NBA Players Salary Prediction

Statistical Learning Final Exam Project

Mahir Selek (2041295), Joi Berberi (2023363)

1. Aim of the project

As it is known, the NBA is one of today's biggest sports organizations, even one of the most popular. It is a huge organization that is regularly followed by so many people around the world every year, and beyond that it is a huge monetary market.

As two sports enthusiast students, we decided to make a project about basketball players playing in this popular league. We wanted to do a forecasting project because it is a big league in terms of money and because the income wages of the players arouse curiosity every year. At the same time, this project should have been linked to the content of our course. That's why we decided it was appropriate for this project to do a regression project on players' statistics and income salaries.

2. Dataset Description

For this project we decided to use the "NBA Players stats since 1950" dataset which is available on Kaggle. However "player_salary" dataset was not provided at Kaggle. So We scraped from '<https://www.basketball-reference.com/contracts/players.html>' website and created by ourselves. The first dataset contains aggregate individual statistics for 67 NBA seasons since 1950. From basic box-score attributes such as points, assists, rebounds etc., to more advanced money-ball like features such as Value Over Replacement. However, unfortunately we didn't focus many of the attributes in the dataset because they were not useful for our project. For this reason we eliminate them in this way we were able to focus on the statistical analysis. We will show that later inside the project.sdf

2.1.Data Components

The reason for the NA values this dataset begins from 1950 and at the beginning some of the values are empty. These are the features that can not be NA because we can not give them a default value or the missing value compromised the future model.

-pos: position of the player
-Tm: Team of the player
-G: Players total game number during one season
-GS: Games Started (available since the 1982 season)
-MP: Total minutes of the player in one season
-BLK: Total block count of the player in one season
-AST: Total assist number of the player in one season
-PTS: Total score of the player in one season
-STL: Total steal number of the player in one season
-TRB: Total Rebound number of the player in one season

-eFG: Effective Field Goal Percentage; the formula is (FG + 0.5 * 3P) / FGA.
This statistic adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal. For example, suppose Player A goes 4 for 10 with 2 threes, while Player B goes 5 for 10 with 0 threes. Each player would have 10 points from field goals, and thus would have the same effective field goal percentage (50%). -FGA: Field Goal Attempts (includes both 2-point field goal attempts and 3-point field goal attempts)
-FT: Free Throws
-2P: 2-Point Field Goals
-2PA: 2-Point Field Goal Attempts
-3P: 3-Point Field Goals (available since the 1979-80 season in the NBA)
-DRB: Defensive Rebounds (available since the 1973-74 season in the NBA) ...

Other categories(columns) containing player data throughout the seasons. As we mentioned above we are not going to use many of them because of they had no value in our analysis.

3. Data Cleaning and Filtering

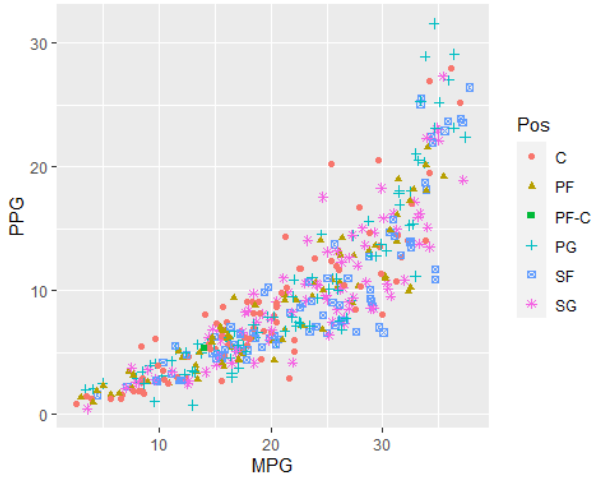
The advantage of the filtering after 2017 is that we don't have any NA or empty feature anymore. Thus, we have transformed our data into a more useful form. distinct functions help us to retain only unique/distinct rows from our input tables. Aim of the mutation is that in the seasons_stats file we don't have stats per game features. So, we mutated all of them to use in our salary prediction project. Our main purpose is to investigate how the stats effect next season's salary the players get.

Then we merged the two dataset that we have. After that we checked out new dataset and we decided to use only necessary features for our models. Our new dataset become clearer and more understandable. We prefer to use specific data belongs only on 2017. Also, we created new variables with mutate function to predict better and understandable data.

Why we choose these 7 parameters after cleaning and filtering? Because these parameters are the numerical parameters on which we can predict what we need for our prediction models

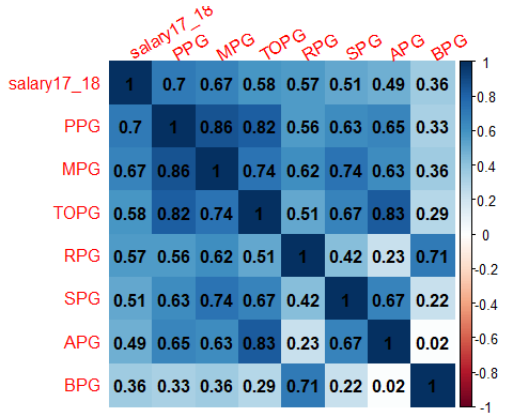
4. Exploratory data analysis

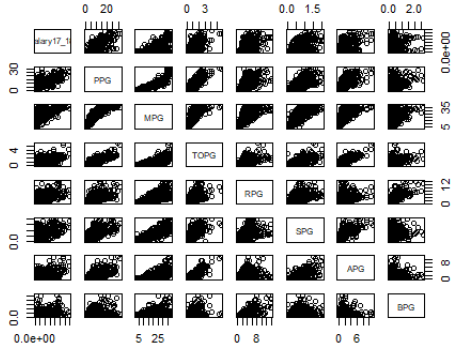
First of all before start in order to get an idea, we wanted to look at the distribution relationship of the numerical data we have with the positions of the players.

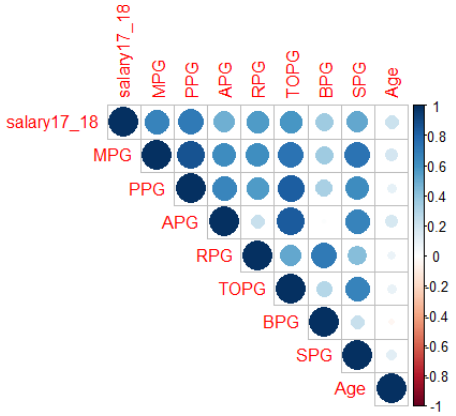


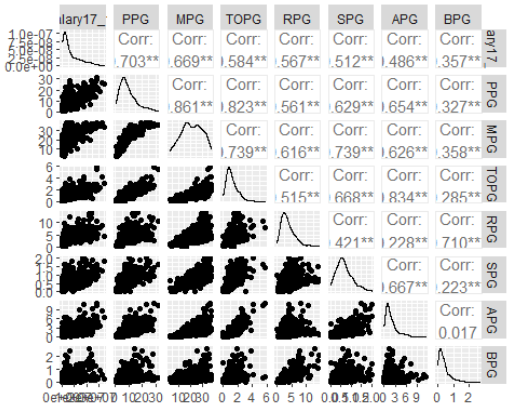
Then we did was to look at the distributions of the continuous variables conditioned on the stats salary. From this summary statistic of the data we noticed that there are some strange values, in particular some values of MPG, BPG and SPG seems too high from a medical point of view

At this point, we preferred to use correlation in order to look at the data we have from the outside. Being able to draw such a straight line helps us not only predict the unknown but also understand the relationship between the variables better.









We also expect some variables to be correlated, it's better to check that before fitting some model. For continuous variables we checked the correlation plots and the correlation matrix to get a visual idea. Thanks to the correlation, it helped us to have an idea about which data would more closely affect our result.

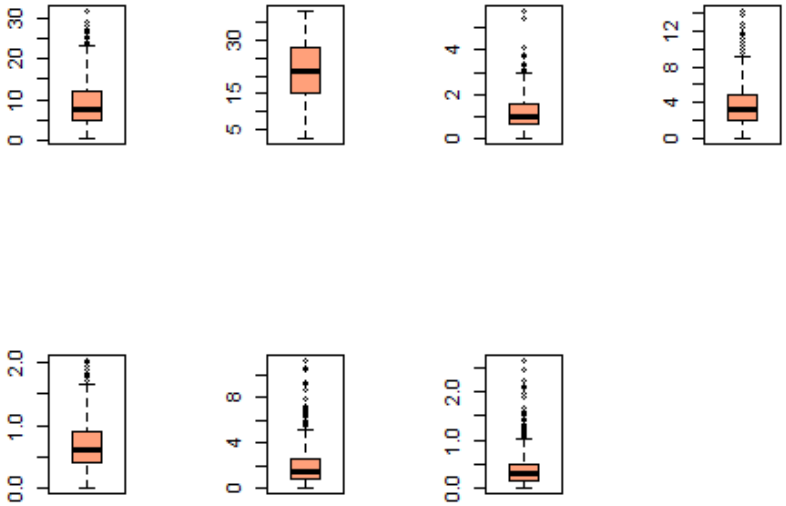
Correlation strength is: PPG > MPG > TOPG > RPG > PER > SPG > APG

The interesting part of this is that the number of turnover players make is linked to their salary, and the relationship has a positive correlation.

So, We interpreted this relationship like this: "the more turnovers they make" means that they are more involved in ball movements in games, which means that players who make turnovers are, at some extend, important to their team. and I thought this could be expressed as "aggressiveness". I know this interpretation could not be appropriate one.

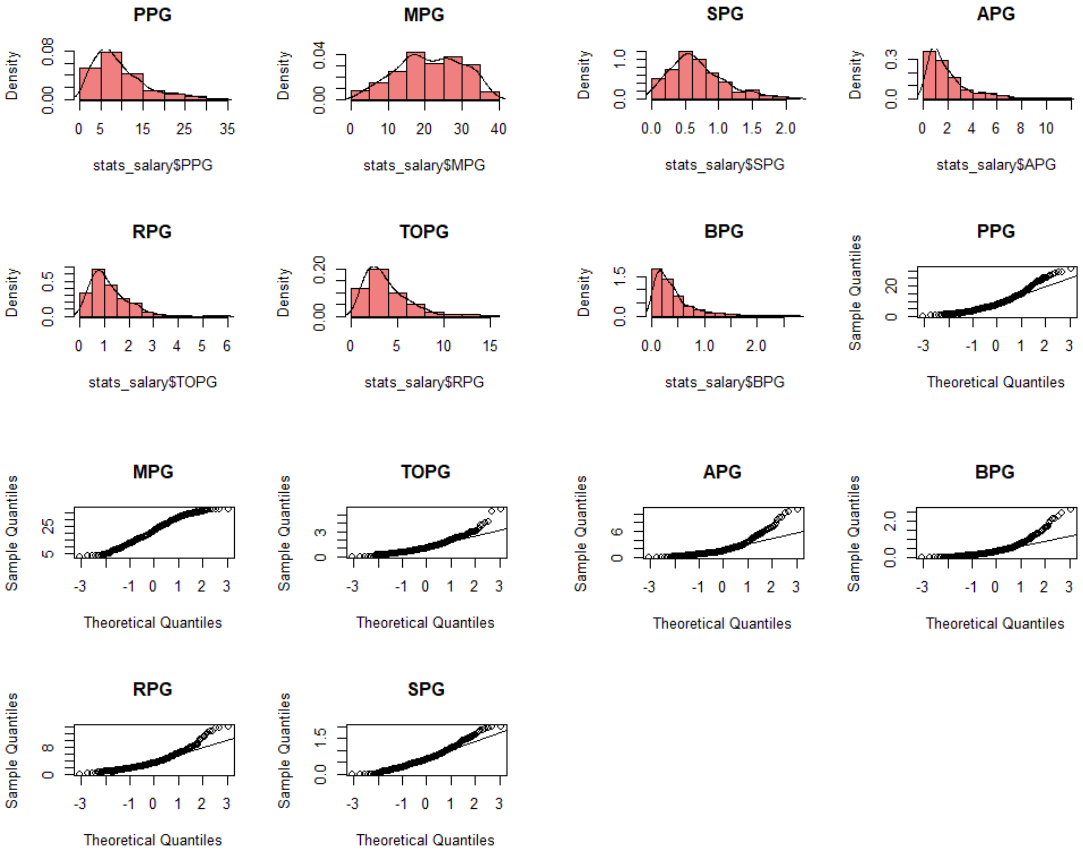
5. Data Visualization

We couldn't put it here because we did an interactive visualization at the beginning of this part.



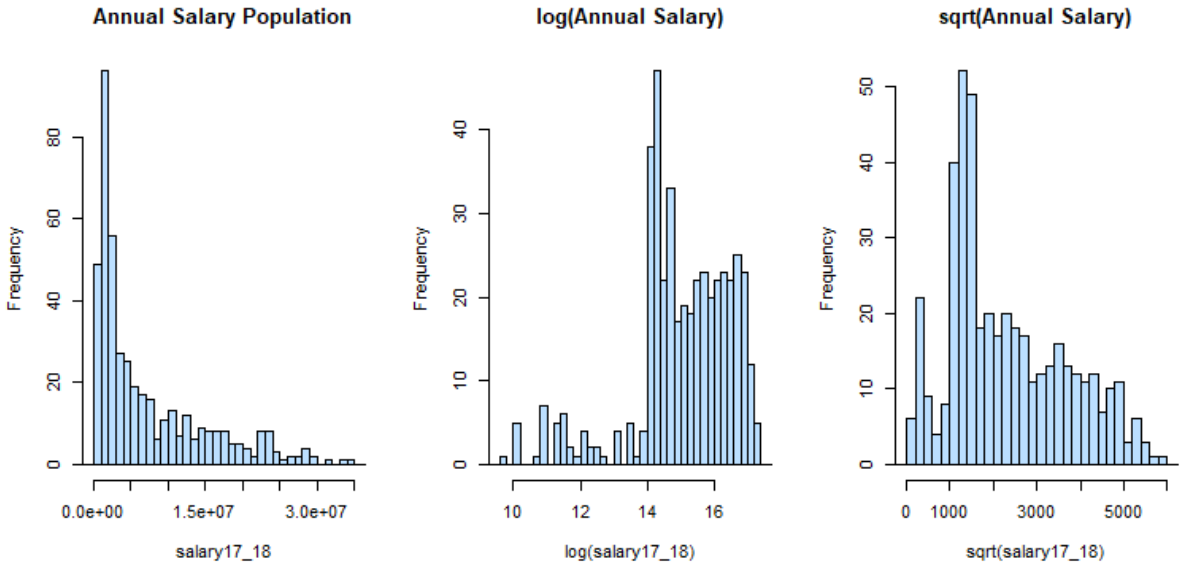
We look to the features independently from the boxplot it can be noticed that the most frequent feature is MPG, as expected APG and BPG represent a minority.

Also we compared the relationship with different independent variables.

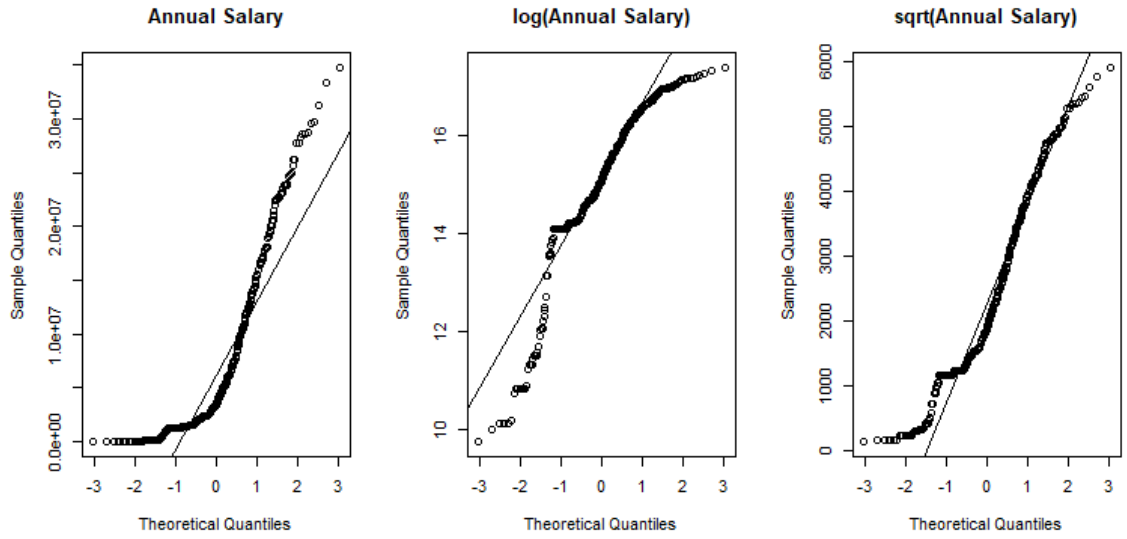


6. Model The Data

We started with Multiple Regression model first.

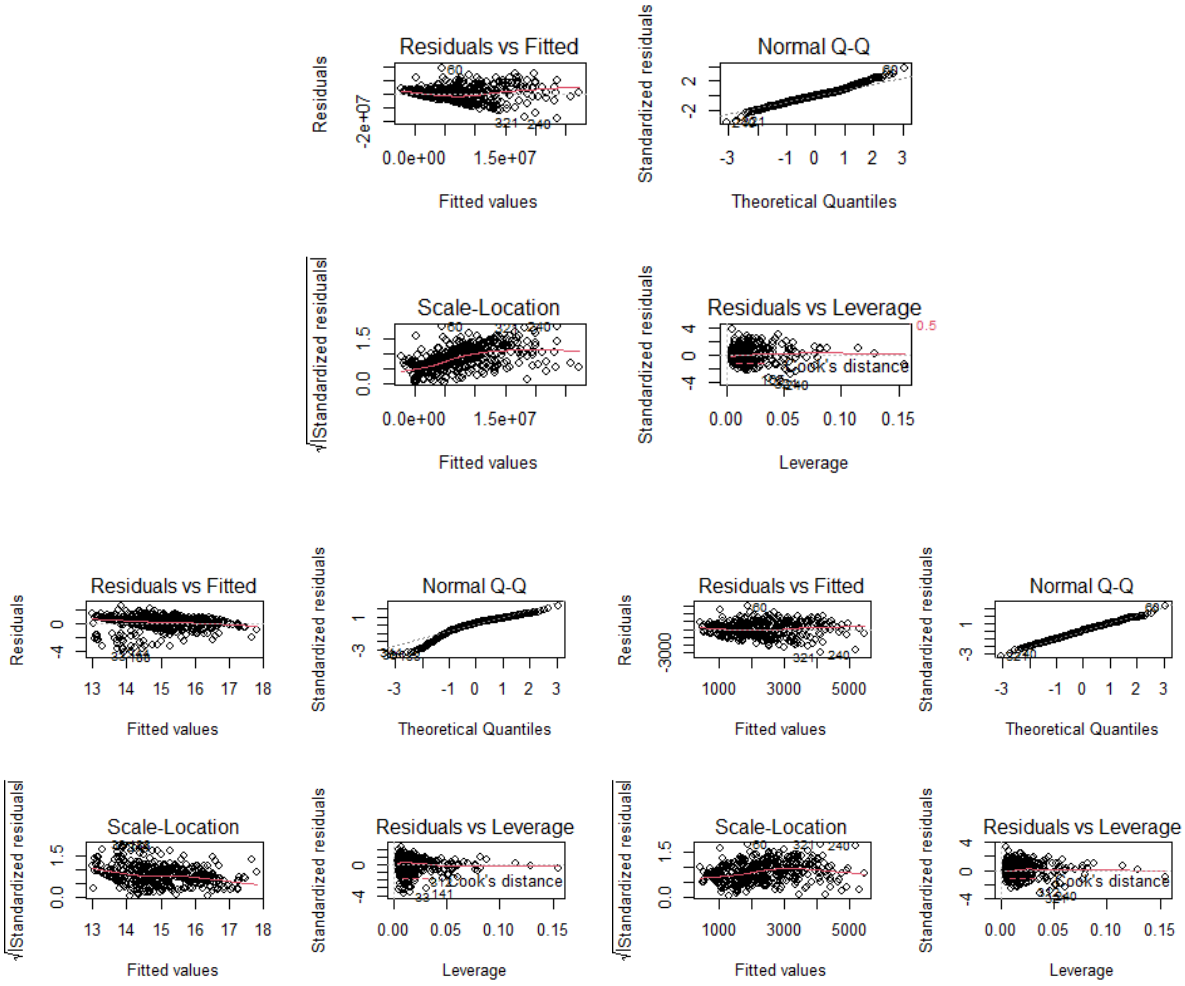


We use Square Root Transform instead of Log Transform, we got slightly better results. First we tried log transform but the results were not satisfactory. Actually, Log transformation is most likely the first thing you should do to remove skewness from the predictor. After that we used squared root transform.



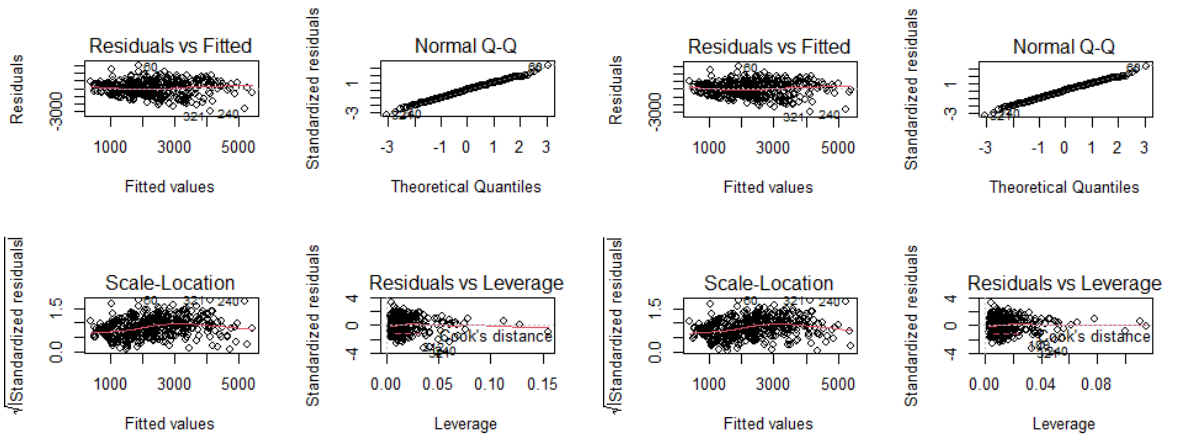
After these short reviews, we set up our model.

Multiple Regression of Most Correlated Predictors with Salary on Different Tests



First model: Linear Model. / Second Model: Converting response variable to log / Third: Converting response variable to sqrt

BACKWARD MODEL SELECTION



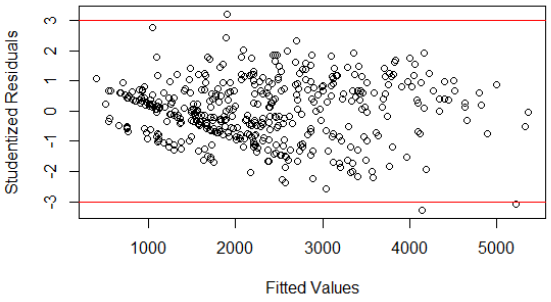
Left model: removing SPG Predictor

Right Model: removing BPG predictor to obtaining BEST MODEL

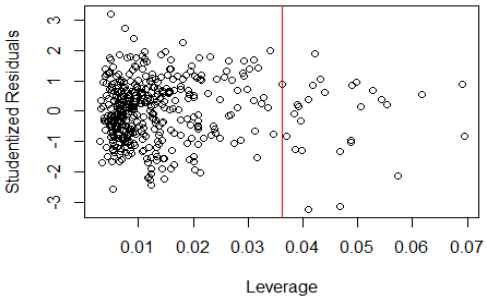
7. Other Possible Problems

- **Outliers:**
  The residual plot identifies some outliers. However, it can be difficult to decide how large a residual need to be before we consider the point to be an outlier.

- **High Leverage Points:**
  A second problem when dealing with regression is the presence of high leverage points. In order to quantify an observations leverage, we compute the leverage statistic. If a given observation has a leverage statistic that greatly exceeds $(p+1)/n$, then we may suspect that the corresponding point has high leverage.

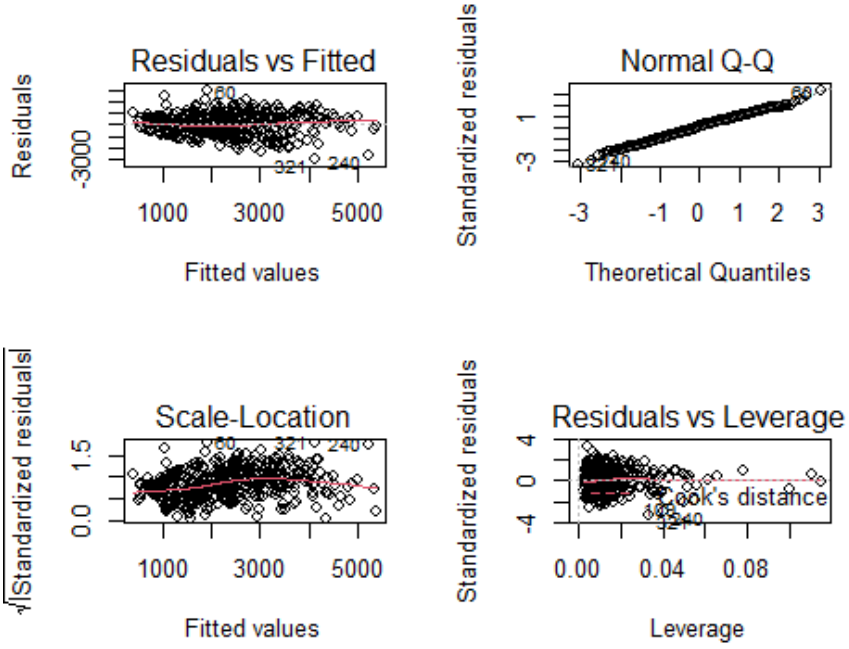


Left model: Outliers (Fitted Values)



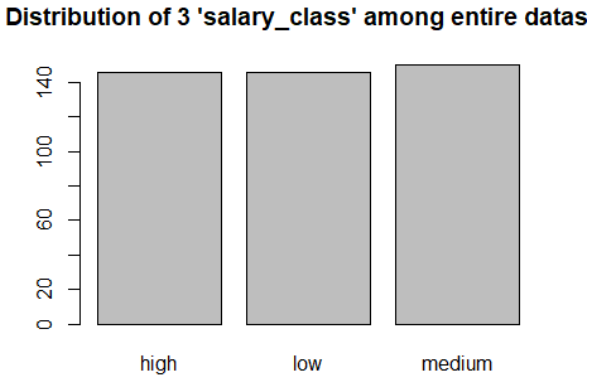
Right Model: High Leverage Points

FINAL REDUCED MODEL



8. COMPARE WITH LDA BUILT-IN FUNCTION

LDA tends to estimate the parameters more efficiently by using more information about the data. Another advantage of LDA is that samples without class labels can be used under the model of LDA. On the other hand, LDA is not robust to gross outliers. Because logistic regression relies on fewer assumptions, it seems to be more robust to the non-Gaussian type of data.




```
          Reference
Prediction high low medium
    high     31   2      6
    low       1  24      5
    medium   20  12     10

Overall Statistics

               Accuracy : 0.5856
                 95% CI : (0.4882, 0.6783)
    No Information Rate : 0.4685
    P-Value [Acc > NIR] : 0.008736

                  Kappa : 0.3827

 Mcnemar's Test P-Value : 0.013132

Statistics by Class:

                     Class: high Class: low Class: medium
Sensitivity               0.5962     0.6316        0.47619
Specificity               0.8644     0.9178        0.64444
Pos Pred Value            0.7949     0.8000        0.23810
Neg Pred Value            0.7083     0.8272        0.84058
Prevalence                0.4685     0.3423        0.18919
Detection Rate            0.2793     0.2162        0.09009
Detection Prevalence      0.3514     0.2703        0.37838
Balanced Accuracy         0.7303     0.7747        0.56032
```

```
          Reference
Prediction high low medium
    high     32   0      7
    low       2  24      4
    medium   11   8     23

Overall Statistics

               Accuracy : 0.7117
                 95% CI : (0.618, 0.7937)
    No Information Rate : 0.4054
    P-Value [Acc > NIR] : 6.435e-11

                  Kappa : 0.5657

 Mcnemar's Test P-Value : 0.2384

Statistics by Class:

                     Class: high Class: low Class: medium
Sensitivity               0.7111     0.7500        0.6765
Specificity               0.8939     0.9241        0.7532
Pos Pred Value            0.8205     0.8000        0.5476
Neg Pred Value            0.8194     0.9012        0.8406
Prevalence                0.4054     0.2883        0.3063
Detection Rate            0.2883     0.2162        0.2072
Detection Prevalence      0.3514     0.2703        0.3784
Balanced Accuracy         0.8025     0.8370        0.7149
```

Due to some disadvantages and shortcomings of the LDA model, we could not get a satisfactory score as a result of this model. However, we think that the accuracy of the LDA model is quite good. We also achieved clearly more efficient results against the QDA model.

9. CONCLUSION

As a result, as we can clearly see, a lot of variables directly affect player salaries. However, two of them caught our attention during this project. The first of these, without doubt, is the time the basketball players take during the match. As the minutes played by the players increase, we can clearly say that their salaries also increase.

For the second pick, our favorite was definitely turnovers. As we mentioned at the beginning of the project, turnovers are closely related to player salaries. Because if a player loses a lot of ball, it means that he plays with the ball a lot. In other words, it means the players who direct the game in the team. In short, we can say that the players with the highest income in the team.

When we combine the effects of these two opposing variables with the number of point statistics of the players, we can see more clearly on the model below how they affect their salaries.

10.REFERENCES

I. An Introduction to Statistical Learning: With Applications in R
II. https://cran.r-project.org/web/packages/ggiraphExtra/vignettes/ggPredict.html
III. https://www.kaggle.com/datasets/drgilermo/nba-players-stats