STSCI 2120/5120 Final Project – Group 5

By: Robert Horn, Winson Dong, Lingyu Zhou, Haoyang Wang and Heyang Dong

Introduction

The data set that we will be analyzing is the NYM Batting data set that takes into account the batting statistics of all New York Mets players from the franchise's creation in 1962 until the end of the 2023 season. This data set refers to many different batting and offensive statistics that help evaluate players. This data set has accrued data from three main categories, demographic statistics, counting statistics, and averaged statistics. Demographic statistics in this data set include a player's dominant hand, whether they are a switch hitter or not, their age and position that they play. The counting statistics are statistics that increase only when a players achieves the feat of what the statistics represents such as Total Bases, Strikeout, Bases on Balls, Runs Scored, Home Runs, Doubles, Triples, Hits, Runs Scored, At Bats, Plate Appearances and Games Played. Finally, the last category is referred to as averaged statistics which use a formula of certain of the counting statistics to get an average that is comparable across all players regardless of how many games they have played during the season.

The topic that we will be analyzing is a player's power which will be analyzed using a player's slugging percentage. Slugging percentage is an average statistics that is calculated by dividing the Total Bases that a player has accrued divided by the total number of at bats they have accrued. This is the best statistic for deriving a players power because it is an average statistic which does not help or penalize any player based on the number of games played or at bats accrued. Additionally, this weights each hit a player gets by its usage of total bases in the numerator, by explaining that a player who has more power should be getting more extra base hits like doubles, triples and home runs worth 2, 3, and 4 bases respectively than singles which are only worth 1 base.

Throughout this process we have been researching the question, what factors affect a players slugging percentage for players on the New York Mets? Our hypothesis is that these demographic factors will significantly affect performance variables, specifically the slugging percentage of a player.

Data Preparation

After the data is loaded in R, we first examine the entire dataset for completeness and consistency. Using colSums(is.na(df)), we found the only NaN values that happen in column “Position”, is a categorical variable that denotes the position of the player without any inherent numerical relationship. Predicting or imputing missing values for such categorical data can introduce inaccuracies and potential biases, as there is no reliable method to infer the missing categories without additional contextual information. Therefore, to preserve the integrity and quality of our analysis, we decided to remove all rows where the "Position" value is missing. This approach ensures the integrity and accuracy of the subsequent analyses and modeling.

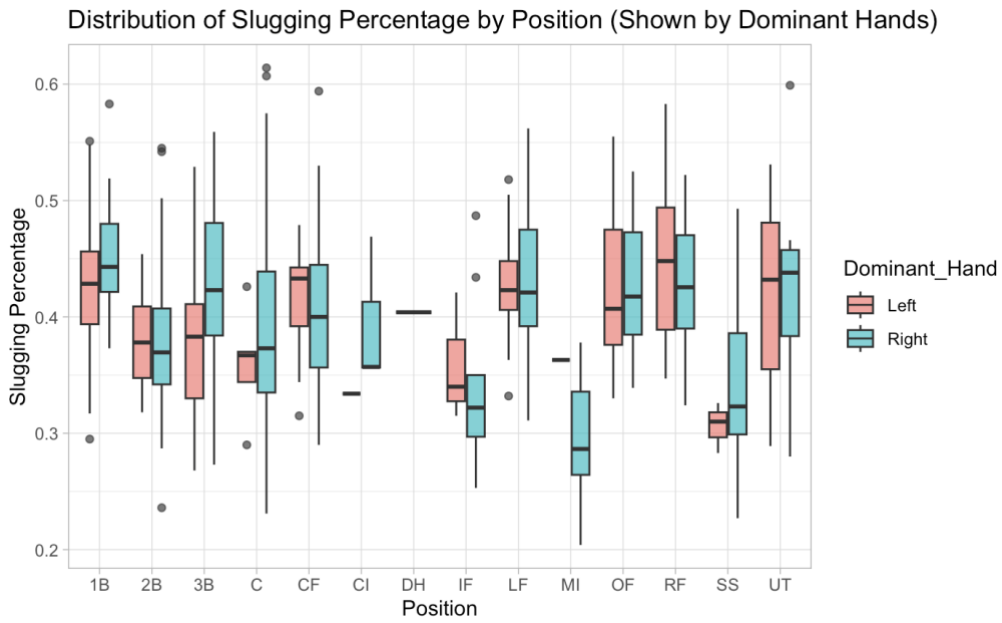
Moving to the next step, we consider other data cleaning steps to ensure data quality, which include checking for and removing duplicate records that may skew analysis, verifying that numeric variables are within realistic and expected ranges, and handling outliers if they are deemed to be data entry errors or anomalies that could adversely affect model performance. We also ensured that categorical variables have consistent levels, such as checking for misspellings or inconsistencies in factor levels (e.g., "Left-Handed" vs. "Left Handed"). Additionally, we standardize the formats of date and text fields to maintain uniformity across the dataset. These preliminary steps are crucial for maintaining data integrity before proceeding with further analysis.

After cleaning the data, we began to do some transformations. We created a new response variable, "Magnitude_Hit," which was derived as the ratio of the "Batting_Average" to "Slugging_Percentage." This variable was intended to provide a more detailed metric for analyzing a player's batting performance. We also created a categorical variable called "Competitiveness," which classified players as either "Competitive" or "Less Competitive" based on their ranking. More specifically, players ranked between 1 and 23 were labeled as "Competitive," while the rest were classified as "Less Competitive." This categorization helped examine the influence of various player attributes on competitiveness levels.

Next, several categorical variables, including "Position," "Dominant_Hand," and "Switch_Hitter," were converted to factors using the ‘as.factors()’ function. This transformation ensured the appropriate recognition of categorical data in subsequent modeling and analysis stages, which facilitates statistical processing and proper handling in regression models. We also excluded certain variables, such as "Name," "Batting_Average," and "Slugging_Percentage," from further analysis to avoid potential issues with multicollinearity and to focus on other predictor variables. After selecting the relevant subset of variables, the data is then ready for exploratory data analysis to understand the distribution of the "Competitiveness" class. Utilizing the ggplot2 package, we can visualize the frequency distribution of the "Competitiveness" classes, providing insight into the balance of our data.

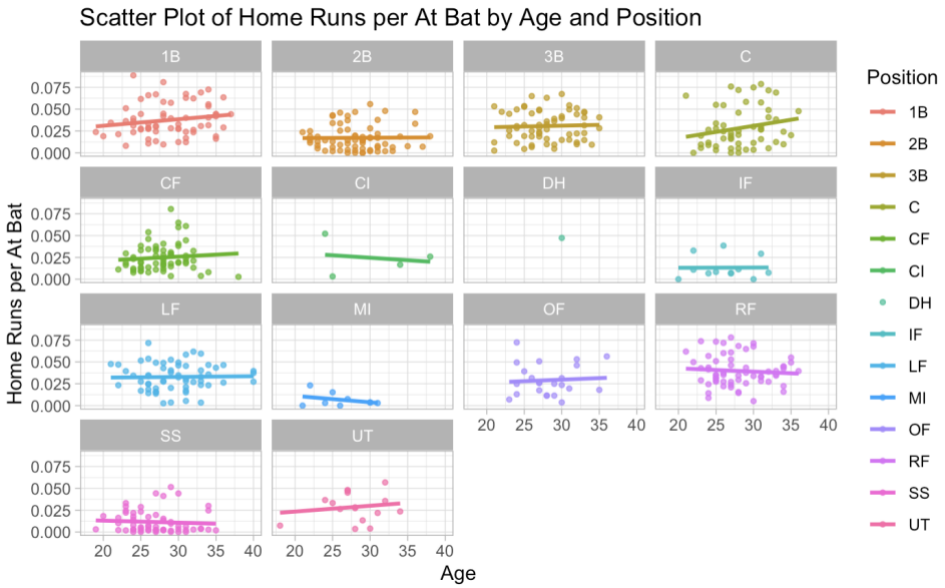
Exploratory Data Analysis

1. Box Plots that showed the distribution of Slugging Percentage with Position by whether the players were batting left-handed or right-handed.



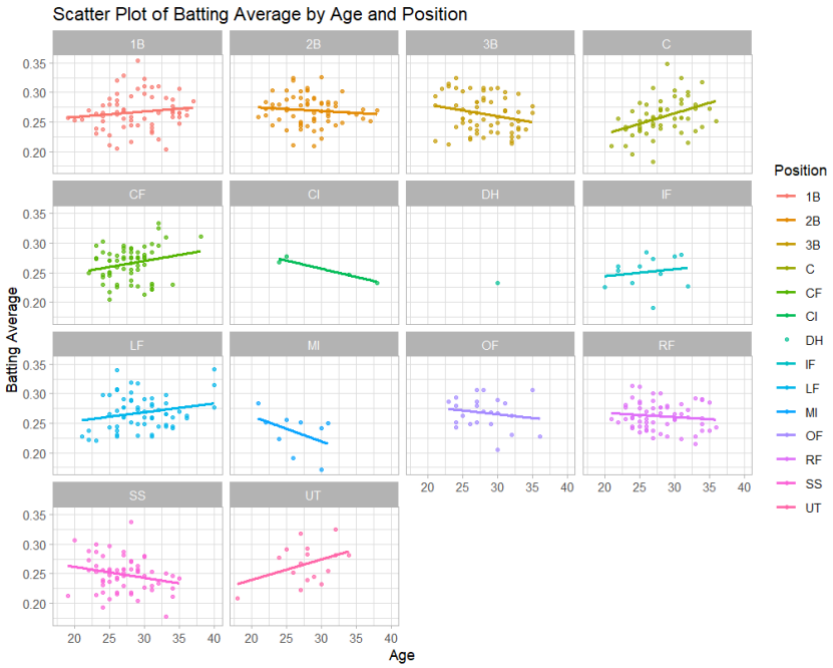
This visualization would help identify any trends in slugging percentage associated with handedness and position. We see that most of the positions have similar median slugging percentages for left and right-handed players. For players that play the 3B and C position, the data shows that the middle 50% of right-handed players have higher slugging percentages than the middle 50% of left-handed players. Additionally, for players that play the C positions, the data for right-handed players is much more spread out with an IQR of about 3 times larger than the left-handed players IQR. This demonstrates that for certain positions, a players' dominant hand can affect their slugging percentage. Additionally, certain positions that are less fielding intensive such as 1B, OF, RF, UT, have higher slugging percentages than players of other positions which are seen as harder to field such as 2B, C, MI, and SS.

2. Scatter Plot of HomeRun Rate and Age by Position



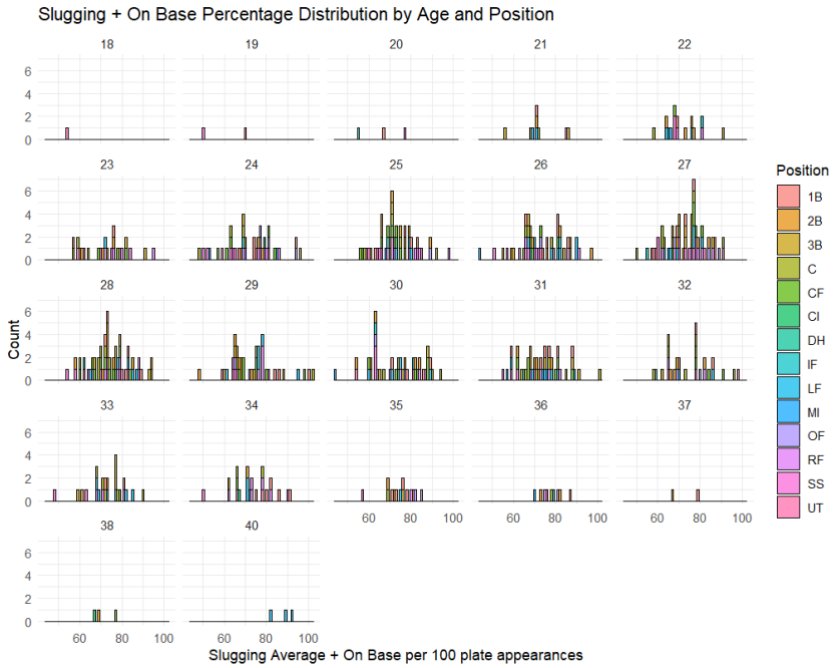
This visualization would help identify any trends in home runs per at bat or home run percentage associated with age. By faceting by position, position wouldn't be a factor in telling whether age affects the home run percentage. Home runs are a counting statistic that represents the most powerful hit a baseball player can obtain being worth 4 bases which differs from slugging percentage as it takes into account and weights every hit a player has by their number of bases attributed to their hit (Single = 1, Double = 2, Triple = 3, Home Run = 4). While slugging percentage is the best power metric, we chose to include home run percentage to see in which percentage of a player's at bats they will achieve the ultimate goal of a home run. For players that play the 1B, CF and C positions, as a player's age increases, so does their home run percentage. Contrastly, for the SS position which is a much more demanding fielding position, players have lower home run percentages on average and decrease as they get older showing how age and position can both impact a players power through home run percentage.

3. Scatter Plot of batting average categorized by age and position



This visualization would help identify any trends in batting average associated with age and position. By faceting by position, position wouldn't be a factor in telling whether age affects the batting average. Batting Average is a batting statistic that is used to determine the percentage of at bats in which a player will get a hit as opposed to recording an out. This statistic is not fully a power statistic but provides a baseline for us to compare their slugging percentages against which should be higher given that each hit is treated the same regardless of the number of bases it is worth. This visualization demonstrates that most positions are very similar in their batting averages hovering between 0.25 and 0.3 showing between a 25-30% that a player obtains a hit each time they have an at bat.

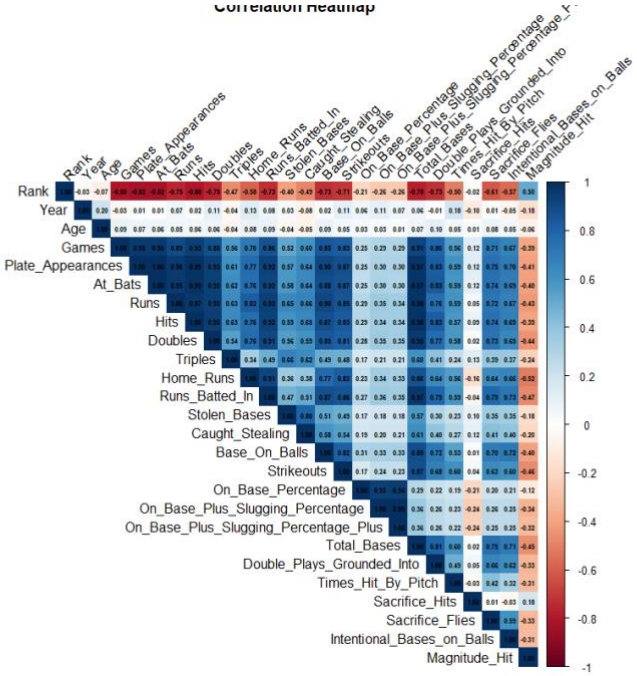
4. Histogram of the Average OnBase+Slugging Percentages (A power statistics that takes into account a player's ability to get on base as well as their power) by Age + Position



This visualization would help identify any trends in age and slugging + on base percentage. By faceting by age we can compare the differences in the percentages throughout the ages to see if there is a significant difference. The plot demonstrates that age can be a factor that affects performance variables, specifically showing that age 28 has the highest number of baseball players, is the median age, and is very close to the mean age of 28.13. This illustrates that the closer a player is to their age 28 season the more likely they will be a major league baseball player due to their age 28 season being thought of as the season where a player gives his prime performance.

Statistical Analysis

The first step in statistical analysis is to create multiple line graphs based on the position that shows age on the x-axis and Slugging Percentage / Batting Average on y-axis to see if there is a linear correlation between age and magnitude of hit.



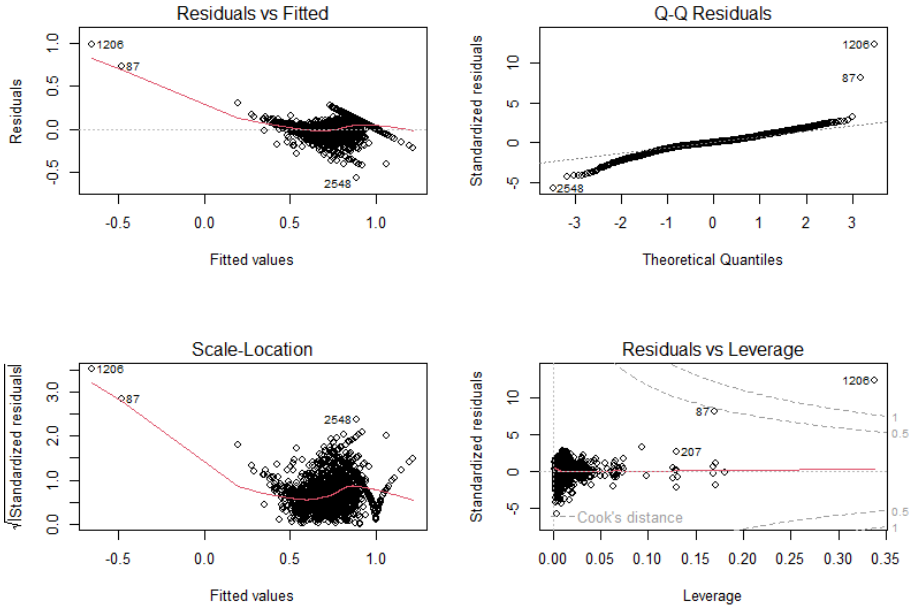
Notice that for Magnitude_Hit (which is the ratio of Batting_Average over Slugging_Percentage), which has a medium level of correlation coefficients with most variables. What's more, the response is negatively correlated with these variables. The highest correlation coefficient it has is with Rank, which is of 0.5.

Next we fit a full model by linear regression. From the summary output we can see that the majority of the variables are significant (by Wald Test), and since we have a p-value of much less than 0.05 in F-test, whole model adequacy is justified. The R-squared value of 0.6596 indicates that about 65.96% of the response can be explained by this model.

We then move onto model selection to refine the model, which is done by backward stepwise selection. Under AIC criteria, Rank, Year, Position, Age, Games, Plate_Appearances, At_Bats, Hits, Doubles, Triples, Home_Runs, On_Base_Percentage, On_Base_Plus_Slugging_Percentage, On_Base_Plus_Slugging_Percentage_Plus, and Sacrifice_Files are considered to be the significant variables, which are the factors that impact players' magnitude of each hit.

```
Call:
lm(formula = Magnitude_Hit ~ Rank + Year + Position + Age + Games +
    Plate_Appearances + At_Bats + Hits + Doubles + Triples +
    Home_Runs + On_Base_Percentage + On_Base_Plus_Slugging_Percentage +
    On_Base_Plus_Slugging_Percentage_Plus + Sacrifice_Flies,
    data = NYM_batting_subset)

Coefficients:
                                (Intercept)                                        Rank                                        Year
                                  1.4965330                                   0.0060380                                  -0.0004077
                                 Position1B                                  Position2B                                  Position3B
                                 -0.1088751                                  -0.0919592                                  -0.1053541
                                  PositionC                                  PositionCF                                  PositionCI
                                 -0.0901962                                  -0.0958694                                  -0.1383440
                                 PositionDH                                  PositionIF                                  PositionLF
                                 -0.1667524                                  -0.1026927                                  -0.1127132
                                 PositionMI                                  PositionOF                                   PositionP
                                 -0.0578121                                  -0.1152016                                  -0.0526002
                                 PositionRF                                  PositionSS                                  PositionUT
                                 -0.1158623                                  -0.0701499                                  -0.1235294
                                        Age                                       Games                           Plate_Appearances
                                 -0.0009136                                  -0.0003356                                  -0.0014112
                                     At_Bats                                        Hits                                     Doubles
                                  0.0012464                                   0.0036506                                  -0.0063274
                                    Triples                                   Home_Runs                            On_Base_Percentage
                                 -0.0116744                                  -0.0053101                                   1.9267430
      On_Base_Plus_Slugging_Percentage  On_Base_Plus_Slugging_Percentage_Plus                                Sacrifice_Flies
                                 -0.7089923                                  -0.0006787                                   0.0048272
```

Further Looking at the diagnostic plot of the refined model, from the Residuals vs. Fitted plot, it seems neither linearity assumption nor constant variance assumption is violated; from QQ plot, since most of the observations fit on the reference line, normality assumption is not violated; from Scale-Location plot, we can see observation 1206, observation 87 are of high leverages; from Residuals vs. Leverage plot, we can see observation 1206 is an outlier.



Now we fit the model based on selected predictors along with dataset that excludes the outlier, i.e., observation 1206, where from the summary we can see R-Squared value is increased by 3% without loss of testing (both t-test and F-test) significance.

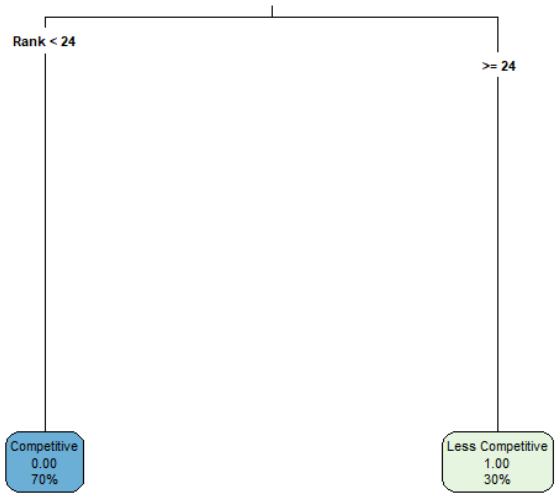
```
> summary(NYM_batting_mod_best)

Call:
lm(formula = Magnitude_Hit ~ Rank + Year + Position + Age + Games +
    Plate_Appearances + At_Bats + Hits + Doubles + Triples +
    Home_Runs + On_Base_Percentage + On_Base_Plus_Slugging_Percentage +
    On_Base_Plus_Slugging_Percentage_Plus + Sacrifice_Flies,
    data = NYM_batting_subset[-1206, ])

Residuals:
     Min       1Q   Median       3Q      Max
-0.55520 -0.04070  0.00310  0.04796  0.97684

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                          1.2825570  0.2661392   4.819 1.55e-06 ***
Rank                                 0.0054214  0.0006482   8.363  < 2e-16 ***
Year                                -0.0002126  0.0001386  -1.534 0.125251
Position1B                          -0.1358993  0.0403567  -3.367 0.000774 ***
Position2B                          -0.1182189  0.0403343  -2.931 0.003419 **
Position3B                          -0.1324229  0.0405005  -3.270 0.001096 **
PositionC                           -0.1194972  0.0397450  -3.007 0.002676 **
PositionCF                          -0.1192821  0.0403784  -2.954 0.003174 **
PositionCI                          -0.1695544  0.0458405  -3.699 0.000223 ***
PositionDH                          -0.1969915  0.0517577  -3.806 0.000146 ***
PositionIF                          -0.1331954  0.0410238  -3.247 0.001187 **
PositionLF                          -0.1385741  0.0404372  -3.427 0.000623 ***
PositionMI                          -0.0870045  0.0422869  -2.057 0.039774 *
PositionOF                          -0.1412590  0.0399940  -3.532 0.000422 ***
PositionP                           -0.0807723  0.0402003  -2.009 0.044650 *
PositionRF                          -0.1415787  0.0403116  -3.512 0.000455 ***
PositionSS                          -0.0938505  0.0405171  -2.316 0.020645 *
PositionUT                          -0.1486497  0.0404696  -3.673 0.000246 ***
Age                                 -0.0007144  0.0005164  -1.383 0.166702
Games                               -0.0002451  0.0001613  -1.519 0.128830
Plate_Appearances                   -0.0017572  0.0002419  -7.265 5.38e-13 ***
At_Bats                              0.0014025  0.0002855   4.912 9.78e-07 ***
Hits                                 0.0040534  0.0003703  10.946  < 2e-16 ***
Doubles                             -0.0055335  0.0007730  -7.159 1.15e-12 ***
Triples                             -0.0106666  0.0016968  -6.286 4.01e-10 ***
Home_Runs                           -0.0036136  0.0005801  -6.229 5.74e-10 ***
On_Base_Percentage                   2.2036245  0.0659454  33.416  < 2e-16 ***
On_Base_Plus_Slugging_Percentage    -1.2115805  0.0939123 -12.901  < 2e-16 ***
On_Base_Plus_Slugging_Percentage_Plus  0.0005801  0.0003506   1.655 0.098158 .
Sacrifice_Flies                      0.0050977  0.0016847   3.026 0.002512 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09417 on 1928 degrees of freedom
  (因为不存在，769个观察量被删除了)
Multiple R-squared:  0.6848,    Adjusted R-squared:  0.6801
F-statistic: 144.4 on 29 and 1928 DF,  p-value: < 2.2e-16
```

Secondly we construct a model that evaluates a player's performance using logistic regression, decision tree. Firstly split the dataset into the training and testing set, followed by fitting and tuning the models. Then, predict testing data by tuned models, compare each model’s performance based on accuracy and F-1 score. Select the best model with the most significant predictors.

We establish modeling players’ competitiveness by logistic regression and decision tree respectively. Since this is a classification problem, we firstly classify player’s competitiveness based on their ranking, where we follow the classification from the data source in Baseball Reference: Players ranking within 24 are labeled as “Competitive”, and players ranking afterwards are labeled as “Less Competitive”. After that, we implement a train test split for training and testing data on a 70:30 basis. We use 10 fold cross validation to tune both logistic

regression and decision tree to find out optimal models respectively. It turns out that both models perform well on the training set, with both models obtaining 100% of accuracy.

With that, we predict test data players' competitiveness based on tuned models, where the decision tree has 100% of predicting accuracy, followed by logistic regression's 98.64%. In terms of F-1 score, the decision tree has 1, while logistic regression is 0.991. Measurements indicate that both models perform well. However, taking a closer look at variable significance, we find that none of the variables in logistic regression is significant; what's more, tuned tree plot is composed of ranking only (on the other hand, however, it indicates rank of each player is the most important factor to consider). Both models indicate that we may need to investigate further into the data (two guesses would be the existence of a large number of missing values and imbalance classes).



Conclusion

We gained valuable insights from our analysis of the NYM Batting dataset about the factors which influence players' power and specifically slugging percentage which is a critical measure of a player's power in baseball. Through our examination of a range of both demographic and performance based variables, we confirmed our hypothesis that these factors significantly impact a players' slugging percentage. The model demonstrated an R-squared value of approximately 0.66 demonstrating that about 66% of the variability in slugging percentage can be explained by the selected variables. This logistic regression and decision tree analysis illustrated a high level of predicted accuracy in its ability to classify players as either "Competitive" or "Less Competitive."

One interesting finding that we discovered was that the more difficult fielding positions such as SS, 2B, MI, CI, and CF had lower home run rate statistics than positions that are much less difficult fielding positions such as 1B, 3B, LF and RF. An insight that could be drawn upon this is that as a player exerts more energy in the field, it causes a negative relationship between that and their power when hitting. Another interesting finding that we discovered was that the age with the largest number of Mets players is age 28 which is the age most often used when discussing a player's peak performances and highest value to their team. We also found this was backed up by our data due to the fact that those ages are very close to the mean age of 28.13 and a median age of 28.

Our analysis has tried to expand upon the multifaceted nature of baseball and its players' performance and highlight the utility of statistical methods in understanding player contribution. Future research ideas could expand upon our data analysis by exploring more factors that are less numerical such as team dynamic or other psychological attributes. Other avenues also include, employing more sophisticated statistical techniques, such as machine learning algorithms, could enhance the understanding of complex variable relationships, particularly in multivariate settings.

References

Topel, M. (2023). *New York Mets Batting and Pitching (1962-2023)* [Data set]. Kaggle. https://www.kaggle.com/datasets/mattop/new-york-mets-batting-and-pitching-1962-2023

Hands-On Programming with R , O'Reilly, 2014 by G.Grolemund

R for Data Science , 2nd Edition, O'Reilly, 2023 by H.Wickham, M.Çetinkaya-Rundel, G.Grolemund.

White, Ian R., et al. "Avoiding Bias Due to Perfect Prediction in Multiple Imputation of Incomplete Categorical Variables." *Computational Statistics & Data Analysis*, vol. 54, no. 10, Apr. 2010, pp. 2267–75. https://doi.org/10.1016/j.csda.2010.04.005.