

Analysis of Moneyball Dataset

Homework 1

Talha Muhammad | Data 621 | March 4th 2018

Contents

[Introduction and Overview 2](#_Toc507974848)

[Exploratory Data Analysis (EDA) 2](#_Toc507974849)

[Data Preparation 9](#_Toc507974850)

[Building Models 13](#_Toc507974851)

[Selecting Models 15](#_Toc507974852)

# Introduction and Overview

The table below summarizes the variables in the dataset. The dataset has some 2,200 observations with each record representing one baseball season from 1871 to 2006, with the statistics adjusted to a 162 game season

Table 1: Variable Definitions & Descriptions



Below we analyze the dataset beginning with the Exploratory Data Analysis (EDA), then prepare the dataset for modeling, develop the models and finally select the models.

# Exploratory Data Analysis (EDA)

*Describe the size and the variables in the moneyball training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren’t doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas.*

*a. Mean / Standard Deviation / Median*

*b. Bar Chart or Box Plot of the data*

*c. Is the data correlated to the target variable (or to other variables?)*

*d. Are any of the variables missing and need to be imputed “fixed”?*

Figure 1 show the target variable, the number of season wins for the baseball team. The figure indicates that the number of wins is slightly skewed to the left with some fat tails.

Figure 1: Season Wins (TARGET)

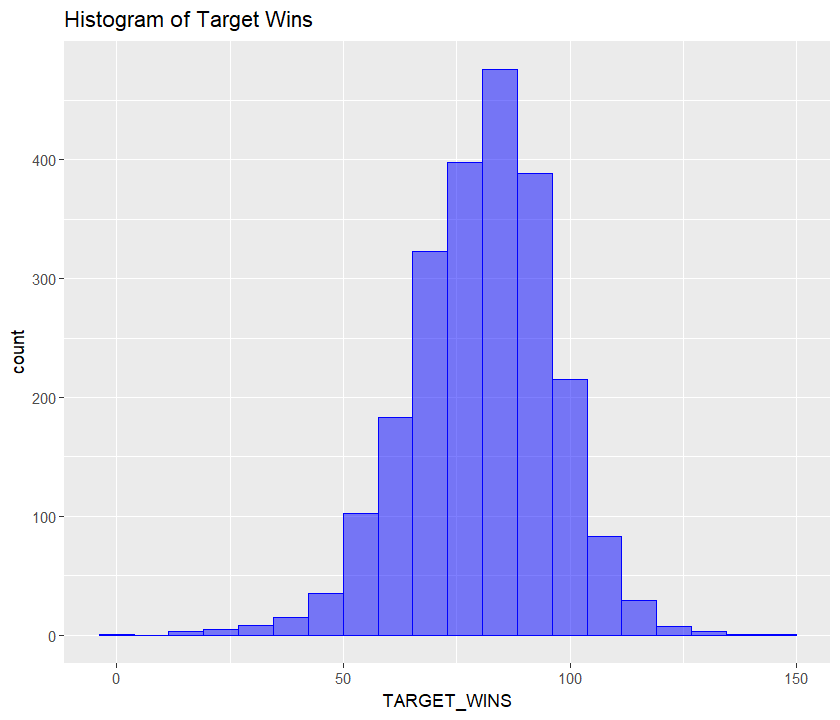


Table *1* below show a summary of the target variable. There are no missing values and the median number of wins is 82, with the maximum at 146.

Table : Target Summary



As part of the EDA analysis we want to understand the distribution of the data variables. Figure 2 shows histogram plots for each of the variables in the dataset. Most of the variables have normal like distributions, which are unimodal, while some appear to be bi-modal and exhibit varying degrees of skew. TEAM\_PITCHING\_H (hits allowed), TEAM\_PITCHING\_SO (strike out by pitchers) and TEAM\_PITCHING\_BB (walks allowed) have long tailed distributions.

Figure : Histograms of Variables

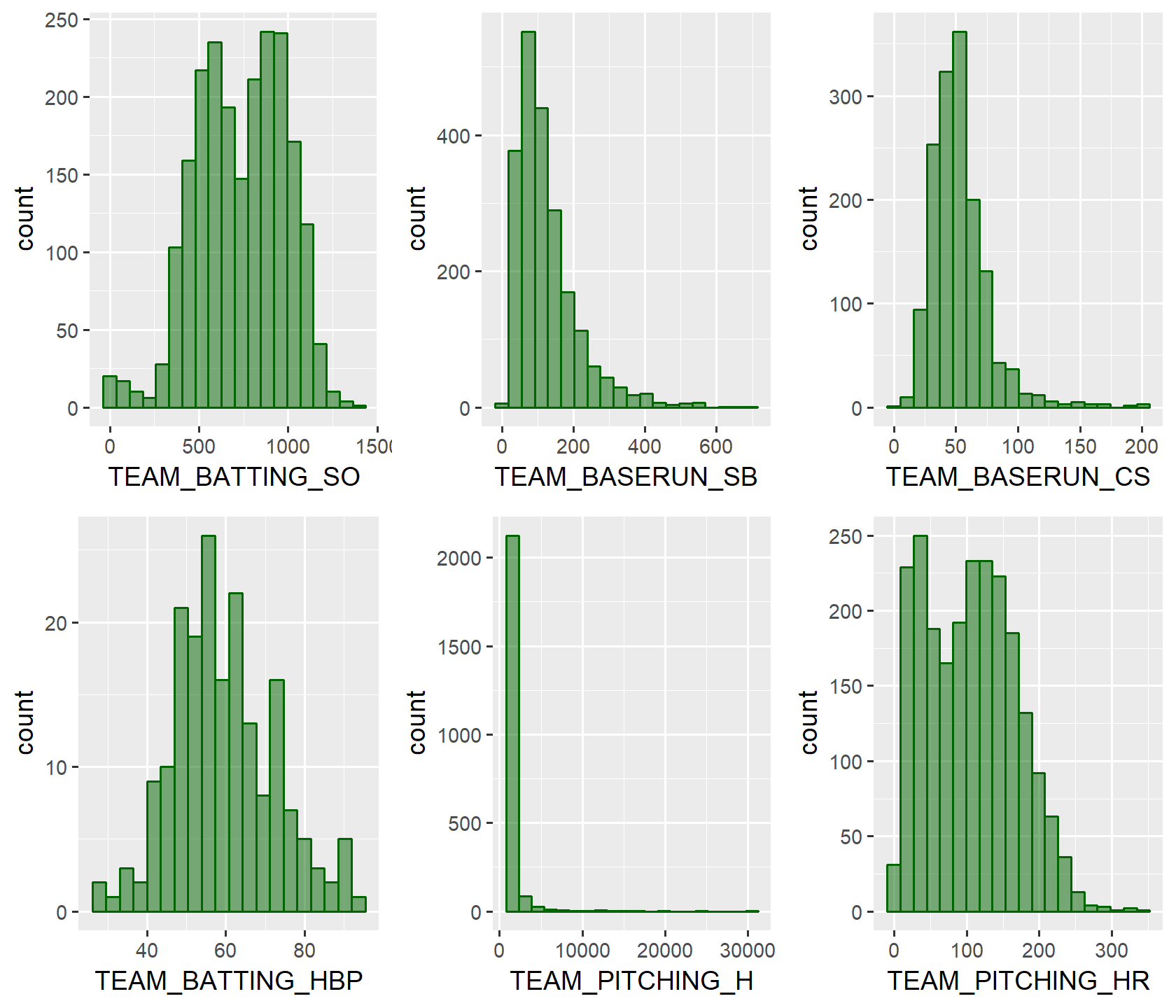
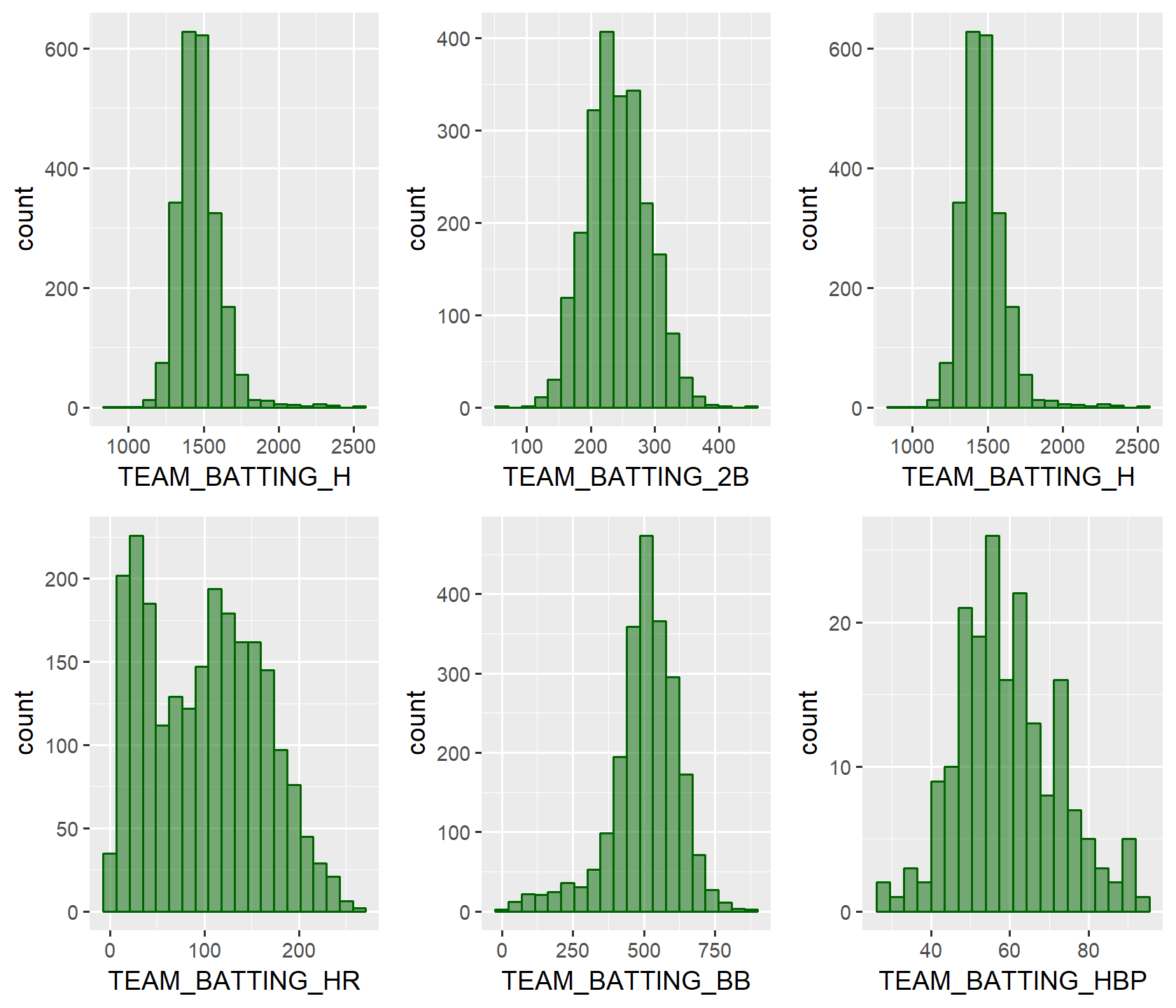


Figure : Histograms of Variables

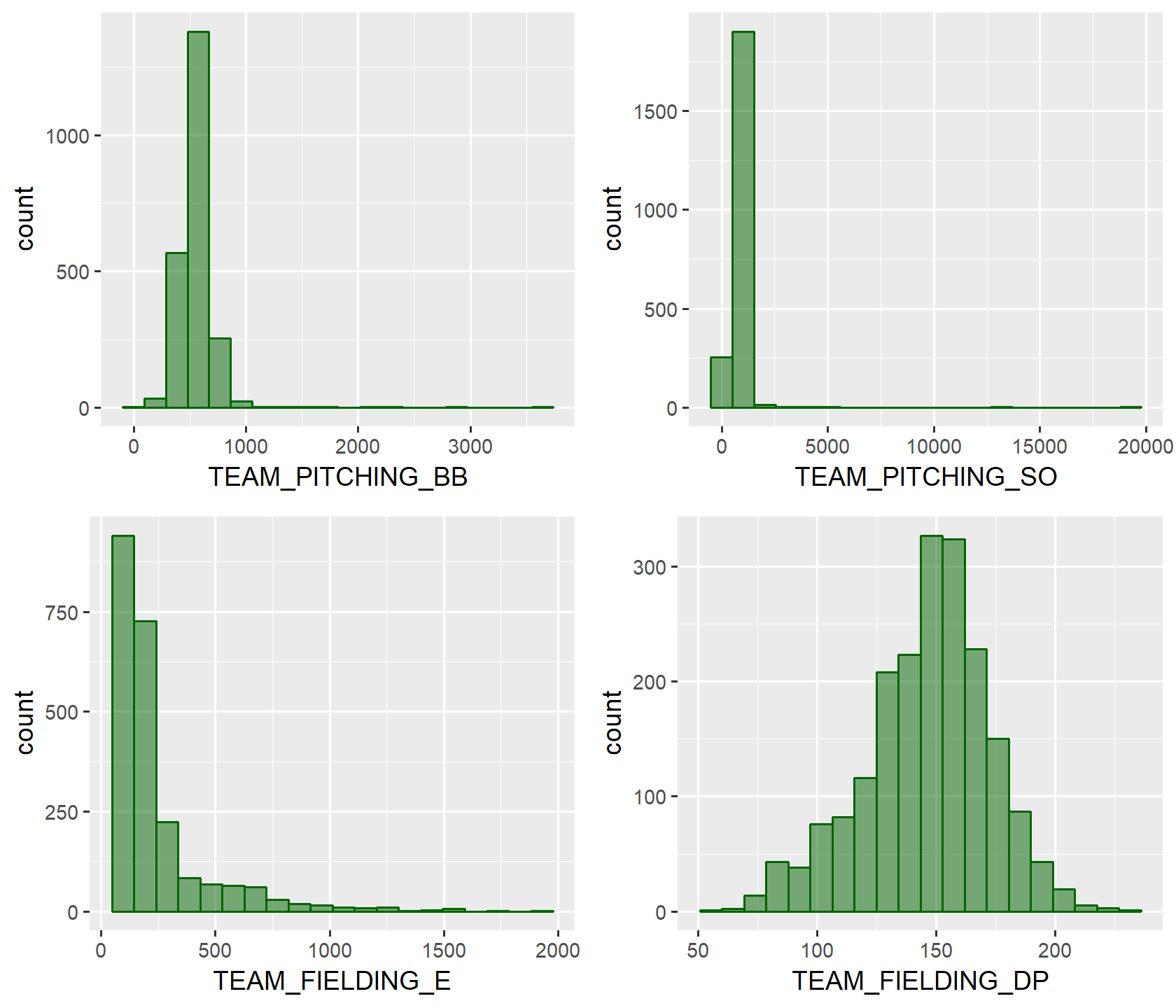


Figure 4 shows box plots of the dataset. TEAM\_BATTING\_H (base hits by batters) has a number of outliers. Figure 5 shows a number of variables have a large number of outliers.

Figure : Box Plot of Variables

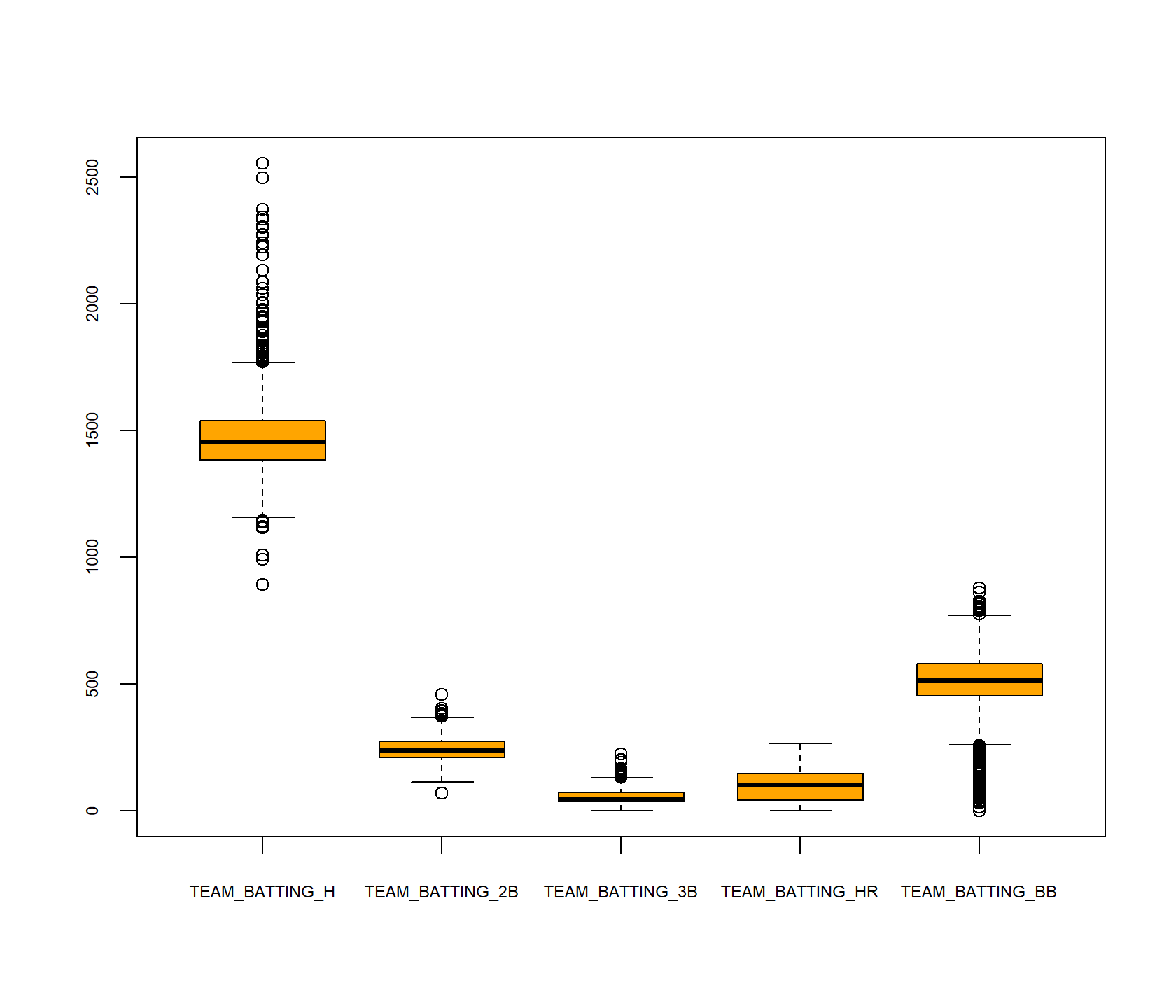


Figure : Box Plots of Variables

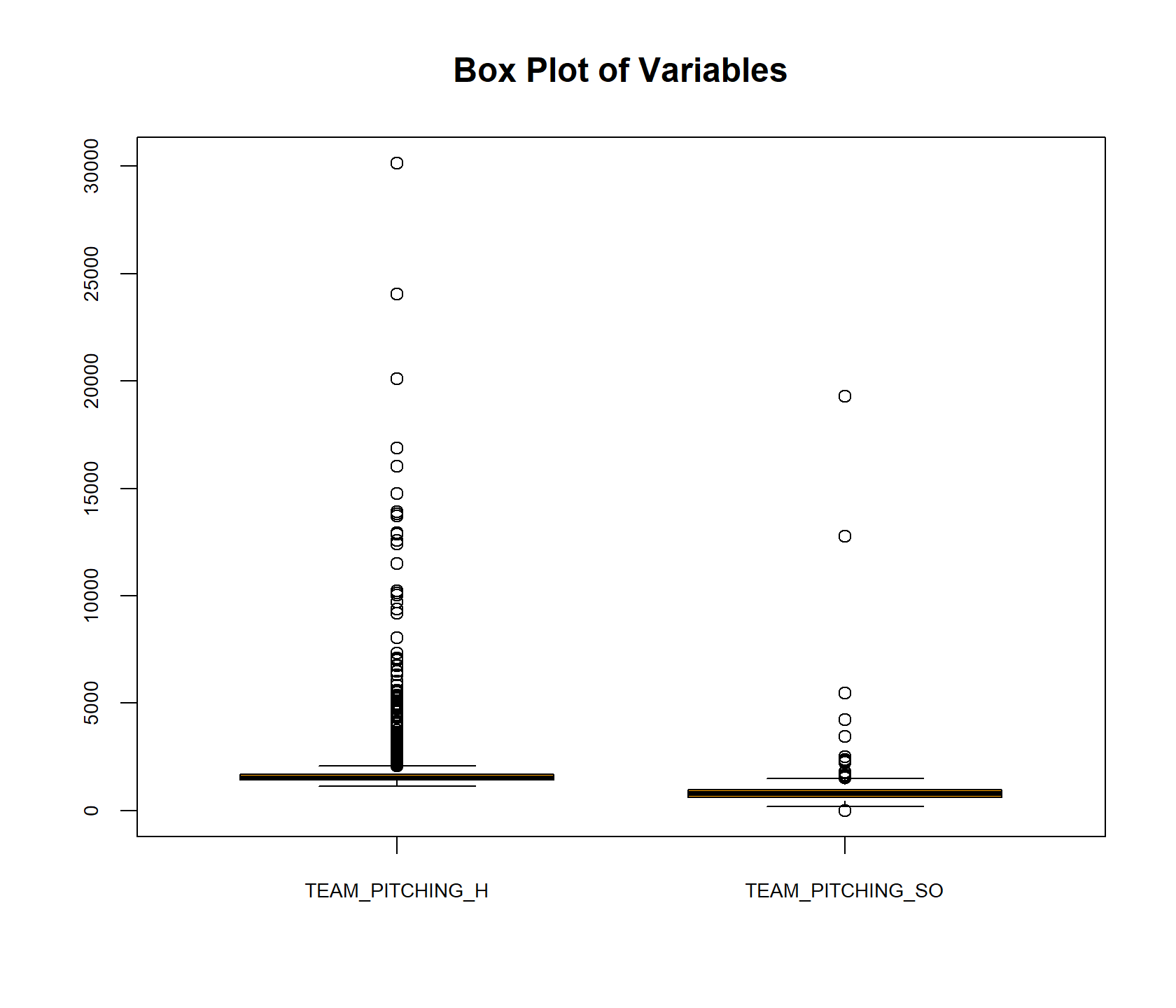
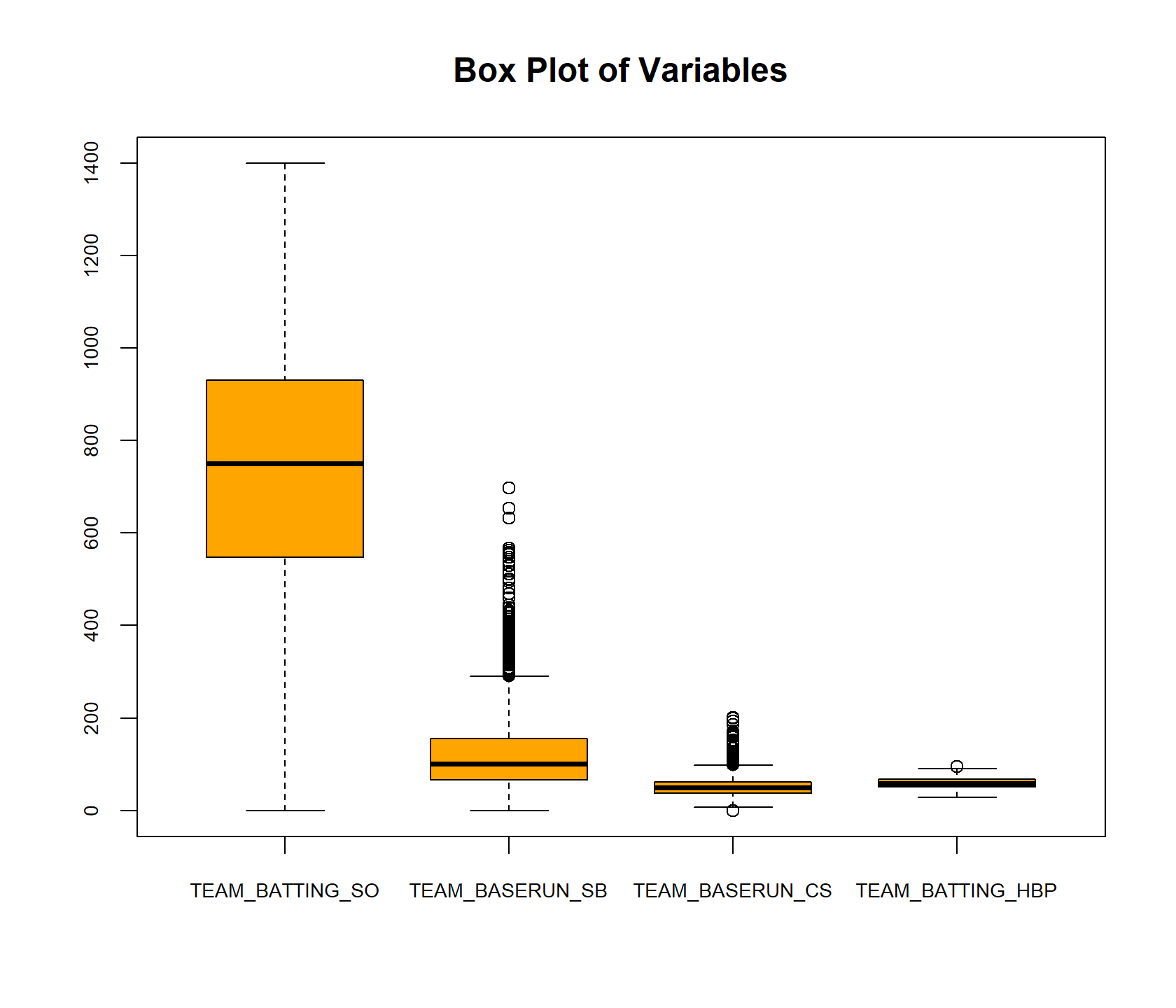


Figure : Box Plots of Variables

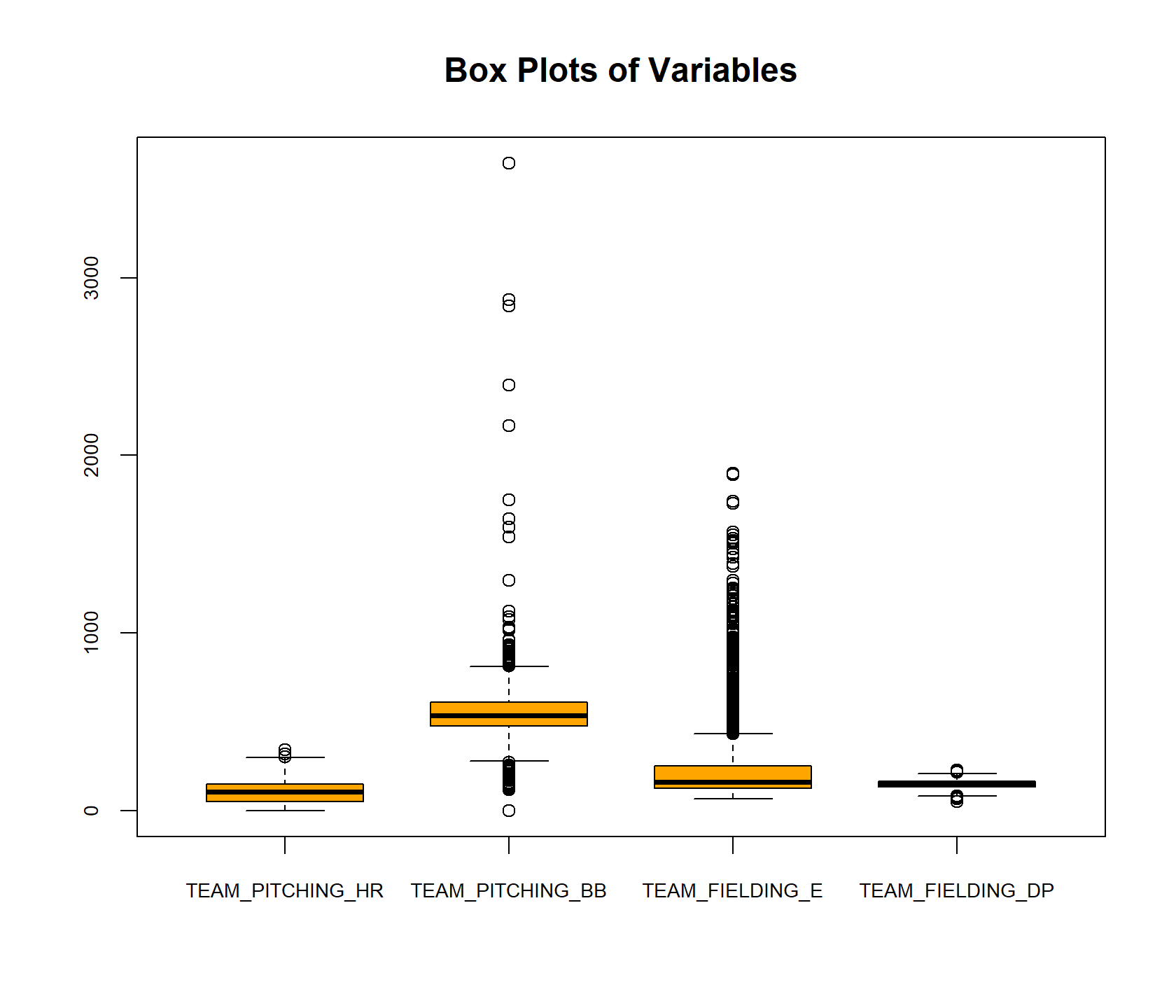


Figure 7 and Figure 8 shows pair plots of the data, versus the target. The pair plots show how the variables, relate the target and to each other. TEAM\_BATTING variables are all positively related to the target variables while other variables are not as strongly related to the target variable, for example TEAM\_BATTING\_HBP (batters hit by pitch).

Figure 7: Pair Plots of Dataset versus Target

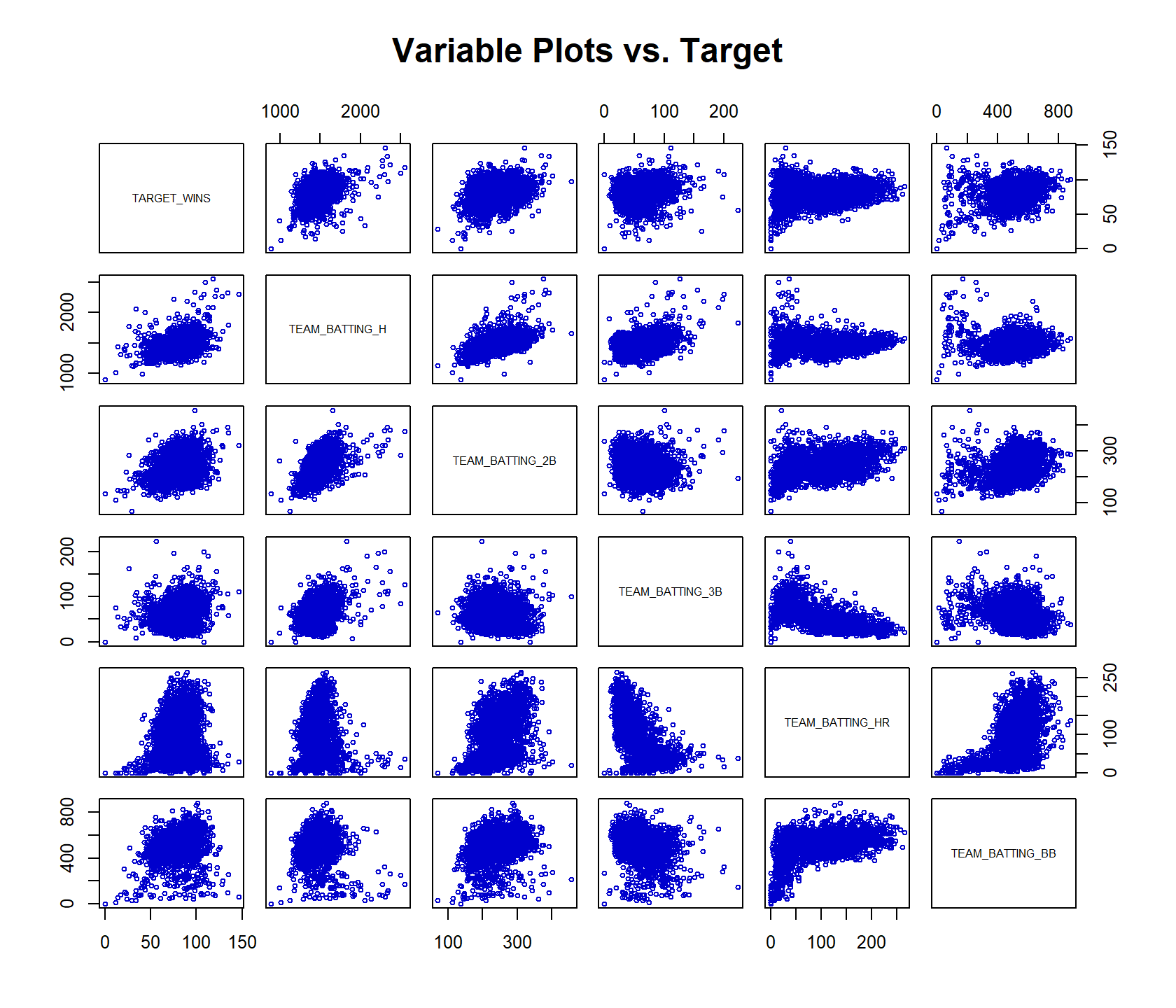


Figure 8: Pair Plots of the Dataset vs. Target

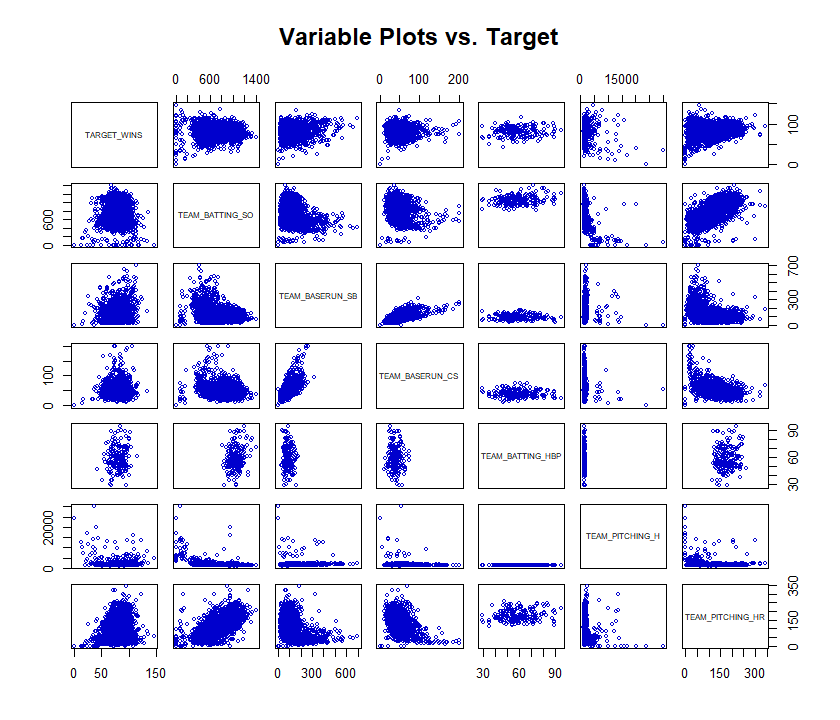
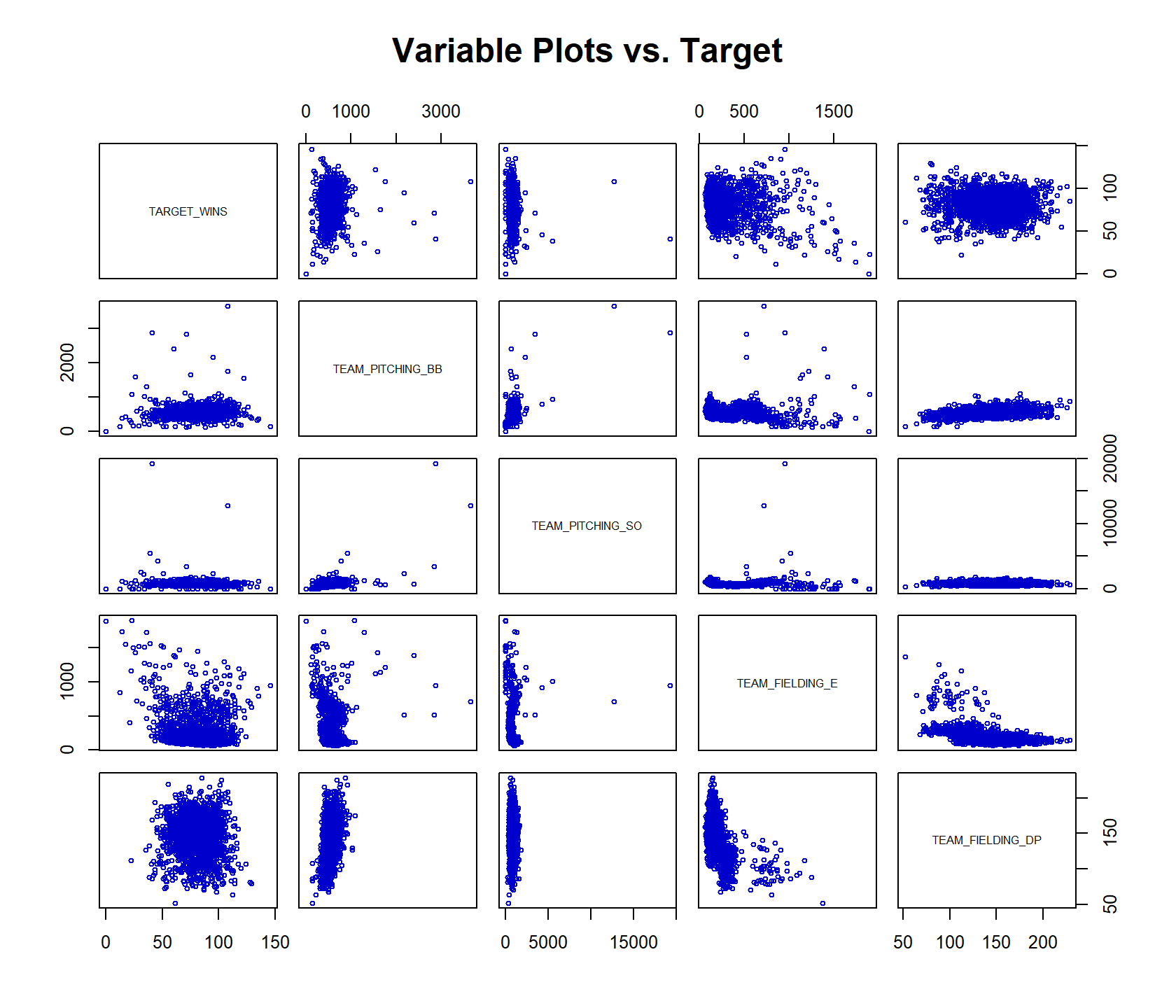


Figure : Variable Plots vs. Target



In the next section we develop the data for modeling purposes.

# Data Preparation

*Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations.*

*a. Fix missing values (maybe with a Mean or Median value)*

*b. Create flags to suggest if a variable was missing*

*c. Transform data by putting it into buckets*

*d. Mathematical transforms such as log or square root (or use Box-Cox)*

*e. Combine variables (such as ratios or adding or multiplying) to create new variables*

Table 3 presents a summary of the variables that have missing values. A number of variables have some missing values, with 3 variables less than 5 percent and 3 variables greater than 10 percent missing values. TEAM\_BATTING\_HBP has by far the most number of missing values, close to 80% values are missing.

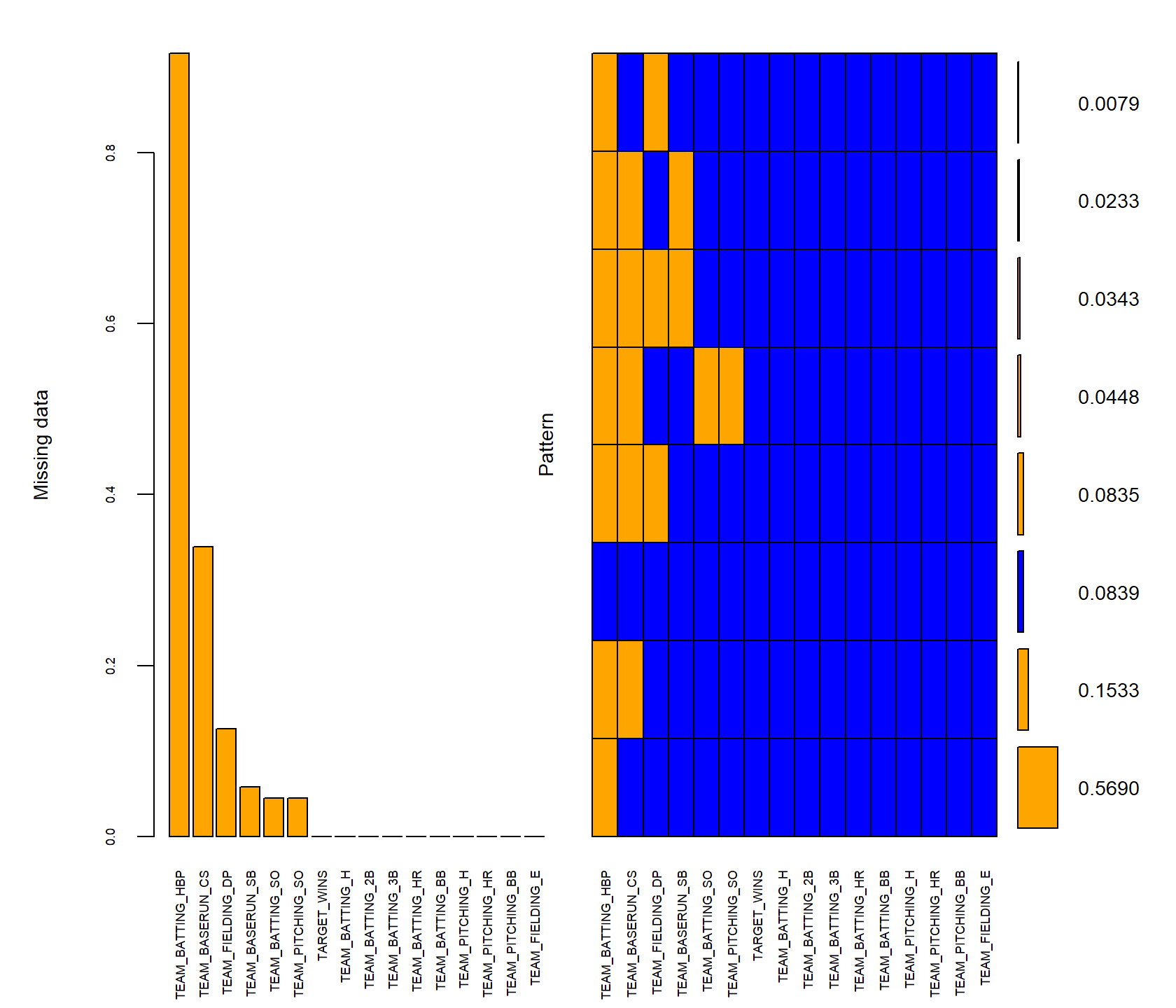
Table : Summary of Missing Value Variables





Figure 10 shows an alternate of way of visualizing the missing values in the dataset. The histogram plot on the left shows the portion of data missing for the individual variables while the plot on the right shows the proportion of missing variables in the dataset. Only 8% of the dataset has compete cases while about 57 percent has only TEAM\_BATTING\_HBP missing and an additional 15% of the dataset has both TEAM\_BATTING\_HBP and TEAM\_BATTING\_CS missing.

Figure : Visualizing Missing Data



As part of the data preparation process we impute missing values of the dataset. For variables, that have less than 10 percent missing values, we impute the median of the dataset. Figure 11 shows a comparison of the missing and imputed datasets for the variable TEAM\_BATTING\_SO where the imputed variables were through the median. For variables that had more than 10 percent missing values (3 variables) partial mean matching procedure was used through multiple imputation. Figure 12 shows the result of the mean matching for only the original dataset and the imputed values in the dataset. Overall the method provides a reasonable approach for imputing missing values. Other missing variables were also checked and they have reasonable imputed values. Using normal Bayesian based regression approaches for imputation were less successful as some of these were yielding negative values. As a result the partial mean matching approach was selected.

Figure : Imputing Data at the Median

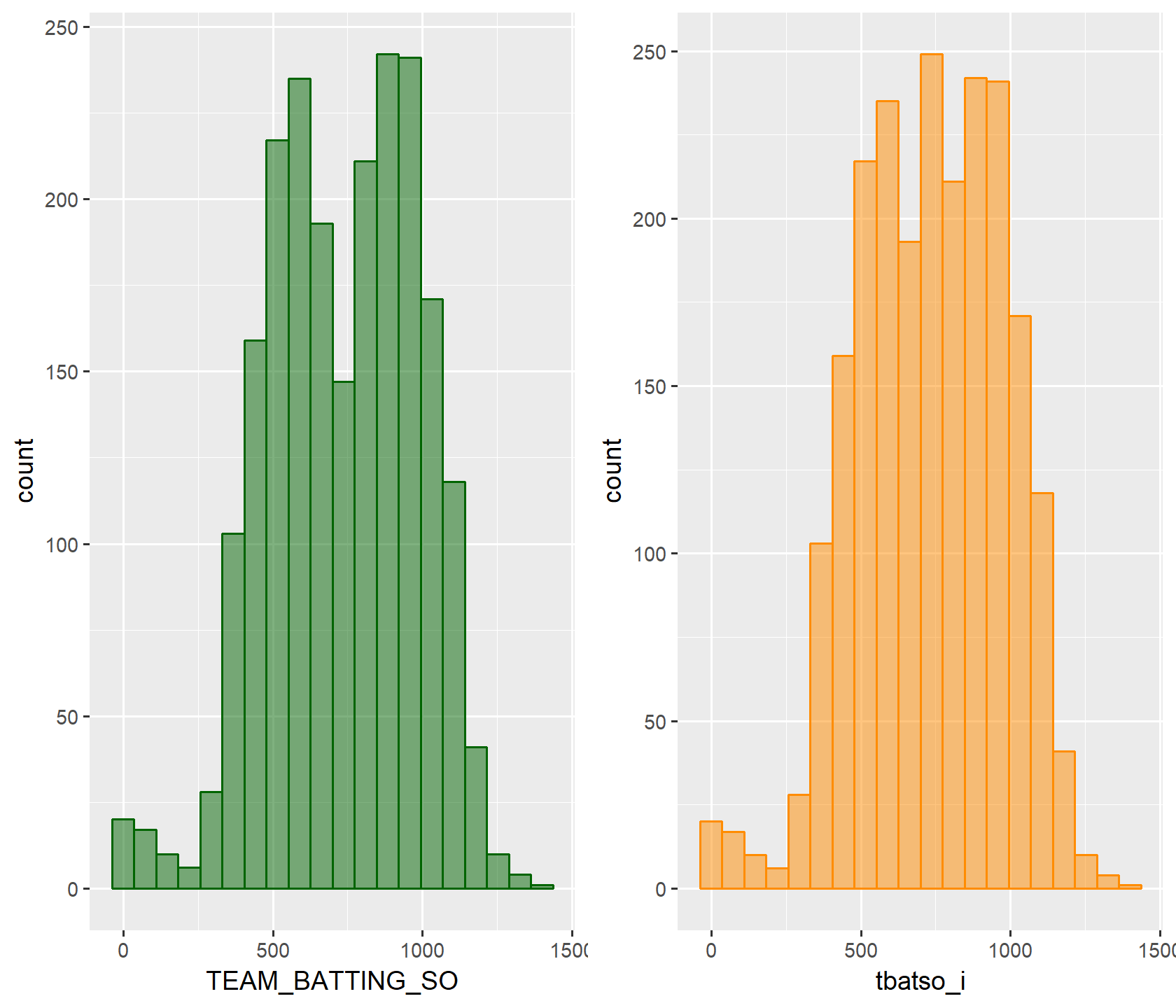
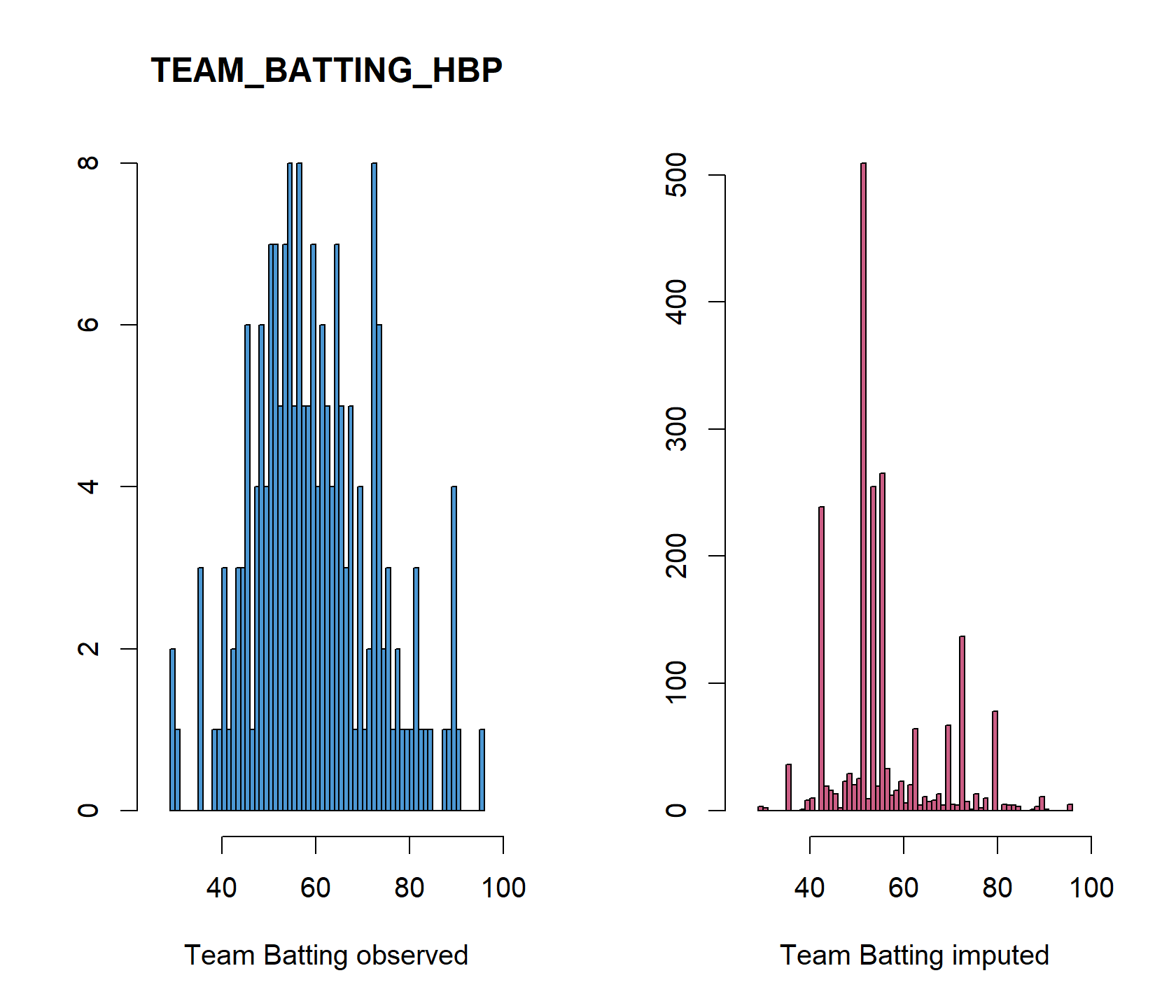


Figure : Multiple Imputation using Partial Mean Matching (TEAM\_BATTING\_HBP)



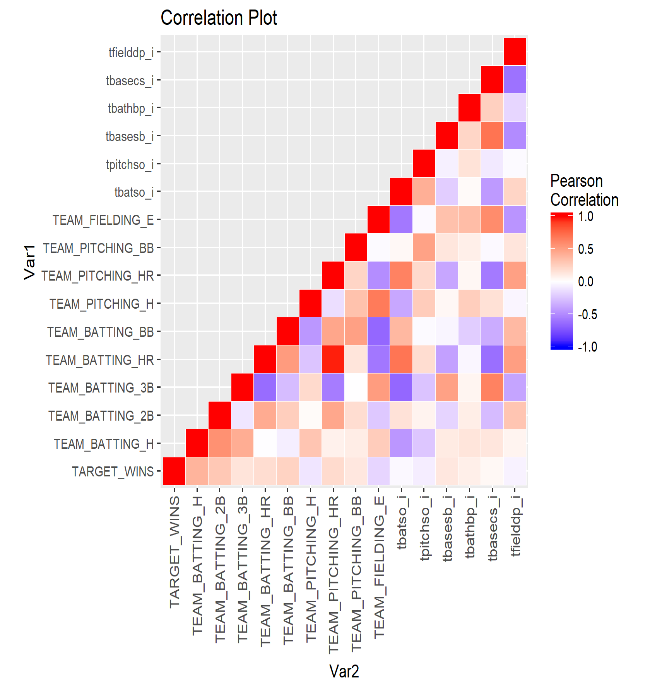
# Building Models

*Using the training data set, build at least three different multiple linear regression models, using different variables (or the same variables with different transformations). Since we have not yet covered automated variable selection methods, you should select the variables manually (unless you previously learned Forward or Stepwise selection, etc.). Since you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done.*

*Discuss the coefficients in the models, do they make sense? For example, if a team hits a lot of Home Runs, it would be reasonably expected that such a team would win more games. However, if the coefficient is negative (suggesting that the team would lose more games), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.*

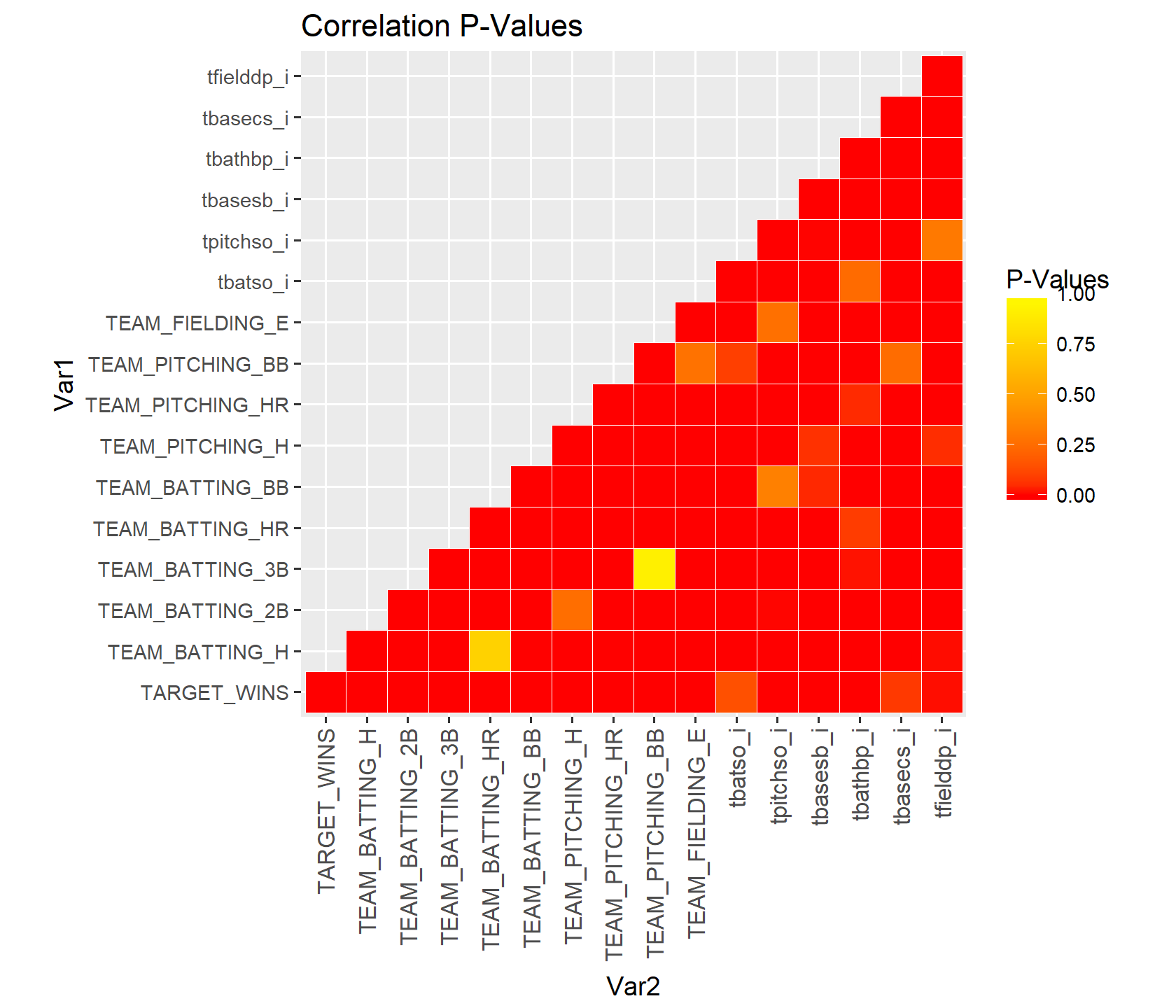
An important step in developing models is to understand correlations between the target variable and the independent variables. In addition, it allows the analyst to see correlations between the independent variables. For example TEAM\_BATTING\_H and TEAM\_BATTING\_2B have high correlations and likely to have multicollinearity issues with the model.

Figure : Correlation Plots



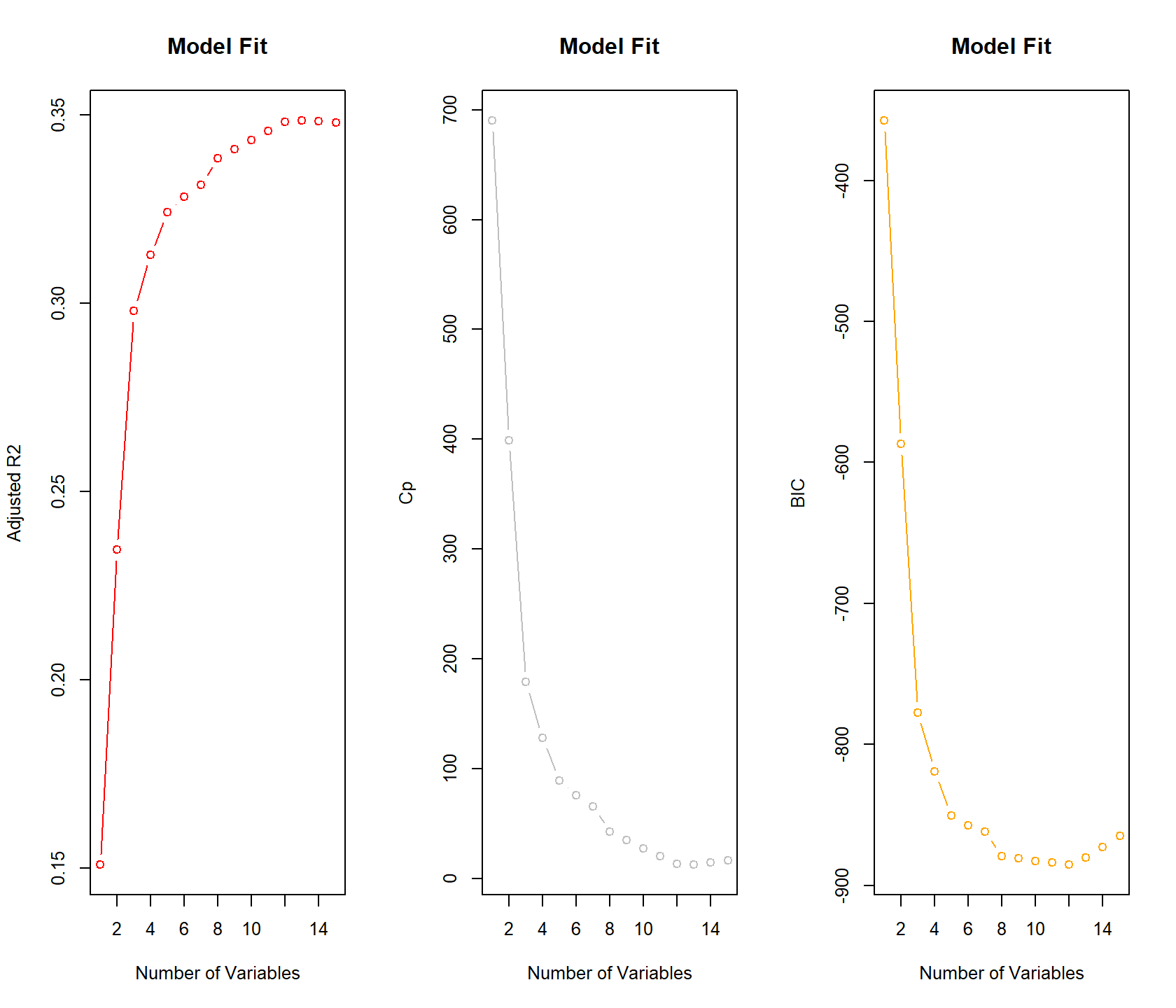
The figure below shows the correlation p-values for the dataset. Overall, most correlations are significant except for a few.

Figure : Correlation P-Values



With the imputed variables, we use forward selection algorithm using linear regression from the “leaps” library in R. The resulting model fits are shown below, with a number of important statistics summarized such as adjusted R-squared and BIC. The results indicate the models with 4 or more variables tend to do quite well, but overall R2 greater than 0.35 would be hard to achieve with the variables in the dataset.

Figure : Forward Selection Fit on All Variables



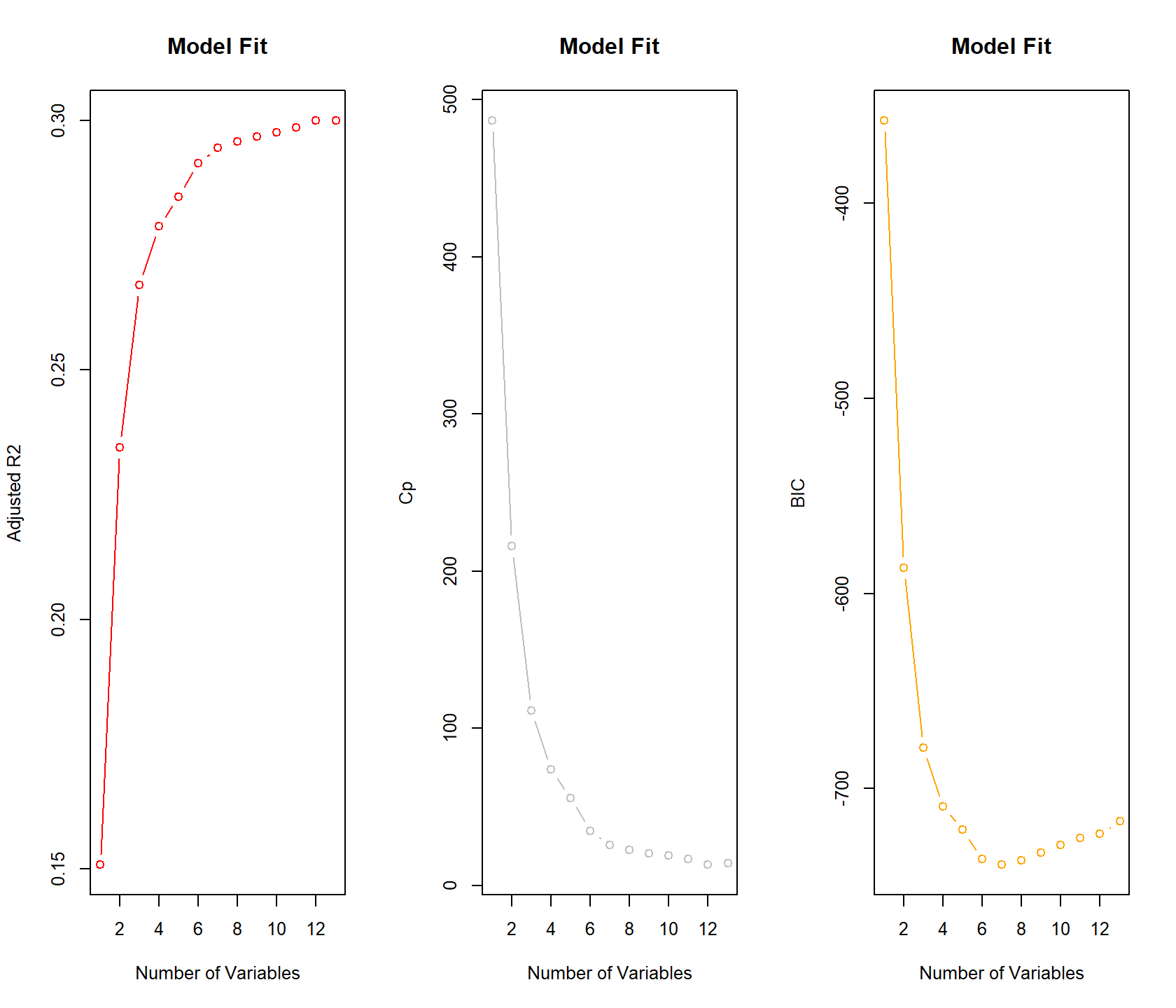
# Selecting Models

*Decide on the criteria for selecting the best multiple linear regression model. Will you select a model with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model.*

*For the multiple linear regression model, will you use a metric such as Adjusted R2, RMSE, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. Using the training data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R2, (c) F-statistic, and (d) residual plots. Make predictions using the evaluation data set.*

Figure 16 shows forward selection fit after removing variables with non-intuitive signs. The results show that the maximum R2 is now lower than before 0.30 with a four to five variable model sufficient for the results.

Figure : Forward Selection Fit After Removing Some Variables



From the forward selection models that are fitted two models are selected based on intuitiveness of estimated model coefficient and model fit. Model 1 has and adjusted R-squared of 0.279 while Model 2 has an R2 of 0.269.

Table : Model Assessment Summary

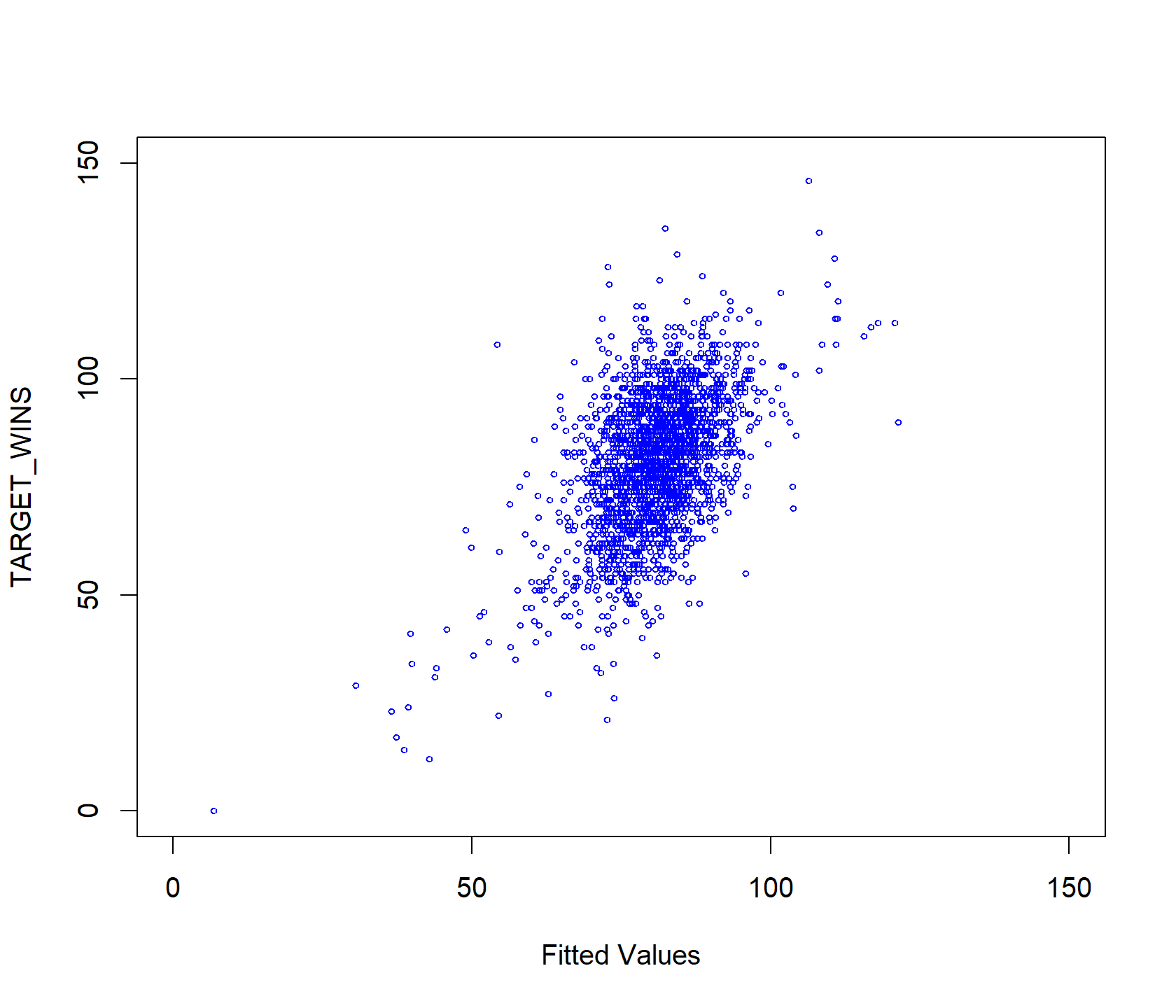
|  |  |  |
| --- | --- | --- |
| Model Name | Adjusted R Squared | F-stat |
| Model 1 | 0.279 | 220.8 |
| Model 2 | 0.269 | 210.3 |

Model 1 results are shown below. Variables have intuitive signs and significant t-statistics.

Table : Model 1 Results



Figure : Model 1 Fit



Model 1 validation plots are shown below. Overall the model fits well and the residuals seem to meet normality assumptions though of the residuals in the tail, seem to be less in-line with normality assumptions.

Figure : Model 1 Validation Plots

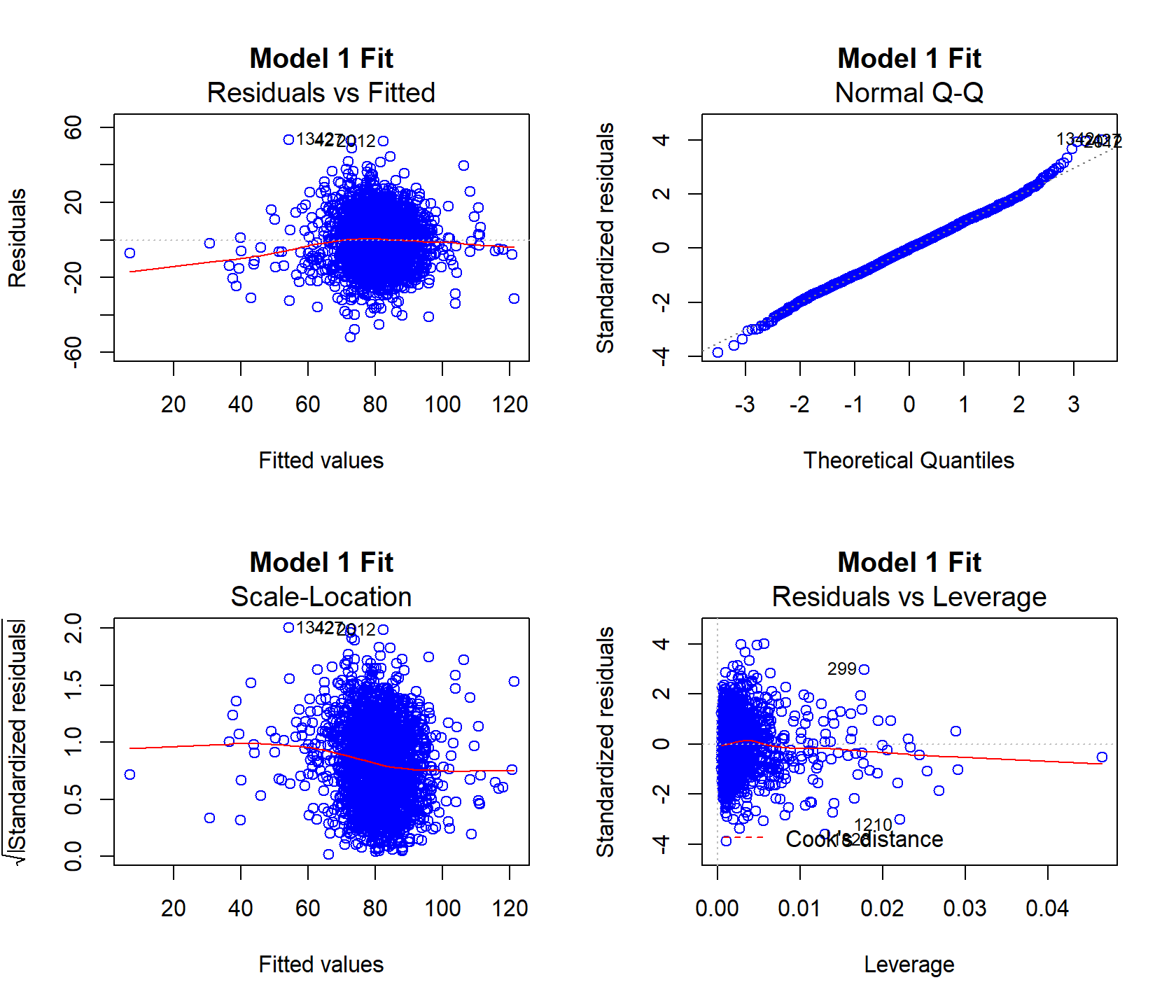


Table : Model 2 Results



Figure : Model 2 Fit

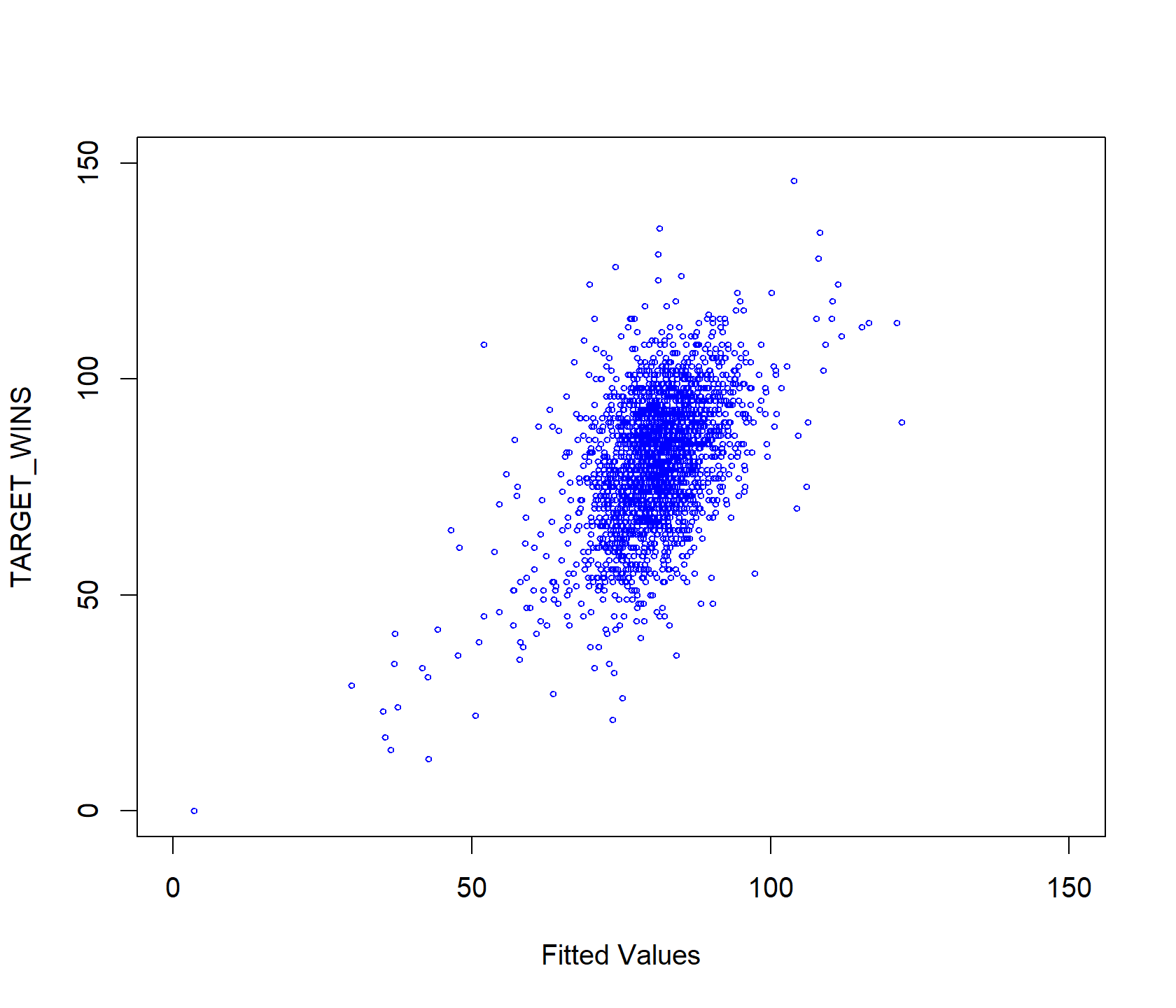


Figure : Model 2 Validation Plots

