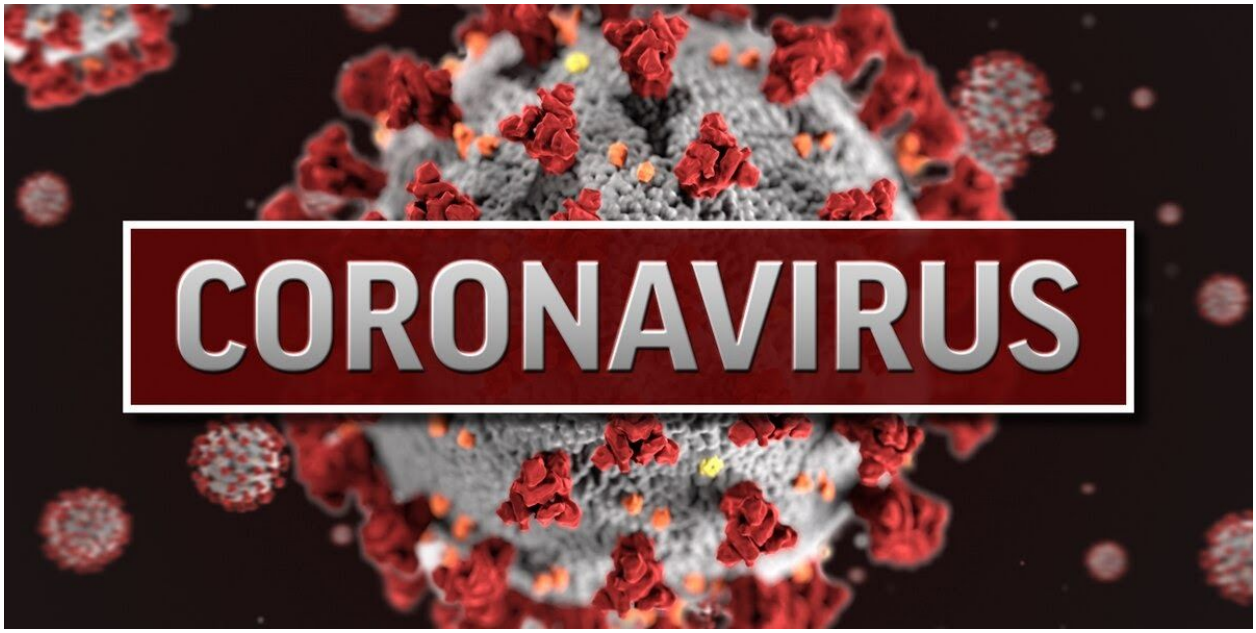
The COVID-19 Project



Group 5

05.14.2019

Project Head: Prof. James Curry

Team Members:

Deepen Patel, Gaurav Kondapuram, Ankit Singh, Smit Ajmera, Akhil Podila, Sweni Thakkar

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)

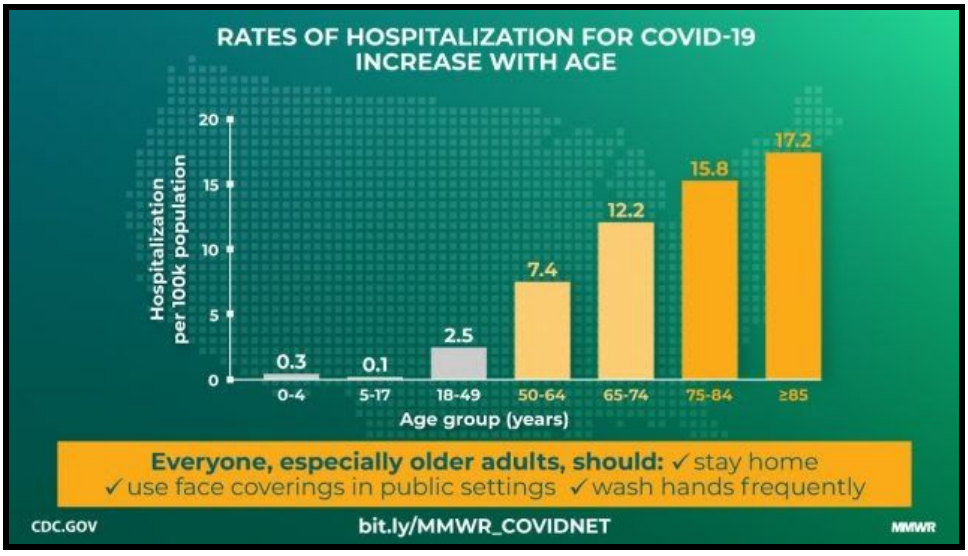
❑ White House Coronavirus Task Force:

The **White House Coronavirus Task Force** is a United States Department of State task force that "coordinates and oversees the Administration's efforts to monitor, prevent, contain, and mitigate the spread" of the coronavirus disease (COVID-19).

Dr. Anthony Fauci the "face of the federal government's response". **Anthony Stephen Fauci** is an American physician and immunologist who has served as the director of the 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 White House Coronavirus Task Force addressing the 2019–20 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 of the United States. Its main goal is to protect public health and safety



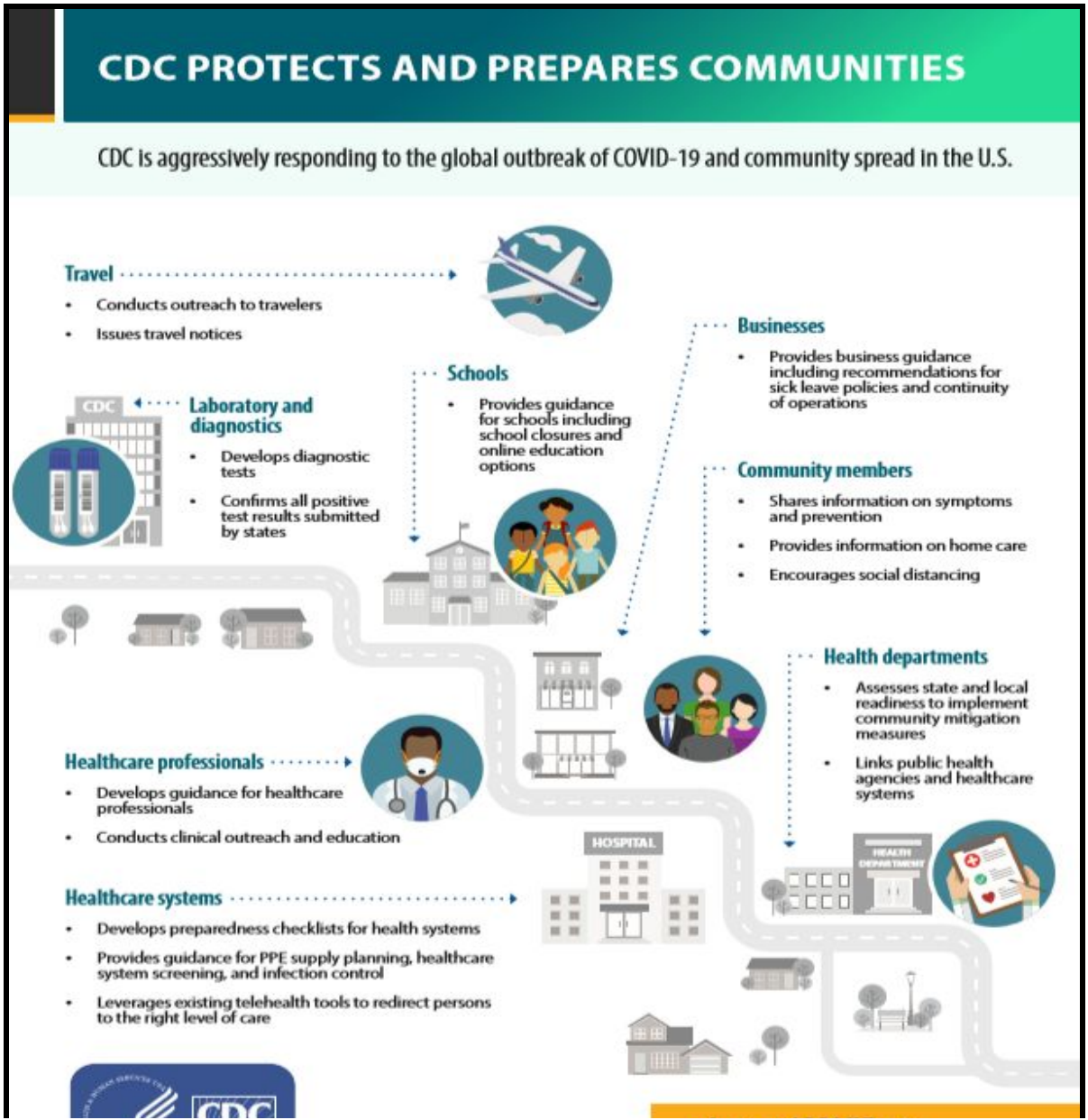
through the control and prevention of disease, injury, and disability in the US and internationally.

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 :

```
```{r}
v2<-all_states_daily%>%select(date,state,positive,negative,death,totalTestResults,hospitalizedCurrently)
str(v2)
```

tibble [1,541 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ date   : num [1:1541] 20200402 20200402 20200402 20200402 20200402 ...
 $ state   : chr [1:1541] "AK" "AL" "AR" "AS" ...
 $ positive   : num [1:1541] 143 1233 643 0 1598 ...
 $ negative   : num [1:1541] 4879 7503 7880 20 21111 ...
 $ death   : num [1:1541] 3 32 12 0 32 203 80 112 12 12 ...
 $ totalTestResults   : num [1:1541] 5022 8736 8523 20 22709 ...
 $ hospitalizedCurrently: num [1:1541] NA NA 66 NA NA ...
 - attr(*, "spec")=
  .. cols(
```

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.

```r
v2<-all_states_daily%>%select(date,state,positive,negative,death,totalTestResults,hospitalizedCurrently)
str(v2)
head(v2)
```

date <dbl>	state <chr>	positive <dbl>	negative <dbl>	death <dbl>	totalTestResults <dbl>	hospitalizedCurrently <dbl>
20200402	AK	143	4879	3	5022	NA
20200402	AL	1233	7503	32	8736	NA
20200402	AR	643	7880	12	8523	66
20200402	AS	0	20	0	20	NA
20200402	AZ	1598	21111	32	22709	NA
20200402	CA	9191	23809	203	33000	1922

6 rows

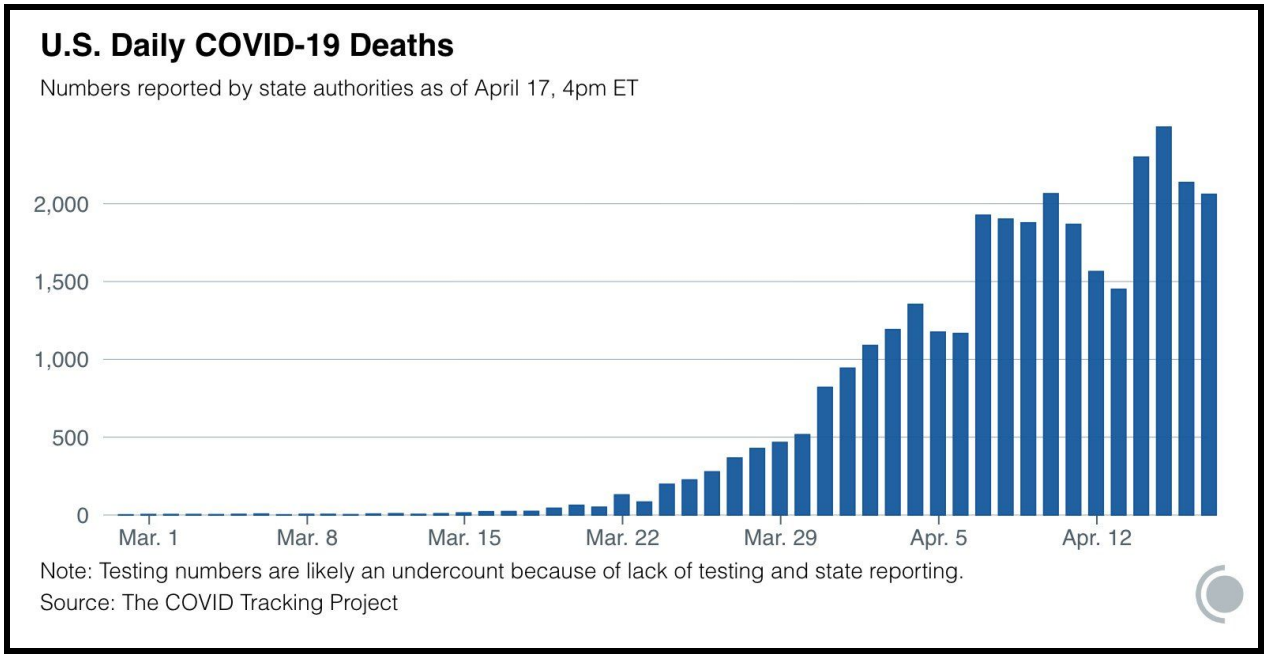
Limitations - Not all the states consistently report their test results and regularly. In such cases, they use other reporting tools like directly asking state officials, watching news conferences, gleaning information from trusted news sources. Moreover, since the symptoms are not visible until 14 days, 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.

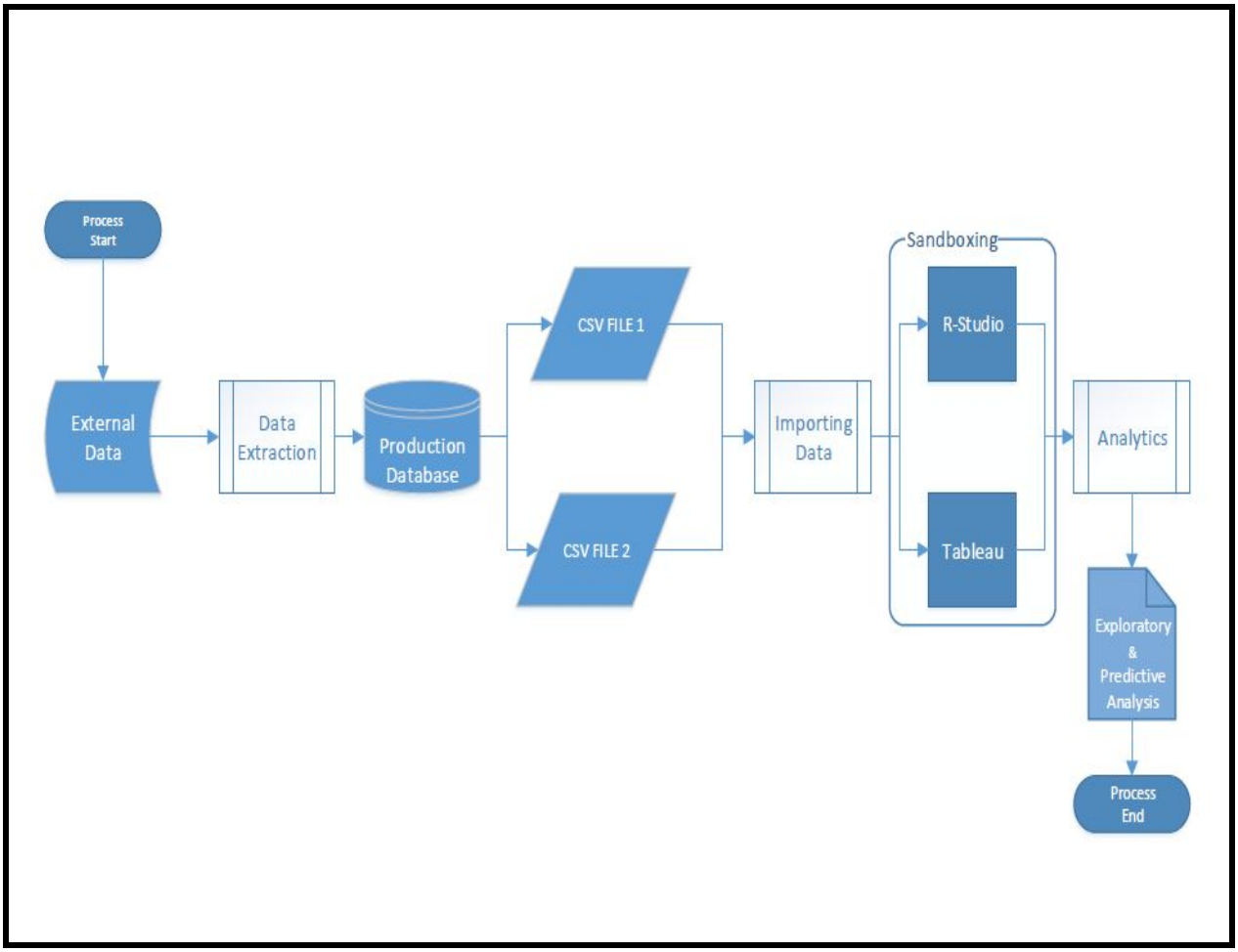
Tableau is an interactive, self-service reporting and analytics tool that enables faculty and staff to integrate and combine data from multiple sources into visualizations and be accessed in a single desktop environment using Tableau Desktop or through a shared dashboard.

- **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.

```
## R Markdown

```{r}
require(tidyverse)
cases=read.csv("C:/Users/Smit Ajmera/Desktop/daily.csv")
states=read.csv("C:/Users/Smit Ajmera/Desktop/info.csv")
cases
states
```
```

data.frame 2769 x 25 | data.frame 56 x 10

date <int>	state <fctr>	positive <int>	negative <int>	pending <int>	hospitalizedCurrently <int>	hospitalizedCumulative <int>
20200423	AK	337	11824	NA	42	NA
20200423	AL	5778	46863	NA	NA	768
20200423	AR	2465	29125	NA	101	291
20200423	AS	0	3	17	NA	NA
20200423	AZ	5769	52928	NA	699	NA
20200423	CA	37369	444728	NA	4929	NA
20200423	CO	10878	39767	NA	859	2123
20200423	CT	23100	48397	NA	1947	NA
20200423	DC	3361	12569	NA	402	NA
20200423	DE	3308	13604	NA	290	NA

1-10 of 2,769 rows | 1-7 of 25 columns Previous 1 2 3 4 5 6 ... 100 Next

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.

```
#Manipulating the data
```{r}
case=cases%>%select(-hash,-total,-inIcuCumulative,-onVentilatorCumulative,-dateChecked,-posNeg)
case
```
```

date <int>	state <fctr>	positive <dbl>	negative <dbl>	pending <dbl>	hospitalizedCurrently <dbl>	hospitalizedCumulative <dbl>
20200423	AK	337	11824	0	42	0
20200423	AL	5778	46863	0	0	768
20200423	AR	2465	29125	0	101	291
20200423	AS	0	3	17	0	0
20200423	AZ	5769	52928	0	699	0
20200423	CA	37369	444728	0	4929	0
20200423	CO	10878	39767	0	859	2123
20200423	CT	23100	48397	0	1947	0
20200423	DC	3361	12569	0	402	0
20200423	DE	3308	13604	0	290	0

1-10 of 2,769 rows | 1-7 of 19 columns Previous 1 2 3 4 5 6 ... 100 Next

❑ 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.

```
## Checking and handling NA values
```{r}
is.na(cases)
cases[is.na(cases)]=0
cases
```
```

date <int>	state <fctr>	positive <dbl>	negative <dbl>	pending <dbl>	hospitalizedCurrently <dbl>	hospitalizedCumulative <dbl>
20200423	AK	337	11824	0	42	0
20200423	AL	5778	46863	0	0	768
20200423	AR	2465	29125	0	101	291
20200423	AS	0	3	17	0	0
20200423	AZ	5769	52928	0	699	0
20200423	CA	37369	444728	0	4929	0
20200423	CO	10878	39767	0	859	2123
20200423	CT	23100	48397	0	1947	0
20200423	DC	3361	12569	0	402	0
20200423	DE	3308	13604	0	290	0

1-10 of 2,769 rows | 1-7 of 25 columns Previous 1 2 3 4 5 6 ... 100 Next

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 5 states in US with highest positive cases and lowest positive cases
```{r}
cases1=cases%>%filter(date=="20200423")
cases1%>%arrange(desc(positive))%>%head(5)   # top 5
cases1%>%arrange(desc(positive))%>%tail(5)   # bottom 5

#Increase in cases in New York
caseny=cases%>%filter(state=="NY")

ggplot(caseny)+geom_line(aes(x=date, y=positiveIncrease))
```
```

	date <int>	state <fctr>	positive <dbl>	negative <dbl>	pending <dbl>	hospitalizedCurrently <dbl>
1	20200423	NY	263460	432460	0	15021
2	20200423	NJ	99989	100159	0	7240
3	20200423	MA	46023	149053	0	3890
4	20200423	CA	37369	444728	0	4929
5	20200423	IL	36934	136382	0	4877

	date <int>	state <fctr>	positive <dbl>	negative <dbl>	pending <dbl>	hospitalizedCurrently <dbl>
52	20200423	WY	326	7241	0	0
53	20200423	GU	135	1180	0	2
54	20200423	VI	54	583	31	0
55	20200423	MP	14	51	0	0
56	20200423	AS	0	3	17	0

Top and bottom 5 states based on Recovery.

```
#Top 5 states with recovered cases
```{r}
cases1%>%select(date,state,positive,negative,recovered)%>%arrange(desc(recovered))%>%head(5)
cases1%>%select(date,state,positive,negative,recovered)%>%arrange(desc(recovered))%>%tail(5)

ggplot(caseny)+geom_line(aes(x=date,y=recovered))
```
```

	date <int>	state <fctr>	positive <dbl>	negative <dbl>	recovered <dbl>
1	20200423	NY	263460	432460	23887
2	20200423	TX	21944	203134	8025
3	20200423	TN	8266	114834	4193
4	20200423	SC	4917	39546	3317
5	20200423	MI	35291	93030	3237

	date <int>	state <fctr>	positive <dbl>	negative <dbl>	recovered <dbl>
52	20200423	OR	2127	41849	0
53	20200423	PA	36647	142061	0
54	20200423	PR	915	9313	0
55	20200423	WA	12494	135459	0
56	20200423	WI	5052	51456	0

Top and bottom 5 states based on deaths.

```
# Top 5 states in US with highest deaths and lowest deaths recorded
```{r}
cases1%>%select(date,state,positive,negative,death)%>%arrange(desc(death))%>%head(5)
cases1%>%select(date,state,positive,negative,death)%>%arrange(desc(death))%>%tail(5)

#It denotes death rate during these period in NY.
caseny=cases%>%filter(state=="NY")
ggplot(caseny)+geom_line(aes(x=date,y=death))
```
```

	date <int>	state <fctr>	positive <dbl>	negative <dbl>	death <dbl>
1	20200423	NY	263460	432460	15740
2	20200423	NJ	99989	100159	5368
3	20200423	MI	35291	93030	2977
4	20200423	MA	46023	149053	2360
5	20200423	IL	36934	136382	1688

	date <int>	state <fctr>	positive <dbl>	negative <dbl>	death <dbl>
52	20200423	WY	326	7241	7
53	20200423	GU	135	1180	5
54	20200423	VI	54	583	3
55	20200423	MP	14	51	2
56	20200423	AS	0	3	0

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.

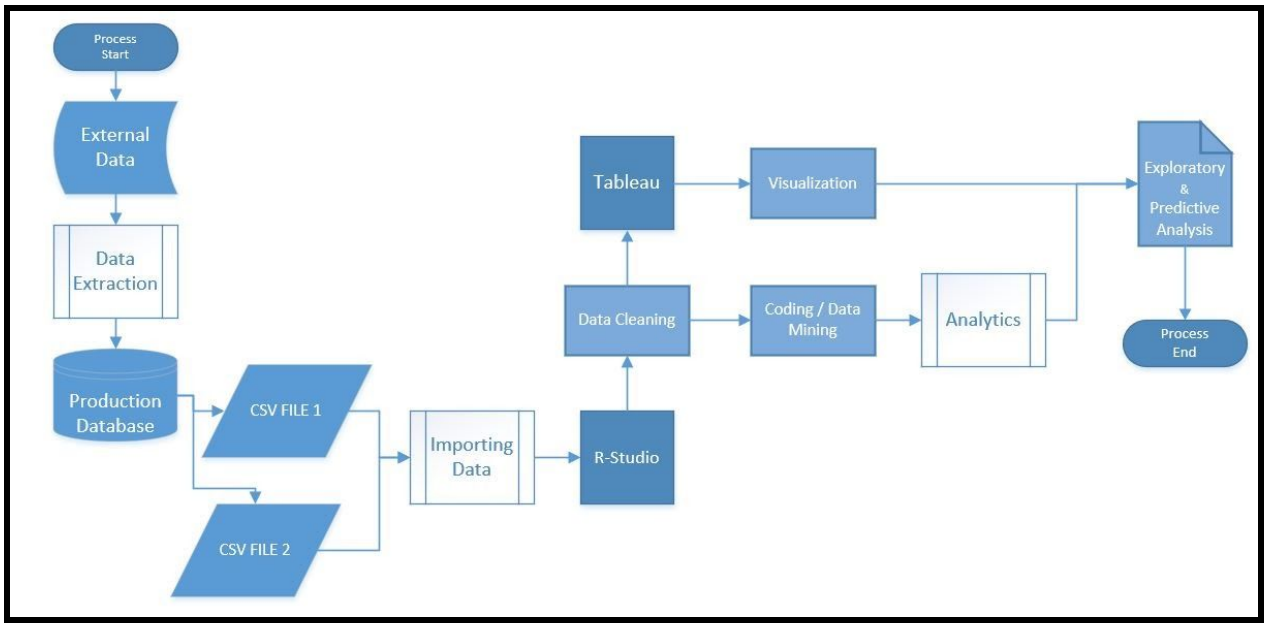
List of States

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

state <chr>	total_positive <dbl>	total_recovered <dbl>	total_death <dbl>
NY	633933	31098	10522
NJ	140344	0	2052
CA	62608	0	1284
MI	58290	0	1674
WA	53442	0	2567
MA	51026	0	615
IL	47637	0	715
FL	47600	0	654
LA	44088	0	1712
PA	36238	0	416
GA	30898	0	924
TX	25693	152	355
CO	22317	0	413
CT	21467	0	465
TN	18356	478	121
OH	17522	0	359
IN	15657	0	383
MD	13018	106	141
WI	12302	0	160
NC	11734	0	54
AZ	10144	0	174
VA	10006	0	228
MO	9619	0	122
SC	8791	0	170
NV	8758	0	163
MS	8030	0	132
AL	7950	0	88
UT	7595	0	34
OR	5930	0	157
MN	5901	774	82
AR	4986	248	56
OK	4666	0	165
DC	4128	729	65
KY	4090	0	105
IA	3910	285	45
RI	3668	105	37
KS	3353	0	76
ID	3027	0	43
ME	2840	407	27
NH	2756	147	22
DE	2596	160	51
VT	2474	45	127
NM	2421	106	23
HI	1714	176	4
PR	1584	0	56
NE	1573	0	18
MT	1503	0	20
WV	1140	0	5
ND	1042	175	13
SD	996	278	18
WY	992	186	0

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.

date	state	positive	negative	pending	hospitalize	hospitalize	inicuCurre	inicuCumu	onVentilat	onVentilat	recovered	hash	dateCheck	death	hospitalize	total	totalTestP	posNeg	fips	deathIncri	hospitalize	negativeIr	positiveIn	totalTestRe
2E+07	AK	337	11824		42						209	59a03ea9	2020-04-2	9		12161	12161	12161	2	0	0	0	2	2
2E+07	AL	5778	46863			768		288		170		78a9b97c	2020-04-2	197	768	52641	52641	52641	1	3	38	3568	313	3881
2E+07	AR	2465	29125		101	291			24	57	902	bd177e23	2020-04-2	45	291	31590	31590	31590	5	3	0	1688	189	1877
2E+07	AS	0	3	17								7c31cc9e	2020-04-23T20:00:00Z			20	3	3	60	0	0	0	0	0
2E+07	AZ	5769	52928		699		305		201		1282	51879cc6	2020-04-2	249		58697	58697	58697	4	20	0	1786	310	2096
2E+07	CA	37369	444728		4929		1531					eccfc506	2020-04-2	1469		482097	482097	482097	6	115	0	14797	1973	16770
2E+07	CO	10878	39767		859	2123						951c9b0f	2020-04-2	508	2123	50645	50645	50645	8	22	120	1510	431	1941
2E+07	CT	23100	48397		1947							59926789	2020-04-2	1639		71497	71497	71497	9	95	0	948	631	1579
2E+07	DC	3361	12569		402		120		200		648	72dcd05b	2020-04-2	139		15930	15930	15930	11	12	0	273	155	428
2E+07	DE	3308	13604		290						643	ac5dc066	2020-04-2	92		16912	16912	16912	10	3	0	251	108	359
2E+07	FL	28832	267876	1301		4693						3ba322cf	2020-04-2	979	4693	298009	296708	296708	12	69	224	7558	523	8081
2E+07	GA	21512	79550			4069						35914be9	2020-04-2	872	4069	101062	101062	101062	13	36	110	6218	772	6990
2E+07	GU	135	1180		2						126	33446da4	2020-04-2	5		1315	1315	1315	66	0	0	94	1	95
2E+07	HI	592	25536			63					444	f9b8c542	2020-04-2	12	63	26128	26128	26128	15	0	7	776	10	786
2E+07	IA	3924	25338		282		102		55		1492	2f57b0c1	2020-04-2	96		29262	29262	29262	19	6	0	842	176	1018
2E+07	ID	1802	16290			162		60			767	b757b69d	2020-04-2	54	162	18092	18092	18092	16	3	4	326	36	362
2E+07	IL	36934	136382		4877		1268		766			c877cdfb	2020-04-2	1688		173316	173316	173316	17	123	0	7144	1826	8970
2E+07	IN	13039	59001				652		333			41aa8a24	2020-04-2	706		72040	72040	72040	18	45	0	1969	601	2570
2E+07	KS	2482	18836			442						8914a9cc	2020-04-2	112	442	21318	21318	21318	20	2	10	844	271	1115
2E+07	KY	3373	32702		301	1105	161	564			1311	e2aacff8a	2020-04-2	185	1105	36075	36075	36075	21	14	29	2566	181	2747
2E+07	LA	25739	117576		1727				274			1ad45a49	2020-04-2	1540		143315	143315	143315	22	67	0	0	481	481
2E+07	MA	46023	149053		3890	4493	1034					bf9b3a5e	2020-04-2	2360	4493	195076	195076	195076	25	178	4493	11535	3079	14614
2E+07	MD	15737	64363		1405	3477	515				1040	99cd5fdf1	2020-04-2	748	3477	80100	80100	80100	24	117	152	2609	962	3571
2E+07	ME	937	16784		42	150	18		11		485	6c869cba	2020-04-2	44	150	17721	17721	17721	23	5	6	0	30	30
2E+07	MI	35291	93030		3611		1148		1027		3237	1e1b7f95	2020-04-2	2977		128321	128321	128321	26	164	0	8771	1325	10096
2E+07	MN	2942	48606		268	712	104	274			1536	77ead2e7	2020-04-2	200	712	51548	51548	51548	27	21	52	1983	221	2204
2E+07	MO	6321	53129		884							48e48812	2020-04-2	218		59450	59450	59450	29	10	0	1110	184	1294
2E+07	MP	14	51								11	f8895b56	2020-04-2	2		65	65	65	69	0	0	0	0	0
2E+07	MS	5153	50236		595	946	156		78			d26dcf04	2020-04-2	201	946	55389	55389	55389	28	8	36	1295	259	1554
2E+07	MT	442	11433		13	59					306	3e192783	2020-04-2	14	59	11875	11875	11875	30	0	0	289	3	292
2E+07	NC	7608	88577		486							bbf9bdc14	2020-04-2	253		96185	96185	96185	37	11	0	5461	388	5849
2E+07	ND	709	15621		18	65					269	7fc1a021	2020-04-2	15	65	16330	16330	16330	38	1	3	711	30	741
2E+07	NE	1813	15547									f2fd230da	2020-04-2	45		17360	17360	17360	31	7	0	590	91	681
2E+07	NH	1588	14424	265	91	213					550	2209e432	2020-04-2	48	213	16277	16012	16012	33	6	7	874	97	971

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.

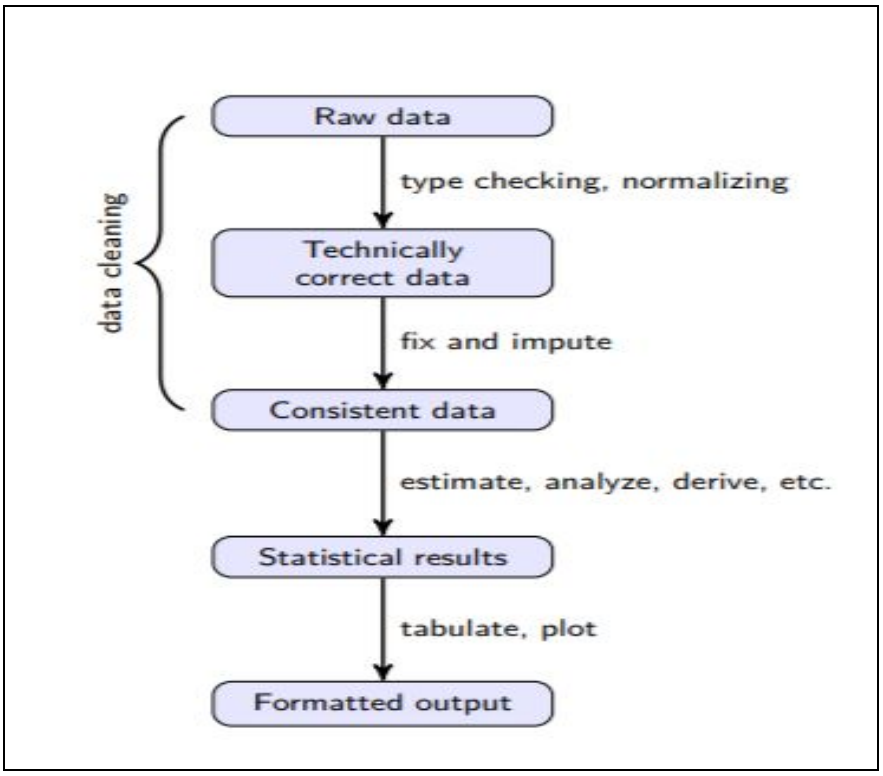
Reduction of files:

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, transforming, and modeling 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 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 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.

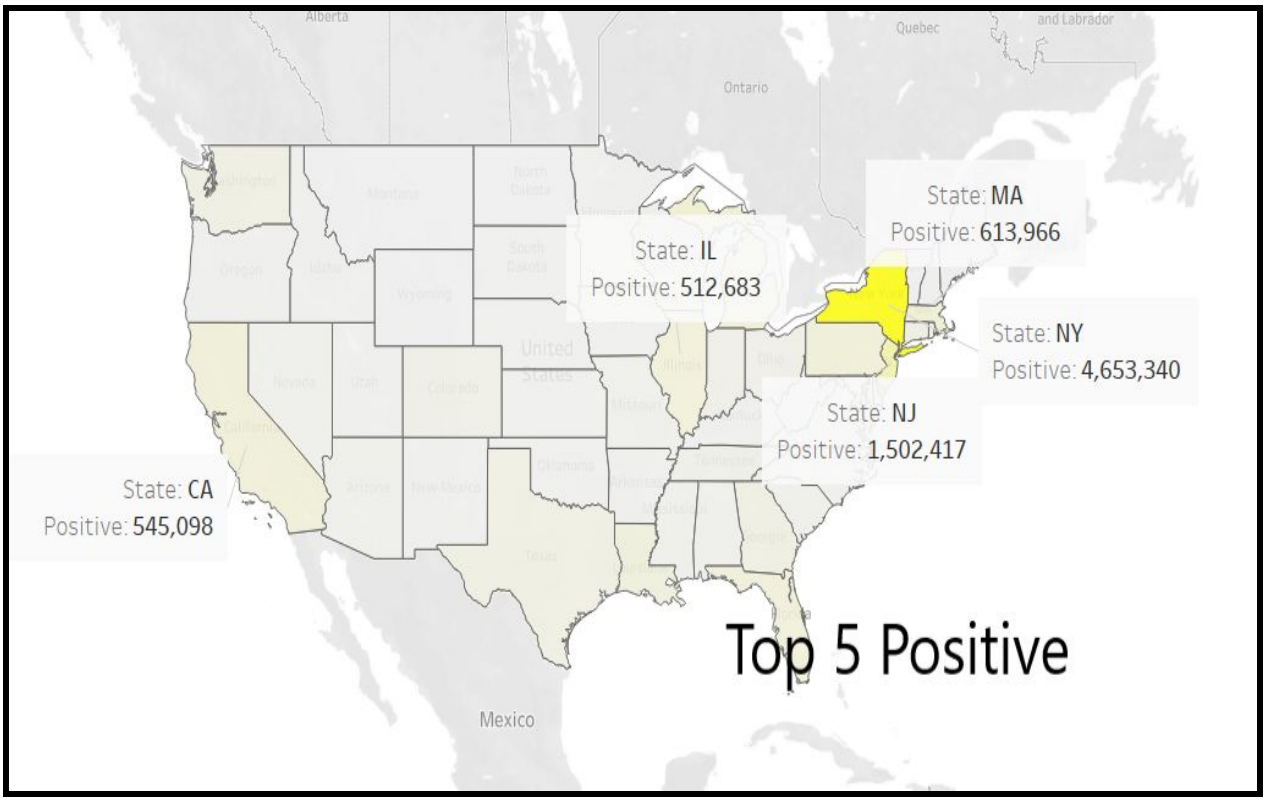


Figure:- Top 5 states detected with positive cases

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

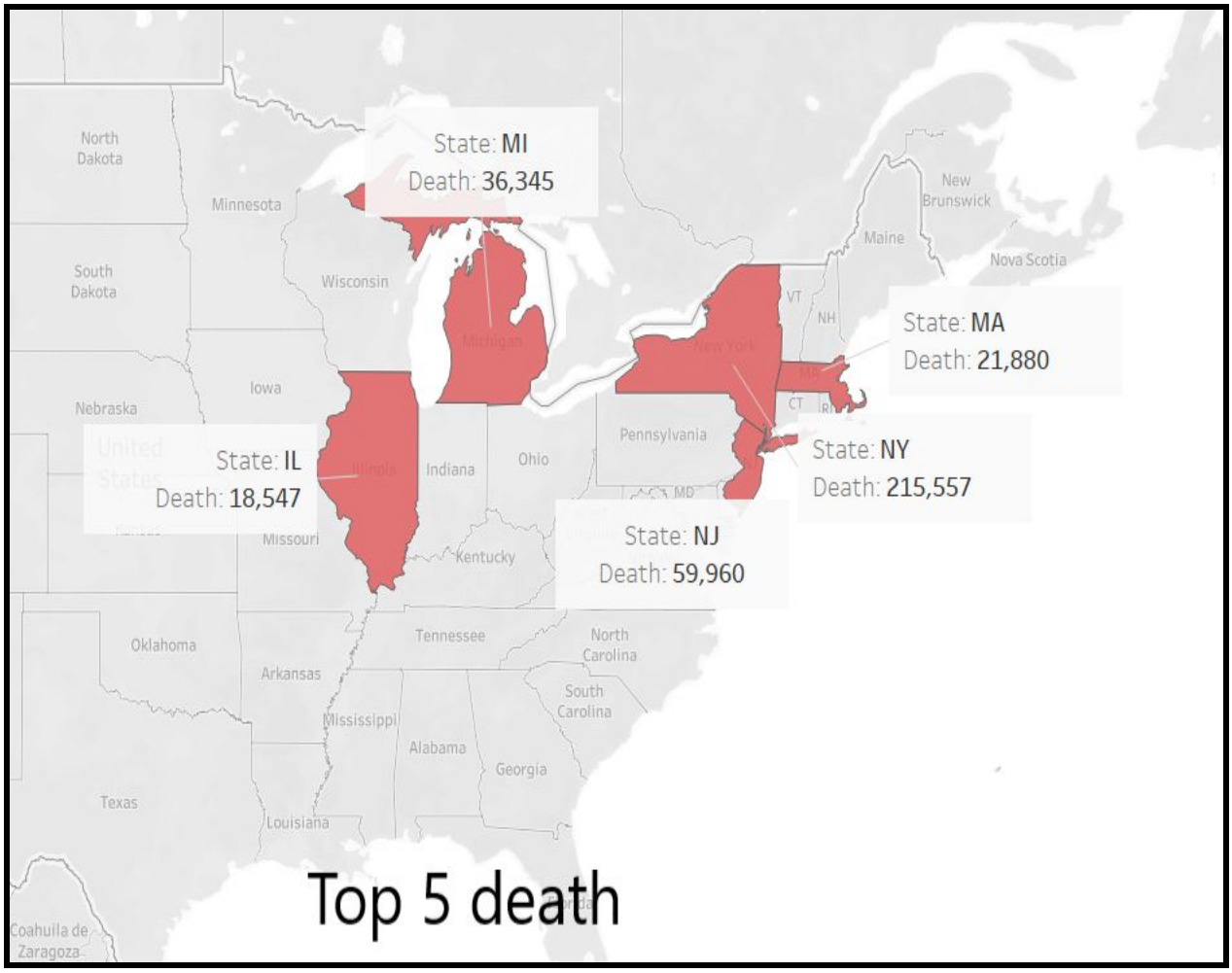


Figure:- Top 5 states that have a higher ratio of deaths

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

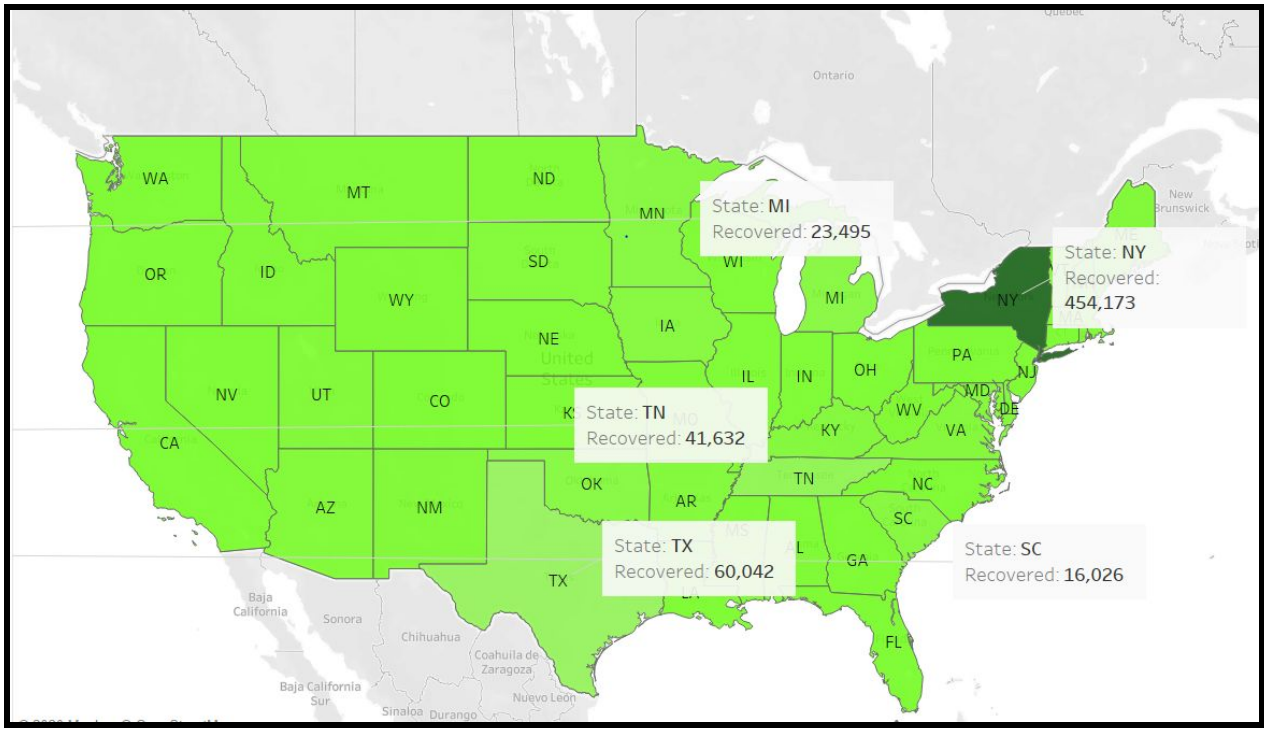


Figure:- Top 5 states that have the highest recovery cases

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).

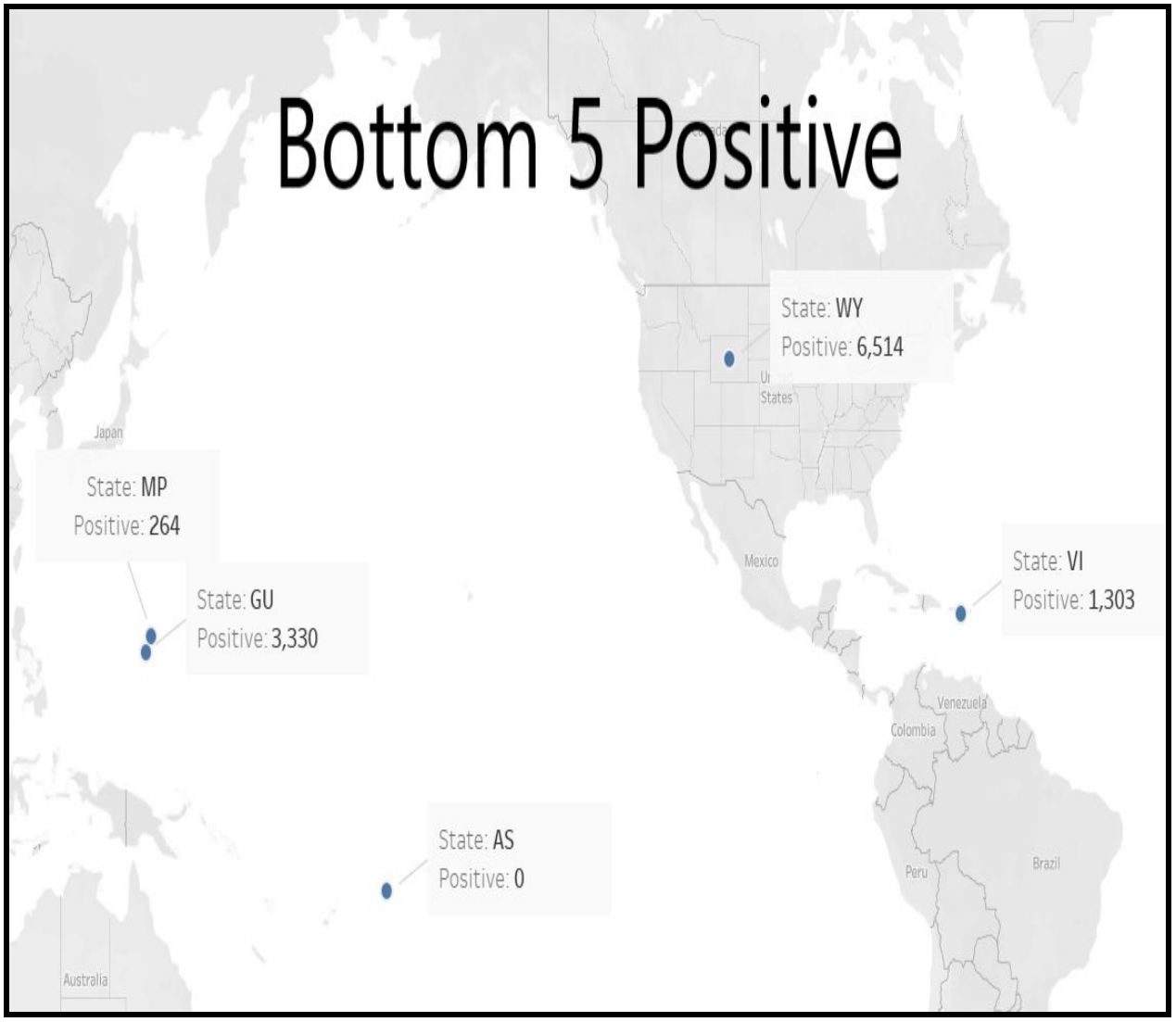


Figure:- Bottom 5 states that have Positive cases

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).

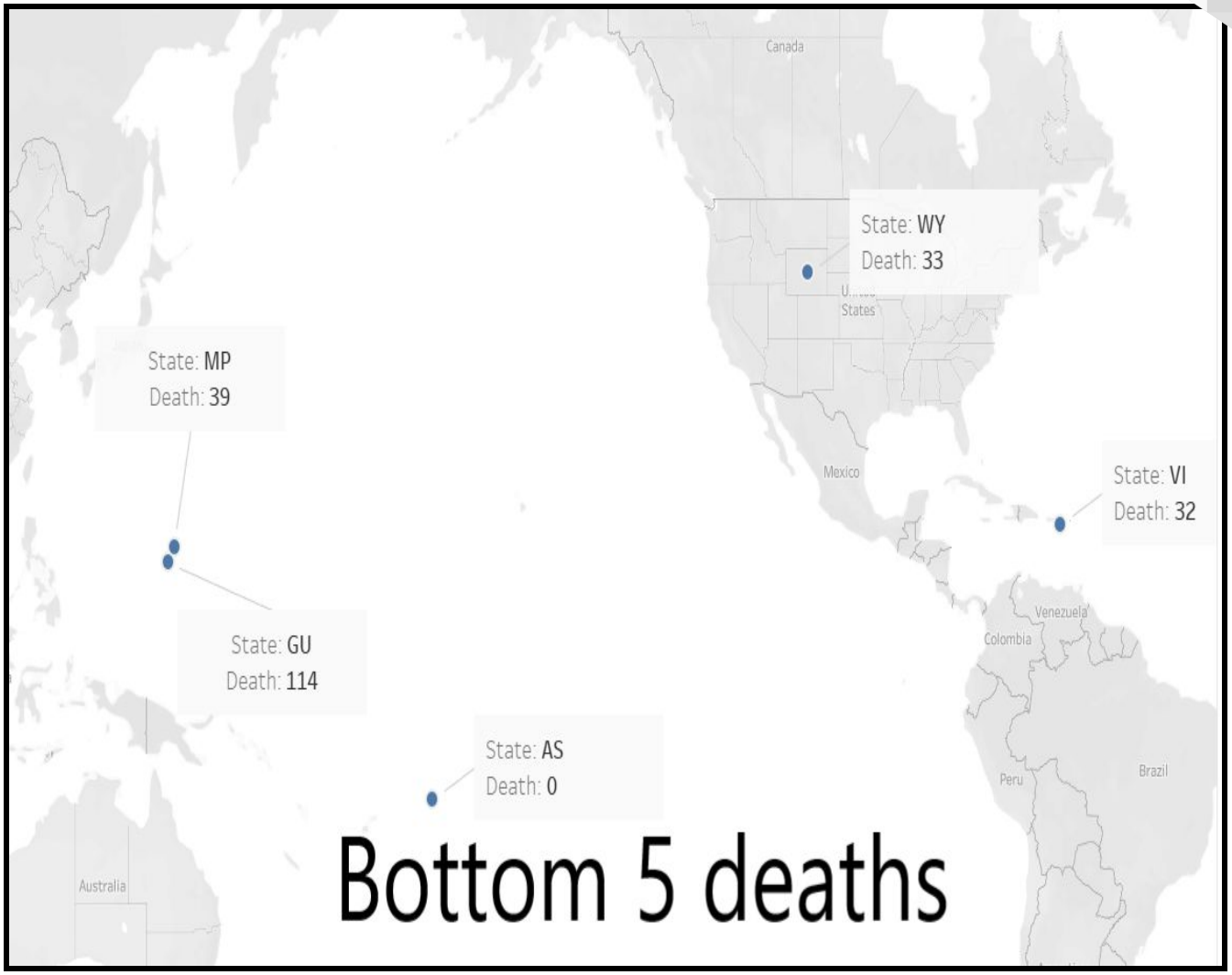


Figure:- Bottom 5 states that have deaths

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.

```
# which state will recover faster from the given dataset
```{r}
#omiting na values
cases%>%is.na()%>%colSums()
clean_cases=cases%>%drop_na()
clean_cases1=clean_cases%>%filter(date=="20200423")
covid19_cases=clean_cases1%>%mutate(ratio=recovered/positive)%>%arrange(desc(ratio))%>%head(25)
recoverstate=covid19_cases%>%select(state,ratio)
recoverstate
ggplot(covid19_cases,aes(x=state,y=ratio))+geom_point()
```
```

R Console | data.frame 25 x 2

	state <fctr>	ratio <dbl>
1	VT	2.0727273
2	GU	0.9333333
3	VI	0.8888889
4	WY	0.8435583
5	MP	0.7857143
6	HI	0.7500000
7	MT	0.6923077
8	SC	0.6745983
9	OK	0.6244614
10	AK	0.6201780

Chunk 7

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.

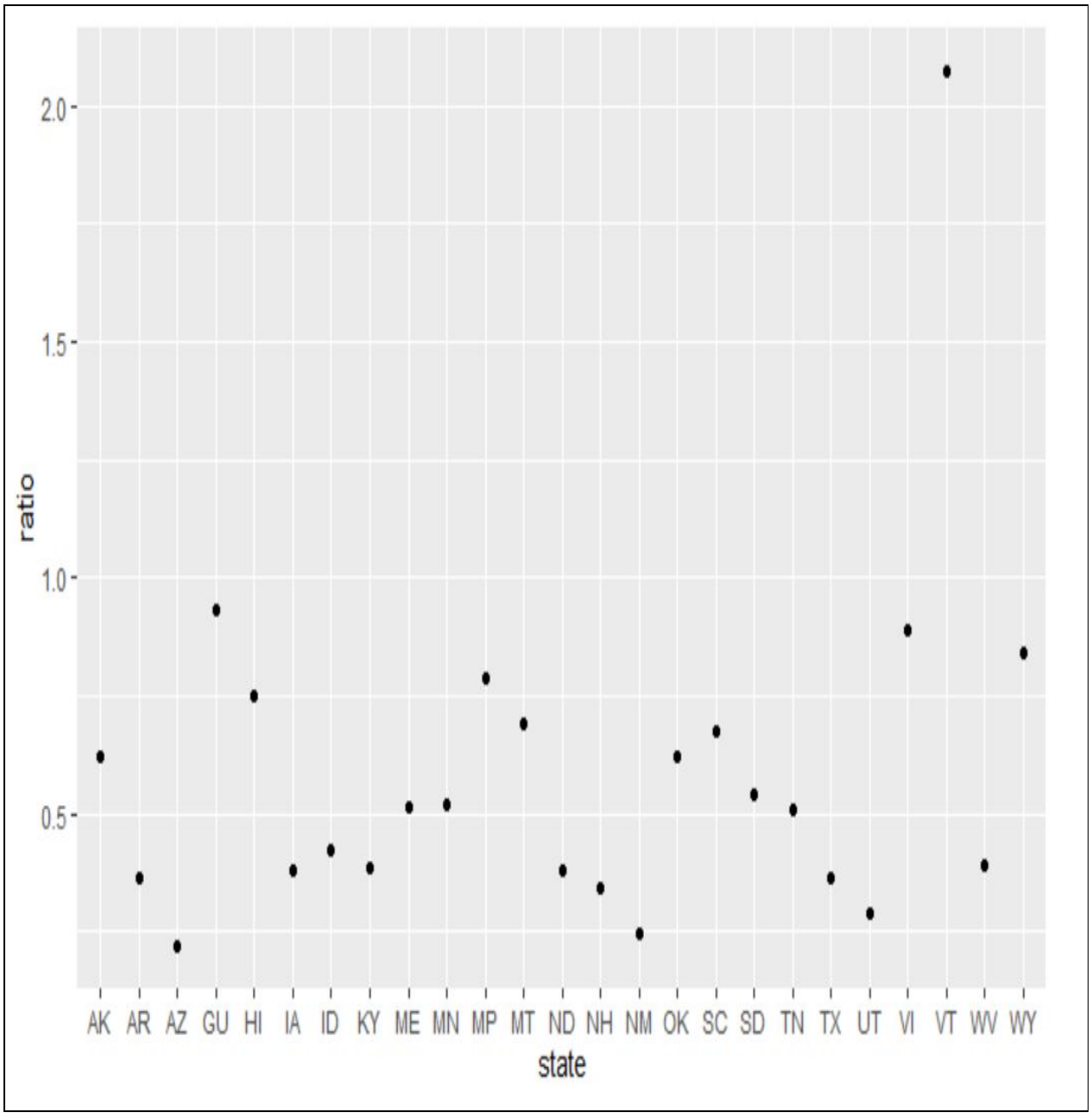


Figure: Graphical representation of higher possibility to recover faster

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.

```
#state that is in critical situation
```{r}

risky=clean_cases1%>%mutate(recover=(recovered+1))%>%arrange(recover)
riskystates=risky%>%mutate(ratio=deathIncrease+1/recover)%>%arrange(desc(ratio))%>%head(25)
riskystates
riskystates%>%select(state,ratio)
ggplot(riskystates,aes(x=state,y=ratio))+geom_point()
```
```

	state <fctr>	ratio <dbl>
1	NY	438.00004
2	NJ	306.00000
3	MA	179.00000
4	MI	164.00031
5	IL	124.00000
6	MD	117.00096
7	CA	116.00000
8	CT	96.00000
9	FL	70.00000
10	LA	68.00000

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.

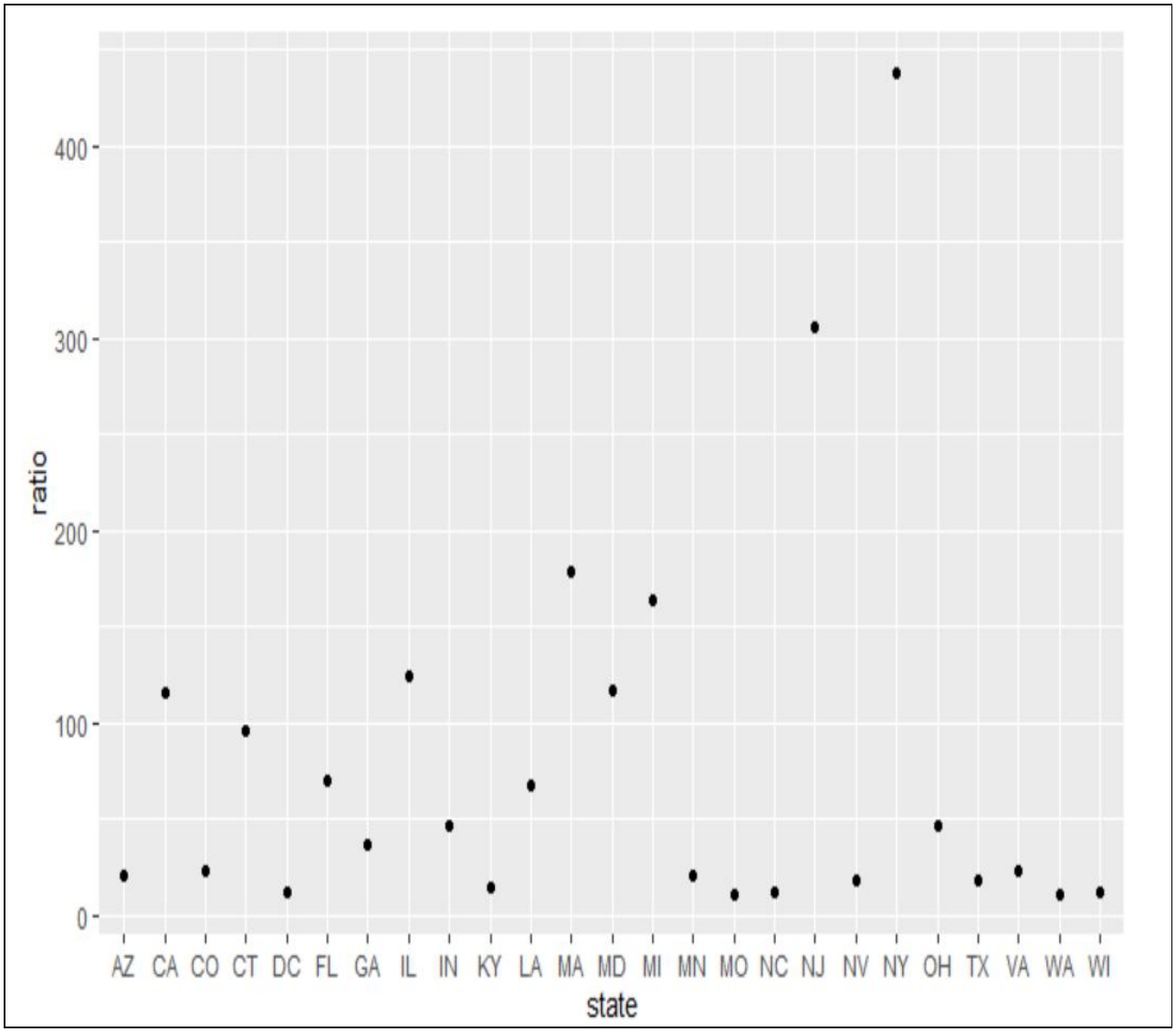


Figure:- Graphical representation of states that are in a critical situation

Project objectives, observations and future recommendations:-

Objectives of the Project:-

➔ The main objective of our project is to get an overview of the current scenario and insights from the COVID-19 pandemic situation in the United States of America, like which states are safer or danger and required to take necessary action.

➔ Additionally, we are trying to highlight the preventive measures to control COVID-19 across the country, given by the two agencies in the United States known as the White House Task Force and Centers for Disease Control and Prevention(CDC).

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.

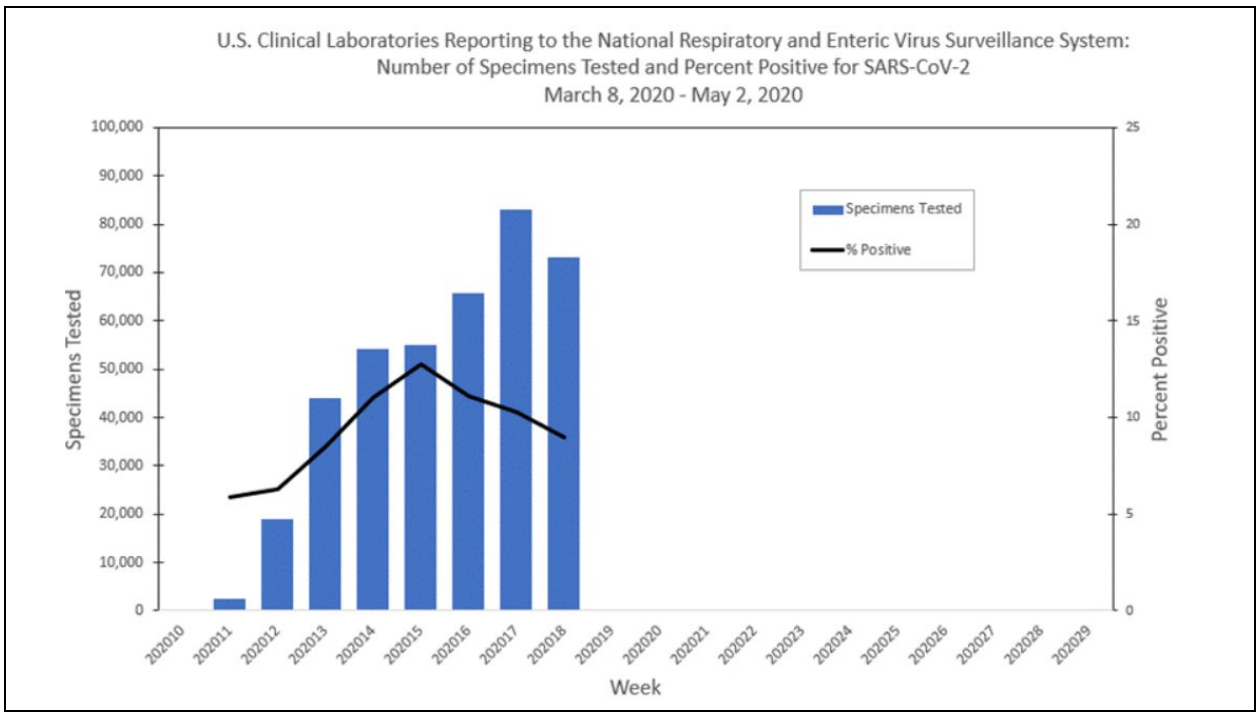
After reading these reports our team came up to the recommendation that we can take to stop this by taking a few steps like:-

- ➔ 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:-**

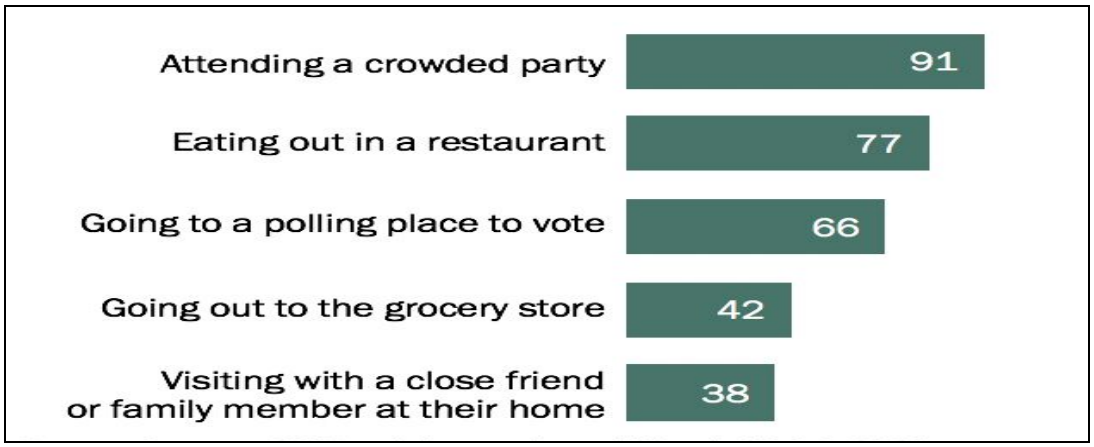


This graph records the covid data till 9th May. The above graph displays the testing kits during the period and the amount of positive cases recorded. As per the graph, the positive test cases were increasing steadily during the first ten week and then the rate was almost doubled after the twelfth week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

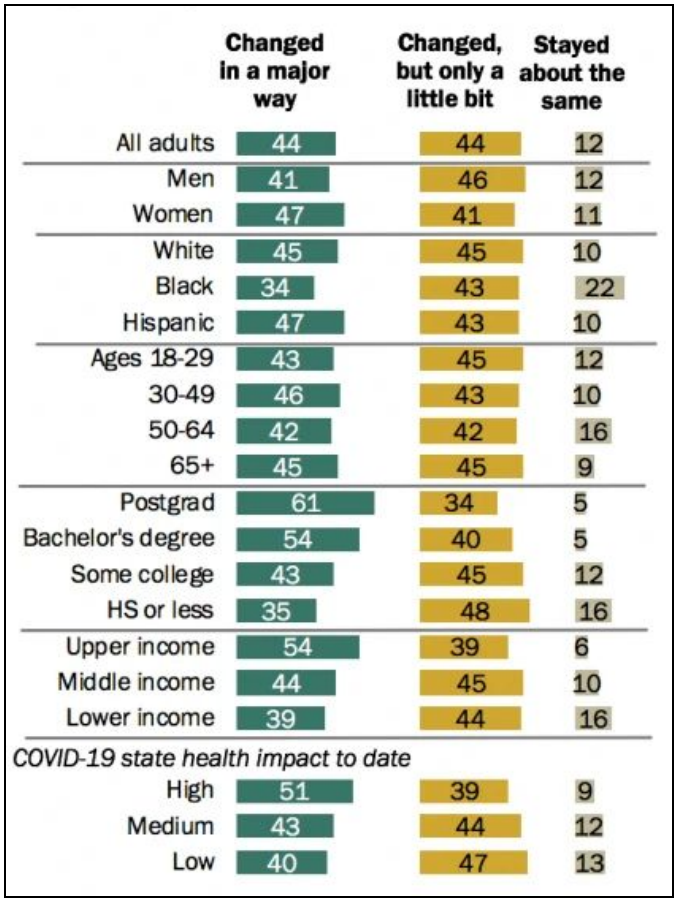
Many infected people can have mild symptoms and hence, it is possible for a person to catch coronavirus as from those people. Apart from that, there have been recordings of some people with no symptoms transmitting the virus. The WHO is currently doing their research on these absurd reports.

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

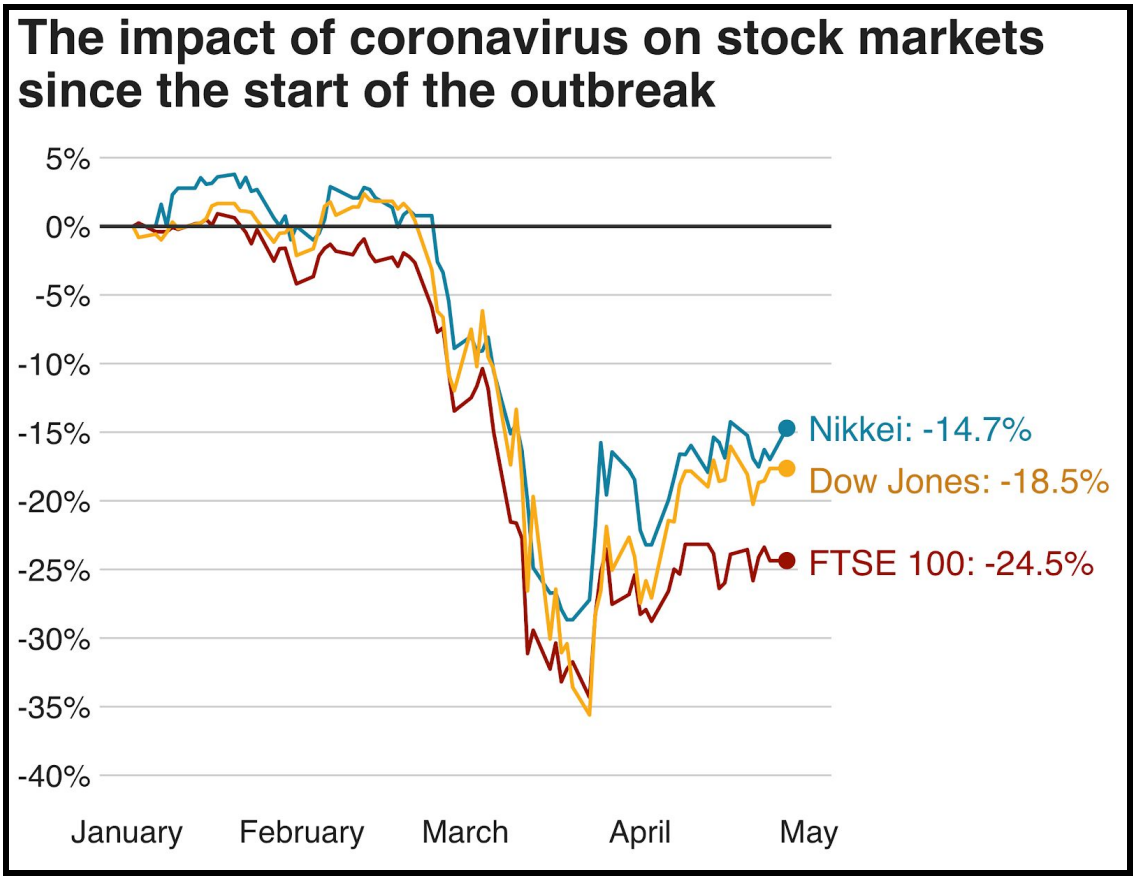
seperate them. They were able to get a downward curve and get their country up and running in a short period of time.



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.



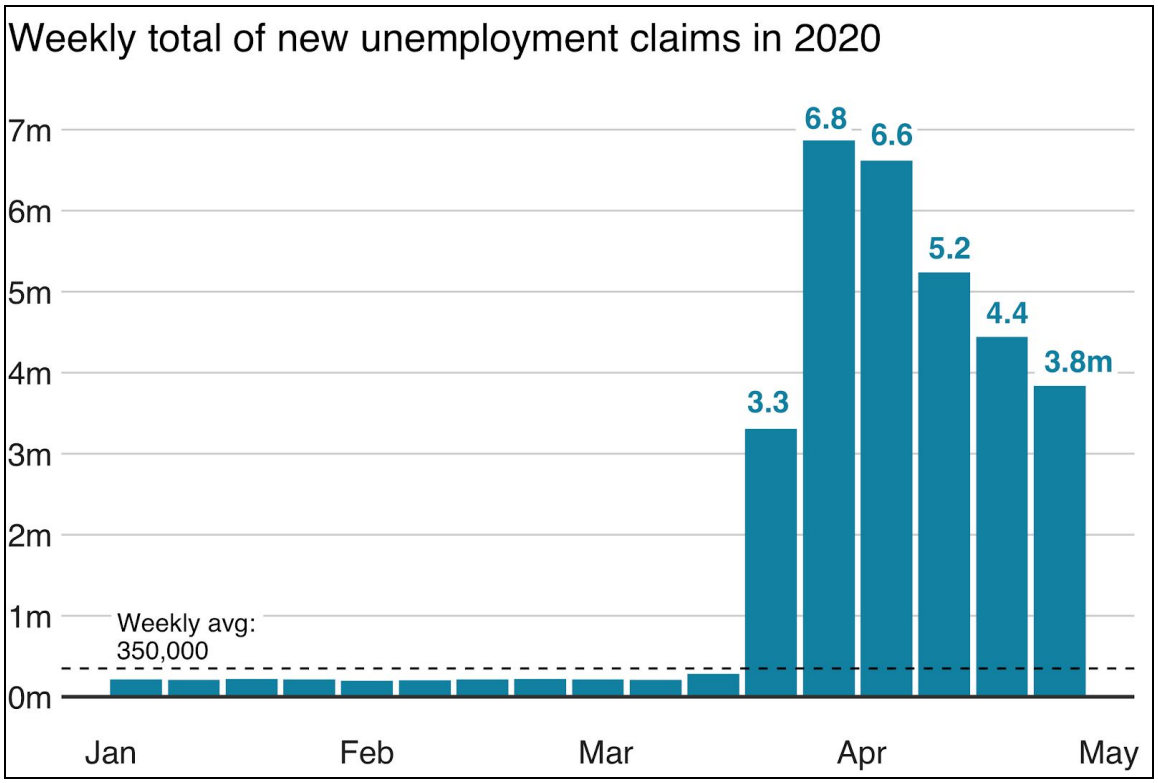
The figure mentioned above basically explained the behavioral changes in various age groups, students, based on their incomes, all the adults, based on their gender, based on their races and effect on health. It basically shows that out of 100 people, the amount of people who changed in a major way,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

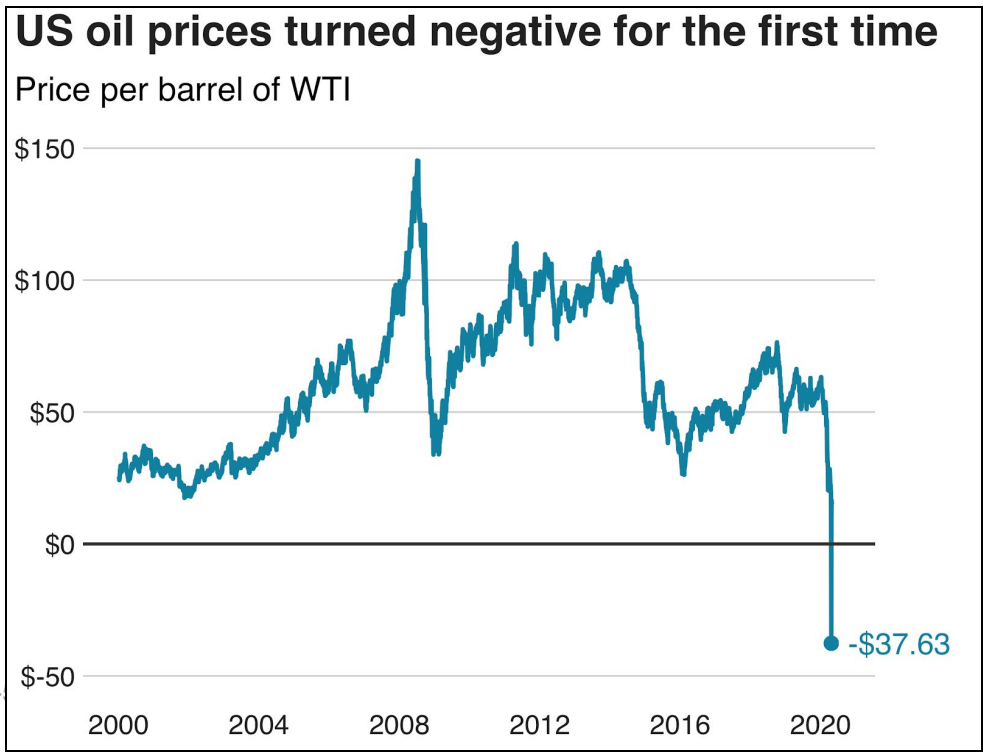


The coronavirus outbreak and the effect of it on the stock markets was observed in January. As per the image, the stock was around 0% and 5% during the January and then the prices dropped below 0 in february and increased a bit then they were plummeted during the mid-March and during the month of april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

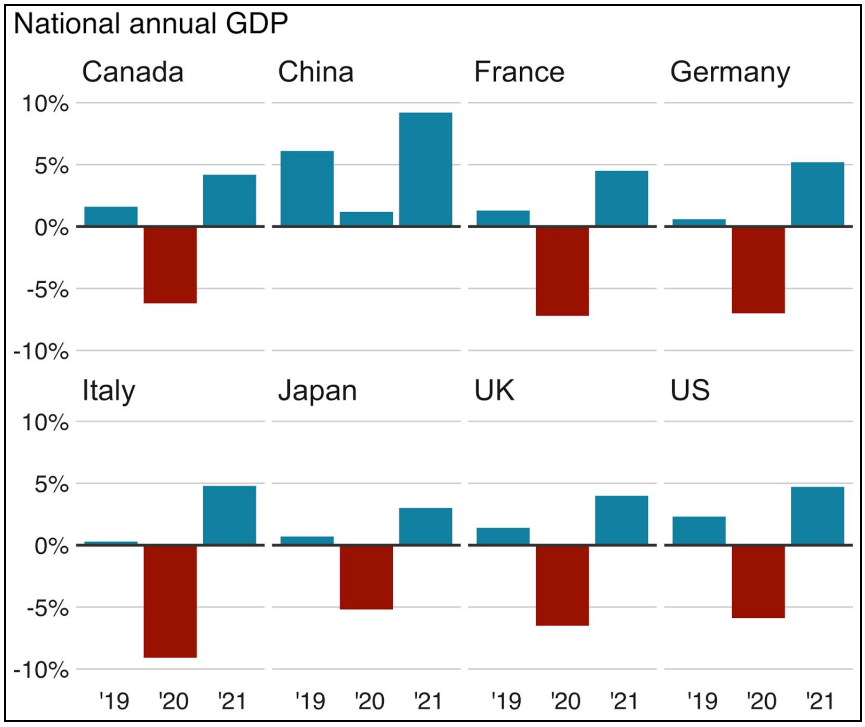
The figure mentioned below, displays the unemployment rate after the outbreak of coronavirus based on weeks. As per the graph, the unemployment was around 350k per week from january to mid-march and during the period of mid-march to may the

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.

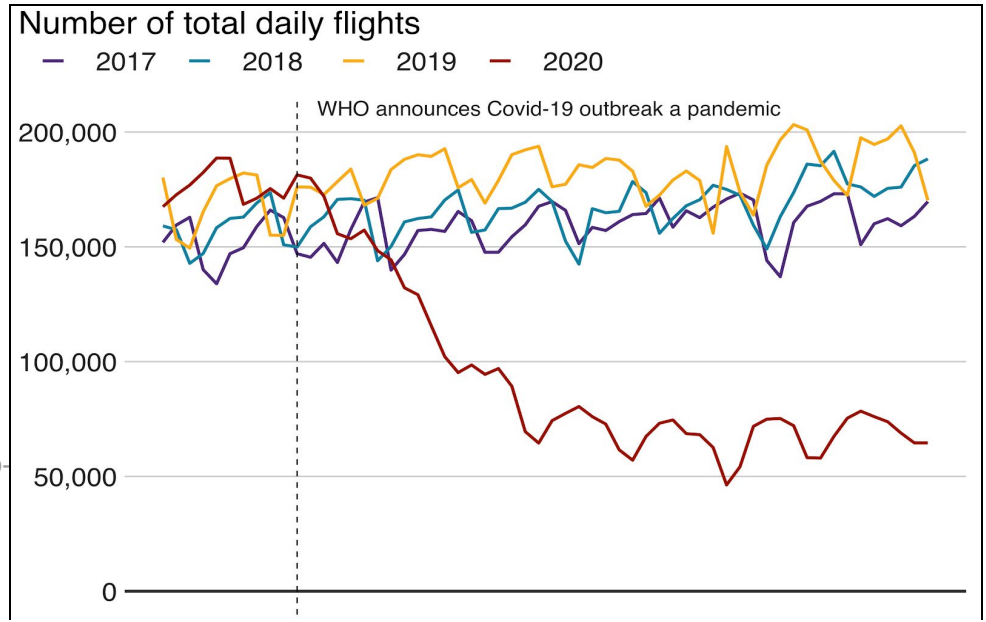




For the first time in history the US oil price per barrel plummeted below $0, as per the figure it is currently at the lowest -$37.63.

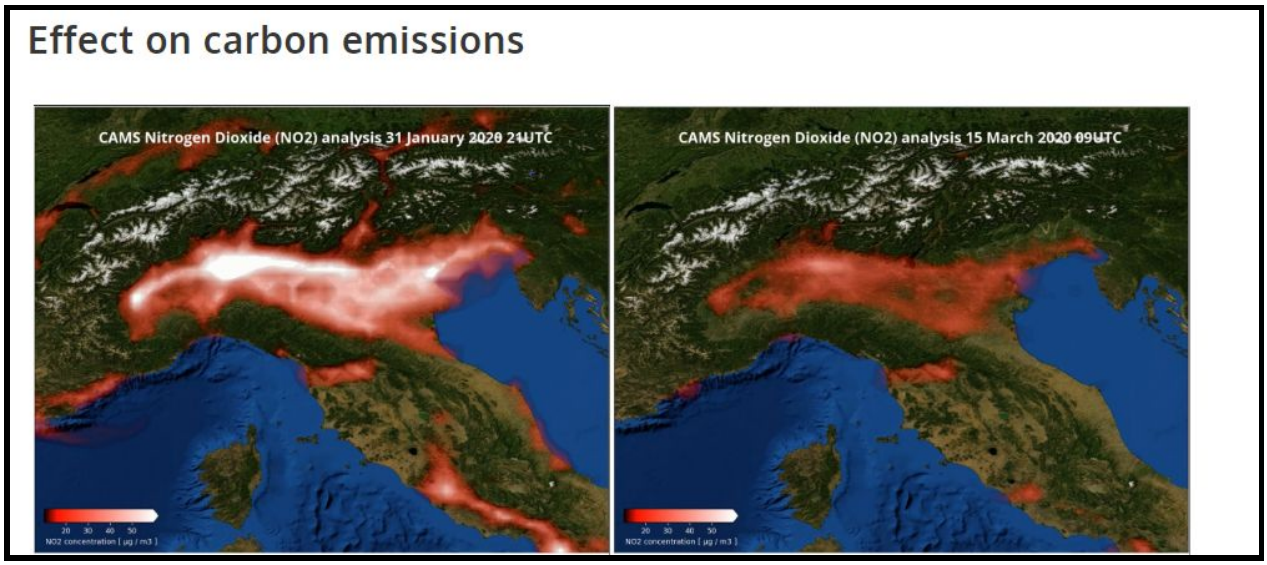


The national annual GDP is mentioned in the figure above. As per the image, every nation that has taken a serious hit from the coronavirus have fallen below -5% and the rest of them have been able to keep their economy around -5% except China who has managed to keep their economy well above 0% although 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 and the TomTom 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.