Name: Ishita More
UID: 2019130039
TE Comps
Subject: Data Analytics

# ISE-Part1

## Aim:

Exploratory Data Analysis

## Objective:

Data exploration (aka exploratory data analysis, or EDA) and the display is a fundamental process of data analysis. EDA is used to summarize your data, visualize patterns in your data, and refine your hypotheses, while data display presents those patterns to others. While EDA is often done ‘on the fly’, and with low-resolution graphics or print-outs, data display is 'presentation-quality' graphics, analogous to what you'd read in a standard scientific presentation, and is what should end up in your thesis or independent project.

## Theory:

Exploratory data analysis (EDA):

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate.

Objectives of EDA:

- Maximize insight into the database/understand the database structure.
- Visualize potential relationships (direction and magnitude) between exposure and outcome variables.
- Detect outliers and anomalies (values that are significantly different from the other observations).
- Develop parsimonious models (a predictive or explanatory model that performs with as few exposure variables as possible) or preliminary selection of appropriate models.
- Extract and create clinically relevant variables.

## Dataset:

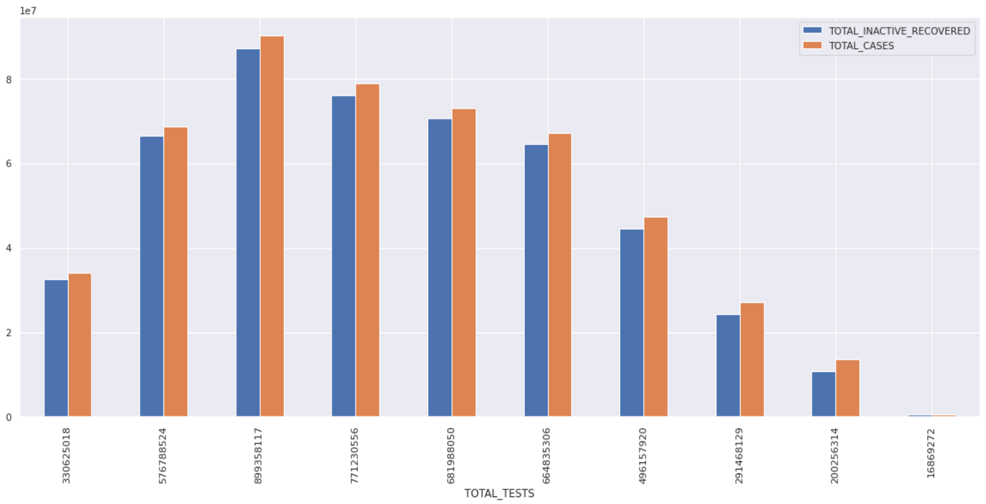
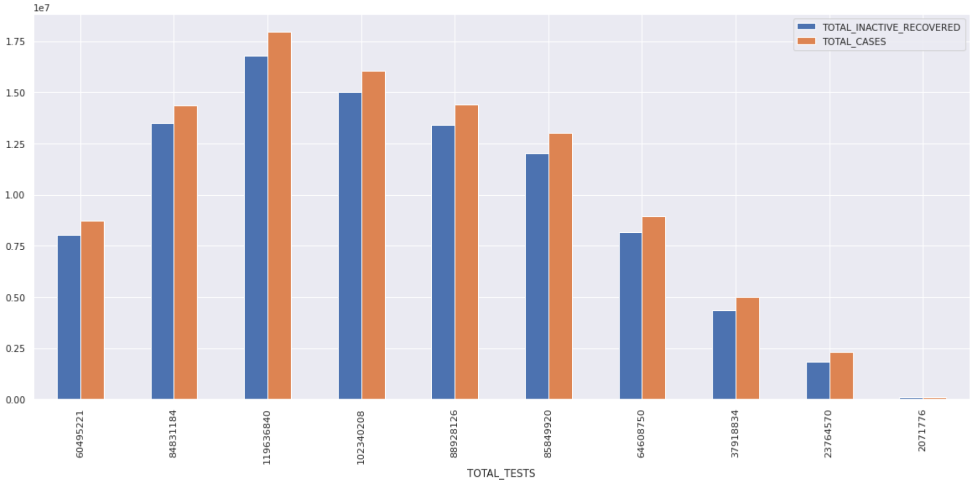
**Daily Age Group Outcomes-**

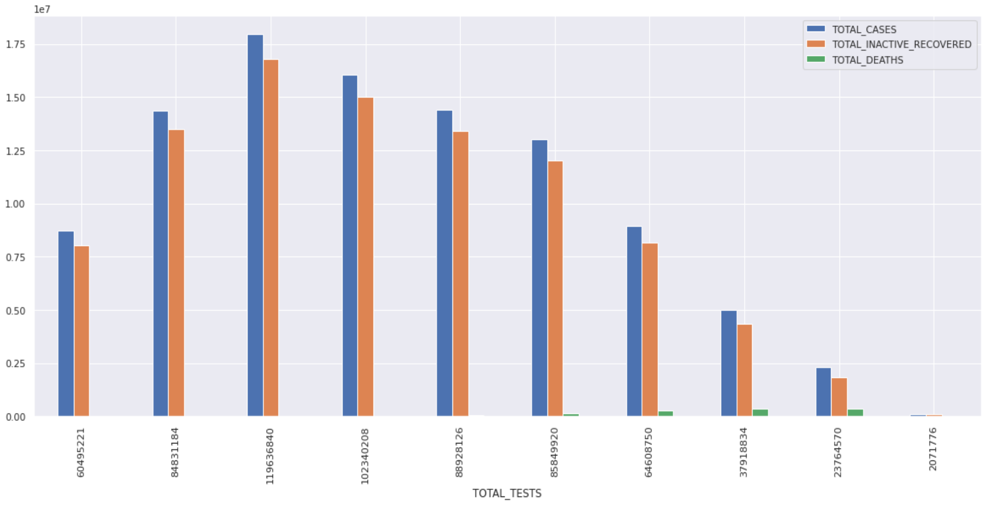
https://www.tn.gov/content/dam/tn/health/documents/cedep/novel-coronavirus/datasets/Public-Dataset-Daily-Age-Group-Outcomes.XLSX

Daily case outcomes by age group (2/25/2022 - 4/9/2020)

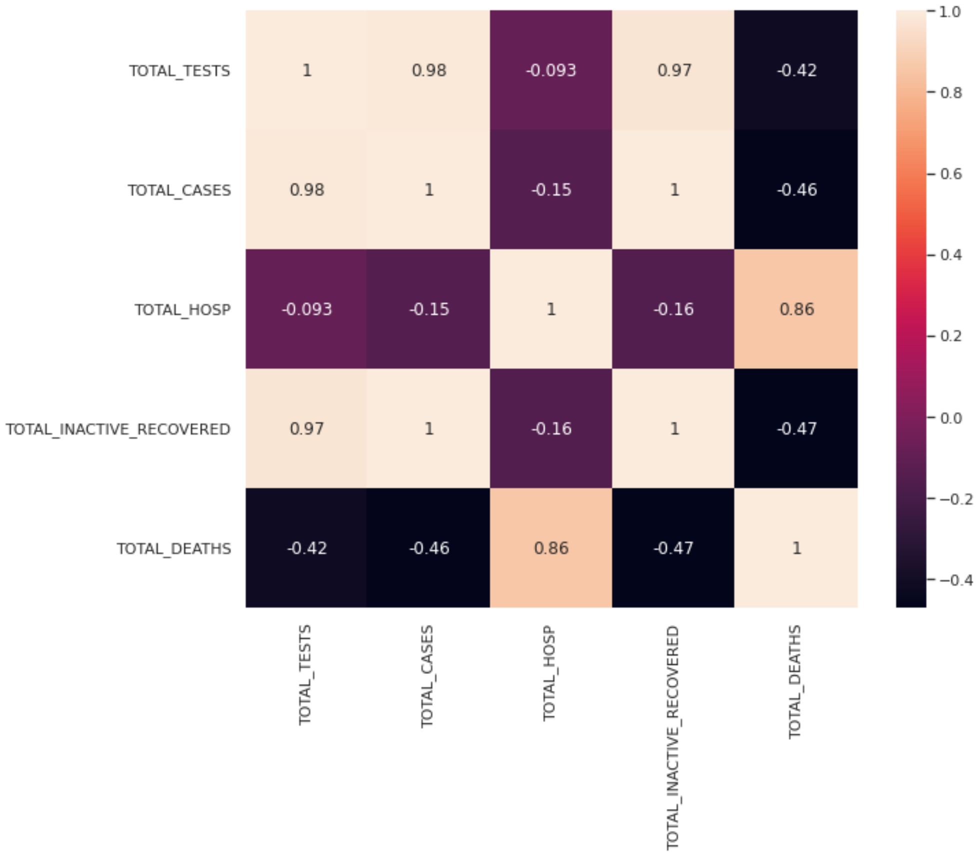
In this Dataset, there is data of total tests done, total cases found, totally recovered, total hospitals, and total deaths on a particular day.

It contains the data from 2/26/2022 to 4/9/2020.

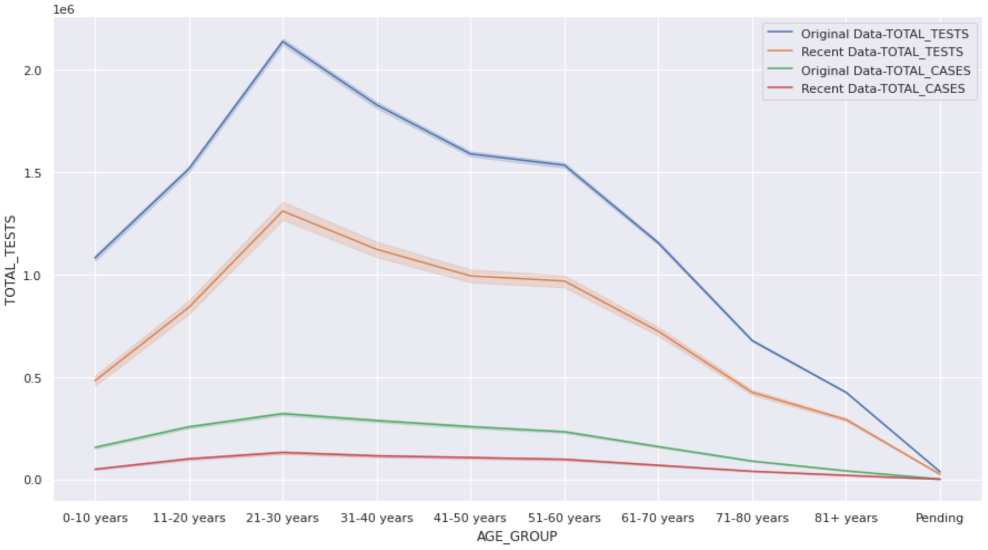
| Graphs | Inference |
|---|---|
|   | Bar graphs are ideal for comparing two or more values, or values over time.<br>Fig 1: Original Data<br>Fig2: Recent Data<br>Fig 1 and 2 is the comparison between recovered and total cases found. Here we can observe that the difference is very low. This implies cases were found daily, but the recovery rate is also good. |



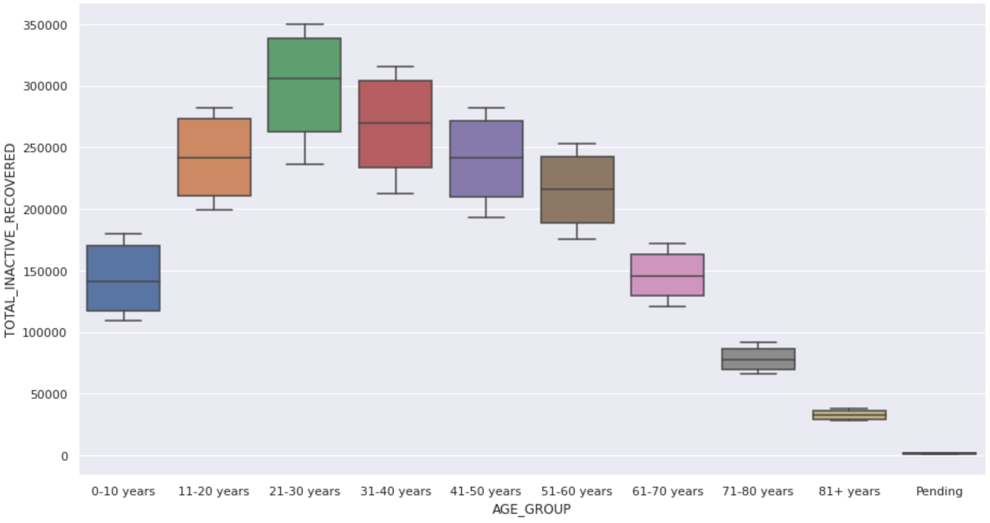
Comparison between Total Cases, Recovery, and Deaths on the basis of a particular age group.
Total Cases found and recovery rate is high for the age group of 20-30 but the death rate is negligible.
Whereas it is the opposite in the case of 70+ age.



As the value of correlation, closer is to 1, for (cases, tests), (recovered, tests), (deaths, hosp) i.e 0.98,0.97,0.86; which implies stronger this relationship is.
A correlation closer to -1 is similar, but instead of both increasing one variable will decrease as the other increases.

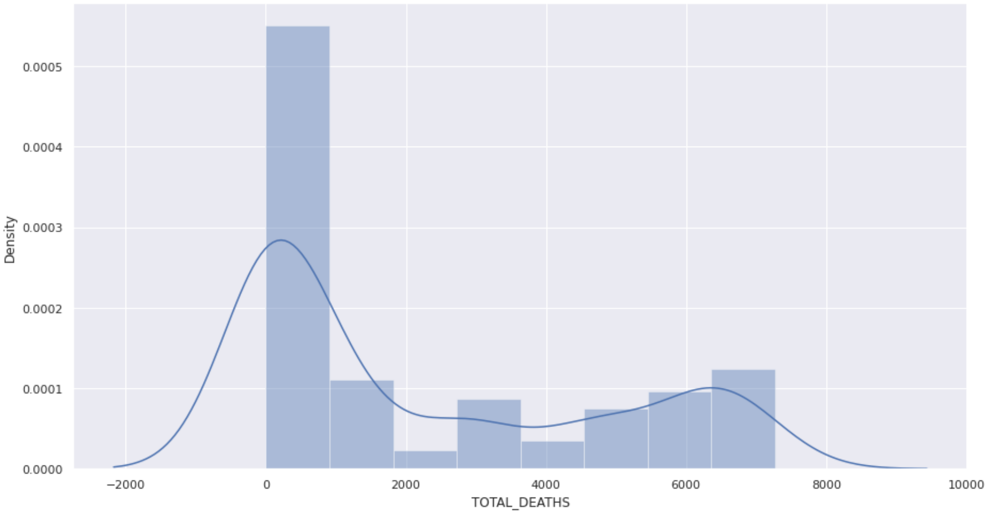


The number of Total tests done has been reduced than before for all age groups. Similarly, the number of Total cases found has been reduced than before for all age groups except the 80+ age group.

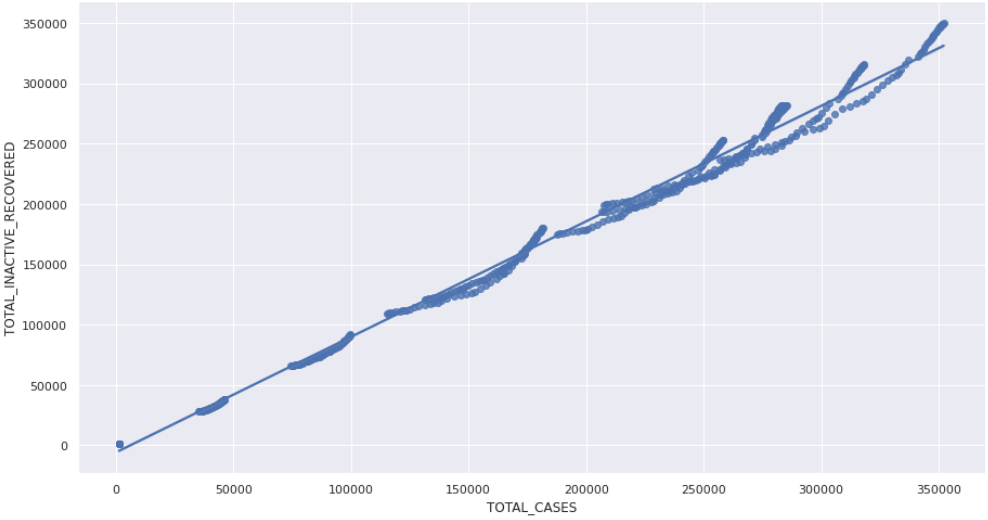
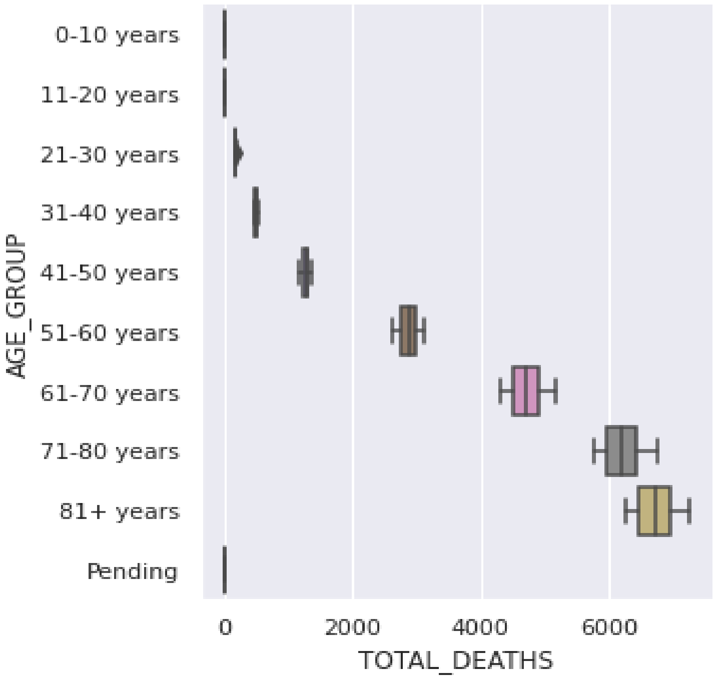


A boxplot is a standardized way of displaying the distribution of data based on a five-number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum").

The recovery rate is high in the case of the 20-30 age group, whereas, it is low in the case of the 60+ age group.



As in this Dist plot, the bar is high at 0-1000, which implies that the number of deaths occurring recently is less i.e. between 0 to 1000.

| | |
|---|---|
|  | As in the Scatter plot, the total cases found and recovery rate is almost a straight line. So, we can say that the total cases found and recovery rate goes hand in hand. |
|  | As in this Cat plot, shows frequencies is high for the 71+ age group, which implies that the number of deaths is high in this age group. |

## Inference:

The Measures of location are the mean, median, and mode.

- Mean is the average of the values. Here, the result tells that the total test that takes place for covid recently is more than before, which implies people are really more aware than in past. Also, as the rate of covid patients is decreased now, the total cases and death also decrease. And the recovery rate is increased.
- Median gives the middle number in an ordered data set.
- The mode gives the most frequent value.

The Measures of spread are the standard deviation, standard error, variance, percentiles, range, and coefficient of variation

- A standard deviation (or σ) is a measure of how dispersed the data is in relation to the mean. Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out. Here, the SD is high which indicates data are more spread out.
- The standard error tells you how much the sample means would vary if you were to repeat a study using new samples from within a single population. Here, Total tests have a high standard error which shows that sample means are widely spread around the population mean whereas total cases, inactive recovered and deaths have a low standard error which shows that sample means are closely distributed around the population mean.
- The 25th percentile is the value at which 75% of the answers lie above 5 and 25% of the answers lie below 5. The 50th percentile of Total Deaths tells that 50 percent of the time total death will be below 117. The 75th percentile is the value at which 25% of the answers lie above 1111 and 75% of the answers lie below 1111.
- The first quartile or Q1 is the value in the data set such that 25% of the data points are less than 5 and 75% of the data set is greater than 5. Similarly, for Q2 it is 5, Q3: 1111, and Q4: 0.
- As the coefficient of variation is low for total tests, cases, recovered and deaths, the smaller the level of dispersion around the mean.

For Original Data:

The interpretation of a 95% confidence interval is that "we are 95% confident that the total tests are between 705060.8593660564 and 730041.6751317602". For Total Cases, it is between 71109.3907736677 and 74627.8881200732, for total recovered is between 67831.91051776875 and 71231.35149096487; for total deaths, it is between 929.8939516786945 and 1006.2990614217422.

For Recent Data:

The interpretation of a 95% confidence interval is that "we are 95% confident that the total tests are between 1145380.8104774102 and 1249067.1502368755". For Total Cases, it is between 171367.74685759106 and 188819.86742812322, for total recovered is between 158093.12442526844 and 174854.2077175887; for total deaths, it is between 2037.0616687672195 and 2463.191902661352.

Graphs:

- Bar graphs are ideal for comparing two or more values, or values over time. Bar Graphs is the comparison between recovered and total cases found. Here we can observe that the difference is very low between total cases and recovered. This implies cases were found daily, but the recovery rate is also good.
- Comparison between Total Cases, Recovery, and Deaths on the basis of a particular age group using a bar graph. Total Cases found and recovery rate is high for the age group of 20-30 and the death rate is negligible. Whereas it is the opposite in the case of 70+ age.
- As the value of correlation, closer is to 1, for (cases, tests), (recovered, tests), (deaths, hosp) i.e 0.98,0.97,0.86; which implies stronger this relationship is. A correlation closer to -1 is similar, but instead of both increasing one variable will decrease as the other increases.
- The number of Total tests done has been reduced than before for all age groups. Similarly, the number of Total cases found has been reduced than before for all age groups except the 80+ age group.
- A boxplot is a standardized way of displaying the distribution of data based on a five-number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). The recovery rate is high in the case of the 20-30 age group, whereas, it is low in the case of the 60+ age group.
- As in this Dist plot, the bar is high at 0-1000, which implies that the number of deaths occurring recently is less i.e. between 0 to 1000.
- As in the Scatter plot, the total cases found and recovery rate is almost a straight line. So, we can say that the total cases found and recovery rate goes hand in hand.
- As in this Cat plot, shows frequencies is high for the 71+ age group, which implies that the number of deaths is high in this age group.

## Conclusion:

Total Number of Tests, Cases, Recovery rate and death cases are high in 21-30 and 31-40 age groups. Also, the recent data (1 Jan - 25 Feb) conclude that covid rate is less in this year than previous one.