

Content Page

1.	Introduction	3
2.	Data Description	3
3	Description and Cleaning of Dataset	4
3.1	Summary statistics for the main variable of interest, salary	4
3.2	Summary statistics for other variables	5
3.2.1	Average career assists per game, <i>AST</i>	5
3.2.2	Career field goal percentage, <i>FGpct</i>	5
3.2.3	Average career point scored per game, <i>PTS</i>	6
3.2.4	Average career rebound per game, <i>TRB</i>	6
3.2.5	Height of the player, <i>height</i>	6
3.2.6	Draft-pick position of the player, <i>draft_pick</i>	7
3.2.7	Draft-pick year of the player, <i>draft_year</i>	7
3.2.8	Primary position played by the player, <i>position</i>	7
3.3	Final Dataset for Analysis	8
4.	Statistical Analysis	8
4.1	Correlations between <i>log(salary)</i> and other Continuous Variables	8
4.2	Statistical Tests	9
4.2.1	Relation between <i>salary</i> and <i>position</i>	9
4.2.2	Relation between <i>salary</i> and <i>height</i>	10
4.2.3	Relation between <i>salary</i> and <i>draft_pick</i>	11
4.2.4	Relation between <i>salary</i> and <i>draft_year</i>	12
4.2.5	The single most important performance measure that is affecting the salary?	14
4.3	Multiple Linear Regression (Optional)	15
5	Conclusion and Discussion	16
6	Appendix	17
7	References	17

1. Introduction

With millions of fans worldwide, the National Basketball Association (NBA) is now one of the most popular and successful professional sports leagues. Huge profits have led to the superstars in the NBA being paid lavishly: in the 2018-19 season, Steph Curry earned over 37.4 million and LeBron James made over 35.6 million on the court.....

In our project, a dataset containing the salaries of NBA players from the 2016-17 season is used, with other variables such as career statistics, draft picks and physical measurements. Based on this dataset, we seek to answer the following popular questions around the NBA:

1. Is the salary of a player dependent of the position he plays in the game?
2. Does the salary depend on the height of the player?
3. Does the salary depend on the draft-pick and draft-year of the player?
4. Is there a single performance measure that is more important in affecting the salary than the others?
5.

This report will cover the data descriptions and analysis using R language. For each of our research objectives, we performed statistical analysis and drew conclusions in the most appropriate approach, together with explanations and elaborations.

2. Data Description

The dataset, titled “NBA Salaries”, is obtained from the online data science community data.world. The original data consists of 2 csv data frames, titled “players.csv” and “salaries_1985to2018.csv”. The dataset was originally posted on basketball-reference.com, the official database partner of the National Basketball Association (NBA), and is open to the public for study and research.

Before proceeding to data analysis, we first performed a preliminary data cleaning to ensure that:

- Irrelevant columns are eliminated, e.g. “birthplace” and “highschool”;
- Players with fewer than 10 games played are treated as unrepresentative anomalies and excluded;
- Redundant information is cut out, e.g. the word “overall” under “draft_pick” column as we only need the number for analysis;
- Only 2016-17 season’s data are included in our dataset
-

After all the preparation, 513 observations (players) with 11 variables are retained for analysis:

1. Sno: serial number
2. X_id: player identity, abbreviated player name
3. AST: career average assists per game

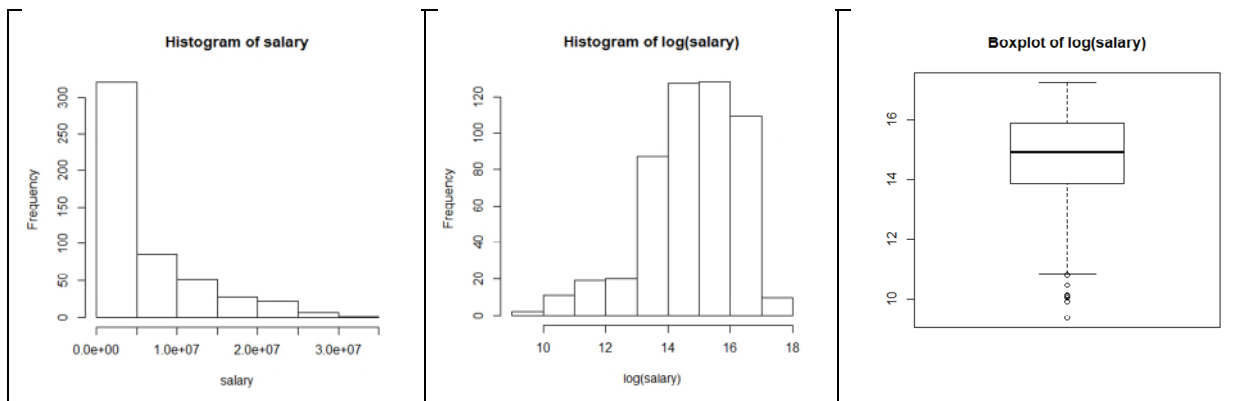
4. FGpct: career field goal percentage
5. PTS: average career point scored per game
6. TRB: average career rebound per game
7. height: height of the player
8. draft_pick: draft-pick position of the player, indicating the perceived value of the player before joining NBA
9. draft_year: draft-pick year of the player, indicating the seniority of the play in NBA
10. position: primary position played by the player (PG, SG, SF, PF or C)
11. salary: total salary in US\$ of the player for the 2016-17 season

3 Description and Cleaning of Dataset

In this section, we shall look into the data in more detail. Each variable is investigated individually to look for possible outliers, and/or to perform a transformation to avoid highly skewed data.

3.1 Summary statistics for the main variable of interest, salary

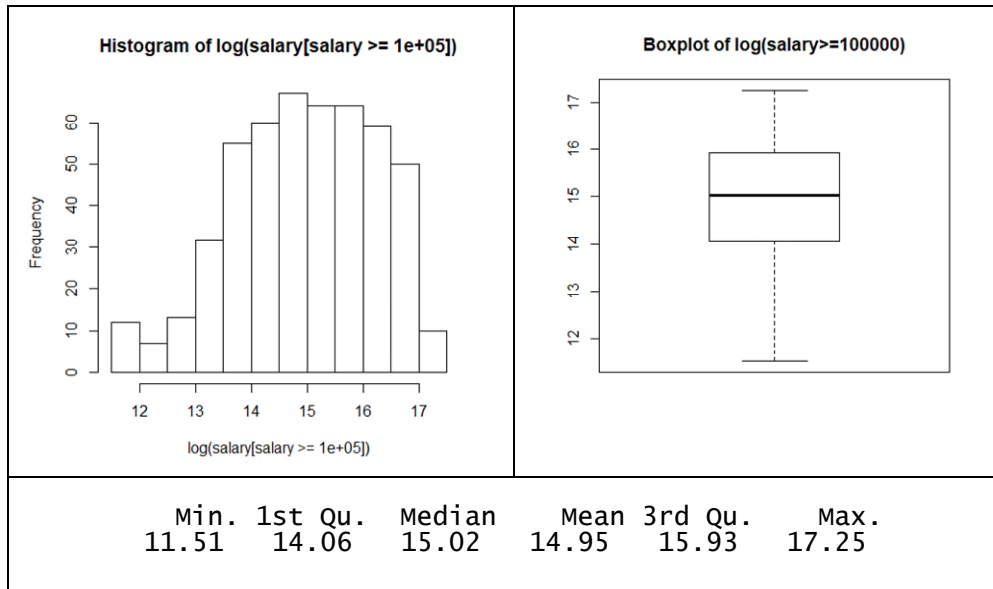
The following plots show the overall distribution of the variable *salary*.



It appears that the variable *salary* is highly skewed, hence we apply a log-transformation (base e) to the variable. The log-transformed data appears to have some outlying values at the left tail. Upon further investigation, we notice that some players were on short-term contracts (a few weeks) during the season. Therefore, we remove those players whose salary are below US\$100,000, approximately 4% of the data.

The histogram and boxplot of the log-transformed variable, with the outliers removed are shown below with summary statistics. The dataset is now more symmetric and does not have any outliers.

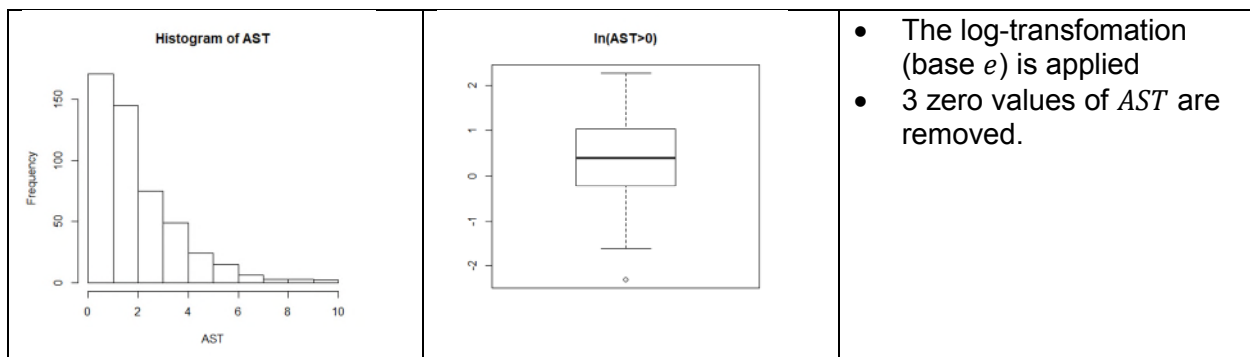
We shall proceed to the next section with this trimmed dataset.



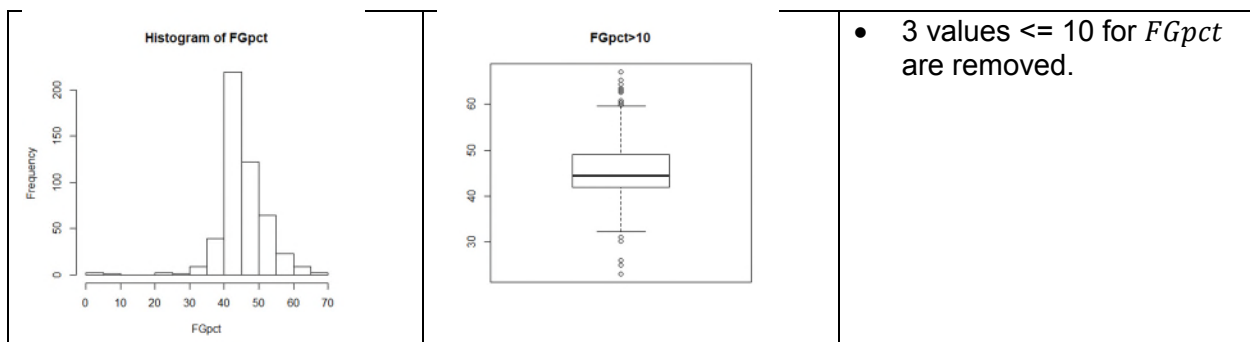
3.2 Summary statistics for other variables

The histogram, the boxplot, the transformation applied and the outliers removed from the variables are tabulated in the following sub-sections.

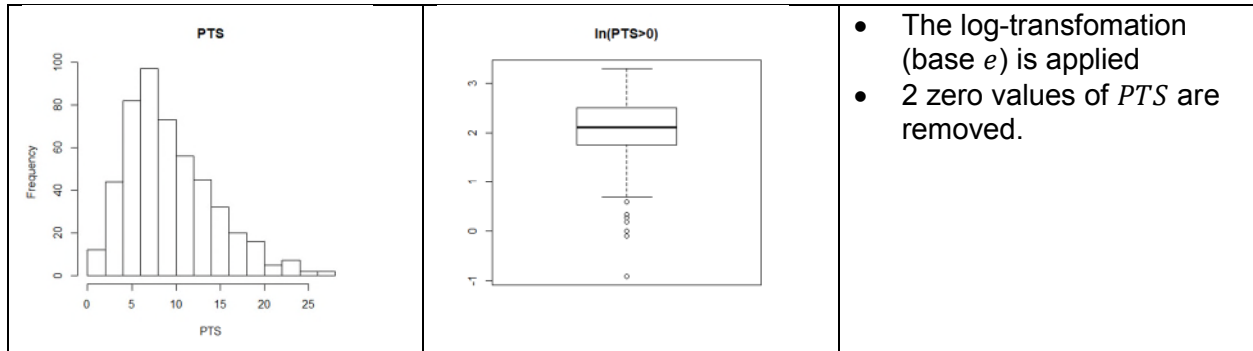
3.2.1 Average career assists per game, *AST*



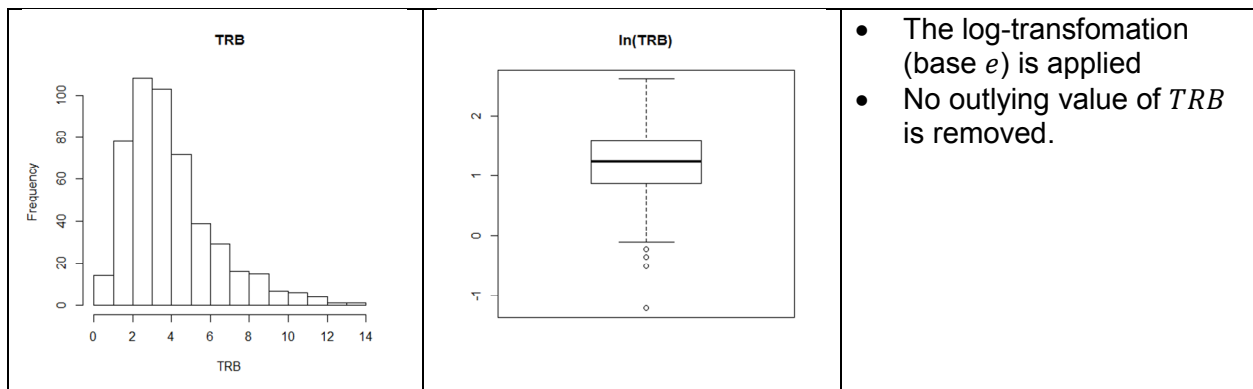
3.2.2 Career field goal percentage, *FGpct*



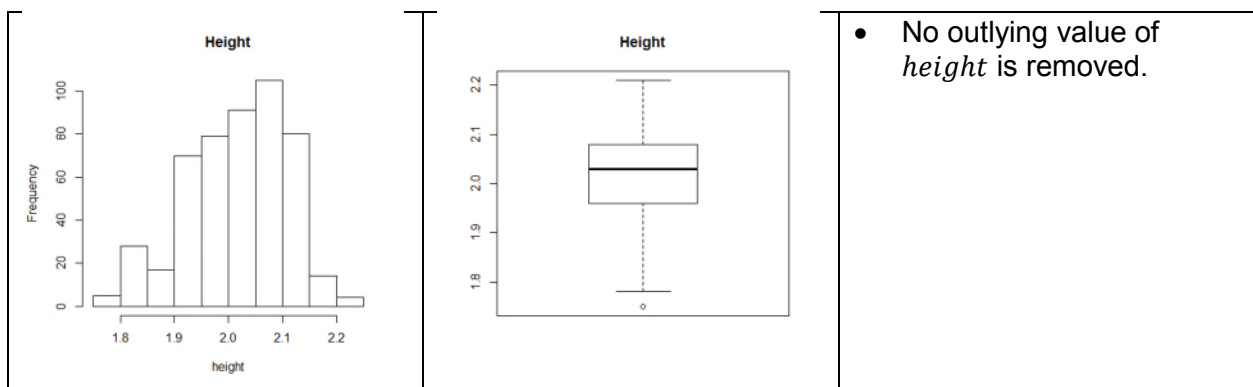
3.2.3 Average career point scored per game, *PTS*



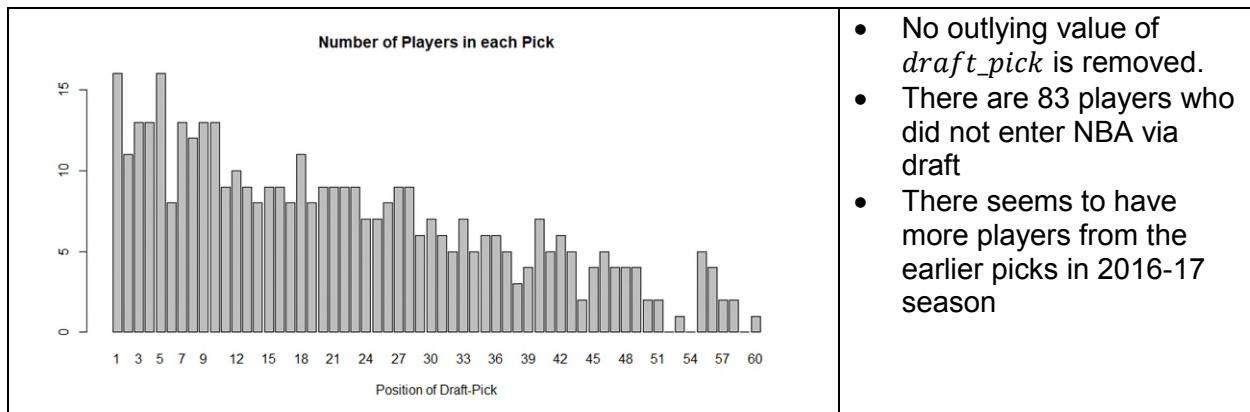
3.2.4 Average career rebound per game, *TRB*



3.2.5 Height of the player, *height*

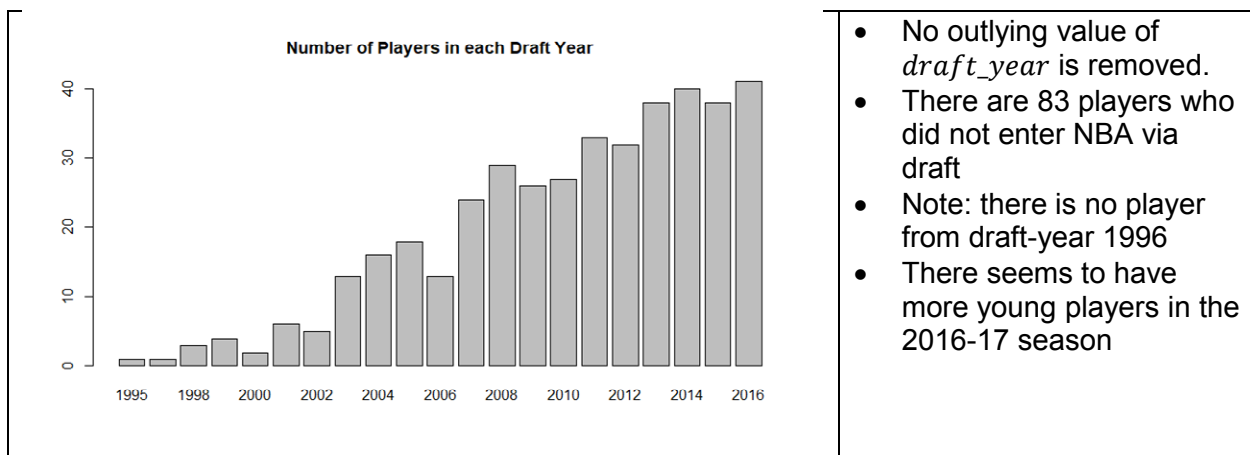


3.2.6 Draft-pick position of the player, *draft_pick*



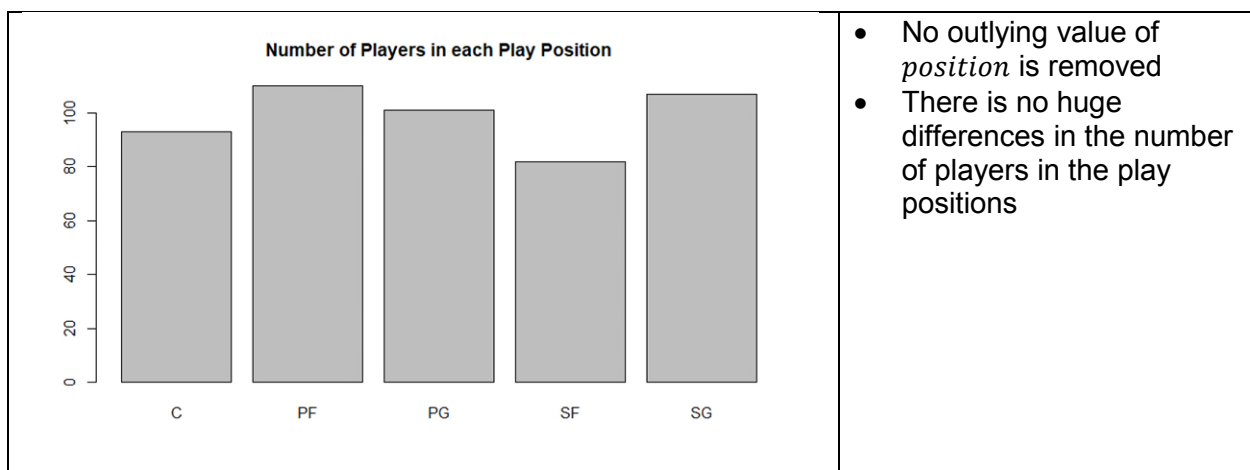
- No outlying value of *draft_pick* is removed.
- There are 83 players who did not enter NBA via draft
- There seems to have more players from the earlier picks in 2016-17 season

3.2.7 Draft-pick year of the player, *draft_year*



- No outlying value of *draft_year* is removed.
- There are 83 players who did not enter NBA via draft
- Note: there is no player from draft-year 1996
- There seems to have more young players in the 2016-17 season

3.2.8 Primary position played by the player, *position*



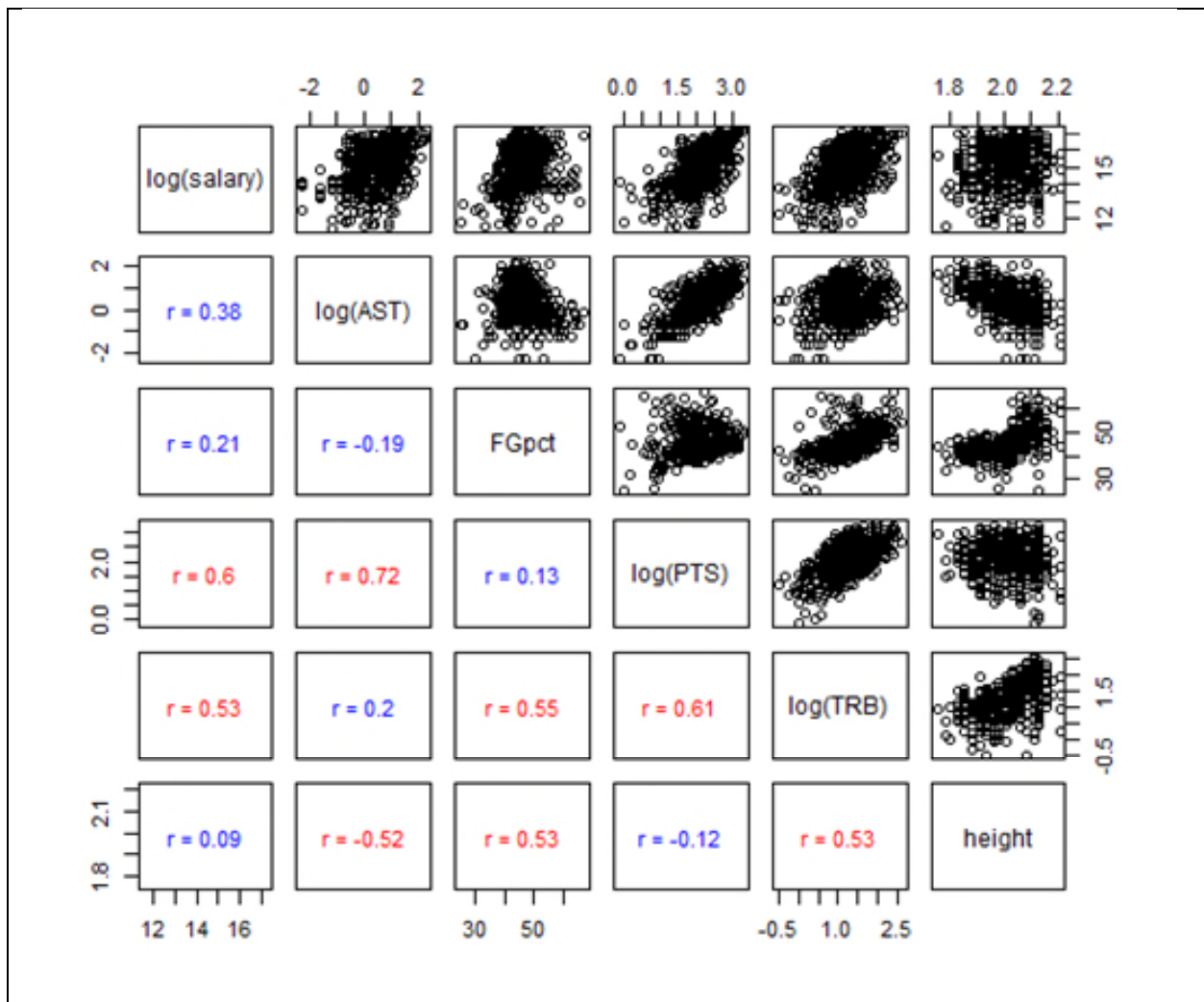
- No outlying value of *position* is removed
- There is no huge differences in the number of players in the play positions

3.3 Final Dataset for Analysis

Based on the above analysis, the dataset is further reduced to 489 observations with the suggested transformations. Namely, log-transformation (base e) to be applied to AST , PTS , TRB , and $salary$.

4. Statistical Analysis

4.1 Correlations between $\log(salary)$ and other Continuous Variables



Scatter plots and correlation coefficients are useful in studying the possible linear relationships between a player's salary and performance indicators.

From the plots, it appears that $\log(salary)$ is more highly correlated to $\log(PTS)$ and $\log(TRB)$ than to other variables.

Among the performance indicators and height, there are a few interesting observations from this tabulation:

- $\log(PTS)$ and $\log(TRB)$ are quite highly correlated ($r = 0.61$)
- $\log(PTS)$ and $\log(AST)$ are positively correlated ($r = 0.72$)
- $FGpct$ and $\log(TRB)$ are positively correlated ($r = 0.55$)
- $height$ is positively correlated $FGpct$ and $\log(TRB)$ ($r = 0.53$), but is negatively correlated to $\log(AST)$

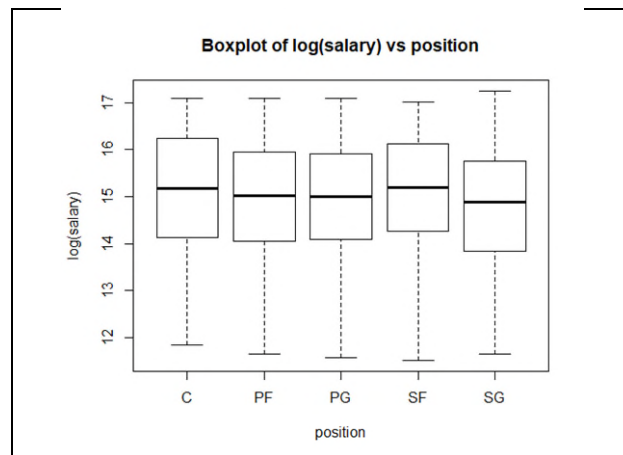
We shall perform some statistical tests to confirm some of our observations in the next section.

4.2 Statistical Tests

4.2.1 Relation between *salary* and *position*

In this section we try to answer the question “Is the salary of a player dependent of the position he plays in the game?”

An analysis of variance (ANOVA) test will be conducted to determine whether $\log(\text{salary})$ is different at each play position, since *position* is a categorical variable. The following plot illustrated the distributions of $\log(\text{salary})$ of the among the play position.



Looking at the boxplot, we see that the spread of $\log(\text{salary})$ are similar for all 5 play positions (factor levels). Hence, the ANOVA test is appropriate for testing the equality of the means (μ_i). We test,

$$H_0: \mu_C = \mu_{PF} = \mu_{PG} = \mu_{SF} = \mu_{SG} \quad \text{against} \quad H_1: \text{not all } \mu_i \text{ are equal}$$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
position	4	8.3	2.074	1.321	0.261
Residuals	484	760.0	1.570		

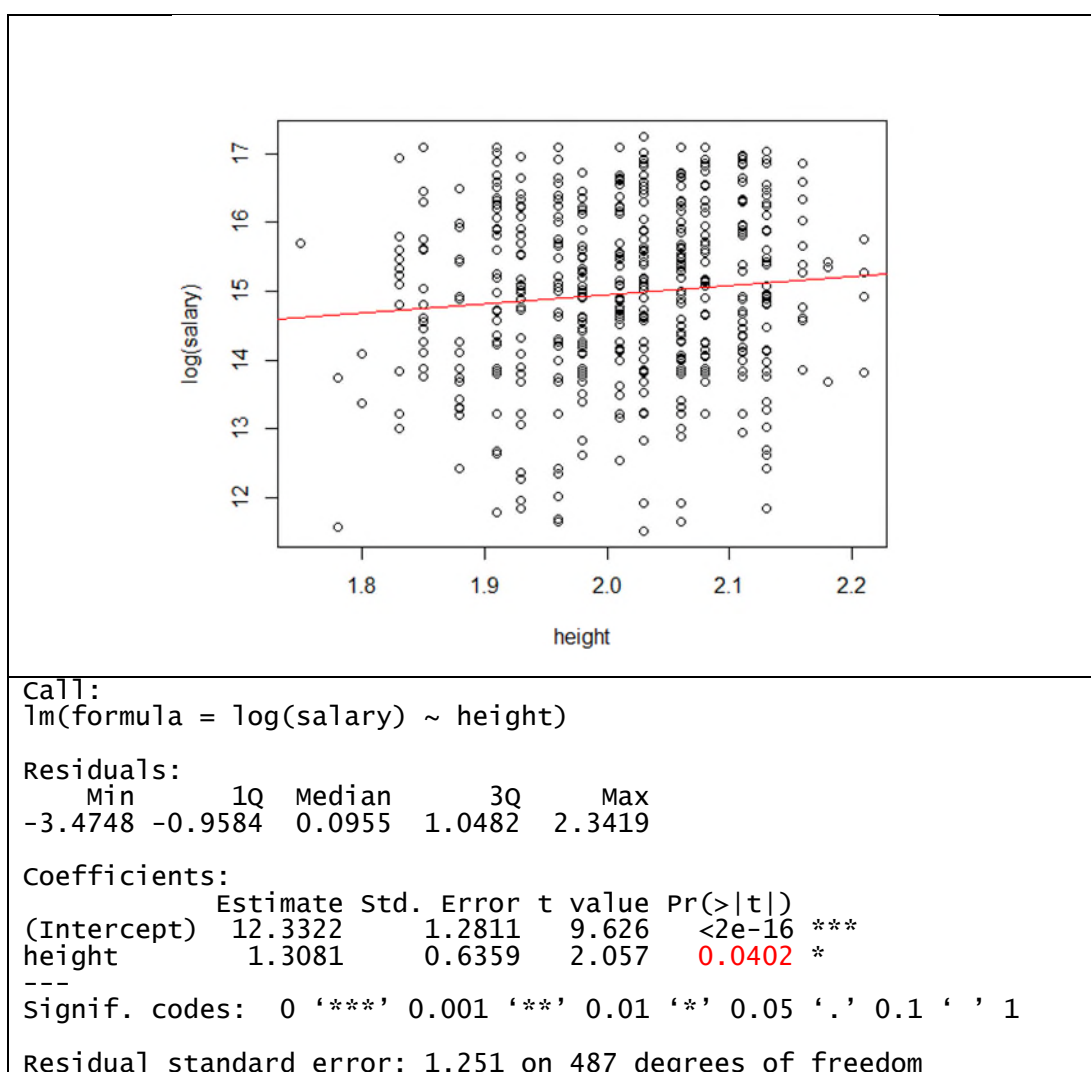
The ANOVA test returns a p-value of 0.261, which shows that the means are not significantly different at a significance level of 0.05. Therefore, we conclude that the salary of NBA player is independent of the position he plays in the game.

4.2.2 Relation between *salary* and *height*

In this section we determine whether the salary of NBA player depends on his height. We perform a simple linear regression between $\log(\text{salary})$ and *height*.

Although the regression model provides a p-value of 0.0402 which indicates a statistically significant relationship between $\log(\text{salary})$ and *height* at 0.05 level of significance, the R-squared for this model is less than 1%, ($R\text{-sq} = 0.0086$), confirming what we have seen in Section 4.1 that the linear correlation between $\log(\text{salary})$ and *height* is only 0.09.

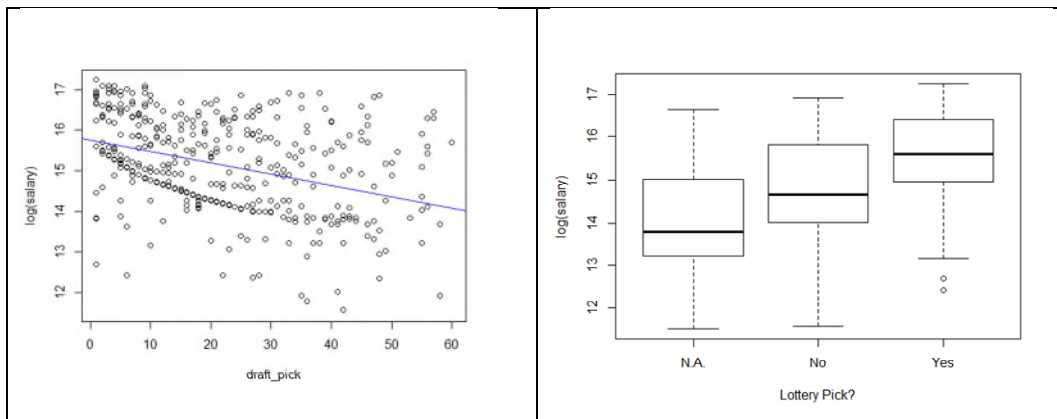
Therefore, we conclude that although the height of NBA player statistically affects his salary, the height only explains less than 1% variation in the $\log(\text{salary})$. It is not practically significant.



Multiple R-squared: 0.008614, Adjusted R-squared: 0.006578
 F-statistic: 4.232 on 1 and 487 DF, p-value: 0.04021

4.2.3 Relation between *salary* and *draft_pick*

Most players enter the league through the annual draft in June. Each year, a total of approximately 60 players from various backgrounds are drafted. Those who display great potential or skill level usually become the “lottery picks”, i.e. top 15 picks.



The scatter plot of $\log(\text{salary})$ versus draft-pick position (above left) shows that there is a downward trend for the salary as the draft position goes large. However, the plot also shows large variation of the salary at each position, and excludes those players who enter NBA not via the draft. In order to understand how draft-pick affects the salary, we regroup the draft-pick into three categories, based on whether the player is in the lottery picks (Yes or No), and the player is not in the draft (N.A.). The boxplot of $\log(\text{salary})$ by the new grouping is illustrated (above right). The boxplot clearly shows a significant different in the distributions of $\log(\text{salary})$ by draft category.

To determine the difference in salary between the lottery and non-lottery picks, excluding the non-draft group, we use a t-test approach to construct a 95% confidence interval (CI). We first determine that the variances of $\log(\text{salary})$ are different in the two groups (Yes and No), p-value approximately 0.0037. Hence, a Welch two-sample t-test is used to constructed a 95% confidence interval for the difference in mean $\log(\text{salary})$, 95% CI: [0.6203 1.0323]. This CI in log-scale translates into a 95% CI for the ratio in geometric means of the player's salary between lottery and non-lottery picks of [1.86, 2.81].

```
> var.test(log(salary[lottery=="Yes"]), log(salary[lottery=="No"]))
      F test to compare two variances

data:  log(salary[lottery == "Yes"]) and log(salary[lottery == "No"])
F = 0.65705, num df = 172, denom df = 233, p-value = 
0.003702
```

```

alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4983400 0.8714897
sample estimates:
ratio of variances
 0.6570531

> t.test(log(salary[lottery=="Yes"]), log(salary[lottery=="No"]), var.equal
=FALSE)

      welch Two Sample t-test

data:  log(salary[lottery == "Yes"]) and log(salary[lottery == "No"])
t = 7.8858, df = 401.54, p-value = 2.984e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.6203142 1.0323051
sample estimates:
mean of x mean of y
 15.63393  14.80762

```

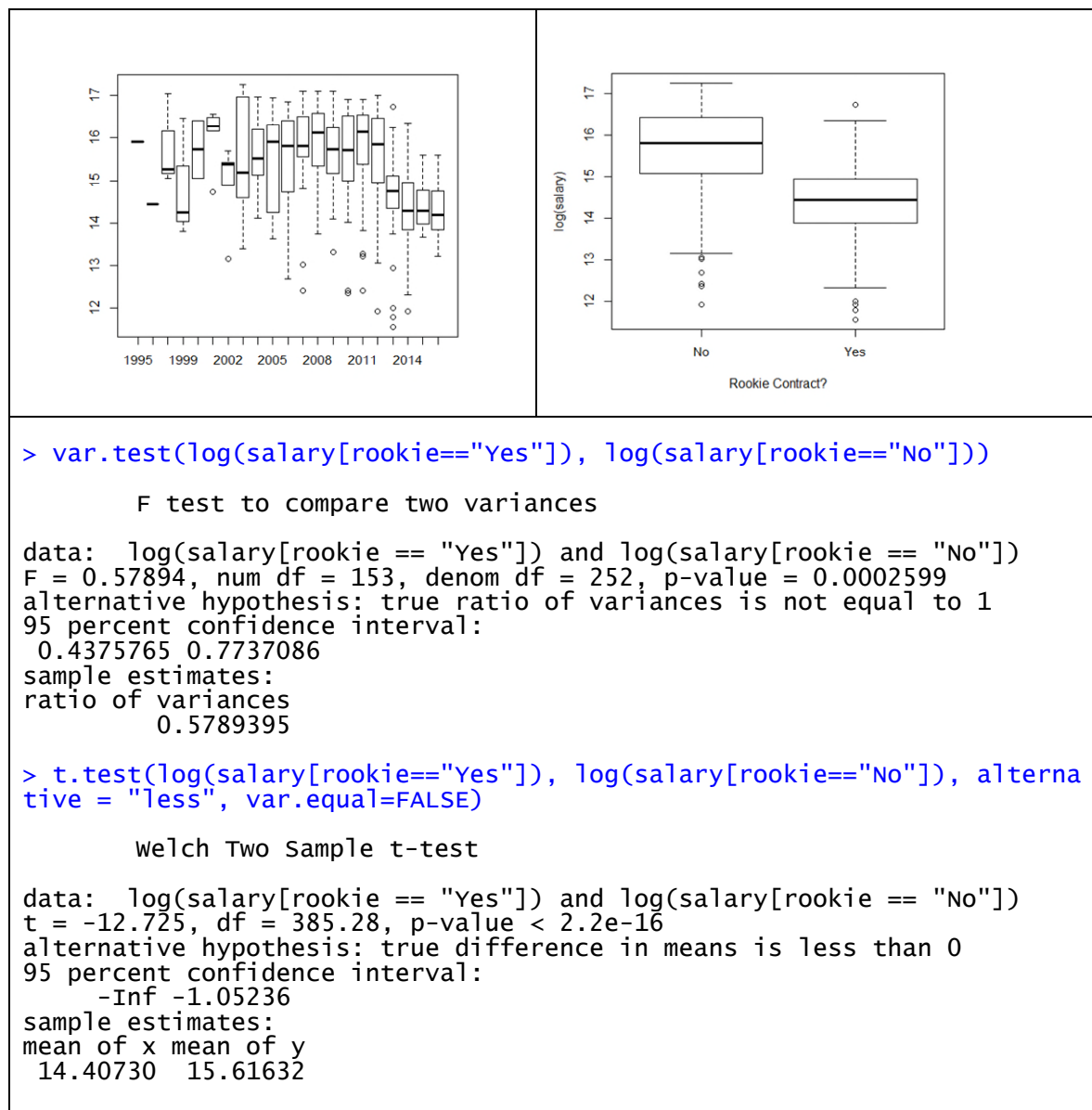
Overall, we conclude that the player's salary depends very on whether the player is in the lottery picks, and whether the player enter NBA via the draft.

4.2.4 Relation between *salary* and *draft_year*

The *draft_year* indicates how longer a player has been in the league, as of 2016. The following boxplot by the draft-year appears to suggest that in the first 4 years of entering the league, players are receiving less salary than their seniors. In fact, when a player enter the league, his pay is determined by a rookie contract that changes year-over-year based on the percentage by which the league raises teams' salary caps. Rookie contracts are guaranteed for the first two years, with teams then having the option to extend the contracts in the third and fourth years as the player's salary increases exponentially each year.

Hence, we are interested in knowing the difference in salary of the player's rookie contract and his subsequent contracts. We re-categorise the draft-year into two category, rookie contract: yes, or No. We have to exclude those who enter the league not via the draft (82 of them), since we do not have the data on the year these players entered the league.

The boxplot of $\log(\text{salary})$ by rookie contract is plotted below. From the diagrams, it is rather straightforward to see that a payroll gap exists between the rookie and non-rookie contracts. We conducted further statistical tests to support this observation.



Variance test

H_0 : Variances of rookie and non-rookie contracts' $\log(\text{salary})$ are equal;

H_1 : Variances of rookie and non-rookie contracts' $\log(\text{salary})$ are not equal;

At a significance level of 0.05, we reject the null hypothesis and conclude that the variances of the two samples are not equal, since $p\text{-value} = 0.00026 < 0.05$.

T-test

H_0 : The mean of $\log(\text{salary})$ under rookie contracts is equal to that under non-rookie contracts;

H_0 : The mean of $\log(\text{salary})$ under rookie contracts is less than that under non-rookie contracts;

Using a sided t-test with unequal variances, the p-value is less than $2.2e-16$. Since $p\text{-value} < 0.05$, we reject the null hypothesis at a significance level of 0.05, and conclude that the mean of $\log(\text{salary})$ under rookie contracts is significantly less than that under non-rookie contracts.

According to the results shown above, we conclude that the salaries of the lottery pick players are significantly higher than those of the non-lottery pick players.

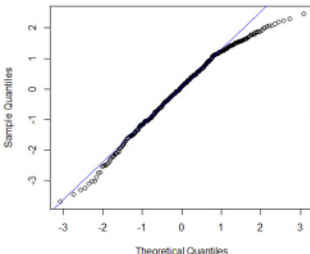
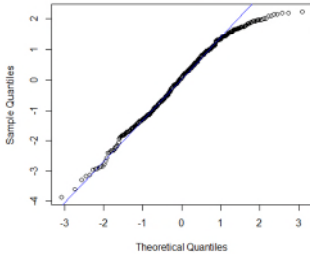
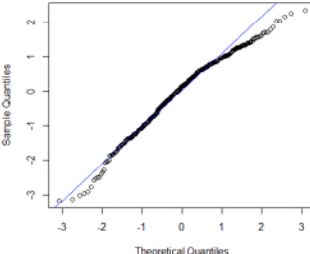
4.2.5 The single most important performance measure that is affecting the salary?

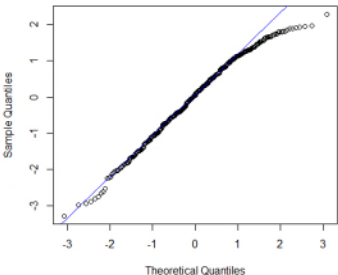
In Section 4.1, we have already seen that the performance measure are quite strong correlated to $\log(\text{salary})$. We now perform a simple linear regression analysis to determine which of the 4 performance measures could be used model $\log(\text{salary})$ in a linear fashion.

$$\log(\text{salary}) = \beta_0 + \beta_1 * X + \varepsilon$$

where X could be any one of $\log(\text{AST})$, FGpct , $\log(\text{PTS})$ or $\log(\text{TRB})$. The summary of the analysis is listed in the table below.

By comparing the R-squared and the residual plot, $\log(\text{PTS})$ is determined to be the single most important performance measure to model the $\log(\text{salary})$ using a simple linear model.

Variable (X)	Fitted Model, with Y being $\log(\text{salary})$	p-value	R-squared	qq-plot of residuals
$\log(\text{AST})$	$\hat{Y} = 14.74 + 0.578X$	<2e-16	0.1436	
FGpct	$\hat{Y} = 12.96 + 0.044X$	4.23e-06	0.0426	
$\log(\text{PTS})$	$\hat{Y} = 12.20 + 1.320X$	<2e-16	0.3654	

$\log(TRB)$	$\hat{Y} = 13.51 + 1.177X$	$<2e-16$	0.2757	
-------------	----------------------------	----------	--------	---

4.3 Multiple Linear Regression (Optional)

In this Section, we attempt to build a multiple linear model for $\log(\text{salary})$ based on the 4 given performance measures, namely $\log(AST)$, $FGpct$, $\log(PTS)$ and $\log(TRB)$. We use a backward elimination method to select the most appropriate model. The result is shown in the R output below.

We conclude that $\log(PTS)$ and $\log(TRB)$ are the significant measures that could be used to modelled $\log(\text{salary})$, whilst $\log(AST)$ and $FGpct$ are not. The fitted model is:

$$\log(\widehat{\text{salary}}) = 12.20 + 0.989 * \log(PTS) + 0.561 * \log(TRB)$$

```
> model15 = lm(log(salary) ~ log(AST) + FGpct + log(PTS) + log(TRB))
> step(model15, direction="backward")
Start: AIC=-23.1
log(salary) ~ log(AST) + FGpct + log(PTS) + log(TRB)

      Df Sum of Sq  RSS   AIC
- log(AST)  1    0.072 457.07 -25.0252
- FGpct     1    0.197 457.19 -24.8911
<none>                    456.99 -23.1017
- log(TRB)  1   16.778 473.77  -7.4704
- log(PTS)  1   40.056 497.05  15.9848

Step: AIC=-25.03
log(salary) ~ FGpct + log(PTS) + log(TRB)

      Df Sum of Sq  RSS   AIC
- FGpct  1    0.158 457.22 -26.856
<none>                    457.07 -25.025
- log(TRB)  1   17.267 474.33  -8.892
- log(PTS)  1   91.804 548.87  62.479

Step: AIC=-26.86
log(salary) ~ log(PTS) + log(TRB)

      Df Sum of Sq  RSS   AIC
<none>                    457.22 -26.856
- log(TRB)  1   30.304 487.53   2.525
- log(PTS)  1   99.265 556.49  67.219

Call:
lm(formula = log(salary) ~ log(PTS) + log(TRB))
```

Coefficients:		
(Intercept)	$\log(\text{PTS})$	$\log(\text{TRB})$
12.2032	0.9885	0.5606

5 Conclusion and Discussion

The National Basketball Association (NBA) has a reputation for being the most innovative of the major professional North American sports leagues, earning money from a combination of television rights, merchandising, ticket sales, and more. In order to sustain its business model, teams will have to attract the most talented players based on their performance, and reward them with attractive salary. In this report, attempt to answer some of the basic questions related to player salary based on 2016-2017 season data, and with very limited number of performance measures.

We conclude that:

- the distribution of salary is independent of the position a player plays
- the height of a player does not affects the salary that he receives
- the geometric mean salary depends on whether the player is in the lottery picks
- among the 4 performance measures, the average number of points scored in a game is the most important single measures that will affect the salary.
-

Additionally, we see that average number of points scored and average number of rebounds in a game, can be used to model the salary via a linear model. The measures on average number of assists and field goal percentage do not seem to have strong correlation with the player salary.

Although the results of this report is interesting, it must be noted that this report is only based on one single season of data published on the internet. Furthermore, with the advancement of data capturing techniques, the NBA has been able to generate more sophisticated performance indexes than what we have considered. Deeper and wider analysis of the NBA data, with advance analytical techniques would be needed to make a stronger statement about the relationship between player's performance and his salary.

6 Appendix

Listing of code and output from R.

7 References

If there is any.