Data Mining - Assignment 1

File structure

1. The assignment is made using *python 3*.
2. Assignment folder contains
    A. Python code for all quetions.
    B. Shell .sh file to run respective python code.
    C. 'assign1.sh' file to run shell script of all quetions.
    D. Data folder - stores all data.
3. Data folder contains
    A. All initial data needed.
    B. All output file genrated by program named as per naming conventions.
    C. Some other data files genrated by program that will be used by further questions

Library needed

Library needed to run assignment are.

1. numpy
2. pandasql
3. pandas
4. pickle
5. datetime
6. re
7. collections
8. json
9. csv
10. matplotlib.pyplot
11. scipy.signal

How to run code

'assign1.sh' is top-level script that runs the entire assignment. It run all the quetion in sequence.

It can be run by

```
bash assign1.sh
```

Other Details

Question 1

1. python file is named as `'neighbor_districts_modifier.py'`
2. shell file is named as `'neighbor_districts_modifier.sh'`
3. Apart from output file it genrate two files 'vaccineData-modified.csv' and 'districts-modified.csv' which are dependency for other questions.

Question 2

1. python file is named as `'edge_generator.py'`

2. shell file is named as `'edge_generator.sh'`

Question 3

1. python file is named as `'case_generator.py'`
2. shell file is named as `'case_generator.sh'`
3. Output file have been named as `'cases_week.csv'` `'cases_month.csv'` `'cases_overall.csv'`
4. Apart from output file it genrate four files 'districts.npz' and 'distToState.npz' 'distToDistKey.pkl' and 'districts-modified-v2.csv' which are dependency for other questions.

Question 4

1. python file is named as `'peaks_generator.py'`
2. shell file is named as `'peaks_generator.sh'`
3. Output file is named as `'district-peaks.csv'` `'state-peaks.csv'` `'overall-peaks.csv'`

Question 5

1. python file is named as `'vaccinated_count_generator.py'`
2. shell file is named as `'vaccinated_count_generator.sh'`
3. Output file is named as `'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'`
4. Apart from output file it genrate 'vaccineData-modified-v2.csv' which are dependency for other questions.

Question 6

1. python file is named as `'vaccination_population_ratio_generator.py'`
2. shell file is named as `'vaccination_population_ratio_generator.sh'`
3. Output file is named as `'district_vaccination_population_ratio.csv'` `'state_vaccination_population_ratio.csv'` `'overall_vaccination_population_ratio.csv'`

Question 7

1. python file is named as `'vaccine_type_ratio_generator.py'`
2. shell file is named as `'vaccine_type_ratio_generator.sh'`
3. Output file is named as `'district_vaccine_type_ratio.csv'` `'state_vaccine_type_ratio.csv'` `'overall_vaccine_type_ratio.csv'`
4. Ratio where Covaxin is zero is written NA

Question 8

1. python file is named as `'vaccinated_ratio_generator.py'`
2. shell file is named as `'vaccinated_ratio_generator.sh'`
3. Output file is named as `'district_vaccinated_dose_ratio.csv'` `'state_vaccinated_dose_ratio.csv'` `'overall_vaccinated_dose_ratio.csv'`

Question 9

1. python file is named as `'complete_vaccination_generator.py'`
2. shell file is named as `'complete_vaccination_generator.sh'`

3. Output file is named as `'complete_vaccination.csv'`