Isa Gaillard
3/17/2020
CP 255

Assignment 2: Importing and Assessing Environmental Justice Databases

Overview:

As mentioned in the beginning of my Assignment 2 Jupyter notebook, the primary focus of my work early in this assignment was on uploading and cleaning the three different versions of CalEnviroScreen (CES): 1.1, 2.0, and 3.0. This took a considerable amount of time because they are large datasets and some of their variables differ in key ways. Once I was able to clean the data, I transitioned to conducting regression analysis, focused mostly on the relationships between air pollution indicators such as ozone, diesel, and particulate matter and their effect on cardiovascular disease and asthma rates across California. Towards the end of my notebook, I created a few data visualizations, running code to conduct simple linear regression between variables of interest. I used scatter plots to visualize these relationships and was left wondering how I can re-code these scatter plot cells in order to improve these visualizations. I've included a "Questions" section at the end of this write-up that I hope can guide the instructor's feedback. I also plan to come to office hours to talk through potential solutions to my questions.

Challenges:

Even though I did not begin analyzing data at a more granular level than the state yet, one major challenge that I've started to consider and that will be important to address in upcoming analysis is that the CES 1.1. Dataset only includes data at the zip code level while 2.0 and 3.0 have data at the census tract level. As mentioned in my assignment notebook, I'm curious to hear any advice the instructors have as to how I can overcome this challenge. I am also open to just conducting geography-specific analysis using CES 2.0 and 3.0 since the data is ready to go for those two datasets and given the limited time I have to work on this assignment. Another challenge was that my scatter plots don't do a good job of displaying the data in a compelling way because the end X and Y axis values aren't the same across different scatter plots. I've included a question about this below.

Successes:

It was very rewarding to have three cleaned and usable datasets that I was able to begin analyzing and running regressions on. I have already started to uncover some interesting connections such as the fact that diesel is more closely correlated with asthma while high ozone emissions are more closely correlated with cardiovascular disease across the state. Successfully visualizing the relationship between CES 3 Score and asthma was a success because it clearly demonstrated the strength of cumulative impacts on health and underscore the importance of comprehensive strategies to address cumulative pollution and socioeconomic burdens on communities.

Questions & Next Steps:

My immediate next steps are to learn how to and begin conducting analysis at the census tract and zip code level, especially where environmental justice policies were put into place in years between when the CES data was gathered. I.e. much of the CES 1.1. data was gathered in 2009 while CES 3.0 data was gathered in 2012 or later. So the data should be especially helpful for shedding light on the efficacy of policies that were implemented between 2009 and 2012. However, there will certainly be limitations to this analysis, so finding out how to overcome this if possible or at least effectively respond to them will be vital. Below, I have listed a few questions I hope to get direct feedback on. If it would be easier to discuss answers to these questions via a zoom office hour meeting or phone call I am open to that and will reach out to set up an appointment soon.

1. There are errors in the last few variables of the cleaned_ces2_data (see Out [7] "education_percentile" through "population_characteristic_score_percentile" towards the end of the data table. I'm not sure how to fix this error.
    a. There are also some nan values in the cleaned_ces3_data file that I'm not sure how to fix. (See Out [9]) ces_3_percentile).
2. How can I improve my scatter plots i.e. how can I change the values on the x and y axis to match across different scatter plots? (See Out [51] & Out [58] scatterplots) The X and Y axis for the scatterplots above for CES 1.1. end at at 300 and 120 and for 2.0 they end at 200 and 150.
    a. I could also use help learning how to improve the dynamic scatterplot (see Out [78]).