Sean McLean

ALY 6010

Final Project – Milestone 1

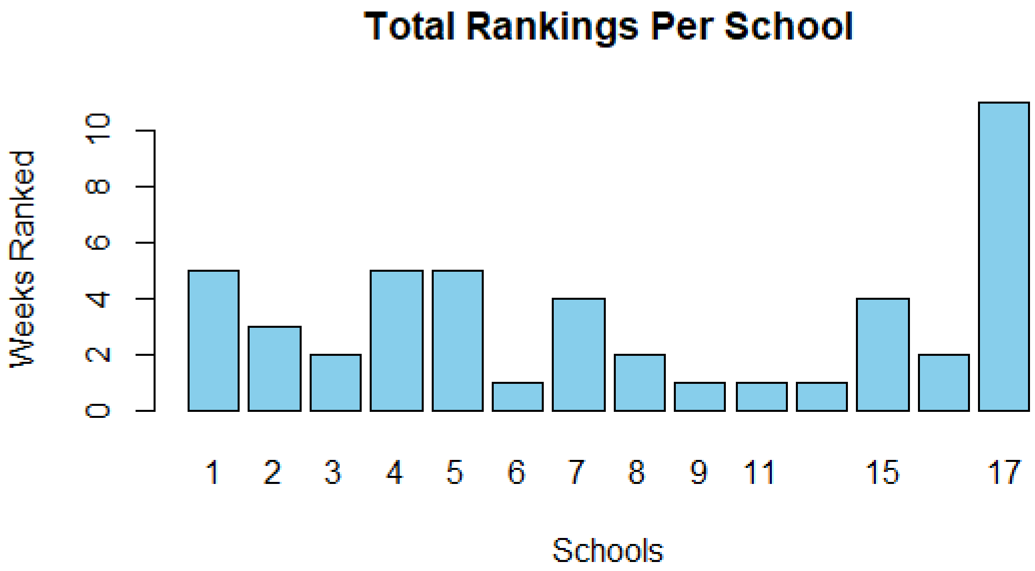
The dataset selected shows all the rankings from every ranking system for every division I school in college basketball during the 2020-2021 season. Overall, it has the primary purpose of providing the rankings for every week of the season and how it changes as the season goes on. The dataset comes from Kaggle which was compiled from all the polling systems used for the college basketball rankings.

After uploading the dataset to R Studio the first noticeable attribute is how large the dataset is, with over 250,000 rows and seven columns or categories. The dataset is just about even in terms of numerical data and categorical data. The numerical data is primarily the weeks of the season when the polling is released and the rankings of each school per week and an identification number for each school. The categorical data consists of the ranking system that is providing the rankings, the abbreviation of the ranking systems or letter codes, and the names of the schools. The other column is for the season but is the same for every row since the dataset is just for the 2020-2021 season.

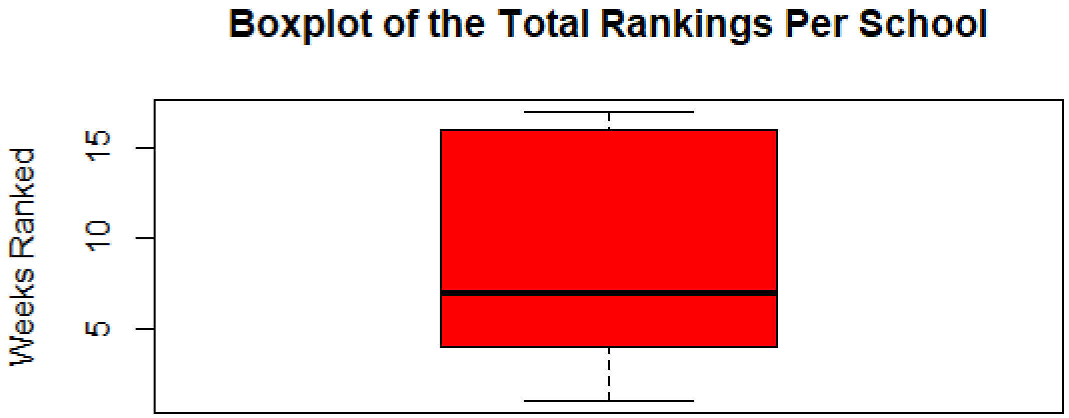
The rankings column orders the schools in terms of record for each week, with many of the ranking systems only ranking the top 25 as is customary in weekly polls for college athletics. Several ranking systems ranked all 357 schools in order in their weekly polls. The metrics of how teams are ranked like record and strength of schedule are not included in this dataset which would explain why teams are ranked in the order they are in. The polls are done each week which is why the dates in the column are every seven days.

With the redundancy of the dataset and the immense size, the first step was to condense it. The season and letter code columns I thought were pretty useless toward doing exploratory data analysis so I decided to remove them from the dataset. This doesn't make the dataset any smaller in terms of the rows, so I decided just to focus on the rankings from one ranking system that only provides the top 25 weekly. I chose the Associated Press simply because they are arguably the most common poll used in college athletics. This shortened the dataset significantly which makes it a lot easier to see what the rankings looked like per school for the season.

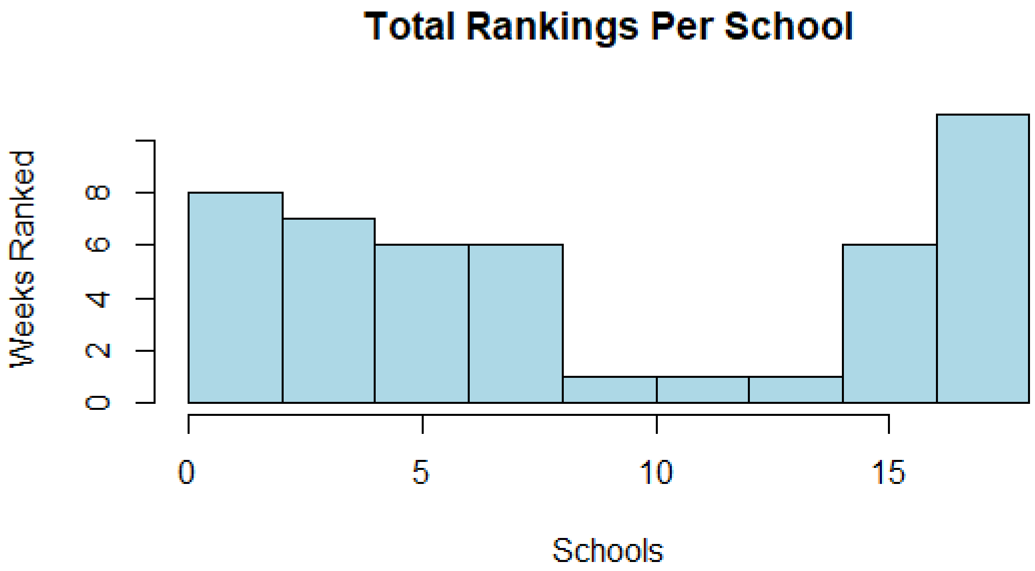
Before I made the dataset smaller, I wanted to know what school appeared the most in the dataset and who appeared the least. The top school was Baylor who was ranked very high all season and the least was Bethune-Cookman who was ranked among the worst schools in the ranking systems that ranked all the schools. The schools that appeared the most were also coincidentally near the top of the rankings most weeks. I then tallied all the schools that were ranked in the top 25 at least once during the season which what I was most curious about from looking at this dataset. Overall, out of the 357 schools, only 47 were ranked at least once in the top 25 polls, ranging from one time to 17 times for an average of just over nine times per school. These results were then put into data visualizations to show how they dispersed between each school and the number of weeks.



Total Rankings Per School



Boxplot of the Total Rankings Per School



Some of the main takeaways when analyzing the data is that there is a small number of schools that are ranked in a season. And since there not too many weeks where rankings are conducted, the average amount of time that schools stay in the top 25 indicates that there is not much diversity in who is ranked and that it is almost the same group of schools that are in the rankings for most of the season. A quarter of those ranked schools are also ranked the entire season, indicating that they are most likely to win the national championship from their body of work. I would like to in the future compare the ranking system used with some others to see if there are any differences in teams or number of weeks ranked. Adding more seasons to this dataset could also be interesting and a question I would have been whether the number of weeks ranked in the polls be a factor or not in winning a national championship. I also would like to incorporate more data visualizations because it would really make the data sets stand out a lot more and easier to understand.

References:

Masseycre. (2021, November 13th). College Football/Basketball/Baseball Rankings. Kaggle. Retrieved on November 11, 2023 from https://www.kaggle.com/datasets/masseyratings/rankings?select=cb2021.csv

Chat GPT. (2023, November 4 th). Default (GPT 3.5). https://chat.openai.com/c/18163088-1d42-4e5b-a8d1-b9736ade02bb