Untitled

Gabrielle Bartomeo

February 11, 2019

```
moneyball_eval <- read.csv(paste(directory, "moneyball-evaluation-data.csv", sep=""))
moneyball_train <- read.csv(paste(directory, "moneyball-training-data.csv", sep=""))[,-1] # use me
```

Training Exploration

Summary()

```
summary(moneyball_train)
```

```
##   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
```

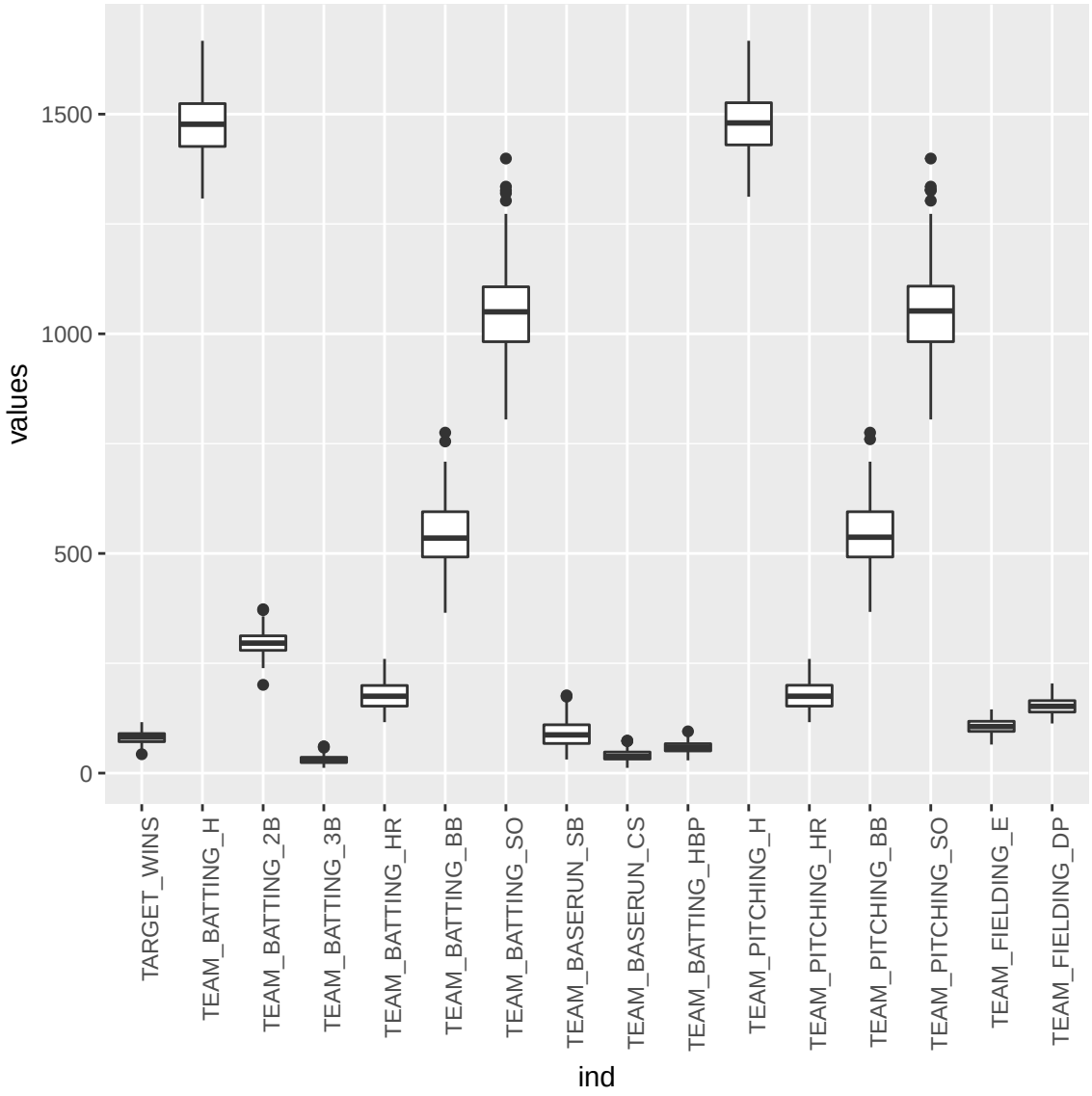
Standard Deviation

```
sapply(moneyball_train[complete.cases(moneyball_train),], sd)
```

```
##      TARGET_WINS   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B
##        12.115013        76.147869        26.329335         9.043878
##  TEAM_BATTING_HR  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB
##        32.413243        74.842133       104.156382        29.916401
##  TEAM_BASERUN_CS TEAM_BATTING_HBP  TEAM_PITCHING_H TEAM_PITCHING_HR
##        11.898334        12.967123        75.788625        32.391678
## TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E TEAM_FIELDING_DP
##        74.916681       104.347209        16.632162        17.611682
```

Boxplots

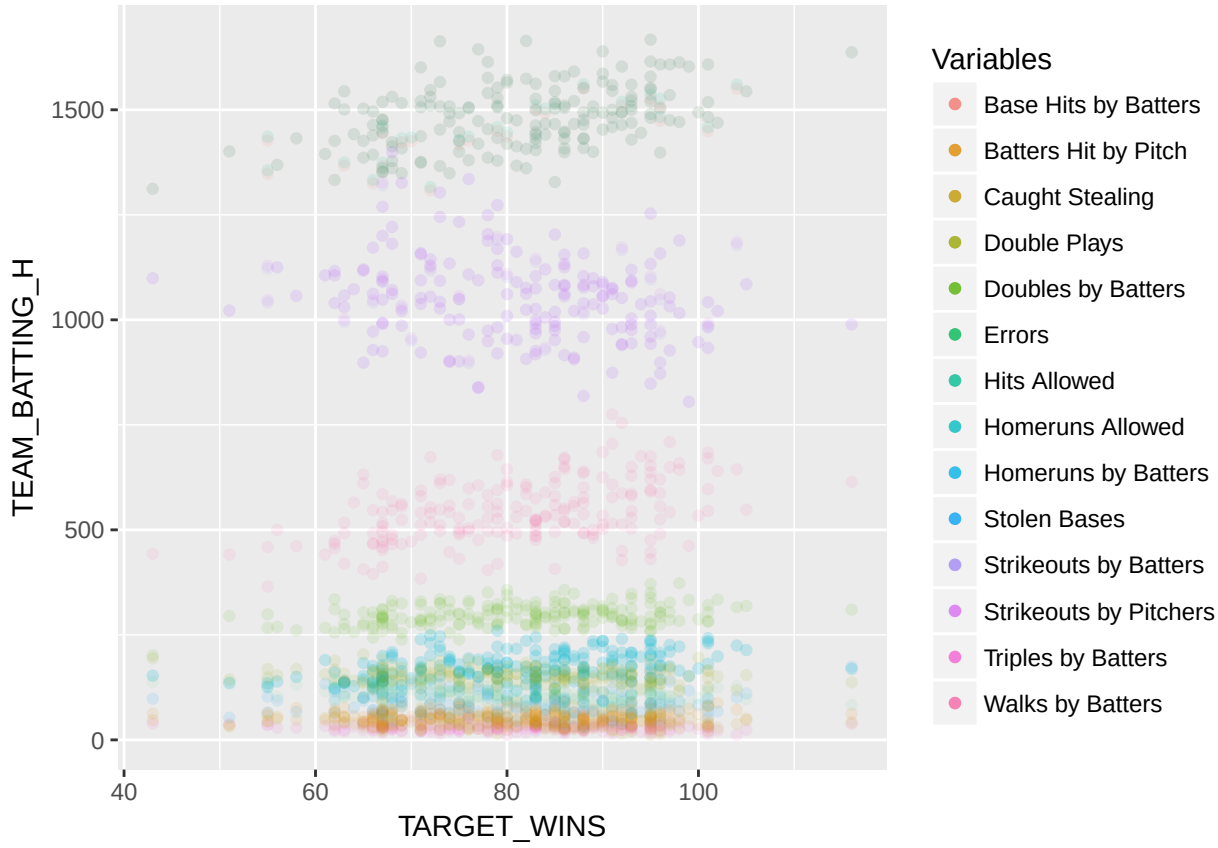
```
ggplot(stack(moneyball_train[complete.cases(moneyball_train),]), aes(x=ind, y=values)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Point plots

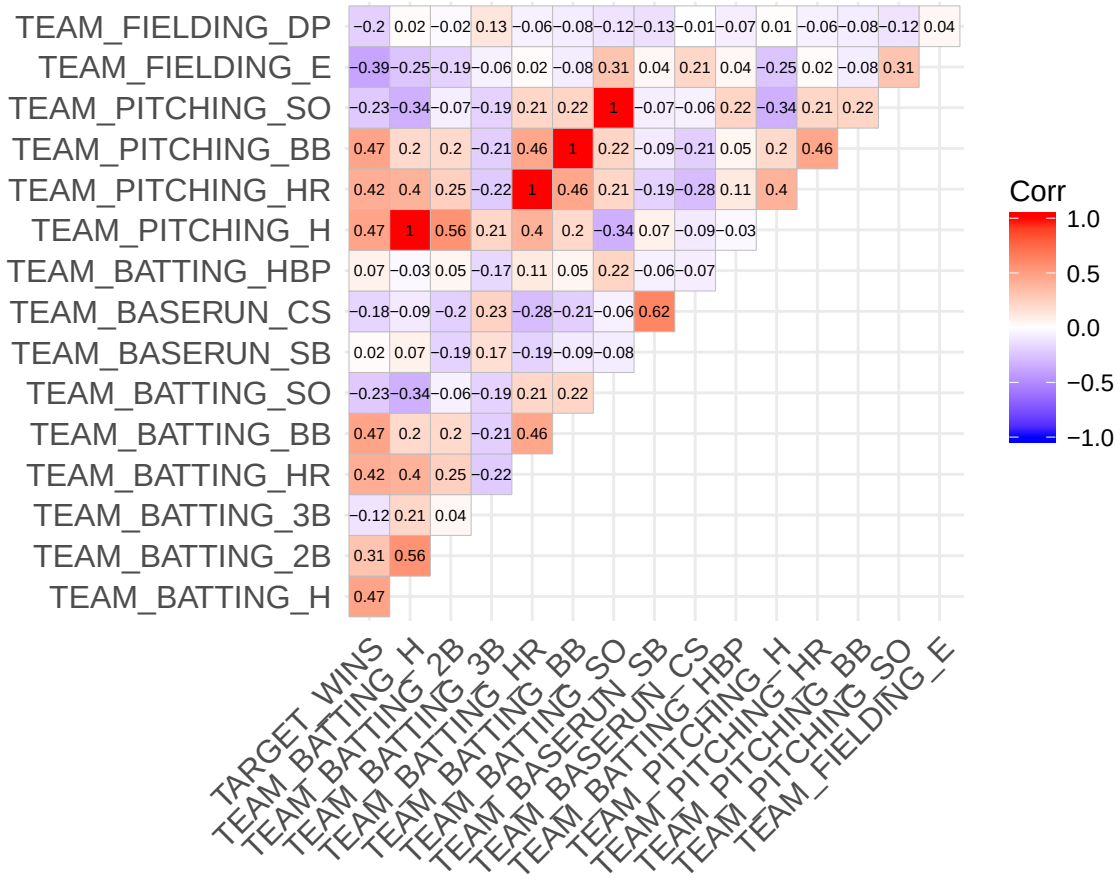
```
ggplot(data=moneyball_train[complete.cases(moneyball_train),], aes(x=TARGET_WINS)) +
  geom_point(aes(y=TEAM_BATTING_H, color="Base Hits by Batters"), alpha=0.1) +
  geom_point(aes(y=TEAM_BATTING_2B, color="Doubles by Batters"), alpha=0.1) +
  geom_point(aes(y=TEAM_BATTING_3B, color="Triples by Batters"), alpha=0.1) +
  geom_point(aes(y=TEAM_BATTING_HR, color="Homeruns by Batters"), alpha=0.1) +
  geom_point(aes(y=TEAM_BATTING_BB, color="Walks by Batters"), alpha=0.1) +
  geom_point(aes(y=TEAM_BATTING_SO, color="Strikeouts by Batters"), alpha=0.1) +
  geom_point(aes(y=TEAM_BASERUN_SB, color="Stolen Bases"), alpha=0.1) +
  geom_point(aes(y=TEAM_BASERUN_CS, color="Caught Stealing"), alpha=0.1) +
  geom_point(aes(y=TEAM_BATTING_HBP, color="Batters Hit by Pitch"), alpha=0.1) +
  geom_point(aes(y=TEAM_PITCHING_H, color="Hits Allowed"), alpha=0.1) +
  geom_point(aes(y=TEAM_PITCHING_HR, color="Homeruns Allowed"), alpha=0.1) +
  geom_point(aes(y=TEAM_PITCHING_SO, color="Strikeouts by Pitchers"), alpha=0.1) +
```

```
  geom_point(aes(y=TEAM_FIELDING_E, color="Errors"), alpha=0.05) +
  geom_point(aes(y=TEAM_FIELDING_DP, color="Double Plays"), alpha=0.1) +
  labs(color="Variables", ylab="Variables")
```



Correlation

```
ggcorrplot(as.data.frame(round(cor(moneyball_train[complete.cases(moneyball_train),]), 3)),
           type="upper", lab=TRUE, lab_size=2)
```



Missing values by variable

```
sapply(moneyball_train, function(x) sum(is.na(x)))
```

```
##      TARGET_WINS   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B
##                0                0                0                0
##  TEAM_BATTING_HR  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB
##                0                0              102              131
##  TEAM_BASERUN_CS TEAM_BATTING_HBP  TEAM_PITCHING_H TEAM_PITCHING_HR
##              772             2085                0                0
## TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E TEAM_FIELDING_DP
##                0              102                0              286
```