

○ PROJECT REPORT ON

- Project Name:- **Data Analysis And Visualization**

(UNDER THE PARTIAL FULFILMENT OF B.Sc. DEGREE COURSE IN  
COMPUTER SCIENCE )

SUBMITTED BY: Mr Nikhil Dashrath Gopale

➤ Guided by **Mrs Deepa Mam**

- DEPARTMENT OF COMPUTER SCIENCE

V.K.K MENON COLLEGE OF COMMERCEAND ECONOMICS & SHARAD  
SHANKAR DIGH COLLEGE OF SCIENCE BHANDUP (EAST), MUMBAI-  
400042

UNIVERSITY OF MUMBAI

➤ 2020-2021

▪ **CERTIFICATE**

➤ **V.K KRISHNA MENON COLLEGE OF COMMERCE AND ECONOMICS  
AND SHARAD SHANKAR DIGHE COLLEGE OF SCIENCE BHANDUP  
EAST, MUMBAI - 400042**

**This is to certify that      Mr. Nikhil Dashrath Gopale**

**Seat No   10   has successfully completed the PROJECT titled**

**“Data Analysis And Visualization ”.**

**For partial fulfillment of Bachelor of Computer Science (SEM-VI) of University  
of Mumbai in academic year 2020 -2021 under the guidance of Mrs Deepa mam**

**.**

**DATE:**

**HEAD OF THE DEPARTMENT**

**PROJECT GUIDE**

**EXAMINER**

-----\*\*\*-----

## INDEX

<b>1</b>	<b>Acknowledgment</b>
<b>2</b>	<b>Overview</b>
2.1	Abstract
2.2	Undertaking
2.3	Objective
2.4	Introduction
2.5	Advantages Of Data Visualization :-
2.6	Requirements and Specification
2.7	Feasibility study
<b>3 .1</b>	<b>System Design</b>
3.2	Activity Diagram
<b>4</b>	<b>Code implementation</b>
<b>5</b>	<b>Results</b>
<b>6</b>	<b>Future scope</b>
<b>7</b>	<b>Reference</b>

➤ **Acknowledgement**

- It gives great pleasure and pride as I present my project on “**Data Analysis And Visualization**”.
- This acknowledgement is a small effort to Express our gratitude to all those who have shown me the path to bring out the various colours of my project With their vast treasure of experience and knowledge.
- I would like to express my sincere thanks to Prof. Mrs.Deepa Mam, who helps me throughout the project for providing me moral support, conducive Work environment and the much needed inspiration to conclude this project on time.
  - I also take this opportunity thank head of the Department Dr Saloni Bhushan and thanks to all of our professions of the Department of Computer Science of V.K. Krishna Menon College for giving me an opportunity to study in the institute and the most needed guidance throughout the duration of the course.
- Prof Mrs.Deepa Mam has provided me with the guidance and necessary support during each phase of the project. He was the Source of continuous encouragement as each milestones was crossed.
  - Last but not the least,I would like to thank my friends and family for the support and encouragement they given me during this course of work.

➤ **Abstract :-**

- The Aim of the project is to provide Data analysis And Visualization of covid-19 (a pandemic started in March 2021). Through plotting of data, various cases have been studied like most affected countries due to this pandemic.

Study of data from various Districts is combined to show the growth of cases and recovery graph. In this project, the predictions on various cases has been done . Comparison graphs has also been plotted to analyse how much Maharashtra is getting affected/recover day by day.

➤ **Undertaking**

- I hereby undertake that this project is made entirely by me with the help of various websites, videos, tutorials and help, suggestions and guidance from my professors, friends and family. I will make sure that this application will be useful and inspiring for other students in each and every way possible.
- This application is made solely for my Final Year Computer Science Project. I will ensure that this program or application will not be misused in any way or I will be responsible for that and no one else.

➤ **Objective**

Data Analysis is an exploratory process of inspecting, cleansing, transforming and modelling data with a specific goal of finding some useful information and usually begins with specific questions.

Data Visualisation involves a visual representation of the data that can enable people with little knowledge regarding the data to understand the information. It reduces the amount of time required by them to process the information and provide valuable insights. Both the processes work simultaneously with the analysis of data taking place before the visual output is produced.

## ➤ Introduction

- ✓ **Data Analysis** is the process of bringing order and structure to collected data. It turns data into information teams can use. Analysis is done using systematic methods to look for trends, groupings, or other relationships between different types of data.
- ✓ **Data visualization** is the process of converting raw data into easily understandable pictorial representation, that enables fast and effective decisions.  
Data visualization is a strategy where we represent the quantitative information in a graphical form.
  - ✓ **Data visualization** is the process of putting data into a chart, graph, or other visual format that helps inform analysis and interpretation. Data visuals present the analyzed data in ways that are accessible to and engage different stakeholders. Data visuals are also used to communicate MEAL results to meet key stakeholder needs. Multiple visuals will likely be needed to understand the larger change process and inform data use.

## ➤ Advantages Of Data Visualization :-

**Some of the main reasons for using data visualization are:**

- to explore sources
- to tell stories
- to predict sales volumes
- to identify areas that need attention or improvement
- to understand what factors influence customers' behavior
- to know which products to place where
- to discover how to increase revenues or reduce expenses
- spreadsheets are hard to visualize
- patterns and trends can be spotted quickly and easily
- saves time and energy
- The main advantages of communicating and analyzing information through visual means are that it is faster for people to grasp meaning, the data is easily interpreted, it is easier for decision makers to see things that were not obvious, and it is simple to share ideas.



➤ **Requirements and Specification**

- ✓ The data required for analysis is based on a question or an experiment. Based on the requirements of those directing the analysis, the data necessary as inputs to the analysis is identified (e.g., Population of people). Specific variables regarding a population (e.g., cases and Deaths) may be specified and obtained. Data may be numerical or categorical.

➤ **Software And Programming Language:**

- **Python programming Language**
- **Jupyter Notebook**
- **MS-Excel**
- **covid-19 Data set**

➤ **Hardware Requirements:**

- ✓ **Laptop/Desktop:-** 64 Bit operating System
- ✓ **Android Phone**

➤ **Feasibility Study:**

✓ **Economic Feasibility:-**

The Developed System is time effective because, you Do not have to enter Data Or any additional information. It is also cost effective because there is no use of paperwork .we can easily Analyse And Visualize the Data and easy to make Decisions.

✓ **Technical Feasibility :-**

The system is economic and it's require only limited software. Such as Databases(MySQL) , programming software(python , R), Visualization software( Tableau ,Power BI) and MS-Excel.

And This software we can easily access. With the help of Data Visualization we can make good decisions in less time.

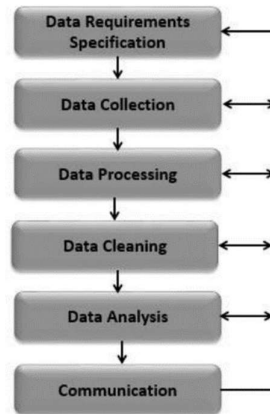
✓ **Market Feasibility :-**

The marketing team of a business performs market feasibility analysis to analyze the demand and requirement of the business product or service and information. Using market research, you can make better decisions for your company as well as for the future or for people..

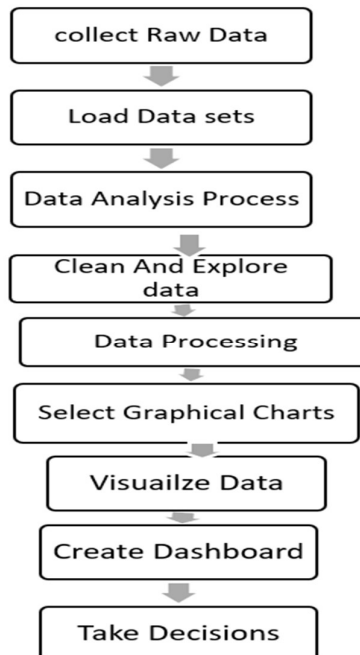
➤ Activity Diagram :-

System Diagram / worl Flow Activity :-

1)\_ Data Analysis Work Flow Activity Process :-



2)\_ Data Visualization Work Flow Activity :-



➤ **Code Implementation :-**

➤ **First install or import python libraries / modules in CMD**

**1. Numpy :-**

NumPy stands for 'Numerical Python' or 'Numeric Python'. It is an open source module of Python which provides fast mathematical computation on arrays and matrices

✓ **Pip install Numpy**

**2. Pandas :-**

Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools. Unlike NumPy library which provides objects for multi-dimensional arrays, Pandas provides in-memory 2d table object called Dataframe. It is like a spreadsheet with column names and row labels.

✓ **Pip install pandas**

**3. Matplotlib :-**

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library

✓ **Pip install matplotlib**

**4. Plotly :-**

Plotly is a Montreal based technical computing company involved in development of data analytics and visualisation tools such as **Dash** and **Chart Studio**.

Code :-

# Data Analysis And Visualization with Python

## Data Analysis process

→ from IPython.display import Image  
Image(r"C:\Users\acer\Desktop\data analysis process.png",width=300,height=200)

```
In [1]: from IPython.display import Image  
        Image(r"C:\Users\acer\Desktop\data analysis process.png",width=300,height=200)  
Out[1]:
```



```
# for data Analysis import libeary  
import pandas as pd  
import numpy as np
```

```
# for Data visualization import libeary  
from matplotlib import pyplot as plt  
import plotly.express as px
```

## Load Data\_CSV File

corona Cases 2/4/2021

→ data=pd.read\_csv("districts cases 2\_4\_2021.csv")  
data.head()

```
In [3]: data=pd.read_csv("districts cases 2_4_2021.csv")
data.head()
```

Out[3]:

	Districts	positive_cases	active_cases	recoverd	deaths	recovery_rate(%)	deaths_rate(%)
0	Ahmednagar	94044	9103	83728.0	1212	89.0	1.3
1	Akola	28368	4050	23850.0	464	84.1	1.6
2	Amravati	49079	2935	45494.0	648	92.7	1.3
3	Aurangabad	83917	19466	63095.0	1342	75.2	1.6
4	Beed	25985	4139	21223.0	614	81.7	2.4

## Check Shape of data

→ data.shape  
(36, 7)

## Find null Values

→ data.isnull().sum()

```
In [5]: data.isnull().sum()
```

```
Out[5]: Districts          0
positive_cases          0
active_cases           0
recoverd                1
deaths                 0
recovery_rate(%)       1
deaths_rate(%)         0
dtype: int64
```

## # find exact location of null value

→ null=data.columns[data.isnull().any()]  
null  
→ print(data[data['recoverd'].isnull()][null])

```
In [6]: # find exact location of null value

null=data.columns[data.isnull().any()]
null

print(data[data['recoverd'].isnull()][null])

    recoverd  recovery_rate(%)
22      NaN                NaN
```

→ data.iloc[22] # iloc-->access specific row

```
In [7]: data.iloc[22] # iloc-->access specific row

Out[7]: Districts          Other States/Country
positive_cases          146
active_cases           48
recoverd              NaN
deaths               96
recovery_rate(%)      NaN
deaths_rate(%)       65.8
Name: 22, dtype: object
```

## Remove null values

→ `remove_null=data.replace(np.nan,0).sum()`  
`remove_null.isnull().sum()`

```
In [8]: remove_null=data.replace(np.nan,0).sum()
```

```
In [9]: remove_null.isnull().sum()
```

```
Out[9]: 0
```

## Drop Unnecessary data

# Drop Other States/Country row

# Drop using Dataframe index

→ `clean_data=data.drop([22])`  
`clean_data.head(25)`

```
In [10]: # Drop Other States/Country row
```

```
# Drop using Dataframe index  
clean_data=data.drop([22])  
clean_data.head(25)
```

12	Jalgaon	85443	6464	77358.0	1596	90.5	1.9
13	Jalna	23006	675	21920.0	410	95.3	1.8
14	Kolhapur	51301	731	48873.0	1694	95.3	3.3
15	Latur	33226	5545	26929.0	748	81.0	2.3
16	Mumbai	414773	49953	352173.0	11690	84.9	2.8
17	Nagpur	230187	46333	179950.0	3859	78.2	1.7
18	Nanded	43245	12268	30199.0	772	69.8	1.8
19	Nandurbar	18032	4525	13215.0	291	73.3	1.6
20	Nashik	178997	33442	143332.0	2222	80.1	1.2
21	Osmanabad	21395	2327	18464.0	587	86.3	2.7
23	Palghar	54072	3377	49711.0	974	91.9	1.8
24	Parbhani	14152	4560	9227.0	354	65.2	2.5
25	Pune	536262	64277	463611.0	8325	86.5	1.6

## find duplicates

→ `clean_data.duplicated().sum()`

```
In [11]: clean_data.duplicated().sum()
```

```
Out[11]: 0
```

## Analyzing Data

### Total Positive And Active Corona Cases

→ `positive_total=clean_data.positive_cases.sum()`

positive\_total

→ active\_total=clean\_data.active\_cases.sum()  
active\_total

→ display(positive\_total,active\_total)

```
In [12]: positive_total=clean_data.positive_cases.sum()
         positive_total

         active_total=clean_data.active_cases.sum()
         active_total

         display(positive_total,active_total)

2812834
356195
```

→ print("The number of Positive\_cases in Maharashtra is {} And Number of  
Active\_cases is {}".format(positive\_total,active\_total))

```
In [13]: print("The number of Positive_cases in Maharashtra is {} And Number of Active_cases is {}".format(positive_total,active_total))

The number of Positive_cases in Maharashtra is 2812834 And Number of Active_cases is 356195
```

### **Recovery Rate in Maharashtra**

→ recoverd\_cases=clean\_data.recoverd.sum()  
→ recovery\_rate= recoverd\_cases / positive\_total \* 100  
recovery\_rate  
→ display(recoverd\_cases,recovery\_rate)

```
In [14]: recoverd_cases=clean_data.recoverd.sum()

         recovery_rate= recoverd_cases / positive_total * 100
         recovery_rate

         display(recoverd_cases,recovery_rate)

2400727.0
85.34904654878318
```

### **Death Rate in Maharashtra**

→ death\_cases = clean\_data.deaths.sum()  
death\_cases  
→ death\_rate = death\_cases / positive\_total \* 100  
death\_rate  
→ display(death\_cases,death\_rate)



```
In [16]: death_cases=clean_data.deaths.sum()
death_cases

death_rate= death_cases / positive_total * 100
death_rate

display(death_cases,death_rate)

54553

1.9394319039090113
```

→ print("The Death\_Rate in Maharashtra is {:.1f}%".format(death\_rate))

```
In [17]: print("The Death_Rate in Maharashtra is {:.1f}%".format(death_rate))

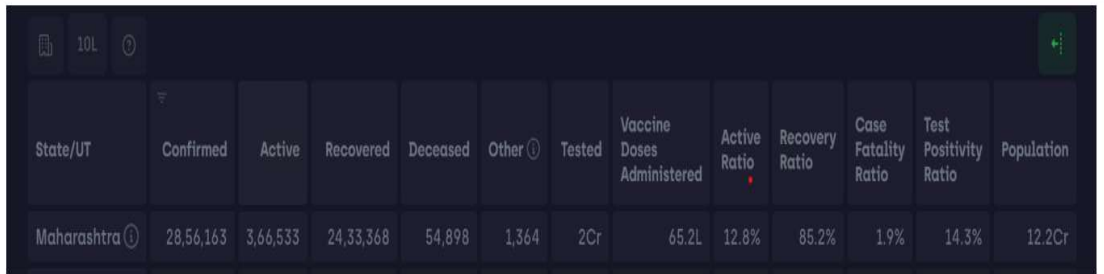
The Death_Rate in Maharashtra is 1.9%
```

## Cross Check Data

### The Data Given below is dated 2/04/2021

```
In [18]: Image(r"C:\Users\acer\Desktop\cross check data.png",width=900,height=500)
```

Out[18]:



State/UT	Confirmed	Active	Recovered	Deceased	Other ⓘ	Tested	Vaccine Doses Administered	Active Ratio	Recovery Ratio	Case Fatality Ratio	Test Positivity Ratio	Population
Maharashtra ⓘ	28,56,163	3,66,533	24,33,368	54,898	1,364	2Cr	65.2L	12.8%	85.2%	1.9%	14.3%	12.2Cr

## Data Visualization

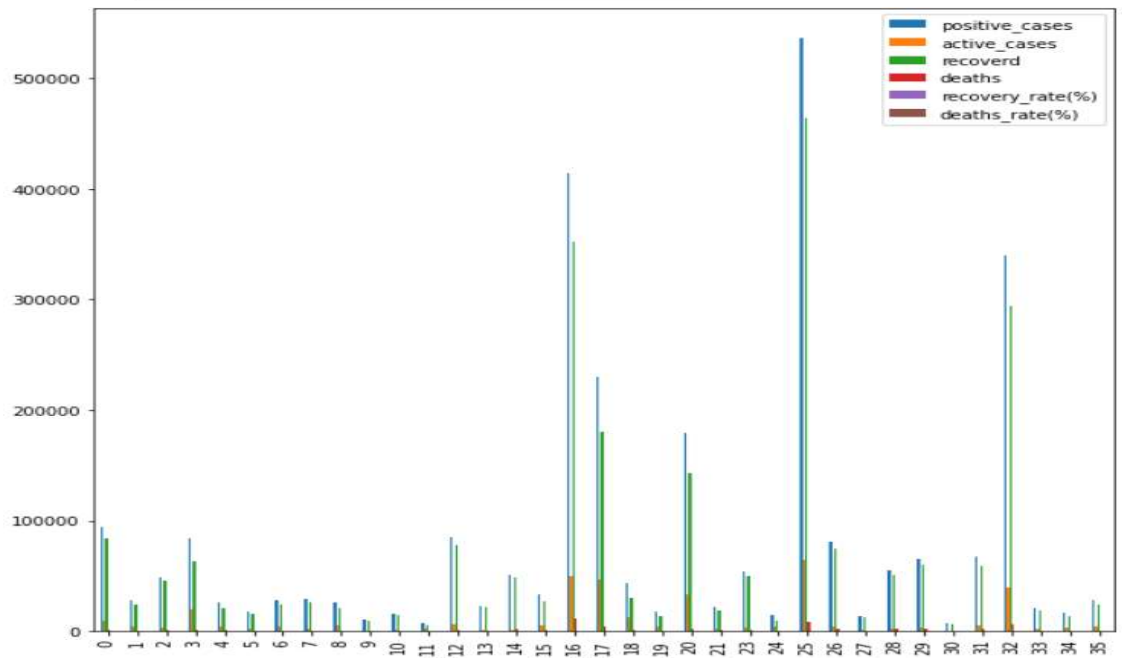
### 1) Bar Graph :-

- ✓ Bar graphs are best used when we need to compare the quantity of categorical values within the same category.
- ✓ Bar graphs should not be used for continuous values.

→

```
fig_dims=(10,10)
fig,ax=plt.subplots(figsize=fig_dims)
name=['Districts']
clean_data.plot.bar(ax=ax)
plt.show()
```

```
In [40]: fig_dims=(10,10)
fig,ax=plt.subplots(figsize=fig_dims)
name=['Districts']
clean_data.plot.bar(ax=ax)
plt.show()
```



### Analyze data by Groupby values

→

```
group_positive=clean_data.groupby('Districts')['positive_cases'].sum()
group_positive
```

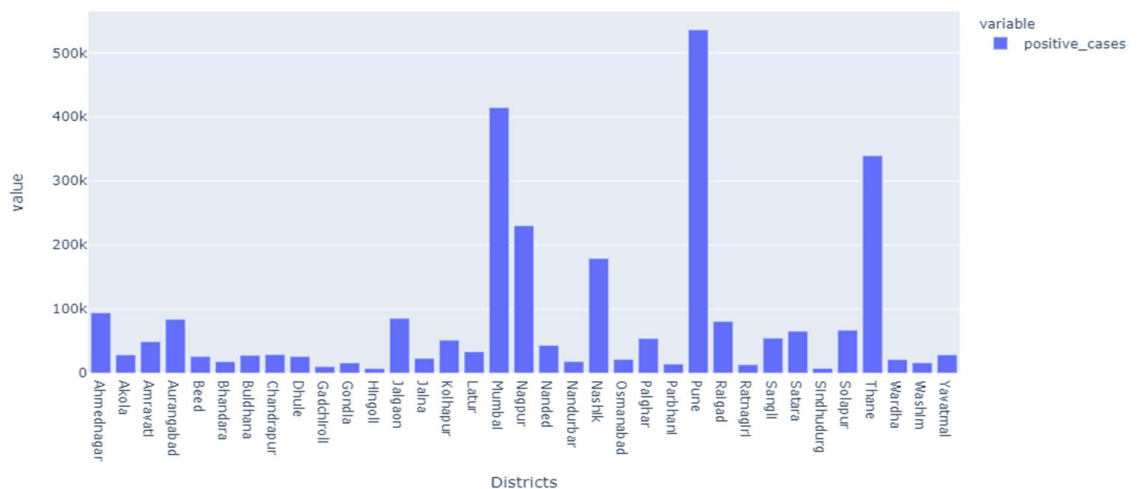
```
In [41]: group_positive=clean_data.groupby('Districts')['positive_cases'].sum()
group_positive.head()
```

```
Out[41]: Districts
Ahmednagar      94044
Akola           28368
Amravati        49079
Aurangabad      83917
Beed            25985
Name: positive_cases, dtype: int64
```

## positive cases Bar Graph

→ px.bar(data\_frame=group\_positive,orientation='v',title='positive cases in Maharashtra')

```
In [42]: px.bar(data_frame=group_positive,orientation='v',title='positive cases in Maharashtra')
```



## Top 5 Districts with highest number of Deaths cases

→ top\_deaths=clean\_data[clean\_data['deaths']>2000]  
top\_deaths

```
In [23]: top_deaths=clean_data[clean_data['deaths']>2000]
top_deaths
```

```
Out[23]:
```

	Districts	positive_cases	active_cases	recovered	deaths	recovery_rate(%)	deaths_rate(%)
16	Mumbai	414773	49953	352173.0	11690	84.9	2.8
17	Nagpur	230187	46333	179950.0	3859	78.2	1.7
20	Nashik	178997	33442	143332.0	2222	80.1	1.2
25	Pune	536262	64277	463611.0	8325	86.5	1.6
32	Thane	339590	39692	293897.0	5970	86.5	1.8

→ `top_5=top_deaths.groupby('Districts')['deaths'].sum()`  
`top_5`

```
In [24]: top_5=top_deaths.groupby('Districts')['deaths'].sum()  
top_5  
Out[24]: Districts  
Mumbai      11690  
Nagpur       3859  
Nashik       2222  
Pune         8325  
Thane        5970  
Name: deaths, dtype: int64
```

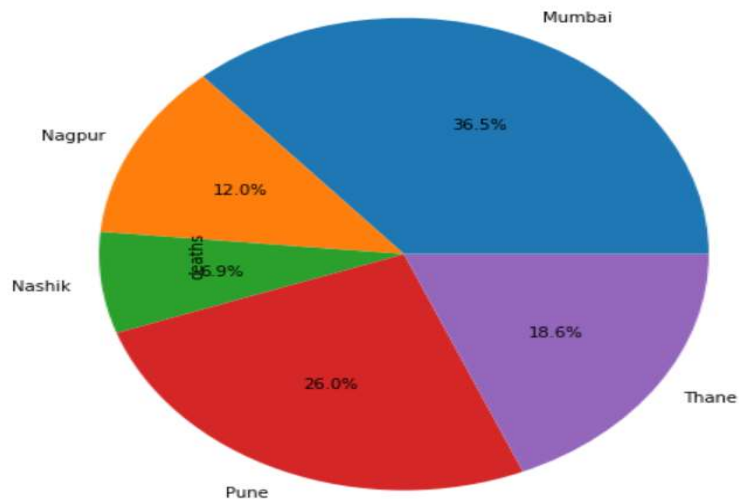
### pie chart of Top 5 Districts with highest number of death cases

#### 2) Pie Chart :-

- ✓ A pie chart is suitable to show the proportional distribution of items within the same category.
- ✓ A pie chart is rendered useless when there are a lot of items within a category. This will decrease the size of each slice and there will be no distinction between the items.

→ `top_5.plot.pie(autopct='%0.1f%%',radius=2,subplots=True)`  
`plt.show()`

```
In [25]: top_5.plot.pie(autopct='%0.1f%%',radius=2,subplots=True)  
plt.show()
```



## Line Graph of Active Cases in Maharashtra

### 3) Line Chart :-

- ✓ A line plot is useful for visualizing the trend in a numerical value over a continuous time interval.

✓

- Using pivot table

→ `top_active=pd.pivot_table(clean_data,values='active_cases',index='Districts')`  
`top_active`

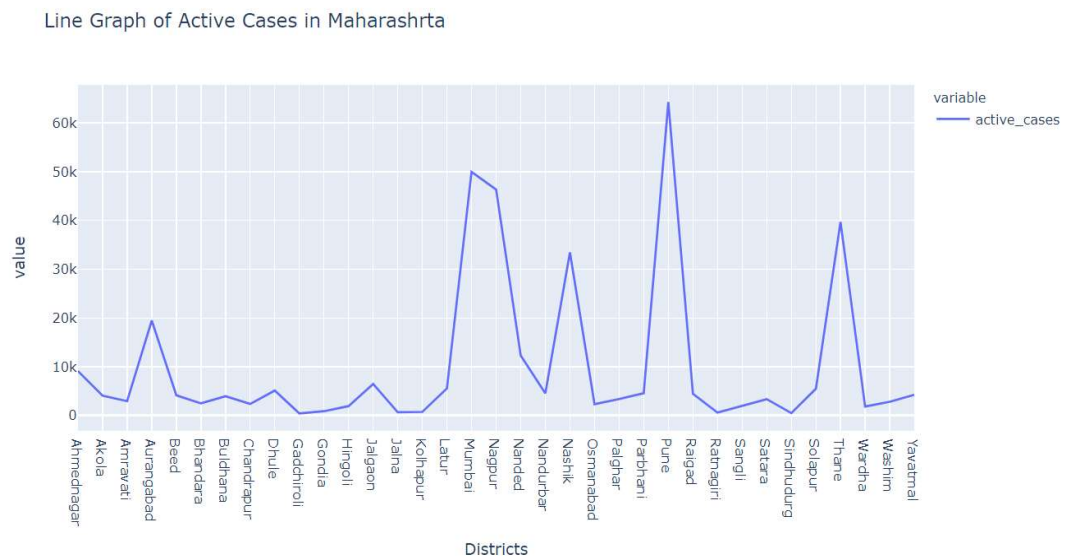
```
In [43]: top_active=pd.pivot_table(clean_data,values='active_cases',index='Districts')
top_active.head()
```

Out[43]:

active_cases	
Districts	
Ahmednagar	9103
Akola	4050
Amravati	2935
Aurangabad	19466
Beed	4139

→ `px.line(data_frame=top_active,orientation='v',title='Line Graph of Active Cases in Maharashtra')`

```
In [28]: px.line(data_frame=top_active,orientation='v',title='Line Graph of Active Cases in Maharashtra')
```



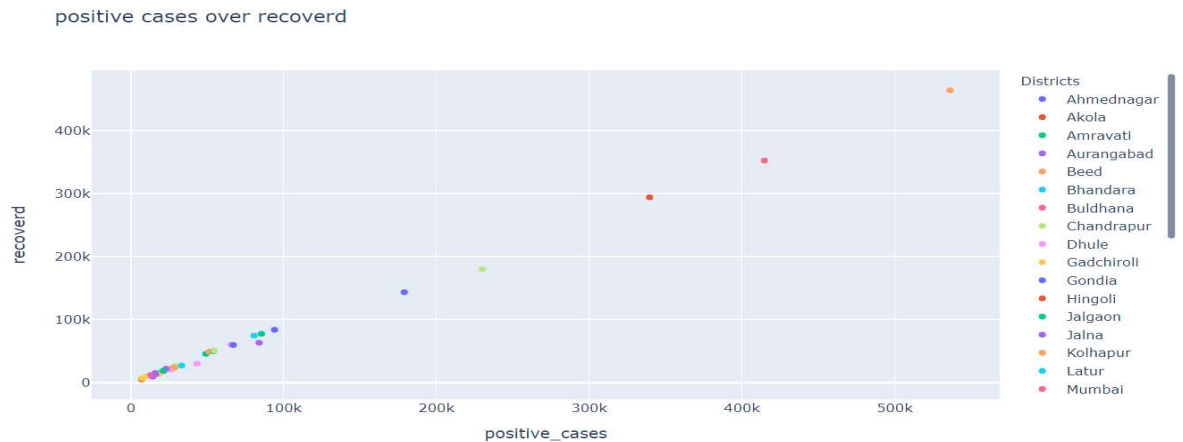
## Scatter plot of positive cases over recoverd

### 4) Scatter plot Chart :-

- ✓ Scatter plots are useful for showing the relationship between two variables. Any correlation between variables or outliers in the data can be easily spotted using scatter plots.

→ `fig=px.scatter(clean_data,x='positive_cases',y='recoverd',color='Districts',orientation='v',title='positive cases over recoverd')`  
`fig.show()`

```
In [29]: fig=px.scatter(clean_data,x='positive_cases',y='recoverd',color='Districts',orientation='v',  
                    title='positive cases over recoverd')  
fig.show()
```



## scatter plot of deaths rate(%) And recovery rate(%)

→ `rate=clean_data[['Districts','deaths_rate(%)','recovery_rate(%)']]`  
`rate`

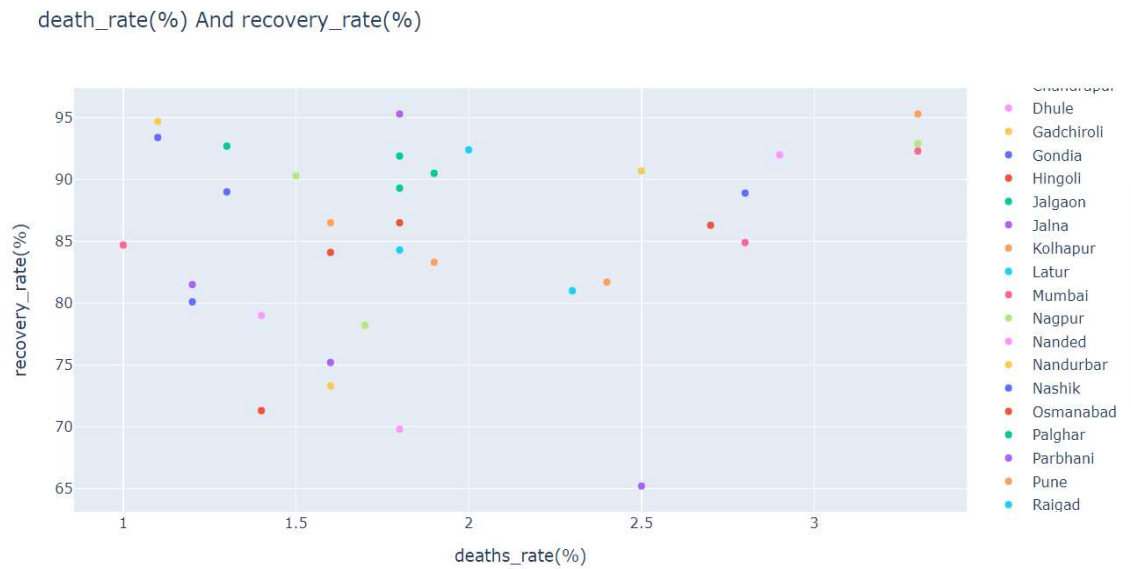
```
In [30]: rate=clean_data[['Districts','deaths_rate(%)','recovery_rate(%)']]  
rate
```

Out[30]:

	Districts	deaths_rate(%)	recovery_rate(%)
0	Ahmednagar	1.3	89.0
1	Akola	1.6	84.1
2	Amravati	1.3	92.7
3	Aurangabad	1.6	75.2
4	Beed	2.4	81.7
5	Bhandara	1.8	84.3
6	Buldhana	1.0	84.7
7	Chandrapur	1.5	90.3
8	Dhule	1.4	79.0
9	Gadchiroli	1.1	94.7
10	Gondia	1.1	93.4

→ fig=px.scatter(rate,x='deaths\_rate(%)',y='recovery\_rate(%)',color='Districts',orientation='v',title='death\_rate(%) And recovery\_rate(%)')  
fig.show()

```
In [31]: fig=px.scatter(rate,x='deaths_rate(%)',y='recovery_rate(%)',color='Districts',orientation='v',
          title='death_rate(%) And recovery_rate(%)')
fig.show()
```



➤ **Benefits that data visualization can bring to your organization's daily life:**

- **Optimizes decision-making:** By simplifying the understanding of information, decision making is made faster, cheaper and with much more foundation. Since the company's KPIs are constantly monitored and analyzed. It is possible to create a more natural decision flow.
- **Identifies trends:** Real-time data allows for intelligent and more reliable reading, enabling analysis of past, current and future trends in the market and industry. In this way, the company obtains competitive advantages and differentiates itself in the market.
- **It promotes insights:** Through the creative exploration of data, decision-makers, as well as the team, who have easy access to reports, now have the ability to tell stories through the insights hidden in the numbers. Visually arranged data allows individuals to see patterns and gain insights more quickly.
- **Reduces the possibility of errors:** The use of a centralized control panel minimizes communication failures and accelerates data consolidation, ensuring that important decisions are based on reliable and up-to-date information.

➤ **Future Scope**

- Data Storing Into Database With a Clear Purpose.
- The trend is clearly for new tools to provide more data to its users, not less. They will be easier to use and therefore widen in reach, as we've seen with the rising interest in Tableau and Power BI.
- Location Based Visualization.
- Intractive Dashboards.
- The data visualization market receives more investments each year and there are several factors that give impetus for it to grow even more in the coming years.
- Visualizing information has existed for a very long time and it will still exist in the future, of course. And maybe even more than yesterday, because today data is an integral part of the global economic system.



## ➤ **Reference :-**

1. Covid Data set :-
  - ✓ [COVID19-India API | api](#)
  - ✓ [mh-covid-summary \(covid19maharashtrgov.in\)](#)
  - ✓ [India Covid-19 - Coronavirus Tracker India \(Live\) - Dashboard - 12589019 confirmed, 743652 Active, 11680229 Recovered and 165138 deceased in India from Coronavirus aka Covid19 Outbreak \(indiacovid-19.in\)](#)
  - ✓ [Coronavirus Outbreak in India - covid19india.org](#)
2. For Data Analysis And Visualization Book :-
  - ✓ Python for Data Analysis :- Wes Mckinney
  - ✓ Python Data Analytics :- Fabio Nelli
3. Data Visualization Python Tutorial using Matplotlib
  - ✓ [Data Visualization Python Tutorial using Matplotlib \(simplifiedpython.net\)](#)
  - ✓ [Matplotlib | Matplotlib For Data Visualization, Exploration \(analyticsvidhya.com\)](#)
4. My Kaggle And Github Project link :-
  - ✓ Kaggle :- <https://www.kaggle.com/nikhilgopale/data-analysis-and-visualization>
  - ✓ GitHub :- <https://github.com/NikhilGopale3008>