The Evolution of NBA Star Players: A Visual Approach

School of Mathematics, Computer Science and Engineering
dept. of Computer Science
City, University of London
London, UK
matheo.hudes@city.ac.uk

Abstract—This research paper explores the diverse characteristics of NBA players throughout the years, visualises the various changes in the way the game is played and used both statistics and visual analytics to answer the following questions: How is the NBA evolving and how different are players nowadays? The paper finds some insights by showing that extreme positions tend to converge towards average height and weight, as well as their shooting trends.

1 Problem Statement

The level and efficiency of a sports athlete is a key measure for sports analysts, team coaches or bookmakers. Lots of different research papers deal with what makes a performing football team or player, by analysing each factor or the game accordingly to each player individually. This analysis problem is yet to be studied statistically for the NBA. The world's most popular basketball league generates billions of revenues every year [insert reference on how popular the NBA is] and the study of its structure can lead to a significant impact on the way the league is being perceived. Major popular sports are constantly evolving and the players as well as the criteria describing them are subject to change. It is essential to assess the correct factors that determine the quality of a player in order to maximise the efficiency of a team. The visual approach of the data used aim to provide meaningful insights regarding the evolution of NBA players, reflecting the evolution of the sport itself, among the arguably best league in the world. Several datasets and API including characteristics about NBA players throughout the years will be used to carry the analysis of the attributes of the best players, alongside the evolution of the sport.

2 State of the Art

The first paper from Goldsberry [1] emphasizes on the spatial dimension of basketball, and all the metrics associated with the sport that are not being accounted. CourtVision brings some more context insight into the basic NBA statistics such as FG%, being the reference to determine whether a player is a good shooter. Even the eFG% accounting for 2-point shots and 3-point shots does not differentiate different types of zones and kinds of shots. This paper brings up a new approach to players' efficiency through visual analytics and shows graphically the best shooting areas of key players as well as their efficiency all around the court with a metric called "Range", based on which Steve Nash is leading between 2006 and 2011. This approach will be our foundation for our research on the evolution in the game. The paper shows that it is through visual analytics that basic statistics can be expanded and lead to better insights, which is what this paper is meant to do.

The second paper from Benito Santos et al. [2] proposes a collective analysis of the performance with all the factors affecting the team members equally, followed by a breakdown with an individual analysis, dealing with the metric affecting individual performance. This paper also uses heatmaps to assess the performance of players accurately over a certain period. Though it is not applied to the same sport, this paper demonstrates how the use of advanced visual techniques lead to more accurate and performing analyses of performance. The diagrams and heatmaps help to complete the numerical statistics and get a better spatial representation of a team's strengths and weaknesses.

The third paper from Zhang et al. [3] shows that team's characteristics tend to change in the beginning and ending of the season and remain stable during the season. It emphasizes on the fact that the dynamical nature of basketball makes it hard to analyse using a univariate approach. This paper captures the metrics month by month from October to April and uses dimensional reduction to plot all the factors such as 2-point and 3-point shots made/attempted, assists, turnovers or personal fouls. The plot of each team then demonstrates similarity within the dominant teams in the league (especially the champion and the runner-up). It could be concluded to some extent that there are better ways to play basketball, even among the best teams in the world. This result will be used as a reference to our research, showing that the evolution of the sport might be because of weaker teams copying the dominant ones, shifting then the general trend.

The last paper from Lorenzo et al. [4] assesses the evolution of in-game attributes of a given player throughout his career. Empirical research shows that players improve their pass skills in 91% of cases and 3-point ability in 59% of cases, which could be interesting to us, since this is the major change in the NBA in the past 15 years.

3 Properties of the Data

The first dataset called "NBA Players" was collected from Kaggle and includes several properties about NBA players drafted between 1996 and 2022. It includes 12844 players (rows) across various seasons, alongside 22 different columns describing the players with their team, height and weight, or season statistics (points, rebounds, assists, etc.) The dataset contains 2551 unique players, 4 categorical variables and 18 numerical ones. It will be used to determine whether we see a trend in the evolution of players' physiques, especially with the emergence of "small ball" [6].

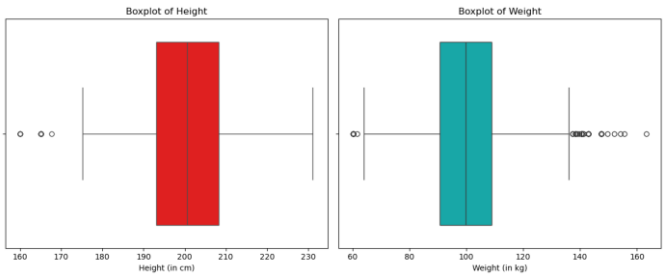

Figure 1. Boxplot of Height and Weight of NBA Players.

The distribution of these 2 variables is quite wide, mainly because the different positions played in basketball imply much different attributes. It is said that the average height of a point guard is around 188cm while a center is about 210cm. [6]

The second source of data comes from the API's of the NBA website. Every statistic can be extracted from this source, but the function returns a dataset for a given player's statistics as well as the league averages on the given year. The player statistics include every shot he has attempted with the X and Y coordinates, the ID of every match he has played, the position of the shot on the court and the time remaining on the clock at the shot moment. There is no definite scale for the shot location X and Y but a quick verification lets us think that it is meant to be drawn on a virtual court with X ranging from -250 to 250 and Y ranging from -50 to 450, the basketball hoop being centred on 0,0.

There are as many data as there are shots from NBA players from the year 1960 to 2022, therefore the shape of the entire dataset remains unknown. The spatial dimension of this dataset makes it useful to bring a visual aspect to our analysis, in order to find additional insights where basic statistics seem to remain similar year after year.

The data from the Kaggle datasets will help to get a first approach and formulate hypotheses that advanced statistics from the NBA API can potentially solve.

4 Analysis

4.1 Analysis Approach

Though every statistic can be used for a specific purpose, our real interest is to combine the basic statistics of the first dataset with the complex ones from the NBA API's. The mixture of both, linked by their temporal variables, help us to assess correctly the trends in the NBA throughout the year, both on the kind of players that rise or the way the game evolved to create those new NBA stars. The analysis from the first dataset is easy to interpret for everyone, but suddenly becomes complex when trying to link the visuals from the API's, because careful consideration of the statistics must be taken and a human expertise of the sport helps pulling out some conclusions that an algorithm cannot solve any other way than with a correlation, which omits lots of factors.

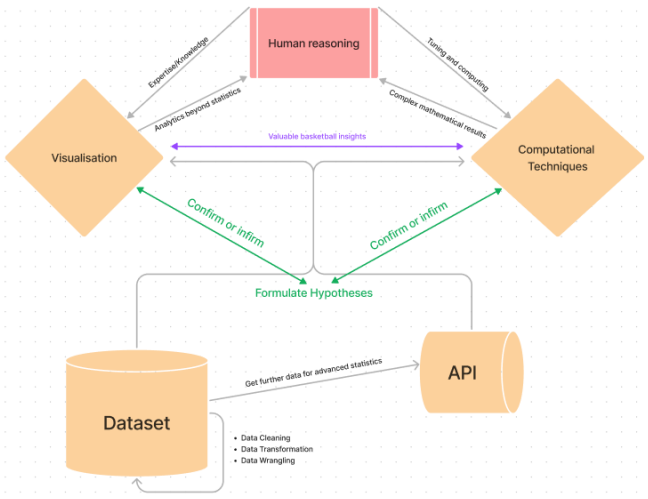

Figure 2. Work frame diagram of this project.

The diagram above illustrates precisely the idea of this project. The combination of our 2 sources of data will create lots of resources that a human cannot deal with mathematically due to their complexity, hence the utility of the computer in this case. On the other hand, some visual techniques employed throughout this paper are going to be easy to interpret for a human conscience (even more with some prior knowledge of basketball), where a computer will not be able to get any insights from the visuals. The goal is to

involve both processes hand in hand in order to find a complex link between them that can lead to a answer to our hypothesis that can be complex enough. The diagram also highlights accurately the data transformation that has to be performed in order to get the mathematical computing, such as cleaning, creating new variables or columns based on the existing ones, scaling the data or concatenating data together.

4.2 Analysis Process

As mentioned in the problem statement, the purpose of this paper is to analyse the evolution of players in the NBA throughout the years, particularly between 2005 and 2022, and try to understand the reasons behind the change in gameplay that occurred during this period of time. The physique statistics provided by the first dataset will help us determine in the first place whether players have changed physically, the impact that this may have on statistics, and finally how we could link the potential change in physical attributes with the gameplay change.

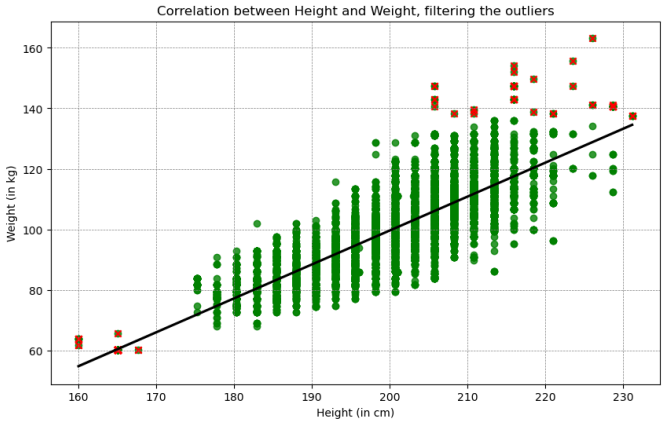

Figure 3. Scatter plot with regression line of Height (in cm) and Weight (in kg).

The first scatter plot of our analysis shows us the trivial result that height and weight are heavily correlated. The outliers are being detected with the IQR method, and only represent a few values, especially for the shortest players.

NBA players are apparently of various heights and weights, as shown on the boxplot in the first place (Figure 1), but how does this impact the game?

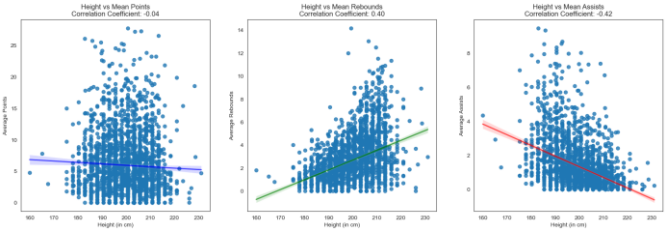

Figure 4. Different scatter plots of basketball statistics and height.

Those 3 scatter plots with regression lines respectively show the correlation between height and points, height and rebounds, height and assists. A linear regression is not adapted to the first plot distribution, as the points seem to follow a kind of normal distribution, suggesting that players with average measurements tend to score the most. Secondly, rebounds are correlated (with a coefficient of 0.40) with height but do not necessarily imply direct causality. Though it remains easier to catch rebounds for taller players, these players are usually assigned to positions closer to the basket, which makes them more likely to get rebounds. Oppositely, the same phenomenon happens for assists, which correlation to height is moderately negative (-0.42). Smaller players do not necessarily have a better game vision than taller ones, but are designed – because of their height – to be point guards, which are in charge of leading the team tactics, as well as delivering assists for their teammates.

A further exploration into our dataset shows that the average height of players is oscillating through the seasons, with no negative slope necessarily.

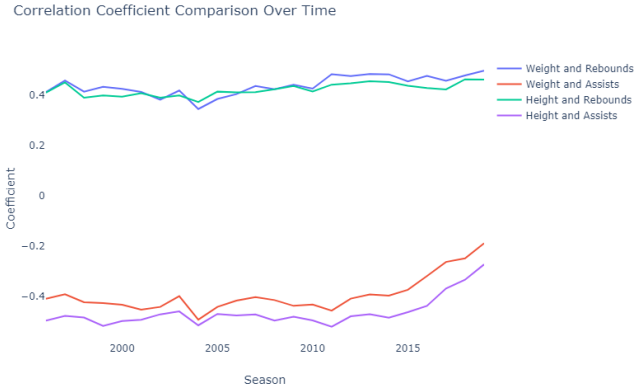

Figure 5. Correlation of Height/Weight and Rebounds/Assists over the seasons [7].

The correlation between measurements and rebounds seems stable over the years. Yet, the correlation between these measurements and the number of assists seems to be decreasing over the years, showing a better ability of the big men to be creative and deliver assists. Since the average height remains stable, this either means that:
- Big men are getting shorter while point guards are getting bigger.
- Big men changed their play style and became more versatile, potentially playing further from the rim.

Performing a clustering technique is a way to provide information on the distribution of data over the years.

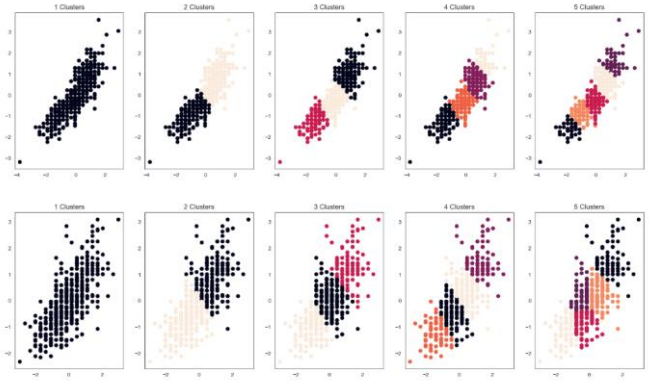

Figure 6. K-Means clustering on using Height and Weight data from 2005 (top) and 2022 (bottom).

The computation of K-means clustering for different number of clusters (1 to 5) on both 2005 and 2022 gives a true analysis of how the data is distributed. The rule of elbow [8] suggests that 2 clusters is an appropriate number for an optimal trade-off between interpretability and accuracy. 3 clusters will be chosen instead, because they are more likely to represent accurately the different measurements of NBA, supposedly grouped in 3 groups:

- The "small men", usually Point Guards and Shooting Guards
- The "standard players", usually Small Forwards, and some tall Shooting guards/small Power Forwards
- The "big men", usually Power Forwards and Centers.

A visual analysis of 3 clusters both in 2005 and 2022 seems already quite revealing. Data from 2005 looks more linear, with less outliers and then an easier ability to cluster, meaning that players in 2022 are more likely to play with atypical measurements.

The visual analysis of the clusters is a first step in evaluating the K-Means clustering technique, and there are several techniques to evaluate it mathematically. We will use the silhouette score. The silhouette score computes a score ranging from -1 to 1 for each data point of each cluster, separated by a colour code. A score approaching +1 indicates that the data point is the furthest from the neighbouring clusters. On the other hand, a value approaching -1 means that the data point is likely to be assigned to the wrong cluster, while a value of 0 means that the data point is extremely close to the boundary between two clusters.
The goal is then to maximise the score for all our data points and avoid negative values, meaning then that values are assigned to clusters that are far from each other.

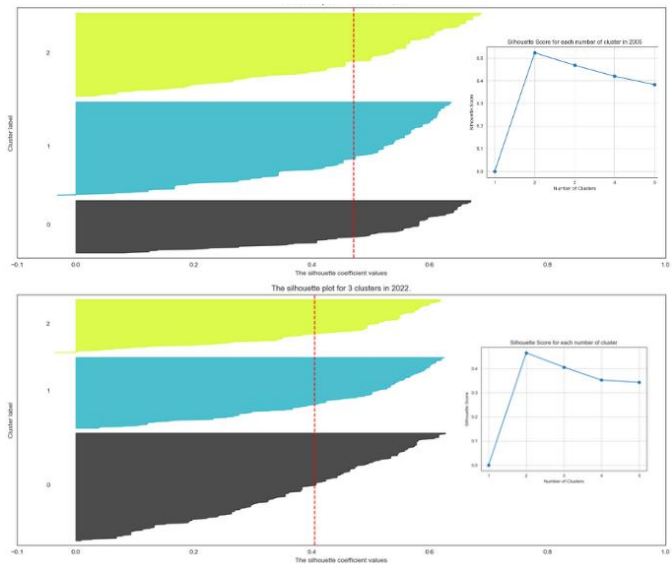

Figure 7. Silhouette score of 3 clusters with their graphical representation, in 2005 (top) and 2022 (bottom)

This figure shows the silhouette score as described above. First thing to notice is that only few data points seem to be misclassified, both in 2005 and 2022. The silhouette plot for clusters 2 and 1 in 2005 seems to be thicker than those from 2022. We can interpret this as saying that the measurements used to be more heterogeneous back then, with only centers being big men and the point guards almost necessarily being smaller than their teammates. On the other hand, The cluster number 2 (the black plot) seems far thicker on the plot from 2022, assuming then that players' dimensions tend to converge. This assumption is backed by the overall silhouette score of both plots, represented by the red dotted line, which is around 0.45 in 2005 and 0.40 in 2022. We have now validated the hypothesis that the emergence of "small ball" is a true phenomenon, that has reduced the disparities between the NBA players in terms of height and weight. New NBA stars include some centers and point guards that are mostly different from those in 2005. It is often said that shorter players are better shooters but what is this assumption worth exactly? Let's examine advanced zone statistics in the league overall.

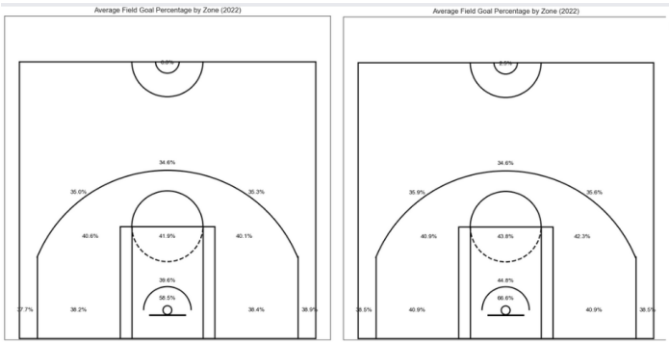

Figure 8. League average shooting accuracy by zone, in 2005 (left) and 2022 (right).

These results show that the shooting accuracy did not get better as a whole in the league, besides the slightly better accuracy close to the rim. A further analysis of the statistics showed that approximately 25,000 3-point shots were taken during the entire 2005 season, while we account for more than 65,000 in 2022. The similar accuracy can be since only 3-pointers specialists were shooting back 2005, the accuracy is brought up while the democratisation of 3-pointers in 2022 brought worse shooters to take these shots anyway.
These hypotheses could be the subject of an entire debate, but the next figure will bring an insightful visual regarding the shooting trends of NBA players.

All the shots taken by the 30 starting point guards and centers in both 2005 and 2022 were scraped, stored in a data frame and visualised.

4.3 Analysis Results

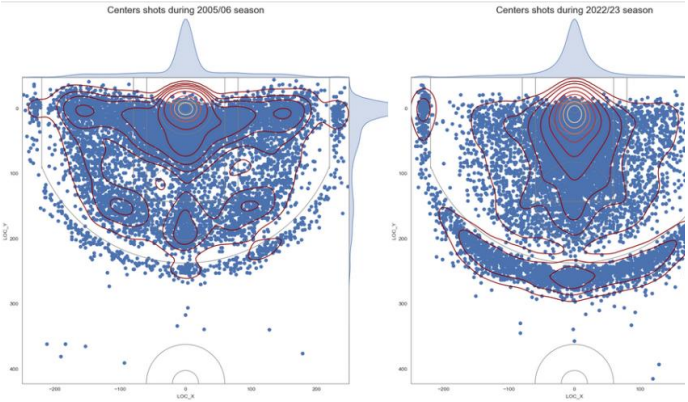

Figure 9. Heatmaps and distribution of starting Centers in 2005 (left) and 2022 (right)

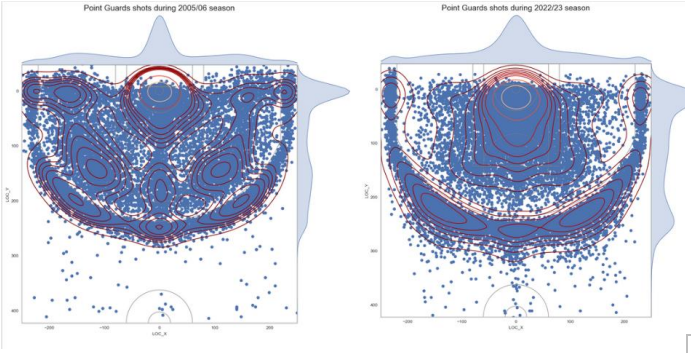

Figure 10. Heatmaps and distribution of starting Point Guards in 2005 (left) and 2022 (right)

These 2 heatmaps summarise and validate our general research hypotheses on the NBA.
The blue dots represent all the made shots while the circles come to show some concentrated locations, like a seismic magnitude. The centers clearly shifted from an insider role to a more versatile position, explaining them their growing likeliness to make assists. 2-point shots close to the 3-point line have been abandoned in favour of 3-point shots, which return a better PPA (Point Per Attempt). Point guards showed a strong tendency to shoot beyond the 3-point line in 2005 as expected, but it is to be seen that they still tend to shot from further and further in 2022, sometimes several meters behind the line while the 2005 shooters tended to stick to the line as much as possible. This, added to the evolution of measurements makes us think that basketball positions are far more versatile than before, and that an NBA star commonly attributed to a position can have a more diverse role and be effective on every aspect of the game.

5 Critical Reflection

The 2-times approach to this problem helped following a clear sequence of tasks in order to reach an answer to our research hypotheses. The analysis goes in order from the simplest basic tasks to the more complex visual analytics provided by the API, as shown by the diagram. The human intervention is needed to interpret the heatmap and read between the lines of the statistics displayed and some knowledge of basketball is almost mandatory to draw conclusions regarding the evolution of the NBA, and especially the key players. A more complex approach including other factors could be computed for K-Means clustering, such as including points, rebounds and assists then performing a PCA and get different results, but the overcomplication of the task in terms of posterior interpretability refrained us to do so. Choosing an API instead of a clear stored dataset is useful because the number of statistics is almost unlimited but is then hard to store properly because choosing which data is more suitable than another becomes almost impossible. A better research and scraping of the API can be considered, in order to deepen the analysis, especially on the visuals due to the high potential of the NBA statistics on this aspect. The complexity from the data used is good enough for analysts in order to draw some basic conclusions, but an extremely advanced analysis would probably require a bigger and more complex dataset, alongside further machine learning techniques, for whoever decides to quantify the actual impact of a player during a given matchup for example. The analysis of defence efficiency is another factor that was not dealt with because of its complexity, both to visualise because of the pace of a match and to quantify. I have myself personally learnt some things about the sport that I have always chosen to see on TV with my own knowledge but never visually and quantitatively verified, which was done throughout this report.

Table of word counts

Problem statement	200/250
State of the art	491/500
Properties of the data	351/500
Analysis: Approach	304/500
Analysis: Process	1160/1500
Analysis: Results	176/200
Critical reflection	326/500
Total	3008/3950

References

The list below provides examples of formatting references.

[1] K. Goldsberry, « CourtVision: New Visual and Spatial Analytics for the NBA », 2012.

[2] A. Benito Santos, R. Theron, A. Losada, J. E. Sampaio, et C. Lago-Peñas, « Data-Driven Visual Performance Analysis in Soccer: An Exploratory Prototype », *Front. Psychol.*, vol. 9, p. 2416, déc. 2018, doi: 10.3389/fpsyg.2018.02416.

[3] S. Zhang, A. Lorenzo, C. T. Woods, A. S. Leicht, et M.-A. Gómez, « Evolution of game-play characteristics within-season for the National Basketball Association », *International Journal of Sports Science & Coaching*, vol. 14, nº 3, p. 355-362, juin 2019, doi: 10.1177/1747954119847171.

[4] J. Lorenzo, A. Lorenzo, D. Conte, et M. Giménez, « Long-Term Analysis of Elite Basketball Players' Game-Related Statistics Throughout Their Careers », *Front. Psychol.*, vol. 10, p. 421, févr. 2019, doi: 10.3389/fpsyg.2019.00421.

[5] https://www.statista.com/statistics/193467/total-league-revenue-of-the-nba-since-2005/

[6] https://www.lines.com/guides/average-height-nba-players/1519

[7] https://www.kaggle.com/code/justinas/nba-height-and-weight-analysis

[8] Humaira, Hestry & Rasyidah, Rasyidah. (2020). Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm. 10.4108/eai.24-1-2018.2292388. J. Sewall, D. Wilkie, and M.C. Lin. Interactive Hybrid Simulation of Large-Scale Traffic, *ACM Transactions on Graphics*, 30(6), Article 135.