621_HW1.Rmd

Shyam BV

February 24, 2018

Contents

---

1 Multiple Linear Regression Model : Predicting The Number Of Wins For The Baseball Team

---

Deliverables:

1. A write-up submitted in PDF format. Your write-up should have four sections. Each one is described below. You may assume you are addressing me as a fellow data scientist, so do not need to shy away from technical details.
2. Assigned predictions (the number of wins for the team) for the evaluation data set.
3. Include your R statistical programming code in an Appendix.

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(corrplot)
library(imputeR)
library(MASS)
library(pls)
library(faraway)
library(VIM)
```

1.1 Data Exploration

Describe the size and the variables in the moneyball training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas.

a. Mean / Standard Deviation / Median
b. Bar Chart or Box Plot of the data and/or Histograms
c. Is the data correlated to the target variable (or to other variables?)
d. Are any of the variables missing and need to be imputed "fixed"?

Dataset contains of 2276 observations,15 predictor variables and 1 dependent variable(`TARGET_WINS`). All the predictor variables are out input for a multipe linear regression model. Below is the summary of the dataset. That will provide all the basic summary stastistic.

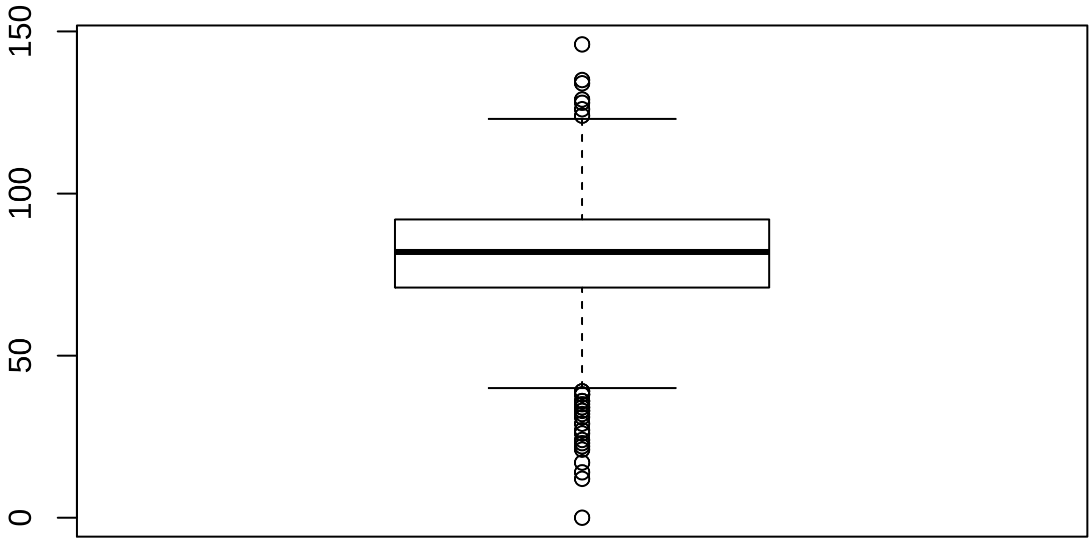
```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  2276
```

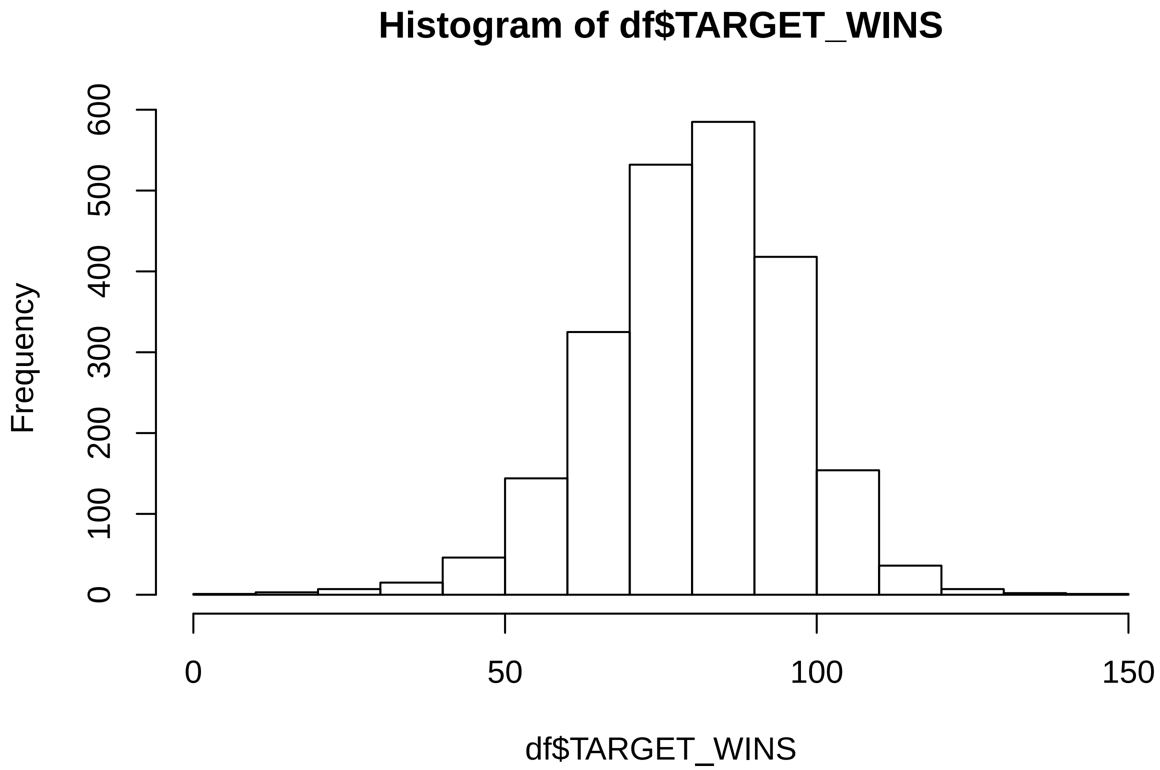
```
##      INDEX         TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B
##  Min.   :   1.0   Min.   :  0.00   Min.   : 891   Min.   : 69.0  
##  1st Qu.: 630.8   1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0  
##  Median :1270.5   Median : 82.00   Median :1454   Median :238.0  
##  Mean   :1268.5   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
```

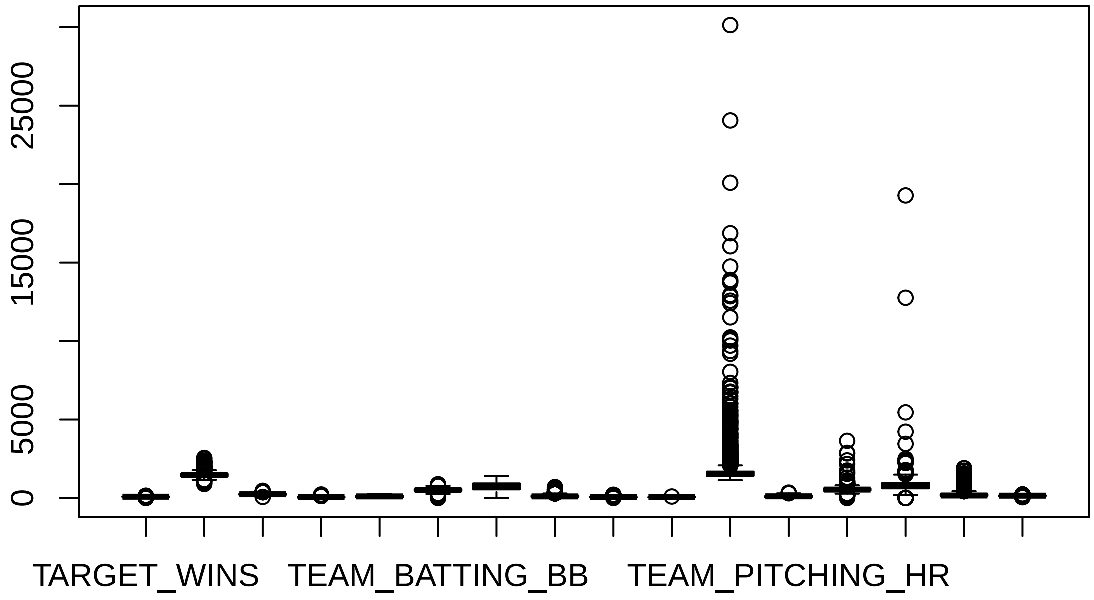
```
##   INDEX TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
## 1     1          39           1445             194              39
## 2     2          70           1339             219              22
## 3     3          86           1377             232              35
## 4     4          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
```

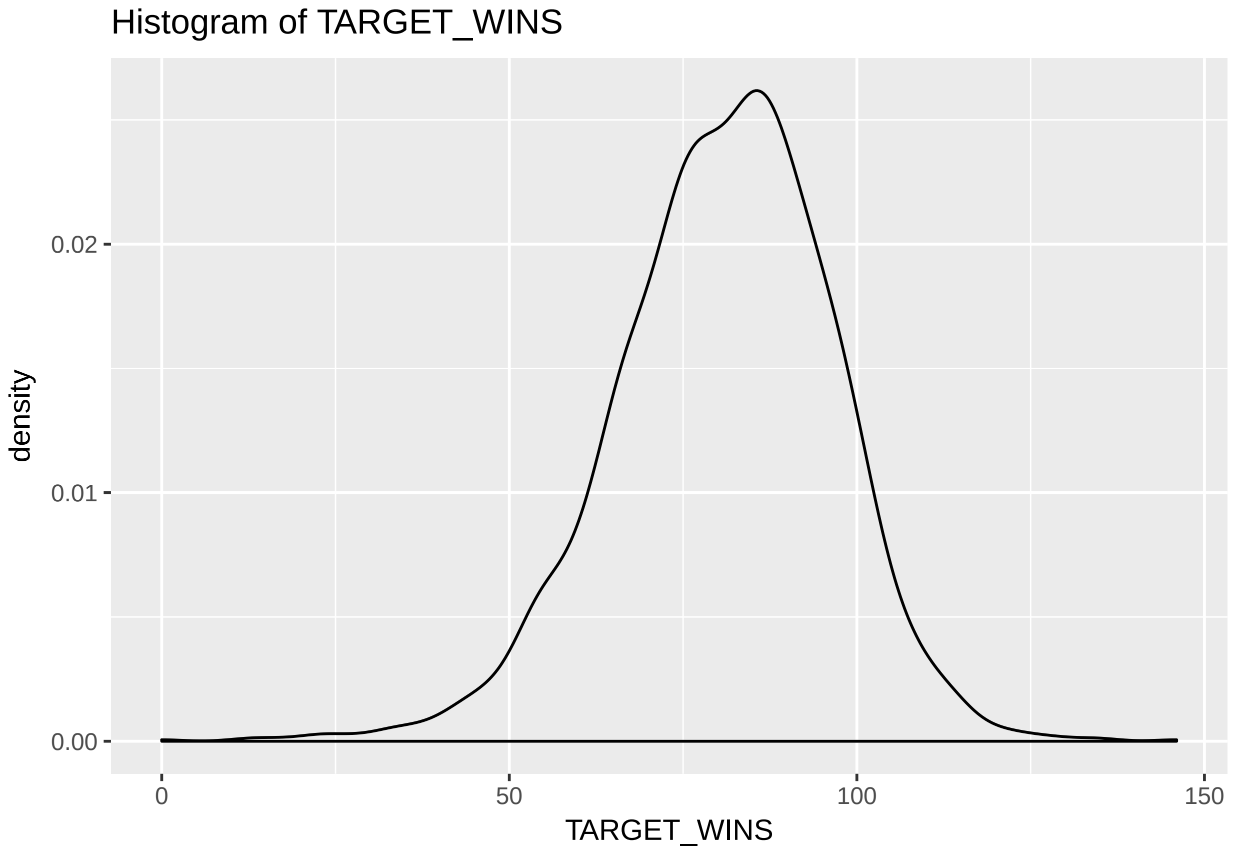
From the above summary, we can see that there are predictor variables some which has `NA` and `INDEX` column is not required as it is just an index of rows. Other predictor variable will be part of our analysis. Below is the distribution of the our dependent variable and all our variables in boxplot.

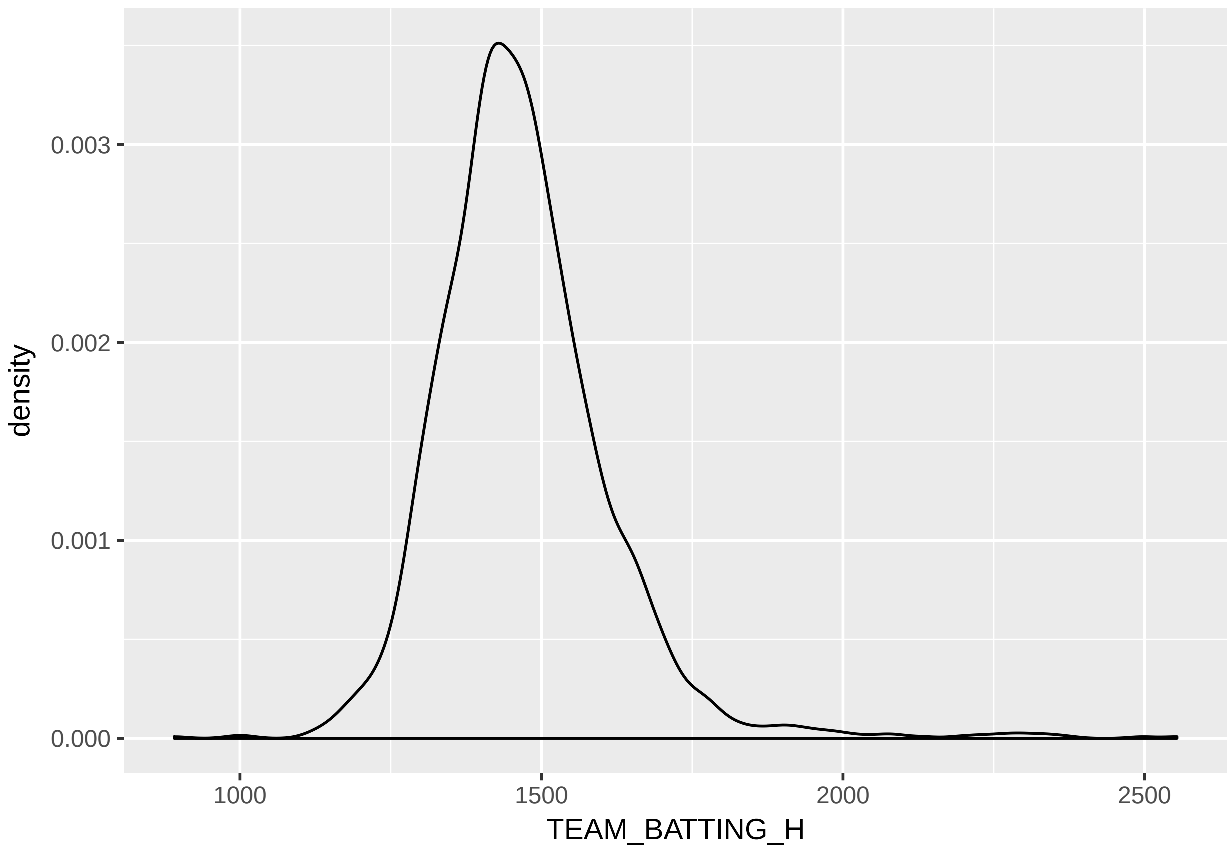

0
50
100
150

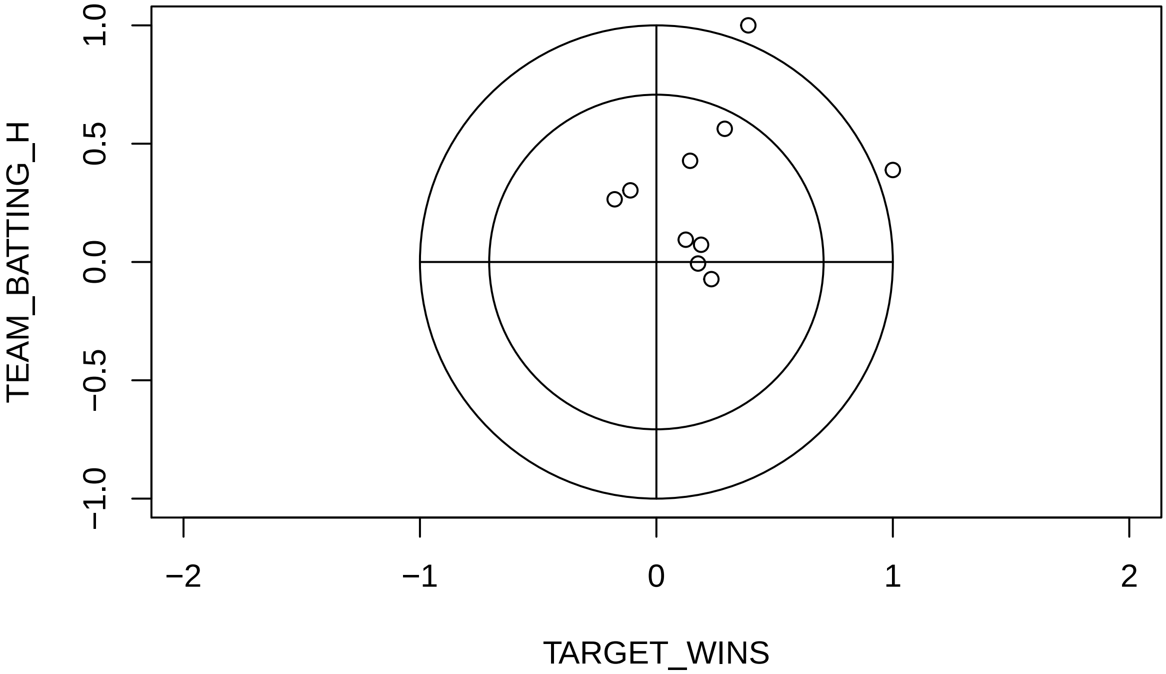

Histogram of df$TARGET_WINS
Frequency
0
100
200
300
400
500
600
0
50
100
150
df$TARGET_WINS



Above plot shows that there are some predictors which has many outliers. Transformation of predictor variables will play a major role in this analysis. Before we perform any transformation, we need to check for multi-collinerity. Below picture shows the correlation between different variables.


Histogram of TARGET_WINS
density
0.02
0.01
0.00
0
50
100
150
TARGET_WINS


0.003
0.002
0.001
0.000
density
1000
1500
2000
2500
TEAM_BATTING_H



It seems some of the predictor variables are highly correlated. These correlation will cause problems in our linear model. This needs to be taken care by various methods. Below shown is the list of exploratory variables which has missing values.

```
##                  NA_count
## TARGET_WINS             0
## TEAM_BATTING_H          0
## TEAM_BATTING_2B         0
## TEAM_BATTING_3B         0
## TEAM_BATTING_HR         0
## TEAM_BATTING_BB         0
## TEAM_BATTING_SO       102
## TEAM_BASERUN_SB       131
## TEAM_BASERUN_CS       772
## TEAM_BATTING_HBP     2085
## TEAM_PITCHING_H         0
## TEAM_PITCHING_HR        0
## TEAM_PITCHING_BB        0
## TEAM_PITCHING_SO      102
## TEAM_FIELDING_E         0
## TEAM_FIELDING_DP      286
```

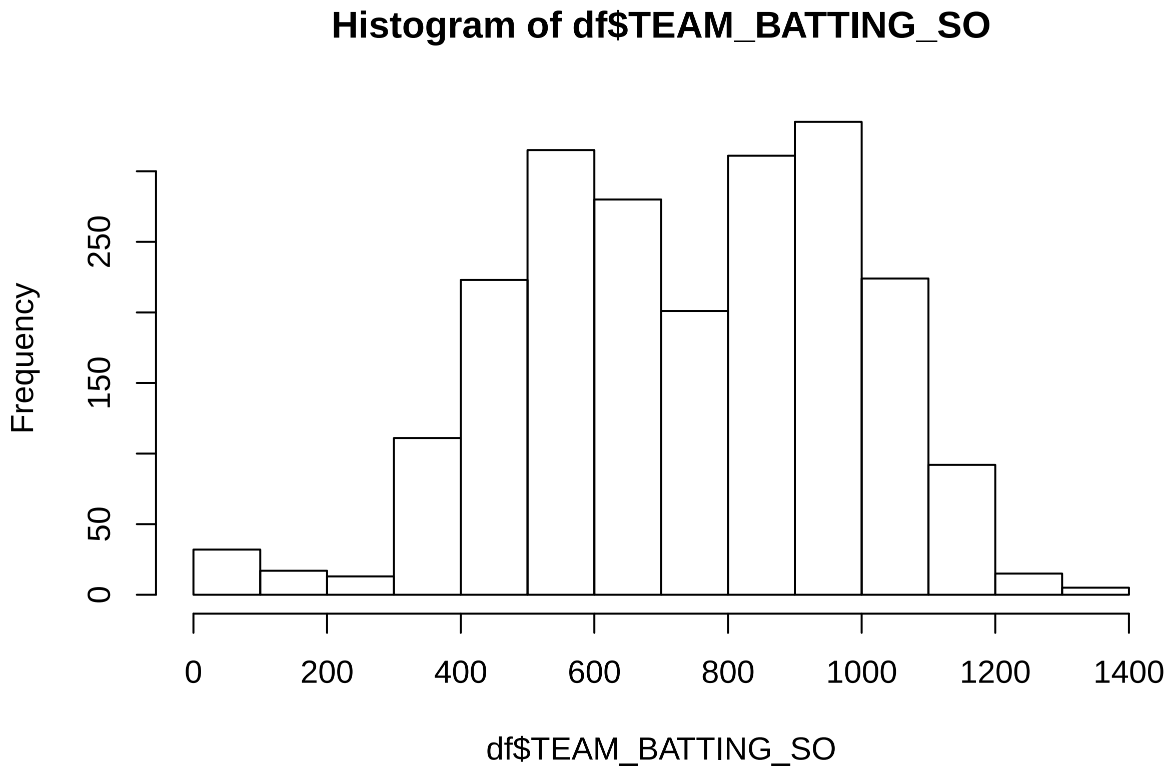
Above mentioned is the general data exploration on the moneyball dataset. Lets deep dive into data preparation part and analyze further.

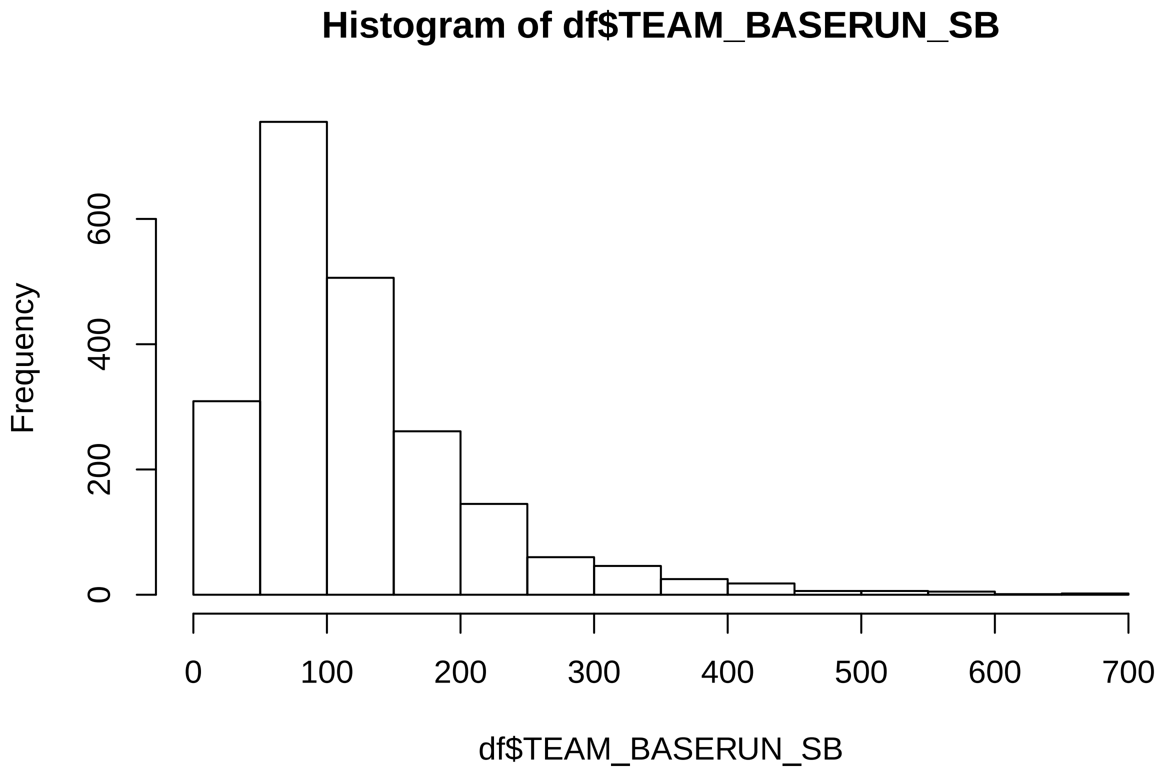
1.2 Data Preparation

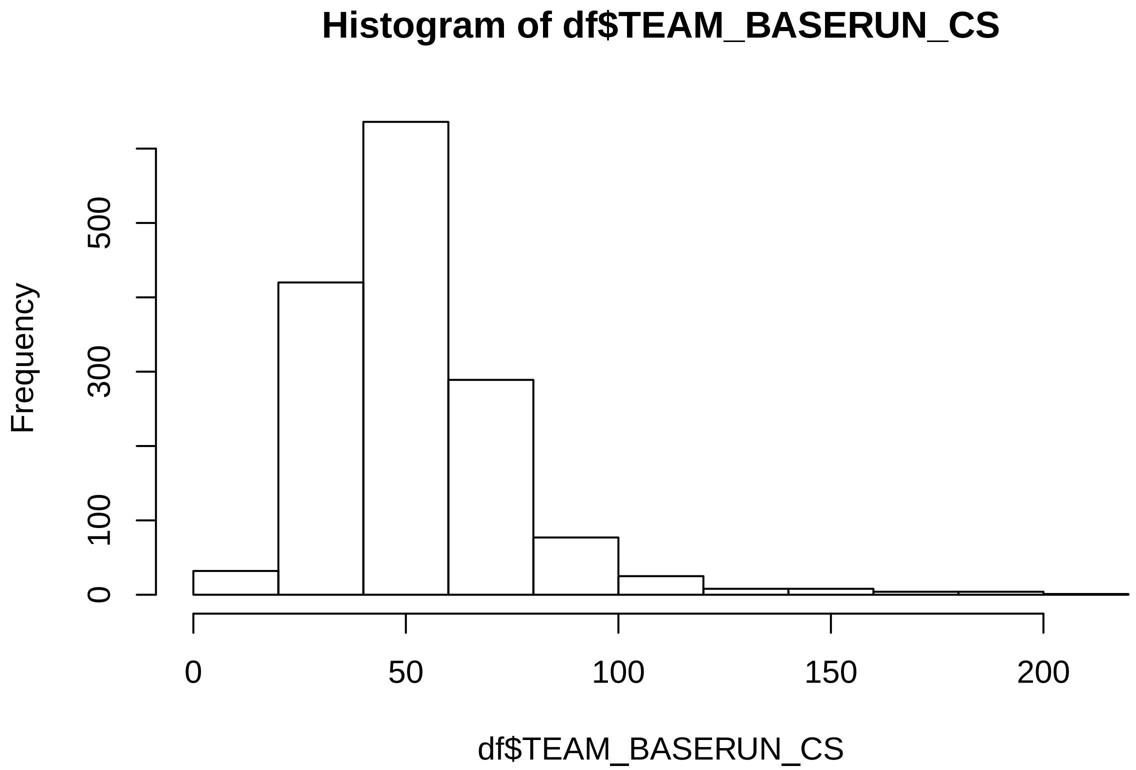
Data preparation is an important step of this analysis. As some of the variables got `NA's`, those needs to be corrected and perform some sort of transformations for the predictors which has many outliers.

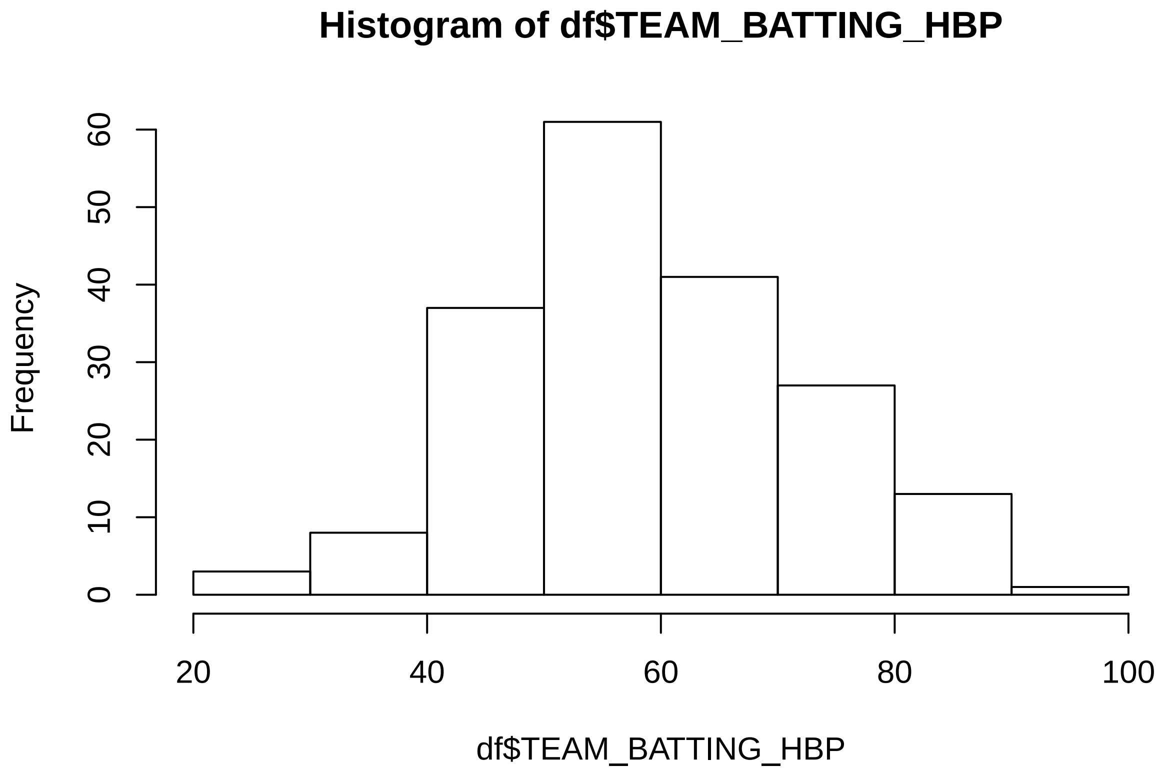
1.2.1 Fix missing values

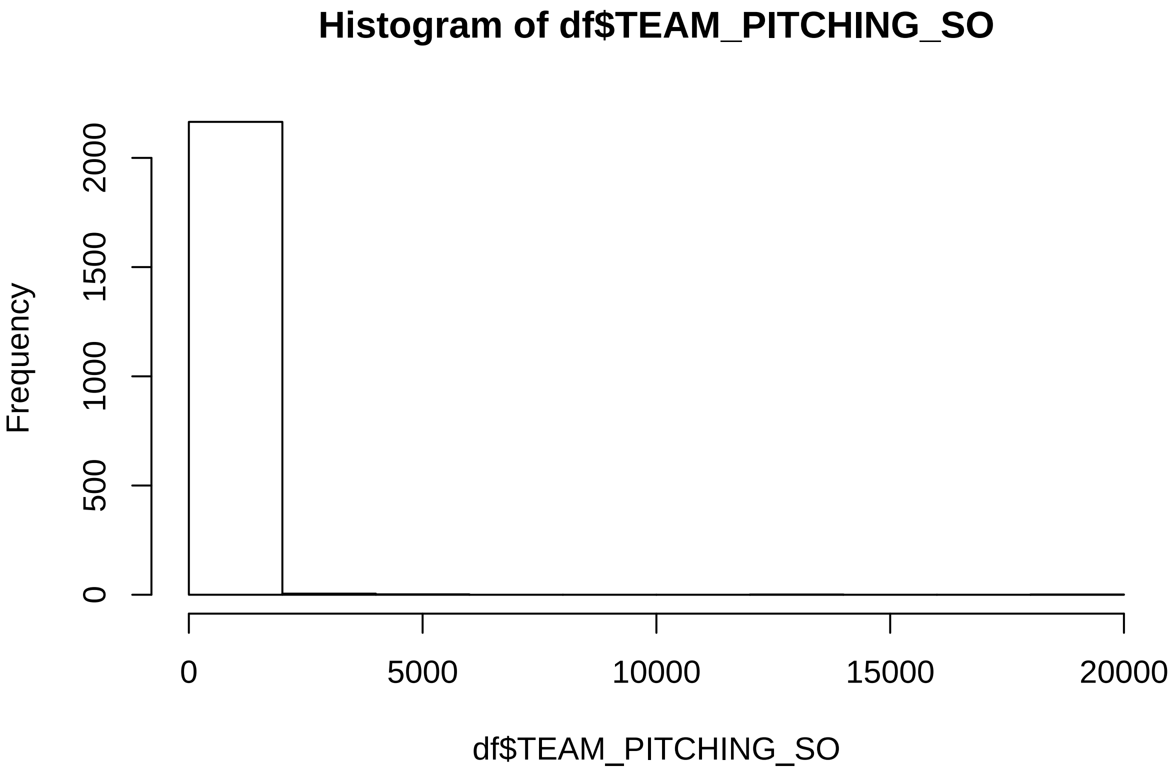
```
##   TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
##  Min.   :  0.00   Min.   : 891   Min.   : 69.0   Min.   :  0.00
##  1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0   1st Qu.: 34.00
##  Median : 82.00   Median :1454   Median :238.0   Median : 47.00
##  Mean   : 80.79   Mean   :1469   Mean   :241.2   Mean   : 55.25
##  3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0   3rd Qu.: 72.00
##  Max.   :146.00   Max.   :2554   Max.   :458.0   Max.   :223.00
##
##  TEAM_BATTING_HR  TEAM_BATTING_BB TEAM_BATTING_SO  TEAM_BASERUN_SB
##  Min.   :  0.00   Min.   :  0.0   Min.   :   0.0   Min.   :  0.0
##  1st Qu.: 42.00   1st Qu.:451.0   1st Qu.: 548.0   1st Qu.: 66.0
##  Median :102.00   Median :512.0   Median : 750.0   Median :101.0
##  Mean   : 99.61   Mean   :501.6   Mean   : 735.6   Mean   :124.8
##  3rd Qu.:147.00   3rd Qu.:580.0   3rd Qu.: 930.0   3rd Qu.:156.0
##  Max.   :264.00   Max.   :878.0   Max.   :1399.0   Max.   :697.0
##                                   NA's   :102      NA's   :131
##  TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR
##  Min.   :  0.0   Min.   :29.00    Min.   : 1137   Min.   :  0.0
##  1st Qu.: 38.0   1st Qu.:50.50    1st Qu.: 1419   1st Qu.: 50.0
##  Median : 49.0   Median :58.00    Median : 1518   Median :107.0
##  Mean   : 52.8   Mean   :59.36    Mean   : 1779   Mean   :105.7
##  3rd Qu.: 62.0   3rd Qu.:67.00    3rd Qu.: 1682   3rd Qu.:150.0
##  Max.   :201.0   Max.   :95.00    Max.   :30132   Max.   :343.0
##  NA's   :772     NA's   :2085
##  TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E  TEAM_FIELDING_DP
##  Min.   :   0.0   Min.   :    0.0   Min.   :  65.0   Min.   : 52.0
##  1st Qu.: 476.0   1st Qu.:  615.0   1st Qu.: 127.0   1st Qu.:131.0
##  Median : 536.5   Median :  813.5   Median : 159.0   Median :149.0
##  Mean   : 553.0   Mean   :  817.7   Mean   : 246.5   Mean   :146.4
##  3rd Qu.: 611.0   3rd Qu.:  968.0   3rd Qu.: 249.2   3rd Qu.:164.0
##  Max.   :3645.0   Max.   :19278.0   Max.   :1898.0   Max.   :228.0
##                   NA's   :102                        NA's   :286
```

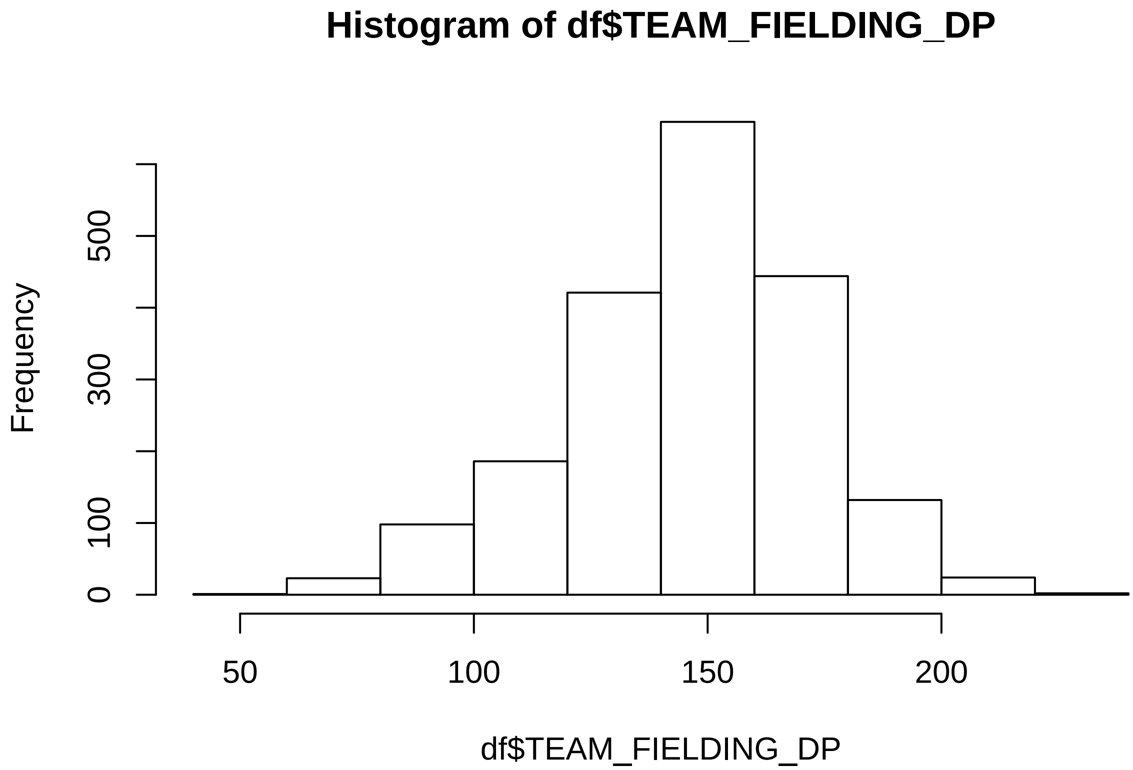

Histogram of df$TEAM_BATTING_SO
Frequency
0
50
150
250
0
200
400
600
800
1000
1200
1400
df$TEAM_BATTING_SO


Histogram of df$TEAM_BASERUN_SB
Frequency
0
200
400
600
0
100
200
300
400
500
600
700
df$TEAM_BASERUN_SB


Histogram of df$TEAM_BASERUN_CS
Frequency
0
100
300
500
0
50
100
150
200
df$TEAM_BASERUN_CS


Histogram of df$TEAM_BATTING_HBP
Frequency
0
10
20
30
40
50
60
20
40
60
80
100
df$TEAM_BATTING_HBP


Histogram of df$TEAM_PITCHING_SO
Frequency
0
500
1000
1500
2000
0
5000
10000
15000
20000
df$TEAM_PITCHING_SO



1.2.1.1 Step 1: Drop Predictors

1. As 95% of the values in `TEAM_BATTING_HBP` predictor is `NA's`, so we will remove that column.
2. `TEAM_BATTING_SB` and `TEAM_BASERUN_CS` have a strong correlation of 65.5%. `TEAM_BASERUN_CS` has around 34% of `NA's`. So we have decided to remove `TEAM_BASERUN_CS` in our analysis.

1.2.1.2 Step 2: Imputation

For other predictors which has `NA`, we have different options to perform imputation. Either we can go with `mean` or `median` or `linear model` imputation. In our case most of the predictor's have approximatly same `mean` and `median`. We have tried `lm` imputation, but it does not predict correctly due to the outliers.

So we have used `kNN` imputation for other missing values in specific predictors. It takes the similar records like it and uses the value for missing observations in TEAM_FIELDING_DP, TEAM_BATTING_SO, TEAM_BASERUN_SB, TEAM_PITCHING_SO. We will perform mean imputation for other mising values.

```
##   TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
##  Min.   :  0.00   Min.   : 891   Min.   : 69.0   Min.   :  0.00
##  1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0   1st Qu.: 34.00
##  Median : 82.00   Median :1454   Median :238.0   Median : 47.00
##  Mean   : 80.79   Mean   :1469   Mean   :241.2   Mean   : 55.25
##  3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0   3rd Qu.: 72.00
##  Max.   :146.00   Max.   :2554   Max.   :458.0   Max.   :223.00
##
##  TEAM_BATTING_HR  TEAM_BATTING_BB TEAM_BATTING_SO  TEAM_BASERUN_SB
##  Min.   :  0.00   Min.   :  0.0   Min.   :   0.0   Min.   :  0.0
##  1st Qu.: 42.00   1st Qu.:451.0   1st Qu.: 548.0   1st Qu.: 66.0  
##  Median :102.00   Median :512.0   Median : 750.0   Median :101.0  
##  Mean   : 99.61   Mean   :501.6   Mean   : 735.6   Mean   :124.8  
##  3rd Qu.:147.00   3rd Qu.:580.0   3rd Qu.: 930.0   3rd Qu.:156.0  
##  Max.   :264.00   Max.   :878.0   Max.   :1399.0   Max.   :697.0  
##                                   NA's   :102      NA's   :131    
##  TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO
##  Min.   : 1137   Min.   :  0.0    Min.   :   0.0   Min.   :    0.0  
##  1st Qu.: 1419   1st Qu.: 50.0    1st Qu.: 476.0   1st Qu.:  615.0  
##  Median : 1518   Median :107.0    Median : 536.5   Median :  813.5  
##  Mean   : 1779   Mean   :105.7    Mean   : 553.0   Mean   :  817.7  
##  3rd Qu.: 1682   3rd Qu.:150.0    3rd Qu.: 611.0   3rd Qu.:  968.0  
##  Max.   :30132   Max.   :343.0    Max.   :3645.0   Max.   :19278.0  
##                                                    NA's   :102      
##  TEAM_FIELDING_E  TEAM_FIELDING_DP
##  Min.   :  65.0   Min.   : 52.0   
##  1st Qu.: 127.0   1st Qu.:131.0   
##  Median : 159.0   Median :149.0   
##  Mean   : 246.5   Mean   :146.4   
##  3rd Qu.: 249.2   3rd Qu.:164.0   
##  Max.   :1898.0   Max.   :228.0   
##                   NA's   :286     
```

1.2.2 Step 3: Transformations and outliers

However, some fields have outlier values. Those variables can be transformed, here we will use `log` transformations on `TEAM_BATTING_H`, `TEAM_FIELDING_E`and `TEAM_PITCHING_SO`. As there are 0 values, we will add a small fraction to avoid `INF`.

As our knowledge on the dataset is limited, we will not remove the outliers. We will use `cook's` distance to remove the outliers in the each model which we build.

```
## [1] 0
```

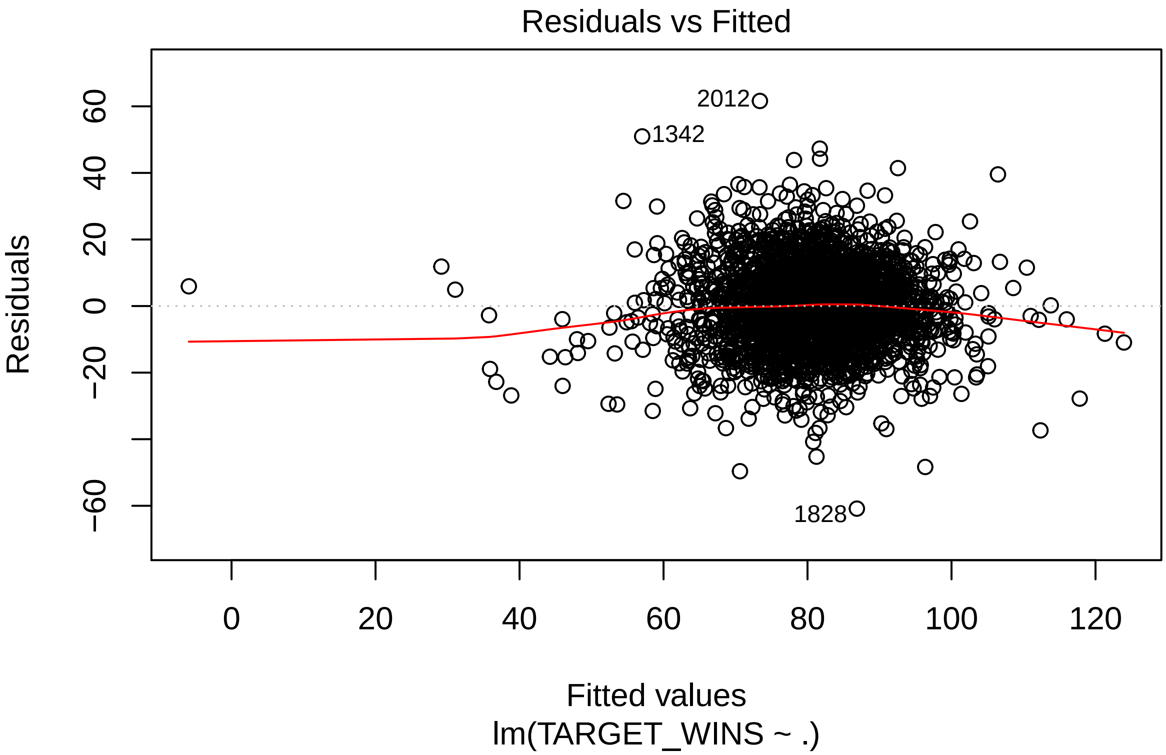
After all the transformations, we have a clean dataset which does not have any missing values.

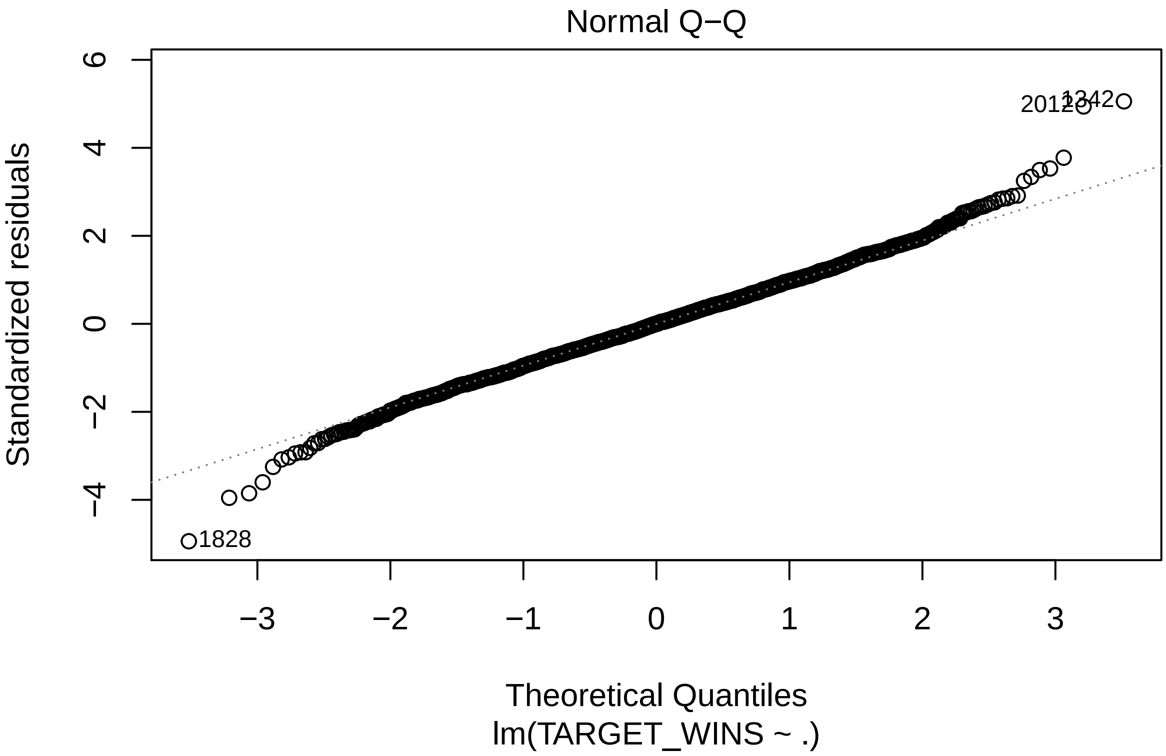
1.3 Build Models

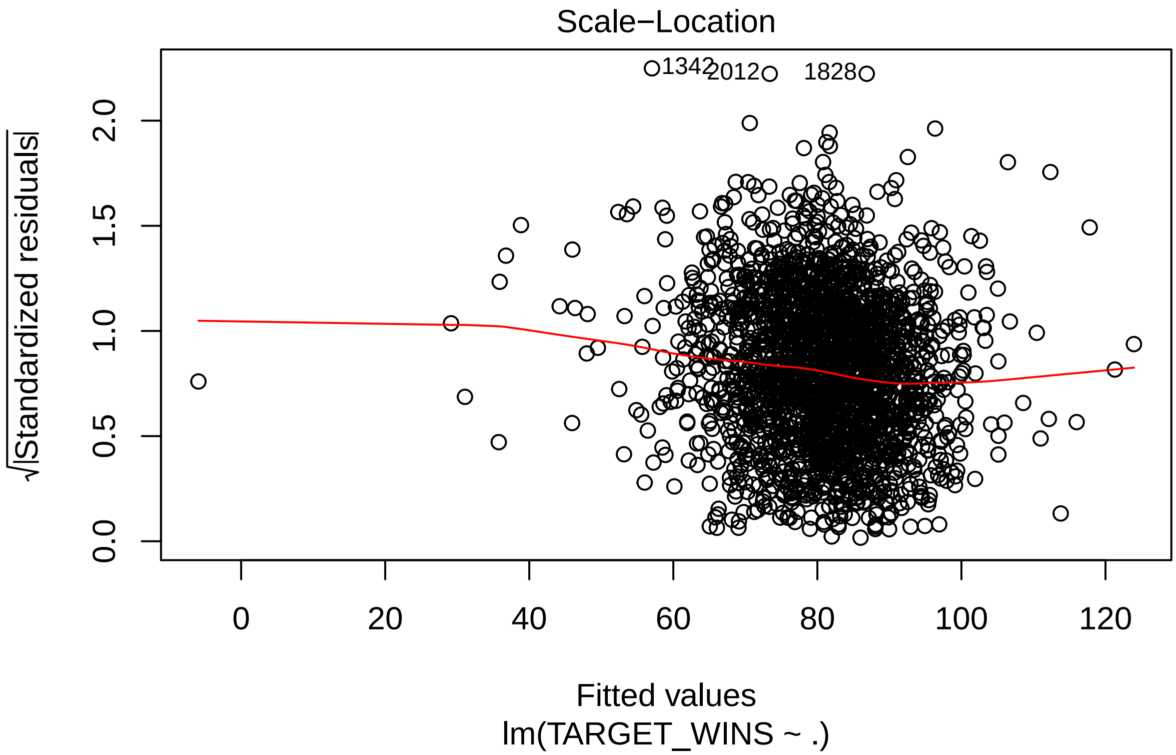
1.3.1 Model 1 - Basic backward elimination

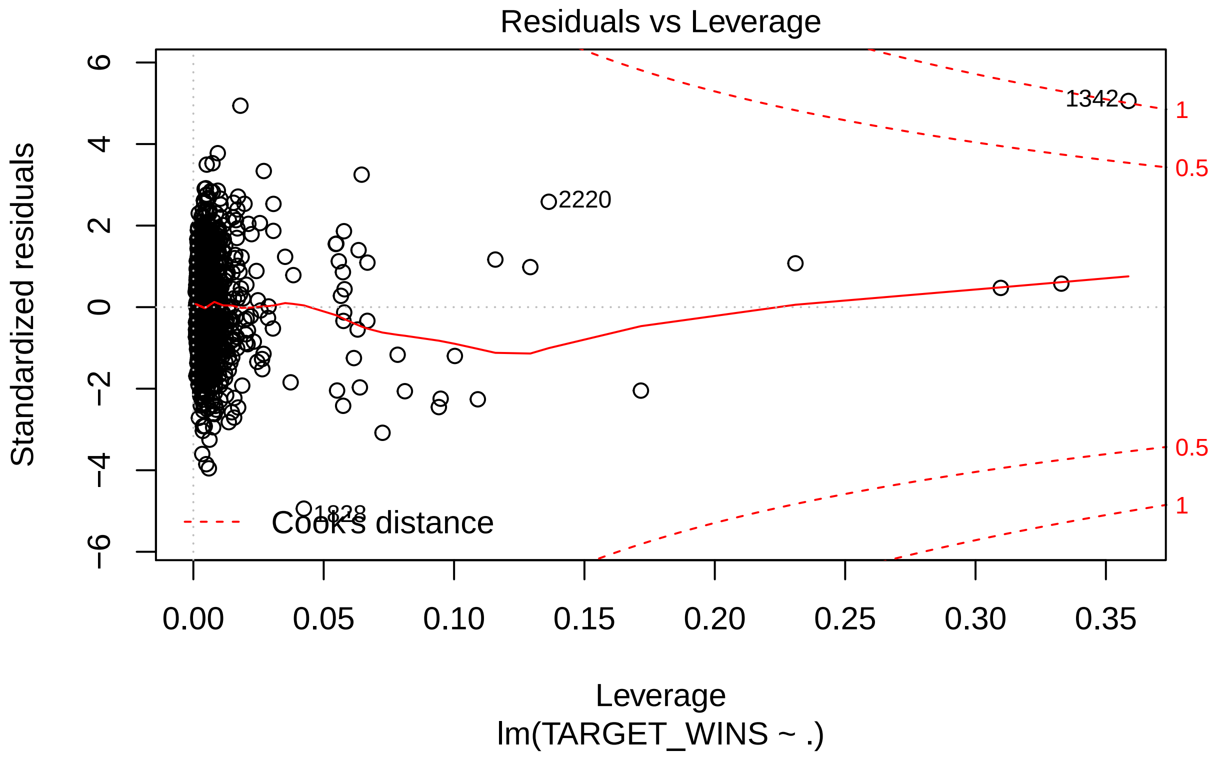
As a first model, we will build a basic model with all the predictors and perform a backward elimination.

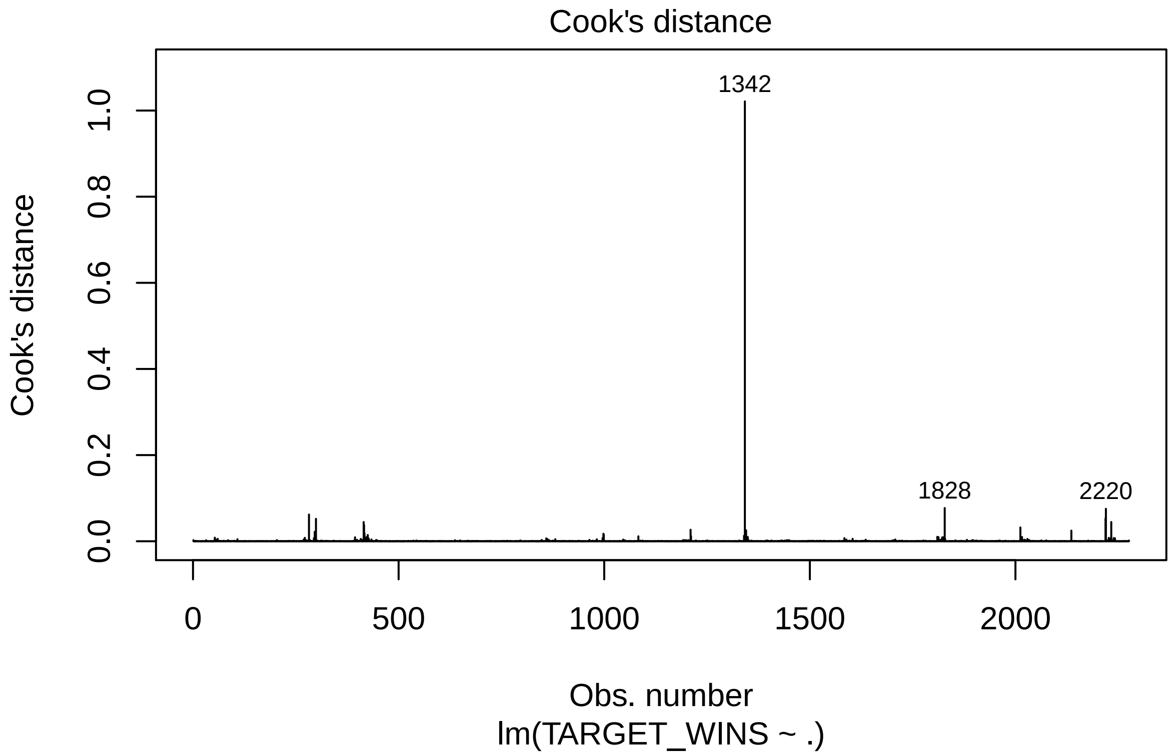
```
## 
## Call:
## lm(formula = TARGET_WINS ~ ., data = df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -60.856  -8.070   0.042   7.996  61.612 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      -3.147e+02  3.861e+01  -8.151 5.93e-16 ***
## TEAM_BATTING_H    6.835e+01  5.445e+00  12.552  < 2e-16 ***
## TEAM_BATTING_2B  -2.254e-02  8.918e-03  -2.527  0.01156 *
## TEAM_BATTING_3B   7.592e-02  1.684e-02   4.509 6.84e-06 ***
## TEAM_BATTING_HR   5.591e-02  2.562e-02   2.182  0.02918 *
## TEAM_BATTING_BB   2.527e-03  4.666e-03   0.542  0.58820
## TEAM_BATTING_SO  -1.500e-02  2.447e-03  -6.132 1.02e-09 ***
## TEAM_BASERUN_SB   4.705e-02  4.364e-03  10.781  < 2e-16 ***
## TEAM_PITCHING_H  -8.899e-04  3.341e-04  -2.664  0.00779 **
## TEAM_PITCHING_HR  7.357e-03  2.218e-02   0.332  0.74013
## TEAM_PITCHING_BB  8.630e-03  3.006e-03   2.871  0.00413 **
## TEAM_PITCHING_SO -3.408e-01  1.368e-01  -2.490  0.01283 *
## TEAM_FIELDING_E  -1.550e+01  1.010e+00 -15.343  < 2e-16 ***
## TEAM_FIELDING_DP -1.645e-01  1.307e-02 -12.587  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.59 on 2262 degrees of freedom
## Multiple R-squared:  0.3653, Adjusted R-squared:  0.3617
## F-statistic: 100.2 on 13 and 2262 DF,  p-value: < 2.2e-16
##
##   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B  TEAM_BATTING_HR
##         3.750130         2.502378         3.178209        34.561298
##  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_H
##         4.705039         5.154090         2.264236         3.173121
## TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E
##        26.548069         3.591864         1.501128         5.527676
## TEAM_FIELDING_DP
##         1.785915
```


Residuals vs Fitted
Residuals
60
40
20
0
−20
−60
2012
1342
1828
0
20
40
60
80
100
120
Fitted values
lm(TARGET_WINS ~ .)

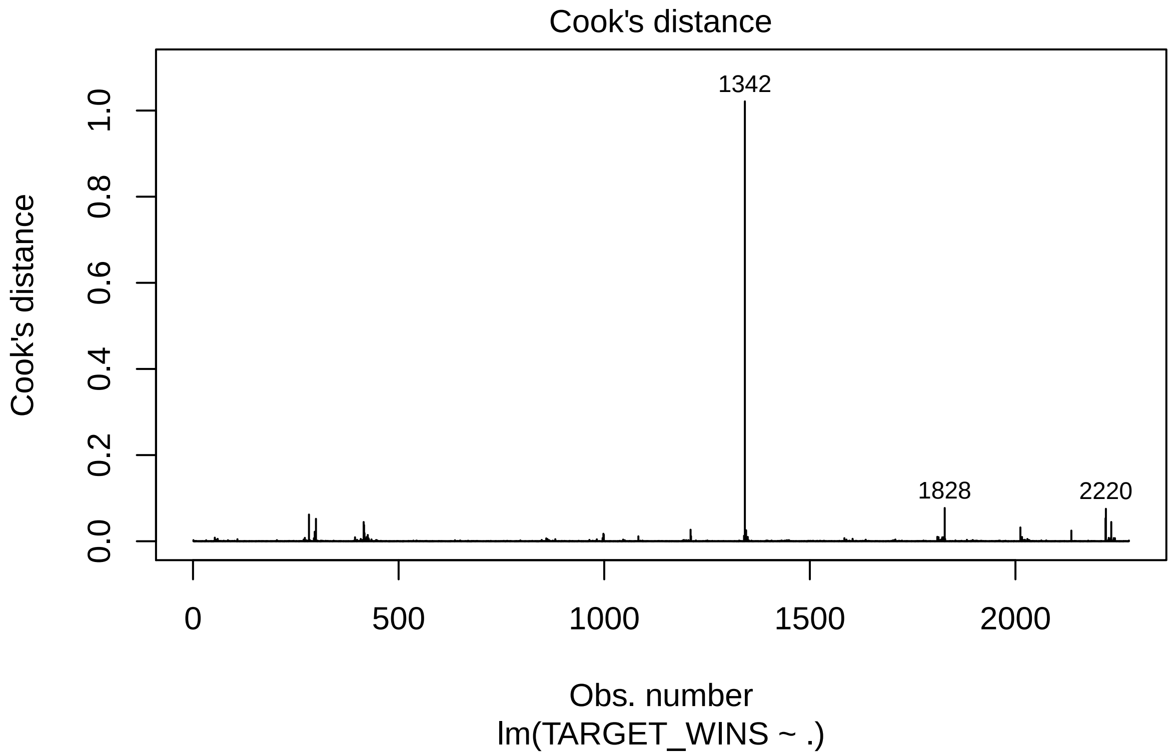

Normal Q−Q
Standardized residuals
6
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
2012
1342
1828
Theoretical Quantiles
lm(TARGET_WINS ~ .)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
1342
2012
1828
0
20
40
60
80
100
120
Fitted values
lm(TARGET_WINS ~ .)


Residuals vs Leverage
Standardized residuals
6
4
2
0
−2
−4
−6
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Leverage
lm(TARGET_WINS ~ .)
1342
2220
1828
Cook's distance
1
0.5
0.5
1



There are some outliers in the dataset, removing them using `cooks's` distance imporves the model and produces an output around `0.37`.



```
## [1] 2220 1828 1342
##      TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
## 2220          97       7.664816             363              71
## 1828          26       7.482119             285             162
## 1342         108       7.080026             338               0
##      TEAM_BATTING_HR TEAM_BATTING_BB TEAM_BATTING_SO TEAM_BASERUN_SB
## 2220              32              45              91              58
## 1828              19             246             194             343
## 1342               0             270             945             307
##      TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO
## 2220           13815              207              292         6.380123
## 1828           11508              123             1594         7.136483
## 1342           16038                0             3645         9.453914
##      TEAM_FIELDING_E TEAM_FIELDING_DP
## 2220        6.838405              161
## 1828        7.262629              122
## 1342        6.573680              107
```

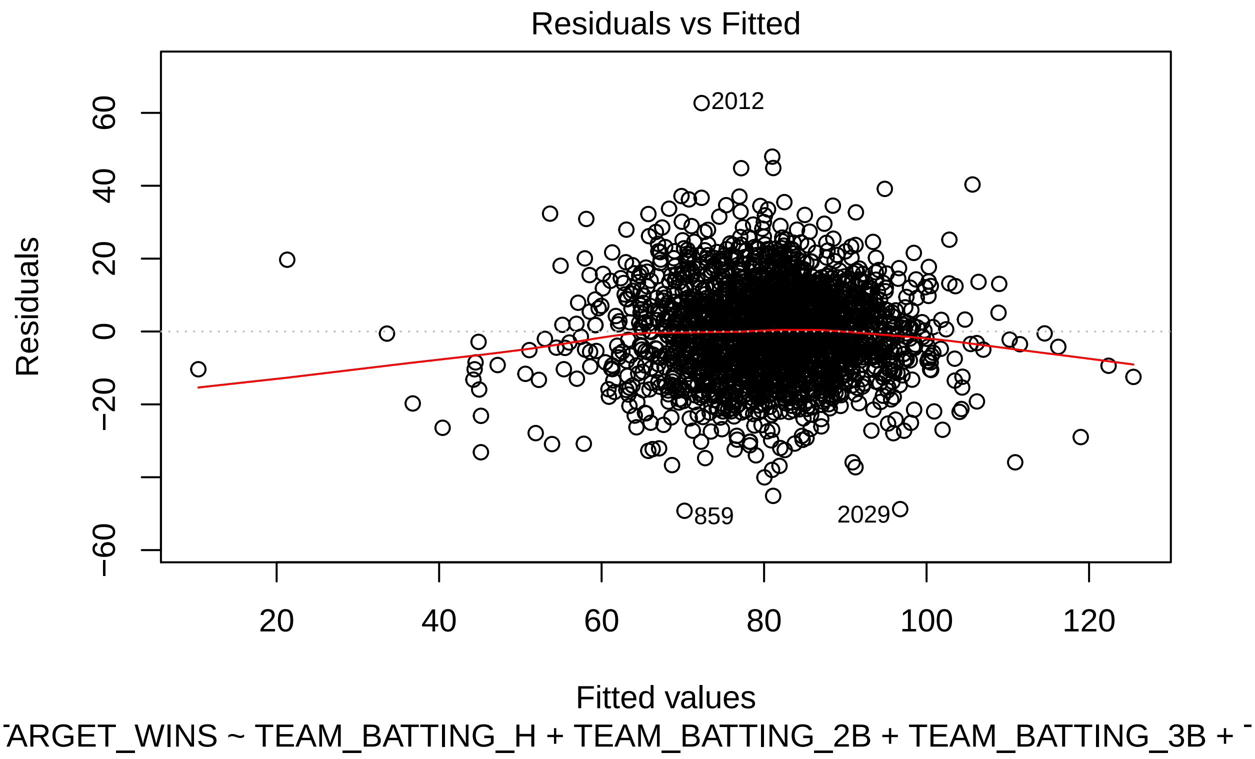
```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = df_mean_out_removed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.679  -8.065   0.002   7.995  61.927
##
## Coefficients:
```

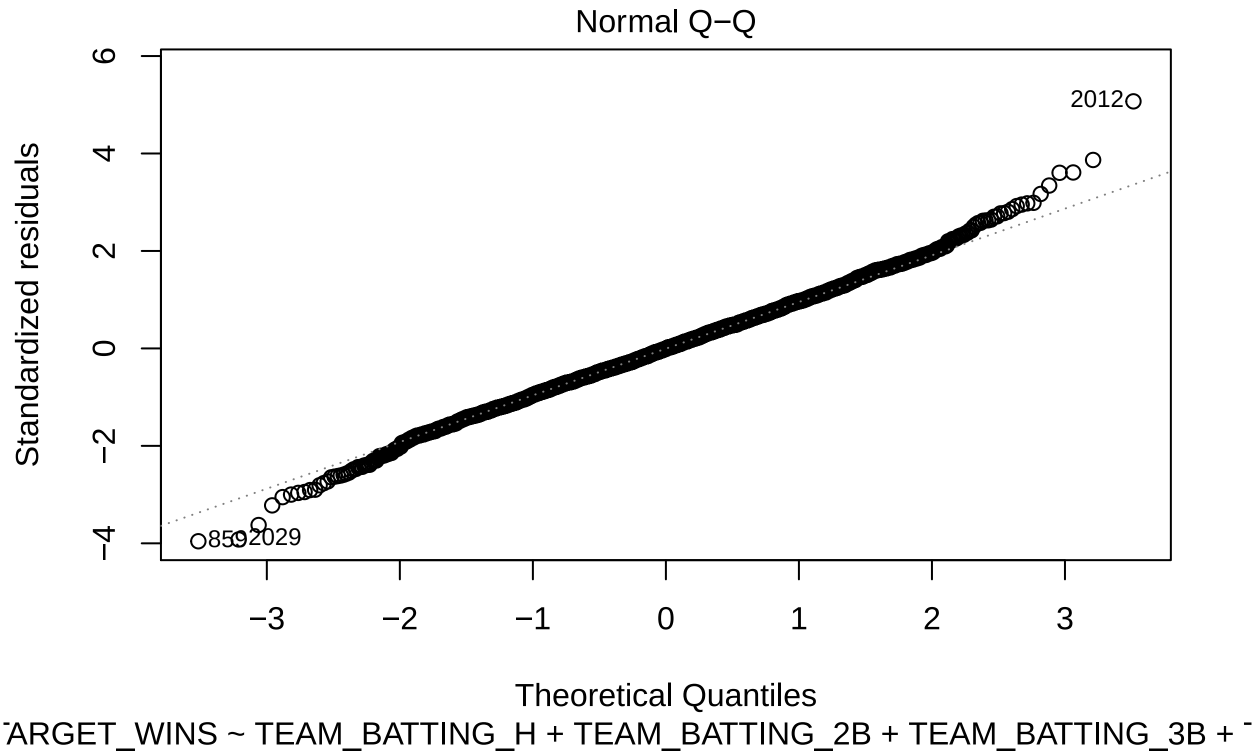
```
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      -3.123e+02  3.825e+01  -8.164 5.31e-16 ***
## TEAM_BATTING_H    6.786e+01  5.393e+00  12.583  < 2e-16 ***
## TEAM_BATTING_2B  -2.497e-02  8.847e-03  -2.822  0.00481 ** 
## TEAM_BATTING_3B   8.454e-02  1.673e-02   5.053 4.71e-07 ***
## TEAM_BATTING_HR   3.157e-02  2.774e-02   1.138  0.25528    
## TEAM_BATTING_BB   9.037e-03  5.220e-03   1.731  0.08356 .  
## TEAM_BATTING_SO  -1.513e-02  2.422e-03  -6.246 5.02e-10 ***
## TEAM_BASERUN_SB   4.756e-02  4.322e-03  11.005  < 2e-16 ***
## TEAM_PITCHING_H  -7.212e-04  3.472e-04  -2.077  0.03788 *  
## TEAM_PITCHING_HR  3.396e-02  2.441e-02   1.391  0.16434    
## TEAM_PITCHING_BB  2.656e-03  3.712e-03   0.716  0.47431    
## TEAM_PITCHING_SO -3.341e-01  1.362e-01  -2.452  0.01426 *  
## TEAM_FIELDING_E  -1.539e+01  1.001e+00 -15.367  < 2e-16 ***
## TEAM_FIELDING_DP -1.646e-01  1.294e-02 -12.724  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 12.46 on 2259 degrees of freedom
## Multiple R-squared:  0.3748, Adjusted R-squared:  0.3712 
## F-statistic: 104.2 on 13 and 2259 DF,  p-value: < 2.2e-16
##
##   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B  TEAM_BATTING_HR 
##         3.714233         2.500814         3.177952        41.274564 
##  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_H 
##         5.956207         5.126899         2.257621         3.153518 
## TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E 
##        32.761163         4.640258         1.517558         5.497411 
## TEAM_FIELDING_DP 
##         1.784629
```

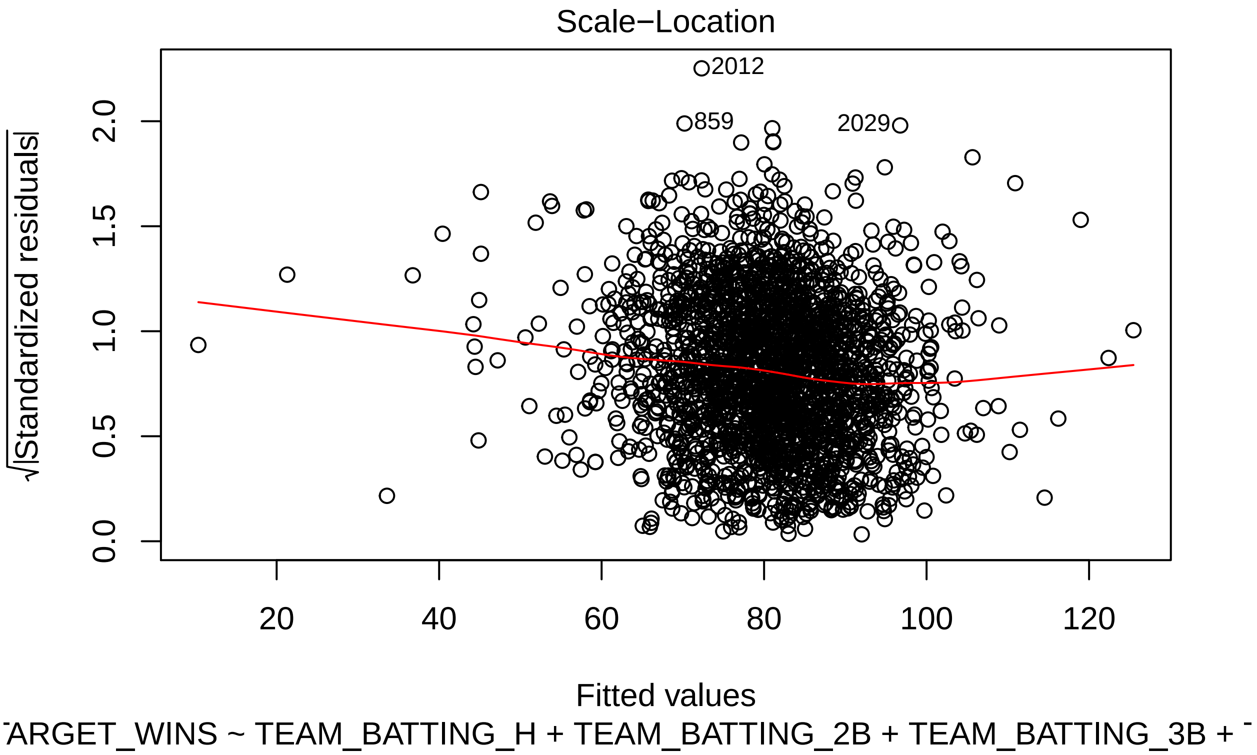
Still the VIF of some predictors as high and some predictors with high p-values.

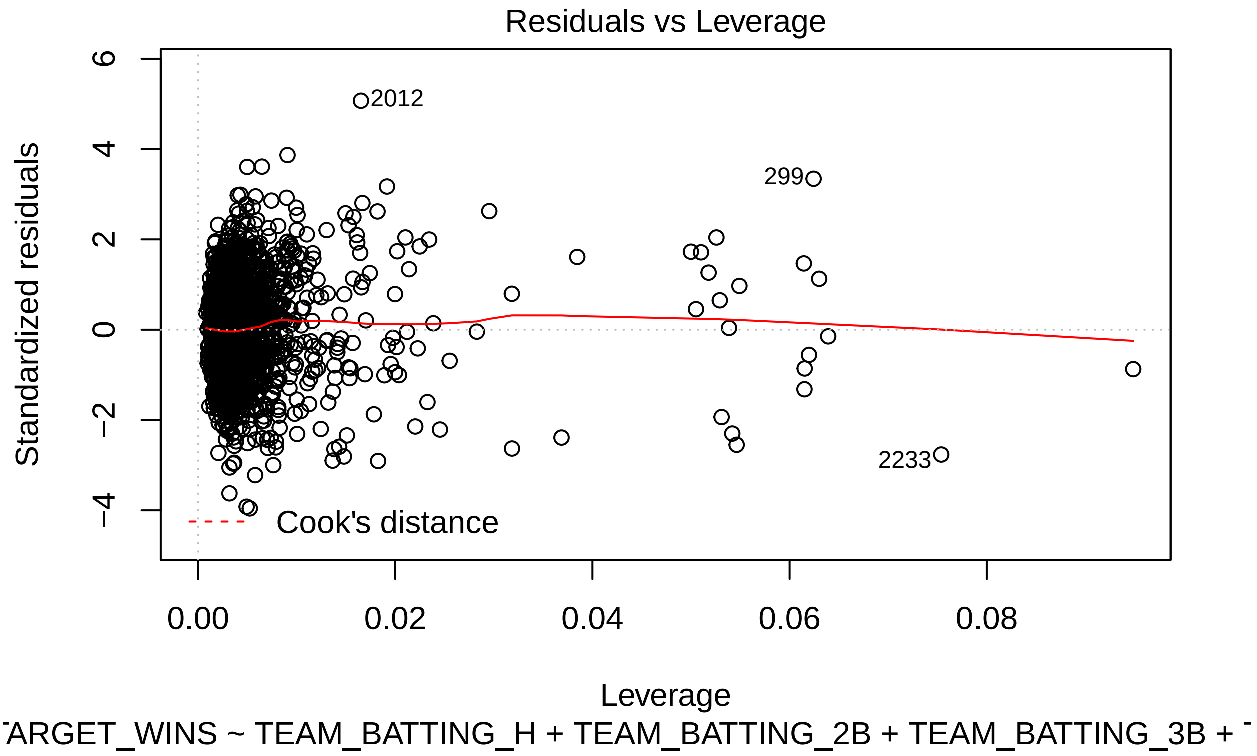
After multiple stepwise removal, we finally got below model that has all predictor variables which are statistically significant and VIF are less than 5.

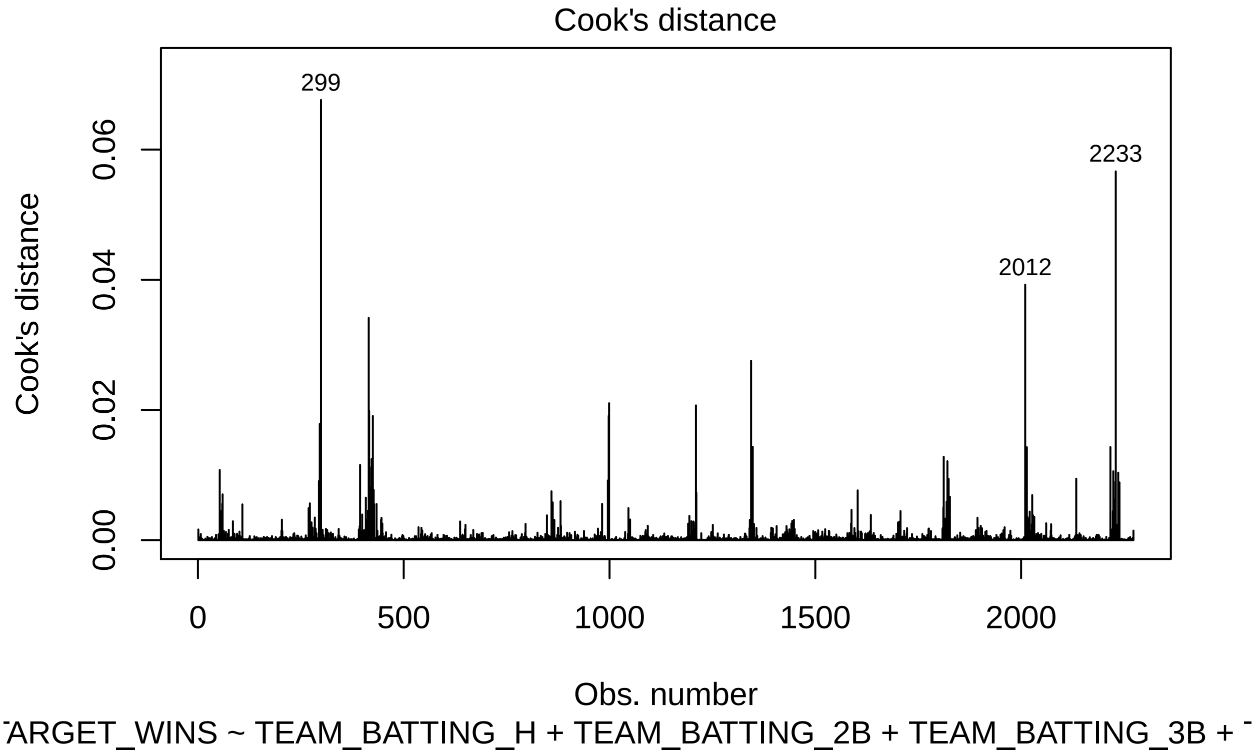
```
## 
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + 
##     TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + 
##     TEAM_BASERUN_SB + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP, 
##     data = df_mean_out_removed)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -49.203  -8.104  -0.096   7.971  62.687 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      -3.103e+02  3.786e+01  -8.196 4.12e-16 ***
## TEAM_BATTING_H    6.751e+01  5.347e+00  12.626  < 2e-16 ***
## TEAM_BATTING_2B  -2.581e-02  8.810e-03  -2.930  0.00343 ** 
## TEAM_BATTING_3B   9.154e-02  1.646e-02   5.560 3.01e-08 ***
## TEAM_BATTING_HR   6.617e-02  9.592e-03   6.899 6.79e-12 ***
## TEAM_BATTING_BB   1.271e-02  3.023e-03   4.203 2.74e-05 ***
## TEAM_BATTING_SO  -1.496e-02  2.421e-03  -6.178 7.68e-10 ***
## TEAM_BASERUN_SB   4.889e-02  4.264e-03  11.466  < 2e-16 ***
## TEAM_PITCHING_SO -2.573e-01  1.317e-01  -1.954  0.05081 .
## TEAM_FIELDING_E  -1.572e+01  9.240e-01 -17.015  < 2e-16 ***
## TEAM_FIELDING_DP -1.657e-01  1.293e-02 -12.822  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.47 on 2262 degrees of freedom
## Multiple R-squared:  0.3727, Adjusted R-squared:   0.37
## F-statistic: 134.4 on 10 and 2262 DF,  p-value: < 2.2e-16
##
##   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B  TEAM_BATTING_HR
##         3.643999         2.475414         3.070580         4.924033
##  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB TEAM_PITCHING_SO
##         1.993548         5.114304         2.193296         1.415675
##  TEAM_FIELDING_E TEAM_FIELDING_DP
##         4.671104         1.777510
```




Normal Q−Q
Standardized residuals
6
4
2
0
−2
−4
2012
859
2029
−3
−2
−1
0
1
2
3
Theoretical Quantiles
ARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +


Scale−Location
2012
859
2029
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
20
40
60
80
100
120
Fitted values
ARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +


Residuals vs Leverage
Standardized residuals
6
4
2
0
−2
−4
2012
299
2233
Cook's distance
0.00
0.02
0.04
0.06
0.08
Leverage
ARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +



Our final model seems to satisfy all the conditions of regression model and has low VIF. However, the adjusted R2 value is around `0.37`, which is quite low. So we will try a different model.

1.3.2 Model 2 - Principle component Regression

Lets take a different approach by creating a principle component regression which zeros-out the multicollinearity. This model uses PCA, which uses the highest variance as principle component.

As our dataset suffers from multi-collinerity, if we try to perform principle component Regression, it will reduce collinearity and produces better output.

```
## Data:    X dimension: 2276 13
##  Y dimension: 2276 1
## Fit method: svdpc
## Number of components considered: 13
## TRAINING: % variance explained
##              1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X             94.888   97.764   99.159   99.612    99.78    99.90    99.96
## TARGET_WINS    1.188    1.194    6.671    6.736    15.85    22.35    22.36
##              8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## X              99.98   100.00    100.00    100.00    100.00    100.00
## TARGET_WINS    25.62    25.99     26.16     26.16     32.09     36.53
```

```
## , , 13 comps
##
##                    TARGET_WINS
## (Intercept)      -3.147004e+02
## TEAM_BATTING_H    6.834777e+01
## TEAM_BATTING_2B  -2.253969e-02
## TEAM_BATTING_3B   7.591899e-02
## TEAM_BATTING_HR   5.591206e-02
## TEAM_BATTING_BB   2.526573e-03
## TEAM_BATTING_SO  -1.500378e-02
## TEAM_BASERUN_SB   4.705119e-02
## TEAM_PITCHING_H  -8.898771e-04
## TEAM_PITCHING_HR  7.357041e-03
## TEAM_PITCHING_BB  8.630060e-03
## TEAM_PITCHING_SO -3.407781e-01
## TEAM_FIELDING_E  -1.549512e+01
## TEAM_FIELDING_DP -1.645092e-01
```

```
## (Intercept)      1 comps      2 comps      3 comps      4 comps
##     0.00000      0.01188      0.01194      0.06671      0.06736
##     5 comps      6 comps      7 comps      8 comps      9 comps
##     0.15849      0.22348      0.22364      0.25624      0.25991
##    10 comps     11 comps     12 comps     13 comps
##     0.26160      0.26160      0.32088      0.36532
```

It seems after adding all the principle components, the R2 is still low. So in the next model, we will try a different approach.

1.3.3 Model 3 - Drop NA

It seems from last two models, any changes is not improving the model. So lets focus on the `NA` data and see if we can improve the model.

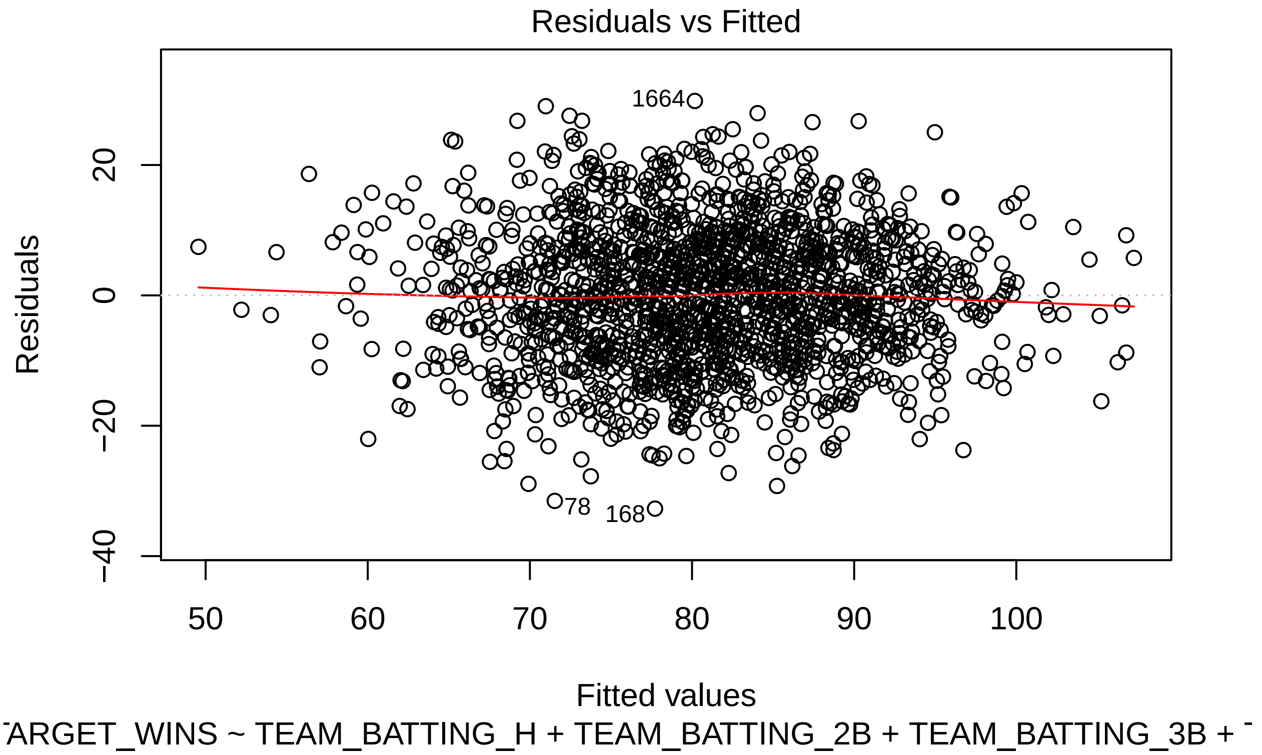
Tried all the `NA` values in different predictors and below strategy for other values. 1. `mean` imputation did not improve much. 2. `median` imputation did not improve much. 3. `kNN` imputation did not improve much.

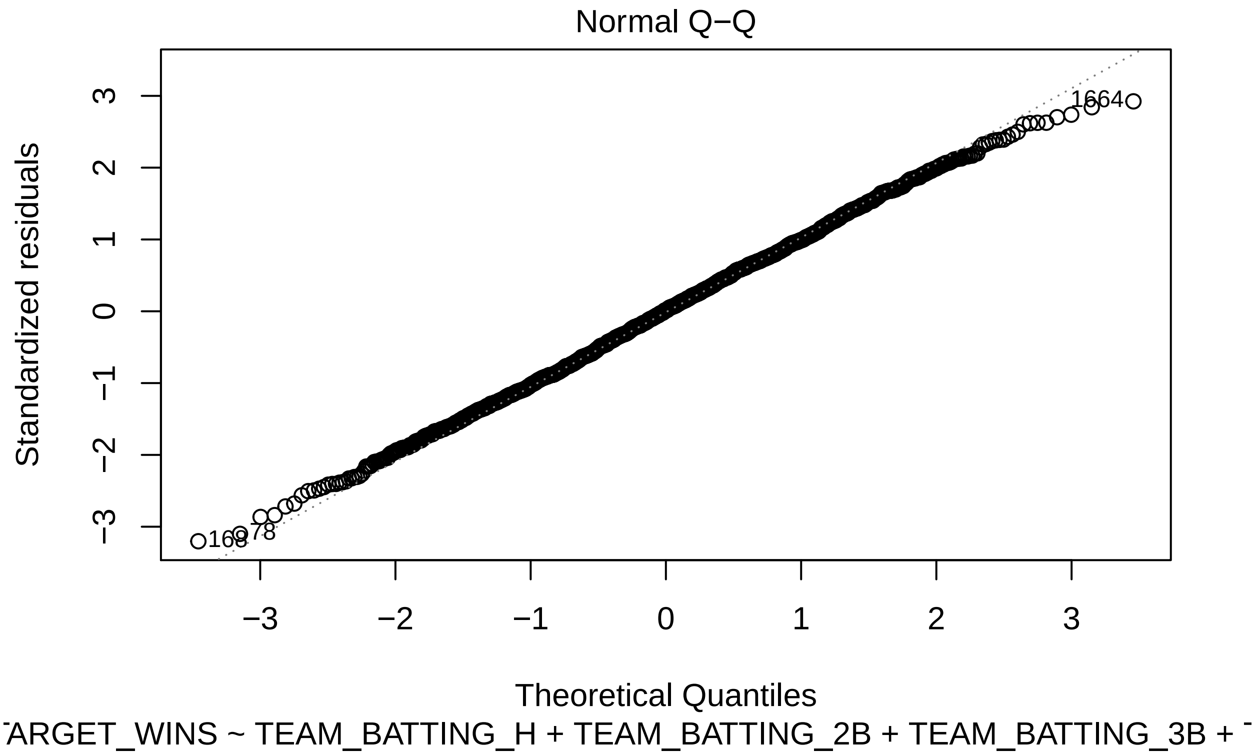
So in this model, we will drop all the `NA` rows and develop in this model.

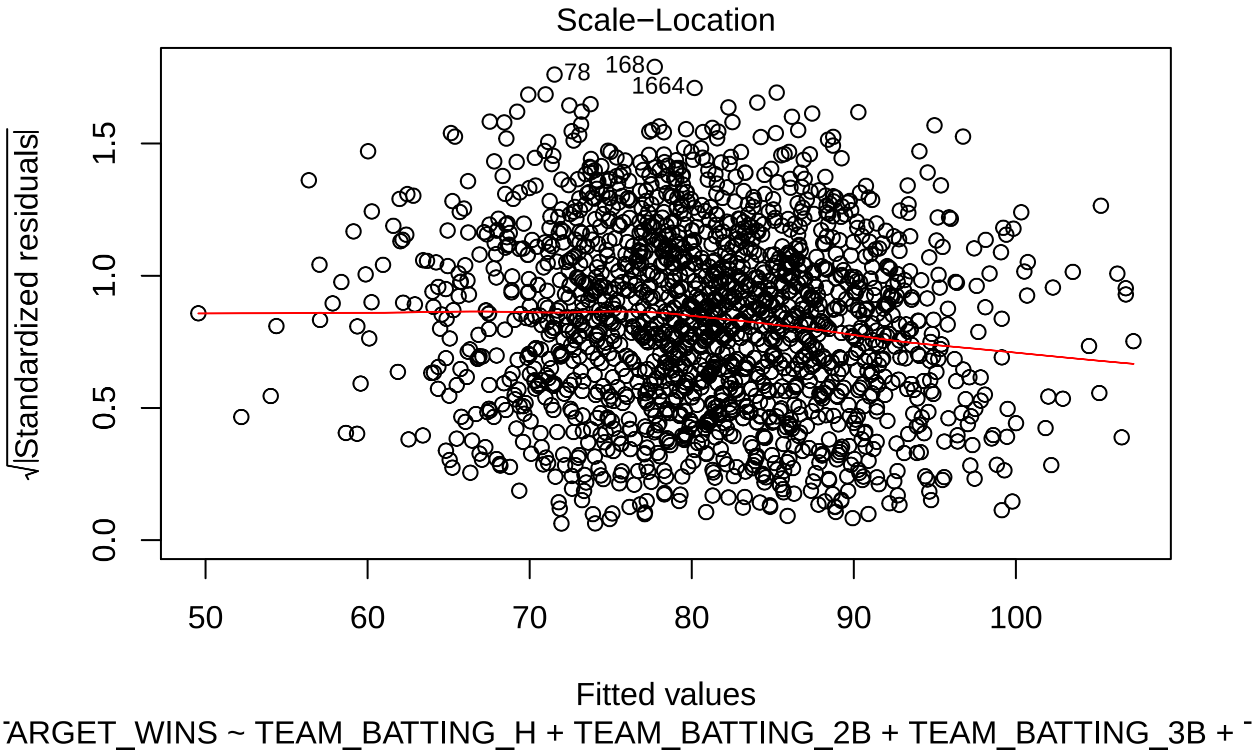
```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.856  -8.070   0.042   7.996  61.612
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -3.147e+02  3.861e+01  -8.151 5.93e-16 ***
## TEAM_BATTING_H    6.835e+01  5.445e+00  12.552  < 2e-16 ***
## TEAM_BATTING_2B  -2.254e-02  8.918e-03  -2.527  0.01156 *
## TEAM_BATTING_3B   7.592e-02  1.684e-02   4.509 6.84e-06 ***
## TEAM_BATTING_HR   5.591e-02  2.562e-02   2.182  0.02918 *
## TEAM_BATTING_BB   2.527e-03  4.666e-03   0.542  0.58820
## TEAM_BATTING_SO  -1.500e-02  2.447e-03  -6.132 1.02e-09 ***
## TEAM_BASERUN_SB   4.705e-02  4.364e-03  10.781  < 2e-16 ***
## TEAM_PITCHING_H  -8.899e-04  3.341e-04  -2.664  0.00779 **
## TEAM_PITCHING_HR  7.357e-03  2.218e-02   0.332  0.74013
## TEAM_PITCHING_BB  8.630e-03  3.006e-03   2.871  0.00413 **
## TEAM_PITCHING_SO -3.408e-01  1.368e-01  -2.490  0.01283 *
## TEAM_FIELDING_E  -1.550e+01  1.010e+00 -15.343  < 2e-16 ***
## TEAM_FIELDING_DP -1.645e-01  1.307e-02 -12.587  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.59 on 2262 degrees of freedom
## Multiple R-squared:  0.3653, Adjusted R-squared:  0.3617
## F-statistic: 100.2 on 13 and 2262 DF,  p-value: < 2.2e-16
##
##   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B  TEAM_BATTING_HR
##         3.750130         2.502378         3.178209        34.561298
##  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_H
##         4.705039         5.154090         2.264236         3.173121
## TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E
##        26.548069         3.591864         1.501128         5.527676
## TEAM_FIELDING_DP
##         1.785915
```

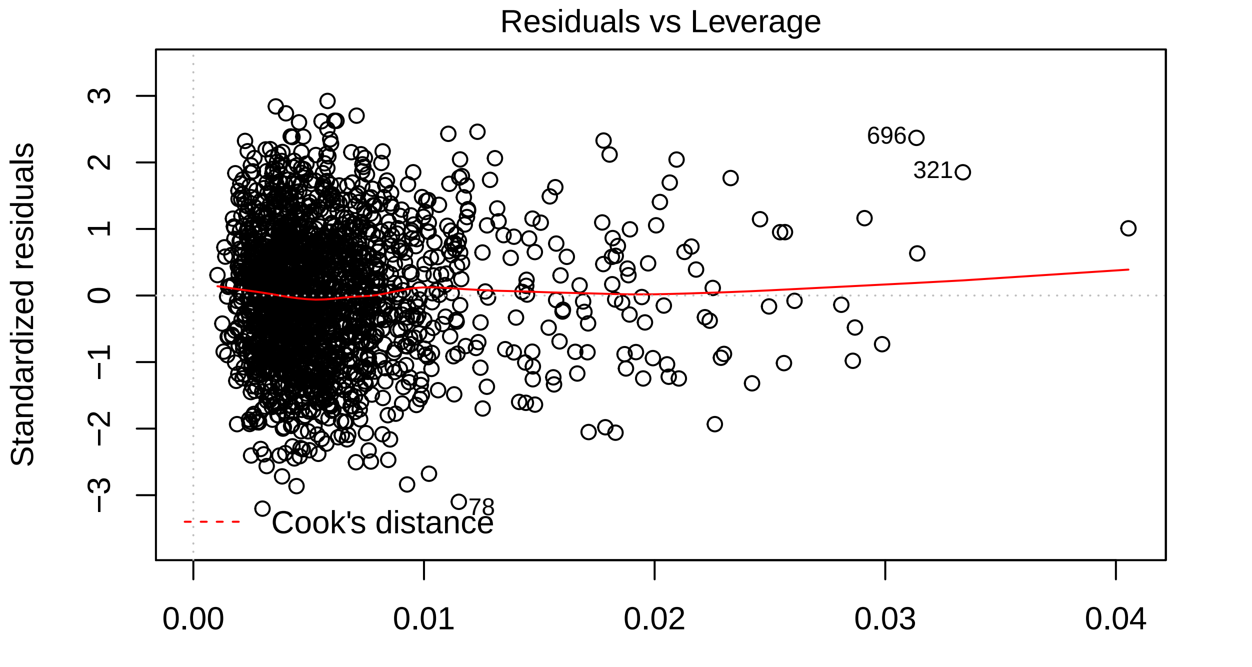
```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = df_na_out_removed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.158  -7.254   0.135   6.945  29.884
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      58.941092   6.030409   9.774  < 2e-16 ***
## TEAM_BATTING_H   -0.031483   0.016426  -1.917  0.05543 .
## TEAM_BATTING_2B  -0.049301   0.008876  -5.554 3.19e-08 ***
## TEAM_BATTING_3B   0.183608   0.018989   9.669  < 2e-16 ***
## TEAM_BATTING_HR   0.141783   0.081347   1.743  0.08151 .
## TEAM_BATTING_BB   0.113365   0.042521   2.666  0.00774 **
## TEAM_BATTING_SO   0.026511   0.021975   1.206  0.22781
## TEAM_BASERUN_SB   0.069369   0.005539  12.525  < 2e-16 ***
## TEAM_PITCHING_H   0.057619   0.014949   3.854  0.00012 ***
## TEAM_PITCHING_HR -0.040017   0.077904  -0.514  0.60754
## TEAM_PITCHING_BB -0.075474   0.040427  -1.867  0.06207 .
## TEAM_PITCHING_SO -0.046960   0.020918  -2.245  0.02489 *
## TEAM_FIELDING_E  -0.119149   0.007145 -16.676  < 2e-16 ***
## TEAM_FIELDING_DP -0.112120   0.012280  -9.131  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.18 on 1821 degrees of freedom
## Multiple R-squared:  0.4059, Adjusted R-squared:  0.4017
## F-statistic: 95.71 on 13 and 1821 DF,  p-value: < 2.2e-16
##
##   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B  TEAM_BATTING_HR
##        55.797863         2.573037         3.008565       348.099446
##  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_H
##       233.994861       402.442710         1.511412       120.066722
## TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E
##       341.908548       274.944168       382.532298         3.018138
## TEAM_FIELDING_DP
##         1.374047
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B +
##     TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO +
##     TEAM_BASERUN_SB + TEAM_PITCHING_H + TEAM_FIELDING_E + TEAM_FIELDING_DP,
##     data = df_na_out_removed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.717  -7.289   0.160   7.018  29.826
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      58.3103927  6.0470398   9.643  < 2e-16 ***
## TEAM_BATTING_H    0.0299013  0.0048307   6.190 7.42e-10 ***
## TEAM_BATTING_2B  -0.0497210  0.0089151  -5.577 2.81e-08 ***
## TEAM_BATTING_3B   0.1785813  0.0190541   9.372  < 2e-16 ***
## TEAM_BATTING_HR   0.1013044  0.0091995  11.012  < 2e-16 ***
## TEAM_BATTING_BB   0.0334030  0.0031434  10.626  < 2e-16 ***
## TEAM_BATTING_SO  -0.0226376  0.0023107  -9.797  < 2e-16 ***
## TEAM_BASERUN_SB   0.0716626  0.0055446  12.925  < 2e-16 ***
## TEAM_PITCHING_H  -0.0005784  0.0020335  -0.284    0.776
## TEAM_FIELDING_E  -0.1109846  0.0069404 -15.991  < 2e-16 ***
## TEAM_FIELDING_DP -0.1157091  0.0123166  -9.395  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.23 on 1824 degrees of freedom
## Multiple R-squared:  0.3987, Adjusted R-squared:  0.3954
## F-statistic: 120.9 on 10 and 1824 DF,  p-value: < 2.2e-16
##
##   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B  TEAM_BATTING_HR
##         4.775520         2.568702         2.997522         4.405461
##  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_H
##         1.265434         4.403454         1.498859         2.198551
##  TEAM_FIELDING_E TEAM_FIELDING_DP
##         2.818066         1.367909
```

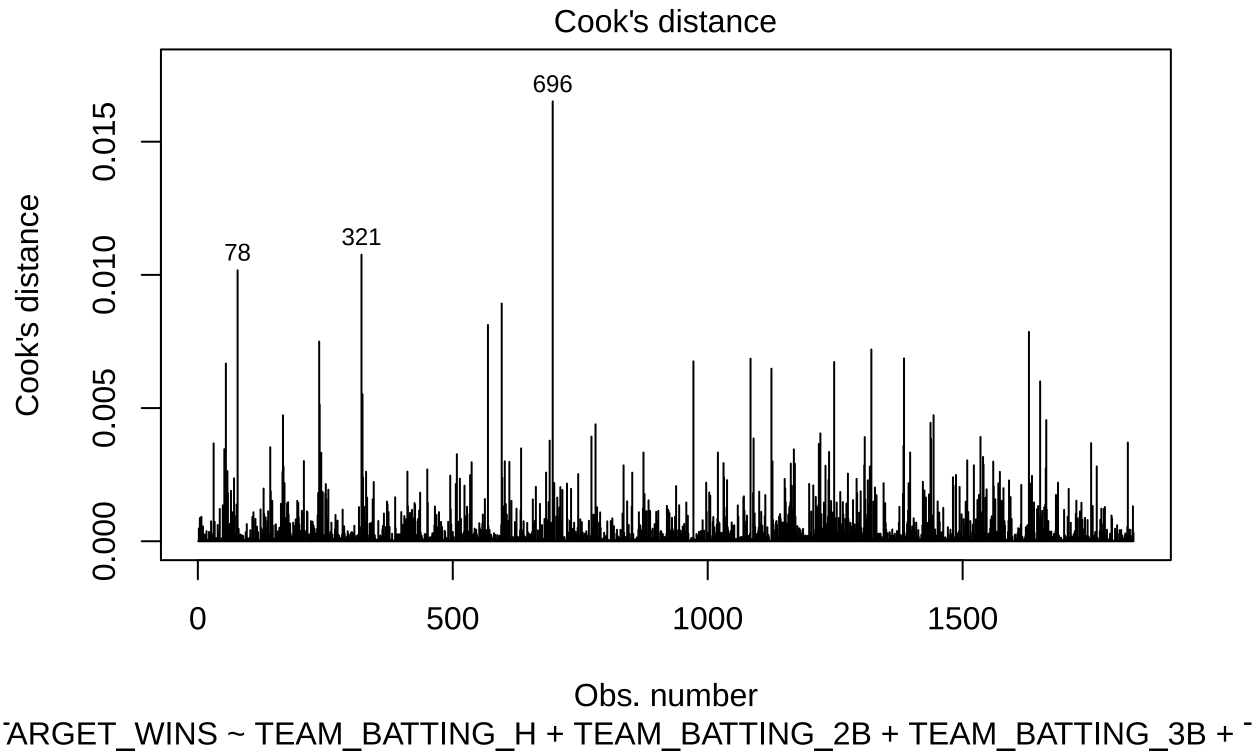

Residuals vs Fitted
Residuals
Fitted values
ARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
20
0
−20
−40
50
60
70
80
90
100
1664
78
168


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
168
78
1664
−3
−2
−1
0
1
2
3
Theoretical Quantiles
ARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +

Scale−Location
78
168
1664
√|Standardized residuals|
0.0
0.5
1.0
1.5
50
60
70
80
90
100
Fitted values
ARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +

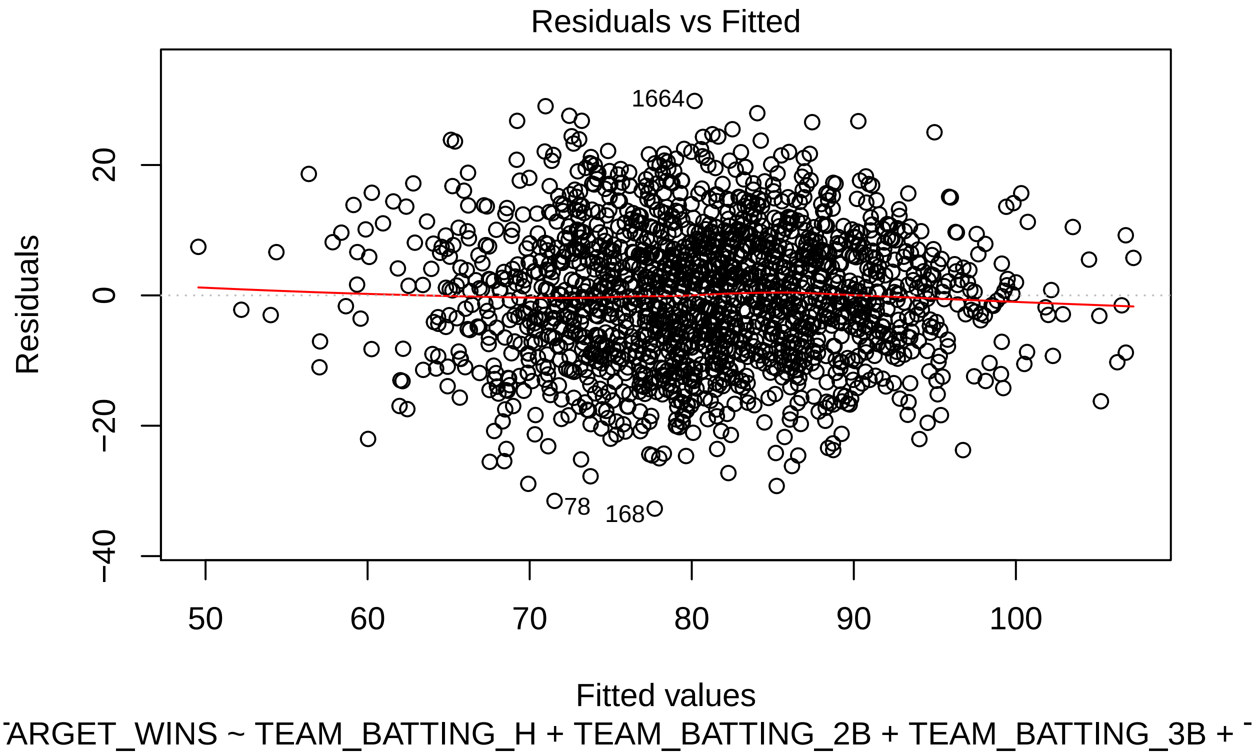


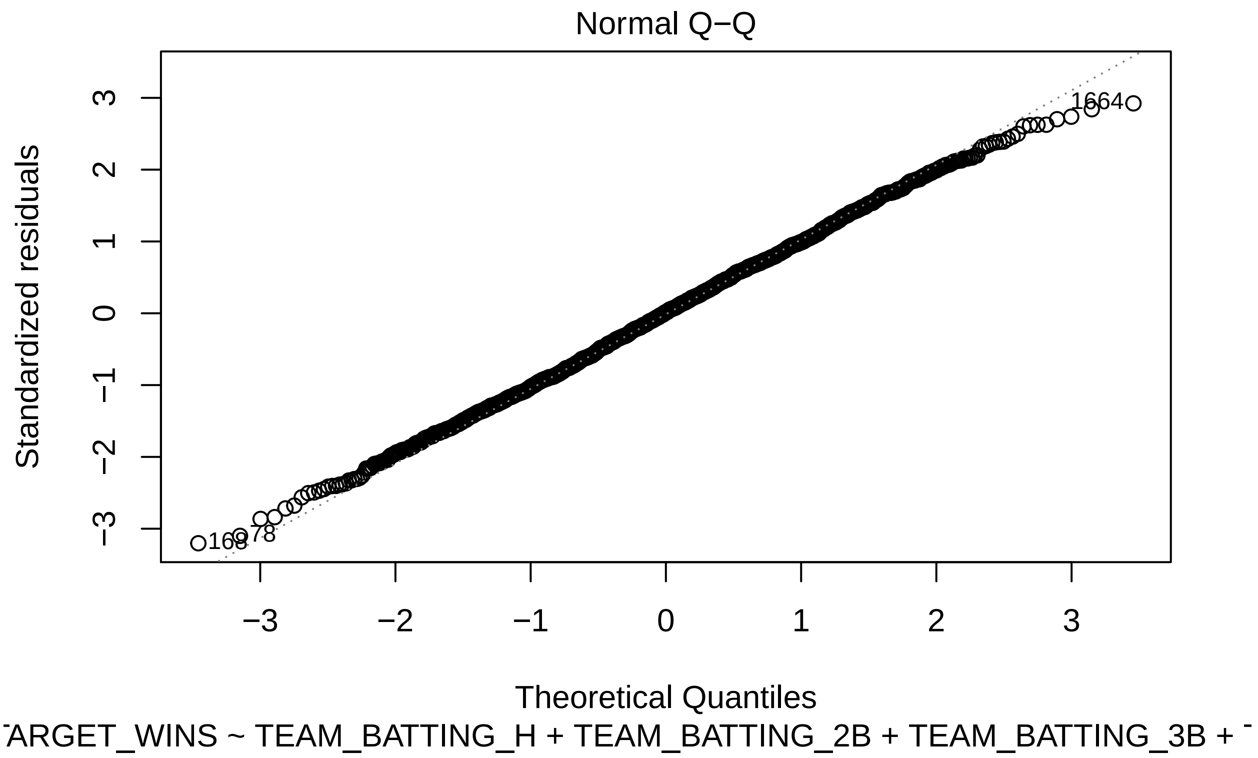
ARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +

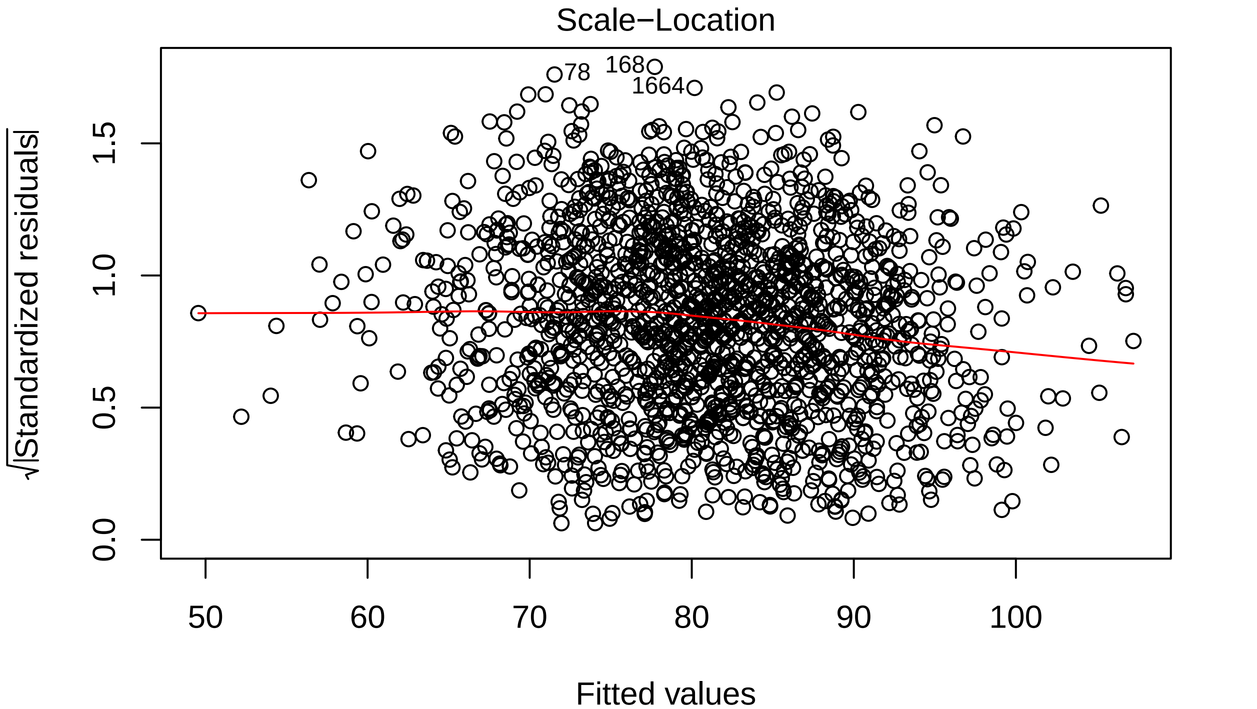


```
## Analysis of Variance Table
##
## Response: TARGET_WINS
##                    Df Sum Sq Mean Sq  F value  Pr(>F)
## TEAM_BATTING_H      1  39391   39391 376.2563 < 2e-16 ***
## TEAM_BATTING_2B     1    231     231   2.2031 0.13791
## TEAM_BATTING_3B     1    143     143   1.3676 0.24238
## TEAM_BATTING_HR     1  22666   22666 216.5028 < 2e-16 ***
## TEAM_BATTING_BB     1  15961   15961 152.4559 < 2e-16 ***
## TEAM_BATTING_SO     1    401     401   3.8283 0.05055 .
## TEAM_BASERUN_SB     1  16700   16700 159.5144 < 2e-16 ***
## TEAM_PITCHING_H     1    322     322   3.0727 0.07978 .
## TEAM_FIELDING_E     1  21537   21537 205.7221 < 2e-16 ***
## TEAM_FIELDING_DP    1   9240    9240  88.2586 < 2e-16 ***
## Residuals        1824 190956     105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

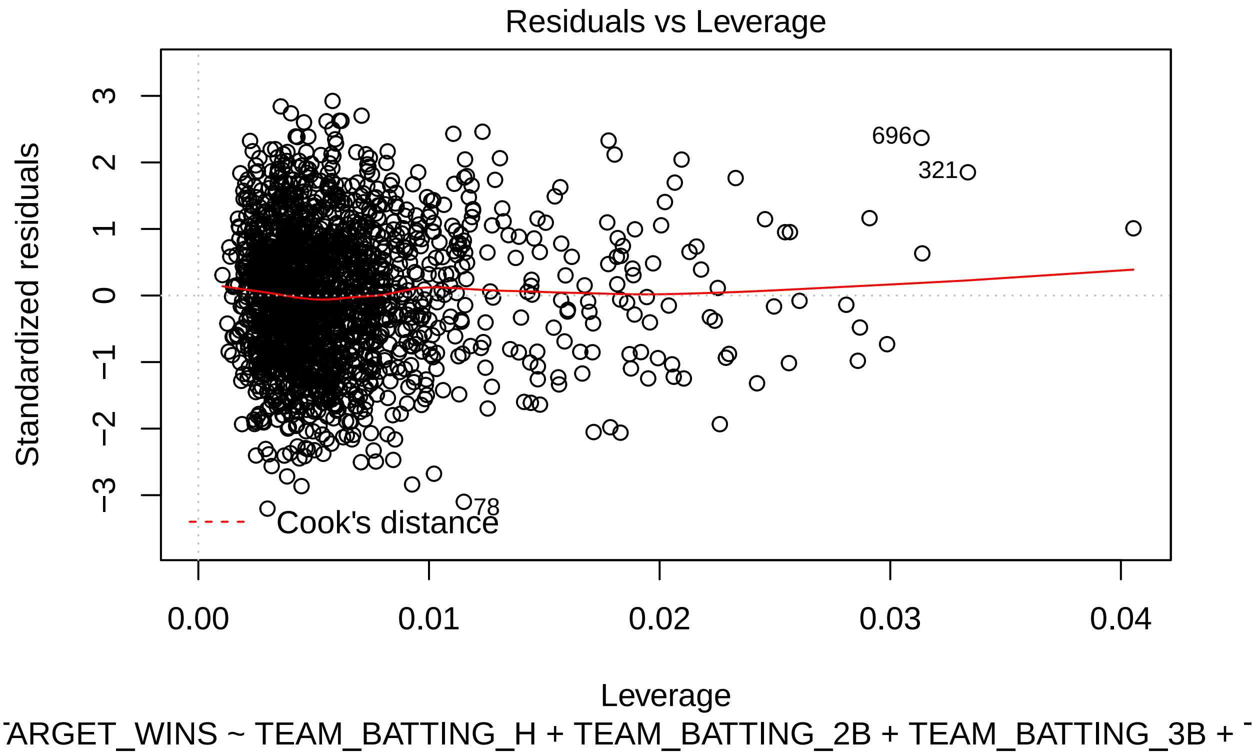
```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B +
##     TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO +
##     TEAM_BASERUN_SB + TEAM_PITCHING_H + TEAM_FIELDING_E + TEAM_FIELDING_DP,
##     data = df_na_out_removed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.717  -7.289   0.160   7.018  29.826
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      58.3103927  6.0470398   9.643  < 2e-16 ***
## TEAM_BATTING_H    0.0299013  0.0048307   6.190 7.42e-10 ***
## TEAM_BATTING_2B  -0.0497210  0.0089151  -5.577 2.81e-08 ***
## TEAM_BATTING_3B   0.1785813  0.0190541   9.372  < 2e-16 ***
## TEAM_BATTING_HR   0.1013044  0.0091995  11.012  < 2e-16 ***
## TEAM_BATTING_BB   0.0334030  0.0031434  10.626  < 2e-16 ***
## TEAM_BATTING_SO  -0.0226376  0.0023107  -9.797  < 2e-16 ***
## TEAM_BASERUN_SB   0.0716626  0.0055446  12.925  < 2e-16 ***
## TEAM_PITCHING_H  -0.0005784  0.0020335  -0.284    0.776
## TEAM_FIELDING_E  -0.1109846  0.0069404 -15.991  < 2e-16 ***
## TEAM_FIELDING_DP -0.1157091  0.0123166  -9.395  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.23 on 1824 degrees of freedom
## Multiple R-squared:  0.3987, Adjusted R-squared:  0.3954
## F-statistic: 120.9 on 10 and 1824 DF,  p-value: < 2.2e-16
##
##   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B  TEAM_BATTING_HR
##         4.775520         2.568702         2.997522         4.405461
##  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_H
##         1.265434         4.403454         1.498859         2.198551
##  TEAM_FIELDING_E TEAM_FIELDING_DP
##         2.818066         1.367909
```

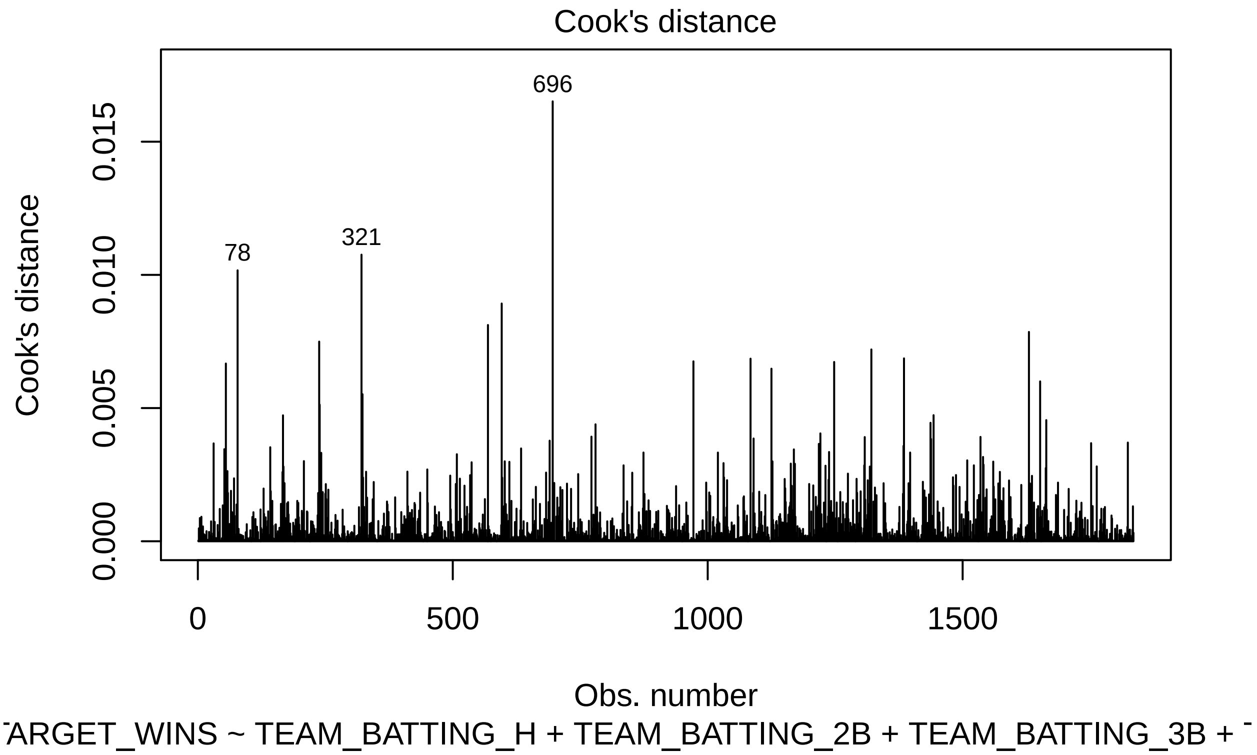

Residuals vs Fitted
Residuals
Fitted values
ARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
1664
78
168
20
0
−20
−40
50
60
70
80
90
100


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
1664
168
78
−3
−2
−1
0
1
2
3
Theoretical Quantiles
ARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +



ARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +

Residuals vs Leverage
Standardized residuals
3
2
1
0
−1
−2
−3
0.00
0.01
0.02
0.03
0.04
Leverage
ARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
696
321
78
Cook's distance

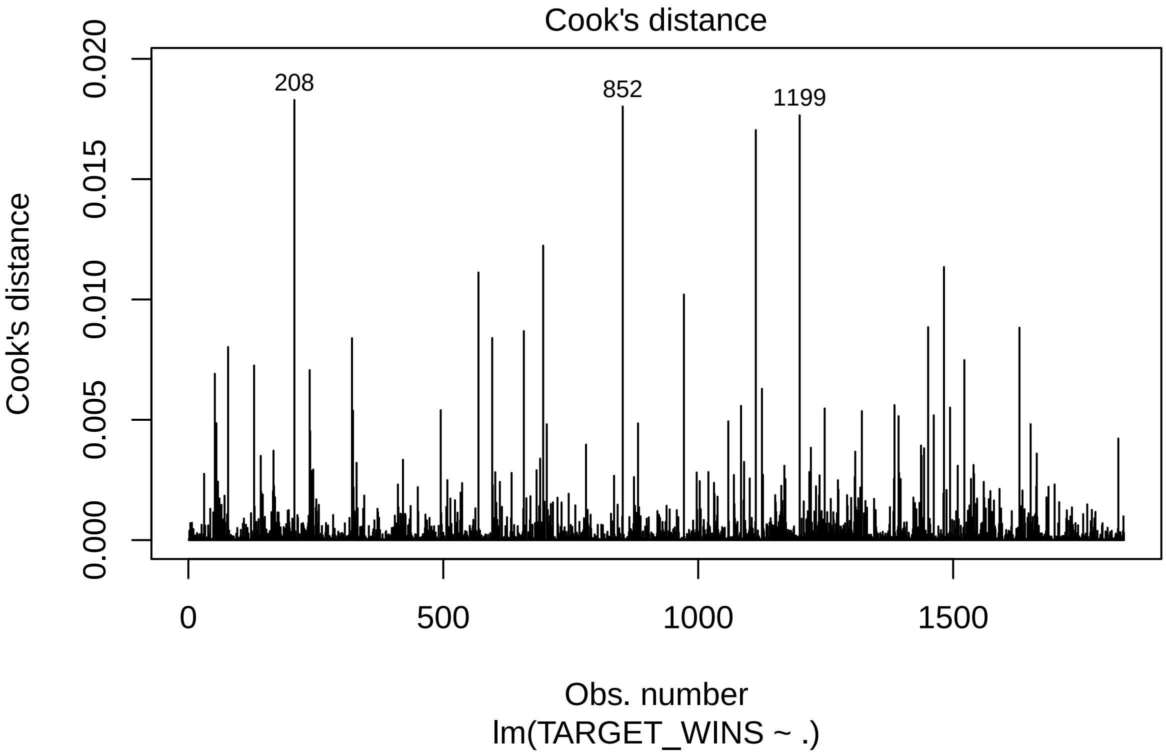


```
## Analysis of Variance Table
##
## Response: TARGET_WINS
##                    Df Sum Sq Mean Sq  F value  Pr(>F)
## TEAM_BATTING_H      1  39391   39391 376.2563 < 2e-16 ***
## TEAM_BATTING_2B     1    231     231   2.2031 0.13791
## TEAM_BATTING_3B     1    143     143   1.3676 0.24238
## TEAM_BATTING_HR     1  22666   22666 216.5028 < 2e-16 ***
## TEAM_BATTING_BB     1  15961   15961 152.4559 < 2e-16 ***
## TEAM_BATTING_SO     1    401     401   3.8283 0.05055 .
## TEAM_BASERUN_SB     1  16700   16700 159.5144 < 2e-16 ***
## TEAM_PITCHING_H     1    322     322   3.0727 0.07978 .
## TEAM_FIELDING_E     1  21537   21537 205.7221 < 2e-16 ***
## TEAM_FIELDING_DP    1   9240    9240  88.2586 < 2e-16 ***
## Residuals        1824 190956     105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally this model shows some level of improvements. It provides an adjusted R2 value of ~`0.39`. Most the predictors are statistically significant and has less VIF.

1.3.4 Model 4 - Scale and Transformations

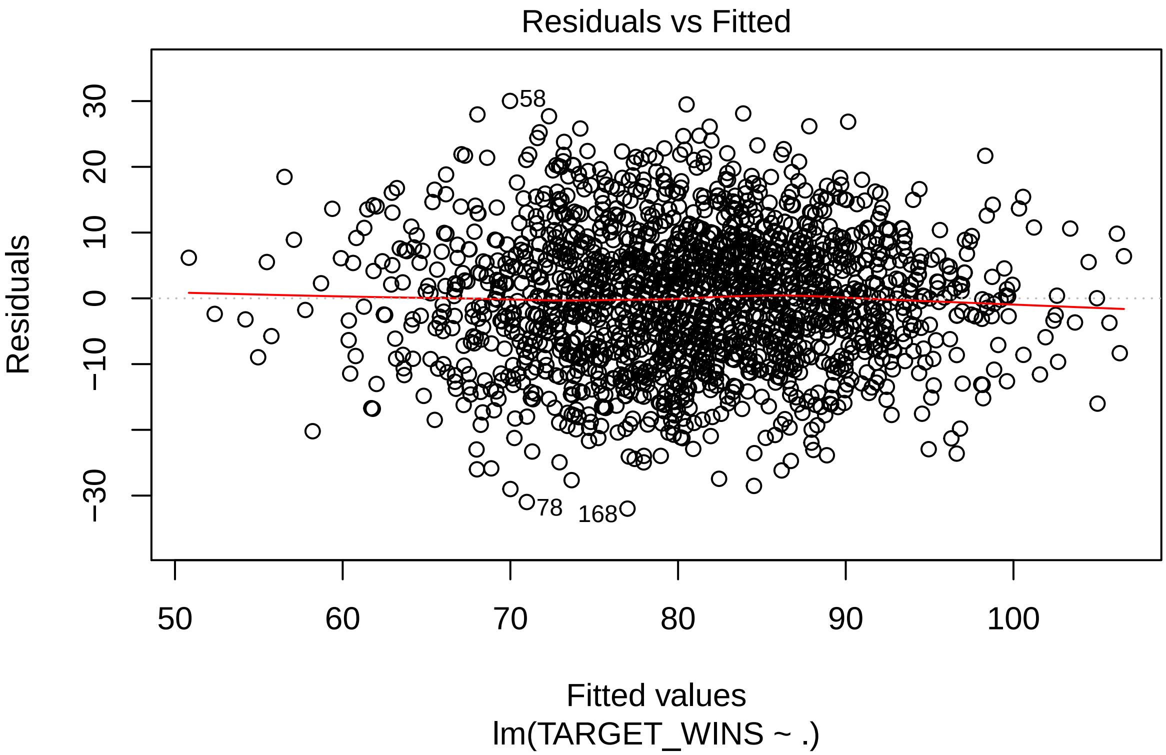
In the previous model, we have not scaled the data. In this model, we will to scale the predictors and remove the outliers.

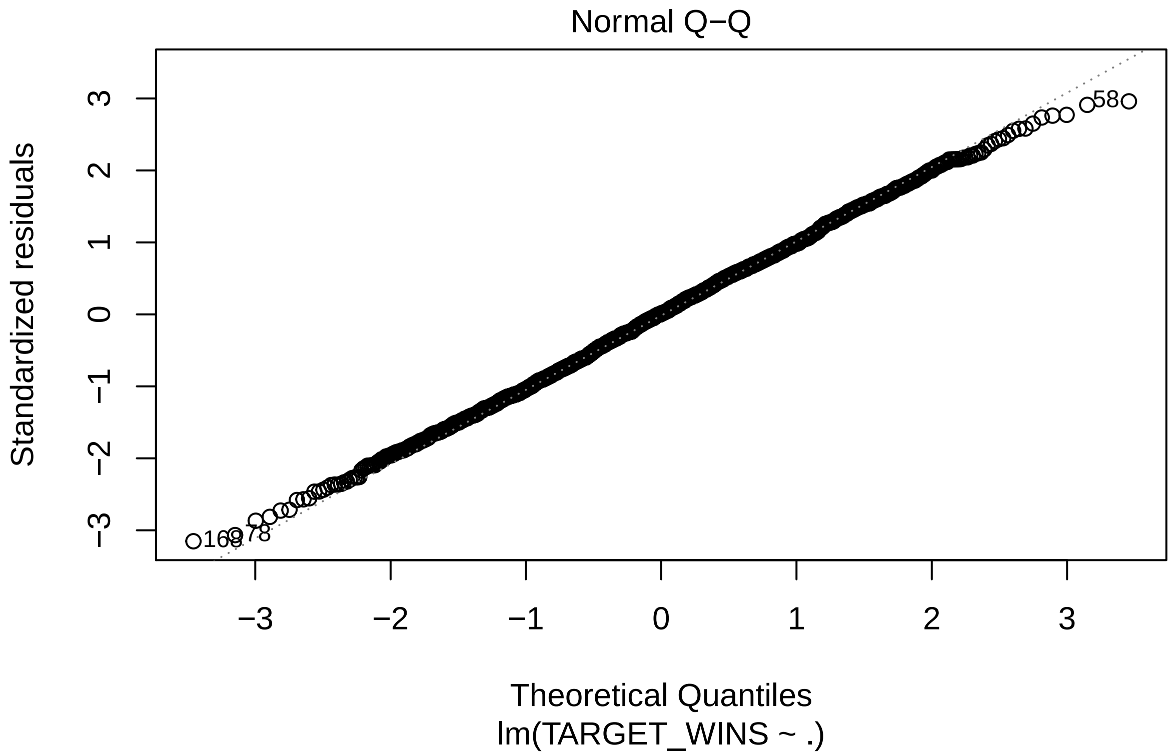


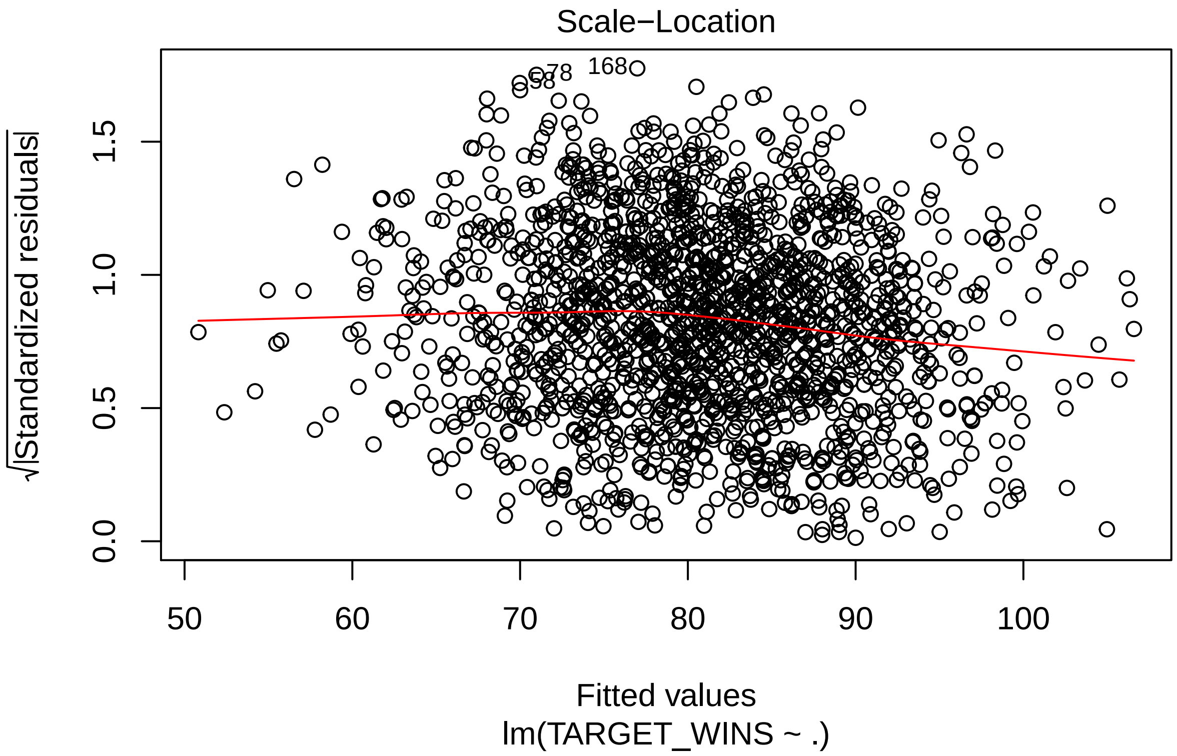
```
## [1] 1199  852  208
##      TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
## 1199         100      1.8408678       2.1690685     -1.14414614
## 852           79      0.3697658       0.4229302     -0.08469578
## 208           91     -0.8792830       0.1668300     -1.42052449
##      TEAM_BATTING_HR TEAM_BATTING_BB TEAM_BATTING_SO TEAM_BASERUN_SB
## 1199       1.5138509       2.6867816       0.7339103      -0.4132380
## 852       -0.3756635      -0.6698948      -0.5888397       1.0652830
## 208        0.3581285       1.0844657      -0.3998754      -0.7165243
##      TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO
## 1199        4.875526        2.9123050         5.498835        2.4210219
## 852         4.766464        0.5369221         1.949481        0.9810264
## 208         3.308483        1.5827996         4.175521        1.1790257
##      TEAM_FIELDING_E TEAM_FIELDING_DP
## 1199      -0.7774007        1.0843038
## 852       -0.8120092       -0.1057594
## 208       -0.9504435        1.1283802
```

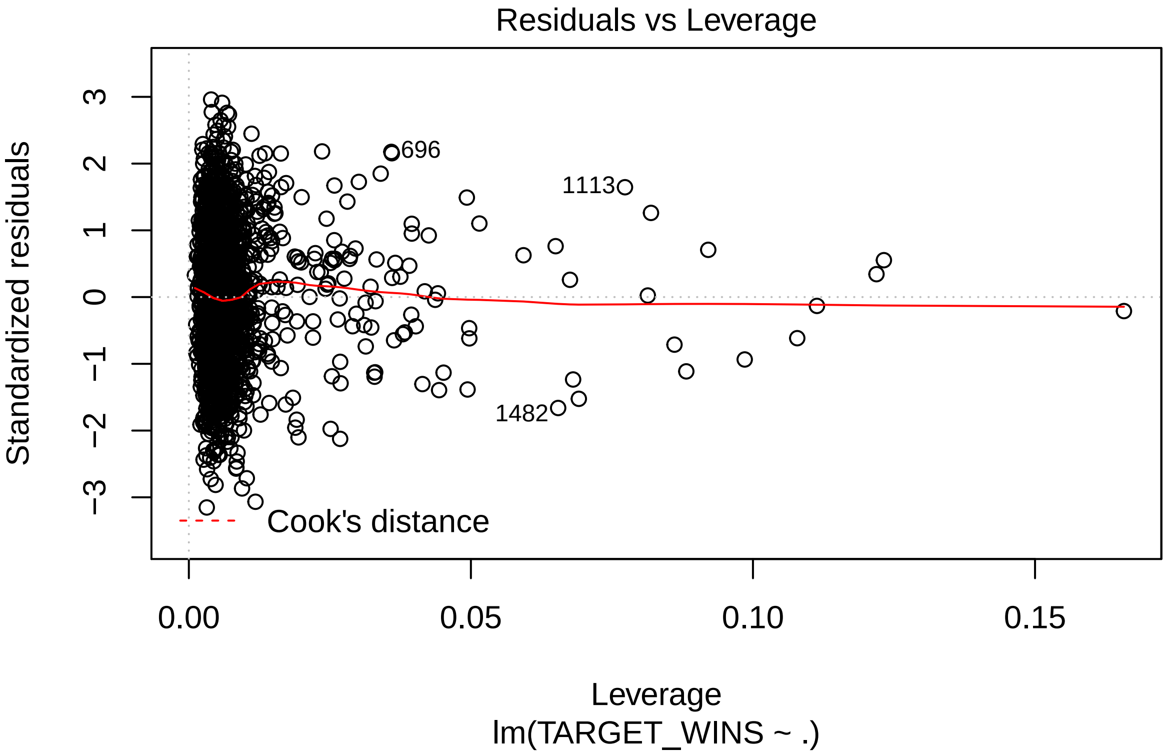
```
## 
## Call:
## lm(formula = TARGET_WINS ~ ., data = df_na_out_scale)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -31.984  -7.216   0.107   6.850  30.021 
## 
## Coefficients:
```

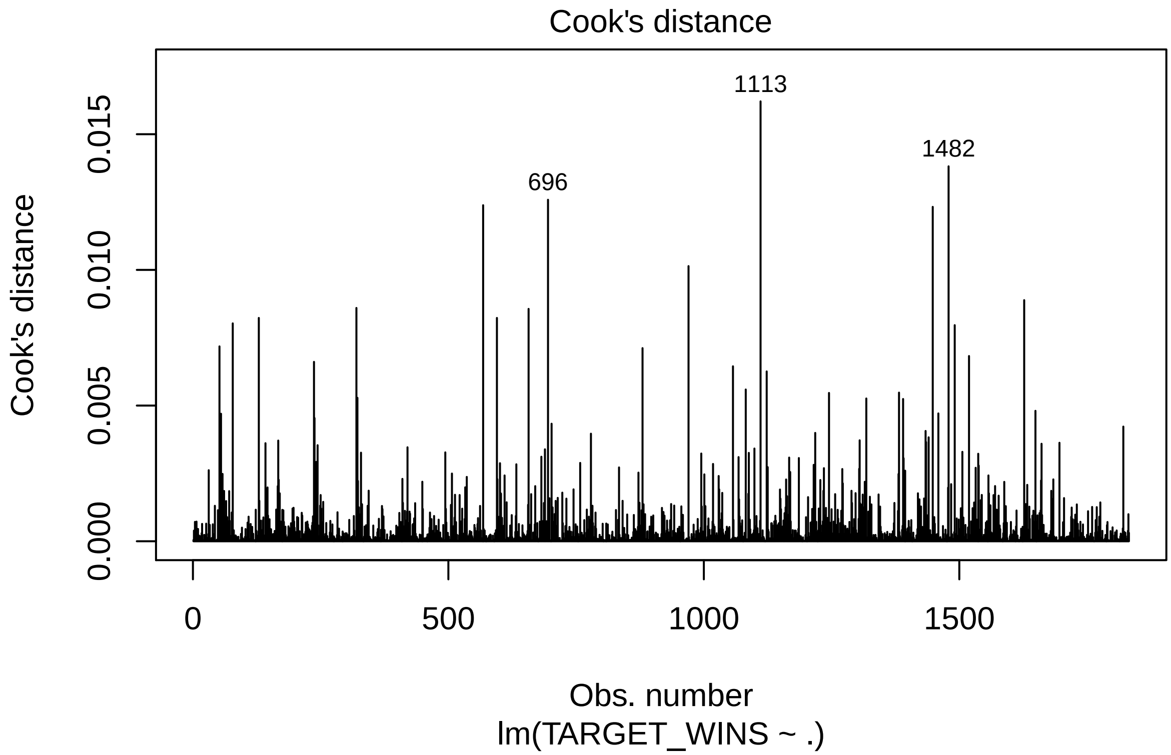
```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       80.9815     0.2375 341.001  < 2e-16 ***
## TEAM_BATTING_H    -4.7180     1.8313  -2.576 0.010066 *
## TEAM_BATTING_2B   -2.1195     0.3807  -5.567 2.98e-08 ***
## TEAM_BATTING_3B    4.0418     0.4124   9.801  < 2e-16 ***
## TEAM_BATTING_HR    8.8160     4.4909   1.963 0.049790 *
## TEAM_BATTING_BB   13.0910     3.8486   3.401 0.000685 ***
## TEAM_BATTING_SO    4.2554     4.9509   0.860 0.390159
## TEAM_BASERUN_SB    3.6623     0.2918  12.551  < 2e-16 ***
## TEAM_PITCHING_H   12.0372     2.6923   4.471 8.27e-06 ***
## TEAM_PITCHING_HR  -3.3581     4.4519  -0.754 0.450760
## TEAM_PITCHING_BB -11.0870     4.1772  -2.654 0.008020 **
## TEAM_PITCHING_SO  -8.9914     4.8304  -1.861 0.062848 .
## TEAM_FIELDING_E   -7.0165     0.4152 -16.897  < 2e-16 ***
## TEAM_FIELDING_DP  -2.5449     0.2782  -9.148  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.16 on 1818 degrees of freedom
## Multiple R-squared:  0.4078, Adjusted R-squared:  0.4036
## F-statistic:  96.3 on 13 and 1818 DF,  p-value: < 2.2e-16
##
##   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B  TEAM_BATTING_HR
##        59.411472         2.566986         3.014102       357.620737
##  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_H
##       261.739515       434.978869         1.510456       124.671919
## TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E
##       349.754257       301.125742       412.450145         3.058127
## TEAM_FIELDING_DP
##         1.372423
```


Residuals vs Fitted
Residuals
30
20
10
0
−10
−30
58
78
168
50
60
70
80
90
100
Fitted values
lm(TARGET_WINS ~ .)


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(TARGET_WINS ~ .)
58
169
78


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
50
60
70
80
90
100
58
78
168
Fitted values
lm(TARGET_WINS ~ .)

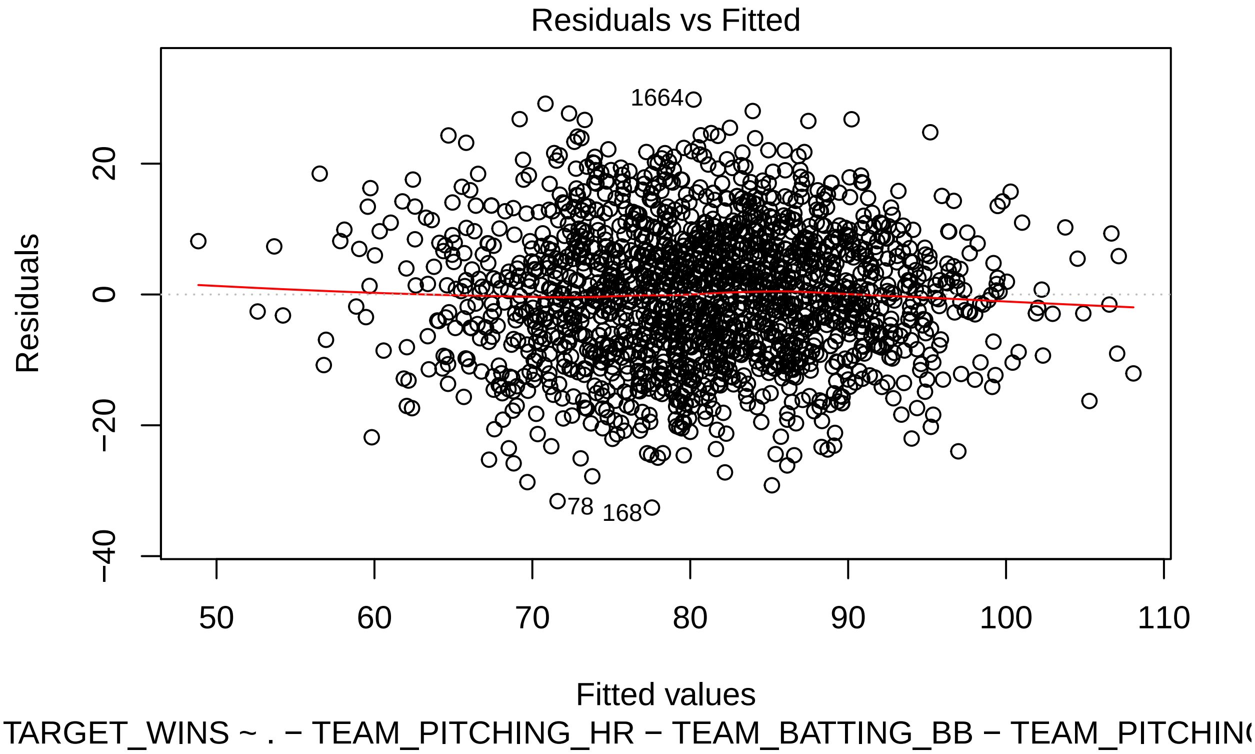

Residuals vs Leverage
Standardized residuals
3
2
1
0
−1
−2
−3
696
1113
1482
Cook's distance
0.00
0.05
0.10
0.15
Leverage
lm(TARGET_WINS ~ .)

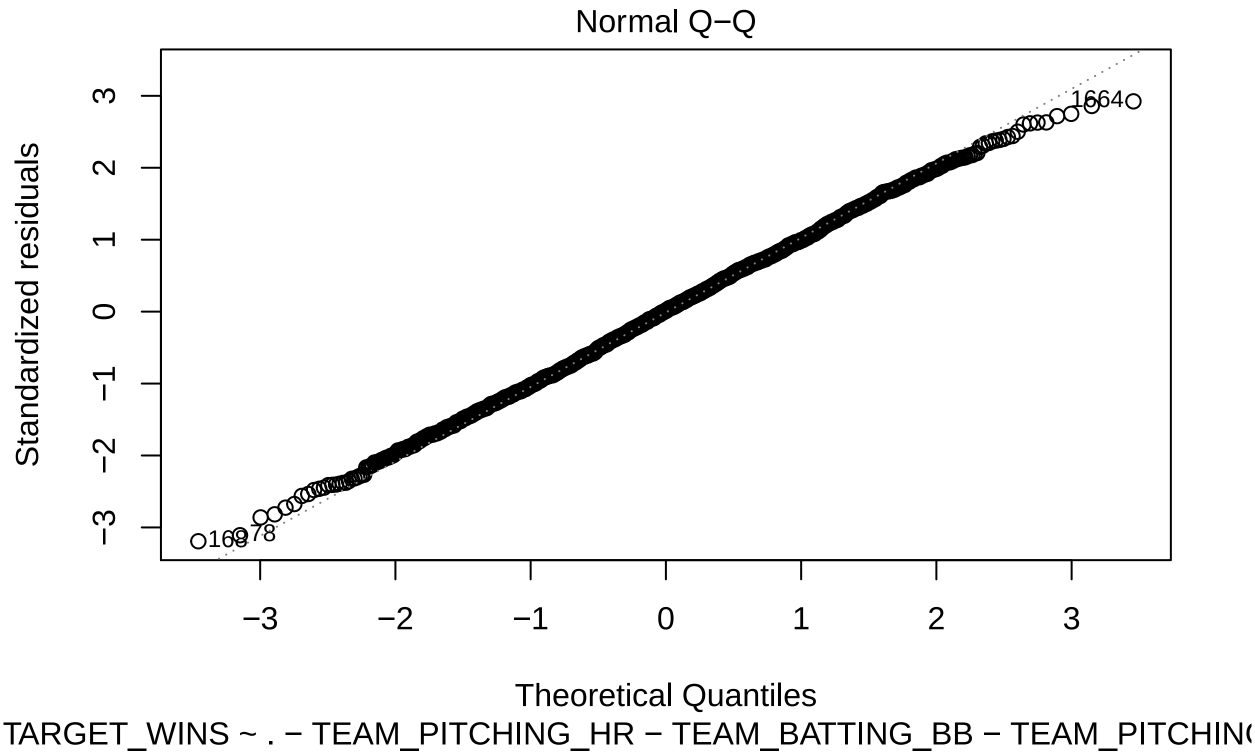


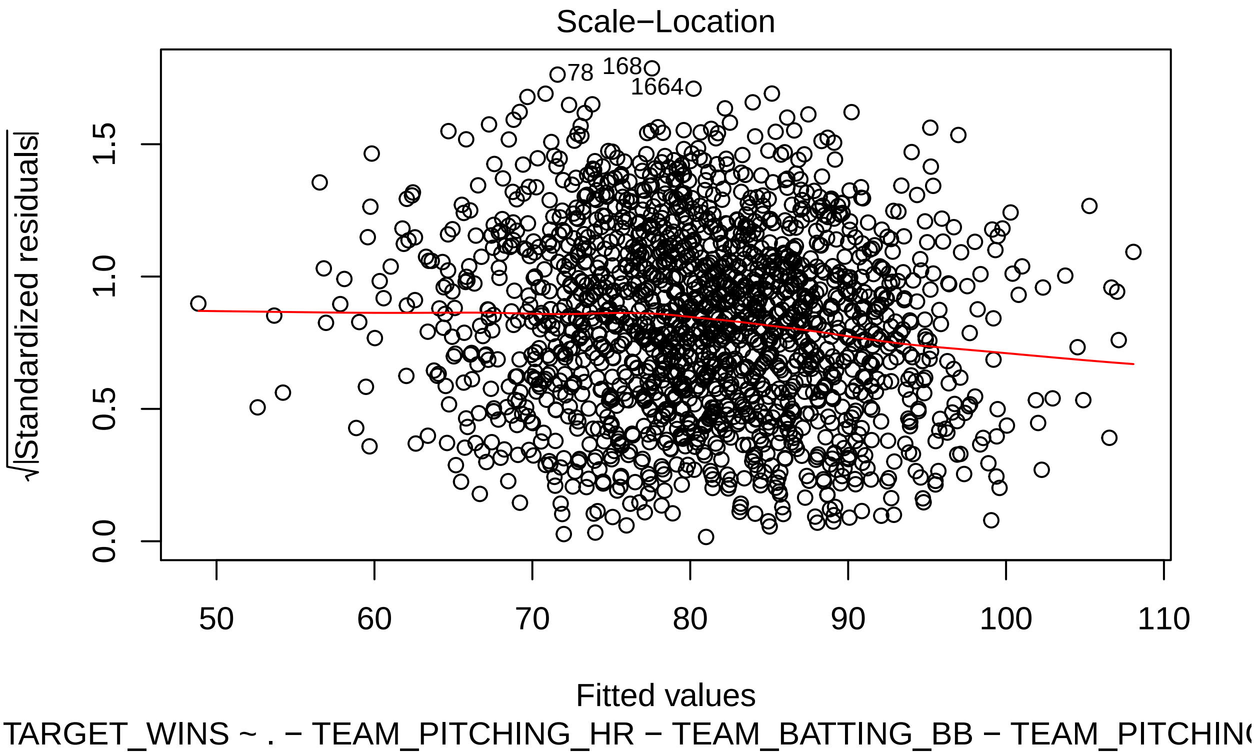
It seems scaling the predictor variables did not improve the model much. But removing 6 outliers has improved the model to ~`0.436`. However, it increased the VIF of predictor variables.

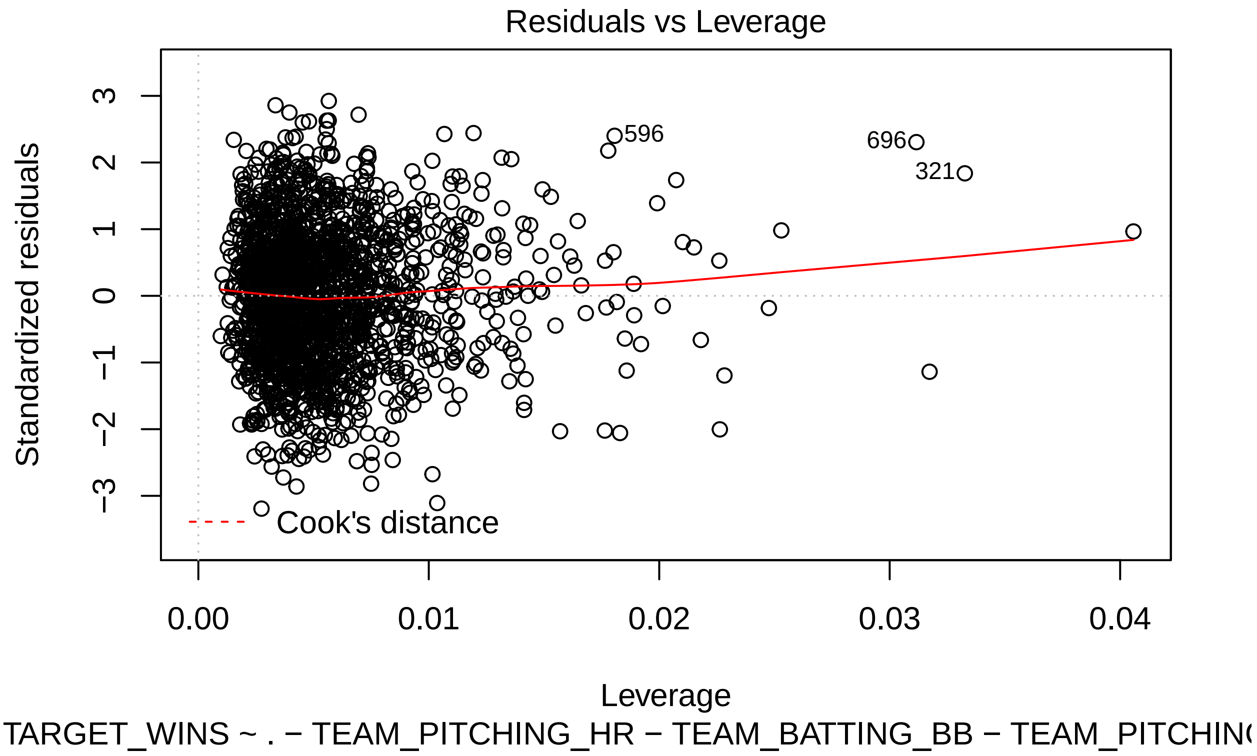
```
##
## Call:
## lm(formula = TARGET_WINS ~ . - TEAM_PITCHING_HR - TEAM_BATTING_BB -
##     TEAM_PITCHING_H - TEAM_BATTING_SO, data = df_na_out_scale)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.575  -7.235   0.112   7.001  29.789
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       80.9863     0.2388 339.199  < 2e-16 ***
## TEAM_BATTING_H     3.1239     0.4509   6.928 5.91e-12 ***
## TEAM_BATTING_2B   -2.1257     0.3772  -5.635 2.03e-08 ***
## TEAM_BATTING_3B    3.8824     0.4126   9.409  < 2e-16 ***
## TEAM_BATTING_HR    5.6124     0.4598  12.207  < 2e-16 ***
## TEAM_BASERUN_SB    3.8115     0.2873  13.265  < 2e-16 ***
## TEAM_PITCHING_BB   3.0413     0.2555  11.904  < 2e-16 ***
## TEAM_PITCHING_SO  -4.9299     0.4273 -11.537  < 2e-16 ***
## TEAM_FIELDING_E   -6.4711     0.3936 -16.439  < 2e-16 ***
## TEAM_FIELDING_DP  -2.6189     0.2790  -9.388  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

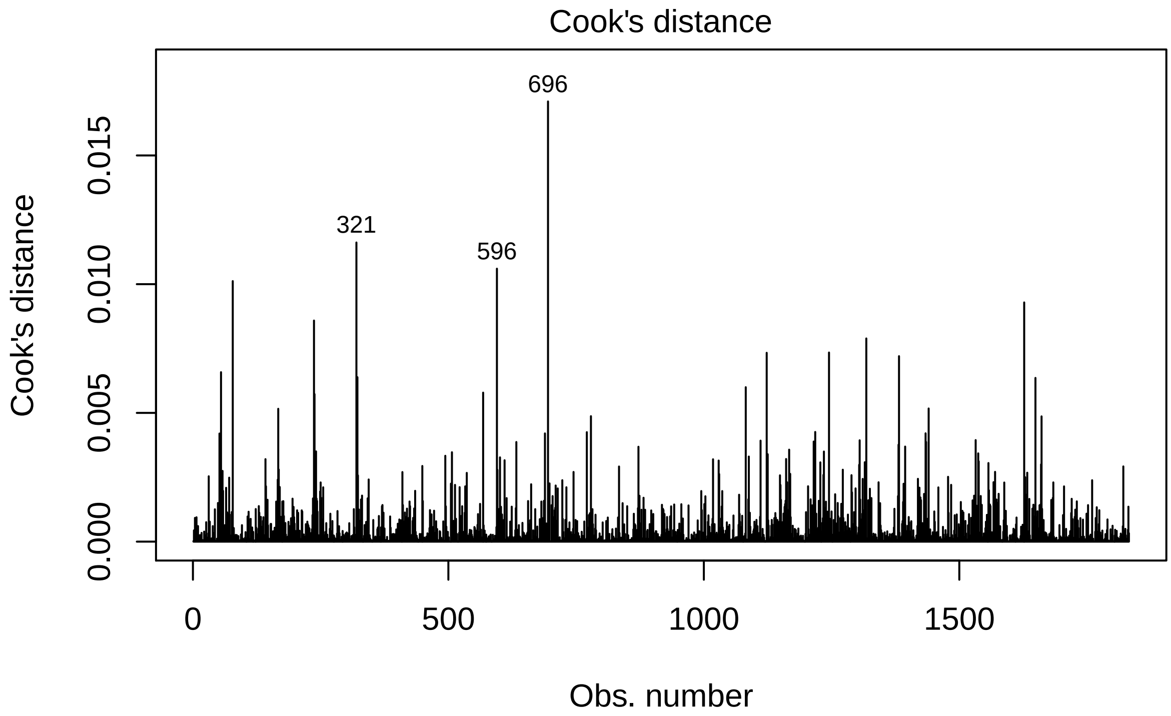
```
## Residual standard error: 10.22 on 1822 degrees of freedom
## Multiple R-squared:  0.3999, Adjusted R-squared:  0.397
## F-statistic: 134.9 on 9 and 1822 DF,  p-value: < 2.2e-16
##
##   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B  TEAM_BATTING_HR
##         3.562785         2.492740         2.984588         3.707053
##  TEAM_BASERUN_SB TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E
##         1.448612         1.114097         3.192329         2.717977
## TEAM_FIELDING_DP
##         1.365040
```




Normal Q−Q
Standardized residuals
−3
−2
−1
0
1
2
3
Theoretical Quantiles
−3
−2
−1
0
1
2
3
168
78
1664
TARGET_WINS ~ . − TEAM_PITCHING_HR − TEAM_BATTING_BB − TEAM_PITCHIN


Scale−Location
78
168
1664
√|Standardized residuals|
0.0
0.5
1.0
1.5
50
60
70
80
90
100
110
Fitted values
TARGET_WINS ~ . − TEAM_PITCHING_HR − TEAM_BATTING_BB − TEAM_PITCHIN


Residuals vs Leverage
Standardized residuals
3
2
1
0
−1
−2
−3
596
696
321
Cook's distance
0.00
0.01
0.02
0.03
0.04
Leverage
TARGET_WINS ~ . − TEAM_PITCHING_HR − TEAM_BATTING_BB − TEAM_PITCHIN



TARGET_WINS ~ . − TEAM_PITCHING_HR − TEAM_BATTING_BB − TEAM_PITCHIN

```
## Analysis of Variance Table
##
## Response: TARGET_WINS
##                    Df Sum Sq Mean Sq  F value Pr(>F)
## TEAM_BATTING_H      1  39247   39247 375.8274 <2e-16 ***
## TEAM_BATTING_2B     1    250     250   2.3960 0.1218
## TEAM_BATTING_3B     1    128     128   1.2237 0.2688
## TEAM_BATTING_HR     1  22747   22747 217.8271 <2e-16 ***
## TEAM_BASERUN_SB     1  11417   11417 109.3329 <2e-16 ***
## TEAM_PITCHING_BB    1  11857   11857 113.5406 <2e-16 ***
## TEAM_PITCHING_SO    1   9199    9199  88.0892 <2e-16 ***
## TEAM_FIELDING_E     1  22765   22765 217.9992 <2e-16 ***
## TEAM_FIELDING_DP    1   9203    9203  88.1253 <2e-16 ***
## Residuals        1822 190267     104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By removing the statistically insignificant predictors we get an adjusted R2 value of ~`0.397`.

1.3.4.1 Explanation of the variables

Practically, `TEAM_BATTING_2B`, `TEAM_PITCHING_SO`, `TEAM_FIELDING_E`, `TEAM_FIELDING_DP` decreases the effect of winning. Other variables increases the chances of winning.

However, the model approximatly explains `TARGET_WINS` around 40% of the time with provided predictor variables.

1.4 Select Models

In the final calculation of RMSE and adjusted R2 for all the models. With that information, all models are almost comparable with each other. If we want to select a model which makes sense, then it will be model 1.

Model 1 is selected because, it did not reject or omit NA observations. If we get more details about the data and have business knowledge, then we can correct the NA values and make a better model. Model 4 rejects the NA data. Often it is costlier to gather the data and reject it.

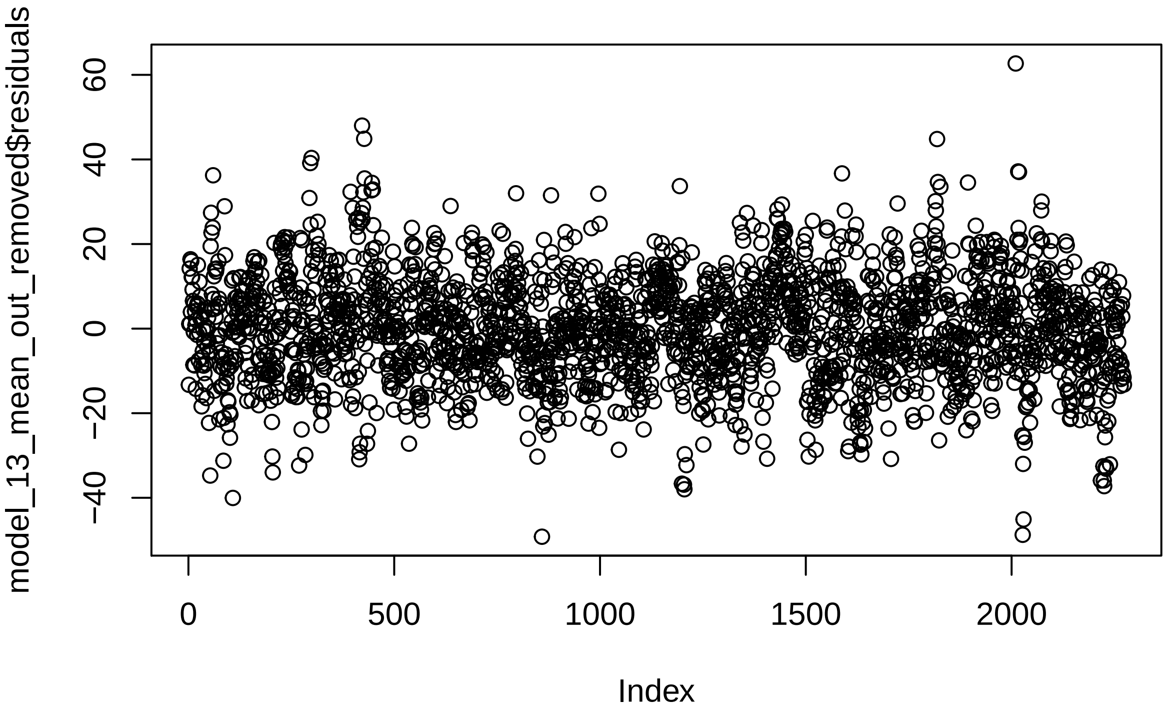
```
## [1] "RMSE: 12.4368108219245"
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B +
##     TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO +
##     TEAM_BASERUN_SB + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP,
##     data = df_mean_out_removed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.203  -8.104  -0.096   7.971  62.687
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -3.103e+02  3.786e+01  -8.196 4.12e-16 ***
## TEAM_BATTING_H    6.751e+01  5.347e+00  12.626  < 2e-16 ***
## TEAM_BATTING_2B  -2.581e-02  8.810e-03  -2.930  0.00343 **
## TEAM_BATTING_3B   9.154e-02  1.646e-02   5.560 3.01e-08 ***
## TEAM_BATTING_HR   6.617e-02  9.592e-03   6.899 6.79e-12 ***
## TEAM_BATTING_BB   1.271e-02  3.023e-03   4.203 2.74e-05 ***
## TEAM_BATTING_SO  -1.496e-02  2.421e-03  -6.178 7.68e-10 ***
## TEAM_BASERUN_SB   4.889e-02  4.264e-03  11.466  < 2e-16 ***
## TEAM_PITCHING_SO -2.573e-01  1.317e-01  -1.954  0.05081 .
## TEAM_FIELDING_E  -1.572e+01  9.240e-01 -17.015  < 2e-16 ***
## TEAM_FIELDING_DP -1.657e-01  1.293e-02 -12.822  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.47 on 2262 degrees of freedom
## Multiple R-squared:  0.3727, Adjusted R-squared:   0.37
## F-statistic: 134.4 on 10 and 2262 DF,  p-value: < 2.2e-16
```

```
## [1] "RMSE: 14.1610018692372"
## Data:    X dimension: 2276 13
##  Y dimension: 2276 1
## Fit method: svdpc
## Number of components considered: 13
## TRAINING: % variance explained
##              1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X             94.888   97.764   99.159   99.612    99.78    99.90    99.96
## TARGET_WINS    1.188    1.194    6.671    6.736    15.85    22.35    22.36
##              8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## X              99.98   100.00    100.00    100.00    100.00    100.00
## TARGET_WINS    25.62    25.99     26.16     26.16     32.09     36.53
## NULL
```

```
## [1] "RMSE: 10.2011322941555"
```

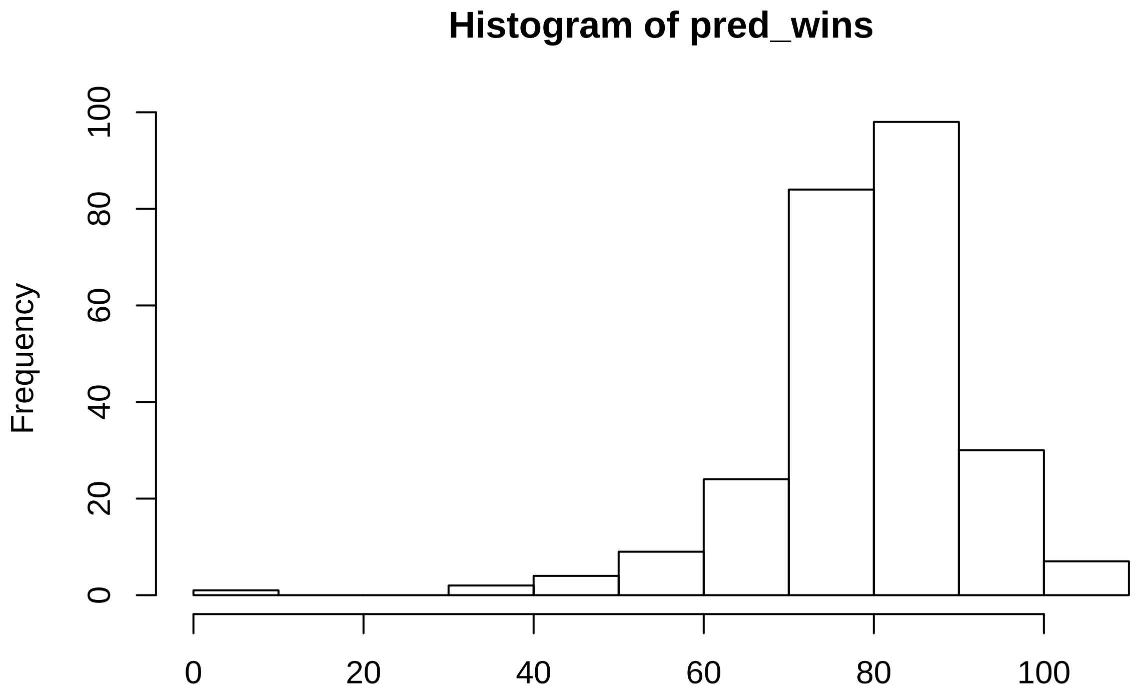
```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B +
##     TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO +
##     TEAM_BASERUN_SB + TEAM_PITCHING_H + TEAM_FIELDING_E + TEAM_FIELDING_DP,
##     data = df_na_out_removed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.717  -7.289   0.160   7.018  29.826
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      58.3103927  6.0470398   9.643  < 2e-16 ***
## TEAM_BATTING_H    0.0299013  0.0048307   6.190 7.42e-10 ***
## TEAM_BATTING_2B  -0.0497210  0.0089151  -5.577 2.81e-08 ***
## TEAM_BATTING_3B   0.1785813  0.0190541   9.372  < 2e-16 ***
## TEAM_BATTING_HR   0.1013044  0.0091995  11.012  < 2e-16 ***
## TEAM_BATTING_BB   0.0334030  0.0031434  10.626  < 2e-16 ***
## TEAM_BATTING_SO  -0.0226376  0.0023107  -9.797  < 2e-16 ***
## TEAM_BASERUN_SB   0.0716626  0.0055446  12.925  < 2e-16 ***
## TEAM_PITCHING_H  -0.0005784  0.0020335  -0.284    0.776
## TEAM_FIELDING_E  -0.1109846  0.0069404 -15.991  < 2e-16 ***
## TEAM_FIELDING_DP -0.1157091  0.0123166  -9.395  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.23 on 1824 degrees of freedom
## Multiple R-squared:  0.3987, Adjusted R-squared:  0.3954
## F-statistic: 120.9 on 10 and 1824 DF,  p-value: < 2.2e-16
```

```
## [1] "RMSE: 10.1910532661882"
##
## Call:
## lm(formula = TARGET_WINS ~ . - TEAM_PITCHING_HR - TEAM_BATTING_BB -
##     TEAM_PITCHING_H - TEAM_BATTING_SO, data = df_na_out_scale)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.575  -7.235   0.112   7.001  29.789
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       80.9863     0.2388 339.199  < 2e-16 ***
## TEAM_BATTING_H     3.1239     0.4509   6.928 5.91e-12 ***
## TEAM_BATTING_2B   -2.1257     0.3772  -5.635 2.03e-08 ***
## TEAM_BATTING_3B    3.8824     0.4126   9.409  < 2e-16 ***
## TEAM_BATTING_HR    5.6124     0.4598  12.207  < 2e-16 ***
## TEAM_BASERUN_SB    3.8115     0.2873  13.265  < 2e-16 ***
## TEAM_PITCHING_BB   3.0413     0.2555  11.904  < 2e-16 ***
## TEAM_PITCHING_SO  -4.9299     0.4273 -11.537  < 2e-16 ***
## TEAM_FIELDING_E   -6.4711     0.3936 -16.439  < 2e-16 ***
## TEAM_FIELDING_DP  -2.6189     0.2790  -9.388  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.22 on 1822 degrees of freedom
## Multiple R-squared:  0.3999, Adjusted R-squared:  0.397
## F-statistic: 134.9 on 9 and 1822 DF,  p-value: < 2.2e-16
```



1.4.1 Predictions

Metric	Model1	Model2	Model3	Model4
RSE	12.43	14.16	10.20	10.19
R^2	0.3727	0.3653	0.3987	0.399
Adj. R^2	0.37	0.3653	0.3954	0.397
F Stat.	134.4	-	120.9	134.9



```
## [1] "Mean predicted wins:"

## [1] 79.62934
```

1.5 Summary

We have perfomed different transformations and created multiple models. Almost all the models are comparable. But we have choosen the best model and compared it with other models. Given the knowledge of the baseball game is limited, we were not able to add many new variables and perform imputation which is relevent. For now, I belive the model can be used to predict the wins for unseen data.