Sandra Duenas - Homework #1 - Moneyball

PREDICT 411 – Section 58 – Winter 2015

Document Navigation Note: **on the left pane, the Bookmarks are available to quickly view the organization of this document as well as to easily navigate to different sections.**

Note: the actual analysis is from page 4 through 31 (28 pages); the rest of the pages contain Bingo work, Appendices, and References.

BINGO BONUS:

If you want Bingo Bonus Points, write a brief section at the top of your Write Up document and tell me exactly what you did and how many points you are attempting.

I completed Bingo points for a total of **40 points**, see yellow highlighted points below.

1. **(20 Points)** Once you select a champion model in Step 4, use PROC GLM and PROC GENMOD to do the OLS Regression. Are the results the same? Are there any differences? Refer to **Bingo Bonus – PROC GLM and PROC**.

2. (20 Points) Use decision tree software such as Angoss or Weka or something else for variable selection or missing value imputation (the more use you make of decision trees, the more points you will receive). Be sure to carefully present your decision tree output so that I can see what you did. Did not do.

3. (20 Points) Recreate as much of the program as you can in "R" Did not do.

4. **(10 Points)** Use SAS Macros or use, in my opinion, good programming technique. *Completed, please refer to section* **Appendix E – SAS Code for EDA Visualization to Detect Outliers**

5. **(10 Points)** Hand in your SCORED FILE as a SAS DATA SET and save me to trouble of converting it. Complied.

6. (?? Points) Roll the dice … think of something creative and run with it. I might give you points. I performed EDA using Simple Regression and also using PCA, but not sure if this would be considered extra ways of ensuring the model is correct.

PENALTY BOX

1. (Lose 10 Points) If you don't have PDF format I have pdf
2. (Lose 10 Points) If you don't have a GOOD Introduction I think this is good
3. (Lose 10 Points) If you don't have a GOOD Conclusion I think this is good
4. (Lose 10 Points) If you don't put your NAME in the file names of any files you hand in this is done
5. (Lose 10 Points) If you don't put your NAME inside of the files you hand in this is done
6. (Lose ?? Points) For anything that I think might annoy your boss ! *not intentionally …. ☺*

Table of Contents

INTRODUCTION:

The goal of the analysis presented in this paper is to explain the process and techniques used for the estimation of a predictive model that accurately predicts the number of victories that a baseball team will have in a regular season.

The analysis uses baseball team data from 1900-1950. Each observation represents one game for a given team. The data set contains 2,276 observations and 17 variables of which one is the target or dependent variable, TargetWins, and 16 variables which are continuous numeric measures of the different statistics of the game.

There is only one numeric continuous variable to be predicted therefore the OLS Regression model estimation technique is used to design the model.

The first step in the process to creating the predictive model is the data exploration step in which simple statistics techniques, such as means, media, percentiles, histograms, and boxplots are used to identify variables with missing values and variables with Outliers.

The next step in the process is the transformation of the variables. The variables with missing values are imputed using its Mean. Imputation or removal of missing values is a requirement for the OLS Regression technique. The variables with outliers are transformed by using either Log10 or Standardization and Trimming mathematical transformations of the values. Both transformations are preceded by capping the values in the outlier variables to either Percentile 1 or Percentile 99, depending on the extreme value.

Interaction variables are created based on the imputed in order to boost the model. Two interaction variables are designed to be included in the process and they are discussed below.

A Simple Regression EDA is performed on each imputed, transformed, as well as variables that did not need transformation in order to assess the strength and direction of the linearity assumption between each predictor and the target variable. From this analysis, the transformed variables based on Log10 or Standardized can be selected prior to creating a final list of variables to put through the OLS Regression Selection process for Stepwise, Forward, and Backward.

The variable selection techniques of Stepwise, Forward, and Backward are used to select the best model in terms of the highest Adjusted R^2, lowest AIC measures, and collinearity with VIF less than 10. Any manual fine tuning to the model is done, such as removing variables with incorrect sign in the coefficient and adding Flag variables left out for included Imputed variables or removing Flag variables when their imputed variable was removed from the final model.

The OLS Regression Assumptions are validated via the Fit Diagnostic Plots to ensure that the model can be as accurate as possible; however, these assumptions can be violated and still produce accurate results.

Using the estimated best model, a Scoring program is created to be used with the Test data set and produce the predictions. The Scoring program implements the exact Imputation and Transformation data preparation techniques as were done in the program that estimated the best model using the Train data set, in that way it ensures that the Test data has the same data preparation as the train data did when used to create the Scoring model.

The evaluation of the Error metric, (TargetWins_actual – TargetWins_estimate), on the scoring of the Train data set should be as close to 0 as possible indicating that the prediction or scoring by the estimated model is accurate.

1. Data Exploration (40 Points)

1.1 Exploring the Structure of the Data Set and the Data

Based on the PROC CONTENTS result, there are 2,276 observations and 17 variables.

Figure 1

The CONTENTS Procedure

Data Set Name	MYDATA.MONEYBALL	Observations	2276
Member Type	DATA	Variables	17
Engine	V9	Indexes	0

The 17 variables are shown in the Figure 2 below. The variable INDEX is the Team ID and it will be removed when a new data set, called *moneyball_train,* is created to be used in the creation of the regression models.

All of the variables are continuous numeric representing counts of measures that describe baseball games and may affect Winning scores negatively or positively.

Figure 2

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Label
1	INDEX	Num	8	
2	TARGET_WINS	Num	8	
10	TEAM_BASERUN_CS	Num	8	Caught stealing
9	TEAM_BASERUN_SB	Num	8	Stolen bases
4	TEAM_BATTING_2B	Num	8	Doubles by batters
5	TEAM_BATTING_3B	Num	8	Triples by batters
7	TEAM_BATTING_BB	Num	8	Walks by batters
3	TEAM_BATTING_H	Num	8	Base Hits by batters
11	TEAM_BATTING_HBP	Num	8	Batters hit by pitch
6	TEAM_BATTING_HR	Num	8	Homeruns by batters
8	TEAM_BATTING_SO	Num	8	Strikeouts by batters
17	TEAM_FIELDING_DP	Num	8	Double Plays
16	TEAM_FIELDING_E	Num	8	Errors
14	TEAM_PITCHING_BB	Num	8	Walks allowed
12	TEAM_PITCHING_H	Num	8	Hits allowed
13	TEAM_PITCHING_HR	Num	8	Homeruns allowed
15	TEAM_PITCHING_SO	Num	8	Strikeouts by pitchers

The first 10 records of the RAW data are shown in Figure 3 below so that a visual inspection of the data can be performed. It can be noticed that some variables have Missing values or a period, ".", such as TEAM_BATTING_HBP. A closer inspection of the variables with Missing values is done next.

Figure 3

Raw Data of the first 10 records

Obs	INDEX	TARGET_WINS	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB	TEAM_BATTING_SO	TEAM_BASERUN_SB	TEAM_BASERUN_CS	TEAM_BATTING_HBP	TEAM_PITCHING_H	TEAM_PITCHING_HR	TEAM_PITCHING_BB	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
1	1	39	1445	194	39	13	143	842	.	.	.	9364	84	927	5456	1011	.
2	2	70	1339	219	22	190	685	1075	37	28	.	1347	191	689	1082	193	155
3	3	86	1377	232	35	137	602	917	46	27	.	1377	137	602	917	175	153
4	4	70	1387	209	38	96	451	922	43	30	.	1396	97	454	928	164	156
5	5	82	1297	186	27	102	472	920	49	39	.	1297	102	472	920	138	168
6	6	75	1279	200	36	92	443	973	107	59	.	1279	92	443	973	123	149
7	7	80	1244	179	54	122	525	1062	80	54	.	1244	122	525	1062	136	186
8	8	85	1273	171	37	115	456	1027	40	36	.	1281	116	459	1033	112	136
9	11	86	1391	197	40	114	447	922	69	27	.	1391	114	447	922	127	169
10	12	76	1271	213	18	96	441	827	72	34	.	1271	96	441	827	131	159

1.2 Exploring for Missing Values

In examining the results from the PROC MEANS for all the variables since they are all continuous numeric variables, we can see that there are 6 variables with Missing values, refer to Figure 4 below, column “N Miss” and “% Miss”. These variables are StrikeOutByBatters_N, StolenBases_P, CaughtStealing_N, BattersHitByPitch_P, StrikeoutsByPitchers_P, and DoublePlays_P.

The explanation of how these 6 variables have their Missing values imputed is explained in the section **Fix missing values**.

Figure 4 - PROC MEANS result matrix of the Raw Data. ***Note*** *that variables have been renamed. The suffix indicates whether the variable has positive impact on the Winning score, _P, or a negative impact, _N.*

Variable	Minimum	25th Pctl	50th Pctl	75th Pctl	Maximum	Sum	Mean	Median	Mode	Std Dev	N Miss	% Miss	Skewness	Kurtosis
TargetWins	-	71	82	92	146	183,880	80.79	82	83	15.75	-	-	(0.40)	1.04
BaseHitsByBattersAllBases_P	891	1,383	1,454	1,538	2,554	3,344,058	1,469.27	1,454	1,458	144.59	-	-	1.57	7.31
DoublesByBatters2Bases_P	69	208	238	273	458	549,078	241.25	238	227	46.80	-	-	0.22	0.01
TriplesByBatters3Bases_P	-	34	47	72	223	125,749	55.25	47	35	27.94	-	-	1.11	1.51
HomerunsByBatters4Bases_P	-	42	102	147	264	226,717	99.61	102	21	60.55	-	-	0.19	(0.96)
WalksByBatters_P	-	451	512	580	878	1,141,548	501.56	512	502	122.67	-	-	(1.03)	2.19
StrikeoutsByBatters_N	-	548	750	930	1,399	1,599,206	735.61	750	-	248.53	102	0.04	(0.30)	(0.32)
StolenBases_P	-	66	101	156	697	267,614	124.76	101	65	87.79	131	0.06	1.98	5.51
CaughtStealing_N	-	38	49	62	201	79,417	52.80	49	52	22.96	772	0.34	1.98	7.66
BattersHitByPitch_P	29	50	58	67	95	11,337	59.36	58	54	12.97	2,085	0.92	0.32	(0.05)
HitsAllowed_N	1,137	1,419	1,518	1,683	30,132	4,049,483	1,779.21	1,518	1,494	1,406.84	-	-	10.34	142.28
HomerunsAllowed_N	-	50	107	150	343	240,570	105.70	107	114	61.30	-	-	0.29	(0.60)
WalksAllowed_N	-	476	537	611	3,645	1,258,646	553.01	537	536	166.36	-	-	6.75	97.27
StrikeoutsByPitchers_P	-	615	814	968	19,278	1,777,746	817.73	814	-	553.09	102	0.04	22.21	673.36
Errors_N	65	127	159	250	1,898	560,990	246.48	159	122	227.77	-	-	2.99	11.01
DoublePlays_P	52	131	149	164	228	291,312	146.39	149	148	26.23	286	0.13	(0.39)	0.19

The result of imputation for missing values is evaluated in Figure 7 below. New fields were added to the moneyball_train data set to store the imputed values and their corresponding Flag indicators. The top area of the table in Figure 7 with white background, are the results prior to imputation. The bottom area of the table, with light blue background are the results after imputation.

It can be seen that the Mean for those 6 variables did no change prior or post imputation, even for BattersHitByPitch which had 92% of its values missing. However, it can be seen in column N that all 2,276 records are accounted for and that there are no missing values in the N Miss column for the imputed columns.

The 6 imputed variables will be used for the regression model rather than their original variables.

Figure 7 – Comparing PROC MEANS Before and After Imputation of Missing values

Variable	N	Minimum	25th Pctl	50th Pctl	75th Pctl	Maximum	Sum	Mean	Median	Mode	Std Dev	N Miss	Skewness	Kurtosis
TargetWins	2276	0	71	82	92	146	183880	80.79	82.00	83.00	15.75	0	-0.40	1.04
BaseHitsByBattersAllBases_P	2276	891	1383	1454	1538	2554	3344058	1469.27	1454.00	1458.00	144.59	0	1.57	7.31
DoublesByBatters2Bases_P	2276	69	208	238	273	458	549078	241.25	238.00	227.00	46.80	0	0.22	0.01
TriplesByBatters3Bases_P	2276	0	34	47	72	223	125749	55.25	47.00	35.00	27.94	0	1.11	1.51
HomerunsByBatters4Bases_P	2276	0	42	102	147	264	226717	99.61	102.00	21.00	60.55	0	0.19	-0.96
WalksByBatters_P	2276	0	451	512	580	878	1141548	501.56	512.00	502.00	122.67	0	-1.03	2.19
StrikeoutsByBatters_N	2174	0	548	750	930	1399	1599206	735.61	750.00	0.00	248.53	102	-0.30	-0.32
StolenBases_P	2145	0	66	101	156	697	267614	124.76	101.00	65.00	87.79	131	1.98	5.51
CaughtStealing_N	1504	0	38	49	62	201	79417	52.80	49.00	52.00	22.96	772	1.98	7.66
BattersHitByPitch_P	191	29	50	58	67	95	11337	59.36	58.00	54.00	12.97	2085	0.32	-0.05
HitsAllowed_N	2276	1137	1419	1518	1683	30132	4049483	1779.21	1518.00	1494.00	1406.84	0	10.34	142.28
HomerunsAllowed_N	2276	0	50	107	150	343	240570	105.70	107.00	114.00	61.30	0	0.29	-0.60
WalksAllowed_N	2276	0	476	537	611	3645	1258646	553.01	536.50	536.00	166.36	0	6.75	97.27
StrikeoutsByPitchers_P	2174	0	615	814	968	19278	1777746	817.73	813.50	0.00	553.09	102	22.21	673.36
Errors_N	2276	65	127	159	250	1898	560990	246.48	159.00	122.00	227.77	0	2.99	11.01
DoublePlays_P	1990	52	131	149	164	228	291312	146.39	149.00	148.00	26.23	286	-0.39	0.19
IMP_StrikeoutsByBatters_N	2276	0	557	736	925	1399	1674238	735.61	735.61	735.61	242.89	0	-0.31	-0.19
MFlag_StrikeoutsByBatters_N	2276	0	0	0	0	1	102	0.04	0.00	0.00	0.21	0	4.40	17.40
IMP_StolenBases_P	2276	0	67	106	151	697	283958	124.76	106.00	124.76	85.23	0	2.03	6.03
MFlag_StolenBases_P	2276	0	0	0	0	1	131	0.06	0.00	0.00	0.23	0	3.80	12.47
IMP_CaughtStealing_N	2276	0	44	53	55	201	120182	52.80	52.80	52.80	18.66	0	2.44	13.12
MFlag_CaughtStealing_N	2276	0	0	0	1	1	772	0.34	0.00	0.00	0.47	0	0.68	-1.54
IMP_BattersHitByPitch_P	2276	29	59	59	59	95	135094	59.36	59.36	59.36	3.75	0	1.11	31.85
MFlag_BattersHitByPitch_P	2276	0	1	1	1	1	2085	0.92	1.00	1.00	0.28	0	-3.00	7.03
IMP_StrikeoutsByPitchers_P	2276	0	626	818	957	19278	1861155	817.73	817.73	817.73	540.54	0	22.72	705.02
MFlag_StrikeoutsByPitchers_P	2276	0	0	0	0	1	102	0.04	0.00	0.00	0.21	0	4.40	17.40
IMP_DoublePlays_P	2276	52	134	146	162	228	333179	146.39	146.39	146.39	24.52	0	-0.42	0.65
MFlag_DoublePlays_P	2276	0	0	0	0	1	286	0.13	0.00	0.00	0.33	0	2.26	3.11

1.3 Exploring for Outliers

The first step to exploring outliers is to run the PROC MEANS on the imputed data set and evaluate the difference between the Median and the Mean as well as the 1 percentile, 5 percentile, 95 percentile, and 99 percentile. Figure 8 below shows these results.

If the Mean is greater than the Median, it indicates that the Outliers are right tailed or that there are more observations with higher values than there are with lower or average values. Likewise on the reverse, when the Mean is lower than the Median, it indicates that there are more observations with lower values than there are with higher or average values.

Based on Figure 8 below, the variables for HitsAllowed, Errors, IMP_StolenBases have much higher Mean values than their Median values, respectively, so these variables have more observations with higher values, right tailed, than with Median values.

Also the 99th Percentile for these 3 variables is much larger than the Median at 7093 for the 99th percentile vs. 1518 for the Median for the *HitsAllowed*, at 1237 for the 99th percentile vs. 159 for the Median for the *Errors*, and at 438 for the 99th percentile vs. 106 for Median for the *IMP_StolenBases*.

Figure 8 – PROC MEANS on the Imputed train data set

Variable	N	Minimum	Maximum	1st Pctl	5th Pctl	50th Pctl	95th Pctl	99th Pctl	Sum	Median	Mean	Mode	Std Dev	N Miss
TargetWins	2276	0	146	38	54	82	104	114	183880	82	81	83	16	0
BaseHitsByBattersAllBases_P	2276	891	2554	1188	1280	1454	1696	1950	3344058	1454	1469	1458	145	0
DoublesByBatters2Bases_P	2276	69	458	141	167	238	320	352	549078	238	241	227	47	0
TriplesByBatters3Bases_P	2276	0	223	17	23	47	108	134	125749	47	55	35	28	0
HomerunsByBatters4Bases_P	2276	0	264	4	14	102	199	235	226717	102	100	21	61	0
WalksByBatters_P	2276	0	878	79	246	512	671	755	1141548	512	502	502	123	0
HitsAllowed_N	2276	1137	30132	1244	1316	1518	2563	7093	4049483	1518	1779	1494	1407	0
HomerunsAllowed_N	2276	0	343	8	18	107	210	244	240570	107	106	114	61	0
WalksAllowed_N	2276	0	3645	237	377	537	757	924	1258646	537	553	536	166	0
Errors_N	2276	65	1898	86	100	159	716	1237	560990	159	246	122	228	0
IMP_StrikeoutsByBatters_N	2276	0	1399	72	363	736	1099	1192	1674238	736	736	736	243	0
IMP_StolenBases_P	2276	0	697	24	36	106	298	438	283958	106	125	125	85	0
IMP_CaughtStealing_N	2276	0	201	18	27	53	83	125	120182	53	53	53	19	0
IMP_BattersHitByPitch_P	2276	29	95	45	59	59	59	75	135094	59	59	59	4	0
IMP_StrikeoutsByPitchers_P	2276	0	19278	208	423	818	1169	1464	1861155	818	818	818	541	0
IMP_DoublePlays_P	2276	52	228	80	100	146	184	202	333179	146	146	146	25	0

An additional outlier exploratory approach is to run a Histogram and a Boxplot on each of the variables.

The SAS Code that creates the Histogram and Boxplot for each variable is show in **Appendix E – SAS Code for EDA Visualization to Detect Outliers.**

The Outlier analysis is shown below in Figure 9 below is based on the results from the EDA_OUTLIER macro above which detail results are shown in **Appendix B – Histogram and Box Plot of Imputed Data Prior to Transformation.**

Figure 9 – Outlier Analysis for each Variable

Variable #	Variable Name	Analysis	Transform (Yes/No)
1	TargetWins	The histogram shows the Normal curve and the Density curve very close to each other which indicates that outliers are not influencing the Mean. However, the Normal curve is slightly more to the left or left tailed indicating a slight overweight of observations with lower TargetWins. The Boxplot shows much more data points to the left of the Minimum wisker confirming the left tail of the Normal curve in the Histogram; however, the Mean is very close to the Median indicating that concerns about Outliers influence should not be very strong and thus the variable **would not need to be transformed.** The boxplot also shows that there are slightly more teams with lower values of TargetWins because the area of the boxplot between the 25th percentile and the 50th percentile is slightly larger than the area above the 50th percentile; however, they seem to be more evenly distributed than not.	No
2	BaseHitsByBattersAllBases	The histogram shows the Normal curve and the Density curve very close to each other which indicates that outliers are not influencing the Mean. However, the Normal curve is slightly more to the right or right tailed indicating a slight overweight of observations with higher BaseHitsByBattersAllBases. The Boxplot shows much more data points to the right of the Maximum wisker confirming the right tail of the Normal curve in the Histogram; however, the Mean is very close to the Median indicating that concerns about Outliers influence should not be very strong and thus the variable **would not need to be transformed.** The boxplot also shows that there are slightly more teams with higher values of DoublesByBatters2Bases because the area of the boxplot between the 50th percentile and the 75th percentile is slightly larger than the area below the 50th percentile; however, they seem to be more evenly distributed than not.	No
3	DoublesByBatters2Bases	The histogram shows the Normal curve and the Density curve very close to each other which indicates that outliers are not influencing the Mean. However, the Normal curve is slightly more to the right or right tailed indicating a slight overweight of observations with higher DoublesByBatters2Bases. The Boxplot shows much more data points to the right of the Maximum wisker confirming the right tail of the Normal curve in the Histogram; however, the Mean is very close to the Median indicating that concerns about Outliers influence should not be very strong and thus the variable **would not need to be transformed.** The boxplot also shows that there are slightly more teams with higher values of DoublesByBatters2Bases because the area of the boxplot between the 50th percentile and the 75th percentile is slightly larger than the area below the 50th percentile; however, they seem to be more evenly distributed than not.	No
4	TriplesByBatters3Bases	The histogram shows the Normal curve and the Density curve with different peak areas which indicates that outliers are indeed influencing the Mean. The Normal curve is significantly more to the right or right tailed indicating a large overweight of observations with higher TriplesByBatters3Bases. The Boxplot shows much more data points to the right of the Maximum wisker as well as the box area between the 50th percentile and the 75th percentile begin bigger than the area of the box to the left of the 50th percentile and thus confirming the right tail of the Normal curve in the Histogram. The Mean is considerable far or higher than the Median indicating that there may be concerns about Outliers influence and thus the variable **may be transformed to reduce the influence of outliers.** The boxplot also shows that there are more teams with higher values of TriplesByBatters3Bases because the area of the boxplot between the 50th percentile and the 75th percentile is much larger than the area below the 50th percentile.	Yes

Variable #	Variable Name	Analysis	Transform (Yes/No)
5	HomerunsByBatters4Bases	The histogram shows the Normal curve with one peak and the Density curve with two peaks indicating a bimodal distribution. There is no indication of outliers influencing the Mean as the both curves tail off evenly on both sides indicating a normal distribution for HomerunsByBatters4Bases. The Boxplot shows no data points to either the right of the Maximum wisker or to the left of the Minimum wisker thus confirming that there are no outliers for this variable. The Mean is almost on the Median indicating that there are no concerns about Outliers influence and thus the variable **does not need to be transformed.** The boxplot also shows that there are more teams with lower values of HomerunsByBatters4Bases because the area of the boxplot between the 25th percentile and the 50th percentile is much larger than the area above the 50th percentile.	No
6	WalksByBatters	The histogram shows the Normal curve and the Density curve overlaying each other except that the Density curve has a higher peak indicating a normal distribution with outliers. Both curves are significantly more to the left or left tailed indicating a large overweight of observations or teams with lower WalksByBatters. The Boxplot shows much more data points to the left of the Minimum wisker confirming the left tail of the Normal and Density curves in the Histogram. The Mean is not too far from the Median indicating that there could be concerns about Outliers and so a closer look at whether this variable **needs to be transformed.** The boxplot also shows that the outliers are not influencing the normal distribution of the data beacuase the areas between the 25th percentile and the 50th percentile is about the same as the area between the 50th percentile and the 75th percentile. This may be another indication that the number of teams with WalksByBatters having low outlier values is not significant to affect the predictive strength of this variable as is.	Yes
7	IMP_BattersHitByPitch	This is an imputed variable which originally had 92% of its observations with Missing values. The imputation was done with the Mean of 59.36. The histogram shows the Normal and density curves with one peak; however there are many outliers. Both curves are significantly more to the left or left tailed indicating and to the right or right tail indicating a large overweight of observations or teams with lower and higher IMP_BattersHitByPitch values. The Boxplot shows no area between Minimum and Maximum wiskers but rather most data points show as outliers on both sides thus confirming the left tail and right tails of the Normal and Density curves in the Histogram indicating strong concerns about outliers. The Mean and the Median are the same due to the imputation of 92% of the observations. This variable **can be transformed to see if the influence of outliers are reduced,** but if not it will need to be thrown out altogether.	Yes
8	IMP_StrikeoutsByBatters	This is an imputed variable which originally had 4% of its observations with Missing values. The imputation was done with the Mean of 735.61. The histogram shows the Normal and Density curves with one peak and evenly distributed; however, the imputed 4% of the observations do show in the histogram as a steep peak. The Boxplot shows one outlier below or to the left of the Minimum wisker but overall the area between the 25th and 50th percentile is the same as the area between the 50th and 75th percentile and the Mean and Median are the same indicating no influence by outliers and thus **no need to tranform this variable.**	No

Variable #	Variable Name	Analysis	Transform (Yes/No)
9	IMP_StolenBases	This is an imputed variable which originally had 6% of its observations with Missing values. The imputation was done with the Mean of 124.76. The histogram shows the Normal and density curves with one peak; however the Density curve more clearly shows the right tail of the distribution indicating Outliers with higher values than the Median for the IMP_StolenBases variable. The Boxplot shows a lot of outliers to the right of the Maximum wisker as well as the Mean to the right of the Median indicating that Outliers are infuencing this variable and so transformation for IMP_StolenBases **needs to be performed.**	Yes
10	IMP_CaughtStealing	This is an imputed variable which originally had 34% of its observations with Missing values. The imputation was done with the Mean of 52.80. The histogram shows the Normal and density curves with one peak; however the Normal curve more clearly shows the right tail of the distribution indicating Outliers with higher values than the Median for the IMP_CaughtStealing variable. The Boxplot shows a lot of outliers to the right of the Maximum wisker as well as to the left of the Minimum wisker. The area between the 25th and 50th percentile is significantly larger than the area between the 50th and 75th percentile indicating an erratic influence of Outliers and imputed values. **This variable will need transformation.**	Yes
11	Errors	The histogram shows both the Normal curve and the Density curve with one peak indicating a normal distribution. However, the two curves do not cover the same area indicating Outliers influence. The Normal curve is right tailed indicating that there are more teams with higher values for the Errors than the Average team. The Boxplot shows a lot of data points to the right of the Maximum wisker thus confirming that there the outliers for this variable have higher values than the Median. The Mean is very far from the Median, located on the 75th percentile marker and thus indicating strong concerns of influence by Outliers and so **this variable needs to be transformed.** The boxplot also shows that there are more teams with higher values of Errors because the area of the boxplot between the 50th and the 75th percentile is much larger than the area below the 50th percentile.	Yes
12	IMP_DoublePlays	This is an imputed variable which originally had 13% of its observations with Missing values. The imputation was done with the Mean of 146.39. The histogram shows the Normal and density curves with one peak; however the Normal curve more clearly shows the right tail and left tail of the distribution indicating Outliers with higher and also with lower values than the Median for the IMP_DoublePlays variable. The Boxplot shows a lot of outliers to the right of the Maximum wisker and to the left of the Minimum wisker. However the Median and the Mean are the same indicating no influence by the Outliers. The area between the 25th and 50th percentile is similar in size as the area between the 50th and 75th percentile indicating a narrow normal distribution. **This variable may not need to be transformed despite the outliers, but an evaluation will have to be done.**	Yes

Variable #	Variable Name	Analysis	Transform (Yes/No)
13	WalksAllowed	The histogram shows the Normal curve and the Density curve with different peak areas which indicates the existence of Outliers. The Normal curve is significantly more to the right or right tailed indicating a large overweight of observations with higher WalksAllowed . The Boxplot shows many more data points to the right of the Maximum wisker however the box area between the 25th and 50th percentiles as well as the area between the 50th and the 75th percentile are the same and the Mean and Median are the same confirming Outliers may not be influencing this variable so **no transformation may be needed.**	No
14	HitsAllowed	The histogram shows the Normal curve and the Density curve with different peak areas which indicates the existence of outliers. The Normal curve is significantly more to the right or right tailed indicating a large overweight of observations with higher WalksAllowed . The Boxplot shows many more data points to the right of the Maximum wisker however the box area between the 25th and 57th percentiles is extremely narrow confirming Outliers may be influencing this variable so **transformation may be needed.**	Yes
15	HomerunsAllowed	The histogram shows the Normal curve with one peak and the Density curve with two peaks indicating a bimodal distribution. There is no indication of outliers influencing the Mean as the both curves tail off evenly on both sides indicating a normal distribution for HomerunsAllowed. The Boxplot shows a few data points to the right of the Maximum wisker thus confirming that there are a few outliers for this variable. The Mean is the same as the Median indicating that there are no concerns about Outliers influence and thus **the variable does not need to be transformed.** The boxplot also shows that there is slightly more teams with lower values of HomerunsAllowed because the area of the boxplot between the 25th percentile and the 50th percentile is slightly larger than the area betweem the 50th and the 75hth percentiles.	No
16	IMP_StrikeOutsByPitchers	This is an imputed variable which originally had 4% of its observations with Missing values. The imputation was done with the Mean of 817.73. The histogram shows the Normal and density curves with one peak; however the Normal curve more clearly shows the right tail of the distribution indicating Outliers with higher values than the Median for the IMP_StrikeOutsByPitchers variable. The Boxplot shows a lot of outliers to the right of the Maximum wisker and only one outlier to the left of the Minimum wisker. The area between the 25th and 50th percentile is slightly larger than the area between the 50th and 75th percentile and the Mean and Median seem to be very close indicating that influence by Outliers is not very significant but this variable **can be transform just to make sure.**	Yes

From the Outlier Analysis just above in Figure 9, we identify nine (9) variables out of 16 that need to be transformed in order to reduce the impact of the Outliers.

The nine (9) variables with Outlier data to be transformed are:

1. TriplesByBatters3Bases_P
2. WalksByBatters_P
3. IMP_BattersHitByPitch_P
4. IMP_StolenBases_P
5. IMP_CaughtStealing_N
6. Errors_N
7. IMP_DoublePlays_P
8. HitsAllowed_N
9. IMP_StrikeoutsByPitchers_P

Refer to section **Transforming Variables with Outlier values** for details on the Outlier Transformation of these 9 variables.

The results of the Outlier transformations are shown Figures 10 through Figure 18 below which display the comparison of the Histogram and Boxplot for each of the nine (9) variables using

1) the data after Imputation but prior to Transformation (left graph),

2) data after Transformation using Cap and Log10 (middle graph), and finally

3) the data after Transformation using Cap and Standardization with Trimming (right graph).

After the display of these nine (9) Figures, a matrix of all the variables with recommendations by the Analyst/Student as to which transformed variable (Log10 vs. Standardized) should be used is presented.

Figure 10 - IMP_StrikeoutsByPitchers_P Outlier Transformation EDA

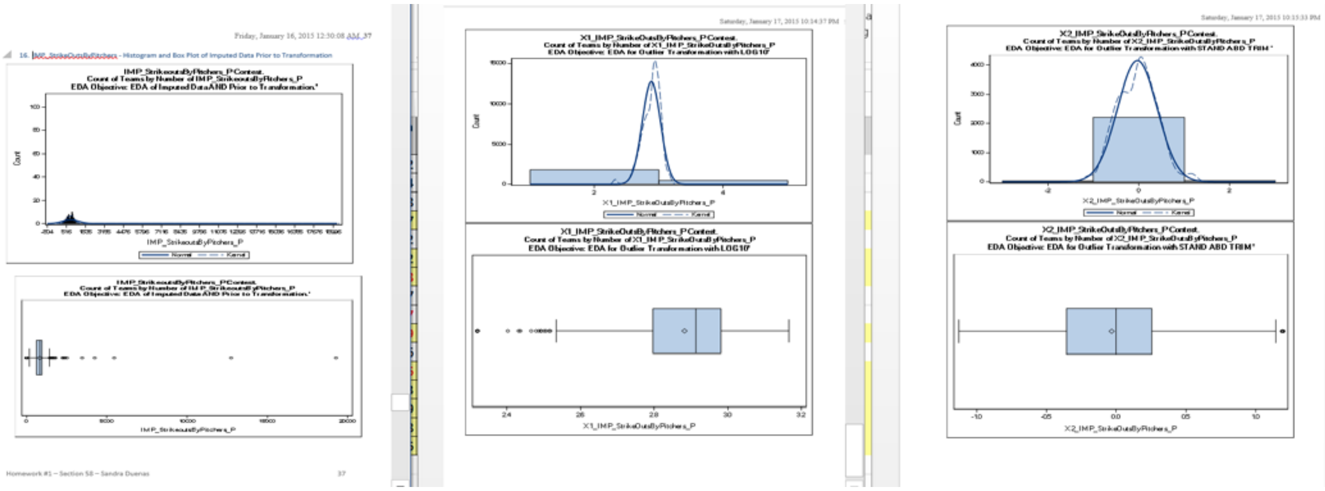


Figure 11 - HitsAllowed_N Outlier Transformation EDA

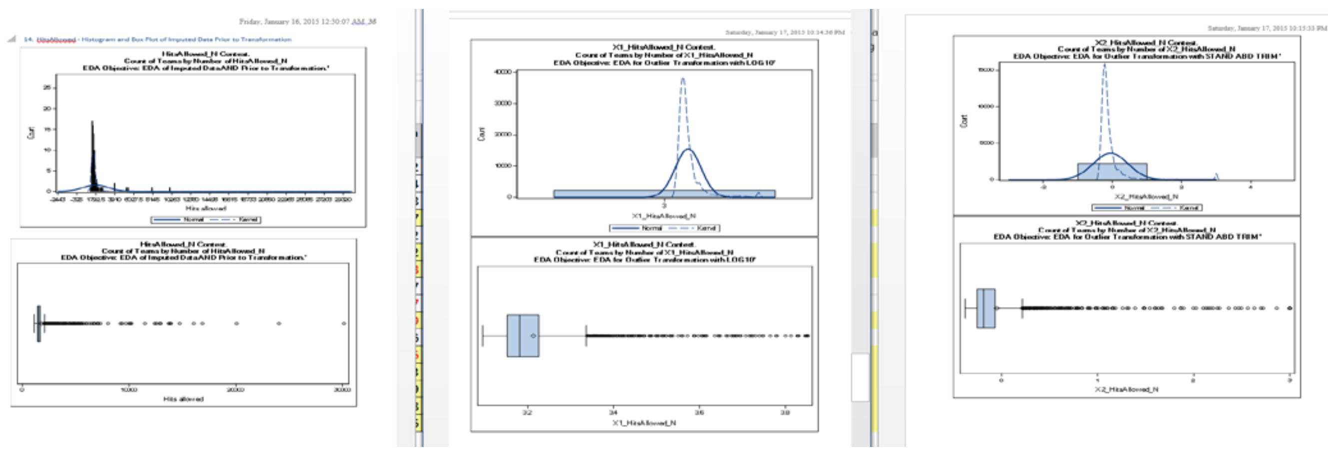


Figure 12 - IMP_DoublePlays_P Outlier Transformation EDA

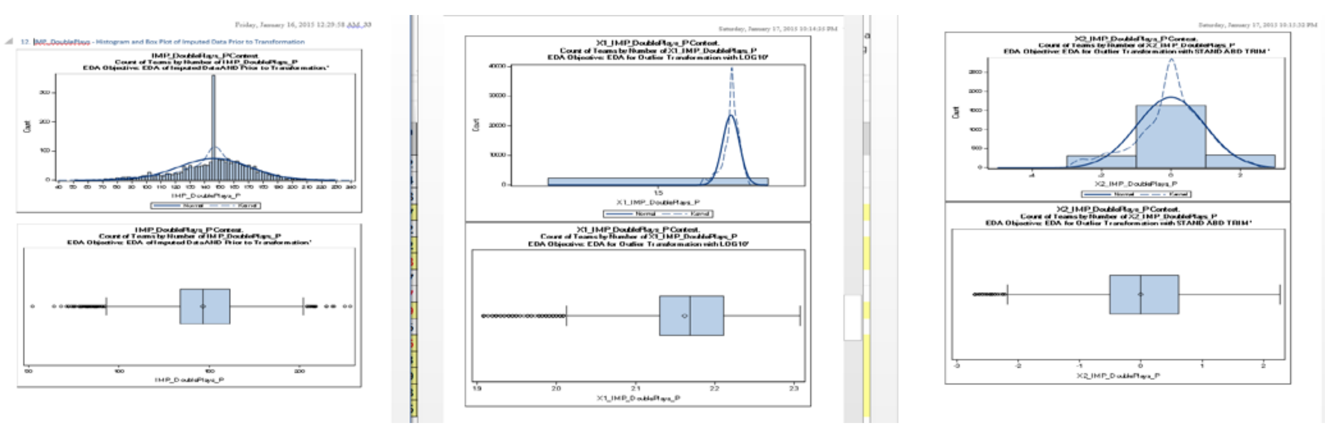


Figure 13 - Errors_N Outlier Transformation EDA

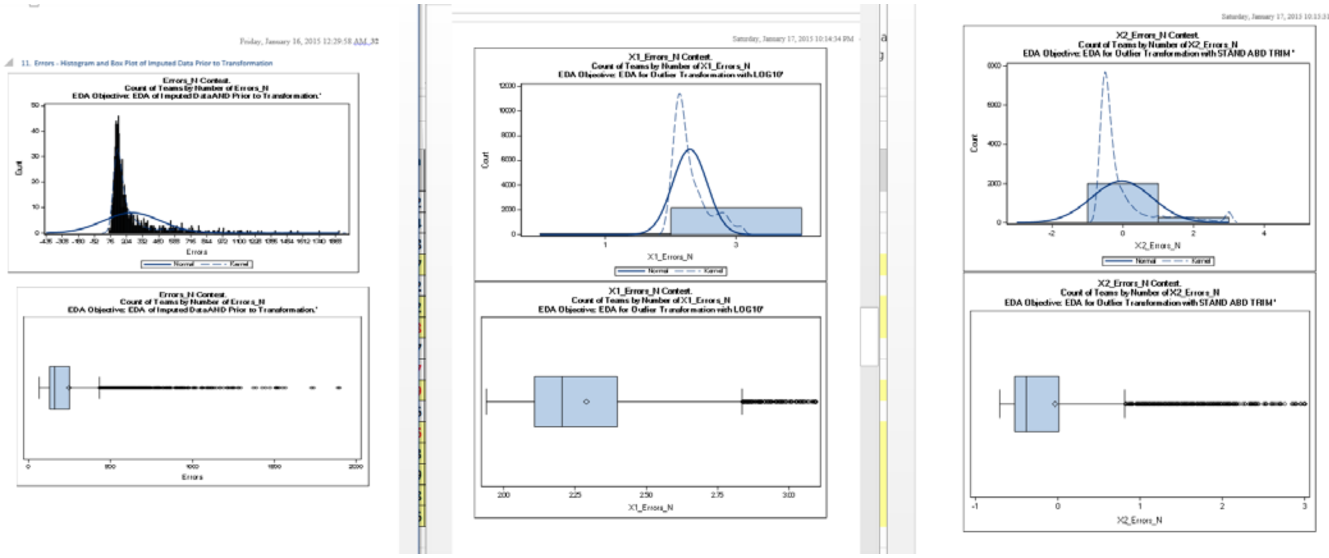


Figure 14 - IMP_CaughtStealing_N Outlier Transformation EDA

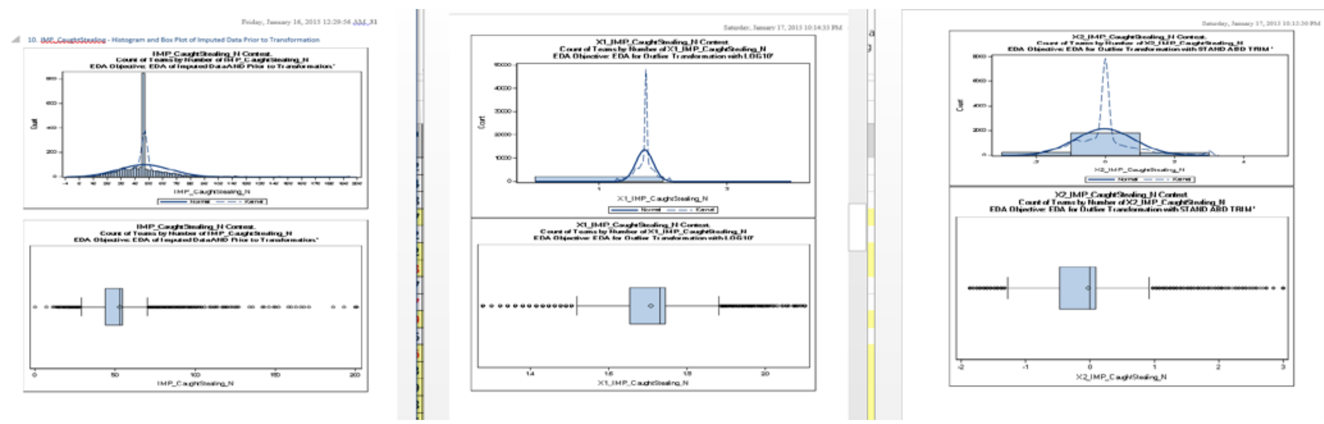


Figure 15 - IMP_StolenBases_P Outlier Transformation EDA

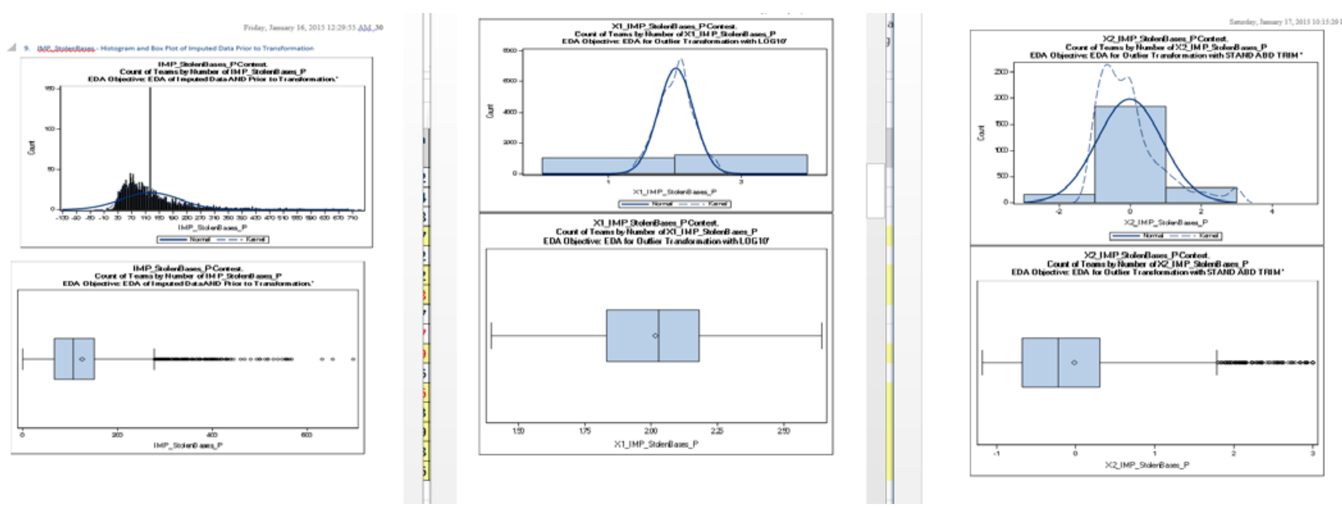


Figure 16 - IMP_BattersHitByPitch_P Outlier Transformation EDA

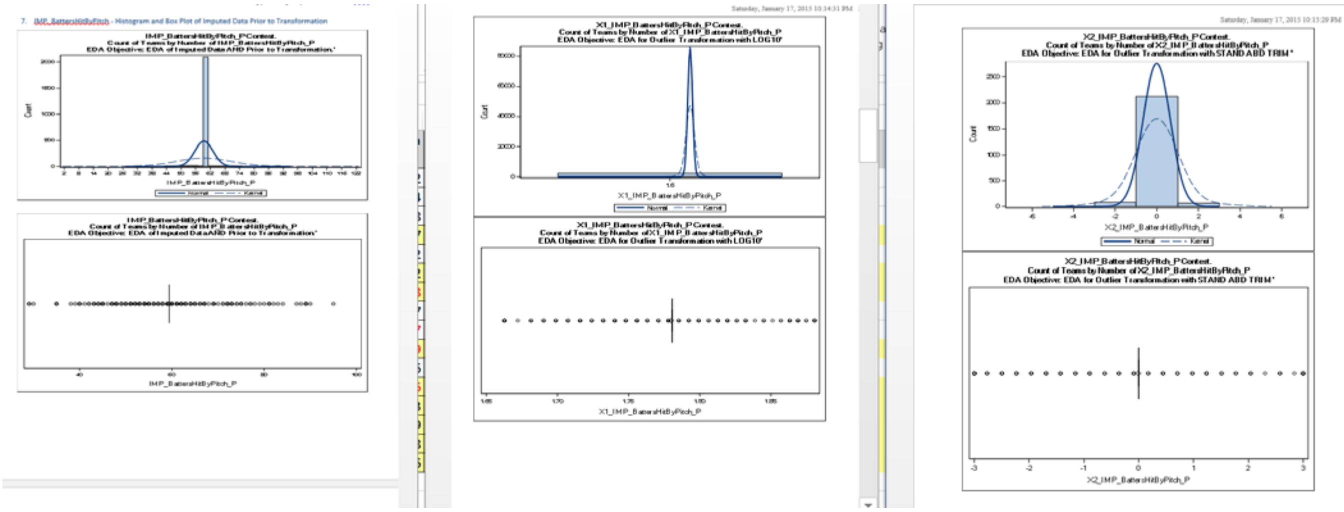


Figure 17 - WalksByBatters_P Outlier Transformation EDA

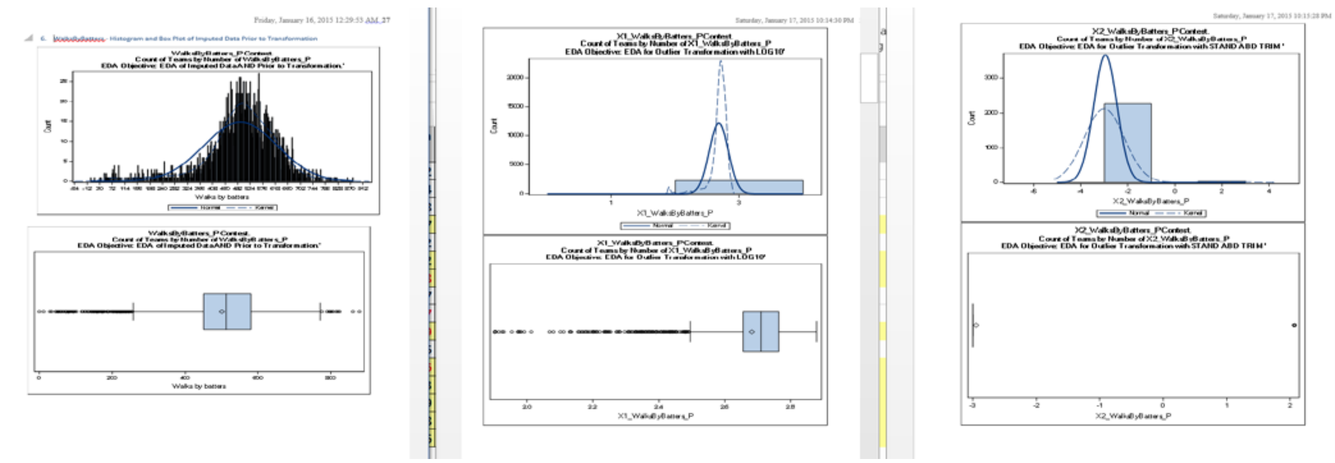
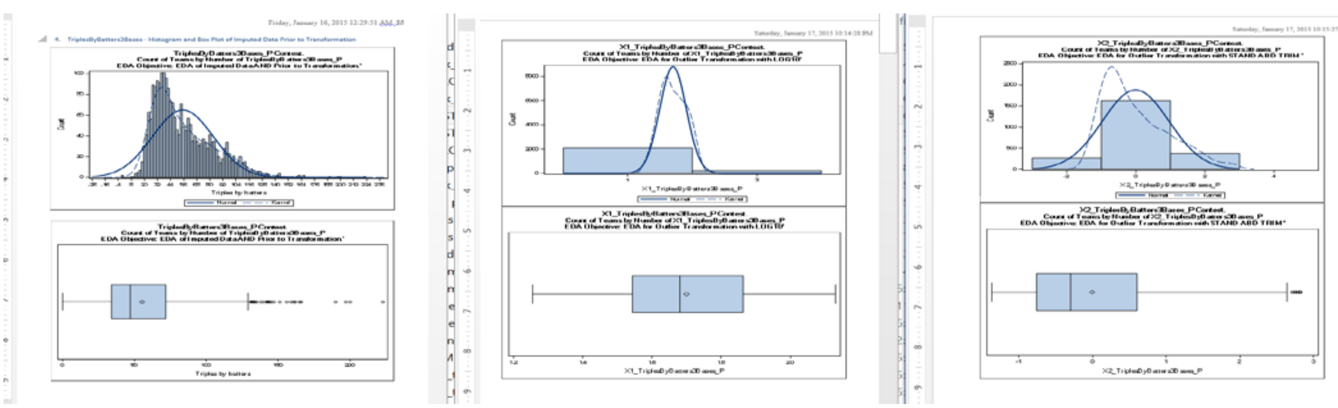


Figure 18 - TriplesByBatters3Bases_P Outlier Transformation EDA



Based on the comparative analysis of the Transformation results above, the following Matrix in Figure 19 below recommends which Transformation variables may be best to use in the model based on the Histogram and Boxplot EDA.

Figure 19 – Analysis of Variable Transformation Results

Variable	N	1st Pctl	99th Pctl	Median	Which Variable to Use? XLog10 or StandTrim?	Reason
TargetWins	2276	38	114	82		
BaseHitsByBattersAllBases_P	2276	1188	1950	1454		
DoublesByBatters2Bases_P	2276	141	352	238		
TriplesByBatters3Bases_P	2276	17	134	47	LOG10 CAPPED	The Histogram shows the Normal and the Density curves close together. The Range is much smaller. The Boxplot shows no Outliers.
HomerunsByBatters4Bases_P	2276	4	235	102		
WalksByBatters_P	2276	79	755	512	STAND AND TRIM CAPPED	The Histogram shows a normal distribution tighter around the Mean. The Boxplot shows no left tail Outliers and only two right tail Outliers.
HitsAllowed_N	2276	1244	7093	1518	STAND AND TRIM CAPPED	The Histogram shows a normal distribution tighter around the Mean and negative left tail that can help offset the higher outliers. The Boxplot shows the Mean further from the Median than the Boxplot from the Log10. So this transformed variable can be swapped for the Log 10.
HomerunsAllowed_N	2276	8	244	107		
WalksAllowed_N	2276	237	924	537		
Errors_N	2276	86	1237	159	LOG10 CAPPED	The Histogram shows the Normal and the Density curves close together. The Range is much smaller. The Boxplot shows less right tailed Outliers and the Mean closer to the Median.
IMP_StrikeoutsByBatters_N	2276	72	1192	736		
IMP_StolenBases_P	2276	24	438	106	LOG10 CAPPED	The Histogram shows the Normal and the Density curves close together. The Range is much smaller. The Boxplot shows less right tailed Outliers and the Mean closer to the Median.
IMP_CaughtStealing_N	2276	18	125	53	STAND AND TRIM CAPPED	The Histogram shows a normal distribution tighter around the Mean and negative left tail that can help offset the higher outliers. The Boxplot shows the Mean closer from the Median than the Boxplot from the Log10.
IMP_BattersHitByPitch_P	2276	45	75	59	LOG10 CAPPED	The Histogram shows the Normal and the Density curves close together. The Range is much smaller. The Boxplot shows a smaller range.
IMP_StrikeoutsByPitchers_P	2276	208	1464	818	STAND AND TRIM CAPPED	The Histogram shows a normal distribution tighter around the Mean. The Boxplot shows no left tail Outliers and only n right tail Outliers. Also, the Mean is closer to the Median.
IMP_DoublePlays_P	2276	80	202	146	STAND AND TRIM CAPPED	The Histogram shows a normal distribution tighter around the Mean. The Boxplot shows fewer left tail Outliers and the Mean closer to the Median.

1.4 EDA for Variable Selection Based on Simple Regression Model

The selected outlier variables listed in Figure 19 above plus the imputed, regular non-imputed, transformed, and interaction variables were further analyzed by running simple OLS Regression models on each variable.

For the transformed variables the comparison of the R^2 between the Log10 (prefixed with X1) and Standardized transformed (prefixed with X2) variables was also done and the best R^2 or Adjusted R^2 was selected. Refer to **Appendix F – SAS Code for Simple OLS Regression Model for Each Variable for EDA of Imputed and Transformed Data** for the SAS Code.

Figure 20 below shows the variables ordered with the highest R^2 based on the Simple OLS Regression models.

Figure 20 – Highest R^2 variables based on Simple Regression model

	Variable	R^2	Adj R^2
1	BaseHitsByBattersAllBases_P	0.1511	0.1508
2	DoublesByBatters2Bases_P	0.0836	0.0832
3	WalksAllowed_N	0.0154	0.0150
4	HomerunsAllowed_N	0.0357	0.0353
5	HomerunsByBatters4Bases_P	0.0310	0.0306
6 **7**	IMP_StrikeoutsByBatters_N MFlag_StrikeoutsByBatters_N	0.0010	0.0001
8	X1_WalksByBatters_P	0.0403	0.0398
9	X1_Errors_N	0.0205	0.0201
10 **11**	X1_IMP_DoublePlays_P MFlag_DoublePlays_P	0.0023	0.0014
12	X1_HitsAllowed_N	0.0006	0.0002
13 **14**	X1_IMP_CaughtStealing_N MFlag_CaughtStealing_N	0.0002	-63E-5
15	X2_TriplesByBatters3Bases_P	0.0206	0.0202
16 **17**	X2_IMP_StolenBases_P MFlag_StolenBases_P	0.0140	0.0131
18 **19**	X2_IMP_StrikeOutsByPitchers_P MFlag_StrikeoutsByPitchers_P	0.0073	0.0064
20 **21**	X2_IMP_BattersHitByPitch_P MFlag_BattersHitByPitch_P	0.0001	-78E-5
22	INT_P	0.0462	
23	INT_N	0.0238	

The analysis from the Simple OLS Regression model resulted in the selection of the variables that will be used in the building the model, which are the ones shown in Figure 20 above.

The model building starts with the Stepwise, Forward, and Backward selection approach on the same variables. Please refer to section **Model Building Using Stepwise, Forward, and Backward Selection** for details on the model building based on the selected variables from this section.

1.5 Principal Component Analysis (PCA) Based on Results from STEPWISE Selection Model

Based on the results from the STEPWISE selection model, twelve (12) variables were kept for the application of Principal Component Analysis. The SAS Code for PCA is located at **Appendix I – SAS Code for PCA EDA based on Stepwise Selected Model.**

Based on the Eigenvalues of the Correlation Matrix shown in Figure 21 below, it can be seen in the Cumulative column that the first eight (8) variables account for 94% of the Variance in TargetWins.

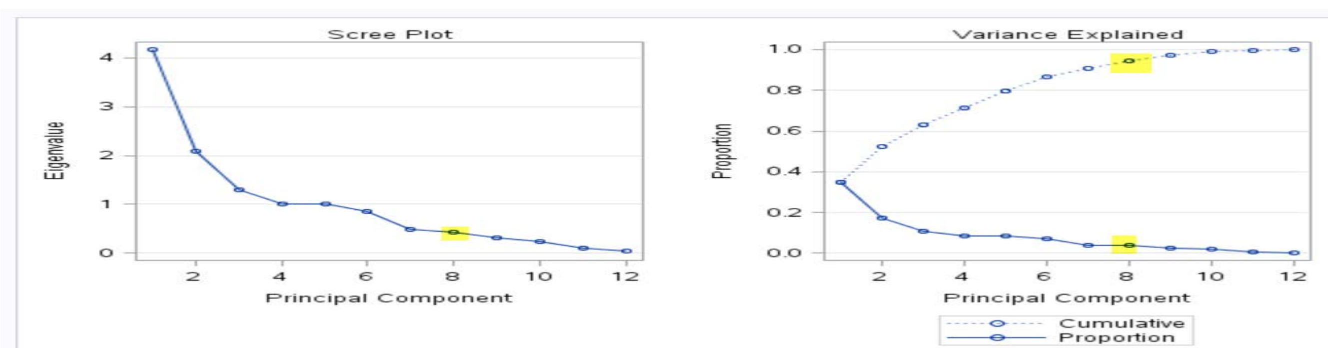
Figure 21 – Correlation Matrix based on STEPWISE selection for the best model proposed for the Scoring program.

	Eigenvalues of the Correlation Matrix				
	Variables	**Eigenvalue**	**Difference**	**Proportion**	**Cumulative**
1	BaseHitsByBattersAllBases_P	4.17083763	2.07582155	0.3476	0.3476
2	MFlag_StolenBases_P	2.09501609	0.79849833	0.1746	0.5222
3	MFlag_CaughtStealing_N	1.29651776	0.29168904	0.1080	0.6302
4	MFlag_BattersHitByPitch_P	1.00482872	0.00349229	0.0837	0.7139
5	X1_WalksByBatters_P	1.00133643	0.15406967	0.0834	0.7974
6	X1_Errors_N	0.84726676	0.36897927	0.0706	0.8680
7	X1_IMP_CaughtStealing_N	0.47828749	0.04103074	0.0399	0.9078
8	X2_TriplesByBatters3Bases_P	0.43725676	0.12889827	0.0364	0.9443
9	X2_IMP_BattersHitByPitch_P	0.30835849	0.07808174	0.0257	0.9700
10	X2_IMP_StolenBases_P	0.23027675	0.13552181	0.0192	0.9892
11	INT_N	0.09475494	0.05949277	0.0079	0.9971
12	INT_P	0.03526217		0.0029	1.0000

Based on the Scree Plot results shown in Figure 22 below, it can be seen that the kink on the Scree plot where the curve flattens is at the 8th component, confirming the observations from the Correlation Matrix above to use the 8th components.

Based on the Variance Explained graph shown below in Figure 22, it can be seen that the first 8 components account for a large proportion of the variance, confirming 94% proportion of the variance that the Correlation Matrix shows.

Figure 22 – Scree Plot and Variance Explained based on STEPWISE selection model proposed for the Scoring program.



Based on this PCA analysis, a PCA Based model is created with only eight (8) variables instead of twelve (12) from the STEPWISE selected model, thus reducing the dimensionality of the model and yet accounting for 94% of the variance. Refer to section **PCA Based Model Building - Model based on the Principal Component Analysis (PCA) Analysis Results** for detail on the building of this additional model.

2. Data Preparation (40 Points)

2.1 Fix missing values

During the data exploration in section **Exploring for Missing Values** above, there were 6 variables identified having Missing values. The Missing values for these 6 variables are imputed or replaced by using the Mean value from the sample data set.

In preparation for the imputation of the Missing values the following steps are performed:

2.1.1 **New variables are created** for each of the 6 variables with missing values. The creation of these new variables is done in the moneyball_train data set and they store the imputed value and a flag indicating whether imputation was done for an observation with a 1 or not done with a 0. Figure 5 below is the new structure of the moneyball_train data set with these additional variables.

Figure 5 – Additional Variables Added for Imputation and Missing Flags

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Label
2	BaseHitsByBattersAllBases_P	Num	8	Base Hits by batters
10	BattersHitByPitch_P	Num	8	Batters hit by pitch
9	CaughtStealing_N	Num	8	Caught stealing
16	DoublePlays_P	Num	8	Double Plays
3	DoublesByBatters2Bases_P	Num	8	Doubles by batters
15	Errors_N	Num	8	Errors
11	HitsAllowed_N	Num	8	Hits allowed
12	HomerunsAllowed_N	Num	8	Homeruns allowed
5	HomerunsByBatters4Bases_P	Num	8	Homeruns by batters
24	IMP_BattersHitByPitch_P	Num	8	
22	IMP_CaughtStealing_N	Num	8	
28	IMP_DoublePlays_P	Num	8	
20	IMP_StolenBases_P	Num	8	
17	IMP_StrikeOutByBatters_N	Num	8	
26	IMP_StrikeoutsByPitchers_P	Num	8	
25	MFlag_BattersHitByPitch_P	Num	8	
23	MFlag_CaughtStealing_N	Num	8	
29	MFlag_DoublePlays_P	Num	8	
21	MFlag_StolenBases_P	Num	8	
19	MFlag_StrikeOutByBatters_N	Num	8	
27	MFlag_StrikeoutsByPitchers_P	Num	8	
8	StolenBases_P	Num	8	Stolen bases
18	StrikeOutByBatters_N	Num	8	
7	StrikeoutsByBatters_N	Num	8	Strikeouts by batters
14	StrikeoutsByPitchers_P	Num	8	Strikeouts by pitchers
1	TargetWins	Num	8	
4	TriplesByBatters3Bases_P	Num	8	Triples by batters
13	WalksAllowed_N	Num	8	Walks allowed
6	WalksByBatters_P	Num	8	Walks by batters

2.1.2 **Data Imputation SAS Code (Macro)** forMissing Values and setting of Missing Flags is performed via the macro as shown in **Appendix C – SAS Code for Data Imputation.** Note that the macro code is later replaced with hardcoding of the imputation in another section, but originally this Macro code is used to quickly get the imputation done.

2.1.3 The **results from the Data Imputation** of the step above are shown in **Appendix A - Imputation Results of Missing Value and Missing Flag setting** for the first 200 observations. It can be seen that the imputed value for each different variable matches the Mean of the PROC MEANS shown in **Figure 4 PROC MEANS result matrix of the Raw Data** above, confirming the Macro is working correctly.

2.2 Transforming Variables with Outlier values

Based on the data exploration for Outliers at **Exploring for Outliers**, there were nine (9) variables identified as having outlier values.

The transformation of the Outlier values was performed using 3 different methods which are explained as follows.

2.3 The **Cap technique,** in which the value for the 1 percentile and the 99 percentile for the given variable was used to cap the lowest possible value and the highest possible value for the variable, respectively.

2.4 Once the capping was done, then that resulting value or the original value if no capping was done was used to apply the mathematical transformation. The **Log10** was applied and results obtained.

2.5 Also, once the capping was done, then separately the **Standardization with Trimming** was applied in parallel to the Log10 transformation.

Refer to **Appendix D – SAS Code for the Transformation of Outlier Values** for details on the SAS Code.

2.3 Creation of Interaction Variables

Two interaction variables are created to bust the accuracy of the model. The idea behind the creation of these two variables is to create one compounded variable with all positive impact in the INT_P as a 'reward' for the teams that have higher scores in measures that increase their winning chance or creating a 'penalty' for teams that have higher scores in measures that decrease their winning chance.

INT_N is created by multiplying all the variables with negative impact on TargetWins.

INT_P is created by multiplying all the variables with positive impact on TargetWins.

Transformations:

INT_N = IMP_StrikeoutsByBatters_N * IMP_CaughtStealing_N * Errors_N;

INT_P = BaseHitsByBattersAllBases_P * WalksByBatters_P * IMP_StolenBases_P * IMP_BattersHitByPitch_P;

These two new interactive features or variables are included in the first model building.

3. Build Models (40 Points)

Model Creation:

3.1 Model Building Using Stepwise, Forward, and Backward Selection

The SAS Code for this model is located at **Appendix G – SAS Code for Model Building using Stepwise, Forward, and Backward Selection on All Variables from Outlier EDA Results.**

Three (3) models are created using the variables listed in Figure 20 in the section **Exploring for Outliers** which include the variables that were imputed and transformed for outliers. These three (3) models are created using the PROC REG selection options for Stepwise, Forward, and Backward. The analysis of these models is explained below.

It can be clearly seen in Figure 23 below that the FORWARD model selection resulted in the highest Adjusted R^2 at 0.43017 and the lowest RMSE of 11.891; however the AIC score of 11290.6 is the same across all three models because they all have the same number of variables.

Figure 23 – Model Validation Metrics for the Stepwise, Forward, and Backward Selection Results using all 23 variables with Transformation of the Variables.

Obs	_MODEL_	_ADJRSQ_	_CP_	_AIC_	_BIC_	_SBC_	_RMSE_	Intercept
1	MODEL_STEPWISE	0.42994	19.063	11290.6	11292.9	11405.2	11.893	253.895
2	MODEL_FORWARD	0.43017	19.148	11290.6	11293	11411	11.891	292.791
3	MODEL_BACKWARD	0.42994	19.063	11290.6	11292.9	11405.2	11.893	253.895

***On a side note,* the measures shown below are from only having Imputed the variables and no Transformation for Outliers was performed. Obviously, the measures above are much better because the Adjusted R^2 is higher and the AIC is lower than the measures shown below resulting in the *conclusion* that Outlier transformation was required after all and that my incorrect finding a few days ago was due to mis-coding the transformation and the model in the Scoring program.**

		results without Transformation but only Imputation						
Obs	_MODEL_	_ADJRSQ_	_CP_	_AIC_	_BIC_	_SBC_	_RMSE_	Intercept
1	MODEL_STEPWISE	0.42201	15.071	11320	11322.3	11423.1	11.976	-14.5887
2	MODEL_FORWARD	0.42214	15.583	11320.5	11322.9	11429.4	11.974	-15.364
3	MODEL_BACKWARD	0.42201	15.071	11320	11322.3	11423.1	11.976	-14.5887

The resulting Parameter Estimates from the three (3) selection approaches are shown below in Figure 24 below.

The Variables with incorrect signed coefficients in their Parameters are in red font. The sign in the coefficients are contrary to the functional effect that the given variable is supposed to have on the dependent variable, TargetWins. For example, DoublesByBatters2Bases is supposed to have an increasing effect on TargetWins, yet the coefficient is negative which does not make sense. All variables with incorrect functional coefficient are in red font in Figure 24 below and they will be removed for the next iteration of building the model.

The Variables with VIF higher than 9.0 are highlighted in yellow. However, these large VIF variables will be left in the next iteration of the model building in which the incorrect signed coefficients will be removed as it is expected the VIF will decrease.

Figure 24 – Parameter Estimates for the Model Selection – First iteration with all variables imputed and transformed.

Parameter Estimates - STEPWISE

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	253.89513	20.41443	12.44	<.0001	0
BaseHitsByBattersAllBases_P	1	0.05197	0.00348	14.94	<.0001	4.06639
DoublesByBatters2Bases_P	1	-0.02872	0.00889	-3.23	0.0013	2.78704
WalksAllowed_N	1	0.00831	0.00196	4.25	<.0001	1.7027
HomerunsAllowed_N	1	0.04908	0.00846	5.8	<.0001	4.32114
MFlag_StrikeoutsByBatters_N	1	8.35129	1.49033	5.6	<.0001	1.52984
X1_WalksByBatters_P	1	16.82267	4.23764	3.97	<.0001	6.39763
X1_Errors_N	1	-60.27238	3.4903	-17.27	<.0001	13.528
X1_IMP_DoublePlays_P	1	-39.36576	4.53316	-8.68	<.0001	1.97204
MFlag_DoublePlays_P	1	4.40097	1.48957	2.95	0.0032	3.92254
X1_IMP_CaughtStealing_N	1	-12.5984	2.69231	-4.61	<.0001	2.10888
MFlag_CaughtStealing_N	1	1.87044	0.86809	2.15	0.0313	2.71778
X2_TriplesByBatters3Bases_P	1	3.07115	0.45786	6.71	<.0001	3.14517
X2_IMP_StolenBases_P	1	49.08885	6.39769	7.67	<.0001	14.0981
MFlag_StolenBases_P	1	34.17604	1.81326	18.85	<.0001	2.86975
X2_IMP_StrikeoutsByPitchers_P	1	-8.96098	1.24162	-7.22	<.0001	4.53496
X2_IMP_BattersHitByPitch_P	1	2.1025	1.12923	1.86	0.0627	1.04986
MFlag_BattersHitByPitch_P	1	6.37818	1.08217	5.89	<.0001	1.44861
INT_P	1	-5.39E-10	2.11E-10	-2.56	0.0107	13.6129
INT_N	1	2.24E-07	9.16E-08	2.44	0.0147	4.95428

Parameter Estimates - FORWARD

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	292.79084	34.73081	8.43	<.0001	0
BaseHitsByBattersAllBases_P	1	0.0551	0.00415	13.28	<.0001	5.7869
DoublesByBatters2Bases_P	1	-0.02894	0.00889	-3.25	0.0012	2.78792
WalksAllowed_N	1	0.01159	0.00307	3.77	0.0002	4.20842
HomerunsAllowed_N	1	0.04997	0.00848	5.89	<.0001	4.34656
MFlag_StrikeoutsByBatters_N	1	8.0947	1.50151	5.39	<.0001	1.55352
X1_WalksByBatters_P	1	10.99683	5.97207	1.84	0.0657	12.7115
X1_Errors_N	1	-58.31827	3.76435	-15.49	<.0001	15.7422
X1_IMP_DoublePlays_P	1	-39.24305	4.53311	-8.66	<.0001	1.97279
MFlag_DoublePlays_P	1	4.30679	1.49083	2.89	0.0039	3.93073
X1_HitsAllowed_N	1	-11.39195	8.23026	-1.38	0.1664	14.9897
X1_IMP_CaughtStealing_N	1	-12.02414	2.70531	-4.44	<.0001	2.13015
MFlag_CaughtStealing_N	1	1.77894	0.87042	2.04	0.0411	2.73355
X2_TriplesByBatters3Bases_P	1	3.06937	0.45777	6.71	<.0001	3.1452
X2_IMP_StolenBases_P	1	47.86345	6.45736	7.41	<.0001	14.3681
MFlag_StolenBases_P	1	34.62489	1.84167	18.8	<.0001	2.96157
X2_IMP_StrikeoutsByPitchers_P	1	-8.61698	1.266	-6.81	<.0001	4.71673
X2_IMP_BattersHitByPitch_P	1	2.03634	1.13001	1.8	0.0717	1.05174
MFlag_BattersHitByPitch_P	1	6.41578	1.08229	5.93	<.0001	1.44952
INT_P	1	-4.97E-10	2.13E-10	-2.33	0.0198	13.8989
INT_N	1	1.90E-07	9.48E-08	2	0.0455	5.31128

Parameter Estimates - BACKWARD

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	253.89513	20.41443	12.44	<.0001	0
BaseHitsByBattersAllBases_P	1	0.05197	0.00348	14.94	<.0001	4.06639
DoublesByBatters2Bases_P	1	-0.02872	0.00889	-3.23	0.0013	2.78704
WalksAllowed_N	1	0.00831	0.00196	4.25	<.0001	1.7027
HomerunsAllowed_N	1	0.04908	0.00846	5.8	<.0001	4.32114
MFlag_StrikeoutsByBatters_N	1	8.35129	1.49033	5.6	<.0001	1.52984
X1_WalksByBatters_P	1	16.82267	4.23764	3.97	<.0001	6.39763
X1_Errors_N	1	-60.27238	3.4903	-17.27	<.0001	13.528
X1_IMP_DoublePlays_P	1	-39.36576	4.53316	-8.68	<.0001	1.97204
MFlag_DoublePlays_P	1	4.40097	1.48957	2.95	0.0032	3.92254
X1_IMP_CaughtStealing_N	1	-12.5984	2.69231	-4.61	<.0001	2.10888
MFlag_CaughtStealing_N	1	1.87044	0.86809	2.15	0.0313	2.71778
X2_TriplesByBatters3Bases_P	1	3.07115	0.45786	6.71	<.0001	3.14517
X2_IMP_StolenBases_P	1	49.08885	6.39769	7.67	<.0001	14.0981
MFlag_StolenBases_P	1	34.17604	1.81326	18.85	<.0001	2.86975
X2_IMP_StrikeoutsByPitchers_P	1	-8.96098	1.24162	-7.22	<.0001	4.53496
X2_IMP_BattersHitByPitch_P	1	2.1025	1.12923	1.86	0.0627	1.04986
MFlag_BattersHitByPitch_P	1	6.37818	1.08217	5.89	<.0001	1.44861
INT_P	1	-5.39E-10	2.11E-10	-2.56	0.0107	13.6129
INT_N	1	2.24E-07	9.16E-08	2.44	0.0147	4.95428

Based on the results above in Figure 24, the Variables in red font, with incorrect signed coefficients are removed from the models. However, the resulting models left in the MFlag_DoublePlays_P without its corresponding variable, so this MFlag variable will be removed, refer to the variables in red font in Figure 25 below.

Also, the resulting models included the imputed variables X1_IMP_CaughtStealing_N but left out its corresponding Flag variable MFlag_CaughtStealing_N shown in green font with yellow highlight in Figure 25 below.

The variables in green font shown in Figure 25 below will be used for the final iteration of model building.

Figure 25 –Variables with correct coefficient signs with respect to functional expected effect on dependent TargetWins variable.

STEPWISE Variables	FORWARD Variables	BACKWARD Variables
BaseHitsByBattersAllBases_P	BaseHitsByBattersAllBases_P	BaseHitsByBattersAllBases_P
X1_WalksByBatters_P	X1_WalksByBatters_P	X1_WalksByBatters_P
X1_Errors_N	X1_Errors_N	X1_Errors_N
MFlag_DoublePlays_P	MFlag_DoublePlays_P	MFlag_DoublePlays_P
X1_IMP_CaughtStealing_N	X1_HitsAllowed_N	X1_IMP_CaughtStealing_N
X2_TriplesByBatters3Bases_P	X1_IMP_CaughtStealing_N	X2_TriplesByBatters3Bases_P
X2_IMP_StolenBases_P	X2_TriplesByBatters3Bases_P	X2_IMP_StolenBases_P
MFlag_StolenBases_P	X2_IMP_StolenBases_P	MFlag_StolenBases_P
X2_IMP_BattersHitByPitch_P	MFlag_StolenBases_P	X2_IMP_BattersHitByPitch_P
MFlag_BattersHitByPitch_P	X2_IMP_BattersHitByPitch_P	MFlag_BattersHitByPitch_P
INT_P	MFlag_BattersHitByPitch_P	INT_P
INT_N	INT_P	INT_N
	INT_N	

MFlag_CaughtStealing_N	MFlag_CaughtStealing_N	MFlag_CaughtStealing_N

3.2 Model Building Based on the Stepwise, Forward, and Backward Selection Results and Without the Variables with Incorrect Signed Coefficients

The creation of three (3) models is based on resulting variables from the Stepwise, Forward, and Backward selection section 3.1 above shown in Figure 25. The SAS Code for this second OLS Stepwise is located at **Appendix H – SAS Code for Model Building Based on the Stepwise, Forward, and Backward Selection Results Above and Without the Variables with Incorrect Signed Coefficients.**

Based on Figure 26 below, it can clearly be seen that both the STEPWISE and BACKWARD models have the highest Adjusted R^2 at 0.38407 and the lowest AIC at 11459.7.

Both, the STEPWISE and FORWARD models are the best model so the Final model building will use the STEPWISE for simplicity.

Figure 26 – Model Validation Metrics for the Stepwise, Forward, and Backward Selection Results After removing incorrectly signed coefficients.

Obs	_MODEL_	_ADJRSQ_	_CP_	_AIC_	_BIC_	_SBC_	_RMSE_	Intercept
1	MODEL_FROM_STPW	0.38407	13	11459.7	11461.9	11534.2	12.363	134.428
2	MODEL_FROM_FORW	0.38381	14	11461.7	11463.9	11541.9	12.365	136.872
3	MODEL_FROM_BACKW	0.38407	13	11459.7	11461.9	11534.2	12.363	134.428

Based on the resulting Parameter Estimates from the final iteration of the three (3) selection approaches shown below in Figure 27, it can be clearly seen that the removal of the Variables that had incorrect signed coefficients improved the models because all the Variables now have correctly signed coefficients per their functional expectations.

Also, the VIF for all variables except 2 are much lower than 9.0, which indicates that the final variables selected have no collinearity issues, except for the X2_IMP_StolenBases_P and INT_P variables, but we'll accept that collinearity.

Given that the STEPWISE and BACKWARD selected models are the ones with the highest Adjusted R^2 and lowest AIC, the STEPWISE will be used for the Scoring or Prediction step.

Also, an additional EDA is performed via Principal Component Analysis, in the next section, in order to evaluate if a simpler model can be created.

Figure 27 – Parameter Estimates for the Final Model Selection

Parameter Estimates - STEPWISE

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	134.4278	15.1905	8.85	<.0001	0
BaseHitsByBattersAllBases_P	1	0.05043	0.00244	20.68	<.0001	1.8498
X1_WalksByBatters_P	1	24.34168	3.86446	6.3	<.0001	4.92427
X1_Errors_N	1	-47.3013	2.71993	-17.39	<.0001	7.60353
X1_IMP_CaughtStealing_N	1	-10.2286	2.42509	-4.22	<.0001	1.58362
MFlag_CaughtStealing_N	1	4.10759	0.85137	4.82	<.0001	2.41947
X2_TriplesByBatters3Bases_P	1	3.1484	0.44296	7.11	<.0001	2.72463
X2_IMP_StolenBases_P	1	57.53961	6.46671	8.9	<.0001	13.3313
MFlag_StolenBases_P	1	30.64349	1.80309	17	<.0001	2.62632
X2_IMP_BattersHitByPitch_P	1	2.18359	1.17147	1.86	0.0625	1.04573
MFlag_BattersHitByPitch_P	1	7.47316	1.06106	7.04	<.0001	1.28894
INT_P	1	-7.15E-10	2.15E-10	-3.32	0.0009	13.1261
INT_N	1	-1.42E-07	7.43E-08	-1.92	0.0552	3.01464

Parameter Estimates - FORWARD

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	136.8721	20.4377	6.7	<.0001	0
BaseHitsByBattersAllBases_P	1	0.05072	0.00293	17.33	<.0001	2.66294
X1_WalksByBatters_P	1	24.09308	4.10771	5.87	<.0001	5.56133
X1_Errors_N	1	-47.0949	2.95516	-15.94	<.0001	8.97173
X1_HitsAllowed_N	1	-0.87883	4.9148	-0.18	0.8581	4.9267
X1_IMP_CaughtStealing_N	1	-10.2471	2.42782	-4.22	<.0001	1.5865
MFlag_CaughtStealing_N	1	4.08305	0.86255	4.73	<.0001	2.48235
X2_TriplesByBatters3Bases_P	1	3.13239	0.45202	6.93	<.0001	2.83598
X2_IMP_StolenBases_P	1	57.40904	6.50918	8.82	<.0001	13.5012
MFlag_StolenBases_P	1	30.71109	1.84267	16.67	<.0001	2.74173
X2_IMP_BattersHitByPitch_P	1	2.17574	1.17254	1.86	0.0636	1.0472
MFlag_BattersHitByPitch_P	1	7.47473	1.06132	7.04	<.0001	1.28903
INT_P	1	-7.10E-10	2.18E-10	-3.26	0.0011	13.3911
INT_N	1	-1.44E-07	7.46E-08	-1.93	0.0542	3.04327

Parameter Estimates - BACKWARD

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	134.4278	15.1905	8.85	<.0001	0
BaseHitsByBattersAllBases_P	1	0.05043	0.00244	20.68	<.0001	1.8498
X1_WalksByBatters_P	1	24.34168	3.86446	6.3	<.0001	4.92427
X1_Errors_N	1	-47.3013	2.71993	-17.39	<.0001	7.60353
X1_IMP_CaughtStealing_N	1	-10.2286	2.42509	-4.22	<.0001	1.58362
MFlag_CaughtStealing_N	1	4.10759	0.85137	4.82	<.0001	2.41947
X2_TriplesByBatters3Bases_P	1	3.1484	0.44296	7.11	<.0001	2.72463
X2_IMP_StolenBases_P	1	57.53961	6.46671	8.9	<.0001	13.3313
MFlag_StolenBases_P	1	30.64349	1.80309	17	<.0001	2.62632
X2_IMP_BattersHitByPitch_P	1	2.18359	1.17147	1.86	0.0625	1.04573
MFlag_BattersHitByPitch_P	1	7.47316	1.06106	7.04	<.0001	1.28894
INT_P	1	-7.15E-10	2.15E-10	-3.32	0.0009	13.1261
INT_N	1	-1.42E-07	7.43E-08	-1.92	0.0552	3.01464

In Figures 27-B below, the best model complies with the the OLS Assumptions of

1. **Linearity** assumption was confirmed during the Simple OLS EDA at **EDA for Variable Selection Based on Simple Regression Model**

2. **Homoscendaticity** or Normality assumption is confirmed in the random pattern of the residual and the predicted value in the 'Residual by Predicted Values' scatter plot and the Quantile graph in Figure 27-B below.

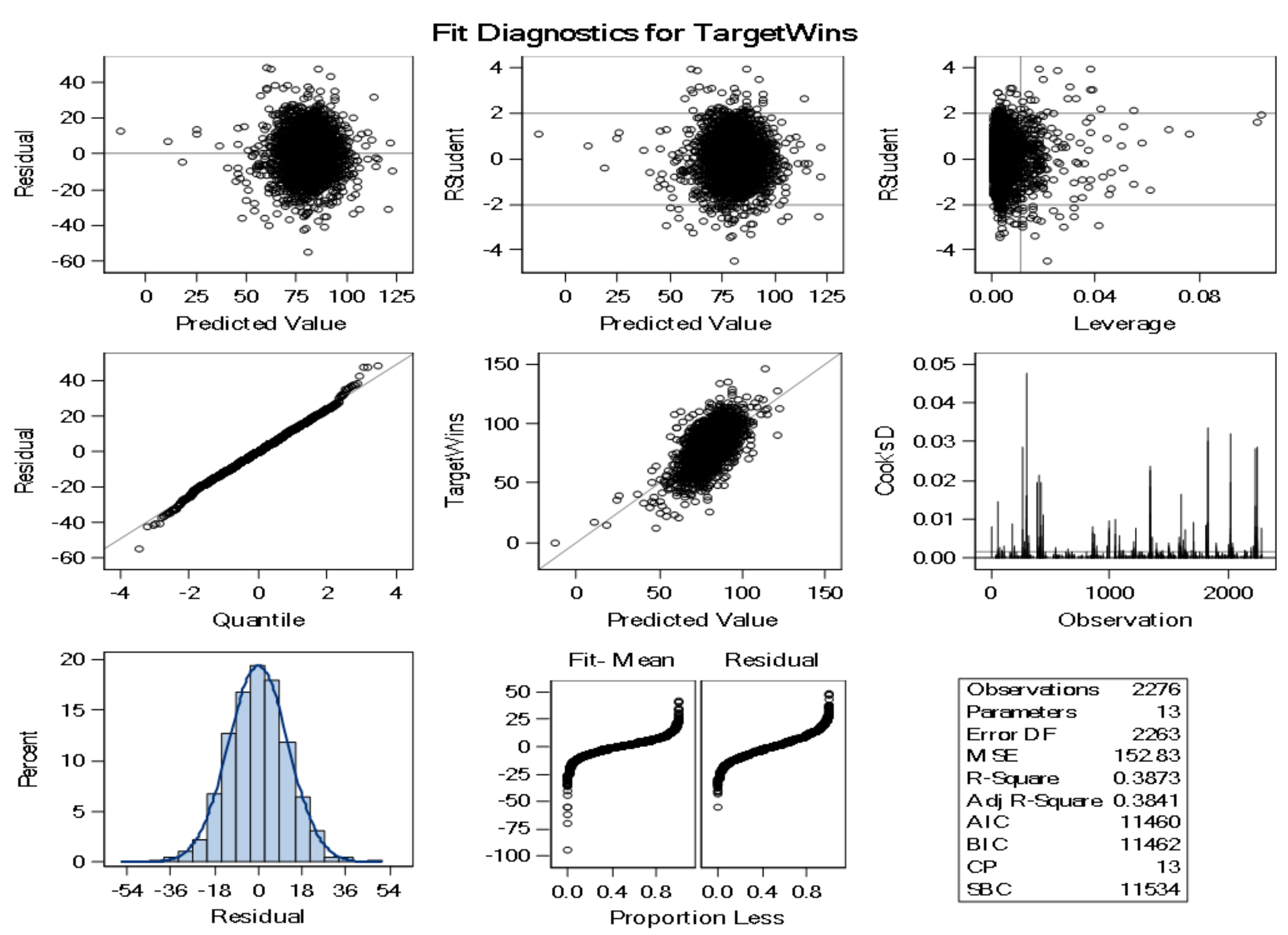
3. **Auto correlation among the Error terms:** none of the graph of the Residuals have a pattern that may indicate autocorrelation.

4. **Predictor correlation with error term is zero** assumption is preserved because the VIF metric is very low for all variables except two.

5. **Error term is normally distributed with Mean = 0 and constant variance:** this assumption is confirmed by the "Percent by Residual" histogram and the "Residual and Quantile" graph in Figure 27-B below.

Figure 27-B – Fit Diagnostics for the Final Model Selection

The REG Procedure
Model: MODEL_FROM_STPW
Dependent Variable: TargetWins



3.3 PCA Based Model Building - Model based on the Principal Component Analysis (PCA) Analysis Results

Based on the PCA analysis in **Principal Component Analysis (PCA) Based on Results from STEPWISE Selection Model**, a PCA Based model was created.

The SAS Code for this *PCA Result based* model is located at **Appendix J – SAS Code for Building PCA Based Model at 94% of Variance with Reduced Dimensionality by Four (4) Variables Less.**

The results shown in Figure 28 clearly show that the two PCA Based models with 94% variance under-perform the STEPWISE selected model with Adj-R^2 of 0. 38407 and the AIC of 11459.7.

Figure 28 – Validation Metrics for PCA Based Model at 94% of Variance.

A	B	O	P	Q	R	S	T	U	V	W	X
Obs	_MODEL_	_P_	_EDF_	_RSQ_	_ADJRSQ_	_CP_	_AIC_	_BIC_	_SBC_	_RMSE_	Intercept
1	MODEL_PCA_BASED_94	9	2267	0.3685	0.36631	9	11520.5	11522.5	11572	12.54	122.274

After this analysis, it can be concluded that the PCA Based model will *not* be used for Scoring.

4. Select Model (40 Points)

Based on the results in Figure 29 below, it can clearly be seen that the model created by the STEPWISE selection process is the best model with the highest Adj-R^2 at 0.38407 and the lowest AIC at 11459.7.

The model from the STEPWISE selection process will be used as the Prediction model to be deployed.

Figure 29 – Criteria for Selecting Best Model for Deployment. STEPWISE can be used.

Obs	_MODEL_	_ADJRSQ_	_CP_	_AIC_	_BIC_	_SBC_	_RMSE_	Intercept
1	MODEL_FROM_STPW	0.38407	13	11459.7	11461.9	11534.2	12.363	134.428
2	MODEL_FROM_FORW	0.38381	14	11461.7	11463.9	11541.9	12.365	136.872
3	MODEL_FROM_BACKW	0.38407	13	11459.7	11461.9	11534.2	12.363	134.428

The model to be used for prediction is shown in Figure 30 below.

Figure 30 – Best Model Results

The REG Procedure
Model: MODEL_FROM_STPW
Dependent Variable: TargetWins

Number of Observations Read	2276
Number of Observations Used	2276

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	218640	18220	119.22	<.0001
Error	2263	345856	152.83092		
Corrected Total	2275	564496			

Root MSE	12.36248	R-Square	0.3873
Dependent Mean	80.79086	Adj R-Sq	0.3841
Coeff Var	15.30183		

Parameter Estimates – STEPWISE						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	134.42779	15.1905	8.85	<.0001	0
BaseHitsByBattersAllBases_P	1	0.05043	0.00244	20.68	<.0001	1.8498
X1_WalksByBatters_P	1	24.34168	3.86446	6.3	<.0001	4.92427
X1_Errors_N	1	-47.30128	2.71993	-17.39	<.0001	7.60353
X1_IMP_CaughtStealing_N	1	-10.22857	2.42509	-4.22	<.0001	1.58362
MFlag_CaughtStealing_N	1	4.10759	0.85137	4.82	<.0001	2.41947
X2_TriplesByBatters3Bases_P	1	3.1484	0.44296	7.11	<.0001	2.72463
X2_IMP_StolenBases_P	1	57.53961	6.46671	8.9	<.0001	13.3313
MFlag_StolenBases_P	1	30.64349	1.80309	17	<.0001	2.62632
X2_IMP_BattersHitByPitch_P	1	2.18359	1.17147	1.86	0.0625	1.04573
MFlag_BattersHitByPitch_P	1	7.47316	1.06106	7.04	<.0001	1.28894
INT_P	1	-7.15E-10	2.15E-10	-3.32	0.0009	13.1261
INT_N	1	-1.42E-07	7.43E-08	-1.92	0.0552	3.01464

The Parameter Estimate matrix shown above in Figure 30 shows most variables with p-values of < .0001. The only exceptions are for variables X2_IMP_BattersHitByPitch_P at 0.0625, which is not much higher than .05; and for variable INT_N at 0.0552, which again is not much higher than .05. These p-values overall show a statistically significant model at 95% confidence level that the model has predictive coefficients that closely match the population results so the model is safe to use for Scoring.

The VIF for two variables, X2_IMP_StolenBases_P and for INT_P, are greater than 10.0 thus indicating some collinearity of these variables. However, the VIF is not much higher than 10.0 so the variables are relatively safe to keep with the awareness that their collinearity may impact the accuracy of these two predictors on the Scoring.

CONCLUSION:

The Scoring or Predictive model was chosen from the best performing model in terms of the highest Adjusted R^2 at 0.38407 and the lowest AIC measure at 11459.72. This model was selected by both the Stepwise and Backward selections; however, manual fine tuning was done to the model in order to remove variables with incorrect sign in the coefficient indicating that the data for those variables were affected by outliers that were not fixed with transformation.

Having left in the model the variables with incorrect coefficients would render the model erroneous because the predicted impact on the TargetWins would have been contrary to the functional or known expected impact, such as more homeruns would normally be expected to increase the chances of TargetWins but a negative coefficient in this variable would have predicted that more Homeruns would decrease the chances of TargetWins. So those variables were manually removed from the final estimated model to ensure the model is designed or specified correctly.

The Simple OLS regression model analysis confirmed the high predictive ability of each of the predictors on the TargetWins.

To ensure that the selected best model had the lowest dimensionality possible, Principal Component Analysis was performed, however, the analysis did not result in a better model at 94% of the variance with eight (8) variables rather than the final twelve (12) variables for the best model.

The selected model was confirmed to meet the OLS Regression Assumptions after post-model estimation analysis was performed using the Fit Diagnostic on TargetWins charts based on which the evaluation of the Residual distribution and Linearity were done.

All variables in the model but for two had p-values of < .0001. The two exception variables had p-values just slightly higher than .05, these were X2_IMP_BattersHitByPitch_P with 0.0625 and INT_N with 0.0552. Since these p-values are not much higher than .05, the variables were left in the model as they made the model more accurate than having taken them out.

The VIF for all variables was below 9.0 except for two variables which had VIF of 13, which would create bias on those two variables, X2_IMP_StolenBases_P and INT_P, however, these variables were left in the model as having taken them out of the model would have decreased the accuracy of the model which does not offset the slight biased produced by this small collinearity.

The Scoring model was confirmed to meet the OLS Assumptions of Homoscedasticity, Auto correlation among the Error terms, predictor correlation with error term is zero, and the error term is normally distributed with Mean = 0 and constant variance.

Testing the above model against the Train data set resulted in the Sum of Errors between the TargetWins – P_TARGET_WINS of -22.22, which is 22 units from 0. Zero (0) would make a perfect predictive model.

The Scoring model produced by this analysis and model estimation process is the result of extensive data exploration, data preparation, model estimation analysis and re-analysis, manual fine tuning, and testing against the train data set, and perhaps some 100 hours of work in the past 14 days!

The scoring results on the Test data set are on the ballpark at 80.2 Wins on Average based on Excel's Average computation of the predicted values.

The analysis and model estimation provided in this paper conclude the creation of the best scoring or predictive model for the Baseball Moneyball use case requiring to predict the Wins of a Team for a given Season.

5. Bingo Bonus – PROC GLM and PROC

PROC GLM

As shown in Figure 31 and 32 below, both PROC steps, REG and GLM, resulted in the same measures.

SAS Code is at **Appendix K – SAS Code for PROC GLM**

Figure 31 – Comparison of Model Measures Between PROC REG and PROC GLM

The REG Procedure
Model: MODEL_FROM_STPW
Dependent Variable: TargetWins

Number of Observations Read	2276
Number of Observations Used	2276

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	218640	18220	119.22	<.0001
Error	2263	345856	152.83092		
Corrected Total	2275	564496			

Root MSE	12.36248	R-Square	0.3873
Dependent Mean	80.79086	Adj R-Sq	0.3841
Coeff Var	15.30183		

GLM Model based on the Stepwise Scoring model

The GLM Procedure
Dependent Variable: TargetWins

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	218640.0888	18220.0074	119.22	<.0001
Error	2263	345856.3611	152.8309		
Corrected Total	2275	564496.4499			

R-Square	Coeff Var	Root MSE	TargetWins Mean
0.387319	15.30183	12.36248	80.79086

Figure 32 – Comparison of Parameter Estimates Between PROC REG and PROC GLM

Parameter Estimates – STEPWISE - FROM PROC REG						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	134.4278	15.1905	8.85	<.0001	0
BaseHitsByBattersAllBases_P	1	0.05043	0.00244	20.68	<.0001	1.8498
X1_WalksByBatters_P	1	24.34168	3.86446	6.3	<.0001	4.92427
X1_Errors_N	1	-47.3013	2.71993	-17.39	<.0001	7.60353
X1_IMP_CaughtStealing_N	1	-10.2286	2.42509	-4.22	<.0001	1.58362
MFlag_CaughtStealing_N	1	4.10759	0.85137	4.82	<.0001	2.41947
X2_TriplesByBatters3Bases_P	1	3.1484	0.44296	7.11	<.0001	2.72463
X2_IMP_StolenBases_P	1	57.53961	6.46671	8.9	<.0001	13.3313
MFlag_StolenBases_P	1	30.64349	1.80309	17	<.0001	2.62632
X2_IMP_BattersHitByPitch_P	1	2.18359	1.17147	1.86	0.0625	1.04573
MFlag_BattersHitByPitch_P	1	7.47316	1.06106	7.04	<.0001	1.28894
INT_P	1	-7.15E-10	2.15E-10	-3.32	0.0009	13.1261
INT_N	1	-1.42E-07	7.43E-08	-1.92	0.0552	3.01464

Parameter Estimates – STEPWISE - FROM PROC GLM						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	
Intercept		134.4278	15.19048	8.85	<.0001	
BaseHitsByBattersAll		0.050427	0.002438	20.68	<.0001	
X1_WalksByBatters_P		24.34168	3.864459	6.3	<.0001	
X1_Errors_N		-47.3013	2.719928	-17.39	<.0001	
X1_IMP_CaughtStealin		-10.2286	2.425095	-4.22	<.0001	
MFlag_CaughtStealing		4.107595	0.851373	4.82	<.0001	
X2_TriplesByBatters3		3.148404	0.442961	7.11	<.0001	
X2_IMP_StolenBases_P		57.53961	6.466709	8.9	<.0001	
MFlag_StolenBases_P		30.64349	1.803086	17	<.0001	
X2_IMP_BattersHitByP		2.183588	1.171472	1.86	0.0625	
MFlag_BattersHitByPi		7.473165	1.061058	7.04	<.0001	
INT_P		0	0	-3.32	0.0009	
INT_N		-1E-07	7E-08	-1.92	0.0552	

The GLM procedure does not offer the VIF measure or the Adjusted R-Square measure so a comparison between these measures is not possible.

PROC GENMOD

As shown in Figure 33 below, the Parameter Estimates between the results from the PROC REG and PROC GENMOD are not different except for the two Interactive variables that have Estimates of zero (0) indicating that they are not found to have an impact on the variability of the dependent variables, TargetWins. However, their p-values between the two models are the same.

SAS Code is at **Appendix L – SAS Code for PROC GENMOD**

Figure 33 – Comparison of Parameter Estimates between PROC REG and PROC GLM

Parameter Estimates – STEPWISE - FROM PROC REG

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	134.42779	15.1905	8.85	<.0001	0
BaseHitsByBattersAllBases_P	1	0.05043	0.00244	20.68	<.0001	1.8498
X1_WalksByBatters_P	1	24.34168	3.86446	6.3	<.0001	4.92427
X1_Errors_N	1	-47.30128	2.71993	-17.39	<.0001	7.60353
X1_IMP_CaughtStealing_N	1	-10.22857	2.42509	-4.22	<.0001	1.58362
MFlag_CaughtStealing_N	1	4.10759	0.85137	4.82	<.0001	2.41947
X2_TriplesByBatters3Bases_P	1	3.1484	0.44296	7.11	<.0001	2.72463
X2_IMP_StolenBases_P	1	57.53961	6.46671	8.9	<.0001	13.3313
MFlag_StolenBases_P	1	30.64349	1.80309	17	<.0001	2.62632
X2_IMP_BattersHitByPitch_P	1	2.18359	1.17147	1.86	0.0625	1.04573
MFlag_BattersHitByPitch_P	1	7.47316	1.06106	7.04	<.0001	1.28894
INT_P	1	-7.15E-10	2.15E-10	-3.32	0.0009	13.1261
INT_N	1	-1.42E-07	7.43E-08	-1.92	0.0552	3.01464

Analysis Of Maximum Likelihood Parameter Estimates – STEPWISE - FROM PROC GENMOD

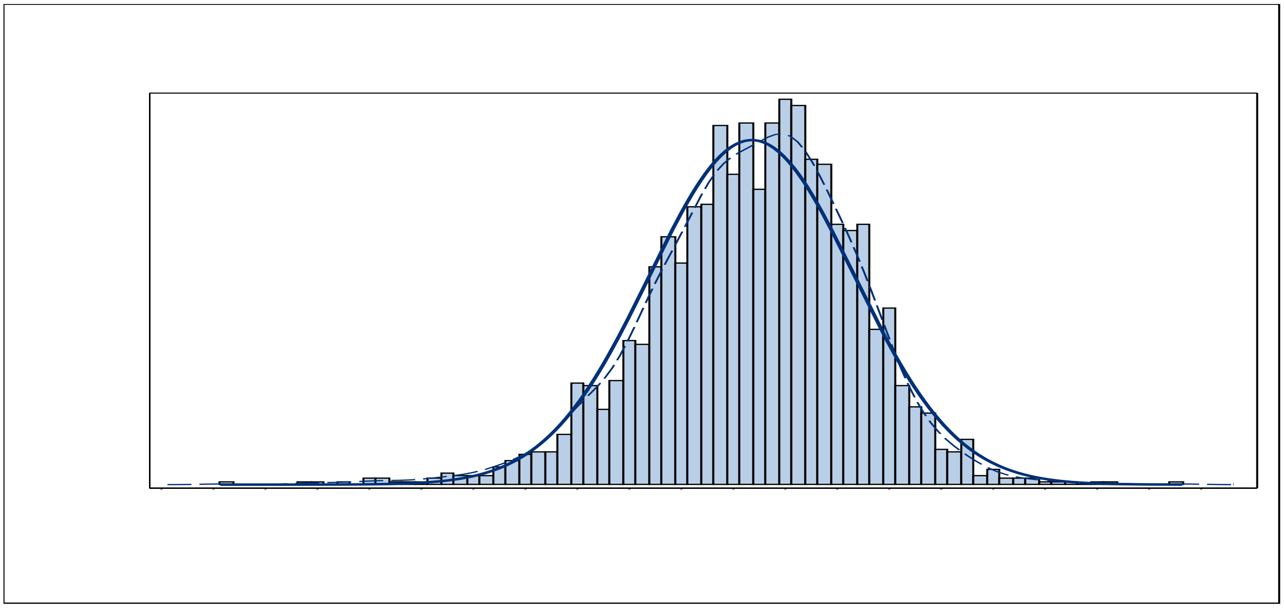
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	134.428	15.147	104.74	164.115	78.76	<.0001
BaseHitsByBattersAll	1	0.0504	0.0024	0.0457	0.0552	430.28	<.0001
X1_WalksByBatters_P	1	24.3417	3.8534	16.7891	31.8942	39.9	<.0001
X1_Errors_N	1	-47.3013	2.7121	-52.617	-41.9856	304.17	<.0001
X1_IMP_CaughtStealin	1	-10.2286	2.4182	-14.9681	-5.4891	17.89	<.0001
MFlag_CaughtStealing	1	4.1076	0.8489	2.4437	5.7715	23.41	<.0001
X2_TriplesByBatters3	1	3.1484	0.4417	2.2827	4.0141	50.81	<.0001
X2_IMP_StolenBases_P	1	57.5396	6.4482	44.9013	70.1779	79.63	<.0001
MFlag_StolenBases_P	1	30.6435	1.7979	27.1196	34.1674	290.49	<.0001
X2_IMP_BattersHitByP	1	2.1836	1.1681	-0.1059	4.4731	3.49	0.0616
MFlag_BattersHitByPi	1	7.4732	1.058	5.3995	9.5469	49.89	<.0001
INT_P	1	0	0	0	0	11.09	0.0009
INT_N	1	0	0	0	0	3.7	0.0544
Scale	1	12.3271	0.1827	11.9742	12.6905		

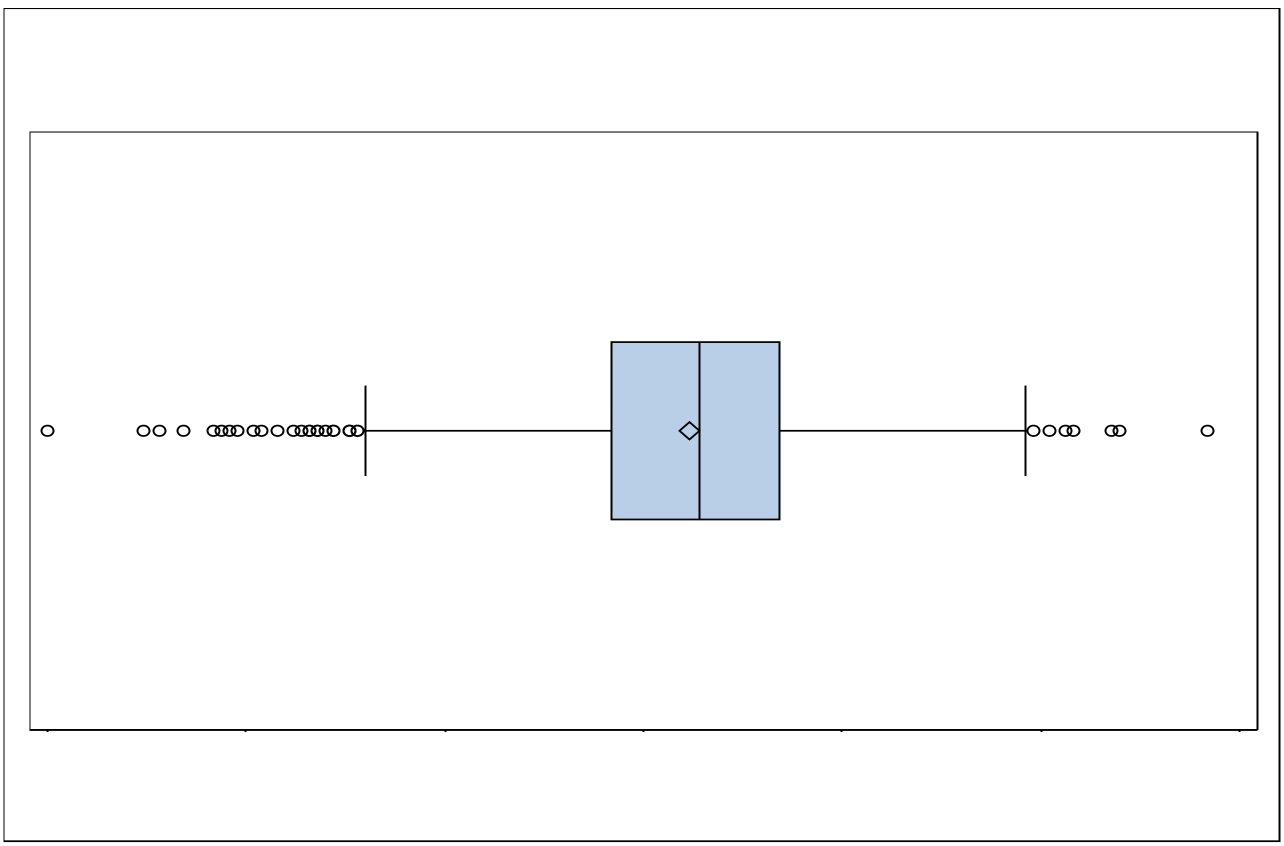
Appendix A - Imputation Results of Missing Value and Missing Flag setting

Obs	StrikeoutsByBatters_N	IMP_StrikeoutsByBatters_N	MFlag_StrikeoutsByBatters_N	StolenBases_P	IMP_StolenBases_P	MFlag_StolenBases_P	CaughtStealing_N	IMP_CaughtStealing_N	MFlag_CaughtStealing_N	BattersHitByPitch_P	IMP_BattersHitByPitch_P	MFlag_BattersHitByPitch_P	StrikeoutsByPitchers_P	IMP_StrikeoutsByPitchers_P	MFlag_StrikeoutsByPitchers_P	DoublePlays_P	IMP_DoublePlays_P	MFlag_DoublePlays_P
1	842.00	842.00	0	.	124.76	1	.	52.80	1	.	59.36	1	5456.00	5456.00	0	.	146.39	1
2	1075.00	1075.00	0	37.00	37.00	0	28.00	28.00	0	.	59.36	1	1082.00	1082.00	0	155.00	155.00	0
3	917.00	917.00	0	46.00	46.00	0	27.00	27.00	0	.	59.36	1	917.00	917.00	0	153.00	153.00	0
4	922.00	922.00	0	43.00	43.00	0	30.00	30.00	0	.	59.36	1	928.00	928.00	0	156.00	156.00	0
5	920.00	920.00	0	49.00	49.00	0	39.00	39.00	0	.	59.36	1	920.00	920.00	0	168.00	168.00	0
6	973.00	973.00	0	107.00	107.00	0	59.00	59.00	0	.	59.36	1	973.00	973.00	0	149.00	149.00	0
7	1062.00	1062.00	0	80.00	80.00	0	54.00	54.00	0	.	59.36	1	1062.00	1062.00	0	186.00	186.00	0
8	1027.00	1027.00	0	40.00	40.00	0	36.00	36.00	0	.	59.36	1	1033.00	1033.00	0	136.00	136.00	0
9	922.00	922.00	0	69.00	69.00	0	27.00	27.00	0	.	59.36	1	922.00	922.00	0	169.00	169.00	0
10	827.00	827.00	0	72.00	72.00	0	34.00	34.00	0	.	59.36	1	827.00	827.00	0	159.00	159.00	0
11	888.00	888.00	0	60.00	60.00	0	39.00	39.00	0	.	59.36	1	928.00	928.00	0	141.00	141.00	0
12	801.00	801.00	0	119.00	119.00	0	79.00	79.00	0	.	59.36	1	801.00	801.00	0	150.00	150.00	0
13	816.00	816.00	0	221.00	221.00	0	109.00	109.00	0	.	59.36	1	821.00	821.00	0	165.00	165.00	0
14	812.00	812.00	0	126.00	126.00	0	80.00	80.00	0	.	59.36	1	812.00	812.00	0	139.00	139.00	0
15	880.00	880.00	0	159.00	159.00	0	89.00	89.00	0	.	59.36	1	880.00	880.00	0	137.00	137.00	0
16	682.00	682.00	0	86.00	86.00	0	69.00	69.00	0	.	59.36	1	682.00	682.00	0	136.00	136.00	0
17	843.00	843.00	0	100.00	100.00	0	53.00	53.00	0	.	59.36	1	843.00	843.00	0	172.00	172.00	0
18	900.00	900.00	0	92.00	92.00	0	64.00	64.00	0	.	59.36	1	911.00	911.00	0	146.00	146.00	0
19	841.00	841.00	0	65.00	65.00	0	49.00	49.00	0	.	59.36	1	1239.00	1239.00	0	177.00	177.00	0
20	760.00	760.00	0	55.00	55.00	0	53.00	53.00	0	.	59.36	1	760.00	760.00	0	171.00	171.00	0
21	835.00	835.00	0	41.00	41.00	0	39.00	39.00	0	.	59.36	1	835.00	835.00	0	190.00	190.00	0
22	928.00	928.00	0	80.00	80.00	0	51.00	51.00	0	.	59.36	1	928.00	928.00	0	170.00	170.00	0
23	902.00	902.00	0	106.00	106.00	0	51.00	51.00	0	.	59.36	1	902.00	902.00	0	202.00	202.00	0
24	860.00	860.00	0	109.00	109.00	0	42.00	42.00	0	.	59.36	1	860.00	860.00	0	156.00	156.00	0
25	926.00	926.00	0	125.00	125.00	0	44.00	44.00	0	.	59.36	1	926.00	926.00	0	162.00	162.00	0
26	819.00	819.00	0	86.00	86.00	0	52.00	52.00	0	.	59.36	1	819.00	819.00	0	175.00	175.00	0
27	1011.00	1011.00	0	89.00	89.00	0	40.00	40.00	0	.	59.36	1	1011.00	1011.00	0	173.00	173.00	0
28	1000.00	1000.00	0	69.00	69.00	0	43.00	43.00	0	.	59.36	1	1000.00	1000.00	0	186.00	186.00	0
29	928.00	928.00	0	94.00	94.00	0	56.00	56.00	0	.	59.36	1	928.00	928.00	0	156.00	156.00	0
30	882.00	882.00	0	160.00	160.00	0	101.00	101.00	0	.	59.36	1	882.00	882.00	0	172.00	172.00	0
31	930.00	930.00	0	169.00	169.00	0	100.00	100.00	0	.	59.36	1	930.00	930.00	0	161.00	161.00	0
32	1007.00	1007.00	0	92.00	92.00	0	76.00	76.00	0	.	59.36	1	1419.00	1419.00	0	155.00	155.00	0
33	993.00	993.00	0	65.00	65.00	0	44.00	44.00	0	.	59.36	1	1109.00	1109.00	0	134.00	134.00	0
34	980.00	980.00	0	53.00	53.00	0	39.00	39.00	0	.	59.36	1	986.00	986.00	0	157.00	157.00	0
35	953.00	953.00	0	126.00	126.00	0	72.00	72.00	0	.	59.36	1	953.00	953.00	0	140.00	140.00	0
36	1028.00	1028.00	0	93.00	93.00	0	45.00	45.00	0	.	59.36	1	1028.00	1028.00	0	146.00	146.00	0
37	1022.00	1022.00	0	71.00	71.00	0	45.00	45.00	0	.	59.36	1	1022.00	1022.00	0	156.00	156.00	0
38	1024.00	1024.00	0	93.00	93.00	0	52.00	52.00	0	47.00	47.00	0	1024.00	1024.00	0	184.00	184.00	0
39	1001.00	1001.00	0	116.00	116.00	0	52.00	52.00	0	77.00	77.00	0	1001.00	1001.00	0	142.00	142.00	0
40	805.00	805.00	0	117.00	117.00	0	51.00	51.00	0	74.00	74.00	0	805.00	805.00	0	151.00	151.00	0
41	838.00	838.00	0	129.00	129.00	0	61.00	61.00	0	56.00	56.00	0	838.00	838.00	0	138.00	138.00	0
42	942.00	942.00	0	143.00	143.00	0	46.00	46.00	0	73.00	73.00	0	942.00	942.00	0	126.00	126.00	0
43	848.00	848.00	0	161.00	161.00	0	57.00	57.00	0	29.00	29.00	0	848.00	848.00	0	139.00	139.00	0
44	1239.00	1239.00	0	73.00	73.00	0	38.00	38.00	0	.	59.36	1	1239.00	1239.00	0	125.00	125.00	0
45	1045.00	1045.00	0	137.00	137.00	0	39.00	39.00	0	.	59.36	1	1045.00	1045.00	0	129.00	129.00	0
46	975.00	975.00	0	97.00	97.00	0	44.00	44.00	0	59.00	59.00	0	975.00	975.00	0	138.00	138.00	0
47	1052.00	1052.00	0	71.00	71.00	0	38.00	38.00	0	57.00	57.00	0	1052.00	1052.00	0	148.00	148.00	0
48	1016.00	1016.00	0	92.00	92.00	0	46.00	46.00	0	50.00	50.00	0	1016.00	1016.00	0	116.00	116.00	0
49	1006.00	1006.00	0	76.00	76.00	0	38.00	38.00	0	45.00	45.00	0	1006.00	1006.00	0	132.00	132.00	0
50	1022.00	1022.00	0	53.00	53.00	0	32.00	32.00	0	35.00	35.00	0	1022.00	1022.00	0	144.00	144.00	0
51	1094.00	1094.00	0	67.00	67.00	0	26.00	26.00	0	55.00	55.00	0	1094.00	1094.00	0	159.00	159.00	0
52	965.00	965.00	0	76.00	76.00	0	30.00	30.00	0	67.00	67.00	0	965.00	965.00	0	172.00	172.00	0
53	99.00	99.00	0	.	124.76	1	.	52.80	1	.	59.36	1	272.00	272.00	0	88.00	88.00	0
54	227.00	227.00	0	.	124.76	1	.	52.80	1	.	59.36	1	525.00	525.00	0	97.00	97.00	0
55	327.00	327.00	0	.	124.76	1	.	52.80	1	.	59.36	1	883.00	883.00	0	97.00	97.00	0
56	428.00	428.00	0	.	124.76	1	.	52.80	1	.	59.36	1	825.00	825.00	0	112.00	112.00	0
57	426.00	426.00	0	.	124.76	1	.	52.80	1	.	59.36	1	822.00	822.00	0	104.00	104.00	0
58	471.00	471.00	0	.	124.76	1	.	52.80	1	.	59.36	1	908.00	908.00	0	71.00	71.00	0
59	699.00	699.00	0	.	124.76	1	.	52.80	1	.	59.36	1	1155.00	1155.00	0	.	146.39	1
60	963.00	963.00	0	.	124.76	1	.	52.80	1	.	59.36	1	1405.00	1405.00	0	.	146.39	1
61	755.00	755.00	0	.	124.76	1	.	52.80	1	.	59.36	1	1092.00	1092.00	0	.	146.39	1
62	744.00	744.00	0	216.00	216.00	0	.	52.80	1	.	59.36	1	1030.00	1030.00	0	.	146.39	1
63	525.00	525.00	0	499.00	499.00	0	.	52.80	1	.	59.36	1	703.00	703.00	0	.	146.39	1
64	633.00	633.00	0	354.00	354.00	0	.	52.80	1	.	59.36	1	765.00	765.00	0	.	146.39	1
65	570.00	570.00	0	419.00	419.00	0	.	52.80	1	.	59.36	1	721.00	721.00	0	.	146.39	1
66	627.00	627.00	0	347.00	347.00	0	.	52.80	1	.	59.36	1	764.00	764.00	0	.	146.39	1
67	632.00	632.00	0	339.00	339.00	0	.	52.80	1	.	59.36	1	742.00	742.00	0	.	146.39	1
68	367.00	367.00	0	305.00	305.00	0	.	52.80	1	.	59.36	1	461.00	461.00	0	.	146.39	1
69	320.00	320.00	0	296.00	296.00	0	.	52.80	1	.	59.36	1	393.00	393.00	0	.	146.39	1
70	292.00	292.00	0	246.00	246.00	0	.	52.80	1	.	59.36	1	361.00	361.00	0	.	146.39	1
71	339.00	339.00	0	298.00	298.00	0	.	52.80	1	.	59.36	1	419.00	419.00	0	.	146.39	1
72	322.00	322.00	0	286.00	286.00	0	.	52.80	1	.	59.36	1	395.00	395.00	0	.	146.39	1
73	329.00	329.00	0	187.00	187.00	0	.	52.80	1	.	59.36	1	358.00	358.00	0	.	146.39	1
74	287.00	287.00	0	197.00	197.00	0	.	52.80	1	.	59.36	1	306.00	306.00	0	.	146.39	1
75	326.00	326.00	0	214.00	214.00	0	.	52.80	1	.	59.36	1	383.00	383.00	0	101.00	101.00	0
76	569.00	569.00	0	223.00	223.00	0	.	52.80	1	.	59.36	1	673.00	673.00	0	106.00	106.00	0
77	.	735.61	1	187.00	187.00	0	.	52.80	1	.	59.36	1	.	817.73	1	104.00	104.00	0
78	.	735.61	1	151.00	151.00	0	.	52.80	1	.	59.36	1	.	817.73	1	96.00	96.00	0
79	.	735.61	1	139.00	139.00	0	.	52.80	1	.	59.36	1	.	817.73	1	94.00	94.00	0
80	.	735.61	1	100.00	100.00	0	.	52.80	1	.	59.36	1	.	817.73	1	109.00	109.00	0
81	.	735.61	1	129.00	129.00	0	.	52.80	1	.	59.36	1	.	817.73	1	140.00	140.00	0
82	.	735.61	1	141.00	141.00	0	.	52.80	1	.	59.36	1	.	817.73	1	95.00	95.00	0
83	.	735.61	1	143.00	143.00	0	.	52.80	1	.	59.36	1	.	817.73	1	107.00	107.00	0
84	572.00	572.00	0	161.00	161.00	0	.	52.80	1	.	59.36	1	606.00	606.00	0	145.00	145.00	0
85	619.00	619.00	0	181.00	181.00	0	.	52.80	1	.	59.36	1	664.00	664.00	0	118.00	118.00	0
86	731.00	731.00	0	145.00	145.00	0	.	52.80	1	.	59.36	1	774.00	774.00	0	137.00	137.00	0
87	687.00	687.00	0	190.00	190.00	0	.	52.80	1	.	59.36	1	737.00	737.00	0	88.00	88.00	0
88	653.00	653.00	0	147.00	147.00	0	.	52.80	1	.	59.36	1	691.00	691.00	0	151.00	151.00	0
89	661.00	661.00	0	129.00	129.00	0	104.00	104.00	0	.	59.36	1	704.00	704.00	0	123.00	123.00	0
90	622.00	622.00	0	164.00	164.00	0	.	52.80	1	.	59.36	1	659.00	659.00	0	129.00	129.00	0
91	572.00	572.00	0	108.00	108.00	0	.	52.80	1	.	59.36	1	747.00	747.00	0	116.00	116.00	0
92	561.00	561.00	0	169.00	169.00	0	.	52.80	1	.	59.36	1	654.00	654.00	0	129.00	129.00	0
93	520.00	520.00	0	94.00	94.00	0	104.00	104.00	0	.	59.36	1	554.00	554.00	0	133.00	133.00	0
94	498.00	498.00	0	100.00	100.00	0	106.00	106.00	0	.	59.36	1	527.00	527.00	0	129.00	129.00	0
95	478.00	478.00	0	71.00	71.00	0	69.00	69.00	0	.	59.36	1	506.00	506.00	0	128.00	128.00	0
96	425.00	425.00	0	60.00	60.00	0	84.00	84.00	0	.	59.36	1	447.00	447.00	0	165.00	165.00	0
97	478.00	478.00	0	78.00	78.00	0	72.00	72.00	0	.	59.36	1	506.00	506.00	0	163.00	163.00	0
98	402.00	402.00	0	82.00	82.00	0	76.00	76.00	0	.	59.36	1	426.00	426.00	0	154.00	154.00	0
99	371.00	371.00	0	86.00	86.00	0	.	52.80	1	.	59.36	1	395.00	395.00	0	160.00	160.00	0
100	382.00	382.00	0	105.00	105.00	0	.	52.80	1	.	59.36	1	402.00	402.00	0	137.00	137.00	0
101	399.00	399.00	0	64.00	64.00	0	.	52.80	1	.	59.36	1	422.00	422.00	0	149.00	149.00	0
102	454.00	454.00	0	68.00	68.00	0	.	52.80	1	.	59.36	1	478.00	478.00	0	154.00	154.00	0
103	418.00	418.00	0	73.00	73.00	0	.	52.80	1	.	59.36	1	440.00	440.00	0	176.00	176.00	0
104	452.00	452.00	0	48.00	48.00	0	.	52.80	1	.	59.36	1	475.00	475.00	0	148.00	148.00	0
105	522.00	522.00	0	38.00	38.00	0	.	52.80	1	.	59.36	1	549.00	549.00	0	153.00	153.00	0
106	450.00	450.00	0	26.00	26.00	0	.	52.80	1	.	59.36	1	473.00	473.00	0	156.00	156.00	0
107	472.00	472.00	0	32.00	32.00	0	.	52.80	1	.	59.36	1	506.00	506.00	0	129.00	129.00	0
108	462.00	462.00	0	21.00	21.00	0	.	52.80	1	.	59.36	1	489.00	489.00	0	107.00	107.00	0
109	612.00	612.00	0	24.00	24.00	0	.	52.80	1	.	59.36	1	644.00	644.00	0	184.00	184.00	0
110	754.00	754.00	0	48.00	48.00	0	.	52.80	1	.	59.36	1	804.00	804.00	0	136.00	136.00	0
111	530.00	530.00	0	44.00	44.00	0	.	52.80	1	.	59.36	1	569.00	569.00	0	191.00	191.00	0
112	619.00	619.00	0	51.00	51.00	0	.	52.80	1	.	59.36	1	660.00	660.00	0	180.00	180.00	0
113	555.00	555.00	0	54.00	54.00	0	.	52.80	1	.	59.36	1	608.00	608.00	0	151.00	151.00	0
114	645.00	645.00	0	59.00	59.00	0	.	52.80	1	.	59.36	1	683.00	683.00	0	147.00	147.00	0
115	535.00	535.00	0	39.00	39.00	0	.	52.80	1	.	59.36	1	563.00	563.00	0	168.00	168.00	0
116	544.00	544.00	0	87.00	87.00	0	.	52.80	1	.	59.36	1	580.00	580.00	0	171.00	171.00	0
117	496.00	496.00	0	64.00	64.00	0	.	52.80	1	.	59.36	1	525.00	525.00	0	137.00	137.00	0
118	526.00	526.00	0	61.00	61.00	0	.	52.80	1	.	59.36	1	553.00	553.00	0	130.00	130.00	0
119	568.00	568.00	0	46.00	46.00	0	.	52.80	1	.	59.36	1	601.00	601.00	0	140.00	140.00	0
120	690.00	690.00	0	29.00	29.00	0	.	52.80	1	.	59.36	1	726.00	726.00	0	151.00	151.00	0
121	648.00	648.00	0	75.00	75.00	0	.	52.80	1	.	59.36	1	682.00	682.00	0	154.00	154.00	0
122	649.00	649.00	0	84.00	84.00	0	36.00	36.00	0	.	59.36	1	683.00	683.00	0	165.00	165.00	0
123	753.00	753.00	0	61.00	61.00	0	36.00	36.00	0	.	59.36	1	797.00	797.00	0	151.00	151.00	0
124	651.00	651.00	0	57.00	57.00	0	33.00	33.00	0	.	59.36	1	685.00	685.00	0	180.00	180.00	0
125	773.00	773.00	0	44.00	44.00	0	28.00	28.00	0	.	59.36	1	813.00	813.00	0	163.00	163.00	0
126	767.00	767.00	0	37.00	37.00	0	17.00	17.00	0	.	59.36	1	807.00	807.00	0	182.00	182.00	0
127	794.00	794.00	0	43.00	43.00	0	15.00	15.00	0	.	59.36	1	825.00	825.00	0	143.00	143.00	0
128	834.00	834.00	0	73.00	73.00	0	39.00	39.00	0	.	59.36	1	877.00	877.00	0	144.00	144.00	0
129	926.00	926.00	0	74.00	74.00	0	45.00	45.00	0	.	59.36	1	974.00	974.00	0	160.00	160.00	0
130	975.00	975.00	0	57.00	57.00	0	27.00	27.00	0	.	59.36	1	975.00	975.00	0	154.00	154.00	0
131	954.00	954.00	0	75.00	75.00	0	52.00	52.00	0	.	59.36	1	954.00	954.00	0	161.00	161.00	0
132	825.00	825.00	0	53.00	53.00	0	41.00	41.00	0	.	59.36	1	825.00	825.00	0	139.00	139.00	0
133	976.00	976.00	0	64.00	64.00	0	37.00	37.00	0	.	59.36	1	976.00	976.00	0	145.00	145.00	0
134	913.00	913.00	0	59.00	59.00	0	47.00	47.00	0	.	59.36	1	913.00	913.00	0	139.00	139.00	0
135	947.00	947.00	0	55.00	55.00	0	45.00	45.00	0	.	59.36	1	947.00	947.00	0	148.00	148.00	0
136	782.00	782.00	0	83.00	83.00	0	44.00	44.00	0	.	59.36	1	782.00	782.00	0	139.00	139.00	0
137	736.00	736.00	0	58.00	58.00	0	34.00	34.00	0	.	59.36	1	736.00	736.00	0	118.00	118.00	0
138	810.00	810.00	0	49.00	49.00	0	37.00	37.00	0	.	59.36	1	852.00	852.00	0	137.00	137.00	0
139	875.00	875.00	0	85.00	85.00	0	40.00	40.00	0	.	59.36	1	880.00	880.00	0	143.00	143.00	0
140	772.00	772.00	0	72.00	72.00	0	44.00	44.00	0	.	59.36	1	772.00	772.00	0	161.00	161.00	0
141	764.00	764.00	0	55.00	55.00	0	38.00	38.00	0	.	59.36	1	769.00	769.00	0	148.00	148.00	0
142	811.00	811.00	0	74.00	74.00	0	61.00	61.00	0	.	59.36	1	811.00	811.00	0	151.00	151.00	0
143	876.00	876.00	0	82.00	82.00	0	53.00	53.00	0	.	59.36	1	876.00	876.00	0	127.00	127.00	0
144	874.00	874.00	0	90.00	90.00	0	65.00	65.00	0	.	59.36	1	874.00	874.00	0	126.00	126.00	0
145	828.00	828.00	0	99.00	99.00	0	51.00	51.00	0	.	59.36	1	838.00	838.00	0	141.00	141.00	0
146	905.00	905.00	0	73.00	73.00	0	52.00	52.00	0	.	59.36	1	911.00	911.00	0	157.00	157.00	0
147	825.00	825.00	0	150.00	150.00	0	60.00	60.00	0	.	59.36	1	1261.00	1261.00	0	142.00	142.00	0
148	869.00	869.00	0	151.00	151.00	0	77.00	77.00	0	.	59.36	1	869.00	869.00	0	186.00	186.00	0
149	847.00	847.00	0	146.00	146.00	0	88.00	88.00	0	.	59.36	1	847.00	847.00	0	176.00	176.00	0
150	896.00	896.00	0	140.00	140.00	0	85.00	85.00	0	.	59.36	1	896.00	896.00	0	153.00	153.00	0
151	849.00	849.00	0	72.00	72.00	0	52.00	52.00	0	.	59.36	1	849.00	849.00	0	197.00	197.00	0
152	910.00	910.00	0	94.00	94.00	0	76.00	76.00	0	.	59.36	1	916.00	916.00	0	182.00	182.00	0
153	839.00	839.00	0	136.00	136.00	0	68.00	68.00	0	.	59.36	1	844.00	844.00	0	171.00	171.00	0
154	859.00	859.00	0	96.00	96.00	0	70.00	70.00	0	.	59.36	1	870.00	870.00	0	140.00	140.00	0
155	1010.00	1010.00	0	92.00	92.00	0	55.00	55.00	0	.	59.36	1	1010.00	1010.00	0	133.00	133.00	0
156	906.00	906.00	0	165.00	165.00	0	76.00	76.00	0	.	59.36	1	906.00	906.00	0	122.00	122.00	0
157	924.00	924.00	0	126.00	126.00	0	60.00	60.00	0	.	59.36	1	924.00	924.00	0	121.00	121.00	0
158	946.00	946.00	0	125.00	125.00	0	48.00	48.00	0	.	59.36	1	946.00	946.00	0	146.00	146.00	0
159	949.00	949.00	0	68.00	68.00	0	44.00	44.00	0	.	59.36	1	1349.00	1349.00	0	121.00	121.00	0
160	1050.00	1050.00	0	82.00	82.00	0	48.00	48.00	0	.	59.36	1	1181.00	1181.00	0	127.00	127.00	0
161	1032.00	1032.00	0	83.00	83.00	0	43.00	43.00	0	.	59.36	1	1032.00	1032.00	0	143.00	143.00	0
162	1160.00	1160.00	0	108.00	108.00	0	58.00	58.00	0	.	59.36	1	1160.00	1160.00	0	136.00	136.00	0
163	1062.00	1062.00	0	98.00	98.00	0	43.00	43.00	0	.	59.36	1	1062.00	1062.00	0	139.00	139.00	0
164	962.00	962.00	0	148.00	148.00	0	66.00	66.00	0	.	59.36	1	962.00	962.00	0	122.00	122.00	0
165	1010.00	1010.00	0	148.00	148.00	0	56.00	56.00	0	59.00	59.00	0	1010.00	1010.00	0	138.00	138.00	0
166	1039.00	1039.00	0	85.00	85.00	0	46.00	46.00	0	45.00	45.00	0	1039.00	1039.00	0	133.00	133.00	0
167	933.00	933.00	0	68.00	68.00	0	22.00	22.00	0	49.00	49.00	0	933.00	933.00	0	166.00	166.00	0
168	1158.00	1158.00	0	86.00	86.00	0	32.00	32.00	0	59.00	59.00	0	1158.00	1158.00	0	171.00	171.00	0
169	1084.00	1084.00	0	92.00	92.00	0	32.00	32.00	0	45.00	45.00	0	1084.00	1084.00	0	170.00	170.00	0
170	1169.00	1169.00	0	52.00	52.00	0	35.00	35.00	0	52.00	52.00	0	1169.00	1169.00	0	146.00	146.00	0
171	454.00	454.00	0	208.00	208.00	0	.	52.80	1	.	59.36	1	537.00	537.00	0	125.00	125.00	0
172	390.00	390.00	0	163.00	163.00	0	.	52.80	1	.	59.36	1	465.00	465.00	0	145.00	145.00	0
173	628.00	628.00	0	118.00	118.00	0	.	52.80	1	.	59.36	1	732.00	732.00	0	110.00	110.00	0
174	649.00	649.00	0	160.00	160.00	0	.	52.80	1	.	59.36	1	692.00	692.00	0	83.00	83.00	0
175	.	735.61	1	240.00	240.00	0	.	52.80	1	.	59.36	1	.	817.73	1	87.00	87.00	0
176	.	735.61	1	153.00	153.00	0	.	52.80	1	.	59.36	1	.	817.73	1	103.00	103.00	0
177	.	735.61	1	134.00	134.00	0	.	52.80	1	.	59.36	1	.	817.73	1	103.00	103.00	0
178	.	735.61	1	147.00	147.00	0	.	52.80	1	.	59.36	1	.	817.73	1	116.00	116.00	0
179	.	735.61	1	178.00	178.00	0	.	52.80	1	.	59.36	1	.	817.73	1	119.00	119.00	0
180	.	735.61	1	133.00	133.00	0	.	52.80	1	.	59.36	1	.	817.73	1	111.00	111.00	0
181	.	735.61	1	185.00	185.00	0	.	52.80	1	.	59.36	1	.	817.73	1	134.00	134.00	0
182	814.00	814.00	0	221.00	221.00	0	.	52.80	1	.	59.36	1	862.00	862.00	0	132.00	132.00	0
183	914.00	914.00	0	247.00	247.00	0	200.00	200.00	0	.	59.36	1	968.00	968.00	0	121.00	121.00	0
184	805.00	805.00	0	212.00	212.00	0	168.00	168.00	0	.	59.36	1	847.00	847.00	0	151.00	151.00	0
185	673.00	673.00	0	246.00	246.00	0	.	52.80	1	.	59.36	1	708.00	708.00	0	126.00	126.00	0
186	568.00	568.00	0	165.00	165.00	0	.	52.80	1	.	59.36	1	598.00	598.00	0	146.00	146.00	0
187	451.00	451.00	0	183.00	183.00	0	.	52.80	1	.	59.36	1	599.00	599.00	0	114.00	114.00	0
188	516.00	516.00	0	86.00	86.00	0	.	52.80	1	.	59.36	1	601.00	601.00	0	114.00	114.00	0
189	359.00	359.00	0	125.00	125.00	0	84.00	84.00	0	.	59.36	1	380.00	380.00	0	126.00	126.00	0
190	428.00	428.00	0	97.00	97.00	0	75.00	75.00	0	.	59.36	1	450.00	450.00	0	134.00	134.00	0
191	401.00	401.00	0	139.00	139.00	0	77.00	77.00	0	.	59.36	1	422.00	422.00	0	166.00	166.00	0
192	451.00	451.00	0	68.00	68.00	0	58.00	58.00	0	.	59.36	1	481.00	481.00	0	155.00	155.00	0
193	372.00	372.00	0	91.00	91.00	0	91.00	91.00	0	.	59.36	1	396.00	396.00	0	150.00	150.00	0
194	489.00	489.00	0	65.00	65.00	0	75.00	75.00	0	.	59.36	1	514.00	514.00	0	176.00	176.00	0
195	445.00	445.00	0	96.00	96.00	0	70.00	70.00	0	.	59.36	1	471.00	471.00	0	176.00	176.00	0
196	504.00	504.00	0	80.00	80.00	0	45.00	45.00	0	.	59.36	1	530.00	530.00	0	154.00	154.00	0
197	579.00	579.00	0	98.00	98.00	0	75.00	75.00	0	.	59.36	1	609.00	609.00	0	160.00	160.00	0
198	610.00	610.00	0	77.00	77.00	0	84.00	84.00	0	.	59.36	1	642.00	642.00	0	168.00	168.00	0
199	555.00	555.00	0	73.00	73.00	0	65.00	65.00	0	.	59.36	1	584.00	584.00	0	164.00	164.00	0
200	597.00	597.00	0	77.00	77.00	0	64.00	64.00	0	.	59.36	1	640.00	640.00	0	174.00	174.00	0

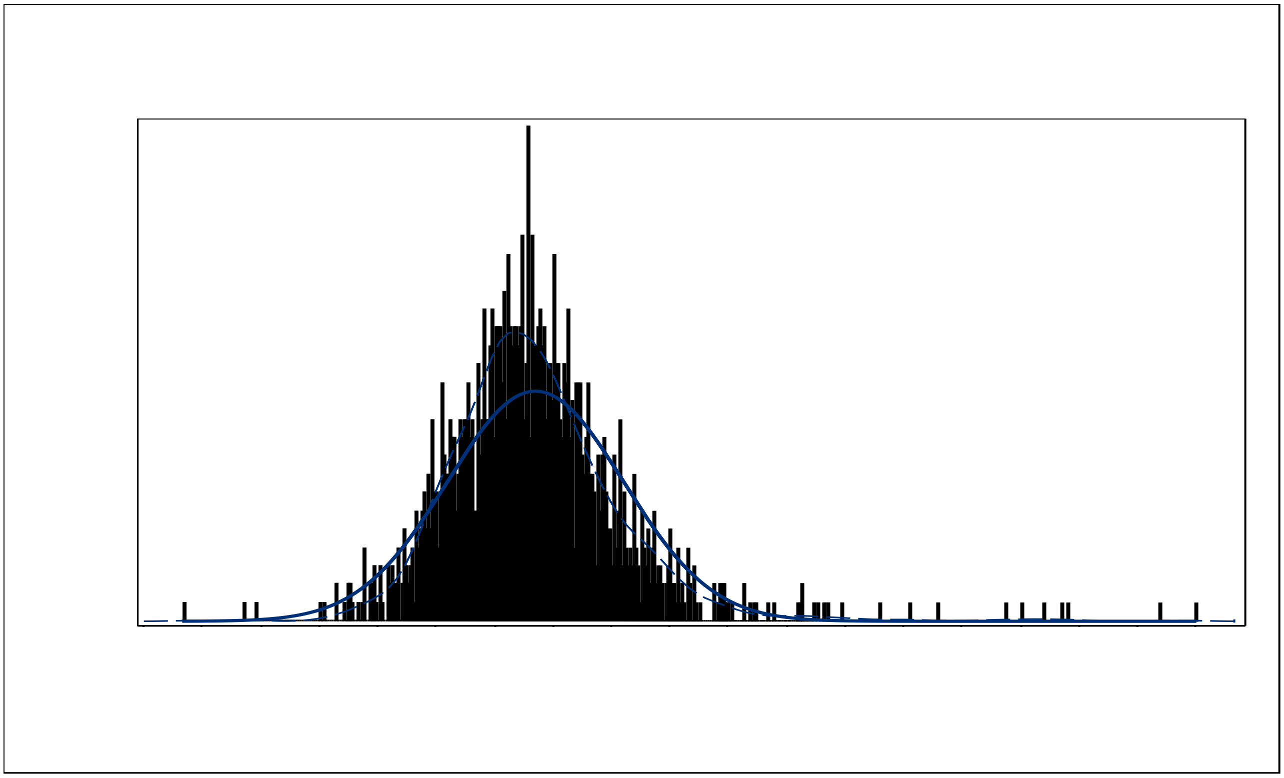
Appendix B – Histogram and Box Plot of Imputed Data Prior to Transformation.

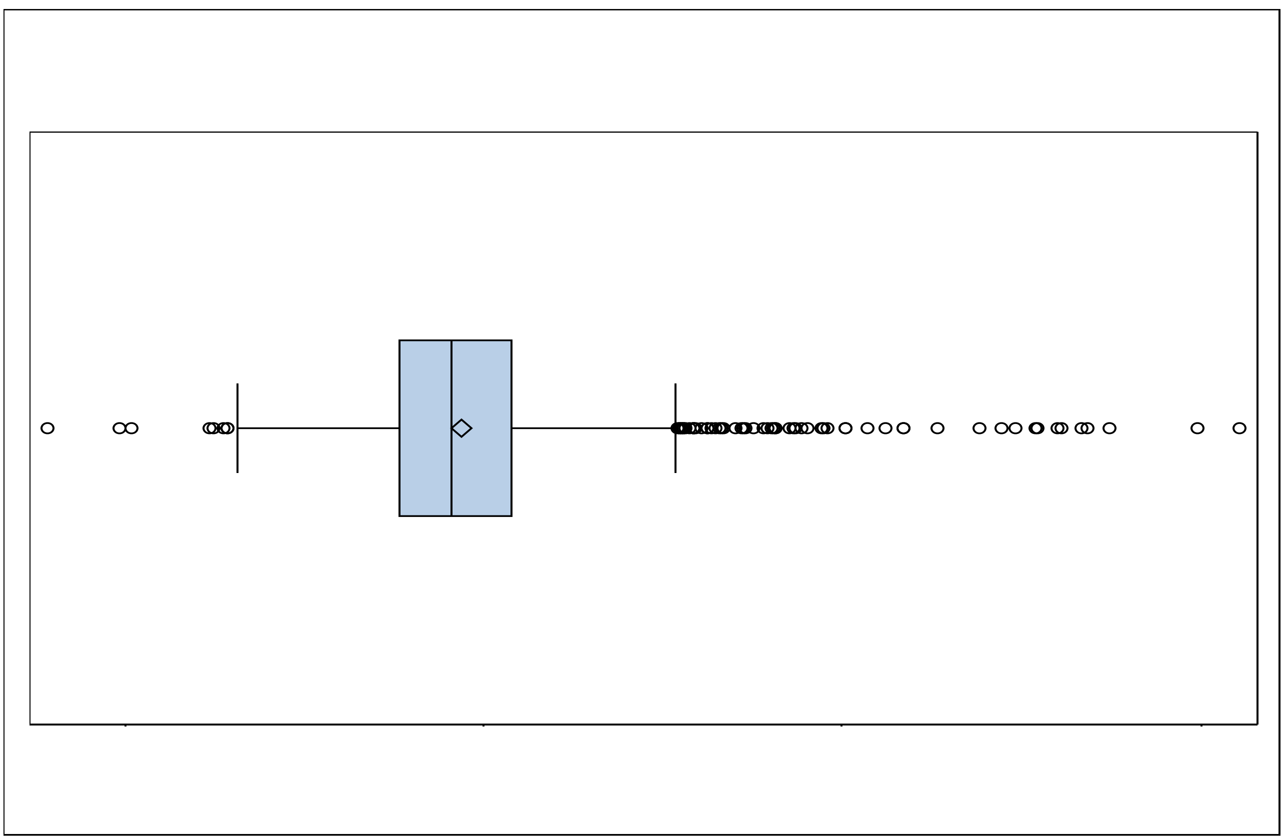
1. TargetWins - Histogram and Box Plot of Imputed Data Prior to Transformation



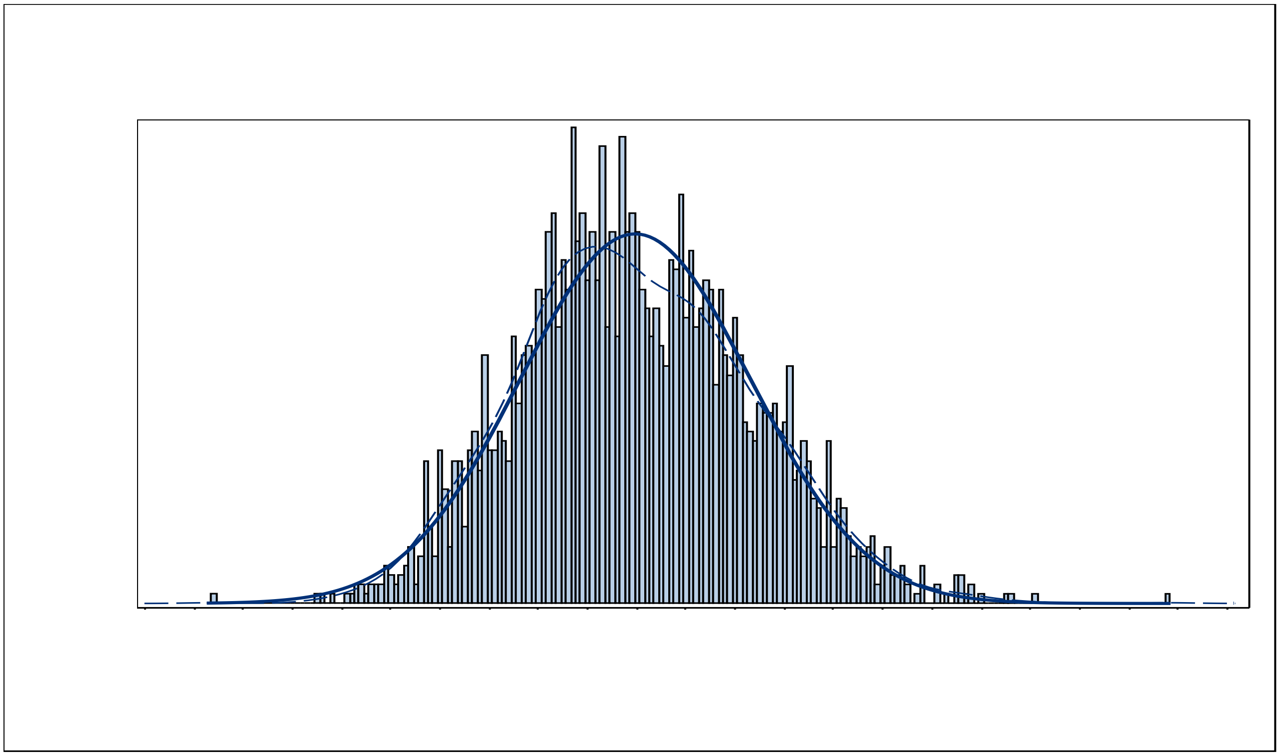


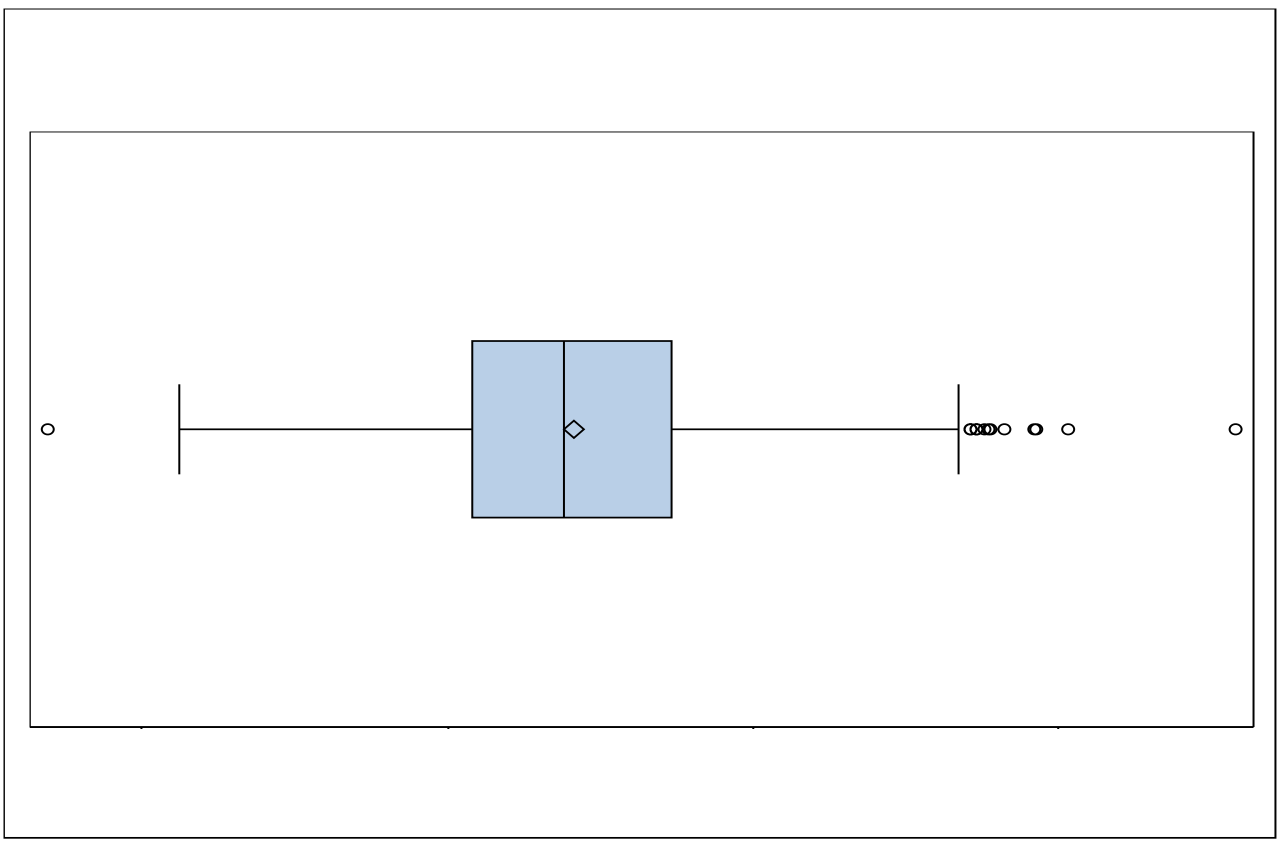
2. BaseHitsByBattersAllBases - Histogram and Box Plot of Imputed Data Prior to Transformation



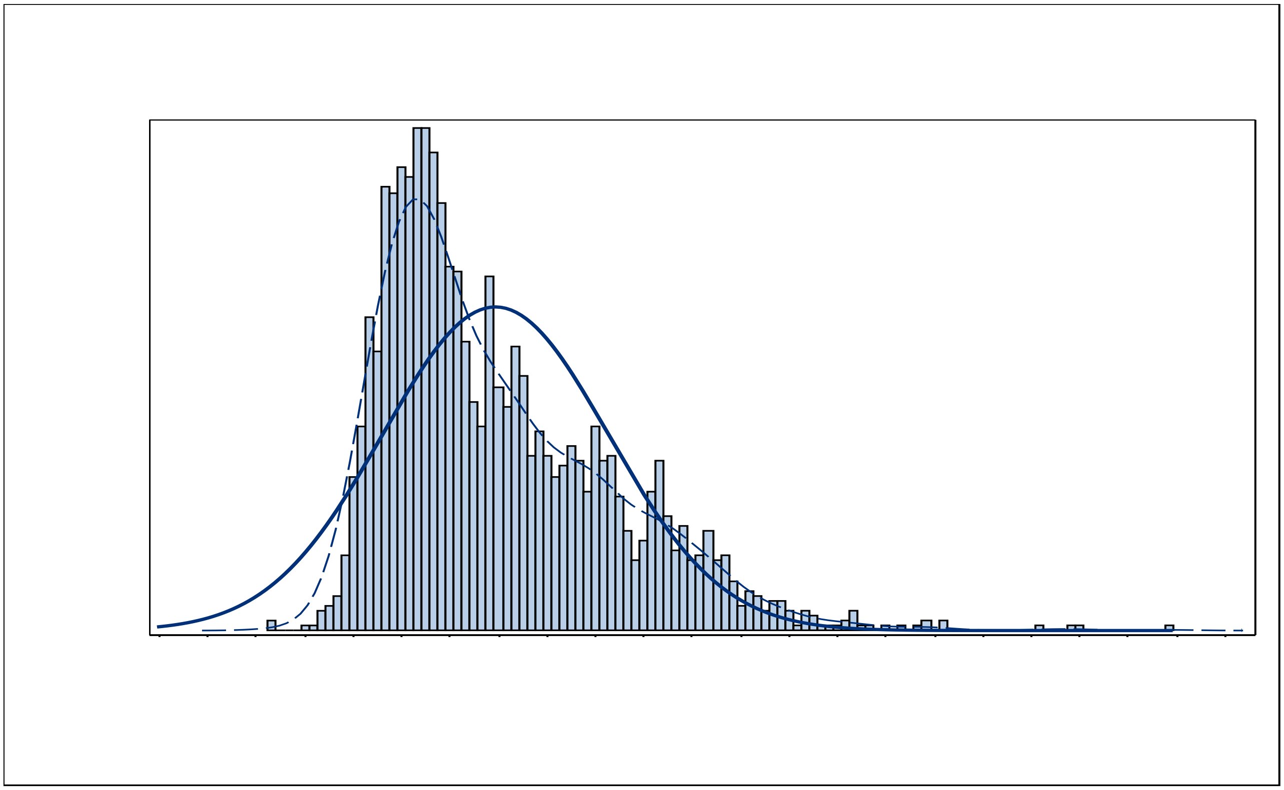


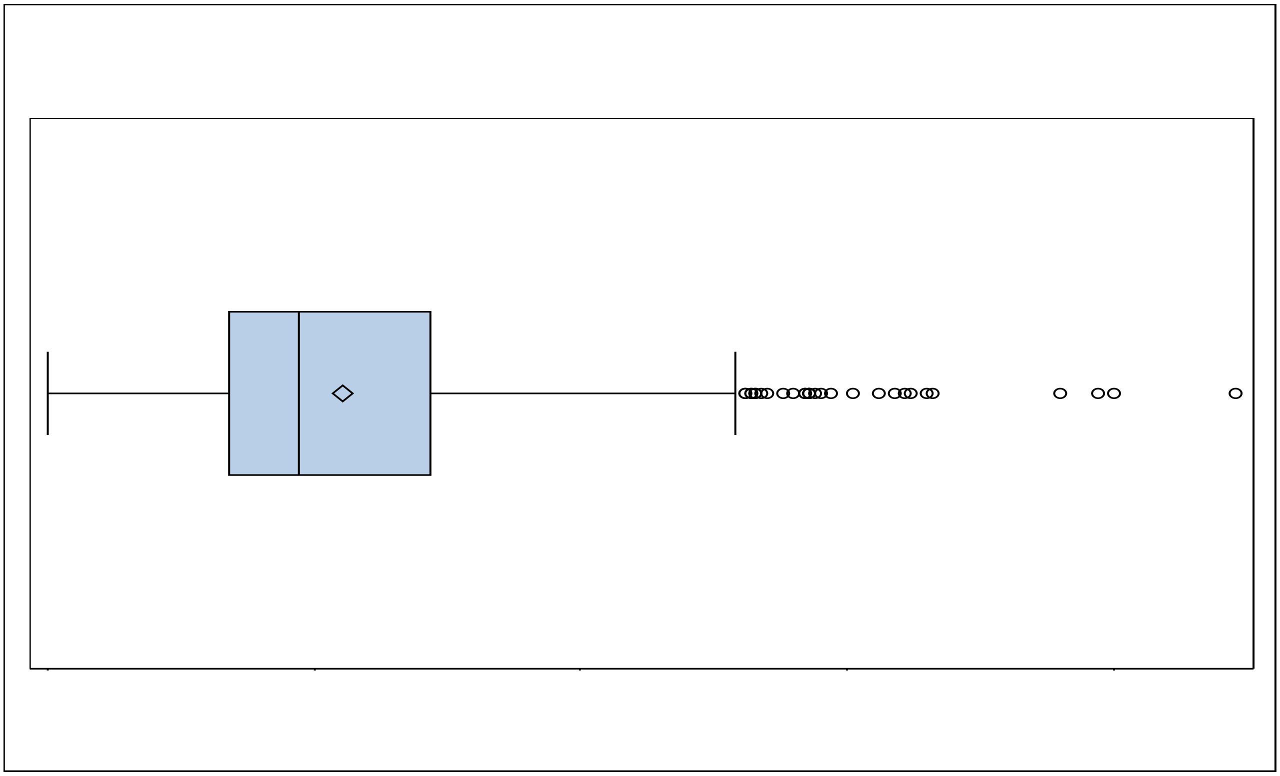
3. DoublesByBatters2Bases - Histogram and Box Plot of Imputed Data Prior to Transformation



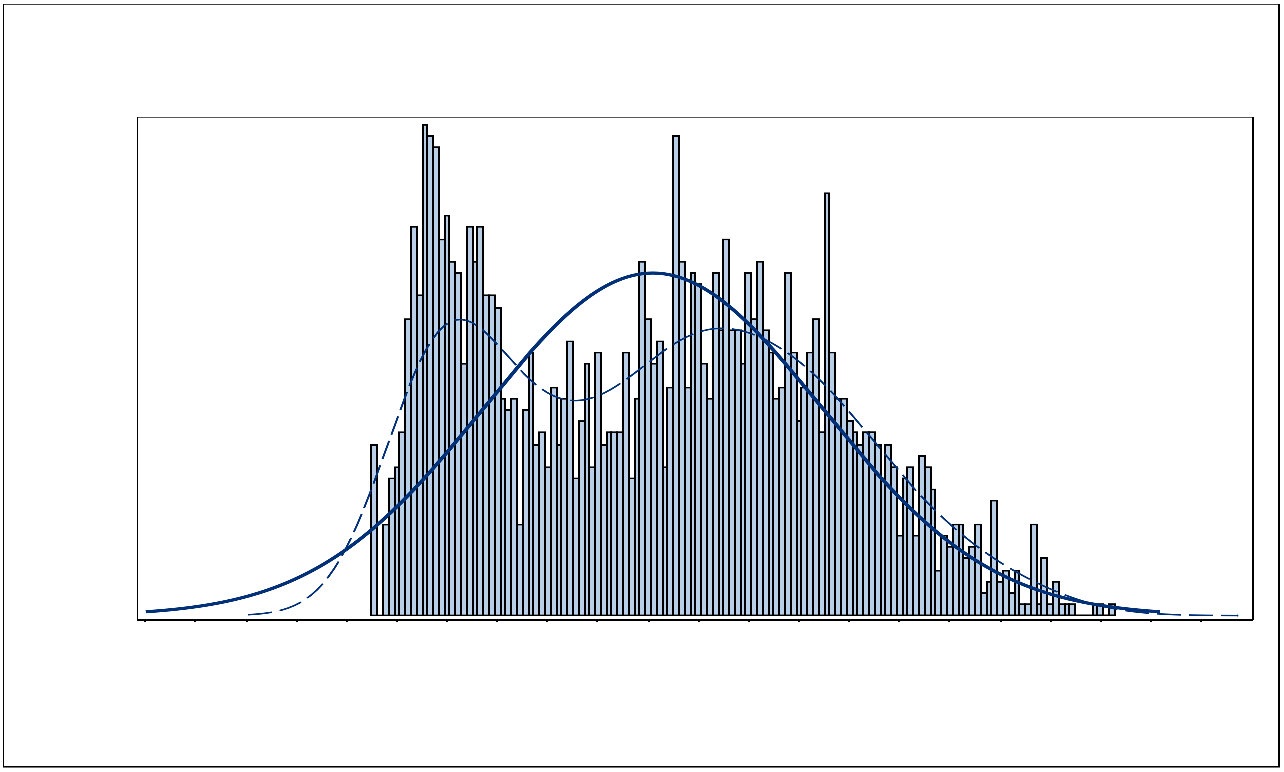


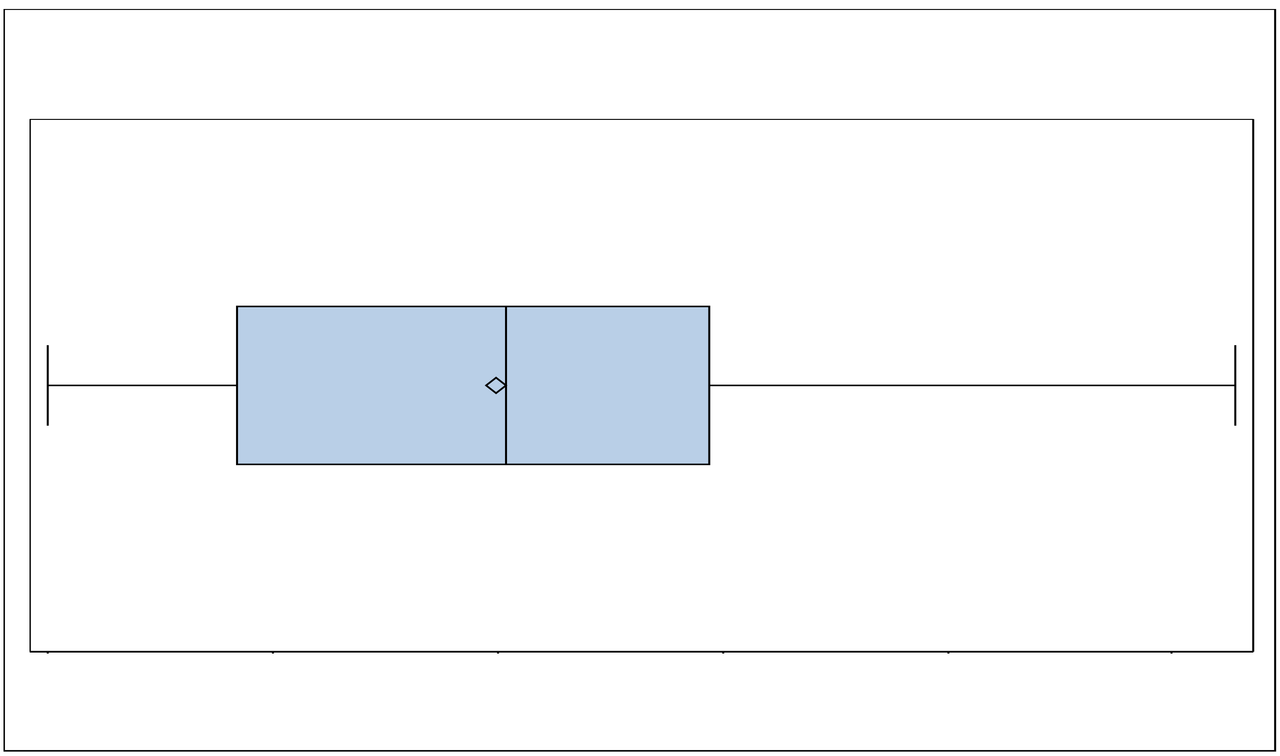
4. TriplesByBatters3Bases - Histogram and Box Plot of Imputed Data Prior to Transformation



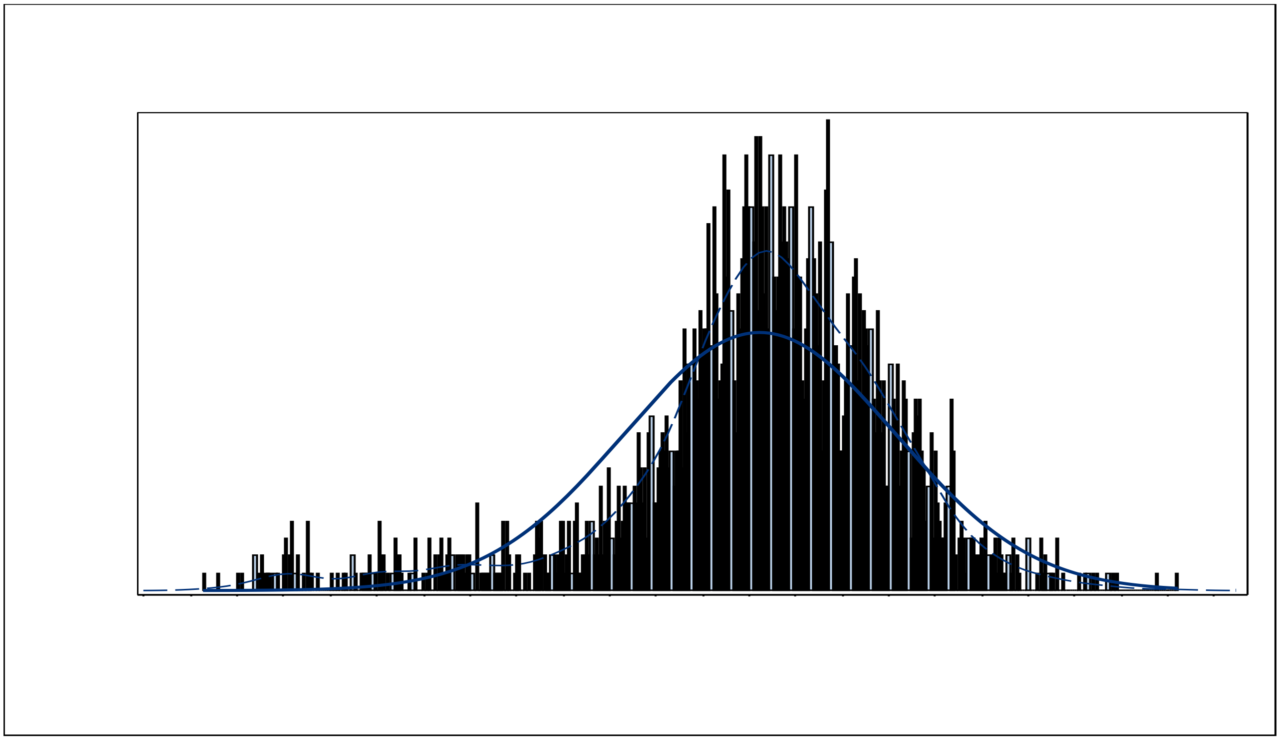


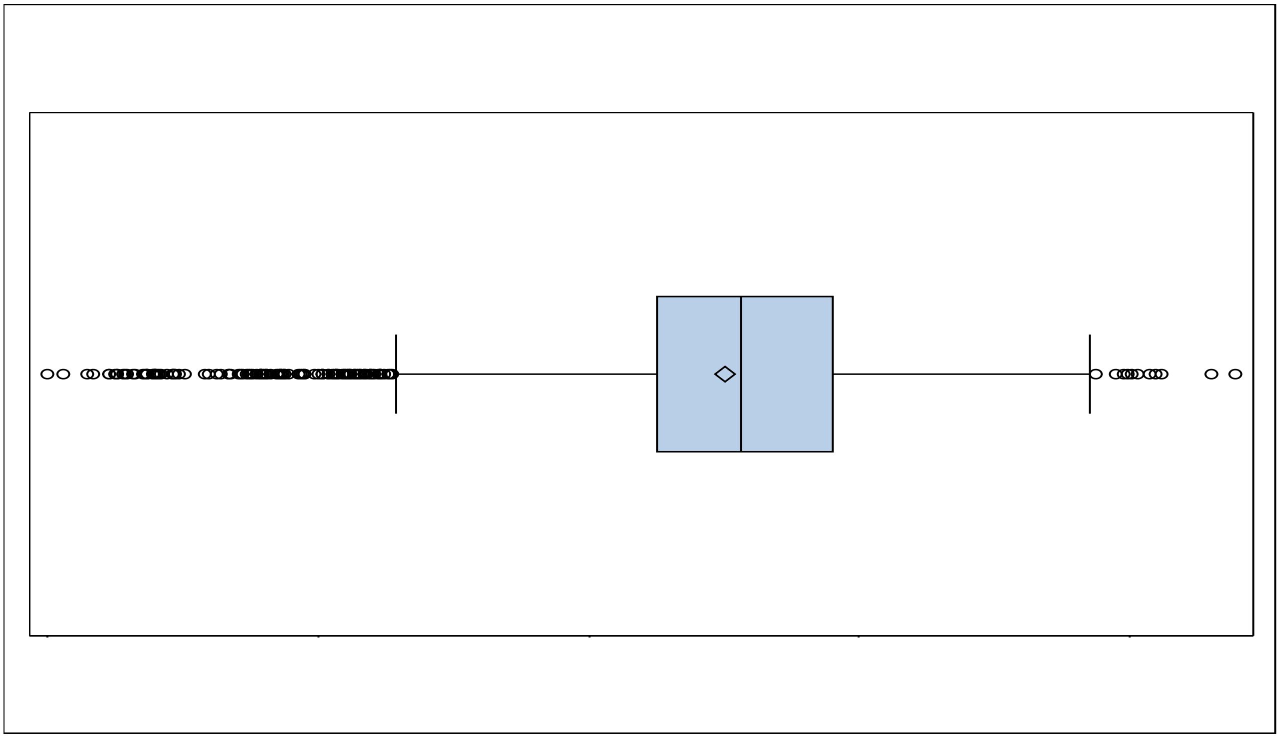
5. HomerunsByBatters4Bases - Histogram and Box Plot of Imputed Data Prior to Transformation



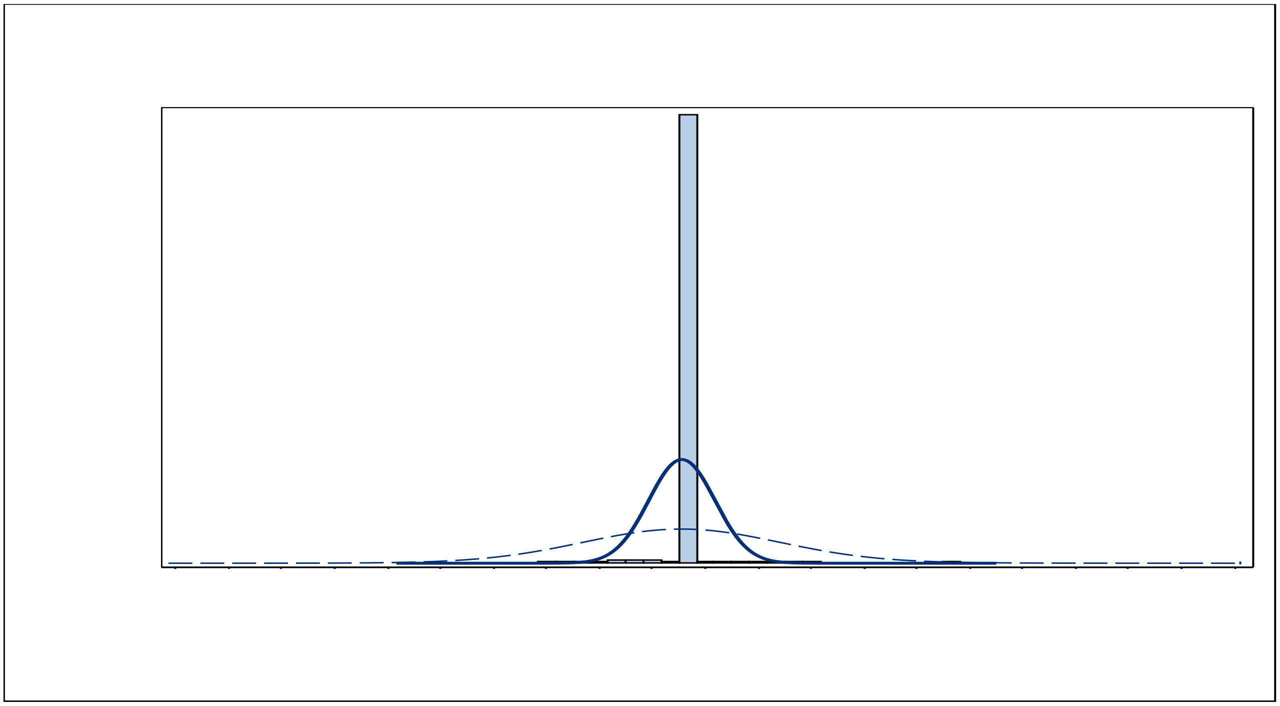


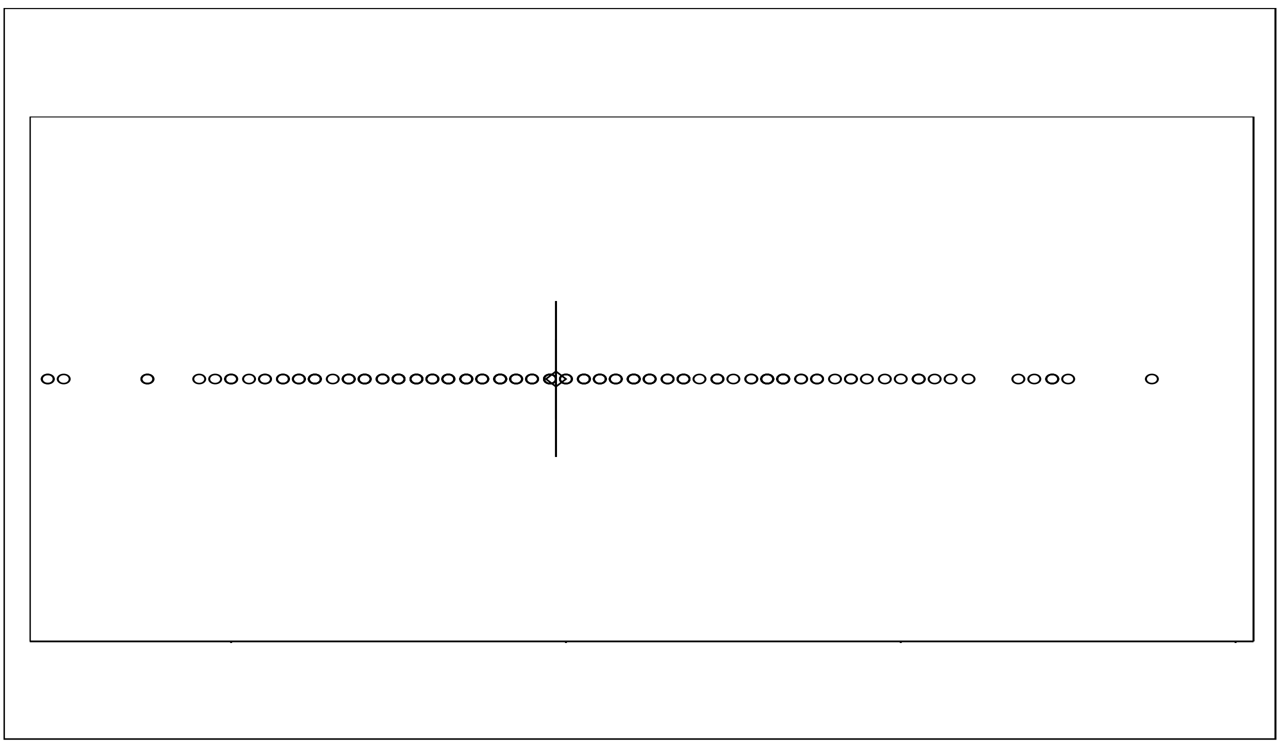
6. WalksByBatters - Histogram and Box Plot of Imputed Data Prior to Transformation



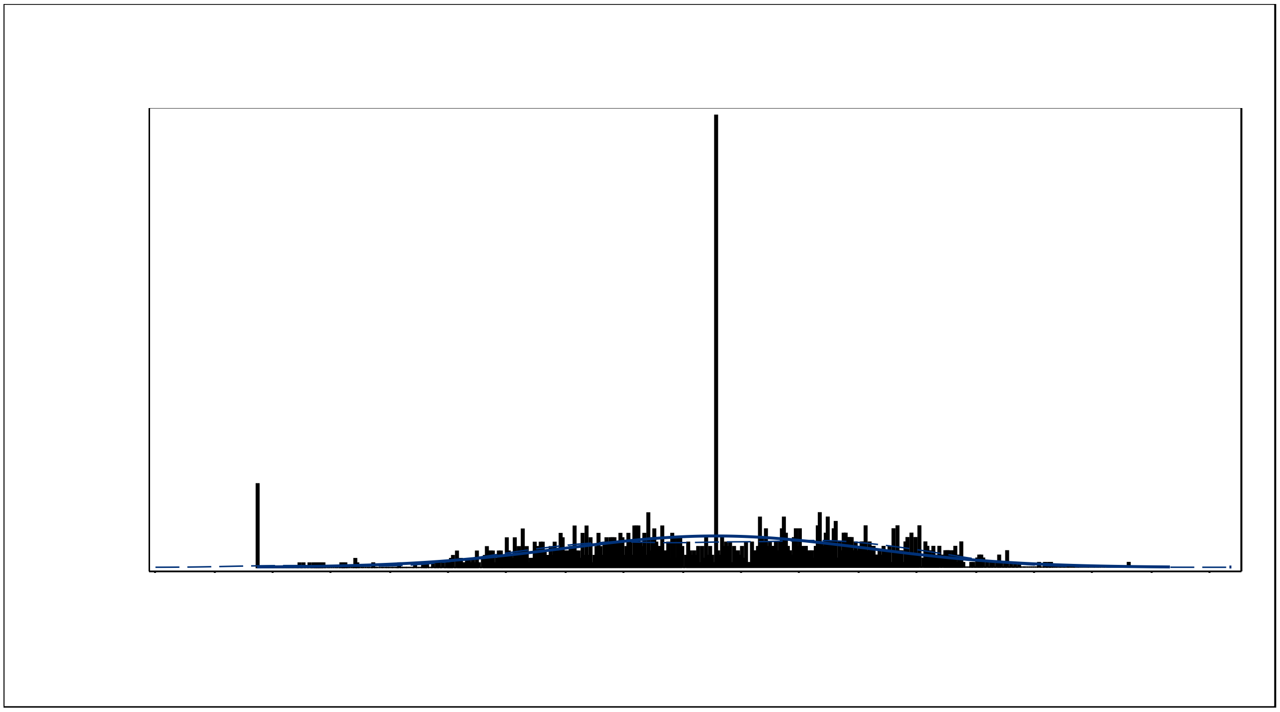


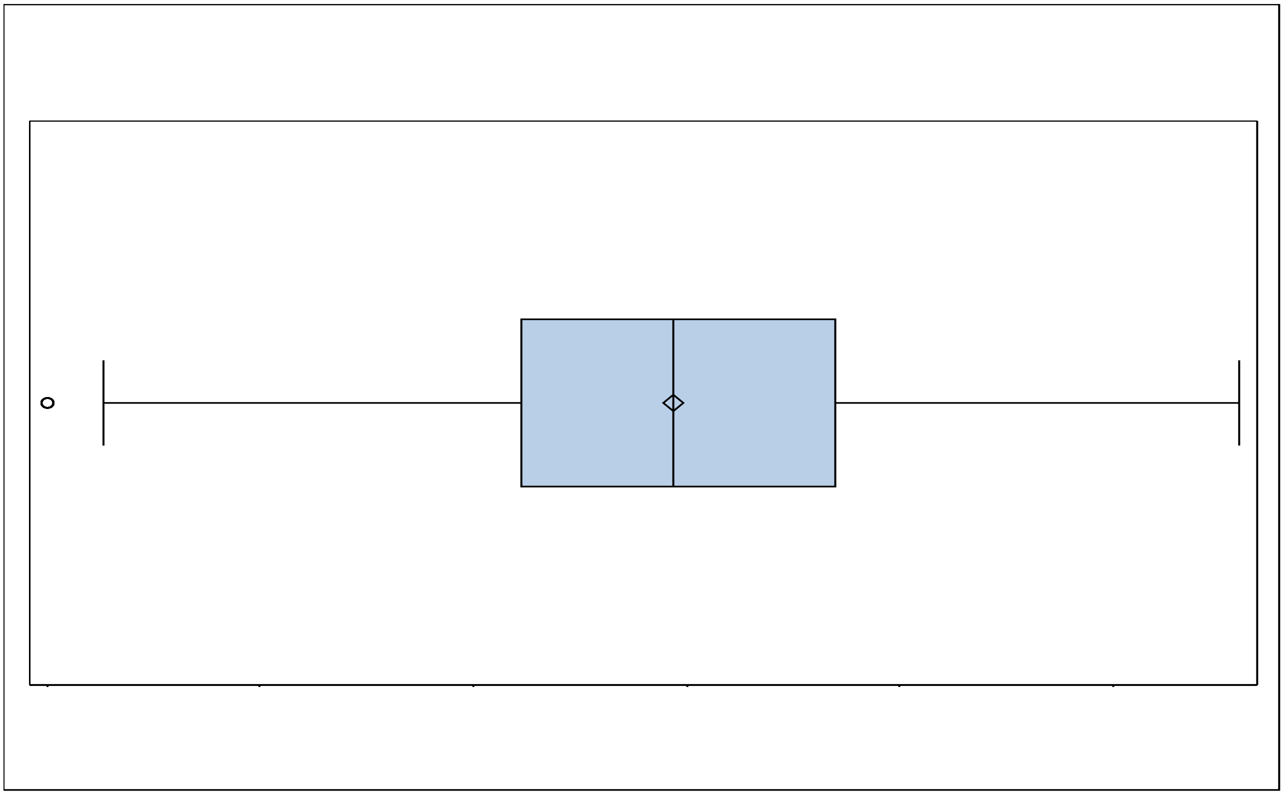
7. IMP_BattersHitByPitch - Histogram and Box Plot of Imputed Data Prior to Transformation



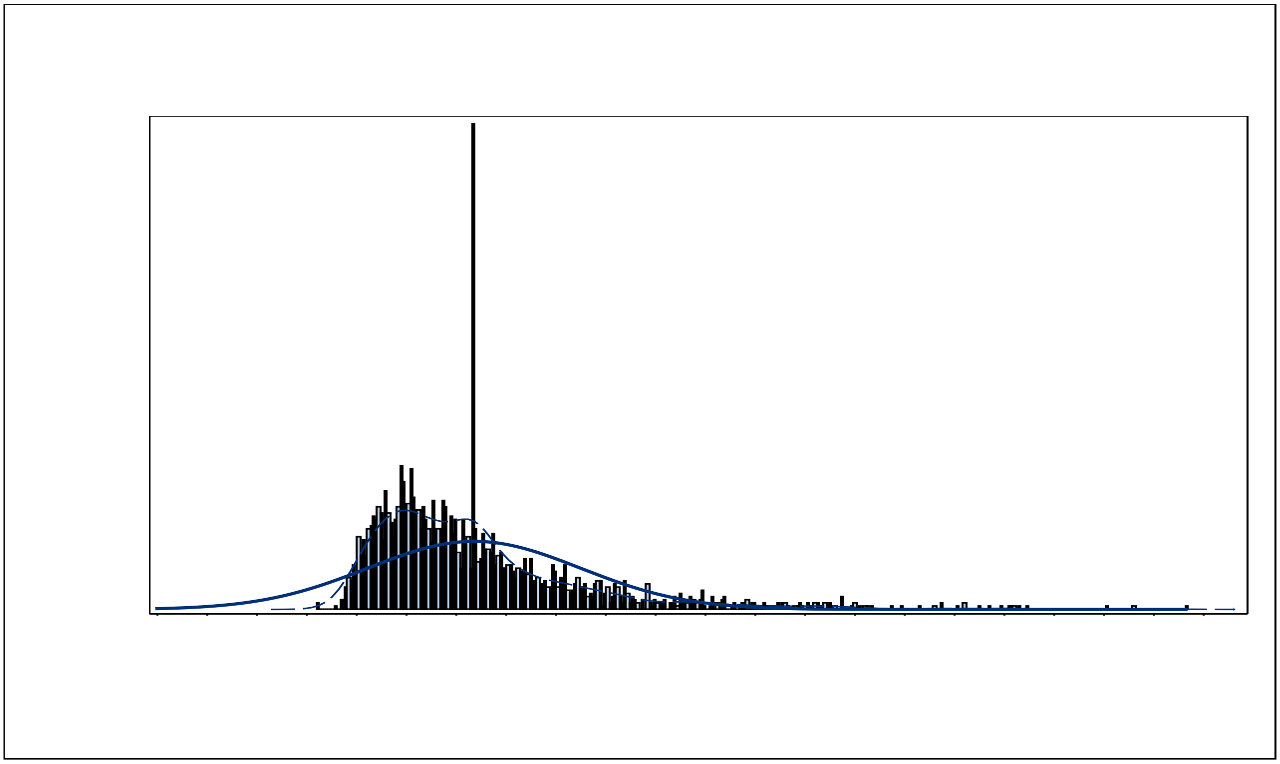


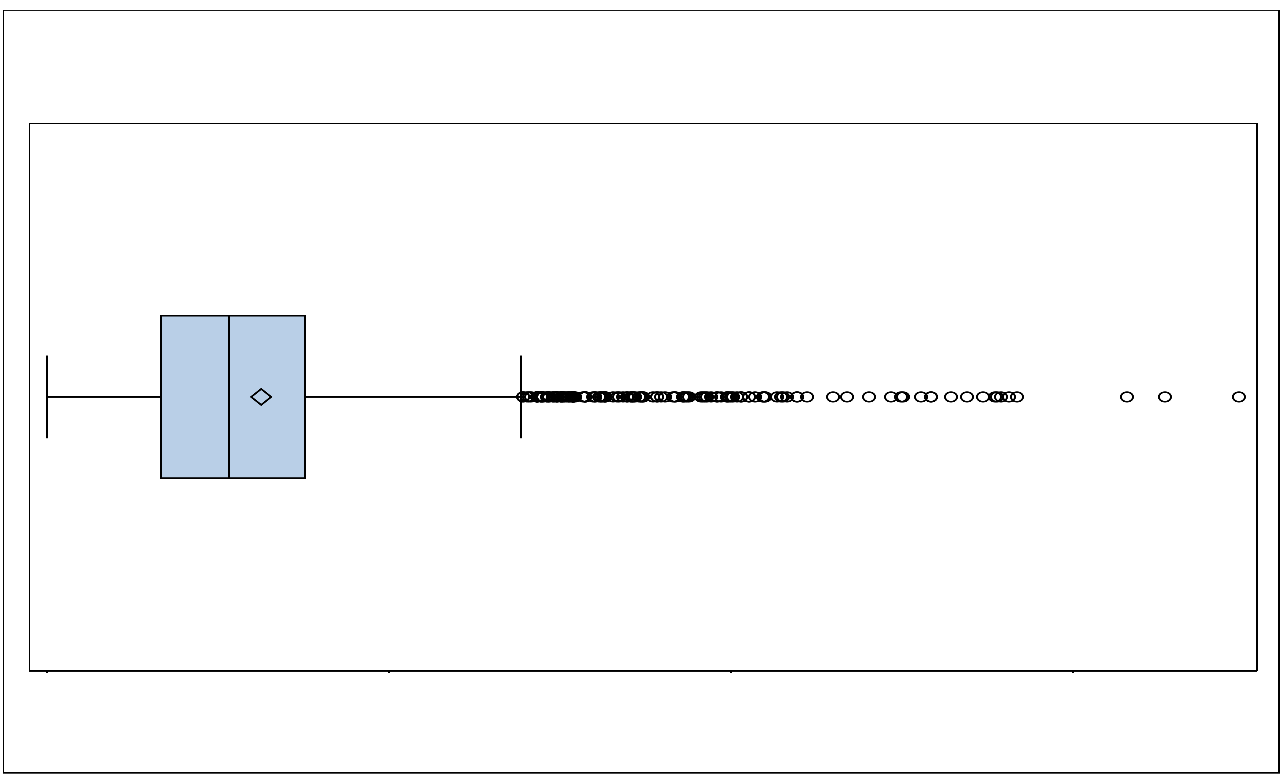
8. IMP_StrikeoutsByBatters - Histogram and Box Plot of Imputed Data Prior to Transformation



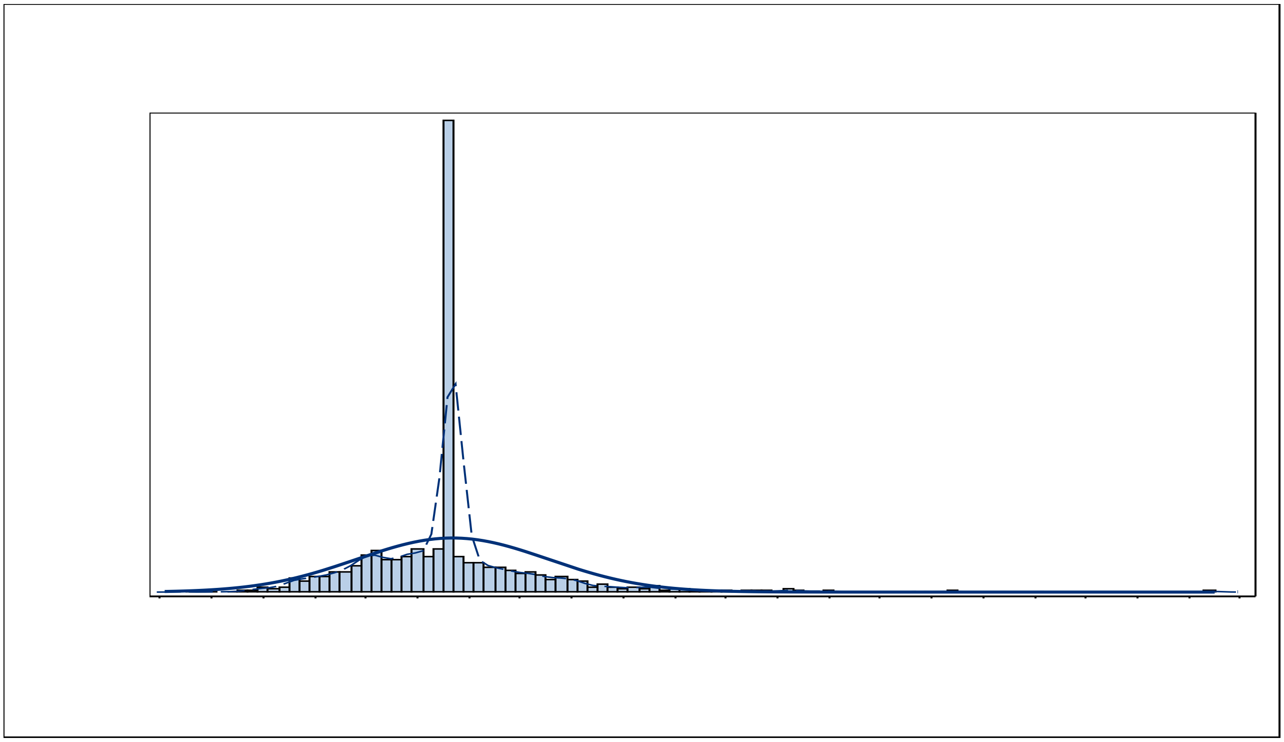


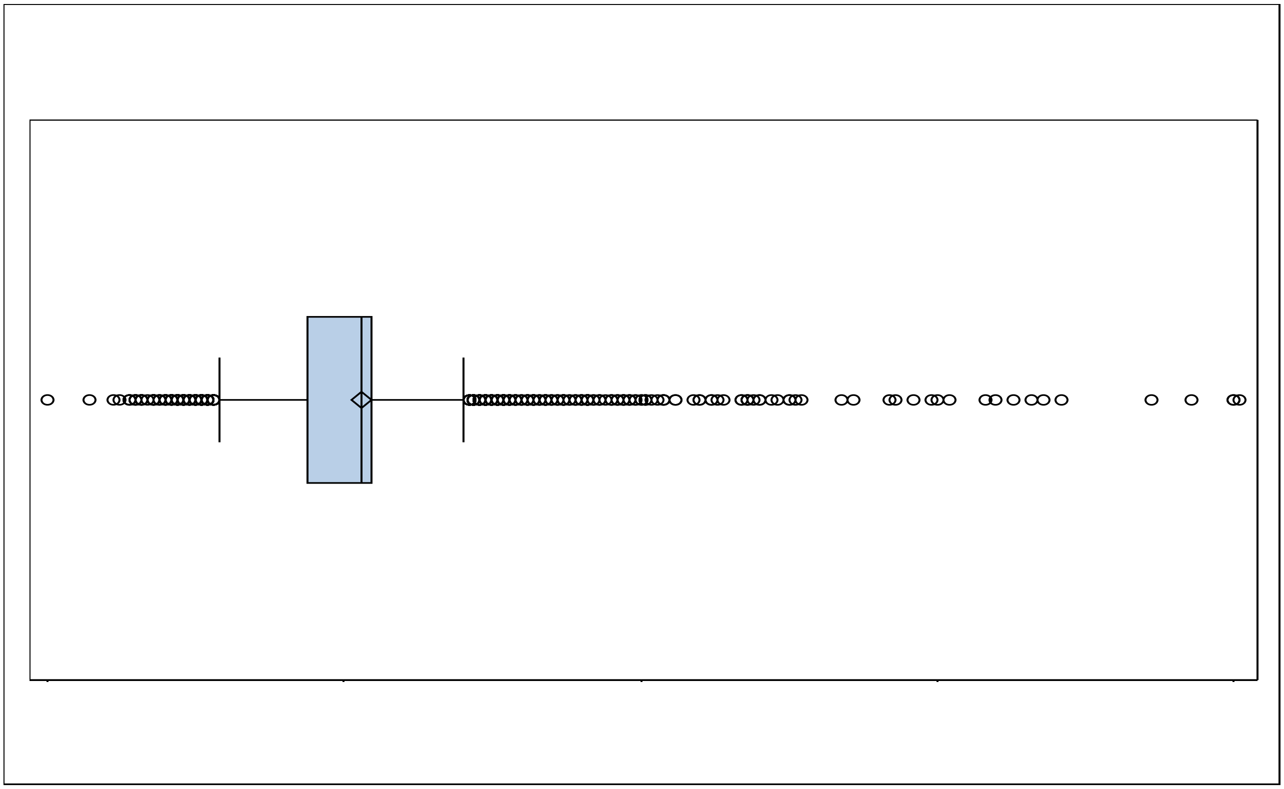
9. IMP_StolenBases - Histogram and Box Plot of Imputed Data Prior to Transformation



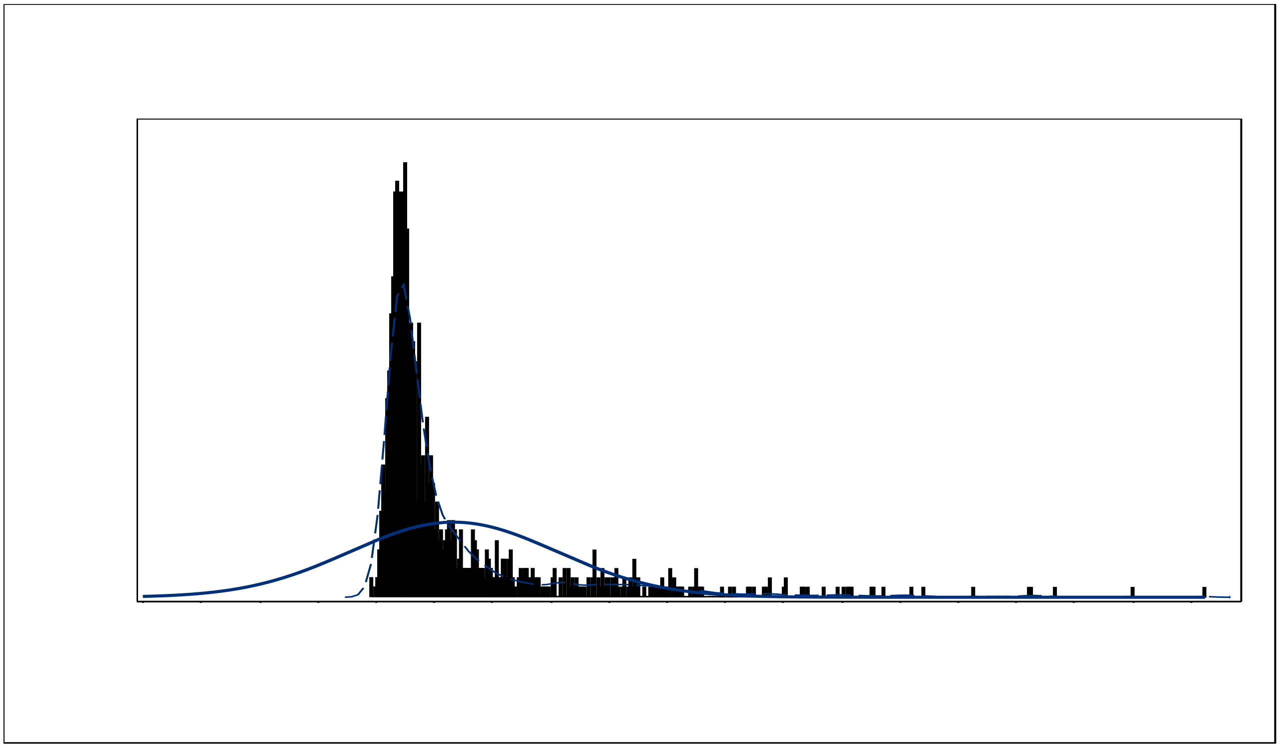


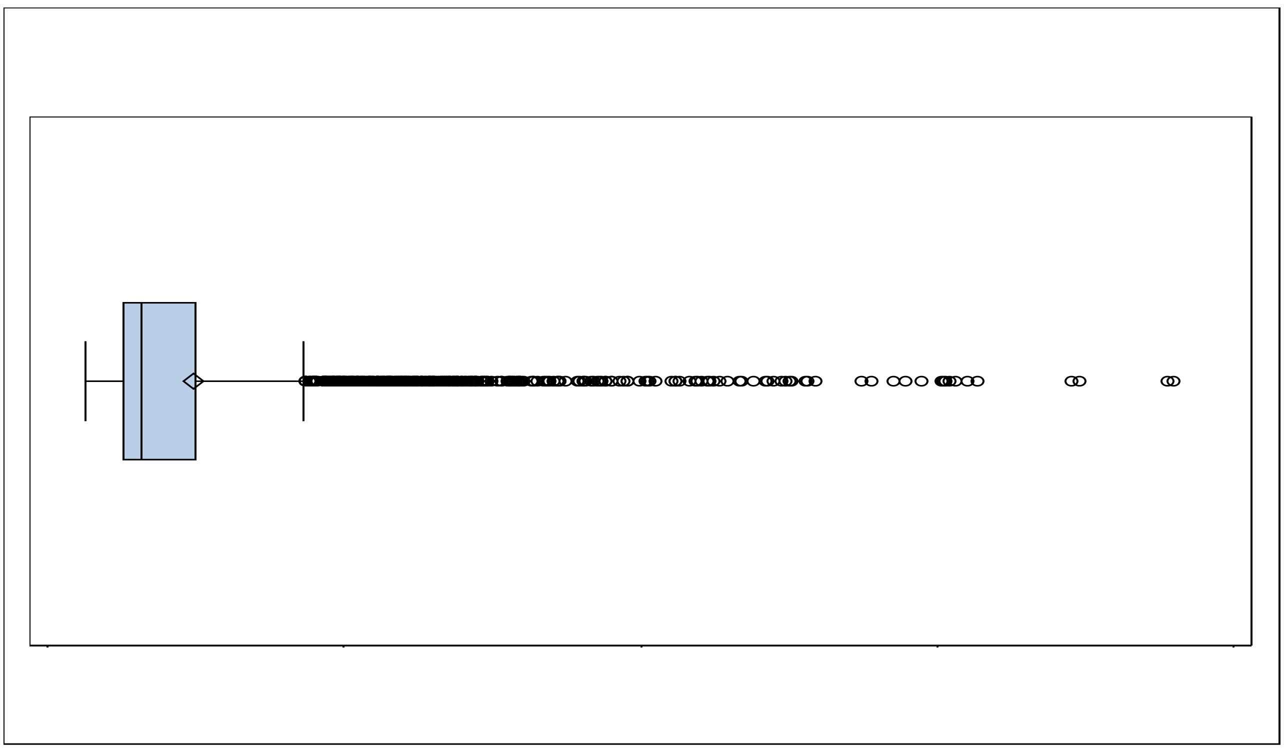
10. IMP_CaughtStealing - Histogram and Box Plot of Imputed Data Prior to Transformation



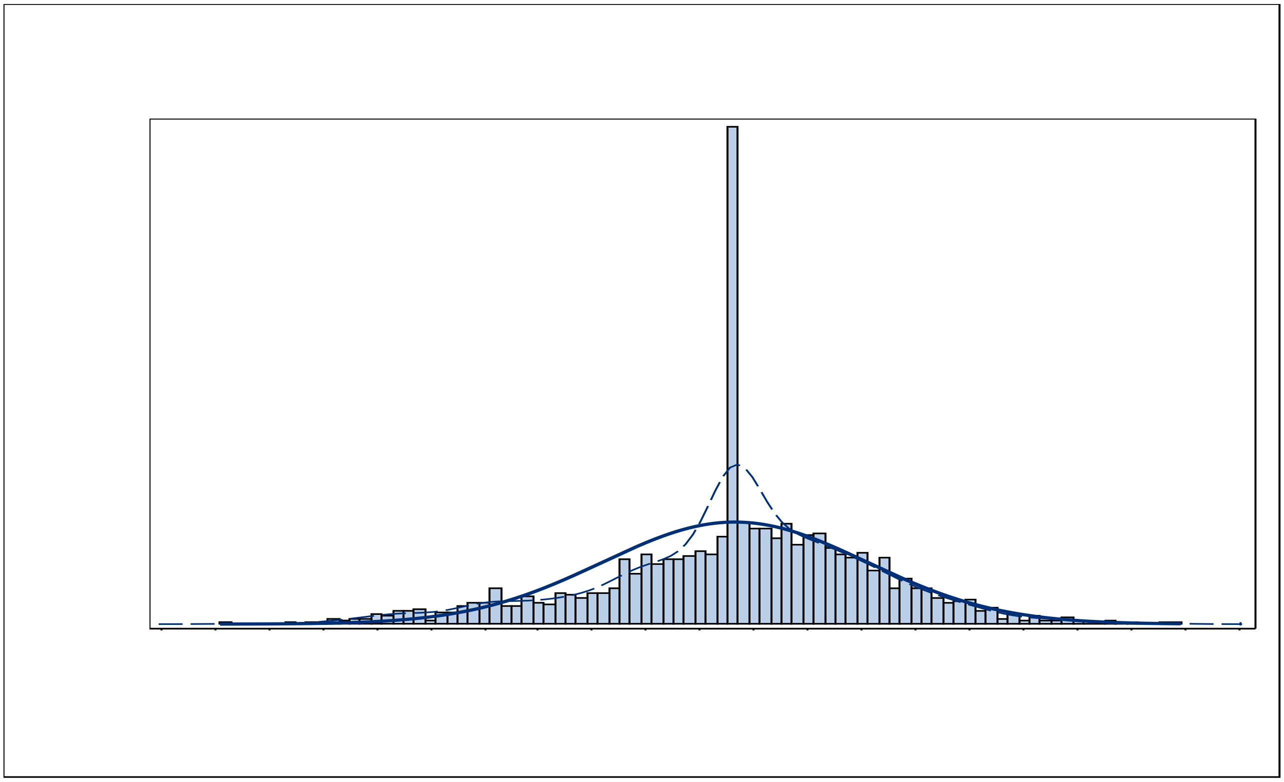


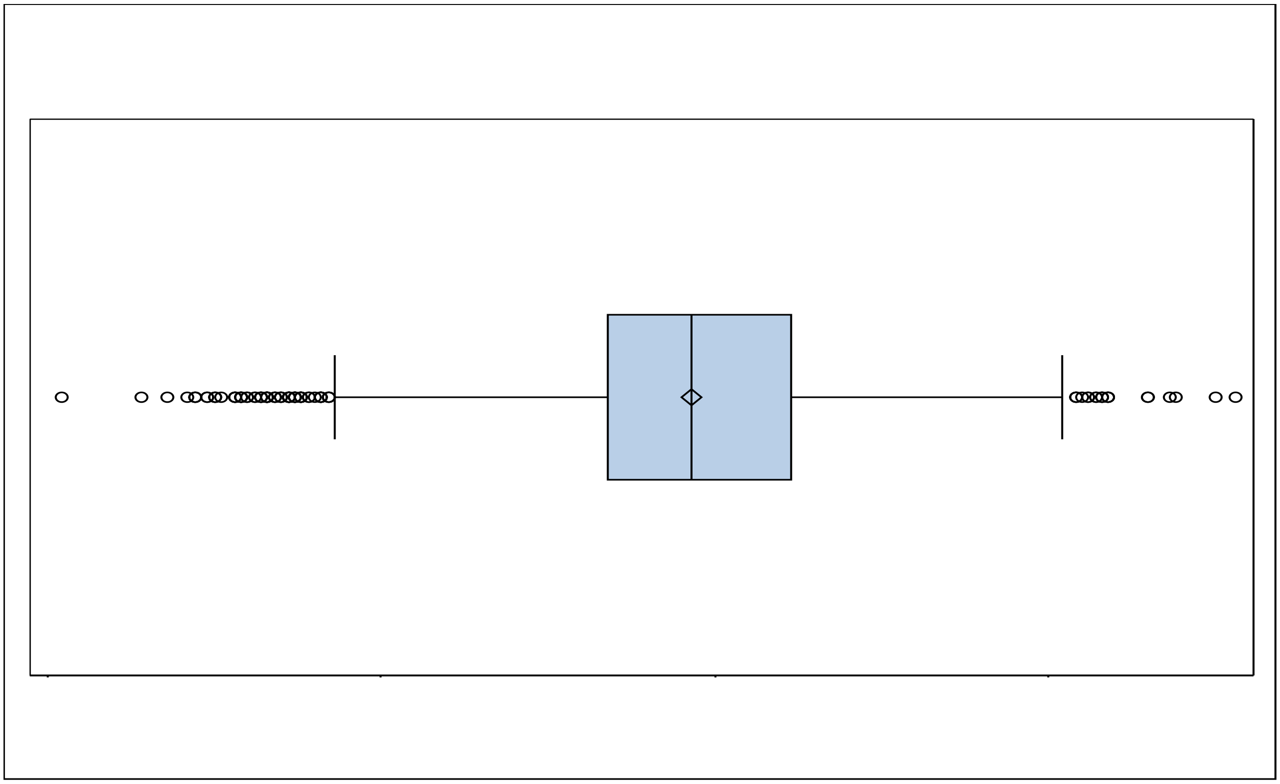
11. Errors - Histogram and Box Plot of Imputed Data Prior to Transformation



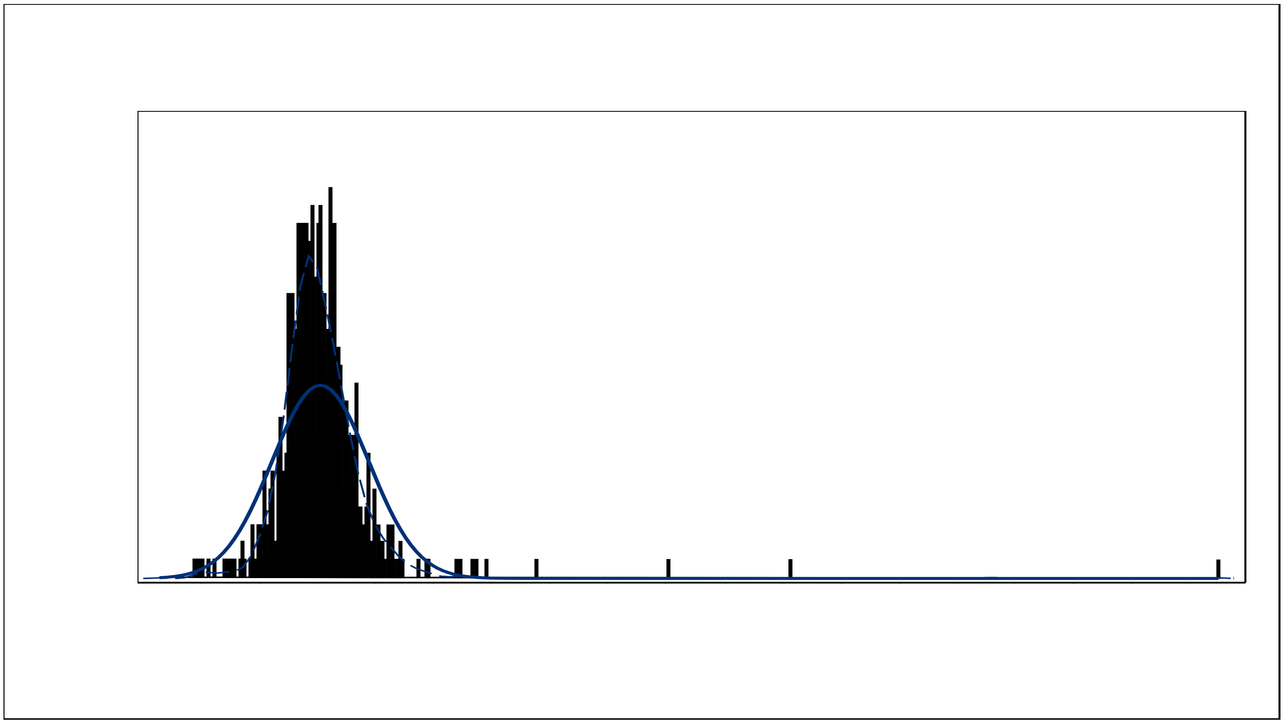


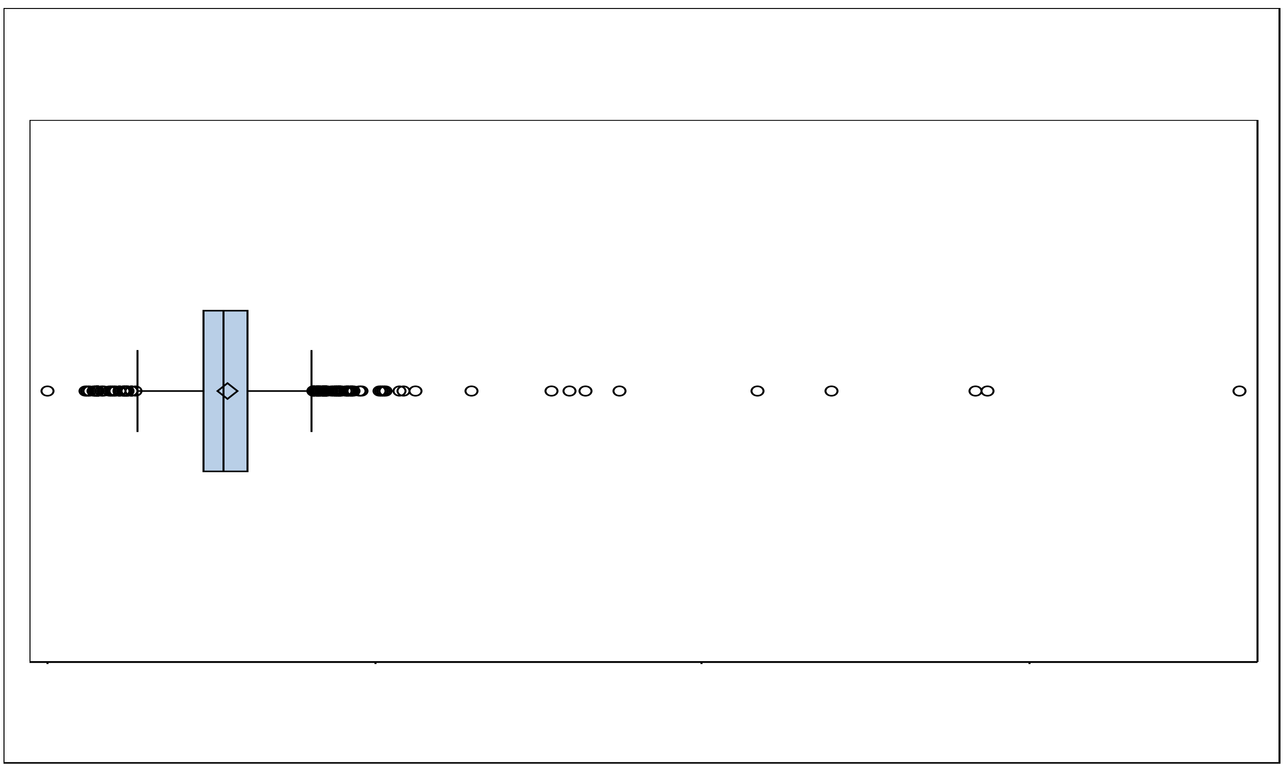
12. IMP_DoublePlays - Histogram and Box Plot of Imputed Data Prior to Transformation



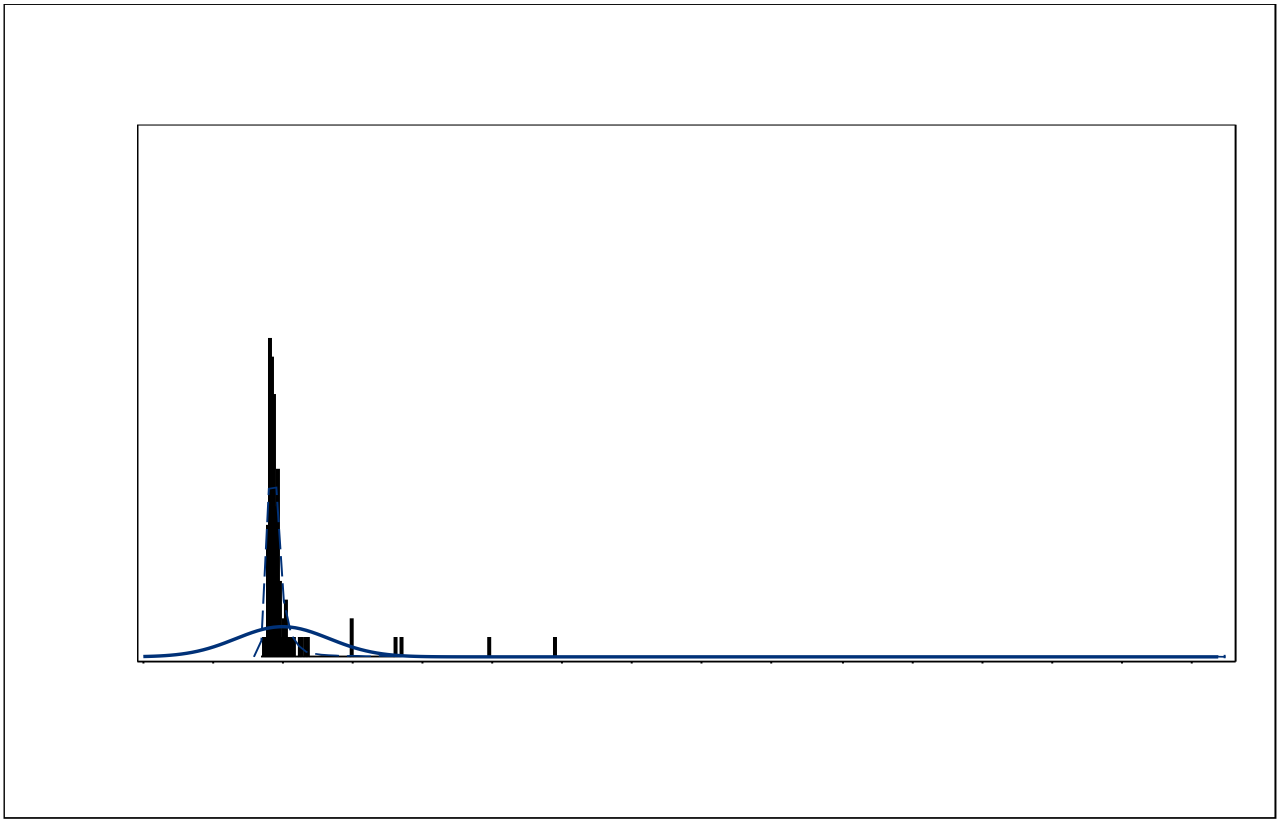


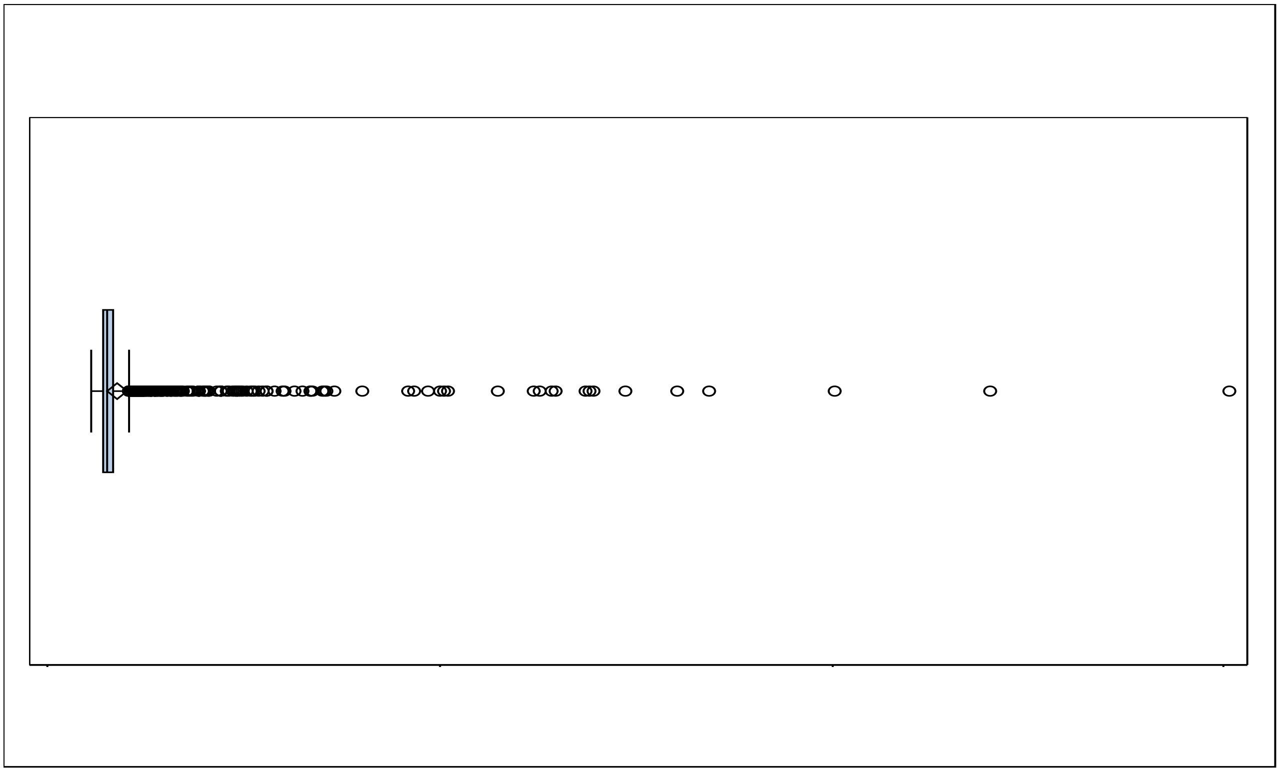
13. WalksAllowed - Histogram and Box Plot of Imputed Data Prior to Transformation



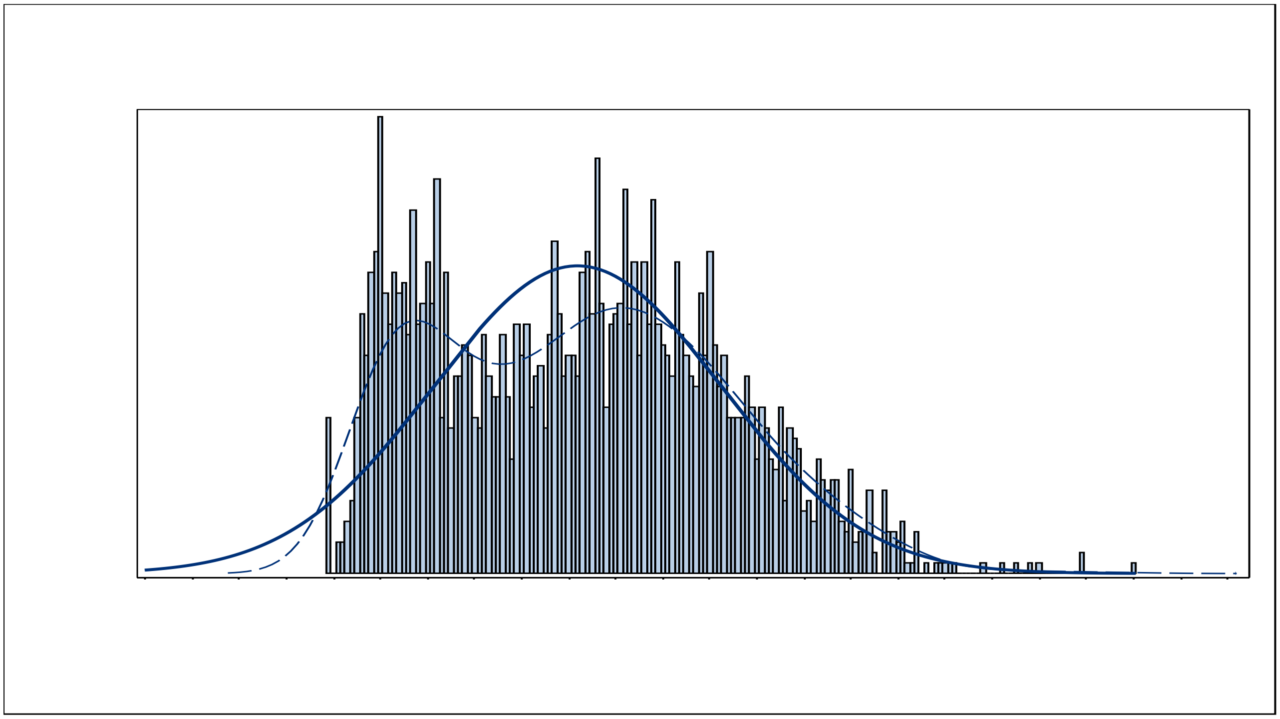


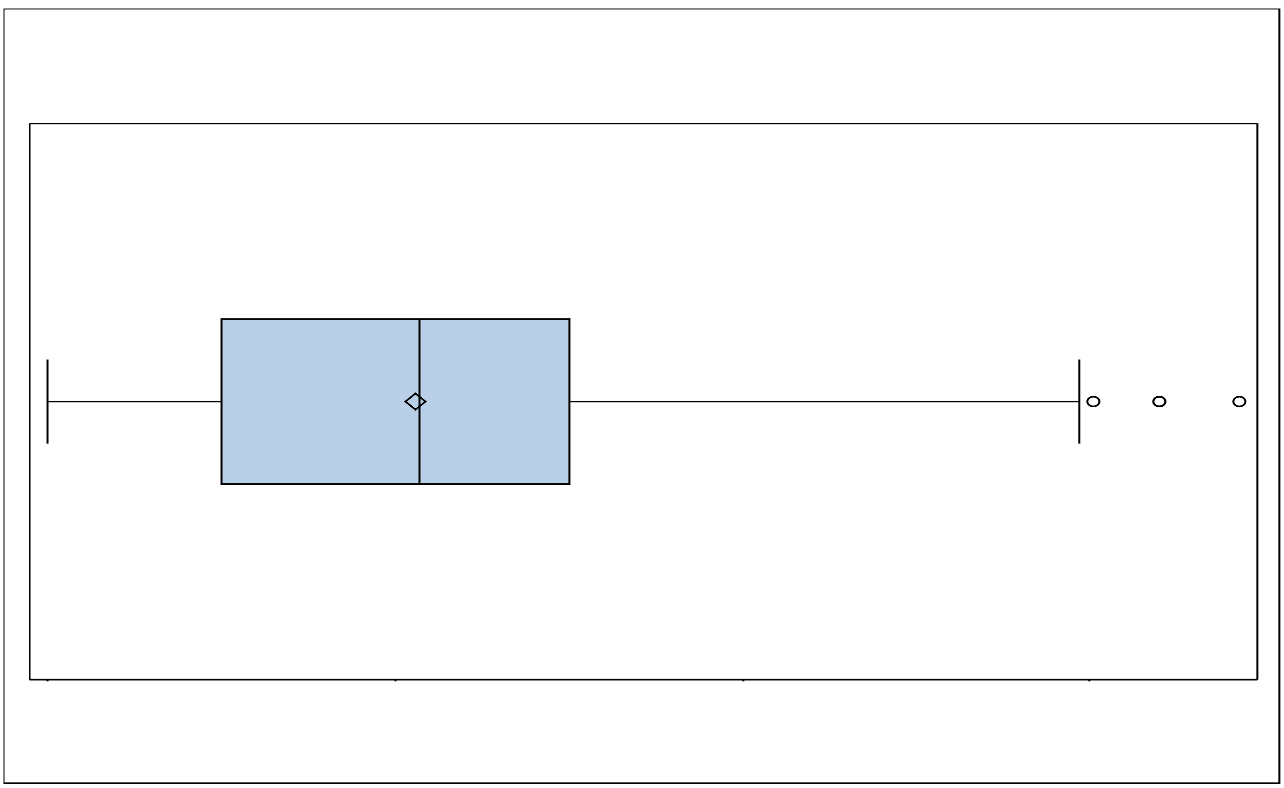
14. HitsAllowed - Histogram and Box Plot of Imputed Data Prior to Transformation



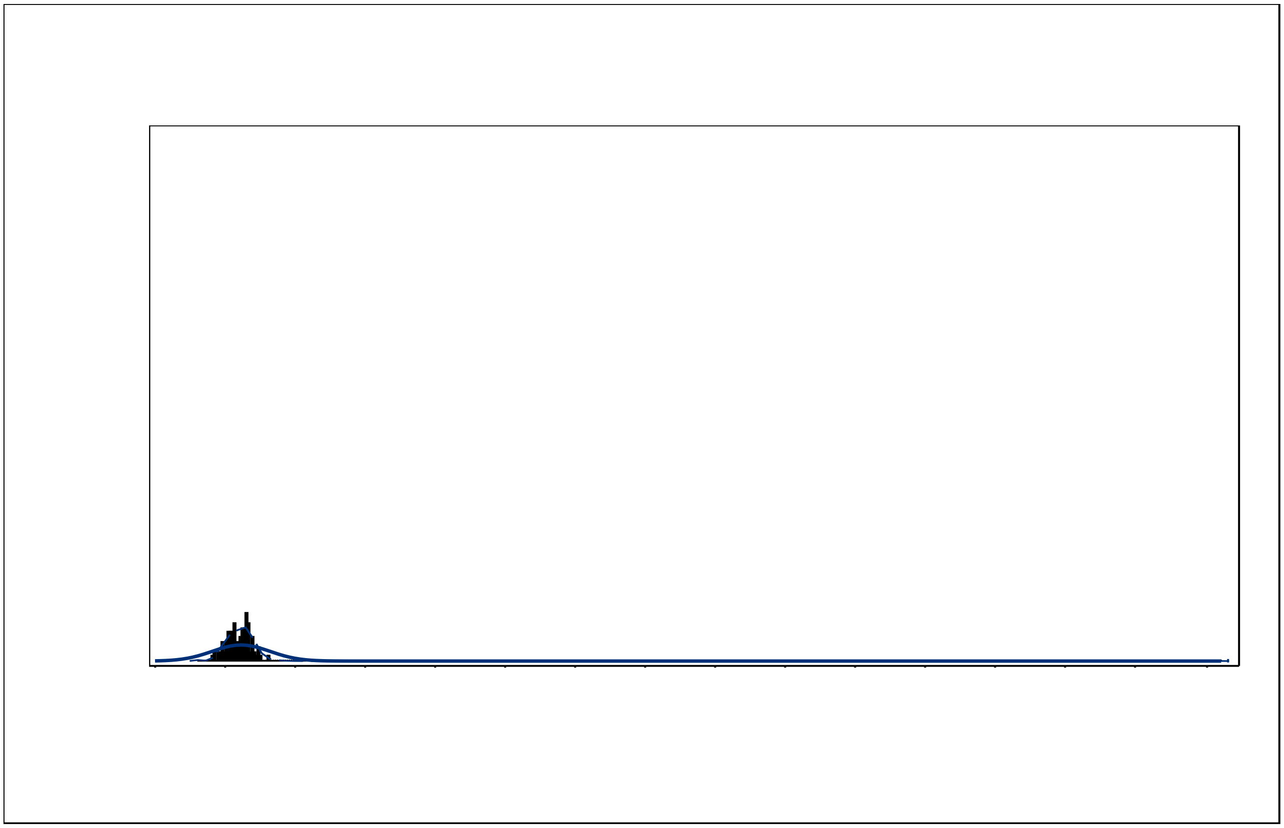


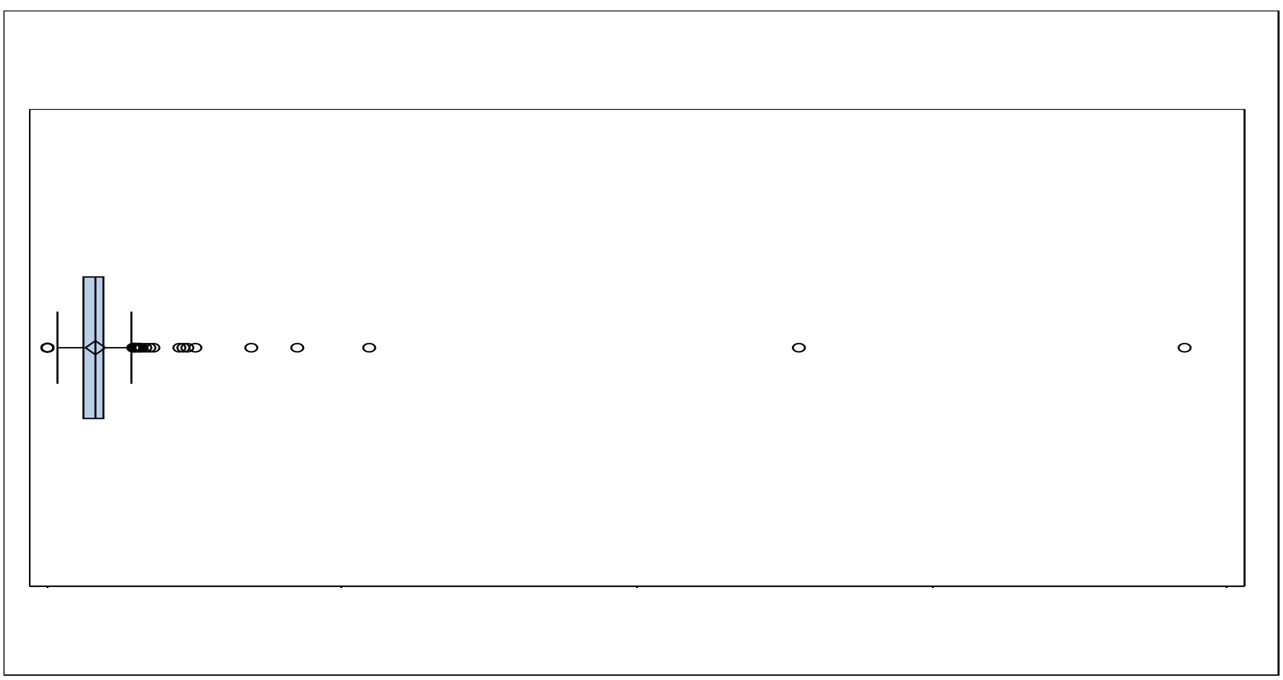
15. HomerunsAllowed - Histogram and Box Plot of Imputed Data Prior to Transformation





16. IMP_StrikeOutsByPitchers - Histogram and Box Plot of Imputed Data Prior to Transformation





Appendix C – SAS Code for Data Imputation of Missing Values

```
************************************************************;

************************************************************;

*	Part 2 - DATA PREPARATION;

************************************************************;

************************************************************;

* The data discovery above indentified 6 variables wiht Missing values;

* The next Data Step creates fields to store the imputed values and;

* a Flag value for each of these 6 variables;

DATA moneyball_train;

SET moneyball_train;

		IMP_StrikeoutsByBatters_N = StrikeoutsByBatters_N;

		MFlag_StrikeoutsByBatters_N = 0;

		IMP_StolenBases_P = StolenBases_P;

		MFlag_StolenBases_P  = 0;

		IMP_CaughtStealing_N = CaughtStealing_N;

		MFlag_CaughtStealing_N = 0;

		IMP_BattersHitByPitch_P = BattersHitByPitch_P;

		MFlag_BattersHitByPitch_P = 0;

		IMP_StrikeoutsByPitchers_P = StrikeoutsByPitchers_P;

		MFlag_StrikeoutsByPitchers_P = 0;

		IMP_DoublePlays_P = DoublePlays_P;

		MFlag_DoublePlays_P = 0;

RUN;

* The code below was added after Synch Session 2 and learning that hard-coding of;

* imputed variables is necessary for this exercise;

* Imputation of missing values using the Mean of each variable;

DATA moneyball_train;

SET moneyball_train;
```

```
        If missing(StrikeoutsByBatters_N) THEN DO
                IMP_StrikeoutsByBatters_N = 735.6053358;
                MFlag_StrikeoutsByBatters_N = 1;
        END;
        IF missing(StolenBases_P) THEN DO
                IMP_StolenBases_P = 124.7617716;
                MFlag_StolenBases_P  = 1;
        END;
        IF missing(CaughtStealing_N) THEN DO
                IMP_CaughtStealing_N = 52.8038564;
                MFlag_CaughtStealing_N = 1;
        END;
        IF missing(BattersHitByPitch_P) THEN DO
                IMP_BattersHitByPitch_P = 59.3560209;
                MFlag_BattersHitByPitch_P = 1;
        END;
        IF missing(StrikeoutsByPitchers_P) THEN DO
                IMP_StrikeoutsByPitchers_P = 817.7304508;
                MFlag_StrikeoutsByPitchers_P = 1;
        END;
        IF missing(DoublePlays_P) THEN DO
                IMP_DoublePlays_P = 146.3879397;
                MFlag_DoublePlays_P = 1;
        END;
RUN;
```

Appendix D – SAS Code for the Transformation of Outlier Values

```
DATA moneyball_train;

SET moneyball_train;

	* FOR LOG10 TRANSFORMATION;

		X1_WalksByBatters_P = WalksByBatters_P;

		X1_HitsAllowed_N = HitsAllowed_N;

		X1_Errors_N = Errors_N;

		X1_IMP_CaughtStealing_N = IMP_CaughtStealing_N;

		X1_IMP_DoublePlays_P = IMP_DoublePlays_P;


	* FOR STANDARDIZED AND TRIM TRANSFORMATION;

		X2_TriplesByBatters3Bases_P = TriplesByBatters3Bases_P;

		X2_IMP_BattersHitByPitch_P = IMP_BattersHitByPitch_P;

		X2_IMP_StrikeoutsByPitchers_P = IMP_StrikeoutsByPitchers_P;

		X2_IMP_StolenBases_P = IMP_StolenBases_P;

RUN;

** USE THE FOLLOWING CODE TO TRANSFORM THE DATA

* Tranformation of Variables that were identified to have Outliers in the train data set;

DATA moneyball_train;

SET moneyball_train;


	******* LOG10 TRANSFORMATION:

	*** For Variable X1_WalksByBatters_P:

	* first, cap any outlier value below p1 or greater than p99;

			IF X1_WalksByBatters_P < 79 THEN X1_WalksByBatters_P = 79;

	ELSE IF X1_WalksByBatters_P > 755 THEN X1_WalksByBatters_P = 755;

	* take the log of the value in order to transform it and minimize influence;

	X1_WalksByBatters_P = sign(X1_WalksByBatters_P) * log10(abs(X1_WalksByBatters_P)+1);


	*** For Variable X1_HitsAllowed_N:

	* first, cap any outlier value below p1 or greater than p99;
```

```
IF X1_HitsAllowed_N < 1244 THEN X1_HitsAllowed_N = 1244;
ELSE IF X1_HitsAllowed_N > 7093 THEN X1_HitsAllowed_N = 7093;
* take the log of the value in order to transform it and minimize influence;
X1_HitsAllowed_N = sign(X1_HitsAllowed_N) * log10(abs(X1_HitsAllowed_N)+1);

*** For Variable X1_Errors_N:
* first, cap any outlier value below p1 or greater than p99;
IF X1_Errors_N < 86 THEN X1_Errors_N = 86;
ELSE IF X1_Errors_N > 1237 THEN X1_Errors_N = 1237;
* take the log of the value in order to transform it and minimize influence;
X1_Errors_N = sign(X1_Errors_N) * log10(abs(X1_Errors_N)+1);

*** For Variable X1_IMP_CaughtStealing_N:
* first, cap any outlier value below p1 or greater than p99;
IF X1_IMP_CaughtStealing_N < 18 THEN X1_IMP_CaughtStealing_N = 18;
ELSE IF X1_IMP_CaughtStealing_N > 125 THEN X1_IMP_CaughtStealing_N = 125;
* take the log of the value in order to transform it and minimize influence;
X1_IMP_CaughtStealing_N = sign(X1_IMP_CaughtStealing_N) * log10(abs(X1_IMP_CaughtStealing_N)+1);

*** For Variable X1_IMP_DoublePlays_P:
* first, cap any outlier value below p1 or greater than p99;
IF X1_IMP_DoublePlays_P < 79 THEN X1_IMP_DoublePlays_P = 79;
ELSE IF X1_IMP_DoublePlays_P > 204 THEN X1_IMP_DoublePlays_P = 204;
* take the log of the value in order to transform it and minimize influence;
X1_IMP_DoublePlays_P = sign(X1_IMP_DoublePlays_P) * log10(abs(X1_IMP_DoublePlays_P)+1);

******* STANDARDIZED AND TRIM TRANSFORMATION:
*** For Variable X2_TriplesByBatters3Bases_P:
* first, cap any outlier value below p1 or greater than p99;
IF X2_TriplesByBatters3Bases_P < 17 THEN X2_TriplesByBatters3Bases_P = 17;
```

```
ELSE IF X2_TriplesByBatters3Bases_P > 134 THEN X2_TriplesByBatters3Bases_P = 134;

STD_X    = (X2_TriplesByBatters3Bases_P - 55.25)/27.938557;  * STANDARDIZING PARAMETER;

X2_TriplesByBatters3Bases_P = max(min(STD_X,3),-3);                        * TRIMMING PARAMETERS;


*** For Variable X2_IMP_BattersHitByPitch_P:

* first, cap any outlier value below p1 or greater than p99;

                    IF X2_IMP_BattersHitByPitch_P < 45 THEN X2_IMP_BattersHitByPitch_P = 45;

ELSE IF X2_IMP_BattersHitByPitch_P > 75 THEN X2_IMP_BattersHitByPitch_P = 75;

STD_X    = (X2_IMP_BattersHitByPitch_P - 59.3560209)/12.9671225;  * STANDARDIZING PARAMETER;

X2_IMP_BattersHitByPitch_P         = max(min(STD_X,3),-3);                                * TRIMMING
PARAMETERS;


*** For Variable X2_IMP_StrikeOutsByPitchers_P:

* first, cap any outlier value below p1 or greater than p99;

                    IF X2_IMP_StrikeOutsByPitchers_P < 205 THEN X2_IMP_StrikeOutsByPitchers_P = 205;

ELSE IF X2_IMP_StrikeOutsByPitchers_P > 1474 THEN X2_IMP_StrikeOutsByPitchers_P = 1474;

STD_X    = (X2_IMP_StrikeOutsByPitchers_P - 817.7304508)/553.0850315;  * STANDARDIZING PARAMETER;

X2_IMP_StrikeOutsByPitchers_P     = max(min(STD_X,3),-3);                                * TRIMMING
PARAMETERS;


*** For Variable X2_IMP_StolenBases_P:

* first, cap any outlier value below p1 or greater than p99;

          IF X2_IMP_StolenBases_P < 24 THEN X2_IMP_StolenBases_P = 24;

ELSE IF X2_IMP_StolenBases_P > 438 THEN X2_IMP_StolenBases_P = 438;

STD_X    = (X2_IMP_StolenBases_P - 817.7304508)/553.0850315;  * STANDARDIZING PARAMETER;

X2_IMP_StolenBases_P     = max(min(STD_X,3),-3);                                * TRIMMING PARAMETERS;

RUN;
```

Appendix E – SAS Code for EDA Visualization to Detect Outliers

```
* EDA VISUALIZATION USING GRAPHS TO DETECT OUTLIERS:;

%MACRO EDA_OUTLIER(varParameter =, varEDAObjective =);

        ods graphics on;

        PROC SGPLOT DATA = moneyball_train;

                HISTOGRAM &varParameter.

                    / BINWIDTH = 2 SHOWBINS SCALE = COUNT;

                DENSITY &varParameter.;

                DENSITY &varParameter. / TYPE = KERNEL;

                TITLE "&varParameter Contest.";

                TITLE2 "Count of Teams by Number of &varParameter.";

                TITLE3 "EDA Objective: &varEDAObjective.";

        RUN;

        ods graphics off;


        ods graphics on;

        PROC SGPLOT DATA = moneyball_train;

                HBOX &varParameter.

                    / MISSING ;

                TITLE "&varParameter Contest.";

                TITLE2 "Count of Teams by Number of &varParameter.";

                TITLE3 "EDA Objective: &varEDAObjective.";

        RUN;

        ods graphics off;

%MEND EDA_OUTLIER;


%EDA_OUTLIER(varParameter = TargetWins, varEDAObjective = "EDA of Imputed Data AND Prior to Transformation.");

%EDA_OUTLIER(varParameter = BaseHitsByBattersAllBases_P, varEDAObjective = "EDA of Imputed Data AND Prior to
Transformation.");
```

```
%EDA_OUTLIER(varParameter = DoublesByBatters2Bases_P, varEDAObjective = "EDA of Imputed Data AND Prior to Transformation.");

%EDA_OUTLIER(varParameter = TriplesByBatters3Bases_P, varEDAObjective = "EDA of Imputed Data AND Prior to Transformation.");

%EDA_OUTLIER(varParameter = HomerunsByBatters4Bases_P, varEDAObjective = "EDA of Imputed Data AND Prior to Transformation.");

%EDA_OUTLIER(varParameter = WalksByBatters_P, varEDAObjective = "EDA of Imputed Data AND Prior to Transformation.");

%EDA_OUTLIER(varParameter = IMP_BattersHitByPitch_P, varEDAObjective = "EDA of Imputed Data AND Prior to Transformation.");

%EDA_OUTLIER(varParameter = IMP_StrikeoutsByBatters_N, varEDAObjective = "EDA of Imputed Data AND Prior to Transformation.");

%EDA_OUTLIER(varParameter = IMP_StolenBases_P, varEDAObjective = "EDA of Imputed Data AND Prior to Transformation.");

%EDA_OUTLIER(varParameter = IMP_CaughtStealing_N, varEDAObjective = "EDA of Imputed Data AND Prior to Transformation.");

%EDA_OUTLIER(varParameter = Errors_N, varEDAObjective = "EDA of Imputed Data AND Prior to Transformation.");

%EDA_OUTLIER(varParameter = IMP_DoublePlays_P, varEDAObjective = "EDA of Imputed Data AND Prior to Transformation.");

%EDA_OUTLIER(varParameter = WalksAllowed_N, varEDAObjective = "EDA of Imputed Data AND Prior to Transformation.");

%EDA_OUTLIER(varParameter = HitsAllowed_N, varEDAObjective = "EDA of Imputed Data AND Prior to Transformation.");

%EDA_OUTLIER(varParameter = HomerunsAllowed_N, varEDAObjective = "EDA of Imputed Data AND Prior to Transformation.");

%EDA_OUTLIER(varParameter = IMP_StrikeoutsByPitchers_P, varEDAObjective = "EDA of Imputed Data AND Prior to Transformation.");
```

Appendix F – SAS Code for Simple OLS Regression Model for Each Variable for EDA of Imputed and Transformed Data

```
***********************************************************;
*         DATA EXPLORATION USING SIMPLE REGRESSION WITH THE;
*                   IMPUTED AND TRANSFORMED DATA;
***********************************************************;
* SIMPLE OLS REGRESSION FOR EACH VARIABLE;
proc reg data=moneyball_train;
model TargetWins = BaseHitsByBattersAllBases_P / selection=rsquare;
run;
quit;


proc reg data=moneyball_train;
model TargetWins = DoublesByBatters2Bases_P / selection=rsquare;
run;
quit;


* gives better Rsquare than X1;
proc reg data=moneyball_train;
model TargetWins = X2_TriplesByBatters3Bases_P / selection=rsquare;
run;
quit;


proc reg data=moneyball_train;
model TargetWins = HomerunsByBatters4Bases_P / selection=rsquare;
run;
quit;


* gives better Rsquare than X2;
```

```
proc reg data=moneyball_train;
model TargetWins = X1_WalksByBatters_P / selection=rsquare;
run;
quit;

* gives better Rsquare than X2;
proc reg data=moneyball_train;
model TargetWins = X1_HitsAllowed_N / selection=rsquare;
run;
quit;

proc reg data=moneyball_train;
model TargetWins = HomerunsAllowed_N / selection=rsquare;
run;
quit;

proc reg data=moneyball_train;
model TargetWins = WalksAllowed_N / selection=rsquare;
run;
quit;

* X1 gives better RSquare;
proc reg data=moneyball_train;
model TargetWins = X1_Errors_N / selection=rsquare;
run;
quit;

proc reg data=moneyball_train;
model TargetWins = IMP_StrikeoutsByBatters_N MFlag_StrikeoutsByBatters_N / selection=rsquare;
```

```
run;

quit;


* the X1 transformation was not giving enough data for the simple OLS;

proc reg data=moneyball_train;

model TargetWins = X2_IMP_StolenBases_P MFlag_StolenBases_P / selection=rsquare;

run;

quit;


* X1 gives better RSquare;

proc reg data=moneyball_train;

model TargetWins = X1_IMP_CaughtStealing_N MFlag_CaughtStealing_N / selection=rsquare;

run;

quit;


* Both X1 and X2 give the same result;

proc reg data=moneyball_train;

model TargetWins = X2_IMP_BattersHitByPitch_P MFlag_BattersHitByPitch_P / selection=rsquare;

run;

quit;


* X2 gives better RSquare;

proc reg data=moneyball_train;

model TargetWins = X2_IMP_StrikeoutsByPitchers_P MFlag_StrikeoutsByPitchers_P / selection=rsquare;

run;

quit;


* X1 gives better RSquare;

proc reg data=moneyball_train;
```

```
model TargetWins = X1_IMP_DoublePlays_P MFlag_DoublePlays_P / selection=rsquare;
run;
quit;

proc reg data=moneyball_train;
model TargetWins = INT_P / selection=rsquare;
run;
quit;

proc reg data=moneyball_train;
model TargetWins = INT_N / selection=rsquare;
run;
quit;
```

Appendix G – SAS Code for Model Building using Stepwise, Forward, and Backward Selection on All Variables from Outlier EDA Results

```
************************************************************;

************************************************************;

*  Part 3:          Model Building and Selection;

************************************************************;

************************************************************;

* THIS STEP IN MODEL CREATION USES ALL IMPUTED AND TRANSFORMED VARIABLES;

ods graphics on;

PROC REG DATA = moneyball_train outest=ESTFILE AIC SBC BIC CP ADJRSQ plots=diagnostics(stats=(default AIC SBC BIC CP ADJRSQ));

MODEL_STEPWISE: MODEL TargetWins =    BaseHitsByBattersAllBases_P

                                      DoublesByBatters2Bases_P

                                      WalksAllowed_N

                                      HomerunsAllowed_N

                                      HomerunsByBatters4Bases_P

                                      IMP_StrikeoutsByBatters_N

                                      MFlag_StrikeoutsByBatters_N

                                      X1_WalksByBatters_P

                                      X1_Errors_N

                                      X1_IMP_DoublePlays_P

                                      MFlag_DoublePlays_P

                                      X1_HitsAllowed_N

                                      X1_IMP_CaughtStealing_N

                                      MFlag_CaughtStealing_N

                                      X2_TriplesByBatters3Bases_P

                                      X2_IMP_StolenBases_P

                                      MFlag_StolenBases_P

                                      X2_IMP_StrikeOutsByPitchers_P

                                      MFlag_StrikeoutsByPitchers_P

                                      X2_IMP_BattersHitByPitch_P
```

```
                         MFlag_BattersHitByPitch_P
                         INT_P
                         INT_N
                         / selection = stepwise VIF AIC SBC BIC CP ADJRSQ;
RUN;

MODEL_FORWARD: MODEL TargetWins =    BaseHitsByBattersAllBases_P
                         DoublesByBatters2Bases_P
                         WalksAllowed_N
                         HomerunsAllowed_N
                         HomerunsByBatters4Bases_P
                         IMP_StrikeoutsByBatters_N
                         MFlag_StrikeoutsByBatters_N
                         X1_WalksByBatters_P
                         X1_Errors_N
                         X1_IMP_DoublePlays_P
                         MFlag_DoublePlays_P
                         X1_HitsAllowed_N
                         X1_IMP_CaughtStealing_N
                         MFlag_CaughtStealing_N
                         X2_TriplesByBatters3Bases_P
                         X2_IMP_StolenBases_P
                         MFlag_StolenBases_P
                         X2_IMP_StrikeOutsByPitchers_P
                         MFlag_StrikeoutsByPitchers_P
                         X2_IMP_BattersHitByPitch_P
                         MFlag_BattersHitByPitch_P
                         INT_P
                         INT_N
                         / selection = forward VIF AIC SBC BIC CP ADJRSQ;
```

```
RUN;

MODEL_BACKWARD: MODEL TargetWins =  BaseHitsByBattersAllBases_P
                                    DoublesByBatters2Bases_P
                                    WalksAllowed_N
                                    HomerunsAllowed_N
                                    HomerunsByBatters4Bases_P
                                    IMP_StrikeoutsByBatters_N
                                    MFlag_StrikeoutsByBatters_N
                                    X1_WalksByBatters_P
                                    X1_Errors_N
                                    X1_IMP_DoublePlays_P
                                    MFlag_DoublePlays_P
                                    X1_HitsAllowed_N
                                    X1_IMP_CaughtStealing_N
                                    MFlag_CaughtStealing_N
                                    X2_TriplesByBatters3Bases_P
                                    X2_IMP_StolenBases_P
                                    MFlag_StolenBases_P
                                    X2_IMP_StrikeOutsByPitchers_P
                                    MFlag_StrikeoutsByPitchers_P
                                    X2_IMP_BattersHitByPitch_P
                                    MFlag_BattersHitByPitch_P
                                    INT_P
                                    INT_N
                                    / selection = backward VIF AIC SBC BIC CP ADJRSQ;
RUN;
ods graphics off;
PROC PRINT DATA = ESTFILE; RUN;
```

Appendix H – SAS Code for Model Building Based on the Stepwise, Forward, and Backward Selection Results Above and Without the Variables with Incorrect Signed Coefficients

```
* BASED ON RESULTS ABOVE FROM STEPWISE, FORWARD, AND BACKWARD SELECTION;

* AND AFTER REMOVING THE INCORRECTLY SIGNED COEFFICIENTS AND INCLUDING A;

* LEFT OUT FLAG VARIABLE AND REMOVING A LEFT IN FLAG VARIABLE AS EXPLAINED;

* IN THE WRITE-UP;

ods graphics on;

PROC REG DATA = moneyball_train outest=ESTFILE AIC SBC BIC CP ADJRSQ plots=diagnostics(stats=(default AIC SBC BIC CP ADJRSQ));

MODEL_FROM_STPW: MODEL TargetWins =          BaseHitsByBattersAllBases_P
                                             X1_WalksByBatters_P
                                             X1_Errors_N
                                             X1_IMP_CaughtStealing_N
                                             MFlag_CaughtStealing_N
                                             X2_TriplesByBatters3Bases_P
                                             X2_IMP_StolenBases_P
                                             MFlag_StolenBases_P
                                             X2_IMP_BattersHitByPitch_P
                                             MFlag_BattersHitByPitch_P
                                             INT_P
                                             INT_N
                                              / VIF;
          TITLE 'MODEL BASED ON STEPWISE RESULTS WITHOUT INCORRECT SIGNED COEFFIENT VARIABLES';


MODEL_FROM_FORW: MODEL TargetWins =          BaseHitsByBattersAllBases_P
                                             X1_WalksByBatters_P
                                             X1_Errors_N
                                             X1_HitsAllowed_N
                                             X1_IMP_CaughtStealing_N
                                             MFlag_CaughtStealing_N
                                             X2_TriplesByBatters3Bases_P
```

```
                                    X2_IMP_StolenBases_P
                                    MFlag_StolenBases_P
                                    X2_IMP_BattersHitByPitch_P
                                    MFlag_BattersHitByPitch_P
                                    INT_P
                                    INT_N
                                     / VIF;
            TITLE 'MODEL BASED ON FORWARD RESULTS WITHOUT INCORRECT SIGNED COEFFIENT VARIABLES';

MODEL_FROM_BACKW: MODEL TargetWins = BaseHitsByBattersAllBases_P
                                    X1_WalksByBatters_P
                                    X1_Errors_N
                                    X1_IMP_CaughtStealing_N
                                    MFlag_CaughtStealing_N
                                    X2_TriplesByBatters3Bases_P
                                    X2_IMP_StolenBases_P
                                    MFlag_StolenBases_P
                                    X2_IMP_BattersHitByPitch_P
                                    MFlag_BattersHitByPitch_P
                                    INT_P
                                    INT_N

                                     / VIF;
            TITLE 'MODEL BASED ON BACKWARD RESULTS WITHOUT INCORRECT SIGNED COEFFIENT VARIABLES';
RUN;
ods graphics off;
PROC PRINT DATA = ESTFILE;  RUN;
```

Appendix I – SAS Code for PCA EDA based on Stepwise Selected Model

```
***********************************************;
*	Principal Component Analysis based on the model output;
*	from the STEPWISE model above;
***********************************************;
* Create a new data set with only the variables to be used for PCA;
* based on the STEPWISE model result;
DATA moneyball_train_pca;
	SET moneyball_train;
	KEEP	BaseHitsByBattersAllBases_P
		X1_WalksByBatters_P
		X1_Errors_N
		X1_IMP_CaughtStealing_N
		MFlag_CaughtStealing_N
		X2_TriplesByBatters3Bases_P
		X2_IMP_StolenBases_P
		MFlag_StolenBases_P
		X2_IMP_BattersHitByPitch_P
		MFlag_BattersHitByPitch_P
		INT_P
		INT_N;
RUN;
* PRINCOMP STEP;
ods graphics on;
title 'Principal Components Analysis using PROC PRINCOMP';
	ITLE1 'based on the STEPWISE model';
proc princomp  data=moneyball_train_pca out=pca_components  outstat=eigenvectors plots=all;
run;  ods graphics off;
```

Appendix J – SAS Code for Building PCA Based Model at 94% of Variance with Reduced Dimensionality by Four (4) Variables Less

```
* BASED ON THE PCA BASED MODEL AT 94% WITH 8 VARIABLES OF THE 12 VARIABLES;

ods graphics on;

PROC REG DATA = moneyball_train outest=ESTFILE AIC SBC BIC CP ADJRSQ plots=diagnostics(stats=(default AIC SBC BIC CP ADJRSQ));

MODEL_PCA_BASED_94: MODEL TargetWins =         BaseHitsByBattersAllBases_P

                                               X1_WalksByBatters_P

                                               X1_Errors_N

                                               X1_IMP_CaughtStealing_N

                                               MFlag_CaughtStealing_N

                                               X2_TriplesByBatters3Bases_P

                                               X2_IMP_StolenBases_P

                                               MFlag_StolenBases_P

                                                       / VIF;

                TITLE 'MODEL BASED ON PCA RESULTS WITH ONLY 8 VARIABLS ACCOUNTING FOR 94% OF VARIANCE';


RUN;

ods graphics off;


PROC PRINT DATA = ESTFILE;

RUN;
```

Appendix K – SAS Code for PROC GLM

```
***********************************************************;
*	BINGO FOR PROC GLM;
***********************************************************;


PROC GLM	DATA = moneyball_train;
			MODEL TargetWins =	BaseHitsByBattersAllBases_P
						X1_WalksByBatters_P
						X1_Errors_N
						X1_IMP_CaughtStealing_N
						MFlag_CaughtStealing_N
						X2_TriplesByBatters3Bases_P
						X2_IMP_StolenBases_P
						MFlag_StolenBases_P
						X2_IMP_BattersHitByPitch_P
						MFlag_BattersHitByPitch_P
						INT_P
						INT_N / SS3;
	TITLE 'GLM Model based on the Stepwise Scoring model';
RUN;
quit;
```

Appendix L – SAS Code for PROC GENMOD

```
***************************************************************;
*	BINGO FOR PROC GENMODE;
***************************************************************;
proc genmod data=moneyball_train;
			MODEL TargetWins =	BaseHitsByBattersAllBases_P
						X1_WalksByBatters_P
						X1_Errors_N
						X1_IMP_CaughtStealing_N
						MFlag_CaughtStealing_N
						X2_TriplesByBatters3Bases_P
						X2_IMP_StolenBases_P
						MFlag_StolenBases_P
						X2_IMP_BattersHitByPitch_P
						MFlag_BattersHitByPitch_P
						INT_P
						INT_N / link=identity dist=normal;
run;
quit;
```

References

Allison, P., (2012) Logistic Regression Using SAS Theory and Application Second Edition. SAS Institute Inc., Cary, NC, USA.

Cody, R., (2011) SAS Statistics by Example. SAS Institute Inc., Cary, NC, USA.

Delwiche, L., Slaughter, S. (2012) The Little SAS Book. SAS Institute Inc., Cary, NC, USA.

Hoffmann, J., (2004) Generalized Linear Models. Pearson Education Inc.

Wedding, D., (2015) PREDICT 411 Generalized Linear Models – PowerPoint Course Content for LinearRegression, LinearRegression_DeployModel, FixMissingValues, Outliers, TransformValues, LinearRegression_ModelValidation. Northwestern University, Evanston, IL, USA.