eda

Jess, Anna and Seth

December 6, 2019

Jess, Anna and Seth Project EDA

12/6/19

```
source('styleguide.R')
```

```
## Loading required package: optimx
## Loading required package: parallel
## Loading required package: minqa
## Loading required package: lme4
## Loading required package: Matrix
## Loading required package: segmented
## Loading required package: dplyr
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
## Loading required package: ggplot2
## Loading required package: hrbrthemes
## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use these themes.
##       Please use hrbrthemes::import_roboto_condensed() to install Roboto Condensed and
##       if Arial Narrow is not on your system, please see http://bit.ly/arialnarrow
## Loading required package: ggcorrplot
```

```
source('helpers.R')
source('packages.R')
source('cleaner.R')
# https://cran.r-project.org/web/packages/segmented/segmented.pdf
```

Read in Clean DF

```
df.clean <- add_time("complete_data_clean.csv")
df.tourney <- add_time("tourney_data_clean.csv")
names(df.tourney)
```

```
##  [1] "School" "G"      "TeamW"  "TeamL"  "W.L."   "SRS"    "SOS"
##  [8] "Tm."    "Opp."   "FTr"    "X3PAr"  "TS."    "TRB."   "AST."
## [15] "BLK."   "eFG."   "TOV."   "FT.FGA" "FG"     "FGA"    "FG."
## [22] "X3P"    "X3PA"   "X3P."   "FT"     "FTA"    "FT."    "ORB"
## [29] "TRB"    "AST"    "STL"    "BLK"    "TOV"    "PF"     "year"
## [36] "time"
```

```
# Check dimensions - len(unique schools) * len(unique years) must equal # of rows
dim_checker(df.clean)
```

```
## [1] "Dim Check Successful"
```

```
dim_checker(df.tourney)
```

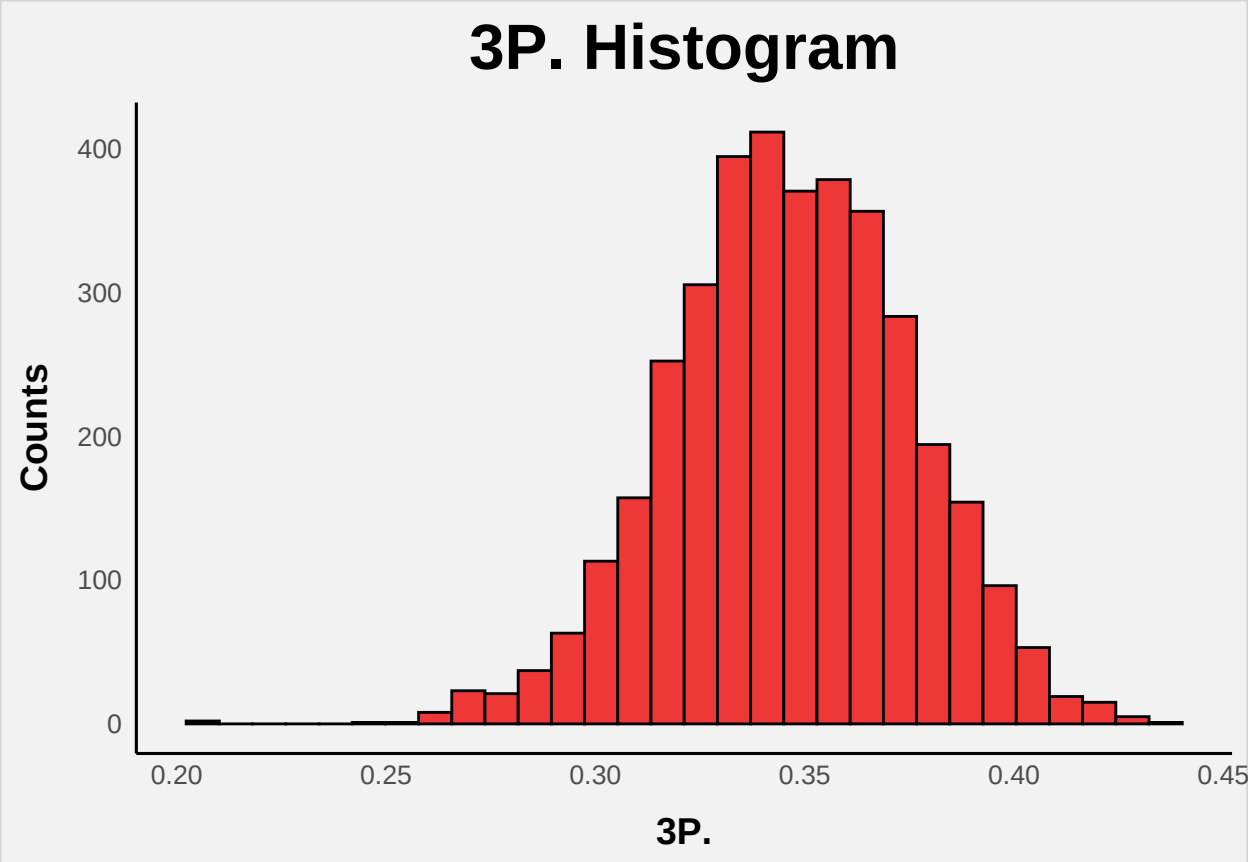
```
## [1] "Dim Check Successful"
```

```
# Per game-ify
get_newprop = cbind(df.tourney$School, get_prop_df(df.tourney))
```

Check assumption of normal distribution

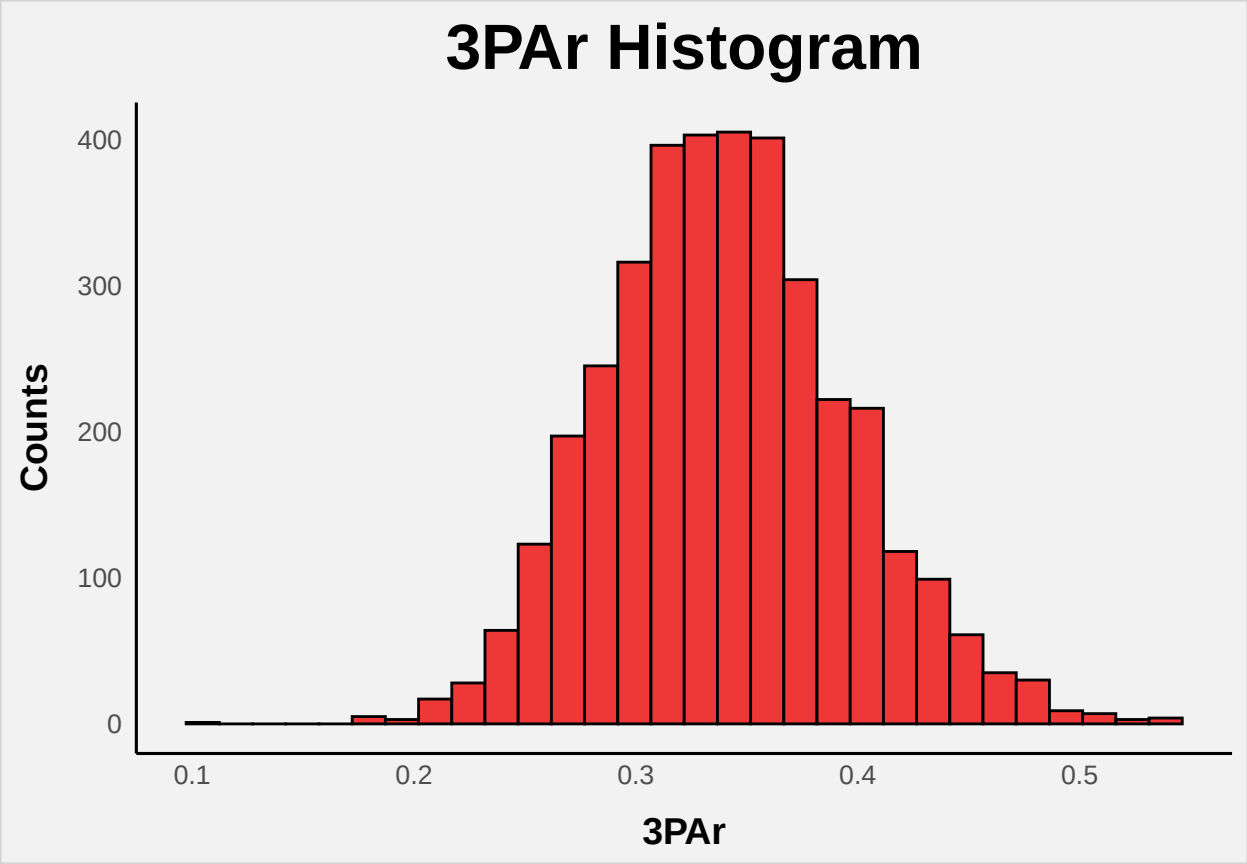
```
p <- ggplot(get_newprop, aes(x=X3P.)) +
  geom_histogram(colour="black", fill='#EE3838') +
  labs(title="3P. Histogram") +
  xlab("3P.") +
  ylab("Counts") +
  theme_hodp()
p
```

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

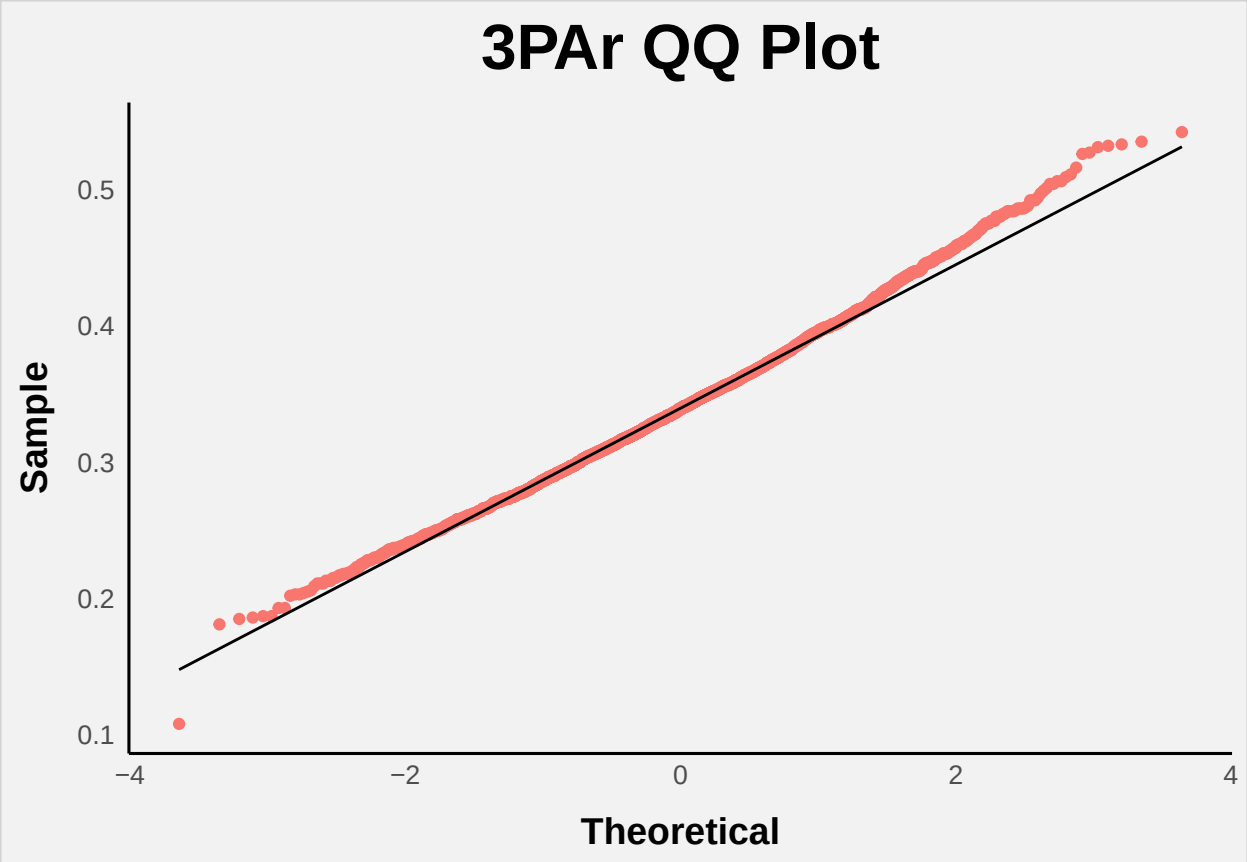



```
#### EDA ####

#### histograms ####

# Let's have X3PAr be our response
# Check assumption of normal distribution
p <- ggplot(df.tourney, aes(x=X3PAr)) +
  geom_histogram(colour="black", fill='#EE3838') +
  labs(title="3PAr Histogram") +
  xlab("3PAr") +
  ylab("Counts") +
  theme_hodp()
p
```

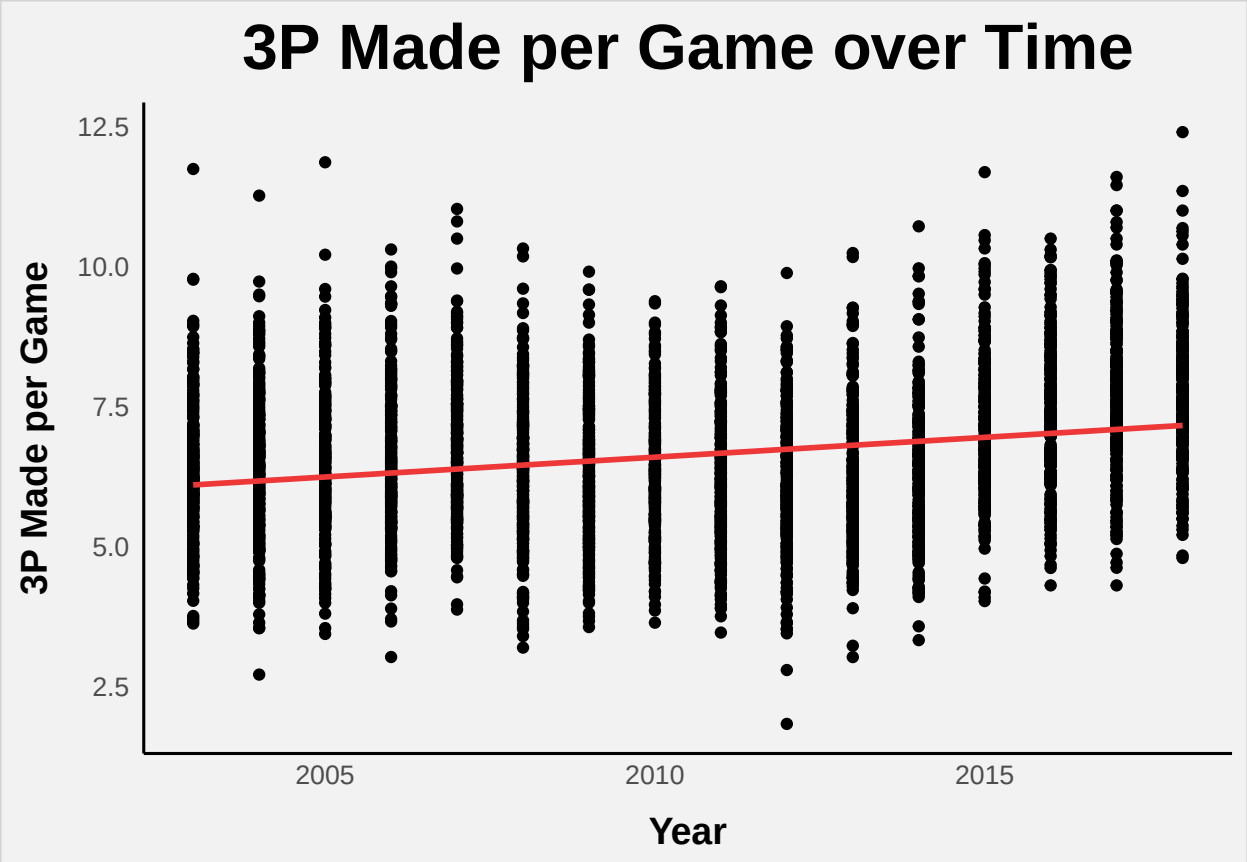
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



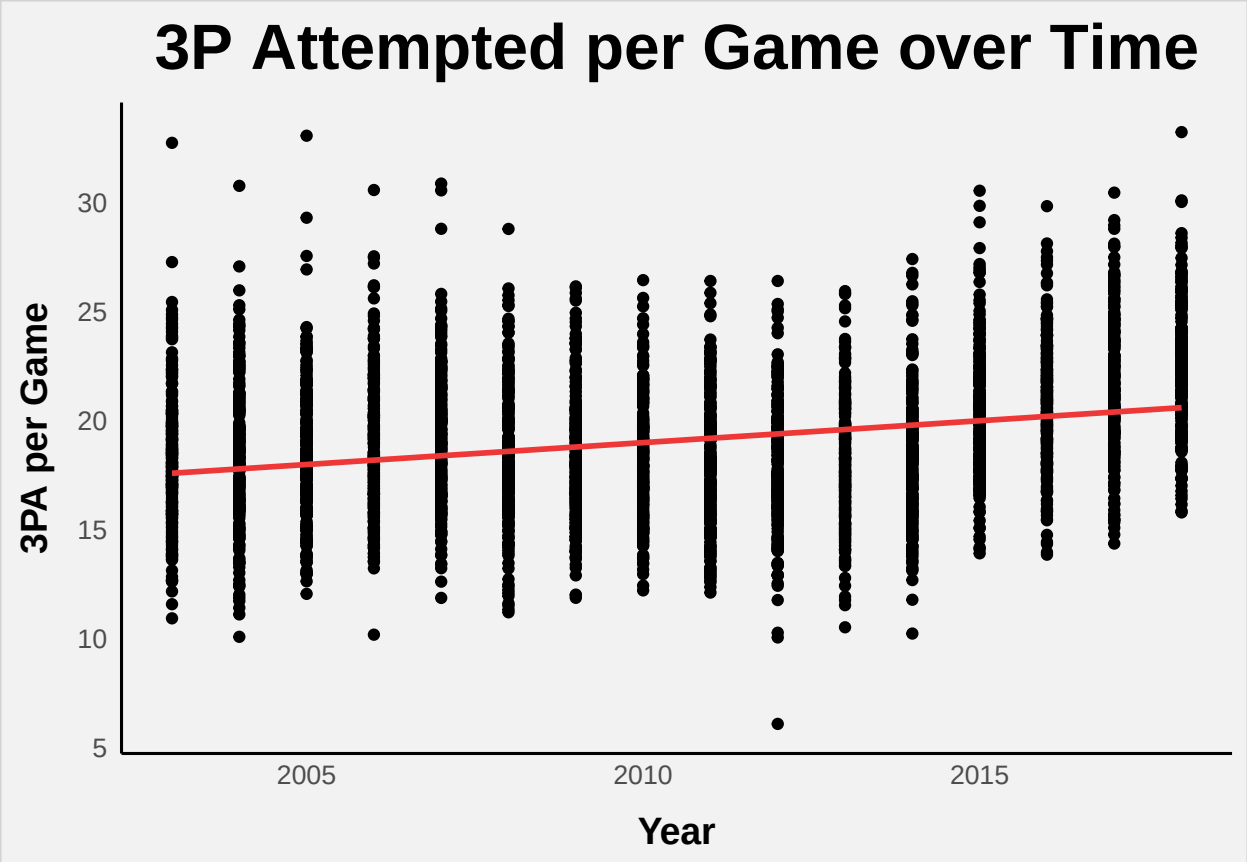

```
# QQ plot
p <- ggplot(df.tourney, aes(sample = X3PAr)) +
  stat_qq(aes(color = '#EE3838')) +
  stat_qq_line() +
  labs(title="3PAr QQ Plot") +
  xlab("Theoretical") +
  ylab("Sample") +
  theme_hodp()+
  theme(legend.position = "none")
p
```



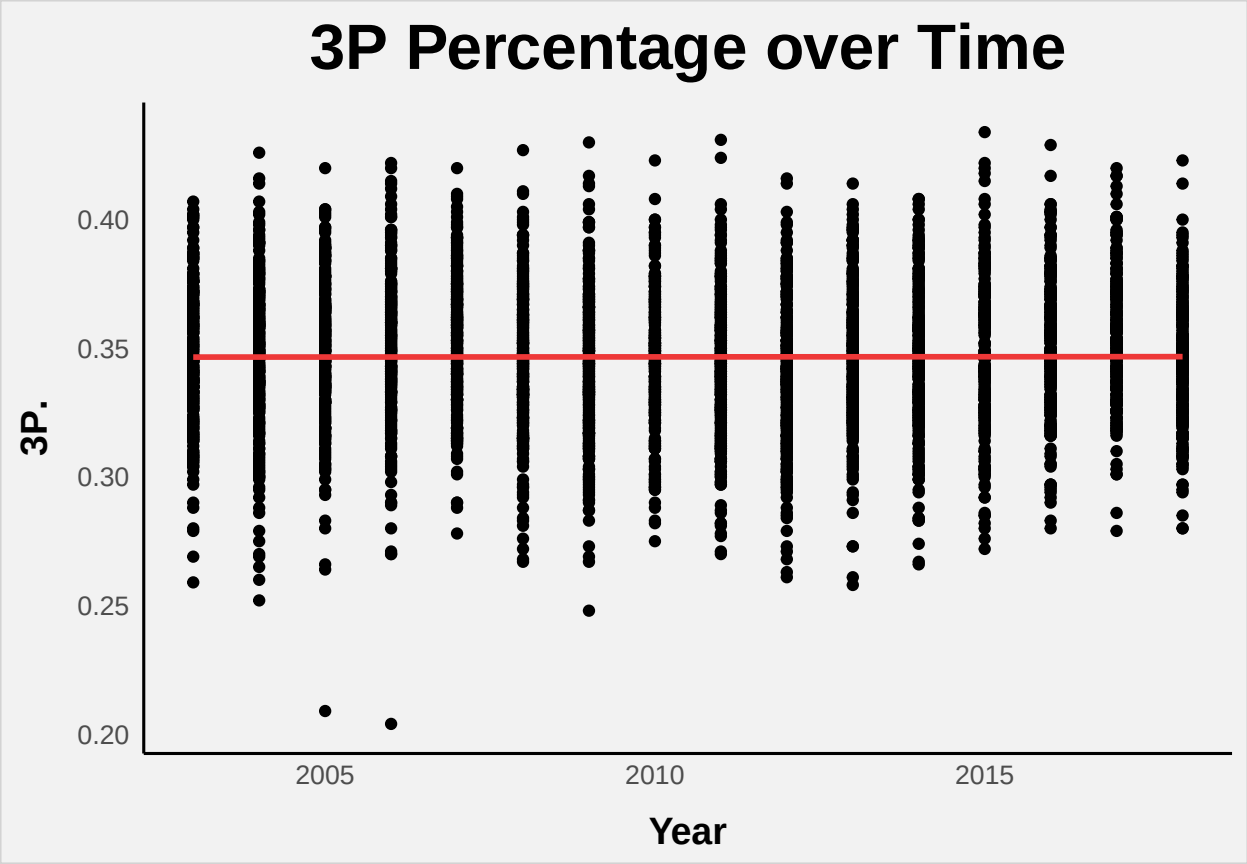

```
# X3P hist
p <- ggplot(get_newprop, aes(x = time + 2003, y = X3P)) +
  geom_point() +
  stat_smooth(method = "lm", col = '#EE3838', se = F) +
  labs(title="3P Made per Game over Time") +
  xlab("Year") +
  ylab("3P Made per Game") +
  #ylim(c(0, 0.6)) +
  theme_hodp()
p
```




```
# 3PA Hist
p <- ggplot(get_newprop, aes(x = time + 2003, y = X3PA)) +
  geom_point() +
  stat_smooth(method = "lm", col = '#EE3838', se = F) +
  labs(title="3P Attempted per Game over Time") +
  xlab("Year") +
  ylab("3PA per Game") +
  #ylim(c(0,0.6)) +
  theme_hodp()
p
```




```
# 3P percentage Hist
p <- ggplot(get_newprop, aes(x = time + 2003, y = X3P.)) +
  geom_point() +
  stat_smooth(method = "lm", col = '#EE3838', se = F) +
  labs(title="3P Percentage over Time") +
  xlab("Year") +
  ylab("3P.") +
  #ylim(c(0,0.6)) +
  theme_hodp()
p
```



Games Increasing

```
# since we know that games are increasing can we make those statistics into
# proportions to control for the specific effect
get_newprop = cbind(df.tourney$School, get_prop_df(df.tourney))
head(get_newprop)
```

```
##    df.tourney$School     TeamW     TeamL      Tm.      Opp.       FG
## 1          Air Force 0.7586207 0.2413793 59.89655 50.86207 20.17241
## 2              Akron 0.4642857 0.5357143 71.75000 72.14286 24.85714
## 3        Alabama A&M 0.4333333 0.5666667 70.43333 74.00000 24.03333
## 4 Alabama-Birmingham 0.6875000 0.3125000 75.43750 69.31250 27.56250
## 5      Alabama State 0.5161290 0.4838710 61.41935 61.70968 20.61290
## 6            Alabama 0.6060606 0.3939394 72.12121 68.03030 24.63636
##        FGA      X3P     X3PA       FT      FTA      ORB      TRB      AST
## 1 41.93103 8.448276 22.31034 11.10345 15.62069  6.00000 21.48276 13.31034
## 2 53.28571 6.178571 17.03571 15.85714 24.57143 11.21429 33.57143 15.10714
## 3 58.73333 6.200000 19.53333 16.16667 23.50000 12.90000 37.06667 12.60000
## 4 62.68750 6.656250 20.18750 13.65625 20.50000 12.53125 34.96875 17.62500
## 5 49.12903 5.580645 16.67742 14.61290 22.29032 11.61290 34.38710 10.16129
## 6 54.60606 7.060606 18.66667 15.78788 22.27273 11.09091 34.78788 11.96970
##         STL      BLK      TOV       PF year time  W.L.    SRS    SOS   FTr
## 1  7.758621 2.344828 10.96552 16.58621 2003    0 0.759   9.12   0.08 0.373
## 2  7.392857 2.535714 14.64286 19.42857 2003    0 0.464  -0.39   0.00 0.461
## 3  7.333333 2.600000 16.23333 21.66667 2003    0 0.433 -15.88 -12.29 0.400
## 4 11.593750 3.593750 13.46875 21.28125 2003    0 0.688  11.76   5.63 0.327
```

```
## 5  7.064516 3.064516 17.32258 18.74194 2003    0 0.516 -11.10 -10.78 0.454
## 6  6.454545 3.363636 12.81818 18.12121 2003    0 0.606  14.29  10.20 0.408
##   X3PAr   TS. TRB. AST. BLK.  eFG. TOV. FT.FGA   FG.  X3P.   FT. id
## 1 0.532 0.607 43.4 66.0  5.7 0.582 18.2  0.265 0.481 0.379 0.711  1
## 2 0.320 0.552 48.6 60.8  4.3 0.524 18.4  0.298 0.466 0.363 0.645  2
## 3 0.333 0.504 48.8 52.4  4.5 0.462 18.8  0.275 0.409 0.317 0.688  3
## 4 0.322 0.521 46.6 63.9  6.5 0.493 15.7  0.218 0.440 0.330 0.666  4
## 5 0.339 0.514 51.9 49.3  5.8 0.476 22.5  0.297 0.420 0.335 0.656  5
## 6 0.342 0.553 50.4 48.6  5.9 0.516 16.4  0.289 0.451 0.378 0.709  6
```

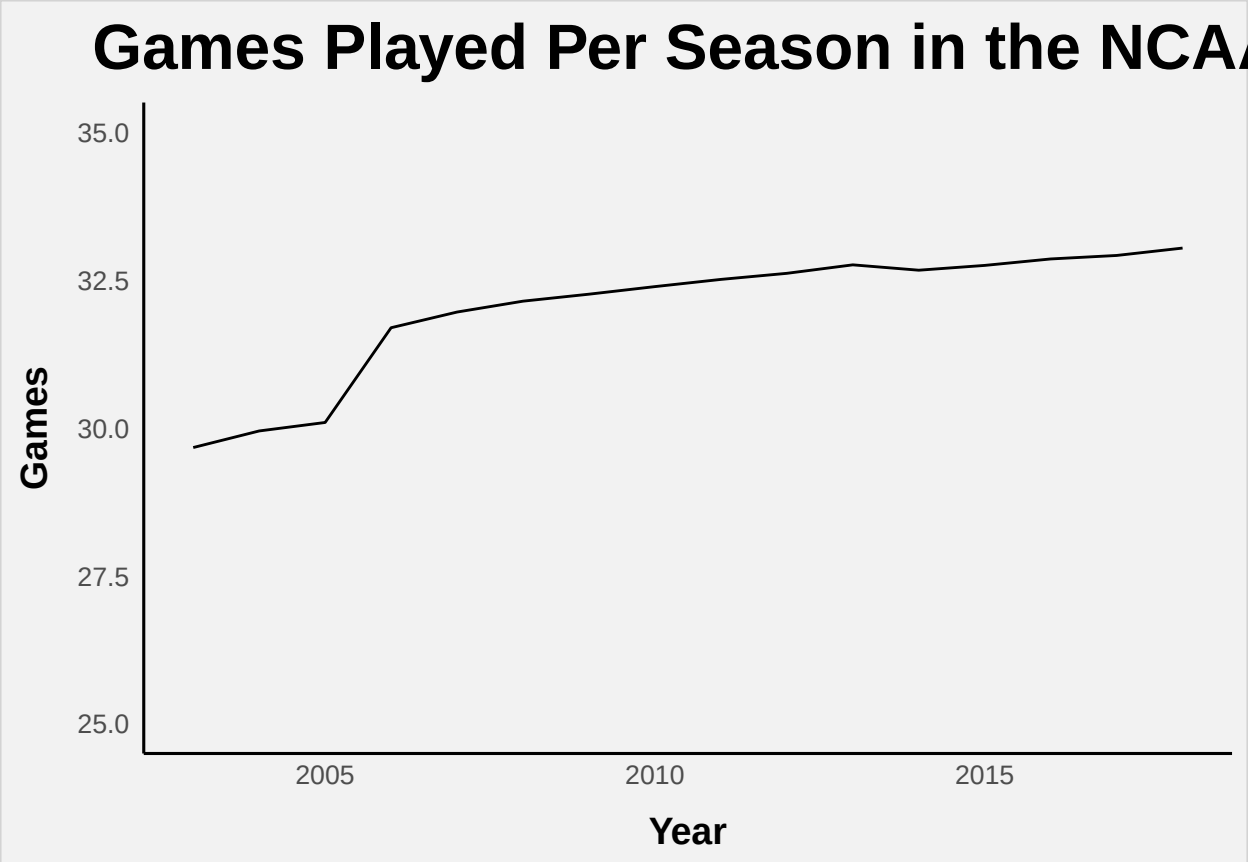
```
df.clean.noschool = df.clean[,2:length(df.clean)]
top_cor_list = cor(df.clean.noschool)[,ncol(df.clean.noschool)-1]
top_cor_list = sort(top_cor_list, decreasing = TRUE)
top_cor_list = top_cor_list[3:length(top_cor_list)]
head(top_cor_list)
```

```
##      Opp.      X3PA       FGA         G       X3P       Tm.
## 0.4650990 0.4003998 0.3902803 0.3778744 0.3437789 0.3405097
```

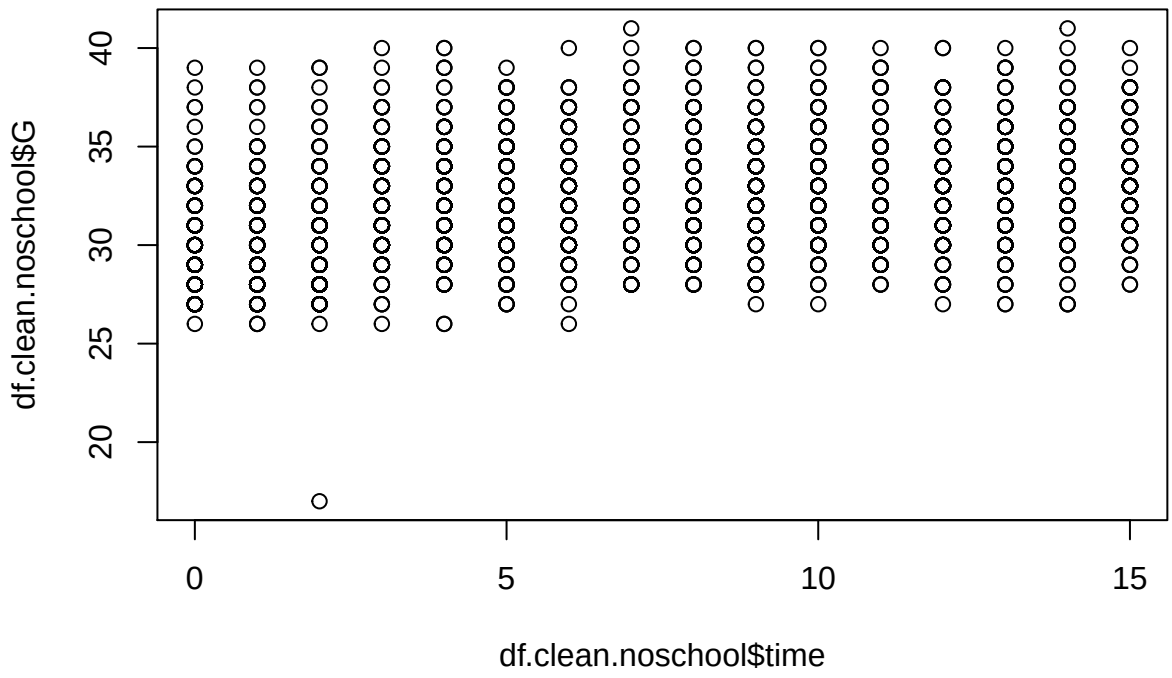
```
list_top = names(top_cor_list)
head(list_top)
```

```
## [1] "Opp." "X3PA" "FGA"  "G"    "X3P"  "Tm."
```

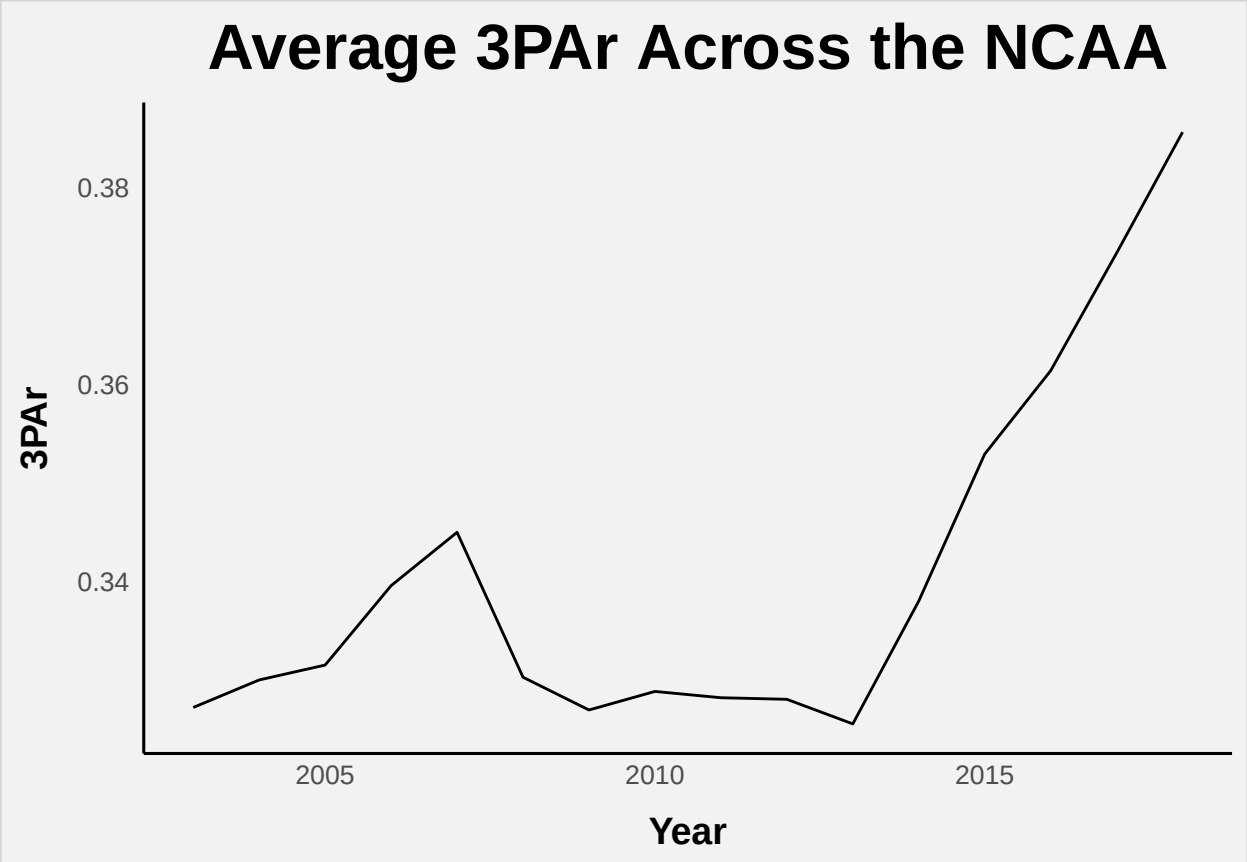
```
# graphs
df.clean.noschool %>%
  group_by(time) %>%
  summarise(mean_games = mean(G)) %>%
  ggplot(df.clean.noschool, mapping = aes(x = time + 2003, y = mean_games)) +
  geom_line(stat="identity") + ggtitle("Games Played Per Season in the NCAA") +
  ylim(25, 35) +
  xlab("Year") +
  ylab("Games")+
  theme_hodp()
```




```
# we noticed that games also increases over time (it's one of the top predictors)
plot(df.clean.noschool$time, df.clean.noschool$G)
```

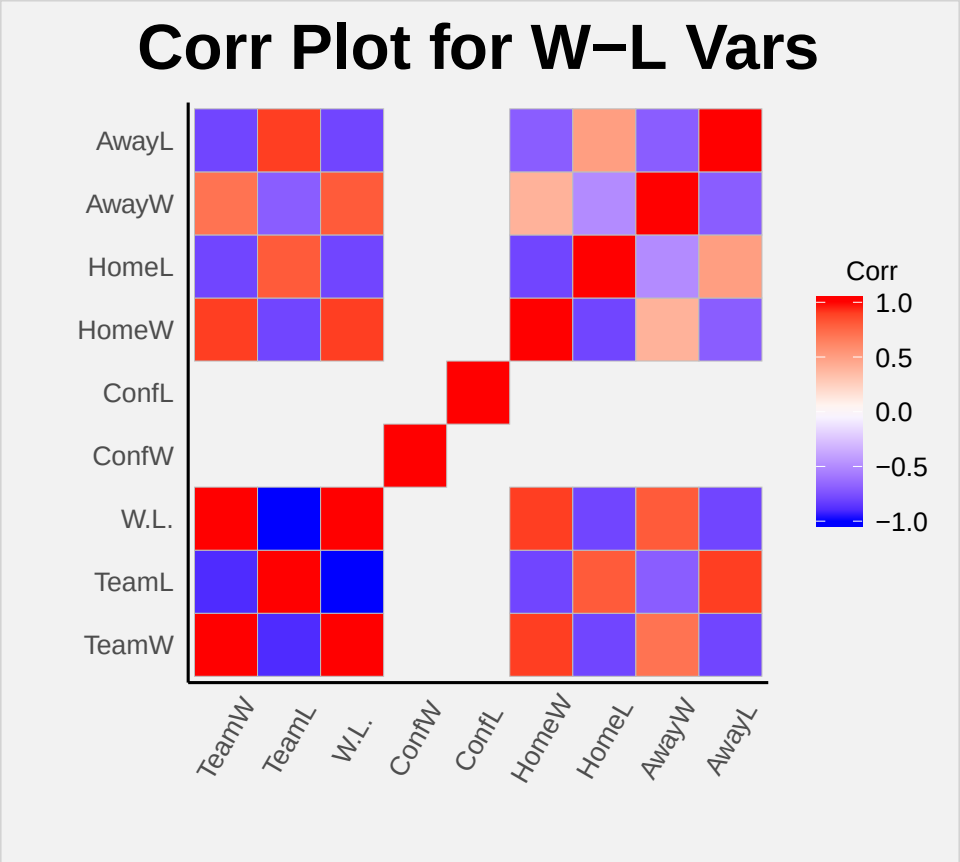



```
### MEANS PLOTS ###

# EDA plot to show how average 3Ar changes with time
df.tourney %>%
  group_by(time) %>%
  summarise(mean_three = mean(X3PAr)) %>%
  ggplot(df.tourney, mapping = aes(x = time + 2003, y = mean_three)) +
  geom_line(stat="identity") + ggtitle("Average 3PAr Across the NCAA") +
  xlab("Year") +
  ylab("3PAr")+
  theme_hodp()
```



CORRELATION PLOT

```
#eda correlation
data <- read.csv('data/full_data_raw.csv')
wl <- data %>% select(TeamW, TeamL, W.L., ConfW, ConfL, HomeW, HomeL, AwayW, AwayL)
cor <- round(cor(wl), 1)
p <- ggcorrplot(cor) +
  labs(title='Corr Plot for W-L Vars') +
  xlab('') + ylab('') +
  theme_hodp() +
  theme(axis.text.x=element_text(angle=60)) +
  theme(legend.position="right")
p
```


Corr Plot for W−L Vars
AwayL
AwayW
HomeL
HomeW
ConfL
ConfW
W.L.
TeamL
TeamW
TeamW
TeamL
W.L.
ConfW
ConfL
HomeW
HomeL
AwayW
AwayL
Corr
1.0
0.5
0.0
−0.5
−1.0