Homework #1: Baseball Analysis

Data 621 Business Analytics and Data Mining

Aadi Kalloo, Nathan Lim, Asher Meyers, Daniel Smilowitz, Logan Thomson

Due June 19, 2016

Data Exploration

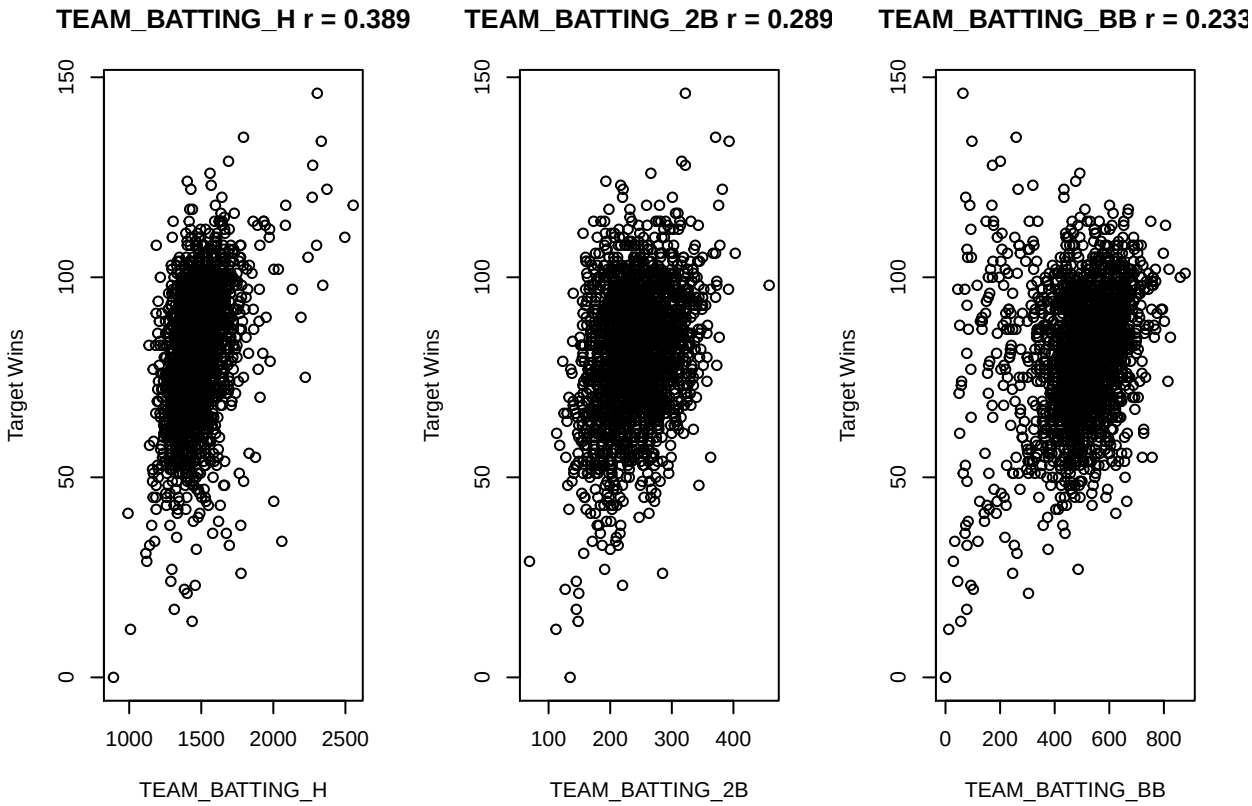
The data analyzed in this report includes 2276 professional baseball teams for the years 1871-2006. In total, 16 variables were present in the data provided. Included below is a summary of descriptive statistics, correlations to wins, and the number of missing values for each variable in the provided data set:

Table 1

	VAR_NAME	MEAN	MEDIAN	CORRELATION TO WINS (r)	NUM_MISSING
2	TARGET_WINS	80.79086	82.0	NA	NA
1	TEAM_BASERUN_CS	52.80386	49.0	0.0224041	772
21	TEAM_BASERUN_SB	124.76177	101.0	0.1351389	131
3	TEAM_BATTING_2B	241.24692	238.0	0.2891036	0
4	TEAM_BATTING_3B	55.25000	47.0	0.1426084	0
5	TEAM_BATTING_BB	501.55888	512.0	0.2325599	0
6	TEAM_BATTING_H	1469.26977	1454.0	0.3887675	0
7	TEAM_BATTING_HBP	59.35602	58.0	0.0735042	2085
8	TEAM_BATTING_HR	99.61204	102.0	0.1761532	0
9	TEAM_BATTING_SO	735.60534	750.0	-0.0317507	102
10	TEAM_FIELDING_DP	146.38794	149.0	-0.0348506	286
11	TEAM_FIELDING_E	246.48067	159.0	-0.1764848	0
12	TEAM_PITCHING_BB	553.00791	536.5	0.1241745	0
13	TEAM_PITCHING_H	1779.21046	1518.0	-0.1099371	0
14	TEAM_PITCHING_HR	105.69859	107.0	0.1890137	0
15	TEAM_PITCHING_SO	817.73045	813.5	-0.0784361	102

It can be seen that there are missing values in 6 of the variables in the data set, and these missing values range from approximately 5-92% of the data provided for their respective variables. However, in only two exceptions do the missing data account for more than 11% of the missing data.

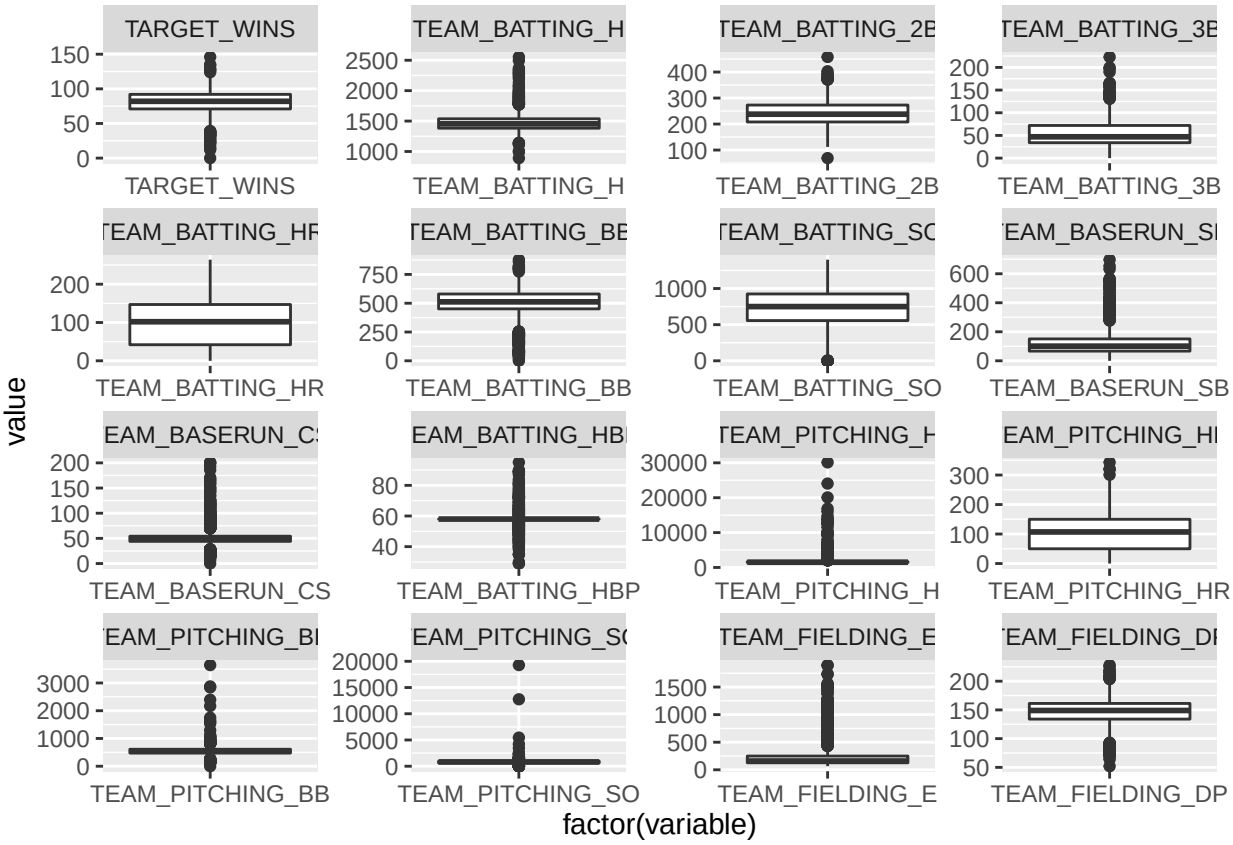
Below are graphs that show the relationship to *Target Wins* for the three variables with the highest correlation coefficient:



As can be seen from Table 1, there are few variables that have any particularly strong correlation with `TARGET_WINS`.The full array of scatterplots representing correlations between `TARGET_WINS` and other variables may be found in Appendix A.

The distribution of values and outliers is also of significant importance in understanding the baseball data set. Here it can be seen that many variables have a skewed distribution:

```
## No id variables; using all as measure variables
```



In summary, the baseball data set provided includes many variables with a skewed distribution, few variables that correlate well with `TARGET_WINS`, and several variables that have missing data and should either require data imputation or should be excluded. The following sections serve to review these issues and go on to create a working regression model that can predict `TARGET_WINS`.

Data Preparation

It was determined that the *Hits By Pitch* variable had too many missing values to be useful for regression, and thus this variable was excluded from the model building process. As shown in Table 1 above, there are several variables that have missing values. The attempted solution to this problem involved imputation using the median for each variable in the data set. A summary of the data is shown here again for inspection and confirmation of similarity between the old and new data sets:

Missing Values Imputed With Median

	VAR_NAME	MEAN	MEDIAN	CORRELATION TO WINS (r)	NUM_MISSING
2	TEAM_BATTING_H	1469.26977	1454.0	NA	NA
1	TEAM_BASERUN_CS	51.51362	49.0	0.0159598	0
21	TEAM_BASERUN_SB	123.39411	101.0	0.1236109	0
3	TEAM_BATTING_2B	241.24692	238.0	0.2891036	0
4	TEAM_BATTING_3B	55.25000	47.0	0.1426084	0
5	TEAM_BATTING_BB	501.55888	512.0	0.2325599	0
6	TEAM_BATTING_HBP	58.11380	58.0	0.0165164	0
7	TEAM_BATTING_HR	99.61204	102.0	0.1761532	0
8	TEAM_BATTING_SO	736.25044	750.0	-0.0305814	0
9	TEAM_FIELDING_DP	146.71617	149.0	-0.0300863	0
10	TEAM_FIELDING_E	246.48067	159.0	-0.1764848	0
11	TEAM_PITCHING_BB	553.00791	536.5	0.1241745	0
12	TEAM_PITCHING_H	1779.21046	1518.0	-0.1099371	0
13	TEAM_PITCHING_HR	105.69859	107.0	0.1890137	0
14	TEAM_PITCHING_SO	817.54086	813.5	-0.0757997	0

The dataset contains 17 columns - an index column (INDEX), a response column (TARGET_WINS) and 15 predictor columns. There are 2,276 observations - but there are many missing values for many of the predictors.

Two predictors in particular stand out:

	Predictor Name	Description	Impact	% Missing	r with Response	p-Value
a	TEAM_BATTING_HBP	Batters hit by pitch (free base)	Positive	91.6%	0.07	0.31
b	TEAM_BASERUN_CS	Strikeouts by batters	Negative	33.9%	0.02	0.39

Including these predictors in our dataset would mean that we would either have to a) forgo a significant chunk of our data (34% or 92%) or b) impute a large number of data points. Their correlation coefficients with the response are less than an absolute value of 7%; the p values of a simple one variable linear regression using them and the response yields models of no statistical significance (i.e. $p > 0.05$). Thus, it seems safe to exclude these predictors from our models. This way, we avoid the twin pitfalls of mass exclusion and imputation.

Further exclusions to the data were made:

Exclusion	Explanation
INDEX == 1347	This row had a suspicious set of zero entries
TEAM_BATTING_BB == 0	Anomalously low walk count (expected occurences of a zero value for this predictor are zero)
TEAM_BATTING_SO	Outside of recognized records link
TEAM_BATTING_HR	Outside of recognized records link

It should be noted that the records excluded from the first two rows of the table above are the same exact points (which would technically make the second exclusion redundant...). That suggests that for whatever reason, strikeouts were not recorded for those rows, but were marked as zero. Those two predictors have the same number of NA values, 102, suggesting their recording method was linked somehow.

Model Creation

Model Summary Table

Model #	# of Predictors	Adj. R^2	F-Statistic	P-Value	Residual Standard Error	Degrees of Freedom
1	11	0.31	95	2.2e-16	13.07	2264
2	8	0.22	82	2.2e-16	13.82	2266
3	7	0.35	157	2.2e-16	11.07	1993
4	4	0.28	225	2.2e-16	11.78	2271
5	8	0.31	125	2.2e-16	13.15	2267

Model 1: Simple Full Linear Regression, With Removal of Non-Significant Predictors

Description: Missing values were replaced with the median values from the associated predictor, to retain all data points for making a regression; a linear regression was fit to all predictors; All non-significant predictors (p<.05) were removed sequentially. The final iteration of this regression model is shown here:

	Estimate	Std. Error	t value	Pr(>|t|)
TEAM_BATTING_H	0.04909	0.00367	13.38	2.469e-39
TEAM_BATTING_2B	-0.02137	0.009163	-2.333	0.01975
TEAM_BATTING_3B	0.06658	0.01662	4.005	6.4e-05
TEAM_BATTING_HR	0.0674	0.009632	6.998	3.399e-12
TEAM_BATTING_BB	0.01155	0.003375	3.421	0.0006342
TEAM_BATTING_SO	-0.008521	0.002453	-3.474	0.0005227
TEAM_BASERUN_SB	0.02492	0.004209	5.92	3.699e-09
TEAM_PITCHING_H	-0.000777	0.0003209	-2.421	0.01555
TEAM_PITCHING_SO	0.002966	0.0006719	4.415	1.059e-05
TEAM_FIELDING_E	-0.01901	0.002392	-7.948	2.972e-15
TEAM_FIELDING_DP	-0.1218	0.01293	-9.419	1.079e-20
(Intercept)	22.34	5.234	4.269	2.043e-05

Table 7: Fitting linear 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_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP

Observations	Residual Std. Error	R^2	Adjusted R^2
2276	13.07	0.3151	0.3117

	Estimate	Std. Error	t value	Pr(>|t|)
TEAM_BATTING_H	0.04909	0.00367	13.38	2.469e-39
TEAM_BATTING_2B	-0.02137	0.009163	-2.333	0.01975
TEAM_BATTING_3B	0.06658	0.01662	4.005	6.4e-05
TEAM_BATTING_HR	0.0674	0.009632	6.998	3.399e-12
TEAM_BATTING_BB	0.01155	0.003375	3.421	0.0006342

	Estimate	Std. Error	t value	Pr(>|t|)
TEAM_BATTING_SO	-0.008521	0.002453	-3.474	0.0005227
TEAM_BASERUN_SB	0.02492	0.004209	5.92	3.699e-09
TEAM_PITCHING_H	-0.000777	0.0003209	-2.421	0.01555
TEAM_PITCHING_SO	0.002966	0.0006719	4.415	1.059e-05
TEAM_FIELDING_E	-0.01901	0.002392	-7.948	2.972e-15
TEAM_FIELDING_DP	-0.1218	0.01293	-9.419	1.079e-20
(Intercept)	22.34	5.234	4.269	2.043e-05

Table 9: Fitting linear 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_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP

Observations	Residual Std. Error	R^2	Adjusted R^2
2276	13.07	0.3151	0.3117

Model 2: SLR Bounded by Recent MLB Data (1962-)

Description: Missing values were replaced with the median values from the associated predictor. Data was compared against records from 1962 and onwards, aka the MLB dataset, and data outside the bounds of that external dataset were replaced with the medians of the associated predictor. Eg, if one of the records in our dataset had more home runs hit than in all of the MLB dataset, then that home run data point was replaced with the median home run figure in our dataset.

Then, a linear regression was fitted to all predictors; predictors were removed in order of significance, to obtain a model with a higher f-statistic.

	Estimate	Std. Error	t value	Pr(>|t|)
TEAM_BATTING_H	0.03724	0.003922	9.496	5.32e-21
TEAM_BATTING_2B	0.01864	0.008774	2.124	0.03376
TEAM_BATTING_BB	0.03099	0.003743	8.28	2.084e-16
TEAM_BATTING_SO	-0.01306	0.00211	-6.191	7.083e-10
TEAM_BASERUN_SB	0.03816	0.0053	7.201	8.087e-13
TEAM_PITCHING_HR	0.05392	0.008428	6.397	1.92e-10
TEAM_FIELDING_E	-0.01014	0.001499	-6.764	1.706e-11
TEAM_FIELDING_DP	-0.1141	0.01365	-8.357	1.107e-16
(Intercept)	25.55	5.856	4.362	1.344e-05

Table 11: Fitting linear model: TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_PITCHING_HR + TEAM_FIELDING_E + TEAM_FIELDING_DP

Observations	Residual Std. Error	R^2	Adjusted R^2
2275	13.82	0.2248	0.2221

Model 3: Data Bounded by 1880-2015 Records

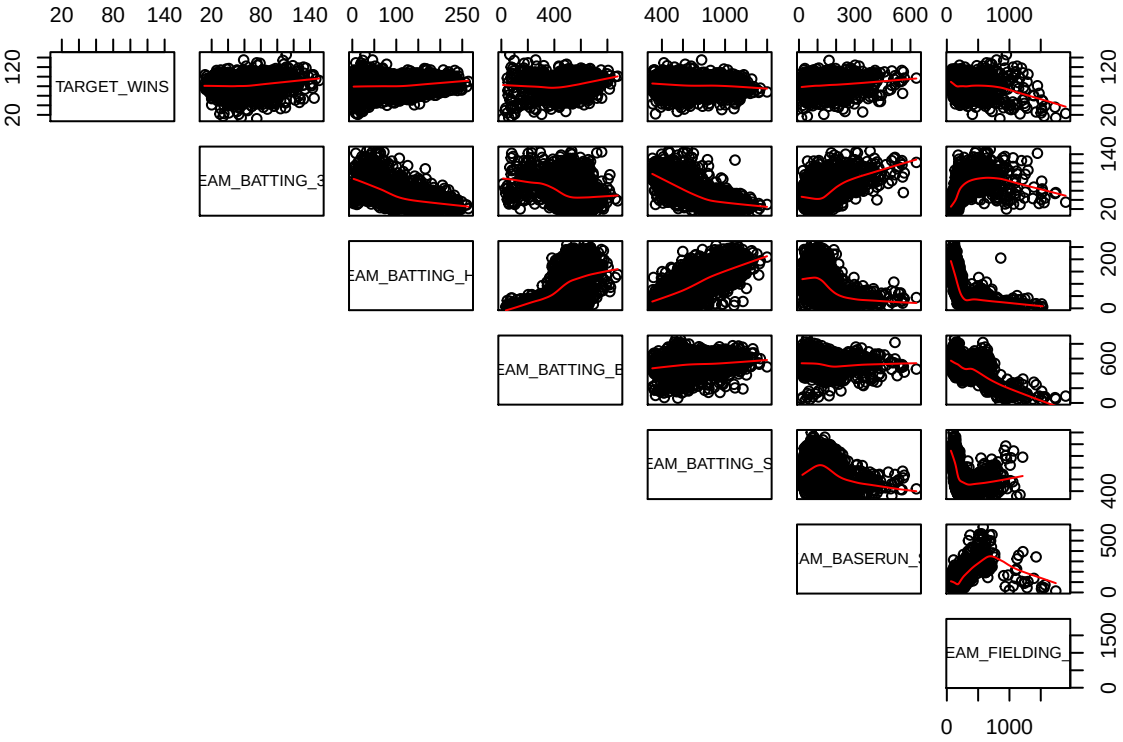
Three predictors were removed: Predictors removed, due to high NA count: TEAM_BATTING_HBP and TEAM_BASERUN_CS Predictor removed, due to lack of relevance: TEAM_FIELDING_DP

Bounds for predictors were set by minimum and maximum all-time MLB records, with citations shown in the code. Where our records were outside the bounds of these external records, they were replaced with NA values. Predictors were removed sequentially removed, in order of significance; multiple predictors were created and tested, in the hopes of improving fit statistics, and one proved useful - the number of 1st base hits. This predictor was then added to the model.

	Estimate	Std. Error	t value	Pr(>|t|)
TEAM_BATTING_3B	0.1804	0.01727	10.45	6.657e-25
TEAM_BATTING_HR	0.1149	0.007506	15.31	4.252e-50
TEAM_BATTING_BB	0.02858	0.003096	9.231	6.64e-20
TEAM_BATTING_SO	-0.017	0.002323	-7.32	3.592e-13
TEAM_BASERUN_SB	0.07938	0.004722	16.81	2.048e-59
TEAM_FIELDING_E	-0.0736	0.003888	-18.93	1.226e-73
TEAM_BATTING_1B	0.02327	0.004018	5.79	8.143e-09
(Intercept)	36.71	5.549	6.616	4.739e-11

Table 13: Fitting linear model: TARGET_WINS ~ TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_FIELDING_E + TEAM_BATTING_1B

Observations	Residual Std. Error	R^2	Adjusted R^2
2001	11.07	0.3557	0.3535



Model 4: Using only significant predictors from model using all variables

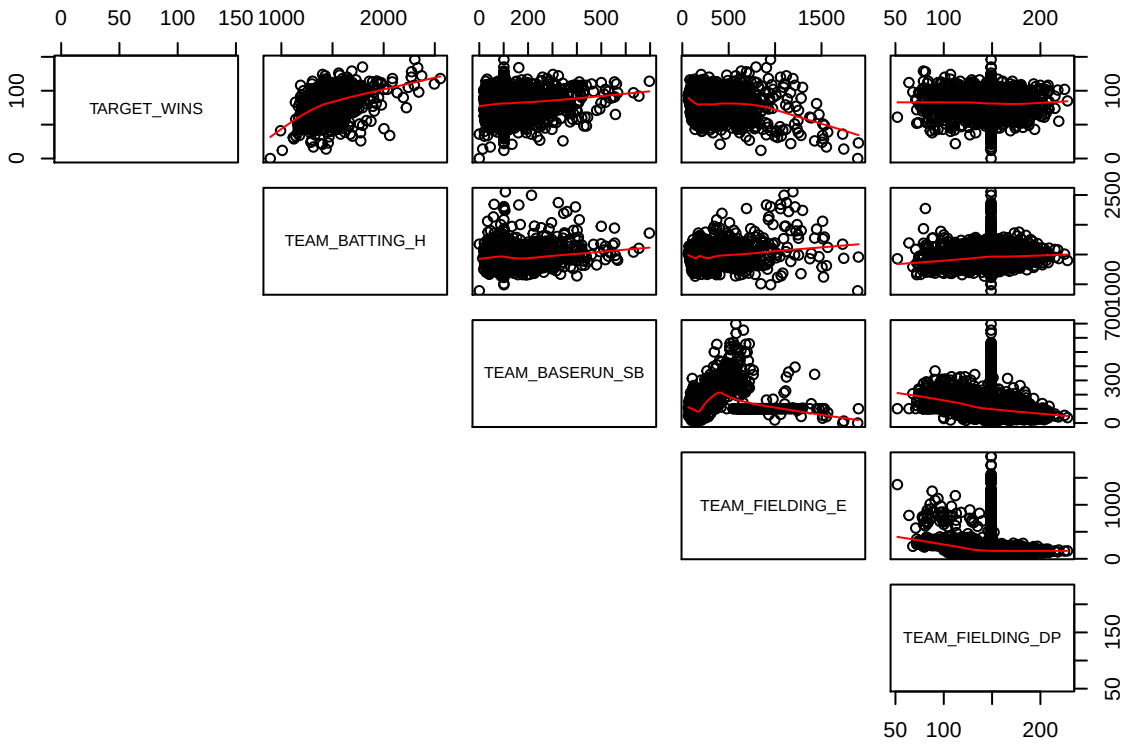
When the model using all variables to predict `TARGET_WINS` was considered, it was decided to create a new model using only the most significant predictors from that regression. This method uses 4 variables only: Base hits by batters, Stolen Bases, Double plays, and Errors. This model produced an F statistic of 225 and $R^2 = 0.283$.

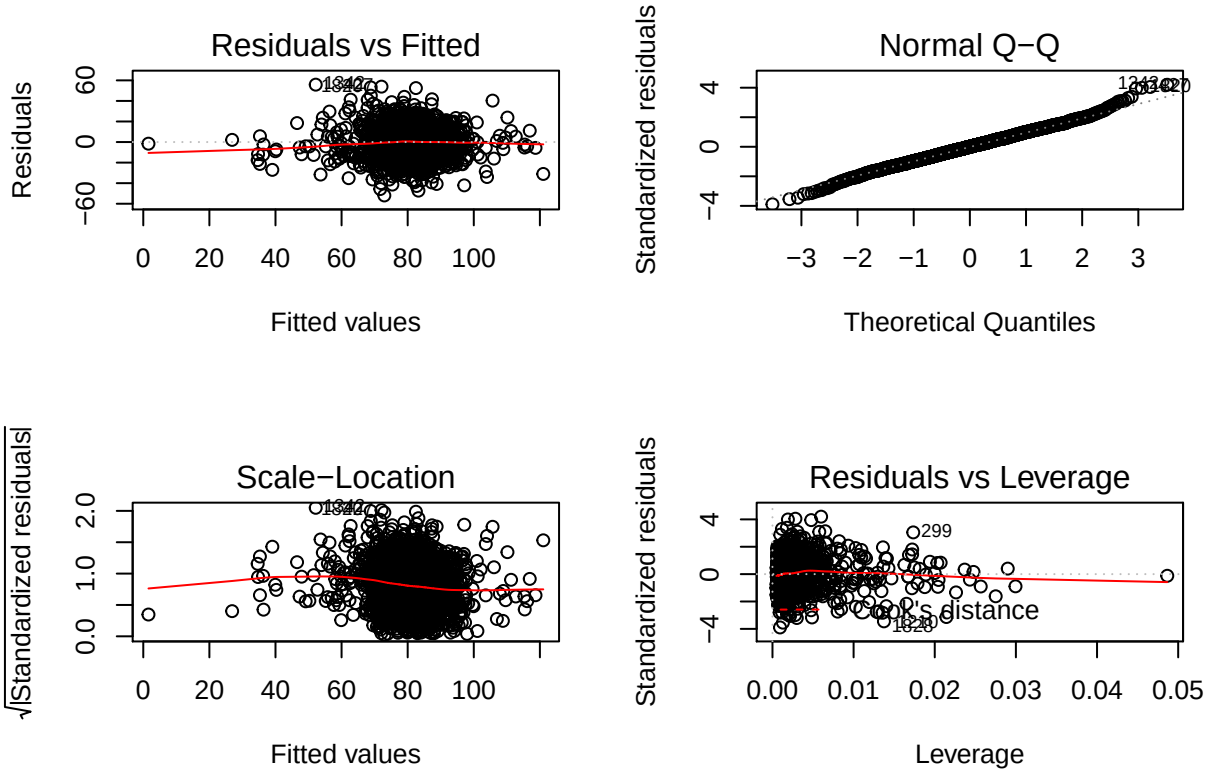
	Estimate	Std. Error	t value	Pr(>|t|)
TEAM_BATTING_H	0.05359	0.002051	26.13	8.068e-132

	Estimate	Std. Error	t value	Pr(>|t|)
TEAM_BASERUN_SB	0.02968	0.003554	8.353	1.144e-16
TEAM_FIELDING_DP	-0.08779	0.01224	-7.17	1.008e-12
TEAM_FIELDING_E	-0.027	0.001367	-19.74	3.322e-80
(Intercept)	17.93	3.198	5.605	2.329e-08

Table 15: Fitting linear model: TARGET_WINS ~ TEAM_BATTING_H + TEAM_BASERUN_SB + TEAM_FIELDING_DP + TEAM_FIELDING_E

Observations	Residual Std. Error	R^2	Adjusted R^2
2276	13.34	0.2841	0.2829





	Estimate	Std. Error	t value	Pr(>|t|)
TEAM_BATTING_H	0.05359	0.002051	26.13	8.068e-132

TEAM_BASERUN_SB 0.02968 0.003554 8.353 1.144e-16

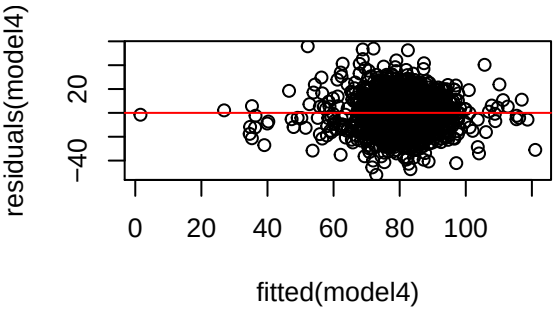
TEAM_FIELDING_DP -0.08779 0.01224 -7.17 1.008e-12

TEAM_FIELDING_E -0.027 0.001367 -19.74 3.322e-80

(Intercept) 17.93 3.198 5.605 2.329e-08

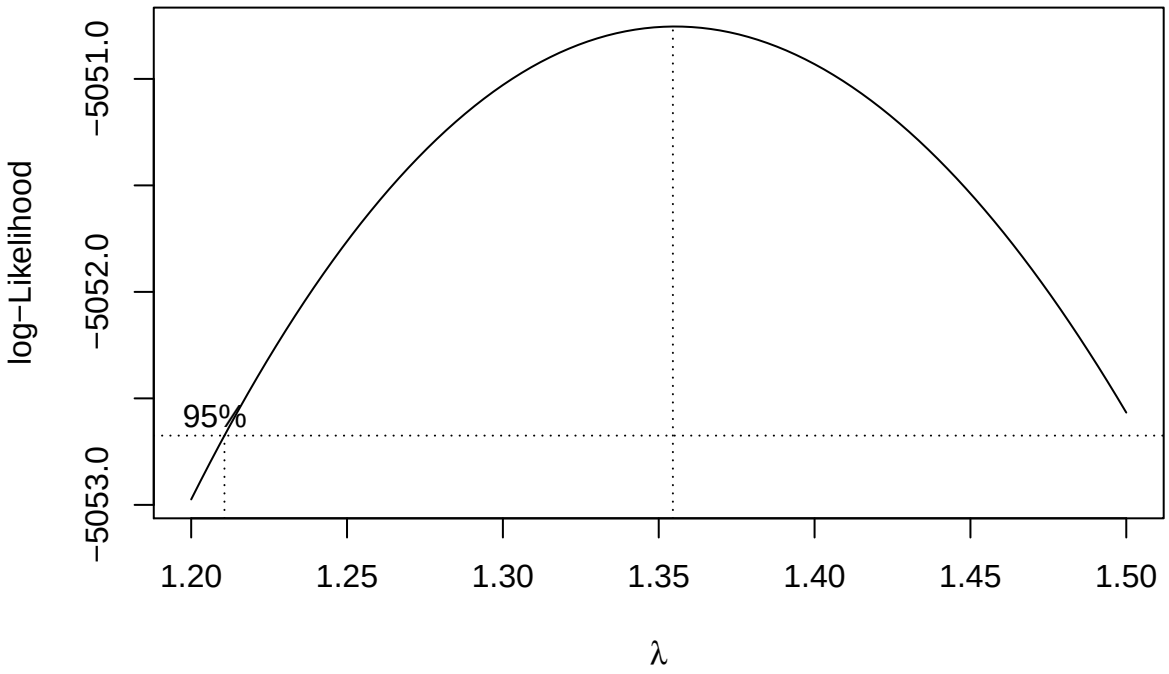
Table 17: Fitting linear model: TARGET_WINS ~ TEAM_BATTING_H + TEAM_BASERUN_SB + TEAM_FIELDING_DP + TEAM_FIELDING_E

Observations	Residual Std. Error	R^2	Adjusted R^2
2276	13.34	0.2841	0.2829



Model 5: Feature engineering and new variables

Many of the more advanced baseball statistics are simply combinations of other statistics (i.e. slugging percentage is total bases divided by at-bats). Using the predictors given in the data set, we wanted to see if combining predictors and/or calculating new values would increase any significance in a model that is trying to predict wins. Total bases and extra bases are both stats that can easily be calculated with the given data. Assuming that the number of doubles, triples and home runs are included in the `TEAM_BATTING_H` variable, we can subtract these out to obtain the number of singles. Adding this to the doubles, triples, and homeruns, each multiplied by the number of bases each is worth (2,3, and 4 respectively) would give total bases. Doing the same thing, but excluding single base hits would give the number of extra bases as well. These two statistics were used in the model, along with many of the other predictors. Using a step-wise methodology, less significant predictors such as caught stealing, pitching walks & hits were left out. To avoid including predictors that were related, the base hits/singles and other multiple base hit predictors were left out. Since the fielding errors predictor was drastically right-skewed, a log transformation was done on this predictor, resulting in a better fit of the model. Most of the predictors have coefficients that behave the way we would expect, given the predictors effect on the game. The extra bases predictor ended up having a negative coefficient, when we would expect there to be a positive one. This may be due to the fact that it is similar stats that make up these predictors, and there is collinearity between the total bases and extra bases predictors. Removing extra bases from the model results in a lower coefficient for total bases, and an overall lower adjusted R^2 value for the model.



	Estimate	Std. Error	t value	Pr(>|t|)
TEAM_BATTING_TOT_BASES	0.04703	0.003591	13.1	7.868e-38

TEAM_BATTING_XTRA_BASE -0.05788 0.01165 -4.969 7.221e-07

TEAM_BATTING_BB	0.01479	0.003173	4.661	3.328e-06
TEAM_BATTING_SO	-0.007654	0.002161	-3.543	0.0004043
TEAM_BASERUN_SB	0.03415	0.004144	8.24	2.869e-16
TEAM_PITCHING_HR	-0.03628	0.01042	-3.483	0.0005049

log(TEAM_FIELDING_E) -8.255 0.853 -9.677 9.833e-22

TEAM_FIELDING_DP	-0.1334	0.01294	-10.31	2.104e-24
(Intercept)	64.92	7.546	8.604	1.413e-17

Table 19: Fitting linear model: TARGET_WINS ~ TEAM_BATTING_TOT_BASES + TEAM_BATTING_XTRA_BASE + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_PITCHING_HR + log(TEAM_FIELDING_E) + TEAM_FIELDING_DP

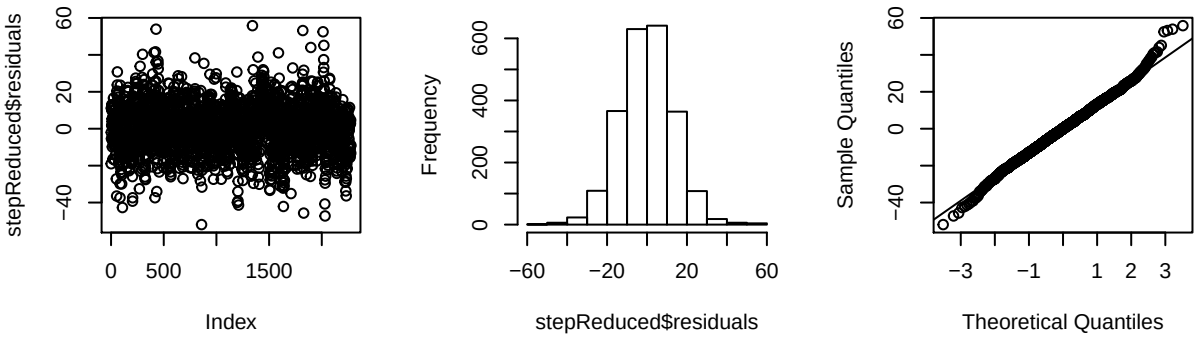
Observations	Residual Std. Error	R^2	Adjusted R^2
2276	13.15	0.3054	0.303

Model Selection and Prediction

The reduced step-wise model described above is selected as the best model for prediction of team wins in a 162-game baseball season. While the R^2 value of this model is not the highest of the models tested, its high F-score indicates that it is the most statistically significant. Additionally, it is one of the more parsimonious models tested, and the simplicity of the model lends itself to easier understanding of the model by possible implementers of the model.

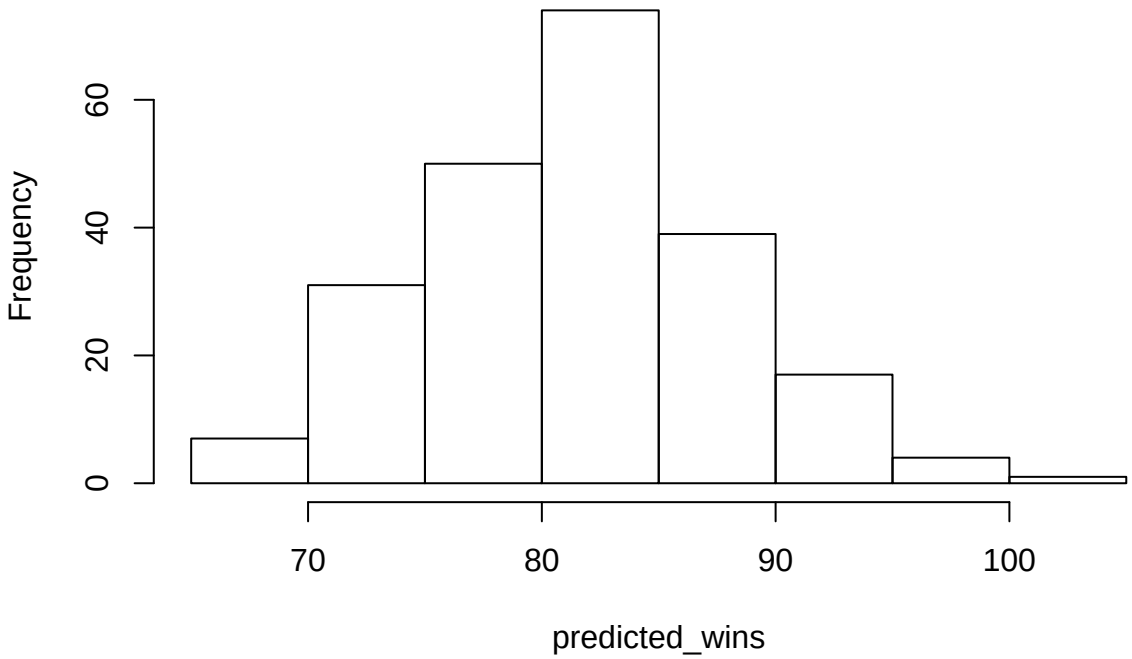
This model has a root-mean-square error of 13.3394, an R^2 of 0.28, and an F-statistic of 225.3. The F statistic has a corresponding p-value of <2.2e-16.

The residual plots for this model are presented below:



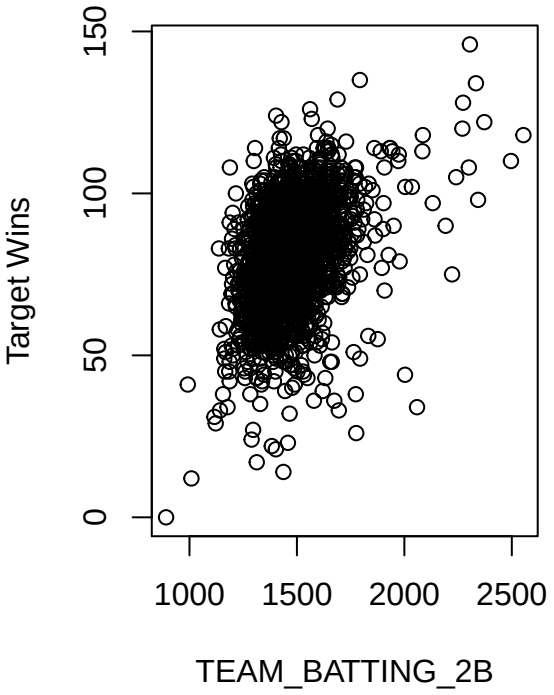
There does not appear to be any pattern in the residuals in the scatterplot, so the condition of linearity can be accepted. The histogram indicates that the residuals are normally distributed. Finally, the scatterplot and Q-Q plot indicate that the residuals indicate near-constant variability. Because the conditions are met, the validity of the use of a linear model is accepted.

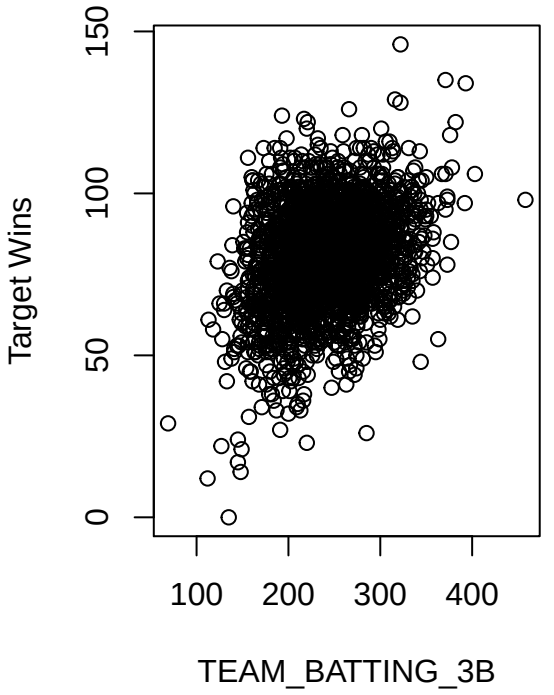
The linear model is applied to an evaluation dataset containing response variables for 259 cases. A histogram of the predicted team wins is presented below.

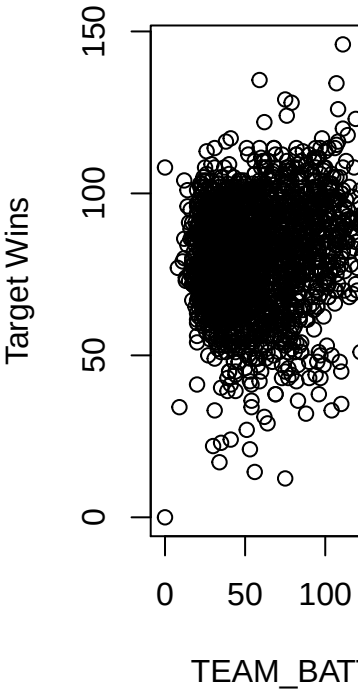


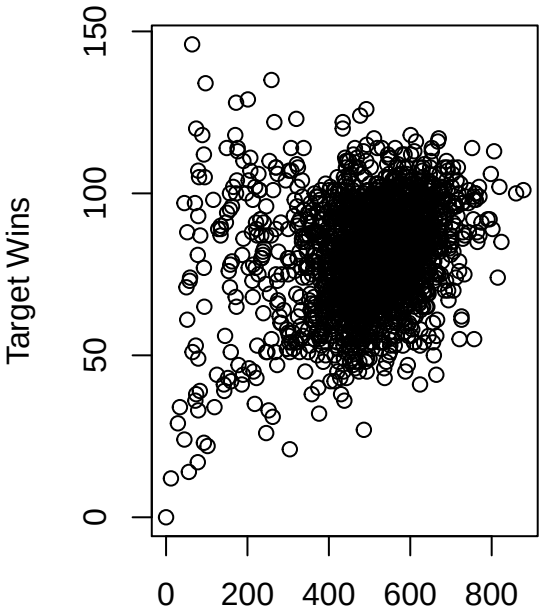
The predicted wins appear roughly normally distributed, with a slight right-skewness. As expected, the distribution is centered near 82, which represents a 0.500 season. Further investigation shows that the median is indeed roughly 82 wins, with the mean slightly lower at roughly 81 wins.

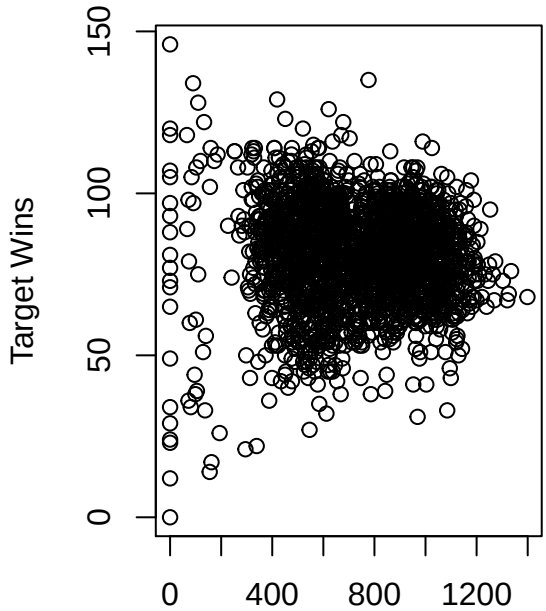
Appendix A

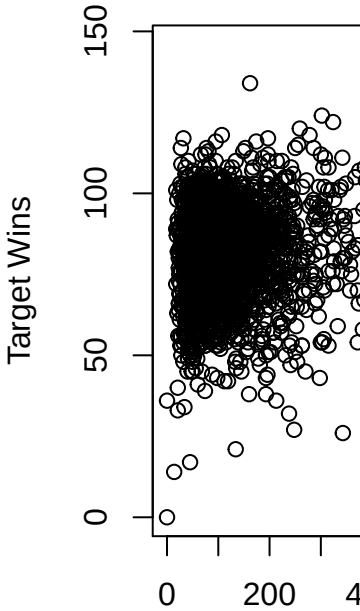


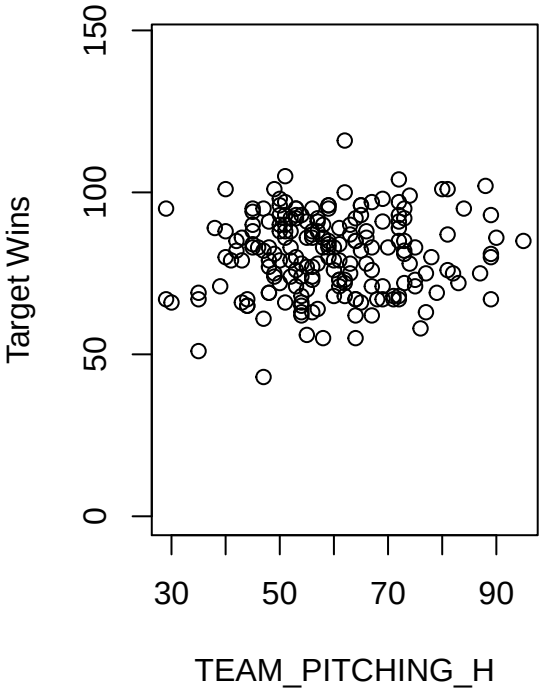


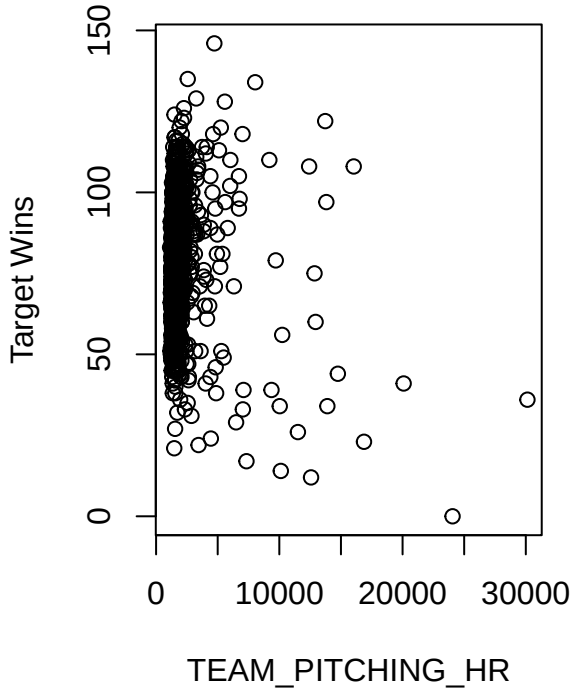


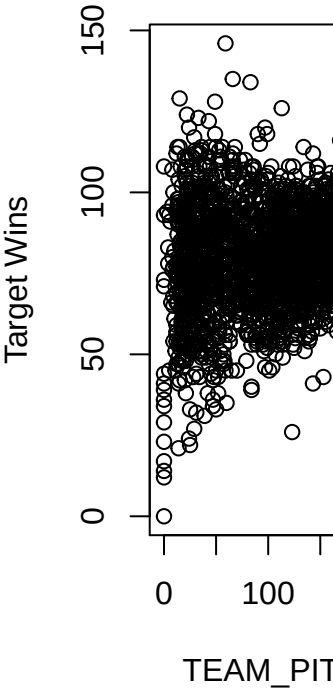

Target Wins
0
50
100
150
0
200
400
600
800
TEAM_BATTING_SO

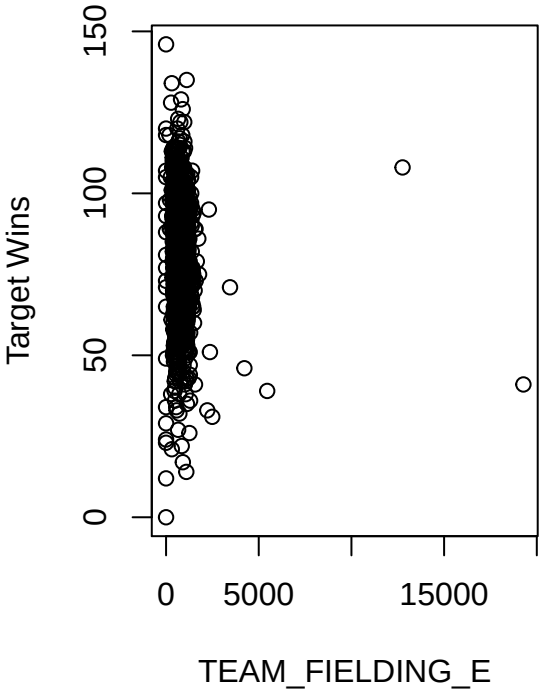

Target Wins
0
50
100
150
0
400
800
1200
TEAM_BASERUN_SB

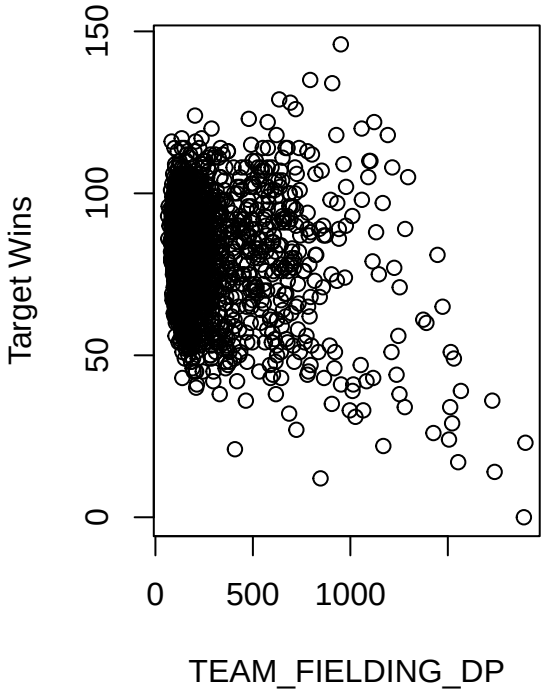

Target Wins
0
50
100
150
0
200
TEAM_BASE

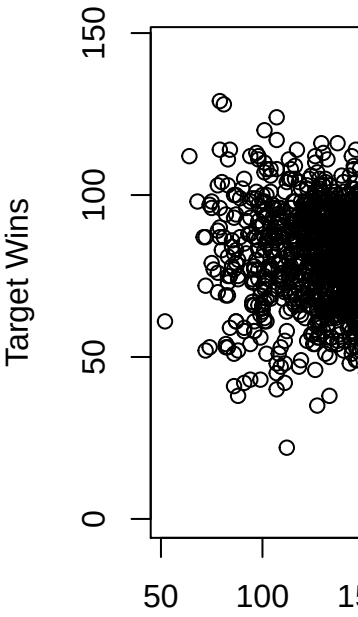

Target Wins
0
50
100
150
30
50
70
90
TEAM_PITCHING_H


Target Wins
0
50
100
150
0
10000
20000
30000
TEAM_PITCHING_HR


Target Wins
0
50
100
150
0
100
TEAM_PIT







Appendix B

Appendix C