Data Visualization Practical Work

NBA Visualization Project

Universidad Politécnica de Madrid
ETSII

Data Science Master

Authors:
Hugo Enrile Lacalle
Alejandro Gouloumis Contreras
Joel Ramirez Moreno

1. Problem Characterization

Firstly, in line with the methodology explained in the theoretical classes, we will start with a first characterisation of the chosen problem, including a brief explanation of the working dataset and the users interested in a project like this. We have chosen to select a set of data from the sports domain, specifically from the basketball competition par excellence: the *National Basketball Association*, better known as ***NBA***.

The world of sport is one of the sectors that is benefiting most from data processing and the hidden knowledge that can be found in it, and the NBA was not going to be an exception. It is already a reality that there are teams of analysts whose function is to recognise patterns in different areas: types of players, scoring trends, team form...

In this case, a dataset has been obtained from the *Kaggle* repository, although previously this content was extracted from the *basketball-reference* data repository. The datasets include data of different types, grouped into five different files. Of these, special use will be made of *games_details.csv* and *games.csv*, which collect detailed information on the records of each player during all the games played between 2003 and 2020. The specific attributes that have been worked on will be exposed throughout the development of the project.

Among the great beneficiaries of this visualization work, we could find scouts, spectators interested in the performance and status of the different teams and players or coaches in the preparation of a certain match, without forgetting the fact that it is also a visualization tool that can be oriented to a more lucrative and entertaining activity. With the proposed tool, it would be possible to answer questions such as: "how has the professional career of a certain player been or is being over time?", "what has been the trend of his game in a certain set of seasons?" or "is this player an outlier in any discipline compared to the rest?".

2. Distribution of the visualization tool

Firstly, it should be noted that the application obtained has been divided into different tabs, in order to be able to answer different questions in a more concrete way and to better satisfy the needs of the user. For this reason, we will try to explain the different design choices made throughout the viewing process according to this division.

2.1. Evolution of Players

In this particular section, we have tried to answer a couple of questions formulated in the previous section: *"How is the trajectory of a player between X seasons?"* and *"Is this player a special piece (or an outlier) compared to the rest in a given field?"*.

Data and Task Abstractions

- <u>Data Abstraction</u>

As mentioned in the previous point, different sets of data contained in *.csv* files have been used, then the starting format of the data is in tables. As regards the items, they will be each of the different rows that are contained in that table, where the columns will refer to the measurable and comparable attributes that have been worked on. In particular, we will talk here about the *games_details.csv* and *games.csv* files. Regarding the first of the files, each of the rows or observations refers to a specific match, which is determined by an ordered attribute that acts as its identifier. This attribute will also be present in the other tables, which has been of special help for the combination of the files in one of a final file. Another important attribute would be the name of the player (*PLAYER_ID*), acting as a categorical attribute. From there, a series of quantitative attributes are available which are those that collect the individual contributions of each player according to different measurable aspects of the game in question: points, assists, rebounds, fouls... All these attributes are needed because we want to summarize the trajectory of the player, so we should consider every aspect that relates to this. The second of the files has been useful to be able to derive a new column (a new attribute) which, as a result of the match *ID*, identifies the season in which it was played, thus creating a temporary attribute. This will not be the only temporary attribute used, but all the matches, in addition to being identified by an ordered attribute (the *ID*), are also identifiable by the exact date on which they were played, thus establishing an order through this other temporary attribute.

- <u>Task Abstraction</u>

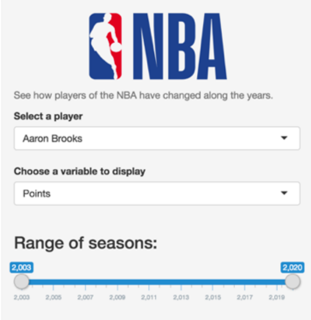
Firstly, a selected player will be treated and his evolution over time will be observed, which allows a great deal of information to be summarised, since the number of matches played over all these seasons is very high. Precisely, this summary of the selected data of each match makes it possible to extract useful information such as the trend of play and detection of outliers or matches where he excelled especially. Due to the great amount of available data, it was decided to present the answer to this question in a more concise way, allowing to compare performance between seasons and again making possible the search for outliers or trends. Finally, to see if the player is different from the others, a final comparison task is performed, which aims to put the performance of the player in question into perspective throughout his career with that of the total number of players in the competition. In summary, it could be said that the abstraction tasks carried out in this block make *exploration*, *comparison* and *summary* possible.

Interaction and Visual Encoding

Once the tasks to be carried out have been defined, it is only necessary to go to the data required to solve them and, depending on the type of attribute used, choose a visualization channel and marks that are appropriate to the nature of the problem, that is, define the idioms selected to deal with it.

The first of the aspects to be covered is that which refers to the exploration task. Specifically, we wanted to carry out a detailed analysis of the player's temporal evolution through his performance in the matches played during the established seasons. To select the player, the categorical attribute that determines his name will be used, which in this case has been ordered using an alphabetical order, which facilitates the task of locating him. As for the statistics of interest, the options are arranged as selectable, this being a quantitative attribute. Finally, it is possible to set the interval of seasons between which the result is to be observed, thus manipulating the view and filtering or adding more information as indicated (Figure 1). Thus, a graph is required to display a quantitative attribute (points, assists...) and an ordered attribute (dates of matches), opting for a line chart (Figure 2). This graph makes it possible to detect the trend of play and also to identify possible outliers by exploring the records stored in the database. Through this representation, the values of the quantitative attribute are codified through points, with a height determined by this value, while on the horizontal axis they are spaced according to the temporal attribute that orders them.

Figure 1

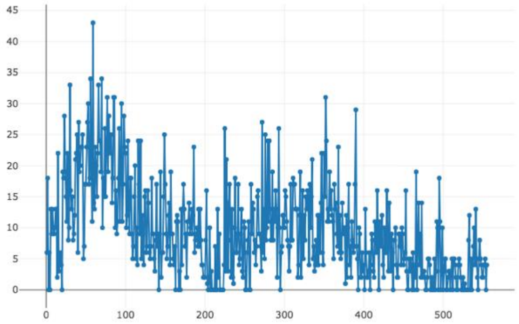


See how players of the NBA have changed along the years.

Select a player: Aaron Brooks

Choose a variable to display: Points

Range of seasons: 2,003 – 2,020



Another of the tasks to be dealt with is that of summarising information. Depending on the specific needs of the different stakeholders, it may not be necessary to have such a detailed arrangement of the records but rather something more concise that allows an idea to be extracted at a glance about the best season, outliers, trends... but more generalized (Figure 3). For this reason, instead of having the time attribute on the horizontal axis referring to the matches played, another time attribute associated with the seasons themselves has been selected. In this case, it was decided to choose the bar chart, coding on the vertical axis the quantitative attribute of the average per season as the height of the bar and spacing these bars horizontally according to the values of the ordered attribute. In addition, to make the visualization more efficient, it has been chosen to use not only a spatial channel but also a colour channel to denote each of the levels of the season attribute.

In order to complete the tasks of the previous section, a final graph is needed that allows a total comparison of a player with respect to the rest. Given that the comparison is made with respect to all the quantitative attributes, a simple scatterplot would not be enough, as its nature limits it to only 2 attributes, so the solution is to choose a graph of parallel coordinates (Figure 4). The coding consists of as many parallel lines as quantitative attributes where the points are given by the value in question of each attribute. In addition, two lines are arranged, one referring to the average of the competition and the other to the selected player. To continue with the interaction that characterizes the application, a small bar is enabled that allows you to select which line you want to highlight (*pop out*) and be able to better appreciate the selected record.

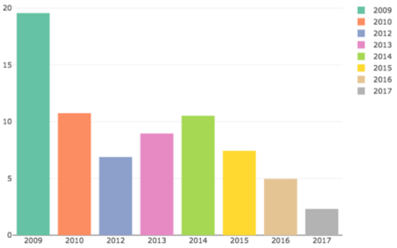


Figure 3

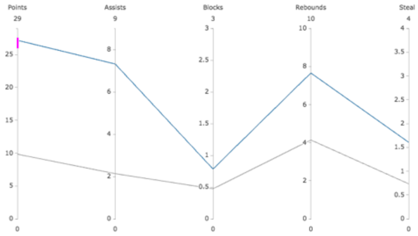


Figure 4

2.2. Compare players.

In this section, the questions about *"How has been a player doing compared to another"* or *"How have their average statistics affected the performance of the team"* are answered.

Data and Task Abstractions

- <u>Data Abstraction</u>

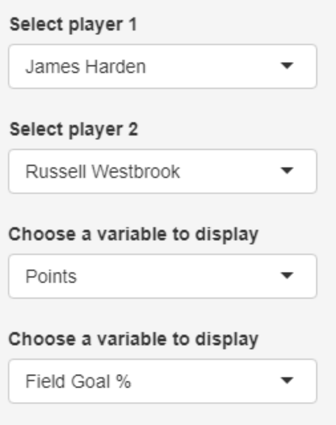
The same tables mentioned in the previous section were used. New attributes were derived: mean aggregate values for the quantitative attributes Points, Assists, Rebounds, Blocks and Plus Minus, keyed by the ordered attribute Season and the categorical Player Name. The resulting items are the average statistics for every player on every season they played. Also, another two quantitative attributes were derived, but instead of averaging, summing: the total number of Field Goals and 3 Point Field Goals Attempted and Made. This sum allows us to extract the shoot accuracy percentage, keyed by each Player and Season.

- <u>Task Abstraction</u>

First, the two selected players evolution over the seasons is observed. This allows us to compare the evolution of both over time, looking for trends and outliers. For example, seasons where the statistics were especially good or bad, as well as the improvement or decadence of the players. The shots accuracy percentage is also compared, and this is especially important because a good selection of the shots is translated to better performance, which is related to the last task: performance comparison. This is an indicator of the players importance to his team. We can measure their contribution to the overall performance of the team. The main task abstraction carried out is *comparison*.

Interaction and Visual Encoding

In order to select the players and statistic to display, a similar menu is displayed as in the previous section, but in this case two players are chosen (Figure 5). This allows us to locate and filter by players and statistics. Each of the players are differentiated by its colour and all the statistics are summarised for every player and season, as mentioned before.



Select player 1: James Harden

Select player 2: Russell Westbrook

Choose a variable to display: Points

Choose a variable to display: Field Goal %

Figure 5

For comparing their evolution over time, a line chart was chosen. This idiom allows us to express the ordered attribute over the horizontal axis (Season), separated for each category, and the qualitative over the vertical one (Points, Assists, Rebounds or Blocks) (Figure 6). We can detect changes in the slope and see how the trend varies with the Season. This translates into player improvement or change of playstyle and see how each player plays a different role on his team (scorer, assister, rebounder, etc.) over the seasons. The tool also allows us to make zoom, screenshots or select points on the line, where the information of the values and keys are shown.

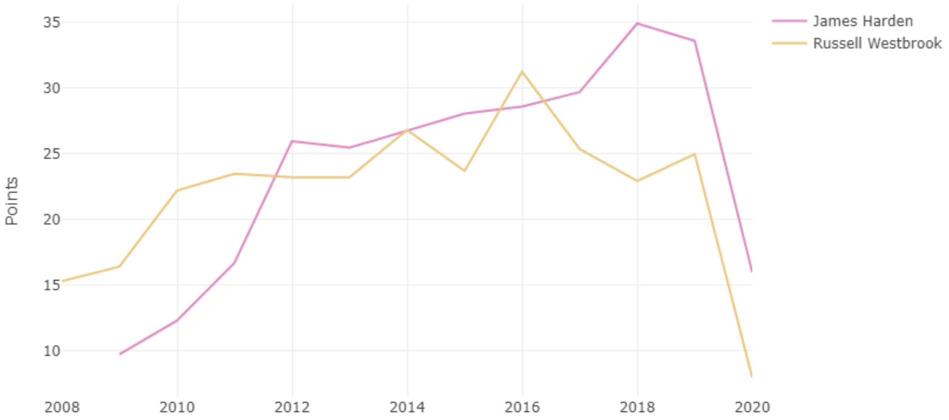


Figure 6

Another way to compare the players performance over time is the shots accuracy percentage, as mentioned before. This can be compared through a stacked horizontal bar chart, where the quantitative attribute (Field Goal % and Field Goal 3 point %) is on the horizontal axis and the ordered attribute (Season) is on the vertical axis, separated for each category (Figure 7). This idiom allows us to compare each players shot accuracy for every season, over the seasons. Also, trends can be found but on a less intuitive way than with the line chart.

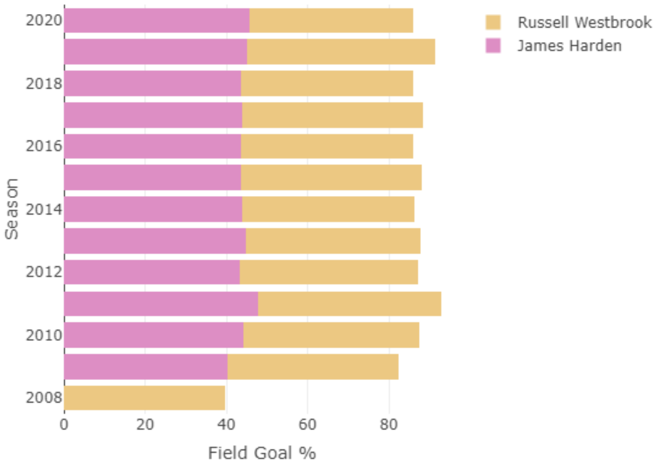


Figure 7

To fulfil the whole comparison task, it is necessary to compare the Plus Minus average of the players for each season. A mentioned before, this represents the players contribution to the performance of the team. A diverging stacked bar chart was chosen, because the Plus Minus metric can have positive and negative values (positive meaning good contribution, negative meaning bad contribution). The quantitative attribute is expressed on the horizontal axis (Field Goal % and Field Goal 3p %) and the ordered attribute over the vertical axis (Season) (Figure 8). This allows us to see on a more efficient way how the players performance goes close or far from the 0 axis, meaning higher or lower contribution to the team.

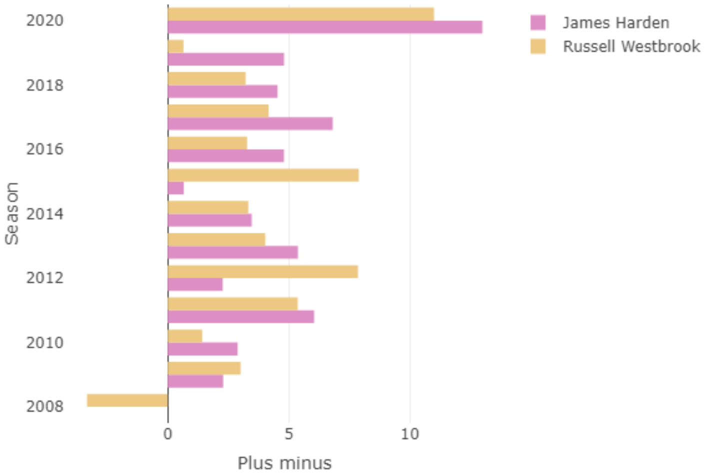


Figure 8

2.3 Clustering of Players

Finally, in this section the following questions are answered: *"Could a pattern be identified in the overall players?"* or "*Could the players be classified depending on their position?"*.

Data and Task Abstractions

- Data Abstraction

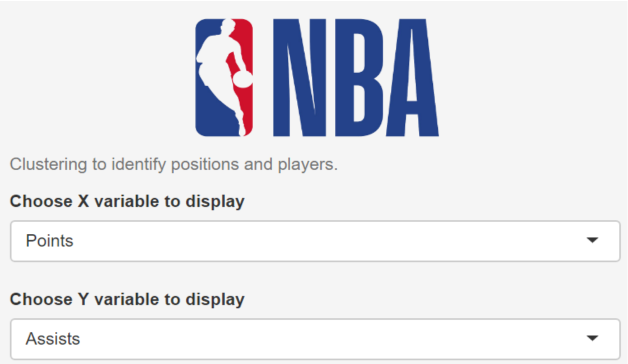
As mentioned in the section before, several tables were used obtain a richer and more accurate clustering. The attributes used to this section were: Points, Assists, Rebounds, Blocks, Free Throws made, Field Goals made and Steals. Once gathered all these stats, mean value of each one of them is computed, grouped by each Player present in all the seasons that have played.

- Task Abstraction

Firstly, and once the clustering method (K Means) is completed, a graphical representation of the clusters is presented, enabling the viewer to compared the different distributions within the clusters, each one of them try represent a position on the field (point guard, shooting guard, small forward, power forward and center). This graph is manipulable, so the different options mentioned before are available to get a better insight about the identification and/or classification of the players, comparing different situations and stats.

Interaction and Visual Encoding

Firstly, different stats are selectable, as mentioned before, those are: Points, Assists, Rebounds, Blocks, Free Throws made, Field Goals made and Steals, to compare all the player under different circumstances.



Clustering to identify positions and players.

Choose X variable to display

Points

Choose Y variable to display

Assists

Figure 9

The first graphical representation presented in the tab is the clustering representation of the algorithm (K Means). Thanks to this graph a clear and simple representation of the different positions and players can be seen, enabling the viewer to interpret the information easily and in a fast way. Five different clusters are represented, each one of them trying to replica the five possible positions a basketball player can play, with each of the combinations possible with the variables represented, a viewer can obtain a deeper analysis in how a random player could be placed in a cluster, for example, if the quantity of points made and the steals performed is high in both cases, the usual representation of an offensive player as shooting guard, should be placed in the rightest side of the graph.

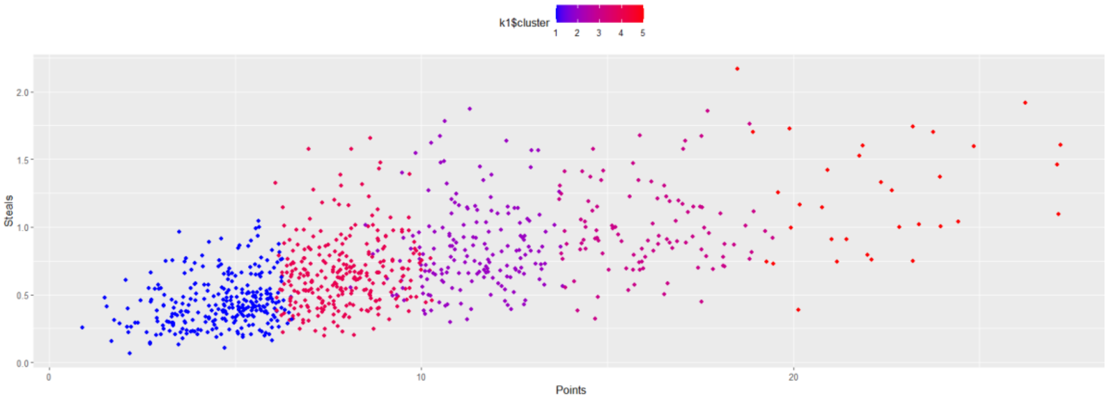


Figure 10

A different representation of the cluster is also served in the lower left side, enabling the viewer an even easier interpretation of the data, here a contour is drawn between the edges of the clusters, facilitating the visualization of the separation between different clusters, as so, a more accurate representation of a player can be recognised and easily tracked. This is a more general representation of the clusters rather than the explained before due to its impossibility of change the variables, as so, a more general insight can be obtained with this graph, and a general pattern could be found with this data.

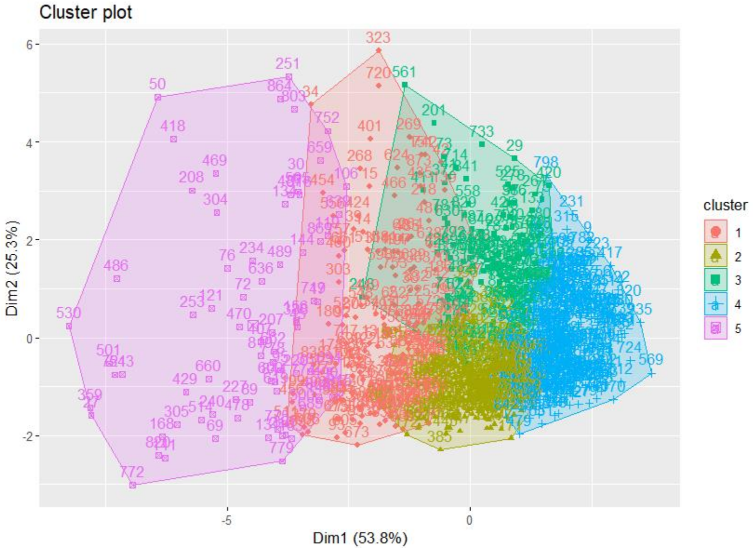


Figure 11

Finally, the last graphical representation is a dendrogram, this helps to visualize the decisions that have been take until a clear separation between the clusters is achieve, with this tree-like graphical representation, an easily interpretation of the general clusters is possible (obviously, due to the great amount of different players present in the dataset, is not possible to track unique players), but this helps to obtain a different vision of the representation, giving the opportunity to the viewer to try to identify factors and values that could lead a specific player into one or another cluster.

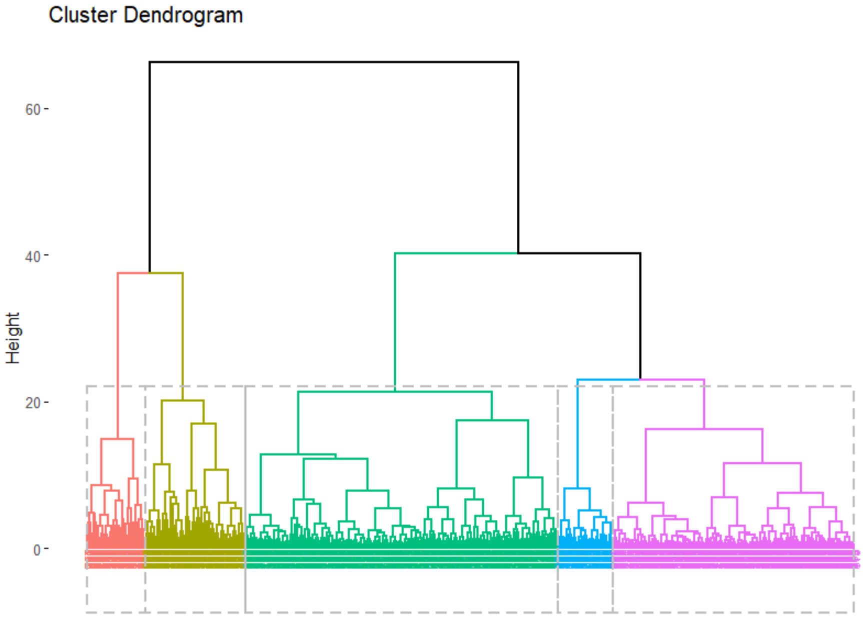


Figure 12

3. Final considerations

Some final considerations should be pointed, the clustering tab, usually takes up to 2 minutes to finish its representation due to the size of the dataset. Also mentioned, the application have been submitted to ShinyApplications under the following url, in case you like to see it: *https://agouloumis.shinyapps.io/NBA-visualization/* also, the github with the application and source is: *https://github.com/hugo-enrile/NBA-visualization*