Noé Flores
Money Ball Analysis

Introduction:
Determining the factors that have the most significant impact on a team's overall wins is paramount for success in the current manifestation of Major League Baseball. Predictive modeling techniques will be incorporated to analyze which of those factors have the highest predictive value for the number of wins a team can expect in a season. To accomplish this goal, we will be performing an exploratory analysis on our data set to gather information regarding the variables we will be incorporating into this research and following up with any data preparation needed to ensure that we are working with optimal data. Finally, we will build and compare several models to determine which has the greatest predictive value, interpretability, and practicality with respect to our goal.

Data Exploration:
A quick examination of the contents of our dataset produces the results below. There are 2,276 total observations and 17 different variables.

Figure 1: Summary table of Data Contents.

Money Ball Data	Contents
Observations	2276
Variables	17
Indexes	0
Observation Length	136
Deleted Observations	0
Compressed	NO
Sorted	NO

In Figure.2 we have a detailed breakdown of the variables.

Figure 2: Summary Data Dictionary and Variable Theoretical Effects.

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

As you can see from the tables in Figures 1 and 2, these observations represent variables of baseball statistics such as hits, types of hits, walks, errors, and strikeouts collected from 1871 to 2006. The variables in Figure 2 represent the actual measurable variables and their theoretically effect or impact on the outcome of baseball games. Given the length of time present in this data set, it's likely that some of the statistical information collected hasn't been consistent throughout the years, meaning that since the data's collection inception, the game and more importantly the metrics of baseball could have changed over this time period. It's likely some record of the collected statistics are inconsistent across the years, missing, or both and will need to be addressed as we go along through our analysis. Also of note, we can clearly see that some of these variables are or should be closely related to one another, such as the different types of base hits by the batters. There are also variables that are related to each other in an inverse fashion with respect to their effect on the outcome of baseball games, such as Walks by Batters vs. Walks Allowed.

In Figure.3 below, we have some basic summary statistics of the data set described above. First and foremost, we can make a quick examination of the data set to look for possible issues, such as missing data or a negative value when we know there shouldn't be any present. The variables in this table contain a breakdown of measures of central tendency and information with respect to missing values. Immediately present in this table is the min and max range between these variables. In some instances, there appears to be a relatively large range. This could be an indication of potential outliers that we may need to address. We also see some instances where there are a large number of missing values, such as in Batters Hit by Pitch and Caught Stealing. In the case of missing or erroneous values, we will need to take steps to impute or fix the data in order to produce a stable predictive model.

We also use the mean and median to get a better understanding of the distribution of our variables. If they are relatively close to each other, it's indicative that the data is likely to be closely divided around the mean or reasonably close to symmetrical. If the data is spread further apart, it's indicative that the distribution for that particular variable is not normally distributed. We will look to examine this a bit further by taking a closer look at the TARGET_WINS variable.

Figure 3: Summary variable types.

Variable	Label	Minimum	Maximum	Mean	Median	Mode	N	N Miss
TARGET_WINS		0	146	80.791	82	83	2276	0
TEAM_BATTING_H	Base Hits by batters	891	2554	1469.3	1454	1458	2276	0
TEAM_BATTING_2B	Doubles by batters	69	458	241.25	238	227	2276	0
TEAM_BATTING_3B	Triples by batters	0	223	55.25	47	35	2276	0
TEAM_BATTING_HR	Homeruns by batters	0	264	99.612	102	21	2276	0
TEAM_BATTING_BB	Walks by batters	0	878	501.56	512	502	2276	0
TEAM_BATTING_SO	Strikeouts by batters	0	1399	735.61	750	0	2174	102
TEAM_BASERUN_SB	Stolen bases	0	697	124.76	101	65	2145	131
TEAM_BASERUN_CS	Caught stealing	0	201	52.804	49	52	1504	772
TEAM_BATTING_HBP	Batters hit by pitch	29	95	59.356	58	54	191	2085
TEAM_PITCHING_H	Hits allowed	1137	30132	1779.2	1518	1494	2276	0
TEAM_PITCHING_HR	Homeruns allowed	0	343	105.7	107	114	2276	0
TEAM_PITCHING_BB	Walks allowed	0	3645	553.01	536.5	536	2276	0
TEAM_PITCHING_SO	Strikeouts by pitchers	0	19278	817.73	813.5	0	2174	102
TEAM_FIELDING_E	Errors	65	1898	246.48	159	122	2276	0
TEAM_FIELDING_DP	Double Plays	52	228	146.39	149	148	1990	286

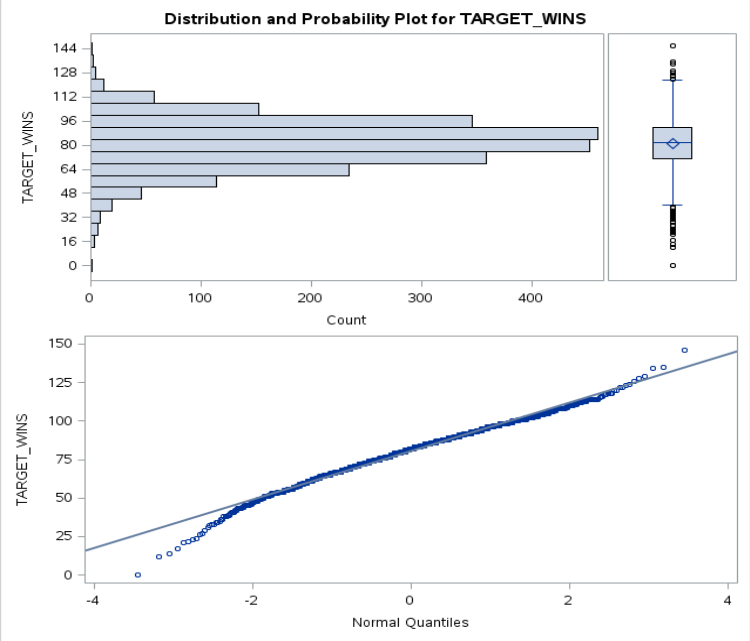
The TARGET_WINS variable information displayed in Figure 4 isolates the measures of central tendency for that specific variable that we previously discussed. The isolation of this variable illustrates just how "average" the average baseball season is for most of Major League Baseball, given the 162 game season currently played.

Figure 4: Summary of Target Wins.

Basic Statistical Measures			
Location		**Variability**	
Mean	80.791	**Std Deviation**	15.752
Median	82	**Variance**	248.130
Mode	83	**Range**	146
		Interquartile Range	21

The distribution for TARGET_WINS looks to be approximately normal. This is what we expected, given the information previously displayed and discussed in Figure .3. The majority of wins fall in the center of the distribution, and this is further confirmed with the Q_Q plot having the bulk of the TARGET_WINS data points fall on the straight line, although we do see some tailing at the bottom left of the plot representing outliers, which is to be expected. With that said, there will always be a possibility of teams performing below and beyond what is typically expected in a baseball season and to be clear, the same is not true for all of our variables. We only focused on TARGET_WINS as an example.

Figure 5: Distribution and Q_Q plot for TARGET_WINS.



Examination of some of the other histograms and Q_Q plots from some of the other variables appear to indicate that we may need to perform some variable transformations in order to improve the linearity and the fit of our model

With our end goal of building a predictive model for TARGET_WINS, we must identify possible predictor variables for TARGET_WINS. One way to asses possible predictors is to examine their correlation with respect to TARGET_WINS. The correlation information is displayed below.

Figure 6: Summary Correlation matrix to TARGET_WINS.

Correlation to TARGET_WINS Varaible and description	Correlation	P-Value	Observations
TEAM_BATTING_H **Base Hits by batters**	0.38877	<.0001	2276
TEAM_BATTING_2B **Doubles by batters**	0.2891	<.0001	2276
TEAM_BATTING_3B **Triples by batters**	0.14261	<.0001	2276
TEAM_BATTING_HR **Homeruns by batters**	0.17615	<.0001	2276
TEAM_BATTING_BB **Walks by batters**	0.23256	<.0001	2276
TEAM_BATTING_SO **Strikeouts by batters**	-0.03175	0.1389	2174
TEAM_BASERUN_SB **Stolen bases**	0.13514	<.0001	2145
TEAM_BASERUN_CS **Caught stealing**	0.0224	0.3853	1504
TEAM_BATTING_HBP **Batters hit by pitch**	0.0735	0.3122	191
TEAM_PITCHING_H **Hits allowed**	-0.10994	<.0001	2276
TEAM_PITCHING_HR **Home runs allowed**	0.18901	<.0001	2276
TEAM_PITCHING_BB **Walks allowed**	0.12417	<.0001	2276
TEAM_PITCHING_SO **Strikeouts by pitchers**	-0.07844	0.0003	2174
TEAM_FIELDING_E **Errors**	-0.17648	<.0001	2276
TEAM_FIELDING_DP **Double Plays**	-0.03485	0.1201	1990

What we see is that there aren't any numbers present with excessively high correlations to TARGET_WINS. There are some variables which are notably more highly correlated than others such as Base Hits by batters, and other team hitting variables. Variables such as Team fielding errors are as we would expect, negatively correlated to overall TARGET_WINS. There are a few surprises though. Double Plays appears to show a negative correlation to wins, and the same negative correlation is present in Strikeouts by pitchers. These results run counter to the proposed theoretical impact on team wins. Also of mention with these results is the fact that some variables have P-Values that would indicate a lack of statistical significance. Typically, a P-Value is considered statistically significant if it is less than or equal to .05. Batters hit by pitch also reminds us that we have to make a decision on the best method for dealing with the missing values.

The missing values we have will have to be addressed before building our models, but it may be prudent for us to have a baseline model with all the variables included for future comparisons. The models we will be developing in this analysis are linear regression models which allow us to model the relationship between our dependent variable TARGET_WINS and one or more independent or predictor variables.

Figure 7: Baseline Model ANOVA and Goodness of Fit results.

Number of Observations Read	2276
Number of Observations Used	191
Number of Observations with Missing Values	2085

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	15	15341	1022.73892	14.27	<.0001
Error	175	12546	71.6908		
Corrected Total	190	27887			

Root MSE	8.46704	**R-Square**	0.5501
Dependent Mean	80.9267	**Adj R-Sq**	0.5116
Coeff Var	10.46261		

The ANOVA table shows an F-test and P-Value that would appear to indicate significance at levels of 14.27 and p<.0001 respectively. The F-test or F-value is a test statistic used to decide whether the model as a whole has statistically significant predictive capability. Generally speaking, we want this number to be high. In order for us to accept our model's predictive ability we need a P-value that less than our specified significance level of .05.

The R-Square in this regression model explains 51% of the variability in TARGET_WINS when using this particular model. We usually look for our models to have an R-Square in the range of 50%. Our adjusted R-Square also indicates how well our model fits, adjusting for the number of variables. Given these results, it would appear that we have a decent model to work with from the onset. These numbers aren't quite what we want and more investigating into this initial model is required to make future baseline comparisons.

The following parameter results were produced as a result of running the initial model described above.

Figure 8: Baseline Model Parameter Estimate Results.

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t-Value**	**Pr > |t|**
Intercept	Intercept	1	60.28826	19.67842	3.06	0.0025
TEAM_BATTING_H	Base Hits by batters	1	1.91348	2.76139	0.69	0.4893
TEAM_BATTING_2B	Doubles by batters	1	0.02639	0.03029	0.87	0.3848
TEAM_BATTING_3B	Triples by batters	1	-0.10118	0.07751	-1.31	0.1935
TEAM_BATTING_HR	Homeruns by batters	1	-4.84371	10.50851	-0.46	0.6454
TEAM_BATTING_BB	Walks by batters	1	-4.45969	3.63624	-1.23	0.2217
TEAM_BATTING_SO	Strikeouts by batters	1	0.34196	2.59876	0.13	0.8955
TEAM_BASERUN_SB	Stolen bases	1	0.03304	0.02867	1.15	0.2507
TEAM_BASERUN_CS	Caught stealing	1	-0.01104	0.07143	-0.15	0.8773
TEAM_BATTING_HBP	Batters hit by pitch	1	0.08247	0.0496	1.66	0.0982
TEAM_PITCHING_H	Hits allowed	1	-1.89096	2.76095	-0.68	0.4943
TEAM_PITCHING_HR	Homeruns allowed	1	4.93043	10.50664	0.47	0.6395
TEAM_PITCHING_BB	Walks allowed	1	4.51089	3.63372	1.24	0.2161
TEAM_PITCHING_SO	Strikeouts by pitchers	1	-0.37364	2.59705	-0.14	0.8858
TEAM_FIELDING_E	Errors	1	-0.17204	0.0414	-4.16	<.0001
TEAM_FIELDING_DP	Double Plays	1	-0.10819	0.03654	-2.96	0.0035

The results above in Figure.8 is where we really start to see the flaws in blindly running a model without utilizing any methods to adjust for missing values in our variables or possibly completely removing some variables that would be too difficult to correct by simply being imputed. This model also lacks any automated variable selection methods that could be incorporated to assist in building an optimal model, such as forward or stepwise selection. The parameter estimates are also troubling in that some of the variables we would assume to have a positive impact on wins actually appear to have a negative overall impact, such as Triples by Batters, Homeruns by Batters, and hits allowed. Going through the exercise of running a base regression model gives us a good benchmark to compare our future models. This exercise also gives credence to the need for fixing and properly preparing our data for modeling.

Data Preparation:
It should now be clear that the missing values below will need to be addressed before we go further in our model building. There are two primary choices when it comes to missing or erroneous data. We can remove the variable from the model altogether, or we can impute them.

Figure 9: Summary of Variables with missing data.

Variable	Label	Mean	Median	Mode	N	N Miss
TEAM_BATTING_HBP	Batters hit by pitch	59.356	58	54	191	2085
TEAM_BASERUN_CS	Caught stealing	52.804	49	52	1504	772
TEAM_FIELDING_DP	Double Plays	146.388	149	148	1990	286
TEAM_BASERUN_SB	Stolen bases	124.762	101	65	2145	131
TEAM_PITCHING_SO	Strikeouts by pitchers	817.730	813.5	0	2174	102
TEAM_BATTING_SO	Strikeouts by batters	735.605	750	0	2174	102

Figure .9 provides a quick summary of the variables we previously identified that will require attention. In the case of Batters hit by a pitch and Caught Stealing, it's clear that there is a substantial amount of missing data. Merely imputing values for 90% of the values as we would have to do for Batters hit by pitch would be a stretch of our abilities and ill advised. Add to this the fact that Batters hit by a pitch and Caught stealing had a very low correlation with respect to TARGET_WINS and statistically insignificant P-values of .3122 and .3853, each well above our .05 threshold, for these reasons; we will remove TEAM_BATTING_HBP and TEAM_BASERUN_CS completely from future models.

Figure 10: Summary Correlations for Variables with missing data.

Correlation to TARGET_WINS Variable and description	Correlation	P-Value	Observations
TEAM_BATTING_HBP Batters hit by pitch	0.0735	0.3122	191
TEAM_BASERUN_CS Caught stealing	0.0224	0.3853	1504
TEAM_FIELDING_DP Double Plays	-0.03485	0.1201	1990
TEAM_BASERUN_SB Stolen bases	0.13514	<.0001	2145
TEAM_PITCHING_SO Strikeouts by pitchers	-0.07844	0.0003	2174
TEAM_BATTING_SO Strikeouts by batters	-0.03175	0.1389	2174

If you look at the table in Figure.10, you will notice that we have a few other variables with missing values, low correlations, and statistically insignificant P-Values very similar to Batters hit by pitch. This is evident in variables such as Caught Stealing Double Plays, and Strikeouts by Batters. In these cases, however, we will not simply delete the variable from our future models. While their correlation and P-Value don't support their inclusion, these variables represent statistics that intuitively feel like they could be relevant in predicting wins in baseball. This is admittedly more perception than science, and we are using what we know about the game of baseball to make that decision. With that said, we will be imputing the missing values for Caught Stealing, Double Plays, Stolen Bases, Strikeouts by Pitchers, and Strikeouts by Batters.

To impute these variables, we will incorporate two different methods. The first method will replace any missing values with the variable's mean value. The second will replace any missing values with the variable's median. For example, TEAM_BATTING_SO has 102 missing values. In each of these cases, we will replace the missing value with the mean which is equal to735.605 or the median, which is equal to 750. These methods should give us a decent solution to our missing data value problems.

With respect to possible outlying data, or data that finds itself to be two or more standard deviations from the mean, we don't want to adjust for too many of those data points at this time. The reason for this is the fact that we feel our data set has enough observations that we can include most of the data points, and still remain confident that our model will adhere to basic regression model assumptions of normality. If we had less than the 2,276 records we have, we would definitely need to consider a solution for outlying data points. Typically the larger the number of observations, the less effect outliers may have. There is one caveat to this decision, and that is our TEAM_PITCHING_H (Hits allowed). The range on this variable is just too significant to ignore and examining this specific variable's quartiles leads us to make the decision that we must adjust these observations.

Figure 11: Basic Statistics and Quartiles for Hits allowed.

Basic Statistical Measures			
	Location	**Variability**	
Mean	1779.21	**Std Deviation**	1407
Median	1518	**Variance**	1979207
Mode	1494	**Range**	28995
		Interquartile Range	264

Quantiles	
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

For this variable, we will use a transformation that caps all observations on the upper end at the 99% value of 7,093. Adjusting for outliers is not an easy decision, and at times may have an adverse effect on our model's accuracy, but as you can see from the information in Figure.11, the values for this variable on the upper end are just too extreme and need to be capped.

Modeling:

Model#1 Imputed with Mean Values for missing Data with normal selection:

This model contains all of the variables except, TEAM_BATTING_HBP TEAM_BASERUN_CS which were removed in our data preparation stage. We imputed any missing observations with the mean of each variable. In this model's case, we adjusted TEAM_BATTING_SO, TEAM_BASERUN_SB, TEMA_PITCHING_SO, and TEAM_FIELDING_DP. We also adjusted TEAM_PITCHING_H and capped the high end observation values at the 99th percentile value to adjust for some significant outlying data points. No automated selection process was incorporated in this model.

Figure.12 Regression model #1 Imputed with MEAN Values

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	180250	13865	81.62	<.0001
Error	2262	384246	169.87		
Corrected Total	2275	564496			

R-Square	0.3193
Adj R-Sq	0.3154

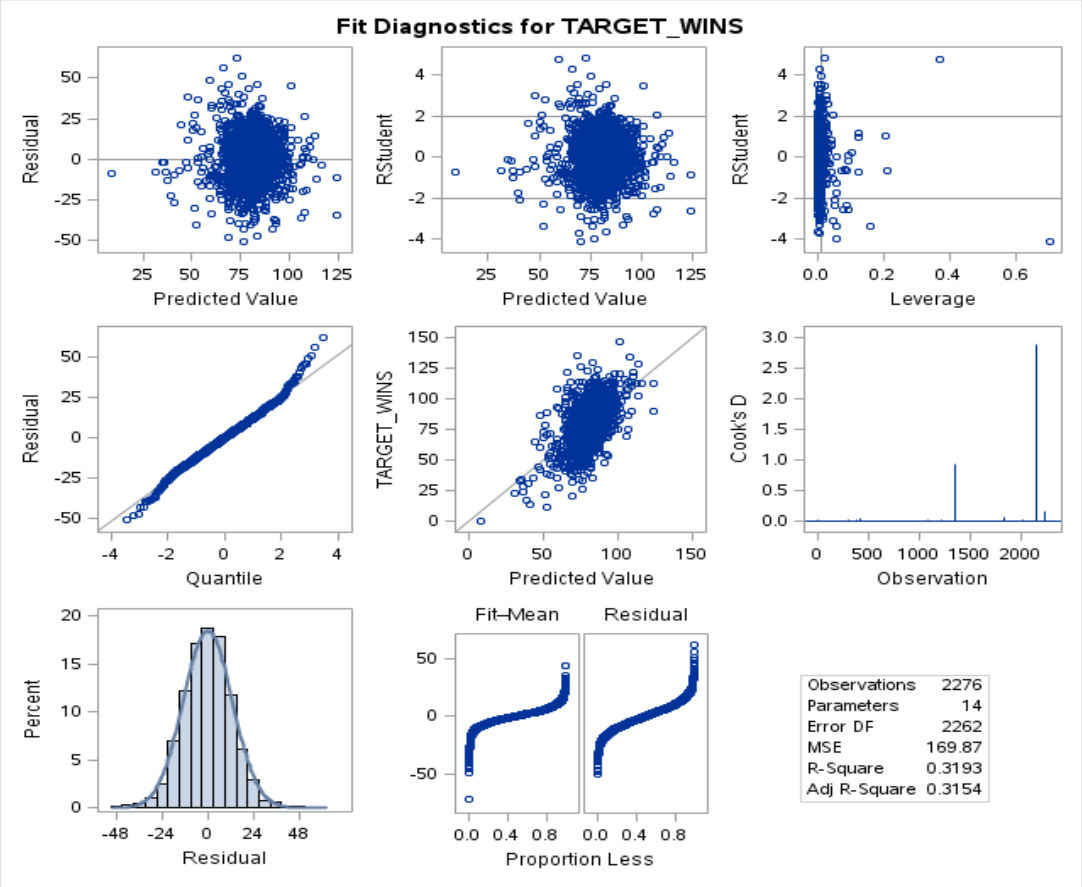
The P-Value for our first model indicates that it is statistically significant, coming in at <.0001 which is well below our .05 threshold. The Adjusted R-Square is .3154 which is below the .50 level we would typically want to see. The F-Value is 81.62, which is good. We typically want F-Value to be positive and large. The mean Square error is 169.87. These three metrics will serve as our benchmarks moving forward for our decision on which model to choose.

Figure.13 Parameter estimates Regression model #1

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	25.038	5.304	4.72	<.0001
TEAM_BATTING_H	Base Hits by batters	1	0.043	0.004	10.48	<.0001
TEAM_BATTING_2B	Doubles by batters	1	-0.019	0.009	-2.11	0.0353
TEAM_BATTING_3B	Triples by batters	1	0.081	0.017	4.65	<.0001
TEAM_BATTING_HR	Homeruns by batters	1	0.061	0.028	2.20	0.0276
TEAM_BATTING_BB	Walks by batters	1	0.022	0.006	3.57	0.0004
TEAM_BASERUN_SB	Stolen bases	1	0.033	0.004	7.48	<.0001
TEAM_PITCHING_H	Hits allowed	1	0.002	0.001	2.43	0.0152
TEAM_PITCHING_HR	Homeruns allowed	1	0.005	0.025	0.22	0.8266
TEAM_PITCHING_BB	Walks allowed	1	-0.008	0.004	-2.00	0.046
TEAM_PITCHING_SO	Strikeouts by pitchers	1	0.003	0.001	3.25	0.0012
TEAM_FIELDING_E	Errors	1	-0.028	0.003	-9.46	<.0001
TEAM_BATTING_SO	Strikeouts by batters	1	-0.009	0.003	-3.41	0.0007
TEAM_FIELDING_DP	Double Plays	1	-0.119	0.013	-9.11	<.0001

For the most part, model#1 looks ok. In most cases, the parameter estimates displayed in Figure.13 coincide with the effect we believe they should have on the outcome of a baseball game. Base hits by batters, Triples by batters, Homeruns by batters all have a positive impact on the number of wins. Strikeouts by batters, Errors, and Walks allowed have a negative impact on wins. However, there are some peculiarities present. The coefficients for Doubles by batters and Double plays show negative values, meaning that our model believes these variables have a negative impact on the outcome of a game. This is counterintuitive to the variable's theoretical effect.

Figure 14: Model#1 Plots and residuals.



The overall diagnostics for our first model indicates a decent fit. The residual plot shows some clustering, which is not what we want to see. Typically we want a random distribution of data points. The histogram shows a fairly normal distribution which is good and our Q_Q plot confirms that observation by having most of the data points lie on the straight line. The Cook's D shows the presence of two outliers, which is consistent with what we see in the Q_Q plot. Overall, this is a good baseline model for us.

Model#2 Imputed with Mean Values for missing Data with Stepwise selection:

This model contains all of the same variables and variable adjustments as Model#1, but we incorporated an automated variable selection process. Up until this point, the procedure we used to build our models started with a fixed set of variables that we included into a regression model. In order to build a more inclusive and robust model, we will incorporate the use of various automated variable selection procedures called Stepwise selection, which adds or removes predictor variables in steps based on the impact to the statistical significance of the model

Figure.15 Regression model #2 Imputed with MEAN Values Stepwise selection

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	180242	15020	88.46	<.0001
Error	2263	384255	169.80		
Corrected Total	2275	564496			

R-Square	0.3193
Adj R-Sq	0.3157

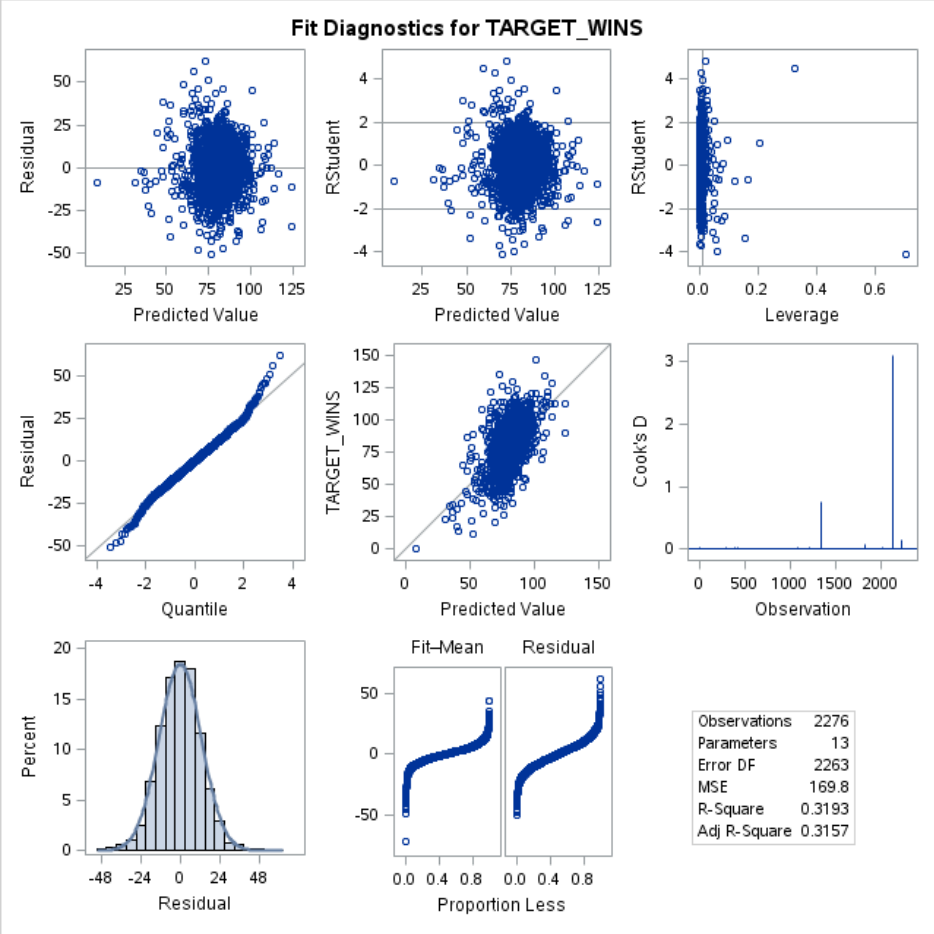
The P-Value for this model is statistically significant, coming in at <.0001 which is well below our .05 threshold. The Adjusted R-Square is .3157 and the F-Value is 88.46 which is better than our first model and expected given the goal of stepwise selection to improve statistical significance. The mean Square error is 169.80

Figure.16 Parameter estimates Regression model #2

Variable		Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept		24.93197	5.28047	3785.3228	22.29	<.0001
TEAM_BATTING_H	Base Hits by batters	0.04279	0.00408	18648	109.83	<.0001
TEAM_BATTING_2B	Doubles by batters	-0.01932	0.00915	756.7622	4.46	0.035
TEAM_BATTING_3B	Triples by batters	0.08145	0.01703	3882.4158	22.86	<.0001
TEAM_BATTING_HR	Homeruns by batters	0.06632	0.00962	8062.167	47.48	<.0001
TEAM_BATTING_BB	Walks by batters	0.0212	0.00579	2273.8376	13.39	3E-04
TEAM_BASERUN_SB	Stolen bases	0.03317	0.00443	9512.0888	56.02	<.0001
TEAM_PITCHING_H	Hits allowed	0.00253	0.001	1082.774	6.38	0.012
TEAM_PITCHING_BB	Homeruns allowed	-0.00803	0.0039	719.83673	4.24	0.04
TEAM_PITCHING_SO	Strikeouts by pitchers	0.00292	0.000878	1883.35	11.09	9E-04
TEAM_FIELDING_E	Errors	-0.02768	0.00289	15536	91.49	<.0001
TEAM_BATTING_SO	Strikeouts by batters	-0.00858	0.00251	1982.4035	11.68	6E-04
TEAM_FIELDING_DP	Double Plays	-0.11872	0.01303	14086	82.96	<.0001

The majority of coefficients continue to fit their expected theoretical results. However, we continue to see some variables running counterintuitive values. Double Plays is displaying a negative value as are doubles by batters. This is something that we should continue to monitor. The major difference in this model is that we have 12 variables as opposed to 13. This is a result of our stepwise selection process.

Figure 17: Model#2 Plots and residuals.



The overall diagnostics for our second model once again shows a residual plot with some clustering. The histogram shows a reasonably normal distribution and the Q_Q plot has most of the data points lying on the straight line. We see two outliers on the Cook's D once again. Overall, this model isn't drastically different from the first, with the exception of the reduced variable (Coefficient).

Model#3 Imputed with Median Values for missing Data Stepwise selection:

Since we had slightly improved results with the stepwise automated selection process in Model#2, we decided to use the automated process once again, but this time, we wanted to test the effect of imputing the missing observation with the variable's median rather than the mean. In this model's case, we adjusted TEAM_BATTING_SO, TEAM_BASERUN_SB, TEMA_PITCHING_SO, and TEAM_FIELDING_DP with median values and we kept the outlier adjustment to TEAM_PITCHING_H.

Figure.18 Regression model #3 Imputed with MEDIAN Values Stepwise selection

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	176856	17686	103.34	<.0001
Error	2265	387640	171.14		
Corrected Total	2275	564496			

R-Square	0.3133
Adj R-Sq	0.3103

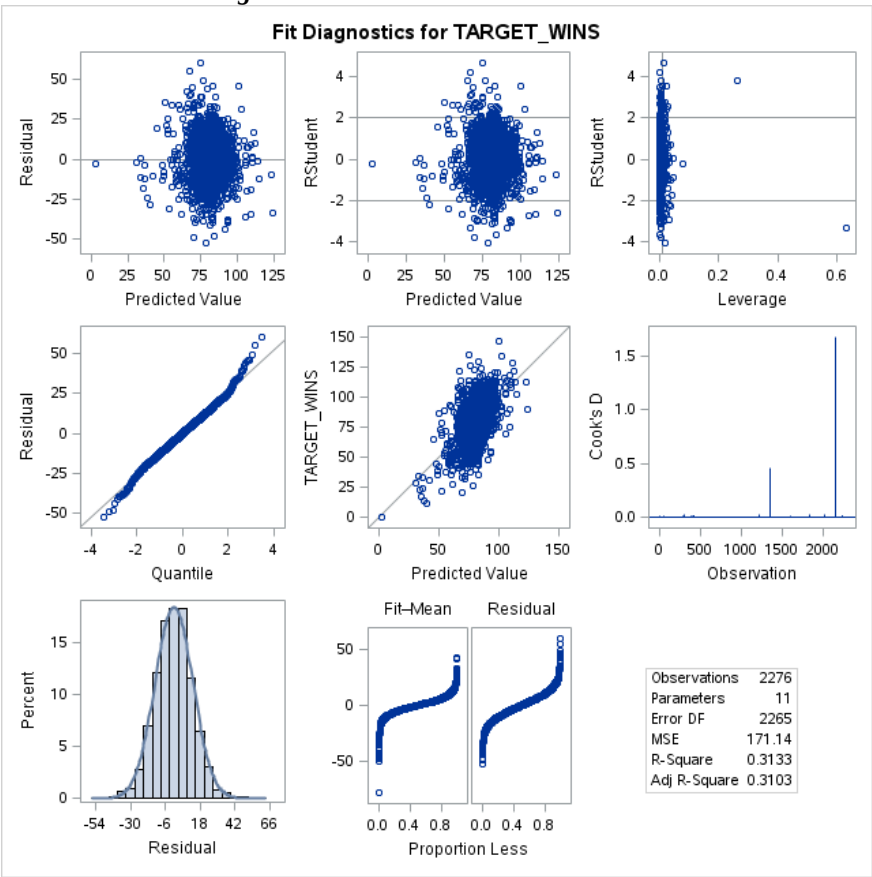
The P-Value for this model is again statistically significant, coming in at <.0001 but the Adjusted R-Square is .3103 which is a drop off from the previous model. The F-Value, on the other hand, came in at 103.34 which is better than each of our first two models. The high F-Value is expected given the use of stepwise selection, but there appears to be a change when using median values for imputation as opposed to mean values. The mean Square error is 171.14 which is a slight drop off in performance from our first two models.

Figure.19 Parameter estimates Regression model #2

Variable		Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept		22.86813	5.23496	3265.8425	19.08	<.0001
TEAM_BATTING_H	Base Hits by batters	0.04783	0.00364	29604	172.98	<.0001
TEAM_BATTING_2B	Doubles by batters	-0.02193	0.00917	978.88187	5.72	0.017
TEAM_BATTING_3B	Triples by batters	0.07444	0.01632	3561.0917	20.81	<.0001
TEAM_BATTING_HR	Homeruns by batters	0.0654	0.00961	7932.6321	46.35	<.0001
TEAM_BATTING_BB	Walks by batters	0.01173	0.00338	2065.3616	12.07	5E-04
TEAM_BASERUN_SB	Stolen bases	0.02597	0.00419	6569.2617	38.38	<.0001
TEAM_PITCHING_SO	Strikeouts by pitchers	0.00222	0.000597	2360.4695	13.79	2E-04
TEAM_FIELDING_E	Errors	-0.02209	0.00203	20335	118.82	<.0001
TEAM_BATTING_SO	Strikeouts by batters	-0.00716	0.00239	1536.7728	8.98	0.003
TEAM_FIELDING_DP	Double Plays	-0.12178	0.01294	15150	88.52	<.0001

We still see two coefficients (variables) outputting contradictory values with respect to their theoretical effect on the outcome of baseball games. Doubles by batters and Double plays may need to be addressed manually in future models. This third model does have fewer coefficients than our previous models. The stepwise process selected only 10 variables in its operation.

Figure 20: Model#3 Plots and residuals.



The fit diagnostics for our third model once again show a residual plot with some clustering. The histogram shows a fairly normal distribution and the Q_Q plot has most of the data points lying on the straight line although there does appear to be more deviation on the upper right side of this plot. We once again see two outliers on the Cook's D , but overall, this model falls in line with the first two, albeit with slight variations in Coefficients, F-Values and Adjusted R-Square.

Model#4 Imputed with Mean Values for missing Data Adjusted R-Square Selection:

For Model#4, we went back to using the data set prepared with imputed observations using mean values and the same variable omission and adjustments from our first two models. The only change we made was to use a different automated variable selection process. We decided to build our model based on automatic Adjusted R-Square selection which creates the model based on the optimal Adjusted R-Square

Figure.21 Regression model #4 Imputed with MEDIAN Values Adjusted R-Square selection

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	180242	15020	88.46	<.0001
Error	2263	384255	169.80		
Corrected Total	2275	564496			

R-Square	0.3193
Adj R-Sq	0.3157

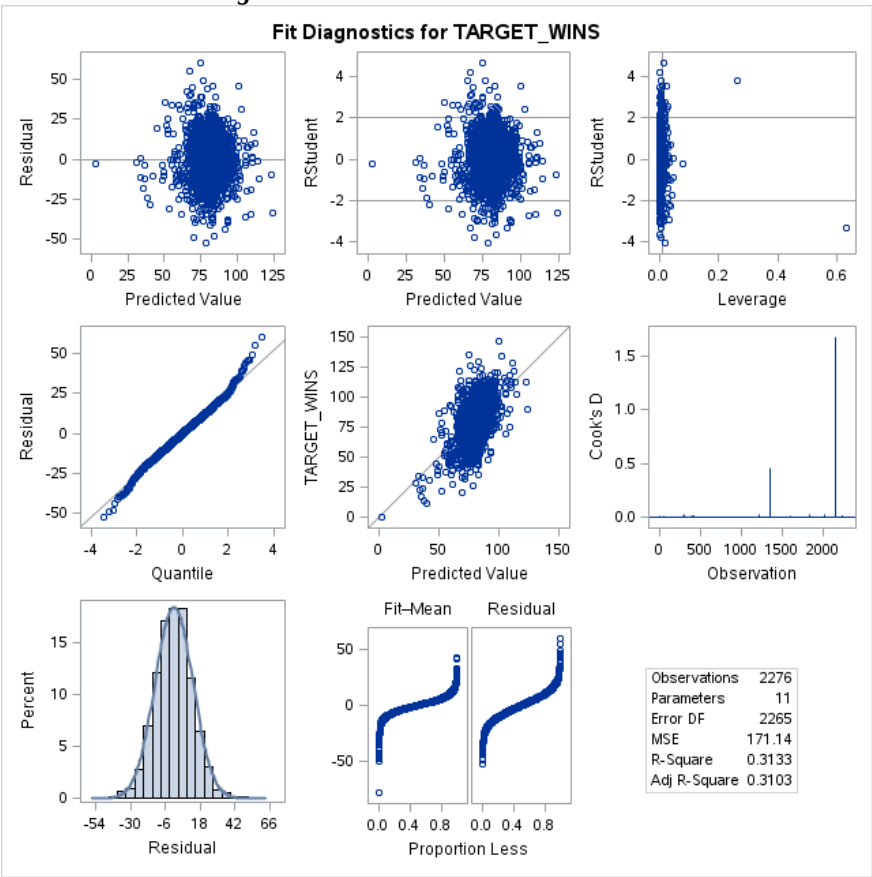
This fourth model produced the exact same results as our second model, which was based off Stepwise selection and the corresponding variable adjustments. The P-Value came in at <.0001 and the Adjusted R-Square of .3157 is precisely the same as Model#2. The F-Value is also the same as Model#2 at 88.46 but it does represent a drop off from Model#3. The mean Square error of 169.80 is also the same as Model#2.

Figure.19 Parameter estimates Regression model #4

Variable		Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept		24.93197	5.28047	3785.3228	22.29	<.0001
TEAM_BATTING_H	Base Hits by batters	0.04279	0.00408	18648	109.83	<.0001
TEAM_BATTING_2B	Doubles by batters	-0.01932	0.00915	756.7622	4.46	0.0349
TEAM_BATTING_3B	Triples by batters	0.08145	0.01703	3882.4158	22.86	<.0001
TEAM_BATTING_HR	Homeruns by batters	0.06632	0.00962	8062.167	47.48	<.0001
TEAM_BATTING_BB	Walks by batters	0.0212	0.00579	2273.8376	13.39	0.0003
TEAM_BASERUN_SB	Stolen bases	0.03317	0.00443	9512.0888	56.02	<.0001
TEAM_PITCHING_H	Hits allowed	0.00253	0.001	1082.774	6.38	0.0116
TEAM_PITCHING_BB	Homeruns allowed	-0.00803	0.0039	719.83673	4.24	0.0396
TEAM_PITCHING_SO	Strikeouts by pitchers	0.00292	0.000878	1883.35	11.09	0.0009
TEAM_FIELDING_E	Errors	-0.02768	0.00289	15536	91.49	<.0001
TEAM_BATTING_SO	Strikeouts by batters	-0.00858	0.00251	1982.4035	11.68	0.0006
TEAM_FIELDING_DP	Double Plays	-0.11872	0.01303	14086	82.96	<.0001

As we experienced in all the previous models, we still see the same two coefficients (variables) outputting contradictory values with respect to their theoretical effect on the outcome of baseball games. This third model does have fewer coefficients than our previous models. The stepwise process selected only 10 variables in its process.

Figure 20: Model#4 Plots and residuals.



The fit diagnostics for our fourth model once again show a residual plot with some clustering. The histogram shows a fairly normal distribution and the Q_Q plot has most of the data points lying on the straight line although there does appear to be more deviation on the upper right side of this plot. We once again see two outliers on the Cook's D , but overall, this model falls in line with the first two albeit with slight variations in Coefficients, F-Values and Adjusted R-Square.

Model# 5 Adjusted R-Square Selection with manual adjustments:

For Model#5, we decided to use the base variables from Model#4,which incorporated the automatic Adjusted R-Square variable selection process, but we decided to manually remove those two variables that continued to show impact on the model counter to what we would expect. This model will now be run without the inclusion of TEAM_BATTING_2B and TEAM_FIELDING_DP, Doubles by batters and Double plays. The results are below.

Figure.21 Regression model #5 Adjusted R-Square selection excluding Doubles and Double Plays

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	165278	16528	93.77	<.0001
Error	2265	399219	176.26		
Corrected Total	2275	564496			

R-Square	0.2928
Adj R-Sq	0.2897

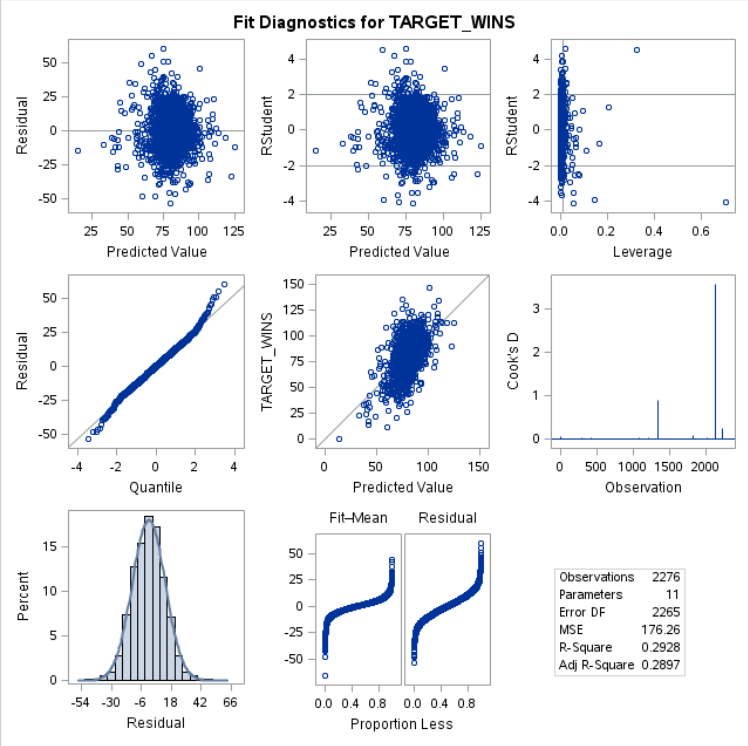
This fifth model produced the worst statistical results thus far. It isn't surprising given the fact that we decided to arbitrarily remove some variables without any scientific backing for doing so other than the fact that they continued to show values that were counterintuitive. The F-Value came in at a good level of 93.77, but the Adjusted R-Square was poor at .2897. Mean Square Error wasn't significantly different from the other models, but it was definitely higher and we prefer that measurement to be lower. It's possible we have some serious issues with collinearity and this process could potentially benefit significantly from a Principal Component Analysis or Factor Analysis in the future.

Figure.22 Parameter estimates Regression Model #5

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	**1**	15.20547	4.85945	3.13	0.0018
TEAM_BATTING_H	Base Hits by batters	**1**	0.03515	0.00326	10.78	<.0001
TEAM_BATTING_3B	Triples by batters	**1**	0.09555	0.01721	5.55	<.0001
TEAM_BATTING_HR	Homeruns by batters	**1**	0.04894	0.00961	5.09	<.0001
TEAM_BATTING_BB	Walks by batters	**1**	0.01731	0.00589	2.94	0.0033
TEAM_BASERUN_SB	Stolen bases	**1**	0.04143	0.00441	9.4	<.0001
TEAM_PITCHING_H	Hits allowed	**1**	0.00328	0.00102	3.23	0.0013
TEAM_PITCHING_BB	Walks allowed	**1**	-0.01045	0.00396	-2.64	0.0084
TEAM_PITCHING_SO	Strikeouts by pitchers	**1**	0.00269	0.0008908	3.02	0.0025
TEAM_FIELDING_E	Errors	**1**	-0.02926	0.00294	-9.97	<.0001
TEAM_BATTING_SO	Strikeouts by batters	**1**	-0.00666	0.00249	-2.68	0.0075

Since we made the manual adjustment, all the coefficients make intuitive sense with respect to the effect on a baseball game, but the model's goodness of fit will be the primary driver of selection and not the way the coefficients are lining up..

Figure 23: Model#5 Plots and residuals.



The fit diagnostics for our fifth model show a not surprisingly similar residual plot with some clustering. The histogram continues to display a fairly normal distribution, and the Q_Q plot mostly confirms that observation with the exception of some mild tailing in the upper right-hand corner of the plot. The Cook's D looks relatively unchanged from the previous models.

Model Selection:
We were able to run five different models based on different compositions, variable adjustments and variable omissions as well as automated selection processes. The table below provides some goodness of fit metrics we previously discussed during the modeling process.

Figure.23 Model Comparison

Model	Number of Variables	Adj-R-Square	Model R-Square	F-Value	Mean Square Error
(1) Mean Imp Variable adjustment/Normal selection	13	0.3154	0.3193	81.62	169.88
(2) Mean Imp Variable adjustment/Stepwise selection	12	0.3157	0.3193	88.46	169.80
(3) Median Imp Variable adjustment/Stepwise selection	10	0.3103	0.3133	103.34	171.14
(4) Mean Imp Variable adjustment/Adj-R-Square selection	12	0.3157	0.3193	88.46	169.80
(5) Adj-R-Square selection with Manual adjustments	10	0.2897	0.2928	93.77	176.26

When we first started this process, we mentioned that we would base our model selection on Adjusted R-Square, F-Value, and Mean Square Error. If those three criteria alone were sufficient for our needs, we could move forward and select Model#2 or Model#4 based on the best overall combination of those three metrics. However, it is difficult to ignore the fact that in each of those models we had coefficients that were counterintuitive to the results we would expect given the nature of baseball games. Our initial analysis of the situation leads us to believe that we would benefit from a Principal Component Analysis or Factor Analysis which would help us control some of the suspected Multicolinearity within the model's variables. Since this analysis is focused on simple OLS regression, we feel compelled to select Model#5 as the right choice. We aren't truly happy with selecting Model#5 but there doesn't appear to be a better solution at the moment. Model#5 remains statistically significant with an F-Value of 93.77, which is second highest among all our models, the Adjusted R-Square isn't terribly different from the others and the Mean Square Error is reasonable with respect to the other models. The biggest factor in selecting this model is that all the coefficients make intuitive sense, and there are only 10 variables making it more user-friendly and instinctive. These decisions are where the data scientist earns their keep.

Conclusion:

After going through and analyzing the data with various modeling techniques we generated five different regression models aiming to predict target wins in baseball. The data was initially examined for inconsistencies, such as missing observations or egregious outliers, and further examined to gather information on the correlation between those variables. The modeling process itself proved to be challenging. The five different models we generated, each had various strengths and weaknesses. We chose a model that was not generated through 100% statistical means, but the model represented the best intuitive solution for our end goal.

With that said, it is our opinion that while there are countless possibilities for building robust models, we find it easier to visualize one that is simple, with fewer parameters, and coefficients that coincide with our beliefs as to their effect on the outcome of baseball games. While the variables need to be statistically significant of course, we can rely on other measures besides Adjusted R-Square, F-Value, and Mean Square Error to assess the model's initial predictive ability and run it through a testing data set to gather further information.

BONUS

I used PROC GLM and PROC GENMOD to produce two new predictive models for our data. I ran the PROCS in SAS using the same variables in Model#5 in my original analysis and in both cases compared the results.

PROC GLM vs PROC REG Model#5

PROC GLM

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	165278	16527.78	93.77	<.0001
Error	2265	399219	176.26		
Corrected Total	2275	564496			

PROC REG #5

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	165278	16528	93.77	<.0001
Error	2265	399219	176.26		
Corrected Total	2275	564496			

PROC GLM
R-Square
0.2928

PROC Reg #5
R-Square
0.2928

PROC GLM

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	15.2055	4.8594	3.13	0.0018
TEAM_BATTING_H	0.0352	0.0033	10.78	<.0001
TEAM_BATTING_3B	0.0955	0.0172	5.55	<.0001
TEAM_BATTING_HR	0.0489	0.0096	5.09	<.0001
TEAM_BATTING_BB	0.0173	0.0059	2.94	0.0033
TEAM_BASERUN_SB	0.0414	0.0044	9.4	<.0001
TEAM_PITCHING_H	0.0033	0.0010	3.23	0.0013
TEAM_PITCHING_BB	-0.0104	0.0040	-2.64	0.0084
TEAM_PITCHING_SO	0.0027	0.0009	3.02	0.0025
TEAM_FIELDING_E	-0.0293	0.0029	-9.97	<.0001
TEAM_BATTING_SO	-0.0067	0.0025	-2.68	0.0075

PROC Reg #5

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	15.2055	4.8595	3.13	0.0018
TEAM_BATTING_H	0.0352	0.0033	10.78	<.0001
TEAM_BATTING_3B	0.0956	0.0172	5.55	<.0001
TEAM_BATTING_HR	0.0489	0.0096	5.09	<.0001
TEAM_BATTING_BB	0.0173	0.0059	2.94	0.0033
TEAM_BASERUN_SB	0.0414	0.0044	9.4	<.0001
TEAM_PITCHING_H	0.0033	0.0010	3.23	0.0013
TEAM_PITCHING_BB	-0.0105	0.0040	-2.64	0.0084
TEAM_PITCHING_SO	0.0027	0.0009	3.02	0.0025
TEAM_FIELDING_E	-0.0293	0.0029	-9.97	<.0001
TEAM_BATTING_SO	-0.0067	0.0025	-2.68	0.0075

When using PROC GLM I got almost the exact same results as the PROC REG procedure using the same selected variables. There might be some discrepancies in thousands place, but overall exactly the same.

I'm aware that this is and should be the case when you compare both outputs and typically if there are differences between the two it might be due to the fact that a variable was incorrectly coded or the simple anser is that we are using two different models, but we just don't know it.

PROC GENMOD vs PROC REG Model#5

PROC GENMOD

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	2265	399218.624	176.26
Scaled Deviance	2265	2276.000	1.00
Pearson Chi-Square	2265	399218.624	176.26
Scaled Pearson X2	2265	2276.000	1.00
Log Likelihood		-9109.652	
Full Log Likelihood		-9109.652	
AIC (smaller is better)		18243.304	
AICC (smaller is better)		18243.442	
BIC (smaller is better)		18312.066	

PROC GENMOD

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	15.2055	4.8477	5.7042	24.7068	9.84	0.0017
TEAM_BATTING_H	1	0.0352	0.0033	0.0288	0.0415	116.72	<.0001
TEAM_BATTING_3B	1	0.0955	0.0172	0.0619	0.1292	30.96	<.0001
TEAM_BATTING_HR	1	0.0489	0.0096	0.0301	0.0677	26.06	<.0001
TEAM_BATTING_BB	1	0.0173	0.0059	0.0058	0.0288	8.69	0.0032
TEAM_BASERUN_SB	1	0.0414	0.0044	0.0328	0.05	88.87	<.0001
TEAM_PITCHING_H	1	0.0033	0.001	0.0013	0.0053	10.46	0.0012
TEAM_PITCHING_BB	1	-0.0104	0.004	-0.0182	-0.0027	6.99	0.0082
TEAM_PITCHING_SO	1	0.0027	0.0009	0.001	0.0044	9.18	0.0024
TEAM_FIELDING_E	1	-0.0293	0.0029	-0.035	-0.0235	99.84	<.0001
TEAM_BATTING_SO	1	-0.0067	0.0025	-0.0115	-0.0018	7.21	0.0073
Scale	1	13.244	0.1963	12.8648	13.6344		

PROC Reg #5

Parameter	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	15.2055	4.8595	3.13	0.0018
TEAM_BATTING_H	0.0352	0.0033	10.78	<.0001
TEAM_BATTING_3B	0.0956	0.0172	5.55	<.0001
TEAM_BATTING_HR	0.0489	0.0096	5.09	<.0001
TEAM_BATTING_BB	0.0173	0.0059	2.94	0.0033
TEAM_BASERUN_SB	0.0414	0.0044	9.4	<.0001
TEAM_PITCHING_H	0.0033	0.0010	3.23	0.0013
TEAM_PITCHING_BB	-0.0105	0.0040	-2.64	0.0084
TEAM_PITCHING_SO	0.0027	0.0009	3.02	0.0025
TEAM_FIELDING_E	-0.0293	0.0029	-9.97	<.0001
TEAM_BATTING_SO	-0.0067	0.0025	-2.68	0.0075

While the output of Proc GENMOD is slightly different from Proc Reg (Missing R-Square, etc), the overall results with respect to the coefficients and Mean Square Errors are exactly the same. This is surely due to the fact that we are running two simple linear models with discreet outcomes in both cases.

Appendix: SAS code

```
/****Load data***/
%let ME =  noeflores20160;
%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;
proc print data=&INFILE.(OBS=10);
run;
/***Check contents of data***/
proc contents data=&INFILE order=varnum;
run;
/***Verify top 10 observations***/
proc print data=&INFILE.(OBS=10);
run;
/***Exploratory data analysis***/
proc means data=&INFILE MIN MAX MEAN MEDIAN MODE std N NMISS;
var TARGET_WINS
TEAM_BATTING_H
TEAM_BATTING_2B
TEAM_BATTING_3B
TEAM_BATTING_HR
TEAM_BATTING_BB
TEAM_BATTING_SO
TEAM_BASERUN_SB
TEAM_BASERUN_CS
```

```
TEAM_BATTING_HBP
TEAM_PITCHING_H
TEAM_PITCHING_HR
TEAM_PITCHING_BB
TEAM_PITCHING_SO
TEAM_FIELDING_E
TEAM_FIELDING_DP;
run;
/***Proc Means Target Wins***/
proc means data=&INFILE MEAN MEDIAN MODE N NMISS STDDEV VARDEF=DF CLM;
var TARGET_WINS;
run;
/***Run PROC UNIVARIATE to examine distributions and outliers***/
PROC UNIVARIATE NORMAL PLOT DATA=&INFILE.;
        VAR TEAM_BATTING_H
                TEAM_BATTING_2B
                TEAM_BATTING_3B
                TEAM_BATTING_HR
                TEAM_BATTING_BB
                TEAM_BATTING_SO
                TEAM_BASERUN_SB
                TEAM_BASERUN_CS
                TEAM_BATTING_HBP
                TEAM_PITCHING_H
                TEAM_PITCHING_HR
                TEAM_PITCHING_BB
                TEAM_PITCHING_SO
                TEAM_FIELDING_E
```

```
                    TEAM_FIELDING_DP
                    ;
RUN;
/***PROC Univariate For Target Wins***/
PROC UNIVARIATE NORMAL PLOT DATA=&INFILE.;
          VAR TARGET_WINS;

/***Proc Corr to check for initial correlations***/
proc corr data=&INFILE;
var     TARGET_WINS
      TEAM_BATTING_H
                    TEAM_BATTING_2B
                    TEAM_BATTING_3B
                    TEAM_BATTING_HR
                    TEAM_BATTING_BB
                    TEAM_BATTING_SO
                    TEAM_BASERUN_SB
                    TEAM_BASERUN_CS
                    TEAM_BATTING_HBP
                    TEAM_PITCHING_H
                    TEAM_PITCHING_HR
                    TEAM_PITCHING_BB
                    TEAM_PITCHING_SO
                    TEAM_FIELDING_E
                    TEAM_FIELDING_DP;
/***Run Model without adjustments to assess baseline fit (AR2 = 0.5116)***/
proc reg data =&INFILE.;
model   TARGET_WINS =
```

```
                TEAM_BATTING_H
                TEAM_BATTING_2B
                TEAM_BATTING_3B
                TEAM_BATTING_HR
                TEAM_BATTING_BB
                TEAM_BATTING_SO
                TEAM_BASERUN_SB
                TEAM_BASERUN_CS
                TEAM_BATTING_HBP
                TEAM_PITCHING_H
                TEAM_PITCHING_HR
                TEAM_PITCHING_BB
                TEAM_PITCHING_SO
                TEAM_FIELDING_E
                TEAM_FIELDING_DP
                /TOL VIF COLLIN;
run;
quit;
/***Proc Means and NMISS summary for missing variables***/
proc means data=&INFILE MEAN MEDIAN MODE N NMISS order=FREQ;
var    TEAM_BATTING_HBP
                TEAM_BASERUN_CS
                TEAM_FIELDING_DP
                TEAM_BASERUN_SB
                TEAM_PITCHING_SO
                TEAM_BATTING_SO;
run;
```

```
/***PROC UNIVARIATE for TEAM PITCHING H***/

PROC UNIVARIATE NORMAL PLOT DATA=&INFILE.;

        VAR TEAM_PITCHING_H;

RUN;

/***Remove TEAM_BATTING_HBP TEAM_BATTING_SO TEAM_BASERUN_CS TEAM_FIELDING_DP***/

data temp;

set &INFILE.;

drop TEAM_BATTING_HBP TEAM_BASERUN_CS;

run;

/***Check contents of temp data**/

proc contents data=temp order=varnum;

run;

/***Verify top 10 observations***/

proc print data=temp (OBS=10);

run;

/***Run Model without adjustments ***/

proc reg data =temp;

model TARGET_WINS = TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_HR
TEAM_BATTING_BB

TEAM_BATTING_SO TEAM_BASERUN_SB TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB
TEAM_PITCHING_SO TEAM_FIELDING_E

TEAM_FIELDING_DP;

run;

/**************Impute MEAN values for missing data*********************************/

data temp1;

set temp;

AVG_TEAM_BATTING_SO = 0;

if missing(TEAM_BATTING_SO)then do;
```

```
AVG_TEAM_BATTING_SO = 1;
TEAM_BATTING_SO = 735.61;
end;
AVG_TEAM_BASERUN_SB = 0;
if missing(TEAM_BASERUN_SB)then do;
AVG_TEAM_BASERUN_SB = 1;
TEAM_BASERUN_SB = 124.76;
end;
AVG_TEAM_PITCHING_SO = 0;
if missing(TEAM_PITCHING_SO)then do;
AVG_TEAM_PITCHING_SO = 1;
TEAM_PITCHING_SO = 817.73;
end;
AVG_TEAM_FIELDING_DP = 0;
if missing(TEAM_FIELDING_DP)then do;
AVG_TEAM_FIELDING_DP = 1;
TEAM_FIELDING_DP = 146.39;
end;
R_TEAM_PITCHING_H = 0;
if (TEAM_PITCHING_H>7093) then do;
R_TEAM_PITCHING_H = 1;
TEAM_PITCHING_H = 7093;
end;
proc MEANS data = temp1;
run;
/***Model NUMBER 1 with imputed MEAN values adjustment***/
proc reg data =temp1;
```

```
model TARGET_WINS = TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_HR
TEAM_BATTING_BB

TEAM_BASERUN_SB TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO
TEAM_FIELDING_E

TEAM_BATTING_SO TEAM_FIELDING_DP;

run;

/***Model NUMBER 2 with imputed MEAN values and Stepwise selection***/

proc reg data =temp1;

model TARGET_WINS = TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_HR
TEAM_BATTING_BB

TEAM_BASERUN_SB TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO
TEAM_FIELDING_E

TEAM_BATTING_SO TEAM_FIELDING_DP/selection=stepwise;

run;

/**************Impute MEDIAN values for missing data*********************************/

data temp2;

set temp;

MED_TEAM_BATTING_SO = 0;

if missing(TEAM_BATTING_SO)then do;

MED_TEAM_BATTING_SO = 1;

TEAM_BATTING_SO = 750;

end;

MED_TEAM_BASERUN_SB = 0;

if missing(TEAM_BASERUN_SB)then do;

MED_TEAM_BASERUN_SB = 1;

TEAM_BASERUN_SB = 101;

end;

MED_TEAM_PITCHING_SO = 0;

if missing(TEAM_PITCHING_SO)then do;
```

```
MED_TEAM_PITCHING_SO = 1;

TEAM_PITCHING_SO = 814;

end;

MED_TEAM_FIELDING_DP = 0;

if missing(TEAM_FIELDING_DP)then do;

MED_TEAM_FIELDING_DP = 1;

TEAM_FIELDING_DP = 149;

end;

R_TEAM_PITCHING_H = 0;

if (TEAM_PITCHING_H>7093) then do;

R_TEAM_PITCHING_H = 1;

TEAM_PITCHING_H = 7093;

end;

proc MEANS data = temp2;

run;

/***Model number 3 with imputed MEDIAN values and Stepwise selection***/

proc reg data =temp2;

model TARGET_WINS = TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_HR
TEAM_BATTING_BB

TEAM_BASERUN_SB TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO
TEAM_FIELDING_E

TEAM_BATTING_SO TEAM_FIELDING_DP/selection=stepwise;

run;

/***Model NUMBER 4 with imputed MEAN values and Adjusted R Square Selection ***/

proc reg data =temp1 outest=rsqest;

model TARGET_WINS = TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_HR
TEAM_BATTING_BB

TEAM_BASERUN_SB TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO
TEAM_FIELDING_E
```

```
TEAM_BATTING_SO TEAM_FIELDING_DP/selection=adjrsq aic bic cp;

run;

proc print data=rsqest;

proc sort data=rsqest; by _rsq_;

proc print data=rsqest;

/***Model NUMBER 4 with imputed MEAN values and Adjusted R Square Selection Stand Alone For
prininting purposes***/

proc reg data =temp1;

model TARGET_WINS = TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_HR
TEAM_BATTING_BB

TEAM_BASERUN_SB TEAM_PITCHING_H TEAM_PITCHING_BB TEAM_PITCHING_SO TEAM_FIELDING_E
TEAM_BATTING_SO

TEAM_FIELDING_DP;

run;

/***Model NUMBER 5 with imputed MEAN values and Adjusted R Square Selection WITHOUT DOUBLES BY
BATTERS

 and TEAM_FIELDING_DP***/

proc reg data =temp1;

model TARGET_WINS = TEAM_BATTING_H TEAM_BATTING_3B TEAM_BATTING_HR TEAM_BATTING_BB

TEAM_BASERUN_SB TEAM_PITCHING_H TEAM_PITCHING_BB TEAM_PITCHING_SO TEAM_FIELDING_E
TEAM_BATTING_SO

;

run;

/**********************Stand alone scoring code************************************/

/*************************************************************************************/

data scorefile;

set HW.moneyball_test;


/***Variable adjustments***/

drop TEAM_BATTING_HBP;
```

```
drop TEAM_BASERUN_CS;
AVG_TEAM_BATTING_SO = 0;
if missing(TEAM_BATTING_SO)then do;
AVG_TEAM_BATTING_SO = 1;
TEAM_BATTING_SO = 735.61;
end;
AVG_TEAM_BASERUN_SB = 0;
if missing(TEAM_BASERUN_SB)then do;
AVG_TEAM_BASERUN_SB = 1;
TEAM_BASERUN_SB = 124.76;
end;
AVG_TEAM_PITCHING_SO = 0;
if missing(TEAM_PITCHING_SO)then do;
AVG_TEAM_PITCHING_SO = 1;
TEAM_PITCHING_SO = 817.73;
end;
AVG_TEAM_FIELDING_DP = 0;
if missing(TEAM_FIELDING_DP)then do;
AVG_TEAM_FIELDING_DP = 1;
TEAM_FIELDING_DP = 146.39;
end;
R_TEAM_PITCHING_H = 0;
if (TEAM_PITCHING_H>7093) then do;
R_TEAM_PITCHING_H = 1;
TEAM_PITCHING_H = 7093;
end;
P_TARGET_WINS = 15.20547
+ 0.03515 * TEAM_BATTING_H
```

```
+ 0.09555 * TEAM_BATTING_3B
+ 0.04894 *       TEAM_BATTING_HR
+ 0.01731 * TEAM_BATTING_BB
+ 0.04143 * TEAM_BASERUN_SB
+ 0.00328 * TEAM_PITCHING_H
- 0.01045 * TEAM_PITCHING_BB
+ 0.00269 * TEAM_PITCHING_SO
- 0.02926 * TEAM_FIELDING_E
- 0.00666 * TEAM_BATTING_SO;
keep INDEX;
keep P_TARGET_WINS;
run;
proc print data=SCOREFILE;
run;
/*******Save to Folder********/
libname NOELIB "/home/noeflores20160/sasuser.v94";
data NOELIB.NOEFILE;
set SCOREFILE;
run;
proc print data=NOELIB.NOEFILE;
run;
proc print data=SCOREFILE;
var INDEX P_TARGET_WINS;
run;
/*****************************************************************/
```

```
/***BINGOOOOOOOOOO BOOOOOOOONUS*****/


/***Proc GLM***/

proc glm data =temp1;

model TARGET_WINS = TEAM_BATTING_H TEAM_BATTING_3B TEAM_BATTING_HR TEAM_BATTING_BB

TEAM_BASERUN_SB TEAM_PITCHING_H TEAM_PITCHING_BB TEAM_PITCHING_SO TEAM_FIELDING_E
TEAM_BATTING_SO

;

run;

/***Proc Genmod***/

proc genmod data =temp1;

model TARGET_WINS = TEAM_BATTING_H TEAM_BATTING_3B TEAM_BATTING_HR TEAM_BATTING_BB

TEAM_BASERUN_SB TEAM_PITCHING_H TEAM_PITCHING_BB TEAM_PITCHING_SO TEAM_FIELDING_E
TEAM_BATTING_SO

;

run;

quit;
```

References

(1) Hoffmann, J. P., (2004). *Generalized Linear Models: An Applied Approach*. Boston, MA: Pearson Education, Inc.

(2) Montgomery, D. C., Peck, E. A., Vinning, G. G., (2012). *Introduction to Linear Regression Analysis* Hoboken, NJ: Wiley.

(3) Cody, R. (2011). *SAS: Statistics by Example*. Carey, NC: SAS Institute Inc.

(4) Evaluating forecast Accuracy. https://www.otexts.org/fpp/2/5 (accessed April 14, 2017)