Sean McLean

ALY 6010

Final Project

Introduction

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.

Data Cleaning/Analysis

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.

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 a data visualization to show how they dispersed between each school and the number of weeks. See Figure 1-1.

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.

Questions to consider

Questions to explore from the initial data set that pertains to the college basketball rankings from the 2020-2021 season.

1. From the 15 teams that were ranked in both the first poll and final ranking, did the means of the team's rankings between the two polls change from the first to final polls?
2. How much of a correlation is there between the rankings by the Associated Press and USA Today Coaches polls for the same week?
3. Do the schools that are ranked between the two ranking systems change consistently over the first and last polls of the season?

Hypothesis

Hypotheses established based off questions 1:

Null: There was no mean difference between the first poll and final poll using a two-sample test for the 15 schools ranked in both.

Alternate: There was a mean difference between the first poll and final poll using a two-sample test for the 15 schools ranked in both.

Hypotheses established based off question 2:

Null: There is no correlation between the rankings of the two ranking systems for the same week.

Alternate: There is either a positive or negative correlation between the rankings of the two ranking systems for the same week.

Hypotheses established based off question 3:

Null: The schools ranked between the two polls do not change between the first and last polls of the season.

Alternate: The schools ranked between the two polls change consistently between the first and last polls of the season.

Findings – Question 1

Research and analysis were done using the data from the first part of the final milestone project and how the questions were formulated. The data for the first and final polls conducted by the Associated Press was extracted to locate the top 25 rankings and see what schools were ranked in both. The 15 schools that were ranked all season or for a portion of the season are

Gonzaga, Baylor, Villanova, Virginia, Iowa, Kansas, Illinois, Crieghton, Texas Tech, West Virginia, Houston, Texas, Florida State, Ohio State, and Michigan. The school's rankings from both polls were then put into new variables that were used for t-tests.

With the usage of several t-tests that used one sample and two sample testing and also at different significance levels, the tests consistently showed high p-values and low t-statistics in each test conducted. The mean ranking of the 15 schools in the final poll was 10.13, about 1.5 points higher than the mean ranking of those schools in the first poll of the season. See Figure 1. 2. While this shows improvement as the season goes on, the high p-value suggests that this small difference is not enough to reject to the null hypothesis because there was nearly no difference in the means. Traditionally the preseason poll is not a reflection of how the final poll will look, with several factors like injuries and other schools that are better than anticipated will alter the polls over the course of a season.

Findings – Question 2

The necessary variables like ranking systems, school ID numbers, rankings, and weeks were extracted from the dataset and used to create a new dataset for analysis. The dependent variable for this question is the rankings, the independent variable is the week being looked at, and the dummy variable is the ranking systems. The strongest relationship when looking at the correlation coefficients is between the polls and the schools. This was also seen after performing a linear regression between ranking systems and the schools, as a low p-value indicates a potential rejection of the null hypothesis.

A scatterplot was created to show all the school's rankings between the two ranking systems for the one week selected. See Figure 1.3. There is a noticeable number of dots that are close together vertically which means that they are the same school ranked close together from the two polls. The positive relationship that was correlation coefficient calculated between the schools and the rankings is evident in the scatterplot with the regression lines for both ranking systems on top of each other and trending upward. The USA Today Coaches poll has more outliers in the scatterplot than the Associated Press which could indicate that they rank schools slightly differently and also have some different schools in the poll. Overall, the two ranking systems rank the top 25 schools more similar than not and could be a microcosm of how it is ranked every week of the season, and enough evidence to reject the null hypothesis.

Findings – Question 3

A new variable was made that featured the first and last weeks of rankings, both polls, and the schools ranked. The dependent variable established is the ranking systems with the independent variables being the schools and the weeks used for the rankings. A regression analysis conducted to see if the relationship between the rankings from the two ranking systems changed over time showed no correlation. Individual scatterplots were produced for each ranking system to compare where the schools fell in the first and last polls of the season.

The scatterplot of the schools ranked in the first and last weeks of the Associated Press poll showed more variety with not as many schools ranked in both polls. See Figure 1.4. Its regression line shows a slightly negative relationship between the weeks. The USA Today Coaches poll scatterplot looks more consistent and appears that more schools are ranked in both polls by the placement of the data points in the plot. See Figure 1.5. The regression line is slightly negative but noticeably less than the regression line in the Associated Press scatter plot. Despite this, there is not enough evidence here for me to favor the alternate hypothesis as it appears the schools are for the most part consistent in both ranking systems.

Figure 1.1 Bar Plot of Total Weeks Ranked Per School

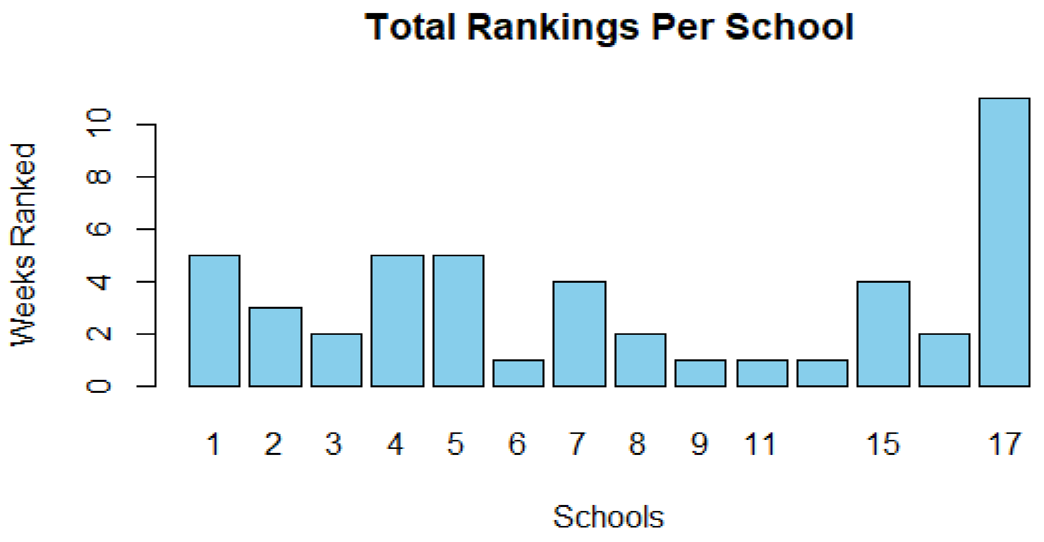


Figure 1.2 Dot Plot of Average Ranks in First and Last Polls

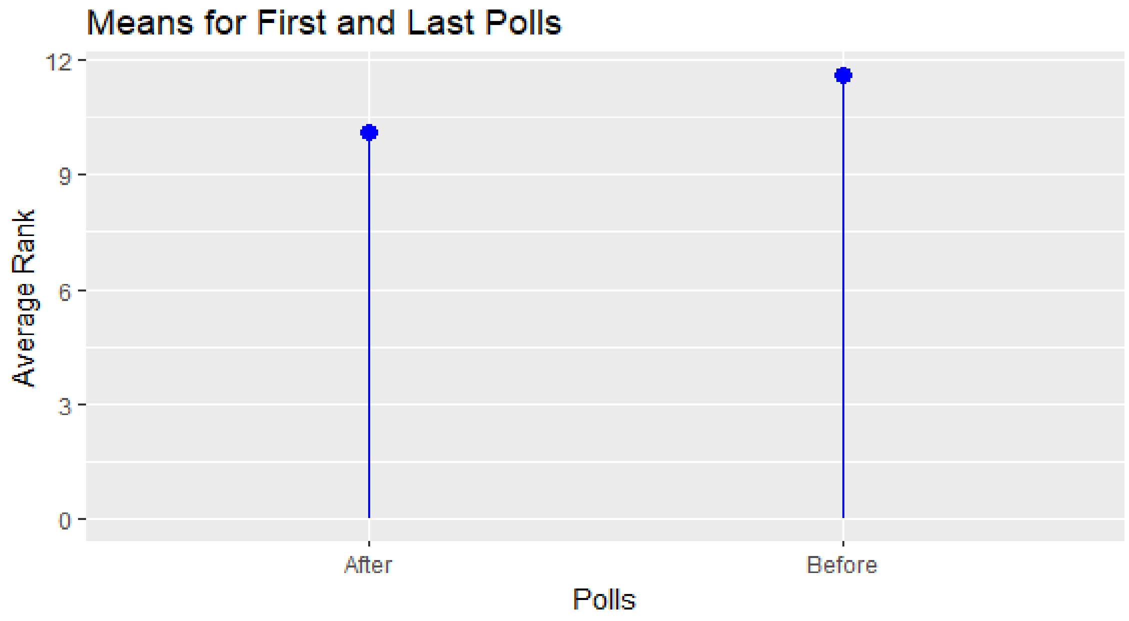


Figure 1.3 Scatterplot of Two Ranking Systems in One Weekly Poll

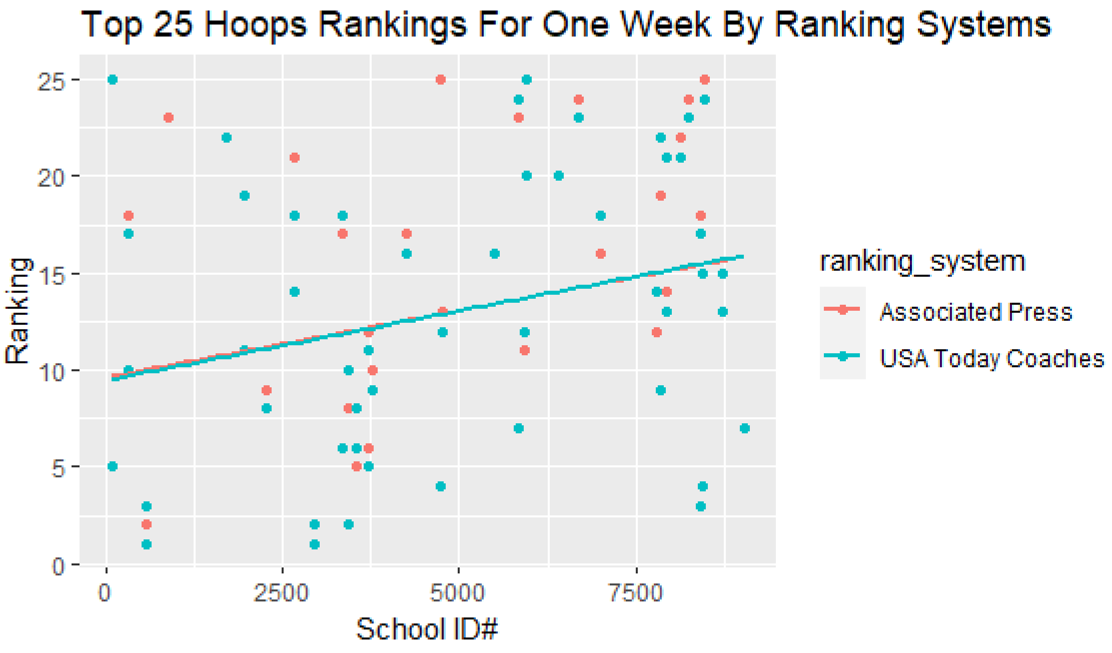


Figure 1.4 Scatterplot of Schools in Associated Press First and Last Polls

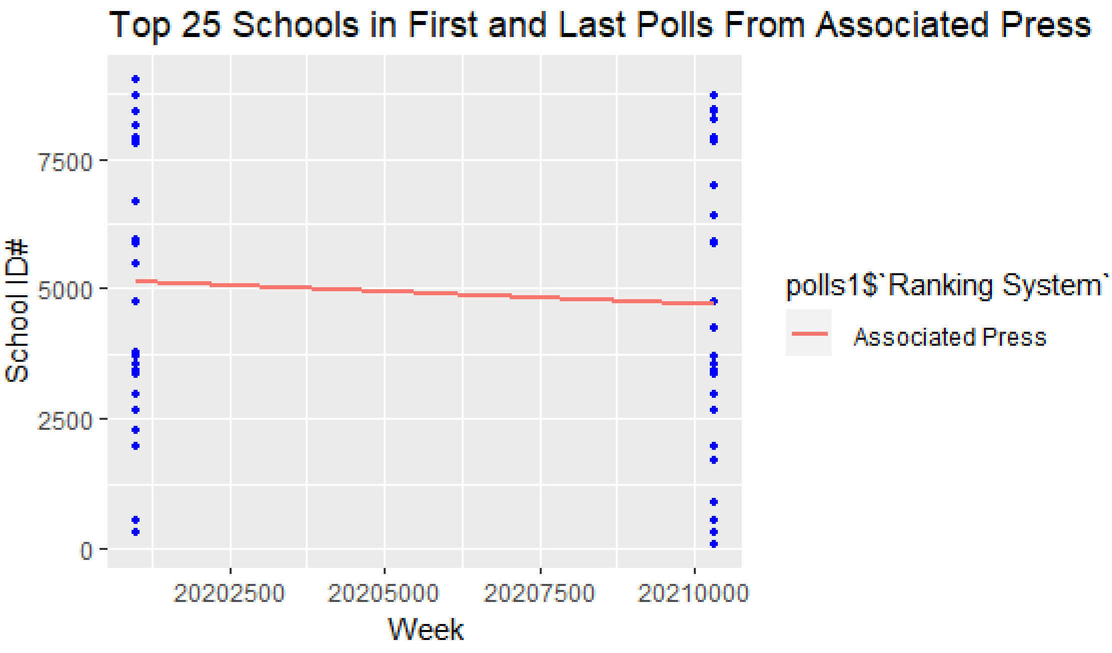
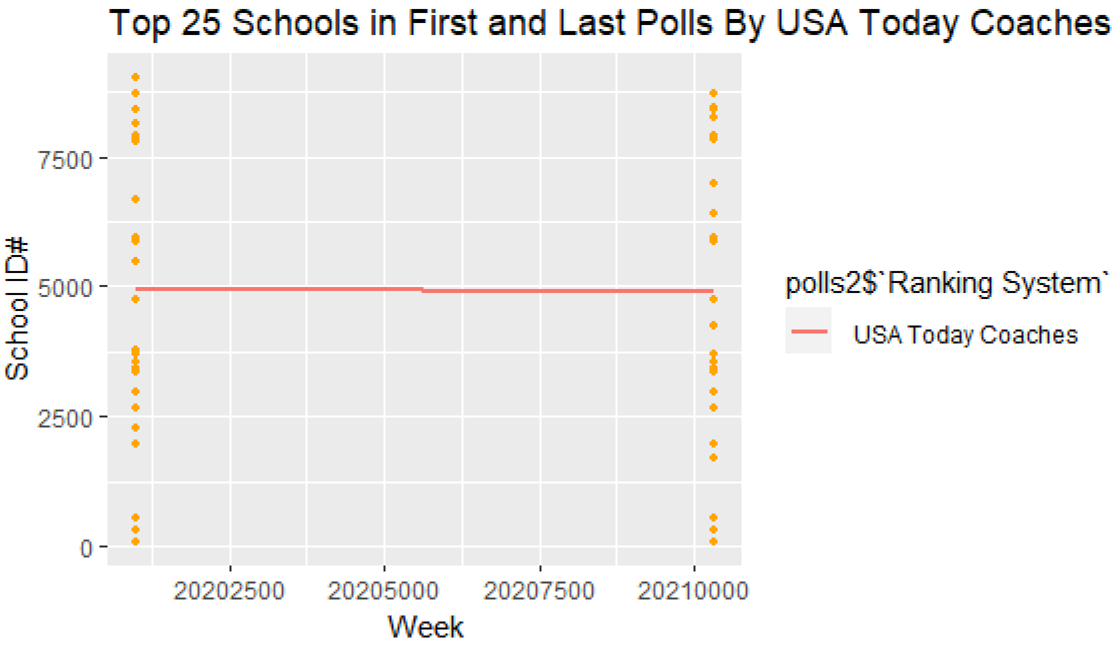


Figure 1.5 Scatterplot of Schools in USA Today Coaches First and Last Polls



References:

Bluman, A. G. (2017). *Elementary Statistics: A Step by Step Approach.* 10th edition. McGraw-Hill Education.

Chat GPT. (2023, November 4 th). Default (GPT 3.5). https://chat.openai.com/

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