Semester Project Write Up

ECON 8320-001

Marissa Czapla

For my end of semester project, I decided to tackle the project option of collecting information about student-athlete NIL deals using the <https://nilcollegeathletes.com> website. To collect the information, I decided to web scrape which I quickly learned was a little more difficult and complex than the web scraping we did in class when looking at Lego sets. At first, I tried to only look at the different names of athletes (using the ‘tr’) to pull the information from. However, this made it hard to pull additional information from the rows and combine the information into a data frame. Instead of creating one data frame, I was having to create multiple data frames and then combine then all together. Though that method did work, it was super complicated. This also led to complications when trying to create additional for loops in the code so that each sponsor per student athlete could have its own row. After about a week I decided to scratch the original code I started with and consult Stack Overflow to get a better idea on how I would be able to do the project.

After researching and reading a copious amount of web scrapping articles, I decided that I would need to start my code by using the find.all for the ‘table’ of the website page and then from there go row by row finding the ‘tr’. From there I was able to web scrape the appropriate information located in the ‘td’, such as the athlete’s name, sport, school and sponsors. This made it a lot easier to include in a data frame and kept me from creating and combing multiple data frames. From there I was also able to easily create an additional for loop within my function that would return each sponsor as its own row.

In additional to the struggle of using the ‘table’ vs using the ‘tr’ as a starting point, I also struggled a lot with creating an if/else statement within my code that would return the ‘next’ pages of data to my data frame. One of the more annoying things about the way the web page was designed, was that each ‘class’ had long and wordy distinct characteristics. For instance, the distinct name for the ‘class’ that was associated with the next button was “"inline-flex items-center border-b-2 border-transparent py-2 pl-1 text-sm font-medium text-gray-500 hover:border-blue-700 hover:text-gray-700". The example we did in class with the LegoSets was just “next”. This created a lot of problems with my code because I spent about 5 hrs. trying to understand why my code would not return all other information than that of the first page. It unfortunately turned out that all I did wrong was include an extra space in the distinct text. An additional area I struggled with was finding a way to concatenate the data from the next pages onto my data frame. This was mainly because I didn’t realize that the ‘href’ associated with the ‘next button’ was only looking at half of a URL rather than the entire URL that would take me to the next page on the website. To fix this issue I had to create a concatenate function that included both URLs: pd.concat([data, collectAthletes('https://nilcollegeathletes.com' + next\_button)],axis=0). By combining the URLs my code finally was able to pull the necessary information I needed to start doing some analysis.

One big take away I had while doing this project was that understanding how a website configuration is set up will make a huge difference. I tried to model my project off what we did in class which often lead to a lot of frustration and errors. I would also make sure I just copy the text of the class distinction rather than typing them up to save me a lot of time in case there are weird white spaces that I wouldn’t be able to catch otherwise. Moving onto the analysis part, I wish I would’ve used a website that overall had more information about the deals done through NIL. All the information from the website is categorical, which isn’t super helpful when wanting to understand the nature of these deals and the effects NIL has on student-athletes. All my analysis was looking at categorical information which helps get a picture about what is going on with college athletics, but it doesn’t really get good detail or clarity on how much these student athletes are making per sponsorship. Another issue I had with the website I used was that it was out of date of what is currently happening with NIL. A few of the athletes, such as Adrian Martinez or Angel Reese, show to still be at their former colleges, rather than their current ones. With that being said, I am sure that all the information on the website is at least a year out of date. Since the NIL rules are new, having outdated information doesn’t tell us anything since so much has changed over the last two years in college sports. I think it would be very interesting to do this project again in the next 5 years to see how different the analysis is, especially with universities now hiring directors of NIL, and new agencies focusing on NIL coming to light. It would also be interesting to see how NIL changes the opportunities for women in sports since many women are becoming the top athletes with NIL sponsors.