

# MSDS597 Final Project - College Basketball

Varun Krishnan

5/7/2021

## Goals and objectives

In this project, I wanted to investigate division 1 ncaa basketball and research what leads teams to be more successful and win games consistently. From my own interests and knowledge about college basketball, I am able to name teams such as Duke and Kentucky that have consistently had high ranked teams year after year and are able to crush the competition extremely often. What can we attribute their success to? Does looking at game stats reveal anything about how these teams play? Do they play better offense or defense? What do they specifically do on offense and defense that leads them to win more games? This is the objective of my analysis and some of the questions that I posed that I hope to answer.

## Data Source

```
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.3.0 --

## v ggplot2 3.3.3      v purrr  0.3.4
## v tibble  3.1.0      v dplyr  1.0.5
## v tidyr   1.1.3      v stringr 1.4.0
## v readr   1.4.0      v forcats 0.5.1

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

team_data <- read.csv("MTeams.csv")
reg_game_data <- read.csv("MRegularSeasonDetailedResults.csv")
tourn_game_data <- read.csv("MNCATourneyDetailedResults.csv")
```

I obtained the data for this project from kaggle in the march mania competition. There are many datasets in this competition for various types of information surrounding college basketball. The main datasets that I focused on is the MTeams dataset which has all division 1 teams, an id used throughout the project and first and last recorded season. I also used the MRegularSeasonDetailedResults and MNCATourneyDetailedResults datasets which contain historical game data including the winner, loser, stats of the winning/losing team, location, date etc. Both of these datasets are formatted the same way but the ncaa tourney set contains game data from only the annual march madness college basketball tournament, while the regular season dataset is just regular season games.

Data source link: <https://www.kaggle.com/c/ncaam-march-mania-2021/data>

## Data cleaning and wrangling

```
combined_game_data <- bind_rows(reg_game_data, tourn_game_data)
```

Since my analysis was not exclusive to the tournament or regular season data, I combined both of the datasets into one larger dataset. This gives more data to work with and examine.

## Data cleaning and wrangling - NA values in game dataset

```
colSums(is.na(combined_game_data))
```

```
## Season DayNum WTeamID WScore LTeamID LScore WLoc NumOT WFGM WFGA
##      0      0      0      0      0      0      0      0      0      0
## WFGM3 WFGA3 WFTM WFTA WOR WDR WAst WTO WStl WBlk
##      0      0      0      0      0      0      0      0      0      0
## WPF LFGM LFGA LFGM3 LFGA3 LFTM LFTA LOR LDR LAst
##      0      0      0      0      0      0      0      0      0      0
## LT0 LStl LBlk LPF
##      0      0      0      0
```

Checked for N/A values in the combined game dataset. There were none.

## Data cleaning and wrangling - NA values in teams dataset

```
colSums(is.na(team_data))
```

```
## TeamID TeamName FirstD1Season LastD1Season
##      0      0      0      0
```

Checked if any teams in the team dataset had N/A values. There were none.

## Data cleaning and wrangling - Merging datasets

```
combined_game_data <- combined_game_data %>% select(-c(Season,DayNum))
```

Since my analysis is going to be more focused about in-game statistics and finding patterns and differences that I could formulate a hypothesis from, I dropped some of the unneeded data from the games dataset. Season number and day number that a game was played on were unneeded.

## Data cleaning and wrangling - Joining team name column

```
combined_game_data <- combined_game_data %>%
  left_join(team_data, by = c("WTeamID" = "TeamID")) %>%
  select(-c(FirstD1Season, LastD1Season)) %>%
  rename(c("WTeamName" = "TeamName")) %>%
  left_join(team_data, by = c("LTeamID" = "TeamID")) %>%
  select(-c(FirstD1Season, LastD1Season)) %>%
  rename(c("LTeamName" = "TeamName")) %>%
  select(WTeamName, LTeamName, everything())
```

For each game in the games dataset, the winner and loser is recorded with their respective team id from the teams dataset. The team id acts as a primary key - foreign key between the teams and games dataset. Instead of having to cross reference the teams dataset every time when looking at the winner/loser of a game, I created a new column for the winning team name and the losing team name in the games dataset.

I left joined the team dataset to the games dataset where winning team id matched team id then dropped the other columns to make a winning team name column. I repeated except matched the losing team id to make a losing team name column.

## Wins vs. Losses

```
## This code chunk takes a few seconds to compile

wins_losses <- data.frame(matrix(ncol=3,nrow=0,
                                dimnames=list(NULL, c("name", "wins", "losses"))))

for(i in 1101:1471) {
  name = team_data %>% filter(TeamID == i) %>% select(TeamName) %>% .[[1]]
  if( nrow(combined_game_data %>% count(WTeamID == i)) == 1){
    next
  }

  wins = combined_game_data %>% count(WTeamID == i) %>%
    select(n) %>% .[[2,1]]

  if( nrow(combined_game_data %>% count(LTeamID == i)) == 1){
    next
  }

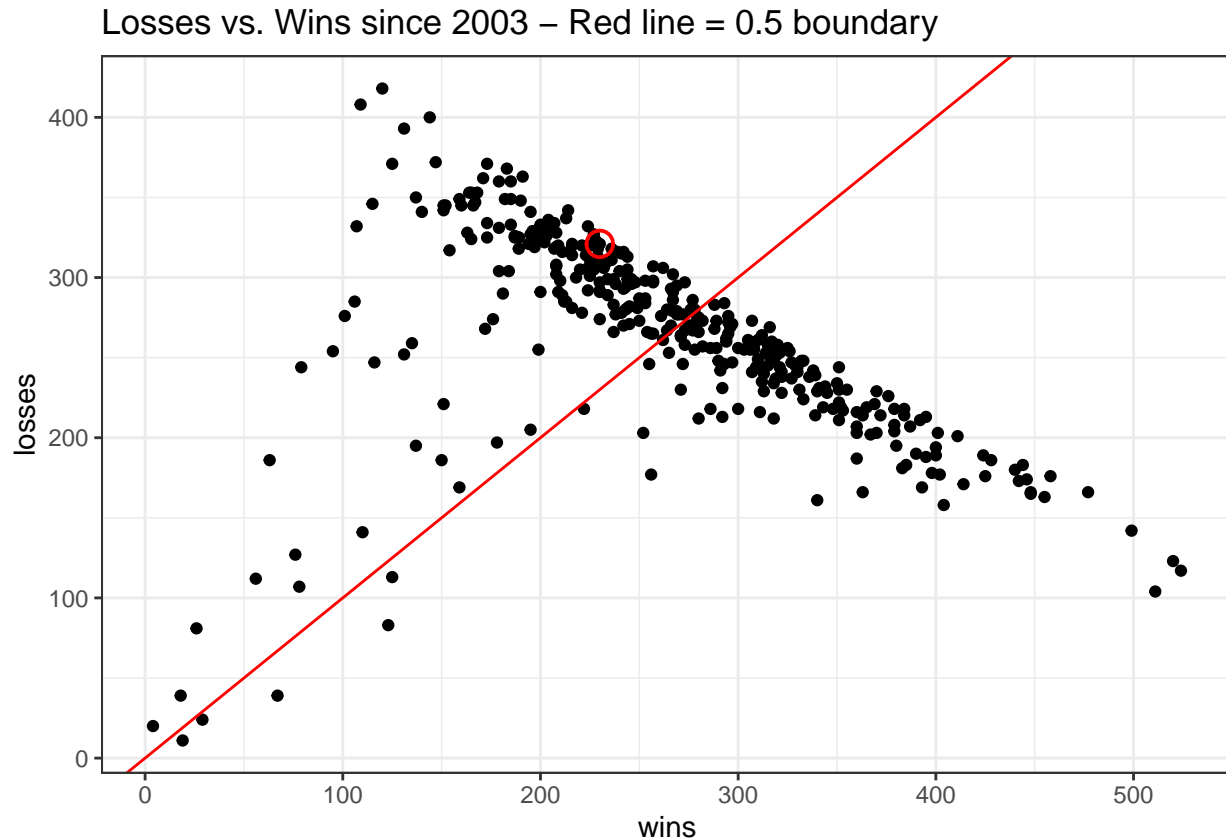
  losses = combined_game_data %>% count(LTeamID == i) %>%
    select(n) %>% .[[2,1]]

  df = data.frame(name = name, wins = wins, losses = losses)
  wins_losses <- rbind(wins_losses,df)
}
```

I wanted to start off by making a wins vs. losses graph with every team to get an idea of how the overall plot looked with teams being all compared. I looped through every team in the teams dataset and calculated the number of wins and losses they had and stored it in a separate dataframe. I had to add a few if condition checks because there were some teams that were in the teams dataset but had no recorded games. This was happening because the teams dataset recorded older teams while the games dataset tracked games up to 2003.

## Wins vs. Losses - Plot

```
library(ggplot2)
ggplot(wins_losses, aes(x=wins, y=losses)) + geom_point() +
  geom_point(data=wins_losses[wins_losses$name == "Rutgers",],
            pch=21, fill=NA, size=4, colour="red", stroke=1) +
  geom_abline(intercept = 0, slope = 1, colour = "red") +
  ggtitle("Losses vs. Wins since 2003 - Red line = 0.5 boundary") +
  theme_bw()
```



The graph looks mostly even, there seems to be more variance in the number of losses a team could have given that they have more losses than wins (left of red line). Teams that have more wins than losses are more tightly knit. I also noticed that there are three teams that have more than 500 wins and less than 150 losses, we know these are the historically very good teams since they have lost less than 150 games since 2003.

## Basketball offense vs. defense analysis

```
final_data <- combined_game_data %>%
  mutate(WFGM2 = WFGM - WFGM3) %>%
  mutate(WFGA2 = WFGA - WFGA3) %>%
  mutate(LFGM2 = LFGM - LFGM3) %>%
  mutate(LFGA2 = LFGA - LFGA3)
write.csv(final_data, "tidy-basketball-data.csv")
```

In basketball there are 2 point and 3 point shots, they are both known as a field goal. The games dataset tracks field goals and 3 point shots so in order to find how many field goals were 2 point shots, I had to take the total field goals and subtract the amount that were 3 point shots. I mutated the games dataset to get this information, I did it for shots made and shots attempted. This expanded the scoring data to now have separate columns for 2 point shots made and 2 point shots attempted for both the winning and losing team.

## Basketball offense analysis - Boxplot

```
boxplot_labels <- data.frame(statid = c("WFGA2","LFGA2",
                                       "WFGM3","LFGM3",
                                       "WFGA3","LFGA3",
                                       "WFTM","LFTM",
                                       "WFTA","LFTA",
                                       "WOR","LOR",
                                       "WAsst","LAsst"),
                             statname = c("2-point attempt","2-point attempt",
                                           "3-point made","3-point made",
                                           "3-point attempt","3-point attempt",
                                           "Foul shot made","Foul shot made",
                                           "Foul shot attempt","Foul shot attempt",
                                           "Off. Rebound","Off. Rebound",
                                           "Assist","Assist"),
                             statoutcome = c("Win","Loss",
                                              "Win","Loss",
                                              "Win","Loss",
                                              "Win","Loss",
                                              "Win","Loss",
                                              "Win","Loss",
                                              "Win","Loss"))

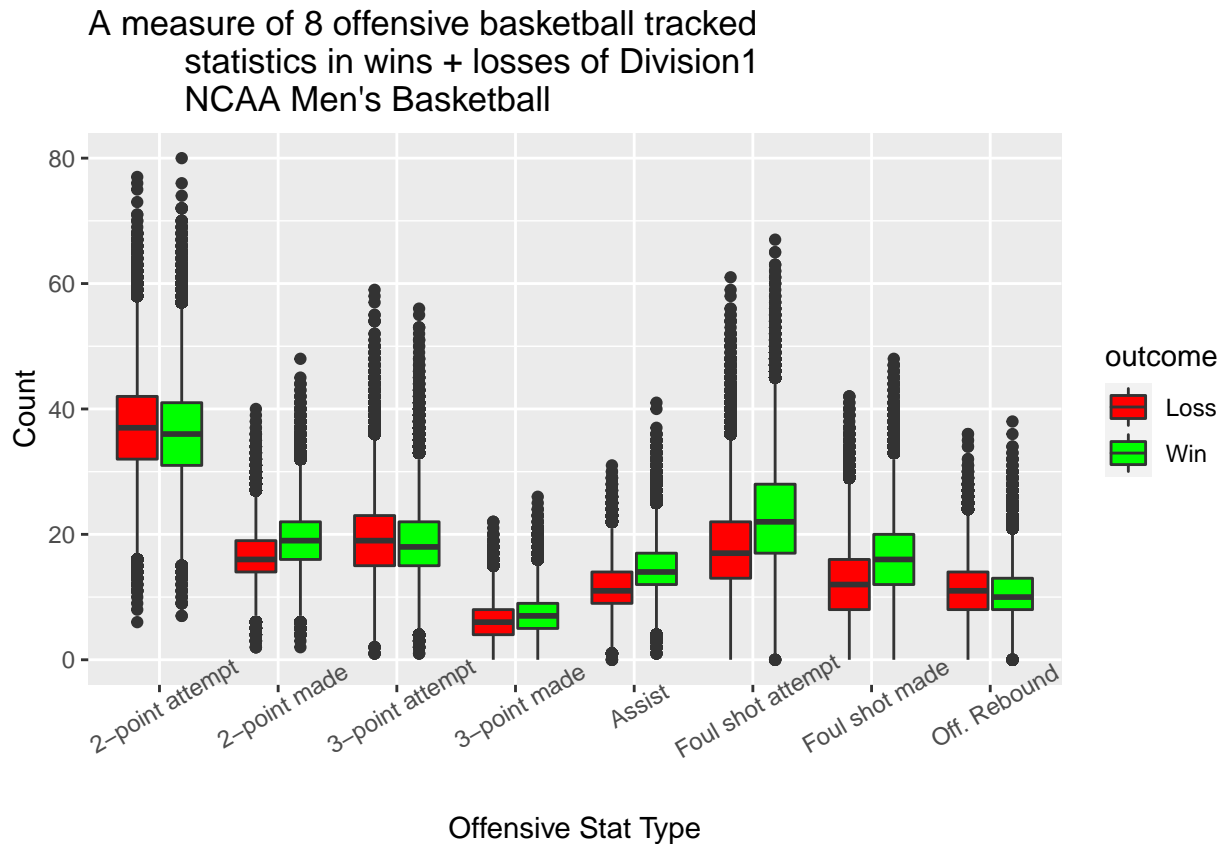
boxplot_data <- final_data %>%
  select(WFGM2) %>%
  rename(c("len" = "WFGM2")) %>%
  mutate(type = "2-point made") %>%
  mutate(outcome = "Win")

boxplot_data <- bind_rows(boxplot_data, final_data %>%
  select(LFGM2) %>%
  rename(c("len" = "LFGM2")) %>%
  mutate(type = "2-point made") %>%
  mutate(outcome = "Loss"))

for(i in seq_len(nrow(boxplot_labels))) {
  boxplot_data <- bind_rows(boxplot_data, final_data %>%
    select(boxplot_labels[i,1]) %>%
    rename(c("len" = boxplot_labels[i,1])) %>%
    mutate(type = boxplot_labels[i,2]) %>%
    mutate(outcome = boxplot_labels[i,3]))
}

ggplot(boxplot_data, aes(x=type, y=len, fill=outcome)) +
  geom_boxplot() + scale_fill_manual(values = c("red","green")) +
```

```
xlab("Offensive Stat Type") + ylab("Count") +
ggtitle("A measure of 8 offensive basketball tracked
statistics in wins + losses of Division1
NCAA Men's Basketball") +
theme(axis.text.x = element_text(angle = 30))
```



I created boxplots for 8 basketball stats on the offense side and grouped the data by Win/Loss so we could compare the spread of the same offensive stat for wins and losses. I gathered all the stats and added a group to each stat to make this plot

I noticed that the median foul shot made + attempted and assists were noticeably higher in the win group vs the loss group. While the difference in medians is not large, I interpreted it as getting foul shot attempts + making them and assists have a larger effect on winning games compared to the other offensive stats.

In basketball terms these results make sense. Being able to draw fouls from the opposing team gives you a free attempt to score some points and more assists are indicative of a team that passes the ball between one another more - harder for opposing team to defend leading to more points scored.

## Basketball defense analysis - Boxplot

```
def_boxplot_labels <- data.frame(statid = c("WSt1", "LSt1",
                                           "WBlk", "LBlk"),
                                statname = c("Steal", "Steal",
                                              "Block", "Block"),
                                statoutcome = c("Win", "Loss",
                                                "Win", "Loss"))
```

```

                                "Win", "Loss"))

def_boxplot_data <- final_data %>%
  select(WDR) %>%
  rename(c("len" = "WDR")) %>%
  mutate(type = "Def. Rebound") %>%
  mutate(outcome = "Win")

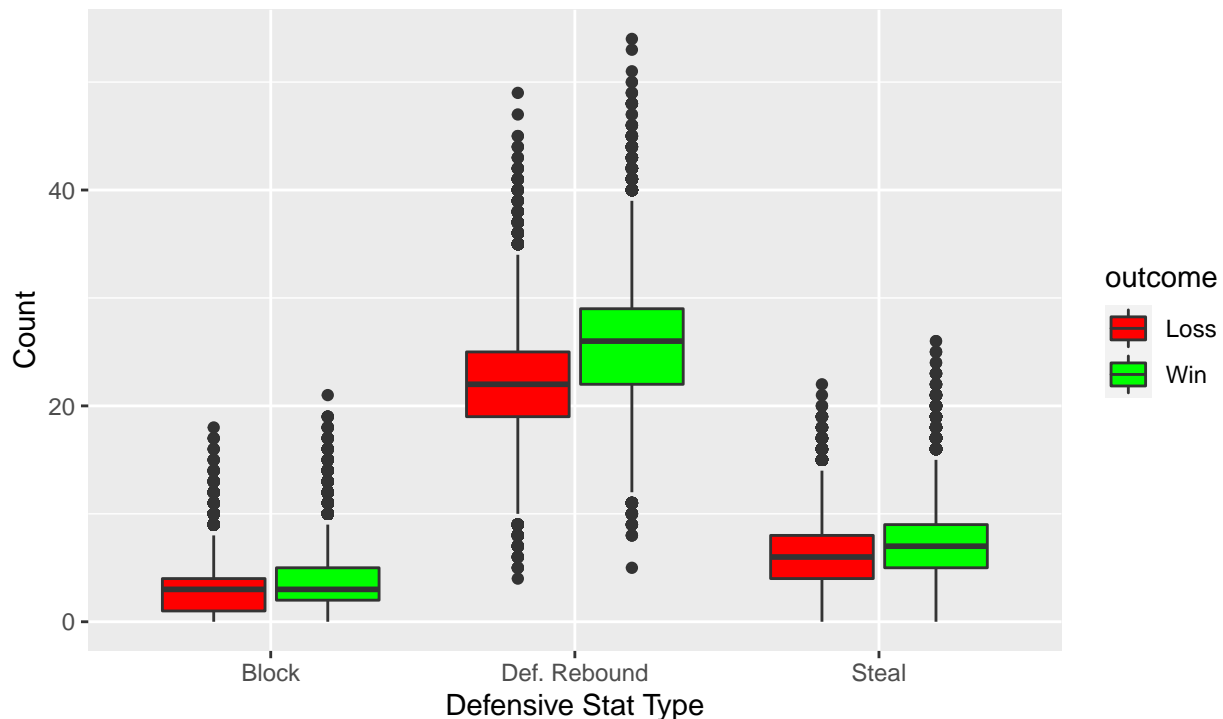
def_boxplot_data <- bind_rows(def_boxplot_data, final_data %>%
  select(LDR) %>%
  rename(c("len" = "LDR")) %>%
  mutate(type = "Def. Rebound") %>%
  mutate(outcome = "Loss"))

for(i in seq_len(nrow(def_boxplot_labels))) {
  def_boxplot_data <- bind_rows(def_boxplot_data, final_data %>%
    select(def_boxplot_labels[i,1]) %>%
    rename(c("len" = def_boxplot_labels[i,1])) %>%
    mutate(type = def_boxplot_labels[i,2]) %>%
    mutate(outcome = def_boxplot_labels[i,3]))
}

ggplot(def_boxplot_data, aes(x=type, y=len, fill=outcome)) +
  geom_boxplot() + scale_fill_manual(values = c("red", "green")) +
  xlab("Defensive Stat Type") + ylab("Count") +
  ggtitle("A measure of 3 defensive basketball tracked
          statistics in wins + losses of Division1
          NCAA Men's Basketball")

```

### A measure of 3 defensive basketball tracked statistics in wins + losses of Division1 NCAA Men's Basketball



I created the same type of grouped boxplots but for defensive basketball stats. Comparing the three, defensive rebounds definitely have a larger impact on the outcome of the game rather than steals or blocks. While its not a large difference, winning teams generally had more defensive rebounds than losing teams.

This again makes sense when thinking about basketball, when the opposing team misses a shot if they get to the ball first they have another opportunity to shoot it. A defensive rebound denies this opportunity for them, leading to the opposing team scoring less points. Blocks and steals don't indicate a difference in winning/losing with the reason being that these are very hard moves to actually perform consistently in a game, and having enough to impact winning is tough.

### Basketball offense analysis - 3 point shots

```
tp_boxplot_data <- final_data %>%
  select(WFGM3) %>%
  rename(c("len" = "WFGM3")) %>%
  mutate(type = "3-point made") %>%
  mutate(outcome = "Win")

tp_boxplot_data <- bind_rows(tp_boxplot_data, final_data %>%
  select(LFGM3) %>%
  rename(c("len" = "LFGM3")) %>%
  mutate(type = "3-point made") %>%
  mutate(outcome = "Loss"))

tp_boxplot_data <- bind_rows(tp_boxplot_data, final_data %>%
```

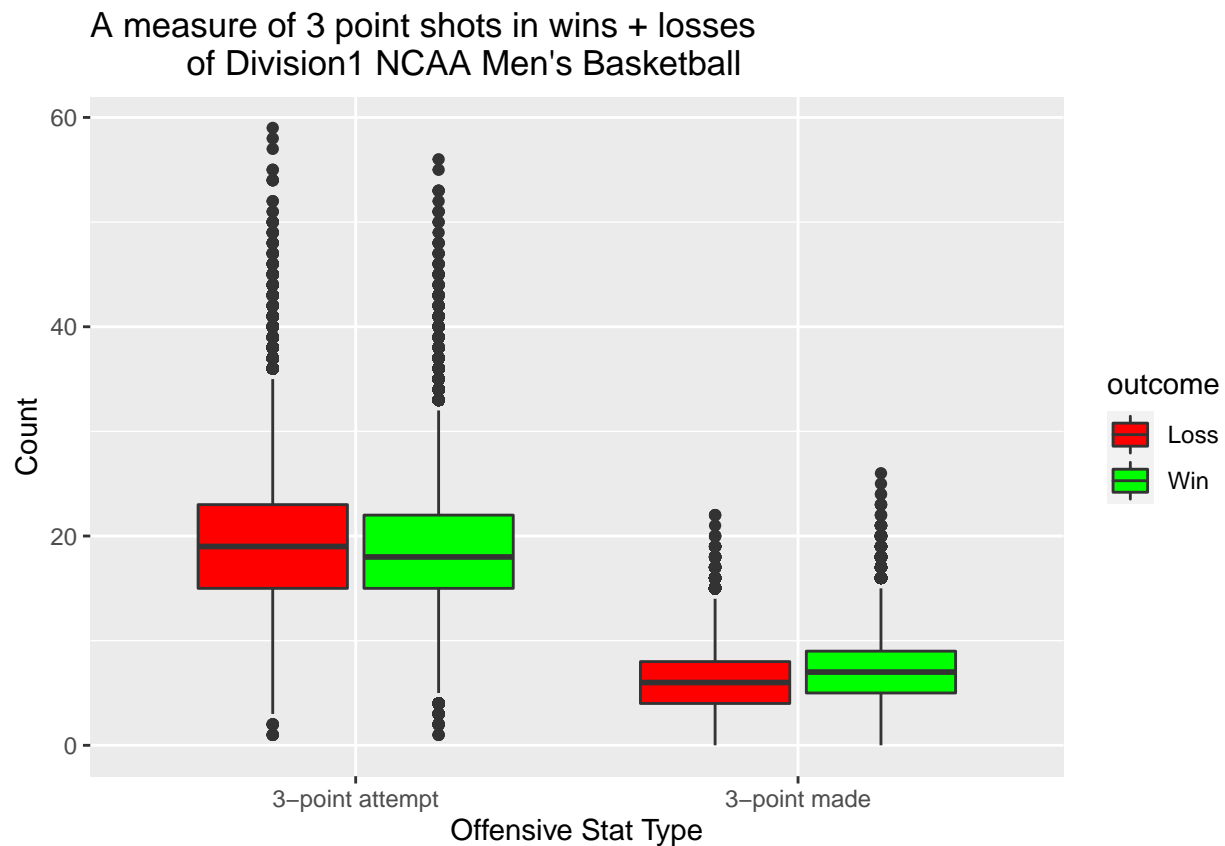
```

select(WFGA3) %>%
  rename(c("len" = "WFGA3")) %>%
  mutate(type = "3-point attempt") %>%
  mutate(outcome = "Win")

tp_boxplot_data <- bind_rows(tp_boxplot_data, final_data %>%
  select(LFGA3) %>%
  rename(c("len" = "LFGA3")) %>%
  mutate(type = "3-point attempt") %>%
  mutate(outcome = "Loss"))

ggplot(tp_boxplot_data, aes(x=type, y=len, fill=outcome)) +
  geom_boxplot() + scale_fill_manual(values = c("red", "green")) +
  xlab("Offensive Stat Type") + ylab("Count") +
  ggtitle("A measure of 3 point shots in wins + losses
    of Division1 NCAA Men's Basketball")

```

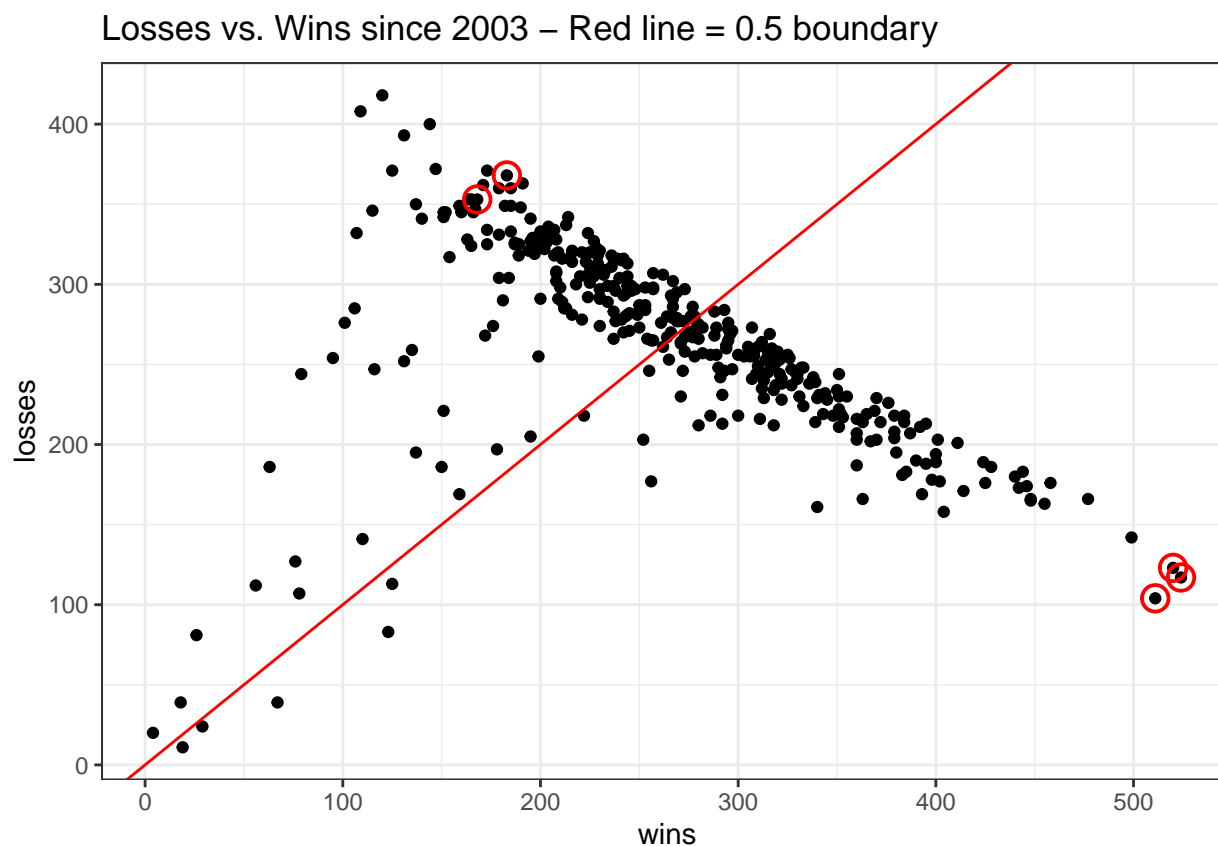


Before I started analyzing the data, I assumed that 3 point shot attempts and 3 point shots made would have a large difference and be a big factor in deciding the outcome of games. My reasoning was simply because a 3 pointer is worth more than a regular 2 point shot, it made sense that going for a shot that is worth more points should improve your chance of winning.

The boxplots showed otherwise and disproved my original assumption. The reasoning for this could be that overall no team in the country is shooting 3 pointers with a high enough success rate to impact the game in their favor. The boxplots show the median 3 point attempts was about 20 while the median 3 point shot made was about 7. This is a 35% success rate which isn't high enough to have an effect on winning.

## Analyzing specific basketball stat moves from best and worst teams

```
ggplot(wins_losses, aes(x=wins, y=losses)) + geom_point() +  
  geom_point(data=wins_losses[wins_losses$name == "Duke",],  
            pch=21, fill=NA, size=4, colour="red", stroke=1) +  
  geom_point(data=wins_losses[wins_losses$name == "Gonzaga",],  
            pch=21, fill=NA, size=4, colour="red", stroke=1) +  
  geom_point(data=wins_losses[wins_losses$name == "Kansas",],  
            pch=21, fill=NA, size=4, colour="red", stroke=1) +  
  geom_point(data=wins_losses[wins_losses$name == "Maine",],  
            pch=21, fill=NA, size=4, colour="red", stroke=1) +  
  geom_point(data=wins_losses[wins_losses$name == "MS Valley St",],  
            pch=21, fill=NA, size=4, colour="red", stroke=1) +  
  geom_abline(intercept = 0, slope = 1, colour = "red") +  
  ggtitle("Losses vs. Wins since 2003 - Red line = 0.5 boundary") +  
  theme_bw()
```



Using the gathered information about specific offensive/defensive moves from the boxplots above, I compared these moves between some of the best and worst schools and the overall national average.

I circled the location of the teams I picked on the wins/losses plot to get an idea of where these teams stand.

## Analyzing specific basketball stat moves from best and worst teams - Barplot

```
bar_data <- data.frame(matrix(ncol=3,
                              nrow=0,
                              dimnames=list(NULL, c("School",
                                                      "Stat",
                                                      "Count"))))

schools_to_plot <- c("Duke",
                     "Kansas",
                     "Gonzaga",
                     "Maine",
                     "MS Valley St")

bar_data <- rbind(bar_data,
                  df = data.frame(School = "Overall",
                                  Stat = "Def. Rebound",
                                  Count = mean(final_data$WDR)))

bar_data <- rbind(bar_data,
                  df = data.frame(School = "Overall",
                                  Stat = "Assist",
                                  Count = mean(final_data$Wast)))

bar_data <- rbind(bar_data,
                  df = data.frame(School = "Overall",
                                  Stat = "Foul shot made",
                                  Count = mean(final_data$WFTM)))

bar_data <- rbind(bar_data,
                  df = data.frame(School = "Overall",
                                  Stat = "Foul shot attempt",
                                  Count = mean(final_data$WFTA)))

for(school in schools_to_plot){
  temp <- final_data %>% filter(WTeamName == school)

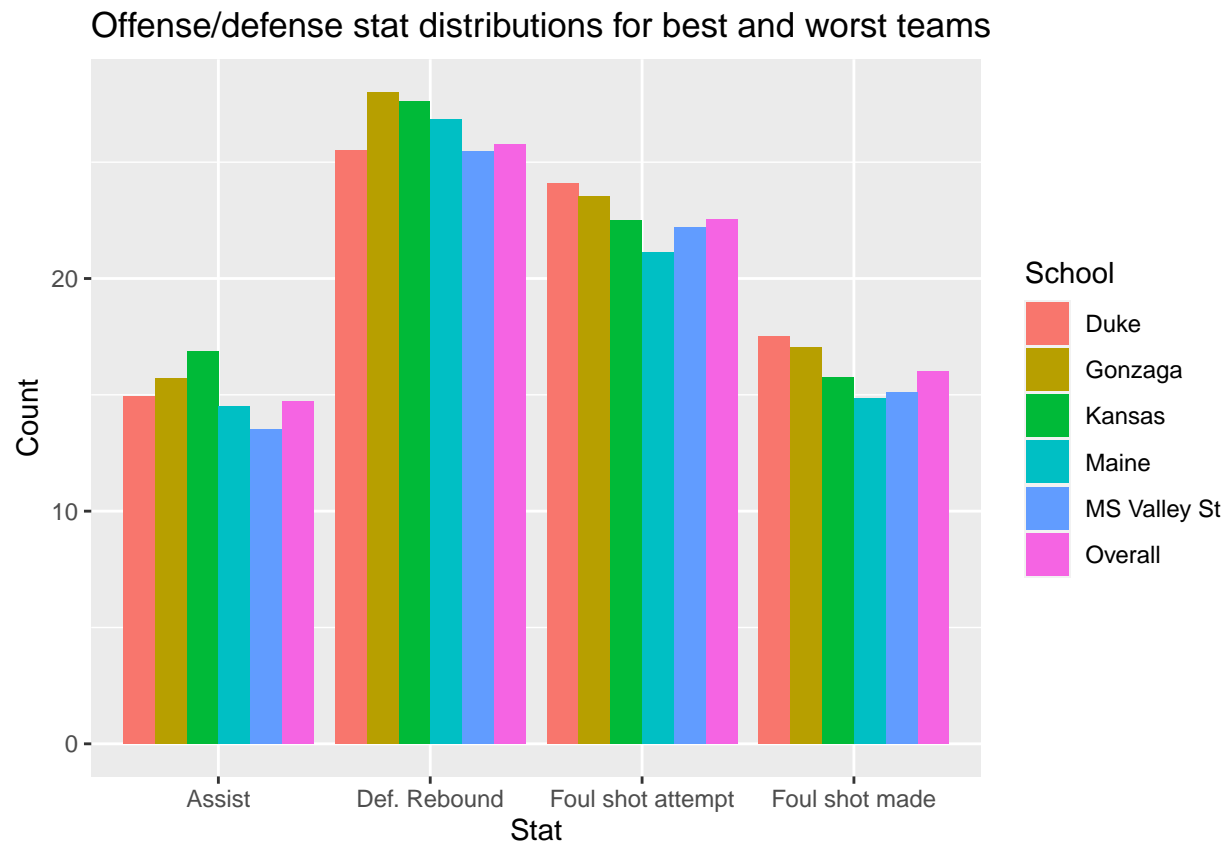
  bar_data <- rbind(bar_data,
                    df = data.frame(School = school,
                                    Stat = "Def. Rebound",
                                    Count = mean(temp$WDR)))

  bar_data <- rbind(bar_data,
                    df = data.frame(School = school,
                                    Stat = "Assist",
                                    Count = mean(temp$Wast)))

  bar_data <- rbind(bar_data,
                    df = data.frame(School = school,
                                    Stat = "Foul shot made",
                                    Count = mean(temp$WFTM)))

  bar_data <- rbind(bar_data,
                    df = data.frame(School = school,
                                    Stat = "Foul shot attempt",
                                    Count = mean(temp$WFTA)))
}
```

```
ggplot(data=bar_data, aes(x=Stat, y=Count, fill=School)) +
  geom_bar(stat="identity", position=position_dodge()) +
  ggtitle("Offense/defense stat distributions for best and worst teams")
```



I plotted the mean amount of those moves and grouped by school. I chose three of the best teams and two of the worst teams and the overall average across all schools.

Our best schools, Duke, Gonzaga and Kansas are at or above the overall average for all 4 of the stats. The two not so good schools, Maine and MS Valley St, are at or below the overall average except for Maine with Def. rebounds. Foul shot attempt and foul shot made are highly correlated and cannot be seen as two independent basketball moves.

## Conlusions and further work

While there are differences suggesting that better teams will have slightly larger amounts of assists, def rebounds and foul shots on average, it is not a huge difference. Basketball is a complicated team sport, there are a lot of factors that influence the outcome that are not quantifiable, such as coaching and how well teammates work with each other. While having more points at the end is how you win, the number stats don't tell the full story which is why I believe the observed differences in comparison to win rate are not as large as initially expected.

Further work for the future could include more team specific analysis. In my analysis, I compared the best offensive/defensive basketball moves to win against national averages. What might yield better results is for example finding the top 10 teams that have the highest average assists per game and looking at how high their win rates are.

Another way I would go more in depth is to examine the data but over splits, where each split represents a time period of 2-3 years. In basketball there are only 5 players per team on the court at once, so the skill level of every individual player is very important. A very good 3 point shooter player might join a college and win a lot of games, but will eventually leave college. Since I compared averages over a long time period, possible effects like this were not revealed.

Overall this was a fun project and being able to work on an analysis about my hobbies and interests was very enjoyable. Thank you for a great semester!