The following is the write up for Project 1 – Moneyball OLS Regression
But first, here are the bonus points I attempted:

- **(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?**
  I ran my model with both proc glm and proc genmod for comparison. Both returned the same values for the coefficients and the intercept. There was no difference other than proc genmod output rounding up to four decimals:

	PROC GLM Output		PROC GENMOD Output	
Parameter	Estimate	Standard Error	Estimate	Standard Error
Intercept	78.02402089	11.06569887	78.024	11.0243
TEAM_BATTING_H	0.04848011	0.00340875	0.0485	0.0034
TEAM_BATTING_2B	-0.03548708	0.00865109	-0.0355	0.0086
TEAM_BATTING_3B	0.05430091	0.01575222	0.0543	0.0157
TEAM_BATTING_HR	0.07218825	0.00941294	0.0722	0.0094
TEAM_BATTING_BB	0.02859148	0.00456927	0.0286	0.0046
imp_team_batting_so	-0.01169451	0.00224199	-0.0117	0.0022
flag_team_batting_so	7.9826305	1.520612	7.9826	1.5149
imp_team_baserun_sb	0.04860853	0.005023	0.0486	0.005
flag_team_baserun_sb	32.94273538	1.79548675	32.9427	1.7888
imp_team_baserun_cs	0.00554873	0.01510471	0.0055	0.015
flag_team_baserun_cs	0.01635282	0.93974646	0.0164	0.9362
TEAM_FIELDING_E	-0.0566747	0.00331815	-0.0567	0.0033
log_imp_team_fieldin	-14.66892828	1.94450846	-14.6689	1.9372
flag_team_fielding_d	4.94571476	1.51316551	4.9457	1.5075
TEAM_PITCHING_BB	-0.00299296	0.00267881	-0.003	0.0027
TEAM_PITCHING_H	0.0020051	0.00038184	0.002	0.0004

- (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.

I created my own decision tree by extracting the values into a csv file and determining a value for team_baserun_cs based on a value for team_baserun_sb. Since both values are somewhat correlated (see write up), I did a little program to get an average for team_baserun_cs based on other known values by matching team_basedrun_sb with other records that did have a team_baserun_cs value and taking the average. I explain this on the writeup below. I did a program in Groovy (Java like language) to do this. Attached at end of write up.

- (10 Points) Hand in your SCORED FILE as a SAS DATA SET and save me to trouble of converting it. Submitting: score_file_ariel_gamino.sas7bdat

Introduction

The goal of this project is to analyze baseball data from the years 1871 to 2006 in order to come up with a model that can predict the number of times a team will win in a season. This will be accomplished by first performing data exploration and data preparation in order to deal with missing values or with values that may affect the model negatively. Secondly, a different set of SAS techniques will be used to select the model that has the best predictive capabilities. These techniques will include Forward, Backward, Stepwise and adjusted r-squared selection. Lastly, three chosen models will be compared via their r-square and AIC values and one final model will be chosen as the best model to use prediction of this data set.

Data Exploration

The data set given is based on baseball statistics from the years 1871 to 2006. It contains 16 performance metrics plus an index. Each record in the data represents a baseball team for a specific year and its respective metrics. These metrics include:

VARIABLE NAME	DEFINITION
INDEX	Identification Variable – unique and sequential for each record.
TARGET_WINS	
TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)
TEAM_BATTING_2B	Doubles by batters (2B)
TEAM_BATTING_3B	Triples by batters (3B)
TEAM_BATTING_HR	Homeruns by batters (4B)
TEAM_BATTING_BB	Walks by batters
TEAM_BATTING_HBP	Batters hit by pitch (get a free base)
TEAM_BATTING_SO	Strikeouts by batters
TEAM_BASERUN_SB	Stolen bases
TEAM_BASERUN_CS	Caught stealing
TEAM_FIELDING_E	Errors
TEAM_FIELDING_DP	Double Plays
TEAM_PITCHING_BB	Walks allowed
TEAM_PITCHING_H	Hits allowed
TEAM_PITCHING_HR	Homeruns allowed
TEAM_PITCHING_SO	Strikeouts by pitchers

Table 1. Data Dictionary for the baseball data set

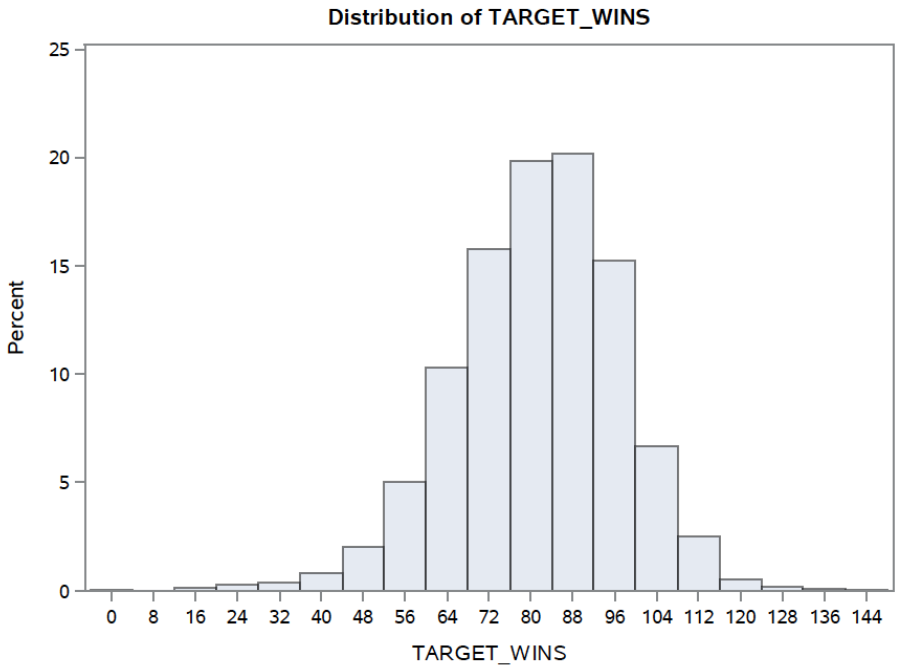
There were a total of 2,276 records. The target_wins variable gives the number of times a team has won for that season and the one that is used as the dependent variable for prediction.

Based on the data set, in average, each team has a won about 81 games. Other statistics including mean, median, standard deviation, maximum, minimum and range can be seen the following table.

Variable	Mean	Median	Std Dev	Minimum	Maximum	Range
TARGET_WINS	80.7908612	82	15.7521525	0	146	146
TEAM_BATTING_H	1469.27	1454	144.5911954	891	2,554	1,663
TEAM_BATTING_2B	241.2469244	238	46.8014146	69	458	389
TEAM_BATTING_3B	55.25	47	27.938557	0	223	223
TEAM_BATTING_HR	99.6120387	102	60.546872	0	264	264
TEAM_BATTING_BB	501.5588752	512	122.6708615	0	878	878
TEAM_BATTING_HBP	59.3560209	58	12.9671225	29	95	66
TEAM_BATTING_SO	735.6053358	750	248.5264177	0	1,399	1,399
TEAM_BASERUN_SB	124.7617716	101	87.791166	0	697	697
TEAM_BASERUN_CS	52.8038564	49	22.9563376	0	201	201
TEAM_FIELDING_E	246.4806678	159	227.7709724	65	1,898	1,833
TEAM_FIELDING_DP	146.3879397	149	26.2263853	52	228	176
TEAM_PITCHING_BB	553.0079086	536.5	166.3573617	0	3,645	3,645
TEAM_PITCHING_H	1779.21	1518	1406.84	1,137	30,132	28,995
TEAM_PITCHING_HR	105.698594	107	61.2987469	0	343	343
TEAM_PITCHING_SO	817.7304508	813.5	553.0850315	0	19,278	19,278

Table 2. Baseball data set basic statistics

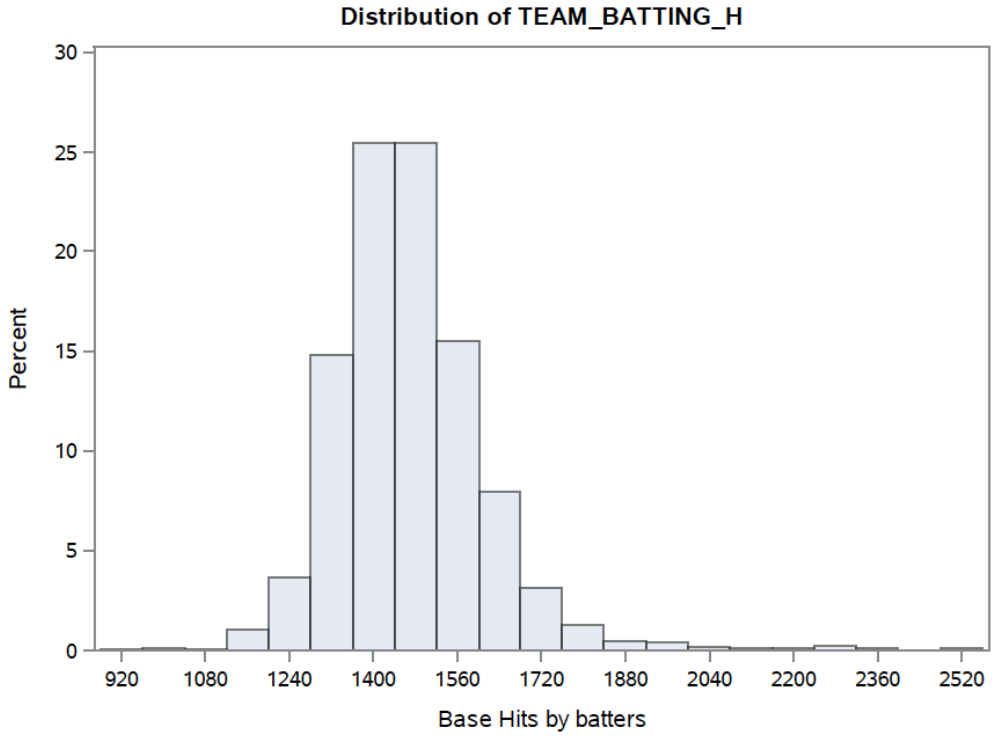
In order to get a feeling for the distribution of the data, I created the following tables, which show each variables distribution along with their percentiles. Beginning with the target variable target_wins, the rest of the variables follow. Target_wins seems to be normally distributed with an average number of wins at 81.



Quantiles (Definition 5)	
Level	Quantile
100% Max	146
99%	114
95%	104
90%	100
75% Q3	92
50% Median	82
25% Q1	71
10%	61
5%	54
1%	38
0% Min	0

Figure 1. target_wins distribution.

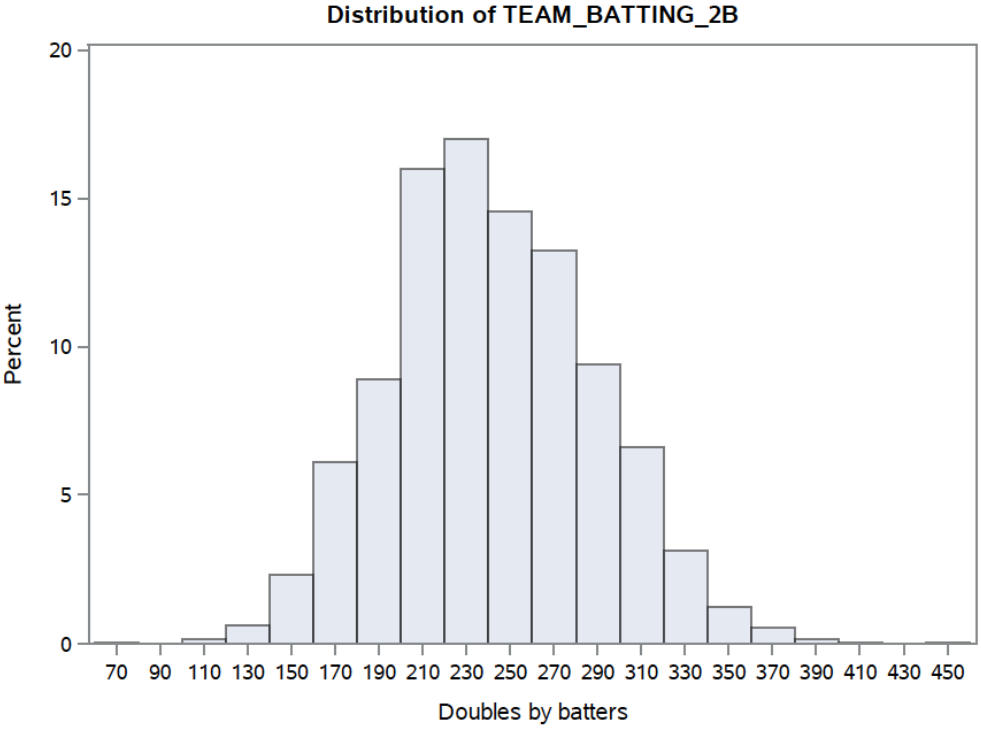
Team_batting_h is skewed to the left as can be seen below. There appear to be a few outliers past the 95% percentile.



Quantiles (Definition 5)	
Level	Quantile
100% Max	2554.0
99%	1950.0
95%	1696.0
90%	1636.0
75% Q3	1537.5
50% Median	1454.0
25% Q1	1383.0
10%	1315.0
5%	1280.0
1%	1188.0
0% Min	891.0

Figure 2. team_batting_h distribution.

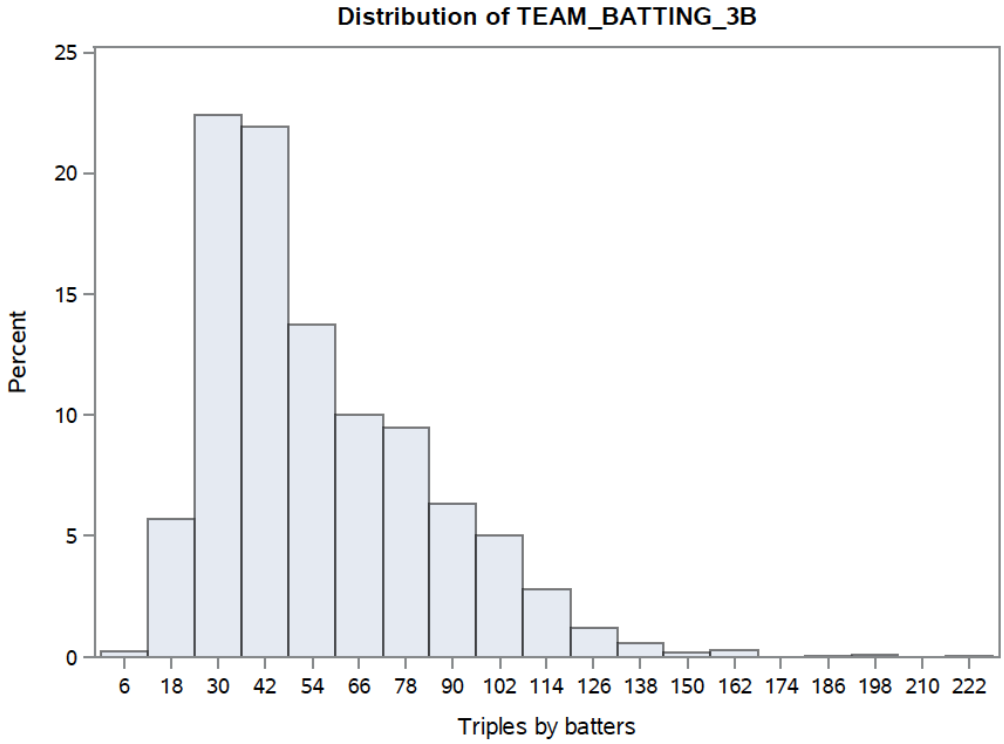
Team_batting_2b is close to a normal distribution but some outliers can be seen pass the 99% percentile as the values jump from 352 to 458.



Quantiles (Definition 5)	
Level	Quantile
100% Max	458
99%	352
95%	320
90%	303
75% Q3	273
50% Median	238
25% Q1	208
10%	182
5%	167
1%	141
0% Min	69

Figure 3. team_batting_2b distribution.

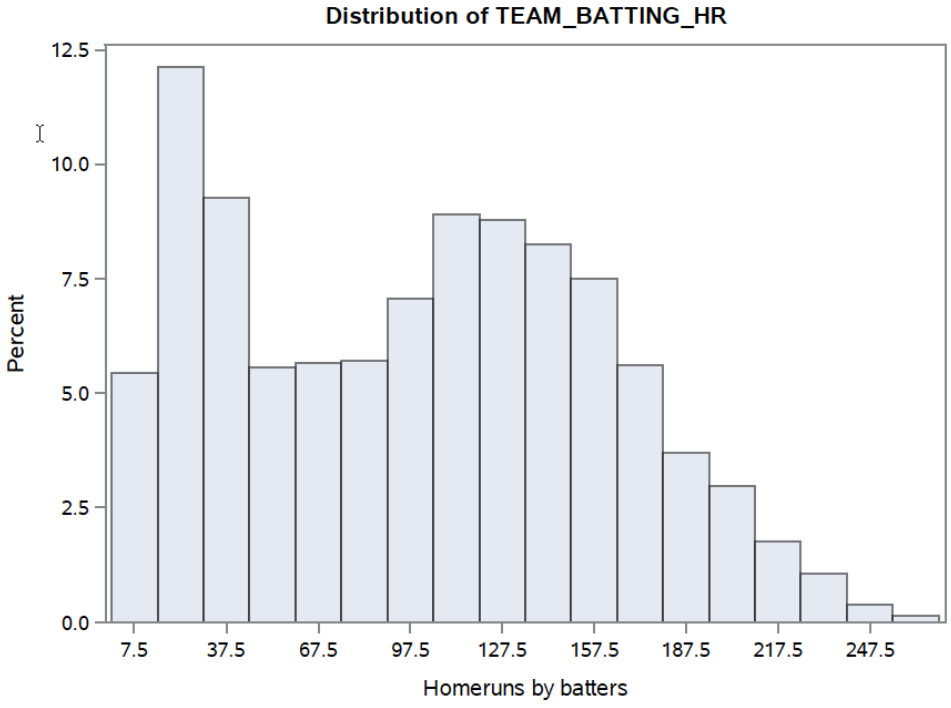
Team_batting_3b is clearly skewed to the left as the median value is 47. There are outliers pass the 50% percentile.



Quantiles (Definition 5)	
Level	Quantile
100% Max	223
99%	134
95%	108
90%	96
75% Q3	72
50% Median	47
25% Q1	34
10%	27
5%	23
1%	17
0% Min	0

Figure 4. Team_batting_3b distribution.

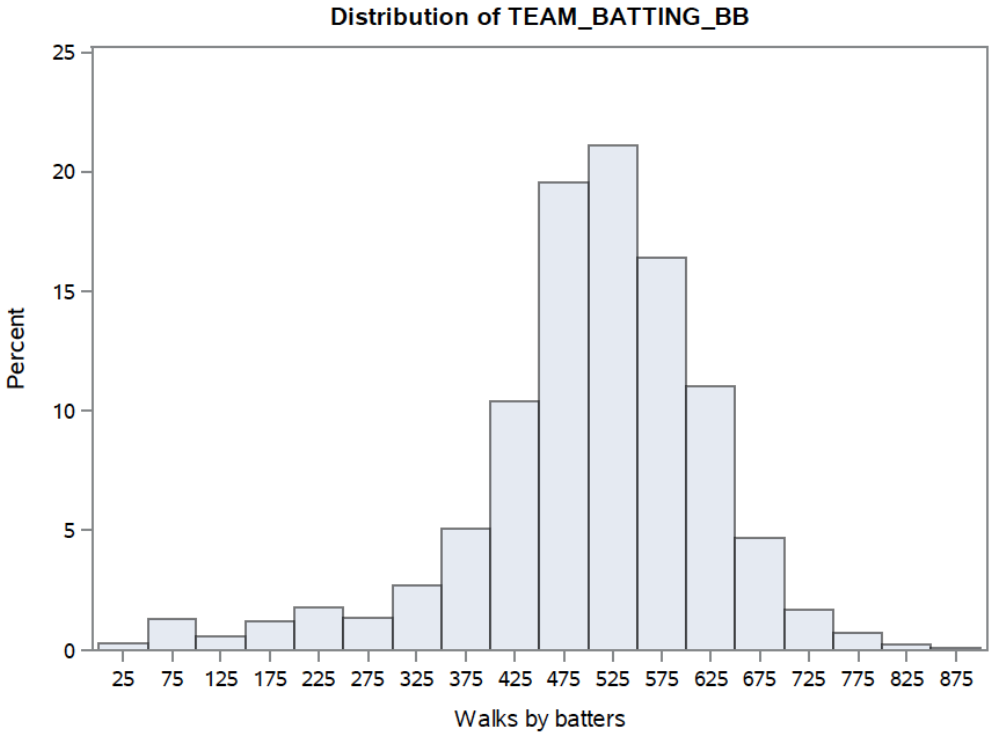
Team_batting_hr has too peaks at around the 25% and around the 75% percentile. Its values drops after the 90% percentile.



Quantiles (Definition 5)	
Level	Quantile
100% Max	264
99%	235
95%	199
90%	180
75% Q3	147
50% Median	102
25% Q1	42
10%	20
5%	14
1%	4
0% Min	0

Figure 5. team_batting_hr distribution.

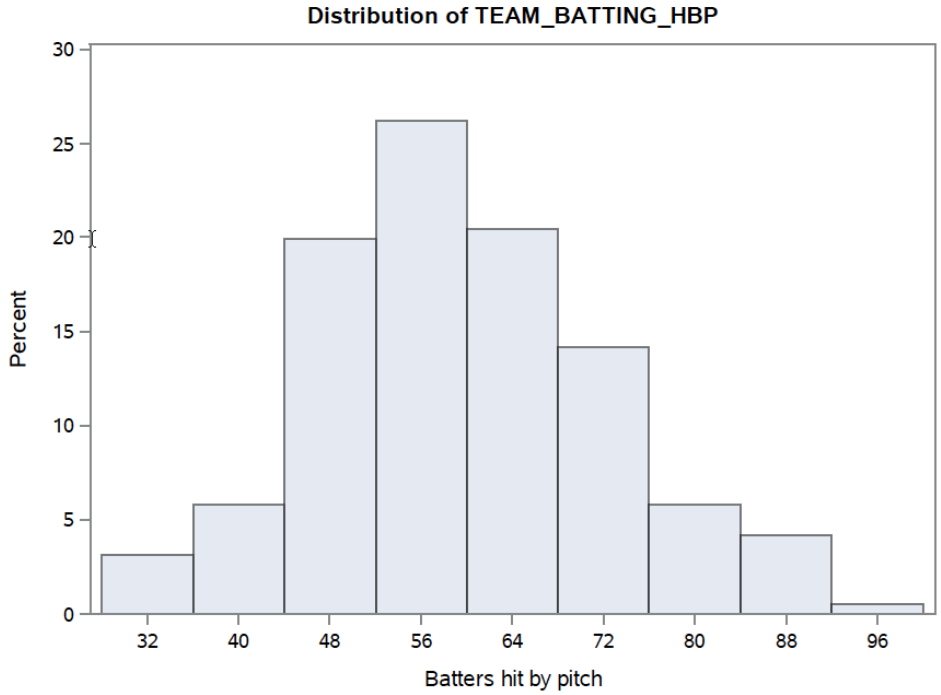
Team_batting_bb appears to have a normal distribution with a few outliers under the 5% percentile.



Quantiles (Definition 5)	
Level	Quantile
100% Max	878
99%	755
95%	671
90%	635
75% Q3	580
50% Median	512
25% Q1	451
10%	363
5%	246
1%	79
0% Min	0

Figure 6. team_batting_bb distribution.

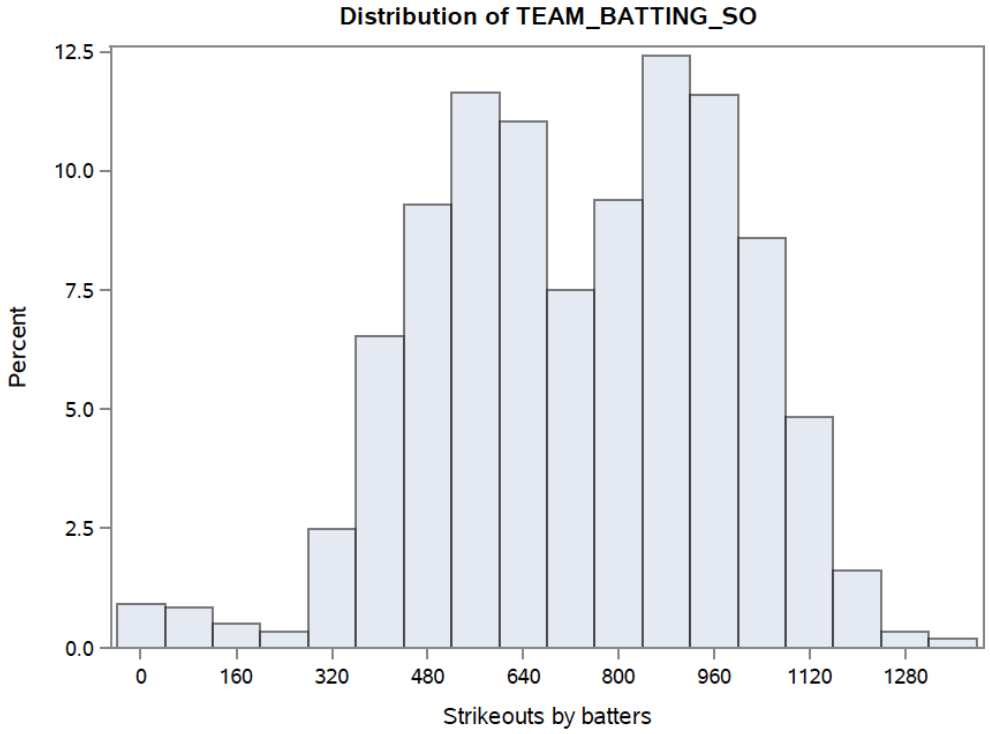
Team_batting_hbp values are normally distributed for the most part. There appears to be a few outliers after the 90% percentile.



Quantiles (Definition 5)	
Level	Quantile
100% Max	95
99%	90
95%	83
90%	76
75% Q3	67
50% Median	58
25% Q1	50
10%	44
5%	40
1%	29
0% Min	29

Figure 7. team_batting_hbp distribution.

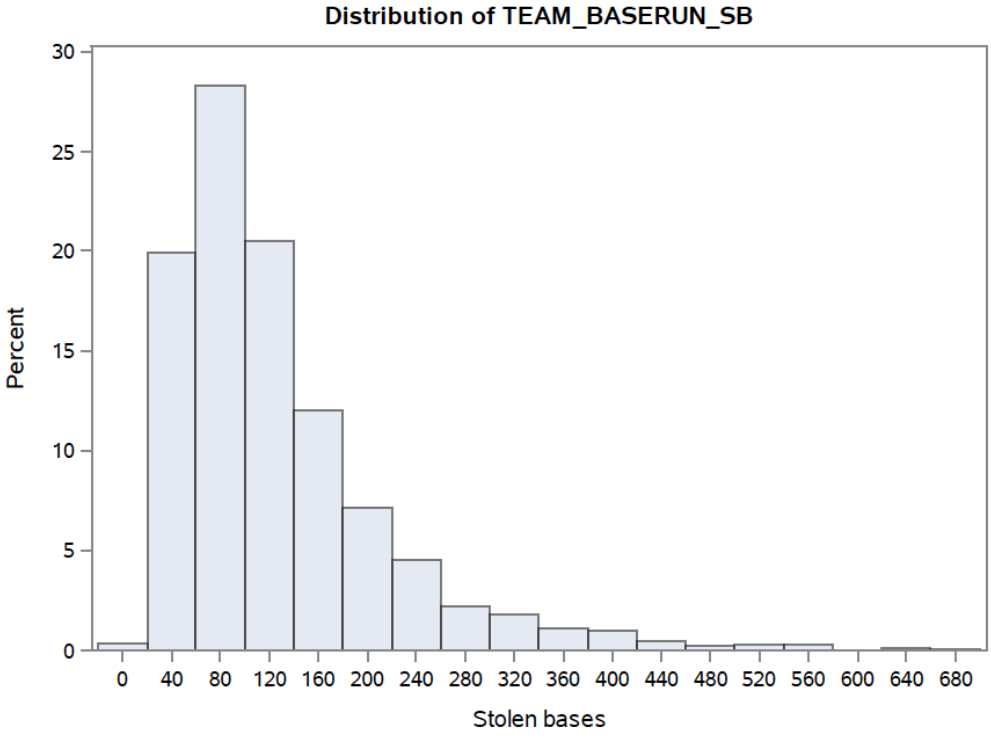
Team_batting_so has too peaks at around 25% and 80% percentile.



Quantiles (Definition 5)	
Level	Quantile
100% Max	1399
99%	1193
95%	1104
90%	1049
75% Q3	930
50% Median	750
25% Q1	548
10%	421
5%	359
1%	67
0% Min	0

Figure 8. team_batting_so distribution.

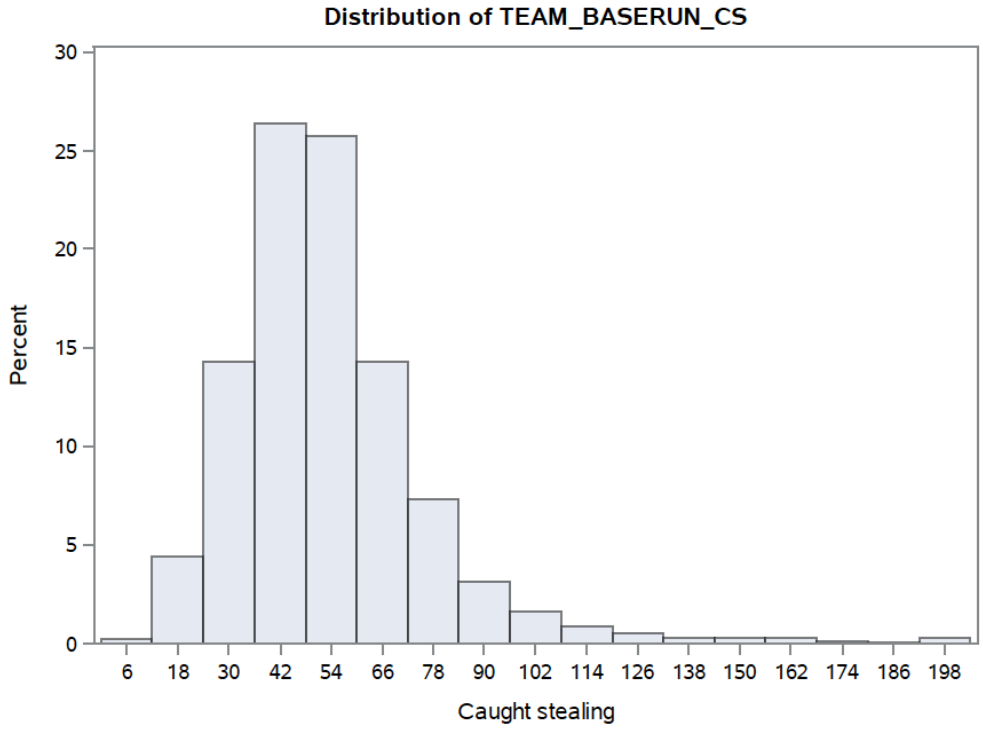
Team_baserun_sb is highly skewed to the left, this is probably due to the outliers found after the 95% percentile.



Quantiles (Definition 5)	
Level	Quantile
100% Max	697
99%	439
95%	302
90%	231
75% Q3	156
50% Median	101
25% Q1	66
10%	44
5%	35
1%	23
0% Min	0

Figure 9. team_baserun_sb distribution.

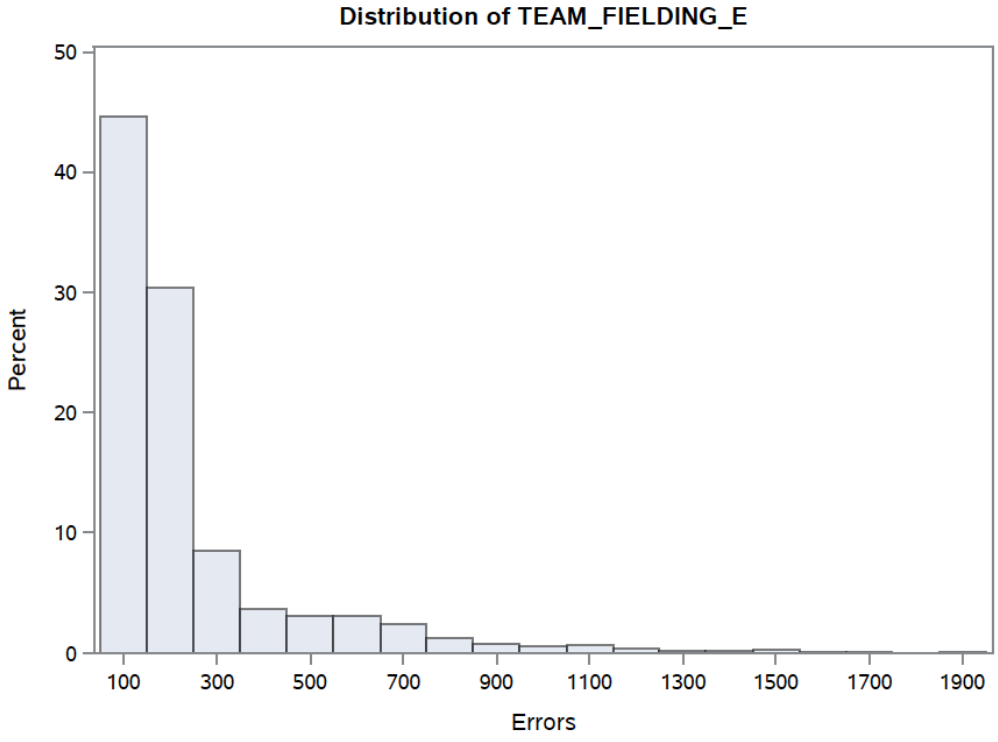
Team_baserun_cs is skewed to the left, most of the outliers forcing this skewedness appear after the 95% percentile.



Quantiles (Definition 5)	
Level	Quantile
100% Max	201
99%	143
95%	91
90%	77
75% Q3	62
50% Median	49
25% Q1	38
10%	30
5%	24
1%	16
0% Min	0

Figure 10. team_baserun_cs distribution.

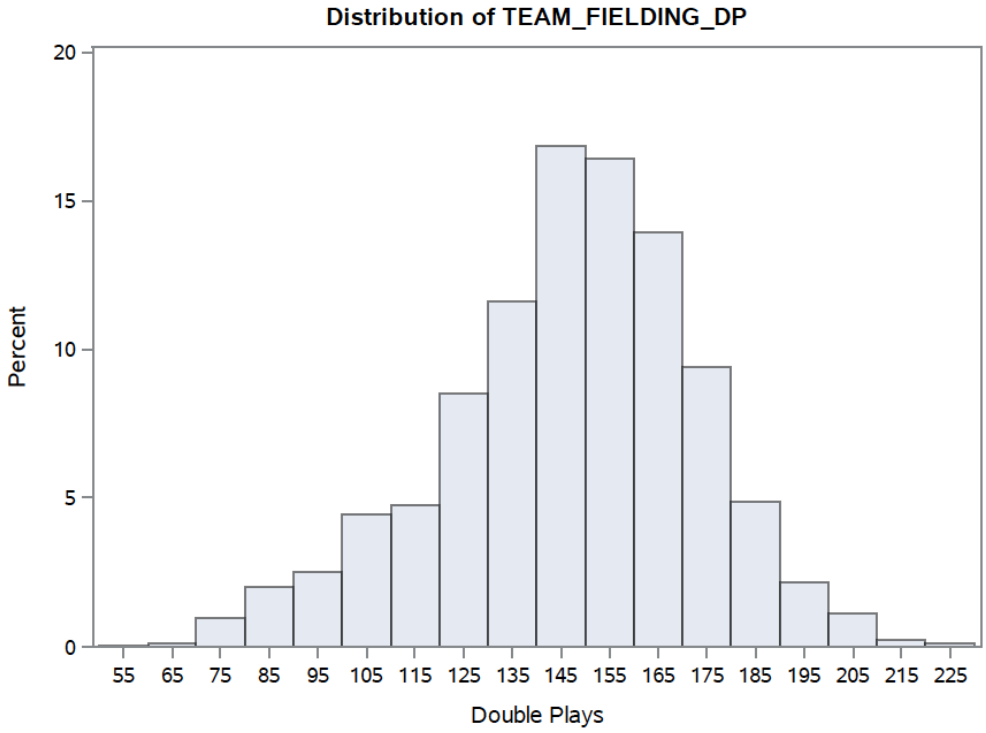
Team_fielding_e is highly skewed to the left, there are plenty of outliers after the 50% percentile.



Quantiles (Definition 5)	
Level	Quantile
100% Max	1898.0
99%	1237.0
95%	716.0
90%	542.0
75% Q3	249.5
50% Median	159.0
25% Q1	127.0
10%	109.0
5%	100.0
1%	86.0
0% Min	65.0

Figure 11. team_fielding_e distribution.

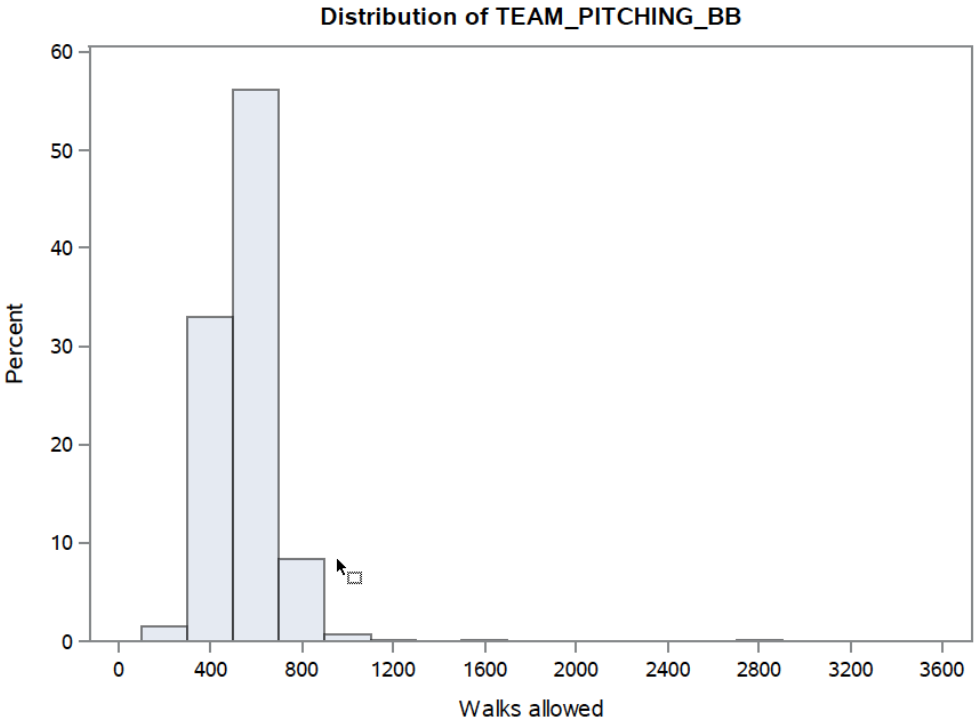
Team_fielding_dp has a normal distributions with a few outliers before the 5% percentile.



Quantiles (Definition 5)	
Level	Quantile
100% Max	228
99%	204
95%	186
90%	178
75% Q3	164
50% Median	149
25% Q1	131
10%	109
5%	98
1%	79
0% Min	52

Figure 12. team_fielding_dp distribution.

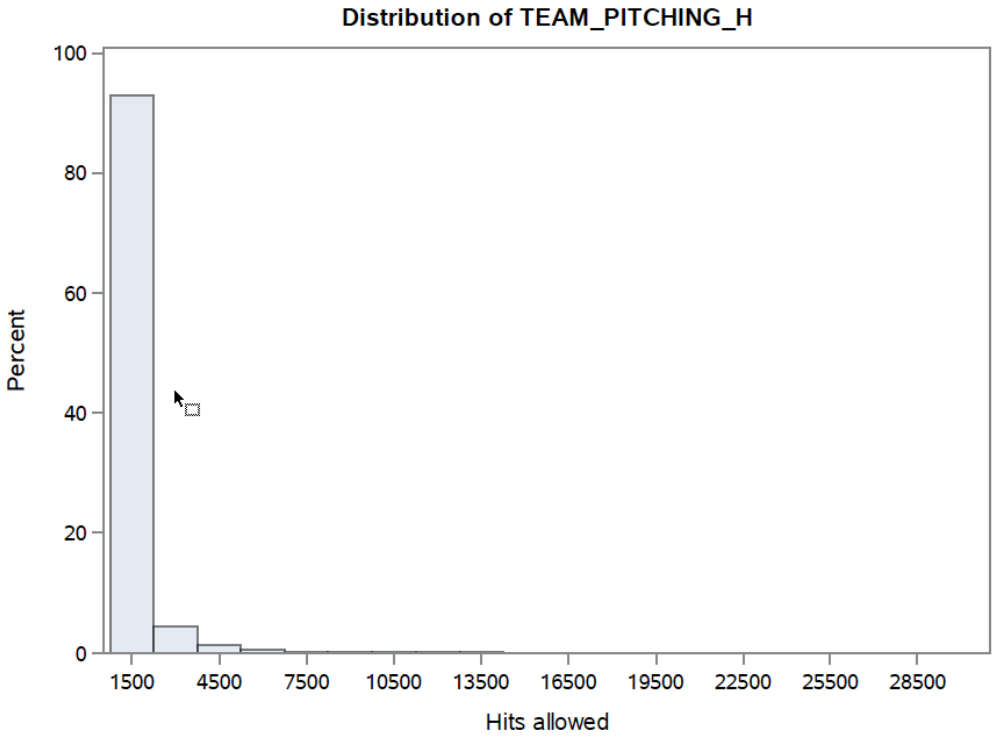
Team_pitching_bb is skewed to the left due to some very high values after the 99% percentile.



Quantiles (Definition 5)	
Level	Quantile
100% Max	3645.0
99%	924.0
95%	757.0
90%	694.0
75% Q3	611.0
50% Median	536.5
25% Q1	476.0
10%	417.0
5%	377.0
1%	237.0
0% Min	0.0

Figure 13. team_pitching_bb distribution.

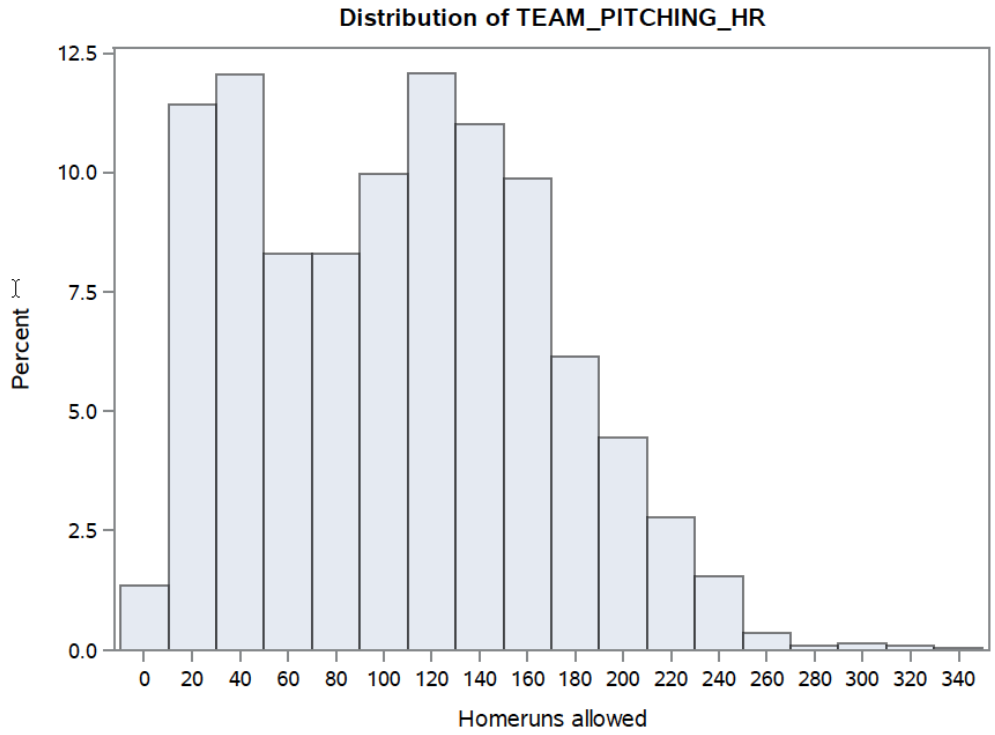
Team_pitching_h is highly skewed to the left with most of its values below the 50% percentile. There are outliers pass the 95% percentile.



Quantiles (Definition 5)	
Level	Quantile
100% Max	30132
99%	7093
95%	2563
90%	2059
75% Q3	1683
50% Median	1518
25% Q1	1419
10%	1356
5%	1316
1%	1244
0% Min	1137

Figure 14. team_pitching_h distribution.

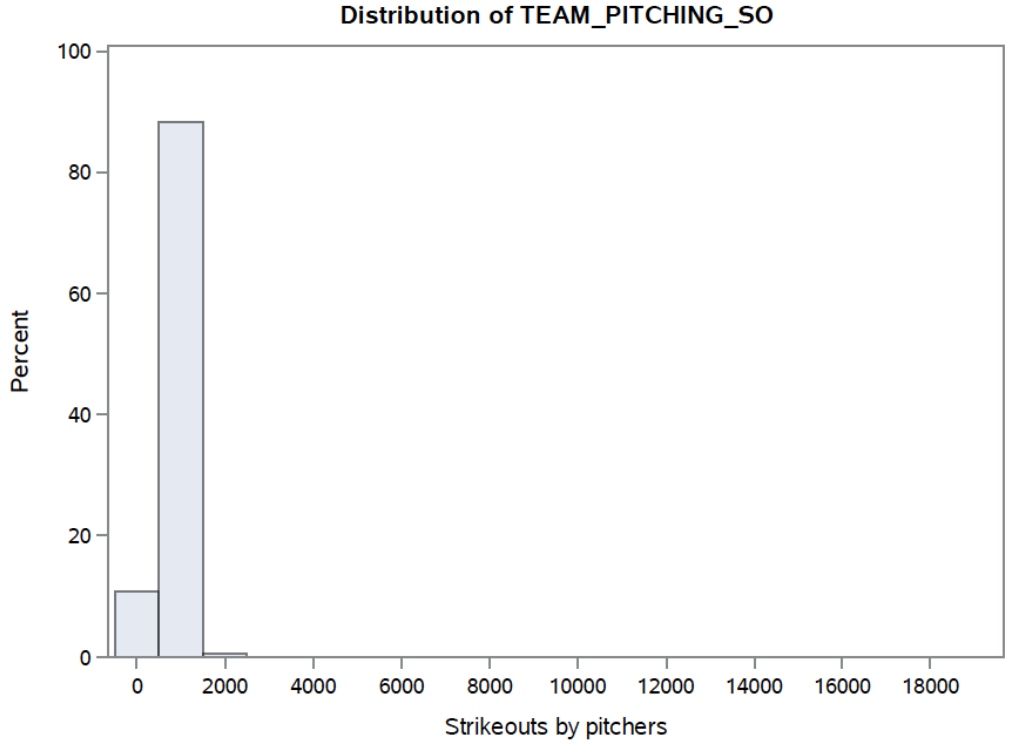
Team_pitching_hr has two peaks around the 25% and the 60% percentile.



Quantiles (Definition 5)	
Level	Quantile
100% Max	343
99%	244
95%	210
90%	187
75% Q3	150
50% Median	107
25% Q1	50
10%	25
5%	18
1%	8
0% Min	0

Figure 15. team_pitching_hr distribution.

Team_pitching_so is highly skewed to the left but very high value outliers pass its 99% percentile.



Quantiles (Definition 5)	
Level	Quantile
100% Max	19278.0
99%	1474.0
95%	1173.0
90%	1095.0
75% Q3	968.0
50% Median	813.5
25% Q1	615.0
10%	490.0
5%	420.0
1%	205.0
0% Min	0.0

Figure 16. team_pitching_so distribution.

These variables will be used to build a model to predict the target variable target_wins and part of this model will be determined by the variables that are correlated to it. The following table shows the amount of correlation between each variable and target_wins.

	TARGET_WINS
TEAM_BATTING_H	0.38877
TEAM_BATTING_2B	0.2891
TEAM_BATTING_BB	0.23256
TEAM_PITCHING_HR	0.18901
TEAM_BATTING_HR	0.17615
TEAM_BATTING_3B	0.14261
TEAM_BASERUN_SB	0.13514
TEAM_PITCHING_BB	0.12417
TEAM_BATTING_HBP	0.0735
TEAM_BASERUN_CS	0.0224
TEAM_BATTING_SO	-0.03175
TEAM_FIELDING_DP	-0.03485
TEAM_PITCHING_SO	-0.07844
TEAM_PITCHING_H	-0.10994
TEAM_FIELDING_E	-0.17648

Table 3. Correlation between target variable and independent variables

Table 3 shows that there is little correlation between most of the variables and target_wins. Two of them, team_batting_h (Base hits by batters) and team_batting_2b (doubles by batters), have the highest correlation at 0.38877 and 0.2891 respectively.

Although there are 2,276 records, not all the variables have all the values. In order to successfully build a linear regression model, those variables with missing values need to either be fixed or eliminated. The following table shows the variables that have values .

Variable	Label	Number of values	Values missing	Percentage Not Missing	Percentage missing
TEAM_BATTING_HBP	Batters hit by pitch	191	2085	8.39%	91.61%
TEAM_BASERUN_CS	Caught stealing	1504	772	66.08%	33.92%
TEAM_FIELDING_DP	Double Plays	1990	286	87.43%	12.57%
TEAM_BASERUN_SB	Stolen bases	2145	131	94.24%	5.76%
TEAM_BATTING_SO	Strikeouts by batters	2174	102	95.52%	4.48%
TEAM_PITCHING_SO	Strikeouts by pitchers	2174	102	95.52%	4.48%

Table 4. Variables with missing values

A missing value could indicate that a data was not collected for a specific metric or that it was lost at some point. It is important to account for these missing values and either eliminate the metric or use a value that would stand in for the missing one. For linear regression to work effectively, all values need to be accounted for.

Data Preparation

It is in the data preparation phase in which we make sure all missing values are fixed and any other necessary transformations are made in order to make a model that is highly predictive. We start by looking at how many of the values are missing for each of these variables. The column 'percentage missing' of table 4 is sorted from highest to smallest, and on it, it can be clearly shown that the variable team_batting_hbp has most of its values missing (at a 91.61% rate). The next variable, team_baserun_cs has about 1/3 (33.92%) of them missing as well. The last three variables, although do have values missing, they are in a smaller scale. team_baserun_sb has only 5.76% missing, team_batting_so 4.48% seam percentage as team_pitching_so.

To fix those missing values, we need to come up with a way of replacing them. For the variable which most of its values are missing, it would be nearly impossible to try to replace them with an accurate representation. As such, the first thing we do is to eliminate the team_batting_hbp variable from consideration when building the model.

Imputed variables are then created for the rest of those variables with missing values. Flag variables are also created to indicate that an imputed variable was used in the creation of the model. The original variable with missing values is then ignored since a newer variable without missing data is introduced.

The following table shows the imputed variables along with their respective flags and the values given in each case there was a missing value.

Original Variable	Imputed Variable	Flag for Imputed Variable	Fix for missing values	Value used for missing record
TEAM_BATTING_HBP	None	None	Eliminate variable from consideration	None
TEAM_BASERUN_CS	imp_team_baserun_cs	flag_team_baserun_cs	Use team_baserun_sb to lookup average value of those records where team_baserun_cs is not missing	Mean of lookup
TEAM_FIELDING_DP	imp_team_fielding_dp	flag_team_fielding_dp	Use mean	146
TEAM_BASERUN_SB	imp_team_baserun_sb	flag_team_baserun_sb	Use mean	125
TEAM_BATTING_SO	imp_team_batting_so	flag_team_batting_so	Use mean	736
TEAM_PITCHING_SO	imp_team_pitching_so	flag_team_pitching_so	Use mean	818

Table 5. Imputed variables and their values for fixing missing values.

The fix for team_fielding_dp, team_baserun_sb, team_batting_so, and team_pitching_so is pretty straightforward. For any records that contain missing values, use the mean of that variable and set such value into the imputed variable. The flag for each one of the imputed variables is set to 1 if a missing value was replaced, 0 if the original value was used. I used this technique because the amount of missing values was small (less than 13%) and using the mean would not influence the model too much.

For the value of team_baserun_cs, in which almost 34% was missing, I created a form of decision tree to generate the imputed value. Since team_baserun_cs (caught stealing) is related to team_baserun_sb (stolen bases), as indicated by their of 0.665524 (calculated through software), I could use one variable to come up with the value of the other one. The idea is as follows, if team_baserun_cs is missing, look up at the value of team_baserun_sb for that record. With that value look up all records with the same team_baserun_sb and calculate the mean for team_baserun_cs. Use this calculated average in the imputed variable. So we are using the average for team_baserun_cs when it's not missing given that the team_baserun_sb is the same.

The following table shows an example of this. The value for team_baserun_cs for the second record is missing. The value for team_baserun_sb for that record, which is 70, is used to look up other records with the same number in that variable. The mean of team_baserun_cs is then calculated and used in the imputed value for the record with the missing team_baserun_cs.

TEAM_BASERUN_SB	TEAM_BASERUN_CS	Value Missing	imp_team_baserun_cs	
70	45	No	45	
70		**Yes**	**45**	<-- 45 is the mean of team_baserun_cs for those records where team_baserun_sb is 70
70	31	No	31	

70	51	No	51
70	36	No	36
70	43	No	43
70	77	No	77
70	40	No	40
70	37	No	37

Table 6. Example of how imp_team_baserun_cs is calculated when missing.

This logic is also shown in the following diagram, which is an excerpt of the decision tree used to come up with the imputed value for team_baserun_cs. In this case, if the team_baserun_cs value is missing, it is checked against the values for teamp_baserun_sb, and the average is then given to the imputed value.

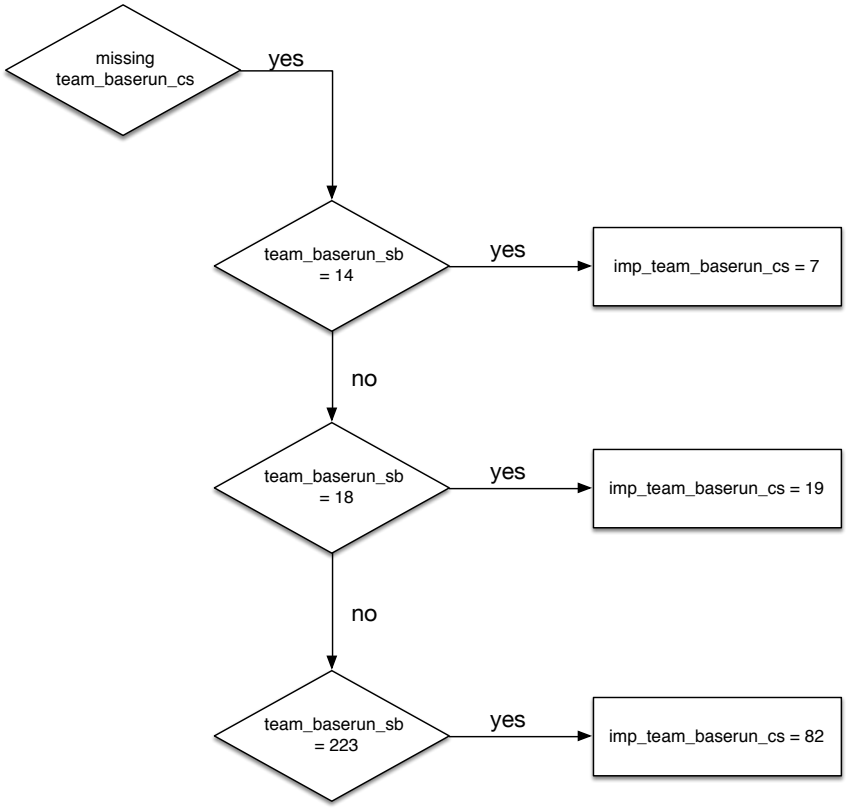


Figure 17. Excerpt of the decision tree used to calculate the imputed value for team_baserun_cs

The other technique used during this data preparation phase was to limit some of the values for some variables that were too skewed to the left or to the right. In other words, outliers were attempted to be eliminated by limiting their amount to greater than the 5% percentile and less that the 95% or 99%. In all cases, these limits covered most of the values found. The following table shows the variables that needed to be capped.

Variable	Capped Variable	Flag for Capped Variable	Value for Capped Variable
TEAM_BATTING_2B	cap_team_batting_2b	flag_cap_team_batting_2b	> 5% and <99%
TEAM_BATTING_3B	cap_team_batting_3b	flag_cap_team_batting_3b	> 5% and <99%
TEAM_PITCHING_H	imp_team_fielding_dp	flag_cap_team_pitching_bb	> 5% and <95%
TEAM_PITCHING_BB	imp_team_baserun_sb	flag_cap_team_fielding_e	> 5% and <99%
TEAM_FIELDING_E	cap_team_fielding_e	flag_cap_team_fielding_e	> 5% and <95%
imp_team_baserun_sb	cap_imp_team_baserun_sb	flag_cap_imp_team_baserun_sb	> 5% and <95%
imp_team_baserun_cs	cap_imp_team_baserun_cs	flag_cap_imp_team_baserun_sb	> 5% and <95%
imp_team_pitching_so	cap_imp_team_pitching_so	flag_camp_imp_team_pitching_so	> 5% and <95%

Table 7. Variables capped and their percentile in which they were capped.

To select the variables that needed to be capped, the original histograms found during the data exploration were studied. For instance figure 4 shows the skewed histogram of variable team_batting_3b. By looking at the percentiles and the values in each one, it was determined that any value less than its 5% percentile should be set to the 5% percentile value. Likewise, any value greater than its 99% percentile should be set to its 99% value. This process was repeated for all other variables that were capped.

Additionally as part of the data preparation, two variables were transformed by using the natural log function. A new variable log_team_batting_h was created based on the natural log of TEAM_BATTING_H. Likewise, the variable imp_team_fielding_dp was transformed into a new variable log_imp_team_fielding_dp. This was done to constrained outlier values in both variables.

Models

Once the data set was examined and the variables were transformed as to fix missing values or reduce the impact from outliers, it was ready for the creation of the models. Different models were built using either Backward, Forward and Stepwise selection as well as adjusted r-square. Based on the statistical values obtained, it was determined that in all cases for this data set, the Forward selection technique provided a better result. Out of all these models, three were selected and shown below.

Model 1

The first model built was created by using the Forward selection method. All variables created during the data preparation method were used, with the exception of those with a large VIF (multicollinearity) which were removed. The summary of the Forward selection is shown in the following figure.

Summary of Forward Selection

Step	Variable Entered	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	TEAM_BATTING_H	Base Hits by batters	1	0.1511	0.1511	592.027	404.89	<.0001
2	TEAM_BATTING_BB	Walks by batters	2	0.0683	0.2195	363.451	199.02	<.0001
3	log_imp_team_fielding_dp		3	0.0346	0.2541	248.712	105.39	<.0001
4	TEAM_BATTING_HR	Homeruns by batters	4	0.0124	0.2665	208.791	38.47	<.0001
5	cap_team_batting_3b		5	0.0123	0.2788	169.381	38.63	<.0001
6	cap_team_fielding_e		6	0.0095	0.2883	139.266	30.35	<.0001
7	cap_imp_team_baserun_sb		7	0.0113	0.2996	103.178	36.55	<.0001
8	cap_team_pitching_h		8	0.0053	0.3048	87.4637	17.12	<.0001
9	cap_team_pitching_bb		9	0.0092	0.3140	58.5509	30.26	<.0001
10	cap_team_batting_2b		10	0.0048	0.3188	44.4299	15.89	<.0001
11	flag_cap_team_pitching_h		11	0.0034	0.3222	34.7958	11.52	0.0007
12	flag_team_batting_so		12	0.0022	0.3245	29.2804	7.46	0.0064
13	flag_cap_team_batting_2b		13	0.0030	0.3274	21.2046	10.04	0.0015
14	flag_team_fielding_dp		14	0.0014	0.3288	18.4728	4.72	0.0298
15	flag_cap_team_pitching_bb		15	0.0013	0.3302	15.9222	4.55	0.0330
16	flag_cap_team_fielding_e		16	0.0006	0.3307	16.0647	1.86	0.1730
17	flag_cap_imp_team_baserun_sb		17	0.0005	0.3312	16.5082	1.56	0.2122
18	flag_cap_team_batting_3b		18	0.0004	0.3316	17.2000	1.31	0.2527
19	imp_team_batting_so		19	0.0003	0.3319	18.0474	1.15	0.2829

Figure 18. Model 1 – Summary of Forward Selection

The Analysis of variance and the parameter estimates are shown in the following figure.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	69.27957	11.62004	5.96	<.0001	0
TEAM_BATTING_H	Base Hits by batters	1	0.03292	0.00376	8.76	<.0001	4.01500
cap_team_batting_2b		1	-0.03119	0.00960	-3.25	0.0012	2.50222
flag_cap_team_batting_2b		1	-4.55978	1.28434	-3.55	0.0004	1.20953
cap_team_batting_3b		1	0.09471	0.01809	5.24	<.0001	3.19227
flag_cap_team_batting_3b		1	-1.37718	1.22181	-1.13	0.2598	1.08672
TEAM_BATTING_HR	Homeruns by batters	1	0.06032	0.01011	5.97	<.0001	5.09757
TEAM_BATTING_BB	Walks by batters	1	0.07136	0.00764	9.34	<.0001	11.94978
imp_team_batting_so		1	-0.00247	0.00230	-1.07	0.2829	4.25803
flag_team_batting_so		1	5.62061	1.61690	3.48	0.0005	1.52373
cap_imp_team_baserun_sb		1	0.03793	0.00593	6.40	<.0001	2.31846
flag_cap_imp_team_baserun_sb		1	1.10844	0.99709	1.11	0.2664	1.19153
cap_team_fielding_e		1	-0.04002	0.00484	-8.27	<.0001	9.29728
flag_cap_team_fielding_e		1	1.71646	1.13714	1.51	0.1313	1.51229
log_imp_team_fielding_dp		1	-15.79218	2.08313	-7.58	<.0001	1.93457
flag_team_fielding_dp		1	3.54040	1.68835	2.10	0.0361	4.26411
cap_team_pitching_bb		1	-0.04650	0.00652	-7.13	<.0001	7.04620
flag_cap_team_pitching_bb		1	3.00376	1.44074	2.08	0.0372	1.57691
cap_team_pitching_h		1	0.02049	0.00256	8.01	<.0001	8.44697
flag_cap_team_pitching_h		1	4.55459	1.14556	3.98	<.0001	1.59172

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	187373	9861.75525	58.99	<.0001
Error	2256	377123	167.16449		
Corrected Total	2275	564496			

Root MSE	12.92921	R-Square	0.3319
Dependent Mean	80.79086	Adj R-Sq	0.3263
Coeff Var	16.00331		

Figure 19. Model 1 – ANOVA table and Parameter Estimates

Model 1 generates the following linear regression equation with 19 variables:

target_wins = 69.27957+ 0.03292 * TEAM_BATTING_H +
-0.03119 * cap_team_batting_2b +
-4.55978 * flag_cap_team_batting_2b +
0.09471 * cap_team_batting_3b +
-1.37718 * flag_cap_team_batting_3b +
0.06032 * TEAM_BATTING_HR +
0.07136 * TEAM_BATTING_BB +
-0.00247 * imp_team_batting_so +
5.62061 * flag_team_batting_so +
0.03793 * cap_imp_team_baserun_sb +
1.10844 * flag_cap_imp_team_baserun_sb +
-0.04002 * cap_team_fielding_e +
1.71646 * flag_cap_team_fielding_e +
-15.79218 * log_imp_team_fielding_dp +
3.5404 * flag_team_fielding_dp +
-0.0465 * cap_team_pitching_bb +
3.00376 * flag_cap_team_pitching_bb +
0.02049 * cap_team_pitching_h +
4.55459 * flag_cap_team_pitching_h

The coefficients and their impact on the model are shown in table 8 below, they are sorted by coefficient size from positive to negative. As it can be seen on the table, the flag variables have a big impact on the model. For instance flag_team_batting_so (Strike out by batters) has a coefficient of 5.62061 (positive impact) while the variable team_batting_so has a coefficient of only -0.00247 (negative impact). Likewise flag_cap_team_pitching_h has a coefficient of 4.55459, while the variable cap_team_pitching_h has a coefficient of 0.02049. Variable flag_team_fielding_dp has a coefficient of 3.5404 while the variable itself, team_fielding_dp, has a negative impact of -15.79218. This was very surprising first, because team_fielding_dp (Double plays) would be thought of having a very positive outcome on the game, but in this model it has a negative one. The other variables that have a relative large negative impact on the model are those for the flags of flag_cap_team_batting_3b (-1.37718) and flag_cap_team_batting_2b (-4.55978) while their respective variables cap_team_batting_3b (0.09471) and cap_team_batting_2b (-0.03119) had relative small impact.

Model 1			
Coefficients	**Variables**	**Original Variable**	**Label**
5.62061	flag_team_batting_so		
4.55459	flag_cap_team_pitching_h		
3.5404	flag_team_fielding_dp		
3.00376	flag_cap_team_pitching_bb		
1.71646	flag_cap_team_fielding_e		

1.10844	flag_cap_imp_team_baserun_sb		
0.09471	cap_team_batting_3b	TEAM_BATTING_3B	Triples by batters (3B)
0.07136	TEAM_BATTING_BB	TEAM_BATTING_BB	Walks by batters
0.06032	TEAM_BATTING_HR	TEAM_BATTING_HR	Homeruns by batters (4B)
0.03793	cap_imp_team_baserun_sb	TEAM_BASERUN_SB	Stolen bases
0.03292	TEAM_BATTING_H	TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)
0.02049	cap_team_pitching_h	TEAM_PITCHING_H	Hits allowed
-0.00247	imp_team_batting_so	TEAM_BATTING_SO	Strikeouts by batters
-0.03119	cap_team_batting_2b	TEAM_BATTING_2B	Doubles by batters (2B)
-0.04002	cap_team_fielding_e	TEAM_FIELDING_E	Errors
-0.0465	cap_team_pitching_bb	TEAMP_PITCHING_BB	Walks allowed
-1.37718	flag_cap_team_batting_3b		
-4.55978	flag_cap_team_batting_2b		
-15.79218	log_imp_team_fielding_dp	TEAM_FIELDING_DP	Double Plays

Table 8. Model 1 variables and their coefficients.

Overall this model gave more weight to the flags than to the variables themselves. The biggest coefficient was a negative one for double plays. Some of the variables that would likely have a bigger impact on number of wins did have a positive impact (cap_team_batting_3b, TEAM_BATTING_BB, TEAM_BATTING_HR, cap_imp_team_baserun_sb, TEAM_BATTING_H) albeit a small one.

Model 2

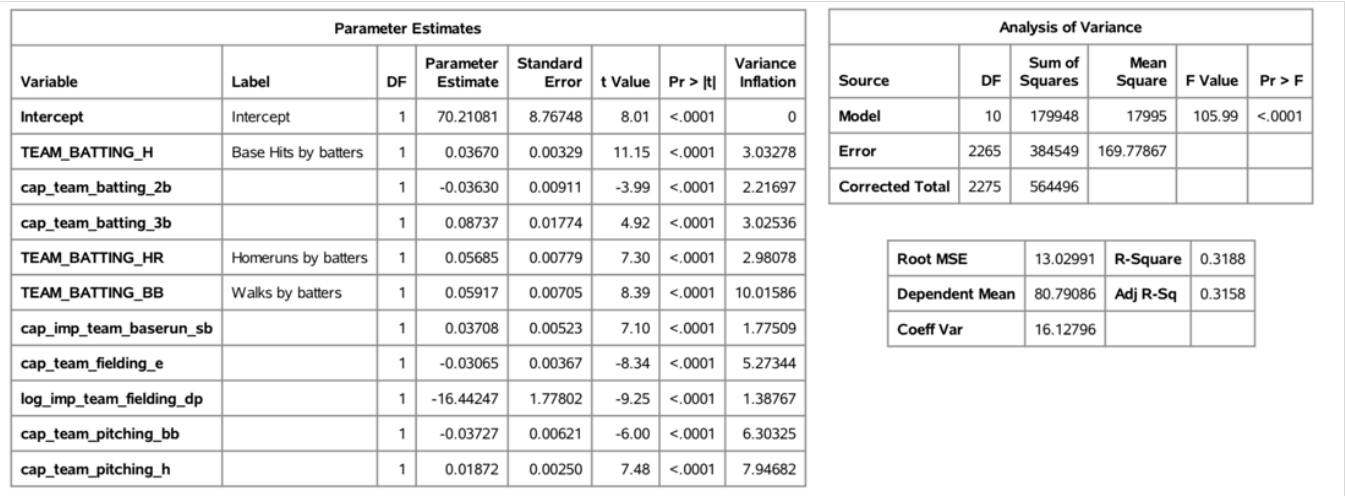
The second model was also built using the Forward selection method. It was, however, built without using any of the flags variables created during the data preparation phase. This was an attempt to see if these flags were indeed needed and whether a better model was built without it. Only 10 variables were generated during the forward selection method. The summary of this is shown in the following figure.

Summary of Forward Selection

Step	Variable Entered	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	TEAM_BATTING_H	Base Hits by batters	1	0.1511	0.1511	548.271	404.89	<.0001
2	TEAM_BATTING_BB	Walks by batters	2	0.0683	0.2195	323.217	199.02	<.0001
3	log_imp_team_fielding_dp		3	0.0346	0.2541	210.262	105.39	<.0001
4	TEAM_BATTING_HR	Homeruns by batters	4	0.0124	0.2665	170.982	38.47	<.0001
5	cap_team_batting_3b		5	0.0123	0.2788	132.205	38.63	<.0001
6	cap_team_fielding_e		6	0.0095	0.2883	102.580	30.35	<.0001
7	cap_imp_team_baserun_sb		7	0.0113	0.2996	67.0737	36.55	<.0001
8	cap_team_pitching_h		8	0.0053	0.3048	51.6302	17.12	<.0001
9	cap_team_pitching_bb		9	0.0092	0.3140	23.1897	30.26	<.0001
10	cap_team_batting_2b		10	0.0048	0.3188	9.3150	15.89	<.0001

Figure 20. Model 2 – Summary of Forward Selection

The ANOVA table and the parameter estimates are shown in the following figure.



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	70.21081	8.76748	8.01	<.0001	0
TEAM_BATTING_H	Base Hits by batters	1	0.03670	0.00329	11.15	<.0001	3.03278
cap_team_batting_2b		1	-0.03630	0.00911	-3.99	<.0001	2.21697
cap_team_batting_3b		1	0.08737	0.01774	4.92	<.0001	3.02536
TEAM_BATTING_HR	Homeruns by batters	1	0.05685	0.00779	7.30	<.0001	2.98078
TEAM_BATTING_BB	Walks by batters	1	0.05917	0.00705	8.39	<.0001	10.01586
cap_imp_team_baserun_sb		1	0.03708	0.00523	7.10	<.0001	1.77509
cap_team_fielding_e		1	-0.03065	0.00367	-8.34	<.0001	5.27344
log_imp_team_fielding_dp		1	-16.44247	1.77802	-9.25	<.0001	1.38767
cap_team_pitching_bb		1	-0.03727	0.00621	-6.00	<.0001	6.30325
cap_team_pitching_h		1	0.01872	0.00250	7.48	<.0001	7.94682

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	179948	17995	105.99	<.0001
Error	2265	384549	169.77867		
Corrected Total	2275	564496			

Root MSE	13.02991	R-Square	0.3188
Dependent Mean	80.79086	Adj R-Sq	0.3158
Coeff Var	16.12796		

Figure 21. Model 2 – ANOVA table and Parameter Estimates

Model 2 generates the following linear regression equation with 10 variables:

target_wins = 70.21081 + 0.0367 * TEAM_BATTING_H +
-0.0363 * cap_team_batting_2b +
0.08737 * cap_team_batting_3b +
0.05685 * TEAM_BATTING_HR +
0.05917 * TEAM_BATTING_BB +
0.03708 * cap_imp_team_baserun_sb +
-0.03065 * cap_team_fielding_e +
-16.44247 * log_imp_team_fielding_dp +
-0.03727 * cap_team_pitching_bb +
0.01872 * cap_team_pitching_h

Similar to our analysis of Model 1, Table 9 shows the coefficients for this model sorted by the ones that had the larger impact. Variables cap_team_batting_3b,TEAM_BATTING_BB, TEAM_BATTING_HR, cap_imp_team_baserun_sb ,TEAM_BATTING_H,cap_team_pitching_h all had positive impact. Although it was a small impact on the model, the only surprise was variable cap_team_pitching_h (Hits allowed), which in theory should have a negative impact on the winning of games. Relative similar to the previous model, the variable team_fielding_dp (Double Plays) had the largest and negative impact for the model (-16.44247).

Model 2			
Coefficients	**Variables**	**Original Variable**	**Label**
0.08737	cap_team_batting_3b	TEAM_BATTING_3B	Triples by batters (3B)
0.05917	TEAM_BATTING_BB	TEAM_BATTING_BB	Walks by batters
0.05685	TEAM_BATTING_HR	TEAM_BATTING_HR	Homeruns by batters (4B)
0.03708	cap_imp_team_baserun_sb	TEAM_BASERUN_SB	Stolen bases
0.0367	TEAM_BATTING_H	TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)

0.01872	cap_team_pitching_h	TEAM_PITCHING_H	Hits allowed
-0.03065	cap_team_fielding_e	TEAM_FIELDING_E	Errors
-0.0363	cap_team_batting_2b	TEAM_BATTING_2B	Doubles by batters (2B)
-0.03727	cap_team_pitching_bb	TEAMP_PITCHING_BB	Walks allowed
-16.44247	log_imp_team_fielding_dp	TEAM_FIELDING_DP	Double Plays

Table 9. Model 2 variables and their coefficients.

Getting rid of the flag variables made for a less complex model, but it created one with coefficients that were very small compared to Model 1.

Model 3

This model was also obtained by using the forward selection method. It was built by removing the capped variables created during the data preparation step. To my surprise, adding the capped variables created a model that was not as accurate as this one. The resulting model contains 14 variables. The summary of the forward selection is shown as follows.

Summary of Forward Selection

Step	Variable Entered	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	TEAM_BATTING_H	Base Hits by batters	1	0.1511	0.1511	987.919	404.89	<.0001
2	TEAM_FIELDING_E	Errors	2	0.0840	0.2351	667.337	249.62	<.0001
3	imp_team_baserun_sb		3	0.0377	0.2728	524.530	117.82	<.0001
4	flag_team_baserun_sb		4	0.0669	0.3397	269.764	229.96	<.0001
5	TEAM_BATTING_BB	Walks by batters	5	0.0137	0.3534	219.213	48.04	<.0001
6	log_imp_team_fielding_dp		6	0.0195	0.3729	146.298	70.58	<.0001
7	TEAM_BATTING_2B	Doubles by batters	7	0.0062	0.3791	124.635	22.51	<.0001
8	TEAM_PITCHING_H	Hits allowed	8	0.0084	0.3875	94.3171	31.15	<.0001
9	TEAM_BATTING_HR	Homeruns by batters	9	0.0045	0.3919	79.1576	16.65	<.0001
10	imp_team_batting_so		10	0.0062	0.3982	57.3065	23.37	<.0001
11	flag_team_batting_so		11	0.0067	0.4049	33.4122	25.65	<.0001
12	TEAM_BATTING_3B	Triples by batters	12	0.0039	0.4088	20.4303	14.93	0.0001
13	flag_team_fielding_dp		13	0.0026	0.4114	12.4284	10.01	0.0016
14	TEAM_PITCHING_BB	Walks allowed	14	0.0003	0.4117	13.1499	1.28	0.2581

Figure 22. Model 3 – Summary of Forward Selection

The ANOVA table and the parameter estimates for model 3 are shown in the following figure.

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	78.21168	10.52044	7.43	<.0001	0
TEAM_BATTING_H	Base Hits by batters	1	0.04854	0.00339	14.32	<.0001	3.72092
TEAM_BATTING_2B	Doubles by batters	1	-0.03559	0.00863	-4.12	<.0001	2.52801
TEAM_BATTING_3B	Triples by batters	1	0.05456	0.01557	3.50	0.0005	2.93186
TEAM_BATTING_HR	Homeruns by batters	1	0.07143	0.00913	7.82	<.0001	4.73615
TEAM_BATTING_BB	Walks by batters	1	0.02854	0.00452	6.31	<.0001	4.76475
imp_team_batting_so		1	-0.01164	0.00216	-5.38	<.0001	4.27584
flag_team_batting_so		1	7.90144	1.47869	5.34	<.0001	1.45047
imp_team_baserun_sb		1	0.04922	0.00465	10.58	<.0001	2.43303
flag_team_baserun_sb		1	32.98826	1.77592	18.58	<.0001	2.65118
TEAM_FIELDING_E	Errors	1	-0.05679	0.00330	-17.19	<.0001	8.76703
log_imp_team_fielding_dp		1	-14.65759	1.90141	-7.71	<.0001	1.83449
flag_team_fielding_dp		1	4.84582	1.47957	3.28	0.0011	3.72718
TEAM_PITCHING_BB	Walks allowed	1	-0.00299	0.00265	-1.13	0.2581	3.00442
TEAM_PITCHING_H	Hits allowed	1	0.00201	0.00037484	5.37	<.0001	4.30751

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	232422	16602	113.04	<.0001
Error	2261	332074	146.87060		
Corrected Total	2275	564496			

Root MSE	12.11902	R-Square	0.4117
Dependent Mean	80.79086	Adj R-Sq	0.4081
Coeff Var	15.00048		

Figure 23. Model 3 – ANOVA table and Parameter Estimates

Model 2 generates the following linear regression equation with 10 variables:

target_wins = 78.2117 +0.048537 * TEAM_BATTING_H +
0.071427 * TEAM_BATTING_HR +
0.028545 * TEAM_BATTING_BB +
-0.011644 * imp_team_batting_so +
7.90144 * flag_team_batting_so +
-14.6576 * log_imp_team_fielding_dp +
4.84582 * flag_team_fielding_dp +
-0.035594 * TEAM_BATTING_2B +
0.054563 * TEAM_BATTING_3B +
0.04922 * imp_team_baserun_sb +
32.9883 * flag_team_baserun_sb +
-0.056792 * TEAM_FIELDING_E +
-0.002994668 * TEAM_PITCHING_BB +
0.00201146 * TEAM_PITCHING_H

Table 10 shows the coefficients for each one of these variables and their corresponding impact on the model sorted from highest to lowest impact. Here we see the largest impact of the three models presented here for a coefficient. The flag variable flag_team_baserun_sb has a coefficient of 32.9883 while its corresponding variable, team_baserun_sb (Stolen Bases) had a much smaller impact of 0.04922. Similar to Model 1, flags used in building models had a much larger impact. Flag variables flag_team_batting_so (7.90144) and flag_team_fielding_dp (4.84582) had a relative high impact. The other variables that had a positive impact on the model TEAM_BATTING_HR (0.071427), TEAM_BATTING_3B (0.054563), imp_team_baserun_sb (0.04922), TEAM_BATTING_H (0.048537), TEAM_BATTING_BB (0.028545) , and TEAM_PITCHING_H (0.00201146), although these were much smaller in comparison to the flag variables. Also similar to the previous two models, variable log_imp_team_fielding_dp had the largest negative impact

with -14.6576. The variables with negative impact on the model had very small coefficients, as can be see on the following table.

Model 3			
Coefficients	**Variables**	**Original Variable**	**Label**
32.9883	flag_team_baserun_sb		
7.90144	flag_team_batting_so		
4.84582	flag_team_fielding_dp		
0.071427	TEAM_BATTING_HR	TEAM_BATTING_HR	Homeruns by batters (4B)
0.054563	TEAM_BATTING_3B	TEAM_BATTING_3B	Triples by batters (3B)
0.04922	imp_team_baserun_sb	TEAM_BASERUN_SB	Stolen bases
0.048537	TEAM_BATTING_H	TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)
0.028545	TEAM_BATTING_BB	TEAM_BATTING_BB	Walks by batters
0.00201146	TEAM_PITCHING_H	TEAM_PITCHING_H	Hits allowed
-0.002994668	TEAM_PITCHING_BB	TEAMP_PITCHING_BB	Walks allowed
-0.011644	imp_team_batting_so	TEAM_BATTING_SO	Strikeouts by batters
-0.035594	TEAM_BATTING_2B	TEAM_BATTING_2B	Doubles by batters (2B)
-0.056792	TEAM_FIELDING_E	TEAM_FIELDING_E	Errors
-14.6576	log_imp_team_fielding_dp	TEAM_FIELDING_DP	Double Plays

Table 10. Model 2 variables and their coefficients.

Out of the variables that would theoretically have a positive impact on the number of games won, for the most part they all had a positive outcome in the model (Homeruns by batters, triples by batters, stolen basis, base hits by batters, walks by batters). The one surprise was the flag flag_team_baserun_sb, which is the flag for a variable with a theoretically negative impact on wins, Strikeouts by batters (imp_team_batting_so), but in this case it had a big positive impact on the model. The flag indicates that a value was imputed (fixed) because of missing values. There's a clear case for stating that missing values had an impact on the model and that imputing the missing values had a positive impact in the overall model.

Model Selection

The tree models can be compared first by their adjusted r-square value. These values give an indication of how much the selected variables have an impact on the target_wins variable. Table 11 shows the r-square and adjusted r-square for each of the models. Figures 19, 21, and 23 previously show the ANOVA tables in full detail.

Model	R-Square	Adj R-Sq
Model 1	0.3319	0.3263
Model 2	0.3188	0.3158
Model 3	0.4117	**0.4081**

Table 11. R-Square and Adj R-Sq model comparison

If we are to use adjusted r-square as a measure for model performance, model 3 performed better than model 1 and model 3. A larger number for adjusted r-square is better.

Also, looking at the AIC values for each of the models, Table 12 shows that model 3 performs better. A smaller number for the calculated AIC gives a clear edge to this model.

Model	AIC
Model 1	11670.71
Model 2	11697.09
Model 3	**11371.17**

Table 12. AIC model comparison

Based on the calculations of adjusted r-square and AIC, Model 3 is a better model and the one selected for predicting the values in the baseball data set. This model performed better and it included flag and imputed variables. It was interesting to see that even though we calculated capped variables, omitting them from the selected model was beneficial, at least in the training data used to build this model.

Conclusion

The goal of this project was to come up with a model that determines the number of times a team will win in a season, was achieved by selecting a predictive model. Linear regression was used to come up with this model and three different models were compared. First, the data was analyzed and cleaned by eliminating missing values and then transformed into flag and capped variables that were used in the creation of the winning model. Although some of the selected variables were surprising, the model was demonstrated to predict to an extent the number of wins for a team based on the generated adjusted r-square and an acceptable AIC number.

Code

```
/*************************************************************************
* UNIT 01 - Predict 411 - Project 01 - Moneyball OLS Regression Baseball
* Data Preparation and Regression Program
* Ariel Gamino - arielgamino2016@u.northwestern.edu
*************************************************************************/

%let ME = arielgamino2016;
%let PATH = /home/&ME./my_courses/donald.wedding/c_8888/PRED411/UNIT01/HW;
%let NAME = HW;
%let LIB = &NAME..;

libname &NAME."&PATH.";

%let INFILE = HW.MONEYBALL;

/** Data Preparation **/
data tempfile_for_regression;
set &INFILE.;
   imp_team_baserun_sb = team_baserun_sb;
   flag_team_baserun_sb = 0;
   imp_team_baserun_cs = team_baserun_cs;
   flag_team_baserun_cs = 0;
   imp_team_fielding_dp = team_fielding_dp;
   flag_team_fielding_dp = 0;
   imp_team_batting_so = team_batting_so;
   flag_team_batting_so = 0;
   imp_team_pitching_so = team_pitching_so;
   flag_team_pitching_so = 0;

   if missing(team_baserun_sb) then do;
     * For missing value just use the average;
     imp_team_baserun_sb  = 125;
     flag_team_baserun_sb = 1;
   end;

   if missing(team_fielding_dp) then do;
     * For missing value just use the average;
     imp_team_fielding_dp  = 146;
     flag_team_fielding_dp = 1;
   end;

   if missing(team_batting_so) then do;
     * For missing value just use the average;
     imp_team_batting_so  = 736;
```

```
    flag_team_batting_so = 1;
  end;

  if missing(team_pitching_so) then do;
    * For missing value just use the average;
    imp_team_pitching_so  = 818;
    flag_team_pitching_so = 1;
  end;


        if missing(team_baserun_cs) then do;
          flag_team_baserun_cs = 1;
          * Poor man's decision tree for team_baserun_sb based on average value for
team_baserun_cs;
    if team_baserun_sb = 0 then imp_team_baserun_cs = 27;
    else if team_baserun_sb = 14 then imp_team_baserun_cs = 7;
    else if team_baserun_sb = 18 then imp_team_baserun_cs = 19;
          else if team_baserun_sb = 19 then imp_team_baserun_cs = 23;
          else if team_baserun_sb = 20 then imp_team_baserun_cs = 23;
          else if team_baserun_sb = 21 then imp_team_baserun_cs = 26;
          else if team_baserun_sb = 22 then imp_team_baserun_cs = 22;
          else if team_baserun_sb = 23 then imp_team_baserun_cs = 37;
          else if team_baserun_sb = 24 then imp_team_baserun_cs = 40;
          else if team_baserun_sb = 25 then imp_team_baserun_cs = 24;
          else if team_baserun_sb = 26 then imp_team_baserun_cs = 22;
          else if team_baserun_sb = 27 then imp_team_baserun_cs = 30;
          else if team_baserun_sb = 28 then imp_team_baserun_cs = 32;
          else if team_baserun_sb = 29 then imp_team_baserun_cs = 26;
          else if team_baserun_sb = 30 then imp_team_baserun_cs = 28;
          else if team_baserun_sb = 31 then imp_team_baserun_cs = 20;
          else if team_baserun_sb = 32 then imp_team_baserun_cs = 28;
          else if team_baserun_sb = 33 then imp_team_baserun_cs = 14;
          else if team_baserun_sb = 34 then imp_team_baserun_cs = 23;
          else if team_baserun_sb = 35 then imp_team_baserun_cs = 32;
          else if team_baserun_sb = 36 then imp_team_baserun_cs = 35;
          else if team_baserun_sb = 37 then imp_team_baserun_cs = 27;
          else if team_baserun_sb = 38 then imp_team_baserun_cs = 32;
          else if team_baserun_sb = 39 then imp_team_baserun_cs = 32;
          else if team_baserun_sb = 40 then imp_team_baserun_cs = 26;
          else if team_baserun_sb = 41 then imp_team_baserun_cs = 38;
          else if team_baserun_sb = 42 then imp_team_baserun_cs = 37;
          else if team_baserun_sb = 43 then imp_team_baserun_cs = 30;
          else if team_baserun_sb = 44 then imp_team_baserun_cs = 35;
          else if team_baserun_sb = 45 then imp_team_baserun_cs = 33;
          else if team_baserun_sb = 46 then imp_team_baserun_cs = 36;
          else if team_baserun_sb = 47 then imp_team_baserun_cs = 32;
          else if team_baserun_sb = 48 then imp_team_baserun_cs = 37;
```

```
else if team_baserun_sb = 49 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 50 then imp_team_baserun_cs = 38;
else if team_baserun_sb = 51 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 52 then imp_team_baserun_cs = 35;
else if team_baserun_sb = 53 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 54 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 55 then imp_team_baserun_cs = 45;
else if team_baserun_sb = 56 then imp_team_baserun_cs = 38;
else if team_baserun_sb = 57 then imp_team_baserun_cs = 39;
else if team_baserun_sb = 58 then imp_team_baserun_cs = 40;
else if team_baserun_sb = 59 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 60 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 61 then imp_team_baserun_cs = 39;
else if team_baserun_sb = 62 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 63 then imp_team_baserun_cs = 41;
else if team_baserun_sb = 64 then imp_team_baserun_cs = 42;
else if team_baserun_sb = 65 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 66 then imp_team_baserun_cs = 41;
else if team_baserun_sb = 67 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 68 then imp_team_baserun_cs = 38;
else if team_baserun_sb = 69 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 70 then imp_team_baserun_cs = 45;
else if team_baserun_sb = 71 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 72 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 73 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 74 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 75 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 76 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 77 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 78 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 79 then imp_team_baserun_cs = 48;
else if team_baserun_sb = 80 then imp_team_baserun_cs = 50;
else if team_baserun_sb = 81 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 83 then imp_team_baserun_cs = 48;
else if team_baserun_sb = 84 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 85 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 86 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 87 then imp_team_baserun_cs = 48;
else if team_baserun_sb = 88 then imp_team_baserun_cs = 40;
else if team_baserun_sb = 89 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 90 then imp_team_baserun_cs = 50;
else if team_baserun_sb = 92 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 93 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 94 then imp_team_baserun_cs = 58;
else if team_baserun_sb = 95 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 97 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 98 then imp_team_baserun_cs = 56;
```

```
else if team_baserun_sb = 100 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 101 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 102 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 103 then imp_team_baserun_cs = 60;
else if team_baserun_sb = 104 then imp_team_baserun_cs = 61;
else if team_baserun_sb = 105 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 106 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 107 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 109 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 110 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 111 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 112 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 113 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 114 then imp_team_baserun_cs = 64;
else if team_baserun_sb = 115 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 116 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 117 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 118 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 119 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 120 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 121 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 122 then imp_team_baserun_cs = 70;
else if team_baserun_sb = 123 then imp_team_baserun_cs = 58;
else if team_baserun_sb = 124 then imp_team_baserun_cs = 61;
else if team_baserun_sb = 125 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 126 then imp_team_baserun_cs = 65;
else if team_baserun_sb = 127 then imp_team_baserun_cs = 68;
else if team_baserun_sb = 128 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 129 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 130 then imp_team_baserun_cs = 56;
else if team_baserun_sb = 131 then imp_team_baserun_cs = 64;
else if team_baserun_sb = 132 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 133 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 134 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 135 then imp_team_baserun_cs = 64;
else if team_baserun_sb = 136 then imp_team_baserun_cs = 56;
else if team_baserun_sb = 137 then imp_team_baserun_cs = 65;
else if team_baserun_sb = 138 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 139 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 140 then imp_team_baserun_cs = 72;
else if team_baserun_sb = 141 then imp_team_baserun_cs = 50;
else if team_baserun_sb = 142 then imp_team_baserun_cs = 66;
else if team_baserun_sb = 143 then imp_team_baserun_cs = 40;
else if team_baserun_sb = 144 then imp_team_baserun_cs = 69;
else if team_baserun_sb = 145 then imp_team_baserun_cs = 71;
else if team_baserun_sb = 146 then imp_team_baserun_cs = 76;
else if team_baserun_sb = 147 then imp_team_baserun_cs = 69;
```

```
else if team_baserun_sb = 148 then imp_team_baserun_cs = 64;
else if team_baserun_sb = 149 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 150 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 151 then imp_team_baserun_cs = 50;
else if team_baserun_sb = 152 then imp_team_baserun_cs = 70;
else if team_baserun_sb = 153 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 154 then imp_team_baserun_cs = 68;
else if team_baserun_sb = 155 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 156 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 157 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 158 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 159 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 160 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 161 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 162 then imp_team_baserun_cs = 60;
else if team_baserun_sb = 163 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 164 then imp_team_baserun_cs = 78;
else if team_baserun_sb = 165 then imp_team_baserun_cs = 69;
else if team_baserun_sb = 166 then imp_team_baserun_cs = 60;
else if team_baserun_sb = 167 then imp_team_baserun_cs = 88;
else if team_baserun_sb = 168 then imp_team_baserun_cs = 42;
else if team_baserun_sb = 169 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 170 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 171 then imp_team_baserun_cs = 69;
else if team_baserun_sb = 172 then imp_team_baserun_cs = 88;
else if team_baserun_sb = 173 then imp_team_baserun_cs = 31;
else if team_baserun_sb = 174 then imp_team_baserun_cs = 61;
else if team_baserun_sb = 175 then imp_team_baserun_cs = 48;
else if team_baserun_sb = 176 then imp_team_baserun_cs = 117;
else if team_baserun_sb = 177 then imp_team_baserun_cs = 94;
else if team_baserun_sb = 178 then imp_team_baserun_cs = 30;
else if team_baserun_sb = 179 then imp_team_baserun_cs = 75;
else if team_baserun_sb = 180 then imp_team_baserun_cs = 65;
else if team_baserun_sb = 182 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 183 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 184 then imp_team_baserun_cs = 41;
else if team_baserun_sb = 186 then imp_team_baserun_cs = 60;
else if team_baserun_sb = 187 then imp_team_baserun_cs = 79;
else if team_baserun_sb = 188 then imp_team_baserun_cs = 35;
else if team_baserun_sb = 189 then imp_team_baserun_cs = 89;
else if team_baserun_sb = 190 then imp_team_baserun_cs = 87;
else if team_baserun_sb = 191 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 192 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 193 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 194 then imp_team_baserun_cs = 74;
else if team_baserun_sb = 195 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 196 then imp_team_baserun_cs = 59;
```

```
        else if team_baserun_sb = 197 then imp_team_baserun_cs = 66;
        else if team_baserun_sb = 198 then imp_team_baserun_cs = 58;
        else if team_baserun_sb = 200 then imp_team_baserun_cs = 46;
        else if team_baserun_sb = 201 then imp_team_baserun_cs = 96;
        else if team_baserun_sb = 202 then imp_team_baserun_cs = 55;
        else if team_baserun_sb = 207 then imp_team_baserun_cs = 56;
        else if team_baserun_sb = 208 then imp_team_baserun_cs = 58;
        else if team_baserun_sb = 209 then imp_team_baserun_cs = 51;
        else if team_baserun_sb = 210 then imp_team_baserun_cs = 29;
        else if team_baserun_sb = 211 then imp_team_baserun_cs = 142;
        else if team_baserun_sb = 212 then imp_team_baserun_cs = 168;
        else if team_baserun_sb = 214 then imp_team_baserun_cs = 72;
        else if team_baserun_sb = 220 then imp_team_baserun_cs = 71;
        else if team_baserun_sb = 221 then imp_team_baserun_cs = 61;
        else if team_baserun_sb = 222 then imp_team_baserun_cs = 51;
        else if team_baserun_sb = 223 then imp_team_baserun_cs = 82;
        else if team_baserun_sb = 231 then imp_team_baserun_cs = 86;
        else if team_baserun_sb = 232 then imp_team_baserun_cs = 87;
        else if team_baserun_sb = 234 then imp_team_baserun_cs = 64;
        else if team_baserun_sb = 235 then imp_team_baserun_cs = 99;
        else if team_baserun_sb = 237 then imp_team_baserun_cs = 82;
        else if team_baserun_sb = 239 then imp_team_baserun_cs = 37;
        else if team_baserun_sb = 245 then imp_team_baserun_cs = 97;
        else if team_baserun_sb = 246 then imp_team_baserun_cs = 100;
        else if team_baserun_sb = 247 then imp_team_baserun_cs = 200;
        else if team_baserun_sb = 248 then imp_team_baserun_cs = 36;
        else if team_baserun_sb = 264 then imp_team_baserun_cs = 140;
        else if team_baserun_sb = 314 then imp_team_baserun_cs = 96;
  else team_baserun_cs = 53;
end;

*Catch all;
if missing(imp_team_baserun_cs) then do;
       imp_team_baserun_cs = 53;
       flag_team_baserun_cs = 1;
end;

*Transform Variables;
cap_team_batting_2b = TEAM_BATTING_2B;
        flag_cap_team_batting_2b = 0;
        cap_team_batting_3b = TEAM_BATTING_3B;
        flag_cap_team_batting_3b = 0;
        cap_team_pitching_h = TEAM_PITCHING_H;
        flag_cap_team_pitching_h = 0;
        cap_team_pitching_bb = TEAM_PITCHING_BB;
        flag_cap_team_pitching_bb = 0;
        cap_team_fielding_e = TEAM_FIELDING_E;
```

```
        flag_cap_team_fielding_e = 0;
        cap_imp_team_baserun_sb = imp_team_baserun_sb;
        flag_cap_imp_team_baserun_sb = 0;
        cap_imp_team_baserun_cs = imp_team_baserun_cs;
        flag_cap_imp_team_baserun_cs = 0;
        cap_imp_team_pitching_so = imp_team_pitching_so;
        flag_camp_imp_team_pitching_so = 0;

        *Cap TEAM_BATTING_2B to 5% > 167, 352 < 99%;
        if TEAM_BATTING_2B < 167.00 then do;
    cap_team_batting_2b = 167.00;
    flag_cap_team_batting_2b = 1;
end;
if TEAM_BATTING_2B > 352.00 then do;
    cap_team_batting_2b = 352.00;
    flag_cap_team_batting_2b = 1;
end;

*Cap TEAM_BATTING_3B to 5% > 23, 134 < 99%;
if TEAM_BATTING_3B < 23.00 then do;
    cap_team_batting_3b = 23.00;
    flag_cap_team_batting_3b = 1;
end;
if TEAM_BATTING_3B > 134.00 then do;
    cap_team_batting_3b = 134.00;
    flag_cap_team_batting_3b = 1;
end;

*Cap TEAM_PITCHING_H to 5% > 1316, 2563 < 95%;
if TEAM_PITCHING_H < 1316.00 then do;
    cap_team_pitching_h = 1316.00;
    flag_cap_team_pitching_h = 1;
end;
if TEAM_PITCHING_H > 2563.00 then do;
    cap_team_pitching_h = 2563.00;
    flag_cap_team_pitching_h = 1;
end;

*Cap TEAM_PITCHING_BB to 5% > 377, 924 < 99%;
if TEAM_PITCHING_BB < 377.00 then do;
    cap_team_pitching_bb = 377.00;
    flag_cap_team_pitching_bb = 1;
end;
if TEAM_PITCHING_BB > 924.00 then do;
    cap_team_pitching_bb = 924.00;
    flag_cap_team_pitching_bb = 1;
end;
```

```
    *Cap TEAM_FIELDING_E to 5% > 100, 716 < 95%;
    if TEAM_FIELDING_E < 100.00 then do;
        cap_team_fielding_e = 100.00;
        flag_cap_team_fielding_e = 1;
    end;
    if TEAM_FIELDING_E > 716.00 then do;
        cap_team_fielding_e = 716.00;
        flag_cap_team_fielding_e = 1;
    end;

    *Cap imp_team_baserun_sb to 5% > 36, 298 < 95%;
    if imp_team_baserun_sb < 36.00 then do;
        cap_imp_team_baserun_sb = 36.00;
        flag_cap_imp_team_baserun_sb = 1;
    end;
    if imp_team_baserun_sb > 298.00 then do;
        cap_imp_team_baserun_sb = 298.00;
        flag_cap_imp_team_baserun_sb = 1;
    end;

    *Cap imp_team_baserun_cs to 5% > 25, 91 < 95%;
    if imp_team_baserun_cs < 25.00 then do;
        cap_imp_team_baserun_cs = 25.00;
        flag_cap_imp_team_baserun_cs = 1;
    end;
    if imp_team_baserun_cs > 91.00 then do;
        cap_imp_team_baserun_cs = 91.00;
        flag_cap_imp_team_baserun_cs = 1;
    end;

    *Cap imp_team_pitching_so to 5% > 423, 1169 < 95%;
    if imp_team_pitching_so < 423.00 then do;
        cap_imp_team_pitching_so = 423.00;
        flag_camp_imp_team_pitching_so = 1;
    end;
    if imp_team_pitching_so > 1169.00 then do;
        cap_imp_team_pitching_so = 1169.00;
        flag_camp_imp_team_pitching_so = 1;
    end;

          log_team_batting_h = log(TEAM_BATTING_H+1);
    log_imp_team_fielding_dp = log(imp_team_fielding_dp+1);

        drop team_batting_HBP; *Too many missing values, drop;
        drop index;
run;
```

```
ods graphics on;
title "Trying forward backward stepwise";
/** Linear Regression for transformed variables **/
/** Model 1 **/
title "Model 1 - Proc reg forward for tempfile_for_regresion FULL model (removed high VIFs)";
proc reg data=tempfile_for_regression outest=model1;
model target_wins =
    team_batting_h
    cap_team_batting_2b
    flag_cap_team_batting_2b
    cap_team_batting_3b
    flag_cap_team_batting_3b
    TEAM_BATTING_HR
    TEAM_BATTING_BB
    imp_team_batting_so
    flag_team_batting_so
    cap_imp_team_baserun_sb
    flag_cap_imp_team_baserun_sb
    cap_imp_team_baserun_cs
    flag_cap_imp_team_baserun_cs
    cap_team_fielding_e
    flag_cap_team_fielding_e
    log_imp_team_fielding_dp
    flag_team_fielding_dp
    cap_team_pitching_bb
    flag_cap_team_pitching_bb
    cap_team_pitching_h
    flag_cap_team_pitching_h
    /selection=forward vif sse aic;
run;

*print values including AIC;
title "Model 1 - Print Values";
proc print data=model1;
run;
quit;

/** Model 2 **/
title "Model 2 - Proc reg forward for tempfile_for_regresion No FLAGS model";
proc reg data=tempfile_for_regression outest=model2;
model target_wins =
    team_batting_h
    cap_team_batting_2b
    cap_team_batting_3b
    TEAM_BATTING_HR
```

```
    TEAM_BATTING_BB
    imp_team_batting_so
    cap_imp_team_baserun_sb
    cap_imp_team_baserun_cs
    cap_team_fielding_e
    log_imp_team_fielding_dp
    cap_team_pitching_bb
    cap_team_pitching_h
    /selection=forward vif sse aic;
run;
quit;

*print values including AIC;
title "Model 2 - Print Values";
proc print data=model2;
run;
quit;

/** Model 3 **/
title "Model 3 - Proc reg forward for tempfile_for_regresion NO CAPS model";
proc reg data=tempfile_for_regression outest=model3;
model target_wins =
    team_batting_h
    TEAM_BATTING_2B
    TEAM_BATTING_3B
    TEAM_BATTING_HR
    TEAM_BATTING_BB
    imp_team_batting_so
    flag_team_batting_so
    imp_team_baserun_sb
    flag_team_baserun_sb
    imp_team_baserun_cs
    flag_team_baserun_cs
    TEAM_FIELDING_E
    log_imp_team_fielding_dp
    flag_team_fielding_dp
    TEAM_PITCHING_BB
    TEAM_PITCHING_H  /selection=forward vif sse aic;
run;
quit;

*print values including AIC;
title "Model 3 - Print Values";
proc print data=model3;
run;
quit;
```

```
/*************************************************************************
* UNIT 01 - Predict 411 - Project 01 - Moneyball OLS Regression Baseball
* Scoring Program
* Ariel Gamino - arielgamino2016@u.northwestern.edu
*************************************************************************/

%let ME = arielgamino2016;
%let PATH = /home/&ME./my_courses/donald.wedding/c_8888/PRED411/UNIT01/HW;
%let NAME = HW;
%let LIB = &NAME..;
libname &NAME."&PATH.";

/************************** Score File ***********************/;

/** Data Preparation **/

data scorefile;
set hw.moneyball_test;

        /** Impute Missing Values **/
  imp_team_baserun_sb = team_baserun_sb;
  flag_team_baserun_sb = 0;
  imp_team_baserun_cs = team_baserun_cs;
  flag_team_baserun_cs = 0;
  imp_team_fielding_dp = team_fielding_dp;
  flag_team_fielding_dp = 0;
  imp_team_batting_so = team_batting_so;
  flag_team_batting_so = 0;
  imp_team_pitching_so = team_pitching_so;
  flag_team_pitching_so = 0;

  if missing(team_baserun_sb) then do;
    * For missing value just use the average;
    imp_team_baserun_sb  = 125;
    flag_team_baserun_sb = 1;
  end;

  if missing(team_fielding_dp) then do;
    * For missing value just use the average;
    imp_team_fielding_dp  = 146;
    flag_team_fielding_dp = 1;
  end;

  if missing(team_batting_so) then do;
    * For missing value just use the average;
    imp_team_batting_so  = 736;
    flag_team_batting_so = 1;
```

```
  end;

  if missing(team_pitching_so) then do;
    * For missing value just use the average;
    imp_team_pitching_so  = 818;
    flag_team_pitching_so = 1;
  end;


        if missing(team_baserun_cs) then do;
          flag_team_baserun_cs = 1;
          * Poor man's decision tree for team_baserun_sb based on average value for
team_baserun_cs;
    if team_baserun_sb = 0 then imp_team_baserun_cs = 27;
    else if team_baserun_sb = 14 then imp_team_baserun_cs = 7;
    else if team_baserun_sb = 18 then imp_team_baserun_cs = 19;
          else if team_baserun_sb = 19 then imp_team_baserun_cs = 23;
          else if team_baserun_sb = 20 then imp_team_baserun_cs = 23;
          else if team_baserun_sb = 21 then imp_team_baserun_cs = 26;
          else if team_baserun_sb = 22 then imp_team_baserun_cs = 22;
          else if team_baserun_sb = 23 then imp_team_baserun_cs = 37;
          else if team_baserun_sb = 24 then imp_team_baserun_cs = 40;
          else if team_baserun_sb = 25 then imp_team_baserun_cs = 24;
          else if team_baserun_sb = 26 then imp_team_baserun_cs = 22;
          else if team_baserun_sb = 27 then imp_team_baserun_cs = 30;
          else if team_baserun_sb = 28 then imp_team_baserun_cs = 32;
          else if team_baserun_sb = 29 then imp_team_baserun_cs = 26;
          else if team_baserun_sb = 30 then imp_team_baserun_cs = 28;
          else if team_baserun_sb = 31 then imp_team_baserun_cs = 20;
          else if team_baserun_sb = 32 then imp_team_baserun_cs = 28;
          else if team_baserun_sb = 33 then imp_team_baserun_cs = 14;
          else if team_baserun_sb = 34 then imp_team_baserun_cs = 23;
          else if team_baserun_sb = 35 then imp_team_baserun_cs = 32;
          else if team_baserun_sb = 36 then imp_team_baserun_cs = 35;
          else if team_baserun_sb = 37 then imp_team_baserun_cs = 27;
          else if team_baserun_sb = 38 then imp_team_baserun_cs = 32;
          else if team_baserun_sb = 39 then imp_team_baserun_cs = 32;
          else if team_baserun_sb = 40 then imp_team_baserun_cs = 26;
          else if team_baserun_sb = 41 then imp_team_baserun_cs = 38;
          else if team_baserun_sb = 42 then imp_team_baserun_cs = 37;
          else if team_baserun_sb = 43 then imp_team_baserun_cs = 30;
          else if team_baserun_sb = 44 then imp_team_baserun_cs = 35;
          else if team_baserun_sb = 45 then imp_team_baserun_cs = 33;
          else if team_baserun_sb = 46 then imp_team_baserun_cs = 36;
          else if team_baserun_sb = 47 then imp_team_baserun_cs = 32;
          else if team_baserun_sb = 48 then imp_team_baserun_cs = 37;
          else if team_baserun_sb = 49 then imp_team_baserun_cs = 37;
```

```
else if team_baserun_sb = 50 then imp_team_baserun_cs = 38;
else if team_baserun_sb = 51 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 52 then imp_team_baserun_cs = 35;
else if team_baserun_sb = 53 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 54 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 55 then imp_team_baserun_cs = 45;
else if team_baserun_sb = 56 then imp_team_baserun_cs = 38;
else if team_baserun_sb = 57 then imp_team_baserun_cs = 39;
else if team_baserun_sb = 58 then imp_team_baserun_cs = 40;
else if team_baserun_sb = 59 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 60 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 61 then imp_team_baserun_cs = 39;
else if team_baserun_sb = 62 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 63 then imp_team_baserun_cs = 41;
else if team_baserun_sb = 64 then imp_team_baserun_cs = 42;
else if team_baserun_sb = 65 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 66 then imp_team_baserun_cs = 41;
else if team_baserun_sb = 67 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 68 then imp_team_baserun_cs = 38;
else if team_baserun_sb = 69 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 70 then imp_team_baserun_cs = 45;
else if team_baserun_sb = 71 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 72 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 73 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 74 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 75 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 76 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 77 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 78 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 79 then imp_team_baserun_cs = 48;
else if team_baserun_sb = 80 then imp_team_baserun_cs = 50;
else if team_baserun_sb = 81 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 83 then imp_team_baserun_cs = 48;
else if team_baserun_sb = 84 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 85 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 86 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 87 then imp_team_baserun_cs = 48;
else if team_baserun_sb = 88 then imp_team_baserun_cs = 40;
else if team_baserun_sb = 89 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 90 then imp_team_baserun_cs = 50;
else if team_baserun_sb = 92 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 93 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 94 then imp_team_baserun_cs = 58;
else if team_baserun_sb = 95 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 97 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 98 then imp_team_baserun_cs = 56;
else if team_baserun_sb = 100 then imp_team_baserun_cs = 52;
```

```
else if team_baserun_sb = 101 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 102 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 103 then imp_team_baserun_cs = 60;
else if team_baserun_sb = 104 then imp_team_baserun_cs = 61;
else if team_baserun_sb = 105 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 106 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 107 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 109 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 110 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 111 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 112 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 113 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 114 then imp_team_baserun_cs = 64;
else if team_baserun_sb = 115 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 116 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 117 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 118 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 119 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 120 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 121 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 122 then imp_team_baserun_cs = 70;
else if team_baserun_sb = 123 then imp_team_baserun_cs = 58;
else if team_baserun_sb = 124 then imp_team_baserun_cs = 61;
else if team_baserun_sb = 125 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 126 then imp_team_baserun_cs = 65;
else if team_baserun_sb = 127 then imp_team_baserun_cs = 68;
else if team_baserun_sb = 128 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 129 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 130 then imp_team_baserun_cs = 56;
else if team_baserun_sb = 131 then imp_team_baserun_cs = 64;
else if team_baserun_sb = 132 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 133 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 134 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 135 then imp_team_baserun_cs = 64;
else if team_baserun_sb = 136 then imp_team_baserun_cs = 56;
else if team_baserun_sb = 137 then imp_team_baserun_cs = 65;
else if team_baserun_sb = 138 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 139 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 140 then imp_team_baserun_cs = 72;
else if team_baserun_sb = 141 then imp_team_baserun_cs = 50;
else if team_baserun_sb = 142 then imp_team_baserun_cs = 66;
else if team_baserun_sb = 143 then imp_team_baserun_cs = 40;
else if team_baserun_sb = 144 then imp_team_baserun_cs = 69;
else if team_baserun_sb = 145 then imp_team_baserun_cs = 71;
else if team_baserun_sb = 146 then imp_team_baserun_cs = 76;
else if team_baserun_sb = 147 then imp_team_baserun_cs = 69;
else if team_baserun_sb = 148 then imp_team_baserun_cs = 64;
```

```
else if team_baserun_sb = 149 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 150 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 151 then imp_team_baserun_cs = 50;
else if team_baserun_sb = 152 then imp_team_baserun_cs = 70;
else if team_baserun_sb = 153 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 154 then imp_team_baserun_cs = 68;
else if team_baserun_sb = 155 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 156 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 157 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 158 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 159 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 160 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 161 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 162 then imp_team_baserun_cs = 60;
else if team_baserun_sb = 163 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 164 then imp_team_baserun_cs = 78;
else if team_baserun_sb = 165 then imp_team_baserun_cs = 69;
else if team_baserun_sb = 166 then imp_team_baserun_cs = 60;
else if team_baserun_sb = 167 then imp_team_baserun_cs = 88;
else if team_baserun_sb = 168 then imp_team_baserun_cs = 42;
else if team_baserun_sb = 169 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 170 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 171 then imp_team_baserun_cs = 69;
else if team_baserun_sb = 172 then imp_team_baserun_cs = 88;
else if team_baserun_sb = 173 then imp_team_baserun_cs = 31;
else if team_baserun_sb = 174 then imp_team_baserun_cs = 61;
else if team_baserun_sb = 175 then imp_team_baserun_cs = 48;
else if team_baserun_sb = 176 then imp_team_baserun_cs = 117;
else if team_baserun_sb = 177 then imp_team_baserun_cs = 94;
else if team_baserun_sb = 178 then imp_team_baserun_cs = 30;
else if team_baserun_sb = 179 then imp_team_baserun_cs = 75;
else if team_baserun_sb = 180 then imp_team_baserun_cs = 65;
else if team_baserun_sb = 182 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 183 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 184 then imp_team_baserun_cs = 41;
else if team_baserun_sb = 186 then imp_team_baserun_cs = 60;
else if team_baserun_sb = 187 then imp_team_baserun_cs = 79;
else if team_baserun_sb = 188 then imp_team_baserun_cs = 35;
else if team_baserun_sb = 189 then imp_team_baserun_cs = 89;
else if team_baserun_sb = 190 then imp_team_baserun_cs = 87;
else if team_baserun_sb = 191 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 192 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 193 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 194 then imp_team_baserun_cs = 74;
else if team_baserun_sb = 195 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 196 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 197 then imp_team_baserun_cs = 66;
```

```
        else if team_baserun_sb = 198 then imp_team_baserun_cs = 58;
        else if team_baserun_sb = 200 then imp_team_baserun_cs = 46;
        else if team_baserun_sb = 201 then imp_team_baserun_cs = 96;
        else if team_baserun_sb = 202 then imp_team_baserun_cs = 55;
        else if team_baserun_sb = 207 then imp_team_baserun_cs = 56;
        else if team_baserun_sb = 208 then imp_team_baserun_cs = 58;
        else if team_baserun_sb = 209 then imp_team_baserun_cs = 51;
        else if team_baserun_sb = 210 then imp_team_baserun_cs = 29;
        else if team_baserun_sb = 211 then imp_team_baserun_cs = 142;
        else if team_baserun_sb = 212 then imp_team_baserun_cs = 168;
        else if team_baserun_sb = 214 then imp_team_baserun_cs = 72;
        else if team_baserun_sb = 220 then imp_team_baserun_cs = 71;
        else if team_baserun_sb = 221 then imp_team_baserun_cs = 61;
        else if team_baserun_sb = 222 then imp_team_baserun_cs = 51;
        else if team_baserun_sb = 223 then imp_team_baserun_cs = 82;
        else if team_baserun_sb = 231 then imp_team_baserun_cs = 86;
        else if team_baserun_sb = 232 then imp_team_baserun_cs = 87;
        else if team_baserun_sb = 234 then imp_team_baserun_cs = 64;
        else if team_baserun_sb = 235 then imp_team_baserun_cs = 99;
        else if team_baserun_sb = 237 then imp_team_baserun_cs = 82;
        else if team_baserun_sb = 239 then imp_team_baserun_cs = 37;
        else if team_baserun_sb = 245 then imp_team_baserun_cs = 97;
        else if team_baserun_sb = 246 then imp_team_baserun_cs = 100;
        else if team_baserun_sb = 247 then imp_team_baserun_cs = 200;
        else if team_baserun_sb = 248 then imp_team_baserun_cs = 36;
        else if team_baserun_sb = 264 then imp_team_baserun_cs = 140;
        else if team_baserun_sb = 314 then imp_team_baserun_cs = 96;
  else team_baserun_cs = 53;
end;

*Catch all;
if missing(imp_team_baserun_cs) then do;
       imp_team_baserun_cs = 53;
       flag_team_baserun_cs = 1;
end;
 /** Transform **/
        cap_team_batting_2b = TEAM_BATTING_2B;
        flag_cap_team_batting_2b = 0;
        cap_team_batting_3b = TEAM_BATTING_3B;
        flag_cap_team_batting_3b = 0;
        cap_team_pitching_h = TEAM_PITCHING_H;
        flag_cap_team_pitching_h = 0;
        cap_team_pitching_bb = TEAM_PITCHING_BB;
        flag_cap_team_pitching_bb = 0;
        cap_team_fielding_e = TEAM_FIELDING_E;
        flag_cap_team_fielding_e = 0;
        cap_imp_team_baserun_sb = imp_team_baserun_sb;
```

```
        flag_cap_imp_team_baserun_sb = 0;
        cap_imp_team_baserun_cs = imp_team_baserun_cs;
        flag_cap_imp_team_baserun_cs = 0;
        cap_imp_team_pitching_so = imp_team_pitching_so;
        flag_camp_imp_team_pitching_so = 0;

        *Cap TEAM_BATTING_2B to 5% > 167, 352 < 99%;
        if TEAM_BATTING_2B < 167.00 then do;
    cap_team_batting_2b = 167.00;
    flag_cap_team_batting_2b = 1;
end;
if TEAM_BATTING_2B > 352.00 then do;
    cap_team_batting_2b = 352.00;
    flag_cap_team_batting_2b = 1;
end;

*Cap TEAM_BATTING_3B to 5% > 23, 134 < 99%;
if TEAM_BATTING_3B < 23.00 then do;
    cap_team_batting_3b = 23.00;
    flag_cap_team_batting_3b = 1;
end;
if TEAM_BATTING_3B > 134.00 then do;
    cap_team_batting_3b = 134.00;
    flag_cap_team_batting_3b = 1;
end;

*Cap TEAM_PITCHING_H to 5% > 1316, 2563 < 95%;
if TEAM_PITCHING_H < 1316.00 then do;
    cap_team_pitching_h = 1316.00;
    flag_cap_team_pitching_h = 1;
end;
if TEAM_PITCHING_H > 2563.00 then do;
    cap_team_pitching_h = 2563.00;
    flag_cap_team_pitching_h = 1;
end;

*Cap TEAM_PITCHING_BB to 5% > 377, 924 < 99%;
if TEAM_PITCHING_BB < 377.00 then do;
    cap_team_pitching_bb = 377.00;
    flag_cap_team_pitching_bb = 1;
end;
if TEAM_PITCHING_BB > 924.00 then do;
    cap_team_pitching_bb = 924.00;
    flag_cap_team_pitching_bb = 1;
end;

*Cap TEAM_FIELDING_E to 5% > 100, 716 < 95%;
```

```
if TEAM_FIELDING_E < 100.00 then do;
    cap_team_fielding_e = 100.00;
    flag_cap_team_fielding_e = 1;
end;
if TEAM_FIELDING_E > 716.00 then do;
    cap_team_fielding_e = 716.00;
    flag_cap_team_fielding_e = 1;
end;

*Cap imp_team_baserun_sb to 5% > 36, 298 < 95%;
if imp_team_baserun_sb < 36.00 then do;
    cap_imp_team_baserun_sb = 36.00;
    flag_cap_imp_team_baserun_sb = 1;
end;
if imp_team_baserun_sb > 298.00 then do;
    cap_imp_team_baserun_sb = 298.00;
    flag_cap_imp_team_baserun_sb = 1;
end;

*Cap imp_team_baserun_cs to 5% > 25, 91 < 95%;
if imp_team_baserun_cs < 25.00 then do;
    cap_imp_team_baserun_cs = 25.00;
    flag_cap_imp_team_baserun_cs = 1;
end;
if imp_team_baserun_cs > 91.00 then do;
    cap_imp_team_baserun_cs = 91.00;
    flag_cap_imp_team_baserun_cs = 1;
end;

*Cap imp_team_pitching_so to 5% > 423, 1169 < 95%;
if imp_team_pitching_so < 423.00 then do;
    cap_imp_team_pitching_so = 423.00;
    flag_camp_imp_team_pitching_so = 1;
end;
if imp_team_pitching_so > 1169.00 then do;
    cap_imp_team_pitching_so = 1169.00;
    flag_camp_imp_team_pitching_so = 1;
end;

      log_team_batting_h = log(TEAM_BATTING_H+1);
log_imp_team_fielding_dp = log(imp_team_fielding_dp+1);

     p_target_wins = 78.2117 +
              0.048537    * TEAM_BATTING_H +
         0.071427    * TEAM_BATTING_HR +
         0.028545    * TEAM_BATTING_BB +
         -0.011644   * imp_team_batting_so +
```

```
            7.90144     * flag_team_batting_so +
            -14.6576    * log_imp_team_fielding_dp +
            4.84582     * flag_team_fielding_dp +
            -0.035594   * TEAM_BATTING_2B +
            0.054563    * TEAM_BATTING_3B +
            0.04922     * imp_team_baserun_sb +
            32.9883     * flag_team_baserun_sb +
            -0.056792   * TEAM_FIELDING_E +
            -0.002994668 * TEAM_PITCHING_BB +
            0.00201146   * TEAM_PITCHING_H;

        keep INDEX;
        keep P_TARGET_WINS;

run;
quit;

title "Score File";
proc print data=scorefile;
run;

/*************************************************************************
* UNIT 01 - Predict 411 - Project 01 - Moneyball OLS Regression Baseball
* Scoring Program - Score Moneball_test
* Ariel Gamino - arielgamino2016@u.northwestern.edu
*************************************************************************/

%let ME = arielgamino2016;
%let PATH = /home/&ME./my_courses/donald.wedding/c_8888/PRED411/UNIT01/HW;
%let NAME = HW;
%let LIB = &NAME..;
libname &NAME."&PATH.";

/*************************** Score File ************************/;

/** Data Preparation **/

data scorefile;
set hw.moneyball_test;

        /** Impute Missing Values **/
   imp_team_baserun_sb = team_baserun_sb;
   flag_team_baserun_sb = 0;
   imp_team_baserun_cs = team_baserun_cs;
   flag_team_baserun_cs = 0;
   imp_team_fielding_dp = team_fielding_dp;
   flag_team_fielding_dp = 0;
```

```
  imp_team_batting_so = team_batting_so;
  flag_team_batting_so = 0;
  imp_team_pitching_so = team_pitching_so;
  flag_team_pitching_so = 0;

  if missing(team_baserun_sb) then do;
    * For missing value just use the average;
    imp_team_baserun_sb  = 125;
    flag_team_baserun_sb = 1;
  end;

  if missing(team_fielding_dp) then do;
    * For missing value just use the average;
    imp_team_fielding_dp  = 146;
    flag_team_fielding_dp = 1;
  end;

  if missing(team_batting_so) then do;
    * For missing value just use the average;
    imp_team_batting_so  = 736;
    flag_team_batting_so = 1;
  end;

  if missing(team_pitching_so) then do;
    * For missing value just use the average;
    imp_team_pitching_so  = 818;
    flag_team_pitching_so = 1;
  end;


        if missing(team_baserun_cs) then do;
          flag_team_baserun_cs = 1;
          * Poor man's decision tree for team_baserun_sb based on average value for
team_baserun_cs;
    if team_baserun_sb = 0 then imp_team_baserun_cs = 27;
    else if team_baserun_sb = 14 then imp_team_baserun_cs = 7;
    else if team_baserun_sb = 18 then imp_team_baserun_cs = 19;
          else if team_baserun_sb = 19 then imp_team_baserun_cs = 23;
          else if team_baserun_sb = 20 then imp_team_baserun_cs = 23;
          else if team_baserun_sb = 21 then imp_team_baserun_cs = 26;
          else if team_baserun_sb = 22 then imp_team_baserun_cs = 22;
          else if team_baserun_sb = 23 then imp_team_baserun_cs = 37;
          else if team_baserun_sb = 24 then imp_team_baserun_cs = 40;
          else if team_baserun_sb = 25 then imp_team_baserun_cs = 24;
          else if team_baserun_sb = 26 then imp_team_baserun_cs = 22;
          else if team_baserun_sb = 27 then imp_team_baserun_cs = 30;
          else if team_baserun_sb = 28 then imp_team_baserun_cs = 32;
```

```
else if team_baserun_sb = 29 then imp_team_baserun_cs = 26;
else if team_baserun_sb = 30 then imp_team_baserun_cs = 28;
else if team_baserun_sb = 31 then imp_team_baserun_cs = 20;
else if team_baserun_sb = 32 then imp_team_baserun_cs = 28;
else if team_baserun_sb = 33 then imp_team_baserun_cs = 14;
else if team_baserun_sb = 34 then imp_team_baserun_cs = 23;
else if team_baserun_sb = 35 then imp_team_baserun_cs = 32;
else if team_baserun_sb = 36 then imp_team_baserun_cs = 35;
else if team_baserun_sb = 37 then imp_team_baserun_cs = 27;
else if team_baserun_sb = 38 then imp_team_baserun_cs = 32;
else if team_baserun_sb = 39 then imp_team_baserun_cs = 32;
else if team_baserun_sb = 40 then imp_team_baserun_cs = 26;
else if team_baserun_sb = 41 then imp_team_baserun_cs = 38;
else if team_baserun_sb = 42 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 43 then imp_team_baserun_cs = 30;
else if team_baserun_sb = 44 then imp_team_baserun_cs = 35;
else if team_baserun_sb = 45 then imp_team_baserun_cs = 33;
else if team_baserun_sb = 46 then imp_team_baserun_cs = 36;
else if team_baserun_sb = 47 then imp_team_baserun_cs = 32;
else if team_baserun_sb = 48 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 49 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 50 then imp_team_baserun_cs = 38;
else if team_baserun_sb = 51 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 52 then imp_team_baserun_cs = 35;
else if team_baserun_sb = 53 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 54 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 55 then imp_team_baserun_cs = 45;
else if team_baserun_sb = 56 then imp_team_baserun_cs = 38;
else if team_baserun_sb = 57 then imp_team_baserun_cs = 39;
else if team_baserun_sb = 58 then imp_team_baserun_cs = 40;
else if team_baserun_sb = 59 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 60 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 61 then imp_team_baserun_cs = 39;
else if team_baserun_sb = 62 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 63 then imp_team_baserun_cs = 41;
else if team_baserun_sb = 64 then imp_team_baserun_cs = 42;
else if team_baserun_sb = 65 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 66 then imp_team_baserun_cs = 41;
else if team_baserun_sb = 67 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 68 then imp_team_baserun_cs = 38;
else if team_baserun_sb = 69 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 70 then imp_team_baserun_cs = 45;
else if team_baserun_sb = 71 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 72 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 73 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 74 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 75 then imp_team_baserun_cs = 49;
```

```
else if team_baserun_sb = 76 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 77 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 78 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 79 then imp_team_baserun_cs = 48;
else if team_baserun_sb = 80 then imp_team_baserun_cs = 50;
else if team_baserun_sb = 81 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 83 then imp_team_baserun_cs = 48;
else if team_baserun_sb = 84 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 85 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 86 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 87 then imp_team_baserun_cs = 48;
else if team_baserun_sb = 88 then imp_team_baserun_cs = 40;
else if team_baserun_sb = 89 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 90 then imp_team_baserun_cs = 50;
else if team_baserun_sb = 92 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 93 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 94 then imp_team_baserun_cs = 58;
else if team_baserun_sb = 95 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 97 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 98 then imp_team_baserun_cs = 56;
else if team_baserun_sb = 100 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 101 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 102 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 103 then imp_team_baserun_cs = 60;
else if team_baserun_sb = 104 then imp_team_baserun_cs = 61;
else if team_baserun_sb = 105 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 106 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 107 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 109 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 110 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 111 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 112 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 113 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 114 then imp_team_baserun_cs = 64;
else if team_baserun_sb = 115 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 116 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 117 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 118 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 119 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 120 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 121 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 122 then imp_team_baserun_cs = 70;
else if team_baserun_sb = 123 then imp_team_baserun_cs = 58;
else if team_baserun_sb = 124 then imp_team_baserun_cs = 61;
else if team_baserun_sb = 125 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 126 then imp_team_baserun_cs = 65;
else if team_baserun_sb = 127 then imp_team_baserun_cs = 68;
```

```
else if team_baserun_sb = 128 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 129 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 130 then imp_team_baserun_cs = 56;
else if team_baserun_sb = 131 then imp_team_baserun_cs = 64;
else if team_baserun_sb = 132 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 133 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 134 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 135 then imp_team_baserun_cs = 64;
else if team_baserun_sb = 136 then imp_team_baserun_cs = 56;
else if team_baserun_sb = 137 then imp_team_baserun_cs = 65;
else if team_baserun_sb = 138 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 139 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 140 then imp_team_baserun_cs = 72;
else if team_baserun_sb = 141 then imp_team_baserun_cs = 50;
else if team_baserun_sb = 142 then imp_team_baserun_cs = 66;
else if team_baserun_sb = 143 then imp_team_baserun_cs = 40;
else if team_baserun_sb = 144 then imp_team_baserun_cs = 69;
else if team_baserun_sb = 145 then imp_team_baserun_cs = 71;
else if team_baserun_sb = 146 then imp_team_baserun_cs = 76;
else if team_baserun_sb = 147 then imp_team_baserun_cs = 69;
else if team_baserun_sb = 148 then imp_team_baserun_cs = 64;
else if team_baserun_sb = 149 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 150 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 151 then imp_team_baserun_cs = 50;
else if team_baserun_sb = 152 then imp_team_baserun_cs = 70;
else if team_baserun_sb = 153 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 154 then imp_team_baserun_cs = 68;
else if team_baserun_sb = 155 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 156 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 157 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 158 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 159 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 160 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 161 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 162 then imp_team_baserun_cs = 60;
else if team_baserun_sb = 163 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 164 then imp_team_baserun_cs = 78;
else if team_baserun_sb = 165 then imp_team_baserun_cs = 69;
else if team_baserun_sb = 166 then imp_team_baserun_cs = 60;
else if team_baserun_sb = 167 then imp_team_baserun_cs = 88;
else if team_baserun_sb = 168 then imp_team_baserun_cs = 42;
else if team_baserun_sb = 169 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 170 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 171 then imp_team_baserun_cs = 69;
else if team_baserun_sb = 172 then imp_team_baserun_cs = 88;
else if team_baserun_sb = 173 then imp_team_baserun_cs = 31;
else if team_baserun_sb = 174 then imp_team_baserun_cs = 61;
```

```
else if team_baserun_sb = 175 then imp_team_baserun_cs = 48;
else if team_baserun_sb = 176 then imp_team_baserun_cs = 117;
else if team_baserun_sb = 177 then imp_team_baserun_cs = 94;
else if team_baserun_sb = 178 then imp_team_baserun_cs = 30;
else if team_baserun_sb = 179 then imp_team_baserun_cs = 75;
else if team_baserun_sb = 180 then imp_team_baserun_cs = 65;
else if team_baserun_sb = 182 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 183 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 184 then imp_team_baserun_cs = 41;
else if team_baserun_sb = 186 then imp_team_baserun_cs = 60;
else if team_baserun_sb = 187 then imp_team_baserun_cs = 79;
else if team_baserun_sb = 188 then imp_team_baserun_cs = 35;
else if team_baserun_sb = 189 then imp_team_baserun_cs = 89;
else if team_baserun_sb = 190 then imp_team_baserun_cs = 87;
else if team_baserun_sb = 191 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 192 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 193 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 194 then imp_team_baserun_cs = 74;
else if team_baserun_sb = 195 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 196 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 197 then imp_team_baserun_cs = 66;
else if team_baserun_sb = 198 then imp_team_baserun_cs = 58;
else if team_baserun_sb = 200 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 201 then imp_team_baserun_cs = 96;
else if team_baserun_sb = 202 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 207 then imp_team_baserun_cs = 56;
else if team_baserun_sb = 208 then imp_team_baserun_cs = 58;
else if team_baserun_sb = 209 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 210 then imp_team_baserun_cs = 29;
else if team_baserun_sb = 211 then imp_team_baserun_cs = 142;
else if team_baserun_sb = 212 then imp_team_baserun_cs = 168;
else if team_baserun_sb = 214 then imp_team_baserun_cs = 72;
else if team_baserun_sb = 220 then imp_team_baserun_cs = 71;
else if team_baserun_sb = 221 then imp_team_baserun_cs = 61;
else if team_baserun_sb = 222 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 223 then imp_team_baserun_cs = 82;
else if team_baserun_sb = 231 then imp_team_baserun_cs = 86;
else if team_baserun_sb = 232 then imp_team_baserun_cs = 87;
else if team_baserun_sb = 234 then imp_team_baserun_cs = 64;
else if team_baserun_sb = 235 then imp_team_baserun_cs = 99;
else if team_baserun_sb = 237 then imp_team_baserun_cs = 82;
else if team_baserun_sb = 239 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 245 then imp_team_baserun_cs = 97;
else if team_baserun_sb = 246 then imp_team_baserun_cs = 100;
else if team_baserun_sb = 247 then imp_team_baserun_cs = 200;
else if team_baserun_sb = 248 then imp_team_baserun_cs = 36;
else if team_baserun_sb = 264 then imp_team_baserun_cs = 140;
```

```
        else if team_baserun_sb = 314 then imp_team_baserun_cs = 96;
  else team_baserun_cs = 53;
end;

*Catch all;
if missing(imp_team_baserun_cs) then do;
       imp_team_baserun_cs = 53;
       flag_team_baserun_cs = 1;
end;
 /** Transform **/
         cap_team_batting_2b = TEAM_BATTING_2B;
         flag_cap_team_batting_2b = 0;
         cap_team_batting_3b = TEAM_BATTING_3B;
         flag_cap_team_batting_3b = 0;
         cap_team_pitching_h = TEAM_PITCHING_H;
         flag_cap_team_pitching_h = 0;
         cap_team_pitching_bb = TEAM_PITCHING_BB;
         flag_cap_team_pitching_bb = 0;
         cap_team_fielding_e = TEAM_FIELDING_E;
         flag_cap_team_fielding_e = 0;
         cap_imp_team_baserun_sb = imp_team_baserun_sb;
         flag_cap_imp_team_baserun_sb = 0;
         cap_imp_team_baserun_cs = imp_team_baserun_cs;
         flag_cap_imp_team_baserun_cs = 0;
         cap_imp_team_pitching_so = imp_team_pitching_so;
         flag_camp_imp_team_pitching_so = 0;

         *Cap TEAM_BATTING_2B to 5% > 167, 352 < 99%;
         if TEAM_BATTING_2B < 167.00 then do;
     cap_team_batting_2b = 167.00;
     flag_cap_team_batting_2b = 1;
 end;
 if TEAM_BATTING_2B > 352.00 then do;
     cap_team_batting_2b = 352.00;
     flag_cap_team_batting_2b = 1;
 end;

 *Cap TEAM_BATTING_3B to 5% > 23, 134 < 99%;
 if TEAM_BATTING_3B < 23.00 then do;
     cap_team_batting_3b = 23.00;
     flag_cap_team_batting_3b = 1;
 end;
 if TEAM_BATTING_3B > 134.00 then do;
     cap_team_batting_3b = 134.00;
     flag_cap_team_batting_3b = 1;
 end;
```

```
*Cap TEAM_PITCHING_H to 5% > 1316, 2563 < 95%;
if TEAM_PITCHING_H < 1316.00 then do;
    cap_team_pitching_h = 1316.00;
    flag_cap_team_pitching_h = 1;
end;
if TEAM_PITCHING_H > 2563.00 then do;
    cap_team_pitching_h = 2563.00;
    flag_cap_team_pitching_h = 1;
end;

*Cap TEAM_PITCHING_BB to 5% > 377, 924 < 99%;
if TEAM_PITCHING_BB < 377.00 then do;
    cap_team_pitching_bb = 377.00;
    flag_cap_team_pitching_bb = 1;
end;
if TEAM_PITCHING_BB > 924.00 then do;
    cap_team_pitching_bb = 924.00;
    flag_cap_team_pitching_bb = 1;
end;

*Cap TEAM_FIELDING_E to 5% > 100, 716 < 95%;
if TEAM_FIELDING_E < 100.00 then do;
    cap_team_fielding_e = 100.00;
    flag_cap_team_fielding_e = 1;
end;
if TEAM_FIELDING_E > 716.00 then do;
    cap_team_fielding_e = 716.00;
    flag_cap_team_fielding_e = 1;
end;

*Cap imp_team_baserun_sb to 5% > 36, 298 < 95%;
if imp_team_baserun_sb < 36.00 then do;
    cap_imp_team_baserun_sb = 36.00;
    flag_cap_imp_team_baserun_sb = 1;
end;
if imp_team_baserun_sb > 298.00 then do;
    cap_imp_team_baserun_sb = 298.00;
    flag_cap_imp_team_baserun_sb = 1;
end;

*Cap imp_team_baserun_cs to 5% > 25, 91 < 95%;
if imp_team_baserun_cs < 25.00 then do;
    cap_imp_team_baserun_cs = 25.00;
    flag_cap_imp_team_baserun_cs = 1;
end;
if imp_team_baserun_cs > 91.00 then do;
    cap_imp_team_baserun_cs = 91.00;
```

```
    flag_cap_imp_team_baserun_cs = 1;
  end;

  *Cap imp_team_pitching_so to 5% > 423, 1169 < 95%;
  if imp_team_pitching_so < 423.00 then do;
    cap_imp_team_pitching_so = 423.00;
    flag_camp_imp_team_pitching_so = 1;
  end;
  if imp_team_pitching_so > 1169.00 then do;
    cap_imp_team_pitching_so = 1169.00;
    flag_camp_imp_team_pitching_so = 1;
  end;

      log_team_batting_h = log(TEAM_BATTING_H+1);
  log_imp_team_fielding_dp = log(imp_team_fielding_dp+1);

      p_target_wins = 78.2117 +
              0.048537    * TEAM_BATTING_H +
          0.071427    * TEAM_BATTING_HR +
          0.028545    * TEAM_BATTING_BB +
          -0.011644   * imp_team_batting_so +
          7.90144     * flag_team_batting_so +
          -14.6576    * log_imp_team_fielding_dp +
          4.84582     * flag_team_fielding_dp +
          -0.035594   * TEAM_BATTING_2B +
          0.054563    * TEAM_BATTING_3B +
          0.04922     * imp_team_baserun_sb +
          32.9883     * flag_team_baserun_sb +
          -0.056792   * TEAM_FIELDING_E +
          -0.002994668 * TEAM_PITCHING_BB +
          0.00201146  * TEAM_PITCHING_H;

   /** Cap Target Wins to Min and Mac values**/
  if p_target_wins < 38 then p_target_wins=38;
  if p_target_wins > 114 then p_target_wins=114;

      keep INDEX;
      keep P_TARGET_WINS;

run;
quit;

title "Score File";
proc print data=scorefile;
run;

/** Save Score File **/
```

```
libname outlib "/home/arielgamino2016/411/Unit01/Project01";
data outlib.Score_File_Ariel_Gamino;
set scorefile;
run;


/*************************************************************
* UNIT 01 - Predict 411 - Project 01 - Moneyball OLS Regression Baseball
* Data Preparation and Regression with GLM and GENMOD
* BINGO BONUS
* Ariel Gamino - arielgamino2016@u.northwestern.edu
*************************************************************************/

%let ME = arielgamino2016;
%let PATH = /home/&ME./my_courses/donald.wedding/c_8888/PRED411/UNIT01/HW;
%let NAME = HW;
%let LIB = &NAME..;

libname &NAME."&PATH.";

%let INFILE = HW.MONEYBALL;

/** Data Preparation **/
data tempfile_for_regression;
set &INFILE.;
   imp_team_baserun_sb = team_baserun_sb;
   flag_team_baserun_sb = 0;
   imp_team_baserun_cs = team_baserun_cs;
   flag_team_baserun_cs = 0;
   imp_team_fielding_dp = team_fielding_dp;
   flag_team_fielding_dp = 0;
   imp_team_batting_so = team_batting_so;
   flag_team_batting_so = 0;
   imp_team_pitching_so = team_pitching_so;
   flag_team_pitching_so = 0;

   if missing(team_baserun_sb) then do;
     * For missing value just use the average;
     imp_team_baserun_sb  = 125;
     flag_team_baserun_sb = 1;
   end;

   if missing(team_fielding_dp) then do;
     * For missing value just use the average;
     imp_team_fielding_dp  = 146;
     flag_team_fielding_dp = 1;
   end;
```

```
  if missing(team_batting_so) then do;
    * For missing value just use the average;
    imp_team_batting_so  = 736;
    flag_team_batting_so = 1;
  end;

  if missing(team_pitching_so) then do;
    * For missing value just use the average;
    imp_team_pitching_so  = 818;
    flag_team_pitching_so = 1;
  end;


        if missing(team_baserun_cs) then do;
          flag_team_baserun_cs = 1;
          * Poor man's decision tree for team_baserun_sb based on average value for
team_baserun_cs;
    if team_baserun_sb = 0 then imp_team_baserun_cs = 27;
    else if team_baserun_sb = 14 then imp_team_baserun_cs = 7;
    else if team_baserun_sb = 18 then imp_team_baserun_cs = 19;
          else if team_baserun_sb = 19 then imp_team_baserun_cs = 23;
          else if team_baserun_sb = 20 then imp_team_baserun_cs = 23;
          else if team_baserun_sb = 21 then imp_team_baserun_cs = 26;
          else if team_baserun_sb = 22 then imp_team_baserun_cs = 22;
          else if team_baserun_sb = 23 then imp_team_baserun_cs = 37;
          else if team_baserun_sb = 24 then imp_team_baserun_cs = 40;
          else if team_baserun_sb = 25 then imp_team_baserun_cs = 24;
          else if team_baserun_sb = 26 then imp_team_baserun_cs = 22;
          else if team_baserun_sb = 27 then imp_team_baserun_cs = 30;
          else if team_baserun_sb = 28 then imp_team_baserun_cs = 32;
          else if team_baserun_sb = 29 then imp_team_baserun_cs = 26;
          else if team_baserun_sb = 30 then imp_team_baserun_cs = 28;
          else if team_baserun_sb = 31 then imp_team_baserun_cs = 20;
          else if team_baserun_sb = 32 then imp_team_baserun_cs = 28;
          else if team_baserun_sb = 33 then imp_team_baserun_cs = 14;
          else if team_baserun_sb = 34 then imp_team_baserun_cs = 23;
          else if team_baserun_sb = 35 then imp_team_baserun_cs = 32;
          else if team_baserun_sb = 36 then imp_team_baserun_cs = 35;
          else if team_baserun_sb = 37 then imp_team_baserun_cs = 27;
          else if team_baserun_sb = 38 then imp_team_baserun_cs = 32;
          else if team_baserun_sb = 39 then imp_team_baserun_cs = 32;
          else if team_baserun_sb = 40 then imp_team_baserun_cs = 26;
          else if team_baserun_sb = 41 then imp_team_baserun_cs = 38;
          else if team_baserun_sb = 42 then imp_team_baserun_cs = 37;
          else if team_baserun_sb = 43 then imp_team_baserun_cs = 30;
          else if team_baserun_sb = 44 then imp_team_baserun_cs = 35;
```

```
else if team_baserun_sb = 45 then imp_team_baserun_cs = 33;
else if team_baserun_sb = 46 then imp_team_baserun_cs = 36;
else if team_baserun_sb = 47 then imp_team_baserun_cs = 32;
else if team_baserun_sb = 48 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 49 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 50 then imp_team_baserun_cs = 38;
else if team_baserun_sb = 51 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 52 then imp_team_baserun_cs = 35;
else if team_baserun_sb = 53 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 54 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 55 then imp_team_baserun_cs = 45;
else if team_baserun_sb = 56 then imp_team_baserun_cs = 38;
else if team_baserun_sb = 57 then imp_team_baserun_cs = 39;
else if team_baserun_sb = 58 then imp_team_baserun_cs = 40;
else if team_baserun_sb = 59 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 60 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 61 then imp_team_baserun_cs = 39;
else if team_baserun_sb = 62 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 63 then imp_team_baserun_cs = 41;
else if team_baserun_sb = 64 then imp_team_baserun_cs = 42;
else if team_baserun_sb = 65 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 66 then imp_team_baserun_cs = 41;
else if team_baserun_sb = 67 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 68 then imp_team_baserun_cs = 38;
else if team_baserun_sb = 69 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 70 then imp_team_baserun_cs = 45;
else if team_baserun_sb = 71 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 72 then imp_team_baserun_cs = 37;
else if team_baserun_sb = 73 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 74 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 75 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 76 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 77 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 78 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 79 then imp_team_baserun_cs = 48;
else if team_baserun_sb = 80 then imp_team_baserun_cs = 50;
else if team_baserun_sb = 81 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 83 then imp_team_baserun_cs = 48;
else if team_baserun_sb = 84 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 85 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 86 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 87 then imp_team_baserun_cs = 48;
else if team_baserun_sb = 88 then imp_team_baserun_cs = 40;
else if team_baserun_sb = 89 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 90 then imp_team_baserun_cs = 50;
else if team_baserun_sb = 92 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 93 then imp_team_baserun_cs = 54;
```

```
else if team_baserun_sb = 94 then imp_team_baserun_cs = 58;
else if team_baserun_sb = 95 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 97 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 98 then imp_team_baserun_cs = 56;
else if team_baserun_sb = 100 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 101 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 102 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 103 then imp_team_baserun_cs = 60;
else if team_baserun_sb = 104 then imp_team_baserun_cs = 61;
else if team_baserun_sb = 105 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 106 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 107 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 109 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 110 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 111 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 112 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 113 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 114 then imp_team_baserun_cs = 64;
else if team_baserun_sb = 115 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 116 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 117 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 118 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 119 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 120 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 121 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 122 then imp_team_baserun_cs = 70;
else if team_baserun_sb = 123 then imp_team_baserun_cs = 58;
else if team_baserun_sb = 124 then imp_team_baserun_cs = 61;
else if team_baserun_sb = 125 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 126 then imp_team_baserun_cs = 65;
else if team_baserun_sb = 127 then imp_team_baserun_cs = 68;
else if team_baserun_sb = 128 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 129 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 130 then imp_team_baserun_cs = 56;
else if team_baserun_sb = 131 then imp_team_baserun_cs = 64;
else if team_baserun_sb = 132 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 133 then imp_team_baserun_cs = 54;
else if team_baserun_sb = 134 then imp_team_baserun_cs = 57;
else if team_baserun_sb = 135 then imp_team_baserun_cs = 64;
else if team_baserun_sb = 136 then imp_team_baserun_cs = 56;
else if team_baserun_sb = 137 then imp_team_baserun_cs = 65;
else if team_baserun_sb = 138 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 139 then imp_team_baserun_cs = 47;
else if team_baserun_sb = 140 then imp_team_baserun_cs = 72;
else if team_baserun_sb = 141 then imp_team_baserun_cs = 50;
else if team_baserun_sb = 142 then imp_team_baserun_cs = 66;
else if team_baserun_sb = 143 then imp_team_baserun_cs = 40;
```

```
else if team_baserun_sb = 144 then imp_team_baserun_cs = 69;
else if team_baserun_sb = 145 then imp_team_baserun_cs = 71;
else if team_baserun_sb = 146 then imp_team_baserun_cs = 76;
else if team_baserun_sb = 147 then imp_team_baserun_cs = 69;
else if team_baserun_sb = 148 then imp_team_baserun_cs = 64;
else if team_baserun_sb = 149 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 150 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 151 then imp_team_baserun_cs = 50;
else if team_baserun_sb = 152 then imp_team_baserun_cs = 70;
else if team_baserun_sb = 153 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 154 then imp_team_baserun_cs = 68;
else if team_baserun_sb = 155 then imp_team_baserun_cs = 49;
else if team_baserun_sb = 156 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 157 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 158 then imp_team_baserun_cs = 51;
else if team_baserun_sb = 159 then imp_team_baserun_cs = 62;
else if team_baserun_sb = 160 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 161 then imp_team_baserun_cs = 52;
else if team_baserun_sb = 162 then imp_team_baserun_cs = 60;
else if team_baserun_sb = 163 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 164 then imp_team_baserun_cs = 78;
else if team_baserun_sb = 165 then imp_team_baserun_cs = 69;
else if team_baserun_sb = 166 then imp_team_baserun_cs = 60;
else if team_baserun_sb = 167 then imp_team_baserun_cs = 88;
else if team_baserun_sb = 168 then imp_team_baserun_cs = 42;
else if team_baserun_sb = 169 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 170 then imp_team_baserun_cs = 55;
else if team_baserun_sb = 171 then imp_team_baserun_cs = 69;
else if team_baserun_sb = 172 then imp_team_baserun_cs = 88;
else if team_baserun_sb = 173 then imp_team_baserun_cs = 31;
else if team_baserun_sb = 174 then imp_team_baserun_cs = 61;
else if team_baserun_sb = 175 then imp_team_baserun_cs = 48;
else if team_baserun_sb = 176 then imp_team_baserun_cs = 117;
else if team_baserun_sb = 177 then imp_team_baserun_cs = 94;
else if team_baserun_sb = 178 then imp_team_baserun_cs = 30;
else if team_baserun_sb = 179 then imp_team_baserun_cs = 75;
else if team_baserun_sb = 180 then imp_team_baserun_cs = 65;
else if team_baserun_sb = 182 then imp_team_baserun_cs = 46;
else if team_baserun_sb = 183 then imp_team_baserun_cs = 44;
else if team_baserun_sb = 184 then imp_team_baserun_cs = 41;
else if team_baserun_sb = 186 then imp_team_baserun_cs = 60;
else if team_baserun_sb = 187 then imp_team_baserun_cs = 79;
else if team_baserun_sb = 188 then imp_team_baserun_cs = 35;
else if team_baserun_sb = 189 then imp_team_baserun_cs = 89;
else if team_baserun_sb = 190 then imp_team_baserun_cs = 87;
else if team_baserun_sb = 191 then imp_team_baserun_cs = 59;
else if team_baserun_sb = 192 then imp_team_baserun_cs = 49;
```

```
        else if team_baserun_sb = 193 then imp_team_baserun_cs = 47;
        else if team_baserun_sb = 194 then imp_team_baserun_cs = 74;
        else if team_baserun_sb = 195 then imp_team_baserun_cs = 51;
        else if team_baserun_sb = 196 then imp_team_baserun_cs = 59;
        else if team_baserun_sb = 197 then imp_team_baserun_cs = 66;
        else if team_baserun_sb = 198 then imp_team_baserun_cs = 58;
        else if team_baserun_sb = 200 then imp_team_baserun_cs = 46;
        else if team_baserun_sb = 201 then imp_team_baserun_cs = 96;
        else if team_baserun_sb = 202 then imp_team_baserun_cs = 55;
        else if team_baserun_sb = 207 then imp_team_baserun_cs = 56;
        else if team_baserun_sb = 208 then imp_team_baserun_cs = 58;
        else if team_baserun_sb = 209 then imp_team_baserun_cs = 51;
        else if team_baserun_sb = 210 then imp_team_baserun_cs = 29;
        else if team_baserun_sb = 211 then imp_team_baserun_cs = 142;
        else if team_baserun_sb = 212 then imp_team_baserun_cs = 168;
        else if team_baserun_sb = 214 then imp_team_baserun_cs = 72;
        else if team_baserun_sb = 220 then imp_team_baserun_cs = 71;
        else if team_baserun_sb = 221 then imp_team_baserun_cs = 61;
        else if team_baserun_sb = 222 then imp_team_baserun_cs = 51;
        else if team_baserun_sb = 223 then imp_team_baserun_cs = 82;
        else if team_baserun_sb = 231 then imp_team_baserun_cs = 86;
        else if team_baserun_sb = 232 then imp_team_baserun_cs = 87;
        else if team_baserun_sb = 234 then imp_team_baserun_cs = 64;
        else if team_baserun_sb = 235 then imp_team_baserun_cs = 99;
        else if team_baserun_sb = 237 then imp_team_baserun_cs = 82;
        else if team_baserun_sb = 239 then imp_team_baserun_cs = 37;
        else if team_baserun_sb = 245 then imp_team_baserun_cs = 97;
        else if team_baserun_sb = 246 then imp_team_baserun_cs = 100;
        else if team_baserun_sb = 247 then imp_team_baserun_cs = 200;
        else if team_baserun_sb = 248 then imp_team_baserun_cs = 36;
        else if team_baserun_sb = 264 then imp_team_baserun_cs = 140;
        else if team_baserun_sb = 314 then imp_team_baserun_cs = 96;
  else team_baserun_cs = 53;
end;

*Catch all;
if missing(imp_team_baserun_cs) then do;
       imp_team_baserun_cs = 53;
       flag_team_baserun_cs = 1;
end;

*Transform Variables;
cap_team_batting_2b = TEAM_BATTING_2B;
        flag_cap_team_batting_2b = 0;
        cap_team_batting_3b = TEAM_BATTING_3B;
        flag_cap_team_batting_3b = 0;
        cap_team_pitching_h = TEAM_PITCHING_H;
```

```
        flag_cap_team_pitching_h = 0;
        cap_team_pitching_bb = TEAM_PITCHING_BB;
        flag_cap_team_pitching_bb = 0;
        cap_team_fielding_e = TEAM_FIELDING_E;
        flag_cap_team_fielding_e = 0;
        cap_imp_team_baserun_sb = imp_team_baserun_sb;
        flag_cap_imp_team_baserun_sb = 0;
        cap_imp_team_baserun_cs = imp_team_baserun_cs;
        flag_cap_imp_team_baserun_cs = 0;
        cap_imp_team_pitching_so = imp_team_pitching_so;
        flag_camp_imp_team_pitching_so = 0;

        *Cap TEAM_BATTING_2B to 5% > 167, 352 < 99%;
        if TEAM_BATTING_2B < 167.00 then do;
    cap_team_batting_2b = 167.00;
    flag_cap_team_batting_2b = 1;
end;
if TEAM_BATTING_2B > 352.00 then do;
    cap_team_batting_2b = 352.00;
    flag_cap_team_batting_2b = 1;
end;

*Cap TEAM_BATTING_3B to 5% > 23, 134 < 99%;
if TEAM_BATTING_3B < 23.00 then do;
    cap_team_batting_3b = 23.00;
    flag_cap_team_batting_3b = 1;
end;
if TEAM_BATTING_3B > 134.00 then do;
    cap_team_batting_3b = 134.00;
    flag_cap_team_batting_3b = 1;
end;

*Cap TEAM_PITCHING_H to 5% > 1316, 2563 < 95%;
if TEAM_PITCHING_H < 1316.00 then do;
    cap_team_pitching_h = 1316.00;
    flag_cap_team_pitching_h = 1;
end;
if TEAM_PITCHING_H > 2563.00 then do;
    cap_team_pitching_h = 2563.00;
    flag_cap_team_pitching_h = 1;
end;

*Cap TEAM_PITCHING_BB to 5% > 377, 924 < 99%;
if TEAM_PITCHING_BB < 377.00 then do;
    cap_team_pitching_bb = 377.00;
    flag_cap_team_pitching_bb = 1;
end;
```

```
if TEAM_PITCHING_BB > 924.00 then do;
    cap_team_pitching_bb = 924.00;
    flag_cap_team_pitching_bb = 1;
end;

*Cap TEAM_FIELDING_E to 5% > 100, 716 < 95%;
if TEAM_FIELDING_E < 100.00 then do;
    cap_team_fielding_e = 100.00;
    flag_cap_team_fielding_e = 1;
end;
if TEAM_FIELDING_E > 716.00 then do;
    cap_team_fielding_e = 716.00;
    flag_cap_team_fielding_e = 1;
end;

*Cap imp_team_baserun_sb to 5% > 36, 298 < 95%;
if imp_team_baserun_sb < 36.00 then do;
    cap_imp_team_baserun_sb = 36.00;
    flag_cap_imp_team_baserun_sb = 1;
end;
if imp_team_baserun_sb > 298.00 then do;
    cap_imp_team_baserun_sb = 298.00;
    flag_cap_imp_team_baserun_sb = 1;
end;

*Cap imp_team_baserun_cs to 5% > 25, 91 < 95%;
if imp_team_baserun_cs < 25.00 then do;
    cap_imp_team_baserun_cs = 25.00;
    flag_cap_imp_team_baserun_cs = 1;
end;
if imp_team_baserun_cs > 91.00 then do;
    cap_imp_team_baserun_cs = 91.00;
    flag_cap_imp_team_baserun_cs = 1;
end;

*Cap imp_team_pitching_so to 5% > 423, 1169 < 95%;
if imp_team_pitching_so < 423.00 then do;
    cap_imp_team_pitching_so = 423.00;
    flag_camp_imp_team_pitching_so = 1;
end;
if imp_team_pitching_so > 1169.00 then do;
    cap_imp_team_pitching_so = 1169.00;
    flag_camp_imp_team_pitching_so = 1;
end;

      log_team_batting_h = log(TEAM_BATTING_H+1);
log_imp_team_fielding_dp = log(imp_team_fielding_dp+1);
```

```
        drop team_batting_HBP; *Too many missing values, drop;
        drop index;
run;


/** Model 3 **/
title "Bingo Bonus - Proc GLM";
proc glm data=tempfile_for_regression  outstat=model_glm;
model target_wins =
    team_batting_h
    TEAM_BATTING_2B
    TEAM_BATTING_3B
    TEAM_BATTING_HR
    TEAM_BATTING_BB
    imp_team_batting_so
    flag_team_batting_so
    imp_team_baserun_sb
    flag_team_baserun_sb
    imp_team_baserun_cs
    flag_team_baserun_cs
    TEAM_FIELDING_E
    log_imp_team_fielding_dp
    flag_team_fielding_dp
    TEAM_PITCHING_BB
    TEAM_PITCHING_H;
run;
quit;

/** Model 3 **/
title "Bingo Bonus - Proc GENMOD";
proc genmod data=tempfile_for_regression;
model target_wins =
    team_batting_h
    TEAM_BATTING_2B
    TEAM_BATTING_3B
    TEAM_BATTING_HR
    TEAM_BATTING_BB
    imp_team_batting_so
    flag_team_batting_so
    imp_team_baserun_sb
    flag_team_baserun_sb
    imp_team_baserun_cs
    flag_team_baserun_cs
    TEAM_FIELDING_E
    log_imp_team_fielding_dp
    flag_team_fielding_dp
```

```
    TEAM_PITCHING_BB
    TEAM_PITCHING_H;;
run;
quit;


/** Code use for my poor's man decision tree made in Groovy **/

//Ariel Gamino - Predict 411 - Section 58
//Project 1
//Calculate average TEAM_BASERUN_CS for each TEAM_BASERUN_SB

def input = new File("/Users/arielgamino/Documents/Ariel/PredictiveAnalytics/Classes/Predict-
411/UNIT01/Project1/baserun_stolen_and_caught_stealing.csv")
def mean_values = [:]
def previous_cs = null
def counter = 1
def currentTotal = 0

input.eachLine{
  def tokens = it.tokenize(',')
  def cs = tokens[0]
  def sb = tokens[1]
  //println "${cs},${sb}"
  //Only use for calculation if sb exists
  if(cs!=previous_cs) {
    //the key has changed, calculate average
    def mean = currentTotal / counter
    if(previous_cs){
      if(mean!=0){
         mean_values[previous_cs] = mean
      } else {
       ///Use average for TEAM_BASERUN_SB which is 53
        mean_values[previous_cs] = 53
      }
    }
    if(sb){
       def sb_int = new Integer(sb)
       currentTotal = sb_int
    }else {
      currentTotal = 0
    }
    counter = 1
  } else {
    if(sb){
      def sb_int = new Integer(sb)
      currentTotal += sb_int
```

```
      counter++
    }
  }
  previous_cs = cs

}

def outfile = new
File("/Users/arielgamino/Documents/Ariel/PredictiveAnalytics/Classes/Predict-
411/UNIT01/Project1/baserun_stolen_and_caught_stealing_averages.csv")
outfile.write("")

mean_values.each{k,v ->
  outfile.append("${k},${v}\n")

}
```