≡

DATA 621 01[46893] : HomeWork1

Code ▾

CUNY_MSDA_DATA 621_Homework

1 Overview

In this homework assignment, you will explore, analyze and model a data set containing approximately 2200 records. Each record represents a professional baseball team from the years 1871 to 2006 inclusive. Each record has the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season.

We have been given a dataset with 2276 records summarizing a major league baseball team's season. The records span 1871 to 2006 inclusive. All statistics have been adjusted to match the performance of a 162 game season.

Your objective is to build a multiple linear regression model on the training data to predict the number of wins for the team. You can only use the variables given to you (or variables that you derive from the variables provided).

Glossary of data

Code

Below is a short description of the variables of interest in the data set:

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_WINS	Number of wins	
TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)	Positive Impact on Wins
TEAM_BATTING_2B	Doubles by batters (2B)	Positive Impact on Wins
TEAM_BATTING_3B	Triples by batters (3B)	Positive Impact on Wins
TEAM_BATTING_HR	Homeruns by batters (4B)	Positive Impact on Wins
TEAM_BATTING_BB	Walks by batters	Positive Impact on Wins
TEAM_BATTING_HBP	Batters hit by pitch (get a free base)	Positive Impact on Wins
TEAM_BATTING_SO	Strikeouts by batters	Negative Impact on Wins
TEAM_BASERUN_SB	Stolen bases	Positive Impact on Wins
TEAM_BASERUN_CS	Caught stealing	Negative Impact on Wins
TEAM_FIELDING_E	Errors	Negative Impact on Wins
TEAM_FIELDING_DP	Double Plays	Positive Impact on Wins
TEAM_PITCHING_BB	Walks allowed	Negative Impact on Wins
TEAM_PITCHING_H	Hits allowed	Negative Impact on Wins
TEAM_PITCHING_HR	Homeruns allowed	Negative Impact on Wins
TEAM_PITCHING_SO	Strikeouts by pitchers	Positive Impact on Wins

2 Deliverables

- 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.
- Assigned predictions (the number of wins for the team) for the evaluation data set.
- Include your R statistical programming code in an Appendix.

3 DATA EXPLORATION

The data set describes baseball team statistics for the years 1871 to 2006 inclusive. Each record in the data set represents the performance of the team for the given year adjusted to the current length of the season - 162 games. The data set includes 16 variables and the training set includes 2,276 records.

##Load the data and understand the data by using some stats and plott

Code

3.1 View rows and columns, variable types

Glimpse of the data shows that all variables are numeric, no ctegorical variable is present here. We do lots of NA for few predcitors in the data set. In our furthe analysis we will try to identify :

- Structure of the each predictors
- How Many NA and Zero , is it significant to remove them or replace them with some predicted value.
- Statistical summary of the data

Code

```
## Observations: 2,276
## Variables: 17
## $ INDEX            <int> 1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 15, 16, 17, 18...
## $ TARGET_WINS      <int> 39, 70, 86, 70, 82, 75, 80, 85, 86, 76, 78, 68, 72...
## $ TEAM_BATTING_H   <int> 1445, 1339, 1377, 1387, 1297, 1279, 1244, 1273, 13...
## $ TEAM_BATTING_2B  <int> 194, 219, 232, 209, 186, 200, 179, 171, 197, 213, ...
## $ TEAM_BATTING_3B  <int> 39, 22, 35, 38, 27, 36, 54, 37, 40, 18, 27, 31, 41...
## $ TEAM_BATTING_HR  <int> 13, 190, 137, 96, 102, 92, 122, 115, 114, 96, 82, ...
## $ TEAM_BATTING_BB  <int> 143, 685, 602, 451, 472, 443, 525, 456, 447, 441, ...
## $ TEAM_BATTING_SO  <int> 842, 1075, 917, 922, 920, 973, 1062, 1027, 922, 82...
## $ TEAM_BASERUN_SB  <int> NA, 37, 46, 43, 49, 107, 80, 40, 69, 72, 60, 119, ...
## $ TEAM_BASERUN_CS  <int> NA, 28, 27, 30, 39, 59, 54, 36, 27, 34, 39, 79, 10...
## $ TEAM_BATTING_HBP <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ TEAM_PITCHING_H  <int> 9364, 1347, 1377, 1396, 1297, 1279, 1244, 1281, 13...
## $ TEAM_PITCHING_HR <int> 84, 191, 137, 97, 102, 92, 122, 116, 114, 96, 86, ...
## $ TEAM_PITCHING_BB <int> 927, 689, 602, 454, 472, 443, 525, 459, 447, 441, ...
## $ TEAM_PITCHING_SO <int> 5456, 1082, 917, 928, 920, 973, 1062, 1033, 922, 8...
## $ TEAM_FIELDING_E  <int> 1011, 193, 175, 164, 138, 123, 136, 112, 127, 131,...
## $ TEAM_FIELDING_DP <int> NA, 155, 153, 156, 168, 149, 186, 136, 169, 159, 1...
```

Sample 6 rows with sample 7 columns

Code

	INDEX <int>	TARGET_WINS <int>	TEAM_BATTING_H <int>	TEAM_BATTING_2B <int>	TEAM_BATTING_3B <int>	TEAM_BATTING_HR <int>
1	1	39	1445	194	39	13
2	2	70	1339	219	22	190
3	3	86	1377	232	35	137
4	4	70	1387	209	38	96
5	5	82	1297	186	27	102
6	6	75	1279	200	36	92

6 rows | 1-7 of 18 columns

Code

3.2 Structure of data

“Dimension of Test dataset is”, 2276 X 17 with

n <int>
2276

1 row

number of observation in test data.

Sumamry of the test data shows very clearly that we have six predictors which has NA and `BATTING_HBP` and `BASERUN_CS` have the max number of NAs in the data set.

Code

```
##      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
```

Code

4 Mean and Median of the data

Code

	INDEX	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB	TEAM
	Min. : 1.0	Min. : 0.00	Min. : 891	Min. : 69.0	Min. : 0.00	Min. : 0.00	Min. : 0.0	Min. : (
	1st Qu.: 630.8	1st Qu.: 71.00	1st Qu.:1383	1st Qu.:208.0	1st Qu.: 34.00	1st Qu.: 42.00	1st Qu.:451.0	1st Qu
	Median :1270.5	Median : 82.00	Median :1454	Median :238.0	Median : 47.00	Median :102.00	Median :512.0	Media
	Mean :1268.5	Mean : 80.79	Mean :1469	Mean :241.2	Mean : 55.25	Mean : 99.61	Mean :501.6	Mean
	3rd Qu.:1915.5	3rd Qu.: 92.00	3rd Qu.:1537	3rd Qu.:273.0	3rd Qu.: 72.00	3rd Qu.:147.00	3rd Qu.:580.0	3rd Qu
	Max. :2535.0	Max. :146.00	Max. :2554	Max. :458.0	Max. :223.00	Max. :264.00	Max. :878.0	Max. :
	NA	NA	NA	NA	NA	NA	NA	NA's ::

`BATTING_HBP` is showing very close mean and median vlaue, and we suspect its due less number of datapoints. Remember we noted highest number of NA in this predictor. Apart from `FIELDING_E` we don't see any big differnce in the mean and median of the data.

4.1 Rename COlumns

Here we removing the `TEAM_` from the column name so that we can disaply it in the plots, and make it easy to read.

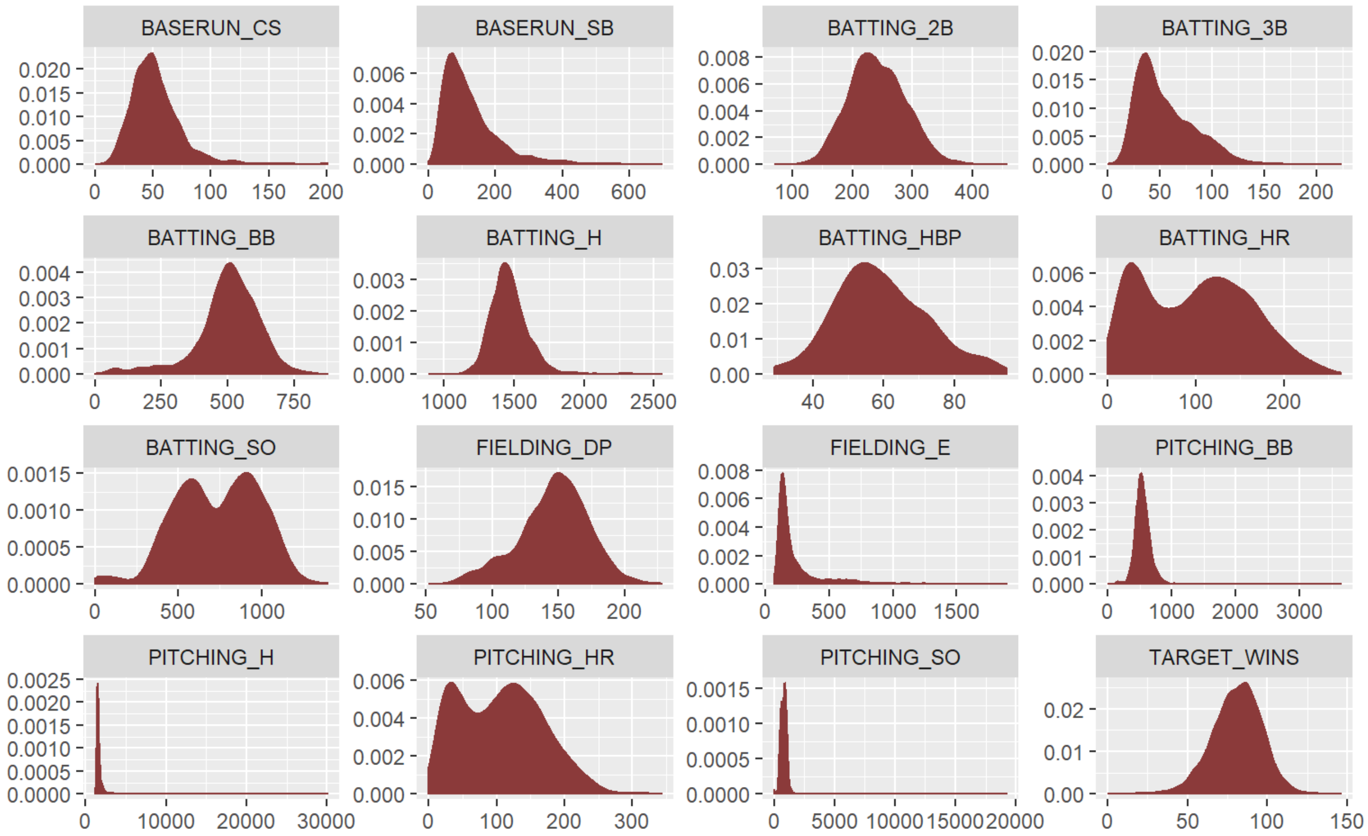
Names Before:

Code

Names After : TARGET_WINS, TEAM_BATTING_H, TEAM_BATTING_2B, TEAM_BATTING_3B, TEAM_BATTING_HR, TEAM_BATTING_BB, TEAM_BATTING_SO, TEAM_BASERUN_SB, TEAM_BASERUN_CS, TEAM_BATTING_HBP, TEAM_PITCHING_H, TEAM_PITCHING_HR, TEAM_PITCHING_BB, TEAM_PITCHING_SO, TEAM_FIELDING_E, TEAM_FIELDING_DP

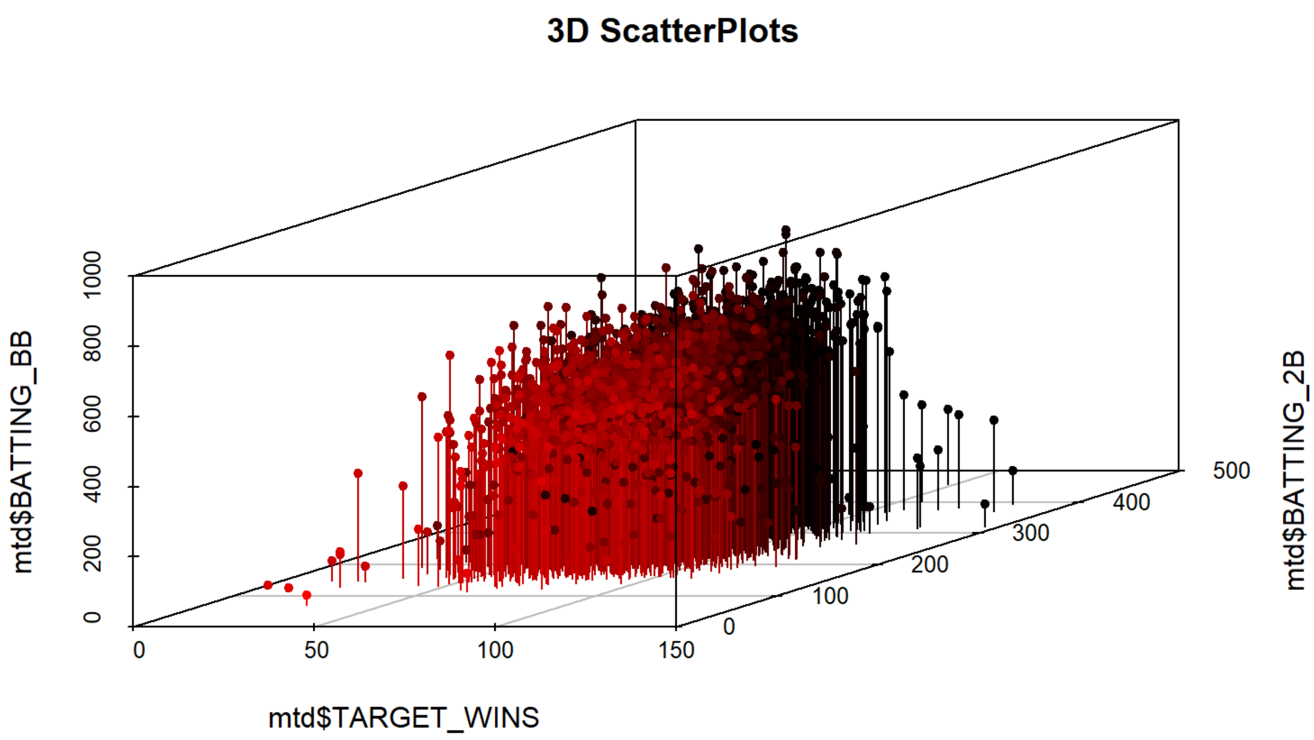
##Visualize the data

Code



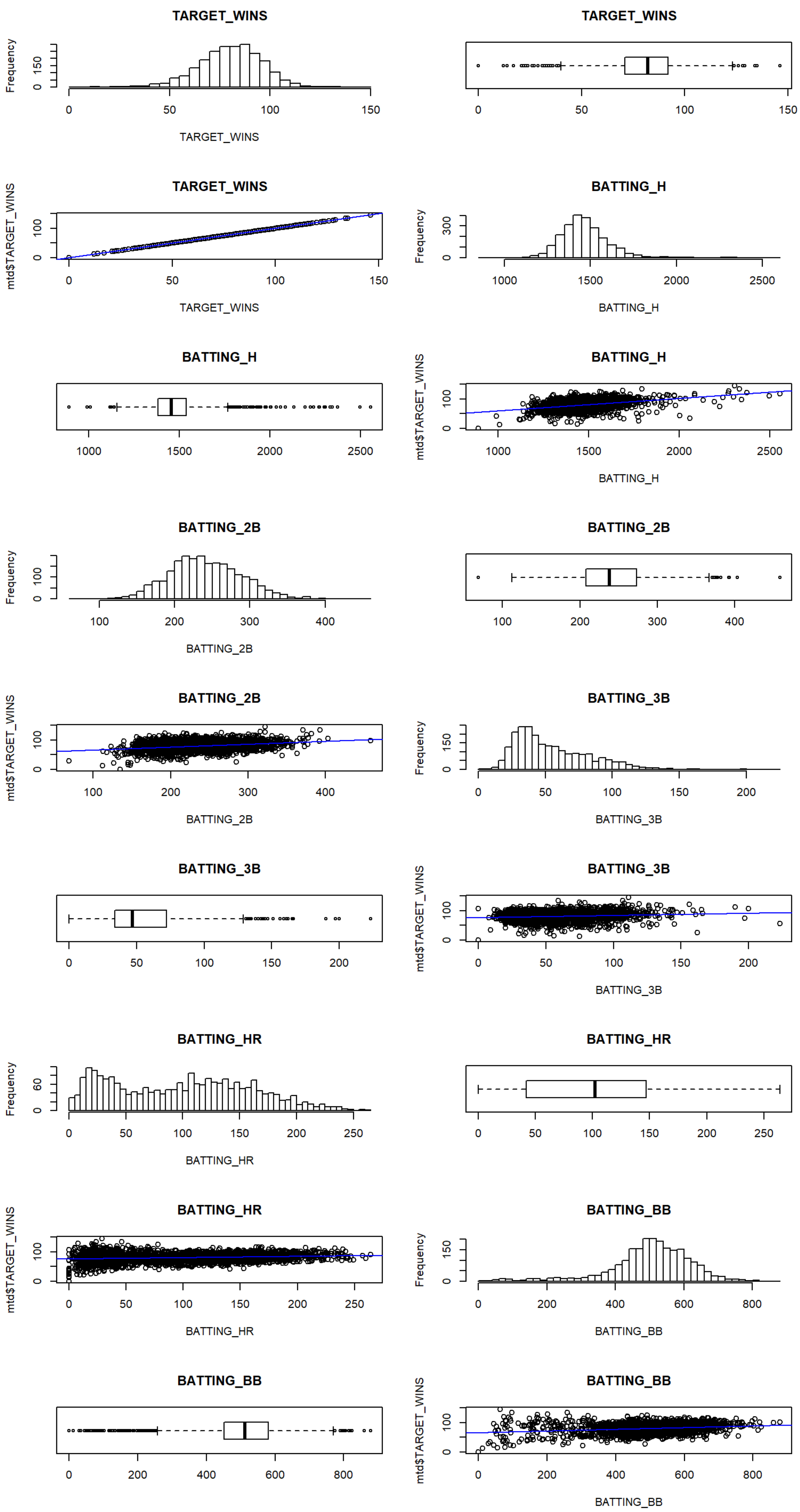
In the histogram plot above, we see that the batting, pitching home-run and batting strike-out variables are bi modal. `TARGET_WINS` and `TEAM_BATTING_2B` has most the normal distribution. `PITCHING_H` and `PITCHING_SO` have the most skewed data distribution. The skewed graphs are all rght-skewed except `BATTING_BB`.

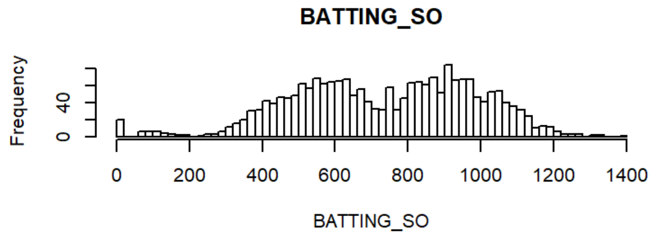
Code

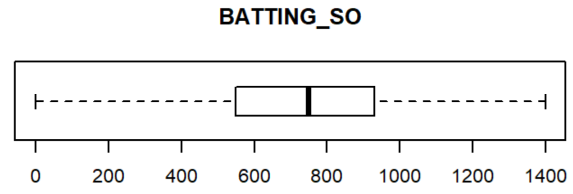


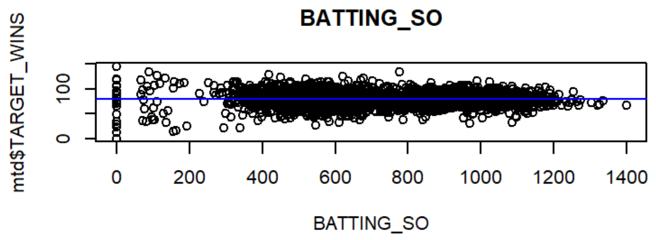
The above 3-D scatter plot, shows the data variance between the `TARGET_WINS`, `TEAM_BATTING_2B` and `TEAM_BATTING_BB` to provide a comparative view.

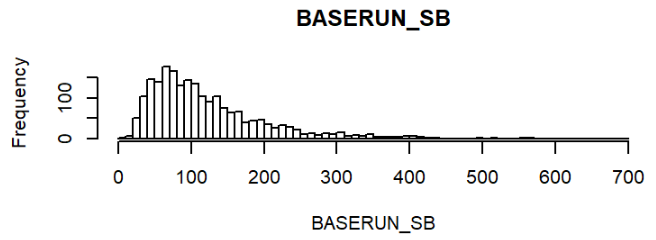
Code

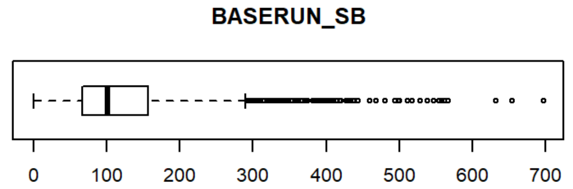

TARGET_WINS
Frequency
TARGET_WINS
TARGET_WINS
TARGET_WINS
mtd$TARGET_WINS
TARGET_WINS
BATTING_H
Frequency
BATTING_H
BATTING_H
BATTING_H
mtd$TARGET_WINS
BATTING_H
BATTING_2B
Frequency
BATTING_2B
BATTING_2B
BATTING_2B
mtd$TARGET_WINS
BATTING_2B
BATTING_3B
Frequency
BATTING_3B
BATTING_3B
BATTING_3B
mtd$TARGET_WINS
BATTING_3B
BATTING_HR
Frequency
BATTING_HR
BATTING_HR
BATTING_HR
mtd$TARGET_WINS
BATTING_HR
BATTING_BB
Frequency
BATTING_BB
BATTING_BB
BATTING_BB
mtd$TARGET_WINS
BATTING_BB

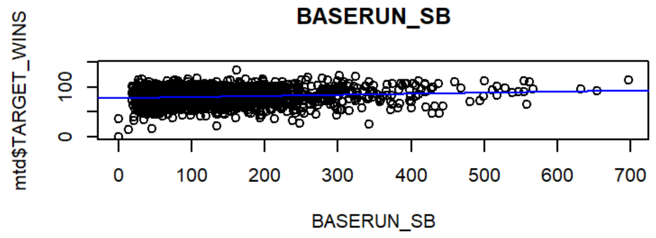

BATTING_SO
Frequency
BATTING_SO

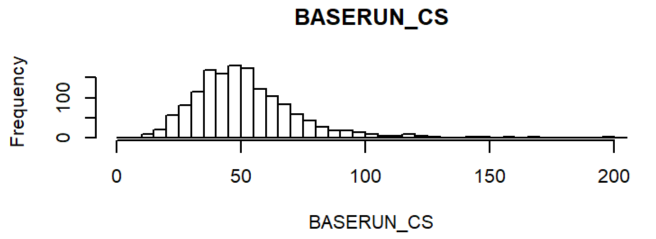

BATTING_SO

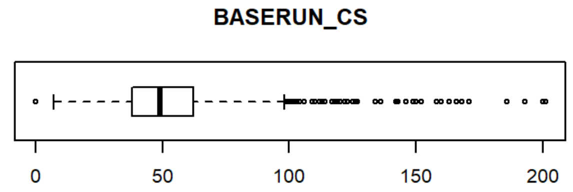

BATTING_SO
mtd$TARGET_WINS
BATTING_SO

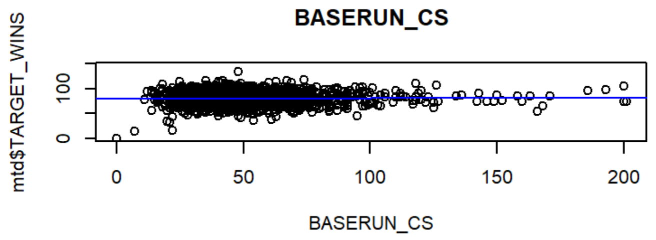

BASERUN_SB
Frequency
BASERUN_SB

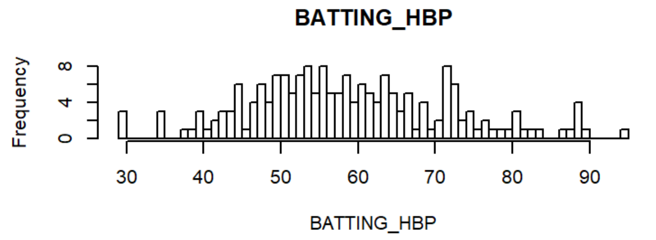

BASERUN_SB

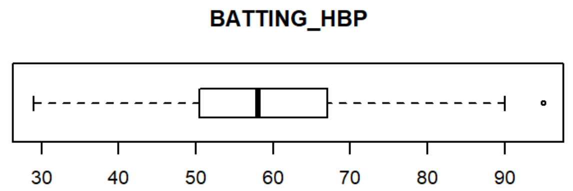

BASERUN_SB
mtd$TARGET_WINS
BASERUN_SB

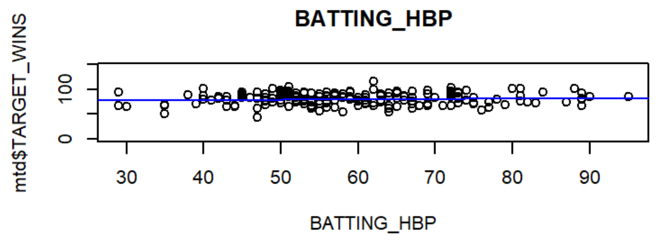

BASERUN_CS
Frequency
BASERUN_CS

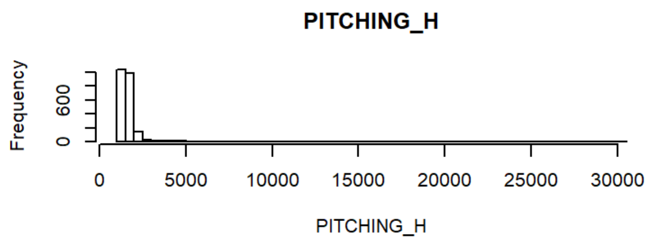

BASERUN_CS

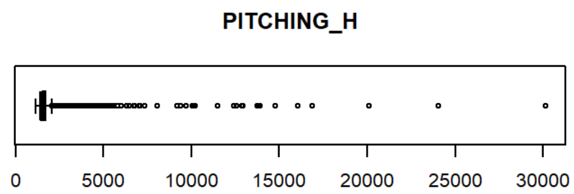

BASERUN_CS
mtd$TARGET_WINS
BASERUN_CS

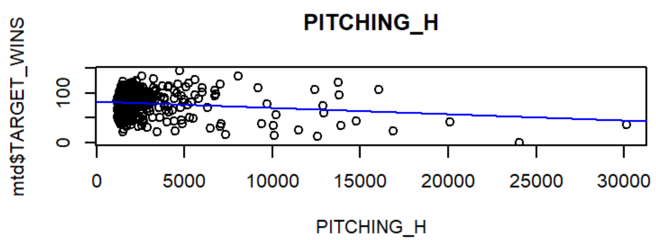

BATTING_HBP
Frequency
BATTING_HBP

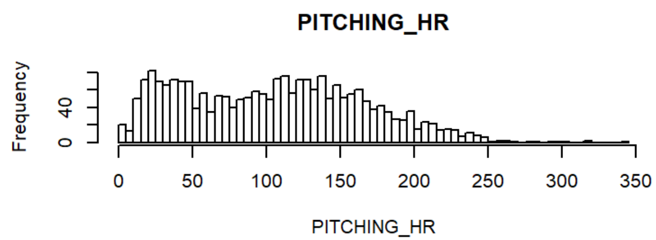

BATTING_HBP

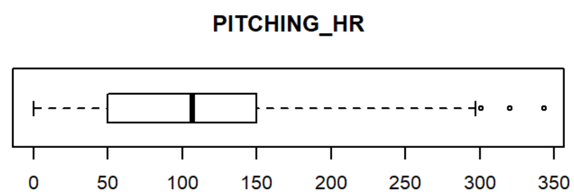

BATTING_HBP
mtd$TARGET_WINS
BATTING_HBP

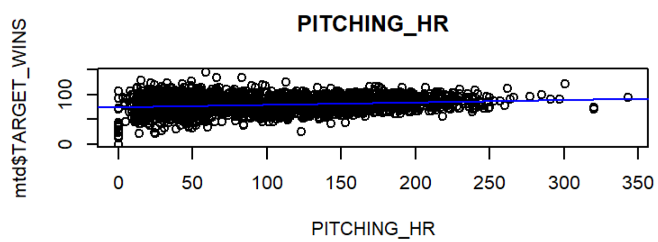

PITCHING_H
Frequency
PITCHING_H

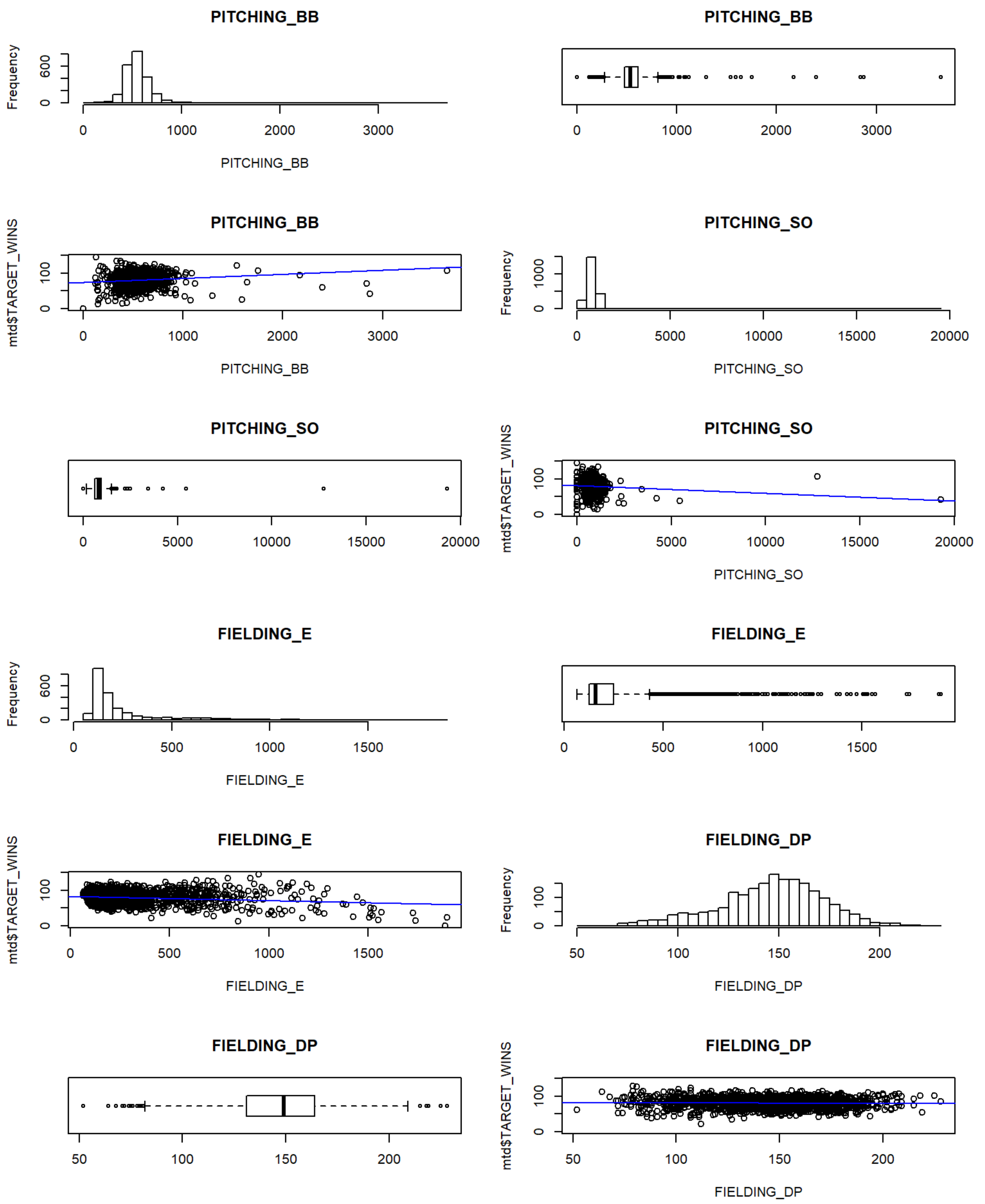

PITCHING_H


PITCHING_H
mtd$TARGET_WINS
PITCHING_H


PITCHING_HR
Frequency
PITCHING_HR


PITCHING_HR


PITCHING_HR
mtd$TARGET_WINS
PITCHING_HR



As can be seen from above histogram, boxplot and scatter plot with regression line shows the spread of the data points. More than half of the variables show skewness. A box-cox transformation may help to mitigate the skewness.

Missing or NA Values

We are trying to see how many `NA` is present in the dataset.

Code

variable	n	percent
BATTING_HBP	2085	92%
BASERUN_CS	772	34%
FIELDING_DP	286	13%
BASERUN_SB	131	5.8%
BATTING_SO	102	4.5%
PITCHING_SO	102	4.5%

The variable `BATTING_HBP` (hit by pitcher) is missing over 90% of it's data.

Zero Values

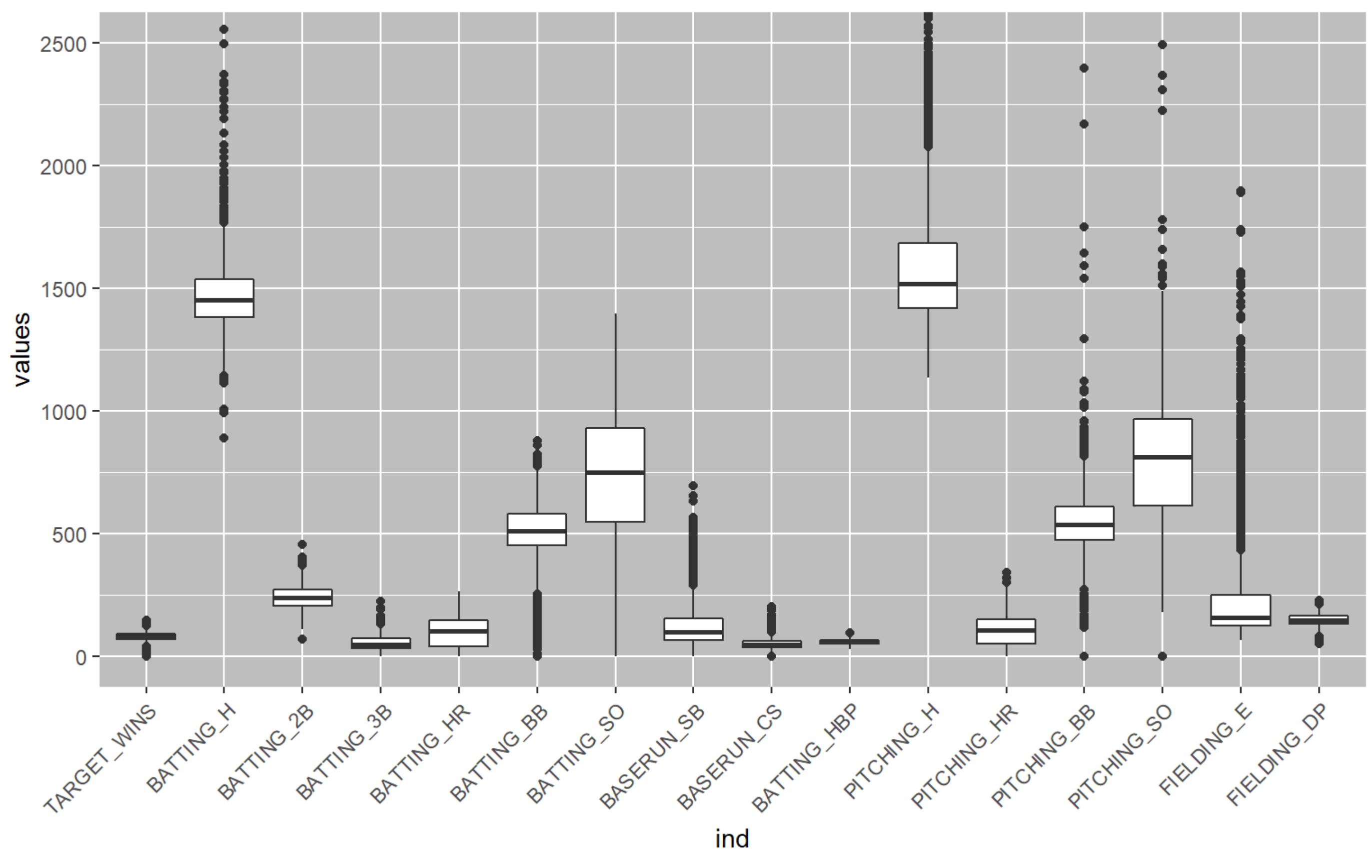
Code

variable	n	percent
BATTING_SO	20	0.9%
PITCHING_SO	20	0.9%
BATTING_HR	15	0.7%
PITCHING_HR	15	0.7%
BASERUN_SB	2	0.1%
BATTING_3B	2	0.1%
BASERUN_CS	1	0%
BATTING_BB	1	0%
PITCHING_BB	1	0%
TARGET_WINS	1	0%

As can be inferred from above, there are Very few zero values exists.

Checking for outliers

Code

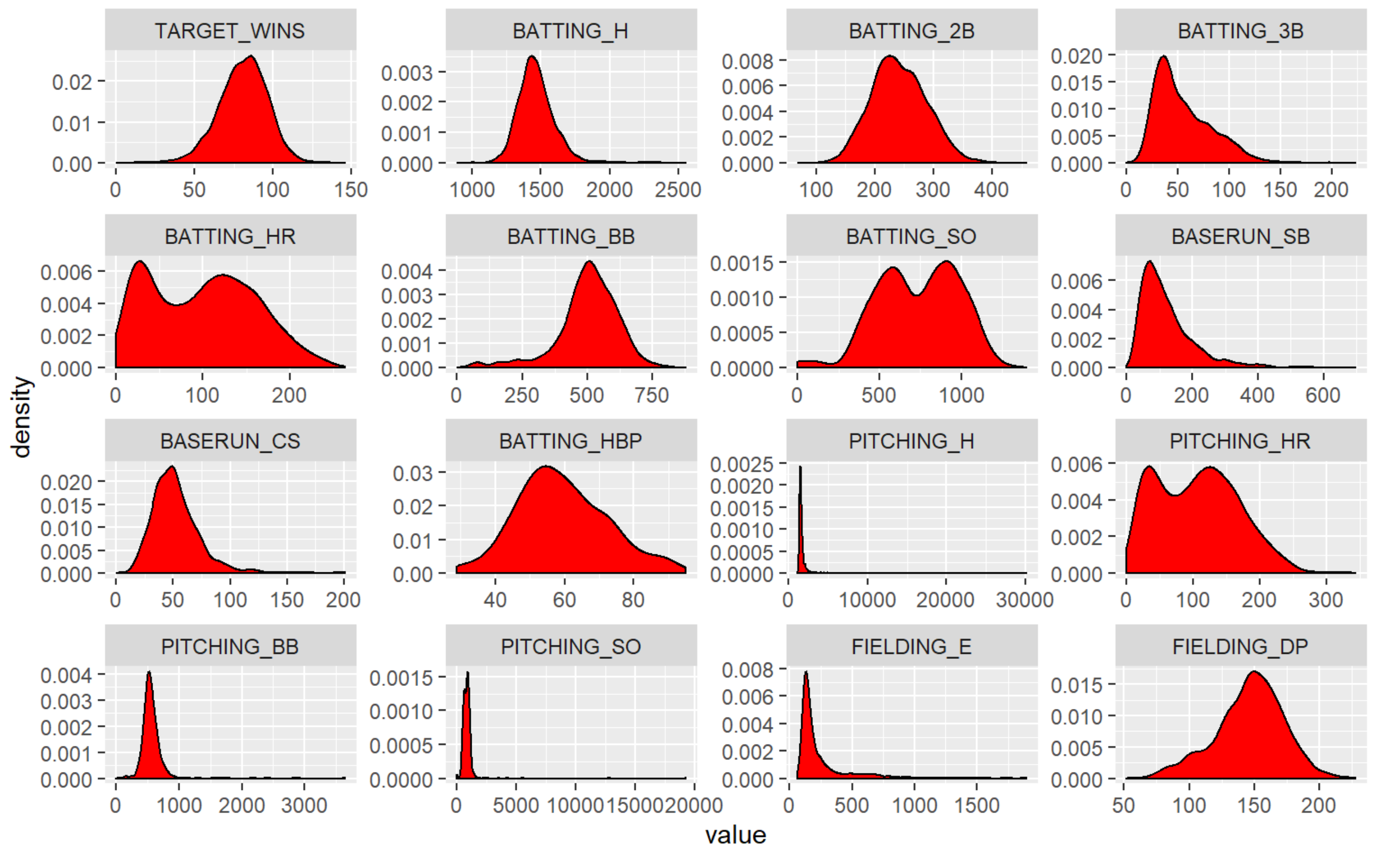


The box plots reveal that a great majority of the explanatory variables have high variances. Many of the medians and means are also not aligned which demonstrates the outliers' effects.

The variance of some of the explanatory variables greatly exceeds the variance of the response "win" variable. The dataset has many outlines with some observations that are more extreme than the 1.5 * IQR of the box plot whiskers.

Checking for skewness in the data

Code



As per above, there are several variables like `PITCHING_H`, `PITCHING_BB`, `PITCHING_SO` and `FIELDING_E` are extremely skewed as there are many outliers.

Finding correlations: Below shows the comparative correlations between the 16 variables as it shows the correlation coefficients and thus find correlated variables. Whichever adhere to a fitted straight red line well, ie. change in synch with each other. If the points lie close to the line but the line is curved, it's good nonlinear association and one can still be defined by other. Each individual plot shows the relationship between the variable in the horizontal vs the vertical of the grid. Each individual plot shows the relationship between the variable in the horizontal vs the vertical of the grid, whereas the diagonal is showing a histogram of each variable.

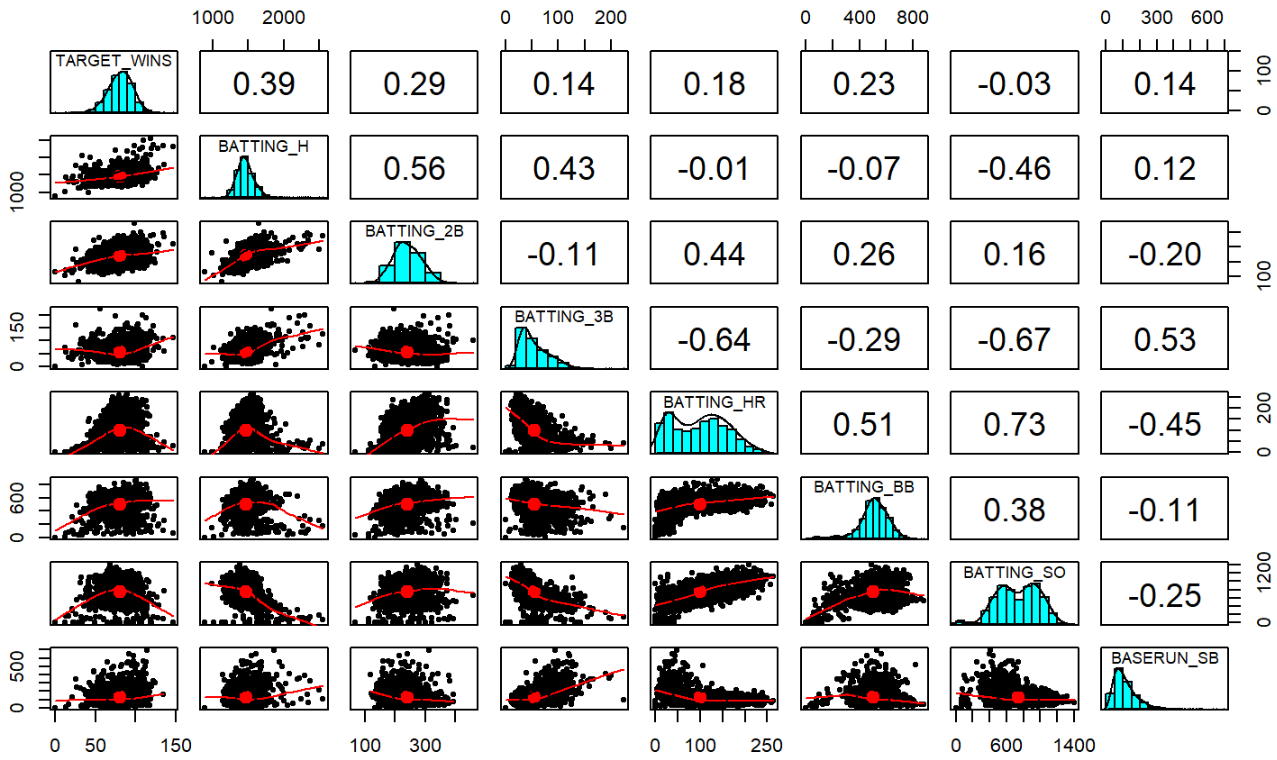
Code

Show 10 entries Search:

	TARGET_WINS	BATTING_H	BATTING_2B	BATTING_3B	BATTING_HR	
TARGET_WINS	1	0.469946650195572	0.312983997280004	-0.124345862964454	0.422416834117253	0.4
BATTING_H	0.469946650195572	1	0.561772855536591	0.213918834444827	0.396275927326416	0.19
BATTING_2B	0.312983997280004	0.561772855536591	1	0.0420344070268067	0.250990454027817	0.19
BATTING_3B	-0.124345862964454	0.213918834444827	0.0420344070268067	1	-0.21879927259709	-0.20
BATTING_HR	0.422416834117253	0.396275927326416	0.250990454027817	-0.21879927259709	1	0.4
BATTING_BB	0.46868792650956	0.197352343885384	0.197492562030794	-0.205843921730807	0.45638161304111	
BATTING_SO	-0.228892727179823	-0.341743283600818	-0.0641512258250527	-0.192918409979567	0.210454439156417	0.21
BASERUN_SB	0.014836392442652	0.0716749520962244	-0.187682787958738	0.169460861525657	-0.190218931518434	-0.088
BASERUN_CS	-0.178755979245533	-0.093775445091233	-0.204138837073558	0.232139777238505	-0.275798375425212	-0.20
BATTING_HBP	0.0735042423086367	-0.029112175684042	0.046084753143314	-0.174247153838812	0.10618116006506	0.047

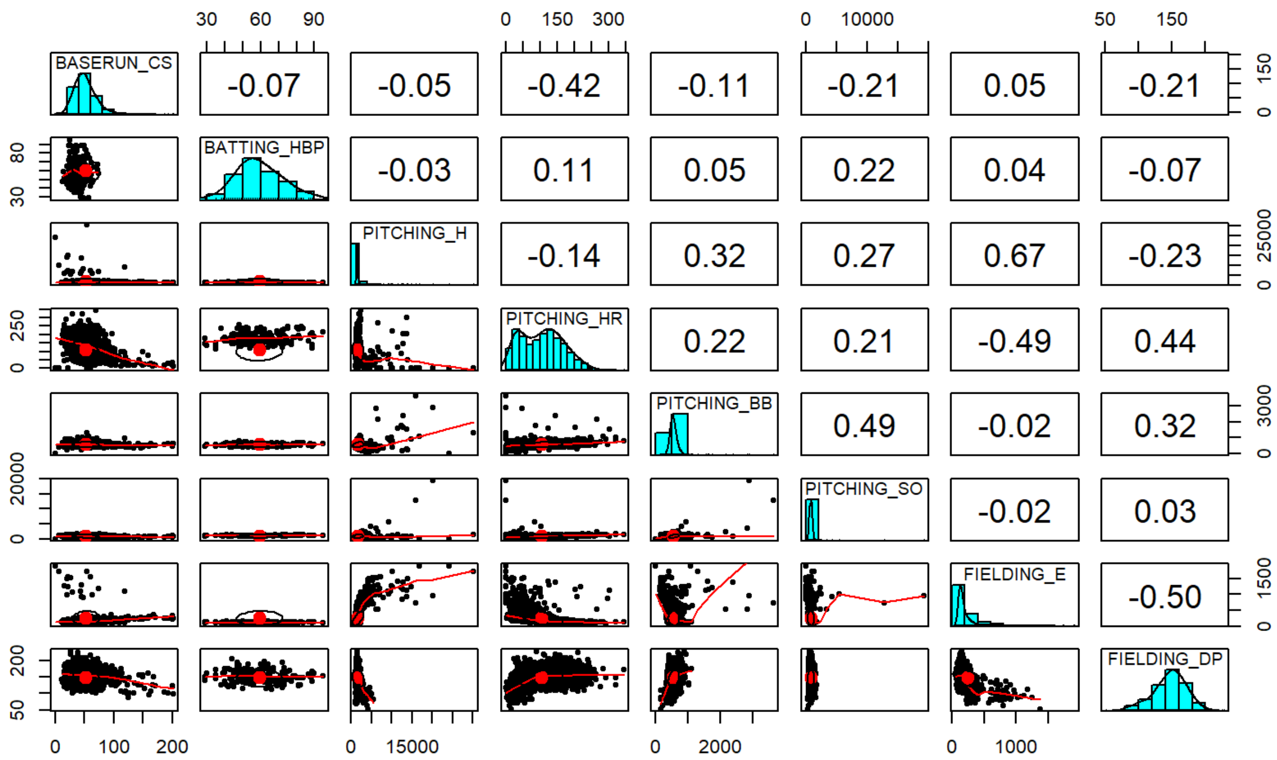
Showing 1 to 10 of 16 entries Previous 1 2 Next

Code



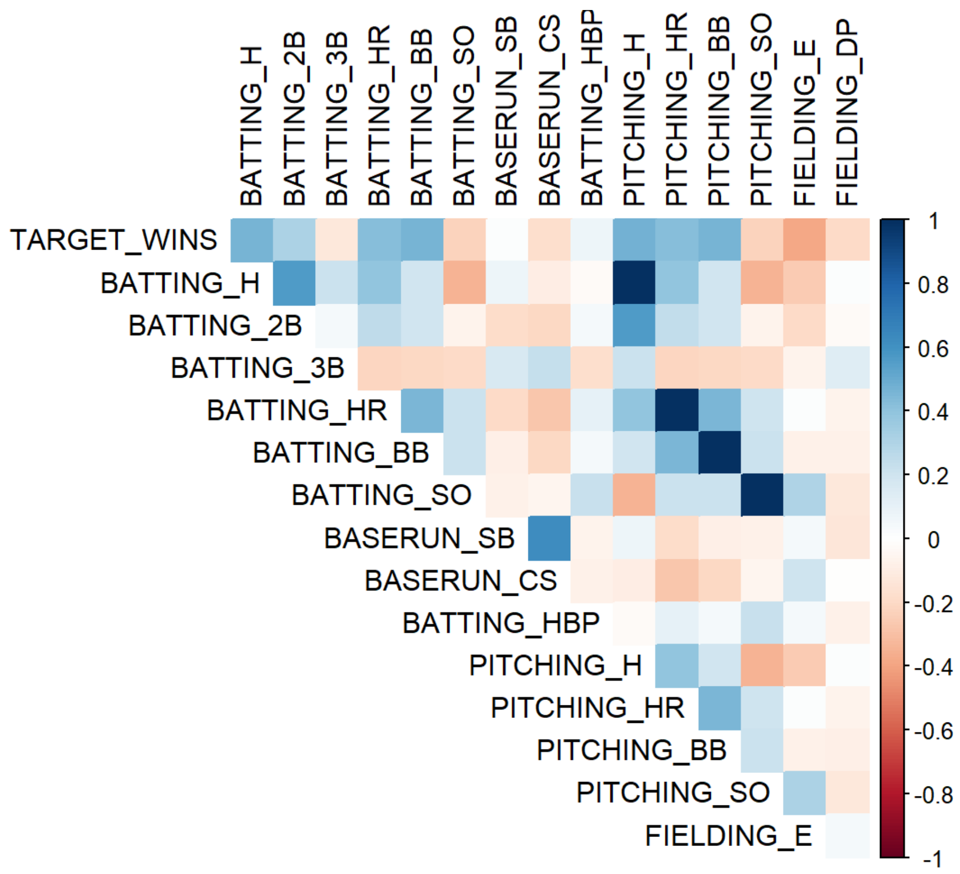
As can be seen from above, `TARGET_WINS` vs `BATTING_2B` is continuous and hence correlated and so is `BATTING_BB` and `BATTING_HR`.

Code



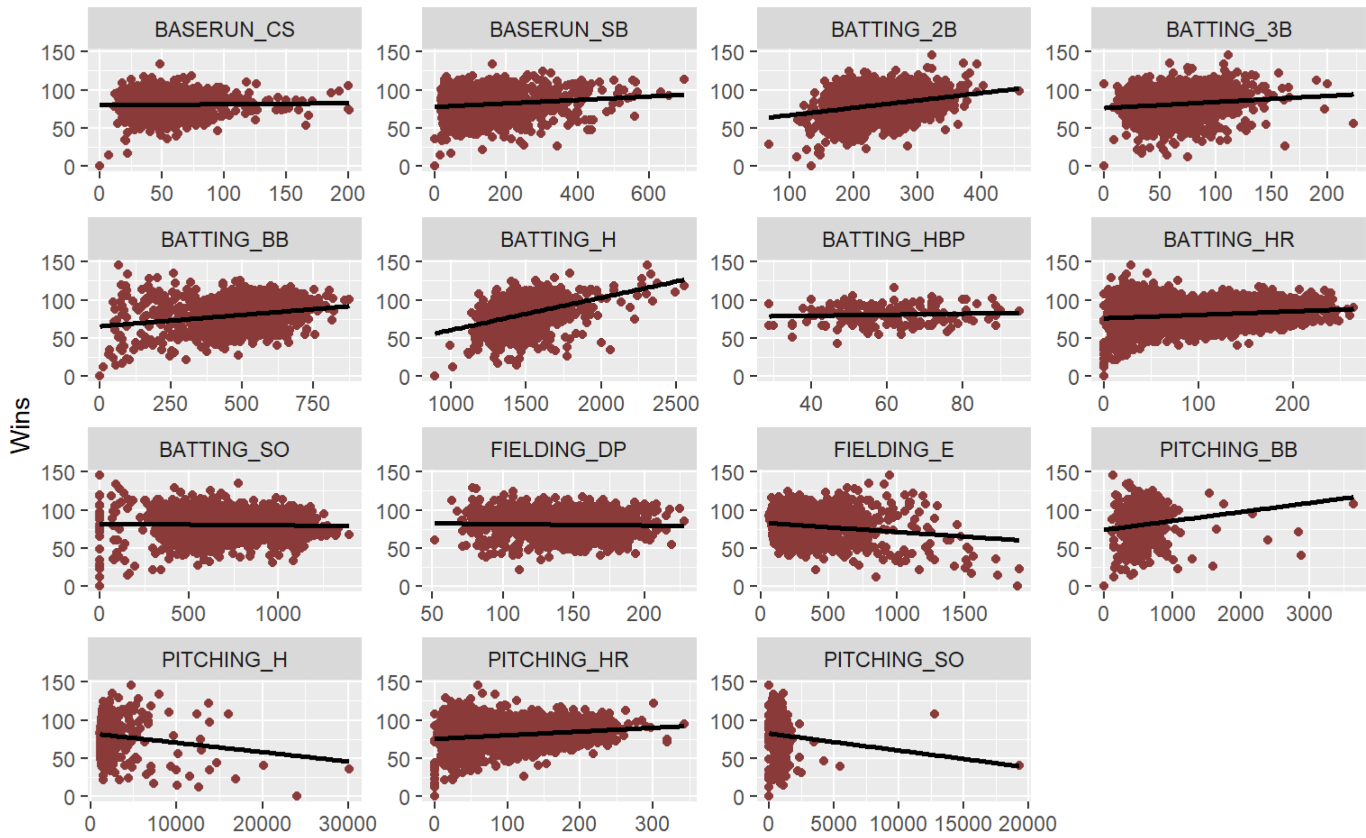
As can be seen from above, `BASERUN_CS` vs `BATTING_HBP` is continuous and hence correlated whereas `PITCHING_SO` and `FIELDING_E` is not correlated at all.

Code



Also, there are some negatively correlated variables. According to the correlation heatmap, the values that correspond most positively are BATTING_H, BATTING_2B, BATTING_HR, BATTING_BB, PITCHING_H, PITCHING_HR, and PITCHING_BB.

Code



Above shows how the data is distributed when compared to the linear regression. Clearly, `PITCHING_H` and `PITCHING_SO` are highly heteroscedastic. Comparatively, `BATTING_HBP` is most homoscedastic.

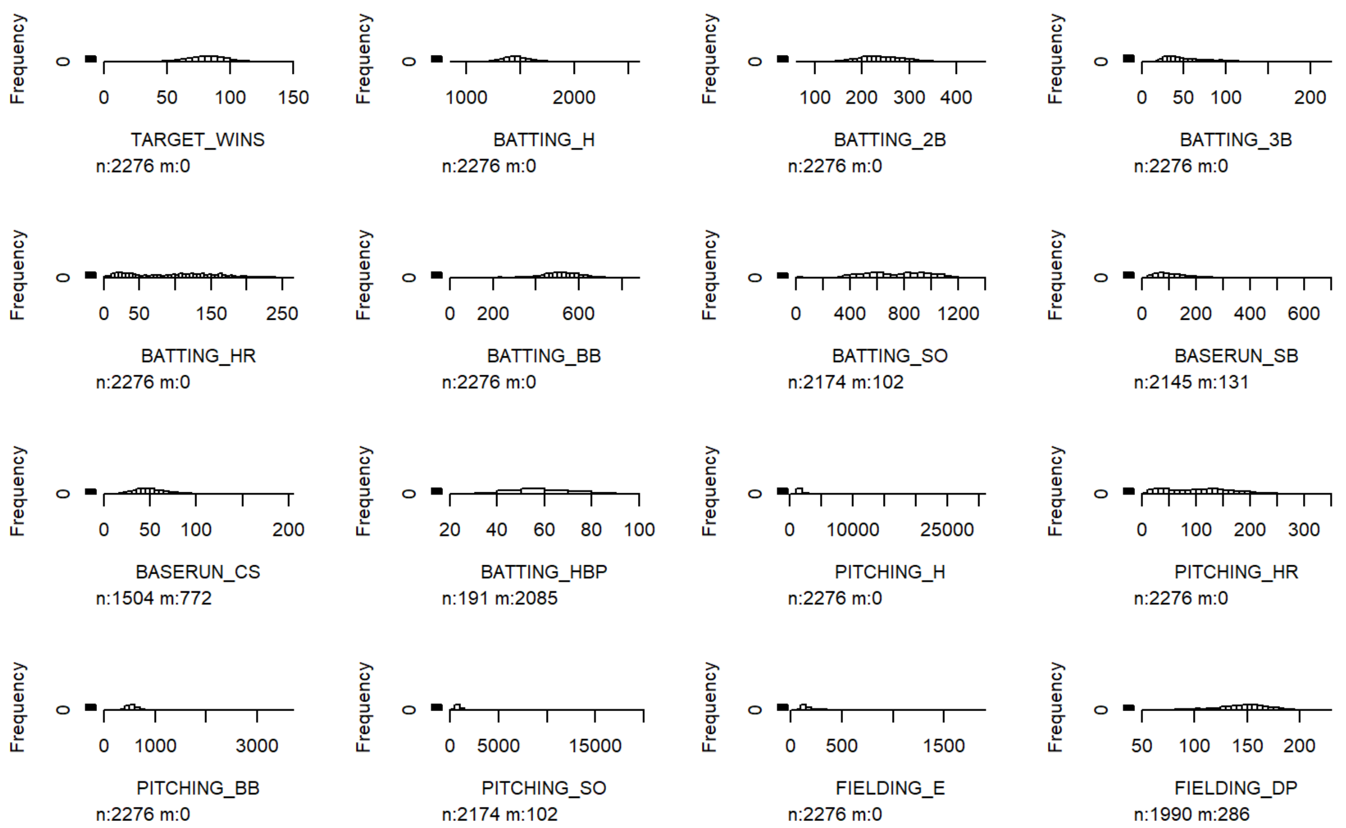
Code

```
##              TARGET_WINS   BATTING_H
## TARGET_WINS  1.00000000  0.46994665
## BATTING_H    0.46994665  1.00000000
## BATTING_2B   0.31298400  0.56177286
## BATTING_3B  -0.12434586  0.21391883
## BATTING_HR   0.42241683  0.39627593
## BATTING_BB   0.46868793  0.19735234
## BATTING_SO  -0.22889273 -0.34174328
## BASERUN_SB   0.01483639  0.07167495
## BASERUN_CS  -0.17875598 -0.09377545
## BATTING_HBP  0.07350424 -0.02911218
## PITCHING_H   0.47123431  0.99919269
## PITCHING_HR  0.42246683  0.39495630
## PITCHING_BB  0.46839882  0.19529071
## PITCHING_SO -0.22936481 -0.34445001
## FIELDING_E  -0.38668800 -0.25381638
## FIELDING_DP -0.19586601  0.01776946
```

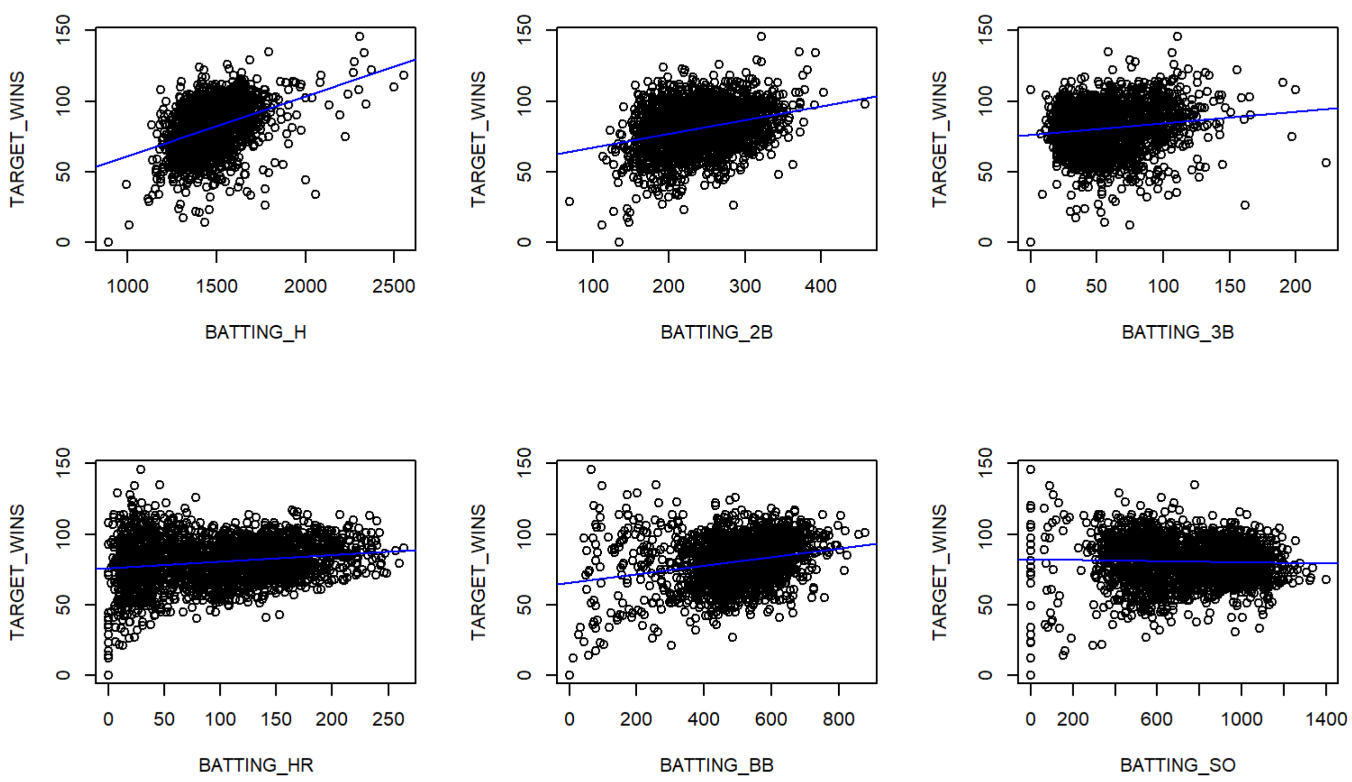
Above shows the correlation coefficient of each variable compared to `TARGET_WINS` and `BATTING_H`.

Histogram of Variables

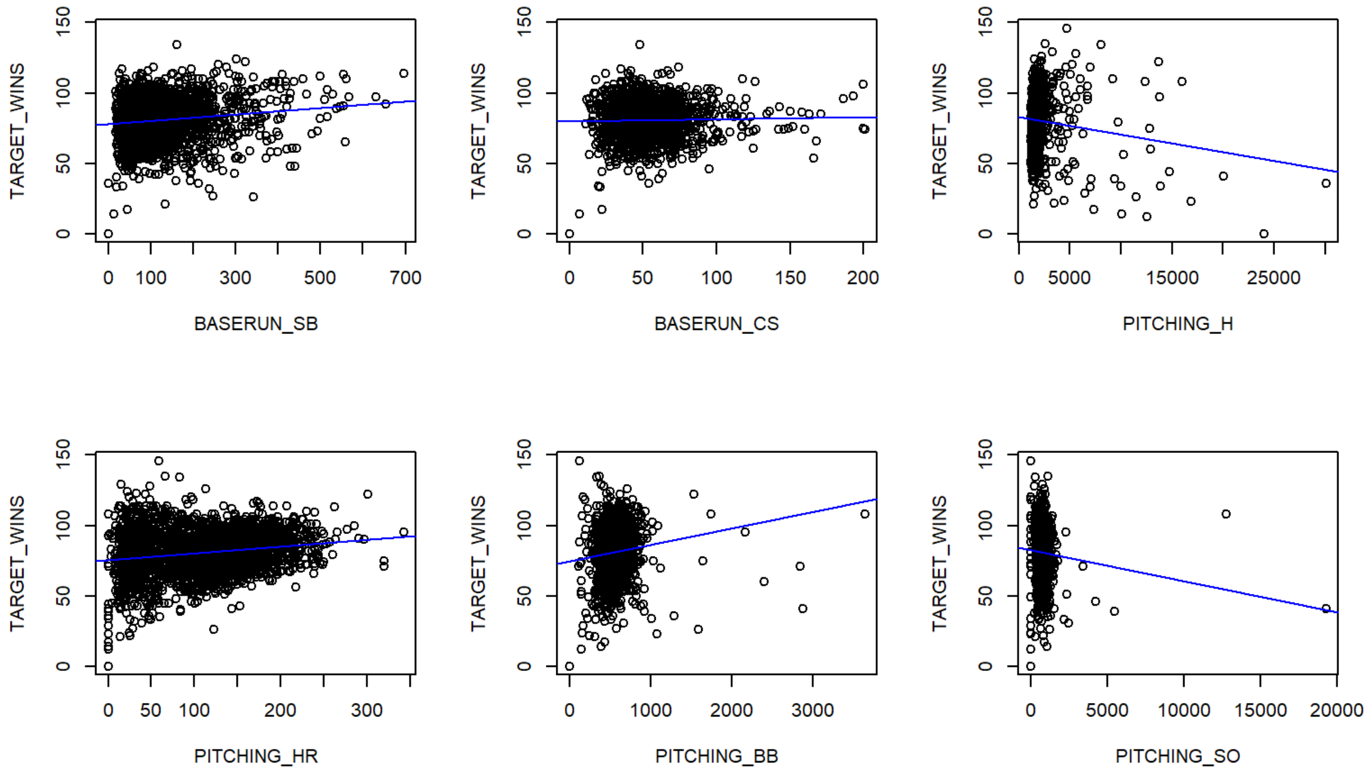
Code



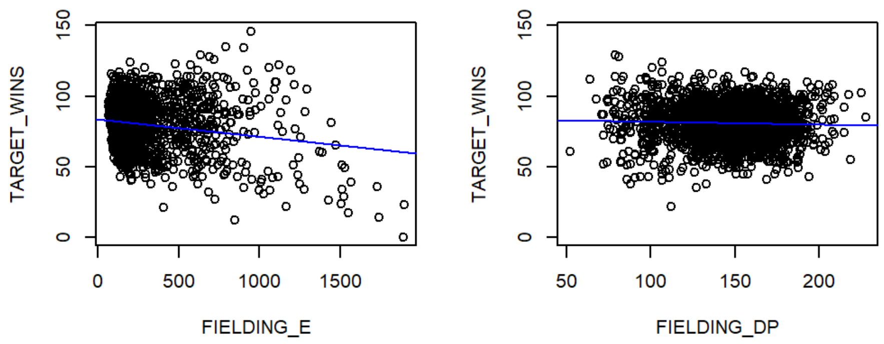
Code



Code



Code

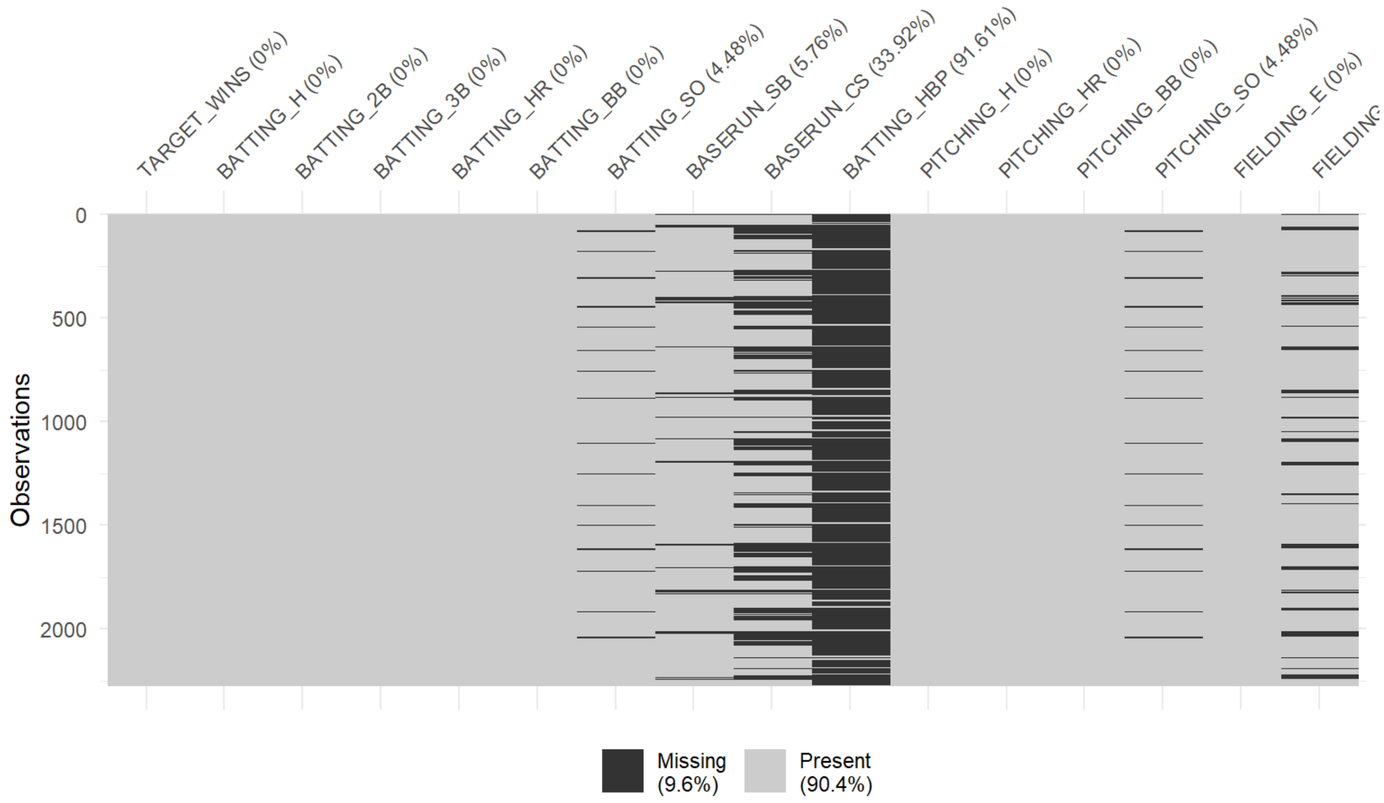


This shows very few variables are normally distributed.

4.1.1 Missing value by Graph

Here will see how much of data is missing in each predictors.

Code



Here from the plots we can see outliers in PITCHING_H,PITCHING_BB and PITCHING_SO

Also, since BATTING_H is a combination of BATTING_2B, BATTING_3B, BATTING_HR (and also includes batted singles), we will create a new variable BATTING_1B equaling BATTING_H - BATTING_2B - BATTING_3B - BATTING_HR and after creating this we will remove BATTING_H

Initial Observations

- Response variable (TARGET_WINS) looks to be normally distributed which means there are good teams, bad teams as well as average teams.
- There are also quite a few variables with missing values. We may need to deal with these in order to have the largest data set possible for modeling.
- A couple variables are bimodal (TEAM_BATTING_HR, TEAM_BATTING_SO, TEAM_PITCHING_HR). This may be a challenge as some of them are missing values and that may be a challenge in filling in missing values.
- Some variables are right skewed (TEAM_BASERUN_CS, TEAM_BASERUN_SB, etc.). This might support the good team theory. It may also introduce non-normally distributed residuals in the model. We shall see.
- Dataset covers a wide time period spanning across multiple "eras" of baseball.

5 DATA PREPARATION

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations. a. Fix missing values (maybe with a Mean or Median value) b. Create flags to suggest if a variable was missing c. Transform data by putting it into buckets d. Mathematical transforms such as log or square root (or use Box-Cox) e. Combine variables (such as ratios or adding or multiplying) to create new variables

Fixing Missing/Zero Values - Remove the invalid data and prep it for imputation. - We could "discard" the TEAM_BATTING_HBP,due to the high percentage of missing data; particularly, replacing it by "ZERO" should not be advisable since the minimum value recorded is 29 and replacing it with a median value would not be much helpful due to high percentage of missing values. We decided not to consider this variable for our study. - A typical professional league baseball game has 9 innings (extra innings come to play in the event of a tie) in length, and in each inning one can only pitch 3 strikeouts. There have been a maximum of 27 potential strikeouts upto a maximum of by 162 games for each of the 30 teams in the American League (AL) and National League (NL), played over approximately six months in Major League Baseball (MLB) season. Therefore having more than 4374 strikeouts (9x3x162) is not possible. Incidentally, the maximum strikeouts in any baseball season has been 513 by Matt Kilroy in the year 1886 as part of Baltimore Orioles within American Association League,

Code

Imputing the values using KNN

Code

6 BUILD MODELS

Using the training data set, build at least three different multiple linear regression models, using different variables (or the same variables with different transformations). Since we have not yet covered automated variable selection methods, you should select the variables manually (unless you previously learned Forward or Stepwise selection, etc.). Since you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done.

Discuss the coefficients in the models, do they make sense? For example, if a team hits a lot of Home Runs, it would be reasonably expected that such a team would win more games. However, if the coefficient is negative (suggesting that the team would lose more games), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

Code

6.1 Model 1 : Kitchen Sink Model/Backward Elimination

With all variables to determine the base model provided. This would allow to see which variables are significant in our dataset, and allows to make other models based on that.

Code

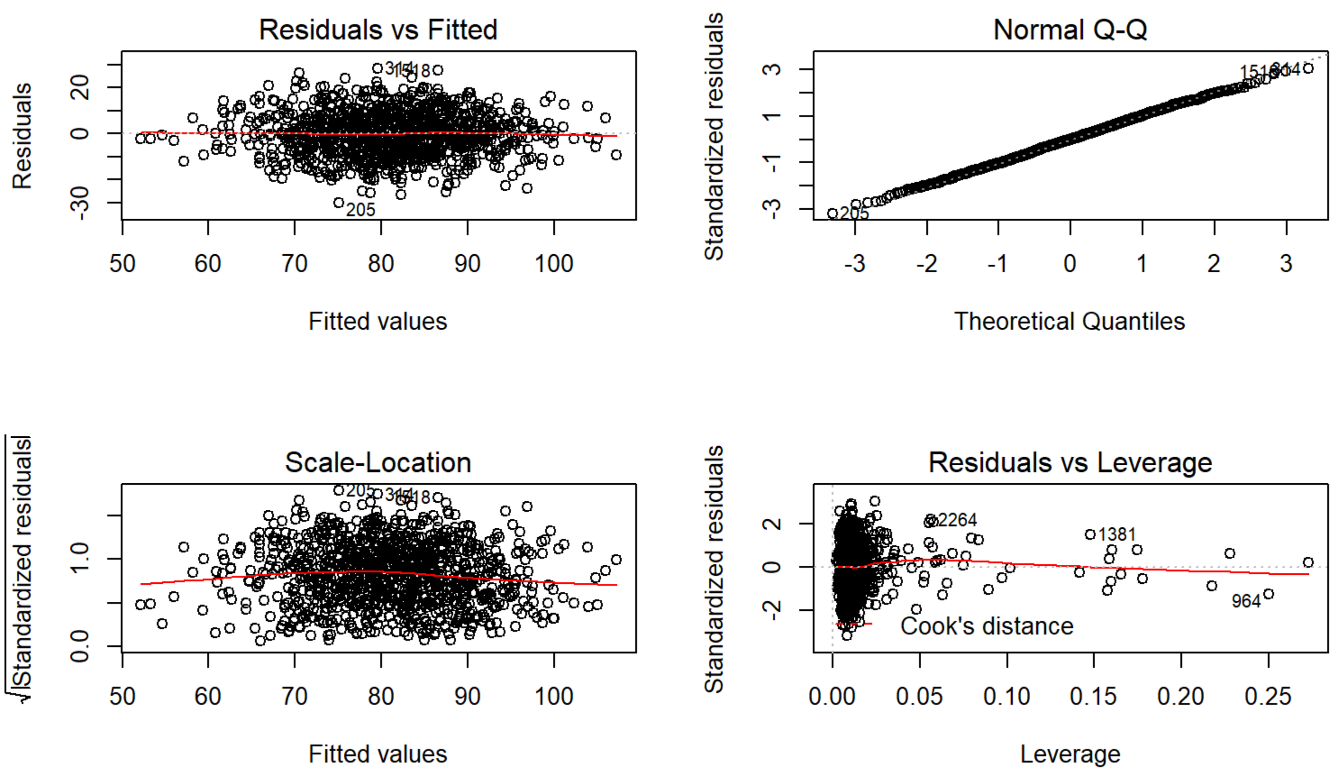
```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = moneyball_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.0724  -6.5828  -0.1407   6.4786  28.3847
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 58.53113    7.79100   7.513 1.25e-13 ***
## BATTING_H    0.01653    0.02346   0.704 0.481330
## BATTING_2B  -0.07540    0.01100  -6.854 1.23e-11 ***
## BATTING_3B   0.17325    0.02552   6.789 1.90e-11 ***
## BATTING_HR   0.13176    0.09460   1.393 0.163944
## BATTING_BB   0.02796    0.05440   0.514 0.607397
## BATTING_SO   0.01254    0.02769   0.453 0.650670
## BASERUN_SB   0.03694    0.01026   3.600 0.000334 ***
## BASERUN_CS   0.05115    0.02196   2.329 0.020032 *
## PITCHING_H   0.01747    0.02210   0.791 0.429325
## PITCHING_HR -0.02926    0.09070  -0.323 0.747075
## PITCHING_BB  0.01110    0.05237   0.212 0.832216
## PITCHING_SO -0.03241    0.02645  -1.225 0.220789
## FIELDING_E  -0.16207    0.01230 -13.176  < 2e-16 ***
## FIELDING_DP -0.10625    0.01545  -6.875 1.07e-11 ***
## BATTING_1B        NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.469 on 1037 degrees of freedom
##   (543 observations deleted due to missingness)
## Multiple R-squared:  0.4421, Adjusted R-squared:  0.4346
## F-statistic:  58.7 on 14 and 1037 DF,  p-value: < 2.2e-16
```

Code

It does a fairly good job predicting, but there are a lot of variables that are not statistically significant. We see the that P-value is less than .05 which makes it one of the possible model but not all the coefficients of the `model1` are significant.

6.1.1 PLOT

Code



6.2 Model 2 : Simple Model

With only the significant variables: Pick variables that had high correlations and include the pitching variables

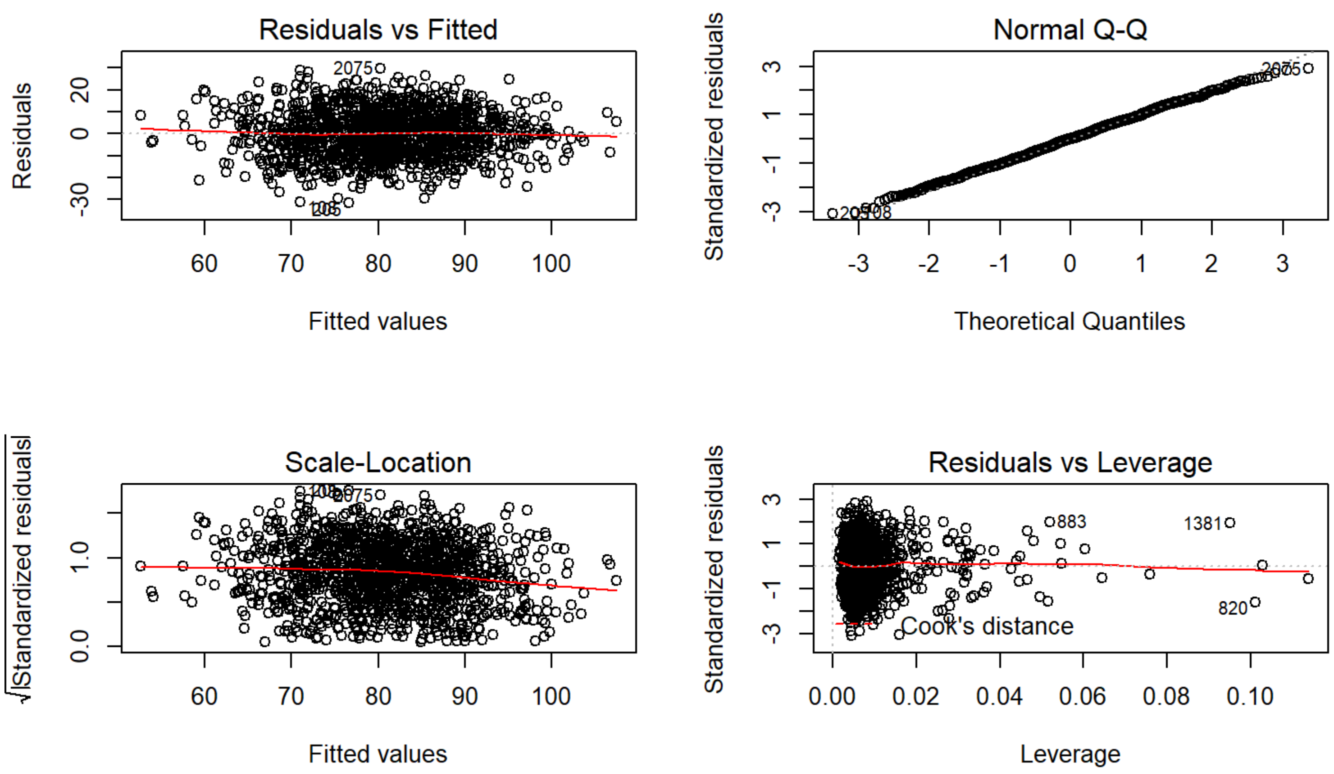
Code

```
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_H + BATTING_3B + BATTING_HR +
##     BATTING_BB + BATTING_SO + BASERUN_SB + PITCHING_SO + PITCHING_H +
##     PITCHING_SO + FIELDING_E + FIELDING_DP, data = moneyball_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.633  -7.407   0.103   7.218  29.771
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 73.346701   6.624503  11.072  < 2e-16 ***
## BATTING_H   -0.036127   0.012857  -2.810 0.005032 **
## BATTING_3B   0.201222   0.022342   9.007  < 2e-16 ***
## BATTING_HR   0.114499   0.010869  10.535  < 2e-16 ***
## BATTING_BB   0.032347   0.003796   8.522  < 2e-16 ***
## BATTING_SO   0.048172   0.020693   2.328 0.020072 *
## BASERUN_SB   0.074635   0.006672  11.186  < 2e-16 ***
## PITCHING_SO -0.071270   0.019581  -3.640 0.000284 ***
## PITCHING_H   0.043819   0.011707   3.743 0.000190 ***
## FIELDING_E  -0.111738   0.008436 -13.245  < 2e-16 ***
## FIELDING_DP -0.105429   0.014630  -7.206 9.77e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.29 on 1286 degrees of freedom
##   (298 observations deleted due to missingness)
## Multiple R-squared:  0.3949, Adjusted R-squared:  0.3902
## F-statistic: 83.92 on 10 and 1286 DF,  p-value: < 2.2e-16
```

Code

6.2.1 PLOT

Code



6.3 Model 3 : Higher Order Stepwise Regression

Only taking the variable from the Model1 that are really significant.

Code

```
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_2B + BATTING_3B + BASERUN_SB +
##     BASERUN_CS + FIELDING_E + FIELDING_DP, data = moneyball_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.0056  -7.9628  -0.3434   8.0241  30.3356
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 93.226932   4.171175  22.350   <2e-16 ***
## BATTING_2B   0.019018   0.008810   2.159   0.0311 *
## BATTING_3B   0.273238   0.025450  10.736   <2e-16 ***
## BASERUN_SB   0.018523   0.011820   1.567   0.1174
## BASERUN_CS   0.007483   0.025892   0.289   0.7726
## FIELDING_E  -0.169187   0.013894 -12.177   <2e-16 ***
## FIELDING_DP -0.043599   0.018145  -2.403   0.0164 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.44 on 1045 degrees of freedom
##   (543 observations deleted due to missingness)
## Multiple R-squared:  0.1794, Adjusted R-squared:  0.1747
## F-statistic: 38.08 on 6 and 1045 DF,  p-value: < 2.2e-16
```

Code

```
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_3B + FIELDING_E + BATTING_2B +
##     FIELDING_DP, data = moneyball_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.154  -9.095   0.359   8.972  47.276
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 73.11824    3.17547  23.026  < 2e-16 ***
## BATTING_3B   0.15080    0.01793   8.411  < 2e-16 ***
## FIELDING_E  -0.02936    0.00371  -7.913 5.08e-15 ***
## BATTING_2B   0.06870    0.00816   8.418  < 2e-16 ***
## FIELDING_DP -0.07547    0.01579  -4.780 1.94e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.17 on 1396 degrees of freedom
##   (194 observations deleted due to missingness)
## Multiple R-squared:  0.1159, Adjusted R-squared:  0.1134
## F-statistic: 45.75 on 4 and 1396 DF,  p-value: < 2.2e-16
```

Code

Further reducing the variables(TEAM_PITCHING_SO and TEAM_BATTING_SO are having high correlation, TEAM_BATTING_H and TEAM_PITCHING_H are also having high correlation, TEAM_BATTING_SO and TEAM_PITCHING_SO are also having high correlation):

Code

```
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_1B + BATTING_2B + BATTING_3B +
##     BATTING_HR + BATTING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS +
##     PITCHING_H + PITCHING_HR + PITCHING_BB + PITCHING_SO + FIELDING_E +
##     FIELDING_DP, data = moneyball_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.0724  -6.5828  -0.1407   6.4786  28.3847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 58.53113    7.79100   7.513 1.25e-13 ***
## BATTING_1B   0.01653    0.02346   0.704 0.481330
## BATTING_2B  -0.05888    0.02461  -2.392 0.016923 *
## BATTING_3B   0.18978    0.03303   5.746 1.20e-08 ***
## BATTING_HR   0.14829    0.10060   1.474 0.140776
## BATTING_BB   0.02796    0.05440   0.514 0.607397
## BATTING_SO   0.01254    0.02769   0.453 0.650670
## BASERUN_SB   0.03694    0.01026   3.600 0.000334 ***
## BASERUN_CS   0.05115    0.02196   2.329 0.020032 *
## PITCHING_H   0.01747    0.02210   0.791 0.429325
## PITCHING_HR -0.02926    0.09070  -0.323 0.747075
## PITCHING_BB  0.01110    0.05237   0.212 0.832216
## PITCHING_SO -0.03241    0.02645  -1.225 0.220789
## FIELDING_E  -0.16207    0.01230 -13.176  < 2e-16 ***
## FIELDING_DP -0.10625    0.01545  -6.875 1.07e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.469 on 1037 degrees of freedom
##   (543 observations deleted due to missingness)
## Multiple R-squared:  0.4421, Adjusted R-squared:  0.4346
## F-statistic:  58.7 on 14 and 1037 DF,  p-value: < 2.2e-16
```

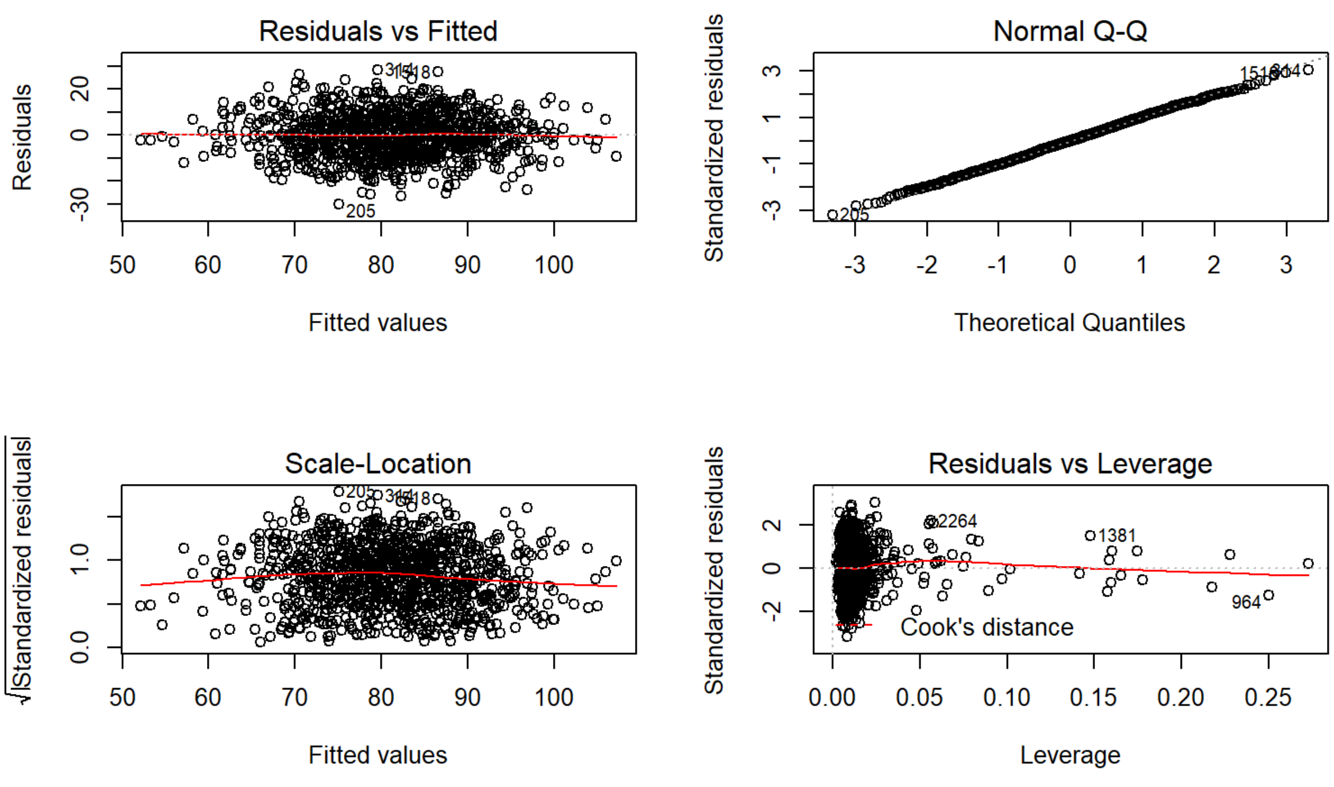
Code

```
##
## Call:
## lm(formula = poly_call[2], data = moneyball_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.0741  -6.5189  -0.0304   6.5548  28.5287
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.226582   7.718003   7.674 3.83e-14 ***
## BATTING_1B   0.021961   0.006883   3.191 0.001462 **
## BATTING_2B  -0.052339   0.008634  -6.062 1.88e-09 ***
## BATTING_3B   0.195353   0.024739   7.897 7.25e-15 ***
## BATTING_HR   0.123437   0.009440  13.077  < 2e-16 ***
## BATTING_BB   0.039462   0.003927  10.048  < 2e-16 ***
## BASERUN_SB   0.036916   0.010210   3.616 0.000314 ***
## BASERUN_CS   0.051264   0.021908   2.340 0.019475 *
## PITCHING_H   0.011846   0.002851   4.155 3.52e-05 ***
## PITCHING_SO -0.020636   0.002747  -7.513 1.25e-13 ***
## FIELDING_E  -0.162363   0.012228 -13.278  < 2e-16 ***
## FIELDING_DP -0.106435   0.015427  -6.899 9.07e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.458 on 1040 degrees of freedom
##   (543 observations deleted due to missingness)
## Multiple R-squared:  0.4418, Adjusted R-squared:  0.4359
## F-statistic: 74.83 on 11 and 1040 DF,  p-value: < 2.2e-16
```

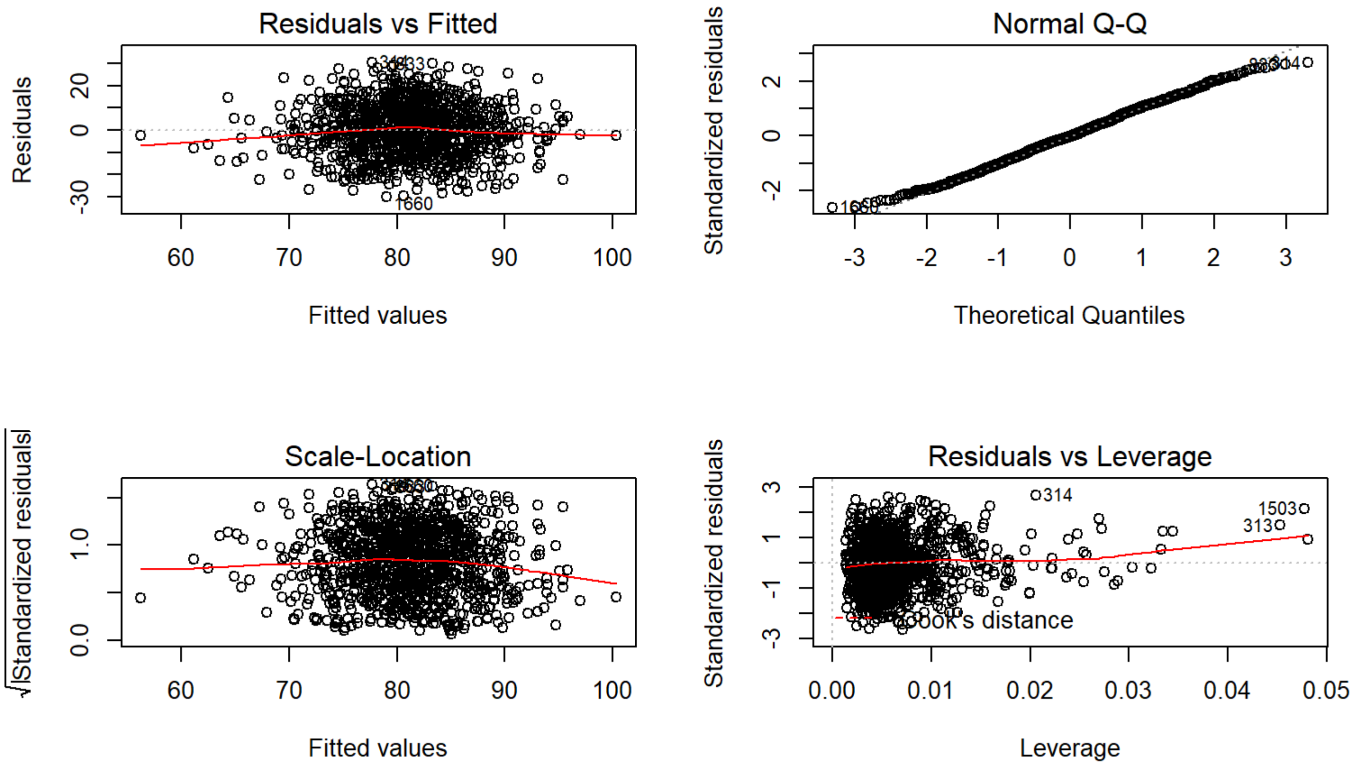
Code

6.3.1 Plot Model 3

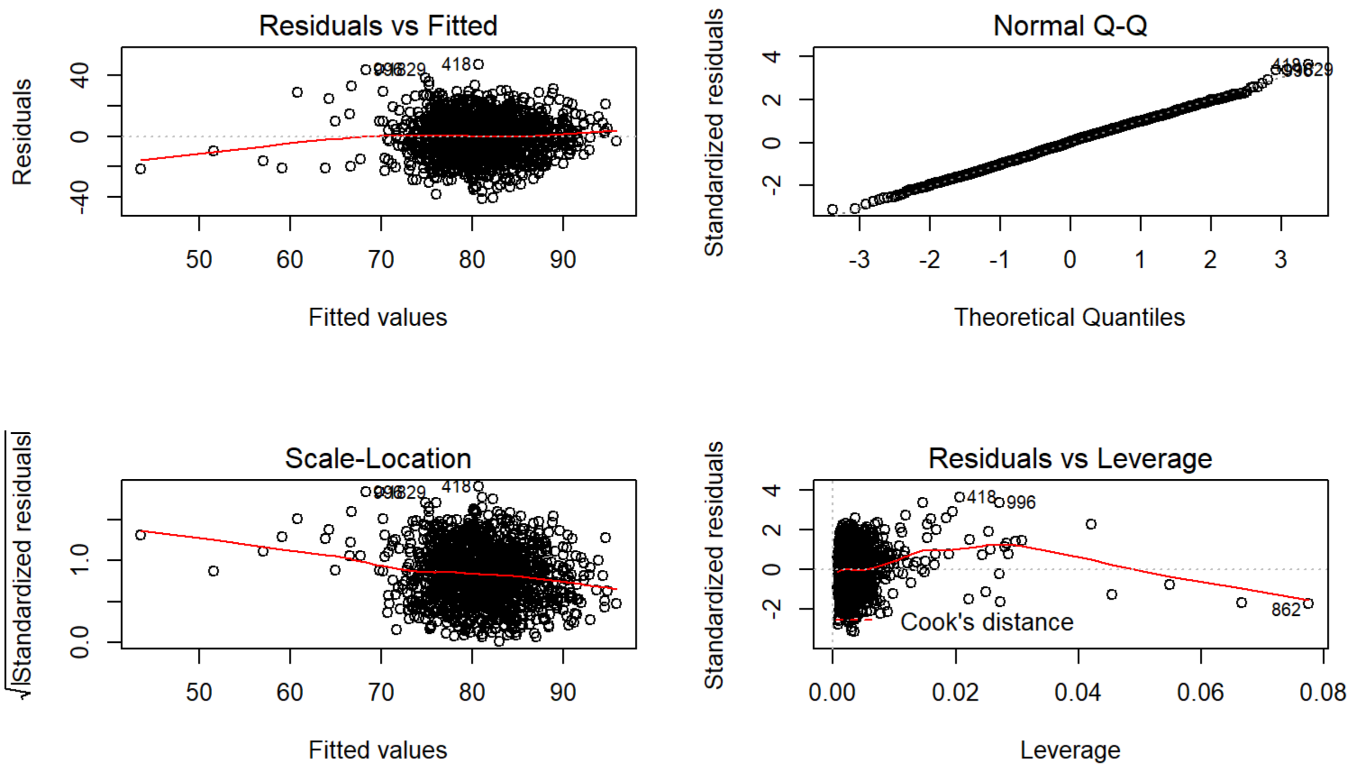
Code



Code



Code



7 SELECT MODELS

We have craeted couple of models in the last step, let's review the result for each our our model:

Code

Show 10 entries Search:

	ModelName	Adjusted.R2	P.Value	AIC	Note
1					
2	model1	0.4346	8.26675339500243e-121	7732.17046271654	BATTING_2B,BATTING_3B,BASERUN_SB ,BASERUN_CS,FIELDING_E,FIELDING_DP
3	model2	0.3902	9.43169458989572e-133	9741.06557425804	All are significant
4	model3a	0.1747	6.064035000153e-42	8122.0744174421	BATTING_3B,FIELDING_E ,BATTING_2B,FIELDING_DP are significant
5	model3b	0.1134	3.7241282367616e-36	11207.2018569633	All are significant

	ModelName	Adjusted.R2	P.Value	AIC	Note
6	model3	0.4346	8.26675339500243e-121	7732.17046271654	Nothing is significant
7	step_back	0.4359	1.77650951196573e-123	7726.72387730447	more vars significant

Showing 1 to 7 of 7 entries

Previous 1 Next

7.0.1 Multicolinearity

Lets Evaluate if we have any multicolinearity in our model1s.Multicollinearity (also collinearity) is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a non-trivial degree of accuracy.

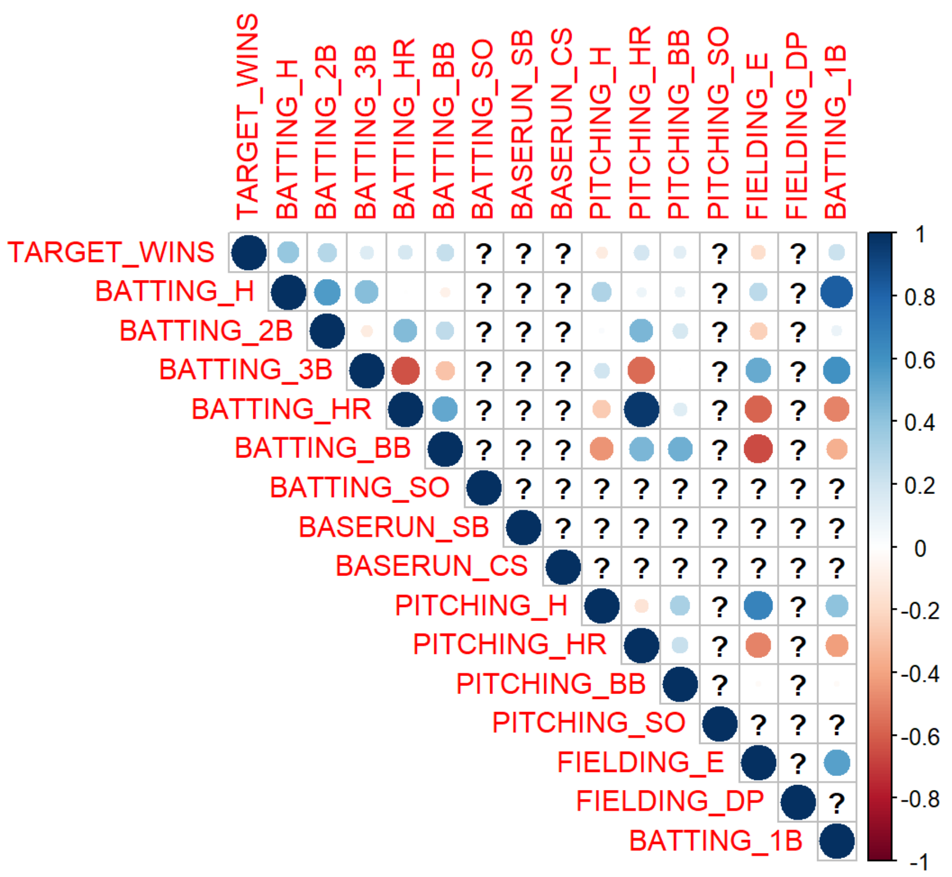
We will user alias function to detect the collinearity of all the predictor in the model1.

7.0.1.1 Model 1

Code

```
## Model :
## TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B + BATTING_HR +
##     BATTING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H +
##     PITCHING_HR + PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP +
##     BATTING_1B
##
## Complete :
##            (Intercept) BATTING_H BATTING_2B BATTING_3B BATTING_HR BATTING_BB
## BATTING_1B  0           1         -1         -1         -1         0
##            BATTING_SO BASERUN_SB BASERUN_CS PITCHING_H PITCHING_HR PITCHING_BB
## BATTING_1B  0          0          0          0          0           0
##            PITCHING_SO FIELDING_E FIELDING_DP
## BATTING_1B  0           0          0
```

Code



Code

Result shows that `BATTING_1B` is corealted with `BATTING_H` , `BATTING_2B` `BATTING_3B` , `BATTING_HR` . Here `+1` and `-1` are indicative of sign of coefecifint of the repstive predictor while stating the value for `BATTING_1B` .

Corrplot also suggest the same except , it doen't show high correlation between `BATTING_H``BATTING_HR` . In our Model2 , we well just follow the p-value significance test and build the model.

Code

RMSE <dbl>	R2 <dbl>
9.804207	0.4255646

1 row

7.0.2 Model 2

Here `alias` doen't suggest any correlated predictor. Now we can run VIF (variance inflation factor), which measures how much the variance of a regression coefficient is inflated due to multicollinearity in the model. The smallest possible value of VIF is one (absence of multicollinearity). Here we will look for VIF value, if that exceeds 5 or 10 indicates a problematic amount of collinearity. "Read More"[' http://www.sthda.com/english/articles/39-regression-model-diagnostics/160-multicollinearity-essentials-and-vif-in-r/ (http://www.sthda.com/english/articles/39-regression-model-diagnostics/160-multicollinearity-essentials-and-vif-in-r/)']

Code

```
## Model :
## TARGET_WINS ~ BATTING_H + BATTING_3B + BATTING_HR + BATTING_BB +
##     BATTING_SO + BASERUN_SB + PITCHING_SO + PITCHING_H + PITCHING_SO +
##     FIELDING_E + FIELDING_DP
```

Code

```
##   BATTING_H  BATTING_3B  BATTING_HR  BATTING_BB  BATTING_SO  BASERUN_SB
##   23.591594    2.924829    4.274146    1.259010  242.802006    1.539592
## PITCHING_SO  PITCHING_H  FIELDING_E FIELDING_DP
##  225.307718   48.406757    2.835717    1.353810
```

VIF output suggest that BATTING_H, PITCHING_H, BATTING_SO,PITCHING_SO are highly impacting model due their colinear relation.

Code

RMSE <dbl>	R2 <dbl>
10.25912	0.3883479

1 row

7.0.2.1 Model 3

Code

RMSE <dbl>	R2 <dbl>
9.804207	0.4255646

1 row

7.0.2.2 Model 4

Code

```
##
## Call:
## lm(formula = TARGET_WINS ~ . - BATTING_H - BATTING_2B - BATTING_3B -
##     BATTING_HR, data = moneyball_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.334  -6.834  -0.136   6.517  29.480
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.857266   8.110353   7.380 3.23e-13 ***
## BATTING_BB   0.006719   0.039339   0.171 0.864410
## BATTING_SO   0.006949   0.022410   0.310 0.756561
## BASERUN_SB   0.035119   0.010675   3.290 0.001036 **
## BASERUN_CS   0.068018   0.022780   2.986 0.002894 **
## PITCHING_H  -0.002634   0.006751  -0.390 0.696514
## PITCHING_HR  0.116181   0.012748   9.113  < 2e-16 ***
## PITCHING_BB  0.030035   0.037698   0.797 0.425796
## PITCHING_SO -0.033549   0.021345  -1.572 0.116309
## FIELDING_E  -0.127737   0.012193 -10.476  < 2e-16 ***
## FIELDING_DP -0.104855   0.016090  -6.517 1.12e-10 ***
## BATTING_1B   0.038734   0.010312   3.756 0.000182 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.86 on 1040 degrees of freedom
##   (543 observations deleted due to missingness)
## Multiple R-squared:  0.3933, Adjusted R-squared:  0.3869
## F-statistic:  61.3 on 11 and 1040 DF,  p-value: < 2.2e-16
```

Code

```
##  BATTING_BB  BATTING_SO  BASERUN_SB  BASERUN_CS  PITCHING_H PITCHING_HR
##  107.539027  216.776484    2.415563    2.721623   14.163628    4.448142
## PITCHING_BB PITCHING_SO  FIELDING_E FIELDING_DP  BATTING_1B
##  144.662915  216.288753    2.187153    1.133447    7.973818
```

Code

RMSE <dbl>	R2 <dbl>
9.922245	0.4109811

1 row

7.0.2.3 Model 5

Code

```
##
## Call:
## lm(formula = TARGET_WINS ~ . - PITCHING_SO - PITCHING_BB - BATTING_H -
##     BATTING_2B - BATTING_3B - BATTING_HR, data = moneyball_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.408  -6.629  -0.164   6.503  29.704
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 60.129049   8.109072   7.415 2.51e-13 ***
## BATTING_BB   0.038506   0.004083   9.430  < 2e-16 ***
## BATTING_SO  -0.027830   0.002911  -9.562  < 2e-16 ***
## BASERUN_SB   0.036013   0.010592   3.400   0.0007 ***
## BASERUN_CS   0.066311   0.022725   2.918   0.0036 **
## PITCHING_H  -0.010813   0.002702  -4.002 6.71e-05 ***
## PITCHING_HR  0.123928   0.010677  11.607  < 2e-16 ***
## FIELDING_E  -0.128182   0.012162 -10.540  < 2e-16 ***
## FIELDING_DP -0.105752   0.016091  -6.572 7.82e-11 ***
## BATTING_1B   0.049404   0.006386   7.737 2.40e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.87 on 1042 degrees of freedom
##   (543 observations deleted due to missingness)
## Multiple R-squared:  0.3909, Adjusted R-squared:  0.3857
## F-statistic: 74.32 on 9 and 1042 DF,  p-value: < 2.2e-16
```

Code

```
##  BATTING_BB  BATTING_SO  BASERUN_SB  BASERUN_CS  PITCHING_H PITCHING_HR
##    1.156266    3.649407    2.373748    2.703075    2.263550    3.113814
##  FIELDING_E FIELDING_DP  BATTING_1B
##    2.171454    1.131320    3.051488
```

Code

RMSE <dbl>	R2 <dbl>
9.991091	0.4029489

1 row

7.0.2.4 Model 6 (Step back)

VIF result suggest that all the predictors in the model `step_back` have no multicolinearirty exist in them.

Code

```
##
## Call:
## lm(formula = poly_call[2], data = moneyball_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.0741  -6.5189  -0.0304   6.5548  28.5287
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.226582   7.718003   7.674 3.83e-14 ***
## BATTING_1B   0.021961   0.006883   3.191 0.001462 **
## BATTING_2B  -0.052339   0.008634  -6.062 1.88e-09 ***
## BATTING_3B   0.195353   0.024739   7.897 7.25e-15 ***
## BATTING_HR   0.123437   0.009440  13.077  < 2e-16 ***
## BATTING_BB   0.039462   0.003927  10.048  < 2e-16 ***
## BASERUN_SB   0.036916   0.010210   3.616 0.000314 ***
## BASERUN_CS   0.051264   0.021908   2.340 0.019475 *
## PITCHING_H   0.011846   0.002851   4.155 3.52e-05 ***
## PITCHING_SO -0.020636   0.002747  -7.513 1.25e-13 ***
## FIELDING_E  -0.162363   0.012228 -13.278  < 2e-16 ***
## FIELDING_DP -0.106435   0.015427  -6.899 9.07e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.458 on 1040 degrees of freedom
##   (543 observations deleted due to missingness)
## Multiple R-squared:  0.4418, Adjusted R-squared:  0.4359
## F-statistic: 74.83 on 11 and 1040 DF,  p-value: < 2.2e-16
```

Code

```
##  BATTING_1B  BATTING_2B  BATTING_3B  BATTING_HR  BATTING_BB  BASERUN_SB
##    3.860683    1.533907    2.592355    2.434721    1.164947    2.401669
##  BASERUN_CS  PITCHING_H PITCHING_SO  FIELDING_E FIELDING_DP
##    2.736003    2.744801    3.892807    2.390615    1.132495
```

Code

RMSE <dbl>	R2 <dbl>
9.802052	0.4258251

1 row

Lets only consider Model with beter RMSE and R2 and check it with AIC test:

Model Name	RMSE	R^2
model1	9.80421	0.42556
model2	10.2591	0.38835
model3	10.0631	0.40604
model4	9.92225	0.41098
model5	9.99109	0.40295
Step Back	9.77083	0.428734

Lets run the AIC weight test to evaluate the best model out of few selected models :

Code

```
##           dAICc df weight
## step_back  0.0  13 1
## model4    87.6  13 <0.001
## model5    87.6  11 <0.001
```

In Both test `Model1` is doing well, but since its not a parsomonious model we decided to check among `model4` and `model5` and `step_back`. Which is a parsomonious model, with no multicolnearity among the predictors. We also note how multicolinearity in models were impacting its effect on overall perfromcne of the model.

Selected Model = `step_back`

7.1 Predict of Eval data

Run the `step_backward` model on Eval data.

Code

```
## Start:  AIC=-9677.33
## BATTING_H ~ BATTING_2B + BATTING_3B + BATTING_HR + BATTING_BB +
##     BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H + PITCHING_HR +
##     PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP + BATTING_1B
##
##               Df Sum of Sq    RSS     AIC
## - BASERUN_CS   1       0.0    0.0 -9680.5
## - PITCHING_BB  1       0.0    0.0 -9679.8
## - FIELDING_E   1       0.0    0.0 -9679.4
## - BATTING_BB   1       0.0    0.0 -9679.3
## - FIELDING_DP  1       0.0    0.0 -9679.1
## - PITCHING_H   1       0.0    0.0 -9678.8
## - BASERUN_SB   1       0.0    0.0 -9678.5
## - PITCHING_HR  1       0.0    0.0 -9677.7
## <none>                        0.0 -9677.3
## - BATTING_SO   1       0.0    0.0 -9674.7
## - PITCHING_SO  1       0.0    0.0 -9673.6
## - BATTING_HR   1     196.1  196.1    52.3
## - BATTING_3B   1    4607.5 4607.5   588.9
## - BATTING_2B   1    4715.2 4715.2   592.9
## - BATTING_1B   1    5029.8 5029.8   603.8
##
## Step:  AIC=-9680.52
## BATTING_H ~ BATTING_2B + BATTING_3B + BATTING_HR + BATTING_BB +
##     BATTING_SO + BASERUN_SB + PITCHING_H + PITCHING_HR + PITCHING_BB +
##     PITCHING_SO + FIELDING_E + FIELDING_DP + BATTING_1B
##
##               Df Sum of Sq    RSS     AIC
## - PITCHING_BB  1       0.0    0.0 -9682.3
## - FIELDING_E   1       0.0    0.0 -9682.3
## - FIELDING_DP  1       0.0    0.0 -9681.8
## - PITCHING_H   1       0.0    0.0 -9681.3
## - BATTING_BB   1       0.0    0.0 -9681.2
## <none>                        0.0 -9680.5
## - BASERUN_SB   1       0.0    0.0 -9680.4
## - PITCHING_HR  1       0.0    0.0 -9679.3
## - PITCHING_SO  1       0.0    0.0 -9676.4
## - BATTING_SO   1       0.0    0.0 -9671.8
## - BATTING_HR   1     196.7  196.7    50.8
## - BATTING_3B   1    4616.4 4616.4   587.3
## - BATTING_2B   1    4778.8 4778.8   593.1
## - BATTING_1B   1    5067.4 5067.4   603.1
##
## Step:  AIC=-9682.32
## BATTING_H ~ BATTING_2B + BATTING_3B + BATTING_HR + BATTING_BB +
##     BATTING_SO + BASERUN_SB + PITCHING_H + PITCHING_HR + PITCHING_SO +
##     FIELDING_E + FIELDING_DP + BATTING_1B
##
##               Df Sum of Sq   RSS     AIC
## - FIELDING_E   1         0     0 -9684.4
## - FIELDING_DP  1         0     0 -9683.8
## <none>                         0 -9682.3
## - BATTING_BB   1         0     0 -9682.2
## - BASERUN_SB   1         0     0 -9682.2
## - PITCHING_HR  1         0     0 -9681.4
## - PITCHING_H   1         0     0 -9680.3
## - PITCHING_SO  1         0     0 -9678.1
## - BATTING_SO   1         0     0 -9673.6
## - BATTING_HR   1       200   200    51.6
## - BATTING_3B   1     14322 14322   777.7
## - BATTING_2B   1     25270 25270   874.3
## - BATTING_1B   1     31677 31677   912.7
##
## Step:  AIC=-9684.37
## BATTING_H ~ BATTING_2B + BATTING_3B + BATTING_HR + BATTING_BB +
##     BATTING_SO + BASERUN_SB + PITCHING_H + PITCHING_HR + PITCHING_SO +
##     FIELDING_DP + BATTING_1B
##
##               Df Sum of Sq   RSS     AIC
## - FIELDING_DP  1         0     0 -9686.3
## <none>                         0 -9684.4
## - BATTING_BB   1         0     0 -9684.3
## - PITCHING_H   1         0     0 -9684.2
## - PITCHING_HR  1         0     0 -9684.0
## - BASERUN_SB   1         0     0 -9683.6
## - PITCHING_SO  1         0     0 -9679.8
## - BATTING_SO   1         0     0 -9675.9
## - BATTING_HR   1       203   203    52.6
## - BATTING_3B   1     15294 15294   786.9
## - BATTING_2B   1     25511 25511   873.9
## - BATTING_1B   1     31824 31824   911.5
```

```
##
## Step:  AIC=-9686.3
## BATTING_H ~ BATTING_2B + BATTING_3B + BATTING_HR + BATTING_BB +
##     BATTING_SO + BASERUN_SB + PITCHING_H + PITCHING_HR + PITCHING_SO +
##     BATTING_1B
##
##               Df Sum of Sq   RSS     AIC
## <none>                         0 -9686.3
## - BASERUN_SB   1         0     0 -9686.3
## - PITCHING_H   1         0     0 -9685.3
## - BATTING_BB   1         0     0 -9685.0
## - PITCHING_HR  1         0     0 -9684.5
## - PITCHING_SO  1         0     0 -9681.5
## - BATTING_SO   1         0     0 -9676.5
## - BATTING_HR   1       204   204    50.9
## - BATTING_3B   1     15432 15432   786.4
## - BATTING_2B   1     25885 25885   874.4
## - BATTING_1B   1     32131 32131   911.1
```

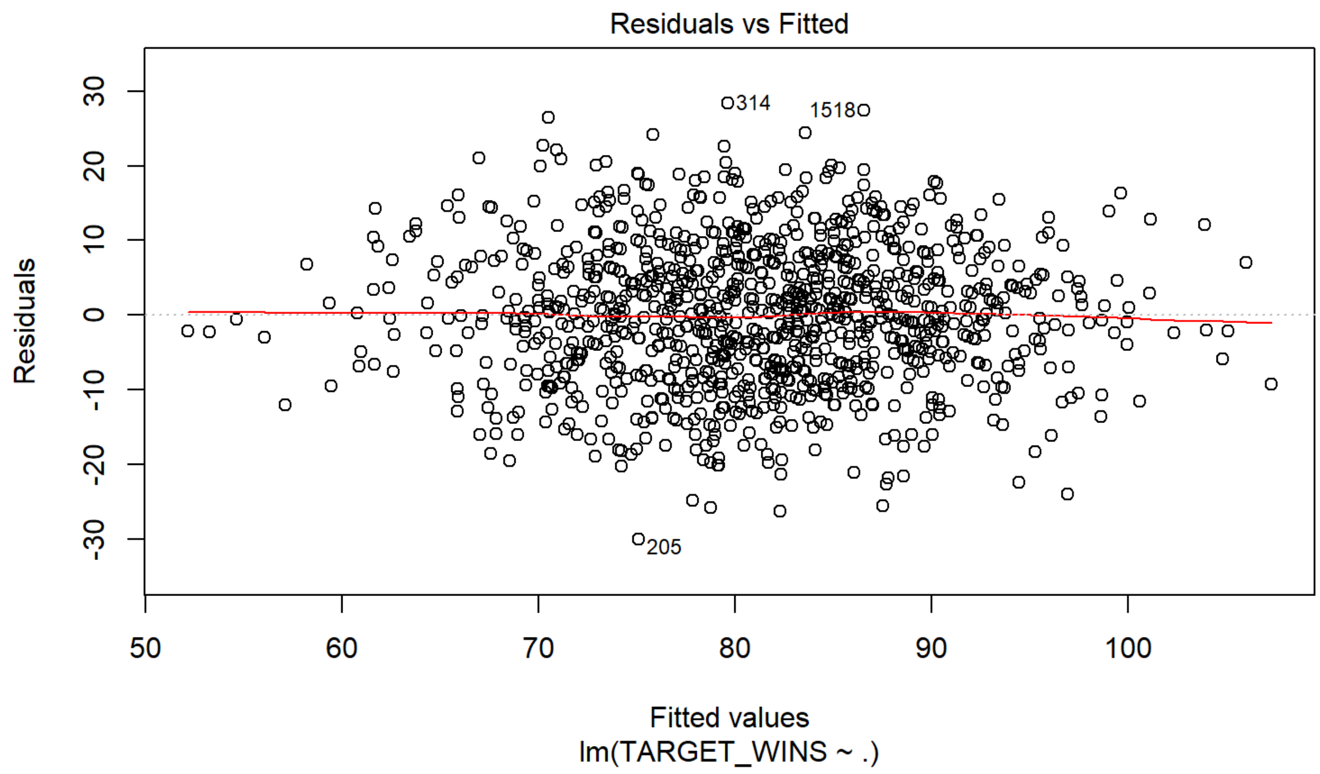
Code

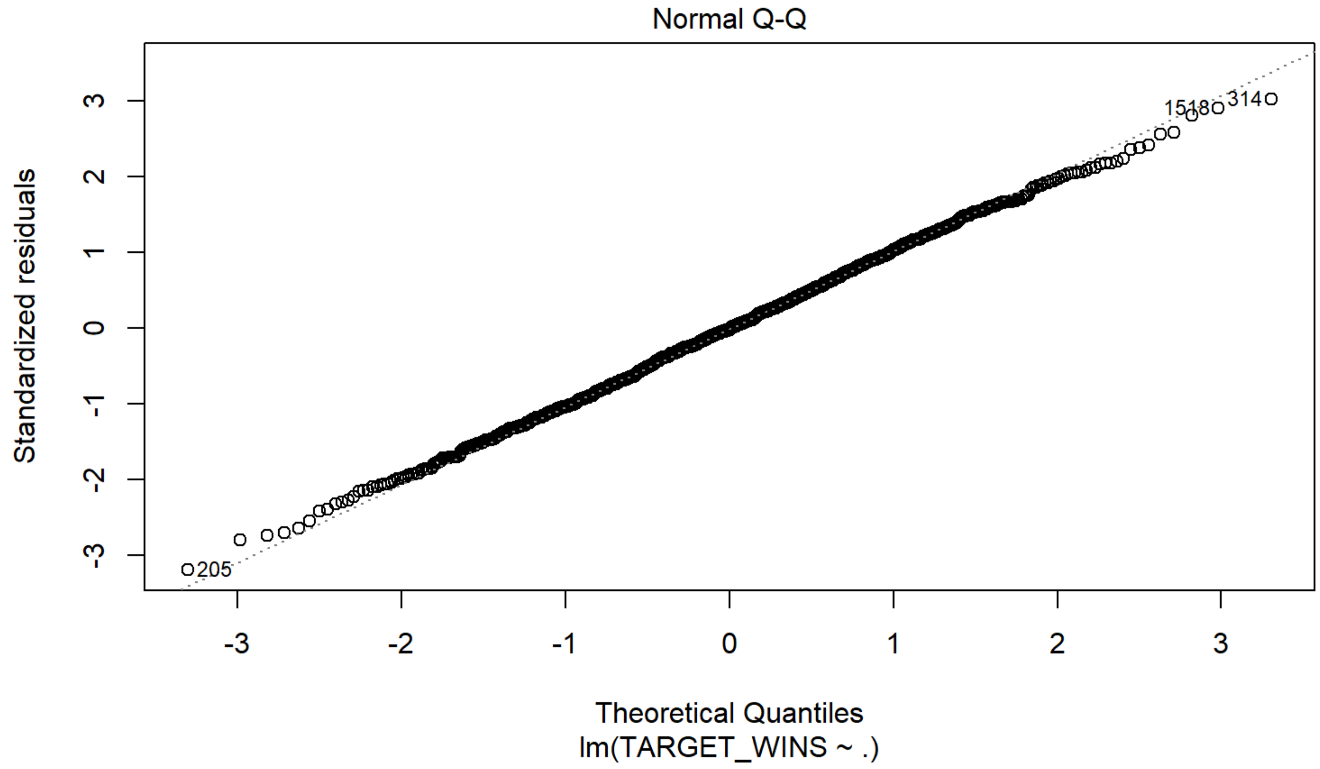
```
##
## Call:
## lm(formula = BATTING_H ~ BATTING_2B + BATTING_3B + BATTING_HR +
##     BATTING_BB + BATTING_SO + BASERUN_SB + PITCHING_H + PITCHING_HR +
##     PITCHING_SO + BATTING_1B, data = med)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -4.866e-12 -5.020e-14  2.600e-14  1.005e-13  5.880e-13
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept) -8.719e-13  7.612e-13 -1.145e+00  0.25374
## BATTING_2B   1.000e+00  2.554e-15  3.915e+14  < 2e-16 ***
## BATTING_3B   1.000e+00  3.308e-15  3.023e+14  < 2e-16 ***
## BATTING_HR   1.000e+00  2.878e-14  3.475e+13  < 2e-16 ***
## BATTING_BB  -1.870e-17  4.134e-16 -4.500e-02  0.96398
## BATTING_SO  -1.405e-14  5.314e-15 -2.643e+00  0.00904 **
## BASERUN_SB   6.607e-16  6.723e-16  9.830e-01  0.32722
## PITCHING_H  -2.819e-15  2.185e-15 -1.290e+00  0.19879
## PITCHING_HR -4.645e-14  2.859e-14 -1.625e+00  0.10613
## PITCHING_SO  1.311e-14  5.172e-15  2.535e+00  0.01221 *
## BATTING_1B   1.000e+00  2.292e-15  4.362e+14  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.109e-13 on 159 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.289e+30 on 10 and 159 DF,  p-value: < 2.2e-16
```

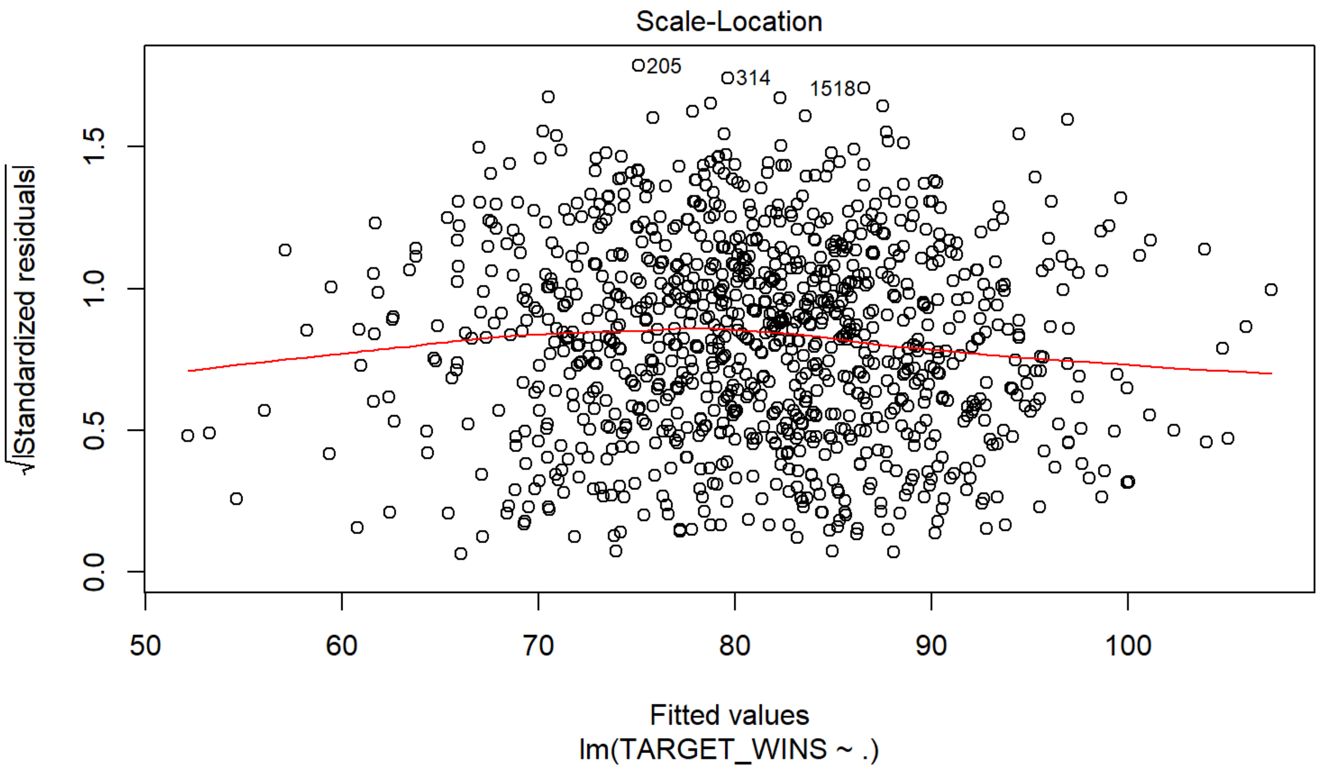
From the three models, model3 is a more parsimonious model. There is no significant difference in R2, Adjusted R2 and RMSE even when i did the treatment for multi-collinearity.

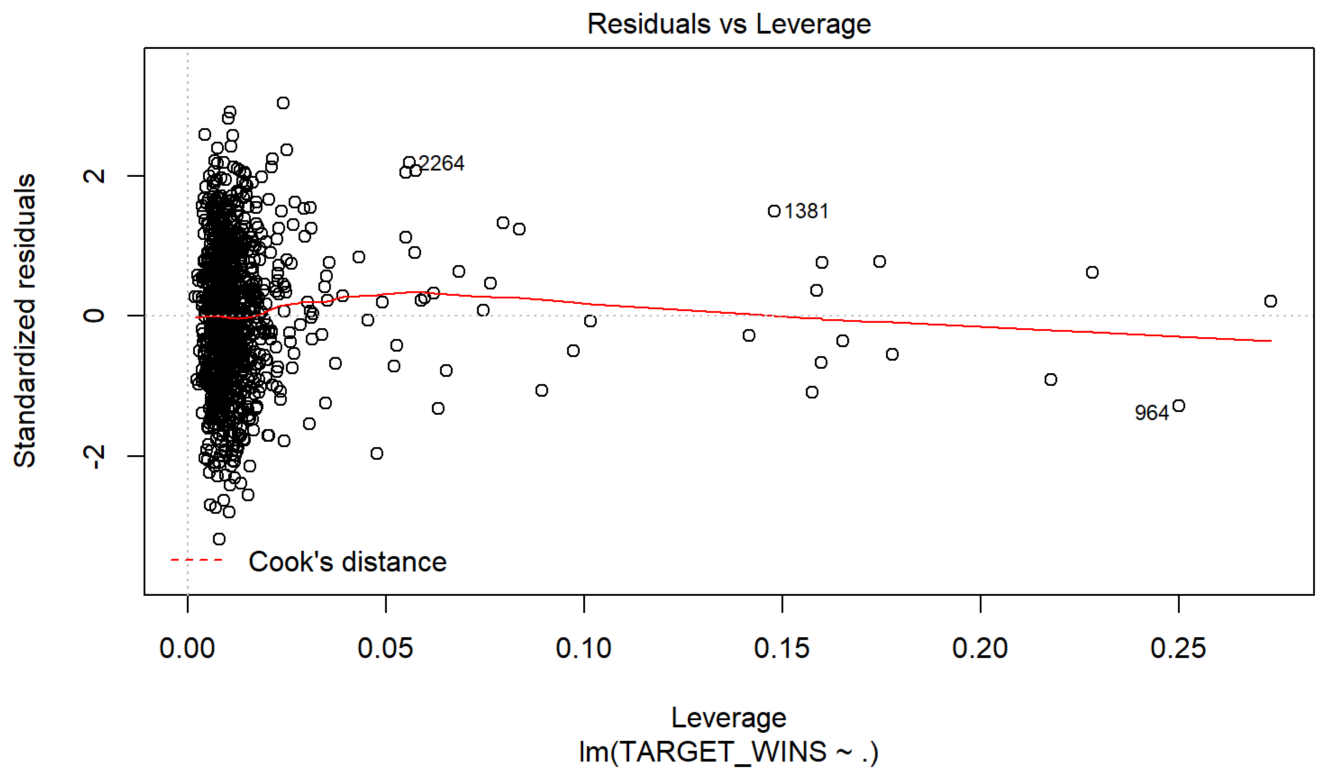
7.1.1 Model 1 : Kitchen Sink Model

Code

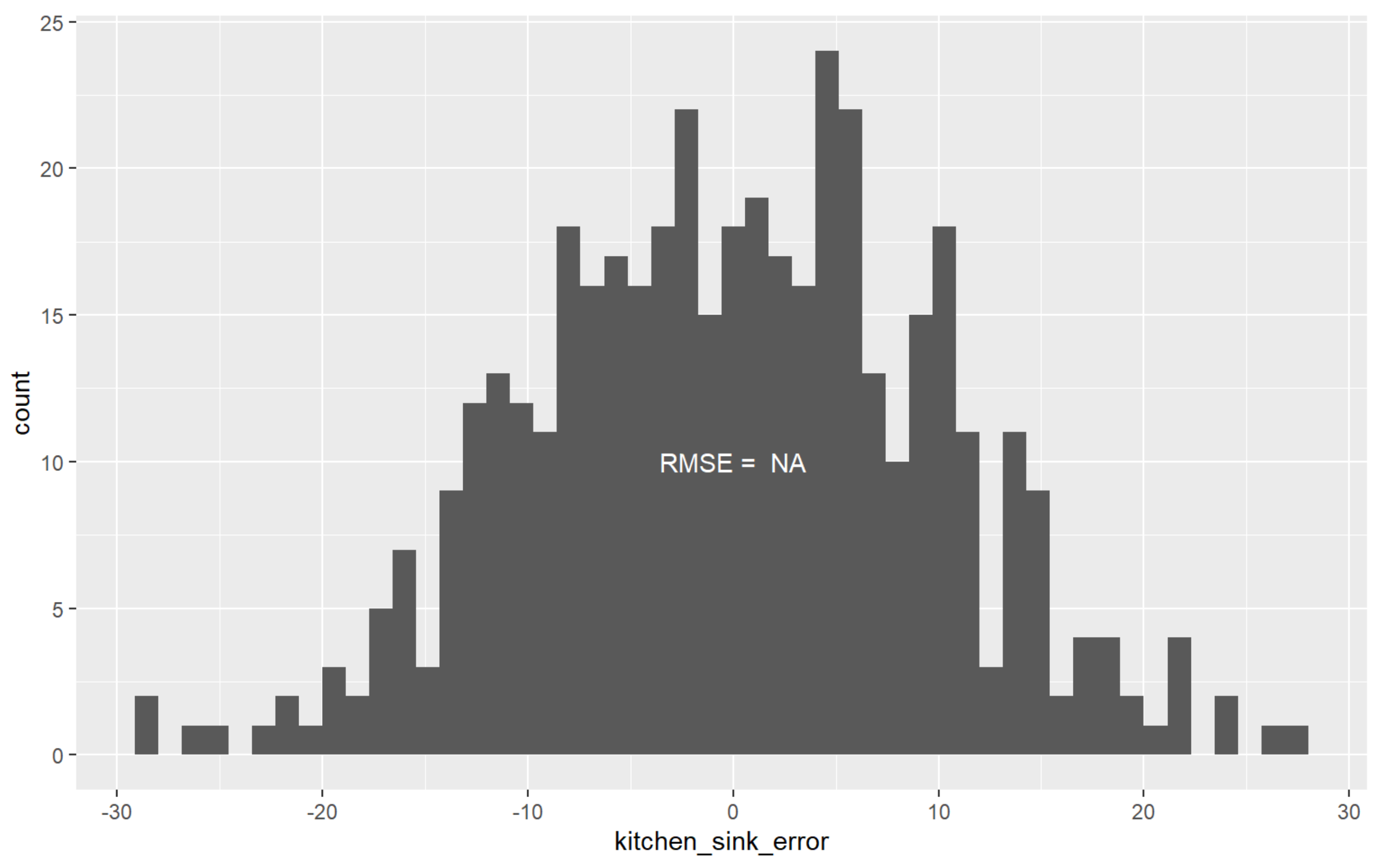

Residuals vs Fitted
Residuals
Fitted values
lm(TARGET_WINS ~ .)
314
1518
205
30
20
10
0
-10
-20
-30
50
60
70
80
90
100


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(TARGET_WINS ~ .)
1518
314
205
3
2
1
0
-1
-2
-3


Scale-Location
√|Standardized residuals|
Fitted values
lm(TARGET_WINS ~ .)
205
314
1518
1.5
1.0
0.5
0.0
50
60
70
80
90
100



Code

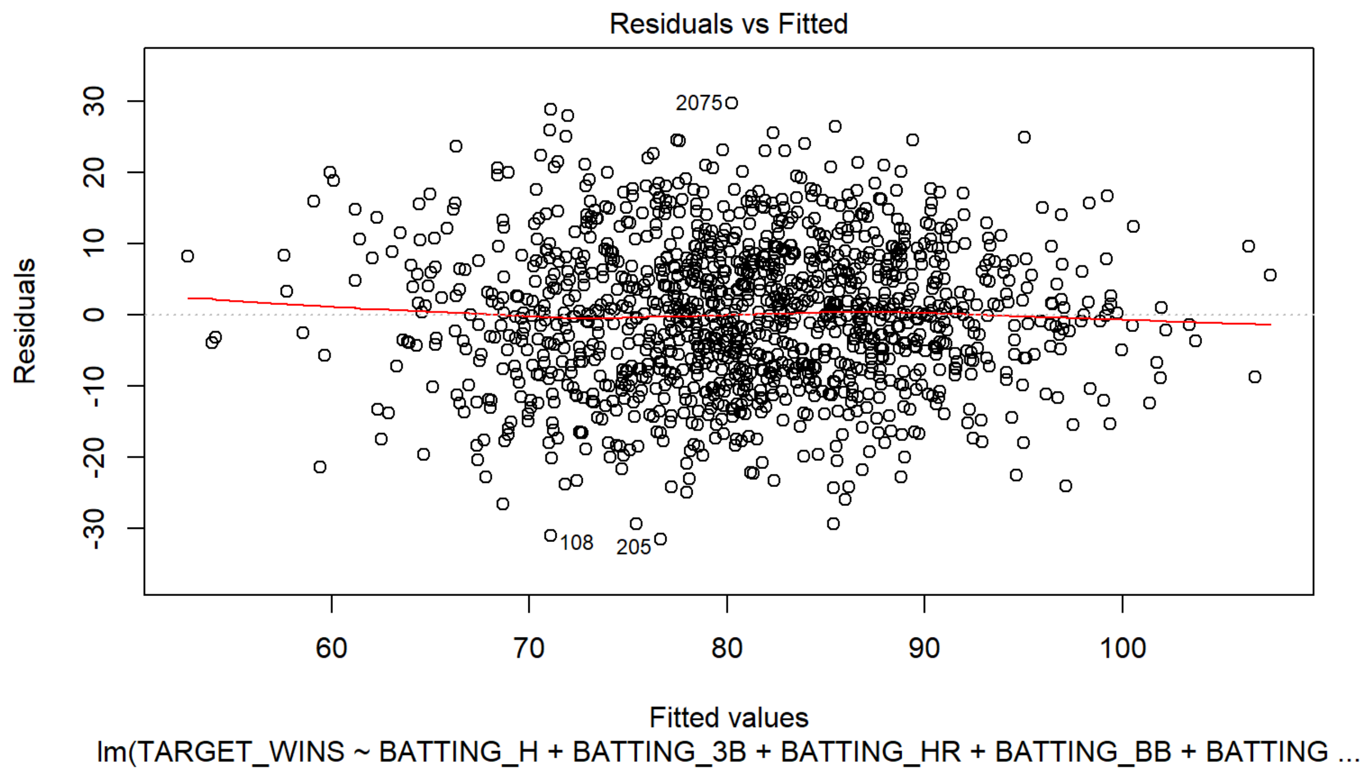


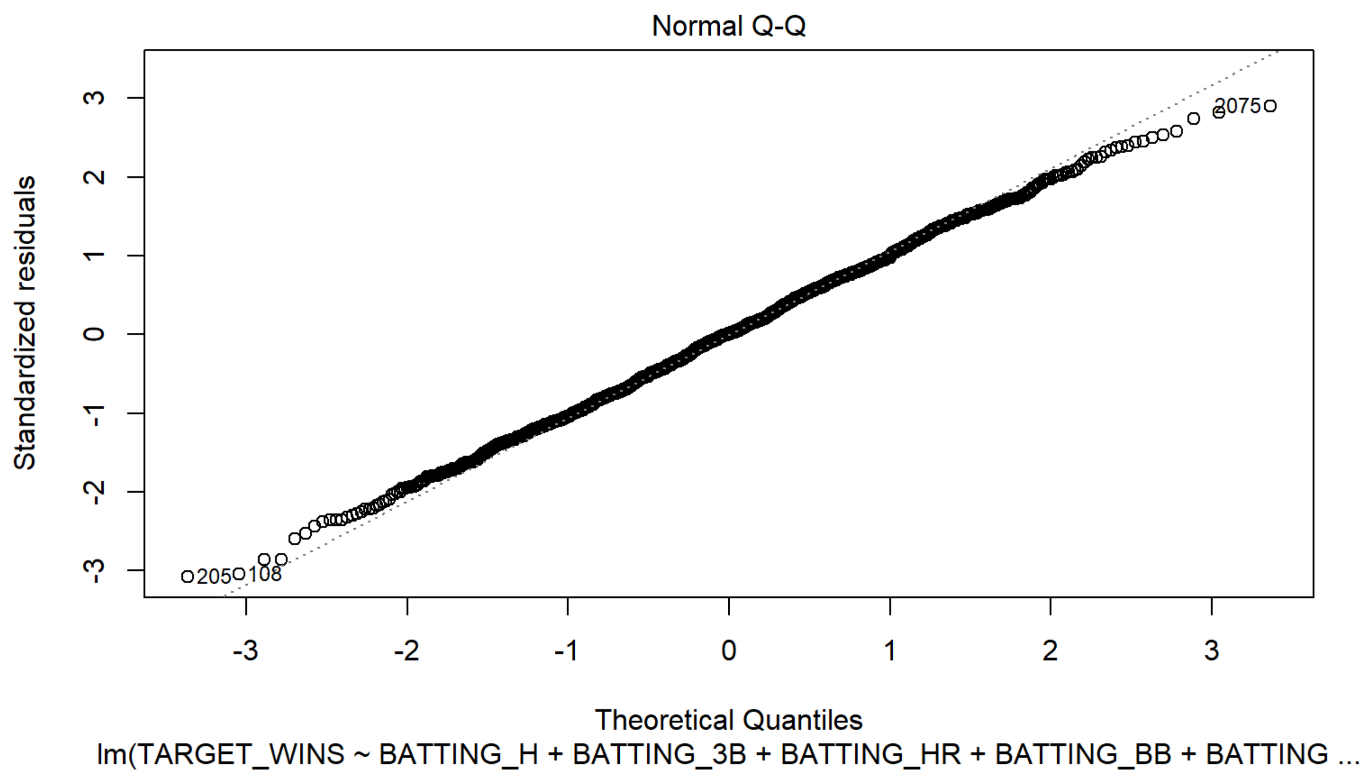
Code

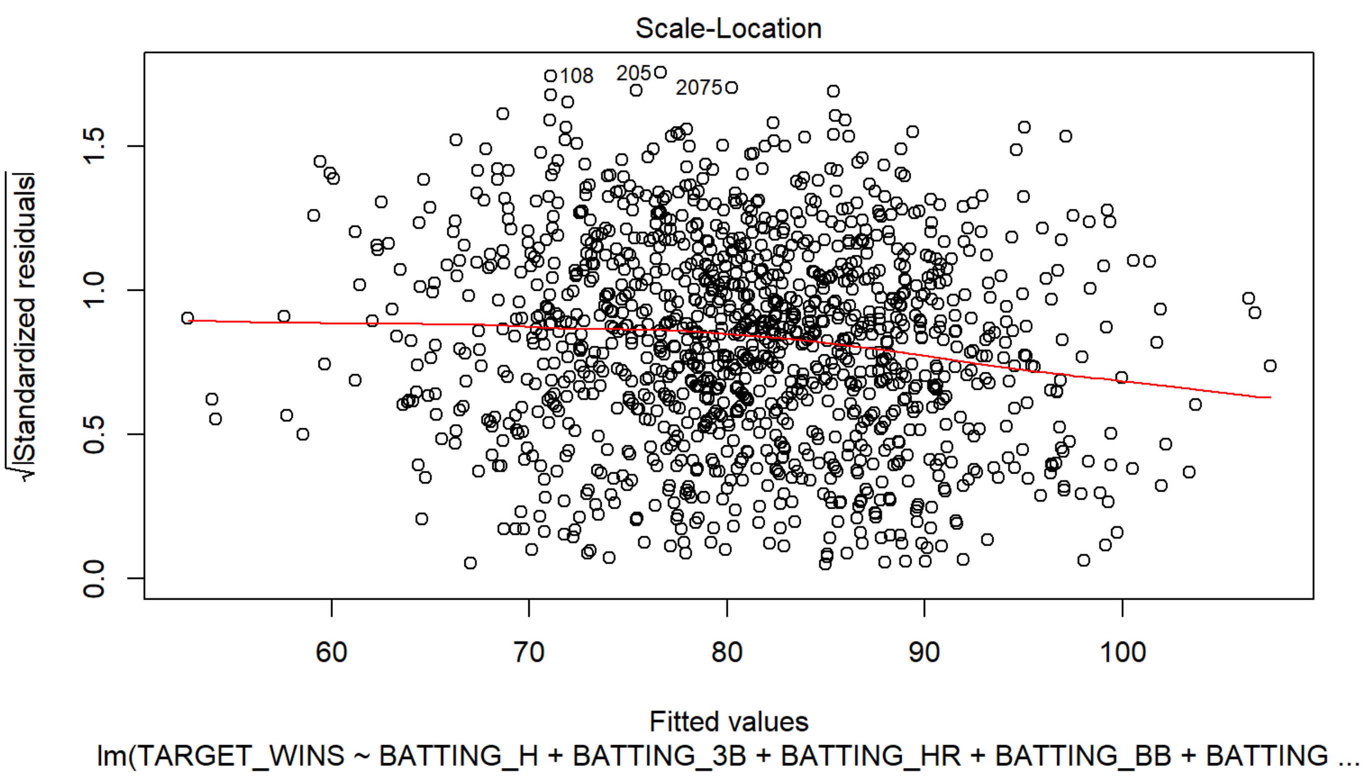
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
## -28.3735  -6.9033  -0.1124  -0.0408   6.4889  27.6495      247
```

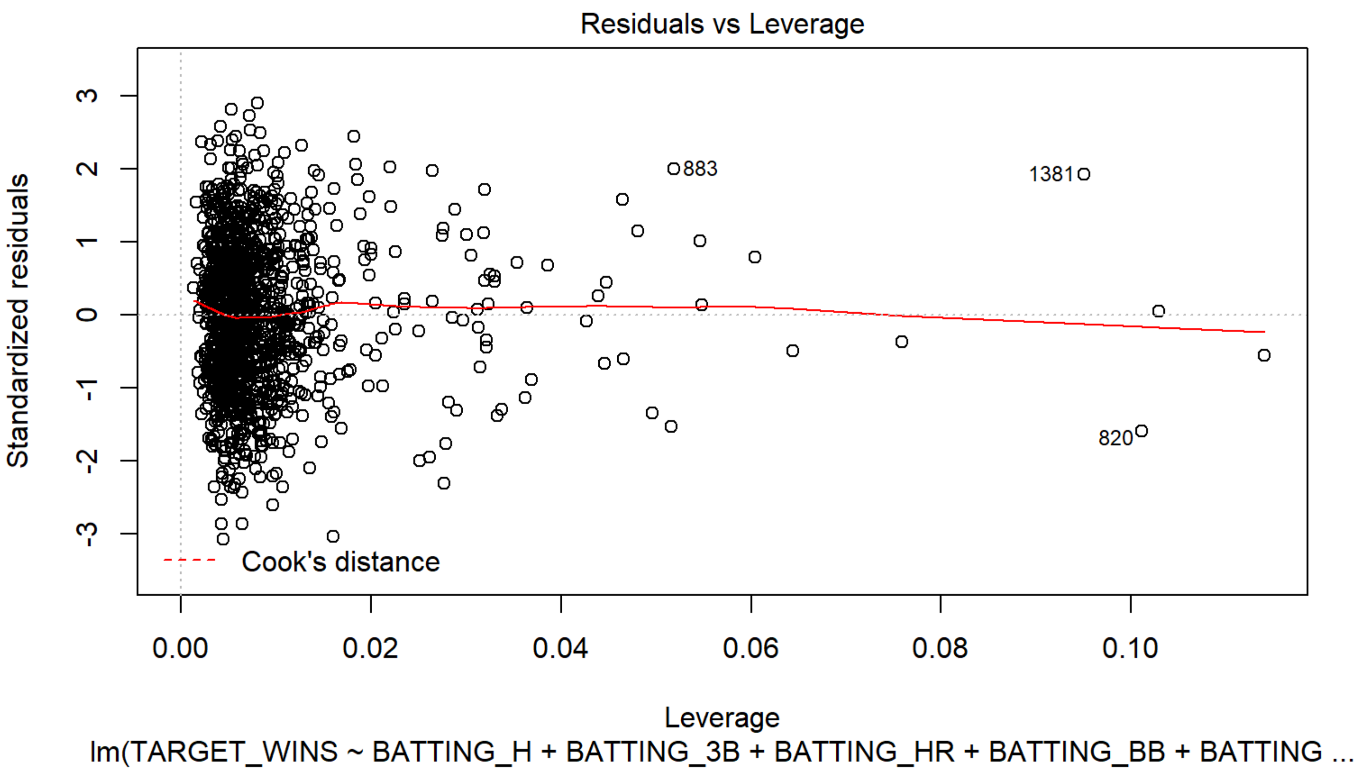
7.1.2 Model 2 : Simple Model

Code

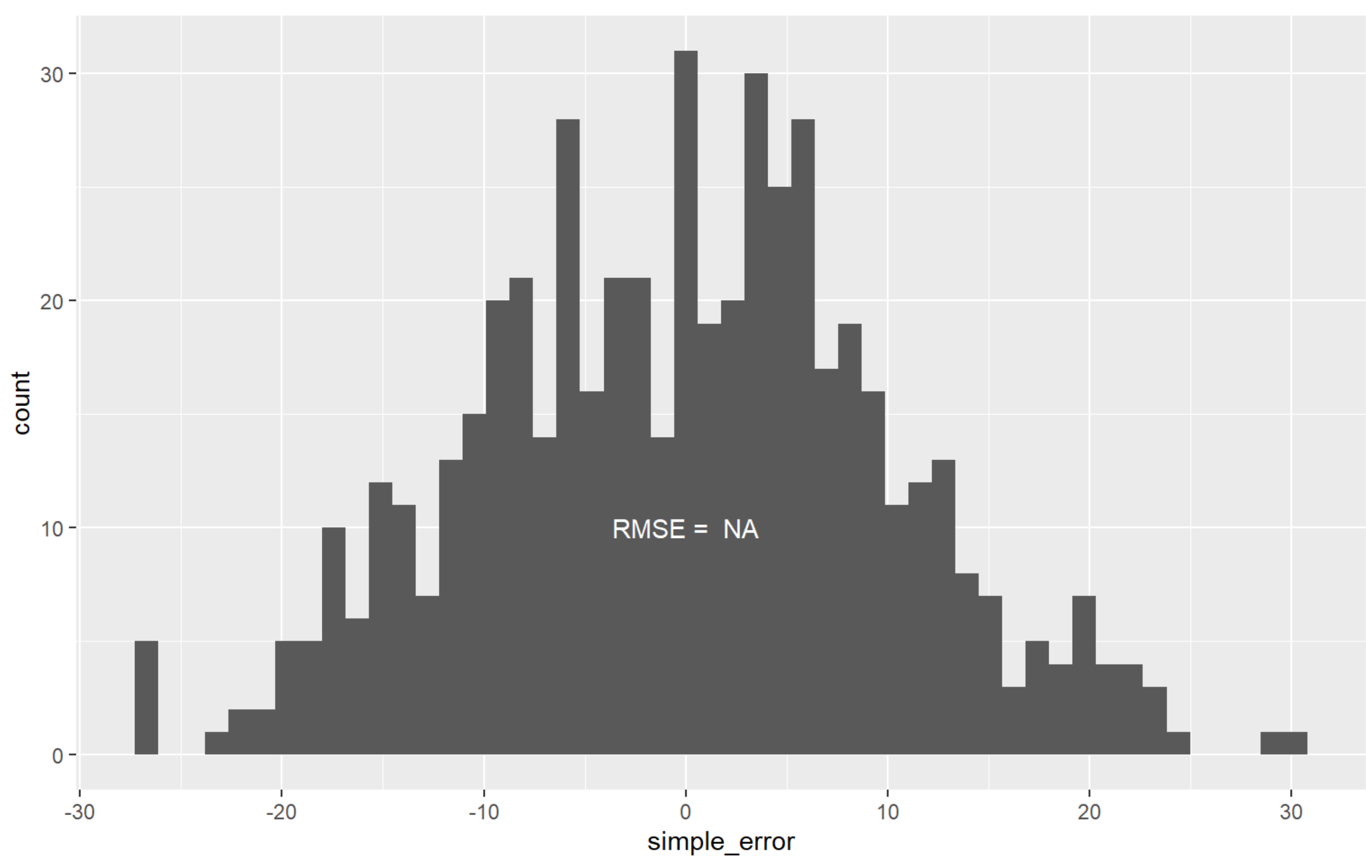

Residuals vs Fitted
Residuals
2075
108
205
Fitted values
lm(TARGET_WINS ~ BATTING_H + BATTING_3B + BATTING_HR + BATTING_BB + BATTING ...


Normal Q-Q
Standardized residuals
2075
205
108
Theoretical Quantiles
lm(TARGET_WINS ~ BATTING_H + BATTING_3B + BATTING_HR + BATTING_BB + BATTING ...


Scale-Location
√|Standardized residuals|
108
205
2075
Fitted values
lm(TARGET_WINS ~ BATTING_H + BATTING_3B + BATTING_HR + BATTING_BB + BATTING ...



Code

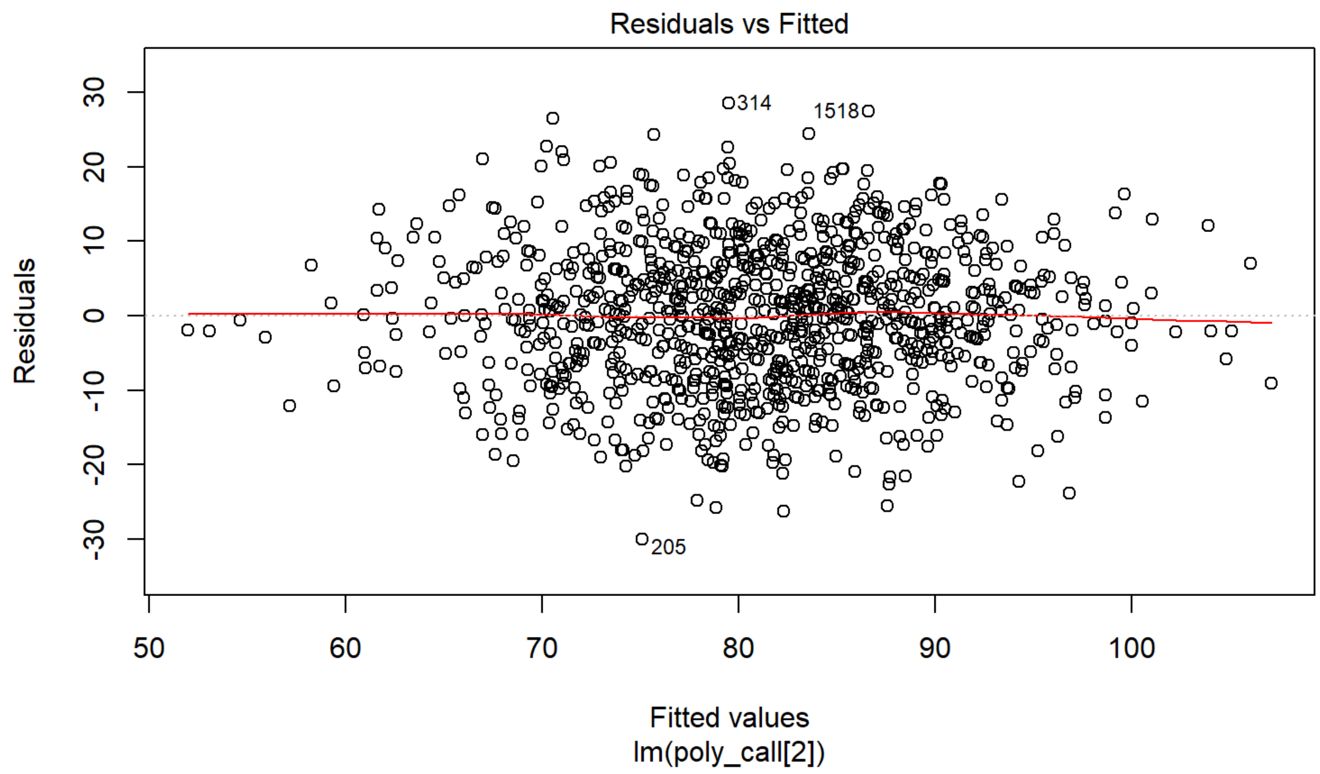


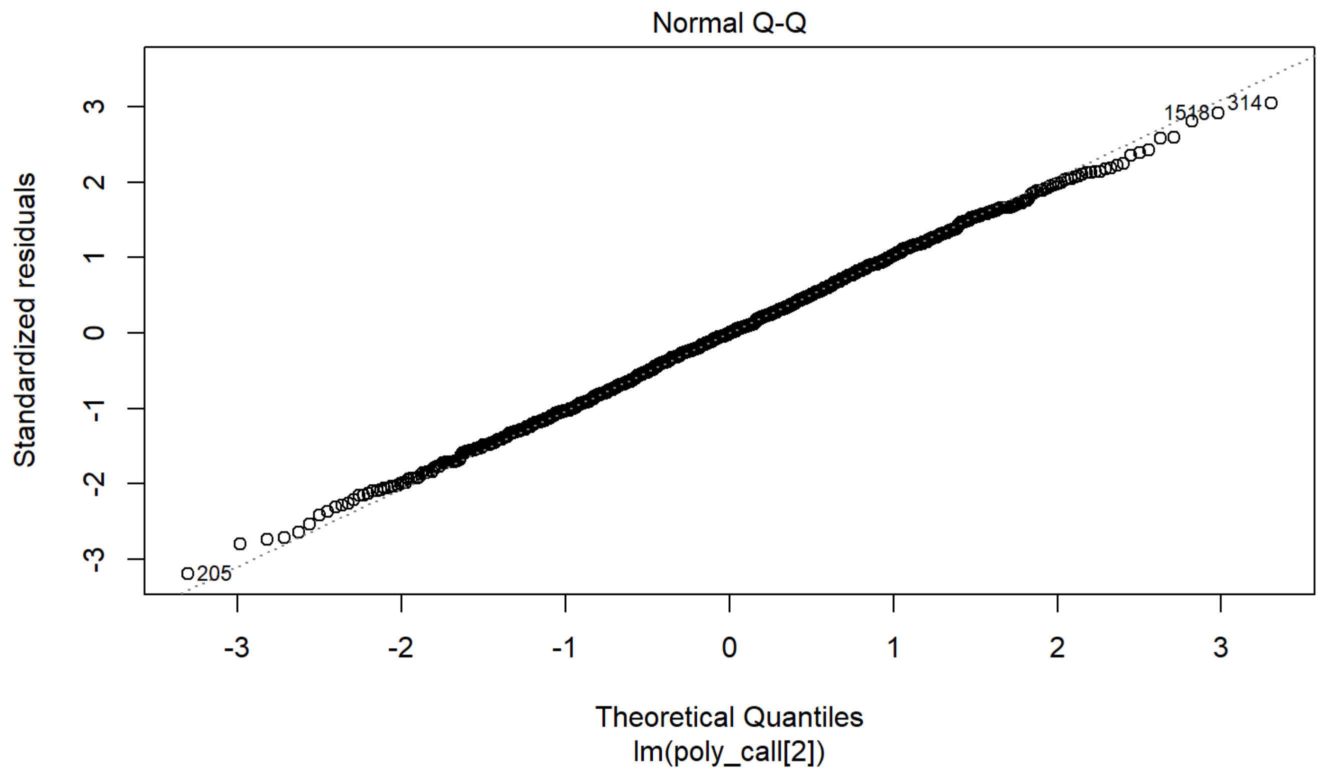
Code

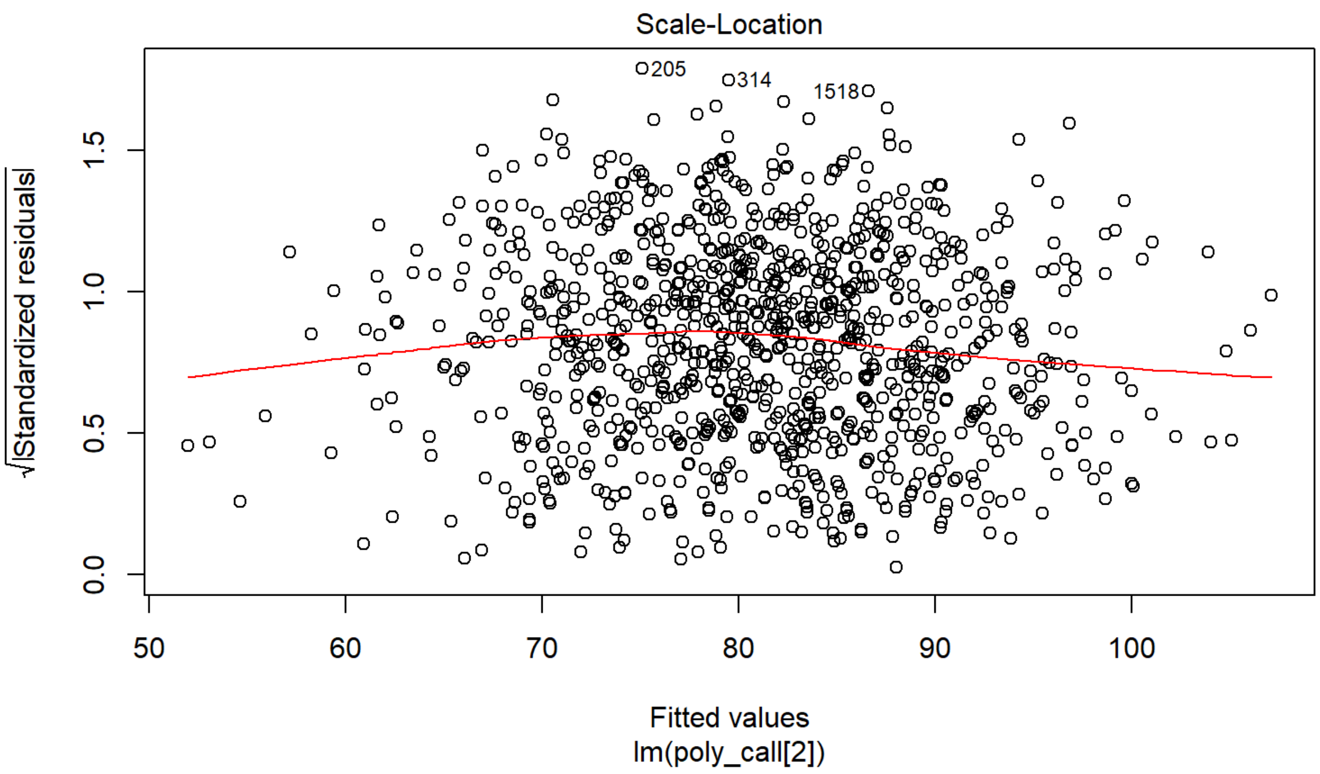
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
## -27.2876  -7.6292   0.2432  -0.1372   6.5731  29.6379      143
```

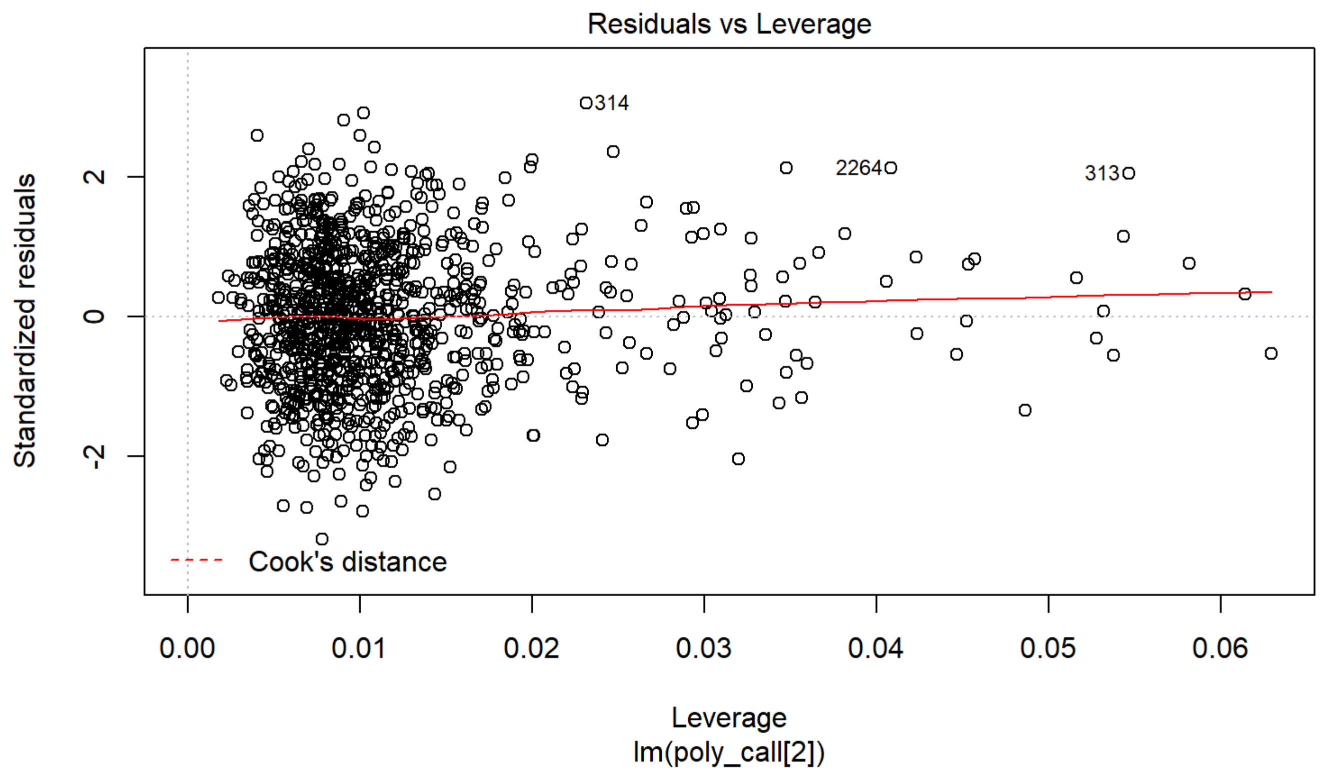
7.1.3 Model 3 : Higher Order Stepwise Regression

Code

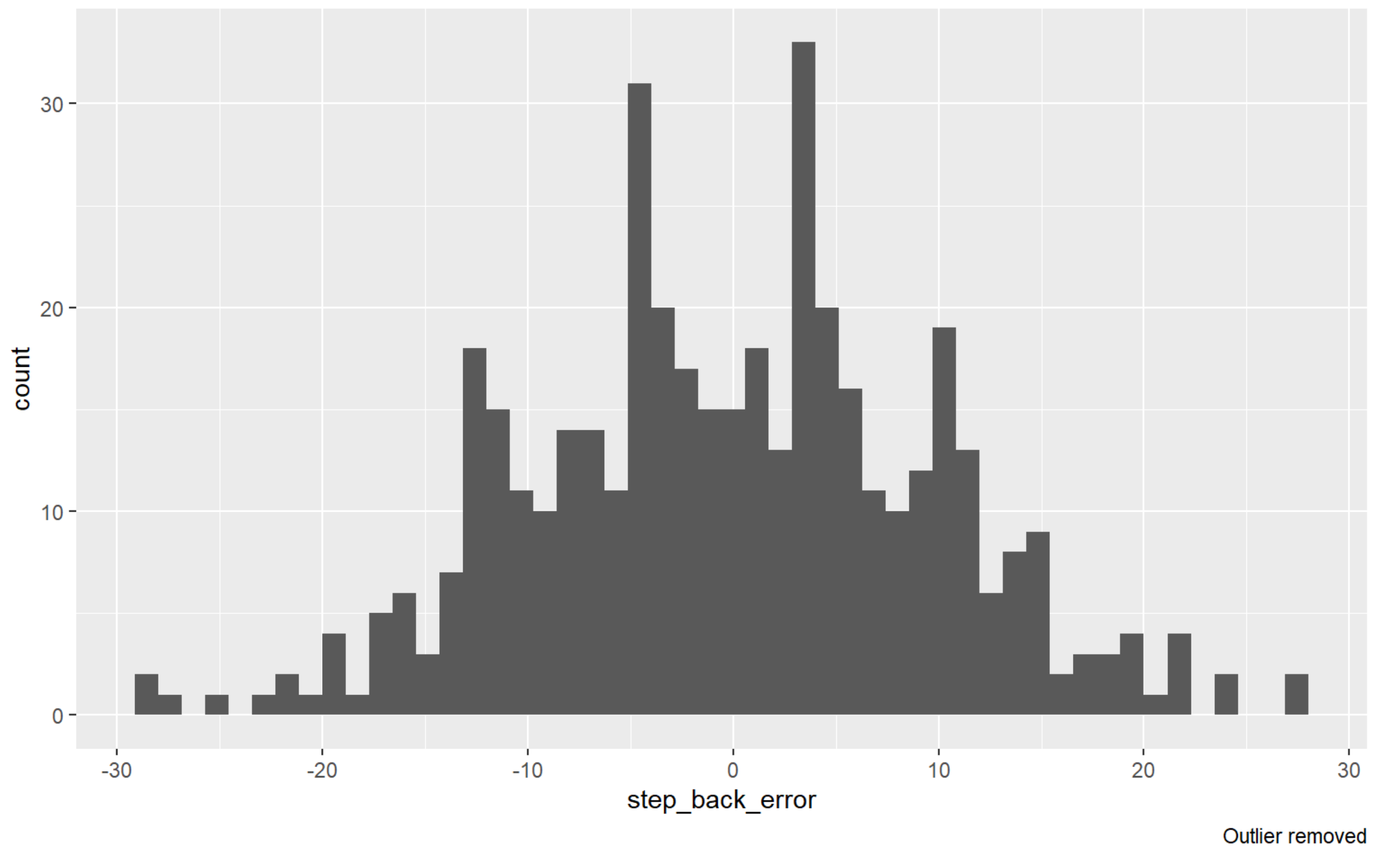

Residuals vs Fitted
Residuals
Fitted values
lm(poly_call[2])
314
1518
205


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(poly_call[2])
1518
314
205


Scale-Location
√|Standardized residuals|
Fitted values
lm(poly_call[2])
205
314
1518



Code



Code

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's 
## -28.00000  -7.00000   0.00000  -0.04147   6.75000  28.00000       247
```

8 CONCLUSION

This report covers an attempt to build a model to predict number of wins of a baseball team in a season based on several offensive and deffensive statistics. Resulting model explained about 36% of variability in the target variable and included most of the provided explanatory variables. Some potentially helpful variables were not included in the data set. For instance, number of At Bats can be used to calculate on-base percentage which may correlate strongly with winning percentage. The model can be revised with additional variables or further analysis.

Code

	kitchen_sink_error	simple_error	step_back_error
	Min. :-28.3735	Min. :-27.2876	Min. :-28.00000
	1st Qu.: -6.9033	1st Qu.: -7.6292	1st Qu.: -7.00000
	Median : -0.1124	Median : 0.2432	Median : 0.00000

	kitchen_sink_error	simple_error	step_back_error
	Mean : -0.0408	Mean : -0.1372	Mean : -0.04147
	3rd Qu.: 6.4889	3rd Qu.: 6.5731	3rd Qu.: 6.75000
	Max. : 27.6495	Max. : 29.6379	Max. : 28.00000
	NA's :247	NA's :143	NA's :247