Baby Shanthi Amaasa
11376627

ANALYSIS AND VISUALIZATION ON COVID-19

I. INTRODUCTION

Coronaviruses (CoV) are an enormous group of infections that cause sickness starting from the normal virus to progressively extreme ailments. Coronaviruses can be transmitted among creatures and individuals. The infection, which was first found in the Chinese city of Wuhan, the capital of China's Hubei province, and has since spread globally, resulting in the ongoing 2019–20 coronavirus pandemic [1]. The official name has been given by WHO as coronavirus disease (COVID-19) and virus as severe acute respiratory syndrome coronavirus 2 (SARS-COV-2). This virus has spread rapidly from one city to the whole world, analysis and visualization of the coronavirus data help to know the spread of the disease and to gain insights.

II. RELATED WORKS

Many organizations have developed interactive dashboards and visualizations on COVID-19. CDC has provided the data and visualizations about the weekly summary of cases, testing, age groups, forecasting [2]. It says that out of total cases 6,814 cases in the United States are travel-related and 14,728 are due to close contacts. WHO provided dashboards on global data and analysis for every region and regions with the highest cases [3].

III. METHODS

A. Visualization Tool

The tool used for visualization is Power BI. It is a business analytics service provided by Microsoft which provides interactive visualization. Power BI provides custom visuals besides general visualizations. Multiple reports can be created in one Power BI file where we can use drag and drop functionality to create the report. Data size is limited to 2 GB. This tool is well suited when the user has security requirements and pricing limitations. And power query has a log of applied steps where we can insert or delete the steps in between at any time.

B. Data Set and Cleaning

The data set is collected from multiple resources. The dataset named COVID_19_Data is collected from Kaggle [4] about deaths, confirmed, and recovered cases for each country. It constitutes of attributes named in table 3.1. The data is from 22nd January to 20th April 2020 with 23,580 observations.

Table 3.1. Attributes of COVID-19_data set

S.no	Attributes
1	Province/State
2	Country
3	Longitude
4	Latitude
5	Date
6	Cumulative_Confirmed
7	Cumulative_Deaths
8	Cumulative_Recovered

The second data set is collected from Our World in Data [5]. This data is the merge of three different data sets and it is about total and daily confirmed cases and deaths in each region. The data sets are merged using a merge query in a power query. The data is from December 31st2019 to April 21st2020. Total observations are 13,994. The data set is named as total and daily cases and consists of attributes listed in Table 3.2.

Table 3.2 Attributes of total and daily cases data set

S.no	Attributes
1	Region
2	Date
3	Total confirmed cases
4	Total confirmed deaths
5	Daily confirmed cases
6	Daily confirmed deaths

The third data set is collected from Our World in Data [6]. It is about the tests performed in each region. The data is from January 21stto April 19th2020. Total observations are 2,212. The data set is named as Test data and consists of attributes named in Table 3.3.

Table 3.3 Attributes of the test data set

S.no	Attributes
1	Region
2	Date
3	Total tests
4	Daily change in total tests
5	Number of tests per confirmed case

Data Cleaning is performed on all the data sets. In the first data set the missing values are replaced with N/A value and in second and third data sets the unnecessary column is removed. In all the data sets the data type of columns has been corrected and columns have been renamed.

C. Exploratory Data Analysis

Exploratory data analysis is used to summarize and analyze data and to see what data can tell us beyond hypothesis testing. EDA is performed on the data sets to

see how it will be useful for the hypothesis and to know the spread of the cases globally.

Spread of cases globally in the map has been visualized as shown in fig 3.1 which shows that Unites States, Italy, Germany, Spain, France are the most affected countries with the disease. It has been visualized in a way that the spread is represented with the color for each region. That is a darker color for the most affected region and a lighter color for the least. The scale for the coloring has been represented. The attributes required for this are confirmed cases and country. Time slicer has been included to see how the spread varies across a period and background color has also been added to make the visualization better. ArcGIS maps have been used to visualize this.

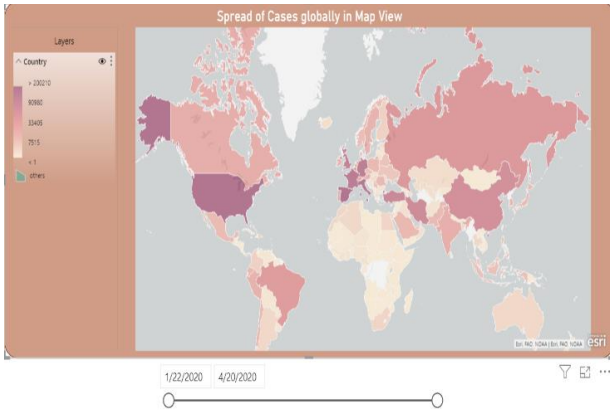

Fig 3.1. Spread of cases globally in map view

Fig 3.2. Zoomed out a picture of layers

Time series for Confirmed, deaths, recovered cases using line chart as it is suitable for time series. The attributes required for this are date, cumulative confirmed cases, cumulative deaths, cumulative recovered cases. The cases have been increasing with time for all confirmed, death, and recovered. We can also observe that there are more recovered cases than deaths. The total number of cases till April 20th globally is 2472253, deaths are 169985, and recovered are 633181.

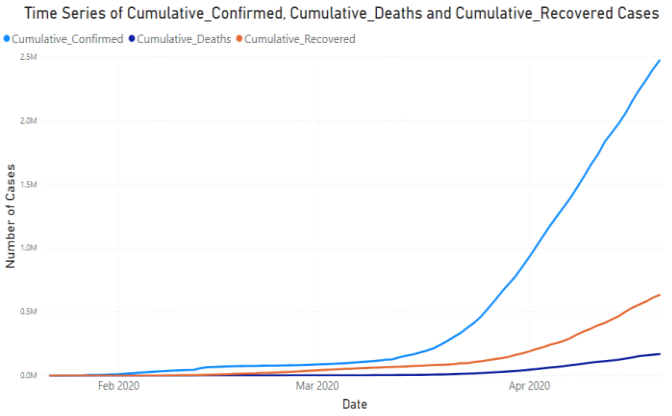

Fig 3.3. Time series for confirmed, recovered, deaths

Confirmed, deaths, recovered country-wise using stacked area chart shown in fig 3.4. Attributes required are cumulative confirmed, deaths, recovered, and country. As the values are cumulative maximum measure is used to consider the value for each country. The U.S is the topmost in deaths, recovery, and confirmed cases. The graph is sorted in descending order of confirmed cases. Therefore, it can be visualized with the top to least affected countries in three categories. The total number of cases in the U.S till April 20th is 784326, deaths are 42094, recovered are 72329.

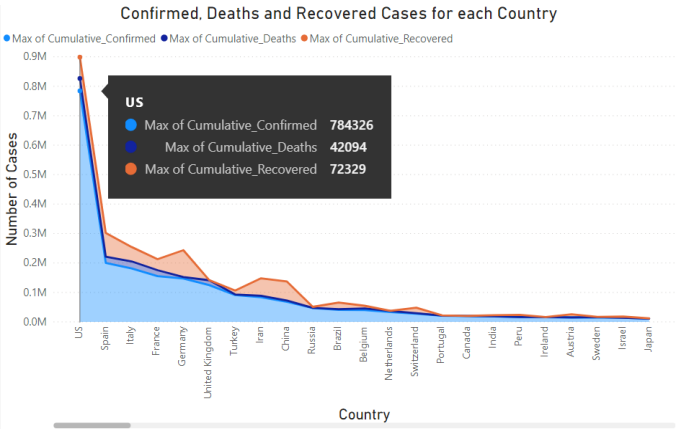

Fig 3.4. Confirmed, Deaths and Recovered cases for each country

Total and daily confirmed cases, deaths per million that are cases relative to the size of the population has been visualized as shown in fig 3.5 using line charts as it is preferable for time series. Total cases and deaths have also been visualized using the card that is total cases are 2 million approximately and deaths are one seventy thousand. Cases and deaths per million across the world are 311.99, 21.79. Daily new cases and deaths per million for April 21st are 9.75 and 0.67 while cases and deaths on April 21st are 76037, 5203. Daily new confirmed cases and deaths have been fluctuating repeatedly from April 5th to 21st 2020.

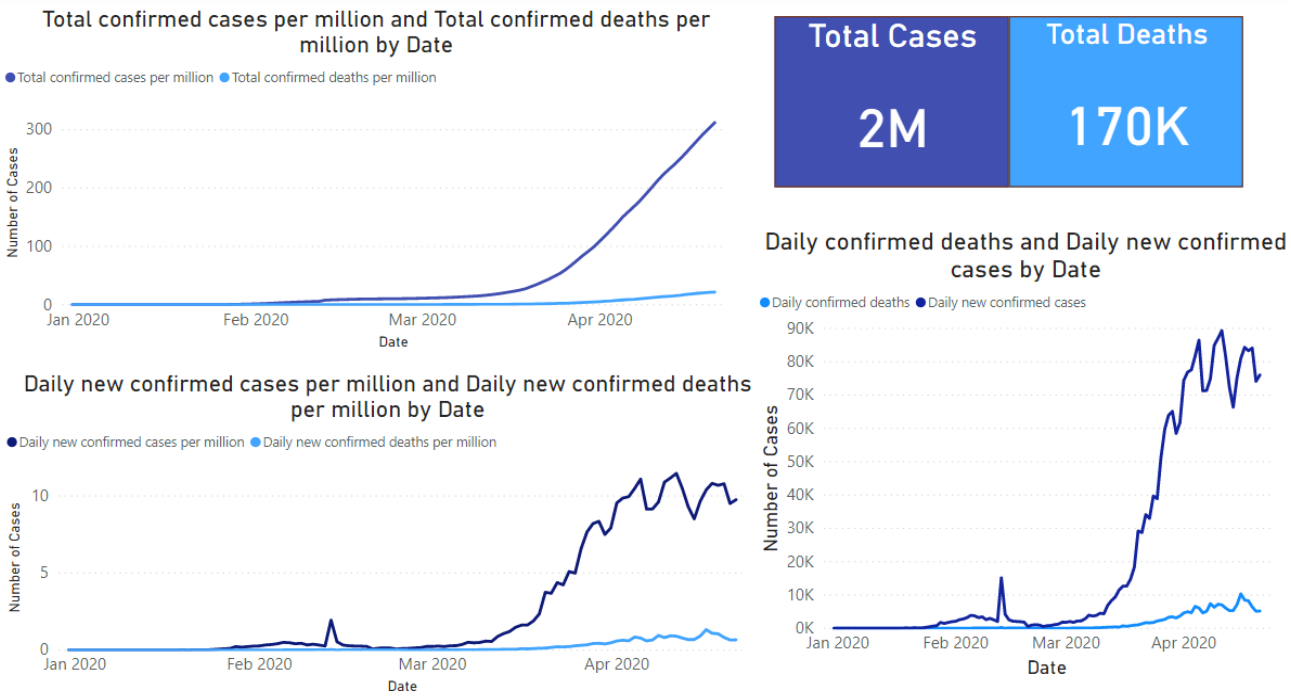

Fig 3.5. Total and daily cases and deaths per million

Daily change in the total tests performed by each country is visualized as shown in fig 3.6 using the line chart where the attributes required are the number of tests and dates. It shows that there is a high number of fluctuations from March 26th to April 13th, 2020. From April 13th to 16th it highly increased and dropped on April 16th to 19th. On April 19th total tests performed are 221,832.

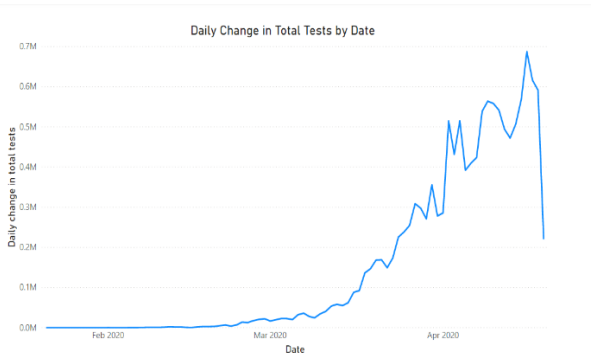

Fig 3.6 Time series of daily change in total tests

D. Hypothesis

- How the disease spread across the continents?
- Is the data about confirmed cases is reliable?
- What is the case fatality rate for the top five affected regions? Do the most effected regions have the highest case fatality rate?

IV. IMPLEMENTATION RESULTS

A. Model A

Total spread and comparison of the cases in each region Africa, North America, South America, Oceania, Europe, Asia excluding China, China across time. Asia has been visualized excluding China as the origin of the disease started from China to see its spread, it is considered as the separate region for this visualization. Attributes needed are region, date, and confirmed cases. This has been visualized using the ribbon chart as it is best to rank the regions and shows the change in percentage along with the original values and changes of ranks can also be observed quickly and clearly without confusion and stress during visualization. The regions are

filtered using the basic filtering in filtered tab and the interactive features are time slicer has been added to visualize the focus of spread for a period, the cursor can be placed in-between days to know the percentage difference and colors for the regions have been chosen attractively as the ranks keep changing and the user may lose the track of particular region along period.

Europe is the most affected region, which is represented as rank 1 followed by North America, Asia excluding China, China, South America, Africa, Oceania. But across the period, the rankings have been changing. China is rank 1 that is the most affected country till March 18th, the rank has been decreasing from march across the period. And Asia excluding china has decreased its rank from March 24th and increased again on March 31st. While all others have been increasing ranks across a period that is getting more effected with time.
It can also be seen from the output that the tooltip shows the percentage difference between consecutive dates from which we can observe the growth rate.

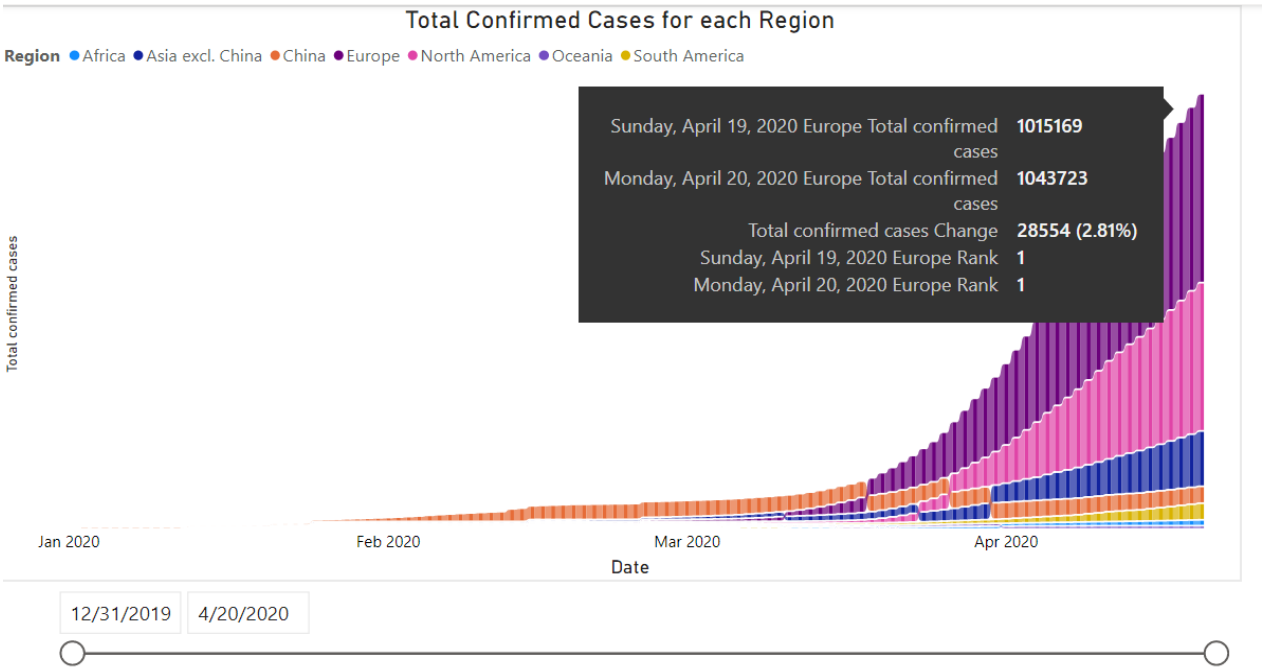

Fig 4.1 Total number of confirmed cases for each region

B. Model B

Test data is important to understand the progress of pandemic. Progress of the pandemic is mainly obtained by the information of confirmed cases that are tested by laboratories. Therefore, testing data is required to assess the spread.

Analyzing the total number of tests performed for selected countries using the line chart as shown in fig 4.2. Some interactive features are attribute slicer that is used to search and select the countries to be displayed and time brush have been used to navigate along period. The advantage of time brush over using default slicer is the time brush visualize the values, therefore we can easily navigate along with the values with time, for example, we can select the period where the trend is increasing or decreasing or having more number of fluctuations. Using attribute slicer, I have selected countries Unites States, Italy, Australia, South Africa, South Korea. Y-axis represents the total number of tests on a logarithmic scale and the x-axis represents the date. The logarithmic scale is chosen as we represent time series data and growth rate can be viewed easily. From the visualization, South Korea has started performing tests early followed by Australia, Italy, United States, and South Korea. By April 18th, 2020.

The United States has performed the highest number of tests compared to other countries. The colors to visualize the trends have been chosen in a way that the user can assess quickly. The background of the plot area has been set to the COVID-19 test image to make it more attractive.

To assess the spread of the confirmed cases we need to compare the number of tests and cases for each country. To make it easier we can visualize the values of the number of tests per confirmed case using a horizontal bar chart. Using the horizontal bar chart, it would be easier to visualize the values of countries with bar size. The values are sorted in descending order to obtain the highest test values to the least. The background color of the bar chart is chosen as the lighter shades of the bar and data labels also have a background color to make it visible and attractive as the data labels are the one which represents the values. The main advantage of the horizontal bar chart in this scenario is no need to compare the values using the y-axis, x-axis as the values represented right after the bar using data labels. Users can quickly grasp the insights without any complications. The attributes used for this visualization are region, total tests per confirmed case, and the date is filtered for April 18th and 19th.

From the graph, it can be visualized that Taiwan performs a greater number of tests per confirmed case on April 18th and 19th that is 134.19 followed by New Zealand and Australia. This may be due to a lack of right test kits or other factors. United states perform 5.27 tests per confirmed case. The value of the number of tests performed will be always greater than 1 because several people can be tested multiple times to confirm whether they are infected. If the value of the number of tests per confirmed case is less than 1 that represents that the spread of the disease in a country is biased and that means the value of confirmed cases is not true. But from the visualization, it can be observed that all values are greater than 1 therefore the data of confirmed cases are reliable.

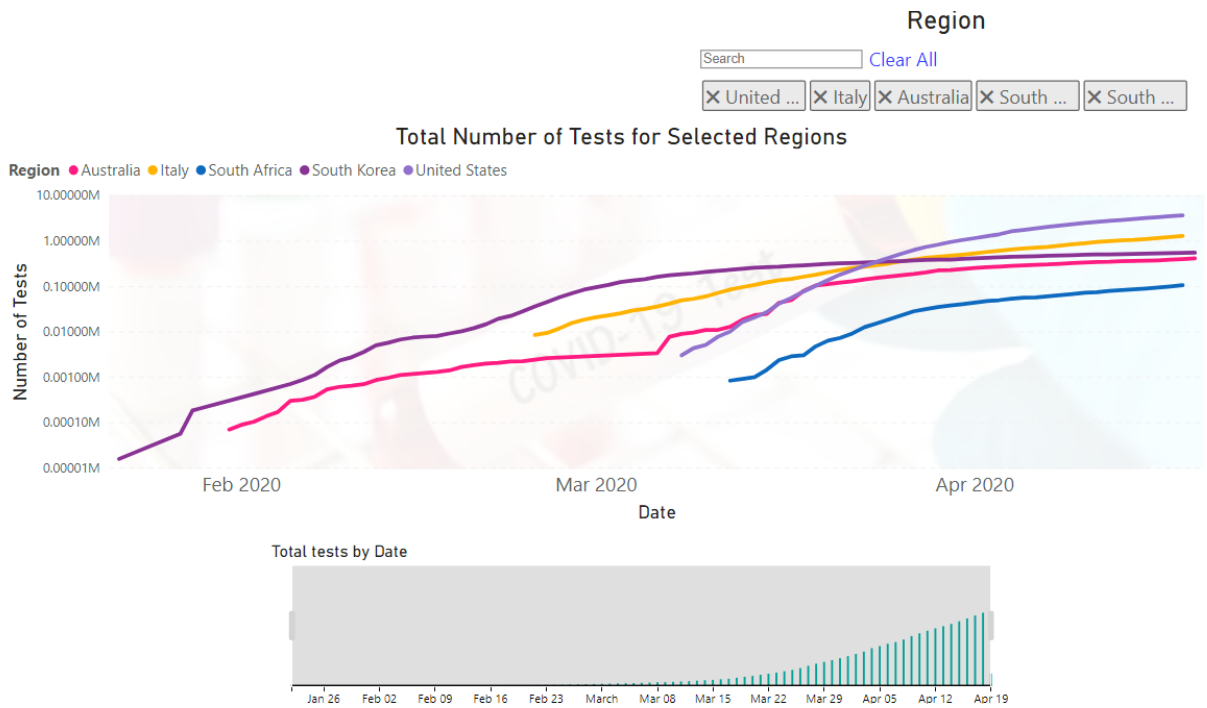

Fig 4.2 Total number of tests for selected regions

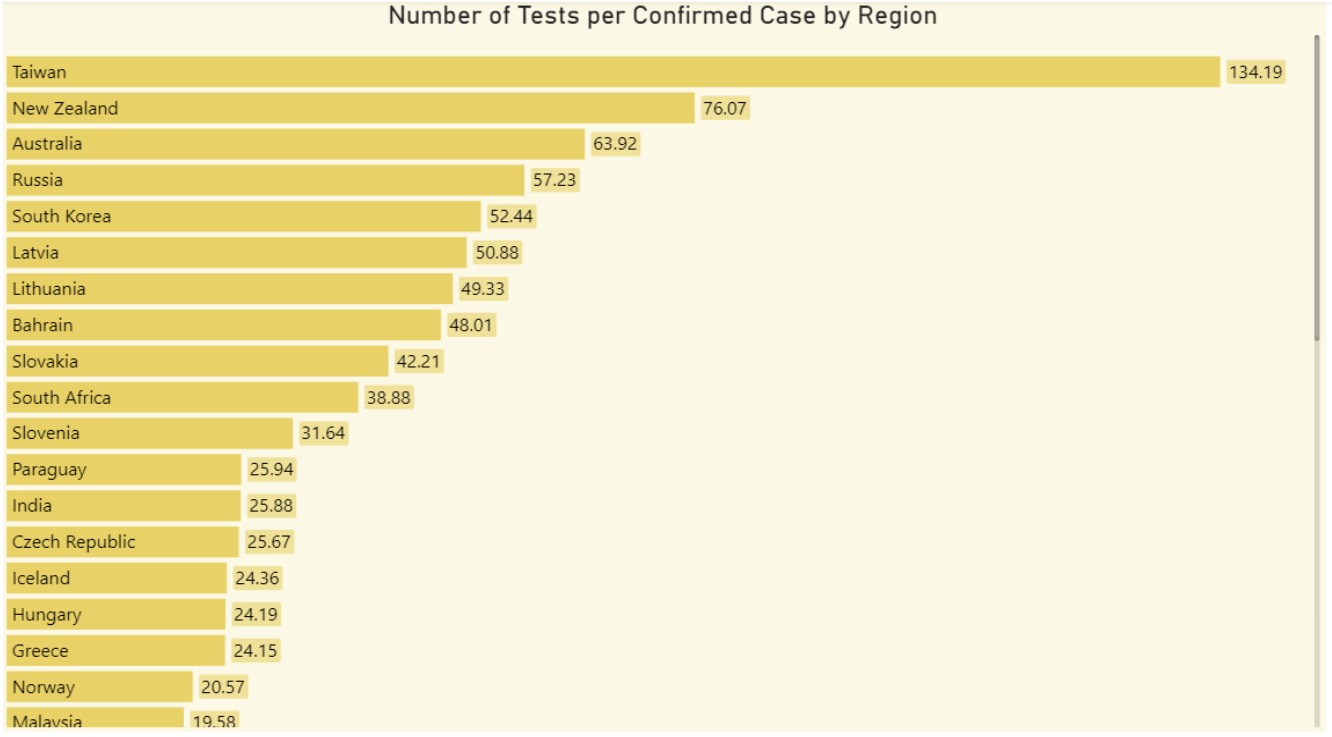

Fig 4.3. Number of tests per confirmed case

C. Model C

Case fatality rate (CFR) is the ratio of the total number of deaths to the total number of confirmed cases in percentage is calculated. This does not represent the actual infection death risk as we do not know the actual confirmed cases as all the people infected [7], dead are not tested, and all the dead people may not have diagnosed. Therefore, the calculating value is not constant, and it changes for period and location.

The case fatality rate for the top five most affected countries is visualized using line charts. The top five countries are filtered using the top n in advanced filtering. The time slicer has been added to navigate along period. The attributes required for this visualization are date, CFR, the total number of deaths, and the total number of confirmed cases. CFR is the measure created using DAX. The expression used is:

CFR=DIVIDE
(sum('Covid_19_Data'[Cumulative_Deaths]),
sum('Covid_19_Data'[Cumulative_Confirmed]))*100.

The attributes are first aggregated and then calculated. From the visualization, we can interpret that the top effected countries are U.S, Spain, Italy, Germany, France. The CFR has first started for France and last for Germany compared to other countries; from this, it can be said that the spread has first started in France among these five countries. The CFR of France has suddenly decreased from February 26th and increased slowly from March 18th, 2020. The Case Fatality Rate of Italy increased continuously from March 1st and it has the highest case fatality rate of 5.37% on April 20th. There are very few fluctuations in CFR in Spain and Germany and it kept increasing continuously. For the U.S the CFR has kept decreasing from March 4th and increased from March 24th.

Most affected countries need not have the highest case fatality rate. For example, the U.S has the highest number of cases till April 20th as per exploratory data analysis in fig 3.4 but does not have the highest case fatality rate in fig 4.4.

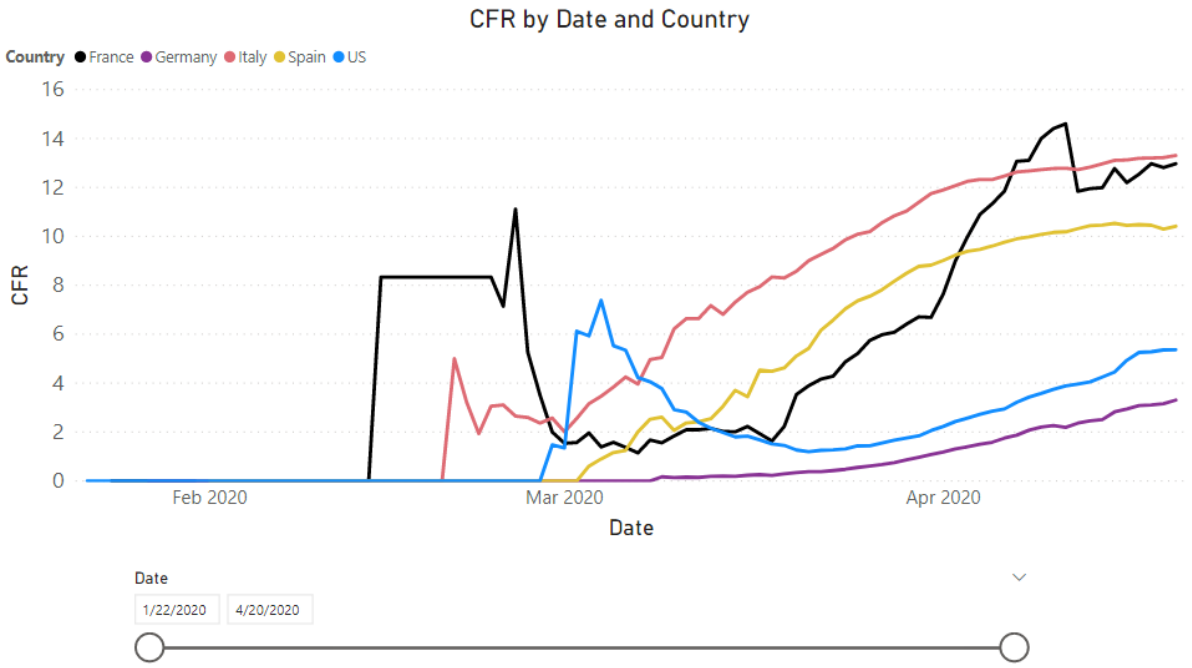

Fig 4.4 Case fatality rate for top five affected countries

V. DISCUSSION AND FUTURE WORK

From model A, users can extract the spread of cases across regions and the percentage of daily change using tooltip. The visual aspects of colors and ribbon chart, time slice makes it easier to track the region with the highest cases along period. From model B, the user can assess the reliability of confirmed cases by the glance of values in a horizontal bar chart and can also know the total number of tests performed in each region with interactive features of search option for the country to be chosen and time brush to choose a period. From model C, the user can see the trends of the case fatality rate of topmost affected countries and can use time slicer to navigate through days, weeks, or months.

Future work includes obtaining the data related to a total number of cases to calculate and visualize the infection fatality rate. Also, data about the age group and prior health conditions of infected people can be used to asses the risk of deaths.

REFERENCES

[1]https://en.wikipedia.org/wiki/Coronavirus_disease_2019

[2]https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/previouscases.html

[3]https://www.who.int/emergencies/diseases/novel-coronavirus-2019

[4]https://www.kaggle.com/imdevskp/corona-virus-report

[5]https://ourworldindata.org/coronavirus#confirmed-deaths

[6] https://ourworldindata.org/covid-testing

[7] https://ourworldindata.org/covid-mortality-risk