PAI 789 Advanced Policy Analysis Final Project (Jiyoung Park)

Research Question

The main idea of this project to see which factor can explain the state of spread of covid-19. I tried to look into the data at a county level. The initial thought was that population would be the most heavily related to the total positive cases and deaths. The next thought is that when it comes to the frequency of travel outside of state, the education attainment level of counties also might have contributed to the spread with the assumption that people with high education attainment are likely to travel more due to their jobs.

Building Codes

This project consists of three steps with corresponding three py files.

1. **Population Data Obtainment**
    1) The 'population.py' file contains codes for obtaining population data as a csv file from the census api. Since the file has the population data for 10 years from 2010 to 2019, the file will be trimmed for analysis at the final stage, right before merging.
    2) The population data is set to be collected by state, based on counties. Therefore, to obtain the data for a certain state, the user should know the state fips code and put it into the codes manually.
        a) Put the state FIPS code in the variables 'in_clause' and 'popfilename.'
        b) Use your api key value to get the data.

2. **Education Attainment Level Data Obtainment**
    1) The 'attainment.py' file has codes for obtaining the education attainment level data of county residents as a csv file from the census api. The same data mining was done in the previous assignment G18. The ratio of highly educated people with more than college degree divided by people with less than high school graduation will be merged with the covid-19 data.
    2) To get a dataset, the file 'census-variable.csv' is needed first to get keys for calling api.
    3) Same with the population data, the education attainment data is set to be collected by state. Therefore, the fips code of the state in interest should be put manually as well like the population data.
        a) Put the state FIPS code in the variables 'in_clause', the filename of 'attain.to_csv' command, and the variable 'edufilename.'
        b) Use your api key value to get the data.

3. **Covid-19 Data Cleaning**
    1) The New York Times covid-19 data is written on a daily basis. I obtained a state-and-county-specified dataset.

[U.S. County-Level Data](us-counties.csv) ([Raw CSV](https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv))]

2) Open a csv file which is in the covid data zip file.
3) Make two separated columns for the state and the county FIPS codes by slicing 5-digit FIPS codes, to merge the dataset with other datasets later: 'fps_stt' for state FIPS codes and 'fps_cty' for county FIPS codes.
4) To sum up the total cases and deaths by month, extracting only month from the date column. In this project, I used the total cases and death only, but it will be also useful to analyze the data by month to see the spreading rate. It creates a new column 'month_a' indicating months as abbreviated names.
5) To analyze the data by state, it is useful to have a dictionary 'state_fps' for the state name and FIPS codes. In the dataset, there are some missing values. To build a dictionary, first clean up missing values and convert a data frame into a dictionary.
6) Next is to build functions in order to slice data by state and calculate monthly cases and deaths by county: a 'daily_by_state' and 'monthly_by_county.' If you put a lowercase state name, the function finds its FIPS code, slices the data, and saves it into a csv file with the assigned file name by the function. The next function is to sum up all daily cases and deaths by month and save it into a csv file. After defining functions, all you need is to have a lowercase state name and run functions. ('statename') Two functions will create two csv files.
7) For creating new columns for total cases and total deaths in the data set created at 6), I built another function called 'total_cases_deaths.' Using a file made at 6), it adds two columns in the data set and saves it to a new csv file. The final covid data set will be 'state_covid.'
8) Now, merge the cleaned covid data with the population and the education attainment datasets.
    a) Read the population and the education attainment csv files and clean the population data to have the 2019 population only. Each dataset will be 'pop_df' and 'attain.'
    b) Merge the 'state_covid' dataset with the 'attain.' I will call the merged dataset as 'covid_ed.' Then drop duplicated and missing values. After merging, calculate fatality rate by dividing the total deaths by total cases and add the result as a new column.
    c) Merge the 'covid_ed' dataset with the population dataset 'pop_df.' The final merged dataset will be called 'covid_ed_pop.' After merging, calculate the total cases per 1,000 population ('cases_per_pop') and add the result as a new column.
    d) Save the final dataset using a variable 'finalfile.'
9) Make scatter plots by using columns to analyze the relationship within a state.