DATA 621 Assignment 1

Kai Lukowiak

2018-02-08

1. Data Exploration:

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

a. Mean / Standard Deviation / Median
b. Bar Chart or Box Plot of the data
c. Is the data correlated to the target variable (or to other variables?)
d. Are any of the variables missing and need to be imputed "fixed"?

Loading the data:

```
library(tidyverse)
library(ggthemes)
library(GGally)
```

```
dfT <- read_csv('moneyball-training-data.csv')
dfT <- dfT %>% select(-INDEX)
glimpse(dfT)
```

```
## Observations: 2,276
## Variables: 16
## $ TARGET_WINS      <int> 39, 70, 86, 70, 82, 75, 80, 85, 86, 76, 78, 6...
## $ TEAM_BATTING_H   <int> 1445, 1339, 1377, 1387, 1297, 1279, 1244, 127...
## $ TEAM_BATTING_2B  <int> 194, 219, 232, 209, 186, 200, 179, 171, 197, ...
## $ TEAM_BATTING_3B  <int> 39, 22, 35, 38, 27, 36, 54, 37, 40, 18, 27, 3...
## $ TEAM_BATTING_HR  <int> 13, 190, 137, 96, 102, 92, 122, 115, 114, 96,...
## $ TEAM_BATTING_BB  <int> 143, 685, 602, 451, 472, 443, 525, 456, 447, ...
## $ TEAM_BATTING_SO  <int> 842, 1075, 917, 922, 920, 973, 1062, 1027, 92...
## $ TEAM_BASERUN_SB  <int> NA, 37, 46, 43, 49, 107, 80, 40, 69, 72, 60, ...
## $ TEAM_BASERUN_CS  <int> NA, 28, 27, 30, 39, 59, 54, 36, 27, 34, 39, 7...
## $ TEAM_BATTING_HBP <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ TEAM_PITCHING_H  <int> 9364, 1347, 1377, 1396, 1297, 1279, 1244, 128...
## $ TEAM_PITCHING_HR <int> 84, 191, 137, 97, 102, 92, 122, 116, 114, 96,...
## $ TEAM_PITCHING_BB <int> 927, 689, 602, 454, 472, 443, 525, 459, 447, ...
## $ TEAM_PITCHING_SO <int> 5456, 1082, 917, 928, 920, 973, 1062, 1033, 9...
## $ TEAM_FIELDING_E  <int> 1011, 193, 175, 164, 138, 123, 136, 112, 127,...
## $ TEAM_FIELDING_DP <int> NA, 155, 153, 156, 168, 149, 186, 136, 169, 1...
```

```
dfE <- read_csv('moneyball-evaluation-data.csv')
dfE <- dfE %>% select(-INDEX)
glimpse(dfE)
```

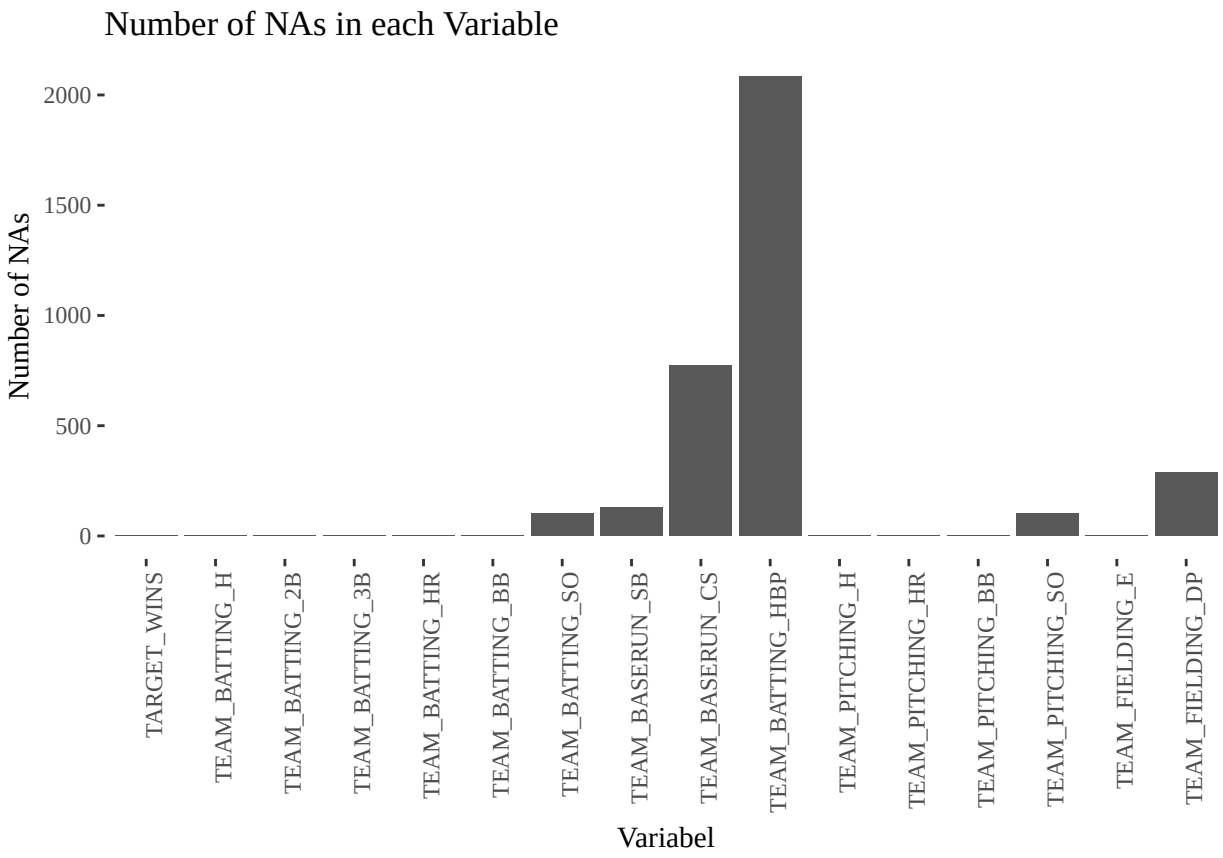
```
## Observations: 259
```

```
## Variables: 15
## $ TEAM_BATTING_H   <int> 1209, 1221, 1395, 1539, 1445, 1431, 1430, 138...
## $ TEAM_BATTING_2B  <int> 170, 151, 183, 309, 203, 236, 219, 158, 177, ...
## $ TEAM_BATTING_3B  <int> 33, 29, 29, 29, 68, 53, 55, 42, 78, 42, 40, 5...
## $ TEAM_BATTING_HR  <int> 83, 88, 93, 159, 5, 10, 37, 33, 23, 58, 50, 1...
## $ TEAM_BATTING_BB  <int> 447, 516, 509, 486, 95, 215, 568, 356, 466, 4...
## $ TEAM_BATTING_SO  <int> 1080, 929, 816, 914, 416, 377, 527, 609, 689,...
## $ TEAM_BASERUN_SB  <int> 62, 54, 59, 148, NA, NA, 365, 185, 150, 52, 6...
## $ TEAM_BASERUN_CS  <int> 50, 39, 47, 57, NA, NA, NA, NA, NA, NA, NA, 2...
## $ TEAM_BATTING_HBP <int> NA, NA, NA, 42, NA, NA, NA, NA, NA, NA, NA, N...
## $ TEAM_PITCHING_H  <int> 1209, 1221, 1395, 1539, 3902, 2793, 1544, 162...
## $ TEAM_PITCHING_HR <int> 83, 88, 93, 159, 14, 20, 40, 39, 25, 62, 53, ...
## $ TEAM_PITCHING_BB <int> 447, 516, 509, 486, 257, 420, 613, 418, 497, ...
## $ TEAM_PITCHING_SO <int> 1080, 929, 816, 914, 1123, 736, 569, 715, 734...
## $ TEAM_FIELDING_E  <int> 140, 135, 156, 124, 616, 572, 490, 328, 226, ...
## $ TEAM_FIELDING_DP <int> 156, 164, 153, 154, 130, 105, NA, 104, 132, 1...
```

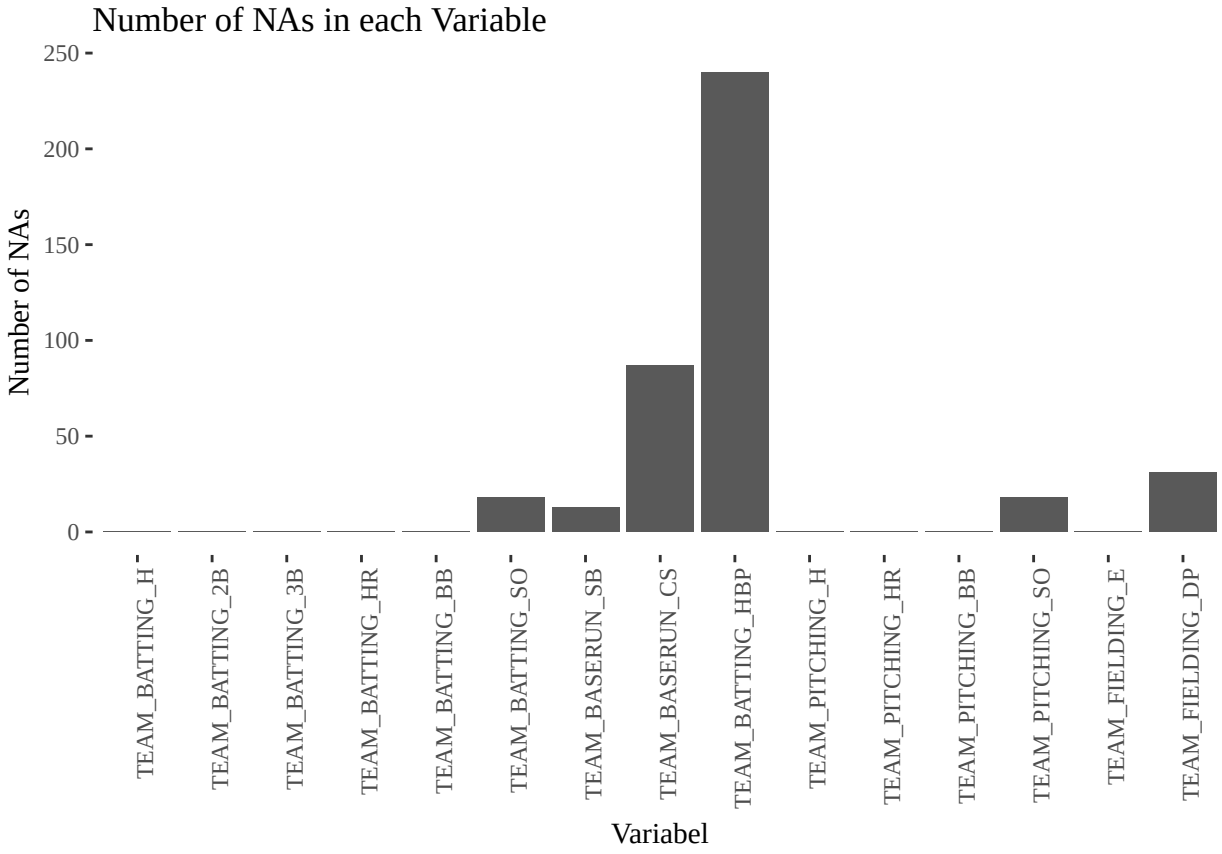
Initial Vizualizaton:

```
naByCol <- function(df){
  x = data.frame(varName = character(),
                 numNA = integer())
  for (i in colnames(df)) {
    y =  sum(is.na(df[,i]))
    newrow = data.frame(varName = i, numNA = y)
    x <- rbind(x, newrow)
  }
  p = ggplot(x, aes(x = varName, y = numNA)) +
    geom_bar(stat = 'identity') +
    xlab("Variabel")+ ylab('Number of NAs')+
    ggtitle("Number of NAs in each Variable") +
    theme_tufte() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))

  return(p)
}
naByCol(dfT)
```


Number of NAs in each Variable
Number of NAs
2000
1500
1000
500
0
TARGET_WINS
TEAM_BATTING_H
TEAM_BATTING_2B
TEAM_BATTING_3B
TEAM_BATTING_HR
TEAM_BATTING_BB
TEAM_BATTING_SO
TEAM_BASERUN_SB
TEAM_BASERUN_CS
TEAM_BATTING_HBP
TEAM_PITCHING_H
TEAM_PITCHING_HR
TEAM_PITCHING_BB
TEAM_PITCHING_SO
TEAM_FIELDING_E
TEAM_FIELDING_DP
Variabel

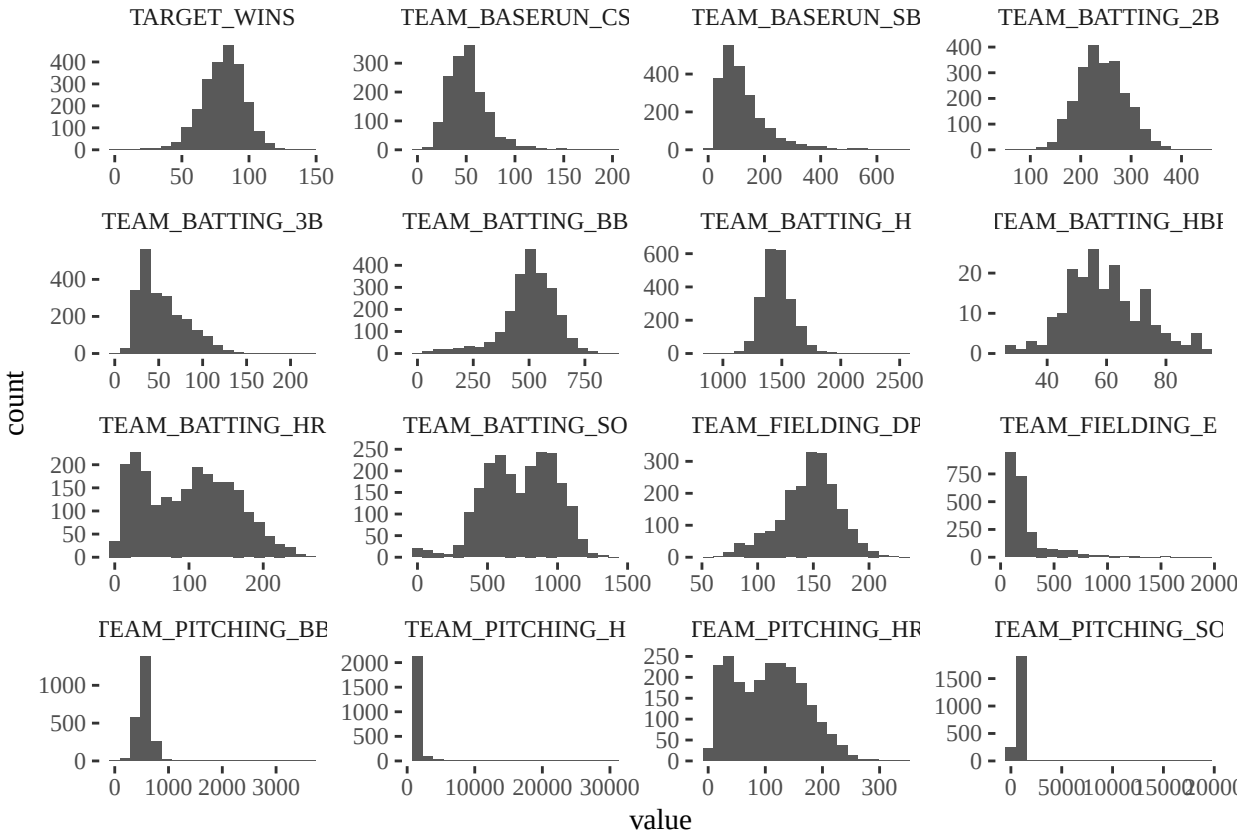
```
naByCol(dfE)
```



Special thanks to this Stack Overflow question

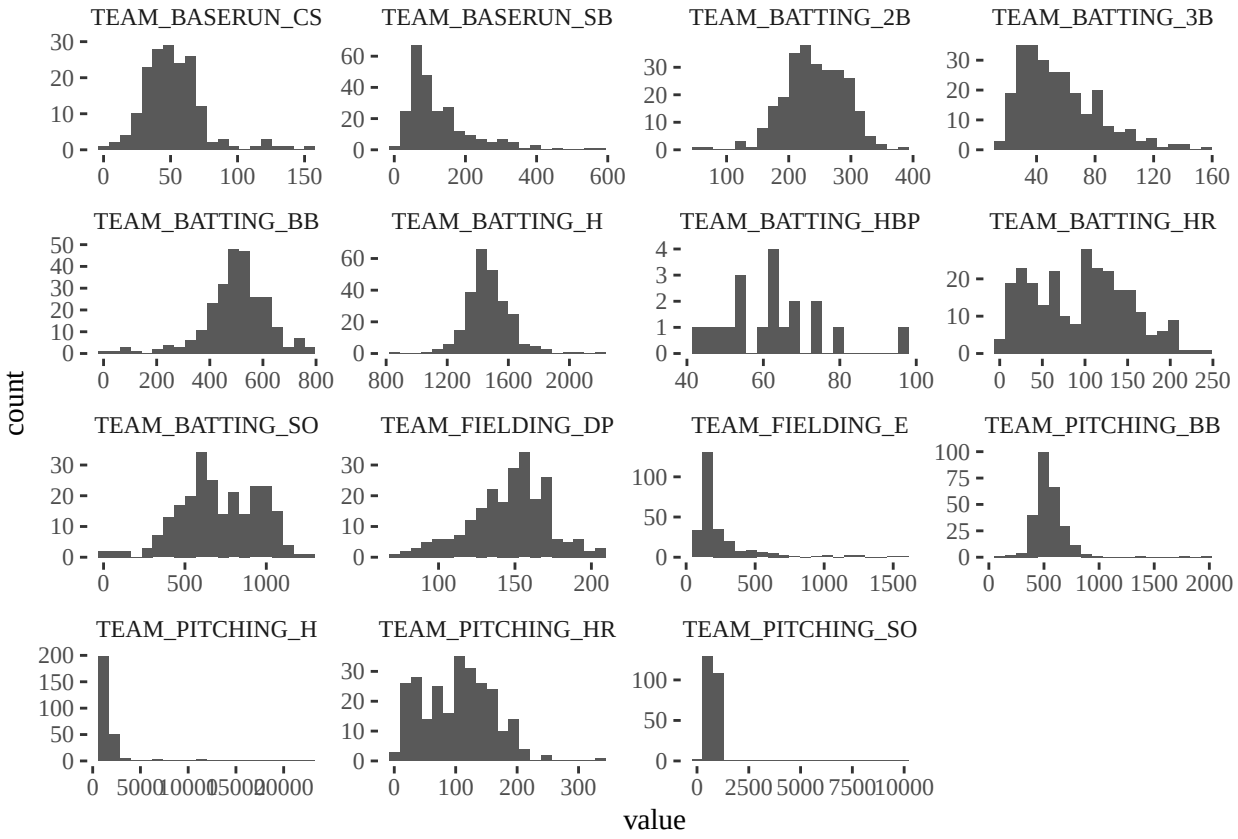
```
ggplot(data = gather(dfT), mapping = aes(x = value)) +
  geom_histogram(bins = 20) + facet_wrap(~key, scales = 'free') +
  theme_tufte()
```

```
## Warning: Removed 3478 rows containing non-finite values (stat_bin).
```




```
ggplot(data = gather(dfE), mapping = aes(x = value)) +
  geom_histogram(bins = 20) + facet_wrap(~key, scales = 'free') +
  theme_tufte()
```

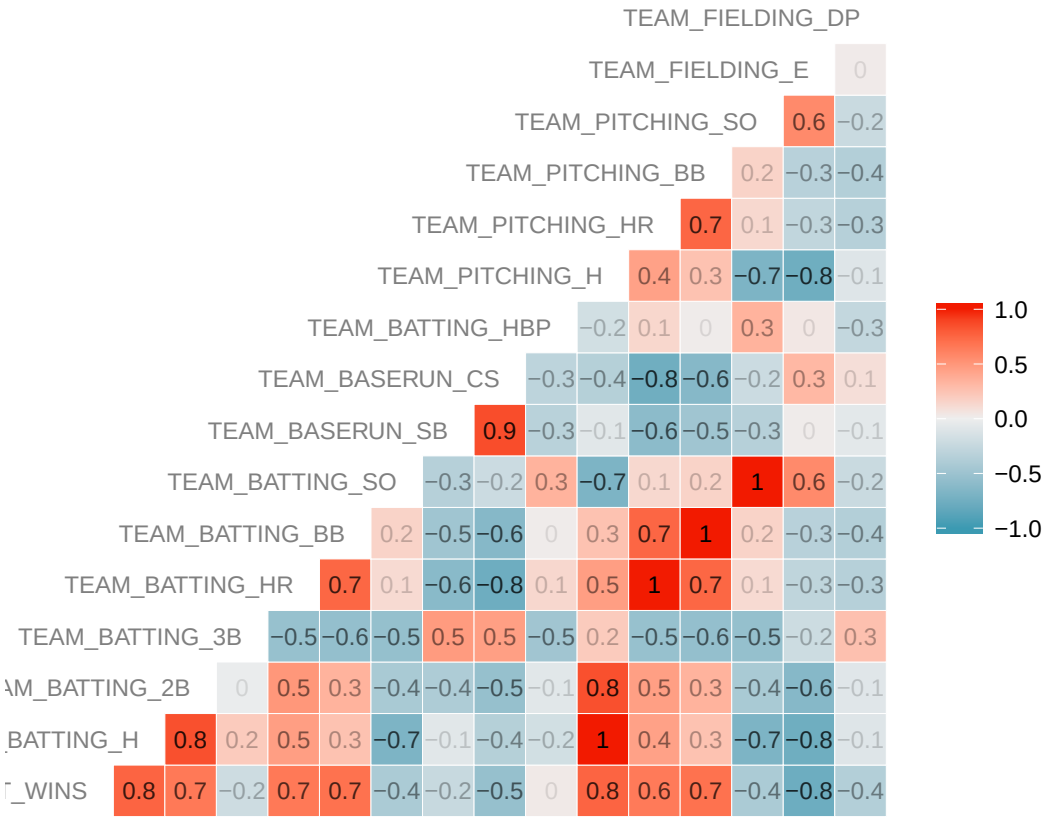
```
## Warning: Removed 407 rows containing non-finite values (stat_bin).
```



From these plots we can see that many variables are aproximetly normally distrubuted. Notable exceptions are `TEAM_BATTING_3B`, `TEAM_BATTING_HR`, `TEAM_PITCHING_H`.

To explore these further: (Also, special thanks to this site)

```
corr <- round(cor(dfT, use = "complete.obs"), 1) # Complete obs b/c of all NAs
ggcorr(corr, hjust = 1, size = 3, color = "grey50",
       layout.exp = 1, label = TRUE, label_size = 3,
       label_alpha = TRUE)
```



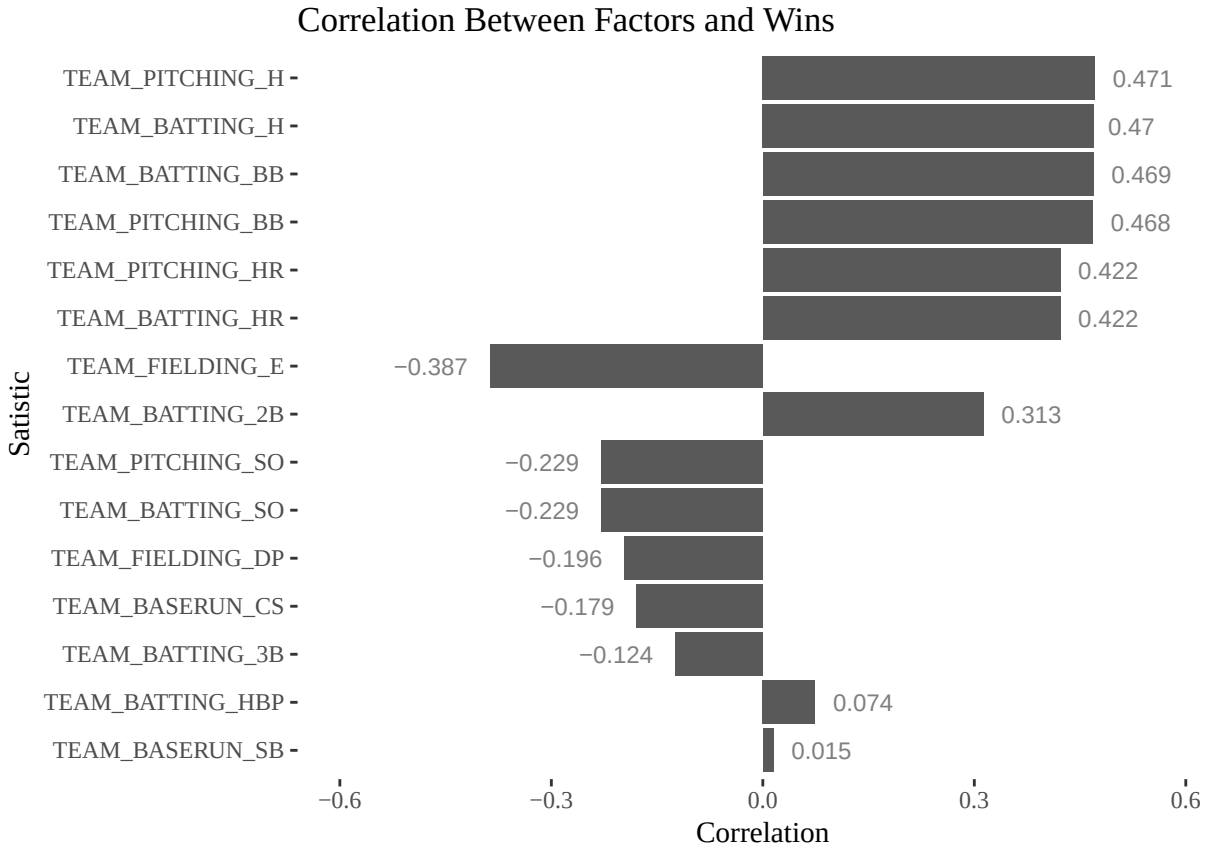

```
summary(dfT)
```

```
##   TARGET_WINS      TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
##  Min.   :  0.00   Min.   : 891   Min.   : 69.0   Min.   :  0.00
##  1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0   1st Qu.: 34.00
##  Median : 82.00   Median :1454   Median :238.0   Median : 47.00
##  Mean   : 80.79   Mean   :1469   Mean   :241.2   Mean   : 55.25
##  3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0   3rd Qu.: 72.00
##  Max.   :146.00   Max.   :2554   Max.   :458.0   Max.   :223.00
##
##  TEAM_BATTING_HR  TEAM_BATTING_BB TEAM_BATTING_SO  TEAM_BASERUN_SB
##  Min.   :  0.00   Min.   :  0.0   Min.   :   0.0   Min.   :  0.0
##  1st Qu.: 42.00   1st Qu.:451.0   1st Qu.: 548.0   1st Qu.: 66.0
##  Median :102.00   Median :512.0   Median : 750.0   Median :101.0
##  Mean   : 99.61   Mean   :501.6   Mean   : 735.6   Mean   :124.8
##  3rd Qu.:147.00   3rd Qu.:580.0   3rd Qu.: 930.0   3rd Qu.:156.0
##  Max.   :264.00   Max.   :878.0   Max.   :1399.0   Max.   :697.0
##                                   NA's   :102      NA's   :131
##  TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR
##  Min.   :  0.0   Min.   :29.00    Min.   : 1137   Min.   :  0.0
##  1st Qu.: 38.0   1st Qu.:50.50    1st Qu.: 1419   1st Qu.: 50.0
##  Median : 49.0   Median :58.00    Median : 1518   Median :107.0
##  Mean   : 52.8   Mean   :59.36    Mean   : 1779   Mean   :105.7
##  3rd Qu.: 62.0   3rd Qu.:67.00    3rd Qu.: 1682   3rd Qu.:150.0
##  Max.   :201.0   Max.   :95.00    Max.   :30132   Max.   :343.0
##  NA's   :772     NA's   :2085
##  TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E  TEAM_FIELDING_DP
##  Min.   :   0.0   Min.   :    0.0   Min.   :  65.0   Min.   : 52.0
```

```
##  1st Qu.: 476.0   1st Qu.:  615.0   1st Qu.: 127.0   1st Qu.:131.0
##  Median : 536.5   Median :  813.5   Median : 159.0   Median :149.0
##  Mean   : 553.0   Mean   :  817.7   Mean   : 246.5   Mean   :146.4
##  3rd Qu.: 611.0   3rd Qu.:  968.0   3rd Qu.: 249.2   3rd Qu.:164.0
##  Max.   :3645.0   Max.   :19278.0   Max.   :1898.0   Max.   :228.0
##                   NA's   :102                        NA's   :286
```

```
corr2 <- round(cor(dfT[,-1], dfT$TARGET_WINS, use = "complete.obs"), 3) # For target:
corr2 <- as.data.frame(corr2) %>% rownames_to_column(var = "Row_name" )%>% as_tibble()
corr2 <- corr2 %>% rename(Correlation = V1)
```

```
corrPlotFunc <- function(corr2){
ggplot(data = corr2,
       aes(x = reorder(Row_name, abs(Correlation)),
           y = Correlation))+
  geom_bar(stat = 'identity') +
  geom_text(aes(label=Correlation),
            hjust = ifelse(corr2$Correlation >= 0, -0.3, 1.3),
            size = 3, color = 'grey50') +
  coord_flip()+
  ylim(-.6, .6)+
  xlab("Satistic")+
  ggtitle("Correlation Between Factors and Wins")+
  theme_tufte()
}
corrPlotFunc(corr2)
```



We also should check if there is correlation between the rows that had tons of NA values:

```
dfMissing <- dfT %>% mutate(isMissing = ifelse(is.na(.$TEAM_BATTING_HBP), 1, 0))
corCoef <- cor(dfMissing$TARGET_WINS, dfMissing$isMissing)
corCoef
```

```
## [1] -0.002610647
```

We can see that the element with the highest missing values is not very correlated

```
tVal <- corCoef * sqrt((nrow(dfMissing - 2)) / (1 - corCoef^2))
tVal
```

```
## [1] -0.1245477
```

We see that the corelation is not significant and futher, even if it was significant, the effect is so small it might be worth deleting the column instead.

2. DATA PREPARATION

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations. a. Fix missing values (maybe with a Mean or Median value) b. Create flags to suggest if a variable was missing c. Transform data by putting it into buckets d. Mathematical transforms such as log or square root (or use Box-Cox) e. Combine variables (such as ratios or adding or multiplying) to create new variables

3. BUILD MODELS

Using the training data set, build at least three different multiple linear regression models, using different variables (or the same variables with different transformations). Since we have not yet covered automated variable selection methods, you should select the variables manually (unless you previously learned Forward or Stepwise selection, etc.). Since you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done. Discuss the coefficients in the models, do they make sense? For example, if a team hits a lot of Home Runs, it would be reasonably expected that such a team would win more games. However, if the coefficient is negative (suggesting that the team would lose more games), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

4. SELECT MODELS

Decide on the criteria for selecting the best multiple linear regression model. Will you select a model with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model. For the multiple linear regression model, will you use a metric such as Adjusted R2, RMSE, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. Using the training data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R2, (c) F-statistic, and (d) residual plots. Make predictions using the evaluation data set