Assessment C996: Taylor Van Daff

1. The Python program extracts the web links from the HTML code of the “Current Estimates,” found in web links section:
   1. The program first defines URL as the current estimates url: 'https://www.census.gov/programs-surveys/popest.html'
   2. The program then uses requests.get() function to retrieve the URL and Beautiful Soup to parse the HTML code into a readable format.
2. To find links that are locators to another HTML page:
   1. When reviewing the HTML code, it was evident that all of the web links were tagged under <a, and under the name href. I created the list, bsurls, in which to place all the web links matching this initial criteria.



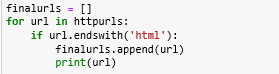
* 1. However, in the webpage there are also <a tags with the name href that do not contain any information. These cases are identified by ‘None’ in the bsurl list. To remove the NaN values and the extraneous input, I developed a new list that loops through bsurls.



* 1. After that, there were some URLs that were missing the first half of their respective address. I developed the httpurls to loop through valurls and add the first half of the URL back onto the link.



* 1. The final criteria I used was to address whether each webpage ended in html. In the case that the websites didn’t, they were removed.



1. To remove duplicates from the URLs:
   1. I used the duplicated function and then selected the keep= ‘first’ method.



1. To ensure URLs that are outbound are separated:
   1. I used the base web address: ‘https://www.census.gov’ to identify the inbound addresses and separate them and the else command to separate the remaining web links into a separate list.



1. Screenshot of final results:

