P8106 - Final Project - NBA Players Salary Prediction

Mingkuan Xu, Mengfan Luo, Yiqun Jin

5/9/2022

Introduction

For any team in the National Basketball Association (NBA), a key strategy to win more games is to properly allocate their salary cap - an agreement that places a limit on the amount of money that a team can spend on players' salaries. How to evaluate the performance of each NBA player and give a suitable level of salary is a therefore complicated problem. In this project, we intend to predict the salary of NBA players in the 2021-2022 season based on their game statistics. We collected game statistics that are commonly used to evaluate players from the NBA official website, built both linear and non-linear models, including linear regression, ridge regression, lasso regression, GAM, MARS, _____, on selected feature variables, and compared these models to determine a final predictive model.

Data preprocessing

We will conduct data analysis and model construction based on two datasets on NBA players' contracted salary [1] and performance statistics per game [2] in 2021-2022. The following steps are included in our data preparation:

- Two original datasets are inner joined by players and teams
- Keep only one record with most number of games played for each of players, given a player may transfer to other teams during the session and have multiple records.
- Remove 5 variables with missing values caused by division of other existing variables.

The final cleaned dataset has 442 records and 24 columns as followed:

- `position` – A categorical variable of the player's position (C, PF, SF, SG, PG)
- `age` – Player's age on February 1 of the season
- `team` – A categorical variable of the player's playig team
- `game` – Number of games played
- `game_starting` – Number of games played as a starter
- `minute` – Minutes played per game
- `field_goal` – Field goals per game
- `fg_attempt` – Field goal attempts per game
- `x3p` – 3-point field goals per game
- `x3p_attempt` – 3-point field goal attempts per game

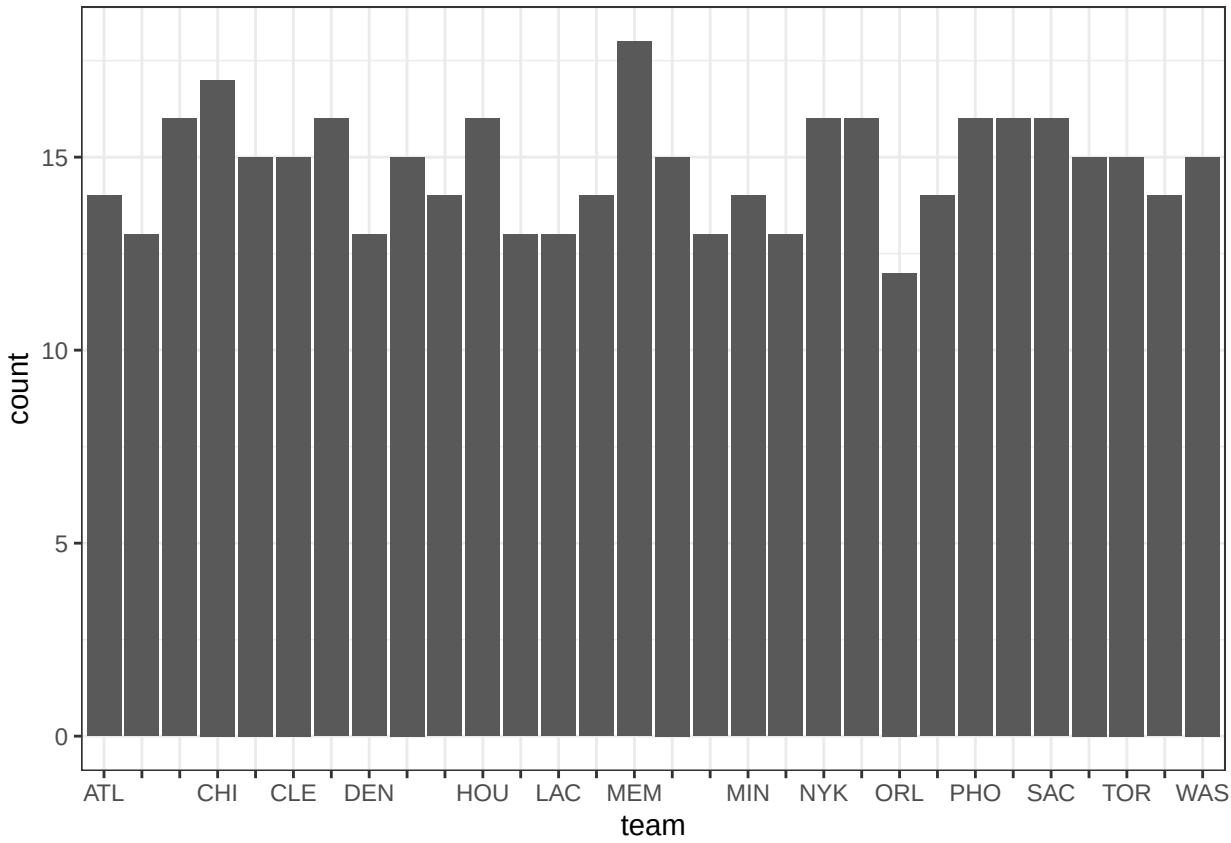
- `x2p` – 2-point field goals per game
- `x2p_attempt` – 2-point field goal attempts per game
- `free_throw` – Free throws per game
- `ft_attempt` – Free throw attempts per game
- `offensive_rb` – Offensive rebounds per game
- `defenssive_rb` – Defensive rebounds per game
- `total_rb` – Total rebounds per game
- `assistance` – Assists per game
- `steal` – Steals per game
- `block` – Blocks per game
- `turnover` – Turnovers per game
- `personal_foul` – Personal fouls per game
- `point` – Points per game
- `salary` – Salary of the player in million

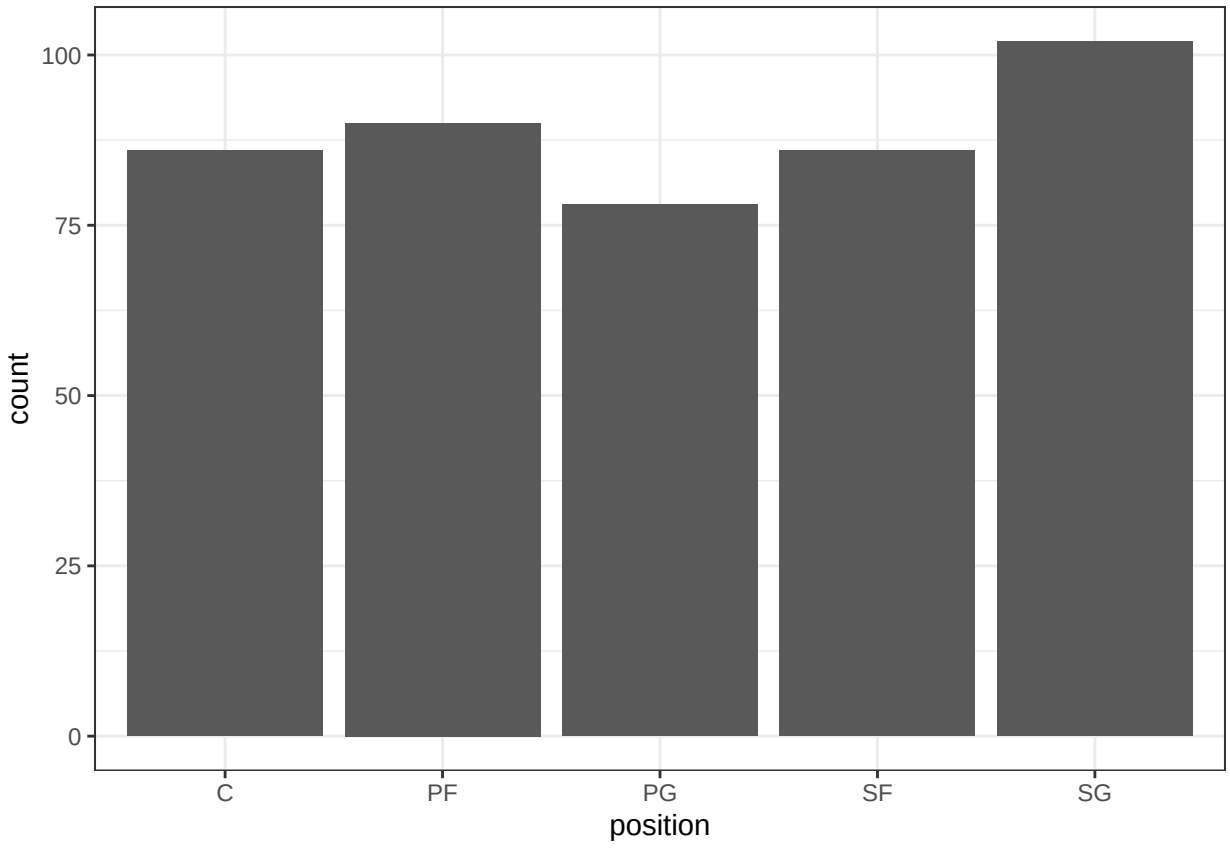
Since `minute` stands for minutes played per game, we will divided variables stands for counts by `minute` to get a rate. These variables includes `field_goal`, `fg_attempt` `x3p`, `x3p_attempt`, `x2p`, `x2p_attempt`, `free_throw`, `ft_attempt`, `offensive_rb` `defenssive_rb`, `total_rb`, `assistance`,`steal`, `block`, `turnover`, `personal_foul` and `point`.

Exploratory Analysis

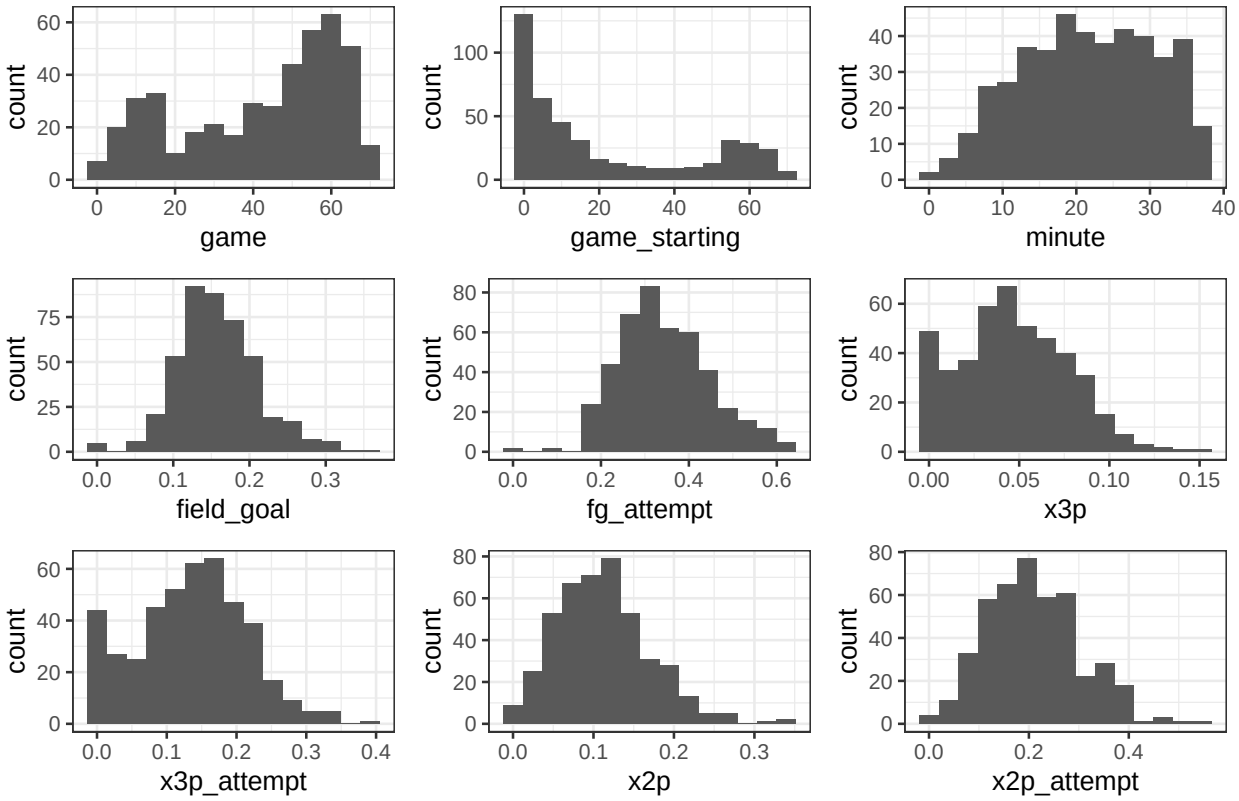
Univariate Analysis

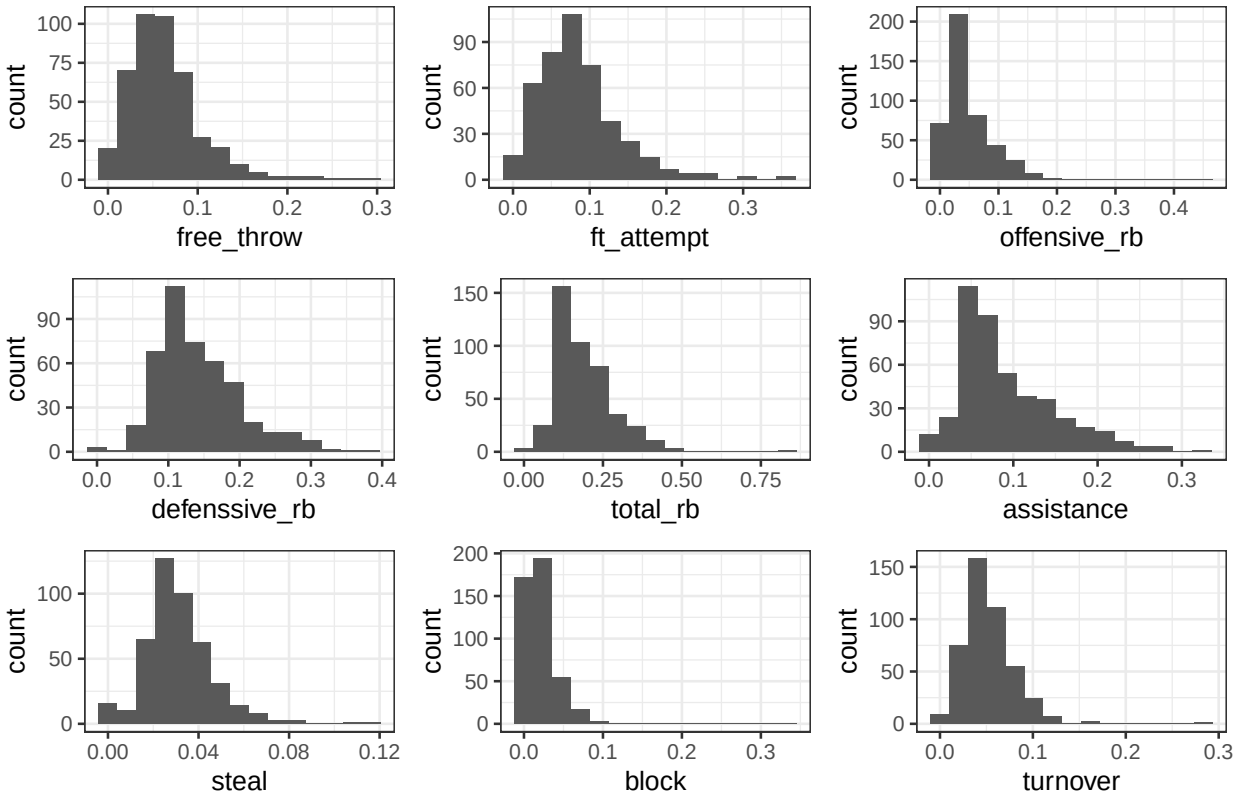
Distributions of the two categorical variables, `team` and `position`.

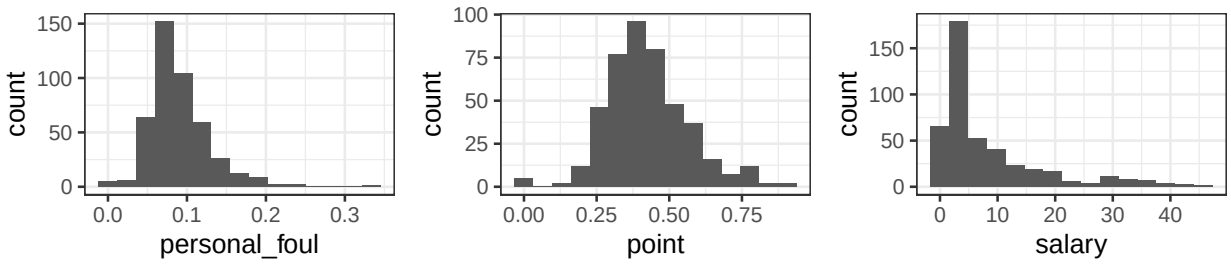

count
15
10
5
0
ATL
CHI
CLE
DEN
HOU
LAC
MEM
MIN
NYK
ORL
PHO
SAC
TOR
WAS
team



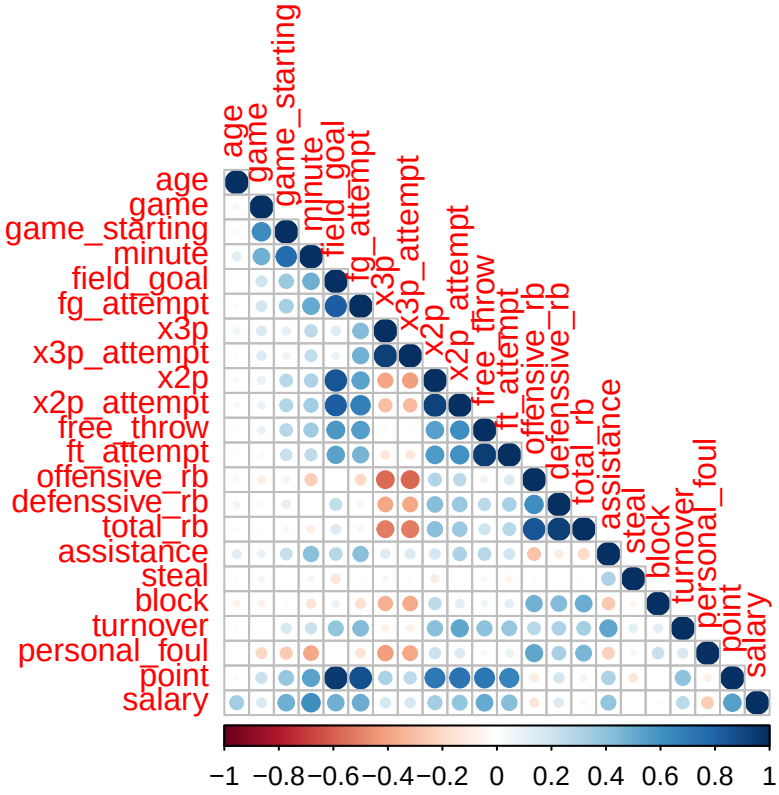
Distributions of other numeric variables.


Histograms of Predictive Variables (Group A)
count
game
game_starting
minute
field_goal
fg_attempt
x3p
x3p_attempt
x2p
x2p_attempt


Histograms of Predictive Variables (Group B)
count
free_throw
ft_attempt
offensive_rb
defenssive_rb
total_rb
assistance
steal
block
turnover

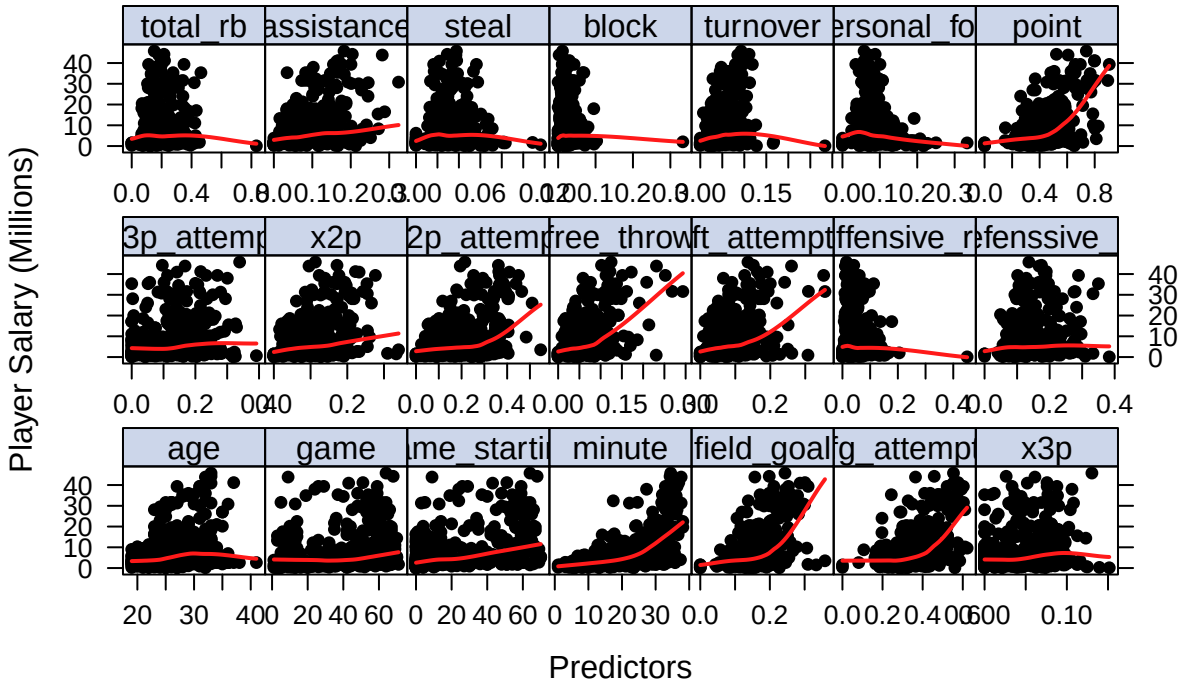

Histograms of Predictive Variables (Group C)
count
150
100
50
0
0.0
0.1
0.2
0.3
personal_foul
count
100
75
50
25
0
0.00
0.25
0.50
0.75
point
count
150
100
50
0
0
10
20
30
40
salary

Correlation Analysis

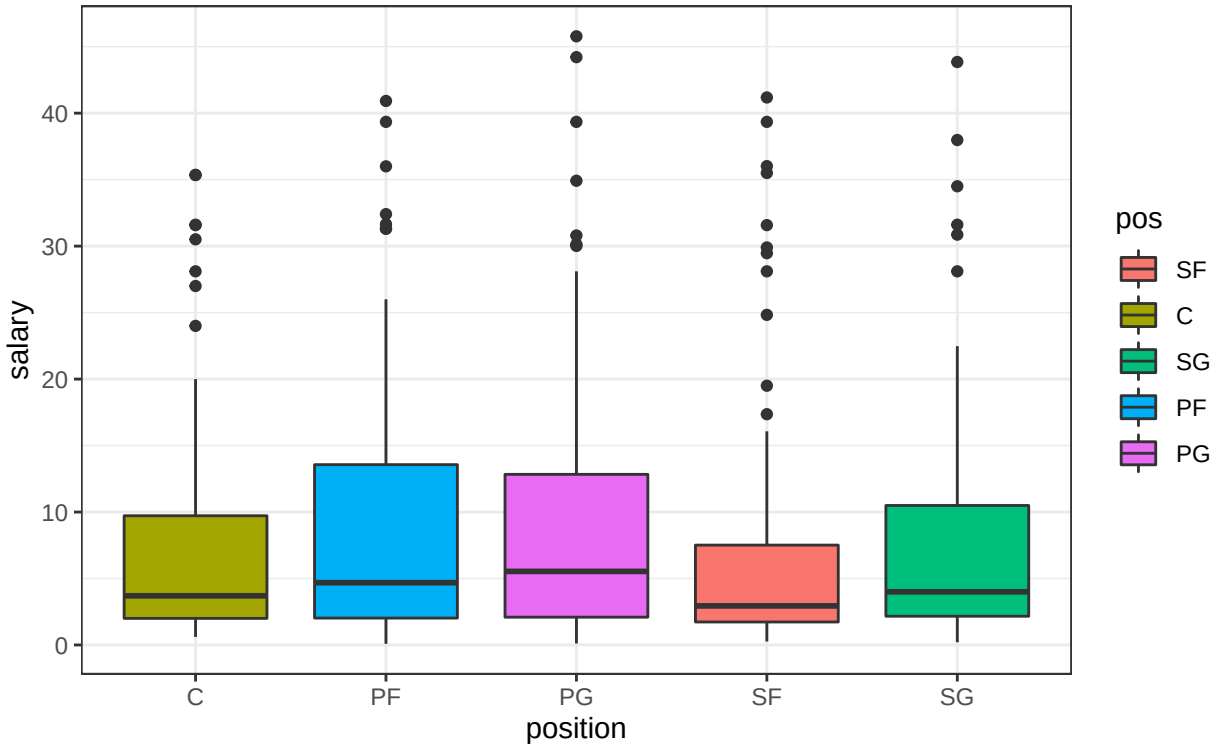


Analyzing trends in data

From numeric variables, we found that `stl`,`x3p`, `age`,`gs` seem to have some non-linear trends.



From categorical variable `position`, extremely high values in salary show in all positions and some teams.



Model Construction

- What predictor variables did you include?
- What technique did you use? What assumptions, if any, are being made by using this technique?
- If there were tuning parameters, how did you pick their values?
- Discuss the training/test performance if you have a test data set.
- Which variables play important roles in predicting the response?
- **Explain/visualize the final model you select.**
- What are the limitations of the models you used (if there are any)? Are the models flexible enough to capture the underlying truth?

Conclusion

- What were your findings? Are they what you expect? What insights into the data can you make?

References

[1]https://www.basketball-reference.com/contracts/players.html

[2]https://www.basketball-reference.com/leagues/NBA_2022_per_game.html