Brandeis University

INTERNATIONAL BUSINESS SCHOOL

# Prediction of NBA Players' Salary Based on Machine Learning


**Group Member:**

Luhan Shen, Jiawei Zhang


Final Project 2

183BUS-211F-4: Analyzing Big Data I

Instructor: Prof. Ahmad Namini

December 19, 2018

## 1. Introduction

As a popular sport, basketball offers players in the NBA truly high salaries and their salaries have attracted lots of attention. 19 Of the 25 highest-paid teams in the world are from the NBA, and overall, the average NBA player makes 7.1 million, top in the world, and up from $6.4 million last year.

We are also very interested in figuring out the determinants of NBA players' salaries. Using data in the past to predict a player's salary is very meaningful, which can serve as a reference to team managers when they decide to sign a contract with a player.

According to economic theory, people get higher profits when they can make higher productions. It's the same to NBA players. If they can contribute more to the team, they will be paid more. So, the determinants of salaries should include the important statistics that can best describe the players' performance and ability.

## 2. Data Collection

We collected data from different sources. Firstly, we got the statistics of players from Kaggle.com, which can comprehensively describe a player's performance. However, the salary in this dataset is just in the 2017-2018 season. Salary is decided before a season begins according to last year's performance. Therefore, we searched and got the players' salary for season 2018-2019 from basketball-reference.com. The bad thing is, there is still an important problem in the dataset, which is missing points per game (PTS) for every player. PTS is the most direct way to judge a player, since the win/loss depends on the score of the team. So, at last, we got the PTS data from NBA official website.

## 3. Data processing

1. Removing useless rows

We observed a lot of data problems in our 3 datasets. In the salary dataset, there are quite a few rows with missing values or just the word "Player" in the player' name column. We directly removed them because they provide little information.

2. Adjusting data type

Some variables are in the character type, violating our expectation of numbers type. In addition, the salary values are all shown in the format "$xxx,xxx". Therefore, we used *as.numeric* function to adjust the data types and *gsub* function to remove the dollar sign and comma sign.

3. Substituting NAs

At last, there are lots of NAs in the Kaggle dataset, which will influence our analysis. In order to keep adequate number of observations, we calculated the average of the columns and substituted the NAs with the average value.

## 4. Data Description

1. Variables

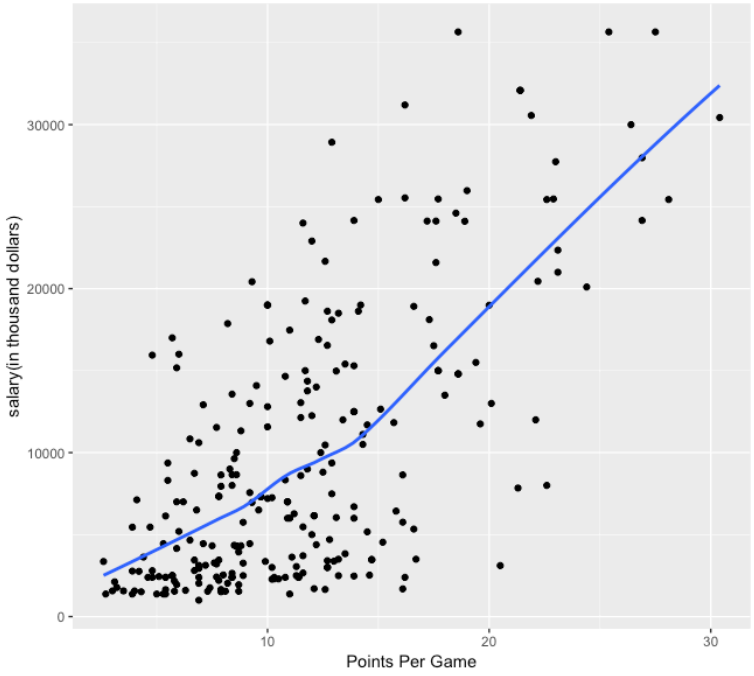
The table below is our variables used in our analysis with the explanation for them. Salary is the dependent variable, and all the others are independent variables.

Table 1: Explanation of variables

| Variable | Explanation |
|---|---|
| DraftNbr | the order of player to be picked by his team |
| Age | age of player |
| Games | games played this season |
| Minutes | total minutes played this season |
| Efficiency | Player Efficiency Rating |
| ShootingPCT | True shooting percentage, a measure of shooting efficiency that takes into account field goals, 3-point field goals, and free throws |
| X3pointsPCT | 3-Point Field Goal Percentage |
| Rebounds | average rebounds per game |
| Assists | average per game |
| Blocks | average blocks per game |
| Turnovers | average turnovers per game |
| Win_Share | an estimate of the number of wins contributed by a player |
| BPM | Box Plus/Minus, a box score estimate of the points |
| VORP | Value Over Replacement Player |
| Pos | Position played |
| Salary | Salary for season 2018-2019 |
| PTS | Average points per game |

We tried to make an initial analysis of our dataset according to our understanding of NBA. After data processing, we have altogether 266 observations with 17 variables, all of which are numeric variables except for the factor variable 'position'. According to common knowledge, points, rebounds, assists, blocks and turnovers are the most important and direct statistics in NBA. So, we draw graphs to visualize the relationship between them.

Figure 1: strong linear correlation between PTS and salary




Among these determinants, points per game have the most obvious positive linear relationship with salary. Assists and rebounds have different effects on salary of players in different positions. PG and SF's salary are not determined by their assists number. And Cs can gain most benefit in the increase of rebounds number. Among all the positions, C has the highest median salary, while PG has the lowest. But the difference between the salary in different positions are rather small. Besides, efficiency, minutes, win share and VORP all represent players' ability and positively increase their salary. It's the same to our expectation, cause these variables are usually used to evaluate a player. What's worth mentioning is turnover data. We suppose it to be negatively relate to salary, because team will lose lots of points. But the graph shows an opposite view. It may be due to players who make many turnovers must have high control of the ball. These players are more likely to be the superstar players.

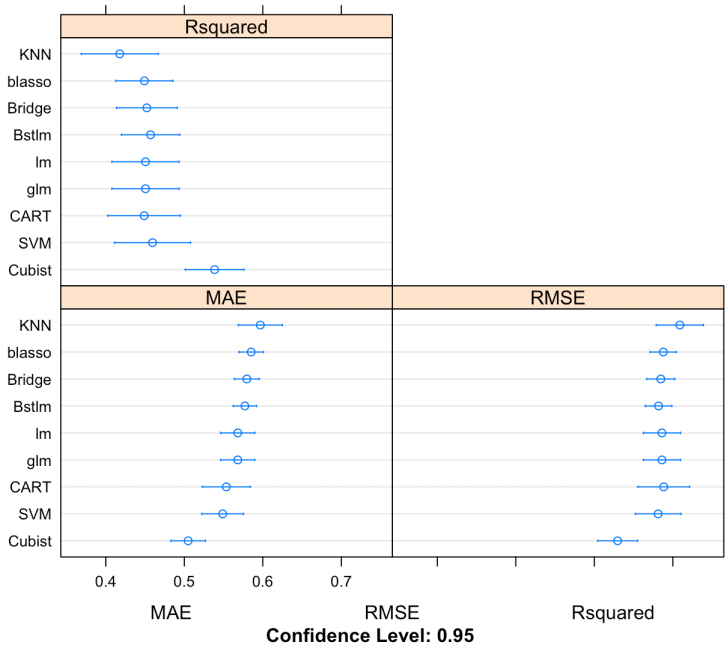
**5. Data transform and split**

To better conduct the analysis, we transformed all the independent variables to normal distribution. And we took the value of salary by its natural log. Then we split the dataset into 80% part for training and 20% for validation.

## 6. Model Training

To conduct linear regression, we first used resampling methods *repeated cross-validation* to create a set of modified data sets from the training samples. We applied five different methods including *lm, glm, svmRadial, rpart, knn, bridge, blassoAveraged, BstLm* and *cubist* to train the model. The results of training are shown in the graph below. The model with *cubist* method has the highest average $R^2$ and lowest average MAE and RMSE value with obvious difference than other algorithms. Therefore, it's the most accurate and effective among the nine models to make prediction.

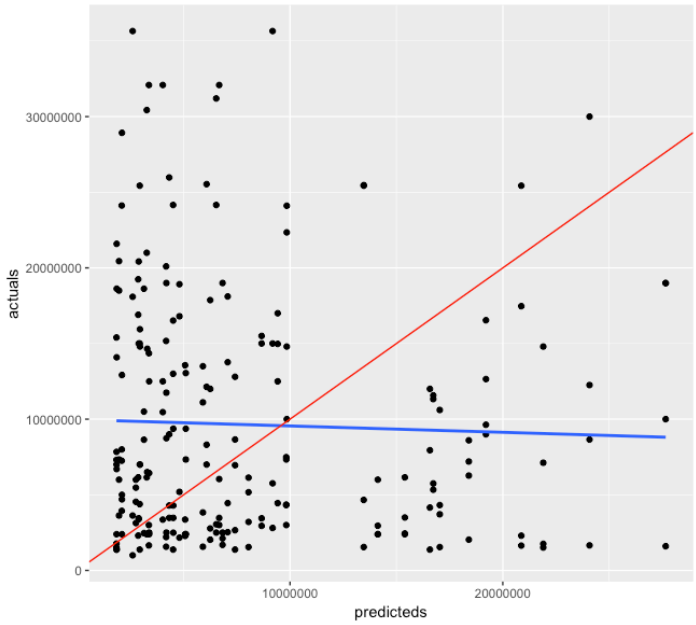
Figure 2: Accuracy of all the algorithms




## 7. Model Validation

In order to test the accuracy of our selected model, we use the validation part of dataset to test the model. We put the predicted salary and observed actual salary in a scatter plot to see their relationship. The blue line in the graph is the fitted line of the points, the slope of which is far from the red 45° line. The prediction of model is rather biased. As simpler methods, linear algorithms are common to have a strong bias. But the advantage is that they are fast to train.

Figure 3: Prediction result




## 8. Improvements

During the process of dealing with NAs, we used the average value to substitute the NAs. It may cause a serious bias to the estimation. Because originally the players with NAs might have totally different level of ability from the average level. By using the average, we overstate or understate their performance. While these rows of data are used to train the model, they may bring wrong information that will mislead the training process and lead to bad results.

We may have omitted variable bias in the model. Players' salaries should also be determined by their team tradition, if the player is liked by local audience and the salary level of other players in the same team. Variables that can describe these aspects may all considered to be included in the model. However, these kinds of data are hard to find or cannot be recorded as numbers.

The last weakness we should mention is that we did not use a large number of algorithms to compare. Maybe there are other more effective algorithms to this analysis.

The large bias of result may also be caused by the small number of samples. After data pre-processing, we have only 214 observations in our dataset, and the training set contains only 80% of it, only 171 observations. So, perhaps the small sample is not representative enough to show the characteristics of NBA players' salary.

## APPENDIX

Figure 4: relation between assists and salary for players in five positions

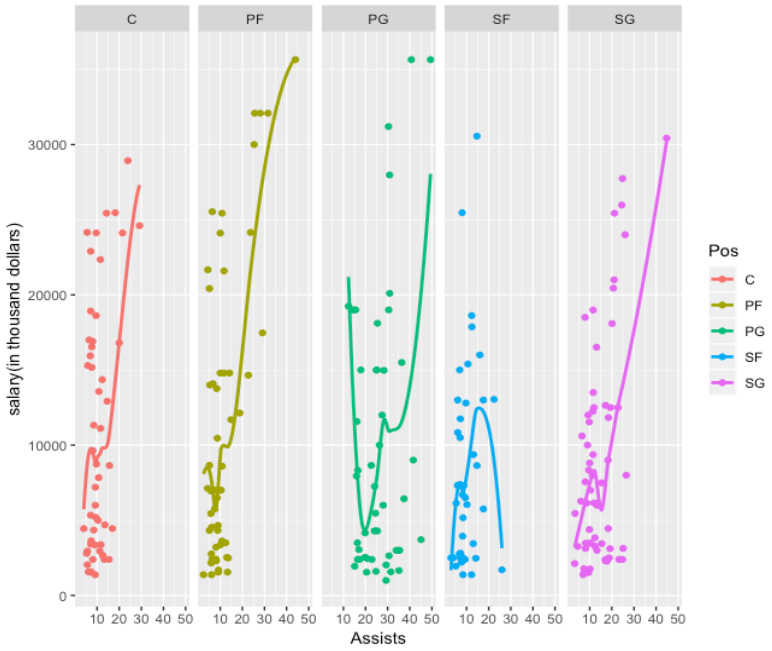



Figure 5: C players have largest benefit to increase number of rebounds

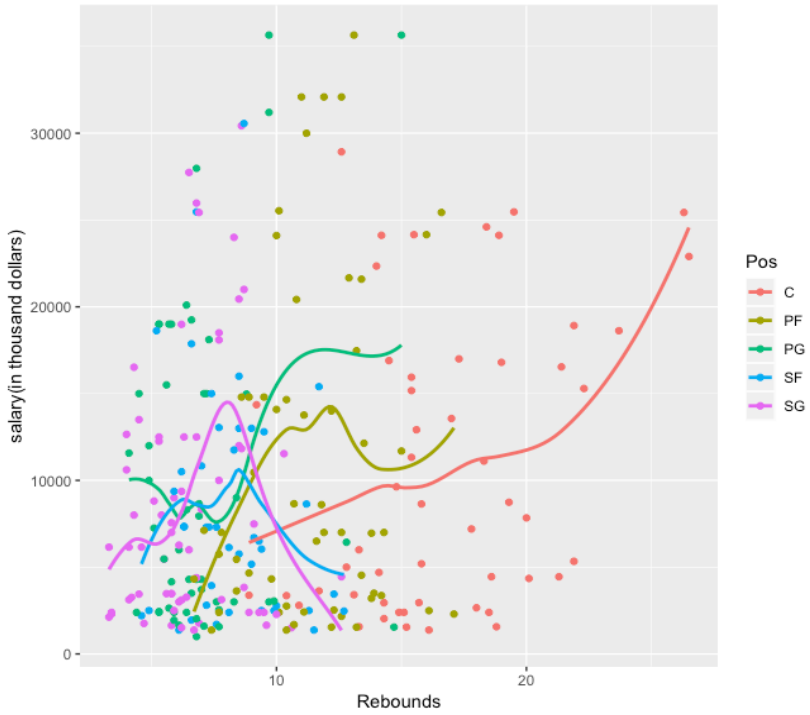


Figure 6: salary level for all positions of players

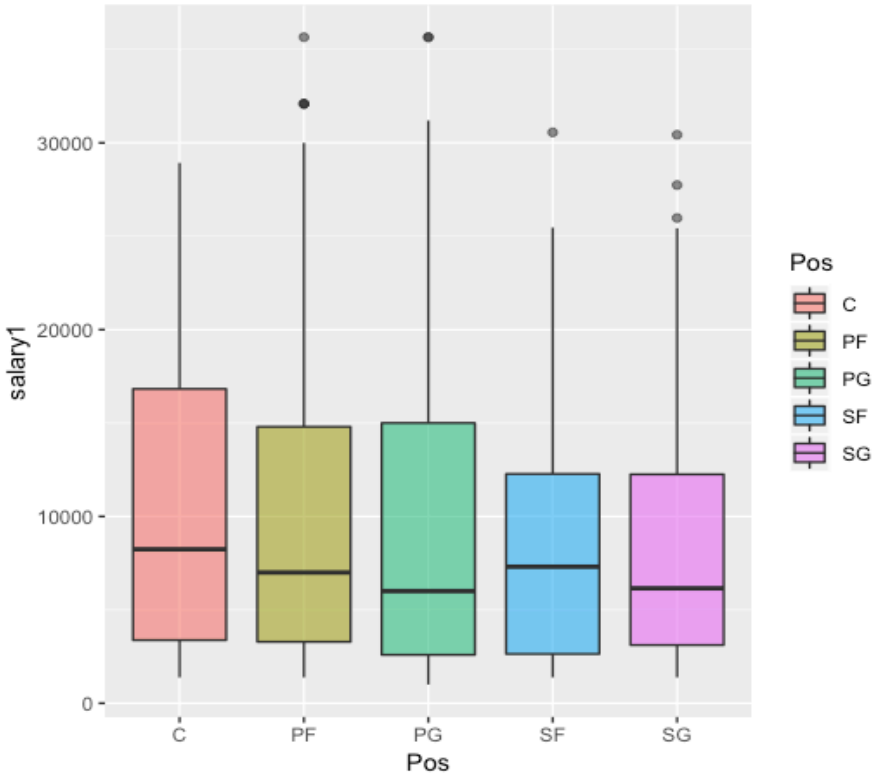



Figure 7: strong positive linear relation between efficiency and salary

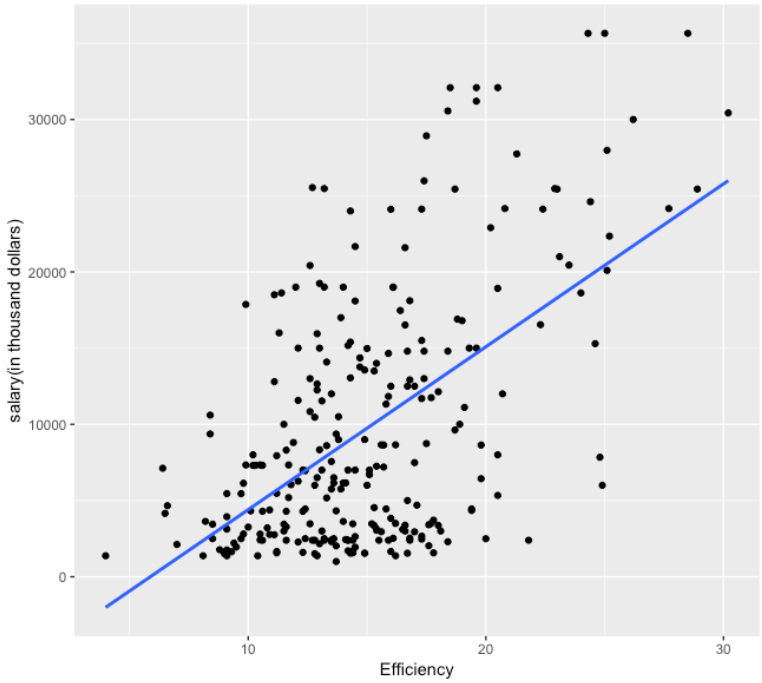


Figure 8: salary has no obvious relationship with turnover numbers

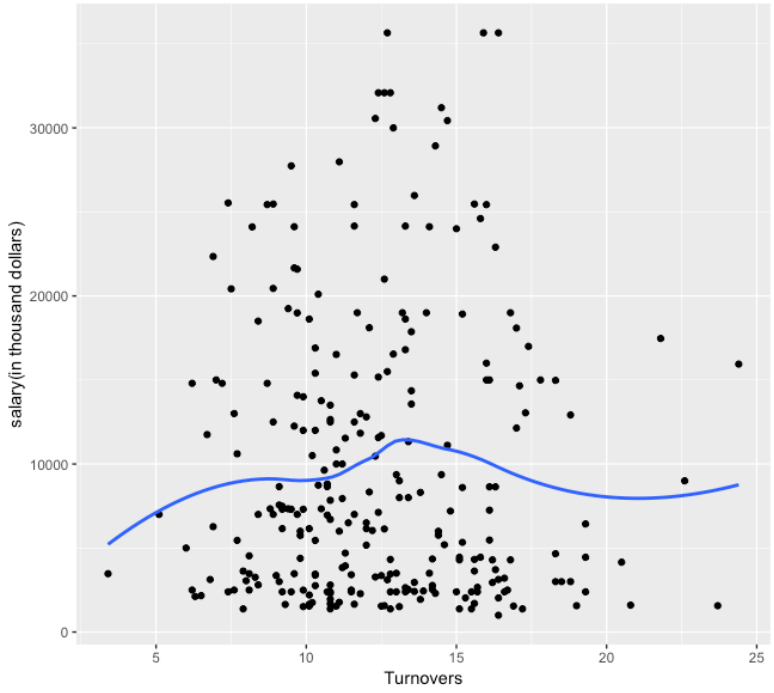



Figure 9: median and range for all the variables

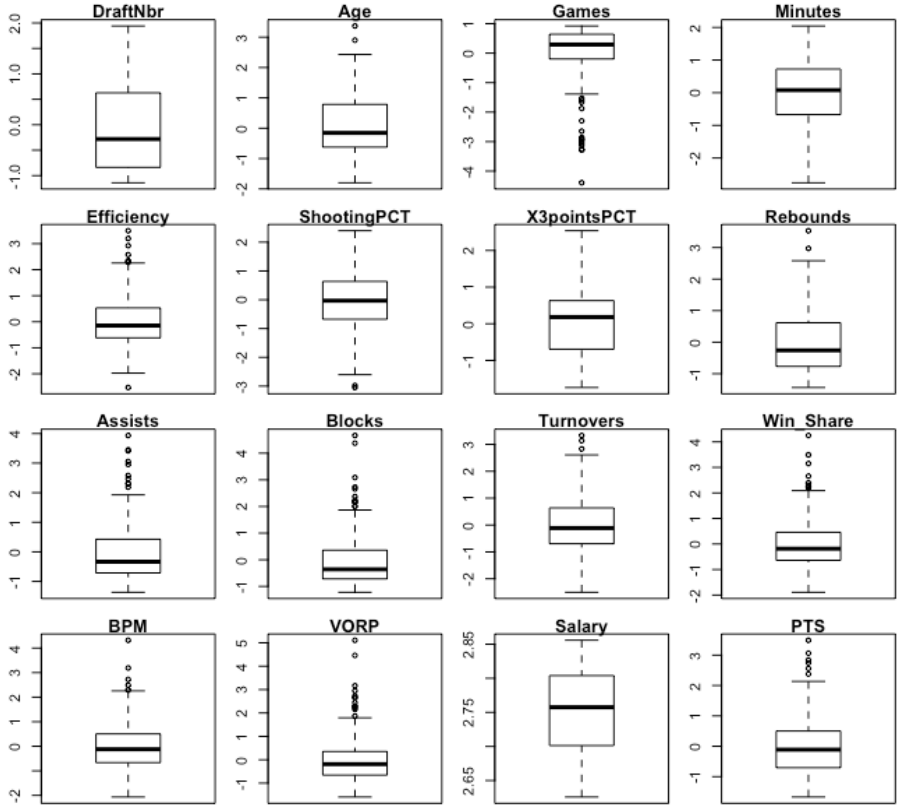


Figure 10: distribution of all variables

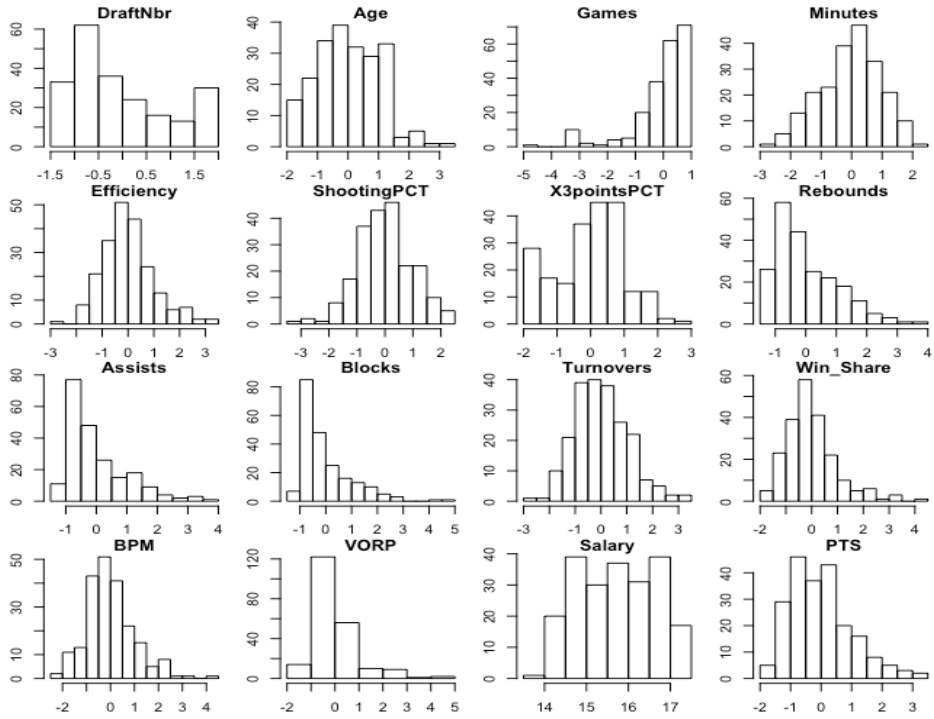



Figure 11: density of all variables after adjusting to 0 mean and 1 standard deviation

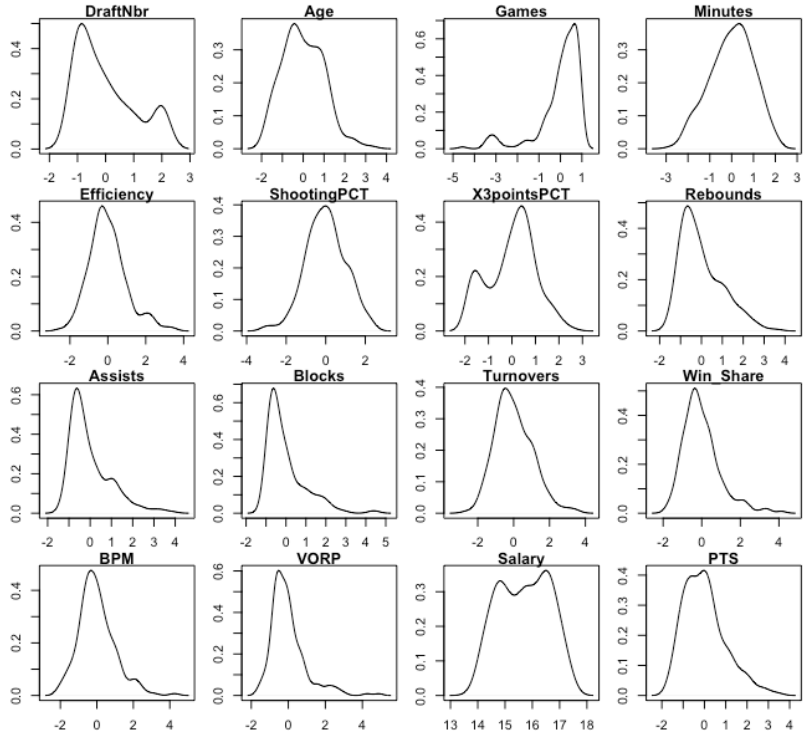


Figure 12: correlation between all variables