MTH 765P Mini-Project:
John Hopkins Covid-19 Daily Reports

Kasi Viswanath Chennupati – 200653723

1 Introduction

The reference to the John Hopkins Covid-19 data is found through the kaggle Covid Data Challenges Data.

The data for Covid-19 cases are maintained by the John Hopkins University github repository.

The data set used for the Data Analysis is the John Hopkins Covid-19 Data repository. The data is downloaded from the John Hopkins github data repository

The Covid Cases are documented and reported by the JohnHopkins University as Daily Reports and Time Series Data sets.To analyze the covid data and monitor the spread of the covid around the world.

2 Scope

To Analyze the Covid-19 confirmed cases in the world and the Top countries at risk.

3 Requirements

In-order to successfully conduct the analysis there are dependencies and requirements to achieve.
!Disclaimer: The Code runs smoothly when executed the .py file rather than jupyter notebook.

The Required Python3 Modules

- Pandas
- Numpy
- Plotly
- OS

The Required modules for the analysis work are installed using the pip3 tool in the Jupyter notebook.

Code :

```
!pip3 install numpy pandas  plotly
```

The installed modules need to be imported in the Jupyter python environment to start working on the data.

Code:

```
import numpy as np
import pandas as pd
import plottly.express as px
```

4 Data Extraction and Description

The Covid19 data from the John Hopkins Data from the John Hopkins University Github Data-Repository

Download the Data from the following link:

JHU_CSSEGIS

JHU_CSSEGIS_COVID_19_daily_reports

The Data is downloaded and saved in the local Current Working Directory in a new folder Data Source.

But the data in the Daily reports folder are CSV files.

The files naming convention:

"MM-DD-YYYY.csv"

The Data contains the csv files of daily reports

They contain the information like

Province_State : Province, state or dependency name

Country_Region : Country, region or sovereignty name. The names of locations included on the Website correspond with the official designations used by the U.S. Department of State

Last_Update

'Confirmed : The number of confirmed covid cases

Deaths : The number of deaths occurred due the covid

Recovered : Number of people recovered from the covid

Active : The current number of active covid cases

Combined_Key: The country and province columns concatenated

Incident_Rate: cases per 100,000 persons.

Case_Fatality_Ratio : Number recorded deaths / Number cases

Latitude : The latitude of the location

Longitude: The longitude data of the location

Since the files follow same naming format and type .The data can be loaded in to a data frame through a loop. The file date column created is not in date time datatype, The File_Date column needs to be converted into the Date-time format/type

```
files = [file for file in os.listdir(path) if file.endswith('.csv')]
        covid_data = pd.DataFrame()
        for file in files:
                current_file = pd.read_csv(filepath_or_buffer = path+file,skip_blank_lines=True)
                current_file["File_Date"] = str(file).strip(" ").strip(".csv")
                covid_data = pd.concat([covid_data,current_file],sort=False)
         covid_data.to_csv("./DataSource/allmonths_data.csv")
```

5 Data Cleaning and Processing

The covid_data data frame is analyzed and identified that the columns Country

There are some duplicated columns due to the change in the naming of column names.
The pandas concat function created a separate column for such scenarios despite the having the same data.

The columns to work and enrich are:

ProvinceState ---------Province/State
Country_Region ------Country/Region
Latitude -----------------Lat
Long ----------------------Longitude

There is also a number of null values identified in the above duplicated columns and the duplicated columns and null values are managed through fillna() method of pandas.

Example:

```
covid_data["Country_Region"] = covid_data["Country_Region"].fillna(covid_data["Country/Region"])
```

for all columns
The columns Active, Confirmed, Deaths and Recovered are loaded into the dataframe as float data type and needed to be loaded as int type. This column type issue is resolved using the astype method of the pandas dataframe method.

Example:

```
covid_data['Active'] = covid_data['Active'].astype(int)
```

For the current analysis we need fewer columns from the covid_data data frame.

also future analysis we add a few enrichment columns like “File_month_year” and “File_Quarter”

These columns are created using the pd.DatetimeIndex and to_period methods from pandas.

And for the current analysis of identifying the top affected countries we need the columns like Active, Confirmed, Deaths, Active, Recovered, Country_Region and File_Date we create a new data frames codif_data_enr and Ranks .

To clearly understand the top affected countries we need to find out the top countries in the Ranks dataframe

Example:

```
Ranks['Confirmed_%']=round((Ranks['Confirmed']covid_data_enr['Confirmed'].sum())*100,2)
```

after the ranking process we see that

'United Kingdom','Spain','France','Russia','Brazil','India','US' are the countries that are most affected.

We create a new data frame covid_filt that contains the data just for the above mentioned top 7 contries.

We have all the data transformations and filtered data frames we need for the visualization.

1.We plot the data for the covid cases progression around the world over time

2. We plot the data for the Top7 countries progression around the world over time

3. A Bar plot of the Top 7 countries over time.

6 Results

From the figure: 1.1 heat map the color scale from 40-80 colored areas we see that the Covid-19 cases are severe in the Europe, America and Asia zones and this confirms the top 7 countries we identified in the analysis phase.

Once the data is plotted and start the animation. We can observe from Figure: 1.1 that after the two months of observation, the most of covid-19 cases are concentrated in the China region.

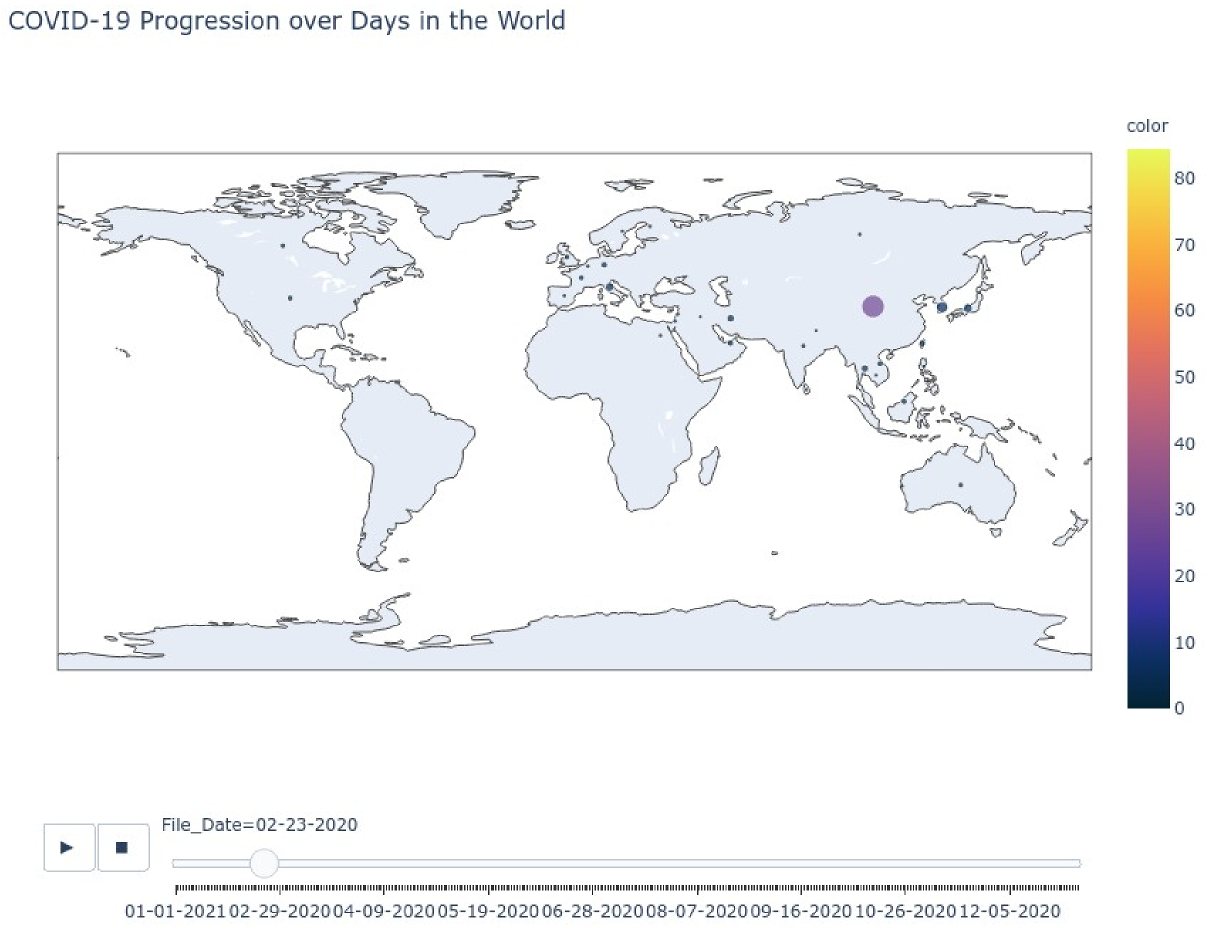

Figure: 1.1

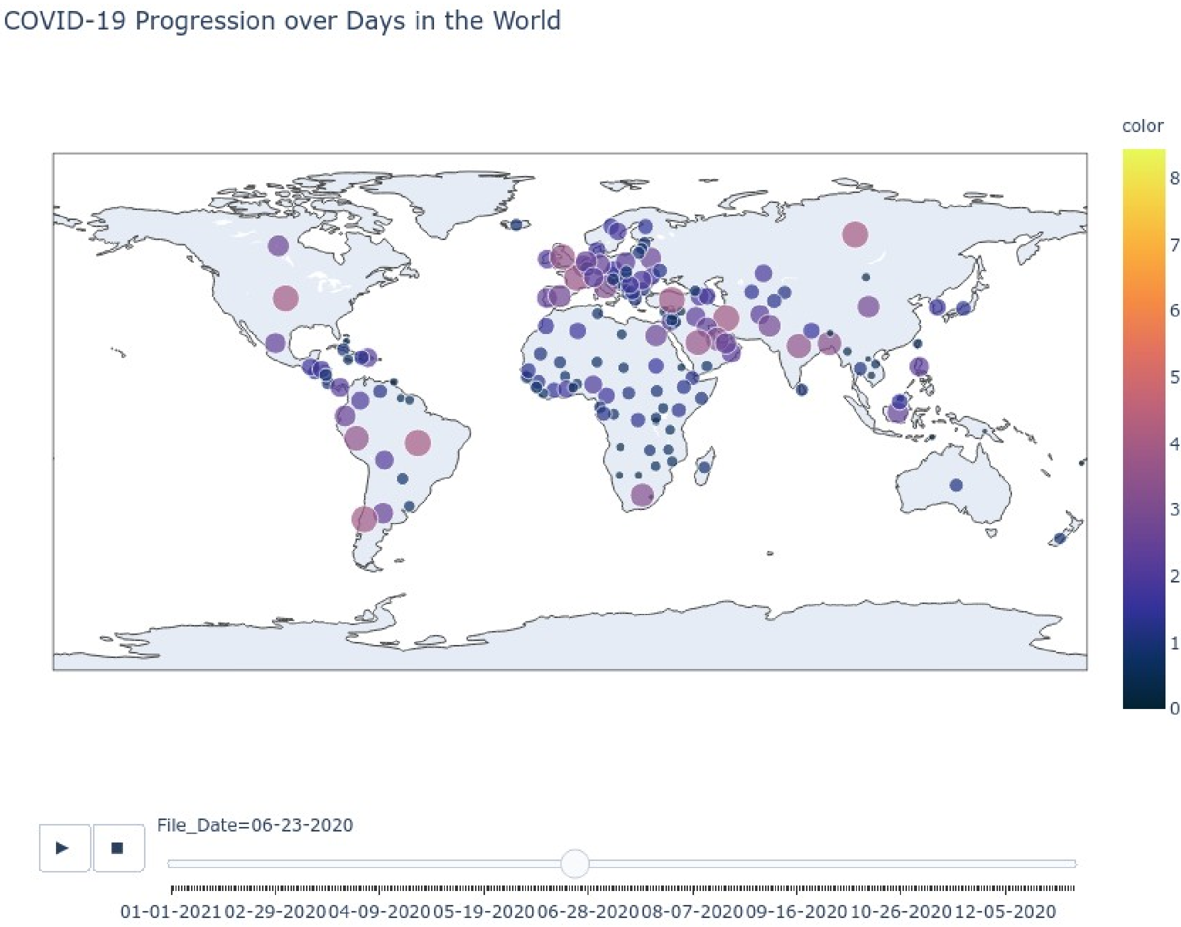

Figure: 1.2

We can observe from Figure: 1.2 that after the six months of observation, the most of covid-19 cases are not concentrated in the China region anymore instead the cases are spread across the world.

We can also observe that the rise of the cases in the china region when compared with the rest of the world the cases in the china region have slowed down.

We also observe that the Europe, Rest of Asia , Middle East, Africa and America region cases have raised exponentially.

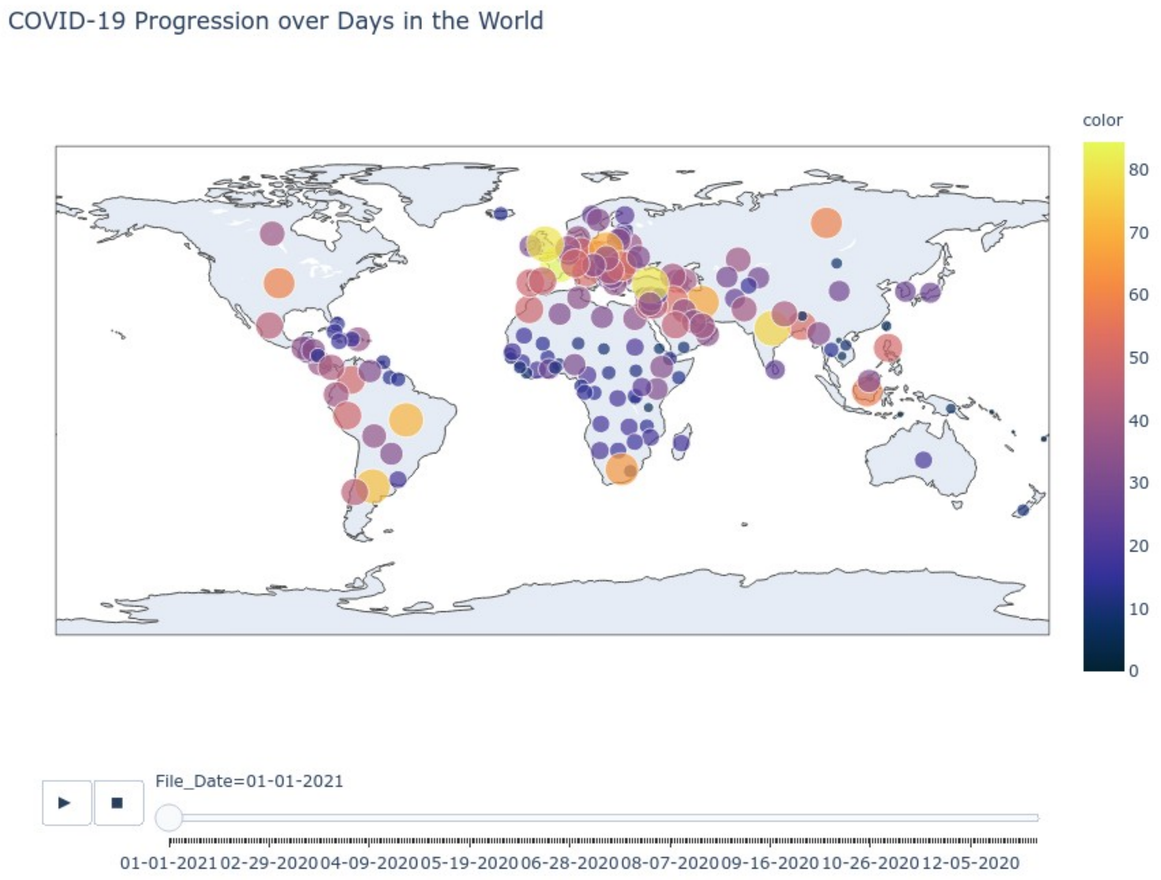

Figure: 1.3

We can observe from Figure: 1.3 that after a year of observation, the most of covid-19 cases are not concentrated in the china anymore as the cases are more in the America, Europe and Asia region excluding china.

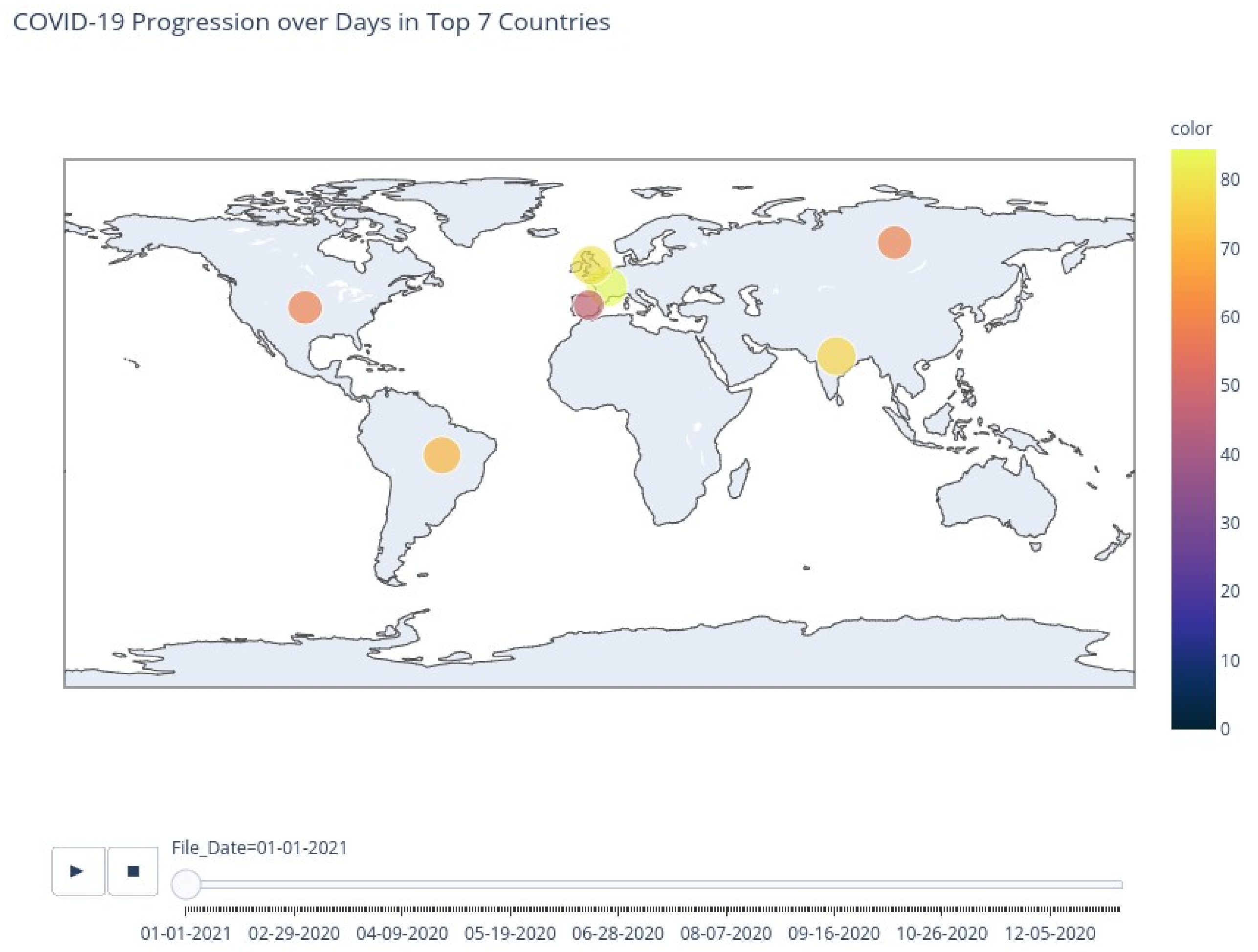

Figure: 1.4

We can observe from the Figure: 1.4 that the United Kingdom, Spain, France, Russia, Brazil, India and US are the individual nations that are most affected as they all fall above the 40 on the color scale.

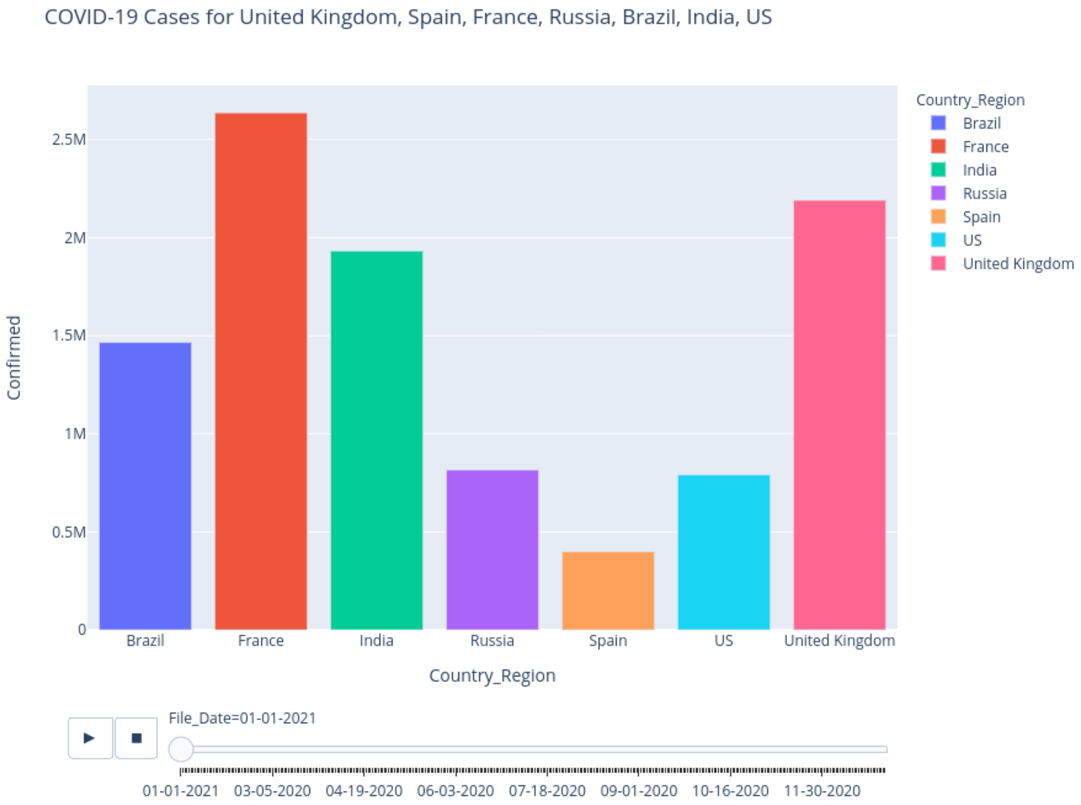

Figure: 1.5

Finally we can see that the Top 7 countries affected the most are in the range of 0.5 million to 2.5 million cases
The most affected countries are France and United Kingdom.

References

COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University
[https://github.com/CSSEGISandData/COVID-19]