

Home Work Assignment - 01

Critical Thinking Group 5

Contents

Overview	2
1 Data Exploration Analysis	2
1.1 Variable identification	2
1.2 Data Summary Analysis	4
1.3 Outliers and Missing Values Identification	5
2. Data Preparation	7
2.1 Outliers treatment	7
2.2 Missing values treatment	11
2.3 Missing Flags	11
2.4 Ratios	12
2.5 Calculated Variables	12
2.6 Correlation for new variables	12
3 Build Models	13
3.1 Model One	15
3.2 Model Two	18
3.3 Model Three	19
3.4 Model Four	21
4 Model Selection	24
4.1 Model selection strategy:	24
4.2 Model diagnostics	25

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.

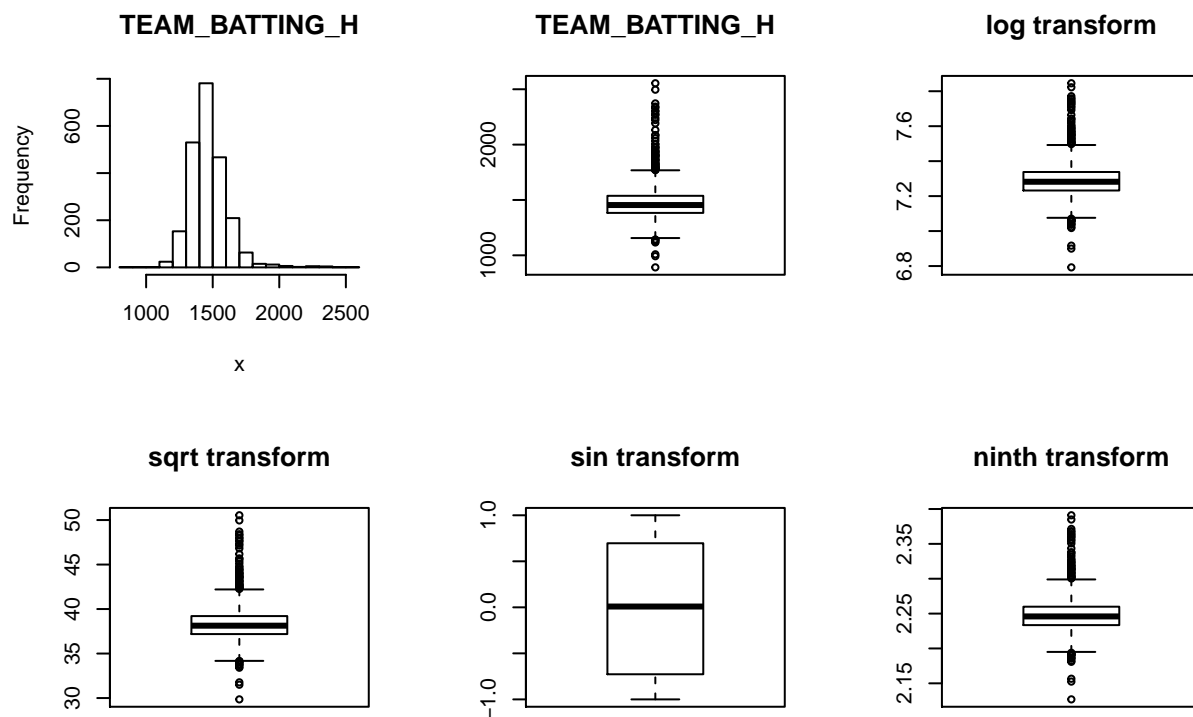
3. Most of the variables seem to indicate a positive / negative correlation in line with the theoretical effect. However, the following stand out as they show a correlation opposite to the theoretical impact: TEAM_BASERUN_CS, TEAM_PITCHING_HR, TEAM_PITCHING_BB, TEAM_PITCHING_SO and TEAM_FIELDING_DP. 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.



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.

***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_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 purging the below steps as guidelines:

- Outliers treatment
- Missing values treatment
- Data transformation

2.1 Outliers treatment

For outliers, we will create 2 sets of variables.

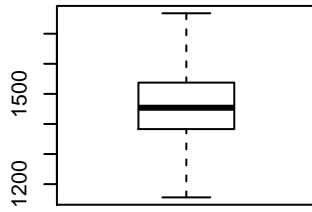
The first set uses the capping method. In this method, we will replace all outliers that lie outside the 1.5 times of IQR limits. We will cap it by replacing those observations less than the lower limit with the value of 5th %ile and those that lie above the upper limit with the value of 95th %ile.

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

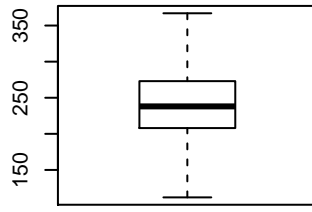
```
TEAM_BATTING_H_NEW  
TEAM_BATTING_2B_NEW  
TEAM_BATTING_3B_NEW  
TEAM_BATTING_BB_NEW  
TEAM_BASERUN_SB_NEW  
TEAM_FIELDING_E_NEW  
TEAM_FIELDING_DP_NEW  
TEAM_PITCHING_BB_NEW  
TEAM_PITCHING_H_NEW  
TEAM_PITCHING_HR_NEW  
TEAM_PITCHING_SO_NEW
```

Lets see how the new variables look in boxplots.

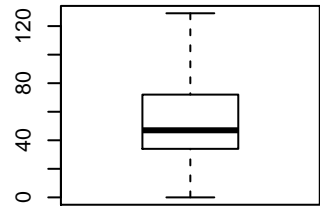
TEAM_BATTING_H_NEW



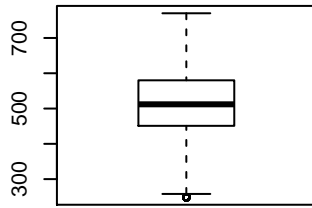
TEAM_BATTING_2B_NEW



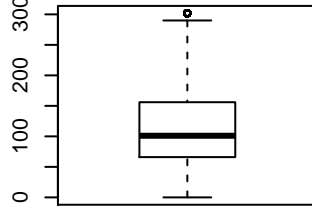
TEAM_BATTING_3B_NEW



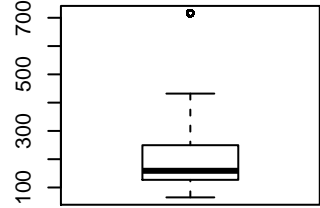
TEAM_BATTING_BB_NEW



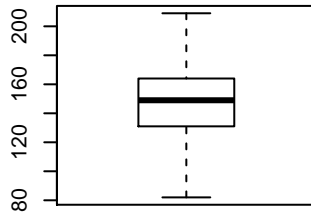
TEAM_BASERUN_SB_NEW



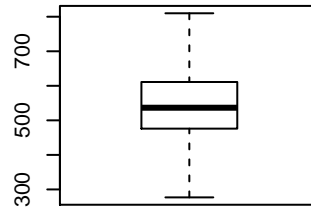
TEAM_FIELDING_E_NEW



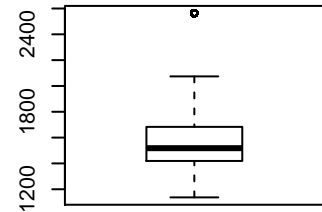
TEAM_FIELDING_DP_NEW



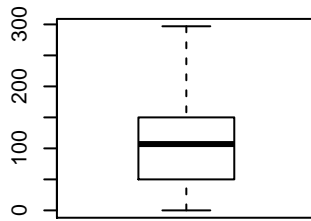
TEAM_PITCHING_BB_NEW



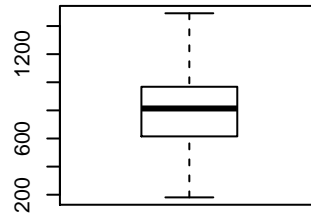
TEAM_PITCHING_H_NEW



TEAM_PITCHING_HR_NEW



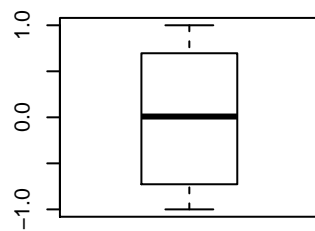
TEAM_PITCHING_SO_NEW



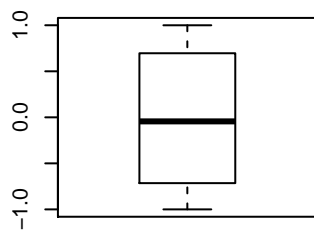
In the second set, we will use the sin transformation and create the following variables:

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

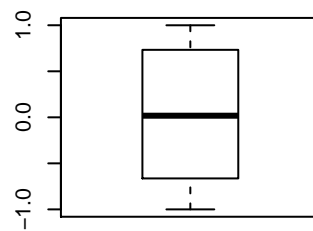
TEAM_BATTING_H_SIN



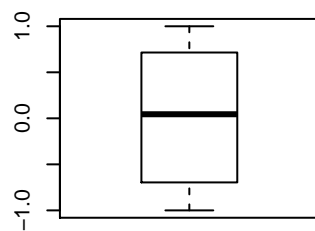
TEAM_BATTING_2B_SIN



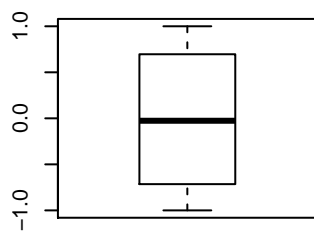
TEAM_BATTING_3B_SIN



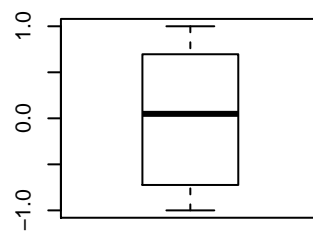
TEAM_BATTING_BB_SIN

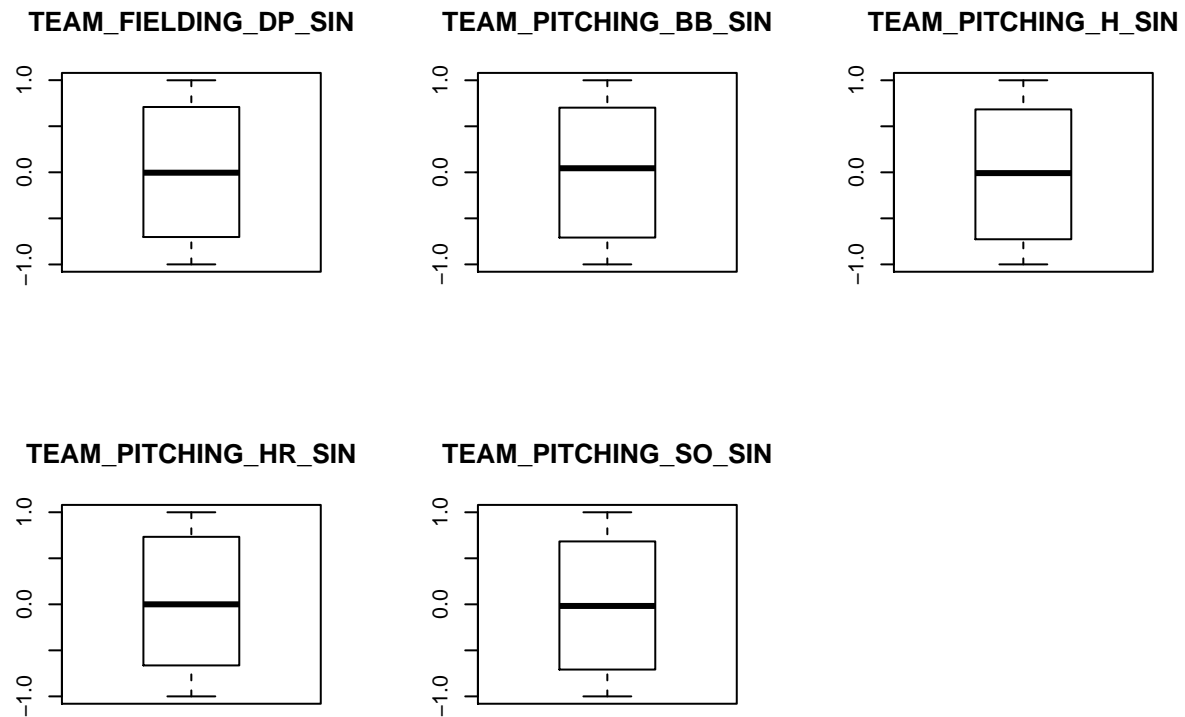


TEAM_BASERUN_SB_SIN



TEAM_FIELDING_E_SIN





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	Y			
TEAM_BATTING_SO	Given	Negative	Y			Y
TEAM_BASERUN_SB	Given	Positive	Y			Y
TEAM_BASERUN_CS	Given	Negative	Y			
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. below the relevant variables for model 1:

```
## Start:  AIC=8529.29
## 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
##
##              Df Sum of Sq    RSS    AIC F value    Pr(>F)
## - TEAM_PITCHING_HR  1      27.3 188679 8527.6   0.2639  0.607544
## - TEAM_BATTING_SO   1     150.8 188802 8528.8   1.4555  0.227805
## <none>              188651 8529.3
```

```

## - TEAM_BATTING_HR 1 314.7 188966 8530.3 3.0379 0.081512 .
## - TEAM_PITCHING_BB 1 361.1 189012 8530.8 3.4854 0.062072 .
## - TEAM_BATTING_H 1 380.6 189032 8531.0 3.6737 0.055434 .
## - TEAM_PITCHING_SO 1 522.1 189173 8532.4 5.0399 0.024890 *
## - TEAM_BATTING_BB 1 736.4 189388 8534.4 7.1080 0.007742 **
## - TEAM_PITCHING_H 1 1539.1 190190 8542.2 14.8569 0.000120 ***
## - TEAM_BATTING_2B 1 3196.2 191847 8558.1 30.8523 3.195e-08 ***
## - TEAM_FIELDING_DP 1 8636.8 197288 8609.4 83.3689 < 2.2e-16 ***
## - TEAM_BATTING_3B 1 9685.4 198337 8619.2 93.4902 < 2.2e-16 ***
## - TEAM_BASERUN_SB 1 16250.9 204902 8678.9 156.8653 < 2.2e-16 ***
## - TEAM_FIELDING_E 1 28809.2 217460 8788.1 278.0877 < 2.2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step: AIC=8527.56
## 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_SO
##
## Df Sum of Sq RSS AIC F value Pr(>F)
## <none> 188679 8527.6
## - TEAM_BATTING_SO 1 374.1 189053 8529.2 3.6125 0.057505 .
## - TEAM_PITCHING_BB 1 432.2 189111 8529.8 4.1740 0.041191 *
## - TEAM_BATTING_H 1 477.1 189156 8530.2 4.6074 0.031967 *
## - TEAM_BATTING_BB 1 847.1 189526 8533.8 8.1801 0.004284 **
## - TEAM_PITCHING_SO 1 1081.9 189760 8536.0 10.4471 0.001250 **
## - TEAM_PITCHING_H 1 1805.9 190484 8543.0 17.4392 3.106e-05 ***
## - TEAM_BATTING_2B 1 3192.1 191871 8556.3 30.8245 3.240e-08 ***
## - TEAM_FIELDING_DP 1 8675.6 197354 8608.0 83.7767 < 2.2e-16 ***
## - TEAM_BATTING_3B 1 9674.9 198353 8617.3 93.4270 < 2.2e-16 ***
## - TEAM_BATTING_HR 1 12421.7 201100 8642.6 119.9518 < 2.2e-16 ***
## - TEAM_BASERUN_SB 1 16314.6 204993 8677.7 157.5437 < 2.2e-16 ***
## - TEAM_FIELDING_E 1 28845.1 217524 8786.6 278.5471 < 2.2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B +
## TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO +
## TEAM_BASERUN_SB + TEAM_FIELDING_E + TEAM_FIELDING_DP + TEAM_PITCHING_BB +
## TEAM_PITCHING_H + TEAM_PITCHING_SO, data = na.omit(data))
##
## Residuals:
## Min 1Q Median 3Q Max
## -32.168 -7.256 0.142 6.945 29.896
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.054832 6.025125 9.801 < 2e-16 ***
## TEAM_BATTING_H -0.033844 0.015767 -2.146 0.03197 *
## TEAM_BATTING_2B -0.049268 0.008874 -5.552 3.24e-08 ***
## TEAM_BATTING_3B 0.183497 0.018984 9.666 < 2e-16 ***

```



```

## TEAM_BATTING_HR    0.100263    0.009155    10.952    < 2e-16 ***
## TEAM_BATTING_BB    0.118363    0.041385     2.860    0.00428 **
## TEAM_BATTING_SO    0.033316    0.017529     1.901    0.05751 .
## TEAM_BASERUN_SB    0.069465    0.005534    12.552    < 2e-16 ***
## TEAM_FIELDING_E   -0.118864    0.007122   -16.690    < 2e-16 ***
## TEAM_FIELDING_DP  -0.112317    0.012271    -9.153    < 2e-16 ***
## TEAM_PITCHING_BB  -0.080309    0.039308    -2.043    0.04119 *
## TEAM_PITCHING_H    0.059813    0.014323     4.176    3.11e-05 ***
## TEAM_PITCHING_SO  -0.053523    0.016559    -3.232    0.00125 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.18 on 1822 degrees of freedom
## Multiple R-squared:  0.4058, Adjusted R-squared:  0.4019
## F-statistic: 103.7 on 12 and 1822 DF,  p-value: < 2.2e-16

```

We notice that model 1 has the following summary characteristics:

- 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

	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-intuitive 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 comparison 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.

We notice that model 2 has the following summary characteristics:

- 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

	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
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-intuitive 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 comparison.

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.

We notice that model 3 has the following summary characteristics:

- 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

	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-intuitive 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 comparison 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.

We notice that model 4 has the following summary characteristics:

-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

	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
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-intuitive 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 comparison 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) Akaike information criterion (AIC) measure has been used to compare relative performance of different models
- (2) Along with that of adjusted R^2 values are also used to compare different models performance
- (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) 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

```
## [1] 13737.06
```

```
## [1] 18290.75
```

```
## [1] 14353.81
```

```
## [1] 13690.51
```

```
##
```

```
## Call:
```

```
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B +  
##     TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BASERUN_SB +  
##     TEAM_FIELDING_E + TEAM_FIELDING_DP + TEAM_PITCHING_BB + TEAM_PITCHING_H +  
##     TEAM_PITCHING_HR + TEAM_PITCHING_SO + TEAM_BATTING_H_NEW +  
##     TEAM_BATTING_2B_NEW + TEAM_FIELDING_E_NEW + TEAM_PITCHING_BB_NEW +  
##     TEAM_BASERUN_SB_SIN + TEAM_BATTING_HBP_Missing + TEAM_BASERUN_CS_Missing +  
##     Hits_R + Strikeout_R, data = na.omit(data))  
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max  
## -32.397  -7.039   0.262   6.793  29.527
```



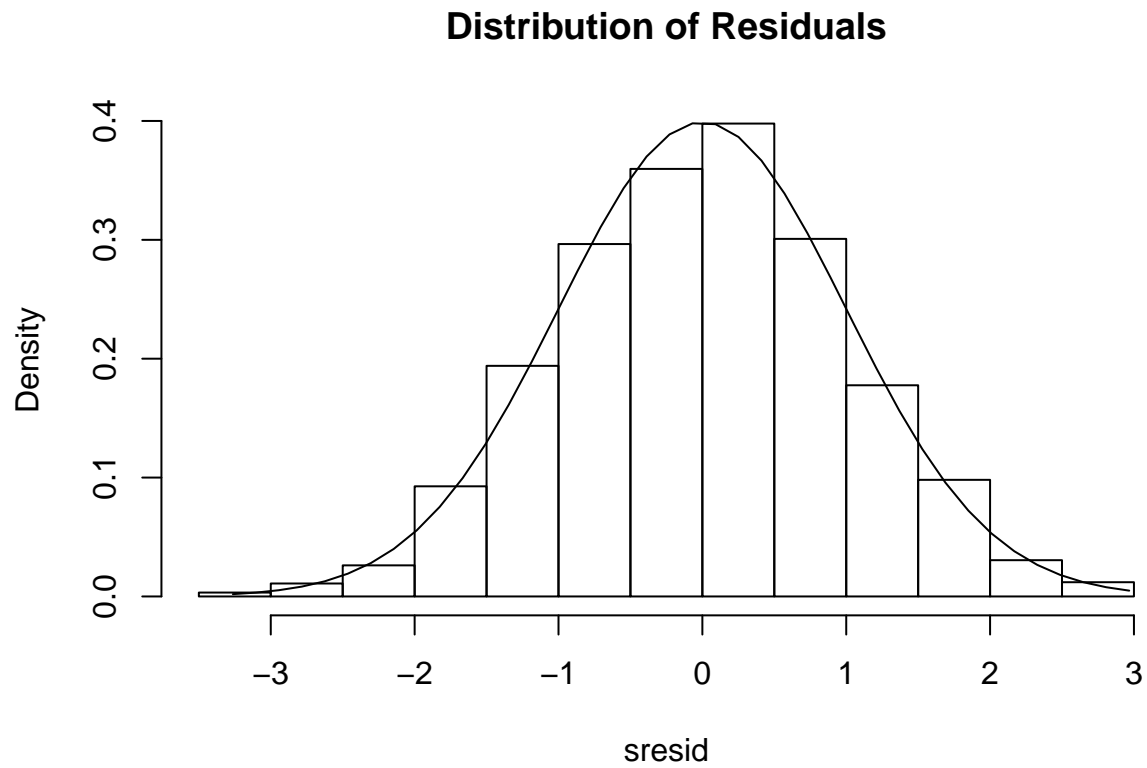
```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.303e+01  3.912e+01   1.356 0.175401
## TEAM_BATTING_H    -1.266e-01  3.658e-02  -3.461 0.000551 ***
## TEAM_BATTING_2B    1.537e-01  6.921e-02   2.221 0.026443 *
## TEAM_BATTING_3B    1.924e-01  1.901e-02  10.120 < 2e-16 ***
## TEAM_BATTING_HR    2.077e-01  6.660e-02   3.119 0.001842 **
## TEAM_BATTING_BB    1.605e-01  5.160e-02   3.110 0.001900 **
## TEAM_BASERUN_SB    6.907e-02  5.545e-03  12.456 < 2e-16 ***
## TEAM_FIELDING_E   -1.974e-01  2.696e-02  -7.320 3.72e-13 ***
## TEAM_FIELDING_DP   -1.051e-01  1.229e-02  -8.551 < 2e-16 ***
## TEAM_PITCHING_BB   -8.095e-02  4.111e-02  -1.969 0.049089 *
## TEAM_PITCHING_H    6.240e-02  2.208e-02   2.825 0.004774 **
## TEAM_PITCHING_HR   -1.035e-01  6.380e-02  -1.622 0.105049
## TEAM_PITCHING_SO   -1.854e-02  2.292e-03  -8.089 1.09e-15 ***
## TEAM_BATTING_H_NEW  9.140e-02  2.840e-02   3.218 0.001312 **
## TEAM_BATTING_2B_NEW -2.020e-01  6.984e-02  -2.892 0.003877 **
## TEAM_FIELDING_E_NEW  6.342e-02  2.355e-02   2.693 0.007144 **
## TEAM_PITCHING_BB_NEW -4.068e-02  1.954e-02  -2.082 0.037522 *
## TEAM_BASERUN_SB_SIN -6.994e-01  3.362e-01  -2.080 0.037657 *
## TEAM_BATTING_HBP_Missing -2.539e+00  9.551e-01  -2.659 0.007913 **
## TEAM_BASERUN_CS_Missing -4.136e+00  8.205e-01  -5.040 5.11e-07 ***
## Hits_R            -1.458e+03  7.100e+02  -2.054 0.040106 *
## Strikeout_R        1.465e+03  7.087e+02   2.067 0.038846 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.02 on 1813 degrees of freedom
## Multiple R-squared:  0.4264, Adjusted R-squared:  0.4197
## F-statistic: 64.17 on 21 and 1813 DF, p-value: < 2.2e-16
```

Looking at the AIC values it appears that models, “step1” & “step 4” are comparatively better models of the pack. “step1” has adjusted R^2 value .5167 which means this model can explain 51.67% variability in data. “step4” has adjusted R^2 value of .5585 and this model can explain 55.85% variability in data. From this two data points model “step4” was picked for further evaluation.

4.2 Model diagnostics

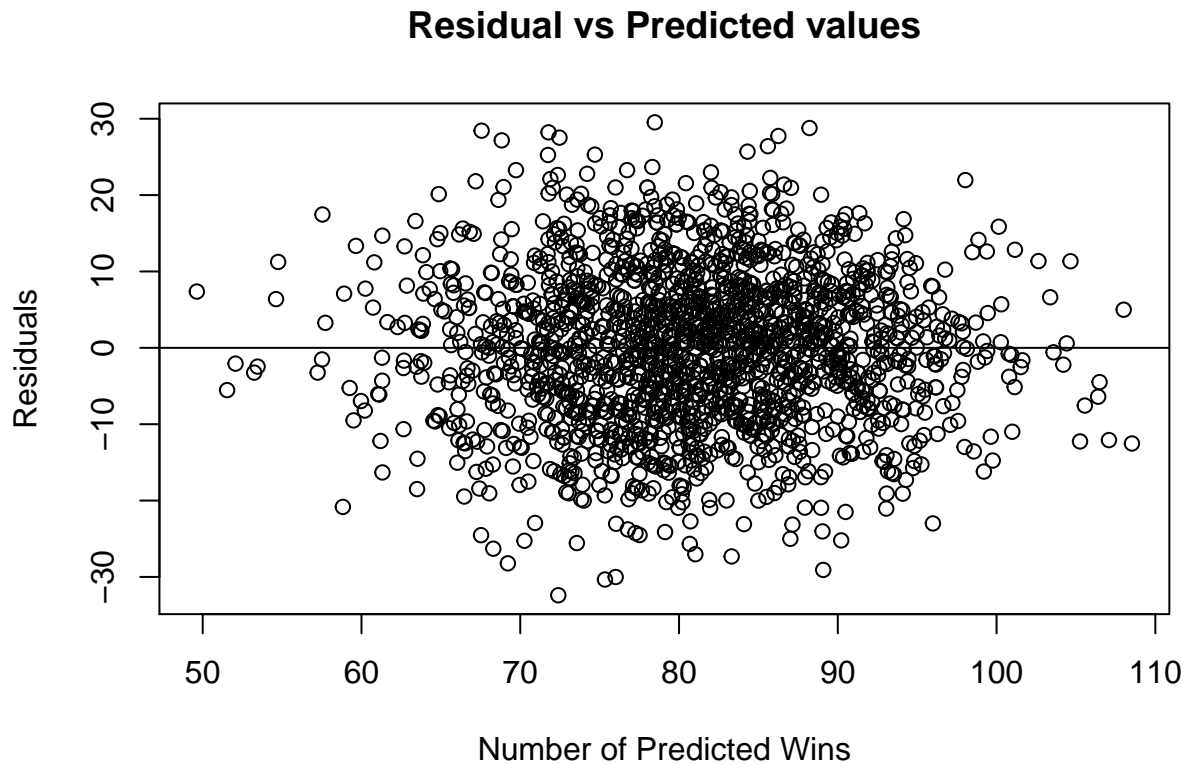
We will create plots to validate the assumption of Linear Regression:

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.

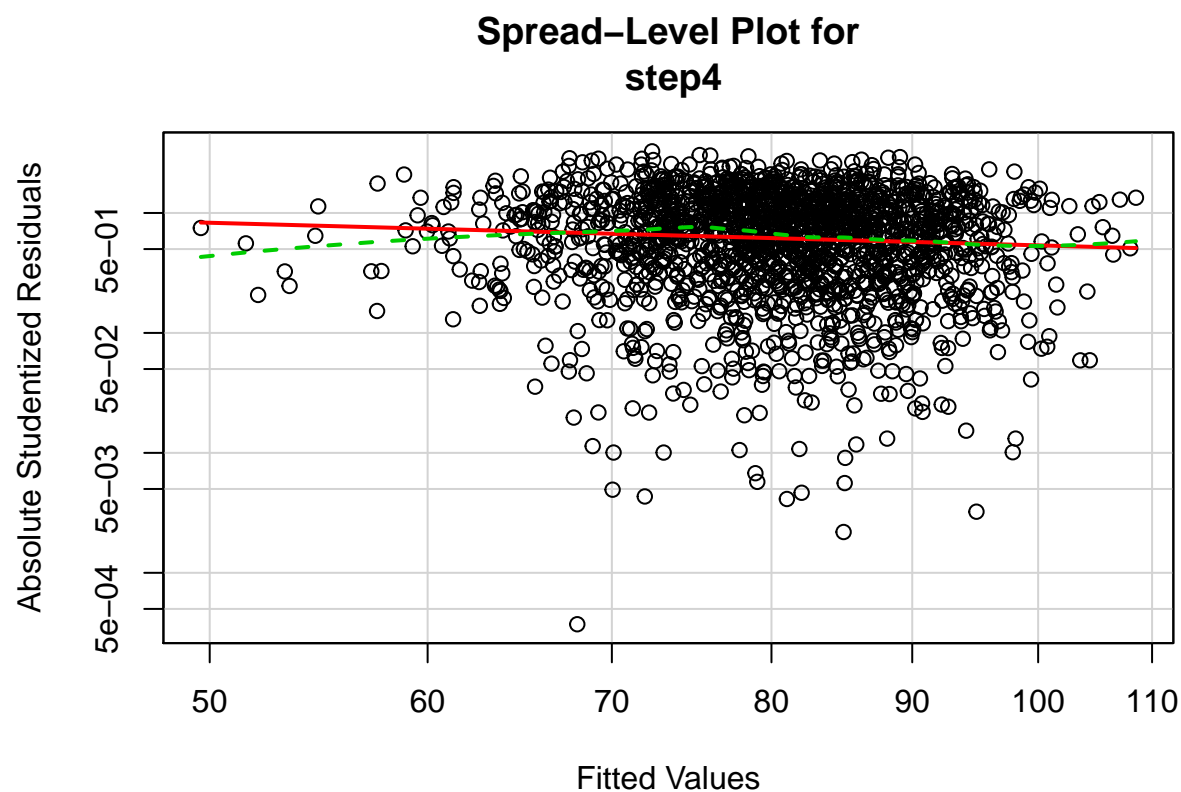
plot residuals with respect to predicted value for randomness:



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 9: 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 2 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 “TEAM_PITCHING_BB” has highest value > 2 and is removed from model, and model is evaluated without that variable. Adjusted R^2 value changed from .5585 to .5562. 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_SO” has highest value > 2 and is removed from model, and model is evaluated without that variable. Adjusted R^2 values changed from .5562 to .5534. Hence this variable is not adding lot of value to the model and can be removed.

pass 3- Based on that variance inflation factors (VIF) following variable “Strikeout_R” has highest value > 2 and is removed from model, and model is evaluated without that variable. Adjusted R^2 value changed from .5534 to .5526. Hence this variable is not adding lot of value to the model and can be removed.

pass 4- Based on that variance inflation factors (VIF) following variable “HomeRuns_R” has highest value > 2 and is removed from model, and model is evaluated without that variable. Adjusted R^2 value changed from .5526 to .5462 which is some compromise with the performance at the cost of reducing complexity. Reduction of variable will simplify the model and hence updated model is selected with some compromise in performance.

```
##      TEAM_BATTING_H      TEAM_BATTING_BB      TEAM_BATTING_HBP
##      1.269620          1.165108          1.037474
##      TEAM_PITCHING_HR      TEAM_PITCHING_SO      TEAM_FIELDING_E
##      1.294692          1.252542          1.086548
##      TEAM_FIELDING_DP      TEAM_BATTING_BB_SIN      TEAM_BASERUN_SB_SIN
##      1.026350          1.034345          1.039926
## TEAM_FIELDING_DP_SIN      Walks_R
##      1.029959          1.024758

##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_BB +
##      TEAM_BATTING_HBP + TEAM_PITCHING_HR + TEAM_PITCHING_SO +
##      TEAM_FIELDING_E + TEAM_FIELDING_DP + TEAM_BATTING_BB_SIN +
##      TEAM_BASERUN_SB_SIN + TEAM_FIELDING_DP_SIN + Walks_R, data = moneyball12)
##
## Residuals:
##      Min        1Q    Median        3Q       Max
## -24.1241  -5.0179  -0.3098   4.7776  22.5184
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.354e+02  2.899e+02   1.157  0.248862
## TEAM_BATTING_H    2.396e-02  9.871e-03   2.427  0.016215 *
```

```

## TEAM_BATTING_BB      5.349e-02  9.217e-03  5.804 2.89e-08 ***
## TEAM_BATTING_HBP     8.191e-02  4.737e-02  1.729 0.085490 .
## TEAM_PITCHING_HR     8.511e-02  2.366e-02  3.596 0.000417 ***
## TEAM_PITCHING_SO    -3.086e-02  7.107e-03 -4.342 2.36e-05 ***
## TEAM_FIELDING_E      -1.810e-01  3.868e-02 -4.679 5.66e-06 ***
## TEAM_FIELDING_DP     -1.168e-01  3.450e-02 -3.385 0.000874 ***
## TEAM_BATTING_BB_SIN  -1.489e+00  8.654e-01 -1.720 0.087109 .
## TEAM_BASERUN_SB_SIN -2.434e+00  8.614e-01 -2.826 0.005252 **
## TEAM_FIELDING_DP_SIN 1.578e+00  8.551e-01  1.846 0.066596 .
## Walks_R              -2.695e+02  2.908e+02 -0.927 0.355256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.161 on 179 degrees of freedom
## (2085 observations deleted due to missingness)
## Multiple R-squared:  0.5725, Adjusted R-squared:  0.5462
## F-statistic: 21.79 on 11 and 179 DF,  p-value: < 2.2e-16

```

Final model was derived after four passes of elimination were carried out. Looking at the VIF values in the final model there is no collinearity among variables < 2 . In this scenario a model with slightly less performance was selected to avoid collinearity effect among variables. But there was no significant reduction of model performance. But this exercise helped to reduce the model complexity.

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_BATTING_SO TEAM_BASERUN_SB TEAM_FIELDING_E
## Min. : 15.0 Min. : 0.0 Min. : 0.0 Min. : 73.0
## 1st Qu.:436.5 1st Qu.: 545.0 1st Qu.: 59.0 1st Qu.: 131.0
## Median :509.0 Median : 686.0 Median : 92.0 Median : 163.0
## Mean :499.0 Mean : 709.3 Mean :123.7 Mean : 249.7
## 3rd Qu.:565.5 3rd Qu.: 912.0 3rd Qu.:151.8 3rd Qu.: 252.0
## Max. :792.0 Max. :1268.0 Max. :580.0 Max. :1568.0
## NA's :18 NA's :13
## TEAM_FIELDING_DP TEAM_PITCHING_BB TEAM_PITCHING_H TEAM_PITCHING_HR
## Min. : 69.0 Min. : 136.0 Min. : 1155 Min. : 0.0
## 1st Qu.:131.0 1st Qu.: 471.0 1st Qu.: 1426 1st Qu.: 52.0
## Median :148.0 Median : 526.0 Median : 1515 Median :104.0
## Mean :146.1 Mean : 552.4 Mean : 1813 Mean :102.1
## 3rd Qu.:164.0 3rd Qu.: 606.5 3rd Qu.: 1681 3rd Qu.:142.5

```

```

## Max. :204.0 Max. :2008.0 Max. :22768 Max. :336.0
## NA's :31
## TEAM_PITCHING_SO TEAM_BATTING_H_NEW TEAM_BATTING_2B_NEW
## Min. : 0.0 Min. :1149 Min. :116.0
## 1st Qu.: 613.0 1st Qu.:1387 1st Qu.:210.0
## Median : 745.0 Median :1455 Median :239.0
## Mean : 799.7 Mean :1467 Mean :242.1
## 3rd Qu.: 938.0 3rd Qu.:1548 3rd Qu.:278.5
## Max. :9963.0 Max. :1775 Max. :376.0
## NA's :18
## TEAM_BATTING_3B_NEW TEAM_BATTING_BB_NEW TEAM_BASERUN_SB_NEW
## Min. : 14.00 Min. :259.0 Min. : 0.0
## 1st Qu.: 35.00 1st Qu.:436.5 1st Qu.: 60.5
## Median : 52.00 Median :509.0 Median : 96.0
## Mean : 55.26 Mean :503.8 Mean :120.2
## 3rd Qu.: 72.00 3rd Qu.:565.5 3rd Qu.:149.0
## Max. :126.00 Max. :746.0 Max. :319.0
##
## TEAM_FIELDING_E_NEW TEAM_FIELDING_DP_NEW TEAM_PITCHING_BB_NEW
## Min. : 73.0 Min. : 83.0 Min. :286.0
## 1st Qu.:131.0 1st Qu.:134.5 1st Qu.:471.0
## Median :163.0 Median :146.4 Median :526.0
## Mean :238.8 Mean :146.4 Mean :542.3
## 3rd Qu.:252.0 3rd Qu.:160.5 3rd Qu.:606.5
## Max. :660.2 Max. :204.0 Max. :805.0
##
## TEAM_PITCHING_H_NEW TEAM_PITCHING_HR_NEW TEAM_PITCHING_SO_NEW
## Min. :1155 Min. : 0.0 Min. : 315.0
## 1st Qu.:1426 1st Qu.: 52.0 1st Qu.: 622.5
## Median :1515 Median :104.0 Median : 764.8
## Mean :1606 Mean :101.6 Mean : 764.8
## 3rd Qu.:1681 3rd Qu.:142.5 3rd Qu.: 927.5
## Max. :2385 Max. :244.0 Max. :1295.0
##
## TEAM_BATTING_H_SIN TEAM_BATTING_2B_SIN TEAM_BATTING_3B_SIN
## Min. : -0.99999 Min. : -0.99976 Min. : -0.9998
## 1st Qu.: -0.67675 1st Qu.: -0.80113 1st Qu.: -0.7271
## Median : 0.05320 Median : 0.01770 Median : 0.1499
## Mean : 0.02619 Mean : -0.01888 Mean : 0.0324
## 3rd Qu.: 0.74501 3rd Qu.: 0.70868 3rd Qu.: 0.7626
## Max. : 0.99991 Max. : 0.99991 Max. : 0.9999
##
## TEAM_BATTING_BB_SIN TEAM_BASERUN_SB_SIN TEAM_FIELDING_E_SIN
## Min. : -0.99999 Min. : -0.99975 Min. : -0.99999
## 1st Qu.: -0.59492 1st Qu.: -0.68318 1st Qu.: -0.81160
## Median : 0.07961 Median : 0.14546 Median : -0.08834
## Mean : 0.06589 Mean : 0.06904 Mean : -0.05169
## 3rd Qu.: 0.73317 3rd Qu.: 0.81676 3rd Qu.: 0.69608
## Max. : 0.99976 Max. : 0.99952 Max. : 0.99991
## NA's :13
## TEAM_FIELDING_DP_SIN TEAM_PITCHING_BB_SIN TEAM_PITCHING_H_SIN
## Min. : -0.999207 Min. : -0.999522 Min. : -0.99999
## 1st Qu.: -0.727143 1st Qu.: -0.670218 1st Qu.: -0.65017
## Median : -0.061920 Median : -0.053114 Median : 0.07940

```

```
## Mean      :-0.008299      Mean      :-0.003834      Mean      : 0.01651
## 3rd Qu.: 0.699225      3rd Qu.: 0.708642      3rd Qu.: 0.68966
## Max.      : 0.997799      Max.      : 0.999757      Max.      : 0.99991
## NA's      :31
## TEAM_PITCHING_HR_SIN TEAM_PITCHING_SO_SIN TEAM_BATTING_HBP_Missing
## Min.      :-0.99999      Min.      :-0.99999      Min.      :0.00000
## 1st Qu.: -0.75099      1st Qu.: -0.71483      1st Qu.:0.00000
## Median    :-0.04424      Median    : 0.01761      Median    :0.00000
## Mean      :-0.03292      Mean      :-0.01170      Mean      :0.07336
## 3rd Qu.: 0.66361      3rd Qu.: 0.68320      3rd Qu.:0.00000
## Max.      : 0.99991      Max.      : 0.99991      Max.      :1.00000
## NA's      :18
## TEAM_BASERUN_CS_Missing Hits_R Walks_R
## Min.      :0.0000      Min.      :0.0679      Min.      :0.06787
## 1st Qu.:0.0000      1st Qu.:0.9382      1st Qu.:0.93871
## Median    :1.0000      Median    :0.9506      Median    :0.95067
## Mean      :0.6641      Mean      :0.9168      Mean      :0.91681
## 3rd Qu.:1.0000      3rd Qu.:1.0000      3rd Qu.:1.00000
## Max.      :1.0000      Max.      :1.0187      Max.      :1.01786
##
## HomeRuns_R Strikeout_R TEAM_BATTING_EB TEAM_BATTING_1B
## Min.      :0.1111      Min.      :0.0679      Min.      : 73.0      Min.      : 657.0
## 1st Qu.:0.9358      1st Qu.:0.9388      1st Qu.:339.5      1st Qu.: 990.5
## Median    :0.9516      Median    :0.9508      Median    :397.0      Median    :1059.0
## Mean      :0.9199      Mean      :0.9204      Mean      :392.9      Mean      :1076.5
## 3rd Qu.:1.0000      3rd Qu.:1.0000      3rd Qu.:450.0      3rd Qu.:1134.0
## Max.      :1.0204      Max.      :1.0189      Max.      :615.0      Max.      :1846.0
## NA's      :1      NA's      :20
```

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

```
summary(step4_final)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_BB +
##     TEAM_BATTING_HBP + TEAM_PITCHING_HR + TEAM_PITCHING_SO +
##     TEAM_FIELDING_E + TEAM_FIELDING_DP + TEAM_BATTING_BB_SIN +
##     TEAM_BASERUN_SB_SIN + TEAM_FIELDING_DP_SIN + Walks_R, data = moneyball2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.1241  -5.0179  -0.3098   4.7776  22.5184
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.354e+02  2.899e+02   1.157 0.248862
## TEAM_BATTING_H    2.396e-02  9.871e-03   2.427 0.016215 *
## TEAM_BATTING_BB    5.349e-02  9.217e-03   5.804 2.89e-08 ***
## TEAM_BATTING_HBP    8.191e-02  4.737e-02   1.729 0.085490 .
## TEAM_PITCHING_HR    8.511e-02  2.366e-02   3.596 0.000417 ***
## TEAM_PITCHING_SO   -3.086e-02  7.107e-03  -4.342 2.36e-05 ***
## TEAM_FIELDING_E   -1.810e-01  3.868e-02  -4.679 5.66e-06 ***
```



```
## TEAM_FIELDING_DP      -1.168e-01  3.450e-02  -3.385  0.000874 ***
## TEAM_BATTING_BB_SIN   -1.489e+00  8.654e-01  -1.720  0.087109 .
## TEAM_BASERUN_SB_SIN  -2.434e+00  8.614e-01  -2.826  0.005252 **
## TEAM_FIELDING_DP_SIN  1.578e+00  8.551e-01   1.846  0.066596 .
## Walks_R               -2.695e+02  2.908e+02  -0.927  0.355256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.161 on 179 degrees of freedom
## (2085 observations deleted due to missingness)
## Multiple R-squared:  0.5725, Adjusted R-squared:  0.5462
## F-statistic: 21.79 on 11 and 179 DF,  p-value: < 2.2e-16
```

```
#pred_test <- predict(step4_final, data, interval="prediction", level=0.95)
#head(pred_test)
```