Report of MA678 Midterm Project

Franky Zhang

11/29/2021

Abstract

The National Basketball Association (NBA) is a professional basketball league in North America, which is absolutely the premier men's professional basketball league in the world. However, in 2020-2021 season, the highest paid player is Golden States top star, Stephen Curry, who earned a salary over $43,000,000 while the lowest paid player's salary is around $150,000. Thus, here comes the problem: which kind of players is favored most in NBA currently or which kind of players can sign a big NBA contract.

Introduction

Usually, boxed score is important description of performance of players(e.g. points, rebounds, assists and so on) and whether a player can sign a big contract is greatly decided by it. Nevertheless, different stages of NBA development seem to have different "most important" positions. For example, during a relatively long period, a truly dominant center could simply dominate the game, like Kareem Abdul-Jabbar and Wilt Chamberlain who could score almost ever time he touched the ball. Then, as Michael Jordan dominated 1990s and won 6 NBA championships, he undoubtedly led a wave of shooting guard. Recently, Warrior's Death lineup was too dominant that nearly the whole league was reorienting around to compete with Hampton Five. Thus, the combo guards with great offensive ability (e.g. Stephen Curry, James Harden, Russell Westbrook) or small forwards(e.g. Lebron James, Kevin Durant and Kawhi Leonard) are absolutely most favored in NBA and have the biggest chance to earn a fairly high salary. On the other hand, the luxury tax threshold is $136,606 million and meanwhile, some teams are willing to take risk of paying high tax luxury to form a better lineup and chase championship while some teams not. It's natural to come to the conclusion that players in big franchise teams are more likely to receive higher salaries. That to say, team is another factor we should take into account.
Therefore, I decide to introduce multilevel models to find out the influences of fixed effects (e.g. points, rebounds, assists and so on) and random effects (teams, positions).

Methods

Data Preprocessing

I found the data set from a public github repository(https://github.com/MattC137/Open_Data/tree/master/Data/Sports/NBA). Firstly, I download 2017-2020 box score and 2021 players information because the players' 2021 salaries are greatly depended on their performances in recent years. Hence, I need to combine them and create the appropriate data frame. Additionally, as the box data is by individual games, after combining the data, I calculate average game statistics(non-Playoff games) for each player and transform `played` information to binary factor (0 or 1) then calculate average `appearances` for each player in non-Playoff games. Additionally, to avoid the bias that some players are labelled with f (forward) and g (guard) while some are labelled more specific with SF, PF, SG, PG, I combined the levels and used foward and guard identically. For the next step, I wiped out 'NA' data in players' information and merge it with aforementioned data frame. Up till now, I get full information of 314 players,which means 11 players for each team on average. Here is the glossary of terms:

column names	explanation
Player_Id_Str	Structured id name of player
Team	Team of player
Position	Position of player(SF, PF, C, SG, PG)
Position1	Position of player(forward, center, guard)
Salary	Annual salary of player (dollars)
Minutes	Average playing time of player (minutes)
Appearance	Average appearance of player (times)
Points	Average points of player
FG_Made	Average field goal made by player
FG_per	Average field goal percentage of player
Threes_Made	Average 3-point field goal made by player
Threes_pre	Average 3-point field goal percentage of player
FT_Made	Average free throws made by player
FT_pre	Average free throws percentage of player
Rebounds	Average rebounds of player
Assists	Average assists of player
Steals	Average steals of player
blocks	Average blocks of player
Turnovers	Average turnovers of player
Fouls	Average fouls of player
Height	Height of player (foot)
weight	Weight of player (lb)
Draft_Pick	Draft pick of player

Exploratory Data Analysis

By aforementioned part, I've got a `NBA_data` with 314 observations and 21 variables, among which there is 1 output `salary` and 20 predictors. However, whether or not to use all of these 20 predictors is depended on following analysis.

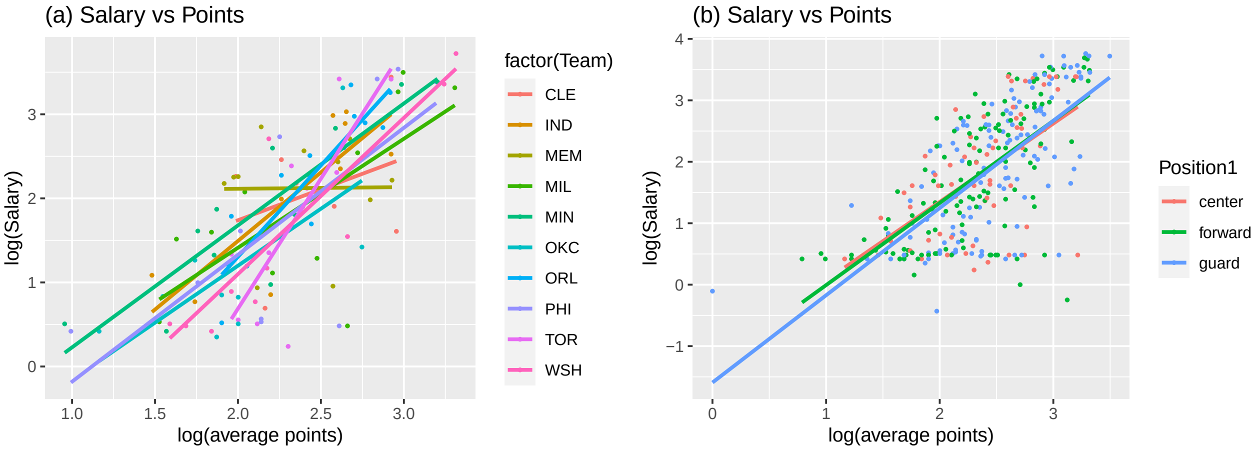

Figure 1: relationship between salaries and points of players

Figure 1 illustrates the relationship between salaries and average points, while fig(a) is in team level and fig(b) is in position level. However, whatever the level, salaries show the increasing trend as points going up. And in different teams and positions, the intercepts and slopes show slights differences. After I draw the graph of salaries versus appearance, rebounds, assists, steals and blocks, the figures are quite similar. Thus I put them in the appendix.

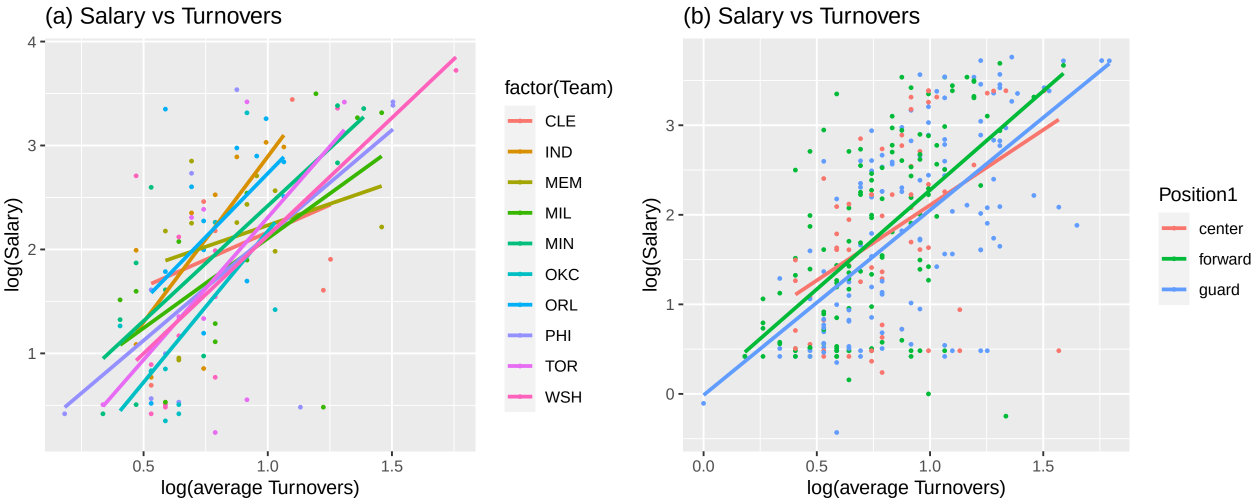

Figure 2: relationship between salaries and turnovers of players.

Figure 2 shows the correlation between players' salaries and turnovers. Similarly, figure(a) is in team level while figure(b) is in postion level. The results is weird because turnover is absolutely a negative statistic on basketball court and no one would sign huge contract with players making over 10 turnovers per game! However, as only those most high-maintenance guards have the chance to make high turnover, merely look at number of turnovers is misleading. Thus, I decided to look into relationship between turnovers and assists.

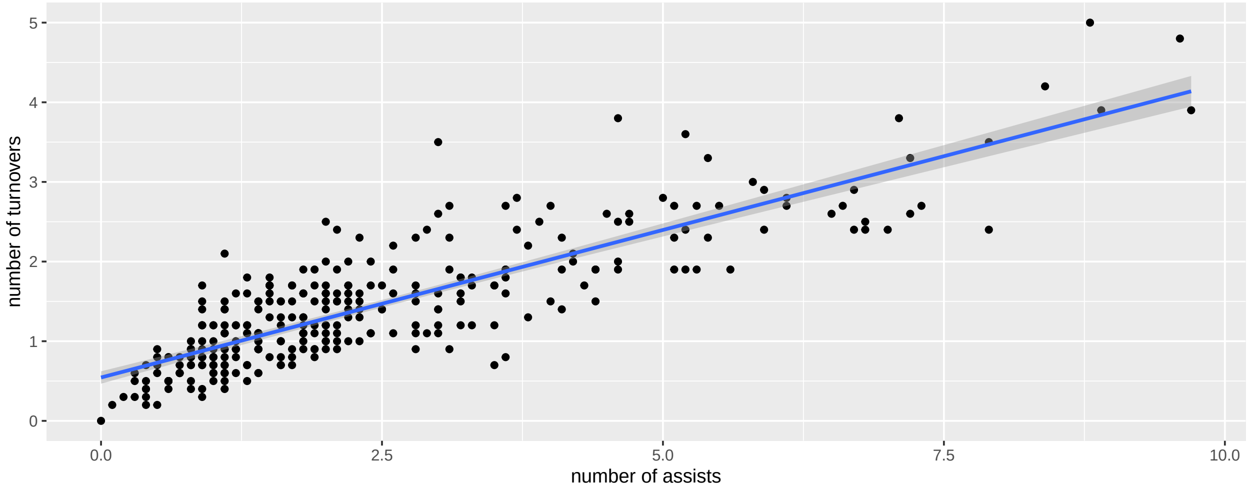

Figure 3: relationship between turnovers and assists of players

Figure 3 verifies that players' turnovers are closely related with their assists while the later is obviously positive stats. What's more, the Pearson covariance between turnovers and assists is over .85! Thus, I decide to exclude variable `Turnovers`. Variable `Fouls` is similar situation that it is highly correlated with `Minutes`, which shows Pearson covariance stats over .75. And `Heights` and `Weights` is another pair of high correlated variables. Hence, I only kept variable `Heights`.

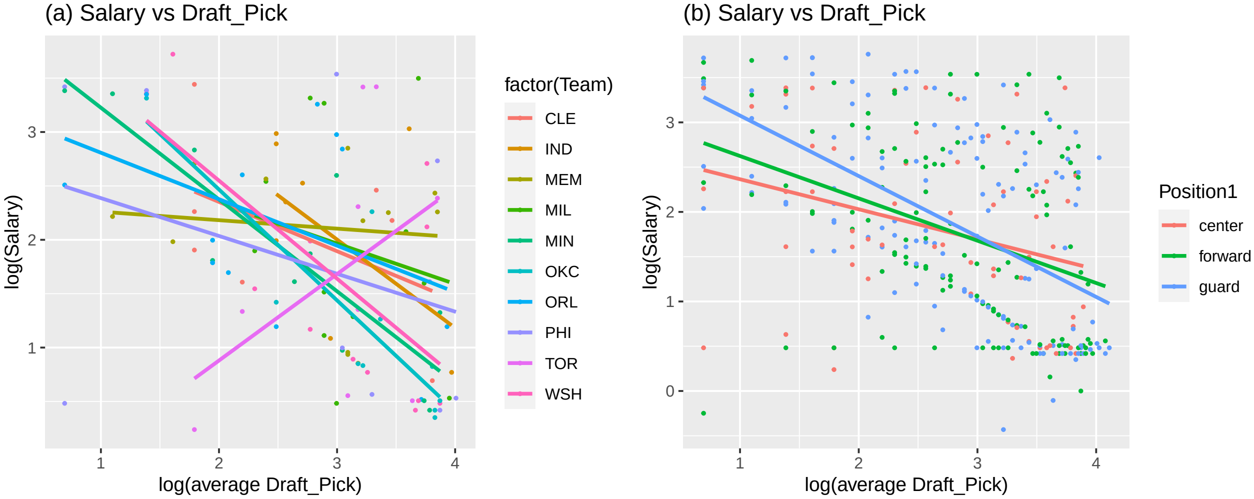

Figure 4: relationship between salaries and draft picks of players

Figure 4 shows that whatever from team or position level, high draft pick players are more favored and usually can get relatively high salary. It is consistent with our intuition that draft pick results reflect the potential of players to some extent. Yet, the slopes and intercepts vary from team to team and that indicate random effects really matter in this case.

Model fitting

Since different teams and positions have quite large impacts on the model, I decide to use multilevel model to fit `NBA_data`. As to selection of variables, I also include `three made per game` because under the crazy trendency of small ball, the ability to shoot 3-points is really important. Meanwhile, since all variables are more or less skewed and have heavy tails, I took `log(variable + 1)` to create new ones. All original distribution plots of variables can be found in Appendix of this report. For the next step, I draw the Pearson correlation matrix to do the predictor selection.

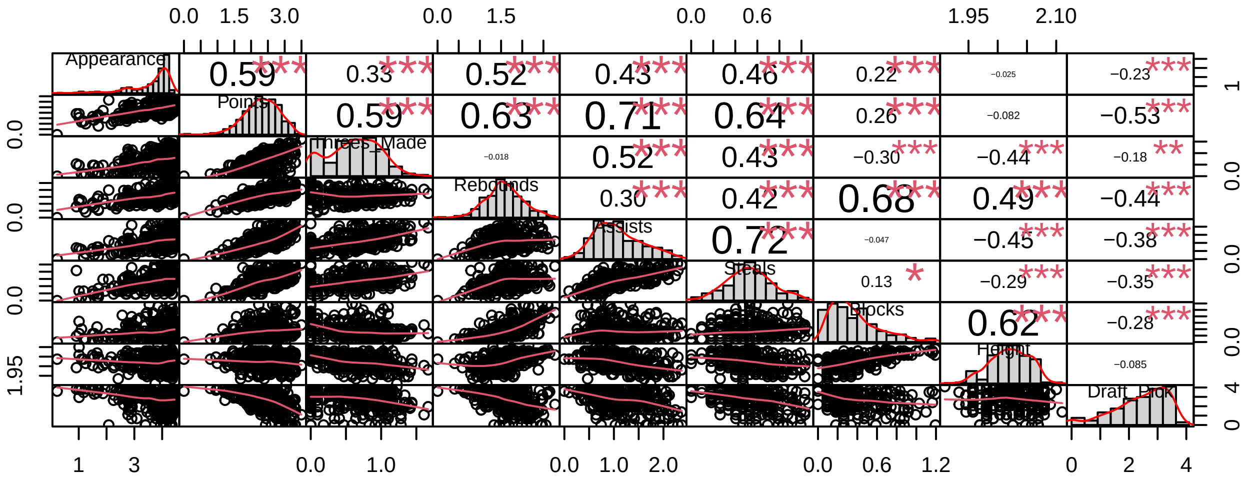

Figure 5: Correlation Matrix

Figure 5 clearly shows Pearson relationship between variables and I use .6 to be a threshold, which indicates variables can be highly related. Since `Points` and `Assists` are two of the most important predictors, we have to keep them anyway. Then, the highest correlation appears between `Assists` & `Steals` and `Blocks` & `Rebounds` as well as `Blocks` & `Height`. Hence, I decided to drop `Rebounds`, `Steals` and `Blocks`.

Additionally, as different teams have quite different on-court strategies and seasonal goals, random effect of teams is quite important for variables: `Appearance, Points, Three_Made and Draft Pick`. On the other hand, every position takes its jobs, like point guard is responsible for handling the ball and triggering team's offense, the preferred `Height` are vary from different positions. Thus, I prefer changing slopes and intercepts of `Height` for different ones. Here is the function I built:

```
model <- lmer(Salary ~ Appearance + Points  + Threes_Made
              + Assists  + Height + Draft_Pick
              + (1 + Appearance + Points + Threes_Made + Assists + Draft_Pick| Team)
              + (1 + Height| Position1),
              data = log_NBA_data)
```

Here is the summary of model(fixed effect) and all variables here are considered as statistically significant at $\alpha = 0.5$ level.

```
##              Estimate Std. Error     df t value Pr(>|t|)
## (Intercept)    -8.95       2.61 170.88   -3.43     0.00
## Appearance      0.26       0.06  92.88    4.61     0.00
## Points          0.38       0.14  42.68    2.63     0.01
## Threes_Made     0.30       0.13  72.08    2.34     0.02
## Assists         0.53       0.13  56.82    4.13     0.00
## Height          4.27       1.29 177.11    3.31     0.00
## Draft_Pick     -0.11       0.05  28.37   -2.21     0.04
```

And the following tables are the summary of random effects. The first one is random effect of Team (only display first ten teams alphabetically) and the second one is Positions.

```
##     (Intercept) Appearance Points Threes_Made Assists Draft_Pick
## ATL       -0.25      -0.02   0.23       -0.15   -0.18       0.00
## BKN       -1.51      -0.12   0.44       -0.05    0.08       0.26
## BOS        0.61       0.05  -0.21        0.04    0.00      -0.10
## CHA        0.33       0.03   0.01       -0.09   -0.15      -0.08
## CHI       -0.18      -0.01   0.04        0.00    0.02       0.03
## CLE        0.46       0.04  -0.13        0.01   -0.04      -0.08
## DAL       -0.46      -0.04   0.21       -0.08   -0.07       0.06
## DEN        0.60       0.05  -0.18        0.03   -0.03      -0.10
## DET       -0.02       0.00  -0.11        0.10    0.14       0.03
## GS         0.25       0.02  -0.17        0.10    0.11      -0.02
```

```
##         (Intercept) Height
## center         0.00   0.00
## forward        0.02  -0.01
## guard         -0.02   0.01
```

Result

Interpretation

Let's take Brooklyn Nets Guard for example. Firstly, we are able to get the following formula of fixed effect:

$$log(salary{+}1) = -8.95{+}0.26{\times}log(Appearance{+}1){+}0.38{\times}log(Points{+}1){+}0.30{\times}log(ThreesMade{+}1){+}0.53{\times}log(Assists{+}1)$$

Then add the random effect of Brooklyn Nets Guard's random effect to the intercepts and slopes and get the estimated formula:

$$log(salary{+}1) = -10.94{+}0.16{\times}log(Appearance{+}1){+}0.78{\times}log(Points{+}1){+}0.29{\times}log(ThreesMade{+}1){+}0.57{\times}log(Assists{+}$$

In the formula, all the parameters are positive except `Draft_Pick`, which indicates they all hvae positve impact on players' salaries. And then look at `Draft_Pick`, it's known to all that the smaller the draft ranking, the larger potential the player being estimated. Thus it's reasonable that the parameter of `Draft_Pick` is negative. Go back to the model, for every increase 1% of player's points, the prediction of salary is 0.78% for Brooklyn Nets' guard line. Of course, as different teams have different on-court characteristics, the favors of players are different, which means all the parameters are changing from team to team.

On top of this, what we can find is that the intercept of Brooklyn Nets players' salaries is far lower then nearly all other teams (-10.94), and this is really understandable because the style of this team is a typical `Big 3` mode, which means over 85% salaries are paid for three super scorers (Kevin Durant, James Harden and Kyrie Irving) and salaries of other rotations are relatively low. Meanwhile, this exactly answers the question why the `Points` parameter of Brooklyn Nets reaches scaring 0.78, far higher then other teams.

What's more, let's take Brooklyn nets all-star guards James Harden for example and input his data into the formula. Then under this model, the prediction salary is 34.46 million dollars and the actual value is 41.25 million.

Model Checking

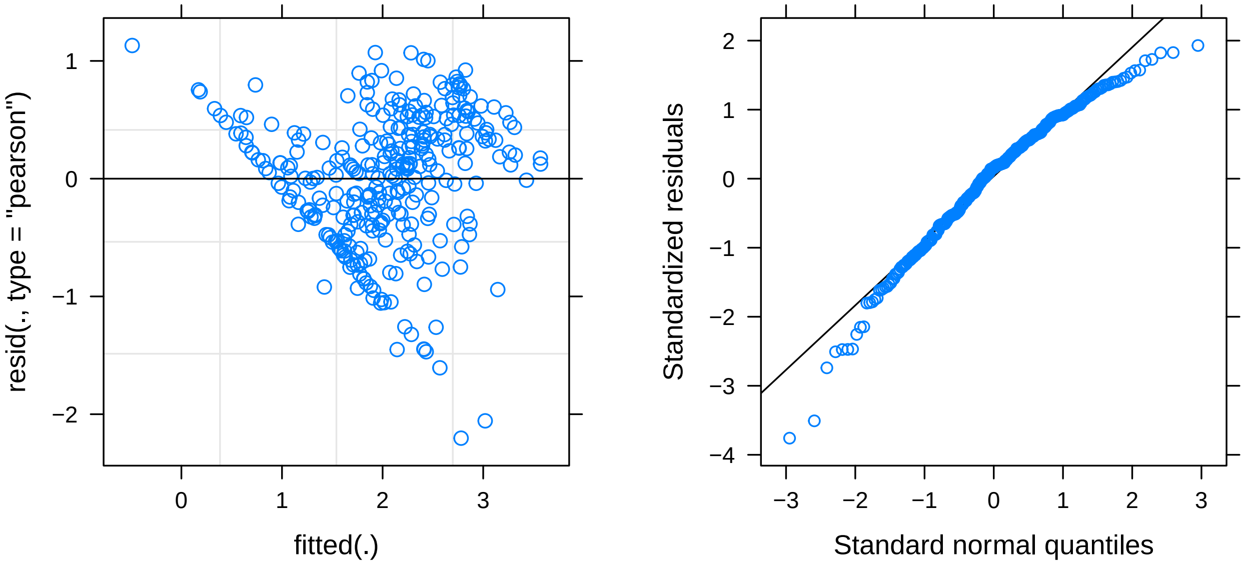

Figure 6: Residual plot and Q-Q plot.

The left plot of Figure 6 is residual plot and the right one is residual Q-Q plot. According to it, the mean value of residuals is approximately 0. Yet as the fitted value close to 0, there's no negative residuals. This phenomenon can be explained by the model we use because we transformed all variables by `log(variable + 1)`. As for Q-Q plot in Figure 6, majority points except tail ones are on the normal distribution line, thus the normality check is acceptable.

Discussion

In this report, multilevel model is used to figure out the relationship between players' salaries and their several basic on-court stats as well as their draft rankings. Also, this model take two kinds of group level into consideration: players' teams and players' position. Generally, from the perspective of fixed effects, all the predictors have positive impacts on players' salaries. For instance, during 2017-2020 seasons, the higher average points (assists, draft ranking and so on) the player scored the higher the estimation of his annual salary by this model. In addition, both in team and position level, the random effects sound reasonable, which means the results can be explained by the characteristics of teams or positions. Finally, several model checks are also good and support the validity of the model.

This report also has some limitations. Firstly, the data set I built needed to be considered more carefully. Although I have considered about the lagging of players' salaries, like the contract the player signed were depended on the performance of years before that, the contract years of players are actually different. Thus the selection of same performance period (2017-2020) might not fair for all players. Additionally, I have already managed to reduce the effects of correlations between predictors by dropping out highly related ones but there still exists over 0.5 Pearson correlation pairs. For the further work, I plan to involve `PCA` to handle high correlated predictors first and then fit multilevel model.

On top of this, our model is under the assumption that the players' salaries is mostly decided by whether their talents are suitable for current NBA basketball style, whatever name them as `samll ball` or `fast paced`. And it's known to all that the luxury cap of NBA has dramatically increased during the past years, which means average income of players has inflated. So it's might not be fair to compare the current income of player since for some their contracts were signed several years ago and some are newly signed and that might not be consistent with players' abilities.

Reference

[1] Micheal, C. *Mixed Models with R.* https://m-clark.github.io/mixed-models-with-R/

[2] Roy, K. (2021, July 1). *What is the most important position in basketball?.* https://simpleblitz.com/nba/important-position-basketball/

Appendix

Variable distributions

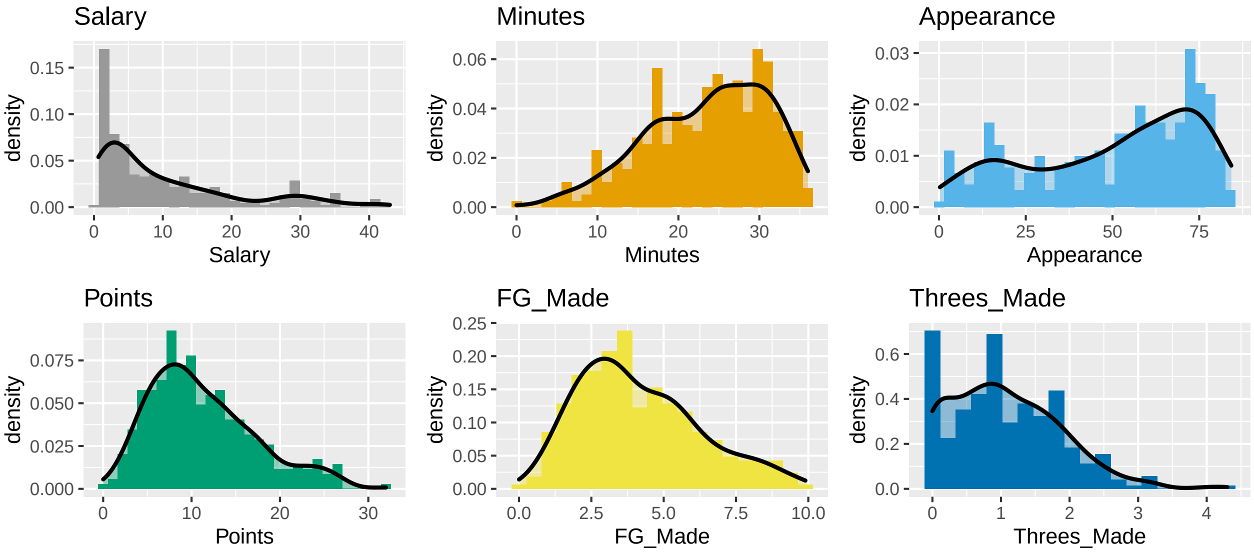

Figure 7: EDA: distribution plots (1)

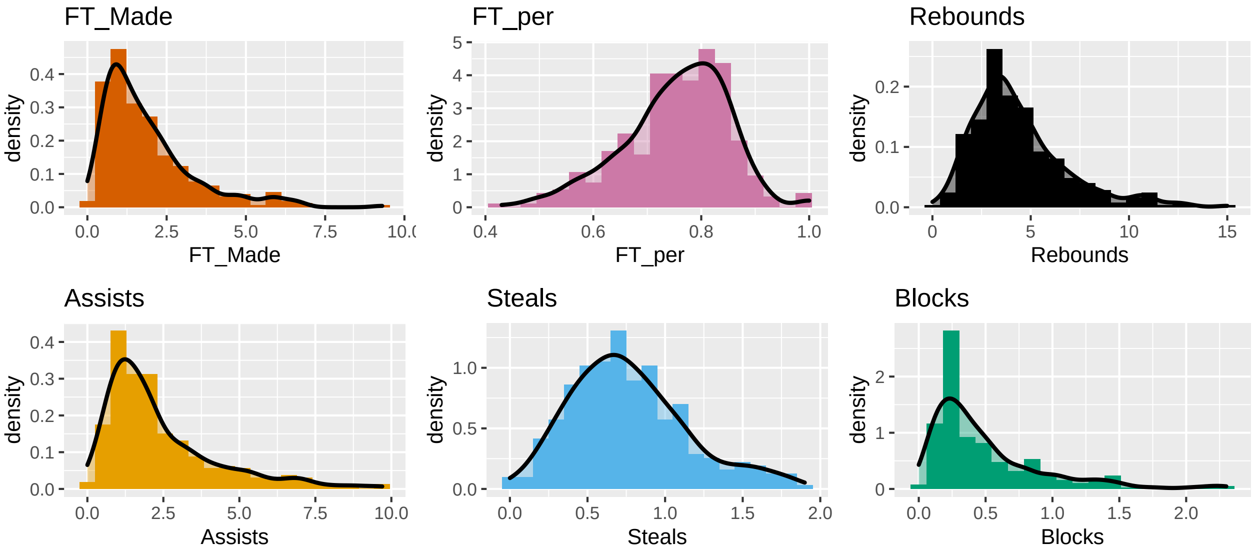

Figure 8: EDA: distribution plots (2)

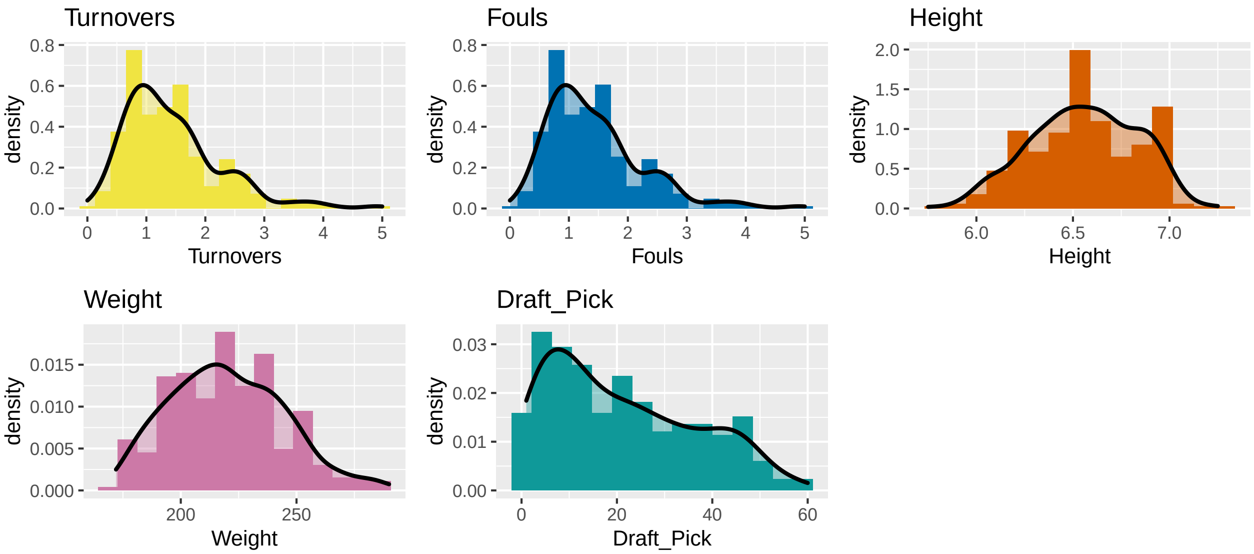

Figure 9: EDA: distribution plots (3)

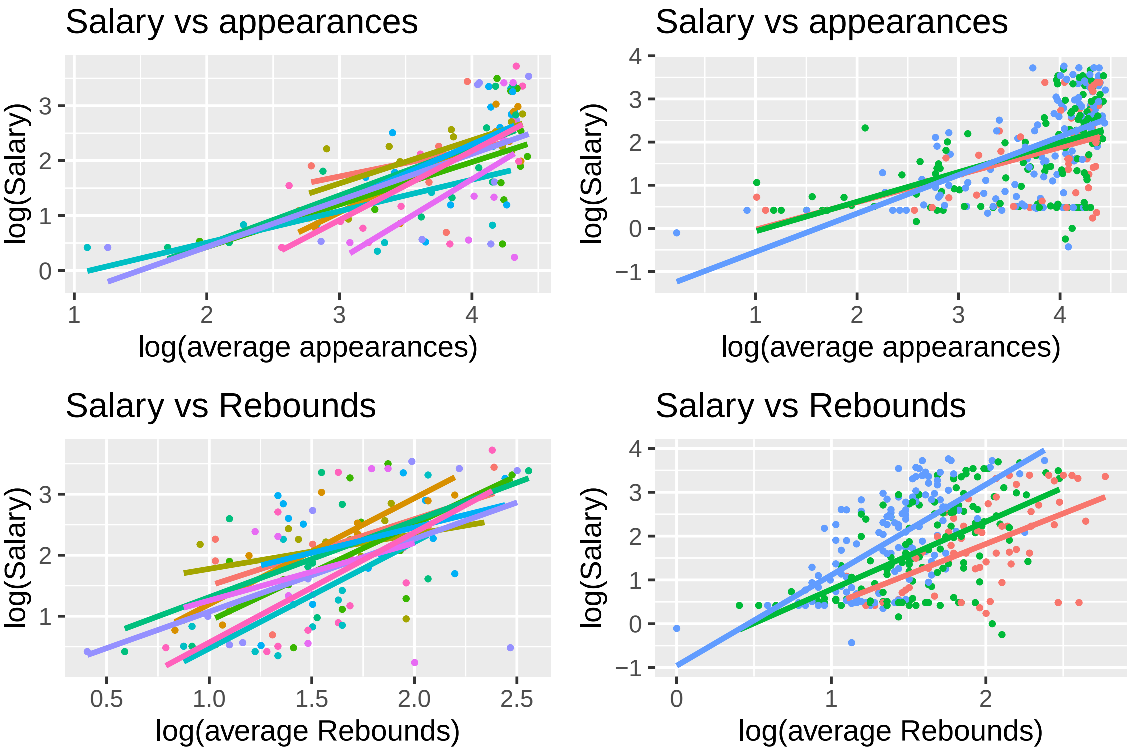

Figure 10: EDA: random effects of teama and positions(1)

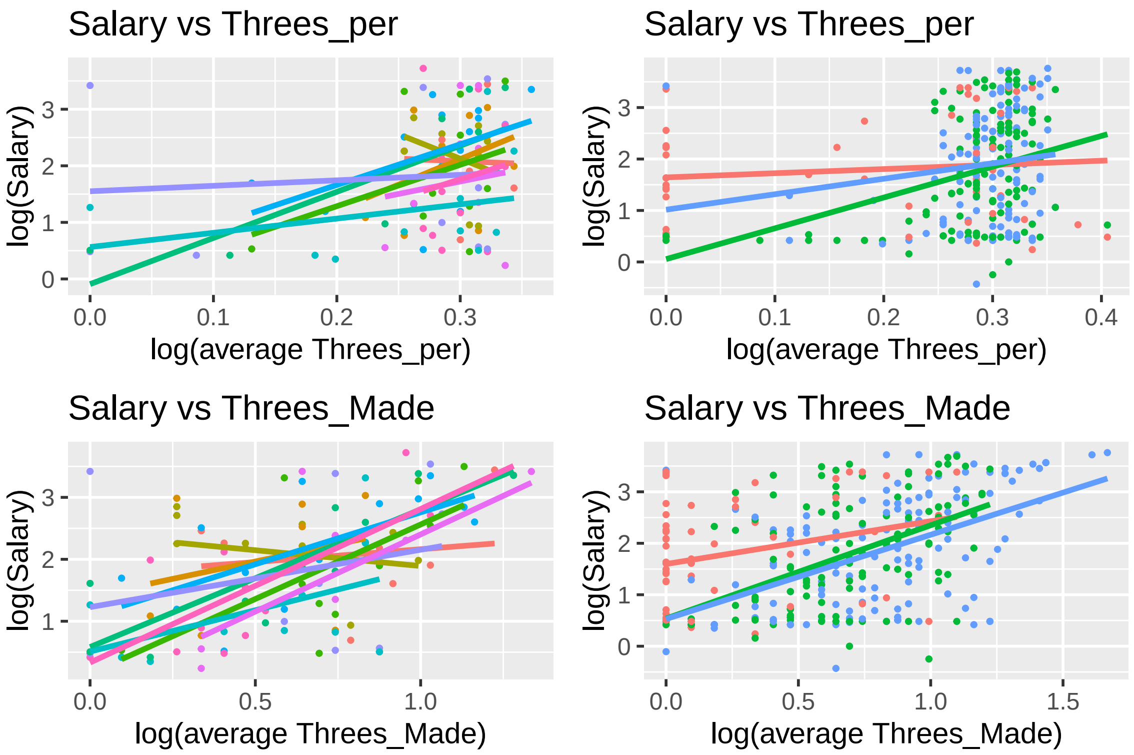

Figure 11: EDA: random effects of teama and positions(2)

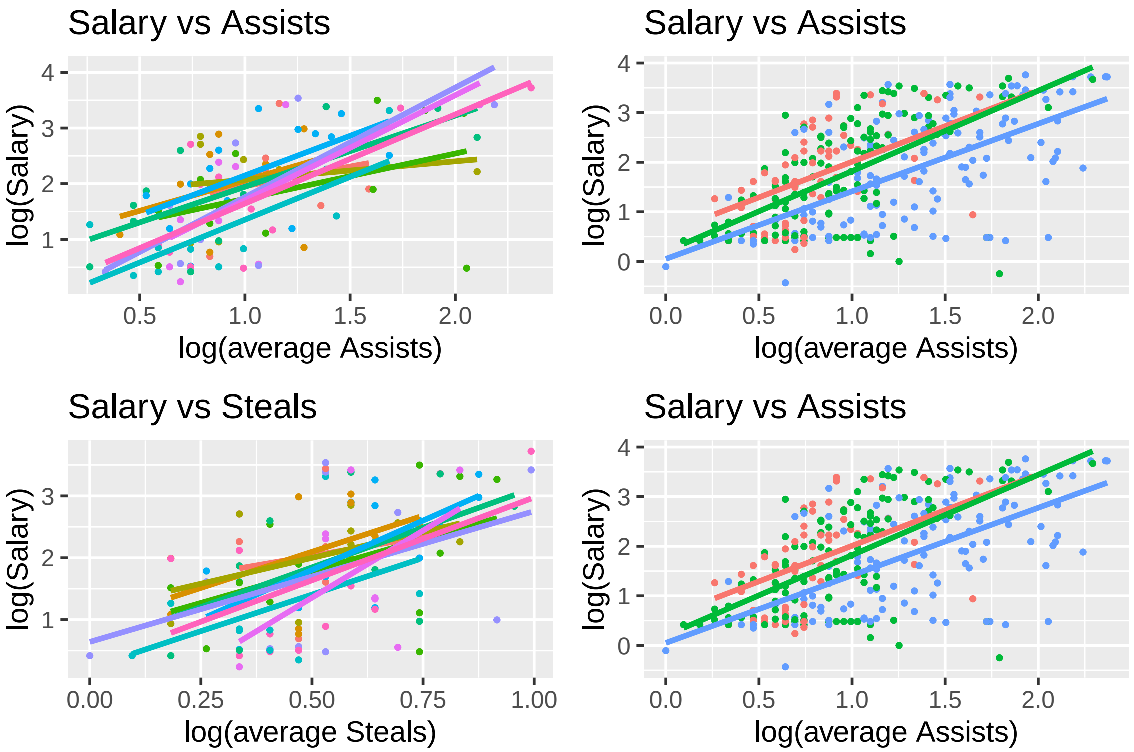

Figure 12: EDA: random effects of teama and positions(3)

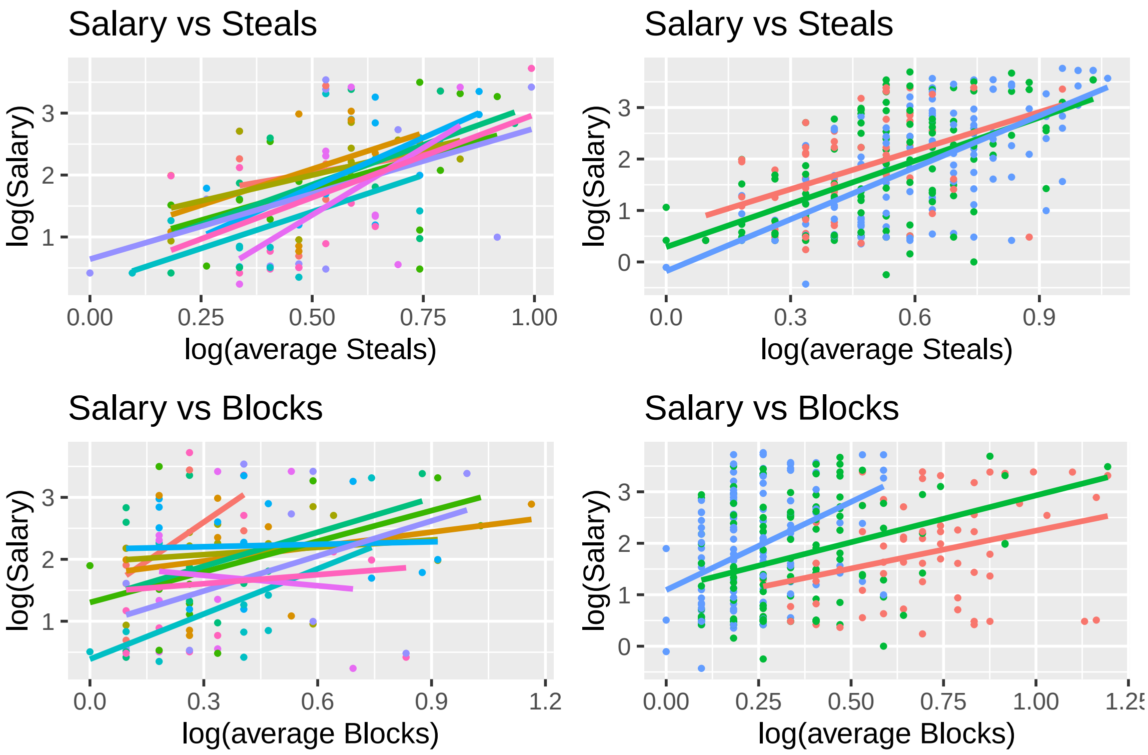

Figure 13: EDA: random effects of teama and positions(4)

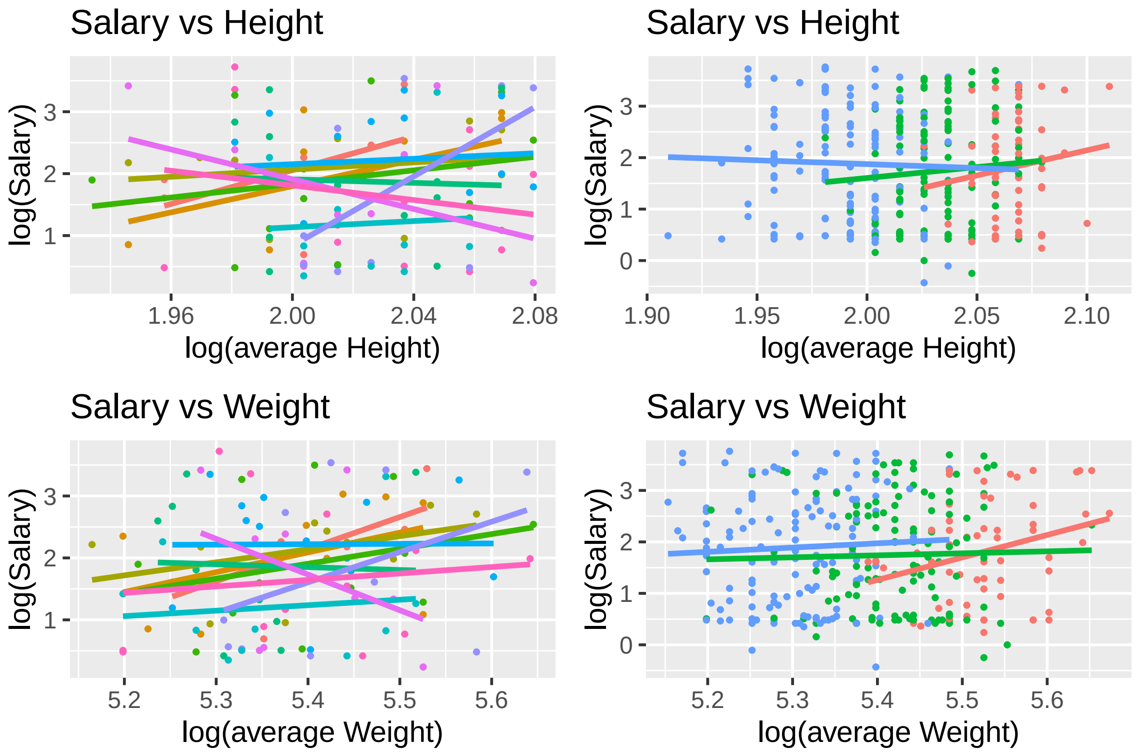

Figure 14: EDA: random effects of teama and positions(5)