

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:

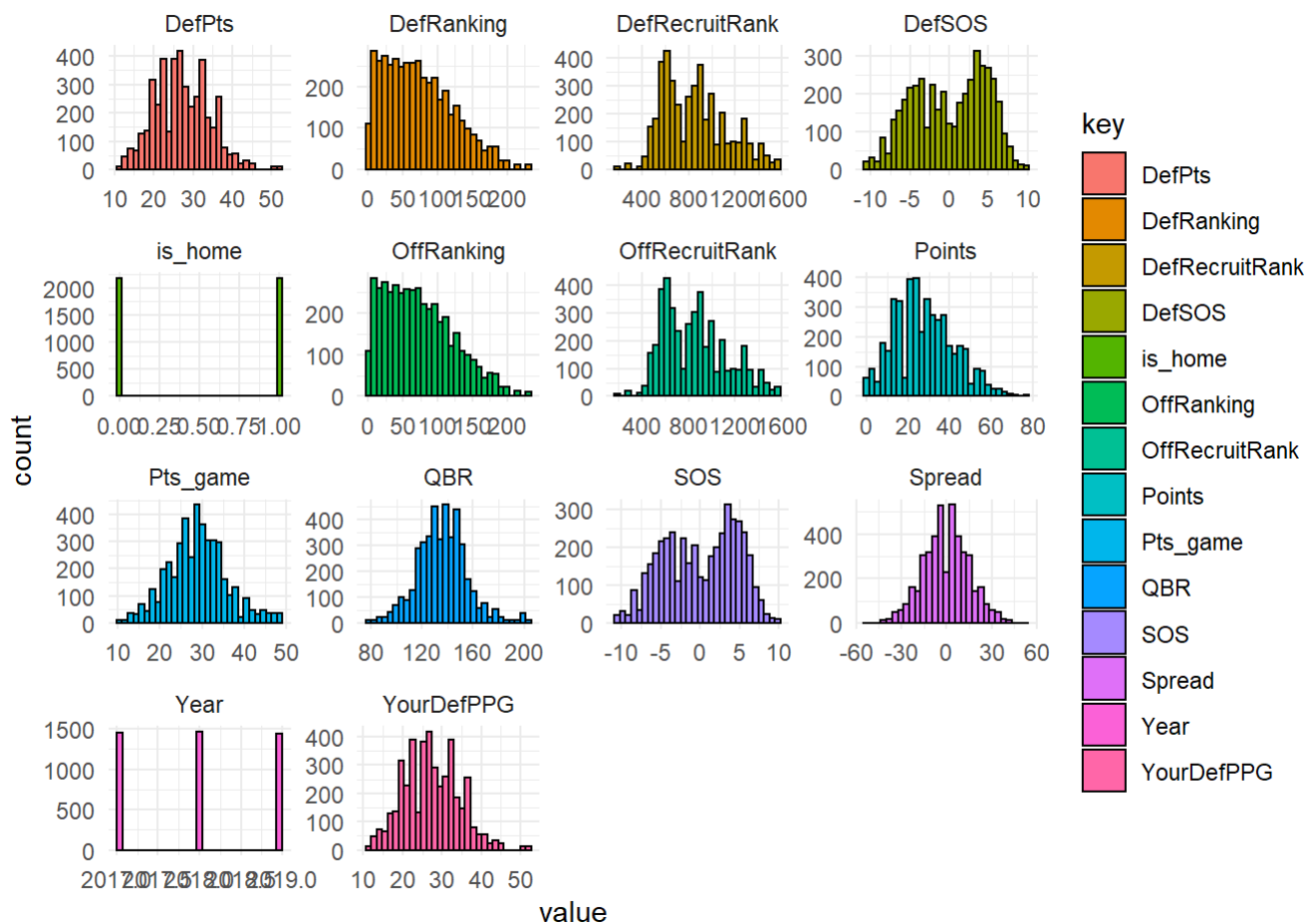
```
# view dataset
glimpse(Orig_Score_Data)
```

```
## 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...
## $ 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)

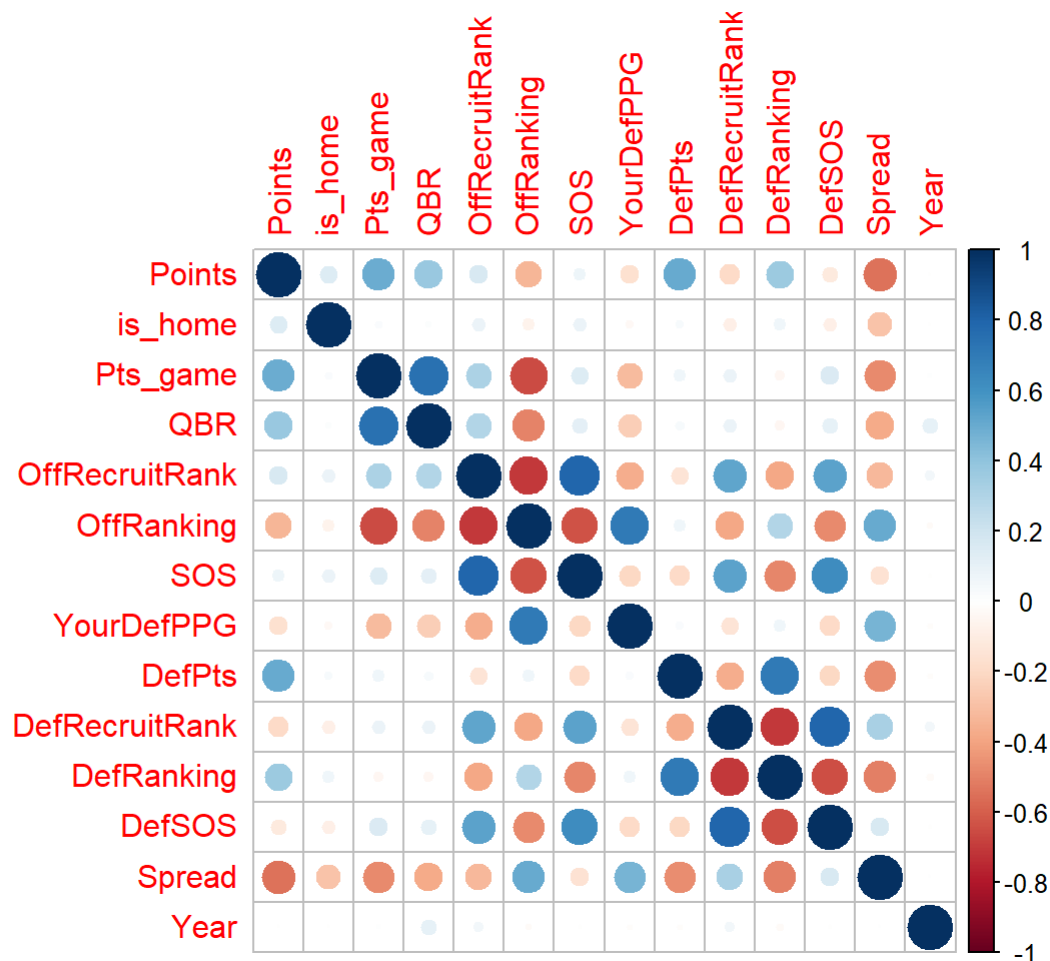
# View data distributions
Orig_Score_Data %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot() +
  geom_histogram(mapping = aes(x=value, fill=key), color="black") +
  facet_wrap(~ key, scales = "free") +
  theme_minimal()
```

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



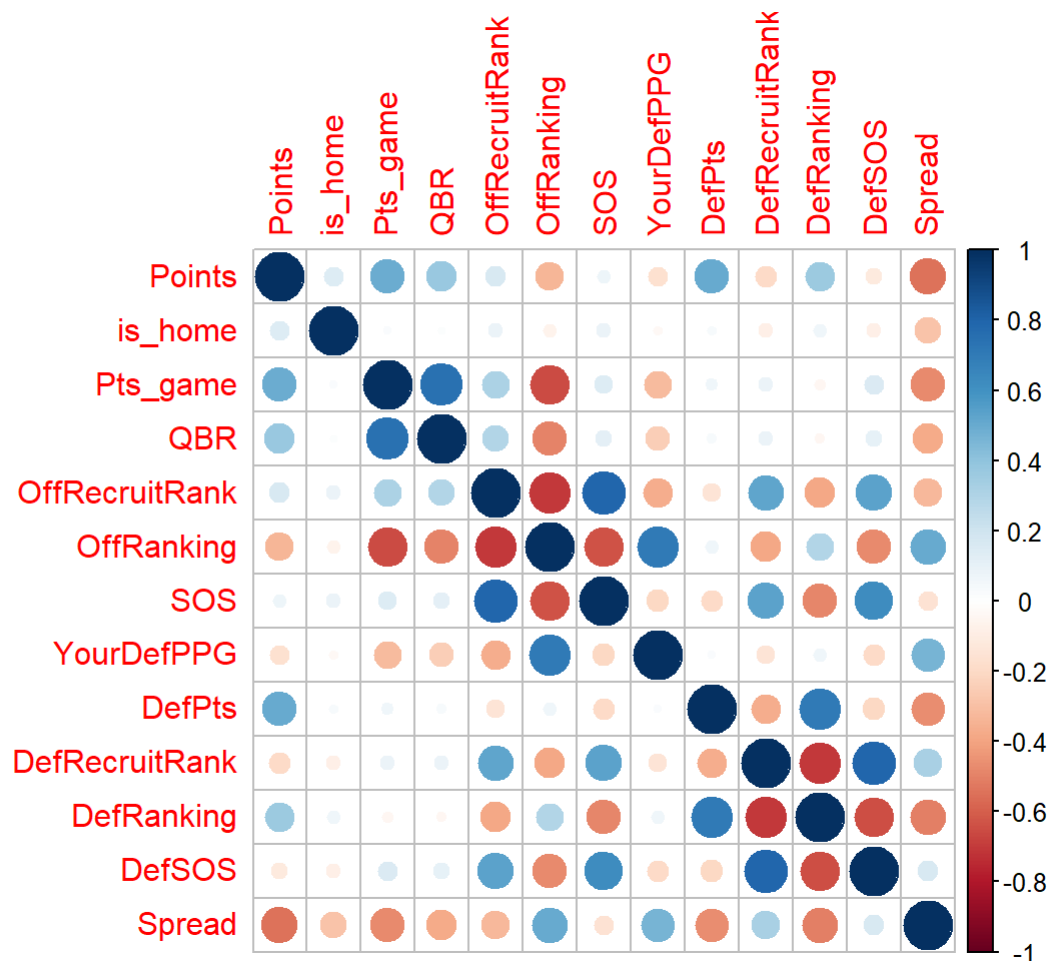
```
## Distributions look fairly normal. Some variables need to change.
```

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



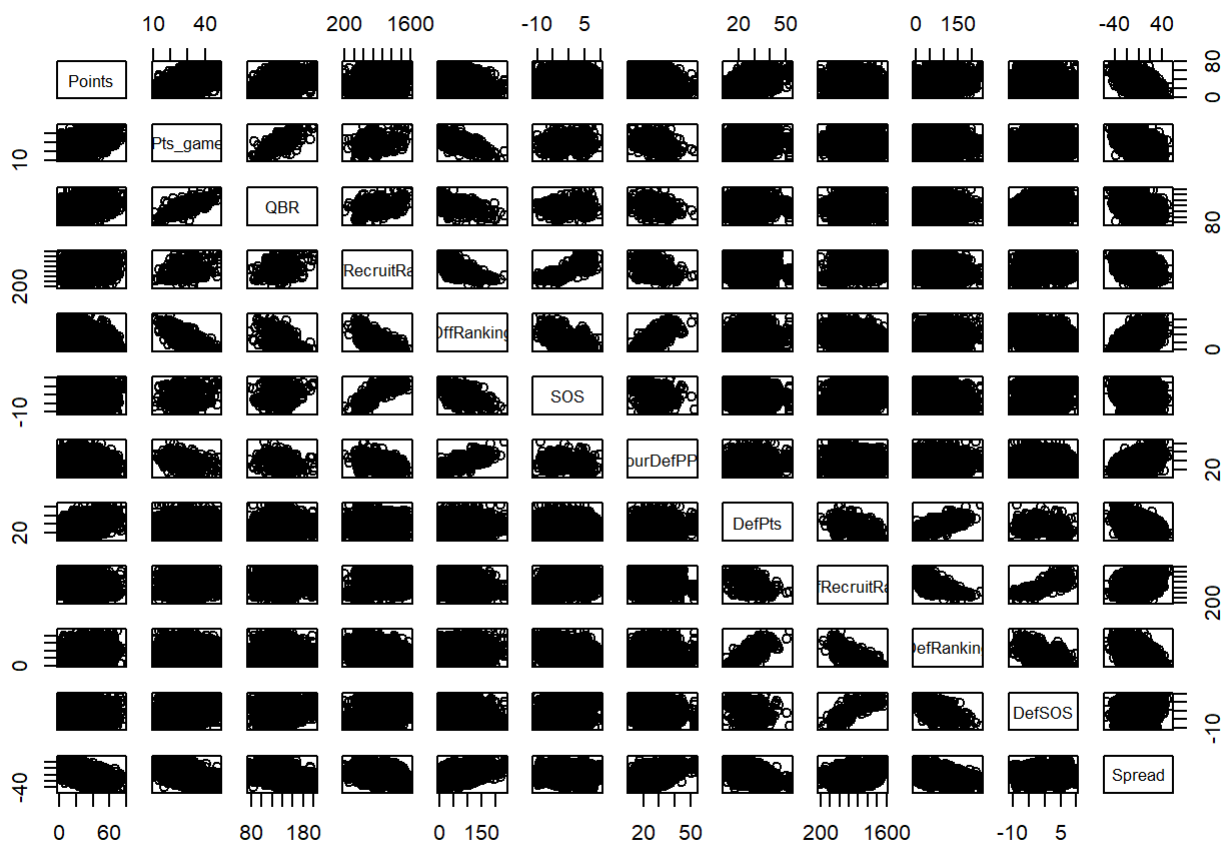
```
#Remove Year
Orig_Score_Data <- Orig_Score_Data %>%
  select(-Year)

# 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 .5316, F-stat = 196
# Variable Importance
imp <- varImp(Orig_Score_Model, scale=FALSE)
imp
```

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

```
# 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 .5315, F-stat = 233.5
## Pretty Decent, let's see which variables add the most value to the model
imp <- varImp(Orig_Score_Model, scale=FALSE)
imp
```



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

```
# Remove DFPI and QBR
```

```
Orig_Score_Data <- Orig_Score_Data %>%
  select(-QBR)
```

```
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
## -32.592  -6.784  -0.214   6.540  44.579
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -29.175802    1.471112  -19.832 < 2e-16 ***
## is_home1         2.803783    0.328801   8.527 < 2e-16 ***
## Pts_game        1.045860    0.029598  35.336 < 2e-16 ***
## OffConfAmerican Athletic -2.727245    0.738222  -3.694 0.000223 ***
## OffConfBig 12    -1.335766    0.719288  -1.857 0.063369 .
## OffConfBig Ten     0.958812    0.664962   1.442 0.149401
## OffConfConference USA -0.398112    0.918215  -0.434 0.664621
## OffConfFBS Independents -1.553601    0.947784  -1.639 0.101246
## OffConfMid-American -0.295729    0.860854  -0.344 0.731216
## OffConfMountain West -0.940827    0.803631  -1.171 0.241775
## OffConfPac-12     -0.118832    0.672805  -0.177 0.859814
## OffConfSEC         0.126589    0.661917   0.191 0.848342
## OffConfSun Belt   -1.893972    0.915189  -2.069 0.038559 *
## SOS              0.784702    0.073898  10.619 < 2e-16 ***
## YourDefPPG       -0.078544    0.029287  -2.682 0.007349 **
## DefPts           0.966735    0.035881  26.943 < 2e-16 ***
## DefRanking        0.025561    0.007759   3.294 0.000994 ***
## DefSOS           -0.725032    0.055056 -13.169 < 2e-16 ***
## Spread           0.088804    0.022744   3.905 9.58e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.893 on 4346 degrees of freedom
## Multiple R-squared:  0.5296, Adjusted R-squared:  0.5277
## F-statistic: 271.9 on 18 and 4346 DF, p-value: < 2.2e-16
```

```
# R-squared is .5307, F-stat = 258.6
```

```
#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 .6133257
sqrt(mean(lm_pred$lm_pred[1:965]-lm_pred$Points[1:965])^2)
```

```
## [1] 0.6133257
```

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

print("Mean:")
```

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

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

```
## [1] -0.6133257
```

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

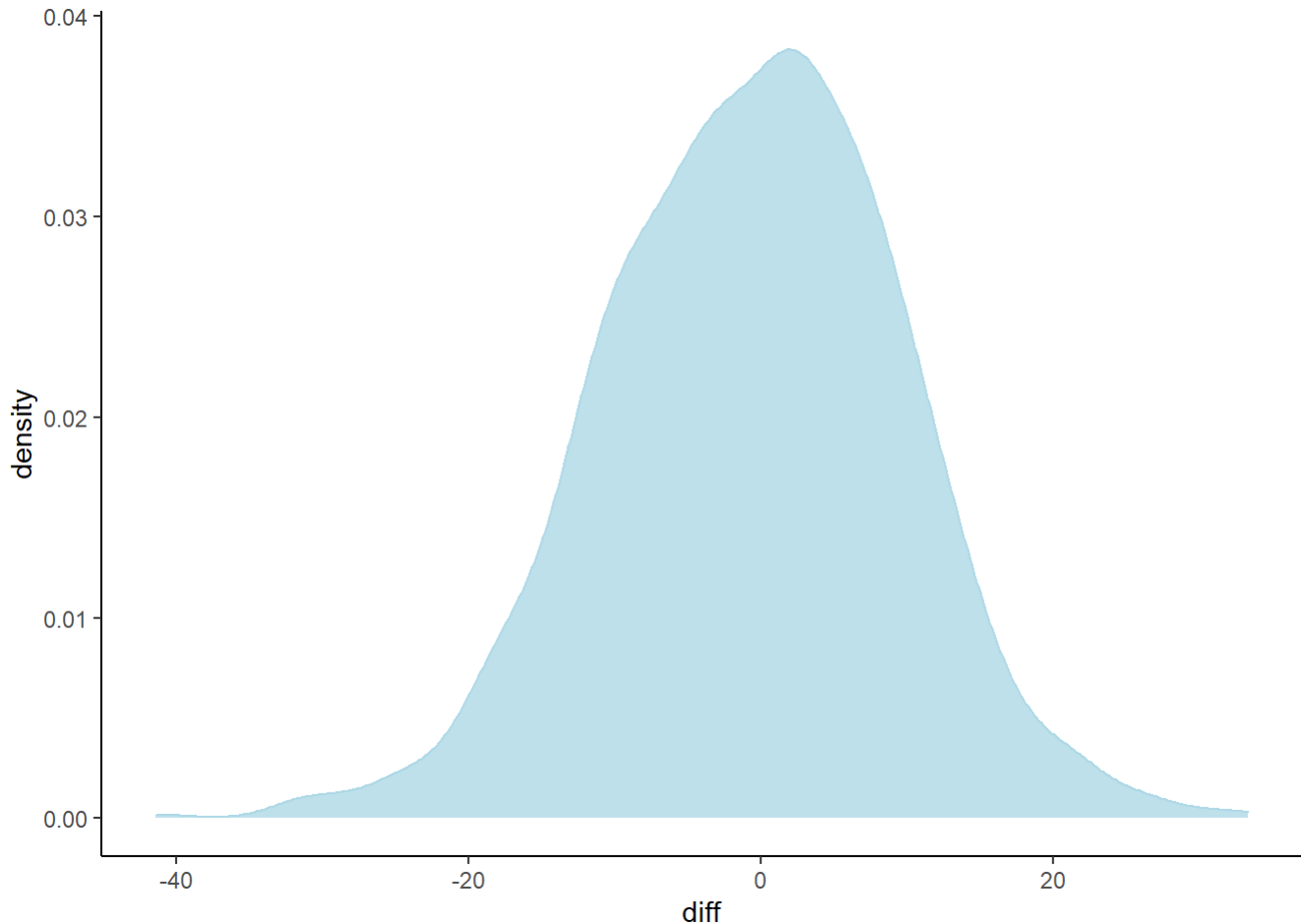
```
## [1] "Median:"
```

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

```
## [1] -0.1365034
```

```
# 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] 10.12732
```

```
# 1 Standard Deviation = 10.11913 points
```

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

```
#lasso  
library(glmnet)
```

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

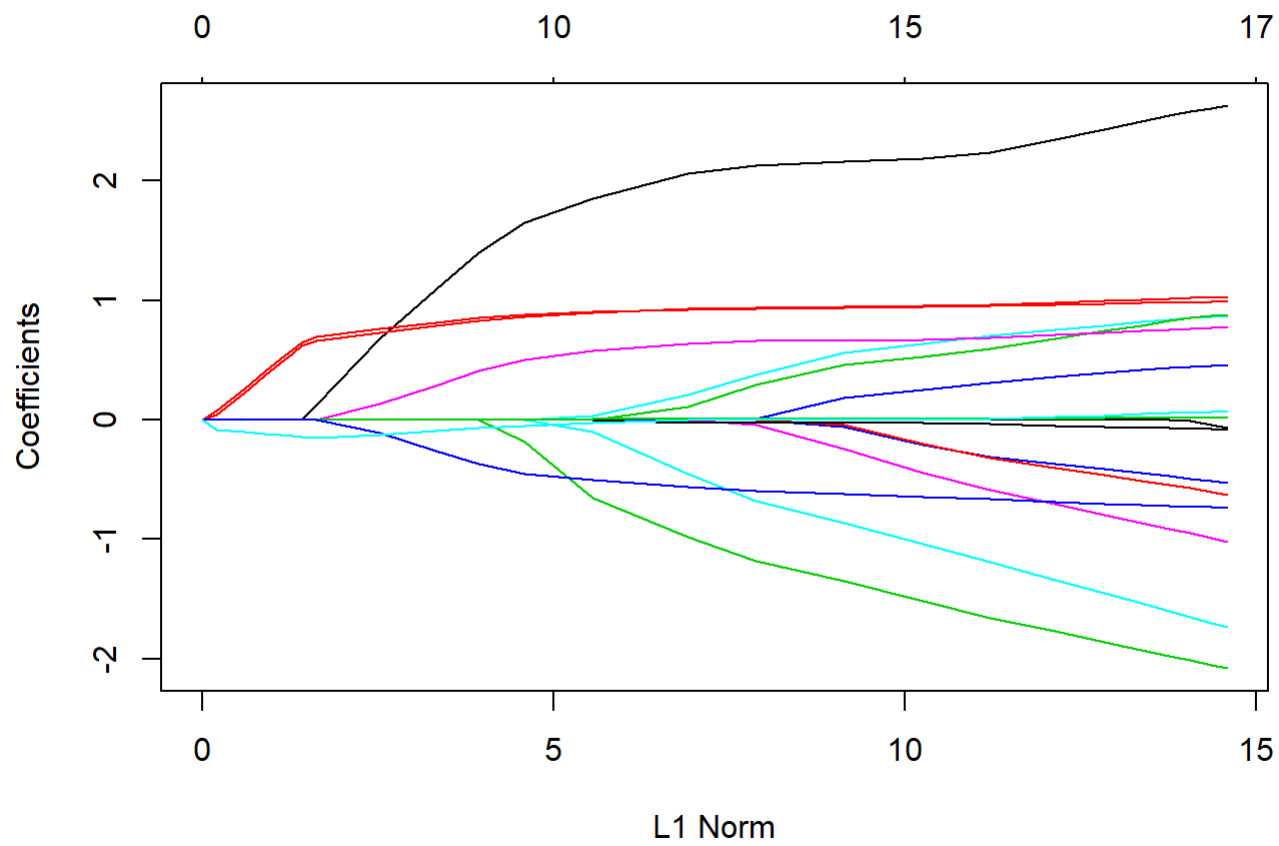
```
##  
## Attaching package: 'Matrix'
```

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

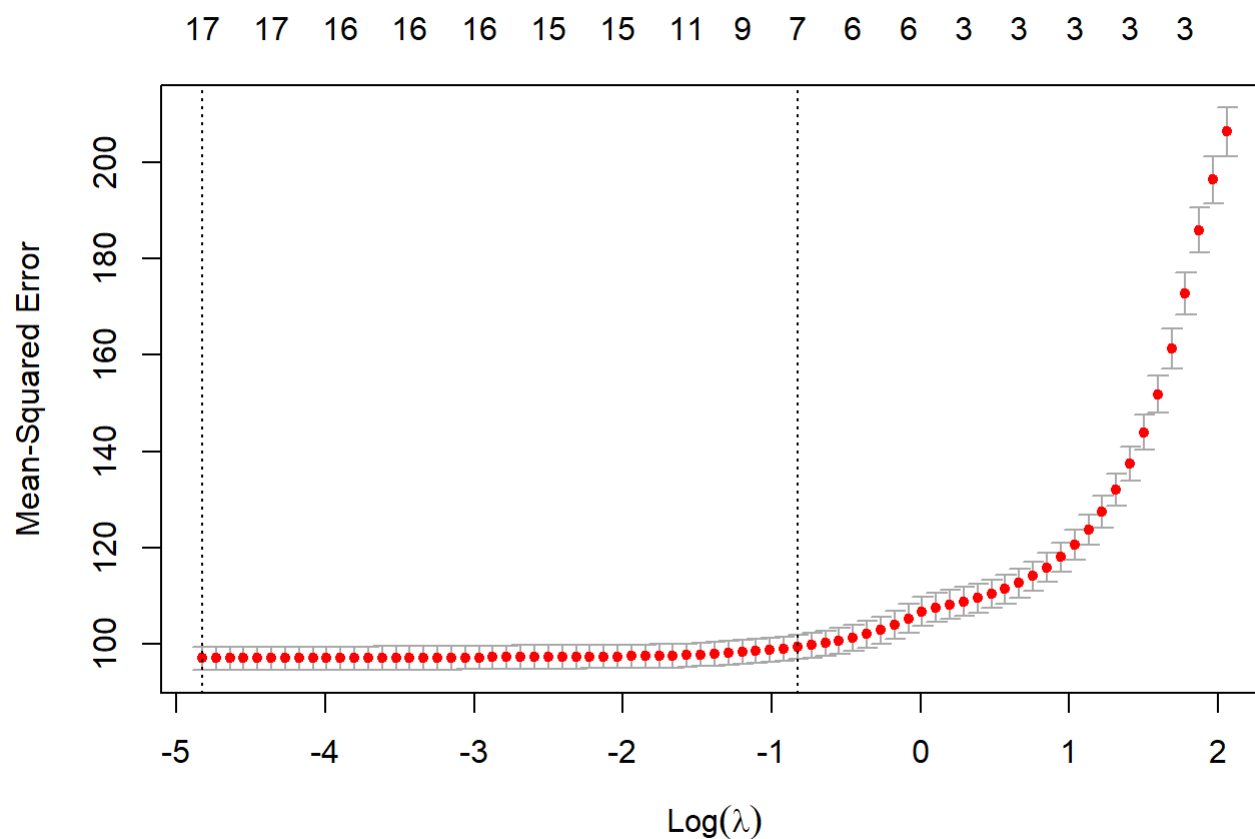
```
## Loaded glmnet 4.0-2
```

```
# 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: .7226794, slightly worse than linear regression
sqrt(mean(lasso.pred-y.test)^2)
```

```
## [1] 0.6037853
```

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

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

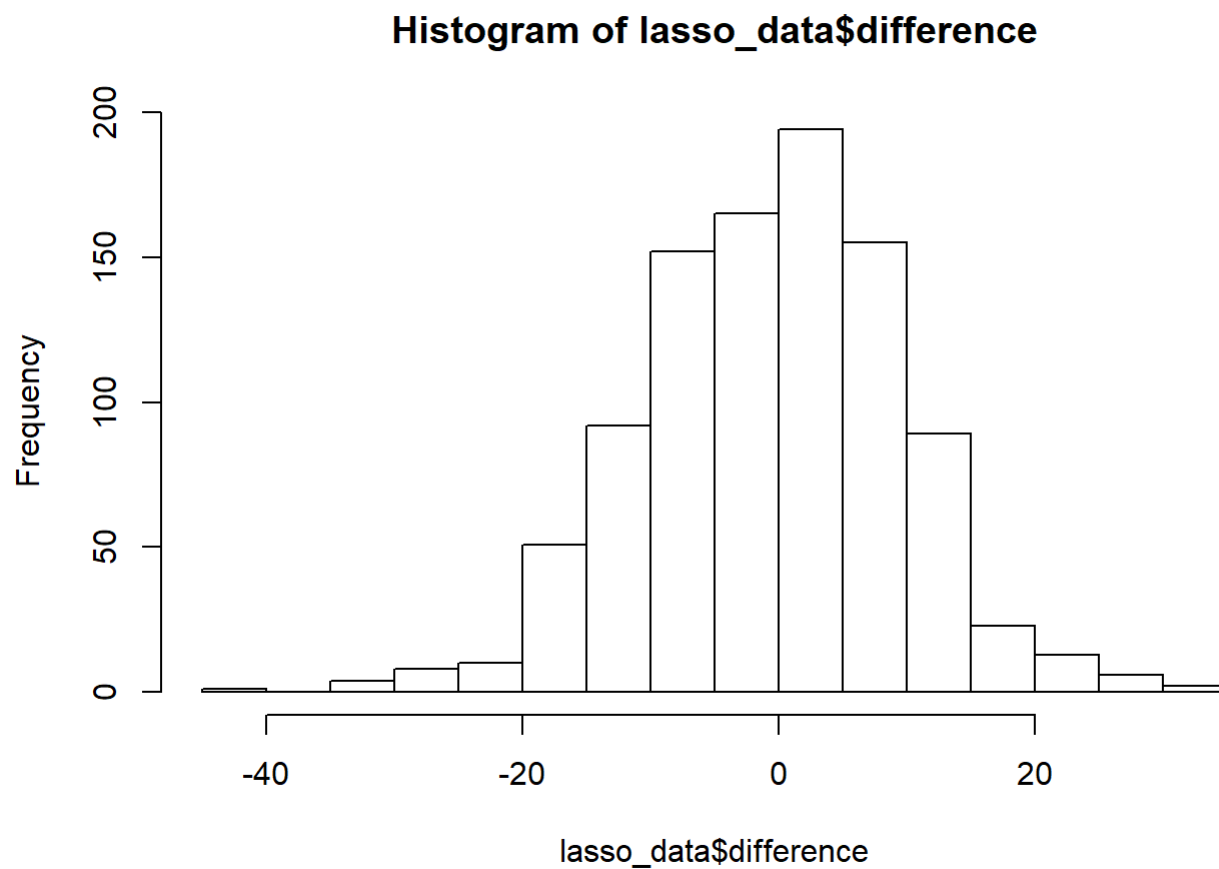
mean(lasso_data$difference)
```

```
## [1] -0.6037853
```

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

```
## [1] -0.0007145845
```

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



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

```
## [1] 10.12872
```

```
# slightly wider distribution than the multiple regression model, # 1 SD = 10.12273 points
```

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

```
lasso.coef
```



```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                                     1
## (Intercept)                      -28.99313348
## is_home1                          2.73018711
## Pts_game                          1.03447869
## OffConfAmerican Athletic         -2.47168019
## OffConfBig 12                    -1.18840327
## OffConfBig Ten                     1.02409020
## OffConfConference USA             -0.05819172
## OffConfFBS Independents          -1.26069382
## OffConfMid-American               .
## OffConfMountain West             -0.66298814
## OffConfPac-12                     .
## OffConfSEC                        0.20497770
## OffConfSun Belt                  -1.54649013
## SOS                              0.78148181
## YourDefPPG                       -0.06888523
## DefPts                           0.96312766
## DefRanking                        0.02320190
## DefSOS                           -0.71108079
## Spread                           0.07612391
```