CSE506 Intro. to Data Mining

Assignment 1

Instructions

- The assignment is to be attempted in pairs.
- Language allowed: Python
- For Plagiarism, institute policy will be followed
- You need to submit the README.pdf, Code file (Code template attached).
- You are allowed to use libraries such as pandas, matplotlib, etc
- Mention methodology, assumptions, and results you may have in README.pdf.
- Submit code and readme in ZIP format with the following name: CSE506_A1_<roll_no>.zip

Dataset: https://api.covid19india.org/states_daily.json Download and save it as a JSON file.

The following link contains live covid19 data arranged in date wise and state wise fashion. Variable Status dictates whether the JSON array has data of “Confirmed”, “Recovered” or “Deceased”.

How to attempt:

- You have to use the given python code template provided.
- For Q1 parts 1-7 and Q2 part 1-3, Your code should be flexible to take any start date and end date (end date > start date). How to define functions and how to run is given in the template itself, Please refer to the python code template.
- You will report numbers and plots in the README.pdf file for the dates mentioned against every question.
- Date format: format='%Y%m%d' 2020-03-14
- For Q2, All 3 plots for 1 question should be in a single plot. refer to tutorial 1.
- For Q3, Report your intercept and slope coefficients in the README.pdf file. Using libraries such as sklearn, TensorFlow is not allowed.

Q1. Data Manipulation

1. Count the total number of “Confirmed”, “Recovered” and “Deceased” from 14-Mar-2020 to 05-Sept-2020 and report the numbers.
2. Count the total number of “Confirmed”, “Recovered” and “Deceased” from 14-Mar-2020 to 05-Sept-2020 for state Delhi (dl)
3. Report total count of “Confirmed”, “Recovered” and “Deceased” count from states Delhi + Maharasthra (Sum of both states count) from 14-Mar-2020 to 05-Sept-2020.
4. Report the highest affected state in terms of “Confirmed”, “Recovered” and “Deceased” with the count till 05-Sept-2020 from 14-Mar-2020.

5. Report the lowest affected state in terms of “Confirmed”, “Recovered” and “Deceased” with the count till 05-Sept-2020 from 14-Mar-2020.
6. Find the day and count with the highest spike in a day in the number of cases for the state Delhi for “Confirmed”, “Recovered” and “Deceased” between dates 14-Mar-2020 and 05-Sept-2020.
7. Report active cases (Assume active = Confirmed - (Recovered + Deceased)) state wise for all states separately on date 05-Sept-2020 (This date only) starting from 14-March-2020.

Q2. Plotting

1. Plot the area trend line for total “Confirmed”, “Recovered” and “Deceased” cases from 14-Mar-2020 to 05-Sept-2010.
2. Plot the area trend line for total “Confirmed”, “Recovered” and “Deceased” cases for the state Delhi (dl) from 14-Mar-2020 to 05-Sept-2020.
3. Plot the area trend line for active cases. Assume active = Confirmed - (Recovered + Deceased) from 14-Mar-2020 to 05-Sept-2020.

Q3. Linear Regression

1. Implement a linear regression on the state Delhi data over dates, separately for “Confirmed”, “Recovered” or “Deceased” and report intercept and slope coefficients for all 3 cases from 14-Mar-2020 to 05-Sept-2020.