The initial scope and goal of my final project was to take weekly scoring from NCAA Division I gymnastics and aggregate it into one program, and then track each team’s weekly progress. My plan heavily relied on the website roadtonationals.com, which is the official NCAA rankings website. However, I soon learned that I would not be able to use this website as it used a mixture of HTML and Javascript. The HTML pandas scraper also uncovered that though this website looks like it uses a table to display information, it is actually just code written to look like a table but isn’t. Therefore pandas couldn’t scrape a table from this website. Because the website also uses Javascript, Beautiful Soup was also unable to scrape much of the data from this website. A screenshot of a computer screen

Description automatically generated with medium confidence

Figure What the roadtonationals site looks like

A screenshot of a computer

Description automatically generated with medium confidence

Figure What the roadtonationals code looks like

Unfortunately as this website is the main NCAA rankings site, there aren’t many other resources for getting an up-to-date rankings list. Most other gymnastics-related websites will simply link to roadtonationals.com than have their own list of updated rankings. I was able to find one fan-made blog that would react to the rankings every week and would include a table of the rankings as well, balancebeamsituation.com. However, after the first five weeks of the season, this blog continuously changes the way it formats its rankings and reactions. Therefore I needed to have a couple different functions to parse through each weeks rankings in order to scrape the main data I was looking for: rank, team name, and overall score.

Table

Description automatically generated

Figure How weeks 1-5 are formatted

Table

Description automatically generated

Figure How weeks 6 and onward are formatted

Then, when I went to add in week 7 (which is formatted similarly to week 6 above), I found that the HTML was extremely messy and teams were not actually being assigned their rank properly in the code. This caused my scraper for week 7 to not return accurate rankings.

Text

Description automatically generated

Figure Even though Michigan has moved down to third in the rankings, this website still has the HTML rank as 1

Text

Description automatically generatedAs you can see in the image below, when I scrape the data from week 7, #1 ranked Oklahoma’s overall score of 198.080 is being attributed to #3 ranked Michigan’s overall score.

Table

Description automatically generated

This is a disappointing revelation that I am probably not going to be able to maintain this program as the NCAA seasons continues, since my source is high inconsistent with its code and formatting. Though I am not surprised as again, this source is a fan-blog and not an official website. My week 1-6 output is very accurate and I am happy with how everything looks. Though you will notice that the change in rankings format for week 6 means I had to go about scraping information in a different way, the rankings will print out as a list within a list.

Text

Description automatically generated

Text

Description automatically generated

I decided to only take the top 15 teams for graphing purposes as having 50 teams represented on one line graph is an absolute mess. I also had to exclude the number 7 ranked Alabama from week 6’s rankings as you will notice that for some reason, their team score is only being scraped as a 19, which was causing huge issues when I then went to graph everything. I wanted to graph each team’s rank and scores over the week to see how much a team has improved or has not improved each week.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Overall I am very happy with how my project turned out. Though I had to overcome many challenges, I am happy I was able to learn a new skill and use it effectively.