CS217 HW 7 – pandas

Overview

In this assignment, we will continue exploring some data related to the COVID-19 pandemic. There are many different ways to access data. In the last assignment, we used data API, which is commonly used to obtain data that are frequently being updated. This time, we will use a CSV file directly, which contains a lumpsum of data.

The data we will use is from New York Times. It is hosted and being maintained at https://github.com/nytimes/covid-19-data. Since it's being updated every day, we will use one specific copy of the data. You need to download the file (`us-counties2020.csv`) from canvas and use pandas to accomplish the following tasks.

Similar to the json assignment, we will mainly focus on two values of COVID-19 statistics:

- `deaths` – an integer referring to the cumulative death count
- `cases` – an integer referring to the cumulative case count

Q0: Load the data. Please load the dataset to a pandas dataframe. Make sure the csv file is located in the directory where your py file can access without specifying the absolute path.

Q1 (1pt): Write a function `num_entries` to extract the total number of rows in this dataset. The function takes no input and returns an integer.

Q2 (1pt): Write a function `num_states` to extract the number of unique state/territories in this dataset. The function takes no input and returns an integer.

Q3 (1pt): Write a function `num_cty` to extract the number of unique counties for a given state in this dataset. The function takes a string input `state` and returns an integer.
Note: there are some data entries with "Unknown" counties. For simplicity, we treat "Unknown" as a unique county name in this assignment.

Some code for verification:

```
>>> num_cty('Alabama')
67
>>> type(num_cty('Alabama'))
<class 'int'>
>>> num_cty('Arizona')
16
>>> num_cty('Arkansas')
76
```

Q4 (1pt): Write a function `num_cases_cty` that returns the cumulative case count of a given county in a given state, till a given date. The function takes three string variables `state`, `county` and `date` and returns an integer.
Hint: there are some cases where no data can be retrieved. It means that the count is 0 for this state, county and date. We should return 0 in this case.

Some code for verification:

```
>>> num_cases_cty("Illinois", "Cook", "2020-04-19")
21272
>>> type(num_cases_cty("Illinois", "Cook", "2020-04-19"))
<class 'int'>
>>> num_cases_cty("Nonexist", "Champaign", "2020-08-15")
0
>>> num_cases_cty("New York", "New York City", "2020-12-01")
319301
```

Q5 (1pt): Write a function `num_cases_state` that returns the cumulative case count of a given state till a given date. The function takes three string variables `state` and `date` and returns an integer.
Hint: we are looking at state-level data. You will need to obtain the sum of confirmed cases in each county.

Some code for verification:

```
>>> num_cases_state("Illinois", "2020-04-19")
30357
>>> num_cases_state("California", "2020-12-19")
1846943
>>> num_cases_state("Nonexist", "2020-08-19")
0
```

Q6 (2pt): Write a function `cty_beyond_thold` to obtain a list of counties with cumulative confirmed cases greater than or equal to the threshold on a given date. The function takes two string variables (`state`, `date`) and an integer variable (`threshold`), and returns a list of county names.

Some code for verification:

```
>>> cty_beyond_thold("Illinois", "2020-04-04", 500)
['Cook', 'DuPage', 'Lake', 'Will']
>>> cty_beyond_thold("Illinois", "2020-03-04", 500)
[]
>>> cty_beyond_thold("Nevada", "2020-10-19", 1000)
['Clark', 'Elko', 'Washoe']
```

Q7 (1pt): Write a function `first_case` to obtain the date of the earliest case in a specified state. The function takes one string parameter `state` and returns a string of the date. The parameter `state` should take a default value of `"Illinois"` and return the date of the earliest case in Illinois by default.

Some code for verification:

```
>>> first_case()
'2020-01-24'
>>> first_case("New Mexico")
'2020-03-11'
>>> first_case("California")
'2020-01-25
```

Q8 (2pt): We define a state as a pivot state if the average death count across all counties is greater than 500 by 2020/12/31. Write a function `pivot_state` to find all pivot states. The function takes no input but returns a list of state names.

Submission:

Submit the completed py file ONLY. No need to rename the file.
DO NOT call any function that you wrote in the py file. That is, please only keep the definition of the functions. Comment out all your function calls.
DO NOT print any information in your functions.
DO NOT change any function names!
DO NOT submit the data file since it is too large!