Final Project Report

Jieming Wei, Jiushuang Guo, Juexiao Wang, Yijun Wu

November 30, 2016

Abstract

This project examines the relationship between the performance and salaries of basketball players from the National Basketball Association (NBA) in the 2015-2016 season. To visualize the different statistics, we have created a shiny app that displays a horizontal bar-chart (one bar per team). Our preliminary finding is: **the team's' performance in general is closely related to the total salary they give out to their players.** Besides, in general, different positions do not have an obvious effect on the correlation between skills and salary. We also provide several interpretations to this result at the end of the report.

Introduction

With the data obtained from **Basketball Reference** by **Sports Reference LLC**, this project studies the relationship between the performance and salaries of basketball players from the National Basketball Association (NBA) in the 2015-2016 season. The shiny app is used to help us visualize different statistics.

This report summarizes the statistical analysis results associated with the basketball players from the National Basketball Association (NBA) in the 2015-2016 season. The purpose of this project is to document both the implemented data collecting, data cleaning and all corresponding modeling and inference techniques used during the statistical analysis. And it presents the result and comments of our findings. By analysing the data obtained from Basketball Reference (by Sports Reference LLC) regarding the salary and performance of NBA basketball players, we intend to examine whether or not the NBA players' performances are related to their salary. If so, how closely or in what pattern are the two factors "performance" and "salary" is related.

The final project involves analyzing data about basketball players from the National Basketball Association (NBA) League in the 2015-2016 season. The central topic behind this project has to do with the salary and performance of NBA players. To be more precise, the motivating question is: " *In the 2015-2016 season, how do the skills of a player relate to his salary?* "

Data

The primary source of data for this project is obtained from **Basketball Reference by Sports Reference LLC**. Basketball Reference is part of the website Sports Reference, which provides comprehensive statistics, up-to-date scores and rankings and history for BNA, ABA, WNBA and top European competitions.

In the project, there are mainly three kinds of raw data tables being used: **roster table**, **totals table**, as well as the **salaries table**.

- For the roster table, it includes the information about each player's name, position, height, weight, birth date, country of origin, years of experience, and attended college.
- For the totals table, which is also known for player statistics, it has the player's statistics during the entire season: age, games played, games started, minutes played, field goals, field goal attempts, etc.
- At last, salaries table, just as its name suggests, consists the salary of the players.

The National Basketball Association (NBA) is the **pre-eminent men's Professional Basketball League in North America**, and is widely considered to be the premier men's professional basketball league in the world. Currently, there are **30 teams in the league** (29 in the United States and 1 in Canada), which is divided into two conferences of three divisions with five teams each. NBA players are the world's best paid sportsmen, by average annual salary per player.

The regular NBA season begins in the last week of October and ends in the middle of April. NBA playoffs begin in late April, with eight teams in each conference competing for the championship. Each team will determine a 12-man active roster (and a 3-man inactive list before the regular season begins. During the regular season, each team plays 82 games, 41 each home and away. **The data we used in this report is selected from the 2015 - 2016 season, the 70th season of NBA, which began on October 27, 2015 and ended on April 13, 2016.** The playoffs started on April 16 and ended with the NBA 2016 Finals on June 19, 2016, with Cleveland winning the championship. The format of our data table is that, **each row represents an athlete, each column corresponds to a specific variable** (e.g height, weight, birth date, country of origin, etc.)

Methodology

Data collection

This is the very first stage of all project. The online dataset website for all NBA data is a great source for analyzing. However, because it's a website for browsing, it does not provide any API or other ways for us to get the data directly. R, as a powerful script language, provides us with powerful data mining package. "XML" package can be used to parse HTML website and extract the element we want. By exploring the HTML tag patterns of basketball-reference-link, we find a way to locate all the data we want including 30 teams' roster data, salaries data and other detailed performance information.

To be more detailed, we scripted all data needed from basketball-reference-link and store the roster, salary and stats data into the data folder. Firstly, we imported functions: scrap_data, installed XML if not installed, and sourced needed function. Then, we created folders roster-data/, stat-data/, and salary-data/. Besides, we also identify nodes with anchor tags for each team and extract the href attribute from the anchor tags. We get the character vector with the team abbreviations. Finally, we loop through team name and export CSV files.

Merging data

First of all, we add a team column to each of the separate tables. We bind all team table for roster, salary and stats data respectively by using `rbind()` function. Then, we inner_join those three tables by column "player" and "team". We found that several players transfer from one team to another. As a result, we got duplicate players. Further cleaning process includes removing duplicate players, renaming column names for better readability, set the proper data type for each column and adding "NA" for all empty entries.

Cleaning data

After we have stored all 90 raw data tables back to disk, we have one more step to enter the analysis phase. Raw dataset is a set that we just obtain from an outside source. We store it and we never change it. We need to clean the raw data so that we can better utilize it. We modified column name, changed date of birth to date format, and changed string height to height in feet. Besides, we also took several steps to have the data written in a conformed format, by filling none value to NA for college, removing $ sign from dollar sign, and changing R in experience to 0. Eventually, we can write out the clean table, in a concise and consistent way.

EDA

For the **Exploratory Data Analysis (EDA)**, which analyzes **the different statistics of both quantitative variables and qualitative variables. Also, this script uses ggplot to visualize the statistics of these variables.** Firstly, we source utility function and install packages, and create a folder called cleandata.
1. For quantitative data, we calculate the Mean, Max, Min, Range, SD, and median for all quantitative data. Then, we create a data frame for all quantitative data such as uniform_number, height, weight, and age. We set row names and column names of dataframe and sink data frame into txt file. In addition, we visualize data by histogram and save as png file; set the path of boxplot to be saved as png file; visualize data by boxplot and save as png file.
2. For qualitative data, we first get a frequency table of qualitative variables. Then, we remove the last row if there is a missing value. We visualize the qualitative data and save them in folders. The angles we analyze are position, team, country, and college.
To visualize data, we can use shiny web apps. By using shiny web apps, we can easily change and see the plot of the relationship between summaries (mean, min, max …) and different teams. You can check this link and play with differents statistics team-salaries-shiny-link

Formating

To get the salary statistics: After reading the clean data, we remove the extra columns, and split the entire data frame into a big list by the team. To make it look nicer, we reorder it into alphabetical order, and assign the name. finally, we create a matrix with data in a csv file.

Measuring

- Compute the efficiency

To evaluate the performance of the players, we usually use a formula that calculates what is known as EFF or "efficiency" statistic. EFF computes performance as an index that takes into account basic individual statistics: points, rebounds, assists, steals, blocks, turnovers, and shot attempts (per game). It is derived by a simple formula:

$EFF = (PTS + REB + AST + STL + BLK - MissedFG - MissedFT - TO)/GP.$

Something needs to be noticed is that: the EFF formula uses Missed FG (missed field goals) and Missed FT (missed free throws). However, these variables are not in the table of statistics. Instead, missed field goals should be individually calculated by deleting field goals from field goal attempts. Likewise, missed free throws should be calculated by deleting free throws from free throw attempts. After getting the values of missed field goals and missed free throws, we should also remember that both of them contribute negatively to EFF. Thus, rather than add them to the sum of the other factors, we should keep in mind that we need to add these variables with a negative sign.
Besides, we should be aware that EFF is not suggesting something on the total value. Instead, it is on the level of Per game, or average performance as we divide each of the statistics like points, rebounds, assists, steals, blocks, turnovers, and shot attempts by the number of games played.
However, in some sense, the EFF is biased, as the statistics it take into account is not complete. One of the issues with EFF is that it favors offense-oriented players over those who specialize in defense, as defense is difficult to quantify with currently tabulated statistics.
Another factor that should be taken into consideration is that: EFF does not take into account the different positions of the players. An alternative, or better version of it is the one considering EFF indices with PCA. As it is mentioned above, EFF ignores the difference depending on the position that the players play. Thus, to compensate such drawbacks, different efficiency indices that

take into account the players' positions should be considered. To calculate these indices, we use Principal Components Analysis (PCA). PCA will give you a weight for each term in the original EFF formula.
To calculate separate EFF indices for each position, you need to subset the player statistics dataset by positions, that is, a subset of C, PF, SF, SG, and PG. There are various ways to compute PCA. Among all these methods, perhaps the simplest way is with the function `prcomp()`. In fact, we perform a PCA on each subset of the data. The coefficients of the EFF index are the weights of the first principal component. In real practice, these weights are in the first column of rotation, from the object returned by `prcomp()`.
One aspect that especially needs attention is: while computing PCA weights, we are using scaled data, such as mean-centered and standardized data rather than the normal one. For the reason that the PCA weights are obtained using standardized variables, in order to get the new EFF value, you have to re-express the weights to **take into account the standard deviation of each variable.**
After we have grasped the method to get the PCA weights, we need to create a new table that called "eff-stats-salary.csv". It would be recommended to set it in the directory data/cleandata/. This table should contain the following variables: player's name, total rebounds, steals, missed field goals, turnovers, efficiency index, total points, assists, blocks, missed free throws, games played, salary.
The next step would be to **visualize the relationship** between all the player statistics — including the EFF* index—and the salary. In order to get this goal, we would have to create a shiny app that displays a scatterplot in which the x-axis corresponds to one statistic among all the variables, and the y-axis corresponds to salary (by default). By using shiny web apps, we can easily change and see the plot of relationship between different statistics/skills and players salary. You can check this link and play with differents statistics stat-salaries-shiny-link . We should also Include a widget that allows us to select the specific statistic for the x-axis. Include another widget that lets us select a variable for the y-axis (by default should be salary). Include another widget to indicate whether the dots should be colored by position. The app should also display the correlation coefficient between the chosen variables. An optional choice would be that, we can add more widgets to further customize the displayed graphic.

- Value of a player
  To measure the "value" of a player, we compute an efficiency over salary ratio obtained as: $value = efficiency/salary$
  In other words, we compute this ratio for each player, using the EFF index (by position), and dividing it by the salary. We use this value to identify the top 20 players , which means most valuable and the bottom 20 players, those with least value. Finally, we save the list of these 40 players in a text file best-worst-value-players.txt inside the data/cleandata directory.

- Finding correlation
  To describe the correlation between the different skills and salary also different positions and salaries, we use the following methods. We plot the least square regression line and correlation coefficient. Firstly, we should import function like install_package_correlation_eda, and we also install packages such as dplyr, stringr, MASS, ggplot2, plyr, readr, reshape2. Then, we read in clean roster data, efficiency data, and add position column to Efficiency table. Finally, we would be able to plot the correlation between different skills and salaries, which would also allow us to calculate the correlation coefficient. Besides, we also plot the correlation between different positions and salaries, finding the least square regression line at last.

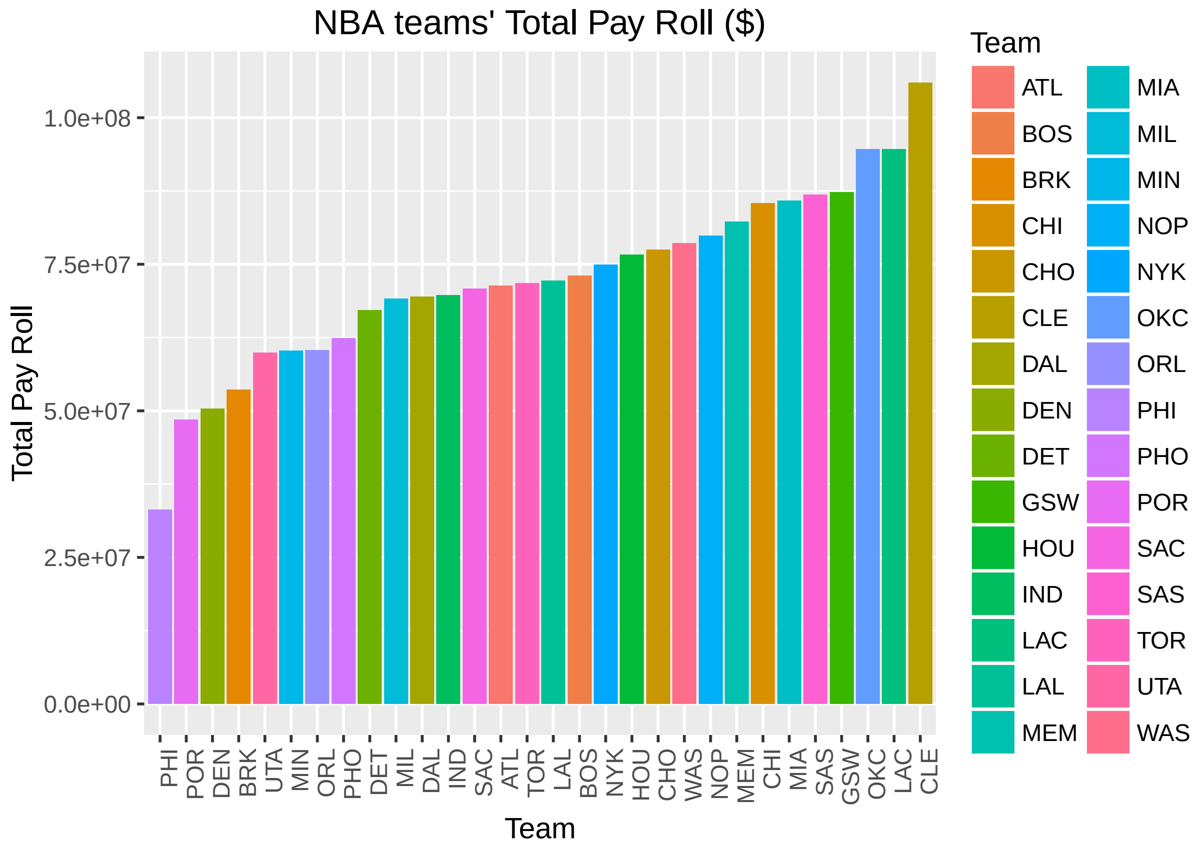
Results

As mentioned in the introduction, the major questions we intend to take is the relation between the salary and basketball players' performances in the 2015 - 2016 season. To get a better idea of the data sets we obtained from basketball references, *we first examined the basic parameters by each team.*

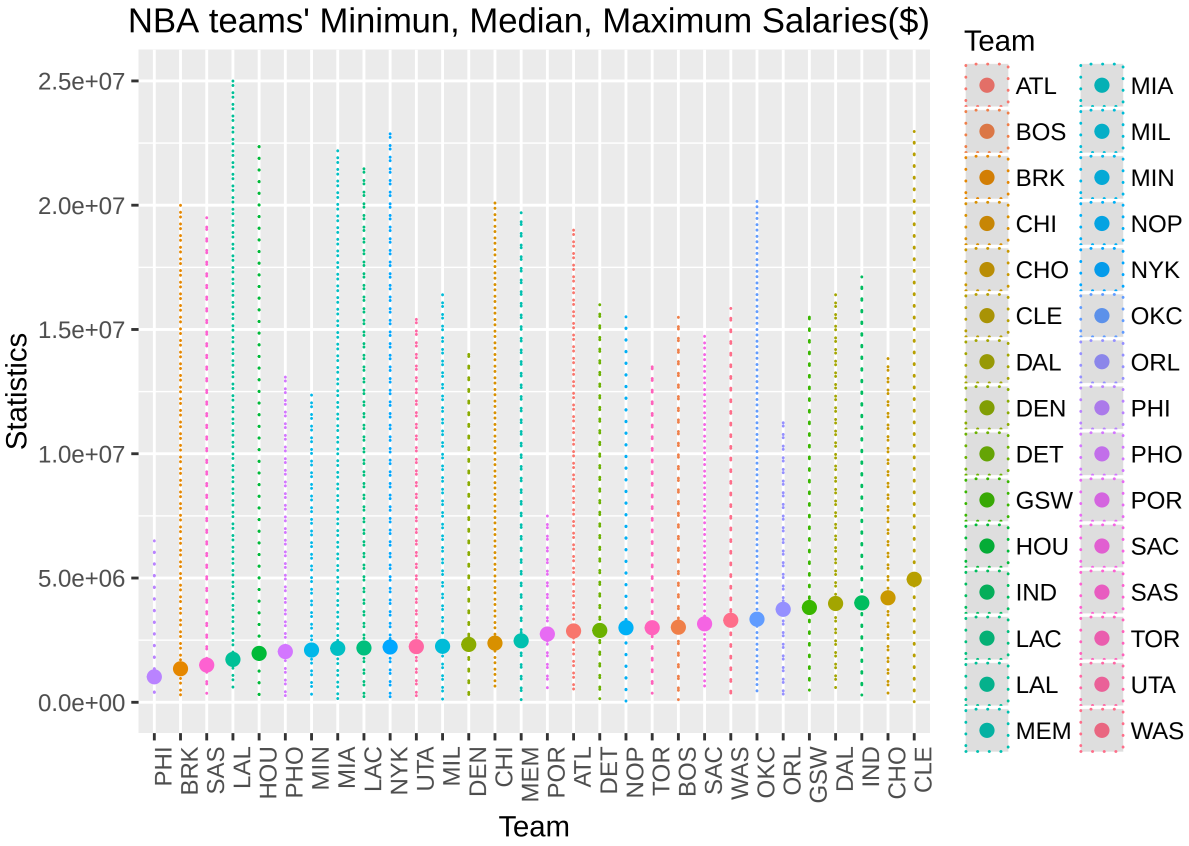
First take a look at the aggregated salary of each team

In the file team-salaries.csv, a summary of each team's salary information is listed. The teams are listed alphabetically.

Below is the plot for total pay roll of different teams in increasing order



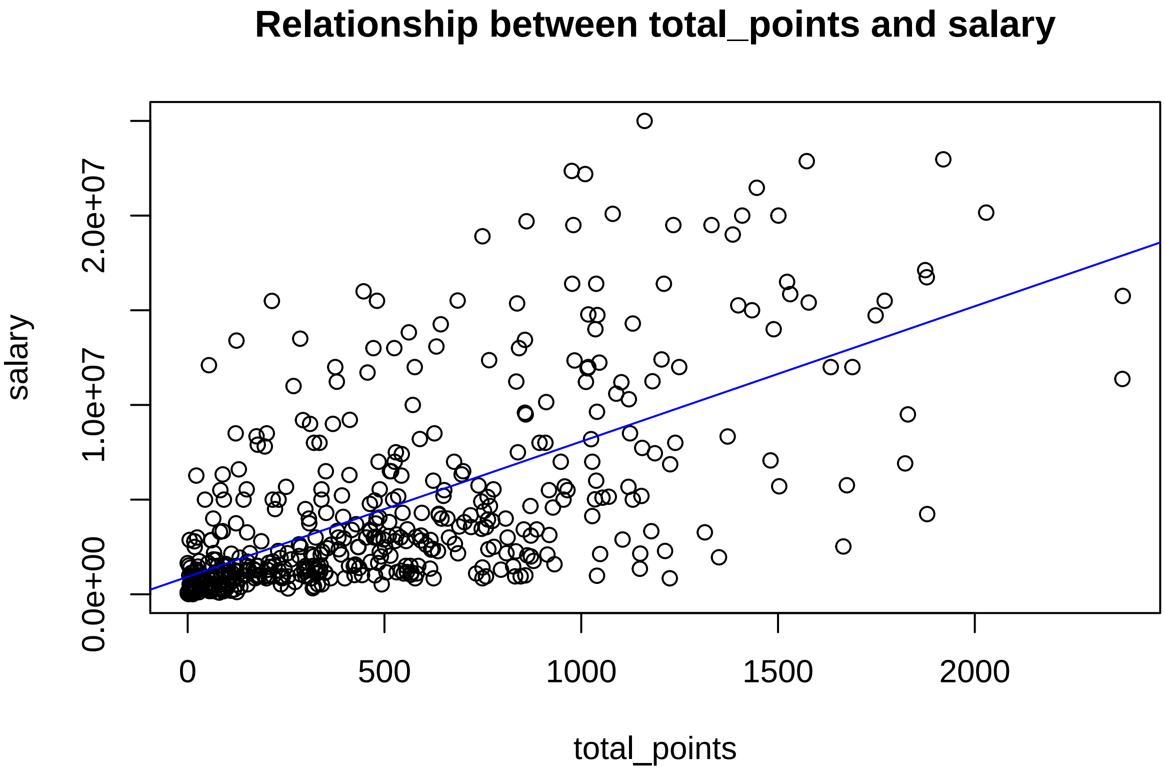
Also let's look at the plot for NBA teams' Minimum, Median, Maximum Salaries($)



From the resource data, we calculated the following summary information of salary including the range (minimum and maximum) of the summary, the first quartile and the third quartile, mean, median, standard deviation, interquartile and the total pay of each team. The lowest salary of NBA players each team that season ranges from $8819 (from Team Cleveland) to $525,093 (from Team Portland, Sacramento, Dallas, LA). The highest salary of NBA players each team that season ranges from $6,500,000 (from Team Philadelphia) to $25,000,000 (from Team LA). The mean of salary each team ranges from $1,842,421(again from Team Philadelphia) to $7,064,168 (again from Team Cleveland) In terms of the total salary, Team Philadelphia has the lowest total salary at $33,163,582, which is less than 10 million more than the highest payment of a single player in the league. The team with the highest total salary is Cleveland, with a total salary of $105,962,520, which is more than triple of the total salary of Team Philadelphia. The total salary of most teams falls into the range of $60,000,000 to $80,000,000. It is noteworthy that Cleveland Cavaliers with the highest aggregated salary among all the teams within the NBA league won the championship that season while Philadelphia 76ers, the team with the lowest aggregated team salary overall ended up the last in the league with only 10 wins out of 82 games that season. *Judging from the final win-lose result of 2015 - 2016 season, our first impression is: the aggregated team salary is highly related to the performances of each teams as a whole, given the fact that teams giving out higher salaries in total tend to be the ones hiring the best players in the league.* After getting a general idea of the relation between the end of season performance, namely the ranking and honors of each team, and their aggregated salary, we now move on to take a closer look at *some specific performances of each player.*

Which skills are more correlated with salary?

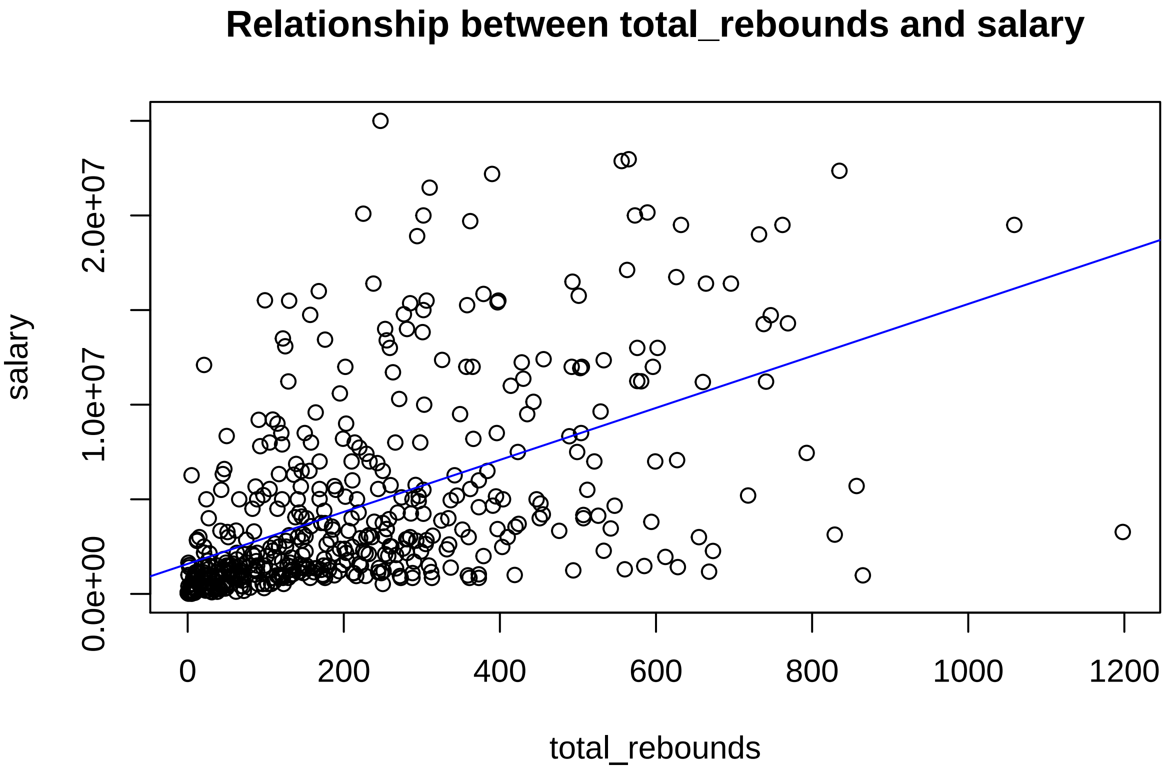
Firstly, we analyzed the relation between total points each player scored that season and their salaries.



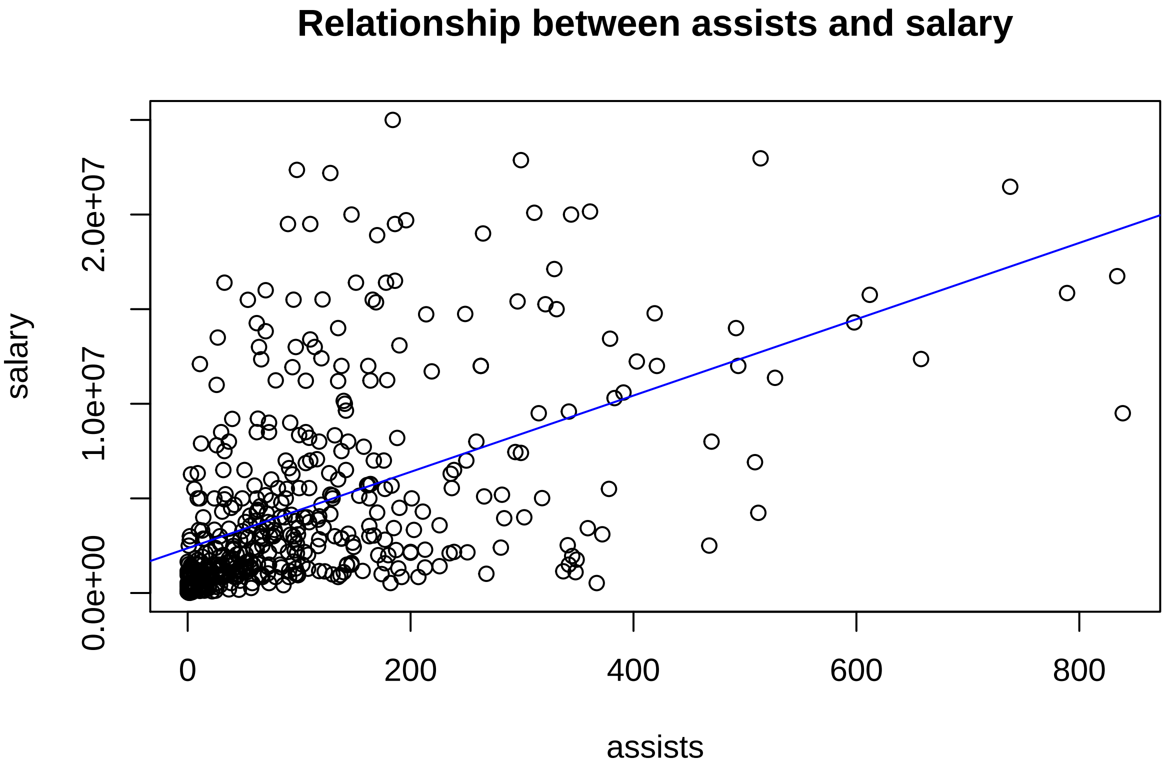
```
## [1] "correlation coefficient is 0.64"
```

As shown in the regression graph “Relationship between total points and salary, the correlation between the two variables total points and salary is 0.64, which shows that the total points scored by players that season is, in general, related to the salaries they received. The highest total points - 2376 is scored by James Harden, who earned a salary of $15,756,438 (ranked the 27th highest among 496 players). *Players receiving higher payments tend to score a higher total-score during the 2015 - 2016 NBA season;* players receiving lower payments in turn do not perform very well in terms of the total points they scored. Despite the general trend, there are a small group of outliers, who receive a very high salary but did not give a top performance in terms of total scores. But the lower points they scored in total could be result from their injury or other unpredictable reasons.

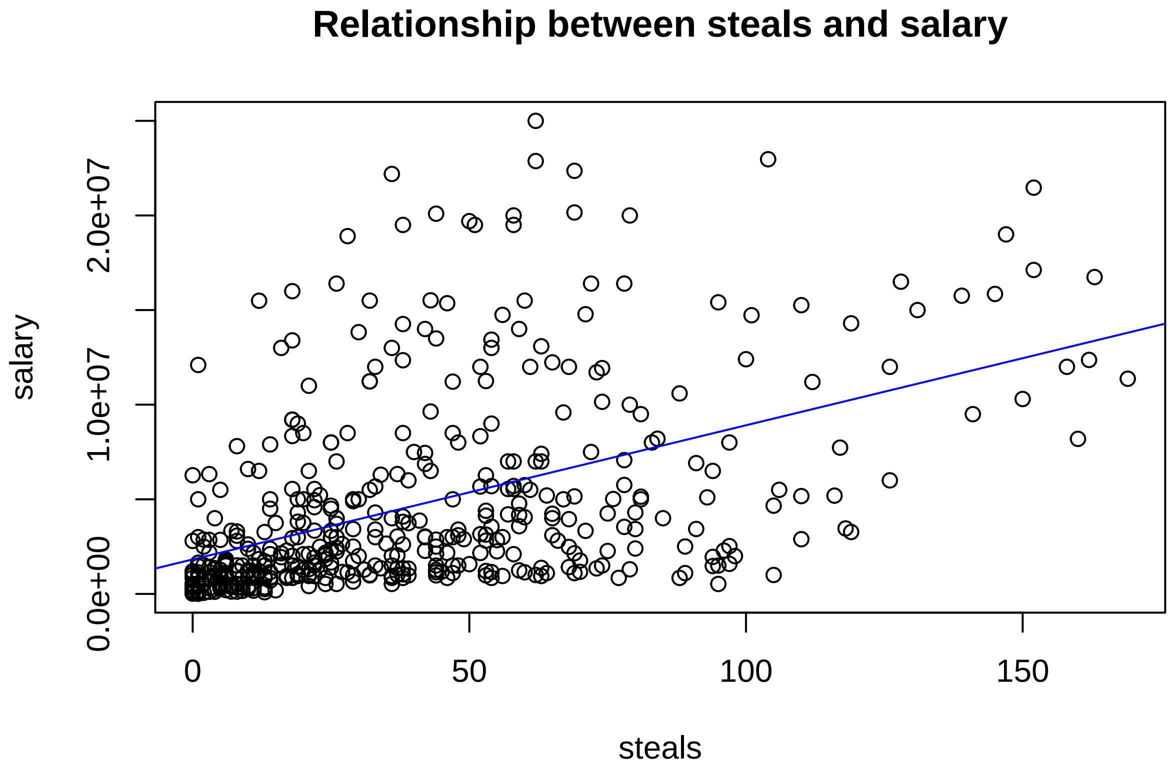
After examining the correlation between total points and player salary, we move on to take a look at some more specific parameters in order to better understand the performance of players.



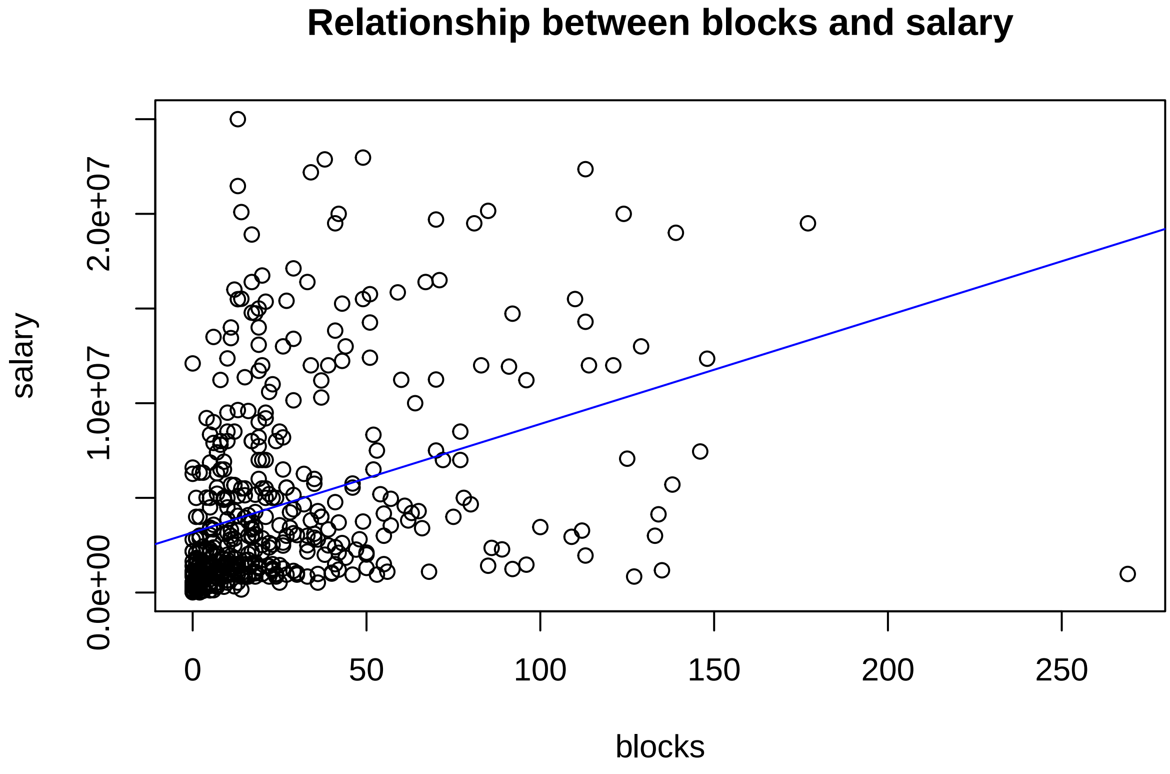
```
## [1] "correlation coefficient is 0.53"
```



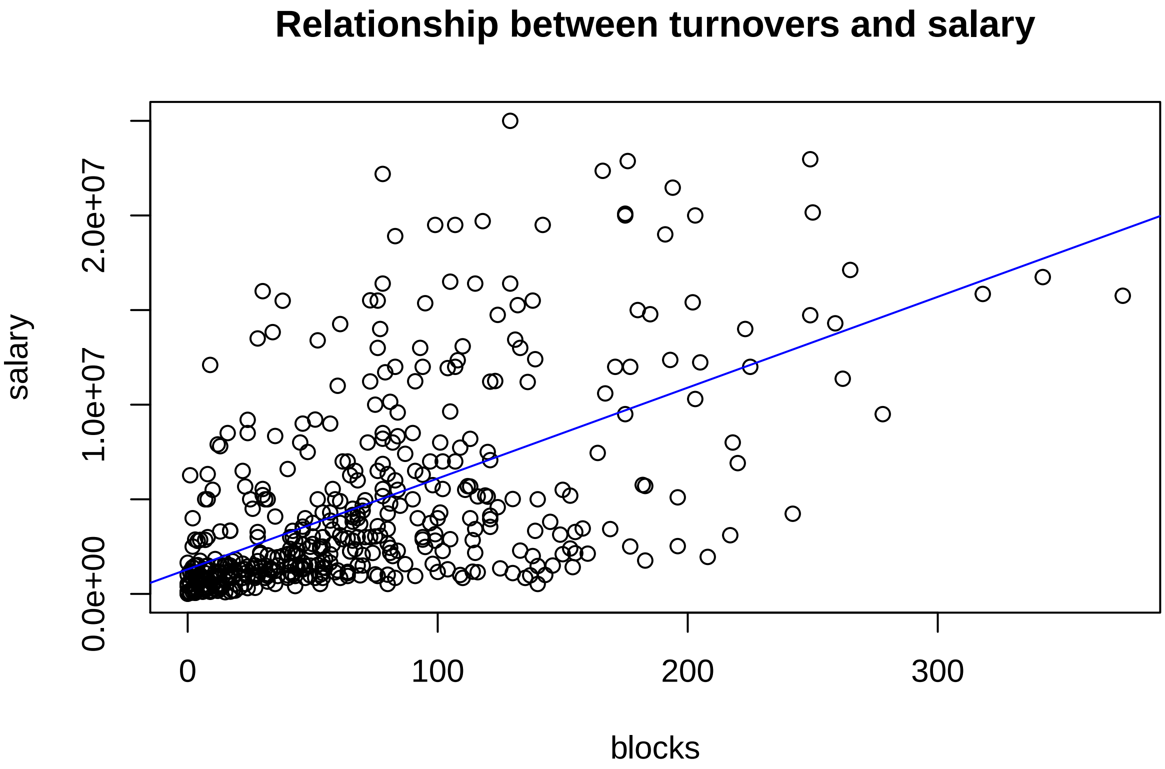
```
## [1] "correlation coefficient is 0.51"
```



```
## [1] "correlation coefficient is 0.49"
```

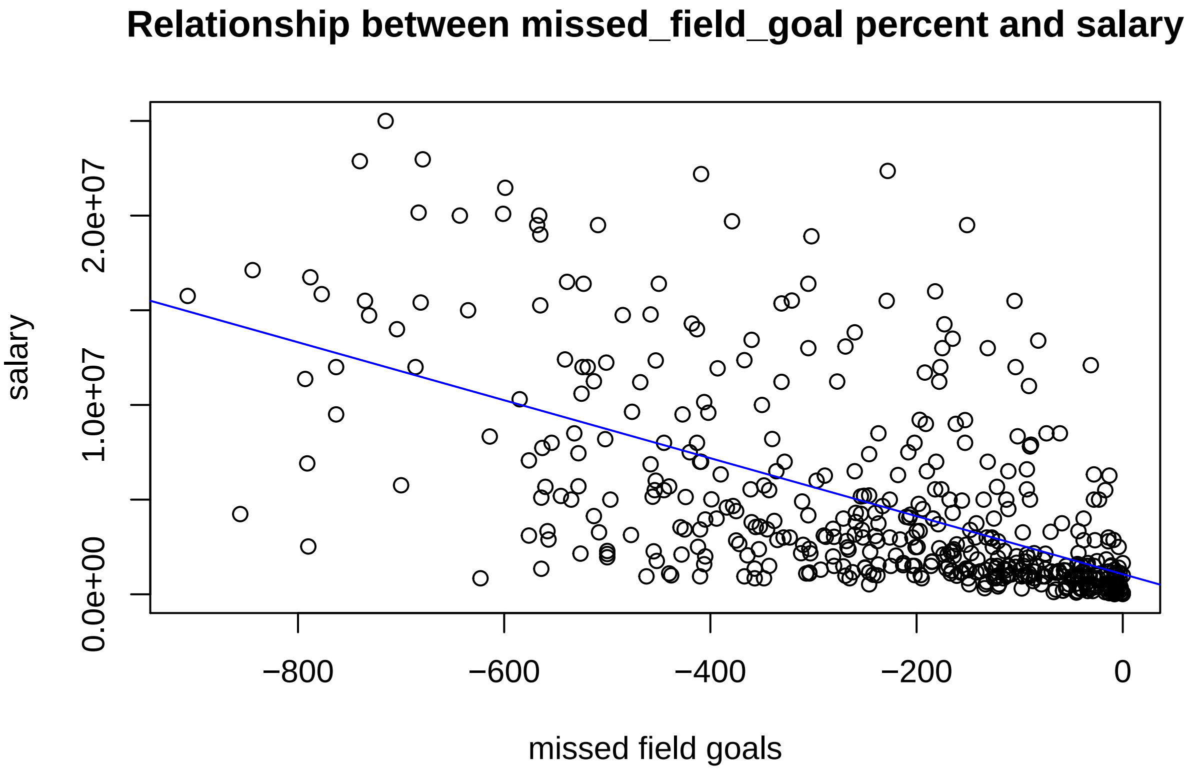


```
## [1] "correlation coefficient is 0.49"
```

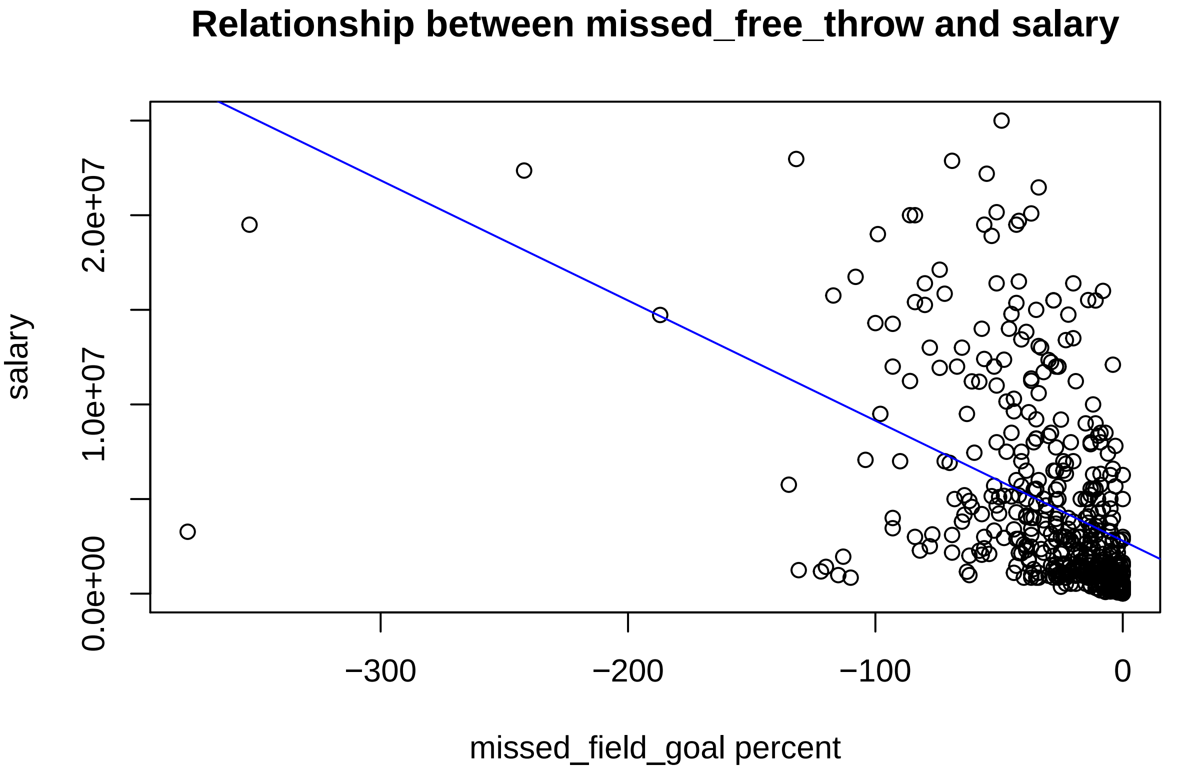


```
## [1] "correlation coefficient is 0.58"
```

Similar to the case of Assist, several other parameters also showed similar correlations. To be more specific, the correlation coefficient between **total rebounds and salary** is 0.53, that of **steals and salary** is 0.49, that of **blocks and salar*y **is 0.49, and that of turnovers and salary is 0.58.** Which means there is a moderate positive relationship between players' salaries and their total rebounds, steals, blocks and turnovers; players with higher salaries in general tend to have higher number of parameters mentioned above.** Despite this general trend, there are several outliers stand out during our data analysis which we think is worth noticing. The player Andre Drummond contributed 1198 times of rebounds, which is the highest among all players within the league, but he only earned a salary of $3,272,091 and he did not rank as high on other parameters. Similar cases are also found in other parameters. For example, Hassan Whiteside contributed the highest number (269 times) of blocks that season, almost 100 times more than the second highest. But his salary is only $981,348, which is only 1/19 of that of the player who contributed the second highest number of assists.*We interpret this phenomenon that certain players who might not be ranked top in terms of total salary of total points scored might be exceptionally good at some skills.

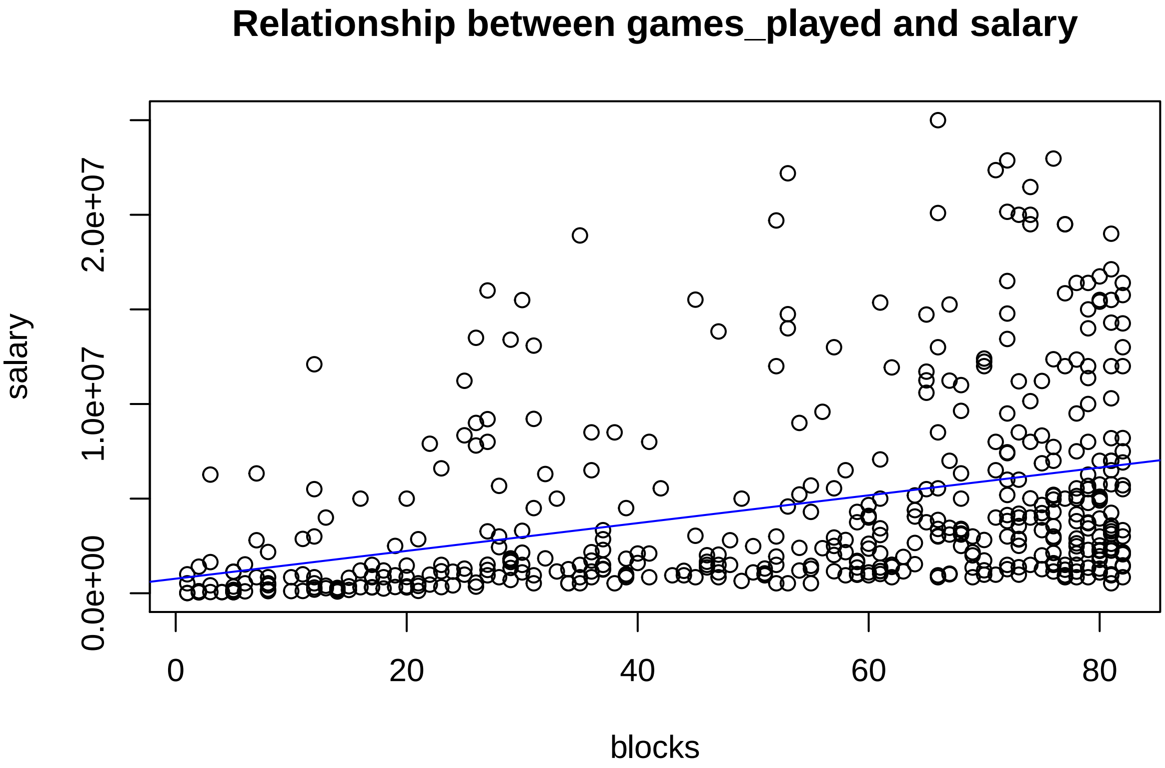

Relationship between missed_field_goal percent and salary

```
## [1] "correlation coefficient is -0.59"
```


Relationship between missed_free_throw and salary

```
## [1] "correlation coefficient is -0.45"
```

Another two parameters we examined are the number of missed field goal and missed free throws. The correlation coefficient of players' salaries and missed field goal is negative because we used negative numbers to interpret each goals missed. To put it in another way, players earned lower salaries tend to miss a lot fewer goals, but as the players' salaries rise, their missed goals varies. The pattern is find in the graph of the relationship between salary and missed field goals. We think such an interesting phenomenon results from the fact that players receive lower salaries in general have less chance to play during the game. Since they did not shoot as much in the first place, they miss fewer goals.



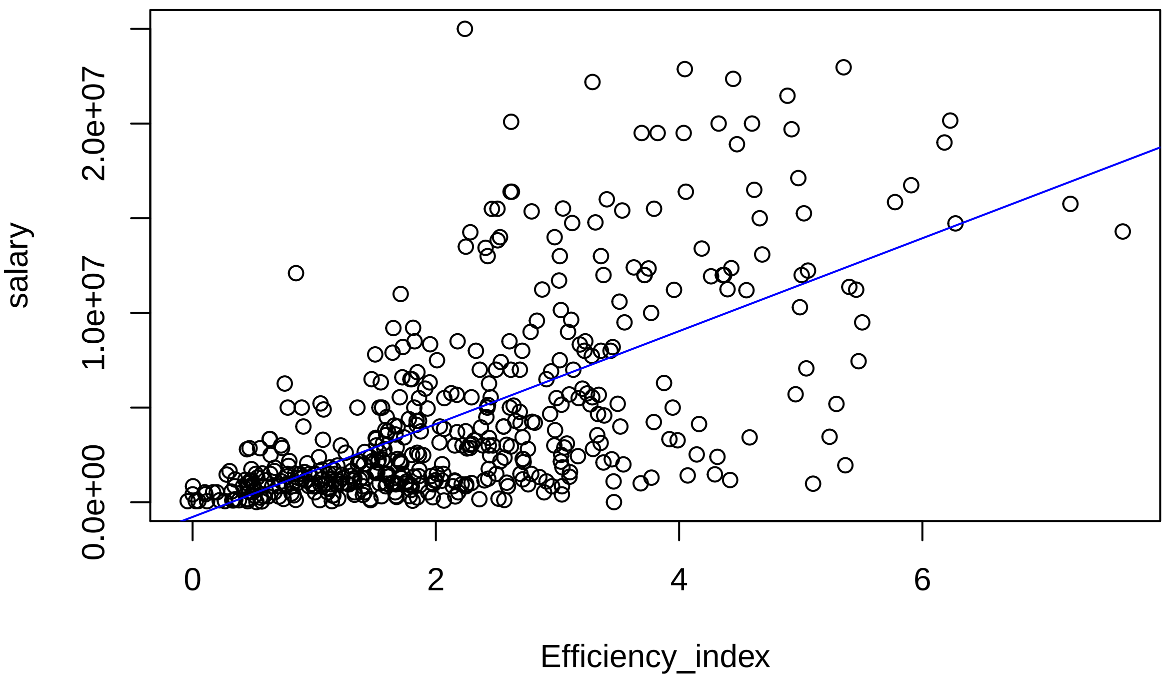
```
## [1] "correlation coefficient is 0.36"
```

However, the correlation coefficient between total games played and the players' salary is 0.36. **Which means the positive relationship between the number of games played and the player's salary is rather weak.**

Are there any differences in skills and salary depending on the players' positions?

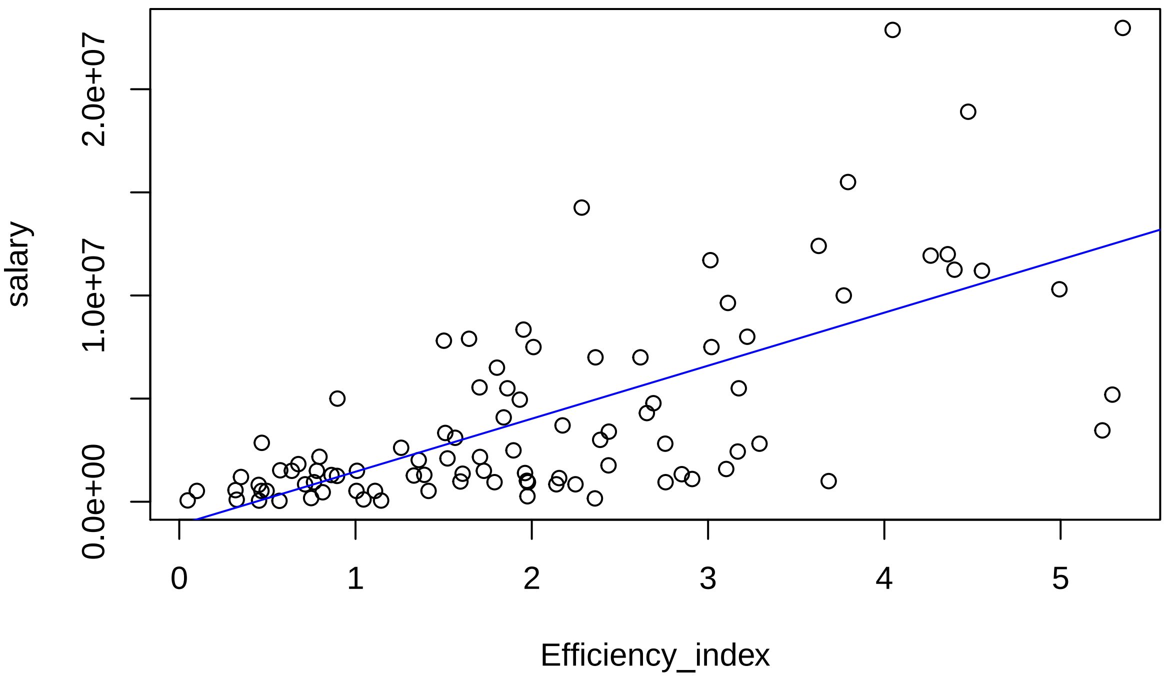
Now that we have a systematic way to analyze the overall performance of NBA players by taking into consideration a wide range of parameters, we are curious if the players' positions have any influence on their performance. To do that, we mainly analyzed the correlation between Efficiency Index and their salaries and have the following findings.

All Position Plot



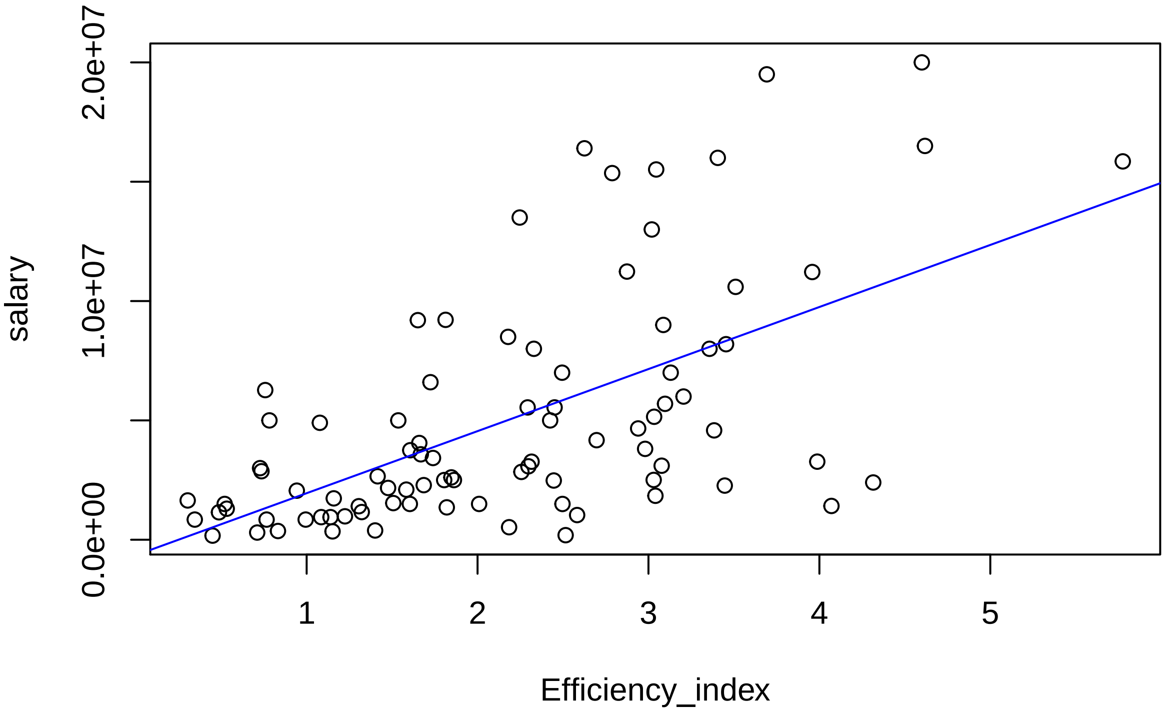
```
## [1] "correlation coefficient is 0.64"
```

Point Guard(PG) Position Plot



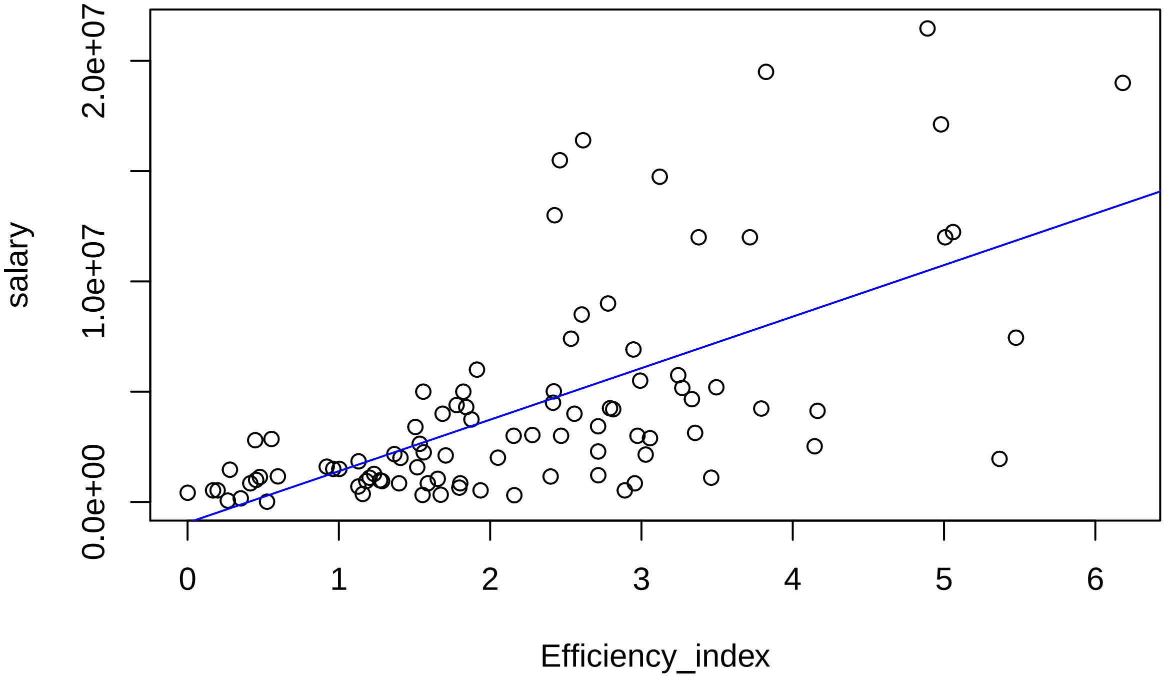
```
## [1] "correlation coefficient is 0.68"
```

Center(C) Position Plot

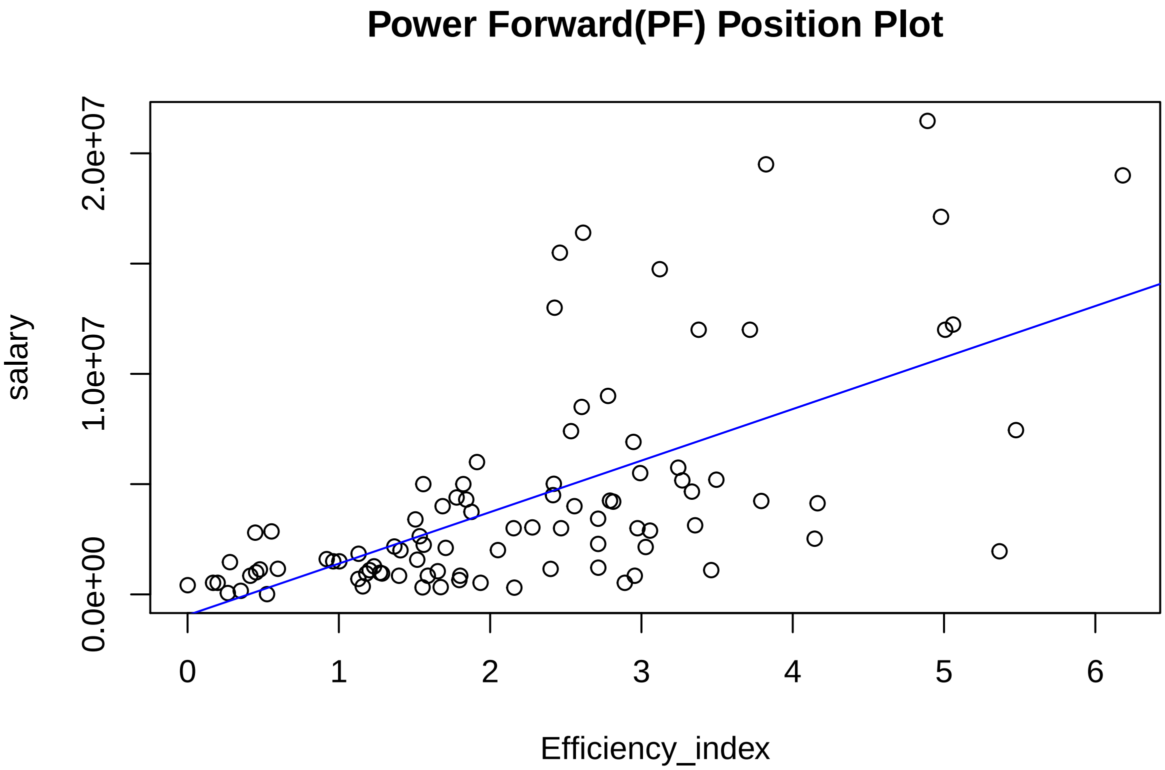


```
## [1] "correlation coefficient is 0.6"
```

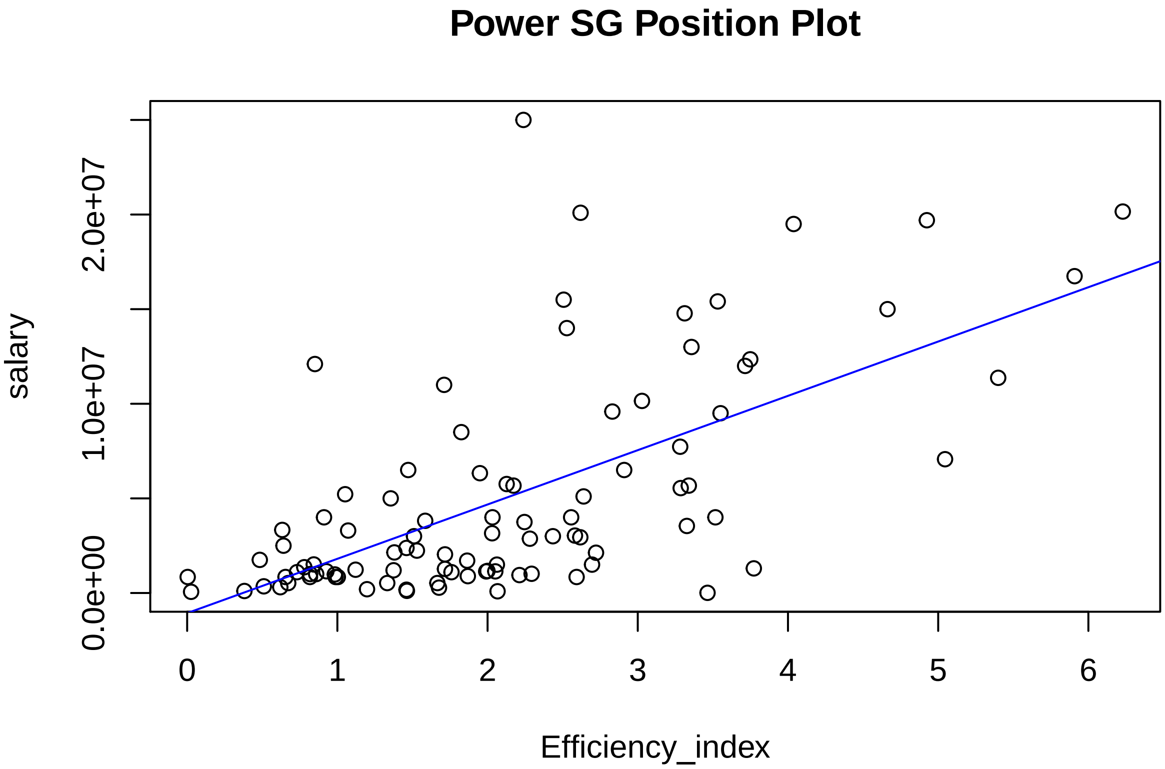
Small Forward(SF) Position Plot



```
## [1] "correlation coefficient is 0.64"
```



```
## [1] "correlation coefficient is 0.64"
```



```
## [1] "correlation coefficient is 0.64"
```

Firstly, **the correlation between the salary of all the players from 2015 - 2016 season and their efficiency index is 0.67, which means there's a pretty obvious positive relationship between players' salaries and their efficiency index.** Secondly, we divide all the players by their position to see if the performance of players from any particular position is more closely related to their salary. It turns out that players from different positions tend to have very similar correlations between their salaries and efficiency index. To be more specific, the correlation coefficient of Point Guard is 0.68, that of Center is 0.6, Small Forward is 0.64, Power Forward is 0.64, Power SG is 0.64. All of their correlations fall around 0.65. **Therefore, we interpret from the graphs that the salary of all the positions have a similar positive relation with the players' performances.** The correlation coefficient between salary and performance of players with Point Guard position is lighter stronger comparing to other positions, but in general, different positions do not have an obvious effect on the correlation between skills and salary.

Are there any undervalued or overvalued players?

To determine whether the players are worth their salary, we calculated an index called "value". Here, we listed players with highest values and lowest ones. The top 20 players (the specific names are in the best-worst-value-players.txt in the cleandata folder) with highest values means they performed reasonably well while earning a modest salary. **These undervalued players contributed to their teams greatly without spending their teams too much money.** However, for the least 20 players(the specific names are in the best-worst-value-players.txt in the cleandata folder), they are considered overvalued because they did not contribute to the team much while earning a very high salary. However, it's noteworthy that some of the overvalued players did not perform well or contribute not because of their own attitude. On the contrary, players like Derrick Rose and Chris Bosh missed most of the games during that season because of their injuries.

Insights

Based on such finding, we would like to bring out two possible reasons behind. Based on the theory in business administration, Performance = f (Ability X Motivation). In this formula, Ability = Skills and Training, and Motivation = Incentives, Rewards, Intrinsic Motivators.

The team-players of various teams start with the same level of ability. However, it is the salary that serves the incentives or rewards that increases the players' motivation, thus resulting in the higher performance. Thus, we bring out a bold statement that the players not necessarily worth the amount of money clubs pay for as whoever get this amount of reward as incentive would have higher motivation and consequently, perform better.

The team-players of various teams start with different level of ability. Plus, both the players and the clubs know who play better and who play worse. The clubs that have more money and provide higher salary would be able to attract those of better ability. That is to say, the higher salary only serve as the factor that attracts those of higher capability. It does not change the players' performance afterwards. Under this assumption, the players really worth the amount of money clubs pay for.

Conclusion

In conclusion, the team's' performance in general is rather closely related to the total salary they give out to their players. If looked upon individually, the players' performance is positively related to their salaries in general while different skills do not impose very different effects on the relationship between the players' salaries and their performances.