**The COVID-19 Project**

**Introduction:-** COVID-19 commonly known as coronavirus was first identified in Wuhan, China in 2019, since then it has spread worldwide and caused a pandemic. It is an infectious disease caused by severe acute respiratory syndrome with a fatality rate of nearly 1 percent.

COVID-19 entered US through following possibilities

1. Imported cases in explorers
2. Cases among close contacts of a known case
3. Community-procured situations where the wellspring of the disease is obscure.



Coronavirus Cases in US

The infection that causes COVID-19 is thought to spread for the most part from individual to individual, predominantly through respiratory beads created when a contaminated individual hacks or wheezes. These beads can land in the mouths or noses of individuals who are close by or potentially be breathed in into the lungs. Spread is more probable when individuals are in close contact with each other (inside around 6 feet).

The coronavirus pandemic is influencing each part of life in the United States now, and with that effect come some hard decisions. Who gets money related assistance from the national government and what amount? By what method will specialists choose who gets treatment and who doesn't if medical clinic assets are inadequate to treat everyone who needs consideration? The share market has fallen, a lot of businesses are closed.

**Agencies responsible for pandemic disease like COVID 19 In US**

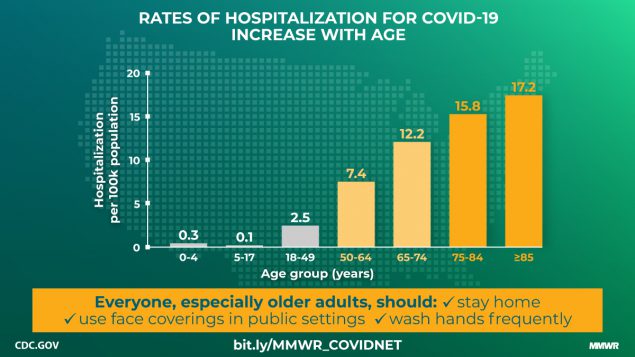
1. White House Coronavirus Task Force
2. Centers for Disease Control and Prevention (CDC)

* **White House Coronavirus Task Force**

The **White House Coronavirus Task Force** is a [United States Department of State](https://en.wikipedia.org/wiki/United_States_Department_of_State) task force that "coordinates and oversees the Administration's efforts to monitor, prevent, contain, and mitigate the spread" of the [coronavirus disease](https://en.wikipedia.org/wiki/Coronavirus_disease_2019) (COVID-19).

Dr. Anthony Fauci the "face of the federal government's response". **Anthony Stephen Fauci** is an American physician and [immunologist](https://en.wikipedia.org/wiki/Immunology) who has served as the director of the [National Institute of Allergy and Infectious Diseases](https://en.wikipedia.org/wiki/National_Institute_of_Allergy_and_Infectious_Diseases) (NIAID) since 1984. Since January 2020, he has been one of the lead members of the [Trump Administration's](https://en.wikipedia.org/wiki/Presidency_of_Donald_Trump) [White House Coronavirus Task Force](https://en.wikipedia.org/wiki/White_House_Coronavirus_Task_Force) addressing the [2019–20 coronavirus pandemic](https://en.wikipedia.org/wiki/2019%E2%80%9320_coronavirus_pandemic) [in the United States](https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_the_United_States).

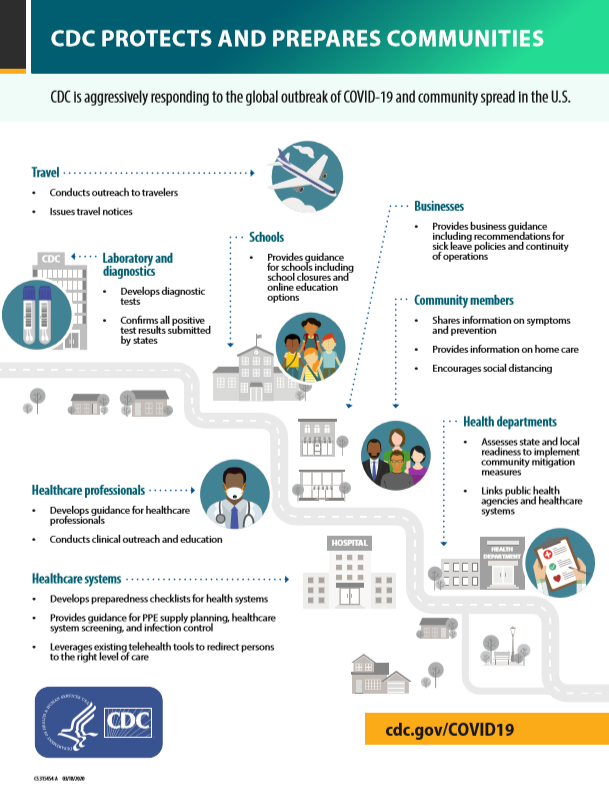
* **Centers for Disease Control and Prevention** (**CDC**)

The **Centers for Disease Control and Prevention** (**CDC**) is the leading [national public health institute](https://en.wikipedia.org/wiki/National_public_health_institutes) of the United States. Its main goal is to protect [public health](https://en.wikipedia.org/wiki/Public_health) and safety through the control and prevention of disease, injury, and [disability](https://en.wikipedia.org/wiki/Disability) in the US and internationally.

**Role of CDC** : CDC is responsible for controlling the introduction and spread of infectious diseases, and provides consultation and assistance to other nations and international agencies to assist in improving their disease prevention and control, environmental health, and health promotion activities.

**Dataset website** - <https://covidtracking.com/data>

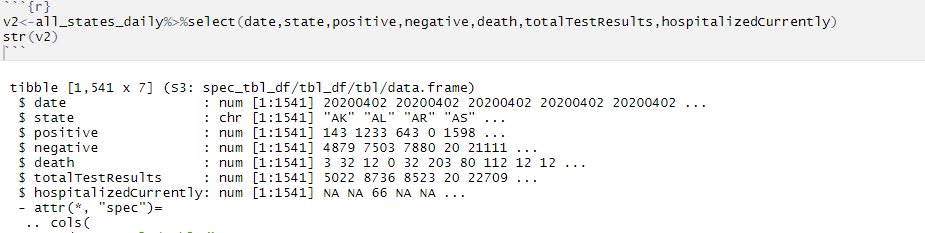
The COVID Tracking Project collects its data from state/district/territory public health authorities—or, occasionally, from trusted news reporting, official press conferences, or (very occasionally) tweets or Facebook updates from state public health authorities or governors.

**Summary of CDC’ action plans:-**

**Description of the data :**

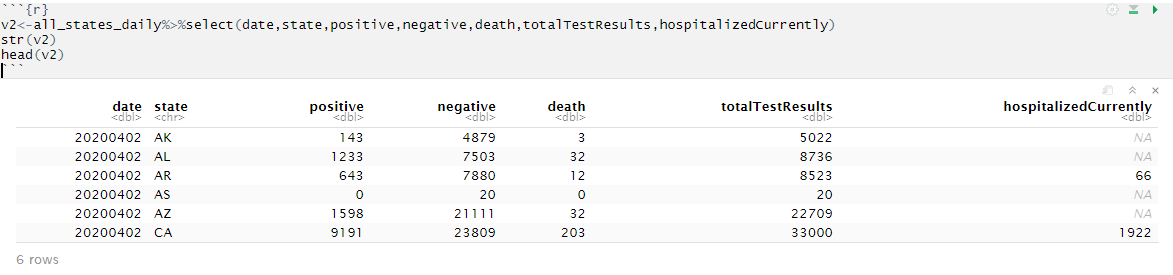
COVID-19 dataset comprises the total number of tests conducted, breaking out positive, negative, and hospitalized patients. The data is collected from ”The COVID Tracking Project” collects its data from state/district/territory public health authorities—or, occasionally, from trusted news reporting, official press conferences, or tweets or Facebook updates from state public health authorities or governors.

**Structure of COVID-19 Dataset in R-studio :**



**Description of columns -**

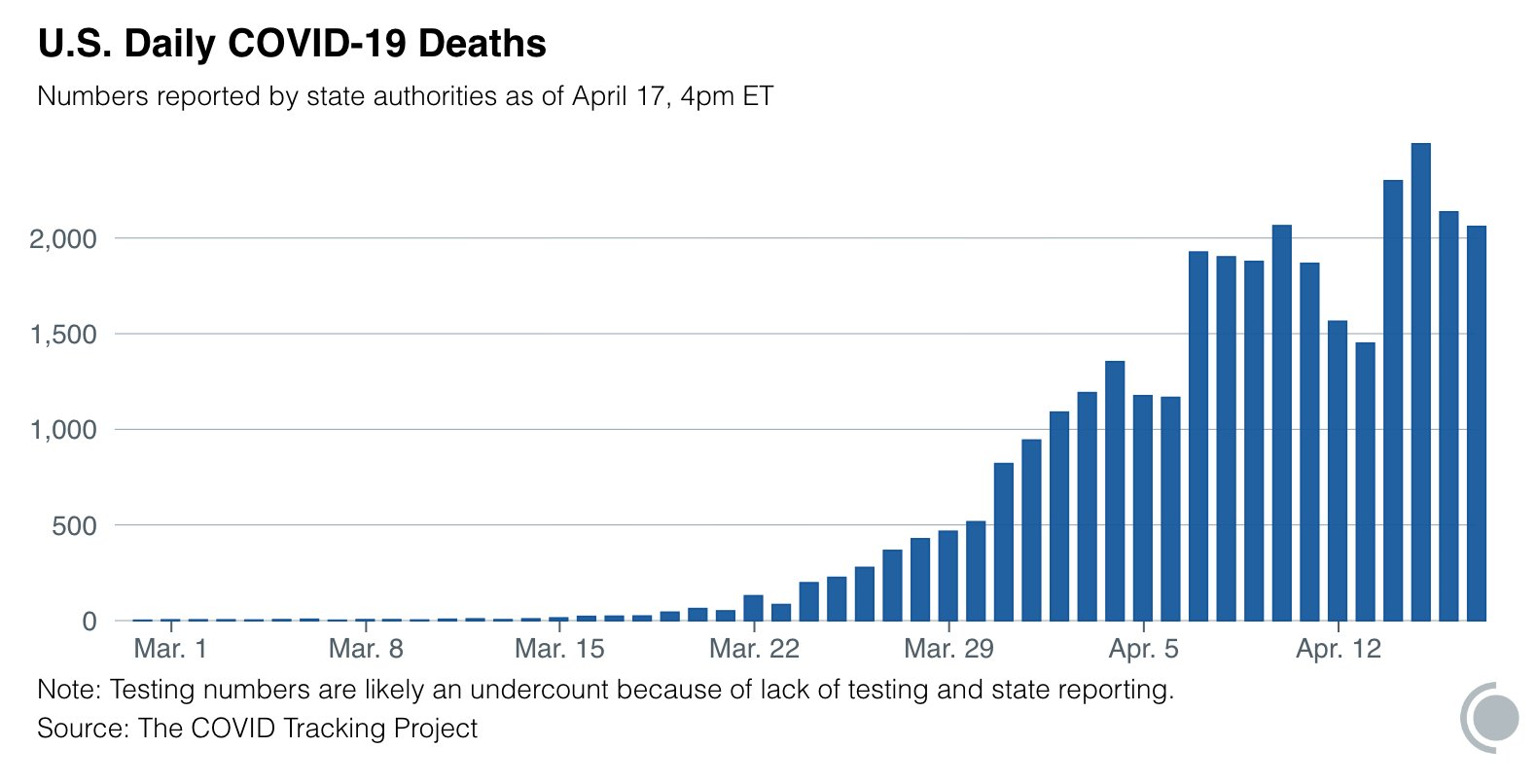
* State - State or territory postal code abbreviation.
* Positive - Total cumulative positive test results.
* Negative - Total cumulative negative test results
* Death - Total cumulative number of people that have died.
* Total test - Total no of tests conducted.
* Hospitalized - Total cumulative number of people hospitalized.

****

**Limitations** - Not all the states consistently report their test results and regularly. In such cases, they use other reporting tools like directly asking state officials, watching news conferences, gleaning information from trusted news sources. Moreover, since the symptoms are not visible until 14 days, the number of actual positive cases may be more than reported to the state/district/territory public health authorities, etc.

**Exploratory analysis of COVID-19 dataset**

This project will conduct exploratory analysis to derive the following insights from the COVID-19 datasets :

****

1. Identifying which state in the US has been hit hardest.
2. Comparing hardest-hit state to the rest of the states of US
3. Drawing a line plot to visualize the confirmed cases in hardest-hit state vs rest of the states
4. Studying the trendline to get more insights about the top 5 states
5. Assessing future problems due to the spread of the virus by computing the growth rate of the spread of the virus
6. Deriving ratios to get insights from confirmed cases, total deaths, recoveries, and hospitalized cases
7. To Predict when the death rate is going to reduce significantly

**Summary :**

Deliverable 3 consists of the details of the tools to be used for data analytics. It would describe the process of how data is extracted from the source website and loaded into the tools for further analysis. Subsequently, the data Cleaning and Manipulation process will be carried out to ensure that data is correct, consistent, and usable by identifying any errors or corruption in the data. Lastly, the basic exploration process of data is initiated to get insights into how different states of the United States are affected.

**Data Analysis Tools: R-studio, Tableau**

The tool that we will use to manipulate the data will be R-studio for data analysis and Tableau for visualizations.

* **R-Studio**:

RStudio is an integrated development environment for R, a programming language for statistical computing and graphics. It provides free and open-source tools for R and enterprise-ready professional software for data science teams to develop and share their work at scale. RStudio makes it easy to set your working directory and access files on your computer.

**R-Studio for Data Analysis:**

R is very important in data science because of its versatility in the field of statistics. R is usually used in the field of data science when the task requires a special analysis of data for standalone or distributed computing.R is also perfect for exploration. It can be used in any kind of analysis work, as it has many tools and is also very extensible. Additionally, it is a perfect fit for big data solutions. Following are some of the highlights which show why R is important for data science:

* Data analysis software.
* Statistical analysis environment
* Open-source

So, most of the development of the R language is done by keeping data science and statistics in mind. As a result, R has become the default choice for data science applications and data science professionals.

* **Tableau:**

Tableau is a powerful and fastest-growing data visualization tool used in the Business Intelligence Industry. It helps in simplifying raw data into a very easily understandable format.

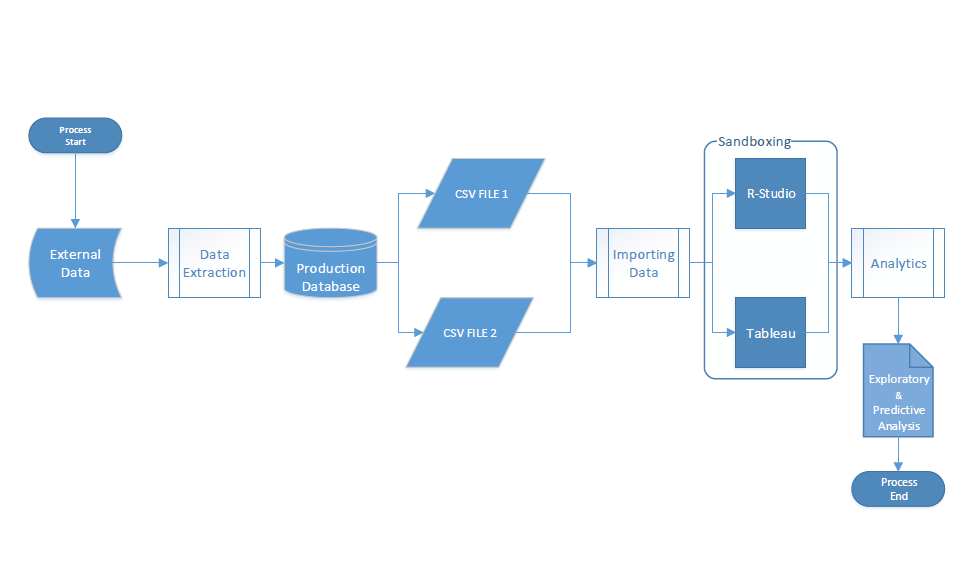
Tableau is an interactive, self-service reporting and analytics tool that enables faculty and staff to integrate and combine data from multiple sources into visualizations and be accessed in a single desktop environment using Tableau Desktop or through a shared dashboard.

**Tableau for visualizations :**

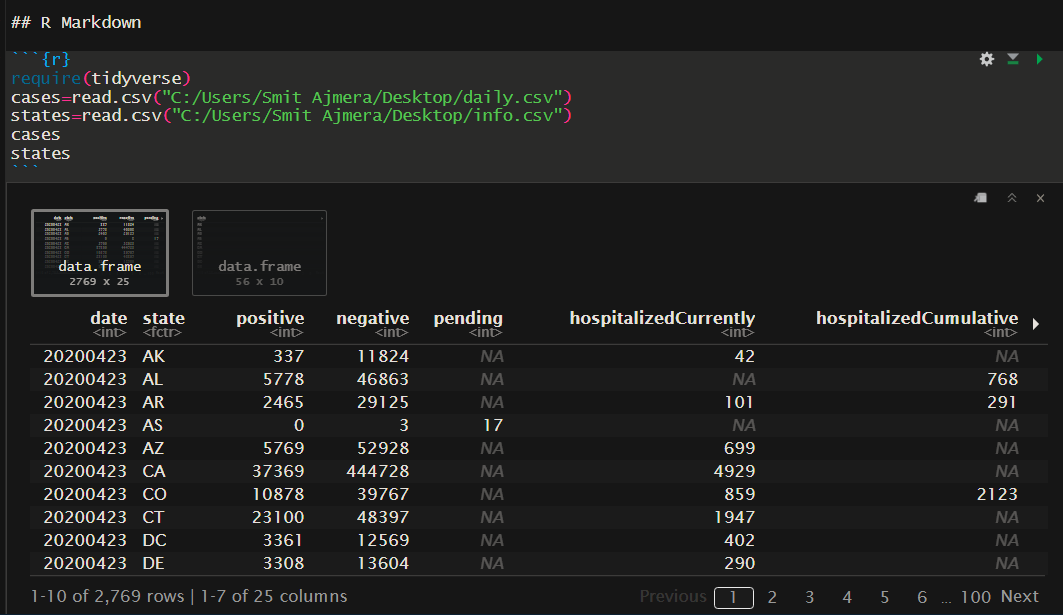
Data analysis is very fast with Tableau and the visualizations created are in the form of dashboards and worksheets. Tableau can help anyone see and understand their data. Connect to almost any database, drag and drop to create visualizations, and share with a click. Tableau has a feature for sourcing configuration that can be connected to several data sources for pulling or crawling. This is particularly good if you are looking to analyze and compare different entities of data.

**Data Flow Chart**

The software Microsoft Visio is used to prepare a process flow chart of how the data moves throughout different phases of the project. Firstly, the data is extracted from covidtracking.com and stored in the production database that is the local server. The data from the Production database is imported into data analytics tools (R-studio and Tableau) for sandboxing, where data cleaning, manipulation, and analysis is carried out. Finally, the Information and insights from the data analysis are documented.



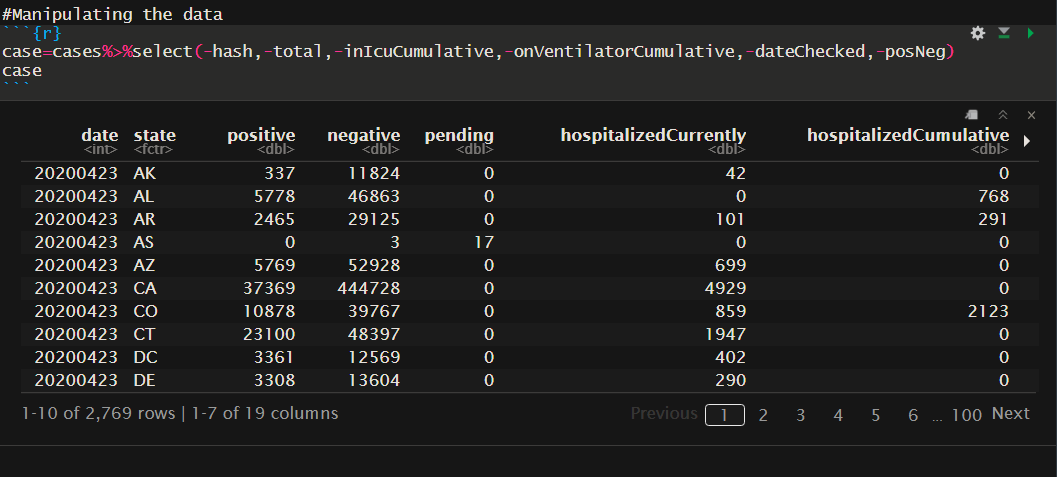
The data of COVID-19 is extracted from covidtracking.com and then loaded into Rstudio. The below image displays the R-Studio interface and how the data is loaded into it.



**Data cleaning and manipulation**

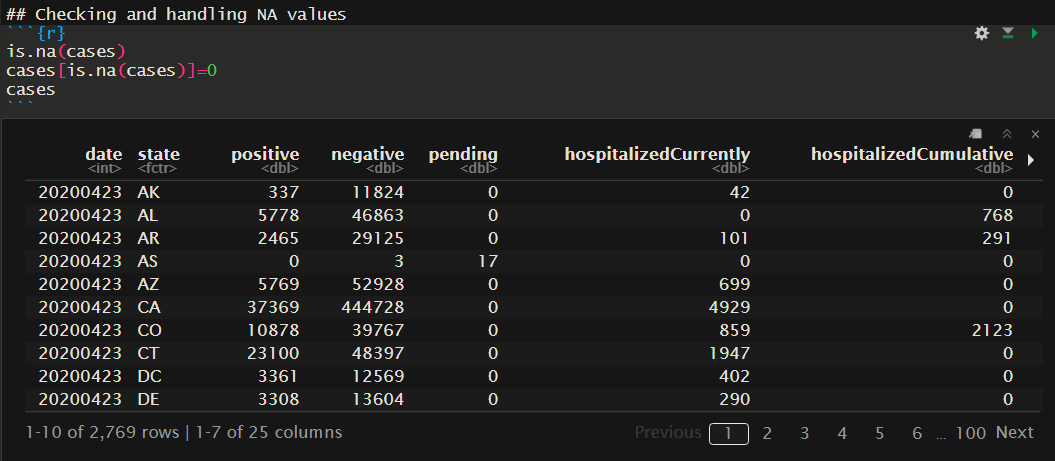
Firstly, the required packages are installed and recalled to load the data into R-studio. Then the columns in the data are analyzed to eliminate the redundant columns from the dataset. Secondly, the data is cleaned, since it will improve the quality of the data and the performance of the model. Then data is checked for any missing values in the columns and subsequently, the missing values are removed from the dataset and by replacing them with 0. Since, we only have to deal with missing values from numerical variables, we will replace it with 0.

* **Manipulation of the dataset**:- We will analyze the data set and eliminate all the columns that will not contribute to prediction.



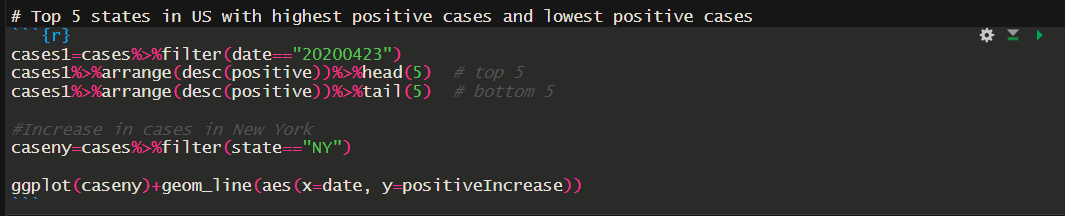
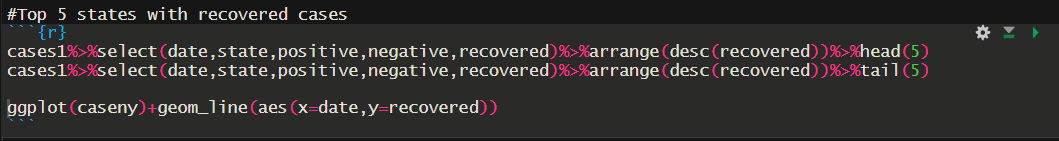
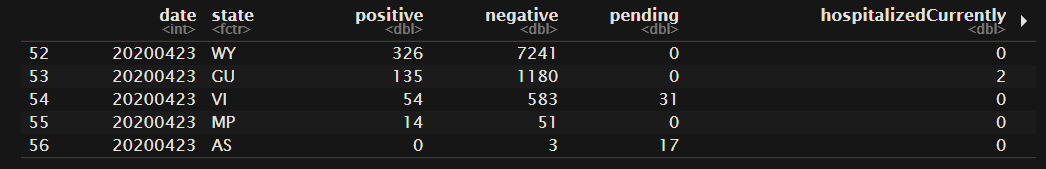
* **Data cleaning :**

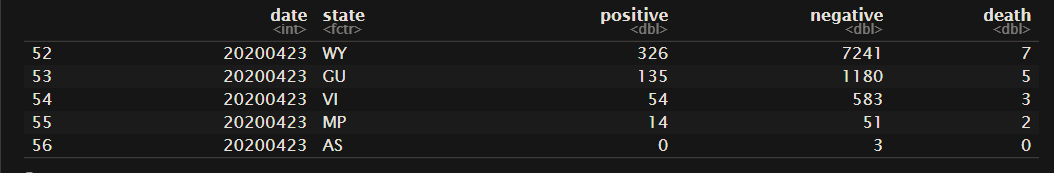
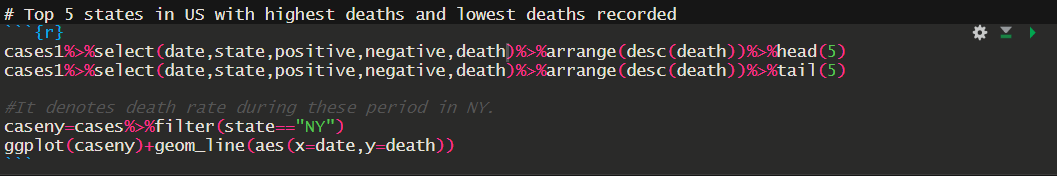
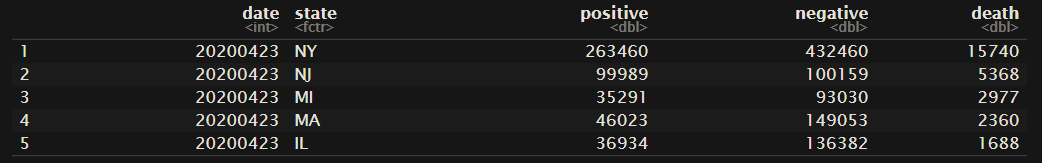
As we can observe that there are a lot of missing values(NA) in the data set and hence we will try to get rid of them by replacing them with 0, using the following code. Since, we only have to deal with missing values from numerical variables, we will replace it with 0.



**Exploring data**

We will perform a basic exploration of data to get an overview of worst-hit states in the US.

* **Top and bottom 5 states based on positive cases.**
* **Top and bottom 5 states based on Recovery.**
* **Top and bottom 5 states based on deaths.**



From the above, we can observe that New York and New Jersey are the hardest hit states with the most number of positive cases and deaths. Further in deliverable 4, we will perform an in-depth analysis of the hardest-hit states while comparing it with others.

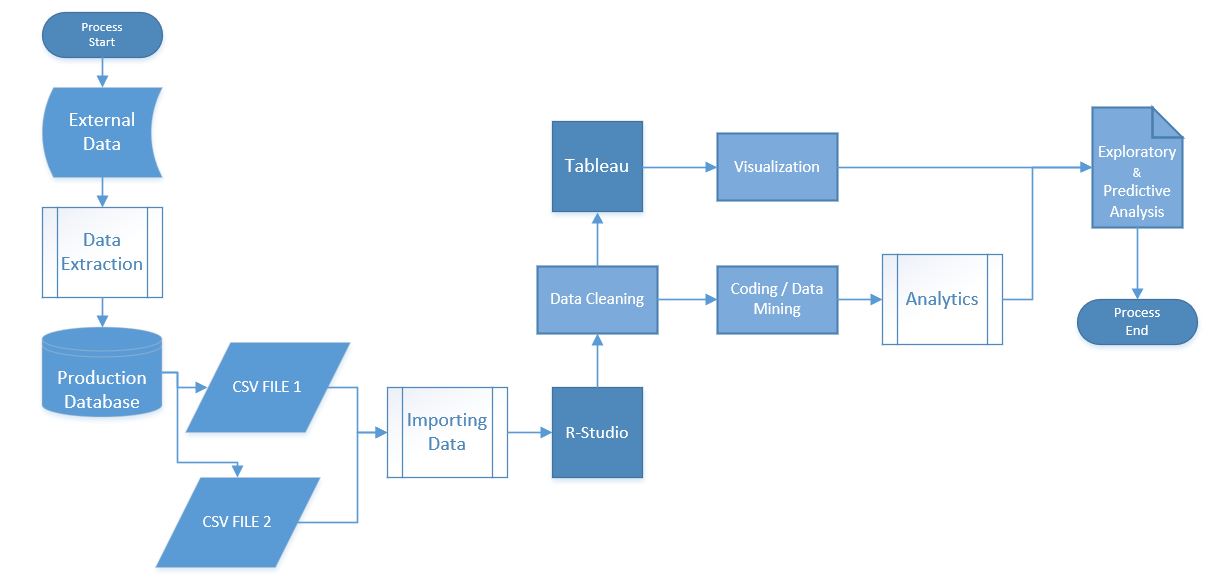
**List of States**

The below list consists of the preview of the states with total positive, total recovered, and total deaths.



**Summary:**

Deliverable 4 consists of Visual Representation of the data and definition of production and analytical Data sets.Visualization is essential for analyzing data and making decisions based on that data. It allows people to quickly and easily see and understand patterns and relationships and spot emerging trends that might go unnoticed with just a table or spreadsheet of raw numbers. And in most cases, no specialized training is required to interpret what’s presented in the graphics, enabling universal understanding.



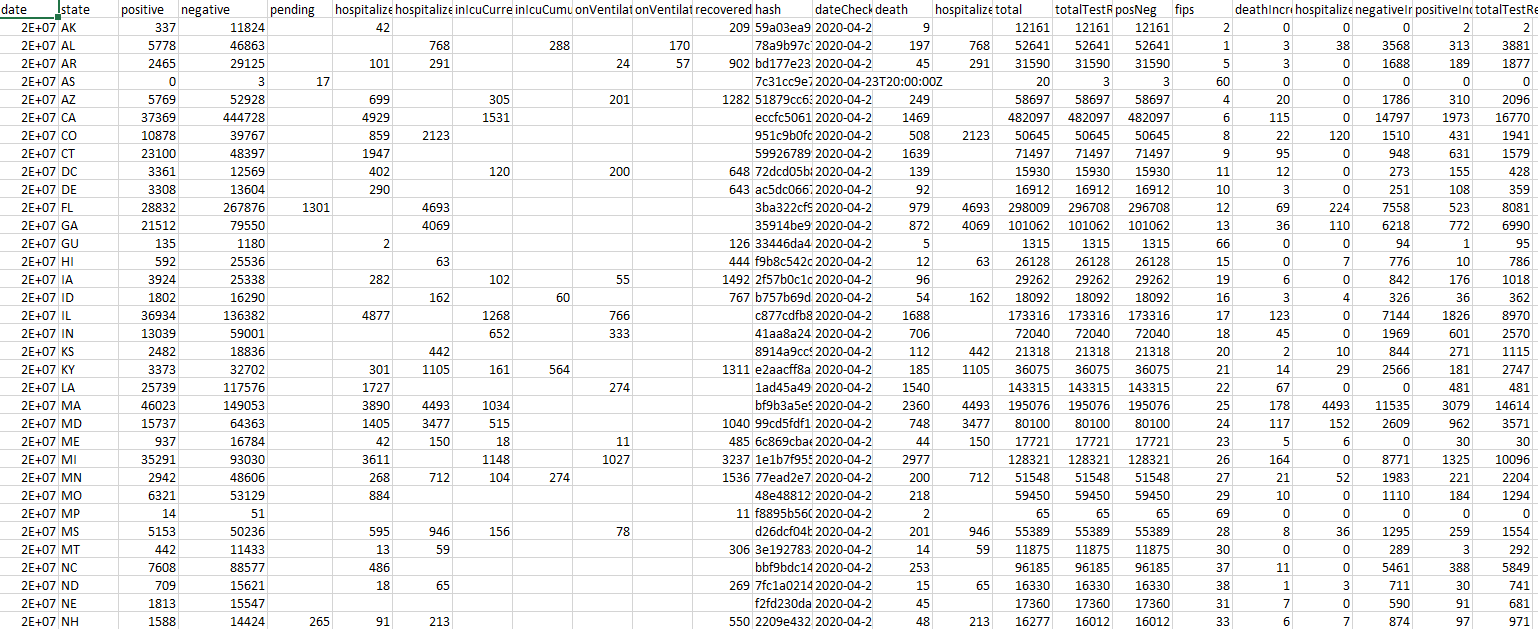
**How data is coming from the production:**

The software Microsoft Visual Studio is used to prepare a process flow chart of how the data moves throughout different phases of the project. Firstly, the data is extracted from covidtracking.com and stored in the production database that is the local server. The data is stored in a CSV file and uploaded on the website, which can be accessed by anyone through the website. The data from the Production database is imported into data analytics tools (R-studio and Tableau) for sandboxing, where data cleaning, manipulation, and analysis is carried out. Finally, the Information and insights from the data analysis are documented.

**Production data set:**

Production data is information that is persistently stored and used by professionals to conduct business processes. It must be accurate, documented, and managed on an on-going basis to ensure its value to the organization.The original data sets that are analyzed using BI tools such as Power BI, Tableau, R studio for getting desired results, the original data set that is being used is known as the Production data set. Production data can include productivity on the amount of product you're making to all the different measurements you must take for a quality check.

Our data set is taken from <https://covidtracking.com/data/us-daily> and it contains positive, negative, pending, death, and hospitalized cases of covid-19 in different states of the US.



**Transformation:**

Most parametric tests require that residuals be normally distributed and that the residuals be homoscedastic.

One approach when residuals fail to meet these conditions is to transform one or more variables to better follow a normal distribution. Often, just the dependent variable in a model will need to be transformed. However, in complex models and multiple regression, it is sometimes helpful to transform both dependent and independent variables that deviate greatly from a normal distribution.

There is nothing illicit in transforming variables, but you must be careful about how the results from analyses with transformed variables are reported. To present means or other summary statistics, you might present the mean of transformed values, or back transform means to their original units. Some measurements in nature are naturally normally distributed. Other measurements are naturally log-normally distributed.

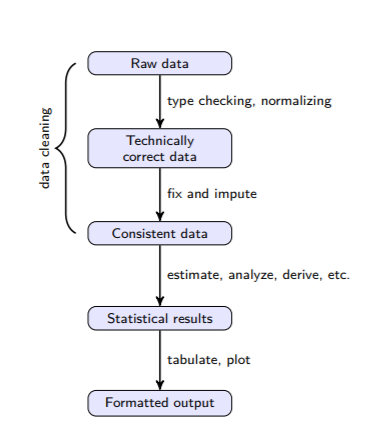
**Reduction of files:**

File compression reduces the amount of space needed to store data. Using compressed files can free up valuable space on a hard drive, or a web server.

There are circumstances when datasets become too large to read directly into R. To overcome this limitation, external stream processing tools can be used to preprocess large text files. And will break a large multi GB file into many chunks, each of which is more manageable for R.

**Cleansing:**

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a recordset, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the coarse data



The first state (Raw data) is the data as it comes in. Raw data files may lack headers, contain wrong data types (e.g. numbers stored as strings), wrong category labels, unknown or unexpected character encoding, and so on. In short, reading such files into an R data.frame directly is either difficult or impossible without some sort of preprocessing. Once this preprocessing has taken place, data can be deemed Technically correct. That is, in this state data can be read into an R data.frame, with correct names, types, and labels, without further trouble. However, that does not mean that the values are error-free or complete. Consistent data is the stage where data is ready for statistical inference. It is the data that most statistical theories use as a starting point. Ideally, such theories can still be applied without taking previous data cleaning steps into account.

**Analytical data set:**

Analytical data is a collection of data that is used to support decision making and/or research. It is historical data that is typically stored in a read-only database that is optimized for data analysis.

Analytics is simply defined as an informational analysis that is derived from the collection of data or statistics. Nearly every single function that exists in business operations can be a data set, and the mere collecting of that data can result in analytics that has an immediate impact on an organization’s bottom line. It all begins with the production of data collection. Analytics are only valuable if the data being collected is accurate, and pertinent to what needs to be analyzed.

We have analyzed the following insights from the COVID-19 datasets:

1. Identifying which state in the US has been hit hardest.
2. Comparing hardest-hit state to the rest of the states of US
3. Drawing a line plot to visualize the confirmed cases in hardest-hit state vs rest of the states
4. Studying the trendline to get more insights about the top 5 states.
5. Deriving ratios to get insights from confirmed cases, total deaths, recoveries, and hospitalized cases
6. To Predict when the death rate is going to reduce significantly
7. Which state will be the first to get out of this pandemic.

R and Tableau are used to perform these analyses. As using Tableau, you would be able to perform predictive analytics with R through leveraging powerful R Packages.

So even non-programmers could simply use these custom Analytics calculations coded in Tableau upon integration with the R server, to derive the output and useful insights through amazing visualizations provided from Tableau. Which allows performing analytics and data visualizations in just one go.

**Analysis of the data:**

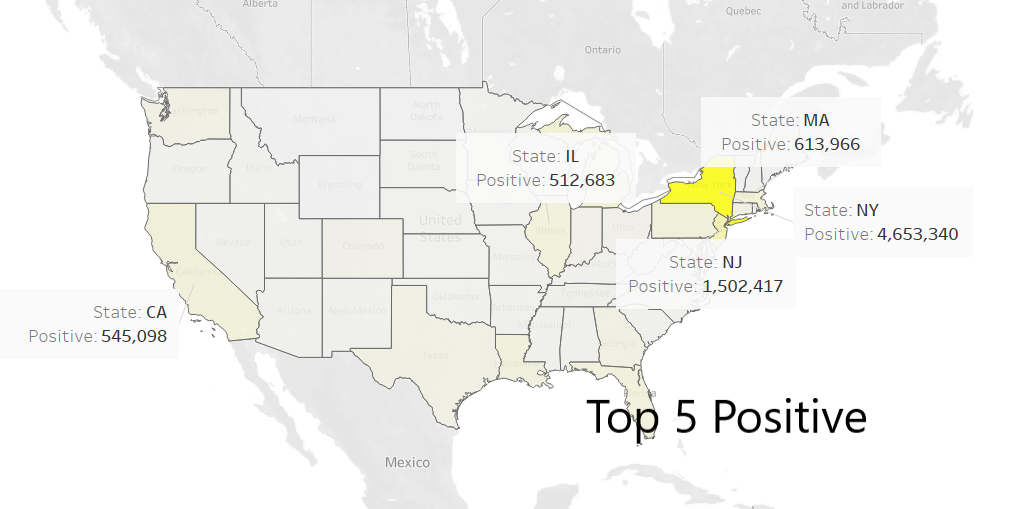
Data analysis is a process of inspecting, [cleansing](https://en.wikipedia.org/wiki/Data_cleansing), [transforming](https://en.wikipedia.org/wiki/Data_transformation), and [modeling](https://en.wikipedia.org/wiki/Data_modeling) [data](https://en.wikipedia.org/wiki/Data) with the goal of discovering useful information, informing conclusions, and supporting decision-making. Analysis refers to breaking a whole into its separate components for individual examination. Data analysis is a [process](https://en.wikipedia.org/wiki/Process_theory) for obtaining raw data and converting it into information useful for decision-making by users. Data is collected and analyzed to answer questions, test hypotheses, or disprove theories.

Here is the analysis that has been done to obtain useful information from the huge dataset of USA COVID-19.

**Following is the tableau visualization of our dataset**

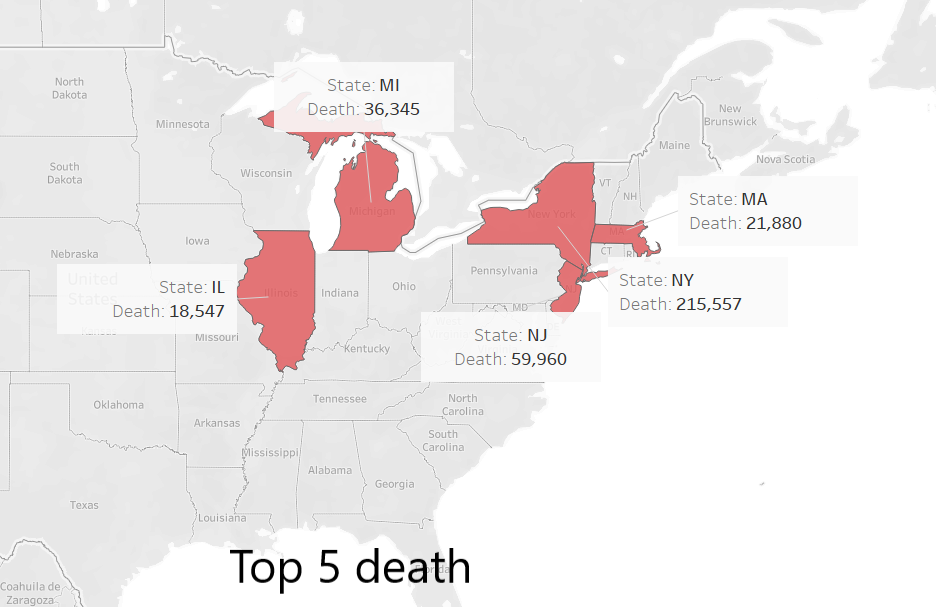
Tableau makes it faster and easier to identify patterns and build practical models by integrating R. This ultimate combination of R with Tableau amplifies data with visual analytics. [Tableau’s visual analytics](https://www.softwebsolutions.com/resources/tableau-an-excellent-tool-to-visualize-data-for-better-insights.html) interface makes analysis of data and interacting with them virtually effortless.

Here we have tried obtaining insights from the data set about the top 5 states that have the highest positive cases of COVID-19 using Tableau. From the below graph we can say that New York is leading in terms of positive cases followed by New Jersey, Massachusetts, California, and Illinois.



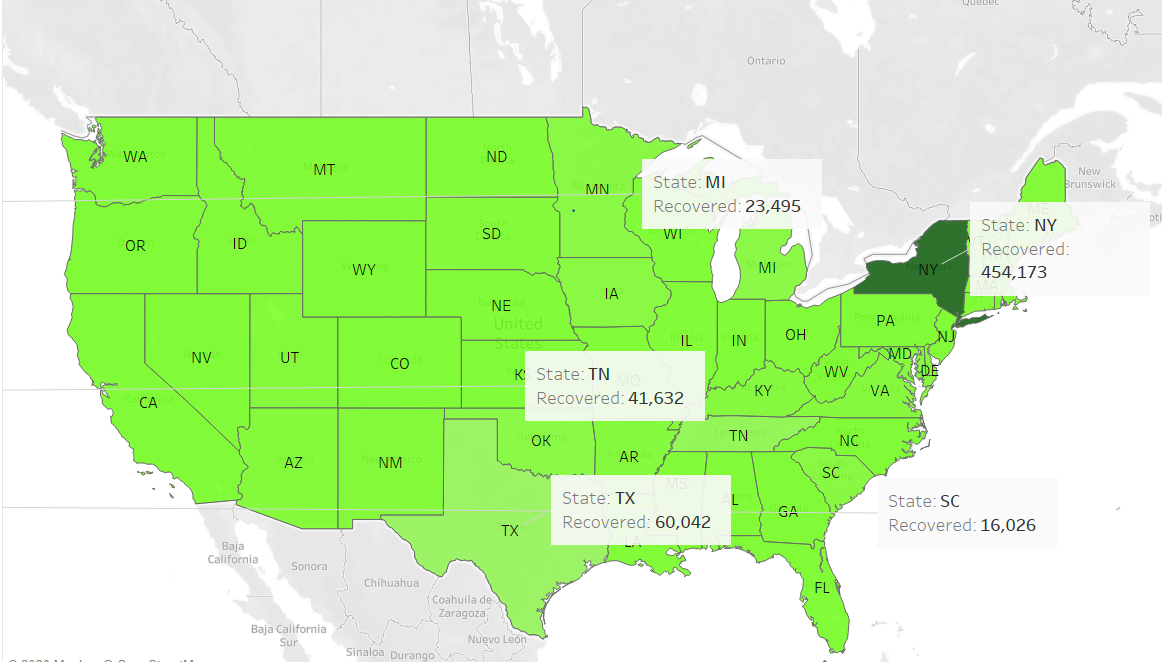
**Figure:-** Top 5 states detected with positive cases

Here we have tried obtaining insights from the data set about the top 5 states that have the highest deaths due to COVID-19 using Tableau. From the below graph we can say that New York is leading in terms of positive cases followed by New Jersey, Massachusetts, Michigan, and Illinois



**Figure:-** Top 5 states that have a higher ratio of deaths

From the below graph we can say that New York has the highest recovery cases followed by Texas, Tennessee, Michigan, and South Carolina.



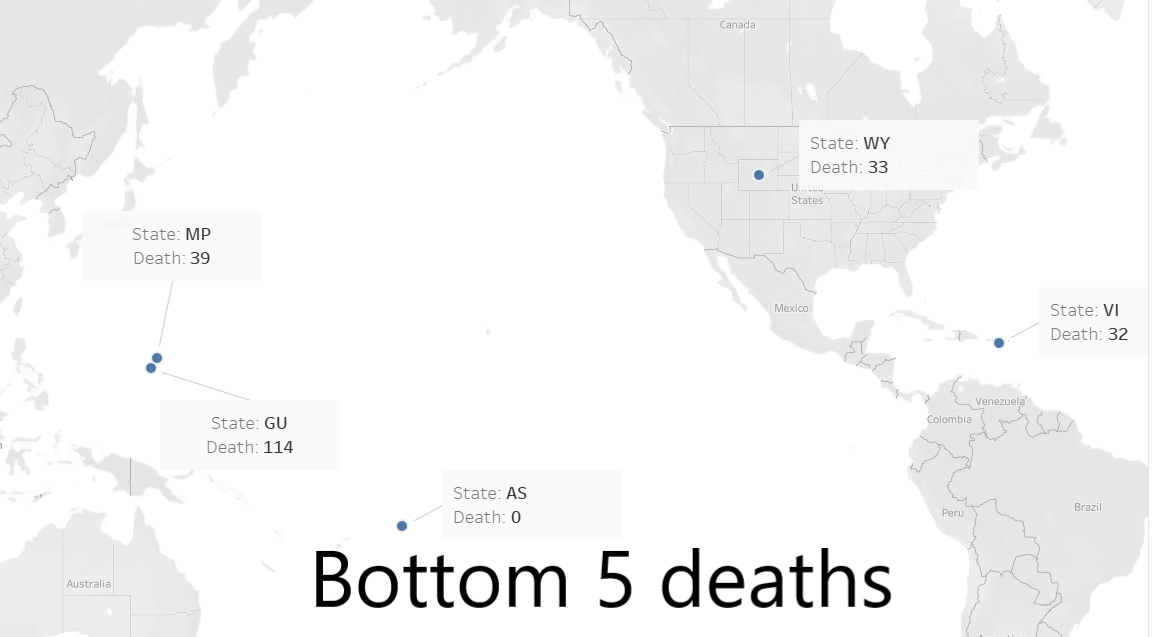
**Figure:-** Top 5 states that have the highest recovery cases

The least count of positive cases in the USA is seen at American Samoa (AS)with zero cases followed Northern Marianas(MP), Guam(GU), Virgin Islands(VI), and Wyoming(WY).



**Figure:-**  Bottom 5 states that have Positive cases

The least death cases are seen at American Samoa (AS)with zero cases followed Northern Marianas(MP), Guam(GU), Virgin Islands(VI), and Wyoming(WY).



**Figure:**- Bottom 5 states that have deaths

**List of states that has higher possibility to recover faster**

We get the list of states that has a higher possibility to recover from this pandemic compared to other states.

For acquiring this result we have taken the ratio of recovered cases to that of positive cases, so when the recovery cases increase and the positive cases decrease that state will have fewer COVID-19 cases and will be the first state to come out of this pandemic.



Below graph shows the graphical representation of the same result. And from the result we can see that Vermont(VT) has the highest possibility to get out of this pandemic and is safe compared to other states.

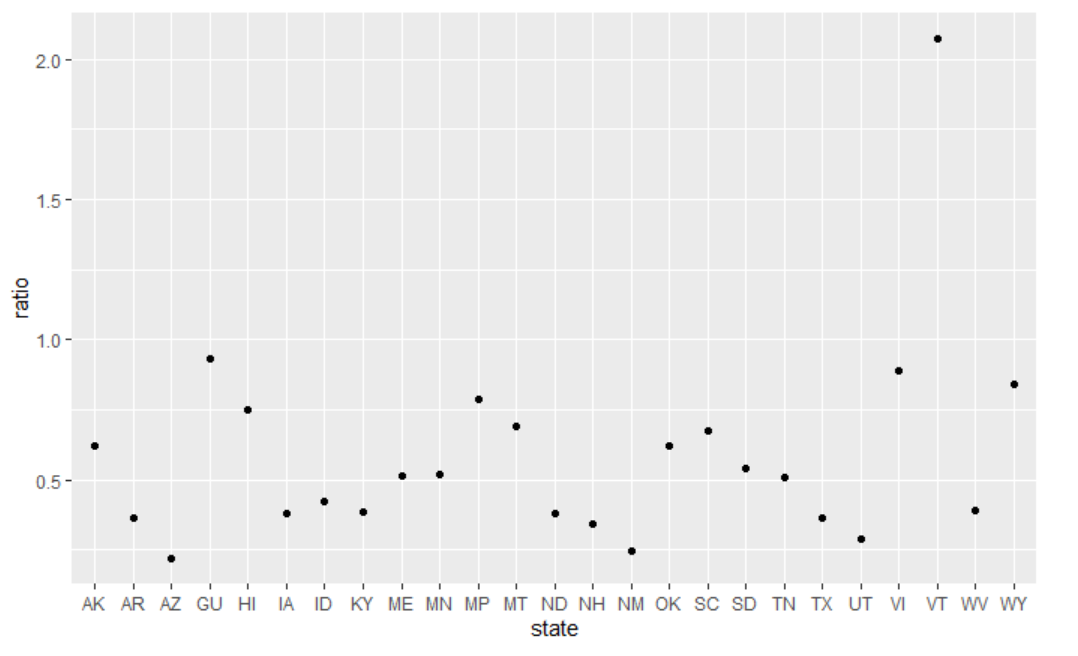
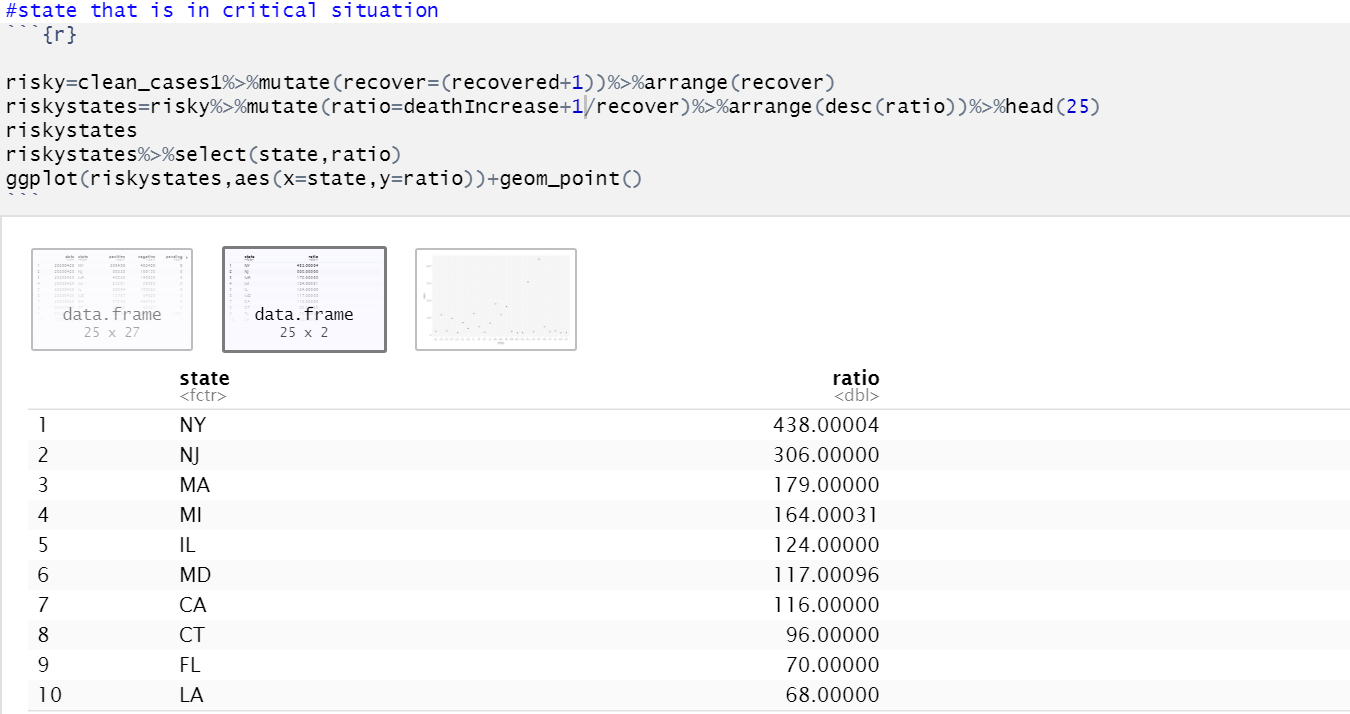


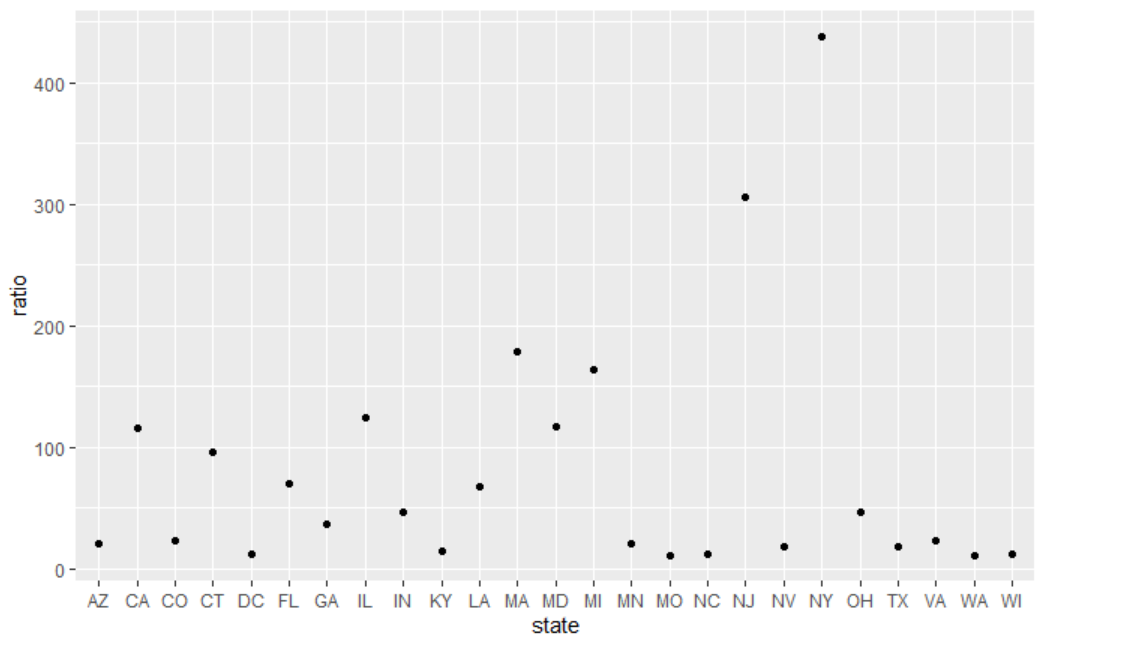
Figure: Graphical representation of higher possibility to recover faster

**List of states that are in a critical situation**

Here, we have derived the list of states that are in critical situations and will be the last to come out from this pandemic. For acquiring this result we have taken the ratio of Death Increase to that of recovery cases so when the death rate is increasing and recovery cases are decreasing that will give the result of states that have the highest death compared to recovery cases and that state is in a critical situation.

****

The below graph shows the graphical representation of the same result. And from the result, we can see that New York(NY) is in a critical situation and New Jersey(NJ) is in the second-highest position.

****

**Figure:-** Graphical representation of states that are in a critical situation

**Final observation and future recommendations:**

**We have driven the data using tools like a tableau. After the analysis using these tools, we came to a few observations:-**

* New York is leading in terms of positive cases followed by New Jersey, Massachusetts, California, and Illinois.
* New York is in critical situations and will be the last to come out from this pandemic. We used the ratio of death increase to that of recovery cases.
* We have observed that New York data might show that it can recover fast but still New York is in a critical situation because of the death ratio.

**After reading these reports our team came up to the recommendation that we can take to stop this by taking a few steps like:-**

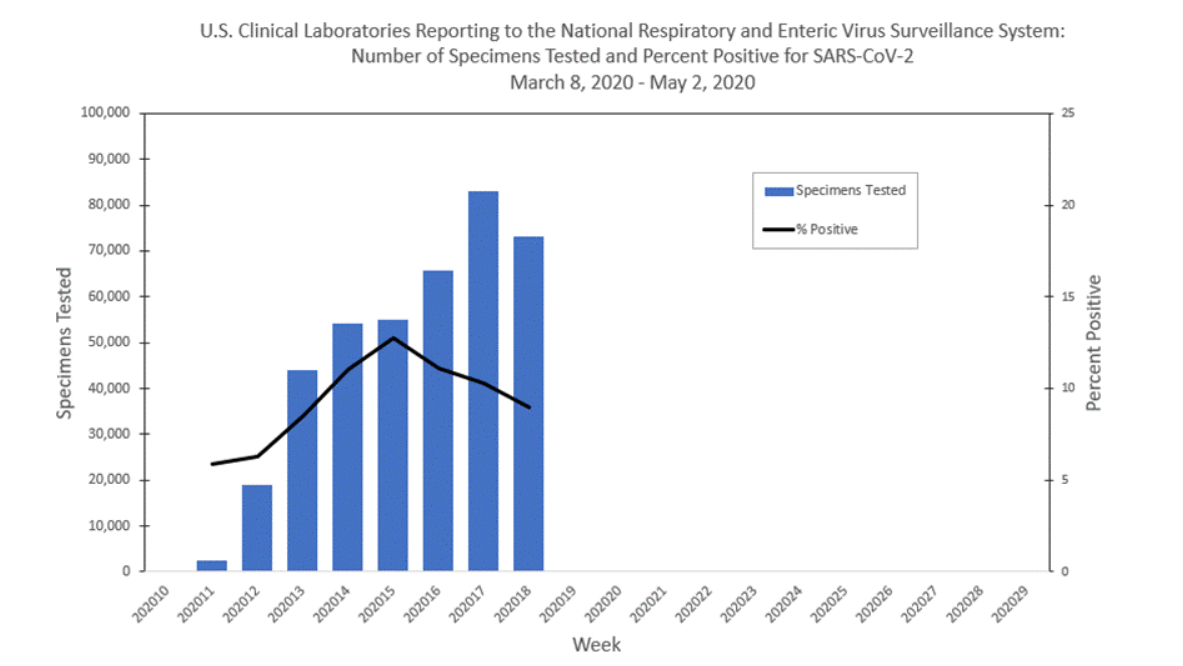
* Keep at least 6 feet between yourself and others.
* Wash your hands with soap and water often.
* Cover your nose and mouth with a tissue or sleeve when sneezing or coughing.
* Do not touch your face with unwashed hands.
* Monitor your health more closely than usual for cold or flu symptoms.

**Future recommendation for the government:-**

* Participatory disaster response strategies, including working with civil society and citizens.
* Citizen-led community responses, including neighborhood volunteer groups and neighborhood associations, clergy, teachers, or others helping to inform the public on the risks and needed steps.
* Building trust between government and citizens, including through strong communications and focusing on reaching vulnerable communities with the information they need.
* Transparency over forecasting models and data that are influencing the government’s strategies.
* Digital platforms or apps to keep citizens informed, enable public participation, and/or offer open data; Digital tools to enable public participation.
* Digital and/or crowdsourced provision of public and government services.
* Protecting data rights and privacy as corporations help lead the response in many countries.
* Tackling misinformation and disinformation online.

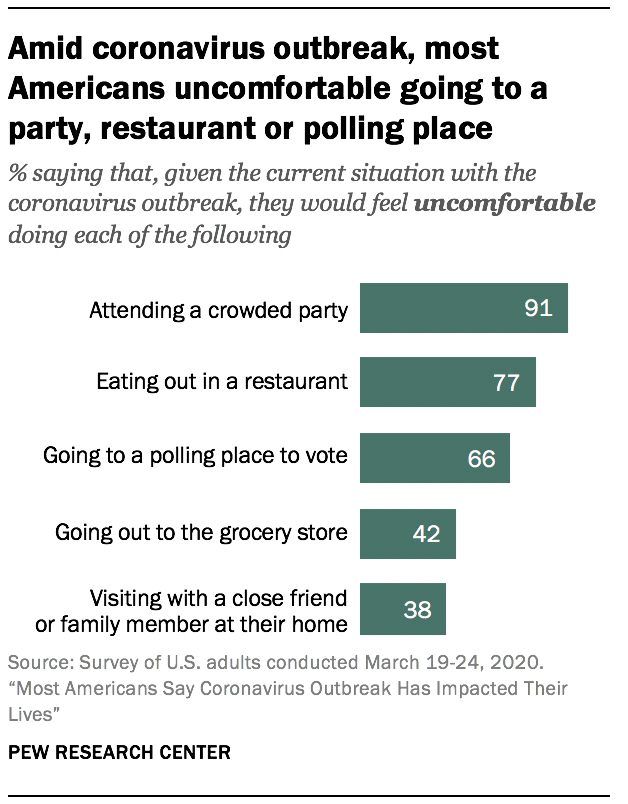
**Artifacts :-**

**About COVID-19:-**

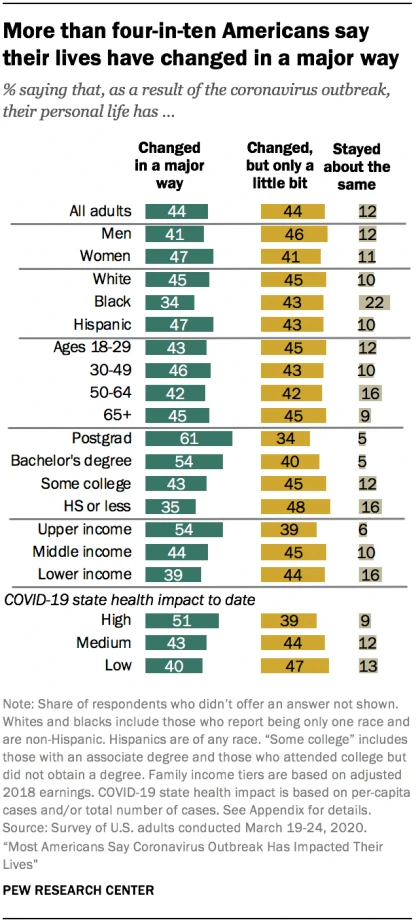
****

This graph records the covid data till 9th May. The above graph displays the testing kits during the period and the amount of positive cases recorded. As per the graph, the positive test cases were increasing steadily during the first ten week and then the rate was almost doubled after the twelfth week and then covid reached its peak during the fifteenth week of 2020 and since then there has been a downfall in the positive cases.

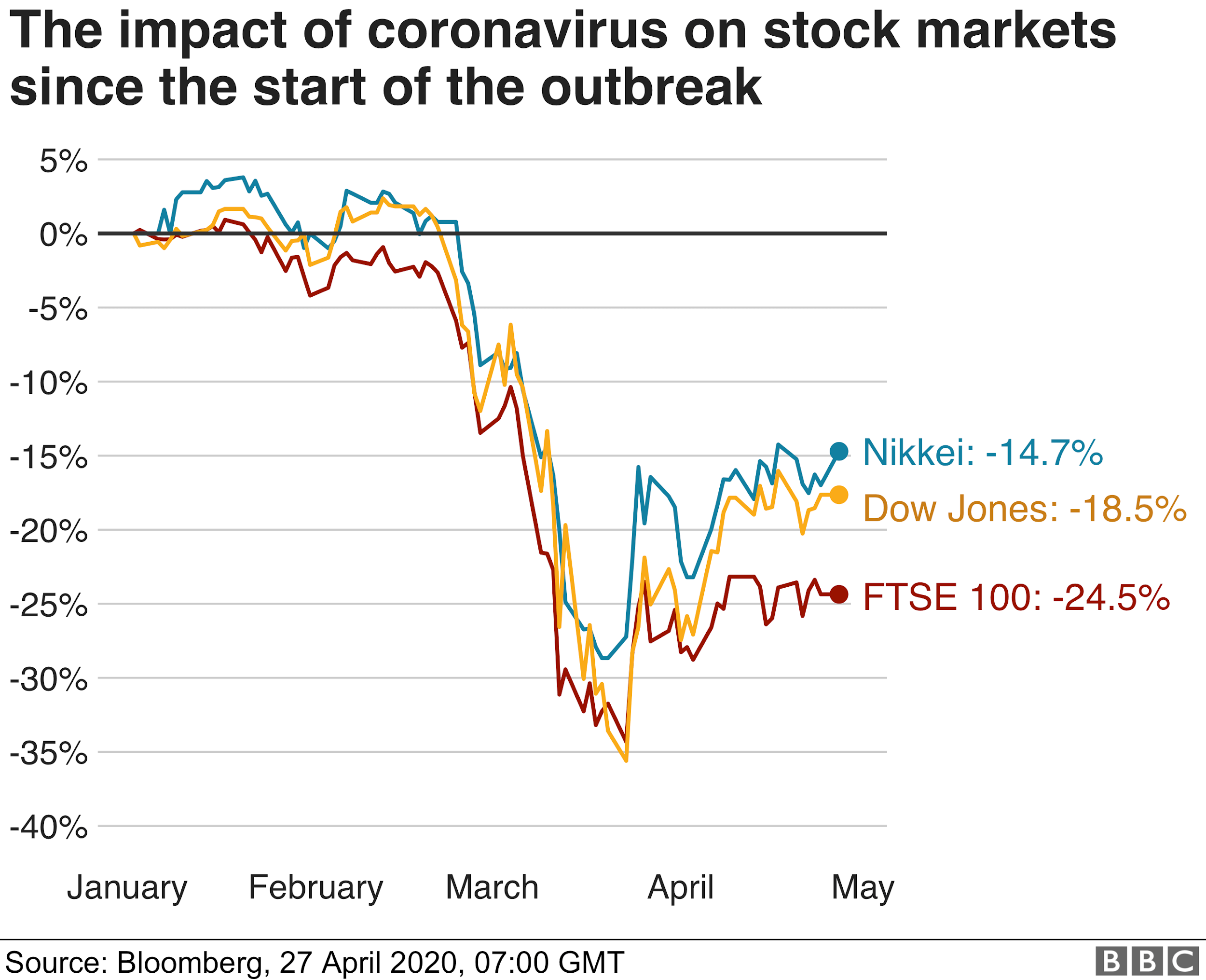
Many infected people can have mild symptoms and hence, it is possible for a person to catch coronavirus as from those people. Apart from that, there have been recordings of some people with no symptoms transmitting the virus. The WHO is currently doing their research on these absurd reports.

This graph might be because of scarcity of the testing kits in the United States. Tests completed by California are 553,409 tests (about 139 per 10,000 individuals), by New York are 826,095 tests (about 425 per 10,000). South Korea learned their lesson from the SARS and they were already prepared, as soon as the outbreak of COVID was confirmed, they started testing potential infected people and they were able to test upto 20000 people per day and by mid-March more than 270,000 people had been tested. Due to which they were able to identify the infected and strict measures were taken to seperate them. They were able to get a downward curve and get their country up and running in a short period of time. 

According to a survey,out of 100 americans 91 are scared to attend a party due to the coronavirus and 77 people are scared to eat out in a restaurant, 66 people are scared to vote,42 are scared to shop their weekly grocery and 38 people are scared to visit a close friend or family member at their home.

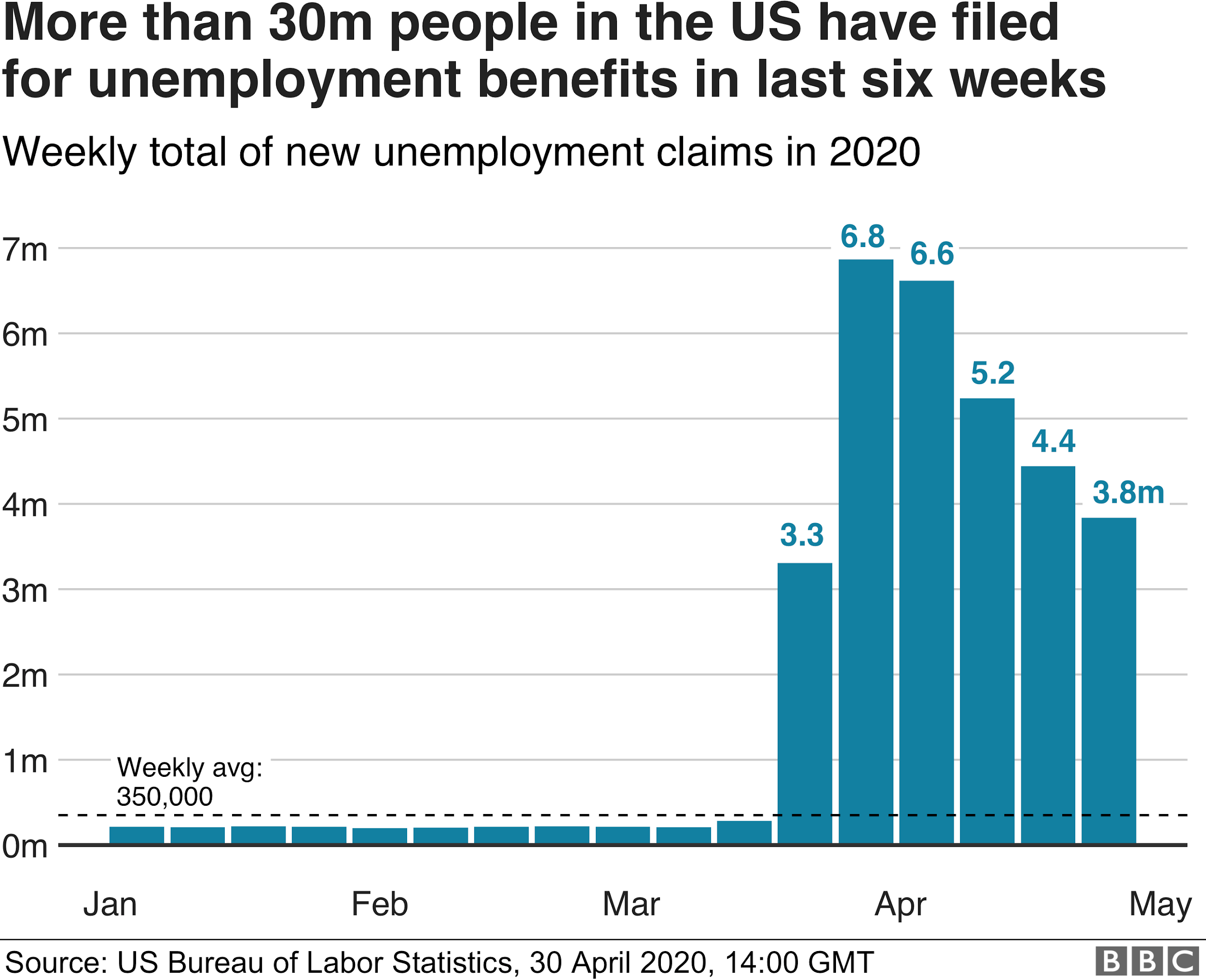


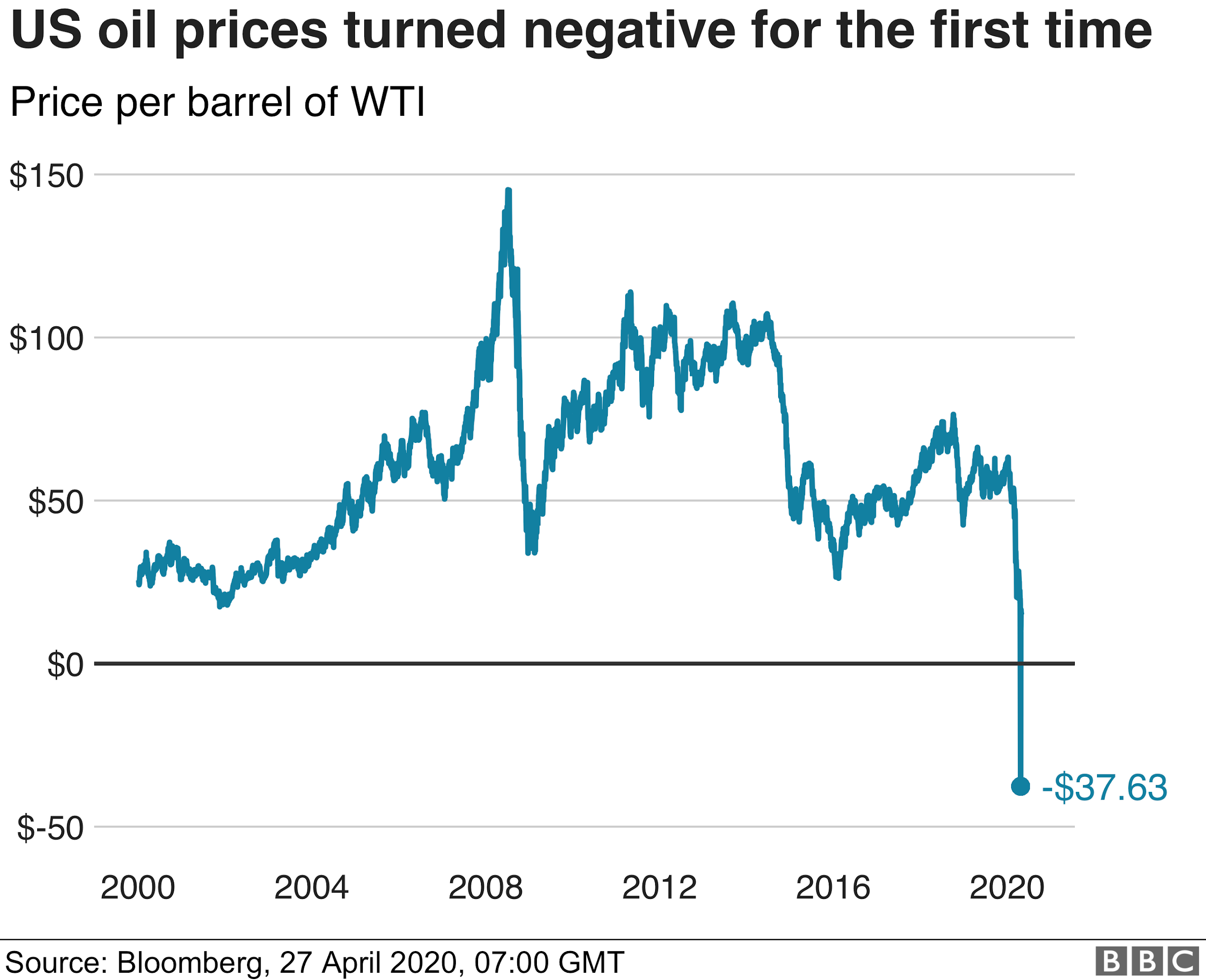
The figure mentioned above basically explained the behavioral changes in various age groups, students, based on their incomes, all the adults, based on their gender, based on their races and effect on health. It basically shows that out of 100 people, the amount of people who changed in a major way, minor way and the people who neglected the spread of corona and stayed the same.



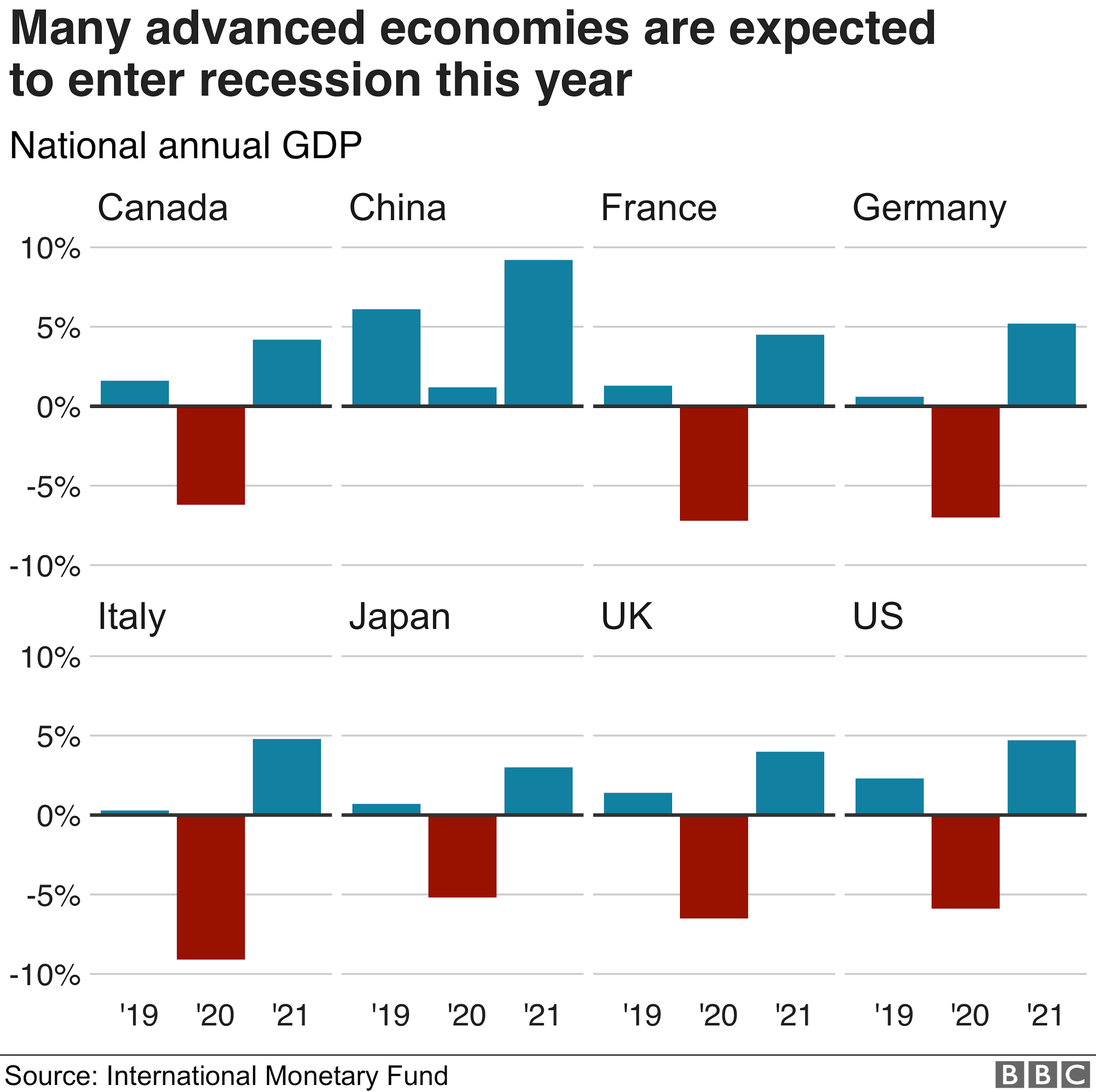
The coronavirus outbreak and the effect of it on the stock markets was observed in January. As per the image, the stock was around 0% and 5% during the January and then the prices dropped below 0 in february and increased a bit then they were plummeted during the mid-March and during the month of april the prices increased by almost 10-15% and they have been increasing since then but at a very low rate.

The figure mentioned below, displays the unemployment rate after the outbreak of coronavirus based on weeks. As per the graph, the unemployment was around 350k per week from january to mid-march and during the period of mid-march to may the unemployment rate has increased significantly, the final week of march records 3.3 million unemployment claims which was 3 million more then it’s previous week, the first week of april records 6.8 million highest till date, then the claims started decreasing and last week the United States had 3.8 million unemployed people.

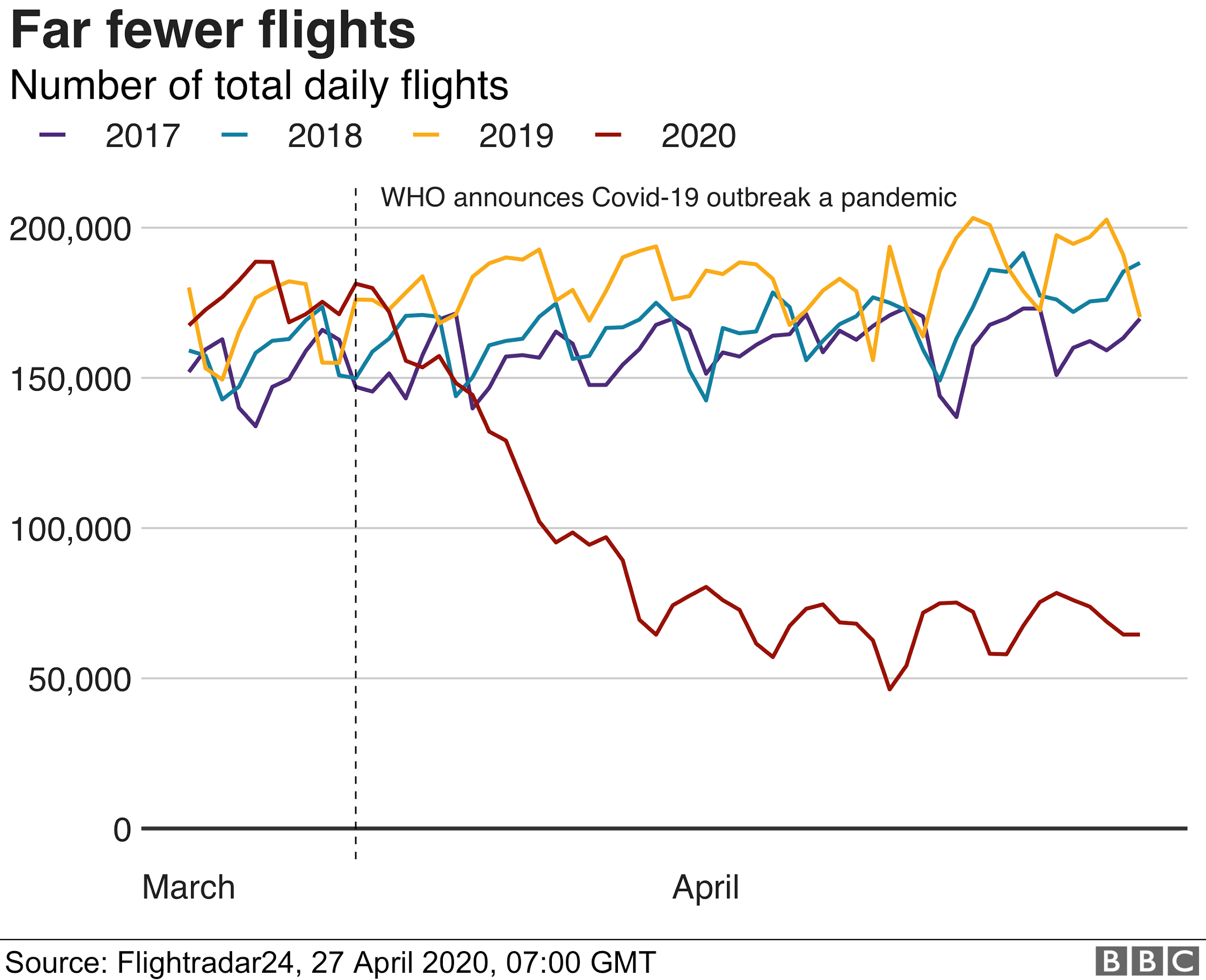




For the first time in history the US oil price per barrel plummeted below $0, as per the figure it is currently at the lowest -$37.63.

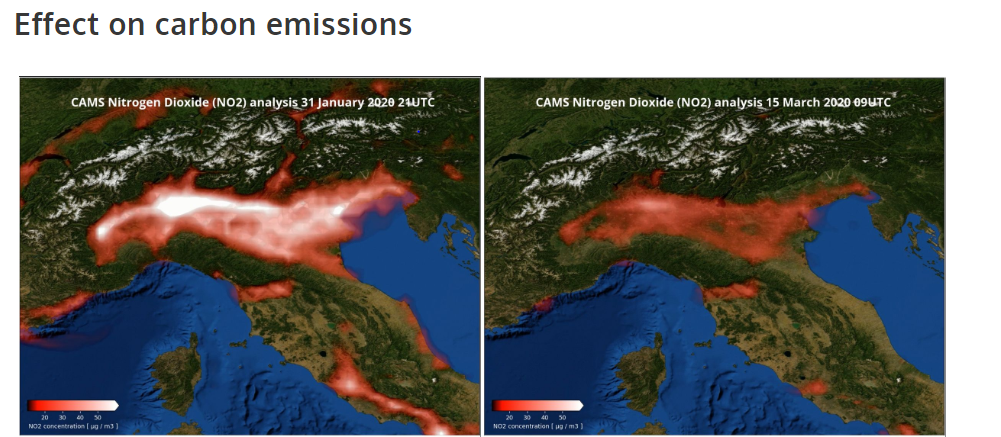


The national annual GDP is mentioned in the figure above. As per the image, every nation that has taken a serious hit from the coronavirus have fallen below -5% and the rest of them have been able to keep their economy around -5% except China who has managed to keep their economy well above 0% although they were the one who took the first hit from the coronavirus. Additionally the prediction has been made for the year of 2021, in which the United states is expected to grow by 10%.



The figure above compares the number of flights that were working during the past 3 years and in the current year. It can be clearly observed that the number of flights that were working during the year of 2018 was increased then the year 2017. Secondly, the graph improved further during the year of 2019. But since due to the lockdown, the year of 2020 has given a serious hit to the airline industries.

The Effect of Coronavirus on the environment, as per the images below In New York, peak congestion went down 47 percent from the 2019 average on the morning of March 23. Los Angeles experienced a 51 percent drop, according to [Fox News](https://www.foxnews.com/auto/traffic-congestion-coronavirus) and the [TomTom Traffic Index](https://www.tomtom.com/en_gb/traffic-index/).



The Carbon dioxide emission due to the airline industry was over 900 million tons, recorded in 2018 and it was projected to be around 2700 million tons by 2050. But due to the lockdown in various countries, the airline industry has observed a significant decrease in their passengers across the world and due to the decrease in demand, few airplanes were in use, due to which during the first three months of 2020 has shown a significant decrease of 40% in the carbon dioxide emission. The environment started healing it’s ozone layer and the skies started to become more clearer.

The 2008 financial crisis led to a 1% dip in the emission of harmful gases, but as the world recovered from the crisis the emissions crept back up with a much faster rate. Similarly, it is expected for the emission of gases to increase and to solve the environmental problem various leaders have suggested the improvement in the technology to be the only solution. Although the recovery in the environment due to the coronavirus will play an important role in slowing the effect of global warming and accelerate the healing of the environment.

