DATA621 Homework #1 - Appendix A

William Aiken, Donald Butler, Michael Ippolito, Bharani Nittala, Leticia Salazar, Santiago Torres

9/25/2022

1. DATA EXPLORATION

Due to the number of fields in this data, I broke the dataset into intuitive sections and explored each section individually.

```
# Load data
eval_df <- read.csv("https://raw.githubusercontent.com/catfoodlover/DATA621/main/HW1/moneyball-evaluation-data.csv")
train_df <- read.csv("https://raw.githubusercontent.com/catfoodlover/DATA621/main/HW1/moneyball-training-data.csv")
```

Base Hits by Batter

- 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)

The means and medians are very similar for the base hits variables implying little skew to the distributions.

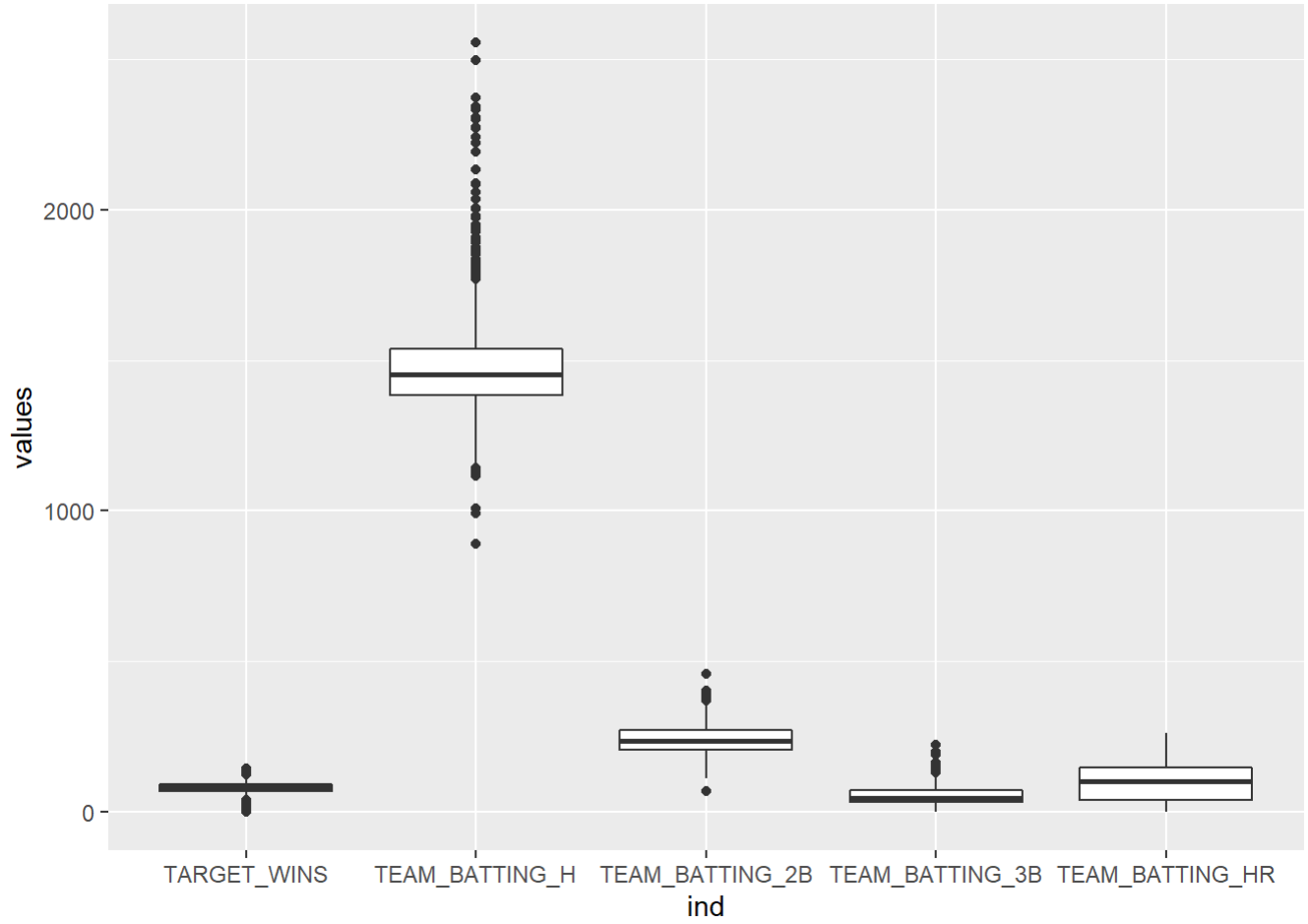
```
# Summary
train_df %>% select(c("TARGET_WINS", "TEAM_BATTING_H","TEAM_BATTING_2B","TEAM_BATTING_3B","TEAM_BATTING_HR")) %>% gtsummary::tbl_summary(statistic =list(c("TARGET_WINS", "TEAM_BATTING_H","TEAM_BATTING_2B","TEAM_BATTING_3B","TEAM_BATTING_HR") ~ "{mean} {median} {sd}"
))
```

Characteristic	**N = 2,276**[1]
TARGET_WINS	81 82 16
TEAM_BATTING_H	1,469 1,454 145
TEAM_BATTING_2B	241 238 47
TEAM_BATTING_3B	55 47 28
TEAM_BATTING_HR	100 102 61

[1] Mean Median SD

We see tight distributions except for all base hits by batters (TEAM_BATTING_H).

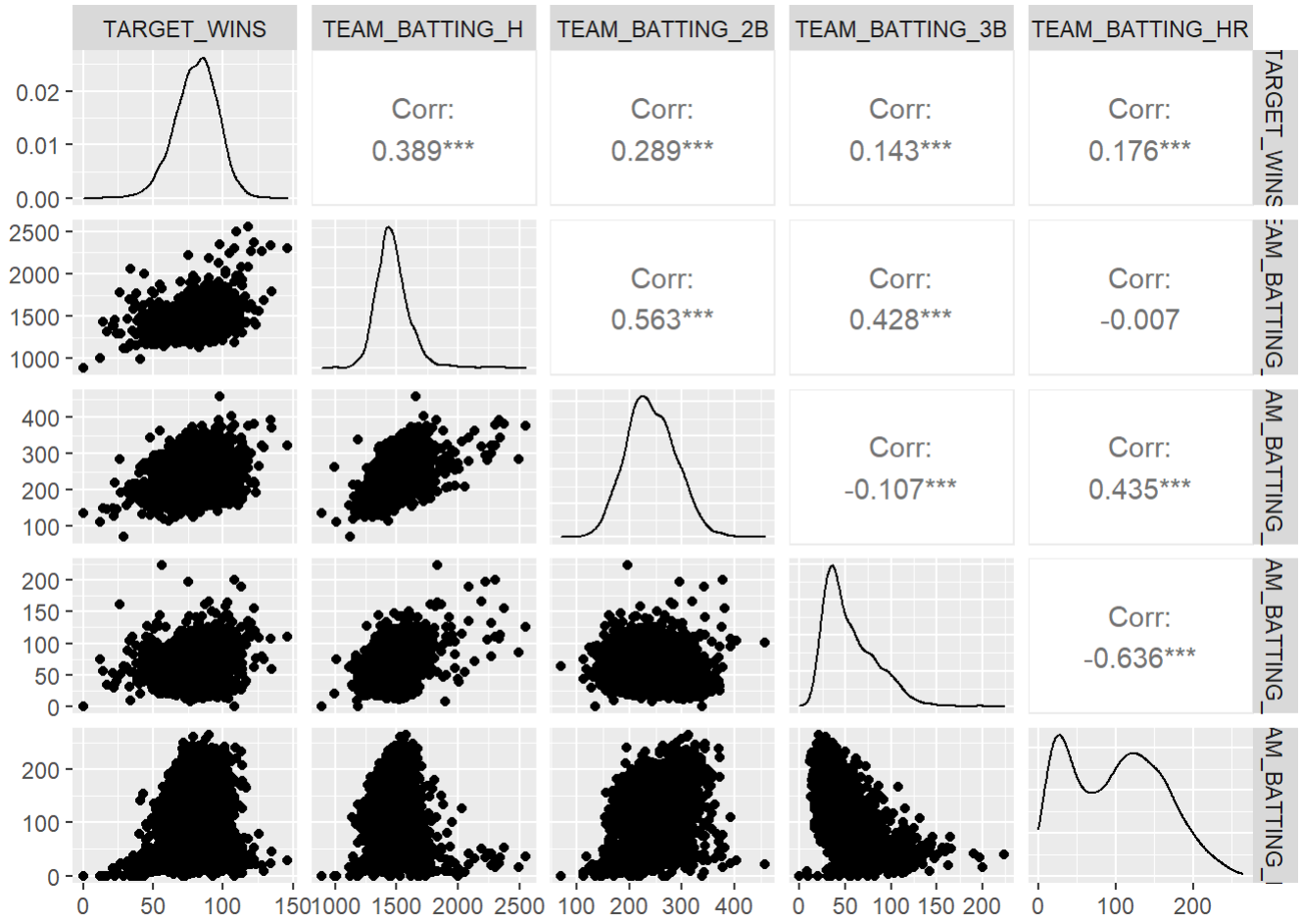
```
# Box plots
temp <- train_df %>% select(c("TARGET_WINS", "TEAM_BATTING_H","TEAM_BATTING_2B","TEAM_BATTING_3
B","TEAM_BATTING_HR"))
ggplot2::ggplot(stack(temp), aes(x = ind, y = values)) +
  geom_boxplot()
```



Unsurprisingly, all possible base hits (TEAM_BATTING_H) is correlated with winning. As you increase the number of bases achieved by an at bat, the correlation decreases.

Interestingly, doubles and triples are correlated with base hits while home runs are not.

```
# Correlation plot
train_df %>% select(c("TARGET_WINS", "TEAM_BATTING_H","TEAM_BATTING_2B","TEAM_BATTING_3B","TEAM_
BATTING_HR")) %>% GGally::ggpairs()
```



Batting

- TARGET_WINS - Number of wins
- 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

The measures of central tendency show us that most of these variable have slight skew to their distributions. Stolen bases has a large right skew to its distribution.

We are missing values for stikeouts, stolen bases and caught stealing.

```
# Summary
train_df %>% select(c("TEAM_BATTING_BB", "TEAM_BATTING_SO", "TEAM_BASERUN_SB", "TEAM_BASERUN_CS"
)) %>% gtsummary::tbl_summary(statistic =list(c("TEAM_BATTING_BB", "TEAM_BATTING_SO", "TEAM_BASE
RUN_SB", "TEAM_BASERUN_CS") ~ "{mean} {median} {sd}"
))
```

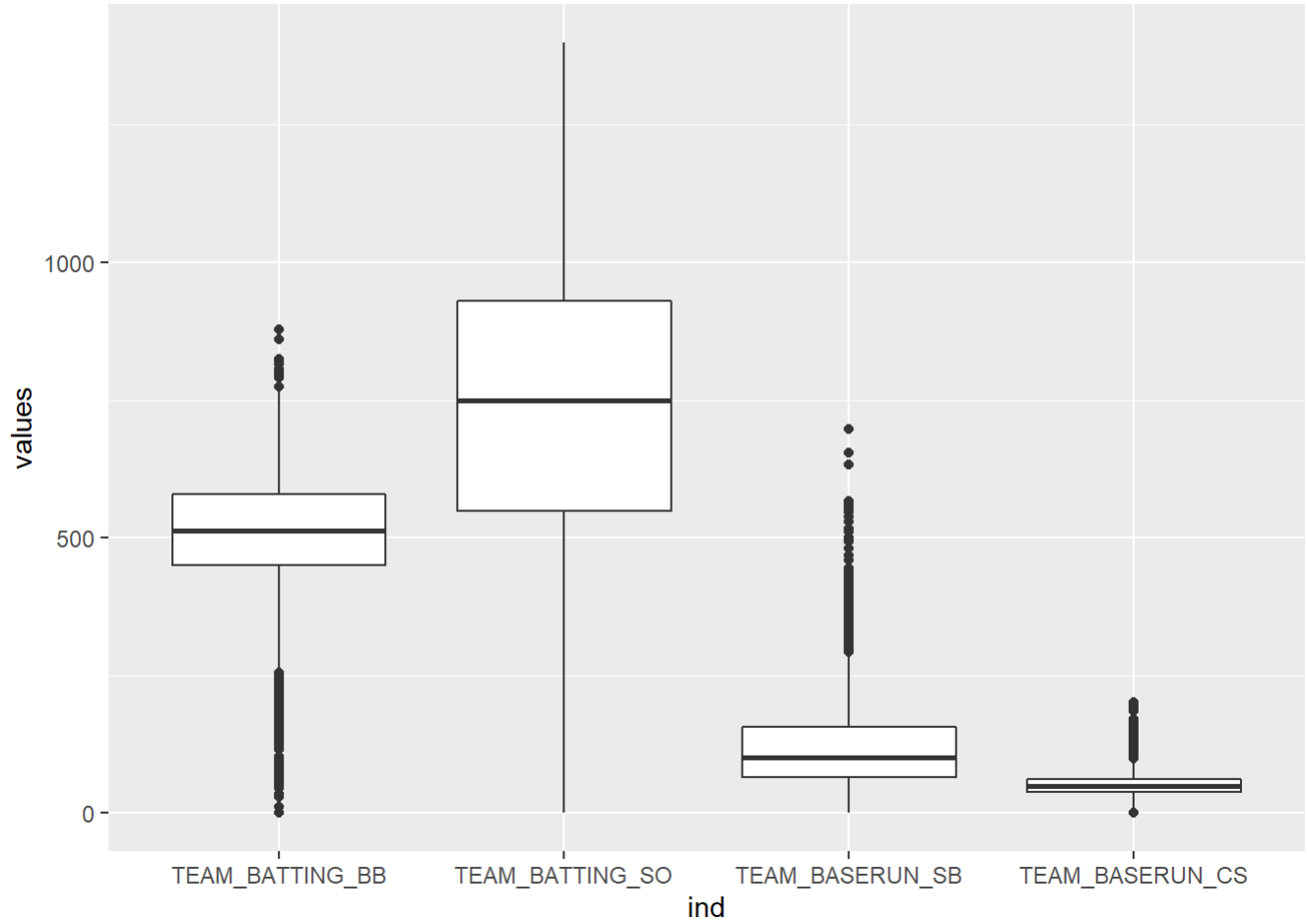
Characteristic	**N = 2,276**[1]
TEAM_BATTING_BB	502 512 123

[1] Mean Median SD

Characteristic	N = 2,276[1]
TEAM_BATTING_SO	736 750 249
Unknown	102
TEAM_BASERUN_SB	125 101 88
Unknown	131
TEAM_BASERUN_CS	53 49 23
Unknown	772

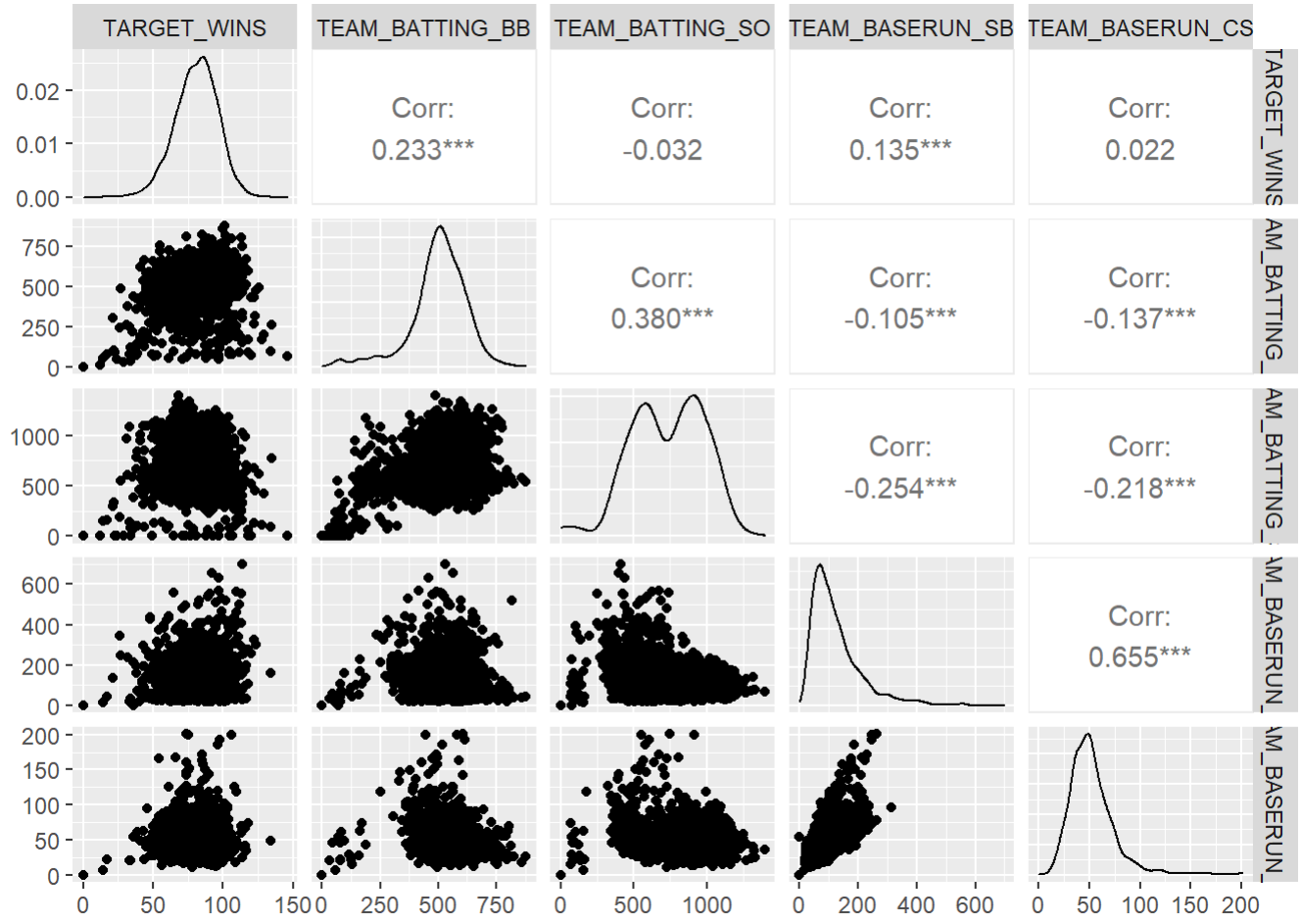
[1] Mean Median SD

```
# Box plots
temp <- train_df %>% select(c("TEAM_BATTING_BB", "TEAM_BATTING_SO", "TEAM_BASERUN_SB", "TEAM_BAS
ERUN_CS"))
ggplot2::ggplot(stack(temp), aes(x = ind, y = values)) +
  geom_boxplot()
```



Of all the batting variables, only walks by batter has a correlation to wins.

```
# Correlation plot
train_df %>% select(c("TARGET_WINS", "TEAM_BATTING_BB", "TEAM_BATTING_SO", "TEAM_BASERUN_SB", "T
EAM_BASERUN_CS")) %>% GGally::ggpairs()
```



Fielding

- TARGET_WINS - Number of wins
- TEAM_FIELDING_E - Errors
- TEAM_FIELDING_DP - Double Plays

The Errors variable(TEAM_FIELDING_E) has an incredibly right skewed distribution. We are missing some Double Plays values.

```
# Summary
train_df %>% select(c("TEAM_FIELDING_E", "TEAM_FIELDING_DP")) %>% gtsummary::tbl_summary(statist
ic =list(c("TEAM_FIELDING_E", "TEAM_FIELDING_DP") ~ "{mean} {median} {sd}"
))
```

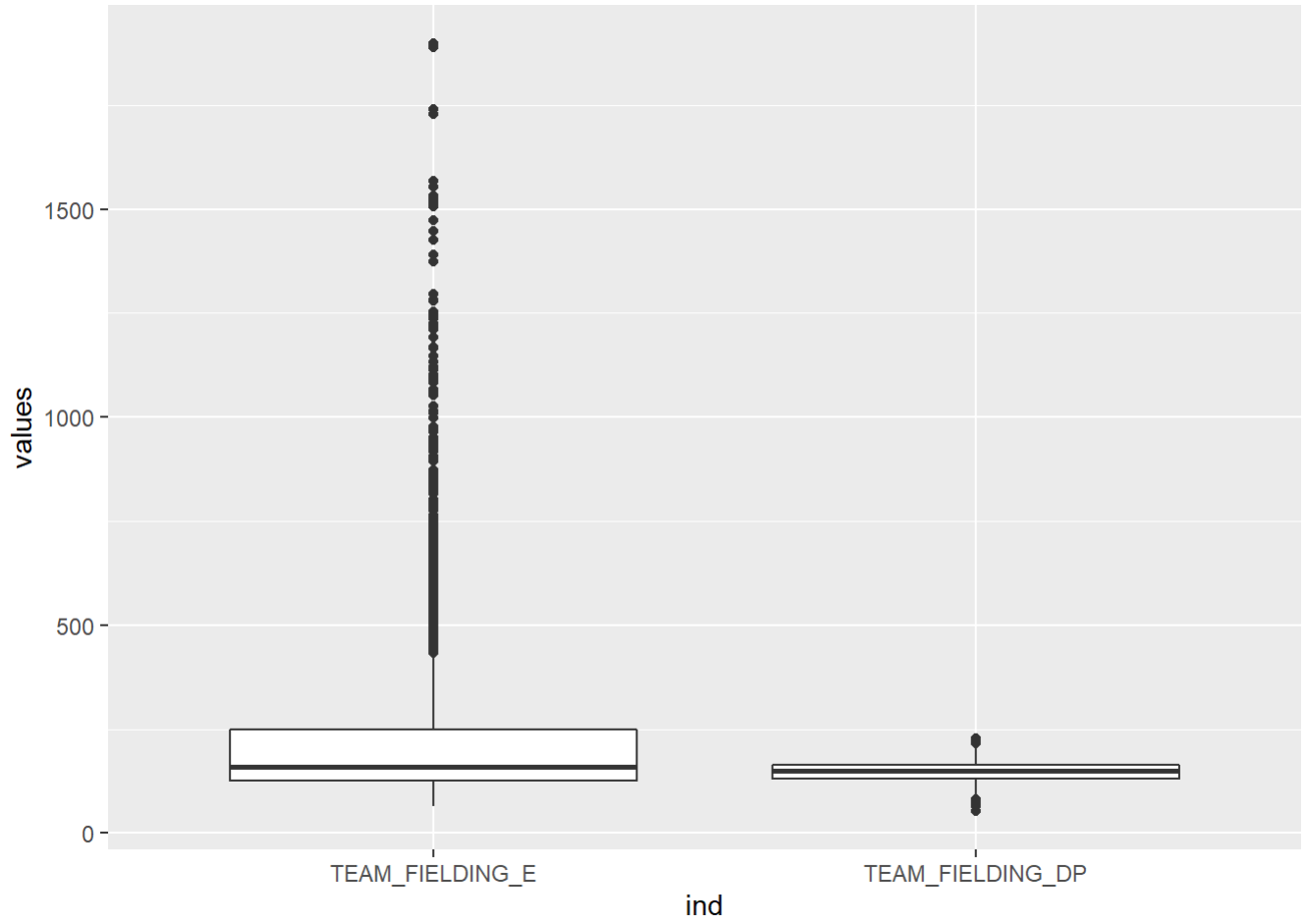
Characteristic	**N = 2,276**[1]
TEAM_FIELDING_E	246 159 228

[1] Mean Median SD

Characteristic	N = 2,276[1]
TEAM_FIELDING_DP	146 149 26
Unknown	286

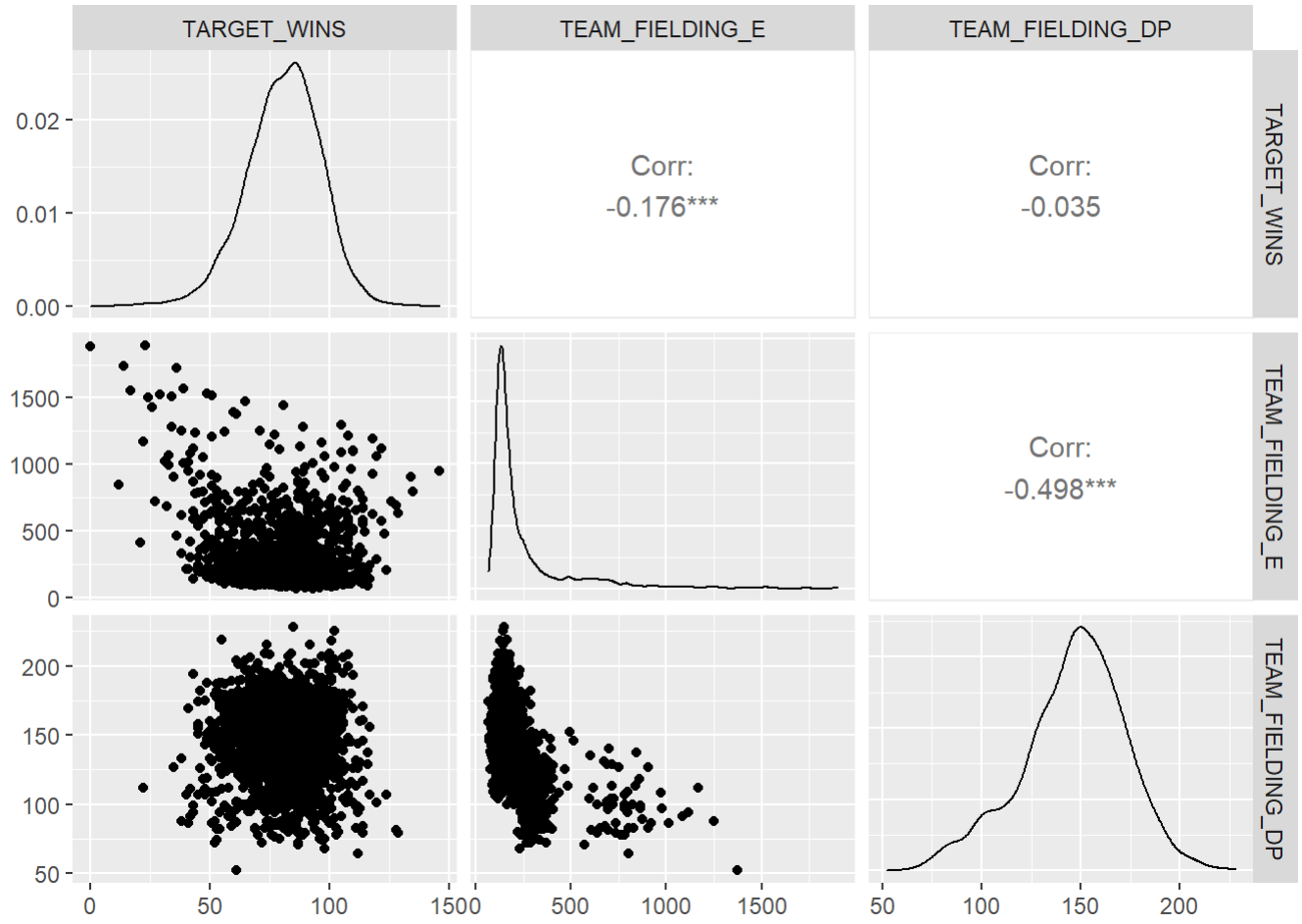
[1] Mean Median SD

```
# Box plots
temp <- train_df %>% select(c("TEAM_FIELDING_E", "TEAM_FIELDING_DP"))
ggplot2::ggplot(stack(temp), aes(x = ind, y = values)) +
  geom_boxplot()
```



Both the Fielding variables are negatively correlated with Wins.

```
# Correlation plot
train_df %>% select(c("TARGET_WINS", "TEAM_FIELDING_E", "TEAM_FIELDING_DP"))%>% GGally::ggpairs
()
```



Pitching

- TARGET_WINS - Number of wins
- TEAM_PITCHING_BB - Walks allowed
- TEAM_PITCHING_H - Hits allowed
- TEAM_PITCHING_HR - Homeruns allowed
- TEAM_PITCHING_SO - Strikeouts by pitchers

Hits allowed (TEAM_PITCHING_H) has a right skew and we are missing some Strikeouts by pitcher (TEAM_PITCHING_SO) values.

```
# Summary
train_df %>% select(c("TEAM_PITCHING_H", "TEAM_PITCHING_HR", "TEAM_PITCHING_BB", "TEAM_PITCHING_
SO")) %>% gtsummary::tbl_summary(statistic =list(c("TEAM_PITCHING_H", "TEAM_PITCHING_HR", "TEAM_
PITCHING_BB", "TEAM_PITCHING_SO") ~ "{mean} {median} {sd}"
))
```

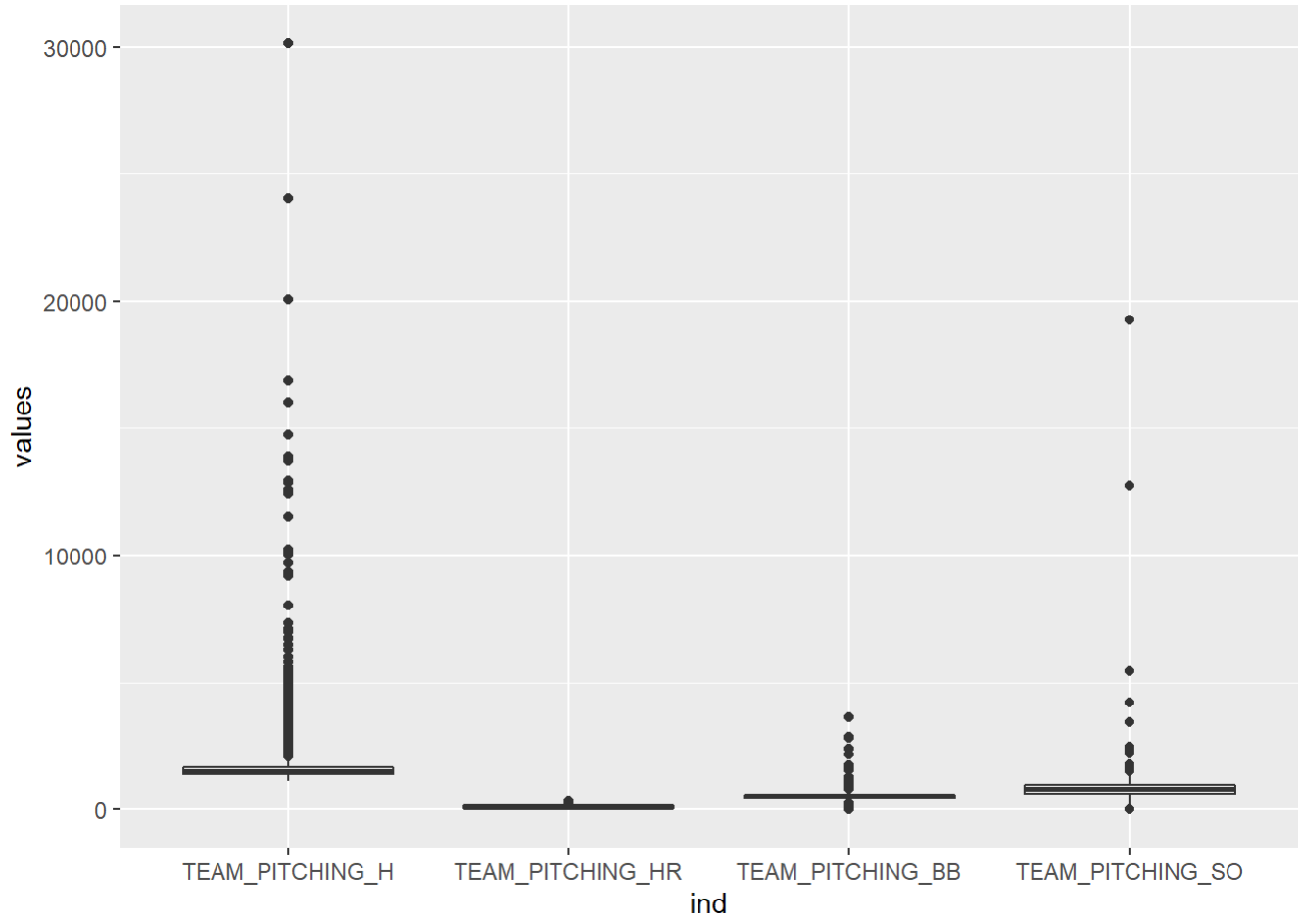
Characteristic	**N = 2,276**[1]
TEAM_PITCHING_H	1,779 1,518 1,407
TEAM_PITCHING_HR	106 107 61

[1] Mean Median SD

Characteristic	N = 2,276[1]
TEAM_PITCHING_BB	553 536 166
TEAM_PITCHING_SO	818 814 553
Unknown	102

[1] Mean Median SD

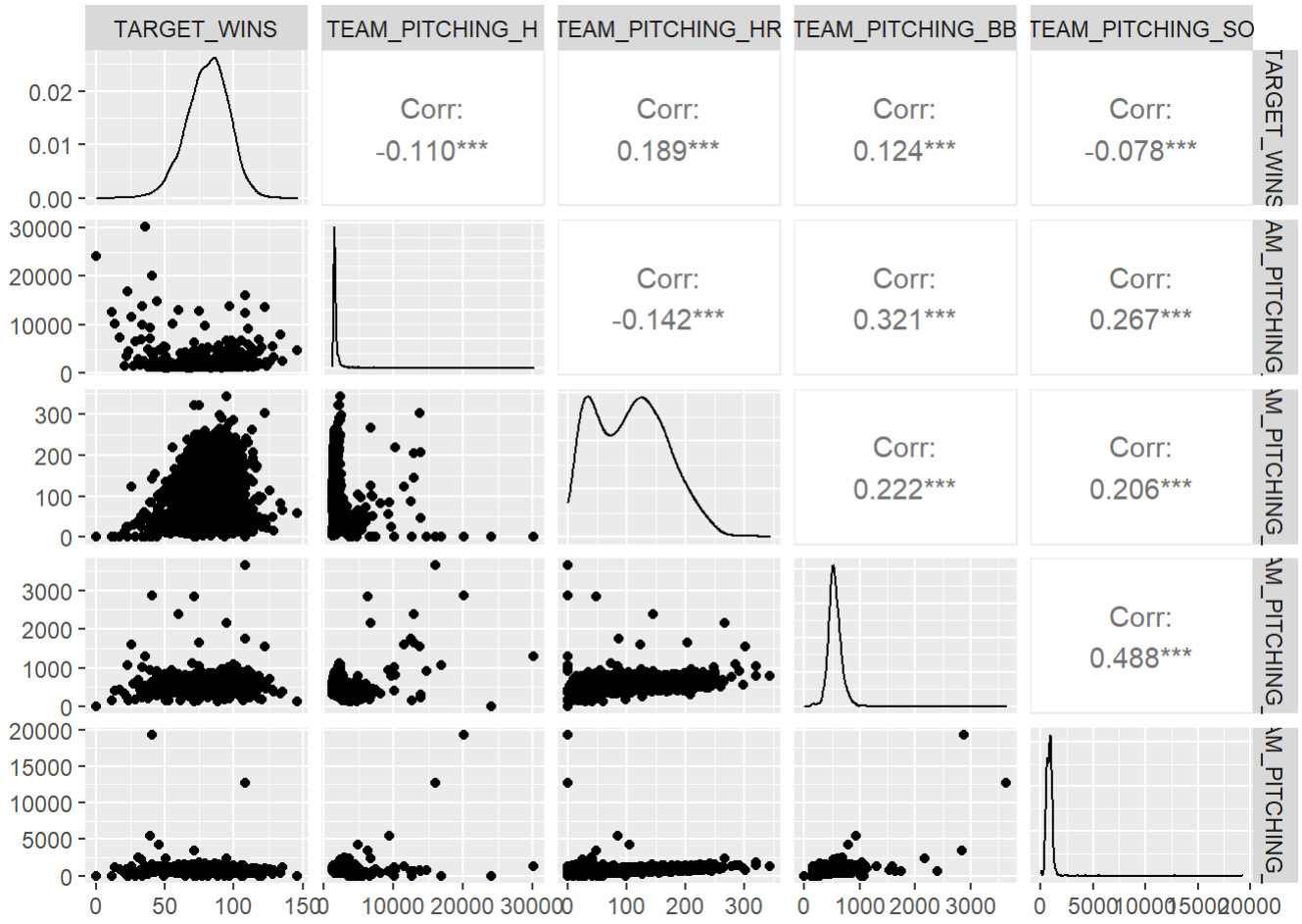
```
# Box plots
temp <- train_df %>% select(c("TEAM_PITCHING_H", "TEAM_PITCHING_HR", "TEAM_PITCHING_BB", "TEAM_P
ITCHING_SO"))
ggplot2::ggplot(stack(temp), aes(x = ind, y = values)) +
  geom_boxplot()
```



Hits allowed is negatively correlated with Winning.

Interestingly, Home runs allowed is positively correlated with Winning.

```
# Correlation plot
train_df %>% select(c("TARGET_WINS","TEAM_PITCHING_H", "TEAM_PITCHING_HR", "TEAM_PITCHING_BB",
"TEAM_PITCHING_SO"))%>% GGally::ggpairs()
```



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:

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_WINS	Number of wins	
TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)	Positive Impact on Wins
TEAM_BATTING_2B	Doubles by batters (2B)	Positive Impact on Wins
TEAM_BATTING_3B	Triples by batters (3B)	Positive Impact on Wins
TEAM_BATTING_HR	Homeruns by batters (4B)	Positive Impact on Wins
TEAM_BATTING_BB	Walks by batters	Positive Impact on Wins
TEAM_BATTING_HBP	Batters hit by pitch (get a free base)	Positive Impact on Wins
TEAM_BATTING_SO	Strikeouts by batters	Negative Impact on Wins
TEAM_BASERUN_SB	Stolen bases	Positive Impact on Wins
TEAM_BASERUN_CS	Caught stealing	Negative Impact on Wins
TEAM_FIELDING_E	Errors	Negative Impact on Wins
TEAM_FIELDING_DP	Double Plays	Positive Impact on Wins
TEAM_PITCHING_BB	Walks allowed	Negative Impact on Wins
TEAM_PITCHING_H	Hits allowed	Negative Impact on Wins
TEAM_PITCHING_HR	Homeruns allowed	Negative Impact on Wins
TEAM_PITCHING_SO	Strikeouts by pitchers	Positive Impact on Wins

https://raw.githubusercontent.com/letisalba/Data_621/master/Homework_1/Images/homework1_table.png (https://raw.githubusercontent.com/letisalba/Data_621/master/Homework_1/Images/homework1_table.png)

- 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 and Preparation

```
# Read data
baseball_df <- read.csv('https://raw.githubusercontent.com/letisalba/Data_621/master/Homework_1/csv/moneyball-training-data.csv')
baseball_eval <- read.csv('https://raw.githubusercontent.com/letisalba/Data_621/master/Homework_1/csv/moneyball-evaluation-data.csv')

# Data overview
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
```

```
dim(baseball_df)
```

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

```
# Data summary
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
```

```
print(paste0('Number of observations: ', nrow(baseball_df)))
```

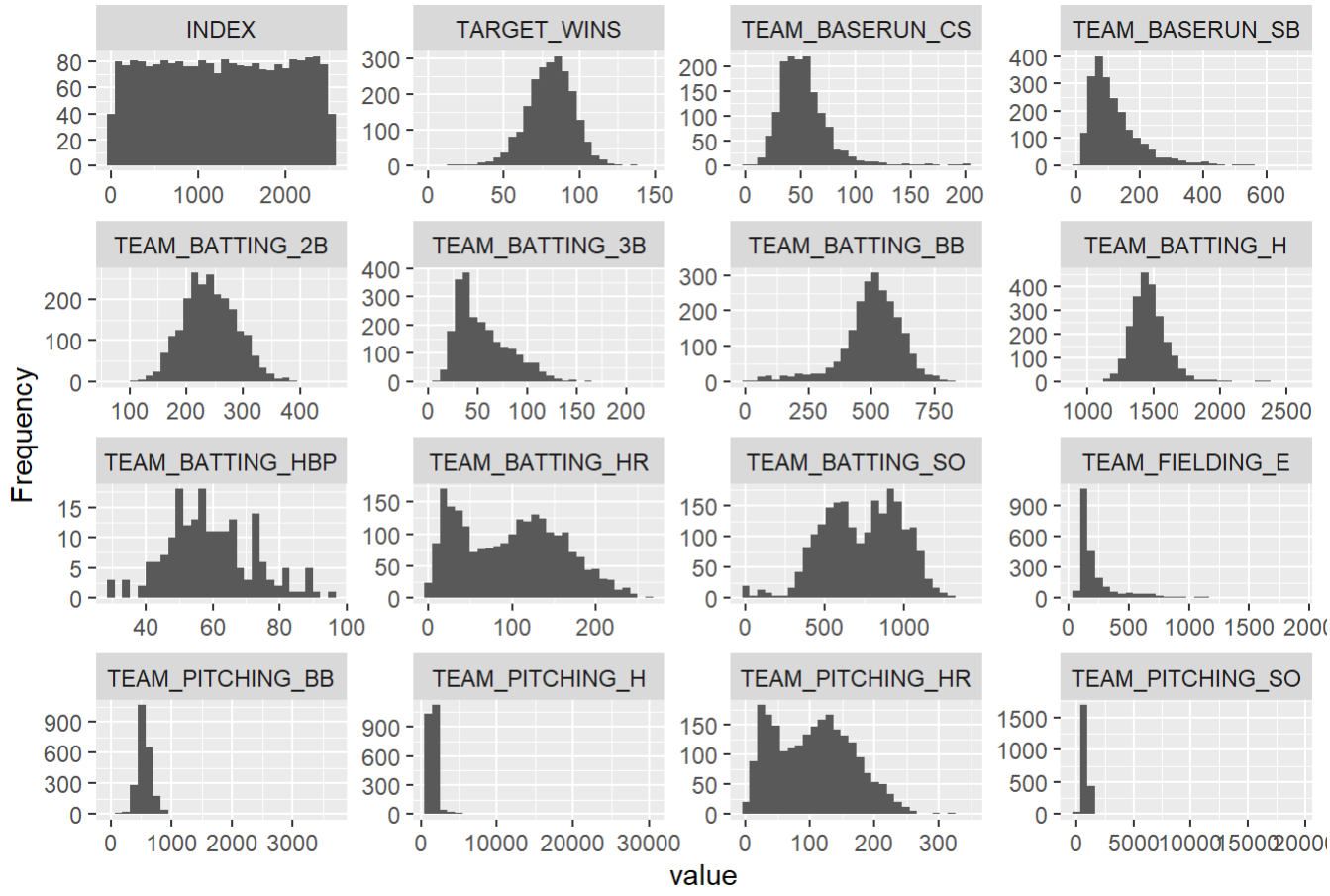
```
## [1] "Number of observations: 2276"
```

```
print(paste0('Observations per year, 1871 - 2006: ',round(nrow(baseball_df)/(2006-1871),2)))
```

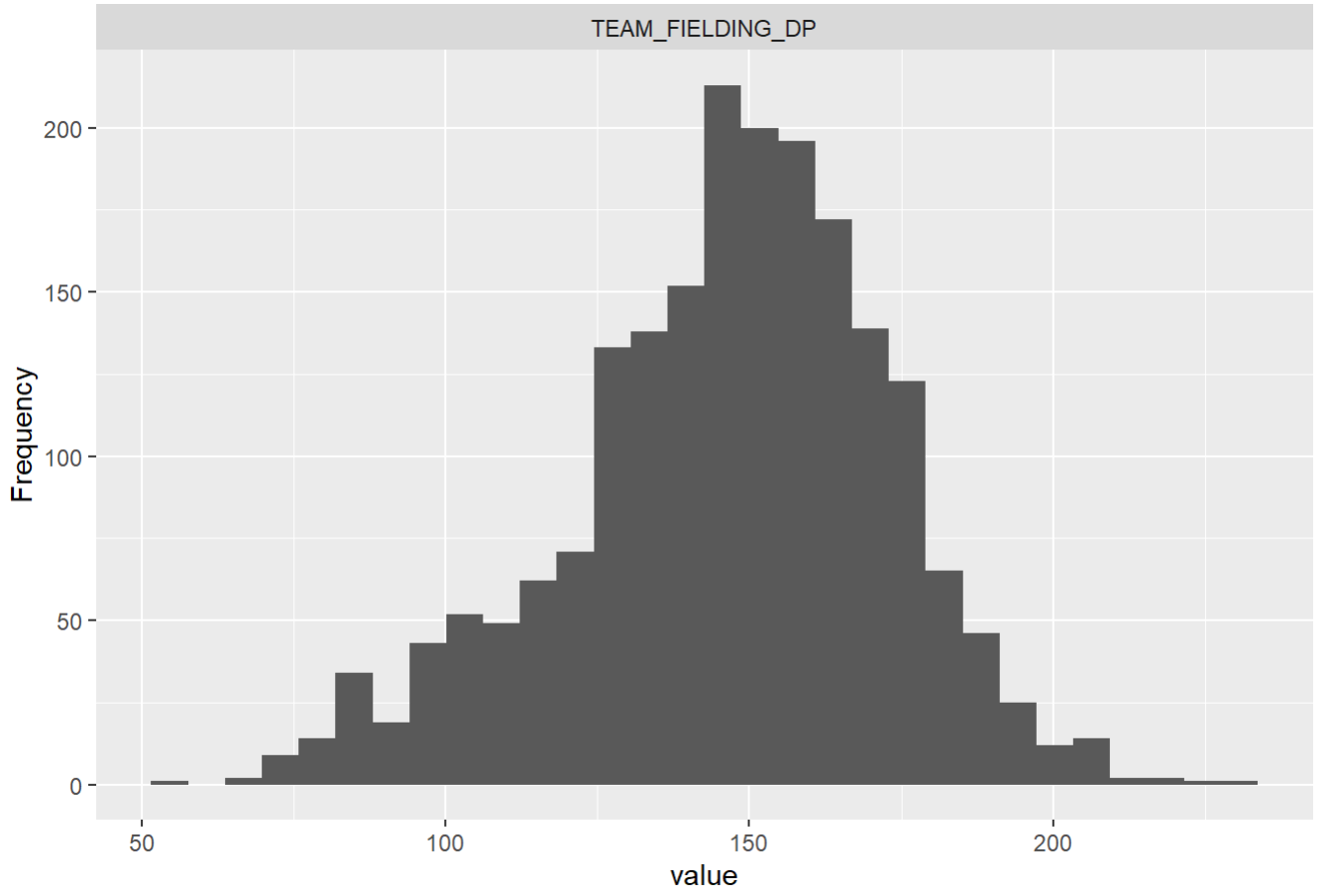
```
## [1] "Observations per year, 1871 - 2006: 16.86"
```

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.

```
# Distribution
plot_histogram(baseball_df)
```


INDEX
TARGET_WINS
TEAM_BASERUN_CS
TEAM_BASERUN_SB
TEAM_BATTING_2B
TEAM_BATTING_3B
TEAM_BATTING_BB
TEAM_BATTING_H
TEAM_BATTING_HBP
TEAM_BATTING_HR
TEAM_BATTING_SO
TEAM_FIELDING_E
TEAM_PITCHING_BB
TEAM_PITCHING_H
TEAM_PITCHING_HR
TEAM_PITCHING_SO
Frequency
value

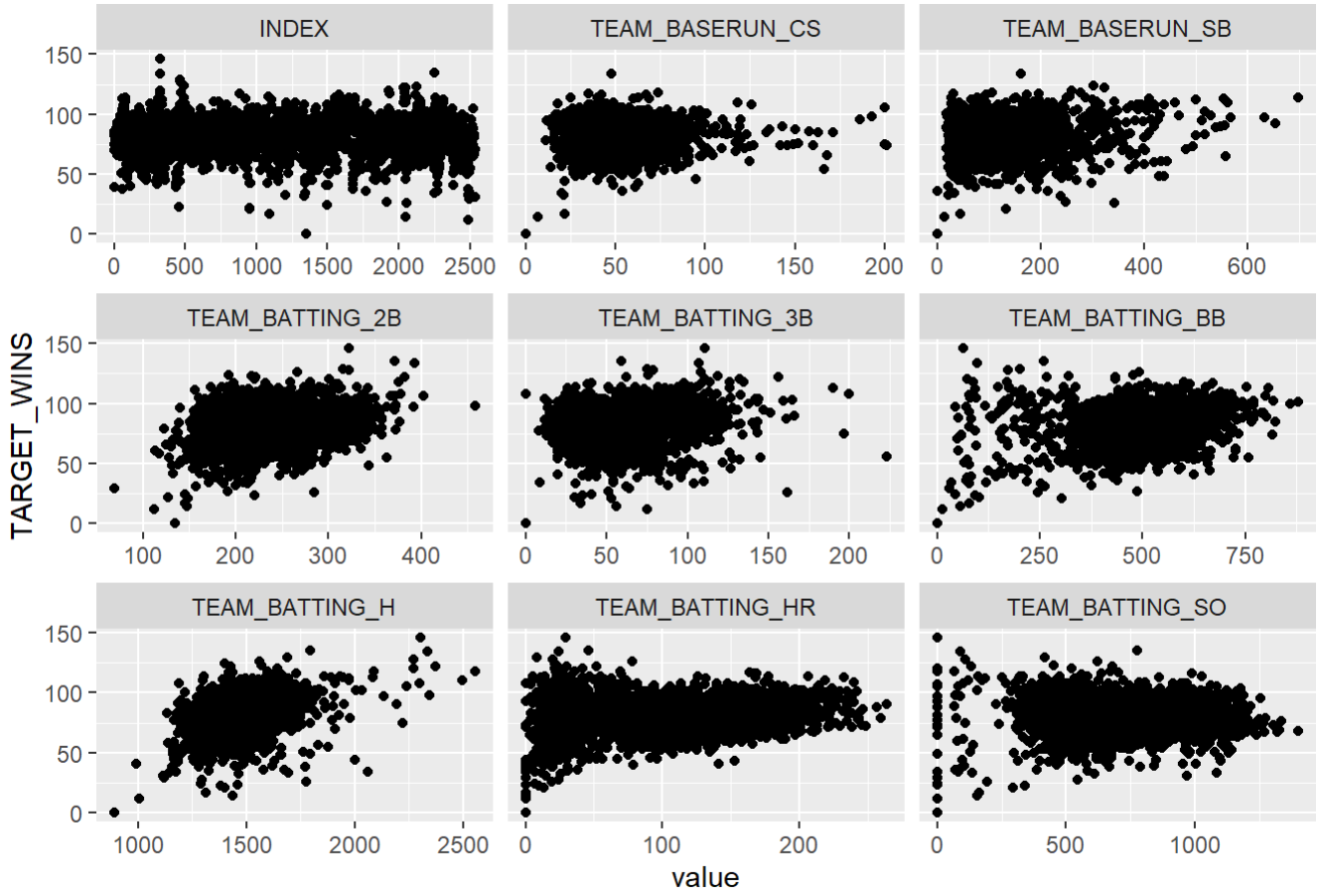
Page 1


TEAM_FIELDING_DP
Frequency
value

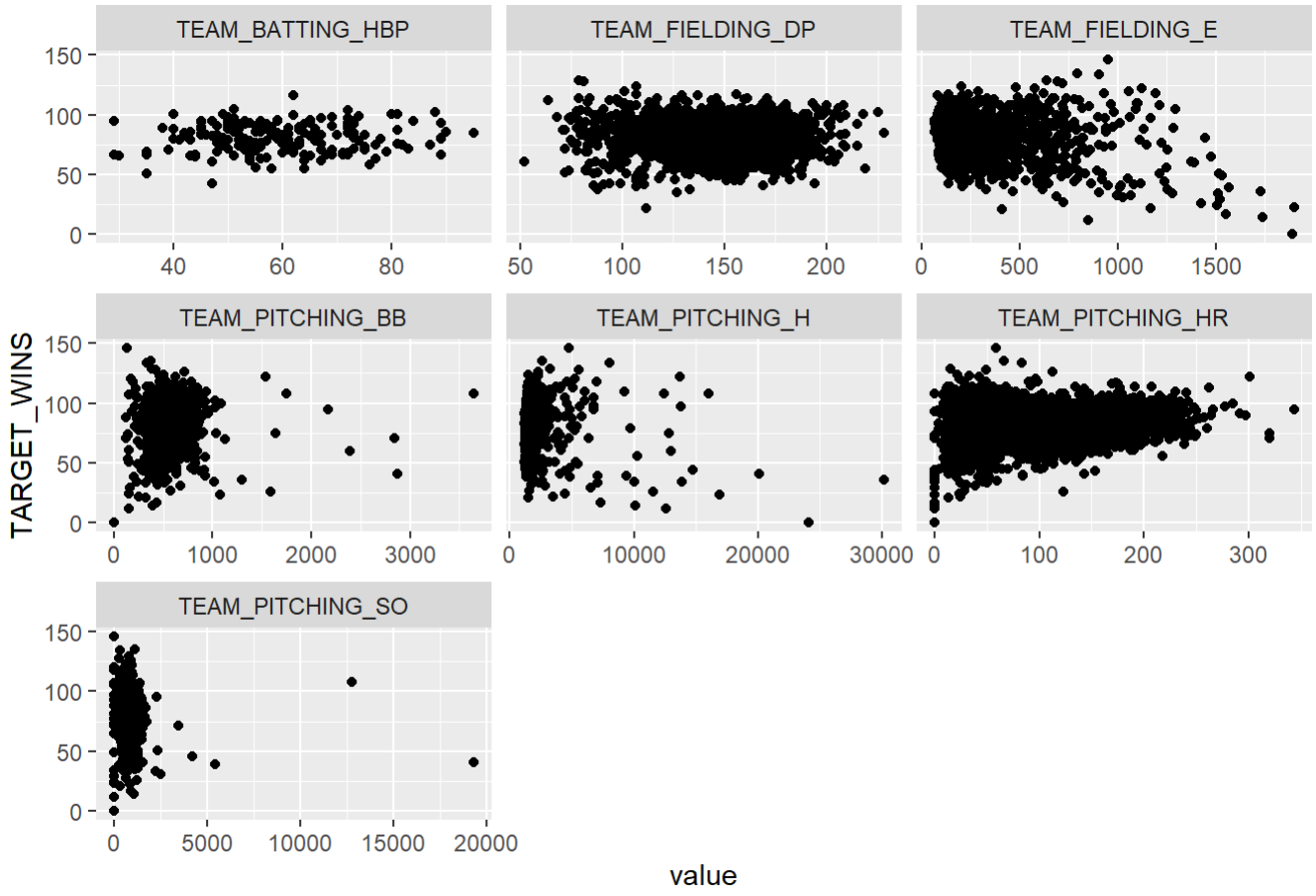
Page 2

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

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

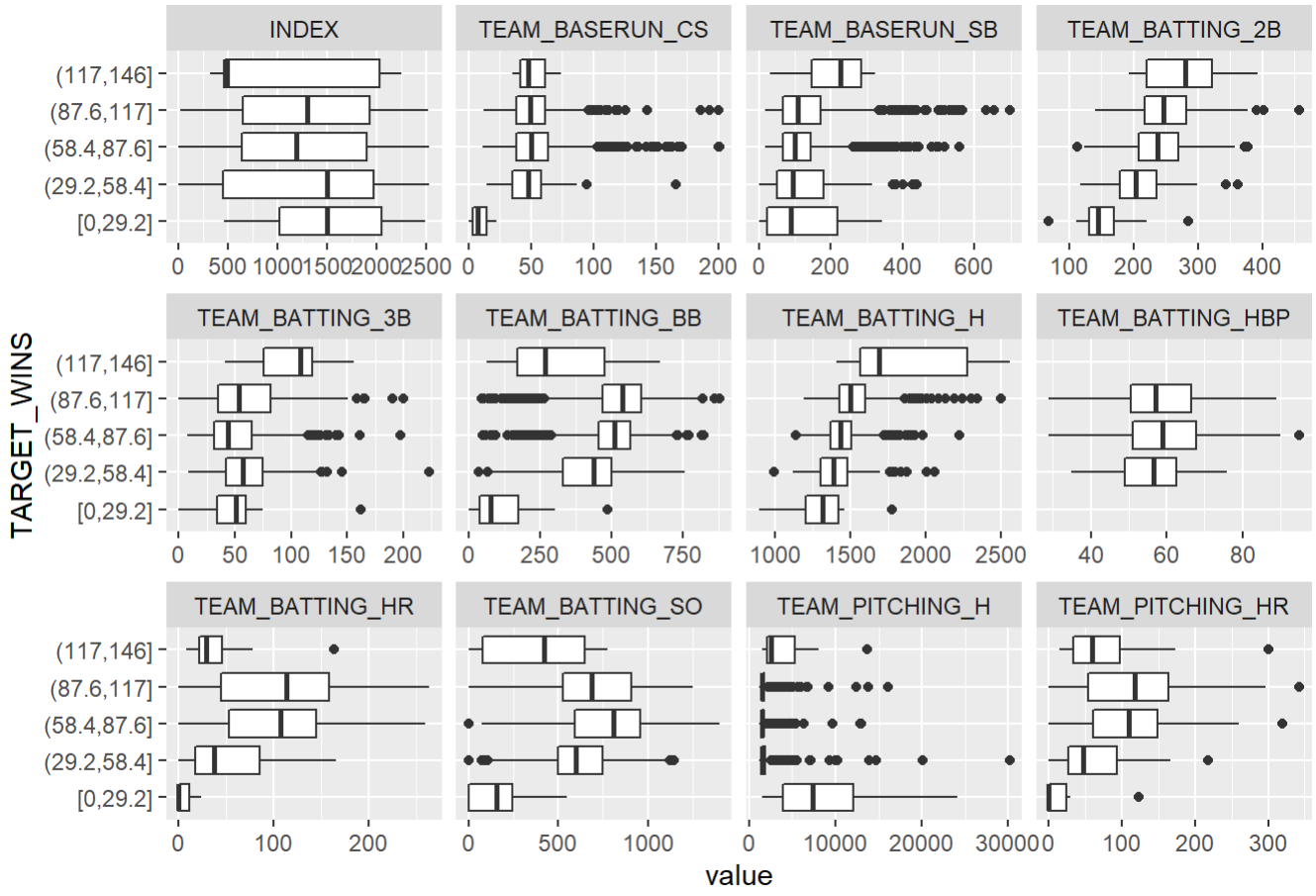



```
## Warning: Removed 2473 rows containing missing values (geom_point).
```

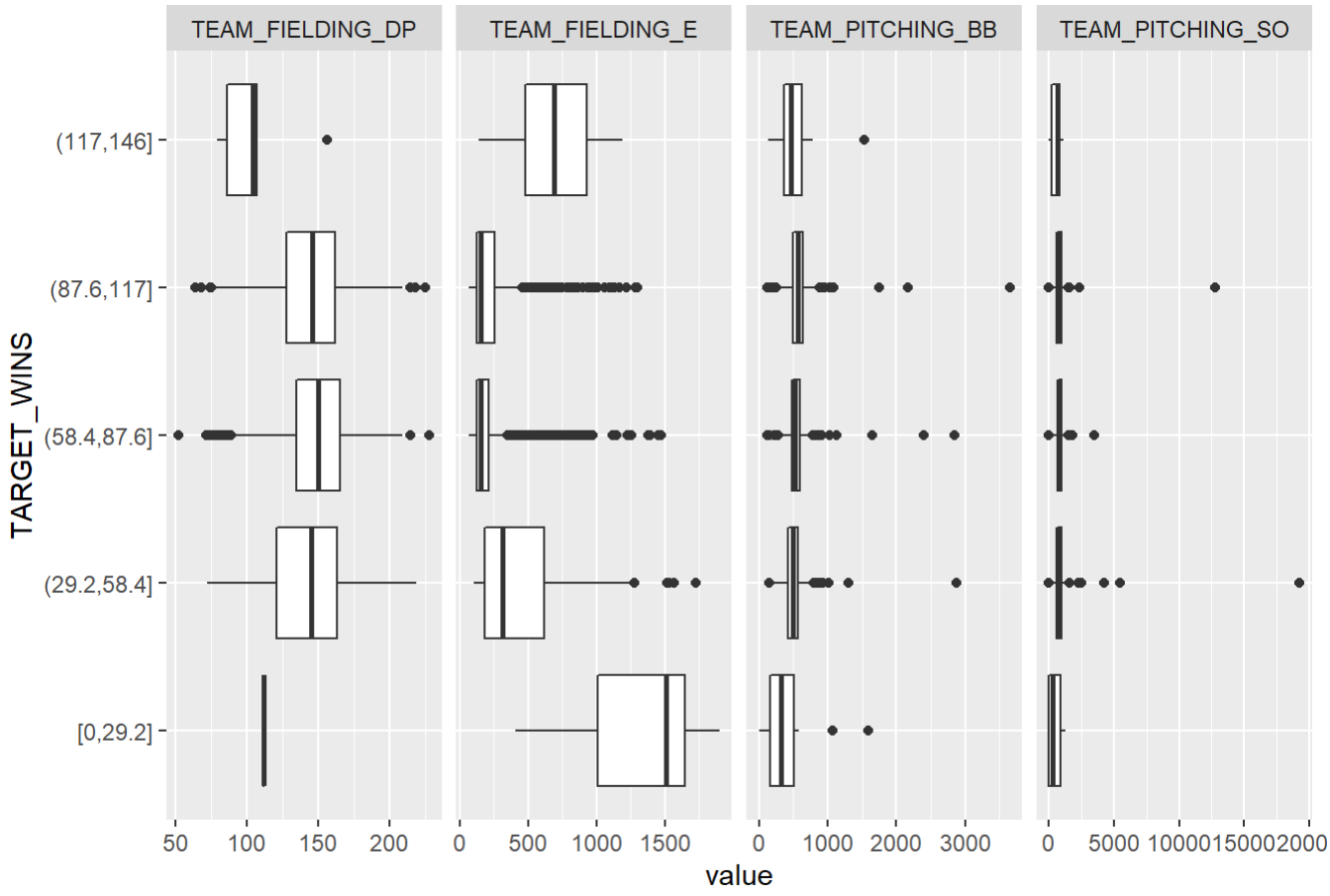



```
# Boxplot for train dataset
plot_boxplot(baseball_df, by = "TARGET_WINS")
```

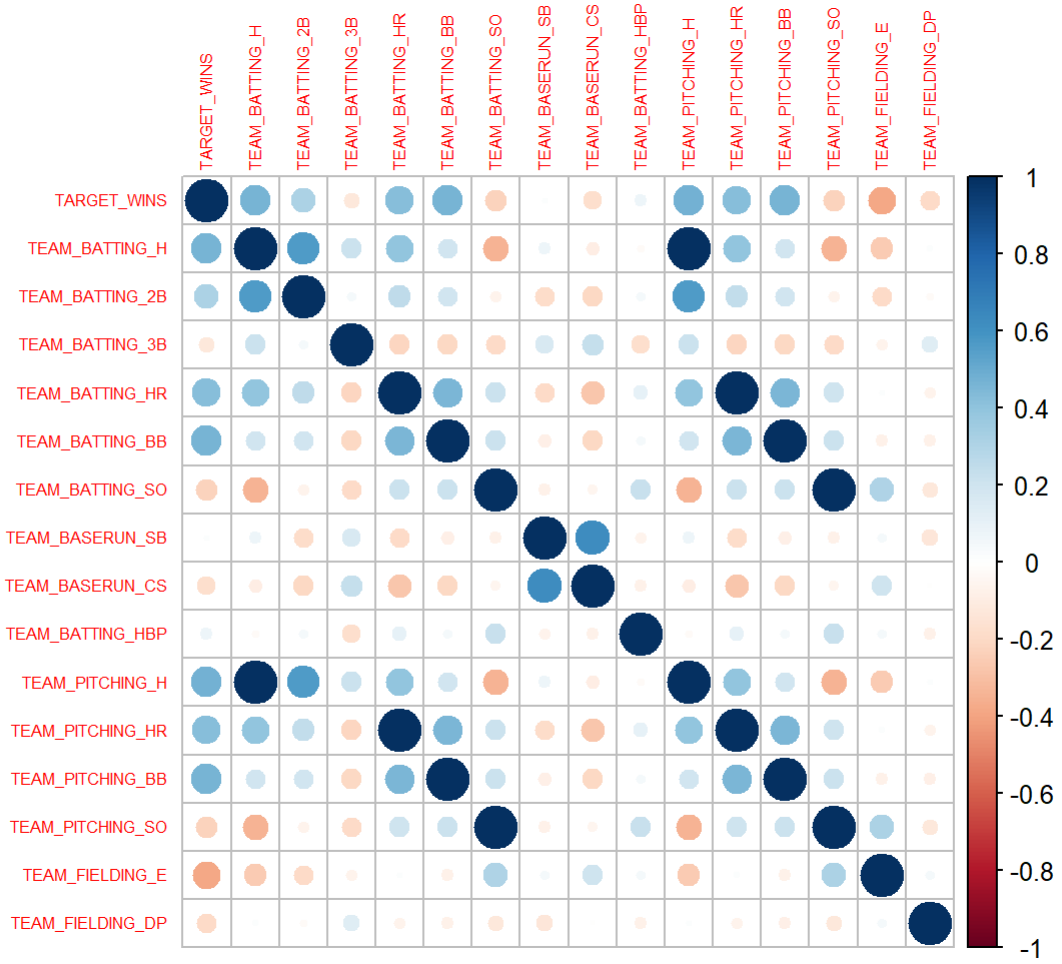
```
## Warning: Removed 3090 rows containing non-finite values (stat_boxplot).
```




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




```
# Correlation plot
corrplot(cor(baseball_df[,2:17], use = 'complete.obs'), tl.cex = 0.5)
```



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. 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

Missing values

```
# Missing values
round(100*colSums(is.na(baseball_df))/nrow(baseball_df),2) %>%
  kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"), font_size = 13)
```

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

In terms of missing values, there are two variables missing many observations. TEAM_BATTING_HBP is missing over 90% of its values, while TEAM_BASERUN_CS is missing just around 30%.

```
#New DF with Missing Removed
baseball_df_mv <- baseball_df[, !names(baseball_df) %in% c('TEAM_BATTING_HBP','TEAM_BASERUN_CS',
'TEAM_FIELDING_DP')]
summary(baseball_df_mv)
```

```
##      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_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB
##  Min.   :  0.0   Min.   : 1137   Min.   :  0.0    Min.   :   0.0
##  1st Qu.: 66.0   1st Qu.: 1419   1st Qu.: 50.0    1st Qu.: 476.0
##  Median :101.0   Median : 1518   Median :107.0    Median : 536.5
##  Mean   :124.8   Mean   : 1779   Mean   :105.7    Mean   : 553.0
##  3rd Qu.:156.0   3rd Qu.: 1682   3rd Qu.:150.0    3rd Qu.: 611.0
##  Max.   :697.0   Max.   :30132   Max.   :343.0    Max.   :3645.0
##  NA's   :131
##  TEAM_PITCHING_SO  TEAM_FIELDING_E
##  Min.   :    0.0   Min.   :  65.0
##  1st Qu.:  615.0   1st Qu.: 127.0
##  Median :  813.5   Median : 159.0
##  Mean   :  817.7   Mean   : 246.5
##  3rd Qu.:  968.0   3rd Qu.: 249.2
##  Max.   :19278.0   Max.   :1898.0
##  NA's   :102
```

```
#Impute NAs with Median
baseball_df_imputed <- mice(baseball_df_mv, m=5, maxit = 5, method = 'pmm')
```

```
##
##  iter imp variable
##   1   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   1   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   1   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   1   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   1   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   2   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   2   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   2   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   2   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   2   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   3   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   3   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   3   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   3   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   3   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   4   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   4   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   4   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   4   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   4   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   5   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   5   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   5   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   5   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
##   5   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_PITCHING_SO
```

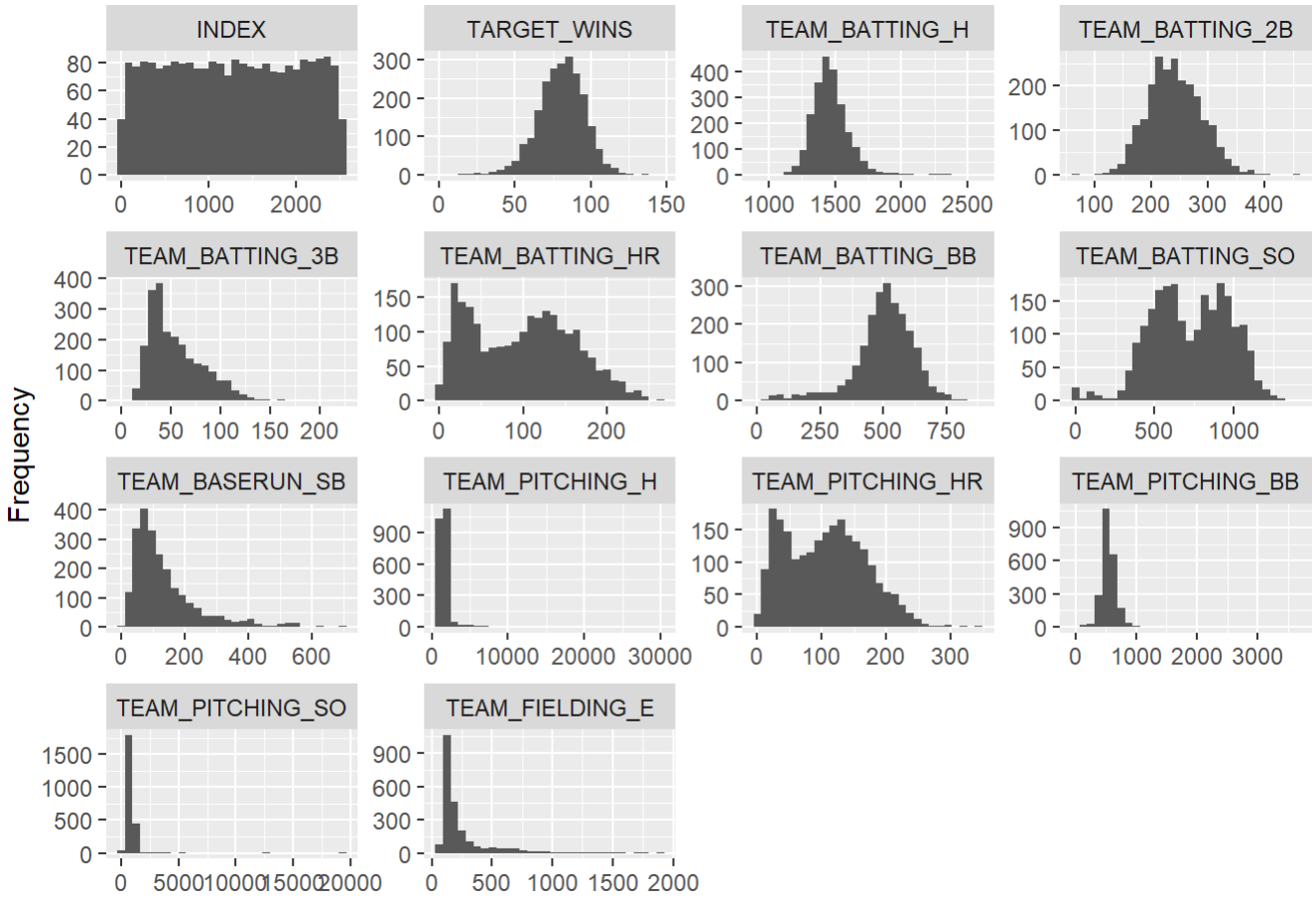
```
baseball_df_final <- complete(baseball_df_imputed)
summary(baseball_df_final)
```

```
##      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.: 541.0
##  Median : 47.00   Median :102.00   Median :512.0   Median : 729.0
##  Mean   : 55.25   Mean   : 99.61   Mean   :501.6   Mean   : 726.9
##  3rd Qu.: 72.00   3rd Qu.:147.00   3rd Qu.:580.0   3rd Qu.: 925.0
##  Max.   :223.00   Max.   :264.00   Max.   :878.0   Max.   :1399.0
##  TEAM_BASERUN_SB TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB
##  Min.   :  0.0   Min.   : 1137   Min.   :  0.0    Min.   :   0.0
##  1st Qu.: 67.0   1st Qu.: 1419   1st Qu.: 50.0    1st Qu.: 476.0
##  Median :105.0   Median : 1518   Median :107.0    Median : 536.5
##  Mean   :136.5   Mean   : 1779   Mean   :105.7    Mean   : 553.0
##  3rd Qu.:169.0   3rd Qu.: 1682   3rd Qu.:150.0    3rd Qu.: 611.0
##  Max.   :697.0   Max.   :30132   Max.   :343.0    Max.   :3645.0
##  TEAM_PITCHING_SO  TEAM_FIELDING_E
##  Min.   :    0.0   Min.   :  65.0
##  1st Qu.:  610.8   1st Qu.: 127.0
##  Median :  799.0   Median : 159.0
##  Mean   :  810.1   Mean   : 246.5
##  3rd Qu.:  957.2   3rd Qu.: 249.2
##  Max.   :19278.0   Max.   :1898.0
```

```
ggplot(melt(baseball_df_final), aes(x=value)) + geom_histogram() + facet_wrap(~variable, scale=
'free') + labs(x='', y='Frequency')
```

```
## Using  as id variables
```

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




```
#Replace Error Maxs
baseball_df_final$TEAM_PITCHING_H[baseball_df_final$TEAM_PITCHING_H > 3*sd(baseball_df_final$TEA
M_PITCHING_H)] <- median(baseball_df_final$TEAM_PITCHING_H)
baseball_df_final$TEAM_PITCHING_BB[baseball_df_final$TEAM_PITCHING_BB > 3*sd(baseball_df_final$T
EAM_PITCHING_BB)] <- median(baseball_df_final$TEAM_PITCHING_BB)
baseball_df_final$TEAM_PITCHING_SO[baseball_df_final$TEAM_PITCHING_SO > 3*sd(baseball_df_final$T
EAM_PITCHING_SO)] <- median(baseball_df_final$TEAM_PITCHING_SO)
baseball_df_final$TEAM_FIELDING_E[baseball_df_final$TEAM_FIELDING_E > 3*sd(baseball_df_final$TEA
M_FIELDING_E)] <- median(baseball_df_final$TEAM_FIELDING_E)
summary(baseball_df_final)
```

```
##      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.: 541.0
##  Median : 47.00   Median :102.00   Median :512.0   Median : 729.0
##  Mean   : 55.25   Mean   : 99.61   Mean   :501.6   Mean   : 726.9
##  3rd Qu.: 72.00   3rd Qu.:147.00   3rd Qu.:580.0   3rd Qu.: 925.0
##  Max.   :223.00   Max.   :264.00   Max.   :878.0   Max.   :1399.0
##  TEAM_BASERUN_SB TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB
##  Min.   :  0.0   Min.   :1137    Min.   :  0.0    Min.   :  0.0
##  1st Qu.: 67.0   1st Qu.:1419    1st Qu.: 50.0    1st Qu.:476.0
##  Median :105.0   Median :1518    Median :107.0    Median :536.5
##  Mean   :136.5   Mean   :1605    Mean   :105.7    Mean   :500.4
##  3rd Qu.:169.0   3rd Qu.:1660    3rd Qu.:150.0    3rd Qu.:536.5
##  Max.   :697.0   Max.   :4134    Max.   :343.0    Max.   :536.5
##  TEAM_PITCHING_SO TEAM_FIELDING_E
##  Min.   :   0.0   Min.   : 65.0
##  1st Qu.: 610.8   1st Qu.:127.0
##  Median : 799.0   Median :159.0
##  Mean   : 788.0   Mean   :198.9
##  3rd Qu.: 954.0   3rd Qu.:215.0
##  Max.   :1600.0   Max.   :681.0
```

Model Building

Model 1 - Full Model

By testing all variables in this first model we are able to see how significant are the variables in our dataset. We will then be able to use this model to base our other models.

```
# Model 1
m1 <- lm(TARGET_WINS ~., data = baseball_df_final, na.action = na.omit)
summ(m1)
```

Observations	2276
Dependent variable	TARGET_WINS
Type	OLS linear regression

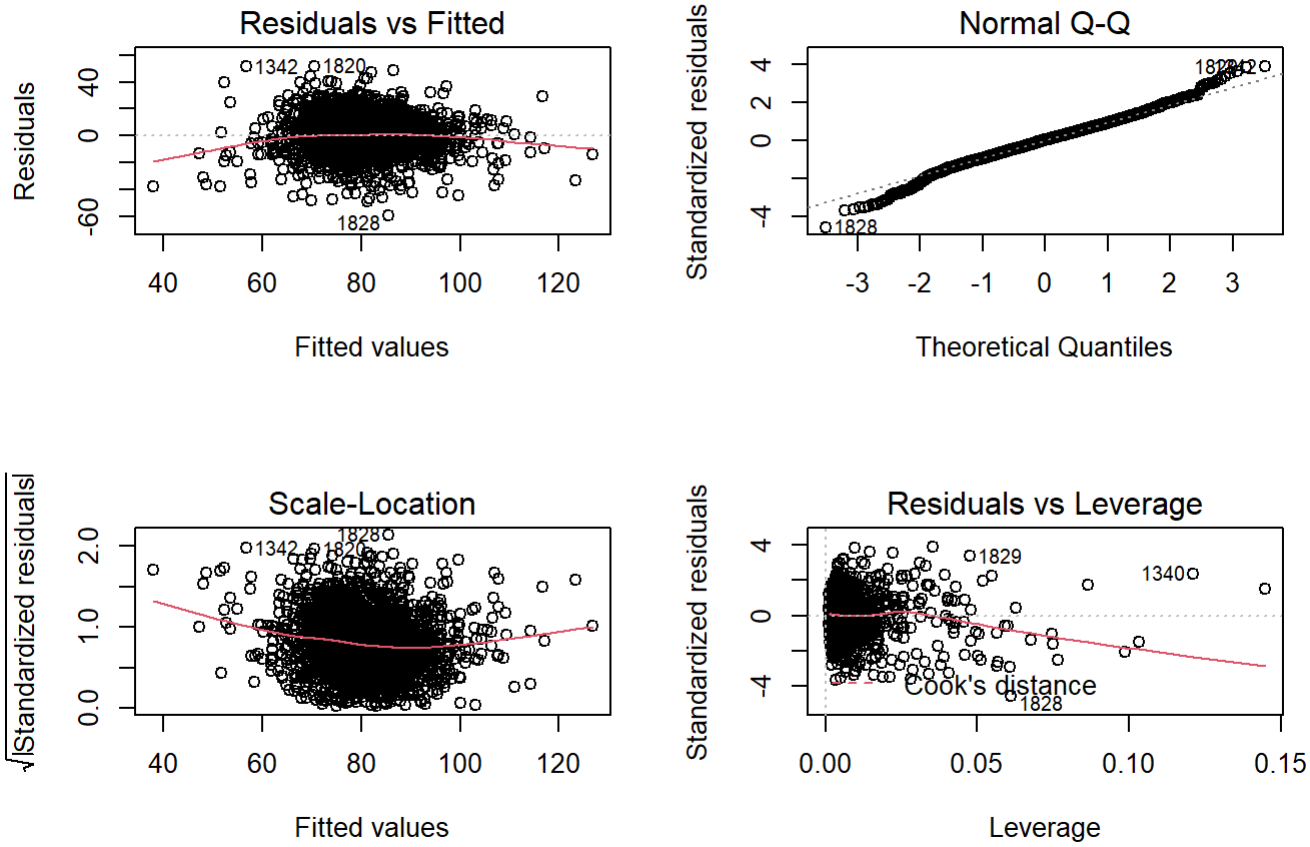
F(13,2262)	67.52

R^2	0.28
Adj. R^2	0.28

	Est.	S.E.	t val.	p
(Intercept)	8.81	5.39	1.63	0.10
INDEX	-0.00	0.00	-1.48	0.14
TEAM_BATTING_H	0.03	0.00	9.36	0.00
TEAM_BATTING_2B	-0.01	0.01	-1.26	0.21
TEAM_BATTING_3B	0.10	0.02	5.49	0.00
TEAM_BATTING_HR	0.10	0.03	3.56	0.00
TEAM_BATTING_BB	0.04	0.00	9.29	0.00
TEAM_BATTING_SO	0.00	0.00	0.74	0.46
TEAM_BASERUN_SB	0.04	0.00	9.65	0.00
TEAM_PITCHING_H	0.00	0.00	2.27	0.02
TEAM_PITCHING_HR	-0.04	0.02	-1.67	0.09
TEAM_PITCHING_BB	-0.02	0.01	-2.58	0.01
TEAM_PITCHING_SO	-0.01	0.00	-1.60	0.11
TEAM_FIELDING_E	-0.02	0.00	-6.67	0.00

Standard errors: OLS

```
# Plot results
par(mfrow=c(2,2))
plot(m1)
```



Model 2: Log transformation

Use of log transformation method which distributes skewness into a more “normally” distributed shape. I applied log transformation for highly skewed variables (less than -1 or greater than 1).

Note: Model 2 was not a successful model compared to model 1. There weren’t any significant changes between the two models therefore discarding this model.

```
# Checking skewness of dataset
sapply(baseball_df_final, function(x) skewness(x)) %>%
  kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"), font_size = 13)
```

	x
INDEX	0.0042149
TARGET_WINS	-0.3987232
TEAM_BATTING_H	1.5713335
TEAM_BATTING_2B	0.2151018
TEAM_BATTING_3B	1.1094652
TEAM_BATTING_HR	0.1860421
TEAM_BATTING_BB	-1.0257599

	x
TEAM_BATTING_SO	-0.2195005
TEAM_BASERUN_SB	1.8985702
TEAM_PITCHING_H	3.3490180
TEAM_PITCHING_HR	0.2877877
TEAM_PITCHING_BB	-2.4931634
TEAM_PITCHING_SO	-0.0204235
TEAM_FIELDING_E	2.1084199

```
# Doing log transformations from model 1
baseball_df_final_log <- baseball_df_final

# Applying log transformation for highly skewed variables
baseball_df_final_log$TEAM_BATTING_H <- log10(baseball_df_final_log$TEAM_BATTING_H + 1)
baseball_df_final_log$TEAM_BATTING_2B <- log10(baseball_df_final_log$TEAM_BATTING_2B + 1)
baseball_df_final_log$TEAM_PITCHING_H <- log10(baseball_df_final_log$TEAM_PITCHING_H + 1)
baseball_df_final_log$TEAM_PITCHING_BB <- log10(baseball_df_final_log$TEAM_PITCHING_BB + 1)
baseball_df_final_log$TEAM_FIELDING_E <- log10(baseball_df_final_log$TEAM_FIELDING_E + 1)
baseball_df_final_log$TEAM_BASERUN_SB <- log10(baseball_df_final_log$TEAM_BASERUN_SB + 1)

# Checking skewness
sapply(baseball_df_final_log, function(x) skewness(x)) %>%
  kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"), font_size = 13)
```

	x
INDEX	0.0042149
TARGET_WINS	-0.3987232
TEAM_BATTING_H	0.7835017
TEAM_BATTING_2B	-0.4041236
TEAM_BATTING_3B	1.1094652
TEAM_BATTING_HR	0.1860421
TEAM_BATTING_BB	-1.0257599
TEAM_BATTING_SO	-0.2195005
TEAM_BASERUN_SB	-0.1114931
TEAM_PITCHING_H	2.1429864
TEAM_PITCHING_HR	0.2877877
TEAM_PITCHING_BB	-13.6001569

	x
TEAM_PITCHING_SO	-0.0204235
TEAM_FIELDING_E	1.0877139

```
# Model 2 log
m2 <- lm(TARGET_WINS ~., data = baseball_df_final_log, na.action = na.omit)

#Summary
summ(m2)
```

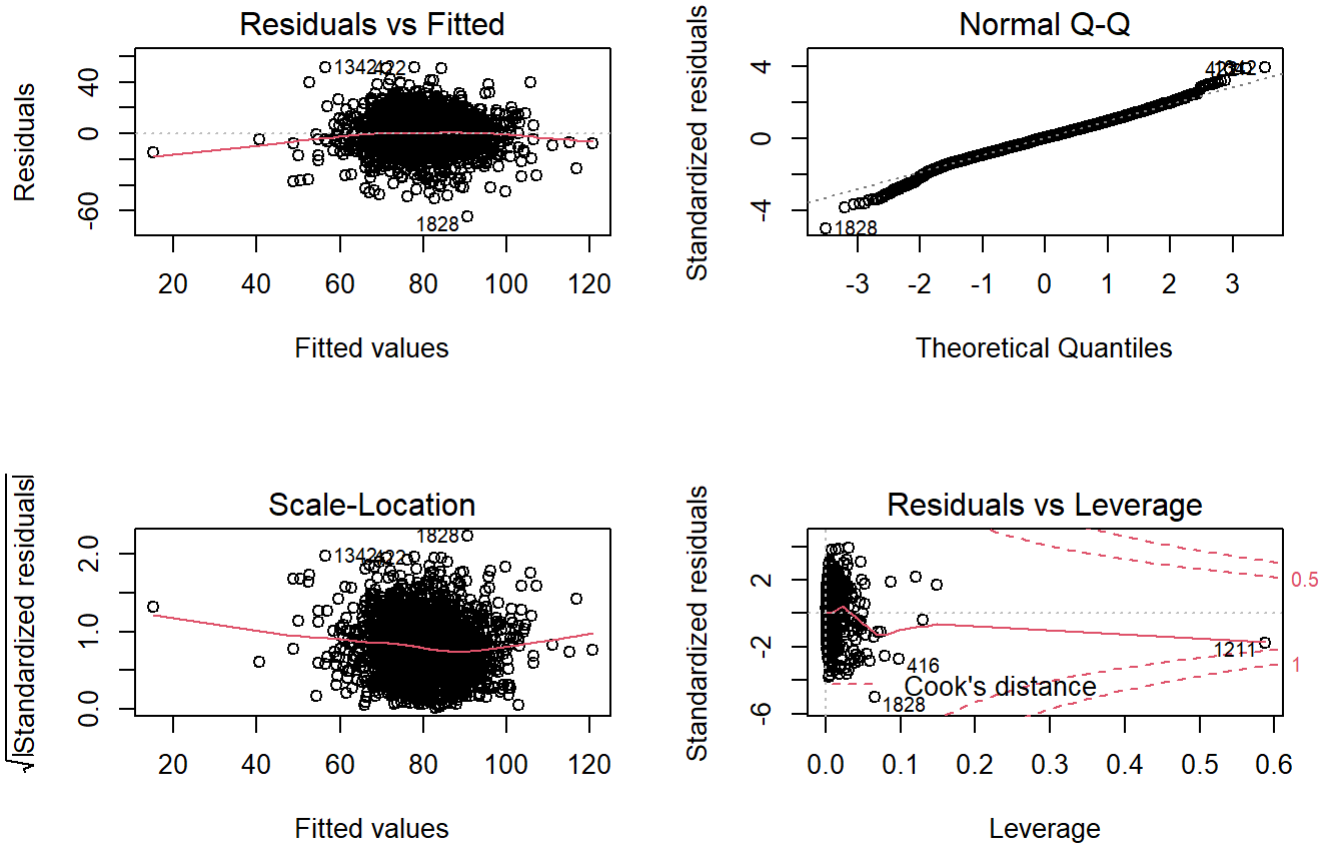
Observations	2276
Dependent variable	TARGET_WINS
Type	OLS linear regression

F(13,2262)	69.74
R²	0.29
Adj. R²	0.28

	Est.	S.E.	t val.	p
(Intercept)	-282.91	37.30	-7.58	0.00
INDEX	-0.00	0.00	-1.67	0.09
TEAM_BATTING_H	108.58	13.73	7.91	0.00
TEAM_BATTING_2B	-5.36	5.25	-1.02	0.31
TEAM_BATTING_3B	0.11	0.02	6.18	0.00
TEAM_BATTING_HR	0.11	0.03	3.89	0.00
TEAM_BATTING_BB	0.03	0.00	8.63	0.00
TEAM_BATTING_SO	-0.00	0.00	-0.81	0.42
TEAM_BASERUN_SB	13.53	1.22	11.05	0.00
TEAM_PITCHING_H	8.59	5.76	1.49	0.14
TEAM_PITCHING_HR	-0.04	0.02	-1.82	0.07
TEAM_PITCHING_BB	-1.95	4.27	-0.46	0.65
TEAM_PITCHING_SO	-0.00	0.00	-1.23	0.22
TEAM_FIELDING_E	-16.98	2.30	-7.37	0.00

Standard errors: OLS

```
# Plot results
par(mfrow=c(2,2))
plot(m2)
```



Model 3: Statistically significant

Focusing on statistically significant values chosen primarily from their R output.

```
# Model 3
m3 <- lm(TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB + TE
AM_BASERUN_SB + TEAM_FIELDING_E, data = baseball_df_final)

# Summary
summ(m3)
```

Observations	2276
Dependent variable	TARGET_WINS
Type	OLS linear regression

F(6,2269)	139.67

R^2	0.27
Adj. R^2	0.27

	Est.	S.E.	t val.	p
(Intercept)	2.43	3.38	0.72	0.47
TEAM_BATTING_H	0.03	0.00	14.57	0.00
TEAM_BATTING_3B	0.09	0.02	5.45	0.00
TEAM_BATTING_HR	0.05	0.01	6.06	0.00
TEAM_BATTING_BB	0.03	0.00	12.02	0.00
TEAM_BASERUN_SB	0.04	0.00	10.06	0.00
TEAM_FIELDING_E	-0.02	0.00	-6.05	0.00

Standard errors: OLS

```
# par(mfrow=c(2,2))
# plot(m3)
```

Model 4: Backwards Elimination

Variables that are not statistically significant are removed to determine a best fit model.

```
# Model 4
m4 <- lm(TARGET_WINS ~ TEAM_BATTING_2B + TEAM_PITCHING_H +  TEAM_PITCHING_HR +  TEAM_PITCHING_BB
+ TEAM_PITCHING_SO, data = baseball_df_final)

# Summary
summ(m4)
```

Observations	2276
Dependent variable	TARGET_WINS
Type	OLS linear regression

F(5,2270)	81.64
R^2	0.15
Adj. R^2	0.15

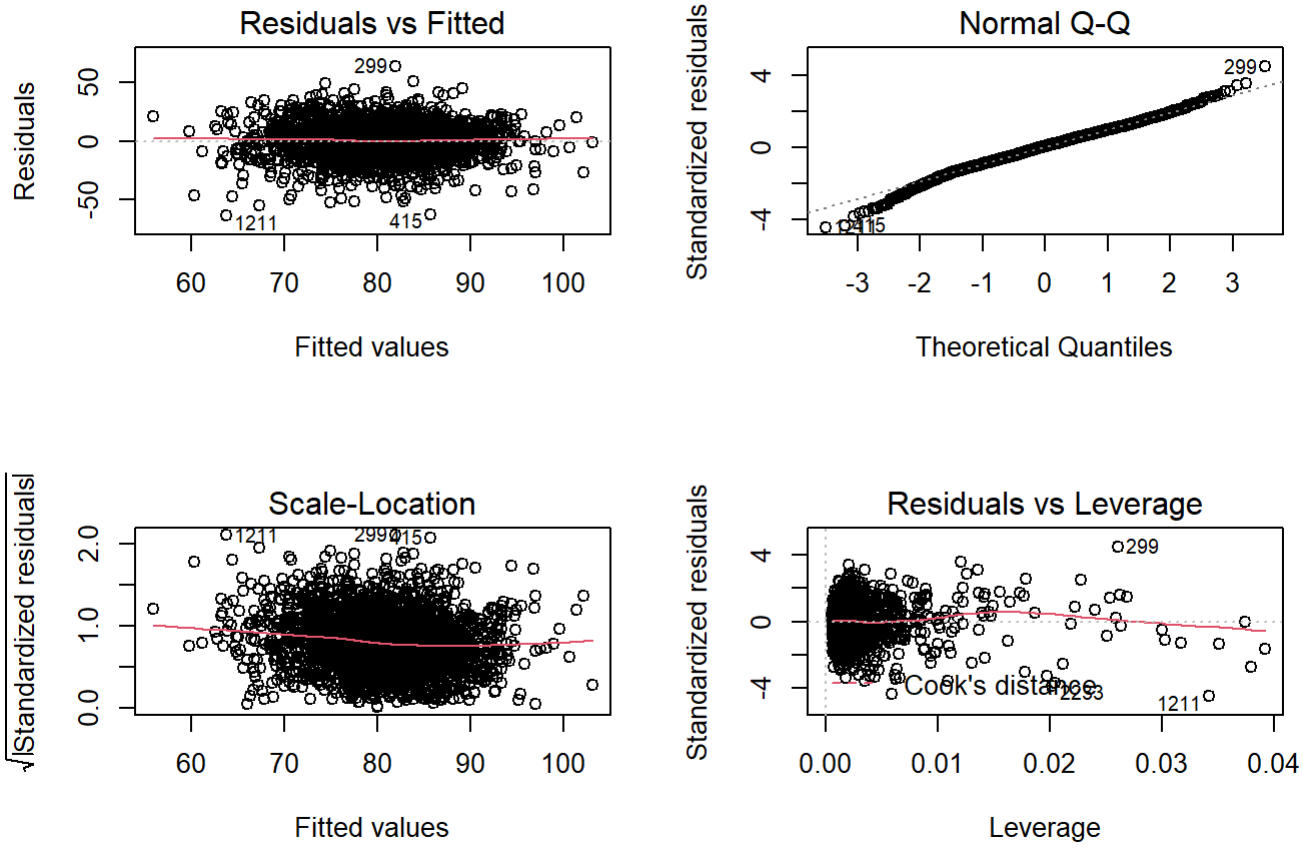
	Est.	S.E.	t val.	p
(Intercept)	44.52	3.44	12.93	0.00

Standard errors: OLS

	Est.	S.E.	t val.	p
TEAM_BATTING_2B	0.06	0.01	7.19	0.00
TEAM_PITCHING_H	0.01	0.00	8.09	0.00
TEAM_PITCHING_HR	0.06	0.01	8.53	0.00
TEAM_PITCHING_BB	0.03	0.01	6.12	0.00
TEAM_PITCHING_SO	-0.02	0.00	-9.76	0.00

Standard errors: OLS

```
# Plot results
par(mfrow=c(2,2))
plot(m4)
```



References

- https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html (https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html)
- https://r-coder.com/correlation-plot-r/ (https://r-coder.com/correlation-plot-r/)