hw1

Shazia Khan

June 6, 2016

1. Exploratory Data Analysis

Each of the record represents a professional baseball team from the years 1871 to 2006 inclusive.

The data preview and summary:

```
## [1] "C:/Users/SR/Documents/Data 621/hw1"
```

```
##   INDEX TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
## 1     1          39           1445             194              39
## 2     2          70           1339             219              22
## 3     3          86           1377             232              35
## 4     4          70           1387             209              38
## 5     5          82           1297             186              27
## 6     6          75           1279             200              36
##   TEAM_BATTING_HR TEAM_BATTING_BB TEAM_BATTING_SO TEAM_BASERUN_SB
## 1              13             143             842              NA
## 2             190             685            1075              37
## 3             137             602             917              46
## 4              96             451             922              43
## 5             102             472             920              49
## 6              92             443             973             107
##   TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR
## 1              NA               NA            9364               84
## 2              28               NA            1347              191
## 3              27               NA            1377              137
## 4              30               NA            1396               97
## 5              39               NA            1297              102
## 6              59               NA            1279               92
##   TEAM_PITCHING_BB TEAM_PITCHING_SO TEAM_FIELDING_E TEAM_FIELDING_DP
## 1              927             5456            1011               NA
## 2              689             1082             193              155
## 3              602              917             175              153
## 4              454              928             164              156
## 5              472              920             138              168
## 6              443              973             123              149
```

```
##      INDEX         TARGET_WINS     TEAM_BATTING_H TEAM_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  
```

```
##
##  TEAM_BATTING_3B  TEAM_BATTING_HR  TEAM_BATTING_BB TEAM_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
##  TEAM_BASERUN_SB TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_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
##  TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_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
##  TEAM_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
```

Exploring the incomplete data

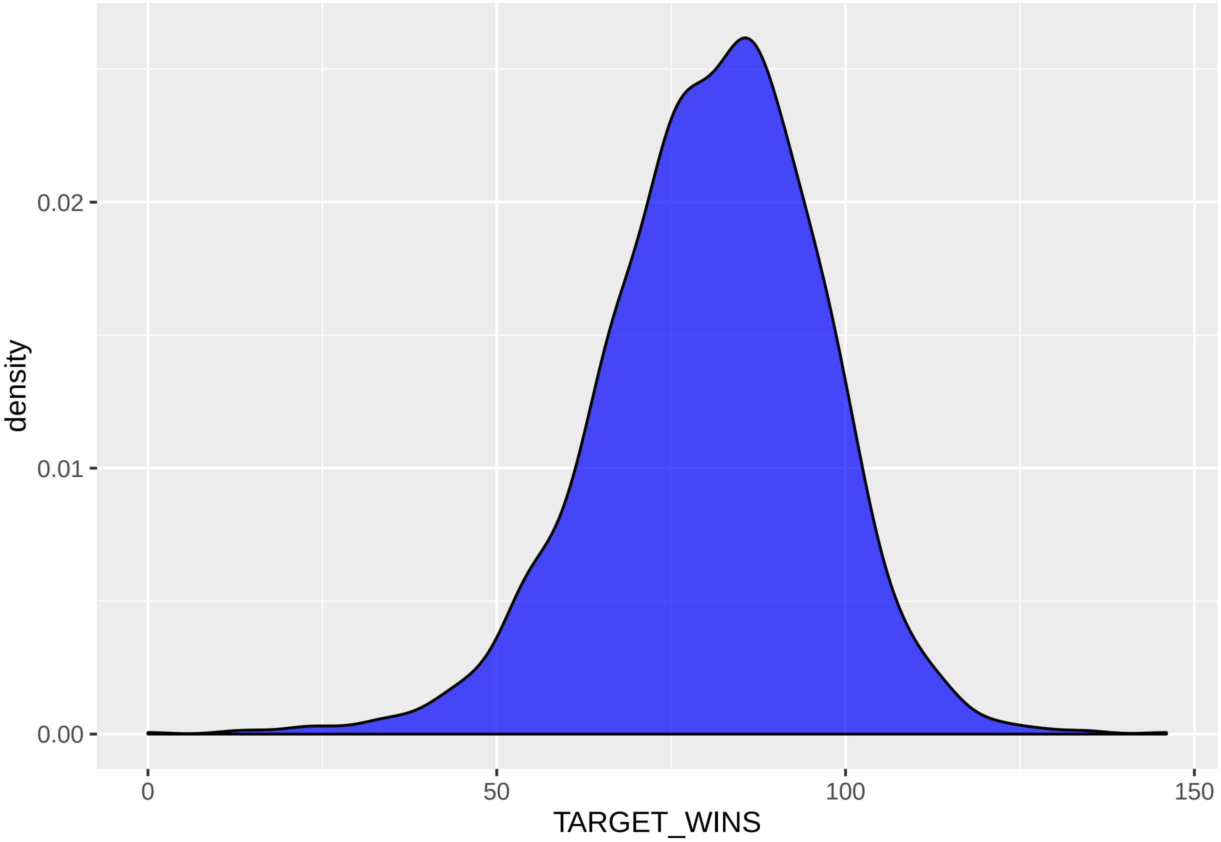
Rows with NULL or NA values are 8.3% of the data

```
nrows <- nrow(tds)
ncomplete <- sum(complete.cases(tds))
ncomplete/nrows*100
```

```
## [1] 8.391916
```

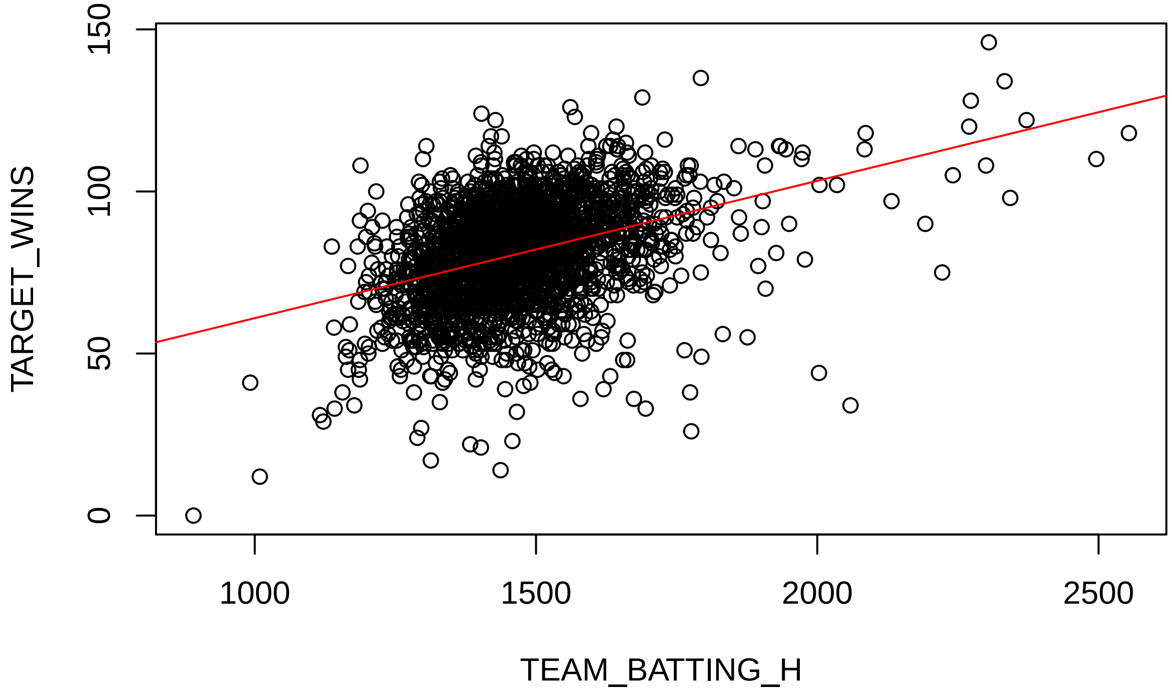
We can see in summary the values of the variables:

1 Dependent Variable : TARGET__WINS values 0 - 146 - Normal Distribution:



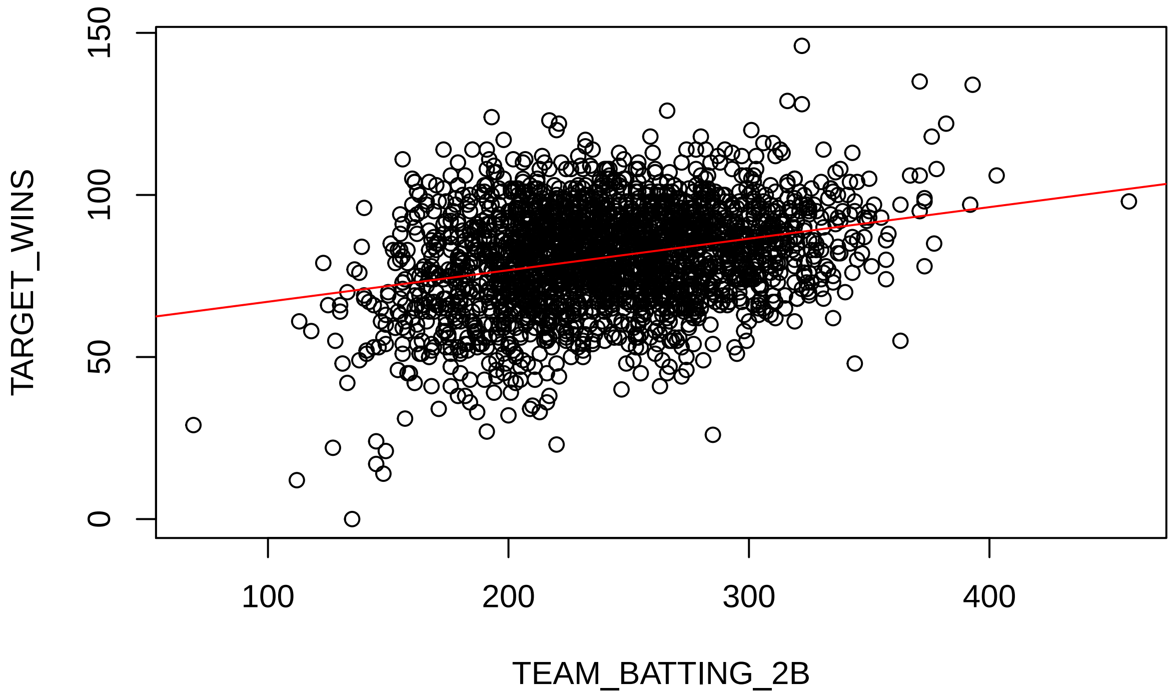
Batting Positive Impact variables:

2 TEAM_BATTING_H - Base Hits by batters Values between 891 - 2554



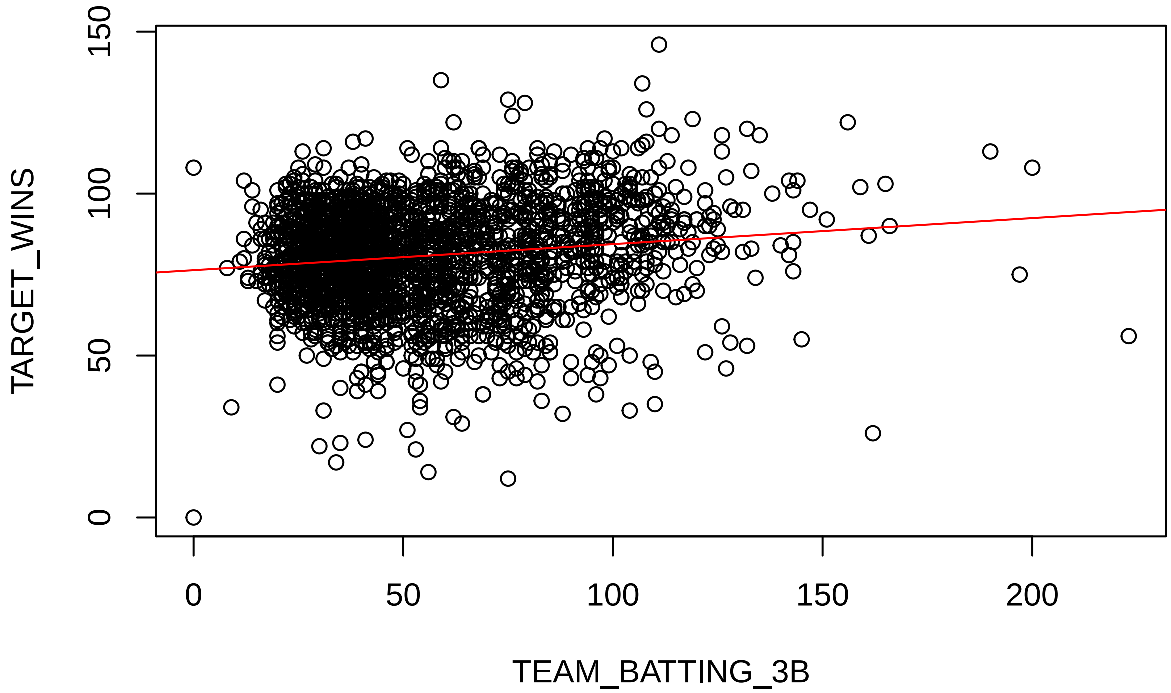
```
##    (Intercept) TEAM_BATTING_H
##    18.56232603     0.04235338
```

3 TEAM_BATTING_2B - Doubles by batters (2B) Values : 69 - 458



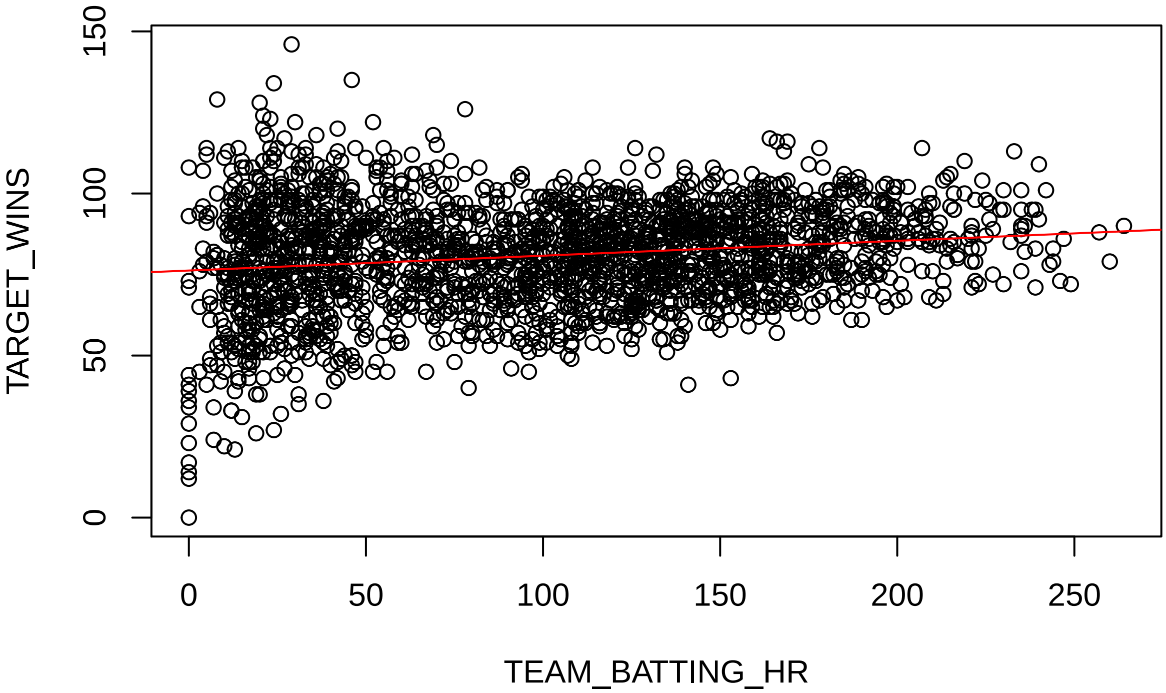
```
##     (Intercept) TEAM_BATTING_2B
##     57.31636499      0.09730485
```

4 TEAM_BATTING_3B - Triples by batters Values 0 - 223



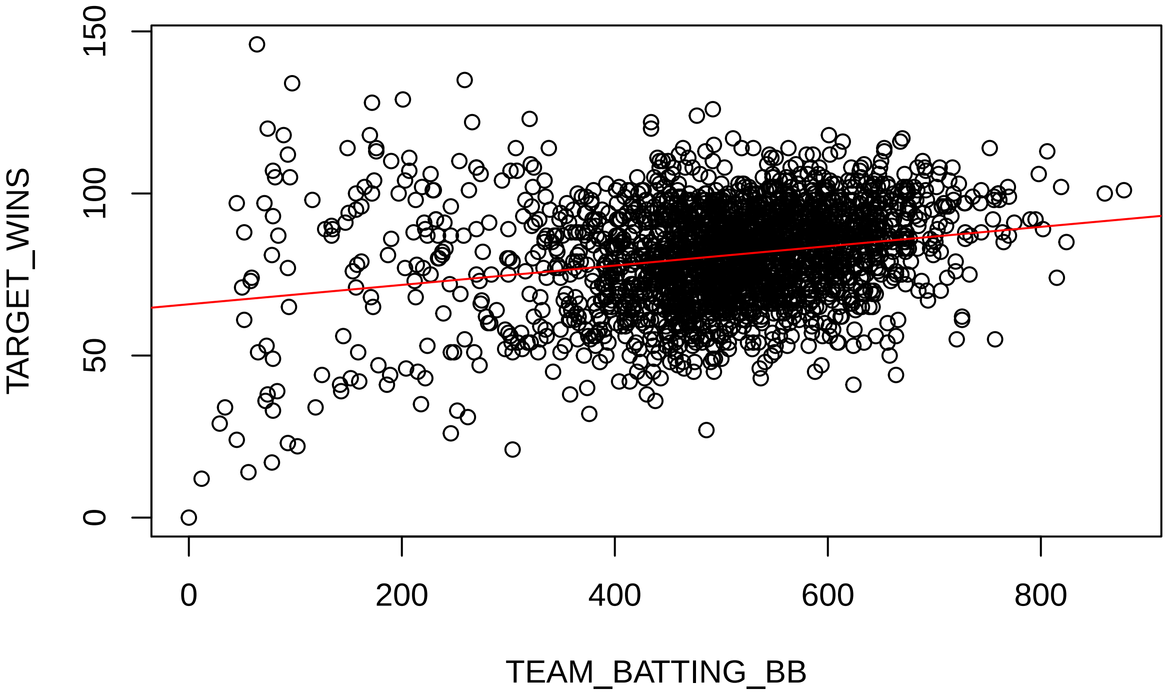
```
##     (Intercept) TEAM_BATTING_3B
##     76.34850517      0.08040463
```

5 TEAM_BATTING_HR - Homeruns by batters values 0 - 264



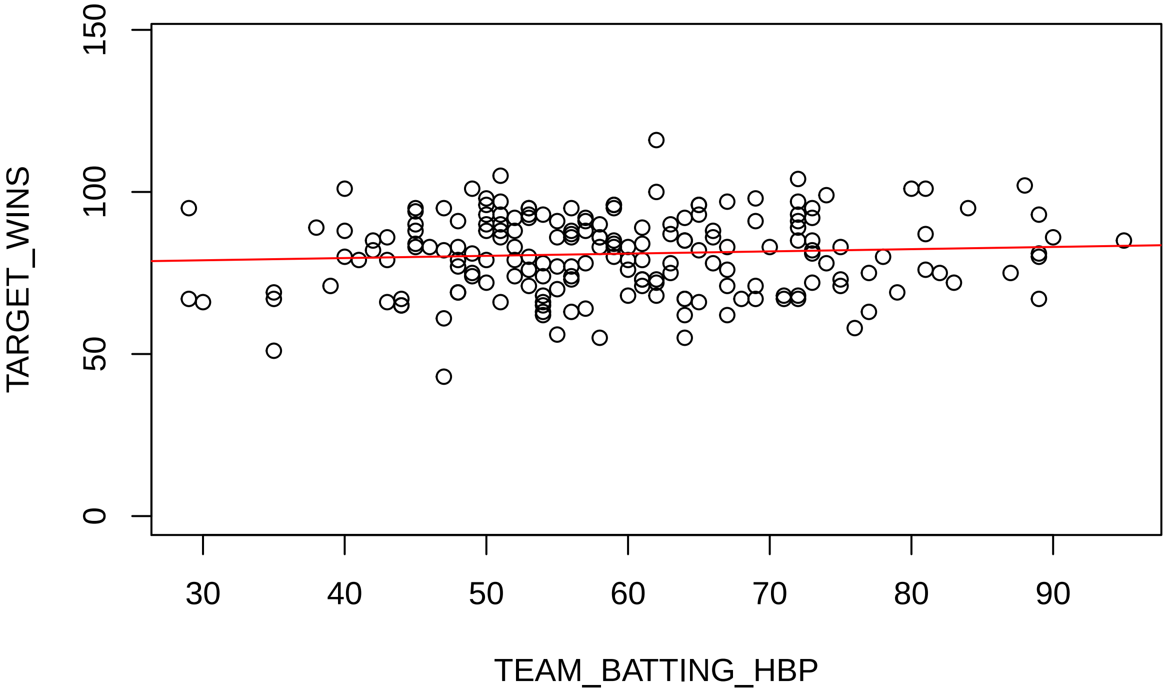
```
##     (Intercept) TEAM_BATTING_HR
##     76.22575835      0.04582883
```

6 TEAM_BATTING_BB - Walks by batters values 0 - 878



```
##     (Intercept) TEAM_BATTING_BB
##     65.81281462      0.02986299
```

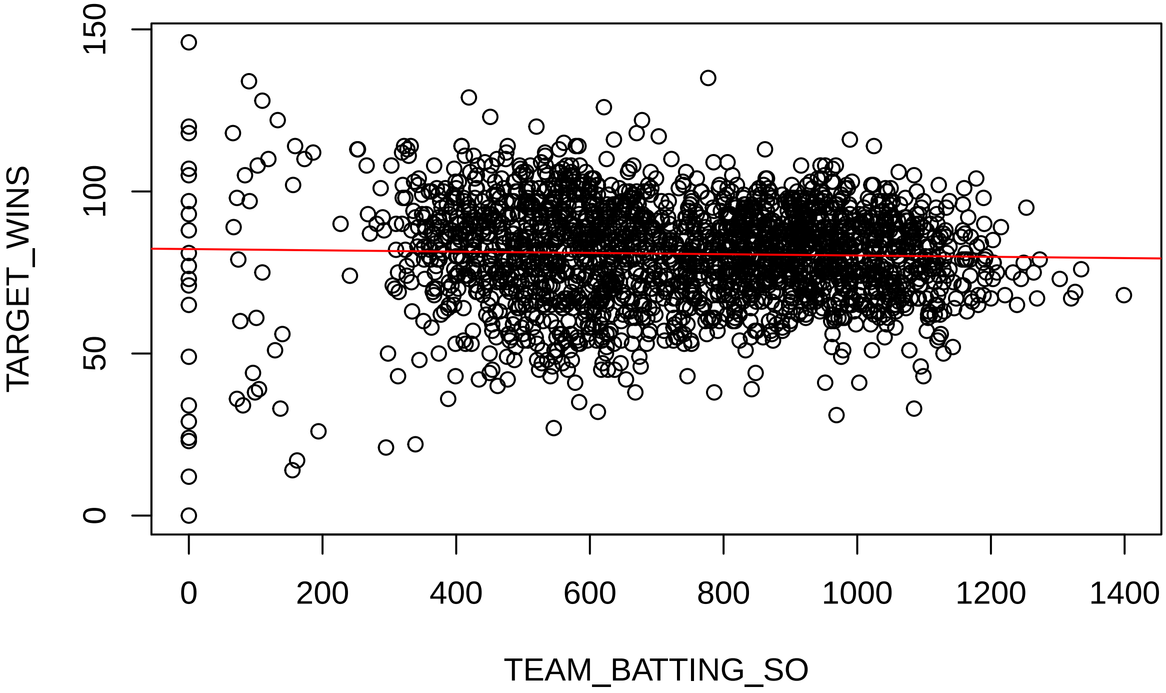
7 TEAM_BATTING_HBP - Batters hit by pitch (get a free base) values 29 - 95 NAs 2085



```
##      (Intercept) TEAM_BATTING_HBP
##      76.85048299       0.06867405
```

Batting Negative Impact variable:

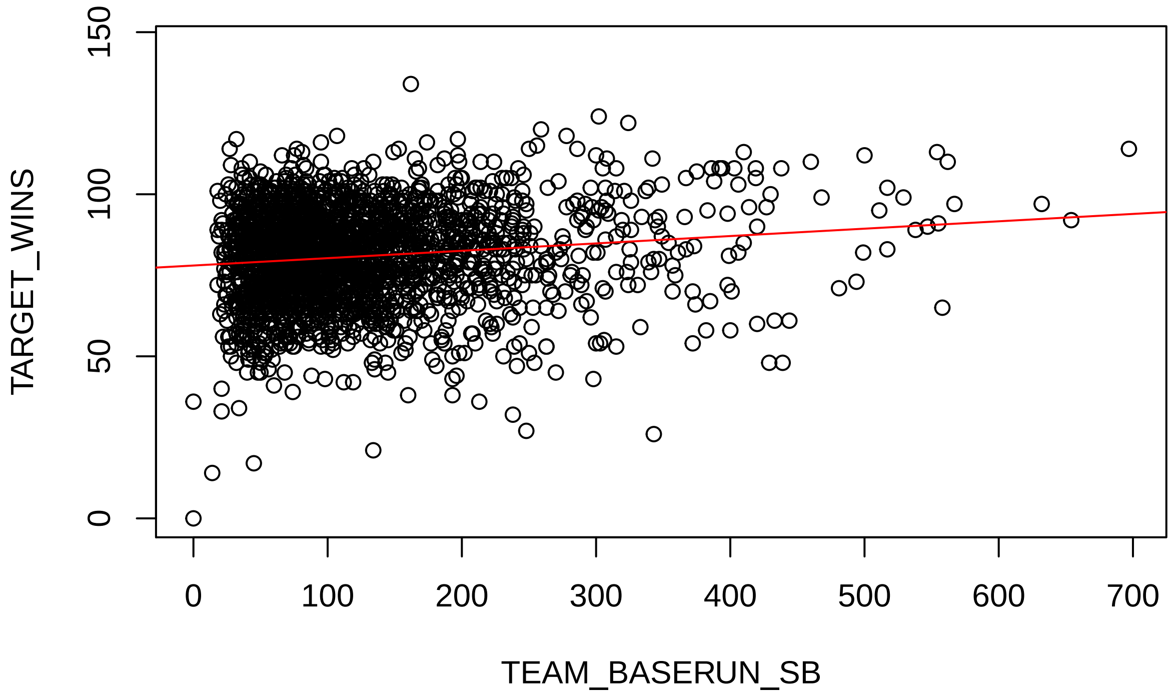
8 TEAM_BATTING_SO - Strikeouts by batters values 0 - 1399 NAs 102



```
##      (Intercept) TEAM_BATTING_SO
##     82.228036497    -0.001989582
```

Baserun Positive Impact variables

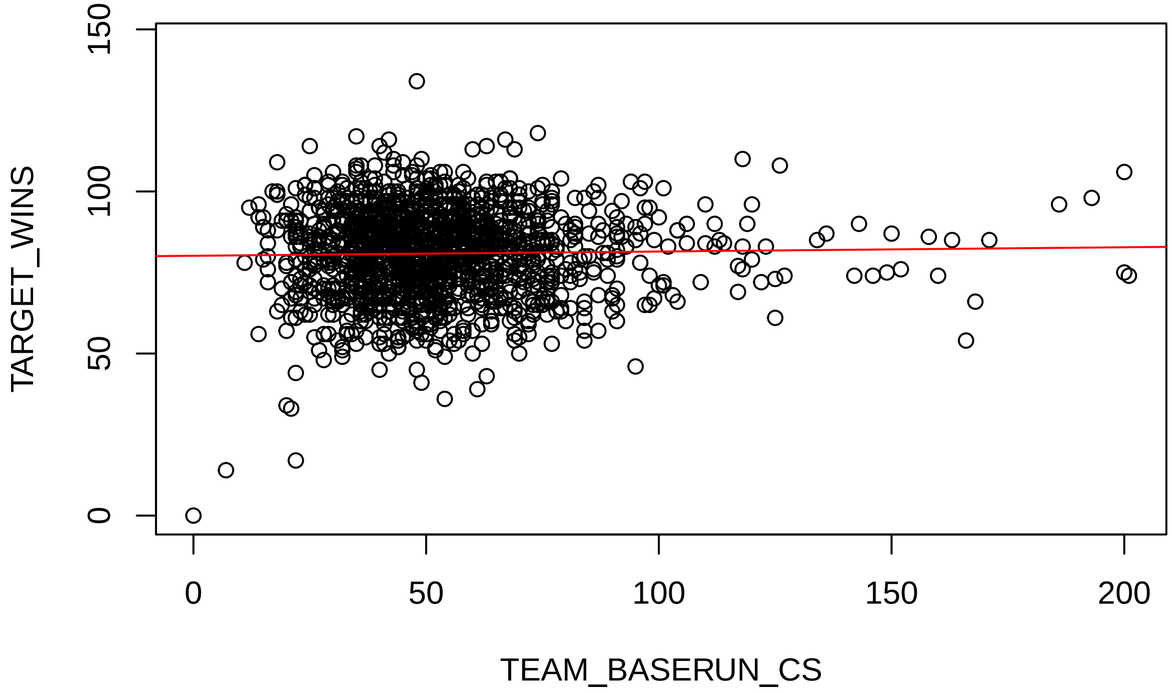
9 TEAM_BASERUN_SB - Stolen basesvalues 0 - 697 NAs 131



```
##      (Intercept) TEAM_BASERUN_SB
##      78.00909052      0.02272863
```

Baserun Negative Impact variables:

10 TEAM_BASERUN_CS - Caught stealing values 0 - 201 NAs 772

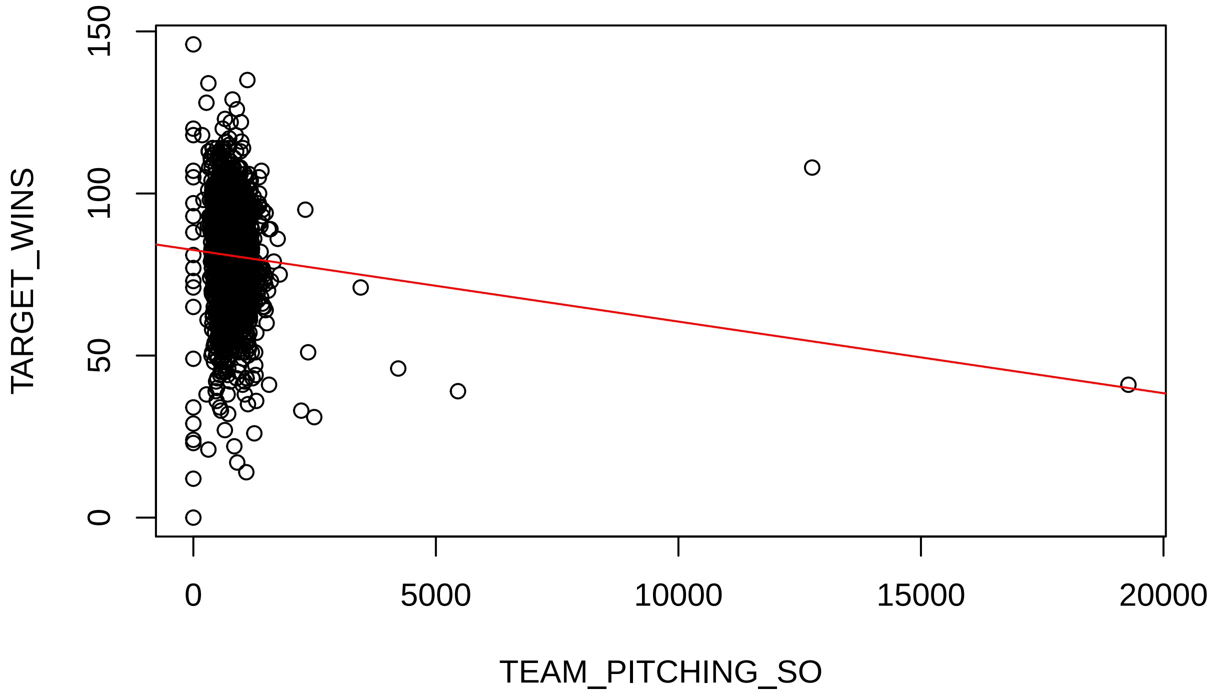


```
##     (Intercept) TEAM_BASERUN_CS
##      80.1519233       0.0131396
```

Pitching Positive Impact variables

11 TEAM_PITCHING_SO - Strikeouts by pitchers values 0 - 19278 NAs 102

the outlier value should be removed to see the actual impact

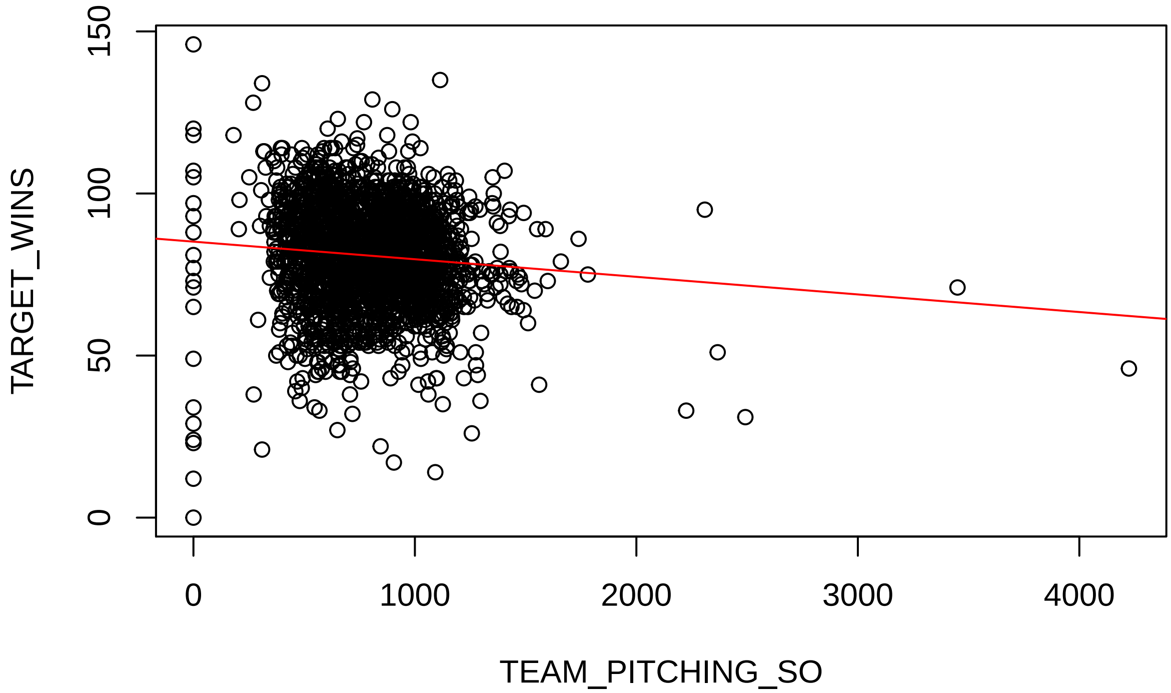


```
##      (Intercept) TEAM_PITCHING_SO
##     82.570478732     -0.002208539
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

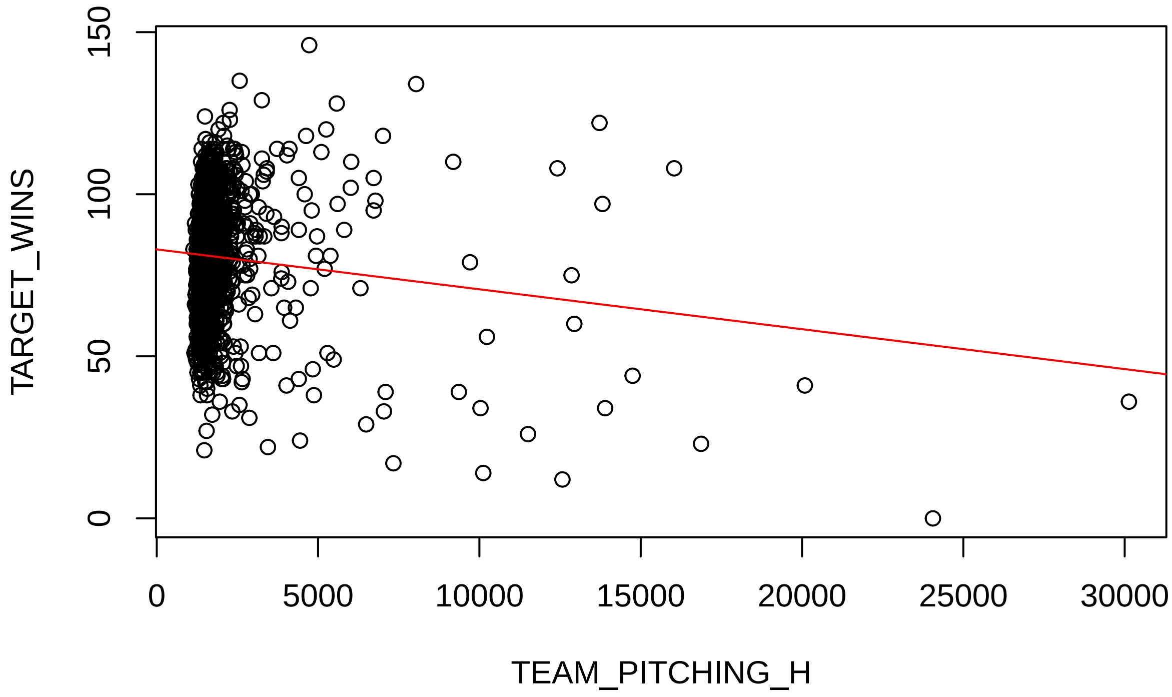
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```



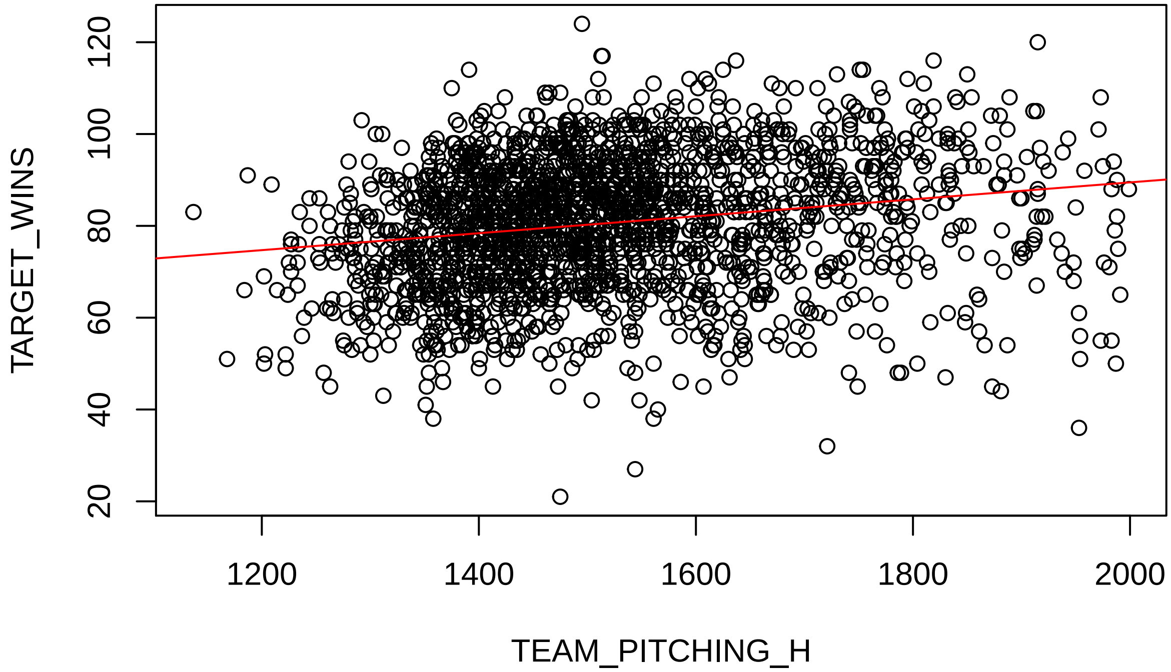
```
##       (Intercept) TEAM_PITCHING_SO
##     85.140324833     -0.005427739
```

Pitching Negative Impact Variables

12 TEAM_PITCHING_H - Hits allowed values 1137 - 30132

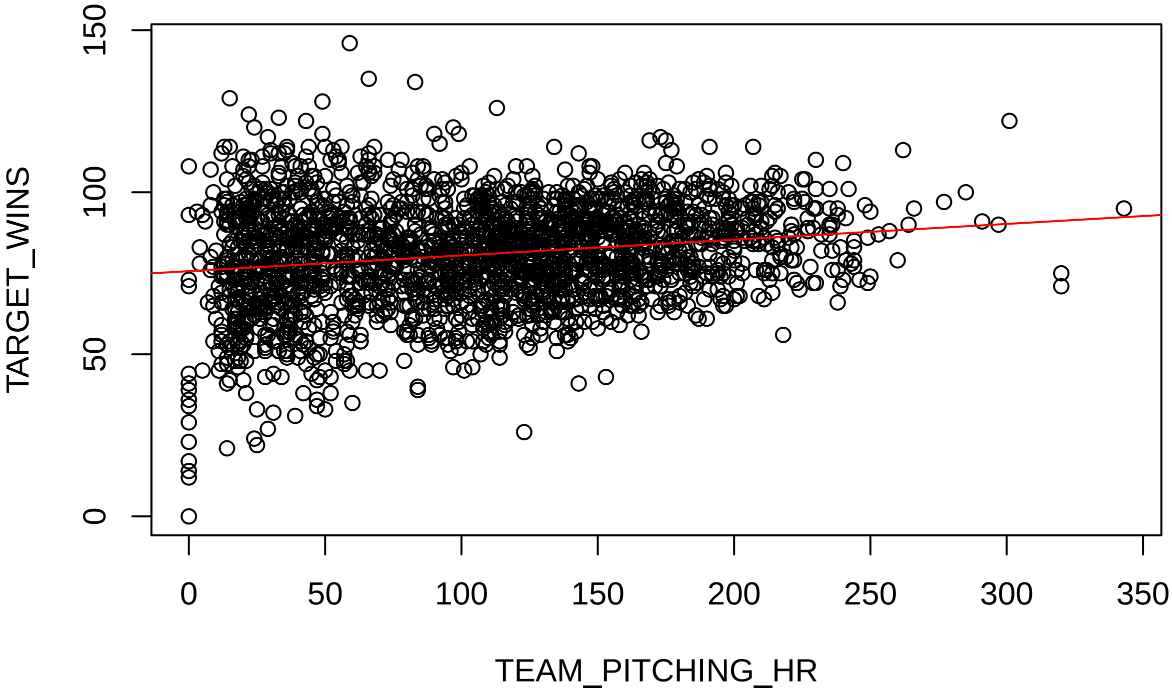


```
##     (Intercept) TEAM_PITCHING_H
##    82.980970075    -0.001230944
```



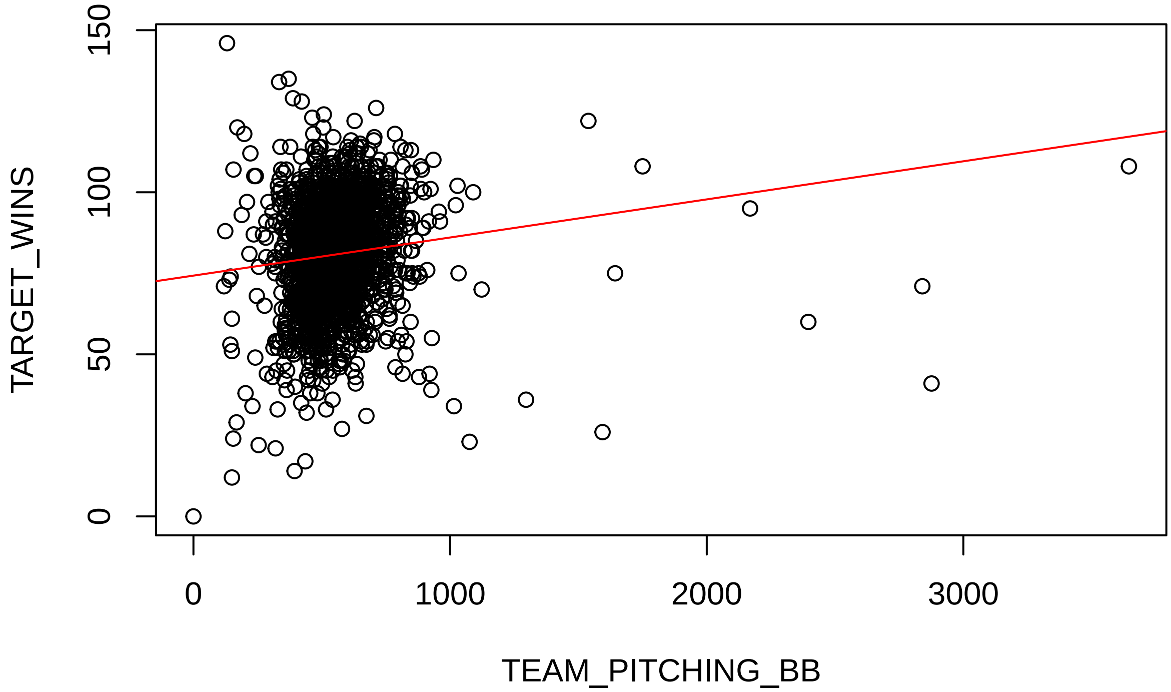
```
##      (Intercept) TEAM_PITCHING_H
##      52.57635915      0.01844559
```

13 TEAM_PITCHING_HR - Homeruns allowed values 0 - 343

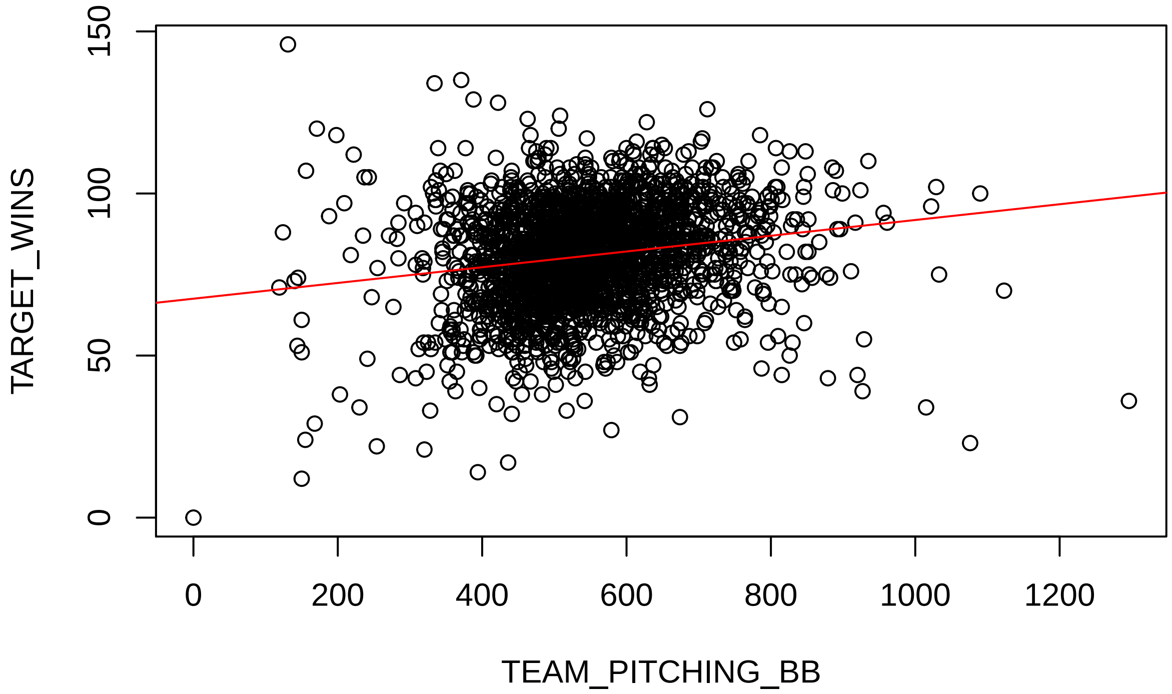


```
##      (Intercept) TEAM_PITCHING_HR
##      75.65692003       0.04857152
```

14 TEAM_PITCHING_BB - Walks allowed values 0 - 3645



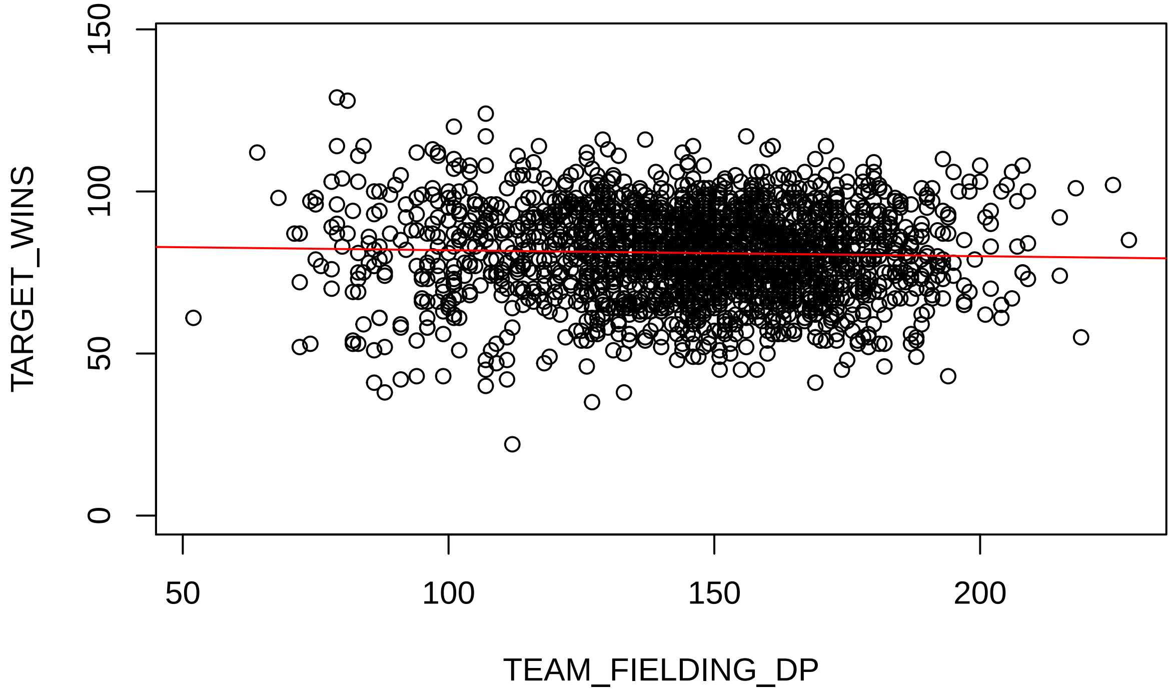
```
##       (Intercept) TEAM_PITCHING_BB
##       74.28863949       0.01175792
```



```
##      (Intercept) TEAM_PITCHING_BB
##       67.5374619        0.0242826
```

Fielding Positive Impact varibales:

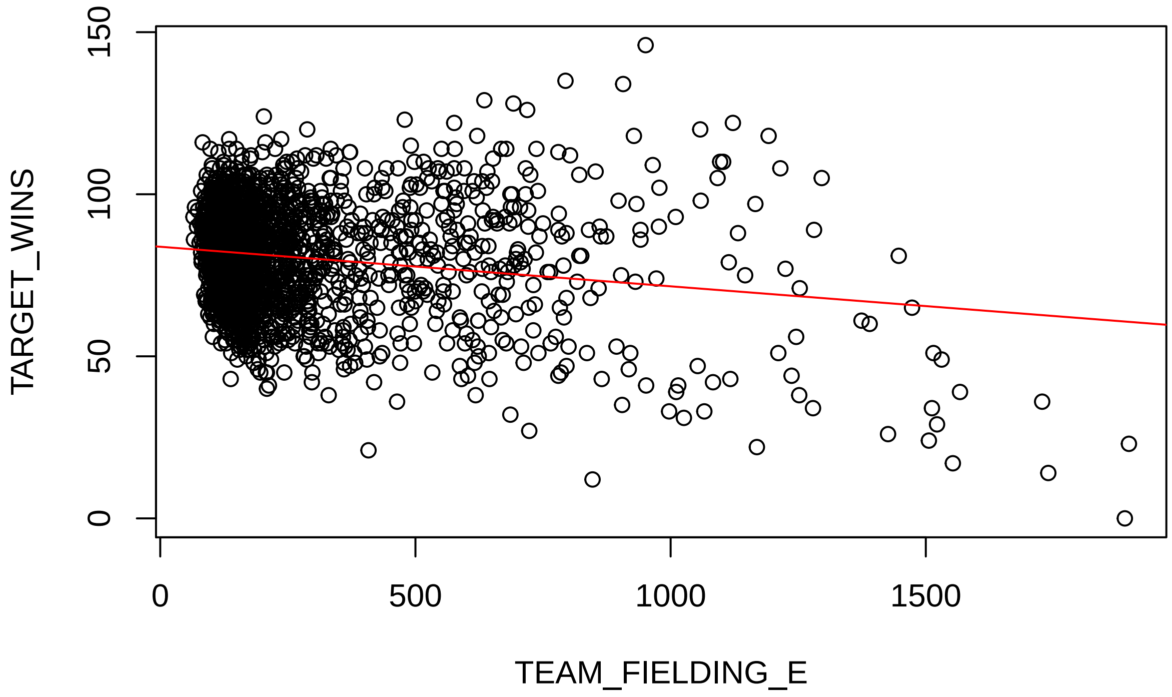
15 TEAM_FIELDING_DP - Double Plays values 52 - 228 NAs 286



```
##      (Intercept) TEAM_FIELDING_DP
##      83.71655180      -0.01852632
```

Fielding Negative Impact variables:

16 TEAM_FIELDING_E - Errors values 65 - 1898



```
##     (Intercept) TEAM_FIELDING_E
##     83.79923359     -0.01220531
```