2K PERFORMANCE

Members: Zheyuan Hu, Shihao Peng, Yang Rong, Suyang Song

UNI: zh2447, sp3905, yr2387, ss6141

Course: IEOR E4523

Term: Fall 2020, Subterm B

Instructor: Uday Menon

1. Background

NBA 2K is a series of basketball sports simulation video games developed and released annually since 1999. Each year's game of this series emulates NBA (National Basketball Association) games. It is extremely popular among basketball fans, and it is one of the most successful and famous sports games in the world. One feature of 2K is that it would update the player rating in newly released games every year based on the performance of players in the previous season.

2. Introduction

We conducted text mining on comments about 2K rating from Twitter and displayed several keywords using a word cloud [Appendix 1]. As the graph indicated, most people have positive feedback such as good, best, and reasonable regarding the 2K rating. While some, on the other hand, showed opposite opinions like arguable, disrespectful, and overrating. It triggered our curiosity about how 2K came up with the rating.

For this project, we anticipated building a statistical connection between player statistics and their corresponding 2K rating using data analytics and machine learning. We took the initiative to clean data first preparing the dataset for processing, and then eliminated features with the least importance. Second, we constructed machine learning models including linear regression, KNN, Random Forest and Decision Tree, and compared pros and cons to determine the best fitting model.

3. Data Collection and Preprocessing

The latest version (NBA 2k21 and NBA statistics for season 2020) is the dataset applied for this project. All data were collected using web scraping from the following two websites:

1) https://hoopshype.com/nba2k
2) https://basketball.realgm.com/nba/stats/2020?Averages/Qualified/points/All/desc/1/Regular_Season

All the 'Nan' values were deleted and the redundant columns ('Unnamed') were dropped before applying a more thorough feature selection process. Also, some star players' data such as Michael Jordan and Kobe Byrant were dropped since they have extremely high ratings, which can be treated as outliers for the dataset. In total, the data set left with twenty-two attributes (FGM, FGA,FG%,3PM,3PA,3P%FTM,FTA,FT%,GP, MPG,TOV, PF,ORB,DRB,RPG,APG,SPG,BPG,PPG). They can be divided into four different collection: Time (GP, MPG); Scores and Hit rate (FGM, FGA, FG%, 3PM, 3PA, 3P%, FTM,FTA,FT%), Contribution (ORB, DRB,PRG,APG,SPG,BPG,PPG), and Mistakes (TOV, PF).

Since there exist regular season and playoff seasons in each year, all the values were averaged for each attribute for players who played both the regular season and playoff season. Groupby function was used for generating a new averaged data frame based on player names and their teams.

4. Feature Selection

Twenty-two input variables are included in the dataset. Heatmap of correlation of player statistics and output variable rating is performed to identify variables which are statistically significant and variables which can be excluded from the analysis. According to Fig 1, the top six correlated variables are PPG, FGM, FTA, FTM, FGA, and TOV, with a correlation larger than 0.8. FT%, 3P%, and GP have a less than 0.2 correlation with the rating variable. Thus, these three variables were removed from the dataset.

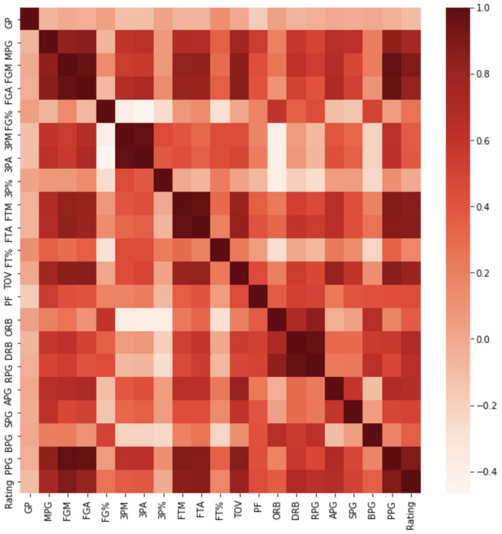
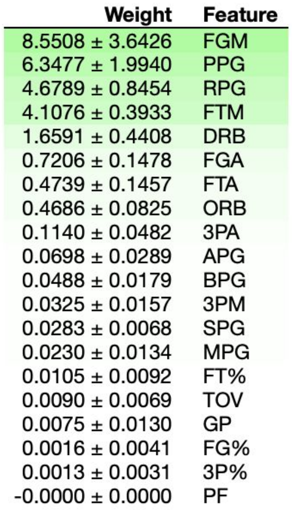


Fig. 1 Heatmap of Correlation Between All Variables

In order to confirm our decision about removing the variables, we also performed the permutation feature importance of linear regression and random forest regression with all 22 variables. The results of the important tests are shown in Fig 2 and 3 and are consistent with the heatmap.



Weight	Feature
8.5508 ± 3.6426	FGM
6.3477 ± 1.9940	PPG
4.6789 ± 0.8454	RPG
4.1076 ± 0.3933	FTM
1.6591 ± 0.4408	DRB
0.7206 ± 0.1478	FGA
0.4739 ± 0.1457	FTA
0.4686 ± 0.0825	ORB
0.1140 ± 0.0482	3PA
0.0698 ± 0.0289	APG
0.0488 ± 0.0179	BPG
0.0325 ± 0.0157	3PM
0.0283 ± 0.0068	SPG
0.0230 ± 0.0134	MPG
0.0105 ± 0.0092	FT%
0.0090 ± 0.0069	TOV
0.0075 ± 0.0130	GP
0.0016 ± 0.0041	FG%
0.0013 ± 0.0031	3P%
-0.0000 ± 0.0000	PF

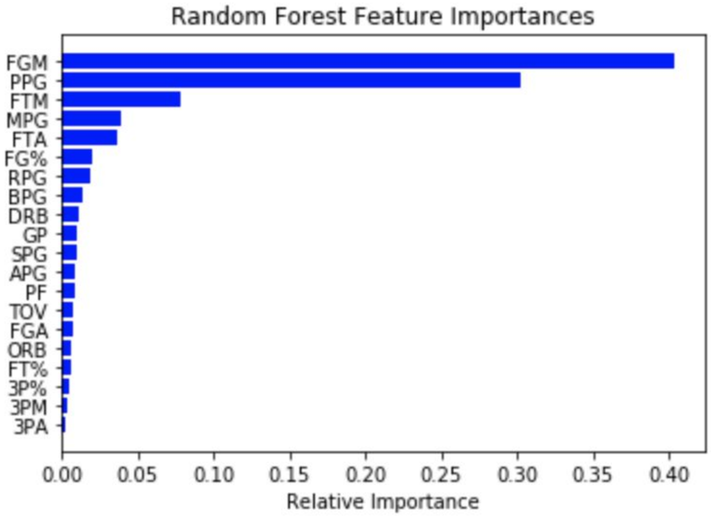


Fig 2. Linear Feature Importances

Fig 3. Random Forest Feature Importances

5. Machine Learning Models

Four different regression models were applied to explore the relationship between all the features and the final ratings of NBA players. All the results for models are based on the train test split size of 80:20, which gives 80% of the total data for training purpose and rest for testing purpose.

1) Linear Regression

Linear Regression model fits a linear model with a constant-coefficient to minimize the residual sum of squares between actual and predicted value. This model gives the accuracy score of 92.66% and a mean square error of 2.37, which implies the data set fits well for a linear approximation. By looking at Fig 4, there are very few outliers. All data points are well distributed.

2) K Nearest Neighbor (KNN) Regressor

In the KNN regression model, two different weight options were tested with the same number of neighbours of 5. Uniform weights give the score of 89.69%, mean square error of 3.32 and R square of 0.9. Distance weights provide slightly better results. Its score is 81.26% and the mean square error is 6.04. R square remains the same. From the comparison, the KNN model weighted by distance is a better fit for the 2K dataset instead of weighing each point equally.

3) Random Forest Regressor

Random Forest is applied with n_estimators of 100, a criterion of mse, and no limitation on max_depth. This test is reliable as it scores 89.13%. Its mean square error is 3.51 and R square is 0.89.

4) Decision Tree Regressor

The result of the Decision Tree model is surprisingly good. The score is 92.66% mean square error is 8.48, and R square is 0.74. By looking at Fig 7, compared to the result of Linear Regression, there might exist the problem of overfitting in this dataset.

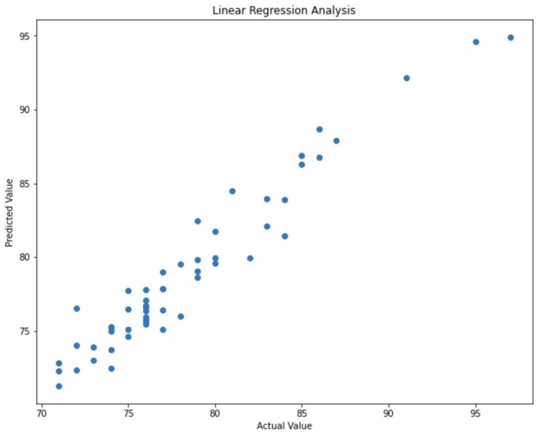


Fig. 4 Linear Regression Analysis

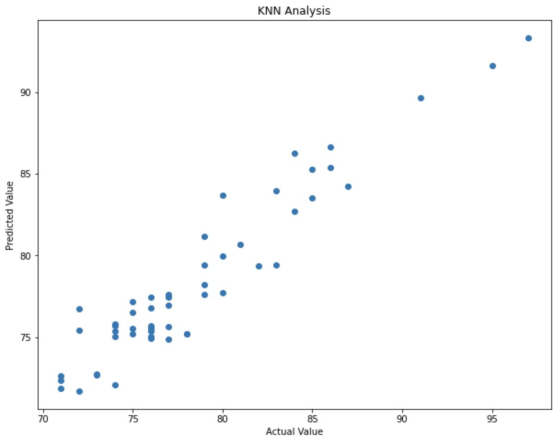


Fig. 5 KNN Analysis

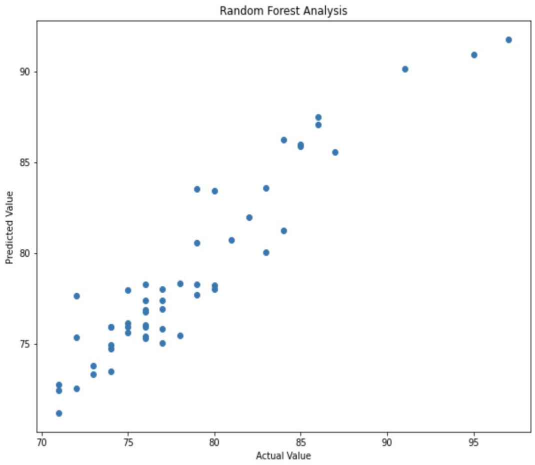


Fig. 6 Random forest Analysis

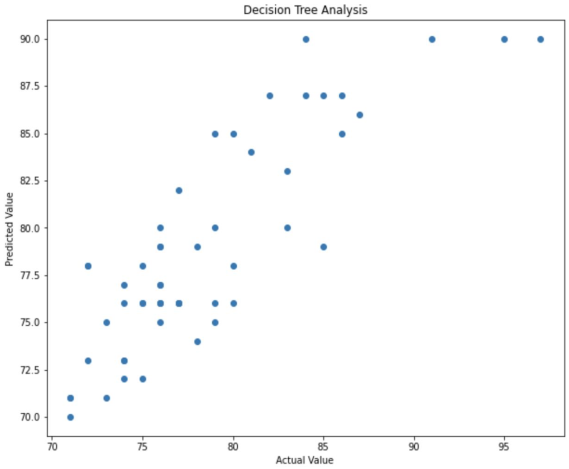


Fig. 7 Decision Tree Analysis

6. Prediction

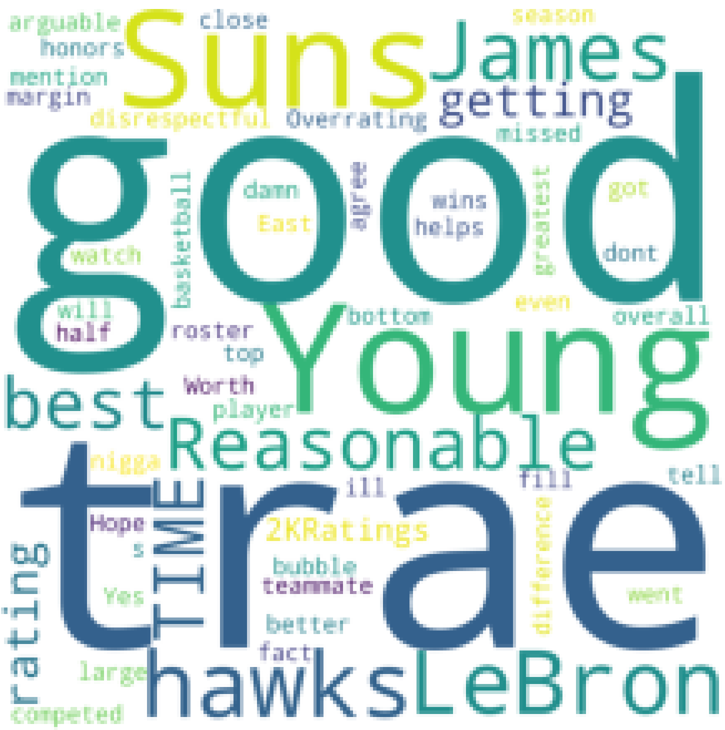
If given the collected player statistics of the new season next year (2020-2021), the model will theoretically predict the 2K player rating of NBA 2K22 with accuracy above 90%. Having the predicted rating earlier before 2K releases can give gamers a rough idea of what to expect in the new NBA 2K game. In addition, we are able to see directly how the performances of players change in the new season by visualizing the fluctuations of ratings throughout the years.

7. Conclusion

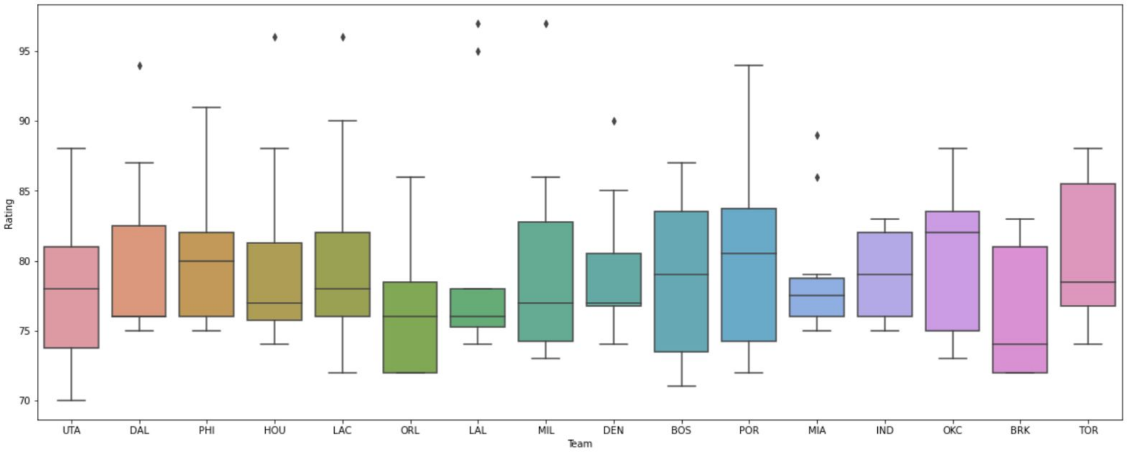
Among the machine learning algorithms we tested, linear regression has the best performance, with a 92.66% accuracy score. The corresponding coefficients of the linear regression model are shown in [Appendix 8]. However, there are still some improvements that can be done to enhance the performance of this model. First, the sample data size is not quite large as we only used one season for the analysis. For future improvement, we can include datasets from multiple seasons. Second, during the data processing, we assumed the same importance of regular and playoff games data while combining and averaging them. But in reality, their importance ratio might be ever-changing from season to season. Thus, testing more combinations of importance ratio may advance the model. Last, there are other potential input variables such as player age, player height, player weight, the injury histories and player positions. These variables in fact will affect the performance in real life for an athlete, thus having a large influence on the rating in the game. Classifying and quantifying these factors and taking them into consideration will be a good direction for further research and enhancement.

Appendix

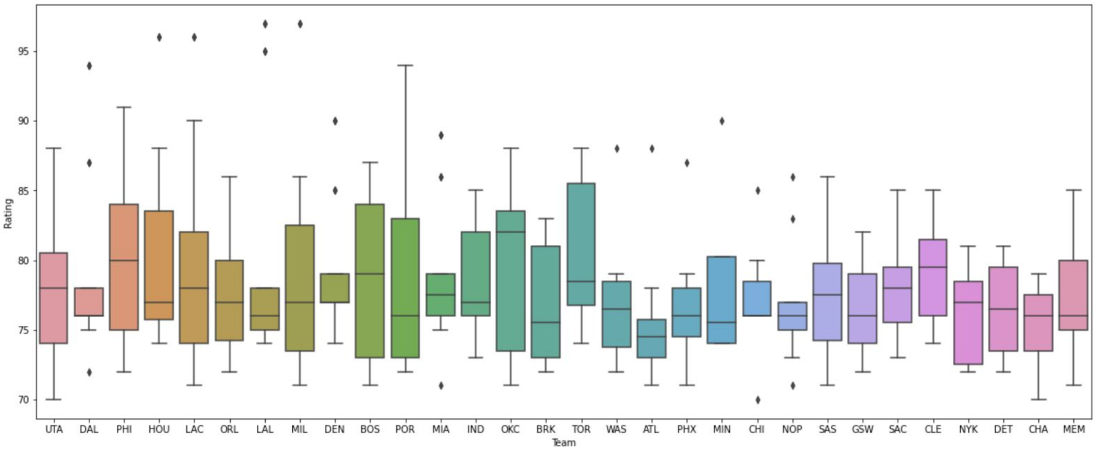
[Appendix 1] 2K Twitter Text Mining



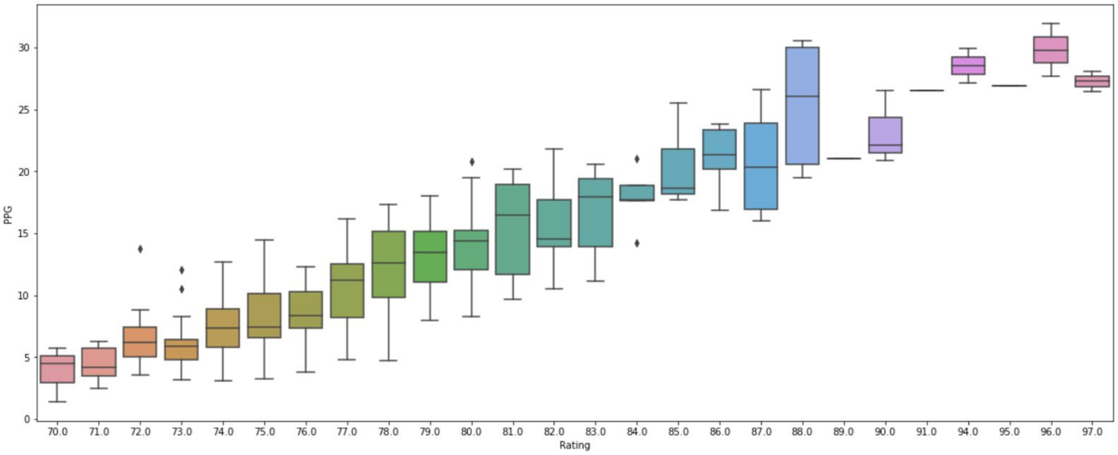
[Appendix 2] Boxplot of Player Rating in Teams which Got into Playoff



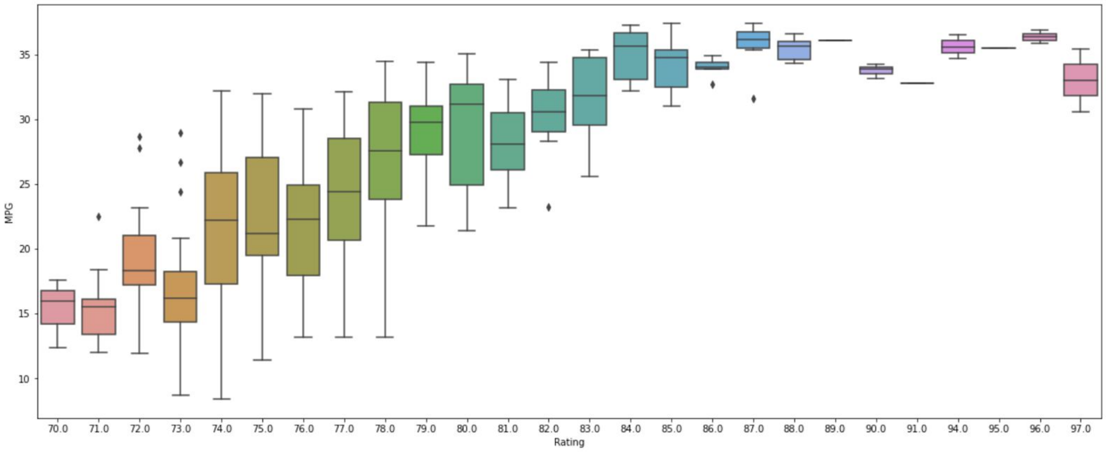
[Appendix 3] Boxplot of Player Rating in teams Which Got Into Playoff



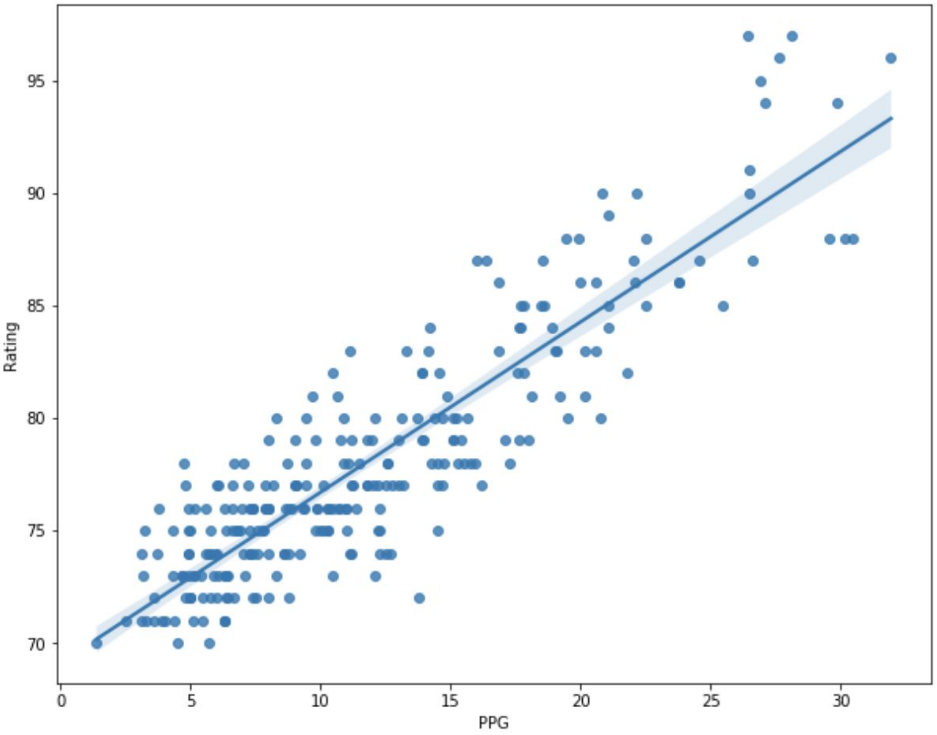
[Appendix 4] Boxplot of Corresponding PPG of Each Rating



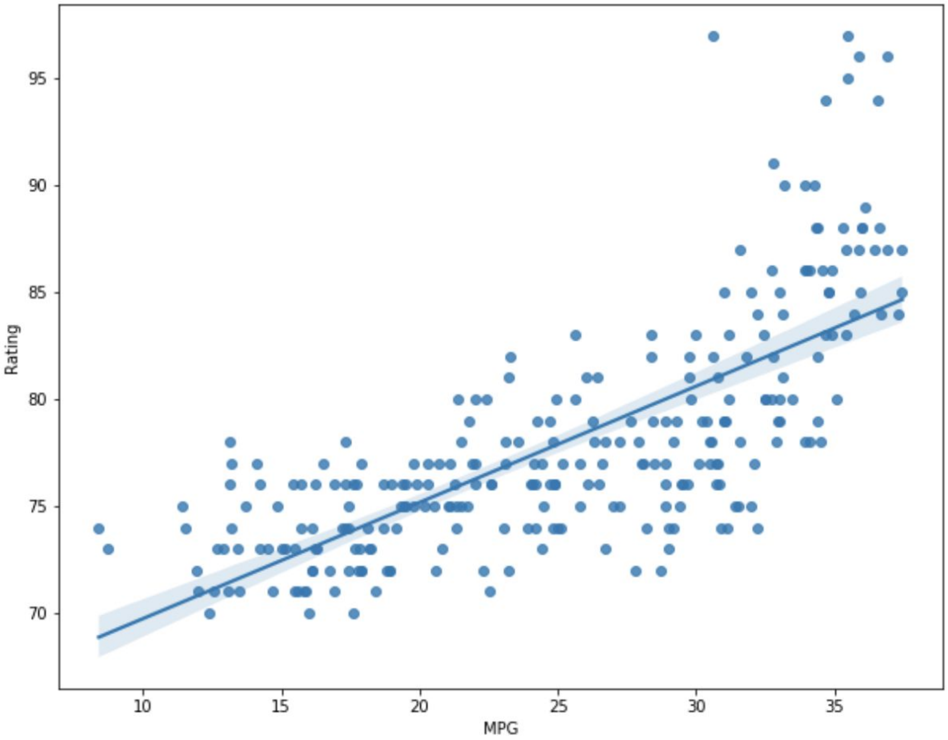
[Appendix 5] Boxplot of Corresponding MPG of Each Rating



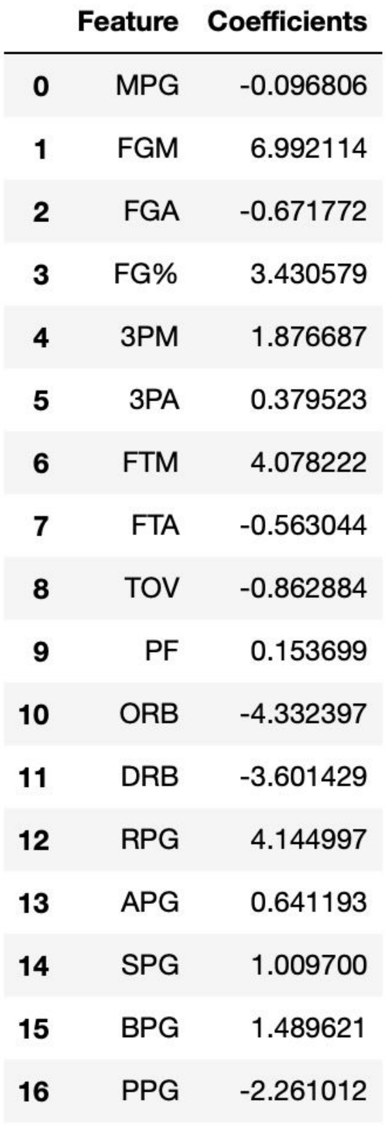
[Appendix 6] Linear Regression Model Fit of Rating versus PPG



[Appendix 7] Linear Regression Model Fit of Rating versus MPG

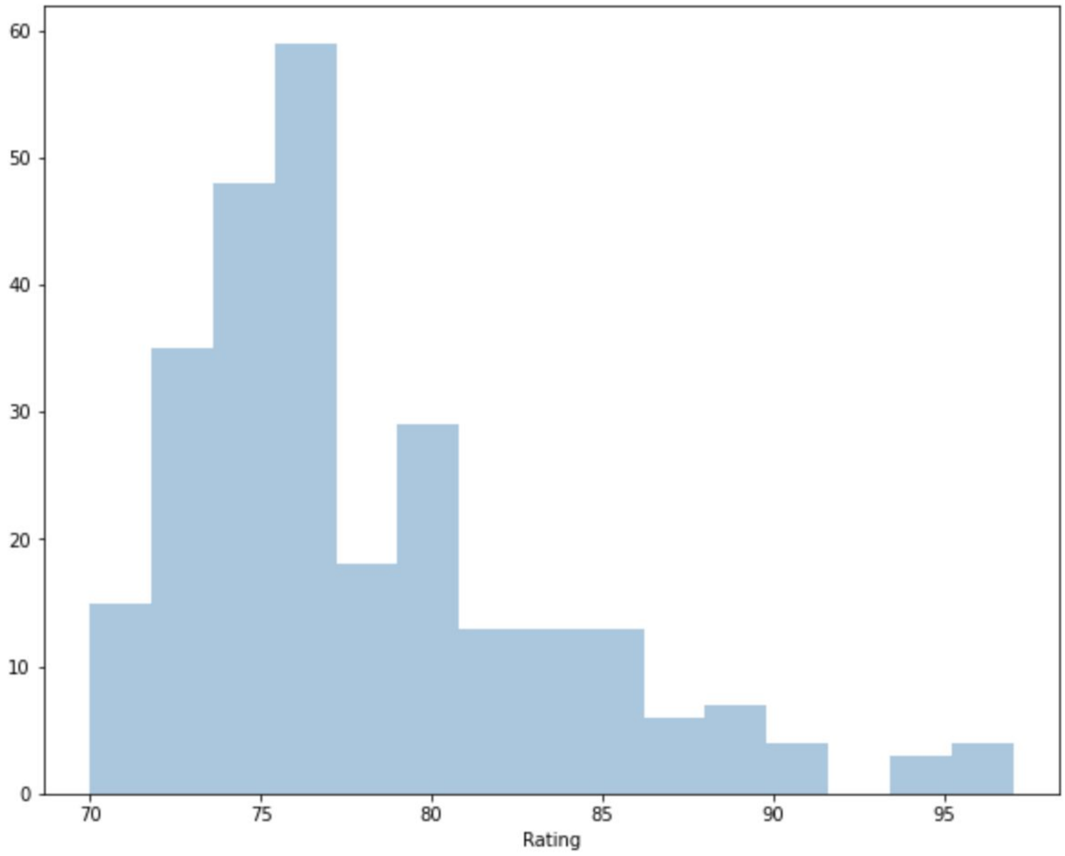


[Appendix 8] Linear Regression Coefficients



	Feature	Coefficients
0	MPG	-0.096806
1	FGM	6.992114
2	FGA	-0.671772
3	FG%	3.430579
4	3PM	1.876687
5	3PA	0.379523
6	FTM	4.078222
7	FTA	-0.563044
8	TOV	-0.862884
9	PF	0.153699
10	ORB	-4.332397
11	DRB	-3.601429
12	RPG	4.144997
13	APG	0.641193
14	SPG	1.009700
15	BPG	1.489621
16	PPG	-2.261012

[Appendix 9] Rating Distribution Histogram


Rating