Name of activity: Module 04 Homework: Python and Web Scraper

Group Member: John Carpenter

Computing ID: jmc7dt

Group Member: Nathan England

Computing ID: nle4bz

Module 04 Homework: Python and Web Scraper Write Up

Introduction:

Our group built a Python program that web scraped game records for the University of Virginia's men's basketball team for the 2018-2019 season. After scraping the data, we wrote the data to a CSV file in order to analyze the scraped data. Upon performing our analysis of the data, we generated some interesting statistics describing the team's performance with regard to free throws and 3 pointers during the season, enabling us to drive insights regarding the team's performance last year. Finally, we decided to further manipulate the data by generating visualizations that lead to more interesting findings in an easily interpretable format.

Approach:

Our approach was to look at some data we found interesting, scrape this data using a Python program, and analyze the data using Python libraries. This allowed us to turn data into information. Our prior experience and background knowledge regarding the data allowed us to drive insights from this information.

Data:

We chose to look at the Virginia Men's basketball team's game records from last year because we both attended the university for our undergraduate degrees and were large UVA basketball fans so found looking at this data would be interesting (especially since we won the national championship last year). Specifically our data was scraped from the game logs from the 2018-2019 season (both regular season and NCAA tournament) which we found on the following webpage: https://www.sports-reference.com/cbb/schools/virginia/2019-gamelogs.html. These records mostly composed of quantitative metrics that are easy to generate summary statistics for (the data includes Virginia's statistics, our opponent's statistics, and the date in which the games were played).

Algorithm and Libraries:

The libraries used in our program were NumPy, pandas, urllib, BeautifulSoup, csv, and matplotlib.We used urllib to make a request to open up the above mentioned webpage. From there we used the BeautifulSoup library to scrape the html data into a more readable format. We

then used the csv library to write the scraped data into a csv file so that we could perform our analysis. In analyzing the data we used NumPy enabling us to perform large mathematical operations on arrays within our dataset. Pandas allowed us to organize and query out statistics from the data. Finally, we used matplotlib to further visualize the data for the Extra Credit part of this assignment.

Applications of Our Program:

Someone might want to use our program to know how specific metrics from UVA's prior season led them to win the national championship last year or potentially even predict what their record will look like this year. Some specific metrics we looked at in analyzing the data were total points scored for Virginia in each game, the number of points scored for the opposing team during each game, the total number of points Virginia scored during the year (2,714 points), the number of 3 pointers made during each game by Virginia, the percentage of total points that were 3 pointers out of Virginia's total points during the season (35.48%), and the number of games where the free throw percentage was greater than 75% (16 games).

Potential Improvements:

One potential improvement would be that our section in which we write data to the csv, we could get rid of the first row to allow us to use the second row as our keys (i.e. headers) when analyzing the data. We could also improve our for loops by implementing regular expressions to pull down extra data from the html code to analyze the entire website.

Extra Credit (see graphs below):

For the extra credit we chose to make two visualizations to draw insights from our data. The first visualizations is a plot of the number of 3 pointers that Virginia made versus their opponent each game. Creating this plot allows us to see which teams were the worst at defending against Virginia's 3 point attempts. As is clear from the plot, Syracuse was the worst at defending against Virginia with regard to 3 pointers (UVA made eighteen 3 pointers that game).

The second plot we generated was the margin of points that Virginia scored over their opponents during each game. Creating this plot allows us to see which games were close and which ones were not close games. As is clear from the plot, the Coppin State game was the largest margin of victory for Virginia while the Auburn game was the smallest margin of victory.

Works Cited:

https://docs.python.org/2/library/csv

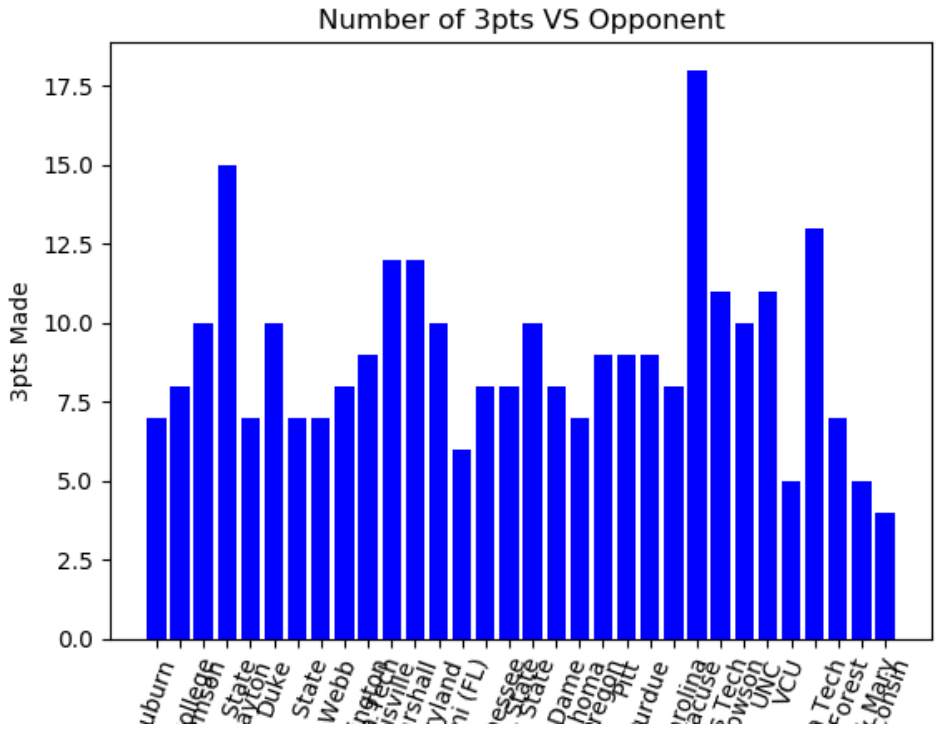
https://matplotlib.org

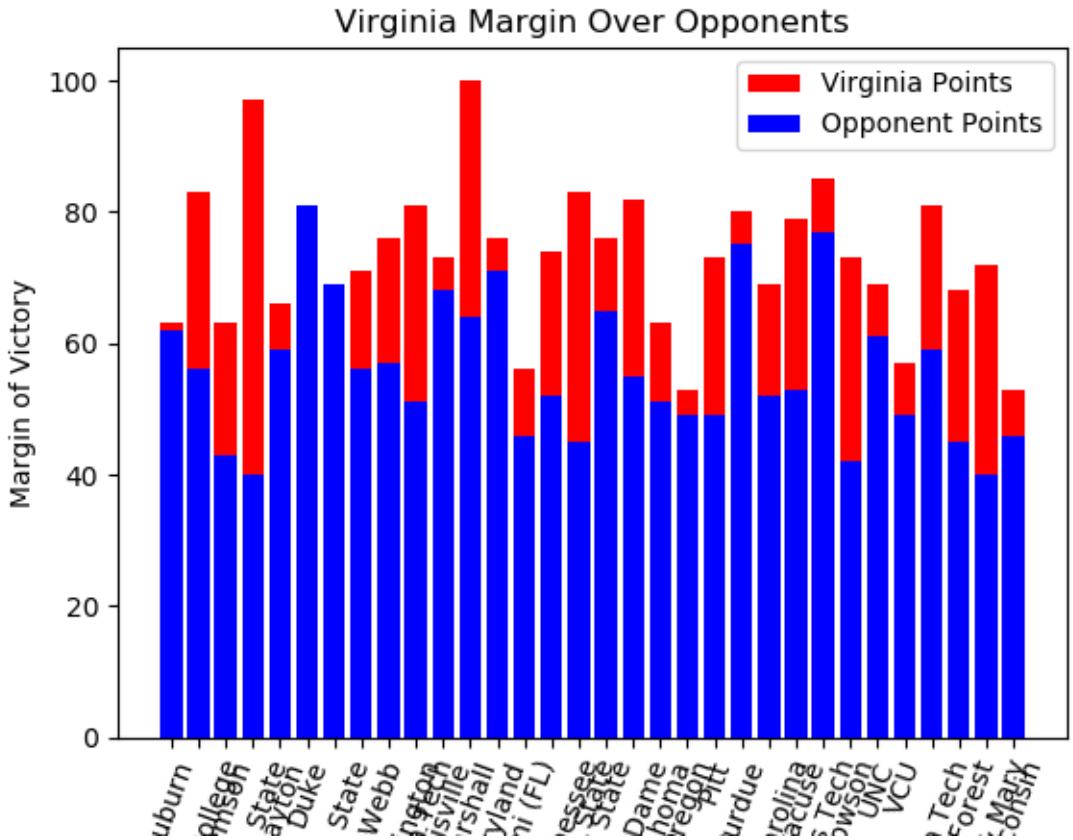
https://numpy.org

https://pandas.pydata.org

https://docs.python.org/2/library/urllib

https://www.pythonforbeginners.com/beautifulsoup/beautifulsoup-4-python




Virginia Margin Over Opponents
Virginia Points
Opponent Points
Margin of Victory
100
80
60
40
20
0
Duke
State
Webb
VCU