Final Project

Raashi Dewan, Corey Huang, Lauren Manning, Grace Stewart, Armando Taveras

Contents

Reading CSV File

Below we read in the .csv file titled "cbb.csv". This file contains basketball statistics for different college teams among different conferences from 2013 to 2019. We have chosen to exclude data from 2020, as that season was cut short and does not include post-season statistics due to the COVID-19 pandemic. We also chose to exclude 2021 data, as the season is still in progress and thus the dataset is not completed yet.

```
bball <- read.csv(file = "cbb.csv")
head(bball)
```

```
##                 TEAM CONF  G  W ADJOE ADJDE BARTHAG EFG_O EFG_D  TOR TORD  ORB  DRB
## 1 North Carolina  ACC 40 33 123.3  94.9  0.9531  52.6  48.1 15.4 18.2 40.7 30.0
## 2      Wisconsin  B10 40 36 129.1  93.6  0.9758  54.8  47.7 12.4 15.8 32.1 23.7
## 3       Michigan  B10 40 33 114.4  90.4  0.9375  53.9  47.7 14.0 19.5 25.5 24.9
## 4     Texas Tech  B12 38 31 115.2  85.2  0.9696  53.5  43.0 17.7 22.8 27.4 28.7
## 5        Gonzaga  WCC 39 37 117.8  86.3  0.9728  56.6  41.1 16.2 17.1 30.0 26.2
## 6       Kentucky  SEC 40 29 117.2  96.2  0.9062  49.9  46.0 18.1 16.1 42.0 29.7
##    FTR FTRD X2P_O X2P_D X3P_O X3P_D ADJ_T  WAB POSTSEASON SEED YEAR
## 1 32.3 30.4  53.9  44.6  32.7  36.2  71.7  8.6        2ND    1 2016
## 2 36.2 22.4  54.8  44.7  36.5  37.5  59.3 11.3        2ND    1 2015
## 3 30.7 30.0  54.7  46.8  35.2  33.2  65.9  6.9        2ND    3 2018
## 4 32.9 36.6  52.8  41.9  36.5  29.7  67.5  7.0        2ND    3 2019
## 5 39.0 26.9  56.3  40.0  38.2  29.0  71.5  7.7        2ND    1 2017
## 6 51.8 36.8  50.0  44.9  33.2  32.2  65.9  3.9        2ND    8 2014
```

Cleaning the Data

In order to increase efficiency and data cleanliness, we removed all columns that were not relevant to the questions that we would like to focus on (which are detailed below).
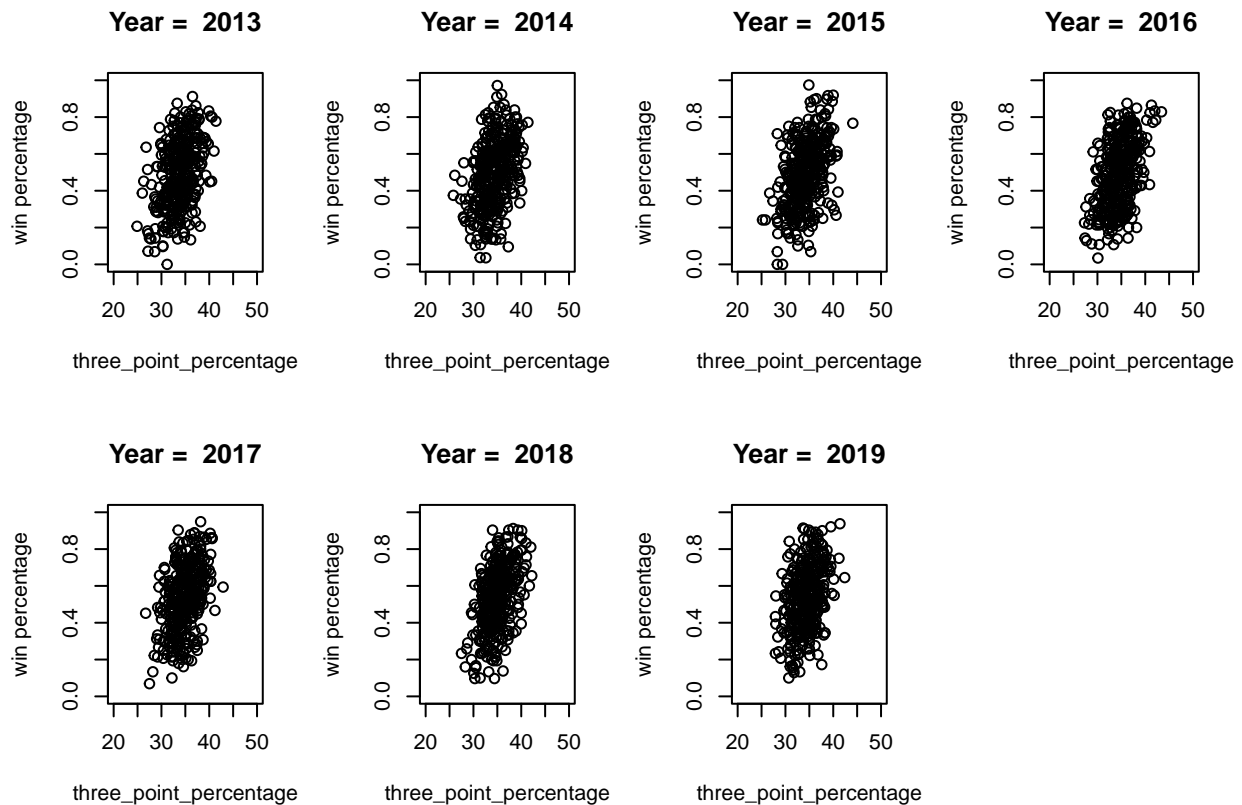
```
bball = bball %>% select("TEAM" | "CONF" | "ADJOE" | "ADJDE" | "W" | "G" | "X3P_D" | "X3P_O" | "YEAR")
head(bball)
```

```
##                TEAM CONF ADJOE ADJDE  W  G X3P_D X3P_O YEAR
## 1 North Carolina  ACC 123.3  94.9 33 40  36.2  32.7 2016
## 2      Wisconsin  B10 129.1  93.6 36 40  37.5  36.5 2015
## 3       Michigan  B10 114.4  90.4 33 40  33.2  35.2 2018
## 4     Texas Tech  B12 115.2  85.2 31 38  29.7  36.5 2019
## 5        Gonzaga  WCC 117.8  86.3 37 39  29.0  38.2 2017
## 6       Kentucky  SEC 117.2  96.2 29 40  32.2  33.2 2014
```

What type of relationship exists between three point shot percentages and win percentages?

In order to examine at the question above, we first looked at each team's three point percentage versus their win percentage per year in order to get a visualization or whether there is a relationship between the two. From the results below, you can see that there does appear to be a bit of a positive relationship between the two factors.

```
x <- c(2013, 2014, 2015, 2016, 2017, 2018, 2019)
y <- c(1:nrow(bball))
par(mfrow=c(2,4))
for (yr in x)
{
  three_point_percentage = c(NA)
  wins = c(NA)
  for (i in y)
  {
    if(bball[i,'YEAR'] == yr )
    {
      newTP <- bball[i,'X3P_O']
      newW <- bball[i,'W'] / bball[i,'G']
      three_point_percentage <- c(three_point_percentage, newTP)
      wins <- c(wins, newW)
    }
  }
  plot(three_point_percentage,wins, xlim=c(20,50),ylim=c(0,1),lty=1, ylab="win percentage", main = paste
}
```

Year = 2013   **Year = 2014**   **Year = 2015**   **Year = 2016**

Year = 2017   **Year = 2018**   **Year = 2019**

Question 1 - Continued

Wanting to look at this question from another angle, we then examined two graphs side-by-side : win percentages by conference and three point percentages by conference. By cross referencing these, you can make the same observation as we did in the section above - that there is a positive relationship between three point percentages and win percentages.

```
par(mfrow=c(1,2))
boxplot(W/G ~ CONF, bball, las = 2, cex.axis = 0.5)
boxplot(X3P_O ~ CONF, bball, las = 2, cex.axis = 0.5)
```
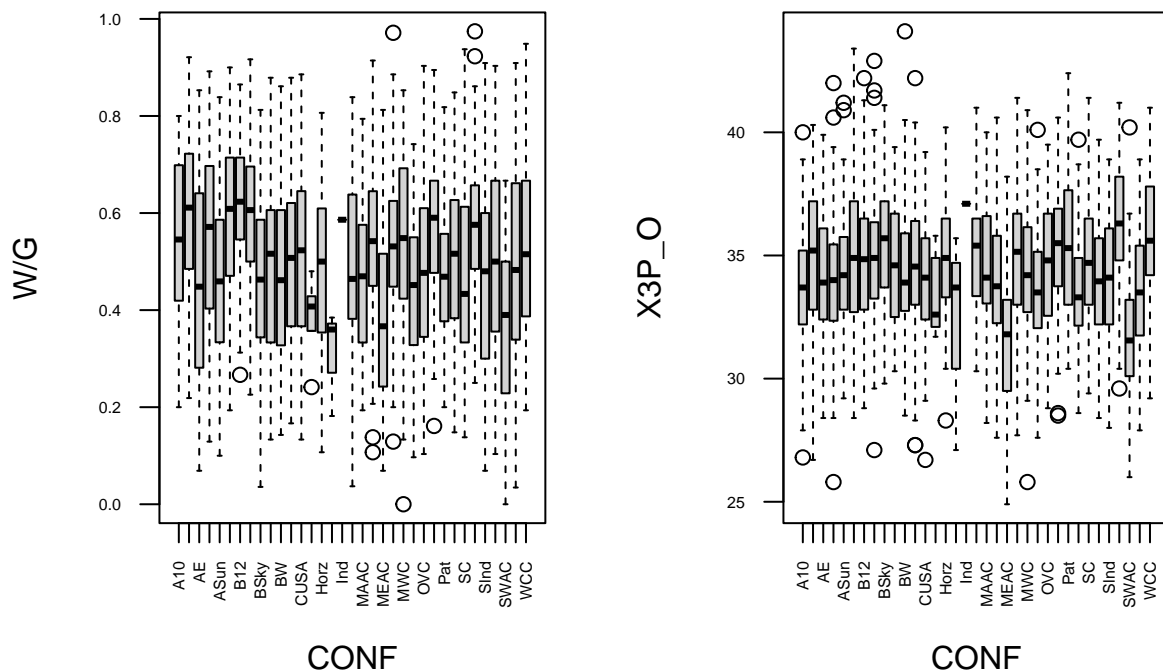
Does offensive efficiency or defensive efficiency better correlate to higher winning percentages?
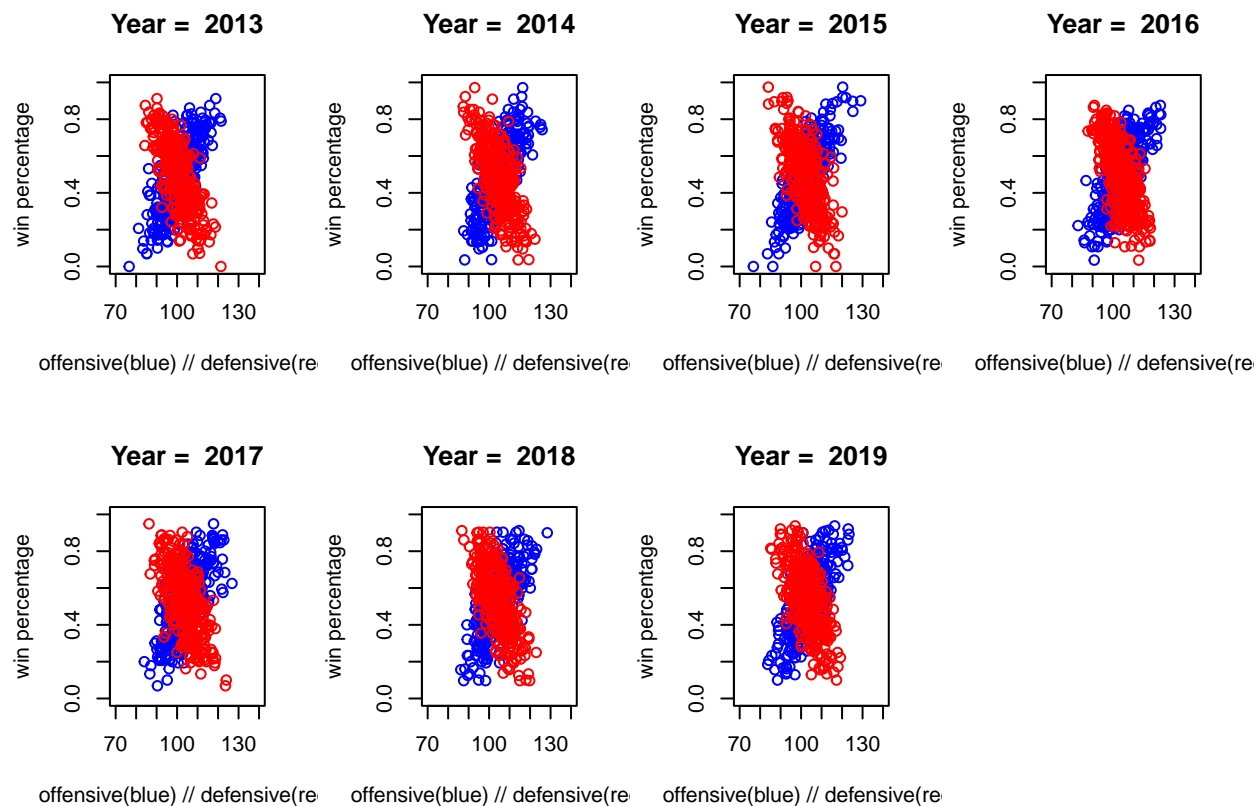
In order to examine this question, we graphed adjusted offensive efficiency versus wins percentage in blue, and adjusted defensive efficiency versus wins percentage in red, all per year. We overlapped these graphs in order to best show the very interesting relationship seen below. While a higher offensive efficiency clearly results in a higher winning percentage, a higher defensive efficiency results in a lower winning percentage. As a result, one would say that a higher offensive efficiency better correlates to a higher winning percentage.

```r
x <- c(2013, 2014, 2015, 2016, 2017, 2018, 2019)
y <- c(1:nrow(bball))
par(mfrow=c(2,4))
for (yr in x)
{
  offensive = c(NA)
  wins = c(NA)
  defensive = c(NA)
  for (i in y)
  {
    if(bball[i,'YEAR'] == yr )
    {
      newO <- bball[i,'ADJOE']
      newD <- bball[i, 'ADJDE']
      newW <- bball[i,'W'] / bball[i,'G']
      offensive <- c(offensive, newO)
```

```
    defensive <- c(defensive, newD)
    wins <- c(wins, newW)
  }
}
plot(offensive,wins, xlim=c(70,140),ylim=c(0,1),lty=1, col='blue',ylab="win percentage", xlab="offensi
points(defensive,wins, xlim=c(70,140),ylim=c(0,1),lty=1, col='red')
}
```

Year = 2013   **Year = 2014**   **Year = 2015**   **Year = 2016**

offensive(blue) // defensive(re   offensive(blue) // defensive(re   offensive(blue) // defensive(re   offensive(blue) // defensive(re

Year = 2017   **Year = 2018**   **Year = 2019**

offensive(blue) // defensive(re   offensive(blue) // defensive(re   offensive(blue) // defensive(re

Conclusions

In order to evaluate what strategies would be most effective for basketball teams to have a higher probability to win, we analyzed different relationships between statistics found in the CSV file and win percentages. These statistics are retrieved from the datasets of 355 college basketball teams from the 2013 to 2019 seasons.

First, we explored the relationship between the three-point shot percentage and win percentage for each team. Looking at the scatter plots for three-point percentage vs win percentage of all teams each year, you can see the overall shape of the graph going slightly in a positive direction. Comparing the boxplots of win percentages by conference with three-point percentages by conference, it is also possible to see a positive correlation by just looking at the medians: most conferences that have a higher median for three-point percentages also have a higher median for win percentages. From these observations, we can conclude that there is a slight but confident positive relationship between three-point percentages and win percentages.

The second relationship we analyzed was which efficiency would be a better correlation to a higher win percentage: offense or defense? Surprisingly, in the scatter plots for each year, we found that while having a better offense (blue) does have a positive relationship with win percentage, having a better defense (red) actually has a negative relationship. This is clearly evident since in the graph the shape of the offensive

efficiency vs win percentage goes in a positive direction while the shape of the defensive efficiency vs win percentage goes in a negative direction.

Based on the data we analyzed, we can first rule out the strategy of having a better defensive efficiency since not only is it ineffective with helping teams win, it decreases the win percentage as well. Although the relationships of three-point percentages vs win percentages and offensive efficiency vs win percentages are both positive, having a better offense has a more effective result. In conclusion, employing a strategy of having a higher three-point percentage and/or offensive efficiency would be more effective for a basketball team to have a higher win percentage, with a better offense being the most effective.