Moneyball - CUNY Data Science 621

Eric Hirsch

2/20/2021

Description of the Dataset

xxxxxx

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

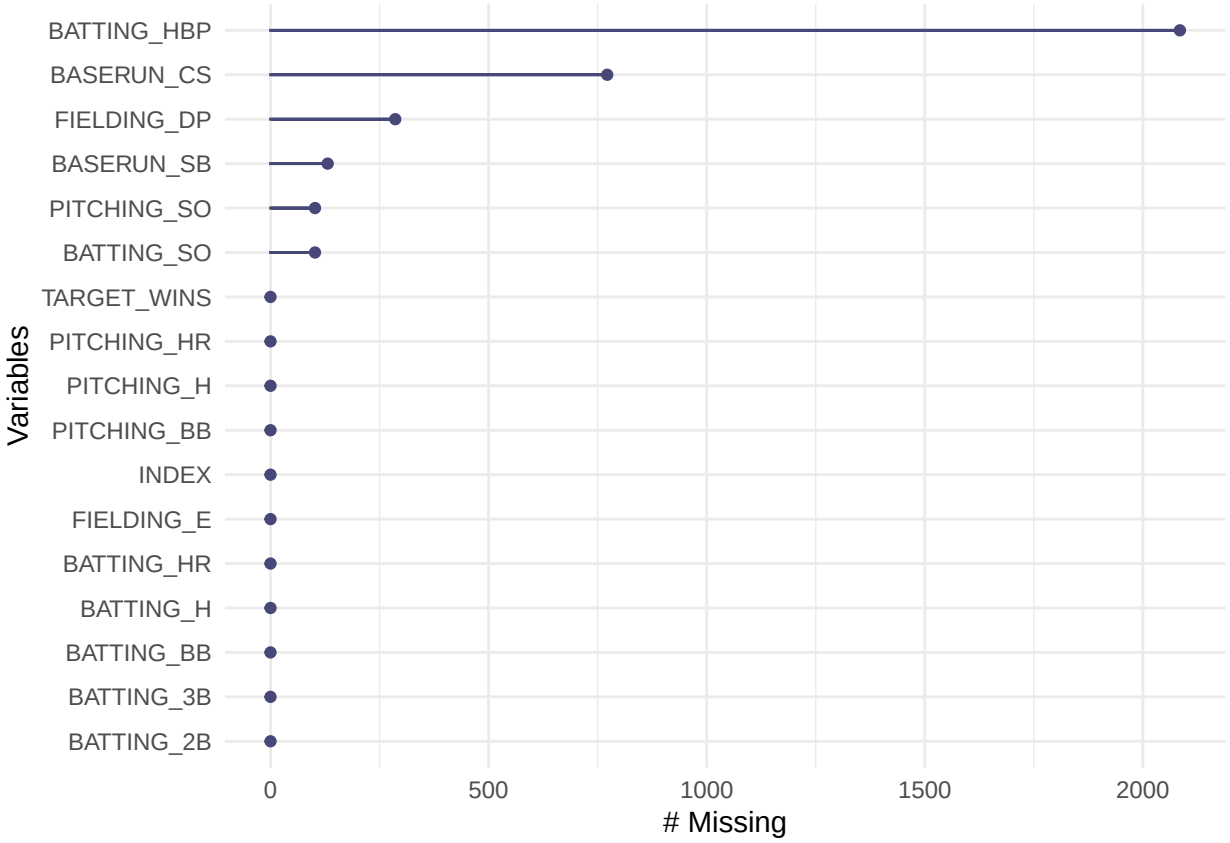
All of the columns in the dataset are numeric. We begin by examining their means, medians and distributions.

```
##      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
```

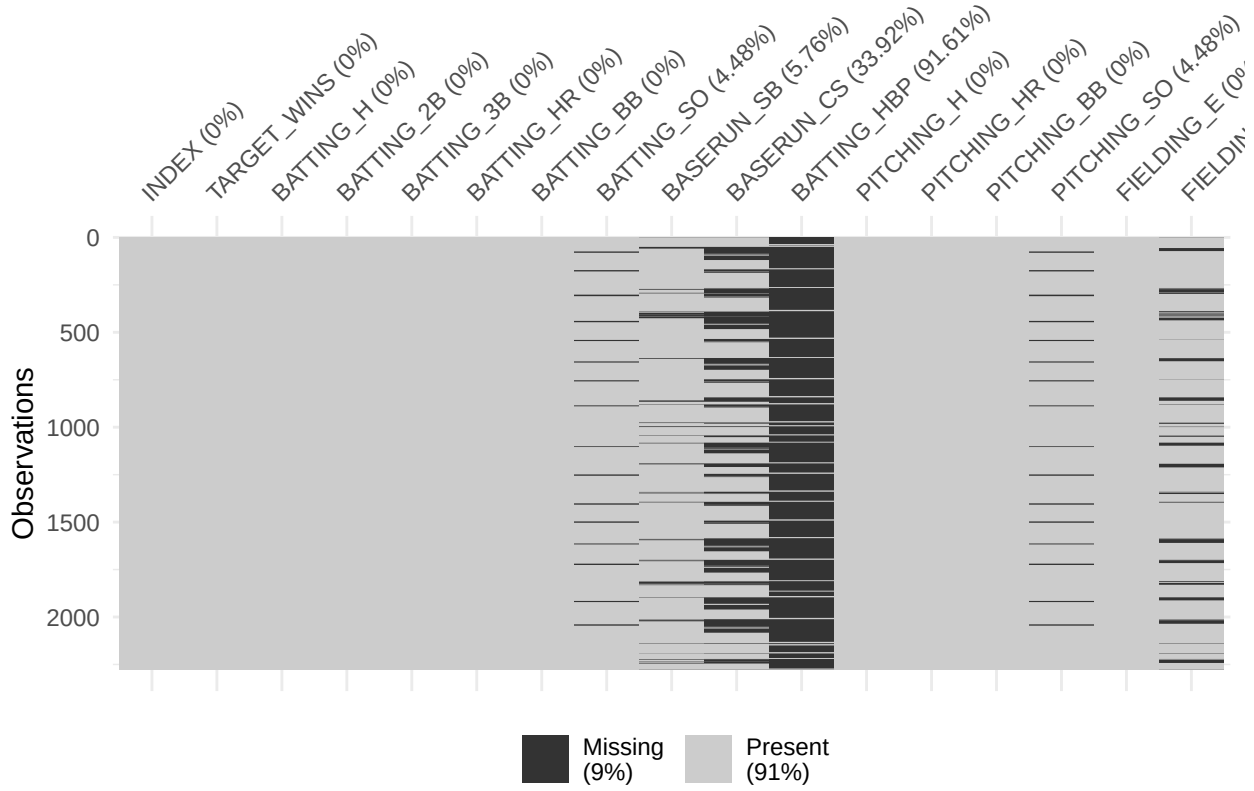
```
##  1st Qu.: 50.0   1st Qu.: 476.0   1st Qu.:  615.0   1st Qu.: 127.0
##  Median :107.0   Median : 536.5   Median :  813.5   Median : 159.0
##  Mean   :105.7   Mean   : 553.0   Mean   :  817.7   Mean   : 246.5
##  3rd Qu.:150.0   3rd Qu.: 611.0   3rd Qu.:  968.0   3rd Qu.: 249.2
##  Max.   :343.0   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.

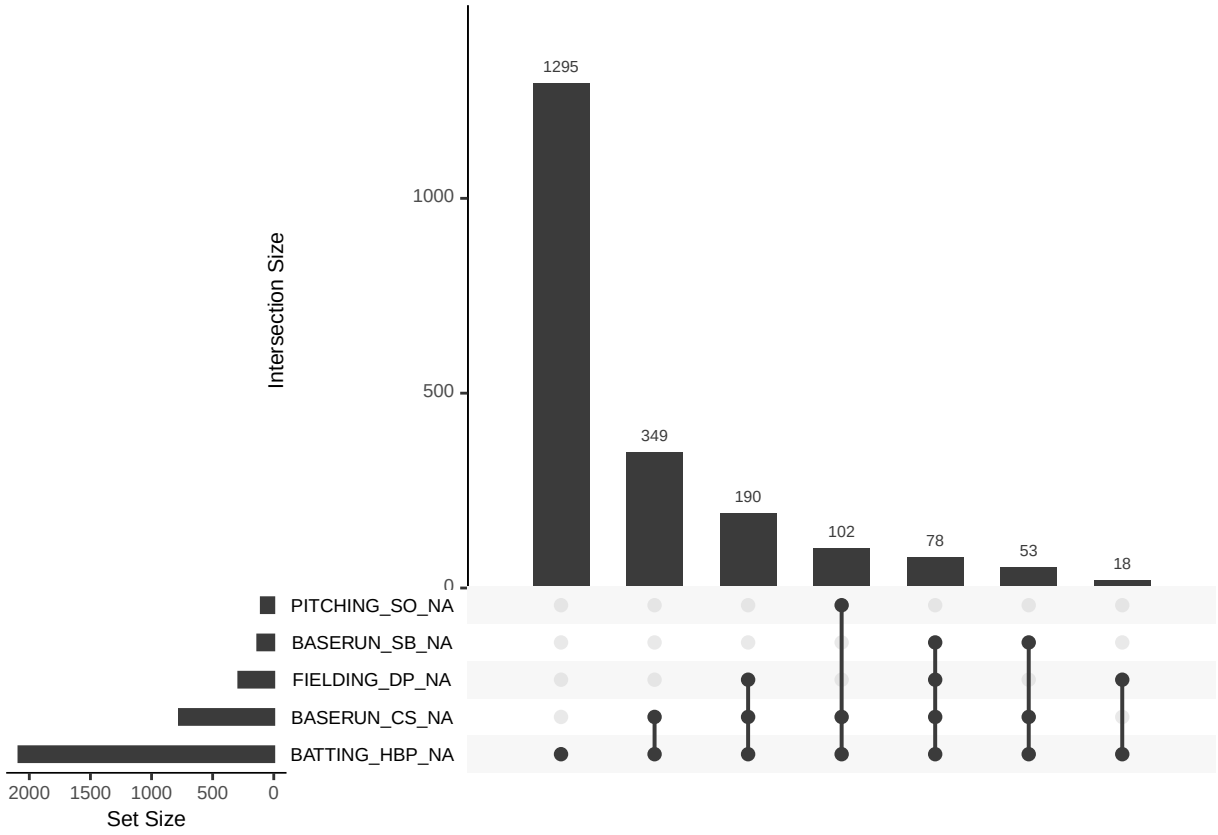
```
## [[1]]
```




```
##
## [[2]]
```


INDEX (0%)
TARGET_WINS (0%)
BATTING_H (0%)
BATTING_2B (0%)
BATTING_3B (0%)
BATTING_HR (0%)
BATTING_BB (0%)
BATTING_SO (4.48%)
BASERUN_SB (5.76%)
BASERUN_CS (33.92%)
BATTING_HBP (91.61%)
PITCHING_H (0%)
PITCHING_HR (0%)
PITCHING_BB (0%)
PITCHING_SO (4.48%)
FIELDING_E (0%)
Observations
0
500
1000
1500
2000
Missing (9%)
Present (91%)

```
##
## [[3]]
```



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

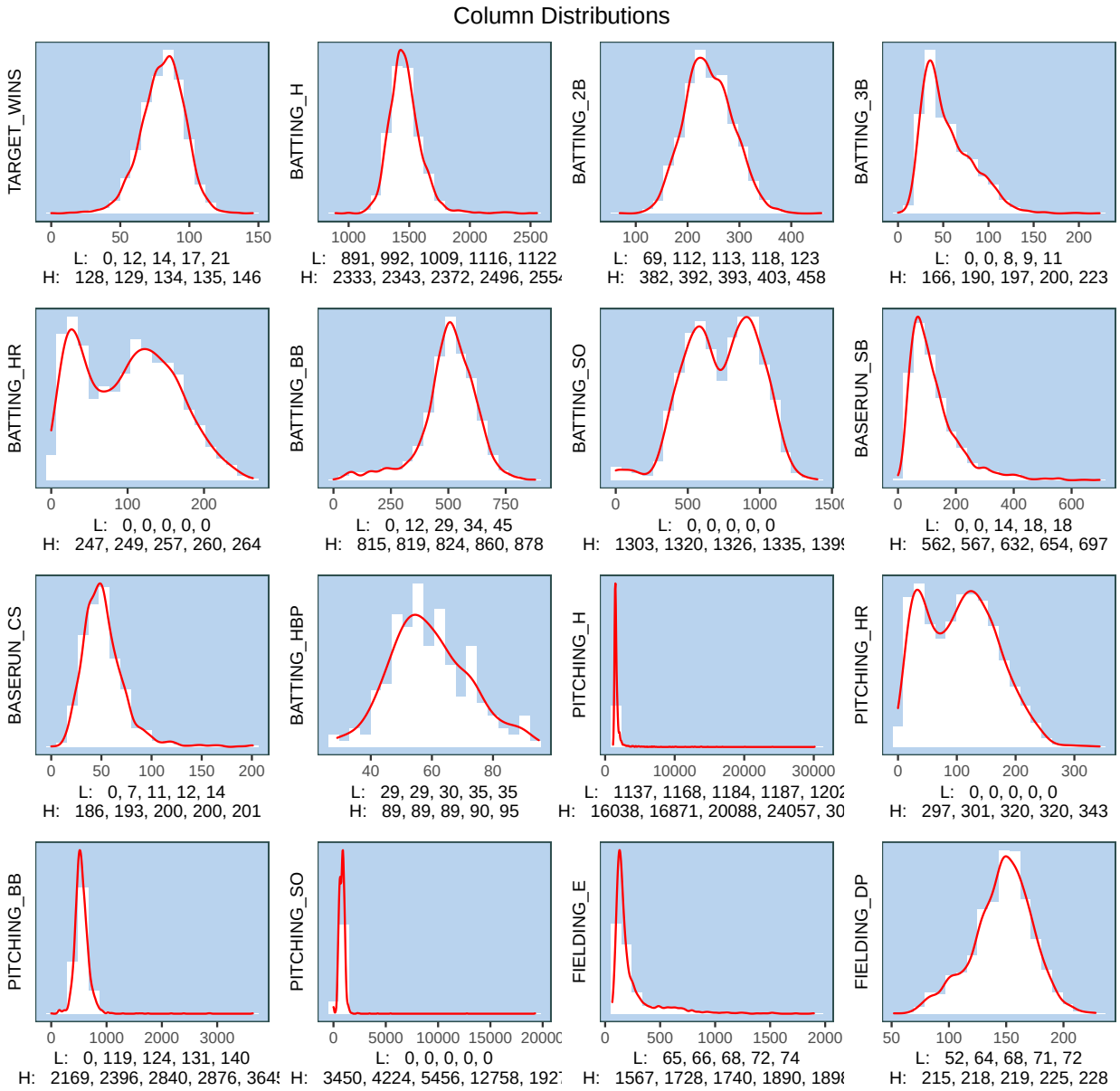


Fig. 1

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):

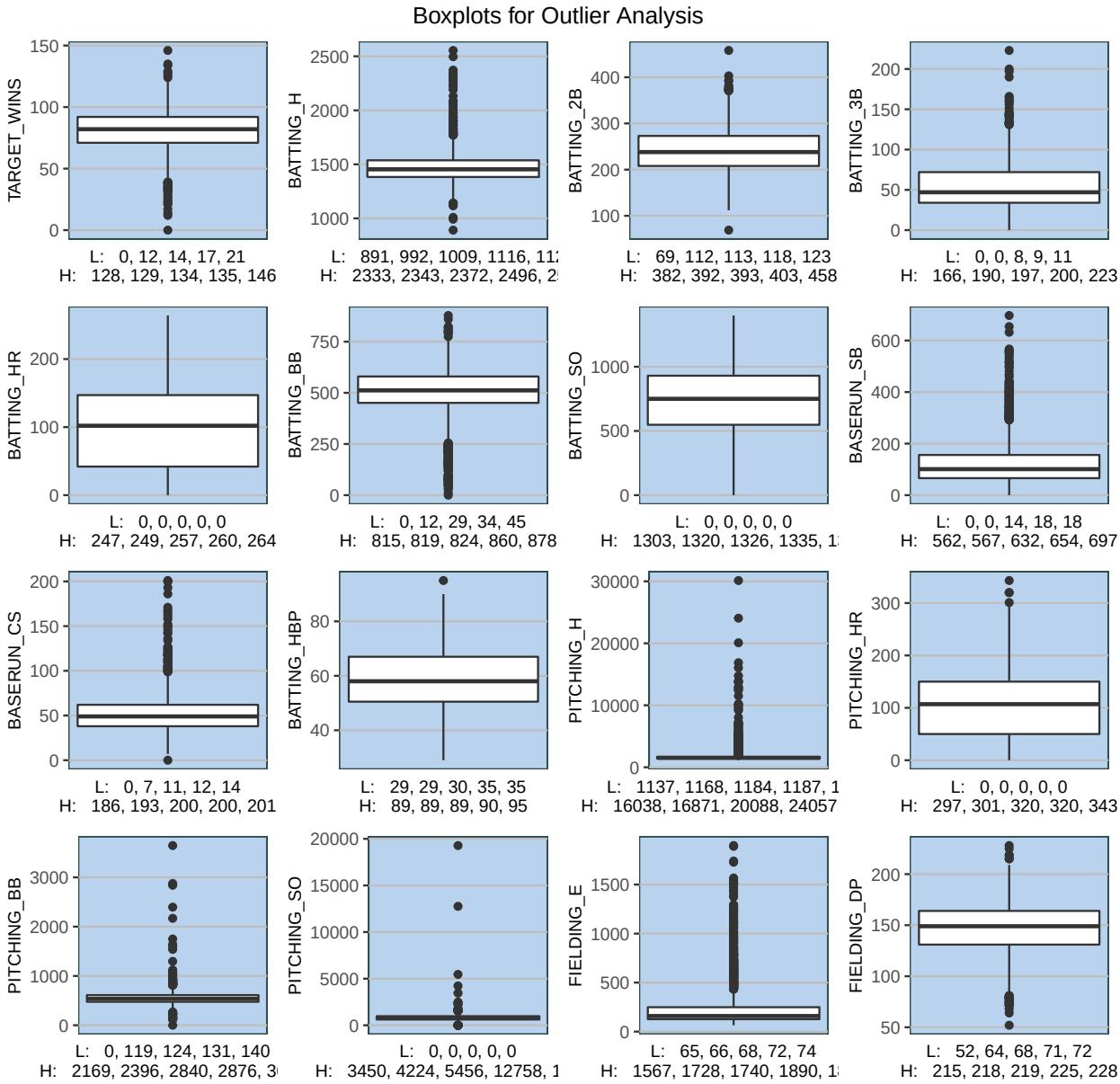


Fig. 2

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).

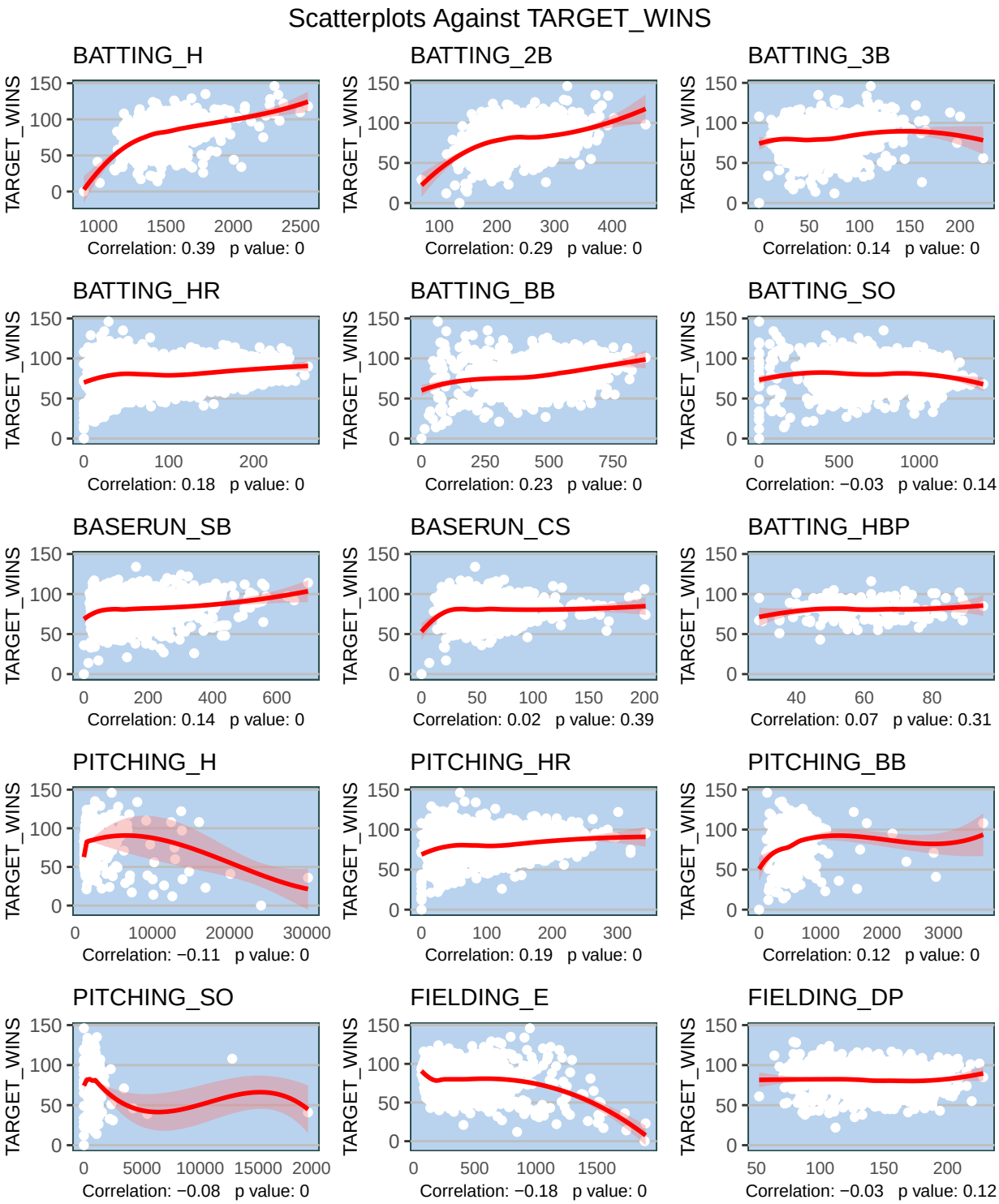


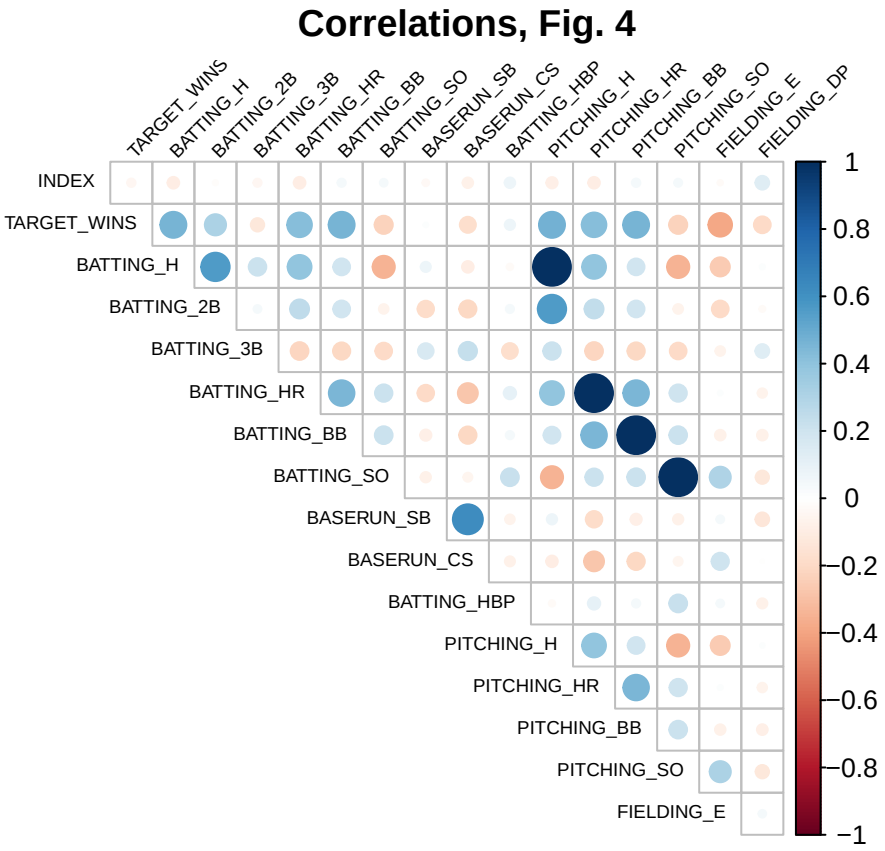
Fig.3

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 picthing 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 and significantly overlap Baserun_SB as well. While not MCAR (missing completely at random), if they are nontheless MAR (missing at random), we can simply 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, there are a number of columns with strong, even extreme interactions (see fig. 5).

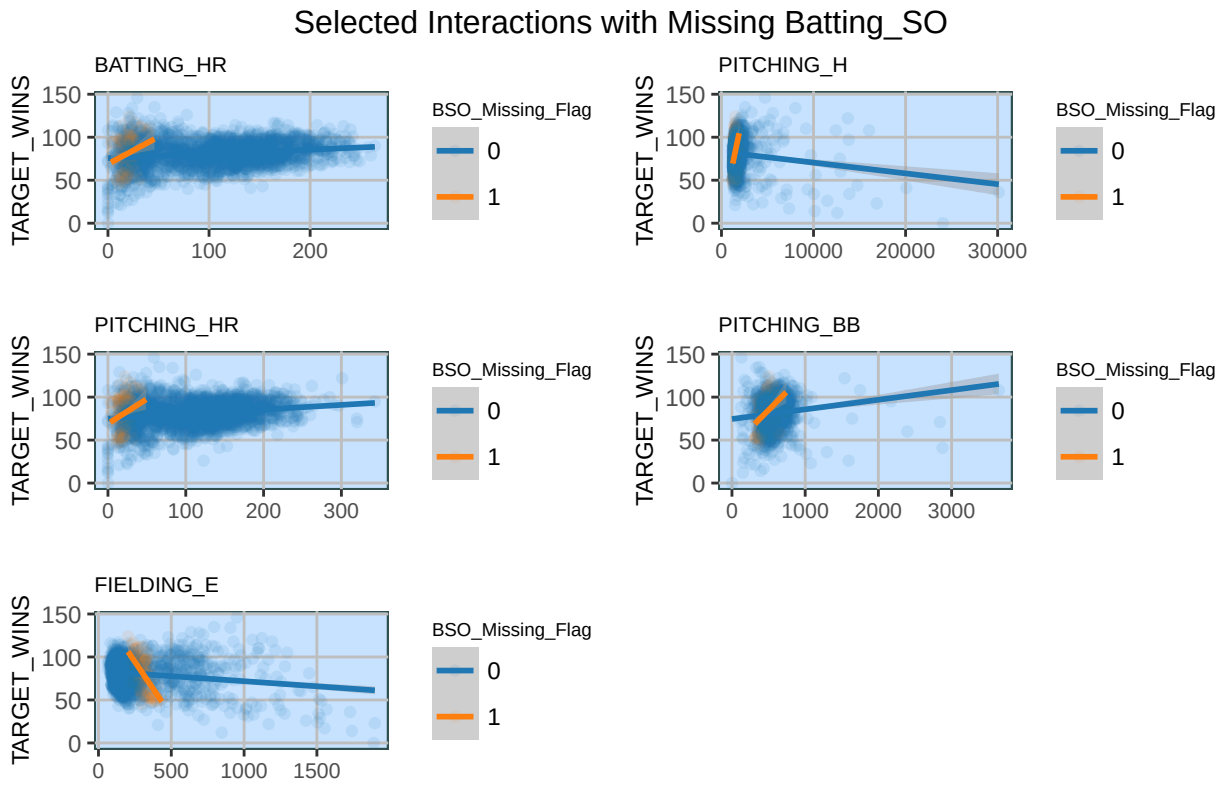


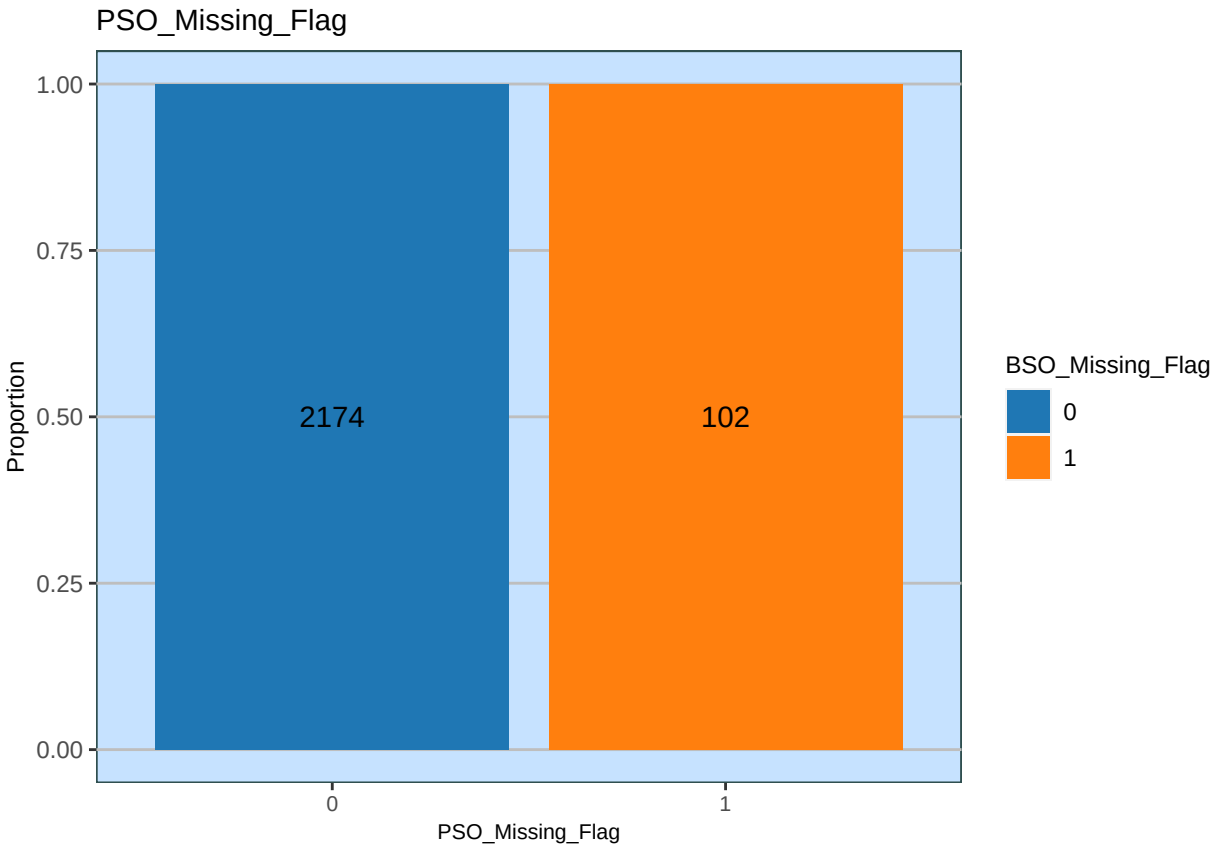
Fig. 5

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) retain a "missing" flag to keep track of the cohort, and 2) add interaction terms where appropriate.

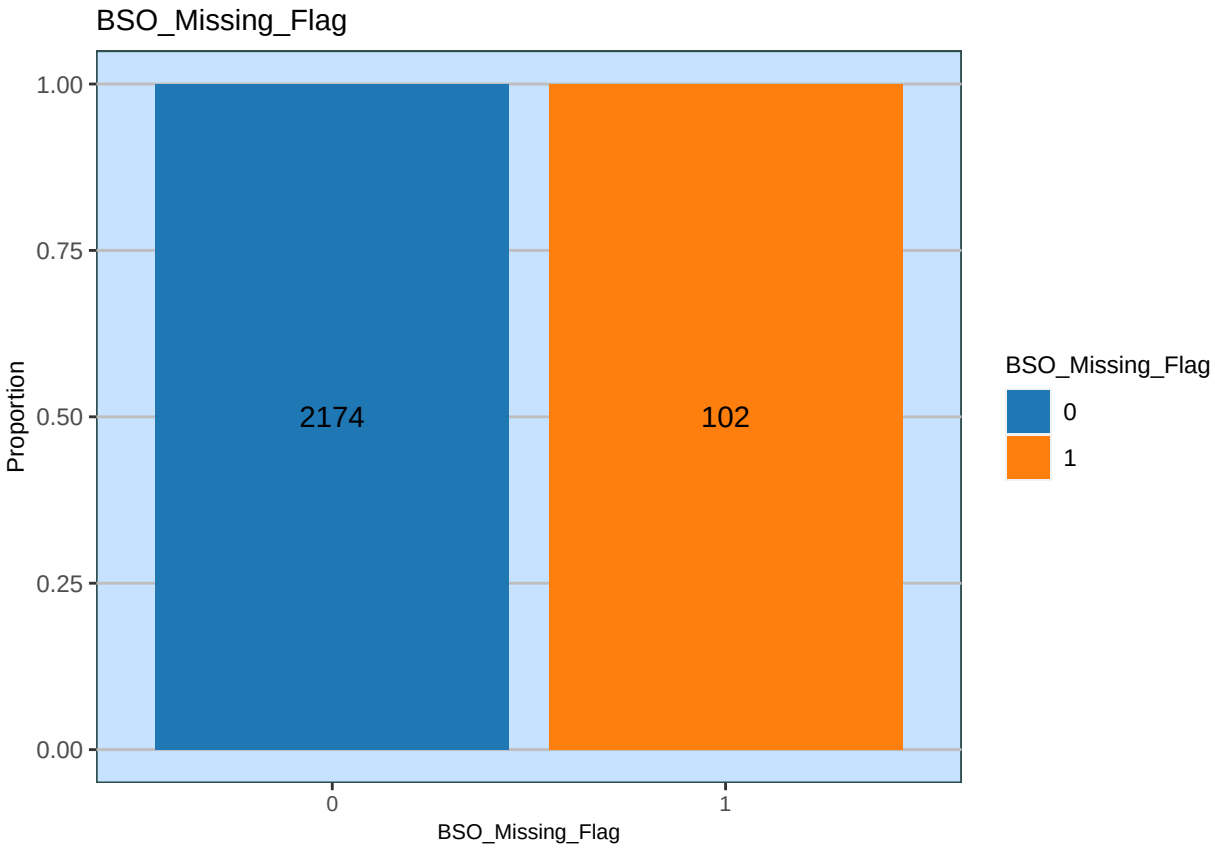
Before we address imputation, we want to work with the implausible zeros in the dataet. In particular, we note that the 0s in Pitching_SO and Batting_SO are a complete overlap, and that the jump between 0 and the next lowest values is not smooth, and so we will treat them as NA's. We do the same with 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.

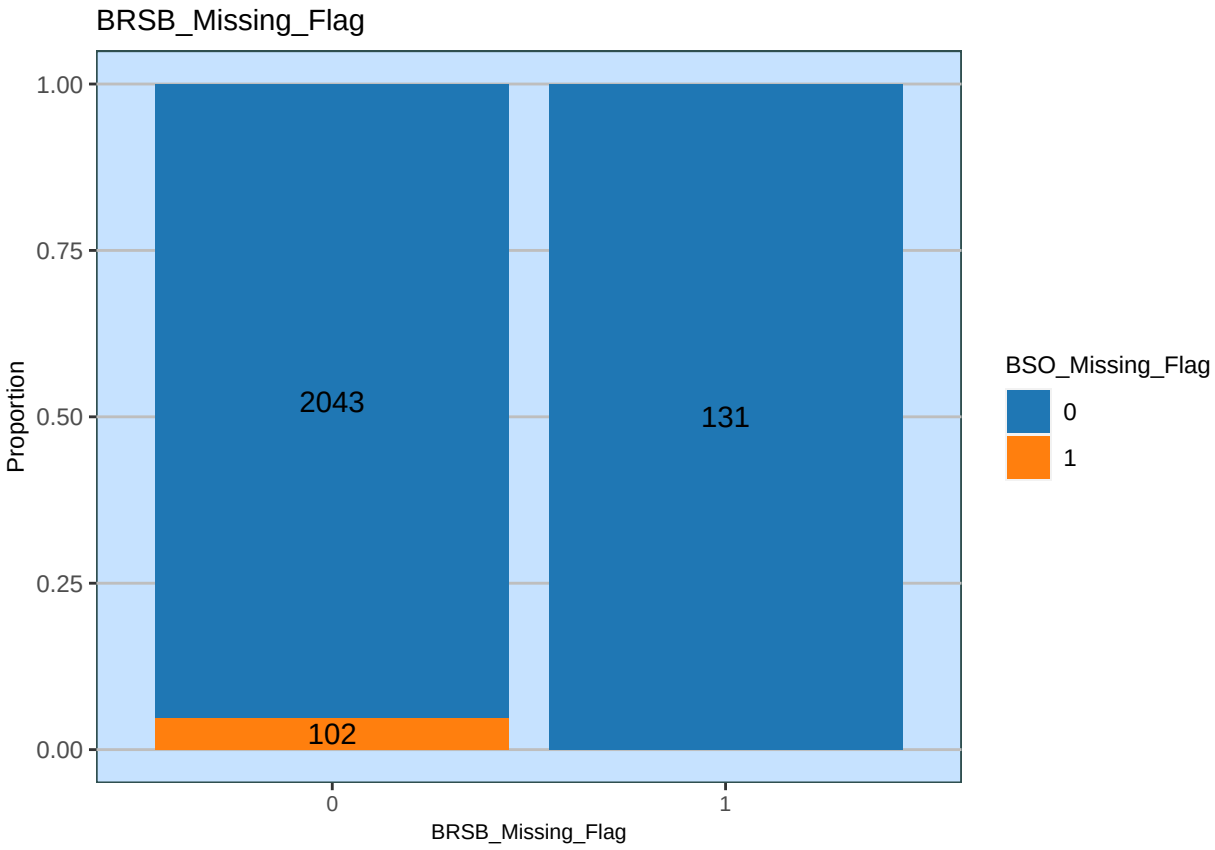
```
## [1] "type:" "mean"
## [1] "r2mean:" "0.4031"
## [1] "r2median:" "0.403"
## [1] "r2omit" "0.4019"

## [[1]]
```


PSO_Missing_Flag
1.00
0.75
0.50
0.25
0.00
Proportion
2174
102
0
1
PSO_Missing_Flag
BSO_Missing_Flag
0
1

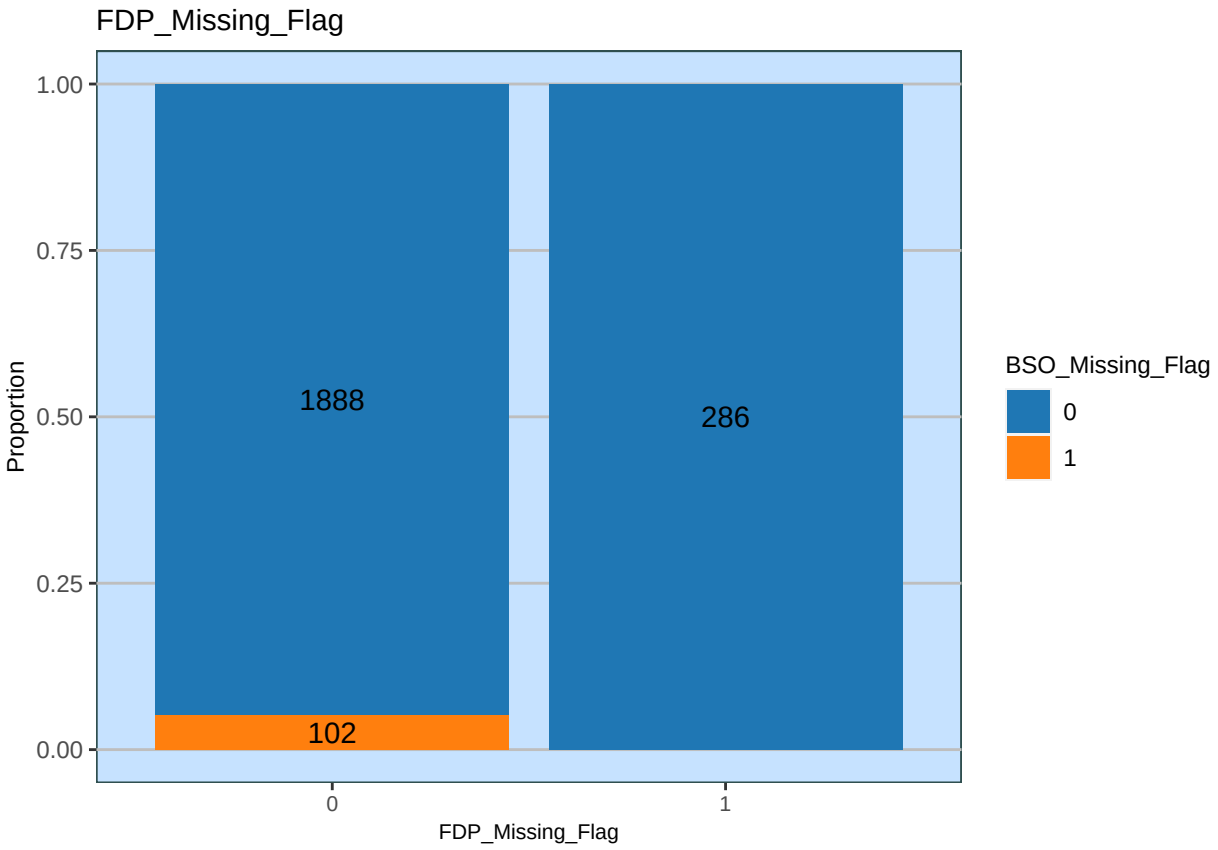
```
##
## [[2]]
```


BSO_Missing_Flag
1.00
0.75
0.50
0.25
0.00
Proportion
2174
102
0
1
BSO_Missing_Flag
BSO_Missing_Flag
0
1

```
##
## [[3]]
```


BRSB_Missing_Flag
1.00
0.75
0.50
0.25
0.00
Proportion
2043
131
102
0
1
BRSB_Missing_Flag
BSO_Missing_Flag
0
1

```
##
## [[4]]
```



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 I am reluctant to eliminate outliers or influential points without good reason. We don't know if extreme numbers are necessarily implausible. Therefore the outliers will remain.

3. Data Modeling

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. 6).

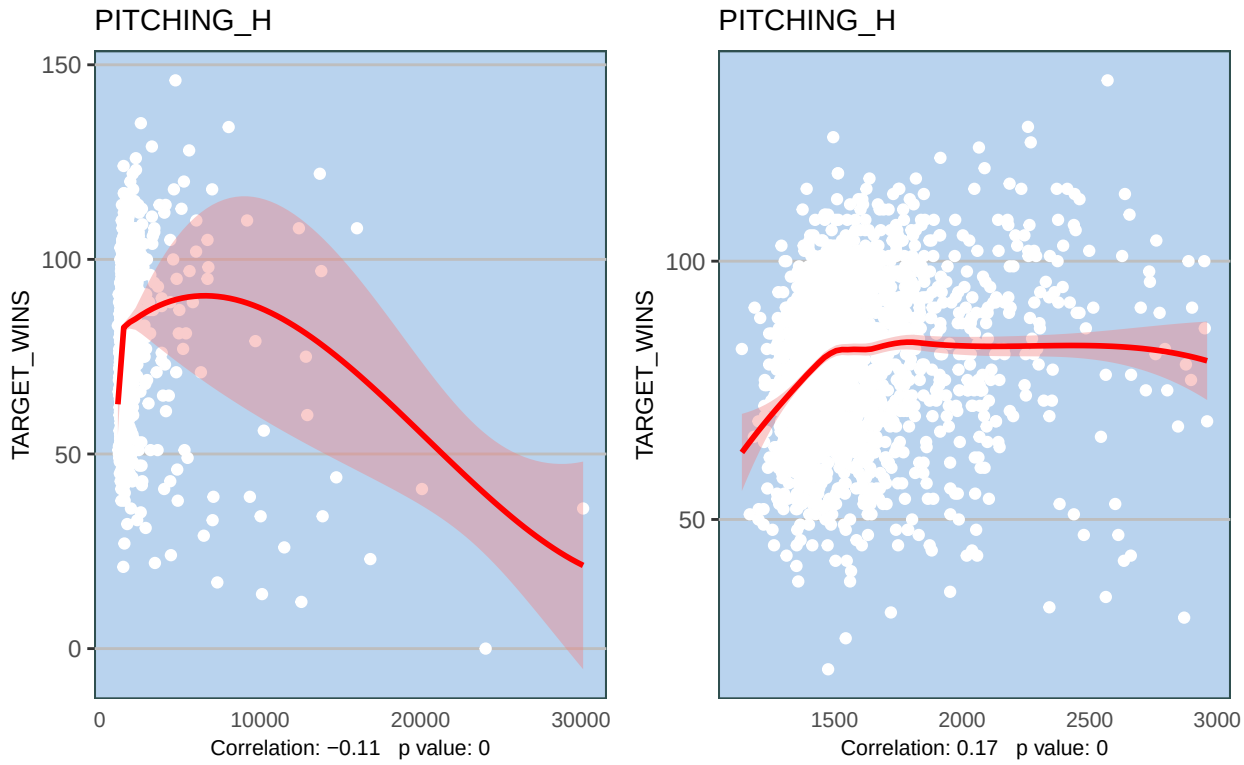
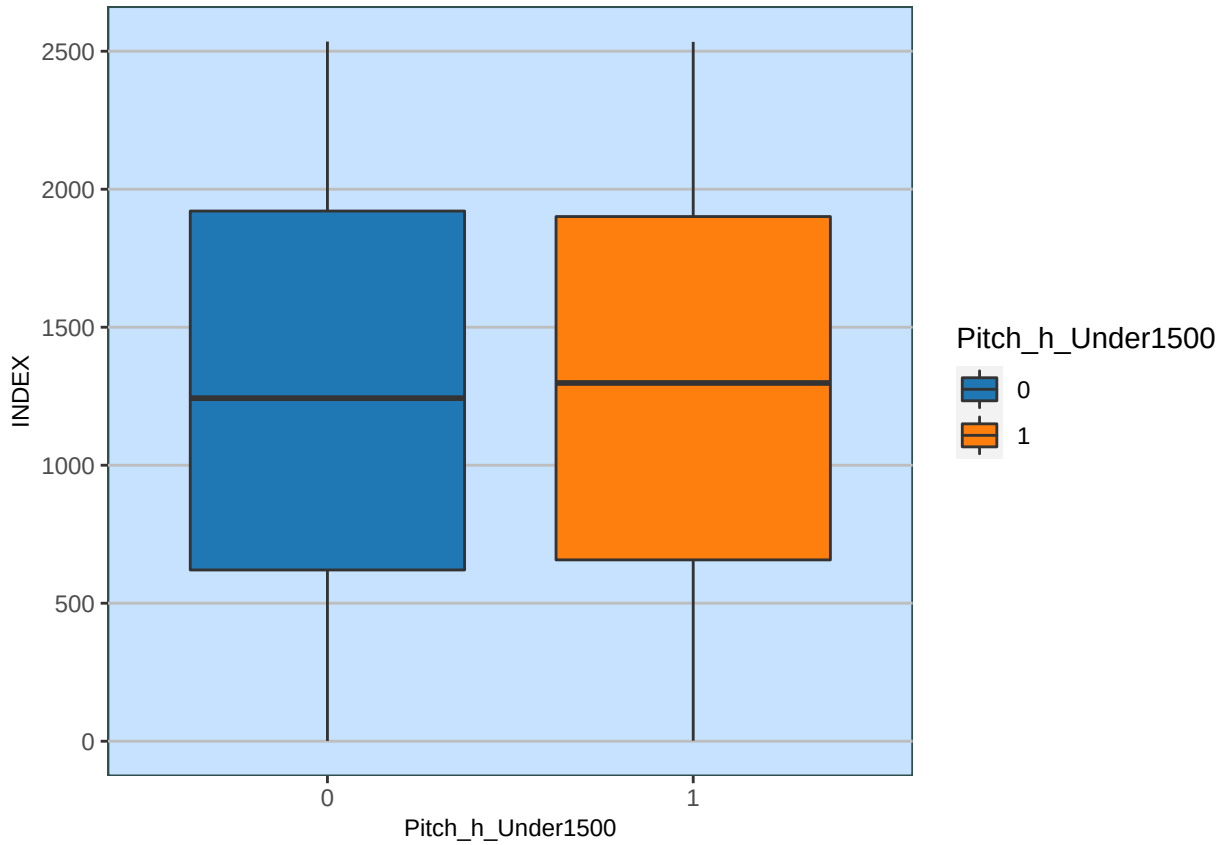
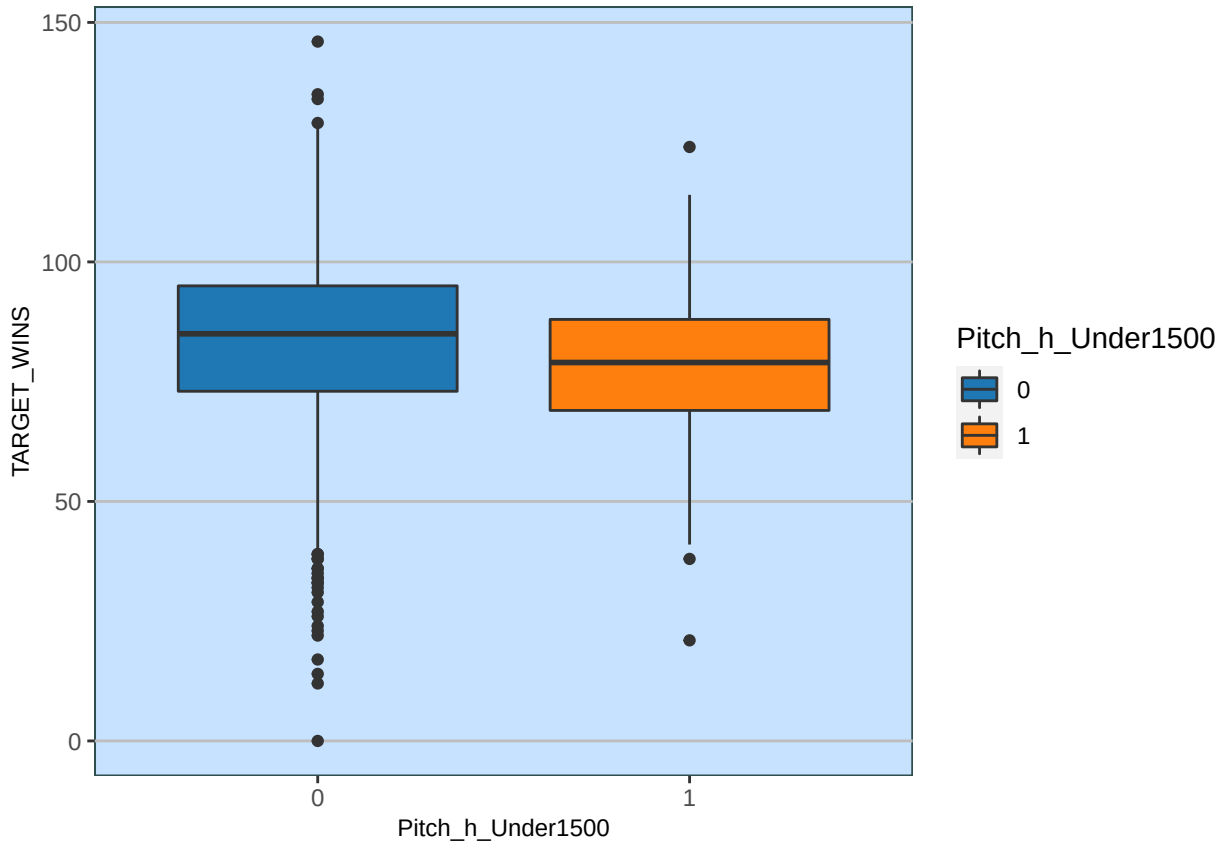

Pitching_H Against Wins, All Records (left) and Hits Below 3000 (right)
PITCHING_H
TARGET_WINS
150
100
50
0
0
10000
20000
30000
Correlation: −0.11 p value: 0
PITCHING_H
TARGET_WINS
100
50
1500
2000
2500
3000
Correlation: 0.17 p value: 0

Fig.6

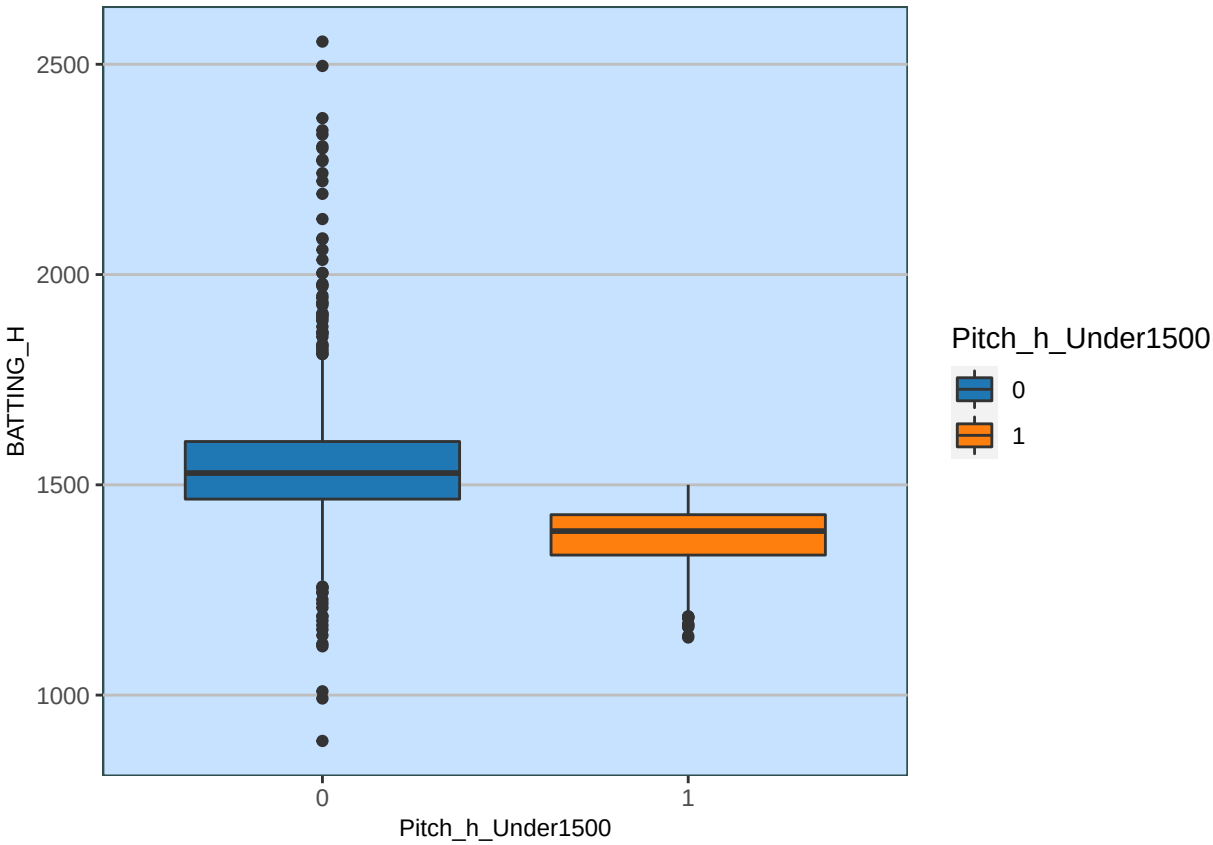
```
## [[1]]
```


2500
2000
1500
1000
500
0
INDEX
0
1
Pitch_h_Under1500
Pitch_h_Under1500
0
1

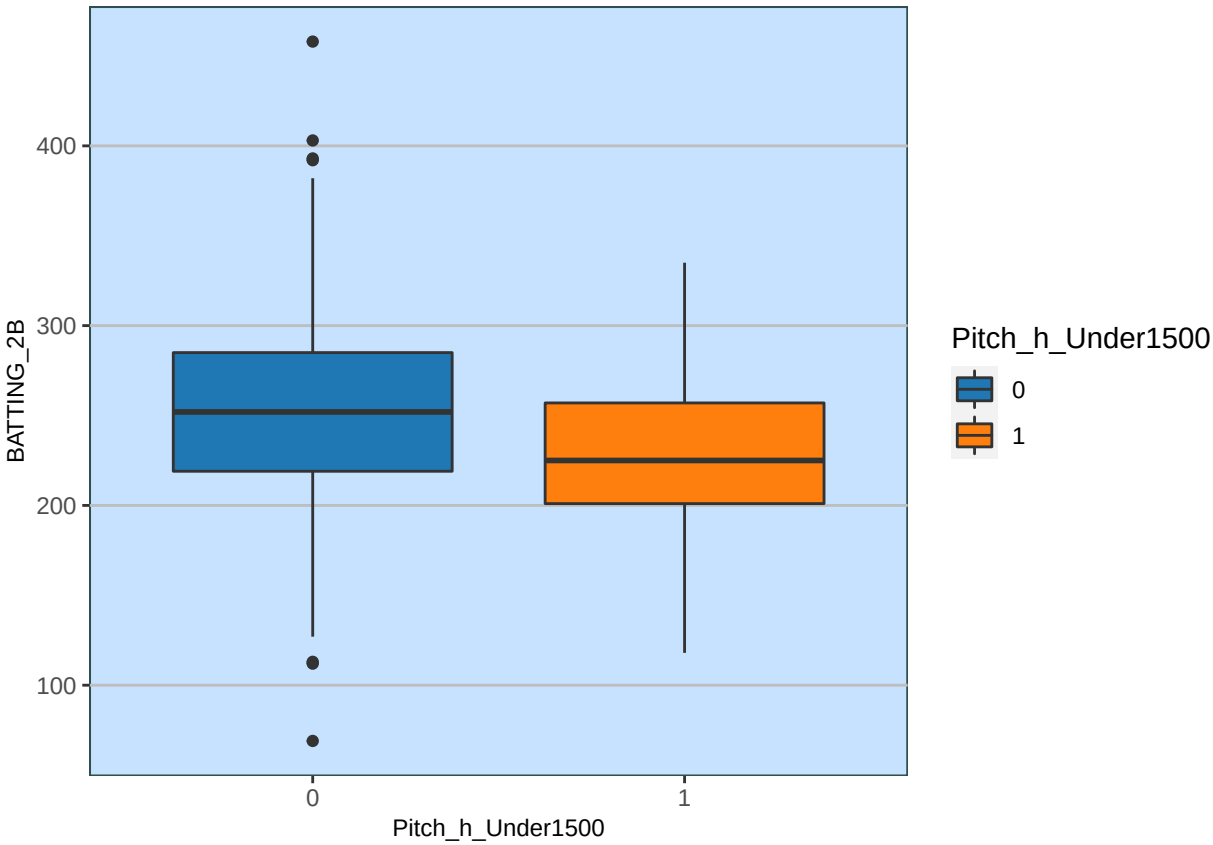
```
##
## [[2]]
```


150
100
50
0
TARGET_WINS
0
1
Pitch_h_Under1500
Pitch_h_Under1500
0
1

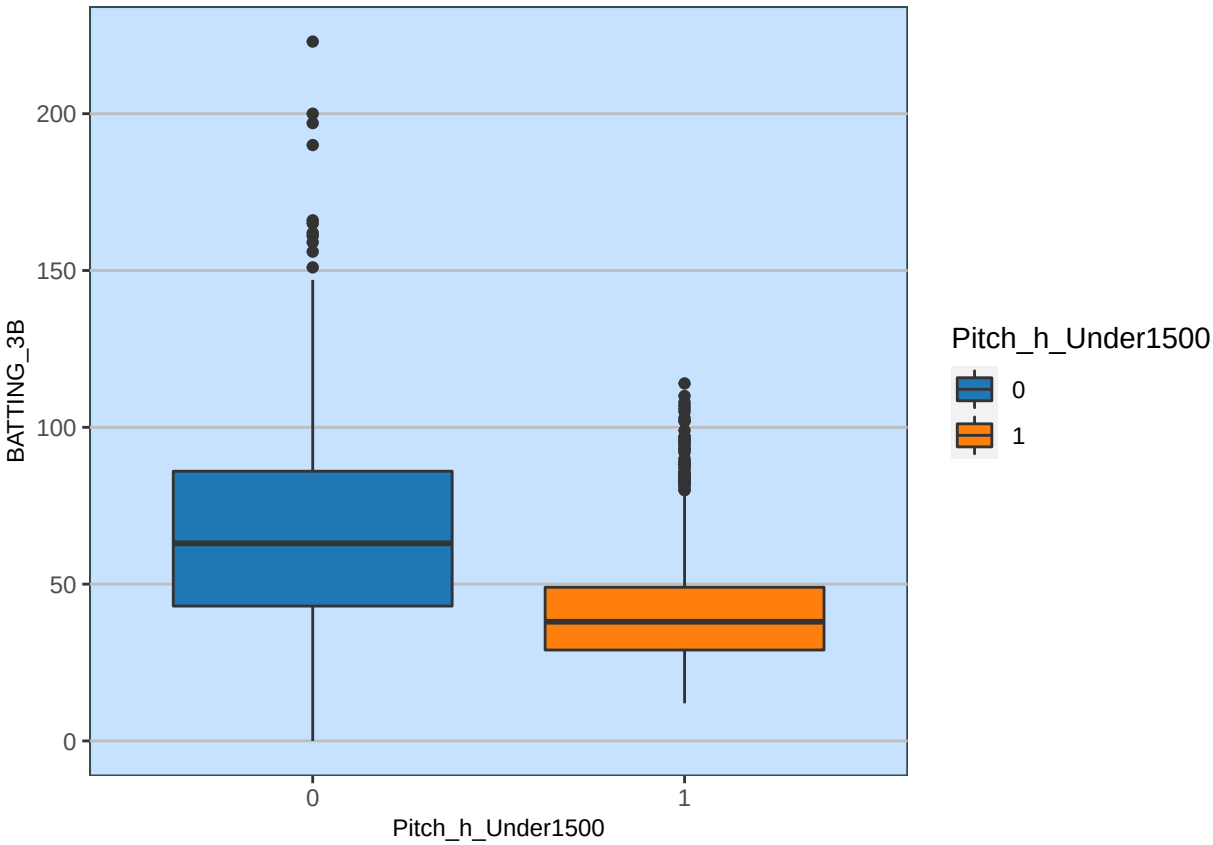
```
##
## [[3]]
```


2500
2000
1500
1000
BATTING_H
0
1
Pitch_h_Under1500
Pitch_h_Under1500
0
1

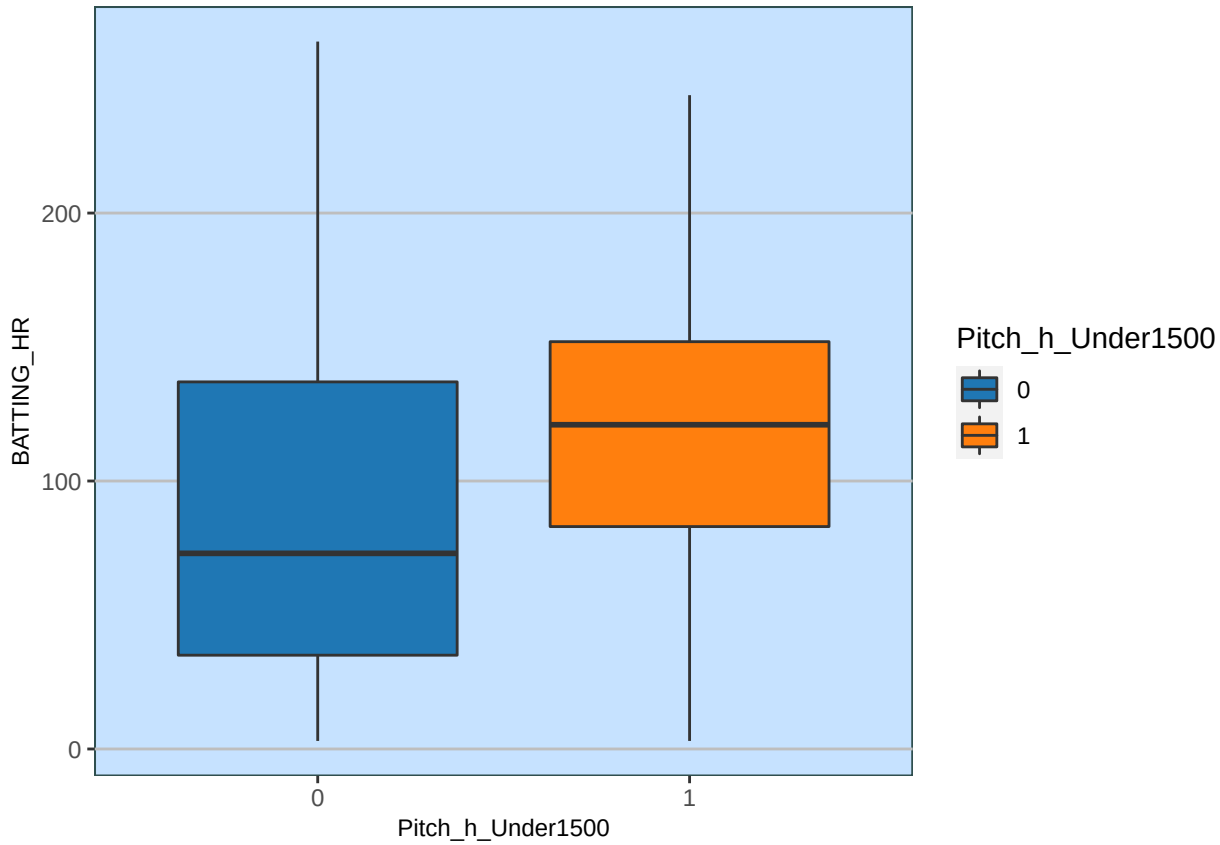
```
##
## [[4]]
```


400
300
200
100
BATTING_2B
0
1
Pitch_h_Under1500
Pitch_h_Under1500
0
1

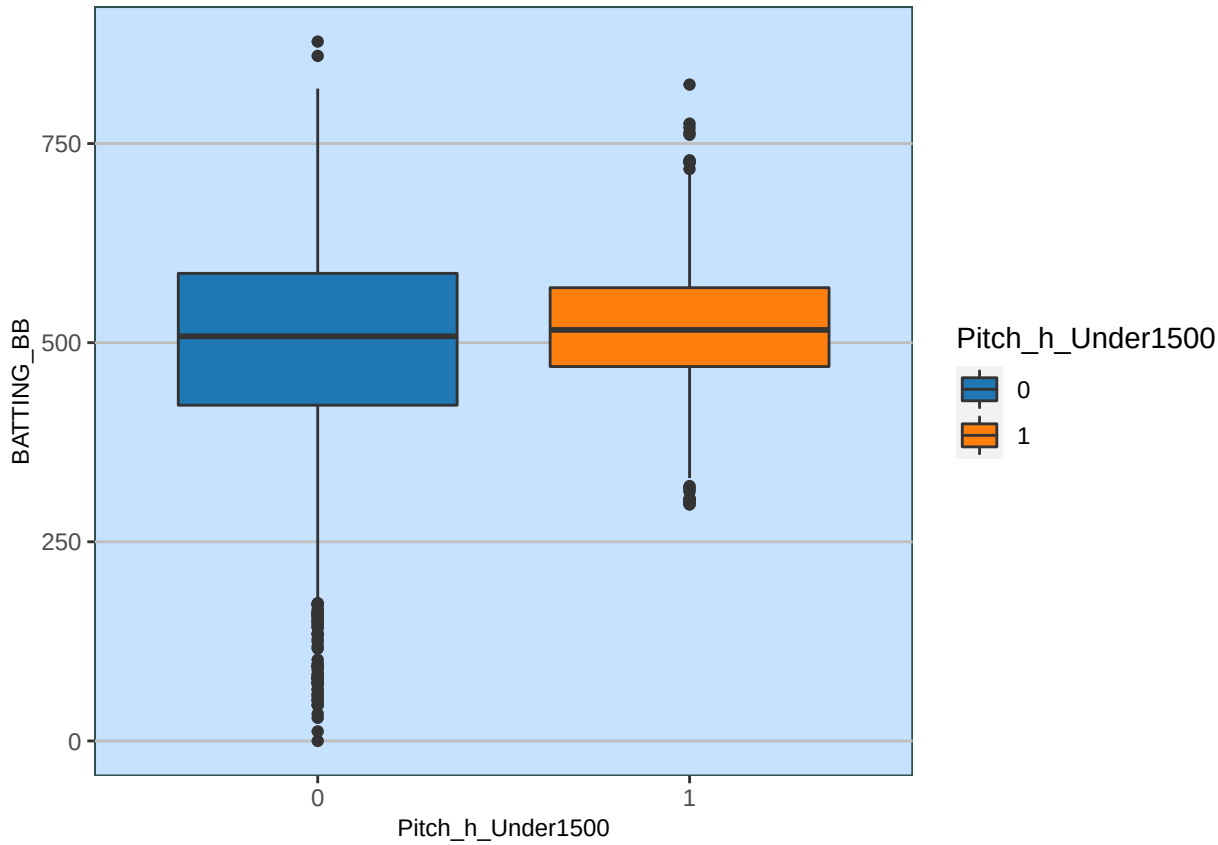
```
##
## [[5]]
```


200
150
100
50
0
BATTING_3B
0
1
Pitch_h_Under1500
Pitch_h_Under1500
0
1

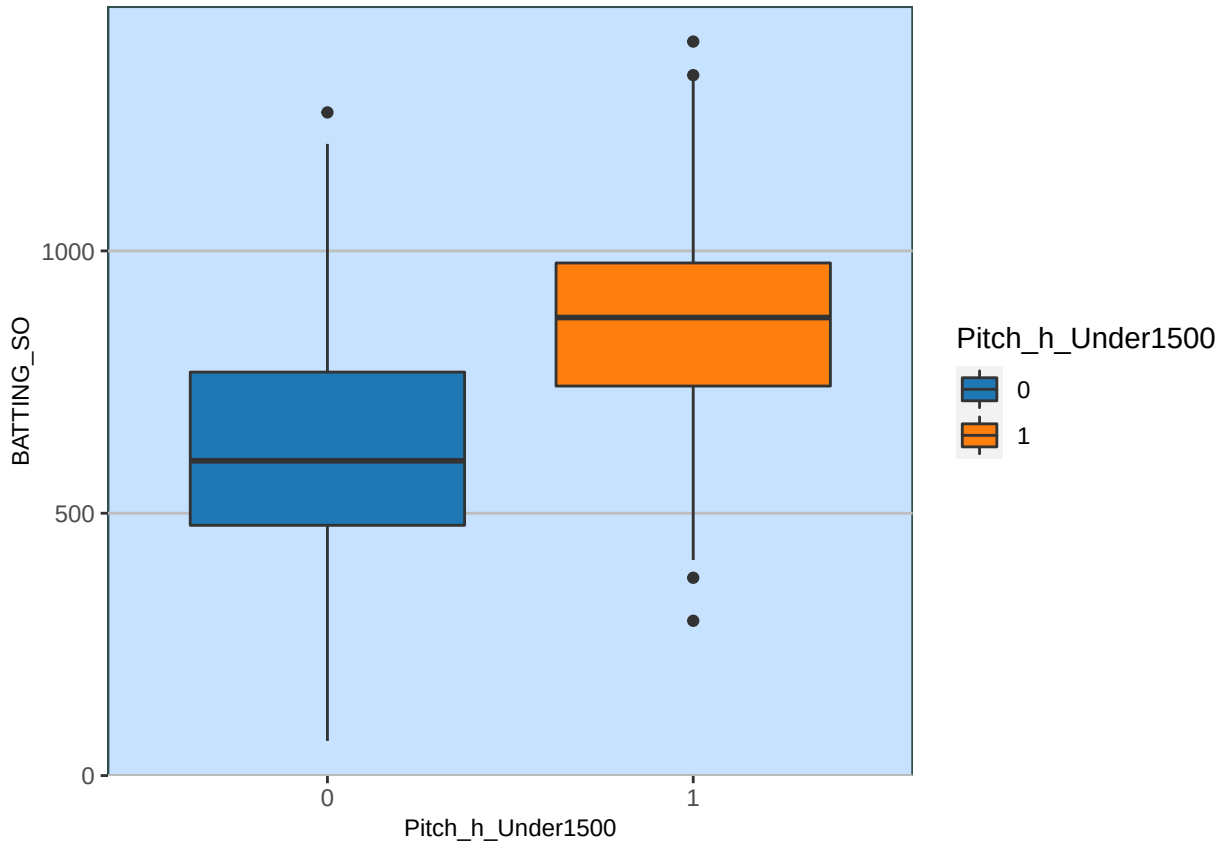
```
##
## [[6]]
```


BATTING_HR
200
100
0
0
1
Pitch_h_Under1500
Pitch_h_Under1500
0
1

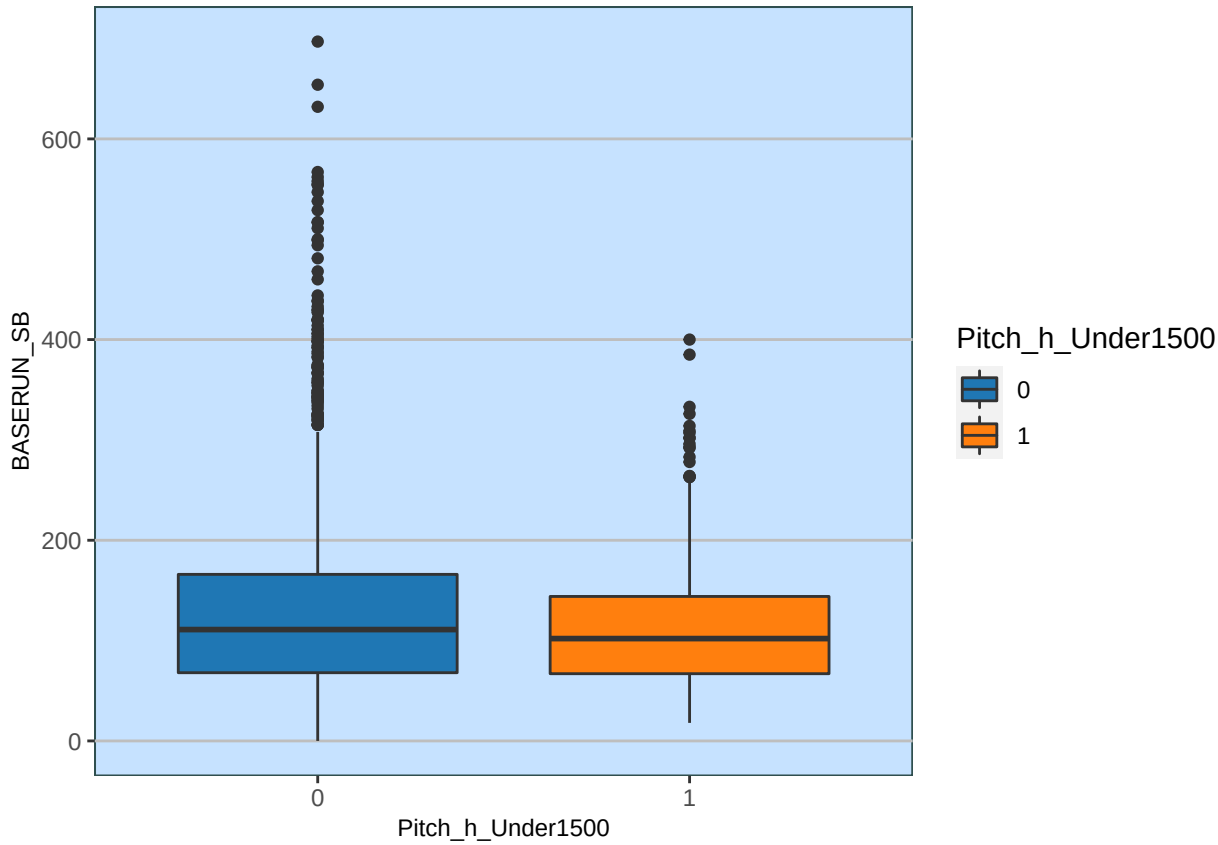
```
##
## [[7]]
```


750
500
250
0
BATTING_BB
0
1
Pitch_h_Under1500
Pitch_h_Under1500
0
1

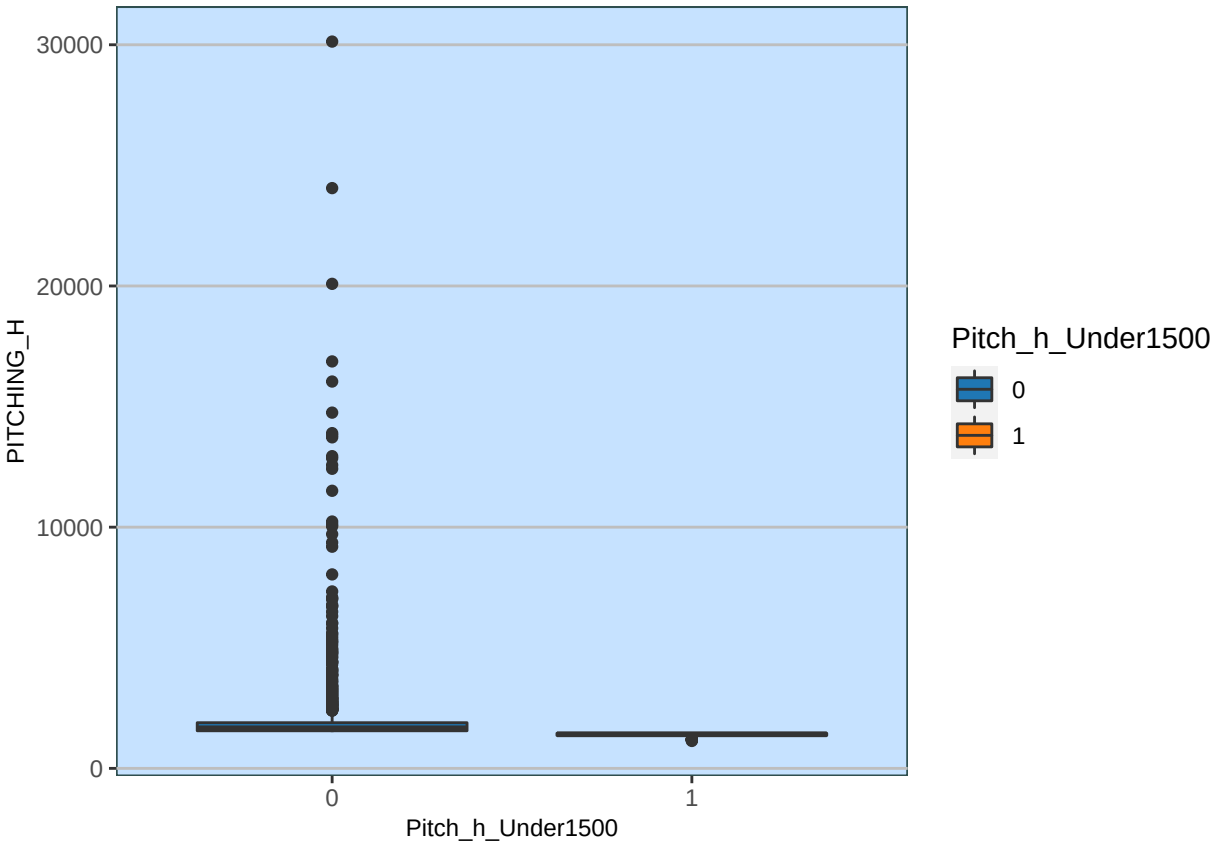
```
##
## [[8]]
```


BATTING_SO
1000
500
0
0
1
Pitch_h_Under1500
Pitch_h_Under1500
0
1

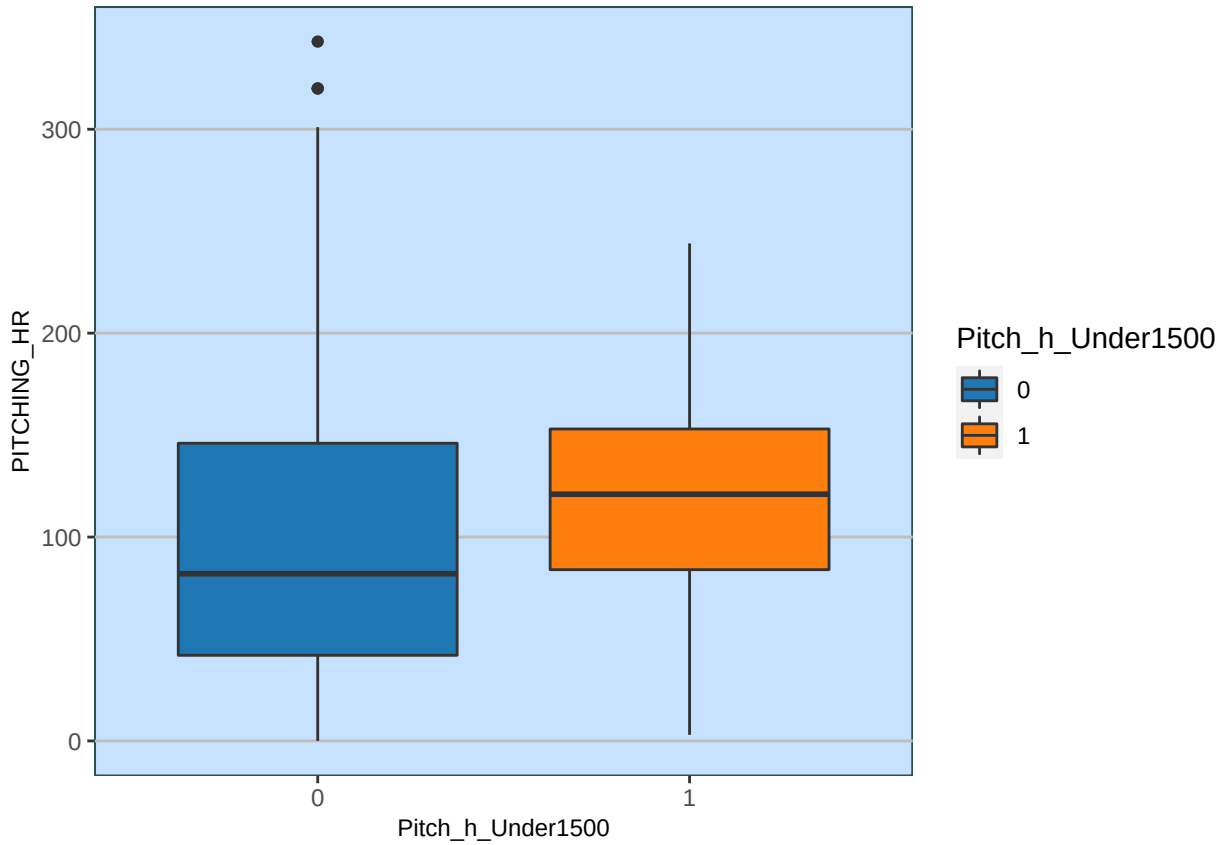
```
##
## [[9]]
```


600
400
200
0
BASERUN_SB
0
1
Pitch_h_Under1500
Pitch_h_Under1500
0
1

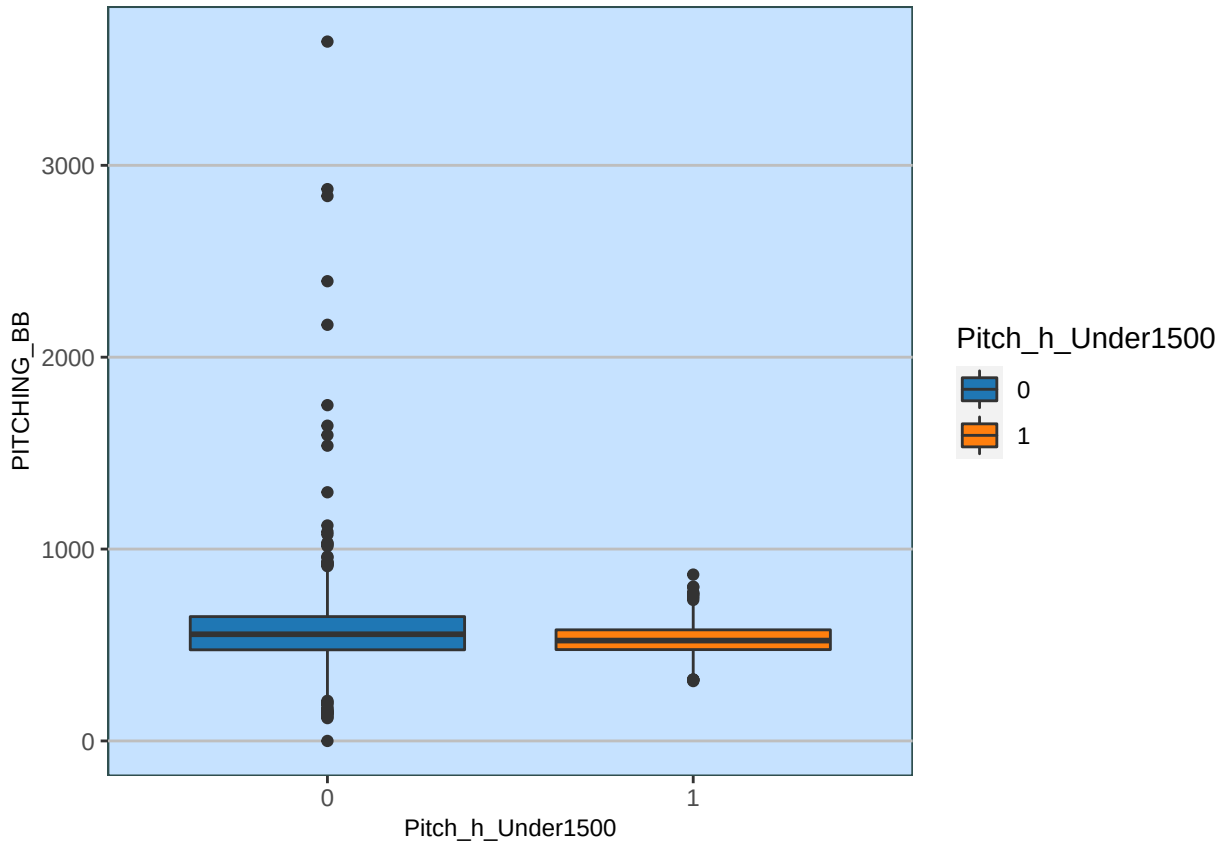
```
##
## [[10]]
```


30000
20000
10000
0
PITCHING_H
0
1
Pitch_h_Under1500
Pitch_h_Under1500
0
1

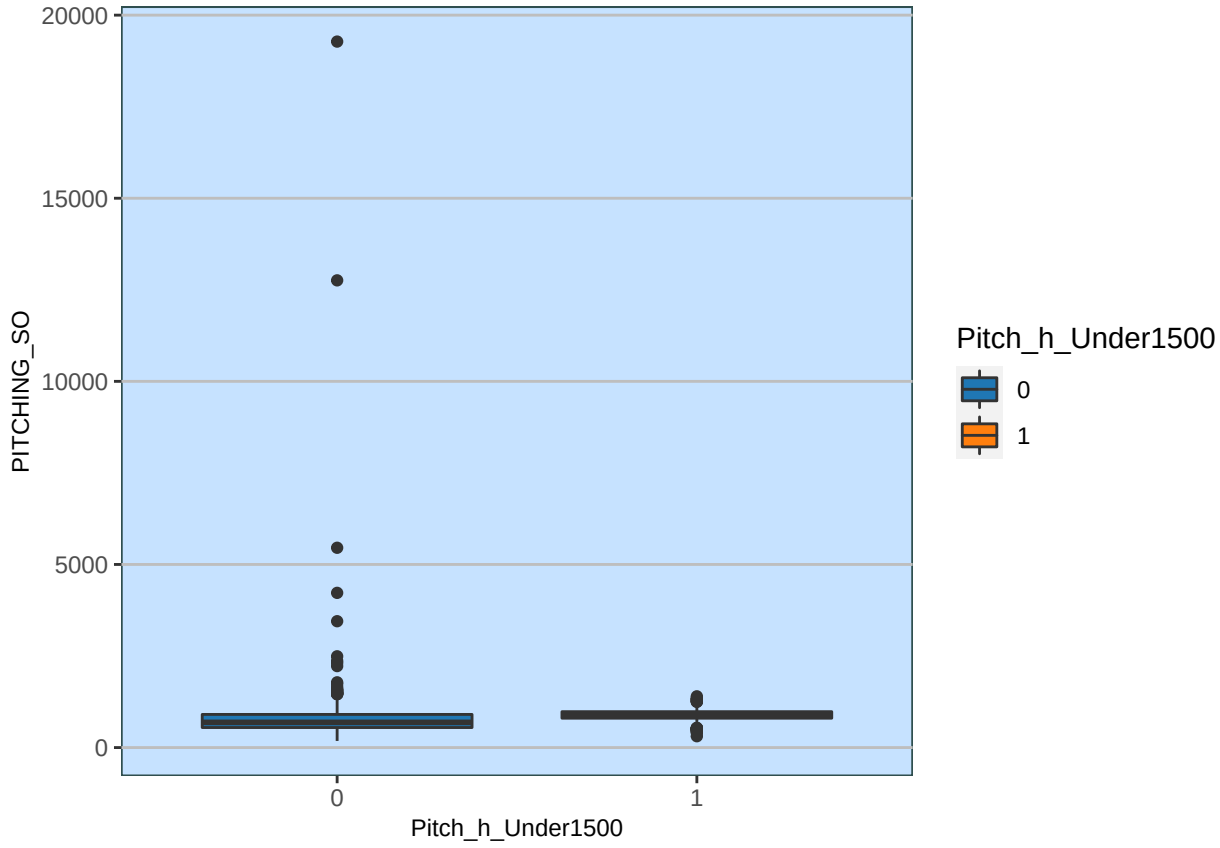
```
##
## [[11]]
```


300
200
100
0
PITCHING_HR
0
1
Pitch_h_Under1500
Pitch_h_Under1500
0
1

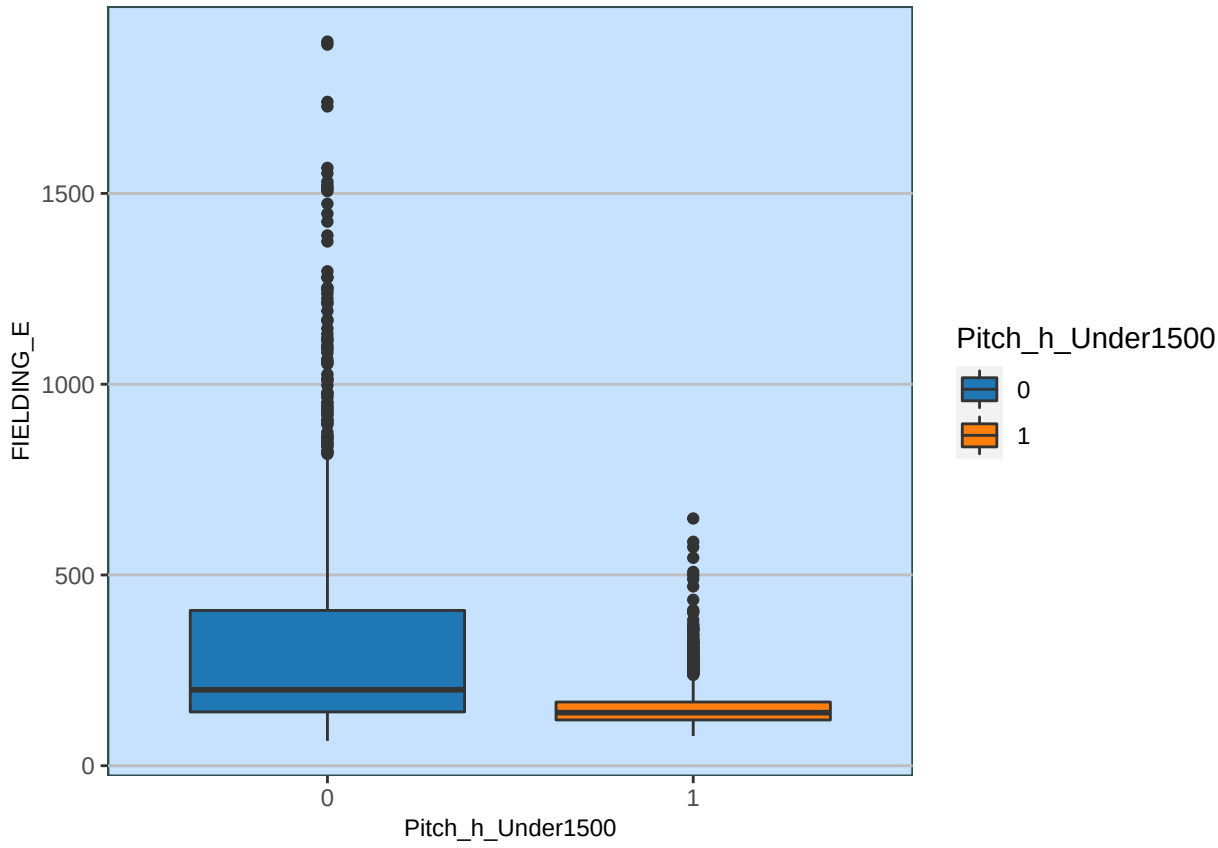
```
##
## [[12]]
```


PITCHING_BB
3000
2000
1000
0
0
1
Pitch_h_Under1500
Pitch_h_Under1500
0
1

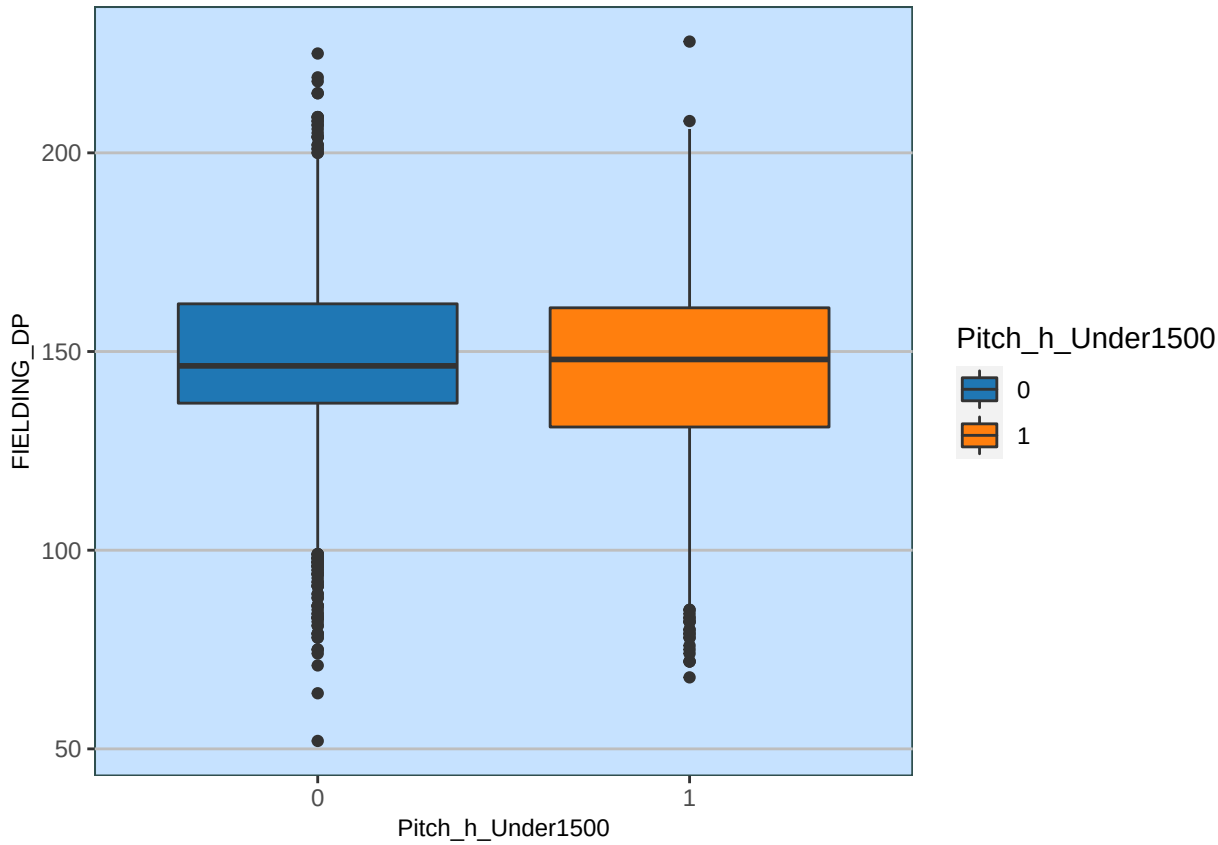
```
##
## [[13]]
```


20000
15000
10000
5000
0
PITCHING_SO
0
1
Pitch_h_Under1500
Pitch_h_Under1500
0
1

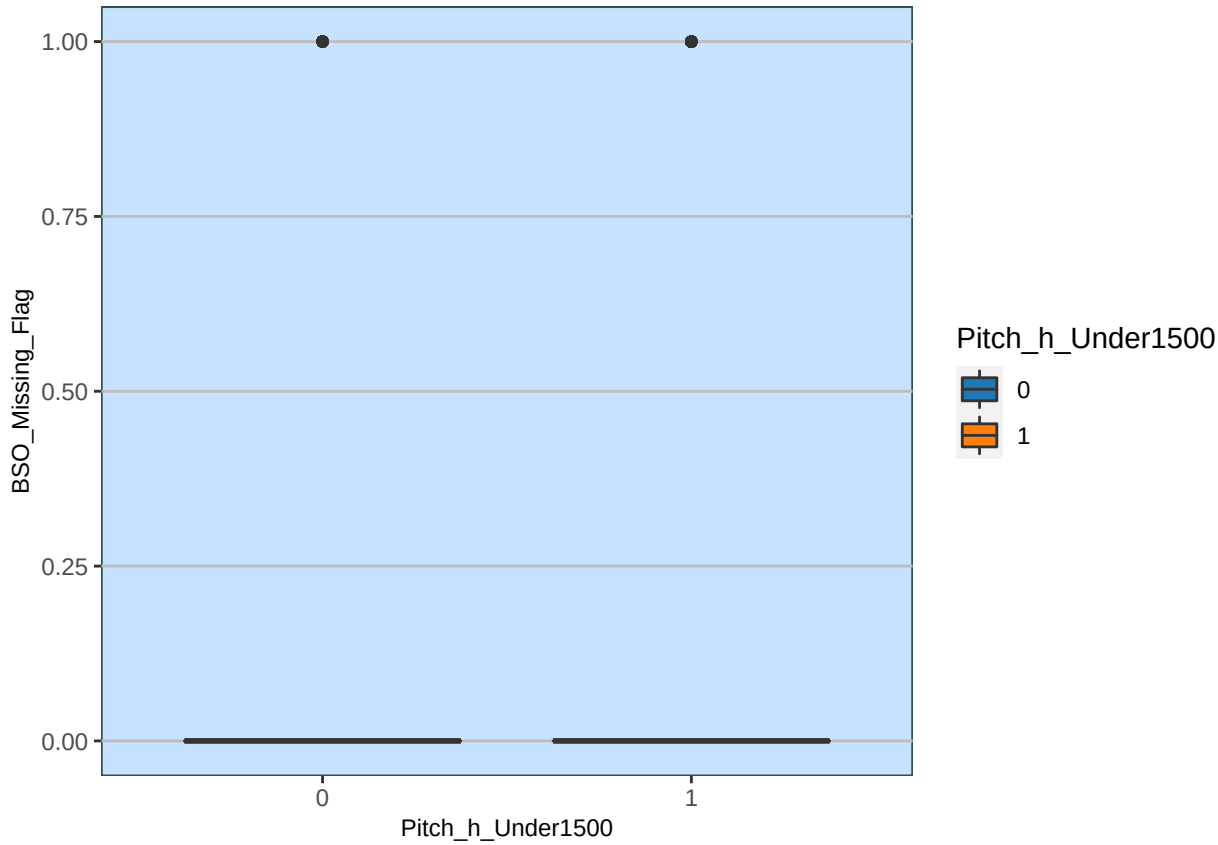
```
##
## [[14]]
```


FIELDING_E
1500
1000
500
0
0
1
Pitch_h_Under1500
Pitch_h_Under1500
0
1

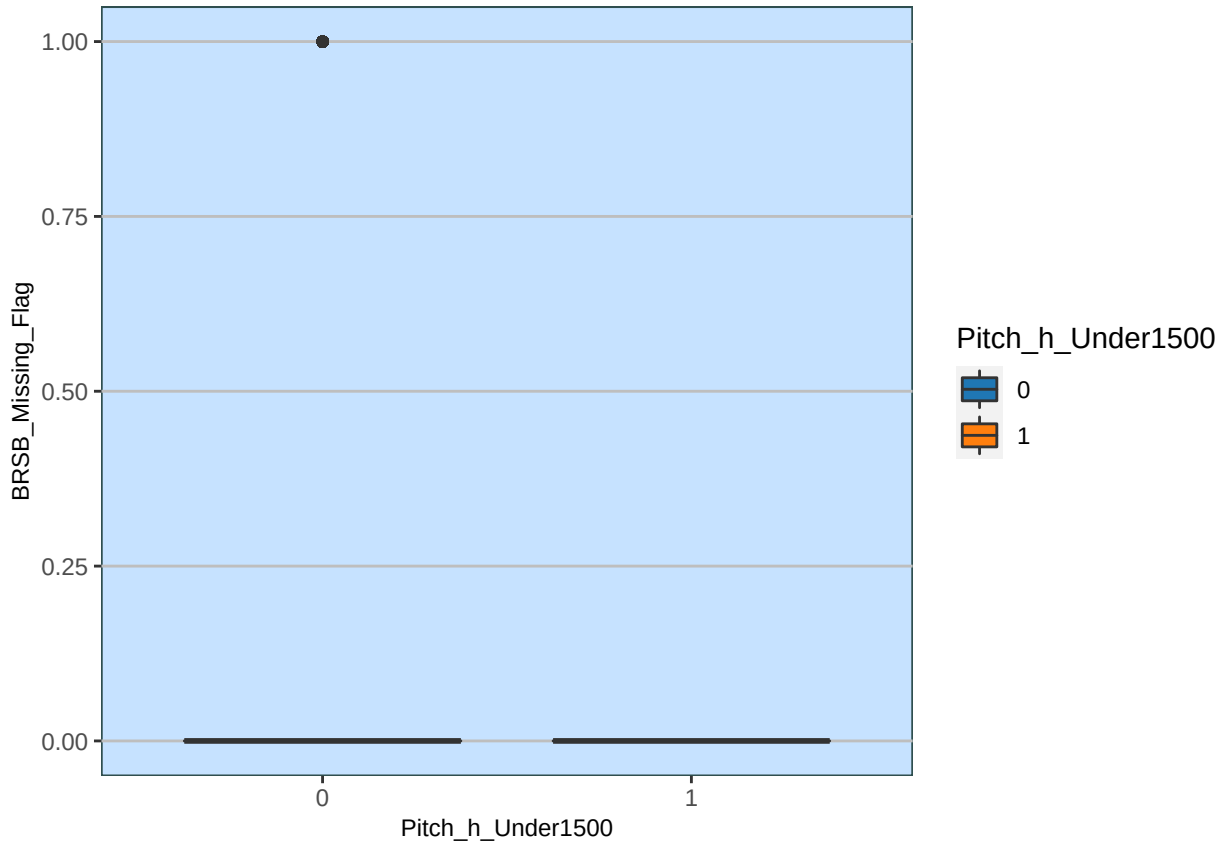
```
##
## [[15]]
```


200
150
100
50
FIELDING_DP
0
1
Pitch_h_Under1500
Pitch_h_Under1500
0
1

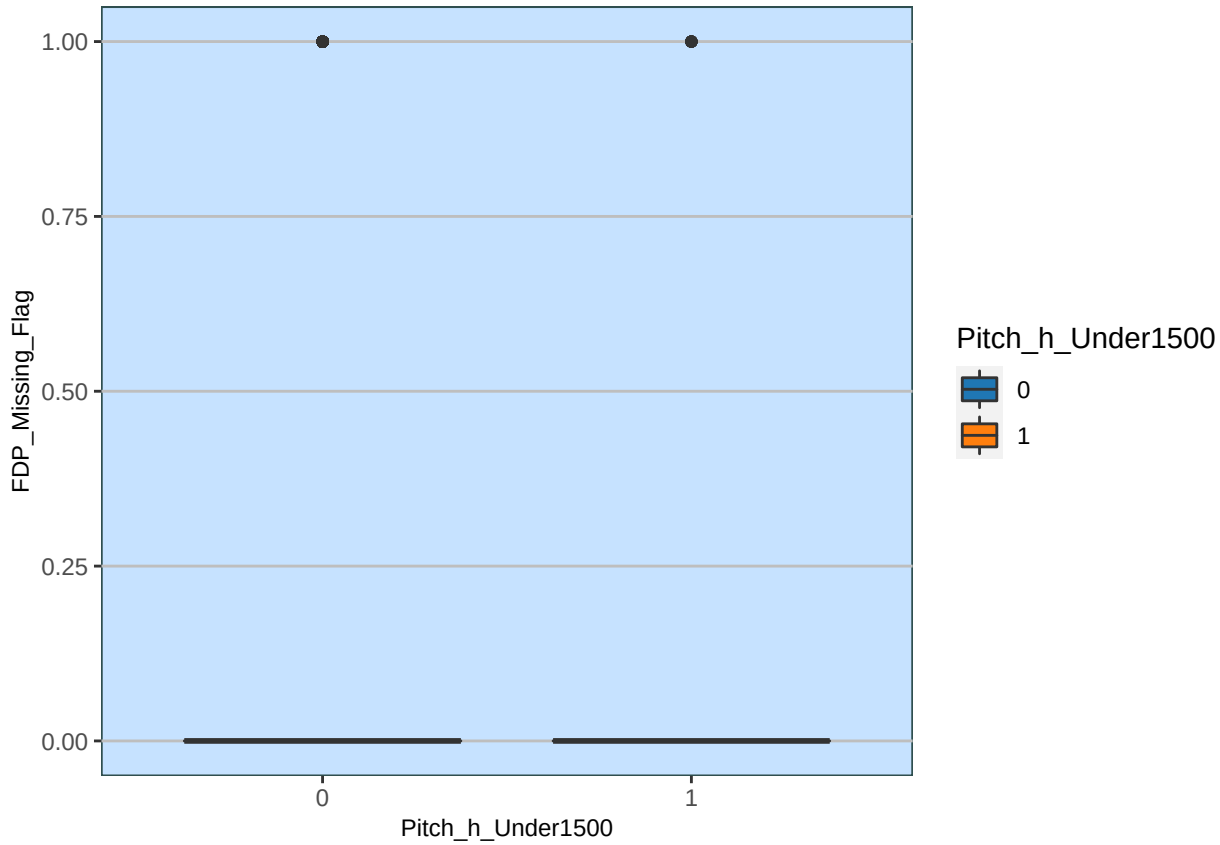
```
##
## [[16]]
```


1.00
0.75
0.50
0.25
0.00
BSO_Missing_Flag
0
1
Pitch_h_Under1500
Pitch_h_Under1500
0
1

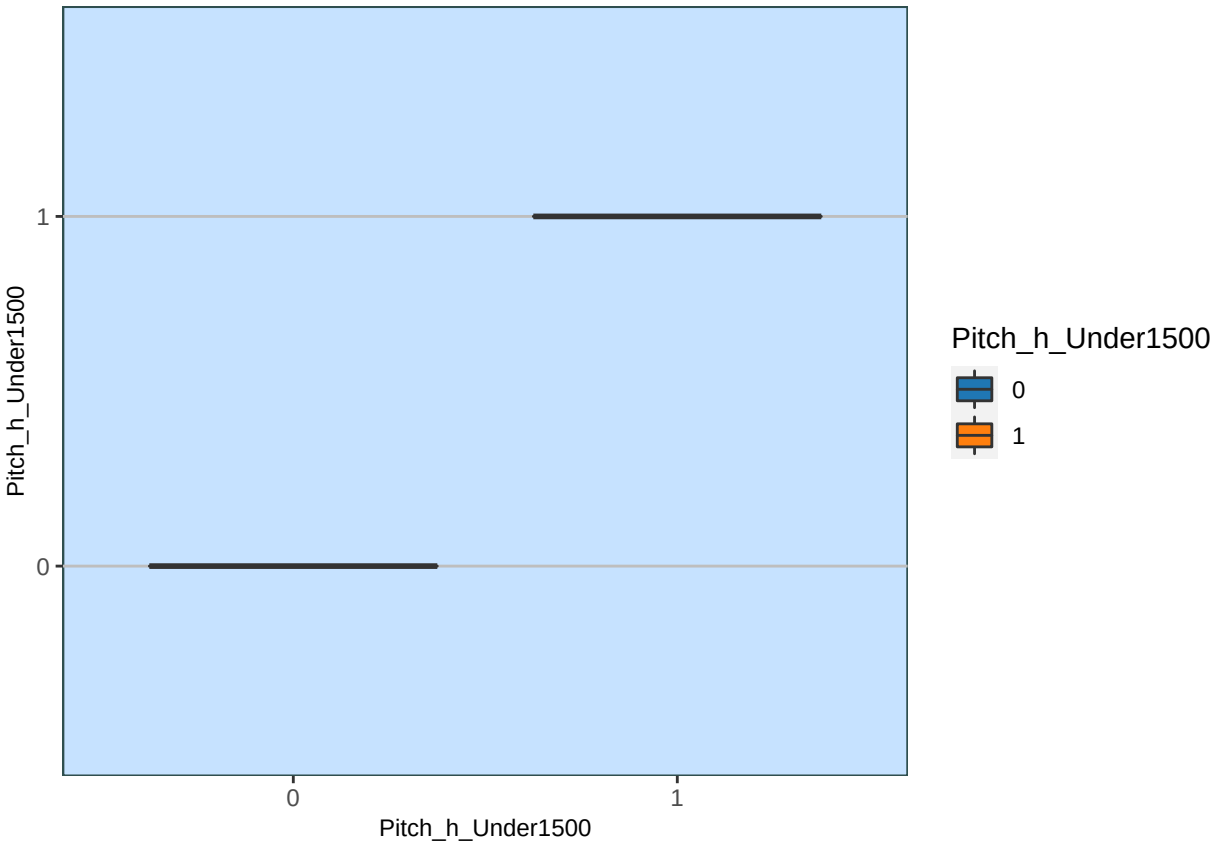
```
##
## [[17]]
```


1.00
0.75
0.50
0.25
0.00
BRSB_Missing_Flag
0
1
Pitch_h_Under1500
Pitch_h_Under1500
0
1

```
##
## [[18]]
```


1.00
0.75
0.50
0.25
0.00
FDP_Missing_Flag
0
1
Pitch_h_Under1500
Pitch_h_Under1500
0
1

```
##
## [[19]]
```



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.

3. We drop PITCHING_HR because it is an implausibly close match with HITTING_HR.

Like many pitching columns, Pitching_HR is unexpectedly positively correlated with wins. However, what makes this column truly implausible is how close a match it is with BATTING_HR. The scatterplot below (Fig. 7) shows that the vast majority of the figures for pitching HR are exactly the same or within 2 or 3 of Batting HR. We therefore drop it since this makes no sense.

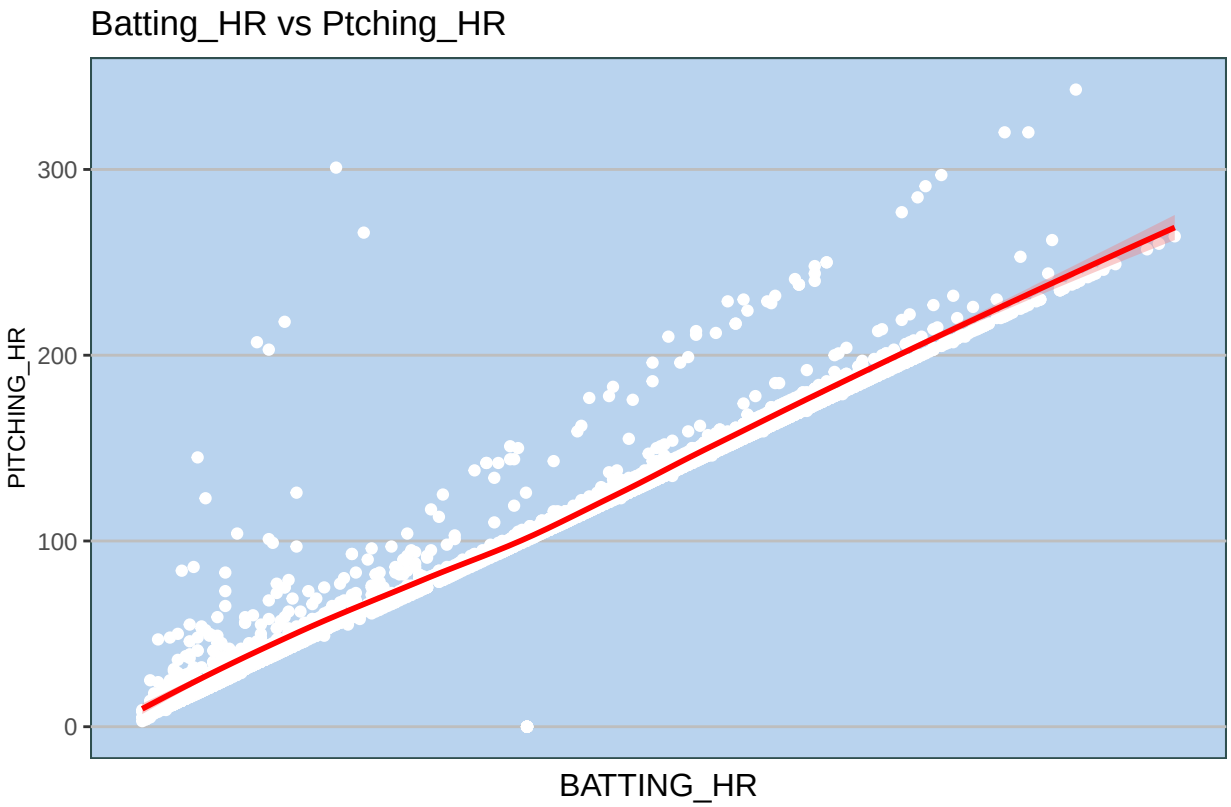


Fig. 7

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

Batting HR has a bimodal distribution (see Fig. 8). 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.

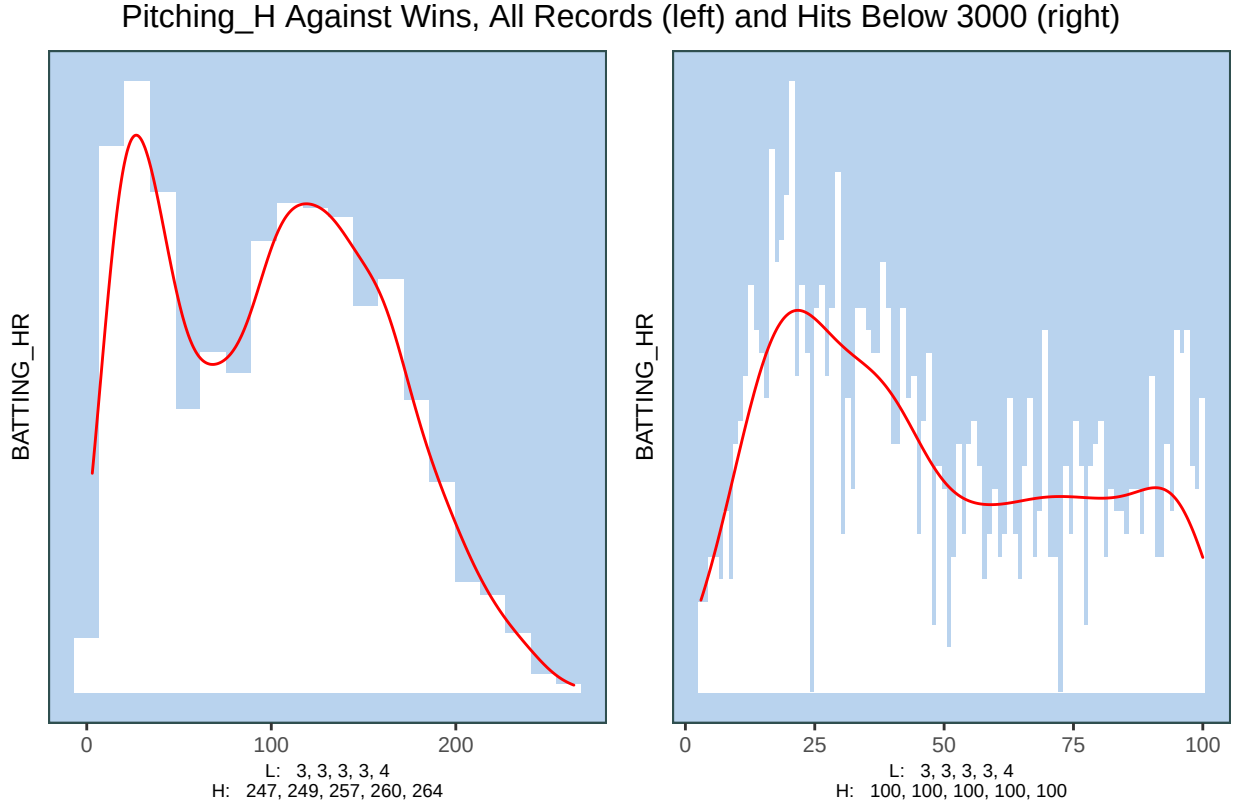


Fig.8

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

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

5. We transform the error variable.

While the distributions of a number of columns suggest possible tranformations, we focus here on errors, which has an upside-down u shape when correlated with wins. We therefore add an error squared term to the dataset.

```
## [1] 0.03072081
```

```
## [1] 0.04825783
```

6. We create interaction terms between the SO missing cohort and the columns identified above - Pitching H, Pitching BB, Batting HR and Fielding E

7. For the sake of legibility, we do not create log terms for the many skewed distributions.

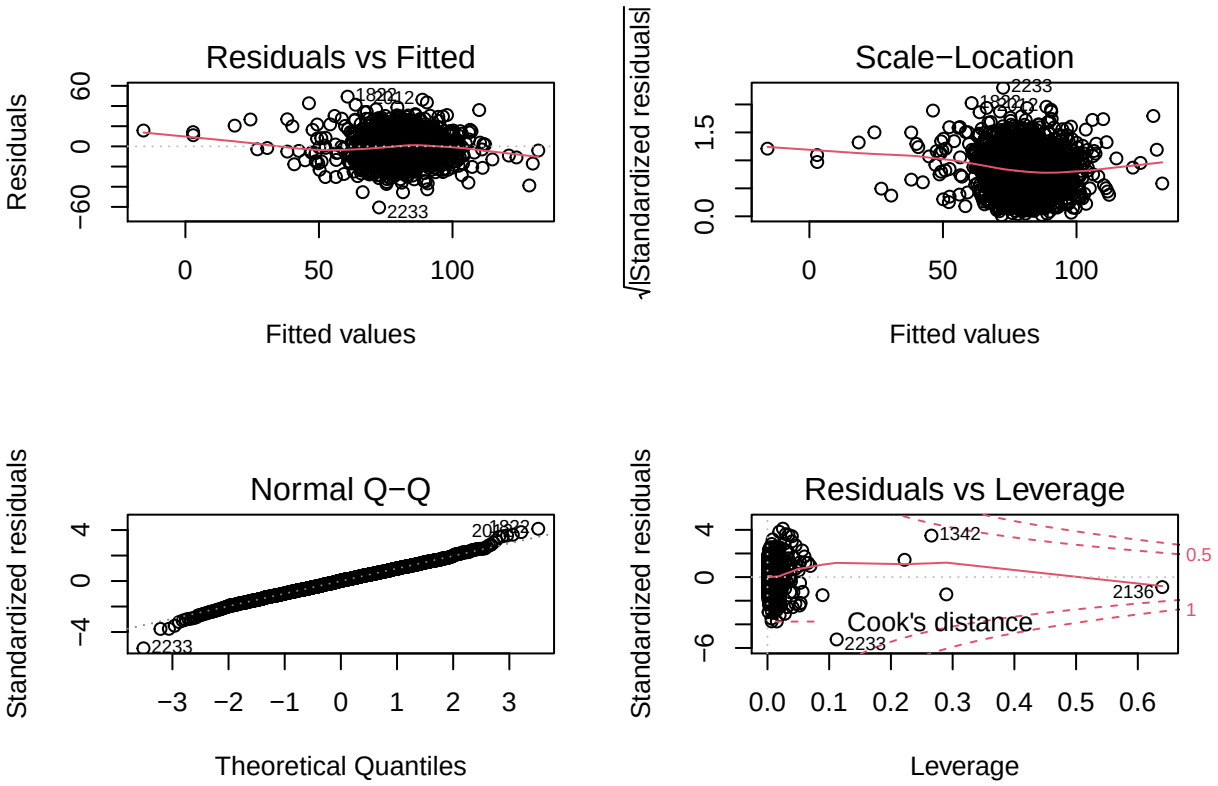
We would normally sacrifice some legibility for improved predicitibilty by trying some log transformations on skewed independent variable distributions. However, legibility is already in serious peril with the odd behavior of the many pitching terms which suggest bad defense wins games. We therefore leave our transformations at those described.

4. Model Selection

Here we build and test our models to gain insight into the dataset and ultimately predict outcomes.

a. Regression 1: Baseline (No transformations except missing flags)

```
##
## 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
```



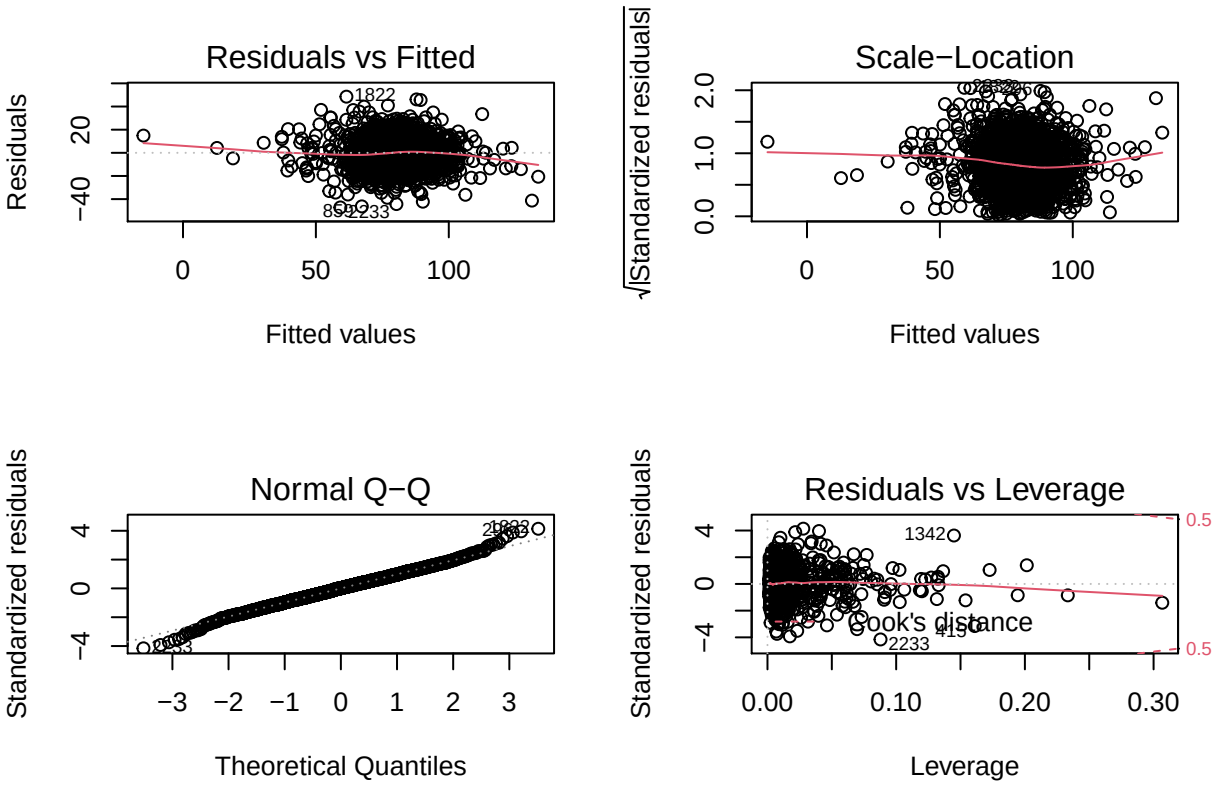

```
## NULL
```

The adjusted r squared is .403. As we expected, many of the signs are in the "wrong" direction.

b. Regression 2: All transformations

```
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B +
##     BATTING_HR + BATTING_BB + BATTING_SO + BASERUN_SB + PITCHING_H +
##     FIELDING_E + FIELDING_DP + BSO_Missing_Flag + BRSB_Missing_Flag +
##     FDP_Missing_Flag + Pitch_h_Under1500 + Prod_DP_H + E_sq +
##     Inter_bb_Cohort + Inter_E_Cohort + Inter_bhr_Cohort + Inter_bbb_Cohort +
##     Inter_bs_Cohort, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.202  -7.806   0.193   7.821  48.504
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.472e+01  6.702e+00   3.688 0.000231 ***
## BATTING_H           5.622e-02  3.302e-03  17.023  < 2e-16 ***
## BATTING_2B         -4.125e-02  8.586e-03  -4.805 1.65e-06 ***
## BATTING_3B          6.743e-02  1.610e-02   4.188 2.93e-05 ***
```

```
## BATTING_HR         5.825e-02  8.978e-03   6.488 1.06e-10 ***
## BATTING_BB         2.593e-02  3.247e-03   7.984 2.23e-15 ***
## BATTING_SO        -1.223e-02  2.218e-03  -5.512 3.95e-08 ***
## BASERUN_SB         5.238e-02  4.795e-03  10.923  < 2e-16 ***
## PITCHING_H        -4.629e-03  2.995e-03  -1.546 0.122287
## FIELDING_E        -8.282e-02  7.453e-03 -11.112  < 2e-16 ***
## FIELDING_DP       -1.646e-01  3.571e-02  -4.610 4.25e-06 ***
## BSO_Missing_Flag   5.042e+01  1.190e+01   4.237 2.36e-05 ***
## BRSB_Missing_Flag  3.794e+01  2.023e+00  18.752  < 2e-16 ***
## FDP_Missing_Flag   5.282e+00  1.713e+00   3.084 0.002064 **
## Pitch_h_Under1500  2.214e+00  6.829e-01   3.242 0.001206 **
## Prod_DP_H          3.671e-05  2.040e-05   1.799 0.072094 .
## E_sq               2.143e-05  4.284e-06   5.002 6.11e-07 ***
## Inter_bb_Cohort    1.336e-01  8.560e-02   1.560 0.118847
## Inter_E_Cohort    -1.938e-01  2.809e-02  -6.899 6.77e-12 ***
## Inter_bhr_Cohort   3.652e-01  1.546e-01   2.362 0.018285 *
## Inter_bbb_Cohort  -1.397e-01  9.536e-02  -1.465 0.143105
## Inter_bs_Cohort    3.896e-02  2.653e-02   1.469 0.142097
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.89 on 2254 degrees of freedom
## Multiple R-squared:  0.4357, Adjusted R-squared:  0.4305
## F-statistic: 82.88 on 21 and 2254 DF,  p-value: < 2.2e-16
##
## [1] "VIF Analysis"
##         BATTING_H        BATTING_2B        BATTING_3B        BATTING_HR
##          3.670470          2.599487          3.258269          4.671215
##        BATTING_BB        BATTING_SO        BASERUN_SB        PITCHING_H
##          2.554246          4.292413          2.688323        285.743188
##        FIELDING_E       FIELDING_DP  BSO_Missing_Flag BRSB_Missing_Flag
##         46.392975         12.345591         97.619815          3.575494
##  FDP_Missing_Flag Pitch_h_Under1500         Prod_DP_H              E_sq
##          5.189563          1.863892        282.770320         24.339858
##   Inter_bb_Cohort    Inter_E_Cohort  Inter_bhr_Cohort  Inter_bbb_Cohort
##       1095.501873         50.949700          7.645153       1173.699962
##   Inter_bs_Cohort
##         21.438192
```




```
## NULL
```

The second model has an adjusted r squared of .4305. This is not significantly better - however the interpretive value of the model is greatly increased, as the coefficient signs are much more reasonable.

```
## Analysis of Variance Table
##
## Model 1: 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
## Model 2: TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B + BATTING_HR +
##     BATTING_BB + BATTING_SO + BASERUN_SB + PITCHING_H + FIELDING_E +
##     FIELDING_DP + BSO_Missing_Flag + BRSB_Missing_Flag + FDP_Missing_Flag +
##     Pitch_h_Under1500 + Prod_DP_H + E_sq + Inter_bb_Cohort +
##     Inter_E_Cohort + Inter_bhr_Cohort + Inter_bbb_Cohort + Inter_bs_Cohort
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   2261 334871
## 2   2254 318536  7     16335 16.513 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Third model, categories of power - batting power and pitching weakness

categories

```
##      INDEX         TARGET_WINS      BATTING_H      BATTING_2B
##  Min.   :   1.0   Min.   :  0.0   Min.   : 891   Min.   : 69.0
##  1st Qu.: 620.2   1st Qu.: 73.0   1st Qu.:1469   1st Qu.:220.0
##  Median :1244.0   Median : 85.0   Median :1530   Median :254.0
##  Mean   :1258.3   Mean   : 82.7   Mean   :1547   Mean   :253.5
##  3rd Qu.:1922.8   3rd Qu.: 94.0   3rd Qu.:1604   3rd Qu.:286.0
##  Max.   :2535.0   Max.   :146.0   Max.   :2554   Max.   :458.0
##    BATTING_3B       BATTING_HR       BATTING_BB      BATTING_SO
##  Min.   :  0.00   Min.   :  3.00   Min.   :  0.0   Min.   :  66.0
##  1st Qu.: 42.00   1st Qu.: 36.25   1st Qu.:425.0   1st Qu.: 474.0
##  Median : 62.00   Median : 76.00   Median :511.0   Median : 594.5
##  Mean   : 66.25   Mean   : 90.55   Mean   :487.8   Mean   : 634.5
##  3rd Qu.: 85.00   3rd Qu.:138.75   3rd Qu.:588.0   3rd Qu.: 784.0
##  Max.   :223.00   Max.   :264.00   Max.   :878.0   Max.   :1264.0
##    BASERUN_SB      PITCHING_H     PITCHING_HR      PITCHING_BB
##  Min.   :  0.0   Min.   : 1501   Min.   :  0.00   Min.   :   0.0
##  1st Qu.: 67.0   1st Qu.: 1560   1st Qu.: 44.00   1st Qu.: 477.0
##  Median :107.5   Median : 1664   Median : 84.00   Median : 557.0
##  Mean   :133.7   Mean   : 2109   Mean   : 99.83   Mean   : 576.4
##  3rd Qu.:156.8   3rd Qu.: 1904   3rd Qu.:148.00   3rd Qu.: 650.8
##  Max.   :697.0   Max.   :30132   Max.   :343.00   Max.   :3645.0
##   PITCHING_SO        FIELDING_E      FIELDING_DP    BRSB_Missing_Flag
##  Min.   :  181.0   Min.   :  65.0   Min.   : 52.0   Min.   :0.000
##  1st Qu.:  542.2   1st Qu.: 140.0   1st Qu.:140.0   1st Qu.:0.000
##  Median :  681.0   Median : 195.0   Median :146.4   Median :0.000
##  Mean   :  776.1   Mean   : 320.0   Mean   :148.6   Mean   :0.109
##  3rd Qu.:  909.8   3rd Qu.: 423.8   3rd Qu.:162.0   3rd Qu.:0.000
##  Max.   :19278.0   Max.   :1898.0   Max.   :225.0   Max.   :1.000
##  FDP_Missing_Flag  category_PH     category_PBB    category_BH
##  Min.   :0.0000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :0.0000   Median :3.000   Median :3.000   Median :3.000
##  Mean   :0.2304   Mean   :2.998   Mean   :2.993   Mean   :2.996
##  3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##   category_BBB    category_BHR   Hitting_Power    Pitching_Weakness
##  Min.   :1.000   Min.   :1.000   Min.   : 2.000   Min.   : 2.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.: 5.000   1st Qu.: 5.000
##  Median :3.000   Median :3.000   Median : 6.000   Median : 6.000
##  Mean   :2.991   Mean   :2.984   Mean   : 5.987   Mean   : 5.992
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.: 7.000   3rd Qu.: 7.000
##  Max.   :5.000   Max.   :5.000   Max.   :10.000   Max.   :10.000
```

The two are correlated

```
## [1] 0.3967352
```

These boxplots show the stronger relationship with batting power

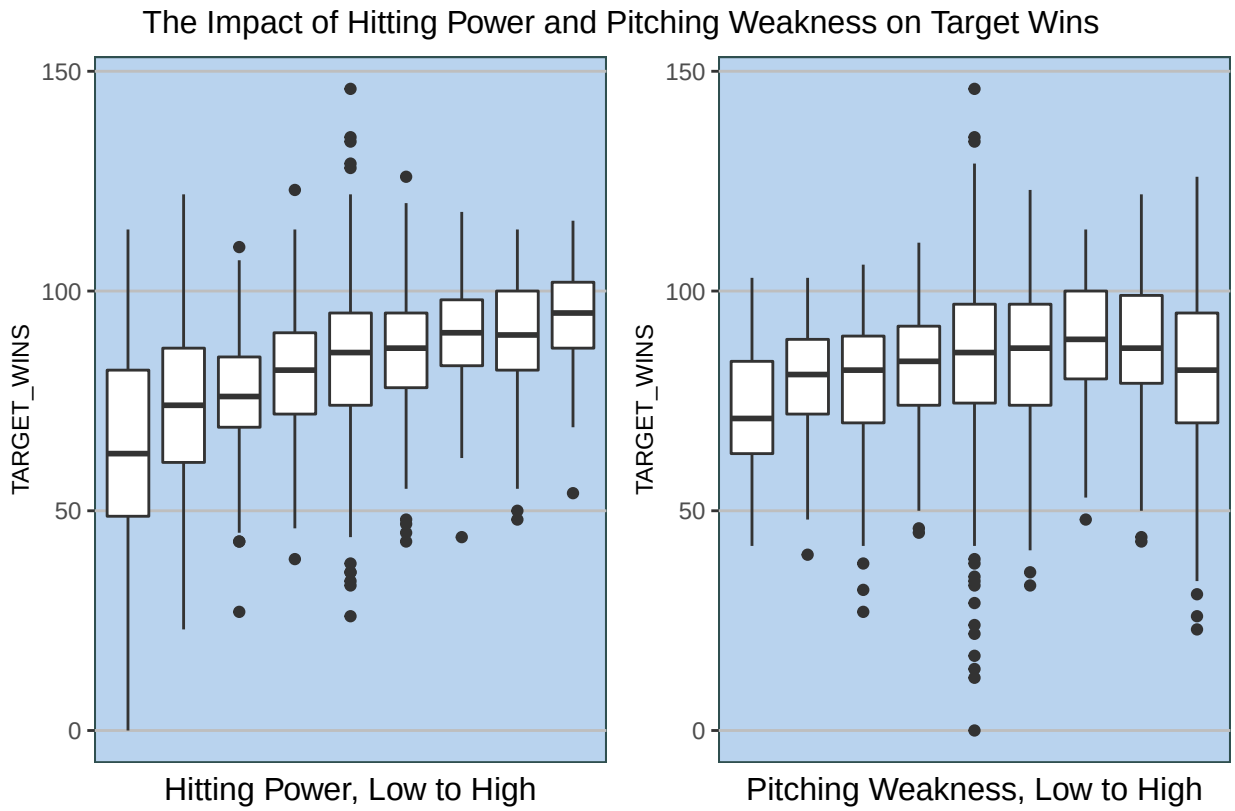


Fig. 8

add categories for eras(?)

k means clustering

we run the regressions

```
##
## Call:
## lm(formula = TARGET_WINS ~ Hitting_Power + Cluster1 + Cluster2,
##     data = dfCat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1169 -0.5336  0.0591  0.5892  3.5113
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.07327    0.04205  -1.742   0.0817 .
## Hitting_Power  0.40373    0.02928  13.790   <2e-16 ***
## Cluster1       0.20627    0.08799   2.344   0.0192 *
## Cluster2       0.09857    0.05860   1.682   0.0928 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9173 on 1198 degrees of freedom
## Multiple R-squared:  0.1607, Adjusted R-squared:  0.1586
## F-statistic: 76.47 on 3 and 1198 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ Hitting_Power + Pitching_Weakness +
##     Cluster1 + Cluster2, data = dfCat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1038 -0.5163  0.0527  0.5879  3.4518
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.06743    0.04208  -1.602  0.10933
## Hitting_Power      0.44118    0.03418  12.907  < 2e-16 ***
## Pitching_Weakness -0.06616    0.03128  -2.115  0.03465 *
## Cluster1           0.25999    0.09146   2.843  0.00455 **
## Cluster2           0.07141    0.05990   1.192  0.23347
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9159 on 1197 degrees of freedom
## Multiple R-squared:  0.1638, Adjusted R-squared:  0.161
## F-statistic: 58.64 on 4 and 1197 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ category_PH + category_PBB + category_BH +
##     category_BBB + category_BHR + Cluster1 + Cluster2, data = dfCat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0387 -0.5154  0.0618  0.5953  3.4084
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.089192   0.055817  -1.598  0.11032
## category_PH  -0.025441   0.043986  -0.578  0.56310
## category_PBB -0.023434   0.053097  -0.441  0.65904
## category_BH   0.340639   0.032324  10.538  < 2e-16 ***
## category_BBB  0.210230   0.058869   3.571  0.00037 ***
## category_BHR -0.001569   0.051933  -0.030  0.97590
## Cluster1      0.142166   0.106801   1.331  0.18340
## Cluster2      0.149966   0.104648   1.433  0.15210
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9144 on 1194 degrees of freedom
## Multiple R-squared:  0.1687, Adjusted R-squared:  0.1639
## F-statistic: 34.62 on 7 and 1194 DF,  p-value: < 2.2e-16
```

Analysis shows good batting and weak pitching are correlated. Poor r squared but significant batting.

Select models

Now we make predictions