CS685A: Assignment 1

Analysis of Covid, Vaccination and Census Data

Problem Description

- Can be found in the Assignment file.

Requirements:

- Python3 should be present on your system and should be accessible using the command `python3`

Running the Project:

- To run the complete project in one go, execute `assign1.sh`

```
./assign1.sh
```

- To run a certain segment, use the following scripts:
    1. `neighbor-districts-modified.sh` : To solve Q1
    2. `edge-generator.sh` : To solve Q2
    3. `case-generator.sh` : To solve Q3
    4. `peaks-generator.sh` : To solve Q4
    5. `vaccinated-count-generator.sh` : To solve Q5
    6. `vaccination-population-ratio-generator.sh` : To solve Q6
    7. `vaccine-type-ratio-generator.sh` : To solve Q7
    8. `vaccinated-ratio-generator.sh` : To solve Q8
    9. `complete-vaccination-generator.sh` : To solve Q9

- The individual scripts ***need to be executed*** in the order in which they are mentioned, as the output generated by one script is needed for the execution of the further scripts.

Directories:

- out/ : Stores the required output json and csv files.
- data/ : Modified and Refined data files used in the project.
- meta/ : Stores the meta data (a set of json files) generated by the execution of scripts. The contents inside this folder ***should not*** be modified when the scripts are being executed. The folder can be deleted before beginning the execution of `assign1.sh`. These files are essentially a transformation of dataset into a more usable format.
- base/ : Contains the initial unmodified data files.
- util/ : Contains python script(s) that have snippets used to clean and refine the original data.

Python Scripts:

- meta.py : Converts the csv files to json format for use by the scripts later on. It also finds issues in multiple data sources (district pairs with edit distances < x, districts contained in another, etc.)
- q1.py : Solves Question 1. Uses `neighbor-districts.json` and `district_wise.csv` to generate `neighbor-districts-modified.json`. Also see **Explaination 1**.
- q2.py : Creates au undirected graph from `neighbor-districts-modified.json`. This graph is printed in edge list format `edge-graph.csv`. Also see **Explaination 2**.
- q3.py : Uses `districts.csv` to determine the number of cases (weekly, monthly and overall) are reported for all districts (in `cases-week.csv`, `cases-month.csv` and `cases-overall.csv` respectively). Also see **Explaination 3**.
- q4.py : Find the 2 peaks on a district, state and overall level and generates `district-peaks.csv`, `state-peaks.csv` and `overall-peaks.csv` respectively. Also see **Explaination 4**.
- q5.py : Uses `cowin_vaccine_data_districtwise_modified.csv` to find number of people in districts and states vaccinated with first and second doses respectively on a weekly, mothly and overall scale. The files containing the output are: `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`. Also see **Explaination 5**.
- q6.py : Use `census.csv` to get various ratios. The files containing the output are: `district-vaccination-population-ratio.csv`, `state-vaccination-population-ratio.csv`, `overall-vaccination-population-ratio.csv`. Also see **Explaination 6**.
- q7.py : Find vaccine-type ratios. The files containing the output are: `district-vaccine-type-ratio.csv`, `state-vaccine-type-ratio.csv`, `overall-vaccine-type-ratio.csv`. Also see **Explaination 7**.
- q8.py : Find dose ratios. The files containing the output are: `district-vaccinated-dose-ratio.csv`, `state-vaccinated-dose-ratio.csv`, `overall-vaccinated-dose-ratio.csv`. Also see **Explaination 8**.
- q9.py : For every state, find the date on which the entire population will get at least one dose of vaccination. Output file is: `complete-vaccination.csv`. Also see **Explaination 9**.

Detailed Explaination:

Also checkout Output File Notes

1. The new json districts does a deep merge of the districts, i.e., if A, B and C are connected such that, originally A is a neighbour of B and B is a neighbour of C but A and C are not connected directly, now if B is to be removed, A and C will be reported as neighbours in the final answer.
2. DFS is used to generate the edge list. During DFS, parent history is maintained to prevent reporting duplicate edges. Edge list is printed directly (no header line is printed in the csv).

3. Calculates the endpoints of the weeks and months along with the weekid and then determines the cases by taking difference between the cumulative cases reported.

4. Confirmed cases are used as a heuristic to determine how many cases are currently active in the week. Using Active cases (and eliminating deaths and recoveries) would not have given the accurate results anyways as recovery and death times differ significantly. However, confirmed cases reported in a particular time period clearly gives an indication of how the pandemic is spreading. Upon completing the analysis, the results obtained are extremely close to the actual peaks which reinforces the hypothesis that confirmed cases are good heuristic to measure the spread.

5. During analysis, following inconsistencies such as (but not limited to) were taken care of: dip in cumulative values, change in the name of headers. In order to deal with dips in cumulative values, max of the values on the consecutive days is taken (as the dip is local and the values rise on subsequent days).

6. State data from census is not used as it would cause issues for Telangana, instead, CoWin data as used as base. This ensures that districts present in Telangana are automatically counted in Telangana when using census data. This helps in improving results of all questions related to census data.

7. "NA" is reported for those districts/states where 0 doses of Covaxin have been administered.

8. Number of people with atleast 1 dose given to them are equal to the number of dose-1 administered (ratio: vaccinateddose1ratio). Number of people who have been given both the doses are equal to the number of dose-2 administered (ratio: vaccinateddose2ratio). In some places, a ratio greater than 1 is observed. This can be justified by a significant population rise in the district over the past 10 years.

9. Weekly rate has been reported in the output (Number of people vaccinated/day, using the last week's data).

Additional Notes:

Modifications to Dataset(s):

`cowin_vaccine_data_districtwise_modified.csv`

- Original Dataset: `cowin_vaccine_data_districtwise.csv`
- There are 3 pairs of districts with same name, a number was augmented to their name to resolve the conflict.
- There are 2 header rows, which were merged into a single header.
- Exact changes made to districts can be found in Appendix-A

`census.csv`

- Original Dataset: `census.csv`
- Districts in this dataset were compared with those present in CoWin dataset.
- All district pairs with edit distances < 3 were examined.
- All district pairs such that one was "contained" in the other's name were examined.
- Some merges were obvious, for others, it was checked if they belong to the same state.

- There are 3 pairs of districts with same name, a number was augmented to their name to resolve the conflict (Appendix-A).
- Exact changes made can be found in Appendix-B

Other than the aforementioned changes, dataset was also modified according to the instructions mentioned in the "NOTE" of Assignment

Output Files:

- Q3, Q4, Q5, Q6, Q7, Q8 and Q9: Have the first line as header.
- Q3 and Q5: The "timeid" portion of the header is replaced with "weekid" in case of weekly analysis ("monthid" for monthly analysis and "overallid" for overall analysis).
- Q5, Q6, Q7 and Q8: The "districtid" portion in the header is replaced with "stateid" for state-wise analysis ("overallid" for overall analysis).
- "overallid" wherever used, is always reported as 1.
- All values for "overall" time period have been aggregated by adding up the difference of cumulative numbers at the end of each month.

Appendix

Appendix-A: CoWin

Original Name	Name Changed To
BR_Aurangabad	BR_Aurangabad1
MH_Aurangabad	MH_Aurangabad2
UP_Balrampur	UP_Balrampur1
CT_Balrampur	CT_Balrampur2
UP_Pratapgarh	UP_Pratapgarh1
RJ_Pratapgarh	RJ_Pratapgarh2

Appendix-B: Census

This list has most (not all) of the changes made.

Original Name	Name Changed To
Gurgaon	Gurugram
Almora\n	Almora
Sri Potti Setnam Nellore	S.P.S. Nellore
Y.S.R.	Y.S.R. Kadapa

Original Name	Name Changed To
All Delhi Districts (8)	Added Delhi to the end of their name
Muktsar	sri_muktsar_sahib
Tehri garhwal	Tehri
Garhwal	pauri_garhwal
Kaimur (bhabua)	kaimur
Jhunjhunu	jhunjhun
Dibang Valley	Upper Dibang Valley
Maldah	Malda
Khargone (West Nimar)	Khargone
Khandwa (East Nimar)	khandwa
Balgot	Bagalkote
The Nilgiris	Nilgiris
Badgam	Budgam
Baramula	baramulla
Bandipore	bandipora
Shupiyan	shopiyan
Lahul and spiti	lahaul_and_spiti
Bara Banki	barabanki
Puruliya	Purulia
Kodarma	Koderma
Jagatsinghapur	Jagatsinghpur
Jajapur	Jajpur
Anugul	Angul
Baudh	Boudh
Banas Kantha	Banaskantha
Sabar Kantha	Sabarkantha
Ahmadabad	Ahmedabad
Buldana	Buldhana
Gondiya	Gondia
Ahmadnagar	Ahmednagar

Original Name	Name Changed To
Rangareddy	Ranga reddy
Mahbubnagar	Mahabubnagar
Chittaurgarh	Chittorgarh
Allahabad	Prayagraj
NSWE District(s)	NSWE Sikkim
Firozpur	ferozepur
Dhaulpur	Dholpur
Darjiling	Darjeeling
Mahesana	Mehsana
Panch Mahals	Panchmahal
Bellary	Bellari
Tumkur	Tumkuru
Mysore	Mysuru
Kanniyakumari	Kanyakumari
Haora	Howrah
Narsimhapur	Narsinghpur
Chikmagalur	Chikkamagaluru
Koch Bihar	Cooch bihar