College Basketball Seasons

Max Stevens and Zack Lasek

Introduction:

College basketball is an extremely popular sport that includes teams from across the country that compete to see who the best team is every year. There are multiple divisions to college basketball, but the highest level is Division 1, and that is what we are going to focus on in this project. Additionally, every single one of those years ends in the NCAA tournament and a champion being crowned. This project will look into data about college basketball teams from the years of 2013-2019 to see what factors go into deciding how those seasons play out, what teams are good and bad, and why, and what went into the champion being the “best team” that season. We are both very big fans of sports in general, as well as specifically college basketball, so we were very interested in diving into this data and hopefully learn more about individual seasons, what specifically makes teams good and bad, what goes into determining who wins a championship the most, and overall trends in college basketball.

Data:

The first source of data we obtained for use in this project is from Kaggle. It is a dataset that gives various metrics about a college basketball team for a certain year (from 2013 to 2021, but we used 2013 to 2019 because there were issues with 2020 and 2021) including their name, conference, games played, wins, various performance metrics, NCAA tournament seed, and their season end result. There are 2,455 data points in total from teams across these 9 years/seasons. The column names seemed good, and the data dictionary provides descriptions for any columns that aren’t perfectly clear from their names, such as a few of the performance metric columns. We removed a couple of columns that were unnecessary, or we felt that we could find from the web scraping that were more useful and easier to understand, such as the TOV column. The data can be found at the link right below.

College Basketball Data | Kaggle

The second source of data we used was scraped using a loop in R and the data was from Sports Reference. The data was from the 2013-2019 college basketball seasons, which matches the data we found and are using from Kaggle. The additional columns/features we added from this website and that we felt would be useful in further analysis are strength of schedule, simple rating system, turnovers, points, and opponent points. A lot of the data from

Kaggle were efficiency metrics and averages, so we felt including some overall totals would be useful, as well as SOS and SRS, which are both very useful, popular, and simple team metrics. The scraping was done using the website below.
2012-13 Men's College Basketball School Stats | College Basketball at Sports-Reference.com

After both sets of data were collected, some cleaning was needed before they could be merged. We had to change some of the names from both data sets to make them match across the two sets and then we had to make sure everything would merge correctly without cutting off teams and years or include extra teams and years that didn't match and introduce NA values. After the data was cleaned and prepared as needed in R, the two datasets were horizontally merged to add together the columns (features) being used from both sources of data. This gives us one complete and final data set that can be used for the analysis questions. The merged data frame has 2,544 observations and 29 variables/columns.

Data Dictionary:

Field	Type	Description
School	Text	The name of the school
Year	Numeric	What season it was based on the year at the time
CONF	Text	The conference that the team belongs to
G	Numeric	The number of games played
W	Numeric	The number of games won
ADJOE	Numeric	Adjusted Offensive efficiency for a team per 100 possessions
ADJDE	Numeric	Adjusted Defensive efficiency for a team per 100 possessions
BARTHAG	Numeric	The chance of beating an average Division 1 team (Power Rating)
EFG_O	Numeric	Effective Field Goal Percentage for a team (all shots included)
EFG_D	Numeric	Effective Field Goal Percentage Allowed for a team (all shots included)

Field	Type	Description
TOR	Numeric	Turnover Percentage Allowed - Turnover Rate for a team in games
TORD	Numeric	The steal rate for a team
ORB	Numeric	The rate a team gets offensive rebounds
DRB	Numeric	The rate a team gives up offensive rebounds (or 1 - their defensive rebound rate)
FTR	Numeric	How often a team shoots free throws
FTRD	Numeric	How often a team fouls a team that end up in them shooting free throws
2P_O	Numeric	Percent shot on 2-point field goal attempts
2P_D	Numeric	Percent that a team allow teams to shoot on 2-point field goal attempts
3P_O	Numeric	Percent shot on 3-point field goal attempts
3P_D	Numeric	Percent that a team allow teams to shoot on 3-point field goal attempts
ADJ_T	Numeric	Estimate of the tempo (number of possessions) for a team on offense per 40 minutes
WAB	Numeric	The wins above the bubble, and the bubble refers to making the NCAA tournament or not. Above bubble means make the tournament and below means not
POSTSEASON	Text	The round where the team was eliminated (made it to but didn’t win if applicable) from the NCAA tournament
SEED	Numeric	The seed number that a team was in NCAA tournament
SOS	Numeric	A rating of the Strength of Schedule of a team
SRS	Numeric	Simple Rating System (takes into account average point differential and SOS)
TOV	Numeric	Total turnovers committed all season
PTS	Numeric	Total points scored by team in a season

Field	Type	Description
OPP_PTS	Numeric	Total points scored against a team in a season

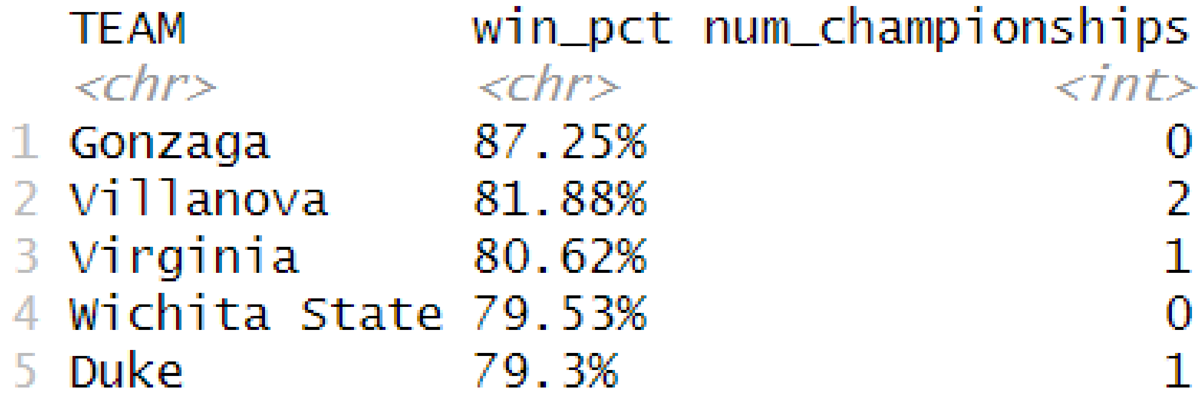
Proposed Analysis:

This project's goal is to determine the success of various teams over these 9 seasons in terms of regular season and postseason success. Additionally, we want to explore the relationships between various performance metrics on team success and some variance in officiating and games themselves across different conferences since that can have a large impact on college basketball games and outcomes at times.

Question 1:

Our first question we looked to explore was which team had the highest average winning percentage throughout the years included in our dataset, and how many times did they win the championship? To answer this question, we constructed a summary table with the top five teams regarding winning percentage. We included the additional four teams for comparison. The summary table shows the team's name, winning percentage, and how many championships they won. It is using data only collected from the Kaggle dataset.

Using the summary table, we were able to conclude that Gonzaga had the best win percentage across the dataset but were not able to win a championship. However, three of the other top five teams did win at least one championship and Villanova won the most championships out of the top 5 with 2 on their own. Seeing that 3 of the top 5 teams in terms of win percentage have won championships, we can conclude that winning games consistently all year is important to winning a championship.



```
  TEAM          win_pct num_championships
  <chr>         <chr>               <int>
1 Gonzaga       87.25%                  0
2 Villanova     81.88%                  2
3 Virginia      80.62%                  1
4 Wichita State 79.53%                  0
5 Duke          79.3%                   1
```

Figure 1: Top Win Percentage

Question 2:

For the next question, we wanted to see how consistent officiating is across the various conferences. To accomplish this, we looked at the conference and FTR (free throw rate) variables. We computed the average FTR for each conference and put them into a summary table. This data is only using the Kaggle dataset data.

To better understand the data, we then constructed a bar plot using the ggplot2 library. Looking at the plot below, there is no alarming difference in the average FTR across the conferences. There seems to be small differences, but officiating does not seem to be different across conferences.

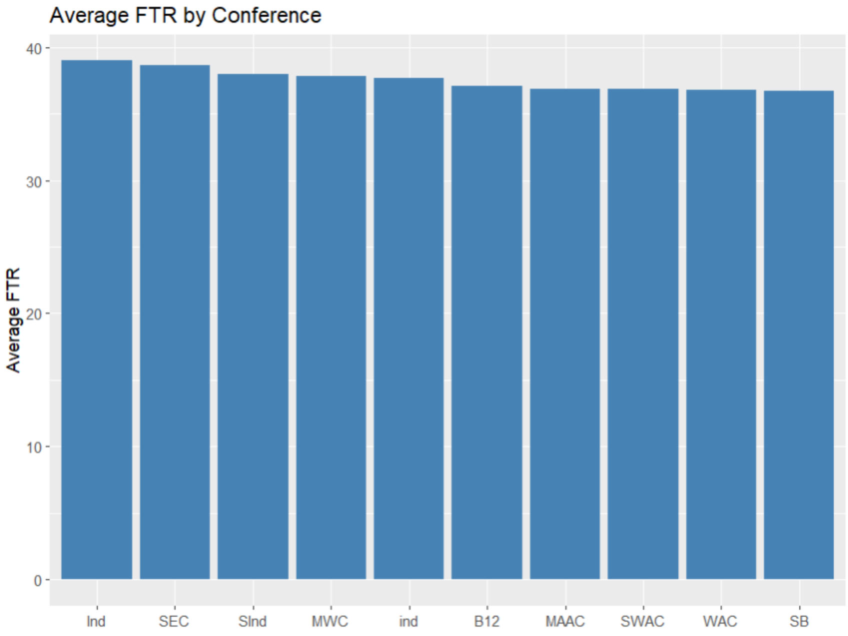


Figure 2: Avg FTR by Conference

Question 3:

The third question that we wanted to dive into deals with success by conference in the NCAA Tournament. We wanted to determine which conference(s) had the most success by determining the lowest average Postseason exit round for each conference. We decided that the best way to look at this would be to list the total number of appearances in each round (Championships, Final 4, Elite 8, Sweet 16) for each conference. This question was only using the Kaggle data and used the CONF and POSTSEASON columns to make the tables for each round.

We wanted to visualize this to make it easier to understand, so we made simple var plots for each round to compare the success of each conference. All the plots can be seen below and in order from Number of Championships (Figure 3) to Number of Sweet 16s (Figure 6). According to the graphs, the ACC and Big East were tied for the most championships with 3, the ACC, Big East, and SEC had the most Final 4 appearances with 4 each, the ACC had the most Elite 8 appearances with 12, and the ACC had the most Sweet 16 appearances with 23.

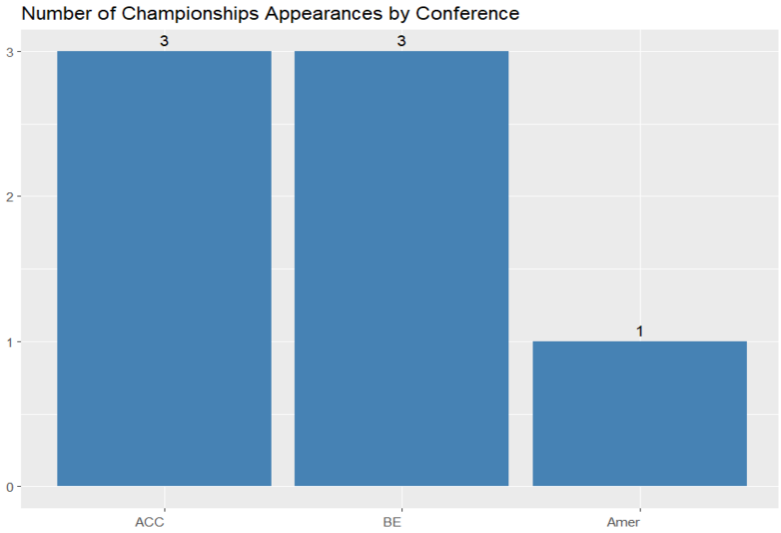


Figure 3: Number of Championships by Conference

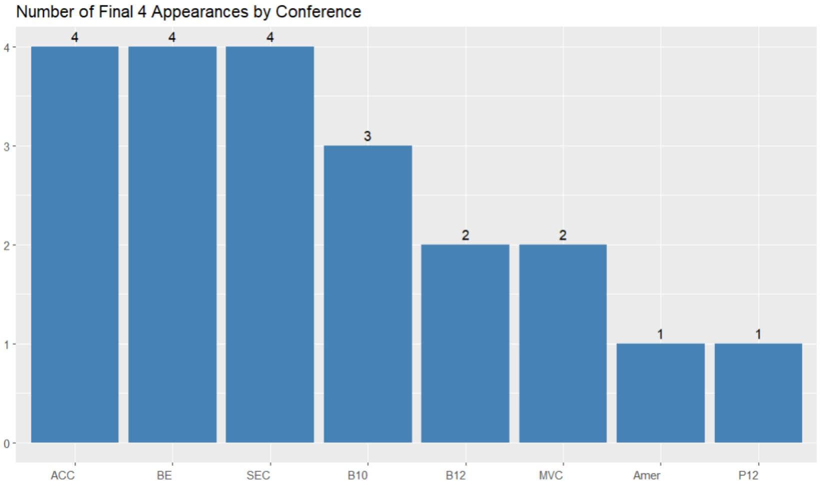


Figure 4: Number of Final 4s by Conference

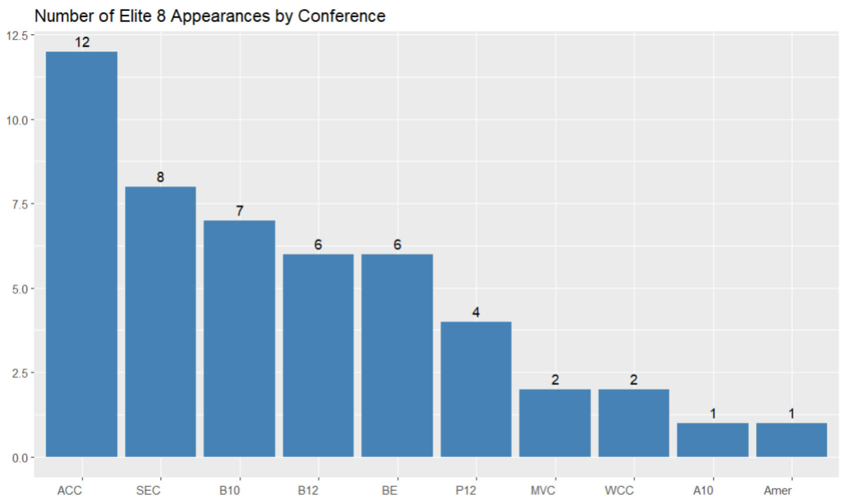


Figure 5: Number of Elite 8s by Conference

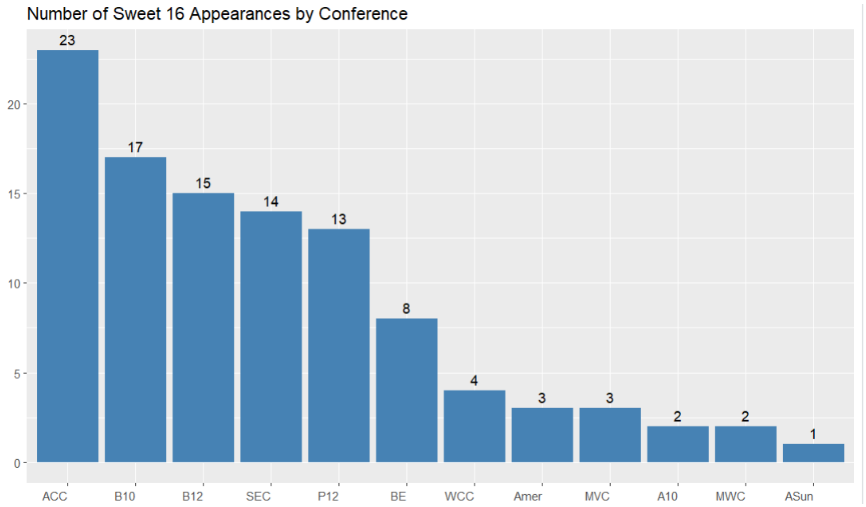


Figure 6: Number of Sweet 16s by Conference

Question 4:

For our next analysis question, we wanted to look into the performance metrics of SOS (strength of schedule) and SRS (simple rating system) for each team by conference to compare these metrics by conference. These are two simple metrics that go into determining what conferences face better competition, often within their own conference because they have better teams, and which conferences have teams that have better overall rating metrics. The columns used for this analysis are School, CONF, SOS, and SRS. Additionally, this analysis question uses data from both data sets collected (Kaggle and College Basketball Reference).

First, we ran summary tables to gather that information and then displayed them in two plots to visually display the data. The two bar graphs show the average SOS and SRS for all conferences, which compares all the conferences to each other. As you can see from the plots below, the graphs are very similar in how the bars look. This makes sense because the metrics are intertwined, a team with a harder schedule probably plays teams that have a higher SRS. It seems to be that the Big 12 has the best SOS and SRS, followed by the Big 10, then the ACC, and then the Big East in both metrics.

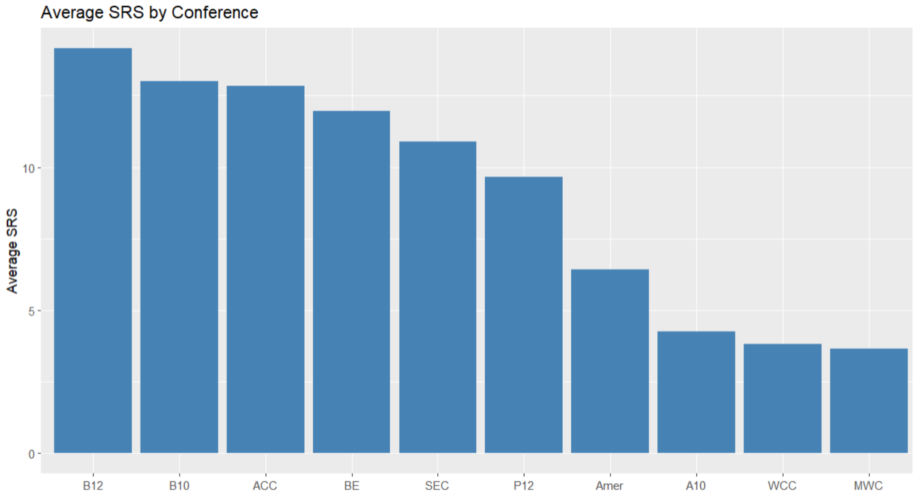


Figure 7: Average SRS by Conference

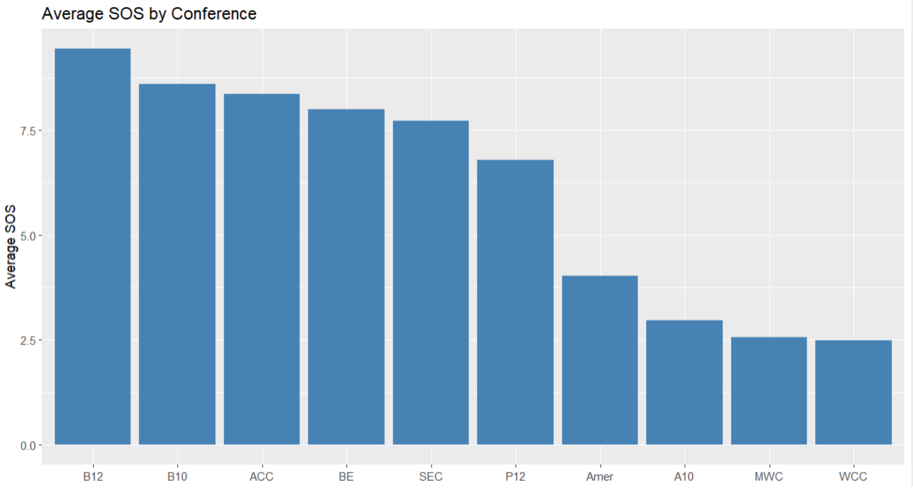


Figure 8: Average SOS by Conference

Question 5:

For our fifth and final analysis question we decided that we wanted to dive into a little about offensive efficiency metrics and how that leads or doesn't lead to more turnovers for a team. Oftentimes, having a high-powered offense means you play with a very fast pace; however, this can often lead to a team committing more turnovers as well due to that fast pace. Therefore, we wanted to see what 10 teams had the best offenses and what 10 teams committed the most turnovers and see if there was any overlap there at all. For this question we needed to use the School, ADJOE, and TOV columns. This analysis question uses data from both data sets.

The plots we created for both can be seen below. The teams with the best offenses are Duke, Villanova, North Carolina, and the rest of the top 10 can be seen below. Additionally, the teams with the most turnovers are Arkansas-Pine Bluff, Savannah State and the rest of the list is below. According to both lists, there is not really any overlap between the top 10 offenses and the top 10 teams who commit the most turnovers.

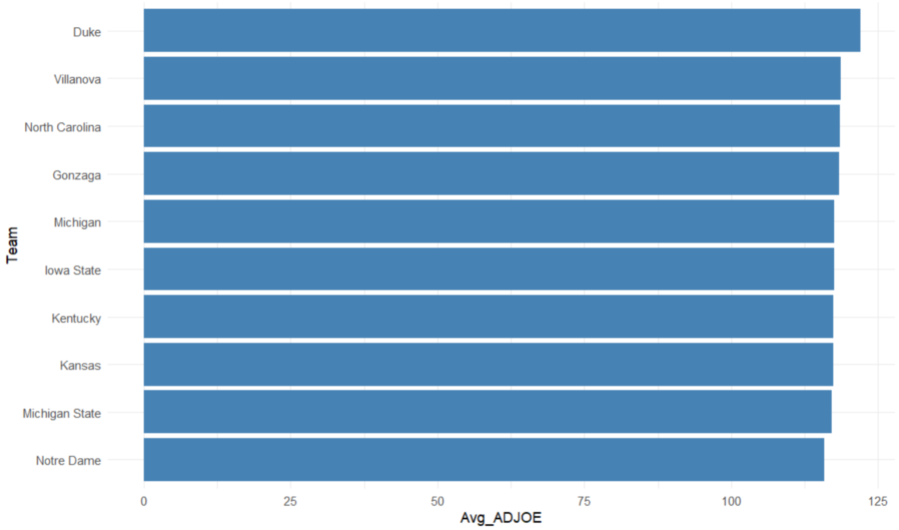


Figure 9: Top 10 Most Efficient Offenses

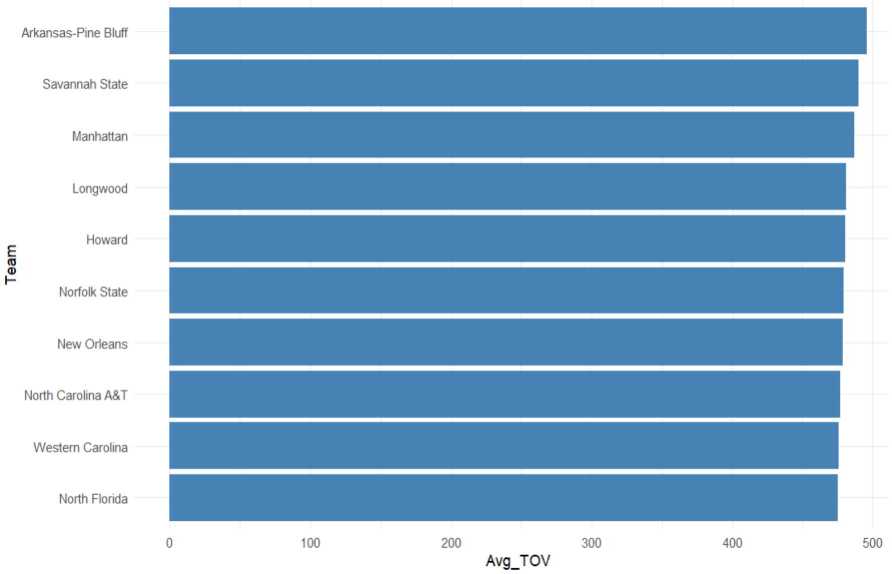


Figure 10: Highest Turnover Offenses

Conclusions:

In the analysis section above, we dove into 5 questions that we thought were interesting and that we wanted to explore further from our data. We examined them not extremely in-depth in the section above, and we displayed and explained what our analysis showed at a baseline level. However, now we will dive a little further into all of those questions and analysis, and see what we can conclude from the results, if anything at all.

1. What teams had the highest winning percentage across all years, and how many championships across those years?
    a. Villanova seems to have had the most success in terms of winning percentage and championships, and some could say that Gonzaga has had disappointing postseason performance since they have won the highest percentage of their games but have not won a title over these years.
2. Is there a significant or noticeable difference in officiating across the conferences in college basketball?
    a. It would be safe to assume that most of the variance in officiating and calls per team is solely different on a team-by-team basis, and maybe there could be and are issues there; however, there does not seem to be an issue or problem based on conference alone.
3. What conferences had the most success in the postseason/NCAA Tournament?
    a. The ACC clearly had the most NCAA tournament success than any other conference. They had the most Sweet 16 and Elite 8 appearances. Also, they were tied for the most national championships over these years with the Big East (3 each) and Final 4s with the Big East and SEC (all had 4). Based on these numbers, it seems fair to say that the ACC has had the most postseason success than any conference, and we would say the Big East and SEC would be next.
4. Based on the criteria of SOS and SRS, which conferences seem to be the best overall and have the best competition?
    a. Based on these two team/performance metrics, the Big 12 is apparently the hardest conference with the best teams. This is very interesting because question 3, which was about NCAA tournament success, says the ACC was clearly the most successful conference. Therefore, does it seem that metrics like these or does postseason success seem to be more important to deciding what

conference is the best? We believe there is no clear and obvious way to decide that (it is up to what someone values more) and nothing definitively can be said or determined based on these metrics and postseason success. It is up for debate or discussion.

5. Is there a correlation in any way between the best offenses in college basketball and teams who commit the most turnovers due to those good offenses with seemingly high paces?
    a. Based on the two graphs in the analysis section, there is no overlap between the most efficient offenses and the teams who commit the most turnovers. Therefore, someone could say there is no correlation between the best and highest paced offenses and turnovers; however, this is not a perfect test or analysis. Most of the teams who have the most turnovers seem to be teams on the lower end of Division 1 teams (teams that most people haven't heard of), so that could possibly be skewing the results. Therefore, this doesn't seem to be the perfect comparison and analysis, so we don't believe a clear and definitive conclusion can be made from this analysis question and results, unfortunately.

A few limitations in the data seem to be that not all performance and team metrics are included in the data. Additionally, some of the lower end Division 1 teams, which clearly are never as good as the best teams in college basketball could be skewing some of the questions people want to answer (such as our question 5), and that could be an issue. Some suggestions for improving the data in future work could be including even more variables and metrics for teams, including more past (and future years at this point if possible with better data), and possibly filtering out teams that never have any success or are kind of irrelevant to the overall seasons in Division 1 basketball if that is possible or good for the data at all, which it very well may not be.