Assignment: Data Manipulation, Cleaning, and Visualization

Due 2/19 11:59pm

February 4, 2021

1. Load RAW_us_deaths.csv and RAW_us_confirmed_cases.csv (they can be found in homework_1 in files on courseworks) into pandas dataframes. Each row in these dataframes contains the timeseries of cumulatives deaths and cases for a municipality, in addition to some information about the city (lat, long/ region, etc).

2. Rename "Province_State" and "Admin2" to state and city respectively. Make these the index

3. Create a new pair of dataframes where you drop all of the columns that are not a date in the time series.

4. Melt these dataframes into long format (the columns with be state, city, date, and value)

5. Add another column called "type" to each of the long dataframes. Set it equal to "deaths" for the death dataframe, and "cases" for the case dataframe.

6. Combine the two long dataframes into a single long dataframe. Use pd.to_datetime to convert the date column from string to datetime objects.

7. Use a groupby and a pivot to create a dataframe that calculates the cumulative deaths and cases for every state for every day. This means the index columns will be state and date, and the value columns will be cases and deaths.

8. Since the the data is given as cumulative deaths/cases, we can use the following code to create a dataframe that has deaths/cases that were recorded on a certain day using the following code (suppose the dataframe from part 7 is called piv):

   piv2 = piv - piv.groupby('state').shift(1).fillna(0.0)

   Explain, in words, what this is doing.

9. Using this dataframe created in part 8, find the dates on which each state recorded the most cases and deaths.

10. RAW_us_deaths.csv contains a field for the population of each state. Use this information, determine for every date, which state had the greatest number of deaths per capita.

11. Make a lineplot for the time series of cumulative deaths for Texas, Louisiana, New York, Florida, New Jersey, Connecticut and California.

12. Create a stacked bar chart with each of the state's from part 11's total cases and total deaths.

13. Write a function that use matplotlib to create a pairplot (histograms on the diagonal, scatterplots on the off-diagonal) for a given dataframe. Use this to make a pairplot for daily deaths in the above states.