Unit 01: "Moneyball Baseball Problem"

Intro:

The assignment detailed below is an analysis on baseball statistics with basic statistics and predictive analytics by way of linear regression. The data is aggregated such that each record represents the results of a season for a given team. Given that the league itself has changed quite considerably since 1871, an adjustment had to be made to the stats so that seasons with less games could be compared to the current 162 game schedule. The quality of the data may be questionable for some seasons so the analysis will include exploration and transformations.

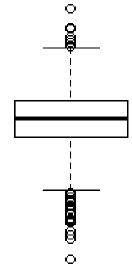
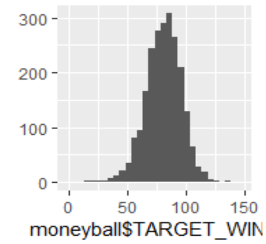
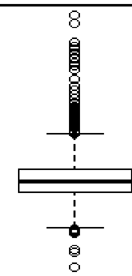
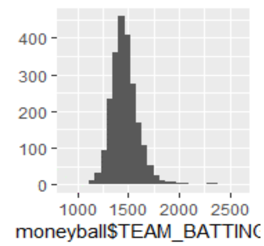
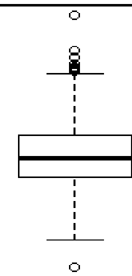
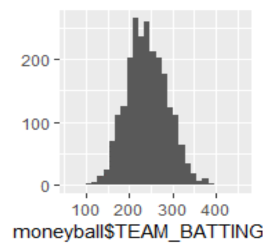
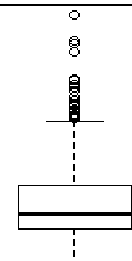
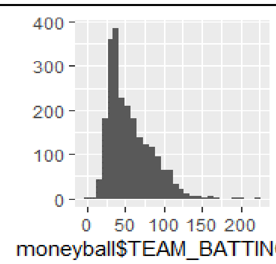
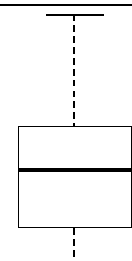
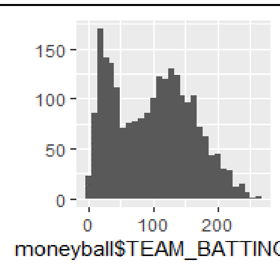
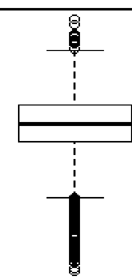
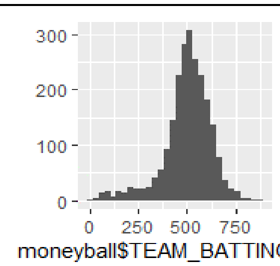
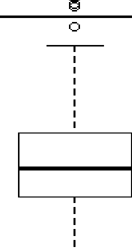
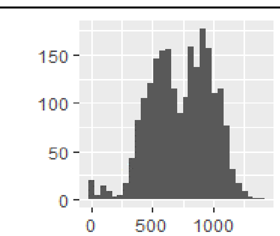
Data:

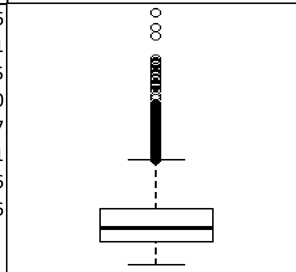
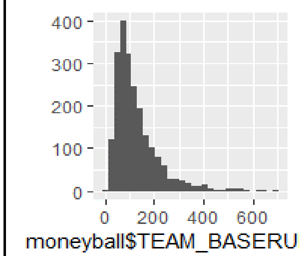
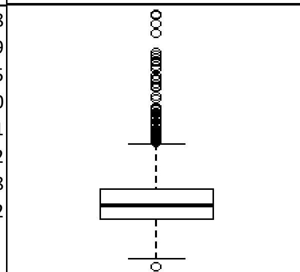
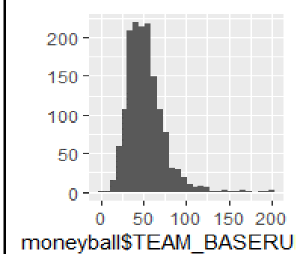
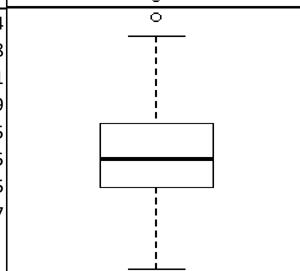
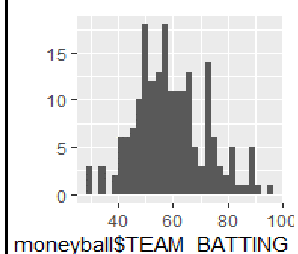
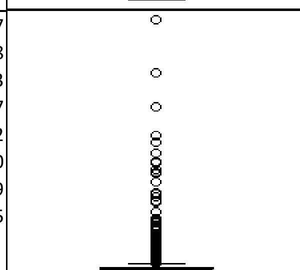
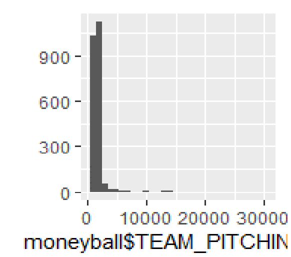
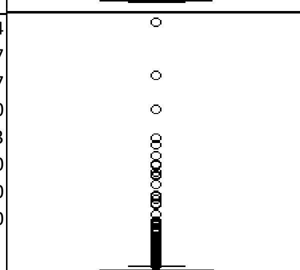
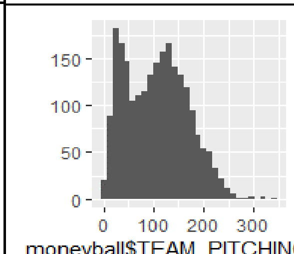
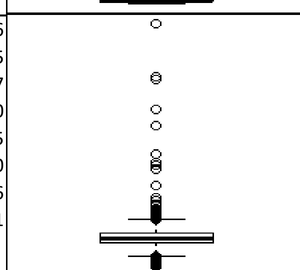
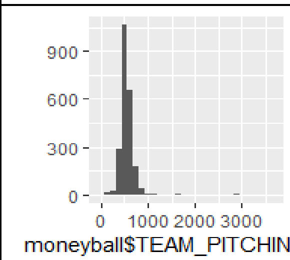
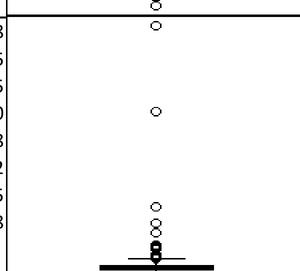
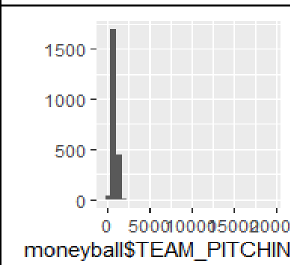
16 variables, excluding the index variable covering three aspects of baseball; hitting, pitching, and fielding. We are only using the variables given in a data set and not supplementing with other data which means certain things will be disregarded that would have an effect, such as the year the season took place. That could be important in comparing the offensive performance of teams and its affect on wins relative to how other teams are performing such as comparing baseball from the early 1900's to that of the late 90's and early 2000's.

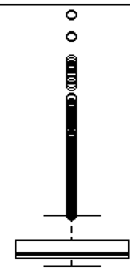
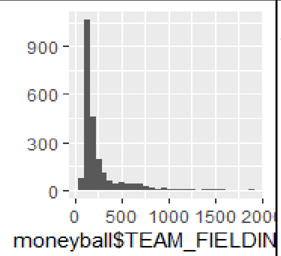
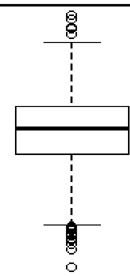
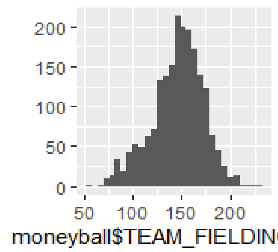
Contents

Bingo Bonus Attempt: Models with single imputations vs models with multiple imputations (pooled summary).
Alternative correlation plots.

Data Exploration

Variable	Basic Stats (Before Transform)	Box Plot	Distribution	Notes
Name: TARGET_WINS Def: Target variable, wins for a given season.	**Mean: 80.79086116** Median: 82 Std Deviation: 15.75215248 Min: 0 Max: 146 NAs: 0 1st quartile: 71 3rd quartile: 92 # Of Outliers: 34			Normal distribution but outliers require fixing. As a matter of practice, we'll set any below the 1st percentile or above the 99th percentile to NA and impute. In general, we will want the results of our linear model to have a mean close to 80.
Name: TEAM_BATTING_H Def: Base Hits by batters (1B,2B,3B,HR)	Mean: 1469.269772 Median: 1454 Std Deviation: 144.5911954 Min: 891 Max: 2554 NAs: 0 1st quartile: 1383 3rd quartile: 1537.25			
Name: TEAM_BATTING_2B Def: Doubles by batters (2B)	Mean: 241.2469244 Median: 238 Std Deviation: 46.8014146 Min: 69 Max: 458 NAs: 0 1st quartile: 208 3rd quartile: 273			
Name: TEAM_BATTING_3B Def: Triples by batters (3B)	Mean: 55.25 Median: 47 Std Deviation: 27.938557 Min: 0 Max: 223 NAs: 0 1st quartile: 34 3rd quartile: 72			Skewed to the left and rapidly declining
Name: TEAM_BATTING_HR Def: Homeruns by batters (4B)	Mean: 99.61203866 Median: 102 Std Deviation: 60.54687197 Min: 0 Max: 264 NAs: 0 1st quartile: 42 3rd quartile: 147			Two local maxima
Name: TEAM_BATTING_BB Def: Walks by batters	Mean: 501.5588752 Median: 512 Std Deviation: 122.6708615 Min: 0 Max: 878 NAs: 0 1st quartile: 451 3rd quartile: 580			
Name: TEAM_BATTING_SO Def: Strikeouts by batters	Mean: 735.6053358 Median: 750 Std Deviation: 248.5264177 Min: 0 Max: 1399 NAs: 0 1st quartile: 102 3rd quartile: 930			

Variable	Basic Stats (Before Transform)	Box Plot	Distribution	Notes
Name: TEAM_BASERUN_SB Def: Stolen bases	Mean: 124.7617716 Median: 101 Std Deviation: 87.79116605 Min: 0 Max: 697 NAs: 131 1st quartile: 66 3rd quartile: 156			Poisson distribution, candidate for transformation
Name: TEAM_BASERUN_CS Def: Caught stealing	Mean: 52.80385638 Median: 49 Std Deviation: 22.95633765 Min: 0 Max: 201 NAs: 772 1st quartile: 38 3rd quartile: 62			Normal distribution, high kurtosis
Name: TEAM_BATTING_HBP Def: Batters hit by pitch (get a free base)	Mean: 59.35602094 Median: 58 Std Deviation: 12.96712251 Min: 29 Max: 95 NAs: 2085 1st quartile: 50.5 3rd quartile: 67			
Name: TEAM_PITCHING_H Def: Hits allowed	Mean: 1779.210457 Median: 1518 Std Deviation: 1406.84293 Min: 1137 Max: 30132 NAs: 0 1st quartile: 1419 3rd quartile: 1682.5			Poisson distribution with some outliers that skew the plot.
Name: TEAM_PITCHING_HR Def: Homeruns allowed	Mean: 105.698594 Median: 107 Std Deviation: 61.29874687 Min: 0 Max: 343 NAs: 0 1st quartile: 50 3rd quartile: 150			
Name: TEAM_PITCHING_BB Def: Walks allowed	Mean: 553.0079086 Median: 536.5 Std Deviation: 166.3573617 Min: 0 Max: 3645 NAs: 0 1st quartile: 476 3rd quartile: 611			Normal distribution, high kurtosis
Name: TEAM_PITCHING_SO Def: Strikeouts by pitchers	Mean: 817.7304508 Median: 813.5 Std Deviation: 553.0850315 Min: 0 Max: 19278 NAs: 102 1st quartile: 615 3rd quartile: 968			Poisson distribution with some outliers that skew the plot.

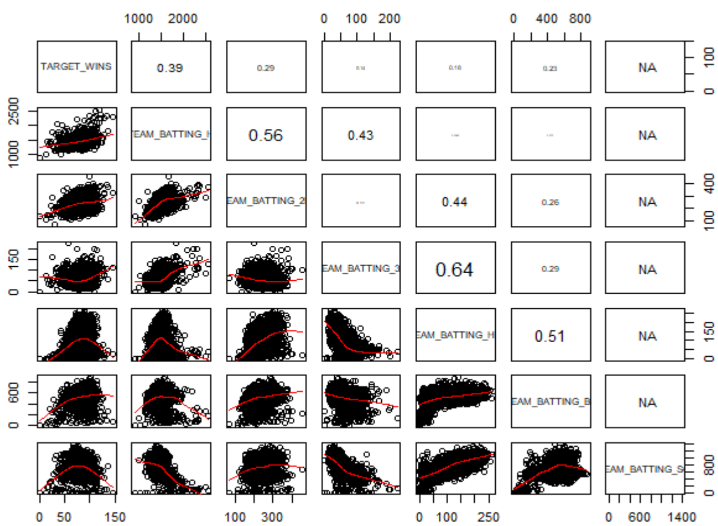
Variable	Basic Stats (Before Transform)	Box Plot	Distribution	Notes
Name: TEAM_FIELDING_E Def: Errors	Mean: 246.4806678 Median: 159 Std Deviation: 227.7709724 Min: 65 Max: 1898 NAs: 0 1st quartile: 127 3rd quartile: 249.25			Poisson distribution, candidate for transformation
Name: TEAM_FIELDING_DP Def: Double Plays	Mean: 146.3879397 Median: 149 Std Deviation: 26.22638525 Min: 52 Max: 228 NAs: 286 1st quartile: 131 3rd quartile: 164			

Correlations

We want to consider how the relationship between variables both between each independent variables and the target variable but also the correlation between the independent variables and other independent variables. We are looking to avoid a situation where collinearity has an adverse affect on the predictive power of our model.

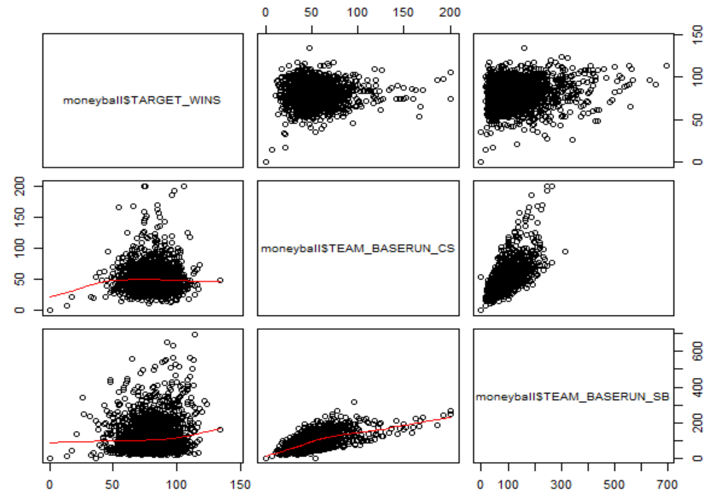
Scatterplot Matrix

Batting



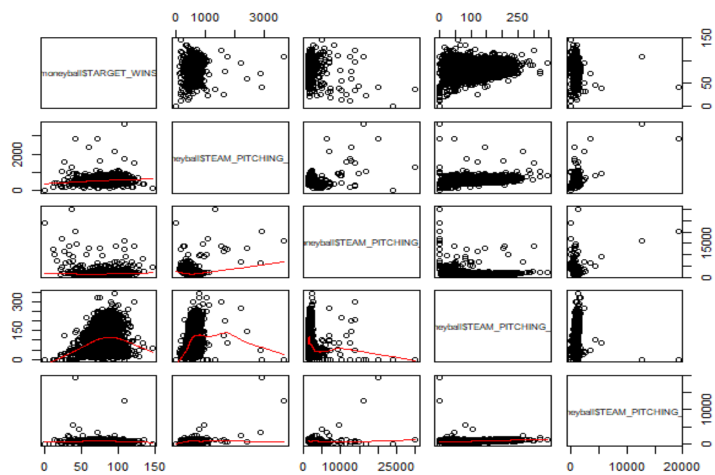
The scatterplot matrix is to read in such as way that the label means on a square signifies the X-axis for every other square in that row and the Y-axis for every square in that column. It's interesting to note here the relationship of triples and homeruns to wins is not as linear as it is for singles and doubles. In the case of homeruns, this may be explained by looking at the relationship between homeruns and strikeouts. While also looking at the base on balls (walks), it may appear on first glance that the conclusion would be that it is more advantageous to look at more pitches rather than what a baseball fan might call "swinging for the fences" (going for the homerun) on every pitch.

Base-running



When considering base-running, there appears to be a weak linear relationship between stolen bases and wins. A new variable may have more predictive value if it were to capture the stolen base percentage, getting caught stealing costs a team a base-runner and potential RBI.

Pitching



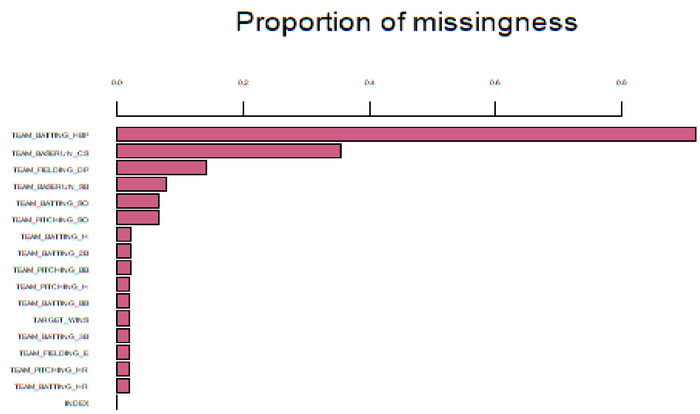
Data Preparation

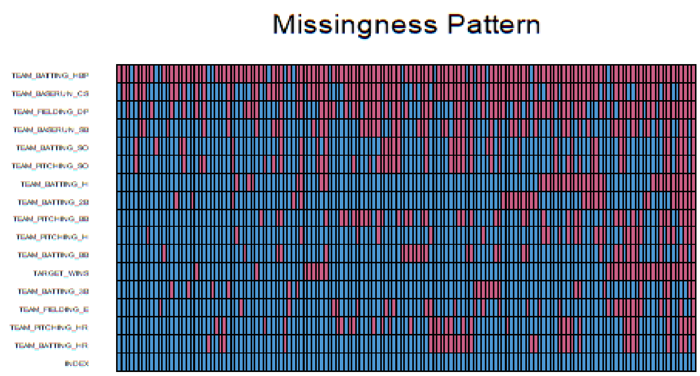
1st step: Identify outliers and set to NA

Outliers can be helpful at times but extreme outliers in a linear regression analysis can be harmful in that they will "pull" the line towards that value. We will take on the general practice of setting those values below the 1st percentile and above the 99th percentile to NA in order to later impute those values using the MICE package.

Team Pitching Hits and Team Fielding Errors are two metrics affected by this but this is helpful, as noted in the distribution section of the data exploration with the histograms heavily skewed to the left because of some extreme outliers. Consider for example that while 98% of the time, a team surrenders less than 10,000 hits in a season there was still a record of a team giving up over 30,000 hits in a season. This same team managed to win 36 games that season while hitting a grand total of 0 homeruns. There are clearly issues with this particular record.

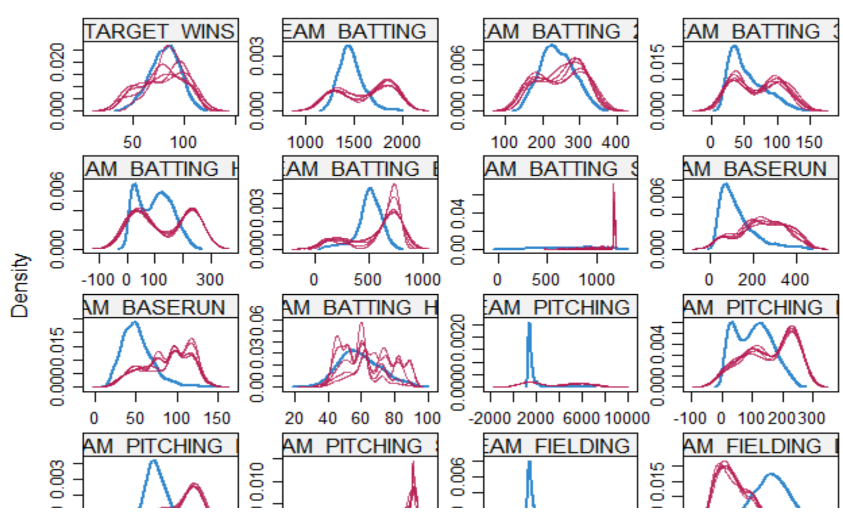
2nd step: Impute missing values

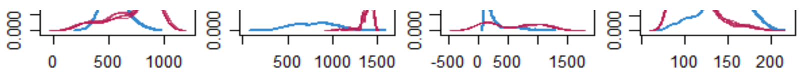




The output above is simply to show how much missing data we have. We can observe that for the majority of the data we have only a small proportion to impute but for TEAM_BATTING_HBP, we have a significant amount of educated guesses to have to make. This analysis will make use of the MICE package and method of imputation will be pmm. PMM stands for Predictive Mean Matching and was chosen on the premise that the imputed values will be plausible according to the other values for the variable in the data.

Before we impute, we will create flags for each of these original variables so that we have a record of which were imputed.
MICE produces a density plot to graph the various imputations over the original plot in blue, but it's much clearer to actually compare the summaries.





When comparing the summaries before and after imputation, it's easier to observe how certain things were fixed such as the previous max of hits allowed went from 19,278 to 1,464. **This document will have single-imputations (using the average) as well as multiple imputations in order to compare the two on the training data set.**

3rd step: Transform the data and add additional variables

Transformations:

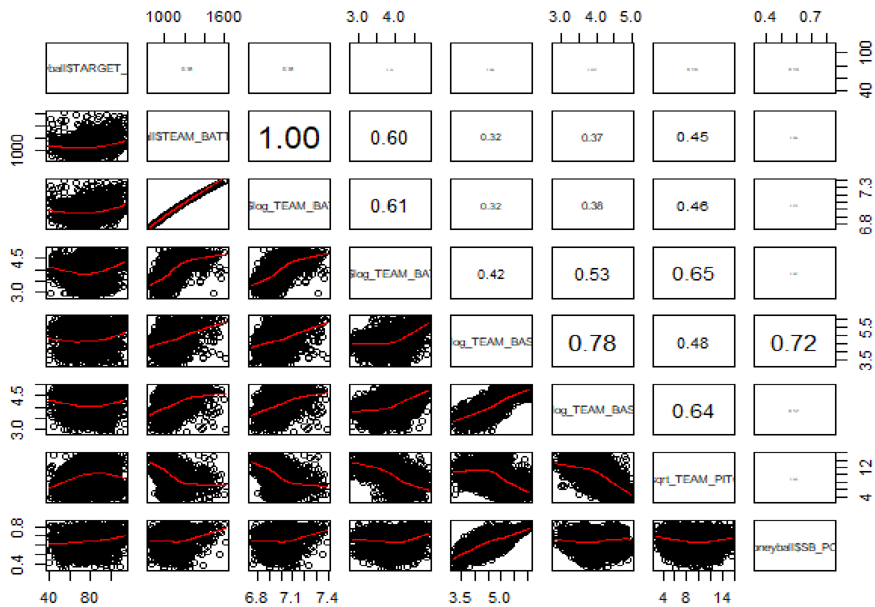
* TEAM_BATTING_1B fixed to equal hits - triples, doubles, and homeruns
* Capping fielding errors at 500, the max was 1225 before.

New variables:

* Logs of singles, triples, stolen bases, and caught stealing metrics
* Square root of homeruns
* Stolen base percentage
* Ratio of walks+hits to strikeouts, in lieu of on base percentage.

Subsetting:

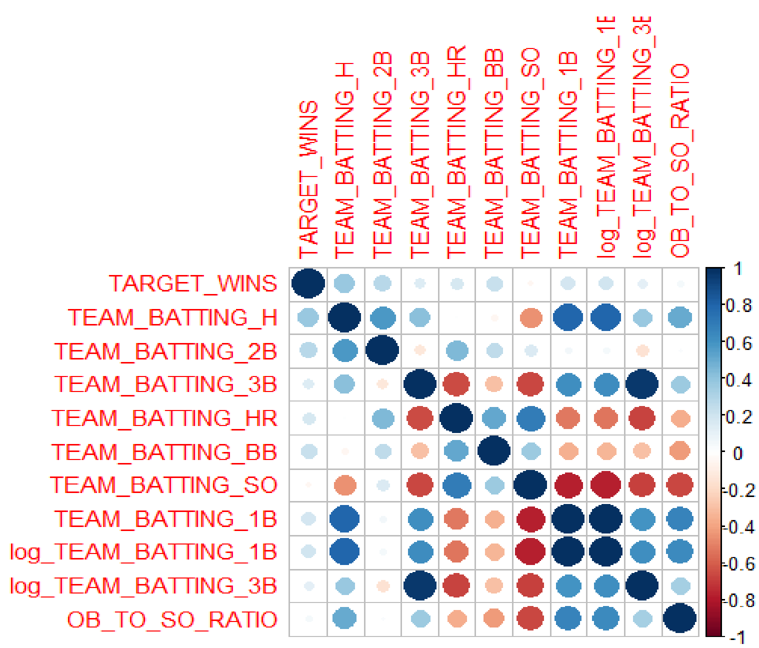
* Only considering those seasons between 21 and 120 wins
* Only considering those seasons where hits were less than 2000.



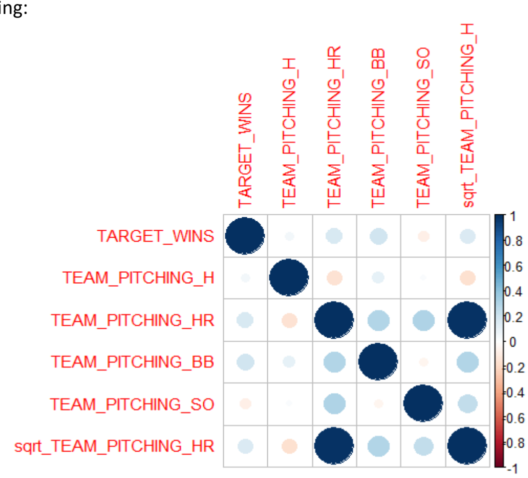
From a quick glance of the scatterplot matrix with TARGET_WINS and these new variables suggests that the transforms may potentially be better performing than the original variables, specifically stolen base percentage as compared to pure stolen bases but it's tough to compare just on the graphs because the original had rows of unavailable data which resulted in not having a correlation value in the grid.

We make use of a simpler corrplot for clarity:

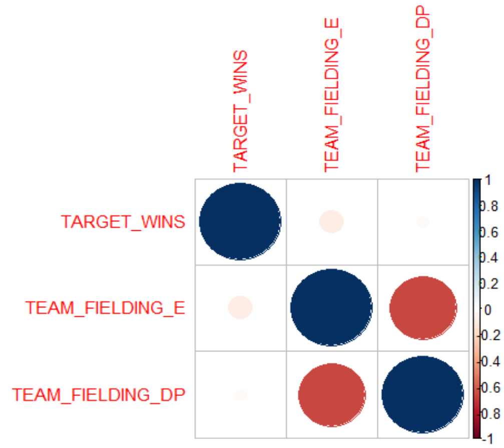
Batting:



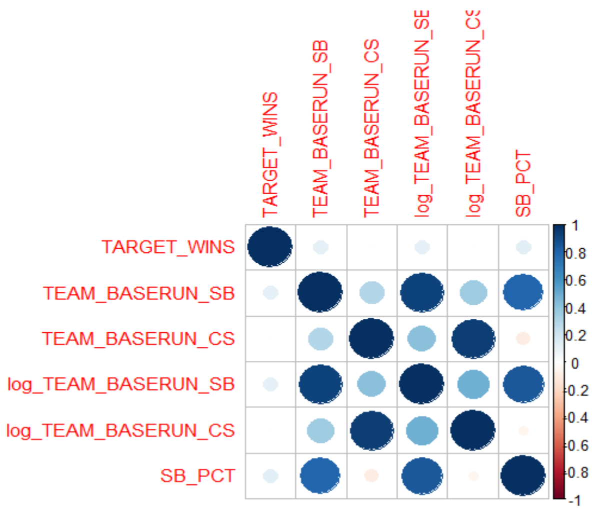
Pitching:



Fielding:



Baserunning:



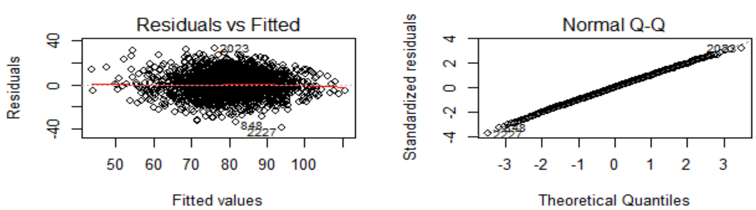
It's interesting to note the positive correlation between walks given up pitching to target wins and that doubles have a stronger correlation with target wins than homeruns do. Most things make sense, such as walks being very favorable and errors being costly.

Build Models

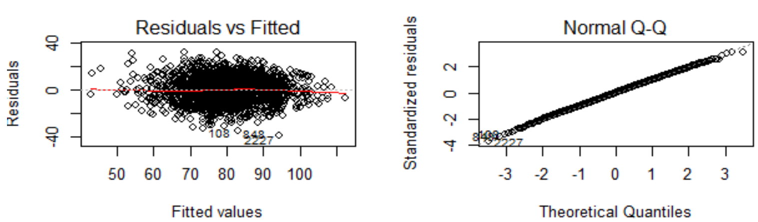
In total between the 8 models were built, 7 with the data imputed with the average and 1 with the data imputed with the MICE package. Unfortunately, the MICE package results based on adjusted R-squared were very poor compared to the results from those with imputations based simply on the mean. PMM and CART methods were used to try to address this issue but to no avail in this particular case.

Model	Imputation	Description	Adjusted R2	AIC	MSE	Notes
1	Avg	Stepwise Approach, forward + back	0.4011	15383.4	116.7048	Performed very well, close contender
2	Avg	All subsets regression	0.3624	15497.47	124.8407	
3	Avg	No transformed variables and excluding HBP	0.3831	15431.93	120.734	
4	Avg	Full model excluded NA flags	0.4023	15379.98	116.2766	
5	Avg	Only those variables from full models with small p values	0.3896	15414.47	119.2212	
6	Avg	Full model including NA flags	**0.429**	**15284.51**	**111.2364**	Best model by every metric
7	Avg	Stepwise Approach, forward + back, including NA flags	0.4275	15289.88	111.5327	2nd best model, chosen because of slightly wider range of fitted
8	MICE	Full model excluding NA flags	0.3319148	#DIV/0!	n/a	Average AIC, highest Adjusted R-Squared

Model 6



Model 7



For the purpose of double-checking visually, a scatterplot of residuals vs fitted and a Q-Q plot were run on the two best performing models on the training data set. The results are encouraging for both.

Select Models

Model 7 was ultimately chosen because of it ranked very favorably in all three metrics being considered and it checked out visually.

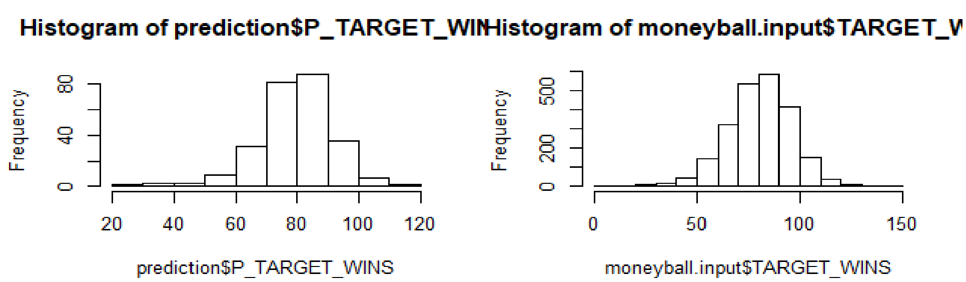
TARGET_WINS -116.7+ (0.2685*TEAM_BATTING_3B)+
(0.3507*TEAM_BATTING_HR)+
(0.03007*TEAM_BATTING_BB)+
(-0.04358*TEAM_BATTING_SO)+
(0.1007*TEAM_BASERUN_SB)+
(0.1018*TEAM_BASERUN_CS)+
(-0.224*TEAM_PITCHING_HR)+
(0.02523*TEAM_PITCHING_SO)+
(-0.1357*TEAM_FIELDING_E)+
(-0.1018*TEAM_FIELDING_DP)+
(35.14*log_TEAM_BATTING_1B)+
(-3.678*log_TEAM_BATTING_3B)+
(-8.478*SB_PCT)+
(-10.15*log_TEAM_BASERUN_CS)+
(8.711*TEAM_BATTING_SO_NATRUE)+
(21.53*TEAM_BASERUN_SB_NATRUE)+
(4.739*TEAM_BASERUN_CS_NATRUE)+
(4.422*TEAM_BATTING_HBP_NATRUE)+
(8.055*TEAM_FIELDING_DP_NATRUE)

The negative intercept may be counterintuitive but you have to consider that it would be impossible for the values to be 0 for the dependent variables such that a negative 116 win season is predicted. There were other models considered that actually did predict negative wins and wins over 162 in the test data set, some of which had positive intercepts and coefficients that were more intuitive.

This linear equation was run on the test data set with the following results from the fitted values.
The minimum number of wins predicted was initially 11 but there was a cap on the bottom such that anything less than 20 gets set to 20.

Min	1st quartile	Median	Mean	3rd Quartile	Max
20	72.8	80.37	79.49	87.07	113.07

We would like to have seen the mean be slightly higher and closer to the median but it's still close in a relative way.
Otherwise the results of the distribution of the fitted values look good (on the left). They are normally distributed much like the input on the right.



Conclusion

This exercise was very good at reinforcing the idea that not one model fits all and that it's very important to come up with multiple models and to test them properly. There were many times a simple and easy to understand model performed well on the training data set and not the test data set. Looking at the summary of the test data input, I could see that the distributions had some important differences. When you build a model using training data with a certain range for each variable, it could be quite disruptive when the test data set has more extreme outliers for instance. Also, it was a reminder that when using variables with some degree of colinearity, there will be times when that means one of the variables which should have a positive effect may unintuitively have a negative effect. However, the times that models were made on the premise of optimizing on the basis of having small VIF values, they actually managed to perform very poorly. Those models were overwritten in order to meet a minimum threshold on the metrics used such as adjusted r-squared and AIC.