Assignment 1

Moneyball OLS Regression Project

Laura Ellis, PREDICT 411 – Section 58

2015

Bingo Bonus Points – Expected (80 + ?? Points)

The Bingo Bonus work starts on page 73

SAS Files Included

The following SAS files have been included with the submission. Although it should be clear from the heading, below is a mapping of what sections the files cover:

1. Part 1 - Exploratory Data Analysis – "Laura_Ellis_Sec58_Assign1_EDA.sas"
2. Part 2 and 3 – Data Prep and Model Trials - "Laura_Ellis_Sec58_Assign1_Data_Prep_Model_Trials.sas"
3. Part 4 – Deploy the Model – "Laura_Ellis_Sec58_Assign1_Deploy_Model.sas"
4. Part 5 – Final Scored SAS File – "laura_ellis_a1_scored_model.sas7bdat"

Introduction

The purpose of this document is to analyze historical professional baseball records from the years 1871 to 2006 and provide the best linear regression model to predict the number of season wins for a team. The nine known variables that were considered as potential predictors were:

TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)
TEAM_BATTING_2B	Doubles by batters (2B)
TEAM_BATTING_3B	Triples by batters (3B)
TEAM_BATTING_HR	Homeruns by batters (4B)
TEAM_BATTING_BB	Walks by batters
TEAM_BATTING_HBP	Batters hit by pitch (get a free base)
TEAM_BATTING_SO	Strikeouts by batters
TEAM_BASERUN_SB	Stolen bases
TEAM_BASERUN_CS	Caught stealing
TEAM_FIELDING_E	Errors
TEAM_FIELDING_DP	Double Plays
TEAM_PITCHING_BB	Walks allowed
TEAM_PITCHING_H	Hits allowed
TEAM_PITCHING_HR	Homeruns allowed
TEAM_PITCHING_SO	Strikeouts by pitchers

Exploratory Data Analysis (EDA) was performed on the data set to further understand the variable properties. Consideration was given to their distribution, correlation to other variables and their relationship with the independent variable; total wins.

Based on the knowledge gained from our EDA, data preparation was performed to optimize the predictor variables for linear regression. Transformations such as binning, trimming, standardization, using tree logic and setting flag variables were used to optimize the variables and mitigate any negative effects from missing data or outliers.

After the data preparation was completed a large number of models were built with a range of approaches to find the best predictive model of baseball team season wins. A number of variable selection techniques such as backward, forward and stepwise regression were used. In addition variables were included in some models based on intuition, their correlation with team wins, presence up in a decision tree and more. Additions were made to the data transformation based on what was learned during building the models such as: introducing alternate ways to remedy missing data and introducing principal component analysis variables.

The best candidate model was selected based on its Adjusted RSquare and AIC values. After selecting the best candidate model, an assessment of the model adequacy was performed by reviewing diagnostic plots.

The model was then made available for production by developing a SAS deployment program. The model was validated by running the deployment program on a test data set to confirm its accuracy on new data.

Results

Part 1: Data Exploration

Variable: TEAM_WINS

We start off by exploring the variable TEAM_WINS. This variable is the independent variable which we are trying to predict. It represents the number of wins within the season for a particular team.

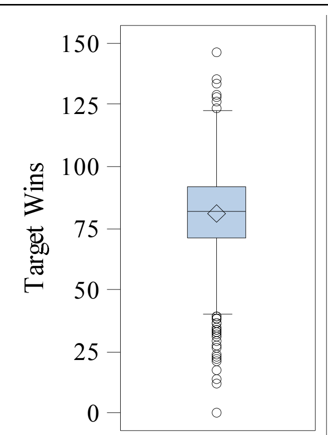
Distribution Details

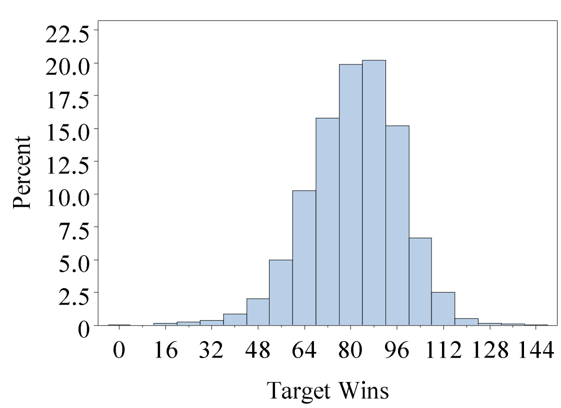
The distribution details for TEAM_WINS is are included in Table 1 below. In looking at the histogram and box plot it is obvious there are a number of outlier values which is not surprising given the large time range that this data takes place over. Given the outlier status, we look to the percentile values and take note that the first percentile is 38 and the 99th is 114. It may be useful to trim to these values in our final model. If we wanted to drop the outlier observations individually we could look at the extreme observations below. This is something we will explore further during the final model assessment. We also notice the standard deviation of 15.75. This will serve as a good baseline when we are trying to see how accurate our decision tree nodes are. Finally we also look at the mean value of approximately 81, this will serve as the value used for the average model. We will rank our model against the average model using this number.

Distribution Metrics

Analysis Variable : TARGET_WINS Target Wins									
N	N Miss	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Minimum	Maximum
276	0	80.7908612	15.7521525	38.0000000	54.0000000	104.0000000	114.0000000	0	146.0000000

Graphical Distribution and Extreme Values





Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	1211	128	418
12	2233	129	422
14	1825	134	296
17	982	135	2012
21	859	146	299

% Missing

0% Missing

Table1: Distribution Details for Variable TEAM_WINS

Correlation Details

Table 2 below illustrates the predictor variables correlation with TARGET_WINS. Note that the highest correlated value is TEAM_BATTING_H (highlighted in yellow) and there are a number of insignificant relationships (highlighted in grey). This may make manual variable selection slightly more challenging as selecting highly or moderately correlated variables is usually a good place to start.

In the table below green is a strong correlation (Pearson correlation value 0.5 +), yellow is a moderate correlation (Pearson correlation value e 0.3-0.5) and grey is an irrelevant or statistically insignificant correlation (p value less than 0.05). This practical view of correlations can be found at the following web page: https://explorable.com/statistical-correlation

TARGET_WINS Correlation	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB
Pearson Correlation	1	0.38877	0.2891	0.14261	0.17615	0.23256
P Value		<.0001	<.0001	<.0001	<.0001	<.0001
N	2276	2276	2276	2276	2276	2276

TARGET_WINS Correlation	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H
Pearson Correlation	-0.03175	0.13514	0.0224	0.0735	-0.10994
P Value	0.1389	<.0001	0.3853	0.3122	<.0001
N	2174	2145	1504	191	2276

TARGET_WINS Correlation	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
Pearson Correlation	0.18901	0.12417	-0.07844	-0.17648	-0.03485
P Value	<.0001	<.0001	0.0003	<.0001	0.1201
N	2276	2276	2174	2276	1990

Table 2: Set of Correlation tables for the variable TEAM_WINS

Variable: Team Batting H

The variable "TEAM_BATTING_H" indicates the number of Base Hits by batters (1B,2B,3B,HR). This metric is seen to positively affect the number of wins.

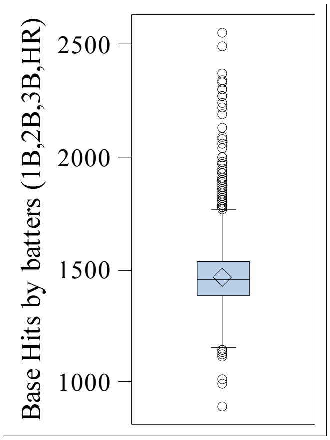
Distribution Details

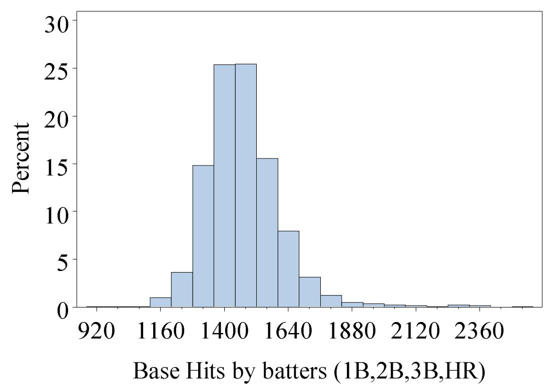
In Table 3 below we can see that base hits by batter are fairly normally distributed with a slightly longer right tail. This is consistent with the majority of box plot outliers occurring on the top spectrum of values. As such we may consider trimming the variable to the 5th and 95th percentile. These values 1280 and 1696 respectively can be obtained from the table below. Note also that there are no missing values in the training data set. However, it will still be useful to know the mean value of 1469.27 as we may use it to error proof our final model against missing values.

Distribution Metrics

Analysis Variable : TEAM_BATTING_H Base Hits by batters (1B,2B,3B,HR)									
N	N Miss	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Minimum	Maximum
2276	0	1469.27	144.5911954	1188.00	1280.00	1696.00	1950.00	891.0000000	2554.00

Graphical Distribution and Extreme Values





Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
891	1211	2333	296
992	2136	2343	273
1009	2233	2372	1810
1116	2276	2496	1811
1122	2239	2554	297

% Missing

0% Missing

Table 3: Distribution Details for Variable TEAM_BATTING_H

Correlation Details

Table 4 below outlines TEAM_BATTING_H's correlation with the other predictor variables variables. TEAM_BATTING_H has a moderate correlation with TARGET_WINS and therefore it may be a good choice for inclusion in the model. Note that TEAM_BATTING_2B appears to have a strong correlation with (highlighted in green). There are also a number of moderate correlations: TEAM_BATTING_3B, TEAM_BATTING_SO, TEAM_PITCHING_H. These relationships suggest that we may need watch out for multicollinearity within our models.

TEAM_BATTING_H Correlation	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB
	0.38877	1	0.56285	0.4277	-0.00654	-0.07246
P Value	<.0001		<.0001	<.0001	0.755	0.0005
N	2276	2276	2276	2276	2276	2276

TEAM_BATTING_H Correlation	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H
Pearson Correlation	-0.46385	0.12357	0.01671	-0.02911	0.30269
P Value	<.0001	<.0001	0.5174	0.6893	<.0001
N	2174	2145	1504	191	2276

TEAM_BATTING_H Correlation	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
Pearson Correlation	0.07285	0.09419	-0.25266	0.2649	0.15538

P Value	0.0005	<.0001	<.0001	<.0001	<.0001
N	2276	2276	2174	2276	1990

Table 4: Set of Correlation tables for the variable TEAM_BATTING_H

In Figure 1 below we can see that the outlier values are affecting the regression line. In fact the LOESS line appears to have very obvious slope changes right around the 5th and 95th percentiles. This further exemplifies why we need to deal with the outliers. Since there are not enough values to build a different model for high and low outliers, it is best that we trim the values to the 5th and 95th percentile.

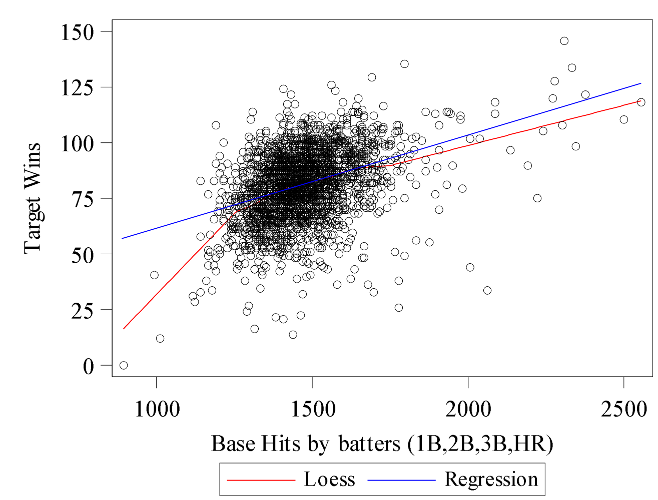


Figure 1: Scatter plot of TEAM_BATTING_H vs Target Wins including Regression and Loess line

Variable: Team Batting 2B

The variable "TEAM_BATTING_2B" indicates the doubles by batters. This metric is seen to positively affect the number of wins.

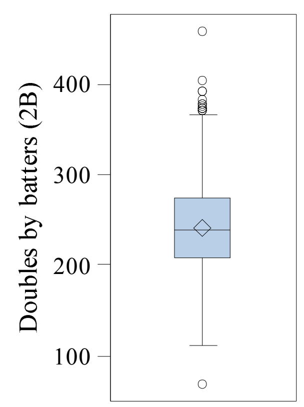
Distribution Details

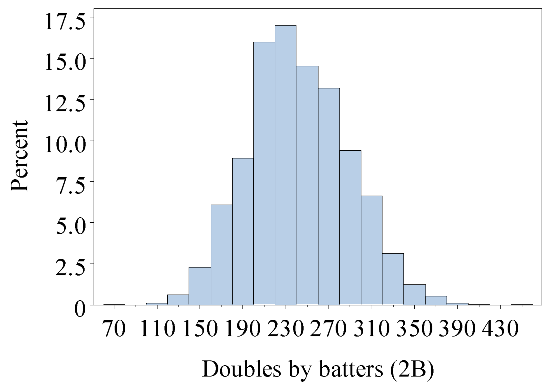
In Table 5 below we can see in the histogram that doubles by batters are fairly normally distributed. In both the histogram and box plot we can see that there are a very small number of outliers. Due to the small number of outliers it is unnecessary to trim to any specific percentile, it is best to trim the individual outlier values. There appear to be 5 major outliers (highlighted in the yellow below). Therefore we will trim the values to be between 112 and 382 to remedy this issue. Note also that there are no missing values in the training data set. However, it will still be useful to know the mean value of 241.247 as we may use it to error proof our final model against missing values.

Distribution Metrics

Analysis Variable : TEAM_BATTING_2B Doubles by batters (2B)									
N	N Miss	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Minimum	Maximum
2276	0	241.2469244	46.801414	141.0000000	167.0000000	320.0000000	352.0000000	69.0000000	458.0000000

Graphical Distribution and Extreme Values





Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
69	2239	382	1810
112	2233	392	2062
113	1397	393	296
118	2237	403	424
123	307	458	425

% Missing

0 % Missing

Table 5: Distribution Details for Variable TEAM_BATTING_2B

Correlation Details

Table 6 below outlines TEAM_BATTING_2B's correlation with the other variables. TEAM_BATTING_2B has a less than moderate correlation with TARGET_WINS and therefore it is not clear at the moment if it will be an obvious choice for the model. As discussed above TEAM_BATTING_H appears to have a strong correlation (highlighted in green). There are also a few moderate correlations: TEAM_BATTING_HR, TEAM_PITCHING_HR. These relationships would suggest that we may need to watch for multicollinearity within our models.

TEAM_BATTING_2B Correlation	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB
Pearson Correlation	0.2891	0.56285	1	-0.10731	0.4354	0.25573
P Value	<.0001	<.0001		<.0001	<.0001	<.0001
N	2276	2276	2276	2276	2276	2276

TEAM_BATTING_2B Correlation	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H
Pearson Correlation	0.16269	-0.19976	-0.09981	0.04608	0.02369
P Value	<.0001	<.0001	0.0001	0.5267	0.2585
N	2174	2145	1504	191	2276

TEAM_BATTING_2B Correlation	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
Pearson Correlation	0.45455	0.17805	0.06479	-0.23515	0.29088
P Value	<.0001	<.0001	0.0025	<.0001	<.0001
N	2276	2276	2174	2276	1990

Table 6: Set of Correlation tables for the variable TEAM_BATTING_2B

In Figure 2 below we can see that the outlier values are affecting the regression line. The LOESS line appears to have slope changes around where the 5 outliers take place. This further exemplifies why we need to deal with the outliers. Since there are not enough outlier values to trim the variable to a specific percentile, it is best that we simply trim the 5 outlier values.

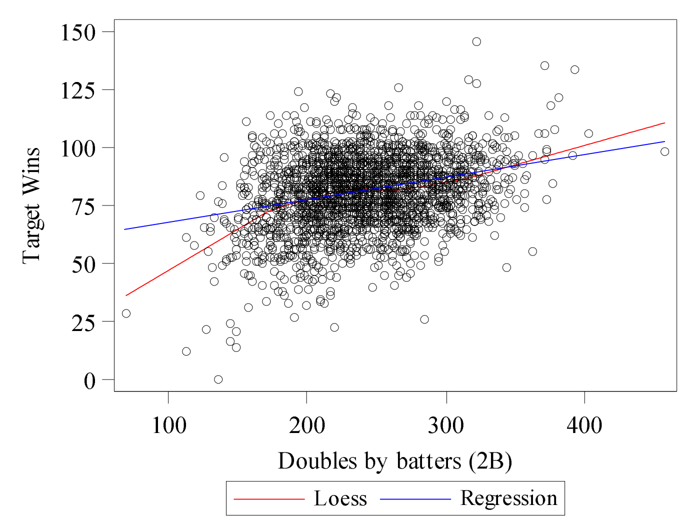


Figure 2: Scatter plot of TEAM_BATTING_2B vs Target Wins including Regression and Loess line

Variable: Triples by batters

The variable “TEAM_BATTING_3B” indicates the Triples hit by batters. This metric is seen to positively affect the number of wins.

Distribution Details

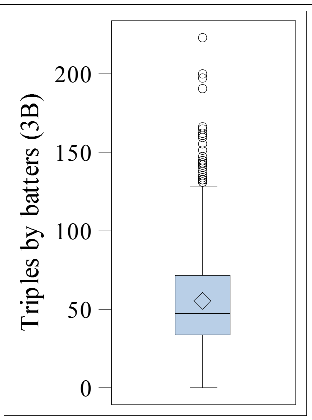
In Table 7 below we can see that the histogram has a long tail and as confirmed by the box plot there appear to be a large number of outliers on the high end of values. As such we should consider trimming to the 99th percentile. We may also want to consider trimming to the 1st percentile but the data presented in Table 7 does not yet present a compelling argument. We should examine the scatterplot in Figure 3 below first.

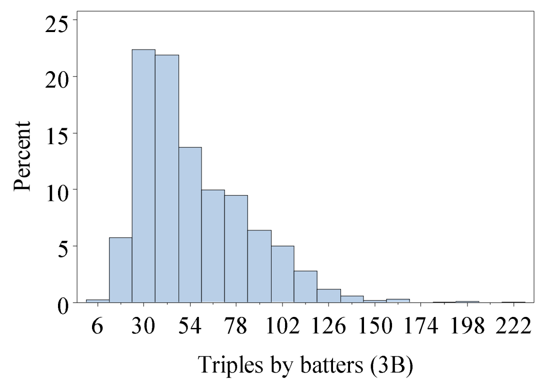
Note also that there are no missing values in the training data set. However, it will still be useful to know the mean value of 52.25 as we may use it to error proof our final model against missing values.

Distribution Metrics

Analysis Variable : TEAM_BATTING_3B Triples by batters (3B)									
N	N Miss	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Minimum	Maximum
2276	0	55.2500000	27.9385570	17.0000000	23.0000000	108.0000000	134.0000000	0	223.0000000

Graphical Distribution and Extreme Values





Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	1342	166	1604
0	1211	190	286
8	860	197	2219
9	2015	200	295
11	262	223	416

% Missing

0% Missing

Table 7: Distribution Details for Variable TEAM_BATTING_3B

Correlation Details

Table 8 below outlines TEAM_BATTING_3B's correlation with the other variables. TEAM_BATTING_3B has a less than moderate correlation with TARGET_WINS and therefore it is not clear at the moment if it will be an obvious choice for the model. It appears to have a strong correlation with TEAM_BATTING_HR, TEAM_BATTING_SO, TEAM_BASERUN_SB, TEAM_PITCHING_HR and TEAM_FIELDING_E. There are also a few moderate correlations: TEAM_BATTING_H, TEAM_BASERUN_CS, TEAM_FIELDING_DP. This variable is highly and moderately correlated to many other variables. These relationships would suggest that this variable could certainly contribute to high multicollinearity within our models.

TEAM_BATTING_3B Correlation	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB
Pearson Correlation	0.14261	0.4277	-0.10731	1	-0.63557	-0.28724
P Value	<.0001	<.0001	<.0001		<.0001	<.0001
N	2276	2276	2276	2276	2276	2276

TEAM_BATTING_3B Correlation	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H
Pearson Correlation	-0.66978	0.53351	0.34876	-0.17425	0.19488
P Value	<.0001	<.0001	<.0001	0.0159	<.0001
N	2174	2145	1504	191	2276

TEAM_BATTING_3B Correlation	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
Pearson Correlation	-0.56784	-0.00222	-0.25882	0.50978	-0.32307
P Value	<.0001	0.9155	<.0001	<.0001	<.0001
N	2276	2276	2174	2276	1990

Table 8: Set of Correlation tables for the variable TEAM_BATTING_3B

In Figure 3 below we can see that the high end outlier values are affecting the regression line. It also appears that the lowest values may also be affecting the regression line. As such a judgment call is made to trim to the 1st and 99th percentiles.

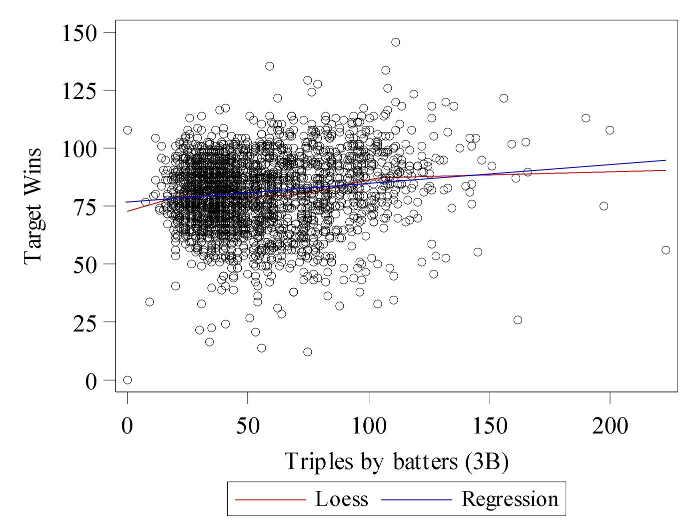


Figure 3: Scatter plot of TEAM_BATTING_3B vs Target Wins including Regression and Loess line

Variable: Homeruns by batters

The variable "TEAM_BATTING_HR" indicates the number of homeruns by batters. This metric is seen to positively affect the number of wins.

Distribution Details

In Table 9 below we can see in the boxplot and histogram that the TEAM_BATTING_HR variable is pretty uniformly distributed with no outliers. Therefore we may want to split the data up into quantiles that all have approximately the same membership to see if it will help our models performance. Figure 10 outlines the quantiles that we would employ to achieve approximately equal membership into each quantile.

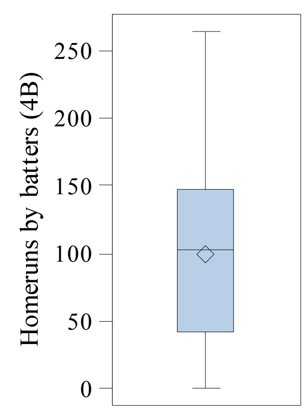
We also notice that while there are not outliers, the values after about 235 are not as frequent as other values. As such we may consider trimming values over 235.

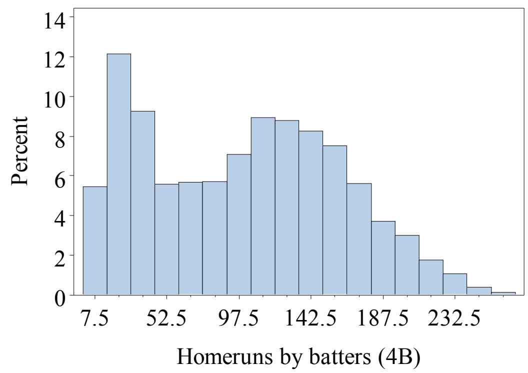
Finally note that there are no missing values in the training data set. However, it will still be useful to know the mean value of 99.61 as we may use it to error proof our final model against missing values.

Distribution Metrics

Analysis Variable : TEAM_BATTING_HR Homeruns by batters (4B)									
N	N Miss	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Minimum	Maximum
2276	0	99.6120387	60.5468720	4.0000000	14.0000000	199.0000000	235.0000000	0	264.0000000

Graphical Distribution and Extreme Values





Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	2239	247	1884
0	2233	249	1038
0	2136	257	260
0	2016	260	2188
0	2015	264	1885

% Missing

0% Missing

*Table 9: Distribution Details for Variabl*e TEAM_BATTING_HR

Analysis Variable : TEAM_BATTING_HR Homeruns by batters (4B)		
Rank for Variable TEAM_BATTING_HR	N Obs	Maximum
0	567	41.0000000
1	565	101.0000000
2	573	146.0000000
3	571	264.0000000

*Table 10: Distribution Details for Variabl*e TEAM_BATTING_HR

Correlation Details

Table 11 below outlines TEAM_BATTING_HR's correlation with the other variables. TEAM_BATTING_HR has a less than moderate correlation with TARGET_WINS and therefore it is not clear at the moment if it will be an obvious choice for the model. It appears to have a strong correlation with TEAM_BATTING_3B, TEAM_BATTING_BB, TEAM_BATTING_SO, TEAM_PITCHING_HR and TEAM_FIELDING_E. There are also a few moderate correlations: TEAM_BATTING_2B, TEAM_BASERUN_SB, TEAM_BASERUN_CS, TEAM_FIELDING_DP. This variable is highly and moderately correlated to many other variables. These relationships would suggest that this variable could certainly contribute to high multicollinearity within our models.

Of particular note is the extremely high correlation to TEAM_PITCHING_HR, with a statistically significant value of 0.96937 (highlighted in red). To reduce the multicollinearity on these variables we may want to consider combining the two variables into one variable.

TEAM_BATTING_HR Correlation	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB
Pearson Correlation	0.17615	-0.00654	0.4354	-0.63557	1	0.51373
P Value	<.0001	0.755	<.0001	<.0001		<.0001
N	2276	2276	2276	2276	2276	2276

TEAM_BATTING_HR Correlation	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H
Pearson Correlation	0.72707	-0.45358	-0.43379	0.10618	-0.25015
P Value	<.0001	<.0001	<.0001	0.1438	<.0001
N	2174	2145	1504	191	2276

TEAM_BATTING_HR Correlation	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
Pearson Correlation	0.96937	0.13693	0.18471	-0.58734	0.44899
P Value	<.0001	<.0001	<.0001	<.0001	<.0001
N	2276	2276	2174	2276	1990

Table 11: Set of Correlation tables for the variable TEAM_BATTING_HR

To ensure that the variables TEAM_BATTING_HR and TEAM_PITCHING_HR, are linearly correlated, we pull a scatterplot of their values. By looking at the graph it is confirmed that the values are linearly correlated. Therefore a new variable COMB_HR is created as a combination of these two variables.

Note: there does appear to be two possible lines in this graph. This is something we may want to explore at a later date but is outside the scope of this assignment.

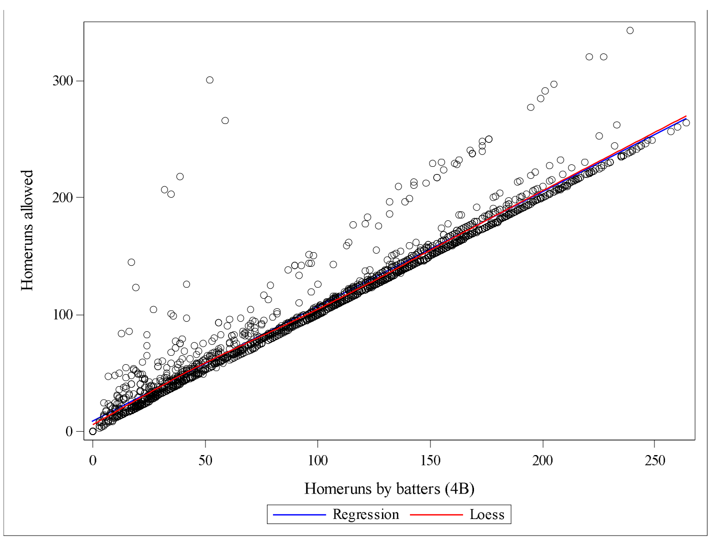


Figure 4: Scatter plot of TEAM_BATTING_HR vs TEAM_PITCHING_HR including Regression and Loess line

In Figure 5 below we can see the scatterplot of TEAM_BATTING_HR vs our predictor variable TARGET_WINS. We notice a very straight line. We also notice that the LOESS line does not appear to deviate too far from the regression line. The only exception is at the very beginning; it appears that the LOESS line shows some deviation from the regression line.

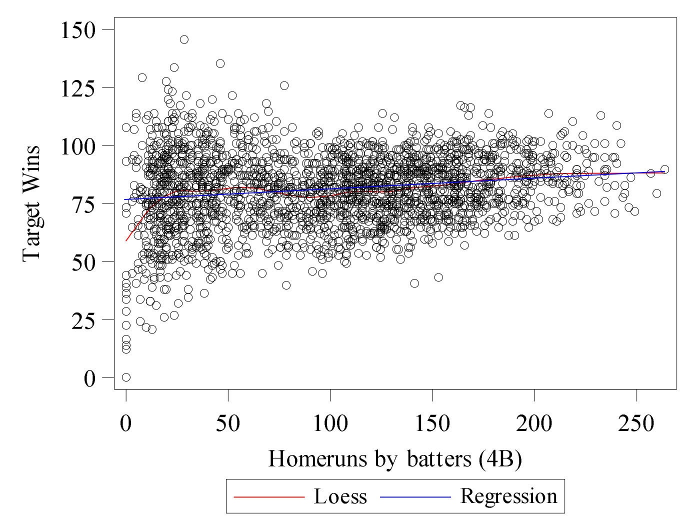


Figure 5: Scatter plot of TEAM_BATTING_HR vs Target Wins including Regression and Loess line

Variable: Walks by Batters

The variable "TEAM_BATTING_BB" indicates the walks by batters. This metric is seen to positively affect the number of wins.

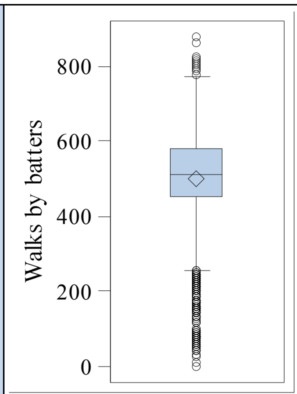
Distribution Details

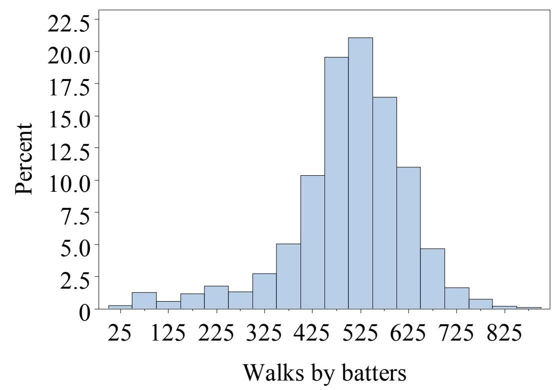
In Table 12 below we can see in the histogram that TEAM_BATTING_BB is a fairly normally distributed variable. In both the histogram and box plot we can see that there outliers on both the high end and the low end. To constrain the outliers we consider a z transform.

Note also that there are no missing values in the training data set. However, it will still be useful to know the mean value of 501.559 as we may use it to error proof our final model against missing values.

Distribution Metrics	Analysis Variable : TEAM_BATTING_BB Walks by batters									
	N	N Miss	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Minimum	Maximum
	2276	0	501.5588752	122.6708615	79.0000000	246.0000000	671.0000000	755.0000000	0	878.0000000

Graphical Distribution and Extreme Values





Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	1211	815	207
12	2233	819	396
29	2239	824	1534
34	1210	860	341
45	2220	878	342

% Missing 0% Missing

Table12: Distribution Details for Variable TEAM_BATTING_BB

We further explore the z transform of TEAM_BATTING_BB by performing a z transform and looking at the resulting histogram. It appears to have constrained the outliers on the high end of the values as nothing exceeds 3, but there still appear to be some outliers on the low end. This is evident by seeing values less than-3. However, it was decided to not further trim the variable because the transformed outliers were not that extreme and not that high in volume.

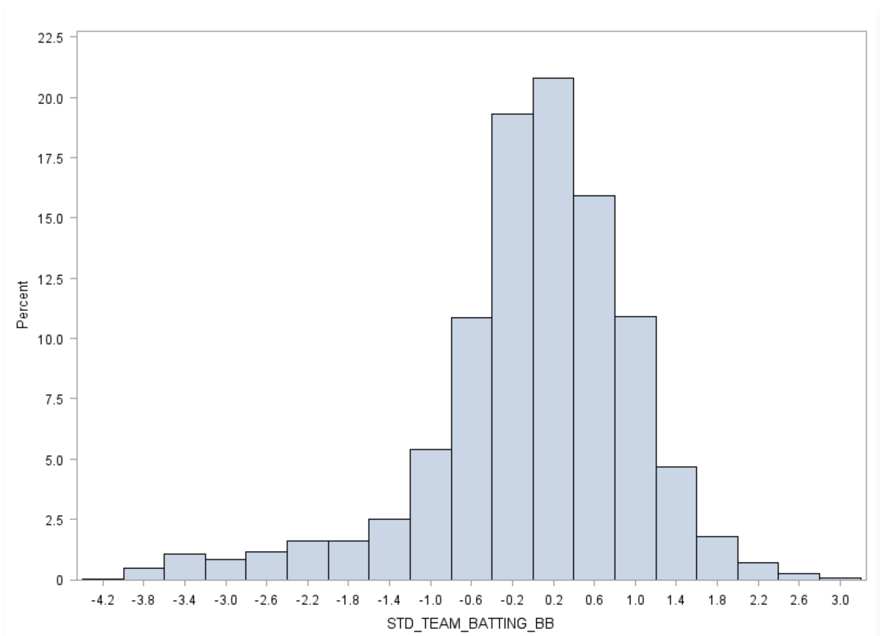


Figure 6: Z Transform of TEAM_BATTING BB

Correlation Details

Table 13 below outlines TEAM_BATTING_BB's correlation with the other variables. TEAM_BATTING_BB has a less than moderate correlation with TARGET_WINS and therefore it is not clear at the moment if it will be an obvious choice for the model. TEAM_BATTING_BB appears to have a few strong correlations (highlighted in green): TEAM_BATTING_HR and TEAM_FIELDING_E.

There are also a few moderate correlations: TEAM_BATTING_SO, TEAM_PITCHING_H, TEAM_PITCHING_HR, TEAM_PITCHING_BB and TEAM_FIELDING_DP. These relationships would suggest that we may need to watch for multicollinearity within our models.

TEAM_BATTING_BB Correlation	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB
Pearson Correlation	0.23256	-0.07246	0.25573	-0.28724	0.51373	1
P Value	<.0001	0.0005	<.0001	<.0001	<.0001	
N	2276	2276	2276	2276	2276	2276

TEAM_BATTING_BB Correlation	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H
Pearson Correlation	0.37975	-0.10512	-0.13699	0.04746	-0.44978
P Value	<.0001	<.0001	<.0001	0.5144	<.0001
N	2174	2145	1504	191	2276

TEAM_BATTING_BB Correlation	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
Pearson Correlation	0.45955	0.48936	-0.02076	-0.65597	0.43088
P Value	<.0001	<.0001	0.3334	<.0001	<.0001
N	2276	2276	2174	2276	1990

Table 13: Set of Correlation tables for the variable TEAM_BATTING_BB

In Figure 7 below we can see that the outlier values are affecting the regression line. The LOESS line appears to deviate from the regression line in both the high end and low end outliers. This further exemplifies why we need to transform the variable to constrain the outliers.

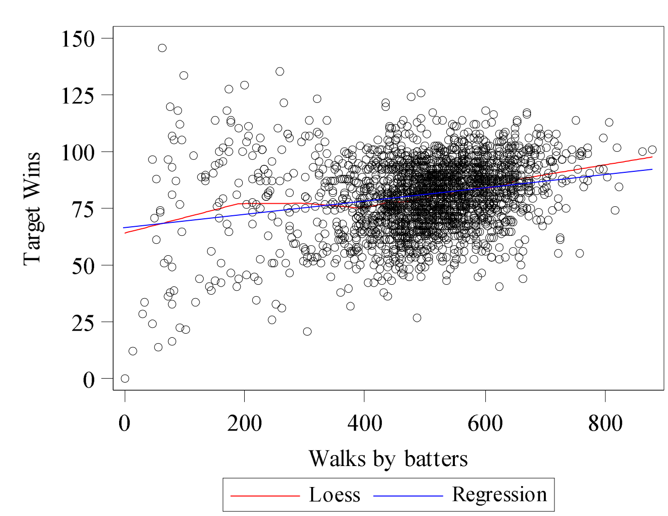


Figure 7: Scatter plot of TEAM_BATTING_BB vs Target Wins including Regression and Loess line

Variable: Batters Hit by a Pitch

The variable "TEAM_BATTING_HBP" indicates the number of batters hit by a pitch (get a free base). This metric is seen to positively affect the number of wins.

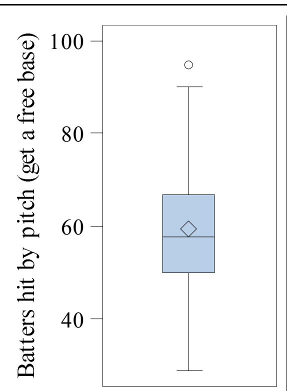
Distribution Details

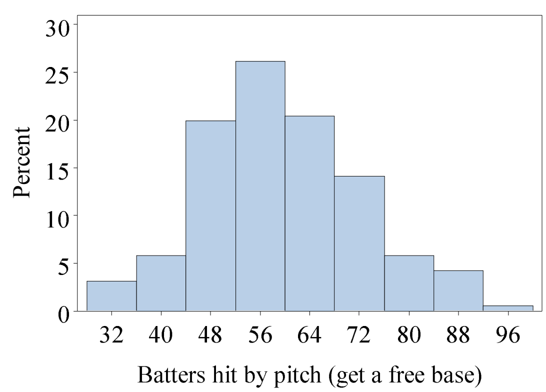
In Table 14 we can see that the data is normally distributed and there is only one outlier on the high end. However, the real issue is that this variable has over 91.6% missing values. Given the large proportion of missing values it does not make sense to replace the missing values. Likely the best idea is to drop the variable all together. There is no need to explore variable correlation further.

Distribution Metrics

Analysis Variable : TEAM_BATTING_HBP Batters hit by pitch (get a free base)									
N	N Miss	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Minimum	Maximum
191	2085	59.3560209	12.9671225	29.0000000	40.0000000	83.0000000	90.0000000	29.0000000	95.0000000

Graphical Distribution and Extreme Values





Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
29	2269	89	1809
29	43	89	2217
30	1861	89	2274
35	2273	90	2215
35	1333	95	1697

% Missing

91.6% Missing

Table14: Distribution Details for Variable TEAM_BATTING_HBP

We take one last look at the variable in Figure 8 through a scatterplot of TARGET_WINS vs TEAM_BATTING_HBP. We can see that there is virtually no slope to the regression and LOESS lines. Thus confirming that the variable has a minimal relationship to TARGET_WINS and we can safely get drop it from our data set.

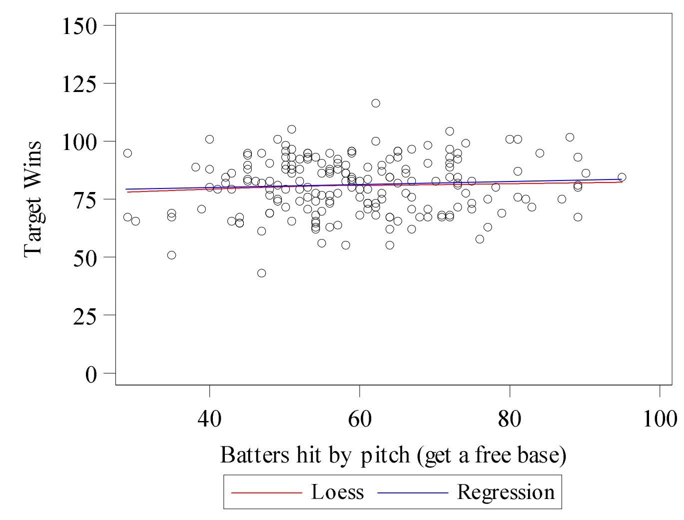


Figure 7: Scatter plot of TEAM_BATTING_HBP vs Target Wins including Regression and Loess line

Variable: Strikeouts by Batters

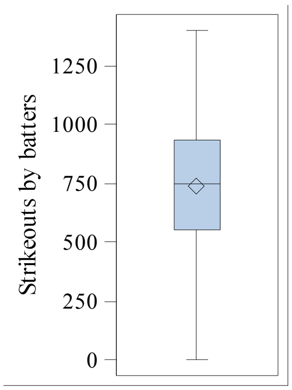
The variable "TEAM_BATTING_SO" indicates the number of strikeouts by batters. This metric is seen to negatively affect the number of wins.

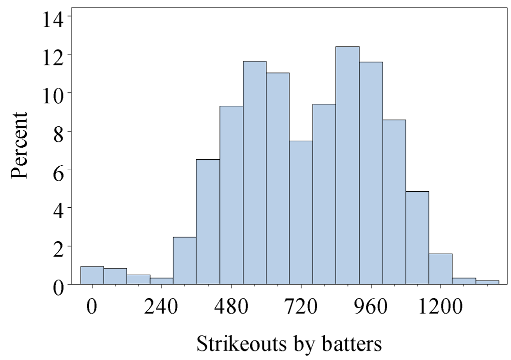
Distribution Details

In Table 15 below we can see in the boxplot and histogram that TEAM_BATTING_SO does not have any outliers. However, we do notice that there are 102 missing values. This equates to 4.5% missing. To handle missing values we explore a few methods. First we can replace the missing value with the mean: 735.6053358. The second method is to employ a decision tree such as in Figure 8 to produce the missing value logic. The decision tree in Figure 8 was produced by SPSS. Multiple trees were produced and reviewed. This tree was selected by looking to maximize tree simplicity and minimize the standard deviation in each tree node to a value as much below the normal TEAM_BATTING_SO standard deviation (248.5264177) as possible.

Distribution Metrics	Analysis Variable : TEAM_BATTING_SO Strikeouts by batters									
	N	N Miss	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Minimum	Maximum
	2174	102	735.6053358	248.5264177	67.0000000	359.0000000	1104.00	1193.00	0	1399.00

Graphical Distribution and Extreme Values





Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	2239	1303	747
0	2233	1320	1243
0	2016	1326	745
0	2015	1335	746
0	1824	1399	1240

% Missing	4.5% Missing

Table15: Distribution Details for Variable TEAM_BATTING_SO

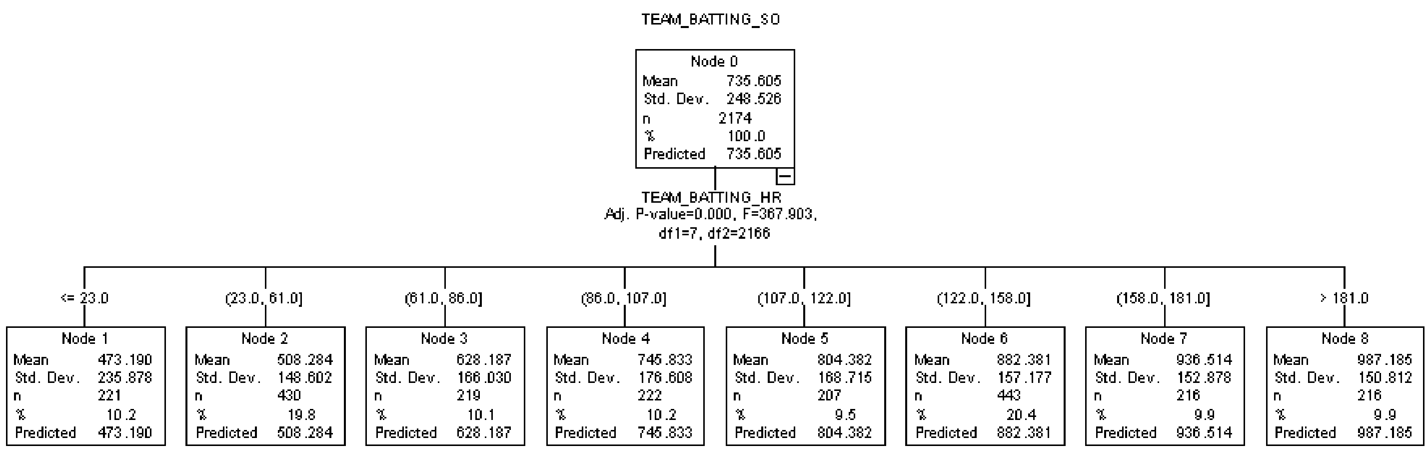


Figure 8: Decision Tree to replace the missing values for TEAM_BATTING_SO

Correlation Details

Table 6 below outlines TEAM_BATTING_SO's correlation with the other variables. TEAM_BATTING_SO has a less than moderate correlation with TARGET_WINS and therefore it is not clear at the moment if it will be an obvious choice for the model. TEAM_BATTING_SO appears to have a strong correlation with a number of variables: TEAM_BATTING_3B, TEAM_BATTING_HR, TEAM_PITCHING_HR and TEAM_FIELDING_E (highlighted in green). There are also a few moderate correlations: TEAM_BATTING_H, TEAM_BATTING_BB, TEAM_PITCHING_H and TEAM_PITCHING_SO. These relationships would suggest that we may need to watch for multicollinearity within our models.

TEAM_BATTING_SO Correlation	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB
Pearson Correlation	-0.03175	-0.46385	0.16269	-0.66978	0.72707	0.37975
P Value	0.1389	<.0001	<.0001	<.0001	<.0001	<.0001

N	2174	2174	2174	2174	2174	2174

TEAM_BATTING_SO Correlation	**TEAM_BATTING_SO**	**TEAM_BASERUN_SB**	**TEAM_BASERUN_CS**	**TEAM_BATTING_HBP**	**TEAM_PITCHING_H**
Pearson Correlation	1	-0.25449	-0.21788	0.22094	-0.37569
P Value		<.0001	<.0001	0.0021	<.0001
N	2174	2043	1504	191	2174

TEAM_BATTING_SO Correlation	**TEAM_PITCHING_HR**	**TEAM_PITCHING_BB**	**TEAM_PITCHING_SO**	**TEAM_FIELDING_E**	**TEAM_FIELDING_DP**
Pearson Correlation	0.66718	0.03701	0.41623	-0.58466	0.15489
P Value	<.0001	0.0845	<.0001	<.0001	<.0001
N	2174	2174	2174	2174	1888

Table 16: Set of Correlation tables for the variable TEAM_BATTING_SO

In Figure 9 below we can see that that although there are no outlier values, there are some values of lower frequency at the low end and high end of the values. This does appear to minimally affect the LOESS regression line. At this point, because there are no outliers, we will choose not to trim this variable.

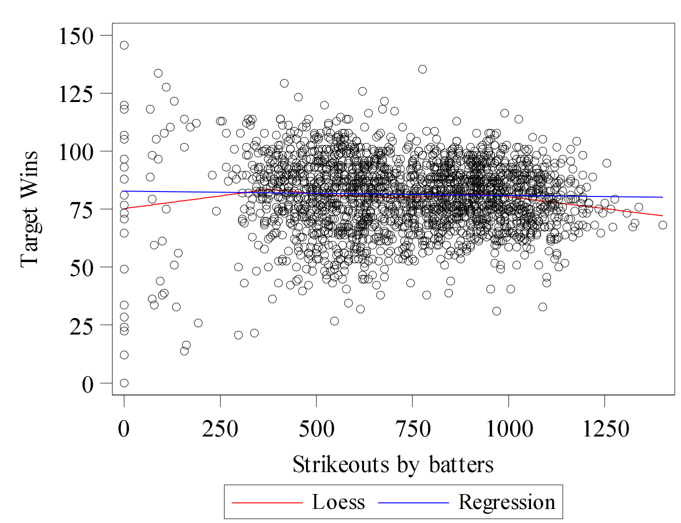


Figure 9: Scatter plot of TEAM_BATTING_SO vs Target Wins including Regression and Loess line

Variable: Stolen Bases

The variable "TEAM_BASERUN_SB" indicates the number of stolen bases by the team. This metric is seen to positively affect the number of wins.

Distribution Details

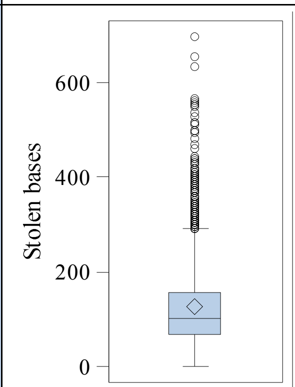
In Table 17 we can see in both the box plot and histogram the large presence of outliers on the high end of values. We may want to explore constraining the outliers with a Z transform and then further trimming them if necessary. We will explore further after looking at the scatterplot.

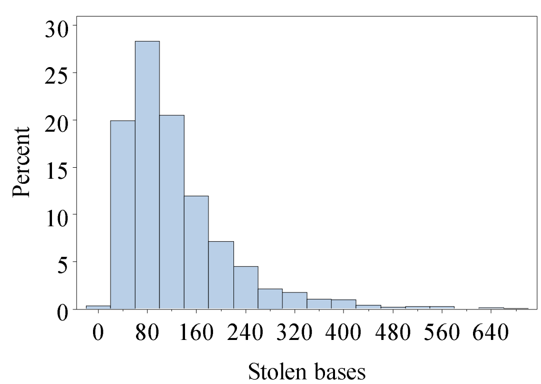
We also notice that there are 131 missing values. This equates to 5.3% missing. To handle missing values we explore a few methods. First we can replace the missing value with the mean: 124.7617716. The second method is to employ a decision tree such as in Figure 10 to produce the missing value logic. The decision tree in Figure 10 was produced by SPSS. Multiple trees were produced and reviewed. This tree was selected by looking to maximize tree simplicity and minimize the standard deviation in each tree node to a value as much below the normal TEAM_BASERUN_SB standard deviation (87.7911660) as possible. Note that 3 out of 4 nodes are well below the standard deviation of 87.7911660 Node 3 contains about 45% of the values and has more than cut the standard deviation in half. Unfortunately node 4 raises the standard deviation, but since it contains less than 10% of the values it was deemed to be a valid trade off.

Distribution Metrics

Analysis Variable : TEAM_BASERUN_SB Stolen bases									
N	N Miss	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Minimum	Maximum
2145	131	124.7617716	87.7911660	23.0000000	35.0000000	302.0000000	439.0000000	0	697.0000000

Graphical Distribution and Extreme Values





Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	1584	562	2023
0	1211	567	643
14	1825	632	642
18	2079	654	279
18	942	697	2022

% Missing

5.3% Missing

Table17: Distribution Details for Variable TEAM_BASERUN_SB

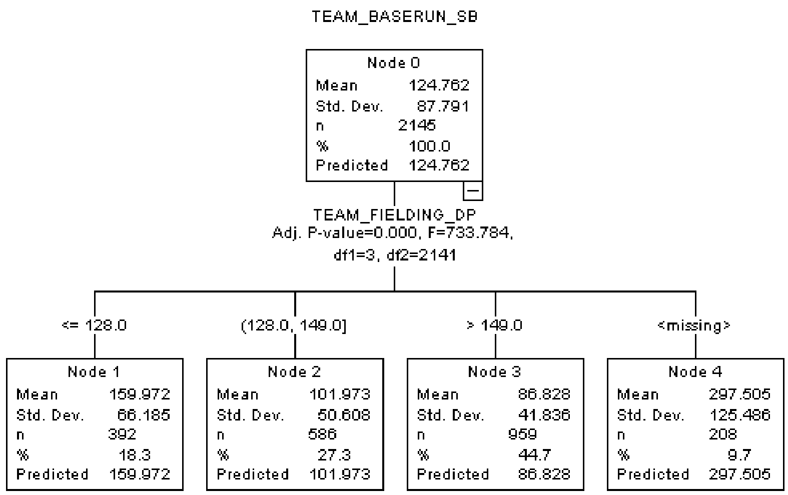


Figure 10: Decision Tree to replace the missing values for TEAM_BASERUN_SB

Correlation Details

Table 18 below outlines TEAM_BASERUN_SB's correlation with the other variables. TEAM_BASERUN_SB has a less than moderate correlation with TARGET_WINS and therefore it is not clear at the moment if it will be an obvious choice for the model. A few other variables appear to have a strong correlation: TEAM_BATTING_3B, TEAM_BASERUN_CS and TEAM_FIELDING_E (highlighted in green). There are also a few moderate correlations: TEAM_BATTING_HR, TEAM_PITCHING_HR and TEAM_FIELDING_DP. These relationships would suggest that we may need to watch for multicollinearity within our models.

TEAM_BASERUN_SB Correlation	**TARGET_WINS**	**TEAM_BATTING_H**	**TEAM_BATTING_2B**	**TEAM_BATTING_3B**	**TEAM_BATTING_HR**	**TEAM_BATTING_BB**
Pearson Correlation	0.13514	0.12357	-0.19976	0.53351	-0.45358	-0.10512
P Value	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
N	2145	2145	2145	2145	2145	2145

TEAM_BASERUN_SB Correlation	**TEAM_BATTING_SO**	**TEAM_BASERUN_SB**	**TEAM_BASERUN_CS**	**TEAM_BATTING_HBP**	**TEAM_PITCHING_H**
Pearson Correlation	-0.25449	1	0.65524	-0.064	0.07329
P Value	<.0001		<.0001	0.379	0.0007
N	2043	2145	1504	191	2145

TEAM_BASERUN_SB Correlation	**TEAM_PITCHING_HR**	**TEAM_PITCHING_BB**	**TEAM_PITCHING_SO**	**TEAM_FIELDING_E**	**TEAM_FIELDING_DP**
Pearson Correlation	-0.41651	0.14642	-0.13713	0.50963	-0.49708
P Value	<.0001	<.0001	<.0001	<.0001	<.0001
N	2145	2145	2043	2145	1937

Table 18: Set of Correlation tables for the variable TEAM_BASERUN_SB

In Figure 11 below we can see that the outlier values are affecting the regression line. In fact the LOESS line appears to have very obvious slope changes around the highest and the lowest values. This further exemplifies why we need to constrain the outliers. We will perform a Z transform to constrain them.

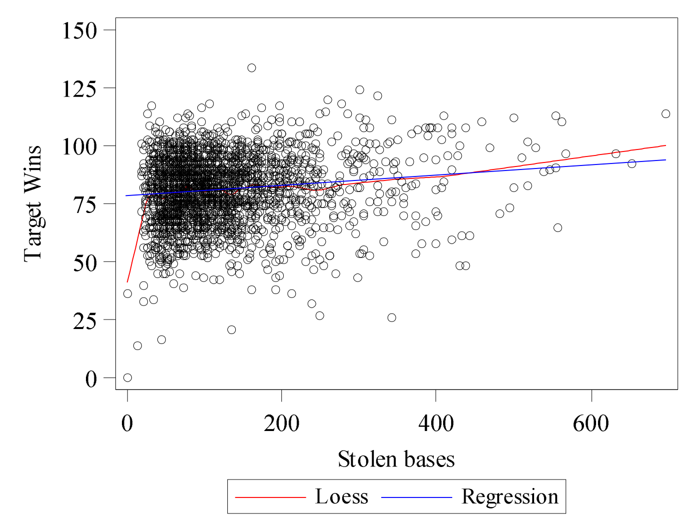


Figure 11: Scatter plot of TEAM_BASERUN_SB vs Target Wins including Regression and Loess line

In Figure 12 we can see how the standardization of the variable has already began to constrain the values (left chart). However, there are still a number of outlier values on the high end reaching as high as 6.6. We decide to take one step further to trim these values to a max of 3. The right chart below shows the newly trimmed and standardized variable.

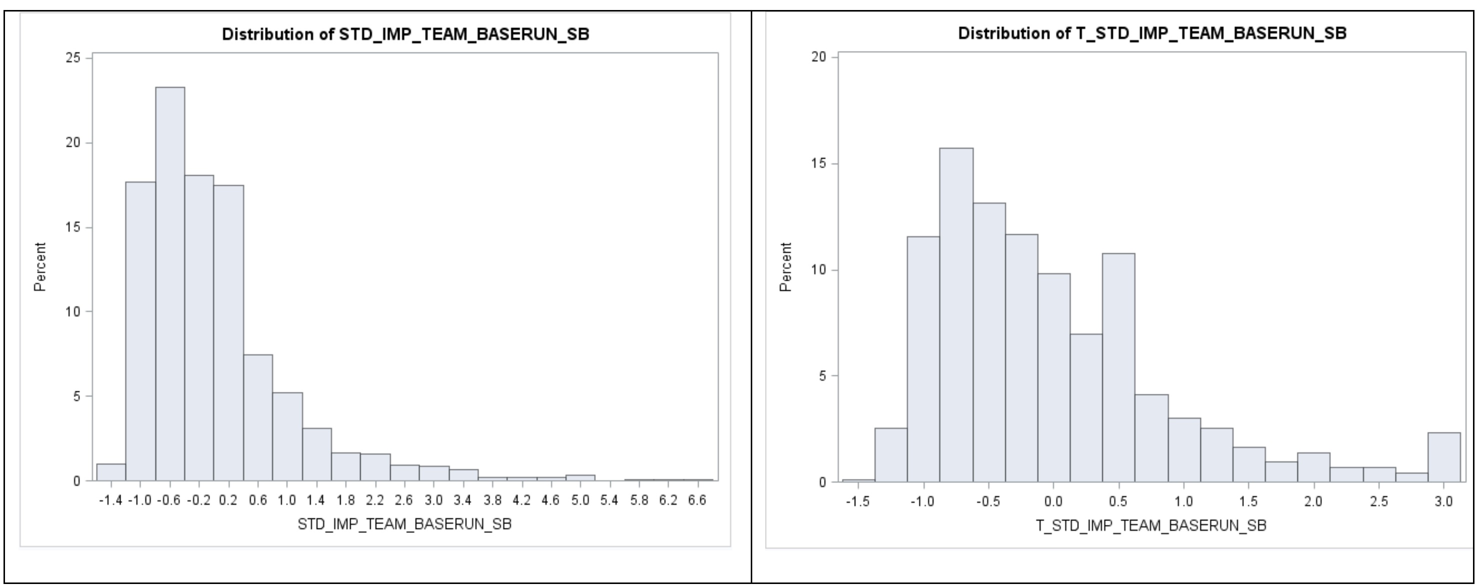


Figure 12: Standardized imputed TEAM_BASE_RUN_SB variable

Variable: TEAM_BASERUN_CS

The variable "TEAM_BASERUN_CS" indicates the number of times the team was caught stealing. This metric is seen to negatively affect the number of wins.

Distribution Details

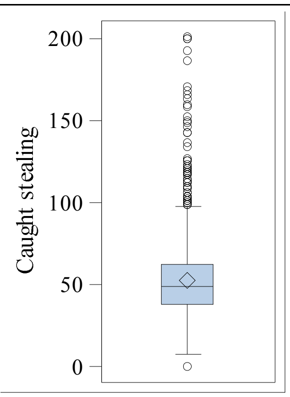
In Table 19 we can see in both the box plot and histogram the large presence of outliers on the high end of values. We may want to explore constraining the outliers with a Z transform and then further trimming them if necessary. We will explore further after looking at the scatterplot.

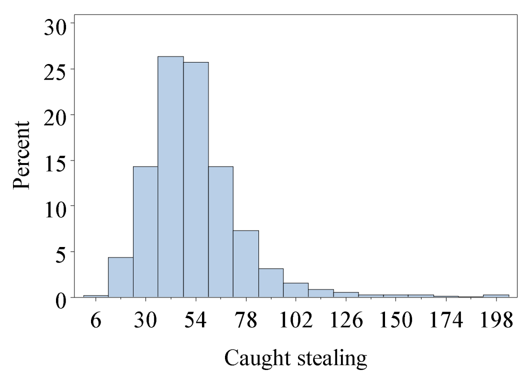
We also notice that there are 772 missing values. This equates to 34% missing. To handle missing values we explore a few methods. First we can replace the missing value with the mean: 52.8038564. The second method is to employ a decision tree such as in Figure 13 to produce the missing value logic. The decision tree in Figure 13 was produced by SPSS. Multiple trees were produced and reviewed. This tree was selected by looking to maximize tree simplicity and minimize the standard deviation in each tree node to a value as much below the normal TEAM_BASERUN_SB standard deviation (22.9563376) as possible. Note that 3 out of 4 nodes are well below the standard deviation of 22.9563376. Unfortunately node 4 raises the standard deviation, but not by a large amount. Since approximately 70% of the values (nodes 1-3) saw a lower standard deviation, it was deemed to be a valid tradeoff.

Distribution Metrics

Analysis Variable : TEAM_BASERUN_CS Caught stealing									
N	N Miss	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Minimum	Maximum
1504	772	52.8038564	22.9563376	16.0000000	24.0000000	91.0000000	143.0000000	0	201.0000000

Graphical Distribution and Extreme Values





Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	1211	186	313
7	1825	193	550
11	802	200	183
12	389	200	1503
14	1767	201	1409

% Missing

34% Missing

Table19: Distribution Details for Variable TEAM_BASERUN_CS

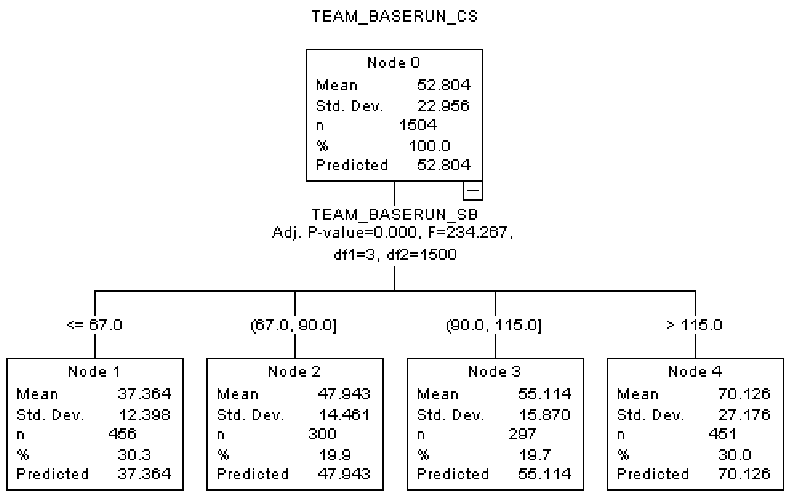


Figure 13: Decision Tree to replace the missing values for TEAM_BASERUN_CS

Correlation Details

Table 6 below outlines TEAM_BASERUN_CS's correlation with the other variables. TEAM_BASERUN_CS has an insignificant moderate correlation with TARGET_WINS and therefore it is not likely to make it into the model. TEAM_BASERUN_SB is the only variable with a strong correlation to TEAM_BASERUN_CS (highlighted in green). There are also a few moderate correlations: TEAM_BATTING_3B, TEAM_BATTING_HR and TEAM_PITCHING_HR. These relationships would suggest that we may need to watch for multicollinearity within our models.

TEAM_BASERUN_CS Correlation	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB
Pearson Correlation	0.0224	0.01671	-0.09981	0.34876	-0.43379	-0.13699
P Value	0.3853	0.5174	0.0001	<.0001	<.0001	<.0001
N	1504	1504	1504	1504	1504	1504

TEAM_BASERUN_CS Correlation	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H
Pearson Correlation	-0.21788	0.65524	1	-0.07051	-0.05201
P Value	<.0001	<.0001		0.3324	0.0437
N	1504	1504	1504	191	1504

TEAM_BASERUN_CS Correlation	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
Pearson Correlation	-0.42257	-0.10696	-0.21022	0.04832	-0.21425
P Value	<.0001	<.0001	<.0001	0.061	<.0001
N	1504	1504	1504	1504	1486

Table 20: Set of Correlation tables for the variable TEAM_BASERUN_CS

In Figure 14 below we can see that only the extreme low end values are affecting the regression line. The high end outlier values appear to make very little effect on the regression line. We will trim the variable to the 1st percentile due to the affect these values have on the LOESS line and we will trim to the 99th percentile due to the sheer number of high end outliers.

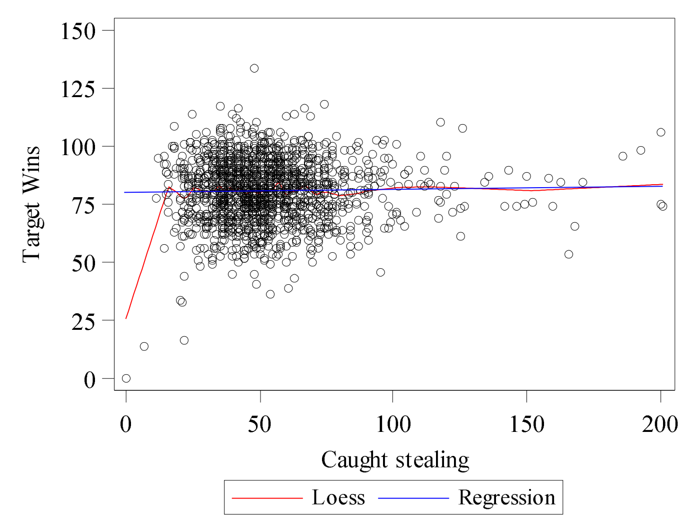


Figure 14: Scatter plot of TEAM_BASERUN_CS vs Target Wins including Regression and Loess line

Variable: Field Errors

The variable "TEAM_FIELDING_E" indicates the field errors. This metric is seen to negatively affect the number of wins.

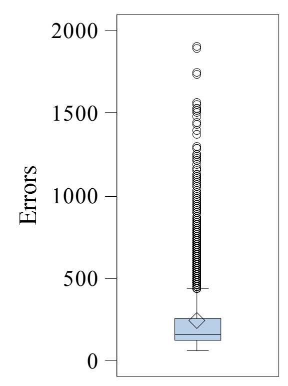
Distribution Details

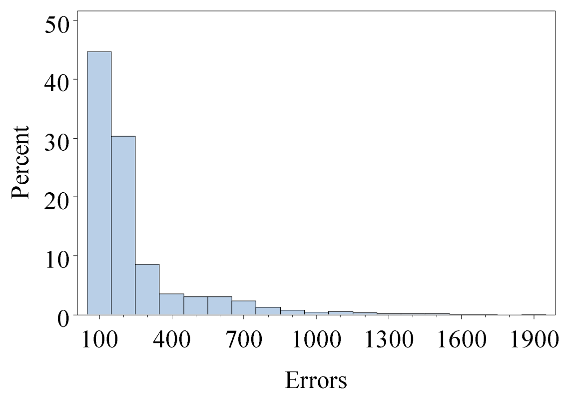
In Table 21 below we can see in the histogram that and the box plot that there are a large number of outliers on the high end of values. Therefore it makes sense to trim the data to the 95th percentile (716) to get rid of the high outliers. While the low end of the values does not exhibit any outliers, there are very few values under 100. Therefore the variable will also be trimmed to the 5th percentile (100)

Note also that there are no missing values in the training data set. However, it will still be useful to know the mean value of 246.4806678 as we may use it to error proof our final model against missing values.

Distribution Metrics	Analysis Variable : TEAM_FIELDING_E Errors									
	N	N Miss	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Minimum	Maximum
	2276	0	246.4806678	227.7709724	86.0000000	100.0000000	716.0000000	1237.00	65.0000000	1898.00

Graphical Distribution and Extreme Values





Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
65	1891	1567	391
66	390	1728	1584
68	1386	1740	1825
72	837	1890	1211
74	1335	1898	415

% Missing	0% Missing

Table 21: Distribution Details for Variable TEAM_FIELDING_E

Correlation Details

Table 22 below outlines TEAM_FIELDING_E's correlation with the other variables. TEAM_FIELDING_E has a less than moderate correlation with TARGET_WINS and therefore it is not clear at the moment if it will be an obvious choice for the model. A number of variables appear to have a strong correlation: TEAM_BATTING_3B, TEAM_BATTING_HR, TEAM_BATTING_BB, TEAM_BATTING_SO, TEAM_BASERUN_SB and TEAM_PITCHING_H (highlighted in green). There are also a few moderate correlations: TEAM_PITCHING_HR, TEAM_FIELDING_DP. These relationships would suggest that we may need to watch for multicollinearity within our models.

TEAM_FIELDING_E Correlation	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB
Pearson Correlation	-0.17648	0.2649	-0.23515	0.50978	-0.58734	-0.65597
P Value	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
N	2276	2276	2276	2276	2276	2276

TEAM_FIELDING_E Correlation	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H
Pearson Correlation	-0.58466	0.50963	0.04832	0.04179	0.66776
P Value	<.0001	<.0001	0.061	0.566	<.0001
N	2174	2145	1504	191	2276

TEAM_FIELDING_E Correlation	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
Pearson Correlation	-0.49314	-0.02284	-0.02329	1	-0.49768
P Value	<.0001	0.2761	0.2777		<.0001
N	2276	2276	2174	2276	1990

Table 22: Set of Correlation tables for the variable TEAM_FIELDING_E

In Figure 15 below we can see that the high end outlier values are affecting the regression line. This further exemplifies why it is necessary to trim the high value outlier values.

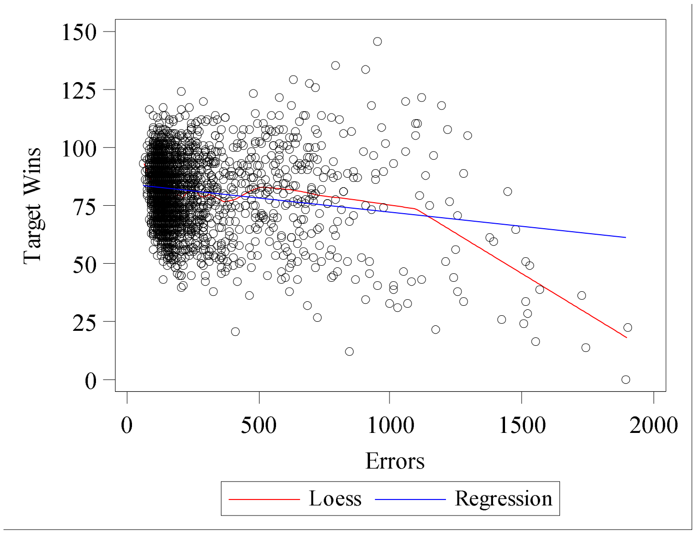


Figure 15: Scatter plot of TEAM_FIELDING_E vs Target Wins including Regression and Loess line

Variable: Fielding Double Plays

The variable "TEAM_FIELDING_DP" indicates the number of team fielding double plays. This metric is seen to positively affect the number of wins.

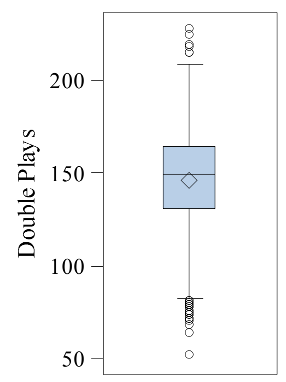
Distribution Details

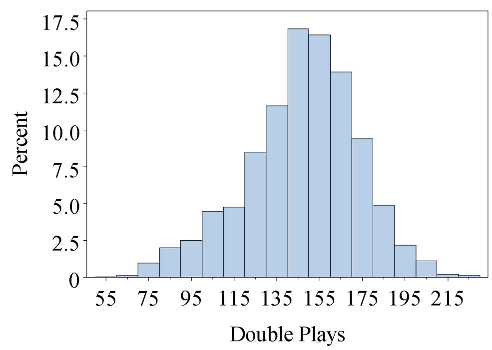
In Table 23 we can see that the variable TEAM_FIELDING_DP has a few outliers on both the high end and the low end of the data set. We will examine their effect on the regression line when we look at the scatter plot.

We can also see that the variable is missing 286 values in total, which is 12.6 percent of the data. The variable is fairly normally distributed with the highest frequency values distributed right around the mean 146.3879397. As such, the missing values will be replaced with the mean.

Distribution Metrics	Analysis Variable : TEAM_FIELDING_DP Double Plays									
	N	N Miss	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Minimum	Maximum
	1990	286	146.3879397	26.2263853	79.0000000	98.0000000	186.0000000	204.0000000	52.0000000	228.0000000

Graphical Distribution and Extreme Values





Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
52	1397	215	1777
64	996	218	342
68	542	219	1535
71	58	225	1445
72	2042	228	1534

% Missing 12.6% Missing

Table 23: Distribution Details for Variable TEAM_FIELDING_DP

Correlation Details

Table 24 below outlines TEAM_FIELDING_DP's correlation with the other variables. TEAM_FIELDING_DP has an insignificant moderate correlation with TARGET_WINS and therefore it is not likely to make it into the model. There are no variables with a strong correlation to TEAM_FIELDING_DP. There are a few moderate correlations: TEAM_BATTING_3B, TEAM_BATTING_HR, TEAM_BATTING_BB, TEAM_BASERUN_SB, TEAM_PITCHING_HR, TEAM_PITCHING_BB and TEAM_FIELDING_E. These relationships would suggest that we may need to watch for multicollinearity within our models.

TEAM_FIELDING_DP Correlation	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB
Pearson Correlation	-0.03485	0.15538	0.29088	-0.32307	0.44899	0.43088
P Value	0.1201	<.0001	<.0001	<.0001	<.0001	<.0001
N	1990	1990	1990	1990	1990	1990

TEAM_FIELDING_DP Correlation	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H
Pearson Correlation	0.15489	-0.49708	-0.21425	-0.07121	-0.22865
P Value	<.0001	<.0001	<.0001	0.3276	<.0001
N	1888	1937	1486	191	1990

TEAM_FIELDING_DP Correlation	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
Pearson Correlation	0.43917	0.32446	0.02616	-0.49768	1
P Value	<.0001	<.0001	0.2559	<.0001	
N	1990	1990	1888	1990	1990

Table 24: Set of Correlation tables for the variable TEAM_FIELDING_DP

In Figure 16 below we can see that the small amounts of outlier values are not affecting the regression line. Therefore no further action will be taken on the outlier values for this variable.

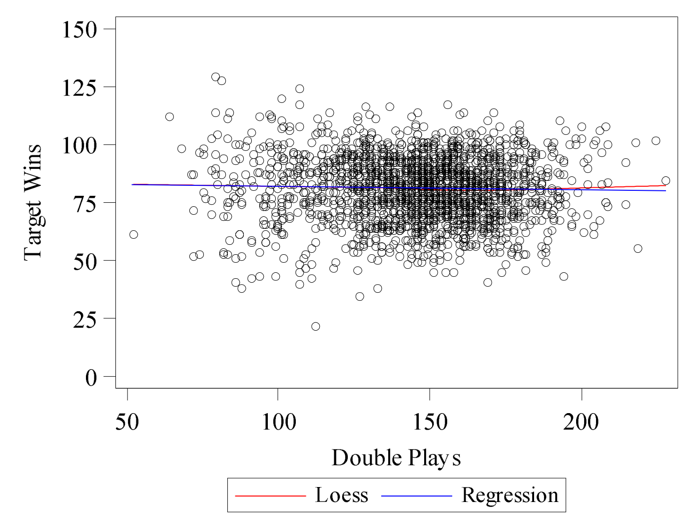


Figure 16: Scatter plot of TEAM_FIELDING_DP vs Target Wins including Regression and Loess line

Variable: Walks Allowed

The variable "TEAM_PITCHING_BB" indicates the number of walks allowed. This metric is seen to negatively affect the number of wins.

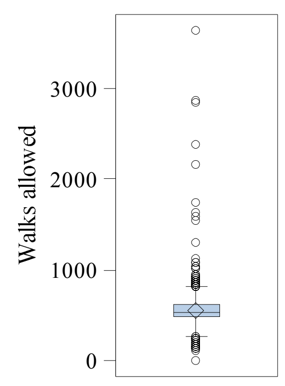
Distribution Details

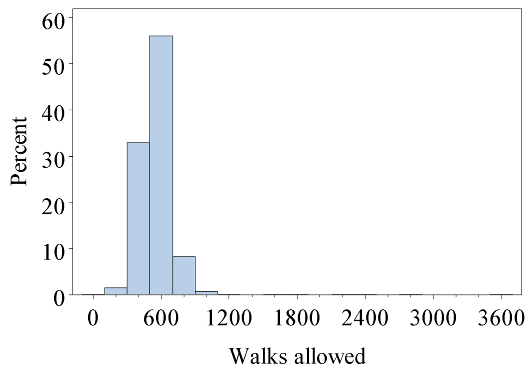
In Table 25 below we can see in both the histogram and box plot there are some outliers on both the high end and low end of values. The histogram has a very long tail and the outliers on the high end can get quite extreme, given that the maximum value is over 3 times that of the 99th percentile. We will examine their effect on the regression line when we look at the scatter plot.

Note also that there are no missing values in the training data set. However, it will still be useful to know the mean value of 553.0079086 as we may use it to error proof our final model against missing values.

Distribution Metrics	Analysis Variable : TEAM_PITCHING_BB Walks allowed									
	N	N Miss	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Minimum	Maximum
	2276	0	553.0079086	166.3573617	237.0000000	377.0000000	757.0000000	924.0000000	0	3645.00

Graphical Distribution and Extreme Values





Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	1211	2169	1340
119	1350	2396	1083
124	1824	2840	282
131	299	2876	2136
140	861	3645	1342

% Missing	0% Missing

Table 25: Distribution Details for Variable TEAM_PITCHING_BB

Correlation Details

Table 26 below outlines TEAM_PITCHING_BB's correlation with the other variables. TEAM_PITCHING_BB has a less than moderate correlation with TARGET_WINS and therefore it is not clear at the moment if it will be an obvious choice for the model. A number of variables appear to have a strong correlation: TEAM_BATTING_BB,TEAM_PITCHING_H, TEAM_PITCHING_SO and TEAM_FIELDING_DP (highlighted in green). There are no moderate correlations. These relationships would suggest that we may need to watch for multicollinearity within our models.

TEAM_PITCHING_BB Correlation	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB
Pearson Correlation	0.12417	0.09419	0.17805	-0.00222	0.13693	0.48936
P Value	<.0001	<.0001	<.0001	0.9155	<.0001	<.0001
N	2276	2276	2276	2276	2276	2276

TEAM_PITCHING_BB Correlation	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H
Pearson Correlation	0.03701	0.14642	-0.10696	0.04785	0.32068
P Value	0.0845	<.0001	<.0001	0.511	<.0001
N	2174	2145	1504	191	2276

TEAM_PITCHING_BB Correlation	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
Pearson Correlation	0.22194	1	0.4885	-0.02284	0.32446
P Value	<.0001		<.0001	0.2761	<.0001
N	2276	2276	2174	2276	1990

Table 26: Set of Correlation tables for the variable TEAM_PITCHING_BB

In Figure 17 below we can see that the outlier values are affecting the regression line. In fact the LOESS line appears to have very obvious slope changes right around the 1th and 99th percentiles. This further exemplifies why we need to deal with the outliers. Since there are not enough values to build a different model for high and low outliers, it is best that we trim the values to the 1th and 99th percentiles.

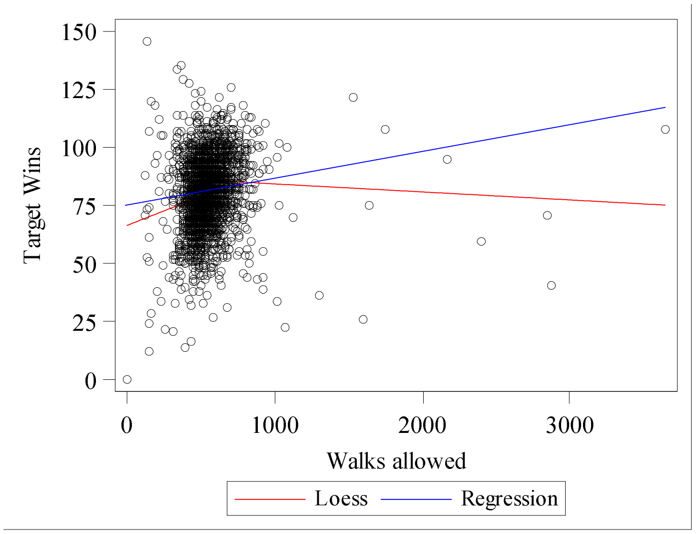


Figure 17: Scatter plot of TEAM_PITCHING_BB vs Target Wins including Regression and Loess line

Variable: Hits Allowed

The variable "TEAM_PITCHING_H" indicates the number of hits allowed. This metric is seen to negatively affect the number of wins.

Distribution Details

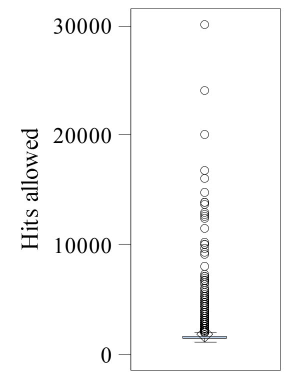
In Table 27 below we can see in both the histogram and box plot there are a lot of outliers on the high end of the values. The histogram has a very long tail and the outliers on the high end can get quite extreme, given that the maximum value is over 4 times that of the 99th percentile. We will examine their effect on the regression line when we look at the scatter plot.

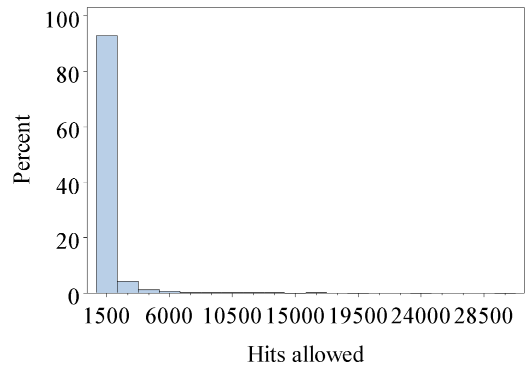
Note also that there are no missing values in the training data set. However, it will still be useful to know the mean value of 1779.21 as we may use it to error proof our final model against missing values.

Distribution Metrics

Analysis Variable : TEAM_PITCHING_H Hits allowed									
N	N Miss	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Minimum	Maximum
2276	0	1779.21	1406.84	1244.00	1316.00	2563.00	7093.00	1137.00	30132.00

Graphical Distribution and Extreme Values





Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1137	1456	16038	1342
1168	1353	16871	415
1184	1001	20088	2136
1187	232	24057	1211
1202	1354	30132	1584

% Missing	0% Missing

Table 27: Distribution Details for Variable TEAM_PITCHING_H

Correlation Details

Table 28 below outlines TEAM_PITCHING_H's correlation with the other variables. TEAM_PITCHING_H has a less than moderate correlation with TARGET_WINS and therefore it is not clear at the moment if it will be an obvious choice for the model. Only TEAM_FIELDING_E appears to have a strong correlation (highlighted in green). There are also a few moderate correlations: TEAM_BATTING_H, TEAM_BATTING_BB, TEAM_BATTING_SO, TEAM_PITCHING_BB. These relationships would suggest that we may need to watch for multicollinearity within our models.

TEAM_PITCHING_H Correlation	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB
Pearson Correlation	-0.10994	0.30269	0.02369	0.19488	-0.25015	-0.44978
P Value	<.0001	<.0001	0.2585	<.0001	<.0001	<.0001
N	2276	2276	2276	2276	2276	2276

TEAM_PITCHING_H Correlation	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H
Pearson Correlation	-0.37569	0.07329	-0.05201	-0.0277	1
P Value	<.0001	0.0007	0.0437	0.7037	
N	2174	2145	1504	191	2276

TEAM_PITCHING_H Correlation	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
Pearson Correlation	-0.14161	0.32068	0.26725	0.66776	-0.22865
P Value	<.0001	<.0001	<.0001	<.0001	<.0001
N	2276	2276	2174	2276	1990

Table 28: Set of Correlation tables for the variable TEAM_PITCHING_H

In Figure 18 below we can see that the high end outlier values are affecting the regression line. The visible outliers appear to be occurring roughly around the 95th percentile (2563). As such the variable will be trimmed to the 95th percentile on the high end. Although there were no low end outliers, the variable will still be trimmed to the 1st percentile (1244) on the low end as there are minimal values less than this amount.

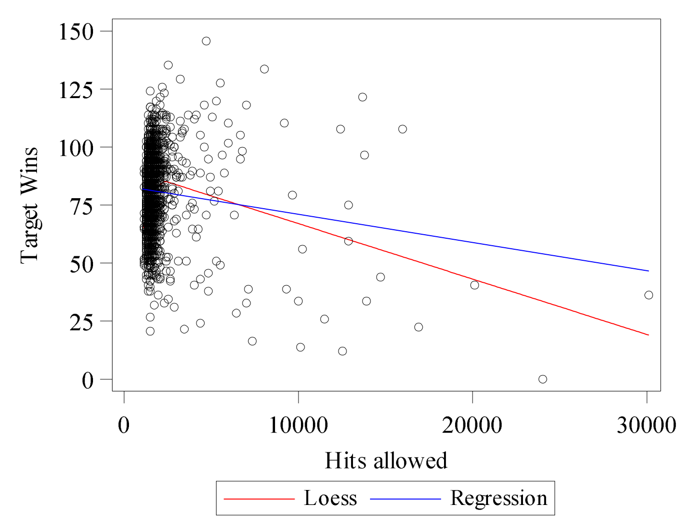


Figure 18: Scatter plot of TEAM_PITCHING_H vs Target Wins including Regression and Loess line

Variable: Homeruns Allowed

The variable "TEAM_PITCHING_HR" indicates the number of homeruns allowed. This metric is seen to negatively affect the number of wins.

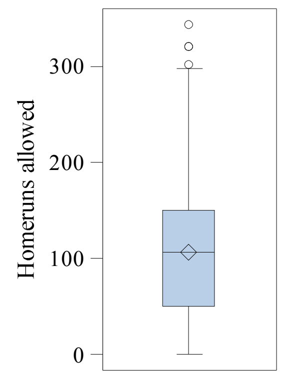
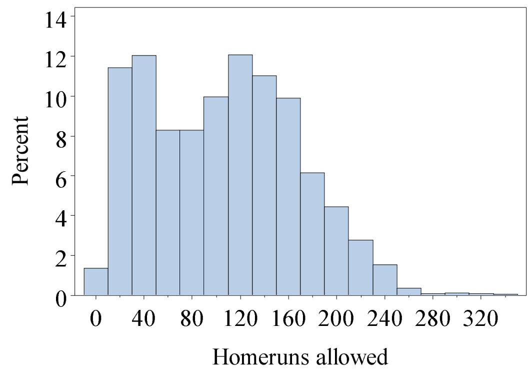
Distribution Details

In Table 29 below we can see in both the histogram and box plot there are a few outliers on the high end of the values. We will examine their effect on the regression line when we look at the scatter plot.

Note also that there are no missing values in the training data set. However, it will still be useful to know the mean value of 1779.21 as we may use it to error proof our final model against missing values.

.

Distribution Metrics	Analysis Variable : TEAM_PITCHING_HR Homeruns allowed									
	N	N Miss	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Minimum	Maximum
	2276	0	105.6985940	61.2987469	8.0000000	18.0000000	210.0000000	244.0000000	0	343.0000000

Graphical Distribution and Extreme Values		
% Missing	0% Missing	

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	2239	297	426
0	2233	301	1810
0	2136	320	964
0	2016	320	1882
0	2015	343	832

Table29: Distribution Details for Variable TEAM_PITCHING_HR

Correlation Details

Table 30 below outlines TEAM_PITCHING_HR's correlation with the other variables. TEAM_PITCHING_HR has a less than moderate correlation with TARGET_WINS and therefore it is not clear at the moment if it will be an obvious choice for the model. It appears to have a strong correlation with TEAM_BATTING_3B, TEAM_BATTING_HR and TEAM_BATTING_SO. There are also a few moderate correlations: TEAM_BATTING_2B, TEAM_BATTING_BB, TEAM_BASERUN_SB, TEAM_BASERUN_CS, TEAM_FIELDING_E and TEAM_FIELDING_DP. This variable is highly and moderately correlated to many other variables. These relationships would suggest that this variable could certainly contribute to high multicollinearity within our models.

Of particular note is the extremely high correlation to TEAM_BATTING_HR, with a statistically significant value of 0.96937 (highlighted in red). As examined in figure 4 it was ensured that the variables have a linear relationship and a new variable COMB_HR was created combining the two variables: TEAM_BATTING_HR and TEAM_PITCHING_HR.

TEAM_PITCHING_HR Correlation	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB
Pearson Correlation	0.18901	0.07285	0.45455	-0.56784	0.96937	0.45955
P Value	<.0001	0.0005	<.0001	<.0001	<.0001	<.0001
N	2276	2276	2276	2276	2276	2276

TEAM_PITCHING_HR Correlation	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H
Pearson Correlation	0.66718	-0.41651	-0.42257	0.10676	-0.14161
P Value	<.0001	<.0001	<.0001	0.1416	<.0001
N	2174	2145	1504	191	2276

TEAM_PITCHING_HR Correlation	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
Pearson Correlation	1	0.22194	0.20588	-0.49314	0.43917
P Value		<.0001	<.0001	<.0001	<.0001

N	2276	2276	2174	2276	1990

Table 30: Set of Correlation tables for the variable TEAM_PITCHING_HR

In Figure 19 below we can see that the extreme high and low end values are affecting the regression line. In fact the LOESS line appears to have very obvious slope changes roughly around the 1th and 99th percentiles. Since there are not enough values to build a different model for the extreme high and low values, it is best that we trim the values to the 1th and 99th percentiles.

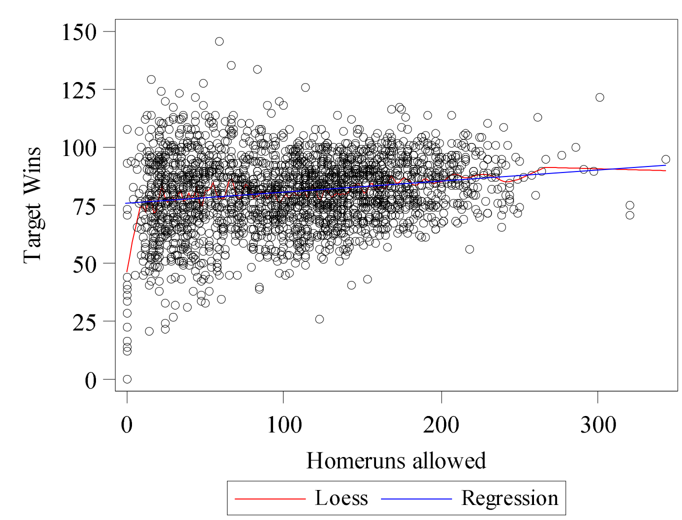


Figure 19: Scatter plot of TEAM_PITCHING_HR vs Target Wins including Regression and Loess line

Variable: Strikeouts by Pitcher

The variable "TEAM_PITCHING_SO" indicates the number of strikes out by pitcher. This metric is seen to positively affect the number of wins.

Distribution Details

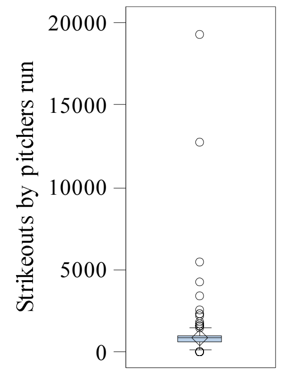
In Table 31 we can see in both the histogram and box plot we can see that there are a very small number of outliers on the low end of values and a greater amount of outliers on the high end of values. Also the actual values of the outliers are very extreme on the high end causing a very long tail in the histogram. To illustrate this, note that the maximum value is more than 13 times the 99th percentile value. We will examine their effect on the regression line when we look at the scatter plot.

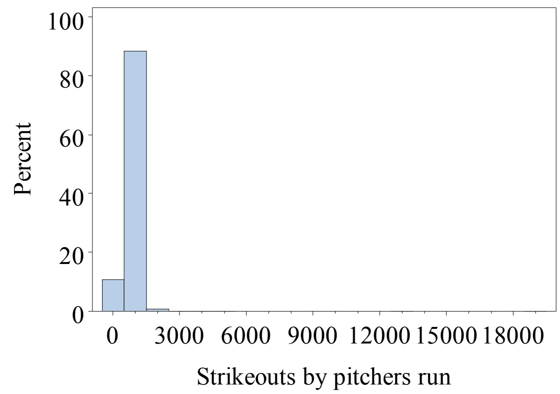
Also note that there are 102 missing values accounting for 4.5 percent of the observations. Given the extreme high values it makes the most sense to trim the values first and then look at replacing the missing values. This is done to prevent from further values being skewed by the extreme values.

Distribution Metrics

Analysis Variable : TEAM_PITCHING_SO Strikeouts by pitchers run									
N	N Miss	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Minimum	Maximum
2174	102	817.7304508	553.0850315	205.0000000	420.0000000	1173.00	1474.00	0	19278.00

Graphical Distribution and Extreme Values





Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	2239	3450	282
0	2233	4224	1826
0	2016	5456	1
0	2015	12758	1342
0	1824	19278	2136

% Missing

4.5% Missing

Table 31 Distribution Details for Variable TEAM_PITCHING_SO

Correlation Details

Table 32 below outlines TEAM_PITCHING_SO's correlation with the other variables. TEAM_PITCHING_SO has a very correlation with TARGET_WINS and therefore it is not clear at the moment if it will be an obvious choice for the model. There are no other variables with a strong correlation to TEAM_PITCHING_SO. There are a few moderate correlations: TEAM_BATTING_SO and TEAM_PITCHING_BB. These relationships would suggest that we may need to watch for multicollinearity within our models.

TEAM_PITCHING_SO Correlation	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB
Pearson Correlation	-0.07844	-0.25266	0.06479	-0.25882	0.18471	-0.02076
P Value	0.0003	<.0001	0.0025	<.0001	<.0001	0.3334
N	2174	2174	2174	2174	2174	2174

TEAM_PITCHING_SO Correlation	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H
Pearson Correlation	0.41623	-0.13713	-0.21022	0.22157	0.26725
P Value	<.0001	<.0001	<.0001	0.0021	<.0001
N	2174	2043	1504	191	2174

TEAM_PITCHING_SO Correlation	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
Pearson Correlation	0.20588	0.4885	1	-0.02329	0.02616
P Value	<.0001	<.0001		0.2777	0.2559
N	2174	2174	2174	2174	1888

Table 32: Set of Correlation tables for the variable TEAM_PITCHING_SO

In Figure 20 below we can see that the extreme high values appear to be pulling the regression line down. Since there are not enough values to build a different model for the extreme high and low values, it is best to trim the values to the 1st and 99th percentiles.

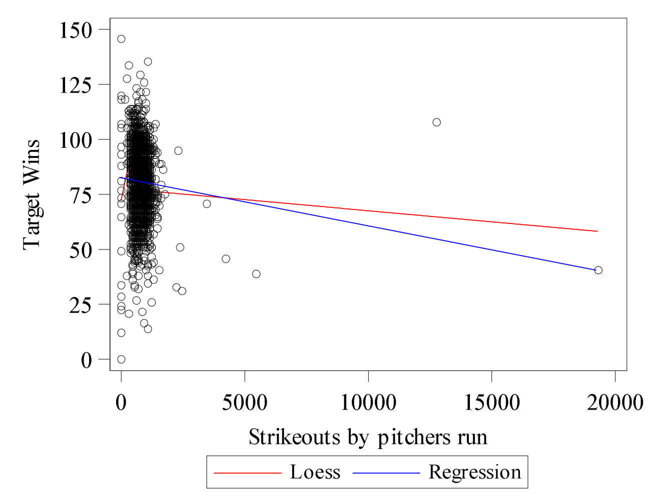


Figure 20: Scatter plot of TEAM_PITCHING_SO vs Target Wins including Regression and Loess line

As seen in Figure 21, after trimming the values, new histogram of the trimmed variable T99_TEAM_PITCHING_SO is pulled to ensure that the distribution has changed significantly. As confirmed by our graph we can now conclude that the distribution no longer includes the extreme values. We then re-pull the distribution metrics for the variable in Table 33 below. The new average is used for replacing missing values. Note that using trees to replace this missing value was investigated. However, a tree with sufficient standard deviation metrics could not be found.

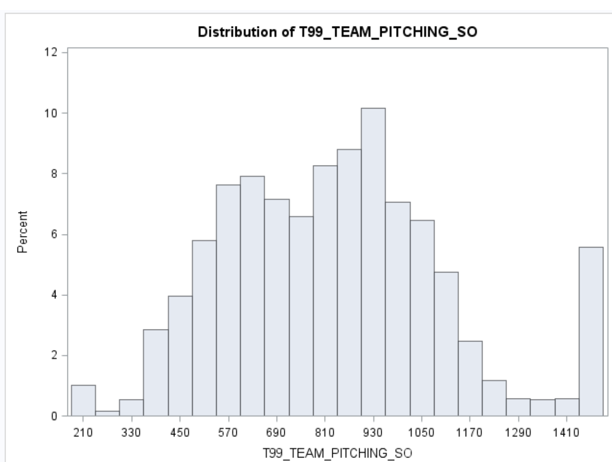


Figure 21: Histogram of new trimmed variable T99_TEAM_PITCHING_SO

Analysis Variable : T99_TEAM_PITCHING_SO									
N	N Miss	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Minimum	Maximum
2276	0	830.2355009	274.5725659	208.0000000	423.0000000	1474.00	1474.00	205.0000000	1474.00

Table 33: Distribution metrics for new trimmed variable T99_TEAM_PITCHING_SO

Part 2: Data Preparation

*New Variables in Blue

TEAM_BATTING_H - Base Hits by Batters

1. **Missing Value Clause**
    a. **Variable:** Contained within the existing variable TEAM_BATTING_H
    b. **Details**: Although there were no missing values in the training set an IF statement was issued to assign the average value 1469.27, in the case of missing values in the test or production set.
2. **Trim to the 5th and 95th Percentiles**
    - **Variable:** T95_TEAM_BATTING_H
    - **Details**: The data was trimmed to the 5th (1280) and 95th (1696) percentile due to the outliers found in the histogram and box plot. The scatterplot and LOESS line also illustrated the impact that the outliers may have on a regression formula. Therefore it was not right to keep the values in the regular model but there were not enough outliers to separate them into a different model. This further confirms the decision to trim to the 5th and 95th percentile.

TEAM_BATTING_2B - Doubles by Batters

3. **Missing Value Clause**
    a. **Variable:** Contained within the existing variable TEAM_BATTING_2B
    b. **Details**: Although there were no missing values in the training set, an IF statement was issued to assign the average value 241.2469244, in the case of missing values in the test or production set.
4. **Trim to the 5 Outlier Values**
    - **Variable:** TRIM_TEAM_BATTING_2B
    - **Details**: The data was trimmed to remove the 5 outlier values found in the histogram and box plot. The LOESS line appears to have slope changes around where the 5 outliers take place. This further exemplifies Details we need to deal with the outliers. Since there are not enough outlier values to trim the variable to a specific percentile, it is best that we simply trim the 5 outlier values.

TEAM_BATTING_3B - Triples by Batters

5. **Missing Value Clause**
    a. **Variable:** Contained within the existing variable TEAM_BATTING_3B
    b. **Details**: Although there were no missing values in the training set, an IF statement was issued to assign the average value 55.25, in the case of missing values in the test or production set.
6. **Trim to the 1st and 99th Percentiles**
    - **Variable:** T99_TEAM_BATTING_3B
    - **Details**: The data was trimmed to the 99th (134) percentile due to the outliers found on the high end in the histogram and box plot. There were no low end outliers found. However, the scatterplot and LOESS line illustrated there may be an impact on the regression formula for the lowest end values. As such a judgment call was made to also trim to the 1st percentile (17).

TEAM_BATTING_HR – Homeruns by Batters

7. **Missing Value Clause**
    a. **Variable:** Contained within the existing variable TEAM_BATTING_HR
    b. **Details**: Although there were no missing values in the training set, an IF statement was issued to assign the average value 99.6120387, in the case of missing values in the test or production set.
8. **Break Into Quantiles**
    - **Variable:** Quant_TEAM_BATTING_HR
    - **Details**: The data appeared to be pretty uniformly distributed and the regression line was relatively flat. As such, the variable was split into quantiles to see if that may expose some differences in the binned value groupings.
9. **Trim the high end values to 235**
    - **Variable:** TRIM_TEAM_BATTING_HR
    - **Details**: As per the histogram, the value frequency is relatively uniform until the value of approximately 235. At that value the frequency tapers off dramatically. As such the variable was trimmed on the high end to 235.
10. **Combine highly correlated variables TEAM_BATTING_HR and TEAM_PITCHING_HR**
    - **Variable:** COMB_HR
    - **Details**: Due to the high correlation between TEAM_BATTING_HR and TEAM_PITCHING_HR (0.96937) the two variables were combined. The values were simply multiplied to create the new variable COMB_HR.

TEAM_BATTING_BB – Walks by Batters

11. **Missing Value Clause**
    - **Variable:** Contained within the existing variable TEAM_BATTING_BB
    - **Details**: Although there were no missing values in the training set, an IF statement was issued to assign the average value 501.5588752, in the case of missing values in the test or production set.
12. **Perform a Z Transform**
    - **Variable:** STD_TEAM_BATTING_BB
    - **Details**: A Z transform was performed on the variable due to the outliers found in the histogram and box plot. The scatterplot and LOESS line also illustrated the impact that the outliers may have on a regression formula

TEAM_BATTING_HBP – Batters Hit by a Pitch

13. **Drop the Variable**
    - **Variable:** N/A
    - **Details**: The variable was missing a value for 91.6% of the observations. Given the large proportion of missing values it does not make sense to replace them. The best idea is to drop the variable all together. One last sanity check was performed: the scatterplot and regression lines were analyzed. The regression lines had virtually no slope indicating a lack of relationship with TARGET_WINs. Therefore we were able to safely drop this variable from our data set.

TEAM_BATTING_SO – Strikeouts by Batters

14. **Missing Value Flag**
    - **Variable:** M_TEAM_BATTING_SO
    - **Details**: A flag variable was created to track the observations with missing values for this variable. If the observation was missing this variable, the flag variable was given a value of 1. Otherwise the flag variable was given a value of 0.
15. **Replace Missing Value with Tree**
    - **Variable:** IMP_TEAM_BATTING_SO
    - **Details**: The decision tree is displayed in figure 8 and it was produced by SPSS. Multiple trees were produced and reviewed. This tree was selected by looking to maximize tree simplicity and minimize the standard deviation in each tree node to a value as much below the normal TEAM_BATTING_SO standard deviation (248.5264177) as possible.
16. **Replace Missing Value with Average**
    - **Variable:** IMP2_TEAM_BATTING_SO
    - **Details**: Missing values were replaced with the mean: 735.6053358. Replacing the missing values with average was employed as a tactic later on in the model building when the earlier models using tree logic replacement were not having much success.

TEAM_BASERUN_SB – Stolen Bases

17. **Missing Value Flag**
    - **Variable:** M_TEAM_BASERUN_SB
    - **Details**: A flag variable was created to track the observations with missing values for this variable. If the observation was missing this variable, the flag variable was given a value of 1. Otherwise the flag variable was given a value of 0.
18. **Replace Missing Value with Tree**
    - **Variable:** IMP_TEAM_BASERUN_SB
    - **Details**: The decision tree is displayed in figure 10 and it was produced by SPSS. Multiple trees were produced and reviewed. This tree was selected by looking to maximize tree simplicity and minimize the standard deviation in each tree node to a value as much below the normal TEAM_BASERUN_SB standard deviation (87.7911660) as possible.
19. **Standardized Imputed Variable - Missing Values Fixed with Tree**
    - **Variable:** STD_IMP_TEAM_BASERUN_SB
    - **Details**: In viewing the box plot and histogram it was clear this variable had a number of outliers on the high end. The scatterplot with the LOESS regression line further confirmed that the regression line was affected by the very high and low values. As such a Z transform was done to constrain outliers.
20. **Trimmed Standardized Imputed Variable - Missing Values Fixed with Tree**
    - **Variable:** T_STD_IMP_TEAM_BASERUN_SB
    - **Details**: Upon reviewing the newly standardized variables (above), it was clear that the high end outliers needed further trimming as they extended out to a value of 6.6. Therefore trimmed to a max value of 3
21. **Replace Missing Value with Average**
    - **Variable:** IMP2_TEAM_BASERUN_SB

- **Details**: Missing values were replaced with the mean: 124.7617716. Replacing the missing values with average was employed as a tactic later on in the model building when the earlier models using tree logic replacement were not having much success.

22. **Standardized Imputed Variable - Missing Values Fixed with Average**
    - **Variable:** STD_IMP2_TEAM_BASERUN_SB
    - **Details**: Following suit with variable 19, standardized the newly imputed variable
23. **Trimmed Standardized Imputed Variable - Missing Values Fixed with Average**
    - **Variable:** T_STD_IMP2_TEAM_BASERUN_SB
    - **Details**: Following suit with variable 20, trimmed the standardized newly imputed variable.

TEAM_BASERUN_CS– Caught Stealing

24. **Missing Value Flag**
    - **Variable:** M_TEAM_BASERUN_CS
    - **Details**: A flag variable was created to track the observations with missing values for this variable. If the observation was missing this variable, the flag variable was given a value of 1. Otherwise the flag variable was given a value of 0.
25. **Replace Missing Value with Tree**
    - **Variable:** IMP_TEAM_BASERUN_CS
    - **Details**: The decision tree is displayed in figure 10 and it was produced by SPSS. Multiple trees were produced and reviewed. This tree was selected by looking to maximize tree simplicity and minimize the standard deviation in each tree node to a value as much below the normal TEAM_BASERUN_CS standard deviation (87.7911660) as possible.
26. **Trimmed Imputed Variable - Missing Values Fixed with Tree**
    - **Variable:** T_STD_IMP_TEAM_BASERUN_CS
    - **Details**: The variable is trimmed to the 1st percentile due to the affect these values have on the LOESS line and further trimmed to the 99th percentile due to the sheer number of high end outliers.
27. **Replace Missing Value with Average**
    - **Variable:** IMP2_TEAM_BASERUN_CS
    - **Details**: Missing values were replaced with the mean: 124.7617716. Replacing the missing values with average was employed as a tactic later on in the model building when the earlier models using tree logic replacement were not having much success.
28. **Trimmed Standardized Imputed Variable - Missing Values Fixed with Average**
    - **Variable:** T_STD_IMP2_TEAM_BASERUN_CS
    - **Details**: Following suit with variable 26, trimmed the newly imputed variable to the 1st and 99th percentile.

TEAM_FIELDING_E – Errors

29. **Missing Value Clause**
    - **Variable:** Contained within the existing variable TEAM_FIELDING_E
    - **Details**: Although there were no missing values in the training set, an IF statement was issued to assign the average value 246.4806678, in the case of missing values in the test or production set.
30. **Perform a Z Transform**

- **Variable:** T95_TEAM_FIELDING_E
- **Details**: The histogram and the box plot show that there are a large number of outliers on the high end of values. Therefore it makes sense to trim the data to the 95th percentile (716) to get rid of the high outliers. While the low end of the values does not exhibit any outliers, there are very few values under 100. Therefore the variable will also be trimmed to the 5th percentile (100)

TEAM_FIELDING_DP – Double Plays

31. Missing Value Flag

- **Variable:** M_TEAM_FIELDING_DP
- **Details**: A flag variable was created to track the observations with missing values for this variable. If the observation was missing this variable, the flag variable was given a value of 1. Otherwise the flag variable was given a value of 0.

32. Replace Missing Value with Average

- **Variable:** IMP_TEAM_FIELDING_ DP
- **Details**: The variable is fairly normally distributed with the highest frequency values distributed right around the mean 146.3879397. Therefore the missing values will be replaced with the mean

TEAM_PITCHING_BB - Walks Allowed

33. Missing Value Clause

- **Variable:** Contained within the existing variable TEAM_PITCHING_BB
- **Details**: Although there were no missing values in the training set an IF statement was issued to assign the average value 553.0079086, in the case of missing values in the test or production set.

34. Trim to the 1st and 99th Percentiles

- **Variable:** T99_TEAM_PITCHING_BB
- **Details**: The data was trimmed to the 1st (237) and 99th (924) percentiles due to the outliers found in the histogram and box plot. The scatterplot and LOESS line also illustrated the impact that the outliers may have on a regression formula. Therefore it was not right to keep the values in the regular model but there were not enough outliers to separate them into a different model. This further confirms the decision to trim to the 1st and 99th percentiles.

TEAM_PITCHING_H - Hits Allowed

35. Missing Value Clause

- **Variable:** Contained within the existing variable TEAM_PITCHING_H
- **Details**: Although there were no missing values in the training set an IF statement was issued to assign the average value 1779.21, in the case of missing values in the test or production set.

36. Trim to the 1st and 95th Percentiles

- **Variable:** T99_TEAM_PITCHING_BB
- **Details**: The data was trimmed to the 1st (1244) and 95th (2563) percentiles due to the outliers found in the histogram and box plot. The scatterplot and LOESS line also illustrated the impact that the high end outliers may have on a regression formula. As there were a significant amount of outliers found on the high end of values the data was trimmed to the 95th percentile. The low end values were only trimmed to the 1st percentile as there did

not appear to be any low end outliers. Also the extreme low end values did not have much of an effect on the regression line.

TEAM_PITCHING_HR - Homeruns Allowed

37. **Missing Value Clause**
    - **Variable:** Contained within the existing variable TEAM_PITCHING_HR
    - **Details**: Although there were no missing values in the training set an IF statement was issued to assign the average value 105.6985940, in the case of missing values in the test or production set.

38. **Trim to the 1st and 99th Percentiles**
    - **Variable:** T99_TEAM_PITCHING_HR
    - **Details**: The data was trimmed to the 1st (8) and 99th (244) percentiles. The histogram and box plot displayed the few high end outliers. However, the scatterplot and LOESS line illustrated the impact that both the high and low extreme values may have on a regression formula. Therefore the variable was trimmed to the 1st (8) and 99th (244) percentiles.

TEAM_PITCHING_SO - Strikeouts by Pitchers

39. **Missing Value Flag**
    - **Variable:** M_TEAM_PITCHING_SO
    - **Details**: A flag variable was created to track the observations with missing values for this variable. If the observation was missing this variable, the flag variable was given a value of 1. Otherwise the flag variable was given a value of 0.

40. **Trim to the 1st and 99th Percentiles and Replace Missing Values**
    - **Variable:** T99_TEAM_PITCHING_SO
    - **Details**: The data was trimmed to the 1st (205) and 99th (1474) percentiles. The histogram and box plot displayed the few low end outliers and the extreme high end outliers. The scatterplot and LOESS line showed how the extreme high values were pulling down the regression line. Therefore the variable was trimmed to the 1st (205) and 99th (1474) percentiles. The variable was then re-examined to gather the new average value. The old average value was skewed by the extreme high values. Any missing values were replaced with the new average value of 830.2355009.

ALL VARIABLES - Principal Component Analysis

41. **Create Principal Component Variables PRIN1 to PRIN21**
    - **Variable:** PRIN1, PRIN2, PRIN3…. PRIN21
    - **Details**: Due to the high correlations between variables, principal component analysis was done to reduce the dimensionality of the data set. PROC PRINCOMP was performed on the transformed variables. Only one variable was present for each base variable (ie. Multiple transformations were not included). TARGET_WINS and INDEX were also removed before performing the procedure.
    - Table 34 contains is a list of Eigenvalues produced. As we are aiming for the highest predictive power possible, the top 9 Eigenvalues were selected (highlighted in yellow) as the best to use since cumulatively they represent 93 percent of the data variability.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	8.12828626	4.93602442	0.3871	0.3871
2	3.19226183	0.93714492	0.1520	0.5391
3	2.25511691	0.60297045	0.1074	0.6465
4	1.65214646	0.16648107	0.0787	0.7251
5	1.48566540	0.47408473	0.0707	0.7959
6	1.01158067	0.32160813	0.0482	0.8441
7	0.68997254	0.12163644	0.0329	0.8769
8	0.56833610	0.08884035	0.0271	0.9040
9	0.47949574	0.13120339	0.0228	0.9268
10	0.34829235	0.04938322	0.0166	0.9434
11	0.29890913	0.05212454	0.0142	0.9576
12	0.24678459	0.05678283	0.0118	0.9694
13	0.19000176	0.04524792	0.0090	0.9784
14	0.14475384	0.03197005	0.0069	0.9853
15	0.11278379	0.02104639	0.0054	0.9907
16	0.09173740	0.05518226	0.0044	0.9951
17	0.03655514	0.00686446	0.0017	0.9968
18	0.02969068	0.00529563	0.0014	0.9982
19	0.02439506	0.01116071	0.0012	0.9994
20	0.01323435	0.01323435	0.0006	1.0000
21	0.00000000		0.0000	1.0000

Table 34: Eigenvalues for Principal Component Analysis

- Figure 22 shows the Scree Plot for the Eigenvalues. At approximately component 9 the graph starts to flatten out, which is another good indicator of where to stop using values.

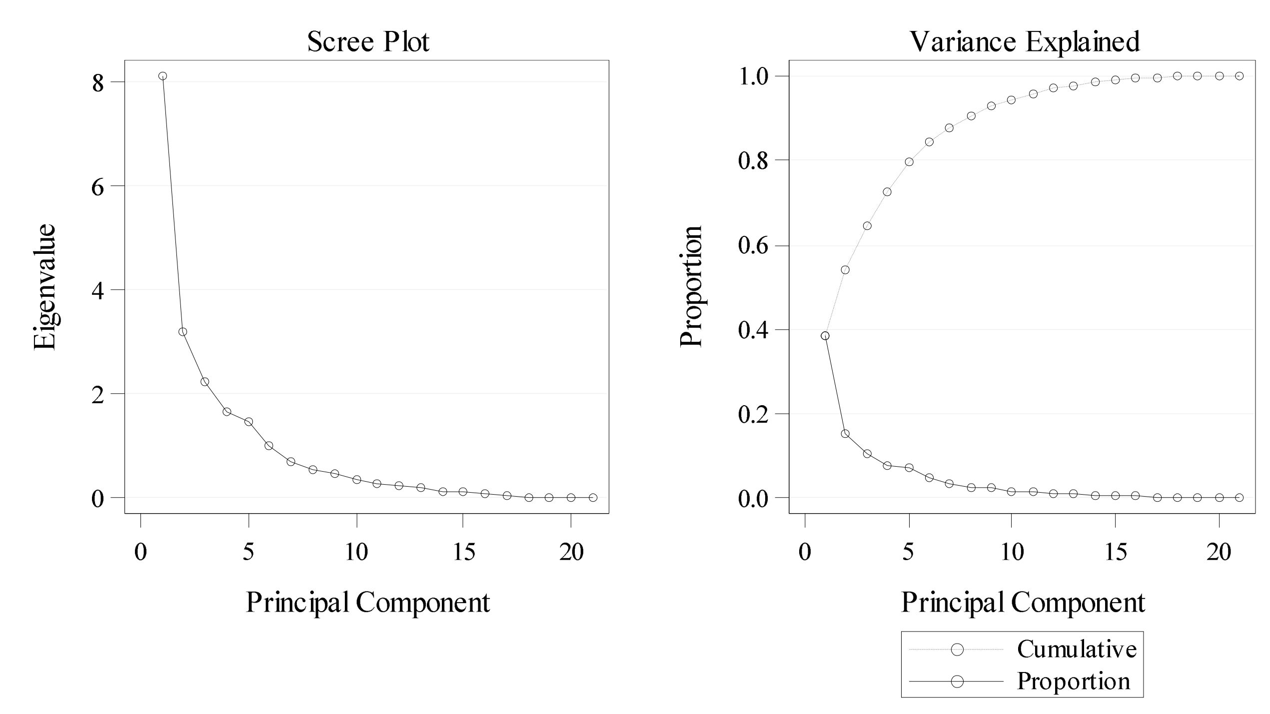


Figure 22: ScreePlot for Principal Component Analysis

- Figure 23 below shows the EigenVectors used to create the variables within a SAS data step.

Eigenvectors

		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13	Prin14	Prin15	Prin16	Prin17	Prin18	Prin19	Prin20	Prin21
TEAM_FIELDING_E	Errors	0.282732	0.194831	-.188775	0.091957	-.067745	-.094837	0.095525	0.071801	-.241779	-.262045	-.095270	-.078613	0.196450	-.194634	0.519649	0.097843	-.354186	0.234441	0.342051	-.119475	0.000000
T95_TEAM_BATTING_H		0.058617	0.361159	0.259760	0.069305	0.426537	0.196581	0.072841	-.039319	-.026966	0.069523	-.035349	-.007846	-.446897	0.424796	0.340880	-.036501	-.171549	-.147396	-.085808	-.057712	0.000000
TRIM_TEAM_BATTING_2B		-.131891	0.287646	0.141254	0.109587	0.377235	0.314832	-.099059	0.194124	0.442070	-.404559	0.041536	0.086716	0.385628	-.224078	-.097517	0.020574	0.031093	0.012213	-.013905	-.021898	0.000000
T99_TEAM_BATTING_3B		0.267028	0.074916	0.231547	0.051696	0.062357	0.076631	-.007979	-.111035	0.155469	0.609373	-.573057	0.039142	0.255658	-.201380	-.046106	0.025516	-.087014	0.053396	0.009293	-.003102	0.000000
QUANT_TEAM_BATTING_HR		-.307325	0.126510	-.072552	0.128572	0.064564	0.023244	0.161052	-.243350	-.170910	0.097241	0.051783	0.170091	0.051353	-.022638	0.005023	0.775004	0.059341	0.215201	-.229784	-.034183	0.000000
COMB_HR		-.279189	0.148311	-.112004	0.251355	0.102783	0.034996	0.136634	-.275448	-.251565	0.142899	0.140276	-.010564	0.197819	-.101344	0.051805	-.593258	-.023851	0.334934	-.310919	-.032358	0.000000
STD_TEAM_BATTING_BB		-.224488	-.021759	0.391776	0.196861	-.121189	-.249351	-.221754	-.006130	0.071615	0.052774	0.030190	0.377620	0.004757	0.138401	0.245425	-.076178	0.371940	0.225447	0.397267	-.217235	0.000000
M_TEAM_BATTING_SO		0.098795	-.385637	-.044371	0.327019	0.363715	-.169208	0.192491	0.147437	-.007916	-.002363	-.008221	0.064761	0.010109	0.026307	-.019752	0.018306	-.005176	0.007894	0.022437	0.080381	0.707107
IMP_TEAM_BATTING_SO		-.281887	-.044000	-.221224	0.173561	-.221849	0.195975	-.018084	0.010834	0.316191	0.055445	-.146368	0.021838	-.057477	0.191777	0.232376	-.015162	-.123740	0.140032	0.143233	0.690489	0.000000
M_TEAM_BASERUN_SB		0.175723	0.207785	-.408211	-.001360	0.034503	-.019142	-.231572	0.249286	0.016274	0.327846	0.320401	0.598151	0.049023	0.077696	-.096406	-.027845	-.216788	-.084684	0.007388	-.032575	0.000000
T_STD_IMP_TEAM_BASERUN_SB		0.210490	-.037521	0.152685	0.390536	-.279024	0.277282	0.139729	-.161687	0.136303	-.049431	0.200450	0.180381	-.472894	-.499674	-.057333	-.012448	-.072330	-.026434	0.006156	-.039960	0.000000
M_TEAM_BASERUN_CS		0.275353	-.030263	0.019055	0.108490	0.047771	-.154517	-.154914	-.440620	0.384937	0.164961	0.517195	-.355927	0.211623	0.157932	0.090152	0.108577	-.056286	0.018049	0.064248	0.001471	0.000000
T99_IMP_TEAM_BASERUN_CS		0.111126	-.193165	0.302429	0.090085	-.227509	0.556923	0.127017	0.297936	-.306819	0.146813	0.275832	-.068201	0.341493	0.255254	0.035308	0.058512	0.018026	0.005243	0.025536	0.007938	0.000000
T95_TEAM_FIELDING_E		0.306197	0.169197	-.145029	0.140636	-.104036	-.061514	0.043727	0.003021	-.039259	-.064065	-.054368	0.079406	0.145597	-.050857	0.373954	-.008633	0.611161	-.346538	-.330661	0.196784	0.000000
M_TEAM_FIELDING_DP		0.218856	0.232110	0.007741	0.213603	-.282996	-.132295	0.387929	-.166408	0.113032	-.283582	-.185918	0.179490	0.136517	0.472595	-.399854	-.055830	-.073469	0.037137	0.007484	-.070560	0.000000
IMP_TEAM_FIELDING_DP		-.156772	0.234721	0.133096	-.171325	-.109649	-.304910	0.594548	0.407318	0.276866	0.264512	0.236342	-.125719	-.022641	-.143834	0.093014	-.009395	0.008885	0.016814	-.013174	0.010675	0.000000
T99_TEAM_PITCHING_BB		-.099322	0.152133	0.340029	0.372354	-.187736	-.379933	-.374617	0.237478	-.165900	-.049295	0.002109	-.144389	0.053889	-.043047	-.078801	0.072881	-.331047	-.173661	-.307200	0.187591	0.000000
T95_TEAM_PITCHING_H		0.217829	0.359137	-.091176	0.152470	0.136098	0.012759	-.133982	0.207442	-.198459	0.092035	0.049181	-.308697	-.217785	-.014042	-.354961	0.049448	0.360289	0.386552	0.249319	0.222065	0.000000
T99_TEAM_PITCHING_HR		-.296584	0.178159	-.087173	0.201667	0.076096	0.018904	0.134880	-.196449	-.222957	0.121839	0.031526	-.087961	0.129911	-.103165	-.136478	-.017382	0.011991	-.615545	0.518090	0.043658	0.000000
T99_TEAM_PITCHING_SO		-.176902	-.033740	-.383446	0.373981	-.168786	0.131699	-.108754	0.248470	0.233976	0.129699	-.177535	-.323016	-.085606	0.132964	0.056965	0.035350	0.067306	-.053010	-.084591	-.556645	0.000000
M_TEAM_PITCHING_SO		0.098795	-.385637	-.044371	0.327019	0.363715	-.169208	0.192491	0.147437	-.007916	-.002363	-.008221	0.064761	0.010109	0.026307	-.019752	0.018306	-.005176	0.007894	0.022437	0.080381	-.707107

Figure 23: Eigenvectors for Principal Component Analysis

- Table 35 below shows the formulas used to create the variables within a SAS data step. The selected PCA variables (PRIN1-PRIN9) were initially used within the models. However, during model trials the remaining PCA variables were added as they increased the models predictive power.

- Note that the variable formulas were calculated automatically with an excel macros and the final formulas were copied into a SAS data step.

PRIN1	0.282732 * TEAM_FIELDING_E + 0.058617 * T95_TEAM_BATTING_H - 0.131891 * TRIM_TEAM_BATTING_2B + 0.267028 * T99_TEAM_BATTING_3B - 0.307325 * QUANT_TEAM_BATTING_HR - 0.279189 * COMB_HR - 0.224488 * STD_TEAM_BATTING_BB + 0.098795 * M_TEAM_BATTING_SO - 0.281887 * IMP_TEAM_BATTING_SO + 0.175723 * M_TEAM_BASERUN_SB + 0.21049 * T_STD_IMP_TEAM_BASERUN_SB + 0.275353 * M_TEAM_BASERUN_CS + 0.111126 * T99_IMP_TEAM_BASERUN_CS + 0.306197 * T95_TEAM_FIELDING_E + 0.218856 * M_TEAM_FIELDING_DP - 0.156772 * IMP_TEAM_FIELDING_DP - 0.099322 * T99_TEAM_PITCHING_BB + 0.217829 * T95_TEAM_PITCHING_H - 0.296584 * T99_TEAM_PITCHING_HR - 0.176902 * T99_TEAM_PITCHING_SO + 0.098795 * M_TEAM_PITCHING_SO
PRIN2	0.194831 * TEAM_FIELDING_E + 0.361159 * T95_TEAM_BATTING_H + 0.287646 * TRIM_TEAM_BATTING_2B + 0.074916 * T99_TEAM_BATTING_3B + 0.12651 * QUANT_TEAM_BATTING_HR + 0.148311 * COMB_HR - 0.021759 * STD_TEAM_BATTING_BB - 0.385637 * M_TEAM_BATTING_SO - 0.044 * IMP_TEAM_BATTING_SO + 0.207785 * M_TEAM_BASERUN_SB - 0.037521 * T_STD_IMP_TEAM_BASERUN_SB - 0.030263 * M_TEAM_BASERUN_CS - 0.193165 * T99_IMP_TEAM_BASERUN_CS + 0.169197 * T95_TEAM_FIELDING_E + 0.23211 * M_TEAM_FIELDING_DP + 0.234721 * IMP_TEAM_FIELDING_DP + 0.152133 * T99_TEAM_PITCHING_BB + 0.359137 * T95_TEAM_PITCHING_H + 0.178159 * T99_TEAM_PITCHING_HR - 0.03374 * T99_TEAM_PITCHING_SO - 0.385637 * M_TEAM_PITCHING_SO
PRIN3	-0.188775 * TEAM_FIELDING_E + 0.25976 * T95_TEAM_BATTING_H + 0.141254 * TRIM_TEAM_BATTING_2B + 0.231547 * T99_TEAM_BATTING_3B - 0.072552 * QUANT_TEAM_BATTING_HR - 0.112004 * COMB_HR + 0.391776 * STD_TEAM_BATTING_BB - 0.044371 * M_TEAM_BATTING_SO - 0.221224 * IMP_TEAM_BATTING_SO - 0.408211 * M_TEAM_BASERUN_SB + 0.152685 * T_STD_IMP_TEAM_BASERUN_SB + 0.019055 * M_TEAM_BASERUN_CS + 0.302429 * T99_IMP_TEAM_BASERUN_CS - 0.145029 * T95_TEAM_FIELDING_E + 0.007741 * M_TEAM_FIELDING_DP + 0.133096 * IMP_TEAM_FIELDING_DP + 0.340029 * T99_TEAM_PITCHING_BB - 0.091176 * T95_TEAM_PITCHING_H - 0.087173 * T99_TEAM_PITCHING_HR - 0.383446 * T99_TEAM_PITCHING_SO - 0.044371 * M_TEAM_PITCHING_SO
PRIN4	0.091957 * TEAM_FIELDING_E + 0.069305 * T95_TEAM_BATTING_H + 0.109587 * TRIM_TEAM_BATTING_2B + 0.051696 * T99_TEAM_BATTING_3B + 0.128572 * QUANT_TEAM_BATTING_HR + 0.251355 * COMB_HR + 0.196861 * STD_TEAM_BATTING_BB + 0.327019 * M_TEAM_BATTING_SO + 0.173561 * IMP_TEAM_BATTING_SO - 0.00136 * M_TEAM_BASERUN_SB + 0.390536 * T_STD_IMP_TEAM_BASERUN_SB + 0.10849 * M_TEAM_BASERUN_CS + 0.090085 * T99_IMP_TEAM_BASERUN_CS + 0.140636 * T95_TEAM_FIELDING_E + 0.213603 * M_TEAM_FIELDING_DP - 0.171325 * IMP_TEAM_FIELDING_DP + 0.372354 * T99_TEAM_PITCHING_BB + 0.15247 * T95_TEAM_PITCHING_H + 0.201667 * T99_TEAM_PITCHING_HR + 0.373981 * T99_TEAM_PITCHING_SO + 0.327019 * M_TEAM_PITCHING_SO
PRIN5	- 0.067745 * TEAM_FIELDING_E + 0.426537 * T95_TEAM_BATTING_H + 0.377235 * TRIM_TEAM_BATTING_2B + 0.062357 * T99_TEAM_BATTING_3B + 0.064564 * QUANT_TEAM_BATTING_HR + 0.102783 * COMB_HR - 0.121189 * STD_TEAM_BATTING_BB + 0.363715 * M_TEAM_BATTING_SO - 0.221849 * IMP_TEAM_BATTING_SO + 0.034503 * M_TEAM_BASERUN_SB - 0.279024 * T_STD_IMP_TEAM_BASERUN_SB + 0.047771 * M_TEAM_BASERUN_CS - 0.227509 * T99_IMP_TEAM_BASERUN_CS - 0.104036 * T95_TEAM_FIELDING_E - 0.282996 * M_TEAM_FIELDING_DP - 0.109649 * IMP_TEAM_FIELDING_DP - 0.187736 * T99_TEAM_PITCHING_BB + 0.136098 * T95_TEAM_PITCHING_H + 0.076096 * T99_TEAM_PITCHING_HR - 0.168786 * T99_TEAM_PITCHING_SO + 0.363715 * M_TEAM_PITCHING_SO
PRIN6	-0.094837 * TEAM_FIELDING_E + 0.196581 * T95_TEAM_BATTING_H + 0.314832 * TRIM_TEAM_BATTING_2B + 0.076631 * T99_TEAM_BATTING_3B + 0.023244 * QUANT_TEAM_BATTING_HR + 0.034996 * COMB_HR - 0.249351 * STD_TEAM_BATTING_BB - 0.169208 * M_TEAM_BATTING_SO + 0.195975 * IMP_TEAM_BATTING_SO - 0.019142 * M_TEAM_BASERUN_SB + 0.277282 * T_STD_IMP_TEAM_BASERUN_SB - 0.154517 * M_TEAM_BASERUN_CS + 0.556923 * T99_IMP_TEAM_BASERUN_CS - 0.061514 * T95_TEAM_FIELDING_E - 0.132295 * M_TEAM_FIELDING_DP - 0.30491 * IMP_TEAM_FIELDING_DP - 0.379933 * T99_TEAM_PITCHING_BB + 0.012759 * T95_TEAM_PITCHING_H + 0.018904 * T99_TEAM_PITCHING_HR + 0.131699 * T99_TEAM_PITCHING_SO - 0.169208 * M_TEAM_PITCHING_SO
PRIN7	+ 0.095525 * TEAM_FIELDING_E + 0.072841 * T95_TEAM_BATTING_H - 0.099059 * TRIM_TEAM_BATTING_2B - 0.007979 * T99_TEAM_BATTING_3B + 0.161052 * QUANT_TEAM_BATTING_HR + 0.136634 * COMB_HR - 0.221754 * STD_TEAM_BATTING_BB + 0.192491 * M_TEAM_BATTING_SO - 0.018084 * IMP_TEAM_BATTING_SO - 0.231572 * M_TEAM_BASERUN_SB + 0.139729 * T_STD_IMP_TEAM_BASERUN_SB - 0.154914 * M_TEAM_BASERUN_CS + 0.127017 * T99_IMP_TEAM_BASERUN_CS + 0.043727 * T95_TEAM_FIELDING_E + 0.387929 * M_TEAM_FIELDING_DP + 0.594548 * IMP_TEAM_FIELDING_DP - 0.374617 * T99_TEAM_PITCHING_BB - 0.133982 * T95_TEAM_PITCHING_H + 0.13488 * T99_TEAM_PITCHING_HR - 0.108754 * T99_TEAM_PITCHING_SO + 0.192491 * M_TEAM_PITCHING_SO
PRIN8	0.071801 * TEAM_FIELDING_E - 0.039319 * T95_TEAM_BATTING_H + 0.194124 * TRIM_TEAM_BATTING_2B - 0.111035 * T99_TEAM_BATTING_3B - 0.24335 * QUANT_TEAM_BATTING_HR - 0.275448 * COMB_HR - 0.00613 * STD_TEAM_BATTING_BB + 0.147437 * M_TEAM_BATTING_SO + 0.010834 * IMP_TEAM_BATTING_SO + 0.249286 * M_TEAM_BASERUN_SB - 0.161687 * T_STD_IMP_TEAM_BASERUN_SB - 0.44062 * M_TEAM_BASERUN_CS + 0.297936 * T99_IMP_TEAM_BASERUN_CS + 0.003021 * T95_TEAM_FIELDING_E - 0.166408 * M_TEAM_FIELDING_DP + 0.407318 * IMP_TEAM_FIELDING_DP + 0.237478 * T99_TEAM_PITCHING_BB + 0.207442 * T95_TEAM_PITCHING_H - 0.196449 * T99_TEAM_PITCHING_HR + 0.24847 * T99_TEAM_PITCHING_SO + 0.147437 * M_TEAM_PITCHING_SO
PRIN9	-0.241779 * TEAM_FIELDING_E - 0.026966 * T95_TEAM_BATTING_H + 0.44207 * TRIM_TEAM_BATTING_2B + 0.155469 * T99_TEAM_BATTING_3B - 0.17091 * QUANT_TEAM_BATTING_HR - 0.251565 * COMB_HR + 0.071615 * STD_TEAM_BATTING_BB - 0.007916 * M_TEAM_BATTING_SO + 0.316191 * IMP_TEAM_BATTING_SO + 0.016274 * M_TEAM_BASERUN_SB + 0.136303 * T_STD_IMP_TEAM_BASERUN_SB + 0.384937 * M_TEAM_BASERUN_CS - 0.306819 * T99_IMP_TEAM_BASERUN_CS - 0.039259 * T95_TEAM_FIELDING_E + 0.113032 * M_TEAM_FIELDING_DP + 0.276866 * IMP_TEAM_FIELDING_DP - 0.1659 * T99_TEAM_PITCHING_BB - 0.198459 * T95_TEAM_PITCHING_H - 0.222957 * T99_TEAM_PITCHING_HR + 0.233976 * T99_TEAM_PITCHING_SO - 0.007916 * M_TEAM_PITCHING_SO
PRIN10	-0.262045 * TEAM_FIELDING_E + 0.069523 * T95_TEAM_BATTING_H - 0.404559 * TRIM_TEAM_BATTING_2B + 0.609373 * T99_TEAM_BATTING_3B + 0.097241 * QUANT_TEAM_BATTING_HR + 0.142899 * COMB_HR + 0.052774 * STD_TEAM_BATTING_BB - 0.002363 * M_TEAM_BATTING_SO + 0.055445 * IMP_TEAM_BATTING_SO + 0.327846 * M_TEAM_BASERUN_SB - 0.049431 * T_STD_IMP_TEAM_BASERUN_SB + 0.164961 * M_TEAM_BASERUN_CS + 0.146813 * T99_IMP_TEAM_BASERUN_CS - 0.064065 * T95_TEAM_FIELDING_E - 0.283582 * M_TEAM_FIELDING_DP + 0.264512 * IMP_TEAM_FIELDING_DP - 0.049295 *

	T99_TEAM_PITCHING_BB + 0.092035 * T95_TEAM_PITCHING_H + 0.121839 * T99_TEAM_PITCHING_HR + 0.129699 * T99_TEAM_PITCHING_SO - 0.002363 * M_TEAM_PITCHING_SO
PRIN11	-0.09527 * TEAM_FIELDING_E - 0.035349 * T95_TEAM_BATTING_H + 0.041536 * TRIM_TEAM_BATTING_2B - 0.573057 * T99_TEAM_BATTING_3B + 0.051783 * QUANT_TEAM_BATTING_HR + 0.140276 * COMB_HR + 0.03019 * STD_TEAM_BATTING_BB - 0.008221 * M_TEAM_BATTING_SO - 0.146368 * IMP_TEAM_BATTING_SO + 0.320401 * M_TEAM_BASERUN_SB + 0.20045 * T_STD_IMP_TEAM_BASERUN_SB + 0.517195 * M_TEAM_BASERUN_CS + 0.275832 * T99_IMP_TEAM_BASERUN_CS - 0.054368 * T95_TEAM_FIELDING_E - 0.185918 * M_TEAM_FIELDING_DP + 0.236342 * IMP_TEAM_FIELDING_DP + 0.002109 * T99_TEAM_PITCHING_BB + 0.049181 * T95_TEAM_PITCHING_H + 0.031526 * T99_TEAM_PITCHING_HR - 0.177535 * T99_TEAM_PITCHING_SO - 0.008221 * M_TEAM_PITCHING_SO
PRIN12	-0.078613 * TEAM_FIELDING_E - 0.007846 * T95_TEAM_BATTING_H + 0.086716 * TRIM_TEAM_BATTING_2B + 0.039142 * T99_TEAM_BATTING_3B + 0.170091 * QUANT_TEAM_BATTING_HR - 0.010564 * COMB_HR + 0.37762 * STD_TEAM_BATTING_BB + 0.064761 * M_TEAM_BATTING_SO + 0.021838 * IMP_TEAM_BATTING_SO + 0.598151 * M_TEAM_BASERUN_SB + 0.180381 * T_STD_IMP_TEAM_BASERUN_SB - 0.355927 * M_TEAM_BASERUN_CS - 0.068201 * T99_IMP_TEAM_BASERUN_CS + 0.079406 * T95_TEAM_FIELDING_E + 0.17949 * M_TEAM_FIELDING_DP - 0.125719 * IMP_TEAM_FIELDING_DP - 0.144389 * T99_TEAM_PITCHING_BB - 0.308697 * T95_TEAM_PITCHING_H - 0.087961 * T99_TEAM_PITCHING_HR - 0.323016 * T99_TEAM_PITCHING_SO + 0.064761 * M_TEAM_PITCHING_SO
PRIN13	0.19645 * TEAM_FIELDING_E - 0.446897 * T95_TEAM_BATTING_H + 0.385628 * TRIM_TEAM_BATTING_2B + 0.255658 * T99_TEAM_BATTING_3B + 0.051353 * QUANT_TEAM_BATTING_HR + 0.197819 * COMB_HR + 0.004757 * STD_TEAM_BATTING_BB + 0.010109 * M_TEAM_BATTING_SO - 0.057477 * IMP_TEAM_BATTING_SO + 0.049023 * M_TEAM_BASERUN_SB - 0.472894 * T_STD_IMP_TEAM_BASERUN_SB + 0.211623 * M_TEAM_BASERUN_CS + 0.341493 * T99_IMP_TEAM_BASERUN_CS + 0.145597 * T95_TEAM_FIELDING_E + 0.136517 * M_TEAM_FIELDING_DP - 0.022641 * IMP_TEAM_FIELDING_DP + 0.053889 * T99_TEAM_PITCHING_BB - 0.217785 * T95_TEAM_PITCHING_H + 0.129911 * T99_TEAM_PITCHING_HR - 0.085606 * T99_TEAM_PITCHING_SO + 0.010109 * M_TEAM_PITCHING_SO
PRIN14	-0.194634 * TEAM_FIELDING_E + 0.424796 * T95_TEAM_BATTING_H - 0.224078 * TRIM_TEAM_BATTING_2B - 0.20138 * T99_TEAM_BATTING_3B - 0.022638 * QUANT_TEAM_BATTING_HR - 0.101344 * COMB_HR + 0.138401 * STD_TEAM_BATTING_BB + 0.026307 * M_TEAM_BATTING_SO + 0.191777 * IMP_TEAM_BATTING_SO + 0.077696 * M_TEAM_BASERUN_SB - 0.499674 * T_STD_IMP_TEAM_BASERUN_SB + 0.157932 * M_TEAM_BASERUN_CS + 0.255254 * T99_IMP_TEAM_BASERUN_CS - 0.050857 * T95_TEAM_FIELDING_E + 0.472595 * M_TEAM_FIELDING_DP - 0.143834 * IMP_TEAM_FIELDING_DP - 0.043047 * T99_TEAM_PITCHING_BB - 0.014042 * T95_TEAM_PITCHING_H - 0.103165 * T99_TEAM_PITCHING_HR + 0.132964 * T99_TEAM_PITCHING_SO + 0.026307 * M_TEAM_PITCHING_SO
PRIN15	0.519649 * TEAM_FIELDING_E + 0.34088 * T95_TEAM_BATTING_H - 0.097517 * TRIM_TEAM_BATTING_2B - 0.046106 * T99_TEAM_BATTING_3B + 0.005023 * QUANT_TEAM_BATTING_HR + 0.051805 * COMB_HR + 0.245425 * STD_TEAM_BATTING_BB - 0.019752 * M_TEAM_BATTING_SO + 0.232376 * IMP_TEAM_BATTING_SO - 0.096406 * M_TEAM_BASERUN_SB - 0.057333 * T_STD_IMP_TEAM_BASERUN_SB + 0.090152 * M_TEAM_BASERUN_CS + 0.035308 * T99_IMP_TEAM_BASERUN_CS + 0.373954 * T95_TEAM_FIELDING_E - 0.399854 * M_TEAM_FIELDING_DP + 0.093014 * IMP_TEAM_FIELDING_DP - 0.078801 * T99_TEAM_PITCHING_BB - 0.354961 * T95_TEAM_PITCHING_H - 0.136478 * T99_TEAM_PITCHING_HR + 0.056965 * T99_TEAM_PITCHING_SO - 0.019752 * M_TEAM_PITCHING_SO
PRIN16	0.097843 * TEAM_FIELDING_E - 0.036501 * T95_TEAM_BATTING_H + 0.020574 * TRIM_TEAM_BATTING_2B + 0.025516 * T99_TEAM_BATTING_3B + 0.775004 * QUANT_TEAM_BATTING_HR - 0.593258 * COMB_HR - 0.076178 * STD_TEAM_BATTING_BB + 0.018306 * M_TEAM_BATTING_SO - 0.015162 * IMP_TEAM_BATTING_SO - 0.027845 * M_TEAM_BASERUN_SB - 0.012448 * T_STD_IMP_TEAM_BASERUN_SB + 0.108577 * M_TEAM_BASERUN_CS + 0.058512 * T99_IMP_TEAM_BASERUN_CS - 0.008633 * T95_TEAM_FIELDING_E - 0.05583 * M_TEAM_FIELDING_DP - 0.009395 * IMP_TEAM_FIELDING_DP + 0.072881 * T99_TEAM_PITCHING_BB + 0.049448 * T95_TEAM_PITCHING_H - 0.017382 * T99_TEAM_PITCHING_HR + 0.03535 * T99_TEAM_PITCHING_SO + 0.018306 * M_TEAM_PITCHING_SO
PRIN17	-0.354186 * TEAM_FIELDING_E - 0.171549 * T95_TEAM_BATTING_H + 0.031093 * TRIM_TEAM_BATTING_2B - 0.087014 * T99_TEAM_BATTING_3B + 0.059341 * QUANT_TEAM_BATTING_HR - 0.023851 * COMB_HR + 0.37194 * STD_TEAM_BATTING_BB - 0.005176 * M_TEAM_BATTING_SO - 0.12374 * IMP_TEAM_BATTING_SO - 0.216788 * M_TEAM_BASERUN_SB - 0.07233 * T_STD_IMP_TEAM_BASERUN_SB - 0.056286 * M_TEAM_BASERUN_CS + 0.018026 * T99_IMP_TEAM_BASERUN_CS + 0.611161 * T95_TEAM_FIELDING_E - 0.073469 * M_TEAM_FIELDING_DP + 0.008885 * IMP_TEAM_FIELDING_DP - 0.331047 * T99_TEAM_PITCHING_BB + 0.360289 * T95_TEAM_PITCHING_H + 0.011991 * T99_TEAM_PITCHING_HR + 0.067306 * T99_TEAM_PITCHING_SO - 0.005176 * M_TEAM_PITCHING_SO
PRIN18	0.234441 * TEAM_FIELDING_E - 0.147396 * T95_TEAM_BATTING_H + 0.012213 * TRIM_TEAM_BATTING_2B + 0.053396 * T99_TEAM_BATTING_3B + 0.215201 * QUANT_TEAM_BATTING_HR + 0.334934 * COMB_HR + 0.225447 * STD_TEAM_BATTING_BB + 0.007894 * M_TEAM_BATTING_SO + 0.140032 * IMP_TEAM_BATTING_SO - 0.084684 * M_TEAM_BASERUN_SB - 0.026434 * T_STD_IMP_TEAM_BASERUN_SB + 0.018049 * M_TEAM_BASERUN_CS + 0.005243 * T99_IMP_TEAM_BASERUN_CS - 0.346538 * T95_TEAM_FIELDING_E + 0.037137 * M_TEAM_FIELDING_DP + 0.016814 * IMP_TEAM_FIELDING_DP - 0.173661 * T99_TEAM_PITCHING_BB + 0.386552 * T95_TEAM_PITCHING_H - 0.615545 * T99_TEAM_PITCHING_HR - 0.05301 * T99_TEAM_PITCHING_SO + 0.007894 * M_TEAM_PITCHING_SO
PRIN19	0.342051 * TEAM_FIELDING_E - 0.085808 * T95_TEAM_BATTING_H - 0.013905 * TRIM_TEAM_BATTING_2B + 0.009293 * T99_TEAM_BATTING_3B - 0.229784 * QUANT_TEAM_BATTING_HR - 0.310919 * COMB_HR + 0.397267 * STD_TEAM_BATTING_BB + 0.022437 * M_TEAM_BATTING_SO + 0.143233 * IMP_TEAM_BATTING_SO + 0.007388 * M_TEAM_BASERUN_SB + 0.006156 * T_STD_IMP_TEAM_BASERUN_SB + 0.064248 * M_TEAM_BASERUN_CS + 0.025536 * T99_IMP_TEAM_BASERUN_CS - 0.330661 * T95_TEAM_FIELDING_E + 0.007484 * M_TEAM_FIELDING_DP - 0.013174 * IMP_TEAM_FIELDING_DP - 0.3072 * T99_TEAM_PITCHING_BB + 0.249319 * T95_TEAM_PITCHING_H + 0.51809 * T99_TEAM_PITCHING_HR - 0.084591 * T99_TEAM_PITCHING_SO + 0.022437 * M_TEAM_PITCHING_SO
PRIN20	-0.119475 * TEAM_FIELDING_E - 0.057712 * T95_TEAM_BATTING_H - 0.021898 * TRIM_TEAM_BATTING_2B - 0.003102 * T99_TEAM_BATTING_3B - 0.034183 * QUANT_TEAM_BATTING_HR - 0.032358 * COMB_HR - 0.217235 * STD_TEAM_BATTING_BB + 0.080381 * M_TEAM_BATTING_SO + 0.690489 * IMP_TEAM_BATTING_SO - 0.032575 * M_TEAM_BASERUN_SB - 0.03996 *

	T_STD_IMP_TEAM_BASERUN_SB + 0.001471 * M_TEAM_BASERUN_CS + 0.007938 * T99_IMP_TEAM_BASERUN_CS + 0.196784 * T95_TEAM_FIELDING_E - 0.07056 * M_TEAM_FIELDING_DP + 0.010675 * IMP_TEAM_FIELDING_DP + 0.187591 * T99_TEAM_PITCHING_BB + 0.222065 * T95_TEAM_PITCHING_H + 0.043658 * T99_TEAM_PITCHING_HR - 0.556645 * T99_TEAM_PITCHING_SO + 0.080381 * M_TEAM_PITCHING_SO
PRIN21	0 * TEAM_FIELDING_E + 0 * T95_TEAM_BATTING_H + 0 * TRIM_TEAM_BATTING_2B + 0 * T99_TEAM_BATTING_3B + 0 * QUANT_TEAM_BATTING_HR + 0 * COMB_HR + 0 * STD_TEAM_BATTING_BB + 0.707107 * M_TEAM_BATTING_SO + 0 * IMP_TEAM_BATTING_SO + 0 * M_TEAM_BASERUN_SB + 0 * T_STD_IMP_TEAM_BASERUN_SB + 0 * M_TEAM_BASERUN_CS + 0 * T99_IMP_TEAM_BASERUN_CS + 0 * T95_TEAM_FIELDING_E + 0 * M_TEAM_FIELDING_DP + 0 * IMP_TEAM_FIELDING_DP + 0 * T99_TEAM_PITCHING_BB + 0 * T95_TEAM_PITCHING_H + 0 * T99_TEAM_PITCHING_HR + 0 * T99_TEAM_PITCHING_SO - 0.707107 * M_TEAM_PITCHING_SO

Table 35: Formulas for the Principal Components PRIN1 – PRIN21

Part 3: Build Models

The chosen model was found through a series of steps. In each step, the hope is to improve upon the predictive power of the last "best model". Below I will outline each step completed and the performance of the resulting model.

STEP 1: Use all variables purposed for the model. Use backward, forward and stepwise selection.

Logic - In Step 1, all final variable transformations are included. This means that if a variable was standardized, the original variable was not included. If the standardized variable was then later trimmed, the original standardized variable was not included etc. The only potential redundancy allowed was to include both the imputed variables and the missing flag variables. In the case that a base variable has 2 imputed values to deal with the missing value - one using tree logic and one using average values, the one using tree logic is selected. On this set of variables forward, backward and stepwise variable selection was performed.

Analysis – The linear model properties for all three models are shown in Figure 24, 25 and 26 below. All models have a few variables that are included in a counter intuitive fashion. All counter intuitive inclusions are highlighted in red below. For example, in Figure 24, IMP_TEAM_FIELDING_DP is included as a negative coefficient, when it should positively contribute to the team wins. COMB_HR is not expected to have any sign coefficient as it is a combination of 2 variables – one that should positively contribute and one that should negatively contribute. Finally, note that the missing values are not expected to follow the sign coefficient of their base variable as it is unclear why a variable may be missing.

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	63.46913	6.78571	13672	87.49	<.0001
M_TEAM_FIELDING_DP	6.02276	1.62844	2137.70065	13.68	0.0002
IMP_TEAM_FIELDING_DP	-0.11007	0.01423	9346.86094	59.81	<.0001
T99_IMP_TEAM_BASERUN_CS	-0.04544	0.01904	890.12613	5.70	0.0171
M_TEAM_BASERUN_SB	35.37373	2.32372	36215	231.74	<.0001
T_STD_IMP_TEAM_BASERUN_SB	6.44412	0.53733	22477	143.83	<.0001
T99_TEAM_BATTING_3B	0.12660	0.01718	8482.39002	54.28	<.0001
STD_TEAM_BATTING_BB	4.82783	0.97796	3808.52058	24.37	<.0001
T95_TEAM_BATTING_H	0.04156	0.00481	11665	74.64	<.0001
TEAM_BATTING_HR	-0.06501	0.04061	400.60881	2.56	0.1095
COMB_HR	-0.00011799	0.00008144	328.08501	2.10	0.1475
M_TEAM_BATTING_SO	4.18990	1.57646	1103.92388	7.06	0.0079
IMP_TEAM_BATTING_SO	-0.01662	0.00217	9187.35924	58.79	<.0001
T95_TEAM_FIELDING_E	-0.07048	0.00510	29902	191.34	<.0001
T99_TEAM_PITCHING_BB	-0.01386	0.00700	612.81789	3.92	0.0478
T95_TEAM_PITCHING_H	-0.00529	0.00324	416.47846	2.66	0.1027
T99_TEAM_PITCHING_HR	0.16967	0.02999	5000.30664	32.00	<.0001

Figure 24: Step 1 Forward Selection Linear Model Properties

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	64.99319	6.70532	14689	93.95	<.0001
M_TEAM_FIELDING_DP	6.17738	1.62533	2258.56560	14.45	0.0001
IMP_TEAM_FIELDING_DP	-0.10734	0.01411	9047.57069	57.87	<.0001
T99_IMP_TEAM_BASERUN_CS	-0.04809	0.01896	1006.01528	6.43	0.0113
M_TEAM_BASERUN_SB	35.63932	2.31704	36991	236.59	<.0001
T_STD_IMP_TEAM_BASERUN_SB	6.41899	0.53718	22326	142.79	<.0001
T99_TEAM_BATTING_3B	0.12567	0.01718	8369.60523	53.53	<.0001
STD_TEAM_BATTING_BB	4.95772	0.97408	4050.24329	25.90	<.0001
T95_TEAM_BATTING_H	0.04268	0.00475	12623	80.74	<.0001
TEAM_BATTING_HR	-0.09893	0.03319	1388.99461	8.88	0.0029
M_TEAM_BATTING_SO	3.63024	1.52878	881.63914	5.64	0.0177
IMP_TEAM_BATTING_SO	-0.01636	0.00216	8968.30960	57.36	<.0001
T95_TEAM_FIELDING_E	-0.07130	0.00507	30977	198.12	<.0001
T99_TEAM_PITCHING_BB	-0.01488	0.00697	713.69599	4.56	0.0327
T95_TEAM_PITCHING_H	-0.00615	0.00319	583.38685	3.73	0.0535
T99_TEAM_PITCHING_HR	0.17252	0.02994	5191.92927	33.21	<.0001

Figure 25: Step 1 Backward Selection Linear Model Properties

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	64.00770	6.27273	16286	104.12	<.0001
M_TEAM_FIELDING_DP	5.93360	1.62374	2088.66290	13.35	0.0003
IMP_TEAM_FIELDING_DP	-0.11085	0.01421	9514.77493	60.83	<.0001
T99_IMP_TEAM_BASERUN_CS	-0.03952	0.01880	691.35113	4.42	0.0356
M_TEAM_BASERUN_SB	33.18150	1.94210	45658	291.91	<.0001
T_STD_IMP_TEAM_BASERUN_SB	6.34414	0.53293	22165	141.71	<.0001
T99_TEAM_BATTING_3B	0.13144	0.01672	9661.49519	61.77	<.0001
STD_TEAM_BATTING_BB	5.04672	0.66163	9100.22439	58.18	<.0001
T95_TEAM_BATTING_H	0.03522	0.00328	18083	115.61	<.0001
COMB_HR	-0.00019204	0.00006657	1301.69685	8.32	0.0040
M_TEAM_BATTING_SO	4.87575	1.53842	1571.07973	10.04	0.0015
IMP_TEAM_BATTING_SO	-0.01679	0.00210	9953.84115	63.64	<.0001
T95_TEAM_FIELDING_E	-0.07104	0.00505	30998	198.18	<.0001
T99_TEAM_PITCHING_BB	-0.01569	0.00451	1894.36010	12.11	0.0005
T99_TEAM_PITCHING_HR	0.12851	0.01724	8695.91138	55.60	<.0001

Figure 26: Step 1 Stepwise Selection Linear Model Properties

In Table 36 below we can see the model metrics for the three models in Step 1. Based on the metrics Stepwise (highlighted in yellow) is selected as the current chosen model as it has the least amount of variables and very close Adjusted R Square and AIC, SBC values.

Model	# Variables	R Squared	Adjusted R Squared	CP	AIC	SBC
Forward_TreeMissing	16	0.37461	0.37018	13.3909	11514.46	11611.87
Backward_TreeMissing	15	0.37403	0.36987	13.4869	11514.57	11606.25
Stepwise_TreeMissing	14	0.37353	0.36965	13.2915	11514.39	11600.34

Table 36: Step 1 Backward, Forward and Stepwise Model Metrics

STEP 2: Try a model with only the variables with a Pearson Correlation Value for TARGET_WINS above 0.1 or below -0.1

Logic - In Step 2, only base variables that have a correlation higher than 0.1 or lower than -0.1 are included. No variable selection is performed, all variables are included in the variable set.

Analysis – The linear model properties are shown in Figure 27. The model has a few variables that are included in a counter intuitive fashion (highlighted in red). For example, TEAM_BATTING_2B is included as a negative coefficient, when it should positively contribute to the team wins.

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	2.93306	3.44988	0.85	0.3953
TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)	1	0.04674	0.00330	14.15	<.0001
TEAM_BATTING_2B	Doubles by batters (2B)	1	-0.05677	0.00890	-6.38	<.0001
TEAM_BATTING_3B	Triples by batters (3B)	1	0.10193	0.01723	5.92	<.0001
TEAM_BATTING_HR	Homeruns by batters (4B)	1	0.00001689	0.02774	0.00	0.9995
TEAM_BATTING_BB	Walks by batters	1	0.03665	0.00556	6.59	<.0001
TEAM_BASERUN_SB	Stolen bases	1	0.05727	0.00404	14.19	<.0001
TEAM_PITCHING_H	Hits allowed	1	0.00287	0.00041219	6.96	<.0001
TEAM_PITCHING_HR	Homeruns allowed	1	0.03820	0.02568	1.49	0.1370
TEAM_PITCHING_BB	Walks allowed	1	-0.01175	0.00403	-2.91	0.0036
TEAM_FIELDING_E	Errors	1	-0.05191	0.00324	-16.02	<.0001

Figure 27: Step 2 Correlated Base Variables Linear Model Properties

In Table 37 below we can see the model metrics for the Step 2 model. Based on the metrics the previous stepwise model from Step 1 above remains the current chosen model.

Model	# Variables	R Squared	Adjusted R Squared	CP	AIC	SBC
CorrelationAbove1	10	0.34383	0.34075	11	10667.14	10729.52

Table 37: Step 2 – Correlated Base Variables Model Metrics

STEP 3: Use a Decision Tree for Initial Variable Selection

Logic - In Step 3, a decision tree was created to predict TARGET_WINS using all variables available. The decision tree in Figure 28 was used as a method of initial variable selection. A regression model was then run with the whole variable set and also the variable set was then further reduced through forward, backward and stepwise regression.
**Please note that a larger image of the decision tree can be made available upon request.*

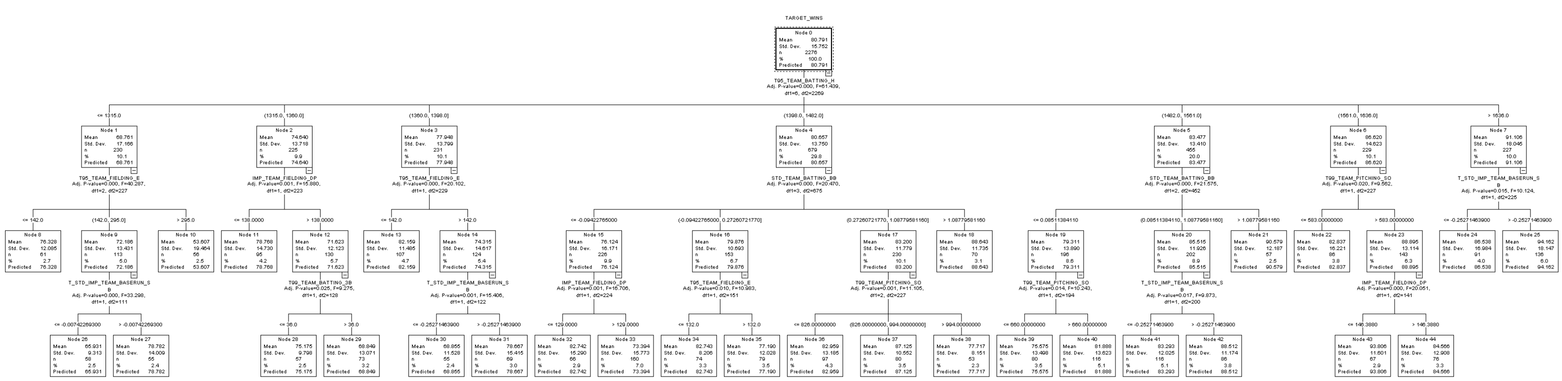


Figure 28: Decision Tree Used for Initial Variable Selection

Analysis – The linear model properties are shown in Figure 29-32. The models each have each kept in the full variable set and produced the same linear model. As such each model has the same one variable that has been included in a counter intuitive fashion (highlighted in red). IMP_TEAM_FIELDING_DP is included as a negative coefficient, when it should positively contribute to the team wins.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	9.33494	4.66069	2.00	0.0453
T95_TEAM_BATTING_H		1	0.05703	0.00292	19.55	<.0001
T_STD_IMP_TEAM_BASERUN_SB		1	2.79543	0.42685	6.55	<.0001
T99_TEAM_PITCHING_SO		1	0.00469	0.00147	3.20	0.0014
STD_TEAM_BATTING_BB		1	2.47916	0.39509	6.27	<.0001
T95_TEAM_FIELDING_E		1	-0.02717	0.00290	-9.36	<.0001
IMP_TEAM_FIELDING_DP		1	-0.09205	0.01342	-6.86	<.0001
T99_TEAM_BATTING_3B		1	0.07320	0.01698	4.31	<.0001

Figure 29: Step 3 Forward Selection Linear Model Properties

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	9.33494	4.66069	2.00	0.0453
T95_TEAM_BATTING_H		1	0.05703	0.00292	19.55	<.0001
T_STD_IMP_TEAM_BASERUN_SB		1	2.79543	0.42685	6.55	<.0001
T99_TEAM_PITCHING_SO		1	0.00469	0.00147	3.20	0.0014
STD_TEAM_BATTING_BB		1	2.47916	0.39509	6.27	<.0001
T95_TEAM_FIELDING_E		1	-0.02717	0.00290	-9.36	<.0001
IMP_TEAM_FIELDING_DP		1	-0.09205	0.01342	-6.86	<.0001
T99_TEAM_BATTING_3B		1	0.07320	0.01698	4.31	<.0001

Figure 30: Step 3 Backward Selection Linear Model Properties

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	9.33494	4.66069	2.00	0.0453
T95_TEAM_BATTING_H		1	0.05703	0.00292	19.55	<.0001
T_STD_IMP_TEAM_BASERUN_SB		1	2.79543	0.42685	6.55	<.0001
T99_TEAM_PITCHING_SO		1	0.00469	0.00147	3.20	0.0014
STD_TEAM_BATTING_BB		1	2.47916	0.39509	6.27	<.0001
T95_TEAM_FIELDING_E		1	-0.02717	0.00290	-9.36	<.0001
IMP_TEAM_FIELDING_DP		1	-0.09205	0.01342	-6.86	<.0001
T99_TEAM_BATTING_3B		1	0.07320	0.01698	4.31	<.0001

Figure 31: Step 3 Stepwise Selection Linear Model Properties

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	9.33494	4.66069	2.00	0.0453
T95_TEAM_BATTING_H		1	0.05703	0.00292	19.55	<.0001
T_STD_IMP_TEAM_BASERUN_SB		1	2.79543	0.42685	6.55	<.0001
T99_TEAM_PITCHING_SO		1	0.00469	0.00147	3.20	0.0014
STD_TEAM_BATTING_BB		1	2.47916	0.39509	6.27	<.0001
T95_TEAM_FIELDING_E		1	-0.02717	0.00290	-9.36	<.0001
IMP_TEAM_FIELDING_DP		1	-0.09205	0.01342	-6.86	<.0001
T99_TEAM_BATTING_3B		1	0.07320	0.01698	4.31	<.0001

Figure 32: Step 3 All Variables Linear Model Properties

In Table 38 below we can see the model metrics for the Step 3 models. Based on these metrics and particularly the low Adjusted R Squared value, the previous stepwise model from Step 1 above remains the current chosen model.

Model	# Variables	R Squared	Adjusted R Squared	CP	AIC	SBC
TVars1Forward	7	0.25672	0.25443	8	11889.5	11935.34
TVars2Backward	7	0.25672	0.25443	8	11889.5	11935.34
TVars3Stepwise	7	0.25672	0.25443	8	11889.5	11935.34
TVars4All	7	0.25672	0.25443	8	11889.5	11935.34

Table 38: Step 3 – Tree Variable Selection Model Metrics

STEP 4: Use Principal Component Analysis Variables Only

Logic - In Step 4, the principal component analysis variables were introduced. Initially a model was run with only the core 9 PCA variables identified in Section 2 (PRIN 1-PRIN9). Next all PCA variables were run in a model. Finally all PCA variables were included in the variable set and further variable reduction was performed through forward, backward and stepwise selection.

Analysis – The linear model properties are shown in Figure 33-37. Note that there are no expectations for the PCA variables to have a positive or negative coefficient as they are simply a combination of all variables. Therefore, the model with the most desirable metrics will be selected.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	33.12318	4.85561	6.82	<.0001
PRIN1		1	0.06373	0.01045	6.10	<.0001
PRIN2		1	-0.01268	0.02342	-0.54	0.5882
PRIN3		1	0.05150	0.00785	6.56	<.0001
PRIN4		1	0.06147	0.02185	2.81	0.0049
PRIN5		1	0.01833	0.02299	0.80	0.4255
PRIN6		1	0.06875	0.01649	4.17	<.0001
PRIN7		1	-0.11675	0.01887	-6.19	<.0001
PRIN8		1	-0.10772	0.01551	-6.95	<.0001
PRIN9		1	0.03089	0.01186	2.60	0.0093

Figure 33: Step 4 Core PCA Variables Linear Model Properties

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	59.10840	7.64550	7.73	<.0001
PRIN1		1	8.49781	0.77800	10.92	<.0001
PRIN2		1	5.95462	1.15773	5.14	<.0001
PRIN3		1	-12.03779	1.07670	-11.18	<.0001
PRIN4		1	5.93161	0.76143	7.79	<.0001
PRIN5		1	-0.55343	1.05829	-0.52	0.6011
PRIN6		1	-1.17112	0.56823	-2.06	0.0394
PRIN7		1	-3.85109	0.88688	-4.34	<.0001
PRIN8		1	7.52951	0.82968	9.08	<.0001
PRIN9		1	1.92629	0.43549	4.42	<.0001
PRIN10		1	8.88585	0.80171	11.08	<.0001
PRIN11		1	10.26393	0.91693	11.19	<.0001
PRIN12		1	23.83229	1.49148	15.98	<.0001
PRIN13		1	-0.52116	0.48896	-1.07	0.2866
PRIN14		1	2.86756	0.99183	2.89	0.0039
PRIN15		1	-5.80558	0.79914	-7.26	<.0001
PRIN16		1	-1.84546	0.64497	-2.86	0.0043
PRIN17		1	-7.36600	0.72137	-10.21	<.0001
PRIN18		1	-2.37874	0.37846	-6.29	<.0001
PRIN19		1	1.49184	0.48677	3.06	0.0022
PRIN20		1	-1.83299	0.37794	-4.85	<.0001
PRIN21		0	0	.	.	.

Figure 34: Step 4 All PCA Variables Linear Model Properties

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	59.10840	7.64550	9268.27421	59.77	<.0001
PRIN1	8.49781	0.77800	18500	119.30	<.0001
PRIN2	5.95462	1.15773	4102.08593	26.45	<.0001
PRIN3	-12.03779	1.07670	19383	125.00	<.0001
PRIN4	5.93161	0.76143	9410.19395	60.69	<.0001
PRIN5	-0.55343	1.05829	42.40572	0.27	0.6011
PRIN6	-1.17112	0.56823	658.67920	4.25	0.0394
PRIN7	-3.85109	0.88688	2923.79349	18.86	<.0001
PRIN8	7.52951	0.82968	12771	82.36	<.0001
PRIN9	1.92629	0.43549	3033.91595	19.57	<.0001
PRIN10	8.88585	0.80171	19049	122.85	<.0001
PRIN11	10.26393	0.91693	19430	125.30	<.0001
PRIN12	23.83229	1.49148	39592	255.33	<.0001
PRIN13	-0.52116	0.48896	176.15922	1.14	0.2866
PRIN14	2.86756	0.99183	1296.17857	8.36	0.0039
PRIN15	-5.80558	0.79914	8183.83846	52.78	<.0001
PRIN16	-1.84546	0.64497	1269.53840	8.19	0.0043
PRIN17	-7.36600	0.72137	16168	104.27	<.0001
PRIN18	-2.37874	0.37846	6125.86731	39.51	<.0001
PRIN19	1.49184	0.48677	1456.48487	9.39	0.0022
PRIN20	-1.83299	0.37794	3647.40800	23.52	<.0001

Figure 35: Step 4 All PCA Variables Forward Selection Linear Model Properties

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	55.77720	6.51250	11372	73.35	<.0001
PRIN1	8.96937	0.67149	27661	178.42	<.0001
PRIN2	5.66253	0.53662	17262	111.35	<.0001
PRIN3	-12.44231	1.02148	23002	148.37	<.0001
PRIN4	6.42488	0.58588	18644	120.26	<.0001
PRIN6	-1.72098	0.34985	3751.57252	24.20	<.0001
PRIN7	-3.51146	0.83860	2718.19576	17.53	<.0001
PRIN8	7.56200	0.69969	18108	116.80	<.0001
PRIN9	2.13827	0.32766	6602.29644	42.59	<.0001
PRIN10	8.96935	0.78115	20439	131.84	<.0001
PRIN11	10.49111	0.89789	21165	136.52	<.0001
PRIN12	24.01419	1.44663	42721	275.56	<.0001
PRIN14	3.62804	0.42030	11552	74.51	<.0001
PRIN15	-6.13436	0.66721	13105	84.53	<.0001
PRIN16	-1.72831	0.63006	1166.53351	7.52	0.0061
PRIN17	-7.59334	0.69696	18402	118.70	<.0001
PRIN18	-2.37869	0.36704	6511.21086	42.00	<.0001
PRIN19	1.48962	0.47200	1544.12577	9.96	0.0016
PRIN20	-1.72189	0.19373	12246	78.99	<.0001

Figure 36: Step 4 All PCA Variables Backward Selection Linear Model Properties

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	36.29485	5.89122	6820.49684	37.96	<.0001
PRIN3	0.08480	0.00473	57738	321.31	<.0001
PRIN4	0.17450	0.01382	28655	159.47	<.0001
PRIN6	0.05638	0.01213	3879.66885	21.59	<.0001
PRIN8	-0.24391	0.02142	23304	129.69	<.0001
PRIN14	0.00060042	0.01374	0.34321	0.00	0.9651
PRIN15	-0.03207	0.00524	6721.80320	37.41	<.0001
PRIN16	0.16933	0.01617	19698	109.62	<.0001
PRIN17	0.05798	0.00464	27997	155.80	<.0001

Figure 37: Step 4 All PCA Variables Stepwise Selection Linear Model Properties

In Table 39 below we can see the model metrics for the Step 4 models. Based on these metrics and particularly the high Adjusted R Squared value and low AIC, SBC values, Backward PCA Only model is now the highest performing model and our current chosen model. The model is highlighted in yellow below.

Model	# Variables	R Squared	Adjusted R Squared	CP	AIC	SBC
CORE_PCA	9	0.25863	0.25569	10	11887.65	11944.95
All_PCA	20	0.38056	0.37507	21	11500.68	11621.02
Forward_PCAOnly	20	0.38056	0.37507	21	11500.68	11621.02
Backward_PCAOnly	18	0.38015	0.37521	18.508	11498.2	11607.08
Stepwise_PCAOnly	9	0.28076	0.2779	362.333	11818.69	11875.99

Table 39: Step 4 - PCA Variables Only

STEP 5: Use Principal Component Analysis Variables and Variables Purposed for the Model (From Step 1)

Logic - In Step 5, we include all principal component analysis variables as well as all variables purposed for the model (all variables from Step 1). Further selection was performed on this variable set through forward, backward and stepwise selection.

Analysis – The linear model properties are shown in Figure 38-40. Both the forward and stepwise models have a few variables included a counter intuitive fashion (highlighted in red). For example: IMP_TEAM_FIELDING_DP is included as a negative coefficient, when it should positively contribute to the team wins.
Note that there are no expectations for the PCA variables to have a positive or negative coefficient. Also note that the missing values are not expected to follow the sign coefficient of their base variable as it is unclear why a variable may be missing

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	60.41520	6.51428	13323	86.01	<.0001
PRIN12	0.02042	0.00498	2601.12081	16.79	<.0001
PRIN16	0.01550	0.00601	1028.98403	6.64	0.0100
PRIN17	0.02728	0.00787	1861.48310	12.02	0.0005
PRIN19	-0.03183	0.01135	1218.13661	7.86	0.0051
M_TEAM_FIELDING_DP	6.48404	1.63775	2427.85486	15.67	<.0001
IMP_TEAM_FIELDING_DP	-0.11061	0.01432	9238.19430	59.64	<.0001
M_TEAM_BASERUN_CS	-0.82684	0.96643	113.38001	0.73	0.3923
T99_IMP_TEAM_BASERUN_CS	-0.04407	0.01984	764.22609	4.93	0.0264
M_TEAM_BASERUN_SB	33.44616	2.26558	33757	217.94	<.0001
T_STD_IMP_TEAM_BASERUN_SB	6.15930	0.55428	19126	123.48	<.0001
T99_TEAM_BATTING_3B	0.11598	0.01718	7062.27521	45.60	<.0001
STD_TEAM_BATTING_BB	2.55390	0.81352	1526.52218	9.86	0.0017
T95_TEAM_BATTING_H	0.04032	0.00379	17529	113.17	<.0001
M_TEAM_BATTING_SO	4.95815	1.69222	1329.69475	8.58	0.0034
IMP_TEAM_BATTING_SO	-0.01046	0.00249	2741.05365	17.70	<.0001
T95_TEAM_FIELDING_E	-0.07453	0.00532	30387	196.19	<.0001
T99_TEAM_PITCHING_HR	0.14521	0.01804	10036	64.80	<.0001

Figure 38: Step 5 All PCA Variables and All Variables Purposed for the Model - Forward Selection Linear Model Properties

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	60.86678	6.91367	12011	77.51	<.0001
PRIN1	-19.09464	1.31862	32494	209.69	<.0001
PRIN2	50.92837	3.39846	34800	224.57	<.0001
PRIN3	-22.65612	2.20523	16356	105.55	<.0001
PRIN4	-68.58101	5.52923	23840	153.84	<.0001
PRIN5	-9.51040	2.29667	2657.17887	17.15	<.0001
PRIN6	-15.46401	2.87393	4486.55533	28.95	<.0001
PRIN7	-42.59652	3.22307	27066	174.67	<.0001
PRIN9	-3.94413	0.32794	22415	144.65	<.0001
PRIN10	17.48837	1.43427	23039	148.67	<.0001
PRIN12	-9.34750	2.70860	1845.53910	11.91	0.0006
PRIN13	52.65548	5.88669	12398	80.01	<.0001
PRIN14	56.20075	6.19157	12767	82.39	<.0001
PRIN15	4.31043	0.44086	14814	95.60	<.0001
PRIN16	-0.69937	0.10866	6419.32520	41.43	<.0001
PRIN20	-6.37006	0.57843	18794	121.28	<.0001
M_TEAM_BASERUN_CS	-19.74616	3.68129	4458.49902	28.77	<.0001
T_STD_IMP_TEAM_BASERUN_SB	105.92795	10.74929	15048	97.11	<.0001
M_TEAM_BATTING_SO	107.95378	6.96777	37197	240.04	<.0001

Figure 39: Step 5 All PCA Variables and All Variables Purposed for the Model - Backward Selection Linear Model Properties

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	59.01536	6.30509	13568	87.61	<.0001
PRIN12	0.02020	0.00498	2552.47534	16.48	<.0001
PRIN16	0.01463	0.00593	943.67950	6.09	0.0136
PRIN17	0.02621	0.00777	1762.93639	11.38	0.0008
PRIN19	-0.03005	0.01116	1123.59768	7.25	0.0071
M_TEAM_FIELDING_DP	6.42889	1.63639	2390.43053	15.43	<.0001
IMP_TEAM_FIELDING_DP	-0.10875	0.01416	9140.39516	59.02	<.0001
T99_IMP_TEAM_BASERUN_CS	-0.03859	0.01878	654.14332	4.22	0.0400
M_TEAM_BASERUN_SB	33.48979	2.26487	33862	218.64	<.0001
T_STD_IMP_TEAM_BASERUN_SB	6.05109	0.53962	19474	125.74	<.0001
T99_TEAM_BATTING_3B	0.11443	0.01708	6952.25779	44.89	<.0001
STD_TEAM_BATTING_BB	2.60620	0.81117	1598.71031	10.32	0.0013
T95_TEAM_BATTING_H	0.04040	0.00379	17604	113.67	<.0001
M_TEAM_BATTING_SO	4.90464	1.69096	1302.93136	8.41	0.0038
IMP_TEAM_BATTING_SO	-0.01016	0.00246	2639.62029	17.04	<.0001
T95_TEAM_FIELDING_E	-0.07464	0.00532	30487	196.85	<.0001
T99_TEAM_PITCHING_HR	0.14703	0.01791	10436	67.38	<.0001

Figure 40: Step 5 All PCA Variables and All Variables Purposed for the Model - Stepwise Selection Linear Model Properties

In Table 40 below we can see the model metrics for the Step 5 models. Based on these metrics and particularly the slightly higher Adjusted R Squared value and lower number of parameters, AIC and SBC values, the model "Forward_PCAIncl_AllTreeMissing" is now the highest performing model and our current chosen model. The model is highlighted in yellow below.

Model	# Variables	R Squared	Adjusted R Squared	CP	AIC	SBC
Forward_PCAIncl_AllTreeMissing	17	0.38043	0.37577	13.4828	11495.17	11598.31
Backward_PCAIncl_AllTreeMissing	18	0.38043	0.37549	15.4955	11497.18	11606.06
Stepwise_PCAIncl_AllTreeMissing	16	0.38023	0.37584	12.2133	11493.91	11591.32

Table 40: Step 5 - All PCA Variables and All Variables Purposed for the Model

STEP 6: Use Principal Component Analysis Variables and Variables Purposed for the Model (From Step 1), Use Averages for Missing Values Instead of Tree Logic

Logic – In Step 6, we examine replacing missing data with the average values vs the decision tree logic that has been used in models up to this point. The same set of variables from Step 5 is used, except those missing values imputed with tree logic are now replaced by missing values imputed with the average. Further selection was then performed on this variable set through forward, backward and stepwise selection.

Analysis – The linear model properties are shown in Figure 41-43. All three models have a few variables included in a counter intuitive fashion (highlighted in red). For example: IMP_TEAM_FIELDING_DP is included as a negative coefficient, when it should positively contribute to the team wins.
Note that there are no expectations for the PCA variables to have a positive or negative coefficient. Also note that the missing values are not expected to follow the sign coefficient of their base variable as it is unclear why a variable may be missing

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	58.51388	7.60833	9142.93746	59.15	<.0001
PRIN12	-0.26648	0.08951	1370.04389	8.86	0.0029
PRIN17	0.11648	0.02721	2833.01949	18.33	<.0001
PRIN20	0.00676	0.01139	54.36225	0.35	0.5532
M_TEAM_FIELDING_DP	6.46522	1.66422	2332.86590	15.09	0.0001
IMP_TEAM_FIELDING_DP	-0.14579	0.01833	9779.05922	63.26	<.0001
M_TEAM_BASERUN_CS	-0.92853	0.95193	147.07101	0.95	0.3295
T99_IMP2_TEAM_BASERUN_CS	-0.06870	0.01928	1962.84413	12.70	0.0004
M_TEAM_BASERUN_SB	37.91603	2.15051	48052	310.86	<.0001
T_STD_IMP2_TEAM_BASERUN_SB	6.00061	0.49584	22639	146.45	<.0001
TRIM_TEAM_BATTING_2B	0.02367	0.01117	694.03582	4.49	0.0342
T99_TEAM_BATTING_3B	0.13113	0.01736	8816.36976	57.04	<.0001
STD_TEAM_BATTING_BB	2.28448	1.00554	797.85102	5.16	0.0232
T95_TEAM_BATTING_H	0.05600	0.00731	9068.29732	58.67	<.0001
M_TEAM_BATTING_SO	5.36471	2.08007	1028.21190	6.65	0.0100
T95_TEAM_FIELDING_E	-0.09775	0.00881	19019	123.04	<.0001
T95_TEAM_PITCHING_H	-0.12898	0.04013	1596.95759	10.33	0.0013
T99_TEAM_PITCHING_HR	0.10600	0.01990	4383.99774	28.36	<.0001
T99_TEAM_PITCHING_SO	-0.09423	0.02705	1875.99620	12.14	0.0005

Figure 41: Step 6 All PCA Variables and All Variables Purposed for the Model - Missing Values Replaced using Average – Forward Selection Linear Model Properties

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	57.88488	7.66905	8810.79085	56.97	<.0001
PRIN1	19.70189	2.88315	7221.85264	46.70	<.0001
PRIN3	-47.45657	3.42263	29733	192.25	<.0001
PRIN4	-3.57580	1.50841	869.11338	5.62	0.0178
PRIN5	19.70865	1.72330	20228	130.80	<.0001
PRIN6	11.42328	0.82968	29317	189.56	<.0001
PRIN7	23.97777	1.71946	30075	194.46	<.0001
PRIN8	30.16730	3.83165	9586.71638	61.99	<.0001
PRIN9	-16.89661	2.92274	5168.77978	33.42	<.0001
PRIN11	-11.19415	2.87231	2349.02372	15.19	0.0001
PRIN12	15.28021	2.44474	6041.71884	39.07	<.0001
PRIN15	-31.15398	2.22681	30271	195.73	<.0001
PRIN17	-33.98709	2.51918	28150	182.02	<.0001
PRIN18	-18.92105	1.35283	30253	195.62	<.0001
PRIN19	-28.39928	2.03145	30225	195.43	<.0001
PRIN20	15.96184	1.13074	30819	199.27	<.0001
M_TEAM_FIELDING_DP	-9.47079	2.34501	2522.60727	16.31	<.0001
M_TEAM_BASERUN_CS	33.80786	6.12372	4713.81868	30.48	<.0001
T99_IMP2_TEAM_BASERUN_CS	-0.11297	0.05690	609.74921	3.94	0.0472
T_STD_IMP2_TEAM_BASERUN_SB	11.91928	2.71177	2987.87078	19.32	<.0001
STD_TEAM_BATTING_BB	71.88986	4.99810	31996	206.88	<.0001
M_TEAM_BATTING_SO	-34.49172	3.94445	11826	76.46	<.0001

Figure 42: Step 6 All PCA Variables and All Variables Purposed for the Model - Missing Values Replaced using Average – Backward Selection Linear Model Properties

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	59.68365	5.92422	15684	101.50	<.0001
PRIN12	-0.20882	0.04499	3328.82382	21.54	<.0001
PRIN17	0.10036	0.01927	4189.75469	27.11	<.0001
M_TEAM_FIELDING_DP	6.51025	1.64637	2416.22922	15.64	<.0001
IMP_TEAM_FIELDING_DP	-0.13660	0.01519	12495	80.86	<.0001
T99_IMP2_TEAM_BASERUN_CS	-0.06000	0.01757	1801.26679	11.66	0.0007
M_TEAM_BASERUN_SB	37.40465	2.06938	50486	326.72	<.0001
T_STD_IMP2_TEAM_BASERUN_SB	5.95361	0.48368	23412	151.51	<.0001
TRIM_TEAM_BATTING_2B	0.01934	0.00969	615.42921	3.98	0.0461
T99_TEAM_BATTING_3B	0.12644	0.01668	8883.75797	57.49	<.0001
STD_TEAM_BATTING_BB	2.85426	0.41613	7269.75184	47.05	<.0001
T95_TEAM_BATTING_H	0.05286	0.00599	12046	77.95	<.0001
M_TEAM_BATTING_SO	4.74191	1.79087	1083.37307	7.01	0.0082
T95_TEAM_FIELDING_E	-0.09379	0.00713	26739	173.04	<.0001
T95_TEAM_PITCHING_H	-0.10348	0.02176	3494.43836	22.61	<.0001
T99_TEAM_PITCHING_HR	0.11394	0.01809	6128.03520	39.66	<.0001
T99_TEAM_PITCHING_SO	-0.07657	0.01363	4879.21271	31.58	<.0001

Figure 43: Step 6 All PCA Variables and All Variables Purposed for the Model - Missing Values Replaced using Average – Stepwise Selection Linear Model Properties

In Table 41 below we can see the model metrics for the Step 6 models. Based on these metrics and particularly the slightly higher Adjusted R Squared value and lower number of parameters, AIC and SBC values, the model “Stepwise_AllVars_AvgMissing” is now the highest performing model and our current chosen model. The model is highlighted in yellow below.

Model	# Variables	R Squared	Adjusted R Squared	CP	AIC	SBC
Forward_AllVars_AvgMissing	18	0.38196	0.37703	14.6017	11491.54	11600.42
Backward_AllVars_AvgMissing	21	0.38247	0.37671	18.7544	11495.68	11621.74
Stepwise_AllVars_AvgMissing	16	0.38162	0.37724	11.8399	11488.79	11586.21

Table 41: Step 6 - All PCA Variables and All Variables Purposed for the Model, Missing Values Replaced using Average

STEP 7: Kitchen Sink

Logic – Since the Adjusted R Squared values seemed to hit a wall around 0.37, new logic was needed. In this step we brute force it and throw in "Everything but the Kitchen Sink". To further clarify, all variables that do not have missing values will be thrown in. This means that redundant variables will be included. For example: IMP_X, STD_X, TRIM_STD_X, IMP2_X etc are all included in the model. Further variable selection will then be performed on this variable set using forward, backward and stepwise selection.

Analysis – The linear model properties are shown in Figure 44-46. All three models have a large number of input variables, which is undesirable. All models also have a few variables included in a counter intuitive fashion (highlighted in red). For example: TEAM_BATTING_3B is included as a negative coefficient, when it should positively contribute to the team wins.
Note that there are no expectations for the PCA variables to have a positive or negative coefficient. Also note that the missing values are not expected to follow the sign coefficient of their base variable as it is unclear why a variable may be missing

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	40.18852	6.98090	4618.65322	33.14	<.0001
IMP_TEAM_BASERUN_CS	0.25029	0.09636	940.25059	6.75	0.0095
IMP_TEAM_BASERUN_SB	0.04726	0.01679	1104.49544	7.93	0.0049
TEAM_BATTING_2B	0.33804	0.13239	908.64576	6.52	0.0107
TEAM_BATTING_3B	-0.43091	0.08528	3557.62467	25.53	<.0001
TEAM_BATTING_BB	0.05667	0.00999	4486.51258	32.19	<.0001
TEAM_BATTING_H	0.07558	0.00547	26605	190.91	<.0001
TEAM_PITCHING_BB	0.00563	0.00327	412.33296	2.96	0.0855
TEAM_PITCHING_H	0.00168	0.00040851	2355.91085	16.91	<.0001
PRIN5	0.11354	0.07486	320.54968	2.30	0.1295
PRIN6	-0.12959	0.15777	94.03230	0.67	0.4115
PRIN12	0.17993	0.06229	1162.64062	8.34	0.0039
PRIN20	0.16507	0.05182	1413.89534	10.15	0.0015
M_TEAM_FIELDING_DP	6.75006	1.59288	2502.54341	17.96	<.0001
IMP_TEAM_FIELDING_DP	-0.10358	0.03624	1138.69956	8.17	0.0043
M_TEAM_BASERUN_CS	1.87136	0.94490	546.60225	3.92	0.0478
T99_IMP2_TEAM_BASERUN_CS	-0.18225	0.05516	1521.19153	10.92	0.0010
M_TEAM_BASERUN_SB	44.09159	2.69483	37306	267.70	<.0001
T_STD_IMP2_TEAM_BASERUN_SB	1.17085	1.70307	65.86707	0.47	0.4918
TRIM_TEAM_BATTING_2B	-0.38047	0.13444	1116.21580	8.01	0.0047
T99_TEAM_BATTING_3B	0.52861	0.09051	4753.78969	34.11	<.0001
T95_TEAM_BATTING_H	-0.04614	0.00833	4271.52200	30.65	<.0001
IMP2_TEAM_BATTING_SO	-0.09260	0.02074	2779.50681	19.95	<.0001
T95_TEAM_FIELDING_E	-0.09497	0.01628	4742.36827	34.03	<.0001
T99_TEAM_PITCHING_BB	-0.06698	0.04742	278.08085	2.00	0.1579
T99_TEAM_PITCHING_HR	0.10155	0.01816	4357.09505	31.27	<.0001
T99_TEAM_PITCHING_SO	0.19621	0.08053	827.21432	5.94	0.0149
M_TEAM_PITCHING_SO	23.87506	5.44812	2676.26503	19.20	<.0001

Figure 44: Step 7 Kitchen Sink, All Variables Without Missing Values – Forward Selection Linear Model Properties

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	66.91029	8.02596	9687.42153	69.50	<.0001
IMP_TEAM_BASERUN_CS	0.25138	0.09758	925.04197	6.64	0.0101
IMP_TEAM_BASERUN_SB	0.05323	0.01715	1342.61938	9.63	0.0019
TEAM_BATTING_2B	0.34271	0.13244	933.26206	6.70	0.0097
TEAM_BATTING_3B	-0.43407	0.08553	3589.80362	25.75	<.0001
TEAM_BATTING_H	0.07552	0.00549	26401	189.41	<.0001
TEAM_PITCHING_BB	0.00570	0.00328	422.19206	3.03	0.0819
TEAM_PITCHING_H	0.00166	0.00041228	2251.14682	16.15	<.0001
PRIN2	104.17426	42.74417	827.91045	5.94	0.0149
PRIN3	-36.23619	13.01014	1081.27785	7.76	0.0054
PRIN4	-92.13336	41.67988	681.07845	4.89	0.0272
PRIN5	-75.50591	32.55835	749.64105	5.38	0.0205
PRIN6	35.72339	15.83350	709.52482	5.09	0.0242
PRIN7	-87.73850	35.82942	835.82837	6.00	0.0144
PRIN8	24.73269	10.33265	798.61241	5.73	0.0168
PRIN10	38.62838	12.73713	1281.99518	9.20	0.0025
PRIN11	26.92501	7.81396	1654.95066	11.87	0.0006
PRIN12	55.24826	16.80045	1507.34208	10.81	0.0010
PRIN16	-49.35490	24.89254	547.94668	3.93	0.0475
PRIN17	-22.36212	8.87397	885.12857	6.35	0.0118
PRIN18	-22.49565	10.61859	625.57496	4.49	0.0342
PRIN19	8.65773	3.57768	816.24652	5.86	0.0156
PRIN20	-20.39908	7.67365	984.99340	7.07	0.0079
M_TEAM_FIELDING_DP	25.41847	7.85069	1461.16587	10.48	0.0012
M_TEAM_BASERUN_CS	25.51970	11.19513	724.28419	5.20	0.0227
T99_IMP2_TEAM_BASERUN_CS	-0.18655	0.05535	1583.46964	11.36	0.0008
M_TEAM_BASERUN_SB	-73.30512	41.14771	442.37806	3.17	0.0750
T_STD_IMP2_TEAM_BASERUN_SB	14.06255	7.51492	488.08476	3.50	0.0614
QUANT_TEAM_BATTING_HR	51.36455	27.31082	493.03051	3.54	0.0601
M_TEAM_PITCHING_SO	230.12414	98.97762	753.47116	5.41	0.0202

Figure 45: Step 7 Kitchen Sink, All Variables Without Missing Values – Backward Selection Linear Model Properties

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	37.21494	6.26816	4911.25956	35.25	<.0001
IMP_TEAM_BASERUN_CS	0.13664	0.05114	994.72460	7.14	0.0076
IMP_TEAM_BASERUN_SB	0.05551	0.00557	13817	99.17	<.0001
TEAM_BATTING_2B	0.33898	0.13230	914.61284	6.56	0.0105
TEAM_BATTING_3B	-0.42732	0.08486	3532.92081	25.36	<.0001
TEAM_BATTING_BB	0.05908	0.00658	11248	80.73	<.0001
TEAM_BATTING_H	0.07616	0.00546	27157	194.91	<.0001
TEAM_PITCHING_BB	0.00606	0.00289	613.11083	4.40	0.0360
TEAM_PITCHING_H	0.00162	0.00040638	2227.65512	15.99	<.0001
PRIN6	0.10567	0.01688	5459.93645	39.19	<.0001
PRIN12	0.08581	0.01019	9876.61403	70.89	<.0001
PRIN20	0.08974	0.01606	4351.09871	31.23	<.0001
M_TEAM_FIELDING_DP	6.77056	1.58931	2528.52807	18.15	<.0001
IMP_TEAM_FIELDING_DP	-0.05515	0.01494	1899.28883	13.63	0.0002
M_TEAM_BASERUN_CS	2.06025	0.92602	689.66816	4.95	0.0262
T99_IMP2_TEAM_BASERUN_CS	-0.21667	0.05047	2567.95897	18.43	<.0001
M_TEAM_BASERUN_SB	45.01469	2.55807	43144	309.66	<.0001
TRIM_TEAM_BATTING_2B	-0.40599	0.13332	1292.11266	9.27	0.0024
T99_TEAM_BATTING_3B	0.51651	0.08994	4594.68258	32.98	<.0001
T95_TEAM_BATTING_H	-0.04932	0.00806	5217.40671	37.45	<.0001
IMP2_TEAM_BATTING_SO	-0.11023	0.01523	7301.00597	52.40	<.0001
T95_TEAM_FIELDING_E	-0.07210	0.00538	25014	179.54	<.0001
T99_TEAM_PITCHING_HR	0.10231	0.01810	4450.19880	31.94	<.0001
T99_TEAM_PITCHING_SO	0.07434	0.01052	6961.84453	49.97	<.0001
M_TEAM_PITCHING_SO	28.10961	4.07059	6644.05830	47.69	<.0001

Figure 46: Step 7 Kitchen Sink, All Variables Without Missing Values – Stepwise Selection Linear Model Properties

In Table 42 below we can see the model metrics for the Step 7 models have greatly improved. Based on these metrics and particularly the much higher Adjusted R Squared value and lower AIC and SBC values, the model "Backward_KITCHEN_SINK" is now the current chosen model. The model is highlighted in yellow below. Note that we did not select the Stepwise model although it has better metrics and lower number of variables. The reason for this is that it's performance did not carry forward to step 8 like the backward selection model's did.

Model	# Variables	R Squared	Adjusted R Squared	CP	AIC	SBC
Forward_KITCHEN_SINK	27	0.44503	0.43837	24.8811	11264.55	11425
Backward_KITCHEN_SINK	29	0.44542	0.43826	27.3112	11266.96	11438.87
Stepwise_KITCHEN_SINK	24	0.44441	0.43849	21.3851	11261.09	11404.34

Table 42: Step 7 Kitchen Sink, All Variables Without Missing Values

STEP 8: Adjusted R Squared Selection

Logic – We now have a high performing model with too many variables. In this step we will take all of the variables selected for that model and perform selection by Adjusted R Squared, we will try selecting the top models with 10, 12 and 15 variables.

Analysis - The resulting best models for each number of variables are listed below. As can be seen below in Figure 46 and 47, models with 10 and 12 variables did not produce similar Adjusted R-Squares to our previous chosen model. As can be seen in Figure 48, models with 15 variables had a very minimal tradeoff to the Adjusted R Squared and were able to discard 14 variables. Therefore the row in Figure 48 highlighted in yellow below is our new chosen model.

Number in Model	Adjusted R-Square	R-Square	Variables in Model
10	0.4083	0.4109	IMP_TEAM_BASERUN_SB TEAM_BATTING_2B TEAM_BATTING_BB TEAM_BATTING_H PRIN2 PRIN3 PRIN8 PRIN12 PRIN16 M_TEAM_BASERUN_SB
10	0.4083	0.4109	IMP_TEAM_BASERUN_SB TEAM_BATTING_BB TEAM_BATTING_H PRIN3 PRIN8 PRIN9 PRIN12 PRIN13 PRIN16 M_TEAM_BASERUN_SB
10	0.4082	0.4108	IMP_TEAM_BASERUN_SB TEAM_BATTING_3B TEAM_BATTING_H PRIN2 PRIN3 PRIN8 PRIN12 T99_IMP2_TEAM_BASERUN_CS M_TEAM_BASERUN_SB T99_TEAM_BATTING_3B
10	0.4079	0.4105	TEAM_BATTING_BB TEAM_BATTING_H PRIN3 PRIN8 PRIN9 PRIN12 PRIN13 PRIN16 M_TEAM_BASERUN_SB T_STD_IMP2_TEAM_BASERUN_SB
10	0.4070	0.4096	IMP_TEAM_BASERUN_SB TEAM_BATTING_H PRIN3 PRIN5 PRIN8 PRIN9 PRIN12 PRIN13 PRIN16 M_TEAM_BASERUN_SB

Figure 46: Step 8 – Adjusted R Square selection for 10 variables

Number in Model	Adjusted R-Square	R-Square	Variables in Model
12	0.4194	0.4225	IMP_TEAM_BASERUN_SB TEAM_BATTING_2B TEAM_BATTING_BB TEAM_BATTING_H TEAM_PITCHING_H PRIN2 PRIN8 PRIN9 PRIN10 PRIN12 PRIN16 M_TEAM_BASERUN_SB
12	0.4194	0.4225	IMP_TEAM_BASERUN_SB TEAM_BATTING_2B TEAM_BATTING_BB TEAM_BATTING_H TEAM_PITCHING_BB PRIN2 PRIN3 PRIN8 PRIN12 PRIN16 M_TEAM_FIELDING_DP M_TEAM_BASERUN_SB
12	0.4192	0.4222	IMP_TEAM_BASERUN_SB TEAM_BATTING_2B TEAM_BATTING_BB TEAM_BATTING_H PRIN2 PRIN3 PRIN8 PRIN12 PRIN16 M_TEAM_FIELDING_DP M_TEAM_BASERUN_SB T99_TEAM_BATTING_3B
12	0.4187	0.4218	IMP_TEAM_BASERUN_SB TEAM_BATTING_2B TEAM_BATTING_BB TEAM_BATTING_H TEAM_PITCHING_H PRIN2 PRIN3 PRIN8 PRIN12 PRIN16 M_TEAM_FIELDING_DP M_TEAM_BASERUN_SB
12	0.4187	0.4218	IMP_TEAM_BASERUN_SB TEAM_BATTING_2B TEAM_BATTING_BB TEAM_BATTING_H TEAM_PITCHING_BB PRIN2 PRIN8 PRIN9 PRIN10 PRIN12 PRIN16 M_TEAM_BASERUN_SB

Figure 47: Step 8 – Adjusted R Square selection for 10 variables

Number in Model	Adjusted R-Square	R-Square	Variables in Model
15	0.4326	0.4363	IMP_TEAM_BASERUN_SB TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_BB TEAM_BATTING_H TEAM_PITCHING_H PRIN2 PRIN3 PRIN8 PRIN12 PRIN16 PRIN17 M_TEAM_FIELDING_DP M_TEAM_BASERUN_SB T99_TEAM_BATTING_3B
15	0.4321	0.4358	IMP_TEAM_BASERUN_CS IMP_TEAM_BASERUN_SB TEAM_BATTING_2B TEAM_BATTING_BB TEAM_BATTING_H TEAM_PITCHING_H PRIN2 PRIN8 PRIN9 PRIN10 PRIN12 PRIN16 M_TEAM_FIELDING_DP T99_IMP2_TEAM_BASERUN_CS M_TEAM_BASERUN_SB
15	0.4314	0.4351	IMP_TEAM_BASERUN_SB TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_BB TEAM_BATTING_H TEAM_PITCHING_BB TEAM_PITCHING_H PRIN2 PRIN3 PRIN8 PRIN12 PRIN16 M_TEAM_FIELDING_DP M_TEAM_BASERUN_SB T99_TEAM_BATTING_3B
15	0.4312	0.4350	TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_BB TEAM_BATTING_H TEAM_PITCHING_H PRIN2 PRIN3 PRIN8 PRIN12 PRIN16 PRIN17 M_TEAM_FIELDING_DP M_TEAM_BASERUN_SB T_STD_IMP2_TEAM_BASERUN_SB T99_TEAM_BATTING_3B
15	0.4310	0.4347	IMP_TEAM_BASERUN_SB TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_BB TEAM_BATTING_H TEAM_PITCHING_H PRIN2 PRIN8 PRIN9 PRIN10 PRIN12 PRIN16 M_TEAM_FIELDING_DP M_TEAM_BASERUN_SB T99_TEAM_BATTING_3B

Figure 48: Step 8 – Adjusted R Square selection for 10 variables

Winning Model Sanity Check - Before going any further, we need to perform a simple sanity check to see the qualities of the chosen model. Further analysis will be performed in the next section.

The linear model properties are shown in Figure 49. As has been consistent with the previous model, this model also has a few variables included in a counter intuitive fashion (highlighted in red). For example: TEAM_BATTING_3B is included as a negative coefficient, when it should positively contribute to the team wins.
Note that there are no expectations for the PCA variables to have a positive or negative coefficient. Also note that the missing values are not expected to follow the sign coefficient of their base variable as it is unclear why a variable may be missing

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	35.98333	4.77270	7.54	<.0001
IMP_TEAM_BASERUN_SB		1	0.06297	0.00435	14.47	<.0001
TEAM_BATTING_2B	Doubles by batters (2B)	1	0.08779	0.01112	7.90	<.0001
TEAM_BATTING_3B	Triples by batters (3B)	1	-0.34454	0.08011	-4.30	<.0001
TEAM_BATTING_BB	Walks by batters	1	0.06096	0.00677	9.01	<.0001
TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)	1	0.06921	0.00419	16.52	<.0001
TEAM_PITCHING_H	Hits allowed	1	0.00213	0.00033205	6.42	<.0001
PRIN2		1	-0.13676	0.00710	-19.26	<.0001
PRIN3		1	0.06545	0.00563	11.63	<.0001
PRIN8		1	-0.33491	0.02852	-11.74	<.0001
PRIN12		1	-0.33794	0.02351	-14.38	<.0001
PRIN16		1	0.11380	0.01348	8.44	<.0001
PRIN17		1	0.02924	0.00865	3.38	0.0007
M_TEAM_FIELDING_DP		1	7.89842	1.48347	5.32	<.0001
M_TEAM_BASERUN_SB		1	38.90386	2.15184	18.08	<.0001
T99_TEAM_BATTING_3B		1	0.42570	0.08440	5.04	<.0001

Figure 49: Winning Model Linear Model Properties

In Table 43 below we can see the model metrics are very close to the "Backward_KITCHEN_SINK" which was our highest performing model. However, in the winning model there are almost half as many variables, so the small performance metric trade off is deemed warranted.

Model	# Variables	R Squared	Adjusted R Squared	CP	AIC	SBC
WinningModel	15	0.43638	0.43264	16	11275.74	11367.43

Table 43: Step 8 Winning Model

Part 4: Select Models

As discussed in the previous section, the winning model was chosen from over 23 candidate models. It was selected for having the highest Adjusted R Squared value and the lowest AIC values for a number of variables within reason. The only models which beat out the predictive power of our winning model were the kitchen sink models. However, they had close to twice as many variables. The slight extra predictive ability in the model was not worth the extra variables and the risk of overfitting.

In the previous section we had a brief look at the coefficients and linear model properties. This topic will be revisited, but for now we begin the model analysis by reviewing model adequacy. We start by examining the Cook's D diagnostic plot in Figure 50. This plot gives us an idea of which observations may be influential points. A point is influential if it has undue influence over the parameters in the model. For example, if its deletion causes a change in the fitted regression model. When examining this plot we are looking to have most points around the same value, no points above the threshold (horizontal line on the chart) and no points greater than one.

While no points are greater than 1, there are many observations above the threshold and many which have a value larger than the rest.

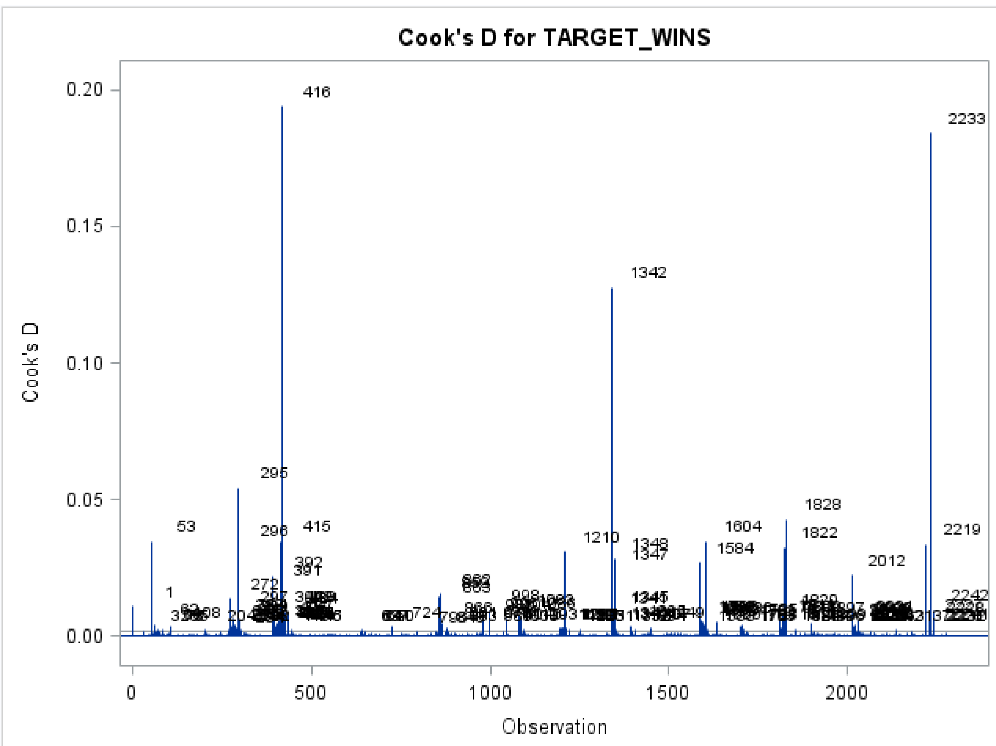


Figure 50: Cook's D Diagnostic Plot for the Winning Model

To remedy this situation, all of the observations with visibly high Cook's D points were examined. The observations were examined for unusual or extreme values in the dependent or predictor variables. Tests were performed to see what effect their deletion had on the regression model. Finally, the following list of influential points was deleted from the model: 1494, 460, 322, 323, 57, 2047, 2345, 2486, 2219, 1211, 1342, 1828, 416,296, 1822
*Note: Observations are referenced by INDEX

The model was then re-pulled on the new data set with leverage points removed. As can be seen below, because of this action the Adjusted R Squared value was raised and the AIC value was lowered. After removing the leverage points, the model was able to be estimated with more predictive power.

Model	# Variables	R Squared	Adjusted R Squared	CP	AIC	SBC
WinningModel -Leverage Removed	15	0.44727	0.44358	16	11112.36	11203

Table 44: Winning Model – Leverage Points Removed

We continue the analysis by re-pulling the Cook's D plot on the new model. It is clear that there are still some a number of leverage points. However, they are not as extreme as previously and a number of the most offending observations have been removed. To dig further into the leverage points is outside of the scope of this assignment.

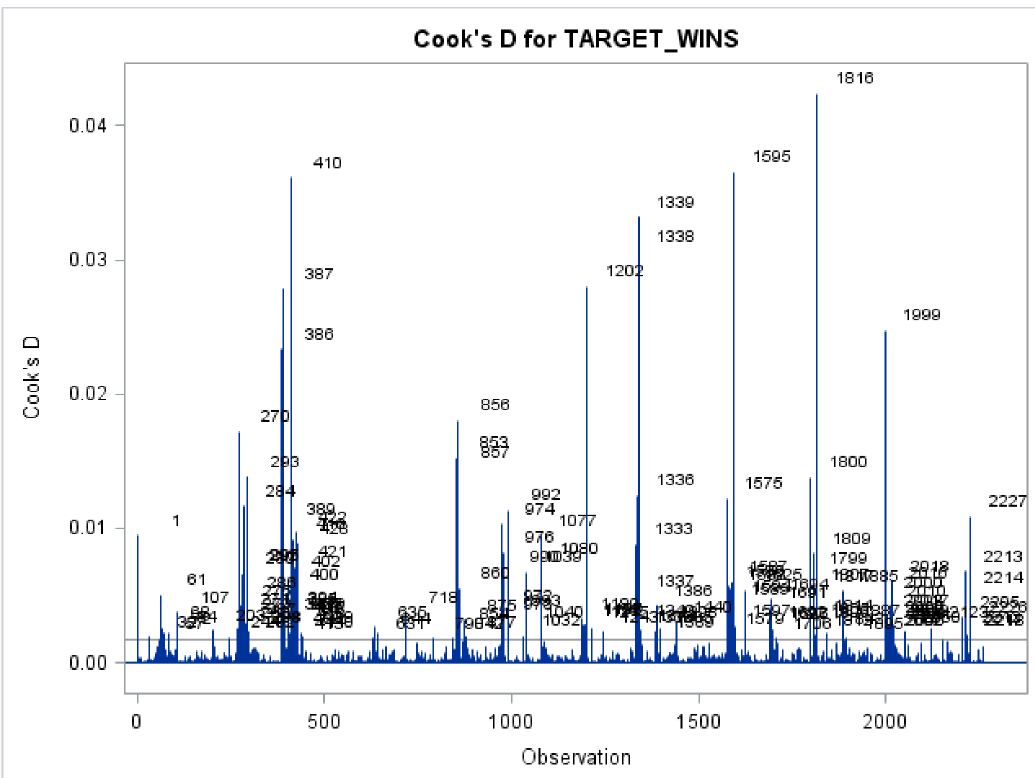


Figure 51: Cook's D Diagnostic Plot for the Winning Model with Leverage Points Removed

Next we review the residuals of the TARGET_WINS vs each of the predictor variables. We are looking for the points to be randomly distributed about zero with no apparent pattern or structure such as a curve to the points. Examples of other patterns would be residuals that increase or decrease with the larger predictive values. The graphs in Figures 52 and 53 below meets both of these conditions. Any groupings of points are as a result of outliers and not due to any other emerging pattern. Therefore, we can keep our assumption that the selected linear regression model is an appropriate fit for the data.

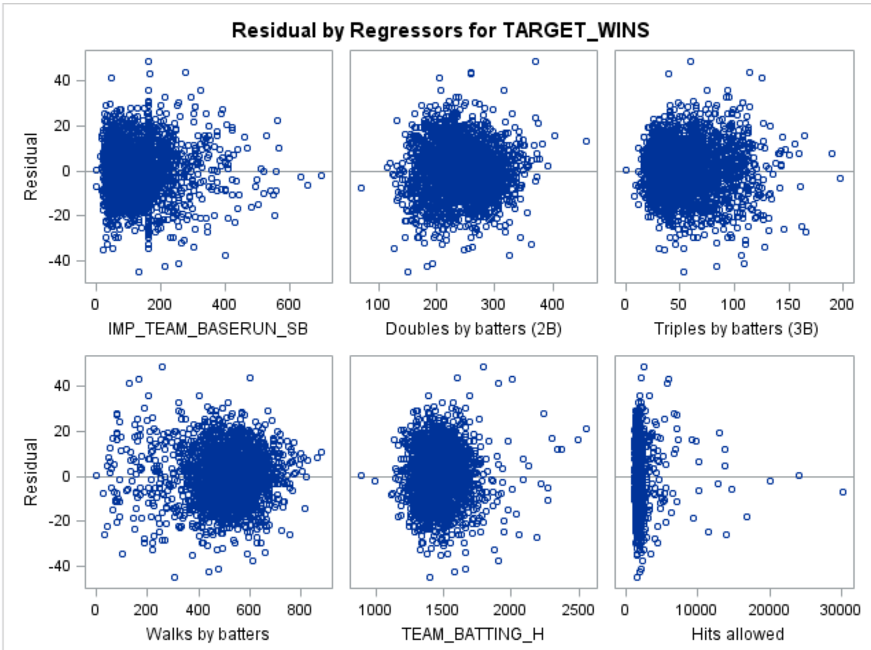


Figure 52: Residual Plots for the Winning Model with Leverage Points Removed

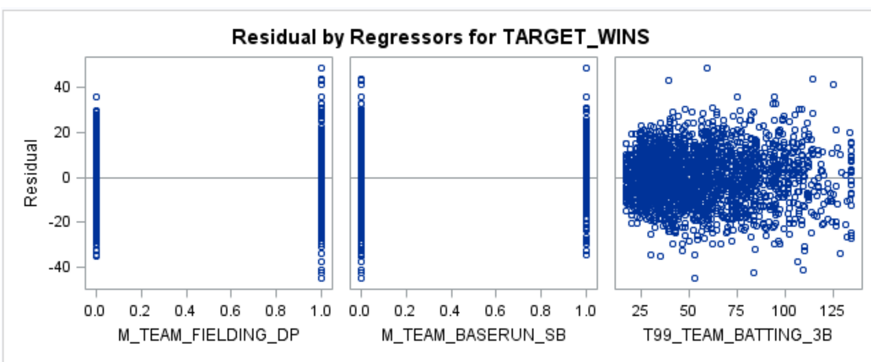


Figure 53: Residual Plots for the Winning Model with Leverage Points Removed Continued

We continue to review the residuals for the principal component variables in Figure 54. Upon looking at these graphs it is clear that each of them is exhibiting a pattern. The variability of the error term is increasing as the independent variables are getting larger. This effect is called Homoscedasticity ("same scatter"). When homoscedasticity occurs, you can find biased standard errors of coefficients and the ability to make inferences form the model is hampered. To correct this issue, often the natural logarithm of the dependent variable is taken and the regression model is re-estimated. Another method is to find the variable causing this issue, transform it and re-run the regression. Both of these exercises are outside of the scope of this assignment.

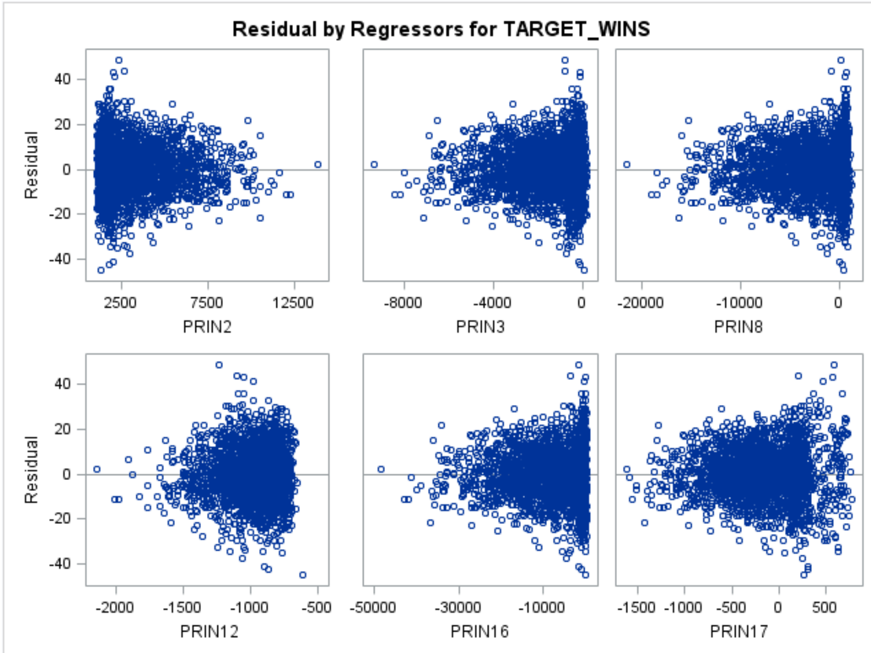


Figure 54: Residual Plots for the Winning Model with Leverage Points Removed Continued

Next we will review the Q-Q plot in Figure 55 to ensure that the residuals are normally distributed. To confirm a normal distribution, the residual values should be distributed closely to a 45 degree line. Any serious deviation from a straight line would suggest that the data is not normally distributed. The values are in fact hugging the 45 degree line and we can conclude that the quantiles of the residual follow the quantiles of a normal distribution. Therefore this graph does not introduce any concern about the adequacy of our chosen optimal model.

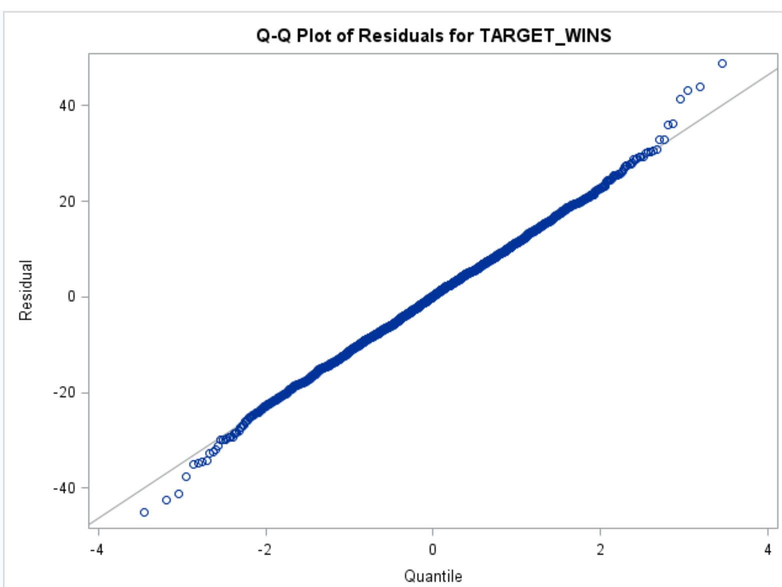


Figure 55: Q-Q Plot for the Winning Model with Leverage Points Removed

Next we look at the estimates for the model to see if it is intuitive. Upon inspection, there are five things (highlighted in red) about this model that can be called into question.

1. **TEAM_BATTING_3B has a negative coefficient** when it should positively contribute to TARGET_WINs and therefore have a positive coefficient. To solve this issue we will examine the affects of removing the variable and re running the model .
2. **TEAM_PITCHING_H has a positive coefficient** when it should negatively contribute to TARGET_WINs and therefore have a negative coefficient. To solve this issue we will examine the affects of removing the variable and re running the model .
3. **The VIF's associated with all of the PRIN variables are extremely high**. This suggests multicollinearity. Multicollinearity can lead to a range of issues including unusual regression coefficients and biased standard errors. One approach is to remedy the issue is to center the variables that comprise the principal component variables and then subsequently re-calculate the principal components and re-pull the model. Another approach is to collect more data and re-pull the model. These actions are outside of the scope of this assignment. However, we will perform trials to see if any of the variables can be removed in an attempt to simplify the model and reduce multicollinearity.
4. **M_TEAM_BASERUN_SB has an unusually high coefficient**. This regression coefficient could cause warning signs as it is a lot of weight to place on a missing value flag. However, when digging into the numbers deeper, it is actually not that much weight relative to the other variables in the regression equation. Many of the PRIN variables can go as high as 10000. When that number is multiplied by their coefficients (0.3 for example) it can produce a value of 3000 to add to the regression equation. As M_TEAM_BASERUN_SB is a flag variable, it's maximum contribution to the regression equation will be 40.39. Therefore, respectively it is not an issue.
5. **T99_TEAM_BATTING_3B is based off of the same variable as TEAM_BATTING_3B** and therefore is redundant. As discussed in bullet 1 we will try removing TEAM_BATTING_3B and re-pulling the model.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	40.49356	4.85055	8.35	<.0001	0
IMP_TEAM_BASERUN_SB		1	0.06550	0.00430	15.22	<.0001	2.25406
TEAM_BATTING_2B	Doubles by batters (2B)	1	0.09190	0.01099	8.36	<.0001	4.36429
TEAM_BATTING_3B	Triples by batters (3B)	1	-0.53924	0.10799	-4.99	<.0001	147.54574
TEAM_BATTING_BB	Walks by batters	1	0.06393	0.00674	9.48	<.0001	11.18005
TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)	1	0.06587	0.00421	15.65	<.0001	5.91836
TEAM_PITCHING_H	Hits allowed	1	0.00208	0.00035908	5.78	<.0001	3.74362
PRIN2		1	-0.13659	0.00729	-18.74	<.0001	3523.15322
PRIN3		1	0.06685	0.00557	11.99	<.0001	1295.19616
PRIN8		1	-0.35997	0.02823	-12.75	<.0001	187865
PRIN12		1	-0.35469	0.02322	-15.27	<.0001	368.39362
PRIN16		1	0.12551	0.01340	9.37	<.0001	198027
PRIN17		1	0.02896	0.00895	3.23	0.0012	226.13668
M_TEAM_FIELDING_DP		1	7.70700	1.47012	5.24	<.0001	3.89372
M_TEAM_BASERUN_SB		1	40.39829	2.13928	18.88	<.0001	4.08348
T99_TEAM_BATTING_3B		1	0.62561	0.11150	5.61	<.0001	149.83411

Figure 55: Parameter Estimates for the Winning Model with Leverage Points Removed

As discussed above TEAM_BATTING_3B and TEAM_PITCHING_H were removed from the model. All PRIN variables were also tested for removal and PRIN17 was removed. Figure 56 shows the parameter estimates for the new model. The model now has intuitive coefficients and there are no redundant variables. The variance inflation values for the PRIN variables are still very high. Techniques have been discussed to address this issue, however they are outside of the scope of this assignment and the model is now accepted as is.

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	37.28529	4.65975	8.00	<.0001	0
IMP_TEAM_BASERUN_SB		1	0.06435	0.00413	15.57	<.0001	2.03108
TEAM_BATTING_2B	Doubles by batters (2B)	1	0.08198	0.01093	7.50	<.0001	4.22107
TEAM_BATTING_BB	Walks by batters	1	0.04386	0.00521	8.42	<.0001	6.52414
TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)	1	0.05756	0.00405	14.21	<.0001	5.35746
PRIN2		1	-0.11056	0.00599	-18.45	<.0001	2328.28828
PRIN3		1	0.06685	0.00556	12.03	<.0001	1259.26471
PRIN8		1	-0.34557	0.02794	-12.37	<.0001	180009
PRIN12		1	-0.34605	0.02266	-15.27	<.0001	343.00667
PRIN16		1	0.12637	0.01318	9.59	<.0001	187265
M_TEAM_FIELDING_DP		1	6.68237	1.47558	4.53	<.0001	3.83607
M_TEAM_BASERUN_SB		1	38.55465	1.92978	19.98	<.0001	3.24945
T99_TEAM_BATTING_3B		1	0.06853	0.01543	4.44	<.0001	2.80744

Figure 56: Parameter Estimates for the Final Model

As can be seen in Table 45, after removing the three variables the Adjusted R Squared went down by approximately 0.01. This is a very small trade off to achieve higher parsimony and a more intuitive model.

Model	# Variables	R Squared	Adjusted R Squared	CP	AIC	SBC
Final Model	15	0.43403	0.43101	13	11159.87	11234.28

Table 45: Final Model Metrics

Conclusion

Several techniques such as variable transformations, principal component analysis, tree logic, adjusted r squared, forward, backward and stepwise variable selection were performed to select the best regression model to predict the number of season wins for a baseball team. The quality metrics Adjusted R-Squared and AIC were compared for all considered models. The winning regression model was then chosen and diagnostic plots such as Q-Q Plot, Cook's D and Residual Plots were examined for model adequacy. As a result of the examination a few small changes were made to the model, such as deleting leverage points and dropping variables. The final winning regression model is defined as:

*TARGET_WINS = 37.28529 + 0.06435 * IMP_TEAM_BASERUN_SB*
*+ 0.08198 * TEAM_BATTING_2B*
*+ 0.04386 * TEAM_BATTING_BB*
*+ 0.05756 * TEAM_BATTING_H*
*-0.11056 * PRIN2*
*+ 0.06685 * PRIN3*
*-0.34557 * PRIN8*
*-0.34605 * PRIN12*
*+ 0.12637 * PRIN16*
*+ 6.68237 * M_TEAM_FIELDING_DP*
*+ 38.55465 * M_TEAM_BASERUN_SB*
*+ 0.06853 * T99_TEAM_BATTING_3B*

This model is intuitive because the variables representing positive team performance such as IMP_TEAM_BASERUN_SB, TEAM_BATTING_2B, TEAM_BATTING_H and T99_TEAM_BATTING_3B all have positive coefficients and therefore contribute positively to the number of team wins. Interestingly there are no variables explicitly representing negative performance. The principal component variables (PRIN2, PRIN3, PRIN8, PRIN12 and PRIN16) are a combination of all base variables and they are therefore not expected to have a positive or negative coefficient. Finally, the missing value flag variables (M_TEAM_FIELDING_DP and M_TEAM_BASERUN_SB) are not expected to follow the sign coefficient of their base variable as it is unclear why a variable may be missing.

The analysis also touched on further examination that could be done to improve the existing model. For example, further analysis and transformations could be performed to reduce the effects of multicollinearity and homoscedasticity. This analysis is outside of the scope of this assignment but should be considered for future examination.

A final note is that this analysis was performed on data from 1871 to 2006. A lot has changed in the way baseball is played and how baseball metrics are gathered during this time period. Therefore, it is recommended to test the model on more recent data (2006-2014) before using it to make present day predictions.

.

Bingo Bonus Points – Expected (80 + ?? Points)

1 Hand in your SCORED FILE as a SAS DATA SET and save me to trouble of converting it. (10/10 points)

- Done! Sent!

2 Use PROC GLM PROC GENMOD to do the OLS Regression (20/20 points)

Summary

Proc GLM produced the same model. It also included a listing of the partial sum of squares and the mean square error for each of the predictor variables.

One of the main differences between PROC GLM and PROC REG is that PROC REG allows for several model statements to be run at once. Also it generates the model adequacy graphs (residuals, cooks, QQ etc).

PROC GENMOD also produced the same model. However its coefficients have been trimmed to only have three decimal places. Proc GENMOD is for generalized linear models. With this procedure you can specify both the link function and the distribution that you would like to use.

Output below

```
proc glm data=dropLeverage;
model TARGET_WINS=IMP_TEAM_BASERUN_SB TEAM_BATTING_2B  TEAM_BATTING_BB TEAM_BATTING_H  PRIN2
PRIN3 PRIN8 PRIN12 PRIN16 M_TEAM_FIELDING_DP M_TEAM_BASERUN_SB T99_TEAM_BATTING_3B;
run;
quit;
```

Source	DF	Type I SS	Mean Square	F Value	Pr > F
IMP_TEAM_BASERUN_SB	1	7886.81991	7886.81991	56.99	<.0001
TEAM_BATTING_2B	1	52915.73373	52915.73373	382.37	<.0001
TEAM_BATTING_BB	1	18126.91688	18126.91688	130.98	<.0001
TEAM_BATTING_H	1	46801.86222	46801.86222	338.19	<.0001
PRIN2	1	5599.70420	5599.70420	40.46	<.0001
PRIN3	1	1118.27239	1118.27239	8.08	0.0045
PRIN8	1	9328.74434	9328.74434	67.41	<.0001
PRIN12	1	13315.93292	13315.93292	96.22	<.0001
PRIN16	1	24928.22517	24928.22517	180.13	<.0001
M_TEAM_FIELDING_DP	1	54.85302	54.85302	0.40	0.5290
M_TEAM_BASERUN_SB	1	55774.84871	55774.84871	403.03	<.0001
T99_TEAM_BATTING_3B	1	2728.60200	2728.60200	19.72	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
IMP_TEAM_BASERUN_SB	1	33568.25203	33568.25203	242.56	<.0001
TEAM_BATTING_2B	1	7789.55552	7789.55552	56.29	<.0001
TEAM_BATTING_BB	1	9810.68190	9810.68190	70.89	<.0001
TEAM_BATTING_H	1	27952.23998	27952.23998	201.98	<.0001
PRIN2	1	47097.15533	47097.15533	340.32	<.0001
PRIN3	1	20016.25481	20016.25481	144.64	<.0001
PRIN8	1	21163.05609	21163.05609	152.92	<.0001
PRIN12	1	32280.20861	32280.20861	233.25	<.0001
PRIN16	1	12729.00414	12729.00414	91.98	<.0001
M_TEAM_FIELDING_DP	1	2838.18663	2838.18663	20.51	<.0001
M_TEAM_BASERUN_SB	1	55238.68225	55238.68225	399.15	<.0001
T99_TEAM_BATTING_3B	1	2728.60200	2728.60200	19.72	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	37.28528796	4.65975161	8.00	<.0001
IMP_TEAM_BASERUN_SB	0.06435140	0.00413187	15.57	<.0001
TEAM_BATTING_2B	0.08198267	0.01092745	7.50	<.0001
TEAM_BATTING_BB	0.04385635	0.00520878	8.42	<.0001
TEAM_BATTING_H	0.05756244	0.00405027	14.21	<.0001
PRIN2	-0.11055841	0.00599304	-18.45	<.0001
PRIN3	0.06684614	0.00555825	12.03	<.0001
PRIN8	-0.34556841	0.02794460	-12.37	<.0001
PRIN12	-0.34604587	0.02265784	-15.27	<.0001
PRIN16	0.12636836	0.01317632	9.59	<.0001
M_TEAM_FIELDING_DP	6.68236890	1.47558082	4.53	<.0001
M_TEAM_BASERUN_SB	38.55464783	1.92978140	19.98	<.0001
T99_TEAM_BATTING_3B	0.06852990	0.01543345	4.44	<.0001

```
proc genmod data=dropLeverage;
model TARGET_WINS=IMP_TEAM_BASERUN_SB TEAM_BATTING_2B  TEAM_BATTING_BB TEAM_BATTING_H  PRIN2
PRIN3 PRIN8 PRIN12 PRIN16 M_TEAM_FIELDING_DP M_TEAM_BASERUN_SB T99_TEAM_BATTING_3B/ link=identity
dist=normal;
run;
```

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	2248	311101.6871	138.3904
Scaled Deviance	2248	2261.0000	1.0058
Pearson Chi-Square	2248	311101.6871	138.3904
Scaled Pearson X2	2248	2261.0000	1.0058
Log Likelihood		-8775.1555	
Full Log Likelihood		-8775.1555	
AIC (smaller is better)		17578.3109	
AICC (smaller is better)		17578.4979	
BIC (smaller is better)		17658.4408	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	37.2853	4.6463	28.1786	46.3919	64.40	<.0001
IMP_TEAM_BASERUN_SB	1	0.0644	0.0041	0.0563	0.0724	243.96	<.0001
TEAM_BATTING_2B	1	0.0820	0.0109	0.0606	0.1033	56.61	<.0001
TEAM_BATTING_BB	1	0.0439	0.0052	0.0337	0.0540	71.30	<.0001
TEAM_BATTING_H	1	0.0576	0.0040	0.0496	0.0655	203.15	<.0001
PRIN2	1	-0.1106	0.0060	-0.1223	-0.0988	342.29	<.0001
PRIN3	1	0.0668	0.0055	0.0560	0.0777	145.47	<.0001
PRIN8	1	-0.3456	0.0279	-0.4002	-0.2910	153.81	<.0001
PRIN12	1	-0.3460	0.0226	-0.3903	-0.3018	234.60	<.0001
PRIN16	1	0.1264	0.0131	0.1006	0.1521	92.51	<.0001
M_TEAM_FIELDING_DP	1	6.6824	1.4713	3.7986	9.5661	20.63	<.0001
M_TEAM_BASERUN_SB	1	38.5546	1.9242	34.7832	42.3261	401.46	<.0001
T99_TEAM_BATTING_3B	1	0.0685	0.0154	0.0384	0.0987	19.83	<.0001
Scale	1	11.7301	0.1744	11.3931	12.0770		

3 Trees (20/20 points)

a) I used SPSS to generate trees to replace missing values for the following variables below:

- TEAM_BATTING_SO (Part 1 Fig 8)
- Team_BASERUN_SB (Part 1 Fig 10) – Note – this variable using decision tree missing value replacement even made it into the model)
- TEAM_BASERUN_CS (Part 1 Fig 13)

b) I also used trees for variable selection - (Part 2, Step3, Figure 28)

4 SAS Macros (Expected 10/10 Points)

I wrote a SAS Macros to generate all of the statistics tables and plots for use in our Exploratory Data Analysis. I then called on the macros for every variable to generate the plots/data automatically.

```
*write a macos to generate statistics;
%macro generateStats(c);

      *generate the descriptive stats and number of missing records;
      proc means data=mb n nmiss mean stddev p1 p5 p95 p99 min max;
      var &c;
      run;

      *generate the bar graph;
      goptions hsize=4in vsize=3in;
      proc univariate data=mb noprint;
      histogram &c;
      run;
```

```
        *generate a box plot;
        ods graphics on / width=1.5in height=2in;
        proc sgplot data= mb;
        vbox &c;
        run;

        *produce a scatterplot;
        ods graphics on / width=3.6in height=2.75in;

        PROC SGPLOT DATA=mb;
        REG X=&c Y=TARGET_WINS /NOMARKERS lineattrs = (color = blue thickness = 4 pattern=solid);
        LOESS X=&c Y=TARGET_WINS / lineattrs = (color = red thickness = 4 pattern=solid);
        run;
        ods graphics on / reset=all;

        *Get the extreme observations for the variable;
        ods select ExtremeObs;
        proc univariate data=mb;
           var &c;
        run;

        ods graphics off;

%mend generateStats;

ods graphics on;

%generateStats(TARGET_WINS);
%generateStats(TEAM_BATTING_H);
%generateStats(TEAM_BATTING_2B);

etc
```

5 Roll the Dice (?? Points) – Using PCA

I used principal component analysis to attempt to deal with the highly correlated variables (covered in Part 2). In the end using PCA did provide higher model performance. Since there were so many variables in the Eigenvectors I was worried I would make a mistake (and go crazy) if I copied and pasted by hand. Instead I created an excel macros to automatically assign the formulas. Here is a snippet of the sheet set up and the code, however I can send the xls if need be.

Eigenvectors

	A	B	C	D	E	F	G	H	I	J	K	L	M
2			Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11
3	TEAM_FIELDING_E	Errors	0.282732	0.194831	-0.18878	0.091957	-0.06775	-0.09484	0.095525	0.071801	-0.24178	-0.26205	-0.09527
4	T95_TEAM_BATTING_H		0.058617	0.361159	0.25976	0.069305	0.426537	0.196581	0.072841	-0.03932	-0.02697	0.069523	-0.03535
	TRIM_T		-0.13189	0.287646	0.141254	0.109587	0.377235	0.314832	-0.09906	0.194124	0.44207	-0.40456	0.041536

	A	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
2		Prin12	Prin13	Prin14	Prin15	Prin16	Prin17	Prin18	Prin19	Prin20	Prin21			
3	TEAM_FIELDING_E	-0.07861	0.19645	-0.19463	0.519649	0.097843	-0.35419	0.234441	0.342051	-0.11948	0			PRIN1
4	T95_TEAM_BATTING_H	-0.00785	-0.4469	0.424796	0.34088	-0.0365	-0.17155	-0.1474	-0.08581	-0.05771	0			PRIN2
	TRIM_T	0.086716	0.385628	-0.22408	-0.09752	0.020574	0.031093	0.012213	-0.01391	-0.0219	0			

```
Sub CREATEPCAFULL()

For i = 3 To 23

   For j = 1 To 21

   Cells(i, 27).Value = Cells(i, 27).Value & " + " & Cells(2 + j, i).Value & " * " & Cells(2 + j, 1).Value

   Next j
```

Next i

End Sub

6 Recreate as much of the program as you can in "R" (20/20 Points)

What: Open the File

Code:

#start by opening the file and loading it into the variable "mb"

mb <- read.csv(file = "C:\\Files\\Masters\\411\\Assignment1\\moneyballTransformed.csv", header=TRUE, sep=",")

require(UsingR)

require(car)

require(MASS)

require(GGally)

require(ggplot2)

require(ggthemes)

#EDA

#List out the variables in the data set

names(mb)

#View first 10 rows

head(mb, n=10)

Results: (I did not include the first 10 rows as it would take up too much space.

```
> names(mb)
 [1] "INDEX"                    "TARGET_WINS"              "TEAM_BATTING_H"
"TEAM_BATTING_2B"
 [5] "TEAM_BATTING_3B"          "TEAM_BATTING_HR"          "TEAM_BATTING_BB"
"TEAM_BATTING_SO"
 [9] "TEAM_BASERUN_SB"          "TEAM_BASERUN_CS"          "TEAM_PITCHING_H"
"TEAM_PITCHING_HR"
[13] "TEAM_PITCHING_BB"         "TEAM_PITCHING_SO"         "TEAM_FIELDING_E"
"TEAM_FIELDING_DP"
[17] "T95_TEAM_BATTING_H"       "TRIM_TEAM_BATTING_2B"     "T99_TEAM_BATTING_3B"
"QUANT_TEAM_BATTING_HR"
[21] "TRIM_TEAM_BATTING_HR"     "COMB_HR"                  "STD_TEAM_BATTING_BB"
"M_TEAM_BATTING_SO"
[25] "IMP_TEAM_BATTING_SO"      "IMP2_TEAM_BATTING_SO"     "M_TEAM_BASERUN_SB"
"IMP_TEAM_BASERUN_SB"
[29] "IMP2_TEAM_BASERUN_SB"     "STD_IMP_TEAM_BASERUN_SB"  "T_STD_IMP_TEAM_BASERUN_SB"
"STD_IMP2_TEAM_BASERUN_SB"
```

```
[33] "T_STD_IMP2_TEAM_BASERUN_SB" "M_TEAM_BASERUN_CS"         "IMP_TEAM_BASERUN_CS"
"IMP2_TEAM_BASERUN_CS"
[37] "T99_IMP_TEAM_BASERUN_CS"    "T99_IMP2_TEAM_BASERUN_CS"  "T95_TEAM_FIELDING_E"
"M_TEAM_FIELDING_DP"
[41] "IMP_TEAM_FIELDING_DP"       "T99_TEAM_PITCHING_BB"      "T95_TEAM_PITCHING_H"
"T99_TEAM_PITCHING_HR"
[45] "T99_TEAM_PITCHING_SO"       "M_TEAM_PITCHING_SO"        "PRIN1"
"PRIN2"
[49] "PRIN3"                      "PRIN4"                     "PRIN5"
"PRIN6"
[53] "PRIN7"                      "PRIN8"                     "PRIN9"
"PRIN10"
[57] "PRIN11"                     "PRIN12"                    "PRIN13"
"PRIN14"
[61] "PRIN15"                     "PRIN16"                    "PRIN17"
"PRIN18"
[65] "PRIN19"                     "PRIN20"                    "PRIN21"
>
> #View first 10 rows
```

What: EDA – Create Histogram with normal curve

Code:

#Create Histogram - Use TARGET_WINS as an example

Y = mb$TARGET_WINS

h<-hist(Y, breaks=10, density=10, col="lightgray", xlab="Accuracy", main="Overall")

#Add normal curve

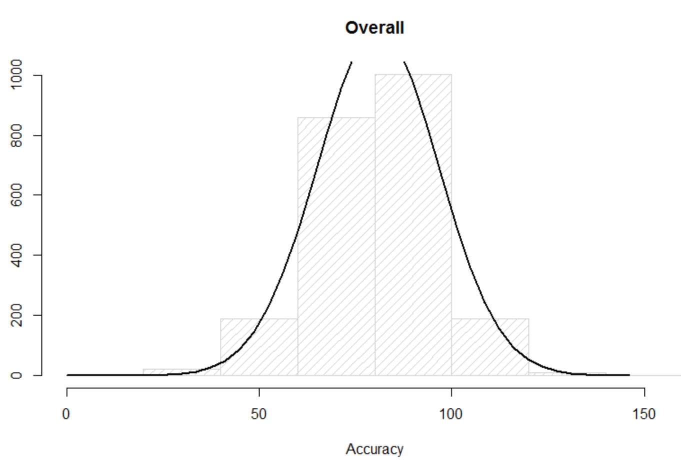
xfit<-seq(min(Y),max(Y),length=40)

yfit<-dnorm(xfit,mean=mean(Y),sd=sd(Y))

yfit <- yfit*diff(h$mids[1:2])*length(Y)

lines(xfit, yfit, col="black", lwd=2)

Results:

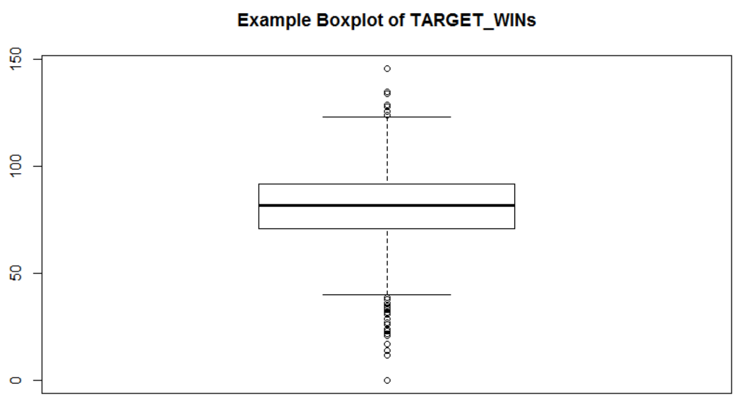


What: EDA – Example BoxPlot

Code:

boxplot(mb$TARGET_WINS,main="Example Boxplot of TARGET_WINs")

Results:



What: Calculate the mean statistics

Code: mean(mb$TARGET_WINS)

quantile(mb$TARGET_WINS, c(.01, .05, .95, .99))

Results:

```
> mean(mb$TARGET_WINS)
[1] 80.79086
> quantile(mb$TARGET_WINS, c(.01, .05, .95, .99))
    1%     5%    95%    99% 
 38.75  54.00 104.00 114.00 
```

What: Calculate Correlation Numbers and Graphs

Code:

#correlation table on all of the base variables

cor(mb[, c(2:25)])

#correlation plot on subset of data b/c the whole data is too large to plot (same as SAS)

require(GGally)

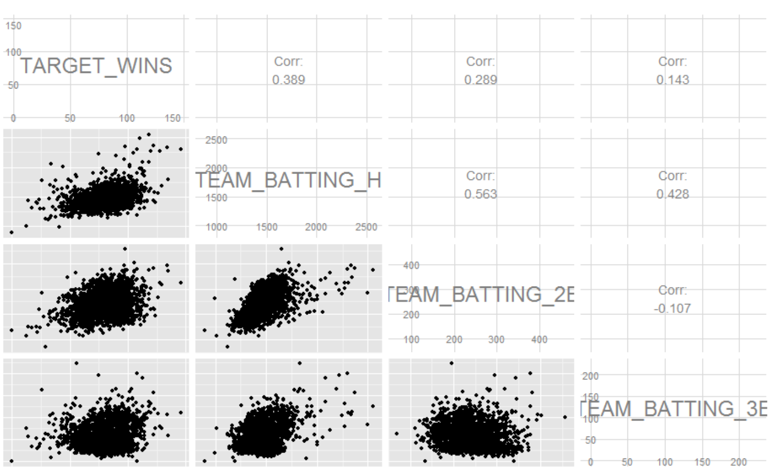
GGally::ggpairs(mb[,c(2:5)], params=list(labelSize=8))

Results:

```
                       TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
TEAM_BATTING_HR TEAM_BATTING_BB TEAM_BATTING_SO
TARGET_WINS            1.000000000    0.388767521      0.28910365     0.142608411
0.176153200     0.232559864              NA
TEAM_BATTING_H         0.388767521    1.000000000      0.56284968     0.427696575     -
0.006544685    -0.072464013              NA
TEAM_BATTING_2B        0.289103645    0.562849678      1.00000000    -0.107305824
0.435397293     0.255726103              NA
TEAM_BATTING_3B        0.142608411    0.427696575     -0.10730582     1.000000000     -
0.635566946    -0.287235841              NA
TEAM_BATTING_HR        0.176153200   -0.006544685      0.43539729    -0.635566946
1.000000000     0.513734810              NA
TEAM_BATTING_BB        0.232559864   -0.072464013      0.25572610    -0.287235841
0.513734810     1.000000000              NA
TEAM_BATTING_SO                 NA             NA              NA              NA
NA              NA                1
TEAM_BASERUN_SB                 NA             NA              NA              NA
NA              NA               NA
TEAM_BASERUN_CS                 NA             NA              NA              NA
NA              NA               NA
TEAM_PITCHING_H       -0.109937054    0.302693709      0.02369219     0.194879411     -
0.250145481    -0.449777625              NA
TEAM_PITCHING_HR       0.189013735    0.072853119      0.45455082    -0.567836679
0.969371396     0.459552072              NA
TEAM_PITCHING_BB       0.124174536    0.094193027      0.17805420    -0.002224148
0.136927564     0.489361263              NA
TEAM_PITCHING_SO                NA             NA              NA              NA
NA              NA               NA
TEAM_FIELDING_E       -0.176484759    0.264902478     -0.23515099     0.509778447     -
0.587339098    -0.655970815              NA
TEAM_FIELDING_DP                NA             NA              NA              NA
NA              NA               NA
T95_TEAM_BATTING_H     0.368992927    0.926574545      0.58929342     0.397713027
0.035386692    -0.004283097              NA
TRIM_TEAM_BATTING_2B   0.287758157    0.562763761      0.99917840    -0.109594178
0.438324645     0.257919578              NA
T99_TEAM_BATTING_3B    0.143500519    0.411831815     -0.11774935     0.993050614     -
0.649585452    -0.290983375              NA
QUANT_TEAM_BATTING_HR  0.165818342   -0.012846819      0.40906358    -0.621785635
0.956900151     0.495097418              NA
TRIM_TEAM_BATTING_HR   0.176386766   -0.007291966      0.43504482    -0.636162608
0.999837180     0.514239690              NA
COMB_HR                0.179668396    0.032857339      0.41591694    -0.566245586
0.953672992     0.431648171              NA
STD_TEAM_BATTING_BB    0.232559864   -0.072464013      0.25572610    -0.287235841
0.513734810     1.000000000              NA
M_TEAM_BATTING_SO      0.007730792   -0.171367201     -0.26060568     0.131790907     -
0.285259870    -0.133620457              NA
IMP_TEAM_BATTING_SO   -0.031042737   -0.403669051      0.20714364    -0.669876844
0.741269794     0.393232663              NA
                      TEAM_BASERUN_SB TEAM_BASERUN_CS TEAM_PITCHING_H TEAM_PITCHING_HR
TEAM_PITCHING_BB TEAM_PITCHING_SO TEAM_FIELDING_E
TARGET_WINS                        NA              NA     -0.10993705       0.18901373
0.124174536               NA     -0.17648476
TEAM_BATTING_H                     NA              NA      0.30269371       0.07285312
0.094193027               NA      0.26490248
TEAM_BATTING_2B                    NA              NA      0.02369219       0.45455082
0.178054204               NA     -0.23515099
TEAM_BATTING_3B                    NA              NA      0.19487941      -0.56783668      -
0.002224148               NA      0.50977845
TEAM_BATTING_HR                    NA              NA     -0.25014548       0.96937140
0.136927564               NA     -0.58733910
TEAM_BATTING_BB                    NA              NA     -0.44977762       0.45955207
0.489361263               NA     -0.65597081
```

```
TEAM_BATTING_SO                    NA              NA              NA                NA
NA               NA                NA
TEAM_BASERUN_SB                     1              NA              NA                NA
NA               NA                NA
TEAM_BASERUN_CS                    NA               1              NA                NA
NA               NA                NA
TEAM_PITCHING_H                    NA              NA      1.00000000       -0.14161276
0.320676162               NA        0.66775901
TEAM_PITCHING_HR                   NA              NA     -0.14161276        1.00000000
0.221937505               NA       -0.49314447
TEAM_PITCHING_BB                   NA              NA      0.32067616        0.22193750
1.000000000               NA       -0.02283756
TEAM_PITCHING_SO                   NA              NA              NA                NA
NA                1                NA
TEAM_FIELDING_E                    NA              NA      0.66775901       -0.49314447      -
0.022837561               NA        1.00000000
TEAM_FIELDING_DP                   NA              NA              NA                NA
NA               NA                NA
T95_TEAM_BATTING_H                 NA              NA      0.22205950        0.08771461
0.094981312               NA        0.19020016
TRIM_TEAM_BATTING_2B               NA              NA      0.02400359        0.45728251
0.179108881               NA       -0.23687709
T99_TEAM_BATTING_3B                NA              NA      0.17662499       -0.59044252      -
0.022709975               NA        0.50709895
QUANT_TEAM_BATTING_HR              NA              NA     -0.23201869        0.92596843
0.135516634               NA       -0.57037625
TRIM_TEAM_BATTING_HR               NA              NA     -0.25070656        0.96922789
0.136921279               NA       -0.58835453
COMB_HR                            NA              NA     -0.15765942        0.95276567
0.167854011               NA       -0.45905782
STD_TEAM_BATTING_BB                NA              NA     -0.44977762        0.45955207
0.489361263               NA       -0.65597081
M_TEAM_BATTING_SO                  NA              NA     -0.04964299       -0.29793696      -
0.123936700               NA        0.05303325
IMP_TEAM_BATTING_SO                NA              NA     -0.35612304        0.68532120
0.062099422               NA       -0.58121209
                      TEAM_FIELDING_DP T95_TEAM_BATTING_H TRIM_TEAM_BATTING_2B
T99_TEAM_BATTING_3B QUANT_TEAM_BATTING_HR TRIM_TEAM_BATTING_HR
TARGET_WINS                          NA        0.368992927           0.28775816
0.14350052             0.16581834           0.176386766
TEAM_BATTING_H                       NA        0.926574545           0.56276376
0.41183181            -0.01284682          -0.007291966
TEAM_BATTING_2B                      NA        0.589293420           0.99917840          -
0.11774935             0.40906358           0.435044820
TEAM_BATTING_3B                      NA        0.397713027          -0.10959418
0.99305061            -0.62178564          -0.636162608
TEAM_BATTING_HR                      NA        0.035386692           0.43832465          -
0.64958545             0.95690015           0.999837180
TEAM_BATTING_BB                      NA       -0.004283097           0.25791958          -
0.29098337             0.49509742           0.514239690
TEAM_BATTING_SO                      NA                 NA                   NA
NA                     NA                    NA
TEAM_BASERUN_SB                      NA                 NA                   NA
NA                     NA                    NA
TEAM_BASERUN_CS                      NA                 NA                   NA
NA                     NA                    NA
TEAM_PITCHING_H                      NA        0.222059503           0.02400359
0.17662499            -0.23201869          -0.250706558
TEAM_PITCHING_HR                     NA        0.087714615           0.45728251          -
0.59044252             0.92596843           0.969227886
TEAM_PITCHING_BB                     NA        0.094981312           0.17910888          -
0.02270998             0.13551663           0.136921279
TEAM_PITCHING_SO                     NA                 NA                   NA
NA                     NA                    NA
TEAM_FIELDING_E                      NA        0.190200163          -0.23687709
0.50709895            -0.57037625          -0.588354533
```

```
TEAM_FIELDING_DP                              1                NA                NA
NA                   NA                      NA
T95_TEAM_BATTING_H                           NA       1.000000000        0.58976224
0.39506108           0.02544365             0.034506558
TRIM_TEAM_BATTING_2B                         NA       0.589762240        1.00000000          -
0.12015917           0.41202453             0.437971228
T99_TEAM_BATTING_3B                          NA       0.395061084       -0.12015917
1.00000000          -0.63572478            -0.650186027
QUANT_TEAM_BATTING_HR                        NA       0.025443654        0.41202453          -
0.63572478           1.00000000             0.958076638
TRIM_TEAM_BATTING_HR                         NA       0.034506558        0.43797123          -
0.65018603           0.95807664             1.000000000
COMB_HR                                      NA       0.071811387        0.41869859          -
0.57977083           0.88413995             0.951947890
STD_TEAM_BATTING_BB                          NA      -0.004283097        0.25791958          -
0.29098337           0.49509742             0.514239690
M_TEAM_BATTING_SO                            NA      -0.183033465       -0.26183934
0.13802599          -0.28956140            -0.285840033
IMP_TEAM_BATTING_SO                          NA      -0.373765571        0.20776525          -
0.67737628           0.71126867             0.741817378
                          COMB_HR STD_TEAM_BATTING_BB M_TEAM_BATTING_SO IMP_TEAM_BATTING_SO
TARGET_WINS            0.17966840         0.232559864       0.007730792         -0.03104274
TEAM_BATTING_H         0.03285734        -0.072464013      -0.171367201         -0.40366905
TEAM_BATTING_2B        0.41591694         0.255726103      -0.260605676          0.20714364
TEAM_BATTING_3B       -0.56624559        -0.287235841       0.131790907         -0.66987684
TEAM_BATTING_HR        0.95367299         0.513734810      -0.285259870          0.74126979
TEAM_BATTING_BB        0.43164817         1.000000000      -0.133620457          0.39323266
TEAM_BATTING_SO                NA                  NA                NA                  NA
TEAM_BASERUN_SB                NA                  NA                NA                  NA
TEAM_BASERUN_CS                NA                  NA                NA                  NA
TEAM_PITCHING_H       -0.15765942        -0.449777625      -0.049642990         -0.35612304
TEAM_PITCHING_HR       0.95276567         0.459552072      -0.297936957          0.68532120
TEAM_PITCHING_BB       0.16785401         0.489361263      -0.123936700          0.06209942
TEAM_PITCHING_SO               NA                  NA                NA                  NA
TEAM_FIELDING_E       -0.45905782        -0.655970815       0.053033254         -0.58121209
TEAM_FIELDING_DP               NA                  NA                NA                  NA
T95_TEAM_BATTING_H     0.07181139        -0.004283097      -0.183033465         -0.37376557
TRIM_TEAM_BATTING_2B   0.41869859         0.257919578      -0.261839337          0.20776525
T99_TEAM_BATTING_3B   -0.57977083        -0.290983375       0.138025990         -0.67737628
QUANT_TEAM_BATTING_HR  0.88413995         0.495097418      -0.289561397          0.71126867
TRIM_TEAM_BATTING_HR   0.95194789         0.514239690      -0.285840033          0.74181738
COMB_HR                1.00000000         0.431648171      -0.216066269          0.67222954
STD_TEAM_BATTING_BB    0.43164817         1.000000000      -0.133620457          0.39323266
M_TEAM_BATTING_SO     -0.21606627        -0.133620457       1.000000000         -0.21079871
IMP_TEAM_BATTING_SO    0.67222954         0.393232663      -0.210798706          1.00000000
```



What: Replace Missing Values with Averages and Transform a variable to a z Score

Code:

#Example Replace missing values with averages

mb$TEAM_BASERUN_SB[mb$TEAM_BASERUN_SB==NA] <- 124.7617716

#Scale and create Z Scores for sample variable;

IMP4_TEAM_BASERUN_SB <- scale(mb$TEAM_BASERUN_SB, center = TRUE, scale = TRUE)

What: Calculate a Regression Model

Code:

#calculate the a regression model

FinalModel <-
lm(TARGET_WINS~IMP_TEAM_BASERUN_SB+TEAM_BATTING_2B+TEAM_BATTING_3B+TEAM_BATTING_BB+TEAM_BATTING_H+TEAM_PITCHING_H+PRIN2+PRIN3+PRIN8+PRIN12+PRIN16+PRIN17+M_TEAM_FIELDING_DP+M_TEAM_BASERUN_SB+T99_TEAM_BATTING_3B,data=mb)

FinalModel

summary(FinalModel)

Results:

```
Call:
lm(formula = TARGET_WINS ~ IMP_TEAM_BASERUN_SB + TEAM_BATTING_2B +
    TEAM_BATTING_3B + TEAM_BATTING_BB + TEAM_BATTING_H + TEAM_PITCHING_H +
    PRIN2 + PRIN3 + PRIN8 + PRIN12 + PRIN16 + PRIN17 + M_TEAM_FIELDING_DP +
    M_TEAM_BASERUN_SB + T99_TEAM_BATTING_3B, data = mb)

Residuals:
    Min      1Q  Median      3Q     Max
-55.647  -7.851   0.167   7.644  53.228

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         35.983331   4.772702   7.539 6.79e-14 ***
IMP_TEAM_BASERUN_SB  0.062973   0.004352  14.468  < 2e-16 ***
TEAM_BATTING_2B      0.087787   0.011117   7.897 4.43e-15 ***
TEAM_BATTING_3B     -0.344544   0.080115  -4.301 1.78e-05 ***
TEAM_BATTING_BB      0.060956   0.006768   9.007  < 2e-16 ***
TEAM_BATTING_H       0.069212   0.004189  16.523  < 2e-16 ***
TEAM_PITCHING_H      0.002133   0.000332   6.423 1.62e-10 ***
PRIN2               -0.136757   0.007100 -19.261  < 2e-16 ***
PRIN3                0.065450   0.005626  11.634  < 2e-16 ***
PRIN8               -0.334908   0.028521 -11.743  < 2e-16 ***
PRIN12              -0.337940   0.023507 -14.376  < 2e-16 ***
PRIN16               0.113797   0.013480   8.442  < 2e-16 ***
PRIN17               0.029244   0.008650   3.381 0.000735 ***
M_TEAM_FIELDING_DP   7.898416   1.483474   5.324 1.11e-07 ***
M_TEAM_BASERUN_SB   38.903861   2.151836  18.079  < 2e-16 ***
```

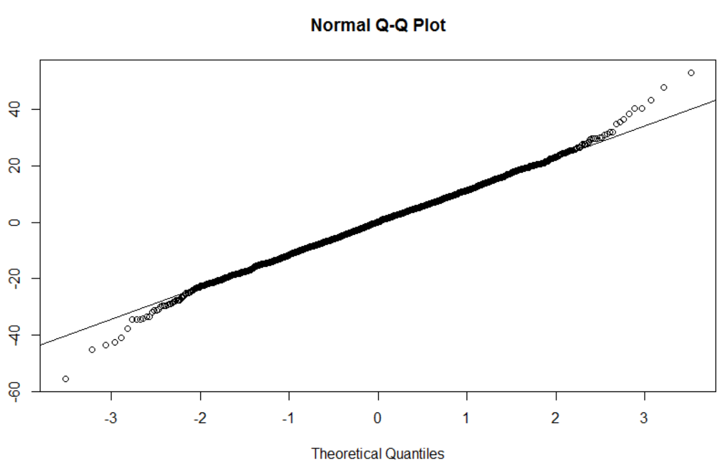
```
T99_TEAM_BATTING_3B  0.425702   0.084400   5.044 4.92e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.87 on 2260 degrees of freedom
Multiple R-squared:  0.4364,   Adjusted R-squared:  0.4326
F-statistic: 116.7 on 15 and 2260 DF,  p-value: < 2.2e-16
```

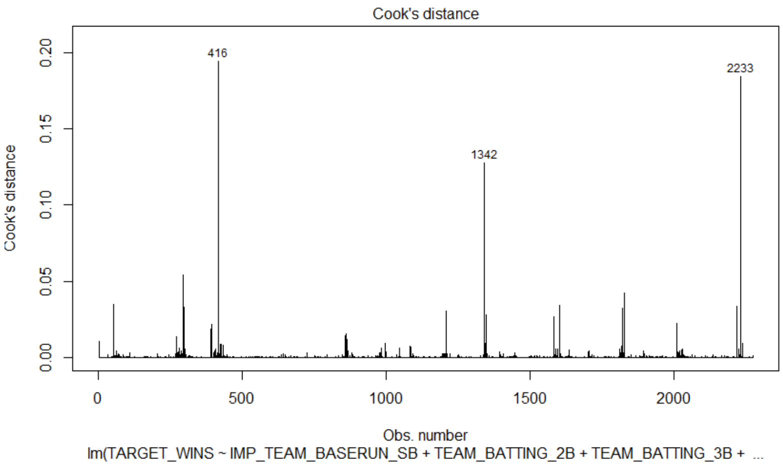
What: Generate the Diagnostic Plots

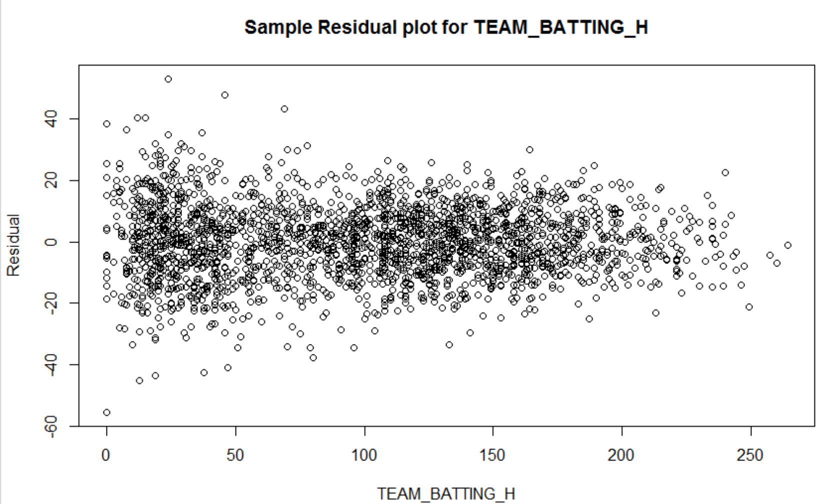
Code:

```
#qqplot

 qqnorm(resid(FinalModel))

 qqline(resid(FinalModel))

   # Cook's D plot

 # identify D values > 4/(n-k-1)

 cutoff <- 4/((nrow(mtcars)-length(FinalModel$coefficients)-2))

 plot(FinalModel, which=4, cook.levels=cutoff)

   #residualplot

 Residual = resid(FinalModel)#Fit X2 as a single predictor

 TEAM_BATTING_H <-mb$TEAM_BATTING_HR

 plot(TEAM_BATTING_H, Residual, main="Sample Residual plot for TEAM_BATTING_H")
```

Results:







What: Backward Regression

Code:

```
#Perform automatic variable selection using backward selection

  YRegrBkwd <-
lm(TARGET_WINS~COMB_HR+IMP_TEAM_BASERUN_SB+TEAM_BATTING_2B+TEAM_BATTING_3B+TEAM_BATTING_BB+TEAM_BATTING_H+TEAM_PITCHING_H+PRIN1+ PRIN2+PRIN3+PRIN4 + PRIN5 + PRIN6 + PRIN7 + PRIN8 +PRIN9 + PRIN10 + PRIN11 + PRIN12+PRIN13 + PRIN14 + PRIN15 +PRIN16+PRIN17 +PRIN18 +PRIN19 +M_TEAM_FIELDING_DP+M_TEAM_BASERUN_SB+T99_TEAM_BATTING_3B,data=mb)

 step <- stepAIC(YRegrBkwd, direction="backward")

 step$anova # display results
```

Results:

```
Initial Model:
TARGET_WINS ~ COMB_HR + IMP_TEAM_BASERUN_SB + TEAM_BATTING_2B +
    TEAM_BATTING_3B + TEAM_BATTING_BB + TEAM_BATTING_H + TEAM_PITCHING_H +
    PRIN1 + PRIN2 + PRIN3 + PRIN4 + PRIN5 + PRIN6 + PRIN7 + PRIN8 +
    PRIN9 + PRIN10 + PRIN11 + PRIN12 + PRIN13 + PRIN14 + PRIN15 +
    PRIN16 + PRIN17 + PRIN18 + PRIN19 + M_TEAM_FIELDING_DP +
    M_TEAM_BASERUN_SB + T99_TEAM_BATTING_3B

Final Model:
TARGET_WINS ~ COMB_HR + IMP_TEAM_BASERUN_SB + TEAM_BATTING_2B +
    TEAM_BATTING_3B + TEAM_BATTING_BB + TEAM_BATTING_H + TEAM_PITCHING_H +
    PRIN1 + PRIN2 + PRIN3 + PRIN7 + PRIN8 + PRIN9 + PRIN10 +
    PRIN11 + PRIN12 + PRIN13 + PRIN14 + PRIN15 + PRIN16 + PRIN17 +
    PRIN18
```

```
                   Step Df    Deviance Resid. Df Resid. Dev      AIC
1                                           2248   316097.9 11284.95
2 - T99_TEAM_BATTING_3B  0   0.0000000      2248   316097.9 11284.95
3   - M_TEAM_BASERUN_SB  1   3.3675013      2249   316101.3 11282.97
4  - M_TEAM_FIELDING_DP  1 131.2606195      2250   316232.6 11281.92
5              - PRIN19  0   0.0000000      2250   316232.6 11281.92
6               - PRIN4  1   0.6237644      2251   316233.2 11279.92
7               - PRIN5  1   0.4668065      2252   316233.7 11277.92
8               - PRIN6  1 141.6125782      2253   316375.3 11276.94
```

What: Forward Regression

Code:

#Perform automatic variable selection using forward selection on the variables in the KITCHENSINK 1 of the report

```
YRegrFwd5 <-
lm(TARGET_WINS~COMB_HR+IMP_TEAM_BASERUN_SB+TEAM_BATTING_2B+TEAM_BATTING_3B+TEAM_BATTING_BB+TEAM_BATTING_H+TEAM_PITCHING_H+PRIN1+ PRIN2+PRIN3+PRIN4 + PRIN5 + PRIN6 + PRIN7 + PRIN8 +PRIN9 + PRIN10 + PRIN11 +
PRIN12+PRIN13 + PRIN14 + PRIN15 +PRIN16+PRIN17 +PRIN18 +PRIN19
+M_TEAM_FIELDING_DP+M_TEAM_BASERUN_SB+T99_TEAM_BATTING_3B,data=mb)

step <- stepAIC(YRegrFwd5, direction="forward")

step$anova # display results

extractAIC(YRegrFwd5)
```

Results:

```
Initial Model:
TARGET_WINS ~ COMB_HR + IMP_TEAM_BASERUN_SB + TEAM_BATTING_2B +
    TEAM_BATTING_3B + TEAM_BATTING_BB + TEAM_BATTING_H + TEAM_PITCHING_H +
    PRIN1 + PRIN2 + PRIN3 + PRIN4 + PRIN5 + PRIN6 + PRIN7 + PRIN8 +
    PRIN9 + PRIN10 + PRIN11 + PRIN12 + PRIN13 + PRIN14 + PRIN15 +
    PRIN16 + PRIN17 + PRIN18 + PRIN19 + M_TEAM_FIELDING_DP +
    M_TEAM_BASERUN_SB + T99_TEAM_BATTING_3B

Final Model:
TARGET_WINS ~ COMB_HR + IMP_TEAM_BASERUN_SB + TEAM_BATTING_2B +
    TEAM_BATTING_3B + TEAM_BATTING_BB + TEAM_BATTING_H + TEAM_PITCHING_H +
    PRIN1 + PRIN2 + PRIN3 + PRIN4 + PRIN5 + PRIN6 + PRIN7 + PRIN8 +
    PRIN9 + PRIN10 + PRIN11 + PRIN12 + PRIN13 + PRIN14 + PRIN15 +
```

```
    PRIN16 + PRIN17 + PRIN18 + PRIN19 + M_TEAM_FIELDING_DP + 
    M_TEAM_BASERUN_SB + T99_TEAM_BATTING_3B


  Step Df Deviance Resid. Df Resid. Dev      AIC
1                     2248   316097.9 11284.95
>     extractAIC(YRegrFwd5)
[1]    28.00 11284.95
```

What: Stepwise Regression

Code:

#Perform automatic variable selection using stepwise selection

YRegrStep <-
lm(TARGET_WINS~COMB_HR+IMP_TEAM_BASERUN_SB+TEAM_BATTING_2B+TEAM_BATTING_3B+TEAM_BATTING_BB+TEAM_BATTING_H+TEAM_PITCHING_H+PRIN1+ PRIN2+PRIN3+PRIN4 + PRIN5 + PRIN6 + PRIN7 + PRIN8 +PRIN9 + PRIN10 + PRIN11 + PRIN12+PRIN13 + PRIN14 + PRIN15 +PRIN16+PRIN17 +PRIN18 +PRIN19 +M_TEAM_FIELDING_DP+M_TEAM_BASERUN_SB+T99_TEAM_BATTING_3B,data=mb)

step <- stepAIC(YRegrStep, direction="both")

step$anova # display results

Results:

```
Stepwise Model Path 
Analysis of Deviance Table

Initial Model:
TARGET_WINS ~ COMB_HR + IMP_TEAM_BASERUN_SB + TEAM_BATTING_2B + 
    TEAM_BATTING_3B + TEAM_BATTING_BB + TEAM_BATTING_H + TEAM_PITCHING_H + 
    PRIN1 + PRIN2 + PRIN3 + PRIN4 + PRIN5 + PRIN6 + PRIN7 + PRIN8 + 
    PRIN9 + PRIN10 + PRIN11 + PRIN12 + PRIN13 + PRIN14 + PRIN15 + 
    PRIN16 + PRIN17 + PRIN18 + PRIN19 + M_TEAM_FIELDING_DP + 
    M_TEAM_BASERUN_SB + T99_TEAM_BATTING_3B

Final Model:
TARGET_WINS ~ COMB_HR + IMP_TEAM_BASERUN_SB + TEAM_BATTING_2B + 
    TEAM_BATTING_3B + TEAM_BATTING_BB + TEAM_BATTING_H + TEAM_PITCHING_H + 
    PRIN1 + PRIN2 + PRIN3 + PRIN7 + PRIN8 + PRIN9 + PRIN10 + 
    PRIN11 + PRIN12 + PRIN13 + PRIN14 + PRIN15 + PRIN16 + PRIN17 + 
    PRIN18


                   Step Df    Deviance Resid. Df Resid. Dev      AIC
1                                           2248   316097.9 11284.95
2 - T99_TEAM_BATTING_3B  0   0.0000000      2248   316097.9 11284.95
3   - M_TEAM_BASERUN_SB  1   3.3675013      2249   316101.3 11282.97
4  - M_TEAM_FIELDING_DP  1 131.2606195      2250   316232.6 11281.92
5              - PRIN19  0   0.0000000      2250   316232.6 11281.92
6               - PRIN4  1   0.6237644      2251   316233.2 11279.92
7               - PRIN5  1   0.4668065      2252   316233.7 11277.92
8               - PRIN6  1 141.6125782      2253   316375.3 11276.94
```