CollegeFootballScores

Steven Mazurski

3/12/2021

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

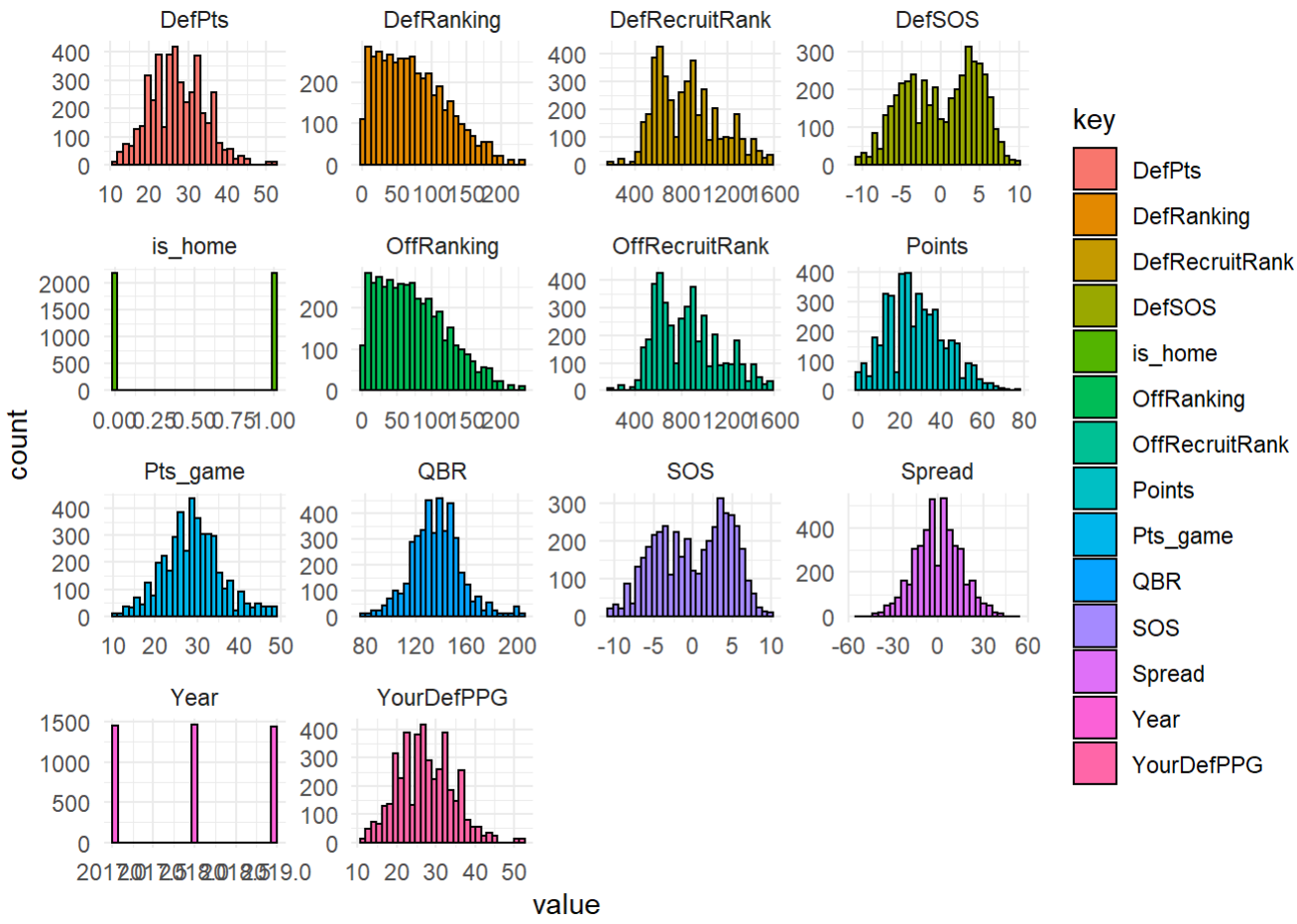
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
# view dataset
glimpse(Orig_Score_Data)
```

```
## Rows: 4,365
## Columns: 18
## $ GameID         <dbl> 401110723, 401114164, 401117854, 401114236, 40111165...
## $ Points         <dbl> 24, 45, 24, 42, 52, 41, 30, 12, 14, 28, 48, 0, 38, 3...
## $ is_home        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ Offense        <chr> "Florida", "Hawaii", "Cincinnati", "Tulane", "Clemso...
## $ Pts_game       <dbl> 33.2, 33.9, 29.6, 33.1, 43.9, 29.5, 24.8, 28.5, 28.5...
## $ QBR            <dbl> 156.1, 147.6, 123.7, 135.2, 166.7, 131.1, 149.2, 130...
## $ OffRecruitRank <dbl> 1276.06, 611.98, 833.80, 689.91, 1347.27, 1294.79, 1...
## $ OffRanking     <dbl> 10, 68, 30, 59, 3, 19, 37, 67, 108, 36, 135, 98, 60,...
## $ OffConf        <chr> "SEC", "Mountain West", "American Athletic", "Americ...
## $ SOS            <dbl> 2.91, -1.75, 2.51, 1.21, 2.70, 5.09, 2.34, -1.23, -4...
## $ YourDefPPG     <dbl> 15.5, 31.9, 20.6, 26.3, 13.5, 22.5, 22.4, 25.5, 23.0...
## $ Defense        <chr> "Miami", "Arizona", "UCLA", "Florida International",...
## $ DefPts         <dbl> 20.2, 35.8, 34.8, 27.2, 32.4, 32.6, 31.8, 15.0, 25.9...
## $ DefRecruitRank <dbl> 1198.88, 913.57, 1195.25, 704.12, 922.20, 599.65, 56...
## $ DefRanking     <dbl> 70, 83, 61, 118, 107, 152, 91, 20, 143, 78, 232, 9, ...
## $ DefSOS         <dbl> -0.18, 3.17, 4.95, -6.96, 2.42, -1.29, -3.73, 2.29, ...
## $ Spread         <dbl> -7.0, 10.5, -2.0, -3.0, -36.0, -33.5, -24.5, 6.0, -2...
## $ Year           <dbl> 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019...
```

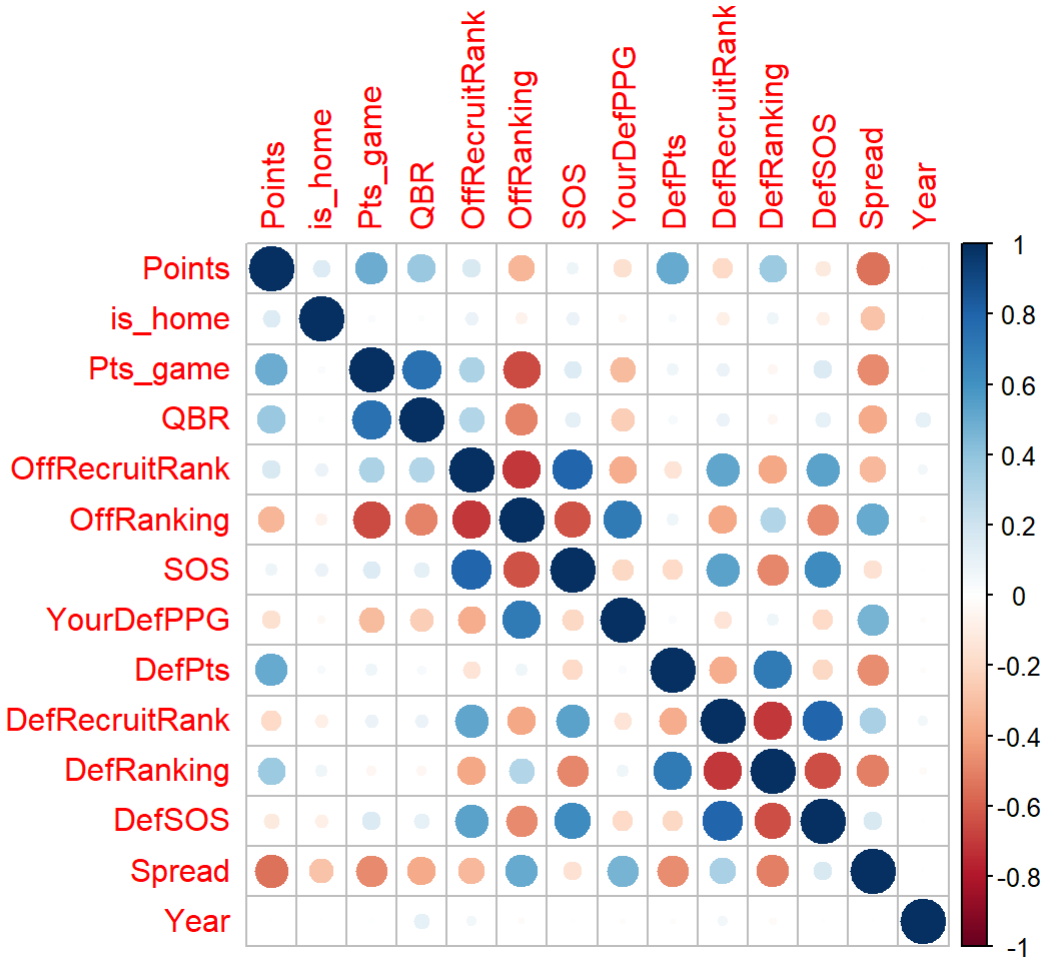
```
# remove non-predictor variables
Orig_Score_Data <- Orig_Score_Data %>%
  select(-GameID, -Offense, -Defense)


# View data distributions
Orig_Score_Data %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot() +
  geom_histogram(mapping = aes(x=value,fill=key), color="black") +
  facet_wrap(~ key, scales = "free") +
  theme_minimal()
```

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

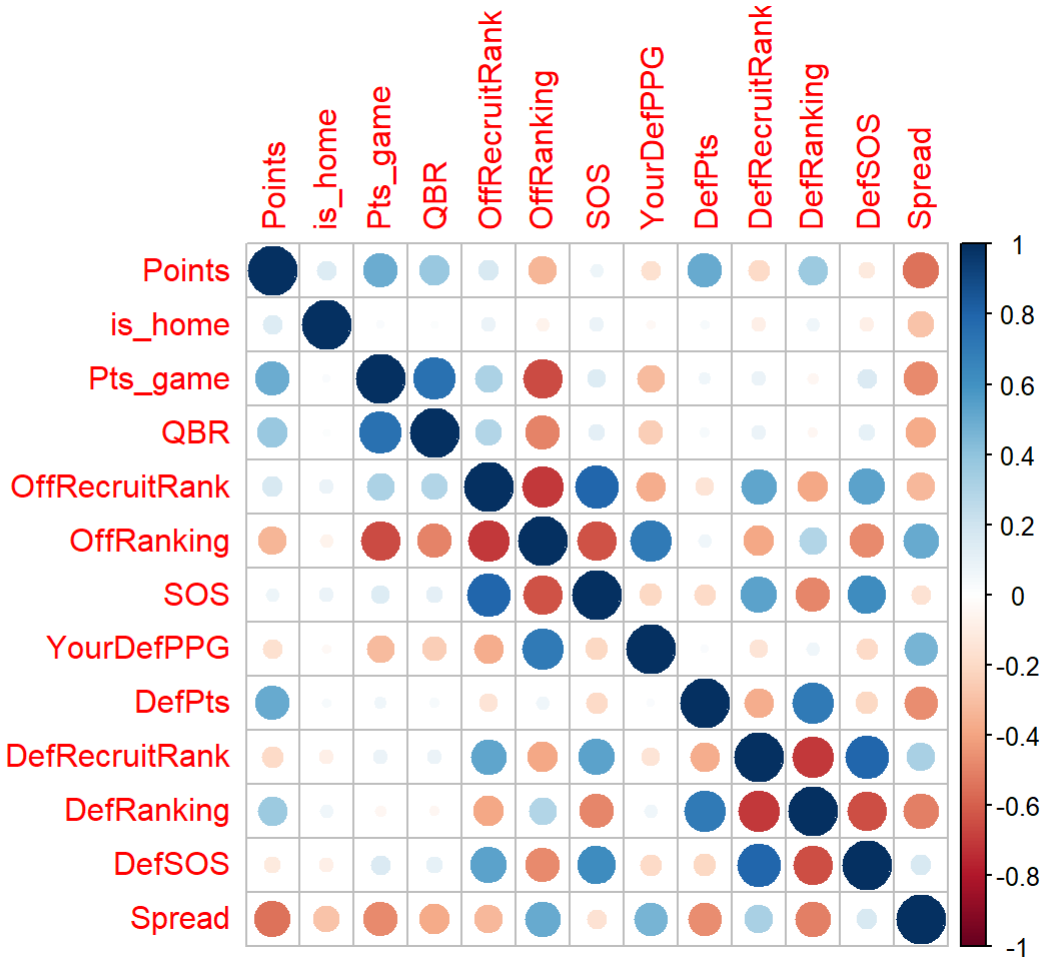



```
## Distributions look fairly normal. Some variables need to change datatypes.
```

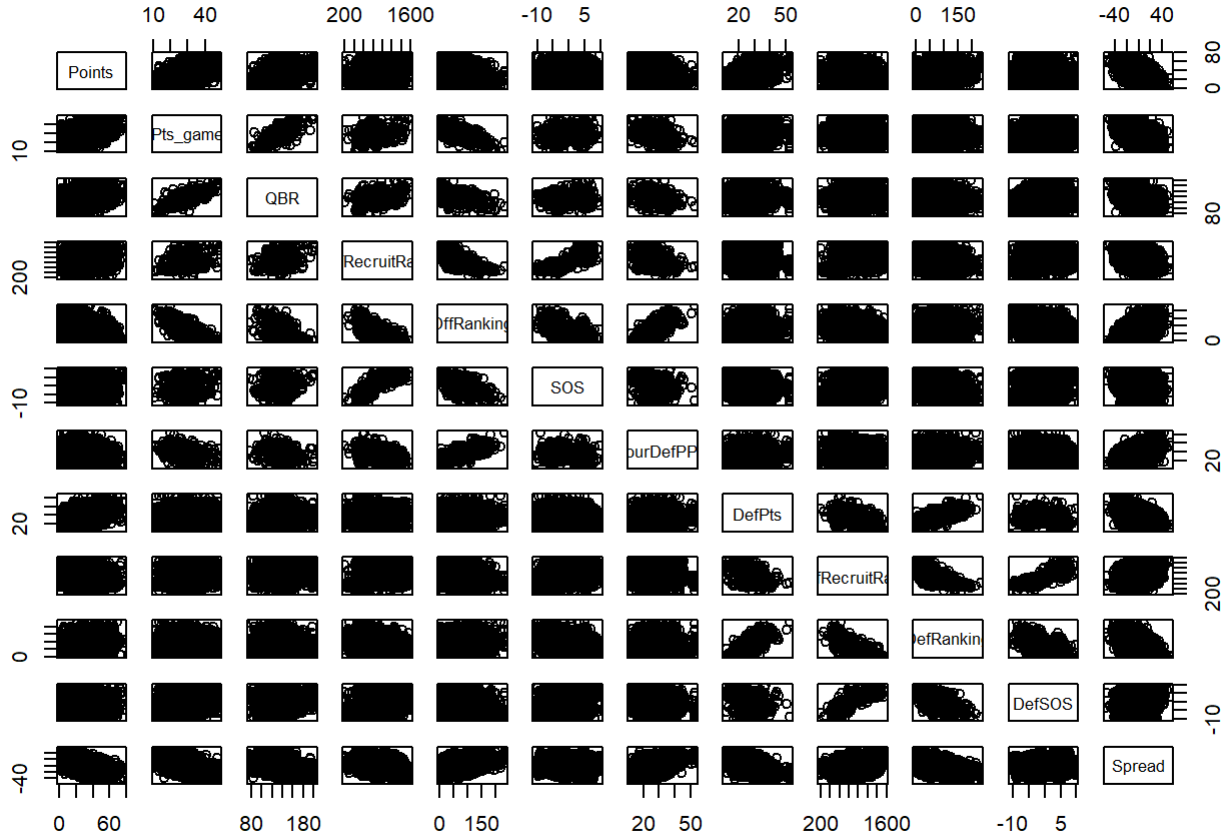
```
# View Corr plot
Orig_Score_Data %>%
  keep(is.numeric) %>%
  cor() %>%
  corrplot()
```




```
#Remove Year
Orig_Score_Data <- Orig_Score_Data %>%
  select(-Year)


# View Corr plot
Orig_Score_Data %>%
  keep(is.numeric) %>%
  cor() %>%
  corrplot()
```




```
# Convert Conference and is_Home to Factor
Orig_Score_Data$is_home <- as.factor(Orig_Score_Data$is_home)
Orig_Score_Data$OffConf <- as.factor(Orig_Score_Data$OffConf)


# View continuous variables in plot
Orig_Score_Data %>%
  keep(is.numeric) %>%
  plot()
```




```
# View in Regression Model
Orig_Score_Model <- lm(Orig_Score_Data$Points~., data = Orig_Score_Data)
summary(Orig_Score_Model)
```

```
##
## Call:
## lm(formula = Orig_Score_Data$Points ~ ., data = Orig_Score_Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.451  -6.770  -0.181   6.591  44.451
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3.145e+01  2.313e+00 -13.595  < 2e-16 ***
## is_home1                  2.882e+00  3.307e-01   8.713  < 2e-16 ***
## Pts_game                  9.433e-01  5.447e-02  17.317  < 2e-16 ***
## QBR                       2.474e-02  1.180e-02   2.096 0.036100 *
## OffRecruitRank            2.111e-03  1.232e-03   1.713 0.086707 .
## OffRanking               -1.723e-02  1.301e-02  -1.325 0.185390
## OffConfAmerican Athletic -2.348e+00  7.817e-01  -3.004 0.002680 **
## OffConfBig 12            -1.585e+00  7.312e-01  -2.167 0.030273 *
## OffConfBig Ten            1.151e+00  6.734e-01   1.710 0.087337 .
## OffConfConference USA     3.653e-02  9.861e-01   0.037 0.970448
## OffConfFBS Independents  -1.096e+00  9.870e-01  -1.111 0.266746
## OffConfMid-American       2.400e-01  9.330e-01   0.257 0.797014
## OffConfMountain West     -5.978e-01  8.539e-01  -0.700 0.483884
## OffConfPac-12            -5.320e-01  6.892e-01  -0.772 0.440245
## OffConfSEC               -1.598e-01  6.947e-01  -0.230 0.818123
## OffConfSun Belt          -1.388e+00  1.003e+00  -1.385 0.166257
## SOS                       6.812e-01  8.983e-02   7.583 4.09e-14 ***
## YourDefPPG               -2.116e-02  5.244e-02  -0.403 0.686614
## DefPts                    9.695e-01  3.588e-02  27.025  < 2e-16 ***
## DefRecruitRank           -7.757e-04  1.135e-03  -0.683 0.494453
## DefRanking                2.783e-02  7.861e-03   3.541 0.000403 ***
## DefSOS                   -7.056e-01  6.499e-02 -10.858  < 2e-16 ***
## Spread                    1.077e-01  2.485e-02   4.333 1.50e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.887 on 4342 degrees of freedom
## Multiple R-squared:  0.5307, Adjusted R-squared:  0.5283
## F-statistic: 223.2 on 22 and 4342 DF,  p-value: < 2.2e-16
```

```
# R-squared is .5307, F-stat = 223.2
# Variable Importance
imp <- varImp(Orig_Score_Model, scale=FALSE)
imp %>%
  arrange(desc(Overall))
```

```
##                              Overall
## DefPts                    27.02471186
## Pts_game                  17.31748983
## DefSOS                    10.85763539
## is_home1                   8.71307504
## SOS                        7.58339613
## Spread                     4.33327835
## DefRanking                 3.54055429
## OffConfAmerican Athletic   3.00392284
## OffConfBig 12              2.16721682
## QBR                        2.09644831
## OffRecruitRank             1.71341913
## OffConfBig Ten             1.71000347
## OffConfSun Belt            1.38456461
## OffRanking                 1.32454997
## OffConfFBS Independents    1.11072940
## OffConfPac-12              0.77185102
## OffConfMountain West       0.70012924
## DefRecruitRank             0.68330084
## YourDefPPG                 0.40348116
## OffConfMid-American        0.25722937
## OffConfSEC                 0.22997336
## OffConfConference USA      0.03704907
```

```
# Remove Insignificant Variables with high p-values

Orig_Score_Data <- Orig_Score_Data %>%
  select(-DefRecruitRank, -OffRanking, -OffRecruitRank, -OffRanking)


Orig_Score_Model <- lm(Orig_Score_Data$Points~., data = Orig_Score_Data)
summary(Orig_Score_Model)
```

```
##
## Call:
## lm(formula = Orig_Score_Data$Points ~ ., data = Orig_Score_Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.367  -6.793  -0.179   6.542  44.903
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -31.203803   1.731845 -18.018  < 2e-16 ***
## is_home1                   2.820718   0.328741   8.580  < 2e-16 ***
## Pts_game                   0.994427   0.037598  26.449  < 2e-16 ***
## QBR                        0.025972   0.011717   2.217 0.026701 *
## OffConfAmerican Athletic  -2.755648   0.738001  -3.734 0.000191 ***
## OffConfBig 12             -1.549938   0.725428  -2.137 0.032688 *
## OffConfBig Ten             1.004046   0.664976   1.510 0.131142
## OffConfConference USA     -0.553624   0.920479  -0.601 0.547571
## OffConfFBS Independents   -1.563055   0.947367  -1.650 0.099037 .
## OffConfMid-American       -0.378903   0.861284  -0.440 0.660011
## OffConfMountain West      -1.060521   0.805082  -1.317 0.187813
## OffConfPac-12             -0.347846   0.680392  -0.511 0.609207
## OffConfSEC                -0.011973   0.664566  -0.018 0.985627
## OffConfSun Belt           -2.085563   0.918851  -2.270 0.023272 *
## SOS                        0.784507   0.073865  10.621  < 2e-16 ***
## YourDefPPG                -0.079069   0.029275  -2.701 0.006941 **
## DefPts                     0.968867   0.035878  27.005  < 2e-16 ***
## DefRanking                 0.026010   0.007758   3.353 0.000807 ***
## DefSOS                    -0.725993   0.055033 -13.192  < 2e-16 ***
## Spread                     0.092086   0.022782   4.042 5.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.888 on 4345 degrees of freedom
## Multiple R-squared:  0.5302, Adjusted R-squared:  0.5281
## F-statistic: 258.1 on 19 and 4345 DF,  p-value: < 2.2e-16
```

```
# R-squared is .5281, F-stat = 258.1
## Let's see which variables add the most value to the model
imp <- varImp(Orig_Score_Model, scale=FALSE)
imp %>%
  arrange(desc(Overall))
```

```
##                             Overall
## DefPts                   27.00479569
## Pts_game                 26.44889783
## DefSOS                   13.19203763
## SOS                      10.62081094
## is_home1                  8.58035273
## Spread                    4.04214950
## OffConfAmerican Athletic  3.73393717
## DefRanking                3.35260891
## YourDefPPG                2.70092796
## OffConfSun Belt           2.26974996
## QBR                       2.21661042
## OffConfBig 12             2.13658401
## OffConfFBS Independents   1.64989272
## OffConfBig Ten            1.50989802
## OffConfMountain West      1.31728275
## OffConfConference USA     0.60145161
## OffConfPac-12             0.51124310
## OffConfMid-American       0.43992795
## OffConfSEC                0.01801607
```

```
# Remove OffConf and QBR
Orig_Score_Data <- Orig_Score_Data %>%
  select(-QBR, -OffConf)


Orig_Score_Model <- lm(Orig_Score_Data$Points~., data = Orig_Score_Data)
summary(Orig_Score_Model)
```

```
##
## Call:
## lm(formula = Orig_Score_Data$Points ~ ., data = Orig_Score_Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.345  -6.910  -0.272   6.630  44.828
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -27.118580   1.307189 -20.746  < 2e-16 ***
## is_home1      2.804329   0.329060   8.522  < 2e-16 ***
## Pts_game      1.007587   0.028847  34.929  < 2e-16 ***
## SOS           0.856795   0.051997  16.478  < 2e-16 ***
## YourDefPPG   -0.105322   0.028400  -3.709 0.000211 ***
## DefPts        0.921080   0.034355  26.810  < 2e-16 ***
## DefRanking    0.031194   0.007452   4.186 2.89e-05 ***
## DefSOS       -0.676552   0.053722 -12.594  < 2e-16 ***
## Spread        0.080912   0.022394   3.613 0.000306 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.919 on 4356 degrees of freedom
## Multiple R-squared:  0.5261, Adjusted R-squared:  0.5252
## F-statistic: 604.5 on 8 and 4356 DF,  p-value: < 2.2e-16
```

```
# R-squared is .5261, F-stat = 604.5
```

```
#Let's test the predictions
Score_Data <- Orig_Score_Data


#Split the dataset
test_data_train <- data.frame(Score_Data[1:3400,])
test_data_test <- data.frame(Score_Data[3401:4365,])

# Create Model
lm_train <- lm(test_data_train$Points~., data = test_data_train)

# Make predictions
lm_pred <- predict(lm_train, test_data_test)


#Convert to dataframe and merge for residuals
lm_pred <- data.frame(lm_pred)
lm_pred$Points <- test_data_test$Points
lm_pred <- na.omit(lm_pred)


#RMSE .5917962
sqrt(mean(lm_pred$lm_pred[1:965]-lm_pred$Points[1:965])^2)
```

```
## [1] 0.5917962
```

```
lm_pred <- lm_pred %>%
  mutate(diff=lm_pred-Points)


print("Mean:")
```

```
## [1] "Mean:"
```

```
mean(lm_pred$diff)
```

```
## [1] -0.5917962
```

```
print("Median:")
```

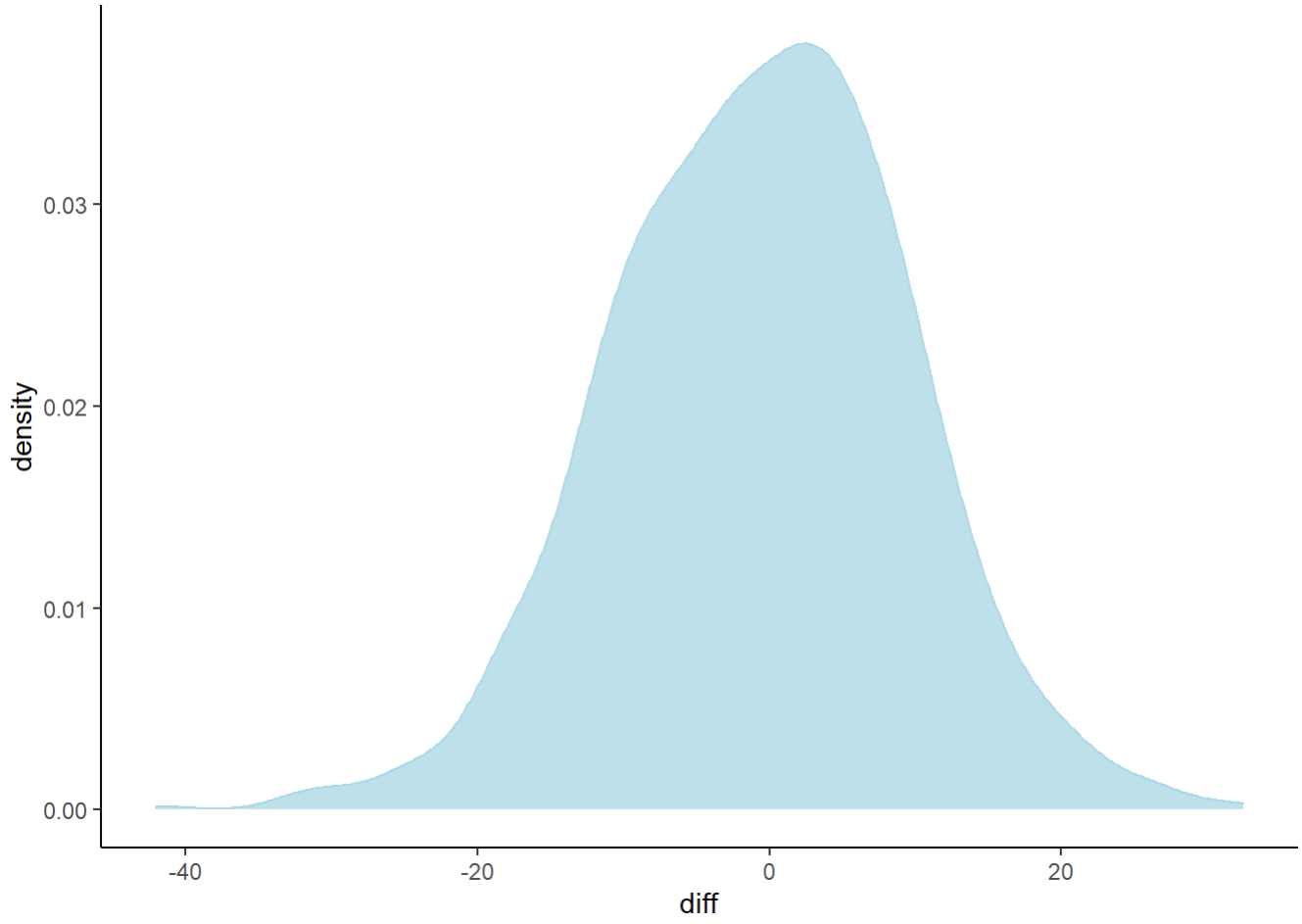
```
## [1] "Median:"
```

```
median(lm_pred$diff)
```

```
## [1] -0.2766613
```

```
# The mean/median predictions distribution are pretty close to zero. Looks like we're slightly
# underestimating the scores
```

```
ggplot(lm_pred, aes(x=diff)) +
  geom_density(alpha=0.8, color = 'lightblue', fill = 'lightblue') +
  theme_classic()
```




```
sd(lm_pred$diff)
```

```
## [1] 10.15588
```

```
 # 1 Standard Deviation = 10.15588 points
```

```
## Let's try a LASSO regression model

#Lasso
library(glmnet)
```

```
## Loading required package: Matrix
```

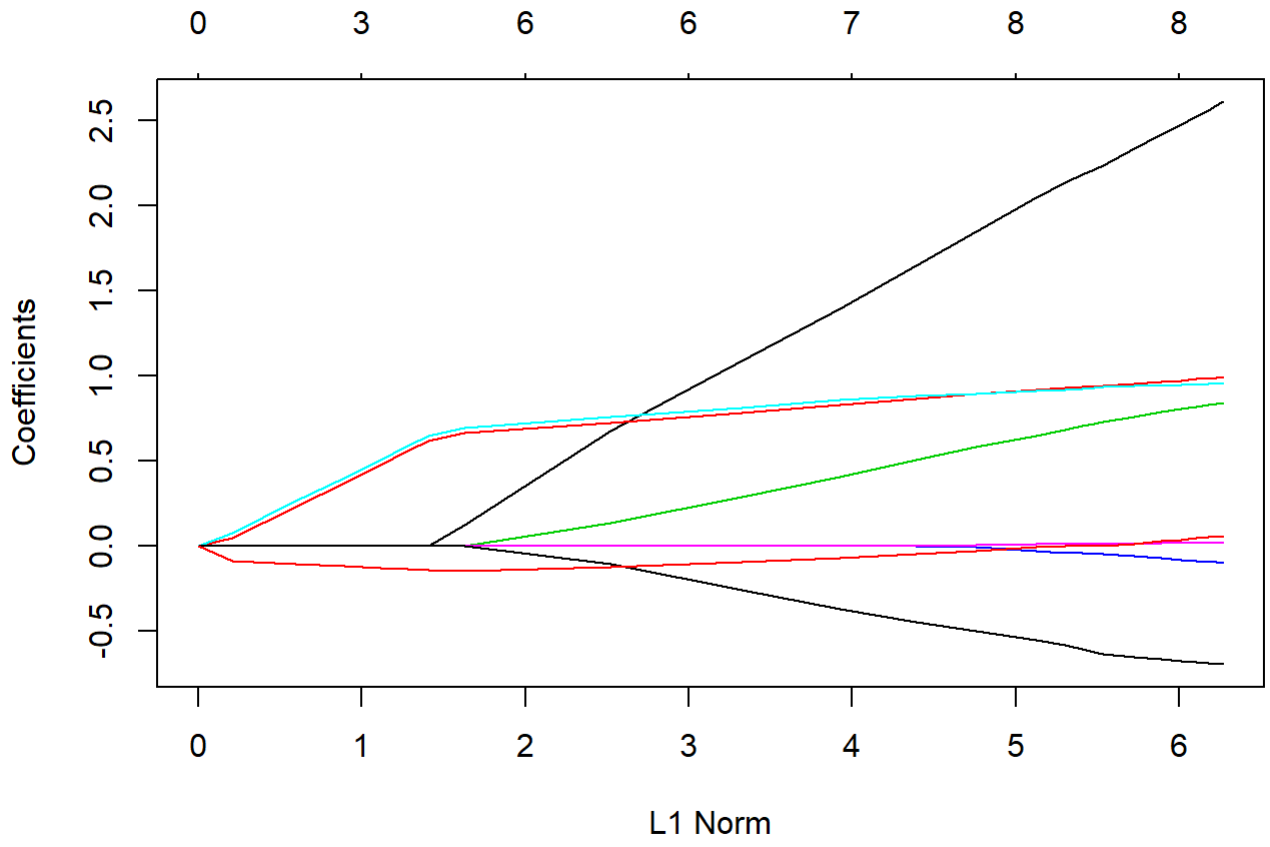
```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

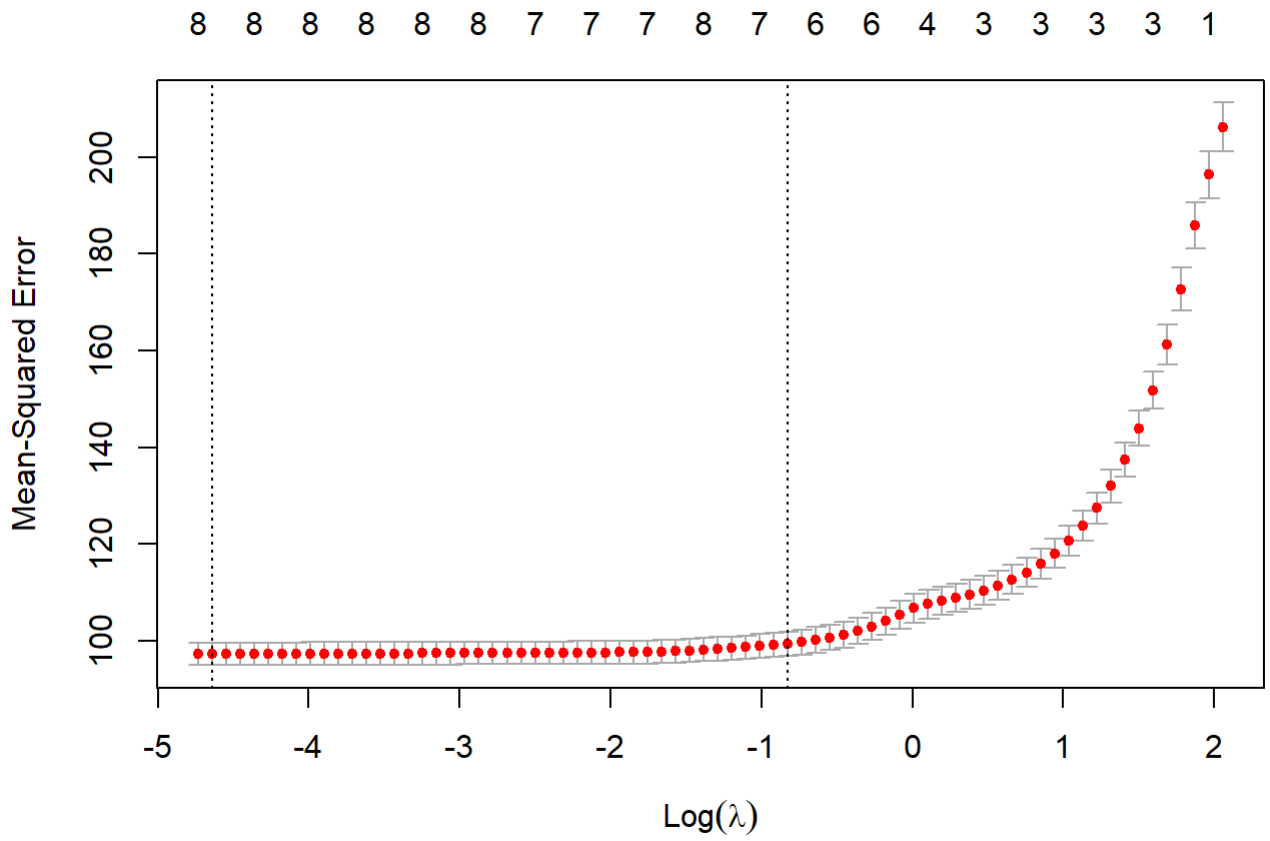
```
## Loaded glmnet 4.0-2
```

```
# Split target variable from features
x <- model.matrix(Score_Data$Points~.,Score_Data)[,-1]
y <- Score_Data$Points

lambda <- 10^seq(10, -2, length = 100)

#split test and training datasets
train=sample(x[1:3400,])
test=(x[3401:4365,])
y.test=y[3401:4365]

# Create Model
lasso.mod=glmnet(x[1:3400,],y[1:3400], alpha=1, lambda = lambda)
plot(lasso.mod)
```

```
## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to unique
## 'x' values
```




```
set.seed(1)
cv.out=cv.glmnet(x[1:3400,],y[1:3400],alpha=1)
plot(cv.out)
```




```
bestlam=cv.out$lambda.min
lasso.pred=predict(lasso.mod,s=bestlam,newx = x[3401:4365,])

# RMSE: .5843381, slightly worse than linear regression
sqrt(mean(lasso.pred-y.test)^2)
```

```
## [1] 0.5843381
```

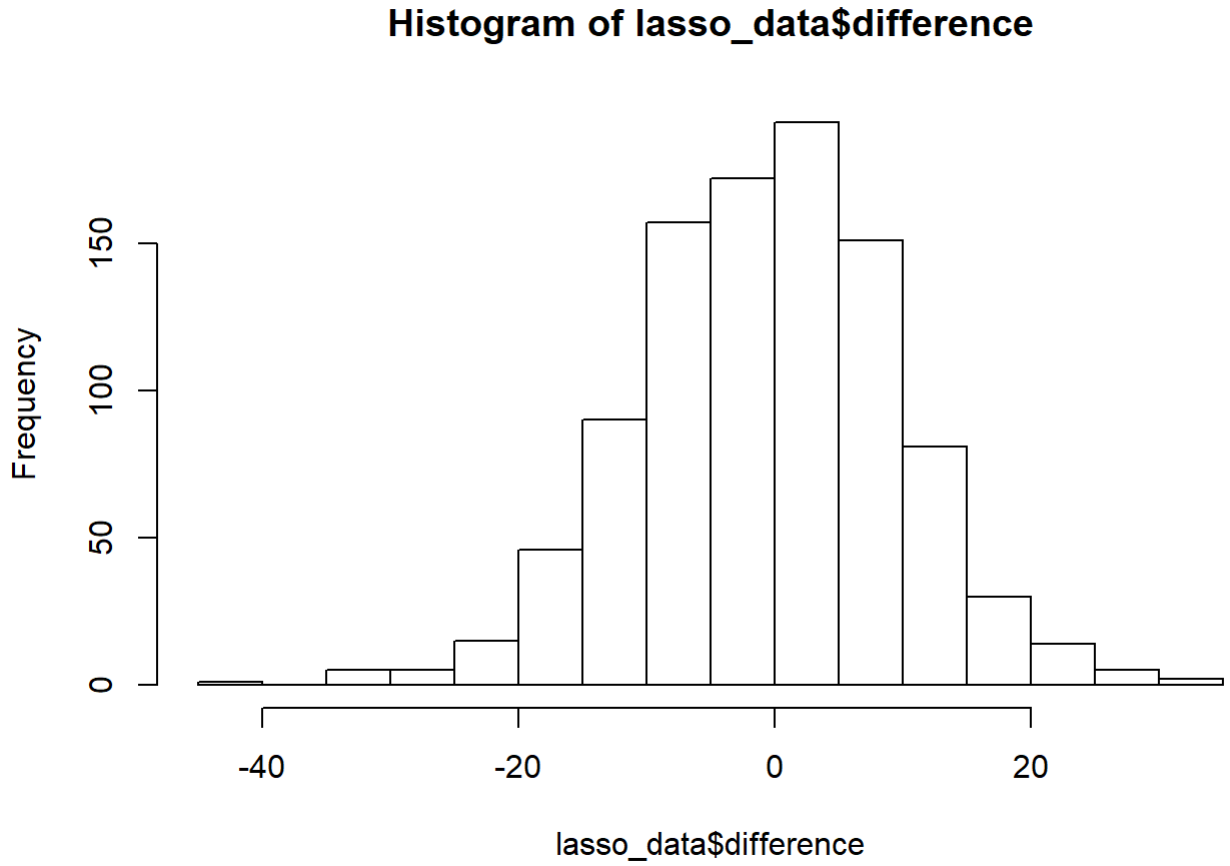
```
lasso_data <- data.frame(lasso.pred, y.test)

lasso_data <- lasso_data %>%
  mutate(difference=X1-y.test)

mean(lasso_data$difference)
```

```
## [1] -0.5843381
```

```
median(lasso_data$difference)
```

```
## [1] -0.2737702
```

```
hist(lasso_data$difference)
```




```
sd(lasso_data$difference)
```

```
## [1] 10.15783
```

```
# slightly wider distribution than the multiple regression model,  # 1 SD = 10.15783 points

out=glmnet(x,y,alpha = 1,lambda = lambda)
lasso.coef=predict(out,type = "coefficients", s=bestlam)

lasso.coef
```

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                        1
## (Intercept) -26.80725508
## is_home1      2.73286698
## Pts_game      0.99770338
## SOS           0.83650120
## YourDefPPG   -0.09594109
## DefPts        0.91781021
## DefRanking    0.02906424
## DefSOS       -0.66669825
## Spread        0.06968715
```