Project Report

IC252: Data Science-II

Submitted by

Name: Ayush Gaurav
Enroll. No.: B21183

Course Instructor

Dr. Satyajit Thakor

INDIAN INSTITUTE OF TECHNOLOGY MANDI

Date: 29th April 2022

Contents

1 Objective

1) Given the "*cases.csv*" dataset, plot the time graph of the I*nfected Fraction* of population (*Infected Fraction* vs *Time in Months*). Do this for Delhi, Mumbai and Kolkata.

a) Compare these graphs. What do you infer from these graphs?

b) Calculate the variance of the *Infected Fraction*.

Hint:-

$$Infected\ Fraction = \frac{|Confirmed - Recovered - Deceased|}{Population}$$

$$Susceptible = \frac{|Population - Confirmed|}{Population}$$

$$Removed = \frac{|Recovered + Deceased|}{Population}$$

Extra work:- Plot the time graph of S*usceptible* and R*emoved* population and compare with the I*nfected Fraction* of the population. Plot all these in a single plane (graph). Do this for Delhi and Mumbai.

2) Given the "*2021_IN_Region_Mobility_Report.csv*" dataset for 2021, plot the following

a) *Retail mobility* of Delhi and Mumbai. Compare them in the same plane.

b) *Transit mobility* of Delhi and Mumbai. Compare them in the same plane.

c) What do you infer from these graphs?

d) Calculate the IQR in each case (Interquartile range).

e) What is the expected value of *Retail* and *Transit mobility* in Delhi and Mumbai?

Hint:-

In given dataset, "retail_and_recreation_percent_change_from_baseline" column represents *Retail mobility* and "transit_stations_percent_change_from_baseline" column represents *Transit mobility*.

Note:-

In descriptive statistics, the interquartile range (IQR) is a measure of statistical dispersion. It is the spread of the data or observations. The IQR may also be called the midspread, middle 50%, or Hspread. It is defined as the spread difference between the 75^{th} and 25^{th} percentiles of the data. The lower quartile corresponds with the 25^{th} percentile and the upper quartile corresponds with the 75th percentile. So, $IQR = Q3 - Q1$. **First, take the median of the data. Then Q3 (median of the lower half of the data) – Q1 (median of the upper half of the data)**

3) Given the "*Cowin_Vaccine_Data_Districtwise.csv*" dataset, do the following.

a) Plot the vaccination coverage of Delhi and Mumbai. (Basically, for each city you have to plot % of people vaccinated with first dose and % of people vaccinated with second dose in the same plane)

b) Calculate the correlation of first dose coverage with the following:

i. $\frac{Sites}{Area\ of\ city}$

ii. $\frac{Sessions}{Area\ of\ city}$

What can you infer from these correlations?

c) Find the state/ UT with the highest vaccination coverage (first dose).

Note:-

- Use the following values of population and area of the cities.
  (Population of Delhi: 20,591,874, Population of Mumbai: 20,667,656, Population of Kolkata: 14,850,000, Area of Delhi : 1400 sq. km , Area of Mumbai: 670 sq.km, Area of Kolkata: 206 sq.km)
- Empty (None) value should be ignored for any dataset.

2 Theoretical Solution

2.1 Question1:

a) Infected Fraction, Susceptible and Removed are calculated as below:

Infected Fraction $= \frac{Confirmed - Recovered - Deceased}{Population}$

Susceptible $= \frac{Population - Confirmed}{Population}$

Removed $= \frac{Removed + Deceased}{Population}$

b) Variance of random variable X is calculated as:

$$\widehat{Var}(\text{X}) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{E}(X))^2$$

2.2 Question2:

a) In the dataframe, find the average of the retail mobility of different regions of Delhi and Mumbai for each date and then, plot it using matplotlib.

b) In the dataframe, find the average of the transit mobility of different regions of Delhi and Mumbai for each date and then, plot it using matplotlib.

c) No theoretical solution for this part.

d) Interquartile range is defined as the spread difference between the 75^{th} and 25^{th} percentiles of the data. The lower quartile corresponds with the 25^{th} percentile and the upper quartile corresponds with the 75^{th} percentile.

So, IQR = $Q_3 - Q_1$

e) Expected value of a random variable is given as:

$$\hat{E}(X) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

2.3 Question3:

a) Vaccination coverage is given as below

$$\text{Vaccination Coverage} = \frac{No.of\ People\ Vaccinated}{Total\ Population} \times 100$$

b) Correlation coefficient between two random variables is given as

$$\text{Correlation coefficient} = \frac{\widehat{Cov}(X,Y)}{\widehat{SD}(\mathrm{X}) \times \widehat{SD}(\mathrm{Y})}$$

where $\widehat{Cov}(X,Y)$: Covariance of random variables X and Y

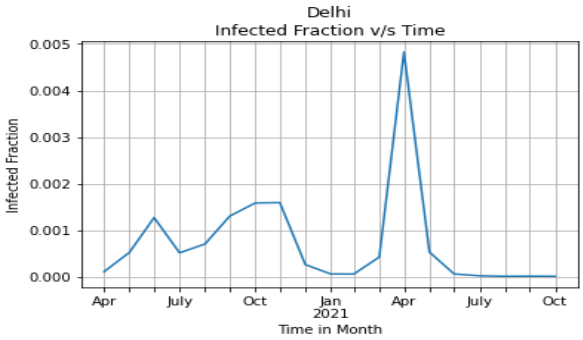
$\widehat{SD}(\mathrm{X})$: Standard deviation of random variable X

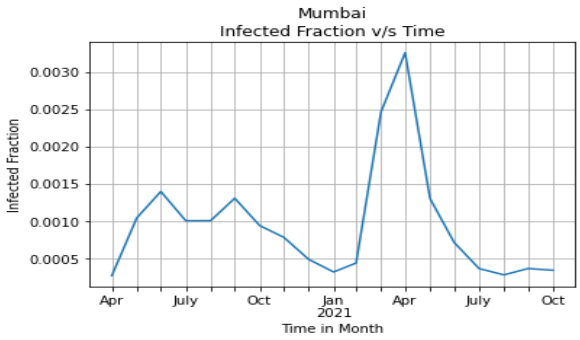
$\widehat{SD}(\mathrm{Y})$: Standard deviation of random variable Y

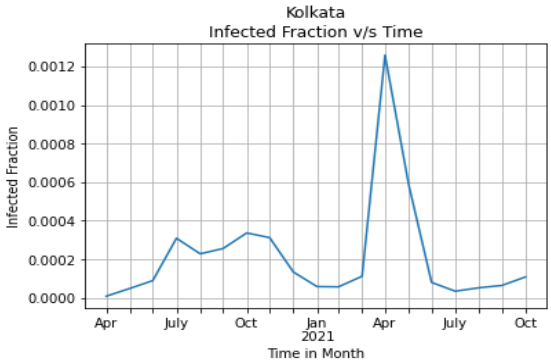
c) No theoretical solution.

3 Simulation/Programming Results and Analysis

3.1 Question1:



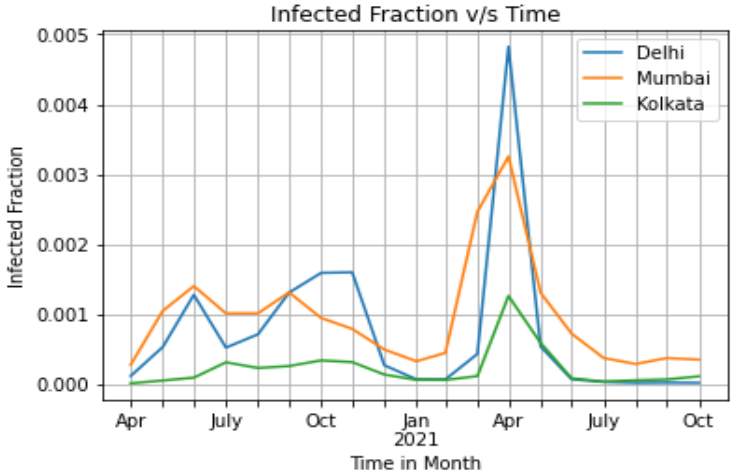




Observation: Time graphs showing infected fraction for Delhi, Mumbai and Kolkata are obtained. The infected fraction plot has three peaks for Mumbai and Kolkata while for Delhi, it has four peaks.

Analysis: Mumbai and Kolkata witnessed three waves of the disease whereas Delhi witnessed four waves of the disease i.e. the infected fraction was very high four times during the given duration for Delhi.

a)



Observation: Time plot showing the infected fraction for Delhi, Mumbai and Kolkata on the same plane is obtained. All three cities have peaks around June 2020, October 2020 and April

2021.Also, April 2021 has the largest peak. In addition to above, Delhi has an additional peak around November 2020.

Analysis: All the three cities experienced three waves of the disease in June 2020, October 2020 and April 2021. Also, the three cities experienced the largest wave in April 2021. In addition to above three, Delhi experienced another wave in November 2020.

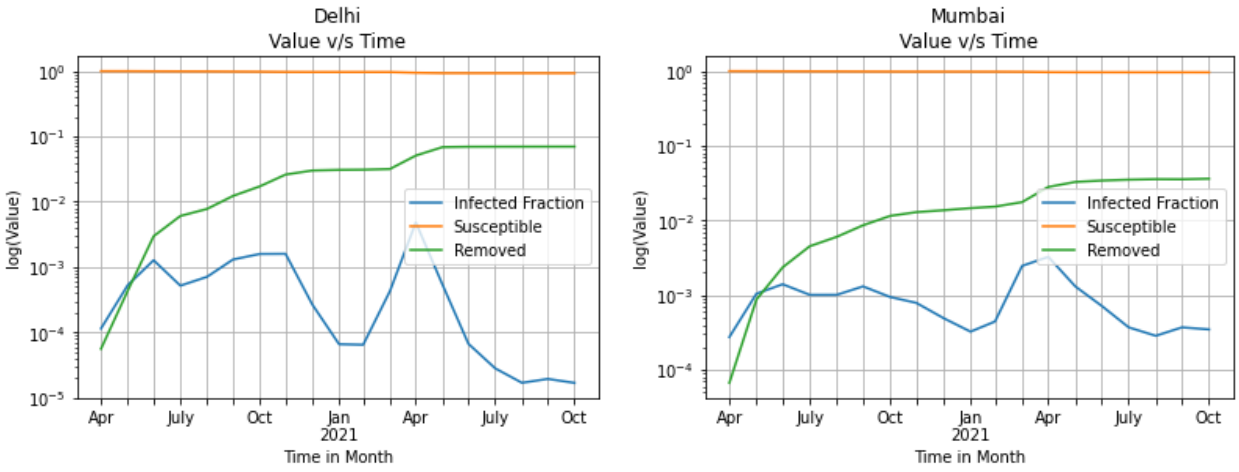
b)

```
Part B
Variance of infected fraction in Delhi is 1.0610918189663683e-06.
Variance of infected fraction in Mumbai is 7.200313627927208e-07.
Variance of infected fraction in Kolkata is 8.40991808053116e-08.
```

Observation: The variance of the infected fraction for the three cities is displayed.

Analysis: Delhi has the highest variance in the infected fraction; Mumbai follows it whereas Kolkata has the least variance in the infected fraction. This means that Kolkata has the least variance in the level of infected fraction whereas Delhi has the highest.

Extra work:



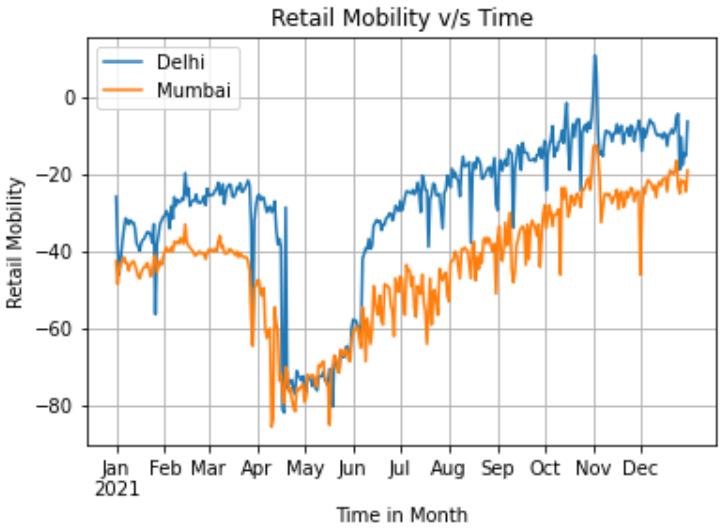
Observation: Time plot showing the infected fraction, susceptible population and removed population on the same plane for the two cities is obtained. We observe that as the time in month increases, infected fractions increases as well as decreases, but the removed population continuously increases whereas susceptible population remains almost constant in both the cities i.e. Delhi and Mumbai. However, the variation in these values is more in case of Delhi as compared to Mumbai.

Analysis: The above observations tell that infected fraction depends on the time of the year and other factors but the susceptible population almost the same whereas the removed population almost always increases. It also tells us that the level of infection in Delhi is more variable as compared to Mumbai.

3 Simulation/Programming Results and Analysis

3.2 Question2:

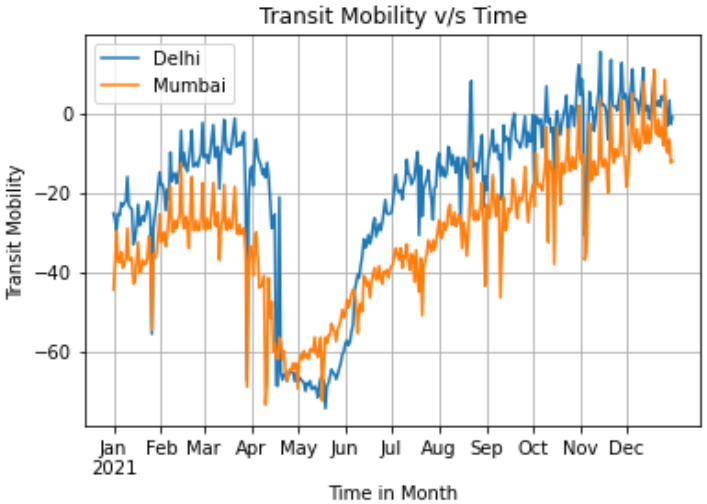
a)



Observation: Plot showing the retail mobility of Delhi and Mumbai with respect to time is obtained. Retail mobility of Delhi is higher than that of Mumbai most of the time. There is a sharp drop in the retail mobility in both the cities from mid-April to June.

Analysis: The sharp drop in retail mobility in both the cities from mid-April to June may be due to high infection rate and lockdown thus, limiting the communication between customers and retailers. The plot tells that the retail mobility of Delhi is higher than Mumbai for most of the times.

b)



Observation: Plot showing the transit mobility of Delhi and Mumbai with respect to time is obtained. Transit mobility of Delhi is higher than that of Mumbai most of the time. There is a sharp drop in the transit mobility for both the cities from mid-April to June.

Analysis: The sharp drop in transit mobility fot both the cities from mid-April to June may be due to high infection rate and lockdown thus, limiting the use of transportation. The plot tells that the transit mobility of Delhi is higher than Mumbai for most of the times.

c) We infer from the graph of retail mobility that the retail mobility of Delhi is higher than that of Mumbai most of the time and there is a sharp drop in the retail mobility for both the cities from mid-April to June.

Also, from the graph of transit mobility, we infer that the transit mobility of Delhi is higher than that of Mumbai most of the time and there is a sharp drop in the transit mobility for both the cities from mid-April to June.

The sharp drop in both the cases may be due to high infected fraction and lockdown between these months.

d)

```
Part D
IQR of Retail mobility in Delhi is 18.36.
IQR of Transit mobility in Delhi is 21.55.
IQR of Retail mobility in Mumbai is 19.0.
IQR of Transit mobility in Mumbai is 22.0.
```

Observation:. Interquartile range of transit mobility is larger than retail mobility for both the cities. Mumbai has larger interquartile range of transit or retail mobility as compared to Delhi.

Analysis: The transit and retail mobilities are more dispersed in case of Mumbai as compared to Delhi. Also, transit mobility has larger dispersion or spread as compared to retail mobility for both the cities.

e)

```
Part E
Expected value of Retail mobility in Delhi is -28.0.
Expected value of Transit mobility in Delhi is -18.78.
Expected value of Retail mobility in Mumbai is -43.69.
Expected value of Transit mobility in Mumbai is -30.89.
```

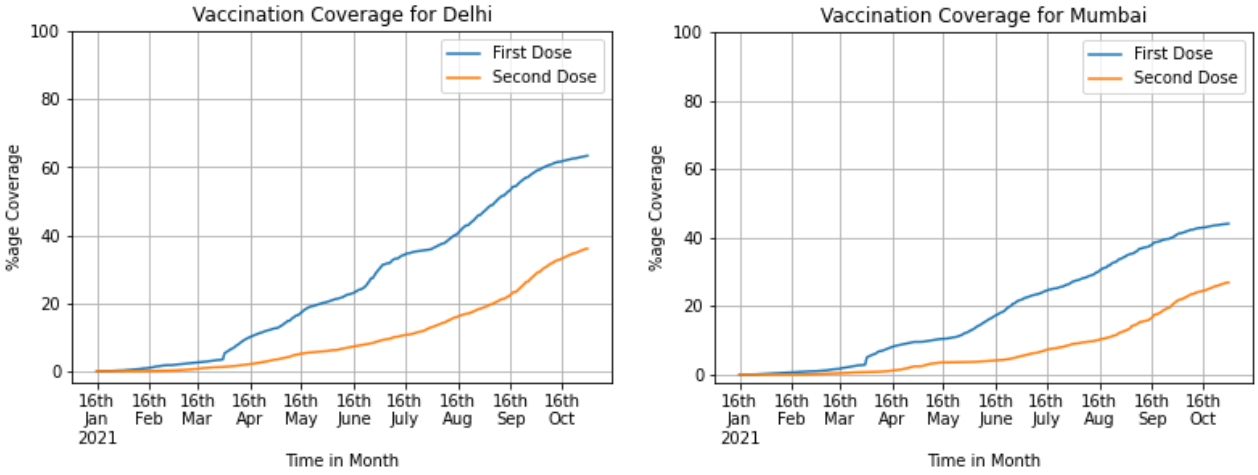
Observation: The expected drop in retail mobility is larger than transit mobility for both the cities. Also, the expected drop in transit mobility or retail mobility is larger for Mumbai as compared to Delhi.

Analysis: Retail mobility drops more than transit mobility from the base line in both the cities. Also, in Mumbai, the drop in transit mobility or retail mobility from the base line is more as compared to Delhi.

3 Simulation/Programming Results and Analysis

3.3 Question3:

a)



Observation: Plot showing vaccination coverage for Delhi and Mumbai is obtained. On any particular date, the value for Mumbai on the y-axis is less than the value for Delhi. Also, on any particular date, the value for First dose is larger than the value for Second dose.

Analysis: The vaccination coverage for Delhi is more than Mumbai at any point of time. Also, the vaccination coverage of the first dose is larger than the second dose as expected for both the cities. Initially, the rate of vaccination is low and it increases with time for both the cities.

b)

```
Part B
The correlation of First Dose Coverage with sites per unit area for Delhi is 0.6433.
The correlation of First Dose Coverage with sessions per unit area for Delhi is 0.8332.
The correlation of First Dose Coverage with sites per unit area for Mumbai is 0.519.
The correlation of First Dose Coverage with sessions per unit area for Mumbai is 0.8091.
```

Observation: For both the cities, sessions per unit area is more strongly related to first dose coverage as compared to sites per unit area. In case of Delhi, correlation of sites per unit area and sessions per unit area with first dose coverage is larger than that in case of Mumbai.

Analysis: To increase the vaccination coverage in any city, increasing the number of sessions per unit area is better option as compared to number of sites per unit area as correlation of first dose coverage is larger with sessions per unit area as compared to sites per unit area.

Also, if sessions per unit area are increased then, in case of Delhi, first dose coverage will increase more in Delhi as compared to Mumbai as correlation is larger in case of Delhi.

c)

```
Part C
The state/UT having highest vaccination is Uttar Pradesh.
```

Observation: The simulation displays the state/UT having the highest vaccination of first dose as Uttar Pradesh.

Analysis: Uttar Pradesh is the state/UT having largest vaccination of first dose.

4 Conclusion

The problems in this assignment and their corresponding solutions above bring us to the conclusion that computer simulation and data science can be used to study the spread of a disease and it's correlation with other factors. Data science can be used to study as well as predict the behaviour of the spread of a certain disease approximately. It can also be used to track the progress of the COVID-19 vaccination program and it tells us how we can make our COVID-19 vaccination programme more effective by giving us useful insights about the correlation of vaccination coverage with various parameters.

----------THANK YOU----------