Green University of Bangladesh

Department of Computer Science and Engineering (CSE)

Faculty of Sciences and Engineering

Semester: (Summer, Year:2022), B.Sc. in CSE (Day)

Course Title: Data Mining Lab

Course Code: CSE-424 **Section:** 191 D4

Lab Project Name: Covid-19 Data Analysis

Student Details

	Name	ID
1.	Md. Mahmudul Hasan	191002351

Submission Date : 12-09-2022

Course Teacher's Name : Arpita Das

Lab Project Status

Marks: …… **Signature: .....**

Comments: ..... **Date: .....**

CONTENTS

Chapter-1

Introduction

1.1 Introduction

During the current coronavirus pandemic, monitoring the evolution of COVID-19 cases is of utmost importance for the authorities to make informed policy decisions and to raise awareness in the general public for taking appropriate public health measures.

At the time of the pandemic outbreak, a lack of laboratory tests, materials, and human resources implied that the evolution of officially confirmed cases did not represent the total number of cases. Even now, there are significant differences across countries in terms of the availability of tests.

1.2 Project overview

This project is designed in the Covid-19 Data Analysis in Bangladesh, India, and World. The Covid-19 data analysis has been done with these three datasets of Bangladesh, India, and the world. This project applied to Data Analysis processes; these five steps are Defining the questions, Collecting the data, Cleaning the data, Analyzing the data, and Visualizing the data.

1.3 Objective

The proposed system for Covid-19 Data Analysis and Visualizations in Jupyter Notebook. This project analysis all the covid-19 cases in Bangladesh, India, and Worldwide.

This project is programmed using python data analysis frameworks Pandas, Matplotlib, and the Machine Learning tool Scikit-learn (Sklearn).

Chapter-2
Implementation of the Project

2.1 Data Sets

2.1.1 Covid-19 in Bangladesh

The purpose of this dataset is to find out how the rates of infection are changing in Bangladesh and whether the deaths and recoveries ratio is alarming for the future.

This dataset contains every single day's data of COVID-19 outbreak in Bangladesh. From the first confirmed case of COVID-19, on 8 March 2020, it contains each confirmed, recovery, and death cases till date, this is a time-series dataset and this dataset will be updated in a daily basis.

Dataset Link: https://www.kaggle.com/datasets/ridoy11/covid19-bangladesh-dataset

2.1.2 Covid-19 in India

This data set contains the data of covid-19 Conformed, Recovered and Deaths in India. This data is taking from the non-governmental organization.

Dataset Link: https://www.kaggle.com/datasets/ravichaubey1506/covid19-india

2.1.3 Covid-19 in the World

This data set contains the data of covid-19 Conformed, Recovered and Deaths in the World.

Dataset Link: https://www.kaggle.com/datasets/imdevskp/corona-virus-report

2.2 Technologies

The Software interfaces are the languages, codes and that programs use to communicate with each other and to the hardware.
Operating System: Windows 8, 10, 11and Linux etc.
Developing Tool: Jupyter Notebook
Python Frameworks: Pandas and Matplotlib.
Machine Learning: Scikit-learn (Sklearn)

2.3 Project Implementation

2.3.1 Data Analysis Process

The data analysis process, or alternately, data analysis steps, involves gathering all the information, processing it, exploring the data, and using it to find patterns and other insights.

The process consists of:

1. **Define the Question:** Ask myself why I'm doing this analysis, what type of data analysis I want to use, and what data I'm planning on analyzing.
2. **Data Collection:** Guided by the requirements I've identified, it's time to collect the data from my sources. Sources include case studies, surveys, interviews, questionnaires, direct observation, and focus groups. Make sure to organize the collected data for analysis.
3. **Data Cleaning:** Not all of the data I collect will be useful, so it's time to clean it up. This process is where I remove white spaces, duplicate records, and basic errors. Data cleaning is mandatory before sending the information on for analysis.
4. **Data Analysis:** Here is where you use data analysis software and other tools to help you interpret and understand the data and arrive at conclusions. Data analysis tools include Excel, Python, R, and Microsoft Power BI.
5. **Data Visualization:** Data visualization is a fancy way of saying, "graphically show our information in a way that people can read and understand it." I can use charts, graphs, maps, bullet points, or a host of other methods.

2.3.2 Data Analysis Python Frameworks

Pandas: Pandas is a high-level Python library for data analysis. In the Python ecosystem, pandas is the most advanced and fast-growing tool for data processing and manipulation. It is necessary for data processing, manipulation, and visualization.

- It's benefits of my project:
  - The data in a suitable way for data analysis and library offers different methods for simple data filtering
  - It has various tools for seamless I/O processes and reads data from CSV, TSV, XLSX files and many more.

Matplotlib: Matplotlib is a standard two-dimensional data visualization library. It is a flexible and easily configurable library that together with NumPy, SciPy, and IPython provides features similar to MATLAB. Matplotlib helps to make static, animated, and interactive plots by writing a few lines of code.

- It's the package used for my project are: Line plot, Scatter plot, Bar chart, Histogram, and Pie chart.

Machine Learning Tools: Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering, and dimensionality reduction via a consistency interface in Python.

- It's used in my project on Distribution Plot.

Chapter-3

Performance Evaluation

3.1 Covid-19 Data Analysis in Bangladesh

In this section, Covid-19 Data Analysis in Bangladesh datasets are Analysis and Visualize for the month of March and April 2020.

- I'm describing the Data Analysis in the Figures caption.

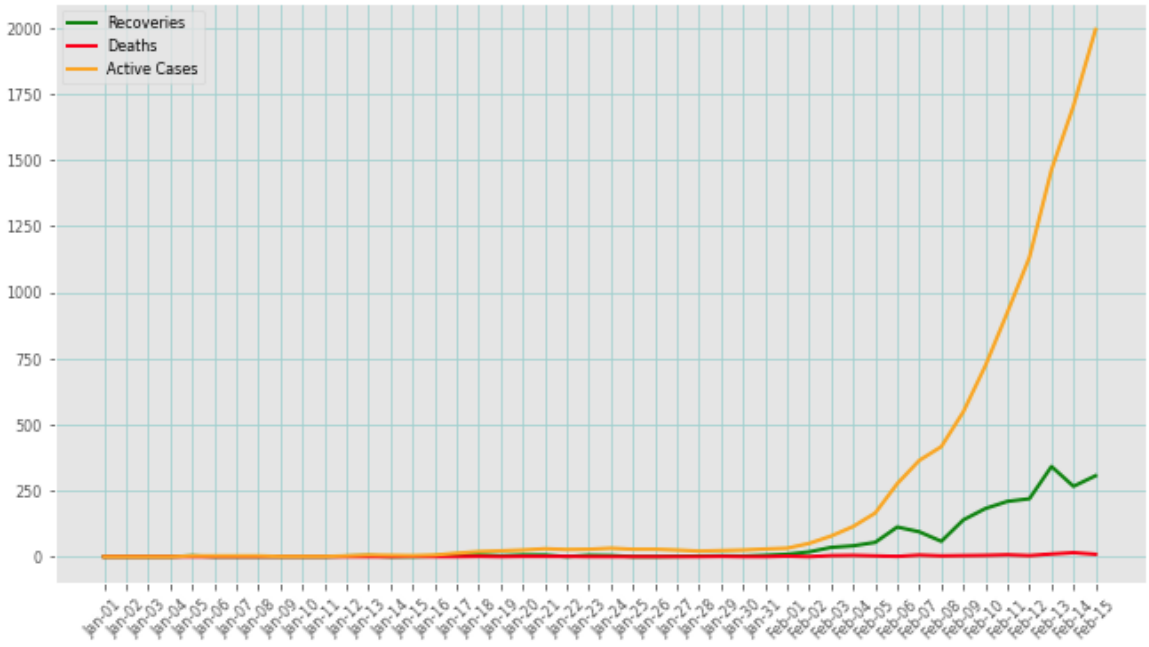

Figure-1: Acive Cases, Recoveries and Deaths by day.

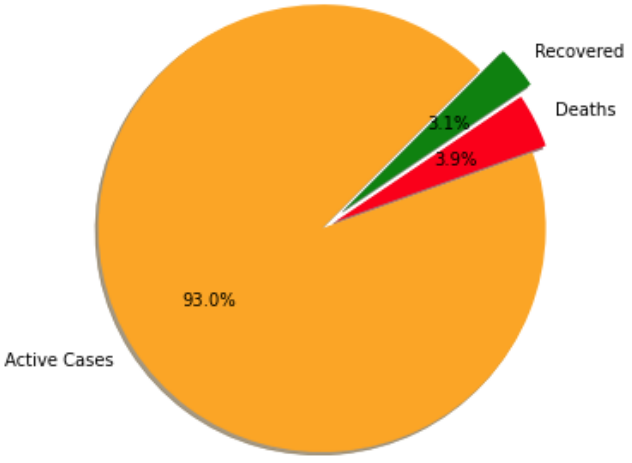

Figure-2: Acive Cases, Recoveries and Deaths by day.

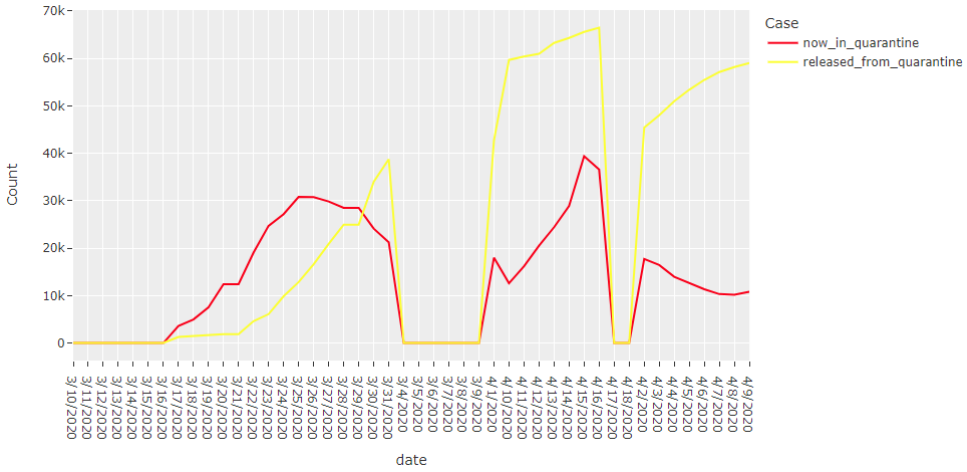

Figure-3: Present Quarantine VS Released from Quarantine by day.

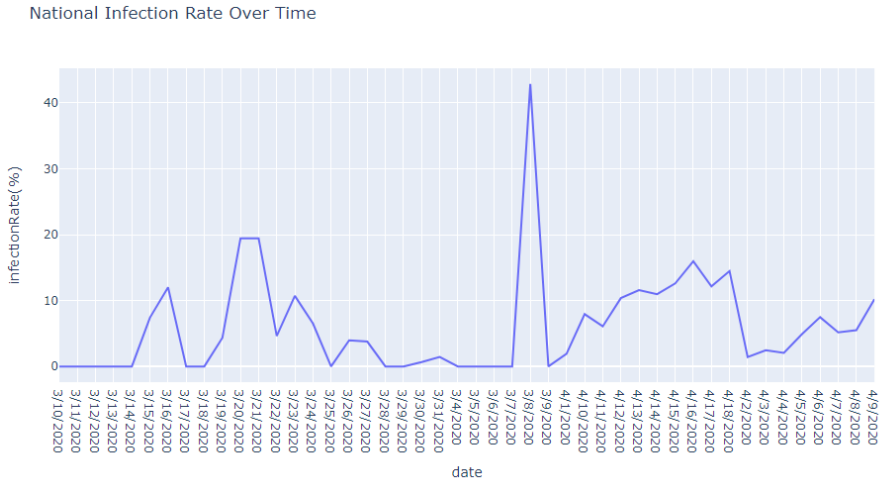

Figure-4: Infection Rate by per day.

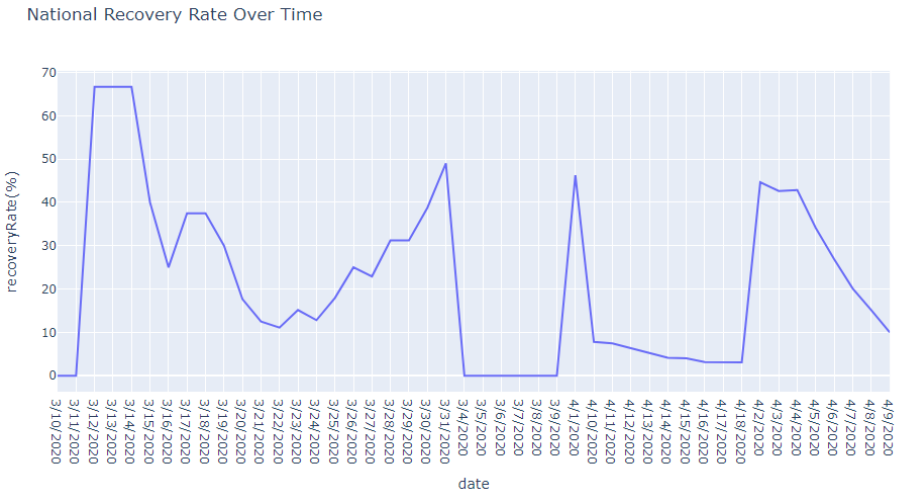

Figure-5: Recovery Rate by day.

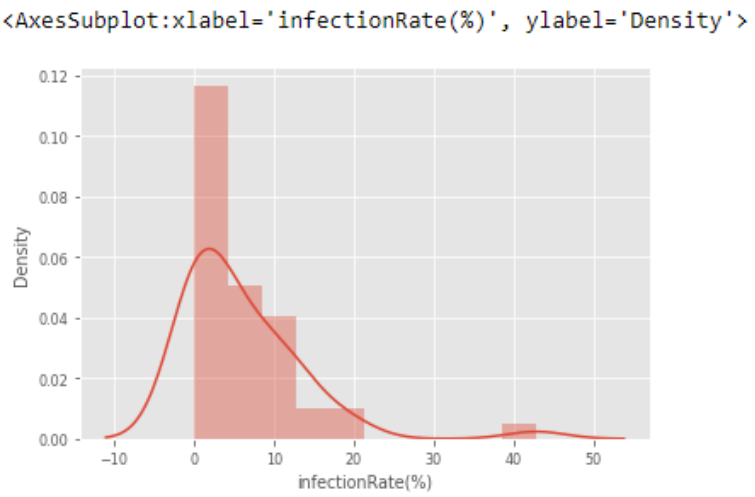

Figure-6:. Distribution plot on infection Rate.

3.2 Covid-19 Data Analysis in India

In this section, Covid-19 Data Analysis in Indai datasets are Analysis and Visualize for the State in India

	Name of State / UT	Total Confirmed cases (Indian National)	Total Confirmed cases (Foreign National)	Cured	Death	Total Cases	Active Cases
0	Andhra Pradesh	12	0	1	0	12	11
1	Chhattisgarh	6	0	0	0	6	6
2	Delhi	38	1	6	1	39	32
3	Gujarat	43	0	0	3	43	40
4	Haryana	16	14	11	0	30	19
5	Himachal Pradesh	4	0	0	1	4	3
6	Karnataka	20	0	3	2	20	15
7	Kerala	131	7	11	0	138	127
8	Madhya Pradesh	23	0	0	1	23	22
9	Maharashtra	144	3	15	4	147	128
10	Odisha	3	0	0	0	3	3
11	Puducherry	1	0	0	0	1	1
12	Punjab	29	0	0	1	29	28
13	Rajasthan	41	2	3	0	43	40
14	Tamil Nadu	32	3	1	1	35	33
15	Telengana	34	11	1	0	45	44
16	Chandigarh	7	0	0	0	7	7

Figure-7:.Calculate Total Cases base on Indian Nationa and Forign Nationa for the State in India.

Name of State / UT	Total Cases
Maharashtra	147
Kerala	138
Telengana	45
Uttar Pradesh	43
Gujarat	43
Rajasthan	43
Delhi	39
Tamil Nadu	35
Haryana	30
Punjab	29

Figure-8:. Total Cases in the per State of India

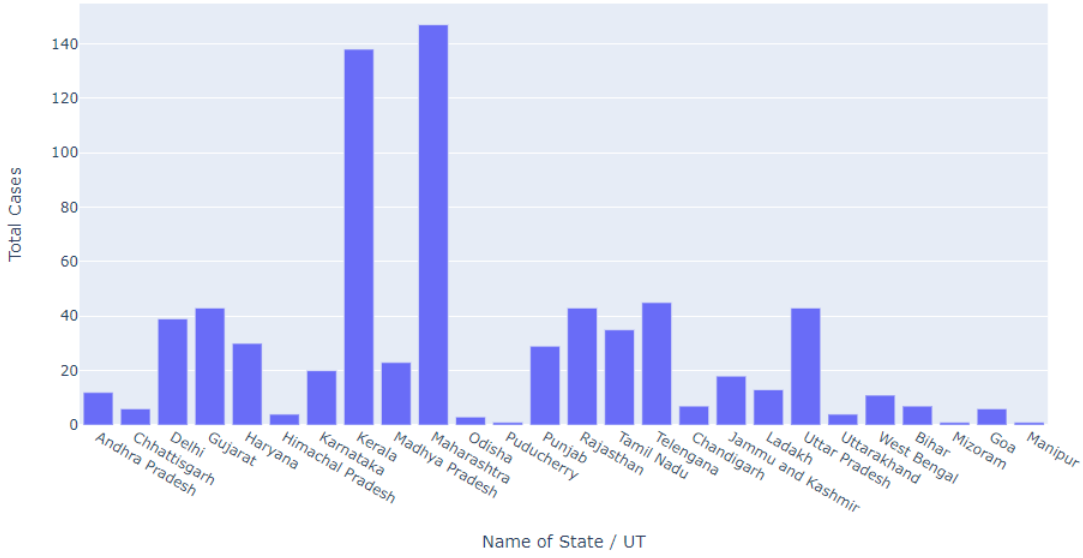

Figure-9:. Total Cases in the per State of India

3.3 Covid-19 Data Analysis in the World

In this section, Covid-19 Data Analysis in the World datasets are Analysis and Visualize for the State in World. Date on 29 April 2020.

- Column Description
  - Date- Specific Date
  - Region-The all countries.
  - Confirmed - The number of confirmed cases
  - Recovered - The number of recovered cases
  - Deaths- The number of death cases

Show the number of Confirmed, Deaths and Recovered cases in each Region(Figure 10)

Region	Confirmed	Deaths	Recovered
Afghanistan	1939	60	252
Albania	766	30	455
Algeria	3848	444	1702
Andorra	743	42	423
Angola	27	2	7
Antigua and Barbuda	24	3	11
Argentina	4285	214	1192
Armenia	1932	30	900

Figure-10:. Confirmed, Deaths and Recovered cases in each Region.

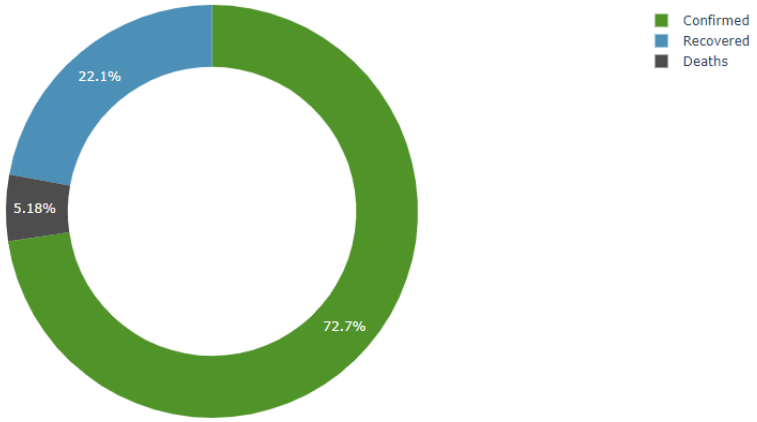

Figure-11:. Confirmed, Deaths and Recovered cases in The world this day (29 April 2020).

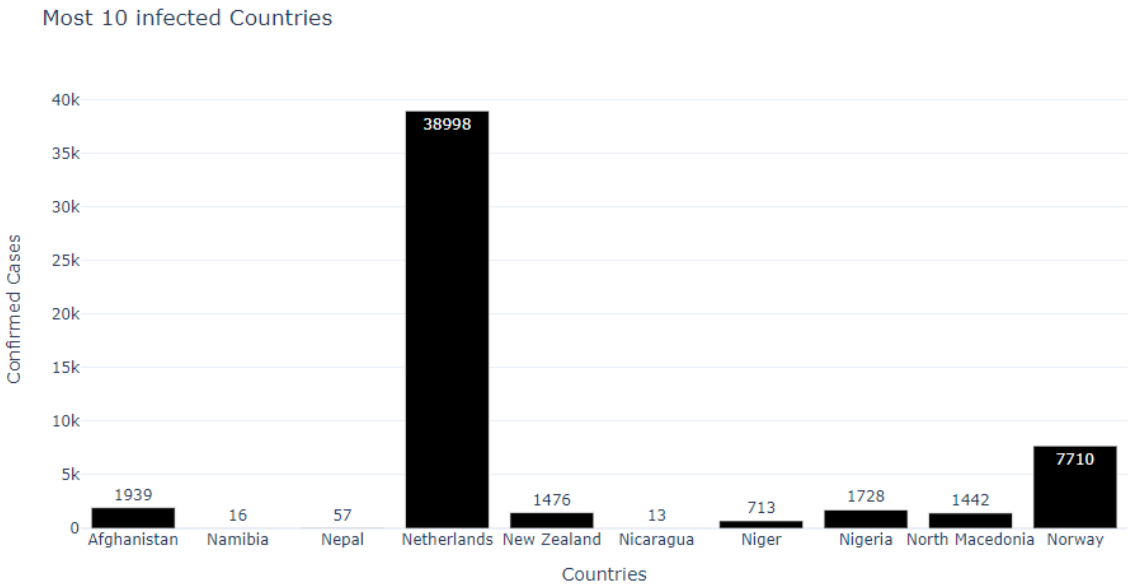

Figure-12:. Confirmed Cases in each country.

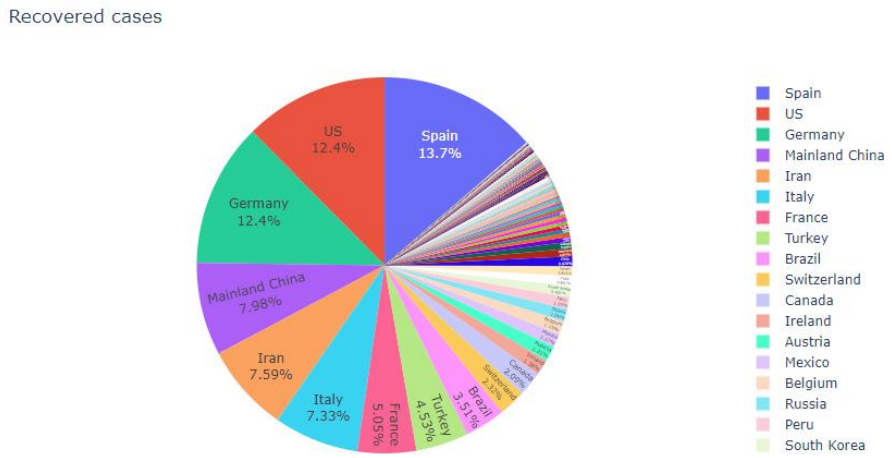

Figure-13:. Recovered cases in each country in the world.

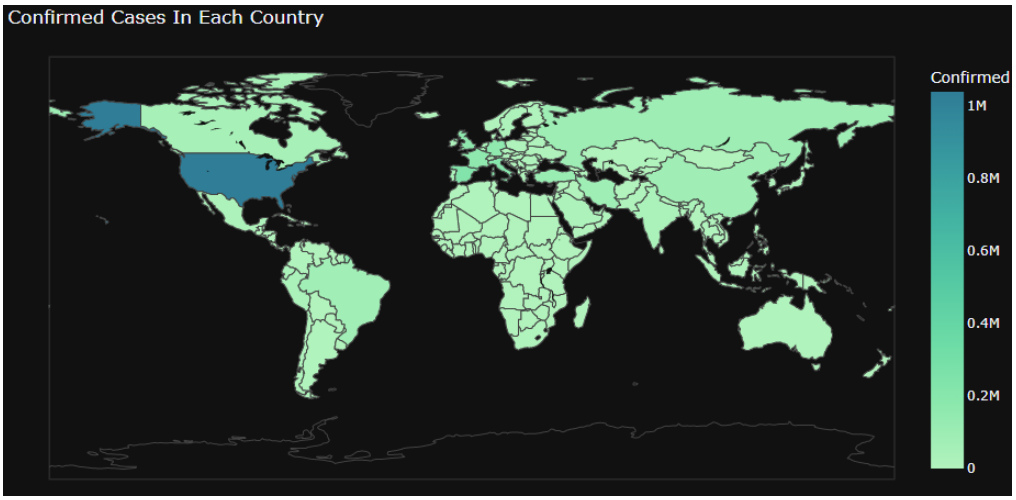

Figure-14:. Confrimed Cases in each Country.

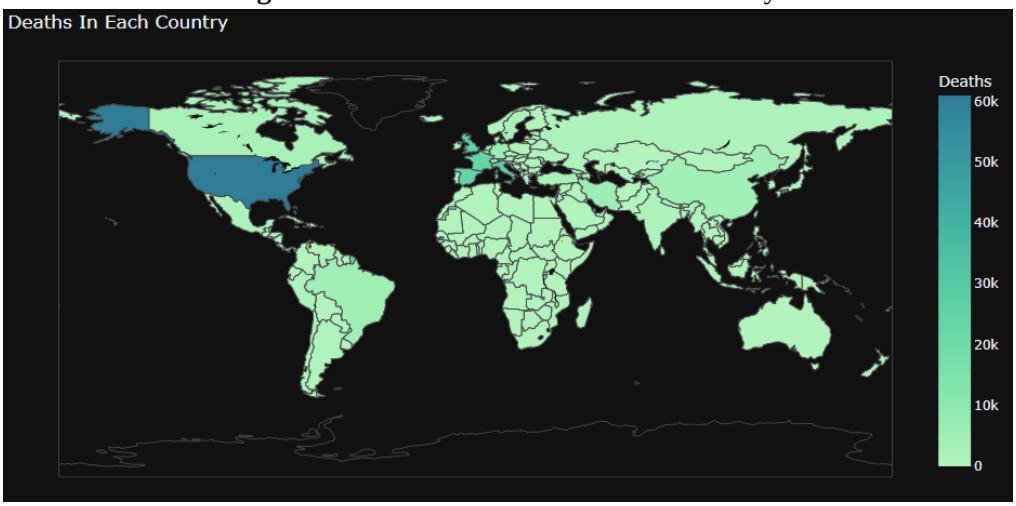

Figure-15:. Deaths in each Country.

Chapter-4

Conclusion

4.1 Introduction

The most important challenge and limitation of Covid-19 is the number of survey responses. My outlier detection methods are heuristic and could, in the future, be improved to be more resilient to malicious responses.

In this project, I have to present the Data Analysis and visualization methods of datasets, which use open analysis to monitor the progress of the COVID-19 pandemic.

4.2 Scope of Future Work

- Advanced Data Analysis and Visualization.
- Dashboards using Power BI

References

1] https://www.kaggle.com/