***Part I: Python***

Develop a web links scraper program in Python that extracts *all* of the unique web links that point out to other web pages from the HTML code of the “Current Estimates” web link and that populates them in a comma-separated values (CSV) file as absolute uniform resource indicators (URIs).

A.  Explain how the Python program extracts the web links from the HTML code of the “Current Estimates” web link.

*The python script utilizes the beautiful soup library to parse the HTML from the link provided. After extracting the HTML from the webpage into memory, the script will use the beautiful soup library to find all of the ‘A’ elements that contain ‘HREF’ links. The list returned is specifically all of the links from the webpage.*

B.  Explain the criteria you used to determine if a link is a locator to another HTML page. Specify the code segment that executes this action as part of your explanation.

*To differentiate between internal links and external links, the following code segment was implemented:*

*tf=a['href'].startswith('http')*

*This sets the variable tf to a Boolean value based on whether the link provided starts with HTTP or not.*

C.  Explain how the program ensures that relative links are saved as absolute URIs in the output file. Specify the code segment that executes this action as part of your explanation.

*To following logic will manipulate the internal links to be outputted in a working format to the csv file.*

*if tf == True:*

*link=a['href']*

*else:*

*link=baseLink + a['href']*

If the boolean logic from 1.b returns TRUE (it starts with HTTP) then the link will export as is, with no change. If the Boolean logic from 1.b returns FALSE then the link will be manipulated to add the base URL to the beginning.

D.  Explain how the program ensures that there are no duplicated links in the output file. Specify the code that executes this action as part of your explanation.

The links found are added into a set, linkSet. The set type will automatically check the existing set for duplicates before adding a new record. This is the way the program reports on each link only one time.

E.  Provide the Python code you wrote to extract *all* the unique web links from the HTML code of the “Current Estimates” web link that point out to other HTML pages.

r = requests.get(search\_link)

raw\_html = r.text

soup = BeautifulSoup(raw\_html, 'html.parser')

for a in soup.find\_all('a', href=True):

The code above will

F.  Provide the HTML code of the “Current Estimates” web page.

Provided as document: htmlFile.txt

G.  Provide the CSV file that your script created.

Provided as document: links.csv

H.  Test your script and provide a screenshot of the successfully executed results.

Provided as document: results.jpg

***Part II: SQL***

I.  Identify the differences in the population size estimates for each U.S. state the Census Bureau provided in two consecutive years using the most current data and the latest historical data datasets for the national total population in two different SQL tables.

Included as document PopDifferences.csv

J.  Write a code to join the two tables on the year and state fields into one SQL table that identifies the absolute differences (in whole rounded hundreds) in the estimates of 10,000 individuals or more between the two datasets. If the earlier estimates are larger than 10,000, the cells should indicate a negative value. Provide a screenshot of your tested code showing successful execution.

Screenshot provided: sqlResults.jpg

K.  Explain how you prepared the data and how the datasets were imported into two SQL tables. Provide a screenshot of the successfully executed SQL code.

The dataset was prepared using formulas in excel spreadsheet corresponding to each row in the excel file.

Screenshot provided: sqlInserts.jpg

The data was prepared using excel formulas in the excel document downloaded from the census website. Within the excel document, the 2015 and 2016 columns were separated and used to create a series of insert statements. Here is an example of the formula used within excel to create an insert:

="insert into pop2015(year, state, population) Values(2015,'"&M10&"',"&N10&");"

Where the M value correlates with the statue column and the N value correlates to the population column. All of the periods were removed from the beginning of the state names using find and replace.

L.  Export the data from the SQL table into a CSV file, with rows representing the states and columns representing the years that both datasets estimate, that only shows the differences between the datasets (in whole rounded tens of thousands) that exceed 10,000 individuals.

popDifferencesPartJ.csv

***Part III: R***

M.  Create a linear regression analysis with R to predict the size of the population for the state you live in for 2020 based on the Current Estimates Data dataset.

N.  Explain how you prepared the data and how the dataset was imported into R, including a screenshot of your results.

O.  Using the estimates for the most recent year in the dataset, create an R script to display a histogram (using one million as the interval size) of the current estimated population size of your state. Provide a screenshot of your results.

P.  Create an R script that will tabulate a statistical description of the estimated 2020 data. Provide a screenshot of your results.

Q.  Predict the population size of your state using a linear regression. Provide a screenshot of your results.

R.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.