Assignment01: Data Analytics on a large simulated dataset for COVID-19

Due Date: 25 November 2020

- Assume that we have a government agency that interacts with multiple network and service providers (like Google and Apple) along with healthcare providers to generate a user mobility and COVID-19 infectiousness dataset.
- The dataset has the locations (GPS and site) of every individual along with their infection status in an urban environment (e.g. a city) during a time of a day on a daily basis
- **Your job is to analyse the dataset and perform two tasks**
    - **1) Find the answers to a set of analytical questions related to the dataset. Specify how you obtained those answers as well.**
    - **2) Visualize the required information regarding the dataset**

Task 1: Data Analytics Questions:

1. How many active cases (infected individuals) were there on 2020-03-01 (01 March 2020)?
2. How many people died on 2020-03-01 (01 March 2020)?
3. What is the total number of infections, recoveries and deaths?
4. How many unique individuals in the dataset?
5. How many unique locations and their names in the dataset?
6. Which school is the closest to a residence ID = 3115? Hint: Use Euclidean distance using the lat and long.
7. What percentage of individuals aged b/w 30 and 60 visit schools at 08:00 am?
8. What is the most dominant daily path of individual (ID= 37415) ? Hint: There are 5 timestamps per day (00:00,08:00,16:00,19:00,22:00). An individual can be at his residence (`currentLocationType`) at given timestamp, then can move to school on the next time stamp and so on. COMPLEX QUESTION! Can skip it if too difficult.
9. What is the proximity (number of individuals sharing the same location) of an individual ID (37415) at a given date and time?
10. What is the age and gender of individual 37415? Did this individual get infected? If yes, how long did the infection last? And did the individual die?
11. What is the infection and mortality rate of health care workers?
12. Does the mortality rate change with age and gender?

Task 2: Visualizations:

1. Plot the total number of infections, recoveries and deaths w.r.t date.
2. Plot a pie chart for the ratio of asymptomatic and symptomatic infections.
3. Plot the variation of proximity of individual (ID=37415) w.r.t date.
4. Plot bar graphs to show the mortality rates and infection rates per age groups (create age groups 0-10, 10-20, and so on).

5. Visualize the number of individuals at a location on a geographical map. (COMPLEX: Can be done with Tableau, Plotly)

You can submit the pdf version of the jupyter notebook (with all answers documented), if the work is done in Python. Otherwise, submit a manually prepared a word/pdf and submit the document with the description of the steps taken as well as code to come up with the answers.

The dataset can be downloaded from below mentioned link:
https://nustedupk0.sharepoint.com/sites/BigDataAnalyticsMSCS-2k19MSDS-2k19/Class%20Materials/Assignment01_Big_Data_Analytics.zip

Data Attributes:

Attribute	Description
Date_Time	The date and time
id	Unique User ID
age	The age of the individual
gender	The gender of the individual. 0 for female, 1 for male
infection_status	susceptible, infected, recovered, deceased
days_since_infection	days since the person got infected
date_of_infection	the date of infection
date_of_symptoms	the date on which she/he started showing covid symptoms
date_of_recovery	The date of recovery
date_of_death	The date of death
date_hospital_check_in	The date of hospital check in
date_hospital_check_out	The data of hospital check out
residence_id	The individual's residence ID
school_id	The individual's school ID (if she/he goes to a school)
workplace_id	The individual's workplace ID (if she/he works)
currentLocationID	The ID of the current location of the individual (the ID is a unique integer that identifies a unique location)
currentLocationType	The type of the current location of the individual (e.g. residence, school, etc)
lat	The latitude of the location
lon	The longitude of the location
zone_id	The zone of the location the city
is_notified_to_isolate	Is the person notified to isolate. 0 for no, 1 for yes
is_symptomatic	Is the person showing COVID-19 symptoms. 0 for no, 1 for yes
incubation_period	a static value for a given individual
days_since_isolation	Days since the person is in quarantine (self-isolation or hospitalization)
isolation_times	Not used
days_since_symptomatic	The days since he became symptomatic
in_hospital	Is hospitalized or not. 0 for no, 1 for yes
wears_mask	0 for no, 1 for yes
type_of_job	The type of the job of the person

morbidity	The comorbidity, if any
exposure_time	Not used
infected_from_user_id	If the person is infected, this tells the ID of the individual from whom she/he got infected
transmitted_count	Not used
asymptomatic_transmission	1 for yes
locationID_of_infection	The location ID where the individual got infected
income	The income of the individual
household_size	The household size of the individual
bmi	The Body Mass Index of the individual
date_of_test	The date of COVID-19 test
test_result	The test result. 0 for negative, 1 for positive