Basketball_analysis

Kai

2022-12-05

Abstract

With 76 years of history and 2.86 billion dollars of worth, NBA (National Basketball Association) is definitely the number one basketball league in the world. The revenue for NBA players on average is about 7.5 million dollars a year, and the highest paid player is Stephen Curry with annual salary of 48,070,014 dollar and lowest is Ishmail Wainright with 633,891 dollar per year. And this made me curious, what type of player could earn big money in NBA? And how does that trend change over the years? I used multilevel model to find the influences of box scores, team and years on salary. And the results shows that different teams in favor of differnet type players, for example the Oklahoma City Thunder pay big money for players who can score, and Toronto Raptors pay big money for players who can dominant the rebound. And that trend changes slightly over the years. There are 5 parts of this report, include Introduction, Methods, Discussion, Results, Conclusion.

Introduction

With a such reputable name, more and more teenagers have a dream of playing in the NBA, but only 1 in 3333 (0.03%) could make it to the league. And among those elites, players with different set of skills in different time are paid differently. for example from 2000 - 2010 the two most paid player is Kevin Garnett (2000 - 04, 2006 - 09) and Shaquille O'Neal (2004 - 06), Kevin Garnett was playing for Minnesota Timberwolves as power forward, his features were incredible scoring ability in the field and intensive defensive ability. And Shaquille O'Neal was playing for Miami Heat as center, and he feature is dominant score and rebound ability, and intense defensive ability. Then from 2009 - 2019 there are three players who broke the cap of highest salaries, they are Koby Bryant (2009 - 16), lebron james(2016 - 17), and Stephen Curry (2017 - present), Among these three players Kobe is shooting guard and Stephen is point guard and Lebron is Power forward. Their features are strong field goal and 3-points ability. From this we could see how it was changed over the years.

Team plays a important role in value a player, and different team has different favors. Some team would pay luxury tax to get the players they want, to win the championship. For example, Oklahoma City Thunder they would pay large salary for players who have a strong field goal ability, and Toronto Raptors would pay high salaries for player who can dominant the rebound. Team is a critical factor that devides the salary. And to prove my point, I used multilevel models to find the influences of these fixed factors and random factors.

Methods

Data Preprocessing

My data was found on Kaggle(https://www.kaggle.com/datasets/patrickhallila1994/nba-data-from-basketball-reference?select=salaries.csv). This NBA data has six tables include boxscore, coaches, games, play_data, player_info, and salaries from 1996 to 2021.

Three tables are maerged to get all the information I need, these three tables are boxscore, games, and salaires. And my analysis is mainly targeting year from 2015 to 2021 so I filter out the data that is before 2015. And I have changed the type of some 'chr' variables into 'num' for some game statistics data like MP, FG, 3P, FT, TRB... etc, and I have transformed the 'num' type of the seasonStartYear into 'chr' for categorical analysis. And I get rid of the null value or strings like "Did Not Played". Then I grouped by playerName and year to get the average of all the game statistics and salaries of the players. Below is the final table that I created for my analysis, it has 16 variables and 2028 observations mainly includes player's name, team, year played, game statistics, and salary.

column names	explanation
playerName	Player's name
seasonStartYear	Team of player
MP	Minutes played in the game
FG	Field goal's made
3P	3-points made
FT	Free throws made
TRB	Total rebounds
AST	Assists
STL	Steals
TOV	Turnovers
BLK	Blocks
PF	Personal fouls
PTS	Points
InflationSalary	Annual Salary (adjusted to inflation)
Salary	Annual salary (dollars, not adjusting to inflation)
teamName	Average steals of player

Exploratory Data Analysis

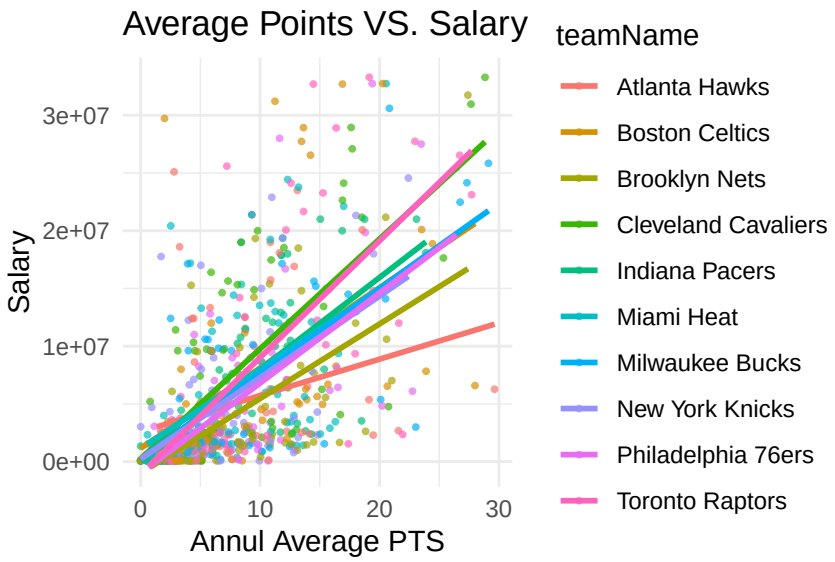

Figure 1: Correlation Between Average Points and Salary Group By Team

Figure 1 shows the correlation between the average points and salary from 2016 - 2019 season. From this graph we could see a clear relationship between the two and all of the team show a positive relationship, and the team with a steeper slope means this team pay more money for players with a strong scoring ability. Among all these teams we could see that the Toronto Raptors has the strongest positive relationship and Atlanta Hawks has the weakest positive relationship.

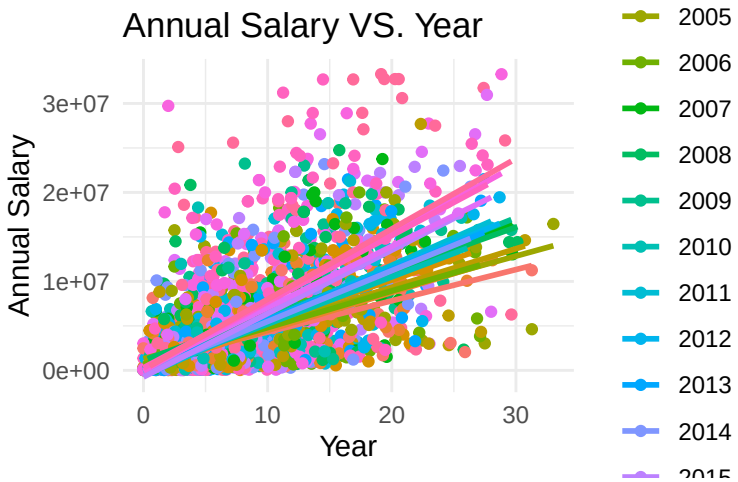

Figure 2: Correlation Between Average Points and Salary Group By Year

Figure 2 shows the correlation between average points a player get and the salary, and this is grouped by years. From this graph we could see that the coefficient are changing with years with no clear pattern, this might due to the global events that happend during that year. But for all these years the correlation between averagePTS and salary didn't change, it has always been a increase trend.

We can see there is a gap between group before - 2014 and after - 2014, the reason that gap was caused is because ESPN and TNT signed contract that will secure the broadcasting rights from 2014 to 2022 (9 Years) for 24 billion dollars, that is a huge number and that for sure lead to huge salary-cap changes for the coming years.

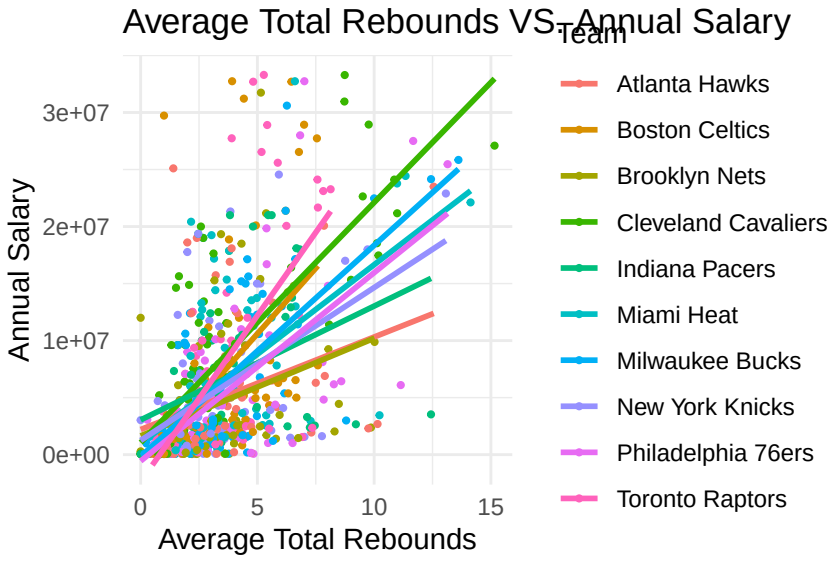

Figure 3: Correlation Between Average Rebounds and Salary Group By Team

Figure 3 shows the correlation between the average total rebounds in each game and the salary from 2016 - 2019 season. From this graph we could see that rebounds plays a important role in pricing a player. In NBA history, the center position always plays a essential role in a game, they are the critical factors of winning. One of the reason is that they could always get rebound and have a second chance to offense, or they could prevent the other team from attack again. The means that if one team dominant the rebound they could have endless chances of attacking until the goal is made. This is why the rebound is such a important factor in waging a player. And here we can see Atlantic Hawks and Toronto Raptors reversed their position, this showed that they truly focused on different aspects of players.

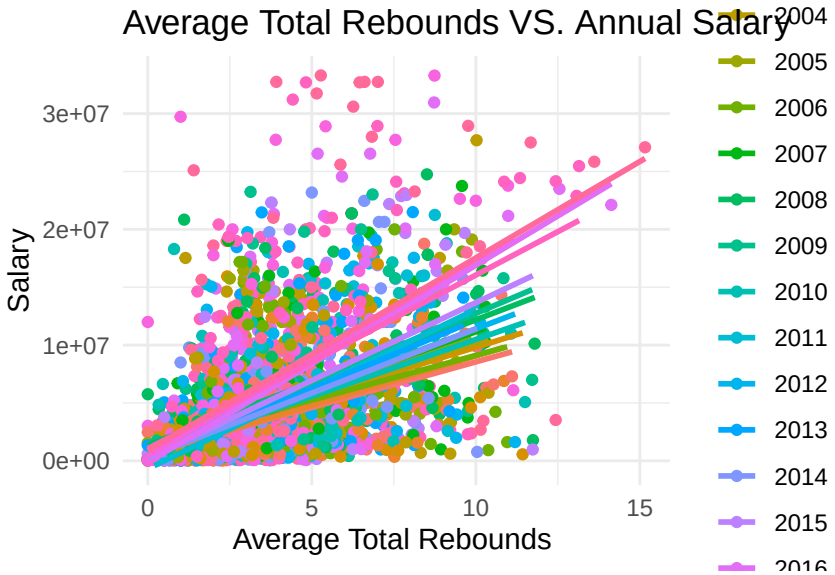

Figure 4: Correlation Between Average Rebounds and Salary Group By Year

This shows the correlation between rebounds and salary over the years. We can see the general trend is not changing all of them have a positive relationship.

Model Fitting

To better fit my multilevel model I decided to alter my variables to log(variable + 1) to minimize the skewness. And for the model I have chose 7 variables that I think are most influensial to salary, include averageFG, average3P, averageTRB, averageAST, averageSTL, averagePTS, and averageBLK.

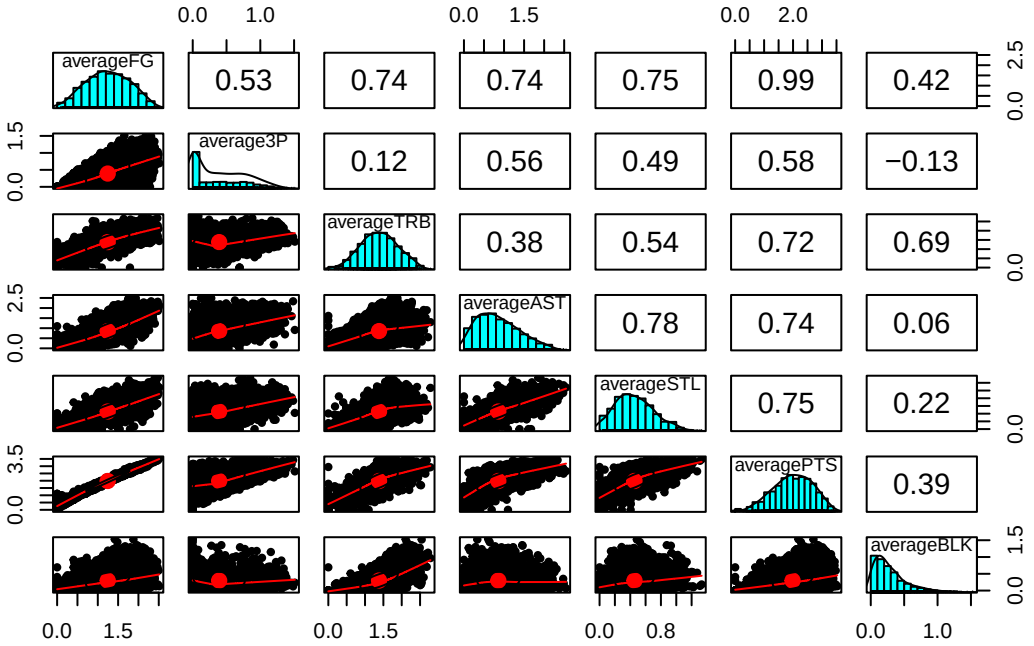

Figure 5: Correlation Matrix

Figure 5 is my correlation graph, and I set 0.7 as my threshold because a good player would perform well in multiple aspects so I set my threshold a bit higher to endorse that. From this graoh we could see there are several variables have a correlation that is higher than the treshold, for example, averageAST VS. averageSTL, averageAST VS. averageFG, and averagePST VS. averageFG. So I decided to drop averagePTS, averageSTL.

My final multilevel model has variables: averageFG, average3P, averageTRB, averageAST, and averageBLK. And because different teams has their preference and focus so the random effect of teams must take into account. And year could have a big impact to players salary too, with new rules been created over the years or economy change, so year is another random effect. And this is my final model down below and the P-values show all my variables are significant with a significant level of a = 0.5:

```
model <- lmer(averageSalary ~ averageFG + average3P  + averageTRB
              + averageAST + averageBLK
              + (averageFG + average3P  + averageTRB + averageAST | teamName)
              + (1 | seasonStartYear),data = log_finaltable)
```

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>
(Intercept)	1.238e+01	1.212e-01	8.544e+00	102.126	1.72e-14 ***
averageFG	5.301e-01	1.237e-01	1.071e+01	4.286	0.00136 **
average3P	4.217e-01	1.064e-01	1.128e+01	3.964	0.00211 **
averageTRB	8.023e-01	1.056e-01	1.273e+01	7.600	4.44e-06 ***
averageAST	2.922e-01	7.379e-02	1.120e+01	3.960	0.00216 **
averageBLK	3.192e-01	1.106e-01	2.768e+03	2.886	0.00393 **

```
ranef(model)
```

This table shows the influence of random effects. As we can see for different teams the coefficient is different and it is same for the year.

Result

Let's take Boston Celtics during 2019 season as a example, this is the formula with the fixed effect:

$$log(salary + 1) = 12.38 + 0.53 \times log(averageFG + 1) + 0.42 \times log(average3P + 1) + 0.80 \times log(averageTRB + 1) + 0.29 \times log(averageAST + 1))$$

This is the formula with the random effect:

$$log(salary + 1) = 12.28 + 0.385 \times log(averageFG + 1) + 0.48 \times log(average3P + 1) + 0.85 \times log(averageTRB + 1) + 0.345 \times log(averageAST + 1))$$

We could see here Boston Celtics has all the coefficients positive and among those the averageTRB is extremely high with a coefficient of 0.85, this means they have a competitive rebound ability in the league. And these coefficients mean that for every unit increase on the log(averageFG + 1) the log(salary + 1) will increase by 0.385, and for every unit increase for log(average3P + 1) the log(salary + 1) will increase by 0.48, and so on. And from team to team this coefficient is different because they have different preferences over players.

And we can see a clear change for the years, some of the year has a negative coefficient and some has positive. This could due to many confounding variables, for example the collapse of the subprime mortgage market in 2010, which led to a economic crisis in the global banking system, and Donald Trump's trade war with China, that slowdown in global economic growth in 2018. All these could be a factor that led to the coefficient change of the year.

Model Checking

Figure 6 is the residual plot, the residual plot show that there is no trend going on and it has a mean value of about 0. And Figure 8 is the Q-Q plot, it has a clear trend with a approximate coefficient of 1 and a mean of 0. So we could say that this model has a normal distribution with mean around 0 and with no skewness.

Discussion

The results supports my initial assumption that players with different skill sets are paid differntly in different teams, and this changes over the years. My multilevel model has include fixed effects include: averageFG, average3P, averageTRB, averageAST, and averageBLK, these are the factors I think could clearly show a player's ability and I think they are closely related to the salary. And two random effects of team and year, I think team plays a key role in setting the salary, they would pay high salary for the talent they are looking for. And year is another random effect that take into consideration of all the major events happened around the world that may impact the salary. I have also used two methods to check the accuracy of my model, and both the residual plot and Q-Q plot confirms the effectiveness of the model.

One thing to improve for future analysis is that the way to deal with missing value. The data I found has many null value because players could sometimes miss games or seasons due to injure. The way I deal with this missing value is just get rid of them, but I think if I could find a trend and use that value to fill those missing value, the results could be more accurate.

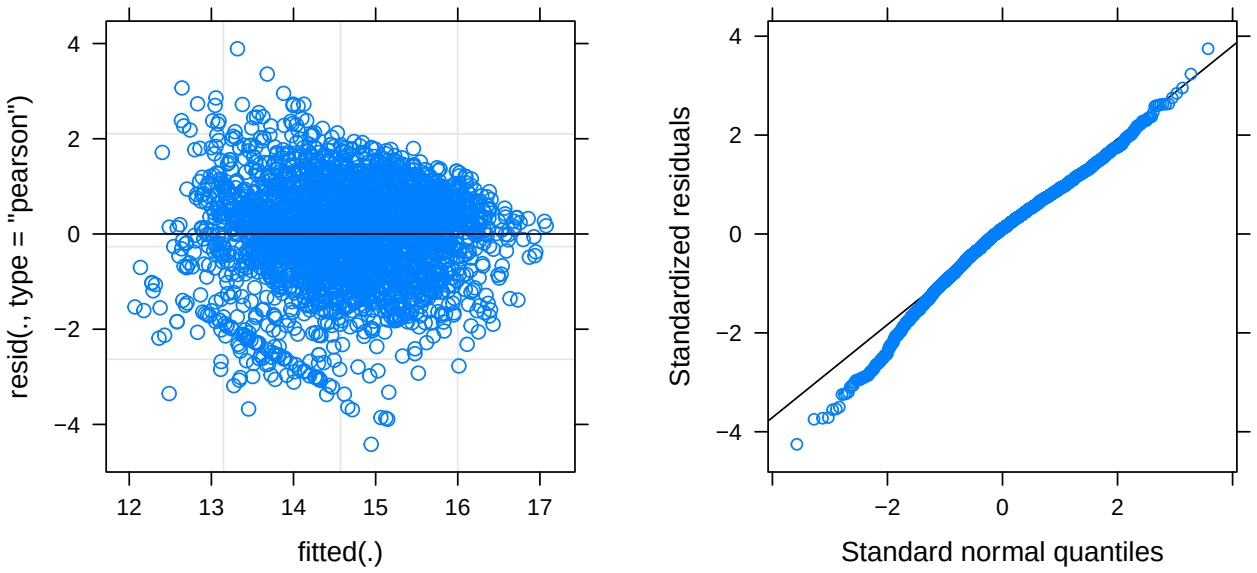

Figure 6: Residuals

Reference

Luke. (2021, October 19). Odds of making it to the NBA? (the math shows how hard it is). Dunk or Three. Retrieved December 11, 2022, from https://dunkorthree.com/odds-of-making-it-to-nba/

Rafferty, S. (2022, December 7). Who are the highest paid NBA players in 2022-23 season? Stephen Curry, Lebron James Top List. Sporting News. Retrieved December 11, 2022, from https://www.sportingnews.com/us/nba/news/highest-paid-nba-players-2022-23-season/bhwhafplgxiuuizmth8yujbu

The Editors of Encyclopaedia Britannica. (n.d.). National Basketball Association. Encyclopædia Britannica. Retrieved December 11, 2022, from https://www.britannica.com/topic/National-Basketball-Association