**Narayani Bakhati**

**May 06th, 2023**

**Extracting Athletes Profile from NIL Student Athletes Website**

For this project, I collected and organized information about student-athlete NIL deals found on NIL sources nilcollegeathletes.com. The website selected for this project has a list of information about NIL Athletes’ records. The objective is to extract and analyze the sponsorship trends among different colleges, sports, and athletes and visualize the results using various visualization techniques to gain insights from the data. The data are saved and stored in a CSV file named, Athlete’s Data Bucket which is user-friendly and easy to navigate.

**Methods**

The first step in this project was to scrape data from the website using Python's requests and the BeautifulSoup library. Since the data was paginated, I started with finding the total pages available scraping the data from the ‘nav’ tag. After I created a list of paginated URLs appending page numbers to the base URL. Further, I used a ThreadPoolExecutor to concurrently iterate over those URLs for performance optimization. The mapping function passed to the iterator scraped the table data for the current page. I had another iterator to go after each table row and scrape the href value embedded in the athlete’s name. The href value was appended to the base URL and the response was retrieved using a Python request. I then extracted the relevant information from the HTML and stored them in a Python dictionary. The value from the dt tags became the keys and the value from the dd tags became the corresponding values. The overall data dictionary was collected in a bucket using Python List. After scraping the data, cleaning and grouping were paramount for a good visualization. I used a sanitizer function to sanitize the data for consistency. Finally, I massaged the data by creating different data frames using the Pandas library and created visualizations using the Plotly library. I created three visualizations: a heatmap showing the correlation between sports and sponsors, a bar chart showing the top sponsors by sport, and a bar chart showing the distribution of sponsors across colleges.

**Results**

The web-scrapping code was successful, and I was able to retrieve the needed information about Athletes. A total of 8390 data of Athletes under NIL were scrapped from the website, consisting of 8348 athletes, 509 colleges, 35 sports, and approximately 67 sponsorship which are unique counts. The data is saved in a CSV file (athlete’s data bucket) for further analysis. The athlete’s data bucket will allow users to filter athletes by the university, sport, social media accounts, and sponsorship details in a list. After analyzing the data, it was found that Barstool Sports is the sponsor with the highest number of sponsorships in the dataset. In terms of sports, football emerged as the sport with the most sponsorships, followed by basketball and golf. Among the colleges included in the dataset, Ohio State University had the most sponsorships, followed by the University of Michigan and Michigan State University. When it comes to individual athletes, Jason Hubbard was found to have the highest number of sponsorships among all the athletes in the dataset. As social media accounts do not provide much insight, I decided not to analyze the list of social media accounts that each athlete had. Below are the total unique counts of each value.

**A picture containing text, screenshot

Description automatically generated**

**Challenges**

During the web scraping process, I encountered some challenges due to data inconsistencies on the website. This made it difficult to create data frames and visualize the data initially. But, after understanding the structure of the webpage, it became possible to scrape the required data. This experience taught me the importance of careful analysis and understanding of the webpage before starting the scraping process. Another challenge was dealing with time and space complexity. The code can take a significant amount of time to run, and it can consume a lot of memory when dealing with large datasets. This is an area where I can improve my skills by optimizing the code to reduce time and space complexity. The visualization aspect of the project could also be improved. Although the current visualizations are informative, they can be enhanced to better communicate the insights from the data. A better visualization would allow the data to be presented in a more concise and understandable way.

In conclusion, I was able to successfully scrape data from the website, clean and organize it, and create informative visualizations to gain valuable insights. However, upon reflection, I acknowledge that there is room for improvement in terms of the time and space complexity of the code, as well as the effectiveness of the visualizations. Despite the challenges I faced during the project, I was able to apply my course knowledge in data scraping, debugging, and problem-solving. The experience taught me the importance of carefully analyzing and optimizing the code, as well as continuously improving my skills in this area. As I continue to develop my skills in this field, I am confident that I will be able to create even more efficient and effective web scraping projects in the future.