University of Luxembourg

Bachelor's Thesis

Statistics in Basketball: Settling NBA Debates analytically

Author:
Oliver Jack

Supervisor:
Prof. Christophe Ley

May 23, 2023

Abstract

The following work looks to showcase the importance of statistics in today's world of sports, in particular in basketball, and how they can help settle certain ongoing discussion topics. Based on the deep analysis that will be presented, the reader will have a better understanding of the great variety of data that can be collected at competitive events and how analytical tools can help shed light on the most specific of details.

I would like to personally thank *Prof.* Christophe Ley for all of his help and guidance throughout my bachelor's thesis.

Contents

1 Introduction

Being one of the oldest ball sports on the planet, basketball has continuously grown in popularity over the last century. Nowadays, it is watched and enjoyed by millions of people worldwide on a daily basis. Meanwhile, the game itself and the way it is played has evolved over time, sometimes making it difficult to compare the new era of basketball to how it was decades ago. This has lead to the arising of heated debates regarding critical topics such as who is the best of all-time or which players are underrated/overrated. Often fuelled by subjective biases and opinions, it can be hard to come to any agreement without conclusive evidence.

The emergence of descriptive statistics and analytical methods can help settle these discussions. Thanks to the constant evolution of technology, most basketball leagues are now able to provide a solid collection of statistics related to players and teams. Naturally, this has opened the door for a deeper analysis of this data, pushing teams to strengthen their knowledge in the field of basketball analytics, in order to gain an advantage over the opponent. While certain information is kept private for the teams, the vast majority of the collected data is shared with the public. This has allowed more and more analysts and fans to form their own opinions based on stats, especially when making their assessments of players' performances.

The aim of this work is to compare former and current players and teams of the NBA (National Basketball Association) at an analytical level, while underlining certain key differences between the so-called "old school" generation and the modern era. This thesis will provide insights into the game of basketball that go beyond traditional metrics such as points, rebounds and assists, while trying to settle debates in a more objective and evidence-based manner. The target audience of this research is not only basketball fanatics, but also anyone who admires the growing power and importance that science can have in sports.

There are several reasons why this research paper is primarily based on the statistics gathered in the NBA, the highest male basketball league in the United States (and arguably in the world). First of all, being one of the oldest basketball leagues worldwide, the NBA has a larger historical database to work with. Next, the NBA has been able to introduce extremely advanced and reliable technology regarding the tracking and collection of data, making it easier than ever to look up any precise stat almost instantly. Finally, the NBA has hosted some of the all-time greatest basketball players and teams over the years, which makes it increasingly interesting to conduct this study at the highest basketball level possible.

To help understand the data that will be used throughout this bachelor's thesis, we will first take a brief look at some of the key basic statistics and how they can be analysed and manipulated using the advanced methods offered by the R package `BasketballAnalyzeR`. After having laid the foundations, I will start off by comparing some of the greatest basketball players of all-time based on their statistical performances alone. By taking a step back, I will then weigh off some of the best single season teams to have played in the NBA. The following section focuses on identifying the main eras in NBA history, highlighting some key analytical differences between the game as we know it today and how it was played many years ago. Finally, the last section analyses which statistical factors play a crucial role in determining a player's salary. I will also briefly touch upon several other methods that are commonly used in basketball analytics nowadays. The main ideas of this work will then be summarised in the final conclusion, followed by a short appendix containing the used code and further interesting information. (Note: The majority of the used data was taken from the source `https://www.basketball-reference.com/`.)

2 Types of Statistics & Metrics in Basketball

When it comes to sports, basketball offers some of the greatest variety in terms of statistics that can be gathered, thanks to the fast pace of the game. First, it is necessary to lay the groundworks by explaining the most fundamental stat lines. Thereafter, some more complex metrics related to individual as well as team performances will be presented. These will all play an essential role in the analysis throughout the thesis.

2.1 Basic Statistics

While the NBA has continuously improved at the technological level, most of the basic data are still being recorded manually today in a so called ”box score” (summary of the most basic statistics). Here is a list of the most important statistics, accompanied by a brief description of what they mean. (Note: These stats can be naturally extended to a team.)

Points (PTS): The number of points a player scores in a game. 1 point is received when scoring a free throw (FT), i.e. a shot that you get after being fouled, 2 points when scoring a 2-point field goal (2FG), i.e. from inside the three point line, and 3 points when scoring a 3-point field goal (3FG), i.e. from behind the three point line. The points often times go hand in hand with the field goal percentage (FG%), which expresses the success rate of the shots.
Assists (AST): The number of times a player passes the ball to a teammate, who then scores.
Rebounds (REB): The number of times a player grabs the ball following a missed shot. They can be subdivided into offensive (OREB) and defensive (DREB) rebounds.
Steals (STL): The number of times a player gains possession of the ball from an opposing player.
Blocks (BLK): The number of times a player blocks a shot of an opposing player.
Turnovers (TO): The number of times a player loses possession of the ball to an opposing player.
Personal fouls (PF): The number of times a player fouls an opposing player.

2.2 Advanced Metrics

Next, we will see some of the more complex metrics used by analysts nowadays. These metrics tend to highlight key information that may be overseen, when only looking at the basic box score of a game. Here a short example: Consider a player who plays most of the game, takes a lot of shots, but hardly contributes on defence. At the end of the game, he might have impressive box score stats, however, his team might have lost due to his lack of defensive commitment. In this case, it is beneficial for coaches or analysts to consider other metrics, when assessing the individual performance of each player.

When it comes to these benchmarks in basketball, there is no ”holy grail” of all stats. In other words, there is no optimal value that can be attributed to a player or to a team to rate their overall performance. Here are some of the most commonly used metrics and the logic behind them. (Note: The exact formula and calculations of the more computational metrics can be found in Section 10). It is important to add that alternative formulas need to be used when determining it for seasons before 1979-1980, since several components used in the calculations were not recorded before then.)

2.2.1 Player Efficiency Rating

Introduced by John Hollinger in 2007, the *Player Efficiency Rating* (PER) was one of the first player evaluation benchmarks to be used on an international level and can be considered as the cornerstone of all-in-one metrics. The aim of the PER is to measure the per-minute rating of a player, based on the pace of the game. The final value is composed of the addition of positive stats (made shots, assists, rebounds, steals and blocks) and the subtraction of negative stats (missed shots, turnovers and fouls). The NBA standard PER is set at 15, making it easier to compare the performances of players. Since it mainly focuses on the offensive display of a player, it neglects the defensive aspects to a certain extent.

2.2.2 Box Plus-Minus

In 2020, Daniel Myers developed the newest version of the *Box Plus-Minus* (BPM), a player evaluation metric that attempts to measure the player’s overall contribution to the team when they are on the court, using only statistics that are widely available. The BPM considers the box score stats of a player, the team’s overall performance and the player’s position. The resulting BPM score estimates the number of points per 100 possessions that a player contributes to the team, above or below the league average. In this sense, a positive BPM would indicate that a player is contributing more than the average player, whereas a negative score would mean that they are contributing less.

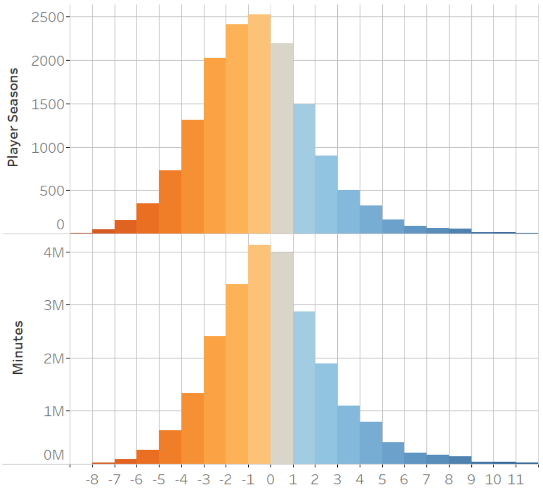


Figure 1: Histogram containing all player seasons from 1974-2019. Source:[9]

Given the bell-shaped curve of the histogram, the distribution seems to follow the normal law, centred at a mean BPM score of 0. This normal distribution would suggest that most NBA players have a relatively average impact on the game (with a BPM score close to the average 0), while fewer players have either a very positive or very negative impact.

2.2.3 Offensive/Defensive Rating

The *Offensive* and *Defensive Rating* (ORtg & DRtg) are two team and player evaluation metrics which underline the performance on both ends of the court. It was first introduced in the scientific paper *A Starting Point for Analyzing Basketball Statistics* by Kubatko et al. (2007)[5]. In a nutshell, it represents a team's efficiency by measuring the number of points scored and allowed per 100 possessions. The reason why we use a per 100 possession index is due to the unpredictable pace of the game. In general, the number of possessions a team has in a game heavily depends on the playstyle and the opponent. Therefore, simply measuring the total points scored or allowed by a team per game would not provide us with a normalised metric, which would make team comparisons ambiguous. To compute the ratings, one first needs to calculate the number of possessions (POSS) per game:

$$POSS = FGA + 0.44FTA + TO - OREB \tag{1}$$

where FGA and FTA stand for field goals/throws attempted, TO stands for turnovers and OREB stands for offensive rebounds. The formula is built on the idea that, by definition, a turnover leads to an automatic possession change, while attempted field goals and free throws only result in a possession change if the attacking team doesn't grab an offensive rebound. This then allows us to calculate the ORtg and DRtg using the following formula:

$$ORtg = \frac{PTS_T}{POSS_T} \tag{2}$$

$$DRtg = \frac{PTS_O}{POSS_O} \tag{3}$$

where T stands for the team itself and O stands for the opponent. Furthermore, this simplified computation gives us the pace of the game:

$$Pace = 5 \cdot \frac{POSS}{MIN} \tag{4}$$

where MIN stands for the total minutes played by all the players. Finally, the difference between the ORtg and the DRtg of a team is referred to as the *Net Rating* (NRtg)

$$NRtg = ORtg - DRtg$$

and will be used in the upcoming sections.

2.2.4 Four Factors

The *Four Factors* (FF) is a team evaluation metric which covers 4 specific aspects of a team's performance: shooting (40%), turnovers (25%), rebounding (20%) and free throws (15%). The idea came in Kubatko et al. (2007)[5], where they tried to find a logical answer to what factors have the greatest impact on the outcome of a game. Moreover, the percentages mentioned above represent the weighted importance of each component.

The formulas of the different factors, which can be computed for both offense (O) and defense (D), are the following:

$$OeFG\% = \frac{2FG_T + 1.5 \cdot 3FG_T}{FGA_T} \qquad DeFG\% = \frac{2FG_O + 1.5 \cdot 3FG_O}{FGA_O} \tag{5}$$

$$OTO\ Ratio = \frac{TO_T}{POSS_T} \qquad DTO\ Ratio = \frac{TO_O}{POSS_O} \tag{6}$$

$$OREB\% = \frac{OREB_T}{OREB_T + DREB_O} \qquad DREB\% = \frac{DREB_T}{OREB_O + DREB_T} \tag{7}$$

$$OFT\ Rate = \frac{FT_T}{FGA_T} \qquad DFT\ Rate = \frac{FT_O}{FGA_O} \tag{8}$$

Here, the subscript T and O stand for team and opponent, while M and A stand for made and attempted.

2.3 Self-developed Metrics

2.3.1 Player Performance Index

In order to measure the performance of players from different generations, I decided to come up with my very own metric, which considers multiple key factors of a player's career. Named the *Player Performance Index* (PPI), the idea behind it is to split it into two main parts and then calculate the product of both. The first part should be a composite metric, composed of three advanced metrics, namely PER, BPM and NRtg. The second part should take several high calibre achievements into consideration, such as the number of Championships, MVP awards, All-NBA and All-Star selections, and playoff wins.

The reason why I selected the three advanced metrics mentioned above is because they individually cover a different aspect of a player's game, as well as being relevant when it comes to determining the greatest player of all-time. While the PER puts more weight on the offensive attributes of a player, the BPM tends to value the individual performance at both ends (offence & defence). Finally, the NRtg considers the difference between the offensive and defensive capabilities of a team when that particular player is playing. Although they emphasise on different areas of the game, they are clearly correlated to a certain extent (e.g. PER & BPM both take individual offensive performance into consideration). While it may be possible to determine the correlation between the metrics theoretically (i.e. by understanding the underlying mathematical relationships between the variables), it can be quite difficult and time-consuming for complex formulas like PER, BPM and NRtg. To avoid redundancy, a dimension reduction technique can be used to linearly combine the three benchmarks, without having to exclude one of them.

In particular, I will be using a well-known dimension reduction method called *Principal Component Analysis* (PCA), allowing me to reduce the dimensionality of the data, while keeping the most important information and relationships between the variables. More precisely, this technique transforms the original variables into a new set of orthogonal variables, called principal components. These components are sorted in descending order, according to their influence on the variance in the data, i.e. the first principal component captures the most variance in the data, followed by the second principal component etc. Then by selecting a subset of the principal components which are responsible for the majority of the total variance, a composite metric can be formed using these main components.

To analyse the correlation between the benchmarks and apply the PCA technique, a data set containing a large number of players, as well as their career PER, BPM and NRtg, is necessary. I decided to simply take the 75 players who were nominated to the *NBA 75th Anniversary Team.* These are considered to be the 75 best NBA players of all-time, making it a great list to apply the PCA to. By collecting and organising this data, I then created a matrix of size 75 × 3, representing the different players and their respective indices (the associated CSV file can be found in the appendix). The next step consists of normalising the data, since the three metrics are measured on different scales and are not equally weighted otherwise. To perform the PCA, it is best to standardise each column, so that it has a mean of 0 and a standard deviation of 1. The standardisation of a random variable X is done by computing the following linear transformation Z:

$$Z = \frac{X - \mu_X}{\sigma_X} \tag{9}$$

where μ_X is the mean of X and σ_X is the standard deviation of X.

Although the collected data might not be perfectly normally distributed, PCA is still applicable, since it is a robust method which can be used on random variables of any distribution. (Using the Shapiro-Wilk test on the BPM and NRtg, the null hypothesis that the data is normally distributed is rejected, meaning that the data is most likely not normally distributed.)

The first step of PCA consists of computing the covariance matrix of the standardised data matrix X, which can be done by using the following formula:

$$Cov(X) = \frac{1}{n-1} X^T X \tag{10}$$

In this case, it is equal to:

$$Cov(X) = \begin{pmatrix} 1.014 & 0.761 & 0.532 \\ 0.761 & 1.014 & 0.649 \\ 0.532 & 0.649 & 1.014 \end{pmatrix}$$

The next step is to simply calculate the eigenvectors and eigenvalues of the obtained matrix. While the eigenvectors determine the direction of the maximum variance, the eigenvalues represent the magnitude of variance associated to each eigenvector. The three orthogonal eigenvectors are regarded as the principal components, each one of them representing a direction in the 3-dimensional space defined by the three advanced metric variables. They are then sorted in descending order, according to their corresponding eigenvalue. In this case, they are the following:

$$v_1 = \begin{pmatrix} 0.573 \\ 0.624 \\ 0.532 \end{pmatrix}, v_2 = \begin{pmatrix} -0.603 \\ -0.119 \\ 0.789 \end{pmatrix}, v_3 = \begin{pmatrix} 0.555 \\ -0.773 \\ 0.308 \end{pmatrix}, \lambda_1 = 0.745, \lambda_2 = 0.186, \lambda_3 = 0.069$$

When it comes to deciding how many principal components should be considered in the final metric, there are several methods that can be used, such as the variance proportion method or the scree plot method. The variance proportion technique consists of selecting a sufficient number of components, in order to encompass enough of the total variance. Generally, this proportion is set between 70%-80%. In our case, the first principal component alone seems to be sufficient, since it is responsible for over 71% of the data. Graphically, the ”elbow” of the scree plot can be used as a good cut off point, when selecting the number of components.

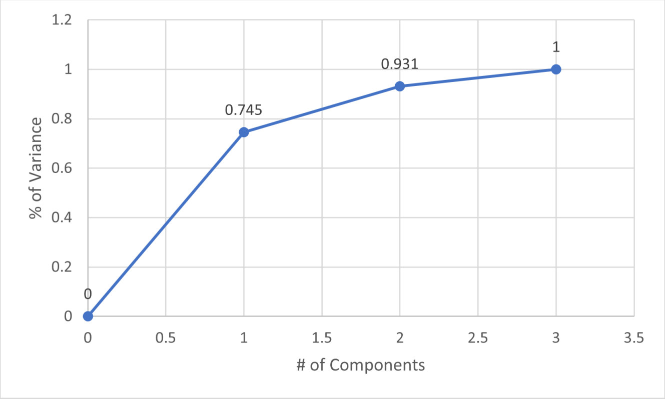


Figure 2: Scree plot related to this example

The values of the first eigenvector can be used as the weights of the normalised variables in the final composite metric, which is given by:

$$CM = 0.573PER + 0.624BPM + 0.532NRtg \tag{11}$$

It is important to note that the standardised values for these variables will need to be used in this formula, instead of the original ones. Furthermore, the standardisation of the data and the complete PCA process entirely depends on the initial set of players you work with, in this case, the list of 75 players and their respective attributes. It would be ideal to work with the set of all former and current NBA players, however, this is not feasible from a data collection stand point.

The second part of the PPI is a weighted linear combination of several individual career achievements, as previously mentioned. Since the main objective of this metric is to help settle the debate of the greatest player of all-time, it is only common sense to give the most value to the greatest team prize in basketball: winning an NBA championship. For many people this is the only criterion that matters, however, there are also other achievements that should be considered in my opinion. The MVP (Most Valuable Player) award is given to a single player who individually performs at a higher level than any other player in the league throughout an entire season and is considered to be the most prestigious individual award in basketball. Therefore, both championships and MVP awards will be weighted equally, with a weight of 1. Next, All-NBA and All-Star selections are also highly recognised achievements for a player, ranking them among the very best in any given season. While the All-NBA teams are made up of 15 players in total and are selected by broadcasters and sportswriters, the All-Star teams are composed of 24 players in total and are voted for by the fans. Finally, the number of postseason wins is a good indicator of how well a player is able to lead their team to victory when it matters most. On average, the best-of-7 series of the NBA playoffs are settled within 5 games, meaning that each series distributes an average of 5 wins. Since there are a total of 15 series during the playoffs, that would add up to 75 playoff wins that are usually distributed every season. Compared to championships or MVP awards, one can see that these achievements are far more attainable, which is why these factors will receive a smaller weight (All-NBA selections weight: $\frac{1}{15}$, All-Star selections weight: $\frac{1}{24}$, Playoff wins weight: $\frac{1}{75}$). An interesting observation is that these accolades are not normalised by dividing by the total number of seasons played. This gives the benefit of the doubt to those players who were able to compete at the highest level for a longer period of time.

Finally, the formula of the PPI is obtained by simply computing the product of these two components:

$$\begin{aligned} PPI =& max\{0.573PER + 0.624BPM + 0.532NRtg, 0\} \cdot \Bigg(Championships + MVPs \\ &+ \frac{All - NBA\ selections}{15} + \frac{All - Star\ selections}{24} + \frac{Playoff\ wins}{75} \Bigg) \end{aligned} \tag{12}$$

The reason why it is logical to take the maximum in the first component, is because the standardised values of the PER, BPM and NRtg can possibly be negative, hence the first component could end being negative. As a result, for two players who have an equal $CM < 0$, the player having achieved more throughout their career would have a lower PPI, which is contradictory. By these means, the floor of the index is simply set at 0.

2.3.2 Team Performance Index

The *Team Performance Index* (TPI) is a self-made team evaluation metric, which is deemed to measure a team's single season performance level, considering both the regular season and the postseason (i.e. playoffs). Once again, the index will be composed of multiple factors, such as advanced metrics and collective achievements, all covering different aspects of a team's success.

The first component of the metric is the ratio of weighted linear combinations of the offensive and defensive Four Factors. Regarding the offensive FF, three of the factors, OeFG%, OREB% and OFT Rate, are positively associated with a team's performance, while a higher OTO Ratio indicates a poor offensive execution by the team. Meanwhile, two of the defensive FF, DeFG% and DFT Rate, are negatively associated with the team's performance, whereas the DTO Ratio and DREB% are positively related. Furthermore, as proposed in Kubatko et al. (2007)[5], both the offensive and defensive FF can be attributed a certain weight, as previously mentioned. This leads to the establishment of the following weighted linear combinations:

$$LC_1 = 0.4 \cdot OeFG\% + 0.25 \cdot (1 - OTO\ Ratio) + 0.2 \cdot OREB\% + 0.15 \cdot OFT\ Rate$$

$$LC_2 = 0.4 \cdot DeFG\% + 0.25 \cdot (1 - DTO\ Ratio) + 0.2 \cdot (1 - DREB\%) + 0.15 \cdot DFT\ Rate$$

To balance the offensive (positively associated) and defensive (negatively associated) elements out, one can consider the ratio of both, which finalises the first of the three components:

$$R = \frac{0.4 \cdot OeFG\% + 0.25 \cdot (1 - OTO\ Ratio) + 0.2 \cdot OREB\% + 0.15 \cdot OFT\ Rate}{0.4 \cdot DeFG\% + 0.25 \cdot (1 - DTO\ Ratio) + 0.2 \cdot (1 - DREB\%) + 0.15 \cdot DFT\ Rate} \tag{13}$$

The next component is simply the ratio between the offensive and defensive rating of the team throughout the entire regular season. Since a higher ORtg is a positive indicator and a higher DRtg is a negative indicator for a team, one can take the ratio due to similar reasons as before. While the previous factor focuses on some specific key aspects of a team's offensive and defensive performance, this ratio attributes a more global value, mainly considering a team's performance over 100 possessions.

The last element of the TPI is composed of several team accolades: championship won, playoff wins, All-NBA and All-Star selections, and regular season wins. Similar to the PPI, these achievements are essential when comparing intergenerational teams and are weighted according to a logical reasoning. Since only one team can win the NBA championship each season, it will once again receive the most important weight. Next, the ratio between playoff wins and playoff games played illustrates how efficient the team is when it truly matters. As before, the number of All-NBA and All-Star players in the team also indicates how well a team is performing as a group of individuals. Here, the same weight reasoning is used as previously explained. Last but not least, the proportion of regular season wins shows how efficient the team has performed over a longer period of time. Every NBA team plays a total of 82 games during the regular season, therefore, the number of wins will be divided by 82 to simply obtain the winning percentage.

Finally, the formula of the TPI is given by the product of these three key factors:

$$\begin{aligned} TPI =& \frac{0.4 \cdot OeFG\% + 0.25 \cdot (1 - OTO\ Ratio) + 0.2 \cdot OREB\% + 0.15 \cdot OFT\ Rate}{0.4 \cdot DeFG\% + 0.25 \cdot (1 - DTO\ Ratio) + 0.2 \cdot (1 - DREB\%) + 0.15 \cdot DFT\ Rate} \\ & \cdot \frac{ORtg}{DRtg} \cdot \Bigg(Championship + \frac{Playoff\ wins}{Playoff\ games\ played} + \frac{All - NBA\ selections}{15} \\ & + \frac{All - Star\ selections}{24} + \frac{Regular\ season\ wins}{82} \Bigg) \end{aligned} \tag{14}$$

3 R Packages `BasketballAnalyzeR` & `nbastatR`

The package `BasketballAnalyzeR` for the statistical language R offers a wide variety of functionalities when it comes to analysing and visualising basketball data. The open-source package was developed by M. Sandri in 2020 and is accompanied by the book *Basketball Data Science: With Applications in R* [16] by P. Zuccolotto and M. Manisera, a good read for those interested in understanding how data science can be used in modern-day basketball analysis. Its development followed a primarily factual approach and was carried out as part of the activities of the international network BDsports. While its main goal is to provide powerful graphical features for scientific research and sports analytics, it is the perfect tool for us to use when comparing intergenerational players and teams analytically, as well as visually.

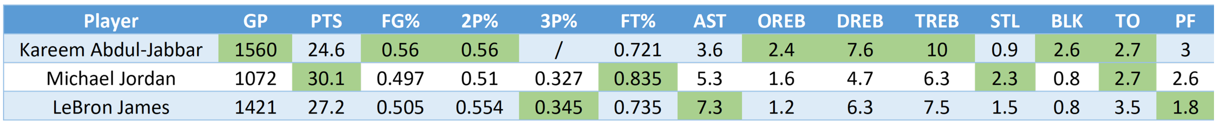
The R package `nbastatR`, originally developed by A. Bresler, is another powerful tool when it comes to the collection and distribution of data related to the NBA. Offering a wide range of basic and advanced stats, it provides reliable data dating back to the 1940s.

4 Greatest Player of all-time

Ever since basketball has become an internationally televised sport, watched by millions around the globe, the age-old "GOAT" (Greatest Of All-Time) debate was born. More often than not, the opinions of fans can be fuelled by emotions and biases towards a specific player/team, which explains why there are so many different points of view on the topic. To settle this ongoing debate to a certain extent, this section is devoted to analytically comparing three of the greatest NBA players, who come up in almost every GOAT discussion. These players are non-other than Kareem Abdul-Jabbar, Michael Jordan and LeBron James. Not only were these players considered to be the best players of their generation, they also presented impressive stats on a consistent basis, making them the perfect candidates to study.

4.1 Comparison of basic Career Statistics

A good starting point is to look at the career per game averages of the most important basic statistics, both for the regular season and playoffs. Although it is clear that a player's performance does not revolve entirely around these stats, they can be seen as a reliable indicator regarding individual offensive and defensive strengths.



Player	GP	PTS	FG%	2P%	3P%	FT%	AST	OREB	DREB	TREB	STL	BLK	TO	PF
Kareem Abdul-Jabbar	1560	24.6	0.56	0.56	/	0.721	3.6	2.4	7.6	10	0.9	2.6	2.7	3
Michael Jordan	1072	30.1	0.497	0.51	0.327	0.835	5.3	1.6	4.7	6.3	2.3	0.8	2.7	2.6
LeBron James	1421	27.2	0.505	0.554	0.345	0.735	7.3	1.2	6.3	7.5	1.5	0.8	3.5	1.8

Figure 3: Career regular season per game stats

While all three players were dominant in most areas, there are some subtle differences that can be observed when comparing these career regular season stats. First of all, Abdul-Jabbar and James have featured in more games, which is mainly due to them both having played for over 20 seasons, while Jordan's playing career lasted 15 seasons. Abdul-Jabbar even holds the record for most minutes played with 57,446, while James is 3rd with 54,092. Concurrently, Jordan is regarded by many as the most prolific scorer in NBA history, not only because of his outstanding career average of 30.1 PTS/G, but also thanks to having won a record-breaking 10 scoring titles, whereas Abdul-Jabbar and James both have 2 and 1 respectively. Meanwhile, the latter were able to show the longevity of their dominance, by both having held the all-time most points scored record (Abdul-Jabbar with 38,387 until James surpassed this in 2023). The natural height advantage of Abdul-Jabbar (2.18m) compared to the rest of the league, allowed him to repeatedly post good numbers in the rebounds and blocks categories, which is a common trend for players at his position (center). Furthermore, most of his field goal attempts came from very close range, where he mainly used his well-known signature move, the "skyhook". This allowed him to maintain an incredible FG% of 56%, making him the 8th most accurate scorer in NBA history. Moreover, Jordan and James played at the shooting guard and small forward positions, where players are expected to be more efficient from distance shooting. On the one hand, Jordan was a very clutch player who thrived in high pressure situations, as was the case for his incredibly high FT% of 83.5%. On the other hand, NBA teams have grown fonder of outside shots, which has translated to an increased focus on these types of shots, as can be seen for James' impressive 3P% of 34.5%. Out of the three players, James seems to be the one who created the most scoring chances for his teammates, which is underlined by the fact that he ranks 4th in all-time assists, with 10,420. At the same time, he also tended to turn the ball over more frequently, which is not surprising, since a linear correlation can generally be observed between assists and turnovers (see Section 7). Finally, Jordan was known for his gritty defence, which lead to him winning the 1988 defensive player of the year award and getting a significant amount of steals (4th of all-time in steals with 2,514).

Graphically, the profiles of these three players can be analysed with a so-called *Radial Plot*. Ideal for comparing multivariate data, each variable is placed on a different axis and its numeric value is represented by the distance to a common centre point. Next, the plotted values of the variables are joined, creating a type of polygon. The more symmetrical this shape is, the more evenly distributed a player's strengths are. Since all of the axes have the same scale, it is important to choose the specific variables carefully, otherwise some variables might be undersized in the final plot, making it hard to interpret differences between players. A way to work around this is to simply standardise the data, which leads to all the variables having the same scale. Similar to before, this normalisation can be done using the list of the 75 players from the *NBA 75th Anniversary Team*. It is also noteworthy that the blue circular line corresponds to the zero value of each standardised variable, simply representing its actual average.

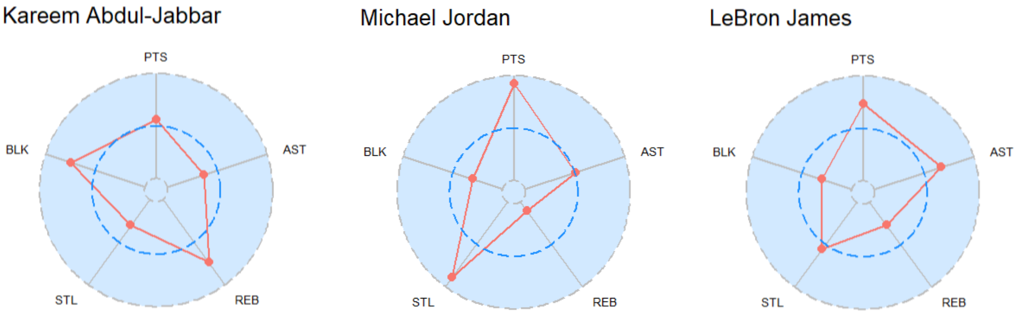
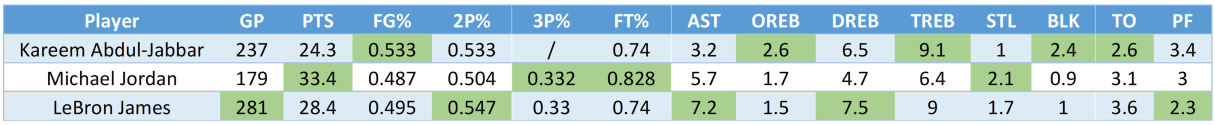


Figure 4: Standardised radial plots representing basic career stats

The three radial plots tend to confirm the previous analysis. On the one hand, Jordan and James rank above average when it comes to points, assists and steals. Jordan even seems to be among the best and most consistent scorers and defenders, since the points for these variables are close to the maximum. On the other hand, Abdul-Jabbar surpasses the norm in the rebounds and blocks categories, clearly illustrating his inside superiority.

While the regular season stats are a good indicator of consistency, the postseason stats show how well a player can perform when it matters most. Therefore, it is interesting to compare these to the regular season averages, to see if there is any change in efficiency when playing with the threat of elimination.



Player	GP	PTS	FG%	2P%	3P%	FT%	AST	OREB	DREB	TREB	STL	BLK	TO	PF
Kareem Abdul-Jabbar	237	24.3	0.533	0.533	/	0.74	3.2	2.6	6.5	9.1	1	2.4	2.6	3.4
Michael Jordan	179	33.4	0.487	0.504	0.332	0.828	5.7	1.7	4.7	6.4	2.1	0.9	3.1	3
LeBron James	281	28.4	0.495	0.547	0.33	0.74	7.2	1.5	7.5	9	1.7	1	3.6	2.3

Figure 5: Career playoffs per game stats

A brief comparison of both tables leads to the conclusion that all three players were able to maintain the same level of consistency throughout the playoffs as for the regular season. Although there is a smaller sample set of games to measure from, it is no coincidence that these three players make up the top three postseason points scorers of all-time (James 1st, Jordan 2nd & Abdul-Jabbar 3rd). In particular, Jordan holds the single game playoffs record for points scored with 63, which isn't surprising, since his scoring ability against the toughest opponents was head and shoulders above the rest of the league. In addition, at least one of them is represented in the top five of all-time playoffs assists (James 2nd), rebounds (James 4th), steals (James 1st & Jordan 3rd) and blocks (Abdul-Jabbar 2nd). Furthermore, James is also closing in on the record for the most postseason triple-doubles (i.e. when a player achieves double-digit values in three positive statistical categories) with 28, which highlights his ability to impact a game in multiple ways.

To simplify the visualisation of the shooting percentages of the three players, one can illustrate them in a bubble plot containing all 75 players from the above-mentioned list. A bubble plot is a 2-dimensional plot where each player is represented by a bubble in the plane. The x-axis, y-axis, bubble colour and size all correspond to one of four variables (here 2P%, 3P%, FT% & total shots attempted respectively). It can also be advantageous to rescale the bubble sizes between 0 and 100, to make the final plot easier to interpret.

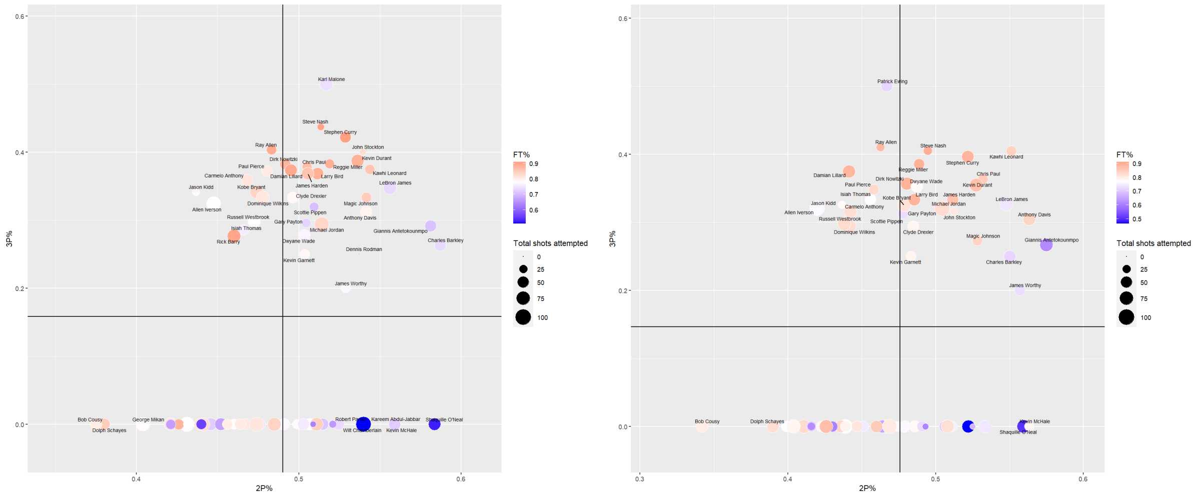


Figure 6: Bubble plots of career shooting percentages in the regular season & playoffs

A first observation that can be made is that all three players shot an above average percentage from 2-point distance, with Abdul-Jabbar and James ranking among the very best. Simultaneously, they seem to be among the players taking the most shots per game, which indicates the individual responsibility they took when it comes to scoring. In general, the smaller the distance to the upper right corner, the more efficient a shooter is. Therefore, James can be regarded as one of the most consistent players, throughout both the regular season and playoffs. It is also noticeable that many players have a 3P% equal to 0. While some players like Shaquille O'Neal only rarely took 3-pointers and hardly ever made them, the most common reason is due to the belated introduction of the 3-point line in 1979. Concerning the FT%, Jordan's percentage was slightly above the standard, while Abdul-Jabbar's and James' were close to the list's average. Overall, one can notice that there are no major differences in the shooting percentages between the regular season and postseason.

(Note: Since James is still an active NBA player at the time of writing, the mentioned stats and records are subject to change, as he continues to strive to be the best player of all-time.)

4.2 In-depth Analysis of individual Strengths

While the basic stats offer a solid foundation for the GOAT debate, it is clear that they are insufficient when it comes to comparing intergenerational players, since they don't tell the full story. Therefore, it may be interesting to analyse the longevity, success and other factors of each player's career.

First of all, longevity is a crucial argument when debating about the greatest player of all-time, since it shows how long a player was able to compete at the highest level. In this sense, a consistent, high-performing career is valued more than a brief, high-peak one. Therefore, one can analyse the PER of a player based on his age, while also taking into account the total number of games played.

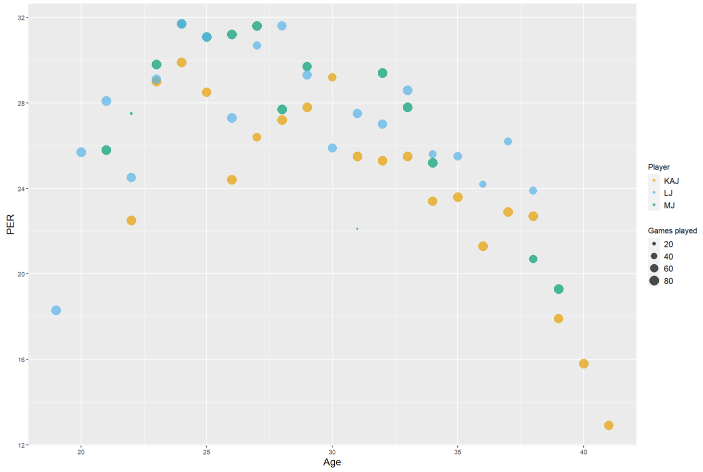


Figure 7: Bubble plot of PER with respect to age

A first look at the plot reveals that both Jordan and James were able to steadily improve their PER up until a certain age (approximately 27-28), before slowly declining towards the tail end of their career. Meanwhile, Abdul-Jabbar's PER tended to fluctuate in his 20s, before decreasing in his 30s. Overall, the evolution of the PER of all three players is remarkable and underlines phenomenal careers, since it is very common for basketball players to be in their prime between the ages of 25-30 (Kalén et al., 2020)[4]. Interestingly, Jordan and James both recorded their highest PER at the age of 24, with a value of 31.7 and were able to outperform Abdul-Jabbar at almost every age. In addition, Jordan still holds the all-time highest career average PER with 27.9. As regards to the number of games played each season, Abdul-Jabbar and Jordan featured in almost all possible games of each season, with the exception of a few outliers due to injury (Abdul-Jabbar in 1977 & Jordan in 1985) and a comeback out of retirement (Jordan in 1994). On the other hand, James' games and minutes played have decreased significantly over the past few years, an unfortunate trend among the majority of NBA teams nowadays, in an attempt to rest their star players for the playoffs.

Another determining factor of a player's greatness is their individual and team success throughout their career. While there are many smaller awards to be won, here are some of the key achievements a player can reach (in descending order of importance according to personal opinion): championships, MVPs, Finals MVPs, scoring titles, All-NBA selections, All-Star selections and game winners. In this case, a bar chart can be used to visually identify the differences between the players in each category.

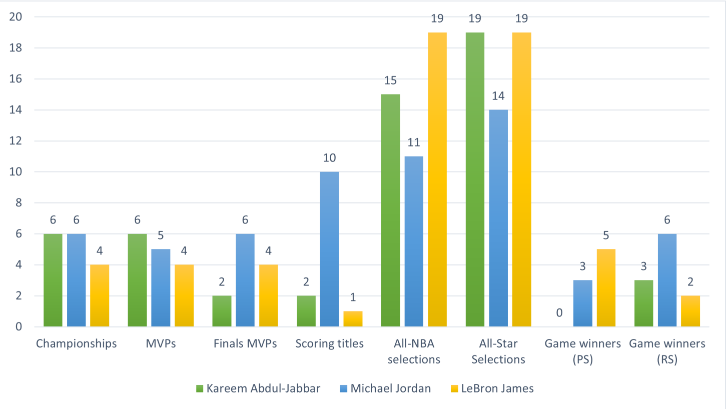


Figure 8: Bar chart with career achievements

Regarding the number of championships won, Abdul-Jabbar and Jordan have both enjoyed an equal amount of team success, while James continues to closely chase them. Next, the MVP and Finals MVP awards highlight the individual dominance of a player, the latter given to outstanding performances when it matters most. While all three players have won a similar amount of MVPs, it is noteworthy that Jordan won the Finals MVP in each of his 6 finals appearances, making him and Shaquille O'Neal the only players in history to hold a 100% record. Moreover, Jordan is also head and shoulders above any other player in the scoring category, having won the scoring title a record 10 times. Abdul-Jabbar and James have featured in the most All-NBA and All-Star selections, both holding the all-time records in these categories. Here, it is important to consider the fact that Jordan played significantly fewer seasons (15) than Abdul-Jabbar (21) and James (20), which has a certain impact on these numbers. Finally, Jordan and James have been able to win the game for their team on several occasions throughout their career, in contrast to Abdul-Jabbar, who only scored a total of 3 game winners. It is essential to distinguish the type of game these game winners are scored in, since the playoffs have a greater significance, giving James a slight edge in terms of being clutch when it matters most. Overall, Jordan currently leads the way in total game winners made with 9. It is important to keep all these numbers in mind, as they play a vital role in the computation of the PPI (see Section 4.3).

In data analysis, it can occasionally be of interest to compare players according to multiple distinct variables at once. This is known as *Multivariate Data Analysis* and was previously already used for the construction of the PPI, using the PCA method. Another frequently used dimensionality reduction tool is the *Multidimensional Scaling* (MDS) method, which aims to find a low-dimensional representation of a dataset. This can lead to a coloured 2-dimensional plot, allowing us to visually determine the pairwise (dis)similarities between players. Similar to the PCA algorithm, the first step of MDS consists of defining a matrix $\Delta^n = (d_{ij})_{i,j \in \mathbb{N}}$, known as the dissimilarity matrix. Here,

$$d_{ij} := \text{distance between } i^{th} \text{ and } j^{th} \text{ player.}$$

While there are many distance metrics to choose from, we will be working with the Euclidean distance, hence

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \tag{15}$$

where x_{ik} and x_{jk} are the values taken by the variable X_k for the players i and j. The main goal of MDS is then to reduce every player to an m-dimensional variable space ($m \ll n$), in a way that the newly obtained dissimilarity matrix Δ^m matches Δ^n as closely as possible. The first step of this reduction is done by computing the *Gram matrix*, a symmetric matrix representing the pairwise (dis)similarities in a lower-dimensional space. It is obtained by squaring each entry of the matrix Δ^n, followed by a centering operation. This operation consists of subtracting the row and columns means from every element, before adding the overall mean to every element. This is an important procedure, as it guarantees that the matrix is centered and has 0 mean, a necessary requirement for the subsequent calculation of the eigenvectors and eigenvalues. Just like for the PCA algorithm, it is then generally of interest to determine the new number of dimensions that should be used for our final dissimilarity matrix Δ^m. While the scree plot method is a valid method for MDS (as for PCA), we will make use of the stress index S, which measures in percentage how well the matrices Δ^n and Δ^m fit. As a general rule, one can remember that the value of the stress index should be close to 0 and should preferably not exceed 20%. After having found the dimension m which minimises the stress index, one can finally compute the matrix Δ^m, which is composed of the m most significant eigenvectors of the Gram matrix. Last but not least, the data can be plotted if $m \in \{1, 2, 3\}$ and can be rotated or reflected to improve readability, since the outcome is rotation and reflection invariant.

In this particular case, we will be using the R function `MDSmap`, for which the dimension ($m = 2$) is already fixed beforehand. More specifically, it is based on Kruskal's non-metric approach, for which the stress index is calculated by the following formula:

$$S = \sqrt{\frac{\sum_{i,j}^{n}(\delta_{ij} - d_{ij})^2}{\sum_{i,j}^{n} d_{ij}^2}} \tag{16}$$

where δ_{ij} is the distance between player i and j in the lower-dimensional space. In our case, the variables will be the career totals of PTS, AST, REB, STL and BLK, applied to the same list of 75 players.

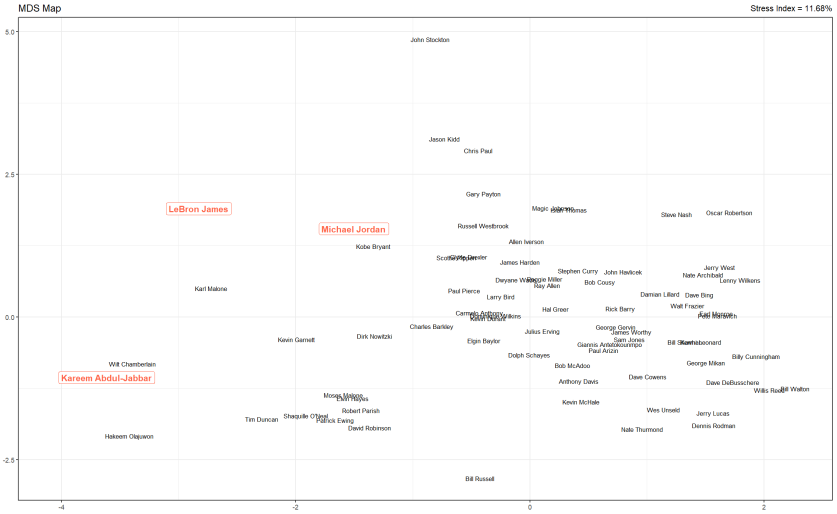


Figure 9: 2-dimensional MDS plot of 75 players

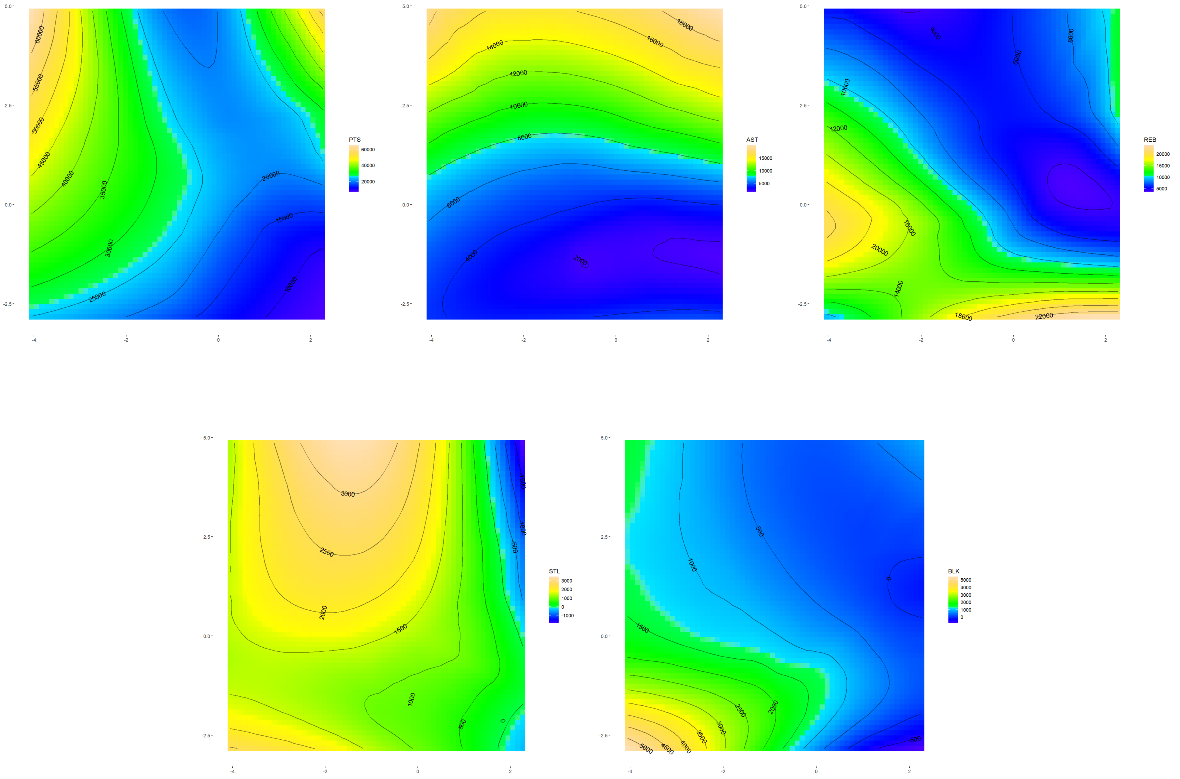
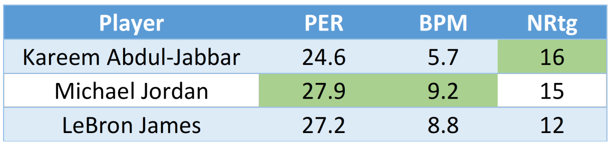


Figure 10: Level plots based on career totals PTS, AST, REB, STL & BLK respectively

The obtained stress index S has a value of 11.68%, validating the choice of 2 dimensions. In the first plot, players are placed based on their similarity level, i.e. the closer they are, the more similar features they share. It is immediately noticeable that there is a large group of players clustered around the origin (0,0) of the plot, while they become more separated as you move further away. Furthermore, one can observe that all three players are so-called *outliers*, meaning that their career totals are extreme values compared to those of the other players on the list. However, this is no surprise, as the previous analysis in Section 4.1 mentions that all three players are among the leaders for most of the basic career stat lines. While the 2D plot on its own can already help understand if a player's performances were irregular, it doesn't necessarily indicate in which way it differs from the rest. Therefore, it can be beneficial to colour the player points according to one of the several variables. As a consequence, one can consider the level plot, which resembles a topographic map. It is obtained by fitting the chosen variable with a surface, which is determined with a method of polynomial local regression, and it has the same axes as the 2D MDS plot. By considering all of the level plots at once, one can get a better understanding of the meaning of a player's position. Overall, the career totals tend to increase from right to left, which visually confirms the previous assessments of Abdul-Jabbar's, Jordan's and James' career achievements. It is however recognisable that Abdul-Jabbar and James are situated further to the left than Jordan, which may seem paradoxical at first. This is largely due to their longer careers, which allowed them to accumulate better total stats. Therefore, they seem to have the edge when it comes to the longevity of their career.

4.3 Comparison of advanced Metrics

A final way to evaluate a player's career is by rating their performance according to an advanced metric, as seen in Section 2.2. Let us first take a look at the advanced metrics PER, BPM and NRtg.



Player	PER	BPM	NRtg
Kareem Abdul-Jabbar	24.6	5.7	16
Michael Jordan	27.9	9.2	15
LeBron James	27.2	8.8	12

Figure 11: Career averages of PER, BPM & NRtg

Firstly, Jordan and James seem to have been the most offensively efficient players out of the three, both having an outstanding career average PER. They also lead the way regarding the BPM, suggesting that they had a greater impact on the team's success when they were on the court. Finally, Abdul-Jabbar's and Jordan's high NRtg show that they were dominant players on both ends of the court, making important offensive and defensive contributions for the team.

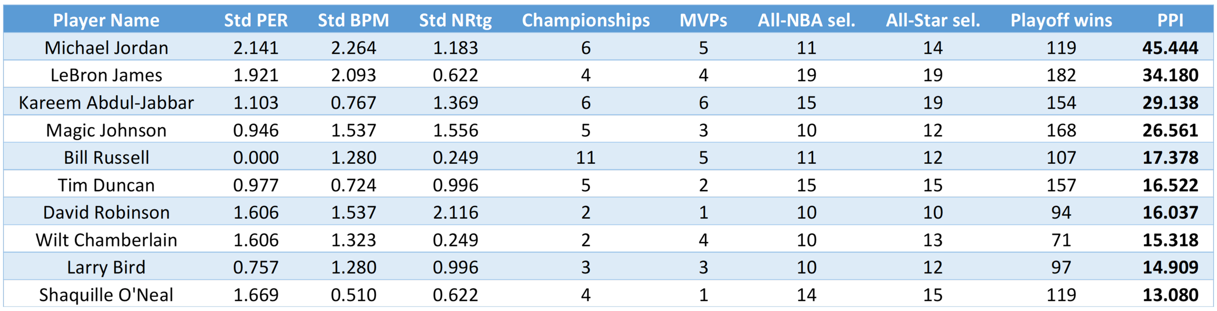
The values of these advanced metrics, together with the previously seen career achievements, can be used to evaluate the players according to the self-developed PPI metric.

$$\begin{aligned} PPI_{KAJ} =& (0.573 \cdot 1.103 + 0.624 \cdot 0.767 + 0.532 \cdot 1.369) \cdot \left(6 + 6 + \frac{15}{15} + \frac{19}{24} + \frac{154}{75}\right) \\ &= 29.138 \end{aligned}$$

$$\begin{aligned} PPI_{MJ} =& (0.573 \cdot 2.141 + 0.624 \cdot 2.264 + 0.532 \cdot 1.183) \cdot \left(6 + 5 + \frac{11}{15} + \frac{14}{24} + \frac{119}{75}\right) \\ &= 45.444 \end{aligned}$$

$$\begin{aligned} PPI_{LJ} =& (0.573 \cdot 1.921 + 0.624 \cdot 2.093 + 0.532 \cdot 0.622) \cdot \left(4 + 4 + \frac{19}{15} + \frac{19}{24} + \frac{182}{75}\right) \\ &= 34.180 \end{aligned}$$

Based on this index alone, Jordan seems to have had the most successful career out of the three, with James coming in at 2nd and Abdul-Jabbar at 3rd. This evaluation appears to be fair, since Jordan delivered strong analytical reports in most of the analysed categories, especially regarding the number of championships he won (arguably the most important factor in the GOAT debate) and his remarkable scoring ability which is still unmatched to this day. The only aspect in which he is inferior to the other two is in terms of longevity and some career total stats, which can be compensated by the incredible consistency he had at his peak, with his 100% win record in the finals. That being said, since James is currently still an active NBA player, this evaluation could possibly still change in the future, making James' last years of his career all the more exciting. One can also take a look at the top 10 players of all-time based on the PPI.



Player Name	Std PER	Std BPM	Std NRtg	Championships	MVPs	All-NBA sel.	All-Star sel.	Playoff wins	PPI
Michael Jordan	2.141	2.264	1.183	6	5	11	14	119	**45.444**
LeBron James	1.921	2.093	0.622	4	4	19	19	182	**34.180**
Kareem Abdul-Jabbar	1.103	0.767	1.369	6	6	15	19	154	**29.138**
Magic Johnson	0.946	1.537	1.556	5	3	10	12	168	**26.561**
Bill Russell	0.000	1.280	0.249	11	5	11	12	107	**17.378**
Tim Duncan	0.977	0.724	0.996	5	2	15	15	157	**16.522**
David Robinson	1.606	1.537	2.116	2	1	10	10	94	**16.037**
Wilt Chamberlain	1.606	1.323	0.249	2	4	10	13	71	**15.318**
Larry Bird	0.757	1.280	0.996	3	3	10	12	97	**14.909**
Shaquille O'Neal	1.669	0.510	0.622	4	1	14	15	119	**13.080**

Figure 12: Ordered list of top 10 players of all-time according to PPI

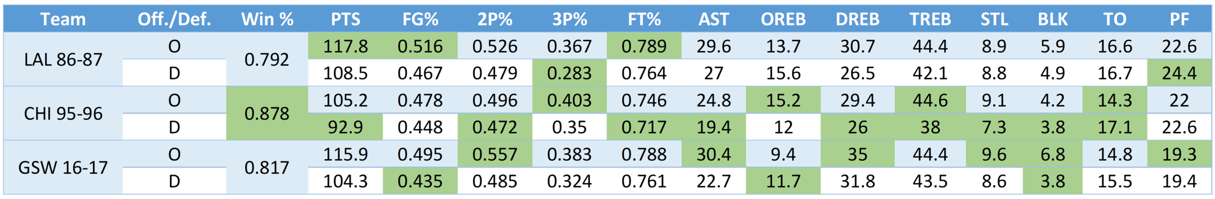
According to this table, it was the right decision to analyse the three chosen players, since they make up the top three. Ultimately, the proposed PPI metric offers multiple strengths. By incorporating several advanced metrics, as well as career accolades in a unique way, a player is evaluated based on diverse factors, each considering a different trait of a player's career success. Moreover, thanks to PCA, one can consider a linear combination of some of the most respected advanced metrics, while limiting the total redundancy. In addition, the weights of the career achievements are attributed according to a valid reasoning, valuing certain achievements more than others. Finally, thanks to the transparency and comprehensiveness of the metric, it is relatively easy to compute and understand the PPI. It is however important to emphasize that there is no single metric that is "correct" and can fully determine the GOAT, since every metric has its up- and downsides. All in all, the GOAT debate will continue to be an open question, discussed by fans all around the world.

5 Greatest Team of all-time

A natural extension of the GOAT debate among players is to analytically determine which single season team was the greatest of all-time so far. In particular, three of the most outstanding teams will be analysed in closer detail: the 1986-1987 Los Angeles Lakers (LAL), the 1995-1996 Chicago Bulls (CHI) and the 2016-2017 Golden State Warriors (GSW). Considered by many as the greatest teams to have ever played and having each won the championship in their respective season, they form the perfect trio to conduct our analysis on. (Note: For simplicity reasons, these single season teams will be referred to as the Lakers, the Bulls and the Warriors.)

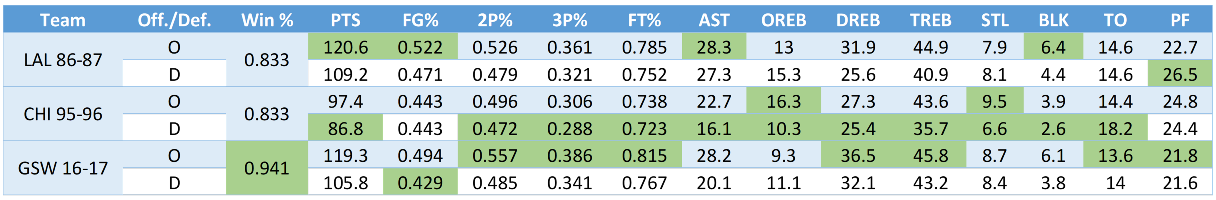
5.1 Comparison of basic Team Statistics

Once again, it is useful to take a look at the basic team statistics throughout the regular season and postseason, to get a better understanding of the strengths of each team. Unlike for players, for teams it is easier to assess both their offensive (O) and defensive (D) performances, offering an additional perspective to the analysis. Here, the defensive stats simply refer to the average performance of their opponents.



Team	Off./Def.	Win %	PTS	FG%	2P%	3P%	FT%	AST	OREB	DREB	TREB	STL	BLK	TO	PF
LAL 86-87	O	0.792	117.8	0.516	0.526	0.367	0.789	29.6	13.7	30.7	44.4	8.9	5.9	16.6	22.6
	D		108.5	0.467	0.479	0.283	0.764	27	15.6	26.5	42.1	8.8	4.9	16.7	24.4
CHI 95-96	O	0.878	105.2	0.478	0.496	0.403	0.746	24.8	15.2	29.4	44.6	9.1	4.2	14.3	22
	D		92.9	0.448	0.472	0.35	0.717	19.4	12	26	38	7.3	3.8	17.1	22.6
GSW 16-17	O	0.817	115.9	0.495	0.557	0.383	0.788	30.4	9.4	35	44.4	9.6	6.8	14.8	19.3
	D		104.3	0.435	0.485	0.324	0.761	22.7	11.7	31.8	43.5	8.6	3.8	15.5	19.4

Figure 13: Regular season team per game stats



Team	Off./Def.	Win %	PTS	FG%	2P%	3P%	FT%	AST	OREB	DREB	TREB	STL	BLK	TO	PF
LAL 86-87	O	0.833	120.6	0.522	0.526	0.361	0.785	28.3	13	31.9	44.9	7.9	6.4	14.6	22.7
	D		109.2	0.471	0.479	0.321	0.752	27.3	15.3	25.6	40.9	8.1	4.4	14.6	26.5
CHI 95-96	O	0.833	97.4	0.443	0.496	0.306	0.738	22.7	16.3	27.3	43.6	9.5	3.9	14.4	24.8
	D		86.8	0.443	0.472	0.288	0.723	16.1	10.3	25.4	35.7	6.6	2.6	18.2	24.4
GSW 16-17	O	0.941	119.3	0.494	0.557	0.386	0.815	28.2	9.3	36.5	45.8	8.7	6.1	13.6	21.8
	D		105.8	0.429	0.485	0.341	0.767	20.1	11.1	32.1	43.2	8.4	3.8	14	21.6

Figure 14: Playoffs team per game stats

First of all, one can notice that all three teams had a high winning percentage during the regular season, which remained consistent throughout the playoffs. The Bulls won an unprecedented 72 out of 82 regular season games, while the Warriors remarkably only lost a single game in the postseason. When it comes to scoring, the Lakers and Warriors delivered exceptional offensive performances, which they even improved on during the playoffs. The Lakers even hold the all-time record for points per game in the playoffs among the teams having reached the finals with 120.6. However, defensively, the Bulls were able to keep their opponents to a far lower average point total, primarily because they had three NBA All-Defensive players in Dennis Rodman, Scottie Pippen and Michael Jordan. Overall, this may suggest that the league-wide scoring average was lower in 1995-1996 than in other years, a trend that will be analysed in Section 6. As regards to the regular season shooting percentages, each team seemed to have their own strengths and weaknesses. Meanwhile, the postseason brought the best out of the Warriors, which lead them to the best 2P%, 3P% and FT% among the three teams, this largely thanks to the phenomenal shooting "splash brothers" Stephen Curry and Klay Thompson, as well as the 4 time scoring champion Kevin Durant. This underlines the dominance they showcased during the playoffs, which goes hand in hand with their winning percentage. On the other hand, the Bulls once again emphasized the importance of their defence, holding their opponents to very low shooting percentages. Similar to the scoring category, the Lakers and Warriors also had a high number of assists per game, showing how well they played collectively. Concurrently, the Bulls had significantly fewer assists, which could be a sign of the team relying more heavily on individual performances (e.g. by Michael Jordan), rather than team efforts. All three teams performed equally as efficient with respect to rebounding, the Bulls being especially dominant on the offensive end, the Warriors grabbing the most defensive boards and the Lakers having a great balance between the two, in large thanks to their star center Kareem Abdul-Jabbar. Another crucial factor of a team's defensive ability is the number of steals and blocks they can get in a game. In this case, the Bulls and the Warriors were able to force a lot of turnovers by repeatedly stealing the ball, whereas the Lakers and the Warriors were extremely efficient at obstructing the opponent's shots by blocking them. Finally, in terms of personal fouls per game, the teams of the 20th century tended to commit more fouls than the Warriors did, hinting to a less aggressive and physical playstyle nowadays.

5.2 Individual Performance Distribution among Teams

After having looked at the overall collective performance of the teams, it is equally as important to study the individual performances. Since basketball is a team sport, it is crucial for multiple players to contribute, without relying too heavily on a one-man-show by their best player. Therefore, it can be interesting to analyse the variability and inequality regarding the shooting and scoring strengths within each team.

In the field of Statistics, variability can be summarised as the extent to which data points in a dataset vary from one another. In other words, variability simply reveals how spread out the data are. There are many different nonnegative indices that can be used to measure the variability of a random variable X, whose values are generally close to 0 if all the data are equal. In this study case, the selected indices are none other than the *Range* (difference between maximal and minimal value in a dataset) and the *Variation Coefficient* (VC), which is simply the ratio of the standard deviation to the mean:

$$VC = \frac{\sigma_X}{|\mu_X|} \tag{17}$$

where N is the number of values x_i ($i = 1, ..., N$) that the variable X takes, μ_X is the mean of these values and

$$\sigma_X^2 = \frac{\sum_{i=1}^{N}(x_i - \mu_X)^2}{N}. \tag{18}$$

The normalised index VC is especially useful when studying the variability of multiple variables with different units of measurement or different means. One can add an additional layer of complexity by weighting the average μ_X and the standard deviation σ_X in a specific way. The weighted variation coefficient (WVC) is particularly useful when comparing the variability of a dataset, composed of variables with different sample sizes and means. The basic computation of the variation coefficient can occasionally be inappropriate, as it gives more weight to data subsets with higher variances. The formulas for the weighted mean and the weighted standard deviation are

$$w\mu_X = \frac{\sum x_i \cdot w_i}{\sum w_i} \tag{19}$$

$$w\sigma_X = \sqrt{\frac{\sum(x_i - w\mu_X)^2 \cdot w_i}{\sum w_i}}. \tag{20}$$

Here, every single observation is weighted differently, based on the values of the weighting variable.

The variability is a useful tool when analysing the individual performances within a team, since it reveals whether it is well-balanced or not. On the one hand, high variability with variables such as points, assists and rebounds, may show that multiple players are making the most of their individual strengths. On the other hand, when it comes to efficiency variables, like shooting percentages, high variability can suggest that a few players are carrying the whole team, while others are performing below the team standards. It is this second category of stats that will be further analysed for all three teams, to see if they functioned as well-oiled machines. In particular, the shooting percentages (2P%, 3P% & FT%) of the players having played more than 500 minutes during the regular season will be plotted in a variability diagram, where each player is represented along a vertical axis by a bubble, of size according to the total number of shots taken. Moreover, the total number of shots attempted for each shot type, is considered as the weighting variable, so that the obtained bubbles are of proportional size.

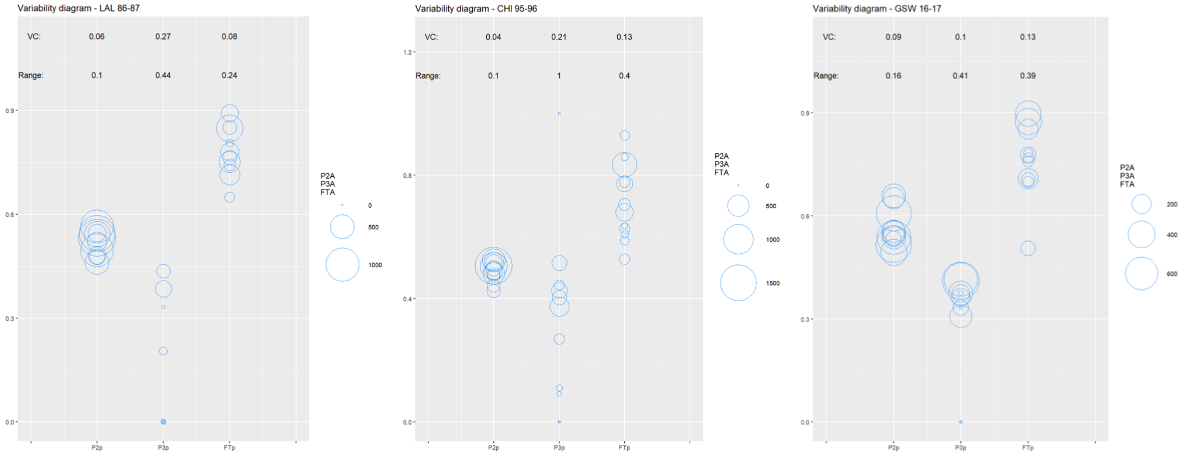


Figure 15: Variability diagrams of individual shooting percentages during regular season

The Lakers and the Bulls present the lowest variation coefficient and range regarding shots from 2-point distance, which is visible due to the formed clusters in the respective diagrams. Even though most of the Warriors players boast an impressive 2P%, they have a higher variability in this category compared to the other two teams. The exact opposite can be said for 3-point shooting, since the Warriors seem to be a well-balanced team from distance, while the variation coefficient and range of the Lakers and the Bulls is relatively high in the 3P% class, mainly due to the 3-point shot not having the same importance then as it does nowadays. It is noteworthy that the maxed-out range of the Bulls is largely due to single players taking only very few shots and then either making or missing them. In addition, for both 2P% and 3P%, it seems like the players taking the majority of the shots were also the most efficient. This validates the idea that players who take more shots tend to be more in rhythm, allowing them to shoot a better percentage in the long run. Finally, the Lakers present the lowest variability with respect to FT%, whereas the percentages of the Bulls and the Warriors were slightly more widespread. Overall, the team with the lowest summed variation coefficients is the Warriors, with a total of 0.32. This observation seems to match the general trend of the NBA over the past few years, where players have been required to become more diverse and have multiple strengths, in particular an efficient shooting ability.

Originating from the field of economics, inequality analysis is used to measure the distribution of income or wealth in a country. It ranges from equal distribution (wealth is distributed evenly) to maximal inequality (one individual holds all the wealth, while all the others are left with nothing). However, neither one of these extremes is attained in practice. The Gini index and the Lorenz curve are tools that are used to identify the extent of the distribution analytically and visually. The Gini coefficient ranges from 0 (perfect equality) to 1 (maximal inequality) and is defined as the ratio of the area between the Lorenz curve and the line of perfect equality, to the total area below the perfect equality line. The Lorenz curve is obtained by plotting the cumulative proportion of wealth held by each percentile of the population, with respect to the cumulative proportion of the population. Generally speaking, the Gini index measures the relative difference between the actual wealth distribution and the perfectly equal distribution.

In the case of basketball, inequality analysis can be conducted on variables, such as points scored by a team. Analogously, the wealth can symbolically be replaced by points, revealing whether a team is well-balanced in terms of scoring or not. For the analysis, the 8 players having scored the most points throughout the regular season are chosen and are listed in a table in increasing order of total points scored. The main reason why only a restricted amount of players are chosen, is because the inclusion of players who hardly play would inflate the observed inequality, leading to misinterpreted results. The next step consists of combining the number of players and the total points scored row for row, giving us the cumulative values Cpl_i and $Cpts_i$ for the player in the i^{th} row. After that, these cumulative values are divided by the total number of players and points scored, to obtain the cumulative percentages $Cpl_i\%$ and $Cpts_i\%$. Here, $Cpts_i\%$ simply indicates the proportion of points scored by the first $Cpl_i\%$ players of the list. The Lorenz curve is then obtained by joining the points plotted at $(Cpl_i\%, Cpts_i\%)$. As previously explained, the Gini coefficient G is non-other than the following ratio:

$$G = \frac{\sum_{i=1}^{N}(Cpl_i\% - Cpts_i\%)}{\sum_{i=1}^{N-1} Cpl_i\%} \tag{21}$$

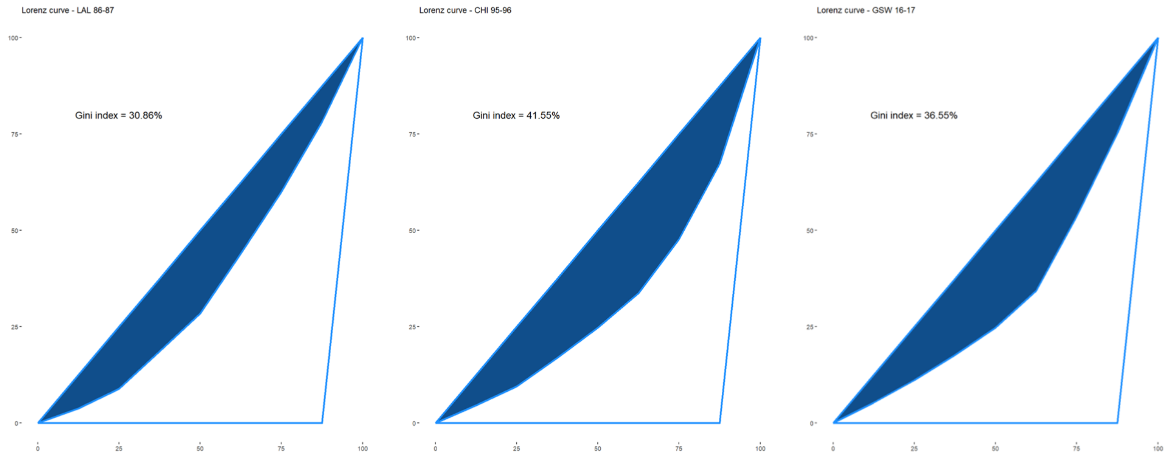


Figure 16: Lorenz curve for points scored during regular season

One can immediately notice that the Gini index of the Bulls is the greatest with a value of 41.55%, suggesting that very few players scored the majority of the points. This turns out to be very true, since Michael Jordan and Scottie Pippen were responsible for a staggering 52.29% of the scoring during the regular season. This sort of reliance on only a handful of players can be detrimental for a team when foul trouble or injury problems strike. The Warriors have a Gini index valued at 36.55%, which is still relatively high, largely due to the fact that they had three of the best scorers of the last decade on their roster. Kevin Durant, Klay Thompson and Stephen Curry together scored 65.62% of the total points in the regular season. Finally, the Lakers seemed to be a more evenly balanced offensive team, posting a value of 30.86%. This is in large part due to their star playmaker Magic Johnson, who tended to create countless scoring opportunities for his teammates, while scoring a healthy amount of points himself. Overall, out of the three teams, the Lakers were the ones that played the best team basketball scoring wise, without relying too heavily on a certain individual.

5.3 Comparison of advanced Metrics

The last step of the all-time greatest teams comparison consists of evaluating them according to multiple advanced metrics, all of them defined in Section 2.2. Let us start off by taking a look at the Offensive/Defensive Rating, as well as the Four Factors of each team.

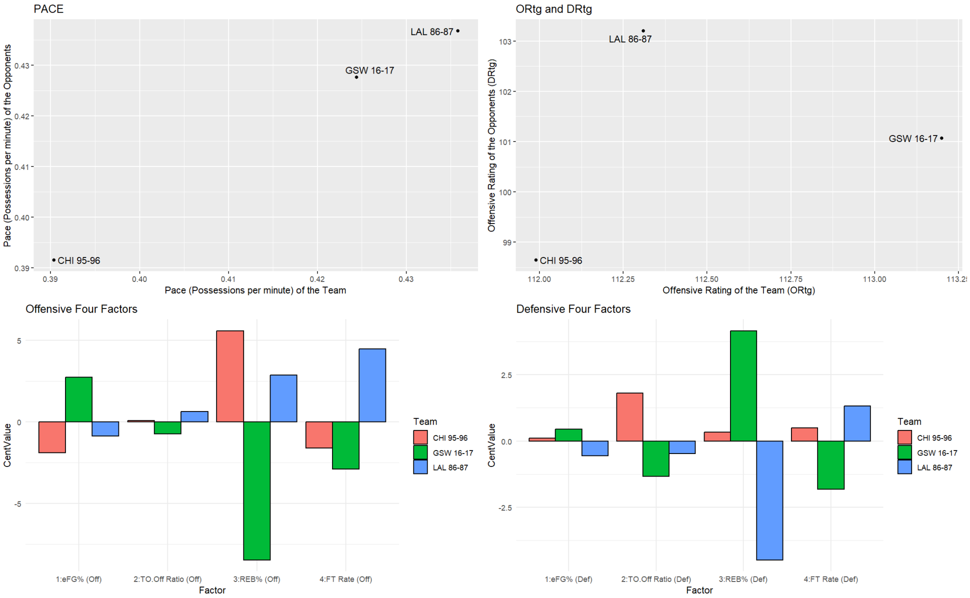


Figure 17: Pace, Offensive/Defensive Rating & Four Factors of each team

The upper left graph presents the pace (possessions per minute) of the opponents with respect to the pace of the teams themselves. The Lakers and the Warriors recorded a very similar pace in the regular season, while the Bulls adopted a significantly slower playstyle. This matches the previous observation that the Bulls scored far fewer points than the other two teams, while also conceding fewer points. In general, one can notice that the own pace was almost matched by the pace of the opponent, possibly implying that the pace of a team depended largely on the league-wide average pace. This alternating trend will be further analysed in Section 6.

The upper right graph shows both the Offensive and Defensive Ratings of the three teams. This is different to the average points scored per game since it is based on the scoring abilities per 100 possessions. In other words, the pace at which the game is played is neglected, providing us with a better metric of how efficient a team was offensively and defensively. Regarding the ORtg, the Warriors have a slight edge as opposed to the Lakers and the Bulls, averaging almost 1 point more per 100 possessions. On the other hand, the Bulls were clearly the more dominant force defensively, being the only team of the three to hold their opponents to below 100 points per 100 possessions. Overall, this graph removes the general misconception that the Bulls were not an efficient scoring team, since they almost coincided with the other teams over 100 possessions.

The two bar plots represent the difference between the offensive and defensive Four Factors of each team and the average of each factor. The value and the size of this difference helps interpret the relative strengths and weaknesses of the three teams. First of all, the Warriors have a slightly better offensive effective field goal percentage than the other two teams, which matches the previous variability analysis done in Section 5.2. Meanwhile, they were all able to keep the opponent's shooting efficiency at a same level. Despite all three teams having a very similar offensive turnover ratio, the Bulls were able to force the opponent to commit more turnovers per 100 possessions. Once again, this underlines how aggressive and committed their defensive playstyle was. The biggest differences can be seen in the rebounding category. While the Warriors were far less efficient at grabbing available rebounds on the offensive end, they showed an impressive dominance in defensive rebounding. This overlaps with the previous assessment in Section 5.1. Finally, the Lakers were able to get to the free throw line the most frequently, by drawing more fouls from the opponent while shooting. At the same time, they were also the team who gave away the most free-throw attempts per field-goal attempts, suggesting that they had to deal with a higher number of sloppy defensive errors. On the other hand, this may have simply been a league-wide trend at the time.

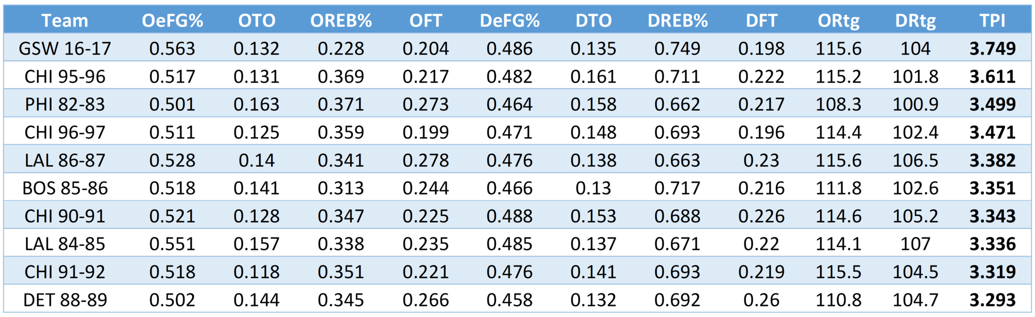
A last step of the comparison consists of computing the self-developed metric TPI for all three teams and seeing how the different teams rank against each other. It is important to remember that the TPI takes several factors into account, such as the Four Factors, the Offensive/Defensive Rating and team achievements.

$$\begin{aligned} TPI_{LAL} =& \frac{0.4 \cdot 0.528 + 0.25 \cdot (1 - 0.14) + 0.2 \cdot 0.341 + 0.15 \cdot 0.278}{0.4 \cdot 0.476 + 0.25 \cdot (1 - 0.138) + 0.2 \cdot (1 - 0.663) + 0.15 \cdot 0.23} \\ & \cdot \frac{115.6}{106.5} \cdot \left(1 + \frac{15}{18} + \frac{3}{15} + \frac{3}{24} + \frac{65}{82}\right) \\ =& \, 3.382 \end{aligned}$$

$$\begin{aligned} TPI_{CHI} =& \frac{0.4 \cdot 0.517 + 0.25 \cdot (1 - 0.131) + 0.2 \cdot 0.369 + 0.15 \cdot 0.217}{0.4 \cdot 0.482 + 0.25 \cdot (1 - 0.161) + 0.2 \cdot (1 - 0.711) + 0.15 \cdot 0.222} \\ & \cdot \frac{115.2}{101.8} \cdot \left(1 + \frac{15}{18} + \frac{2}{15} + \frac{3}{24} + \frac{72}{82}\right) \\ =& \, 3.611 \end{aligned}$$

$$\begin{aligned} TPI_{GSW} =& \frac{0.4 \cdot 0.563 + 0.25 \cdot (1 - 0.132) + 0.2 \cdot 0.228 + 0.15 \cdot 0.204}{0.4 \cdot 0.486 + 0.25 \cdot (1 - 0.135) + 0.2 \cdot (1 - 0.749) + 0.15 \cdot 0.198} \\ & \cdot \frac{115.6}{104} \cdot \left(1 + \frac{16}{17} + \frac{4}{15} + \frac{4}{24} + \frac{67}{82}\right) \\ =& \, 3.749 \end{aligned}$$

According to this index, the 2016-2017 Warriors seem to have been the most dominant single season team in NBA history. This comes as no surprise, as they were a team with a record-breaking 4 All-Stars in Kevin Durant, Stephen Curry, Klay Thompson and Draymond Green. This incredible group of individuals, combined with impressive collective stats, makes up the perfect contender for the greatest team of all-time. An ordered list of the top 10 teams according to the TPI is given down below.



Team	OeFG%	OTO	OREB%	OFT	DeFG%	DTO	DREB%	DFT	ORtg	DRtg	TPI
GSW 16-17	0.563	0.132	0.228	0.204	0.486	0.135	0.749	0.198	115.6	104	**3.749**
CHI 95-96	0.517	0.131	0.369	0.217	0.482	0.161	0.711	0.222	115.2	101.8	**3.611**
PHI 82-83	0.501	0.163	0.371	0.273	0.464	0.158	0.662	0.217	108.3	100.9	**3.499**
CHI 96-97	0.511	0.125	0.359	0.199	0.471	0.148	0.693	0.196	114.4	102.4	**3.471**
LAL 86-87	0.528	0.14	0.341	0.278	0.476	0.138	0.663	0.23	115.6	106.5	**3.382**
BOS 85-86	0.518	0.141	0.313	0.244	0.466	0.13	0.717	0.216	111.8	102.6	**3.351**
CHI 90-91	0.521	0.128	0.347	0.225	0.488	0.153	0.688	0.226	114.6	105.2	**3.343**
LAL 84-85	0.551	0.157	0.338	0.235	0.485	0.137	0.671	0.22	114.1	107	**3.336**
CHI 91-92	0.518	0.118	0.351	0.221	0.476	0.141	0.693	0.219	115.5	104.5	**3.319**
DET 88-89	0.502	0.144	0.345	0.266	0.458	0.132	0.692	0.26	110.8	104.7	**3.293**

Figure 18: Ordered list of top 10 teams of all-time according to TPI

While the 1995-1996 Bulls finish close behind in 2^{nd} place, the 1982-1983 76ers are surprisingly ranked in 3^{rd}. Based on the TPI alone, it turns out that the 1986-1987 Lakers were not among the three best single season teams of all-time. Furthermore, Michael Jordan's Bulls teams appear an impressive four times in the top 10 ranking. This either means that Jordan was fortunate enough to play in some of the best teams to have ever existed, or that he was almost single-handedly able to lead his team to success over and over again. According to the analysis in Section 4, the latter seems to be more likely. The main logic behind the TPI index coincides with the one of the PPI, taking both advanced metrics and team accolades into account to simultaneously value multiple strengths of a team. Generally speaking, this makes it a strong benchmark for comparing intergenerational teams and evaluating the greatest teams of all-time. Once again, it is important to state that every metric has its limitations and that there is no such thing as the "perfect" metric to measure a team's success. All in all, the greatest team of all-time will continue to be up for discussion for many years to come.

6 Identifying the different Eras in NBA History

Having covered some controversial discussion topics, the next section deals with something that most fans can agree on: the game in today's age is vastly different from how it was decades ago. The introduction of the 3-point line, the frequent rule changes and the continuously evolving playstyle are just some of the changes that we have witnessed over the years. Therefore, it may be interesting to determine the different NBA eras based on the statistics alone. With the help of k-means clustering, the aim is to group seasons into clusters with common traits and, consequently, identify the specific eras.

Cluster Analysis is an unsupervised classification method, where the goal is to divide individual data points into groups (i.e. clusters) based on the similarity of their attributes. In other words, cases that are grouped in the same cluster are supposed to share common characteristics with one another, while being significantly different to the cases in other clusters. Moreover, this natural grouping method can be regarded as a type of dimensionality reduction technique, as large data sets can be reduced to several homogeneous groups, making it easier to interpret the data. A commonly used partitioning technique is the so-called *k-means Clustering* method. In general, it can be divided into two main steps: identifying the number of clusters to be defined and determining the respective clusters.

An important notion for the first step is *Pearson's Correlation Coefficient* η^2 (also known as *Explained Variance*), which measures the clusterisation quality with respect to the number of clusters. More specifically, it is the ratio of the *Between Deviance* (BD) to the *Total Deviance* (TD). Let N be the number of individual cases grouped into k clusters, then for a given variable X we have

$$BD = \sum_{j=1}^{k} (\mu_j - \mu)^2 n_j \tag{22}$$

where μ is the overall mean of X, μ_j is the mean of X only for the cases belonging to the j^{th} cluster and n_j is the number of cases in the j^{th} cluster,

$$TD = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \mu)^2 \tag{23}$$

where x_{ij} is the value of the i^{th} case of the j^{th} cluster. This ratio, ranging from 0 to 1, helps determine the overall clusterisation quality, with a higher value indicating a better separation between clusters. Concurrently, the number of clusters should remain relatively low to simplify interpretation.

The second step consists of a repetitive algorithm that tends to optimise the partitioning of the cases into k distinct clusters. First of all, k cluster centers are chosen randomly, then each case is assigned to its closest cluster, according to a well-defined distance (usually the Euclidean distance). Next, the center of each cluster is recalculated so that the sum of distances between the cases in a particular cluster and its center is minimal. The re-assignment of the cases and the re-computation of the centers is then repeated until the centers become relatively stable.

In our case, the considered variables will be a collection of basic stats (2-pt. field goals made, 2-pt. field goals attempted, 3-pt. field goals made, 3-pt. field goals attempted, free throws made, free throws attempted, off. rebounds, def. rebounds, assists, steals, blocks, turnovers, personal fouls, points & Pace), while the observations will simply be the seasons from 1980-2023. With the help of the `kclustering` method in R, the graph summarising the average Pearson correlation coefficient with respect to the number of clusters is obtained.

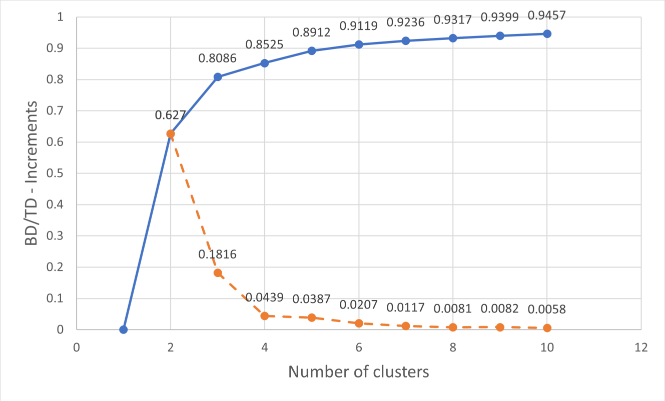


Figure 19: Quality of clusterisation wrt. number of clusters

The explained variance naturally improves with the number of clusters. A general rule of thumb is that values higher than 50% are deemed as acceptable. Keeping in mind that the number of clusters should be held low, one can consider the percentage increase of the explained variance, which is represented by the dashed line. In particular, a threshold should be fixed, which determines whether an additional cluster is justified or not. In our case, the threshold will be set at 10%, meaning that an increase below this percentage doesn't warrant an extra cluster. Looking at the figure, one can recognise that the optimal number of clusters seems to be 3, with a clusterisation quality of $\eta^2 = 62.7\%$.

The next step of the clustering algorithm can be executed by running the `kclustering` function once again, this time specifying the number of clusters. As a result, the following 3 clusters are obtained, each of which supposedly represents a different era.

Era 1	**Era 2**	**Era 3**
1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994	1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015	2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023

Table 1: 3 different eras based on k-means clustering

A first positive observation is that none of the eras contains a gap in between its seasons, guaranteeing a smooth transition from one era to the next. Interestingly enough, this property of "smoothness" remains true if the number of clusters k is increased to 4 (1980-1993, 1994-2005, 2006-2016 & 2017-2023) or 5 (1980-1989, 1990-1994, 1995-2005, 2006-2016 & 2017-2023). Nevertheless, we will stick to the 3 originally obtained eras for the remainder of the analysis. Although the second era is visibly the longest according to the results, it is important to note that this could alter based on how far back the seasons are considered and due to the ongoing 3rd era. In order to analyse the different characteristics of each cluster, one can take a look at the corresponding standardised radial plots.

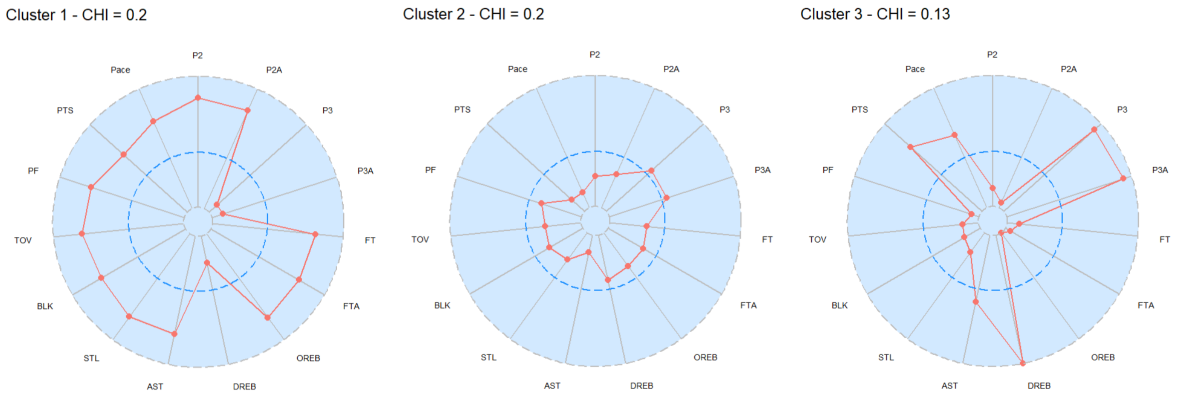


Figure 20: Standardised radial plots of average profiles in each era

As proposed by Rocha da Silva & Rodrigues (2021)[12] (a good read for those interested in deepening this topic), the three eras can be referred to as the ”Classic Era” (era 1), the ”Transitional Era” (era 2) and the ”Modern Era” (era 3). In the Classic Era, the 2-point shot seemed to be a large part of every team’s playstyle, despite the 3-point line having already been introduced. During this era, teams had above average stats in almost every category apart from 3-point shooting and defensive rebounds. This could largely be down to the relatively high pace of the game, leading to more possessions and more opportunities to record certain stats. The Transitional Era seemed to present close to average stats in most categories, compared to the other two eras. A clear change was that teams started to rely more on 3-point shots, as it became a reliable resource to win games. Meanwhile, the average pace tended to drop significantly, explaining why most of the basic stats averages were lower than in the Classic Era. Lastly, the Modern Era was captivated by the high number of 3-point attempts and makes, as well as points scored, defensive rebounds and overall pace. According to the LASSO regressions by Rocha da Silva & Rodrigues (2021)[12], in the Modern Era, the converted 2-point shots became a negative factor to win games, while the converted 3-point shots became a positive factor. Overall, one can notice clear differences between the eras: the Classic Era mainly focused on close distance shots, the Transitional Era started the shooting transition from the rim to the 3-point line and the Modern Era is characterised by an improved shot selection, shifting from less efficient 2-point shots to more rewarding 3-point shots.

Finally, the *Cluster Heterogeneity Index* (CHI), measuring the variability within a cluster, is given by the formula

$$CHI_j = \frac{\sum_{i=1}^{p} s_{ij}^2}{p} \tag{24}$$

where

$$s_{ij}^2 = \frac{\sum_{h=1}^{n_j}(x_{hi} - \mu_{ij})^2}{n_j - 1} \tag{25}$$

is the estimated sample variance of the i^{th} $(i = 1, ..., p)$ variable X_i in the j^{th} cluster of size n_j. Here, x_{hi} is the value of the variable X_i for the h^{th} data point in cluster j and μ_{ij} is the mean of the variable X_i for all the data points in cluster j. The closer this value is to 0, the better the clusterisation quality is. A general threshold can be fixed at 50%, which may vary depending on the number of individual cases.

$$CHI_1 = 0.2,\ CHI_2 = 0.2,\ CHI_3 = 0.13$$

It seems that all three clusters have a relatively low CHI, leading us to believe that the derived clusterisation is of satisfactory standard and that each cluster is more or less homogeneous.

(Note: One could further analyse the differences between eras by conducting $ANOVA$ (Analysis of Variance) tests on variables from different seasons, to determine if their means are significantly dissimilar.)

7 Key Factors influencing Player Salaries

Over the past few decades, the average player salary in the NBA has continuously risen. This is largely due to external factors such as the inflation of the USD ($), as well as the growing league and team revenue. Simultaneously, the so-called *Salary Cap*, the league-wide salary limit that a team can offer to their players, has gone up accordingly. Therefore, it may be interesting to normalise the data by considering the player salaries as a percentage of the salary cap, when looking for the key factors that influence a player's salary. In particular, the analysis will be conducted on all NBA players from the seasons 1994-1995 to 2015-2016, using the following set of variables: position, minutes played, points, assists, rebounds, steals, blocks, turnovers, personal fouls, PER, BPM, years of experience & team wins. The core information of the used dataset is provided by John Rosson (2019)[13]. Using multiple linear regression, the global aim is to understand which underlying stats are valued the most by NBA teams upon negotiating contracts and how these can help predict future salaries.

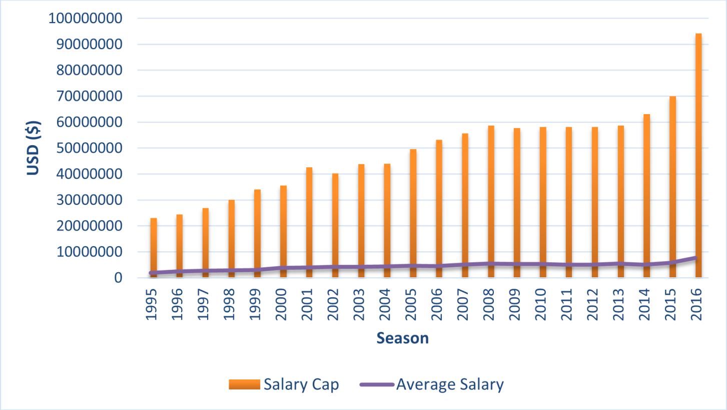


Figure 21: Salary cap & average salary from 1995-2016

Multiple linear regression is a well-known statistical modelling technique that is used to model the relationship between multiple independent variables (i.e. predictors) $X_1, ..., X_n$ and one dependent variable (i.e. response variable) Y. This relationship is expressed by a linear equation of the following form:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \epsilon \tag{26}$$

where $\beta_0, ..., \beta_n$ are the respective slope coefficients and ϵ is the model's error term, which expresses the part of Y that cannot be explained by the predictors $X_1, \ldots, X_n$. As for the quality of the modelling, it is often measured with tools such as the *Coefficient of Determination* R^2, representing the proportion of the variance in the dependent variable that is explained by the independent variables. Let $y_1, ..., y_N$ be a sample of N observations of the dependent variable Y with mean $\bar{y}$ and with corresponding predictions $\hat{y_1}, ..., \hat{y_N}$. Then we have

$$R^2 = 1 - \frac{RSS}{TSS} \tag{27}$$

where

$$RSS = \sum_{i=1}^{N} (y_i - \hat{y_i})^2 \tag{28}$$

is the *Residual Sum of Squares*, and

$$TSS = \sum_{i=1}^{N} (y_i - \bar{y})^2 \tag{29}$$

is the *Total Sum of Squares*. R^2 typically ranges from 0 to 1, with a higher value indicating a better fit of the model to the data. When the predicted model is a strictly worse predictor than the mean, we have $RSS > TSS$, which leads to R^2 taking a negative value. In this case, the model is not suited for the given data.

After having collected additional information, such as the annual salary cap, and having cleaned the dataset, an important step is to select the independent variables to be included in the model. To do this, one can first compute the correlation matrix and analyse the pairwise correlation of the variables with the help of a heat map. When choosing the independent variables that should be included in the model, it is generally a good idea to choose variables having a strong correlation with the dependent variable. A positive correlation coefficient represents a positive relationship between the variables, while a negative correlation coefficient indicates a negative relationship.

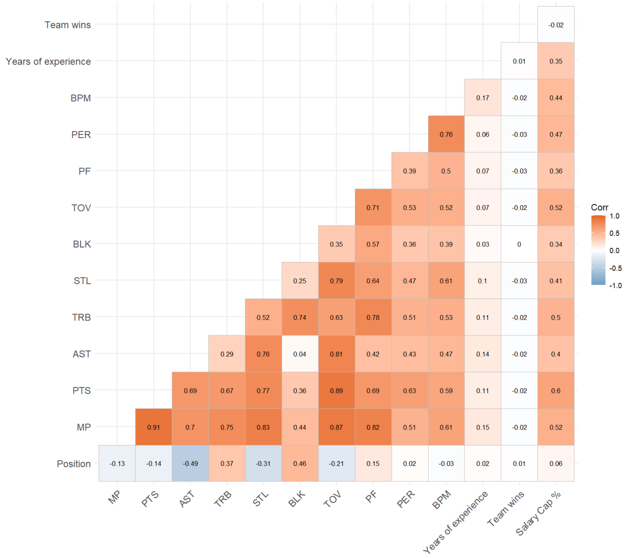
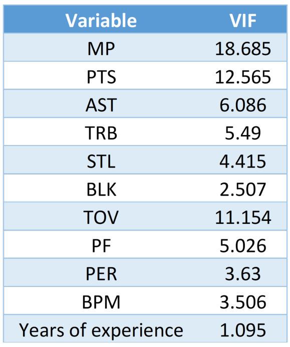


Figure 22: Heat map of correlation matrix

First of all, one can notice that two of the variables (Team wins & Position) have a correlation coefficient close to 0 with the dependent variable Salary cap % (SC%), meaning that there is little to no linear relationship between the variables. Hence, these variables will be dropped for the remainder of the analysis. Next, one also needs to consider having a low multicollinearity between the independent variables, as this is a key assumption when selecting them. Therefore, one can use the correlation matrix to compute the *Variance Inflation Factor* (VIF), which measures the multicollinearity between the predictors. More specifically, the VIF determines to which extent the variance of a predicted regression coefficient is increased due to multicollinearity in the model and should ideally be no greater than 5 for an independent variable. Its formula is given by

$$VIF_i = \frac{1}{1 - R_i^2} \tag{30}$$

where R_i^2 is the obtained coefficient of determination by considering a linear regression model with the i^{th} independent variable taking the role of the dependent variable, while all the other independent variables remain as predictors.

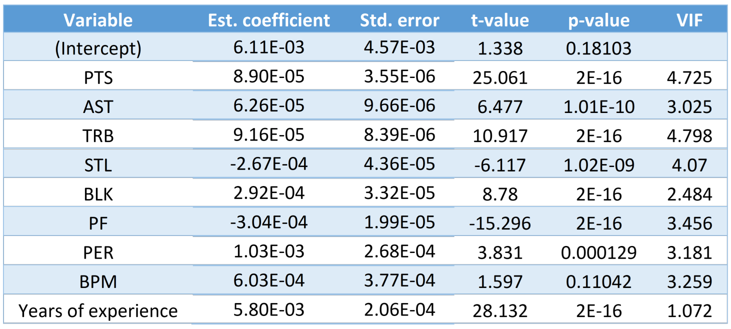


Variable	VIF
MP	18.685
PTS	12.565
AST	6.086
TRB	5.49
STL	4.415
BLK	2.507
TOV	11.154
PF	5.026
PER	3.63
BPM	3.506
Years of experience	1.095

Figure 23: VIF values for each independent variable

Based on this table, the variable MP has the highest VIF value ($\gg 5$), indicating that it is strongly correlated with the other predictors. In other words, considering it would most likely not add any new information to the model. This is not surprising since a player who plays more minutes tends to be more involved in the game, allowing him to record greater values in each statline. Due to its high multicollinearity, the variable MP is removed from the analysis. After the removal, the VIF for the remaining variables all drop below 10, apart from the variable TOV ($VIF_{TOV} = 10.687$). A brief look at the correlation heat map helps us understand that its pairwise correlation with several predictors is relatively high, especially with PTS and AST. Therefore, TOV is also removed as independent variable, finally leading to all considered VIF values dropping below 5.

The next step consists of fitting the model by using a regression analysis tool to optimally estimate the regression coefficients for each independent variable. A commonly used method is the *Ordinary Least Squares* (OLS) regression, whose goal is to minimise the sum of the squared deviations. The R function `lm` provides an estimated model containing the estimated coefficients, the standard errors, the t- and p-values, all of which can be summarised in a table.



Variable	Est. coefficient	Std. error	t-value	p-value	VIF
(Intercept)	6.11E-03	4.57E-03	1.338	0.18103	
PTS	8.90E-05	3.55E-06	25.061	2E-16	4.725
AST	6.26E-05	9.66E-06	6.477	1.01E-10	3.025
TRB	9.16E-05	8.39E-06	10.917	2E-16	4.798
STL	-2.67E-04	4.36E-05	-6.117	1.02E-09	4.07
BLK	2.92E-04	3.32E-05	8.78	2E-16	2.484
PF	-3.04E-04	1.99E-05	-15.296	2E-16	3.456
PER	1.03E-03	2.68E-04	3.831	0.000129	3.181
BPM	6.03E-04	3.77E-04	1.597	0.11042	3.259
Years of experience	5.80E-03	2.06E-04	28.132	2E-16	1.072

Figure 24: Results of estimated model

The estimated coefficients describe the possible linear relationships between the independent variables and the response variable. For example, keeping all the other predictors constant, an increase of 1 unit in PTS would presumably lead to an increase of 8.904e-05 ($\approx$ 0.008904%) in the salary cap %. Overall, PER and Years of experience have the largest estimated coefficients with 1.03e-03 and 5.80e-03 respectively. In addition, the estimated coefficient of the intercept is 0.006114, meaning that the dependent variable equals 0.006114 when all the independent variables are set to 0. Meanwhile, the standard error, the t- and p-values all measure the quality of the estimated coefficient. In general, the p-value is the most commonly used out of the three, since a value below a certain significance level (usually 0.05) reveals that a predictor is statistically significant to forecast the response variable. In this case, all the independent variables are statistically significant, apart from the variable BPM. Furthermore, the value $R^2 = 49.75\%$ confirms a relatively good fitting, as almost half of the dependent variable's variance can be explained by the estimated model. Finally, the obtained linear regression equation is given by:

$$\begin{aligned} Y_{SC\%} =& 6.11\text{e-}03 + 8.90\text{e-}05 X_{PTS} + 6.26\text{e-}05 X_{AST} + 9.16\text{e-}05 X_{TRB} - 2.67\text{e-}04 X_{STL} \\ &+ 2.92\text{e-}04 X_{BLK} - 3.04\text{e-}04 X_{PF} + 1.03\text{e-}03 X_{PER} + 6.03\text{e-}04 X_{BPM} \\ &+ 5.80\text{e-}03 X_{YoE} + \epsilon \end{aligned} \tag{31}$$

In the case where a player's values of the predictors are known, his future salary can be predicted with the help of this equation and the corresponding annual salary cap.

8 Going further: Additional Approaches to Basketball Analytics

This section consists of multiple advanced methods, allowing us to discover patterns in data. Requiring a large number of modern-day stats, this analysis can primarily only be done on games, teams and players from the 21st century. Generally speaking, these tools are only some of the modern day techniques that are used by basketball analysts and experts to understand the hidden mechanisms behind the data.

8.1 Shooting Density Estimation

When analysing a game, it can be of interest to measure the occurrence frequency of a certain event, as proposed by Zuccolotto & Manisera (2020)[16]. For example, the shooting frequency in time and space can give teams and players key insights as to which types of shots are the most effective. *Density Estimation Methods*, such as histograms and kernel (KDE), are tools that estimate the cumulative distribution function or probability density function of a random variable. With the help of the R method `densityplot` and play-by-play data of the games, the kernel density estimation of shots with respect to the total time (total amount of seconds played), the play length (seconds in to the 24-second shot clock) and the shot distance (distance from the rim in feet) can be computed and plotted. In particular, the `densityplot` function uses the common Gaussian kernel density estimation method, as well as a bandwidth, whose default value is simply the standard deviation of the kernel. The kernel is a smooth weighting function that spreads the influence of each data point across its neighborhood. Meanwhile, the bandwidth controls the width of the kernel function and determines to which extent each data point influences the density estimate. Thanks to the play-by-play data provided by Zuccolotto & Manisera (2020)[16], we can consider an example of density estimation, looking at the field goal density of the 2017-2018 regular season Golden State Warriors with respect to the three previously mentioned variables.

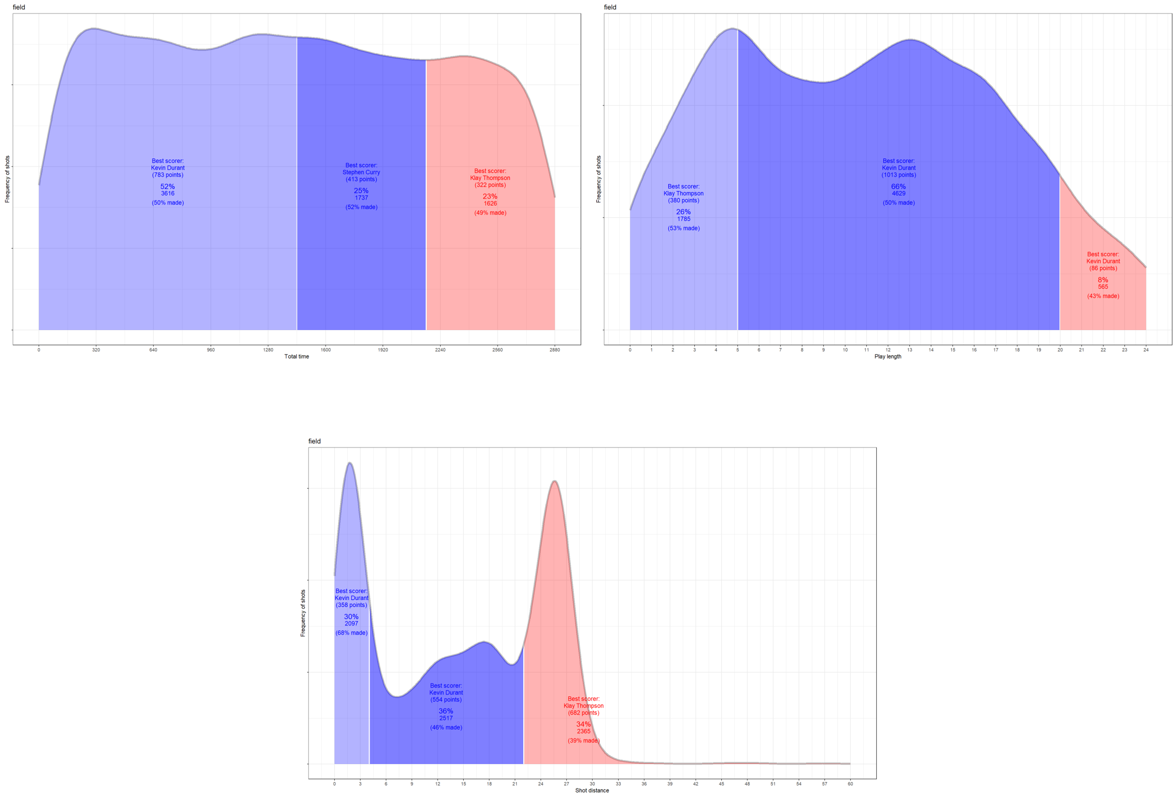


Figure 25: Density estimation of GSW' shooting performance in 2017-2018 with respect to total time, play length & shot distance

Based on the plots of the estimated densities, the Warriors tend to shoot slightly more field goals in the first half (52%) than in the second half (48%). While their shooting percentages seem to be fairly consistent throughout the entire game ($\approx$ 50%), the best scorer on the team varies from one period to the next. Furthermore, they tend to shoot relatively fast, with 26% of their shots coming within the first five seconds of the shot clock, maintaining a high field goal percentage for these shots (53%). On the other hand, they only rarely have to force shots near the end of the shot clock (8%), making an impressive 43% of these shots nonetheless. In terms of the shot distance, the Warriors tend to mainly take two types of shots, either within 4 feet of the rim (30%) or from behind the 3-point line (34%). Their remarkable shooting percentages are 68% and 39% respectively, with Kevin Durant being the best scorer from inside the 3-point line and Klay Thompson being the best from behind the arc.

A further way to represent the spatial distribution of data is by using graphical tools, such as density polygons, rasters and hexbins. A commonly used graph in basketball analytics is the so-called *Shot Chart*, illustrating the estimated spatial shot density according to different zones on the court. Using the same data as before, one can consider several types of shot charts for arguably the best shooter of all-time, Stephen Curry.

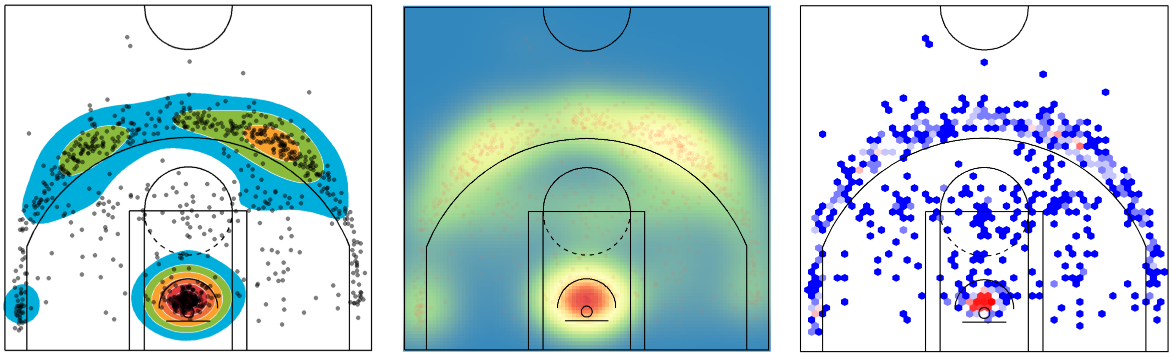


Figure 26: Spatial density estimation of Stephen Curry's shooting performance in 2017-2018

The three plotted shot charts all take the same data as input, while displaying the obtained spatial density estimates slightly differently. Overall, one can notice that Curry tends to take most of his shots either from the 3-point line or from very close range. Moreover, the majority of his 3-point shots come from the top of the arc, with a slight preference for his left-hand side.

8.2 High-pressure Shooting

Another interesting subject to analyse is the classification of shots according to a specific game situation. In particular, this can provide us with information on which players are the best at dealing with high-pressure conditions. In Zuccolotto & Manisera (2020)[16], they studied the impact of high-pressure game situations on the shooting performance of players by means of *CART* (Classification and Regression Trees), a popular decision tree algorithm in machine learning. The first step of their study consisted of clearly defining the different high-pressure game situations. Incorporating the suggestions of several basketball experts, they came up with the following situations:

> When the shot clock is going to expire, when the score difference with respect to the opponent is small, when the team as a whole has performed poorly during the match up to that particular moment in the game, when the player who is shooting has missed his previous shot, and when the time left on the clock is running out. (Zuccolotto & Manisera, 2020, p. 101)[16]

More specifically, the authors used CART to examine the influence of these high-pressure conditions on a player's scoring probability by considering all the joint associations between the variables. Based on the obtained resulting CART models, they could conclude that not all shots are alike, which led them to develop their own shooting performance measure, focusing on the shot conditions. This new metric should supposedly value a made shot more when the scoring probability is lower (e.g. near the end of the shot clock). The authors introduced the following measure:

> For each shot type T (2P: 2-point, 3P: 3-point, FT: free-throw), let J_T be the set of attempted shots of type T and x_{ij} the indicator assuming value 1 if the jth shot of the ith player scored a basket and 0 otherwise. The new shooting performance of player i for shot type T is given by
>
> $$P_i(T) = av_{j \in J_T}(x_{ij} - \pi_{ij}) \tag{32}$$
>
> where $av_{j \in J_T}(\cdot)$ denotes averaging over all of the shots of type T attempted by player i and π_{ij} is the scoring probability assigned by the CART model to the jth shot of the ith player, that is to a shot of the same type and attempted in the same game situation as the jth shot of the ith player. For each shot, the difference $x_{ij} - \pi_{ij}$ can be used as a performance measure of the shot. In fact, the difference is positive if the shot scored a basket and negative if it missed. (Zuccolotto & Manisera, 2020, p. 106)[16]

The measure P_i can be either positive or negative and is ideally compared between players with the help of a bubble plot, with the x-axis, y-axis and colour representing the $P_i(2P)$, $P_i(3P)$ and $P_i(FT)$ respectively. In their analysis, the authors used data from the *Olympic Basketball Tournament Rio 2016*, only considering the players having attempted over 15 shots of each type.

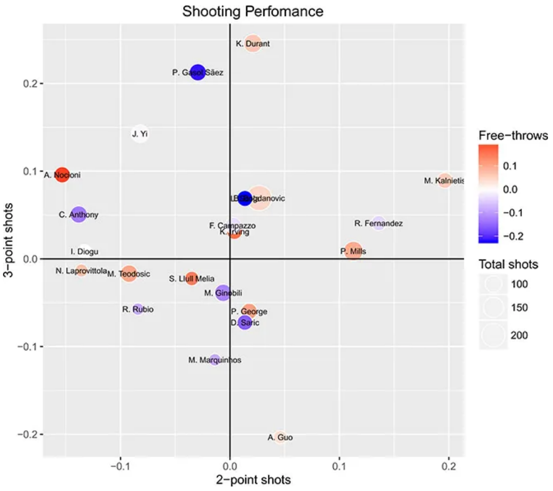


Figure 27: Bubble plot of players' shooting performance values from *Rio 2016*. Source:[17]

The bubble plot reveals how players perform compared to the average, taking both the shot type and game situation into consideration. The top right plane presents the players shooting above average from both 2-points and 3-points, while the bottom left plane presents the players shooting below average in both categories. Concurrently, the more red a player's bubble is, the better their free throw shooting is. An interesting feature of this new measure, is that it highlights key differences between players having the same standard shooting percentages. As mentioned by the authors, B. Bogdanovic and A. Nocioni both have very similar 3-point field goal percentages (≈ 45%), however, according to the new metric, Nocioni shoots slightly better than Bogdanovic, based on the shot conditions. Contrarily, the measure also identifies the players having different standard shooting percentages, while sharing a similar value in the shooting performance measure. As mentioned by the authors once again, N. Laprovittola and C. Anthony are evaluated almost equally regarding 2-point shots, even though they have different 2-point field goal percentages (36.8% & 38.5% respectively).

9 Conclusion

The primary goal of my work was to shed some light on some common NBA discussion topics, based solely on statistics. For that, I used several well-known analytical methods, such as Principal Component Analysis, Multidimensional Scaling, k-means clustering and multiple linear regression.

After having logically developed my own player and team evaluation metrics (PPI & TPI), the main focus was to determine the greatest player and the greatest single season team of all-time. Overall, the analysis and metric values revealed that Michael Jordan can be considered as the GOAT, while the dominant 2016-2017 Golden State Warriors topped the charts in the team category.

Next, I attempted to identify and classify the key eras in NBA history, with the help of k-means clustering. According to the obtained results, the NBA seasons from 1980-2023 can be grouped into three eras, all having slightly different characteristics: the Classic Era (1980-1994), the Transitional Era (1995-2015) and the Modern Era (2016-2023).
By means of multiple linear regression, it turns out that all of the basic season stats, as well as the PER and the experience of a player, are statistically significant and explain approximately half of the variance, when it comes to determining the proportion of a player's salary with respect to the annual salary cap.

Finally, some advanced analytical methods, such as shooting density estimation and the development of a true shooting performance measure, were presented, to give the reader an idea of how modern-day teams and players can benefit from basketball analytics.

On a personal note, I was able to get a first insight into collecting, manipulating and interpreting large sets of data, while doing research on a topic I thoroughly enjoy. Not only could I apply my previous Sports and Statistics knowledge to the research, I also learned several new data analysis techniques, such as PCA and MDS, which will undoubtedly be useful to me in the near future.

10 Appendix

10.1 Calculating PER

John Hollinger's *Player Efficiency Rating* (PER) needs to be calculated in two steps, according to (*Calculating PER*, n.d.)[1]. First, we determine the uPER (unadjusted PER), using the following formula:

$$\begin{aligned} uPER = \frac{1}{MP}\Big[&3FG + \frac{2}{3}AST + \Big(2 - factor \cdot \frac{team\ AST}{team\ FG}\Big) \cdot FG \\ &+ FT \cdot 0.5 \cdot \Big(1 + \Big(1 - \frac{team\ AST}{team\ FG}\Big) + \frac{2}{3} \cdot \frac{team\ AST}{team\ FG}\Big) \\ &- VOP \cdot TO - VOP \cdot DREB\% \cdot (FGA - FG) \\ &- 0.44VOP \cdot (0.44 + 0.56DREB\%) \cdot (FTA - FT) \\ &+ VOP \cdot (1 - DREB\%) \cdot (REB - OREB) \\ &+ VOP \cdot DREB\% \cdot (OREB + BLK) + VOP \cdot STL \\ &- PF \cdot \Big(\frac{league\ FT}{league\ PF} - 0.44\frac{league\ FTA}{league\ PF} \cdot VOP\Big)\Big] \end{aligned} \tag{33}$$

where:

- MP stands for minutes played
- FT(A) stands for free throws made (attempted)
- $factor = \frac{2}{3} - \frac{1}{4}\frac{league\ AST \cdot league\ FT}{(league\ FG)^2}$
- $VOP = \frac{league\ PTS}{league\ FGA - league\ OREB + league\ TO + 0.44 \cdot league\ FTA}$
- $DREB\% = \frac{league\ REB - league\ OREB}{league\ REB}$

Hollinger chose the different statistics included in PER according to their ability to describe a player's impact on the game. Considering the determination of the respective weights, Hollinger primarily used a regression analysis to deduce the relative importance of each statistical category, while also taking his own judgment and experience as a basketball analyst into account.

Secondly, the obtained uPER needs to consider the pace of the team, as well as the league. Therefore, the uPER is standardised, giving us the wanted PER value:

$$PER = 15 \cdot \frac{league\ Pace \cdot uPER}{team\ Pace \cdot league\ uPER} \tag{34}$$

Calculating the PER for seasons before 1979-1980 can be problematic, since the NBA hadn't introduced the 3-point shot and wasn't tracking some of the basic stats before then. However, by simply disregarding these statistics, one can still come up with a relatively close approximation to what the actual PER value would have been for a player playing prior to 1979-1980.

10.2 Calculating BPM

Similar to the PER, Daniel Myers' *Box Plus-Minus* (BPM) is calculated in several steps, by adding 9 different terms. Each individual term takes a different basic statistic into account, allowing us to cover most of the box score stats. Here are the formulas for these particular terms, according to Cappe (2020)[2]:

$$rBPM_1 = 0.123391 \cdot \frac{MP}{GP + 2} \tag{35}$$

$$rBPM_2 = 0.119597 \cdot OREB\% \tag{36}$$

$$rBPM_3 = -0.151287 \cdot DREB\% \tag{37}$$

$$rBPM_4 = 1.255644 \cdot STL\% \tag{38}$$

$$rBPM_5 = 0.531838 \cdot BLK\% \tag{39}$$

$$rBPM_6 = -0.305868 \cdot AST\% \tag{40}$$

$$rBPM_7 = 0.921292 \cdot USG\% \cdot TO\% \tag{41}$$

$$\begin{aligned} rBPM_8 =& 0.711217 \cdot USG\% \cdot (1 - TO\%) \cdot [2 \cdot (TS\% - team\ TS\%) \\ &+ 0.017022 \cdot AST\% + 0.297639 \cdot (3PTS\% - league\ 3PTS\%) \\ &- 0.213485] \end{aligned} \tag{42}$$

$$rBPM_9 = 0.72593 \cdot \sqrt{AST\% \cdot REB\%} \tag{43}$$

where:

- MP & GP stand for minutes & games played respectively
- USG% stands for Usage percentage and is calculated in the following way:

$$USG\% = \frac{team\ MP \cdot (FGA + 0.44FTA + TO)}{5MP \cdot (team\ FGA + 0.44 \cdot team\ FTA + team\ TO)}$$

- TS% stands for True Shooting percentage and is calculated in the following way:

$$TS\% = \frac{50PTS}{FGA + 0.44FTA}$$

By adding all of these terms together, we get the so-called "raw" BPM:

$$rBPM = \sum_{i=1}^{9} rBPM_i \tag{44}$$

Once again, this value needs to be normalised to each team individually. Therefore, the final BPM value is given by considering a corrective term, namely the team adjusted coefficient:

$$BPM = rBPM + \frac{1.2 \cdot team\ NRtg - 5 \cdot rBPM \cdot \frac{MP}{team\ MP}}{5} \tag{45}$$

As already explained before, this value can be either positive or negative. Myers developed the majority of the formulas by repetitively testing different linear combinations of advanced statistics, in order to find the combination that led to the smallest error in his regression analysis. Further details can be found in the article *About Box Plus/Minus (BPM)* [9], written by the developer Myers himself.

10.3 Python Code

First of all, this is the Python code for the PCA algorithm, that was primarily used for the development of the self-invented metrics:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

# Load data from CSV file
data = pd.read_csv("Top 75 players advanced metrics.csv")

# Replace the N/A values by column average
values = {"PER": data["PER"].mean(), "BPM": data["BPM"].mean(), "NRtg":
    data["NRtg"].mean()}
data = data.fillna(value=values)

# Standardize the data
scaler = StandardScaler()
data_std = scaler.fit_transform(data.drop(’Player Name’, axis=1))

# Compute covariance matrix of data
print(np.cov(np.transpose(data_std)))

# Perform the other steps of PCA
pca = PCA()
pca.fit(data_std)

# Print principal components
print(pca.components_)

# Get the explained variance ratios
variance_ratios = pca.explained_variance_ratio_
print(variance_ratios)

# Plot scree plot
plt.plot(np.cumsum(pca.explained_variance_ratio_))
plt.xlabel(’# of Components’)
plt.ylabel(’% of Variance’)
plt.show()

# Extract the first component
first_component = pca.components_[0]

# Create a DataFrame from the first component
component_df = pd.DataFrame(first_component, index=data.columns[1:],
    columns=[’PC1’])

# Print the DataFrame
print(component_df)
```

Listing 1: PCA code

The following code can be used to execute the MDS algorithm without using the R method `MDSmap`:

```python
from sklearn.manifold import MDS
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import pairwise_distances
import pandas as pd
import matplotlib.pyplot as plt

# Load data from Excel file and select desired columns
df = pd.read_excel("Top 75 players career regular season averages.xlsx",
                   usecols=["trebPerGame", "astPerGame", "stlPerGame", "blkPerGame", "ptsPerGame"])

# Replace missing values with column mean
column_mean = df.mean()
df.fillna(column_mean, inplace=True)

# Compute pairwise Euclidean distances between data points
distances = pairwise_distances(StandardScaler().fit_transform(df), metric='euclidean')

# Instantiate MDS object with desired parameters
mds = MDS(n_components=2, dissimilarity="precomputed", random_state=42)

# Fit MDS model to distance matrix
fitting = mds.fit_transform(distances)

# Plot final fitting
plt.scatter(fitting[:, 0], fitting[:, 1])
plt.title("MDS Plot")
plt.show()
```

Listing 2: MDS code

10.4 Datasets

As I was unable to include the full details of every single used dataset, I have decided to upload all of them to an online folder, which is linked down below:

https://drive.google.com/drive/folders/14GS7B4kyhkg4h0C94rEmXGfqBC3x_L0Z?usp=share_link

Glossary

2FG 2-point Field Goals made
2P% 2-point Field Goal Percentage
3FG 3-point Field Goals made
3P% 3-point Field Goal Percentage

AST Assists

BD Between Deviance
BLK Blocks
BPM Box Plus-Minus

CART Classification and Regression Trees
CHI Chicago Bulls
CHI Cluster Heterogeneity Index

DREB Defensive Rebounds
DRtg Defensive Rating

eFG% Effective Field Goal Percentage

FF Four Factors
FG Field Goals made
FG% Field Goal Percentage
FGA Field Goals attempted
FT Free Throws made
FT% Free Throw Percentage
FTA Free Throws attempted

GOAT Greatest Of All-Time
GP Games Played
GSW Golden State Warriors

KDE Kernel Density Estimation

LAL Los Angeles Lakers
LASSO Least Absolute Shrinkage & Selection Operator

MDS Multidimensional Scaling

MIN Minutes
MP Minutes Played
MVP Most Valuable Player

NBA National Basketball Association
NRtg Net Rating

OLS Ordinary Least Squares
OREB Offensive Rebounds
ORtg Offensive Rating

PCA Principal Component Analysis
PER Player Efficiency Rating
PF Personal Fouls
POSS Possessions
PPI Player Performance Index
PTS Points

rBPM Raw Box Plus-Minus
REB Rebounds
RSS Residual Sum of Squares

SC% Salary Cap Percentage
STL Steals

TD Total Deviance
TO Turnovers
TPI Team Performance Index
TS% True Shooting Percentage
TSS Total Sum of Squares

uPER Unadjusted Player Efficiency Rating
USG% Usage Percentage

VC Variation Coefficient
VIF Variance Inflation Factor
VOP Value of Possession

WVC Weighted Variation Coefficient

References

[1] *Calculating PER.* (n.d.). Basketball Reference. Retrieved April 5, 2023, from https://www.basketball-reference.com/about/per.html

[2] Cappe. (2020, February 29). *Learn a Stat: Box Plus Minus and VORP.* Hack a Stat. https://hackastat.eu/en/learn-a-stat-box-plus-minus-and-vorp/

[3] *Four Factors.* (n.d.). Retrieved April 8, 2023, from https://www.basketball-reference.com/about/factors.html

[4] Kalén, A., Pérez-Ferreirós, A., Costa, P.B., Rey, E. (2020), Effects of age on physical and technical performance in National Basketball Association (NBA) players, *Research in Sports Medicine*, 29(3):277-288.

[5] Kubatko, J., Oliver, D., Pelton, K., Rosenbaum, D.T. (2007), A Starting Point for Analyzing Basketball Statistics, *Journal of Quantitative Analysis in Sports*, 3(3):1-22.

[6] Ley, C. & Dominicy, Y. (2020), Science Meets Sports: When Statistics Are More Than Numbers, *Cambridge Scholars Publishing.*

[7] *Linear regression.* (2023, May 8). Wikipedia. https://en.wikipedia.org/wiki/Linear_regression

[8] *Multidimensional scaling.* (2023, March 28). Wikipedia. https://en.wikipedia.org/wiki/Multidimensional_scaling

[9] Myers, D. (2020, February). *About Box Plus/Minus (BPM).* Basketball Reference. https://www.basketball-reference.com/about/bpm2.html

[10] *Principal component analysis.* (2023, April 18). Wikipedia. https://en.wikipedia.org/wiki/Principal_component_analysis

[11] Ramzai, J. (2020, April 27). *Clearly Explained: Gini coefficient and Lorenz curve.* Towards Data Science. https://towardsdatascience.com/clearly-explained-gini-coefficient-and-lorenz-curve-fe6f5dcdc07

[12] Rocha da Silva, J.V. & Rodrigues, P.C. (2021), The three Eras of the NBA regular seasons: Historical trend and success factors, *Journal of Sports Analytics*, 7(4):263-275.

[13] Rosson, J. *NBA Salary Predictions using Data Science and Linear Regression.* (2019, May 2). Towards Data Science. https://towardsdatascience.com/nba-salary-predictions-4cd09931eb55

[14] *A whole new ball game: Quantifying changes in NBA basketball over the past 30 years.* (2018, October 19). Thinking Machines. https://stories.thinkingmachines/nba-in-30-years/

[15] Zhao, Y. (2022), Model Prediction of Factors Influencing NBA Players' Salaries Based on Multiple Linear Regression, *Proceedings of the 2022 2nd International Conference on Economic Development and Business Culture (ICEDBC 2022).*

[16] Zuccolotto, P. & Manisera, M. (2020), Basketball Data Science: With Applications in R, *CRC Press.*

[17] Zuccolotto, P., Manisera, M., Sandri, M. (2018), Big data analytics for modeling scoring probability in basketball: The effect of shooting under high-pressure conditions, *International Journal of Sports Science & Coaching*, 13(4):569-589.