A Very Extensive NCAA Exploratory Analysis

Troy Walters

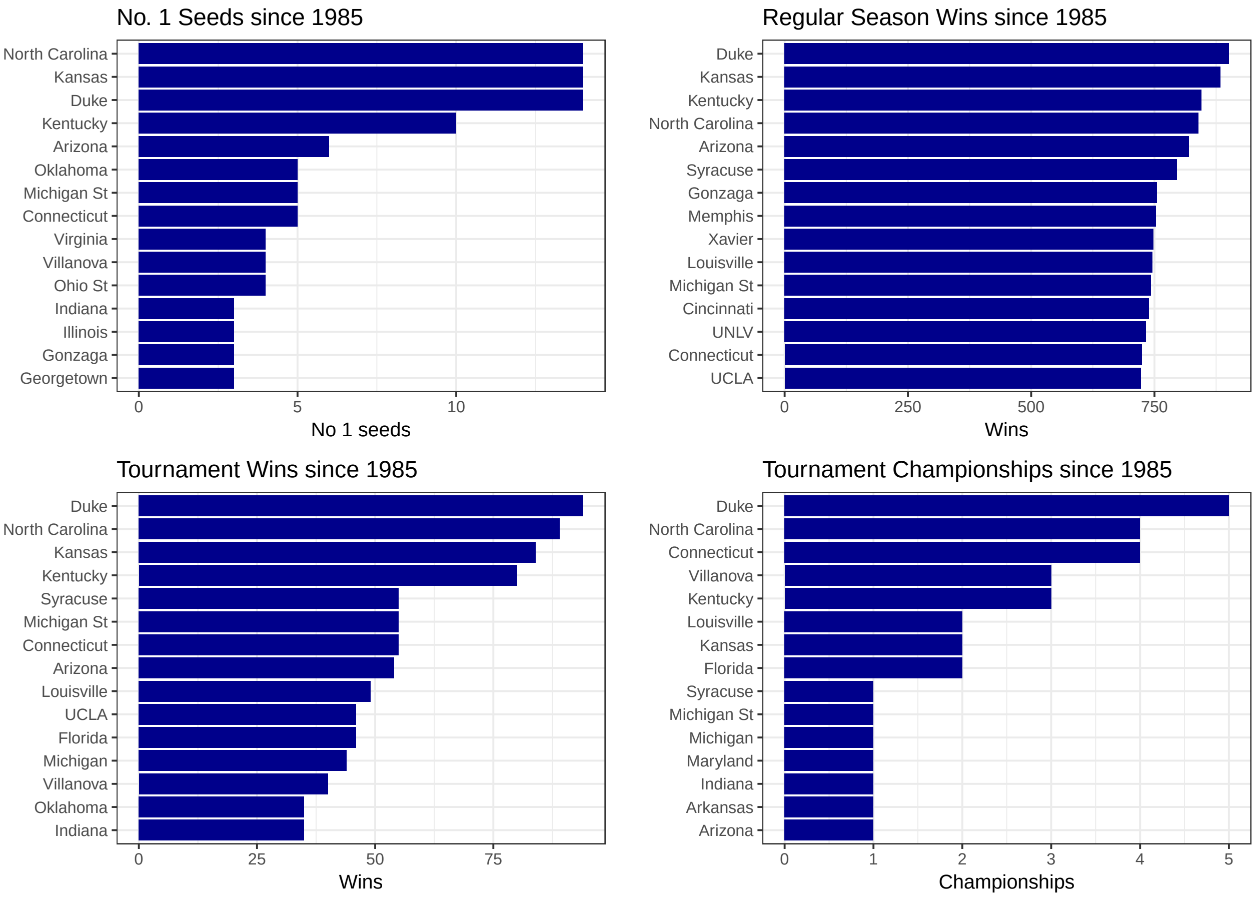
February 20, 2018

Libraries

```
library(data.table)
library(dplyr)
library(magrittr)
library(ggplot2)
library(gridExtra)
library(ggExtra)
library(corrplot)
library(factoextra)
library(stringr)
library(FactoMineR)

theme_set(theme_bw())
```

```
dir <- '/Users/JaviFerrando/Desktop/MLProject/input/'
setwd("/Users/JaviFerrando/Desktop/MLProject")

# Data Section 1
teams <- fread(paste(dir,'Teams.csv',sep=''))
seasons <- fread(paste(dir,'Seasons.csv',sep=''))
seeds <- fread(paste(dir,'NCAATourneySeeds.csv',sep=''))
seas_results <- fread(paste(dir,'RegularSeasonCompactResults.csv',sep=''))
tour_results <- fread(paste(dir,'NCAATourneyCompactResults.csv',sep=''))
seas_detail <- fread(paste(dir,'RegularSeasonDetailedResults.csv',sep=''))
tour_detail <- fread(paste(dir,'NCAATourneyDetailedResults.csv',sep=''))
conferences <- fread(paste(dir,'Conferences.csv',sep=''))
team_conferences <- fread(paste(dir,'TeamConferences.csv',sep=''))
coaches <- fread(paste(dir,'TeamCoaches.csv',sep=''))
tour_enrich <- fread(paste(dir,'NCAATourneyDetailedResultsEnriched.csv',sep=''))
seas_enrich <- fread(paste(dir,'NCAASeasonDetailedResultsEnriched.csv',sep=''))
```

```
setkey(teams, TeamID)
setkey(seeds, TeamID)

g1 <-
    teams[seeds][, one_seed := as.numeric(substr(Seed, 2, 3)) == 1][, sum(one_seed), by = TeamName][orde
    ggplot(aes(x = reorder(TeamName, V1), y = V1)) +
    geom_bar(stat = 'identity', fill = 'darkblue') +
    labs(x = '', y = 'No 1 seeds', title = 'No. 1 Seeds since 1985') +
    coord_flip()

setkey(seas_results, WTeamID)

g2 <-
    seas_results[teams][, .(wins = .N), by = TeamName][order(-wins)][1:15,] %>%
    ggplot(aes(x = reorder(TeamName, wins), y = wins)) +
    geom_bar(stat = 'identity', fill = 'darkblue') +
    labs(x = '', y = 'Wins', title = 'Regular Season Wins since 1985') +
    coord_flip()

setkey(tour_results, WTeamID)

g3 <-
    tour_results[teams][, .(wins = .N), by = TeamName][order(-wins)][1:15,] %>%
    ggplot(aes(x = reorder(TeamName, wins), y = wins)) +
    geom_bar(stat = 'identity', fill = 'darkblue') +
    labs(x = '', y = 'Wins', title = 'Tournament Wins since 1985') +
    coord_flip()

g4 <-
    tour_results[teams][DayNum == 154, .(wins = .N), by = TeamName][order(-wins)][1:15,] %>%
    ggplot(aes(x = reorder(TeamName, wins), y = wins)) +
    geom_bar(stat = 'identity', fill = 'darkblue') +
    labs(x = '', y = 'Championships', title = 'Tournament Championships since 1985') +
    coord_flip()

grid.arrange(g1, g2, g3, g4, nrow = 2)
```



## Historical Performance {#historical}

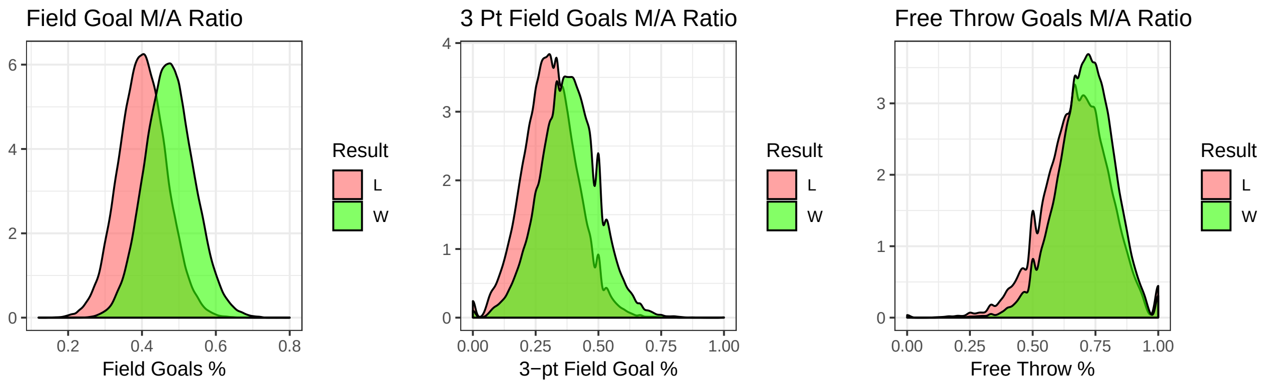
Indicators of Regular Season Success

Let's now turn to the regular season game statistics. We are interested in knowing how certain statistics correlate with winning vs losing. We will take the regular season detail and first convert it to a more 'long' format with only 1 column of TeamIDs and a factor indicating whether that row corresponds to a win or a loss. Here I also add some additional game statistcs. These include field goal percentage, free throw percentage, offensive/defensive rebounding efficiency, and possessions. These last two come from Laksan Nathan's kernel here.

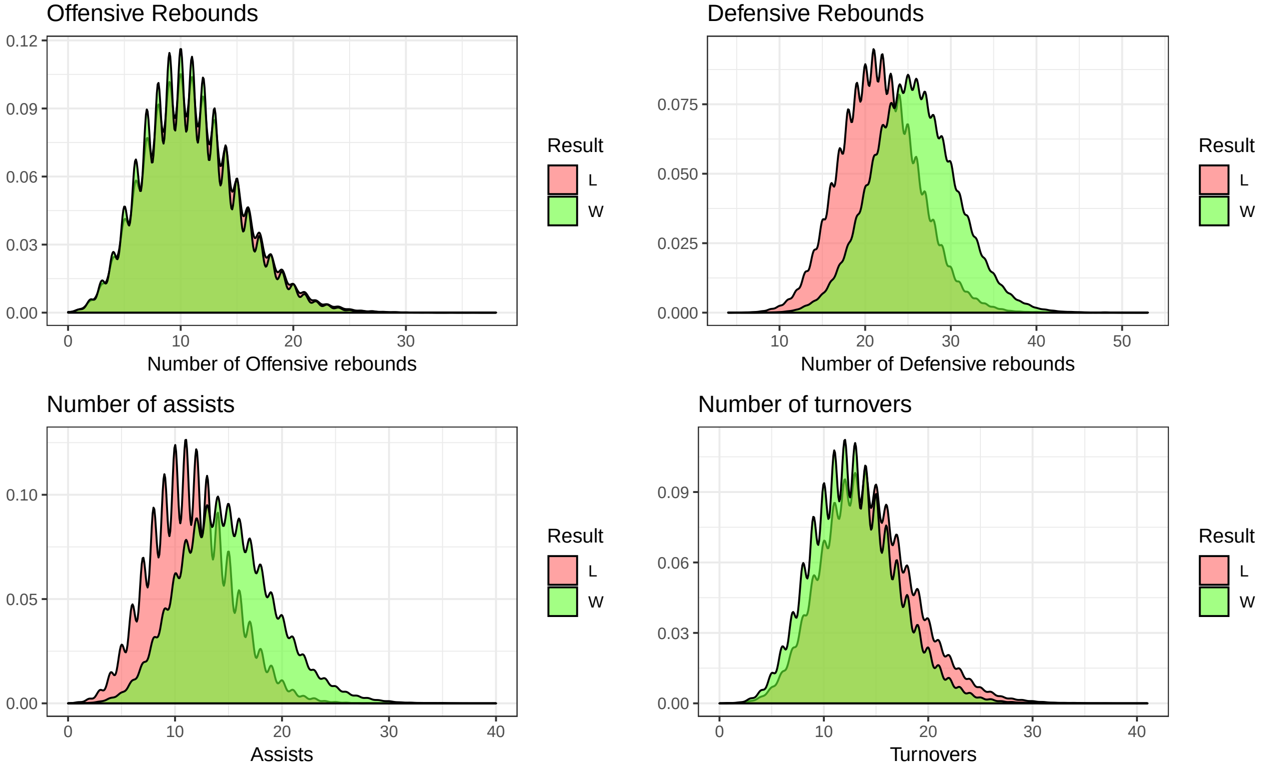
```r
win_stats <- seas_enrich[, .(
    Season,
    TeamID = WTeamID,
    Result = rep('W', .N),
    FGM = WFGM,
    FGA = WFGA,
    FGP = WFGM / WFGA,
    FGP2 = (WFGM - WFGM3) / (WFGA - WFGA3),
    FGM3 = WFGM3,
    FGA3 = WFGA3,
    FGP3 = WFGM3 / WFGA3,
    FTM = WFTM,
    FTA = WFTA,
    FTP = WFTM / WFTA,
    OR = WOR,
    DR = WDR,
    AST = WAst,
    TO = WTO,
    STL = WStl,
    BLK = WBlk,
    PF = WPF,
    PIE = WPIE,
    ORP = WOR / (WOR + LDR),
    DRP = WDR / (WDR + LOR),
    eFG = WeFGP,
    NetRTG = WNetRtg,
    FTAR = WFTAR,
    POS = 0.96 * (WFGA + WTO + 0.44 * WFTA - WOR)
    )]

los_stats <- seas_enrich[, .(
    Season,
    TeamID = LTeamID,
    Result = rep('L', .N),
    FGM = LFGM,
    FGA = LFGA,
    FGP = LFGM / LFGA,
    FGP2 = (LFGM - LFGM3) / (LFGA - LFGA3),
    FGM3 = LFGM3,
    FGA3 = LFGA3,
    FGP3 = LFGM3 / LFGA3,
    FTM = LFTM,
    FTA = LFTA,
    FTP = LFTM / LFTA,
    OR = LOR,
    DR = LDR,
    AST = LAst,
    TO = LTO,
    STL = LStl,
    BLK = LBlk,
    PF = LPF,
    PIE = LPIE,
    ORP = (LOR / (LOR + WDR)),
    DRP = LDR / (LDR + WOR),
    eFG = LeFGP,
    NetRTG = LNetRtg,
    FTAR = LFTAR,
    POS = 0.96 * (LFGA + LTO + 0.44 * LFTA - LOR)
    )]

stats_all <- rbindlist(list(win_stats, los_stats))
```

Now let's take a look at the distributions of these statistics for winning and losing teams.

```
g1 <- stats_all %>%
    ggplot(aes(x = FGP, fill = Result)) +
    geom_density(alpha = 0.6) +
    scale_fill_manual(values = c('#FF6666', '#33FF00')) +
    labs(x = 'Field Goals %', y = '', title = 'Field Goal M/A Ratio')

g2 <- stats_all %>%
    ggplot(aes(x = FGP2, fill = Result)) +
    geom_density(alpha = 0.6) +
    scale_fill_manual(values = c('#FF6666', '#33FF00')) +
    labs(x = '2-pt Field Goal %', y = '', title = '2 Pt Field Goals M/A Ratio')

g3 <- stats_all %>%
    ggplot(aes(x = FGP3, fill = Result)) +
    geom_density(alpha = 0.6) +
    scale_fill_manual(values = c('#FF6666', '#33FF00')) +
    labs(x = '3-pt Field Goal %', y = '', title = '3 Pt Field Goals M/A Ratio')

g4 <- stats_all %>%
    ggplot(aes(x = FTP, fill = Result)) +
    geom_density(alpha = 0.6) +
    scale_fill_manual(values = c('#FF6666', '#33FF00')) +
    labs(x = 'Free Throw %', y = '', title = 'Free Throw Goals M/A Ratio')


grid.arrange(g1, g3, g4, ncol = 3)
```



```r
g5 <- stats_all %>%
    ggplot(aes(x = OR, fill = Result)) +
    geom_density(alpha = 0.6) +
    scale_fill_manual(values = c('#FF6666', '#66FF33')) +
    labs(x = 'Number of Offensive rebounds', y = '', title = 'Offensive Rebounds')

g6 <- stats_all %>%
    ggplot(aes(x = DR, fill = Result)) +
    geom_density(alpha = 0.6) +
    scale_fill_manual(values = c('#FF6666', '#66FF33')) +
    labs(x = 'Number of Defensive rebounds', y = '', title = 'Defensive Rebounds')

g7 <- stats_all %>%
    ggplot(aes(x = AST, fill = Result)) +
    geom_density(alpha = 0.6) +
    scale_fill_manual(values = c('#FF6666', '#66FF33')) +
    labs(x = 'Assists', y = '', title = 'Number of assists')

g8 <- stats_all %>%
    ggplot(aes(x = TO, fill = Result)) +
    geom_density(alpha = 0.6) +
    scale_fill_manual(values = c('#FF6666', '#66FF33')) +
    labs(x = 'Turnovers', y = '', title = 'Number of turnovers')

g9 <- stats_all %>%
    ggplot(aes(x = STL, fill = Result)) +
    geom_density(alpha = 0.6) +
    scale_fill_manual(values = c('#FF6666', '#66FF33')) +
    labs(x = 'Steals', y = '', title = 'Steals per Game')

g10 <- stats_all %>%
    ggplot(aes(x = BLK, fill = Result)) +
    geom_density(alpha = 0.6) +
    scale_fill_manual(values = c('#FF6666', '#66FF33')) +
    labs(x = 'Blocks', y = '', title = 'Blocks per Game')

grid.arrange(g5, g6, g7, g8, ncol = 2)
```



Unsurprisingly, we see that winning teams tend to have a higher mean (or lower in the case of turnover) in pretty much every metric. The last few plots are bit spikey due to the more discrete nature of the data.

We don't have final game statistics until we have the Result, so we obviously can't use these statistics in this form to predict the winners of tournament matchups. However, we can use regular season aggregate statistics to predict the winner in tournament matchups. Let's take a look at that next.

Predictors of Tournament Success

One of the obvious predictors for how deep a team goes in the tournament would be regular season wins. Let's see how regular season wins correlate to tournament progress each year.

```
wins_s <- seas_results[, .(rsW = .N), by = c('WTeamID', 'Season')] #Wins of a team per season

wins_t <- tour_results[!(DayNum %in% c(134, 135)), .(tW = .N), by = c('WTeamID', 'Season')]

#max(wins_t$tW)

wins_teams <- wins_s[wins_t][teams]
```

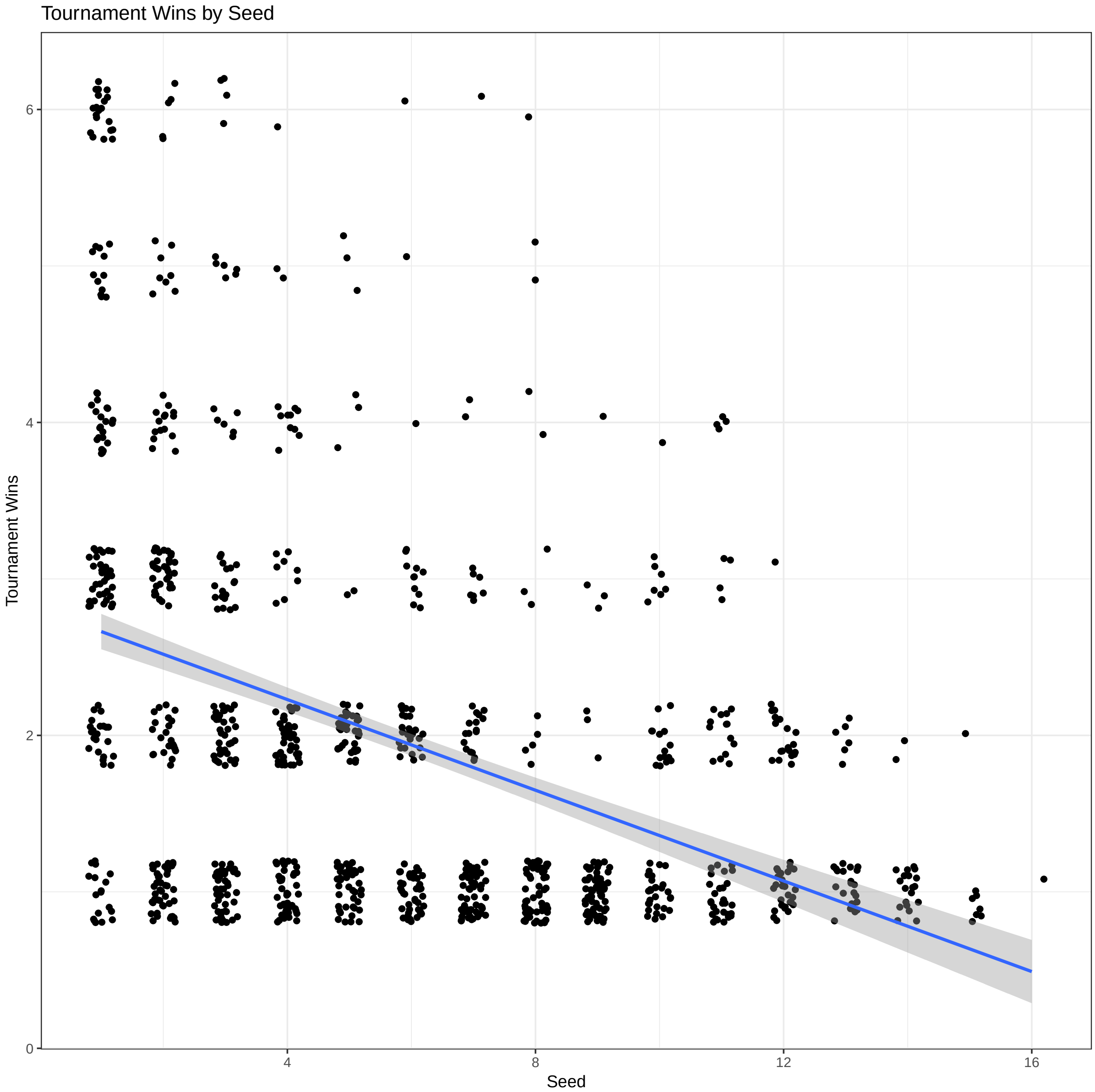
In nearly every year, tournament wins is positively correlated with regular season wins. Of course, there are some exceptions - for example in 2000, the relationship is slightly negative! Single-elimination tournaments produce some variations as they leave little room for error. Sometimes strong favorites don't get as far as expected.

The problem with using regular season wins is that in college basketball, not every team plays the same number of games in a regular season. Let's do something similar to see if average scores during regular season are associated with better tournament progress.

```
wins <- seas_results[, .(n_games = .N, sum_score = sum(WScore)), by = c('WTeamID', 'Season')]
```

```
losses <- seas_results[, .(n_games = .N, sum_score = sum(LScore)), by = c('LTeamID', 'Season')]

all_games <- rbindlist(list(wins, losses))

all_games <- all_games[, .(rs_ppg = sum(sum_score) / sum(n_games)), by = c('WTeamID', 'Season')]
```

```
seeds[, .(Season, WTeamID = TeamID, seed_num = as.numeric(substr(Seed, 2, 3)))
      ][wins_t, on = c('Season', 'WTeamID')] %>%
    ggplot(aes(x = seed_num, y = tW)) +
    geom_jitter(width = 0.2, height = 0.2) +
    geom_smooth(method = 'lm') +
    labs(
        x = 'Seed',
        y = 'Tournament Wins',
        title = 'Tournament Wins by Seed')
```

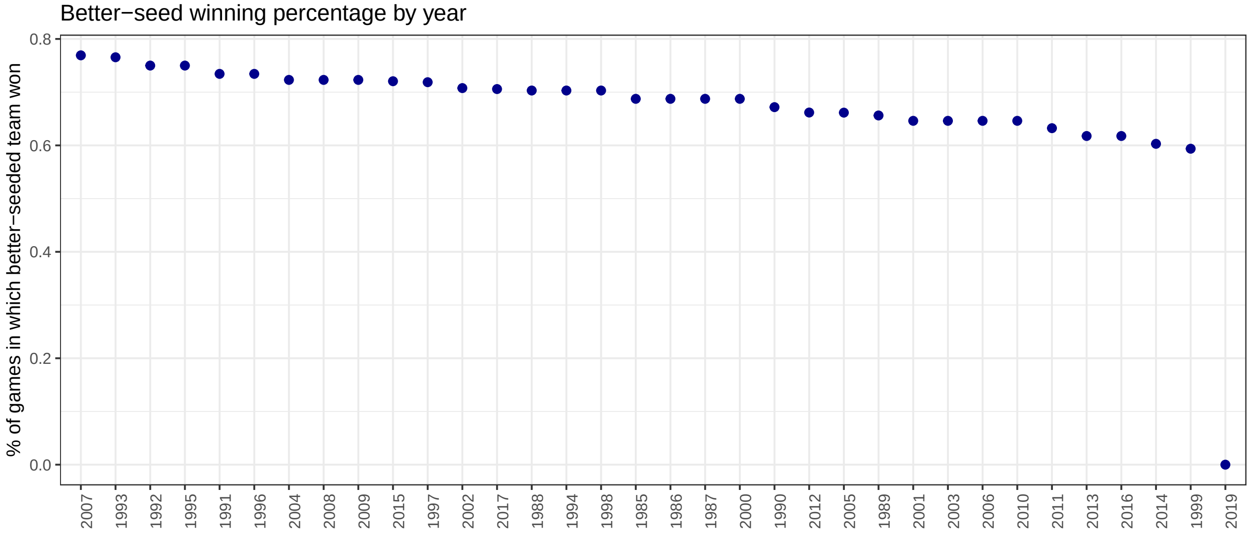


I've introduced some jiter to this plot to avoid overplotting. It exhibits a strong negative relationship between seed and tournament progress - the lower a team's seed, the deeper they go into the tournament (as measured by tournament wins). We see that a 16 seed has never made it past the first round of the tournament. From the plot we can also determine that the lowest seed to ever win the tournament was a number 8. A vast majority of teams that have won the tournament since 1985 have been number 1 seeds.

We may also wonder how likely it is that a better-seeded (i.e. lower number) team will win any particular tournament matchup. Let's look at the percentage of times the better-seeded team won by season.

```
tour_results_seeds <- seeds[, .(Season, WTeamID = TeamID, winner_seed = as.numeric(substr(Seed, 2, 3)))
                            ][tour_results, on = c('Season', 'WTeamID')
                              ][seeds[, .(Season, LTeamID = TeamID, loser_seed = as.numeric(substr(Seed
                                      ], on = c('Season', 'LTeamID')]
```

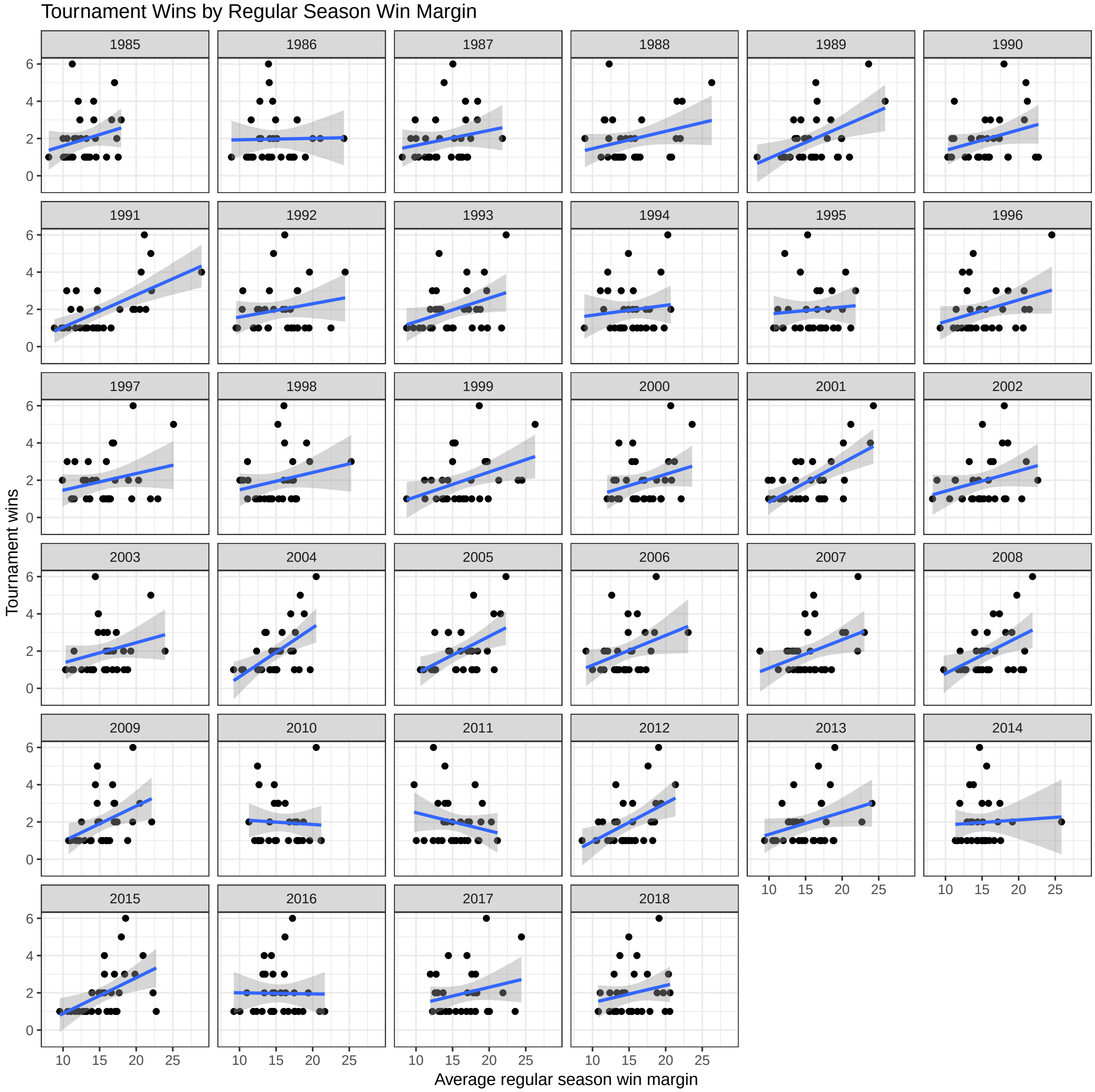
```
tour_results_seeds[Season != 2018, .(Season, low_seed_win = ifelse(winner_seed < loser_seed, 1, 0))
                   ][, sum(low_seed_win, na.rm = TRUE) / .N, by = Season] %>%
    ggplot(aes(x = reorder(Season, -V1), y = V1)) +
    geom_point(color = 'darkblue', size = 2) +
    labs(
        x = '',
        y = '% of games in which better-seeded team won',
        title = 'Better-seed winning percentage by year') +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



When examining these data by season, we see that the better-seeded team won games at a rate that varies between 0.79 in the 2007 tournament to approximately 0.59 in the 1999 tournament.

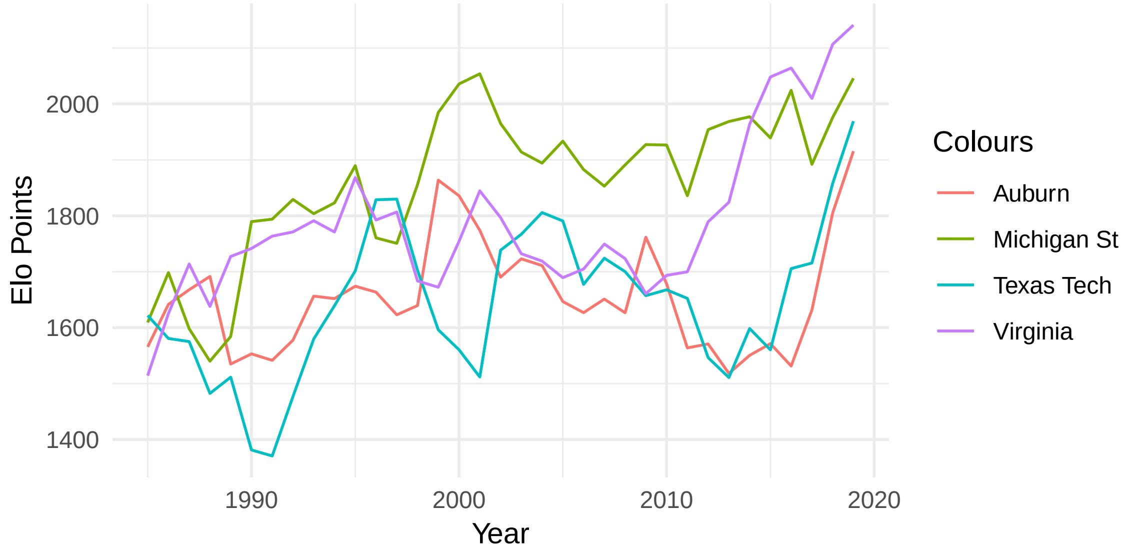
Now let's examine the relationship between a team's regular season win margin and its tournament performance.

```
seas_results[, .(avg_win_marg = mean(WScore - LScore)), by = c('WTeamID', 'Season')
             ][wins_t, on = c('WTeamID', 'Season')] %>%
    ggplot(aes(x = avg_win_marg, y = tW)) +
    geom_point() +
    geom_smooth(method = 'lm') +
    labs(
        x = 'Average regular season win margin',
        y = 'Tournament wins',
        title = 'Tournament Wins by Regular Season Win Margin') +
    facet_wrap(~Season)
```

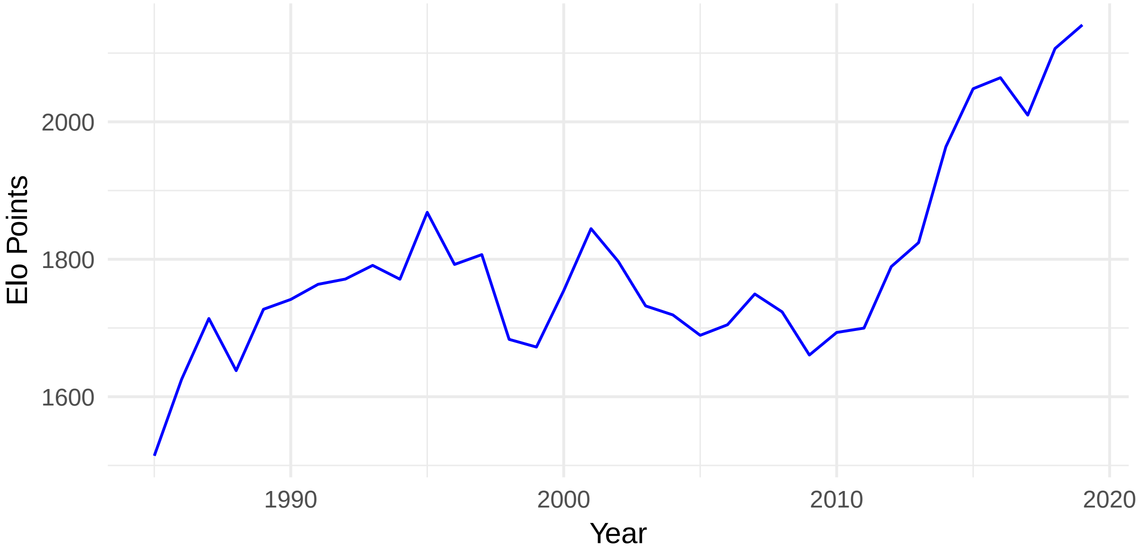


Now let's move beyond the basic stats and use some of the box score data as well. To start, let's create a standardized data frame in wide format with all of the teams regular season stats. We'll create some additional statistics such as various shooting percentages, rebounds per game, steals per game, etc. Because of the format of the data, we first need to get the stats for the games winning teams and losing teams seperately. Then we will bind these row-wise and group by Season and TeamID to calculate the stats.

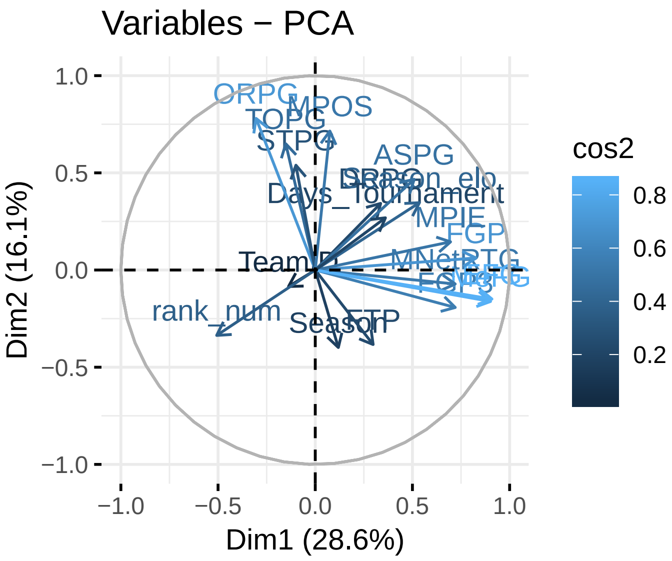
```
#stats_season -> average
stats_season <- stats_all[, .(
    FGP = sum(FGM) / sum(FGA),
    FGP3 = sum(FGM3) / sum(FGA3),
    FTP = sum(FTM) / sum(FTA),
    ORPG = mean(OR),
    DRPG = mean(DR),
    ASPG = mean(AST),
    TOPG = mean(TO),
    STPG = mean(STL),
    #MFTAR = mean(FTAR),
    #BLPG = mean(BLK),
    #PFPG = mean(PF),
    MeFG = mean(eFG),
    MNetRTG = mean(NetRTG),
    #MORP = mean(ORP),
    MPIE = mean(PIE),
    MPOS = mean(POS),
    EFG = (mean(FGM)+0.5*mean(FGM3))/mean(FGA))

    , by = c('TeamID', 'Season')]
seas_elos <- fread(paste(dir,'season_elos.csv',sep=''))
seas_elo_duke <-seas_elos[seas_elos$team_id==1181]
seas_elo_virginia <-seas_elos[seas_elos$team_id==1438]
seas_elo_virginia[seas_elo_virginia==1438] <- "Virginia"
seas_elo_texas <-seas_elos[seas_elos$team_id==1403]
seas_elo_texas[seas_elo_texas==1403] <- "Texas Tech"
seas_elo_michigan <-seas_elos[seas_elos$team_id==1277]
seas_elo_michigan[seas_elo_michigan==1277] <- "Michigan St"
seas_elo_auburn <-seas_elos[seas_elos$team_id==1120]
seas_elo_auburn[seas_elo_auburn==1120] <- "Auburn"

elos_plot <- rbind(seas_elo_virginia,seas_elo_texas,seas_elo_michigan,seas_elo_auburn)
elos_plot$team_id <- as.factor(elos_plot$team_id)


ggplot(data=elos_plot, aes(x=season, y=season_elo, group=team_id)) +
  #geom_point(aes(x=seas_elos$season, y=seas_elos$season_elo), size = 3) +
  geom_line(aes(color=team_id))+
  #scale_color_brewer(palette="Paired")+
  theme_minimal() +
  scale_fill_discrete(name = "Teams") +
  labs(x = "Year") + labs(y = "Elo Points")+
  labs(color="Colours")
```



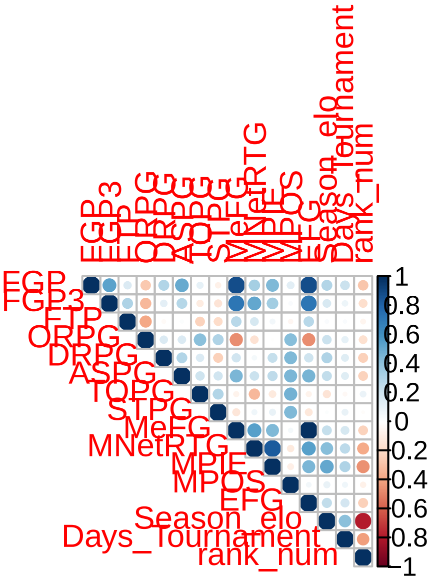
```
ggplot(data=seas_elo_virginia, aes(x=seas_elo_virginia$season, y=seas_elo_virginia$season_elo)) +
  #geom_point(aes(x=seas_elos$season, y=seas_elos$season_elo), size = 3) +
  geom_line(color="blue")+
  scale_color_brewer(palette="Paired")+
  theme_minimal() +
  labs(x = "Year") + labs(y = "Elo Points")+
  labs(color="Colours")
```



```
seas_elos <- seas_elos[seas_elos$season>=2003,]

colnames(seas_elos) <-c("TeamID","Season","Season_elo")
stats_season <- merge(x = stats_season, y = seas_elos, by=c("TeamID","Season"), all = FALSE)
wins_t_2003 <- wins_t[wins_t$Season>=2003,]
colnames(wins_t_2003) <-c("TeamID","Season","Days_Tournament")
stats_season <- merge(x = stats_season, y = wins_t_2003, by=c("TeamID","Season"), all = FALSE)
dseeds_tournament <- fread(paste(dir,'NCAATourneySeeds.csv',sep=''))
dseeds_tournament <- dseeds_tournament %>%
  mutate(ranking = as.factor((str_replace(Seed, "[A-Z]",""))),
         rank_num = as.numeric(str_replace(ranking, ".[a-z]","")))
dseeds_tournament <- dseeds_tournament[ , -which(names(dseeds_tournament) %in% c("Seed","ranking"))]
#names(dseeds_tournament) <- tolower(names(dseeds_tournament))
stats_season <- merge(x = stats_season, y = dseeds_tournament, by=c("TeamID","Season"), all = FALSE)

pca_crypto <- princomp(~Season_elo+EFG+MPOS+Days_Tournament+MPIE+MNetRTG,cor=TRUE, data=stats_season)
pca_crypto <- PCA(stats_season,scale.unit = TRUE,graph = FALSE)
fviz_pca_var(pca_crypto, col.var = "cos2")
```



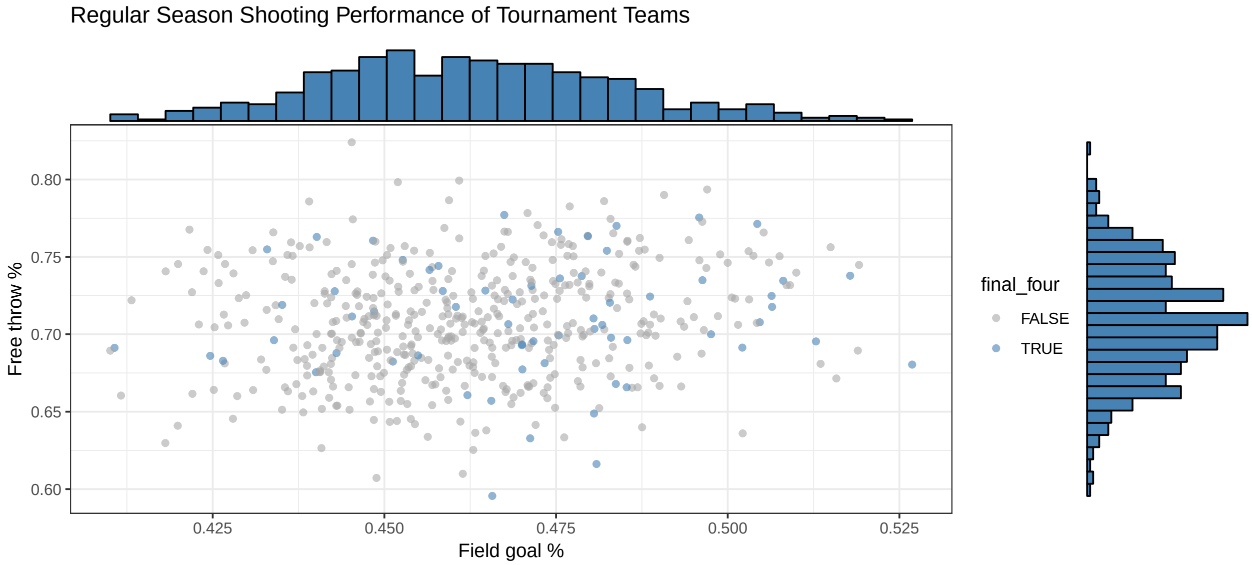
```
M <- cor(scale(stats_season %>% select (3:18)),method="pearson")
corrplot(M, method = "circle",type = "upper")
```



```
#corrplot(stats_season, type = "upper", order = "hclust",
         #tl.col = "black")
```

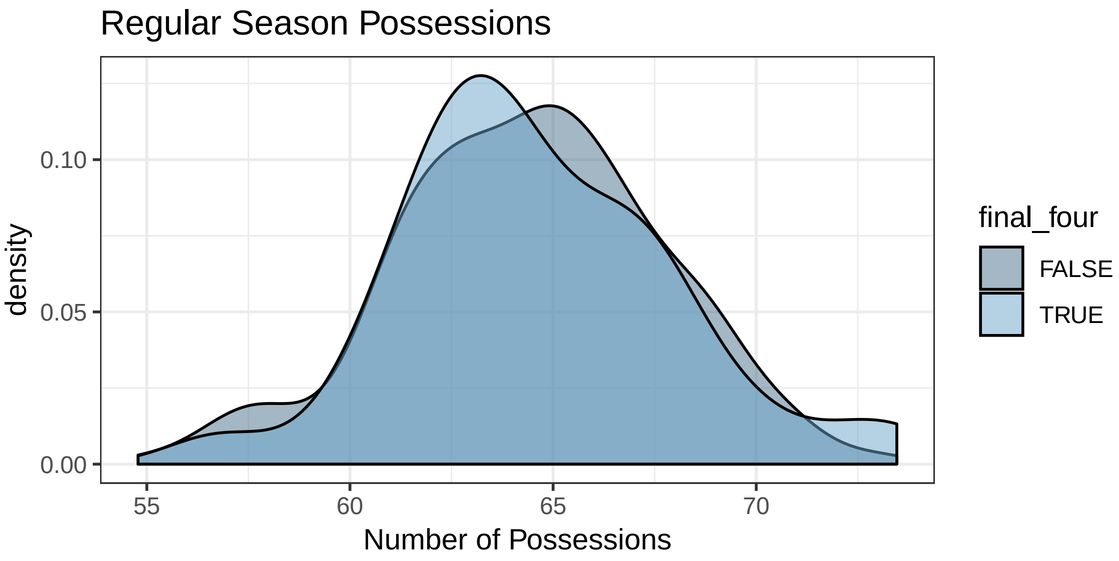
Can we use a team's regular season game statistics to predict tournament success. First let's look at field goal % and free throw % during the regular season and see if these equate to tournament success. We'll define success in the case as making the Final Four.

```
g1 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = FGP, y = FTP, color = final_four)) +
    geom_point(alpha = 0.6) +
    labs(
        x = 'Field goal %',
        y = 'Free throw %',
        title = 'Regular Season Shooting Performance of Tournament Teams') +
    scale_color_manual(values = c('darkgrey', 'steelblue'))

ggMarginal(g1, type = 'histogram', fill = 'steelblue')
```

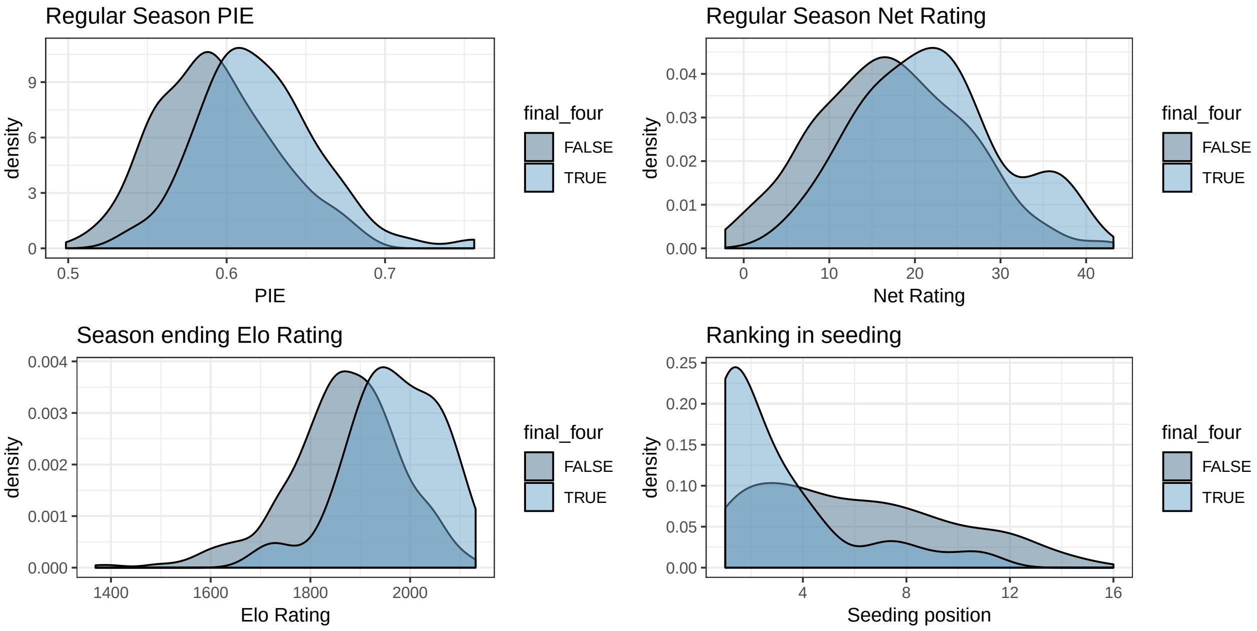


We see that the distribution of field goal % appears to have a peak around 0.45. The distribution of free throw percentage peaks near 72%. Interestingly in terms of shooting %, there does not seem to be much of a difference between teams that make the Final Four and the rest of the tournament field in terms of their regular season performance; however it is hard to tell from this plot type. To double-check, let's plot the densities of these two statistics for Final Four teams and the rest of the field.

```
g1 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
                ][, final_four := tW >= 4] %>%
    ggplot(aes(x = MPOS, fill = final_four)) +
    scale_fill_manual(values = c('skyblue4', 'skyblue3')) +
    geom_density(alpha = 0.5) +
    labs(x = 'Number of Possessions', title = 'Regular Season Possessions')
grid.arrange(g1)
```

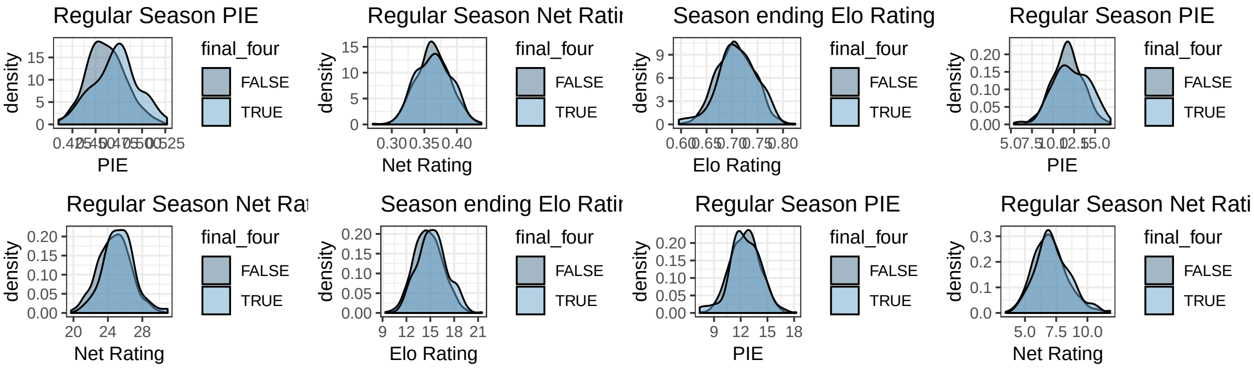


```
g1 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = MPIE, fill = final_four)) +
    scale_fill_manual(values = c('skyblue4', 'skyblue3')) +
    geom_density(alpha = 0.5) +
    labs(x = 'PIE', title = 'Regular Season PIE')

g2 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = MNetRTG, fill = final_four)) +
    scale_fill_manual(values = c('skyblue4', 'skyblue3')) +
    geom_density(alpha = 0.5) +
    labs(x = 'Net Rating', title = 'Regular Season Net Rating')

g3 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = Season_elo, fill = final_four)) +
    scale_fill_manual(values = c('skyblue4', 'skyblue3')) +
    geom_density(alpha = 0.5) +
    labs(x = 'Elo Rating', title = 'Season ending Elo Rating')

g4 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = rank_num, fill = final_four)) +
    scale_fill_manual(values = c('skyblue4', 'skyblue3')) +
    geom_density(alpha = 0.5) +
    labs(x = 'Seeding position', title = 'Ranking in seeding')


grid.arrange(g1, g2, g3, g4, ncol = 2)
```



```
g1 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = FGP, fill = final_four)) +
    scale_fill_manual(values = c('skyblue4', 'skyblue3')) +
    geom_density(alpha = 0.5) +
    labs(x = 'PIE', title = 'Regular Season PIE')

g2 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = FGP3, fill = final_four)) +
    scale_fill_manual(values = c('skyblue4', 'skyblue3')) +
    geom_density(alpha = 0.5) +
    labs(x = 'Net Rating', title = 'Regular Season Net Rating')

g3 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = FTP, fill = final_four)) +
    scale_fill_manual(values = c('skyblue4', 'skyblue3')) +
    geom_density(alpha = 0.5) +
    labs(x = 'Elo Rating', title = 'Season ending Elo Rating')
g4 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = ORPG, fill = final_four)) +
    scale_fill_manual(values = c('skyblue4', 'skyblue3')) +
    geom_density(alpha = 0.5) +
    labs(x = 'PIE', title = 'Regular Season PIE')

g5 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = DRPG, fill = final_four)) +
    scale_fill_manual(values = c('skyblue4', 'skyblue3')) +
    geom_density(alpha = 0.5) +
    labs(x = 'Net Rating', title = 'Regular Season Net Rating')

g6 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = ASPG, fill = final_four)) +
    scale_fill_manual(values = c('skyblue4', 'skyblue3')) +
    geom_density(alpha = 0.5) +
    labs(x = 'Elo Rating', title = 'Season ending Elo Rating')
g7 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = TOPG, fill = final_four)) +
    scale_fill_manual(values = c('skyblue4', 'skyblue3')) +
    geom_density(alpha = 0.5) +
    labs(x = 'PIE', title = 'Regular Season PIE')

g8 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = STPG, fill = final_four)) +
    scale_fill_manual(values = c('skyblue4', 'skyblue3')) +
    geom_density(alpha = 0.5) +
    labs(x = 'Net Rating', title = 'Regular Season Net Rating')
```

```r
grid.arrange(g1, g2, g3, g4, g5, g6, g7, g8, ncol = 4)
```



From the density plots, it actually appears that Final Four teams do shoot better from the floor during the regular season. Non Final Four teams shoot around 0.45 during the regular season and Final Four teams seem to shoot around 0.475. It is very important to keep in mind however that the sample size for Final Four teams is much smaller than the sample size for the rest of the tournament field. Therefore its unclear whether we can consider this difference statistically significant. For free throw percentage, there does not appear to be much of a difference.

Let's get a better idea of whether the difference in field goal percentage is real. We can use a two-sample t-test to determine if there is a difference in the sample means. Because the sample size of the two are different (and hence the variance), we can use Welch's two-sample t-test.

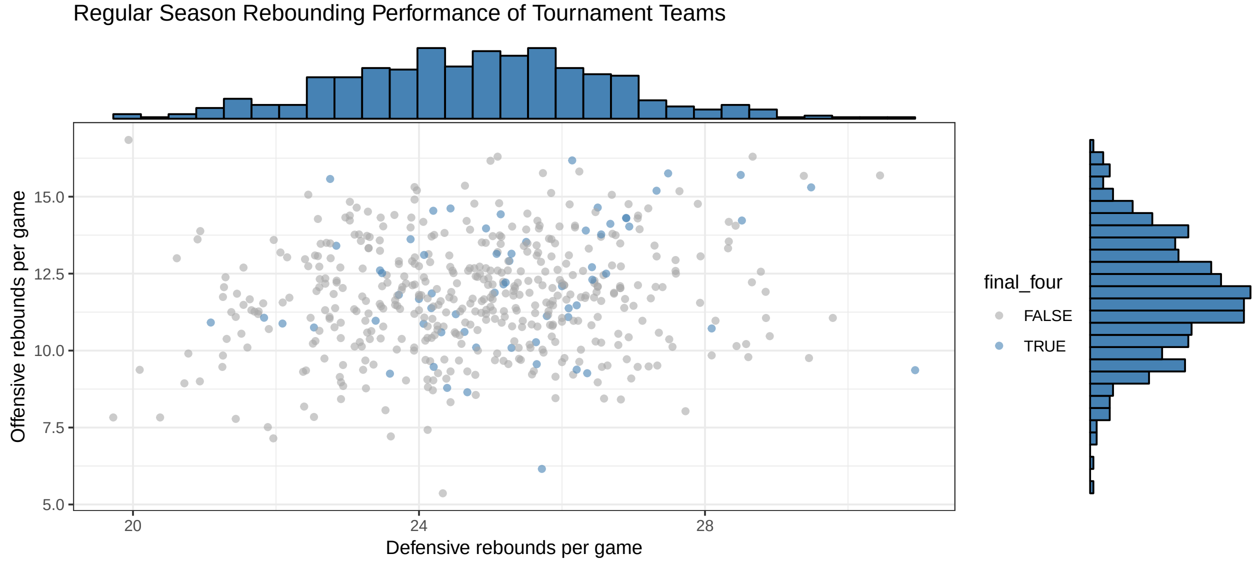
```r
fgp_noff <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4
                ][final_four == FALSE,  rank_num]

fgp_ff <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4
                ][final_four == TRUE,  rank_num]

t.test(fgp_noff, fgp_ff, alternative = 'two.sided', var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  fgp_noff and fgp_ff
## t = 8.1177, df = 102.29, p-value = 1.116e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.299050 3.785771
## sample estimates:
## mean of x mean of y
##  6.042411  3.000000
```

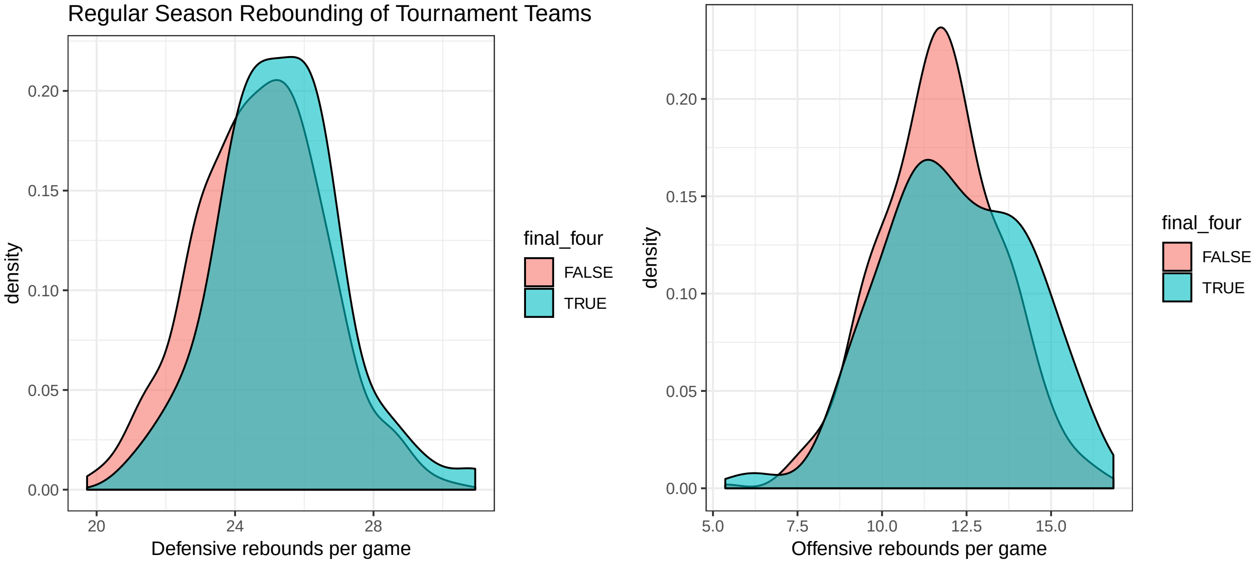
When doing so, we get a test statistic of -3.1443 and a p-value of 0.002417. At the 95% significance level therefore, we can reject the null hypothesis of a zero difference in mean and accept evidence of the alternative hypothesis that there is a difference in the mean field goal percentage of Final Four teams and non-Final Four teams. That difference appears to be about one percentage point.

Now let's do the same thing for rebounding performance.

```r
g1 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = DRPG, y = ORPG, color = final_four)) +
    geom_point(alpha = 0.6) +
    labs(
        x = 'Defensive rebounds per game',
        y = 'Offensive rebounds per game',
        title = 'Regular Season Rebounding Performance of Tournament Teams') +
    scale_color_manual(values = c('darkgrey', 'steelblue'))

ggMarginal(g1, type = 'histogram', fill = 'steelblue')
```

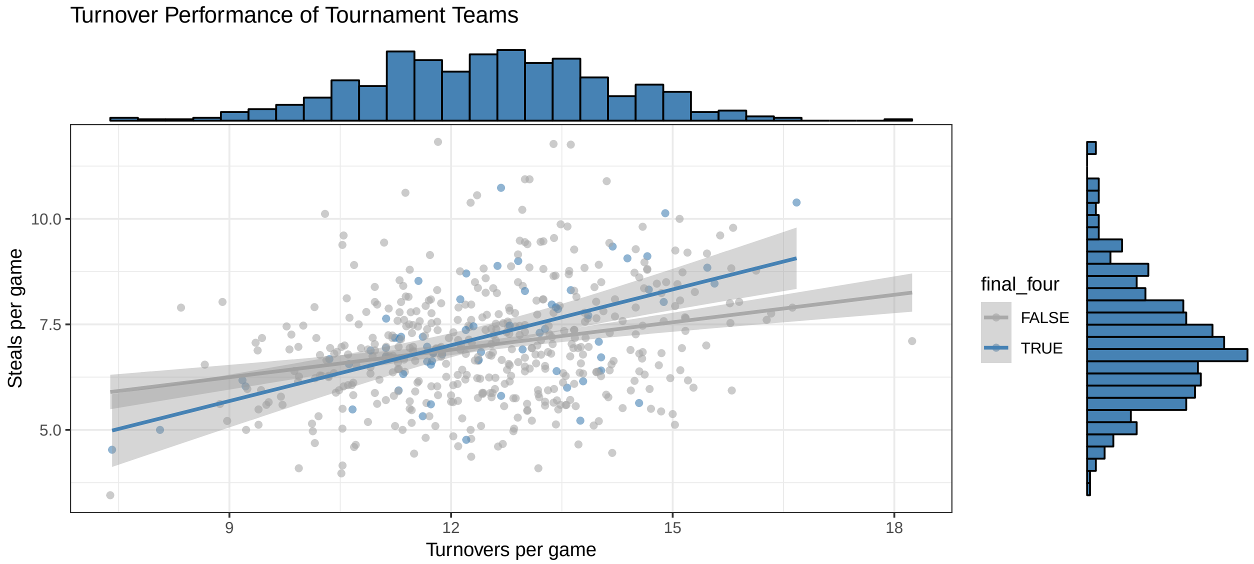


```
g1 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = DRPG, fill = final_four)) +
    geom_density(alpha = 0.6) +
    labs(x = 'Defensive rebounds per game', title = 'Regular Season Rebounding of Tournament Teams')

g2 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = ORPG, fill = final_four)) +
    geom_density(alpha = 0.6) +
    labs(x = 'Offensive rebounds per game')

grid.arrange(g1, g2, ncol = 2)
```



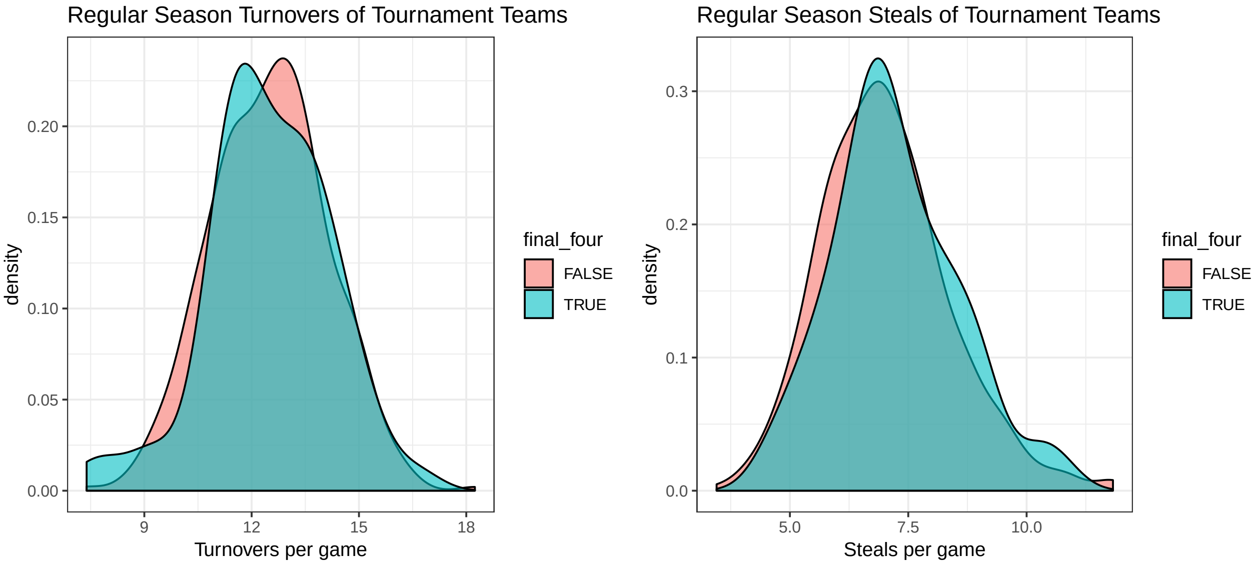
In terms of defensive rebounding, there does not appear to be much separation between Final Four teams and the rest of the field. The same goes for offensive rebounding, however the appears to be a skew in the distribution for Final Four teams, perhaps an artifact of limited sample size.

```
g1 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
       ][, final_four := tW >= 4] %>%
  ggplot(aes(x = TOPG, y = STPG, color = final_four)) +
  geom_point(alpha = 0.6) +
  geom_smooth(aes(color = final_four), method = 'lm') +
  labs(
    x = 'Turnovers per game',
    y = 'Steals per game',
    title = 'Turnover Performance of Tournament Teams') +
  scale_color_manual(values = c('darkgrey', 'steelblue'))

ggMarginal(g1, type = 'histogram', fill = 'steelblue')
```



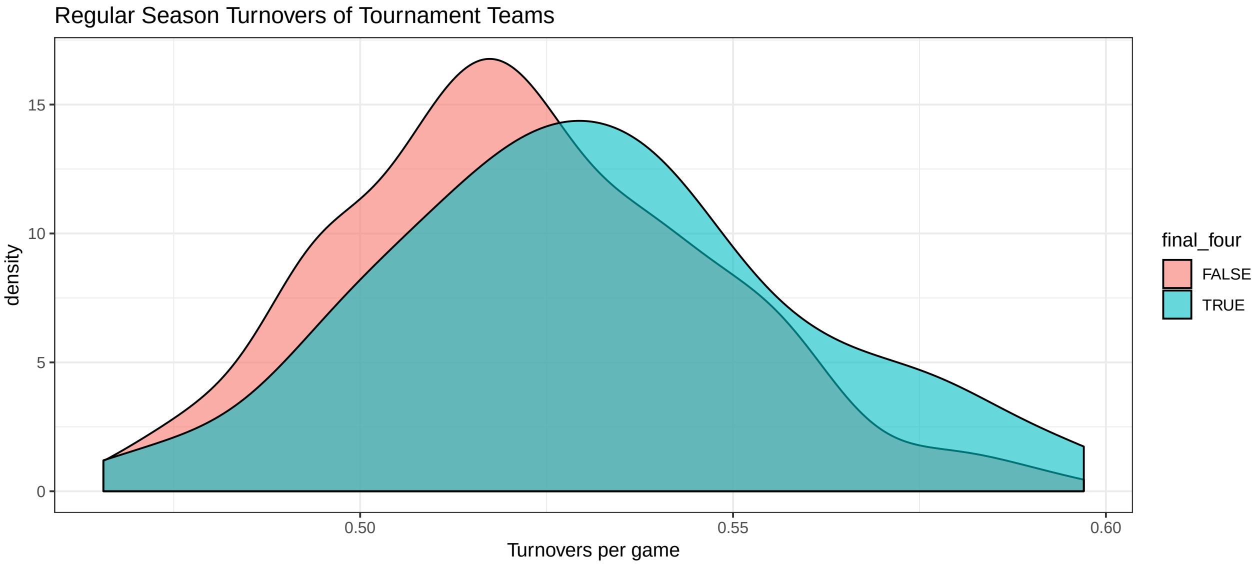
The ratio of steals to turnovers is positive for all tournament teams, however the relatioship appears to be stronger for Final Four teams indicating that this ratio be be a good predictor of tournament success.

```
g1 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = TOPG, fill = final_four)) +
    geom_density(alpha = 0.6) +
    labs(x = 'Turnovers per game', title = 'Regular Season Turnovers of Tournament Teams')

g2 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = STPG, fill = final_four)) +
    geom_density(alpha = 0.6) +
    labs(x = 'Steals per game',  title = 'Regular Season Steals of Tournament Teams')

grid.arrange(g1, g2, ncol = 2)
```



There seems to be some separation of means between all tournament teams and Final Four teams for regular season turnovers per game, however its inconclusive whether or not the difference is significant.

```
g1 <- stats_season[wins_t, on = c(TeamID = 'WTeamID', 'Season'), nomatch = 0
              ][, final_four := tW >= 4] %>%
    ggplot(aes(x = EFG, fill = final_four)) +
    geom_density(alpha = 0.6) +
    labs(x = 'Turnovers per game', title = 'Regular Season Turnovers of Tournament Teams')


grid.arrange(g1)
```


Regular Season Turnovers of Tournament Teams
density
0
5
10
15
0.50
0.55
0.60
Turnovers per game
final_four
FALSE
TRUE