Keeno Glanville  
Homework1 Analysis of Baseball

**DATA EXPLORATION**

Within this dataset there are 2276 observations of 16 variables. The main focal point of this data is that we want to predict the target wins that a team will have over a given parameters.

A table with text and images

Description automatically generated with medium confidence

To first attack the dataset there was some basic cleaning to remove the unnecessary naming within the columns. We then did some exploration summary of each column as well as the missing values within each. (ALL ACTIONS DONE TO TRAINING SETS DONE TO TESTING). These missing values were eventually imputed utilizing the MICE package.

A graph with numbers and lines

Description automatically generated

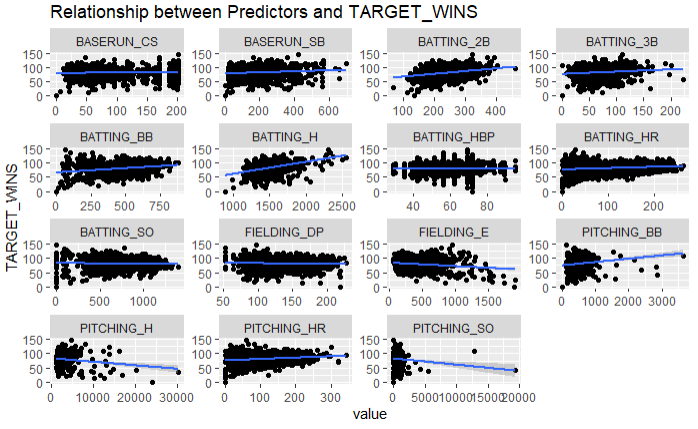
**DATA PREPARATION**

In the preparation of the data for analysis I utilized various techniques to best decide how I would proceed with my analysis. One of these methods included creating a correlation heat map that would be essential in allowing me to better understand the data. This is significant because it would allow me to make a more informed decision as to what type of model I would create on the data set.

A diagram of a number of points

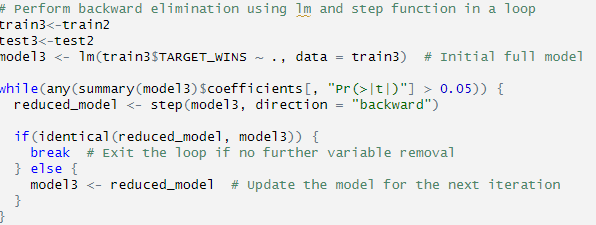
Description automatically generated with medium confidence

In continuation of preparation, I utilized plots of all the variables against the target variable to see any specific linear relationships between them. Overall, there weren’t very much direct linear relationships.



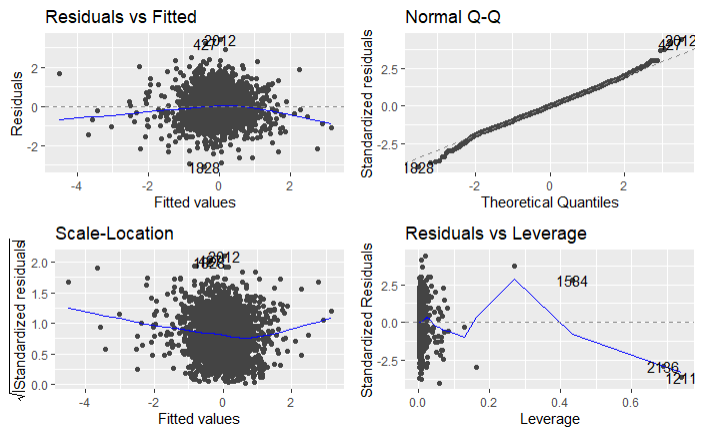
BUILD MODELS

To build the build the models I will go with three approaches. The first will be a basic approach that will give us a model that is not tampered with. The second model chosen was normalized as well as scaled. This would be able to give us a model that had the stronger assumptions of regressions. The final model would be one that incorporated backward propagation. This would be one that removed variables one at a time with p values > 0.05.



SELECT MODELS

The model selection here we will go with will be the second model. In terms of selecting a model we will always go for the best performance because that should give us the best results in real world scenarios. We don’t want to be biased in our decision as it could hinder us going further. What we notice in the model however is that we didn’t have a perfectly normal dataset through the residual plots. The Q-Q plot also showed various skewedness through the tail ends. Overall, this was similar throughout the models so through choosing the strong R-squared value we selected the model with the strongest predictor of future variables.



**Appendix**

<https://rpubs.com/kglan/1079300>

https://github.com/kglan/MSDS/blob/main/DATA621/Assignment1/Assingment1.Rmd