Homework 2: Data Science 621

Eric Hirsch

3/20/2021

```
knitr::opts_chunk$set(echo = FALSE, warning =  FALSE, message = FALSE)
```

Description of the Dataset

a. ASSIGNMENT: In this assignment we explore, analyze and model a data set containing approximately 2276 records, each representing a professional baseball team from the years 1871 to 2006 inclusive. Each record has the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season.

We will build a multiple linear regression model on the training data to predict the number of wins for the team.

b. THE ISSUE OF HIDDEN GROUPINGS: An issue with the data is hidden groupings. Records may not be independent of each other, as team data in one year will be related to team data in the next year. We know that if some records were adjusted to match a longer season, there may be an "eras of baseball" effect as teams from earlier years behave differently from later ones. Finally, within the record, columns may not be independent. In particular, teams with high offensive stats (like hitting) may have lower defensive stats (like pitching), as the teams on limited budgets make strategic choices between the two. We will attempt to address some of these issues in this analysis.

1. Data Exploration

All of the columns in the dataset are numeric. We begin by examining their means, medians and distributions.

```
##      INDEX         TARGET_WINS      BATTING_H      BATTING_2B
##  Min.   :   1.0   Min.   :  0.00   Min.   : 891   Min.   : 69.0
##  1st Qu.: 630.8   1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0
##  Median :1270.5   Median : 82.00   Median :1454   Median :238.0
##  Mean   :1268.5   Mean   : 80.79   Mean   :1469   Mean   :241.2
##  3rd Qu.:1915.5   3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0
##  Max.   :2535.0   Max.   :146.00   Max.   :2554   Max.   :458.0
##
##    BATTING_3B       BATTING_HR       BATTING_BB      BATTING_SO
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.0   Min.   :   0.0
##  1st Qu.: 34.00   1st Qu.: 42.00   1st Qu.:451.0   1st Qu.: 548.0
##  Median : 47.00   Median :102.00   Median :512.0   Median : 750.0
##  Mean   : 55.25   Mean   : 99.61   Mean   :501.6   Mean   : 735.6
##  3rd Qu.: 72.00   3rd Qu.:147.00   3rd Qu.:580.0   3rd Qu.: 930.0
##  Max.   :223.00   Max.   :264.00   Max.   :878.0   Max.   :1399.0
```

```
##                                                  NA's   :102
##    BASERUN_SB      BASERUN_CS     BATTING_HBP      PITCHING_H
##  Min.   :  0.0   Min.   :  0.0   Min.   :29.00   Min.   : 1137
##  1st Qu.: 66.0   1st Qu.: 38.0   1st Qu.:50.50   1st Qu.: 1419
##  Median :101.0   Median : 49.0   Median :58.00   Median : 1518
##  Mean   :124.8   Mean   : 52.8   Mean   :59.36   Mean   : 1779
##  3rd Qu.:156.0   3rd Qu.: 62.0   3rd Qu.:67.00   3rd Qu.: 1682
##  Max.   :697.0   Max.   :201.0   Max.   :95.00   Max.   :30132
##  NA's   :131     NA's   :772     NA's   :2085
##   PITCHING_HR     PITCHING_BB      PITCHING_SO        FIELDING_E
##  Min.   :  0.0   Min.   :   0.0   Min.   :    0.0   Min.   :  65.0
##  1st Qu.: 50.0   1st Qu.: 476.0   1st Qu.:  615.0   1st Qu.: 127.0
##  Median :107.0   Median : 536.5   Median :  813.5   Median : 159.0
##  Mean   :105.7   Mean   : 553.0   Mean   :  817.7   Mean   : 246.5
##  3rd Qu.:150.0   3rd Qu.: 611.0   3rd Qu.:  968.0   3rd Qu.: 249.2
##  Max.   :343.0   Max.   :3645.0   Max.   :19278.0   Max.   :1898.0
##                                   NA's   :102
##   FIELDING_DP
##  Min.   : 52.0
##  1st Qu.:131.0
##  Median :149.0
##  Mean   :146.4
##  3rd Qu.:164.0
##  Max.   :228.0
##  NA's   :286
```

We note that a number of columns have NAs. Batting_SO and Pitching_SO have the same number of NA's and may be related.

We more closely examine the distribution of columns in the dataset (fig. 1):

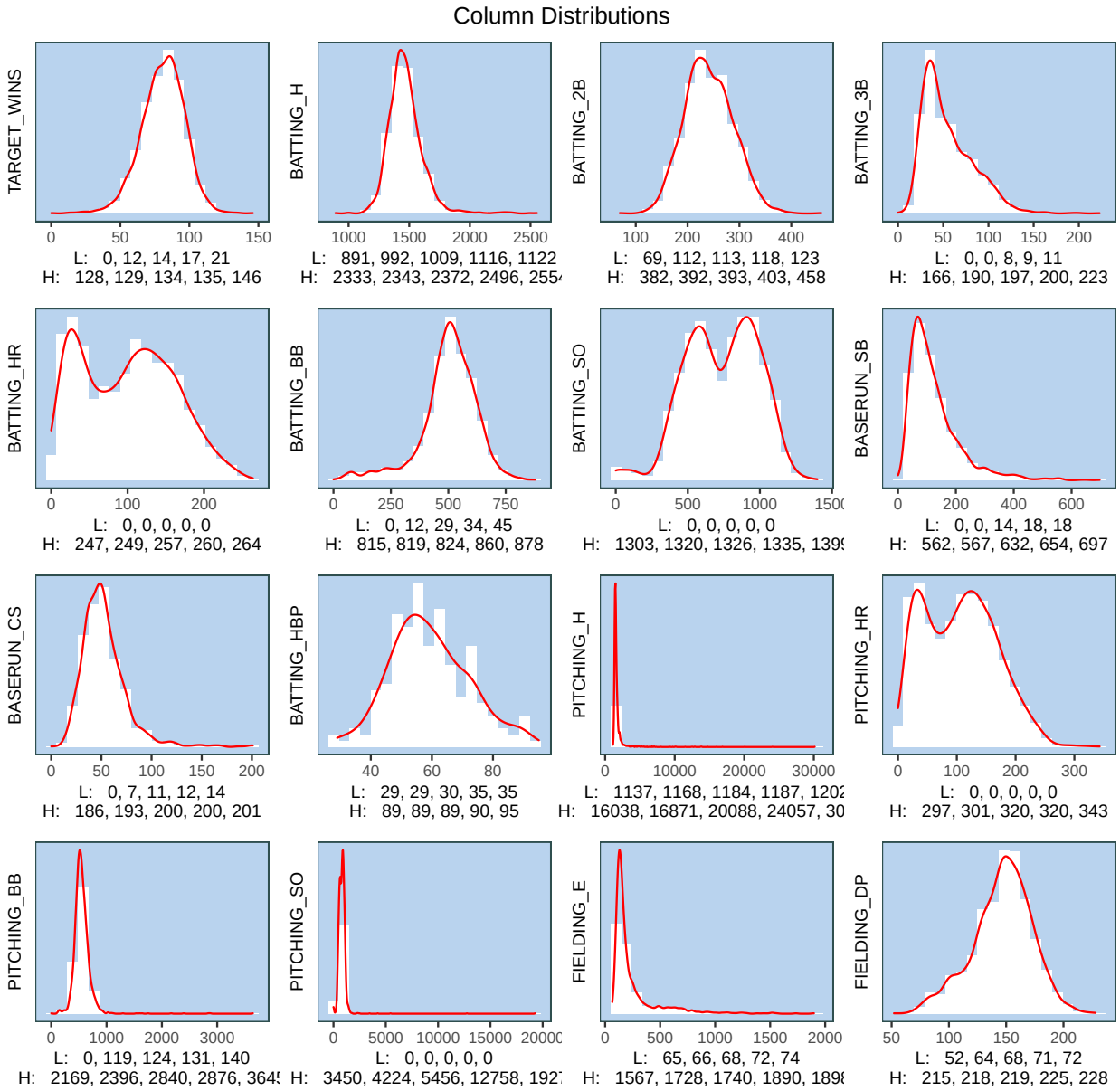


Fig. 1

Our dependent variable (Target Wins) appears to be normally distributed. However, a number of columns are severely skewed (Errors, Strikeouts, Pitching_H, etc.) A few columns (Batting SO, Pitching_HR and Batting_HR) have a bimodal distribution. This might point to some hidden groupings in the dataset.

Boxplots help us identify outliers (fig. 2):

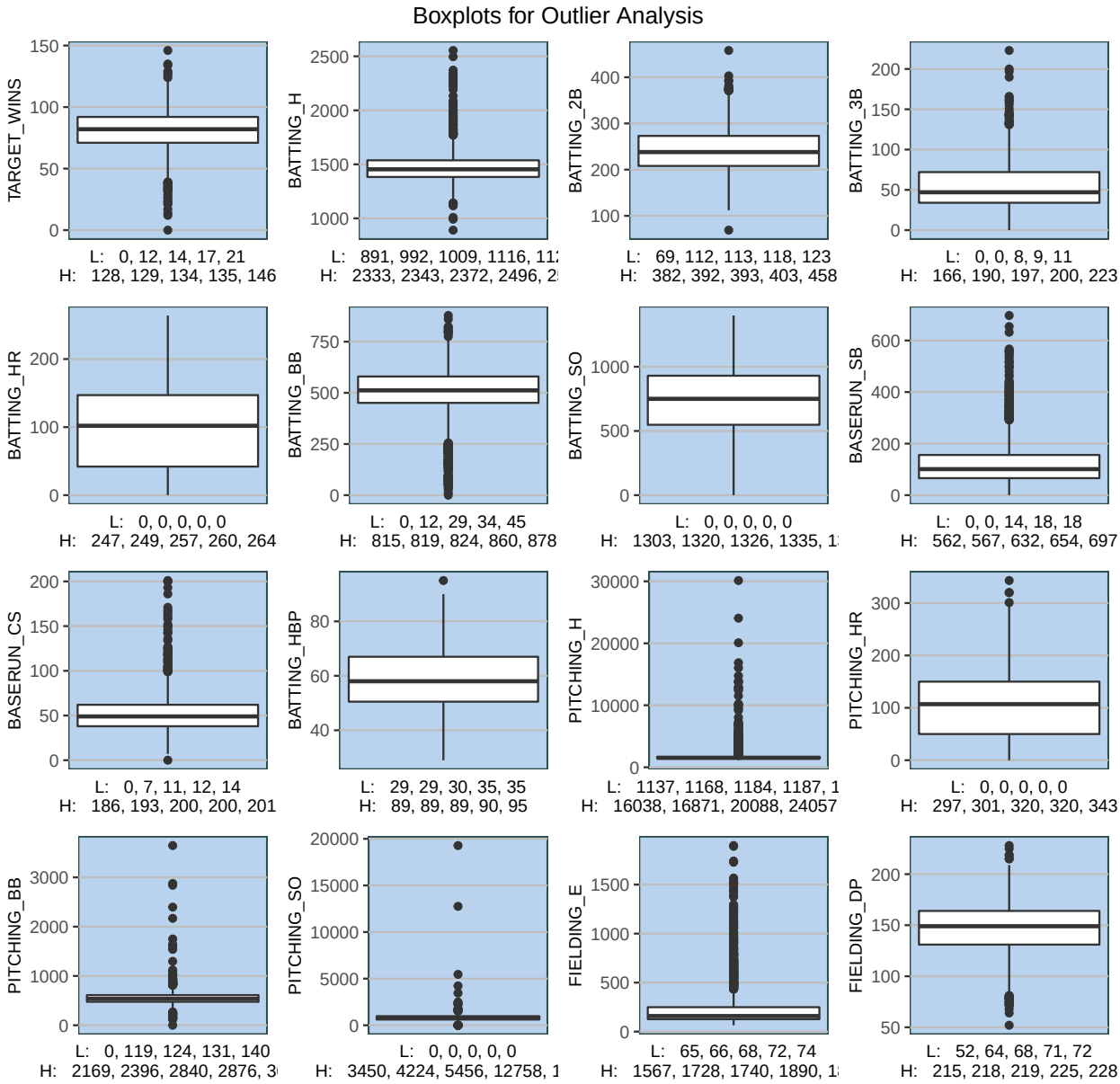


Fig. 2

There a number of outliers, both high and low. For example, there are many zeros, which may be implausible. In addition, many of the ranges appear extreme, such as giving up between 3,500 hits and 19,000 hits, or getting from 12 to over 800 walks.

We investigate correlations in the dataset, both between the dependent variable and the other variables (fig. 3), and between the dependent variables and each other (fig. 4).

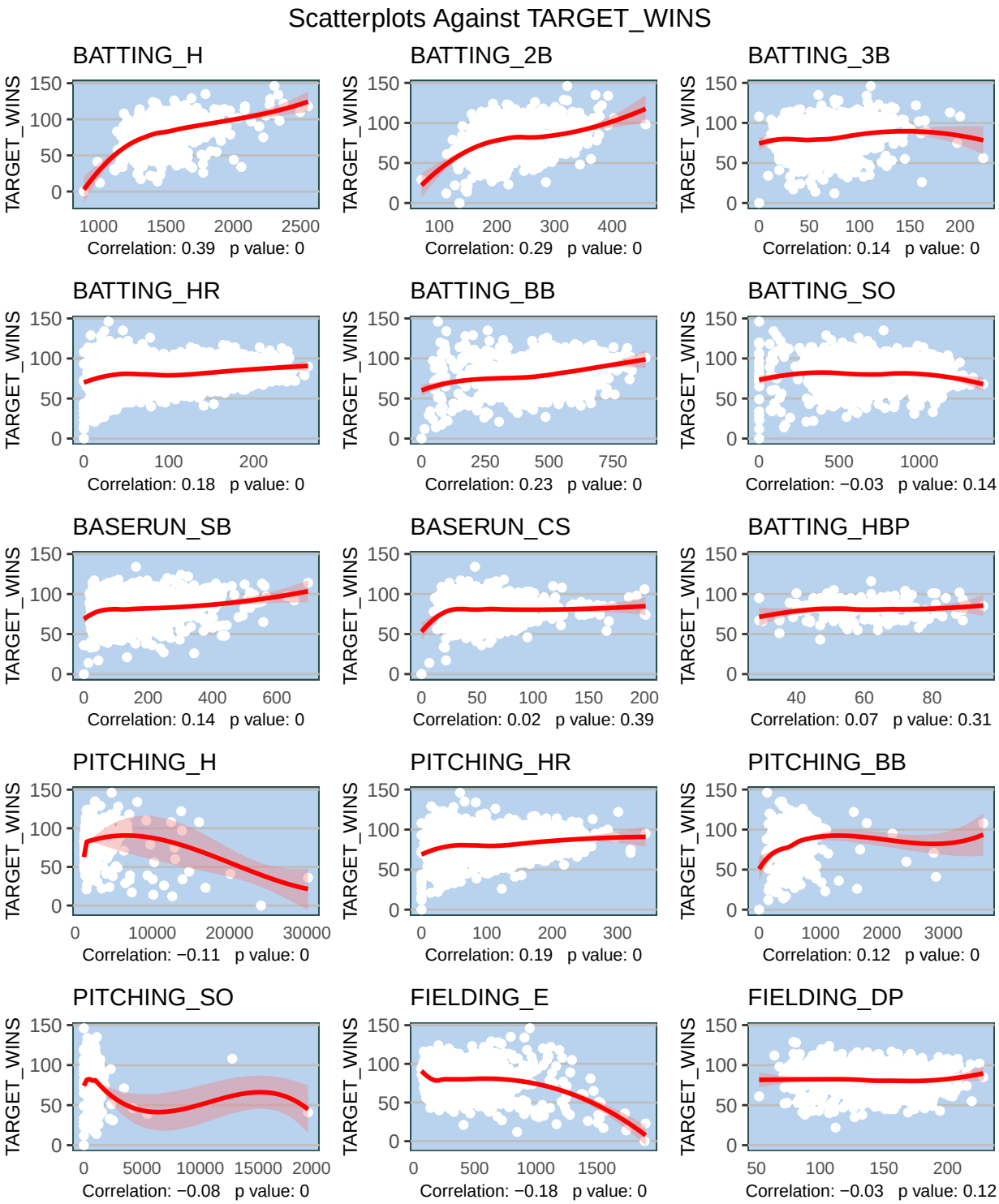


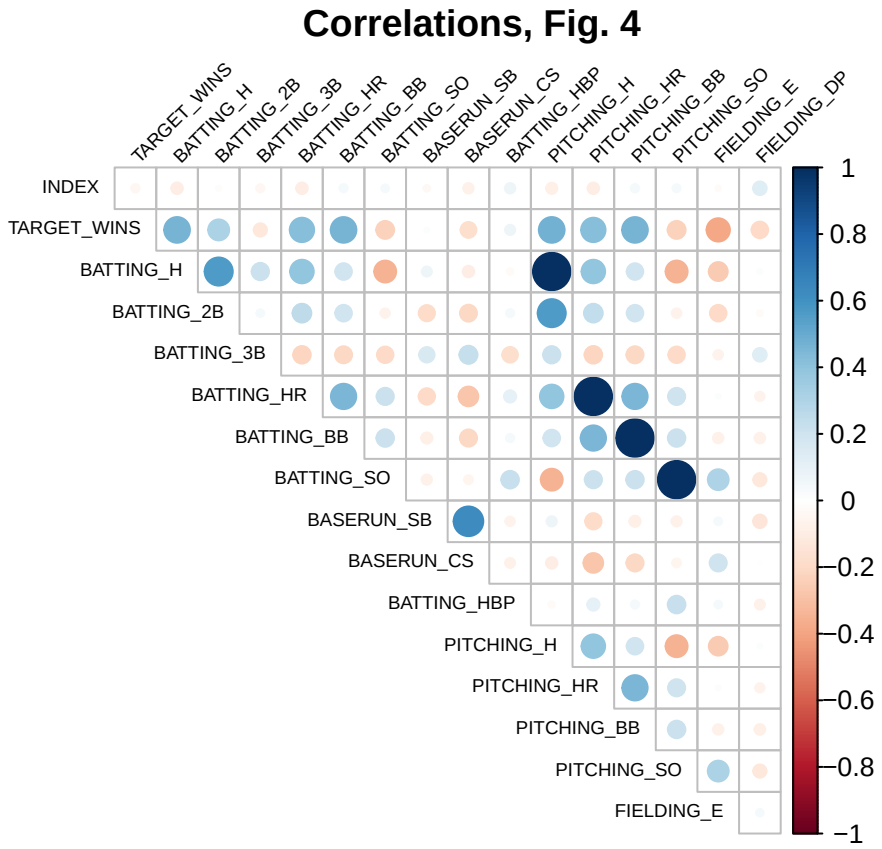
Fig.3

Here we see a number of puzzles, mainly among the pitching correlations. Hits should show a much stronger negative correlation, and in fact appear positive for a portion. Making double plays is surprisingly neutral, as are strikeouts. Pitching_HR is also positive when we would expect negative.

We do need to acknowledge here the possibility of strategy groupings (defense and offense) which may contribute to these anomalies. In other words, a team with poor pitching may have strong hitting, which

then wins games.

We can look for evidence of this possibility by examining multicollinearity:



Indeed, the pitching categories are strongly correlated with their hitting counterparts. All four of the pitching categories follow this pattern.

2. Data Preparation

We begin by devising a strategy for the NAs. We can eliminate the Batting_HBP and Baserun_CS columns because they have too many NA's. We also create flags for the other columns with significant NA's.

We are particularly interested in the SO columns because they do not appear random, and investigation establishes that they have complete overlap with each other. Fielding_DP and BR_SB aslo have some overlap. These may relate to eras of baseball when certain statistics were not collected. (see Fig. 5)

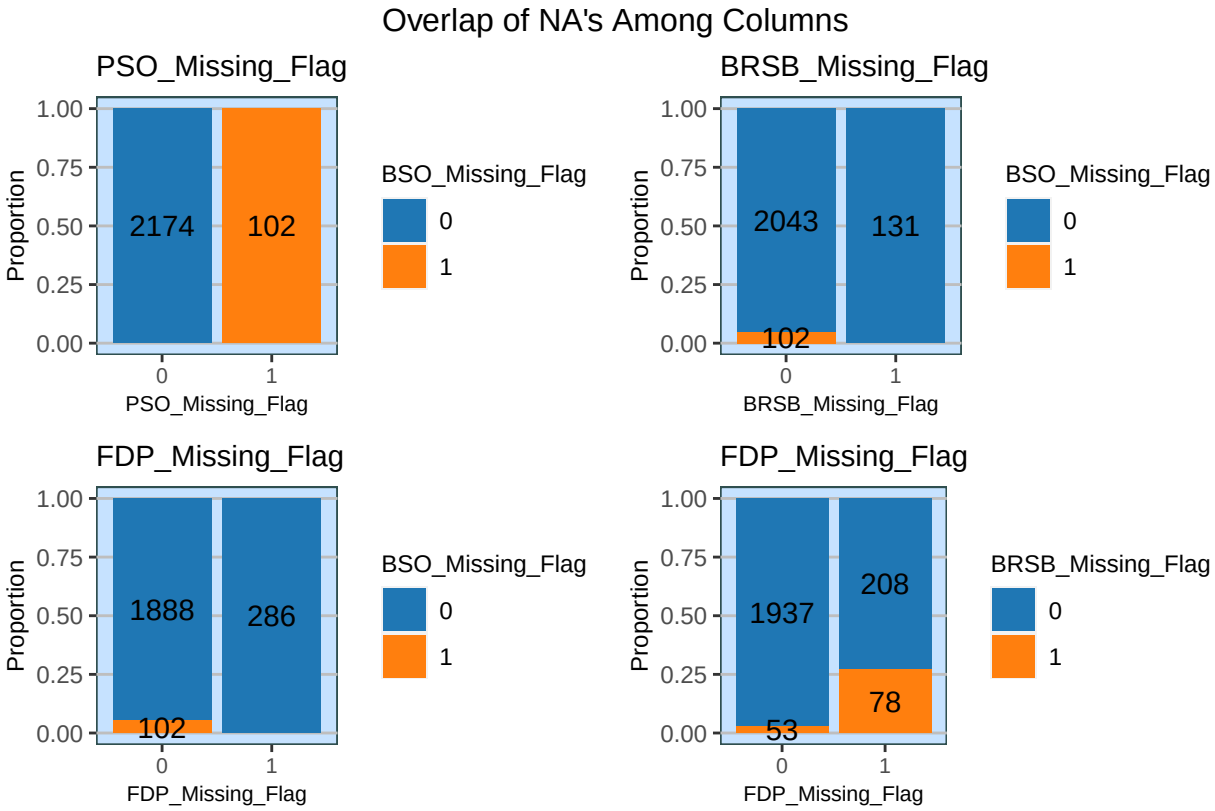


Fig. 5

We eliminate the pitching SO column because it is redundant. While not MCAR (missing completely at random), if the Batting_SO column is MAR (missing at random), we may be bale to eliminate these rows, as there are not so many (5% of the total).

One way to investigate the randomness of this missing cohort is to look for interactions between the cohort and other dataset columns. In fact, we see that there are a number of columns with strong, even extreme interactions (see fig. 6).

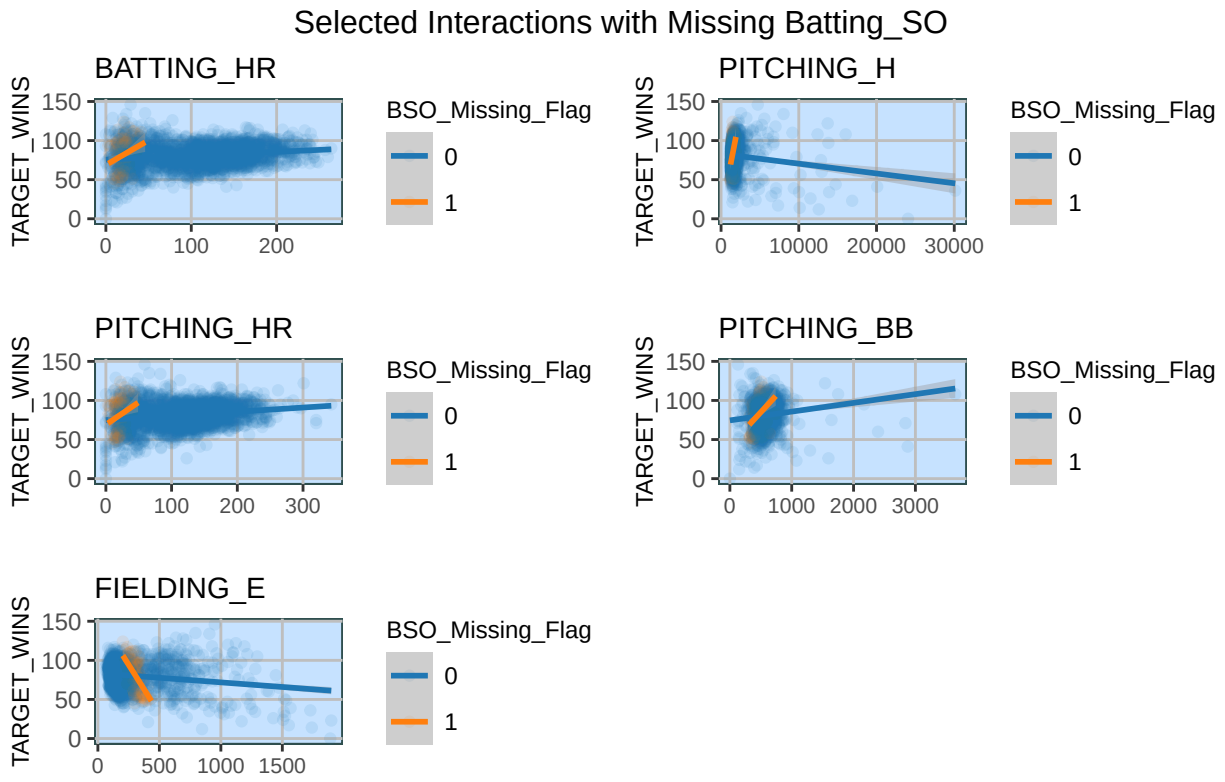


Fig. 6

It is possible this cohort represents a different baseball era when such statistics were not collected. In any case, we cannot eliminate these rows without losing critical data, so we employ the following strategy: 1) retain the rows and impute a value, 2) create a "missing" flag to keep track of the cohort, and 2) add interaction terms where appropriate.

Before we address imputation, we want to work with the implausible zeros in the dataet. In particular, we note that the 0s in Pitching_SO and Batting_SO are a complete overlap, and we can see from the histograms that the jump between 0 and the next lowest value is not smooth, and so we will treat them as NA's. We do the same with batting and pitching HR, since there is also a jump up after zero which suggests it is being used as an indicator of missing value.

Just so we have some reasonable criteria for imputation strategy, we compare the r-squared of three regressions - with NA's imputed as means, with NA's imputed as medians, and with NA rows eliminated altogether.

```
## [1] "type:" "mean"
## [1] "r2mean:" "0.4031"
## [1] "r2median:" "0.403"
## [1] "r2omit" "0.4019"
```

The mean and median have the same r-squared, while the elimination of the rows has a smaller r-squared. We therefore choose to impute the mean.

Not surprisingly, the evaluation dataset shows the same results:

```
## [1] "type:" "mean"
## [1] "r2mean:" "0.4031"
```

```
## [1] "r2median:" "0.403"
## [1] "r2omit" "0.4019"
```

Although outliers and possible bad data appear in a number of places, without domain knowledge we are reluctant to eliminate any other outliers or influential points at this point without good reason. We don't know if extreme numbers are necessarily implausible. Therefore the outliers will remain.

3. Data Transformation

1. We create a flag for hits under 1500

As previously noted, Pitching_H is surprisingly weak in it's relationship to wins, and in fact appears positive for a large portion of its distribution. We examine more closely the relationship between pitching hits and wins, paying particular attention to the portion of the relationship where hits are below 3,000 (fig. 7).

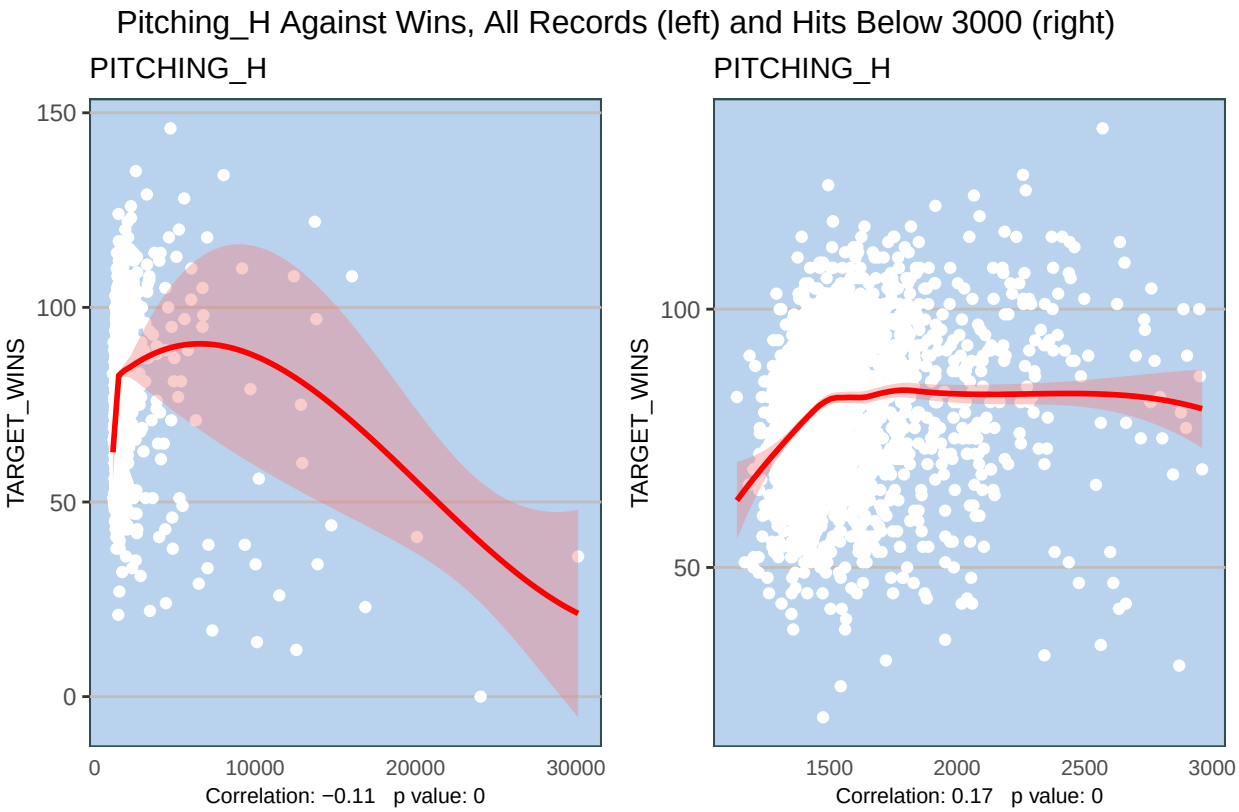


Fig.7

We can see here the positive correlation between pitching_h and wins. While we can't explain the phenomenon, we can account for it statistically by adding a binary flag for records with hits under 1500.

2. We create an interaction between Fielding_DP and hits.

The Fielding_DP correlation with Target Wins is surprising, since making double plays should help a team win. On the other hand, a team that makes double plays is also a team that gives up hits.

We therefore create an interaction term for Fielding_DP and Pitching_H.

4. We create a flag to account for the bimodal distribution of Batting HR.

Batting HR has a bimodal distribution (see Fig. 9). We don't explain this, but speculate that it may be related to different eras of baseball. Therefore, we create a flag to separate records with less than 80 HR form those with more.

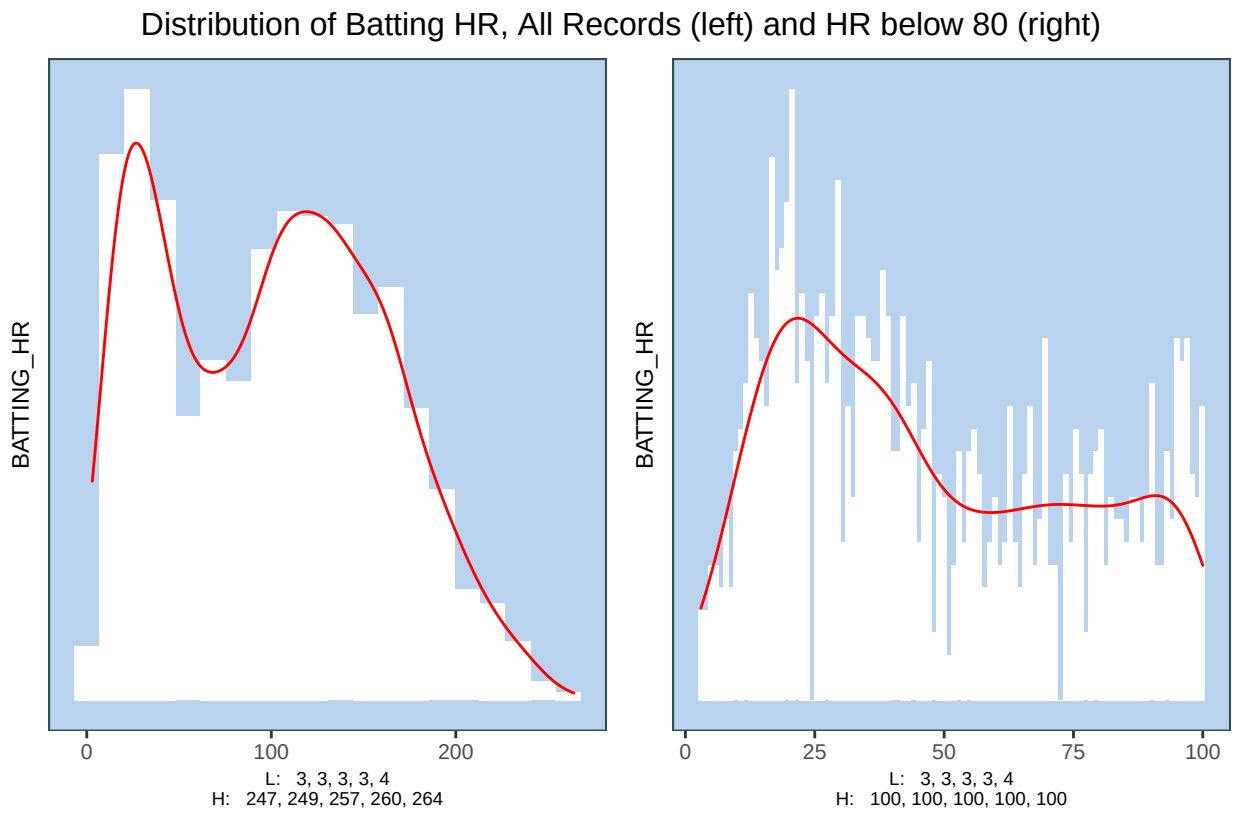


Fig.9

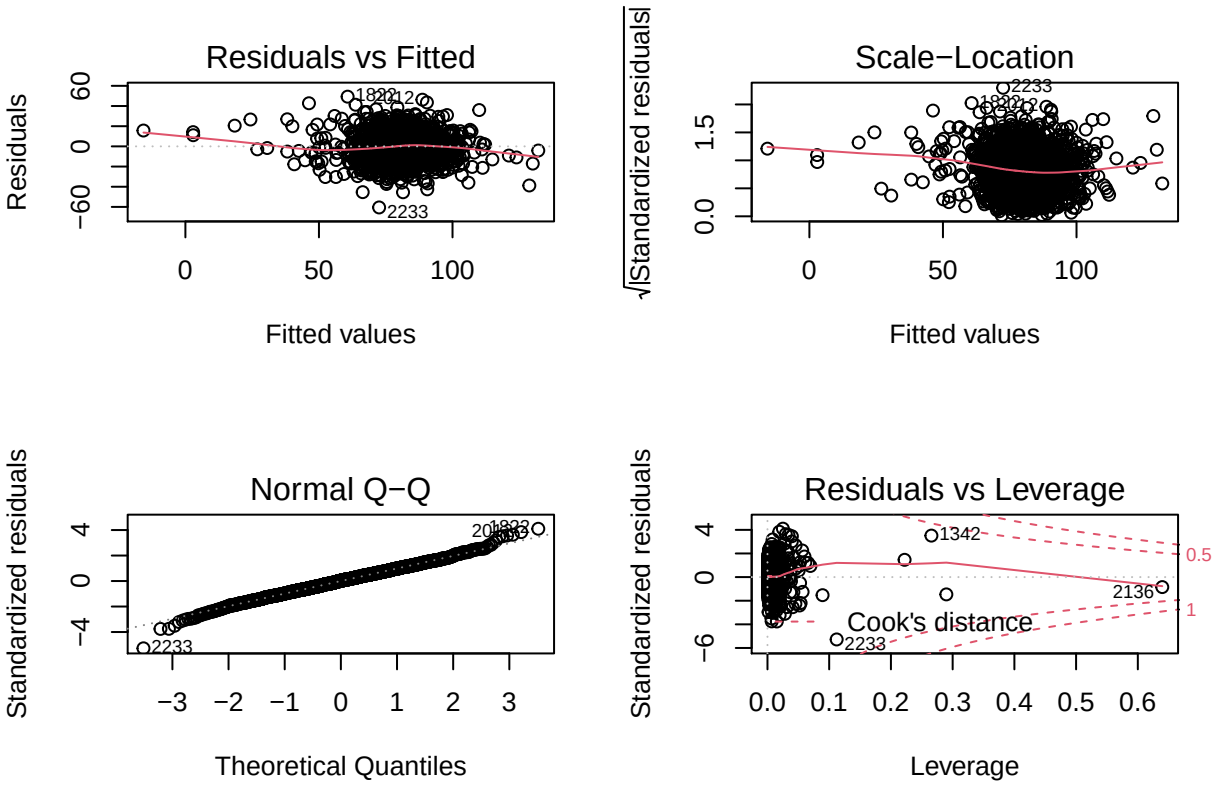
```
## [1] 0.02231354
```

```
## [1] 0.03503598
```

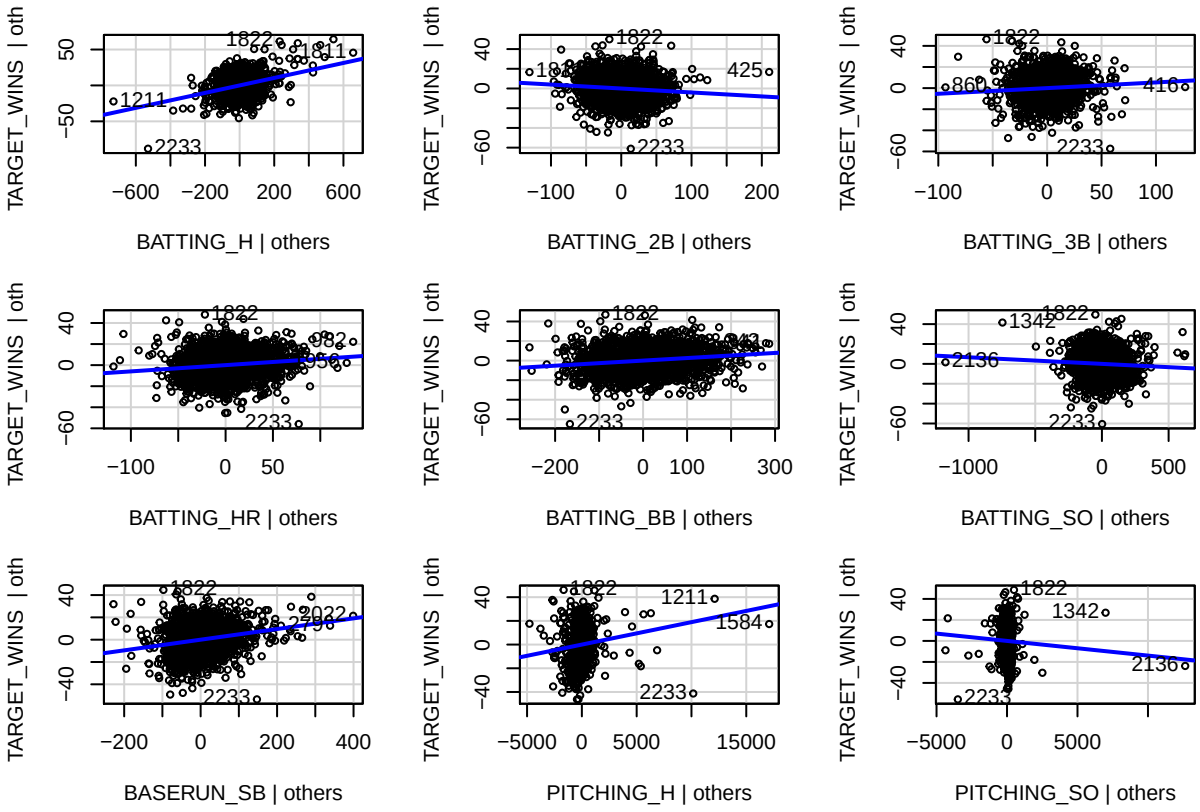
4. Data Modeling

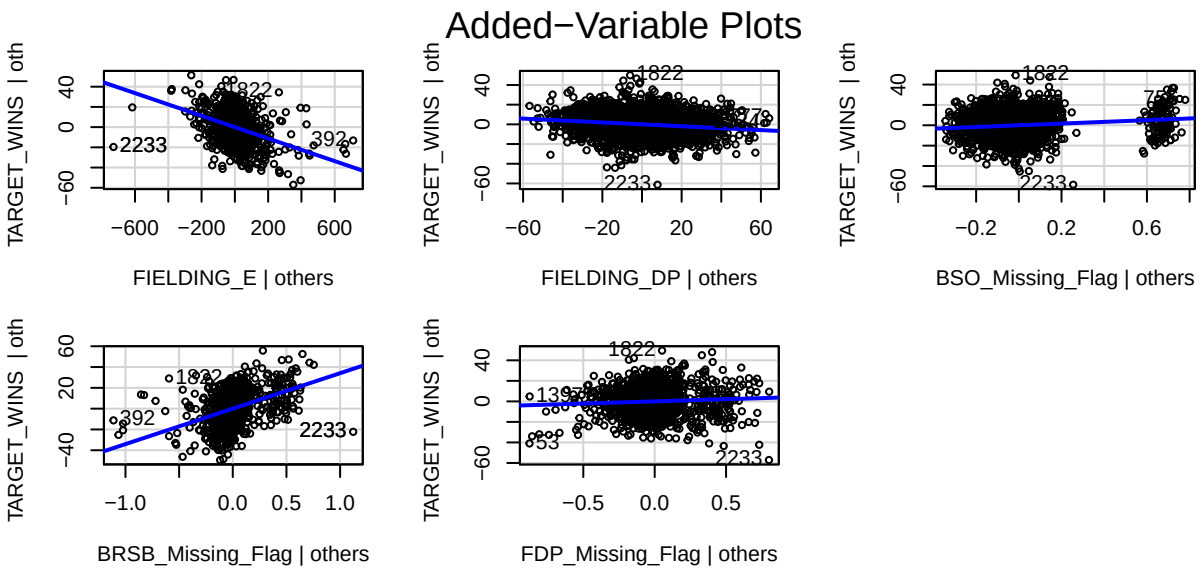
```
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B +
##     BATTING_HR + BATTING_BB + BATTING_SO + BASERUN_SB + PITCHING_H +
##     PITCHING_SO + FIELDING_E + FIELDING_DP + BSO_Missing_Flag +
##     BRSB_Missing_Flag + FDP_Missing_Flag, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.531  -8.063   0.330   8.075  49.266
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       13.7948052  5.0143117   2.751  0.00599 **
## BATTING_H          0.0521109  0.0033520  15.546  < 2e-16 ***
```

```
## BATTING_2B        -0.0401259  0.0086621  -4.632 3.82e-06 ***
## BATTING_3B         0.0537762  0.0158617   3.390  0.00071 ***
## BATTING_HR         0.0595856  0.0089648   6.647 3.75e-11 ***
## BATTING_BB         0.0260490  0.0032618   7.986 2.20e-15 ***
## BATTING_SO        -0.0066440  0.0022278  -2.982  0.00289 **
## BASERUN_SB         0.0477764  0.0046194  10.343  < 2e-16 ***
## PITCHING_H         0.0018926  0.0003398   5.569 2.86e-08 ***
## PITCHING_SO       -0.0013966  0.0006654  -2.099  0.03593 *
## FIELDING_E        -0.0560670  0.0033748 -16.613  < 2e-16 ***
## FIELDING_DP       -0.0969459  0.0134629  -7.201 8.10e-13 ***
## BSO_Missing_Flag   8.3474206  1.4721894   5.670 1.61e-08 ***
## BRSB_Missing_Flag 34.1064444  1.8484454  18.451  < 2e-16 ***
## FDP_Missing_Flag   4.2303099  1.4669785   2.884  0.00397 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.17 on 2261 degrees of freedom
## Multiple R-squared:  0.4068, Adjusted R-squared:  0.4031
## F-statistic: 110.7 on 14 and 2261 DF,  p-value: < 2.2e-16
##
## [1] "VIF Analysis"
##         BATTING_H        BATTING_2B        BATTING_3B        BATTING_HR
##          3.608349          2.524443          3.016545          4.444255
##        BATTING_BB        BATTING_SO        BASERUN_SB        PITCHING_H
##          2.459248          4.131567          2.380761          3.511045
##       PITCHING_SO        FIELDING_E       FIELDING_DP  BSO_Missing_Flag
##          1.946738          9.076248          1.674220          1.425731
## BRSB_Missing_Flag  FDP_Missing_Flag
##          2.848146          3.633432
```




```
## NULL
##
##  studentized Breusch-Pagan test
##
## data:  step3
## BP = 306.77, df = 14, p-value < 2.2e-16
##
##
##  Shapiro-Wilk normality test
##
## data:  step3$residuals
## W = 0.99701, p-value = 0.0002005
```


TARGET_WINS | others
BATTING_H | others
BATTING_2B | others
BATTING_3B | others
BATTING_HR | others
BATTING_BB | others
BATTING_SO | others
BASERUN_SB | others
PITCHING_H | others
PITCHING_SO | others
1822
1211
2233
425
1342
2136
1584



5. Model Selection

6. Predictions

7. Conclusion