Data Analysis and Visualization on Covid 19 in India state wise

Major Project Report

Submitted to the Centurion University

In partial fulfillment of requirements for the award of the degree

Bachelor of Technology

in

Computer Science and Engineering

By

M Satya Dinesh

211801360005

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CENTURION UNIVERSITY OF TECHNOLOGY AND MANAGEMENT

(CUTM-AP) VIZIANAGARAM

535003

CENTURION UNIVERSITY OF TECHNOLOGY AND MANAGEMENT

<div align="center">

(CUTM-AP) VIZIANAGARAM

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

</div>

<div align="center">

CERTIFICATE

</div>

This is to certify that their report entitled "Data Analysis and Visualization on COVID-19 IN INDIA Statewise" was submitted by M Satya Dinesh(211801360005) in the Department of Computer Science and Engineering in partial fulfillment of the B.Tech. Degree in Computer Science and Engineering is a bonafide record of the seminar work carried out by him under our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

(Project Guide and HOD)   **External Examiner**

Dr. Lakshman Rao

Department of Computer Science

And Engineering

Centurion University (CUTM-AP)

Abstract

This Project is about the updates on covid-19 in India up to 24th March 2022.

No. of confirmed cases, no. of active cases, no of persons who are cured/discharged, and no of deaths in different states of India are included.

This project is done by using python libraries like pandas, NumPy, matplotlib, etc… The main theme of this project is to analyze the data and visualize it using python.

INDEX

LIST OF FIGURES

1. Introduction

In this project, you will learn how to preprocess and merge datasets to calculate needed measures and prepare them for an Analysis. In this project, we are going to work with the COVID19 in India statewide dataset. Python is a versatile language. It is used for general programming and development purposes, and complex tasks like Machine Learning, Data Science, and Data Analytics as well. Data science is the study of data to extract knowledge and insights from the data and apply knowledge and actionable insights. As the stats are increasing every day, the analysis and visualization of data will change. This project tells us how to analyze the data and visualize it through python.

2. Existing and Proposed System

This project is about the updates on covid 19 in India. In the already existing project, there are only just basic graphs and they cannot understand by the user. These graphs are represented using NumPy and I would like to develop better visualization of the previous project. I can make more no effective graphs to understand easily by using Plotly, pandas, and some other python libraries. So, it will be useful for the users to interact with the interface and develop the most interactive and I'm furtherly developing the most interactive and friendly nature with the user interface.

3. System Requirements

3.1 Software Requirements

To read the data and to run and execute the following code the user minimum software requirements are:

1. Windows 8 / Ubuntu 18.04

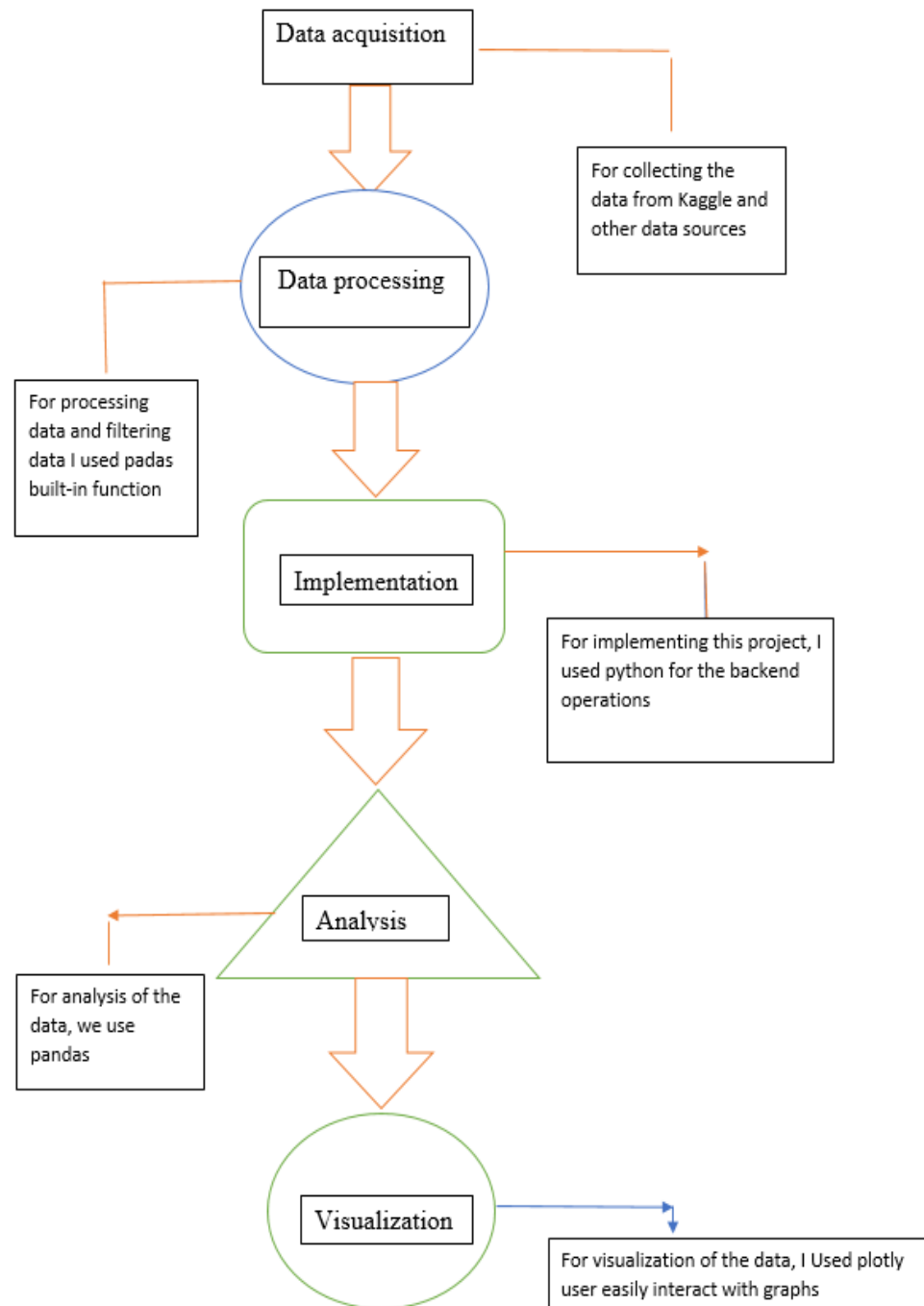2. Python idle / Jupiter notebook

3. MS EXCEL

3.2 Hardware Requirements

To read the data and to run and execute the following code the user's minimum hardware requirements are:

1. Intel core(i3) processor or Ryzen 3 processor

2. RAM: 8 GB

3. Intel iris integrated graphics card

4. Methodology

4.1 Modular Design

Data acquisition

For collecting the data from Kaggle and other data sources

Data processing

For processing data and filtering data I used padas built-in function

Implementation

For implementing this project, I used python for the backend operations

Analysis

For analysis of the data, we use pandas

Visualization

For visualization of the data, I Used plotly user easily interact with graphs

4.1.1 Data Acquisition

For collecting the required data for my project, I visited online sources like Kaggle and other useful websites, etc.

4.1.2 Data Processing

Now the acquisition data is processed in MS Excel for cleaning the garbage and duplicate values. The data is adjusted in a better manner to understand and utilize easily.

4.1.3 Implementation

After processing the collected data, it is imported to python related interpreters (Jupyter notebook) for easy analysis of the data.

4.1.4 Analysis

After the implementation, the processed data has to be analyzed for visualization. In the analysis stage, the user can analyze the data in the form of rows and columns for a clear understanding.

4.1.5 Visualization

After analysis of the data, the final stage is a visualization of data in graphical representation. For the visualization, we used some python built-in libraries like pandas, matplotlib, plot, etc. These modules are required for a better representation of the data.

4.2 What is Python

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. It was created by Guido van Rossum, and released in 1991. Its high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components. Python's simple, easy-to-learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed. Python has a simple syntax similar to the English language. Python has a syntax that allows developers to write programs with fewer lines than some other programming languages. Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc.). Python can be treated procedurally, in an object-oriented way, or in a functional way.

The main applications of python are:

1. Web applications
2. Big data applications
3. Testing
4. Automation
5. Data science, machine learning, and AI
6. Desktop software
7. Mobile applications

The main uses of python language are:

1. Python can be used on a server to create web applications.
2. Python can be used alongside software to create workflows.
3. Python can connect to database systems. It can also read and modify files.

4.3 Modules:

In this project for the data analysis and visualization, the following modules are used.

4.3.1 Numpy: NumPy is a Python library used for working with arrays. It also has functions for working in the domain of linear algebra, Fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open-source project and you can use it freely. NumPy stands for Numerical Python. Using NumPy, mathematical and logical operations on arrays can be performed.

4.3.2 Pandas: Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008. Pandas allow us to analyze big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant. Relevant data is very important in data science.

4.3.3 Matplotlib: Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose toolkits. Matplotlib is a low-level graph plotting library in python that serves as a visualization utility. Matplotlib was created by John D. Hunter. Matplotlib is open source and we can use it freely.

4.3.4 Plotly: Python Plotly library is an open source library that can be used for data visualization and understanding data simply and easily. Plotly supports various types of plots like line charts, scatter plots, histograms, box plots, etc. Plotly has hover tool capabilities that allow us to detect any outliers or anomalies in a large no of data points. It allows us for the endless customization of our graphs that makes our plot more meaningful and understandable for others

5. Data Set

	State/UT	Confirmed Cases	Active Cases	Cured/Discharged	Death
0	Andaman and Nicobar Islands	10031	1	9901	129
1	Andhra Pradesh	2319328	450	2304148	14730
2	Arunachal Pradesh	64484	2	64186	296
3	Assam	724193	1364	716190	6639
4	Bihar	830435	35	818144	12256
5	Chandigarh	91894	21	90708	1165
6	Chhattisgarh	1151995	132	1137829	14034
7	Dadra and Nagar Haveli and Daman and Diu	11441	0	11437	4
8	Delhi	1864135	490	1837497	26148
9	Goa	245265	50	241385	3830
10	Gujarat	1223806	299	1212565	10942
11	Haryana	984788	390	973798	10600
12	Himachal Pradesh	284412	205	280075	4132
13	Jammu and Kashmir	453595	112	448733	4750
14	Jharkhand	435069	70	429684	5315
15	Karnataka	3944970	1844	3903084	40042
16	Kerala	6529049	5918	6455655	67476
17	Ladakh	28215	26	27961	228
18	Lakshadweep	11402	0	11350	52
19	Madhya Pradesh	1040940	129	1030077	10734
20	Maharashtra	7872817	5089	7723959	143769
21	Manipur	137057	78	134859	2120
22	Meghalaya	93738	50	92096	1592
23	Mizoram	223343	1412	221253	678
24	Nagaland	35467	32	34677	758
25	Odisha	1287285	492	1277677	9116
26	Puducherry	165773	10	163801	1962
27	Punjab	758981	147	741096	17738
28	Rajasthan	1282745	321	1272873	9551
29	Sikkim	39131	18	38662	451
30	Tamil Nadu	3452534	505	3414004	38025
31	Telangana	790989	637	786241	4111
32	Tripura	100871	1	99951	919
33	Uttar Pradesh	2070441	700	2046249	23492
34	Uttarakhand	437092	592	428809	7691
35	West Bengal	2016976	805	1994974	21197

df.shape

The data set used in this project is about updates on covid-19 in India up to 24th March 2022.

The No. of confirmed cases, no. of active cases, no of persons who are cured/discharged, and no of deaths in different states of India are included. I collected this data set from the most useful website KAGGLE. I used this data set for the analysis of players' stats and visualized it in the form of graphs like bars, pie charts, histograms, etc.

6. Implementation

Code:

```
#importing required python libraries

import NumPy as np

import pandas as pd

import matplotlib.pyplot as plt

import plotly. express as ps

import os

os.getcwd()
'C:\\Users\\REVALLA LOKESH'
 #reading the CSV file which contains data

df=pd.read_csv("COVID_19 INDIA - Sheet1.csv")

df.head(10)

 df.info()

#representing the data in a bar graph

x=df.head(10)["State/UT"]

y=df.head(10)["Confirmed Cases"]

plt.bar(x,y,width=0.6,color='g')

font1={"family":"serif","color":"b","size":"20"}
```

```python
font2={"family":"serif","color":"r","size":"15"}

plt.xlabel("State/UT",fontdict=font2)

plt.ylabel("Confirmed Cases",fontdict=font2)

plt.xticks(rotation=90)

plt.title("covid 19",fontdict=font1)

plt.legend(y)

plt.show()

 #displaying the data in 3D graph representation

fig=ps.scatter_3d(df.head(20),"State/UT","Confirmed Cases","Death")

fig.show()

#representation of data in the form of a pie chart

x=df.head(10)["State/UT"]

y=df.head(10)["Confirmed Cases"]

myexplode=[0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1]

plt.pie(df.head(10)["ConfirmedCases"],labels=x,explode=myexplode,shado
w=True)

plt.title("covid 19",color='blue',size=20,loc='left')

plt.legend(y)

plt.show()
#visualization of data as a  grid plot  graph
```

```python
x=df.head(10)["State/UT"]

y=df.head(10)["Confirmed Cases"]

plt.plot(x,y,'r',label='line one',linewidth=1)

font1={"family":"serif","color":"b","size":"20"}

font2={"family":"serif","color":"r","size":"15"}

plt.xlabel("State/UT",fontdict=font2)

plt.ylabel("Confirmed Cases",fontdict=font2)

plt.xticks(rotation=90)

plt.title("multi plot graph",fontdict=font1)

plt.legend()

plt.grid(color='g',linestyle='-',linewidth=3)

plt.show()

 #scattering the given raw data

x=df.head(10)["State/UT"]

y=df.head(10)["Confirmed Cases"]

colors = np.random.randint(10, size=(10))

sizes=np.random.randint(300, size=(10))

plt.scatter(x,y,c=colors,s=sizes,alpha=1,cmap='nipy_spectral')

plt.colorbar()

plt.xticks(rotation=90)
```
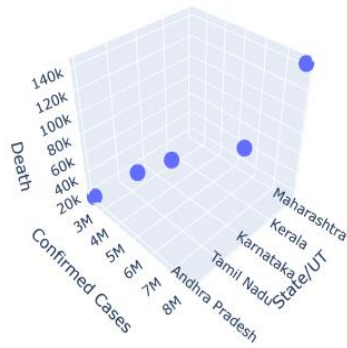
```python
font1={"family":"serif","color":"b","size":"20"}

font2={"family":"serif","color":"r","size":"15"}

plt.xlabel("State/UT",fontdict=font2)

plt.ylabel("Confirmed Cases",fontdict=font2)

plt.title("scattering graph",font1)

plt.legend(y)

plt.show()
```
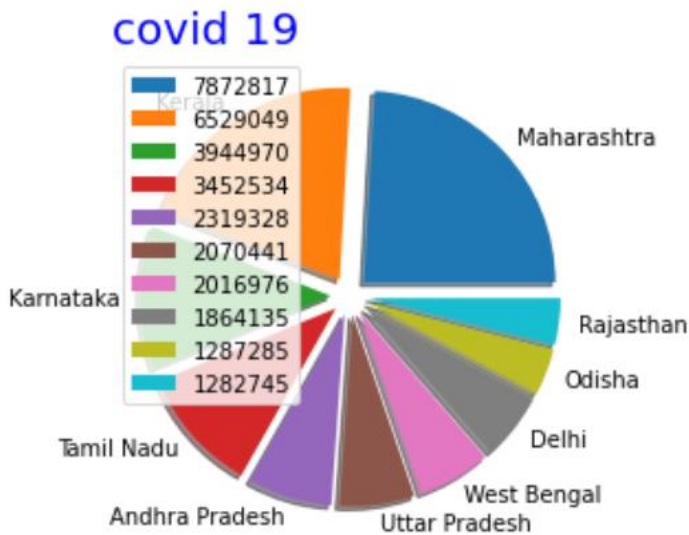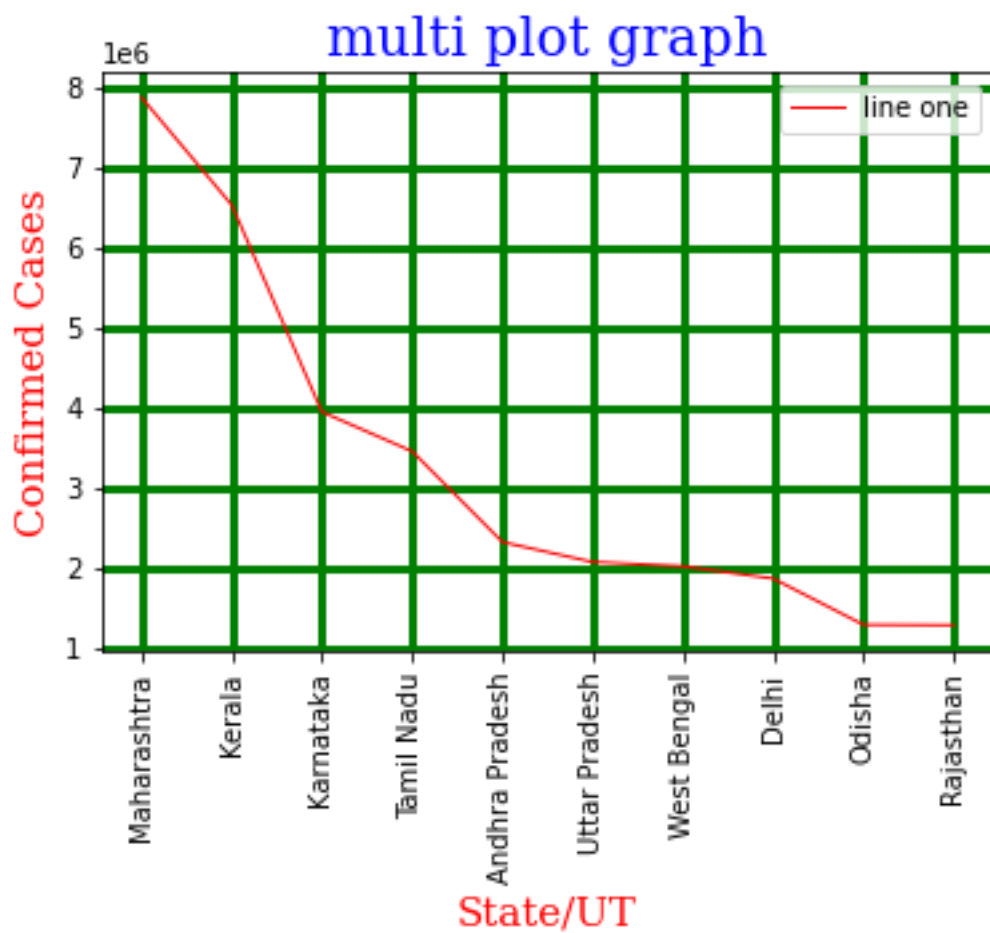
7. Results
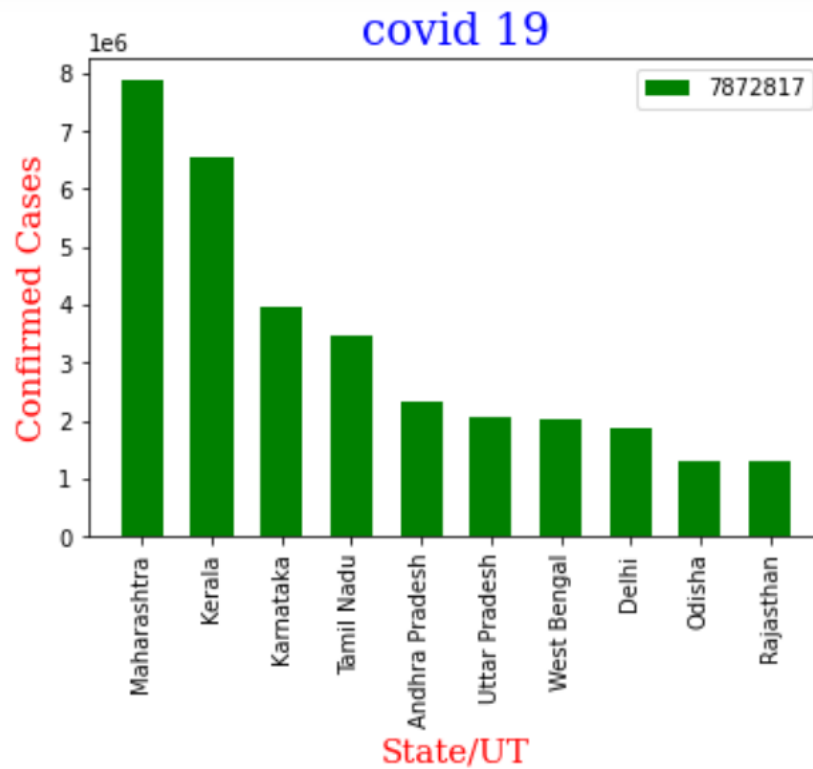
in Fig.7.1 (bar chart showing the number of confirmed cases in the
top 10 states in india)

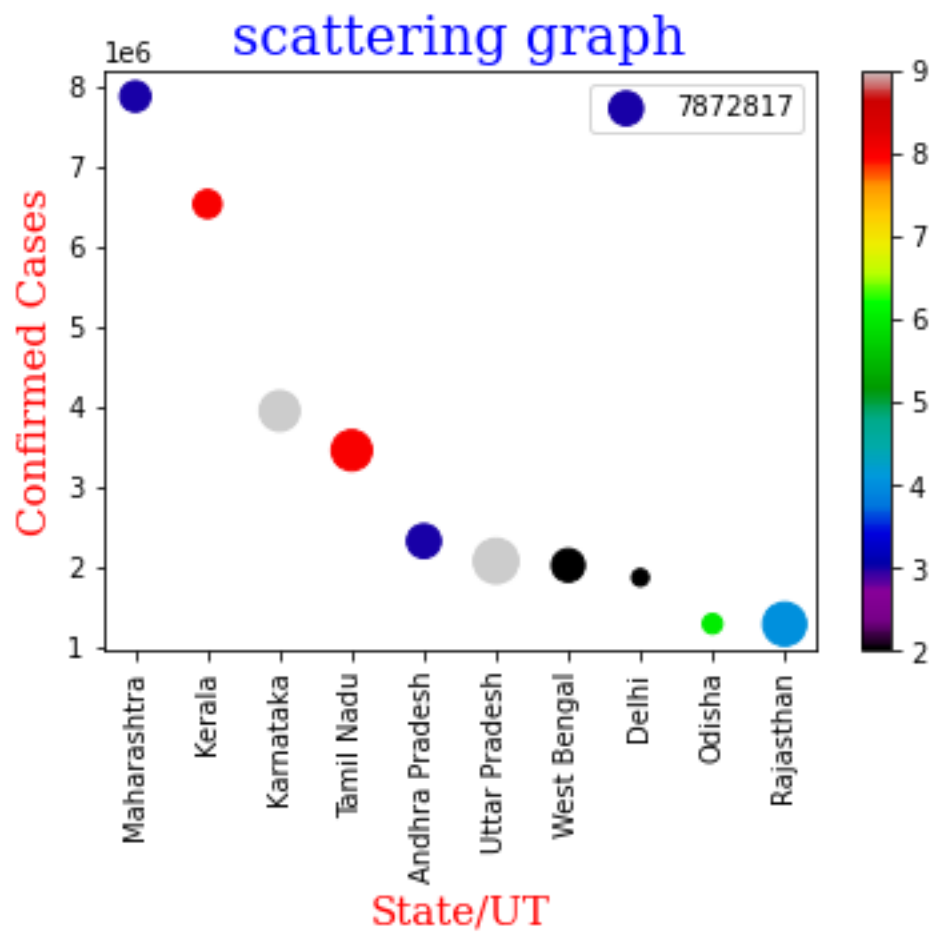**In Fig .7.2 (3D graph showing the number of confirmed cases and deaths in the top 5 states in India)**

In Fig.7.3 (pie chart showing the no of confirmed cases in the top 10 states in India)

multi plot graph

1e6

Confirmed Cases

State/UT

1

In Fig.7.4 (grid plot showing the no of confirmed cases in the top 10 states in India)

In Fig.7.5 (scatter plot showing the no of confirmed cases in the top 10 states in India)

8. Conclusion

In this project, I conclude that python is a very useful programming language for the analysis and visualization of data. This project helps us to know the details about covid 19 in India state-wise through graphical representation bar graphsraphs, pie charts, etc… This project just listed some basics to medium-advanced analysis over here, to give you an idea of how to use the data set. Through this project, I learned that visualization of any raw data is simple with the python libraries. People show more interest in cricket and this type of visualization attracts them more. Based on this project, further many new projects with an advanced level of visualizations can develop.

9. Reference

1. https://www.kaggle.com/datasets/aryakrishnanar/covid-data-india-statewise

2. https://courses.csail.mit.edu/6.867/wiki/images/3/3f/Plot-python

3. https://www.geeksforgeeks.org/python-programming-language/

4. https://www.tutorialspoint.com/python/index.htm