{‘Team_8’ : [‘Riya Prajapati’, ‘Vivek Chowdhury’, ‘Jason Riffaterre’]}

Group Project Question 03

The final python script written to complete the requirements for CIS 9650’s Group Project Question 3 was called “scraping_teams.py”. This code is designed to scrape the names and player heights from multiple team rosters in four sports: men’s swimming, women's swimming, men's volleyball, and women’s volleyball. There were 20 webpages of team rosters in total, and these were evenly distributed between the sports.

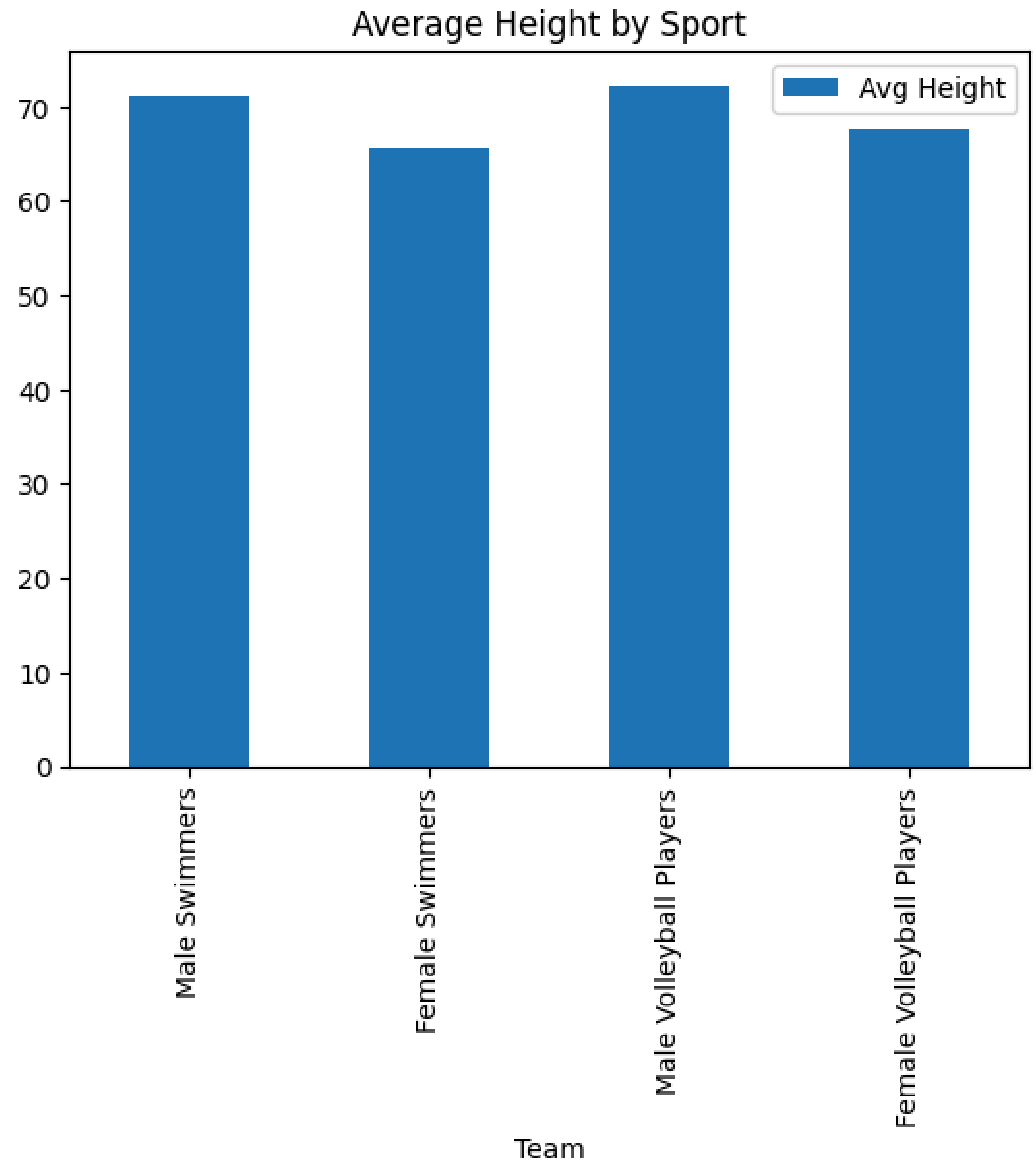
The product includes 4 pandas data frames that include a heading with the name of the sport, a list of the players' names and their heights in inches. Summary data for each sport also prints to screen. In each case, the data about the team names and heights is saved to a csv file that is named after the sport. For each dataframe (and in addition to the summary data obtained through the built in .describe() function) the code produces 2 tables [8 in all] of the five shortest and five tallest players, and it names the average height as a calculated value within the code. Each data frame also outputs its data to a database file. The integrity of this data was tested by opening the database files in VS code with sqlite viewer extension and comparing it to the original. Finally, a bar graph was produced to show the average height distribution across the four different sports.

Four dictionaries were created to track the websites where the rosters could be found for each sport. A dictionary format that included team names as keys was chosen over a simple list of URLs in order to support debugging. Since processing heights of players seemed the correct situation in which to use a class object, the ‘Tallness’ class was included as an experiment. In this context, it likely increased complexity without much value added, but was worth it for the practice. Because 40 team pages were being scraped, several soup methods and methods of processing this soup data made combining much of the code into one or two functions somewhat awkward and hard to read. As such, each step in the process has been separated into its own function to call. This aided in debugging so many methods. An example of how web pages differed in their handling of data is in how heights were originally displayed in feet and inches. The function ‘feetToInches’ was designed to transform height text objects from soup into arguments that could be applied in functions. This function required the use of various logic statements in order to select the correct method to translate the display values for heights. For example, 6-2, 6'2, and 6'2" were variations for height that needed to be parsed into two indexable values: [6, 2]. Similar logic was required to select the appropriate tags to use in searching for the names and heights of players within the roster.

There was also a challenge with web pages for Medgar Evers that were down at the time we were writing this code. But these cases were handled in python through network status checks prior to attempting to process soup data. Teams with broken pages were omitted from the final product as a result of this 404 error checking.

The *n* for these data sets was reasonably high (200 male and 154 female swimmers, as well as 126 male and 136 female volleyball players), so one would expect the results of this study to be fairly reliable. But this data revealed some unexpected results. The average height of a male swimmer was 71 inches, while that of a male volleyball player was 72 inches. For female swimmers it was 65 inches, while for female volleyball players it was 67 inches. The difference in height between players in the two sports seemed negligible compared to what one might have expected. It is easy to imagine that volleyball teams would have shown a preference for particularly tall players, and so one would have expected the volleyball players to be notably taller. But the graph included in this report shows that this was not the case. Perhaps at the elite levels in these sports there would be a greater differential between the average heights of swimmers and volleyball players, since they would be drawing from an ideal pool of athletes. And it appears that colleges may not have this same luxury of choice. But a similar scrape of elite athletic team data would need to be conducted to further this argument. If nothing else, this exercise revealed the importance of using data rather than intuition when drawing conclusions.

{‘Team_8’ : [‘Riya Prajapati’, ‘Vivek Chowdhury’, ‘Jason Riffaterre’]}


Average Height by Sport
Avg Height
Male Swimmers
Female Swimmers
Male Volleyball Players
Female Volleyball Players
Team