Moneyball - CUNY Data Science 621

Eric Hirsch

2/20/2021

Description of the Dataset

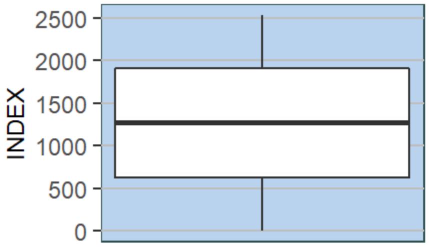
xxxxxx

Data Exploration

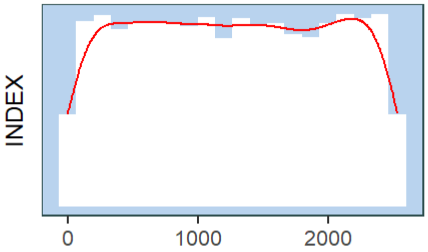
summary info

```
##      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
```

boxplots and correlations

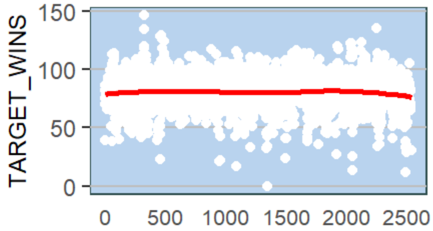


L: 1, 2, 3, 4, 5\
H: 2531, 2532, 2533, 2534, 2535

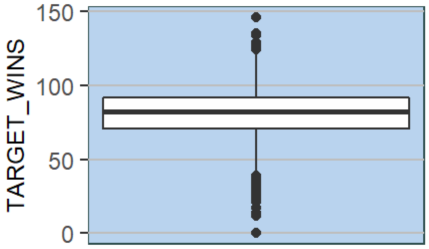


L: 1, 2, 3, 4, 5\
H: 2531, 2532, 2533, 2534, 2535

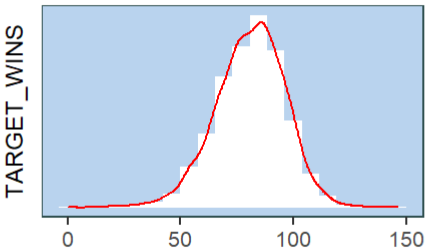
INDEX



Correlation: -0.02 p value: 0.32

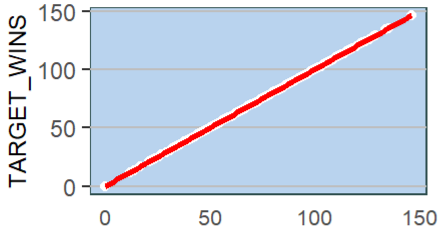


L: 0, 12, 14, 17, 21\
H: 128, 129, 134, 135, 146

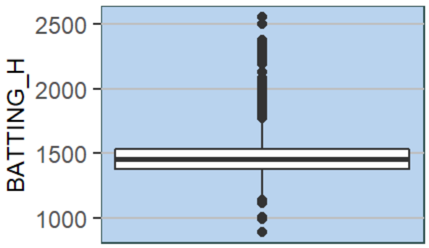


L: 0, 12, 14, 17, 21\
H: 128, 129, 134, 135, 146

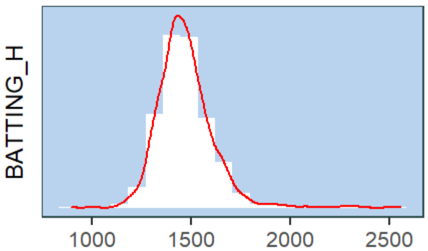
TARGET_WINS



Correlation: 1 p value: 0

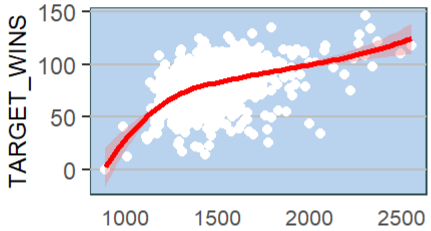


L: 891, 992, 1009, 1116, 1122\
H: 2333, 2343, 2372, 2496, 2554

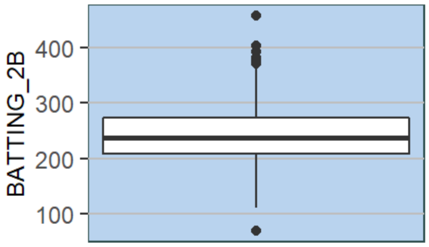


L: 891, 992, 1009, 1116, 1122\
H: 2333, 2343, 2372, 2496, 2554

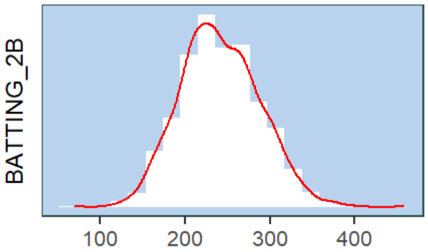
BATTING_H



Correlation: 0.39 p value: 0

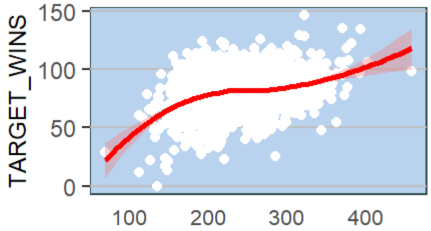


L: 69, 112, 113, 118, 123\
H: 382, 392, 393, 403, 458

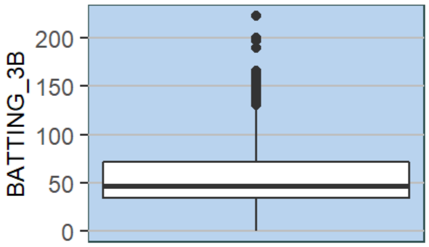


L: 69, 112, 113, 118, 123\
H: 382, 392, 393, 403, 458

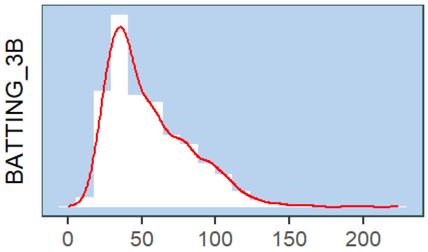
BATTING_2B



Correlation: 0.29 p value: 0

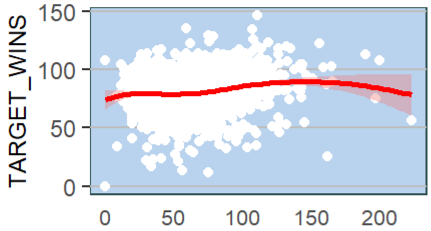


L: 0, 0, 8, 9, 11\
H: 166, 190, 197, 200, 223

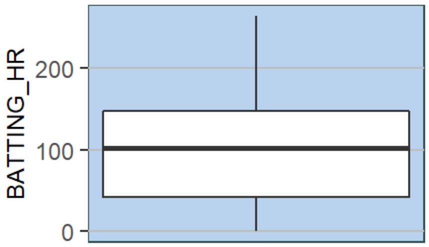


L: 0, 0, 8, 9, 11\
H: 166, 190, 197, 200, 223

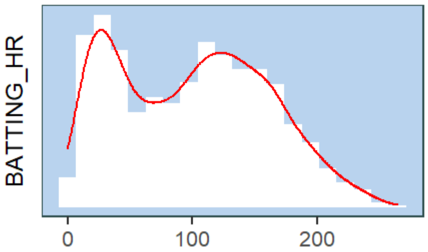
BATTING_3B



Correlation: 0.14 p value: 0

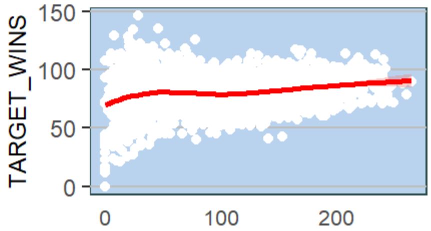


L: 0, 0, 0, 0, 0\
H: 247, 249, 257, 260, 264

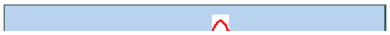


L: 0, 0, 0, 0, 0\
H: 247, 249, 257, 260, 264

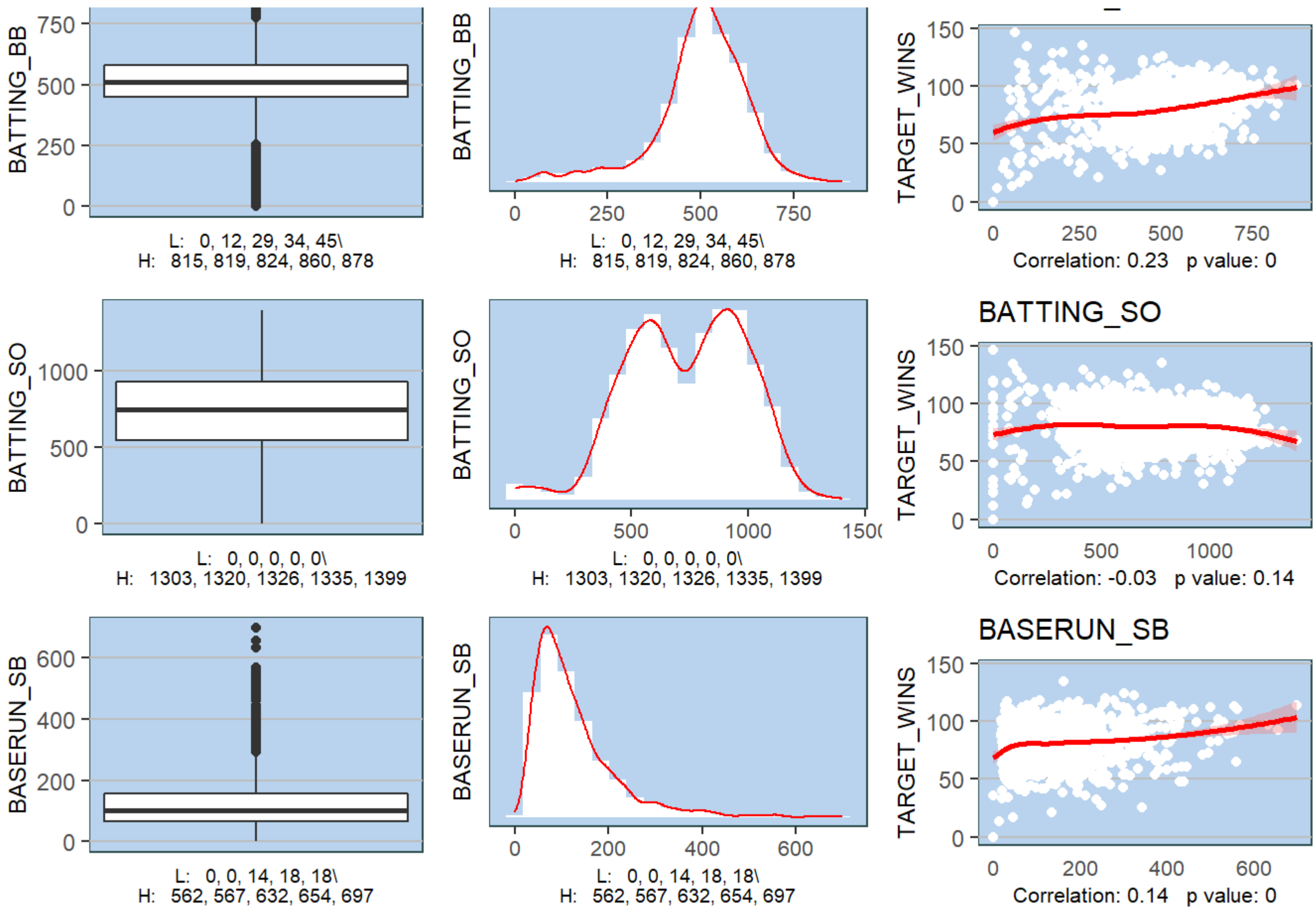
BATTING_HR

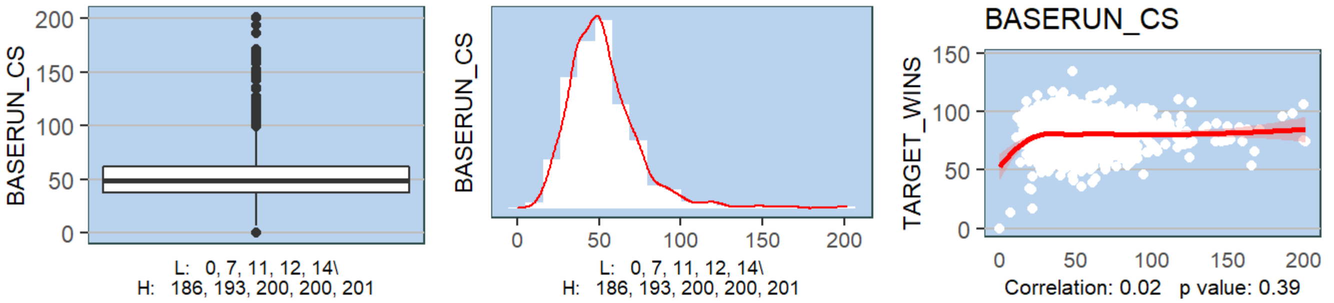


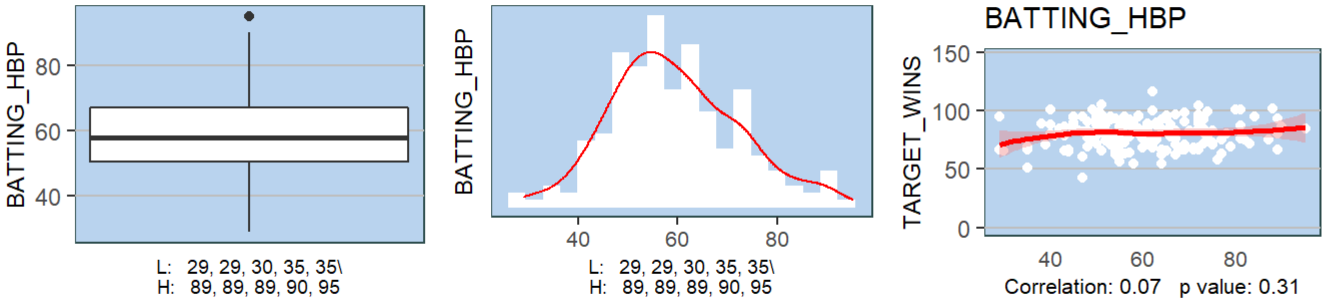
Correlation: 0.18 p value: 0

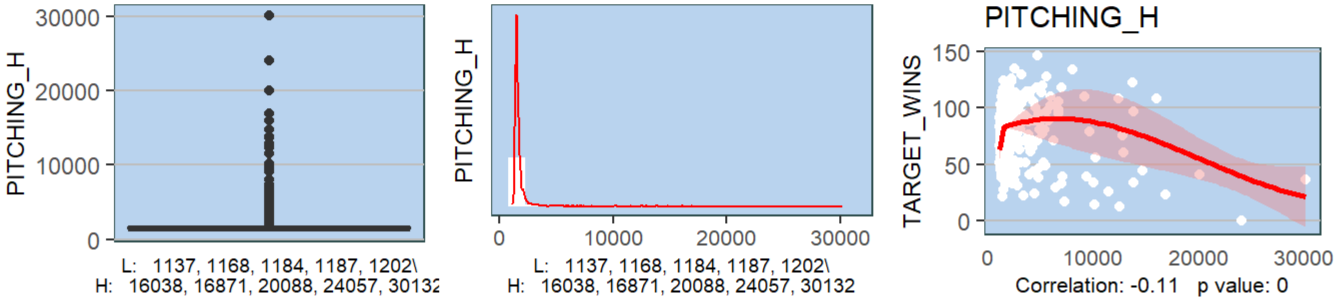


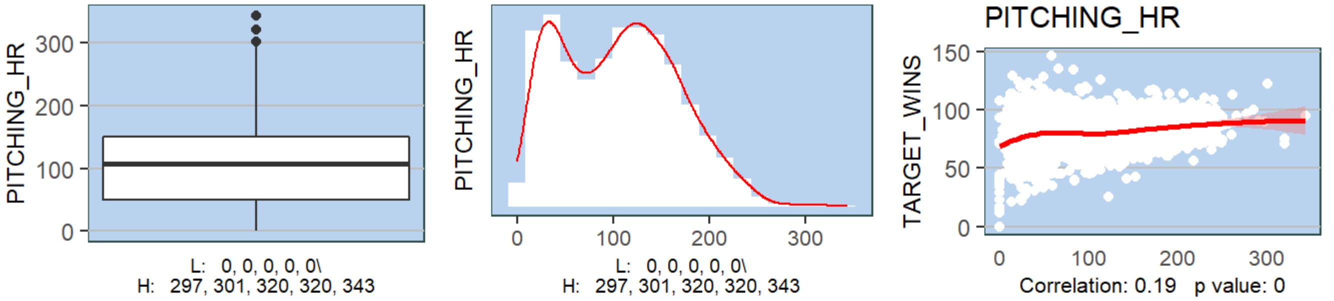
BATTING BB

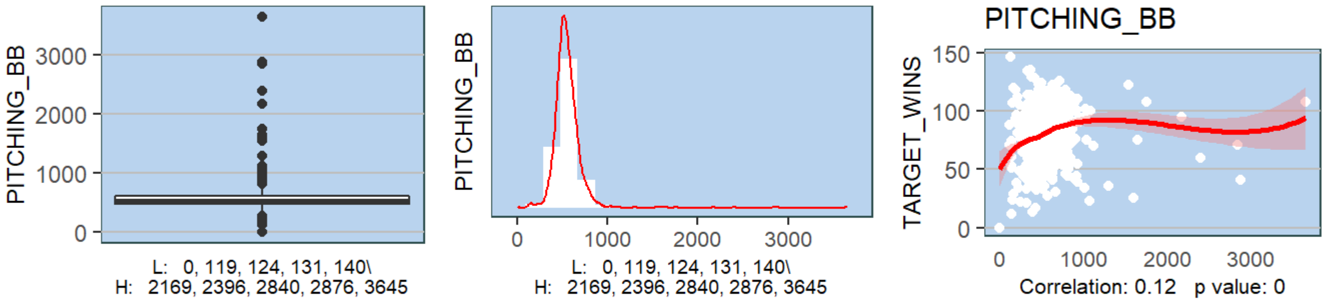

BATTING_BB
750
500
250
0
L: 0, 12, 29, 34, 45\
H: 815, 819, 824, 860, 878
BATTING_BB
0
250
500
750
L: 0, 12, 29, 34, 45\
H: 815, 819, 824, 860, 878
TARGET_WINS
150
100
50
0
0
250
500
750
Correlation: 0.23 p value: 0
BATTING_SO
BATTING_SO
1000
500
0
L: 0, 0, 0, 0, 0\
H: 1303, 1320, 1326, 1335, 1399
BATTING_SO
0
500
1000
150(
L: 0, 0, 0, 0, 0\
H: 1303, 1320, 1326, 1335, 1399
TARGET_WINS
150
100
50
0
0
500
1000
Correlation: -0.03 p value: 0.14
BASERUN_SB
BASERUN_SB
600
400
200
0
L: 0, 0, 14, 18, 18\
H: 562, 567, 632, 654, 697
BASERUN_SB
0
200
400
600
L: 0, 0, 14, 18, 18\
H: 562, 567, 632, 654, 697
TARGET_WINS
150
100
50
0
0
200
400
600
Correlation: 0.14 p value: 0

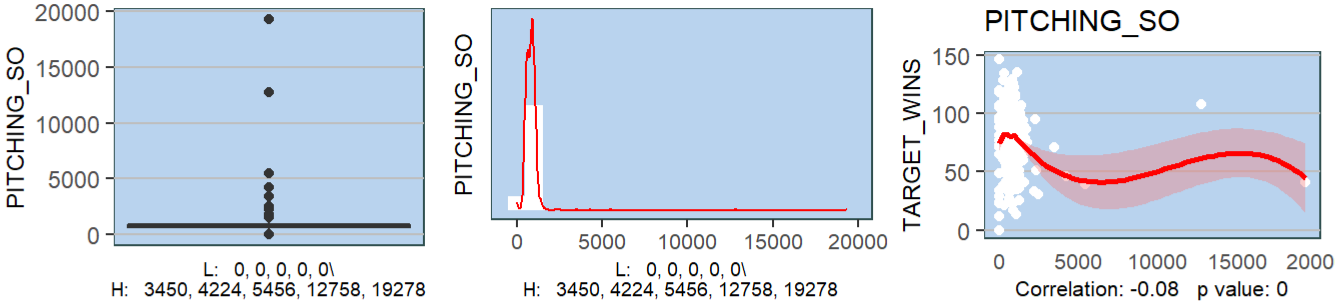

BASERUN_CS
200
150
100
50
0
L: 0, 7, 11, 12, 14\
H: 186, 193, 200, 200, 201
BASERUN_CS
0
50
100
150
200
L: 0, 7, 11, 12, 14\
H: 186, 193, 200, 200, 201
BASERUN_CS
TARGET_WINS
150
100
50
0
0
50
100
150
200
Correlation: 0.02 p value: 0.39

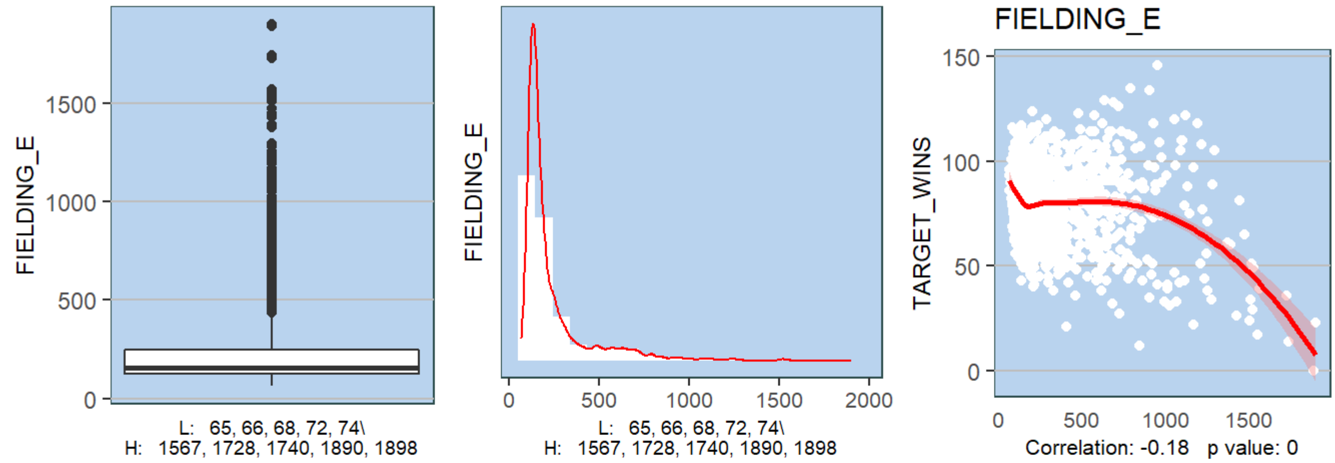

BATTING_HBP
80
60
40
L: 29, 29, 30, 35, 35\
H: 89, 89, 89, 90, 95
BATTING_HBP
40
60
80
L: 29, 29, 30, 35, 35\
H: 89, 89, 89, 90, 95
BATTING_HBP
TARGET_WINS
150
100
50
0
40
60
80
Correlation: 0.07 p value: 0.31

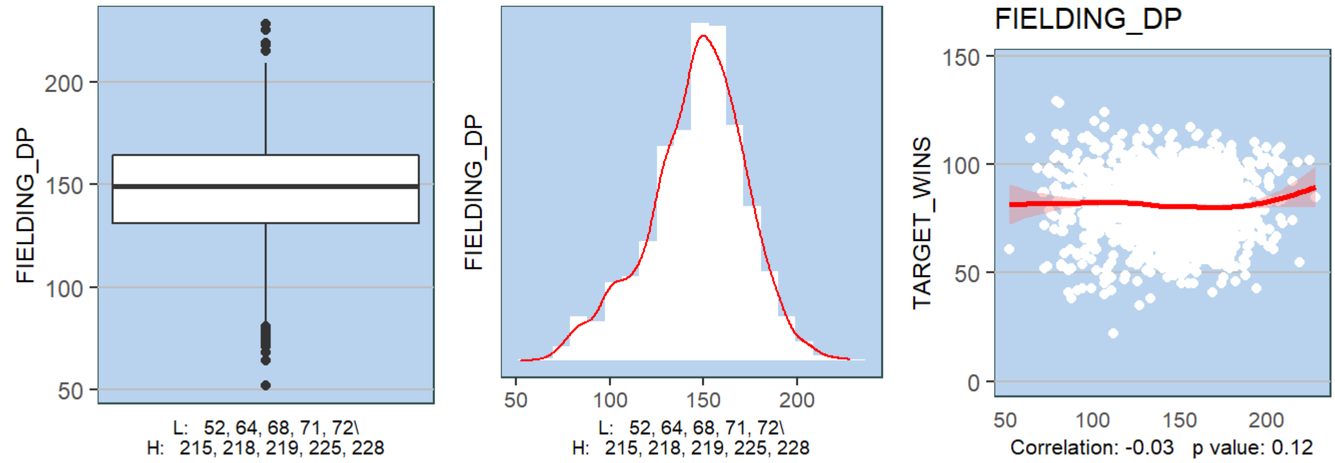

PITCHING_H
30000
20000
10000
0
L: 1137, 1168, 1184, 1187, 1202\
H: 16038, 16871, 20088, 24057, 30132
PITCHING_H
0
10000
20000
30000
L: 1137, 1168, 1184, 1187, 1202\
H: 16038, 16871, 20088, 24057, 30132
PITCHING_H
TARGET_WINS
150
100
50
0
0
10000
20000
30000
Correlation: -0.11 p value: 0


PITCHING_HR
300
200
100
0
L: 0, 0, 0, 0, 0\
H: 297, 301, 320, 320, 343
PITCHING_HR
0
100
200
300
L: 0, 0, 0, 0, 0\
H: 297, 301, 320, 320, 343
PITCHING_HR
TARGET_WINS
150
100
50
0
0
100
200
300
Correlation: 0.19 p value: 0


PITCHING_BB
3000
2000
1000
0
L: 0, 119, 124, 131, 140\
H: 2169, 2396, 2840, 2876, 3645
PITCHING_BB
0
1000
2000
3000
L: 0, 119, 124, 131, 140\
H: 2169, 2396, 2840, 2876, 3645
PITCHING_BB
TARGET_WINS
150
100
50
0
0
1000
2000
3000
Correlation: 0.12 p value: 0

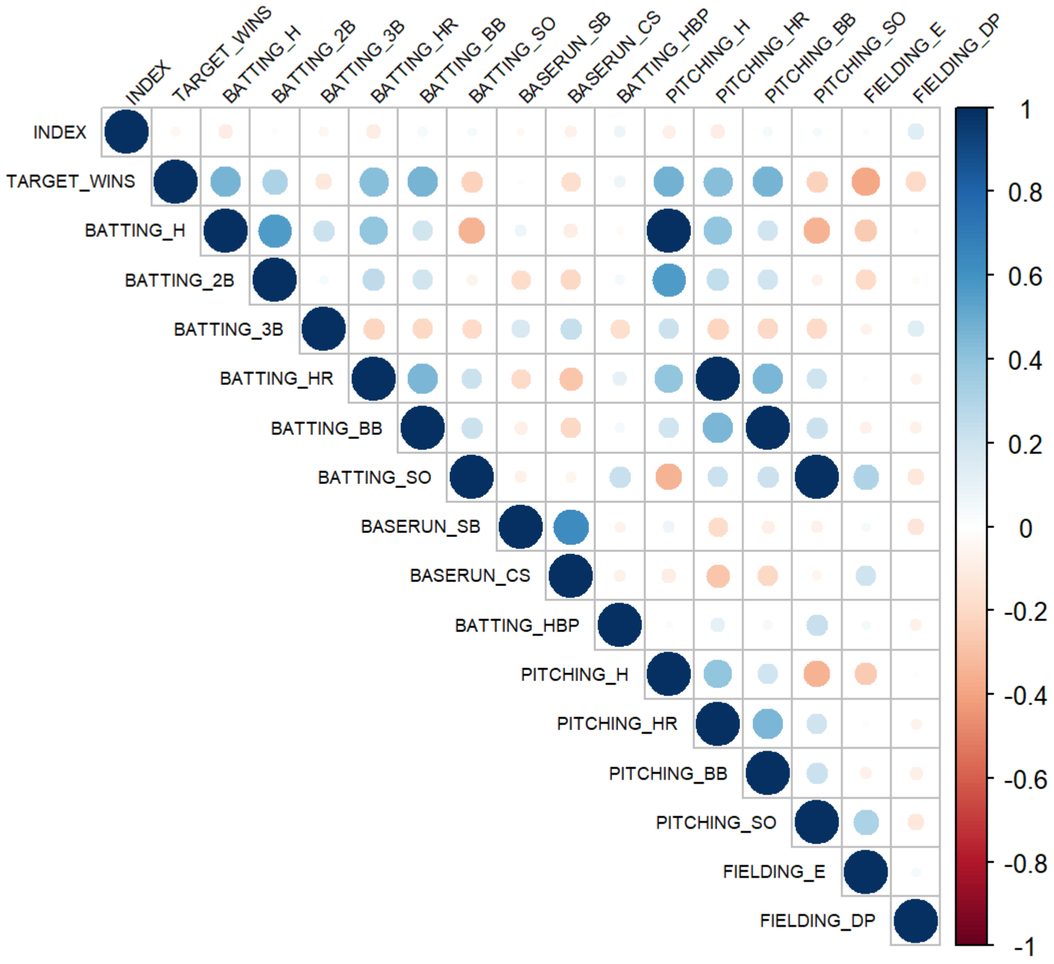

PITCHING_SO
20000
15000
10000
5000
0
L: 0, 0, 0, 0, 0\
H: 3450, 4224, 5456, 12758, 19278
PITCHING_SO
0
5000
10000
15000
20000
L: 0, 0, 0, 0, 0\
H: 3450, 4224, 5456, 12758, 19278
PITCHING_SO
TARGET_WINS
150
100
50
0
0
5000
10000
15000
2000
Correlation: -0.08 p value: 0





dealing with outliers

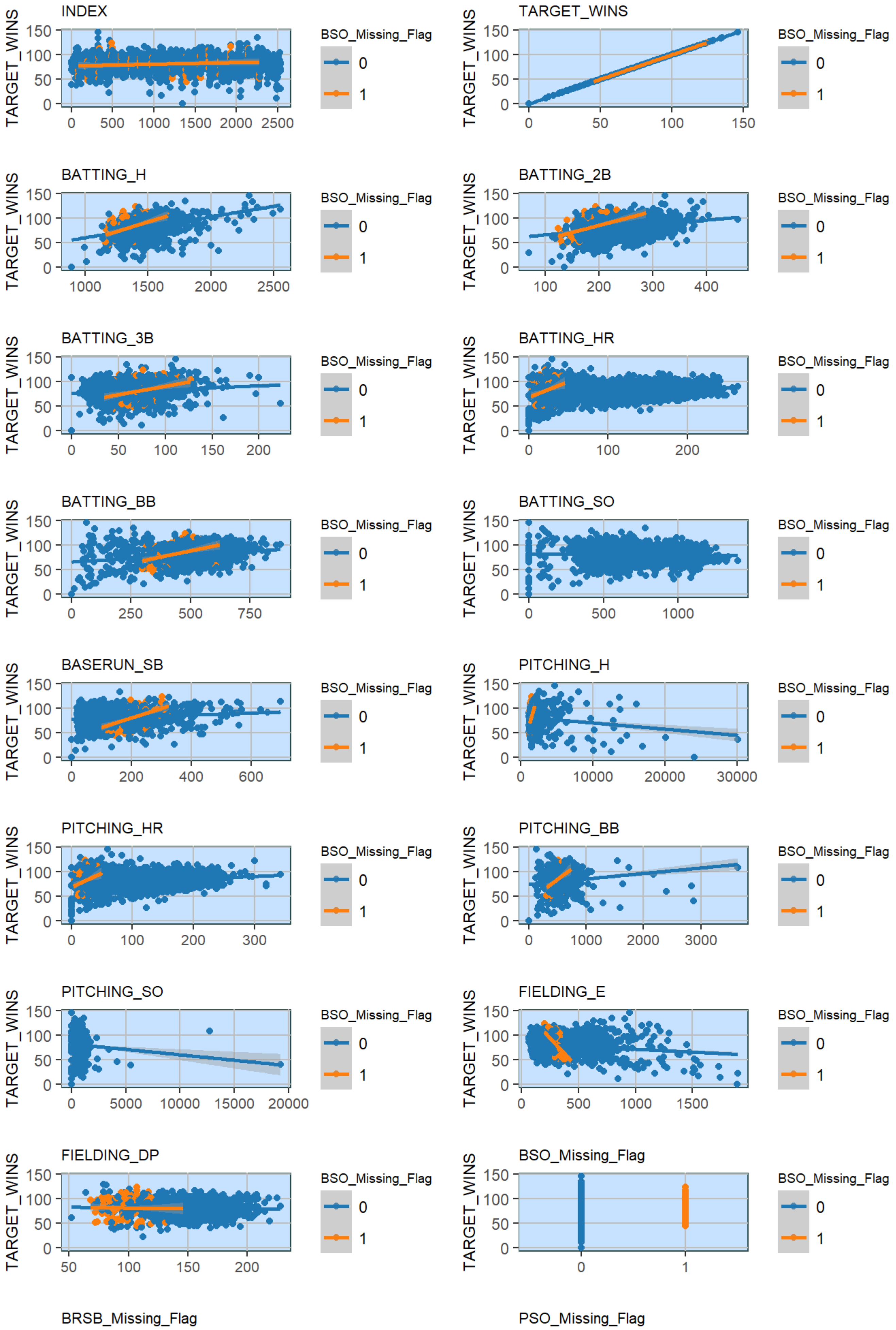
looking at mutlicollinearlity

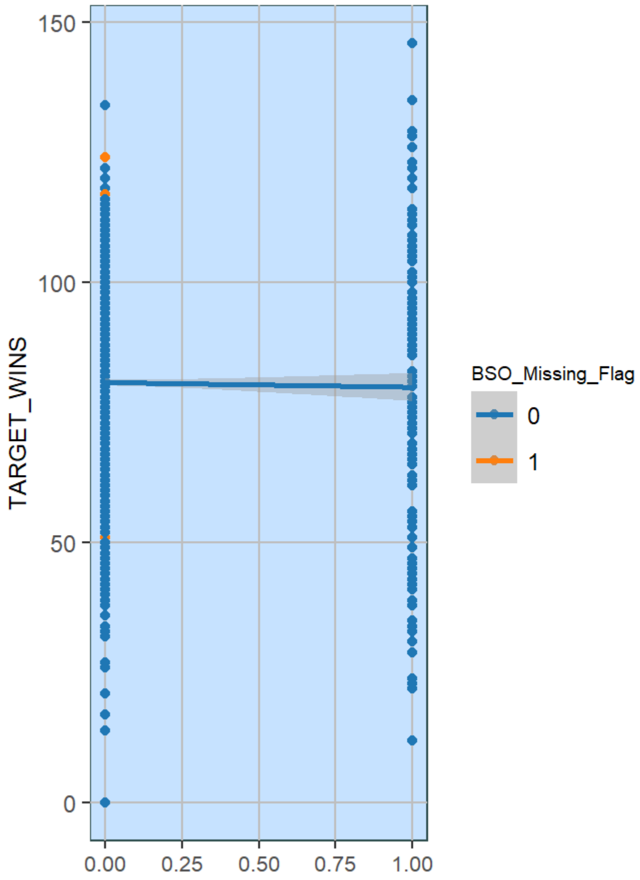


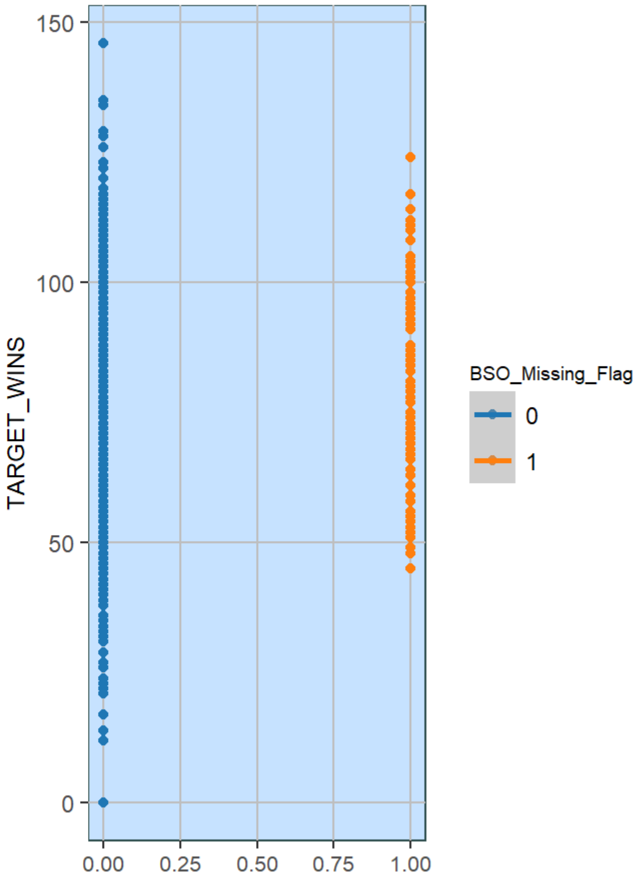
####

Data Preparation

Dealing with na - get rid of hbp and cs, create flags, check if Batting_SO is random


INDEX
TARGET_WINS
150
100
50
0
0
500
1000
1500
2000
2500
BSO_Missing_Flag
0
1
TARGET_WINS
TARGET_WINS
0
50
100
150
BATTING_H
1000
1500
2000
2500
BATTING_2B
100
200
300
400
BATTING_3B
0
50
100
150
200
BATTING_HR
0
100
200
BATTING_BB
0
250
500
750
BATTING_SO
0
500
1000
BASERUN_SB
0
200
400
600
PITCHING_H
0
10000
20000
30000
PITCHING_HR
0
100
200
300
PITCHING_BB
0
1000
2000
3000
PITCHING_SO
0
5000
10000
15000
20000
FIELDING_E
0
500
1000
1500
FIELDING_DP
50
100
150
200
BSO_Missing_Flag
0
1
BRSB_Missing_Flag
PSO_Missing_Flag





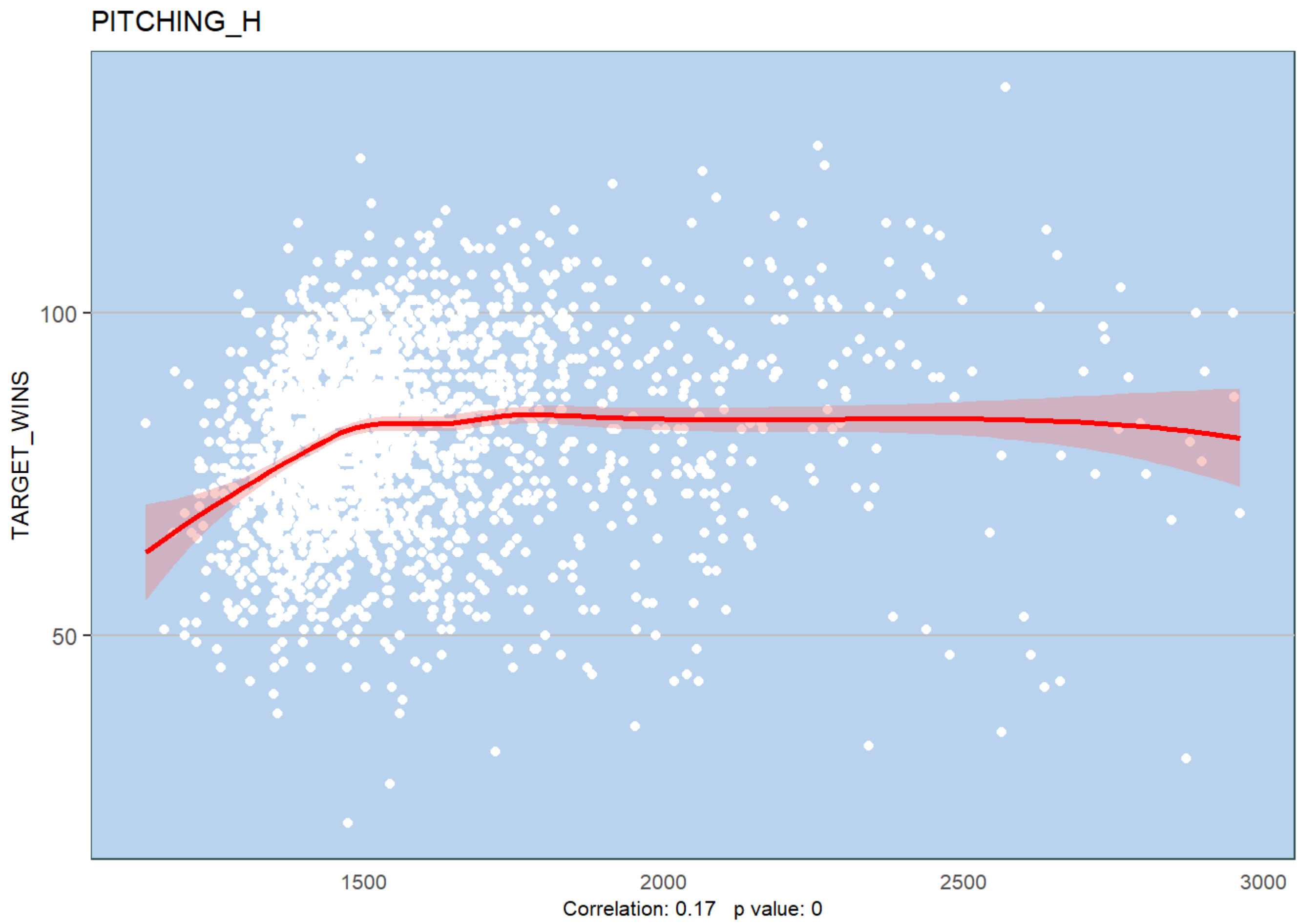
Imputation missing values

```
## [1] "type:" "mean"
## [1] "r2mean:" "0.4074"
## [1] "r2median:" "0.4074"
## [1] "r2omit" "0.4019"
```

```
## [1] "type:" "mean"
## [1] "r2mean:" "0.4074"
## [1] "r2median:" "0.4074"
## [1] "r2omit" "0.4019"
```

Transformations:

pitching_h - deal with odd behavior



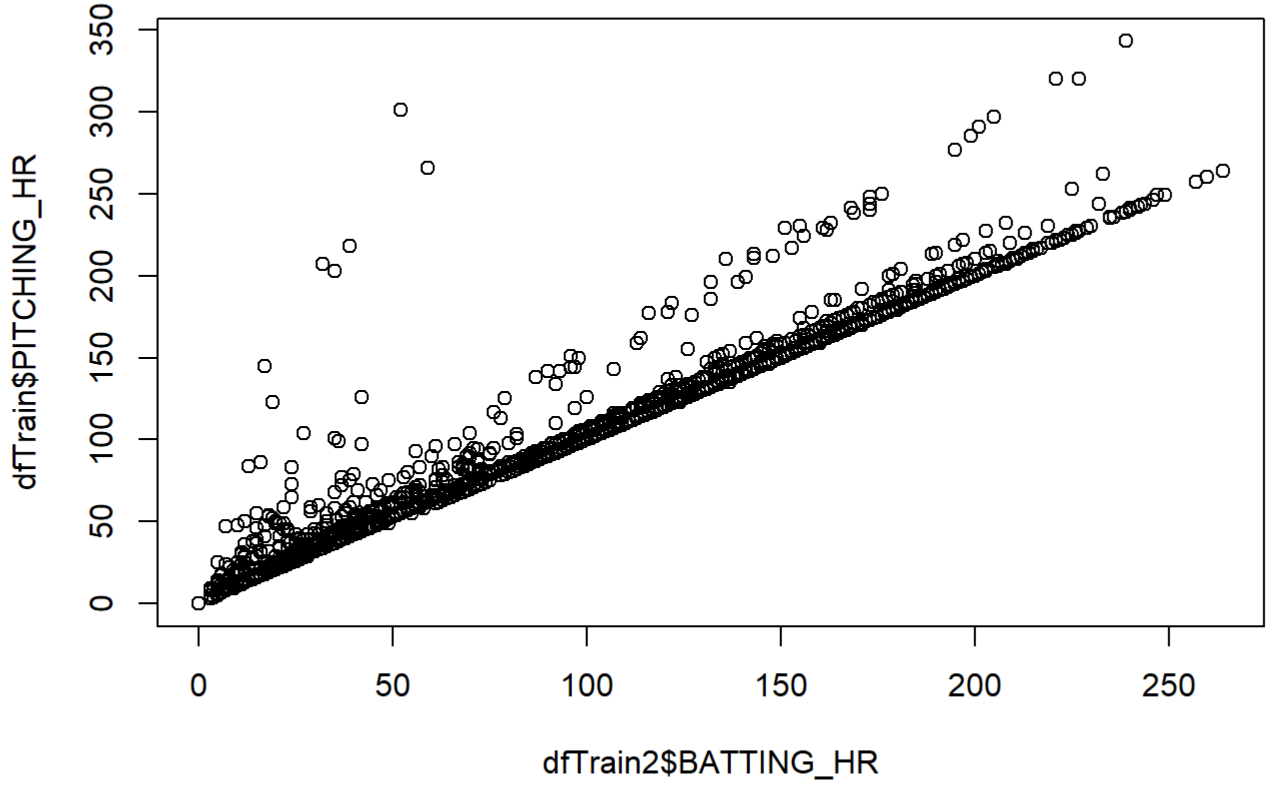
F_DP - deal with relationship to hits

```
## [1] 0.0003924305
```

```
## [1] 0.01398409
```

```
## [1] 0.02142144
```

Home runs - deal with implausibility (drop pitch)



deal

with bimodaility(create HR under 60)

```
## [1] 0.03060384
```

```
## [1] 0.03617586
```

create new errors field

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

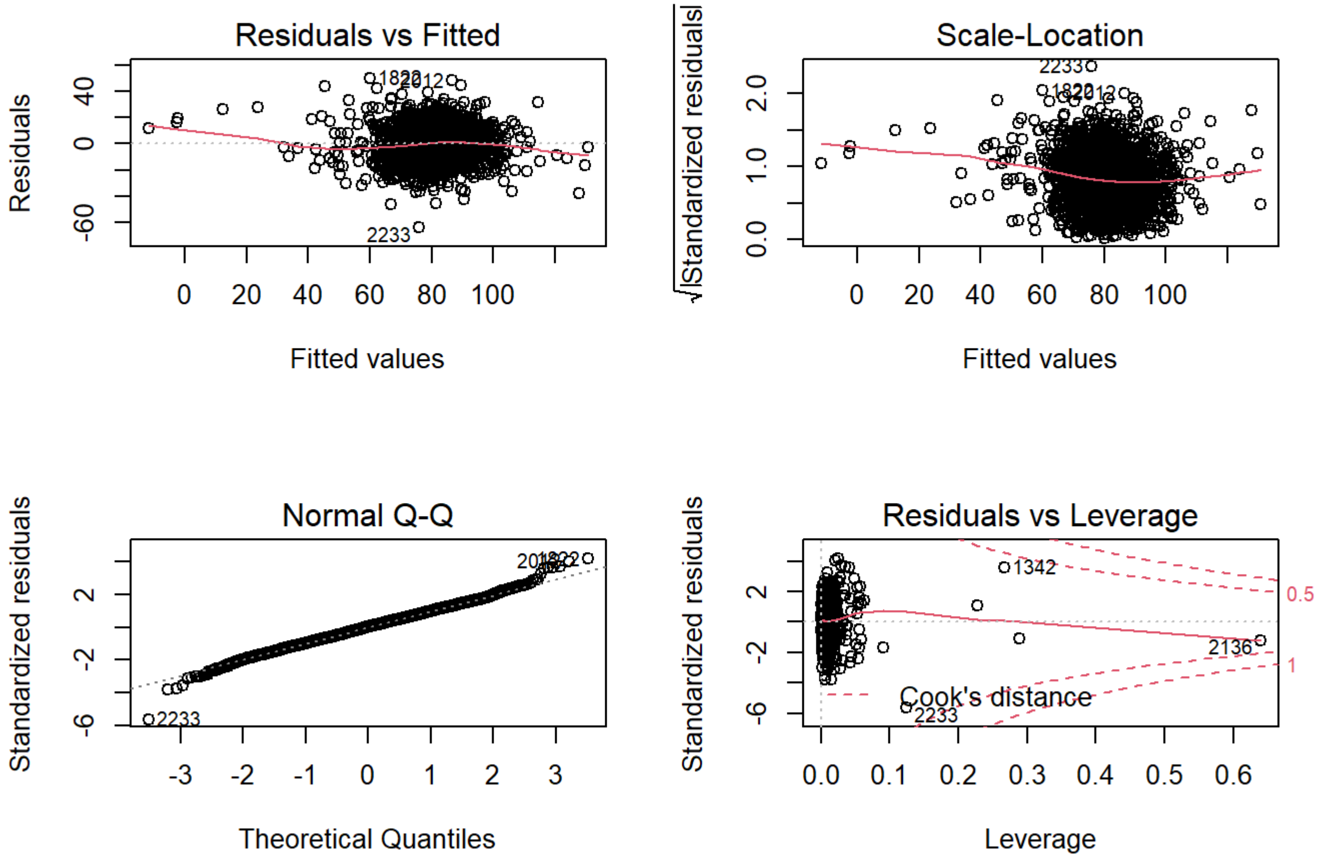
```
## [1] 0.04825783
```

Create Missing “cohort” interactions

Build Models

Create regression 1 - no transformations except missing flags

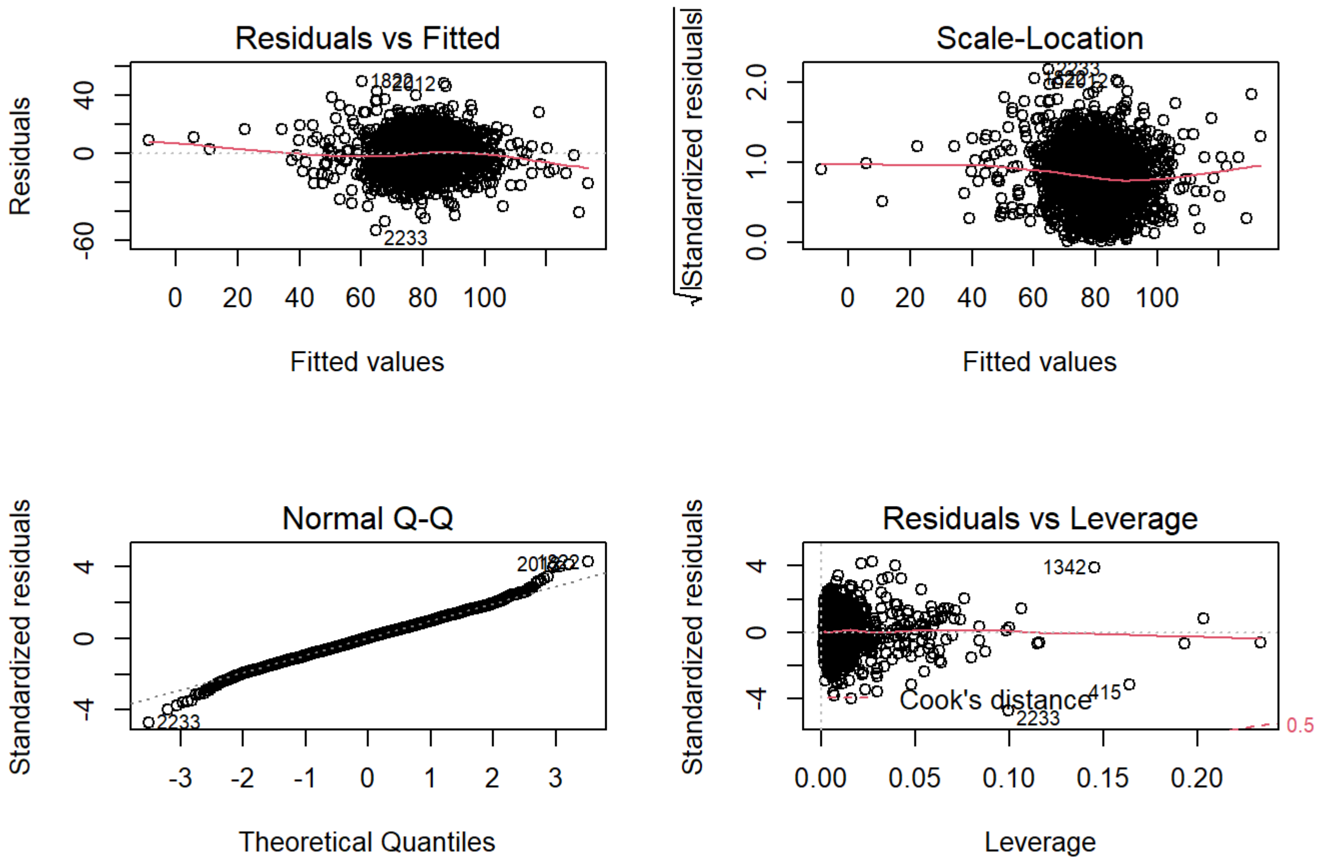
```
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B +
##     BATTING_HR + BATTING_BB + BATTING_SO + BASERUN_SB + PITCHING_H +
##     PITCHING_SO + FIELDING_E + FIELDING_DP + BSO_Missing_Flag +
##     BRSB_Missing_Flag + FDP_Missing_Flag, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.849  -7.994   0.338   7.926  49.898
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       20.7188116  5.1141823   4.051 5.27e-05 ***
## BATTING_H          0.0481387  0.0034185  14.082  < 2e-16 ***
## BATTING_2B        -0.0357462  0.0086576  -4.129 3.78e-05 ***
## BATTING_3B         0.0550541  0.0155444   3.542 0.000406 ***
## BATTING_HR         0.0711042  0.0091701   7.754 1.34e-14 ***
## BATTING_BB         0.0253016  0.0032449   7.797 9.58e-15 ***
## BATTING_SO        -0.0107507  0.0023244  -4.625 3.95e-06 ***
## BASERUN_SB         0.0497807  0.0046076  10.804  < 2e-16 ***
## PITCHING_H         0.0019828  0.0003352   5.916 3.81e-09 ***
## PITCHING_SO       -0.0009736  0.0006627  -1.469 0.141921
## FIELDING_E        -0.0570291  0.0033828 -16.859  < 2e-16 ***
## FIELDING_DP       -0.1003640  0.0134425  -7.466 1.17e-13 ***
## BSO_Missing_Flag   8.7007564  1.4580508   5.967 2.79e-09 ***
## BRSB_Missing_Flag 33.7150508  1.8190459  18.534  < 2e-16 ***
## FDP_Missing_Flag   4.2670950  1.4589567   2.925 0.003482 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.13 on 2261 degrees of freedom
## Multiple R-squared:  0.411,  Adjusted R-squared:  0.4074
## F-statistic: 112.7 on 14 and 2261 DF,  p-value: < 2.2e-16
```



```
## NULL
##         BATTING_H        BATTING_2B        BATTING_3B        BATTING_HR
##          3.779947          2.540104          2.918055          4.769509
##        BATTING_BB        BATTING_SO        BASERUN_SB        PITCHING_H
##          2.451529          4.931438          2.385757          3.440215
##       PITCHING_SO        FIELDING_E       FIELDING_DP  BSO_Missing_Flag
##          1.985417          9.184956          1.681234          1.408612
## BRSB_Missing_Flag  FDP_Missing_Flag
##          2.778256          3.619847
```

Model 2, all the 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 + E_sq + Inter_E_Cohort +
##     Inter_bhr_Cohort + Inter_bs_Cohort, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.914  -7.727   0.159   7.657  49.775
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.036e+01  5.340e+00   3.812 0.000141 ***
## BATTING_H          5.197e-02  3.369e-03  15.428  < 2e-16 ***
## BATTING_2B        -3.576e-02  8.564e-03  -4.176 3.08e-05 ***
## BATTING_3B         7.470e-02  1.573e-02   4.749 2.17e-06 ***
## BATTING_HR         6.411e-02  9.163e-03   6.997 3.44e-12 ***
## BATTING_BB         2.574e-02  3.186e-03   8.079 1.05e-15 ***
## BATTING_SO        -1.307e-02  2.163e-03  -6.046 1.74e-09 ***
## BASERUN_SB         5.110e-02  4.674e-03  10.934  < 2e-16 ***
## PITCHING_H         1.020e-03  2.955e-04   3.452 0.000566 ***
## FIELDING_E        -8.132e-02  7.033e-03 -11.563  < 2e-16 ***
## FIELDING_DP       -1.063e-01  1.333e-02  -7.972 2.46e-15 ***
## BSO_Missing_Flag   4.677e+01  9.614e+00   4.864 1.23e-06 ***
## BRSB_Missing_Flag  3.562e+01  1.855e+00  19.204  < 2e-16 ***
## FDP_Missing_Flag   6.209e+00  1.507e+00   4.119 3.94e-05 ***
## Pitch_h_Under1500  2.018e+00  6.760e-01   2.985 0.002864 **
## E_sq               1.897e-05  4.097e-06   4.631 3.85e-06 ***
## Inter_E_Cohort    -1.801e-01  2.620e-02  -6.874 8.05e-12 ***
## Inter_bhr_Cohort   3.514e-01  1.515e-01   2.320 0.020404 *
## Inter_bs_Cohort    4.870e-02  2.397e-02   2.032 0.042287 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.87 on 2257 degrees of freedom
## Multiple R-squared:  0.437,  Adjusted R-squared:  0.4325
## F-statistic: 97.34 on 18 and 2257 DF,  p-value: < 2.2e-16
```



```
## NULL
##          BATTING_H        BATTING_2B        BATTING_3B        BATTING_HR 
##           3.833033          2.595576          3.120382          4.973356 
##         BATTING_BB        BATTING_SO        BASERUN_SB        PITCHING_H 
##           2.468116          4.457772          2.563766          2.792528 
##         FIELDING_E       FIELDING_DP  BSO_Missing_Flag BRSB_Missing_Flag 
##          41.459797          1.726987         63.952655          3.017211 
##   FDP_Missing_Flag Pitch_h_Under1500              E_sq     Inter_E_Cohort 
##           4.035107          1.833498         22.339941         44.481566 
##   Inter_bhr_Cohort   Inter_bs_Cohort 
##           7.360403         17.564481
```

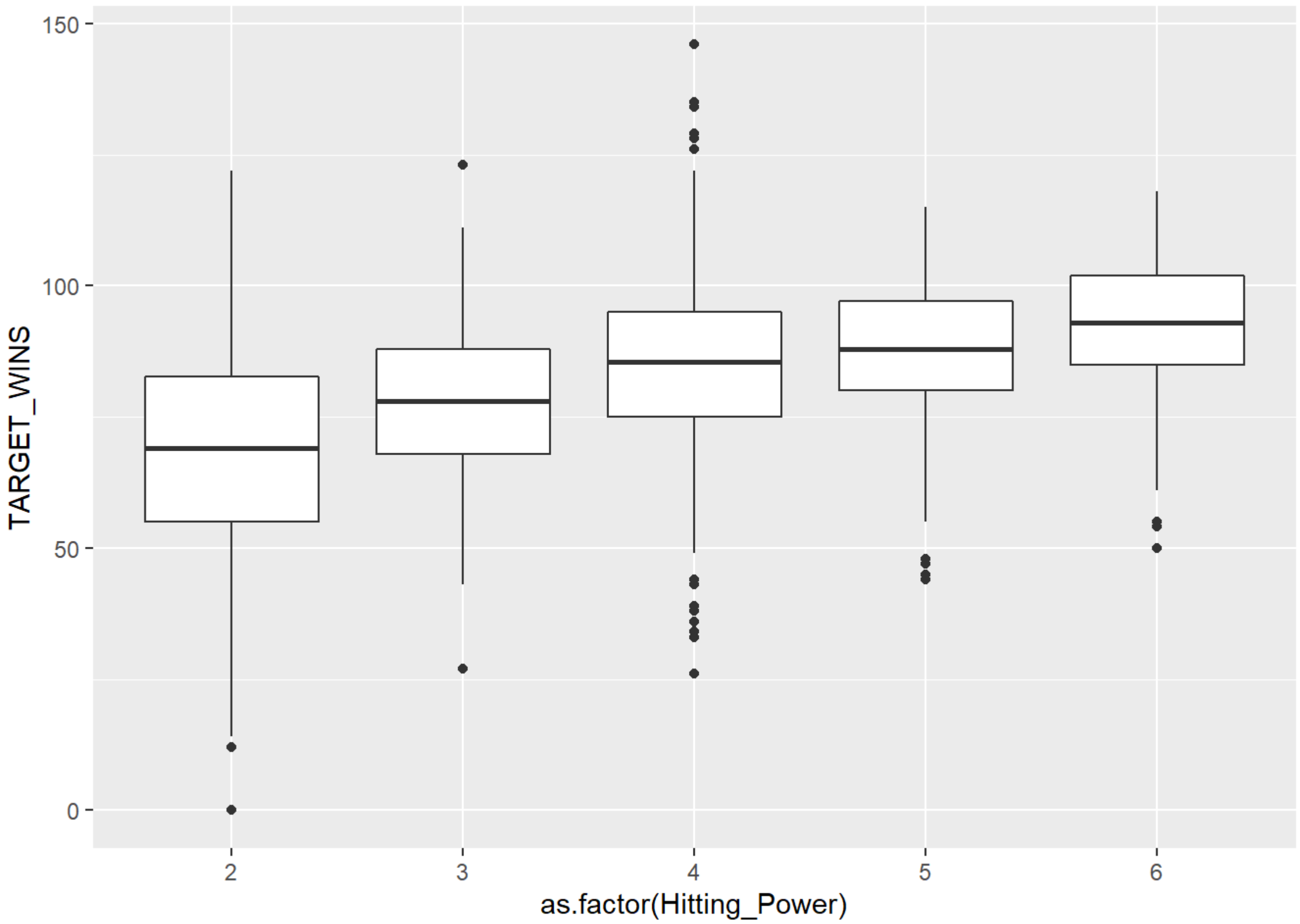
second model explains better, but does not necessarily perform lot better.

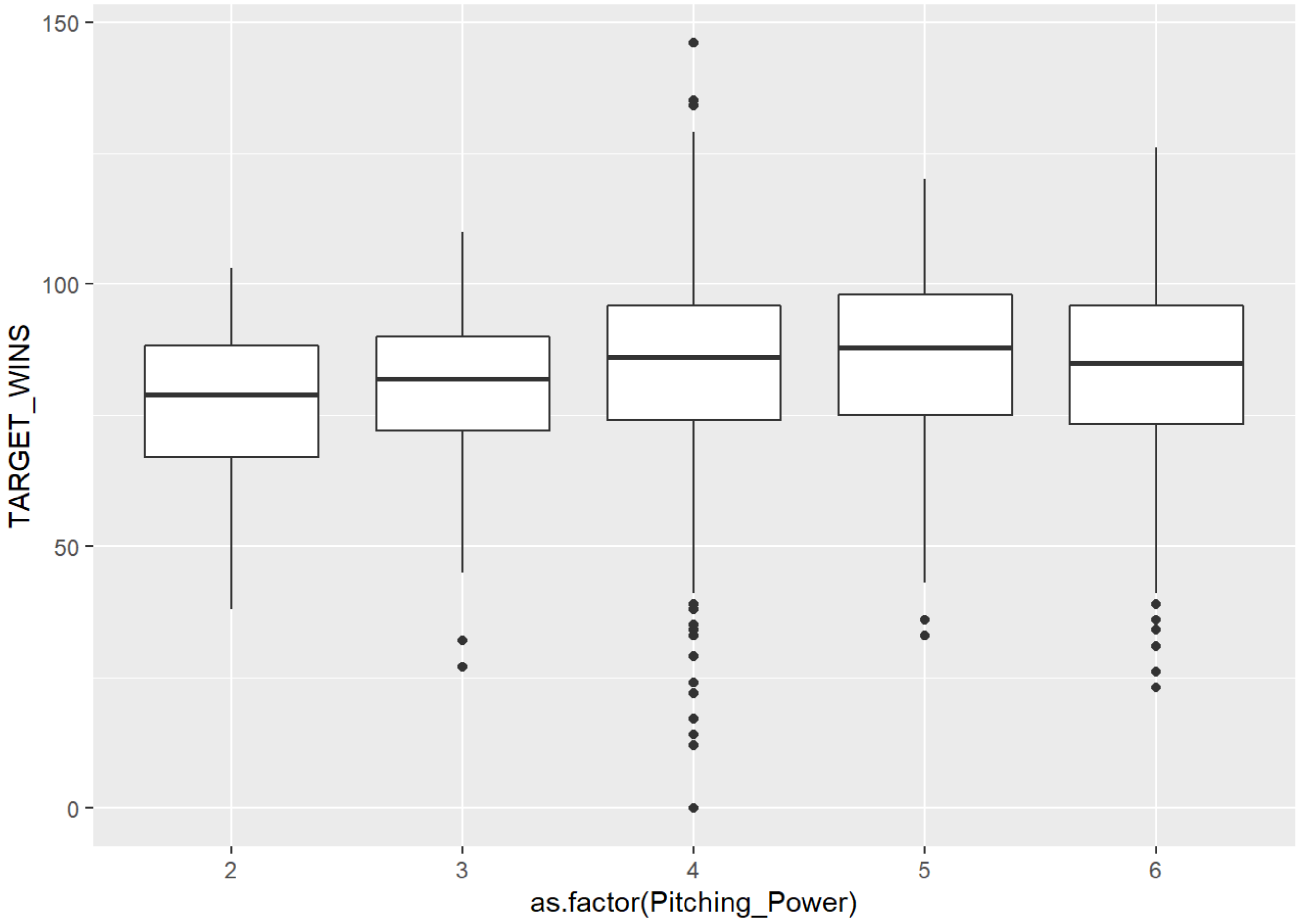
Third model, categories of power - batting power and pitching weakness

categories

```
##
## Call:
## lm(formula = TARGET_WINS ~ Hitting_Power + Pitching_Power, data = dfCat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -70.746  -9.493   0.777   9.752  63.271
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     61.7831     1.9586   31.55   <2e-16 ***
## Hitting_Power    5.9918     0.4292   13.96   <2e-16 ***
## Pitching_Power  -0.7552     0.4364   -1.73   0.0838 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.02 on 1199 degrees of freedom
## Multiple R-squared:  0.1501, Adjusted R-squared:  0.1487
## F-statistic: 105.9 on 2 and 1199 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ category_PH + category_PBB + category_BH +
##     category_BBB + category_BHR, data = dfCat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -70.394  -9.489   0.810   9.921  62.641
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  58.38917    2.32797  25.082  < 2e-16 ***
## category_PH  -0.08504    0.75686  -0.112 0.910557
## category_PBB -0.27344    0.96749  -0.283 0.777514
## category_BH   6.48270    0.64973   9.978  < 2e-16 ***
## category_BBB  3.89931    1.03826   3.756 0.000181 ***
## category_BHR  2.15128    0.70765   3.040 0.002417 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.94 on 1196 degrees of freedom
## Multiple R-squared:  0.1604, Adjusted R-squared:  0.1568
## F-statistic: 45.68 on 5 and 1196 DF,  p-value: < 2.2e-16
```


150
100
50
0
TARGET_WINS
2
3
4
5
6
as.factor(Hitting_Power)


150
100
50
0
TARGET_WINS
2
3
4
5
6
as.factor(Pitching_Power)

```
## [1] 0.3946243
```

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

Select models

Now we make predictions