Project 1. Crime and Border

Mother should I build
the wall?

Pink Floyd, "The Wall"

There is a continuous discussion about whether or not the [illegal] immigrants infiltrating the USA from Mexico positively affect the crime rate in the border neighborhoods. Your assignment is to cast some light on the situation using data from Wikipedia.

Write a program that shall:

1. Download and extract crime-related data (namely, the total violent crime rate per 100,000 people) from the major US cities from the List of US cities by crime rate. https://en.wikipedia.org/wiki/List_of_United_States_cities_by_crime_rate
2. For each city, follow the link to its Wikipedia page and extract its coordinates. Calculate the distance to the border as the smallest distance from each city to San Ysidro, Yuma, Tucson, El Paso, Laredo, Del Rio, and Brownsville, TX. You are allowed to hardcode the coordinates of these locations. Use the great-circle distance formula.
3. Save the city names, crime rates, and smallest distances to the border in a 3-column CSV file.
4. Calculate and display the correlation between the two numbers (use scipy.corrcoeff()).

Finally, scatter plot crime rates against the distances. You can use Excel, Google Sheets, LibreOffice, or matplotlib. Add a trend line if you want. Write a one-page report that explains in layperson's terms the procedure of your discovery, your findings (especially the correlation), and their significance. Include the scatter plot in the report. For the report, use MS Word, Google Docs, or LibreOffice.

Save each downloaded HTML document into the directory data. If the directory does not exist, create it. If a needed document has been previously downloaded, it shall be retrieved from the file, not from the Web[1].

[1] Consider writing function getDocument(URL) that checks if the document has been stored in the data directory. If it is, the function shall read it from the file. If it is not or the directory does not exist yet, the function shall create the directory, download the document, save it to an appropriately named file, and return its content. Do not use any absolute file paths in your code, as I will not be able to reproduce your results.

You must use modules CSV and BeautifulSoup. You will probably need module urllib. You are not allowed to use NumPy or Pandas. The use of module matplotlib is optional.

Deliverable (to be submitted to BlackBoard): Your code and the report. Do not include the HTML files.