Project Part 3

Introduction

Data Description and Question

Each year, the division one college basketball tournament is played in March. The event, more well known as "March Madness," is a tournament that brings in fans from a national level in the United States. Fans across the country make predictions of the tournament bracket before the first game and compete with each other on the accuracy of their brackets.

The seeding of the tournament depends on each team's regular season record, and a higher seed (lower numbering) is given to a team that is better than a lower seeded team. In recent trends of college basketball and basketball at all levels, scoring efficency and increased three point shooting volume have lead to increased scoring averages. Assuming that better teams recruit more efficient players that can score more, and other teams do not have the luxury of high volume efficient scorers, these assumptions can lead to a question that fans could ask about tournament trends from year-to-year: Has the scoring difference between higher seeded teams and lower seeded teams been widened? The answer to this question can provide insight on whether the increase in three point shooting and efficiency scoring increases the scoring gap between higher and lower seeded teams.

The dataset for this report is a comprehensive NCAA March Madness Basketball Tournament game results .csv file from 1985 to 2019. The data is provided by user michaelaroy from data.world. The dataset begins at 1985 because that is the first year the tournament expanded to 64 teams, from 53 teams in the previous year's format. This time cutoff allows the dataset to be consistent in all features such as number of regions, number of teams, number of games, and number of rounds played.

Data Relevance

Each observation (row) of the dataset is a game played in a given tournament, providing data on features such as seeding of the two teams and the score of the respective teams. These two features can give a basic overview of the scoring of the game and can be used to determine which teams score more in games. Using seeding as ranking for the teams and score difference to determine the winner of the respctive games, a test can be conducted to

find if an average higher seed can be indicative of scoring more points than an average lower seed.

Generalization

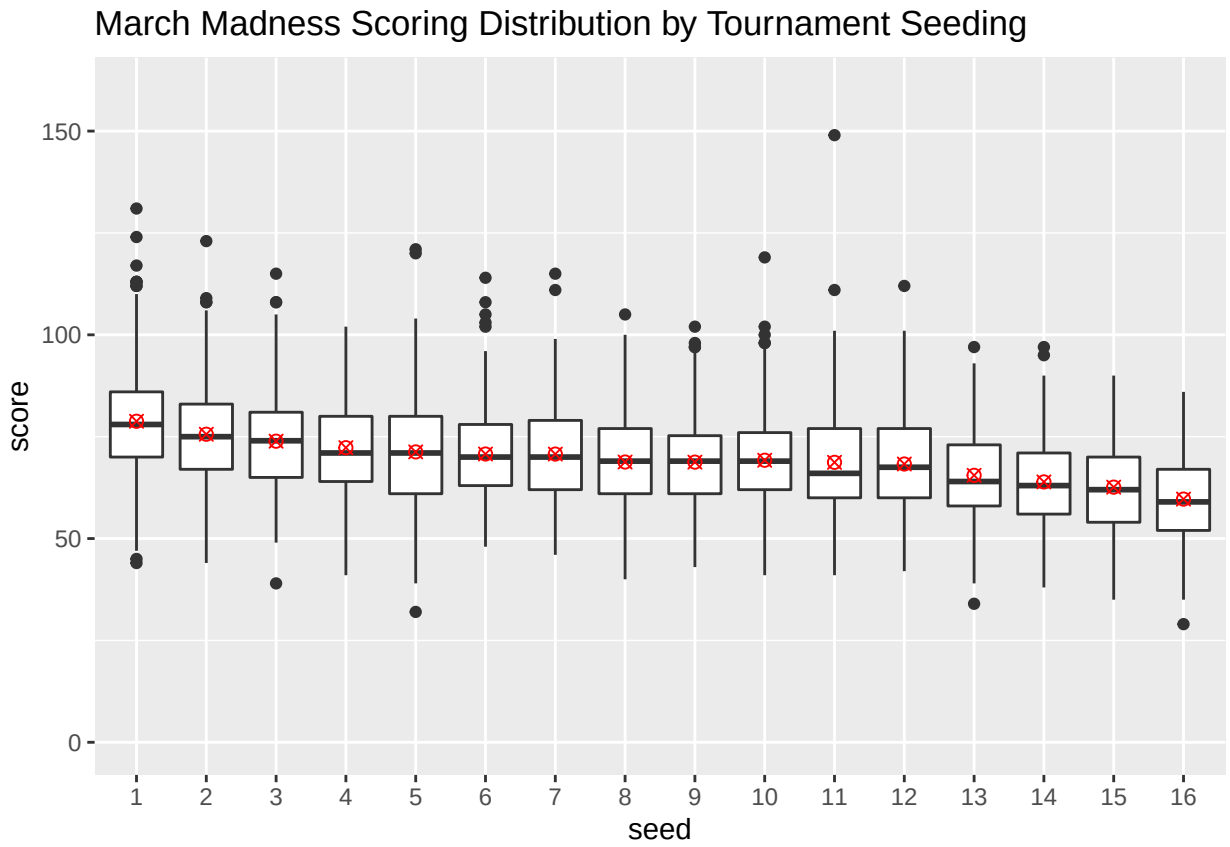
The outcome of the test would be able to tell whether the increase in volume shooting has changed the scoring landscape of college basketball. The result can be interpreted as a trend of future developent in scoring in college basketball tournaments. If the test rejects the null hypothesis and is significant, then it will mean that better programs are recruiting better players to score more in comparison to lower seeded teams in the tournament. If the test does not reject the null hypothesis and is significant, that means college basketball's scoring gap between higher seeds and lower seeds is not increasing.

Method

Test Determination Through Data Exploration

Before testing, the dataset should be evaluated on whether higher seeded teams do score more than lower seeded teams in general, passing an eye test.

```
seed.score <- data.frame(seed = c(bigdance[,"Seed"],bigdance[,"Seed.1"]),
                         score = c(bigdance[,"Score"],bigdance[,"Score.1"]),
                         year = c(bigdance[,"Year"]))
seed.score$seed <- as.factor(seed.score$seed)
ggplot(seed.score, aes(x=seed, y=score)) +
  geom_boxplot() +
  ylim(min=0, max=160) +
  ggtitle("March Madness Scoring Distribution by Tournament Seeding") +
  stat_summary(fun.y="mean", color="red", geom = "point", shape=13, size=2)
```




```
# Reference 1, 2
```

The graph created above shows that higher seeded teams definitely score more points (mean and median tested) than lower seeded teams. I decided to find the median seeding in all games in the tournaments to conduct a two-sample z-test, which gives a fair split between appearances of higher seeded teams and lower seeded teams in the tournament. This method is decided on because higher seeded teams usually win more games, which can lead to disproportionate amounts of higher seeded teams if the seeding was cut at 8 and 9.

I decided to use 2013 as the year to compare recent basketball scoring in touraments and compare with the population of scoring averages in years before 2013. The reason behind this is because this is the year Stephen Curry, one of the most efficient volume three point shooters in basketball history, appeared in his first all star game in the NBA (Reference 3). He is seen as a historic player who changed the game by shooting more threes at high efficiency and volume. Many basketball players now have modeled their game after Curry, which in turn should develop more efficient scoring in college basketball.

Conducing Two-Sample Z Test

```
seed.score$seed <- as.numeric(seed.score$seed)
median.seed <- median(seed.score$seed)
higherseed <- data.frame(seed.score[seed.score$seed <= median.seed,])
lowerseed <- data.frame(seed.score[seed.score$seed > median.seed,])
high1319 <- which(higherseed$year > 2013)
low1319 <- which(lowerseed$year > 2013)
highbefore13 <- which(higherseed$year <= 2013)
lowbefore13 <- which(lowerseed$year <= 2013)
seedhigh <- higherseed$score[high1319]
seedlow <- lowerseed$score[low1319]
stdhigh <- sd(higherseed$score[highbefore13])
stdlow <- sd(lowerseed$score[lowbefore13])
meanhigh <- mean(higherseed$score[highbefore13])
meanlow <- mean(lowerseed$score[lowbefore13])
meanbefore13 <- meanhigh-meanlow
z.test(seedhigh, seedlow, mu=meanbefore13, sigma.x=stdhigh, sigma.y=stdlow,
       alternative="greater")
```

```
##
##  Two-sample z-Test
##
## data:  seedhigh and seedlow
## z = -2.1512, p-value = 0.9843
## alternative hypothesis: true difference in means is greater than 7.558415
## 95 percent confidence interval:
##  4.092363       NA
## sample estimates:
## mean of x mean of y
##  73.22025  67.62604
```

The null hypothesis of the test is that the scoring average between higher seeded teams and lower seeded teams have stayed the same from before 2013 and after 2013. The alternative hypothesis is that the scoring average is greater between higher seeded teams and lower seeded teams from before 2013 and after 2013. After conducting the test, it was surprising

to find that the average score difference between higher and lower seeded teams have become narrower. So I decided to conduct a second two-sample z test.

```
z.test(seedhigh, seedlow, mu=meanbefore13, sigma.x=stdhigh, sigma.y=stdlow,
       alternative="less")
```

```
##
##  Two-sample z-Test
##
## data:  seedhigh and seedlow
## z = -2.1512, p-value = 0.01573
## alternative hypothesis: true difference in means is less than 7.558415
## 95 percent confidence interval:
##        NA 7.096066
## sample estimates:
## mean of x mean of y
##  73.22025  67.62604
```

In the second test, the alternative is the the scoring average is more narrow after 2013 in comparison to before 2013. The p-value of the test is below 0.05, which means the test is statistically significant, and the true difference in average of scoring difference after 2013 is less than the population mean before 2013.

Test Conclusion

The initial hypothesis that I made about scoring gap between higher seeded teams and lower seeded teams have become wider since 2013 has been proven not true. So a second test which tests if the scoring gap has become narrower was conducted, which proved to be statistically significant, meaning through the test, a trend was found and can be applied into the general population. This second test statistically found a trend that recent scoring averages between higher and lower seeded teams have become narrower, which means the competitive landscape of college basketball has become even more competitive, as lower seeded teams are getting closer to how much the higher seeded teams are scoring on average.

References

1. http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visuali
2. https://stackoverflow.com/questions/40830832/add-one-column-below-another-in-a-data-frame-in-r
3. https://www.basketball-reference.com/players/c/curryst01.html