average and, int the other hand, Jonny Flynn or Brandon Jennings are way below the average.

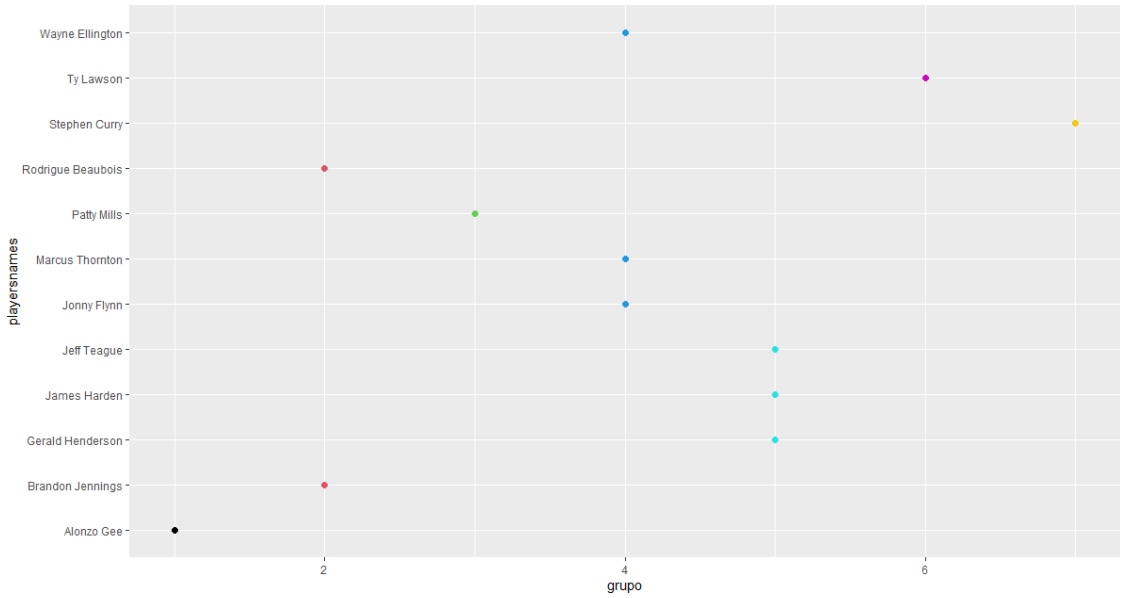

Figure 20: Clustering for NBA 2009 dataset Kmeans (7 clusters)

The Figure 20 shows a clustering with 7 clusters for the 2009 dataset. This number was selected as adequate by the system explained in the previous chapter. In this chart we can observe that Stephen Curry is all alone in the shooting graphic. Every player is grouped according to their shooting stats. So that means that every player that is alone has at least one shooting characteristic where they are unique. This doesn't imply been better on shooting, but that can tell us about something special on this player

It is interesting the analysis of the merge dataset 2009 + 2018 because the last data on Draft classes is added and we can extract some conclusions of the new clustering. We believe that the comparison of new players with old ones may be of interest to conclude the variables that best characterize the players and about the evolution in Draft competitions. After merging the data and choosing a set of numerical variables common to the two datasets, we have studied the behaviour of $ `betweenss` and $ `tot.withinss` to determine the K value of the clustering. The results can be seen in Figure 21. It follows that k = 13 can be an adequate number of groups. In total we have 64 players, so a homogeneous distribution in number would be 5 which is not unreasonable.

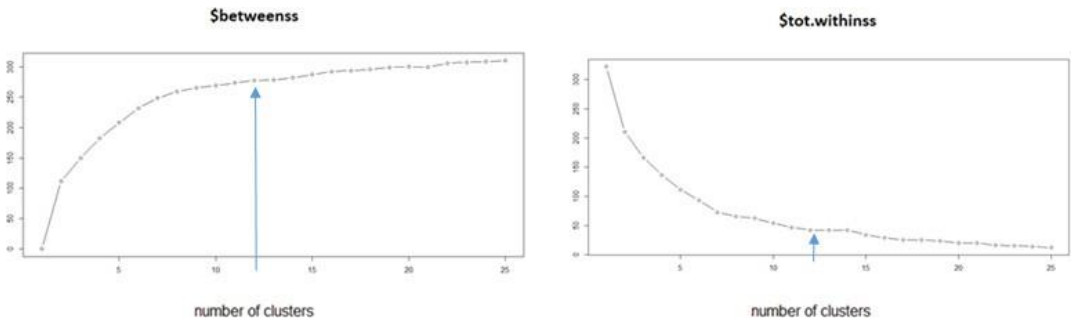

Figure 21: Selection of K for the merge dataset 2009 & 2018

Figure 22 shows the clustering vector for NBA dataset reported by R statement `datos.kmeans[["cluster"]]`. The list of the names and the number of the cluster assigned to

each player is shown. In a different way Figure 23 plots the same information into a chart. This Figure it is more visual because every cluster has a different colour.

```
Clustering vector:
      Bam Adebayo      De'Aaron Fox   Frank Ntilikina    Furkan Korkmaz
                9                 6                 3                10
    Jarrett Allen      John Collins    Jonathan Isaac      Josh Jackson
                1                 1                 3                11
  Lauri Markkanen        Malik Monk    Markelle Fultz        OG Anunoby
                4                 8                 9                12
Terrance Ferguson     Thomas Bryant      Tony Bradley      Zach Collins
                3                13                 7                13
       Alonzo Gee       Austin Daye  Brandon Jennings     Byron Mullens
                3                 3                 4                12
   Chase Budinger    DaJuan Summers       Danny Green  Dante Cunningham
                3                13                 8                12
     DeJuan Blair     DeMar DeRozan     Derrick Brown        Earl Clark
                9                11                 2                13
 Gerald Henderson   Hasheem Thabeet      James Harden     James Johnson
                6                 9                 4                12
       Jeff Ayres       Jeff Teague      Jon Brockman     Jonas Jerebko
                9                 6                 9                 3
      Jonny Flynn   Marcus Thornton       Omri Casspi       Patty Mills
                3                 8                12                 8
Rodrigue Beaubois       Serge Ibaka     Stephen Curry Terrence Williams
                3                 6                 5                12
        Ty Lawson      Tyreke Evans   Wayne Ellington
                6                11                 8
```

Figure 22: Clustering vector for NBA 2009 & 2018 dataset

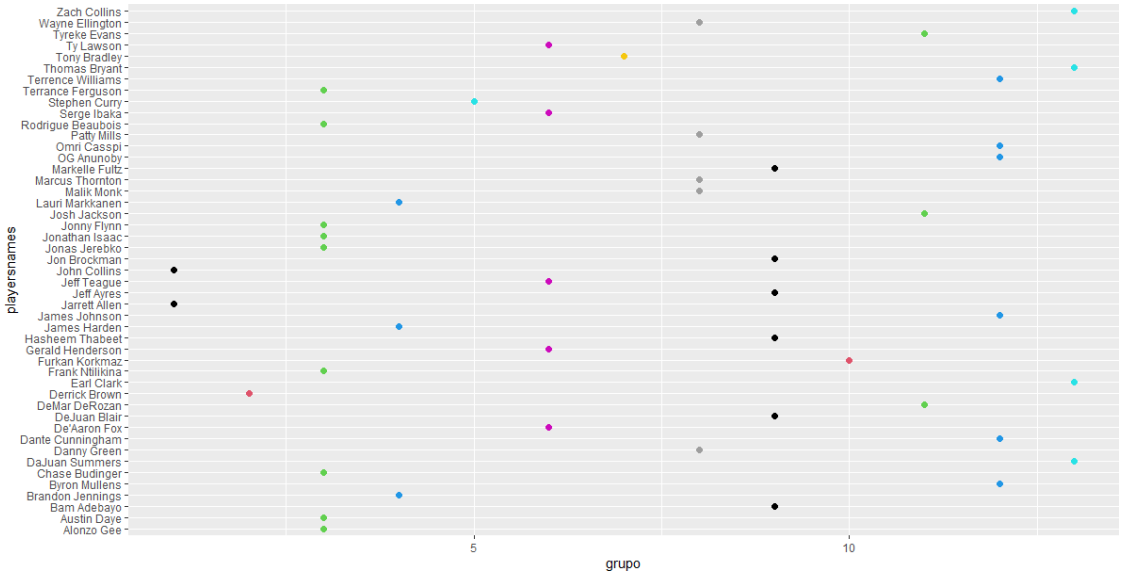

Figure 23: Clustering chart for NBA 2009 & 2018 dataset

In Figure 23 the clustering for only NBA stats is shown. The first thing that catches our eye is the appearance of Stephen Curry, once again, alone. Out of a total of 56 players studied, only 4 players group alone, one of which is Stephen Curry.

The same study over NCAA data creates Figures 24 and 25 in the same way:

```
> #INERCIA ENTRE GRUPOS:  mayor es mejor
> datosNCAA.kmeans$betweenss
[1] 279.2771
> #INERCIA INTRA GRUPOS:  menor es mejor
> datosNCAA.kmeans$withinss
 [1]  0.7342007  0.0000000  1.7454304  1.9670432 12.3376871  1.7508007  6.7361645 10.2810675
 [9]  0.6324035  2.7277570  0.0000000  2.5178927  1.2924231
> #INERCIA TOTAL INTRA GRUPOS:  menor es mejor
> datosNCAA.kmeans$tot.withinss
[1] 42.72287
```

```
Clustering vector:
      Bam Adebayo      De'Aaron Fox   Frank Ntilikina    Furkan Korkmaz     Jarrett Allen
                5                 1                 8                 2                 3
     John Collins    Jonathan Isaac      Josh Jackson   Lauri Markkanen        Malik Monk
               13                 8                 1                 6                 7
   Markelle Fultz       OG Anunoby Terrance Ferguson     Thomas Bryant      Tony Bradley
                5                 8                 7                12                11
     Zach Collins       Alonzo Gee       Austin Daye  Brandon Jennings     Byron Mullens
               12                 8                 8                 6                 8
   Chase Budinger    DaJuan Summers       Danny Green  Dante Cunningham      DeJuan Blair
                7                12                 7                 8                 5
    DeMar DeRozan     Derrick Brown        Earl Clark  Gerald Henderson   Hasheem Thabeet
                9                 3                12                10                 5
     James Harden     James Johnson       Jeff Ayres      Jeff Teague      Jon Brockman
                4                 8                 5                10                 5
    Jonas Jerebko       Jonny Flynn   Marcus Thornton       Omri Casspi       Patty Mills
                8                10                 6                 8                 7
Rodrigue Beaubois       Serge Ibaka     Stephen Curry Terrence Williams         Ty Lawson
               10                13                 4                 8                10
     Tyreke Evans   Wayne Ellington
                9                 7
```

Figure 24: Clustering vector for NCAA dataset

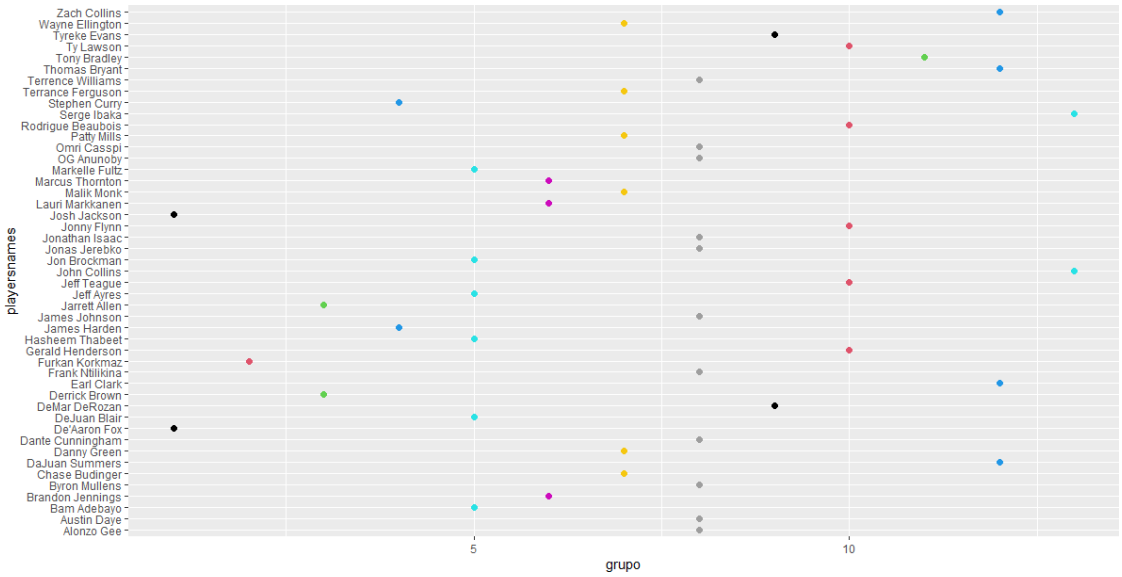

Figure 25: Clustering chart for NCAA dataset

The first thing we are going to observe is the behavior of Stephen Curry, to verify that the study shows something out of the ordinary, and it is coincidence.

This time, in the graph obtained through the NCAA shooting data study, we can see that the player is not alone, another player is along with him. The player who accompanies him is James Harden. Currently a top 3 player in the NBA, 2017 - 2018 MVP and NBA's best scorer of the last three seasons.

James Harden was chosen number 3 of the 2009 Draft by Oklahoma City Thunder, a team in which he had difficulties at the beginning in his career, due to having two stars ahead of him such as Kevin Durant or Russell Westbrook. Finally, he was traded to the Houston Rockets in 2014, where he could show all its potential.

These data demonstrate that there are common NCAA characteristics and statistics among NBA superstars. What makes incidence once again in the vital importance of studying the shot (especially statistics of percentages of different types of shots, volume of shots) before other characteristics related to the players physique. Factor that also makes us realize that a player with a great physique can dominate in the NCAA, but it is not usually the case in the NBA.

5.2 Clustering Visualizations

- **Chernoff Faces**

The guidelines that have been followed to apply the Chernoff faces method have been divided into the following:

1. Dataset segmentation: When deciding on which variables to relate and study, the original datasets are modified to obtain the most specific relationships possible. Data can be ordered by years, playing position, statistics, etc.

2. Preparation of the environment: installing the necessary libraries for the development of the work. In this case, the 'aplpack' library is used, which includes the 'faces' function that allows the representation of the vector created with the Chernoff algorithm.

In the 2009 draft, the choices of each team had special relevance for their franchises, given that they are considered one of the best generations of the Draft in recent years. Stars like James Harden, Stephen Curry or Blake Griffin, were selected in this class of the Draft. In this Draft class, bad predictions were given that marked the future of the teams, by letting players pass who turned out to be better than the choice made. Stephen Curry, Demar Derozan and Jrue Holiday, were drafted players greatly underrated by teams and would end up being stars.

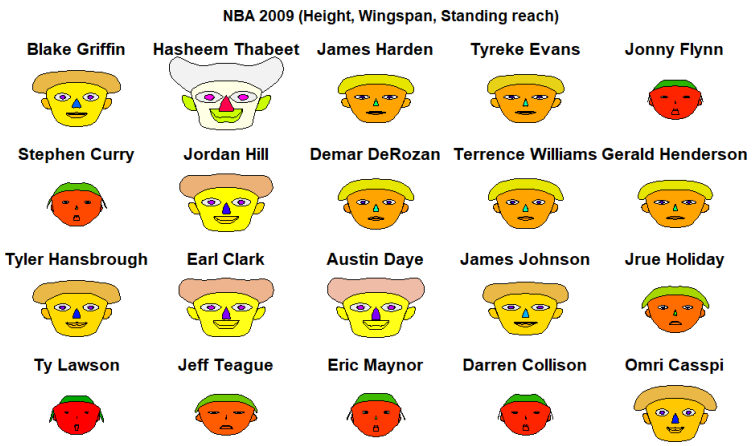

Figure 26: Chernoff faces NBA 2009 for Height, Wingspan, Standing reach

We establish the relationship between the first 20 players of the 2009 Draft in terms of the physical characteristics of height, wingspan, and standing reach (Figure 26). Above them stands out a player who would end up being number 2 in the Draft, Hasheem

Thabeet. This player would end up being a mediocre player, so not always that he excels in certain variables implies greater value in the player.

In the previous representation, it can be seen that except in the case of number 2 (Hasheem Thabeet), the rest of the players have normal characteristics in terms of general dimensions (taking into account their playing position). Besides, the rest of the graphs are shown (Figures 27 and 28) that refer to the physical aspects of the players (weight, body fat, vertical jump, physical tests).

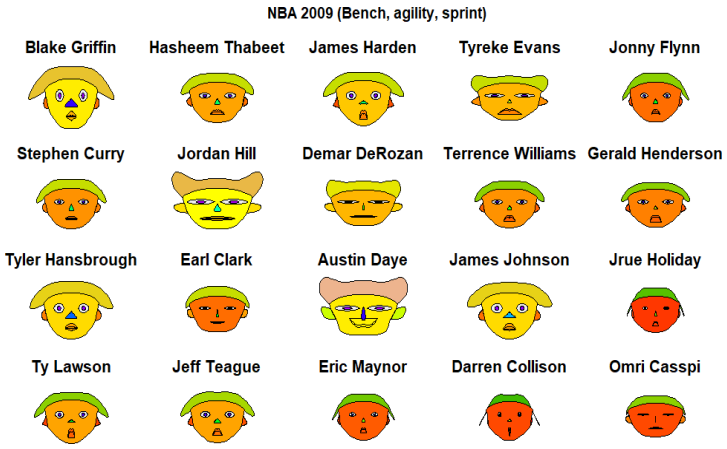

Figure 27: Chernoff faces NBA 2009 for bench, agility and sprint

In the graphics, we can see that in terms of physical evidence, the most similar player to number 1 in the Draft was James Johnson.

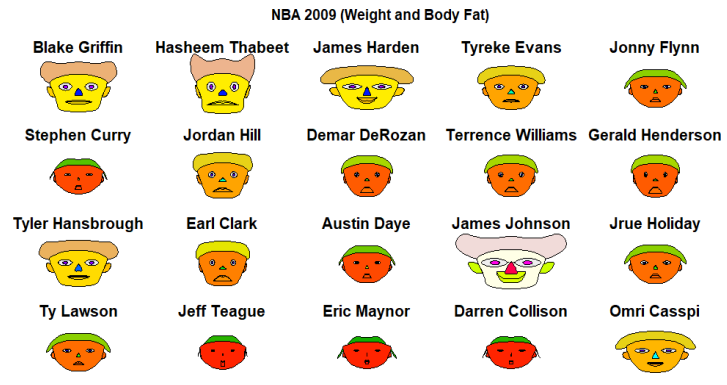

Figure 28: Chernoff faces NBA 2009 for weight and body fat

Figure 28 represents the relationship of weight and body fat. James Johnson, the player who stood out in terms of physical tests, this time stands out negatively in this regard.

This can tell us why James Jhonson struggled with his Draft pick and with the first years of NBA career.

Figure 29: Chernoff faces NBA 2009 for vertical tryouts

The variables that refer to vertical jump are represented, and the same 'small' players (Eric Maynor, Darren Collison) follow the trend in terms of low physical conditions (height, weight, jump), but in this characteristic (vertical jump) a player changes the trend concerning his position. Jonny Flynn shows very good jumping qualities compared to the other players who tended to be similar.

From the representations that have been exposed, conclusions can be drawn about several things that could be considered in the Draft elections about the physique of the players:

1. James Johnson, despite having a very high percentage of body fat, achieved better results in physical tests than many other players.

2. Jonny Flynn showed excellent physical condition compared to the other players in his position.

3. Hasheem Thabeet possesses well above average physical qualities in terms of height and size.

- **Graphics and Dashboards in Power BI**

Thanks to the correct crossing of data between the different tables, different interactive visualizations have been obtained. The graphs represent the variety of relationships between the statistics studied (age, physical qualities, shooting percentages, game statistics, etc.). One of the greatest advantages of this tool is the possibility of manipulating the graphics, which can be adapted to suit the user. It can be filtered by years, players, positions and more. The most important examples of the study are described below.

The first actions carried out with this tool have consisted on the creation of a panel through which we can see graphs relating the percentages of the different types of shots. Filters can be added to study specific players, seasons or positions within the data sets (see Figure 30).

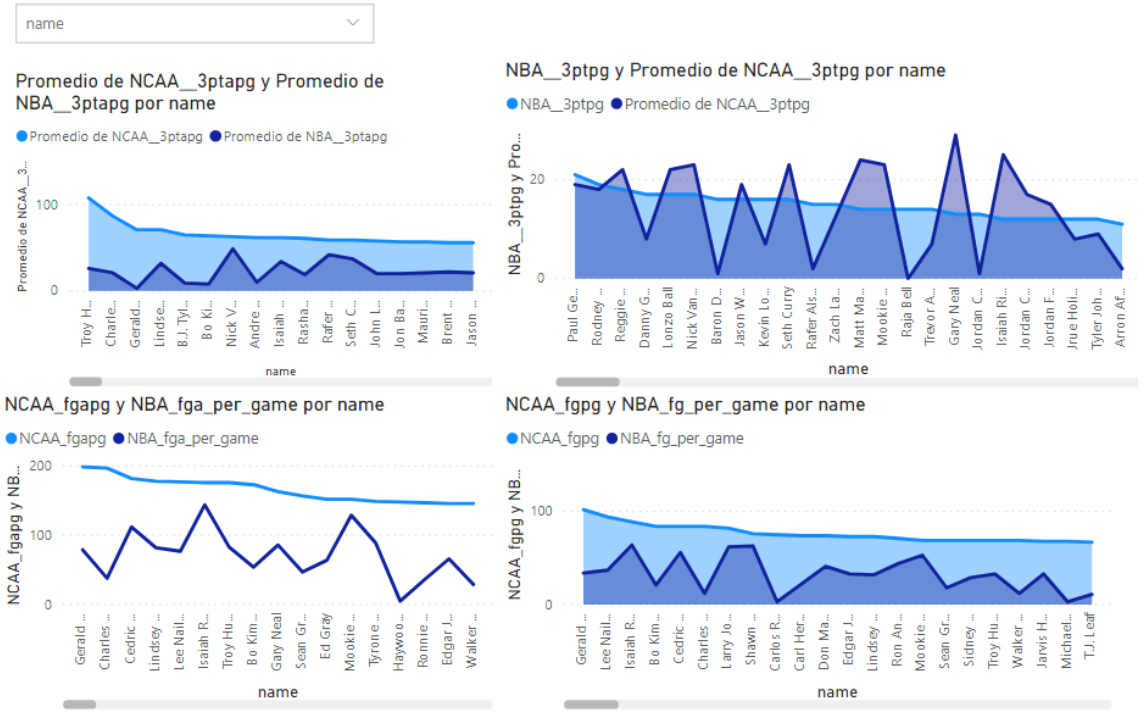

Figure 30. PowerBI**:** Different shooting stats among a great set of players.

Regarding the case studied in Draft 2009, the graphs shown in Figure 31 have been obtained.

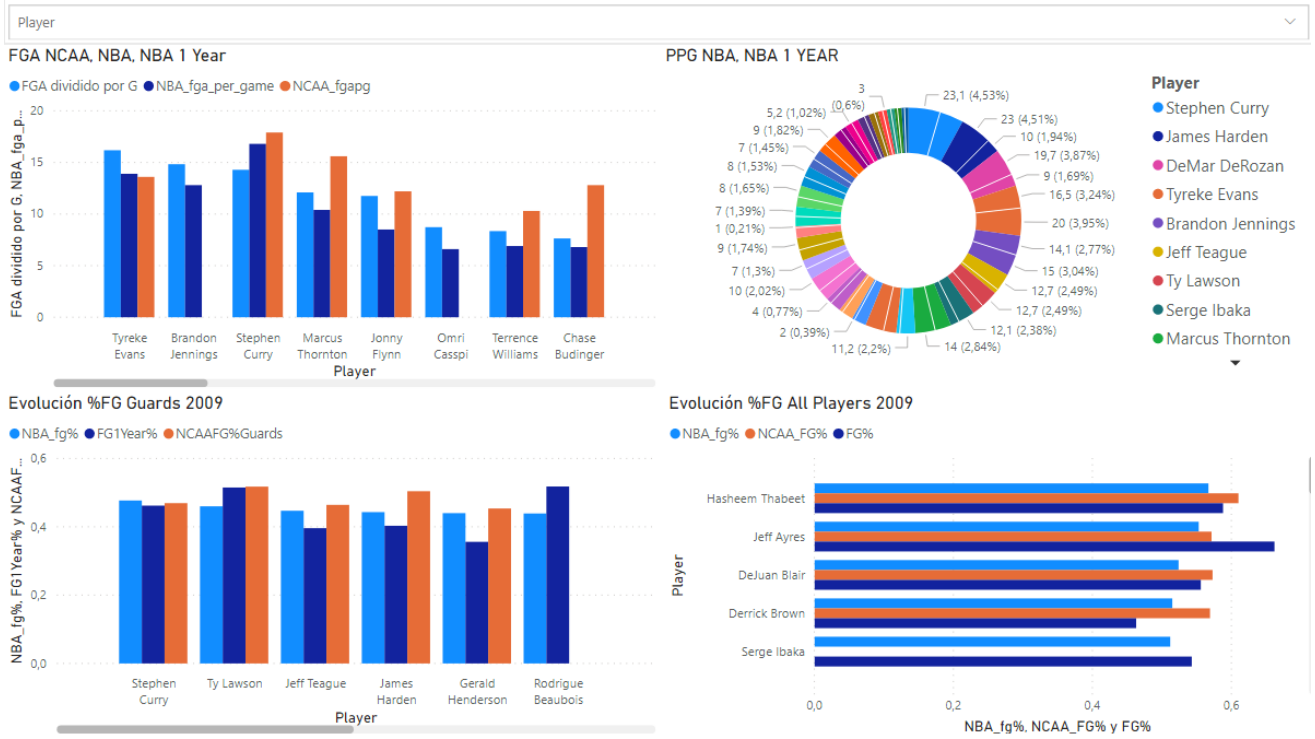

Figure 31. PowerBI: Results over Draft 2009

In Figure 31, the upper graph represents the evolution in the shooting of the players of the 2009 Draft class in NCAA, the first year of NBA and NBA career. Stephen Curry is a player who is consistently leading the charts.

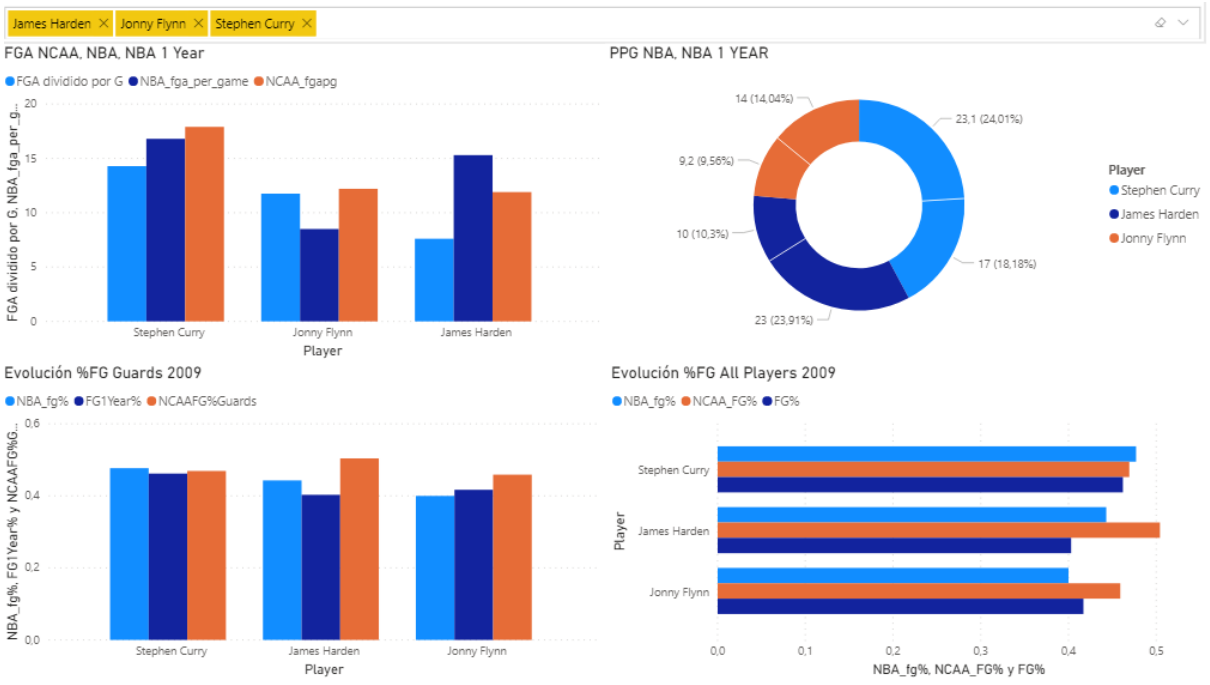

Figure 32. PowerBi**:** Filtered results over 2009 Draft

As it’s shown in Figure 32, filtering by players (current stars and an example of an unsuccessful high-choice player), we can see the difference between players and Jonny Flynn's negative trend in shooting percentages and volume. Looking at the graphs above, we can interpret that players with a high shooting percentage combined with a high volume of shots in the NCAA tend to have better NBA careers, especially among outside players. On the other hand, if you have a high volume of shots, but low percentages of effectiveness, they tend not to succeed in the long term. Players like James Harden do not show the outstanding percentages in the first year of NBA of Stephen Curry, but they do show a constancy that has allowed them to follow a positive long-term trend.

In addition, by filtering by those stars, together with a player selected in a high position, who ended up being a mediocre player (Jonny Flynn), we can observe the differences and extract the information that the Minnesota Timberwolves did not take into account when making the mistake in the election of the player.

Chapter 6. Conclusions and Future Word

This work exposes an ongoing problem in the basketball industry. The analysis of current and historical data on the players is proposed to make clustering adapted to the requirements of the teams at all times. In this way, coaches can better understand the profile of new players by comparing themselves with others with similar characteristics and making better decisions. The work shows interesting results about the 2009 Draft that exemplifies its usefulness. It also shows the results for a more recent Draft class.

The most relevant result of this study has to do with the comparison of shooting statistics in the NCAA between players of different generations. This result shows the grouping of two players who are current stars in the NBA based only on statistics during their college days. This could indicate that there is a common pattern that teams should consider when making their Draft choices.

As future work, the improvement and settlement of the relationships studied in this work can be determined, as well as the automation of the study in general. Carry out a program that can help teams in the search for new players and in making decisions about the direction that each franchise should take.

Chapter 7. References

1. Hoover SJ, Winner RK, McCutchan H, et al. (2017) Mood and Performance Anxiety in High School Basketball Players: A Pilot Study. Int J Exerc, Sci. 2017;10(4):604-618. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5466400/

2. Akenhead, R. and Nassis, G. P. (2016) Training Load and Player Monitoring in High-Level Football: Current Practice and Perceptions, International Journal of Sports Physiology and Performance, 2016, 11, 587 -593 http://dx.doi.org/10.1123/ijspp.2015-0331

3. NbaDraft web site, https://www.nbadraft.net/, last access, August 30, 2020

4. Tankathon web site, http://www.tankathon.com/ , last access, August 30, 2020

5. Barzilai, A. (2007) Model Assessing the Relative Value of Draft Position in the NBA Draft, article plublish on http://www.82games.com/barzilai1.htm, last access, August 30, 2020

6. Stats NBA https://stats.nba.com/, last access, August 30, 2020

7. Basketball Reference, https://www.basketball-reference.com/, last access, August 30, 2020

8. Watave, A. (2016) Relative Value of Draft Position in the NBA, Undergraduate Senior Thesis, B.A Economics, University of California, Berkeley.

9. Chernoff, H. (1971) The use of faces to represent points in k-dimensional space graphically. Technical Report 71, Department of Statistics, Stanford University

10. Astel, A. et al. (2006) Classification of drinking water samples using the Chernoff's faces visualization approach. Polish Journal of Environmental Studies, vol 15 pp. 691-697

11. Raciborski, R. (2009) Graphical representation of multivariate data using Chernoff faces, The Stata Journal, vol 9, N.3 pp.374-387

12. Morris CJ, Ebert DS, Rheingans P. (1999) An Experimental Analysis of the Pre-Attentiveness of Features in Chernoff Faces, SPIE proceedings of Applied Imagery Pattern Recognition: 3D Visualization for Data Exploration and Decision Making

13. Reyes, J., (2019) Ideas for the use of Chernoff faces in school cartography, MTA-ELTE Research Group on Cartography and GIS, National Office for Research and Technology of Hungary and Hungarian Scientific Research Fund (OTKA).
14. Standford, Michael. Hierarchical Cluster Analysis. http://www.econ.upf.edu/~michael/stanford/maeb7.pdf
15. Murtagh F, Contreras P (2012). Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. Wiley Online Library. 2 (1): 86–97. doi:10.1002/widm.53.
16. Everitt B (2011). Cluster analysis. Chichester, West Sussex, U.K: Wiley. ISBN 9780470749913
17. Krijnen, W.P. (2009) Applied Statistics for Bioinformatics https://cran.r-project.org/doc/contrib/Krijnen-IntroBioInfStatistics.pdf
18. Golub, et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science, 286(5439):531-7. DOI: 10.1126/science.286.5439.531
19. R Cran Project https://cran.r-project.org/, last access August 30, 2020
20. Rstudio web site www.rstudio.com, last access August 30, 2020
21. PowerBI https://powerbi.microsoft.com/ , last access August 30, 2020
22. GitHub www.github.com, last access August 30, 2020
23. Hartigan, J.A. & Wong, M.A. (1975). A k-means clustering algorithm. Applied Statistics, 28, 100-108.
24. Pollard, D. (1981). Strong consistency of K-means clustering. Annals of statistics, 9, 135-140.