BMIS2542-Data Programming Essentials with Python

Sessions-2&3 ***Challenge*** option:Data Wrangling with Pandas and EPA dataset

Use [EPA datasets posted here](https://pitt-my.sharepoint.com/:f:/g/personal/narayanr_pitt_edu/EqOZpO62G2xOu9DcAl5x3mEBbp0kt6sD8vNZt6i1d2fsUQ?e=OkZyo5), *EPA-data-2015.csv* and *EPA-data-2016.csv*, for this exercise. Read the datasets into Python as Pandas dataframes and answer the following questions. You may use any libraries available in the Python ecosystem:

1. Comparing the 2015 data with 2016 data:
   1. Have the overall missing values in the EPA data increased or decreased between 2015 and 2016?
   2. List the states for which the missing values in the data have increased and the states for which the missing values in the data have decreased.
   3. Create a new Pandas dataframe with the following columns, populate the dataframe, and write it out as a CSV file:
      * Year
      * Month
      * StateName
      * CountyName
      * Missing\_Proportion: this is ratio of missing observations to the total observations reported for the county in the corresponding month and year.
2. Use “AQI” and “Arithmetic Mean” (of PM 2.5) as indicators of air quality to answer the following questions:
   1. On average, from 2015 to 2016 has pollution in the U.S. increased or decreased?
   2. Insert a column in the 2015 and 2016 dataframes and call it “AQI category.” Populate the values for AQI category as following:

|  |  |
| --- | --- |
| AQI Category | AQI value |
| Good | 0-50 |
| Moderate | 51-100 |
| Unhealthy-sensitive | 101-150 |
| Unhealthy | 151-200 |
| Very unhealthy | 201-300 |
| Hazardous | >300 |

* 1. When AQI levels are in the 0-50 range, air quality is considered as “Good.” How many days did Pittsburgh not meet the good air quality standard in 2015? How about in 2016?
  2. Compare and comment on the air quality levels between Pittsburgh and Philadelphia.
  3. In 2015 and 2016, on the days when Pittsburgh did not have good air quality, did Philadelphia have good air quality?
  4. If you rank order the U.S. cities featured in the datasets according to air quality, where does Pittsburgh feature in 2015 and 2016? Provide the absolute rank and also the rank in percentile terms.

1. Use the “Event Type” variable in the dataset to answer the following questions:
   1. Create an indicator (or dummy) variable in the datasets and name it “eventdummy”; eventdummy should be set to 0 if Event Type=”None”, else it should be set as 1.
   2. Using the eventdummy variable, find out if at any given “Local Site” air quality is impacted by the events that the site experiences. Group the states into commonly known regions of the U.S. [[1]](#footnote-1), and see if there is variation in your finding across the different regions.
2. Comment on the presence of outliers in the data.
3. Combine the two datasets and then create the following:
   1. A new variable that is a combination of “State Code” and “County Name”; call it “state-county.”
   2. Create a pivot table with “state-county,” “month” and “year” as the indices. The values of AQI must be averaged in the pivot table. Then, query the pivot table you created for the values of cities with the phrase “New” in them (New York, Newport, etc.).

1. For example, this Wikipedia page lists the different U.S. regions considered in the U.S. Census: <https://en.wikipedia.org/wiki/List_of_regions_of_the_United_States> [↑](#footnote-ref-1)