# CS685: Data Mining

Assignment 1 (100 marks)

Due on: 13th September, 2021, 11:00pm

Explore the website `https://www.covid19india.org/`. The APIs needed to access the data are available at `https://api.covid19india.org/`. Use csv files since the JSON files are no longer updated.

For data sources, see towards the end of the document.

The district data of India is available as a JSON file from `neighbor-districts.json`. It is in an adjacency list format.

Submit all the necessary components of your data as a single `zip` file named `rollno-assign1.zip` in the `hello.iitk.ac.in` portal within the deadline.

If you do not follow the naming conventions, marks for that question will be automatically 0 (zero).

While the programs/scripts should be named `.sh` files, you can invoke any program from within the shell file. The programs should run in the Linux operating system.

All the output files should be sorted by the first field.

1. (10 marks) Modify the neighbor districts data according to the districts found from the Covid19 portal. A neighbor of a larger district is a combination of all the neighbors of its components. Output the new data as `neighbor-districts-modified.json`. Use the state code and district codes from vaccination data as their ids. Arrange all the district and state keys in alphabetical order. Only include common districts from vaccination data and covid data.

2. (10 marks) Construct an undirected graph of districts out of this new file. In the graph, every district is a node. A district node is connected by an edge to every adjacent district of it, and vice versa.

   Output the graph in an edge list format. If district $i$ has an edge with district $j$, output $i, j$.

   Call this output file `edge-graph.csv` and the script/program to generate this `edge-generator.sh`.

3. (10 marks) For every district $i$, find the number of cases from the Covid-19 portal. Take the time-period of analysis from 15th March, 2020 to 14th August, 2021.

   Output the total number of cases per week for every district in the following manner: $districtid$, $timeid$, $cases$, where $timeid$ is the id of the time (week/month/overall) starting from 1.

   Call this output file `cases-time.csv` and the script/program to generate this `case-generator.sh` where `time` is `week`, `month`, and `overall`.

4. (10 marks) For every district, state and overall, find the week and month having peak (highest) number of active cases for wave-1 and wave-2. The output file contains columns: $districtid$, $wave1 - weekid$, $wave2 - weekid$, $wave1 - monthid$, $wave2 - monthid$.

   Call this output file `peaks.csv` and the script/program to generate this `peaks-generator.sh`.

   A week starts from Sunday and runs till Saturday. The next week starts from Thursday and ends in the next Wednesday. Thus, two consecutive weeks overlap.

A wave starts when cases start rising, and ends when cases flatten out. The peak of a wave is its highest point. Identify the two most important peaks. (Roughly, wave-1 was in the summer of 2020, while wave-2 was in April-May of 2021.)

5. (10 marks) Find the number of people vaccinated with 1 or 2 doses of any vaccine, and sort the output file with district id and state id. Output this for all districts and all states weekly, monthly and overall in the following manner: $districtid$, $timeid$, $dose1$, $dose2$.

   Call this output file `vaccinated-count-time.csv` and the script/program to generate this `vaccinated-count-generator.sh` where time is week, month, and overall.

6. (10 marks) For each state, district and overall, find the following ratios: total number of females vaccinated (either 1 or 2 doses) to total number of males vaccinated (same). For that district/state/country, find the ratio of population of females to males. (If a district is absent in 2011 census, drop it from analysis.) Now find the ratio of the two ratios, i.e., vaccination ratio to population ratio.

   Output them in the following manner: $districtid$, $vaccinationratio$, $populationratio$, $ratioofratios$.

   Call this output file `vaccination-population-ratio.csv` and the script/program to generate this `vaccination-population-ratio-generator.sh`.

   Sort the output by the final ratio.

7. (10 marks) For each state, district and overall, find the following ratios: total number of Covishield vaccinated persons (either 1 or 2 doses) to total number of Covaxin vaccinated persons (same).

   Output them in the following manner: $districtid$, $vaccineratio$.

   Call this output file `vaccine-type-ratio.csv` and the script/program to generate this `vaccine-type-ratio-generator.sh`.

   Sort the output by the ratio.

8. (10 marks) For each state, district and overall, find the following ratio: total number of persons vaccinated (both 1 and 2 doses) to total population. (If a district is absent in 2011 census, drop it from analysis.)

   Output them in the following manner: $districtid$, $vaccinateddose1ratio$, $vaccinateddose2ratio$.

   Call this output file `vaccinated-dose-ratio.csv` and the script/program to generate this `vaccinated-ratio-generator.sh`.

   Sort the output by the dose-1 ratio.

9. (10 marks) For every state, find the date on which the entire population will get at least one does of vaccination. Assume the same rate of vaccination as in the week ending on 14th Aug, 2021. (Do not treat children separately, and assume the same rate of vaccination.)

   Output them in the following manner: $stateid$, $populationleft$, $rateofvaccination$, $date$.

   Call this output file `complete-vaccination.csv` and the script/program to generate this `complete-vaccination-generator.sh`.

10. (10 marks) Write a manual that describes how to use your code. Include all the programs, their plugins, and dependencies needed to run the program. Include a top-level script `assign1.sh` that runs the entire assignment. Call this manual `README.txt`.

**NOTE:**

1) For questions specifying **for every/each district, state and/or overall**, output separate files with respective columnid. ex: Q4 output files district-peaks.csv, state-peaks.csv, overall-peaks.csv

2) Each month starts from 15th and ends at 14th of next month in calendar.

3) FOR QUESTION 1

a) Use state code and district key from vaccination data as identfiers (IDs).

b) Some district names in `neighbor-district.json` have old names; replace them with new district names as per CoWin vaccination data.

c) Some Districts in CoWin data has been repeated, merge them into one while calculating the results for all questions. For example, for Ahmedabad and Ahmedabad corporation, merge them into Ahmedabad.

d) Avoid considering the below districts

FROM COWIN DATA:

Chengalpattu, Gaurela Pendra Marwahi, Nicobars, North and Middle Andaman, Saraikela-Kharsawan, South Andaman, Tenkasi, Tirupathur, Yanam

FROM `neighbor-district.json`, remove all entries of:

Kheri, Konkan division, Niwari, Noklak, Parbhani, Pattanamtitta

4) FOR QUESTION 3:

The cases from district.csv is cumulative , so consider the number of cases in present day = present day cases - previous day cases

5) For QUESTION 4 ONLY :

The week overlap that is Sunday to Saturday is week 1 and Thursday to Wednesday is week2 is *only* for finding peaks. For other questions consider non-overlapping weeks from Sunday to Saturday.

6) LINKS

VACCINE DATA: `http://data.covid19india.org/csv/latest/cowin_vaccine_data_districtwise.`

DISTRICT CASE DATA: `https://data.covid19india.org/csv/latest/districts.csv`

CENSUS DATA: `http://censusindia.gov.in/pca/DDW_PCA0000_2011_Indiastatedist.xlsx`.