Homework #1

Zachary Safir, Mario Pena, Sie Siong Wong, hector Santana

9/16/2021

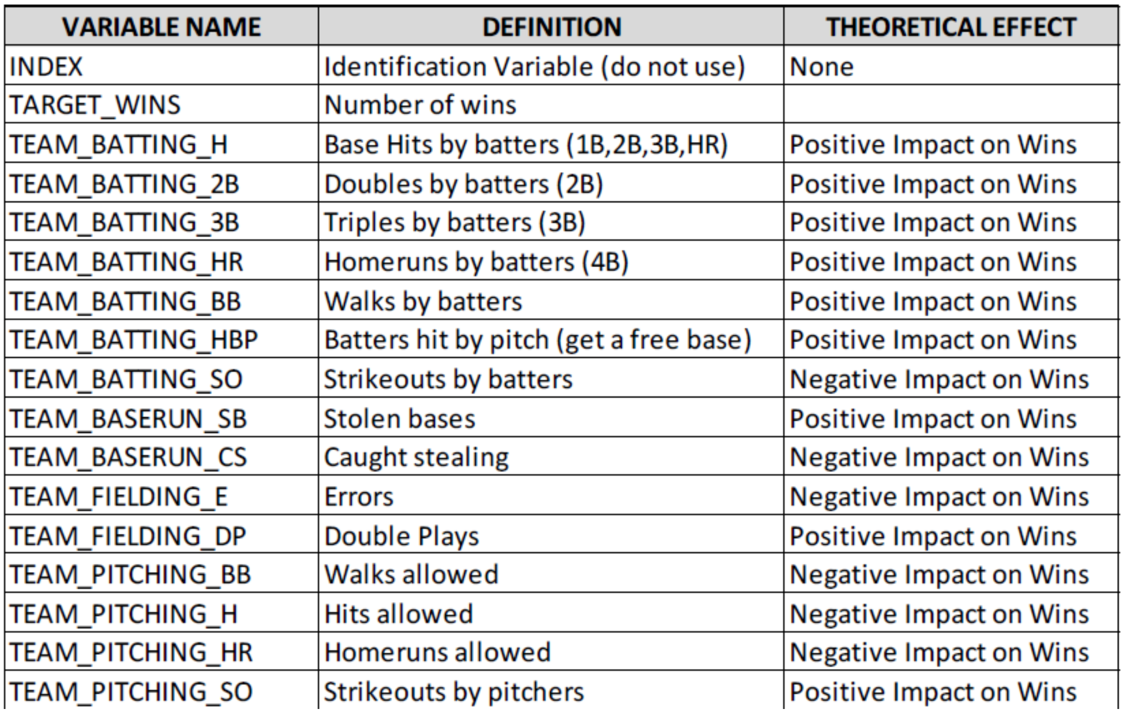
Contents

1 Introduction

For this assignment, we were tasked to explore, analyze and model a data set containing 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.

Our objective is to build a multiple linear regression model on the training data to predict the number of wins for the team. We can only use the variables given to us (or variables that we derive from the variables provided). Below is a short description of the variables of interest in the data set:

For all resources used in the making of this assignment, including the code used for all our findings, please use this link here to find our GitHub page.



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

2 Data Exploration

2.1 Initial Summary Statistics

As shown below, we can being exploring the initial state of our training data. We notice that we have a mix of normal and skewed variables. TEAM_PITCHING_H appears to be very right skewed.

We also notice that there are variables with lots of missing values, with the most extreme being TEAM_BATTING_HBP, which appear to be almost entirely missing.

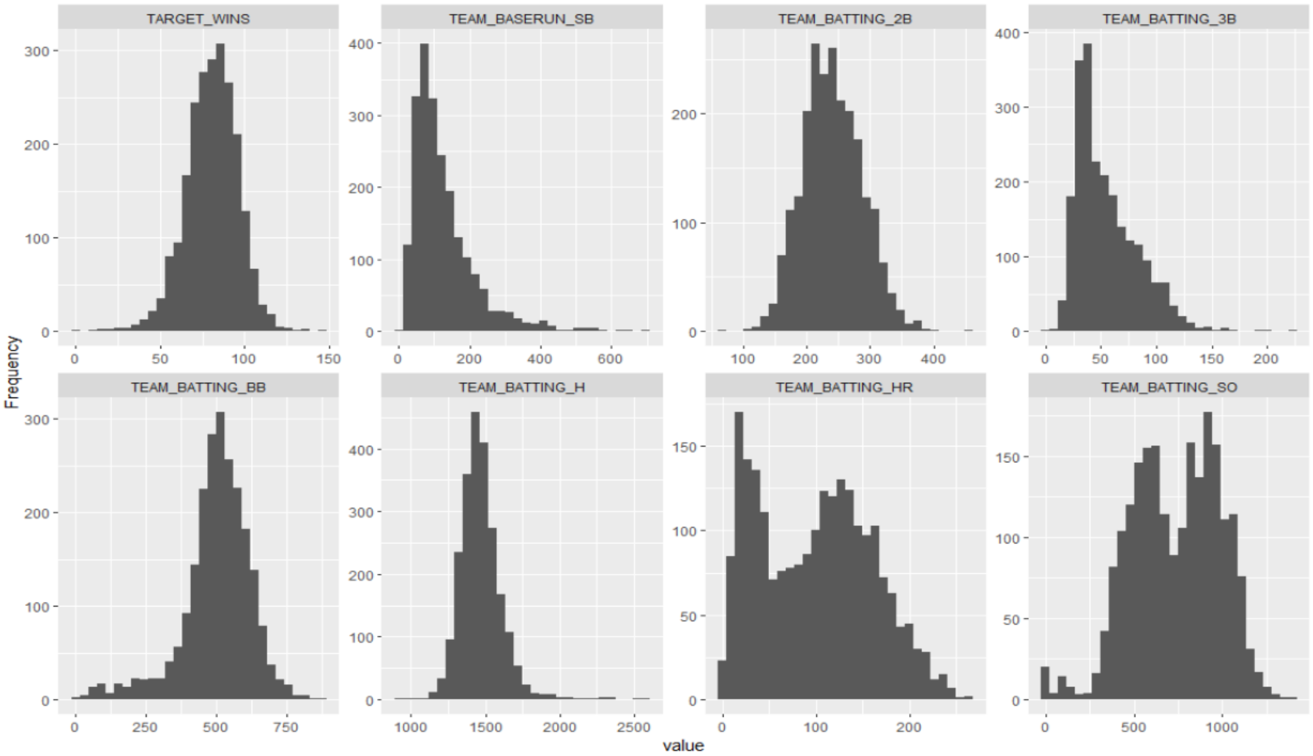
We also notice some odd values for TEAM_WINS. According to our research, the highest number of wins by a baseball team, is a tie at 116. However, there appears to be a team in our data that supposedly has 146 wins. There also appears to be a team that has 0 wins, which is also impossible. No such team has ever gone without winning a single game. There are hardly any teams with below 40 wins, and only one team that appears to have only won 20 games, but no teams below that.

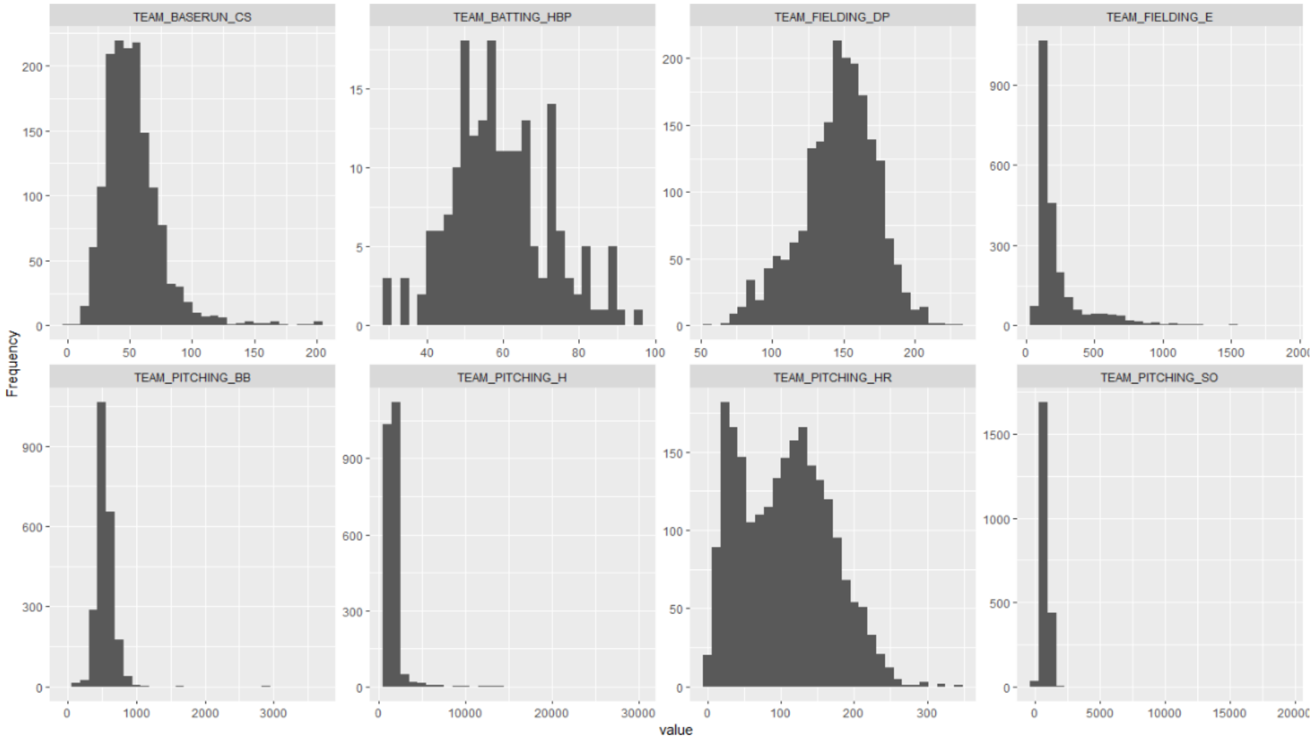
	Overall (N=2276)
TARGET_WINS	
- Mean (SD)	80.79 (15.75)
- Median (Q1, Q3)	82.00 (71.00, 92.00)
- Range	0.00 - 146.00
TEAM_BATTING_H	
- Mean (SD)	1469.27 (144.59)
- Median (Q1, Q3)	1454.00 (1383.00, 1537.25)
- Range	891.00 - 2554.00
TEAM_BATTING_2B	
- Mean (SD)	241.25 (46.80)
- Median (Q1, Q3)	238.00 (208.00, 273.00)
- Range	69.00 - 458.00
TEAM_BATTING_3B	
- Mean (SD)	55.25 (27.94)
- Median (Q1, Q3)	47.00 (34.00, 72.00)
- Range	0.00 - 223.00
TEAM_BATTING_HR	
- Mean (SD)	99.61 (60.55)
- Median (Q1, Q3)	102.00 (42.00, 147.00)
- Range	0.00 - 264.00
TEAM_BATTING_BB	
- Mean (SD)	501.56 (122.67)
- Median (Q1, Q3)	512.00 (451.00, 580.00)
- Range	0.00 - 878.00
TEAM_BATTING_SO	
- Mean (SD)	735.61 (248.53)
- Median (Q1, Q3)	750.00 (548.00, 930.00)
- Range	0.00 - 1399.00
- N-Miss	102
TEAM_BASERUN_SB	
- Mean (SD)	124.76 (87.79)
- Median (Q1, Q3)	101.00 (66.00, 156.00)
- Range	0.00 - 697.00
- N-Miss	131

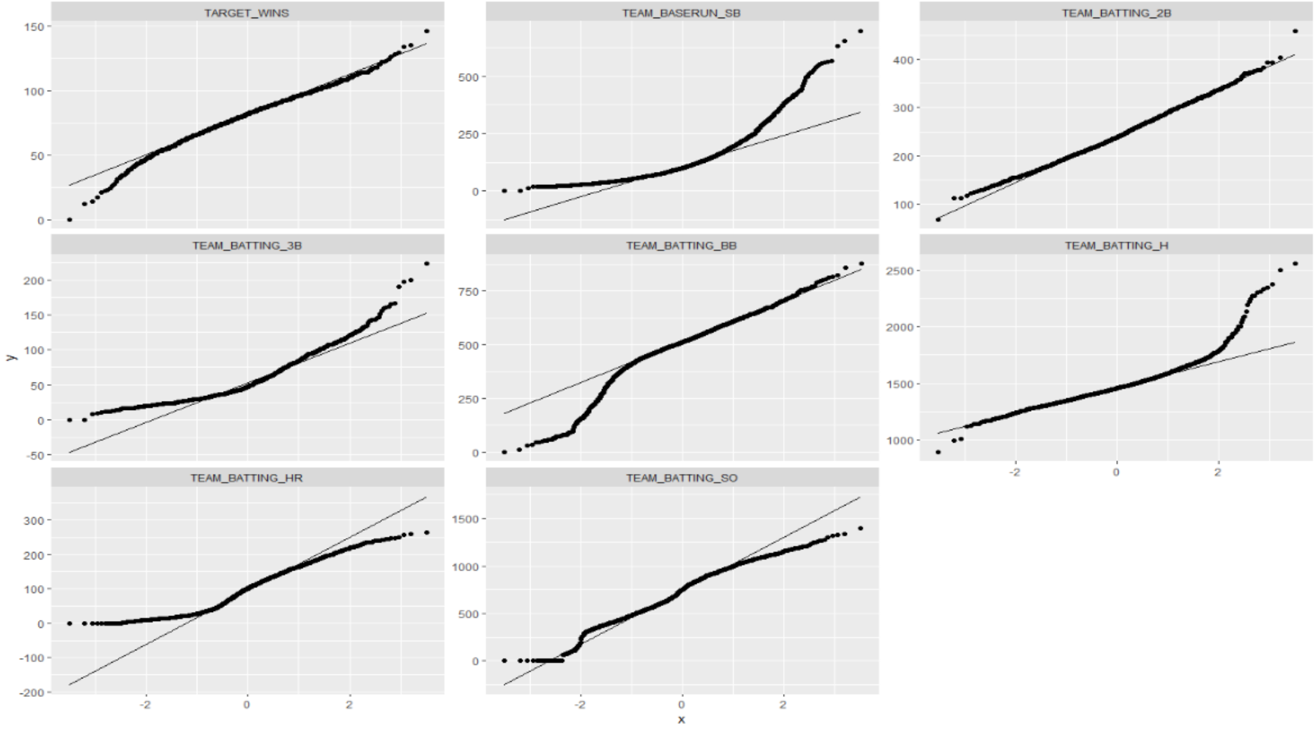
	Overall (N=2276)
TEAM_BASERUN_CS	
- Mean (SD)	52.80 (22.96)
- Median (Q1, Q3)	49.00 (38.00, 62.00)
- Range	0.00 - 201.00
- N-Miss	772
TEAM_BATTING_HBP	
- Mean (SD)	59.36 (12.97)
- Median (Q1, Q3)	58.00 (50.50, 67.00)
- Range	29.00 - 95.00
- N-Miss	2085
TEAM_PITCHING_H	
- Mean (SD)	1779.21 (1406.84)
- Median (Q1, Q3)	1518.00 (1419.00, 1682.50)
- Range	1137.00 - 30132.00
TEAM_PITCHING_HR	
- Mean (SD)	105.70 (61.30)
- Median (Q1, Q3)	107.00 (50.00, 150.00)
- Range	0.00 - 343.00
TEAM_PITCHING_BB	
- Mean (SD)	553.01 (166.36)
- Median (Q1, Q3)	536.50 (476.00, 611.00)
- Range	0.00 - 3645.00
TEAM_PITCHING_SO	
- Mean (SD)	817.73 (553.09)
- Median (Q1, Q3)	813.50 (615.00, 968.00)
- Range	0.00 - 19278.00
- N-Miss	102
TEAM_FIELDING_E	
- Mean (SD)	246.48 (227.77)
- Median (Q1, Q3)	159.00 (127.00, 249.25)
- Range	65.00 - 1898.00
TEAM_FIELDING_DP	
- Mean (SD)	146.39 (26.23)
- Median (Q1, Q3)	149.00 (131.00, 164.00)
- Range	52.00 - 228.00
- N-Miss	286

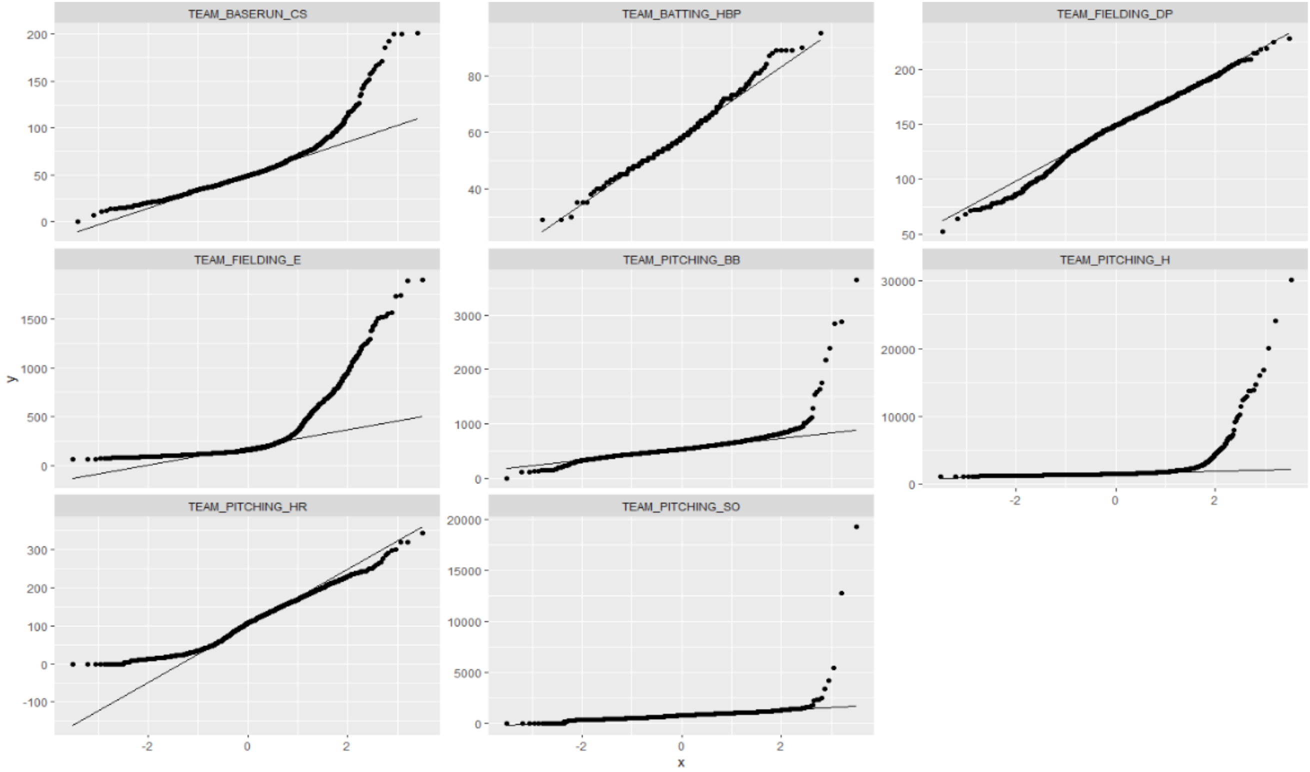
2.2 Distributions

The graphics below allow us to explore the distribution of our values in our training data. We can now more clearly explore the distributions of our variables and notice any skewing in our data more clearly.





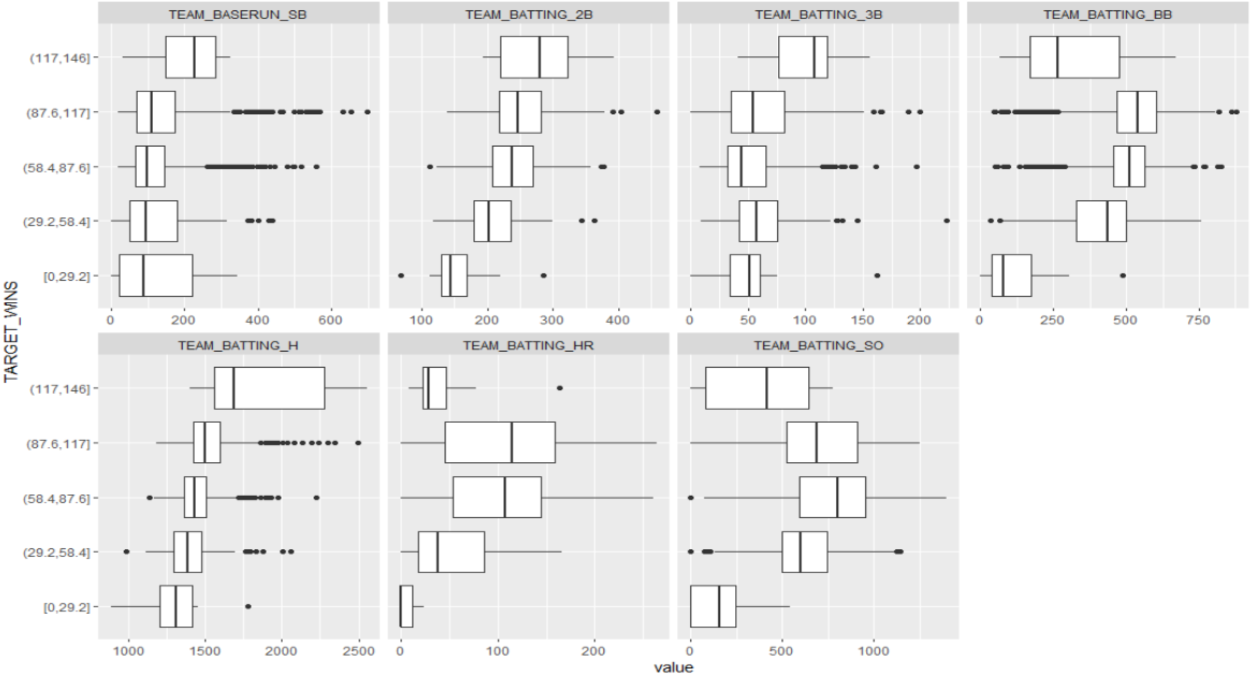

TARGET_WINS
150
100
50
0
TEAM_BASERUN_SB
500
250
0
TEAM_BATTING_2B
400
300
200
100
TEAM_BATTING_3B
200
150
100
50
0
-50
y
TEAM_BATTING_BB
750
500
250
0
TEAM_BATTING_H
2500
2000
1500
1000
TEAM_BATTING_HR
300
200
100
0
-100
-200
TEAM_BATTING_SO
1500
1000
500
0
-2
0
2
x

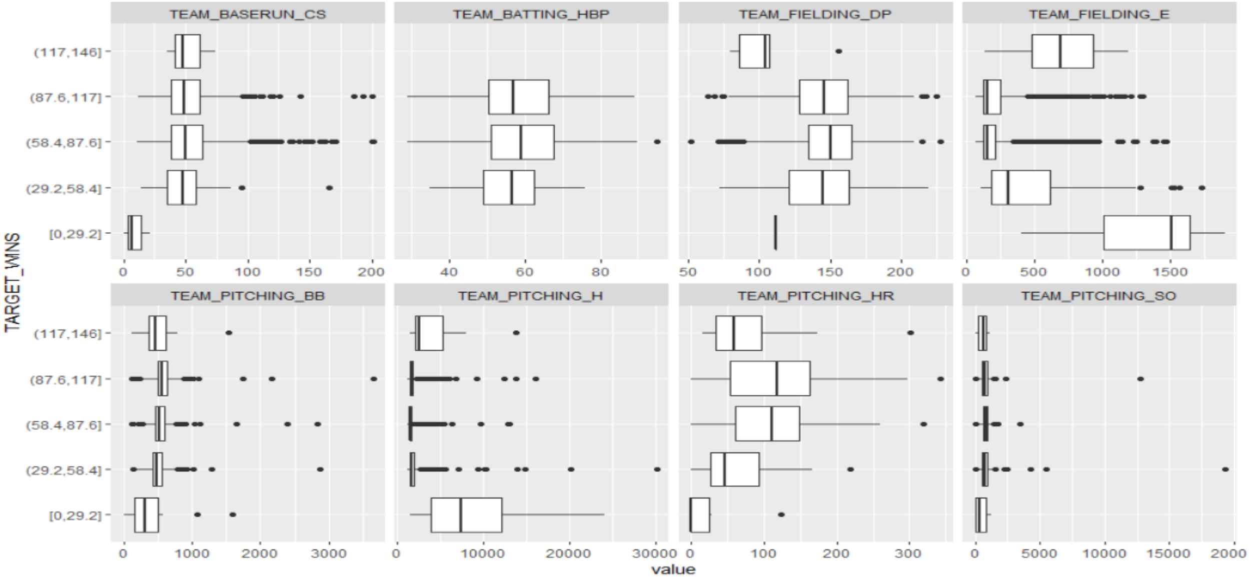

TEAM_BASERUN_CS
200
150
100
50
0
TEAM_BATTING_HBP
80
60
40
TEAM_FIELDING_DP
200
150
100
50
TEAM_FIELDING_E
1500
1000
500
0
y
TEAM_PITCHING_BB
3000
2000
1000
0
TEAM_PITCHING_H
30000
20000
10000
0
TEAM_PITCHING_HR
300
200
100
0
-100
TEAM_PITCHING_SO
20000
15000
10000
5000
0
-2
0
2
x

2.3 Boxplots

As shown below, we can begin exploring the trends in our data when split up into different win groupings. We first notice that the two groups with wins above and below what should be possible both have weird distributions relative to the other groups.

We also make note here, that the our main variables of interest, the variables showing number of bases scored by batters, appear to be relatively normal in distribution when they are split up by group like this.





2.4 Missing Values

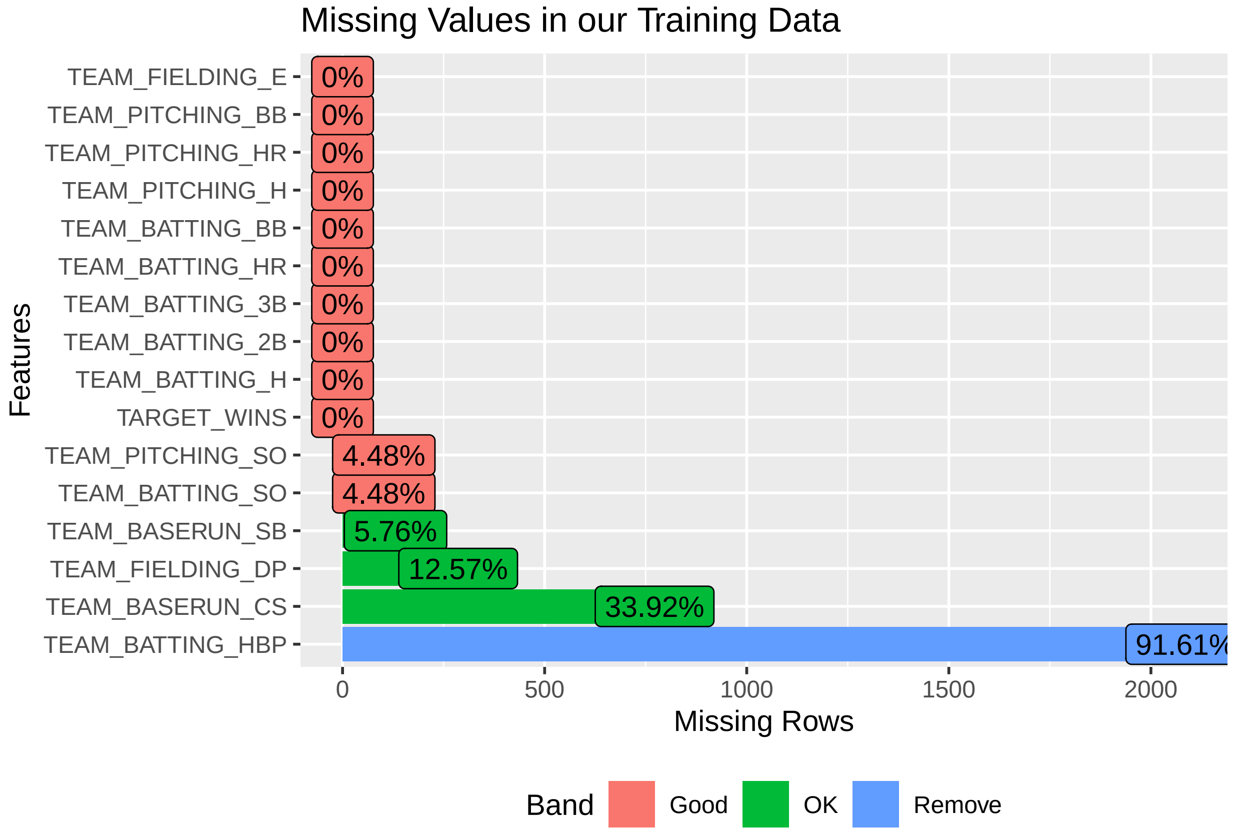
We can see in the figure below that quite a few of the variables in both the training and evaluation set have missing values. The two primary methods to handle these missing values are imputation or data removal. There are couple options to deal with these missing values.

- remove rows containing missing values
- replace missing values with an average value
- replace missing value with a median value

From our missing values chart, we can see that there are 91% missing values in TEAM_BATTING_HBP variable, 34% missing values in TEAM_BASERUN_CS and so on.

The data also showed a few variables with the value of "zero", which made us question their validity, but after further investigation we concluded that these values are legitimate. For instance, we have 20 zero values for batting and pitching strikeouts. It is reasonable to assume that a team had zero pitching strikeouts or that a team had zero batting strikeouts during a season, and in a span of 136 years it is not surprising to see a few teams with zero strikeouts.

What is interesting though is that we see a correlation between the number of zero strikeouts from both the batting and pitching performance. We can also observe the same for the variables right below in batting and pitching home runs, as well as walks by batters and walks allowed. This could just be a coincidence but there is no way of telling with the data provided to us as we can't see which statistics belong to what team. However, knowing or not knowing this information should not affect our analyses or the performance of our models, thus we decided to keep these values for the analyses.



3 Data Preparation

After the data exploration phase was completed, we transitioned into preparing the data for our regression models.

Given the number of data abnormalities, we needed to incorporate some data transformations prior to imputation. First we dropped the INDEX and TEAM_BATTING_HBP variables as they did not serve as useful inputs for our models. The rationale is that the INDEX variable provided no valuable information, and the TEAM_BATTING_HBP was missing nearly all of its data points.

We also dropped variables that exhibited large amounts of skew and nonlinearity. We determined that they had lower correlation with the dependent variable and would also cause the models to inaccurately predict TARGET_WIN values. Skew and nonlinear data points can cause a large amount of variability in a regression model. Thus, we decided that if they had no application to the models ability to interpret the dataset it was better to remove them from the process.

After applying these transformations, we strategized on how to handle the missing data points. We assumed that the missing data was missing at random and chose to impute the data points using the predictive mean matching (PMM) algorithm in the MICE library. The reason that we chose this method rather than replacing it with the mean or median is because of the large number of missing values. The PMM methodology allowed us to maintain the underlying structure of the existing data variation while reducing the possibility for implausible values being calculated. It is our view that this method reduced the risk of further skewing the values of our variables while respecting the heteroscedastic nature of the dataset.

4 Model Building

Following the data preparation phase, we brainstormed how best to construct an appropriate model design process. Our initial step was to partition the transformed dataset with the imputed data points from our PMM methodology. This was done in order to use 70% of it in the models and then evaluate their performance with the predictions against the remaining 30%.

Following the partition, we constructed a saturated regression model which contained all variables of the transformed dataset. This gave us a starting point to analyze the statistical significance of each variable and their associated correlations to the dependent variable.

Using the backward elimination methodology we were able to discern variables that did not contribute to our regression model. Producing a statistical summary of our saturated model we can see that four variables had p-values that were statistically insignificant. The variables are as follows:

- TEAM_BATTING_2B
- TEAM_BASERUN_CS
- TEAM_PITCHING_HR
- TEAM_PITCHING_BB

To confirm these p-values, we conducted a hypothesis test to determine whether the null hypothesis of each coefficient was equal to zero. The result of the test indicated that the p-value of the F-test was 0.1262. This allowed us to conclude that the null hypothesis cannot be rejected.

We also conducted a permutation test which resulted in similar findings to our statistical summary. The permutation test results were similar to the normal theory-based value of 2.22e-16. This conclusion allowed us to confirm that the 4 predictors, which we deemed statistically insignificant, had weak relationships to the response variable. Based on these findings we removed these four variables and created our first linear regression model.

For our second model we referred to the summary result from model 1. We saw that TEAM_BATTING_BB was the least strongest relationship with the response variable as compared to the rest of predictors. We drop this predictor to build our second model.

In order to construct our final model we analyzed the correlations associated with our saturated model, first model, and second model. Based on the analysis we were able to remove the variables TEAM_BASERUN_SB, TEAM_PITCHING_SO, and TEAM_BATTING_3B as their correlation to TARGET_WINS was nearly 0 when running the correlations on model 1. This allowed us to infer that the impact from these variables were negligible. From a real world perspective this makes sense as stolen bases and triples are rarer events, as teams who can consistently put players on the bases through singles and doubles seemingly have greater wins. However, we could not conclude model superi

5 Selecting Models

To evaluate our models we constructed a residuals plot, QQ plot, and also applied the Shapiro-Wilk normality test. This allowed us to check whether the residuals were normal. Models one and two met the required conditions of linearity, nearly normal residuals, and constant variability of the least squares regression. The Shapiro-Wilk normality test also proved the normality of the data in those two models with a p-value greater than 0.05.

The slight variations in models 1 and 2 was the kurtosis and skew in the tails. However they did not appear to be extreme deviations from the mean prompting us to further explore other performance metrics. Model three appeared to have the least normal residuals.

Our second test of our models' performances was the RMSE test. The corresponding visualizations of the RMSE test indicated that the model 1 had the lowest RMSE. However, the variation in the models was minimal indicating that model 1 only had slightly fewer errors than the other two models. This required a test of other key metrics to confirm.

Our final test analyzed the min-max accuracy, mean absolute percentage error, mean standard error, and mean absolute error. Now these tests produced consistent results with Model 1 having the greatest performance accuracy and the best scores on the rest of the metrics. However, the differences in each test were marginal. Thus, the question arose, how do you select the best model of the three? Given the higher accuracy in model one, the normality of the residuals, and the inclusion of additional variables we felt that model one would best handle variability in new datasets. The marginal differences in the rest of the performance metrics indicated that model one was nearly in line with the rest of the models and despite a higher level of kurtosis in the tails the residuals were nearly normal. In addition, we felt that the additional variables would capture the volatile nature of baseball and be most applicable to real world dynamics.

Selecting model 1 as the primary model, we constructed a prediction set from the evaluation dataset and saved this to a CSV file. The corresponding results will be provided with this report on our GitHub page for this assignment.

6 Conclusion

This assignment had us work with an interesting set of data, with many odd values. From the impossible number of wins and vast number of missing values in our data, this assignment was both a challenge and an excellent learning experience. Every member in our group had to also overcome the fact that none of us knew much of anything when it comes to baseball.

Given our current limited knowledge on feature selection and model building, we feel that our approach was done very well. However, we could have possibly explored the relationships between our variables more and used that to create new combinations of our values to better understand the overall shape of the data. If we had more time, we would love to explore any kind of new method that would improve our model.

Overall, we are happy with our final results and look forward to seeing how well our model performs. There is still much to learn and improve on, but this concludes the work we did for this assignment.