**RAMNIRANJAN JHUNJHUNWALA COLLEGE**

**Department of DSAI**

**Ghatkopar (West), Mumbai - 86**



**2021-2022**

**Project Report  
On**

**Exploratory Data Analysis on Covid-19**

**In partial fulfillment of M.Sc. (DSAI)**

**By**

**Ms Grishma Shah**

**Project Guide**

**Prof. Bharati Bhole**

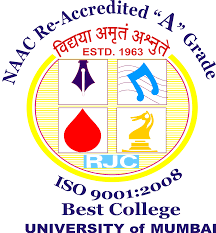
**RAMNIRANJAN JHUNJHUNWALA COLLEGE**

**(AUTONOMOUS)**

**(Affiliated to University of Mumbai)**

**GHATKOPAR(WEST), 400086.**

**Certificate**



This is to certify that the Project entitled, “Exploratory Data Analysis on Covid-19

” is bonafide work of Ms Grishma Shah bearing **Seat No: - 27** submitted in partial fulfilment of the requirements for the award of Degree Master of Science in DSAI,

Signature of Internal Guide Sign of Co-Ordinator

Examiner

Date: College Seal

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **SR.NO** | **TOPIC** | PAGE NO |
| 1 | ACKNOWLEDGEMENT | 4 |
| 2 | ABSTRACT | 5 |
| 2 | **INTRODUCTION** | 6 |
| 3 | OBJECTIVE & PROPOSED SYSTEM | 7 |
| 4 | **EDA AND DESCRIPT** | 8 |
| 5 | **Covid-19 DATABASE** | 8 |
| 6 | GANTT CHART | 9 |
| 7 | FUNCTIONALITY AND SCOPE | 10 |
| 8 | SRS | 11 |
| 9 | DETAILS OF THE Covid-19 | 12 |
| 10 | **AUTOMATED PROCESSES** | 13 |
| 11 | DATA PREPROCESSING | 14 |
| 12 | RESEARCH METHODOLOGY | 14 |
| 13 | DETAILS OF Covid-19 | 15 |
| 14 | PLOT COVID-19 CASES ON DIFFERENT CATEGORIES | 16 |
| 15 | DATA VISUALIZATION USING POWR BI | 22 |
| 16 | POWER BI DASHBOARD | 23 |
| 17 | INSIGHTS & ANALYSIS  PROJECT DETAILS | 25 |
| 18 | CONCLUSION & SUGGESTION | 25 |
| 19 | REFERENCE | 26 |
| 20 | PROJECT CODE LINK | 26 |
|  |  |  |

**ACKNOWLEDGEMENT**

Before we get into thick of things, we would like to add a few heartfelt words for the people who were part of **Exploratory Data Analysis on Covid-19**project in numerous ways, people who gave unending support right from the stage the project idea was conceived. A project report is such a comprehensive coverage; it would not have been materialized without the help of many.

The four things that go on to make a successful endeavour are dedication, hard work, patience and correct guidance. Able and timely guidance not only helps in making an effort fruitful but also transforms the whole process of learning and implementing into an enjoyable experience.

In particular, I would like to thank our principal **Dr. Himanshu Dawda**

& R.J college.

I would like togive a very special honour and respect to our teacher, **Prof. BHARATI BHOLE** who took keen interest in checking the minute details of the project work and guided us throughout the same. A sincere quote of thanks to the non-teaching staff for providing us software & their time.

Last but not least I wish to avail myself of this opportunity, express a sense of gratitude and my friend Master. Madhusudan Tiwari for his support, strength and help for everything.

**ABSTRACT**

Since December 2019 the world is experiencing a deadly disease caused by a novel coronavirus termed as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease associated with this virus is known as COVID-19. This paper focuses on COVID-19 based on freely available datasets including the ones in Kaggle repository. Data analytics is provided on a number of aspects of COVID-19 including the symptoms of this disease, the difference of COVID-19 with other diseases caused by severe acute respiratory syndrome (SARS), Middle East respiratory syndrome (MERS), and swine flu. The impact of temperature on the spread of COVID-19 is also discussed based on the datasets. Moreover, data visualization is provided on the comparison of infections in males/females which shows that males are more prone to this disease and the older people are more at risk. Based on the data, the pattern in the increase of confirmed cases is found to be an exponential curve in nature. Finally, the relative number of confirmed, recovered and death cases in different countries are shown with data visualization.

**INTRODUCTION**

Enveloped, single stranded positive-sense ribonucleic acid (RNA) viruses named coronaviruses contain one of the largest viral genomes which are around 32 kbp in length. They can infect humans as well as a wide range of animals [1]. The 2019 novel coronavirus termed as SARS-CoV-2 caused pneumonia outbreak in Wuhan, China resulting in the 2019- 2020 coronavirus pandemic declared by World Health Organization (WHO). It belongs to the Ortho coronavirinae subfamily. It is distinct from Middle East respiratory syndrome (MERS) and severe acute respiratory syndrome corona virus (SARS-CoV).Wuhan’s Huanan Seafood Wholesale Market trades a variety of live animal species which includes fish, poultry, marmots, snakes and bats which linked the outbreak.

In order to prevent the spread of this infection, the WHO recommends frequent hand washing, keeping unwashed hands away from the face, social distancing, and covering coughs and sneezes with a tissue or inner elbow. Some national health authorities recommend masks for the suspects and their caregivers. So far no vaccine or antiviral treatment is available for COVID-19 .

In order to stop this pandemic, the route of transmission of the virus to humans via animals, identification of the reservoirs, the incubation period of the virus, the characteristics of the susceptible population and their survival rates need to be identified. The analysis of the clinical information regarding age, gender, source of the virus, incubation period, transmission route and treatment response will help researchers to develop vaccines against COVID19. As of 13 April 2020, COVID-19 has affected more than 1,858,800 patients in 210 countries and territories around the world and has become a major global health concern. This paper analyses COVID-19 based on the currently available data. Analytics is provided on a number of aspects of COVID-19 including the symptoms, the difference with other viruses, and the impact of temperature. Moreover, data visualization is provided on the comparison of infections on male/females, and the pattern in the increase of confirmed cases and the relative number of confirmed /recovered/death cases in different countries. The rest of the paper is organized as follows. Section 2 describes the different aspects of COVID-19 using tabular data. Section 3 provides visualization of how the infection has spread across the world using pie charts and bar charts. Section 4 provides concluding remarks.

**PROPOSED SYSTEM**

1. In the proposed system, we are going to use the Covid-19 Database to analyze and derive insights on the Confirmed ,recovered and deaths cases. We will pre-process the database to get reliable and accurate patterns and insights and use descriptive statistics to draw conclusions such as — major causes of deaths, major countries, states are effected.

2. Using EDA, we will show the trend of the growth of covid-19 in the world and how it spread to various parts of the world. Using graphs, we study how one factor behind terrorism is interrelated to the other.

3. We also do predictive analytics as to what the magnitude of the cases acts can be in the upcoming months,

**OBJECTIVE OF ANALYSIS AND STUDY**

**The objectives of this project are:**

* To study and analysis the covid-19 cases that have taken place in various parts of the world, at different times.
* the various causes behind covid cases
* Covid cases according to country
* Range of covid-19 cases which is increasing or decreasing
* People who died due to covid-19

**EDA & DESCRIPTIVE ANALYTICS**

**Descriptive analytics** is the interpretation of historical data to better understand changes that have occurred in a business. **Descriptive analytics** describes the use of a range of historic data to draw comparisons. A statistical method that is used to search and summarize historical data in order to identify patterns or meaning. Preliminary stage of data processing that creates a summary.

**Exploratory data analysis**(**EDA**) is an approach to analysing data sets to summarize their main characteristics, often with visual methods. EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. It’s where the researcher takes a bird’s eye view of the data and tries to make some sense of it.

**COVID-19 DATABASE**

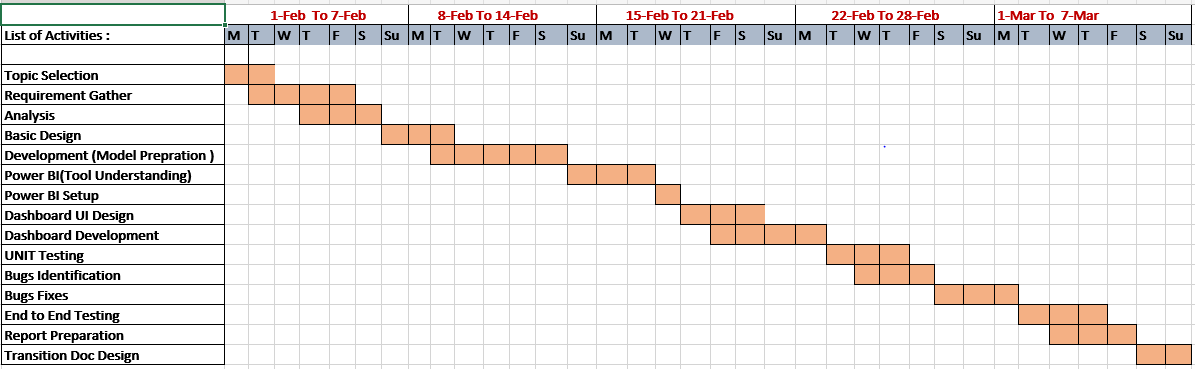
**About the dataset**

In this document, the many linked charts, our COVID-19 Data Explorer, and the Complete COVID-19 dataset report and visualize the data on confirmed cases , recovered cases and deaths from Johns Hopkins University (JHU). We make the data used in our charts and tables downloadable as a complete and structured .csv, .xlsx and json file.

The Johns Hopkins University dashboard and dataset is maintained by a team at its Center for Systems Science and Engineering (CSSE). It has been publishing updates on confirmed cases and deaths for all countries since January 22, 2020. A feature on the JHU dashboard and dataset was published in The Lancet in early May 2020. This has allowed millions of people across the world to track the course and evolution of the pandemic.

JHU updates its data multiple times each day. This data is sourced from governments, national and subnational agencies across the world — a full list of data sources for each country is published on Johns Hopkins GitHub site. It also makes its data publicly available there.

**Gantt Chart**

****

**Functionality and Scope**

* We can get the different types of python libraries which are as follows:
* **Pandas**
* **NumPy:**
* Matplot
* Seaborn
* Pyplot
* Line interpolation / smoothing
* Bi-polar bar chart
* Simple pie chart
* Pie chart with custom labels
* Animating a Donut with SVG animate
* We can also get data which is available on Kaggle.
* We will able to find out in which area the tweets are more trending with help of world map.
* We can visually represent all insightful data with the help of creating dashboard.

**SRS (Software Requirement Specification)**

**Internal Interface Requirement:**

The recent explosion in data pertaining to users on social media has created a great interest in performing data analysis on this data using Dataset and Machine Learning principles to understand the covid cases across all world. This project intends to perform the same tasks. The difference between this project and other analysis tools is that, it will perform real time analysis of data and create attractive dashboard. Describe the context and origin of the product being specified in this SRS. For example, state whether which terrorism rate is higher in which country or state. If the SRS defines a component of a larger system, relate the requirements of the larger system to the functionality of this software and identify interfaces between the two. A simple diagram that shows the major components of the overall system, subsystem interconnections, and external interfaces can be helpful. The data functions are:

* Collect data from Kaggle website
* Remove redundant information from these collected data
* Do the development on Jupyter notebook
* Perform exploratory data Analysis on the Covid-19 dataset to classify their nature viz. positive, negative and so on.

**External Interface Requirement:**

We classify External Interface in 3 types, those are:

**User Interface:** Describe the logical characteristics of each interface between the software product and the users. This include sample screen images, GUI standards or product family style guides that are to be followed, screen layout constraints, standard buttons and functions that will appear on every screen, keyboard shortcuts, error message display standards, and so on. Define the software components for which a user interface is needed.

**Hardware interface:** Describe the logical and physical characteristics of each interface between the software product and the hardware components of the system. This include the supported device types, the nature of the data and control interactions between the software and the hardware, and communication protocols to be used.

**Software Interface:** Describe the connections between this product and other specific software components (name and version), including databases, operating systems, tools, libraries, and integrated commercial components. Identify the data items or messages coming into the system and going out and describe the purpose of each. Describe the services needed and the nature of communications. If the data sharing mechanism must be implemented in a specific way (for example, use of a global data area in a multitasking operating system), specify this as an implementation constraint.

Python (Pandas Libraries) – Backend process

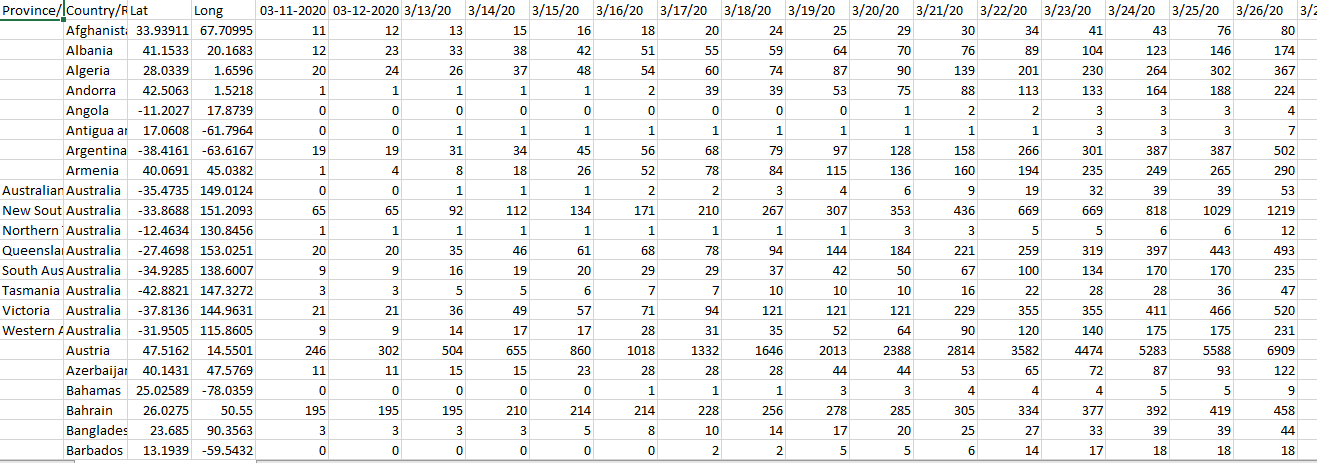
Power Bi– Data visualization

Dataset – Kaggle

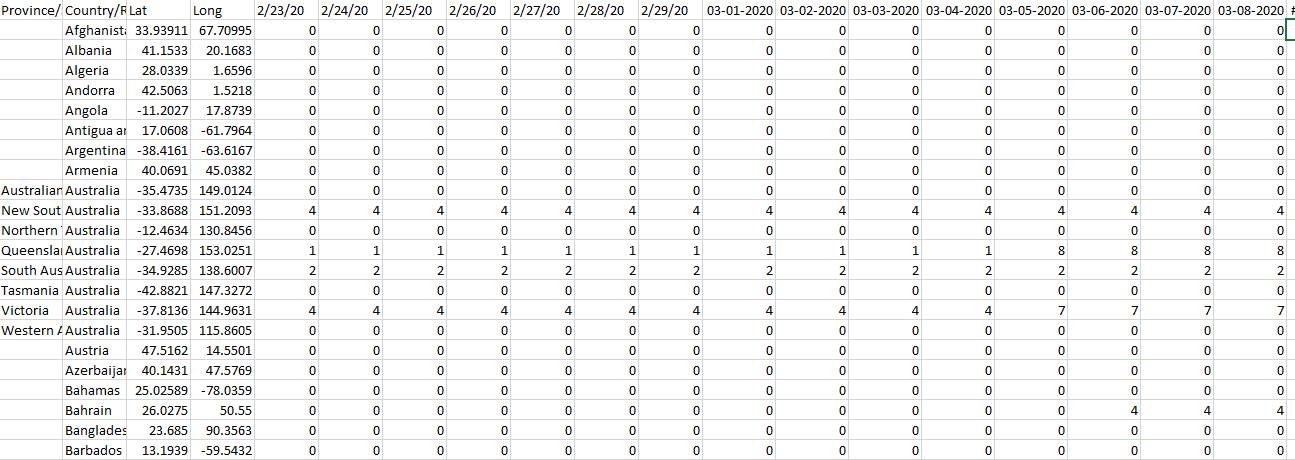
**CHARACTERISTICS OF THE COVID-19**

* Contains all cases which is confirmed, recovered and died
* the most of the cases are increasing day by day
* Includes information on starting from 22nd January 2020 to 15th May 2020.
* Includes information of all the countries in that world
* Daily Articles are and data are getting updated to for covid-19

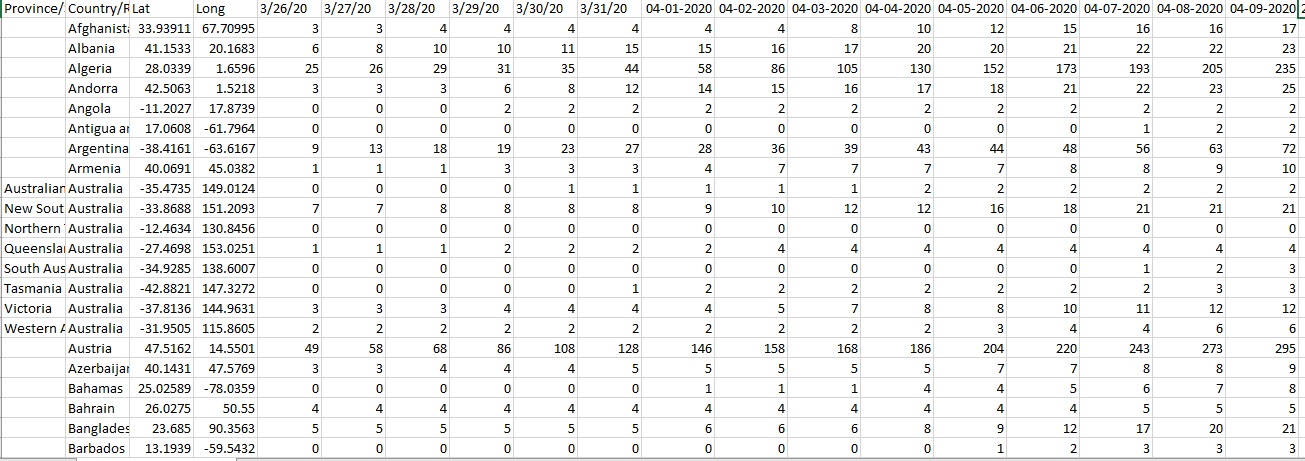
**Dataset for confirmed Cases-**



**Dataset for Recovered Cases-**



**Dataset for Deaths-**



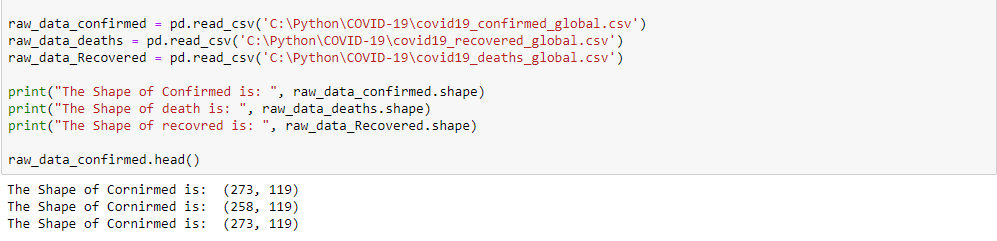
**Automated Processes**

* Initial Boolean filters of articles.
* Remove duplicate articles
* Location identification
* Review relevant source articles; apply covid-19 dataset inclusion criteria to identify ratio of cases
* Populate database with confirmed, recovered and death cases characteristics according to established coding rules.
* Present analysts with high-validity, topically clustered source articles.

**Data pre-processing**

The Data contains 3 csv files for confirmed, recovered and deaths, which are numeric.Now the since the project contains different types of Graphs and different aspects of it have been explored so different type of data pre-processing has been done before plotting the respective graph, which could be making of a new column or a new data frame itself.

**In order to calculate the casualties in the dataset I introduced:**



**RESEARCH METHODOLOGY**

**Exploratory Data Analysis**

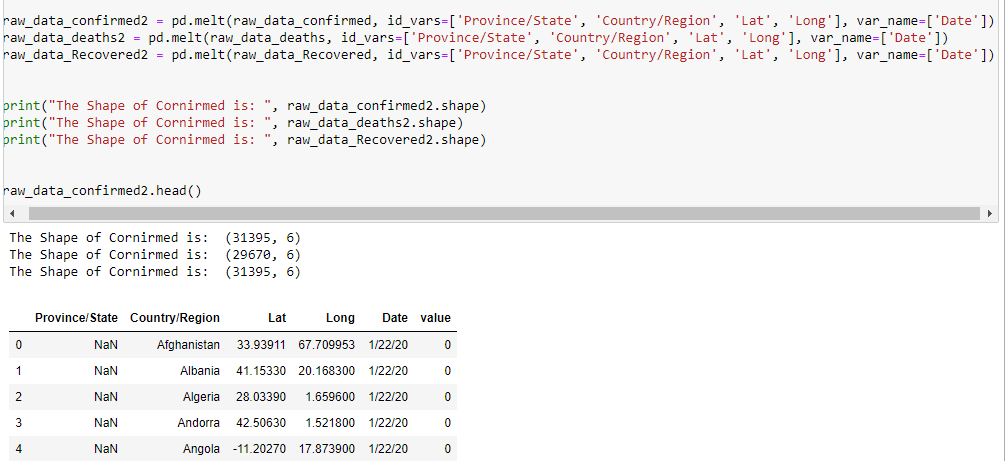
Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

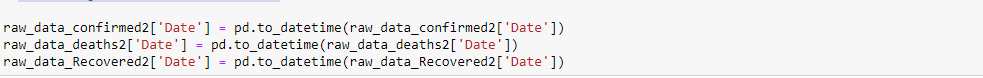
* maximize insight into a data set
* uncover underlying structure
* extract important variables
* develop parsimonious models
* determine optimal factor settings.

**Details of Covid-19 Dataset**

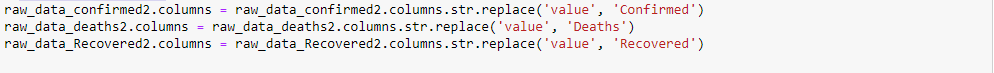
**Un-pivoting the Data**



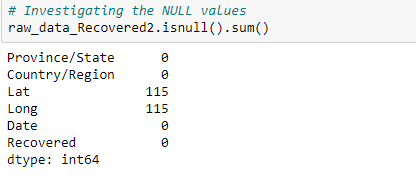
**Converting the new column to dates**



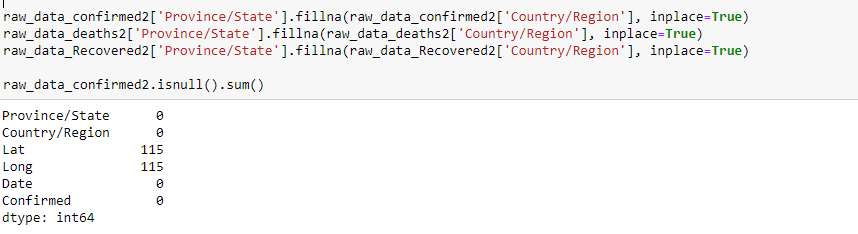
**Renaming the Values**



**Checking null values**



**Dealing with NULL values**



**Joining all 3 tables in 1 table**

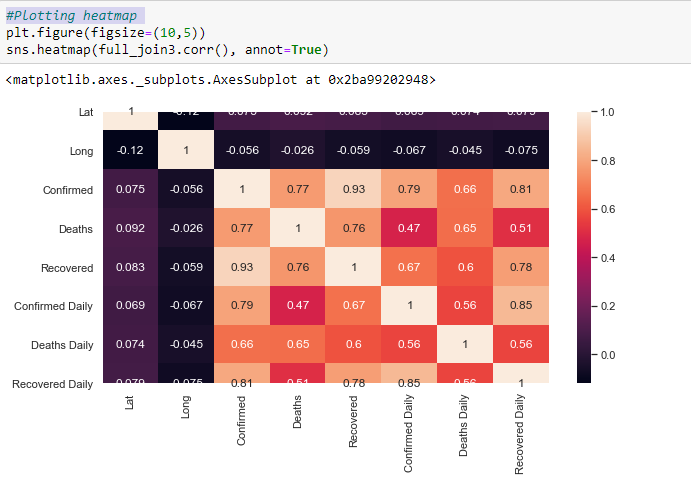


**Adding Month and Year as a new Column**

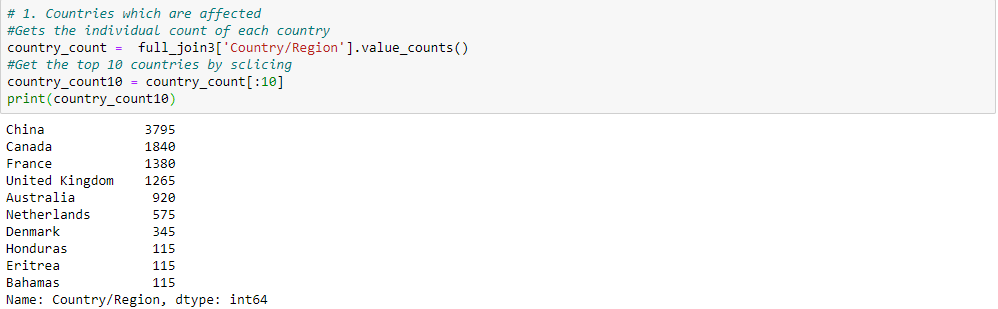


**PLOT COVID CASES BASED ON DIFFERENT CATEGORIES**

* **Plotting heatmap**

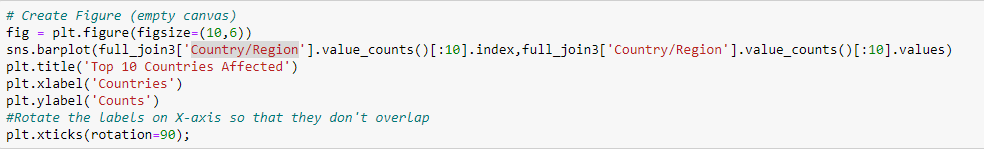


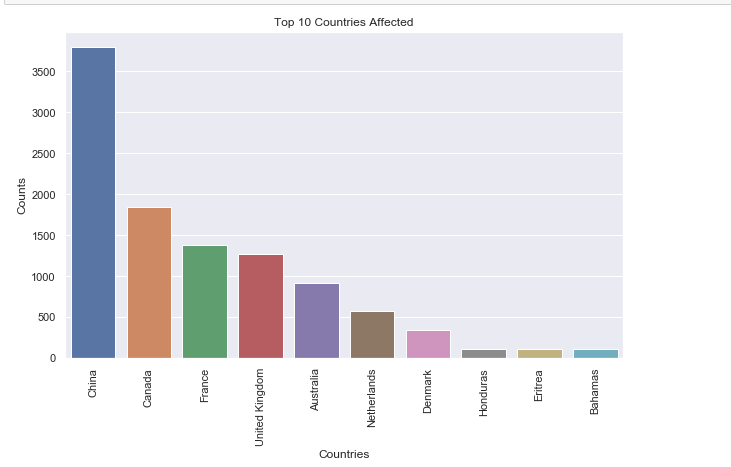
* **Countries which are affected most.**



* **Create Figure (empty canvas)**

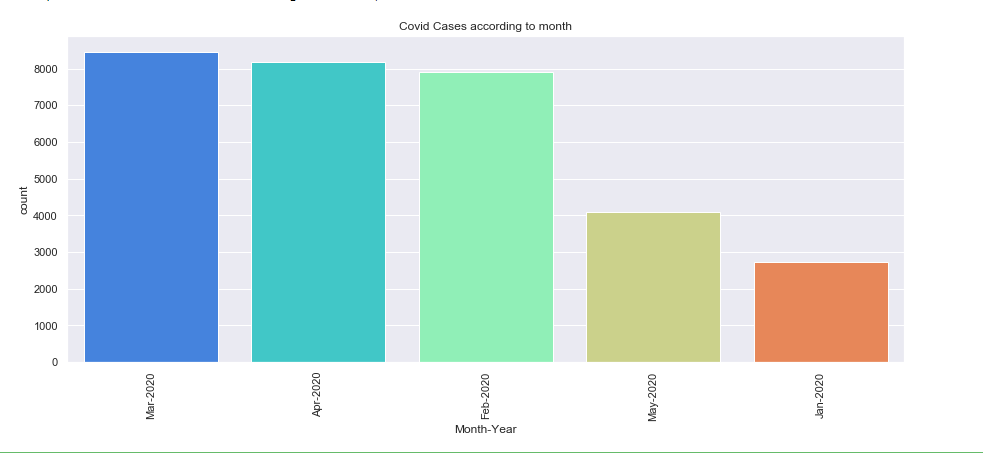
**Top 10 Countries Affected Most in Graph.**





* **Cases Activities According to the Year**





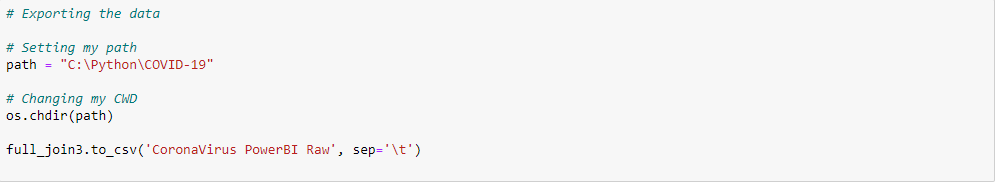
**DATA VISUALIZATION USING POWR BI**

**What is Power BI?**

Power BI is a business analytics service by Microsoft. It aims to provide interactive visualizations and business intelligence capabilities with an interface simple enough for end users to create their own reports and dashboards.

Here I have performed EDA (Exploratory Data analysis) using python script and extract insightful data in power bi to create the dashboard.

* **Python Script for connecting power BI instance.**

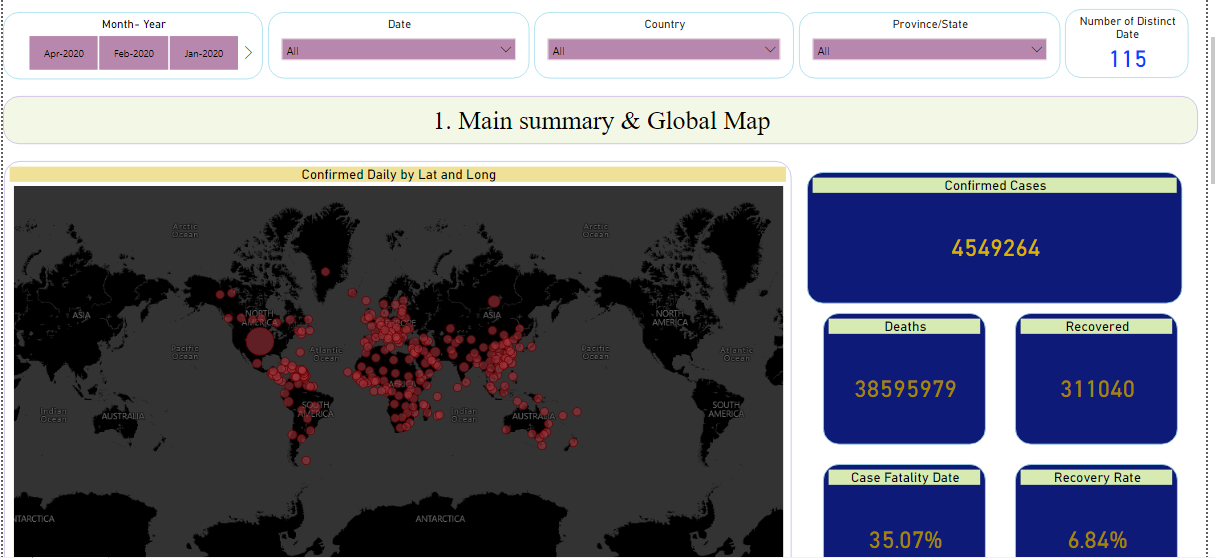


**POWER BI DASHBOARD**

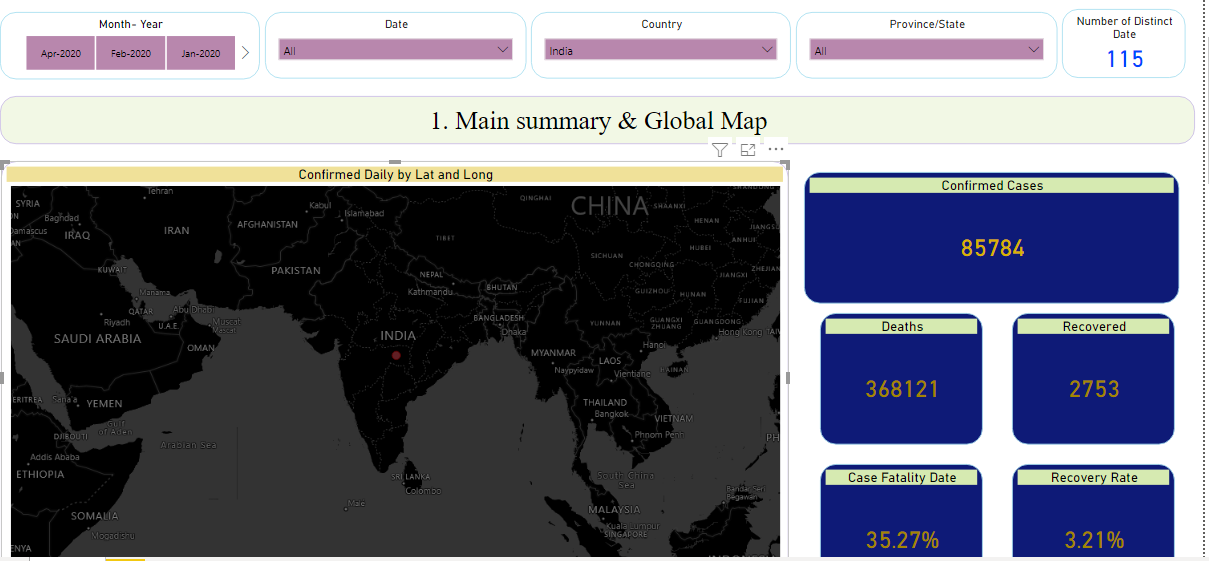
**Header**



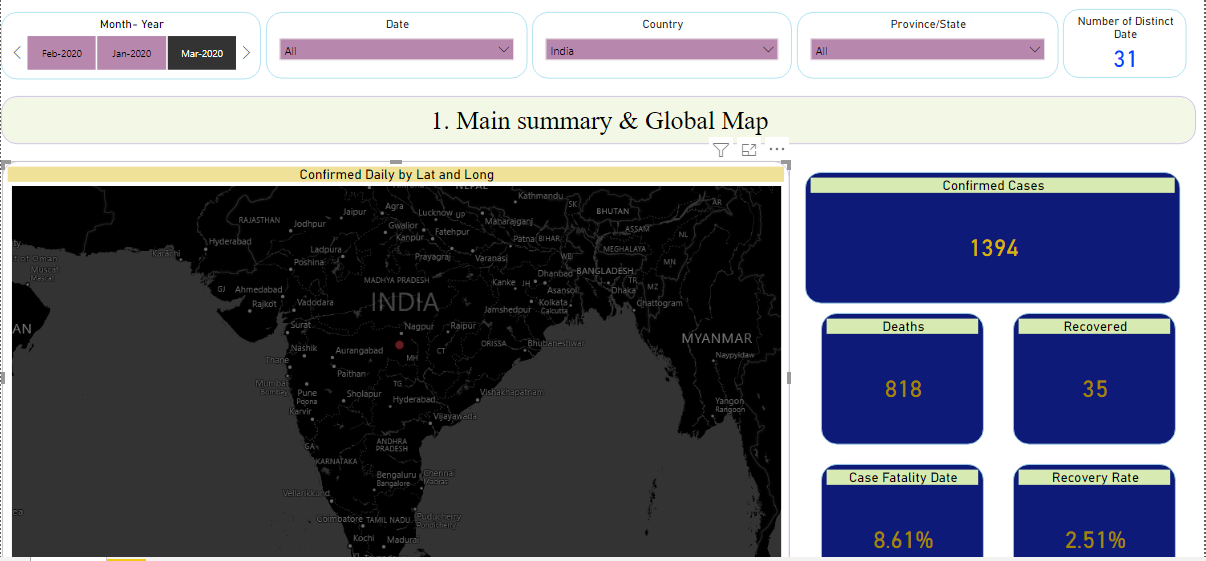
**Here you can see the total number of confirmed, Recovered and Death cases**



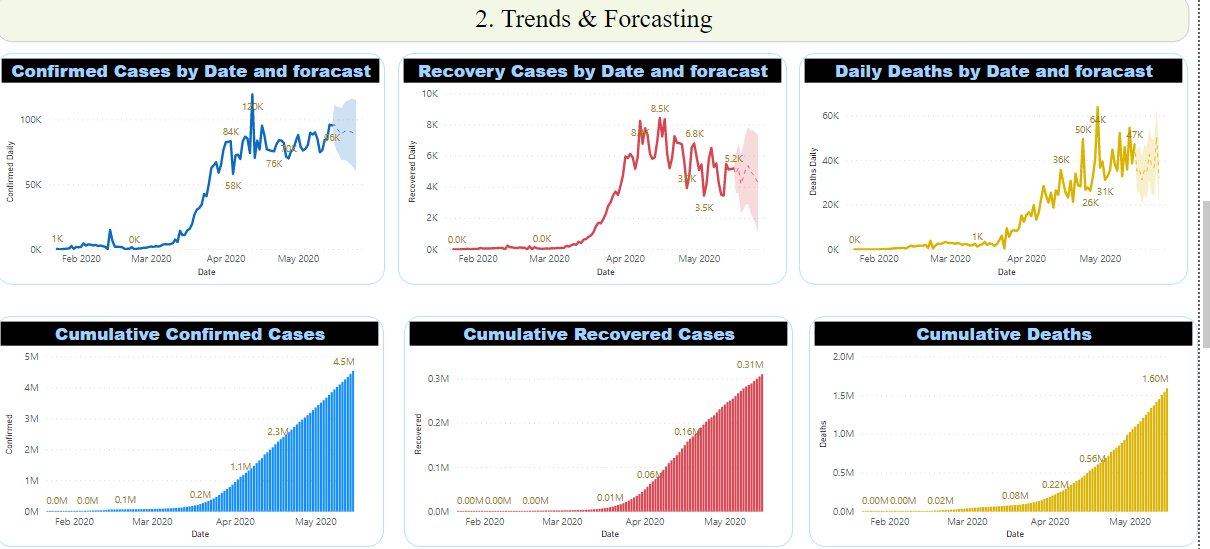
**You can also search for particular country which you wants to see the insights.(For ex. India)**



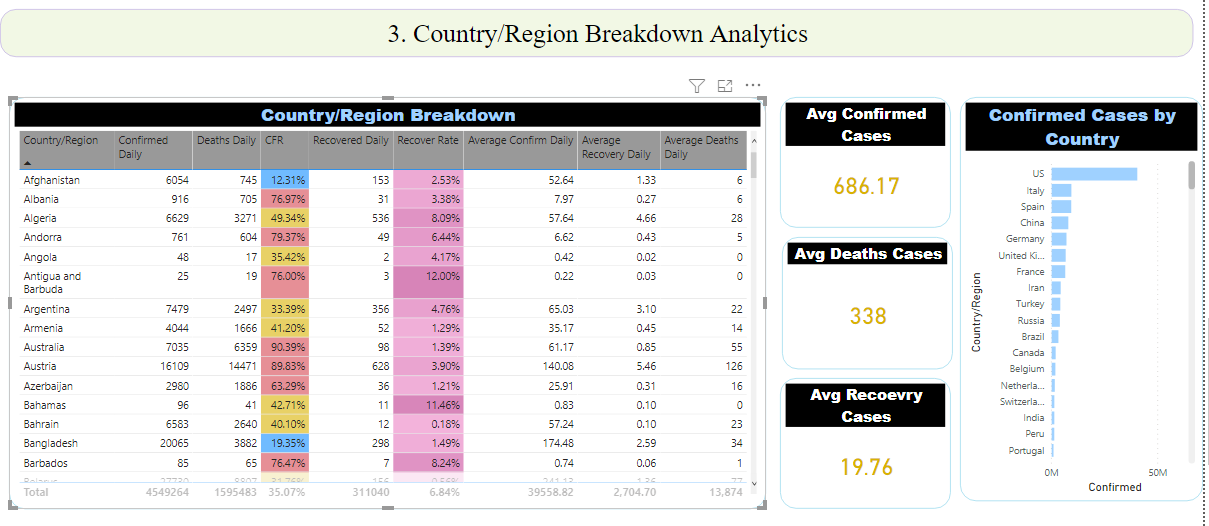
**Here you can also filter out content based on state, month, year and date.(For ex. March Month)**



**Now in below you can see trends and forecasting all over the world**



**Now you can see country/Region breakdown analytics**



**Insights & Analysis**

* The data is started from Jan 2020 to 15 may 2020
* From above datasets we can see that from starting to latest dates corona cases are Increasing day by day.
* Wheather it is confirmed case, recovered case or death case all cases are Increasing.
* We have visulized the data using power bi and created the dashboard
* From data visualization mostly effected country is US till may 15.

**CONCLUSION AND SUGGESTIONS**

We exploited the resources of the Python language and its libraries for our use and have managed to complete our main objective of Exploratory Data Analysis of the Covid-19 Dataset.

The comprehensive analysis done for the world as well as specifically for India gives us enough information, which can help us to conclude and trace some very significant insights and conclusions.

The maps used are interactive in nature and dynamic, which make them more interesting but at no point have we compromised with the information in order to accommodate the aesthetics.

Any peculiar or not expected trend has been examined in order to make it more readable and understandable for the readers.

Now this alone might be insignificant but the in depth analysis that we have done breaks it down to the core so as to get the maximum information available from the dataset and hence use those for Predictive and finally prescriptive analysis which can help the Governments and Private organisations. We can update latest data set and can check the updated value also.

**REFRENCE**

1. <https://towardsdatascience.com/global-terror-and-the-eda-visualization-rabbit-hole-a953c998b7e9>
2. <https://www.youtube.com/watch?v=tzRPDQ4uayQ>
3. <https://www.kaggle.com/sudalairajkumar/covid19-in-india>

**PROJECT CODE LINK**

1. <https://github.com/Grishmashah121/Covid-19_DataAnalysis>
2. <https://drive.google.com/file/d/1BwyvjBYllh43m9UdsuUoEt1RQAzRTcqL/view>