C742 Project

Chase Clark

# Census Analysis

## Part I – Python

### A.

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 the ‘A’ elements that contain ‘HREF’ links. The list returned is specifically the links from the webpage

### B.

To differentiate between internal 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. If the links starts with http it is a good indicator that it is an external link.

### C.

The 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, then the link will export as is, with no changes. If the Boolean logic from 1.B returns False, then the link will be manipulated to add the base URL to the beginning.

### D.

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.

See scraper.py

### F.

See htmlFile.txt

### G.

See links.csv

### H.

See results.jpg

## Part II – SQL

### I.

See popDifferences.csv

### J.

See popDifferencesPartJ.csv

See sqlQueryResults.jpg

### K.

See sqlInsert.jpg

### L.

See sqlResults.jpg

See sqlInsert.csv

The dataset was prepared using the current population estimates spreadsheet from the census website. There were periods leading all of the state names that were scrubbed from the data. Afterwards the cells containing state names and the cells containing their population were related using an excel formula. The excel formula would relate the cells and format them in a SQL insert statement. First the formula was applied to the 2015 data and then the 2016 data to create two sets of inserts for two tables. Here is one insert from each data set.

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

="insert into pop2016(year, state, population) Values(2016,'"&M10&"',"&N10&");"

Here M was the column that containing the name of each state and N was the column containing population values for that year. First the N column contained 2015 population data. After the inserts were created, the N column was replaced with 2016 data and a new set of inserts were created. The formula needed to be slightly changed between sets to account for the new year and table name.

## Part III – R

### M.

See 2020est.png

Based on the linear regression in 2020est.png, one can infer the population of Florida would increase by 290,376 residents per year. By this logic, it is inferred that the population for Florida will be 21,692,307 in the year 2020.

### N.

The data was prepared by selecting the population data and adding it into a .tab file. All of the commas from the population values were purged with find/replace. Each population value was entered on a new row and the corresponding year was added in tandem on the same row, one tab apart.

See floPop.tab

### O.

### P.

### Q.

See 2020est.png

Based on the linear regression in 2020est.png, one can infer the population of Florida would increase by 290,376 residents per year. By this logic, it is inferred that the population for Florida will be 20,821,178 in the year 2017.

### R.