Data621 - HW1

Devin Teran, Atina Karim, Tom Hill, Amit Kapoor

2/26/2021

Contents

```r
library(dplyr)
library(DataExplorer)
library(GGally)
library(ggplot2)
library(readr)
library(reshape2)
library(purrr)
library(tidyr)
library(corrplot)
library(MASS)
library(caret)
```

Data Overview

The data set contains approximately 2276 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.Below is a short description of the variables

- INDEX - Identification Variable
- TARGET_WINS - Number of 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_HBP - Batters hit by pitch (get a free base)
- TEAM_BATTING_SO - Strikeouts by batters
- TEAM_BASERUN_SB - Stolen bases
- TEAM_BASERUN_CS - Caught stealing
- TEAM_FIELDING_E - Errors
- TEAM_FIELDING_DP - Double Plays
- TEAM_PITCHING_BB - Walks allowed
- TEAM_PITCHING_H - Hits allowed
- TEAM_PITCHING_HR - Homeruns allowed
- TEAM_PITCHING_SO - Strikeouts by pitchers

Objective

To build a multiple linear regression model on the training data to predict *TARGET_WINS*, which is the number of wins for the team.

Data Exploration

```
# read data
baseball_df <- read.csv(’https://raw.githubusercontent.com/hillt5/DATA_621/master/HW1/moneyball-training
baseball_eval <- read.csv(’https://raw.githubusercontent.com/hillt5/DATA_621/master/HW1/moneyball-evalua

head(baseball_df)
```

```
##   INDEX TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
## 1     1          39           1445             194              39
## 2     2          70           1339             219              22
## 3     3          86           1377             232              35
## 4     4          70           1387             209              38
## 5     5          82           1297             186              27
## 6     6          75           1279             200              36
##   TEAM_BATTING_HR TEAM_BATTING_BB TEAM_BATTING_SO TEAM_BASERUN_SB
## 1              13             143             842              NA
## 2             190             685            1075              37
## 3             137             602             917              46
## 4              96             451             922              43
## 5             102             472             920              49
## 6              92             443             973             107
##   TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR
## 1              NA               NA            9364               84
## 2              28               NA            1347              191
## 3              27               NA            1377              137
## 4              30               NA            1396               97
## 5              39               NA            1297              102
## 6              59               NA            1279               92
##   TEAM_PITCHING_BB TEAM_PITCHING_SO TEAM_FIELDING_E TEAM_FIELDING_DP
## 1              927             5456            1011               NA
## 2              689             1082             193              155
## 3              602              917             175              153
## 4              454              928             164              156
## 5              472              920             138              168
## 6              443              973             123              149
```

```r
dim(baseball_df)
```

```
## [1] 2276   17
```

```r
summary(baseball_df)
```

```
##      INDEX         TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B
##  Min.   :   1.0   Min.   :  0.00   Min.   : 891   Min.   : 69.0  
##  1st Qu.: 630.8   1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0  
##  Median :1270.5   Median : 82.00   Median :1454   Median :238.0  
##  Mean   :1268.5   Mean   : 80.79   Mean   :1469   Mean   :241.2  
##  3rd Qu.:1915.5   3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0  
##  Max.   :2535.0   Max.   :146.00   Max.   :2554   Max.   :458.0  
##                                                                  
##  TEAM_BATTING_3B  TEAM_BATTING_HR  TEAM_BATTING_BB TEAM_BATTING_SO 
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.0   Min.   :   0.0  
##  1st Qu.: 34.00   1st Qu.: 42.00   1st Qu.:451.0   1st Qu.: 548.0  
##  Median : 47.00   Median :102.00   Median :512.0   Median : 750.0  
##  Mean   : 55.25   Mean   : 99.61   Mean   :501.6   Mean   : 735.6  
##  3rd Qu.: 72.00   3rd Qu.:147.00   3rd Qu.:580.0   3rd Qu.: 930.0  
##  Max.   :223.00   Max.   :264.00   Max.   :878.0   Max.   :1399.0  
##                                                    NA’s   :102     
##  TEAM_BASERUN_SB TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_PITCHING_H
##  Min.   :  0.0   Min.   :  0.0   Min.   :29.00    Min.   : 1137  
##  1st Qu.: 66.0   1st Qu.: 38.0   1st Qu.:50.50    1st Qu.: 1419  
##  Median :101.0   Median : 49.0   Median :58.00    Median : 1518  
##  Mean   :124.8   Mean   : 52.8   Mean   :59.36    Mean   : 1779  
##  3rd Qu.:156.0   3rd Qu.: 62.0   3rd Qu.:67.00    3rd Qu.: 1682  
##  Max.   :697.0   Max.   :201.0   Max.   :95.00    Max.   :30132  
##  NA’s   :131     NA’s   :772     NA’s   :2085                    
##  TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E 
##  Min.   :  0.0    Min.   :   0.0   Min.   :    0.0   Min.   :  65.0  
##  1st Qu.: 50.0    1st Qu.: 476.0   1st Qu.:  615.0   1st Qu.: 127.0  
##  Median :107.0    Median : 536.5   Median :  813.5   Median : 159.0  
##  Mean   :105.7    Mean   : 553.0   Mean   :  817.7   Mean   : 246.5  
##  3rd Qu.:150.0    3rd Qu.: 611.0   3rd Qu.:  968.0   3rd Qu.: 249.2  
##  Max.   :343.0    Max.   :3645.0   Max.   :19278.0   Max.   :1898.0  
##                                    NA’s   :102                       
##  TEAM_FIELDING_DP
##  Min.   : 52.0   
##  1st Qu.:131.0   
##  Median :149.0   
##  Mean   :146.4   
##  3rd Qu.:164.0   
##  Max.   :228.0   
##  NA’s   :286
```

```r
print(’Number of observations:’)
```

```
## [1] "Number of observations:"
```

```r
nrow(baseball_df)
```

```
## [1] 2276
```

```r
print('Observations per year, 1871 - 2006:')
```

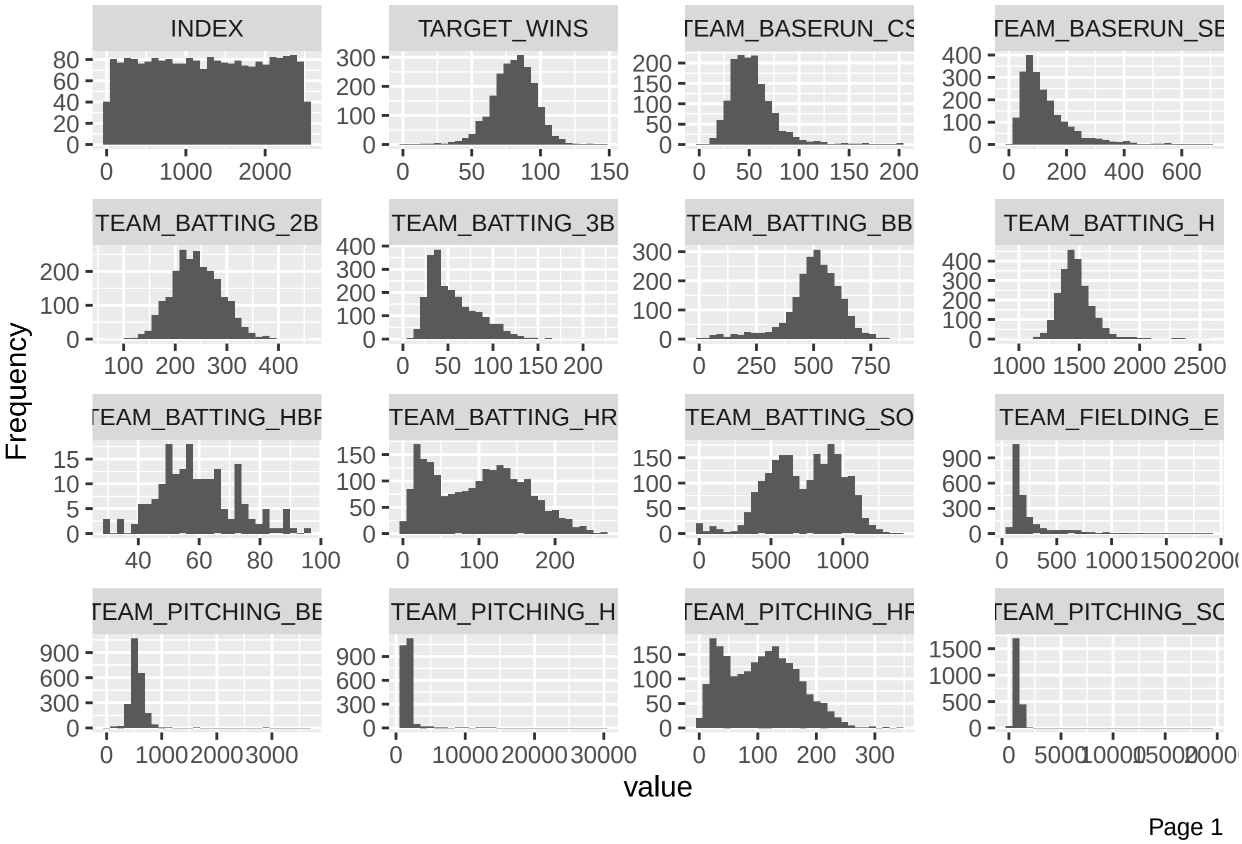
```
## [1] "Observations per year, 1871 - 2006:"
```

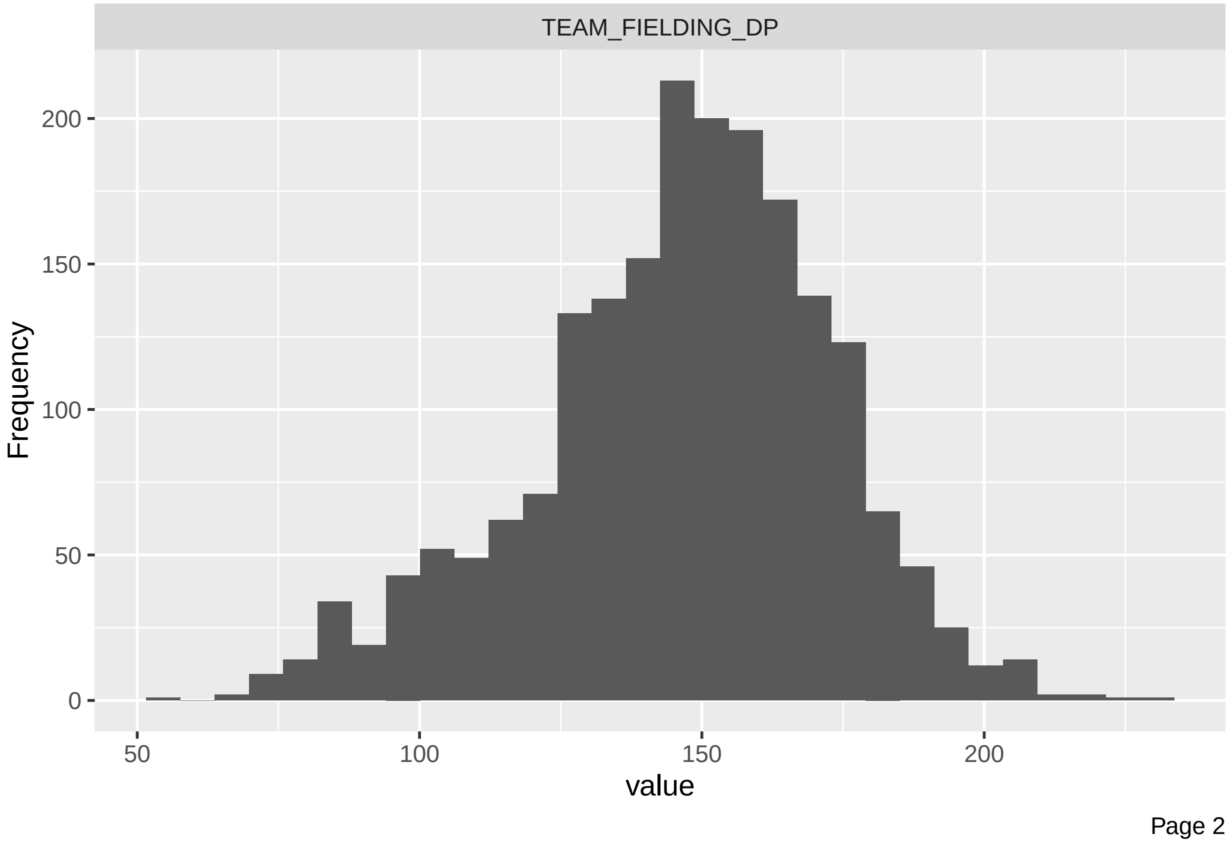
```r
round(nrow(baseball_df)/(2006-1871),2)
```

```
## [1] 16.86
```

Some columns have maximum values that are clearly outliers, like TEAM_PITCHING_H AND TEAM_PITCHING_HR. The assignment mentions that some of the season records were adjusted to match the performance during a 162-game season. There are 2276 seasons in the training set. Observations span 128 years, with an average of 17 teams playing per year. Based on a quick Google search, there were initially 8 teams in the MLB, and 30 teams in 2006. The MLB has two leagues of the same size since 1901, with the league sizes increasing in the late 20th century.

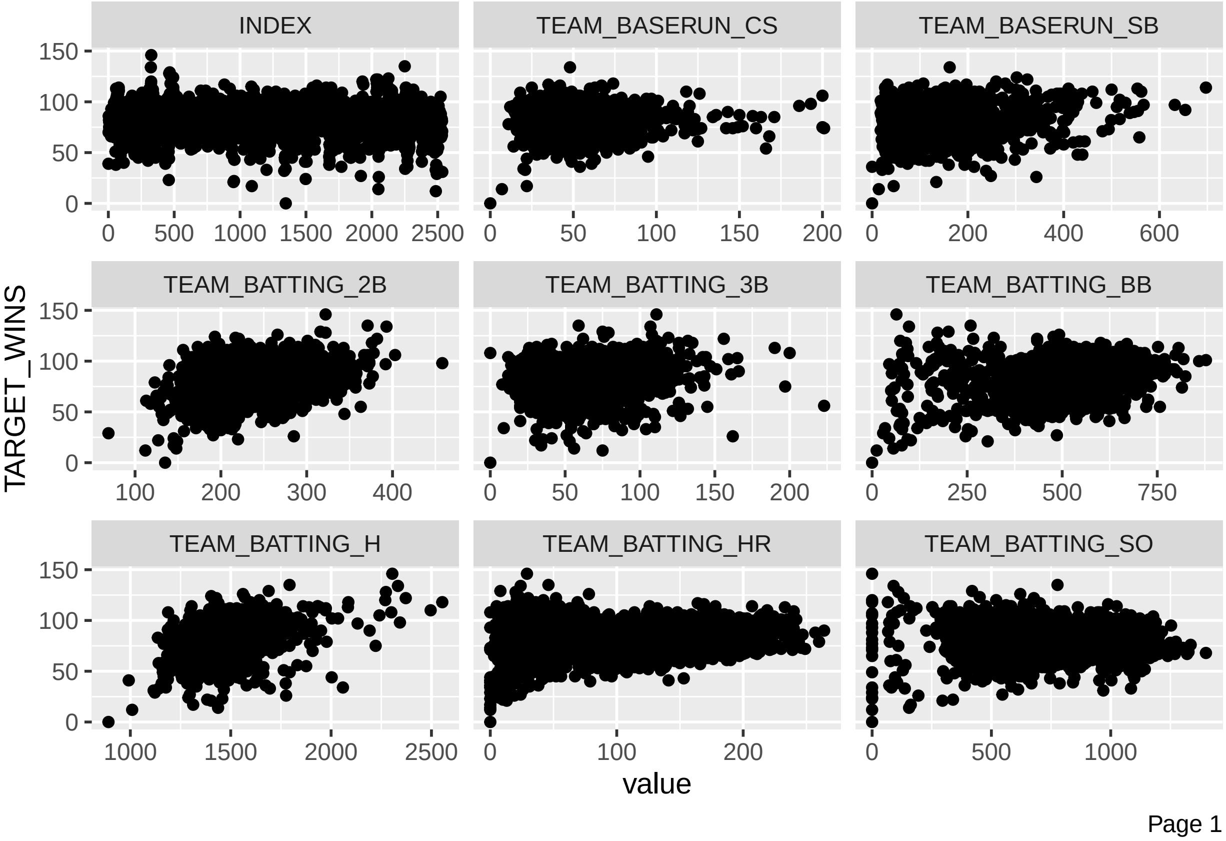
```r
# distribution
plot_histogram(baseball_df)
```



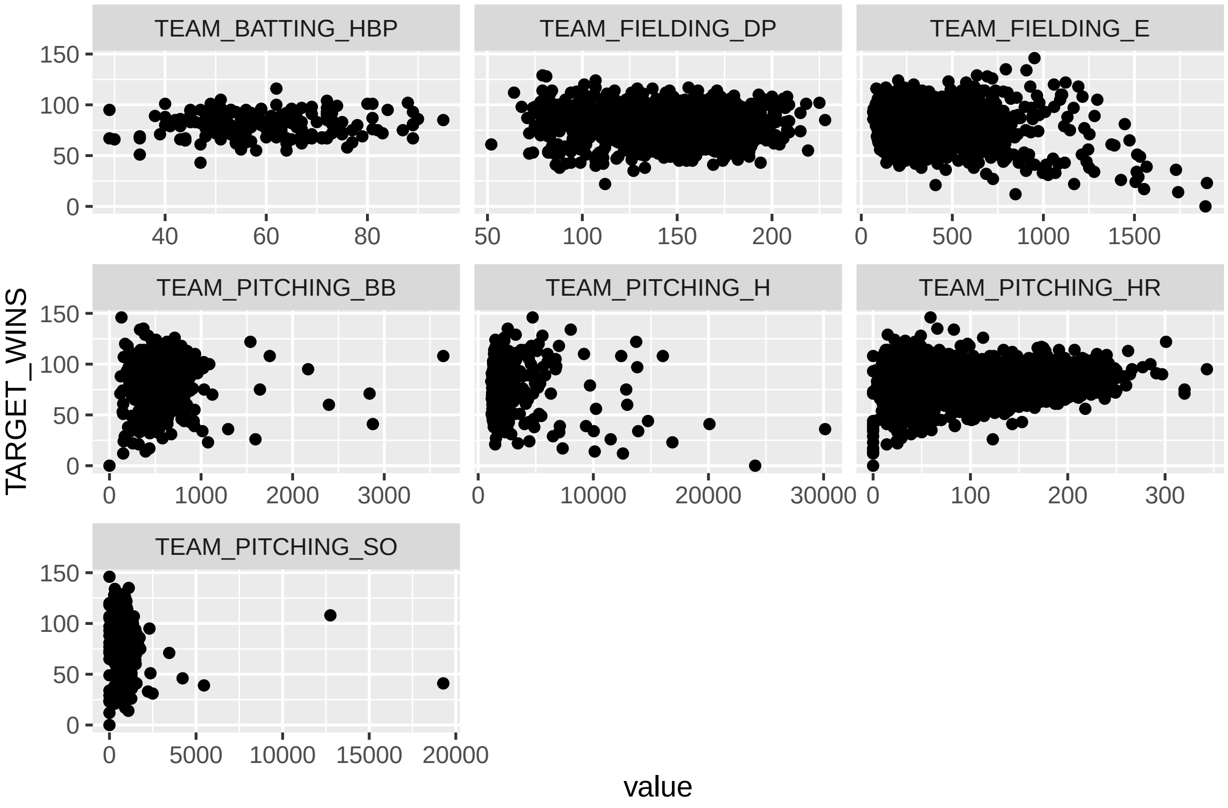


```
# against the response variable
plot_scatterplot(baseball_df, by = "TARGET_WINS")
```

```
## Warning: Removed 1005 rows containing missing values (geom_point).
```



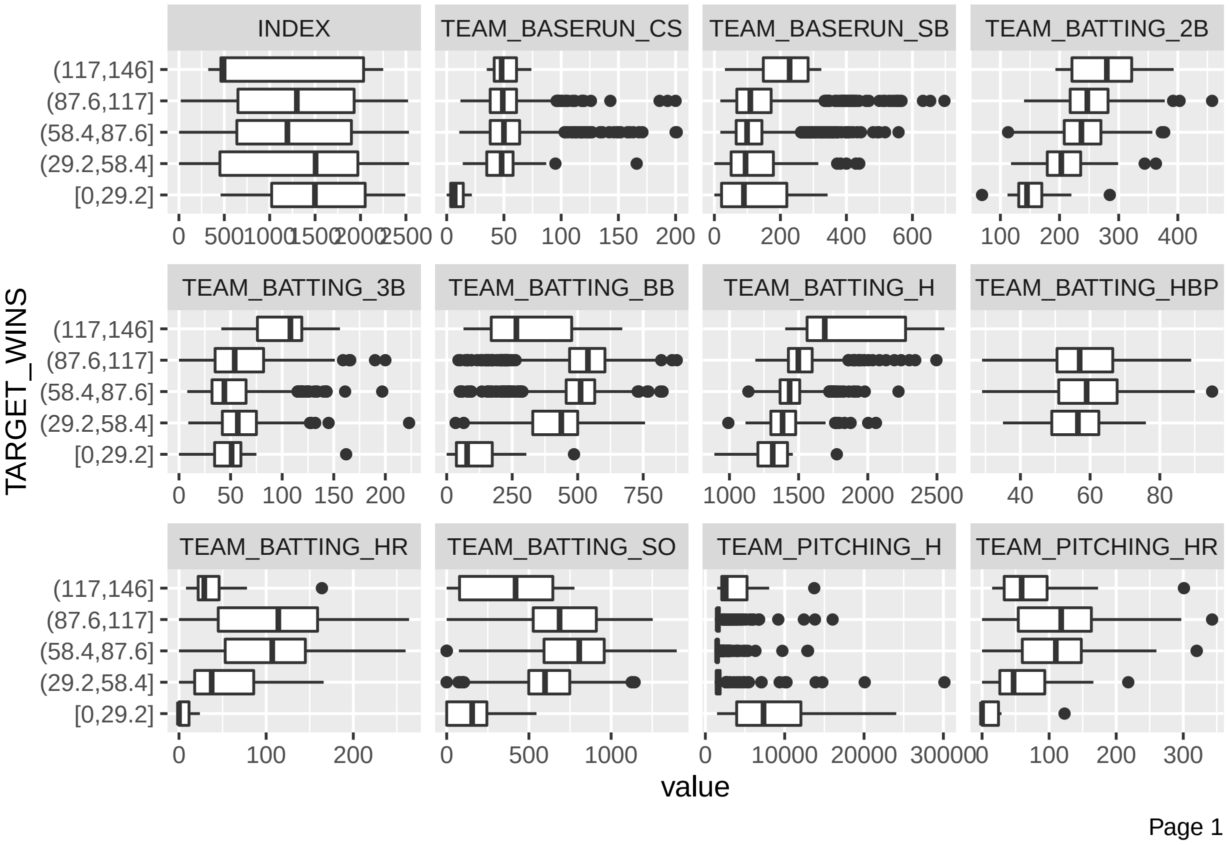
```
## Warning: Removed 2473 rows containing missing values (geom_point).
```



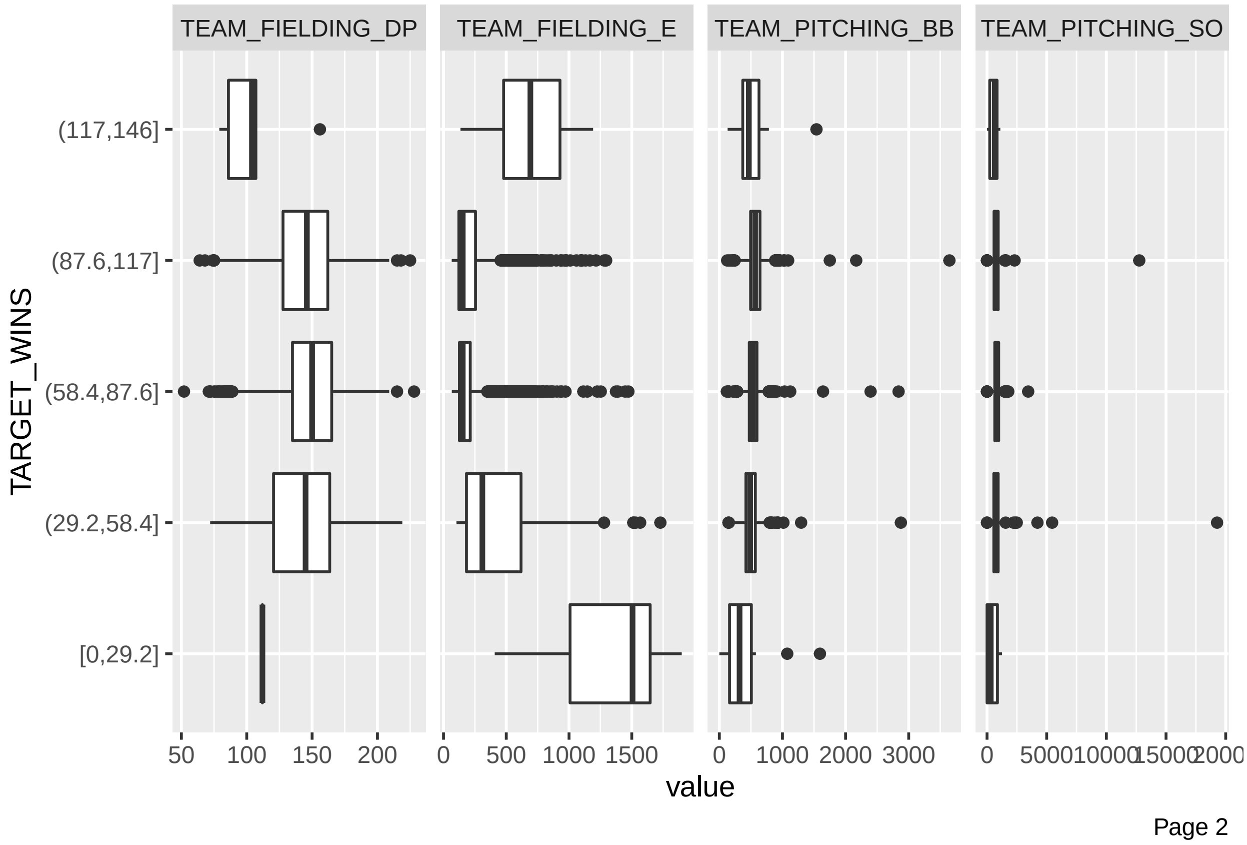
Page 2

```
# boxplot for train dataset
plot_boxplot(baseball_df, by="TARGET_WINS")
```

```
## Warning: Removed 3090 rows containing non-finite values (stat_boxplot).
```



```
## Warning: Removed 388 rows containing non-finite values (stat_boxplot).
```

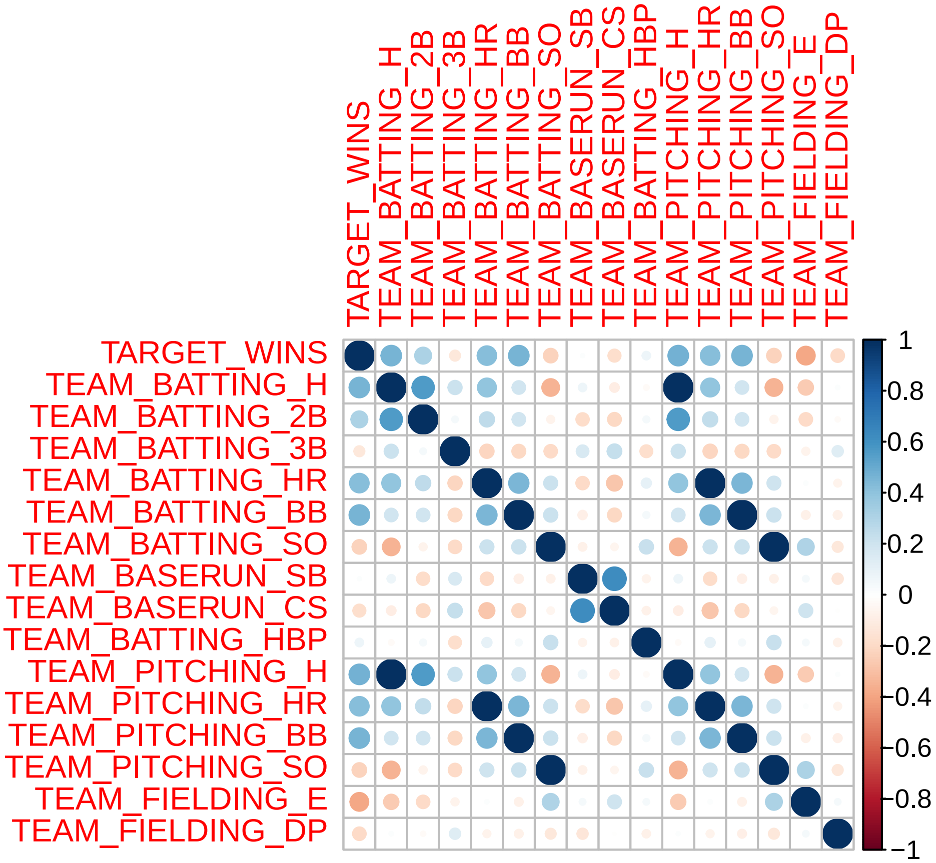


Feature Boxplots and Histograms

```
#baseball_df %>%
#  keep(is.numeric) %>%
#  gather() %>%
#  ggplot(aes(value)) +
#    facet_wrap(~ key, scales = "free") +
#    geom_boxplot()


#baseball_df %>%
#  keep(is.numeric) %>%
#  gather() %>%
#  ggplot(aes(value)) +
#    facet_wrap(~ key, scales = "free") +
#    geom_histogram()
```

Based on this quick look of boxplots and histograms, there are several variables with skewed distributions that would benefit from transformation. Additonally, there are a few variables with bimodal distributions.

```
corrplot(cor(baseball_df[,2:17], use = 'complete.obs'))
```



Looking at the correlation plot, there appear to be several strong correlations between explanatory variables and the target.

From an initial inspection, it appears the team should focus on getting players on base through hits or walks. Contrary to what I would expect, teams can still win if the pitchers allow homeruns, hits and walks to the other team. *Variables with Highest Positive Correlation with TARGET_WINS:*
* TEAM_BATTING_H = 0.47 * TEAM_BATTING_HR = 0.42 * TEAM_BATTING_BB = 0.47 * TEAM_PITCHING_H = 0.47 * TEAM_PITCHING_HR = 0.42 * TEAM_PITCHING_BB = 0.47

To win more games it makes sense the team will need to make fewer errors.
Variables with Strongly Negative Correlation with TARGET_WINS:

There were several batting variables which were related.
Positive Correlations between variables:
* TEAM_PITCHING_H and TEAM_BATTING_H = 0.99
* TEAM_PITCHING_HR and TEAM_BATTING_HR = 0.99
* TEAM_PITCHING_BB and TEAM_BATTING_BB = 0.99
* TEAM_PITCHING_SO and TEAM_BATTING_SO = 0.99

The pitchers who have more strikeouts allow fewer hits, which makes sense. It's interesting that pitchers who have fewer strikeouts have fewer team batting hits. Potentially due to the game being over in fewer innings and lower score games. This would be an interesting point to look into more.
Negative Correlations between variables:
* TEAM_PITCHING_SO and TEAM_BATTING_H = -0.34
* TEAM_PITCHING_SO and TEAM_PITCHING_H = -0.34

Missing values

```
round(100*colSums(is.na(baseball_df))/nrow(baseball_df),2)
```

```
##            INDEX      TARGET_WINS   TEAM_BATTING_H  TEAM_BATTING_2B
##             0.00             0.00             0.00             0.00
##  TEAM_BATTING_3B  TEAM_BATTING_HR  TEAM_BATTING_BB  TEAM_BATTING_SO
##             0.00             0.00             0.00             4.48
##  TEAM_BASERUN_SB  TEAM_BASERUN_CS TEAM_BATTING_HBP  TEAM_PITCHING_H
##             5.76            33.92            91.61             0.00
## TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E
##             0.00             0.00             4.48             0.00
## TEAM_FIELDING_DP
##            12.57
```

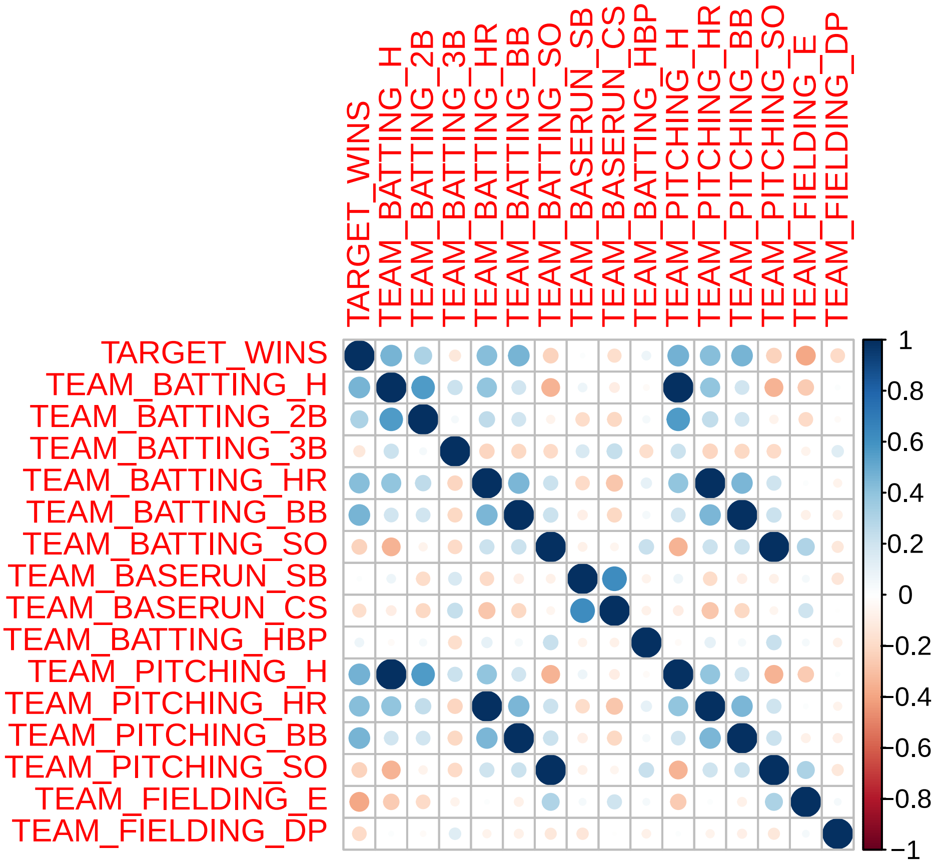
In terms of missing values, there are two variables missing many obserations. TEAM_BATTING_HBP is missing over 90% of its values, while TEAM_BASERUN_CS is missing just around 30%. Since so many observations are missing, imputing values could change the distributions considerably. To retain as many features as possible, I think it makes sense to explore these two variables first. The other affected missing values only have 5-10% misisng values. None of these appear to be stand-ins for 'zero' values, so mean values can be used insead.

I'll start by looking at TEAM_BATTING_HBP.

```
baseball_no_hbp <- baseball_df %>%
  filter(is.na(TEAM_BATTING_HBP)) %>% #missing values for hbp
  dplyr::select(-TEAM_BATTING_HBP)  ## select all rows except hbp



baseball_hbp <- baseball_df %>%
  filter(!is.na(TEAM_BATTING_HBP)) #not missing values for hbp
```

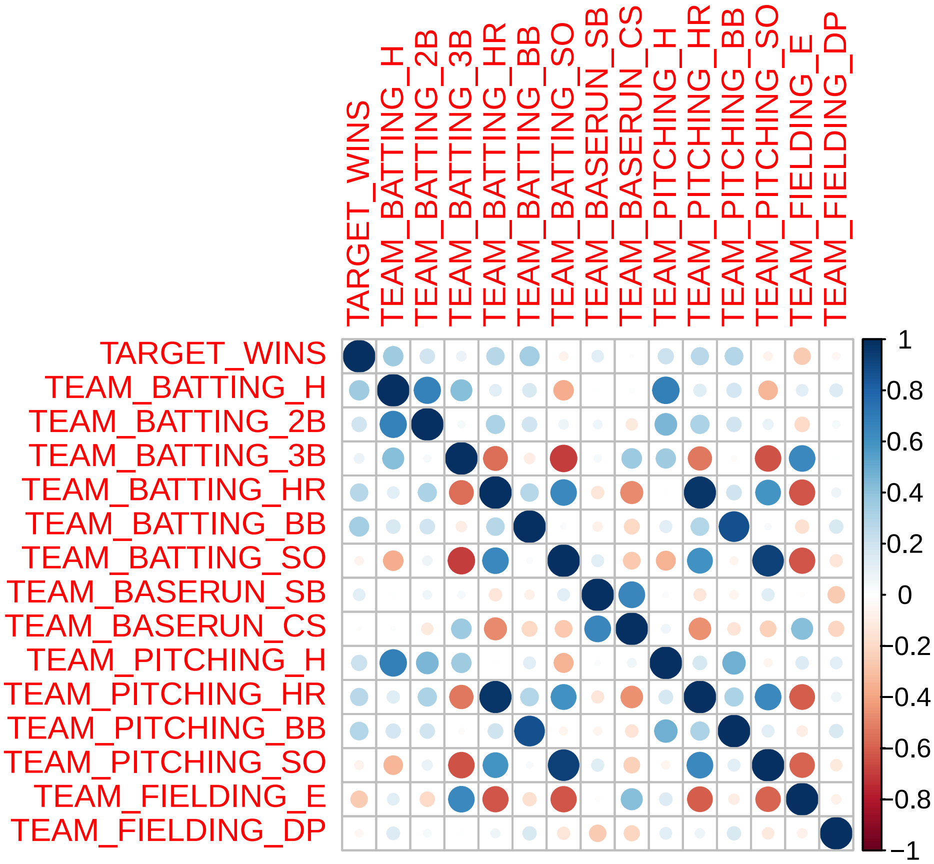
I separated training data into two smaller dataframes, one with complete values for HBP and one omitting this variable.

```
corrplot(cor(baseball_df[,2:17], use = 'complete.obs'))

corrplot(cor(baseball_hbp[,2:17], use = 'complete.obs'))
```

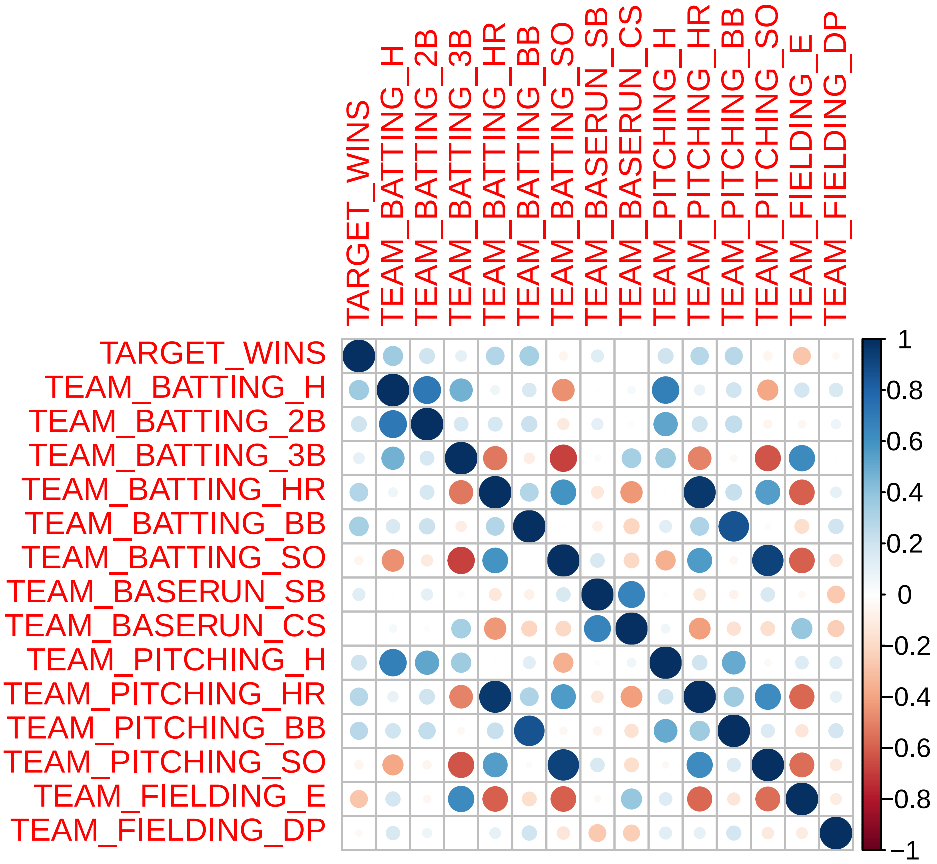


When HBP has values, it appears that there are no major changes in correlations.

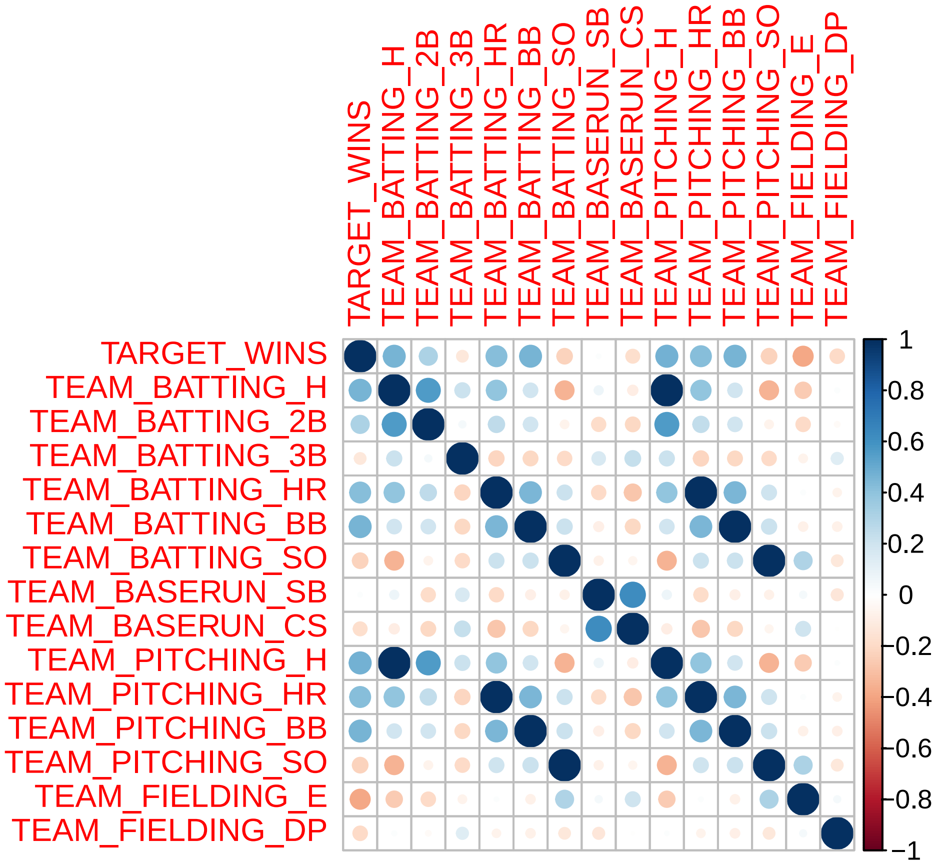
```
corrplot(cor(baseball_df[,-c(1,11)], use = 'complete.obs'))  #all rows
```



```
corrplot(cor(baseball_no_hbp[,2:16], use = 'complete.obs')) #only rows  missing values for hbp
```



```r
corrplot(cor(baseball_hbp[,-c(1,11)], use = 'complete.obs')) #only rows with values for hbp
```



There are three new correlaton plots being considered: the first is all datapoints, then a plot with missing hbp values, and finally a plot for rows with hbp values same as the previous pair. There appear to be no major discrepancies between missing values and the overall set. However, comparing missing values to available values does illustrate there are some distinct changes correlation when the hbp was recorded. This may be because HBP only represents only a small proportion of entries and has more variation. However, there also appear to be stronger correlations when HBP is added, which may help predict wins better than omitting altogether.

```
hbp_lm <- lm(baseball_hbp, formula = TARGET_WINS ~.-INDEX-TEAM_BATTING_HBP-TEAM_BATTING_SO-TEAM_BATTING_


summary(hbp_lm)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ . - INDEX - TEAM_BATTING_HBP - TEAM_BATTING_SO -
##     TEAM_BATTING_HR - TEAM_BASERUN_CS - TEAM_BATTING_H - TEAM_BASERUN_SB -
##     TEAM_PITCHING_BB - TEAM_BATTING_2B - TEAM_BATTING_3B, data = baseball_hbp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.8123  -5.9942  -0.0737   5.3098  22.2025
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       62.916227  19.219854   3.274 0.001269 **
## TEAM_BATTING_BB   0.055959   0.009466   5.912 1.61e-08 ***
## TEAM_PITCHING_H   0.026147   0.010184   2.567 0.011041 *
## TEAM_PITCHING_HR  0.091571   0.024033   3.810 0.000189 ***
## TEAM_PITCHING_SO -0.028772   0.007191  -4.001 9.13e-05 ***
## TEAM_FIELDING_E  -0.173897   0.039905  -4.358 2.18e-05 ***
## TEAM_FIELDING_DP -0.121570   0.035338  -3.440 0.000719 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.473 on 184 degrees of freedom
## Multiple R-squared:  0.5263, Adjusted R-squared:  0.5109
## F-statistic: 34.07 on 6 and 184 DF,  p-value: < 2.2e-16
```

```
baseball_hbp_dummy <- baseball_df %>%
  mutate(TEAM_HBP_YES_NO = case_when(!is.na(TEAM_BATTING_HBP) ~ 1, is.na(TEAM_BATTING_HBP) ~ 0)) %>%
  dplyr::select(-TEAM_BATTING_HBP)


summary(lm(baseball_hbp_dummy, formula = TARGET_WINS ~.-INDEX-TEAM_PITCHING_BB-TEAM_PITCHING_HR-TEAM_BA
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ . - INDEX - TEAM_PITCHING_BB - TEAM_PITCHING_HR -
##     TEAM_BATTING_H, data = baseball_hbp_dummy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.3981  -6.6295   0.1545   6.4842  28.2220
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      58.585629   6.442796   9.093  < 2e-16 ***
## TEAM_BATTING_2B  -0.060017   0.009747  -6.158 9.50e-10 ***
## TEAM_BATTING_3B   0.166293   0.022021   7.552 7.51e-14 ***
## TEAM_BATTING_HR   0.100869   0.009172  10.998  < 2e-16 ***
## TEAM_BATTING_BB   0.038251   0.003366  11.363  < 2e-16 ***
## TEAM_BATTING_SO   0.040704   0.009102   4.472 8.35e-06 ***
## TEAM_BASERUN_SB   0.034100   0.008689   3.924 9.10e-05 ***
## TEAM_BASERUN_CS   0.052980   0.018176   2.915  0.00361 **
## TEAM_PITCHING_H   0.031740   0.004269   7.435 1.76e-13 ***
## TEAM_PITCHING_SO -0.058995   0.007547  -7.817 1.02e-14 ***
## TEAM_FIELDING_E  -0.158154   0.009939 -15.912  < 2e-16 ***
## TEAM_FIELDING_DP -0.112916   0.013095  -8.623  < 2e-16 ***
## TEAM_HBP_YES_NO  -2.456525   0.923761  -2.659  0.00792 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.532 on 1473 degrees of freedom
##   (790 observations deleted due to missingness)
## Multiple R-squared:  0.4407, Adjusted R-squared:  0.4361
## F-statistic: 96.71 on 12 and 1473 DF,  p-value: < 2.2e-16
```

I compared two preliminary linear models that I arrived at through backward selection. Looking only at

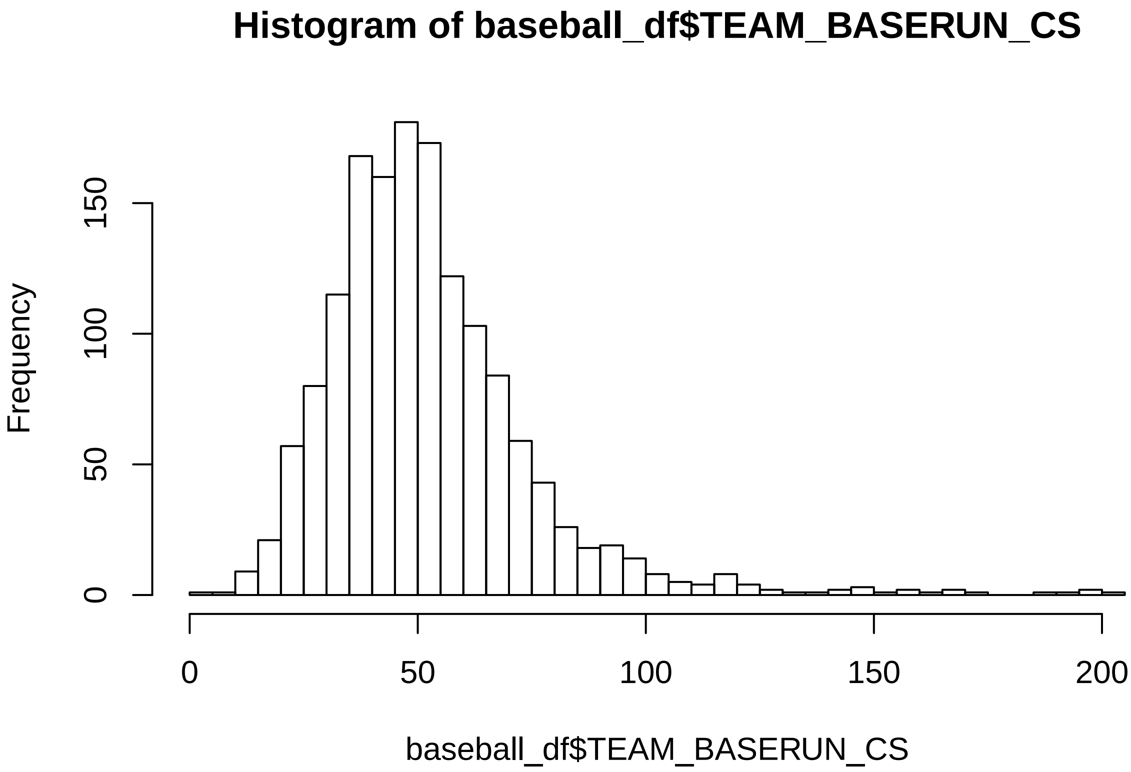
HBP-containing observations, there's a small increase in r-squared compared to a model that uses a dummy variable to consider whether values were available.

Next, I'll look at TEAM_BASERUN_CS, which was missing about 30% of its values.

```
sum(baseball_df$TEAM_BASERUN_CS == 0, na.rm = TRUE)
```

```
## [1] 1
```

```
hist(baseball_df$TEAM_BASERUN_CS, breaks = 30)
```

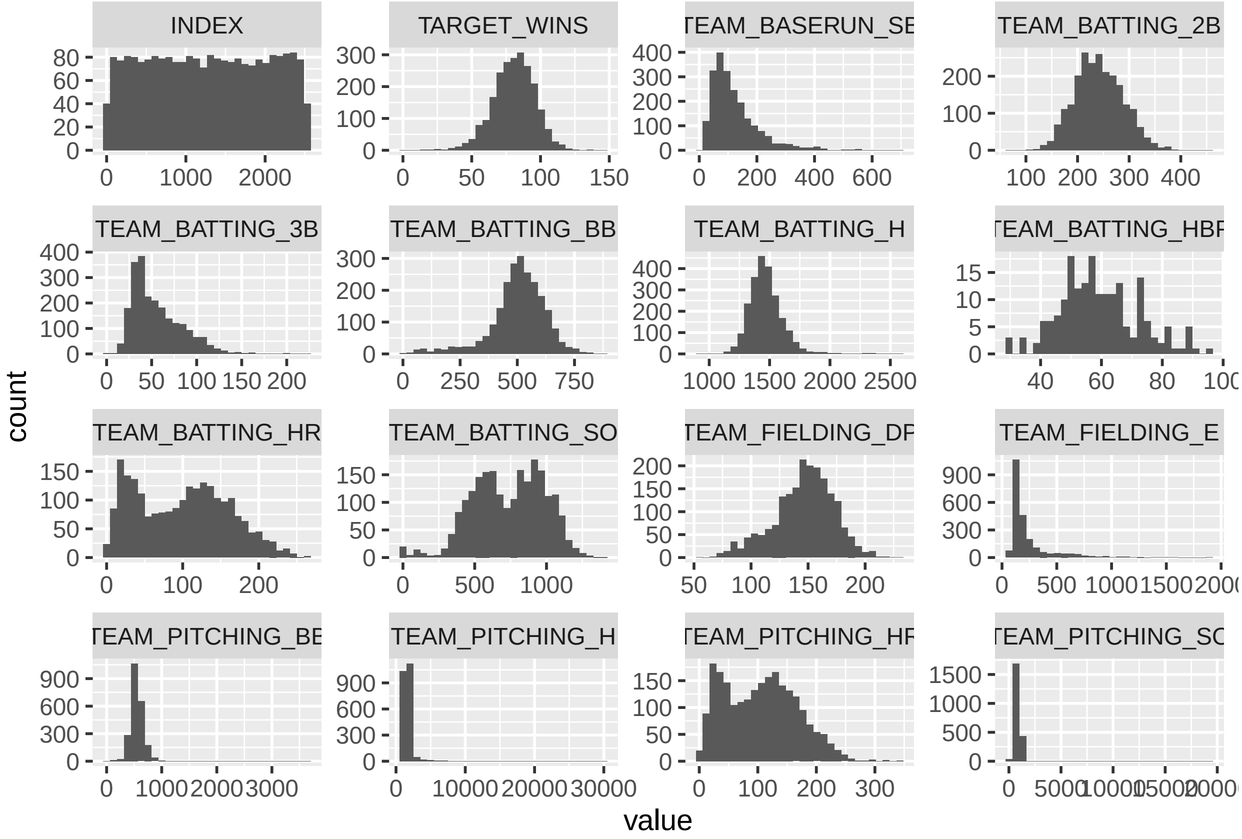


```
baseball_no_cs <- baseball_df %>%
  filter(is.na(TEAM_BASERUN_CS)) %>% #missing values for hbp
  dplyr::select(-TEAM_BASERUN_CS)  ## select all rows except hbp



baseball_cs <- baseball_df %>%
  filter(!is.na(TEAM_BASERUN_CS)) #not missing values for hbp
```

Same as HBP, it appears CS did not miscode values of 0 as NA. I'm going to separate the dataset in the same way as HBP to see if there are differences in its distribution and correlation plots.

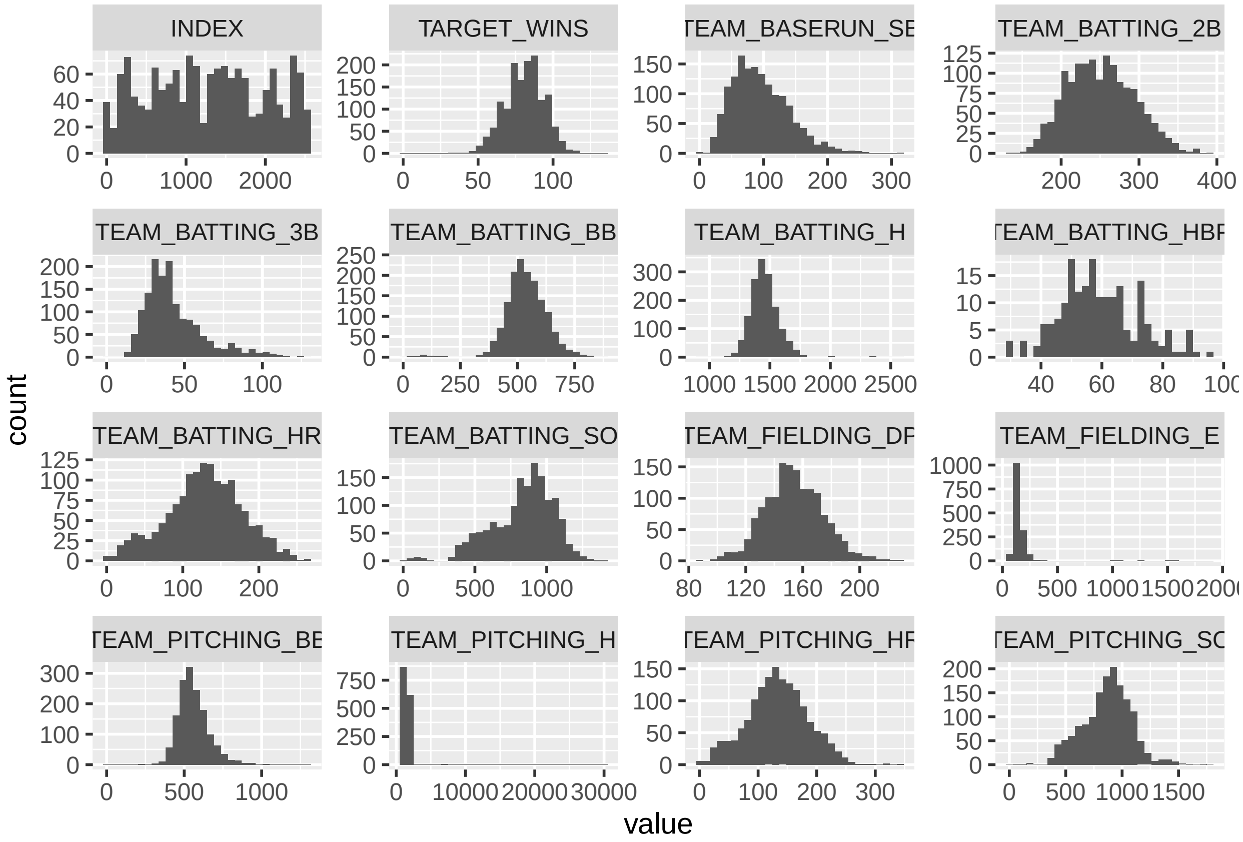
```
baseball_df %>% ##original histograms
  dplyr::select(-TEAM_BASERUN_CS) %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2706 rows containing non-finite values (stat_bin).
```



```
baseball_cs %>% ##historgrams with seasons having CS statistics
  dplyr::select(-TEAM_BASERUN_CS) %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram()
```

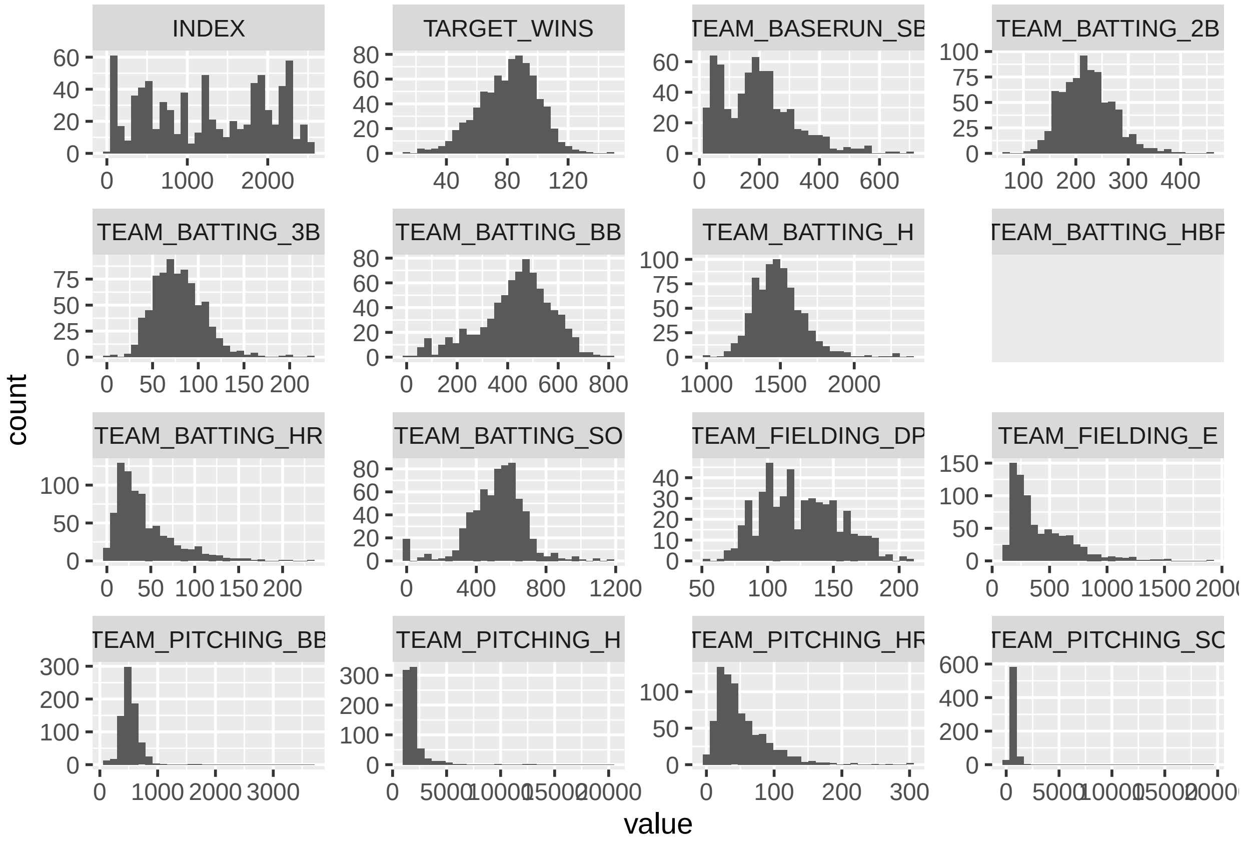
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1331 rows containing non-finite values (stat_bin).
```



```
baseball_no_cs %>% #histograms missing CS statistics
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

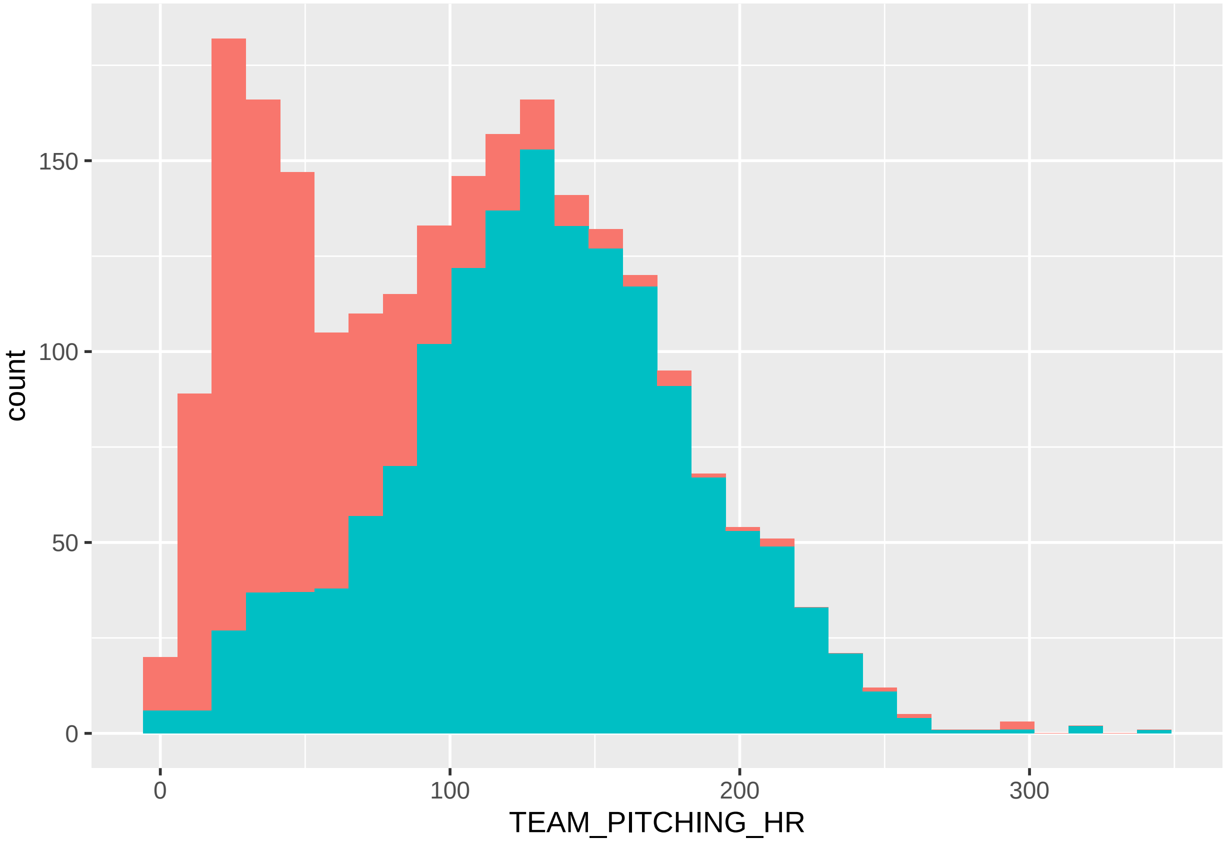
```
## Warning: Removed 1375 rows containing non-finite values (stat_bin).
```



After subsetting for availability of CS statistics, an interesting pattern emerges: our three bimodal variables, TEAM_PITCHING_HR, TEAM_BATTING_SO, and TEAM_BATTING_HR, are no longer bimodal.

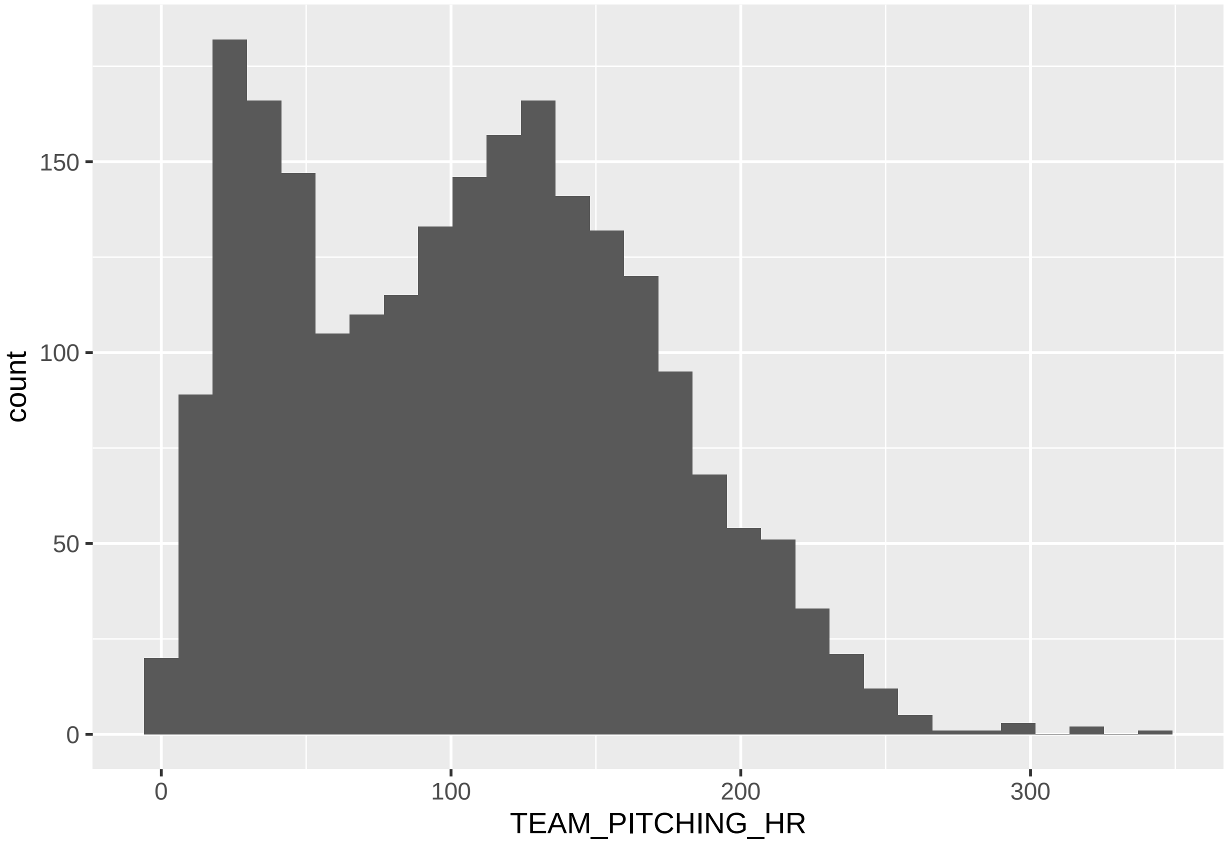
```
baseball_hbp_dummy <- baseball_hbp_dummy %>%
  mutate(TEAM_CS_YES_NO = case_when(!is.na(TEAM_BASERUN_CS) ~ '1', is.na(TEAM_BASERUN_CS) ~ '0'))

ggplot(baseball_hbp_dummy, aes(x = TEAM_PITCHING_HR, fill = TEAM_CS_YES_NO)) +
  geom_histogram() +
  theme(legend.position = 'none')
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



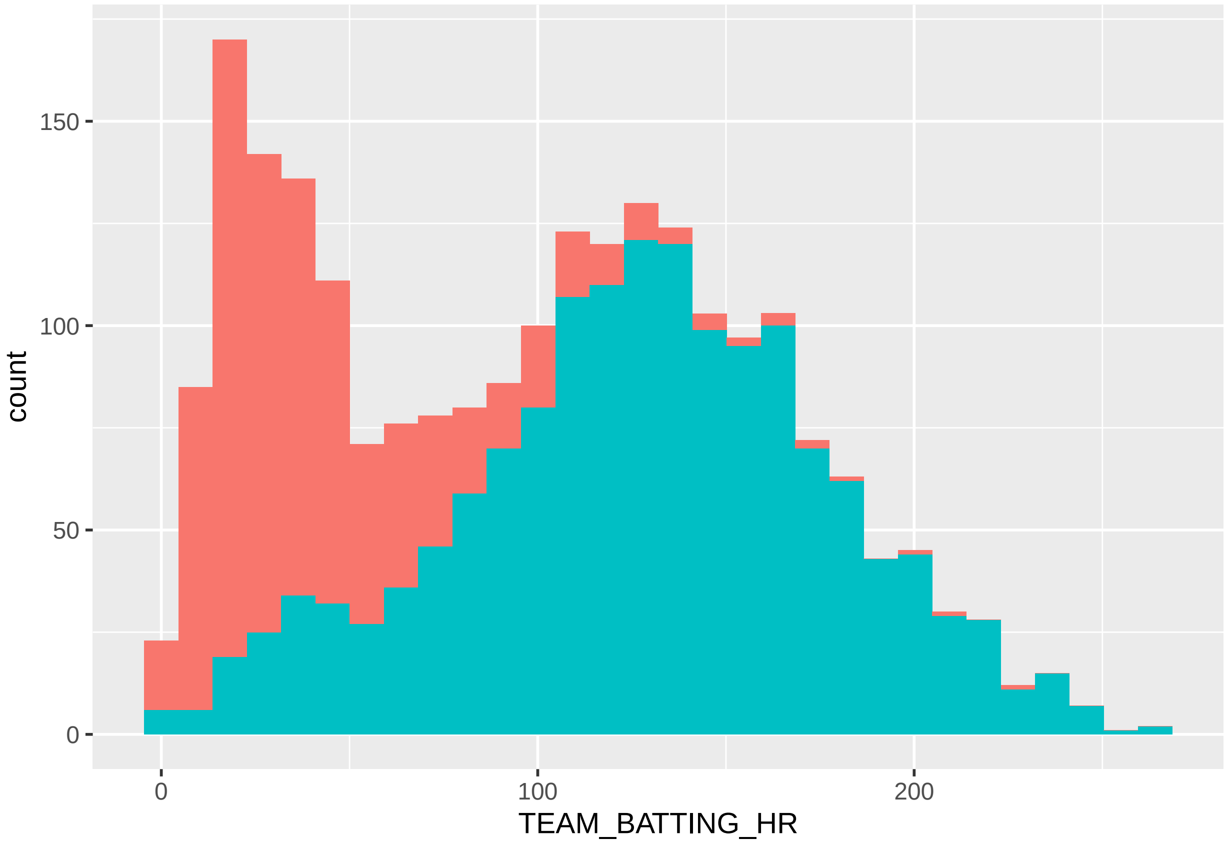
```r
ggplot(baseball_df, aes(x = TEAM_PITCHING_HR)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



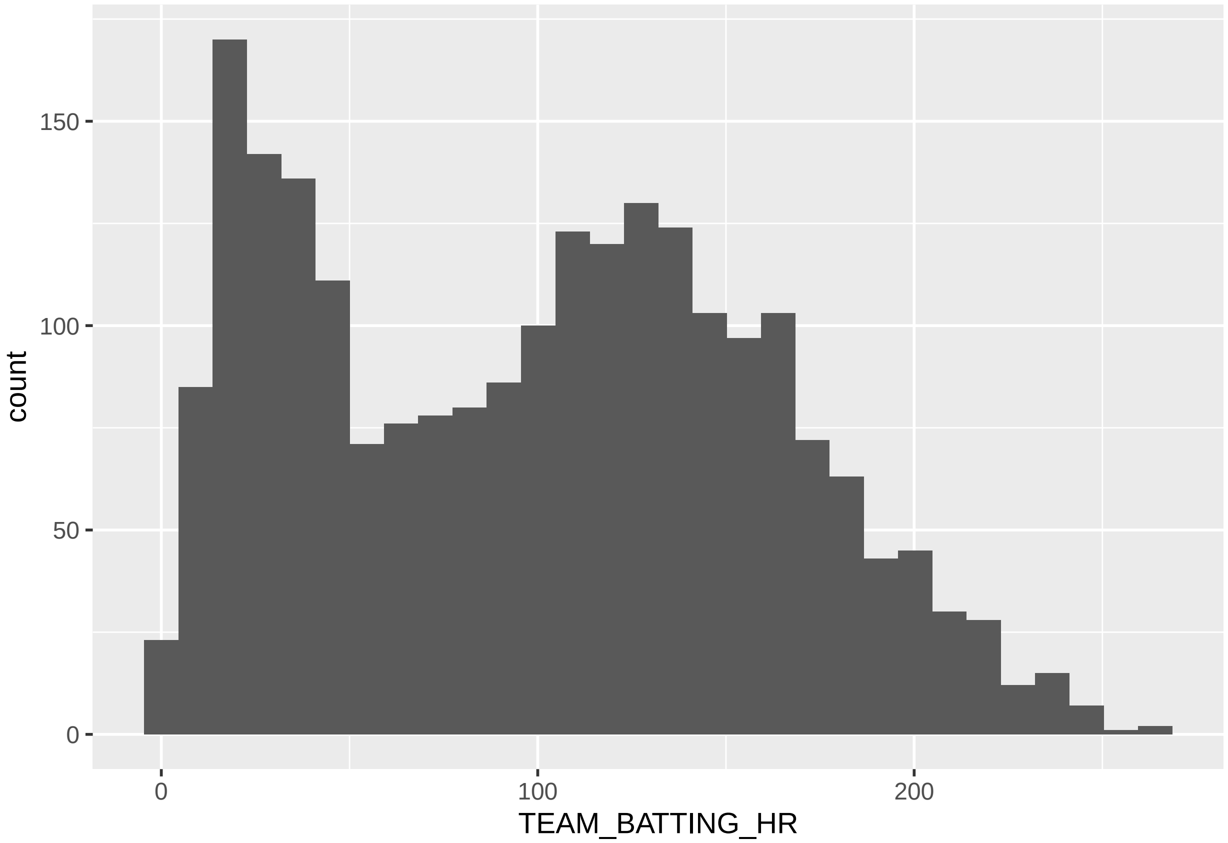
```r
ggplot(baseball_hbp_dummy, aes(x = TEAM_BATTING_HR, fill = TEAM_CS_YES_NO)) +
  geom_histogram() +
  theme(legend.position = 'none')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(baseball_df, aes(x = TEAM_BATTING_HR)) +
  geom_histogram()
```

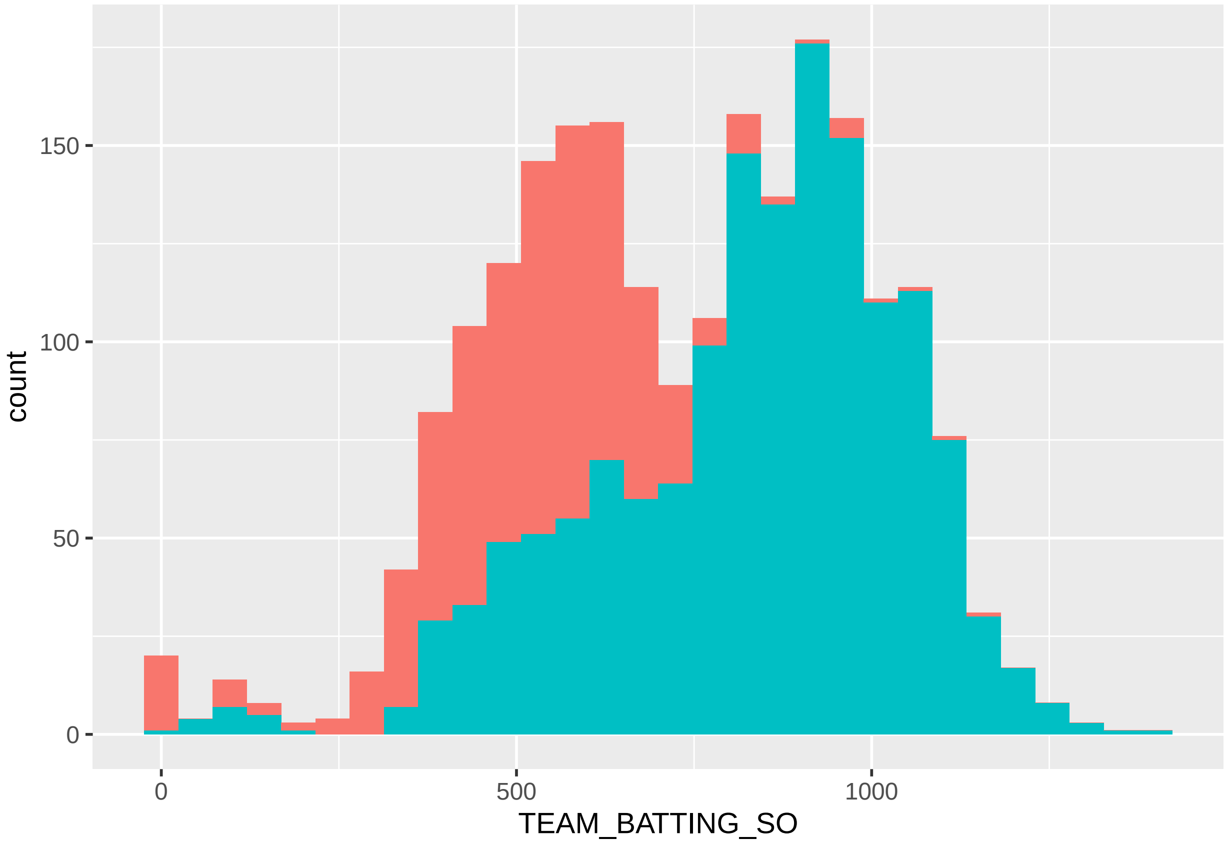
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



```
ggplot(baseball_hbp_dummy, aes(x = TEAM_BATTING_SO, fill = TEAM_CS_YES_NO)) +
  geom_histogram() +
  theme(legend.position = 'none')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

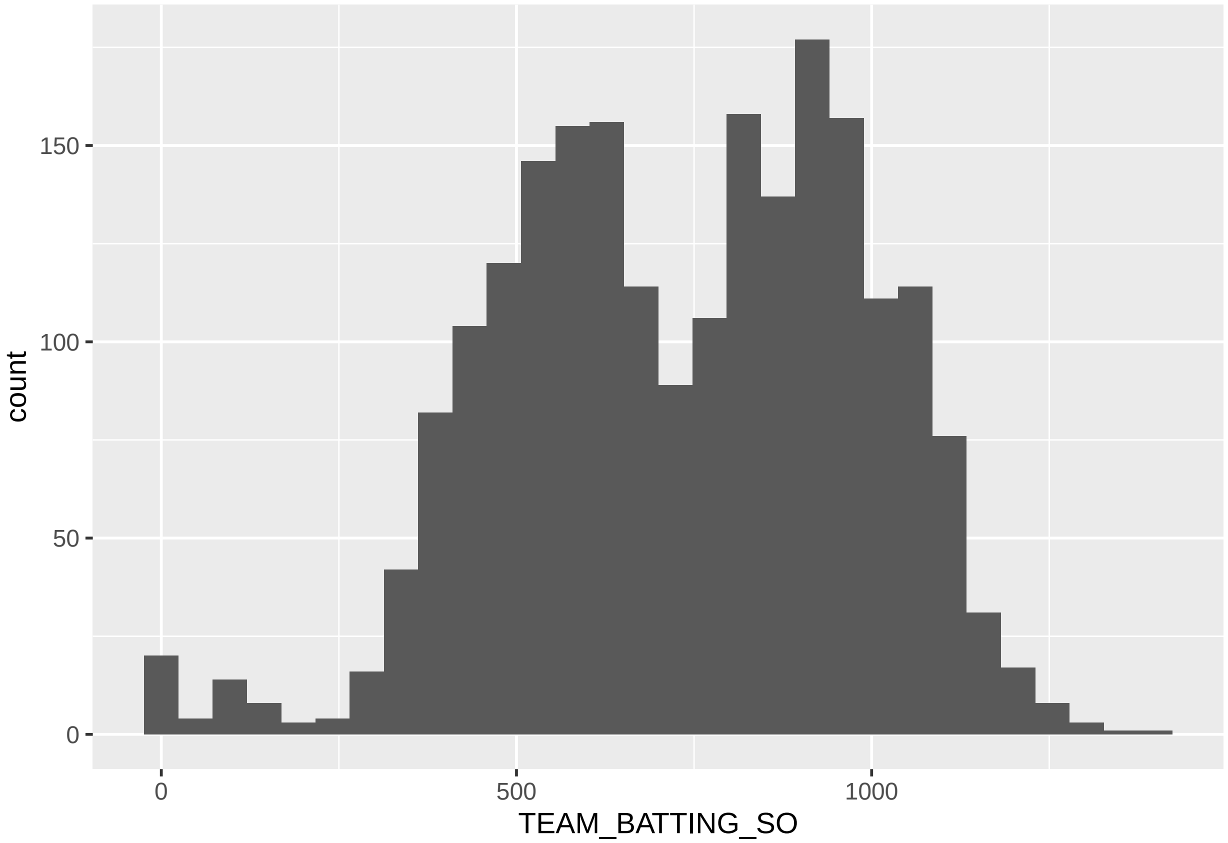
```
## Warning: Removed 102 rows containing non-finite values (stat_bin).
```



```
ggplot(baseball_df, aes(x = TEAM_BATTING_SO)) +
  geom_histogram()
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

```
## Warning: Removed 102 rows containing non-finite values (stat_bin).
```



As these three histograms illustrate, the bimodal distributions are explained by missing CS values or not. Missing values explain both modes present in the overall histogram.

```
# Devin - start- baseball_df_fix not defined yet - did you want to move to after? I added the code here
baseball_df_fix <- baseball_df %>%
  mutate(TEAM_CS_YES_NO = case_when(!is.na(TEAM_BASERUN_CS) ~ 1, is.na(TEAM_BASERUN_CS) ~ 0)) %>%
  mutate(TEAM_HBP_YES_NO = case_when(!is.na(TEAM_BATTING_HBP) ~ 1, is.na(TEAM_BATTING_HBP) ~ 0)) %>%
  dplyr::select(-c(TEAM_BATTING_HBP, INDEX, TEAM_BASERUN_CS))
# Devin - end- baseball_df_fix not defined yet - did you want to move to after? I added the code here f

#Devin - still erroring out?
#image(is.na(baseball_df_fix),axes=FALSE,col=gray(1:0))
#axis(2, at=0:17/17, labels=colnames(baseball_df_fix))
#axis(1, at=0:2275/2275, labels=row.names(baseball_df_fix),las=2)
```

To better visualize the missing values, it looks like two of them overlap perfectly.

```
baseball_df_fix[rowSums(is.na(baseball_df_fix)) > 0,] %>%
  dplyr::select(TEAM_PITCHING_SO, TEAM_FIELDING_DP, TEAM_BATTING_SO, TEAM_BASERUN_SB)
```

```
##      TEAM_PITCHING_SO TEAM_FIELDING_DP TEAM_BATTING_SO TEAM_BASERUN_SB
## 1                5456               NA             842              NA
## 53                272               88              99              NA
## 54                525               97             227              NA
## 55                883               97             327              NA
## 56                825              112             428              NA
## 57               822              104             426              NA
## 58               908               71             471              NA
## 59              1155               NA             699              NA
## 60              1405               NA             963              NA
## 61              1092               NA             755              NA
## 62              1030               NA             744             216
## 63               703               NA             525             499
## 64               765               NA             633             354
## 65               721               NA             570             419
## 66               764               NA             627             347
## 67               742               NA             632             339
## 68               461               NA             367             305
## 69               393               NA             320             296
## 70               361               NA             292             246
## 71               419               NA             339             298
## 72               395               NA             322             286
## 73               358               NA             329             187
## 74               306               NA             287             197
## 77                NA              104              NA             187
## 78                NA               96              NA             151
## 79                NA               94              NA             139
## 80                NA              109              NA             100
## 81                NA              140              NA             129
## 82                NA               95              NA             141
## 83                NA              107              NA             143
## 175               NA               87              NA             240
## 176               NA              103              NA             153
## 177               NA              103              NA             134
## 178               NA              116              NA             147
## 179               NA              119              NA             178
## 180               NA              111              NA             133
## 181               NA              134              NA             185
## 269              552               NA             450             196
## 272              252               NA              84             105
## 273              208               NA              72              61
## 274             1059               91             477              NA
## 275              943               NA             559              NA
## 276             1273               NA             833              NA
## 277             1168               NA             786              NA
## 278              923               NA             746             333
## 279              481               NA             401             654
## 280              669               NA             566             385
## 281              772               NA             643             373
## 282             3450               NA             724             481
## 283              797               NA             664             410
## 284              583               NA             529             250
## 285              502               NA             403             290
## 286              320               NA             253             410
## 287              378               NA             303             386
## 288              316               NA             252             554
## 289              398               NA             319             500
## 290              374               NA             344             272
## 291              459               NA             419             398
## 294             1552               NA            1006              NA
## 295               556                NA             103             394
## 296               310                NA              90             162
## 297               181                NA              66             107
## 298                 0                NA               0              NA
## 299                 0                NA               0              NA
## 304                NA                80              NA             140
## 305                NA                88              NA             104
## 306                NA               112              NA             136
## 307                NA                75              NA             176
## 308                NA               102              NA             231
## 309                NA                85              NA             205
## 310                NA                98              NA             201
## 311                NA                94              NA             197
## 391               460                NA             105              74
## 392               387                NA             129              54
## 393                 0                NA               0              NA
## 394              1739                NA            1170              NA
## 395              1221                NA             746             298
## 396               686                NA             546             517
## 399              1354               129             652              NA
## 400              1276               109             646              NA
## 401              1029                94             527              NA
## 402               849                81             440              NA
## 403               954                NA             571              NA
## 404               975                NA             668              NA
## 405               796                NA             550              NA
## 406               556                NA             453             334
## 407              1144                NA             685              NA
## 408               909                NA             606              NA
## 409               836                83             423              NA
## 410               881               100             457              NA
## 411               833               131             432              NA
## 412              1019               140             516              NA
## 413              1044                NA             625              NA
## 414              1205                NA             833              NA
## 415                 0                NA               0              NA
## 416               782                NA             140             101
## 417               707                NA              96              88
## 418               270                81             110              NA
## 419               837               118             305              NA
## 420              1145               100             424              NA
## 421              1237               107             603              NA
## 422               808                79             419              NA
## 423               833               104             432              NA
## 424               974               104             505              NA
## 425              1091                NA             660              NA
## 426               981                NA             678              NA
## 427               898                NA             621              NA
## 428               875                NA             670             278
## 429               718                NA             536             511
## 430               811                NA             676             344
## 431               777                NA             633             298
## 432               658                NA             548             286
## 433               594                NA             535             259
## 434               449               NA             366             401
## 435               464               NA             367             420
## 436               477               NA             389             324
## 437               460               NA             426             238
## 438               409               NA             374             270
## 442                NA               92              NA             304
## 443                NA               94              NA             240
## 444                NA              105              NA             283
## 445                NA              107              NA             302
## 446                NA              117              NA             250
## 447                NA               80              NA             223
## 448                NA              101              NA             198
## 539               574               NA             485             326
## 542                NA               68              NA             207
## 543                NA               86              NA             230
## 544                NA              108              NA             188
## 545                NA               87              NA             223
## 546                NA              108              NA             225
## 547                NA              105              NA             216
## 637               836               83             413              NA
## 638               712               NA             431              NA
## 639               892               NA             600              NA
## 640               879               NA             608              NA
## 641               872               NA             743             217
## 642               527               NA             439             632
## 643               811               NA             671             567
## 644               695               NA             596             538
## 645               568               NA             463             383
## 646               614               NA             519             289
## 647               553               NA             512             292
## 648               410               NA             324             301
## 649               391               NA             314             268
## 650               386               NA             310             406
## 651               329               NA             268             238
## 652               341               NA             320             176
## 653               345               NA             319             246
## 655                NA               98              NA             168
## 656                NA               86              NA             190
## 657                NA              129              NA             192
## 658                NA              104              NA             182
## 659                NA              125              NA             167
## 660                NA               76              NA             206
## 749              1011               NA             624             266
## 754                NA               88              NA             198
## 755                NA              118              NA             215
## 756                NA              146              NA             206
## 757                NA              116              NA             184
## 758                NA              114              NA             221
## 759                NA              131              NA             205
## 844               536               NA             430             224
## 845               839               NA             715             357
## 846               814               NA             673             427
## 847               741               NA             627             331
## 848               709               NA             573             439
## 849              842               NA             686             433
## 850              714               NA             582             187
## 851              631               NA             541             282
## 852              634               NA             583             245
## 853              367               NA             290             319
## 854              475               NA             378             276
## 855              561               NA             450             233
## 856              506               NA             400             221
## 857              526               NA             425             224
## 858              362               NA             333             101
## 859              310               NA             295             134
## 860                0               NA               0              NA
## 861                0               NA               0              NA
## 862              845              112             339              NA
## 863              890               94             313              NA
## 864             1029              100             381              NA
## 865              848               97             419              NA
## 866             1096               99             541              NA
## 881             1148               NA             744              NA
## 882             1600               NA             899              NA
## 886               NA               85              NA             137
## 887               NA               94              NA             224
## 888               NA              101              NA             175
## 889               NA               82              NA             290
## 890               NA               96              NA             286
## 976             1173               86             601              NA
## 977             1033               NA             625              NA
## 978             1016               NA             677              NA
## 979              739               NA             561             256
## 980              440               NA             337             349
## 981              606               NA             490             239
## 982              905               NA             162              45
## 996              443               64             186              NA
## 997              757               89             271              NA
## 998                0               NA               0              NA
## 999                0               NA               0              NA
## 1044            1436              100             532              NA
## 1045            1284               NA             848              NA
## 1046             626               NA             487             429
## 1047             708               NA             590             420
## 1048             653               NA             540             305
## 1049             909               NA             741             315
## 1050             875               NA             740             558
## 1082            2225               NA            1085              NA
## 1083             657               NA              77             136
## 1084            1013               NA             650              NA
## 1085             678               NA             469              NA
## 1086             731               NA             618             293
## 1087             533               NA             441             494
## 1088             560               NA             474             460
## 1089             569               NA             453             438
## 1090             608               NA             514             398
## 1091             560               NA             532             430
## 1092             475               NA             375             270
## 1093              450               NA             364             349
## 1094              486               NA             393             226
## 1095              412               NA             333             245
## 1096              384               NA             313             230
## 1097              392               NA             351             145
## 1101               NA              117              NA             325
## 1102               NA              108              NA             198
## 1103               NA               78              NA             187
## 1104               NA               91              NA             149
## 1191              592              108             241              NA
## 1192             1021              100             378              NA
## 1193              850               NA             509              NA
## 1194              918               NA             612              NA
## 1195              937               NA             648              NA
## 1196              792               NA             665             241
## 1197              505               NA             424             555
## 1198              870               NA             725             382
## 1199              718               NA             612             238
## 1200              693               NA             565             419
## 1201              661               NA             559             272
## 1202              577               NA             541             293
## 1203              515               NA             397             263
## 1204              566               NA             454             270
## 1205              493               NA             399             193
## 1206              653               NA             528             241
## 1207              704               NA             565             243
## 1208              494               NA             460             252
## 1209              426               NA             400             248
## 1210              547               NA              81              34
## 1211                0               NA               0               0
## 1249               NA               82              NA             181
## 1250               NA               84              NA             252
## 1251               NA               74              NA             239
## 1252               NA              107              NA             145
## 1253               NA              106              NA             206
## 1254               NA               95              NA             226
## 1255               NA               98              NA             292
## 1340             2309               NA             513             212
## 1341             1561               86             578              NA
## 1342            12758               NA             945              NA
## 1345                0               NA               0              NA
## 1346              363               NA              74             226
## 1347              468               NA             156             168
## 1348              205               NA              67              46
## 1349                0               NA               0              NA
## 1350                0               NA               0              NA
## 1351              556               NA             450             286
## 1393              737               NA             437              NA
## 1394              722               NA             477              NA
## 1395              833               NA             694             144
## 1396              687               NA             564             372
## 1397              292               52             101              NA
## 1401               NA               96              NA             217
## 1402               NA               74              NA             206
## 1403               NA               86              NA             225
## 1404               NA               82              NA             243
## 1405               NA              102              NA             309
## 1406               NA              106              NA             287
## 1407               NA               82              NA             263
## 1496               NA              104              NA             208
## 1497               NA               96              NA             184
## 1498               NA               75              NA             153
## 1499               NA               72              NA             123
## 1500               NA               97              NA             217
## 1501               NA              126              NA             224
## 1502               NA              107              NA             242
## 1584             1296               NA              72               0
## 1585              601               NA             475             315
## 1588              765               78             354              NA
## 1589              732               NA             443              NA
## 1590              974               NA             643              NA
## 1591              858               NA             593              NA
## 1592             1003               NA             836             341
## 1593              702               NA             576             529
## 1594              814               NA             663             374
## 1595             1072               NA             741              NA
## 1596             1042               NA             733             321
## 1597              601               NA             456             468
## 1598              753               NA             604             307
## 1599              574               NA             450             343
## 1600              616               NA             498             414
## 1601              576               NA             487             274
## 1602              577               NA             545             229
## 1603              529               NA             421             254
## 1604              392               NA             310             346
## 1605              461               NA             370             238
## 1606              450               NA             367             200
## 1607              451               NA             415             198
## 1608              387               NA             363             226
## 1612               NA               91              NA             144
## 1613               NA               88              NA             191
## 1614               NA              115              NA             170
## 1615               NA               79              NA             210
## 1616               NA              103              NA             196
## 1698             2367               NA             979              NA
## 1699              491               NA             397             251
## 1700              745               NA             639             174
## 1701              789               83             380              NA
## 1702              942               NA             570              NA
## 1703              926               NA             617              NA
## 1704             1144               NA             784              NA
## 1705              997               NA             843             307
## 1706              651               NA             498             289
## 1707              852               NA             705             347
## 1708              650               NA             546             248
## 1709              508               NA             480             235
## 1710              431               NA             343             264
## 1711              450               NA             367             315
## 1712              451               NA             359             273
## 1713              511               NA             413             210
## 1714              410               NA             375             117
## 1718               NA              116              NA             199
## 1719               NA               98              NA             188
## 1720               NA              119              NA             214
## 1721               NA              115              NA             172
## 1722               NA               79              NA             278
## 1723               NA               78              NA             196
## 1724               NA              107              NA             197
## 1810              770               NA             133             324
## 1811              637               NA             173             214
## 1812                0               NA               0              NA
## 1813                0               NA               0              NA
## 1814             1590              113             589              NA
## 1815              640               79             332              NA
## 1816              692              102             359              NA
## 1817              797              127             413              NA
## 1818              949              129             492              NA
## 1819              845               NA             511              NA
## 1820              981               NA             678              NA
## 1821              932               NA             633              NA
## 1822              364               NA             119             134
## 1823                0               NA               0              NA
## 1824                0               NA               0              NA
## 1825             1092               NA             155              14
## 1826             4224               NA            1095              NA
## 1827              890               NA             692             399
## 1828             1257               NA             194             343
## 1829              402               84             159              NA
## 1830             1072               78             397              NA
## 1895              845               NA             501              NA
## 1896              652               NA             451              NA
## 1897              760               NA             558             211
## 1898              565               NA             429             547
## 1899              697               NA             564             388
## 1900              638               NA             496             375
## 1901              732               NA             592             357
## 1902              594               NA             484             275
## 1903              540               NA             503             323
## 1904              420               NA             342             367
## 1905              326               NA             266             392
## 1906              446               NA             361             361
## 1907              414               NA             335             339
## 1908              500               NA             404             406
## 1909              434               NA             402             231
## 1910              420               NA             389             253
## 1914               NA              101              NA             308
## 1915               NA               98              NA             300
## 1916               NA               98              NA             308
## 1917               NA               90              NA             307
## 1918               NA               83              NA             190
## 1919               NA              105              NA             248
## 2012             1114               NA             777              NA
## 2013               927               NA              618              NA
## 2014              1157               NA              871             207
## 2015                 0               NA                0              NA
## 2016                 0               NA                0              NA
## 2017               927               83              458              NA
## 2018               656               NA              397              NA
## 2019               777               NA              513              NA
## 2020               590               NA              408              NA
## 2021               577               NA              495             392
## 2022               490               NA              408             697
## 2023               750               NA              625             562
## 2024               686               NA              572             403
## 2025               696               NA              584             366
## 2026               624               NA              524             337
## 2027               572               NA              530             226
## 2028               378               NA              308             307
## 2029               427               NA              345             254
## 2030               466               NA              374             231
## 2031               480               NA              388             213
## 2032               469               NA              434             112
## 2033               301               NA              281             225
## 2037                NA              131               NA             202
## 2038                NA               87               NA             209
## 2039                NA               87               NA             170
## 2040                NA               99               NA             119
## 2041                NA              111               NA             132
## 2042                NA               72               NA             158
## 2043                NA               96               NA             172
## 2136             19278               NA              952              NA
## 2137              1275               86              551              NA
## 2138               784               NA              615             372
## 2190               841               NA              685             517
## 2191              1313               NA              843              NA
## 2219               636               NA              110             359
## 2220               590               NA               91              58
## 2221               990              113              507              NA
## 2222               893              135              463              NA
## 2223              1136              137              582              NA
## 2224               698               NA              582             263
## 2225               636               NA              593             296
## 2226               374               NA              298             193
## 2227               565               NA              460             306
## 2228               634               NA              501             300
## 2229               524               NA              427             255
## 2230               438               NA              411             210
## 2231               387               NA              363             188
## 2232               569               NA              137              21
## 2233                 0               NA                0              NA
## 2234              1167               NA              807              NA
## 2235              1061               NA              786             193
## 2236               598               NA              450             444
## 2237               729               NA              603             400
## 2238               779               NA              596             303
## 2239                 0               NA                0              NA
## 2240             1060             96            543              NA
## 2241              660             99            334              NA
## 2242             1126            127            584              NA
## 2276             2492             NA            969              NA
```

```
summary( lm(baseball_df_fix, formula = TEAM_BATTING_SO ~.-TARGET_WINS))
```

```
## 
## Call:
## lm(formula = TEAM_BATTING_SO ~ . - TARGET_WINS, data = baseball_df_fix)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -112.307   -3.515   -0.733    2.989  100.385 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      20.011705   6.396987   3.128  0.00179 ** 
## TEAM_BATTING_H    0.287453   0.016182  17.763  < 2e-16 ***
## TEAM_BATTING_2B   0.004292   0.009945   0.432  0.66608    
## TEAM_BATTING_3B  -0.024591   0.020178  -1.219  0.22312    
## TEAM_BATTING_HR   2.173703   0.071019  30.607  < 2e-16 ***
## TEAM_BATTING_BB   0.264486   0.044878   5.893 4.50e-09 ***
## TEAM_BASERUN_SB   0.017420   0.005876   2.964  0.00307 ** 
## TEAM_PITCHING_H  -0.281532   0.014499 -19.417  < 2e-16 ***
## TEAM_PITCHING_HR -2.072216   0.068079 -30.438  < 2e-16 ***
## TEAM_PITCHING_BB -0.253951   0.042640  -5.956 3.10e-09 ***
## TEAM_PITCHING_SO  0.945191   0.002419 390.665  < 2e-16 ***
## TEAM_FIELDING_E   0.007461   0.007813   0.955  0.33973    
## TEAM_FIELDING_DP -0.011230   0.013192  -0.851  0.39475    
## TEAM_CS_YES_NO   -0.242665   0.854894  -0.284  0.77655    
## TEAM_HBP_YES_NO   4.013018   1.027459   3.906 9.74e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 10.81 on 1820 degrees of freedom
##   (441 observations deleted due to missingness)
## Multiple R-squared:  0.9975, Adjusted R-squared:  0.9975 
## F-statistic: 5.264e+04 on 14 and 1820 DF,  p-value: < 2.2e-16
```

```
summary( lm(baseball_df_fix, formula = TEAM_PITCHING_SO ~.-TARGET_WINS))
```

```
## 
## Call:
## lm(formula = TEAM_PITCHING_SO ~ . - TARGET_WINS, data = baseball_df_fix)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -109.203   -2.921    0.575    3.335  121.537 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       0.602953   6.745982   0.089  0.92879    
## TEAM_BATTING_H   -0.316956   0.016873 -18.785  < 2e-16 ***
## TEAM_BATTING_2B   0.005710   0.010459   0.546  0.58521
## TEAM_BATTING_3B   0.011791   0.021229   0.555  0.57866
## TEAM_BATTING_HR  -2.299366   0.074470 -30.876  < 2e-16 ***
## TEAM_BATTING_BB  -0.269092   0.047228  -5.698 1.41e-08 ***
## TEAM_BATTING_SO   1.045519   0.002676 390.665  < 2e-16 ***
## TEAM_BASERUN_SB  -0.007573   0.006193  -1.223  0.22149
## TEAM_PITCHING_H   0.299708   0.015210  19.704  < 2e-16 ***
## TEAM_PITCHING_HR  2.218126   0.070945  31.265  < 2e-16 ***
## TEAM_PITCHING_BB  0.255925   0.044882   5.702 1.38e-08 ***
## TEAM_FIELDING_E  -0.011783   0.008215  -1.434  0.15163
## TEAM_FIELDING_DP  0.005163   0.013877   0.372  0.70989
## TEAM_CS_YES_NO    1.134114   0.898748   1.262  0.20715
## TEAM_HBP_YES_NO  -3.263506   1.082434  -3.015  0.00261 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.37 on 1820 degrees of freedom
##   (441 observations deleted due to missingness)
## Multiple R-squared:  0.9974, Adjusted R-squared:  0.9974
## F-statistic: 4.991e+04 on 14 and 1820 DF,  p-value: < 2.2e-16
```

It appears that TEAM_PITCHING_SO and TEAM_BATTING_SO are missing all of the same rows. By quickly running a linear model for either column shows that it's possible to approximate values from other season records.

```
baseball_df%>%
  dplyr::filter(TEAM_PITCHING_SO < 5)
```

```
##    INDEX TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
## 1    325         120           2270             301             132
## 2    326         146           2305             322             111
## 3    435          65           1464             147              32
## 4    459          23           1458             220              35
## 5    952          77           1895             244               8
## 6    953          73           1685             206              31
## 7   1106          49           1794             281              58
## 8   1107         107           1725             194              67
## 9   1347           0            891             135               0
## 10  1498          24           1289             145              41
## 11  1502         105           1767             249              77
## 12  1503          71           1491             200              57
## 13  2037          97           1903             256              50
## 14  2038         118           2086             280             135
## 15  2048          81           1927             207             142
## 16  2049          88           1622             155              67
## 17  2253          34           1177             171               9
## 18  2254          93           1527             200              64
## 19  2486          12           1009             112              75
## 20  2493          29           1122              69              64
##    TEAM_BATTING_HR TEAM_BATTING_BB TEAM_BATTING_SO TEAM_BASERUN_SB
## 1               42              74               0              NA
## 2               29              64               0              NA
## 3                3             94               0              NA
## 4                0             93               0              NA
## 5                8             93               0              NA
## 6                0             58               0              NA
## 7                6             79               0              NA
## 8                4             79               0              NA
## 9                0               0               0               0
## 10               7             45               0              NA
## 11              20             95               0              NA
## 12              17             50               0              NA
## 13              18             71               0              NA
## 14              22             89               0              NA
## 15               8             78               0              NA
## 16              12             52               0              NA
## 17               0            119               0              NA
## 18               0             79               0              NA
## 19               0             12               0              NA
## 20               0             29               0              NA
##    TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR
## 1               NA               NA            5253               97
## 2               NA               NA            4727               59
## 3               NA               NA            4312                9
## 4               NA               NA           16871                0
## 5               NA               NA            5203               22
## 6               NA               NA            4074                0
## 7               NA               NA            5484               18
## 8               NA               NA            3408                8
## 9                0               NA           24057                0
## 10              NA               NA            4443               24
## 11              NA               NA            4404               50
## 12              NA               NA            3552               41
## 13              NA               NA            5605               53
## 14              NA               NA            4629               49
## 15              NA               NA            5382               22
## 16              NA               NA            3864               29
## 17              NA               NA           10035                0
## 18              NA               NA            3638                0
## 19              NA               NA           12574                0
## 20              NA               NA            6492                0
##    TEAM_PITCHING_BB TEAM_PITCHING_SO TEAM_FIELDING_E TEAM_FIELDING_DP
## 1               171                0            1058               NA
## 2               131                0             951               NA
## 3               277                0            1473               NA
## 4              1076                0            1898               NA
## 5               255                0            1225               NA
## 6               140                0             931               NA
## 7               241                0            1531               NA
## 8               156                0             853               NA
## 9                 0                0            1890               NA
## 10              155                0            1506               NA
## 11              237                0            1092               NA
## 12              119                0            1253               NA
## 13              209                0            1166               NA
## 14              198                0             928               NA
## 15              218                0            1447               NA
## 16              124                0            1132               NA
## 17             1015                0            1279               NA
## 18              188                0            1010               NA
## 19              150                0             847               NA
## 20              168                0            1522               NA
```

```
baseball_df%>%
  dplyr::filter(TEAM_BATTING_SO < 5)
```

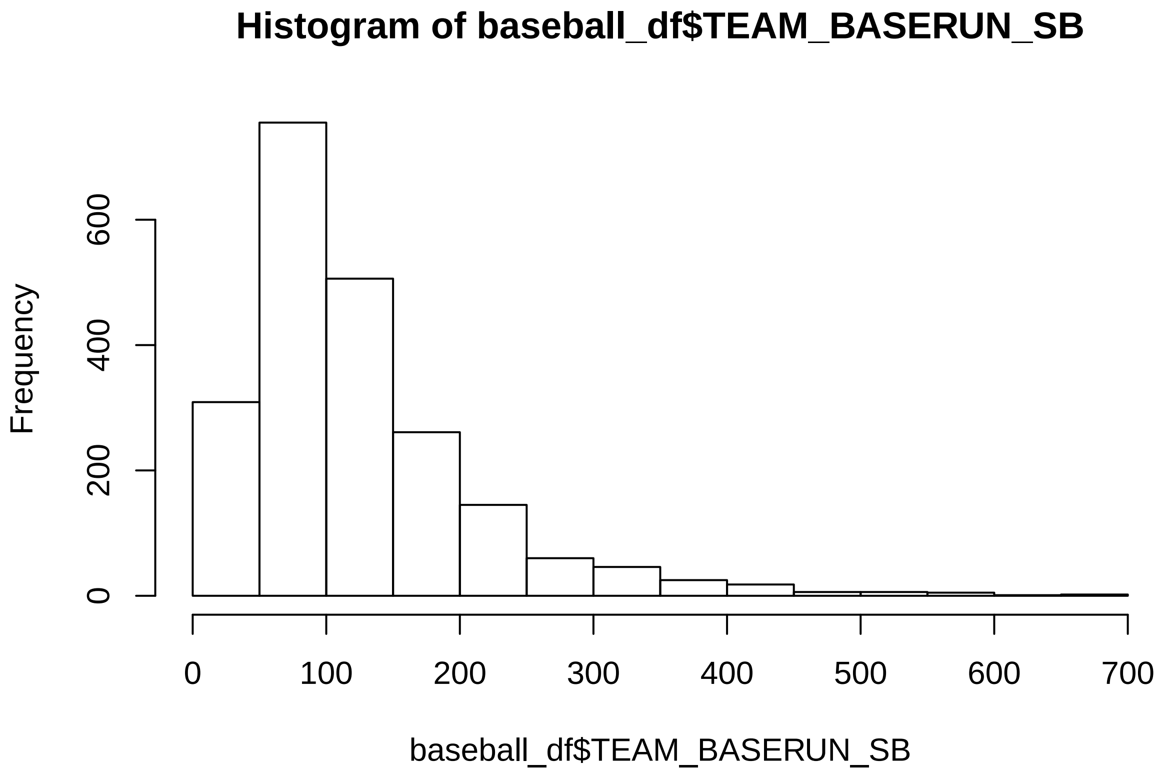
```
##    INDEX TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
## 1    325         120           2270             301             132
## 2    326         146           2305             322             111
## 3    435          65           1464             147              32
## 4    459          23           1458             220              35
## 5    952          77           1895             244               8
## 6    953          73           1685             206              31
## 7   1106          49           1794             281              58
## 8   1107         107           1725             194              67
## 9   1347           0            891             135               0
## 10  1498          24           1289             145              41
## 11  1502         105           1767             249              77
## 12  1503          71           1491             200              57
## 13  2037          97           1903             256              50
## 14  2038         118           2086             280             135
## 15  2048          81           1927             207             142
## 16  2049          88           1622             155              67
## 17  2253          34           1177             171               9
## 18  2254          93           1527             200              64
## 19  2486          12           1009             112              75
## 20  2493          29           1122              69              64
##    TEAM_BATTING_HR TEAM_BATTING_BB TEAM_BATTING_SO TEAM_BASERUN_SB
## 1               42              74               0              NA
## 2               29              64               0              NA
## 3                3              94               0              NA
## 4                0              93               0              NA
## 5                8              93               0              NA
## 6                0              58               0              NA
## 7                6              79               0              NA
## 8                4              79               0              NA
## 9                0               0               0               0
## 10               7              45               0              NA
## 11              20              95               0              NA
## 12              17              50               0              NA
## 13              18              71               0              NA
## 14              22              89               0              NA
## 15               8              78               0              NA
## 16              12              52               0              NA
## 17               0             119               0              NA
## 18               0              79               0              NA
## 19               0              12               0              NA
## 20               0              29               0              NA
##    TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR
## 1               NA               NA            5253               97
## 2               NA              NA             4727               59
## 3               NA              NA             4312                9
## 4               NA              NA            16871                0
## 5               NA              NA             5203               22
## 6               NA              NA             4074                0
## 7               NA              NA             5484               18
## 8               NA              NA             3408                8
## 9                0              NA            24057                0
## 10              NA              NA             4443               24
## 11              NA              NA             4404               50
## 12              NA              NA             3552               41
## 13              NA              NA             5605               53
## 14              NA              NA             4629               49
## 15              NA              NA             5382               22
## 16              NA              NA             3864               29
## 17              NA              NA            10035                0
## 18              NA              NA             3638                0
## 19              NA              NA            12574                0
## 20              NA              NA             6492                0
##    TEAM_PITCHING_BB TEAM_PITCHING_SO TEAM_FIELDING_E TEAM_FIELDING_DP
## 1               171                0            1058               NA
## 2               131                0             951               NA
## 3               277                0            1473               NA
## 4              1076                0            1898               NA
## 5               255                0            1225               NA
## 6               140                0             931               NA
## 7               241                0            1531               NA
## 8               156                0             853               NA
## 9                 0                0            1890               NA
## 10              155                0            1506               NA
## 11              237                0            1092               NA
## 12              119                0            1253               NA
## 13              209                0            1166               NA
## 14              198                0             928               NA
## 15              218                0            1447               NA
## 16              124                0            1132               NA
## 17             1015                0            1279               NA
## 18              188                0            1010               NA
## 19              150                0             847               NA
## 20              168                0            1522               NA
```

Lookin closer at these two variables, there are also some values that may be omitted if they are implausibly small. Going a whole season with zero strikeouts, pitching or batting, seems unlikely. It may make sense to recode these as NA and impute values onto them.

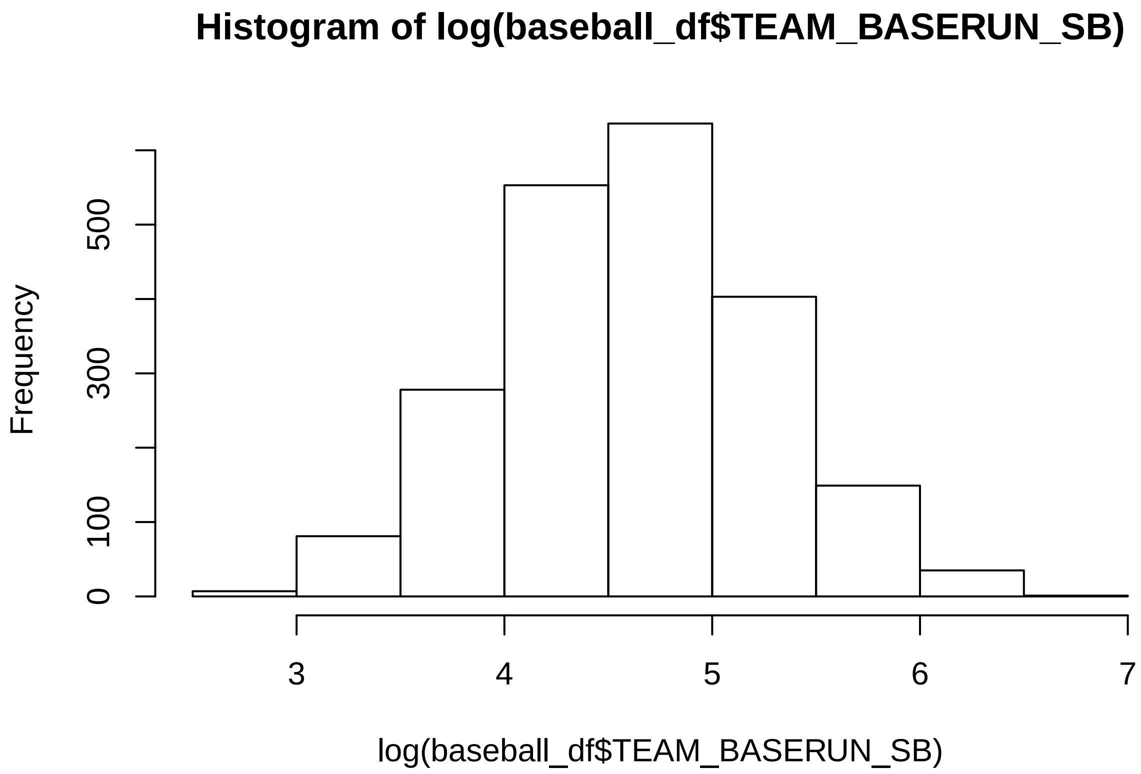
```r
baseball_df %>%
  dplyr::select(TEAM_BASERUN_CS, TEAM_BASERUN_SB) %>%
  filter(is.na(TEAM_BASERUN_SB))
```

```
##     TEAM_BASERUN_CS TEAM_BASERUN_SB
## 1                NA              NA
## 2                NA              NA
## 3                NA              NA
## 4                NA              NA
## 5                NA              NA
## 6                NA              NA
## 7                NA              NA
## 8                NA              NA
## 9                NA              NA
## 10               NA              NA
## 11               NA              NA
## 12               NA              NA
## 13               NA              NA
## 14               NA              NA
## 15               NA              NA
## 16               NA              NA
## 17               NA              NA
## 18               NA              NA
## 19               NA              NA
## 20               NA              NA
## 21               NA              NA
## 22               NA              NA
## 23               NA              NA
## 24               NA              NA
## 25               NA              NA
## 26               NA              NA
## 27               NA              NA
## 28               NA              NA
## 29               NA              NA
## 30               NA              NA
## 31               NA              NA
## 32               NA              NA
## 33               NA              NA
## 34               NA              NA
## 35               NA              NA
## 36               NA              NA
## 37               NA              NA
## 38               NA              NA
## 39               NA              NA
## 40               NA              NA
## 41               NA              NA
## 42               NA              NA
## 43               NA              NA
## 44               NA              NA
## 45               NA              NA
## 46               NA              NA
## 47               NA              NA
## 48               NA              NA
## 49               NA              NA
## 50               NA              NA
## 51               NA              NA
## 52               NA              NA
## 53               NA              NA
## 54               NA              NA
## 55               NA              NA
## 56               NA              NA
## 57               NA              NA
## 58               NA              NA
## 59               NA              NA
## 60               NA              NA
## 61               NA              NA
## 62               NA              NA
## 63               NA              NA
## 64               NA              NA
## 65               NA              NA
## 66               NA              NA
## 67               NA              NA
## 68               NA              NA
## 69               NA              NA
## 70               NA              NA
## 71               NA              NA
## 72               NA              NA
## 73               NA              NA
## 74               NA              NA
## 75               NA              NA
## 76               NA              NA
## 77               NA              NA
## 78               NA              NA
## 79               NA              NA
## 80               NA              NA
## 81               NA              NA
## 82               NA              NA
## 83               NA              NA
## 84               NA              NA
## 85               NA              NA
## 86               NA              NA
## 87               NA              NA
## 88               NA              NA
## 89               NA              NA
## 90               NA              NA
## 91               NA              NA
## 92               NA              NA
## 93               NA              NA
## 94               NA              NA
## 95               NA              NA
## 96               NA              NA
## 97               NA              NA
## 98               NA              NA
## 99               NA              NA
## 100              NA              NA
## 101              NA              NA
## 102              NA              NA
## 103              NA              NA
## 104              NA              NA
## 105              NA              NA
## 106              NA              NA
## 107              NA              NA
## 108              NA              NA
## 109              NA              NA
## 110              NA              NA
## 111              NA              NA
## 112              NA              NA
## 113              NA              NA
## 114              NA              NA
## 115              NA              NA
## 116              NA              NA
## 117              NA              NA
## 118              NA              NA
## 119              NA              NA
## 120              NA              NA
## 121              NA              NA
## 122              NA              NA
## 123              NA              NA
## 124              NA              NA
## 125              NA              NA
## 126              NA              NA
## 127              NA              NA
## 128              NA              NA
## 129              NA              NA
## 130              NA              NA
## 131              NA              NA
```

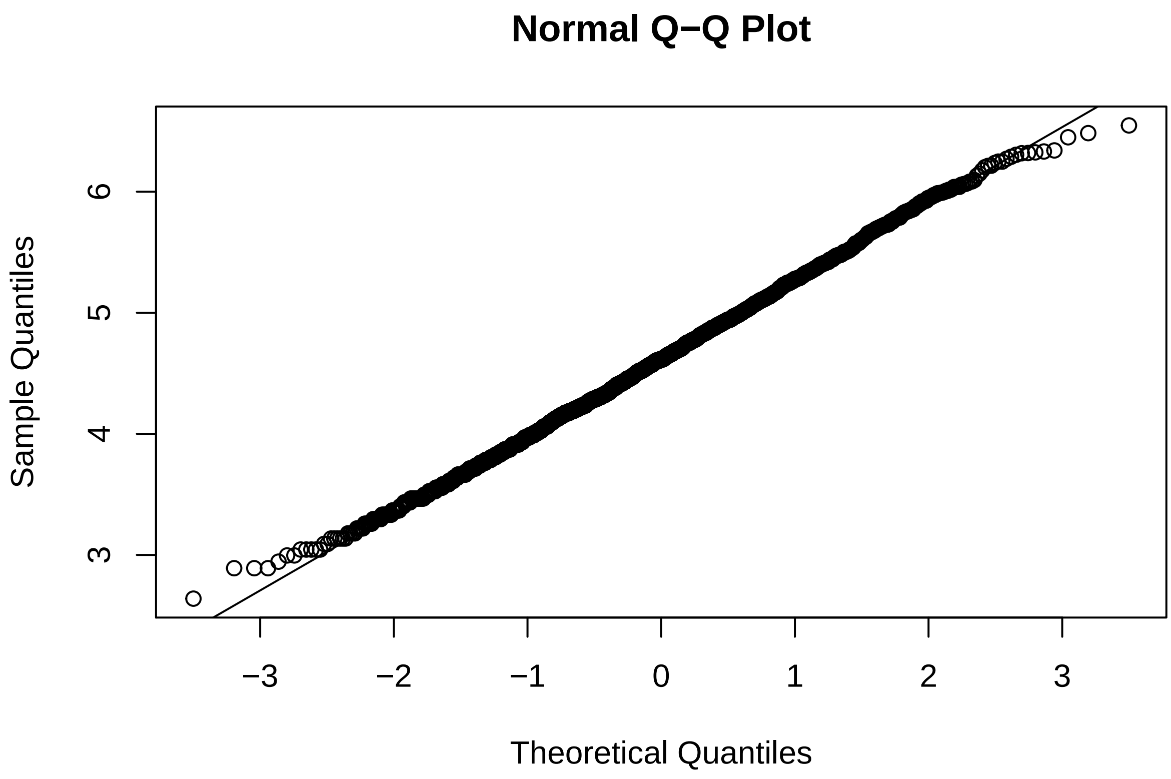
```
hist(baseball_df$TEAM_BASERUN_SB)
```



```
hist(log(baseball_df$TEAM_BASERUN_SB))
```



```
baseball_log_sb <- baseball_df %>%
  filter(!is.na(TEAM_BASERUN_SB)) %>%
  filter(TEAM_BASERUN_SB != 0) %>%
  mutate(LOG_BASERUN_SB = log(TEAM_BASERUN_SB))



qqnorm((baseball_log_sb$LOG_BASERUN_SB))
qqline((baseball_log_sb$LOG_BASERUN_SB))
```



The column TEAM_BASERUN_SB is partly correlated with TEAM_BASERUN_CS in the training set. However, there are many misisng values so single imputation may not be an option. In this case, TEAM_BASERUN_SB may qualify for multiple imputation after log transform to make it normally distributed.

```
baseball_df_fix %>%
  filter(is.na(TEAM_FIELDING_DP))
```

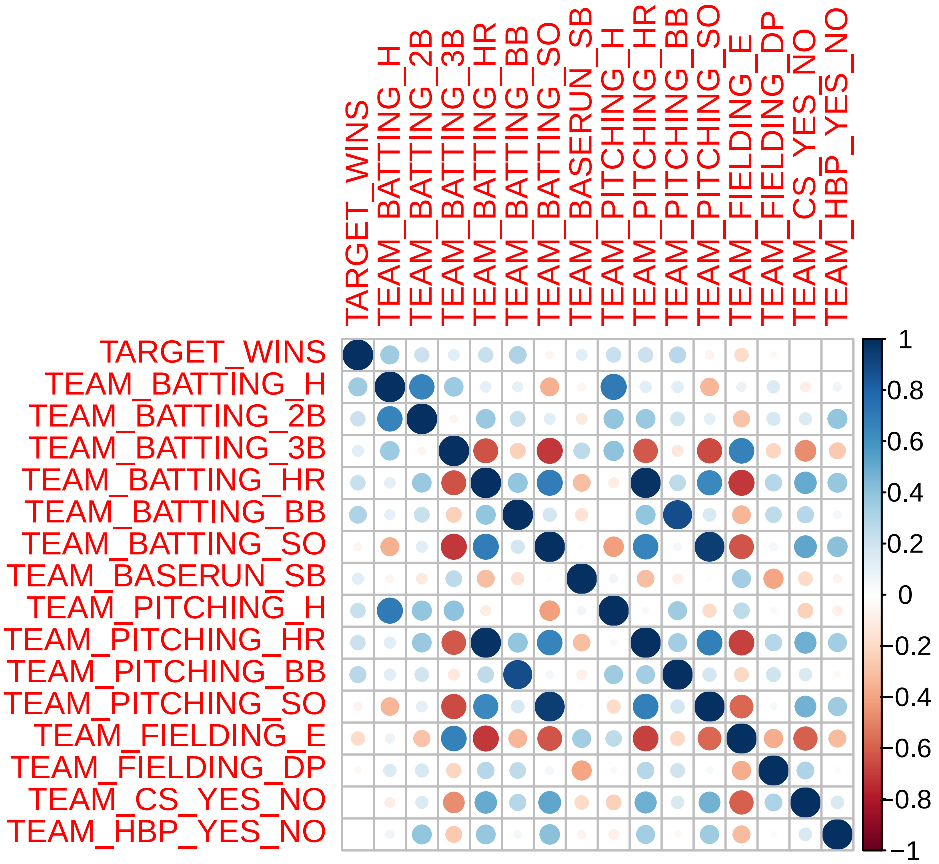
```
##     TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_HR
## 1            39           1445             194              39              13
## 2           104           1670             345             142              56
## 3           107           1551             261              88              53
## 4            67           1323             208              77              32
## 5            78           1502             209              82              33
## 6            82           1680             248             126              71
## 7            85           1430             202             108              68
## 8           105           1583             248              68              53
## 9            93           1486             213              76              38
## 10          102           1484             212              94              63
## 11          108           1705             224              63              82
## 12          102           2035             334             115             126
## 13           88           1693             244              70              67
## 14           92           1751             216              92              45
## 15          114           1932             282             102              55
## 16          111           1665             207              60              58
## 17          101           1617             190              96              42
## 18           44           1533             221              79              25
## 19          105           2241             282             105              42
## 20           98           2343             344             113              35
## 21           47            1468             211              83               8
## 22           96            1369             203             128              49
## 23           61            1244             184              88              25
## 24           59            1169             153              63              10
## 25           92            1604             242             120              37
## 26           67            1263             192              83              22
## 27           84            1451             186              82              24
## 28           71            1325             162              76              10
## 29           85            1460             170             119              36
## 30           51            1479             176             122              33
## 31           75            1596             204             107              34
## 32          113            2084             343             190              42
## 33          108            1907             293             111              31
## 34          113            1944             260             126              29
## 35          112            1974             303              82              24
## 36          104            1722             167              84              13
## 37           94            1652             223              78              19
## 38           89            1469             231              40              26
## 39          108            2300             378             200              16
## 40          134            2333             393             107              24
## 41          118            2554             376             126              36
## 42          120            2270             301             132              42
## 43          146            2305             322             111              29
## 44           39            1620             201              44               0
## 45           51            1764             180              81              18
## 46           65            1464             147              32               3
## 47           86            1379             250              48              28
## 48           43            1258             190              77              21
## 49          102            1640             280              95              68
## 50           87            1767             307              99              13
## 51           93            1604             238             101              57
## 52           55            1418             216              72              33
## 53           93            1662             228             114              42
## 54           53            1426             169             132              25
## 55          104            1352             161             144              60
## 56           92            1423             307              63              13
## 57           51            1351             213              71              23
## 58           23            1458             220              35               0
## 59           56            1832             196             223              39
## 60           44            2003             272              44               0
## 61           98            1653             458             101              21
## 62           90            1701             234              72             205
## 63          126            1561             266             108              78
## 64          118            1598             259             114              69
## 65           95            1576             238             131             107
## 66           92            1441             176             114              92
## 67           82            1564             226              81              97
## 68           98            1480             191             106              72
## 69           78            1318             165             102              29
## 70           70            1908             325             106              80
## 71           90            1659             230             123              43
## 72           72            1664             232             119              47
## 73           92            1545             189              91              19
## 74           82            1563             189              90              30
## 75           89           1550             236             112             37
## 76          101           1589             202             122             56
## 77          101           1541             162             143             54
## 78           91           1513             156             111             38
## 79           76           1438             170             112             53
## 80           97           1542             215             122             44
## 81           97           1404             160              99             39
## 82           89           1523             230             112             61
## 83           95           1478             184             147             33
## 84           66           1369             175             106             47
## 85           89           1391             167              81             48
## 86           82           1512             204              82             37
## 87           69           1712             279              83             76
## 88           82           1738             293             131             45
## 89           93           1609             269              85             27
## 90           98           1543             221             108             20
## 91           90           1554             210             113             14
## 92           70           1358             170              94             45
## 93           69           1711             265             117             34
## 94           70           1464             201             112             42
## 95           96           1481             192              93             19
## 96           72           1316             182              72             24
## 97           48           1447             220              95             17
## 98           61           1321             157              72             15
## 99           54           1317             162              72             26
## 100          76           1508             213             103             26
## 101         101           1495             213             104             28
## 102          92           1804             281             124             41
## 103          85           1811             303             113             46
## 104         105           1773             242              83             36
## 105         101           1852             262              91             35
## 106          85           1699             237             109             20
## 107          88           1499             176              61             20
## 108          21           1402             149              53             13
## 109          77           1895             244               8              8
## 110          73           1685             206              31              0
## 111         106           1585             182              97             40
## 112          73           1321             226              46             18
## 113          66           1539             271              79             21
## 114          62           1376             224              99             38
## 115         115           1660             232             107             70
## 116         103           1834             278             165             72
## 117          84           1577             219              89             63
## 118          17           1313             145              34              0
## 119          49           1794             281              58              6
## 120         107           1725             194              67              4
## 121          44           1347             195              94             30
## 122          48           1389             208              90             42
## 123          60           1320             216              40             41
## 124          71           1639             276              42             75
## 125          53           1227             174              75             23
## 126          65           1485             192              90             21
## 127          33           1142             213              31             12
## 128          60           1518             162              68             17
## 129          62            1316             174              73              25
## 130          77            1397             175              94              20
## 131          90            1491             232              95              19
## 132          73            1549             242              99              30
## 133         110            1496             222              93              56
## 134         108            1464             231              94              54
## 135          72            1458             236              82              27
## 136         100            1514             193             110              32
## 137          82            1519             219             105              57
## 138          87            1864             282             161              52
## 139          88            1645             234              95              48
## 140          72            1598             215             108              35
## 141          75            1648             248              88              29
## 142          60            1468             174              74              19
## 143          87            1488             190             107              23
## 144         102            1506             228             104              26
## 145          77            1426             182             120              27
## 146          79            1541             217             105              24
## 147          91            1691             231             117              32
## 148          58            1412             220              80              17
## 149          32            1466             200              88              26
## 150         108            1608             191              80              18
## 151          64            1475             154              82              20
## 152          67            1287             142              65              19
## 153          65            1536             229              95              25
## 154          45            1503             216             110              52
## 155          43            1632             211              90              42
## 156          47            1480             176              99              46
## 157          65            1492             199              87              50
## 158          75            1490             161              76              34
## 159          80            1582             205              72              43
## 160          34            2059             209              54               7
## 161           0             891             135               0               0
## 162          95            1494             261              68              59
## 163         108            1188             338               0               0
## 164          24            1289             145              41               7
## 165          79            1978             211             103               5
## 166         102            2004             258              39              12
## 167          89            1901             205             125              15
## 168         105            1767             249              77              20
## 169          71            1491             200              57              17
## 170          92            1723             252             120              82
## 171          91            1490             187              98              10
## 172         114            1593             235              97              33
## 173          64            1256             130              86              22
## 174          54            1458             235              80              26
## 175          36            1674             216              54               0
## 176          76            1509             213             143              44
## 177         109            1607             246              83              33
## 178          92            1600             253             151              39
## 179          80            1590             244             110              43
## 180          76            1370             230              98              25
## 181          99            1473             223             108              38
## 182          66            1297             222              63              29
## 183          56           1351             216              56              20
## 184         101           1387             206              94              37
## 185          99           1671             281             117              62
## 186          86           1272             188              57              20
## 187          80           1592             274              66              56
## 188          96           1567             272              96              28
## 189          80           1471             213              60              25
## 190          92           1504             238             101              53
## 191          90           1950             309             113             100
## 192          90           2192             319             166              51
## 193          77           1722             292             105              61
## 194          68           1708             261             102              49
## 195          85           1556             259              88              36
## 196         100           1719             257              88              33
## 197          51           1494             261              85              17
## 198          84           1667             231             140              61
## 199          85           1516             212             143              64
## 200          51           1475             198              96              21
## 201          45           1166             158              75               3
## 202          82           1394             180             115               7
## 203          95           1385             220             114              19
## 204          72           1491             239             102              26
## 205          80           1294             181              59              17
## 206          27           1296             191              51              24
## 207          85           1364             151             114              40
## 208         102           1817             221             159              46
## 209          87           1656             233             109              32
## 210          83           1722             212             118              34
## 211          74           1566             173             134              31
## 212          79           1437             153              96              15
## 213         122           2372             382             156              52
## 214         110           2496             284              85              15
## 215          97           1903             256              50              18
## 216         118           2086             280             135              22
## 217          96           1655             312              98              35
## 218         122           1428             221              62              30
## 219          78           1208             168              44               9
## 220         110           1972             254              61              24
## 221          81           1927             207             142               8
## 222          88           1622             155              67              12
## 223          14           1437             148              56               0
## 224          46           1254             154             127              27
## 225          81           1399             168              82              40
## 226          26           1776             285             162              19
## 227          78           1519             235             116              41
## 228         123           1569             217             119              23
## 229         102           1574             238              93              29
## 230          90           1658             220             122              63
## 231         104           1421             161              94              68
## 232         107           1696             267              99              67
## 233          78           1546             257             110              31
## 234          87           1560             232              88              56
## 235          76           1423             186              91              42
## 236          83           1748             223             124              75
## 237         108           1775             242             118              53
## 238          82           1637             236             111              40
## 239          79           1710             197             108              49
## 240         103           1792             232             104              38
## 241          83           1536             205              93              37
## 242          65           1545             174              70              25
## 243         135           1793             371              59              46
## 244          54           1244             182              32              12
## 245          57           1329             243              61              40
## 246          34           1177             171               9               0
## 247          93           1527             200              64               0
## 248         107           1475             195              76              12
## 249         101           1493             229              91              17
## 250         114           1416             191              82              25
## 251         108           1591             240              99              23
## 252         114           1860             313              94              47
## 253         110           1427             179              56              43
## 254         108           1574             253              77              70
## 255          93           1558             212              87              57
## 256         101           1584             201              61              69
## 257          60           1282             149              57              49
## 258          70           1581             187             120              12
## 259          48           1662             192             109              47
## 260          50           1448             167              97              46
## 261          36           1579             184              83              38
## 262          42           1393             161              59              14
## 263          90           1624             185              94              50
## 264          41            992             263              20               0
## 265          70           1477             193              75              18
## 266          83           1414             187             133              29
## 267          72           1338             238              75              12
## 268          75           2222             295             197              35
## 269          97           2132             363              71              32
## 270          53           1420             176             101              23
## 271          62           1336             160              84              40
## 272          50           1582             226             104              29
## 273          55           1616             268             145              72
## 274          54           1663             262             128              70
## 275          75           1689             238              95              44
## 276          54           1517             189              85              38
## 277          58           1523             173              93              50
## 278          33           1695             187             104              12
## 279          12           1009             112              75               0
## 280          68           1347             174              38               6
## 281          38           1156             182              69              31
## 282          61           1380             198              84              62
## 283          58           1141             118              59              36
## 284          54           1444             197              74              33
## 285          29           1122              69              64               0
## 286          31           1116             157              62              15
##      TEAM_BATTING_BB TEAM_BATTING_SO TEAM_BASERUN_SB TEAM_PITCHING_H
## 1                143             842              NA            9364
## 2                203             699              NA            2761
## 3                302             963              NA            2264
## 4               275             755              NA            1914
## 5               346             744             216            2080
## 6               455             525             499            2249
## 7               341             633             354            1729
## 8               596             570             419            2003
## 9               646             627             347            1810
## 10              625             632             339            1742
## 11              705             367             305            2141
## 12              657             320             296            2498
## 13              618             292             246            2094
## 14              512             339             298            2165
## 15              519             322             286            2371
## 16              440             329             187            1810
## 17              459             287             197            1723
## 18              664             450             196            1881
## 19               81              84             105            6723
## 20              116              72              61            6778
## 21              273             559              NA            2477
## 22              322             833              NA            2092
## 23              415             786              NA            1849
## 24              469             746             333            1446
## 25              563             401             654            1925
## 26              352             566             385            1493
## 27              502             643             373            1741
## 28              596             724             481            6313
## 29              661             664             410            1752
## 30              550             529             250            1630
## 31              669             403             290            1989
## 32              653             253             410            2638
## 33              442             303             386            2376
## 34              485             252             554            2441
## 35              545             319             500            2460
## 36              564             344             272            1872
## 37              458             419             398            1808
## 38              222            1006              NA            2266
## 39              324             103             394           12420
## 40               97              90             162            8041
## 41              170              66             107            7013
## 42               74               0              NA            5253
## 43               64               0              NA            4727
## 44               83             105              74            7093
## 45              159             129              54            5292
## 46               94               0              NA            4312
## 47              190            1170              NA            2050
## 48              537             746             298            2059
## 49              819             546             517            2060
## 50              246             571              NA            2951
## 51              314             668              NA            2341
## 52              259             550              NA            2051
## 53              616             453             334            2040
## 54              224             685              NA            2382
## 55              294             606              NA            2028
## 56              232             625              NA            2377
## 57              246             833              NA            1954
## 58               93               0              NA           16871
## 59              145             140             101           10234
## 60              125              96              88           14749
## 61              213             660              NA            2733
## 62              382             678              NA            2460
## 63              492             621              NA            2258
## 64              601             670             278            2088
## 65              545             536             511            2110
## 66              348             676             344            1729
## 67              636             633             298            1919
## 68              631             548             286            1776
## 69              474             535             259            1462
## 70              609             366             401            2342
## 71              518             367             420            2100
## 72              528             389             324            2042
## 73              514             426             238            1669
## 74              444             374             270            1711
## 75              582             485             326            1833
## 76              230             431              NA            2627
## 77              229             600              NA            2290
## 78              221             608              NA            2188
## 79              439             743             217            1688
## 80              458             439             632            1850
## 81              417             671             567            1697
## 82              527             596             538            1775
## 83              531             463             383            1814
## 84              490             519             289            1619
## 85              543             512             292            1502
## 86              673             324             301            1914
## 87              633             314             268            2133
## 88              516             310             406            2166
## 89              466             268             238            1975
## 90              485             320             176            1645
## 91              524             319             246            1678
## 92              693             624             266            2200
## 93              634             430             224            2132
## 94              595             715             357            1719
## 95              659             673             427            1790
## 96              626             627             331            1556
## 97              464             573             439            1789
## 98              387             686             433            1621
## 99              610             582             187            1616
## 100             605             541             282            1758
## 101             600             583             245            1625
## 102             673             290             319            2283
## 103             591             378             276            2274
## 104             588             450             233            2209
## 105             552             400             221            2344
## 106             538             425             224            2101
## 107             593             333             101            1630
## 108             304             295             134            1475
## 109              93               0              NA            5203
## 110              58               0              NA            4074
## 111             227             744              NA            2445
## 112               212               899                NA                2352
## 113               274               625                NA                2544
## 114               324               677                NA                2064
## 115               493               561               256                2186
## 116               460               337               349                2396
## 117               380               490               239                1950
## 118                78               162                45                7335
## 119                79                 0                NA                5484
## 120                79                 0                NA                3408
## 121               189               848                NA                2039
## 122               386               487               429                1786
## 123               283               590               420                1584
## 124               456               540               305                1981
## 125               353               741               315                1506
## 126               508               740               558                1756
## 127               252              1085                NA                2342
## 128               281                77               136               12943
## 129               279               650                NA                2050
## 130               344               469                NA                2021
## 131               512               618               293                1763
## 132               551               441               494                1873
## 133               650               474               460                1769
## 134               649               453               438                1839
## 135               549               514               398                1724
## 136               662               532               430                1593
## 137               599               375               270                1922
## 138               576               364               349                2305
## 139               491               393               226                2034
## 140               425               333               245                1976
## 141               431               313               230                2023
## 142               366               351               145                1640
## 143               234               509                NA                2485
## 144               219               612                NA                2259
## 145               220               648                NA                2063
## 146               488               665               241                1836
## 147               519               424               555                2014
## 148               386               725               382                1694
## 149               376               612               238                1721
## 150               503               565               419                1973
## 151               524               559               272                1744
## 152               461               541               293                1372
## 153               629               397               263                1991
## 154               436               454               270                1873
## 155               428               399               193                2018
## 156               459               528               241                1830
## 157               461               565               243                1859
## 158               402               460               252                1599
## 159               465               400               248                1686
## 160                34                81                34               13898
## 161                 0                 0                 0               24057
## 162               482               513               212                6723
## 163               270               945                NA               16038
## 164                45                 0                NA                4443
## 165               162                74               226                9710
## 166             165             156             168            6012
## 167             128              67              46            5811
## 168              95               0              NA            4404
## 169              50               0              NA            3552
## 170             601             450             286            2131
## 171             240             437              NA            2514
## 172             307             477              NA            2412
## 173             396             694             144            1507
## 174             535             564             372            1776
## 175              72              72               0           30132
## 176             720             475             315            1910
## 177             321             443              NA            2656
## 178             232             643              NA            2422
## 179             323             593              NA            2300
## 180             454             836             341            1644
## 181             369             576             529            1794
## 182             583             663             374            1592
## 183             302             741              NA            1954
## 184             401             733             321            1971
## 185             507             456             468            2201
## 186             334             604             307            1585
## 187             501             450             343            2031
## 188             646             498             414            1938
## 189             570             487             274            1739
## 190             559             545             229            1592
## 191             588             421             254            2449
## 192             628             310             346            2774
## 193             546             370             238            2146
## 194             490             367             200            2096
## 195             513             415             198            1692
## 196             470             363             226            1832
## 197             249             979              NA            3612
## 198             533             397             251            2061
## 199             521             639             174            1767
## 200             268             570              NA            2438
## 201             215             617              NA            1749
## 202             276             784              NA            2034
## 203             565             843             307            1638
## 204             417             498             289            1948
## 205             235             705             347            1564
## 206             486             546             248            1544
## 207             461             480             235            1444
## 208             674             343             264            2282
## 209             461             367             315            2032
## 210             486             359             273            2163
## 211             444             413             210            1937
## 212             368             375             117            1573
## 213             266             133             324           13724
## 214             254             173             214            9190
## 215              71               0              NA            5605
## 216              89               0              NA            4629
## 217             246             511              NA            2736
## 218             434             678              NA            2066
## 219             390             633              NA            1779
## 220             190             119             134            6028
## 221              78               0              NA            5382
## 222              52               0              NA            3864
## 223              56             155              14           10121
## 224             204            1095              NA            4837
## 225             573             692             399            1799
## 226             246             194             343           11508
## 227             214             501              NA            2563
## 228             320             451              NA            2269
## 229             323             558             211            2143
## 230             475             429             547            2184
## 231             334             564             388            1757
## 232             692             496             375            2181
## 233             433             592             357            1912
## 234             538             484             275            1915
## 235             547             503             323            1527
## 236             619             342             367            2145
## 237             584             266             392            2178
## 238             561             361             361            2024
## 239             543             335             339            2115
## 240             500             404             406            2216
## 241             462             402             231            1659
## 242             418             389             253            1669
## 243             259             777              NA            2570
## 244             321             618              NA            1866
## 245             312             871             207            1765
## 246             119               0              NA           10035
## 247              79               0              NA            3638
## 248             207             397              NA            2438
## 249             263             513              NA            2260
## 250             338             408              NA            2048
## 251             466             495             392            1854
## 252             530             408             697            2232
## 253             492             625             562            1712
## 254             592             572             403            1889
## 255             565             584             366            1856
## 256             744             524             337            1887
## 257             656             530             226            1385
## 258             643             308             307            1940
## 259             475             345             254            2055
## 260             414             374             231            1804
## 261             438             388             213            1953
## 262             414             434             112            1504
## 263             502             281             225            1742
## 264             142             952              NA           20088
## 265             583             615             372            1884
## 266             596             685             517            1735
## 267             245             843              NA            2084
## 268             284             110             359           12856
## 269              45              91              58           13815
## 270             562             582             263            1704
## 271             568             593             296            1433
## 272             658             298             193            1987
## 273             757             460             306            1983
## 274             656             501             300            2105
## 275             459             427             255            2073
## 276             394             411             210            1617
## 277             373             363             188            1623
## 278              79             137              21            7041
## 279              12               0              NA           12574
## 280             171             807              NA            1948
## 281             358             786             193            1561
## 282             357             450             444            1832
## 283             297             603             400            1379
## 284             609             596             303            1887
## 285              29               0              NA            6492
## 286             262             969              NA            2870
##     TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO TEAM_FIELDING_E
## 1                  84              927             5456            1011
## 2                  93              336             1155             631
## 3                  77              441             1405             546
## 4                  46              398             1092             644
## 5                  46              479             1030             644
## 6                  95              609              703             699
## 7                  82              412              765             597
## 8                  67              754              721             523
## 9                  46              787              764             436
## 10                 74              734              742             420
## 11                103              885              461             443
## 12                155              806              393             509
## 13                 83              764              361             450
## 14                 56              633              419             455
## 15                 68              637              395             334
## 16                 63              478              358             325
## 17                 45              489              306             314
## 18                 31              815              552             603
## 19                126              243              252            1296
## 20                101              336              208            1059
## 21                 14              461              943            1053
## 22                 75              492             1273             705
## 23                 37              617             1168             623
## 24                 12              580              923             648
## 25                 44              676              481             659
## 26                 26              416              669             545
## 27                 29              602              772             643
## 28                 48             2840             3450             519
## 29                 43              793              797             604
## 30                 36              606              583             644
## 31                 42              834              502             479
## 32                 53              826              320             371
## 33                 39              551              378             359
## 34                 36              609              316             372
## 35                 30              679              398             345
## 36                 14              613              374             354
## 37                 21              501              459             337
## 38                 40              343             1552             941
## 39                 86             1750              556            1215
## 40                 83              334              310             907
## 41                99              467              181            1192
## 42                97              171                0            1058
## 43                59              131                0             951
## 44                 0              363              460            1567
## 45                54              477              387            1515
## 46                 9              277                0            1473
## 47                42              282             1739             941
## 48                34              879             1221             645
## 49                85             1029              686             576
## 50                22              411              954             743
## 51                83              458              975             676
## 52                48              375              796             671
## 53                52              756              556             652
## 54                42              374             1144             894
## 55                90              441              909             650
## 56                22              387             1044             650
## 57                33              356             1205             741
## 58                 0             1076                0            1898
## 59               218              810              782            1246
## 60                 0              920              707            1237
## 61                35              352             1091             898
## 62               297              553              981             861
## 63               113              712              898             719
## 64                90              785              875             621
## 65               143              730              718             632
## 66               110              418              811             500
## 67               119              781              777             568
## 68                86              757              658             476
## 69                32              526              594             470
## 70                98              747              449             555
## 71                54              656              464             464
## 72                58              648              477             484
## 73                21              555              460             445
## 74                33              486              409             468
## 75                44              688              574             582
## 76                93              380              712             595
## 77                80              340              892             612
## 78                55              320              879             603
## 79                62              515              872             681
## 80                53              550              527             581
## 81                47              504              811             551
## 82                71              614              695             513
## 83                41              652              568             468
## 84                56              579              614             484
## 85                52              586              553             434
## 86                47              852              410             406
## 87                95              789              391             523
## 88                56              643              386             470
## 89                33              572              329             335
## 90                21              517              341             346
## 91                15              566              345             366
## 92                73             1123             1011             630
## 93                42              790              536             663
## 94                49              698              839             573
## 95               23              797              814            475
## 96               28              740              741            448
## 97               21              574              709            712
## 98               18              475              842            589
## 99               32              749              714            497
## 100              30              705              631            565
## 101              30              652              634            441
## 102              52              852              367            492
## 103              58              742              475            432
## 104              45              733              561            434
## 105              44              699              506            354
## 106              25              665              526            318
## 107              22              645              362            322
## 108              14              320              310            408
## 109              22              255                0           1225
## 110               0              140                0            931
## 111              62              350             1148            821
## 112              32              377             1600            817
## 113              35              453             1033            734
## 114              57              486             1016            668
## 115              92              649              739            491
## 116              94              601              440            502
## 117              78              470              606            573
## 118               0              436              905           1553
## 119              18              241                0           1531
## 120               8              156                0            853
## 121              45              286             1284            780
## 122              54              496              626            616
## 123              49              340              708            539
## 124              91              551              653            507
## 125              28              433              909            622
## 126              25              601              875            722
## 127              25              517             2225           1066
## 128             145             2396              657           1390
## 129              39              435             1013            791
## 130              29              498              678            631
## 131              22              605              731            721
## 132              36              666              533            679
## 133              66              769              560            498
## 134              68              815              569            401
## 135              32              649              608            511
## 136              34              696              560            419
## 137              72              758              475            487
## 138              64              712              450            482
## 139              59              607              486            402
## 140              43              526              412            367
## 141              36              529              384            447
## 142              21              409              392            373
## 143              38              391              850            787
## 144              39              329              918            639
## 145              39              318              937            665
## 146              29              581              792            706
## 147              38              618              505            684
## 148              20              463              870            731
## 149               31              441              718             686
## 150               22              617              693             466
## 151               24              620              661             542
## 152               20              491              577             500
## 153               32              815              515             425
## 154               65              543              566             533
## 155               52              529              493             590
## 156               57              568              653             587
## 157               62              574              704             492
## 158               36              431              494             410
## 159               46              496              426             407
## 160               47              230              547            1512
## 161                0                0                0            1890
## 162              266             2169             2309             522
## 163                0             3645            12758             716
## 164               24              155                0            1506
## 165               25              795              363            1114
## 166               36              495              468             978
## 167               46              391              205            1281
## 168               50              237                0            1092
## 169               41              119                0            1253
## 170              101              743              556             555
## 171               17              405              737             660
## 172               50              465              722             668
## 173               26              475              833             654
## 174               32              652              687             765
## 175                0             1296             1296            1728
## 176               56              911              601             648
## 177               55              531              732             965
## 178               59              351              974             693
## 179               62              467              858             697
## 180               30              545             1003             764
## 181               46              449              702             579
## 182               36              716              814             556
## 183               29              437             1072             775
## 184               53              570             1042             558
## 185               82              668              601             620
## 186               25              416              753             528
## 187               71              639              574             594
## 188               35              799              616             490
## 189               30              674              576             524
## 190               56              592              577             416
## 191              126              738              529             399
## 192               65              795              392             428
## 193               76              680              461             390
## 194               60              601              450             363
## 195               39              558              451             412
## 196               35              501              387             404
## 197               41              602             2367            1211
## 198               75              659              491             631
## 199               75              607              745             453
## 200               35              443              942             836
## 201                5              323              926             785
## 202               10              403             1144             616
## 203               22              668              997             576
## 204               34              545              651             555
## 205               21              284              852             503
## 206               29              579              650             723
## 207               42              488              508             508
## 208               58              846              431             435
## 209               39              566              450             481
## 210               43              610              451             398
## 211               38              549              511             428
## 212               16              403              410             372
## 213              301             1539              770            1122
## 214               55              935              637            1097
## 215               53              209                0            1166
## 216               49              198                0             928
## 217               58              407              845             693
## 218               43              628              981             576
## 219               13              574              932             676
## 220               73              581              364            1103
## 221               22              218                0            1447
## 222               29              124                0            1132
## 223                0              394             1092            1740
## 224              104              787             4224             918
## 225               51              737              890             535
## 226              123             1594             1257            1426
## 227               69              361              845             790
## 228               33              463              652             479
## 229               39              440              760             489
## 230               83              626              565             566
## 231               84              413              697             532
## 232               86              890              638             561
## 233               38              535              732             544
## 234               69              660              594             471
## 235               45              587              540             606
## 236               92              760              420             530
## 237               65              717              326             544
## 238               49              694              446             542
## 239               61              671              414             451
## 240               47              618              500             491
## 241               40              499              434             483
## 242               27              451              420             468
## 243               66              371             1114             794
## 244               18              482              927             597
## 245               53              414             1157             600
## 246                0             1015                0            1279
## 247                0              188                0            1010
## 248               20              342              656             641
## 249               26              398              777             740
## 250               36              489              590             551
## 251               27              543              577             576
## 252               56              636              490             577
## 253               52              590              750             516
## 254               84              710              686             526
## 255               68              673              696             562
## 256               82              886              624             555
## 257                53               708               572               489
## 258                15               789               378               488
## 259                58               587               427               470
## 260                57               516               466               430
## 261                47               542               480               464
## 262                15               447               469               419
## 263                54               539               301               428
## 264                 0              2876             19278               952
## 265                23               744               784               499
## 266                36               731               841               514
## 267                19               382              1313               731
## 268               203              1643               636              1146
## 269               207               292               590               933
## 270                28               674               698               707
## 271                43               609               636               587
## 272                36               826               374               624
## 273                88               929               565               612
## 274                89               830               634               562
## 275                54               563               524               453
## 276                41               420               438               471
## 277                53               398               387               430
## 278                50               328               569               997
## 279                 0               150                 0               847
## 280                 9               247              1167               796
## 281                42               483              1061               618
## 282                82               474               598               623
## 283                44               359               729               573
## 284                43               796               779               678
## 285                 0               168                 0              1522
## 286                39               674              2492              1026
##     TEAM_FIELDING_DP TEAM_CS_YES_NO TEAM_HBP_YES_NO
## 1                 NA              0               0
## 2                 NA              0               0
## 3                 NA              0               0
## 4                 NA              0               0
## 5                 NA              0               0
## 6                 NA              0               0
## 7                 NA              0               0
## 8                 NA              0               0
## 9                 NA              0               0
## 10                NA              0               0
## 11                NA              0               0
## 12                NA              0               0
## 13                NA              0               0
## 14                NA              0               0
## 15                NA              0               0
## 16                NA              0               0
## 17                NA              0               0
## 18                NA              0               0
## 19                NA              1               0
## 20                NA              1               0
## 21                NA              0               0
## 22                NA              0               0
## 23                NA              0               0
## 24                NA               0               0
## 25                NA               0               0
## 26                NA               0               0
## 27                NA               0               0
## 28                NA               0               0
## 29                NA               0               0
## 30                NA               0               0
## 31                NA               0               0
## 32                NA               0               0
## 33                NA               0               0
## 34                NA               0               0
## 35                NA               0               0
## 36                NA               0               0
## 37                NA               0               0
## 38                NA               0               0
## 39                NA               0               0
## 40                NA               1               0
## 41                NA               1               0
## 42                NA               0               0
## 43                NA               0               0
## 44                NA               1               0
## 45                NA               1               0
## 46                NA               0               0
## 47                NA               0               0
## 48                NA               0               0
## 49                NA               0               0
## 50                NA               0               0
## 51                NA               0               0
## 52                NA               0               0
## 53                NA               0               0
## 54                NA               0               0
## 55                NA               0               0
## 56                NA               0               0
## 57                NA               0               0
## 58                NA               0               0
## 59                NA               0               0
## 60                NA               1               0
## 61                NA               0               0
## 62                NA               0               0
## 63                NA               0               0
## 64                NA               0               0
## 65                NA               0               0
## 66                NA               0               0
## 67                NA               0               0
## 68                NA               0               0
## 69                NA               0               0
## 70                NA               0               0
## 71                NA               0               0
## 72                NA               0               0
## 73                NA               0               0
## 74                NA               0               0
## 75                NA               0               0
## 76                NA               0               0
## 77                NA               0               0
## 78                NA              0               0
## 79                NA              0               0
## 80                NA              0               0
## 81                NA              0               0
## 82                NA              0               0
## 83                NA              0               0
## 84                NA              0               0
## 85                NA              0               0
## 86                NA              0               0
## 87                NA              0               0
## 88                NA              0               0
## 89                NA              0               0
## 90                NA              0               0
## 91                NA              0               0
## 92                NA              0               0
## 93                NA              0               0
## 94                NA              0               0
## 95                NA              0               0
## 96                NA              0               0
## 97                NA              0               0
## 98                NA              0               0
## 99                NA              0               0
## 100               NA              0               0
## 101               NA              0               0
## 102               NA              0               0
## 103               NA              0               0
## 104               NA              0               0
## 105               NA              0               0
## 106               NA              0               0
## 107               NA              0               0
## 108               NA              0               0
## 109               NA              0               0
## 110               NA              0               0
## 111               NA              0               0
## 112               NA              0               0
## 113               NA              0               0
## 114               NA              0               0
## 115               NA              0               0
## 116               NA              0               0
## 117               NA              0               0
## 118               NA              1               0
## 119               NA              0               0
## 120               NA              0               0
## 121               NA              0               0
## 122               NA              0               0
## 123               NA              0               0
## 124               NA              0               0
## 125               NA              0               0
## 126               NA              0               0
## 127               NA              0               0
## 128               NA              0               0
## 129               NA              0               0
## 130               NA              0               0
## 131               NA              0               0
## 132               NA              0               0
## 133               NA              0               0
## 134               NA              0               0
## 135               NA              0               0
## 136               NA              0               0
## 137               NA              0               0
## 138               NA              0               0
## 139               NA              0               0
## 140               NA              0               0
## 141               NA              0               0
## 142               NA              0               0
## 143               NA              0               0
## 144               NA              0               0
## 145               NA              0               0
## 146               NA              0               0
## 147               NA              0               0
## 148               NA              0               0
## 149               NA              0               0
## 150               NA              0               0
## 151               NA              0               0
## 152               NA              0               0
## 153               NA              0               0
## 154               NA              0               0
## 155               NA              0               0
## 156               NA              0               0
## 157               NA              0               0
## 158               NA              0               0
## 159               NA              0               0
## 160               NA              1               0
## 161               NA              1               0
## 162               NA              0               0
## 163               NA              0               0
## 164               NA              0               0
## 165               NA              0               0
## 166               NA              1               0
## 167               NA              1               0
## 168               NA              0               0
## 169               NA              0               0
## 170               NA              0               0
## 171               NA              0               0
## 172               NA              0               0
## 173               NA              0               0
## 174               NA              0               0
## 175               NA              1               0
## 176               NA              0               0
## 177               NA              0               0
## 178               NA              0               0
## 179               NA              0               0
## 180               NA              0               0
## 181               NA              0               0
## 182               NA              0               0
## 183               NA              0               0
## 184               NA              0               0
## 185               NA              0               0
## 186               NA              0               0
## 187               NA              0               0
## 188               NA              0               0
## 189               NA              0               0
## 190               NA              0               0
## 191               NA              0               0
## 192               NA              0               0
## 193               NA              0               0
## 194               NA              0               0
## 195               NA              0               0
## 196               NA              0               0
## 197               NA              0               0
## 198               NA              0               0
## 199               NA              0               0
## 200               NA              0               0
## 201               NA              0               0
## 202               NA              0               0
## 203               NA              0               0
## 204               NA              0               0
## 205               NA              0               0
## 206               NA              0               0
## 207               NA              0               0
## 208               NA              0               0
## 209               NA              0               0
## 210               NA              0               0
## 211               NA              0               0
## 212               NA              0               0
## 213               NA              0               0
## 214               NA              1               0
## 215               NA              0               0
## 216               NA              0               0
## 217               NA              0               0
## 218               NA              0               0
## 219               NA              0               0
## 220               NA              1               0
## 221               NA              0               0
## 222               NA              0               0
## 223               NA              1               0
## 224               NA              0               0
## 225               NA              0               0
## 226               NA              0               0
## 227               NA              0               0
## 228               NA              0               0
## 229               NA              0               0
## 230               NA              0               0
## 231               NA              0               0
## 232               NA              0               0
## 233               NA              0               0
## 234               NA              0               0
## 235               NA              0               0
## 236               NA              0               0
## 237               NA              0               0
## 238               NA              0               0
## 239               NA              0               0
## 240               NA              0               0
## 241               NA              0               0
## 242               NA              0               0
## 243               NA              0               0
## 244               NA              0               0
## 245               NA              0               0
## 246               NA              0               0
## 247               NA              0               0
## 248               NA              0               0
## 249               NA              0               0
## 250               NA              0               0
## 251               NA              0               0
## 252               NA              0               0
## 253               NA              0               0
## 254               NA              0               0
## 255               NA              0               0
## 256               NA              0               0
## 257               NA              0               0
## 258               NA              0               0
## 259               NA              0               0
## 260               NA              0               0
## 261               NA              0               0
## 262               NA              0               0
## 263               NA              0               0
## 264               NA              0               0
## 265               NA              0               0
## 266               NA              0               0
## 267               NA              0               0
## 268               NA              0               0
## 269               NA              1               0
## 270               NA              0               0
## 271               NA              0               0
## 272               NA              0               0
## 273               NA              0               0
## 274               NA              0               0
## 275               NA              0               0
## 276               NA              0               0
## 277               NA              0               0
## 278               NA              1               0
## 279               NA              0               0
## 280               NA              0               0
## 281               NA              0               0
## 282               NA              0               0
## 283               NA              0               0
## 284               NA              0               0
## 285               NA              0               0
## 286               NA              0               0
```

```r
baseball_df_na_dp <- baseball_df_fix %>%
  filter(!is.na(TEAM_FIELDING_DP))

summary(lm(baseball_df_na_dp, formula = TEAM_FIELDING_DP~.-TARGET_WINS))
```

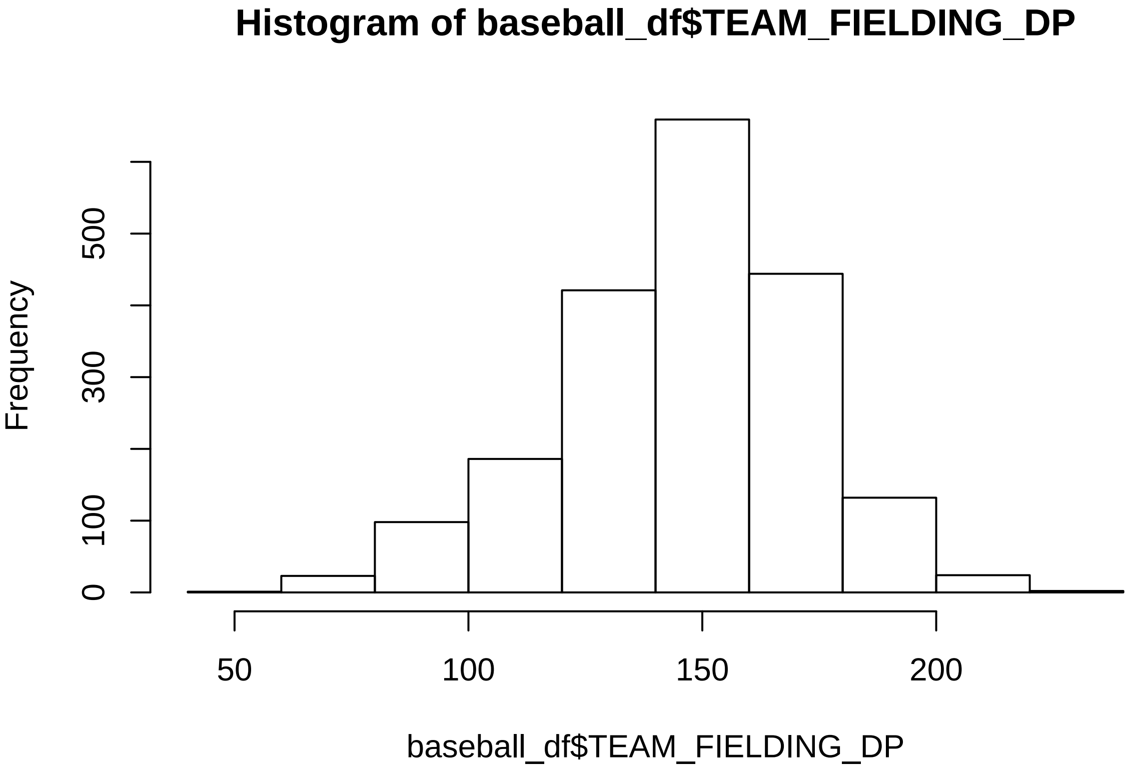
```
##
```

```
## Call:
## lm(formula = TEAM_FIELDING_DP ~ . - TARGET_WINS, data = baseball_df_na_dp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.601 -12.873  -0.814  12.674  61.123
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      127.88475   10.99343  11.633  < 2e-16 ***
## TEAM_BATTING_H     0.07151    0.03110   2.300  0.02158 *
## TEAM_BATTING_2B   -0.02827    0.01766  -1.601  0.10955
## TEAM_BATTING_3B   -0.10305    0.03578  -2.880  0.00402 **
## TEAM_BATTING_HR   -0.12579    0.15525  -0.810  0.41791
## TEAM_BATTING_BB   -0.01834    0.08048  -0.228  0.81978
## TEAM_BATTING_SO   -0.03544    0.04163  -0.851  0.39475
## TEAM_BASERUN_SB   -0.10282    0.01018 -10.097  < 2e-16 ***
## TEAM_PITCHING_H   -0.03362    0.02829  -1.188  0.23484
## TEAM_PITCHING_HR   0.12631    0.14854   0.850  0.39526
## TEAM_PITCHING_BB   0.04148    0.07648   0.542  0.58761
## TEAM_PITCHING_SO   0.01473    0.03959   0.372  0.70989
## TEAM_FIELDING_E   -0.08529    0.01374  -6.208 6.63e-10 ***
## TEAM_CS_YES_NO     9.51649    1.50229   6.335 2.99e-10 ***
## TEAM_HBP_YES_NO   -1.95290    1.83235  -1.066  0.28666
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.21 on 1820 degrees of freedom
##   (155 observations deleted due to missingness)
## Multiple R-squared:  0.2889, Adjusted R-squared:  0.2834
## F-statistic: 52.81 on 14 and 1820 DF,  p-value: < 2.2e-16
```

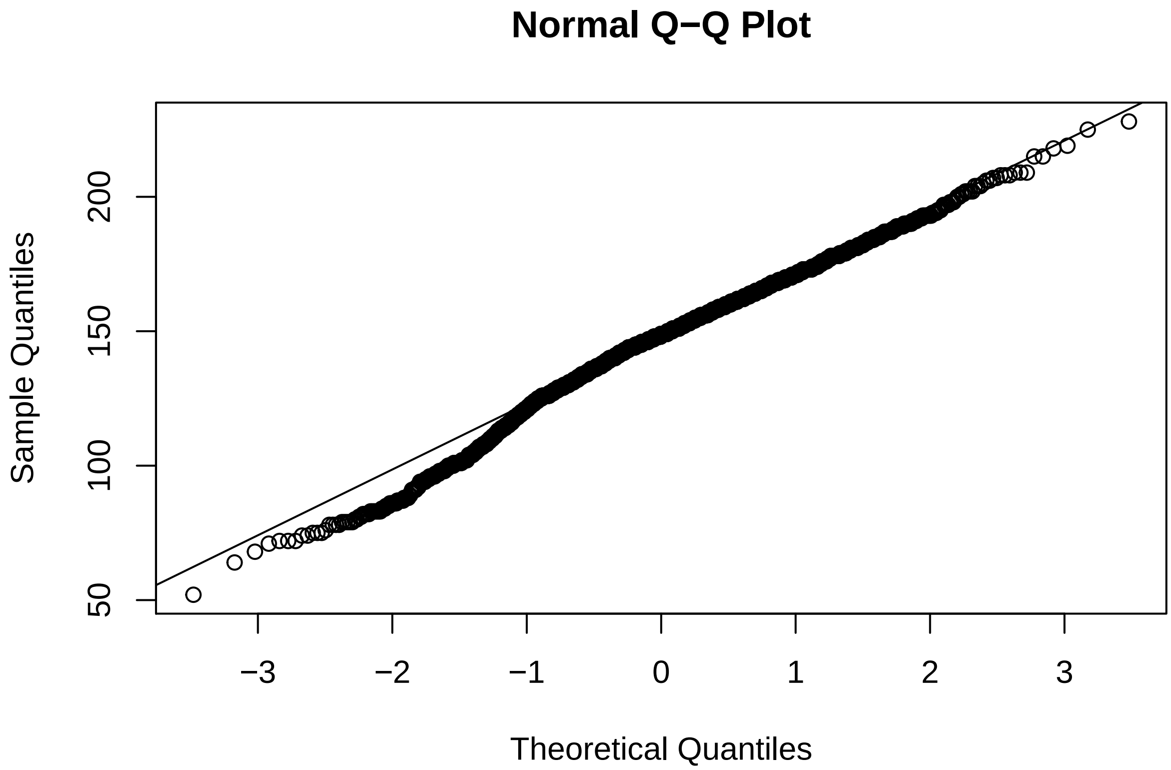
```
corrplot(cor(baseball_df_na_dp, use = 'complete.obs'))
```



```
hist(baseball_df$TEAM_FIELDING_DP)
```



```
qqnorm(baseball_df$TEAM_FIELDING_DP)
qqline(baseball_df$TEAM_FIELDING_DP)
```



Fitting a Linear Model

My first change to the data was to eliminate the index and, replace HBP and BASERUN_CS with dummy variables.

```
baseball_df_fix <- baseball_df %>%
  mutate(TEAM_CS_YES_NO = case_when(!is.na(TEAM_BASERUN_CS) ~ 1, is.na(TEAM_BASERUN_CS) ~ 0)) %>%
  mutate(TEAM_HBP_YES_NO = case_when(!is.na(TEAM_BATTING_HBP) ~ 1, is.na(TEAM_BATTING_HBP) ~ 0)) %>%
  dplyr::select(-c(TEAM_BATTING_HBP, INDEX, TEAM_BASERUN_CS))

baseball_lm <- lm(baseball_df_fix, formula = TARGET_WINS ~.)

summary(baseball_lm)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = baseball_df_fix)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.424  -6.972   0.192   6.983  28.645
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)      57.987326   5.994875   9.673  < 2e-16 ***
## TEAM_BATTING_H   -0.027606   0.016383  -1.685 0.092156 .
## TEAM_BATTING_2B  -0.043602   0.009296  -4.691 2.93e-06 ***
## TEAM_BATTING_3B   0.186256   0.018867   9.872  < 2e-16 ***
## TEAM_BATTING_HR   0.155277   0.081692   1.901 0.057493 .
## TEAM_BATTING_BB   0.102630   0.042342   2.424 0.015456 *
## TEAM_BATTING_SO   0.030619   0.021908   1.398 0.162398
## TEAM_BASERUN_SB   0.068643   0.005505  12.469  < 2e-16 ***
## TEAM_PITCHING_H   0.053979   0.014889   3.625 0.000296 ***
## TEAM_PITCHING_HR -0.052190   0.078164  -0.668 0.504413
## TEAM_PITCHING_BB -0.064794   0.040239  -1.610 0.107522
## TEAM_PITCHING_SO -0.047628   0.020830  -2.286 0.022341 *
## TEAM_FIELDING_E  -0.127819   0.007304 -17.499  < 2e-16 ***
## TEAM_FIELDING_DP -0.104483   0.012332  -8.472  < 2e-16 ***
## TEAM_CS_YES_NO   -3.839845   0.799028  -4.806 1.67e-06 ***
## TEAM_HBP_YES_NO  -2.647365   0.964312  -2.745 0.006104 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 10.1 on 1819 degrees of freedom
##   (441 observations deleted due to missingness)
## Multiple R-squared:  0.4152, Adjusted R-squared:  0.4103
## F-statistic: 86.09 on 15 and 1819 DF,  p-value: < 2.2e-16
```

The initial linear model explains 41% of variation. Next, I’ll add some log transformations of skewed columns: TEAM_PITCHING_BB, TEAM_PITCHING_SO, TEAM_BASERUN_SB, and TEAM_FIELDING_E.

```
baseball_log_lm <- lm(baseball_df_fix, formula = TARGET_WINS ~.+log(TEAM_FIELDING_E) + log(TEAM_PITCHIN

summary(baseball_log_lm)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ . + log(TEAM_FIELDING_E) + log(TEAM_PITCHING_BB) +
##     log(TEAM_PITCHING_SO) + log(TEAM_BASERUN_SB), data = baseball_df_fix)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.639  -6.850   0.083   6.851  29.725
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           209.367276  78.152118   2.679 0.007452 **
## TEAM_BATTING_H         -0.029153   0.016874  -1.728 0.084208 .
## TEAM_BATTING_2B        -0.038719   0.009338  -4.146 3.54e-05 ***
## TEAM_BATTING_3B         0.195250   0.019028  10.261  < 2e-16 ***
## TEAM_BATTING_HR         0.142934   0.082824   1.726 0.084560 .
## TEAM_BATTING_BB         0.140220   0.044797   3.130 0.001775 **
## TEAM_BATTING_SO         0.013917   0.022007   0.632 0.527220
## TEAM_BASERUN_SB         0.084057   0.015811   5.316 1.19e-07 ***
## TEAM_PITCHING_H         0.056607   0.015418   3.671 0.000248 ***
## TEAM_PITCHING_HR       -0.048334   0.079109  -0.611 0.541295
## TEAM_PITCHING_BB       -0.035418   0.041146  -0.861 0.389466
## TEAM_PITCHING_SO       -0.053456   0.021981  -2.432 0.015117 *
## TEAM_FIELDING_E        -0.069984   0.021909  -3.194 0.001426 **
## TEAM_FIELDING_DP       -0.104523   0.012260  -8.525  < 2e-16 ***
## TEAM_CS_YES_NO         -3.568860   0.803825  -4.440 9.54e-06 ***
## TEAM_HBP_YES_NO        -3.061485   1.007413  -3.039 0.002408 **
## log(TEAM_FIELDING_E)  -11.718073   4.037533  -2.902 0.003749 **
## log(TEAM_PITCHING_BB) -36.670662  12.641719  -2.901 0.003767 **
## log(TEAM_PITCHING_SO)  17.416259   6.179754   2.818 0.004881 **
## log(TEAM_BASERUN_SB)   -2.068123   1.483764  -1.394 0.163538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.03 on 1815 degrees of freedom
##   (441 observations deleted due to missingness)
## Multiple R-squared:  0.4248, Adjusted R-squared:  0.4188
## F-statistic: 70.55 on 19 and 1815 DF,  p-value: < 2.2e-16
```

This model explains more variation, but the F-statistic decreased relative to the original model. Next, I'm going to add a few features I'm curious about. TEAM_BATTING_H considers all base hits, including 2B, 3B,and HR. I will create a new variable only looking at singles called TEAM_BATTING_1B. Related to this, I will also incorporate an approximation of an important baseball statistic, slugging. Because some base hits convert to runs at different rates, slugging weighs, singles, doubles, triples and home runs with increasing weight. Usually, slugging also has a denominator of at-bats, which is unavailable in this dataset. Instead, I'll approximate this by dividing by the number of hits. The weights I'm assigning are proportional to the number of bases, so 1 for single, 2 for double… 4 for HR.

```
baseball_df_fix <- baseball_df_fix %>%
  mutate(TEAM_BATTING_1B = TEAM_BATTING_H - TEAM_BATTING_2B - TEAM_BATTING_3B - TEAM_BATTING_HR) %>%
  mutate(TEAM_BATTING_SLG = (TEAM_BATTING_H + TEAM_BATTING_2B + 2 * TEAM_BATTING_3B + 3 *TEAM_BATTING_H


baseball_vars_lm <- lm(baseball_df_fix, formula = TARGET_WINS ~.+log(TEAM_FIELDING_E) + log(TEAM_PITCHI

summary(baseball_vars_lm)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ . + log(TEAM_FIELDING_E) + log(TEAM_PITCHING_BB) +
##     log(TEAM_PITCHING_SO) + log(TEAM_BASERUN_SB), data = baseball_df_fix)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.610  -6.830   0.047   6.788  29.845
##
## Coefficients: (1 not defined because of singularities)
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           178.96333   99.72432   1.795 0.072887 .
## TEAM_BATTING_H         -0.02227    0.02194  -1.015 0.310078
## TEAM_BATTING_2B        -0.05084    0.02640  -1.926 0.054310 .
## TEAM_BATTING_3B         0.17007    0.05471   3.109 0.001909 **
## TEAM_BATTING_HR         0.10812    0.10905   0.991 0.321574
## TEAM_BATTING_BB         0.13896    0.04488   3.096 0.001989 **
## TEAM_BATTING_SO         0.01206    0.02233   0.540 0.589305
## TEAM_BASERUN_SB         0.08407    0.01581   5.316 1.19e-07 ***
## TEAM_PITCHING_H         0.05546    0.01560   3.556 0.000386 ***
## TEAM_PITCHING_HR       -0.05053    0.07925  -0.638 0.523812
## TEAM_PITCHING_BB       -0.03328    0.04138  -0.804 0.421329
## TEAM_PITCHING_SO       -0.05302    0.02200  -2.410 0.016069 *
## TEAM_FIELDING_E        -0.07007    0.02191  -3.197 0.001411 **
## TEAM_FIELDING_DP       -0.10475    0.01227  -8.536  < 2e-16 ***
## TEAM_CS_YES_NO         -3.62119    0.81103  -4.465 8.50e-06 ***
## TEAM_HBP_YES_NO        -3.00343    1.01454  -2.960 0.003112 **
## TEAM_BATTING_1B              NA         NA      NA       NA
## TEAM_BATTING_SLG       18.12471   36.91502   0.491 0.623497
## log(TEAM_FIELDING_E)  -11.66550    4.03980  -2.888 0.003927 **
## log(TEAM_PITCHING_BB) -37.13340   12.67944  -2.929 0.003447 **
## log(TEAM_PITCHING_SO)  18.47861    6.54882   2.822 0.004829 **
## log(TEAM_BASERUN_SB)   -2.03114    1.48598  -1.367 0.171838
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.03 on 1814 degrees of freedom
##   (441 observations deleted due to missingness)
## Multiple R-squared:  0.4249, Adjusted R-squared:  0.4185
## F-statistic: 67.01 on 20 and 1814 DF,  p-value: < 2.2e-16
```

Neither of these features offered additional significance. Finally, I'll use back-selection to eliminate non-contributing variables.

```
baseball_back_lm <- lm(baseball_df_fix, formula = TARGET_WINS ~.-TEAM_BATTING_1B+log(TEAM_FIELDING_E) +

summary(baseball_back_lm)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ . - TEAM_BATTING_1B + log(TEAM_FIELDING_E) +
##     log(TEAM_PITCHING_BB) + log(TEAM_PITCHING_SO) + log(TEAM_BASERUN_SB) -
##     TEAM_BATTING_SLG - TEAM_PITCHING_H - TEAM_BATTING_BB - TEAM_BATTING_SO -
##     TEAM_PITCHING_HR - TEAM_PITCHING_BB - TEAM_FIELDING_E, data = baseball_df_fix)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -31.5371  -6.9106   0.1119   7.0369  28.4639
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -26.850089  38.695974  -0.694 0.487850
## TEAM_BATTING_H           0.029999   0.004279   7.010 3.34e-12 ***
## TEAM_BATTING_2B         -0.040948   0.009282  -4.412 1.09e-05 ***
## TEAM_BATTING_3B          0.199025   0.019098  10.421  < 2e-16 ***
## TEAM_BATTING_HR          0.094197   0.008831  10.666  < 2e-16 ***
## TEAM_BASERUN_SB          0.073484   0.015580   4.717 2.58e-06 ***
## TEAM_PITCHING_SO        -0.039449   0.006805  -5.797 7.94e-09 ***
## TEAM_FIELDING_DP        -0.103899   0.012282  -8.459  < 2e-16 ***
## TEAM_CS_YES_NO          -3.536451   0.786013  -4.499 7.25e-06 ***
## TEAM_HBP_YES_NO         -3.574125   0.935744  -3.820 0.000138 ***
## log(TEAM_FIELDING_E)  -23.337095   1.331275 -17.530  < 2e-16 ***
## log(TEAM_PITCHING_BB)  17.535210   1.507901  11.629  < 2e-16 ***
## log(TEAM_PITCHING_SO)  16.615041   5.730249   2.900 0.003782 **
## log(TEAM_BASERUN_SB)   -1.478015   1.483772  -0.996 0.319324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.1 on 1821 degrees of freedom
##   (441 observations deleted due to missingness)
## Multiple R-squared:  0.4148, Adjusted R-squared:  0.4106
## F-statistic: 99.28 on 13 and 1821 DF,  p-value: < 2.2e-16
```

Before moving to the final model, I want to try creating a simple model with fewer predictors to see how it performs compared to our other models. To starts I chose a few variables that were highly positively and negatively correlated with TARGET_WINS.
* TEAM_BATTING_H * TEAM_BATTING_HR * TEAM_BATTING_BB * TEAM_PITCHING_H * TEAM_PITCHING_HR * TEAM_PITCHING_BB * TEAM_FIELDING_E * TEAM_FIELDING_DP * TEAM_BATTING_SO * TEAM_CS_YES_NO

From there I removed multiple predictors at once. To do this we need to construct a null hypothesis test which states that removing the variables doesn't make a better model. We construct a F-test and compare both versions of the model. If the p-value is under 0.05 we reject the null hypothesis, which indicates our new model isn't different than the first model. If the p-value is greater than 0.05, the model isn't better with those variables, so I will remove them. The simpler the model the better.

To determine which variables I removed, I chose the variable that was not proving to be significant in the linear regression (where the p-value was greater than 0.05). While this doesn't mean the variable itself isn't signficiant, it means the variable alongside the other combination of variables in the model is not significant.

Steps: * Remove TEAM_PITCHING_BB *

- TEAM_PITCHING_BB & TEAM_BATTING_BB
- TEAM_PITCHING_SO & TEAM_BATTING_SO
- TEAM_PITCHING_HR & TEAM_BATTING_HR
- TEAM_PTICHING_H & TEAM_BATTING_H

Steps: * Remove TEAM_PITCHING_BB & TEAM_PITCHING_SO

```
m1 <- lm(TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_HR +TEAM_BATTING_BB + TEAM_BATTING_SO  + TEAM_PITC
summary(m1)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_HR +
##     TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_PITCHING_H + TEAM_PITCHING_HR +
##     TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP +
##     TEAM_CS_YES_NO, data = baseball_df_fix)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.525  -7.848   0.161   7.880  40.171
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      54.830197   5.962944   9.195  < 2e-16 ***
## TEAM_BATTING_H    0.014811   0.004169   3.553 0.000391 ***
## TEAM_BATTING_HR   0.254967   0.057271   4.452 9.01e-06 ***
## TEAM_BATTING_BB   0.052542   0.019298   2.723 0.006534 **
## TEAM_BATTING_SO  -0.027645   0.009317  -2.967 0.003043 **
## TEAM_PITCHING_H   0.018217   0.002332   7.813 9.22e-15 ***
## TEAM_PITCHING_HR -0.194561   0.054263  -3.586 0.000345 ***
## TEAM_PITCHING_BB -0.018421   0.017909  -1.029 0.303814
## TEAM_PITCHING_SO  0.011678   0.008444   1.383 0.166820
## TEAM_FIELDING_E  -0.061318   0.005580 -10.988  < 2e-16 ***
## TEAM_FIELDING_DP -0.149379   0.013160 -11.351  < 2e-16 ***
## TEAM_CS_YES_NO   -3.046049   0.874662  -3.483 0.000508 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 11.38 on 1876 degrees of freedom
##   (388 observations deleted due to missingness)
## Multiple R-squared:  0.3036, Adjusted R-squared:  0.2995
## F-statistic: 74.35 on 11 and 1876 DF,  p-value: < 2.2e-16
```

```
#remove TEAM_PITCHING_BB & TEAM_PITCHING_SO
m2<- lm(TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_HR +TEAM_BATTING_BB + TEAM_BATTING_SO  + TEAM_PITCHI

summary(m2)
```

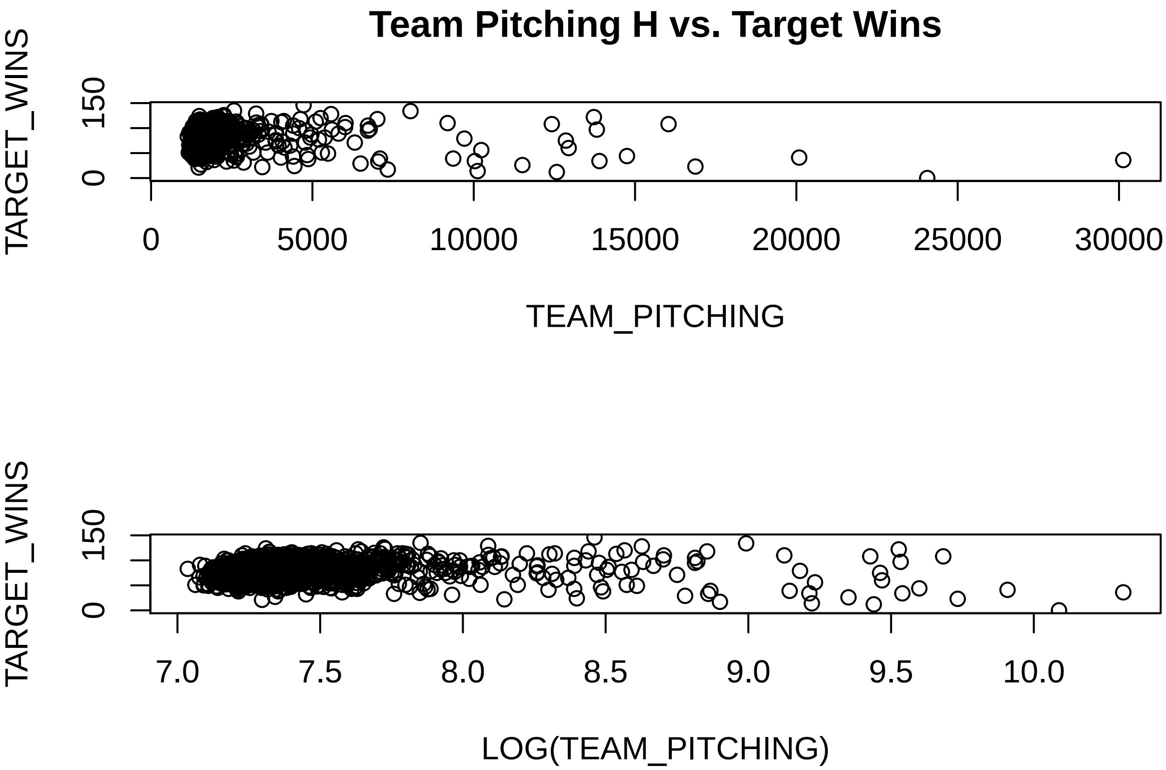
```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_HR +
##     TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_PITCHING_H + TEAM_PITCHING_HR +
##     TEAM_FIELDING_E + TEAM_FIELDING_DP + TEAM_CS_YES_NO, data = baseball_df_fix)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.397  -7.831   0.217   7.924  40.201
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      54.329471   5.937886   9.150  < 2e-16 ***
## TEAM_BATTING_H    0.014275   0.003800   3.756 0.000178 ***
## TEAM_BATTING_HR   0.254520   0.032747   7.772 1.26e-14 ***
## TEAM_BATTING_BB   0.032968   0.003447   9.564  < 2e-16 ***
## TEAM_BATTING_SO  -0.015132   0.002258  -6.701 2.72e-11 ***
## TEAM_PITCHING_H   0.018472   0.001841  10.035  < 2e-16 ***
## TEAM_PITCHING_HR -0.193336   0.031119  -6.213 6.40e-10 ***
## TEAM_FIELDING_E  -0.059177   0.005370 -11.020  < 2e-16 ***
## TEAM_FIELDING_DP -0.147447   0.013090 -11.264  < 2e-16 ***
## TEAM_CS_YES_NO   -2.979270   0.865032  -3.444 0.000586 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 11.38 on 1878 degrees of freedom
##   (388 observations deleted due to missingness)
## Multiple R-squared:  0.3028, Adjusted R-squared:  0.2995
## F-statistic: 90.65 on 9 and 1878 DF,  p-value: < 2.2e-16
```

```
anova(m1, m2)
```

```
## Analysis of Variance Table
##
## Model 1: TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_HR + TEAM_BATTING_BB +
##     TEAM_BATTING_SO + TEAM_PITCHING_H + TEAM_PITCHING_HR + TEAM_PITCHING_BB +
##     TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP + TEAM_CS_YES_NO
## Model 2: TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_HR + TEAM_BATTING_BB +
##     TEAM_BATTING_SO + TEAM_PITCHING_H + TEAM_PITCHING_HR + TEAM_FIELDING_E +
##     TEAM_FIELDING_DP + TEAM_CS_YES_NO
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1876 243143
## 2   1878 243407 -2   -263.79 1.0176 0.3616
```

- Took the log of TEAM_PITCHING_H it’s relationship to TARGET_WINS more linear

```
par(mfrow=c(2,1))
plot(baseball_df_fix$TEAM_PITCHING_H,baseball_df_fix$TARGET_WINS,xlab = ’TEAM_PITCHING’,ylab = ’TARGET_
plot(log(baseball_df_fix$TEAM_PITCHING_H),baseball_df_fix$TARGET_WINS,xlab = ’LOG(TEAM_PITCHING)’,ylab =
```



```
#log TEAM_PITCHING_H
m3 <- lm(TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_HR +TEAM_BATTING_BB + TEAM_BATTING_SO  + log(TEAM_
summary(m3)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_HR +
##     TEAM_BATTING_BB + TEAM_BATTING_SO + log(TEAM_PITCHING_H) +
##     TEAM_PITCHING_HR + TEAM_FIELDING_E + TEAM_FIELDING_DP + TEAM_CS_YES_NO,
##     data = baseball_df_fix)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.877  -7.714   0.053   8.006  37.267
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -3.068e+02  3.376e+01  -9.087  < 2e-16 ***
## TEAM_BATTING_H       -1.708e-03  4.855e-03  -0.352  0.72499
## TEAM_BATTING_HR       3.880e-01  4.108e-02   9.447  < 2e-16 ***
## TEAM_BATTING_BB       3.305e-02  3.440e-03   9.608  < 2e-16 ***
## TEAM_BATTING_SO      -1.409e-02  2.243e-03  -6.281 4.18e-10 ***
## log(TEAM_PITCHING_H)  5.639e+01  5.432e+00  10.382  < 2e-16 ***
## TEAM_PITCHING_HR     -3.254e-01  3.980e-02  -8.176 5.36e-16 ***
## TEAM_FIELDING_E      -6.592e-02  5.775e-03 -11.415  < 2e-16 ***
## TEAM_FIELDING_DP     -1.492e-01  1.307e-02 -11.414  < 2e-16 ***
## TEAM_CS_YES_NO       -2.390e+00  8.585e-01  -2.784  0.00543 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.36 on 1878 degrees of freedom
##   (388 observations deleted due to missingness)
## Multiple R-squared:  0.3053, Adjusted R-squared:  0.302
## F-statistic: 91.72 on 9 and 1878 DF,  p-value: < 2.2e-16
```

- Remove TEAM_BATTING_H

```
#Remove TEAM_BATTING_H
m4 <- lm(TARGET_WINS ~ TEAM_BATTING_HR +TEAM_BATTING_BB + TEAM_BATTING_SO  + log(TEAM_PITCHING_H) + TEA
summary(m4)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_HR + TEAM_BATTING_BB +
##     TEAM_BATTING_SO + log(TEAM_PITCHING_H) + TEAM_PITCHING_HR +
##     TEAM_FIELDING_E + TEAM_FIELDING_DP + TEAM_CS_YES_NO, data = baseball_df_fix)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.874  -7.721   0.062   7.971  37.408
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -2.985e+02  2.407e+01 -12.400  < 2e-16 ***
## TEAM_BATTING_HR       3.786e-01  3.101e-02  12.208  < 2e-16 ***
## TEAM_BATTING_BB       3.294e-02  3.425e-03   9.618  < 2e-16 ***
## TEAM_BATTING_SO      -1.378e-02  2.069e-03  -6.663 3.51e-11 ***
## log(TEAM_PITCHING_H)  5.487e+01  3.280e+00  16.728  < 2e-16 ***
## TEAM_PITCHING_HR     -3.168e-01  3.130e-02 -10.122  < 2e-16 ***
## TEAM_FIELDING_E      -6.466e-02  4.526e-03 -14.287  < 2e-16 ***
## TEAM_FIELDING_DP     -1.491e-01  1.307e-02 -11.413  < 2e-16 ***
## TEAM_CS_YES_NO       -2.379e+00  8.578e-01  -2.773   0.0056 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.36 on 1879 degrees of freedom
##   (388 observations deleted due to missingness)
## Multiple R-squared:  0.3053, Adjusted R-squared:  0.3023
## F-statistic: 103.2 on 8 and 1879 DF,  p-value: < 2.2e-16
```

```
anova(m3, m4)
```

```
## Analysis of Variance Table
##
## Model 1: TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_HR + TEAM_BATTING_BB +
##     TEAM_BATTING_SO + log(TEAM_PITCHING_H) + TEAM_PITCHING_HR +
##     TEAM_FIELDING_E + TEAM_FIELDING_DP + TEAM_CS_YES_NO
## Model 2: TARGET_WINS ~ TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO +
##     log(TEAM_PITCHING_H) + TEAM_PITCHING_HR + TEAM_FIELDING_E +
##     TEAM_FIELDING_DP + TEAM_CS_YES_NO
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1878 242538
## 2   1879 242554 -1   -15.989 0.1238  0.725
```

This leaves a model with an R-squared value of ~30, which means the model accounts for 30% of the variance in the data.

Final Model using all Training Data

For my final model I considered, I originally modeled all of the dummy variables but they ended up not contributing anything to the model. This final model eliminates several features altogether, transforms three, and considers four different interaction effects.

```
baseball_interactions <- lm(baseball_df_fix, formula = TARGET_WINS ~ (TEAM_BATTING_H * TEAM_BATTING_2B

summary(baseball_interactions)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ (TEAM_BATTING_H * TEAM_BATTING_2B +
##     TEAM_BATTING_H * TEAM_BATTING_3B + TEAM_BATTING_H * TEAM_BATTING_HR),
##     data = baseball_df_fix)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.807  -8.816   0.569   9.585  58.270
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -7.961e+00  1.291e+01  -0.616 0.537654
## TEAM_BATTING_H                  4.868e-02  9.039e-03   5.386 7.97e-08 ***
## TEAM_BATTING_2B                -2.802e-02  5.708e-02  -0.491 0.623581
## TEAM_BATTING_3B                 4.678e-01  1.016e-01   4.605 4.34e-06 ***
## TEAM_BATTING_HR                 2.462e-01  6.928e-02   3.553 0.000388 ***
## TEAM_BATTING_H:TEAM_BATTING_2B  2.323e-05  3.814e-05   0.609 0.542590
## TEAM_BATTING_H:TEAM_BATTING_3B -2.231e-04  6.497e-05  -3.434 0.000606 ***
## TEAM_BATTING_H:TEAM_BATTING_HR -1.108e-04  4.622e-05  -2.397 0.016596 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.05 on 2268 degrees of freedom
## Multiple R-squared:  0.2074, Adjusted R-squared:  0.2049
## F-statistic: 84.77 on 7 and 2268 DF,  p-value: < 2.2e-16
```

```
baseball_lm2 <- lm(baseball_df_fix, formula = TARGET_WINS ~.-TEAM_BATTING_1B+log(TEAM_FIELDING_E) + log


summary(baseball_lm2)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ . - TEAM_BATTING_1B + log(TEAM_FIELDING_E) +
##     log(TEAM_PITCHING_BB) + log(TEAM_PITCHING_SO) - TEAM_BATTING_SLG -
##     TEAM_PITCHING_H - TEAM_BATTING_BB - TEAM_BATTING_SO - TEAM_PITCHING_HR -
##     TEAM_PITCHING_BB - TEAM_FIELDING_E + log(TEAM_FIELDING_E) +
##     log(TEAM_PITCHING_SO) + TEAM_BATTING_3B:TEAM_BATTING_HR +
##     TEAM_BATTING_2B:TEAM_BATTING_HR + TEAM_BATTING_H:TEAM_BATTING_HR +
##     TEAM_BATTING_H:TEAM_BATTING_3B - TEAM_BATTING_3B - TEAM_BATTING_SO -
##     TEAM_BATTING_2B - TEAM_BATTING_BB - TEAM_BATTING_HR - TEAM_BATTING_H -
##     TEAM_BATTING_HR - TEAM_PITCHING_HR, data = baseball_df_fix)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -31.0990  -7.0521   0.1861   6.9307  27.4218
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -1.369e+01  3.690e+01  -0.371  0.71058
## TEAM_BASERUN_SB                 5.819e-02  5.596e-03  10.400  < 2e-16 ***
## TEAM_PITCHING_SO               -4.162e-02  6.988e-03  -5.955 3.11e-09 ***
## TEAM_FIELDING_DP               -9.923e-02  1.223e-02  -8.116 8.79e-16 ***
## TEAM_CS_YES_NO                 -3.889e+00  7.742e-01  -5.023 5.59e-07 ***
## TEAM_HBP_YES_NO                -3.183e+00  1.000e+00  -3.182  0.00149 **
## log(TEAM_FIELDING_E)           -2.358e+01  1.313e+00 -17.952  < 2e-16 ***
## log(TEAM_PITCHING_BB)           1.788e+01  1.497e+00  11.946  < 2e-16 ***
## log(TEAM_PITCHING_SO)           1.829e+01  5.822e+00   3.142  0.00171 **
## TEAM_BATTING_3B:TEAM_BATTING_HR -7.815e-04  2.553e-04  -3.061  0.00224 **
## TEAM_BATTING_2B:TEAM_BATTING_HR -3.637e-04  7.112e-05  -5.114 3.48e-07 ***
## TEAM_BATTING_H:TEAM_BATTING_HR   1.526e-04  1.597e-05   9.556  < 2e-16 ***
## TEAM_BATTING_H:TEAM_BATTING_3B   1.923e-04  1.760e-05  10.930  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.07 on 1822 degrees of freedom
##   (441 observations deleted due to missingness)
## Multiple R-squared:  0.4177, Adjusted R-squared:  0.4139
## F-statistic: 108.9 on 12 and 1822 DF,  p-value: < 2.2e-16
```

The R-squared statistic indicates that this model predicts less than half of the variation in wins with the included features. For a next step, I hope to use cross-validation techniques to split the training data further and allow me to compare RMSE of various models.

```
#Tom - still working on this
#set.seed(123)

#baseball_cv <-


#cv_model <- train(TARGET_WINS ~ ., baseball_df_fix, method = ’lm’, trControl = trainControl(method = ’
```

Evaluation Data

I also loaded the evaluation data and predicted the wins using my final model. Since the actual wins are withheld, I compared the distribution of predictions to the actual wins in the training set. The means were similar but the training data included much more variation between teams. It’s also worth mentioning as well that using the predict function creates missing values as the evaluation data is missing. In fact, for TEAM_BATTING_HBP, over 90% of rows are missing entries.

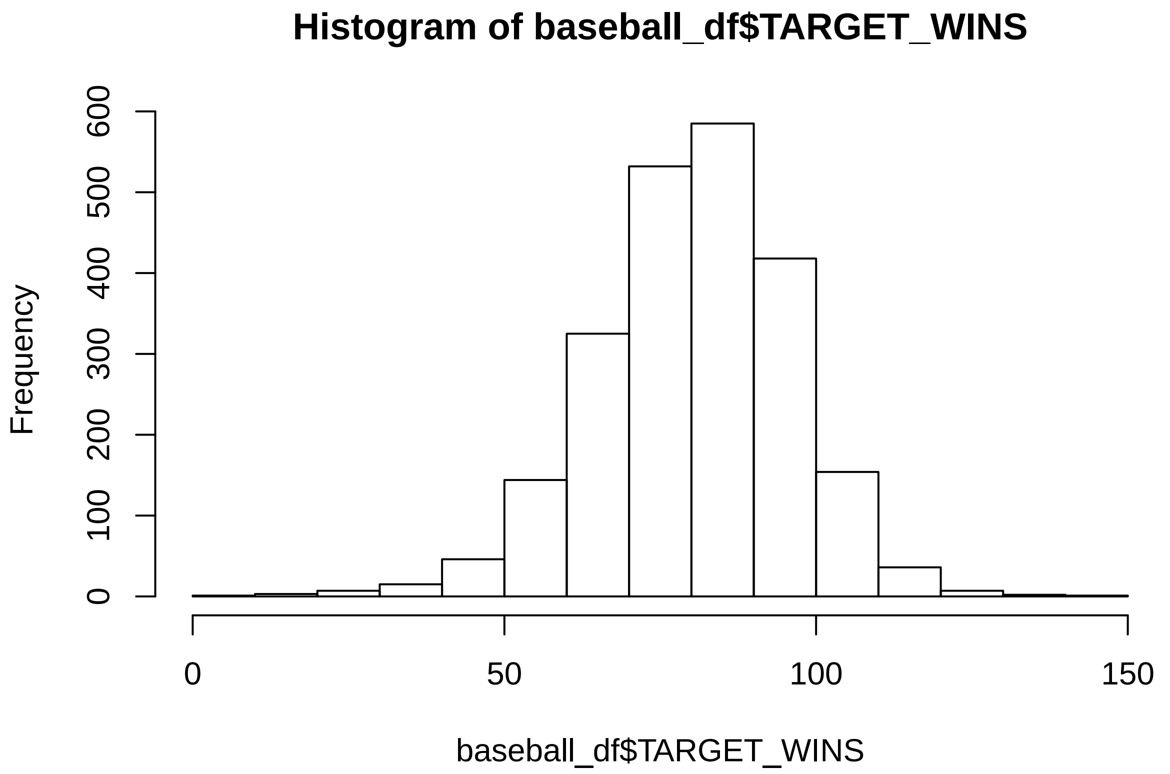
```
round(100*colSums(is.na(baseball_eval))/nrow(baseball_eval),2)
```

```
##            INDEX   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B
##             0.00             0.00             0.00             0.00
##  TEAM_BATTING_HR  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB
##             0.00             0.00             6.95             5.02
##  TEAM_BASERUN_CS TEAM_BATTING_HBP  TEAM_PITCHING_H TEAM_PITCHING_HR
##            33.59            92.66             0.00             0.00
## TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E TEAM_FIELDING_DP
##             0.00             6.95             0.00            11.97
```

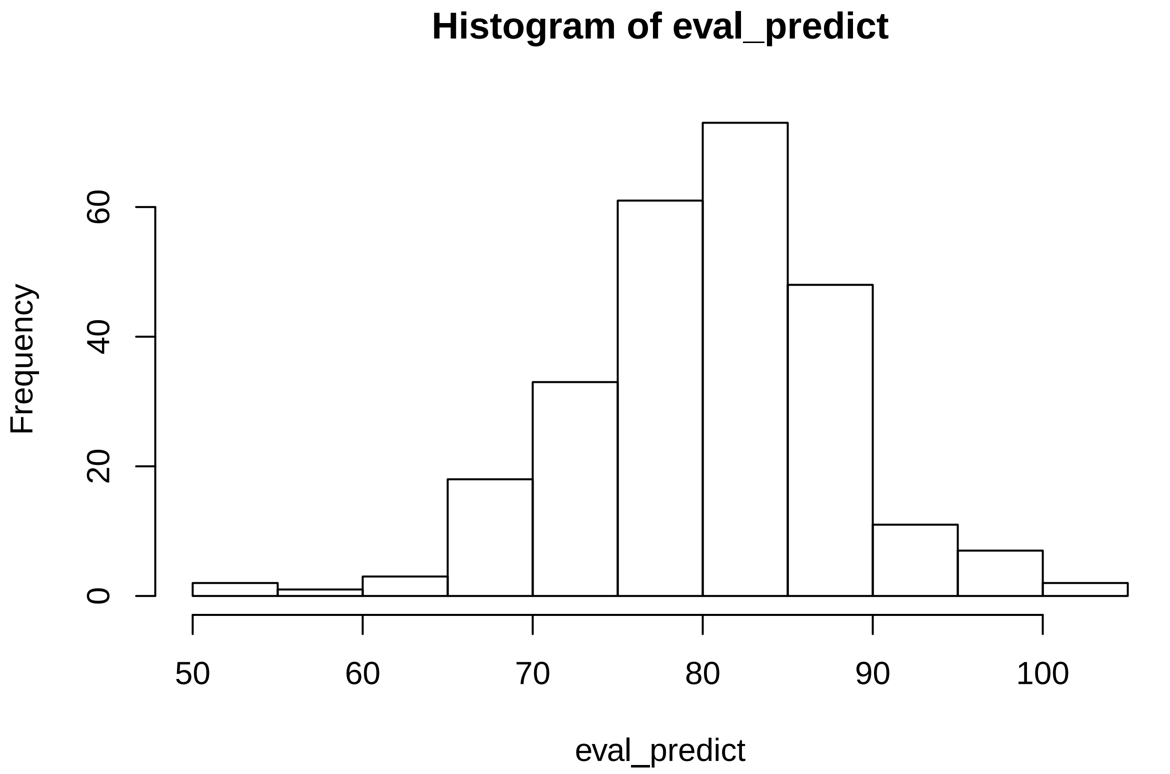
The prediction data also has missing values, which are approximately the same as the training data.

```
baseball_vars <- baseball_eval %>%
  dplyr::select(TEAM_PITCHING_H, TEAM_PITCHING_HR, TEAM_FIELDING_DP, TEAM_BATTING_3B, TEAM_FIELDING_E, T

eval_predict <- predict(baseball_interactions, newdata = baseball_eval)
```

```
hist(baseball_df$TARGET_WINS)
```



```r
hist(eval_predict)
```



```
summary(eval_predict)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   54.33   75.52   80.85   80.47   85.46  102.01
```

```
sd(eval_predict)
```

```
## [1] 7.711588
```

```
summary(baseball_df$TARGET_WINS)
```

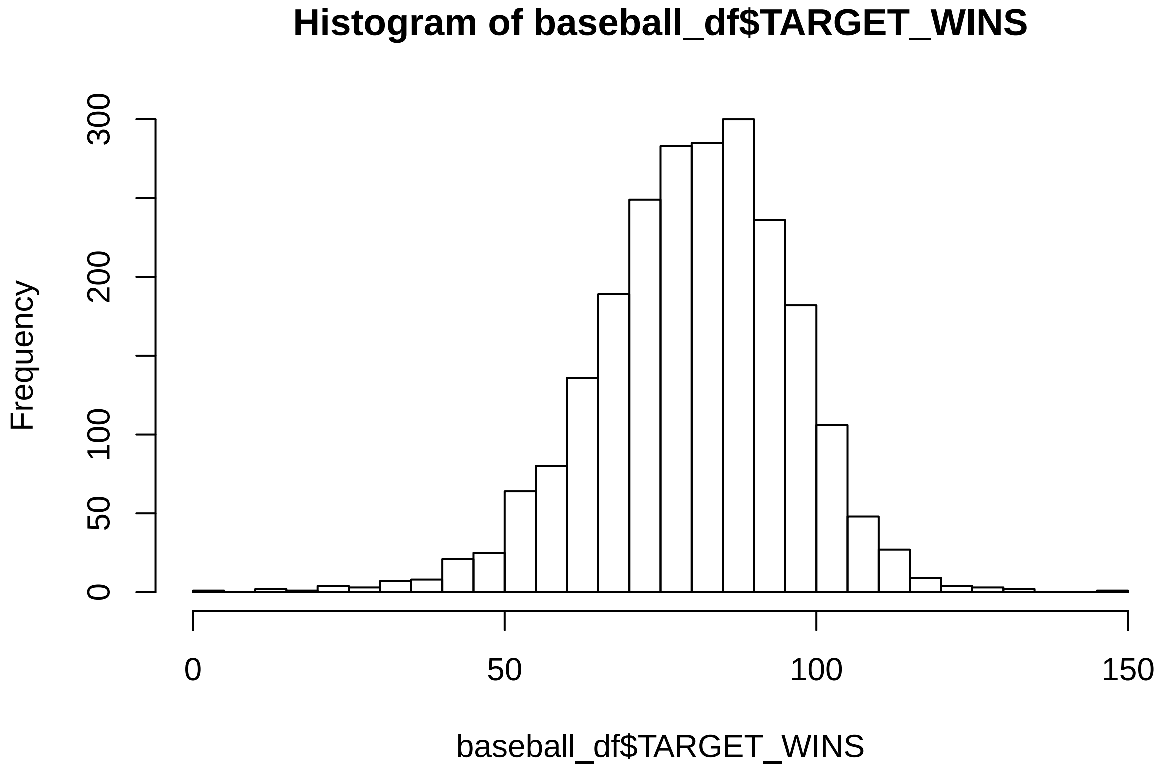
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   71.00   82.00   80.79   92.00  146.00
```

```
sd(baseball_df$TARGET_WINS)
```

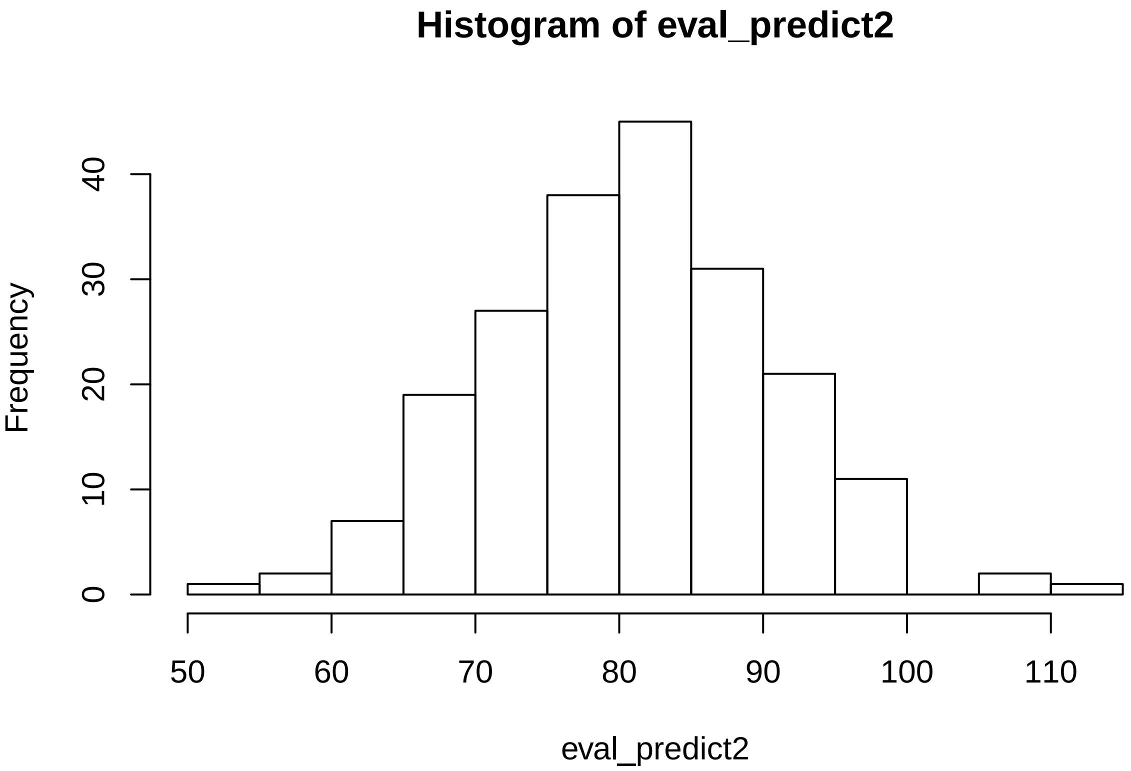
```
## [1] 15.75215
```

```
baseball_eval <- baseball_eval %>% ##added new features to eval data so predict could run
  mutate(TEAM_CS_YES_NO = case_when(!is.na(TEAM_BASERUN_CS) ~ 1, is.na(TEAM_BASERUN_CS) ~ 0)) %>%
  mutate(TEAM_HBP_YES_NO = case_when(!is.na(TEAM_BATTING_HBP) ~ 1, is.na(TEAM_BATTING_HBP) ~ 0))  %>%
  mutate(TEAM_BATTING_1B = TEAM_BATTING_H - TEAM_BATTING_2B - TEAM_BATTING_3B - TEAM_BATTING_HR) %>%
  mutate(TEAM_BATTING_SLG = (TEAM_BATTING_H + TEAM_BATTING_2B + 2 * TEAM_BATTING_3B + 3 *TEAM_BATTING_H
```

```
eval_predict2 <- predict(baseball_lm2, newdata = baseball_eval)

hist(baseball_df$TARGET_WINS, breaks = 40)
```



```
hist(eval_predict2)
```



```
summary(eval_predict2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   54.80   74.15   80.69   80.76   87.47  111.48      54
```

```
sd(eval_predict2, na.rm = T)
```

```
## [1] 9.677568
```

```
n_test <-nrow(baseball_eval)
n_train <- nrow(baseball_df)

summary(baseball_df$TARGET_WINS)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   71.00   82.00   80.79   92.00  146.00
```

```
sd(baseball_df$TARGET_WINS)
```

```
## [1] 15.75215
```