Untitled

```
library(tufte)
library(ggplot2)
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 3.5.2
```

```
library(ggcorrplot)
library(Matrix)
library(gridExtra)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------- tidyverse 1.2.1 --
```

```
## v tibble  1.4.2     v purrr   0.2.5
## v tidyr   0.8.2     v dplyr   0.7.8
## v readr   1.2.1     v stringr 1.3.1
## v tibble  1.4.2     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::combine() masks gridExtra::combine()
## x tidyr::expand()  masks Matrix::expand()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(matrixcalc)
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:scales':
##
##     alpha, rescale
```

```
## The following objects are masked from 'package:ggplot2':
```

```
##
##     %+%, alpha
```

```
library(GGally)
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
library(ggpubr)
```

```
## Loading required package: magrittr
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(leaps)
```

```
# Load in data
mb_train <- read.csv("./data/moneyball-training-data.csv")[,-1] # use me
```

```
# Removes all rows with missing data
mb_complete <- mb_train[complete.cases(mb_train),]
```

```
Means <- sapply(mb_complete, mean)
Stan_Dev <- sapply(mb_complete, sd)
```

```
Data_Summary <- summary(mb_train)
```

```
Boxplots <- ggplot(stack(mb_complete), aes(x=ind, y=values)) +
    geom_boxplot() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
Point_plots <- ggplot(data=mb_complete, 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(mb_complete), 3)),
           type="upper", lab=TRUE, lab_size=2)
```

```
Missing_values <- sapply(mb_train, function(x) sum(is.na(x)))
```

```
Histograms <- mb_train %>%
    gather() %>%
    ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram()
```

```
# Pairs <- pairs(mb_complete)  This doesn't work, not sure why

Corr_matrix <- round(cor(mb_complete),2)
```

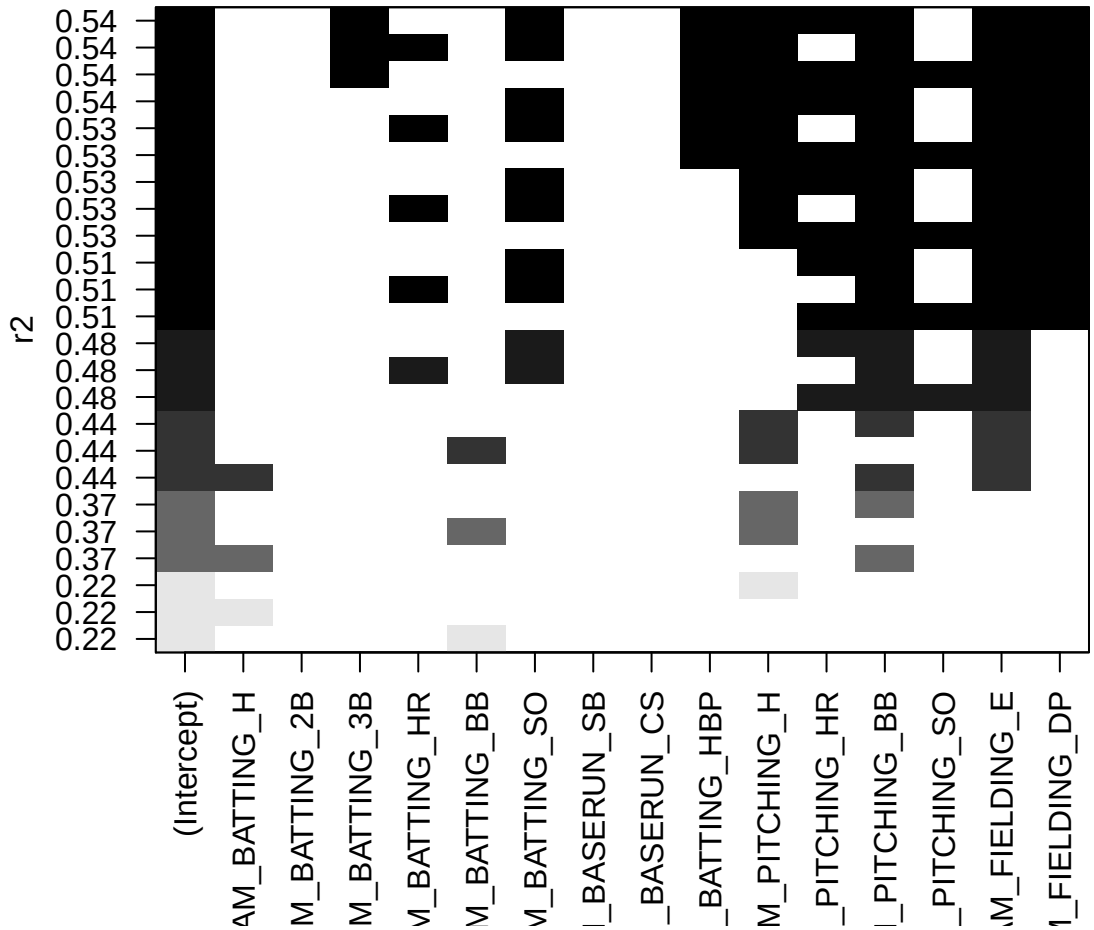
Linear Models

Unscaled

```
# Basic linear model with all variables
mb_lm <- lm(TARGET_WINS ~ ., mb_train)
```

```
LM_Summary <- summary(mb_lm)
```

```
# All Subsets Regression from leaps package
Leaps <- regsubsets(x=mb_complete[,2:16], y=mb_complete[,1], nbest=3)
# plot a table of models showing variables in each model.
# models are ordered by the selection statistic.
Leaps_plot <- plot(Leaps, scale="r2")
```

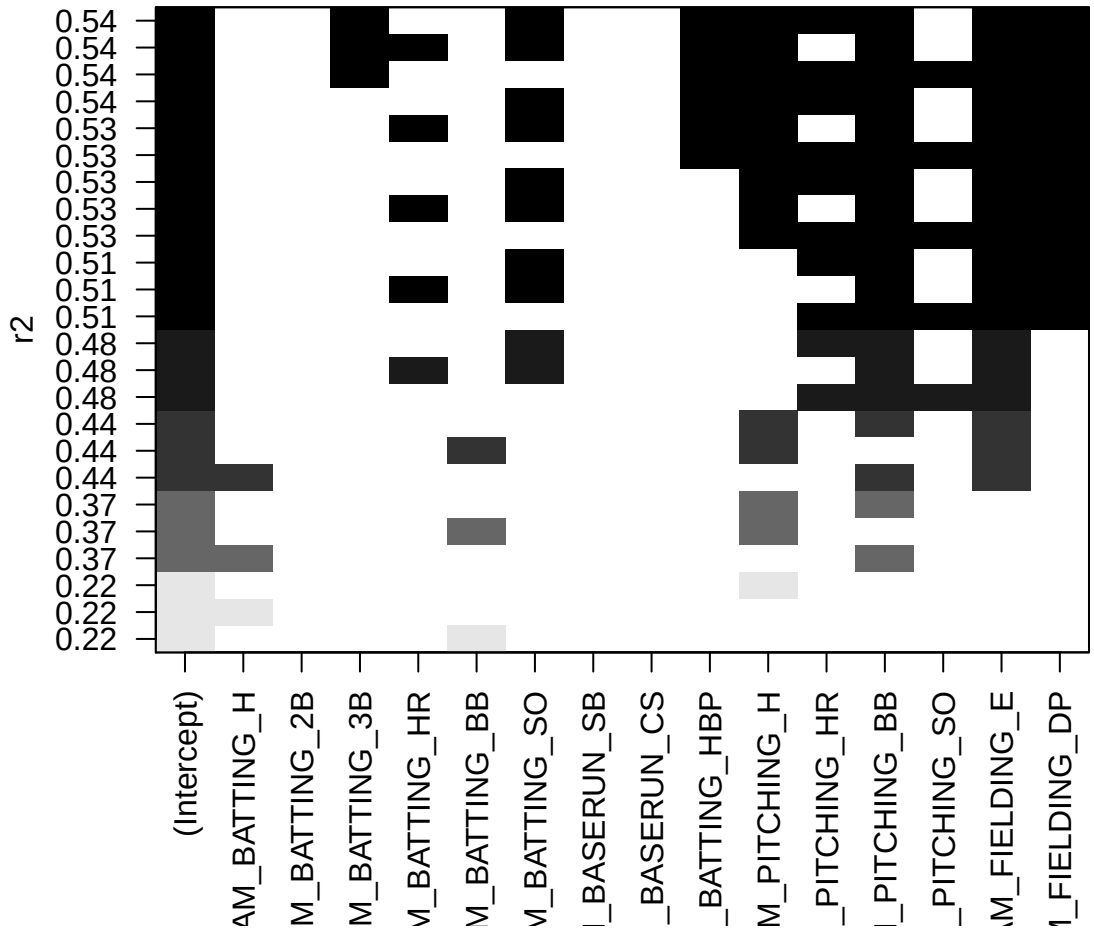


Scaled

```
# Scale all the predictor variables
mb_scaled <- as.data.frame(mb_complete[,2:16], center=Means, scale=Stan_Dev)
mb_scaled$TARGET_WINS <- mb_complete[,1]
# Linear model using all scaled predictors
mb_scaled_lm <- lm(TARGET_WINS ~ ., mb_scaled)
```

```
Scaled_LM_Summary <- summary(mb_scaled_lm)
```

```
# All Subsets Regression from leaps package on SCALED data
Scaled_Leaps <- regsubsets(x=mb_complete[,1:15], y=mb_scaled[,16], nbest=3)
# plot a table of models showing variables in each model.
# models are ordered by the selection statistic.
Scaled_Leaps_plot <- plot(Leaps, scale="r2")
```


r2
0.54
0.54
0.54
0.54
0.53
0.53
0.53
0.53
0.53
0.51
0.51
0.51
0.48
0.48
0.48
0.44
0.44
0.44
0.37
0.37
0.37
0.22
0.22
0.22
(Intercept)
4M_BATTING_H
M_BATTING_2B
M_BATTING_3B
M_BATTING_HR
M_BATTING_BB
M_BATTING_SO
_BASERUN_SB
_BASERUN_CS
_BATTING_HBP
M_PITCHING_H
_PITCHING_HR
_PITCHING_BB
_PITCHING_SO
M_FIELDING_E
1_FIELDING_DP