# Data Analysis Report On COVID 19

**Prepared by :**

**Shivani Sheth**

**shivani29111998@yahoo.com**

# Table of Contents

# Acknowledgement

I would like to take this special opportunity to express my deep sense of gratitude to all those people without whom this project wouldn't have been possible. Most importantly, I would like to thank my parents for their inexhaustible source of inspiration and motivation, without whom this project would have been challenging.

I would also like to extend my gratitude to the makers of Tableau Public for providing me with the right platform to visualize data easily and effectively. The platform is highly user friendly and consists of powerful tools and visualization techniques which helped me understand its working very smoothly. I would also like to thank the Johns Hopkins Whiting School of Engineering for providing the CSSE COVID-19 Dataset to the public for analysis on the recent COVID 19. The Python platforms, such as Jupyter Notebook also facilitated in merging the data files into one combined file.

Lastly, I would like to thank my friends for their valuable feedback which was very helpful in the various phases of the completion of the project. The feedback and guidance ensured that this project was made scrupulously through the reiteration of the modules and charts of the projects.

Shivani Sheth

14 March 2020

# Chapter One : Introduction

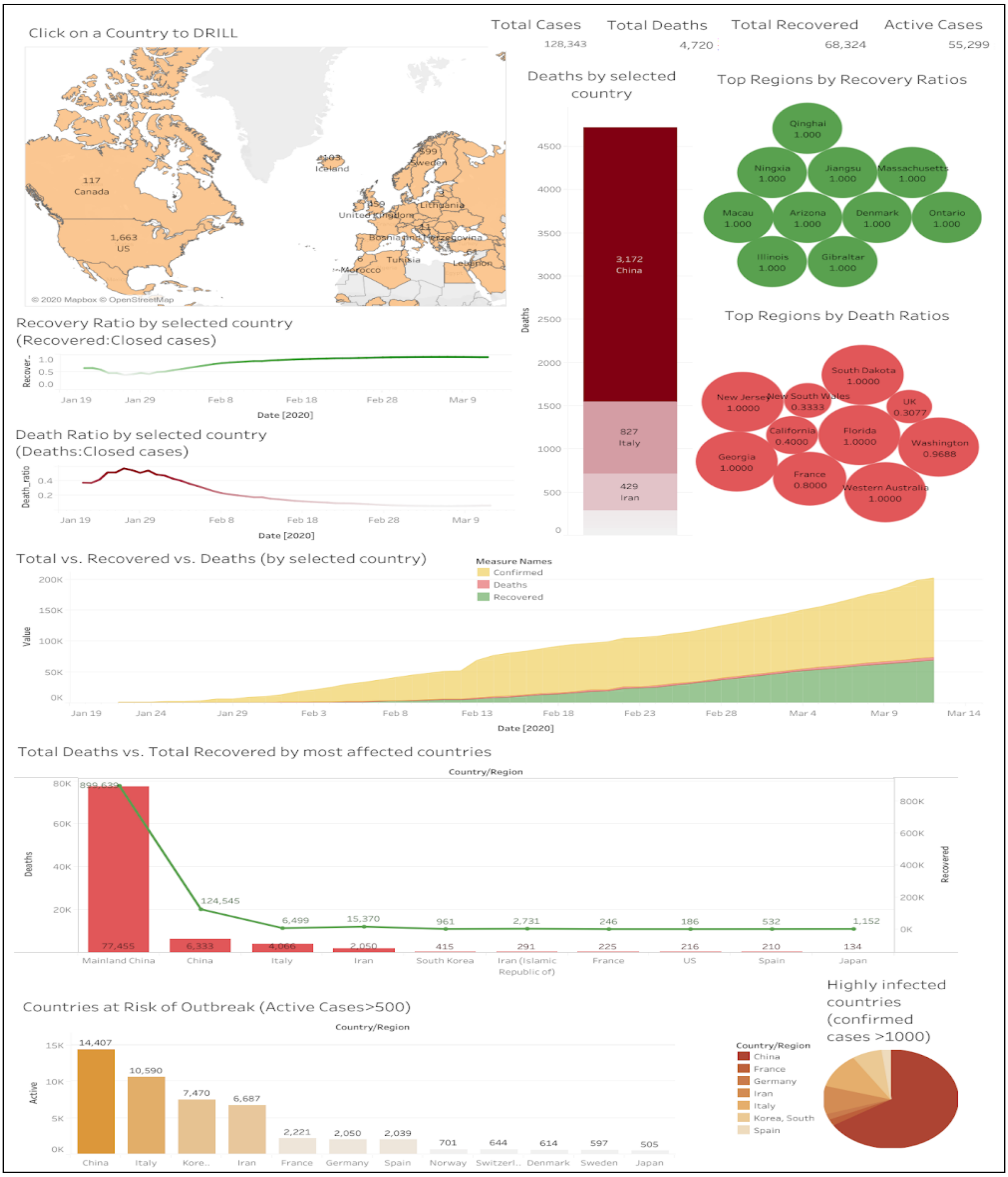
COVID-19 is the infectious disease caused by the novel virus known as the Coronavirus. It originated from Wuhan, China, in December 2019 and is one of the most infectious diseases of all time. The disease rapidly spread, resulting in an epidemic throughout China, followed by an increasing number of cases in other countries throughout the world. In February 2020, the World Health Organization designated the disease COVID-19, which stands for coronavirus disease in 2019. Since the first reports of cases from Wuhan, a city in the Hubei Province of China, at the end of 2019, more than 80,000 COVID-19 cases have been reported in China, with the majority of those from Hubei and surrounding provinces. A joint World Health Organization (WHO)-China fact-finding mission estimated that the epidemic in China peaked between late January and early February 2020, and the rate of new cases decreased substantially by early March.

As of present (14th March 2020), there are 156,099 confirmed cases, 5,819 deaths, and 72624 recovered cases, as quoted by the World Health Organization (WHO). Person-to-person spread of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is thought to occur mainly via respiratory droplets, resembling the spread of influenza. With droplet transmission, virus released in the respiratory secretions when a person with infection coughs, sneezes, or talks can infect another person if it makes direct contact with the mucous membranes. The infection can also occur when a person touches a surface or object that has the virus on it, then touches their eyes, nose, or mouth. The incubation period for COVID-19 is thought to be within 14 days following exposure, with most cases occurring approximately four to five days after exposure. The disease causes respiratory illness (like the flu) with symptoms such as a cough, fever, and in more severe cases, difficulty in breathing. Severe illness can occur in otherwise healthy individuals of any age, but it predominantly occurs in adults with advanced age or underlying medical comorbidities

Infection control to limit transmission is an essential component of care in patients with suspected or documented COVID-19. Currently, there is no specific medicine to prevent or treat the coronavirus disease and the best prevention here is a precaution. Hence, there is a strong need to spread awareness among the masses through figures and interactive charts. In this way, the dashboard made on Tableau here is an effective medium to enlighten individuals about the effect of COVID 19 all over the world.

# Chapter Two : Overview

The data source used for this project is the CSSE COVID-19 Dataset by Johns Hopkins CSSE. For the visualization of the data, this project utilizes the Tableau platform due to many reasons. First and foremost, Tableau is not only user friendly but also has some great tools for connecting different charts to a global worksheet. It can also combine two or more datasets, which is quite helpful while using different types of data from different tables or sources.



The interactive dashboard as shown in the figure above uses two types of datasets. First is the combined dataset defined by the named 'Combined' in this project, and is made by integrating the original data files from 22nd January 2020 to 12th March 2020 using Python to merge CSV files. It has been used in the project to plot charts with respect to each day to show an increase or decrease of a trend. Second is the dataset consisting of data on the last day of the dataset used, which is 12th March 2020, defined by '03-12-2020' in this project to plot and compare total data country-wise. This is used to contrast a trend between different countries and identify the most affected countries. The Key Performance Indicators (KPIs) used by the dashboard are the total number of confirmed cases, deaths, recovered cases, and active cases. The recovery ratios and death ratios are also computed to display relative data between different provinces of the world.

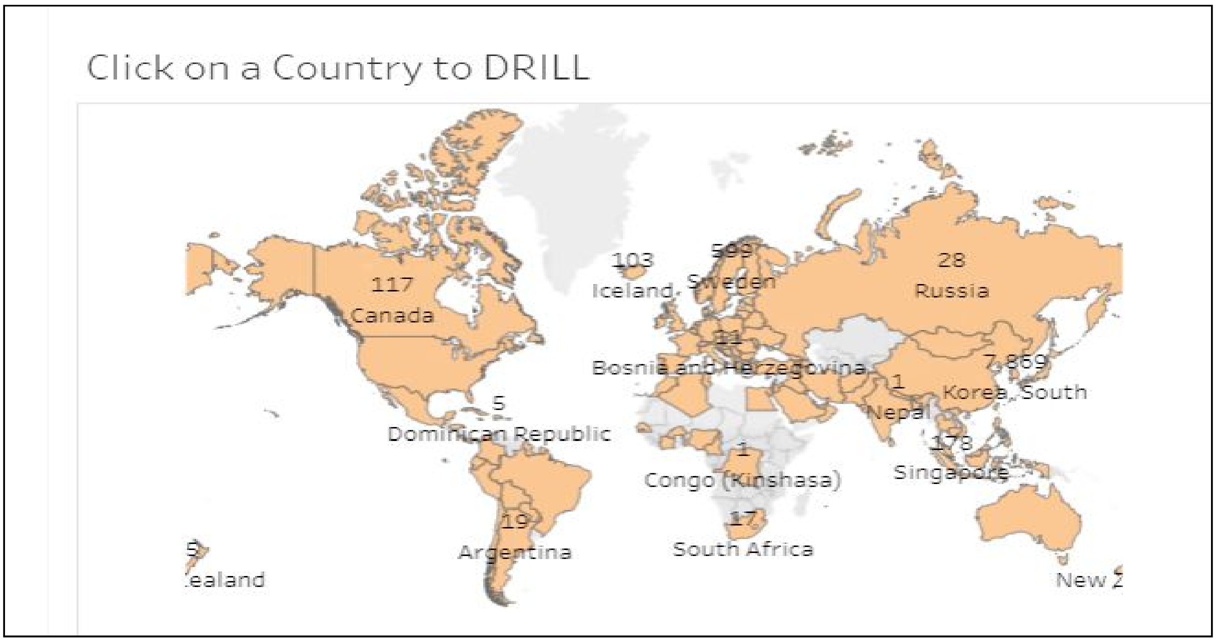
The dashboard created is an interactive one, which means that the user can click on a given country and the information related to the country will display on the other worksheets added. This is known as 'drilling down' to specific data of a country when a given country is selected by the user. On the overview of the dashboard given below, the first chart that a user can see on the dashboard is a world-wide map attached to the top left-hand side of the dashboard. This map encompasses all the countries of the world in a two-dimensional figure. The user can choose a country, to observe the detailed information of the chosen country in the neighboring charts. Upon selecting a country, its total confirmed cases, deaths, recovered and active cases are shown at a glimpse on the tiles right next to the world map on the top left corner of the dashboard. These give an overview of the effect of COVID on the given country. Next is a bar chart that shows the total number of deaths in the given country. On the extreme right are the highest recovered ratio and highest death ratio given by the provinces in the country. These give an overview of the most and least affected provinces in the country. Below the World Map, are the total death and the total recovered ratios of the country given on each day of the given period plotted on line charts. This shows us an increase or decrease in the deaths or recovered cases over time. Below that is an area chart that displays the confirmed cases, the deaths, and the recovered cases of the given country sorted by date. Hence it shows us the rise or fall of these cases for a given nation. These are the charts that are drilled down based on a specific country. If no country is selected, then the charts display the overall values worldwide, that is, the combined values of all the countries in the world.

The next graph that follows is a mixed chart that plots the total number of deaths and the total number of recovered cases for different countries, on two different axes of the same diagram. This compares the difference between the two rates among the nations most affected by COVID 19. The bar graph on the bottom left of the dashboard displays the total active cases for countries having more than 500 active cases and hence at a risk of an outbreak. The last chart on the bottom right displays the number of confirmed cases in highly infected countries on a pie chart.

Detailed analysis of each chart, along with its dimensions and measures, is expounded in the next chapter.

# Chapter Three : Detailed Analysis of Each Graph

## 1. World Map by Confirmed Cases




**(Dimension: Longitude; Measure: Latitude; Aggregation: SUM; DataSource: 03-12-2020)**

In the above chart, a world-wide map is plotted using the Latitude and Longitude columns as given in the data source ‘03-12-2020’ and a data label with the total confirmed cases are plotted, that is, each country also displays the total confirmed cases along with the country name. The user can click on a specific country to drill down the other charts based on the selected country, and hence get an insight into its specific data. The charts that will show data related to the specific nation selected are: The total number of cases, total deaths, provinces by highest recovery and death ratios, day-wise recovery and death ratios, and the day-wise total number of confirmed, death and recovered cases. These are essential to identify the provinces that have the highest death rates and take immediate measures or supply more resources to those provinces in the country. The total number of confirmed cases can also recognize the stage of the outbreak that the specific country has reached. Lastly, the day-wise data of the total number of deaths, recovered, and confirmed cases help in observing the spread of the disease in the country. It can help in deciding if the measures taken so far have helped the country and if extra measures are needed to contain the disease.

## 2. Total Confirmed, Deaths, Recovered, and Active cases.


**(Drilled by Selected Country; DataSource: 03-12-2020)**

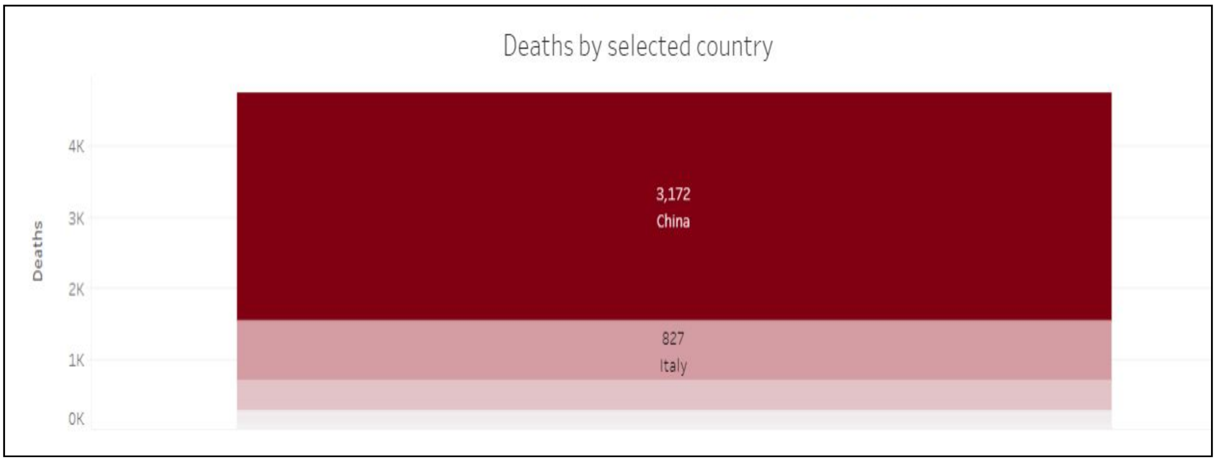
After selecting a specific country, the first thing a user sees is the KPI tiles at the top right of the dashboard. These KPI tiles are individual worksheets that hold the total confirmed cases, total deaths, total recoveries, and total active cases. The sum of confirmed cases is added in the text mark, which displays the total confirmed cases below the respective heading of each file. These tiles represent the data of the respected selected country in numeric values, to get an overview of the spread of the disease in the country. Here, active cases are the cases that haven't reached an outcome (death or recovery) yet. They are calculated by the number of confirmed cases subtracted by the closed cases. If no country is selected, then it will show the worldwide total values (aggregated values of all the countries).

**Calculations:**

Total Active Cases = Sum(confirmed) - ( Sum(deaths)+Sum(recovered) )

Where the latter represents the <u>Total Closed Cases</u> or the cases with a final outcome (either death or recovered).

## 3. Death Split Up by Country

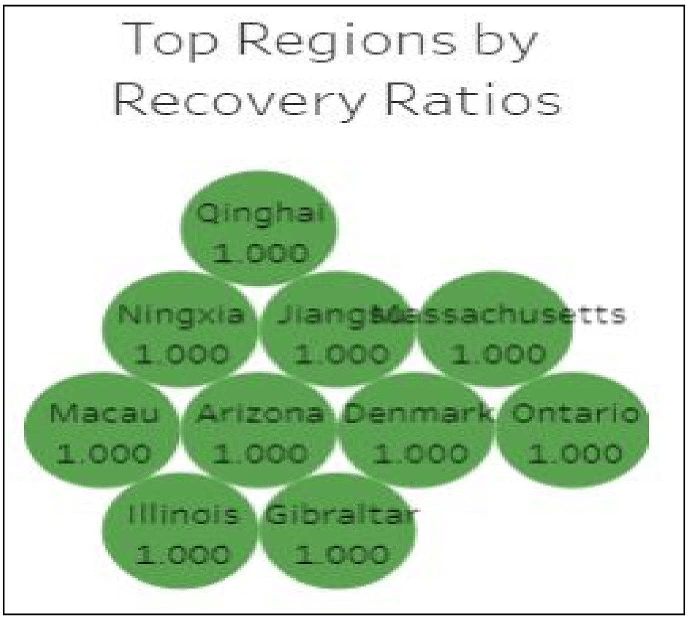



**(Dimension:Country; Measure:Deaths; Aggregation: SUM; Drilled By Selected Country; DataSource: 03-12-2020)**

In the above KPI, a column chart is used to represent the number of deaths in different countries across the globe. The graph is sliced by different countries to give a visualization of the extent of deaths and a comparative measure between the various nations. Also, the magnitude of deaths is represented by different shades of red to understand the intensity. This was done by setting the sum(deaths) variable to the color mark. Hence, the nation with the highest number of deaths will be viewed in a darker shade of red as compared to the countries with a lower number of deaths.

This chart will also get drilled based on the selected country from the map and show the death toll for the specific country. Hence one can identify the countries where the virus has proven to be deadly.

## 4. Top 10 Provinces with Highest Recovery Ratios




**(Dimension: Province/region; Measure: recovery_ratio; Aggregation: none; Drilled By Selected Country; DataSource: 03-12-2020)**

In the above KPI, a packed bubbles chart is used to represent the recovery ratios of different regions or provinces within the selected country. Also, a filter is applied to the province dimension to limit the result to 'Top 10'. So, we will see the top 10 regions with the highest recovery ratios within the country. This will help us to identify the provinces where the disease has either been contained or the measures taken by the province have proved to be effective. Hence, similar measures can be taken in the regions where either the active cases or the death rates are increasing.

Here, provinces with fewer confirmed cases can also have high recovery rates which means that out of the closed cases, higher people have recovered and fewer people have died. Hence, the treatment policies taken by the province are either effective, or the province has healthier members of the country as compared to other regions. If no country is selected, then this chart will show the top 10 regions of the world that have the highest recovery ratios.

**Calculations:**

Recovery ratio = Sum(recovered) / ( sum(Recovered) + sum(deaths) )

Where the denominator represents the <u>Total Closed Cases</u> or the cases with a final outcome (either death or recovered).

## 5. Top 10 Provinces with Highest Death Ratios


**(Dimension: Province/region; Measure: death_ratio; Aggregation: none; Drilled By Selected Country; DataSource: 03-12-2020)**

In the above KPI, a packed bubbles chart is used to represent the death ratios of different regions or provinces within the selected country. Again, we filter the province dimension to display the 'Top 10' results and hence the user can see the top 10 regions with the highest death ratios in the country. This chart will give us an insight into the regions that have been severely affected by the disease within a country and hence is in a dire need of immediate protection against the virus.

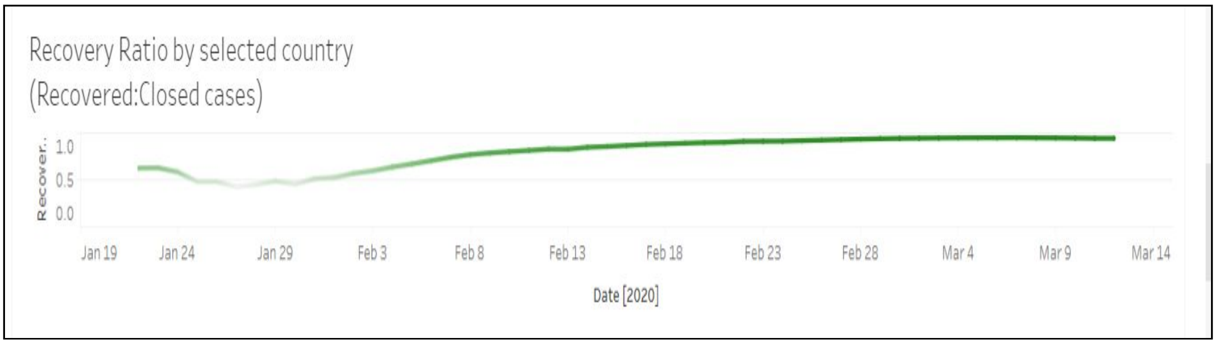
Provinces that have few confirmed cases can also have high death ratios as this means that out of all the closed cases, more people have died and fewer people have recovered. Hence, it ultimately shows a need for better treatment policies in a particular region of the country. If no country is selected, then this will show the top 10 provinces of the world that have the highest death ratios.

**Calculations:**

Death ratio = Sum(deaths) / ( sum(Recovered) + sum(deaths) ) ]

Where the denominator represents the <u>Total Closed Cases</u> or the cases with a final outcome (either death or recovered).

## 6. Recovery Ratio by date



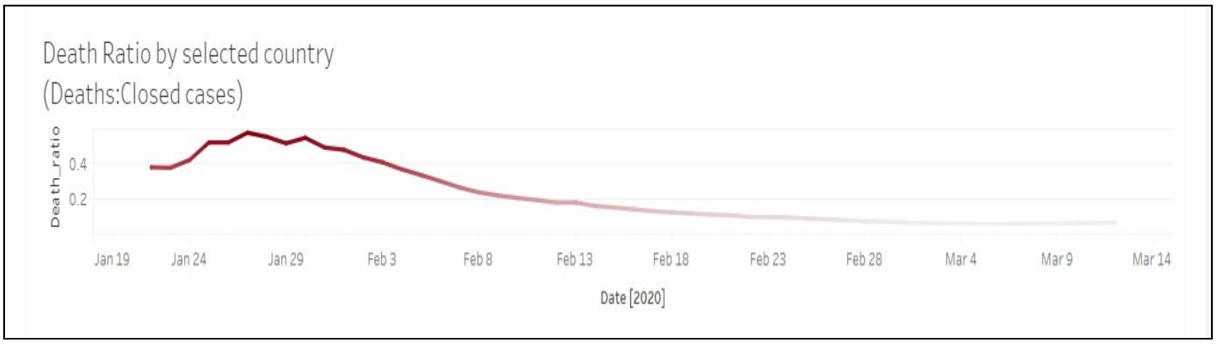

**(Dimension: Date; Measure: recovered_ratio; Aggregation: none; Drilled By Selected Country; DataSource: Combined)**

In the above KPI, a line chart is used to show the trend in the rise/fall of the recovered_ratio by date for the selected country. Also, a color mark is added on the ratio measure to represent the intensity of the magnitude of the ratio. Hence higher recovery rates are displayed by darker shades of green and lower recovery rates are shown by lighter grades of green.

This shows a trend in the recovery rates of the country and hence can give us an insight into the effectiveness of the actions taken by the country. For example, after testing a new potential cure for the virus, we can see whether the potential cure is effective or not. We can also judge the current strategies taken by the country based on the recovery rates, as higher the curve, more is the recovery ratio which means that the current steps taken by the country are helping its citizens recover.

If no country is selected, then this chart will show the trend in the rise/fall of the recovered_ratio of all the countries combined. Hence, a display of the worldwide trend of recovery rates over the given period will give us an idea of the effect of the virus on the health of individuals all over the world.

## 7. Death Ratio by date



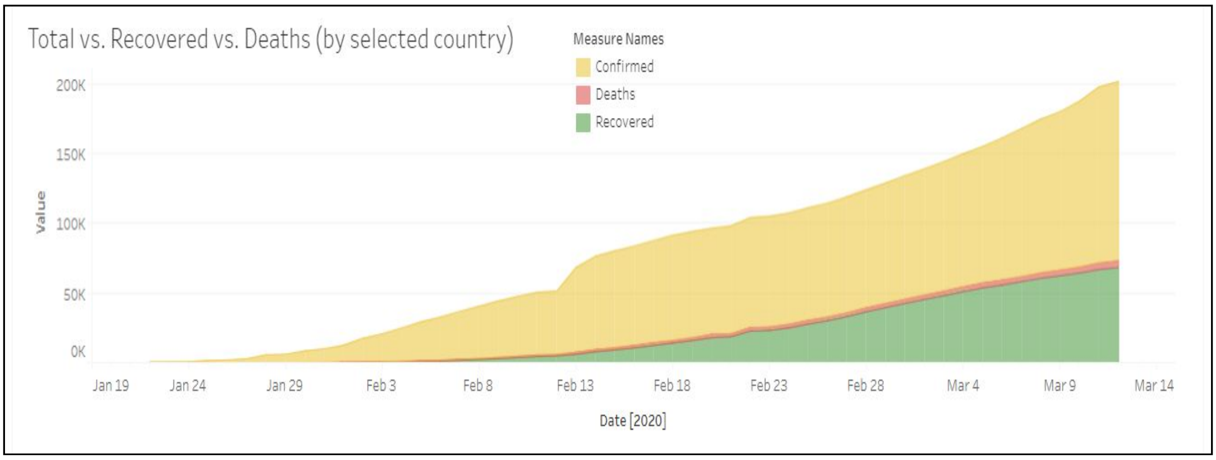

**(Dimension: Date; Measure: death_ratio; Aggregation: none; Drilled By Selected Country; DataSource: Combined)**

In the above KPI, a line chart is used to show the trend in the rise/fall of the death_ratio by date for the selected country. Also, a color mark is added on the measure to represent the intensity of the magnitude of the ratio. Hence, darker shades of red refer to higher death ratios and lighter shades of red refer to lower death ratios.

This color-intensive chart can give us an insight on which provinces need extra resources and measures to control its growing death rates and which provinces have effective strategies shown by their decline in the death ratios. A consistent rise in the death ratio refers to the poor management of the country to fight against the virus and hence it requires extra measures and resources to keep the country's citizens and economy safe.

If no country is selected, then it will show the trend in the rise/fall of the death_ratio of all the countries combined. A worldwide death ratio will also give an insight into how deadly the virus is, based on the number of deaths all over the world. Higher the curve on the chart when no specific country is selected, more deadly will be the virus. Hence based on the deadliness of the disease, extra precautions should be taken by each country to protect its citizens, and hence the economy of the country.

## 8. Total Confirmed, Deaths, Recovered by date



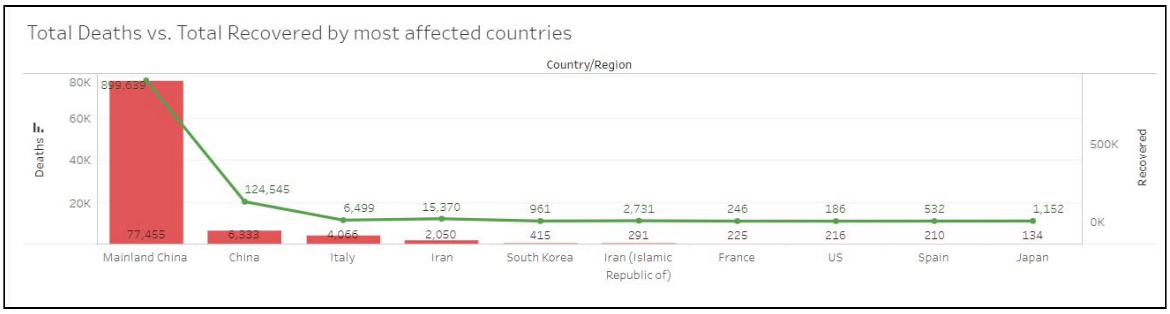

**(Dimension: Date; Measure: Confirmed, deaths, recovered; Aggregation: SUM; Drilled By Selected Country; DataSource: Combined)**

In the above KPI, an area chart is used to show the growth in the total number of recovered, deaths, and confirmed cases, by date for the selected country. The three different measures are plotted on a common axis to show relative growth. Also, all three measures are color-marked for the user to distinguish and easily identify. The magnitude of a measure depends upon the area covered by that measure. Hence, the larger the area covered, the higher the magnitude of the measure. Here, as we see from the above graph, yellow color covers the largest area and hence has the highest number of confirmed cases in the country, followed by green or the total number of cases recovered, and red or the total number of deaths.

The day-wise graph also shows us an increase or decrease in the total number of cases. This gives us an insight into the spread of the disease in the country. From the rise or fall of the total number of cases, we can also see how effective a country has been in containing the spread of the disease. For example, after a complete lockdown of the nation for a few weeks, we should see a decline in the number of confirmed cases. However, it will not affect the death or recovered cases as these are only affected by the country's ability to treat the disease.

If no country is selected, then it plots the worldwide growth of confirmed, deaths, and recovered cases. Specifically, in our chart above, we see a massive rise in the total number of confirmed cases worldwide, as in the given period, the disease had just started to spread all over the world and hence we see a steep incline in the confirmed cases.

## 9. Deaths vs. Recovered by highest affected countries

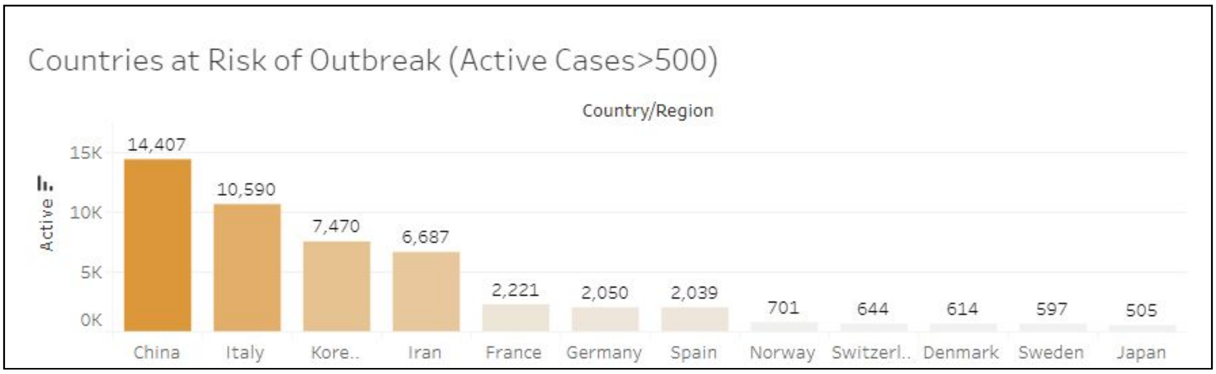



**(Dimension: Country; Measure: deaths, recovered; Aggregation: SUM; DataSource: 03-12-2020)**

In the above KPI, a mixed chart is used to represent the trend in total deaths and recovered cases, by countries with most casualties. Total deaths and total recoveries are plotted on the same axis to give a relative comparison. The deaths are represented using bars and the recoveries are represented using a line chart. Two different axes are used for the two measures because of the significant difference in absolute values. Color mark is used for both the measures to make it distinguishable for the user; red color signifies the total number of deaths and the green color signifies the total number of recoveries. Also, data labels showing the number of deaths and recovered are marked for both measures. A filter on the country dimension is applied to show the top 10 countries with maximum deaths. Hence, only the top nations with the highest casualties are displayed in the above chart.

This chart will not be drilled down based on the selected country on the map. It shows a comparison between different countries in the same view and the data used is as of the last date of our given dataset, i.e 12th March 2020.

## 10. Active cases by countries at risk of an outbreak




**(Dimension: Country; Measure: Active; Aggregation: SUM; DataSource: 03-12-2020)**

In the above KPI, a bar chart is plotted to represent the countries at high risk of an outbreak. This is done by plotting the countries with respect to active cases above a threshold value. Hence, a filter on the active measure is applied to show only countries with active cases greater than 500. A color mark is added on the active dimension to show the intensity of the magnitude of active cases. Thus, countries with higher active cases have darker color shades and those with lower active cases have lighter color shades. These countries are most likely to see an outbreak or exponential growth in the spread of the disease in the near future.

A rise in the number of active cases leads to an increase in the number of confirmed cases and hence it also shows the spread of the disease among different countries in the world. The nations are displayed in descending order of active cases and hence this chart displays the countries with the highest spread of the disease as of 12th March 2020.

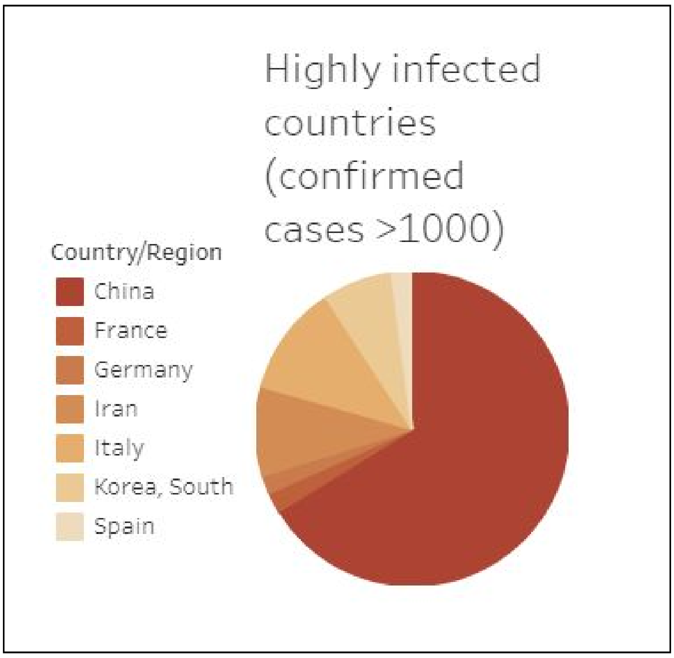
This chart is not drilled down based on the selected country (map). It shows a comparison between different countries in the same view.

**Calculations:**

Total Active Cases = Sum(confirmed) - ( Sum(deaths)+Sum(recovered) )

Where the latter represents the Total Closed Cases or the cases with a final outcome (either death or recovered).

## 11. Total cases by highly infected countries




**(Dimension: Country; Measure: Confirmed; Aggregation: SUM; DataSource: 03-12-2020)**

In the above KPI, a pie chart represents the sum of confirmed cases in highly infected countries. This is done to show the volume percentage of confirmed cases between different countries. Also, the respective countries are color-marked to represent the intensity of the magnitude of the confirmed measure. Hence the nations with higher numbers of confirmed cases would have darker shades of brown and those with fewer confirmed cases would have lighter shades of brown. A filter is added on the confirmed measure to display only those countries with confirmed cases greater than 1000. This is done to show the countries which are highly infected and the relative portion of total cases by these nations.

This chart will not be drilled down based on the selected country (map). It shows a comparison between different countries in the same view, and the data used is as of 12th March 2020.

# Chapter Four : Conclusion

This project gives a detailed analysis of the effect of COVID 19 based on different Key Performance Indicators (KPIs). It gives a country-wise insight on the current scenario and trends of each country to evaluate the growing or declining rate of measures and suggests if the country requires additional measures to contain the disease. It also compares the countries that are either highly affected by the disease or are likely to witness an outbreak. This will help us to know the nations that are successful in reducing the spread of the disease, and hence one can follow the steps taken by that specific country to contain the outbreak. It will also help us identify if a particular measure taken by a country has had favorable effects or unfavorable ones by observing the slope of the confirmed cases in the nation. Hence, the effectiveness of a measure taken can be seen practically through the charts and can be implemented accordingly.

Additionally, researchers and analysts can utilize the graphs used in this project to further analyze the effect of certain preventive measures taken by a country, such as a lockdown. This will help recognize the effectual countermeasures to prevent the spread of COVID 19 and can play a major role in reducing the number of active cases in the country. The data used in this project can also be used by data scientists wherein machine learning algorithms can be applied to this dashboard or other similar dashboards for predictive analysis on the future consequences of the disease. Since there does not exist a cure at present, the prediction can warn and alert against the increase in the spread of the disease and steps can be taken to face the inevitable. Other machine learning algorithms such as clustering and neural networks can also be applied to the trends observed in this project to analyze and prevent such catastrophic diseases in the future.

In conclusion, I hope that my work on this dashboard will help tackle the spread of the novel COVID 19 using data-driven decisions. Such decisions can not only ensure a quicker triumph, but also an effective one. The right decision at the right time is very crucial to contain a pandemic such as this and can help save a country's economy, along with many lives of its citizens. Hence, I hope this dashboard inspires many other data enthusiasts to come together and collectively construe a potential solution.