Database Selection: NY Mets Batting Statistics 1962-2023

Data Set

Research Question:

What factors affect a player's Power (Slugging Percentage) for the New York Mets Players?

- Some factors to consider are Age (Do older/younger players have more power on average) and Dominant Swings
- Position (Do certain positions aka 1B or SS have more power on average?)
- Year (How Does Power Change Over Time?)

Using Age, Year, and Position as factors and (Slugging Percentage / Batting Average) as → magnitude of each hit → power of the player

Introduction

The data set that we will be analyzing is the NYM Batting data set which 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.

This 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? We believe that the factors are

Exploratory Data Analysis:

1. Fields overview to get familiar with their meaning, data type and underlying relations
2. Ensuring integrity and availability by data cleaning: imputing missing data and outliers, correcting data types and removing unnecessary or unusable data.
3. Univariate analysis by plotting numerical and categorical variables.
4. Analyze relationships between two and more variables, use two-way, scatter, box plot for bivariate analysis and heatmap multivariate analysis

Data Visualization:

1. Multiple Box Plots to show the distribution of Slugging percentage and position + Distribution with Leftie/Rightie
2. Scatter Plot of HomeRun Rate and Age by Position
3. Scatter Plot of batting average categorized by age and position
4. Histogram of the Average OnBase+Slugging Percentages by Age
5. Heatmap of all variables to show their correlation and potential multicollinearity

Statistical Analysis

Code:

```
Creates the boxplot with Dominant Hands
ggplot(NYM_batting_At_least_250_at_Bats, aes(x = Position, y =
Slugging_Percentage, fill = Dominant_Hand)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Distribution of Home Runs by Position(Shown with Dominant
Hands)",
       x = "Position",
       y = "Home Runs") +
  theme_light()
```

```
ggplot(NYM_batting_At_least_250_at_Bats, aes(x = Age, y
= (Home_Runs/At_Bats), color = Position)) +
  geom_point(size = 1, alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Scatter Plot of Average Home Runs by Age
and Position",
       x = "Age",
       y = "Average Home Runs") +
  facet_wrap(~ Position) +
  theme_light()

ggplot(NYM_batting_At_least_250_at_Bats, aes(x = Age, y =
(Slugging_Percentage/Batting_Average), color = Position)) +
  geom_point(size = 1, alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE)+
  labs(title = "Scatter Plot of Magnitude of the Hites by Age and Position",
       x = "Age",
       y = "Magnitude of Hits") +
  facet_wrap(~ Position) +
  theme_light()
```

```
ggplot(NYM_batting_At_least_250_at_Bats, aes(x = Slugging_Percentage * 100, fill = factor(Position))) +
  geom_histogram(binwidth = 1, color = "black", alpha = 0.7) +
  labs(title = "Slugging Percentage Distribution by Age and Position",
       x = "Slugging Average per 100 plate appearances",
       y = "Count",
       fill = "Position") +  # Add legend title
  theme_minimal()
```

```
mets_data <- read.csv("NYM_batting.csv")
filter_mets_data <- subset(mets_data, At_Bats > 250)

numeric_cols <- sapply(filter_mets_data, is.numeric)
filtered_corr <- cor(filter_mets_data[, numeric_cols], use = "complete.obs")

filtered_corr_melted <- melt(filtered_corr)


ggplot(data = filtered_corr_melted, aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1, 1), space = "Lab",  name="Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, size = 12, hjust = 1)) +
  coord_fixed() +
  geom_text(aes(label = round(value, 2)), color = "black", size = 4)
```

```
#STSCI 5120 Final Project
#Statistical Analysis
library(ggplot2)
library(corrplot)
library(MASS)
library(caret)
library(rpart)
library(rpart.plot)
library(e1071)

#If running the code after knit,use
#NYM_batting <- read,csv("NYM_batting.csv")
NYM_batting <- read.csv(file.choose(), header = T)

#Response for linear regression
NYM_batting$Magnitude_Hit <- NYM_batting$Batting_Average /
NYM_batting$Slugging_Percentage

#Based on the data source: https://www.baseball-reference.com/teams/NYM/2023.shtml
#We classify players' powerfulness into two categories:
#1. Those with rank between 1- 23: Competitive
#2. Those with rank after 23: Less Competitive
#Response for logistic regression, decision tree etc.
NYM_batting$Competitiveness <- ifelse(NYM_batting$Rank >= 1 & NYM_batting$Rank <=
23,
                                      "Competitive", "Less Competitive")

colnames(NYM_batting)

#factorization: Position, Dominant_Hand, Switch_Hitter
NYM_batting[c("Position", "Dominant_Hand", "Switch_Hitter", "Competitiveness")] <-
lapply(NYM_batting[c("Position", "Dominant_Hand", "Switch_Hitter", "Competitiveness")],
as.factor)

#Predictors to consider: ~ - Name - Batting_Average - Slugging_Percentage
NYM_batting_subset <- NYM_batting[ , !(names(NYM_batting) %in% c("Name",
"Batting_Average", "Slugging_Percentage"))]

ggplot(NYM_batting_subset, aes(x = Competitiveness)) +
```

```
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Number of Each Competitiveness Class", x = "Competitiveness", y = "Count") +
  theme_minimal()

#1. Linear Regression Analysis for Batting_Average / Slugging_Percentage as the response
numeric_columns <- sapply(NYM_batting_subset, is.numeric)
NYM_batting_numeric <- NYM_batting_subset[ , numeric_columns]
pairs(NYM_batting_numeric)

#Correlation heatmap
correlation_matrix <- cor(NYM_batting_numeric, use = "complete.obs")
corrplot(correlation_matrix, method = "color", type = "upper",
         tl.col = "black", tl.srt = 45, addCoef.col = "black",
         number.cex = 0.5, title = "Correlation Heatmap")

NYM_batting_mod <- lm(Magnitude_Hit~.- Competitiveness, data = NYM_batting_subset)
summary(NYM_batting_mod)

#backward selection
stepAIC(NYM_batting_mod, direction = 'backward')

#diagnostic plots
par(mfrow = c(2,2))
plot(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))
par(mfrow=c(1,1))

NYM_batting_mod_best <- 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,])
summary(NYM_batting_mod_best)

#Before going to 2 and 3, we firstly split the data into train, test set at 70:30 level
set.seed(2024)
```

```
NYM_batting_subset <- na.omit(NYM_batting_subset)
samples <- sample(1:nrow(NYM_batting_subset),nrow(NYM_batting_subset)*0.7,replace=F)
NYM_batting_subset_train <- NYM_batting_subset[samples,]
NYM_batting_subset_test <- NYM_batting_subset[-samples,]



#2. Logistic regression and decision tree
#2.1 logistic regression
NYM_batting_logistic_model <- train(
  Competitiveness ~ .,
  data = NYM_batting_subset_train,
  method = "glm",
  family = "binomial",
  trControl = trainControl(method = "cv", number = 10)
)

train_pred_logistic <- predict(NYM_batting_logistic_model, NYM_batting_subset_train)

actual <- factor(NYM_batting_subset_train$Competitiveness, levels = c("Competitive", "Less
Competitive"))
predicted <- factor(train_pred_logistic, levels = c("Competitive", "Less Competitive"))

true_positive <- sum(predicted == "Competitive" & actual == "Competitive")
false_positive <- sum(predicted == "Competitive" & actual == "Less Competitive")
true_negative <- sum(predicted == "Less Competitive" & actual == "Less Competitive")
false_negative <- sum(predicted == "Less Competitive" & actual == "Competitive")

conf_matrix <- matrix(c(true_positive, false_negative, false_positive, true_negative),
                      nrow = 2,
                      byrow = T,
                      dimnames = list("Prediction" = c("Competitive", "Less Competitive"),
                                      "Reference" = c("Competitive", "Less Competitive")))

print(conf_matrix)



#2.2 decision tree
NYM_batting_tree_model <- train(
  Competitiveness ~ .,
```

```
  data = NYM_batting_subset_train,
  method = "rpart",
  tuneLength = 10,
  trControl = trainControl(method = "cv", number = 10)
)

train_pred_tree <- predict(NYM_batting_tree_model, NYM_batting_subset_train)

actual_tree <- factor(NYM_batting_subset_train$Competitiveness, levels = c("Competitive",
"Less Competitive"))
predicted_tree <- factor(train_pred_tree, levels = c("Competitive", "Less Competitive"))

true_positive_tree <- sum(predicted_tree == "Competitive" & actual_tree == "Competitive")
false_positive_tree <- sum(predicted_tree == "Competitive" & actual_tree == "Less
Competitive")
true_negative_tree <- sum(predicted_tree == "Less Competitive" & actual_tree == "Less
Competitive")
false_negative_tree <- sum(predicted_tree == "Less Competitive" & actual_tree ==
"Competitive")

conf_matrix_tree <- matrix(c(true_positive_tree, false_negative_tree, false_positive_tree,
true_negative_tree),
                           nrow = 2,
                           byrow = T,
                           dimnames = list("Prediction" = c("Competitive", "Less Competitive"),
                                           "Reference" = c("Competitive", "Less Competitive")))

print(conf_matrix_tree)



#predict on testing set
test_pred_logistic <- predict(NYM_batting_tree_model, NYM_batting_subset_test)
test_conf_logistic <- confusionMatrix(test_pred_logistic,
NYM_batting_subset_test$Competitiveness)


test_pred_logistic <- predict(NYM_batting_logistic_model, NYM_batting_subset_test)
test_pred_logistic <- factor(test_pred_logistic, levels = c("Competitive", "Less Competitive"))
```

```
actual_test <- factor(NYM_batting_subset_test$Competitiveness, levels = c("Competitive",
"Less Competitive"))
test_conf_logistic <- confusionMatrix(test_pred_logistic, actual_test)
test_conf_logistic


test_pred_tree <- predict(NYM_batting_tree_model, NYM_batting_subset_test)
test_pred_tree <- factor(test_pred_tree, levels = c("Competitive", "Less Competitive"))
actual_test_tree <- factor(NYM_batting_subset_test$Competitiveness, levels = c("Competitive",
"Less Competitive"))
test_conf_tree <- confusionMatrix(test_pred_tree, actual_test_tree)
test_conf_tree

#performance comparison based on accuracy, F1 score
logistic_accuracy <- test_conf_logistic$overall['Accuracy']
logistic_f1 <- test_conf_logistic$byClass['F1']

tree_accuracy <- test_conf_tree$overall['Accuracy']
tree_f1 <- test_conf_tree$byClass['F1']

#significant predictors in logistic regression (tuned)
summary(NYM_batting_logistic_model$finalModel)

#tree plot and variable importance
best_tree_model <- NYM_batting_tree_model$finalModel
rpart.plot(best_tree_model, type = 3, fallen.leaves = T)

importance <- varImp(NYM_batting_tree_model)
print(importance)
```

Data cleaning and transformation:

After loading the data, we first examine the entire dataset for completeness and consistency. Since there are no missing values, we consider other data cleaning steps to ensure data quality. These steps 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 ensure that categorical variables have consistent levels, such as checking for misspellings or inconsistencies in factor levels (e.g., "Left-Handed" vs. "Left Handed").

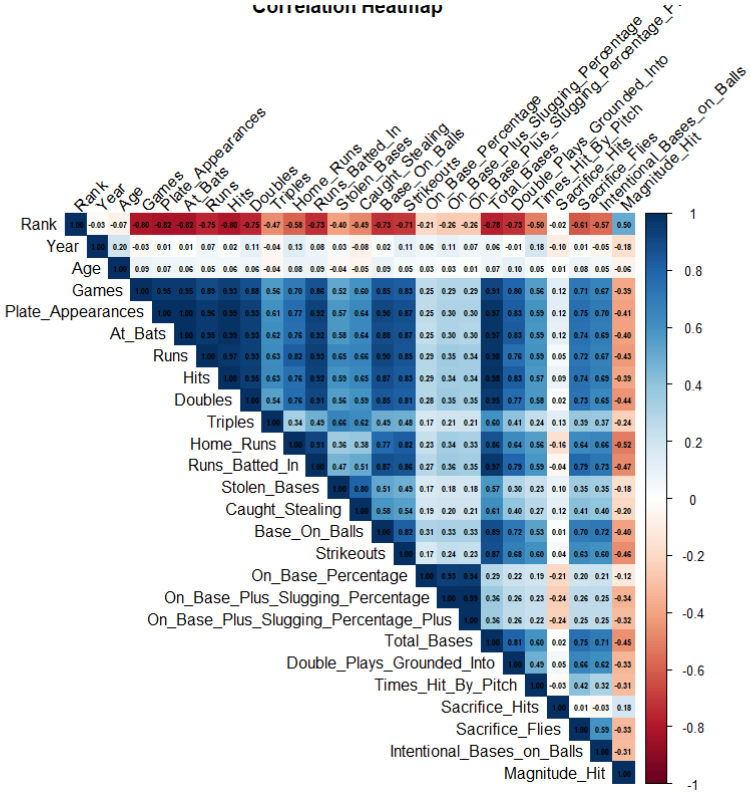
Additionally, we standardize the formats of date and text fields to maintain uniformity across the dataset. These preliminary steps are crucial for maintaining data integrity before proceeding with further analysis.

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.

Statistical Analysis:

1. 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_Avergae 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.

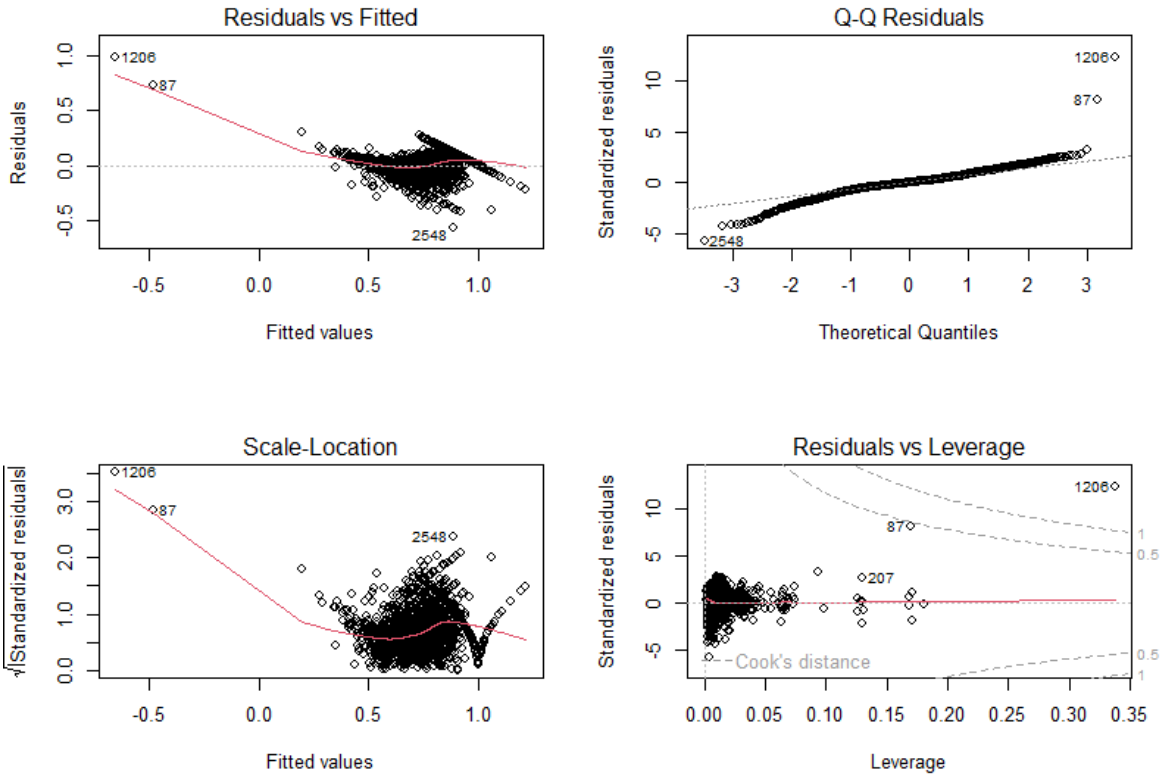
```
Coefficients: (1 not defined because of singularities)
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                          1.490e+00  2.923e-01   5.097 3.79e-07 ***
Rank                                 6.053e-03  6.986e-04   8.665  < 2e-16 ***
Year                                -4.043e-04  1.516e-04  -2.666  0.00774 **
Position1B                          -1.064e-01  4.224e-02  -2.520  0.01181 *
Position2B                          -9.253e-02  4.212e-02  -2.197  0.02814 *
Position3B                          -1.017e-01  4.234e-02  -2.403  0.01636 *
PositionC                           -8.899e-02  4.153e-02  -2.143  0.03227 *
PositionCF                          -9.426e-02  4.222e-02  -2.233  0.02568 *
PositionCI                          -1.359e-01  4.792e-02  -2.837  0.00461 **
PositionDH                          -1.613e-01  5.414e-02  -2.979  0.00293 **
PositionIF                          -1.009e-01  4.280e-02  -2.357  0.01855 *
PositionLF                          -1.096e-01  4.229e-02  -2.592  0.00961 **
PositionMI                          -5.695e-02  4.412e-02  -1.291  0.19686
PositionOF                          -1.136e-01  4.178e-02  -2.718  0.00663 **
PositionP                           -5.225e-02  4.209e-02  -1.241  0.21462
PositionRF                          -1.138e-01  4.216e-02  -2.698  0.00703 **
PositionSS                          -7.043e-02  4.231e-02  -1.665  0.09613 .
PositionUT                          -1.226e-01  4.226e-02  -2.902  0.00376 **
Age                                 -9.354e-04  5.451e-04  -1.716  0.08631 .
Games                               -3.163e-04  1.700e-04  -1.860  0.06304 .
Plate_Appearances                   -6.021e-03  1.455e-02  -0.414  0.67904
At_Bats                              5.876e-03  1.455e-02   0.404  0.68626
Runs                                 1.686e-05  5.741e-04   0.029  0.97658
Hits                                 3.497e-03  4.748e-04   7.365 2.61e-13 ***
Doubles                             -6.098e-03  8.444e-04  -7.222 7.36e-13 ***
Triples                             -1.202e-02  1.969e-03  -6.103 1.26e-09 ***
Home_Runs                           -4.663e-03  1.193e-03  -3.909 9.60e-05 ***
Runs_Batted_In                      -2.253e-04  5.422e-04  -0.416  0.67782
Stolen_Bases                         4.639e-04  6.683e-04   0.694  0.48771
Caught_Stealing                      9.020e-05  1.611e-03   0.056  0.95535
Base_On_Balls                        4.486e-03  1.456e-02   0.308  0.75810
Strikeouts                          -1.706e-04  2.029e-04  -0.841  0.40054
On_Base_Percentage                   1.927e+00  6.543e-02  29.449  < 2e-16 ***
On_Base_Plus_Slugging_Percentage    -7.098e-01  9.049e-02  -7.844 7.20e-15 ***
On_Base_Plus_Slugging_Percentage_Plus -6.735e-04  3.566e-04  -1.889  0.05906 .
Total_Bases                                NA         NA      NA       NA
Double_Plays_Grounded_Into           5.965e-04  1.001e-03   0.596  0.55150
Times_Hit_By_Pitch                   5.737e-03  1.459e-02   0.393  0.69425
Sacrifice_Hits                       5.334e-03  1.455e-02   0.367  0.71393
Sacrifice_Flies                      9.662e-03  1.468e-02   0.658  0.51055
Intentional_Bases_on_Balls           1.208e-03  1.250e-03   0.966  0.33395
Dominant_HandRight                  -1.419e-03  5.386e-03  -0.263  0.79222
Switch_HitterYes                     3.268e-03  8.229e-03   0.397  0.69134
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0983 on 1917 degrees of freedom
  (因为不存在，769个观察量被删除了)
Multiple R-squared:  0.6596,    Adjusted R-squared:  0.6523
F-statistic: 90.58 on 41 and 1917 DF,  p-value: < 2.2e-16

> 
```

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



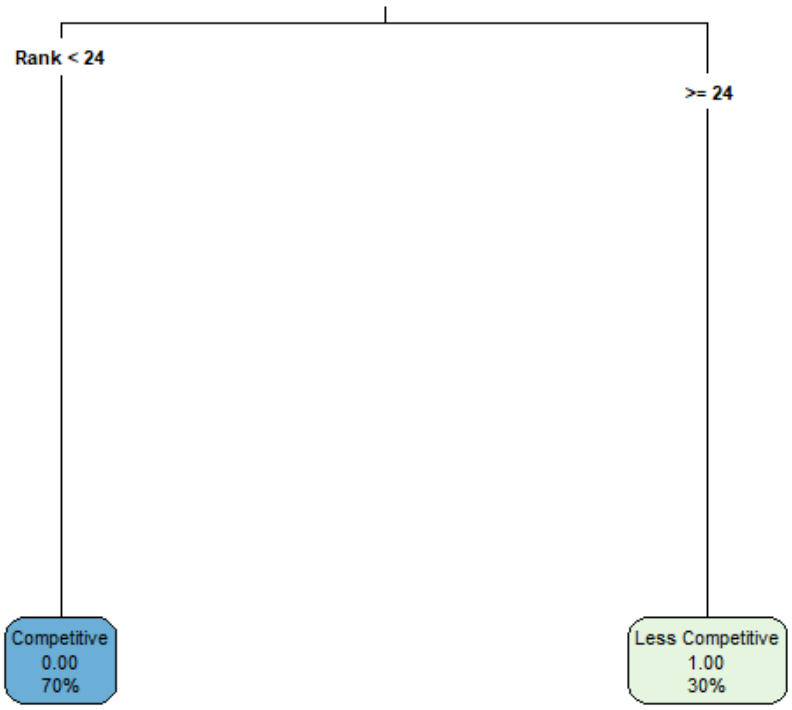
Now we fit the model based on selected predictors along with dataset that excludes the outlier, i.e., observation 1206, where from the summary we can see R-Squared value is increased by 3% without loss of testing (both t-test and F-test) significance.

```
> 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
```

2. Construct a model that evaluates player's performance using logistic regression (multinomial), decision tree, and other models. 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.

   Now we consider 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. 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).



Explanation for new packages used in statistical analysis

Corrplot: A package in **R** that is specifically designed for visualizing correlation matrices. It provides a variety of methods for displaying correlations. We used it to create the correlation heatmap.

Rpart: Generate a visual representation of the decision tree, helping us to interpret the model and understand how different predictors contribute to predicting the Slugging Percentage.

<u>Conclusions and Insights:</u>

From the analysis we conducted

Limitations and future research: For future studies, increasing sample sizes and ensuring higher data quality at the collection stage would improve the reliability of the findings. Moreover, employing more sophisticated statistical techniques, such as machine learning algorithms, could enhance the understanding of complex variable relationships, particularly in multivariate settings.

5. R Code:
    - All your analysis should be done in R, and your project should include well-commented code that is easy to follow.
    - Make sure the code runs smoothly from start to finish, with all necessary libraries and functions loaded at the beginning.
6. Report Structure: Your final project should be presented as a report with the following sections:
    - Introduction: Describe the dataset, the research question, and your hypothesis.
    - Data Preparation: Explain the data cleaning and transformation steps.
    - Exploratory Data Analysis (EDA): Include your summary statis- tics and visualizations.
    - Statistical Analysis: Detail the statistical tests or models you used, along with their results.
    - Conclusion: Summarize your key findings and insights. 2

• References: Cite any sources for data or external libraries used. 10. Submission:

- Submit a well-formatted report (PDF) along with your R script (Rmd file with Html output).
- The report should be between 5-10 pages, not including code.
- Ensure that all plots and tables are properly integrated into the re- port (as well as in the notebook).

Tests: Slugging Percentage / Batting Average → shows the magnitude of the hit for each player (2 → usually gets around a double per hit)

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

Report Structure

- Introduction: Describe the dataset, the research question, and your hypothesis.
- Data Preparation: Explain the data cleaning and transformation steps.
- Exploratory Data Analysis (EDA): Include your summary statistics and visualizations.
- Statistical Analysis: Detail the statistical tests or models you used, along with their results.
- Conclusion: Summarize your key findings and insights.
- References: Cite any sources for data or external libraries used.
- Appendices: List of all codes for reference

Filters out all of the outliers by only having players who have played at least about half of the season.

```
NYM_batting_At_least_250_at_Bats <- NYM_batting %>%
```

filter(At_Bats > 250)

Appendices

Statistical Analysis

1. Summary of the full linear model

```
Call:
lm(formula = Magnitude_Hit ~ . - Competitiveness, data = NYM_batting_subset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.55602 -0.04206  0.00276  0.04847  0.98693

Coefficients: (1 not defined because of singularities)
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        1.490e+00  2.923e-01   5.097 3.79e-07 ***
Rank                               6.053e-03  6.986e-04   8.665  < 2e-16 ***
Year                              -4.043e-04  1.516e-04  -2.666  0.00774 **
Position1B                        -1.064e-01  4.224e-02  -2.520  0.01181 *
Position2B                        -9.253e-02  4.212e-02  -2.197  0.02814 *
Position3B                        -1.017e-01  4.234e-02  -2.403  0.01636 *
PositionC                         -8.899e-02  4.153e-02  -2.143  0.03227 *
PositionCF                        -9.426e-02  4.222e-02  -2.233  0.02568 *
PositionCI                        -1.359e-01  4.792e-02  -2.837  0.00461 **
PositionDH                        -1.613e-01  5.414e-02  -2.979  0.00293 **
PositionIF                        -1.009e-01  4.280e-02  -2.357  0.01855 *
PositionLF                        -1.096e-01  4.229e-02  -2.592  0.00961 **
PositionMI                        -5.695e-02  4.412e-02  -1.291  0.19686
PositionOF                        -1.136e-01  4.178e-02  -2.718  0.00663 **
PositionP                         -5.225e-02  4.209e-02  -1.241  0.21462
PositionRF                        -1.138e-01  4.216e-02  -2.698  0.00703 **
PositionSS                        -7.043e-02  4.231e-02  -1.665  0.09613 .
PositionUT                        -1.226e-01  4.226e-02  -2.902  0.00376 **
Age                               -9.354e-04  5.451e-04  -1.716  0.08631 .
Games                             -3.163e-04  1.700e-04  -1.860  0.06304 .
Plate_Appearances                 -6.021e-03  1.455e-02  -0.414  0.67904
```

```
At_Bats                                  5.876e-03  1.455e-02   0.404  0.68626
Runs                                     1.686e-05  5.741e-04   0.029  0.97658
Hits                                     3.497e-03  4.748e-04   7.365 2.61e-13 ***
Doubles                                 -6.098e-03  8.444e-04  -7.222 7.36e-13 ***
Triples                                 -1.202e-02  1.969e-03  -6.103 1.26e-09 ***
Home_Runs                               -4.663e-03  1.193e-03  -3.909 9.60e-05 ***
Runs_Batted_In                          -2.253e-04  5.422e-04  -0.416  0.67782
Stolen_Bases                             4.639e-04  6.683e-04   0.694  0.48771
Caught_Stealing                          9.020e-05  1.611e-03   0.056  0.95535
Base_On_Balls                            4.486e-03  1.456e-02   0.308  0.75810
Strikeouts                              -1.706e-04  2.029e-04  -0.841  0.40054
On_Base_Percentage                       1.927e+00  6.543e-02  29.449  < 2e-16 ***
On_Base_Plus_Slugging_Percentage        -7.098e-01  9.049e-02  -7.844 7.20e-15 ***
On_Base_Plus_Slugging_Percentage_Plus -6.735e-04  3.566e-04  -1.889  0.05906 .
Total_Bases                                   NA         NA      NA       NA
Double_Plays_Grounded_Into               5.965e-04  1.001e-03   0.596  0.55150
Times_Hit_By_Pitch                       5.737e-03  1.459e-02   0.393  0.69425
Sacrifice_Hits                           5.334e-03  1.455e-02   0.367  0.71393
Sacrifice_Flies                          9.662e-03  1.468e-02   0.658  0.51055
Intentional_Bases_on_Balls               1.208e-03  1.250e-03   0.966  0.33395
Dominant_HandRight                      -1.419e-03  5.386e-03  -0.263  0.79222
Switch_HitterYes                         3.268e-03  8.229e-03   0.397  0.69134
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0983 on 1917 degrees of freedom
  (因为不存在，769个观察量被删除了)
Multiple R-squared:  0.6596,    Adjusted R-squared:  0.6523
F-statistic: 90.58 on 41 and 1917 DF,  p-value: < 2.2e-16
```

2. Backward selection based on full model

```
> #backward selection
> stepAIC(NYM_batting_mod, direction = 'backward')
Start:  AIC=-9046.99
Magnitude_Hit ~ (Rank + Year + Position + Age + Games + Plate_Appearances +
    At_Bats + Runs + Hits + Doubles + Triples + Home_Runs + Runs_Batted_In +
    Stolen_Bases + Caught_Stealing + Base_On_Balls + Strikeouts +
    On_Base_Percentage + On_Base_Plus_Slugging_Percentage + On_Base_Plus_Slugging_Percentage_Plus +
    Total_Bases + Double_Plays_Grounded_Into + Times_Hit_By_Pitch +
    Sacrifice_Hits + Sacrifice_Flies + Intentional_Bases_on_Balls +
    Dominant_Hand + Switch_Hitter + Competitiveness) - Competitiveness


Step:  AIC=-9046.99
Magnitude_Hit ~ Rank + Year + Position + Age + Games + Plate_Appearances +
    At_Bats + Runs + Hits + Doubles + Triples + Home_Runs + Runs_Batted_In +
    Stolen_Bases + Caught_Stealing + Base_On_Balls + Strikeouts +
    On_Base_Percentage + On_Base_Plus_Slugging_Percentage + On_Base_Plus_Slugging_Percentage_Plus +
    Double_Plays_Grounded_Into + Times_Hit_By_Pitch + Sacrifice_Hits +
    Sacrifice_Flies + Intentional_Bases_on_Balls + Dominant_Hand +
    Switch_Hitter

                                       Df Sum of Sq    RSS     AIC
- Runs                                  1    0.0000 18.525 -9049.0
- Caught_Stealing                       1    0.0000 18.526 -9049.0
- Dominant_Hand                         1    0.0007 18.526 -9048.9
- Base_On_Balls                         1    0.0009 18.526 -9048.9
- Sacrifice_Hits                        1    0.0013 18.527 -9048.8
- Times_Hit_By_Pitch                    1    0.0015 18.527 -9048.8
- Switch_Hitter                         1    0.0015 18.527 -9048.8
- At_Bats                               1    0.0016 18.527 -9048.8
- Plate_Appearances                     1    0.0017 18.527 -9048.8
- Runs_Batted_In                        1    0.0017 18.527 -9048.8
- Double_Plays_Grounded_Into            1    0.0034 18.529 -9048.6
- Sacrifice_Flies                       1    0.0042 18.530 -9048.5
- Stolen_Bases                          1    0.0047 18.530 -9048.5
- Strikeouts                            1    0.0068 18.532 -9048.3
- Intentional_Bases_on_Balls            1    0.0090 18.535 -9048.0
<none>                                         18.525 -9047.0
- Age                                   1    0.0285 18.554 -9046.0
- Games                                 1    0.0334 18.559 -9045.5
- On_Base_Plus_Slugging_Percentage_Plus 1    0.0345 18.560 -9045.3
- Year                                  1    0.0687 18.594 -9041.7
- Home_Runs                             1    0.1477 18.673 -9033.4
- Position                             15    0.5853 19.111 -9016.1
- Triples                               1    0.3599 18.885 -9011.3
- Doubles                               1    0.5040 19.029 -8996.4
- Hits                                  1    0.5243 19.050 -8994.3
- On_Base_Plus_Slugging_Percentage      1    0.5946 19.120 -8987.1
- Rank                                  1    0.7255 19.251 -8973.7
- On_Base_Percentage                    1    8.3809 26.906 -8317.9

Step:  AIC=-9048.99
Magnitude_Hit ~ Rank + Year + Position + Age + Games + Plate_Appearances +
    At_Bats + Hits + Doubles + Triples + Home_Runs + Runs_Batted_In +
    Stolen_Bases + Caught_Stealing + Base_On_Balls + Strikeouts +
    On_Base_Percentage + On_Base_Plus_Slugging_Percentage + On_Base_Plus_Slugging_Percentage_Plus +
    Double_Plays_Grounded_Into + Times_Hit_By_Pitch + Sacrifice_Hits +
    Sacrifice_Flies + Intentional_Bases_on_Balls + Dominant_Hand +
    Switch_Hitter
```

	Df	Sum of Sq	RSS	AIC
- Caught_Stealing	1	0.0000	18.526	-9051.0
- Dominant_Hand	1	0.0007	18.526	-9050.9
- Base_On_Balls	1	0.0009	18.526	-9050.9
- Sacrifice_Hits	1	0.0013	18.527	-9050.8
- Times_Hit_By_Pitch	1	0.0015	18.527	-9050.8
- Switch_Hitter	1	0.0015	18.527	-9050.8
- At_Bats	1	0.0016	18.527	-9050.8
- Plate_Appearances	1	0.0017	18.527	-9050.8
- Runs_Batted_In	1	0.0017	18.527	-9050.8
- Double_Plays_Grounded_Into	1	0.0034	18.529	-9050.6
- Sacrifice_Flies	1	0.0042	18.530	-9050.5
- Stolen_Bases	1	0.0056	18.531	-9050.4
- Strikeouts	1	0.0069	18.532	-9050.3
- Intentional_Bases_on_Balls	1	0.0093	18.535	-9050.0
<none>			18.525	-9049.0
- Age	1	0.0286	18.554	-9048.0
- Games	1	0.0335	18.559	-9047.5
- On_Base_Plus_Slugging_Percentage_Plus	1	0.0348	18.560	-9047.3
- Year	1	0.0687	18.594	-9043.7
- Home_Runs	1	0.1766	18.702	-9032.4
- Position	15	0.5859	19.111	-9018.0
- Triples	1	0.3691	18.895	-9012.3
- Doubles	1	0.5143	19.040	-8997.3
- On_Base_Plus_Slugging_Percentage	1	0.5973	19.123	-8988.8
- Hits	1	0.6064	19.132	-8987.9
- Rank	1	0.7289	19.254	-8975.4
- On_Base_Percentage	1	8.3912	26.917	-8319.1

```
Step:  AIC=-9050.98
Magnitude_Hit ~ Rank + Year + Position + Age + Games + Plate_Appearances +
    At_Bats + Hits + Doubles + Triples + Home_Runs + Runs_Batted_In +
    Stolen_Bases + Base_On_Balls + Strikeouts + On_Base_Percentage +
    On_Base_Plus_Slugging_Percentage + On_Base_Plus_Slugging_Percentage_Plus +
    Double_Plays_Grounded_Into + Times_Hit_By_Pitch + Sacrifice_Hits +
    Sacrifice_Flies + Intentional_Bases_on_Balls + Dominant_Hand +
    Switch_Hitter

                                       Df Sum of Sq    RSS     AIC
- Dominant_Hand                         1    0.0007 18.526 -9052.9
- Base_On_Balls                         1    0.0009 18.526 -9052.9
- Sacrifice_Hits                        1    0.0013 18.527 -9052.8
- Times_Hit_By_Pitch                    1    0.0015 18.527 -9052.8
- Switch_Hitter                         1    0.0015 18.527 -9052.8
- At_Bats                               1    0.0016 18.527 -9052.8
- Plate_Appearances                     1    0.0017 18.527 -9052.8
- Runs_Batted_In                        1    0.0018 18.527 -9052.8
- Double_Plays_Grounded_Into            1    0.0034 18.529 -9052.6
- Sacrifice_Flies                       1    0.0042 18.530 -9052.5
- Strikeouts                            1    0.0069 18.532 -9052.3
- Stolen_Bases                          1    0.0088 18.534 -9052.1
- Intentional_Bases_on_Balls            1    0.0093 18.535 -9052.0
<none>                                         18.526 -9051.0
- Age                                   1    0.0287 18.554 -9049.9
- Games                                 1    0.0334 18.559 -9049.5
- On_Base_Plus_Slugging_Percentage_Plus 1    0.0348 18.560 -9049.3
- Year                                  1    0.0703 18.596 -9045.6
- Home_Runs                             1    0.1766 18.702 -9034.4
```

```
- Position                                15    0.5893 19.115 -9019.6
- Triples                                  1    0.3690 18.895 -9014.3
- Doubles                                  1    0.5144 19.040 -8999.3
- On_Base_Plus_Slugging_Percentage         1    0.5973 19.123 -8990.8
- Hits                                     1    0.6094 19.135 -8989.6
- Rank                                     1    0.7308 19.256 -8977.2
- On_Base_Percentage                       1    8.3915 26.917 -8321.1

Step:  AIC=-9052.91
Magnitude_Hit ~ Rank + Year + Position + Age + Games + Plate_Appearances +
    At_Bats + Hits + Doubles + Triples + Home_Runs + Runs_Batted_In +
    Stolen_Bases + Base_On_Balls + Strikeouts + On_Base_Percentage +
    On_Base_Plus_Slugging_Percentage + On_Base_Plus_Slugging_Percentage_Plus +
    Double_Plays_Grounded_Into + Times_Hit_By_Pitch + Sacrifice_Hits +
    Sacrifice_Flies + Intentional_Bases_on_Balls + Switch_Hitter

                                        Df Sum of Sq    RSS     AIC
- Base_On_Balls                          1    0.0009 18.527 -9054.8
- Switch_Hitter                          1    0.0011 18.527 -9054.8
- Sacrifice_Hits                         1    0.0013 18.527 -9054.8
- Times_Hit_By_Pitch                     1    0.0014 18.528 -9054.8
- At_Bats                                1    0.0015 18.528 -9054.8
- Plate_Appearances                      1    0.0016 18.528 -9054.7
- Runs_Batted_In                         1    0.0018 18.528 -9054.7
- Double_Plays_Grounded_Into             1    0.0031 18.529 -9054.6
- Sacrifice_Flies                        1    0.0041 18.530 -9054.5
- Strikeouts                             1    0.0073 18.533 -9054.1
- Stolen_Bases                           1    0.0085 18.535 -9054.0
- Intentional_Bases_on_Balls             1    0.0095 18.536 -9053.9
<none>                                          18.526 -9052.9
- Age                                    1    0.0290 18.555 -9051.8
- Games                                  1    0.0330 18.559 -9051.4
- On_Base_Plus_Slugging_Percentage_Plus  1    0.0343 18.560 -9051.3
- Year                                   1    0.0697 18.596 -9047.6
- Home_Runs                              1    0.1765 18.703 -9036.3
- Position                              15    0.5924 19.119 -9021.2
- Triples                                1    0.3684 18.895 -9016.3
- Doubles                                1    0.5142 19.040 -9001.3
- On_Base_Plus_Slugging_Percentage       1    0.6038 19.130 -8992.1
- Hits                                   1    0.6092 19.135 -8991.5
- Rank                                   1    0.7302 19.256 -8979.2
- On_Base_Percentage                     1    8.3928 26.919 -8322.9

Step:  AIC=-9054.82
Magnitude_Hit ~ Rank + Year + Position + Age + Games + Plate_Appearances +
    At_Bats + Hits + Doubles + Triples + Home_Runs + Runs_Batted_In +
    Stolen_Bases + Strikeouts + On_Base_Percentage + On_Base_Plus_Slugging_Percentage +
    On_Base_Plus_Slugging_Percentage_Plus + Double_Plays_Grounded_Into +
    Times_Hit_By_Pitch + Sacrifice_Hits + Sacrifice_Flies + Intentional_Bases_on_Balls +
    Switch_Hitter

                                        Df Sum of Sq    RSS     AIC
- Switch_Hitter                          1    0.0011 18.528 -9056.7
- Runs_Batted_In                         1    0.0019 18.529 -9056.6
```

```
- Double_Plays_Grounded_Into             1    0.0030 18.530 -9056.5
- Sacrifice_Hits                         1    0.0062 18.533 -9056.2
- Strikeouts                             1    0.0075 18.535 -9056.0
- Stolen_Bases                           1    0.0086 18.536 -9055.9
- Times_Hit_By_Pitch                     1    0.0093 18.536 -9055.8
- Intentional_Bases_on_Balls             1    0.0096 18.537 -9055.8
<none>                                          18.527 -9054.8
- Age                                    1    0.0289 18.556 -9053.8
- Games                                  1    0.0338 18.561 -9053.2
- On_Base_Plus_Slugging_Percentage_Plus  1    0.0342 18.561 -9053.2
- Sacrifice_Flies                        1    0.0643 18.591 -9050.0
- Year                                   1    0.0702 18.597 -9049.4
- Home_Runs                              1    0.1757 18.703 -9038.3
- At_Bats                                1    0.1852 18.712 -9037.3
- Plate_Appearances                      1    0.2936 18.821 -9026.0
- Position                              15    0.5922 19.119 -9023.2
- Triples                                1    0.3698 18.897 -9018.1
- Doubles                                1    0.5137 19.041 -9003.2
- On_Base_Plus_Slugging_Percentage       1    0.6043 19.131 -8993.9
- Hits                                   1    0.6096 19.137 -8993.4
- Rank                                   1    0.7309 19.258 -8981.0
- On_Base_Percentage                     1    8.3948 26.922 -8324.7

Step:  AIC=-9056.71
Magnitude_Hit ~ Rank + Year + Position + Age + Games + Plate_Appearances +
    At_Bats + Hits + Doubles + Triples + Home_Runs + Runs_Batted_In +
    Stolen_Bases + Strikeouts + On_Base_Percentage + On_Base_Plus_Slugging_Percentage +
    On_Base_Plus_Slugging_Percentage_Plus + Double_Plays_Grounded_Into +
    Times_Hit_By_Pitch + Sacrifice_Hits + Sacrifice_Flies + Intentional_Bases_on_Balls

                                       Df Sum of Sq    RSS     AIC
- Runs_Batted_In                        1    0.0019 18.530 -9058.5
- Double_Plays_Grounded_Into            1    0.0028 18.531 -9058.4
- Sacrifice_Hits                        1    0.0060 18.534 -9058.1
- Strikeouts                            1    0.0076 18.536 -9057.9
- Times_Hit_By_Pitch                    1    0.0086 18.537 -9057.8
- Intentional_Bases_on_Balls            1    0.0096 18.538 -9057.7
- Stolen_Bases                          1    0.0105 18.539 -9057.6
<none>                                          18.528 -9056.7
- Age                                   1    0.0287 18.557 -9055.7
- Games                                 1    0.0343 18.562 -9055.1
- On_Base_Plus_Slugging_Percentage_Plus 1    0.0348 18.563 -9055.0
- Sacrifice_Flies                       1    0.0642 18.592 -9051.9
- Year                                  1    0.0696 18.598 -9051.4
- Home_Runs                             1    0.1752 18.703 -9040.3
- At_Bats                               1    0.1841 18.712 -9039.3
- Plate_Appearances                     1    0.2927 18.821 -9028.0
- Position                             15    0.5993 19.127 -9024.3
- Triples                               1    0.3690 18.897 -9020.1
- Doubles                               1    0.5146 19.043 -9005.0
- On_Base_Plus_Slugging_Percentage      1    0.6032 19.131 -8995.9
- Hits                                  1    0.6086 19.137 -8995.4
```

```
- Rank                                    1    0.7303 19.258 -8983.0
- On_Base_Percentage                      1    8.4137 26.942 -8325.3

Step:  AIC=-9058.51
Magnitude_Hit ~ Rank + Year + Position + Age + Games + Plate_Appearances +
    At_Bats + Hits + Doubles + Triples + Home_Runs + Stolen_Bases +
    Strikeouts + On_Base_Percentage + On_Base_Plus_Slugging_Percentage +
    On_Base_Plus_Slugging_Percentage_Plus + Double_Plays_Grounded_Into +
    Times_Hit_By_Pitch + Sacrifice_Hits + Sacrifice_Flies + Intentional_Bases_on_Balls

                                       Df Sum of Sq    RSS     AIC
- Double_Plays_Grounded_Into            1    0.0022 18.532 -9060.3
- Sacrifice_Hits                        1    0.0063 18.536 -9059.8
- Strikeouts                            1    0.0077 18.538 -9059.7
- Times_Hit_By_Pitch                    1    0.0089 18.539 -9059.6
- Intentional_Bases_on_Balls            1    0.0089 18.539 -9059.6
- Stolen_Bases                          1    0.0114 18.541 -9059.3
<none>                                             18.530 -9058.5
- Age                                   1    0.0288 18.559 -9057.5
- On_Base_Plus_Slugging_Percentage_Plus 1    0.0343 18.564 -9056.9
- Games                                 1    0.0343 18.564 -9056.9
- Sacrifice_Flies                       1    0.0673 18.597 -9053.4
- Year                                  1    0.0693 18.599 -9053.2
- At_Bats                               1    0.1879 18.718 -9040.7
- Plate_Appearances                     1    0.2981 18.828 -9029.2
- Position                             15    0.6008 19.131 -9026.0
- Triples                               1    0.3689 18.899 -9021.9
- Home_Runs                             1    0.4534 18.983 -9013.2
- Doubles                               1    0.5410 19.071 -9004.1
- On_Base_Plus_Slugging_Percentage      1    0.6058 19.136 -8997.5
- Hits                                  1    0.6265 19.157 -8995.4
- Rank                                  1    0.7285 19.259 -8985.0
- On_Base_Percentage                    1    8.4125 26.942 -8327.2

Step:  AIC=-9060.27
Magnitude_Hit ~ Rank + Year + Position + Age + Games + Plate_Appearances +
    At_Bats + Hits + Doubles + Triples + Home_Runs + Stolen_Bases +
    Strikeouts + On_Base_Percentage + On_Base_Plus_Slugging_Percentage +
    On_Base_Plus_Slugging_Percentage_Plus + Times_Hit_By_Pitch +
    Sacrifice_Hits + Sacrifice_Flies + Intentional_Bases_on_Balls

                                       Df Sum of Sq    RSS     AIC
- Sacrifice_Hits                        1    0.0059 18.538 -9061.7
- Strikeouts                            1    0.0085 18.541 -9061.4
- Times_Hit_By_Pitch                    1    0.0087 18.541 -9061.4
- Stolen_Bases                          1    0.0094 18.542 -9061.3
- Intentional_Bases_on_Balls            1    0.0095 18.542 -9061.3
<none>                                             18.532 -9060.3
- Age                                   1    0.0283 18.561 -9059.3
- Games                                 1    0.0342 18.566 -9058.7
- On_Base_Plus_Slugging_Percentage_Plus 1    0.0354 18.568 -9058.5
- Sacrifice_Flies                       1    0.0680 18.600 -9055.1
- Year                                  1    0.0695 18.602 -9054.9
```

```
- At_Bats                                  1    0.1979 18.730 -9041.5
- Plate_Appearances                        1    0.3046 18.837 -9030.3
- Position                                15    0.6021 19.134 -9027.6
- Triples                                  1    0.3758 18.908 -9022.9
- Home_Runs                                1    0.4544 18.987 -9014.8
- Doubles                                  1    0.5484 19.081 -9005.1
- On_Base_Plus_Slugging_Percentage         1    0.6036 19.136 -8999.5
- Hits                                     1    0.6379 19.170 -8996.0
- Rank                                     1    0.7263 19.259 -8987.0
- On_Base_Percentage                       1    8.4204 26.953 -8328.5

Step:  AIC=-9061.65
Magnitude_Hit ~ Rank + Year + Position + Age + Games + Plate_Appearances +
    At_Bats + Hits + Doubles + Triples + Home_Runs + Stolen_Bases +
    Strikeouts + On_Base_Percentage + On_Base_Plus_Slugging_Percentage +
    On_Base_Plus_Slugging_Percentage_Plus + Times_Hit_By_Pitch +
    Sacrifice_Flies + Intentional_Bases_on_Balls

                                       Df Sum of Sq    RSS     AIC
- Times_Hit_By_Pitch                    1    0.0076 18.546 -9062.9
- Intentional_Bases_on_Balls            1    0.0079 18.546 -9062.8
- Strikeouts                            1    0.0092 18.547 -9062.7
- Stolen_Bases                          1    0.0093 18.547 -9062.7
<none>                                             18.538 -9061.7
- Age                                   1    0.0281 18.566 -9060.7
- Games                                 1    0.0363 18.574 -9059.8
- On_Base_Plus_Slugging_Percentage_Plus 1    0.0376 18.576 -9059.7
- Sacrifice_Flies                       1    0.0656 18.604 -9056.7
- Year                                  1    0.0688 18.607 -9056.4
- At_Bats                               1    0.1921 18.730 -9043.5
- Plate_Appearances                     1    0.3033 18.841 -9031.9
- Triples                               1    0.3771 18.915 -9024.2
- Position                             15    0.7125 19.251 -9017.8
- Home_Runs                             1    0.5236 19.062 -9009.1
- Doubles                               1    0.5728 19.111 -9004.0
- On_Base_Plus_Slugging_Percentage      1    0.5979 19.136 -9001.5
- Hits                                  1    0.6327 19.171 -8997.9
- Rank                                  1    0.7270 19.265 -8988.3
- On_Base_Percentage                    1    8.4313 26.969 -8329.3

Step:  AIC=-9062.85
Magnitude_Hit ~ Rank + Year + Position + Age + Games + Plate_Appearances +
    At_Bats + Hits + Doubles + Triples + Home_Runs + Stolen_Bases +
    Strikeouts + On_Base_Percentage + On_Base_Plus_Slugging_Percentage +
    On_Base_Plus_Slugging_Percentage_Plus + Sacrifice_Flies +
    Intentional_Bases_on_Balls

                                       Df Sum of Sq    RSS     AIC
- Intentional_Bases_on_Balls            1    0.0050 18.551 -9064.3
- Stolen_Bases                          1    0.0072 18.553 -9064.1
- Strikeouts                            1    0.0081 18.554 -9064.0
<none>                                             18.546 -9062.9
- Age                                   1    0.0299 18.576 -9061.7
```

```
- Games                                  1    0.0361 18.582 -9061.0
- On_Base_Plus_Slugging_Percentage_Plus  1    0.0364 18.582 -9061.0
- Sacrifice_Flies                       1    0.0619 18.608 -9058.3
- Year                                  1    0.0635 18.609 -9058.2
- At_Bats                               1    0.1849 18.731 -9045.4
- Plate_Appearances                     1    0.2964 18.842 -9033.8
- Triples                               1    0.3835 18.929 -9024.8
- Position                             15    0.7081 19.254 -9019.5
- Home_Runs                             1    0.5164 19.062 -9011.0
- Doubles                               1    0.5693 19.115 -9005.6
- On_Base_Plus_Slugging_Percentage      1    0.6048 19.150 -9002.0
- Hits                                  1    0.6311 19.177 -8999.3
- Rank                                  1    0.7309 19.276 -8989.1
- On_Base_Percentage                    1    8.4451 26.991 -8329.7

Step:  AIC=-9064.32
Magnitude_Hit ~ Rank + Year + Position + Age + Games + Plate_Appearances +
    At_Bats + Hits + Doubles + Triples + Home_Runs + Stolen_Bases +
    Strikeouts + On_Base_Percentage + On_Base_Plus_Slugging_Percentage +
    On_Base_Plus_Slugging_Percentage_Plus + Sacrifice_Flies

                                       Df Sum of Sq    RSS     AIC
- Stolen_Bases                          1    0.0072 18.558 -9065.6
- Strikeouts                            1    0.0108 18.561 -9065.2
<none>                                             18.551 -9064.3
- Age                                   1    0.0300 18.581 -9063.2
- Games                                 1    0.0349 18.585 -9062.6
- On_Base_Plus_Slugging_Percentage_Plus 1    0.0350 18.586 -9062.6
- Sacrifice_Flies                       1    0.0638 18.614 -9059.6
- Year                                  1    0.0657 18.616 -9059.4
- At_Bats                               1    0.1801 18.731 -9047.4
- Plate_Appearances                     1    0.2925 18.843 -9035.7
- Triples                               1    0.3823 18.933 -9026.4
- Position                             15    0.7135 19.264 -9020.4
- Home_Runs                             1    0.5277 19.078 -9011.4
- Doubles                               1    0.5754 19.126 -9006.5
- On_Base_Plus_Slugging_Percentage      1    0.6133 19.164 -9002.6
- Hits                                  1    0.6288 19.180 -9001.0
- Rank                                  1    0.7313 19.282 -8990.6
- On_Base_Percentage                    1    8.4416 26.992 -8331.6

Step:  AIC=-9065.56
Magnitude_Hit ~ Rank + Year + Position + Age + Games + Plate_Appearances +
    At_Bats + Hits + Doubles + Triples + Home_Runs + Strikeouts +
    On_Base_Percentage + On_Base_Plus_Slugging_Percentage + On_Base_Plus_Slugging_Percentage_Plus +
    Sacrifice_Flies

                                       Df Sum of Sq    RSS     AIC
- Strikeouts                            1    0.0096 18.567 -9066.5
<none>                                             18.558 -9065.6
- Age                                   1    0.0310 18.589 -9064.3
- On_Base_Plus_Slugging_Percentage_Plus 1    0.0351 18.593 -9063.9
- Games                                 1    0.0367 18.595 -9063.7
```

```
- Games                                   1    0.0367 18.595 -9063.7
- Sacrifice_Flies                         1    0.0608 18.619 -9061.2
- Year                                    1    0.0641 18.622 -9060.8
- At_Bats                                 1    0.1753 18.733 -9049.1
- Plate_Appearances                       1    0.2869 18.845 -9037.5
- Triples                                 1    0.4033 18.961 -9025.4
- Position                               15    0.7085 19.266 -9022.2
- Home_Runs                               1    0.5436 19.102 -9011.0
- Doubles                                 1    0.5693 19.127 -9008.4
- On_Base_Plus_Slugging_Percentage        1    0.6151 19.173 -9003.7
- Hits                                    1    0.6479 19.206 -9000.3
- Rank                                    1    0.7591 19.317 -8989.0
- On_Base_Percentage                      1    8.4733 27.031 -8330.8

Step:  AIC=-9066.55
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

                                        Df Sum of Sq    RSS     AIC
<none>                                              18.567 -9066.5
- Age                                    1    0.0278 18.595 -9065.6
- On_Base_Plus_Slugging_Percentage_Plus  1    0.0360 18.604 -9064.8
- Games                                  1    0.0385 18.606 -9064.5
- Sacrifice_Flies                        1    0.0728 18.640 -9060.9
- Year                                   1    0.0776 18.645 -9060.4
- At_Bats                                1    0.1693 18.737 -9050.8
- Plate_Appearances                      1    0.3056 18.873 -9036.6
- Triples                                1    0.4206 18.988 -9024.7
- Position                              15    0.7119 19.279 -9022.8
- Doubles                                1    0.5980 19.165 -9006.5
- On_Base_Plus_Slugging_Percentage       1    0.6109 19.178 -9005.1
- Rank                                   1    0.7735 19.341 -8988.6
- Home_Runs                              1    0.7834 19.351 -8987.6
- Hits                                   1    0.8679 19.435 -8979.1
- On_Base_Percentage                     1    8.4698 27.037 -8332.3
```

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

3. Fit the refined linear regression:

```
> 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
```

4. Confusion matrix of logistic regression on training data

```
> print(conf_matrix)
                    Reference
Prediction          Competitive Less Competitive
  Competitive               966                 0
  Less Competitive            0               405
```

5. Confusion matrix of decision tree on testing data

```
> print(conf_matrix_tree)
                    Reference
Prediction          Competitive Less Competitive
  Competitive               966                 0
  Less Competitive            0               405
```

6. Performance metrics for logistic regression on testing data

```
Confusion Matrix and Statistics

                  Reference
Prediction         Competitive Less Competitive
  Competitive              418                 0
  Less Competitive           8               162

               Accuracy : 0.9864
                 95% CI : (0.9734, 0.9941)
    No Information Rate : 0.7245
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9664

 Mcnemar's Test P-Value : 0.01333

            Sensitivity : 0.9812
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9529
             Prevalence : 0.7245
         Detection Rate : 0.7109
   Detection Prevalence : 0.7109
      Balanced Accuracy : 0.9906

       'Positive' Class : Competitive
```

7. Performance metrics of decision tree on testing data

```
Confusion Matrix and Statistics

                Reference
Prediction         Competitive Less Competitive
  Competitive              426                0
  Less Competitive           0              162

              Accuracy : 1
                95% CI : (0.9937, 1)
    No Information Rate : 0.7245
    P-Value [Acc > NIR] : < 2.2e-16

                 Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.7245
         Detection Rate : 0.7245
   Detection Prevalence : 0.7245
      Balanced Accuracy : 1.0000

       'Positive' Class : Competitive
```

8. Summary of refined logistic regression

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.420e+02	3.329e+05	-0.003	0.998
Rank	3.477e+01	4.504e+03	0.008	0.994
Year	5.795e-02	1.683e+02	0.000	1.000
Position1B	-1.502e+01	3.201e+04	0.000	1.000
Position2B	3.259e+01	2.698e+04	0.001	0.999
Position3B	4.698e+01	2.674e+04	0.002	0.999
PositionC	-2.024e+01	1.777e+04	-0.001	0.999
PositionCF	-1.341e+01	2.883e+04	0.000	1.000
PositionCI	-3.901e+01	7.486e+04	-0.001	1.000
PositionDH	-2.139e+01	1.043e+05	0.000	1.000
PositionIF	-6.486e+00	3.617e+04	0.000	1.000
PositionLF	-3.381e+01	2.698e+04	-0.001	0.999
PositionMI	-4.204e+01	5.287e+04	-0.001	0.999
PositionOF	-2.230e+01	2.196e+04	-0.001	0.999
PositionP	-1.833e+01	2.426e+04	-0.001	0.999
PositionRF	4.991e+00	6.878e+04	0.000	1.000
PositionSS	1.469e+00	2.153e+04	0.000	1.000
PositionUT	-5.848e+00	2.089e+04	0.000	1.000
Age	-3.589e-02	6.043e+02	0.000	1.000
Games	-3.093e-02	4.376e+02	0.000	1.000
Plate_Appearances	4.514e+01	1.475e+05	0.000	1.000
At_Bats	-4.523e+01	1.475e+05	0.000	1.000
Runs	-4.959e+00	2.444e+03	-0.002	0.998
Hits	9.505e-02	1.428e+03	0.000	1.000
Doubles	8.924e+00	4.095e+03	0.002	0.998
Triples	1.703e+00	9.499e+03	0.000	1.000
Home_Runs	1.397e+01	7.190e+03	0.002	0.998
Runs_Batted_In	-9.746e-01	1.153e+03	-0.001	0.999
Stolen_Bases	-5.510e-01	4.069e+03	0.000	1.000
Caught_Stealing	-2.738e+00	1.255e+04	0.000	1.000
Base_On_Balls	-4.181e+01	1.478e+05	0.000	1.000
Strikeouts	6.043e-01	4.110e+02	0.001	0.999
On_Base_Percentage	-4.582e+01	9.287e+04	0.000	1.000
On_Base_Plus_Slugging_Percentage	8.699e+00	1.933e+05	0.000	1.000
On_Base_Plus_Slugging_Percentage_Plus	-9.779e-03	6.779e+02	0.000	1.000
Total_Bases	NA	NA	NA	NA
Double_Plays_Grounded_Into	-2.774e-01	2.932e+03	0.000	1.000
Times_Hit_By_Pitch	-4.838e+01	1.468e+05	0.000	1.000
Sacrifice_Hits	-4.595e+01	1.475e+05	0.000	1.000
Sacrifice_Flies	-3.800e+01	1.499e+05	0.000	1.000
Intentional_Bases_on_Balls	2.791e-01	7.324e+03	0.000	1.000
Dominant_HandRight	-1.582e+00	8.402e+03	0.000	1.000
Switch_HitterYes	7.906e+00	1.276e+04	0.001	1.000
Magnitude_Hit	4.201e+01	3.542e+04	0.001	0.999

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.6642e+03 on 1370 degrees of freedom
Residual deviance: 6.6781e-07 on 1328 degrees of freedom
AIC: 86

Number of Fisher Scoring iterations: 25

9. Variable significance for untuned tree

rpart variable importance

only 20 most important variables shown (out of 43)

	Overall
Rank	100.00
PositionP	74.57
Total_Bases	49.62
Runs	47.34
Hits	46.14
At_Bats	0.00
Age	0.00
PositionOF	0.00
PositionCF	0.00
PositionC	0.00
Caught_Stealing	0.00
PositionMI	0.00
Strikeouts	0.00
Magnitude_Hit	0.00
PositionIF	0.00
Triples	0.00
Runs_Batted_In	0.00
Intentional_Bases_on_Balls	0.00
Sacrifice_Hits	0.00
Position1B	0.00

10. Competitiveness class distribution