Lorena Vasquez
Computer Information Systems
April 16th 2022
Homework 2
Report on Web Scraping

To begin this project, I initiated a new python notebook on Google Collab. After opening the notebook, I have to install the corresponding libraries: BeautifulSoup, requests and pandas. BeautifulSoup is important to this project so that the user/reader is able to read the HTML from the website. Requests are important since we are using a URL. we want python to request the URL, much like how we request to go to Facebook.com. Lastly Pandas is useful as it allows me to create dataframes, manipulate the data frames. In doing so, being able to use Pandas allows us to retrieve the data from the websites into lists and then be able to produce them nicely within a dataframe.

To make it easier, I chose a website that had a pretty nice table to work with. The table has multiple columns and all I had to do is be able to read the data. I then asked the program to be able to request the URL to be able to inspect the page.

Once having the request of the url, I am able to process it and read it through BeautifulSoup. BeautifulSoup to be able to extract the page.content of the webpage within the html format. Soup prettify () enables the user to even further be able to read the data much nicer to be able to code through the data. After that, we find the table that we are wanting to extract.

I then found the table, and was able to retrieve the ‘tr’. The ‘table’ tag is just to find the table, once having the table the ‘tr’ tag allows us to be able to find the information within the table rows.

I then created an empty list for each of the columns. I then created a variable for the rows which found all the rows in the table. I then created a for loop to be able to reiterate through the rows and find all the cells where there is a ‘td’ tag. I know from my data, that there are 9 columns, meaning there should be in total 9 cells to be able to reiterate and fine the data. I then was able to find the cell information through indexing the lists, finding the text and making the text into a float. Finally all of that information is then inserted into the empty lists that I have created. I have the columns and the rows all together now in doing this loop.

One of the columns had a % attached to each one of the elements within the list. I then created a for loop to reiterate through the list and be able to split it from the number to the % sign. I then zipped all together the lists. I could have created a data series here, but I lit zipping the lists together. I then created a new columns list and was freely able to put any characters of my liking. I then temporarily set the index to the year to have a nice dataframe to present. Lastly I did statistical analytics in utilizing the describe method in finding the 8 statistical data points. Not all of my columns were able to be into the describe command, as they are not numerical values.

Lastly, the last step is to export my data frame as a CSV file, in doing so, I wrote a line of code with the file name and made sure I added ‘.csv’ at the end of my filename!