James Hamlette

ECON 8320-001 Spring 2023

Semester Project Write-Up

The project that I decided to work on this semester was the scraping and analysis of NIL college athlete data. Very few websites provide this data readily and even fewer provide data on college athletes in an extractable format. Recently, NIL deals and associated valuations seem to have created a larger amount of interest in the topic as a market is being created and various companies are partaking. While I was largely successful overall in scraping and analyzing some data, some of my python code and thought process between the project proposal and now have changed. I proposed to match roster website information with the NIL data but found this hard to generalize overall in the end considering that each roster website for NCAA teams can be immensely different. In concluding results, I show that most top 100 NIL deals are either Football or Baseball players. Also, I show graphical analysis of the major positions that are common in basketball and football in garnering NIL deals. Furthermore, I assess that the most popular sponsor is Barstool Sports.

In my scraping procedures, I focus on two primary sources of NIL deal data. This includes the NIL College Athletes website and the On3 top 100 valuations and rankings data[[1]](#footnote-1). Within the On3, subsets of alternative top 100s are analyzed such as the top 100 football NIL deals. Two different sources are used for a few reasons. Firstly, the NIL College Athletes website doesn’t appear to update regularly with recent NIL deal announcements. However, there is much more accessible data on the NIL College Athletes website that can be reached using the Beautiful Soup library in python. The On3 website provided more quantitative variables that extends my analysis options so by including both websites, I paint a picture of some general facts about NIL athletes overall from the major available databases.

Coding wise, I defined 4 different functions to scrape the data. My first function (*collectNames)* first finds NIL College Athlete website data on all the names in a page and appends them to a list. To get the sponsors, I used a find\_next\_sibling() method to append the sponsors to each athlete name. However, something that I was not able to figure out in this process was how to get multiple observations for each athlete-sponsor combination. Instead, my scraping method for the sponsors of this data has all the sponsors for one athlete in a single value/cell along with their name. While this is not ideal and will need to be changed to run more efficiently, less than a few hundred out of the over 8,000 observations had more than 1 sponsor so for my initial analysis here, I stuck with this. Moving forward through the function I used for loops on parsed tags to gather full-column data on sport and university. I attempted to pull twitter links from the sub-pages on this website but was unable to. My code for scraping the twitter links ran by itself correctly, but once put into the function it didn’t allow my function to finish scraping before timing out. After appending the sport and university data to the names and sponsors in one data frame, my code then proceeds to scrape through and concatenate the next pages while there is a next page using a recursive function. My analysis that appears in my slides is then from this data frame. No cleaning was done on this. Other problems that I was never able to resolve included getting the values from this website as well as correcting for names with hyphens or symbols in either the University or Name fields. Sometimes these values were unrecognizable in the data frame after scraping.

Following, I used three functions to scrape On3 data on the top 100 NIL athlete deals, top 100 football, and top 100 basketball NIL deals respectively. The analysis for each of these is fundamentally the same. I gathered data on name, position, valuation, followers, and made a variable for rank for each of the three data frames. From there I did valuation by position graphical analysis and position distribution summaries for each. For the overall top 100 athletes I made a graph to show Valuation by rank to see if there was an association that could be observed. The On3 data needed some cleaning prior to analysis as well. Many of the values for followers and valuation were represented with dollar signs and “M” for millions, “K” for thousands, etc. With that, I wrote a function to change the value in each data frame by replacing the string value and multiplying the number by the corresponding factor. Then this function also changed the values to floats of themselves. Something that didn’t work out in this part of the scraping procedure was my attempt on the main page of On3’s deal tracker. This page has an interactive feature using a “load more” button which when clicked didn’t change the page or link so the Beautiful Soup library was not much help here.

Overall, my project results are in a single Jupyter notebook file, LaTeX slides, and described briefly here in this write-up. The data is also in csv format to provide for replicable coding. Something that I think I would consider if I were to try this again is making my code more efficient. The NIL College Athlete scraping function didn’t take more than 5 minutes to run but it had issues when trying to optimize it to take more values being scraped such as the twitter links for each athlete. Also, I parsed through the entire body of the table instead of doing one row at a time and looping through the rows and subsequent pages. This might have saved me more time trying to make sure things lined up correctly when joining columns to each other.

1. NIL College Athletes website: <https://nilcollegeathletes.com/athletes> ; On3 Top 100 websites: <https://www.on3.com/nil/rankings/player/nil-100/> [↑](#footnote-ref-1)