Project Part I

Haoxuan Derek Liu (hdl5hz)

Reading Data File

```
## Read original csv file for 1985-2019 NCAA Basketball Tournament. ##

bigdance <- read.csv("Big_Dance_CSV.csv")
```

Data Description

Background Information

The Big_Dance_CSV.csv dataset is provided by data.world's user michaelaroy. The dataset includes all NCAA March Madness Tournament game result information from 1985 to the current year, 2019. 1985 was the first year in which the tournament expanded to 64 teams, from 53 teams in the previous year's tournament. The format of the tournament remains the same throughout 2019, making the dataset consistent in terms of number of regions, number of rounds, number of teams, and number of games each year.

March Madness is an annual Division I collegiate basketball single-elimination tournament, deciding the national champion of Division I basketball.

Data Organization

```
dim(bigdance)
```

```
## [1] 2205   10
```

```
names(bigdance)
```

```
##  [1] "Year"          "Round"         "Region.Number" "Region.Name"
##  [5] "Seed"          "Score"         "Team"          "Team.1"
##  [9] "Score.1"       "Seed.1"
```

The dataset is organized as a time series. Each row represents a game played in a tournament, with 10 varaibles recorded: Year, Round, Region Number, Region Name, Seed (Team 1), Score (Team 1), Team (Team 1), Team.1 (Team 2), Score.1 (Team 2), Seed.1 (Team 2). Each of the 10 column variables / features are self explanatory, as they describe what is recorded in the column. Region Name, Team and Team.1 are Factors as they are recorded using characters, the rest of the variables are integers. From top to bottom in terms of indexing, each row is exactly ordered by year, round, region and lastly by the higher seeded team's seeding number. A total of 2205 games are recorded, as indicated by the dimensions of the dataframe 'bigdance'. A total of 63 games are played in each year's tournament, and there are 35 years in between 1985 - 2019, which multiplies to 2205, confirming that there are no missing games in the dataset.

Data Collection

The dataset represents the population data between 1985-2019, as all games under the tournament are recorded in the set. The collection of the data is most likely from the box score of each game, while adding explanatory elements such as round or region for each game. The author did not specify their source of the data, an educated guess would be the data is scrapped from a major sports statistics website such as espn, fivethirtyeight, or the NCAA basketball official website's records.

Potential Issues

Looking at the general tournament rules of the March Madness tournament, there are a few issues that arise from the dataset that could potentially skew the data points and its calculations.

The first issue at hand is that since 2011, the NCAA added 4 more teams into the tournament, in which these 4 teams compete in another earlier round of single elimination game to secure their seeding in the 64 team tourney. The 4 added teams' games are similar to Major League Baseball's playoff Wildcard game. The dataset, however, does not include these games from 2011-2019, which means there are a handful of pre-tournament games that are missing from the set. By definition, the four teams and their games are a part of the tournament, but not in the round of 64. So if calculations are made, it should be made abundantly clear that these games were not used in methods applied later.

Another issue that could cause potential issues in future method applications could be the column names. Since each game (row) has two teams playing against each other head to head, there are two team columns, two score columns and two seeding columns. I must be very careful not to mix these columns up when grouping or indexing to cause calculation errors. A potential solution to this is to make sure the column names are mutated to distinct names in the future.

A third issue is that the dataset does not account for rule changes during the period of 1985-2019, such as the three point line moving further away, which could be an underlying factor that is not recorded in the dataset, that could affect the trends of the data.

Data Presentation

Numerical Representation

```
head(summary(bigdance$Team), 5)
```

```
##           Duke North Carolina         Kansas       Kentucky        Arizona 
##            102             97             90             85             62 
```

```
head(summary(bigdance$Team.1), 5)
```

```
## Michigan St      Xavier    Syracuse      Kansas      Temple 
##          31          31          30          27          27 
```

```
summary(bigdance$Score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    32.0    65.0    74.0    74.3    82.0   131.0 
```

```
summary(bigdance$Score.1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   29.00   59.00   67.00   68.05   76.00  149.00 
```

The first function displays the top 5 teams that appear in the tournament as higher seeds in each game. As the output shows, the top 4 appearing teams are Duke, UNC, Kansas and Kentucky. These 4 teams are all ‘Blue Bloods’ in college basketball, having historically been strong basketball programs, which makes sense for them appear as the top seeded teams in games of tournaments.

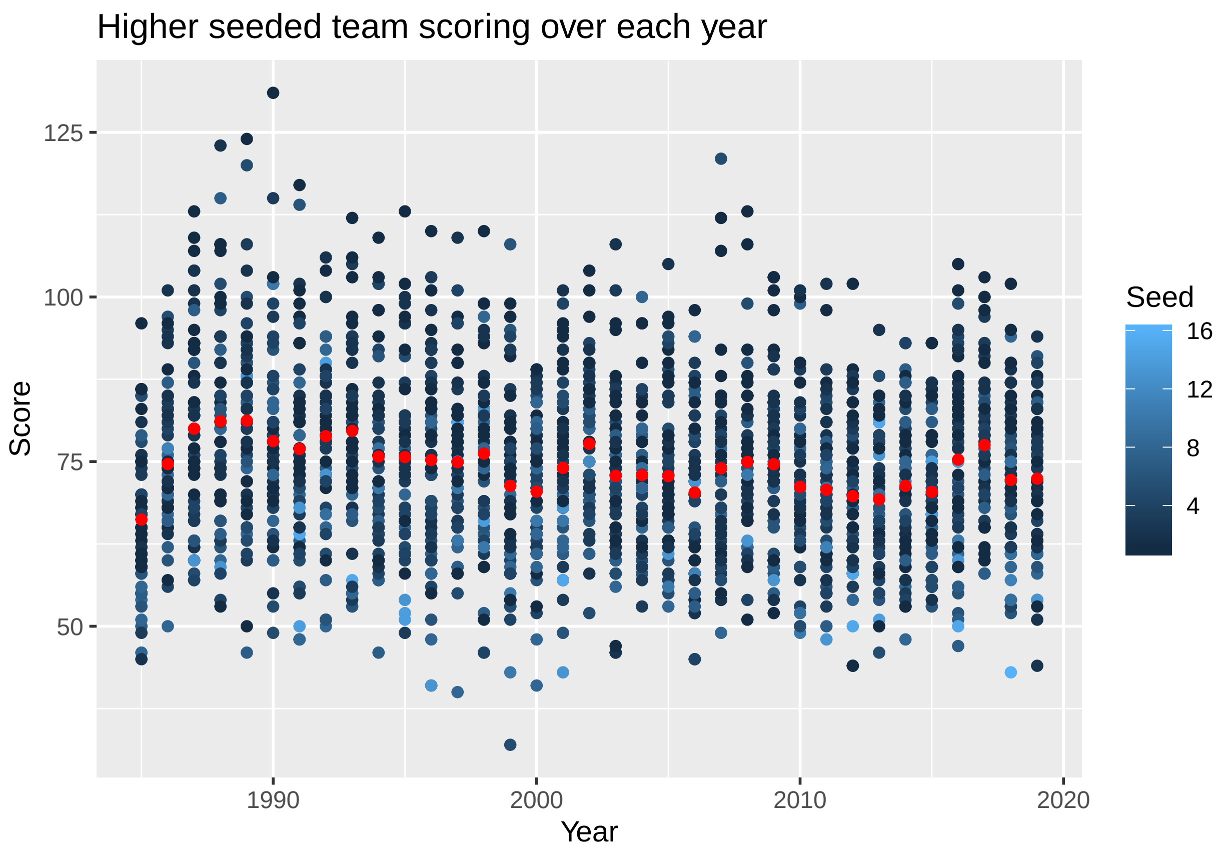
The second functions is the same as the first function, instead it displays the lower seeded team that appears the most in games of all tournaments. These teams are not necessarily not good in the tournament, they may just be facing more difficult opponents. As the output shows, Kansas is also included in this top 5. This shows that being in the second column does not necessarily mean the team isn't good, as they could be a 2 seed that appear in later rounds and play against 1 seeds frequently over the years.

The third function shows the quartile range, median and mean of the of the final scores of teams with the higher seed. This is useful to gauge the ability of the team with the higher seed in each game, and is even more useful when used as comparison with the lower seeded team.

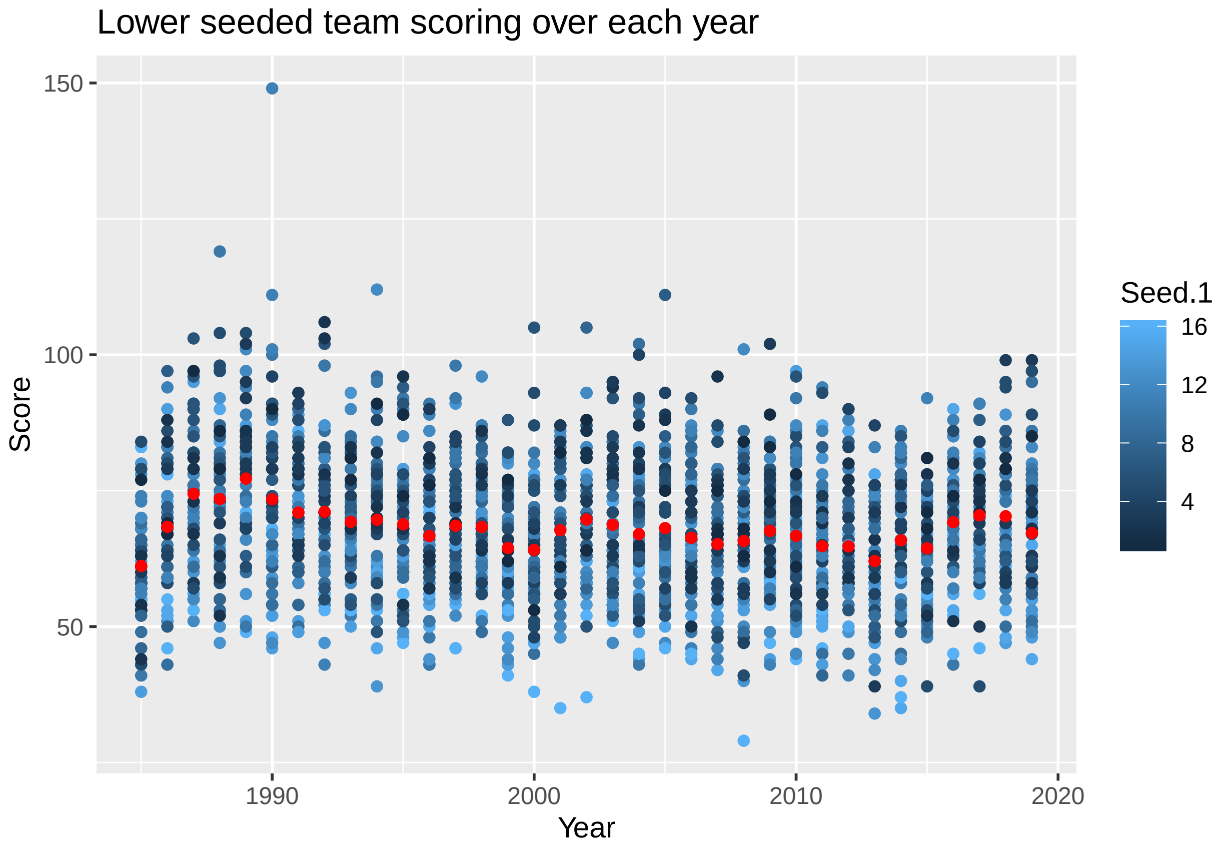
The fourth function is the same as the third function, but for the lower seeded team. As expected, the mean and median scores are lower than the higher seeded team, as it is logical the higher seeded team are better and can score more. What's surprising is that max points scored by a lower seeded team is 149, in comparison to 131 for the higher seeded team. This could be an outlier that could be interesting to investigate.

Graphical Representation

```
library(ggplot2)
bdgraph1 <- ggplot(bigdance,
                   aes(x=Year,
                       y=Score,
                       color=Seed)) +
  geom_point() +
  stat_summary(fun.y=mean,
               geom="point",
               color="red") +
  labs(title="Higher seeded team scoring over each year",
       y="Score",
       x="Year")
bdgraph1
```



```
bdgraph2 <- ggplot(bigdance,
                   aes(x=Year,
                       y=Score.1,
                       color=Seed.1)) +
  geom_point() +
  stat_summary(fun.y=mean,
               geom="point",
               color="red") +
  labs(title="Lower seeded team scoring over each year",
       y="Score",
       x="Year")
bdgraph2
```



One of the issues mentioned before is that due to rule changes, scoring trends could increase in mean over the years. However, by plotting the scoring of the higher seeded team and the lower seeded teams, the scoring trend doesn't have a clear upward trend in the mean, as denoted by the red dots on the graphs. The distribution of points scored seem to have a pattern that resembles the cosine curve, so there could be a significance in this visualization. Maybe there are annual changes in how the game is played strategically that causes the fluctuation in variance of the scores by all the teams.

Conclusion

Through initial investigation, we can see that the dataset is a time series with 10 variables for each row (game) that is recorded in the March Madness tournament since 1985-2019. Through further numerical and graphical summary, it can be inferred that a higher seeded team scores more than a lower seeded team on average and measured by median. Graphically, there is not a clear trend of higher scoring as the years pass by as hypothesized in the potential data issues, but does show a fluctuation that could indicate other factors affecting gameplan by teams each year.

References

1. https://data.world/michaelaroy/ncaa-tournament-results
2. https://en.wikipedia.org/wiki/NCAA_Division_I_Men%27s_Basketball_Tournament