Report on

Covid-19 Outbreak Analysis and Visualization

1. ABSTRACT

The project was based on the current and mounting pandemic named Corona Virus. It was able to visualize how this virus had originated in China and spreading rapidly on a global scale. With the help of Bar and Donut charts, it was easier to perceive how the virus augmented and extended during the span of 3 months in various regions of China; some of them which includes Guangdong, Zhejiang, Henan and Hunan provinces. Since the disease is contagious and has affected various other countries, with the help of multiline charts comparisons are done for confirmed, deaths and recoveries across severely affected nations such as USA, Italy, Spain, Germany, UK, France, Belgium, Russia, Iran and Turkey. Map and bubble chart also enabled one to visualize the impact of the virus universally. Additionally, the virus has impacted the one of the leading nations in the world more harshly out of all and with the help of the visualization like choropleth and bubble chart included in this project, one can envision the confirmed, death and recovered cases across all the states of United States of America.

2. GOALS

- To visualize the initiation and rapid expansion of the virus across provinces of China.
- To visualize the spread of the virus globally and to compare the impact of it among different countries.
- To visualize the ongoing pandemic in the most affected nation and how it's still active in all of the states of USA.

3. SPECIFICATION

Data sources of the datasets used in this project are from the following –

- https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/
- https://www.coronainusa.com
- https://www.kaggle.com/imdevskp/corona-virus-report
- https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases

The packages of ggplot2, ggrepel, map, mapproj, plotly and ggthemes, lubridate, dplyr and readxl from the statistical programming language R was used for Data Visualization in terms of Grouped Bar, Donut, Multiline, Map and Bubble Charts. Choropleth map in d3.js was also used.

4. INTRODUCTION

The coronavirus disease 2019 (COVID-19) is a highly contagious disease caused by severe acute respiratory syndrome. The known symptoms of this virus are flu like and can affect a human being within a span of 2-14 days of exposure. Since this a current pandemic and as of today there are 2402076 reported confirmed cases, 165106 deaths and 623911 recovered cases worldwide.

Confirmed Cases = Active Cases + Deaths + Recovered Cases

The visualization used in this project explored the initiation, expansion and impact of the virus by breaking it down from a global scale to a regional scale based on which it drew inferences in terms of statistical numbers and how it has an influence on the mankind which has resulted to a lockdown in almost every other nation where the virus continues to exist.

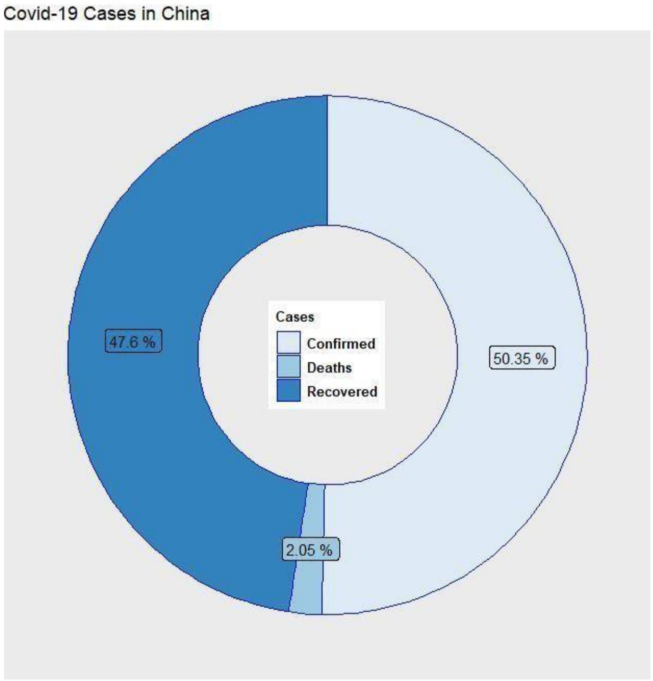
5. ORIGINATION OF THE VIRUS

COVID-19 Timeline

31st Dec 2019: China reported a cluster of cases of pneumonia in Wuhan, Hubei Province. A novel coronavirus was eventually identified.
4th Jan 2020: WHO (World Health Organization) reported on social media that there was a cluster of pneumonia cases – with no deaths – in Wuhan, Hubei province.
10th Jan 2020: WHO issued a comprehensive package of technical guidance online with advice to all countries on how to detect, test and manage potential cases, based on what was known about the virus at the time. This guidance was shared with WHO's regional emergency directors to share with WHO representatives in countries.
13th Jan 2020: Officials confirm a case of COVID-19 in Thailand, the first recorded case outside of China.
22nd Jan 2020: WHO mission to China issued a statement saying that there was evidence of human-to-human transmission in Wuhan, but more investigation was needed to understand the full extent of transmission.
11th March 2020: Deeply concerned both by the alarming levels of spread and severity, and by the alarming levels of inaction, WHO made the assessment that COVID-19 can be characterized as a pandemic.
19th March 2020: China reports no new locally spread infections for the first time since the pandemic began.
7th April: China reports no COVID-19 deaths for first time.
China is concerned a second wave of infections could be brought in by foreign arrivals.
It has already shut its border to foreigners including those with visas or residence permits.
International flights have been reduced with both Chinese and foreign airlines only allowed to operate one international flight a week. Flights must not be more than 75% full.

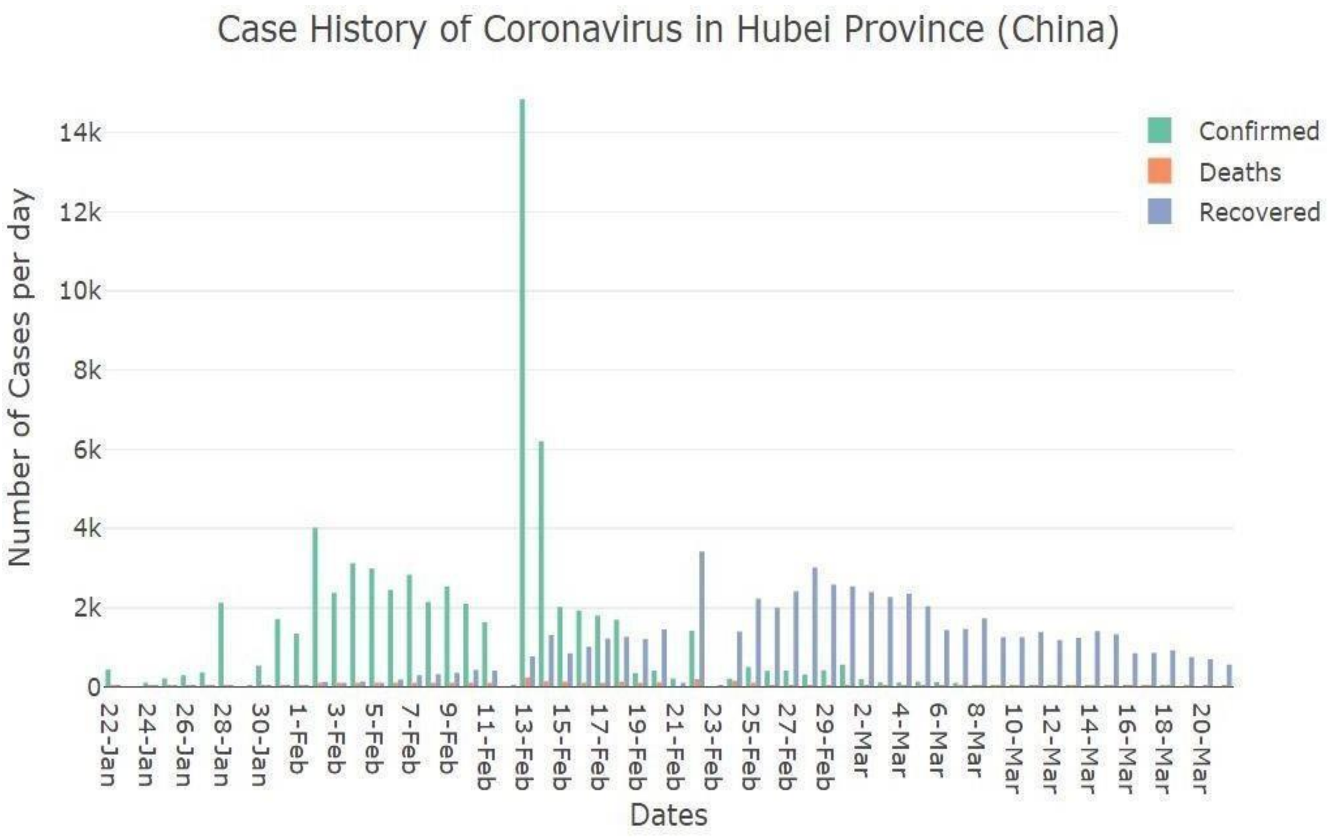
China vs COVID-19

The virus is known to have infected roughly 2.4 million people around the world and killed more than 165106 of them. Additionally, the virus has also spread to at least 185 countries and regions around the world in just four months. This virus that had originated in Wuhan city, China has total cases of 82052, 3339 and 77575 for confirmed, deaths and recovery. As shown in the chart below that represents the total percentage of how many people have attacked by this virus, one can notice that most of them have recovered and there are less deaths.



6. PROVINCES OF CHINA IMPACTED BY COVID-19

The origin of the virus and how it propagated is cleared by the above. However, the measures taken by China to defeat COVID-19 can be represented by the graph below. Elaboration on how Coronavirus has affected China and how china has recovered from this deadly virus incredibly while most of the world is still struggling with it is done below. Starting with Hubei province whose capital Wuhan city is now known as the epic center of the coronavirus crisis. In this plot the cases have been recorded from 22nd January – 21st March 2020 which is approximately two months of data related to Hubei province alone. Dates are on the x-axis and the number of cases that includes confirmed, deaths and recovered are on y- axis.

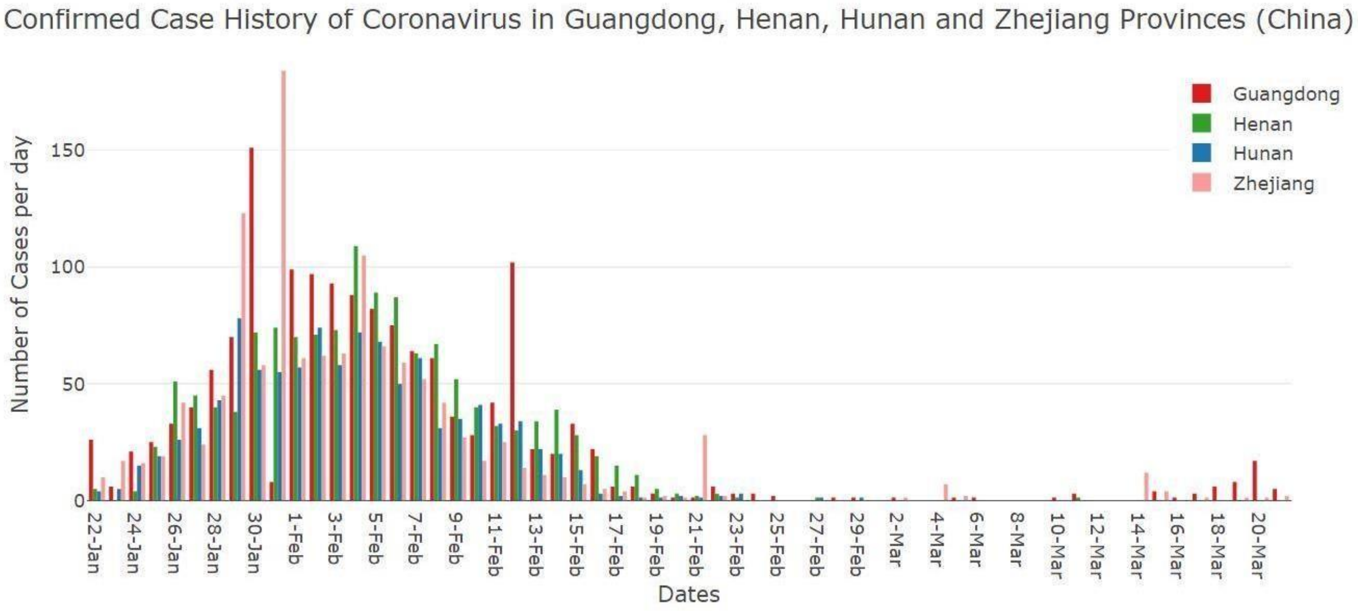


On 22nd January there were 444 confirmed cases while lockdown was implemented on 23rd January and since then growth of confirmed cases continued to increase until 4th February where 4024 confirmed cases were recorded and again cases got decreased for 10 days but we observe that there is a sudden increase on 13th February making a record of nearly 15000 cases (14840) in just one day but then the count of the confirmed cases that are recorded daily has continued to decrease drastically from the very next day.

From the overall graph one can notice that it took nearly 20 days for the count to reach the peak stage i.e. from 22 Jan-13th Feb and then it for it to get stable it almost took 35 days. There's a need to observe how the recovered graph has increased from 13th Feb until 21st Mar with a maximum count recorded of 3418 cases. And the obvious result of recovery was dropping of death count that are recorded daily.

The population of Hubei province is 58.8 million and in that 67800 people were infected i.e. 1.15% and death and recovered rates are 4.73% and 94.54% respectively.

Next, how was the effect in other provinces of China was illustrated. In this plot death rate and recovered rate are absent as there is no much significance since most of the china deaths and recovered have occurred in Hubei province itself as explained previously. This plot demonstrates confirmed cases of Guangdong, Zhejiang, Henan and Hunan provinces on y-axis and dates on x-axis. Observing that cases have been recorded more in the period from 29th Jan - 12th Feb where peak stages of all the four provinces fall. In this, province Zhejiang had two peak stages one on 29th Jan and other on 31st and a sudden downfall on 30th from 123 – 58 cases per day and sudden increase to 184 the next day i.e. triple the value of previous day and It took 7 days to reach peak stage from normal count and approximately 12 days to become stable (i.e. from 1st – 12th Feb). Guangdong province also has two peak stages on 30th Jan marking its highest count 151 and other on 12th Feb recording 102 cases. Observe that after reaching a peak point the graph is falling down immediately i.e. the trend is not continuing for long and they have situation under control by this one can imagine at what rate they had been stopping the spread. The other two provinces Henan and Hunan does not have their graph changing drastically but they still do have ups and downs recording highest count of 109 and 78 per day respectively. Considering this, Hubei province also didn't takemuch time for China government to take the situation under control and eradicate the virus.



MEASURES TAKEN BY CHINA TO OVERCOME COVID-19

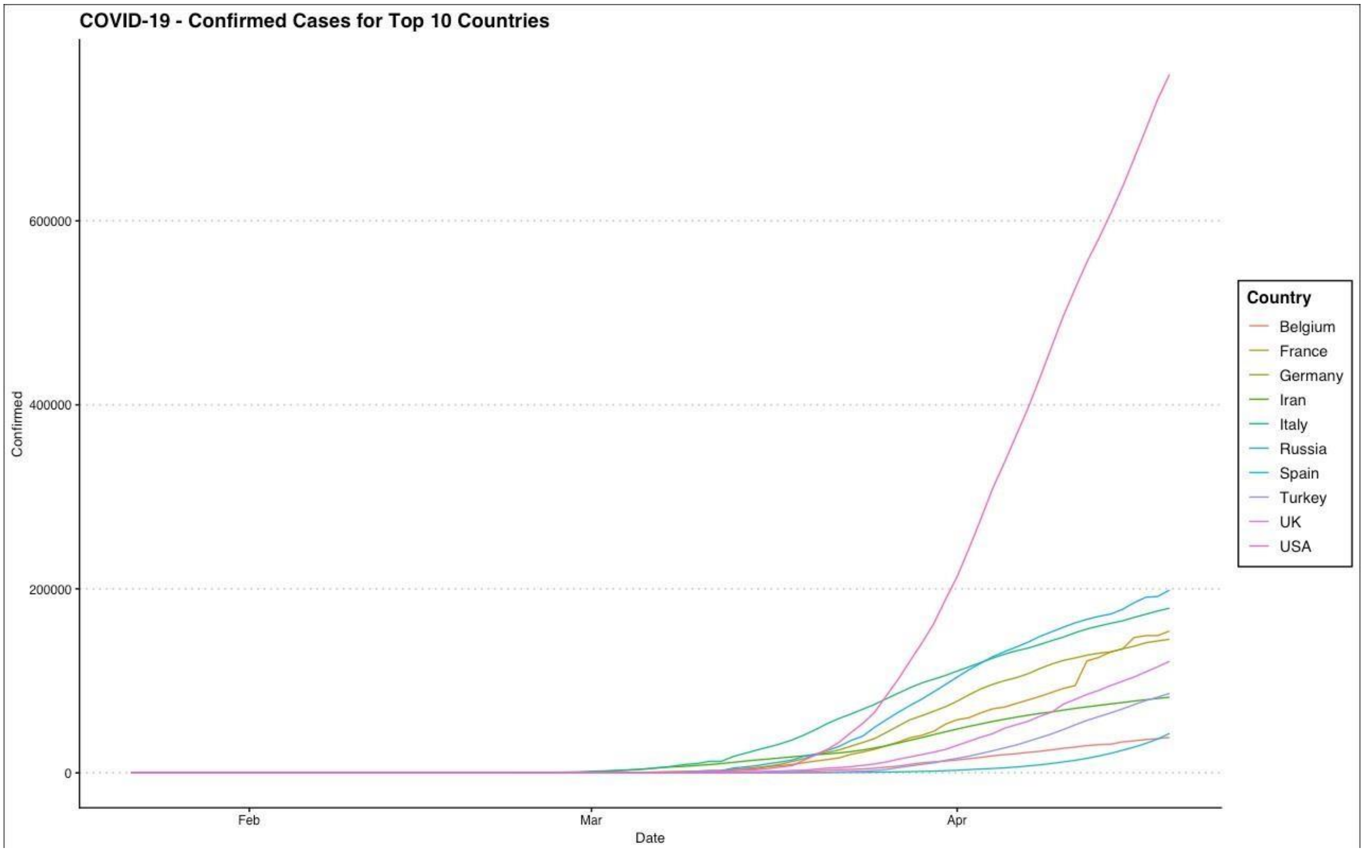
One may wonder how China could control the spread of this deadly virus in such huge population of Hubei which is almost double the population of Texas with such less death rate. Taking into account that it is easier for Chinese government to control the spread as they know what is the lose end unlike the rest of the world. Since they knew where the virus is born, they contaminated that place, Wuhan city for 76 days immediately after they are aware that virus was spreading through human to human interaction. By doing this they made sure that virus was not spreading to other provinces, so instead China concentrated only on the recovery rate which reduced the new count of death or confirmed cases.

REASONS BEHIND THE FAST RECOVERY RATE IN CHINA

This is because China might be aware of something is coming their way but not the severity of the situations. The facts considered by China was that it is spreading through human interaction and the right measures were taken to overcome the problem. Cases had started to register in China since December 2019 while most of the world was unaware that virus would reach them too. It was only after WHO issued a statement on Jan 22nd, 2020, that virus is spreading through human interaction and the world reacted but then was too late as people from China had travelled to different places around the world and interacted with many people through some form or the other. It was clear that rest of the nations were not well prepared for what was coming their way and by the time they had started to take effective measures it was too late, and the situation was out of control. For example, in Spain, the first case that is registered was on 31st January when a German tourist tested positive and by 13th Feb Spain confirmed multiple cases but still the government was reluctant to implement lockdown. They implemented lockdown on 14th March after all the 50 provinces in the country had registered positive cases. Spain was the not the only country that was unaware of the severity of the situation and that's why the next topic of this project illustrates how the virus has impacted the rest of the world.

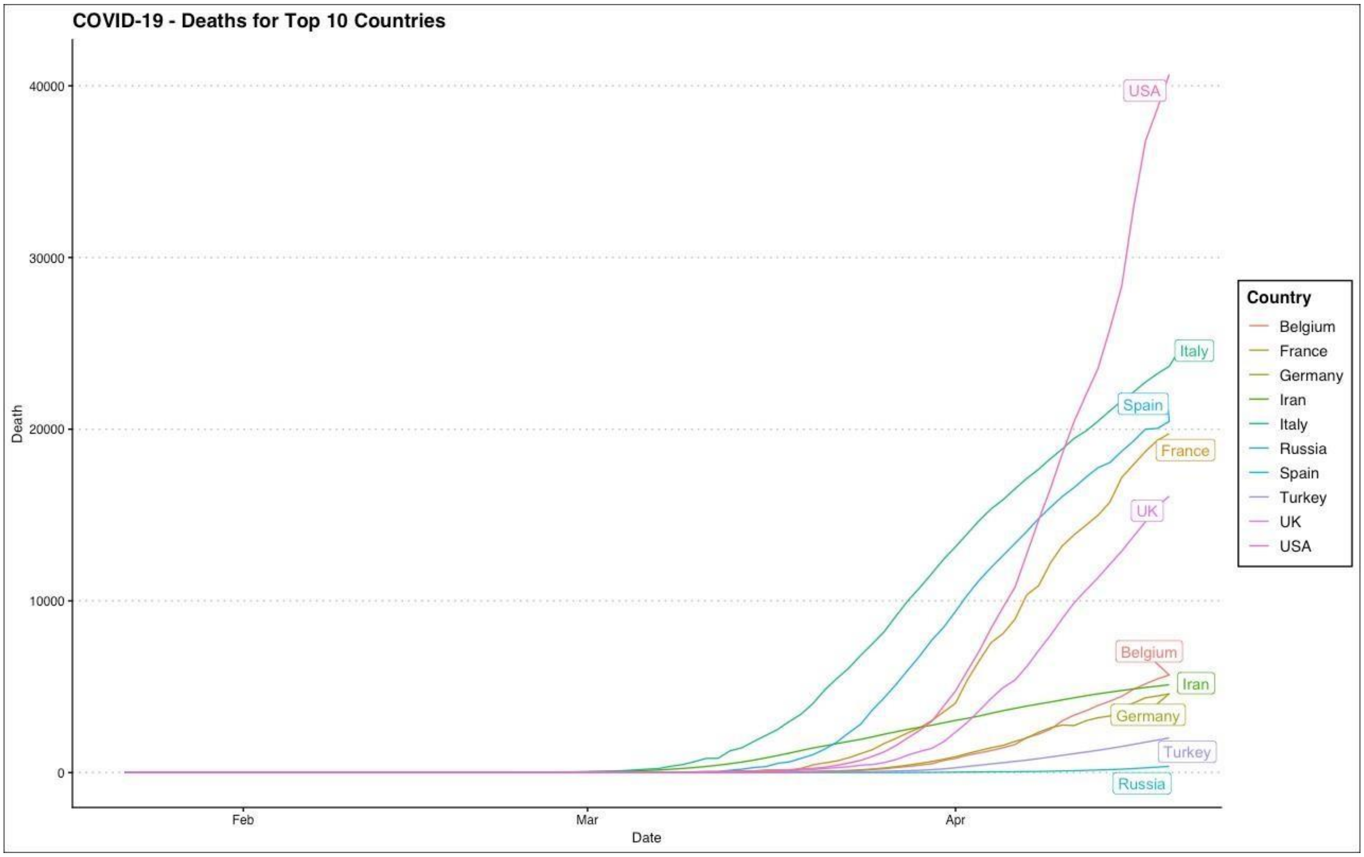
7. TOP 10 COUNTRIES AFFECTED BY COVID-19

The 2019-20 global pandemic was recognized as a pandemic on 11th March 2020. As explained above, by this time, the virus had reached almost 185 countries with more than 2.4 million cases and more than 165000 deaths overall. The recovery rate was observed to be increasing with more than 624000 people but there was always room for a deterioration and reinfection depending on the current scenario. Since it was declared a global pandemic, many countries had put in the lockdown protocol for the safety of its citizens and to take control over the situation. However, taking factors like population and overseas travel into account, some countries' numbers were rising at a constant rate. The plot below represents the total number of confirmed cases for top 10 countries where the outbreak continues to rise.



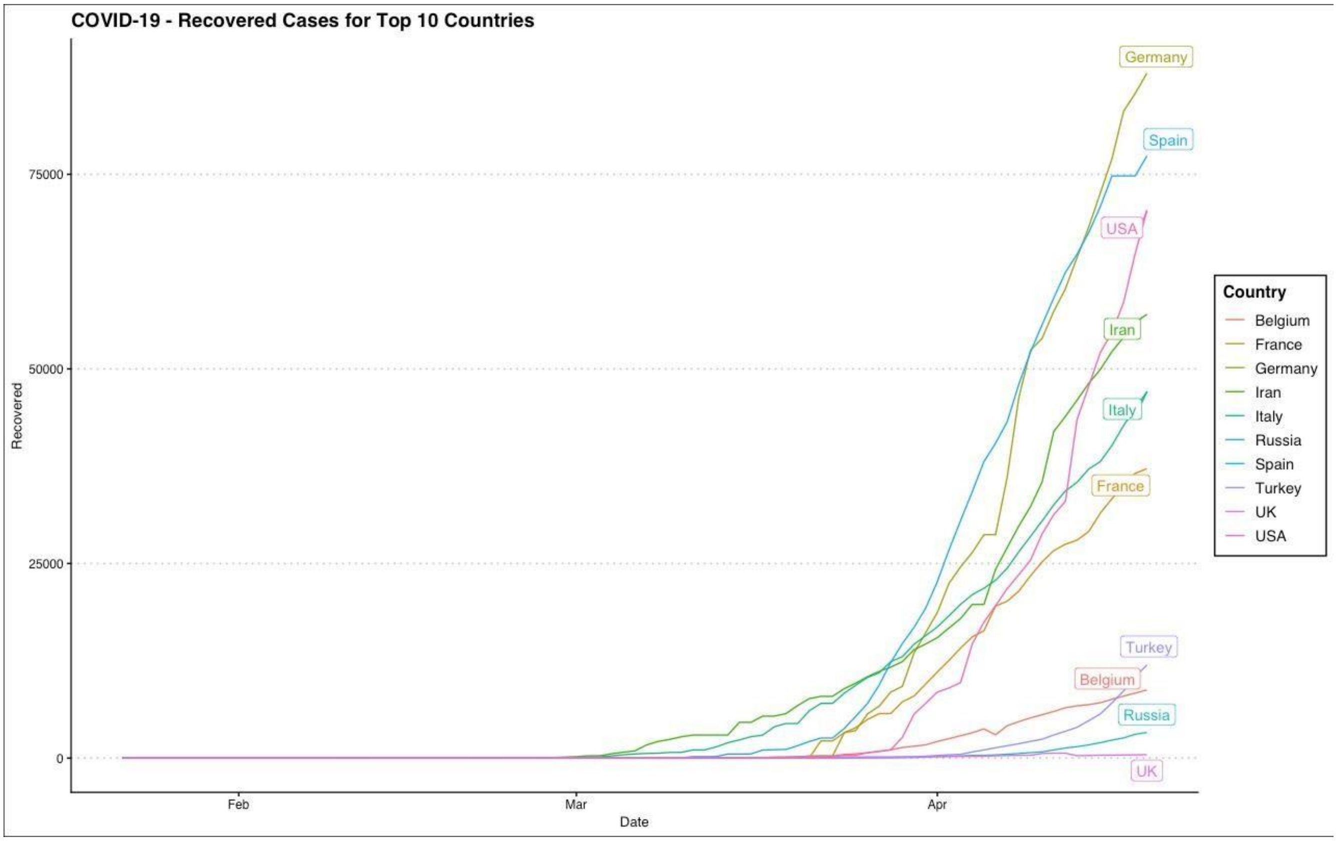
The above plot was made by extracting data from Kaggle source and cleaned the data by only taking into account of the highest number of cases in the world. After this accessing the new dataset and using packages of ggplot2 and ggthemes from R programming the above multiline plot was created where x-axis is the Date which is in an interval of months, starting from 22nd January till 19th April and y-axis consists of the number of confirmed cases. Since R default format for axis labels is in the scientific order of exponential, change of parameter by applying *options(scipen=999)* was needed. Because of this, we can see clearly from the above visualization that USA marks to be above the 600000 line and is approximately more than 764000 as of today making it the highest number of confirmed cases in the entire world. Likewise, Spain ranks second, Italy third and so on while Russia ranks lowest out of the top 10 countries as seen by the color mapping of the legend.

However, since there only a finite set of colors and more the number of lines on the multiline chart, it tends to get difficult to differentiate among them. This issue was resolved by using the ggrepel package which uses label right next to the line on the and was implemented for the plot of highest number of Deaths per country.



With number of deaths on y-axis and Date (22nd January to 19th April) on x-axis, it can be observed that USA has the highest number of Deaths more than 40000 and comparing it with the number of confirmed cases, that is, nearly 36% death rate. However, it can also be seen that, Italy has passed Spain in terms of number of deaths and so on while Russia still remains to be the lowest for deaths out of the top 10 countries. It can be clearly noticed that all the countries represented in the plot are countries with high population or have high tourism rate. Because of this reason, the virus which had originated from China and any individual travelling from one country to a high tourist country, the chances of peak in the numbers were likely. Since the virus is highly contagious, one can unknowingly get infected and pass it on from person to another which again results in a greater number of cases and high mortality rate. It can also be observed that the line on the plot didn't start to rise until mid-March and that's why COVID-19 was declared a pandemic. Due to this reason, all of the countries above which were known for its tourism and various other attractions now follow a strict travel ban or a 14-day quarantine mandatory for whoever travelling overseas to take control over the situation.

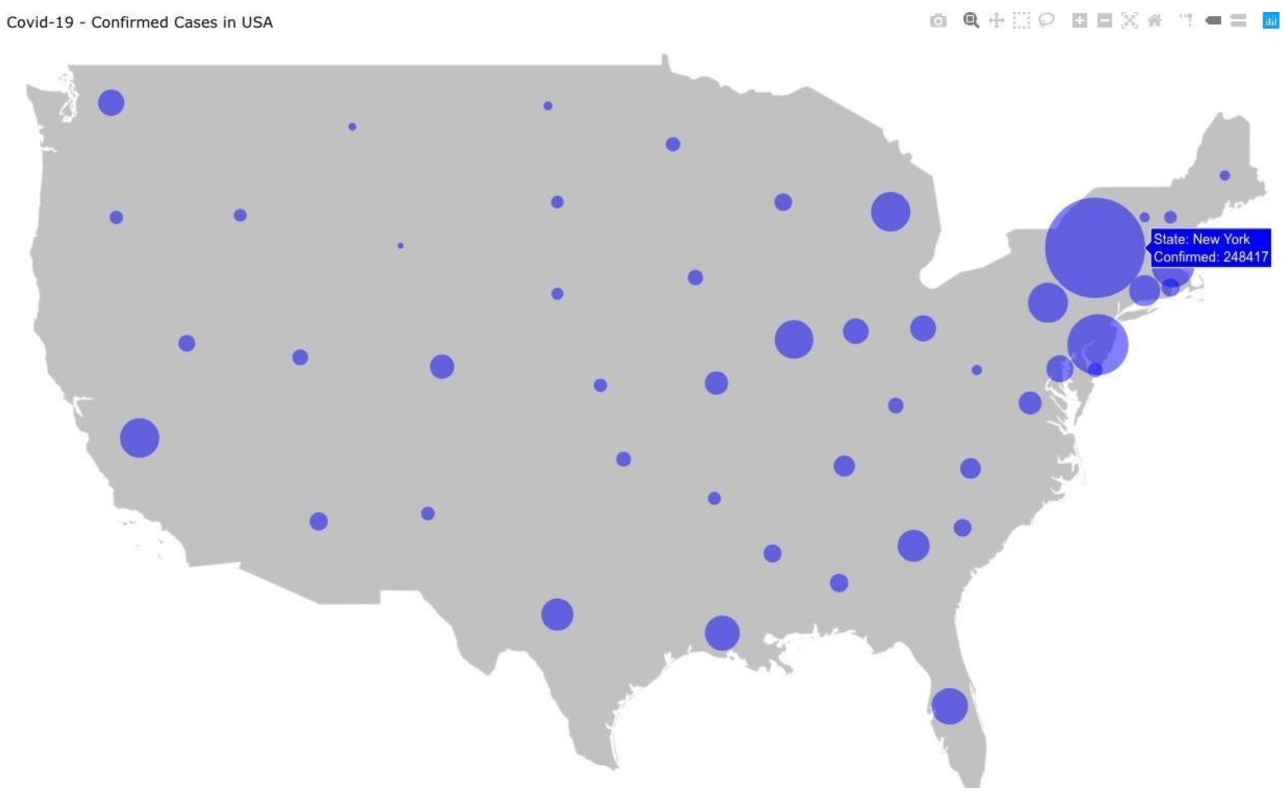
Next, recovery rate should also be taken into account in order to see how a nation is progressing and dealing with the current pandemic. The graph below represents number of recovered cases on y-axis and Date on x-axis (22nd January to 19th April).



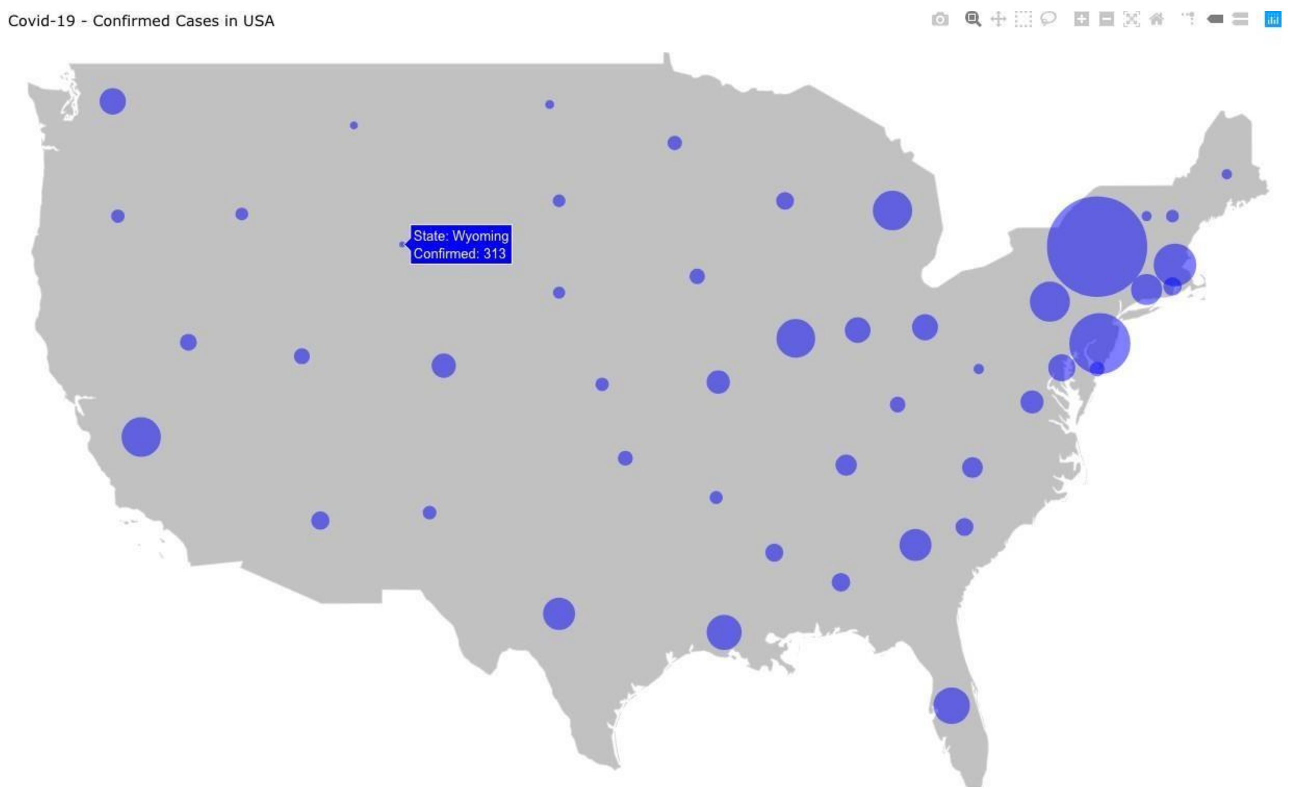
It can be observed that Germany triumphs over all the other countries in the aspect of recovery rate with more than 88000 recovered cases. Spain ranks second with over 77000, USA third over 70000 and so on. However, UK has significantly low recovery rate which is near to 0. The reason why the top 9 countries are improving in recovery rate could be because of the response measures taken by each government. These response methods vary from having and using enough testing kits, containment by wearing surgical masks, maintaining social distance in public spaces and more importantly each of these nations having a strong task force to deal with the situation in a scientific perspective and forming a law and order in terms of lockdown or curfews for the general public to stay safe.

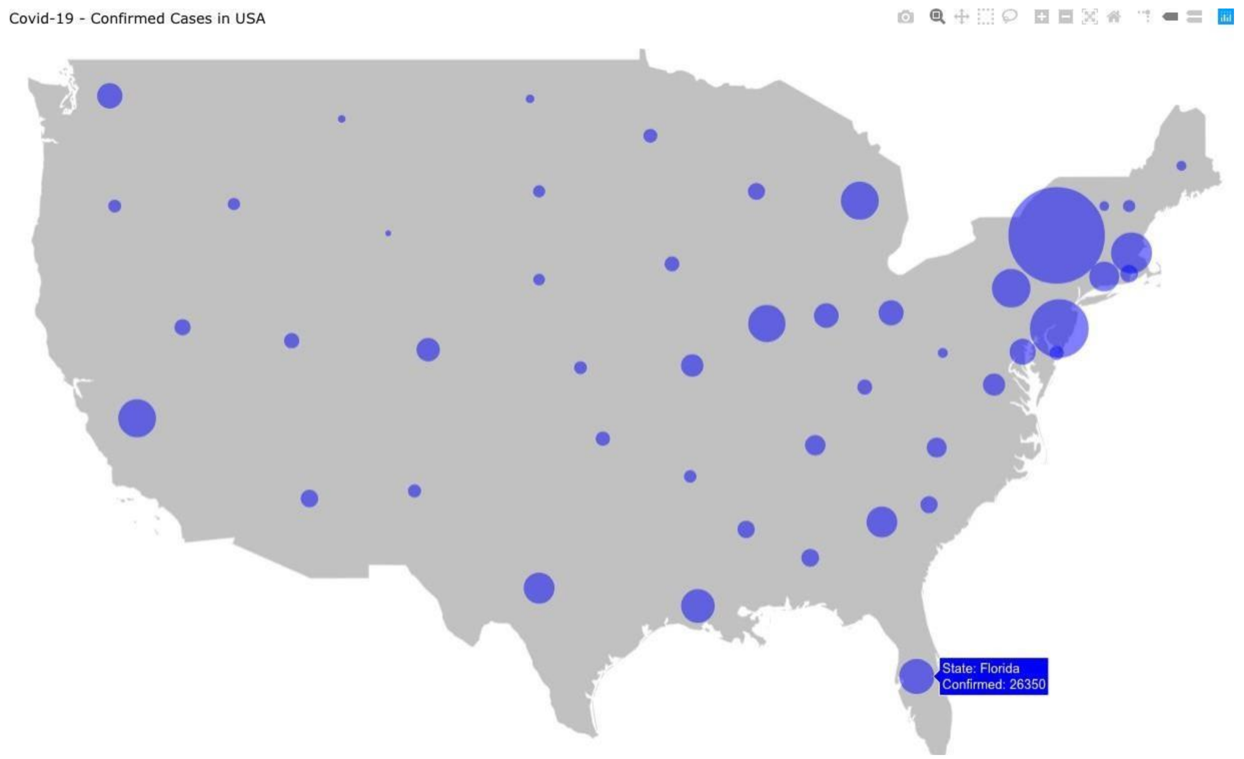
8. USA – CONFIRMED CASES

Since it has been noted clear that the figures suggest USA being the only country with maximum number of confirmed cases and death rate compared to the rest of the world, a visualization is implemented in this project to envision where and how all the states of the country are dealing with COVID-19. This visualization is done using the same packages mentioned above as well as the addition of maps, mapproj and the graphical interactive package, plotly. The map package in R already has a list of features which includes longitude and latitude for each count, country, and the world as a whole. Since, one needs to visualize how only USA's states and counties are dealing with virus, this can be achieved by using the filter function and setting it to USA only. The same can be achieved for every other country. After this, the data in R is matched to the data extracted from https://www.coronainusa.com to get the exact figures of each state in terms of Confirmed cases, Deaths and Recovered cases which is mapped to data in R with longitude and latitude of each state. Since, longitude and latitude for every county within the state differs, this project had implemented a bubble chart to show within the center of the state.



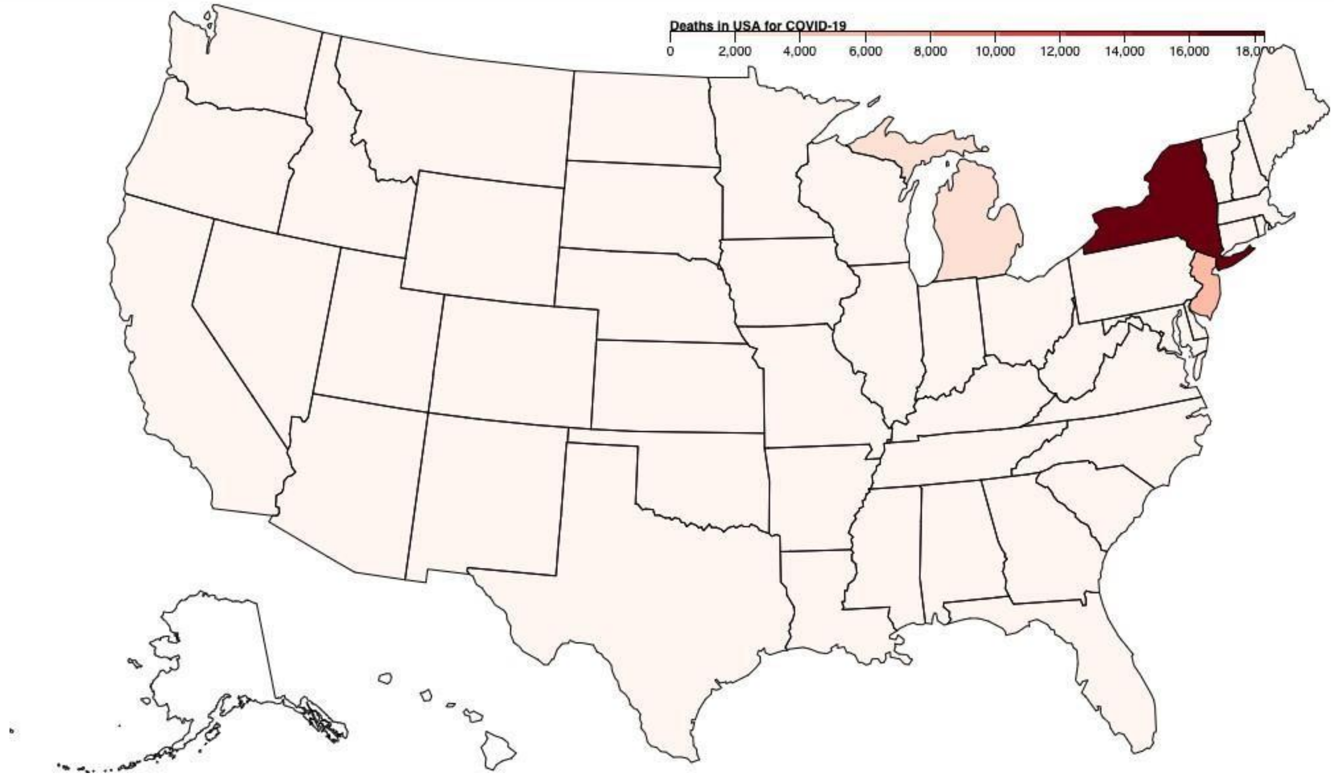
When one hovers over the map, each state with its name and the figure for confirmed cases is visible. By this representation, it was interpreted that the state of New York has the largest bubble which suggested it has the maximum number of confirmed cases, that is, 248417. Likewise, the smallest bubble was observed to be the state of Wyoming which suggested it has the least number of Confirmed cases, that is, 313. Another example of Florida which has 25528 confirmed cases as of today and is represented below for better understanding of the bubble chart on the map of USA.


Covid-19 - Confirmed Cases in USA
State: Wyoming
Confirmed: 313


Covid-19 - Confirmed Cases in USA
State: Florida
Confirmed: 26350

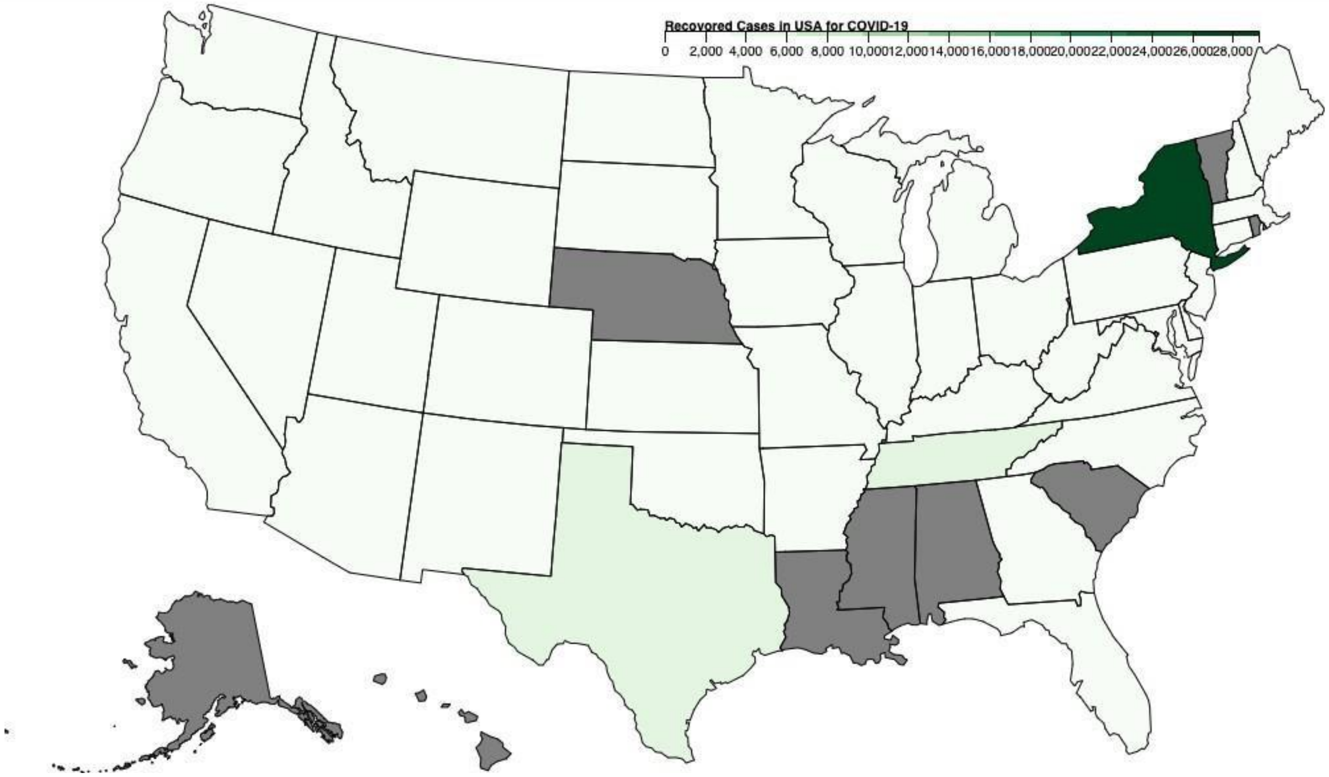
9. USA – DEATHS AND RECOVERED

To visualize the death and recovered cases of COVID-19 for the population of USA, this project has implemented geo visualization choropleth using d3.js. Using this choropleth map gives one an option of a good visualization by the contrast of colors. Two kinds of datafiles are used to create this, us- states.json which has all the coordinates and names of the each state and USA-deaths.csv whose data was extracted and cleaned from https://www.coronainusa.com. With the help of projections (geoAlbersUsa), pathGenerator and other functions in d3 which includes scaling the colorScale, mapping and user defined functions for the feature selection and extraction of data points, the below maps were created.



The above map represents the Deaths in USA and with the help of the legend and color scale of schemes of Red right above the map it signifies that the state of New York has deaths over 18000 which is the highest compared to any of the other states. After that, New Jersey has more than 4000 deaths and Michigan has more than 2000 death cases. These three states have the highest number of deaths compared to the rest of the country as the other states are all near to 1000 or below.

The map below represents the Recovered cases in USA and with the help of the legend and color scale of schemes of Green signifies that the state of New York along with the high death rate also has the highest recovery rate which is more than 28000 also making it the highest compared to any of the other states. Meanwhile, states like Tennessee and Texas have more than 3300 recovered cases. The states which are marked grey, that is, Alabama, Mississippi, Louisiana, Nebraska, South Carolina, Vermont, Rhode Island and Alaska represent 0 number of recovered cases while the rest of the states are below 2000.

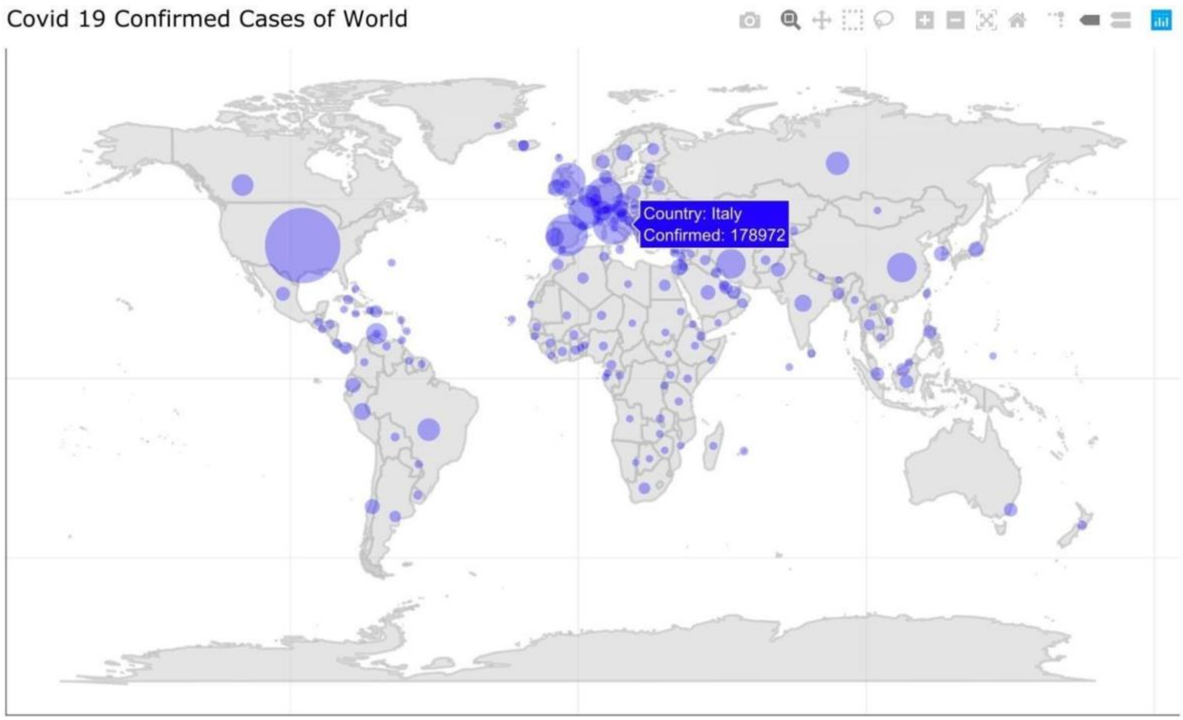


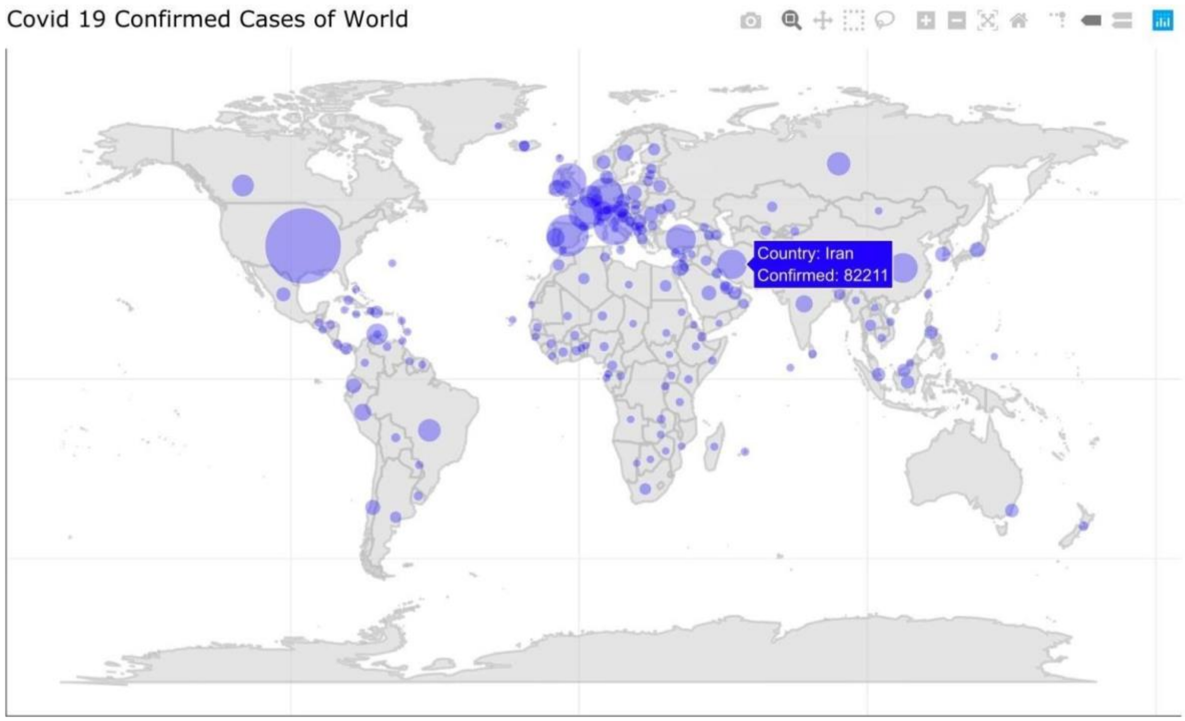
10. COVID-19 GLOBALLY

The coronavirus outbreak has impacted the entire world on a large scale as many choose to stay indoors or panic buy resulting in shortage of supplies and struggling with necessities of livelihood. Since the outbreak is worldwide, it has led to disruption in many global events in terms of sports and political. Meanwhile, educational institutes are also suffering which has led to the unemployment rate rising rapidly because of which the entire globe is facing the worst recession in history of mankind. To show what we are dealing with; this project has applied geovisualization techniques with the help of map, mapproj, plotly, ggplot2 and ggthemes used in R programming to highlight the various regions of the world where the virus continues to exist and rising day after another.

The following maps are example of the three categories: Confirmed, Death and Recovered.

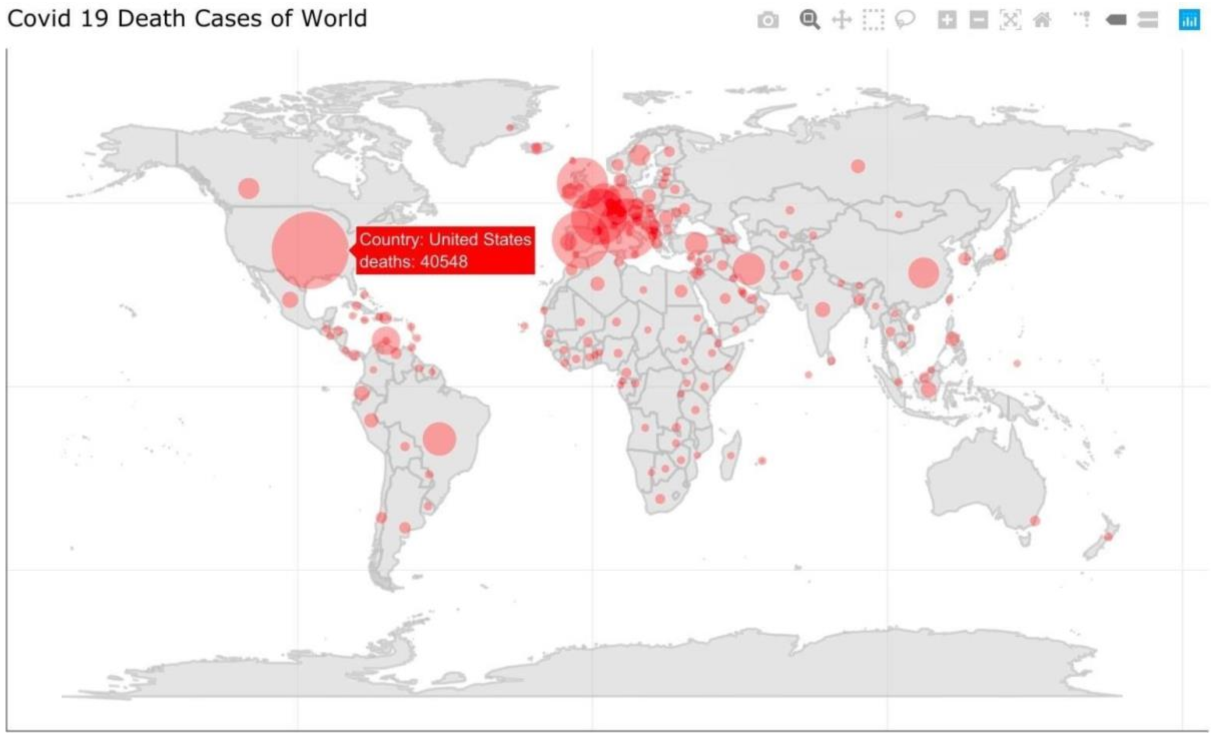
CONFIRMED CASES

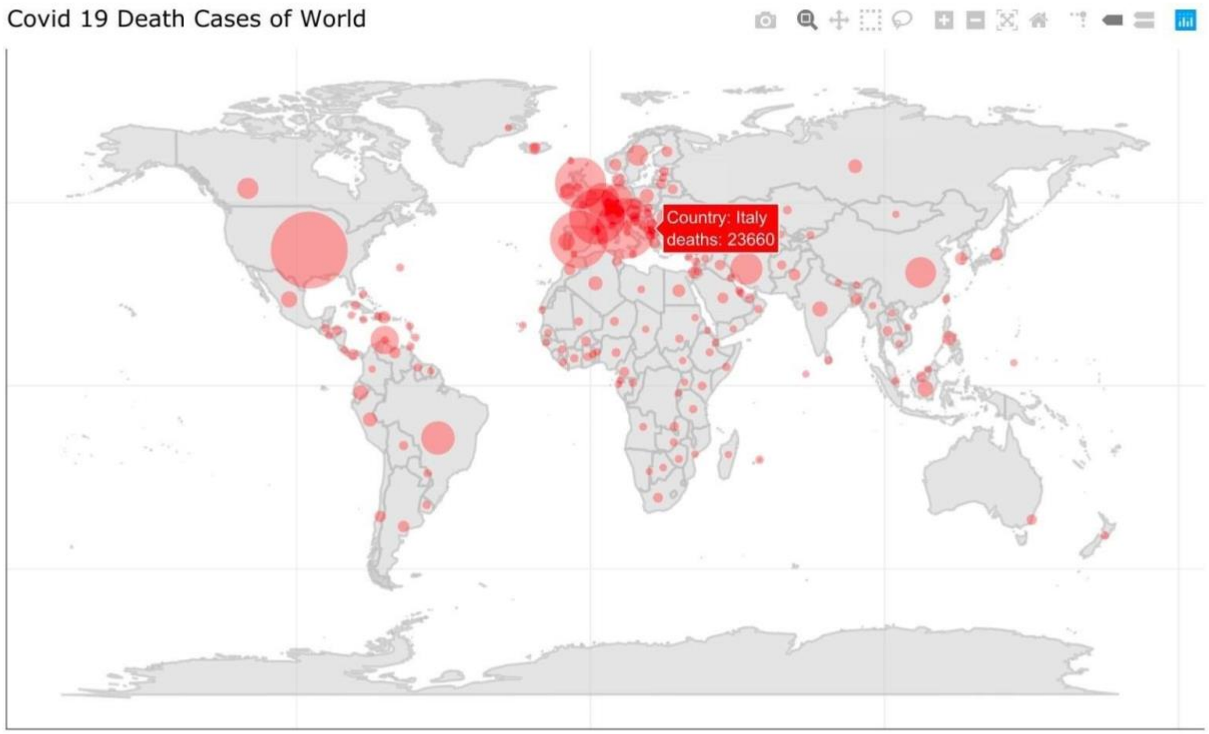




The blue bubbles represent the number of confirmed cases. Larger the size of the bubble equivalents to high number of confirmed cases and vice versa.

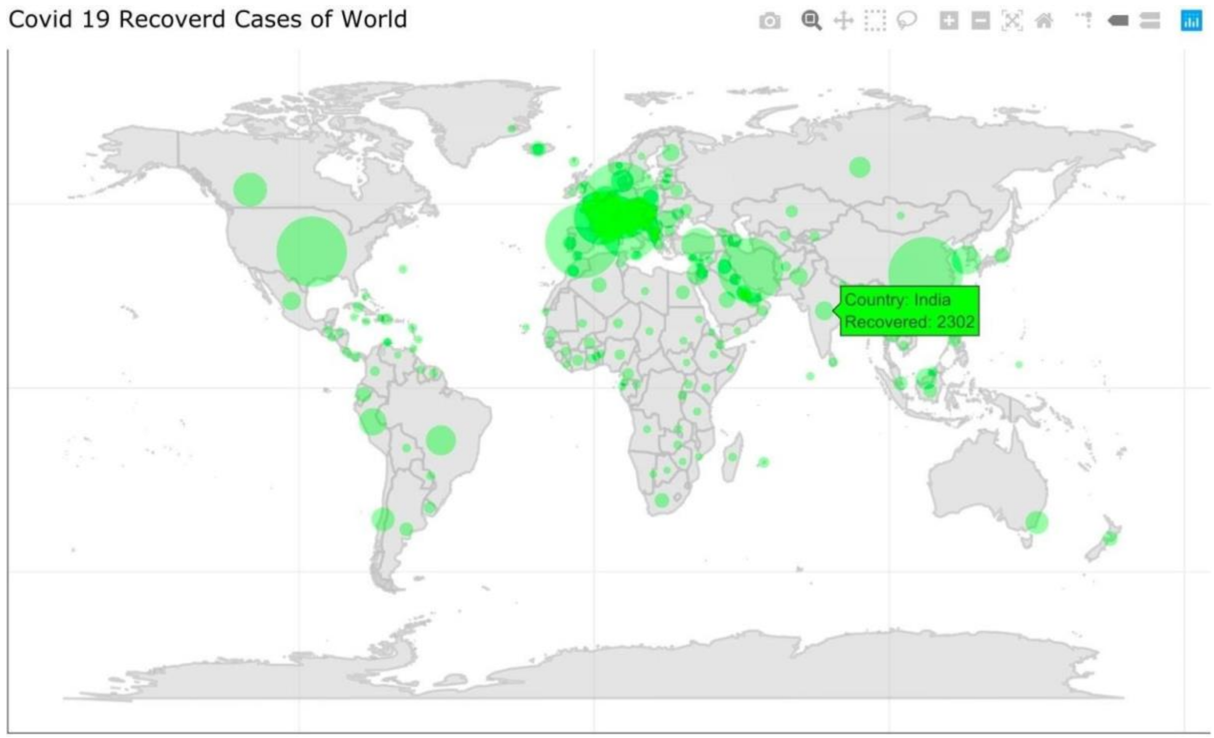
DEATHS

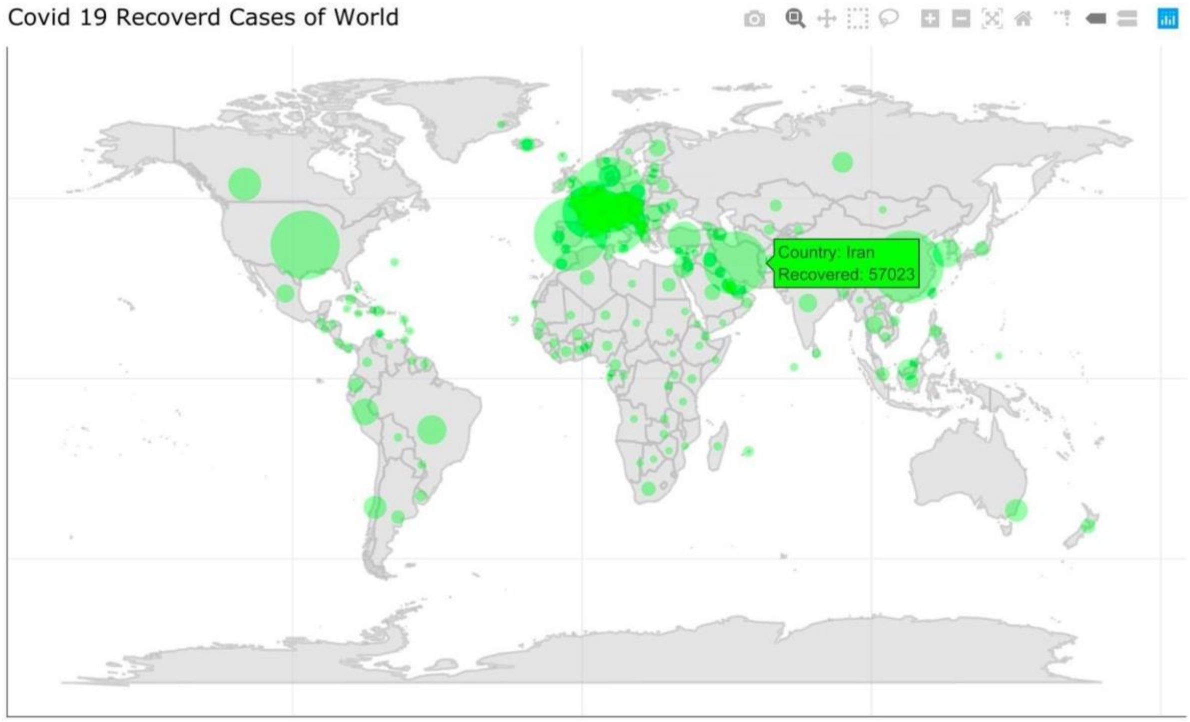




The red bubbles represent the number of death cases. Larger the size of the bubble equivalents to high number of deaths and vice versa.

RECOVERED





The green bubbles represent the number of recovered cases. Larger the size of the bubble equivalents to high number of recovered cases and vice versa.

11. CONCLUSION

The following inferences were made after analyzing data and the plots used in this project –

- The virus which originated in Wuhan City, China although high in number before, the daily active new cases have been decreasing and their current mortality rate out of the confirmed cases is observed to be 6%
- The country with the maximum reported confirmed cases and deaths is USA and its still rising.
- Out of the top 10 countries affected most by COVID-19, the country with the maximum recovered cases out of confirmed is Germany standing strong at 95%.
- Geo visualization using bubble map has proved to be ideal for keeping a track of COVID-19.

12. REFERENCES

- Medium. 2020. *Create A Web App W/ Coronavirus Case Data In A Blink Of An Eye—Using R, Shiny And Plotly*. [online] Available at: <https://towardsdatascience.com/create-a-coronavirus-app-using-r-shiny-and-plotly-6a6abf66091d> [Accessed 20 April 2020].
- En.wikipedia.org. 2020. *2019–20 Coronavirus Pandemic*. [online] Available at: <https://en.wikipedia.org/wiki/2019–20_coronavirus_pandemic#International_responses> [Accessed 20 April 2020].
- Lecture slides and program examples given by the faculty.