CS685: Assignment 1

Neil Rajiv Shirude
Roll no:170429
neilrs@iitk.ac.in

1 Solution to Part 10: Overall Conclusions

1.1 Choice of Covid-19 Data

The folllowing datasets were available at https://api.covid19india.org/-

1. Entire data: `https://api.covid19india.org/v4/data-all.json`
2. Raw data: `https://api.covid19india.org/raw_data1.json`
3. Data in CSV format: `https://api.covid19india.org/csv/latest/districts.csv`

I chose to use entire data (data-all.json), because of the following merits-

- Availability of data over entire period of analysis, unlike the data in CSV format (districts.csv)
- No repetition in the number of cases and no double counting, as in the case of raw data
- Well-documented 'confirmed' number of cases. (On the other hand, raw data has reliable 'hospitalized' data. But according to WHO, the hospitalization rate for COVID-patients is not uniform and is less than 20%.

1.2 Missing Values

1.2.1 Identification of Missing Values

1. For certain days, particularly in the beginning of the analysis period, district-level data was not available. Instead, state-level data was available.
2. Certain states considered certain special COVID-19 cases as districts. These included-
   - State Pool
   - Foreign Evacuees
   - Airport Quarantine
   - Railway Quarantine
   - BSF Camp
   - Italians
   - Other Districts
   - Others, etc.

1.2.2 Handling of Missing Values

1. For a certain day, if district-level data is not available and only state-level data is available, I mapped the cases for a state to the district in the state with the highest population.
2. Over the period of analysis, there were approximately 38 lakh confirmed COVID-19 cases in India. The sum of the number of cases mapped to these absurd districts was less than 30,000 cases. As these absurd cases account for less than 1% of the total casesover the analysis period, I ignored these cases (after consulting with sir during a discussion hour).

1.3 Overall Conclusions

1.3.1 Overall Results

Over the entire period of analysis, the following districts were identified as the top-5 hotspots-

- Neighborhood analysis hotspots-
    1. Bengaluru urban
    2. Bhopal
    3. Delhi
    4. Surat
    5. Ahmedabad
- Neighborhood analysis coldspots-
    1. Upper Dibang Valley
    2. Krishna
    3. South Tripura
    4. Kalimpong
    5. Unokoti
- State analysis hotspots-
    1. Puducherry
    2. Bengaluru urban
    3. Chennai
    4. Raipur
    5. Patna
- State analysis coldspots-
    1. Diu
    2. Krishna
    3. Lahaul and Spiti
    4. Vizianagaram
    5. Wayanad

1.3.2 Interpretation of Overall Results

According to this analysis, the districts with very high cases in comparison to their neighboring districts/ other districts in their state are identified as hotspot districts. If the spread of COVID-19 cases in these hotspot districts can be contained, the impact would be very high, since the disease has not spread in the vicinity of these districts.

Districts such as Pune and Mumbai that have reported the highest number of cases in India cannot be predominantly seen in these lists since their neighbors/ other districts in their state have also witnessed large scale spread of COVID-19.

Similarly, the districts having very low cases in comparison to their neighboring districts/ other districts in their state are identified as coldspot districts. These districts have done a better job at containing the spread of COVID-19 as compared to their neighbors.