Delft University of Technology

Data Visualization 2018
IN4086-14

InfoVis Project Group 31

Authors:
Achilleas Vlogiaris (4875974)
Lisanne Hupkens (4283171)
Rafail Skoulos (4847482)

December 19, 2018

Contents

1 Introduction

The National Basketball Association (NBA) is the professional basketball league in North America since 1946. This league consists of 30 teams split up in the Western and Eastern Conference. These teams are playing 82 games in total per season[1]. The NBA keeps already record of data since 1946. Besides the simple data, such as the point counts, the NBA made also complex metrics. An example of a complex metrics is the PER. These statistics are introduced to study basketball in a objective way[2].Unfortunately, the NBA does not present the data in a user friendly way (see Figure 1[3]). The many different matches, teams, players, and metrics makes the data unclear. This report will show the process of visualizing a part of the data of this domain in a more efficient way and provide a tool to individuals who want to use these data.
In the National Basketball Association, the General Manager or GM of a team typically controls player transactions and bears the primary responsibility on behalf of the team during contract negotiations with players. The general manager is also normally the person who hires and fires the coaching staff, including the head coach. The exact title and responsibilities held by a general manager can vary from team to team.
The main concept of this project is to provide a valuable tool to the General Managers of NBA, this tool will help the GMs to make more targeted decisions and increase the performance of their teams.

[1] https://www.nba.com/news/faq
[2] https://stats.nba.com/help/faq/
[3] https://stats.nba.com/scores/12/05/2018https://stats.nba.com/scores/12/05/2018

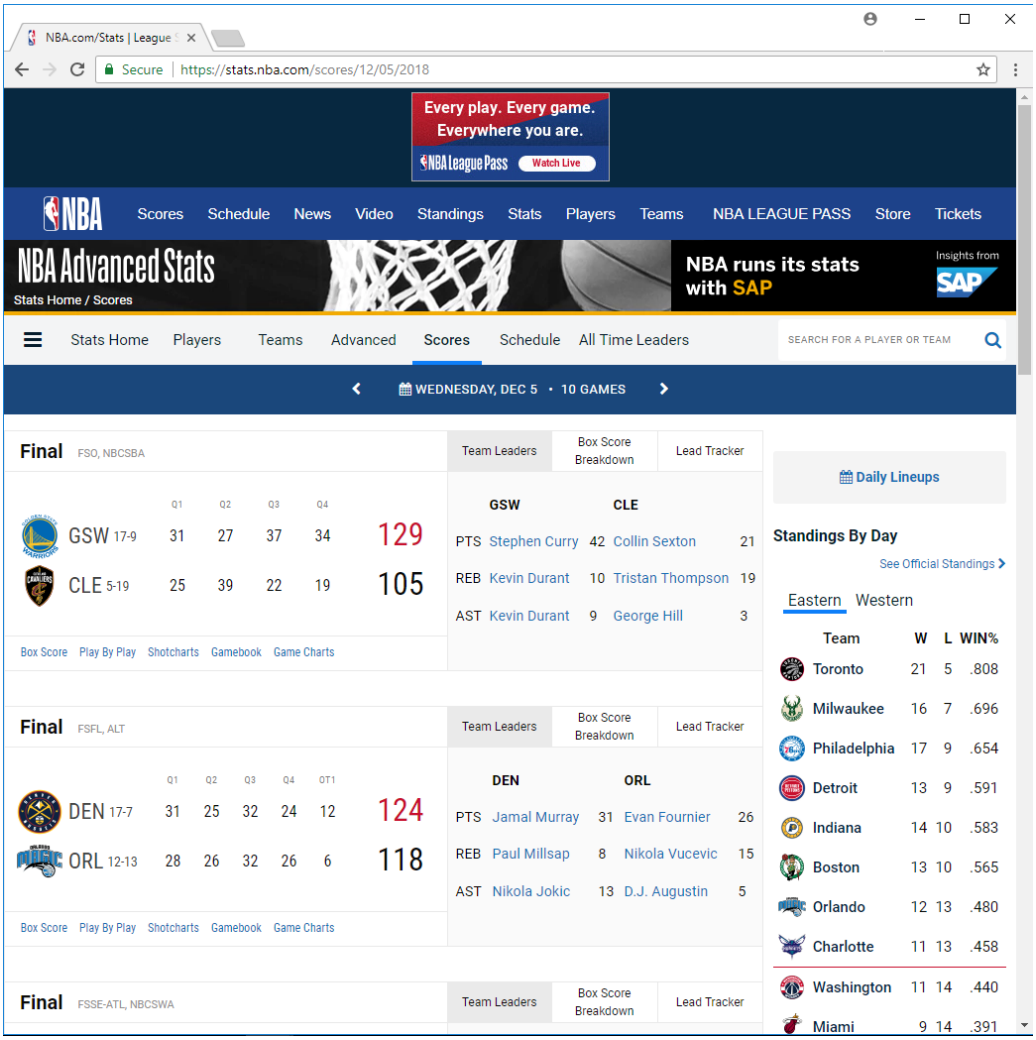

Figure 1: The existing tool of the NBA is not clear

2 The Data

2.1 What is the data about?

The data is about NBA (National Basketball Association) which is the basketball league of USA, the best one in the worlds. Especially, our data contains statistics for each team and each player of the league. These statistics can be used to measure the performance of the teams and the players. The statistics contained are points, assists, rebounds, turnovers, the percentage of shots made and many others which can be found there along with their description. Next to the simple metrics, the more complex metric Efficiency is also included. The Efficiency is used as a metric to measure the player's or team's overall performance, both in offense and defense, and it contains not only the "positive" achievements, but also their wrong actions.For example, in general a efficiency greater than 20 for a player is considered as good. Finally, the salary of each player for a specific year is included.

2.2 What is the level of complexity?

Due to the popularity of the NBA and the attention they pay to every single statistic of each player,a great amount of data can be found about them. As a result, for every player and team, hundreds of different statistics can be found[4]. This makes the data complex and not transparent, thus characterizing a team or a player as good or bad requires the extraction of the correlation between the statistics.

[4]https://stats.nba.com/help/glossary/

2.3 Preprocessing

Because of the plethora of different statistics that are recorded about each NBA game, many datasets can be found online. However the datasets about the last season (2017-2018) were not sufficient, so we decided to use data until the season 2016-2017. The implemetation of the charts chose required data that were not included in one dataset, so the most important task we did on the preprocessing of the data was the combination of three datasets.

The first dataset[5] contains the statistics for each player from 1950 until 2017. We choose to keep only the ones for 2007 and above in order to have results that are representative. Also modifications had to be done so as to fix the format of the names (contained abbrevations) and the dates, and calculate the per game statistics as it contained the ones for the whole season. Furthermore, the efficiency metric has to be calculated for each player by combining his components.

The second dataset[6] contains the salary of each player from 2000 to 2019. As in the first dataset, we keep only the salaries from 2007 to 2017. There we had to group the players by their team and by the year, so as to extract the salary of the whole team for every season.

The third one[7] is about the win percentage of each team for every year, which did not demand any preprocess.

For each of the graphs we implemented, we had to make different combinations of each of the datasets above. For the processing of the datasets (CSV files), we used the *pandas* package of Python programming language, because it makes it easy to manipulate that type of files and also we had experience in using that.

3 Data Analysis Task

3.1 Why is this data interesting?

This data could be used in many different ways. For example, the coach of a team could use it to adapt the game strategy against another team. Moreover, a team manager can use it during the recruiting of a new player. Another application would be to use it by sport betting. By all these activities it is advantages if this data is can be use to make a objective decision.

3.2 What questions can a analyst answer with this data?

The tool will help to answer following questions:

1. **What is the performance of each player through the years?**
   Is the performance of a player fluctuating, steady, rising or descending? During the requirement of a new player it is not only important to know the player metrics of the last season. You should understand the past, to predict the future.

2. **What is the correlation between efficiency and the win percentage of the team over the years?**
   Often there has been said that there is a strong correlation between the player efficiency rate and the win percentage of a team per season [8]. But is this really true for all the years? And if this is true, what statics are making the outliers a outlier? In this way a coach can know if he can rely on the PER metrics.

3. **What is the correlation between salary and the win percentage of the team over the years?**
   The salaries of NBA players are without a doubt super high. Some players are earning more than thirty million dollar per year[9]. It is important to understand the influence of the salary on the success of a team, because it can be used in the negotiation process during the recruitment of a new player.

4. **How can we evaluate the salary of the player in comparison with his performance ?**
   The salary of NBA players are currently skyhigh. But the success of a player depends on many different factors. The large number of players and attributes will be summarised in one tool. This tool could help a coach to by recruiting a player from another team. We are interested about players that get more than 5 million dollars and have positive efficiency.

[5] https://www.kaggle.com/drgilermo/nba-players-stats
[6] `https://www.kaggle.com/hrfang1995/nba-salaries-by-players-of-season-2000-to-2019`
[7] https://www.kaggle.com/druswick/nba-team-records-historic
[8] `http://www.espn.com/nba/columns/story?columnist=hollinger_john&id=2850240`
[9] `http://www.espn.com/nba/salaries`

4 Interactive Visualization Techniques

Building the tool The code is build with the help of HTML, CSS, JavaScript, Python and the D3 library. The code can be found in the attached folder. A screen-cast that shows the final result together with commentary can also be found in this folder. The tool is existing out of: (1) A tree to select the teams and the players. This is a tool to make the navigation between the long list of players more easy, see Figure 2 (2) A stacked bar chart to evaluate the performance of a player through the years, see Figure 3. (3) A scatter plot to find the correlation between efficiency of salary and the win percentage, see Figure 4. (4) A Line graph yo show the trend of the efficiency, salary or the win percentage of years, see Figure 5. (5) And a box plot to see the distribution of the efficiency per team, see Figure 9. These separate graphs are explained in more detail in the sections underneath.

Figure 2: The collapsible tree that shows all the teams and players in a clear way acts as a selection tool.

1. **What is the performance of each player through the years?**
   The stacked bar chart was chosen to visualize the data that is applicable to the answer of this question, see Figure 3. The different seasons are the ordered key attribute and placed on the x-axis. It could be seen inappropriate to use a bar chart instead of a line graph, because the seasons are sequential. This choice was made because in this way we can visualize the summation of the attributes, something that couldn't be visualized sufficiently with a line graph. The layers of the chart are representing different types of positive offensive and defensive actions per player. Therefore they are the categorical key attributes. The colours of the key attributes are chosen in such a way that they are easy to distinguish, even if you suffer from any kind of color blindness whatsoever[10]. To make the colors less intensive there was chosen to turn down the opacity again. The countable number of these positive accomplishments is the quantitative value attribute. In this way it is possible to see the contribution of a positive player's action to the overall score in one overview. So, this chart can help to see the performance trend of a player concerning both the overall score and the specific attributes.

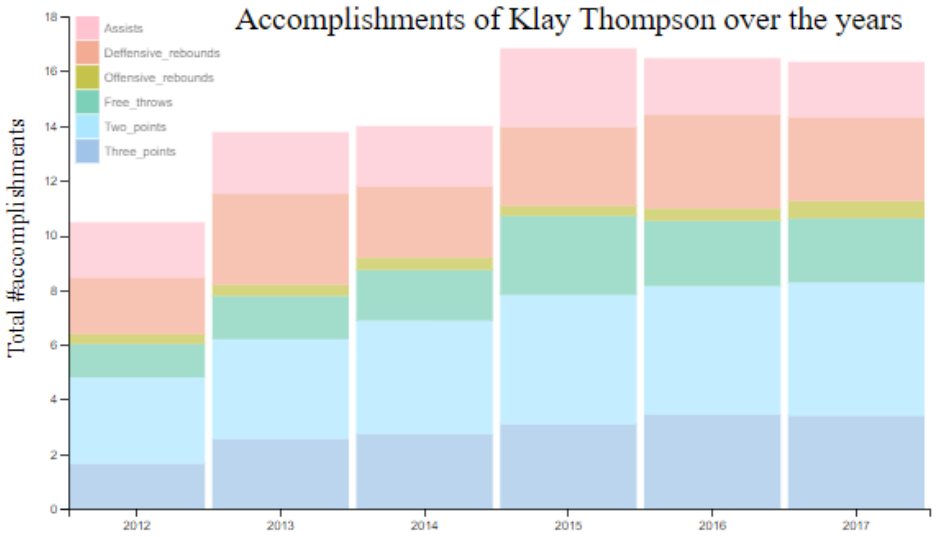

Figure 3: This stacked bar chart shows how specific attributes contribute to the entire performance of a player

[10] `https://personal.sron.nl/~pault/`

2. **What is the correlation between efficiency and the win percentage of the team over the years?**
A scatter plot was the logical graph to answer this question, because this graph can show both a correlation and outliers, see Figure4. The two quantitative attributes placed on the axis are the player efficiency rate and the win percentage per team. Every dot in the scatter plot represents a team. The team can be found by hover over the dots. By using the drop-down menu the year that will be analyzed can be chosen. A delightful transition is added to the change between years to make it possible to follow the development of a team. There is chosen to use for every team another colour to separate the teams. Some dots were hard to recognize in the beginning, because some dots were overlapping. This problem has been solved by turning the opacity down and adding a black outline. In this way the overlapping dots will look more dark. An increase in radius when hovering over a dot, makes it even more easy to select the right team.
Moreover, the change of the PER and the win percentage over the years can be found back in the line graph that is coupled to the scatterplot, see Figure 5. By clicking on a dot of a team in the scatterplot, the statics of this team over time will be visible in the line graph. The years are placed on the x-axis in a fixed way. With the help of a drop-down menu the PER or the win percentage per team can be set out over time. In this way both trends can be show by the use of only one graph. These trends are emphasized by adding a line between the measured data points. The data points are having a darker color to indicate the importance compared with the assumed line between the data points. Next to these two graphs, a boxplot is added to the whole. This boxplot shows the distribution of the player efficiency rate within one team of all the teams next to each other. The colors of the dots in the scatter plot are corresponding to the colors in the boxplot. This boxplot will help to understand the team in more detail in a condensed way.

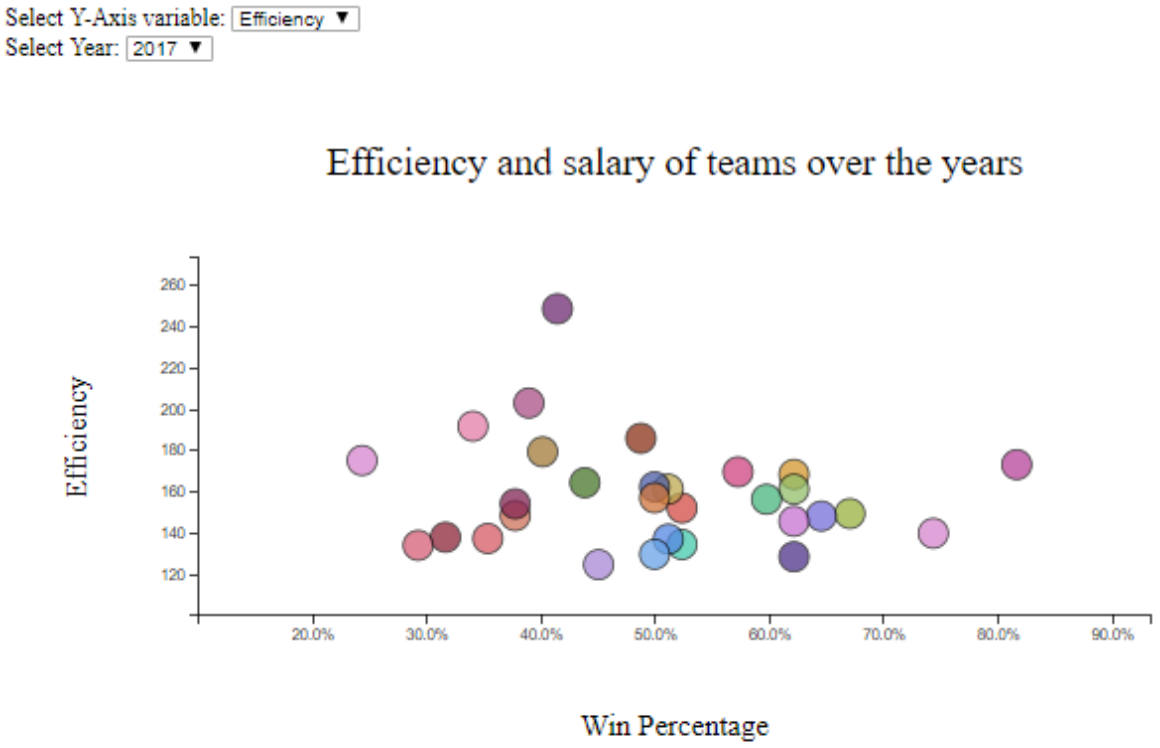

Figure 4: This scatterplot shows the correlation between the PER or the salary against the win percentage per team for several years.

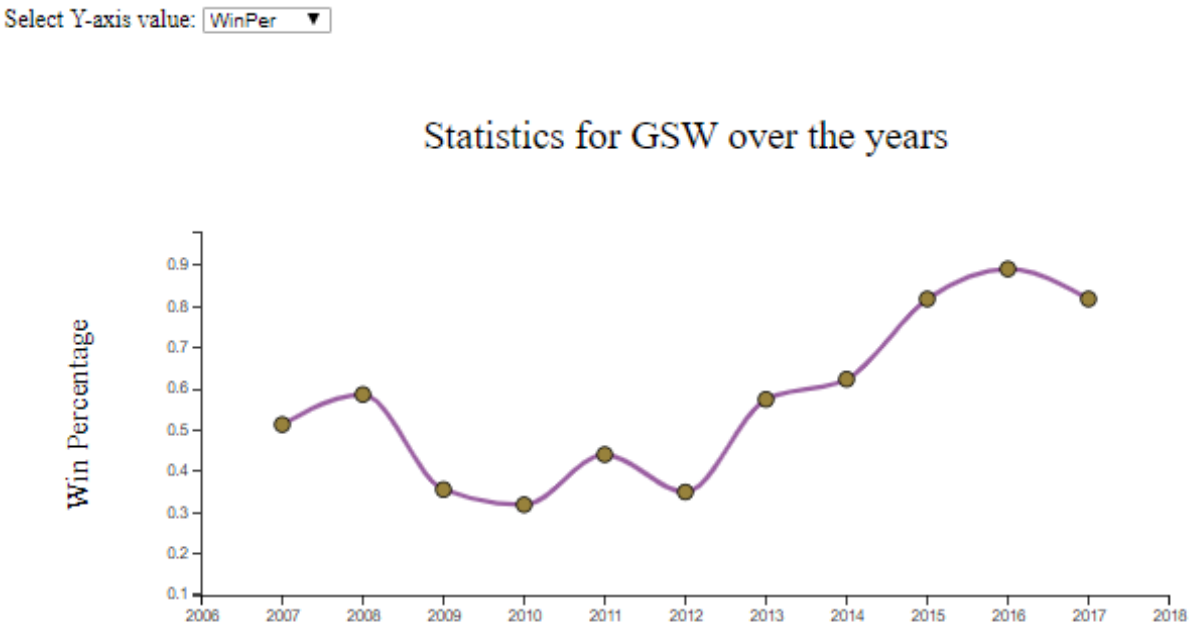

Figure 5: This line graph shows the trend PER, the salary or the win percentage per team over the years.

3. **What is the correlation between the salary of and the win percentage of the team over the years?**
   The metrics that are involved by answering this question are similar to the metrics of the third question. That is why there is chosen to incorporate this question in the scatterplot and line graph. By the use of the drop down menu the salary attribute can be chosen to analyze, see Figure6. It can be seen that there no important correlation in the salary of the teams with their win percentage. For example, there are teams with small salary that have more wins than others with greater salary. However, from the Figure 7 and 8 it easily observed that from GSW team the win percentage increased the last 2 years that their salary increased. The controversial results can be explained by the fact that many players that have players with small contract for many years, start having good performance in the middle of their contracts years so they should wait until the contracts expires to get a better one. So players with high performance have low salary and that affects the correlation between the salary and the wins of a team.

   We also choose to use a box plot so it can be observed how the distribution of salaries of the players affects the win percentage of a team. From the Figure 9 it can be seen that there is no specific pattern in the salary of the players that affects the win percentage of the team. Neither the teams that had smaller range of salaries nor the ones with greater range follow a specific trend. For example ORL and GSW has almost the same range of salaries but the former was in the 26th place while the latter was 1st.

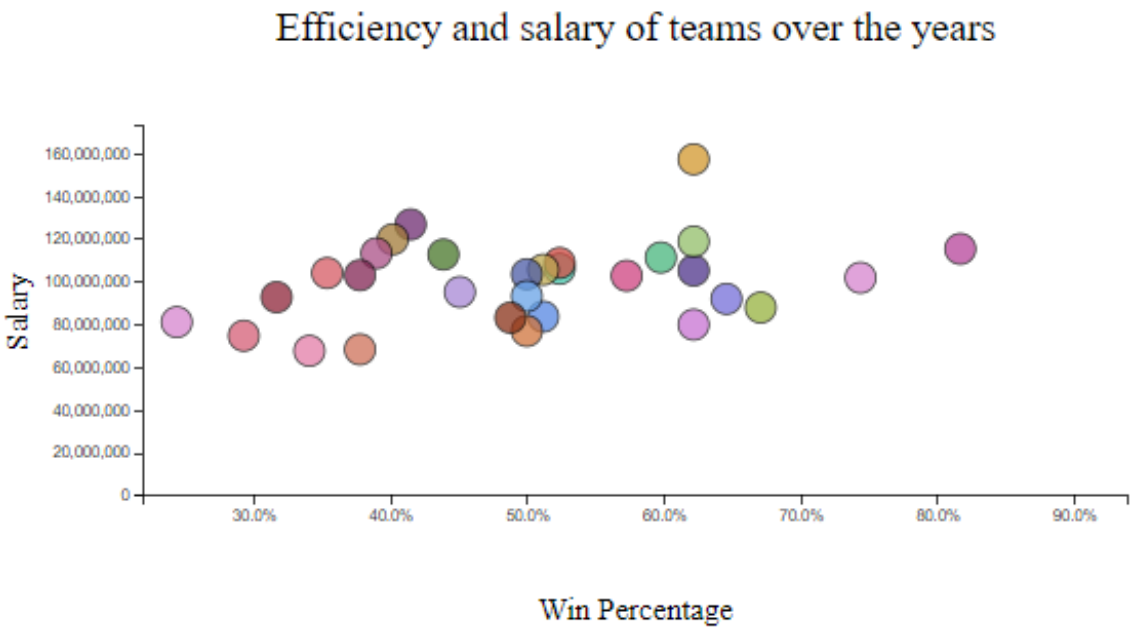

Figure 6: This scatter plot shows the distribution of the salaries for each team in the season 2016-2017, according to their win percentage.

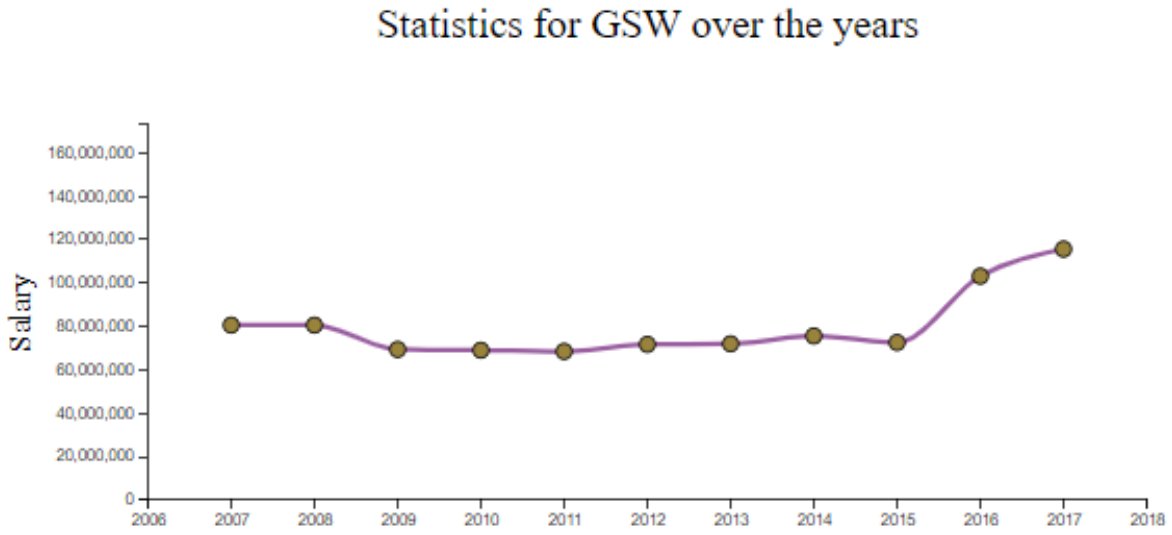

Figure 7: This line graph shows the salary for the selected team through the years

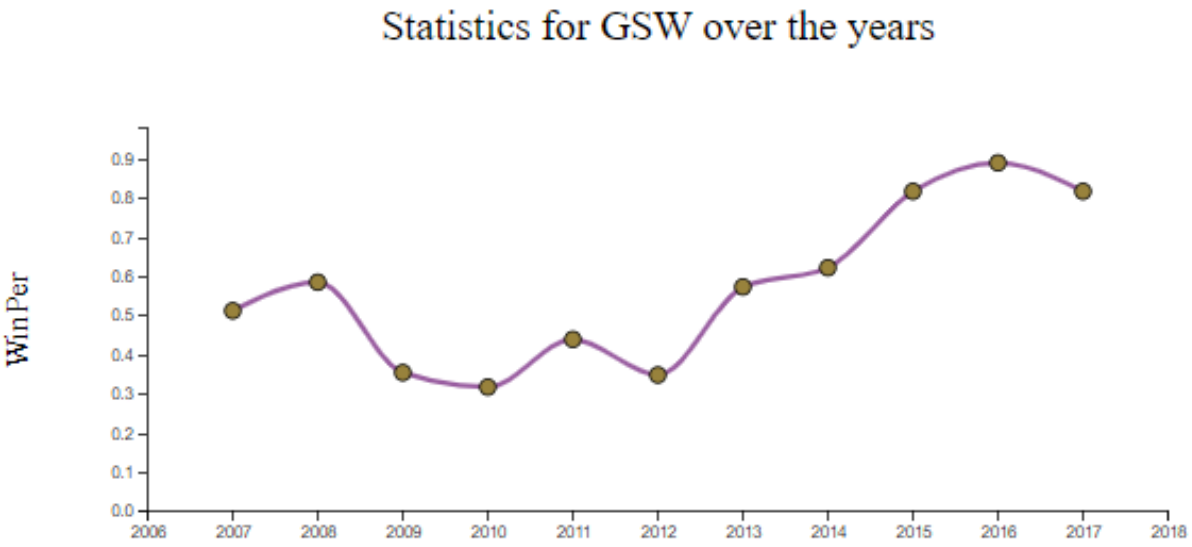

Figure 8: This box plot shows the win percentage for the selected team through the years.

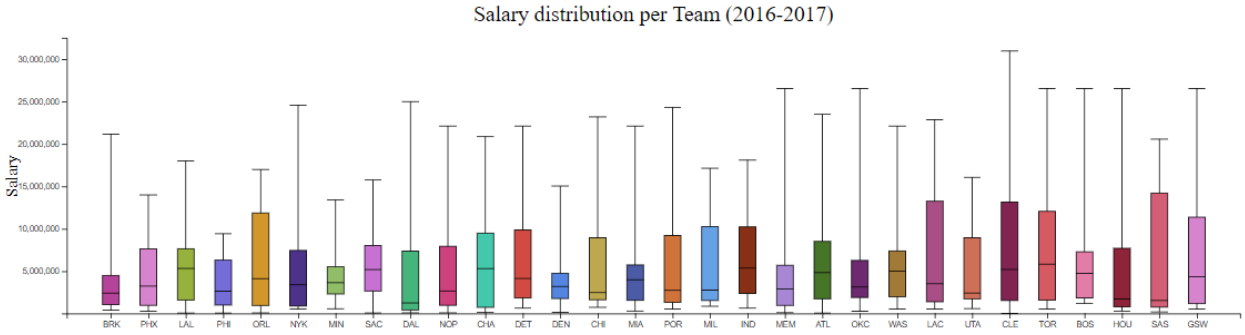

Figure 9: This boxplot shows the distribution of the Salaries for each team in the season 2016-2017.

4. **How can we evaluate the salary of the player in comparison with his performance ?**
   Both the efficiency metric and the salary statistics are used to evaluate what the most valuable players. This is done by dividing the efficiency by the salary, this metric isn't a common metric that somebody can see in an NBA graph. We chose this combinatorial metric to distinguish the most valuable players. A bar chart is chosen to visualize this answer to the question, see Figure 10. Firstly We suppose that the GM will choose a specific team according the assumption that he will make from the scatterplots 4 and the boxplot 9. More specifically for the bar chart, each bar represents different player, the players are grouped per team. The selection between teams can be made by using the selection tree tool.
   As we said the quantitative attribute is the efficiency divided by the salary, and can be found back as the length of the bar. Due to the fact that the efficient is a small number and the salary a large number the values are normalized so we can get a more reasonable and comparable value. The overall average of this value is represented by the horizontal line. This means that if a player is relatively cheap for his performance he will pass this line with his bar. Only a selection of players per team are shown to make the tool more clear and targeted. The selection of the players are based on their performance. Only the players whose their efficiency is positive get visualized. After all, we can say that only these players must be recruited. Lastly, because of the large surface of the bars, there is chosen to make the colour less bright by turning down the opacity.

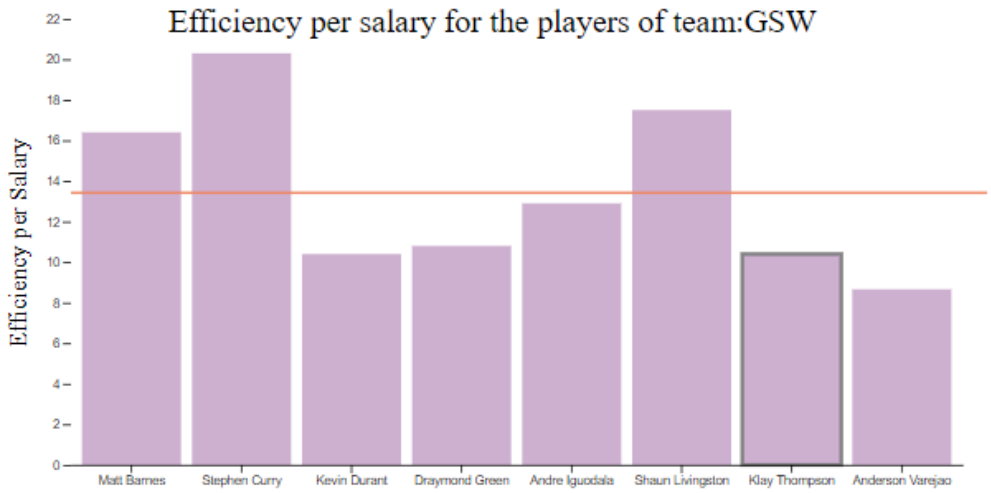

Figure 10: This bar chart shows the ratio between the salary and the PER for each player.

5 Results and Evaluation

This tool is a major improvement compared with the data visualisation that is provided by the NBA. In this way the data is shown more user friendly and the navigation is done more easily. Although these manny improvements, there are also recommendations for further development. The most important issues that should be reconsidered are:

- The choice of the stacked bar chart above the stacked line chart: A line chart can better visualise the fact the seasons are ordered in a sequential way.
- The drop down menu of the line graph: There was chosen to show all line graphs with in one graph. This was done to save space. However for the analysis of the data it could be more convenient to show the graphs underneath each other. A verticale pointer could be used to highlight the same time in every graph.
- Checkboxes by the stacked bar chart: If you want to recruit a player with a specific skills it would be nice to be able to hide the attributes that are not of interest. In this way the tool is more goal oriented.
- Furthermore we don't visualize players with low efficiency and the graphs that we made can provide valuable information only about players that a GM could buy and not to sell.

6 Conclusion

Since a very high number of NBA team and players exist in the league, it is very challenging for a GM to choose specific players to buy. Looking at our visualizations a GM can make several interesting observation about the teams and the players. For example, we can observe correlations about the teams between efficiency and wins, salary and wins, the efficiency of a team over the years or the route of the budget that a team spent over the years. Furthermore, we can make observations about each players of each team and see if a player get payed too much in comparison with his efficiency. In addition we can notice if a player with high or low efficiency per salary performs better or worse through the years. With these observations a GM can make the right choices. As we mentioned in the introduction most general managers don't care about players with big salaries or rookies with very low salaries, the selection of a new player is very important for a NBA team but our tool can't provide any help. For the selection of the new players a mechanism already exist (the NBA draft[11]). We can assume that the most valuable players are the players that aren't rookies neither super stars. The general manager of a

[11] https://en.wikipedia.org/wiki/NBA_draft

team will need a player who can make an impact to the whole team without spending a fortune and probably that player can be easily observed from our visualizations.

7 Individual Reflection

Generally through the design and the decision-making phase of this project we all worked together. When we knew what we wanted to do and how we would approach it, we decided that each of us should work individually on a specific task and we would help each other as it became necessary. However, we met at least once per week in order to keep track of what the others were doing, and also to solve problems that may have appeared. All of us wanted to deal with every aspect of the project and that is why each one created a different visualization. Lastly, we would like to note that although each of us wrote a specific part of the report the content was approved and decided by all of us.

7.1 Rafail Skoulos

The first task which was assigned to me by the team is the prepossessing of the datasets. Because of my prior experience in python and pandas package especially, I was able to do all the necessary modifications so as to provide the proper data for the implementation of the graphs. The other tasks were the implementation of scatter plots, line graphs and the box plot. At first I have to communicate with my teammates and decide how all graphs will be connected so as to be in the same line and be able to provide a satisfying result. For both of them, I firstly had to implement them in their basic form by using d3. Then, I had to make the dropdown lists so as to change the y-axis variable and the year (for the scatter plot and line graph). Also I had to combine the line graph and scatter plot so they can have the proper interaction. Finally for all of them I had to choose the appropriate colors so they can support their results.

7.2 Achilleas Vlogiaris

The main part that I had to implement was the bar chart and the stacked bar chart. As we mentioned before the main concept is that a General Manager makes some significant assumptions about which teams had good performance in the previous years, then the GM will be able to select from specific teams players with high efficiency per salary. That's the reason I made these two graphs. Furthermore I had to filter the dataset and get all the players with salary over 5 million dollars and also players with positive efficiency. The second graph appears if we choose a specific player from the bar chart or a specific player from the selection tree and the second graph shows the performance of the player through the years. In addition, I participated in the process of combining the two bar graphs(bar chart and stacked bar chart) with the selection tree, first for the selection of which team is going to be visualized from the bar chart and second for the selection of the player that is going to be visualized on the stacked bar chart.

7.3 Lisanne Hupkens

My background is not computer science, but Industrial Engineering and Biomedical Engineering. Besides the couple of times I worked with Matlab, my experience with coding was zero. This made it difficult for me to keep up with the high level of programming. I made an attempt to build the selection tree tool. I discovered that it was not hard to find a good D3 template, but it was hard to connect the data to the template in a proper way. The data we used had another format than the format needed. My first reaction was to edit the data by hand in excel. Luckily someone pointed me out that this was a really bad method to edit the data and introduced me to Pandas of Python. The tool worked in the end after the use of Pandas on its own. I should practice more with coding, but my interest in data visualization and its possibility is definitely aroused.