DATA MINING ASSIGNMENNT 1

COVID CASES AND VACCINE DATA ANALYSIS

Requirements:

- Python version 3.8 or 3.9
- **Python Libraries needed:**
  - Pandas 1.3.2
  - Numpy 1.21.1
  - Python-dateutil 2.8.2
  - Json
  - Csv

Submission Structure:

1) Dataset

   a) Calendar.csv – Contains date and their corresponding weekid and monthid for analysis of no of COVID cases

   b) Calendar2.csv – Contains date and their respective weekid and monthid for analysis of Vaccine data.

   c) Census1.csv – Modified Census data consisting of only those columns needed for analysis and with some data preprocessing to make it suitable for analysis like combining districts of Telangana, Assam etc.

   d) districts_2.csv – Modified District data with data preprocessing done to match district names across all datasets and addition of new column required for analysis.

   e) neighbor-districts-modified.json – Modified neighbor-districts.json to match district names across datasets and some districts combined due to lack of data available for analysis in either Vaccine or Covid cases dataset.

HARSH AGARWAL
21111030
M.TECH CSE

   f) vaccine.csv – Modified Cowin-vaccine data such that row timeseries is converted into column timeseries with only a subset of columns from original required for analysis and some districts are combined.

2) Output – Contains csv files generated as output in Q2 - Q9. In total 23 csv files are generated as output.

   a) Q2 – edge-graph.csv
   b) Q3 – cases-week.csv, cases-month.csv, cases-overall.csv
   c) Q4 – district-peaks.csv, state-peaks.csv, overall-peaks.csv
   d) Q5 – district-vaccinated-count-week.csv, district-vaccinated-count-month.csv, district-vaccinated-count-overall.csv, state-vaccinated-count-week.csv, state-vaccinated-count-month.csv, state-vaccinated-count-overall.csv
   e) Q6 – district-vaccination-population-ratio.csv, state-vaccination-population-ratio.csv, overall-vaccination-population-ratio.csv
   f) Q7 – district-vaccine-type-ratio.csv, state-vaccine-type-ratio.csv, overall-vaccine-type-ratio.csv
   g) Q8 – district-vaccinated-dose-ratio.csv, state-vaccinated-dose-ratio.csv, overall-vaccinated-dose-ratio.csv
   h) Q9 – complete-vaccination.csv

3) Python – Contains python scripts for each question

   a) Q2 – edge-generator.py
      Input : neighbor-districts-modified.json
   b) Q3 – case-generator.py
      Input : districts_2.csv and Calendar.csv
   c) Q4 – peaks-generator.py
      Input : districts_2.csv and Calendar.csv
   d) Q5 – vaccinated-count-generator.py
      Input : vaccine.csv and Calendar2.csv
   e) Q6 – vaccination-population-ratio-generator.py
      Input : Census1.csv and vaccine.csv

f) Q7 – vaccine-type-ratio-generator.py
Input : vaccine.csv

g) Q8 – vaccinated-dose-ratio-generator.py
Input : Census1.csv and vaccine.csv

h) Q9 – complete-vaccination.py
Input : vaccine.csv , Census1.csv and Calendar2.csv

4) Script – Contains .sh files to execute python scripts

a) Complete Assignment – assign1.sh
b) Q2 – edge-generator.sh
c) Q3 – case-generator.sh
d) Q4 – peaks-generator.sh
e) Q5 – vaccinated-count-generator.sh
f) Q6 – vaccination-population-ratio-generator.sh
g) Q7 – vaccine-type-ratio-generator.sh
h) Q8 – vaccinated-ratio-generator.sh
i) Q9 – complete-vaccination-generator.sh

Steps To Execute Assignment:

1. Move into the ‘Scripts’ folder and open Terminal/Command Prompt etc.
2. To run complete assignment, type ‘bash assign1.sh’, all the generated csv files will be stored in the ‘Output’ folder.

Note: You may delete one or more csv files from the ‘Output’ folder but do not delete ‘Output’ folder itself.

3. You may even run particular question script individually by typing ‘bash filename.sh’ and generated csv is stored in ‘Output’ folder.

NOTE:

1. **Do not delete ‘Output’ folder or else scripts will start generating error as all the csv files generated are stored in the ‘Output’ folder.**

2. **Files in Dataset are pre-processed and formatted for use in analysis, so do not replace or modify them.**
3. Most of the data cleaning and processing was done using Microsoft Excel and some part was done using python scripts. In Excel, formulas like ‘Concatenate’, ‘VLOOKUP’ and other formulas and filters were used to do the above task.
4. Covid data for majority of districts is considered from 26-04-2020 to 14-08-2021 as data was available from that point of time and for some districts covid cases data was available from even later date, so for those districts analysis is available from corresponding weekid and monthid onwards.
5. In all cases, where generated numeric value was negative, it has been replaced with 0, to avoid erroneous analysis and have suitable output.
6. To find the date corresponding to a weekid or monthid, refer Calendar.csv – weekid and monthid for Q3 and Calendar2.csv for Q5 and Q9.
7. For Q4 , To find the date corresponding to weekid mentioned in district-peaks.csv or state-peaks.csv refer Calendar.csv, if weekid (n) is even then refer weekid2 column with value (n/2) or if weekid (n) is odd then refer weekid column with value ((n+1)/2). For monthid simply refer monthid column with the value in output file.
8. Q1, Variations in district names between neighbor-district.json and Vaccine data was found and corrected using python dictionary and string functions like split and then remaining difference in district name were corrected manually using Sublime Text ‘find and replace’ functionality. Some districts were combined into corresponding states like for Delhi, Assam, Telangana, Manipur, Goa, Andaman and Nicobar Islands and Sikkim.
9. Q1, neighbor-districts-modified.json though output for Q1, is stored in Dataset folder as it is input for Q2.
10. Q2, Generated edge list contains only (district i, district j) and not (district j, district i) that is, only unique district pairs are present to avoid addition of redundant edge data.
11. Q4, to find peaks, daily confirmed cases were used to find weekly and monthly totals.
12. Q7, for some districts ratio is given as inf in output signifying the fact that 0 Covaxin Dose were administered in that region during considered period.

HARSH AGARWAL
21111030
M.TECH CSE