Regression Analysis of Baseball Team Performance

Abdellah AitElmouden | Gabriel Abreu | Jered Ataky | Patrick Maloney

2/12/2021

Abstract

To see how regression will help us evaluate baseball team performance, this project is designed to explore whether a teams success in any given season can be predicted or explained by any number of statistics in that season. Our goal is to build a multiple linear regression model on the training data to predict the number of wins for the team. we will explore, analyze and model a historical baseball data set containing approximately 2200 records. Each record represents a professional baseball team from the years 1871 to 2006 inclusive, and the data include the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season.

While correlation does not imply causation, it is suggested that a focus on some of the variables such as single hits or triple or more hits to the exclusion of doubles might be worth pursuing. Also the data suggests that a focus on home runs allowed may not be worth giving up a number of more normal hits.

Introduction

Because baseball is so numbers-heavy, there are many different statistics to consider when searching for the best predictors of team success. There are offensive statistics (offense meaning when a team is batting) and defensive statistics (defense meaning when a team is in the field). These categories can be broken up into many more subcategories. However, for the purpose of the this project we will use the available data to build a multiple linear regression model on the training data to predict the number of wins for the team.

To see how regression will help us predict the number of wins for the team, we actually don't need to understand all the details about the game of baseball, which has over 100 rules. Here, we distill the sport to the basic knowledge one needs to know how to effectively attack the data science problem. The goal of a baseball game is to score more runs (points) than the other team. Each team has 9 batters that have an opportunity to hit a ball with a bat in a predetermined order. After the 9th batter has had their turn, the first batter bats again, then the second, and so on. Each time a batter has an opportunity to bat, we call it a plate appearance (PA). At each PA, the other team's pitcher throws the ball and the batter tries to hit it. The PA ends with an binary outcome: the batter either makes an out (failure) and returns to the bench or the batter doesn't (success) and can run around the bases, and potentially score a run (reach all 4 bases). Each team gets nine tries, referred to as innings, to score runs and each inning ends after three outs (three failures).

Data Exploration

The dataset we will be using was provided in csv file. The files contain approximately 2200 records. Each record represents a professional baseball team from the years 1871 to 2006 inclusive. Each record has the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season. The game statistics that will be used in this study are the following:

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_WINS	Number of wins	Outcome Variable
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

The initial steps are to download the data and take a quick glimpse of the columns, their data types, number of columns, and rows. Based on initial observations, the data contains 2276 teams with a variety of baseball performance statistics.

At first glance, the column BATTING_HBP has numerous NA values that will need to be addressed before building a model. Figure 1 show summary statistics of the target wins. The noteworthy statistics are the average number of wins in a season is 81 games, the median number of wins in a season is 82 games, and the standard deviation is 16 games.

Figure 1 : Summary Statistics

Characteristic	**N = 2,276**
TARGET_WINS	81 (16) 82 0 146
TEAM_BATTING_H	1,469 (145) 1,454 891 2,554
TEAM_BATTING_2B	241 (47) 238 69 458
TEAM_BATTING_3B	55 (28) 47 0 223
TEAM_BATTING_HR	100 (61) 102 0 264
TEAM_BATTING_BB	502 (123) 512 0 878
TEAM_BATTING_HBP	59 (13) 58 29 95
TEAM_BATTING_SO	736 (249) 750 0 1,399
TEAM_BASERUN_SB	125 (88) 101 0 697
TEAM_BASERUN_CS	53 (23) 49 0 201
TEAM_FIELDING_E	246 (228) 159 65 1,898
TEAM_FIELDING_DP	146 (26) 149 52 228
TEAM_PITCHING_BB	553 (166) 536 0 3,645
TEAM_PITCHING_SO	818 (553) 814 0 19,278

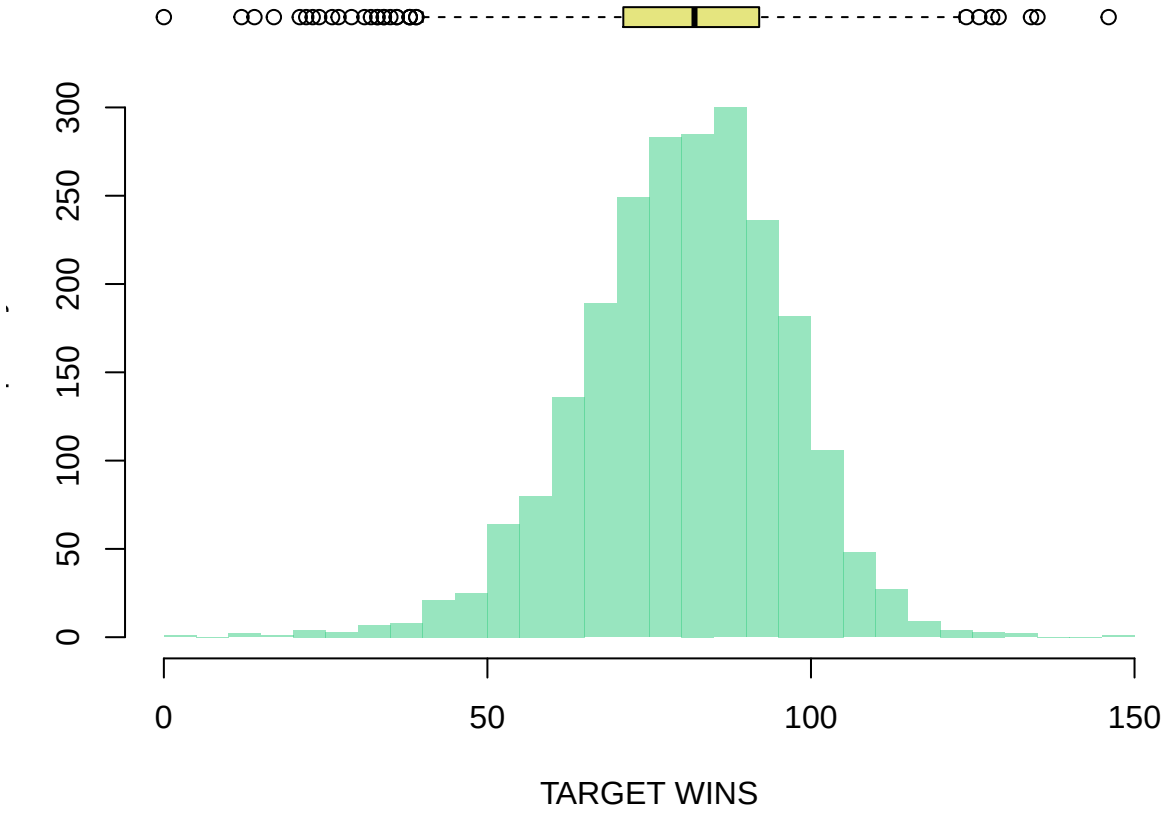
Mean (SD) Median Minimum Maximum

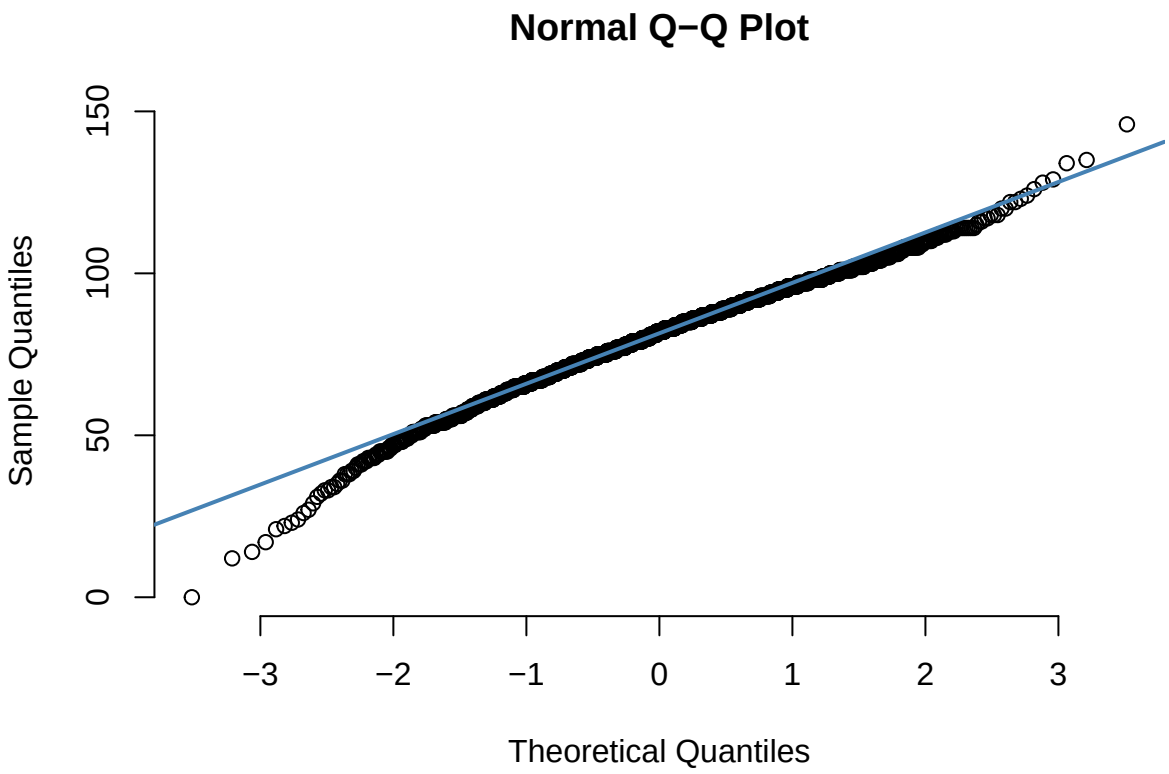
By examining the target wins variable in detail, there is a clear guideline of how many wins each team should approximately win. Most teams will likely win the average number of games (81), but there will be some variability from the average with some teams winning more or less than 81 games.

The other variables also play an important role in understanding the data. In Figure 1, summary statistics are presented for all the variables. it is sufficient in getting the gist of each variable's distribution. For example, the average Base Hits by batters per team is 1469 with the minimum base hits at 891 and maximum base hits at 2554. Remember that the dataset contains baseball statistics on 2276 teams. Missing values were excluded from the summary and they will be dealt with in the data preparation section of this report.

A quick look at Figure 2 will reveal the distribution of the target wins. The distribution is approximately normal with a majority of the target wins falling in the center of the distribution.The approximate normal distribution is confirmed by the QQ plot below the distribution plot. Most of the target wins fall on the line in the QQ plot with some data points diverging at the ends. This indicates possibility of outliers where some teams are winning more games or losing more games than what is expected in the normal range.In the boxplot, there are points that fall outside the whiskers which confirms our suspicions of outliers seen in the QQ plot.

Figure 2 : Distribution and Probability Plot for TARGET_WINS



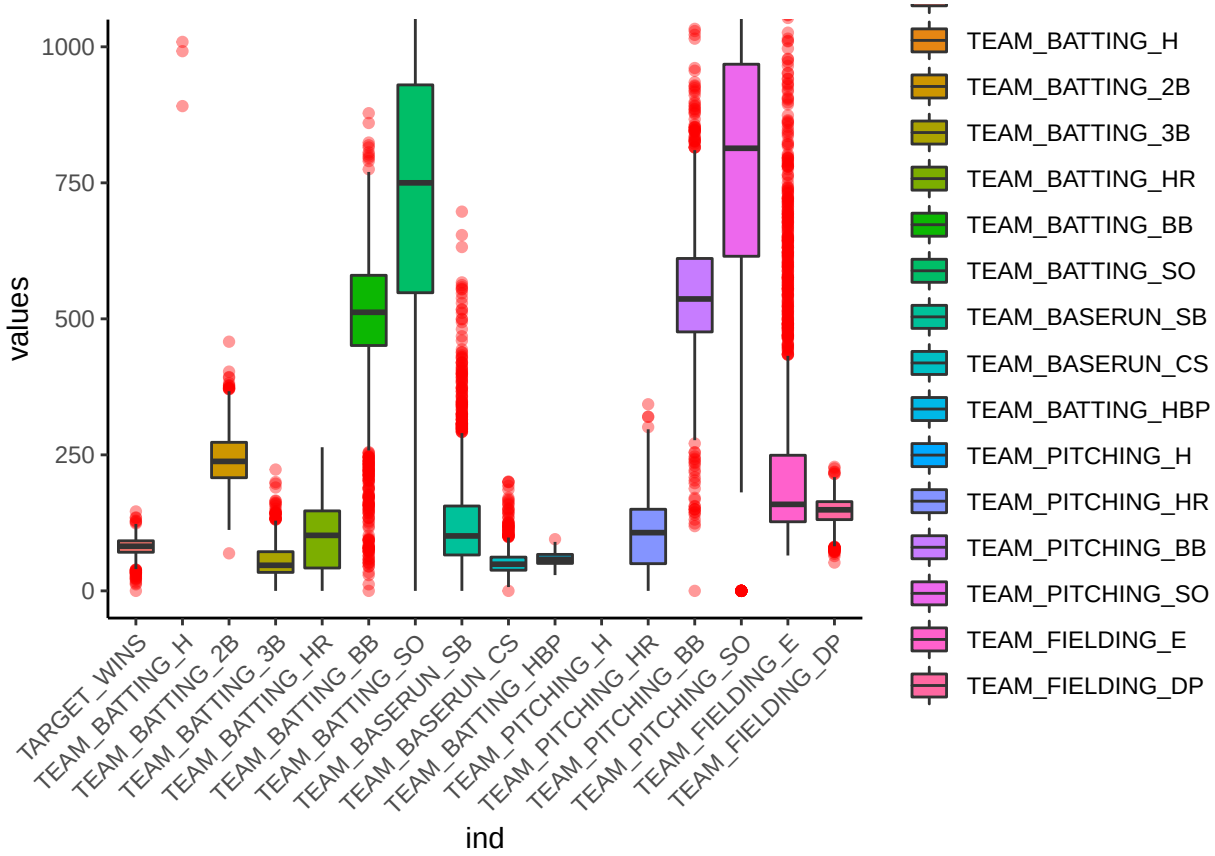


Now in order to build our models properly It's worth exploring for other columns with NA values.

Columns_w_NA	Percent_NA
team_batting_so	4.481547
team_baserun_sb	5.755712
team_baserun_cs	33.919156
team_batting_hbp	91.608084
team_pitching_so	4.481547
team_fielding_dp	12.565905

Outliers

The following diagram shows the outliers for all the variables, both dependent and independent.



As we can see from the graph only 4 of the 16 variables are normally or close to normally distributed. the other 12 variables have a significant skew. The response variable Target_wins seems to be normally distributed. Batting_Hr, Batting_SO and Pitching_HR are bi-modal. 10 of the 16 variables have a minimum value of 0. This is not a major concern as the total % of 0 in each column is less than 1%. The variables Batting_BB, Batting_CS, Baserun_SB, Pitching_BB and Fielding_E have a significant number of outliers.

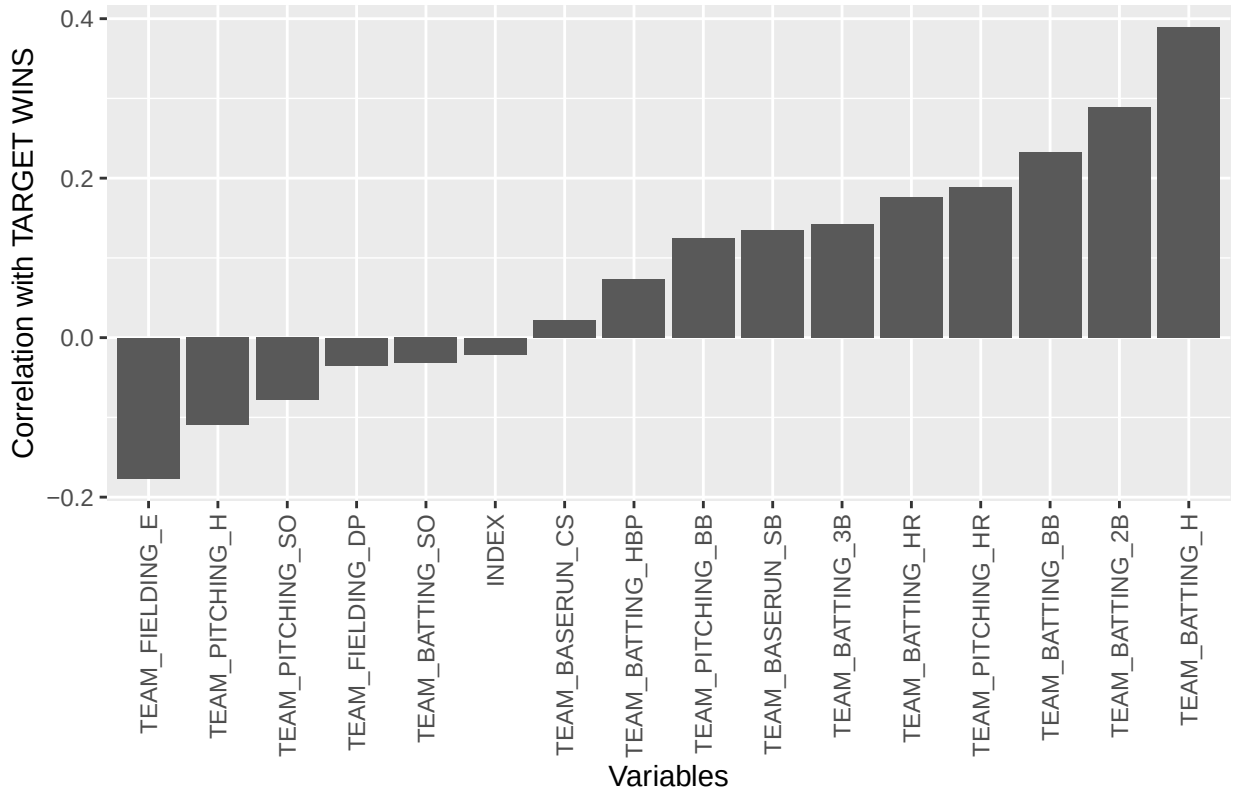
Correlations among predictors and Variable Selection

It is possible that not all variables will need to be used in creating an accurate model. In Figure 4, a correlation value is computed for each variable against target wins. Some variables are highly correlated with target wins, while other variables are not. For example, Base Hits by batters has a value of 0.38877 which is high while Caught stealing is barely correlated with target wins with a value of 0.0224.There is also a column for p-values which indicates whether the correlations are significant. We can use a decision rule of 95% meaning any variable with a p-value of less than 0.05 is significant. It appears that Strike-outs by batters (TEAM_BATTING_SO),Caught stealing(TEAM_BASERUN_CS), Batters hit by pitch (TEAM_BATTING_HBP), and Double plays (TEAM_FIELDING_DP)do not meet our decision rule and could be excluded from use.

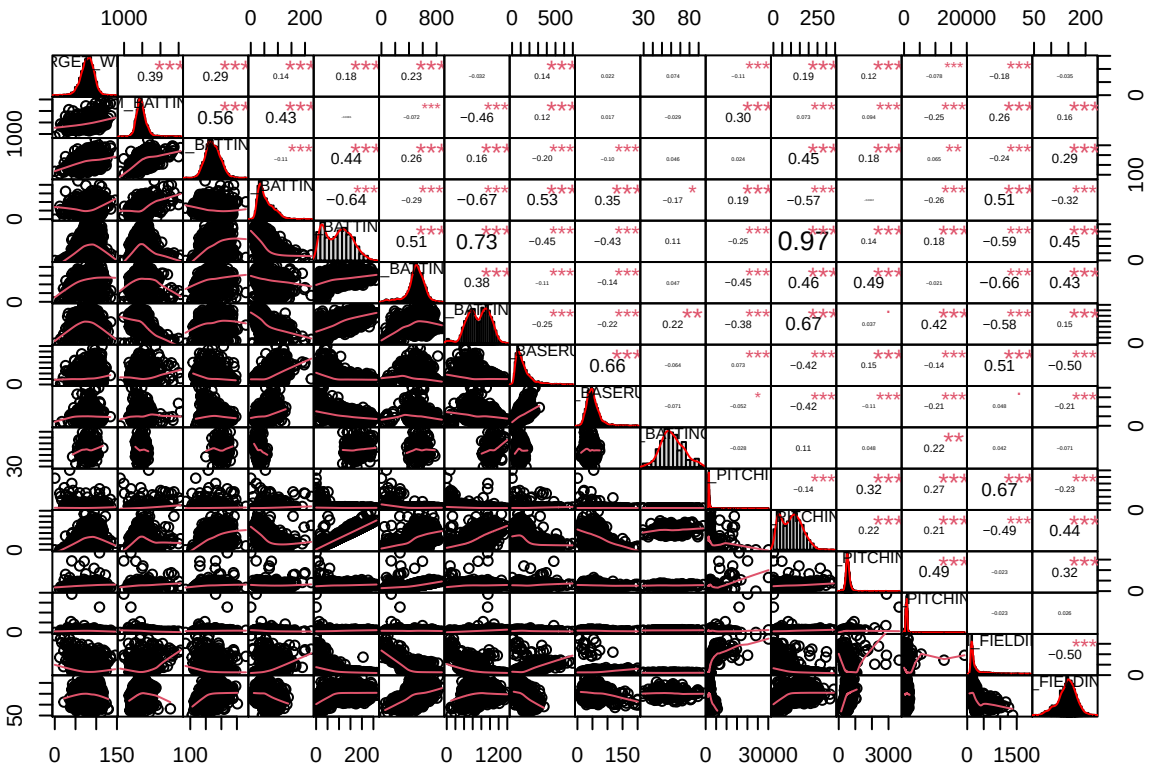
term	TARGET_WINS
INDEX	-0.02105643
TEAM_BATTING_H	0.38876752
TEAM_BATTING_2B	0.28910365
TEAM_BATTING_3B	0.14260841
TEAM_BATTING_HR	0.17615320
TEAM_BATTING_BB	0.23255986
TEAM_BATTING_SO	-0.03175071

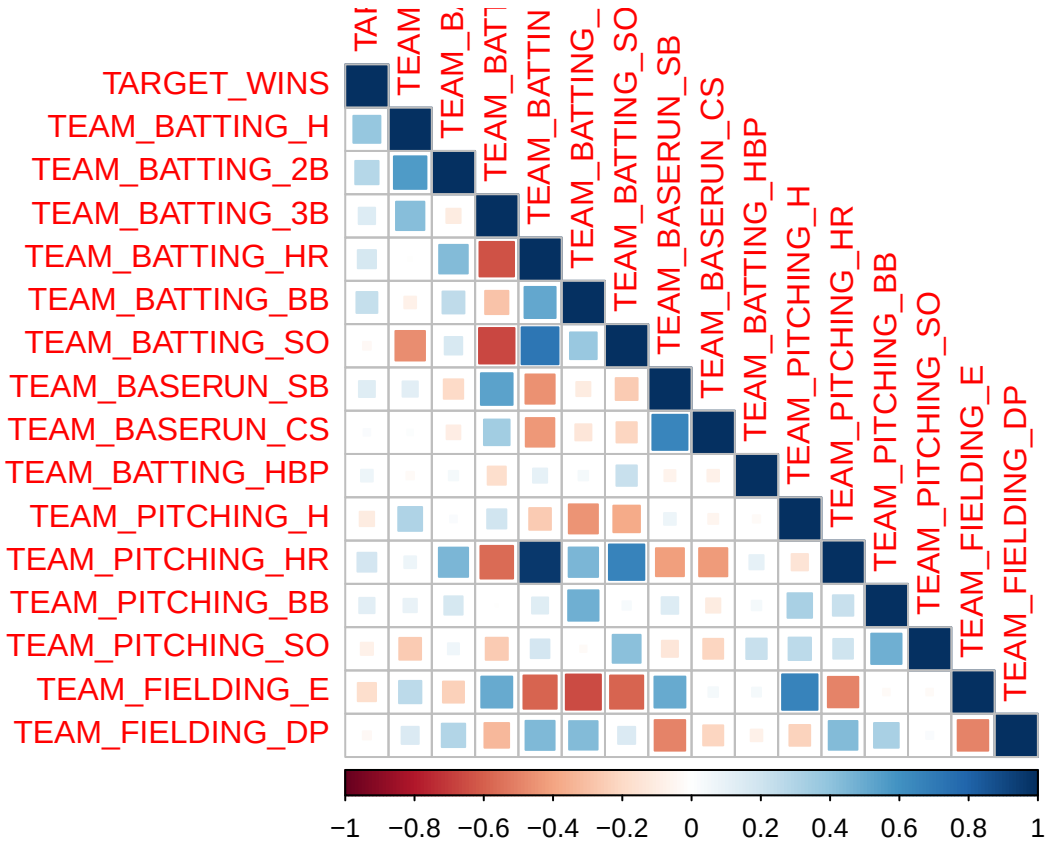
TEAM_BASERUN_SB	0.13513892
TEAM_BASERUN_CS	0.02240407
TEAM_BATTING_HBP	0.07350424
TEAM_PITCHING_H	-0.10993705
TEAM_PITCHING_HR	0.18901373
TEAM_PITCHING_BB	0.12417454
TEAM_PITCHING_SO	-0.07843609
TEAM_FIELDING_E	-0.17648476
TEAM_FIELDING_DP	-0.03485058

Figure 4: Correlation Against Target Win



Before entirely excluding variables, it is a good idea to transform the data by fixing missing values or combining variables and reexamine the viability of those variables for predicting wins.





From the table we can see that there are positive or negative correlations among the predictors. If we look at the numerical correlations with the response variable. We can see that the predictors Batting_H, Batting_HR, Batting_BB, Pitching_H, and Pitching_HR are more correlated and should be included in our regression.

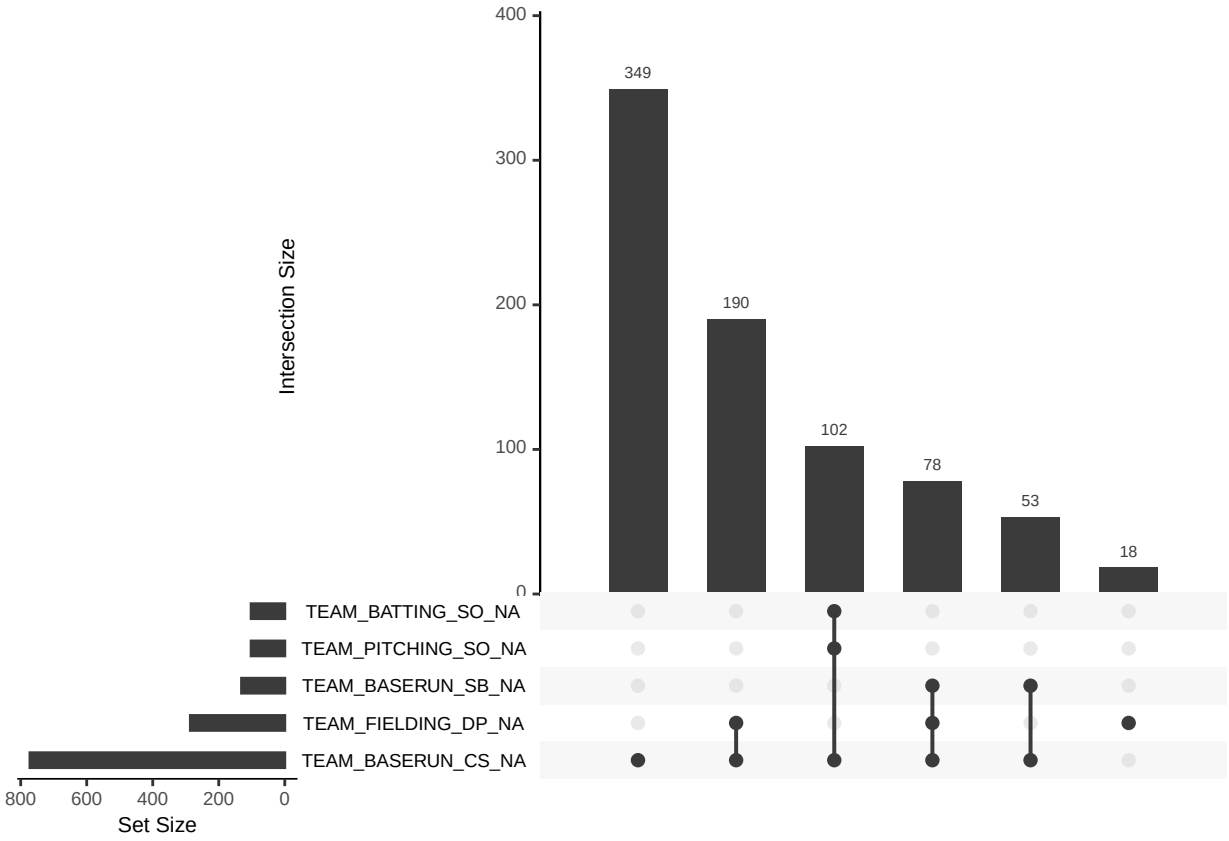
Also Examining significant correlations among the independent variables, we see that four of the pairs have a correlation close to 1. This can lead to multicollinearity issues in our analysis.

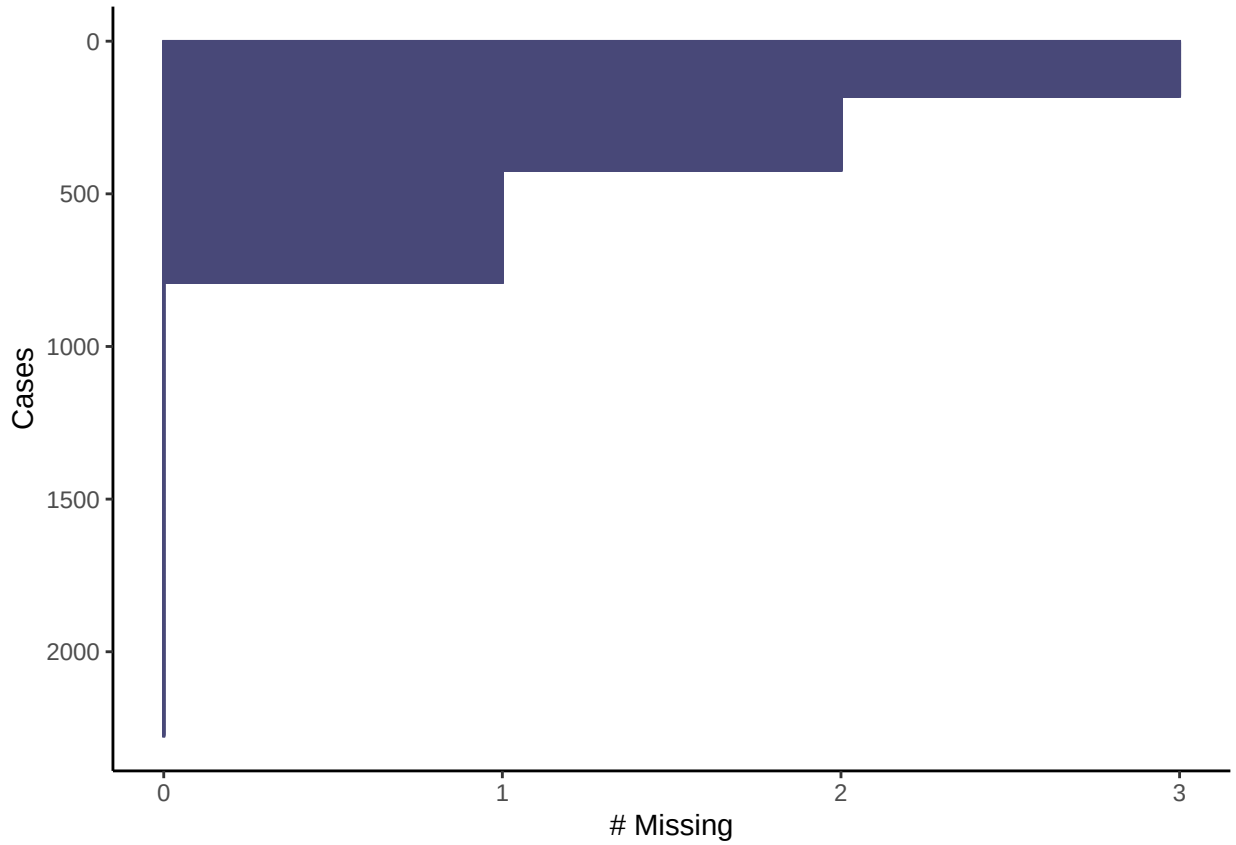
Data Preparation

Missing values need to be handled before building models.They can be handled by either dropping the records, dropping the entire variable, or imputation. In this case, it was determined that Batters hit by pitch variable should be dropped altogether prior to model building because it has too many missing values to properly impute.All other variables with missing values will be considered for the model because a majority of the records are not missing.These variables will be imputed.

First we will remove Batting_HBP (Hit by Pitch) which has 92% missing values.

We will look at the patterns and intersections of missingness among the variables, using the naniar package. We can see that only 22 of the observations have all 5 variables missing, we will just delete these cases. The pattern suggests that the variables are Missing at Random (MAR)


Intersection Size
400
300
200
100
0
349
190
102
78
53
18
TEAM_BATTING_SO_NA
TEAM_PITCHING_SO_NA
TEAM_BASERUN_SB_NA
TEAM_FIELDING_DP_NA
TEAM_BASERUN_CS_NA
800
600
400
200
0
Set Size



By looking at the patterns and intersections of missing data among the variables. We can see that 5 variables have missing values, Team_BATTING has the most missing values so we are completely removing these observations. Overall, the pattern suggests that the variables are Missing at Random (MAR).

When it comes to fixing missing values,there are several methods at our disposal. The first technique is to fill the missing values with the mean values of each variable. We'll use the Hmisc R Package to fill the missing data with the mean, most of the time, mean imputation will lead to good results. The same procedure will be used for the other variables with missing values in Model 2 but by using the Median instead of the mean

The second technique for imputing missing values is to use a decision tree. This is slightly more involved, but will likely give the better results. A decision tree will be created for each variable with missing values. In mean imputation, a fixed value is used for missing values of an entire variable whereas in decision tree imputation, a value is used based on certain conditions.

Build Models

MODEL 1: MEAN FULL MODEL

This is a full model containing all the variables with the meanusred for missing values. This is a good starting model to determine how well each variable helps predict wins. The mean is generally an adequate guess for missing values. In this model, no selection technique is used. All variables are manually included.

To the that we used the Hmisc R package to imputes missing value using user defined statistical method (mean in our case)

```
##
## Call:
## lm(formula = 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_PITCHING_H + TEAM_PITCHING_HR +
##     TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP,
##     data = train_model1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.113  -7.633  -0.018   7.324  45.214
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      59.237509   6.113895   9.689  < 2e-16 ***
## TEAM_BATTING_H   -0.002070   0.005441  -0.380 0.703633
## TEAM_BATTING_2B  -0.023765   0.008808  -2.698 0.007034 **
## TEAM_BATTING_3B   0.168568   0.018723   9.003  < 2e-16 ***
## TEAM_BATTING_HR   0.321816   0.055324   5.817 6.98e-09 ***
## TEAM_BATTING_BB   0.093055   0.018164   5.123 3.30e-07 ***
## TEAM_BATTING_SO  -0.049966   0.008917  -5.603 2.40e-08 ***
## TEAM_BASERUN_SB   0.081187   0.006161  13.178  < 2e-16 ***
## TEAM_BASERUN_CS  -0.059350   0.014616  -4.061 5.09e-05 ***
## TEAM_PITCHING_H   0.025458   0.002263  11.251  < 2e-16 ***
## TEAM_PITCHING_HR -0.216833   0.052137  -4.159 3.33e-05 ***
## TEAM_PITCHING_BB -0.060526   0.016914  -3.578 0.000354 ***
## TEAM_PITCHING_SO  0.028051   0.007993   3.509 0.000459 ***
## TEAM_FIELDING_E  -0.084678   0.005292 -16.001  < 2e-16 ***
## TEAM_FIELDING_DP -0.120255   0.012485  -9.632  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.96 on 1975 degrees of freedom
##   (286 observations deleted due to missingness)
## Multiple R-squared:  0.3858, Adjusted R-squared:  0.3814
## F-statistic: 88.61 on 14 and 1975 DF,  p-value: < 2.2e-16
```

The overall p-value for Model 1is less than 0.0001,which indicates a significant model in predicting wins. The Adjusted R-Squared and Mean Square Error (MSE) will be the metrics used to determine the best model.A higher Adjusted R-Squared is better and a lower MSE is better.In this case, the Adjusted R-Squared is 0.3814 and MSE is 168.91, which will be the current benchmark.

The resulting equation for Model 1 is :

WINS =+ 25.06831
+0.0473146*Base Hits by batters
-0.0209806*Doubles by batters
+0.0692224*Triples by batters
+0.0680963*Homeruns by batters
+0.0108690*Walks by batters
-0.0081244*Strikeouts by batters
+0.0299345*Stolen bases
-0.01173*Caught stealing
-0.00073149*Hits allowed
+0.01481*Homeruns allowed
+0.00008066*Walks allowed
+0.0026600*Strikeouts by pitchers
-0.02118*Errors
-0.1208451*Double plays

Most of Model 1 makes senseas positive measures of success like Base Hits, Triples, Homeruns, Walks by batters, and Stolen bases are positive coefficients in the equation while negative measures of success like Strikeouts by batters, Caught stealing, Hits allowed, and Errors are negative coefficients in the equation.All these values make intuitive sense.On the other hand, Doubles and Double plays are shown as negative coefficients when they should have a positive impact on wins. Also, Homeruns allowed and Walks allowed are shown as positive coefficients when they should have negative impact on wins.These values are counter intuitive.The counter intuitive parts of the model may need to be further investigated if this model were to be chosen for deployment.However, for now, the model will be kept as a benchmark despite certain measures not making sense.

MODEL 2: MEDIAN WITH STEPWISE

Earlier in the exploration of the data, the analysis revealed the possibility of outliers present in the data. Because the mean is highly influenced by outliers, this model attempts to remedy that by using the median to impute missing values.Model 2 is a significant model based on the p-value of less than 0.0001. The Adjusted R-Squared is 0.3147 and MSE is 169.799.

```
##
## Call:
## lm(formula = 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_PITCHING_H + TEAM_PITCHING_HR +
##     TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP,
##     data = train_model2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.141  -7.562  -0.055   7.303  45.741
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      58.081468   6.120416   9.490  < 2e-16 ***
## TEAM_BATTING_H   -0.002408   0.005486  -0.439 0.660776
## TEAM_BATTING_2B  -0.023924   0.008849  -2.704 0.006915 **
## TEAM_BATTING_3B   0.169910   0.018789   9.043  < 2e-16 ***
## TEAM_BATTING_HR   0.316324   0.055363   5.714 1.27e-08 ***
## TEAM_BATTING_BB   0.089135   0.018298   4.871 1.20e-06 ***
## TEAM_BATTING_SO  -0.045012   0.008991  -5.006 6.04e-07 ***
## TEAM_BASERUN_SB   0.078868   0.006138  12.850  < 2e-16 ***
## TEAM_BASERUN_CS  -0.060070   0.014307  -4.198 2.81e-05 ***
## TEAM_PITCHING_H   0.026060   0.002291  11.374  < 2e-16 ***
## TEAM_PITCHING_HR -0.213394   0.052207  -4.087 4.54e-05 ***
## TEAM_PITCHING_BB -0.056704   0.017043  -3.327 0.000893 ***
## TEAM_PITCHING_SO  0.024406   0.008105   3.011 0.002635 **
## TEAM_FIELDING_E  -0.082493   0.005281 -15.621  < 2e-16 ***
## TEAM_FIELDING_DP -0.121035   0.012572  -9.627  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11 on 1975 degrees of freedom
##   (286 observations deleted due to missingness)
## Multiple R-squared:  0.3816, Adjusted R-squared:  0.3772
## F-statistic: 87.05 on 14 and 1975 DF,  p-value: < 2.2e-16
```

The resulting equation for Model 2 is :

WINS =+ 25.0602
+0.04824*Base Hits by batters
-0.02006*Doubles by batters
+0.06047*Triples by batters
+0.05299*Homeruns by batters
+0.01042*Walks by batters
-0.009349*Strikeouts by batters
+0.02949*Stolen bases
-0.01188*Caught stealing
-0.0007342*Hits allowed
+0.01480*Homeruns allowed
+0.00008891*Walks allowed
+0.002843*Strikeouts by pitchers
-0.02112*Errors
-0.1210*Double plays

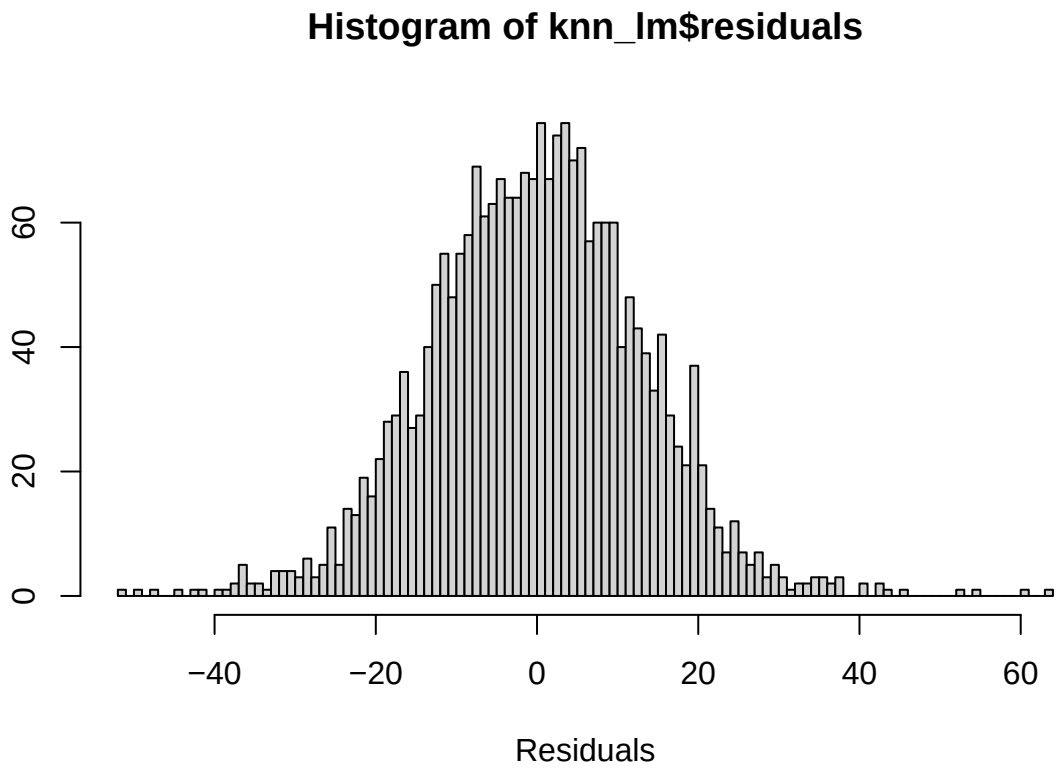
MODEL 3: knn Imputation

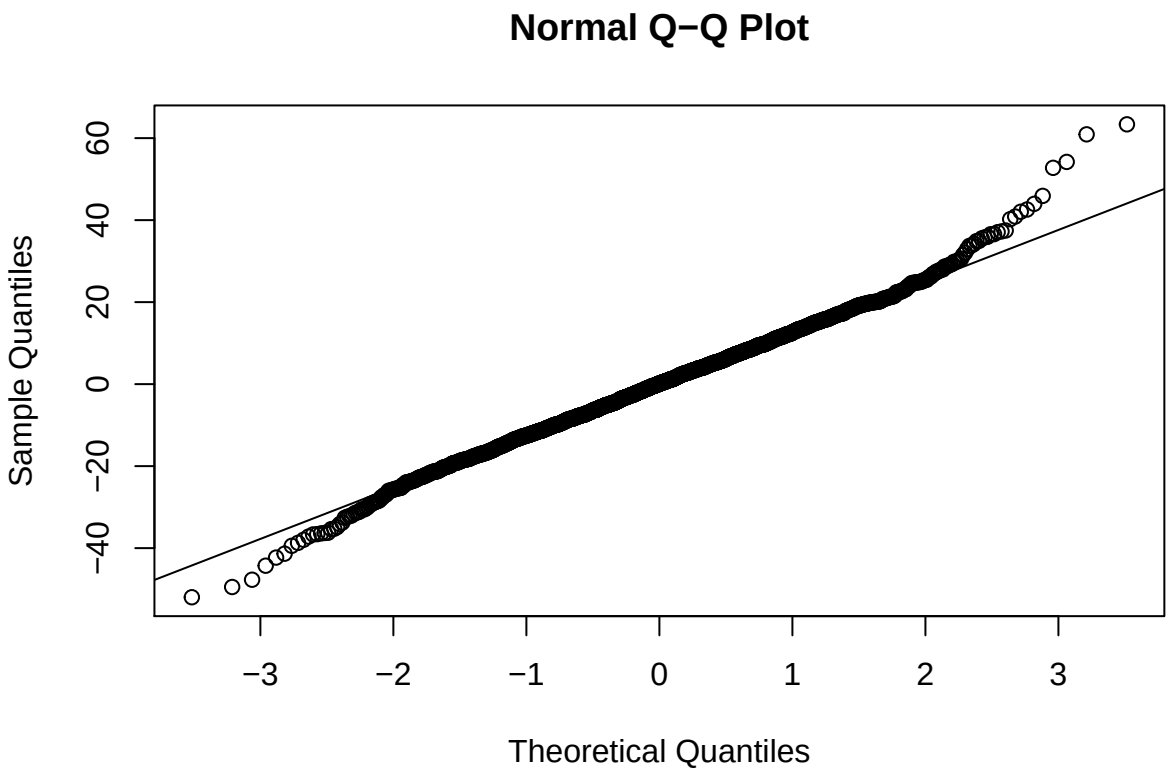
Building the stepwise regression model with knn imputed values

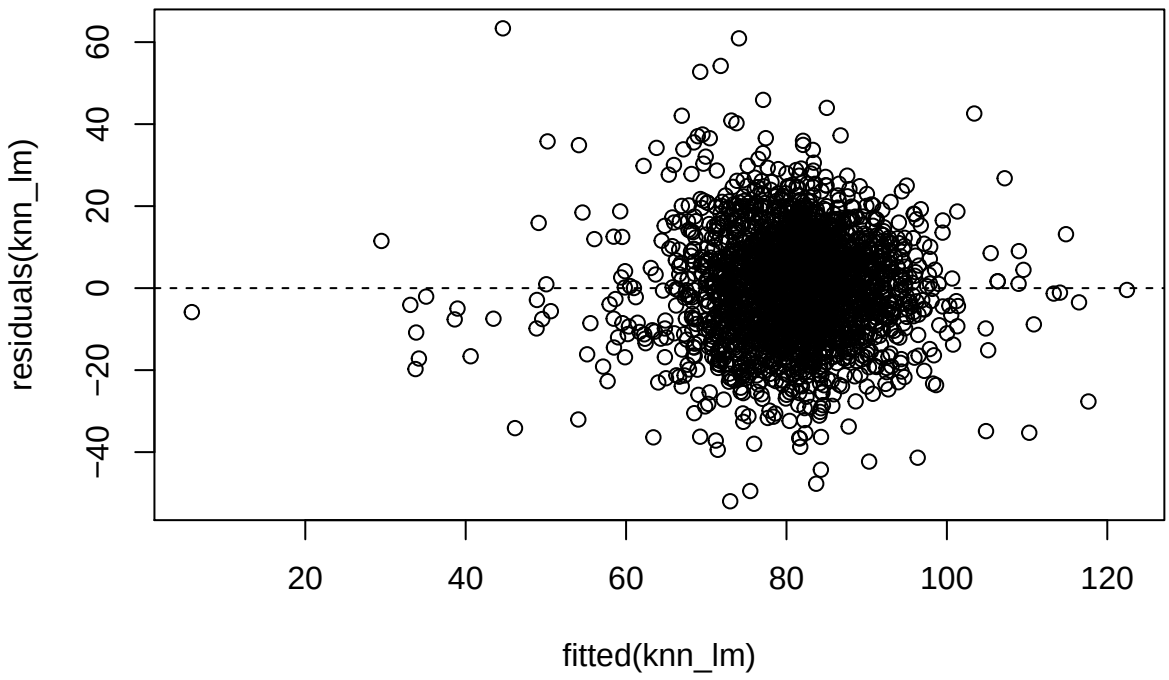
```
##    nvmax     RMSE  Rsquared      MAE    RMSESD RsquaredSD     MAESD
## 1      2 13.79212 0.2334200 10.88609 0.4274216 0.07365865 0.3267934
## 2      3 13.42654 0.2724130 10.64926 0.3496118 0.06603760 0.2728521
## 3      4 13.24085 0.2925472 10.53556 0.3679441 0.06416916 0.3015155
## 4      5 13.20091 0.2968303 10.46976 0.3657979 0.05783949 0.2993156
## 5      6 13.07741 0.3094892 10.31952 0.4446855 0.06491801 0.3105508
## 6      7 13.06613 0.3114016 10.26766 0.4351503 0.06375965 0.2756727
## 7      8 13.16812 0.3006475 10.31135 0.4888167 0.06662962 0.3453242
## 8      9 13.25753 0.2916576 10.36201 0.5564611 0.06760619 0.3322944
## 9     10 13.20108 0.2987041 10.28312 0.5924854 0.06776726 0.3507628
## 10    11 13.13176 0.3049392 10.26027 0.5115151 0.07370665 0.3321067
## 11    12 13.32504 0.2843850 10.43739 0.4710437 0.06738657 0.3153434
## 12    13 13.21363 0.2963206 10.34155 0.4282910 0.06856800 0.2930328
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_SO +
##     TEAM_BASERUN_SB + TEAM_PITCHING_HR + TEAM_FIELDING_E + TEAM_FIELDING_DP,
##     data = knn_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.980  -8.547   0.148   8.398  63.362
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      35.203403   4.656401   7.560 5.80e-14 ***
## TEAM_BATTING_H    0.044863   0.002599  17.263  < 2e-16 ***
## TEAM_BATTING_SO  -0.010323   0.002091  -4.937 8.52e-07 ***
## TEAM_BASERUN_SB   0.037616   0.003784   9.942  < 2e-16 ***
## TEAM_PITCHING_HR  0.059940   0.007971   7.520 7.85e-14 ***
## TEAM_FIELDING_E  -0.027353   0.001582 -17.289  < 2e-16 ***
## TEAM_FIELDING_DP -0.117420   0.013006  -9.028  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 13.13 on 2269 degrees of freedom
## Multiple R-squared:  0.3069, Adjusted R-squared:  0.3051
## F-statistic: 167.5 on 6 and 2269 DF,  p-value: < 2.2e-16
```




Normal Q−Q Plot
Sample Quantiles
−40
−20
0
20
40
60
−3
−2
−1
0
1
2
3
Theoretical Quantiles



Now lets run the model…

Select A Model

As mentioned earlier, the main decision criterionis the Adjusted R-Squared. A higher Adjusted R-Squared is indicative of better performance. MSE is also used as a secondary criterion which measures the difference between actual and predicted values. A lower MSE is better.Aside from these criteria, the goal is also to have a highly interpretable model that makes sense.

Measures	Model 1	Model 2	Model 3
r_squared	0.3858	0.3816	0.3067
adj_rsquared	0.3814	0.3772	0.3049
rse	10.9600	11.0000	13.1300
f_stat	88.6100	87.0500	167.3000

Looking at the three models, we are going to use model 1, and the model is selected based on the criteria we had mentioned.

Prediction

The test data should go to to the same cleaning process with the training data.

After, we have cleaned the test dataset, here's how the result looks like on the table above which include the predicted values, and the prediction intervals.

```
eval_model1 <- drop_na(eval_model1)
predict1 <- predict(model1, newdata = eval_model1, interval="prediction")
kable(predict1)
```

fit	lwr	upr
60.07336	38.49515	81.65158
65.67460	44.11320	87.23601
71.23288	49.68348	92.78227
84.86087	63.31623	106.40550
103.93799	81.28573	126.59026
71.85566	50.14909	93.56223
72.34177	50.75179	93.93174
75.22567	53.67294	96.77841
70.76950	49.23281	92.30620
64.63909	43.08353	86.19464
83.07114	61.51171	104.63057
85.21038	63.63224	106.78852
82.79881	61.22043	104.37720
85.81195	64.21798	107.40592
74.96797	53.40346	96.53247
71.29929	49.76256	92.83602
76.99432	55.46692	98.52171
68.15789	46.58147	89.73432
86.08610	64.50586	107.66633
85.29403	63.73352	106.85453
82.19979	60.64507	103.75451
83.06310	61.51523	104.61096
70.06209	48.52136	91.60282
78.84225	57.29569	100.38881
86.07804	64.51782	107.63826
68.95160	47.39845	90.50475
82.72954	61.13728	104.32179
67.20395	45.60529	88.80262
91.75254	70.18634	113.31875
86.83842	65.30544	108.37141
84.97968	63.42602	106.53334
85.02571	63.44522	106.60620
79.97374	58.43932	101.50816
86.02490	64.46247	107.58733
77.33769	55.81314	98.86224
90.51117	68.94900	112.07334
79.17794	57.40528	100.95060
88.95546	67.38655	110.52437
81.77011	60.20082	103.33939
92.94351	71.37142	114.51559
72.58169	50.93529	94.22809
68.45749	46.91301	90.00196
75.56876	53.92197	97.21555
73.74384	52.19562	95.29207
81.66188	60.09615	103.22760
74.53919	52.99974	96.07864
74.49560	52.94412	96.04708
69.84047	48.31247	91.36848
77.20263	55.66987	98.73539
98.68540	76.94744	120.42337
80.59269	59.00785	102.17753
66.38084	44.79346	87.96822
80.95703	59.39260	102.52145
93.48978	71.92510	115.05447
80.78763	59.17746	102.39781
87.42718	65.89639	108.95796
84.14919	62.57919	105.71919

Conclusion

3 models were generated from baseball team data from 1871 to 2006 to predict number of wins. Prior to generating the models, the baseball data was analyzed to better understand the relationship between variables. The chosen model was Model 1 which was created using mean….

Appendix

```
knitr::opts_chunk$set(cache =TRUE)

# 0.  Librairies

library(corrplot)
library(tidyverse)
library(Hmisc)
library(PerformanceAnalytics)
library(corrplot)
library(mice)
library(gt)
library(DMwR)
library(caret)
library(bnstruct)
library(VIM)
library(corrr)
library(tidyverse)
library(gtsummary)
library(kableExtra)
```

```
# 1. Data Exploration

## Import data

train_data <- read.csv("https://raw.githubusercontent.com/aaitelmouden/DATA621/master/Project1/moneybal
glimpse(train_data)


## Summary table

wins <- train_data %>% select(TARGET_WINS, TEAM_BATTING_H, TEAM_BATTING_2B, TEAM_BATTING_3B, TEAM_BATTI
table1 <- tbl_summary(wins,
          statistic = list(all_continuous() ~ "{mean} ({sd}) {median} {min} {max}"), missing = "no")
table1


## Create more variables

singles <- train_data$TEAM_BATTING_H - (train_data$TEAM_BATTING_2B + train_data$TEAM_BATTING_3B + train_

train_data$TEAM_BATTING_SLG <- ((train_data$TEAM_BATTING_HR *4)+ (train_data$TEAM_BATTING_3B*3) + (trai
```

```
## Plots

### Layout to split the screen
layout(mat = matrix(c(1,2),2,1, byrow=TRUE),  height = c(1,8))
```

```r
####  Draw the boxplot and the histogram
par(mar=c(0, 3.1, 1.1, 2.1))
boxplot(train_data$TARGET_WINS ,main="Figure 2 : Distribution and Probability Plot for TARGET_WINS",cex
par(mar=c(4, 3.1, 1.1, 2.1))
hist(train_data$TARGET_WINS , breaks=40 , col=rgb(0.2,0.8,0.5,0.5) , border=F , main="" , xlab="TARGET
```

```r
### qq plots

qqnorm(train_data$TARGET_WINS, pch = 1, frame = FALSE)
qqline(train_data$TARGET_WINS, col = "steelblue", lwd = 2)


max_obs <- 2276
batting_so_na <- ((102/max_obs) * 100)
baserun_sb_na <- (131/max_obs) * 100
baserun_cs_na <- (772/max_obs) * 100
batting_hbp_na <- (2085/max_obs) * 100
pitching_so_na <- (102/max_obs) * 100
fielding_dp_na <- (286/max_obs) * 100

df_percent_na <- data.frame(Columns_w_NA = c("team_batting_so", "team_baserun_sb", "team_baserun_cs", "
```

```r
### the largest islands in the world

gt_tbl <- gt(data = df_percent_na, )


### Show the gt Table

gt_tbl
```

```r
### Outlliers

ggplot(stack(train_data[,-1]), aes(x = ind, y = values, fill=ind)) +
  geom_boxplot(outlier.colour = "red",  outlier.alpha=.4) +
  coord_cartesian(ylim = c(0, 1000)) +
  theme_classic()+
  theme(axis.text.x=element_text(angle=45, hjust=1))


### Correlation

COR <- train_data %>%
  correlate() %>%
  focus(TARGET_WINS)
gt(COR)

COR %>%
  mutate(rowname = factor(term, levels = term[order(TARGET_WINS)])) %>%  # Order by correlation strengt
  ggplot(aes(x = rowname, y = TARGET_WINS)) +
    geom_bar(stat = "identity") +
    ylab("Correlation with TARGET WINS") +
    xlab("Variables") + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+ ggtitle("F
```

```r
#### pairwise.complete.obs ignores NA values and computes correlation on complete observations
#### we might have to run these corrplots again after we handle the NA values

chart.Correlation(train_data[-c(1)], histograme=TRUE, method= "pearson", use="pairwise.complete.obs")

data.corr <- cor(train_data[-c(1)], use="pairwise.complete.obs")

corrplot(data.corr, type = "lower", method="square")

flattenCorrMatrix <- function(cormat, pmat) {
  ut <- upper.tri(cormat)
  data.frame(
    row = rownames(cormat)[row(cormat)[ut]],
    column = rownames(cormat)[col(cormat)[ut]],
    cor  =(cormat)[ut],
    p = pmat[ut]
    )
}

#### Eliminate INDEX from data frame
data_no_index <- train_data[-c(1)]

cor_matrix <- rcorr(as.matrix(data_no_index))

flattenCorrMatrix(cor_matrix$r, cor_matrix$P)
```

```r
#. 2 Data preparation

train_data <- train_data[-11]


par(mfrow=c(1,2))
gg_miss_upset(train_data,
              nsets = 5,
              nintersects = NA)
gg_miss_case(train_data)+
  theme_classic()


par(mfrow=c(1,2))
gg_miss_upset(train_data,
              nsets = 5,
              nintersects = NA)
gg_miss_case(train_data)+
  theme_classic()
```

```r
# 3. Build models

### MODEL 1: MEAN FULL MODEL

### Filling missing values with Mean using the impute package

train_model1 <- train_data
train_model1$TEAM_BATTING_SO[is.na(train_model1$TEAM_BATTING_SO)] = mean(train_model1$TEAM_BATTING_SO, r
```

```r
train_model1$TEAM_BASERUN_SB[is.na(train_model1$TEAM_BASERUN_SB)] = mean(train_model1$TEAM_BASERUN_SB, r
train_model1$TEAM_BASERUN_CS[is.na(train_model1$TEAM_BASERUN_CS)] = mean(train_model1$TEAM_BASERUN_CS, r
train_model1$TEAM_PITCHING_SO[is.na(train_model1$TEAM_PITCHING_SO)] = mean(train_model1$TEAM_PITCHING_SO

model1 <- lm(TARGET_WINS ~
               TEAM_BATTING_H +   # Base Hits by batters (1B,2B,3B,HR)
               TEAM_BATTING_2B +  # Doubles by batters (2B)
               TEAM_BATTING_3B +  # Triples by batters (3B)
               TEAM_BATTING_HR +  # Homeruns by batters (4B)
               TEAM_BATTING_BB +  # Walks by batters
               TEAM_BATTING_SO +  # Strikeouts by batters
               TEAM_BASERUN_SB +  # Stolen bases
               TEAM_BASERUN_CS +  # Caught stealing
               TEAM_PITCHING_H +  # Hits allowed
               TEAM_PITCHING_HR + # Homeruns allowed
               TEAM_PITCHING_BB + # Walks allowed
               TEAM_PITCHING_SO + # Strikeouts by pitchers
               TEAM_FIELDING_E +  # Errors
               TEAM_FIELDING_DP,  # Double Plays
             data=train_model1)
summary(model1)

### Mean Square Error (MSE)

mean(model1$residuals^2)

layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(model1)
```

```r
### MODEL 2: MEDIAN WITH STEPWISE

train_model2 <- train_data

#### Filling missing values with Mean using the impute package

train_model2$TEAM_BATTING_SO[is.na(train_model2$TEAM_BATTING_SO)] = median(train_model2$TEAM_BATTING_SO
train_model2$TEAM_BASERUN_SB[is.na(train_model2$TEAM_BASERUN_SB)] = median(train_model2$TEAM_BASERUN_SB
train_model2$TEAM_BASERUN_CS[is.na(train_model2$TEAM_BASERUN_CS)] = median(train_model2$TEAM_BASERUN_CS
train_model2$TEAM_PITCHING_SO[is.na(train_model2$TEAM_PITCHING_SO)] = median(train_model2$TEAM_PITCHING_


model2 <- lm(TARGET_WINS ~
               TEAM_BATTING_H +   # Base Hits by batters (1B,2B,3B,HR)
               TEAM_BATTING_2B +  # Doubles by batters (2B)
               TEAM_BATTING_3B +  # Triples by batters (3B)
               TEAM_BATTING_HR +  # Homeruns by batters (4B)
               TEAM_BATTING_BB +  # Walks by batters
               TEAM_BATTING_SO +  # Strikeouts by batters
               TEAM_BASERUN_SB +  # Stolen bases
               TEAM_BASERUN_CS +  # Caught stealing
               TEAM_PITCHING_H +  # Hits allowed
               TEAM_PITCHING_HR + # Homeruns allowed
               TEAM_PITCHING_BB + # Walks allowed
               TEAM_PITCHING_SO + # Strikeouts by pitchers
```

```
                TEAM_FIELDING_E +  # Errors
                TEAM_FIELDING_DP,  # Double Plays
              data=train_model2)
summary(model1)

mean(model2$residuals^2)

layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(model2)
```

```
### MODEL 3: kNN Imputation

train_data_impute <- select(train_data, -c(INDEX))


#### using default values

knn_data <- knnImputation(train_data_impute)

summary(knn_data)

train.control <- trainControl(method ='cv', number=10)

step.model <- train(TARGET_WINS ~., data = knn_data,
                    method = "leapSeq",
                    tuneGrid = data.frame(nvmax = 2:13),
                    trControl = train.control
                    )

step.model$results

step.model$bestTune

summary(step.model$finalModel)

coef(step.model$finalModel, 6)


knn_lm <- lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_PITCHING_

summary(knn_lm)


hist(knn_lm$residuals, xlab = "Residuals", ylab = "", breaks=100)
qqnorm(knn_lm$residuals)
qqline(knn_lm$residuals)

plot(fitted(knn_lm), residuals(knn_lm))
abline(h=0, lty = 2)
```

References

«««< HEAD

-Dealing with Missing Data using R : https://medium.com/coinmonks/dealing-with-missing-data-using-r-3ae428da2d17 -Decision Tree : http://www.learnbymarketing.com/tutorials/rpart-decision-trees-in-r/ -Decision Tree : https://www.datacamp.com/community/tutorials/decision-trees-R -Introduction to Data Science (Case Study Moneyball): https://rafalab.github.io/dsbook/linear-models.html#case-study-moneyball ======= »»»> bf59d52d90d6b4ab9b840f4b7fa1cd132debf1fb