Home Work Assignment - 01

Critical Thinking Group 5

Contents

Overview

The data set contains 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 will be exploring, analyzing, and modeling the data set to predict a number of wins for a team using Ordinary Least Square (OLS).
To attain our objective, we will be following the below best practice steps and guidelines:
1 -Data Exploration
2 -Data Preparation
3 -Build Models
4 -Select Models

1 Data Exploration Analysis

In section we will explore and gain some insights into the dataset by pursuing the below high level steps and inquiries:
-Variable identification
-Variable Relationships
-Data summary analysis
-Outliers and Missing Values Identification

1.1 Variable identification

First let's display and examine the data dictionary or the data columns as shown in table 1.

Table 1: Variable Definition

VARIABLE_NAME	DEFINITION	THEORETICAL_EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_WINS	Number of wins	Target
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

We notice that all variables are numeric. The variable names seem to follow certain naming pattern to highlight certain arithmetic relationships. In other words, we can compute the number of '1B' hits by taking the difference between overall hits and '2B', '3B', 'HR'. Although such naming and construct is not recommended in normalized database design (as it violates third normal form), it is very frequent practice in the data analytics.

Our predictor input is made of 15 variables. And our dependent variable is one variable called TARGET_WINS.

Please note that we will not be using INDEX variable as it serves as just an identifier for each row. And has no relationships to other variables.

1.2 Data Summary Analysis

In this section, we will create summary data to better understand the initial relationship variables have with our dependent variable using correlation, central tendency, and dispersion As shown in table 2.

Table 2: Data Summary

	mean	sd	median	trimmed
TARGET_WINS	80.79086	15.75215	82.0	81.31229
TEAM_BATTING_H	1469.26977	144.59120	1454.0	1459.04116
TEAM_BATTING_2B	241.24692	46.80141	238.0	240.39627
TEAM_BATTING_3B	55.25000	27.93856	47.0	52.17563
TEAM_BATTING_HR	99.61204	60.54687	102.0	97.38529
TEAM_BATTING_BB	501.55888	122.67086	512.0	512.18331
TEAM_BATTING_SO	735.60534	248.52642	750.0	742.31322
TEAM_BASERUN_SB	124.76177	87.79117	101.0	110.81188
TEAM_BASERUN_CS	52.80386	22.95634	49.0	50.35963
TEAM_BATTING_HBP	59.35602	12.96712	58.0	58.86275
TEAM_PITCHING_H	1779.21046	1406.84293	1518.0	1555.89517
TEAM_PITCHING_HR	105.69859	61.29875	107.0	103.15697
TEAM_PITCHING_BB	553.00791	166.35736	536.5	542.62459
TEAM_PITCHING_SO	817.73045	553.08503	813.5	796.93391
TEAM_FIELDING_E	246.48067	227.77097	159.0	193.43798
TEAM_FIELDING_DP	146.38794	26.22639	149.0	147.57789

Table 3: Missing Data and Data Correlation

	Missing	Correlation
TARGET_WINS	0	1.0000000
TEAM_BATTING_H	0	0.3887675
TEAM_BATTING_2B	0	0.2891036
TEAM_BATTING_3B	0	0.1426084
TEAM_BATTING_HR	0	0.1761532
TEAM_BATTING_BB	0	0.2325599
TEAM_BATTING_SO	102	-0.0317507
TEAM_BASERUN_SB	131	0.1351389
TEAM_BASERUN_CS	772	0.0224041
TEAM_BATTING_HBP	2085	0.0735042
TEAM_PITCHING_H	0	-0.1099371
TEAM_PITCHING_HR	0	0.1890137
TEAM_PITCHING_BB	0	0.1241745
TEAM_PITCHING_SO	102	-0.0784361
TEAM_FIELDING_E	0	-0.1764848
TEAM_FIELDING_DP	286	-0.0348506

Based on table 2 and Table 3, we can make the below observations:

1.Some of the variables like TEAM_PITCHING_H, TEAM_PITCHING_SO and TEAM_FIELDING_E seem to have outliers which is evident from the mean, median and trimmed mean values.

2.TEAM_BATTING_HBP and TEAM_BASERUN_CS seems to be missing a lot of values which casts doubt on its usefulness as a predictor. Maybe a flag for presense or absense of TEAM_BATTING_HBP and TEAM_BASERUN_CS might be a better predictor. Also given the fact that there is low correlation, we decided to exclude these 2 variables from any missing value or outlier treatment.

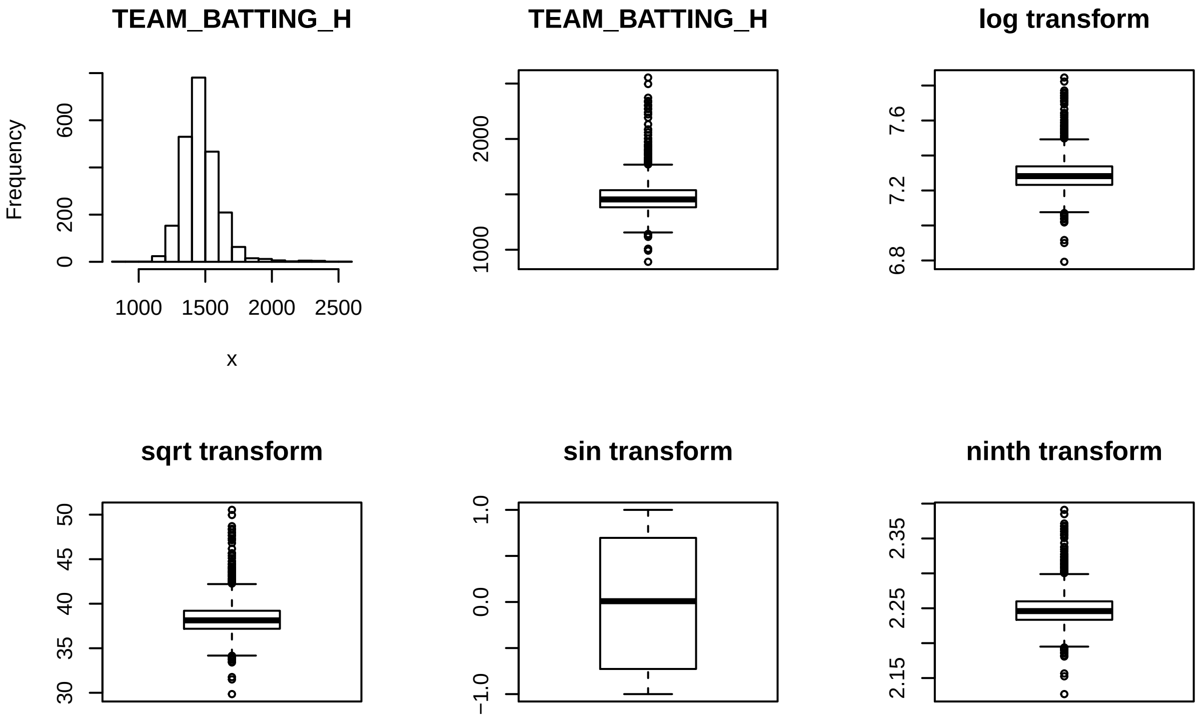
3.Most of the variables seem to indicate a positive / negative correlation in line with the theoretical effect. However, the following stand out as they show a correlation opposite to the theoretical impact: TEAM_BASERUN_CS, TEAM_PITCHING_HR, TEAM_PITCHING_BB, TEAM_PITCHING_SO and TEAM_FIELDING_DP. Lets evaluate these variables further once we fix any missing values or outliers.

4. We will impute the missing values in TEAM_BATTING_SO, FIELDING_DP, BASERUN_SB and TEAM_PITCHING_SO since it has lesser missing values even though there is low correlation. So we will create new variables that will have the respective missing values handled.

1.3 Outliers and Missing Values Identification

In this section we look at boxplots to determine the outliers in variables and decide on whether to act on the outliers.

Lets do some univariate analysis. We will look at the Histogram and Boxplot for each variable to detect outliers if any and treat it accordingly.



***Please note that we have created similar figures to figure 1 above for each remaining variable. However, we hid the remaining figures for ease of streamlining the report as they have similar shapes. However, we have drawn the below observations from each remaining figure.

- For TEAM_BATTING_H, we can see that there are quite a few outliers, both at the upper and lower end. Accordingly, we decide to create a new variable that will have the outlier fixed.

- For TEAM_BATTING_2B, we can see that there are quite a few outliers, both at the upper and a single outlier at the lower end. For this variable we decide to create a new variable that will have the outliers fixed.
- For TEAM_BATTING_3B, we can see that there are quite a few outliers at the upper end. For this variable we decide to create a new variable that will have the outliers fixed.
- For TEAM_BATTING_HR, we can see that there are no outliers.
- For TEAM_BATTING_BB, we can see that there are quite a few outliers, both at the upper and lower end. For this variable we decide to create a new variable that will have the outlier fixed.
- For TEAM_BATTING_SO, we can see that there are no outliers. No further action needed for this variable.
- For TEAM_BASERUN_SB, we can see that there are quite a few outliers at the upper end. For this variable we decide to create a new variable that will have the outlier fixed.
- For TEAM_FIELDING_E, we can see that there are quite a few outliers at the upper end. For this variable we decide to create a new variable that will have the outlier fixed.
- For TEAM_FIELDING_DP, we can see that there are quite a few outliers, both at the upper and lower end. For this variable we decide to create a new variable that will have the outlier fixed.
- For TEAM_PITCHING_BB, we can see that there are quite a few outliers, both at the upper and lower end. For this variable we decide to create a new variable that will have the outlier fixed.
- For TEAM_PITCHING_H, we can see that there are quite a few outliers at the upper end. For this variable we decide to create a new variable that will have the outlier fixed.
- For TEAM_PITCHING_HR, we can see that there only 3 outliers at the upper end. For this variable we decide to create a new variable that will have the outlier fixed.
- For TEAM_PITCHING_SO, we can see that there are quite a few outliers at the upper and a single outlier on the lower end. For this variable we decide to create a new variable that will have the outlier fixed.

Please note that, in most of the cases above, we see that a SIN transformation seems to work well to take care of the outliers. We will go ahead and create these new variables respectively.

2. Data Preparation

Now that we have completed the preliminary analysis, we will be cleaning and consolidating data into one dataset for use in analysis and modeling. We will be puring the belwo steps as guidlines:
- Outliers treatment
- Missing values treatment
- Data transformation

2.1 Outliers treatment

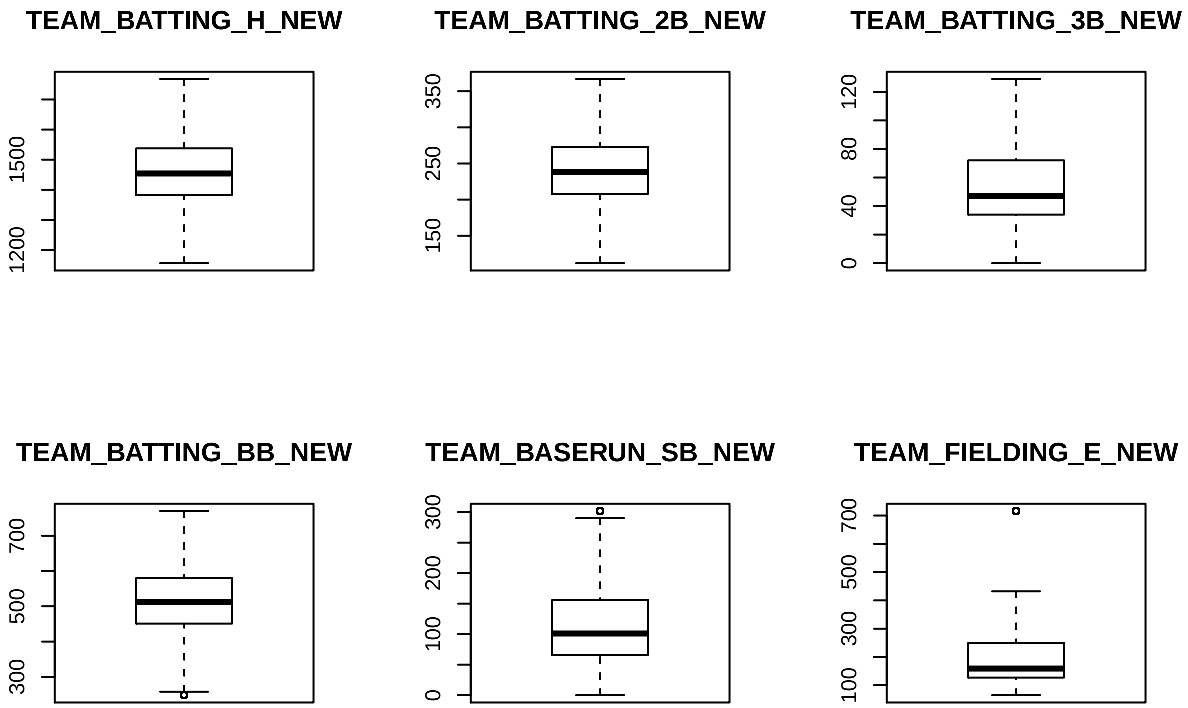
For outliers, we will create 2 sets of variables.

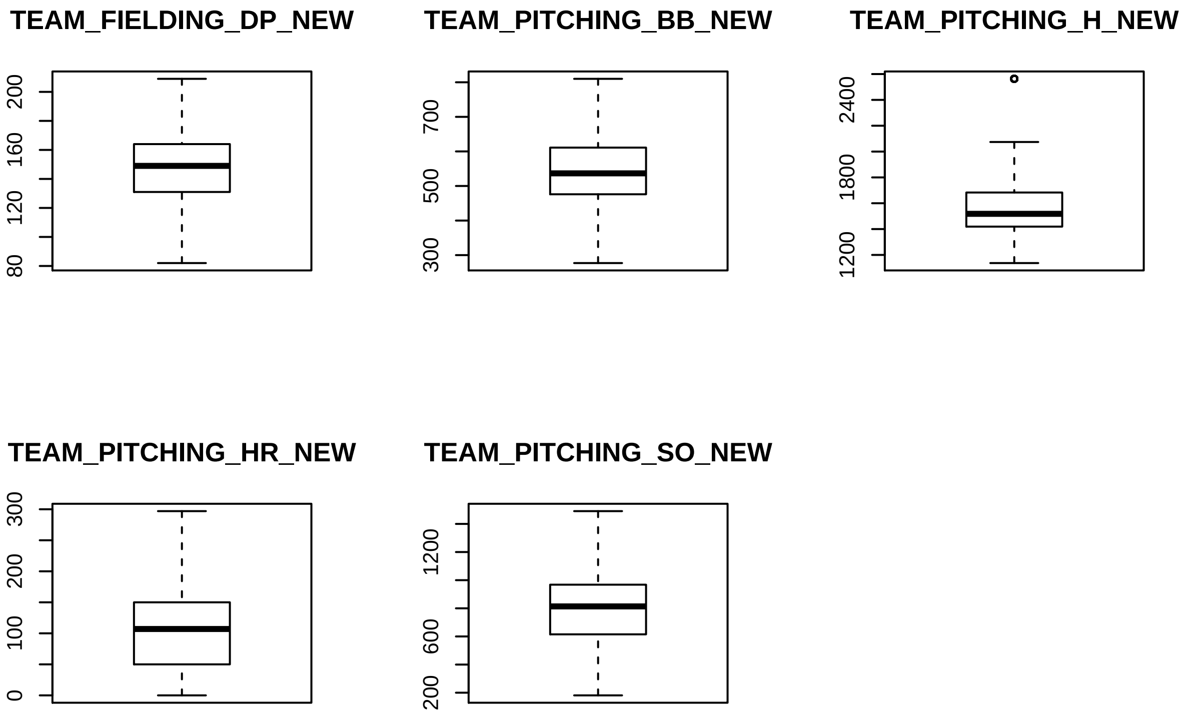
The first set uses the capping method. In this method, we will replace all outliers that lie outside the 1.5 times of IQR limits. We will cap it by replacing those observations less than the lower limit with the value of 5th %ile and those that lie above the upper limit with the value of 95th %ile.

Accordingly we create the following new variables while retaining the original variables.

TEAM_BATTING_H_NEW
TEAM_BATTING_2B_NEW
TEAM_BATTING_3B_NEW
TEAM_BATTING_BB_NEW
TEAM_BASERUN_SB_NEW
TEAM_FIELDING_E_NEW
TEAM_FIELDING_DP_NEW
TEAM_PITCHING_BB_NEW
TEAM_PITCHING_H_NEW
TEAM_PITCHING_HR_NEW
TEAM_PITCHING_SO_NEW

Lets see how the new variables look in boxplots.


TEAM_BATTING_H_NEW
1200
1500
TEAM_BATTING_2B_NEW
150
250
350
TEAM_BATTING_3B_NEW
0
40
80
120
TEAM_BATTING_BB_NEW
300
500
700
TEAM_BASERUN_SB_NEW
0
100
200
300
TEAM_FIELDING_E_NEW
100
300
500
700


TEAM_FIELDING_DP_NEW

TEAM_PITCHING_BB_NEW

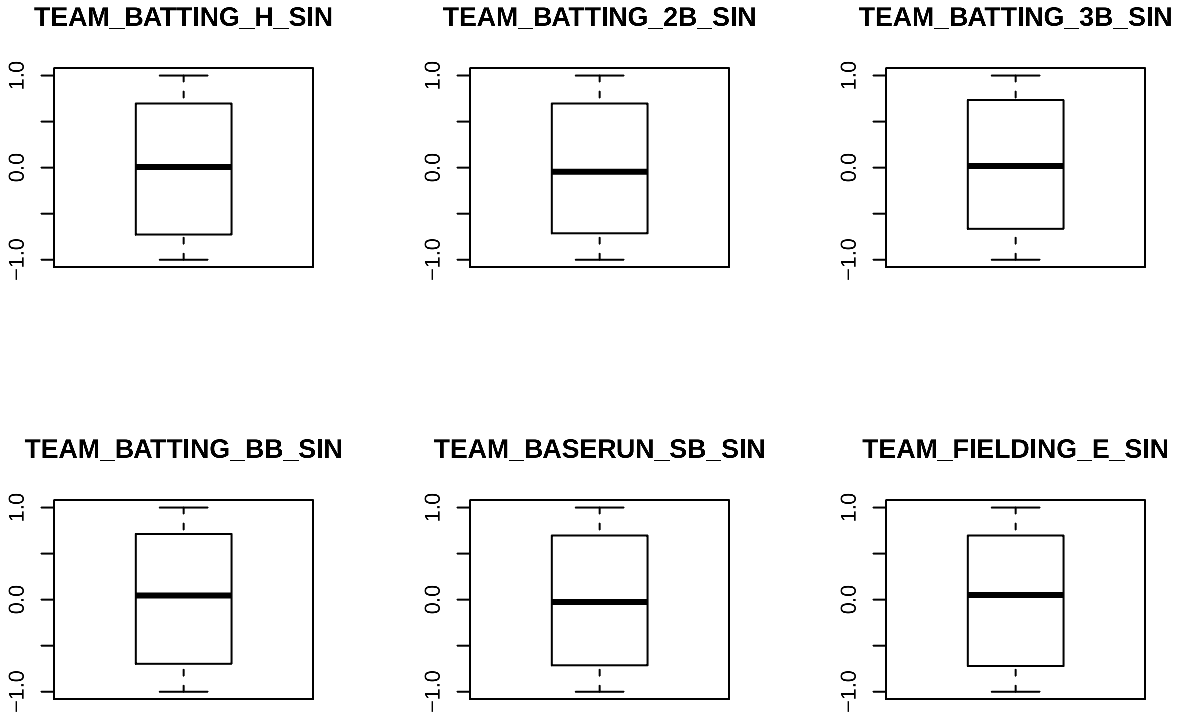
TEAM_PITCHING_H_NEW

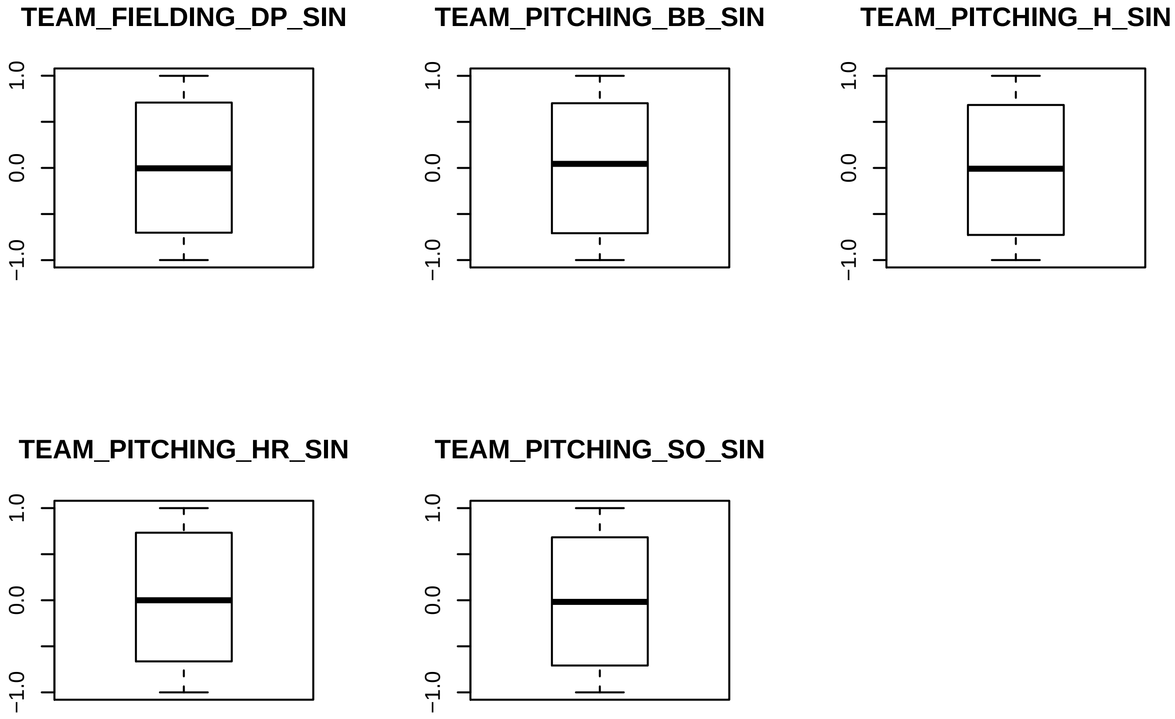
TEAM_PITCHING_HR_NEW

TEAM_PITCHING_SO_NEW

In the second set, we will use the sin transformation and create the following variables:

TEAM_BATTING_H_SIN
TEAM_BATTING_2B_SIN
TEAM_BATTING_3B_SIN
TEAM_BATTING_BB_SIN
TEAM_BASERUN_SB_SIN
TEAM_FIELDING_E_SIN
TEAM_FIELDING_DP_SIN
TEAM_PITCHING_BB_SIN
TEAM_PITCHING_H_SIN
TEAM_PITCHING_HR_SIN
TEAM_PITCHING_SO_SIN


TEAM_BATTING_H_SIN
TEAM_BATTING_2B_SIN
TEAM_BATTING_3B_SIN
TEAM_BATTING_BB_SIN
TEAM_BASERUN_SB_SIN
TEAM_FIELDING_E_SIN
1.0
0.0
−1.0



2.2 Missing values treatment

Next we impute missing values. Since we have handled outliers, we can go ahead and use the mean as impute values. As with outliers, we will go ahead and create new variables for the following:

TEAM_BATTING_SO_NEW

We will re-use the already created new variables for fixing the missing values for the below:

TEAM_PITCHING_SO_NEW
TEAM_BASERUN_SB_NEW
TEAM_FIELDING_DP_NEW

Lets now create some additional variables that might help us in out analysis.

2.3 Missing Flags

First we create flag variables to indicate whether TEAM_BATTING_HBP and TEAM_BASERUN_CS and missing. If the value is missing, we code it with 0 and if the value is present we code it with 1.
We will name our missing flag variables as follow:
TEAM_BATTING_HBP_Missing
TEAM_BASERUN_CS_Missing

2.4 Ratios

Next we create some additional variables, that we think may be useful with the prediction. Here we create the following ratios:
Hits_R = TEAM_BATTING_H/TEAM_PITCHING_H
Walks_R = TEAM_BATTING_BB/TEAM_PITCHING_BB
HomeRuns_R = TEAM_BATTING_HR/TEAM_PITCHING_HR
Strikeout_R = TEAM_BATTING_SO/TEAM_PITCHING_SO

2.5 Calculated Variables

Finally, we will also create calculated variables as below:

1. TEAM_BATTING_EB (Extra Base Hits) = 2B + 3B + HR
2. TEAM_BATTING_1B (Singles by batters) = TEAM_BATTING_H - TEAM_BATTING_EB

2.6 Correlation for new variables

Lets see how the new variables stack up against wins.

```
## TEAM_BATTING_HBP_Missing  TEAM_BASERUN_CS_Missing                   Hits_R
##              0.002610647              0.004864215              0.095800033
##                  Walks_R               HomeRuns_R              Strikeout_R
##              0.083660245              0.013440964              0.063193881
##          TEAM_BATTING_EB          TEAM_BATTING_1B
##              0.344958150              0.217430135
```

All new variables seem to have a positive correlation with wins. However, some of them do not seem to have a strong correlation. Lets see how they perform while modeling.

3 Build Models

In this phase, we will build four models. The models independent variables will be based initially on the original data set variables, derived dataset variables, transformed dataset variables, and all variables in the dataset. In addition, for each model, we will perform a stepwise selection and stop at a point where we retain only those variables that have lower AIC (Akaike An Information Criterion). Recall (AIC) is a measure of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Lower AIC leads to better quality model.

Below is a summary table showing models and their respective variables.

VARIABLE_NAME	Comments	Theoretical.Effect	Model1	Model2	Model3	Model4
TEAM_BATTING_H	Given	Positive	Y			Y
TEAM_BATTING_2B	Given	Positive	Y			Y
TEAM_BATTING_3B	Given	Positive	Y			Y
TEAM_BATTING_HR	Given	Positive	Y			Y
TEAM_BATTING_BB	Given	Positive	Y			Y
TEAM_BATTING_HBP	Given	Positive				
TEAM_BATTING_SO	Given	Negative	Y			Y
TEAM_BASERUN_SB	Given	Positive	Y			Y
TEAM_BASERUN_CS	Given	Negative				
TEAM_FIELDING_E	Given	Negative	Y			Y
TEAM_FIELDING_DP	Given	Positive	Y			Y
TEAM_PITCHING_BB	Given	Negative	Y			Y
TEAM_PITCHING_H	Given	Negative	Y			Y
TEAM_PITCHING_HR	Given	Negative	Y			Y
TEAM_PITCHING_SO	Given	Positive	Y			Y
TEAM_BATTING_H_NEW	Derived	Positive		Y		Y
TEAM_BATTING_2B_NEW	Derived	Positive		Y		Y
TEAM_BATTING_3B_NEW	Derived	Positive		Y		Y
TEAM_BATTING_BB_NEW	Derived	Positive		Y		Y
TEAM_BASERUN_SB_NEW	Derived	Positive		Y		Y
TEAM_FIELDING_E_NEW	Derived	Negative		Y		Y
TEAM_FIELDING_DP_NEW	Derived	Positive		Y		Y
TEAM_PITCHING_BB_NEW	Derived	Negative		Y		Y
TEAM_PITCHING_H_NEW	Derived	Negative		Y		Y
TEAM_PITCHING_HR_NEW	Derived	Negative		Y		Y
TEAM_PITCHING_SO_NEW	Derived	Positive		Y		Y
TEAM_BATTING_H_SIN	Derived	Positive			Y	Y
TEAM_BATTING_2B_SIN	Derived	Positive			Y	Y
TEAM_BATTING_3B_SIN	Derived	Positive			Y	Y
TEAM_BATTING_BB_SIN	Derived	Positive			Y	Y
TEAM_BASERUN_SB_SIN	Derived	Positive			Y	Y
TEAM_FIELDING_E_SIN	Derived	Negative			Y	Y
TEAM_FIELDING_DP_SIN	Derived	Positive			Y	Y
TEAM_PITCHING_BB_SIN	Derived	Negative			Y	Y
TEAM_PITCHING_H_SIN	Derived	Negative			Y	Y
TEAM_PITCHING_HR_SIN	Derived	Negative			Y	Y
TEAM_PITCHING_SO_SIN	Derived	Positive			Y	Y
TEAM_BATTING_HBP_Missing	Derived				Y	Y
TEAM_BASERUN_CS_Missing	Derived				Y	Y
Hits_R	Derived				Y	Y
Walks_R	Derived				Y	Y
HomeRuns_R	Derived				Y	Y
Strikeout_R	Derived				Y	Y
TEAM_BATTING_EB	Derived				Y	Y
TEAM_BATTING_1B	Derived				Y	Y

3.1 Model One

In this model, we will be using the original variables. We will create model and we will highlight the variables that being recommended using the AIC value.
First we will produce the summary model as per below:

```
##
## 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_FIELDING_E + TEAM_FIELDING_DP + TEAM_PITCHING_BB +
##     TEAM_PITCHING_H + TEAM_PITCHING_HR + TEAM_PITCHING_SO, data = na.omit(data))
##
## 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_FIELDING_E  -0.119149   0.007145 -16.676  < 2e-16 ***
## TEAM_FIELDING_DP -0.112120   0.012280  -9.131  < 2e-16 ***
## TEAM_PITCHING_BB -0.075474   0.040427  -1.867  0.06207 .
## TEAM_PITCHING_H   0.057619   0.014949   3.854  0.00012 ***
## TEAM_PITCHING_HR -0.040017   0.077904  -0.514  0.60754
## TEAM_PITCHING_SO -0.046960   0.020918  -2.245  0.02489 *
## ---
## 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
```

Next, we will step thru this model (model 1) and retain only those variables that have the most impact.

Based on the backward stepwise selection, below are the characteristics of the refined model :

- The Residual standard error is 10.18
- Multiple R-squared: 0.4058
- Adjusted R-squared: 0.4019
- F-statistic: 103.7 on 12 and 1822 DF
- p-value: < 2.2e-16

Table 5: Coefficients for the refined model 1

	Coefficients
(Intercept)	59.0548324
TEAM_BATTING_H	-0.0338435
TEAM_BATTING_2B	-0.0492679
TEAM_BATTING_3B	0.1834965
TEAM_BATTING_HR	0.1002629
TEAM_BATTING_BB	0.1183635
TEAM_BATTING_SO	0.0333161
TEAM_BASERUN_SB	0.0694647
TEAM_FIELDING_E	-0.1188641
TEAM_FIELDING_DP	-0.1123169
TEAM_PITCHING_BB	-0.0803085
TEAM_PITCHING_H	0.0598130
TEAM_PITCHING_SO	-0.0535232

Based on the above coefficients, we can see that some of the coefficients are counter-intutive to the Theoretical impact.

- TEAM_BATTING_H (-0.034), TEAM_BATTING_2B (-0.049), TEAM_FIELDING_DP (-0.112), TEAM_PITCHING_SO (-0.054) have a negative coefficient even though they are theoretically supposed to have a positive impact on wins. This means that a unit change in each of these variables will decrease the number of a wins.
- Similarly, TEAM_BATTING_SO (0.033), TEAM_PITCHING_H (0.06) have a positive coefficient even though they are theoretically supposed to have a negative impact on wins. This means that a unit change in each of these variables will increase the number of a wins.
- TEAM_BATTING_3B (0.183), TEAM_BATTING_HR (0.1), TEAM_BATTING_BB (0.118), TEAM_BASERUN_SB (0.069), TEAM_FIELDING_E (-0.119), TEAM_PITCHING_BB (-0.08) have the intended theoretical impact on wins. This means that a unit change in each of these variables will either decrease or increase the number of a wins as intended by the theoretical impact.

Since we have already seen this result in our data exploration phase, we will retain this model as is for comparision with other models.

3.2 Model Two

In this model (model2), we will be using the adjusted values based on our outlier treatment process. We will create model and we will highlight the variables that being recommended using the AIC value. First we will produce the summary model as per below:

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H_NEW + TEAM_BATTING_2B_NEW +
##     TEAM_BATTING_3B_NEW + TEAM_BATTING_BB_NEW + TEAM_BASERUN_SB_NEW +
##     TEAM_FIELDING_E_NEW + TEAM_FIELDING_DP_NEW + TEAM_PITCHING_BB_NEW +
##     TEAM_PITCHING_H_NEW + TEAM_PITCHING_HR_NEW + TEAM_PITCHING_SO_NEW,
##     data = na.omit(data))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.032  -8.396   0.269   8.411  70.493
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          21.9620608  6.0558154   3.627 0.000294 ***
## TEAM_BATTING_H_NEW    0.0260878  0.0051203   5.095 3.78e-07 ***
## TEAM_BATTING_2B_NEW  -0.0003544  0.0096157  -0.037 0.970603
## TEAM_BATTING_3B_NEW   0.1257703  0.0182053   6.908 6.35e-12 ***
## TEAM_BATTING_BB_NEW   0.0511574  0.0083740   6.109 1.18e-09 ***
## TEAM_BASERUN_SB_NEW   0.0442051  0.0055102   8.022 1.65e-15 ***
## TEAM_FIELDING_E_NEW  -0.0216626  0.0029143  -7.433 1.49e-13 ***
## TEAM_FIELDING_DP_NEW -0.1041769  0.0140840  -7.397 1.95e-13 ***
## TEAM_PITCHING_BB_NEW -0.0314461  0.0074625  -4.214 2.61e-05 ***
## TEAM_PITCHING_H_NEW   0.0103825  0.0020683   5.020 5.58e-07 ***
## TEAM_PITCHING_HR_NEW  0.0751211  0.0089378   8.405  < 2e-16 ***
## TEAM_PITCHING_SO_NEW -0.0055230  0.0020750  -2.662 0.007830 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.42 on 2264 degrees of freedom
## Multiple R-squared:  0.2779, Adjusted R-squared:  0.2744
## F-statistic: 79.21 on 11 and 2264 DF,  p-value: < 2.2e-16
```

Lets now step thru this model and retain only those variables that have the most impact.

Based on the backward stepwise selection, below are the characteristics of the refined model :

- The Residual standard error is 13.42
- Multiple R-squared: 0.2779
- Adjusted R-squared: 0.2747
- F-statistic: 87.16 on 10 and 2265 DF
- p-value: < 2.2e-16

Table 6: Coefficients for the refined model 2

	Coefficients
(Intercept)	22.0443242
TEAM_BATTING_H_NEW	0.0259818
TEAM_BATTING_3B_NEW	0.1258334
TEAM_BATTING_BB_NEW	0.0511472
TEAM_BASERUN_SB_NEW	0.0442132
TEAM_FIELDING_E_NEW	-0.0216441
TEAM_FIELDING_DP_NEW	-0.1041916
TEAM_PITCHING_BB_NEW	-0.0314459
TEAM_PITCHING_H_NEW	0.0103832

	Coefficients
TEAM_PITCHING_HR_NEW	0.0751231
TEAM_PITCHING_SO_NEW	-0.0055425

Based on the above coefficients, we can see that some of the coefficients are counter-intutive to the Theoretical impact.

- TEAM_FIELDING_DP_NEW (-0.104), TEAM_PITCHING_SO_NEW (-0.006) have a negative coefficient even though they are theoretically supposed to have a positive impact on wins. This means that a unit change in each of these variables will decrease the number of a wins.
- Similarly, TEAM_PITCHING_H_NEW (0.01), TEAM_PITCHING_HR_NEW (0.075) have a positive coefficient even though they are theoretically supposed to have a negative impact on wins. This means that a unit change in each of these variables will increase the number of a wins.
- TEAM_BATTING_H_NEW (0.026), TEAM_BATTING_3B_NEW (0.126), TEAM_BATTING_BB_NEW (0.051), TEAM_BASERUN_SB_NEW (0.044), TEAM_FIELDING_E_NEW (-0.022), TEAM_PITCHING_BB_NEW (-0.031) have the intended theoretical impact on wins. This means that a unit change in each of these variables will either decrease or increase the number of a wins as intended by the theoretical impact.

However, since the correlation seems to have a minor impact, we will go ahead and retain this model for further comparision.

3.3 Model Three

In this model (model3), we will be using the derived values based on our variable transformation process. We will create model and we will highlight the variables that being recommended using the AIC value. First we will produce the summary model as per below:

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H_SIN + TEAM_BATTING_2B_SIN +
##     TEAM_BATTING_3B_SIN + TEAM_BATTING_BB_SIN + TEAM_BASERUN_SB_SIN +
##     TEAM_FIELDING_E_SIN + TEAM_FIELDING_DP_SIN + TEAM_PITCHING_BB_SIN +
##     TEAM_PITCHING_H_SIN + TEAM_PITCHING_HR_SIN + TEAM_PITCHING_SO_SIN +
##     TEAM_BATTING_HBP_Missing + TEAM_BASERUN_CS_Missing + Hits_R +
##     Walks_R + HomeRuns_R + Strikeout_R + TEAM_BATTING_EB + TEAM_BATTING_1B,
##     data = na.omit(data))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.340  -8.347   0.419   8.589  38.453
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.085e+01  6.556e+00   3.180   0.0015 **
## TEAM_BATTING_H_SIN       4.656e-01  4.279e-01   1.088   0.2767
## TEAM_BATTING_2B_SIN      3.123e-01  3.948e-01   0.791   0.4290
## TEAM_BATTING_3B_SIN     -2.905e-01  4.043e-01  -0.719   0.4725
## TEAM_BATTING_BB_SIN     -7.876e-01  4.296e-01  -1.834   0.0669 .
## TEAM_BASERUN_SB_SIN     -6.811e-01  4.048e-01  -1.682   0.0927 .
## TEAM_FIELDING_E_SIN      -2.182e-01  4.022e-01  -0.542   0.5876
## TEAM_FIELDING_DP_SIN     -9.013e-02  3.986e-01  -0.226   0.8211
## TEAM_PITCHING_BB_SIN      5.520e-01  4.334e-01   1.274   0.2029
## TEAM_PITCHING_H_SIN      -1.649e-02  4.309e-01  -0.038   0.9695
## TEAM_PITCHING_HR_SIN     -4.350e-01  4.028e-01  -1.080   0.2803
## TEAM_PITCHING_SO_SIN      3.438e-01  3.985e-01   0.863   0.3884
## TEAM_BATTING_HBP_Missing -5.576e+00  1.070e+00  -5.213 2.07e-07 ***
## TEAM_BASERUN_CS_Missing  -2.016e+00  8.386e-01  -2.405   0.0163 *
## Hits_R                   -5.752e+02  1.034e+03  -0.557   0.5779
## Walks_R                  -8.704e+02  7.047e+02  -1.235   0.2170
## HomeRuns_R                6.358e+01  6.320e+01   1.006   0.3145
## Strikeout_R               1.389e+03  8.584e+02   1.619   0.1057
## TEAM_BATTING_EB           7.352e-02  4.457e-03  16.497  < 2e-16 ***
## TEAM_BATTING_1B           2.391e-02  3.585e-03   6.668 3.42e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.06 on 1815 degrees of freedom
## Multiple R-squared:  0.1682, Adjusted R-squared:  0.1595
## F-statistic: 19.32 on 19 and 1815 DF,  p-value: < 2.2e-16
```

Lets now step thru this model and retain only those variables that have the most impact.

Based on the backward stepwise selection, below are the characteristics of the refined model :

- The Residual standard error is 12.05
- Multiple R-squared: 0.1648
- Adjusted R-squared: 0.1612
- F-statistic: 45.05 on 8 and 1826 DF
- p-value: < 2.2e-16

Table 7: Coefficients for the refined model 3

	Coefficients
(Intercept)	20.7930798
TEAM_BATTING_BB_SIN	-0.5872247
TEAM_BASERUN_SB_SIN	-0.7099891
TEAM_BATTING_HBP_Missing	-5.6654377
TEAM_BASERUN_CS_Missing	-2.0192389
Walks_R	-1074.4800683
Strikeout_R	1081.7869878
TEAM_BATTING_EB	0.0736664
TEAM_BATTING_1B	0.0239674

Based on the above coefficients, we can see that some of the coefficients are counter-intutive to the Theoretical impact.

- TEAM_BATTING_BB_SIN (-0.587), TEAM_BASERUN_SB_SIN (-0.71) have a negative coefficient even though they are theoretically supposed to have a positive impact on wins. This means that a unit change in each of these variables will decrease the number of a wins.

- TEAM_BATTING_EB (0.074), TEAM_BATTING_1B (0.024) have the intended theoretical impact on wins. This means that a unit change in each of these variables will either decrease or increase the number of a wins as intended by the theoretical impact.
- The newly derived variables TEAM_BATTING_HBP_Missing (-5.665) and TEAM_BASERUN_CS_Missing (-2.019) seem to a negative impact on wins. This means that a missing value will decrease the number of a wins.
- The newly derived variables, Walks_R (-1074.48), Strikeout_R (1081.787) seem to have a huge impact on the wins. A unit change in each of these variables seems to have a huge impact on the wins.

At this point, we will retain this model as is for comparision with other models.

3.4 Model Four

In this model (model4), we will be using all variables original, adjusted, and derived values. We will create model and we will highlight the variables that being recommended using the AIC value. First we will produce the summary model as per below:

```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = na.omit(data))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.748  -7.039   0.112   6.909  29.178
##
## Coefficients: (2 not defined because of singularities)
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               4.423e+01  4.366e+01   1.013 0.311203
## TEAM_BATTING_H           -1.402e-01  3.878e-02  -3.616 0.000308 ***
## TEAM_BATTING_2B           1.537e-01  6.962e-02   2.207 0.027425 *
## TEAM_BATTING_3B           3.517e-01  2.626e-01   1.339 0.180661
## TEAM_BATTING_HR           1.217e-01  9.643e-02   1.262 0.207083
## TEAM_BATTING_BB           1.635e-01  5.465e-02   2.992 0.002806 **
## TEAM_BATTING_SO           2.215e-02  2.615e-02   0.847 0.397000
## TEAM_BASERUN_SB           1.818e-01  1.186e-01   1.533 0.125539
## TEAM_FIELDING_E          -1.933e-01  2.737e-02  -7.063 2.32e-12 ***
## TEAM_FIELDING_DP         -1.882e-01  1.150e-01  -1.636 0.101954
## TEAM_PITCHING_BB         -8.359e-02  4.401e-02  -1.899 0.057671 .
## TEAM_PITCHING_H           7.938e-02  2.883e-02   2.753 0.005959 **
## TEAM_PITCHING_HR         -7.308e-02  8.605e-02  -0.849 0.395808
## TEAM_PITCHING_SO         -3.346e-02  2.258e-02  -1.482 0.138546
## TEAM_BATTING_H_NEW        9.404e-02  2.862e-02   3.286 0.001037 **
## TEAM_BATTING_2B_NEW      -2.018e-01  7.026e-02  -2.872 0.004120 **
## TEAM_BATTING_3B_NEW      -1.642e-01  2.649e-01  -0.620 0.535506
## TEAM_BATTING_BB_NEW      -4.941e-03  2.864e-02  -0.173 0.863049
## TEAM_BASERUN_SB_NEW      -1.133e-01  1.197e-01  -0.947 0.344003
## TEAM_FIELDING_E_NEW       6.068e-02  2.390e-02   2.539 0.011193 *
## TEAM_FIELDING_DP_NEW      8.459e-02  1.166e-01   0.726 0.468156
## TEAM_PITCHING_BB_NEW     -3.642e-02  2.534e-02  -1.437 0.150797
## TEAM_PITCHING_H_NEW      -6.147e-03  7.775e-03  -0.791 0.429258
## TEAM_PITCHING_HR_NEW      5.246e-02  6.002e-02   0.874 0.382197
## TEAM_PITCHING_SO_NEW     -6.145e-03  1.374e-02  -0.447 0.654713
## TEAM_BATTING_H_SIN        4.414e-01  3.594e-01   1.228 0.219586
## TEAM_BATTING_2B_SIN       8.668e-02  3.310e-01   0.262 0.793439
## TEAM_BATTING_3B_SIN      -1.245e-01  3.411e-01  -0.365 0.715178
## TEAM_BATTING_BB_SIN      -3.094e-01  3.605e-01  -0.858 0.390981
## TEAM_BASERUN_SB_SIN      -6.894e-01  3.391e-01  -2.033 0.042222 *
## TEAM_FIELDING_E_SIN      -1.562e-01  3.377e-01  -0.462 0.643801
## TEAM_FIELDING_DP_SIN     -2.464e-01  3.351e-01  -0.735 0.462272
## TEAM_PITCHING_BB_SIN      5.706e-01  3.629e-01   1.572 0.116039
## TEAM_PITCHING_H_SIN      -1.723e-02  3.603e-01  -0.048 0.961859
## TEAM_PITCHING_HR_SIN     -2.779e-01  3.381e-01  -0.822 0.411094
## TEAM_PITCHING_SO_SIN      4.924e-02  3.358e-01   0.147 0.883432
## TEAM_BATTING_HBP_Missing -2.466e+00  9.666e-01  -2.551 0.010812 *
## TEAM_BASERUN_CS_Missing  -4.111e+00  8.326e-01  -4.938 8.64e-07 ***
## Hits_R                   -1.090e+03  8.744e+02  -1.247 0.212590
## Walks_R                  -3.925e+02  5.960e+02  -0.659 0.510269
## HomeRuns_R                1.590e+01  5.354e+01   0.297 0.766568
## Strikeout_R               1.482e+03  7.228e+02   2.050 0.040488 *
## TEAM_BATTING_EB                  NA         NA      NA       NA
## TEAM_BATTING_1B                  NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.05 on 1793 degrees of freedom
## Multiple R-squared:  0.4294, Adjusted R-squared:  0.4164
## F-statistic: 32.92 on 41 and 1793 DF,  p-value: < 2.2e-16
```

Lets now step thru this model and retain only those variables that have the most impact.

Based on the backward stepwise selection, below are the characteristics of the refined model :

- The Residual standard error is 10.02
- Multiple R-squared: 0.4264
- Adjusted R-squared: 0.4197
- F-statistic: 64.17 on 21 and 1813 DF
- p-value: < 2.2e-16

Table 8: Coefficients for the refined model 4

	Coefficients
(Intercept)	53.0329880
TEAM_BATTING_H	-0.1265832
TEAM_BATTING_2B	0.1537429
TEAM_BATTING_3B	0.1923598
TEAM_BATTING_HR	0.2077476
TEAM_BATTING_BB	0.1604808
TEAM_BASERUN_SB	0.0690665
TEAM_FIELDING_E	-0.1973740

	Coefficients
TEAM_FIELDING_DP	-0.1050952
TEAM_PITCHING_BB	-0.0809532
TEAM_PITCHING_H	0.0623973
TEAM_PITCHING_HR	-0.1034570
TEAM_PITCHING_SO	-0.0185433
TEAM_BATTING_H_NEW	0.0913967
TEAM_BATTING_2B_NEW	-0.2019549
TEAM_FIELDING_E_NEW	0.0634176
TEAM_PITCHING_BB_NEW	-0.0406747
TEAM_BASERUN_SB_SIN	-0.6993951
TEAM_BATTING_HBP_Missing	-2.5394433
TEAM_BASERUN_CS_Missing	-4.1357181
Hits_R	-1458.4638983
Strikeout_R	1465.0397960

Based on the above coefficients, we can see that some of the coefficients are counter-intutive to the Theoretical impact.

- TEAM_BATTING_H (-0.127), TEAM_FIELDING_DP (-0.105), TEAM_PITCHING_SO (-0.019), TEAM_BATTING_2B_NEW (-0.202), TEAM_BASERUN_SB_SIN (-0.699) have a negative coefficient even though they are theoretically supposed to have a positive impact on wins. This means that a unit change in each of these variables will decrease the number of a wins.
- TEAM_PITCHING_H (0.062), TEAM_FIELDING_E_NEW (0.063) has a positive coefficient even though they are theoretically supposed to have a negative impact on wins. This means that a unit change in each of these variables will increase the number of a wins.
- TEAM_BATTING_2B (0.154), TEAM_BATTING_3B (0.192), TEAM_BATTING_HR (0.208), TEAM_BATTING_BB (0.16), TEAM_BASERUN_SB (0.069), TEAM_FIELDING_E (-0.197), TEAM_PITCHING_BB (-0.081), TEAM_PITCHING_HR (-0.103), TEAM_BATTING_H_NEW (0.091), TEAM_PITCHING_BB_NEW (-0.041) have the intended theoretical impact on wins. This means that a unit change in each of these variables will either decrease or increase the number of a wins as intended by the theoretical impact.
- The newly derived variables TEAM_BATTING_HBP_Missing (-2.539), TEAM_BASERUN_CS_Missing (-4.136) seem to a negative impact on wins. This means that a missing value will decrease the number of a wins.
- The newly derived variables, Hits_R (-1458.464), Strikeout_R (1465.04) seem to have a huge impact on the wins. A unit change in each of these variables seems to have a huge impact on the wins.

At this point, we will retain this model as is for comparision with other models and further refining.

4 Model Selection

In section we will further examine all four models. We will apply a model selection strategy by comparing models' AIC, R-squared, and VIF (variance inflation factors).
In addition, we will perform diagnostics to validate the assumption of Linear Regression.

4.1 Model selection strategy:

Following model selection strategy has been used for this assignment:

1. (1) Akaike information criterion (AIC) measure has been used to compare relative performance of different models
2. (2) Along with that of adjusted R^2 values are also used to compare different models performance
3. (3) Different regression model diagnostics plots has been used to test assumptions for regression- (a) test for normality of residuals (b) plot for randomness of residuals, (c) evaluation of homoscedasticity
4. (4) Finally model has been tested for collinearity and enhanced by removing collinearity with the use of variance inflation factors (VIF)

Compare models by AIC measures and adjusted R^2 values Below are the AIC Scores for the 4 models that we built earlier:

Table 9: Model AIC Scores

models	AIC
Model1	13737.06
Model2	18290.75
Model3	14353.81
Model4	13690.51

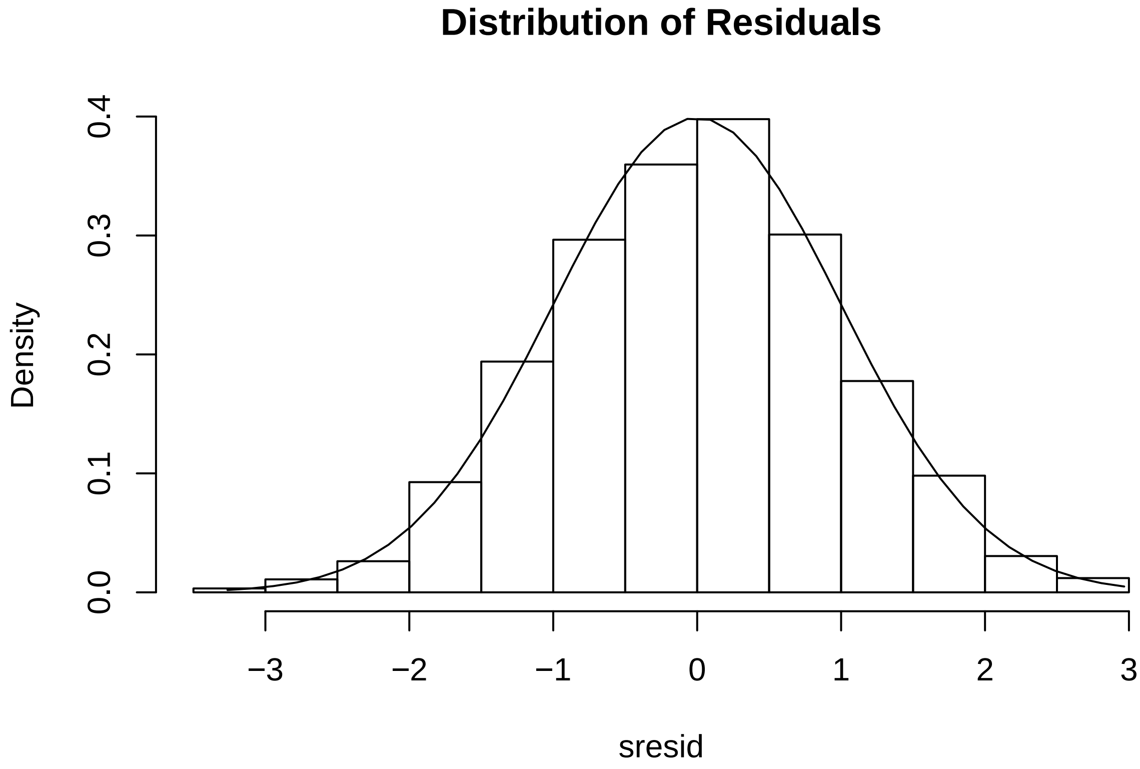
Looking at the AIC values it appears that models, "step1" & "step 4" are comparatively better models of the pack.

Below are the analysis of the adjusted R^2 values:
"step1" has adjusted R^2 value 0.4019 which means this model can explain 40.19% variability in data. "step4" has adjusted R^2 value of 0.4197 and this model can explain 41.97% variability in data. From this two data points model "step4" was picked for further evaluation.

4.2 Model diagnostics

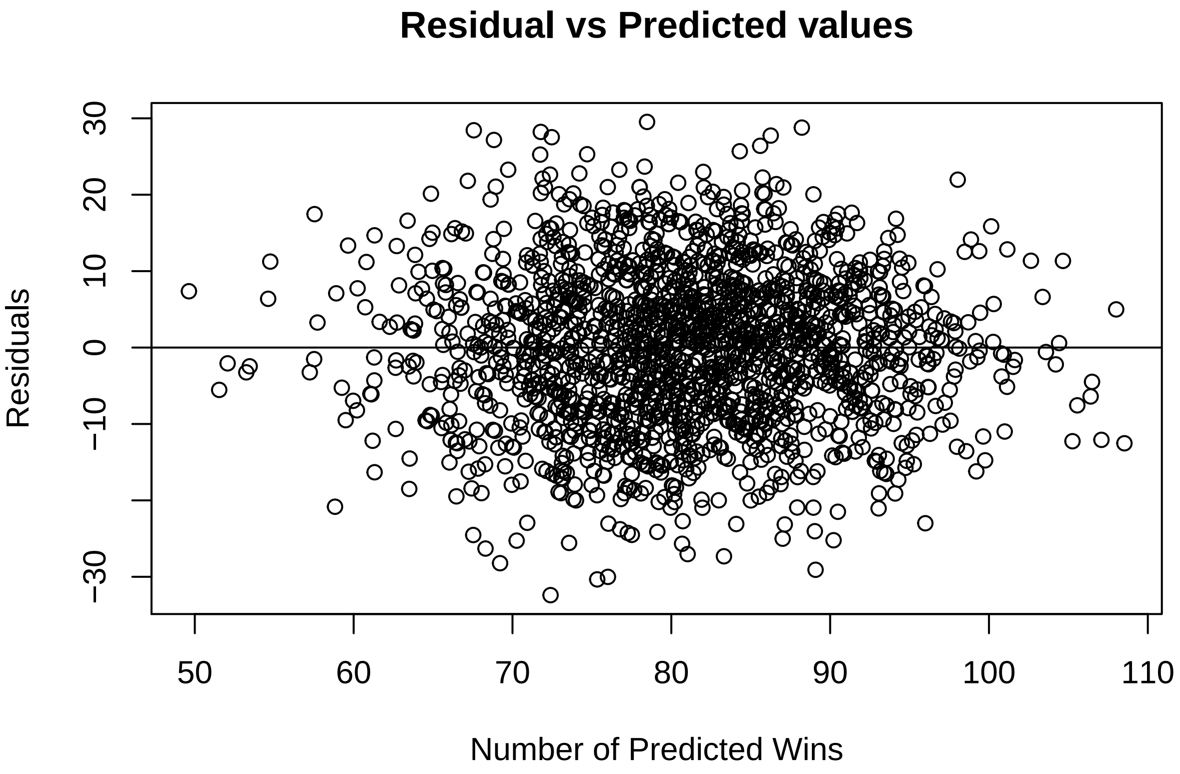
We will create plots to validate the assumption of Linear Regression:

Normality check of residual values:



Based on the normality plot it appears that residual distribution is normal. This indicates the mean of the difference between our predictions.

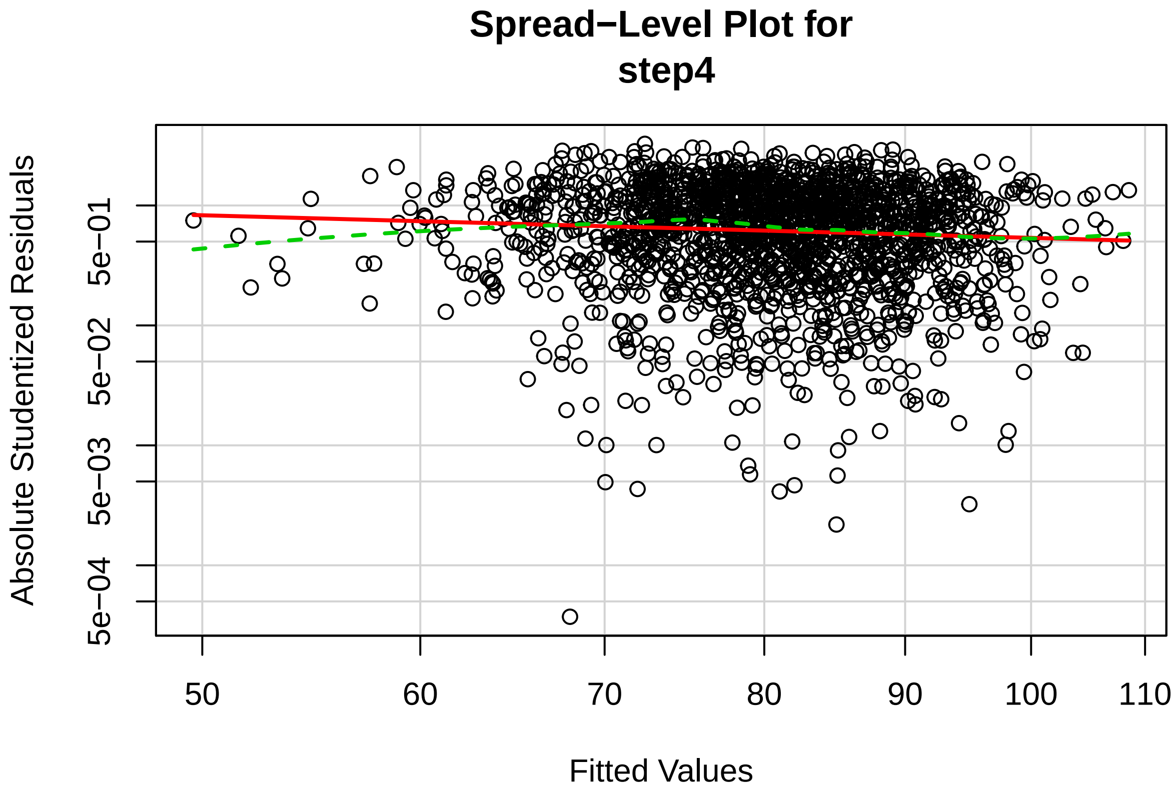
plot residuals with respect to predicted value for randomness:



Distribution of residual values are random around base line and do not show any pattern around base line.

Evaluate homoscedasticity:

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 10.99181    Df = 1     p = 0.0009151536
```



```
##
## Suggested power transformation:  1.626347
```

The test confirms the non-constant error variance test. It also has a p-value higher than a significance level of 0.05.

Analysis of collinearity:

Table 10: Analysis of collinearity

TEAM_BATTING_H	16.889895
TEAM_BATTING_2B	12.700340
TEAM_BATTING_3B	1.763097
TEAM_BATTING_HR	15.511898
TEAM_BATTING_BB	18.849834
TEAM_BASERUN_SB	1.249784
TEAM_FIELDING_E	6.657638
TEAM_FIELDING_DP	1.191288
TEAM_PITCHING_BB	17.122218
TEAM_PITCHING_H	16.437776
TEAM_PITCHING_HR	15.376065
TEAM_PITCHING_SO	2.176509
TEAM_BATTING_H_NEW	12.929265
TEAM_BATTING_2B_NEW	12.666855

TEAM_FIELDING_E_NEW	6.182960
TEAM_PITCHING_BB_NEW	7.639399
TEAM_BASERUN_SB_SIN	1.006246
TEAM_BATTING_HBP_Missing	1.246500
TEAM_BASERUN_CS_Missing	1.376209
Hits_R	187.163514
Strikeout_R	186.803538

Variables have been tested with variance inflation factors (VIF). If any variable has value which is greater than 3 then the highest value variable been removed from model and model performance has been evaluated. Following are the out comes from this assessment steps-

pass 1- Based on that variance inflation factors (VIF) following variable "Hits_R" has highest value < 3 and is removed from model, and model is evaluated without that variable. Adjusted R^2 value has changed from 0.4197 to 0.4187 due to removal of this variable. Hence this variable is not adding lot of value to the model and can be removed.

pass 2- Based on that variance inflation factors (VIF) following variable "TEAM_BATTING_BB" has highest value < 3 and is removed from model, and model is evaluated without that variable. Adjusted R^2 values changed from 0.4187 to 0.4159. Hence this variable is not adding lot of value to the model and can be removed.

pass 3-pass 9- Based on that variance inflation factors (VIF) step 3- step 9 was followed by removing one variable at a time to reduce the VIF measure below 3 for all variable and without compromising too much on model performance(adjusted R^2 value). In final model adjusted R^2 value is 0.4037. That means around 40.37 % variability can be explained by this model. Also all the variables are relevant and having p value less than 0.05.

```
##           TEAM_BATTING_H          TEAM_BATTING_2B          TEAM_BATTING_3B
##                16.863430                12.698393                 1.763056
##          TEAM_BATTING_HR          TEAM_BATTING_BB          TEAM_BASERUN_SB
##                15.501371                18.849596                 1.247755
##          TEAM_FIELDING_E         TEAM_FIELDING_DP         TEAM_PITCHING_BB
##                 6.657283                 1.191067                17.121561
##          TEAM_PITCHING_H         TEAM_PITCHING_HR         TEAM_PITCHING_SO
##                16.395815                15.362614                 2.174531
##       TEAM_BATTING_H_NEW      TEAM_BATTING_2B_NEW      TEAM_FIELDING_E_NEW
##                12.928182                12.663764                 6.182897
##     TEAM_PITCHING_BB_NEW      TEAM_BASERUN_SB_SIN TEAM_BATTING_HBP_Missing
##                 7.639379                 1.006246                 1.246453
##  TEAM_BASERUN_CS_Missing              Strikeout_R
##                 1.375620                10.545014
```

```
##         TEAM_BATTING_3B         TEAM_BASERUN_SB         TEAM_FIELDING_E
##                1.727409                1.224382                1.767441
##        TEAM_FIELDING_DP        TEAM_PITCHING_HR        TEAM_PITCHING_SO
##                1.182640                2.069750                2.066239
##      TEAM_BATTING_H_NEW     TEAM_BATTING_2B_NEW    TEAM_PITCHING_BB_NEW
##                1.949900                1.597064                1.192712
##     TEAM_BASERUN_SB_SIN TEAM_BASERUN_CS_Missing             Strikeout_R
##                1.003535                1.329079                1.284631
```

```
##
## Call:
```

```
## lm(formula = TARGET_WINS ~ . - Hits_R - TEAM_BATTING_BB - TEAM_BATTING_H -
##     TEAM_BATTING_HR - TEAM_BATTING_2B - TEAM_PITCHING_H - TEAM_PITCHING_BB -
##     TEAM_FIELDING_E_NEW - TEAM_BATTING_HBP_Missing, data = data_step4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.740  -7.022   0.108   7.101  28.685
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             40.283276   8.906321   4.523 6.49e-06 ***
## TEAM_BATTING_3B          0.172848   0.018879   9.155  < 2e-16 ***
## TEAM_BASERUN_SB          0.071631   0.005507  13.008  < 2e-16 ***
## TEAM_FIELDING_E         -0.120338   0.007257 -16.583  < 2e-16 ***
## TEAM_FIELDING_DP        -0.106677   0.012368  -8.625  < 2e-16 ***
## TEAM_PITCHING_HR         0.094122   0.008705  10.812  < 2e-16 ***
## TEAM_PITCHING_SO        -0.019500   0.002206  -8.839  < 2e-16 ***
## TEAM_BATTING_H_NEW       0.032683   0.004342   7.528 8.05e-14 ***
## TEAM_BATTING_2B_NEW     -0.056302   0.008926  -6.307 3.55e-10 ***
## TEAM_PITCHING_BB_NEW     0.033202   0.003093  10.736  < 2e-16 ***
## TEAM_BASERUN_SB_SIN     -0.668686   0.339924  -1.967 0.049316 *
## TEAM_BASERUN_CS_Missing -4.062390   0.803315  -5.057 4.69e-07 ***
## Strikeout_R             17.050630   4.940275   3.451 0.000571 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.16 on 1822 degrees of freedom
##   (441 observations deleted due to missingness)
## Multiple R-squared:  0.4076, Adjusted R-squared:  0.4037
## F-statistic: 104.5 on 12 and 1822 DF,  p-value: < 2.2e-16
```

Final model was derived after number of iterations of variable eliminations were carried out. VIF values in the final model among variables < 3. In this scenario a model with slightly less performance was selected to avoid collinearity effect among variables and reduced complexity. Final model all the variables are relevant and having p value less than 0.05.

Test Data

We will now run the final model on the test data.
Prior to this, we still have to carry out the same transformations that we did to the train dataset to the test dataset. Below is a quick summary after the transformations:

```
##  TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B  TEAM_BATTING_HR
##  Min.   : 819   Min.   : 44.0   Min.   : 14.00   Min.   :  0.00
##  1st Qu.:1387   1st Qu.:210.0   1st Qu.: 35.00   1st Qu.: 44.50
##  Median :1455   Median :239.0   Median : 52.00   Median :101.00
##  Mean   :1469   Mean   :241.3   Mean   : 55.91   Mean   : 95.63
##  3rd Qu.:1548   3rd Qu.:278.5   3rd Qu.: 72.00   3rd Qu.:135.50
##  Max.   :2170   Max.   :376.0   Max.   :155.00   Max.   :242.00
##
##  TEAM_BATTING_BB TEAM_BASERUN_SB TEAM_FIELDING_E  TEAM_FIELDING_DP
##  Min.   : 15.0   Min.   :  0.0   Min.   :  73.0   Min.   : 69.0
##  1st Qu.:436.5   1st Qu.: 59.0   1st Qu.: 131.0   1st Qu.:131.0
##  Median :509.0   Median : 92.0   Median : 163.0   Median :148.0
##  Mean   :499.0   Mean   :123.7   Mean   : 249.7   Mean   :146.1
##  3rd Qu.:565.5   3rd Qu.:151.8   3rd Qu.: 252.0   3rd Qu.:164.0
##  Max.   :792.0   Max.   :580.0   Max.   :1568.0   Max.   :204.0
##                  NA's   :13                       NA's   :31
##  TEAM_PITCHING_BB TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_SO
##  Min.   : 136.0   Min.   : 1155   Min.   :  0.0    Min.   :   0.0
##  1st Qu.: 471.0   1st Qu.: 1426   1st Qu.: 52.0    1st Qu.: 613.0
##  Median : 526.0   Median : 1515   Median :104.0    Median : 745.0
##  Mean   : 552.4   Mean   : 1813   Mean   :102.1    Mean   : 799.7
##  3rd Qu.: 606.5   3rd Qu.: 1681   3rd Qu.:142.5    3rd Qu.: 938.0
##  Max.   :2008.0   Max.   :22768   Max.   :336.0    Max.   :9963.0
##                                                    NA's   :18
##  TEAM_BATTING_H_NEW TEAM_BATTING_2B_NEW TEAM_FIELDING_E_NEW
##  Min.   :1149       Min.   :116.0       Min.   : 73.0
##  1st Qu.:1387       1st Qu.:210.0       1st Qu.:131.0
##  Median :1455       Median :239.0       Median :163.0
##  Mean   :1467       Mean   :242.1       Mean   :238.8
##  3rd Qu.:1548       3rd Qu.:278.5       3rd Qu.:252.0
##  Max.   :1775       Max.   :376.0       Max.   :660.2
##
##  TEAM_PITCHING_BB_NEW TEAM_BASERUN_SB_SIN TEAM_BATTING_HBP_Missing
##  Min.   :286.0        Min.   :-0.99975    Min.   :0.00000
##  1st Qu.:471.0        1st Qu.:-0.68318    1st Qu.:0.00000
##  Median :526.0        Median : 0.14546    Median :0.00000
##  Mean   :542.3        Mean   : 0.06904    Mean   :0.07336
##  3rd Qu.:606.5        3rd Qu.: 0.81676    3rd Qu.:0.00000
##  Max.   :805.0        Max.   : 0.99952    Max.   :1.00000
##                       NA's   :13
##  TEAM_BASERUN_CS_Missing     Hits_R        Strikeout_R
##  Min.   :0.0000          Min.   :0.0679   Min.   :0.0679
##  1st Qu.:0.0000          1st Qu.:0.9382   1st Qu.:0.9388
##  Median :1.0000          Median :0.9506   Median :0.9508
##  Mean   :0.6641          Mean   :0.9168   Mean   :0.9204
##  3rd Qu.:1.0000          3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000          Max.   :1.0187   Max.   :1.0189
##                                           NA's   :20
```

Now that we have the test data prepared, we will go ahead and run the final model on this dataset.

```r
summary(step4_final)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ . - Hits_R - TEAM_BATTING_BB - TEAM_BATTING_H -
##     TEAM_BATTING_HR - TEAM_BATTING_2B - TEAM_PITCHING_H - TEAM_PITCHING_BB -
##     TEAM_FIELDING_E_NEW - TEAM_BATTING_HBP_Missing, data = data_step4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.740  -7.022   0.108   7.101  28.685
```

```
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             40.283276   8.906321   4.523 6.49e-06 ***
## TEAM_BATTING_3B          0.172848   0.018879   9.155  < 2e-16 ***
## TEAM_BASERUN_SB          0.071631   0.005507  13.008  < 2e-16 ***
## TEAM_FIELDING_E         -0.120338   0.007257 -16.583  < 2e-16 ***
## TEAM_FIELDING_DP        -0.106677   0.012368  -8.625  < 2e-16 ***
## TEAM_PITCHING_HR         0.094122   0.008705  10.812  < 2e-16 ***
## TEAM_PITCHING_SO        -0.019500   0.002206  -8.839  < 2e-16 ***
## TEAM_BATTING_H_NEW       0.032683   0.004342   7.528 8.05e-14 ***
## TEAM_BATTING_2B_NEW     -0.056302   0.008926  -6.307 3.55e-10 ***
## TEAM_PITCHING_BB_NEW     0.033202   0.003093  10.736  < 2e-16 ***
## TEAM_BASERUN_SB_SIN     -0.668686   0.339924  -1.967 0.049316 *
## TEAM_BASERUN_CS_Missing -4.062390   0.803315  -5.057 4.69e-07 ***
## Strikeout_R             17.050630   4.940275   3.451 0.000571 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.16 on 1822 degrees of freedom
##   (441 observations deleted due to missingness)
## Multiple R-squared:  0.4076, Adjusted R-squared:  0.4037
## F-statistic: 104.5 on 12 and 1822 DF,  p-value: < 2.2e-16
```

```
pred_test <- predict(step4_final, data_test, interval="prediction", level=0.95)

#pred_test

final_predicted_DS <- cbind(data_test, pred_test)
```

Below is the result of the prediction for the 259 cases that we had for evaluation.

```
kable(final_predicted_DS)
```

TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTIN
1209	170	33	83	
1221	151	29	88	
1395	183	29	93	
1539	309	29	159	
1445	203	68	5	
1431	236	53	10	
1430	219	55	37	
1385	158	42	33	
1259	177	78	23	
1397	212	42	58	
1427	243	40	50	
1496	239	55	164	
1420	223	57	186	
1460	232	22	176	
1411	195	22	141	
1434	192	30	153	
1297	204	22	130	

TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTIN
1446	284	25	166	
1276	162	52	17	
1715	322	72	116	
1520	295	68	49	
1597	291	38	98	
1453	256	67	105	
1378	225	26	118	
1516	277	24	152	
1556	288	20	164	
1499	183	28	3	
1464	263	58	47	
1558	318	66	32	
1502	308	36	39	
1596	320	58	130	
1546	260	59	110	
1516	282	53	115	
1550	275	47	146	
1447	260	54	148	
1450	252	28	203	
1347	239	36	130	
1561	260	56	214	
1578	252	26	135	
1598	259	45	181	
1497	322	21	145	
1569	310	39	124	
1119	118	33	7	
1609	196	120	62	
1514	175	70	80	
1657	237	119	41	
1746	213	106	69	
1319	224	70	56	
1293	204	70	18	
1420	235	70	36	
1496	269	54	76	
1625	289	38	80	
1391	239	50	145	
1319	203	43	130	
1411	251	35	107	
1420	221	41	104	
1552	206	106	38	
1280	203	72	15	
1120	122	61	7	
1390	183	84	18	
1554	252	81	29	
1410	218	69	45	
1507	262	28	159	
1481	284	19	242	
1450	253	23	200	
1637	260	93	26	
1436	202	82	44	
1600	218	89	21	
1348	168	76	23	

TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTIN
1460	191	111	22	
1621	255	126	37	
1433	241	49	45	
1440	232	48	155	
1479	211	34	232	
1573	281	36	106	
1558	224	42	171	
1385	225	46	130	
1419	250	27	164	
1284	198	61	19	
1403	200	68	10	
1631	358	48	105	
1666	343	82	98	
1804	376	86	129	
1534	284	53	74	
1472	222	52	156	
1489	229	21	134	
1367	198	21	156	
1485	222	46	101	
1458	225	32	109	
1530	334	30	198	
1421	160	72	30	
1869	301	122	58	
1400	169	66	26	
1494	193	81	12	
1449	223	62	20	
1385	200	76	29	
1443	218	99	24	
1825	284	106	61	
1627	296	95	38	
1623	299	106	54	
1556	298	82	60	
1381	228	39	80	
1556	272	46	114	
1416	206	32	168	
1413	257	21	204	
1504	253	102	33	
1193	165	68	45	
1461	325	30	166	
1458	294	36	187	
1295	237	64	25	
1431	263	58	118	
1469	305	59	98	
1633	266	59	115	
1603	295	58	132	
1487	269	52	117	
1474	318	44	101	
1594	296	52	152	
1415	285	42	140	
1445	289	34	126	
1362	199	81	29	
1572	195	106	30	

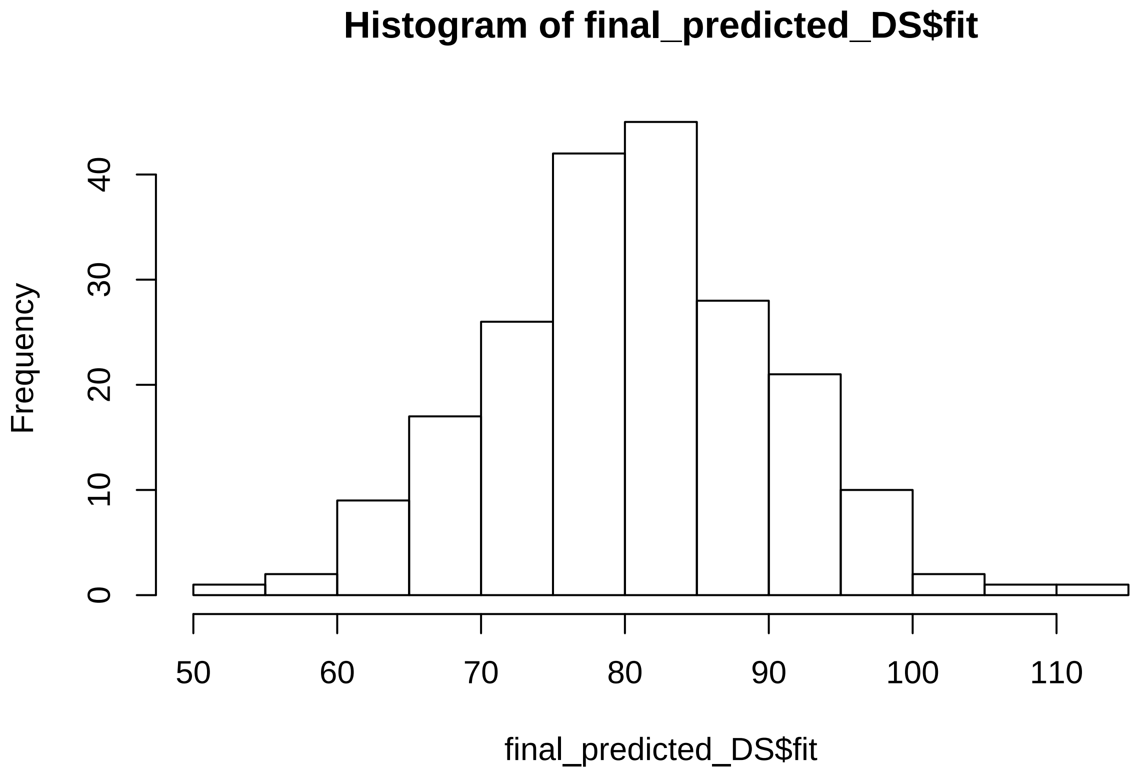
TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTIN
1209	168	56	16	
1242	155	69	20	
1098	116	63	29	
1235	175	77	26	
1651	247	80	59	
1712	265	85	68	
1391	206	78	41	
1625	299	73	105	
1740	319	77	128	
1626	303	55	84	
1471	277	36	65	
1373	232	14	130	
1466	215	35	158	
1450	226	30	203	
1474	223	57	18	
1335	228	49	120	
1455	233	36	97	
1477	272	35	82	
1426	240	25	125	
1255	183	61	11	
1264	141	79	9	
1695	310	89	66	
1460	274	66	63	
1349	237	46	53	
1340	226	40	117	
1396	257	42	150	
1472	259	47	82	
1544	256	46	112	
1453	282	41	141	
1446	257	39	196	
1468	289	30	106	
1546	44	29	0	
1372	195	31	103	
1365	203	29	98	
1314	172	26	112	
1469	323	41	200	
1382	185	86	32	
1642	218	135	29	
1324	153	65	17	
1770	313	116	160	
1765	293	83	164	
1590	277	76	113	
1775	334	88	193	
1635	297	77	183	
1557	264	79	146	
1485	210	57	153	
1461	229	41	152	
1322	208	19	147	
1462	281	18	163	
1537	217	115	23	
1495	236	85	35	
1468	280	70	66	

TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTIN
1689	296	74	59	
1533	301	59	104	
1379	229	55	64	
1373	223	37	94	
1394	215	43	118	
1371	223	36	116	
1400	210	28	148	
1327	209	33	114	
1432	263	33	199	
1474	251	22	156	
1450	279	28	205	
2025	292	140	32	
1669	281	102	35	
1631	291	79	52	
1420	299	79	5	
1312	230	52	29	
2058	336	90	75	
1351	181	58	25	
1452	199	87	17	
1466	242	57	68	
1534	256	44	64	
1609	311	38	61	
1344	207	28	59	
1438	239	41	96	
1368	225	53	139	
1381	218	52	127	
1498	250	59	130	
1389	206	53	145	
1448	224	49	117	
1307	225	58	102	
1517	250	38	104	
1417	245	25	112	
1352	209	45	125	
1458	296	34	106	
1390	290	35	116	
1475	257	80	52	
1378	178	85	35	
1817	277	155	60	
1711	213	133	29	
1415	217	112	52	
1263	190	32	97	
1328	221	63	96	
1571	248	59	126	
1522	235	70	130	
1550	278	57	133	
1412	237	33	98	
1344	243	46	111	
1441	276	30	141	
1395	271	35	107	
1506	320	31	168	
1437	269	39	143	
2170	241	70	13	

TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTIN
1324	194	53	94	
1442	239	25	136	
1413	279	37	157	
1416	269	39	130	
1523	216	97	33	
1294	169	51	24	
1668	251	98	79	
1422	215	53	140	
1524	231	31	200	
1392	227	41	134	
1318	200	44	80	
1499	229	26	112	
1345	215	48	141	
1620	210	139	66	
1339	185	80	34	
1621	272	86	95	
1585	288	62	105	
1576	269	46	67	
1541	300	49	101	
1149	175	18	59	
1626	265	27	125	
1461	228	29	121	
1472	284	39	181	
1366	218	39	99	
1489	287	36	195	
1457	305	38	187	
1454	220	52	9	
1642	221	98	56	
819	72	72	18	
1251	162	23	95	
1345	190	23	125	
1381	263	37	102	
1410	270	36	122	
1423	339	34	172	

The histogram of the predicted wins is as below:

```
hist(final_predicted_DS$fit)
```


Histogram of final_predicted_DS$fit
Frequency
0
10
20
30
40
50
60
70
80
90
100
110
final_predicted_DS$fit

Conclusion: