COVID – 19 Analysis

Data Warehousing and Power BI

Team members:
Akanksha Shetty
Siddharth Shetty

Introduction

About Covid-19:

Coronavirus disease 2019 is an infectious disease caused by severe acute respiratory syndrome coronavirus 2(SARS-CoV-2) that was found in late 2019, however it was not until March 11 that it was declared as a pandemic. Coronaviruses are groups of viruses that cause illness in animals and humans. The symptoms associated with COVID 19 are Shortness of breath or difficulty breathing, Fever, Repeated shaking with chills, Muscle pain, Headache, Sore throat and new loss of taste or smell. However, the most common ones are shortness of breath, fever and cough. As of 11th May there are more than 4.14 million cases across 187 countries and territories with a death count of more than 284,000.

For our project we have decided to perform our analysis on the factors influencing COVID-19, the areas most affected by it, the growth rate of the disease, the age bracket most vulnerable to the disease and we have also compared different diseases and external factors with the respective geographical locations to find correlations between the primitive factors. We have done so by collating datasets ranging from John Hopkins dataset to regional datasets released by officials. Other than these demographic data is also used from sources such as data.gov and cdc.gov

Objectives:

This project aims to analyze how the spread of coronavirus is related to other parameters.
The questions we tried to answer with our analysis are stated below:

- Which region is the most affected by the disease?
- What are the Top-N countries affected by the disease?
- What are the Top-N states in the USA affected by the disease?
- Did the stay-at-home order in the respective states have any positive/ negative effect?
- How does the stay-at-home order affect the growth of coronavirus cases across the states in the USA?
- Out of the Top-N which race has been the most affected?
- Out of the different age groups, which of them were the most vulnerable to the disease?
- Does gender have any correlation to the disease?
- Is there any pattern between Influenza and Covid-19? Any pattern?
- What is the difference between the recovery rate and the rate tested positive for Covid-19?
- Is there any association with the timestamp? Is it progressing over time?
- Does the population density have any correlation with the number of confirmed cases?
- Do the total tax collections of the states have any correlation with the number of confirmed cases?
- How can the number of confirmed cases be classified by race?
- Which states have seen the worst unemployment rates due to COVID-19?

About the Dataset:-

Our primary dataset is **John Hopkins University CSSE COVID-19 Dataset**. The link for the same can be found here –

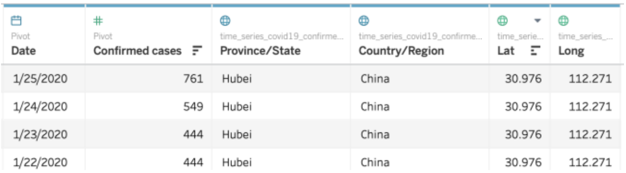
https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data

The repository can be found on GitHub and is updated on a daily basis. This repository consists of various datasets ranging from data of all countries to data specific to the United dates of America.

The files used to fetch the confirmed cases, the death count and recovered cases on a global level i.e all countries are:

Confirmed – time_series_covid19_confirmed_global.csv

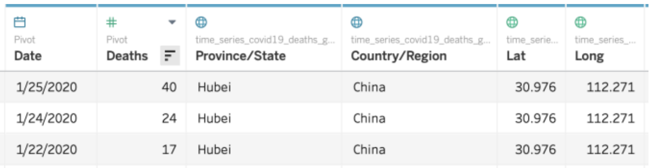
Preview: *After pivoting columns to rows*



Pivot	Pivot	time_series_covid19_confirme...	time_series_covid19_confirme...	time_serie...	time_series_...
Date	Confirmed cases	Province/State	Country/Region	Lat	Long
1/25/2020	761	Hubei	China	30.976	112.271
1/24/2020	549	Hubei	China	30.976	112.271
1/23/2020	444	Hubei	China	30.976	112.271
1/22/2020	444	Hubei	China	30.976	112.271

Deaths – time_series_covid19_deaths_global.csv

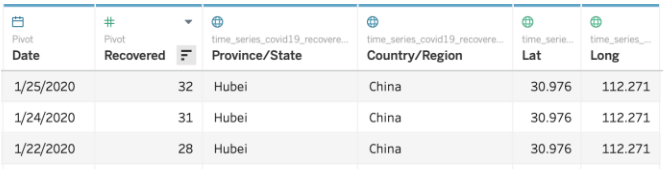
Preview: *After pivoting columns to rows*



Pivot	Pivot	time_series_covid19_deaths_g...	time_series_covid19_deaths_g...	time_serie...	time_series_...
Date	Deaths	Province/State	Country/Region	Lat	Long
1/25/2020	40	Hubei	China	30.976	112.271
1/24/2020	24	Hubei	China	30.976	112.271
1/22/2020	17	Hubei	China	30.976	112.271

Recovered - time_series_covid19_recovered_global.csv

Preview: *After pivoting columns to rows*



Pivot	Pivot	time_series_covid19_recovere...	time_series_covid19_recovere...	time_serie...	time_series_...
Date	Recovered	Province/State	Country/Region	Lat	Long
1/25/2020	32	Hubei	China	30.976	112.271
1/24/2020	31	Hubei	China	30.976	112.271
1/22/2020	28	Hubei	China	30.976	112.271

Analysis on country specific data i.e. about the United States of America can be found below

Confirmed cases- time_series_covid19_confirmed_US.csv

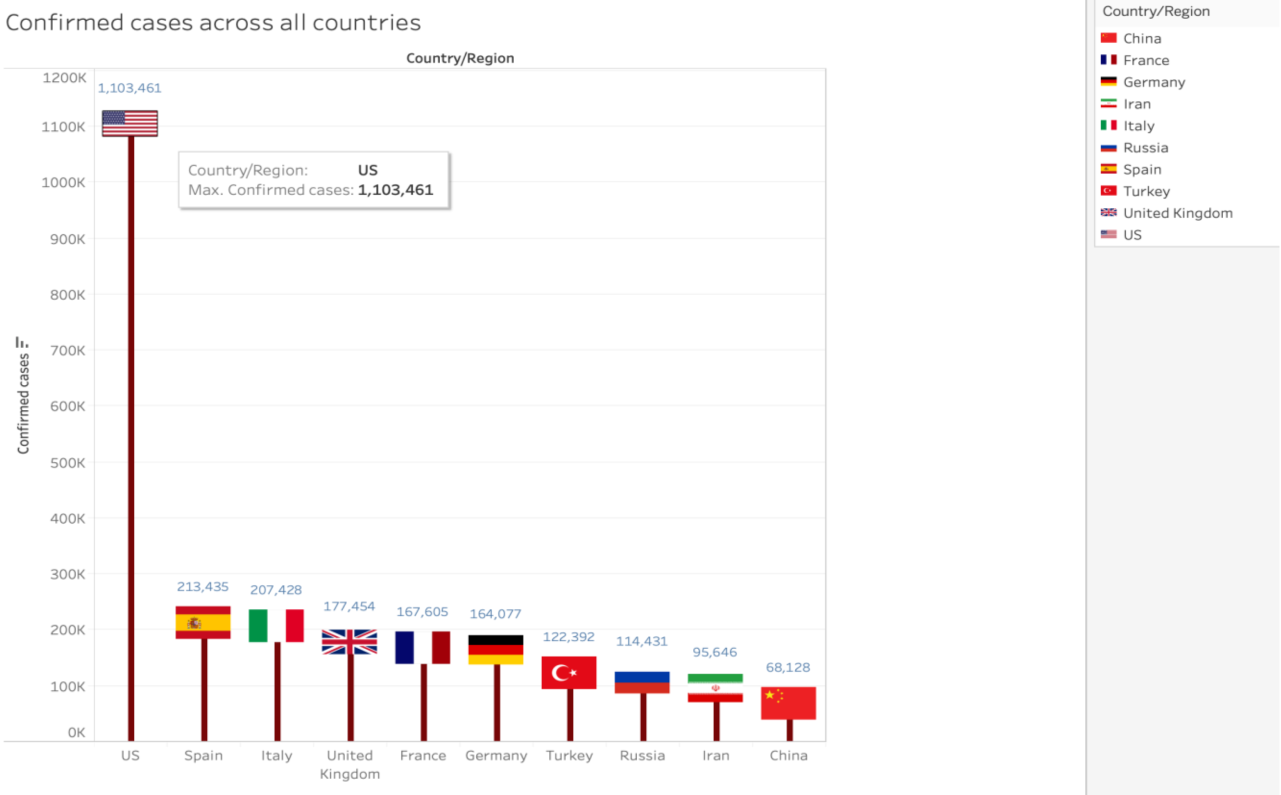
Deaths - time_series_covid19_deaths_US.csv

Additionally, the following datasets are downloaded to analyze our primary data set from different angles:

- https://worldpopulationreview.com/states/
- https://www.cdc.gov/nchs/nvss/vsrr/covid_weekly/index.htm
- https://www.cdc.gov/nchs/nvss/vsrr/covid19/index.htm
- https://data.census.gov/cedsci/table?g=0100000US.04000.001&tid=GOVSTIMESERIES.GS00TC01&hidePreview=true&vintage=2018&layer=VT_2018_040_00_PY_D1&cid=S0102_C01_001E

Visualizations:

Let’s explore the number of cases across different countries.
The visualization we created is:

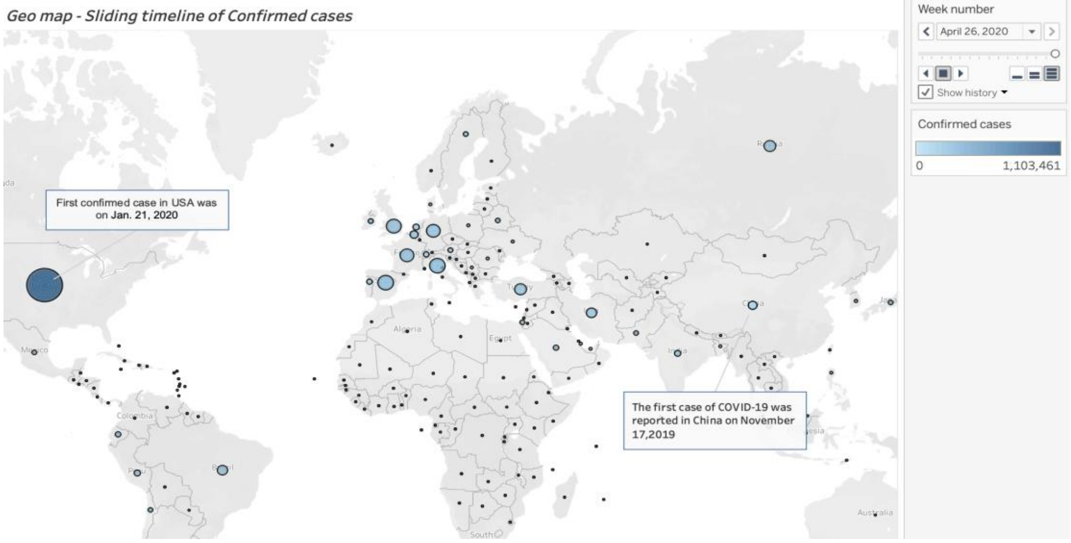


The data in the above visualization is dated as of 1st May 2020 and consists of only the Top 10 countries affected by a coronavirus.

The USA has the highest number of cases as compared to other countries, followed by Spain and then Italy. However, the data and statistics keep changing by the day. The top 5 countries as of 11th may are USA, Spain, UK, Russia and then Italy. Therefore, the above sheet is only valid for that particular date due to the variability of the disease.

One important observation is that China has lesser cases compared to other countries, the first few cases were initially found in China; therefore, it is surprising to see that the transmission in other countries has amplified at a humongous rate.

We learnt above that **time plays a very important role**, therefore let's get an overview of all the cases by associating it with **timeline.**



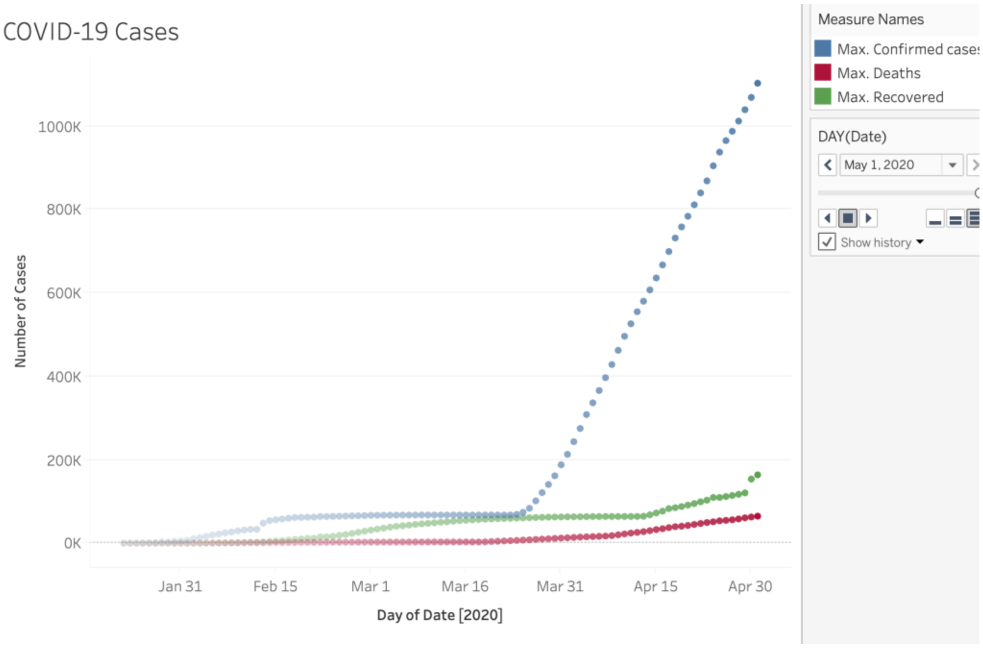
Geo map - Sliding timeline of Confirmed cases

First confirmed case in USA was on Jan. 21, 2020

The first case of COVID-19 was reported in China on November 17,2019

This is an **interactive worksheet** with a **play button**, it has a slider where the users can adjust the timeline by **dragging the slider**. The size of the bubbles indicates the intensity in the number of cases and are also color coordinated according to it. We have added annotations for more information accordingly.

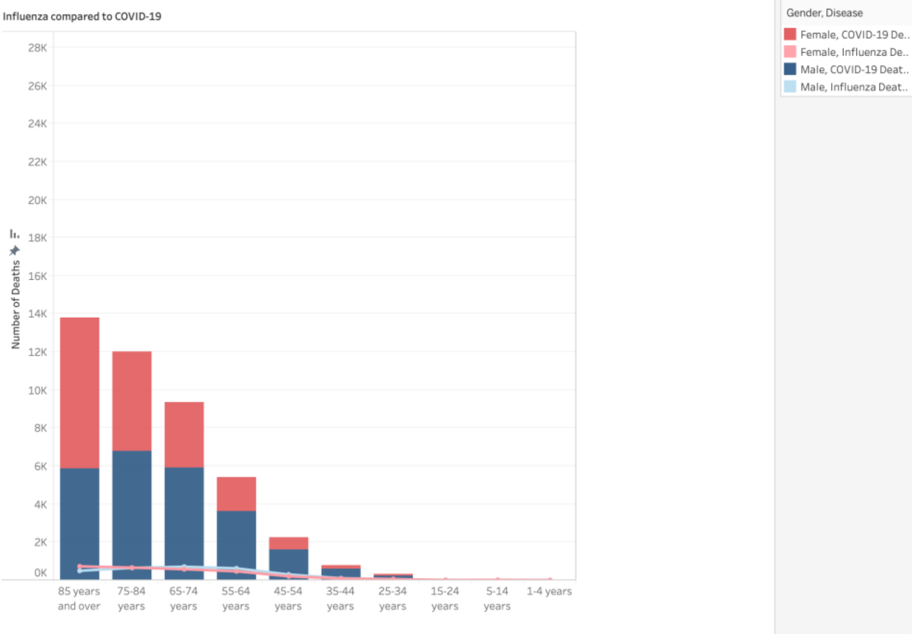
The users can see the **pattern in the growth** of the number of cases by the week. We can see that initially; the bubble starts to get bigger in China. However, it gets more prominent for the USA later on. As the USA has the highest number of cases, **we focus our analysis here after to the USA** to see the different factors that may or may not affect the spread of the pandemic.



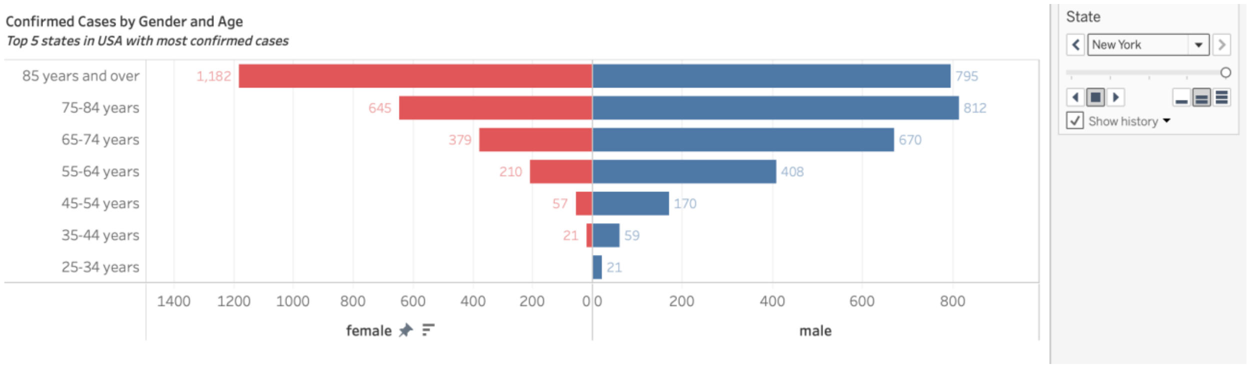
COVID-19 Cases

From the above chart we can see that, **after mid-March the number of confirmed cases starts to grow** exponentially.

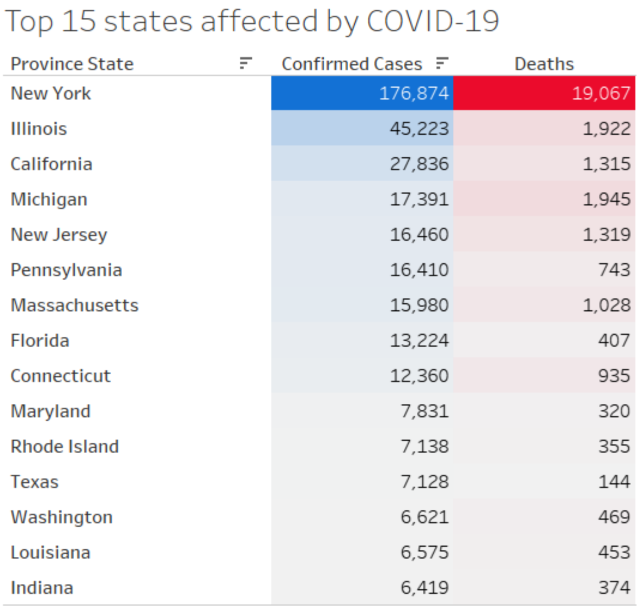
We try to **analyze** the distribution of **COVID-19 cases based on genders and compare it with the numbers of Influenza cases.** Here we compare the annual Influenza cases with the COVID-19 deaths so far.



We can comprehend from the above chart that the deaths due to **COVID-19 are way higher than the Influenza cases.** Also, we can observe that in most cases the number of **deaths is more for the male population** than the female population, ***except for the age group of 85 years and over***. The classification of positive cases of COVID – 19 between Male and Female population is more evident in the below chart. Here we can see that overall **Male population has higher numbers** of COVID-19 cases than the female population. The chart below is at a **much granular level** since it focuses on the **top 5 states that are most affected by COVID-19.**



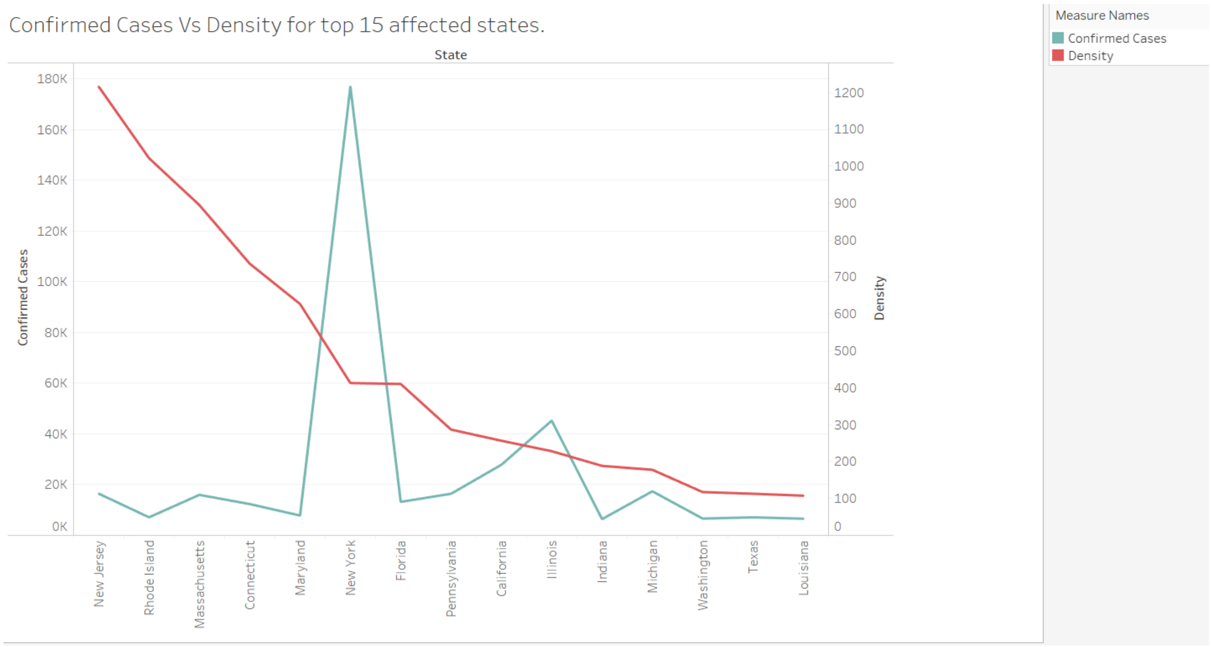
Following is the table consisting of the **confirmed cases and deaths of the top 15 most affected states**. We can see that **New York** tops the table with the **highest number of cases and deaths** followed by Illinois, California, Michigan etc.



Top 15 states affected by COVID-19

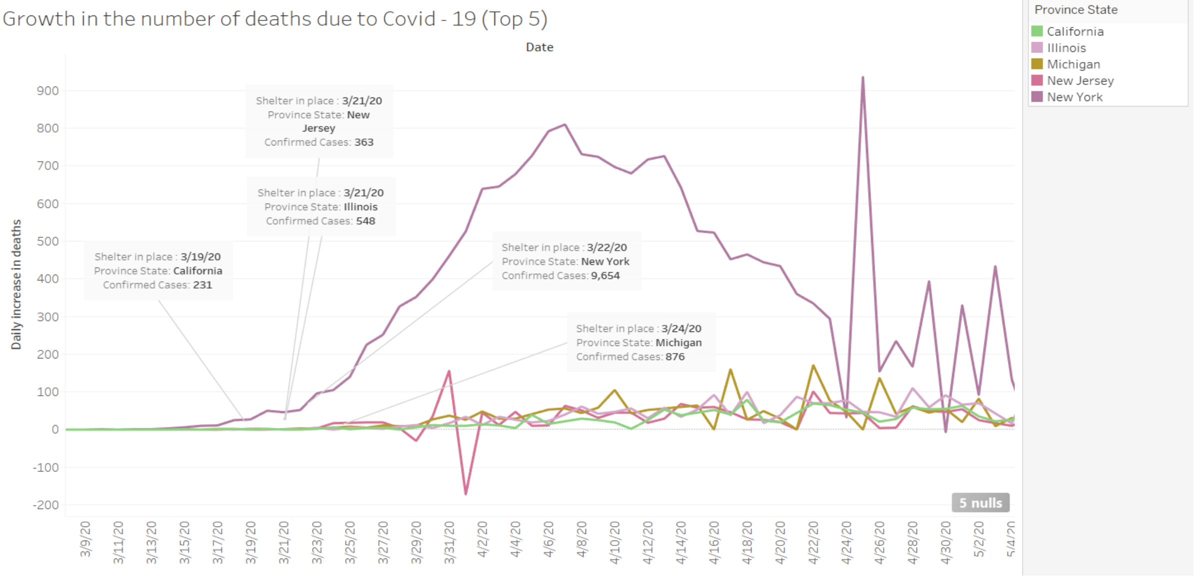
Province State	Confirmed Cases	Deaths
New York	176,874	19,067
Illinois	45,223	1,922
California	27,836	1,315
Michigan	17,391	1,945
New Jersey	16,460	1,319
Pennsylvania	16,410	743
Massachusetts	15,980	1,028
Florida	13,224	407
Connecticut	12,360	935
Maryland	7,831	320
Rhode Island	7,138	355
Texas	7,128	144
Washington	6,621	469
Louisiana	6,575	453
Indiana	6,419	374

Thereafter, we try to analyze if there is a **relation between** the number of **cases in** a **state** and its **density**. From the following line plot, we can see that the relation between confirmed cases and density of a state ***cannot be determined***.



Confirmed Cases Vs Density for top 15 affected states.

Furthermore, we try to see the **Effects of Shelter in Place Order** imposed by the different states and whether it helped in mitigating the adverse effects caused by the pandemic.


Growth in the number of deaths due to Covid - 19 (Top 5)

Shelter in place : 3/21/20
Province State: New Jersey
Confirmed Cases: 363

Shelter in place : 3/21/20
Province State: Illinois
Confirmed Cases: 548

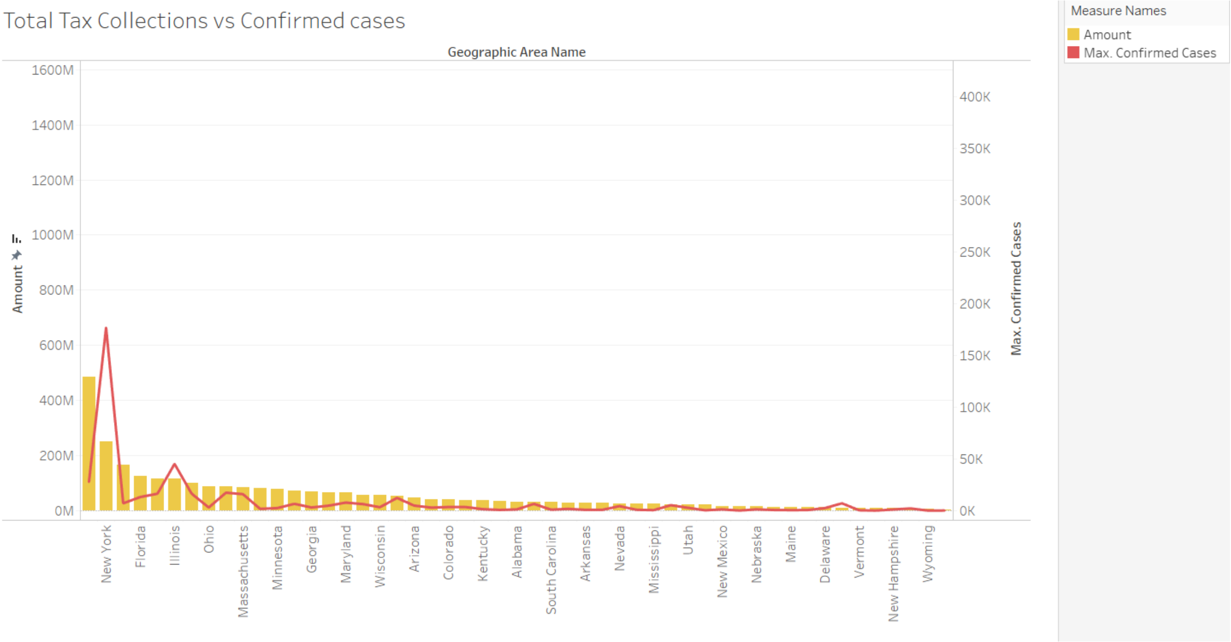
Shelter in place : 3/19/20
Province State: California
Confirmed Cases: 231

Shelter in place : 3/22/20
Province State: New York
Confirmed Cases: 9,654

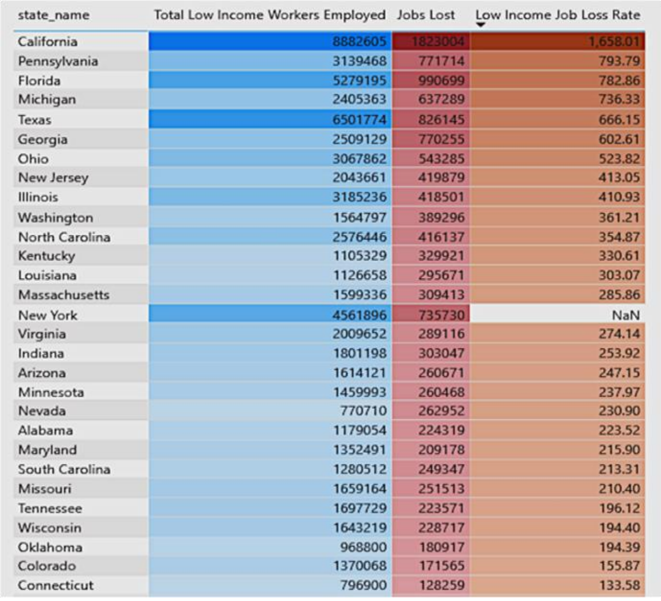
Shelter in place : 3/24/20
Province State: Michigan
Confirmed Cases: 876

We can see that in terms of quick response from the Government with respect to the COVID-19 pandemic, **New York was too far behind** the other top most affected states. New York already had 9654 cases when it implemented the Shelter-in-place order, whereas Illinois had 548, New Jersey had 363, California had 231 and Michigan had 836. This number is significantly less for the other states compared to New York. We can see from the above plot that the ***growth in the number of deaths for New York is exponential even after the Stay- at-home order was imposed by the State Government.***

Furthermore, we have tried to **analyze if tax collections of a state are correlated with the Number of confirmed cases**. We can see from the below plot that there is no significant relation between them

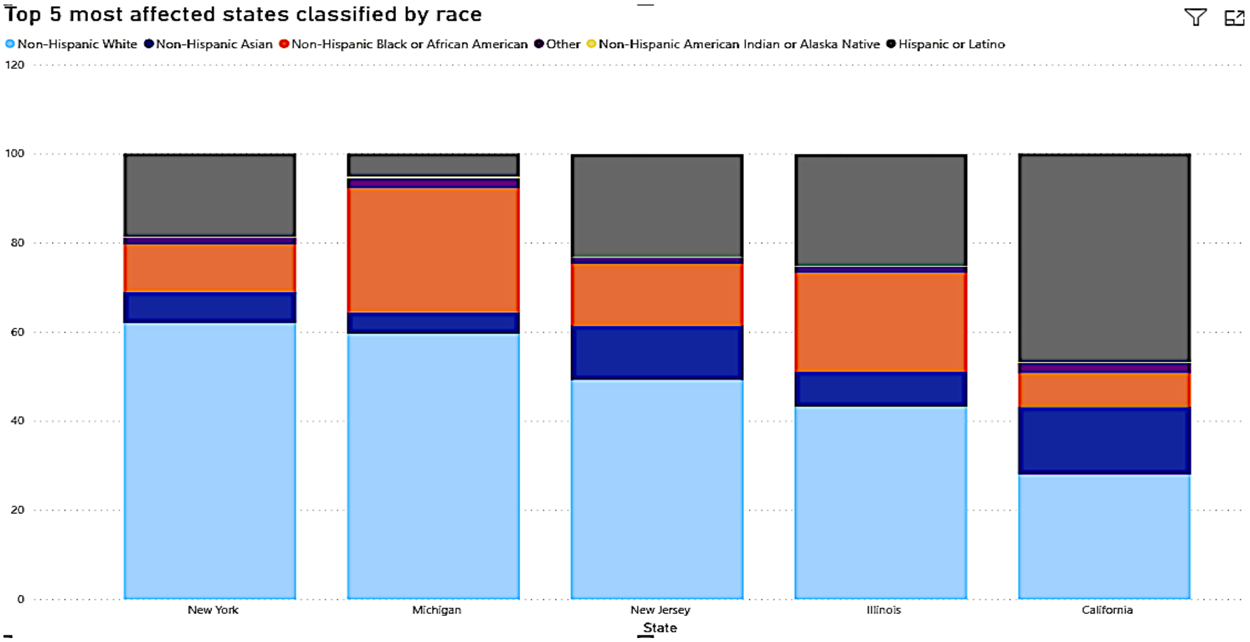

Total Tax Collections vs Confirmed cases

We have used **Power BI to generate reports** and **process the data for Low Income Employees across the USA** as well to classify the confirmed cases based on the race.
We can see from the below table that **California has the highest number of employees who fall in the Low income salary group** (<$40,000 salary). The **number of jobs lost was also the highest for California.** States such as Pennsylvania, Florida, Michigan, Texas, Georgia , Ohio etc. follow California in the list of states where low income jobs are most affected after COVID-19.



state_name	Total Low Income Workers Employed	Jobs Lost	Low Income Job Loss Rate
California	8882605	1823004	1,658.01
Pennsylvania	3139468	771714	793.79
Florida	5279195	990699	782.86
Michigan	2405363	637289	736.33
Texas	6501774	826145	666.15
Georgia	2509129	770255	602.61
Ohio	3067862	543285	523.82
New Jersey	2043661	419879	413.05
Illinois	3185236	418501	410.93
Washington	1564797	389296	361.21
North Carolina	2576446	416137	354.87
Kentucky	1105329	329921	330.61
Louisiana	1126658	295671	303.07
Massachusetts	1599336	309413	285.86
New York	4561896	735730	NaN
Virginia	2009652	289116	274.14
Indiana	1801198	303047	253.92
Arizona	1614121	260671	247.15
Minnesota	1459993	260468	237.97
Nevada	770710	262952	230.90
Alabama	1179054	224319	223.52
Maryland	1352491	209178	215.90
South Carolina	1280512	249347	213.31
Missouri	1659164	251513	210.40
Tennessee	1697729	223571	196.12
Wisconsin	1643219	228717	194.40
Oklahoma	968800	180917	194.39
Colorado	1370068	171565	155.87
Connecticut	796900	128259	133.58

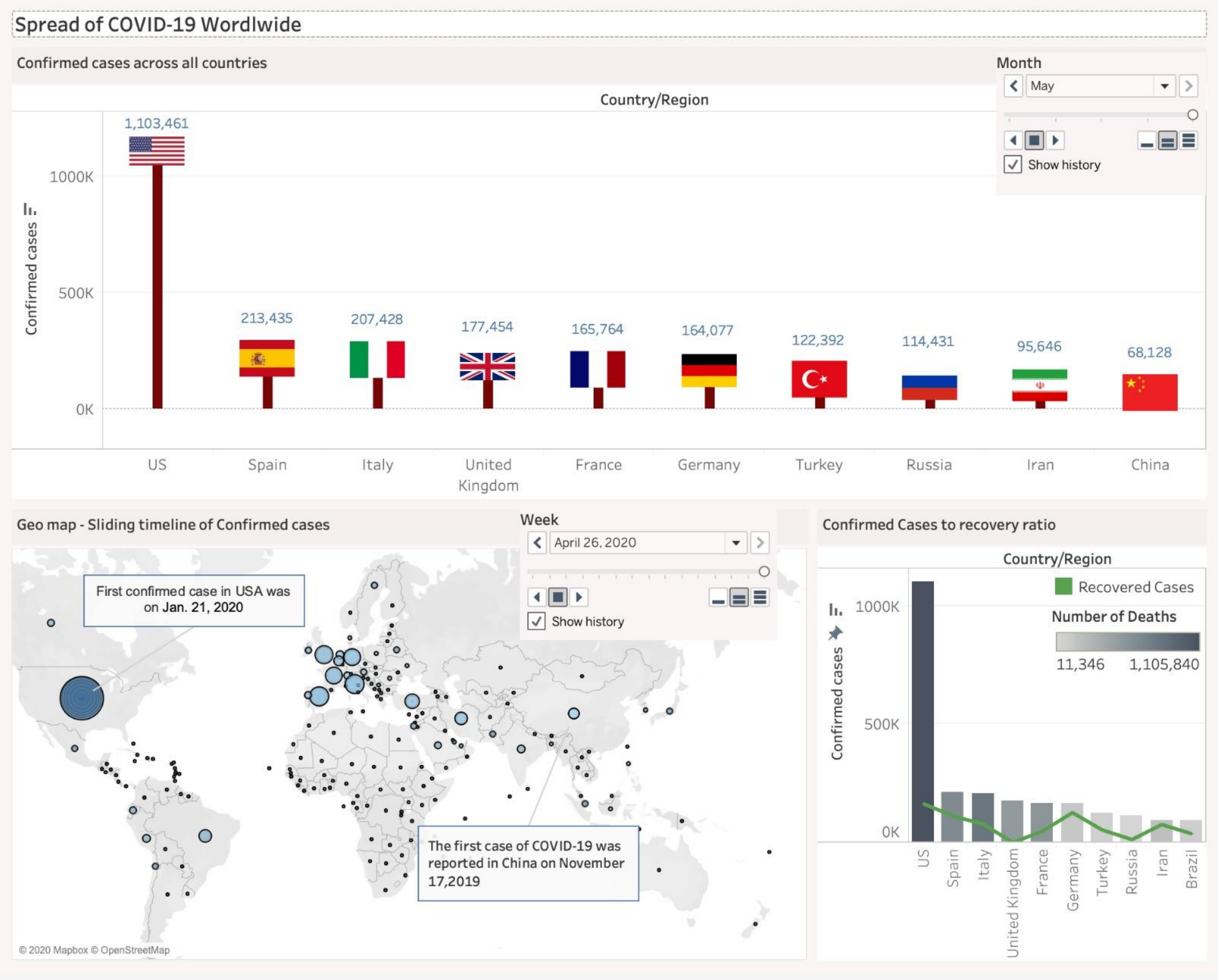
From the following visualization we can see the distribution of **COVID-19 cases based on race for the top 5 most affected states in the USA**. We can see that **Non hispanic whites are most affected in by COVID-19**, whereas in California, hispanic or latino communities have been hit the worst due to the pandemic.



Dashboards:

Finally, we collated selected sheets from the workbook into our Dashboard.

Dashboard 1:



Spread of COVID-19 Wordlwide

Confirmed cases across all countries

Month: May

☑ Show history

Geo map - Sliding timeline of Confirmed cases

Week: April 26, 2020

☑ Show history

First confirmed case in USA was on Jan. 21, 2020

The first case of COVID-19 was reported in China on November 17,2019

Confirmed Cases to recovery ratio

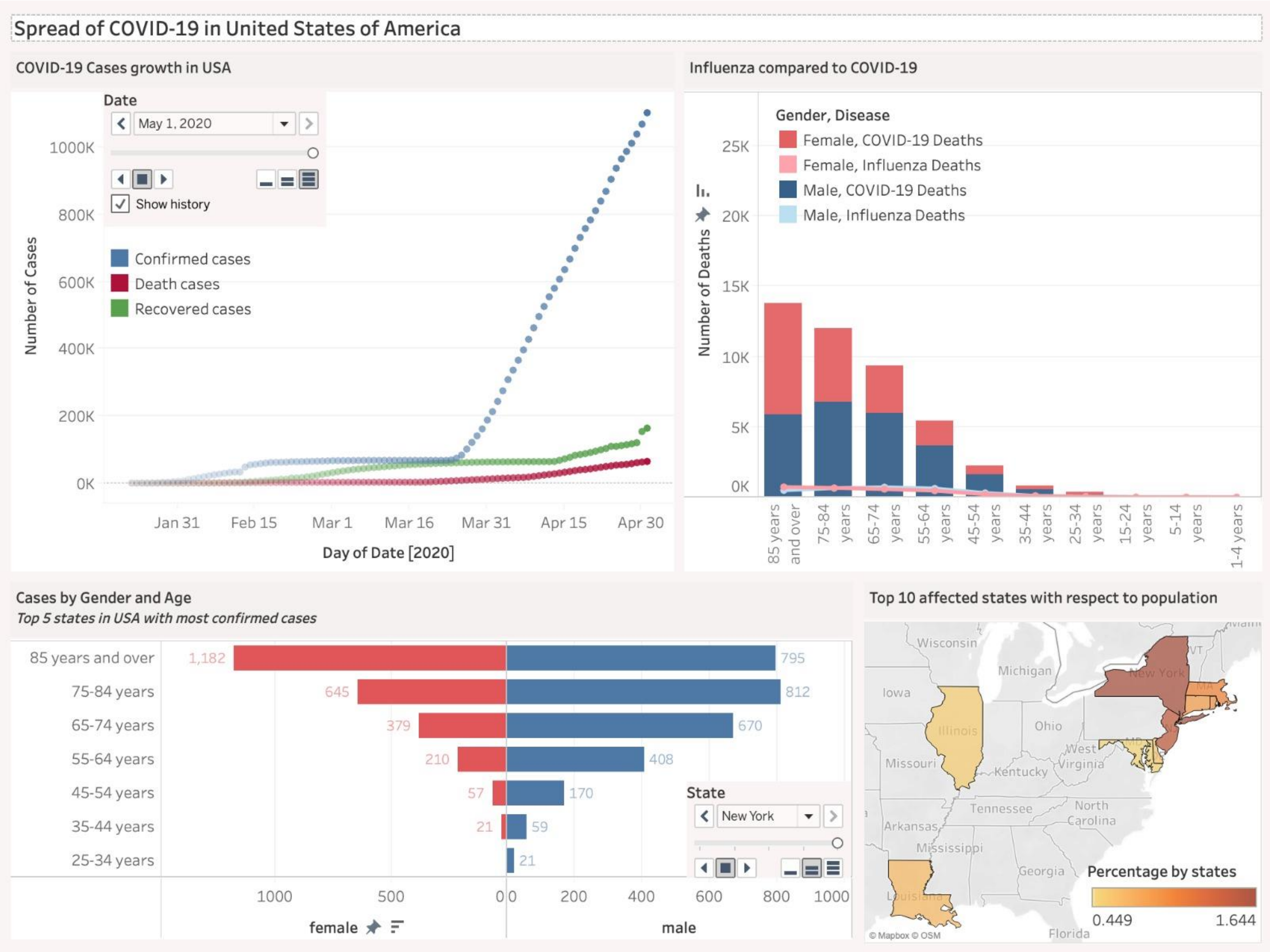
All the worksheets can be found in the attached Tableau packaged workbook.

Focus:

The focus of this dashboard is to give an overview of COVID-19 since it has been declared as a pandemic, it is essential to know where the world stands and how everyone is dealing with the disease. This above dashboard gives information on the confirmed cases in the world along with the deaths and recovered cases. A timeline has been associated in 2 tiles of this dashboard and can be adjusted according to the user's requirement. We wanted to make this dashboard interactive because cases of COVID-19 keep changing by the minute, it is not possible for users to process so much information, therefore they can drag the slider back and forth as per their needs.

Dashboard 2:

As mentioned above we further did our analysis on the cases in United States of America, dashboard for the same can be found below :



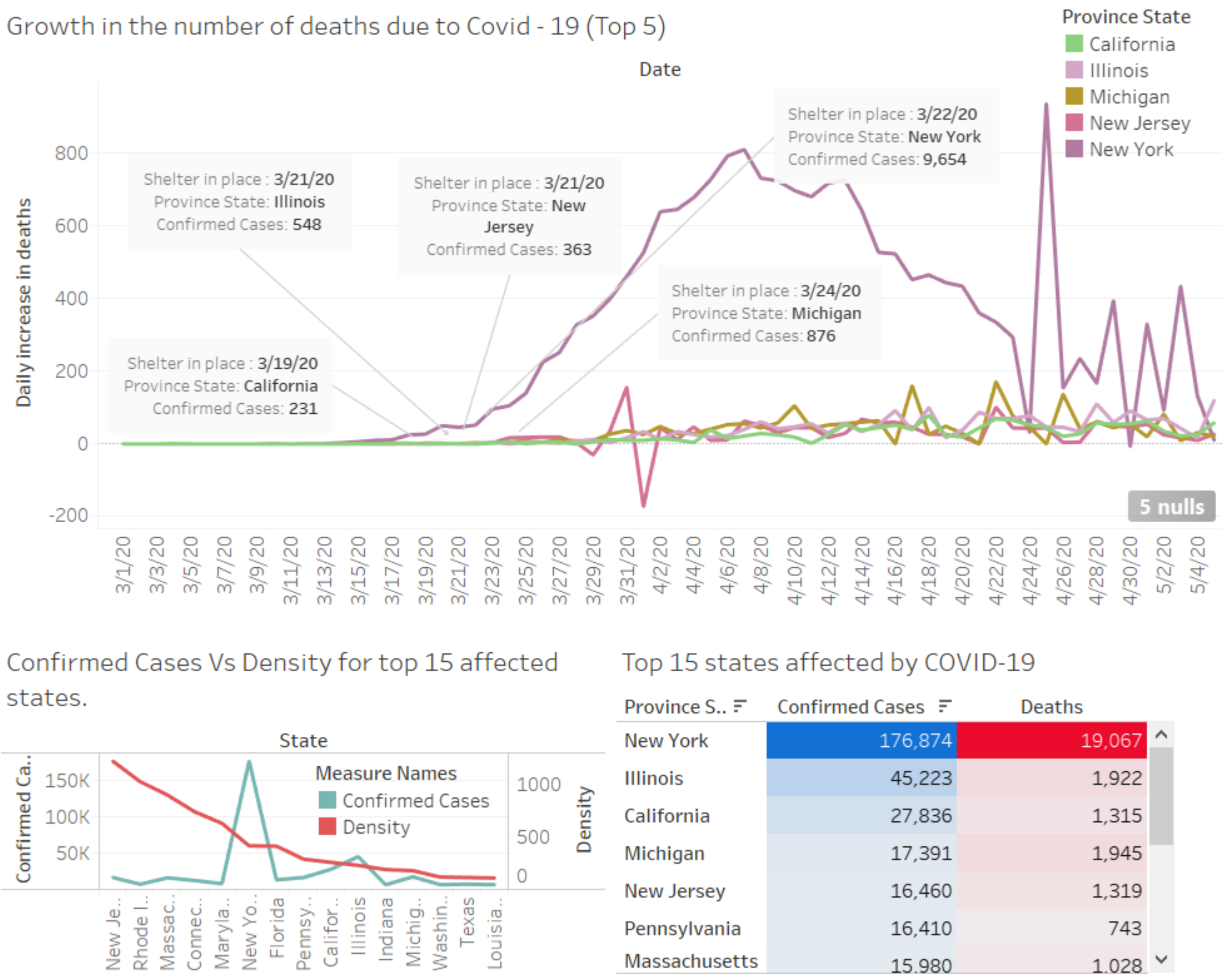
All the worksheets can be found in the attached Tableau packaged workbook.

Focus:

The focus of this dashboard is to get a granular perspective. Here we are focusing on the United States of America and associating the cases with Influenza, gender and population. Two of the tiles have a play button options for users to analyze and experiment. The explanation and observations of each have been explained above. In the last tile, we computed the most affected states by comparing the number of cases with the population of that particular state.

Dashboard 3 :

We can learn more about the spread of COVID-19 in the USA from the dashboard below:



Growth in the number of deaths due to Covid - 19 (Top 5)

Shelter in place : 3/22/20
Province State: New York
Confirmed Cases: 9,654

Shelter in place : 3/21/20
Province State: Illinois
Confirmed Cases: 548

Shelter in place : 3/21/20
Province State: New Jersey
Confirmed Cases: 363

Shelter in place : 3/24/20
Province State: Michigan
Confirmed Cases: 876

Shelter in place : 3/19/20
Province State: California
Confirmed Cases: 231

Confirmed Cases Vs Density for top 15 affected states.

Top 15 states affected by COVID-19

Province S..	Confirmed Cases	Deaths
New York	176,874	19,067
Illinois	45,223	1,922
California	27,836	1,315
Michigan	17,391	1,945
New Jersey	16,460	1,319
Pennsylvania	16,410	743
Massachusetts	15,980	1,028

Focus:

The focus of this dashboard is to understand the spread and growth of COVID-19 cases across the different states of the USA. As the US is one of the most affected countries in the world, we focused on this particular country to understand the spread of the virus. In the above dashboard, we can see plots for the growth in the number of deaths in the top 5 most affected states. We can also see the table where states are sorted based on the number of confirmed cases. Furthermore, we have tried to find the correlation between the Density and Confirmed cases in the 15 most affected states in the country.

Attachments:

3 Tableau Dashboards
1 Power BI report

We used Tableau to create the dashboards and add interactivity for better understanding. PowerBI was used to generate reports and a few visualizations. We also used PowerBI to go through all our datasets thoroughly in order to perform our analysis.

Thank you