Determinants of NBA Players' Salary

Stat 133, Fall 2016

Albert L Hu, Sarah Hu, Ming Ho Cheung, Zihan Chen

December 2, 2016

Determinants of NBA Players' Salary

Abstract

Exploring the relationship between the salary and performance of NBA players in the National Basketball Association (NBA) is of great significance for team managers to make reasonable decisions on a player's salary. Our project serves to highlight the performance variables, such as points, turnovers, and blocks, that are strongly correlated to a player's salary in the NBA. In order to do so, our project involved an exploratory phase, referred to as Exploratory Data Analysis (EDA), to analyze multiple data sets collected from the website 'Basketball Reference' that are dedicated to displaying statistics of NBA players in various seasons. Furthermore, we calculated efficiency (EFF) indices that take into account players' positions by utilizing Principal Components Analysis (PCA). Additionally, we calculated correlations between performance variables and players' salary. In this project, we focused on investigating performance variables of 471 NBA players from 30 teams in the 2015-2016 season. An emergent feature that our group discovered was all performance variables of a player are positively correlated to his salary, but some are more strongly correlated than others. To be more specific, total points and field goals gained by a player are the two most correlated performance variables. We presented some of the more meaningful findings and implications of the results in detail in the following demonstration.

Introduction

In the 2013-2014 National Basketball Association (NBA) season, the Los Angeles Laker's Kobe Bryant earned $30.4 million whereas the average NBA player salary earned about $5 million. Is Kobe that much better than a player who earns $5 million a year? This leads to the question of whether or not NBA players are overpaid, which is problematic in the NBA. Statistically, some players are overpaid based on poor or sub-par total performance, but are still highly valued. These players are highly sought after because they are proficient in one aspect of the game, such as three point shooting, or they may be a strong defensive force. However, these same specialized players may be a major liability in other areas. For example, the three point shooter may be a poor defender and the defensive force may be a poor free throw shooter. Whatever the issue, overpayment in the NBA is a concern, and team managers aim to fairly compensate a player for his performance.

However, in the NBA it seems as if compensation is a reward for past performance and anticipated or expected future performance. However, past performance and expected performance may not be good indicators of fair compensation. NBA owners and general managers often overspend for a player that they feel will meet an immediate need. Therefore, for managers of different teams in NBA, it is of great significance to understand the relationship between player's performance and salary. The exploratory research question for this project is "In the 2015-2016 season, what is the interpretation of the performance statistics of a player and what is the relationship between performance variables and player's salary?"

To answer this problem, we first used exploratory data analysis to explore individual performance statistic and see the relationship between different performance variables and salary. Then we calculated the efficiency index by position to better evaluate players' performance with different positions.

Data

Introduction of NBA

National Basketball Association (NBA) is one of the four major professional sports leagues in the United States and Canada. It is widely considered to be the premier men's professional basketball league in the world, and has 30 teams, including 29 U.S teams and 1 Canadian team. These teams are split into the Eastern or Western Conference. The regular NBA season is from the last week of October to the mid April. The NBA Playoffs begin in late April, wherein eight teams from each conference compete for the Championship.

Introduction of data

The primary source of data for this project is from Basketball Reference, which is a website that provides statistics, scores, and history for the National Basketball Association(NBA), American basketball Association (ABA), Women's National Basketball Association(WNBA), and top European competition. In this project, we focused on analyzing the statistics of all 30 NBA teams in the 2015-2016 season.

The project starts with data acquisition. Our team scraped three kinds of raw data tables for each team from the website: Roster, Player Statistics, and Salaries, and stored those tables in a comma separated file(CSV).

The Roster tables contain information about each player's name, position, height, weight, birth date, years of experience and attended college. There are some missing values for attended college in the raw tables. They were all converted to 'NA' values when we started our data acquisition process. The Statistics tables contain players' statistics during the entire season: age, games played, games started, minutes played, field goals, field goals attempts, etc. They are all quantitative variables that we will use to reach an answer for the central topic: the relationship between salary and performances of NBA players. The salaries table simply contains the salary of the players.

In the Roster table, there's a 'Position' columns that contains five different positions: 'C', 'SF', 'SG', 'PF' and 'PG'. They are the abbreviated names of 'Center', 'Small Forward', 'Shooting Guard', 'Power Forward' and 'Point Guard', respectively. Players at the center position are skilled at gathering rebounds and contesting shots. The small forward position is considered to be the most versatile of the main five basketball positions. They are typically skilled at drawing fouls and shooting from long-range. Most shooting guards are good shooters from three-point-range. Power forwards share a similiar role to that of the center; they also play close to the basket and are often noted for the accurate mid-range jump-shots. They also defend within close to the basket. Lastly, point guards are the team's best ball handler and passer. Since they make the plays for the team, they are often good at assists and steals. Since players in various positions have various skills, we took their positions into consideration when calculating the efficiency index and the values of players.

Methodology

In analyzing all of the data for this project, we utilized the programming language R. Our main source of data was this website. First of all, we had to scrape three different tables from every team's page in order to gather the roster, salary, and in-game statistics of each player for every team. There were a few issues that we ran into while doing this; some players started out on a certain team but sometime during the course of the season were transferred onto another team. However, most of these players that ended being transferred weren't on the salary or stats tables for the new team, so we were able to avoid dealing with a player appearing on multiple teams by cross-referencing the three different tables for each team. This was achieved by joining all of the tables on the common column "Player" for each team, and then at the very end, once we got rid of duplicates within each team, we simply combined all of the joined tables with each other to get all of the values together.

However, there were still some problems that weren't conducive to our analysis of the data. For example, the mode of several of the columns would make plotting the variable extremely inefficient (e.g. the salary column had a mode of factor, when it really should have had a mode of numeric). As such, after combining the data from all of the different teams together, we had to clean it. Additionally, we changed the names of the columns of the table so that a reader will have a better understanding of what kind of data each variable/column contains. After applying these changes to our dataset, it was written into roster-salary-stats.csv.

Upon having a complete dataset that can be used to form the basis of our analysis, we to get a feel for the data. In doing so, we created a script called eda-script.R, which generates summary statistics for each of the variables within roster-salary-stats.csv, all of which can be seen in data/cleandata/eda-output.txt. Additionally, we also generated box plots, histograms, and bar charts within ed-script.R, all of which can be viewed in the images folder.

In addition to this period of data exploration, we looked at the aggregated salaries for each team, along with their additional statistics. To do this, we created make-salary-stats-script.R, which essentially iterates through all of the teams and creates vectors for each of the values in team-salaries.csv (teams, total_payroll , min_salary, max_salary, first_quartile_salary, median_salary, third_quartile_salary, average_salary, interquartile_range, and standard_deviation). Then, these values are then put into a dataframe and written to team-salaries.csv. To aid with understanding the data and providing a visual representation of our results, we created a web app using shiny that plots a bar chart of a specific statistic relating to salary per team.

After performing an analysis of the salaries of each team, we decided to tackle the heart of the matter: giving each player a statistic that measures how well they perform in their games and how much they contribute to the team. In doing this, we decided to generate the efficiency of each player, which is typically calculated by summing up the total points, rebounds, assists, steals, and blocks, subtracting missed field goals, missed free throws, and turnovers, and then dividing the resulting number by the games the specific player has played. However, the issue with this method is that offense-oriented players are heavily favored; the number of points a player scores per game vastly outweighs the rebounds, assists, steals, and blocks that they can rack up. To solve this issue, we decided to use principal components analysis(PCA) to give a more fair analysis. This way, we can compare apples to apples, so to speak. Ultimately, we wrote the variables used to calculated our efficiency index (position, total points, rebounds, assists, steals, blocks, missed field goals, missed free throws, turnovers, games played, and salary) and the efficiency index to eff-stats-salary.csv.

As with the analysis of salaries per team, we also created a web app using shiny. For this particular app, we wanted the user to be able to compare any two statistics with each other and visually see the correlation between the two. As such, we provided the ability to plot any of the two of total points, rebounds, assists, steals, blocks, missed field goals, missed free throws, turnovers, games played, and salary against each other. To aid in understanding our scatter plot, we gave the option of color-coding the plot by position, as well as an additional option of displaying the linear regression line on the plot.

Finally, we wanted to explore which players had the best “value”, where value was defined to be a player’s efficiency index divided by their salary. This way, we can see how a player’s value is tied to the amount of money that they make. We then found the top 20 most “valuable” players, as well as the 20 least “valuable” players, and saved our findings in best-worst-value-players.txt.

Result

Players Analysis

From the position bar chart we can see that there are roughly equal number of players in each position. However, there are slightly more players who play power forward(PF) than players who play other positions. This make sense because the power forward’s primary goal is shot blocking and short range shooting. They defend opposite power forwards, stay close to the basket, and help teammates in scoring points.

From histogram of player’s ages, we can see that most players are in their mid 20’s, with the average age being 26.6. This make sense because most professional players who finish college are about that age when they join the NBA. There are few players in their later 30’s because many of players retire due to injuries

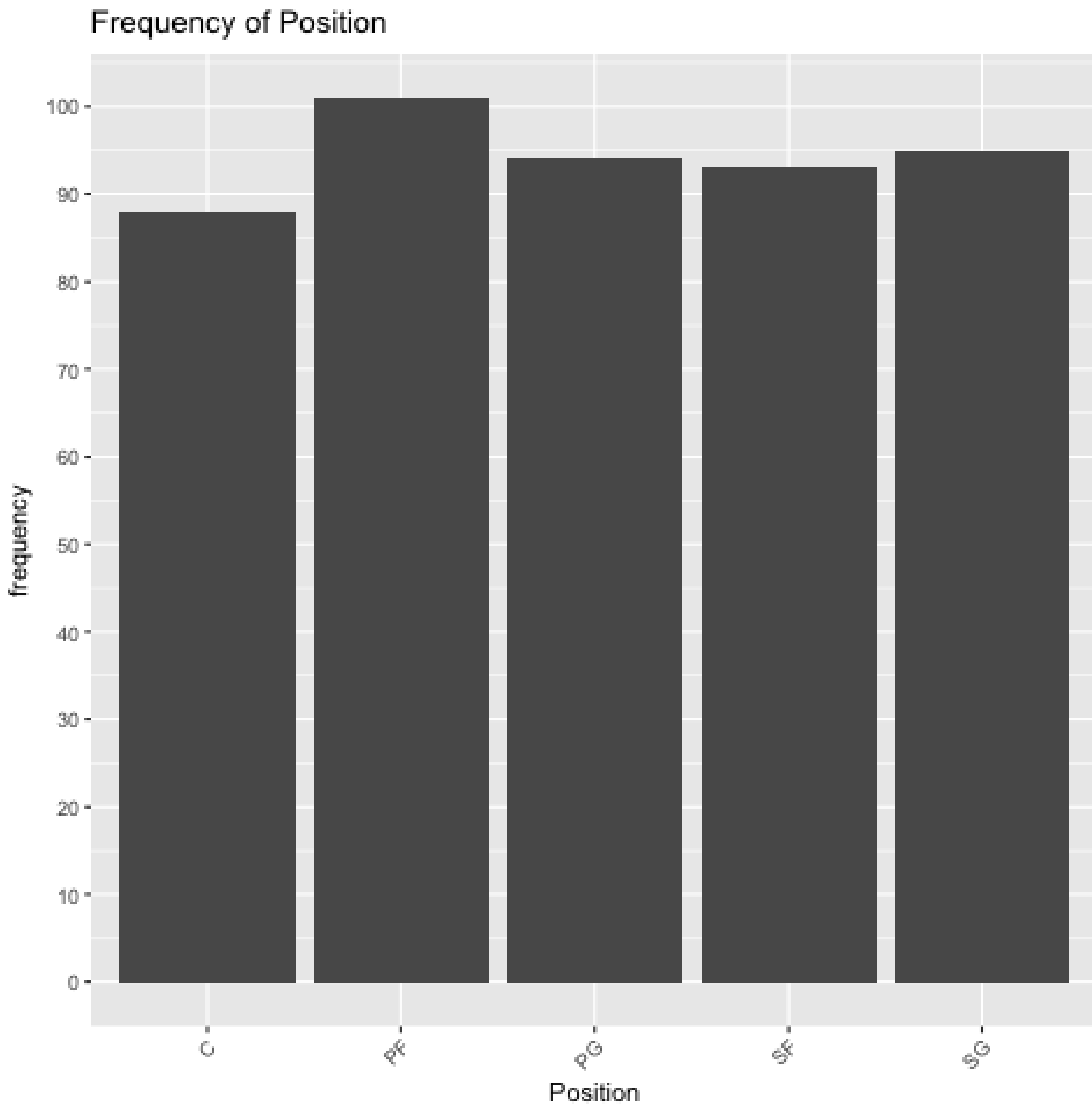


Figure 1: *Bar Chart of Players' Position*

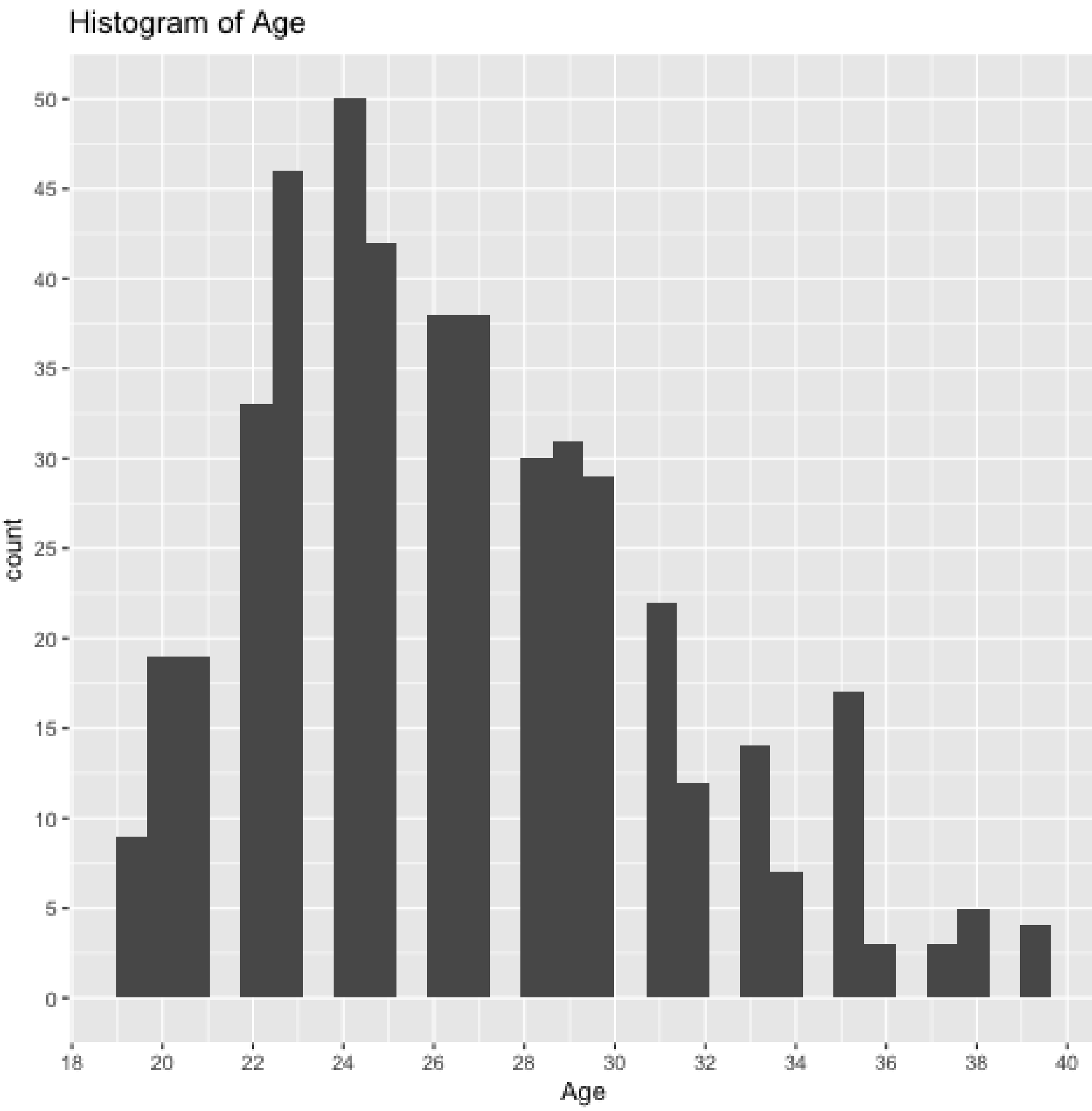


Figure 2: *Frequency of Players' Age*

after playing in the NBA for over a decade. Additionally, as people age, they lose their stamina and strength, and cannot keep up with others who are younger than they are.

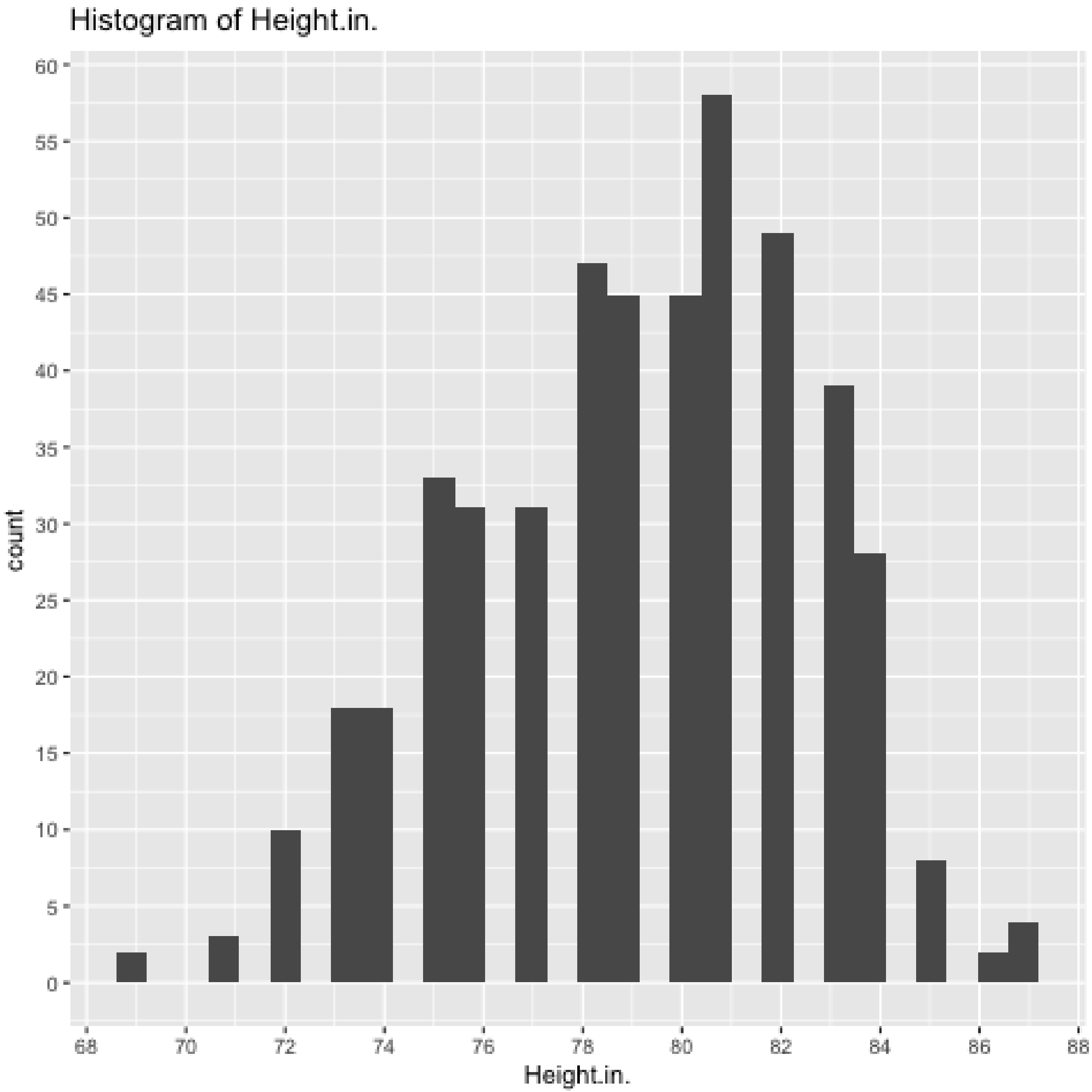


Figure 3: *Frequency of Players' Height*

From the histogram of height we can see that height of players are roughly normally distributed. The relatively taller players are in the position of center because their height enables them to gather rebounds more effectively.

Skills and Position analysis

We used boxplots for exploratory analysis to see the relationship between each skill measurement and position. Boxplots show that point guards focus more on assist and steals. Players in position center focus more on blocks, defensive rebounds, offensive rebounds, free throws, and two point field goals. This make sense because center players are positioned close to the basket. Shooting guards focus on making three point field goals because they are always far away from the basket.

Skills and Salary analysis

Figure 4: *Scatterplot of years of experience and salary*

Multiple regression analysis was conducted to determine which explanatory variables were predictors of NBA player salaries. From the scatterplot of the number of years players have played in NBA and their salary, we observe that there is a weak positive correlation of 0.36 between years played in NBA. About 13.12% of the variations in salaries can be explained by years played in NBA. This makes sense because salaries should be more correlated to player's performance. Although more experience tends to lead to better performance, it is not the cause, since a player's performance can drop off as they age. Thus the number of years players have played in the NBA does not appears to be a main factor that influences the salary of a player.

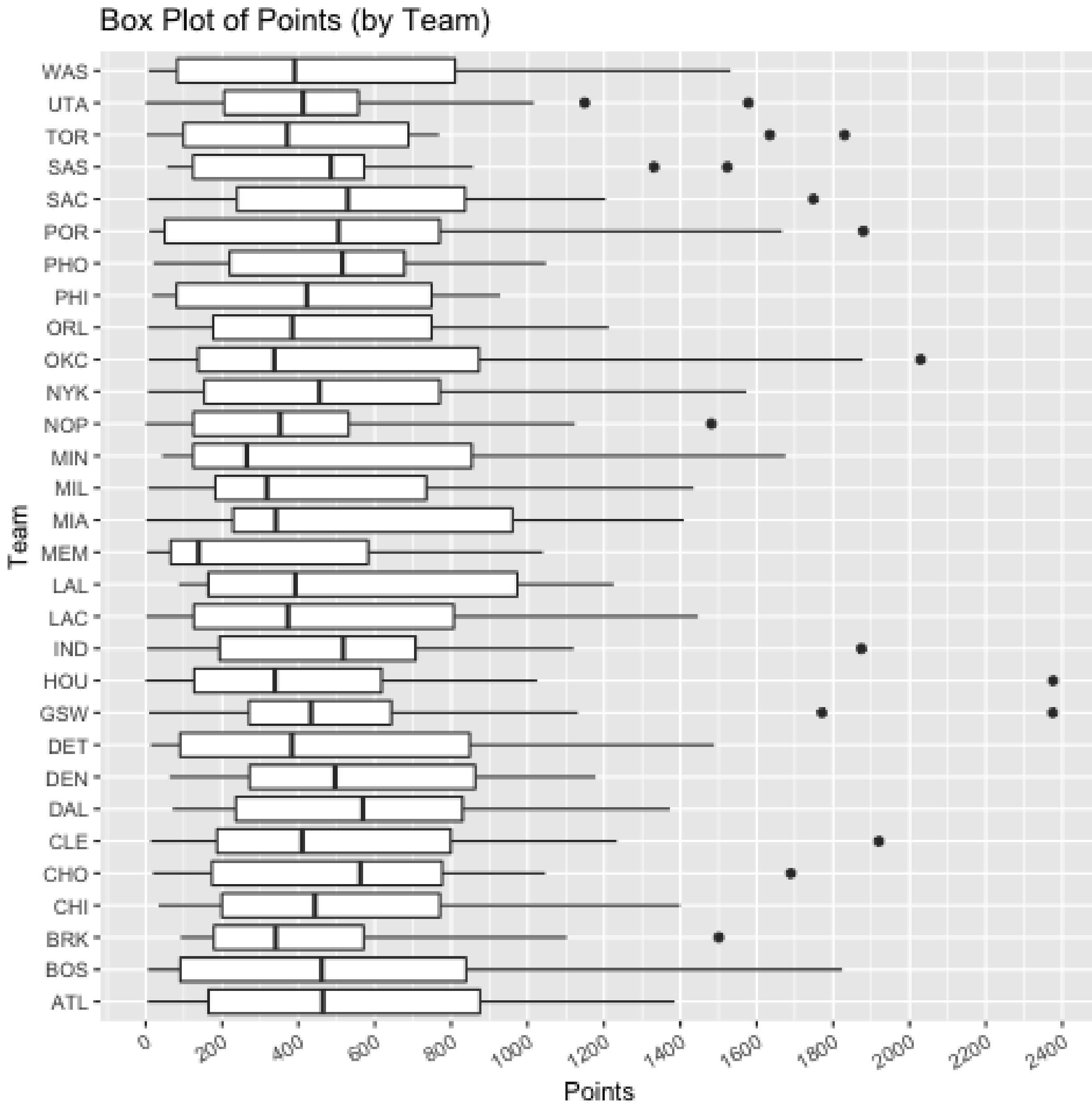


Figure 5: *Boxplot of Points by Team*

From the boxplot of points we can see that team Memphis Grizzlies has the lowest average total points score in the season. Also, we can see that there are outliers in several teams. There are two people who scored the most points in the entire season, one from the Golden State Warriors(GSW) and one from the Houston Rockets (HOU). After checking the data, we identified these two people to be Stephen Curry, with 2375 points, and James Harden, with 2376 points. The person who scored the second-most points in the entire season is from team Oklahoma City Thunder(OKC). This player is Kevin Durant (who coincidentally is playing for GSW in the 2016-2017 season). However, people who contribute large number of points in the season do not get paid significantly more than other players. Thus, we decide to check the relationship between points score in the season by players and their salaries.

Figure 6: *Scatterplot of Point and Salary*

From the scatter plot of points scored in the season and salary, we can see that there is a moderately strong positive correlation of 0.64 between points scored in the season and salaries. About 40.84% of the variation in salaries can be explained by point scored in the season. This make sense because the higher points scored in a season indicates better performance. Team managers would be willing to pay more for players that can help them win games. However, we expected to see an even stronger correlation between points scored in the season and salaries. Thus we decide to see what variable are related to points, how these variables are related to salaries and investigate what other factors determines how much a player gets paid.

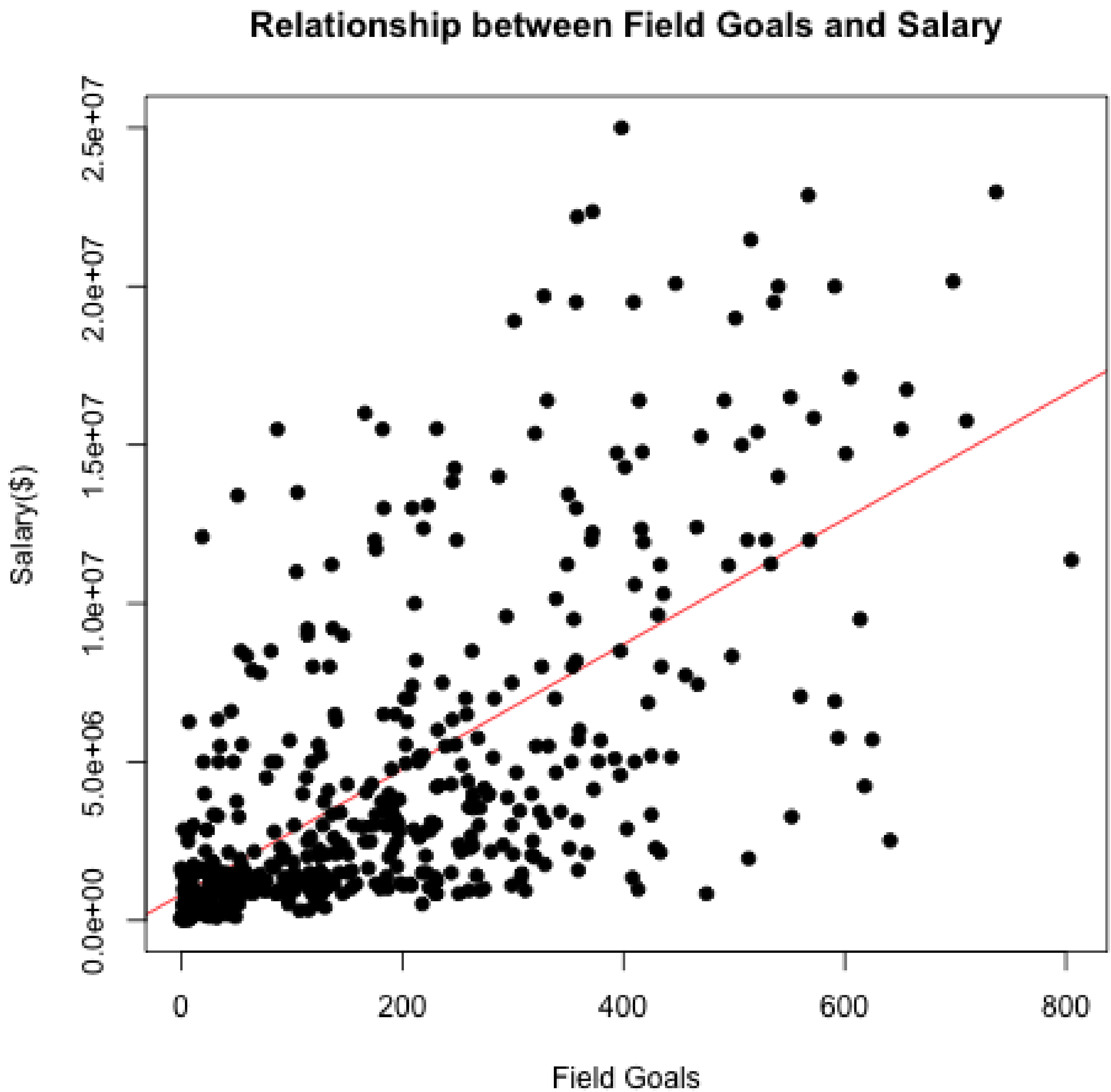


Figure 7: *Scatterplot of Field Goals and Salary*

There are four variables related to points: field goals, three point goals, two point goals, and free throws. The scatterplot of the number of field goals and salaries appears to have a moderately strong positive correlation of 0.64. About 40.97% of the variation of salaries can be explained by the number of field goals. Thus the number of points a player contributes appears to be a main factor that influences his salary. This make sense because a team wins a game by scoring more points than its opponent, and players get points by making field goals. Therefore, a player who is able to contribute more field goals than others will be favored, and team manager will be willing to pay high salaries for these players. However, there are two types of field goals: two points and three points. Thus we further analyze to see whether three or two point field goals influence salary more.

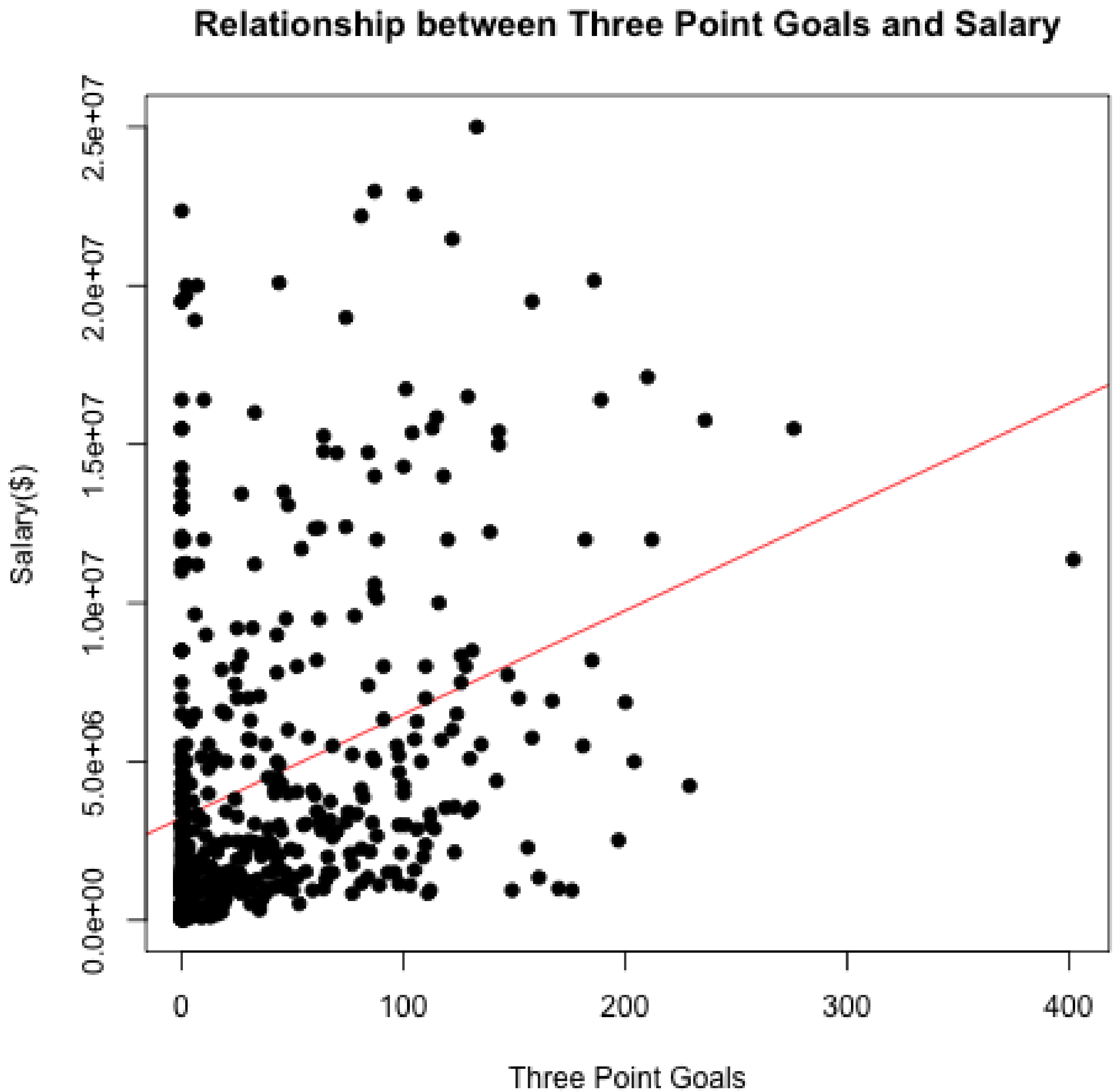


Figure 8: *Scatterplot of Three Point Goals and Salary*

The scatterplot of the number of three point goals and salaries appears to be very scattered and has a weak positive correlation of 0.34. Only 11.34% of the variation of salaries can be explained by the number three point goals. Thus the number of three point field goals a player contributes does not appears to be a main factor that influence the amount of salaries a player get, which is understandable because typically a person's three point percentage is much lower than than their two point percentage. This means that a player will have to make more attempts (and miss more shots) to receive a high number of three point goals. If the team is spending its possessions on missed three pointers, it has fewer opportunities to make two point baskets, which are easier to make and may be more conducive to scoring more points in the long run.

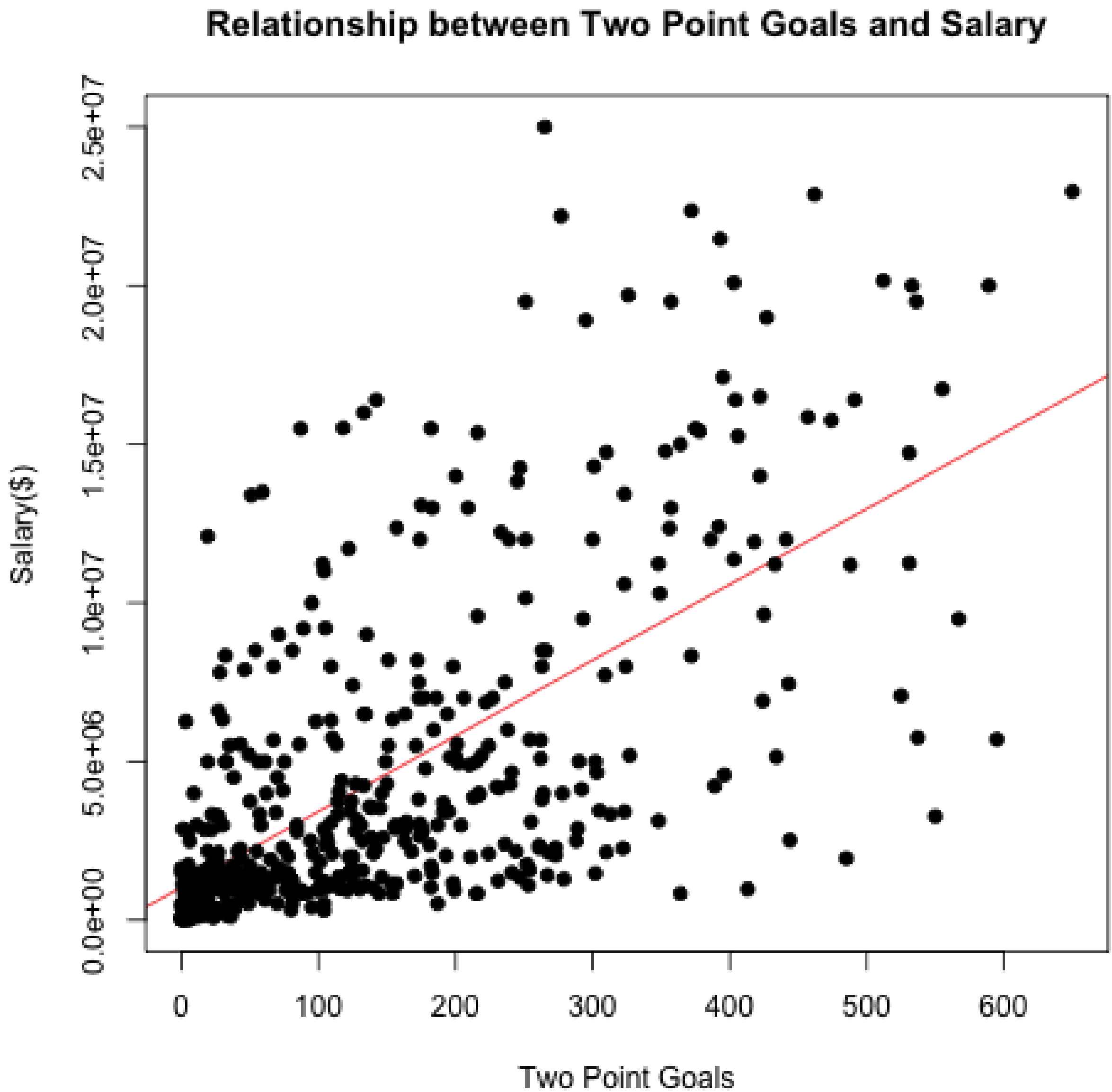


Figure 9: *Scatterplot of Two Point Goals and Salary*

Checking the scatterplot of the number of two point goals and salaries, we found it to be less scattered than the three point field goals and it has a moderately strong positive correlation of 0.64. About 41.14% of the variation of salaries can be explained by the number two point goals. Thus the number of two point goals a player contributes appears to be a main factor that influences the salary of a player.

Another way a player contributes points is by making successful free throws. The scatterplot of the number of free throws made and salaries appears to have a moderately strong positive correlation of 0.6213. About 38.47% of the variation of salaries can be explained by the number of free throws. Thus the number of free throws a player makes also appears to be a main factor that influences the salary of a player.

Figure 10: *Scatterplot of Free Throws and Salary*

Figure 11: *Scatterplot of Total Rebounds and Salary*

Since basketball is a team game, other skills that help the team such as rebounds, blocks, assist and steals are also critical. The scatterplot of the number of total rebounds and salary appears to have a moderate positive correlation of 0.5299. About 27.92% of the variation of salaries can be explained by the number of total rebounds. Thus the number of total rebounds a player contributes may be a main factor that influences a player's salary. We further investigated two types of rebounds: offensive and defensive.

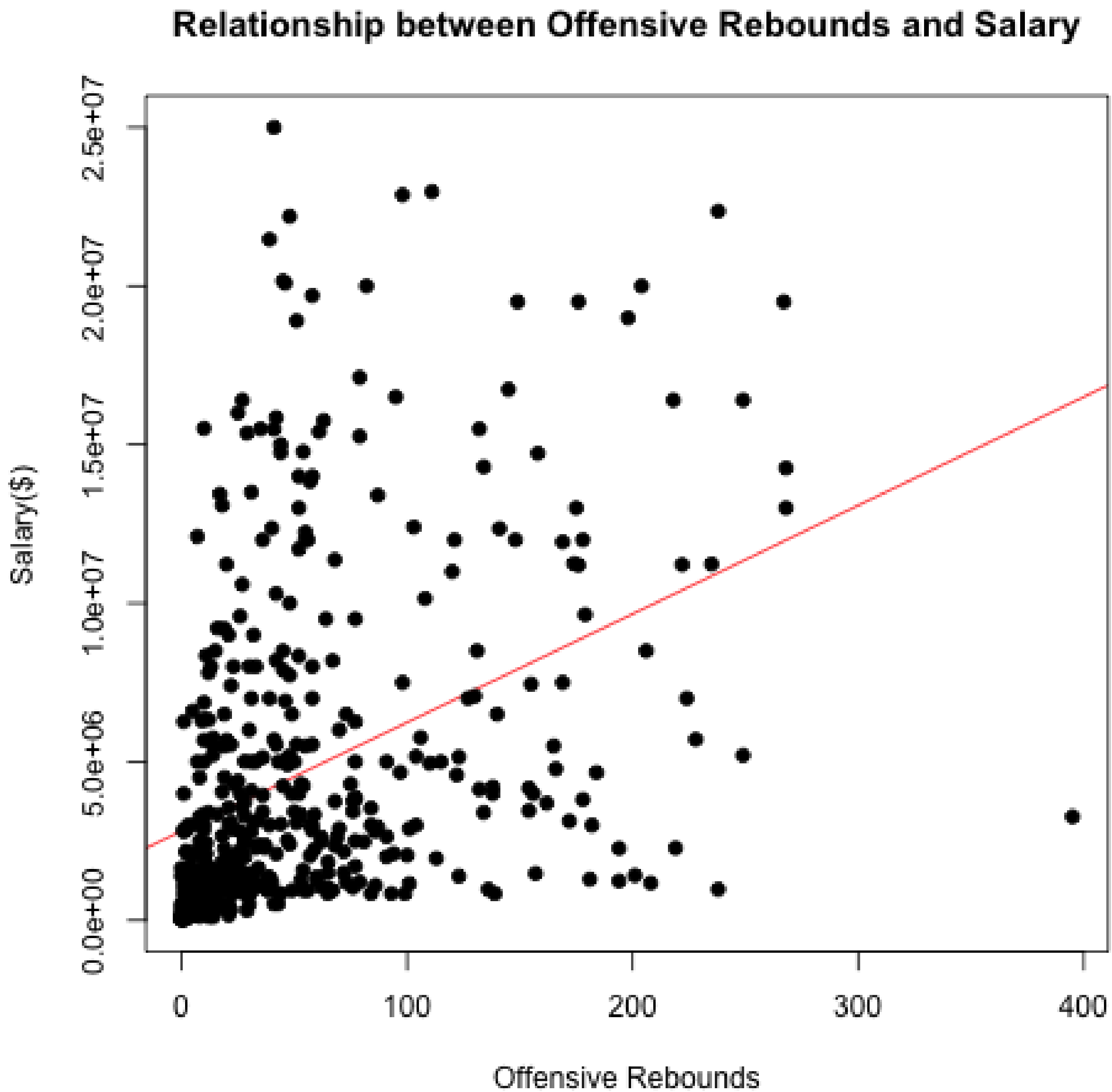


Figure 12: *Scatterplot of Offensive Rebounds and Salary*

The scatterplot of the number of offensive rebounds and salaries appears to have a weak positive correlation of 0.3935. Only 15.31% of the variation of salaries can be explained by the number of offensive rebounds. Thus the number of offensive rebounds a player contributes does not appears to be a main factor that influences a player's salary.

The scatterplot of the number of defensive rebounds and salaries appears to have a moderate positive correlation of 0.5592. About 31.12% of the variation of salaries can be explained by the number of offensive

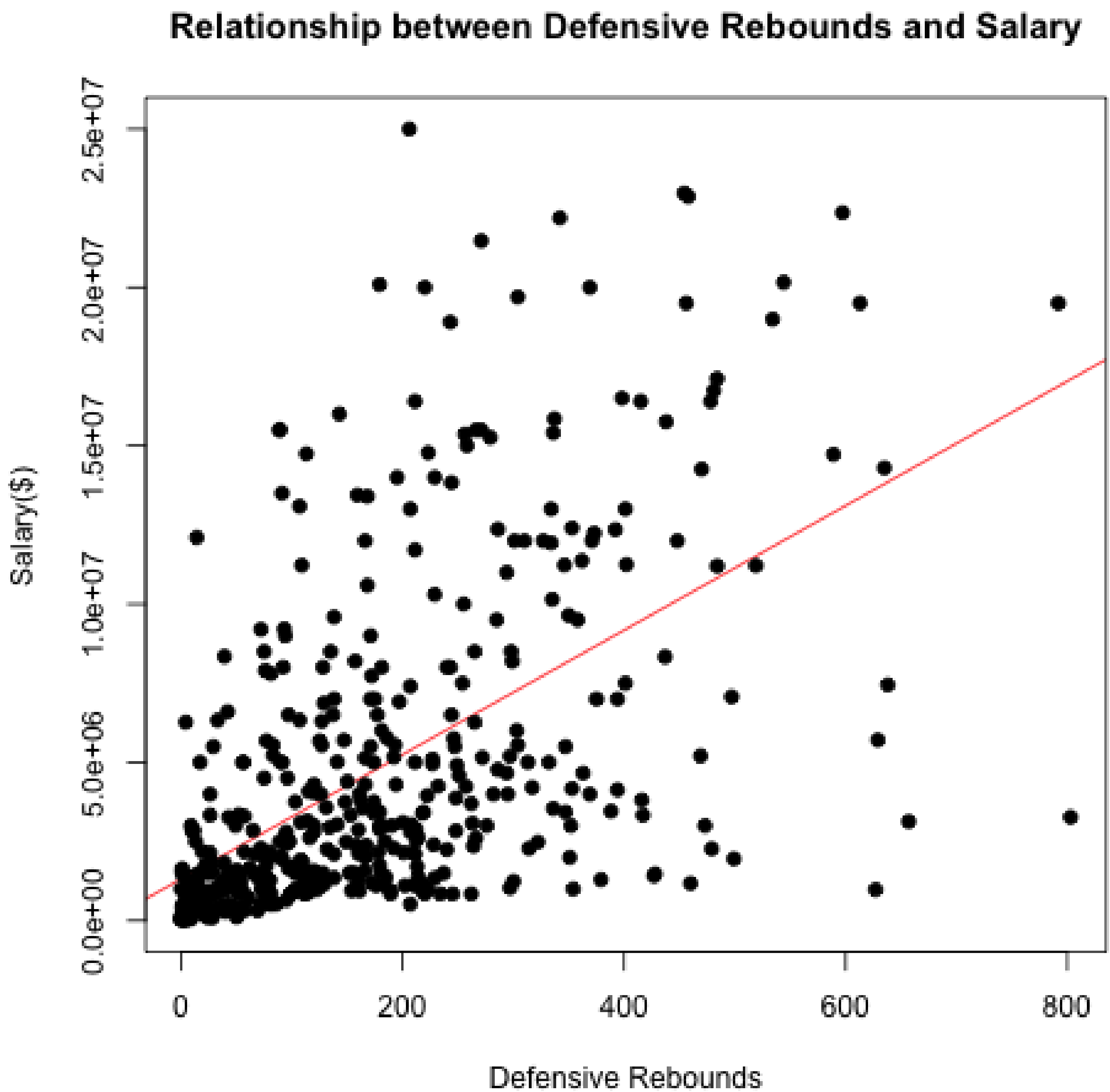


Figure 13: *Scatterplot of Defensive Rebounds and Salary*

rebounds. Thus the number of defensive rebounds a player contributes appears to be a main factor that influences a player's salary. It should also be noted that defensive rebounds seem to be more closely related to a player's salary than offensive rebounds. This is to be expected, since it usually is more important to stop an opposing team from getting a rebound close to the basket than rebounding for your own team.

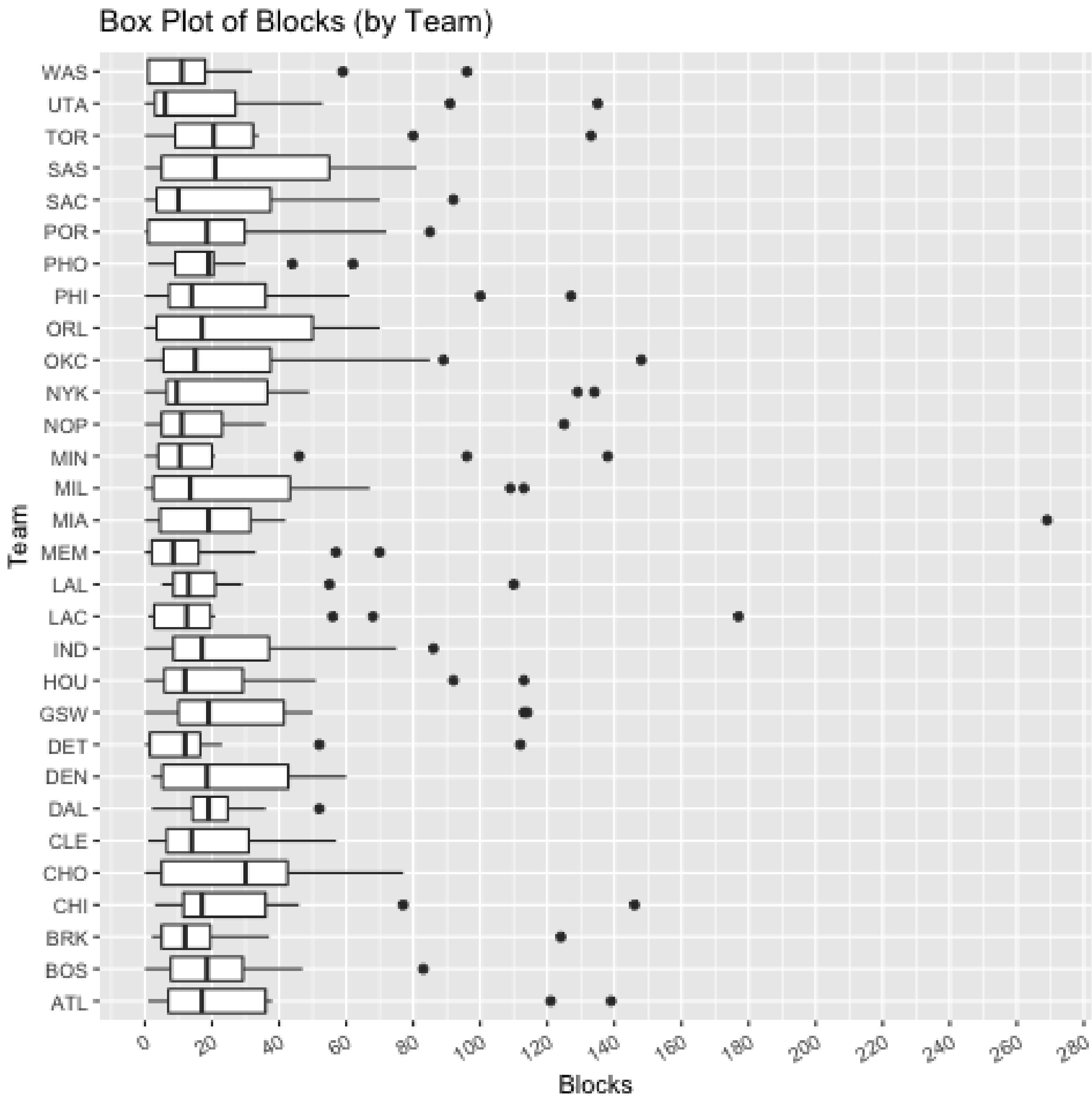


Figure 14: *Boxplot of Blocks and Salary*

Beside rebounds, a player can also help the team by blocking its opponents shoots. From the boxplot of blocks we can see that most team have average blocks around 20. The Charlotte Hornets (CHO) have the highest number of blocks in the season. Many teams have a player that contributes a lot more blocks than the other players in the team. The player who generates the most number of blocks is Hassan Whiteside (269 blocks) from the Miami Heat (MIA). However, after checking the data, this person does not have a high salary compare to players who scored many points in the season. Thus we decide to check the correlation between blocks and salaries.

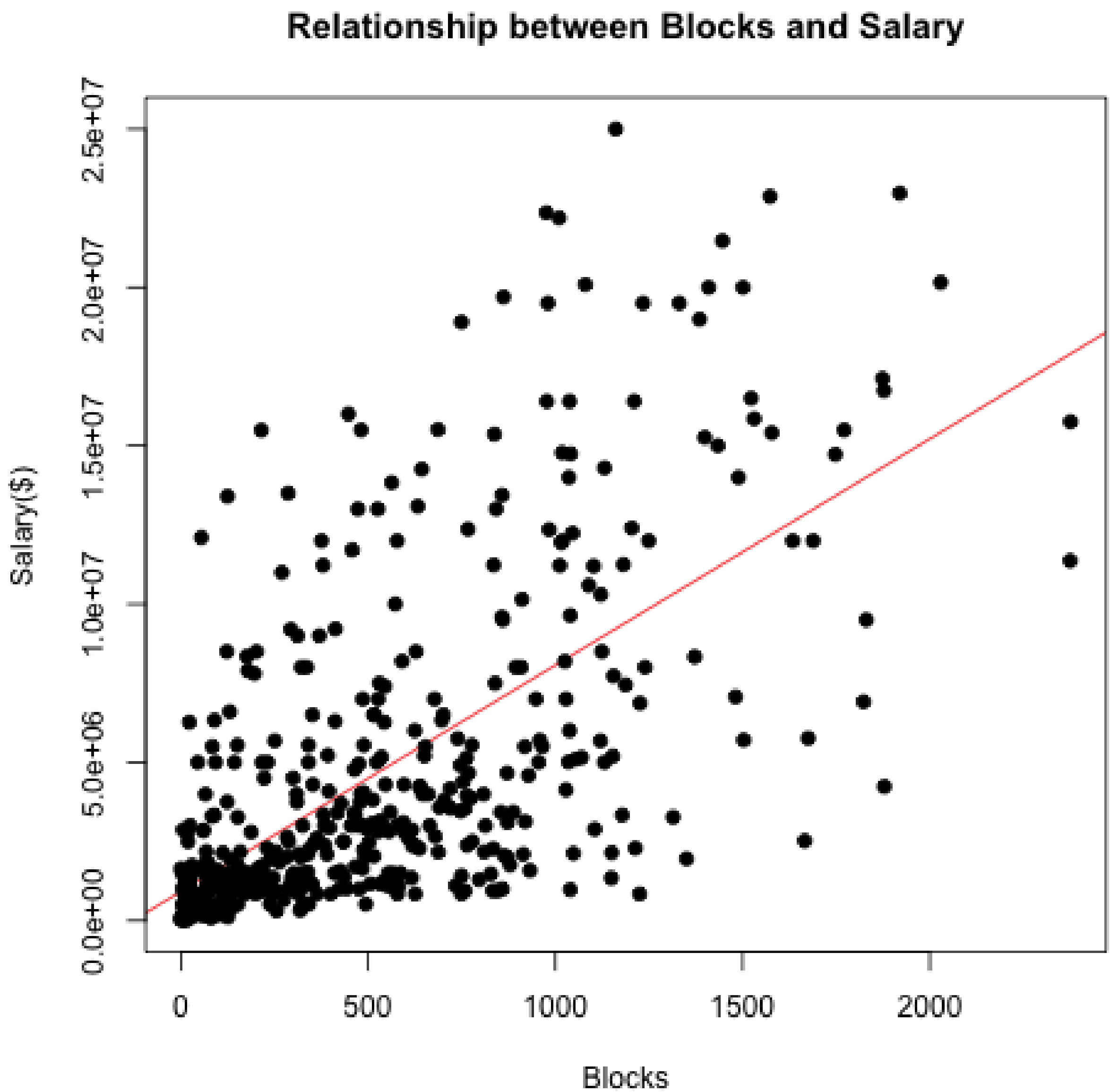


Figure 15: *Scatterplot of Number of Blocks and Salary*

The scatterplot of the number of blocks and salaries appears to have a weak positive correlation of 0.3589. Only 12.69% of the variation of salaries can be explained by the number blocks. Thus the number of blocks a player contributes does not appears to be a main factor that influences their salary.

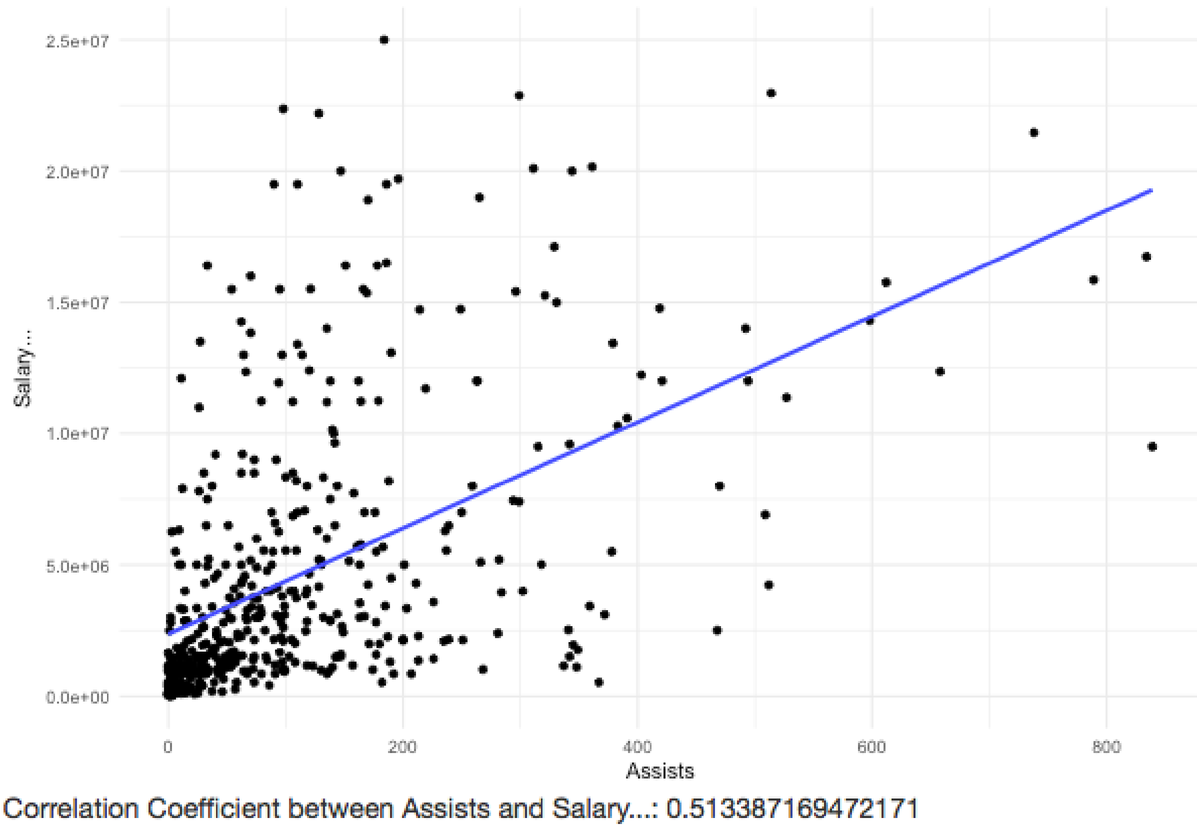
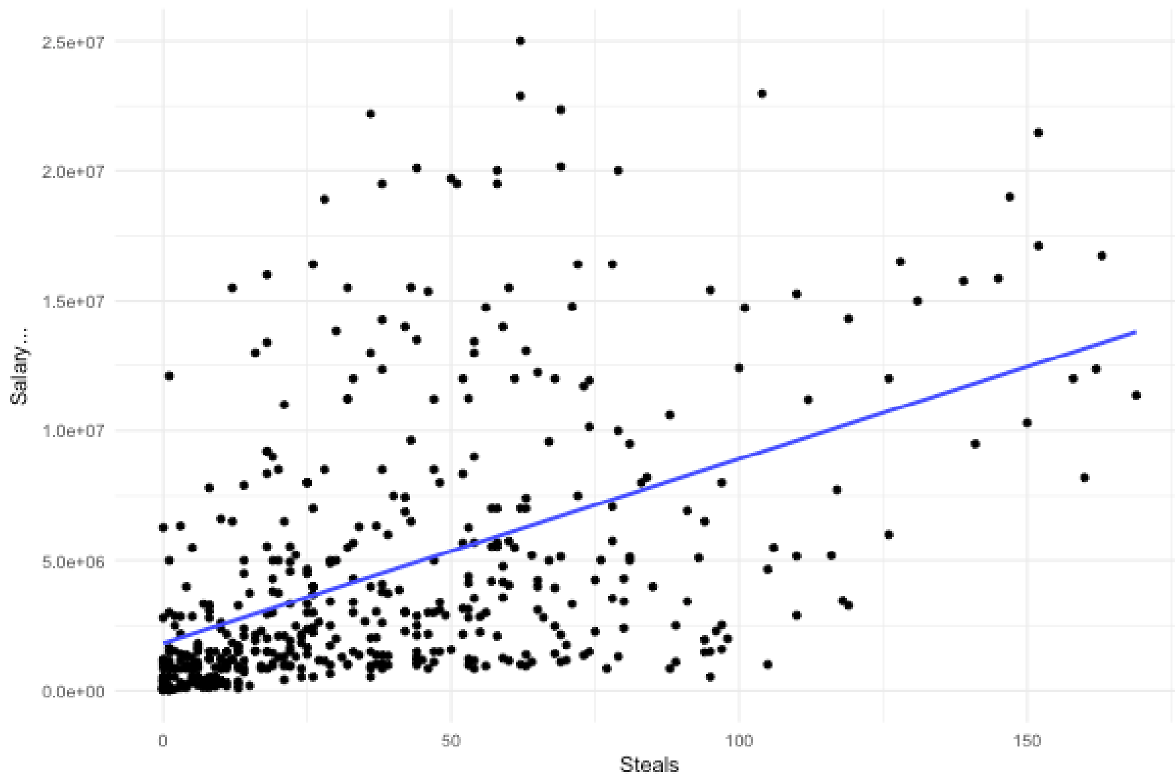


Figure 16: *Scatterplot of Number of Assists and Salary*

The scatterplot of the number of assists and salary appears to have a moderate positive correlation of 0.5134. Only 26.20% of the variation of salaries can be explained by the number of total assist. Thus the number of assists a player contributes appears to be a main factor that influence a player's salary.

The scatterplot of the number of steals and salary appears to have a moderate positive correlation of 0.4857. About 23.43% of the variation of salaries can be explained by the number of steals . Thus the number of steals a player contributes may be a main factor that influence the salary of a player.

Besides scoring and assistant skills, mistakes players made have surprising relationship with salary. Measurements for mistakes are turnovers and personal fouls. The scatterplot of the number of turnovers and salary appears to have a moderately strong positive correlation of 0.5828. About 33.83% of the variation of salaries can be explained by the number of turnovers. This result is surprising at first because normally people would think the more turnovers a player has, the worst his basketball skills are, and thus the lower he will be paid. However, if we approach these findings with a different mindset, it makes sense. More turnovers means that the player have more possession of ball, and score more points, meaning that the player has better skills and a hihger salary. However, handling the ball more often means that if you have the same turnover percentage as other players, you will have more total turnovers. The relatively high correlation means that the number of turnovers a player contributes is directly correlated to the number of points they scored, and this appears to be a factor influences a player's salary.



Correlation Coefficient between Steals and Salary...: 0.485732140216002

Figure 17: *Scatterplot of Number of Steals and Salary*

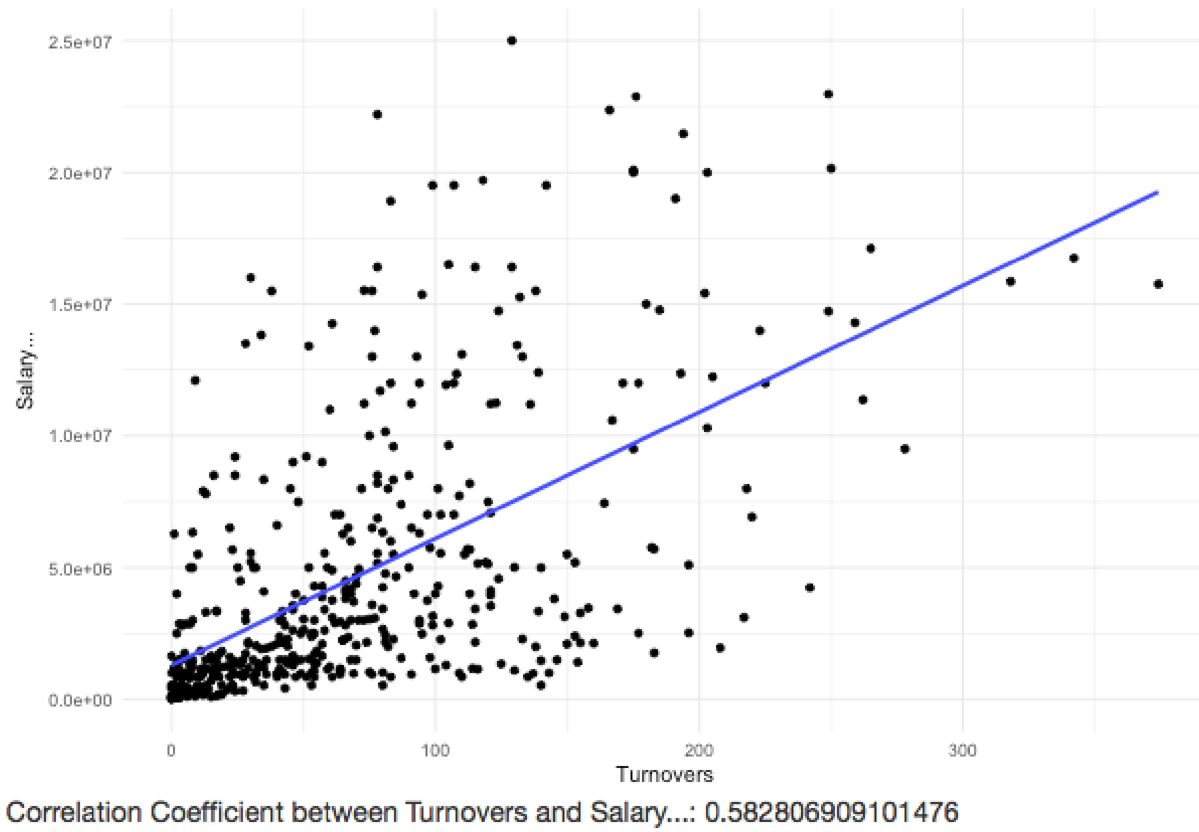


Figure 18: *Scatterplot of Number of Turnovers and Salary*

Figure 19: *Scatterplot of Number of Personal Fouls and Salary*

The scatterplot of the number of personal fouls and salary appears to have a moderate positive correlation of 0.4451. About 19.64% of the variation of salaries can be explained by the number of total rebounds. Similarly to the turnover statistics, this result doesn't make sense at first. However, large amounts of personal fouls may also mean that the player is aggressive and will probably contribute more points. Teams would want to pay higher salaries for these players. Thus the number of personal fouls a player draws may influence their salary.

Efficiency and Value Analysis

Exploratory data analysis of measurements for skills shows that total points, field goals, especially two point goals, and free throws were the three main contributors to player's salary. Moreover, rebounds, especially defensive rebounds, assists, turnovers, and personal fouls are also correlated with player's salary. However, players in different positions have a different skill focus. For example, a player who plays shooting guard will have much higher number of field points recorded than a player who plays power forward. Thus we further calculated weighted efficiency by position to account the different role each position has. Then we graphed a scatter plot between efficiency and salary to see the relationship between a player's efficiency and his salary.

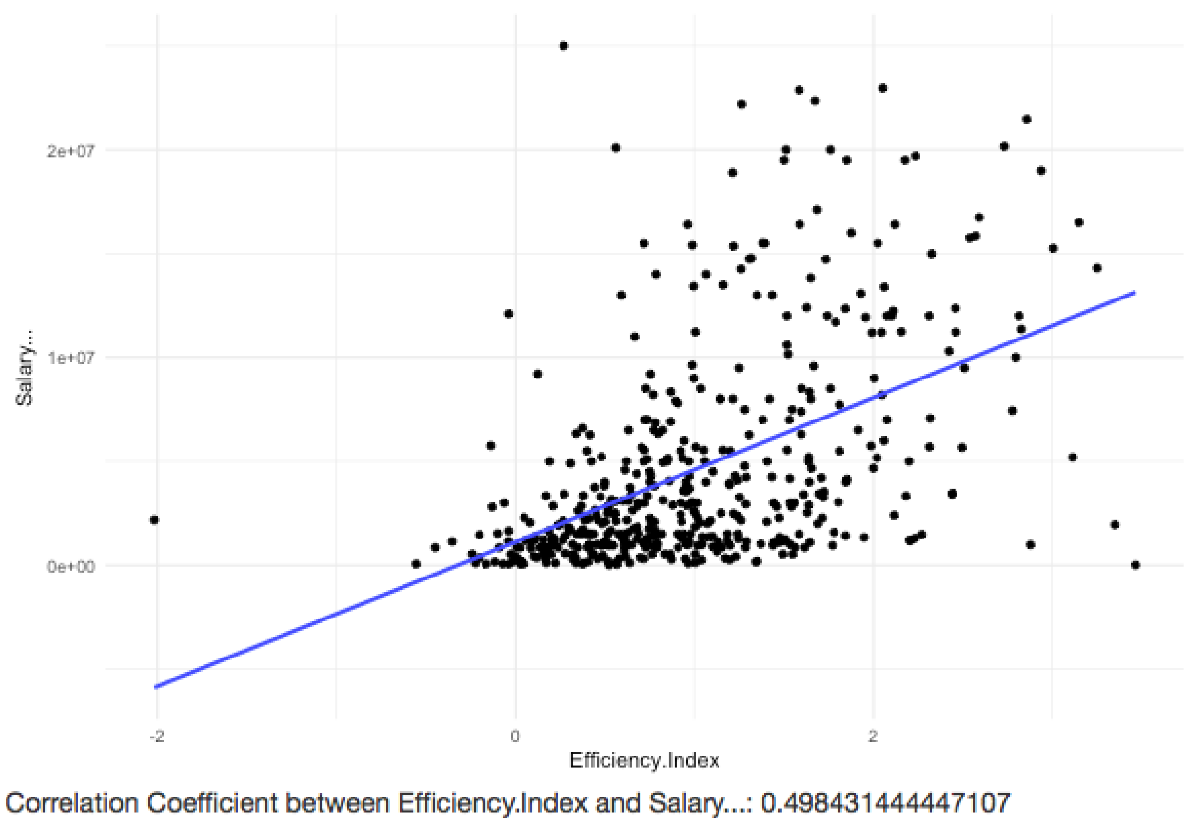


Figure 20: *Scatterplot of EFF Indices and Salary*

The graph above illustrates the relationship between efficiency of a player and his salary. Each point represents a player and is colored by his position. Hence there are total 471 points across the x axis. The line on the scatterplot is the regression line, which represents the "standard salary" a player should get with different efficiency. Dots above the regression line represents players who receive higher salary than the average salary a player gets with the same efficiency, which may indicate an overpayment. Similarly, dots below the regression line presents players who receive less salary than the average salary earned by players with the same efficiency, which may indicate an underpayment. The further a dot is from the regression line, the more differentiation is the player's salary from the "standard salary".

At the first glance, we observe that dots are pretty spread and there is an positive correlation of 0.498 between efficiency and salary. This means that there is an positive relationship between efficiency and salary; however, efficiency cannot account for all the variations in salaries. In fact, efficiency can only account for about 25% of variations in salaries. After taking a closer look, we can see that the majority of players have salaries below \$1 million. In addition, there are also couple outliers. The most obvious one is a blue dot on top of the graph that represents a player in the small forward position who has an efficiency of less than 0.5 but receives the highest salary. After checking the data set, we found this player to be Kobe Bryant. Another outlier we spotted is a green dot which represents a player in position power guard that has the lowest efficiency but not the lowest salary. After checking the data, we found this person to be Tony Wroten. These two outliers are possible players that get over paid. As for players that might be underpaid, there is a blue dot on the rightmost of graph which represents a play in position small forward that has the highest efficiency (3.464) but gets a extremely low salary (\$8819). This player turns out to be Dahntay Jones. However, this player could be an outlier since he only played one game, but maybe had a really good game, which artificially drives up his statistics. After research online, we found that Dahntay Jones has over 10 years of NBA experience and is actually one of the highest paid NBA player, who received 1.5 million salary in year 2015. A possible explanation is that the data on Basketball Reference is wrong. Another player that might be underpaid is Giannis Antetokounmpo who is represented by the green dot on the leftmost of the graph under the regression line. He has the second highest efficiency (3.350) but receives a salary much below the "standard salary". One possible reason is that as a 21-year-old player, he is new to the NBA. Although people start to see him to shine as the primary player for the Milwaukee Bucks, Giannis still need to play more games to gain attention and show how valuable he is.

Best and Worst Value Players

	Best.Player	Worst.Player
1	Dahntay Jones	Jimmer Fredette
2	Briante Weber	Orlando Johnson
3	Nate Robinson	Justin Harper
4	Xavier Munford	Coty Clarke
5	Jordan Hamilton	Tony Wroten
6	Henry Sims	Keith Appling
7	Jordan Farmar	Elliot Williams
8	Jared Cunningham	Nazr Mohammed
9	Alan Williams	Spencer Dinwiddie
10	Axel Toupane	Luis Montero
11	Ty Lawson	Josh Huestis
12	Marcus Thornton	Mitch McGary
13	J.J. O'Brien	JaKarr Sampson
14	Jason Thompson	Jordan McRae
15	Erick Green	Bruno Caboclo
16	Michael Beasley	Aaron Harrison
17	Bryce Dejean-Jones	Sonny Weems
18	Joe Johnson	Archie Goodwin
19	T.J. McConnell	Andrew Wiggins
20	Hassan Whiteside	Lou Amundson

Figure 21: *List of Best and Worst Value Players*

From the list of 20 best and worst players, we can see that the best values player is Dahntay Jones. This make sense because we calculate values using the formula value = efficiency/ salary. Dahntay has the highest efficiency and an extremely low salary. There is no doubt that he will be the best valued player. But as mentioned above, he is an outlier.

Conclusion

The purpose of this project is to analyze the performance of a player and investigate the relationship between skills and salaries. We found that total points, field goals (especially two point goals), and free throws were the three main contributors to player's salary. Moreover, rebounds (especially defensive rebounds), assists, turnovers, and personal fouls were also statistically significant. In regards to assists, teams may be focusing on a player's ability to contribute to scoring. Additionally, in the case of rebounds, a player's value can be enhanced if he is able to either prevent the opponent from another scoring chance by grabbing defensive rebounds and conversely, providing additional scoring chances for his team by grabbing offensive rebounds. Turnovers were also found to be a significant contributor in this study. The reason that a player who has a large number of turnover is that he has more possession of the ball. As for personal fouls, higher number of fouls means this player plays aggressively and maybe able to contribute more points. Thus, a player who does accumulate fouls may still be a valuable asset to his team.

Overpaid and Underpaid Players

We further calculated weighted efficiency that accounts for role difference in different positions as a representation of a player's skills. We found a positive relationship between efficiency and salaries and also spotted outliers that represent players who were overpaid or underpaid. Namely, Kobe Bryant and Tony Wroten may be overpaid since dots represented them are way higher than the regression line. Dahntay Jones appears to be underpaid since he has the highest efficiency but an extremely low salary. However, this could be result from an data error. We believe that Giannis Antetokounmpo is underpaid because he has the second highest efficiency but much lower salary than players with lower efficiency indices.

Application and Further Analysis

With the burgeoning field of basketball analytics, teams are focusing on a multitude of metrics and developing formulas to determine player efficiency. This practice in theory should influence NBA team management decisions when it comes to determining player salary. This may signal a change in thinking among NBA front office personnel. Further possible analysis is calculating long term efficiency of a player to see his potential and consistency of his performance. Other possible analysis includes relationship between college and efficiency and region and efficiency.