International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 08 | Aug 2021 www.irjet.net p-ISSN: 2395-0072

A Comparative Study of Various Data Visualization Techniques using COVID-19 Data

Sai Vasavi Harsha Vardhan Gupta Somisetty*[1], Akhil Songa[1], Sai Tanmai Raavi[1], Sri Teja Kumar Reddy Tetali[1], Sailesh Edara[1], Dr Bhavani Madireddy[2]

[1]Student, Department of Computer Science and Engineering, GITAM University, Vizag, Andhra Pradesh, 530045, India
[2]Assistant Professor, Department of Computer Science and Engineering, GITAM University, Vizag, Andhra Pradesh, 530045, India.

Abstract :

Objectives:
To apply various data visualization techniques on COVID-19 datasets that help to get insights, make accurate decisions, find trends and patterns, and present valuable information.

Methods:
The following data visualization techniques, i.e., Bar graph, Box plot, Bubble chart, Choropleth map, Density plot, Heat map, Histogram, Line graph, Network analysis, Parallel coordinates, Pie chart, Scatter plot, Scatter plot matrices, Timeline chart, Time series plot, Tree map, Violin plot, Word cloud are used on various Covid-19 datasets taken from Kaggle.

Findings:
Analyzing vast volumes of data is a tedious task and practically impossible, and hence data visualization is needed to make those datasets meaningful. There are various types of data like categorical, numerical, spatial, time-series, and many more. For analyzing various types of data, different data visualization techniques are required. On the same data, different techniques provide different insights and different ways of visualization. On COVID-19 data different techniques gave different insights, for instance, density plot gave a clear idea about the distribution of COVID-19 cases, Pie chart about recovery, death, and discharge rate, though box plot and violin plot are drawn using the same data their visualizations are completely different. Every data visualization technique has its own pros and cons, so knowing various techniques is essential.

Applications:
Data visualization finds its application in each and every industry. This paper discusses how data visualization can be applied to Covid-19 data and how meaningful insights can be drawn.

Keywords: *Data Visualization, Data Visualization Techniques, COVID-19, Exploratory Data Analysis (EDA), Explanatory Data Analysis*

1.INTRODUCTION

From the past couple of years, data has been increasing enormously. From an individual to a large enterprise, everyone continuously consumes and leaves behind vast amounts of data. Nearly 500 hours of content is being uploaded on YouTube every minute [1], Google nearly processes 20 petabytes of data every day [2], and it is estimated that the retail giant, Walmart collects more than 2.5 Petabytes of customer data every single hour [3]. This gives a clear picture of how much data is being generated over the internet every single hour.

Such a large volume of data left untouched is of no use, but exploring and understanding such data without any aid seems next to impossible. So a way to get insights into data at one glance is needed. Data visualization is the technique that helps

to represent massive datasets in simple yet highly informational pictures and graphs. Data visualization represents the relationship between various characteristics of data effectively. It helps to make quick, effective, and accurate decisions that cannot be made directly from raw data. Data visualization can be applied to any industry but is primarily used in Finance, Health, Agriculture, Retail, Travel, and many more. For example, a research article named " Agricultural Data Visualization for Prescriptive Crop Planning " analyzed the relationship between crop growth and various weather parameters like the temperature, rainfall, and the levels of reservoirs in the state of Karnataka with the help of data visualization and predicted proper weather conditions for generating a better yield of the crop [4]. Data visualization is a very powerful tool that can be applied to any dataset irrespective of the domain.

Data visualization is now a buzzword, but its concepts and implementations trace back to the 17th Century. Following is Figure 1, which depicts the evolution of Data visualization.

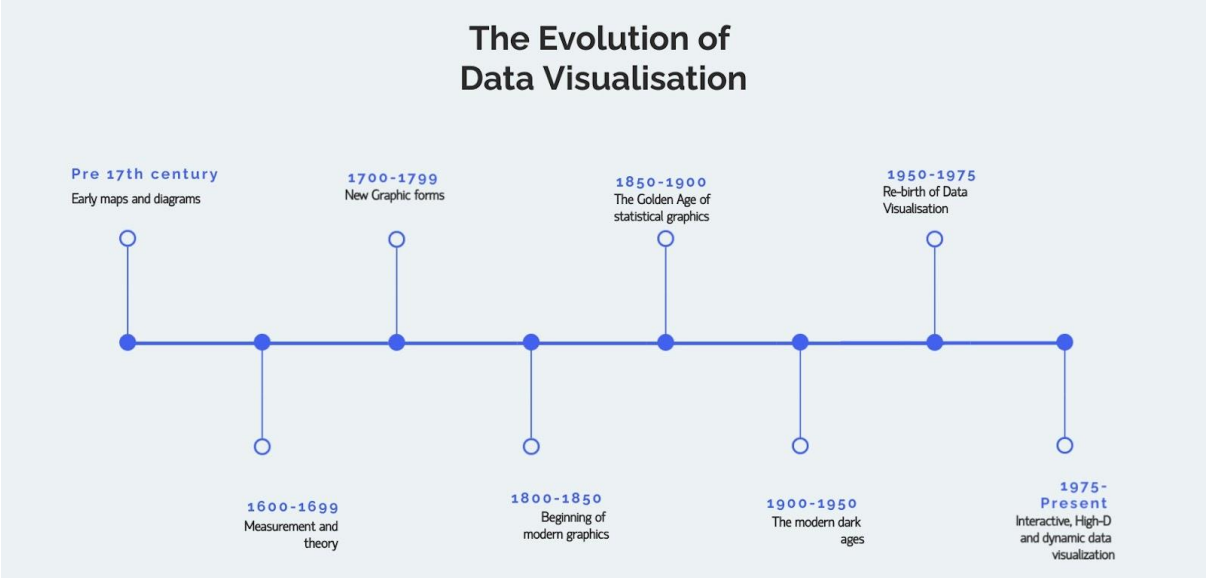


Figure 1. Evolution of Data Visualization [30]

Primarily, two types of analysis can be made using data visualization. Those are Exploratory data analysis(EDA) and Explanatory data analysis. EDA deals with exploring the data and helps to answer questions regarding data behavior, whether data contains any noise, how the features are dependent on each other, and many more.

An explanatory analysis is performed after the development of the model. The explanatory analysis explains the outcomes of an analysis made on the given data set, i.e., it helps to present something observed in the data to the stakeholders. Also, it helps to present the predictions made by the developed model. Data visualization is an integral concept of Data Science, without which the ability of a data scientist to explain the facts is hindered [5].

Figure 2 is the flow of development of a general computation model. The stages of Exploratory and Explanatory analysis are highlighted in the flow chart.

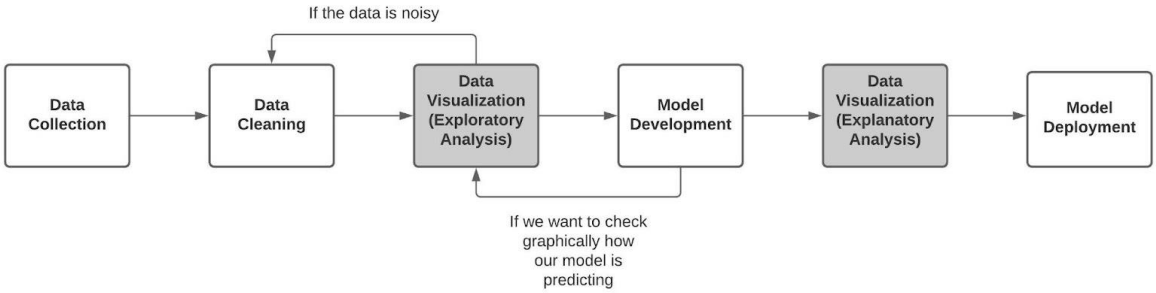


Figure 2. Flow chart showing data-driven computational model development

"By visualizing information, we turn it into a landscape that you can explore with your eyes, a sort of information map. And when you're lost in information, an information map is kind of useful."

—David McCandless

2. Types of data visualization

Figure 3 Clearly shows the various types of data visualization.

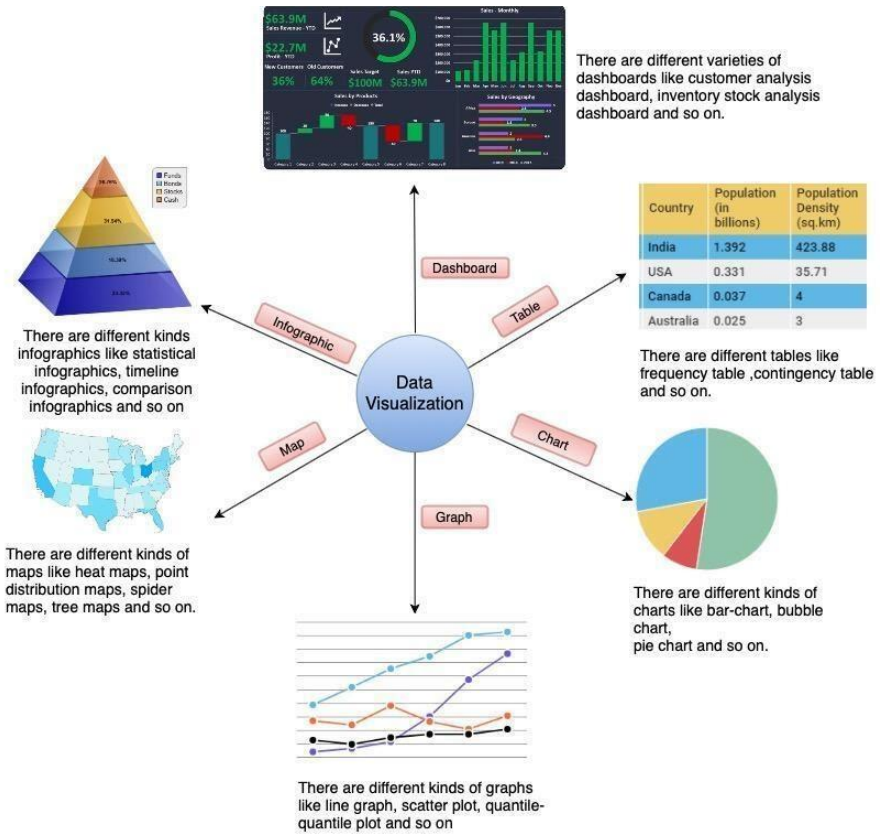


Figure 3. Types of Data Visualization

2.1. Table

A table organizes the data in rows and columns. Generally, in a table, rows represent the data, and columns represent the features or properties. Tables can represent n-dimensional data like 2-D, 3-D, 4-D. Drawing a table requires significantly less effort when compared with other data visualization methods. Tables are flexible depending on the data. So usually, tables are the first visualization technique used to get quick insights into the data.

2.2. Chart

A chart is a pictorial representation where the data is represented using symbols like lines, bars, slices. This helps the users to understand the given data and helps them to predict future trends. Some examples of charts are Pie charts, histograms.

2.3. Graphs

A graph is a diagram that shows the relation between any variable quantities. Generally, graphs are drawn between two variables. It is a representation of numerical data where it shows the mathematical relation between different data variables. Basically, a graph is a subset of a chart, which means every graph is a type of chart but not every chart is a graph. Some Examples of graphs are scatter plots, quantile plots.

2.4. Maps or geographical visualization

Map visualization is used for analyzing geospatial data and visualizing it as maps. This type of data presentation is more straightforward and intuitive. The distribution or proportion of data in each region is seen graphically [6]. Some Examples of Maps are choropleth maps, Heat maps.

2.5. Infographics

Infographics is a fusion of "Information" and "Graphics". It is the visualization of data that converges complex information into a simple, easy-to-understand format for the audience by using visual cues. Infographics use multiple data visualization techniques and also additional elements like graphics, illustrations, and typography [7].

2.6.Dashboard

A Dashboard is a data visualization tool used to manage and collectively display all the vital information in a graphical way. It displays various visualization techniques in a single window which helps to compare the insights from multiple techniques in a simple manner. A dashboard also shows historical data and brings multiple metrics together. It may look like a report, but the main difference between a dashboard and a report is, dashboards are interactive, whereas reports are static. It can also be used to predict trends [8]. A best well-known example of a dashboard is google analytics.

3. Data Visualization Techniques

Data visualization techniques are used to unlock the benefits and make accurate decisions from the vast data. Following is the list of various data visualization techniques.

3.1. Bar Graph:

A Bar graph helps to visualize categorical data. It represents the data using bars where its size varies proportionately with the value of the data. Typically, the higher the value, the more significant is the size of the bar. Bar graphs are used for quantitative analysis between different groups because they allow the analyzer to recognize patterns and trends quickly. Popularly used various types of bar graphs are Horizontal bar graphs, Vertical bar graphs, Grouped bar graphs, and Stacked bar graphs [9].

In figure 4, X-axis represents the states of India, and the Y-axis represents the number of deaths. Each state of India is a category, and the length of the bar depends on the number of deaths recorded in a particular state.

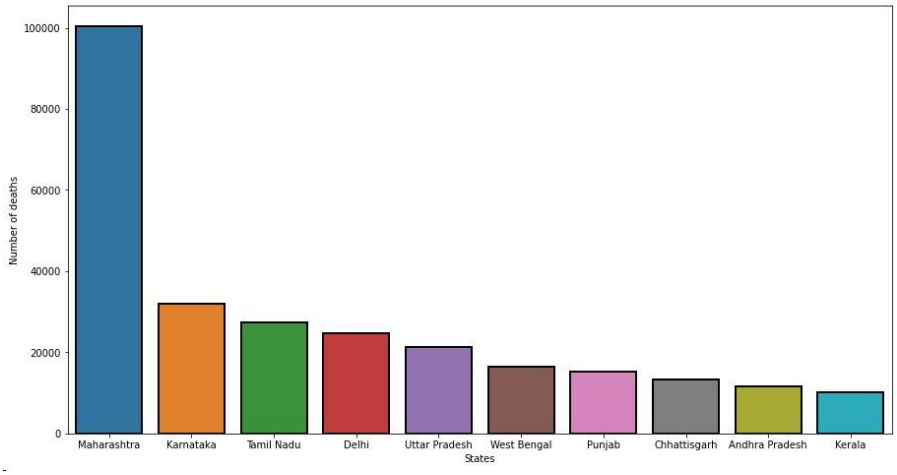


Figure 4. Top ten states with the total number of deaths in India as of 8th June 2021.

3.1.1. Gained Insights:

From the graph in figure 4, it can be seen that Maharashtra has recorded the highest number of deaths compared to the other nine states in India. Based on this information, the Government can know which states have high cases and take precautions accordingly.

3.2. Box Plot:

A box plot is used to represent numerical data. It is made of two parts, a box and a set of whiskers. Refer to figure 5(a) and 5(b) to understand the parts of the box plot. There are two ways of drawing a box plot. In the first method, a box plot is drawn without considering any outliers and drawn directly. In the second method, a box plot is drawn along with the outliers. Firstly, the interquartile range has to be applied, and then the outliers are to be identified, and finally, the plot is to be drawn. Box plots are used to view the spread of the entire data, distribution of data from the median, detect the outliers present in the data, and compare data from multiple data sets [10].
For instance, consider Figure 5(c). Here the box plot is generated by considering marks of students from three different schools.

School-1 : [90,100,50,45,80,79,20,79,60,30]
School-2 : [99,100,65,45,80,79,70,79,90,50]
School-3 : [10,90,50,75,80,79,60,79,60,30].

In the figure, the average, highest, lowest values of each school are clearly depicted, and using the box plots, the school that is performing the best among the three can be easily identified. For the given data, the performance of school 2 is better, compared to other schools.

In figure 5(d), X-axis represents the month, and the Y-axis represents the confirmed Covid-19 cases. The box plot is drawn after applying the Interquartile Range and finding out the outline.

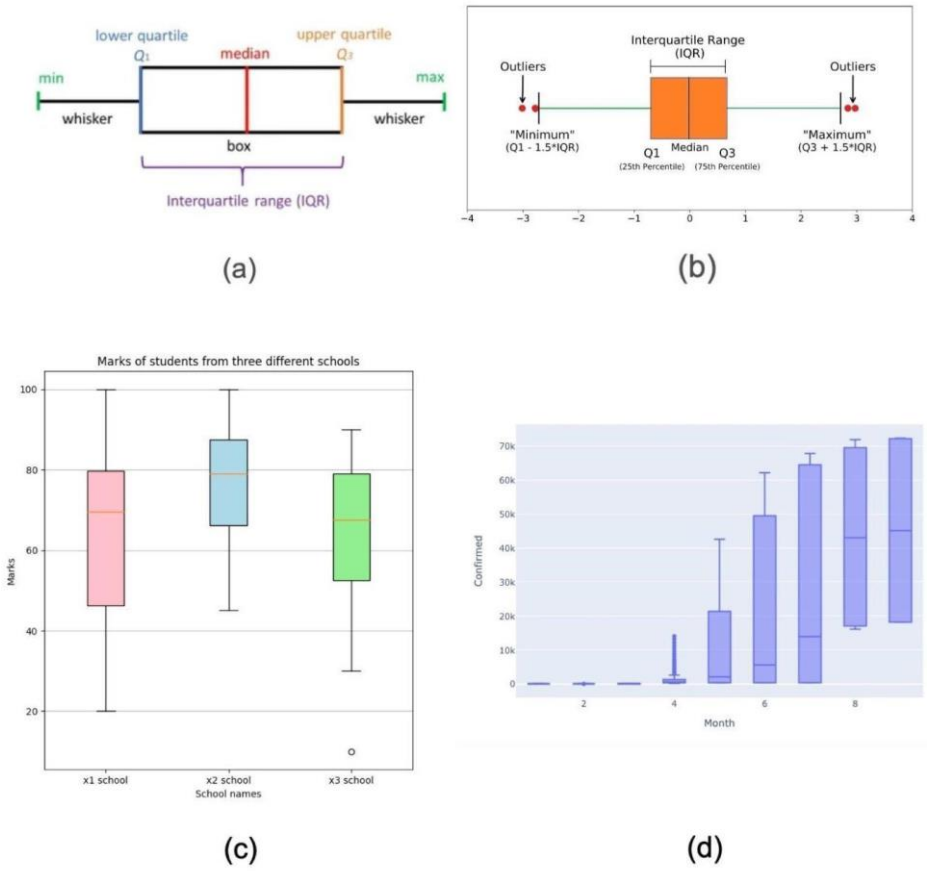


Figure 5.
a) Boxplot without outliers
b) Boxplot with outliers
c) Comparison of three schools
d) Confirmed Covid-19 cases from January 2020 to September 2020 in Guatemala

3.2.1. Gained insights:

Figure 5(d) clearly shows that the cases kept on increasing with time. It can be observed from the graph that the cases have started increasing from the fourth month. The presence of many outliers in the box plot of the fourth month suggests that the cases could have increased due to some anomalous events. The government needs to study for the sudden spike in the cases and make sure the same situation does not repeat in the future. Also, one can easily observe the maximum, minimum, and average cases confirmed each month. In the eighth and ninth months, minimum confirmed cases are far higher when compared with previous months.

3.3. Bubble Chart:

A bubble chart is a variation of the scatter plot. If dots in the scatter plot are replaced with bubbles, it becomes a bubble chart. Unlike scatter plots, a bubble chart uses 3-dimensional data for plotting, where the two dimensions(x,y) are used for pointing in the cartesian plane, and the third dimension is used to specify the size of the bubble. The size of the bubble is directly proportional to the data associated with that bubble.

There are different ways to scale the size of a bubble, i.e., scaling by the diameter or the bubble area. Generally, Scaling is done using the diameter. Apart from simple cartesian planes, maps, some graphics can be used to plot bubble charts [11]. There are different varieties of bubble charts like simple bubble charts, labeled bubble charts, 3d bubble charts, packed bubble charts, bubble map charts, multivariate bubble charts, and many more. The advantage of bubble map charts over

the choropleth map is that the choropleth map uses different colors to represent the data, where multiple shades are hard to differentiate.
The main disadvantage of a bubble chart is that having many data bubbles nearby can make the chart harder to read.

Figure 6 is a bubble chart describing the cases recorded in various countries around the world. The bubble's position represents the country, and the size of the bubble depends on the number of cases recorded.

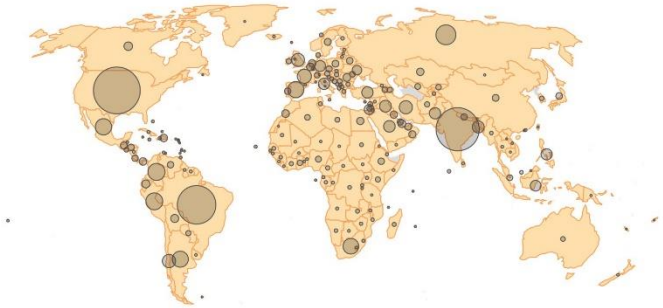


Figure 6. Confirmed covid-19 cases worldwide till July 26, 2021

3.3.1. Gained insights:

From the bubble chart in figure 6 it is clear that the US, Brazil, and India have the highest number of confirmed cases compared with the rest of the world.

3.4. Choropleth maps:

Choropleth maps are geographical maps used to depict the quantitative data or disparity in quantitative data among different geographical regions like countries, states, or any other geographical locations. Visual factors like colors, shades and brightness are used to show the data or variation of data. Choropleth maps are the most used thematic maps because of their easy understanding capability. In Choropleth maps, generally, the darker the shade, the higher the value will be [12].
Figure 7 shows the choropleth map regarding total deaths due to covid 19 in India.

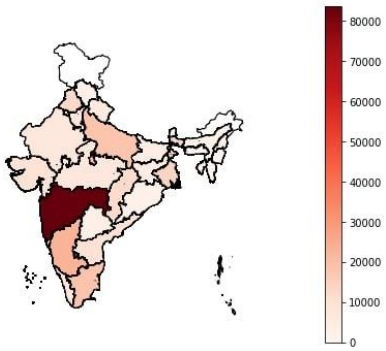


Figure 7. Total Deaths due to Covid-19 in India till 1st week of March 2021

3.4.1. Gained insights:

From figure 7, it is clear that the state of Maharashtra has higher deaths due to Covid-19 when compared with other states in India. Once the government knows which states suffer from higher death rates, the government can study whether deaths are caused due to the shortage of oxygen, shortage of beds in hospitals, or a new variant of the Covid virus.

3.5. Density plot:

A density curve is used to represent the distribution of numeric data. It plots the distribution of the entire data in a single continuous curve. It is a smoothened version of a histogram. The main difference between a histogram and a density plot is that a histogram uses bins to show distribution, whereas a density plot uses a continuous curve to show the distribution. There are different types of distributions like uniform distribution, triangular distribution, normal distribution, multimodal distribution, and many more. Depending on the type of distribution, the shape of the density plot varies [13]. Density plots are better for large data sets compared to histograms.

Figure 8 shows the total distribution of covid-19 cases in India. The x-axis represents the date, and the y-axis represents the number of cases(lakhs).

Figure 8. Total covid-19 cases in India till 25-July-2021

3.5.1. Gained insights:

From figure 8, the highest number of cases recorded in India during the first wave was around 1 lakh, whereas the highest cases recorded were around four lakhs during the second wave. During the second wave, the cases increased drastically in a very short span of time compared to the first wave. Daily recorded cases during the second wave were very high when compared to the first wave. India was affected very badly during the second wave.

3.6. Heat Map Chart:

A heat map represents the numeric data using color variations done by changing the shade, intensity, or hue of the color. Generally, there are two different types of heat maps: Cluster heat maps and Spatial heat maps.

A cluster heat map uses a grid or tabular structure to represent the data, and the insights can be derived by observing the intensity of the color of a particular cell. Whereas spatial heat map uses geographical locations on a map. The main difference between the spatial heat map and the choropleth map is that in a choropleth map, there are predefined geographical boundaries like countries, districts, cities, streets, which are colored based on the data variation, whereas in a spatial heat map, there are no predefined boundaries [14]. See Figure 9(b), which shows a heat map of real estate prices in and around Warsaw city, the capital of Poland.

The heat map can be used for the feature selection mechanism of a computational model by visualizing the correlation between the features(there are various algorithms for calculating the correlation between the features). Also, Not a Number(NaN) values can be found out from a dataset with the help of a heat map. For instance, consider figure 9(a), a heat map generated by using a dataset of Tesla stock price taken from NASDAQ. From the figure, it is clear that there are no NaN values in the dataset as the heat map is filled with the color corresponding to '0' in the given scale.

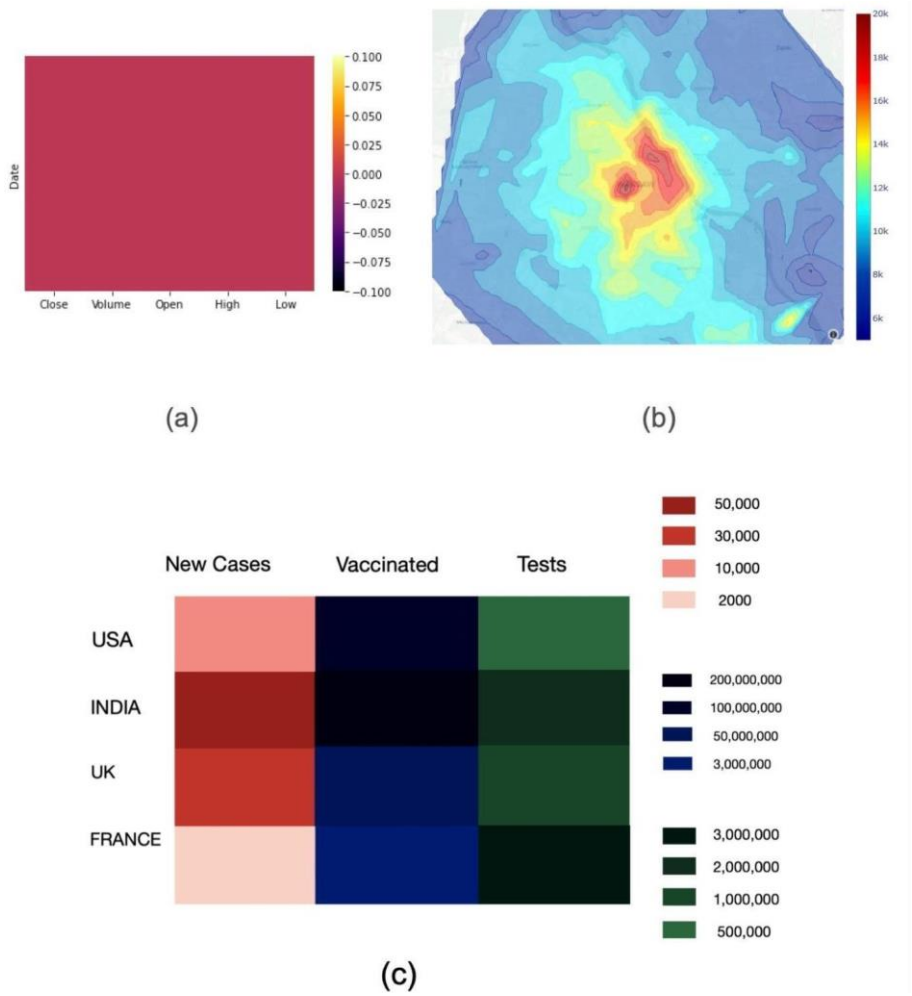


Figure 9.

a) Not a Number(NaN) values present in tesla stock price data set

b) Heat map of real estate prices in and around Warsaw city

c) Heat map comparing the new cases registered, number of people vaccinated, and the rate of tests performed in four countries: the USA, India, the UK, and France till July 2021.

3.6.1 Gained insights:

From figure 9(c), it can be seen that new cases registered were highest in India and lowest in France. Regarding the total number of citizens vaccinated, India and the USA are on the top, while France has done the least vaccinations. France tops the list regarding the number of tests done, and the USA has the least number of tests done.

3.7 Histogram:

Histograms represent the distribution of continuous numeric data for a single quantitative variable like age, weight, marks, covid cases, and many more. In histograms, total distribution is divided into several bins; the width of the bin plays a vital role in the histogram. The size depends on the data associated with that particular bin. Greater the bin width, the smaller the number of bins, and vice versa. Bin width controls the resolution of the histogram, if the bin width is high, the histogram becomes over smoothened, and if the bin width is lower, the histogram becomes under smoothened. If the bin width is too wide or narrow, it is not easy to see the true data distribution. Figure 10(a) shows histograms with various bin widths on the same data. In the figure, X axis represents number of participants and Y axis represents their mean Systolic BP. Even though the histogram and bar graph look visually similar, the histogram represents the frequency distribution of continuous data. However, the bar graph represents the count of discrete or categorical variables. Histogram can be used to find outliers, bimodality, skew, other useful features of shape in the distribution and compare subgroups present in the

data [15].

In Figure 10(b) x-axis represents the age, and the y-axis represents the count. The light purple color represents females, whereas the other represents the male.

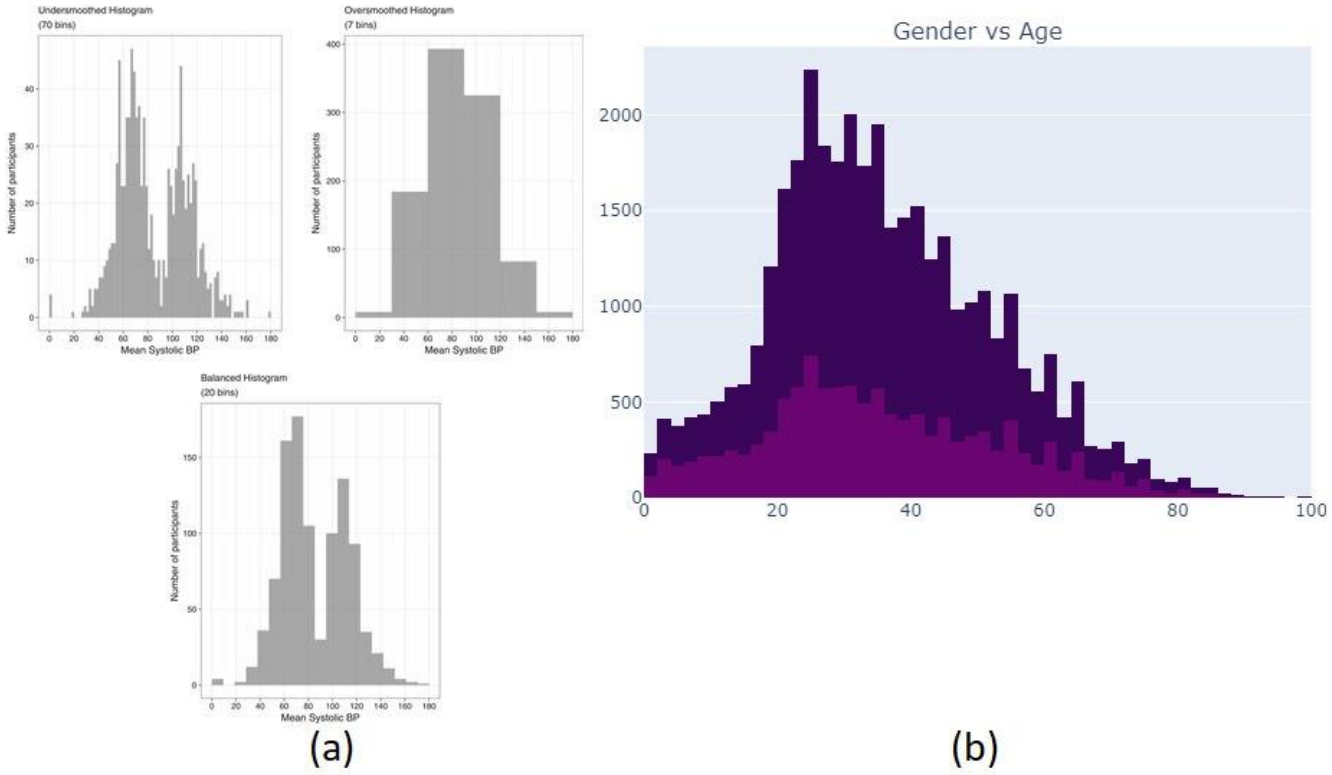


Figure 10.
a) Histograms with various bin widths [15]
b) Covid-19 deaths on 26th April 2021 in India.

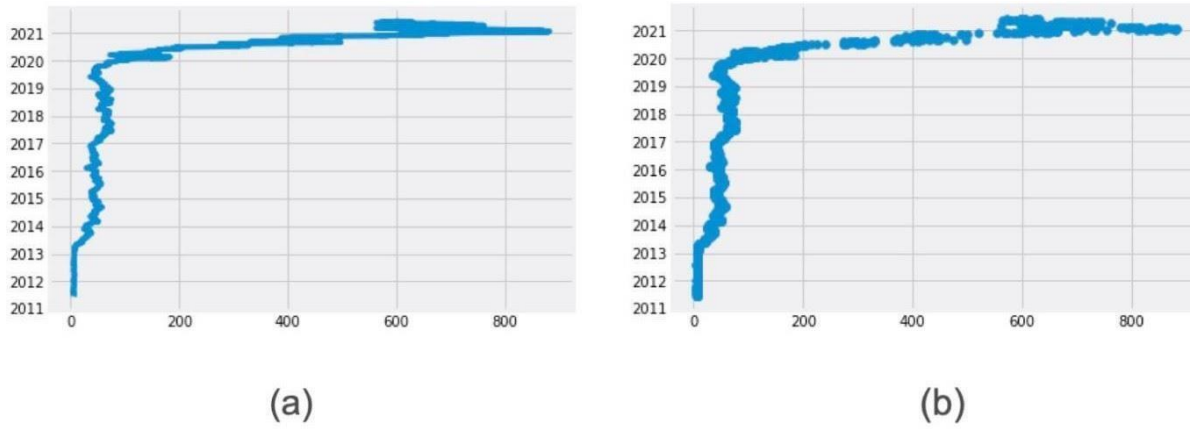
3.7.1 Gained insights:

From figure 10(b), it can be seen that there are more deaths in both male and female patients belonging to the age group 20 - 40. Also, more men were prone to death compared to women.

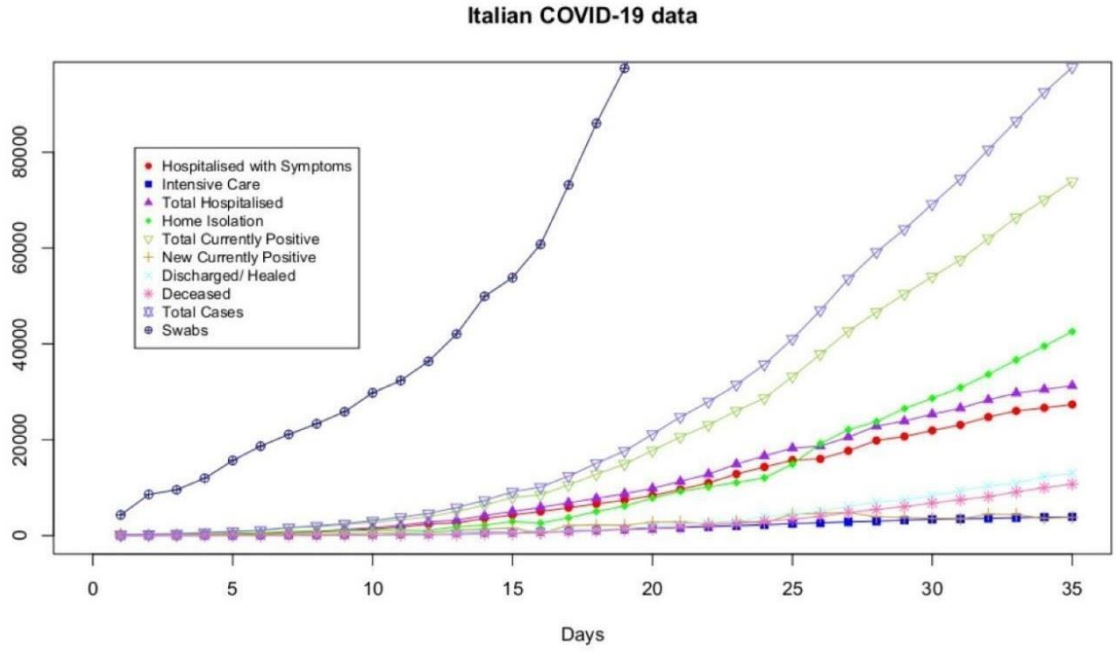
3.8 Line Graph:

A line graph is used to represent continuous data; mostly, it is used for displaying the information that changes over time. A line graph is almost similar to a scatter plot; the main difference is that line segments connect data points in line graphs. Because of the line segment, the variation between points can easily be compared [16]. Line graphs are best for small data. If line graphs are plotted on huge time series data, the line graph will almost look like a scatter plot as the line segments between the data points are very small and not visible.

Figure 11(a) and 11(b) is plotted by taking the Tesla closing price from 1st January 2011 to 1st July 2021, and there is not much difference between both the graphs. Line graphs represent small datasets very effectively as the line segment between each point in the graph gives a great understanding of how the data varies with time.



(a) (b)



(c)

Figure 11

a) Line graph representing Tesla Stock Price from Oct 6th, 2011 to Sept 9th, 2021.
b) Scatter plot representing Tesla Stock Price from Oct 6th, 2011 to Sept 9th, 2021.
c) The number of people hospitalized with symptoms, people who need intensive care, current positive cases, deceased number, and other features in Italy from 24 February to. 29 March 2020 [31].

3.8.1 Gained insights:

From figure 11(c), it can be seen that there is no drastic change in the number of people who require intensive care though the number of people hospitalized has gradually increased during 35 days. Also, from day 25, the home isolation count has increased compared to hospitalization. Considering the swabs curve, the slope of the line segment joining the points on day 1,2 is higher than the slope of the line segment joining the points on day 2,3. It implies that the total swabs taken from days 2, 3 were comparatively less when compared to days 1,2. Similarly, based on the slope of the line segment between the points, other features like total currently positive, New currently positive, and Deceased can be compared.

3.9. Network analysis:

From molecules within a substance to relationships among human beings, everything forms a Network of connections. The process of analyzing networks is known as network analysis.

Some more examples for networks like Mutual friends in Facebook, LinkedIn connections, the spread of disease, food chain, World Wide Web (WWW) can be considered. Any network can be categorized as either a centralized network or a decentralized network. Understanding whether a network is centralized or decentralized is critical because, in a centralized network, there is a root node from where the network begins. Every new node that adds to the network will be under the root node; every node connects to the root node. The disadvantage of such a system is that if the root node is disconnected from the network, every node of the network will get isolated.

As the name suggests, a Decentralized network is a system that is strictly against the policy of a single root node acting as a representative of the network. Here, all the nodes of the network are interconnected. Even if a single node is disconnected, the entire network remains stable [17], [18].

Network analysis is used to identify the type of network, identify the nodes connected together in a network, know the layout of the network, and find interesting patterns that lead to more valuable insights, and many more.

Consider Figure 12. It is a plot representing the network analysis of how Covid-19 infection was spread among different age groups in Karnataka. The node's shape in the figure represents the gender, while the color indicates the person's age group. The larger the node's size, the higher the possibility that people belonging to that gender and age group were more responsible for spreading the infection. The smaller nodes connected to the large nodes are those classes of citizens who got infected due to the large node they are connected to.

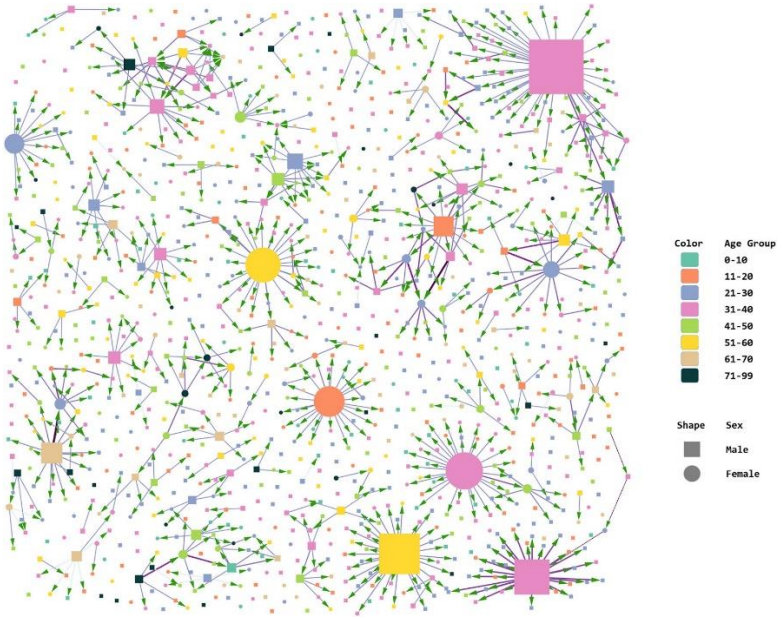


Figure 12. Network Analysis of Covid-19 infection transmission among different age groups in Karnataka from 9 March to 17 May 2020 [32]

3.9.1 Gained insights:

From figure 12, it is clear that both men and women belonging to the age groups of 31-40 and 51-60 played a significant role in the transmission of the COVID-19 in Karnataka.

3.10. Parallel coordinates:

The parallel coordinate system is a particular type of constructing a graph, where the axes of the graph are placed parallel to each other. In the traditional cartesian system, the axes of the graph are perpendicular to each other. In the regular cartesian system, the data is plotted as a point in the graph. However, in a parallel coordinate system, the data is represented using a unique curve known as a polygonal line.

A polygonal line, refer to figure 13(a), is a line that connects various points. Conventionally, a point is plotted using a normal cartesian system by considering the perpendicular distance from the X and Y-axis depending on the X and Y-coordinate values. The same is not the case with the parallel coordinate system. In parallel coordinates, the nth coordinate value is plotted on that respective axis. When all such plotted coordinates are joined together, it forms a polygonal line.

In the parallel coordinate system, as the graph's axes are parallel, visualizing multidimensional data becomes more accessible and interactive. It is used to compare various features, i.e. multidimensional data, common to a category of items and deduce valuable insights from them [19], [20]. Consider figure 13(c), for instance. It is a parallel coordinate plot between different varieties of food items and a set of most frequently found nutritional values in most foods. Here, the leftmost axis contains the food items being studied, and each of the subsequent axes represents the nutritional value under study.

Here, there are fourteen different types of nutritional values, or rather it can be said that the parallel coordinate plot here is a 14-Dimensional plot. If the polygonal line is higher on a particular axis, it can be interpreted that the food being considered is rich in that particular nutritional value. On the other hand, if the polygonal line points to a lower value on some axis, that particular food is deficient in that nutritional value. On a general note, by observing the graph, it can be said that most of the food under study has high amounts of water and carbohydrates in them.

A parallel coordinate graph is highly complex to analyze directly, so techniques like brushing, bundling, progressive rendering are applied on the various dimensions available to filter out the required information. The paper "Smart Brushing for Parallel Coordinates" discussed the brushing approaches that can be applied on parallel coordinates to gain more insights easily [21]. Upon applying brushing on figure 13(c), another refined plot is obtained (figure 13(d)), which has only those food items which have 0-20 grams of fiber, 0.25-1 gram of Vitamin C, 0-5 grams of potassium.

Figure 13(b) shows the parallel coordinate plot of positive PCT(procalcitonin) levels of patients in India, recorded on three different days with a time gap of one month each. Each of the coloured circle represents States and Union Territories of India. Each of these three days is a dimension in the given plot. The variation in the PCT serum level of a person can be attributed to the criticalness of the patients affected due to Covid-19. If a patient is highly critical, his/her PCT levels are high compared to those who are moderately affected [22]. Comparing patients from different states with respect to different dates becomes easier in parallel coordinates

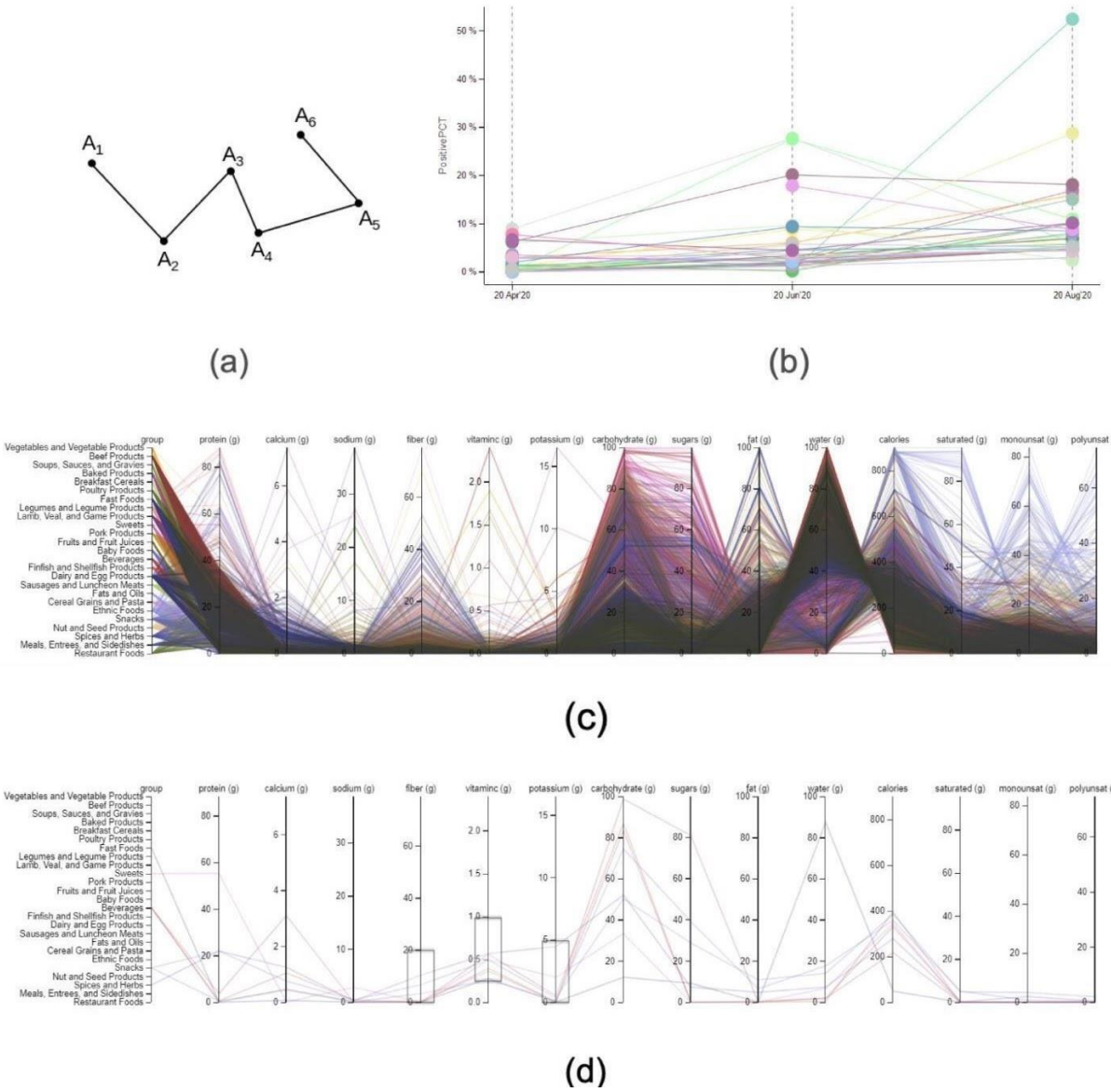


Figure 13

a) Polygonal Line joining five different points.
b) PCT levels of patients diagnosed with Covid-19 from various states in India.
c) Parallel Coordinates representing 14 different nutritional values in commonly consumed food items.
d) A newly filtered parallel coordinate plot after brushing.

3.10.1 Gained sights:

From figure 13(b) it is clear that Delhi has a PCT level of 8.9% on 20th April 2020, whereas Ladakh has 0.3 on 20th April. On 20th June, Delhi had 27.7%, whereas Ladakh had 27.6%, and On 20th August, Delhi had 6.1% whereas Ladakh had 10.9%. (This shows that even though Delhi has started with a higher PCT level, it eventually managed to decrease their level within August when compared to Ladakh)

3.11. Pie Chart:

A pie chart is used to visualize categorical data. It is visually represented in a circular shape; A pie chart is divided into slices of various sizes, each slice representing a numerical proportion of data of a certain kind [23]. There are different pie charts like the Doughnut chart, Exploded pie chart, Polar area diagram, Ring chart, and many more.

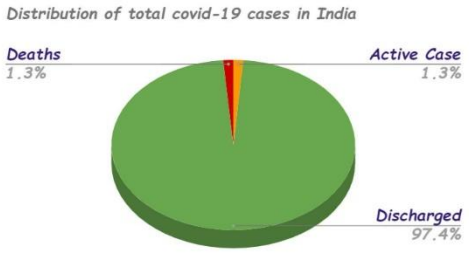


Figure 14. Pie chart showing the distribution of the percentage of total deaths, active cases, and discharge rate in India till July 2021

3.11.1 Gained Insights:

From figure 14 it can be observed that out of the total covid cases recorded, the discharge rate is 97.4%, deaths are 1.3%, and active cases are 1.3%. The space occupied by the discharge rate in the pie chart indicates that the discharge rate in India is very high compared to the percentage of active cases and the percentage of deaths.

3.12. Scatter plot :

A scatter plot is used to represent numerical data by plotting it on cartesian coordinates, where data is primarily represented as dots. There are different markers for representing a data point on the graph. However, the most common marker used is the dot. There are n-dimensional scatter plots, but 2-D scatter plots are commonly used. Consider a 2-D scatter plot drawn by plotting the (x,y) data points on the graph. Here X is the horizontal coordinate, and Y is the vertical coordinate. A scatter plot can be used for the Identification of the correlation between variables, identifying different clusters present in the data, Useful to explore randomness in the given data, Estimate the accuracy of a trained model (figure 15(b) is taken from the article "Community Risk Factors in the COVID-19 Incidence and Mortality in Catalonia (Spain). A Population-Based Study". The figure compares the difference between the prediction of a developed model, i.e., the expected covid-19 cases and the observed covid-19 cases) [24].

In Figure 15(a) the X-axis represents various days between April 2021 and May 2021 on which cases were recorded and the Y-axis is the number of cases associated with the particular date.

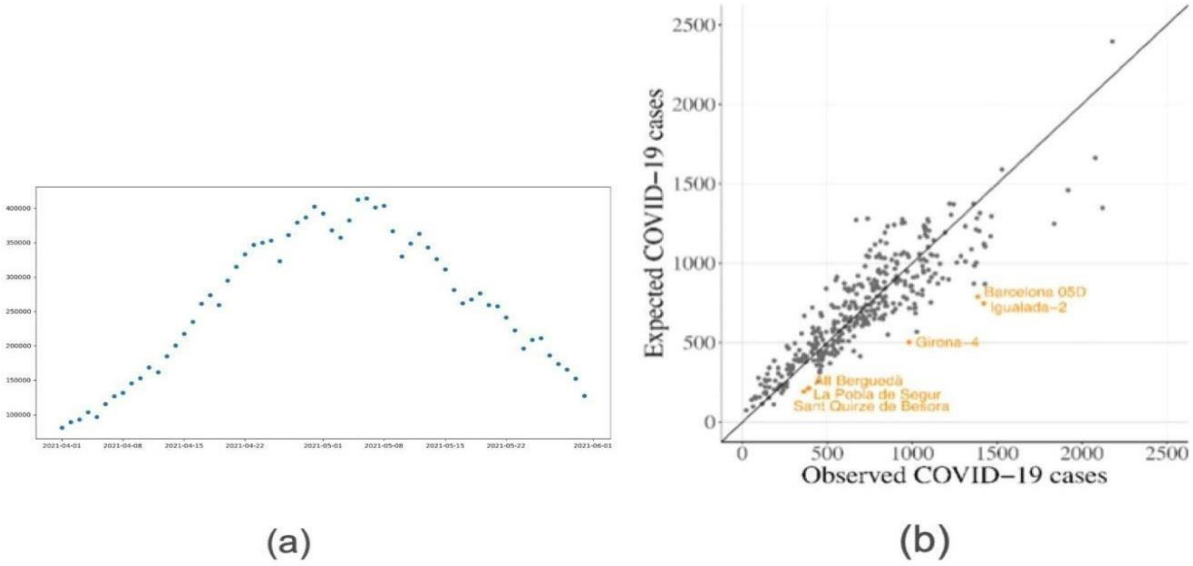


Figure 15.
a) scatter plot showing covid-19 cases from April 2021 to May 2021
b) Scatter plot between observed covid-19 cases and expected covid-19 cases [33]

3.12.1 Gained insights:

Figure 15(a) gives a clear picture of how the number of cases varies over a period of time. The number of cases increased rapidly from April to the first week of May and decreased gradually until June.

3.13 Scatter plot Matrices:

A scatter plot matrix is used to represent numerical data that contain more than two variables or features. It is also called a pair plot. A Scatter Plot matrix is a matrix or a 2-dimensional grid of various Scatter plots. Each scatterplot in the matrix is generated by considering the combinations of all the variables in the data set. Given that a dataset has N different dimensions or variables, the scatter plot matrix generated has (N*N)-N scatter plots. Using a scatter plot matrix, identifying the correlation between more than two variables is easy and can easily compare multiple scatter plots [25].

Figure 16 is plotted by grouping various scatter plots containing data about the recorded Covid-19 cases. The scatter plots are drawn between the UK and Spain, UK and Italy, Spain and Italy, and many more. Covid cases of these countries can be compared and traced efficiently in a single scatterplot matrix. A bar graph can also be used to visualize the same data, but it can only show the count of cases recorded in a country. The variation of the cases registered with respect to time can be traced using a scatter plot matrix, thereby generating more insights.

In the above scatter plot matrix, blue points indicate the data corresponding to the country on X-axis, and red points indicate the data corresponding to the country on Y-axis.

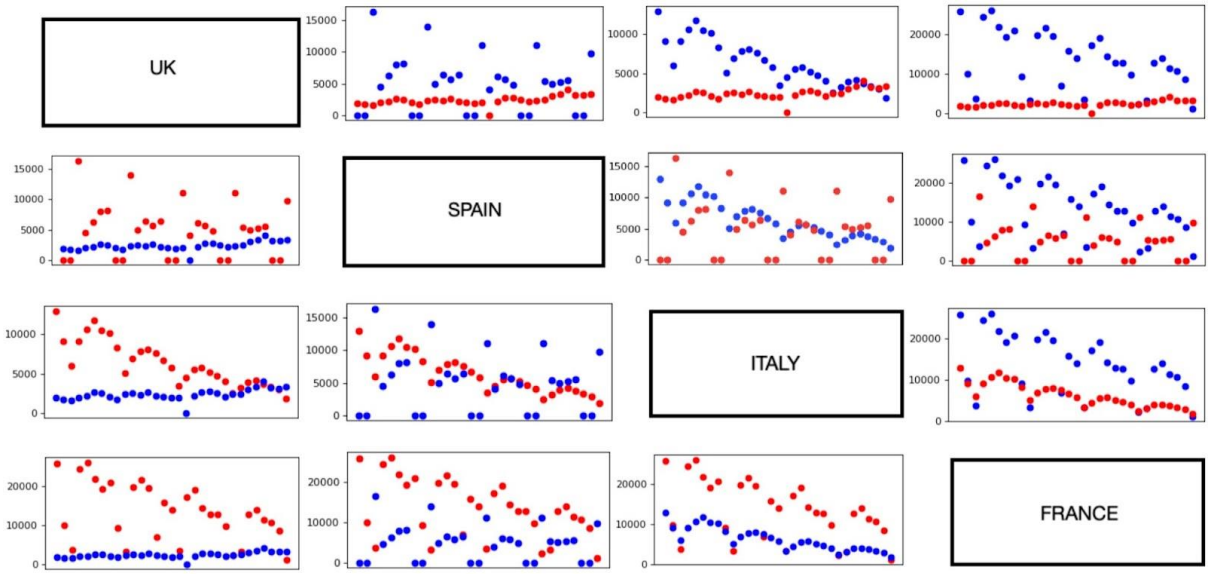


Figure 16. Scatter plot matrix comparing Covid-19 cases recorded in UK, Spain, Italy, France on May 2021

3.13.1 Gained insights:

From figure 16 it is clear that the UK has recorded the lowest number of cases compared to other countries. France has recorded the highest number of cases. The number of new cases recorded in Italy has drastically dropped between 1st May 2021 to 31st May 2021.

3.14 TimeLine chart:

A timeline chart is used to represent the time series data. A *time series* is a record taken at equally spaced intervals of time. Time series data can be collected yearly, quarterly, monthly, weekly, daily, or even hourly. In a time series, the data is always maintained in chronological order. TimeLine chart is invented to manifest a series of events in sequential order. It represents how the resources are used over a period.

Three types of commonly used timeline charts are,

- Gantt chart

- Standard timeline chart
- Time series graph

The standard timeline chart is represented as a long bar, labelled with the date and list of events. These charts are usually simple to create once the data required for the timeline is available [26].

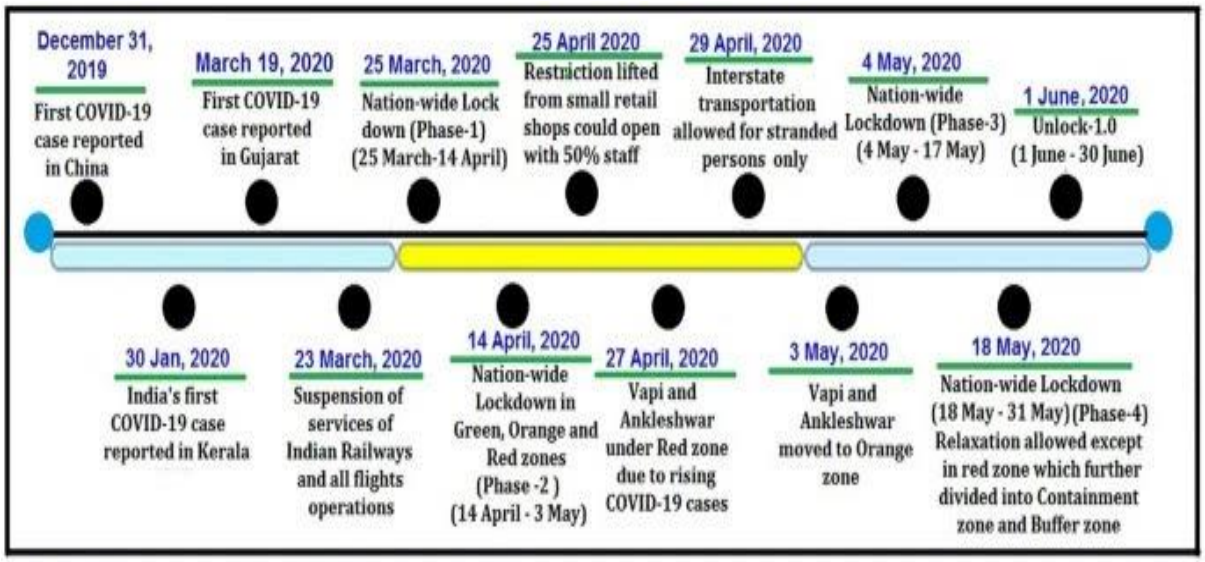


Figure 17. Timeline plot showing various events related to Covid-19 that happened in India from 31st December 2019 to 30th June 2020 [34]

3.14.1. Gained insights:

From figure 17, one can track the entire history of how Covid-19 affected India and also know about the series of decisions taken by the Indian Government to prevent the transmission of infection. From the timeline, it is clear that India imposed a travel ban nearly 50 days after the first case. The first lockdown was announced in India nearly 52 days after the first case. By visualizing the timeline chart, the Government can identify those steps that led to unwanted outcomes and rule out a specific option if the same situation arises in the future.

3.15. Time Series plot:

The graphical representation of time series data is called a time series plot. Time series data have different behaviors like trend, seasonality, cycles, unexplained variation, and irregularities(if data contains outliers). The time series plot shows the behavior of the time series data, which has been recorded over a period of time [27]. Time series plots are used in the Financial industry, Medical industry, Educational Institutions, Disaster management, and many more to get insights into how data varies over a specific period of time.

Figure 18 is a time series plot generated by taking the time series data that contains the number of Covid-19 cases recorded for 250 days in four different countries: the United States of America, Russia, Brazil, and India. Here, the X-axis represents the number of days, and Y-axis represents the number of Covid-19 cases recorded.

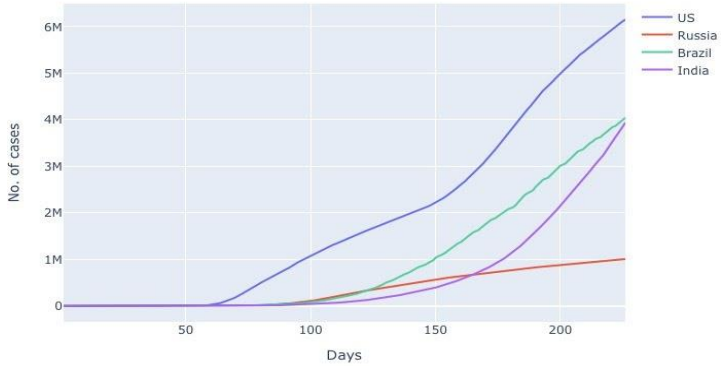


Figure 18. Time series plot between Covid-19 cases recorded in four different countries for a duration of 250 days from Jan 2020 to Aug 2020

3.15.1 Gained Insights:

The graph in figure 18 shows that the new cases recorded in India, Brazil, and the United States kept increasing proportionally with time. However, cases in the USA were increasing rapidly compared to India and Brazil after 50 days.

3.16. Tree Map:

A tree map represents hierarchical data by the use of nested figures(primarily rectangles). The structure of a tree map is similar to the tree data structure. Generally, a tree data structure contains a root node, various branches, inner nodes(child nodes except leaf nodes), and leaf nodes.

In a tree map, the root node is the largest rectangle within which other, comparatively small rectangles are nested. Each of those small rectangles represents a branch for a given node. Such nested rectangles are drawn until the leaf node is reached [28].

Refer to figure 19(a). It represents a hierarchical structure of how India is subdivided into various states and how these states are further classified into cities based on the recorded Covid-19 cases. The same tree structure, when visualized using a Tree map, eases the complexity to get a brief understanding of how each state has been affected by Covid-19, refer to figure 19(b). Also, one can dwell deep into a particular state and understand the distribution of cases in each city of that respective state.

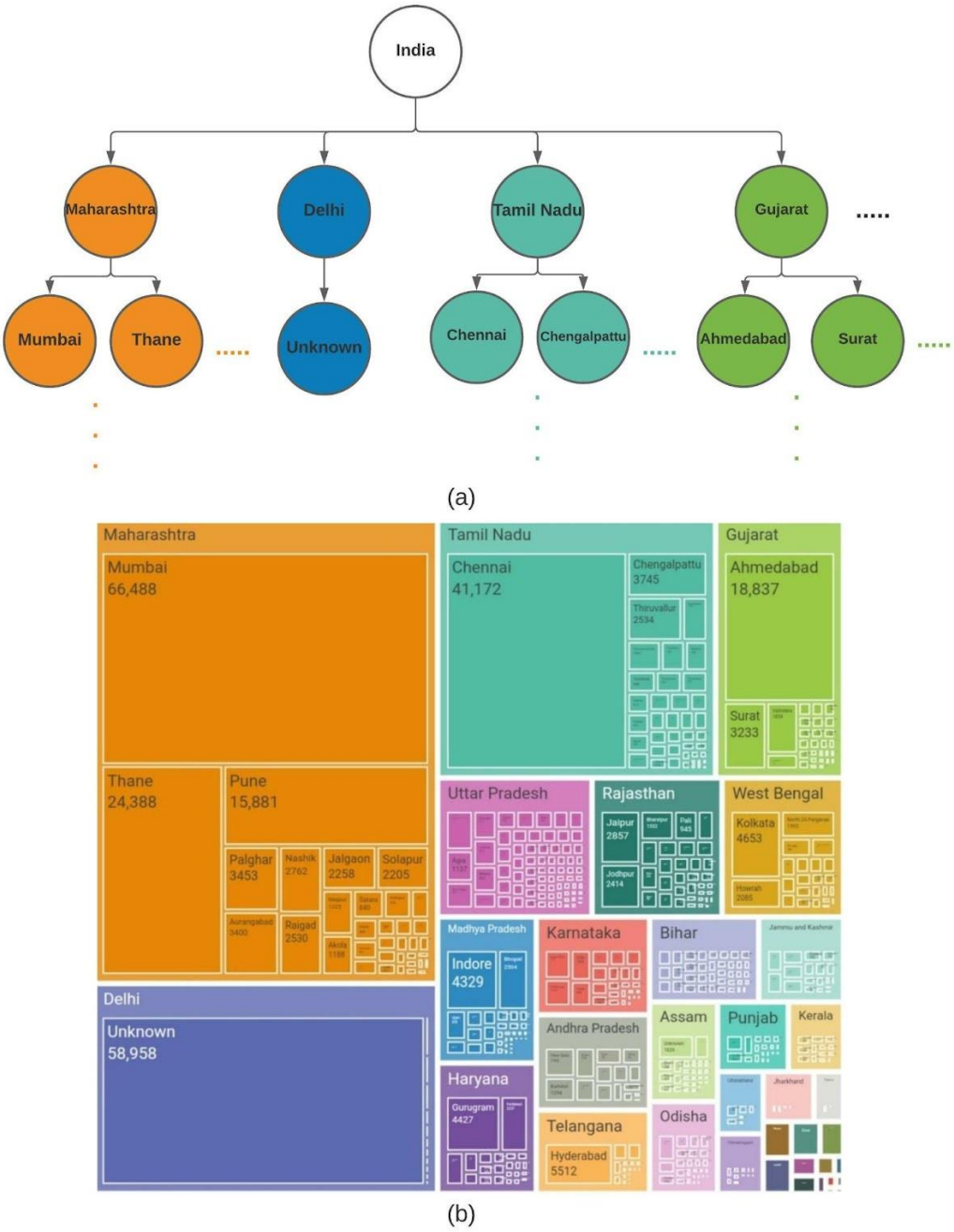


Figure 19.
a) Hierarchical structure of various states and cities of India
b) TreeMap showing the distribution of Covid-19 cases among States and cities in India as of 6th May,2021

3.16.1 Gained Insights:

From figure 19 it is clear that Maharashtra has the highest number of cases in India, and in the state of Maharashtra, Mumbai city has recorded the highest cases than other cities.

3.17. Violin Plot:

A Violin plot represents the numeric data; it is the combination of a box plot and a density plot. Refer to figure 20(a) to understand the structure of a violin plot. The main disadvantage with a box plot is that it does not give any idea regarding the data distribution. From figure 20(b), it can be seen that the box plot does not clearly show whether the distribution of data is bimodal, uniform, or normal. In contrast, upon drawing a violin plot, the variation in data distribution can be clearly seen.

A Violin plot is designed by placing the distribution curve vertically to the left and right sides of a boxplot. For a single data distribution, the curve is symmetric in both directions. When two variables are compared, the curve is not necessarily symmetric; refer to figure 20(c), where the left distribution belongs to the female and the right belongs to the male.

A violin plot is highly beneficial as it combines the advantages of both a box plot and a distribution curve, thereby deriving better insights. A Violin plot clarifies where the data is concentrated and provides a great visualization of the distribution of data and the outliers. It can be used to identify peaks, valleys, bumps and clusters [13]. In Figure 20(c), the X-axis represents countries in the dataset; Y-axis represents the age groups of Covid-19 patients. The left curve represents information about female patients, whereas the right curve describes the male patients.

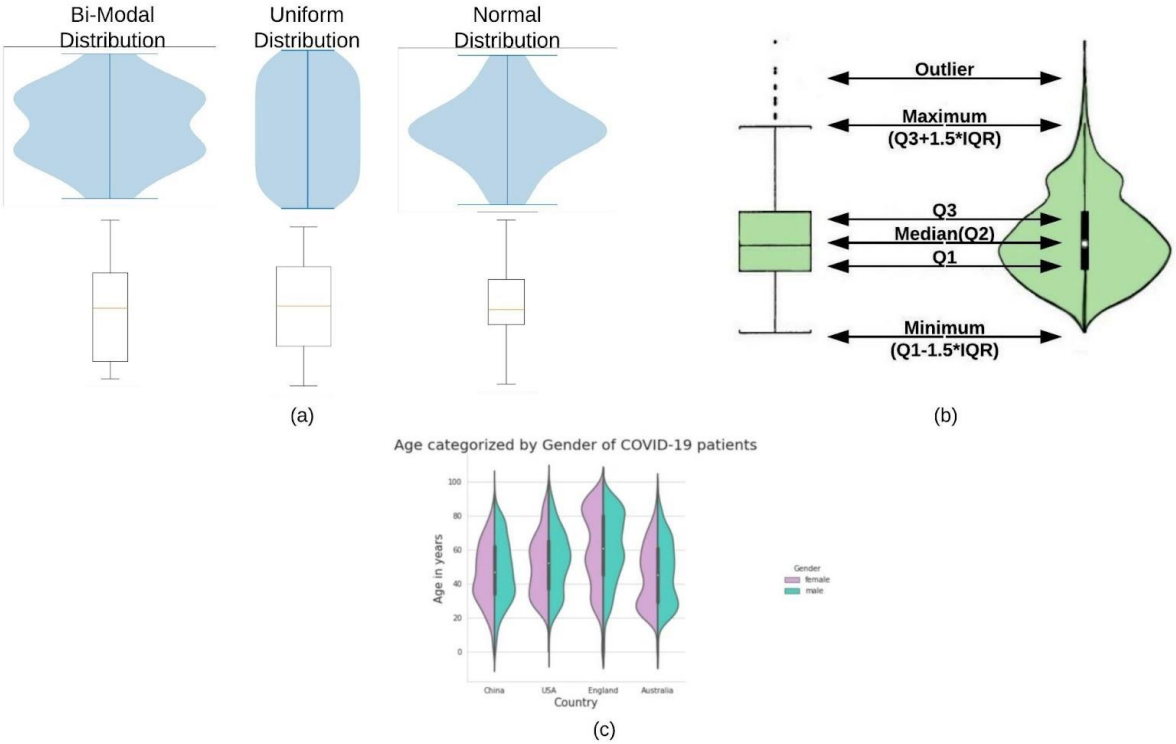


Figure 20.
a) Comparison between Box Plots and Violin Plots on various distributions
b) Structure of a Violin Plot comparing with boxplot
c) Violin plot showing age groups of male and female patients affected by Covid-19 in four different countries

3.17.1. Gained insights:-

From figure 20(c), it can be seen that people above the age of 65 in England were highly affected compared with other countries. In Australia, young citizens between the age groups 20-40 were highly affected. Based on these insights, the Governments of the respective countries can take precautionary measures to minimize deaths in the future.

3.18. Word cloud:

A Word cloud is a key visualization technique in providing a user with the ability to identify the most frequently appeared words in a text document with ease.

Here, the color, size, and boldness of a word depend on the category the word belongs to and its frequency in the given dataset. A group of related words, i.e., words belonging to a similar category, are represented using the same color. Higher the frequency of the word in the given data set, the larger the word's size in the visualization. Word cloud is the first point for starting a deeper text analysis. A world cloud helps in summarizing the entire dataset in an easily understandable graphical form [29].

Figure 21 is a word cloud generated by considering tweets data from Twitter India. The dataset for generating this visualization has been taken from Kaggle.

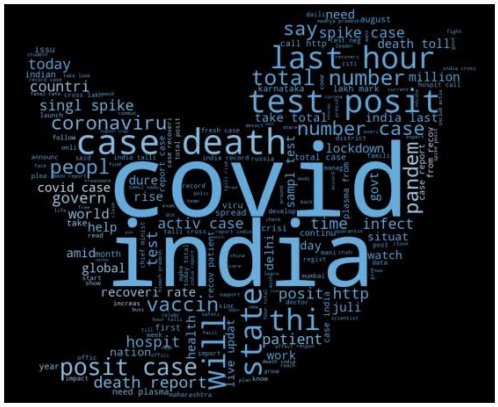


Figure 21. Covid-19 tweets in India between 25th July 2020 to 30th July 2020

3.18.1 Gained Insights:

From the world cloud in figure 21, it is clear that Covid and India are the most tweeted tweets in India, followed by case, death, test, pandemic, and many more. By knowing the most tweeted tweets, people's concern during the adversity of Covid-19 can be understood easily.

4. Conclusion

Data visualization has become the go-to solution for analyzing and interpreting complex data from various sources. Besides large-scale businesses, data visualization is also popularly used by scientists and research personnel to analyze experimental data, draw insights and help them produce and present desired results of their research more effectively. Knowing about different visualization techniques is important because different kinds of data require different visualization techniques. For instance, Categorical data uses a pie chart; Numerical data uses a scatter plot; geospatial data use choropleth maps. Also, for tasks like Outlier detection, box plots are used, for understanding distribution histograms and density curves are used, for detecting keywords, a word cloud is used. As discussed in the paper, for the Covid-19 data, various visualization techniques gave different insights like timeline chart described the series of decisions taken by the Government of India, the bar graph represented the top 10 states that contributed the most to COVID-19 related deaths in India and the world cloud clearly depicted the concern of people during the pandemic. So it can be concluded that there is a need to use various visualization techniques as the data being generated has become very diverse and to get a variety of insights from the data.

REFERENCES

[1] T. Marcoux, N. Agarwal, R. Erol, A. Obadimu, and M. N. Hussain, "Analyzing Cyber Influence Campaigns on YouTube using YouTubeTracker," in Big Data and Social Media Analytics, Springer, 2021, pp. 101–111.

[2] S. Papadimitriou and J. Sun, "Disco: Distributed co-clustering with map-reduce: A case study towards petabyte-scale end-to-end mining," 2008, pp. 512–521.

[3] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton, "Big data: the management revolution," Harv. Bus. Rev., vol. 90, no. 10, pp. 60–68, 2012.

[4] SJB Institute of Technology, H. B P, A. R, A. shek S, and R. Vibhu C, "Agricultural Data Visualization for Prescriptive Crop Planning," Int. J. Comput. Trends Technol., vol. 49, no. 3, pp. 183–188, Jul. 2017, doi: 10.14445/22312803/IJCTT-V49P129.

[5] C. N. Knaflic, Storytelling with data: A data visualization guide for business professionals. John Wiley & Sons, 2015.

[6] D. Mashima, S. Kobourov, and Y. Hu, "Visualizing dynamic data with maps," IEEE Trans. Vis. Comput. Graph., vol. 18, no. 9, pp. 1424–1437, 2011.

[7] J. Lankow, J. Ritchie, and R. Crooks, Infographics: The power of visual storytelling. John Wiley & Sons, 2012.

[8] S. Few, Information dashboard design: The effective visual communication of data, vol. 2. O'reilly Sebastopol, CA, 2006.

[9] R. Donnelly and W. M. Kelley, The Humongous Book of Statistics Problems: Nearly 900 Statistics Problems with Comprehensive Solutions for All the Major Topics of Statistics. Penguin, 2009.

[10] D. F. Williamson, "The Box Plot: A Simple Visual Method to Interpret Data," Ann. Intern. Med., vol. 110, no. 11, p. 916, Jun. 1989, doi: 10.7326/0003-4819-110-11-916.

[11] T. Onorati, P. Díaz, T. Zarraonandia, and I. Aedo, "The Immersive Bubble Chart: a Semantic and Virtual Reality Visualization for Big Data," 2018, pp. 176–178.

[12] J. Stewart and P. J. Kennelly, "Illuminated Choropleth Maps," Ann. Assoc. Am. Geogr., vol. 100, no. 3, pp. 513–534, Jun. 2010, doi: 10.1080/00045608.2010.485449.

[13] J. L. Hintze and R. D. Nelson, "Violin Plots: A Box Plot-Density Trace Synergism," Am. Stat., vol. 52, no. 2, p. 181, May 1998, doi: 10.2307/2685478.

[14] JS. Zhao, Y. Guo, Q. Sheng, and Y. Shyr, "Advanced Heat Map and Clustering Analysis Using Heatmap3," BioMed Res. Int., vol. 2014, pp. 1–6, 2014, doi: 10.1155/2014/986048.

[15] R. L. Nuzzo, "Histograms: A Useful Data Analysis Visualization," PM&R, vol. 11, no. 3, pp. 309–312, Mar. 2019, doi: 10.1002/pmrj.12145.

[16] R. L. Nuzzo, "Histograms: A Useful Data Analysis Visualization," PM&R, vol. 11, no. 3, pp. 309–312, Mar. 2019, doi: 10.1002/pmrj.12145.

[17] M. Khan and S. S. Khan, "Data and information visualization methods, and interactive mechanisms: A survey," Int. J. Comput. Appl., vol. 34, no. 1, pp. 1–14, 2011.

[18] M. Diani, "Network analysis," Methods Soc. Mov. Res., pp. 173–200, 2002.

[19] A. Inselberg, "The plane with parallel coordinates," Vis. Comput., vol. 1, no. 2, pp. 69–91, 1985.

[20] J. Johansson and C. Forsell, "Evaluation of Parallel Coordinates: Overview, Categorization and Guidelines for Future Research," IEEE Trans. Vis. Comput. Graph., vol. 22, no. 1, pp. 579–588, Jan. 2016, doi: 10.1109/TVCG.2015.2466992.

[21] R. C. Roberts, R. S. Laramee, G. A. Smith, P. Brookes, and T. D'Cruze, "Smart brushing for parallel coordinates," IEEE Trans. Vis. Comput. Graph., vol. 25, no. 3, pp. 1575–1590, 2018.

[22] R. Hu, C. Han, S. Pei, M. Yin, and X. Chen, "Procalcitonin levels in COVID-19 patients," Int. J. Antimicrob. Agents, vol. 56, no. 2, p. 106051, Aug. 2020, doi: 10.1016/j.ijantimicag.2020.106051.

[23] I. Spence, "No humble pie: The origins and usage of a statistical chart," J. Educ. Behav. Stat., vol. 30, no. 4, pp. 353–368, 2005.

[24] A. Sarikaya and M. Gleicher, "Scatterplots: Tasks, Data, and Designs," IEEE Trans. Vis. Comput. Graph., vol. 24, no. 1, pp. 402–412, Jan. 2018, doi: 10.1109/TVCG.2017.2744184.

[25] Q. Cui, M. O. Ward, and E. A. Rundensteiner, "Enhancing scatterplot matrices for data with ordering or spatial attributes," San Jose, CA, Jan. 2006, p. 60600R. doi: 10.1117/12.650409.

[26] P. H. Nguyen, K. Xu, R. Walker, and B. W. Wong, "TimeSets: Timeline visualization with set relations," Inf. Vis., vol. 15, no. 3, pp. 253–269, Jul. 2016, doi: 10.1177/1473871615605347.

[27] D. Gerbing, "Time series components," Portland State Univ., p. 9, 2016.

[28] L. K. Long, L. C. Hui, G. Y. Fook, and W. M. N. Wan Zainon, "A Study on the Effectiveness of Tree-Maps as Tree Visualization Techniques," Procedia Comput. Sci., vol. 124, pp. 108–115, 2017, doi: 10.1016/j.procs.2017.12.136.

[29] S. Lohmann, F. Heimerl, F. Bopp, M. Burch, and T. Ertl, "Concentri Cloud: Word Cloud Visualization for Multiple Text Documents," in 2015 19th International Conference on Information Visualisation, Barcelona, Spain, Jul. 2015, pp. 114–120. doi: 10.1109/iV.2015.30.

[30] M. Friendly, "A Brief History of Data Visualization," in Handbook of Data Visualization, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 15–56. doi: 10.1007/978-3-540-33037-0_2.

[31] M. Milano and M. Cannataro, "Statistical and Network-Based Analysis of Italian COVID-19 Data: Communities Detection and Temporal Evolution," Int. J. Environ. Res. Public. Health, vol. 17, no. 12, p. 4182, Jun. 2020, doi: 10.3390/ijerph17124182.

[32] S. Saraswathi, A. Mukhopadhyay, H. Shah, and T. S. Ranganath, "Social network analysis of COVID-19 transmission in Karnataka, India," Epidemiol. Infect., vol. 148, p. e230, 2020, doi: 10.1017/S095026882000223X.

[33] Q. Zaldo-Aubanell et al., "Community Risk Factors in the COVID-19 Incidence and Mortality in Catalonia (Spain). A Population-Based Study," Int. J. Environ. Res. Public. Health, vol. 18, no. 7, p. 3768, Apr. 2021, doi: 10.3390/ijerph18073768.

[34] R. Nigam, K. Pandya, A. J. Luis, R. Sengupta, and M. Kotha, "Positive effects of COVID-19 lockdown on air quality of industrial cities (Ankleshwar and Vapi) of Western India," Sci. Rep., vol. 11, no. 1, p. 4285, Dec. 2021, doi: 10.1038/s41598-021-83