Titolo del progetto

Domenico Plantamura, Eduardo David Lotto, Manuel D'Alterio Grazioli, Gabriele Fugagnoli

Contents

Introduction

Objective of the project

Our goal is to investigate whether the salaries earned by the NBA players during the 2023-2024 season are fair in proportion to their performance during the current year's Regular season. To analyse performance, we selected several statistics: from the most common such as points, rebounds, assists to advanced metrics like Usage, Player Impact Estimated and Winning Shares. The idea is to deep dive into the relationship between salaries and performance through different models in order to understand what kind of relationship there is and which model best fits the data. Finally, we will compare actual salaries with those predicted by our models to find out which players (according to the models) are the most overpaid or underpaid.

Steps followed

To perform our analysis we followed these steps:

1. Data collection;
2. Data exploration;
3. Analysis;
4. Interpretation.

We now explain in depth each step.

Data collection

We performed a web scraping operation from the Official NBA Stats website, from which we collected most of the stats. Additionally, we downloaded data about the salaries from Hoopshype and other stats of interest from Basketball reference. All data concerns the 2023-2024 NBA Regular Season.

Why consider only Regular Season data?

Considering only data about Regular Season without considering players performance during playoffs limits a bit the potential of our analysis. On one hand, it's reasonable to infer that player performance during playoffs should have an important weight in determining his salary. On the other hand, considering playoffs in the analysis carries different issues. There are teams (and consequently players) that go further than others: 14 out of 30 teams can't qualify for the playoffs. For the teams which qualify, playoff stats are calculated on a number of games that could differ greatly between different teams (e.g. if a team loses in the first round, it plays from 4 to 7 games. If a team reaches the finals, it plays from 16 to 28 games). During Regular Season every team plays a fixed number of games, 82. Additionally, coaches usually rotate players at their disposal in a different way during playoffs: for instance, during regular season approximately 10-12 players for each team take part in the game; during playoffs it is not uncommon to observe only 7-8 players that come into play for each team. Furthermore, usually in a playoff game the best players are more involved compared to Regular season games. It means that, first of all, they play several more minutes. Moreover, they have the ball in their hands for a lot of time and consequently their stats grow a lot; hence, it could happen that few players record a large part of the entire team's statistics. Considering this, including playoffs data in the analysis could lead to an overestimation of performance of 2-3 players and to an underestimation of the performance of the rest of the team.

All in all, it is undeniable that playoffs are a fundamental part of the season. It is also obvious that if a player has more responsabilities in that phase he probably deserves a higher salary. But we think that for the purposes of our analysis, the addition of statistics collected on a small sample of matches, different for practically every team, with highly polarised data between the various players may lead to biases if not handled properly. We think that considering only the regular season, although leading to a limited analysis, may be sufficient to grasp the main relationships between salaries and performance.

Glossary

- **PLAYER NAME**: players' name
- **SALARY**: salary earned by a player for 2023-2024 season (collected from Hoopshype)
- **AGE**: players' age
- **POS**: "Position", states the playing position of a player

Traditional stats (collected from the NBA website)

- **GP**: "Games played", the number of games played by a player during the 2023-2024 regular season
- **FG_PCT**: "Field Goal Percentage", The percentage of field goal attempts that a player makes. Formula: (FGM)/(FGA)
- **FG3_PCT**: "3 Points "Field Goal Percentage", The percentage of 3pt field goal attempts that a player makes.
- **FT_PCT**: "Free throws Percentage", the percentage of free throws attempts that a player makes
- **OREB**: "Offensive Rebounds", The number of rebounds a player or team has collected while they were on offense
- **DREB**: "Defensive Rebounds", The number of rebounds a player or team has collected while they were on defense
- **REB**: "Rebounds"; A rebound occurs when a player recovers the ball after a missed shot. This statistic is the number of total rebounds a player or team has collected on either offense or defense

- **AST**: “Assists”, The number of assists – passes that lead directly to a made basket – by a player
- **TOV**: “Turnovers”; A turnover occurs when the player or team on offense loses the ball to the defense
- **STL**: “Steals”, Number of times a defensive player or team takes the ball from a player on offense, causing a turnover
- **BLK**: “Blocks”, A block occurs when an offensive player attempts a shot, and the defense player tips the ball, blocking their chance to score
- **BLKA**: “Blocks Against”, The number of shots attempted by a player or team that are blocked by a defender
- **PF**: “Personal fouls”, The number of personal fouls a player or team committed
- **PFD**: “Personal fouls drawn”, The number of personal fouls that are drawn by a player or team
- **PTS**: “Points”, the number of points scored by a player
- **MIN**: “Minutes played”, number of minutes played by a player during the 2023-2024 Regular season
- **MIN_G**: “Minutes played per game”

Advanced stats (collected from the NBA website)

- **OFF_RATING**: “Offensive Rating”, Measures a team’s points scored per 100 possessions. On a player level this statistic is team points scored per 100 possessions while they are on court. Formula: 100*((Points)/(POSS)
- **DEF_RATING**: “Defensive Rating”, The number of points allowed per 100 possessions by a team. For a player, it is the number of points per 100 possessions that the team allows while that individual player is on the court. Formula: 100*((Opp Points)/(Opp POSS))
- **NET_RATING**: “Net Rating”, Measures a team’s point differential per 100 possessions. On player level this statistic is the team’s point differential per 100 possessions while they are on court. Formula: OFFRTG - DEFRTG
- **AST_TO**: “Assist to Turnover Ratio”, The number of assists for a player or team compared to the number of turnovers they have committed
- **TS_PCT**: “True Shooting Percentage”, A shooting percentage that factors in the value of three-point field goals and free throws in addition to conventional two-point field goals. Formula: Points/ [2*(Field Goals Attempted+0.44*Free Throws Attempted)]
- **USG_PCT**: “Usage Percentage”, The percentage of team plays used by a player when they are on the floor. Formula: (FGA + Possession Ending FTA + TO) / POSS
- **PIE**: “Player Impact Estimate”, PIE measures a player’s overall statistical contribution against the total statistics in games they play in. PIE yields results which are comparable to other advanced statistics (e.g. PER) using a simple formula. Formula: (PTS + FGM + FTM - FGA - FTA + DREB + (.5 * OREB) + AST + STL + (.5 * BLK) - PF - TO) / (GmPTS + GmFGM + GmFTM - GmFGA - GmFTA + GmDREB + (.5 * GmOREB) + GmAST + GmSTL + (.5 * GmBLK) - GmPF - GmTO)

The stats below are collected from Basketball Reference:

- **WS**: “Win Shares”; Win Shares is a player statistic which attempts to divvy up credit for team success to the individuals on the team. It is calculated using player, team and league-wide statistics and the sum of player win shares on a given team will be roughly equal to that team’s win total for the season (more details here https://www.basketball-reference.com/about/ws.html).
- **BPM**: “Box Plus/Minus”; a box score estimate of the points per 100 possessions that a player contributed above a league-average player, translated to an average team
- **VORP**: “Value Over Replacement Player”; a box score estimate of the points per 100 TEAM possessions that a player contributed above a replacement-level (-2.0) player, translated to an average team and prorated to an 82-game season. Multiply by 2.70 to convert to wins over replacement.

BPM and VORP are calculated per 100 possessions; MIN and WS are calculated over the whole regular season, MIN_G is calculated per game. The other stats are considered per 48 minutes.

BPM and VORP are calculated per 100 possessions; MIN and WS are calculated over the whole regular season, MIN_G is calculated per game. The other stats are considered per 48 minutes.

Why statistics per 48 minutes?

Considering most statistics projected over 48 minutes avoids overestimating performance for players who play, on average, more minutes in a game. In this way we think that the contribution of each player is fairly evaluated and not distorted by the minutes played.

Data integration and cleaning

Once we had obtained the tables of interest, we selected from each table the statistics useful for analysis (those given in the glossary) and then merged the slices of the various datasets.

At this stage, the data were cleaned:

- NA removal
- Matching players' names
- Transforming the Salary column into a numeric one
- Removing players who played less than 480 minutes during the entire regular season

The reason why we selected players with at least 480 minutes played is that we wanted to avoid considering stats taken on a too small amount of minutes. After these operation, the final dataset consists of 360 rows and 31 columns.

```
data_st <- merge(data_salary, data_traditional_per48, by = "PLAYER_NAME", all = TRUE)
data_ast <- merge(data_st, data_advanced, by = "PLAYER_NAME", all = TRUE)
data_mast <- merge(data_ast, data_miscellaneous, by = "PLAYER_NAME", all = TRUE)
data_mastt <- merge(data_mast, data_trad_tot, by = "PLAYER_NAME", all = TRUE)
final_dataset <- merge(data_mastt, data_vorp, by = "PLAYER_NAME", all = TRUE)
```

Data exploration

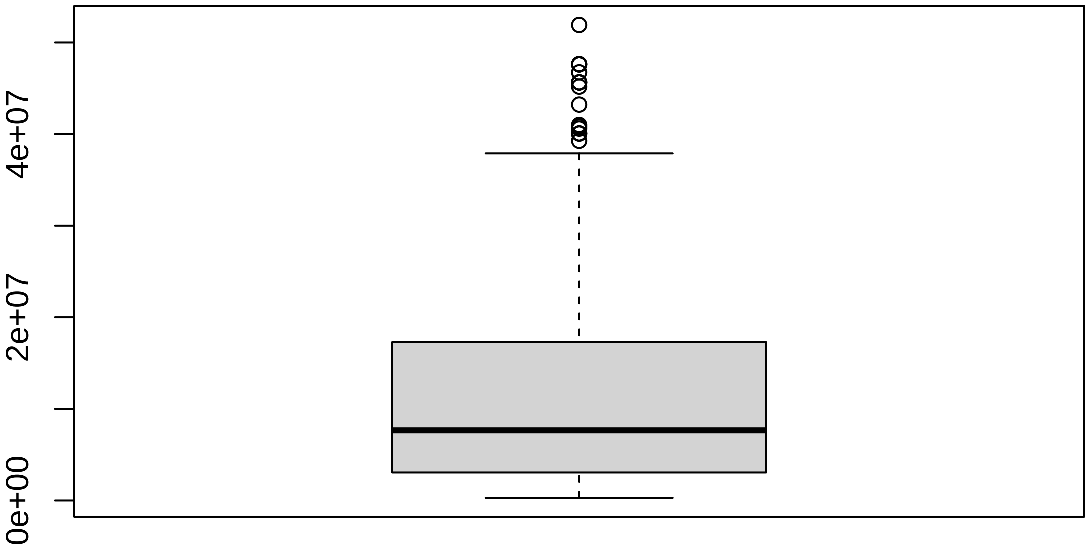
Before studying the data with formal models, we got an overview through an exploratory data analysis.

	PLAYER_NAME	Salary	AGE	GP	FG_PCT	FG3_PCT	FT_PCT	OREB	DREB	REB
3	Aaron Gordon	22266182	28	73	0.556	0.290	0.658	3.6	6.2	9.8
4	Aaron Holiday	2346614	27	78	0.446	0.387	0.921	0.9	3.8	4.7
5	Aaron Nesmith	5634257	24	72	0.496	0.419	0.781	1.5	5.1	6.6
6	Aaron Wiggins	1836096	25	78	0.562	0.492	0.789	2.3	4.9	7.3
12	Al Horford	10000000	37	65	0.511	0.419	0.867	2.3	9.1	11.4

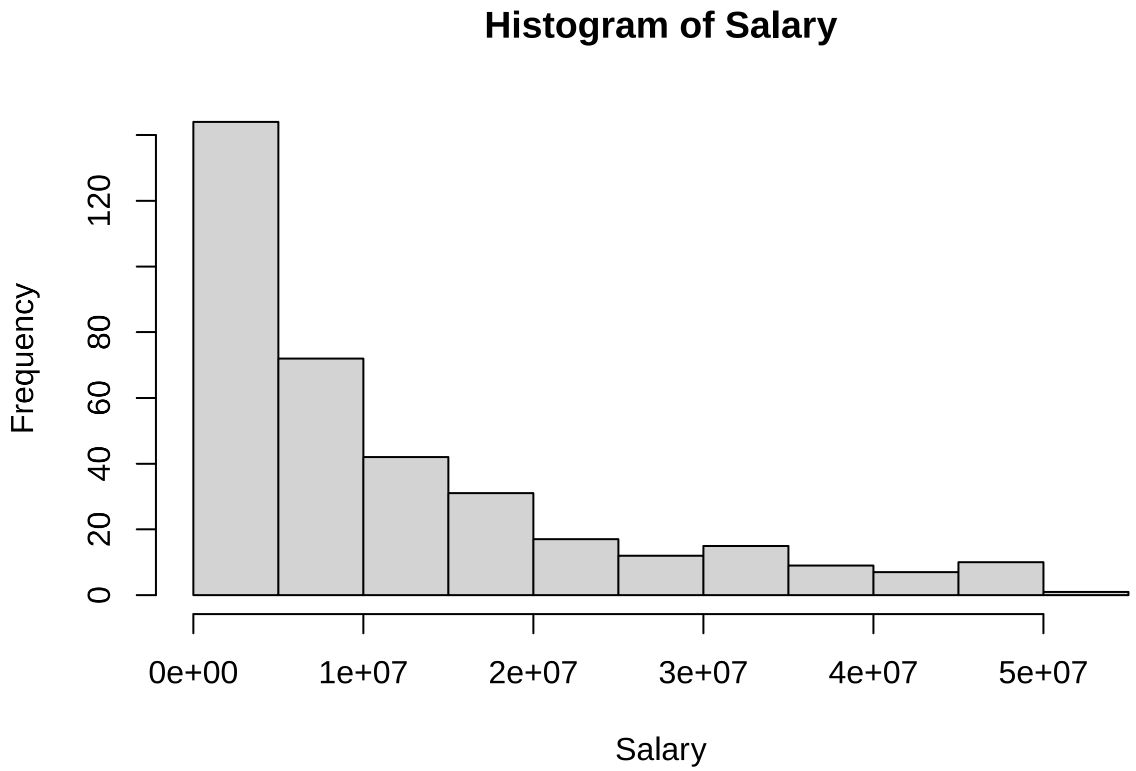
	AST	TOV	STL	BLK	BLKA	PF	PTS	OFF_RATING	DEF_RATING	NET_RATING
3	5.4	2.2	1.2	0.9	1.2	3.0	21.2	119.8	111.1	8.7
4	5.3	2.0	1.6	0.2	0.8	4.7	19.4	110.5	107.6	2.9
5	2.6	1.5	1.6	1.2	1.2	5.8	21.1	119.3	115.0	4.3
6	3.4	2.2	2.2	0.7	1.3	3.6	21.2	115.6	110.0	5.7
12	4.6	1.3	1.0	1.7	0.3	2.6	15.5	120.9	109.5	11.4

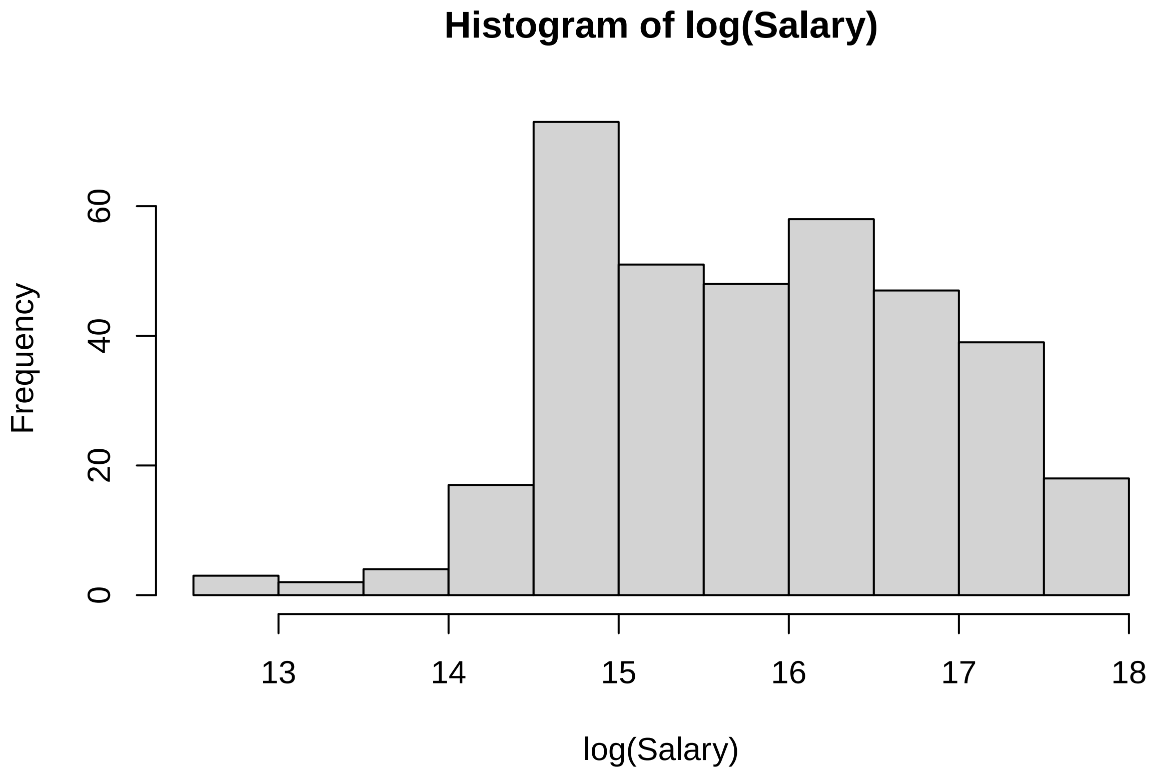
	AST_TO	TS_PCT	USG_PCT	PIE	PFD	MIN	MIN_G	Pos	WS	BPM	VORP
3	2.47	0.607	0.174	0.103	4.7	2296.810	31.46315	PF	7.1	1.3	1.9
4	2.64	0.578	0.158	0.078	2.5	1269.297	16.27303	PG	2.5	-1.5	0.2
5	1.69	0.631	0.158	0.071	3.5	1994.655	27.70354	SF	4.1	-0.5	0.8
6	1.54	0.664	0.163	0.096	2.3	1227.938	15.74280	SG	3.7	0.7	0.8
12	3.50	0.650	0.119	0.105	0.8	1739.797	26.76610	C	6.2	3.6	2.5

Firstly, an analysis of the variable Salary that will be the dependent variable in the models.



```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   289542  3065128  7657240 12061891 17271922 51915615
```


Histogram of Salary
Frequency
0
20
40
60
80
120
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
Salary



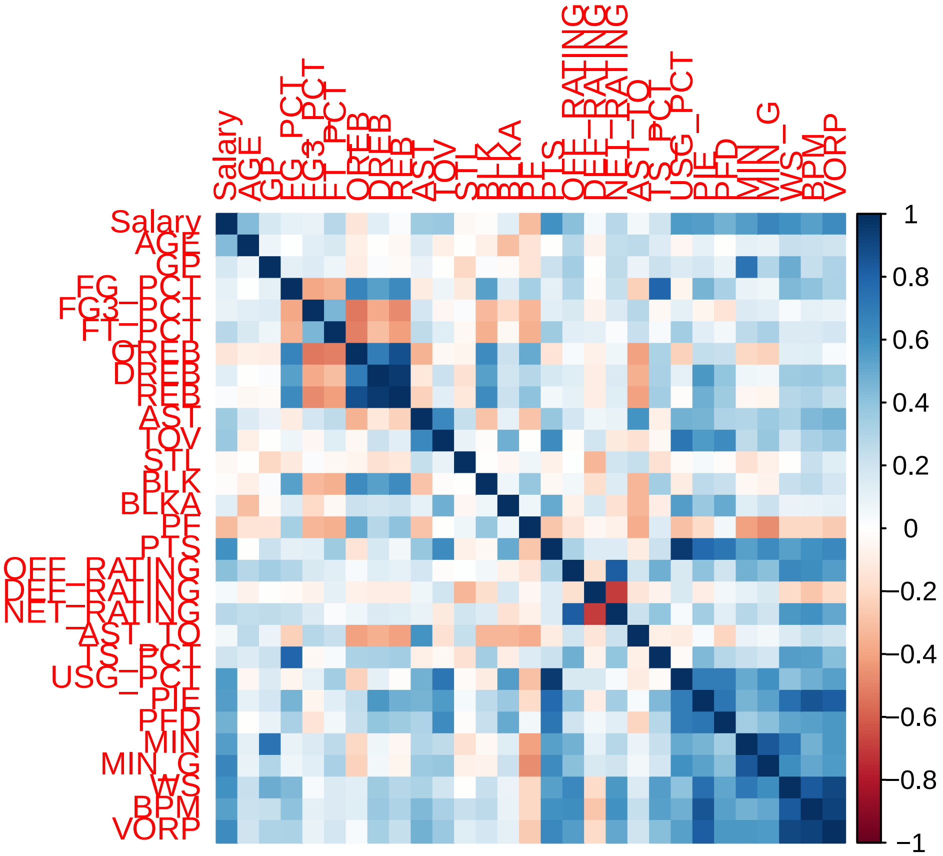
The boxplot shows that the salary distribution is right skewed, with some outliers in the right side. We expected this kind of distribution, the outliers are the players earning the highest salaries. The histogram also highlights the right skewed distribution. The asimmetry could be mitigated by applying a logarithmic transformation to the variable.

In order to study correlations between the variables that will be the independent ones in the models, we used the functions corrplot and pairs.

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(cor(fd_numeric), method = 'color')
```



```
#corrplot(cor(fd_numeric), method = 'ellipse')
```