DATA ANALYSIS OF COVID-19 IN DIFFERENT COUNTRIES

Chandigarh College of Engg. & Tech.

CCET

Submitted by

ABOLI

(CO17305)

PIYUSH MALHOTRA

(CO17346)

SIDDHARTH SAMBER

(CO17358)

Under the supervision of

(Dr. Ankit Gupta)

CONTENTS

- ACKNOWLEDGEMENT
- ABSTRACT
- OVERVIEW
- INDIA
- SOUTH KOREA
- USA
- ITALY

ACKNOWLEDGEMENT

I would like to put forth my regardful thanks to all those who had helped me in guiding me to get the most out of this golden opportunity to make the data mining assignment on topic Covid-19 at Chandigarh College of Engineering & Technology Sec-26, Chandigarh.

I pay my special thanks to Dr. Ankit Gupta (Assistant Professor, CSE Department), who gave me a chance to work with him and gave me excellent knowledge. Also, I put my special regards to the management, team members and friends who have always been so supporting and ready to help.

I also put forth my special thanks to all the concerned persons as well as Dr. Sunil K. Singh, (HOD, CSE), Dr. Ankit Gupta(Training In-charge, CSE) and Dr. Manpreet Singh, Principal C.C.E.T. (Degree Wing), Chandigarh, who have enabled me to have an opportunity to work at the prestigious organization.

ABSTRACT

In this report, we are representing the analysis,observations and details of the impact of COVID 19 in the countries India,South Korea, USA and Italy.

It consists of detailed description of the various datasets,statistics and impact of the pandemic over the different countries whose data we are analysing here.

The purpose of the study was to perform data mining on the impact of the virus over the countries and analyse it so as to handle the situation better and in an efficient manner. Necessary precautions could be taken accordingly.

Observations and conclusions could be derived from the visualisations presented.

OVERVIEW

COVID 19

COVID-19 is a respiratory illness caused by a new virus. Symptoms include fever, coughing, sore throat and shortness of breath. The virus can spread from person to person, but good hygiene can prevent infection.

It spreads faster than other diseases, like common cold. Every virus has a Basic Reproduction number (R0) which implies how many people will get the disease from the infected person.

Currently the goal of all scientists around the world is to "Flatten the Curve". COVID-19 currently has an exponential growth rate around the world which we will be seeing in the notebook ahead. Flattening the Curve typically implies even if the number of Confirmed Cases are increasing but the distribution of those cases should be over a longer timestamp.

Every Pandemic has four stages:

- Stage 1: Confirmed Cases come from other countries
- Stage 2: Local Transmission Begins
- Stage 3: Communities impacted with local transmission
- Stage 4: Significant Transmission with no end in sight

Italy, USA, UK and France are some countries which are currently in Stage 4. While India is on the edge of Stage 3.

TOOLS USED

- ORANGE
- PYTHON SCRIPT
- MICROSOFT EXCEL

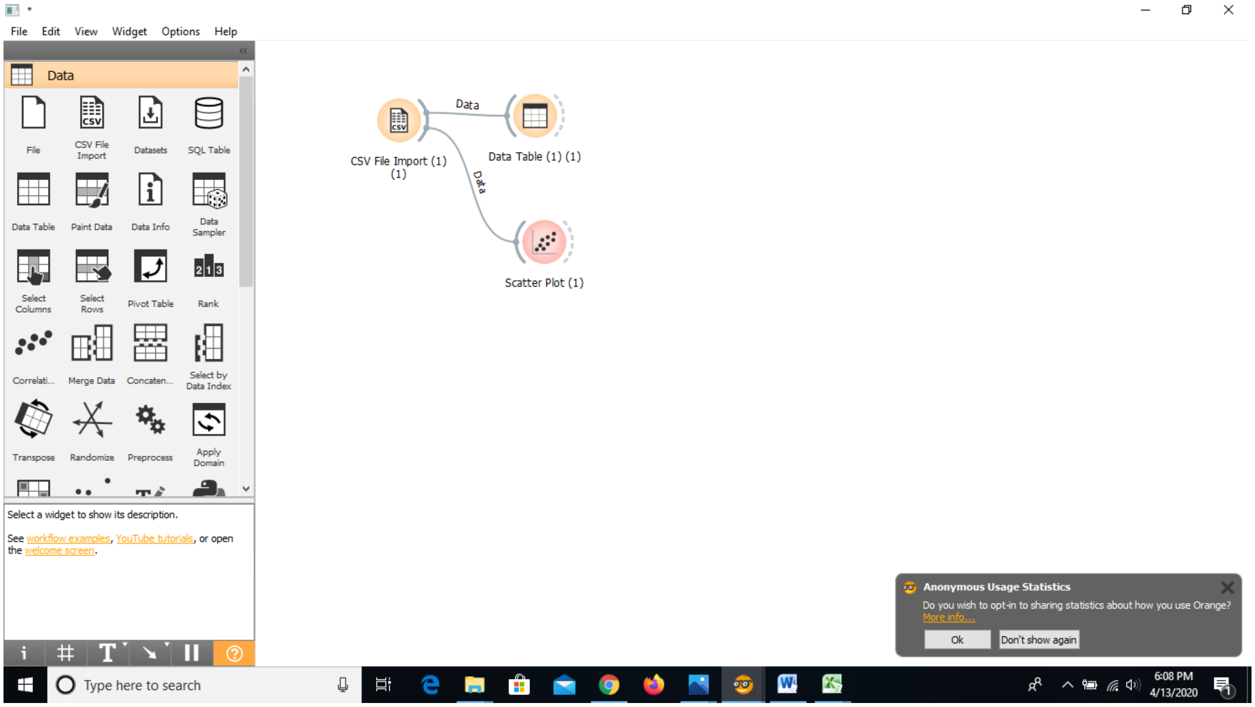
ORANGE DATA MINING TOOL

Orange is a C++ core object and routines library that incorporates a huge variety of standard and non-standard machine learning and data mining algorithms. It is an open-source data visualization, data mining, and machine learning tool. Orange is a scriptable environment for quick prototyping of the latest algorithms and testing patterns. It is a group of python-based modules that exist in the core library. It implements some functionalities for which execution time is not essential, and that is done in Python.

Orange is a set of graphical widgets that utilizes strategies from the core library and orange modules and gives a decent user interface. The widget supports digital-based communication and can be gathered together into an application by a visual programming tool called an orange canvas.

The objective of Orange is to provide a platform for experiment-based selection, predictive modeling, and recommendation system. It primarily used in bioinformatics, genomic research, biomedicine, and teaching. In education, it is used for providing better teaching methods for data mining and machine learning to students of biology, biomedicine, and informatics.

Demo workflow in orange:



Select a widget to show its description.

See workflow examples, YouTube tutorials, or open the welcome screen.

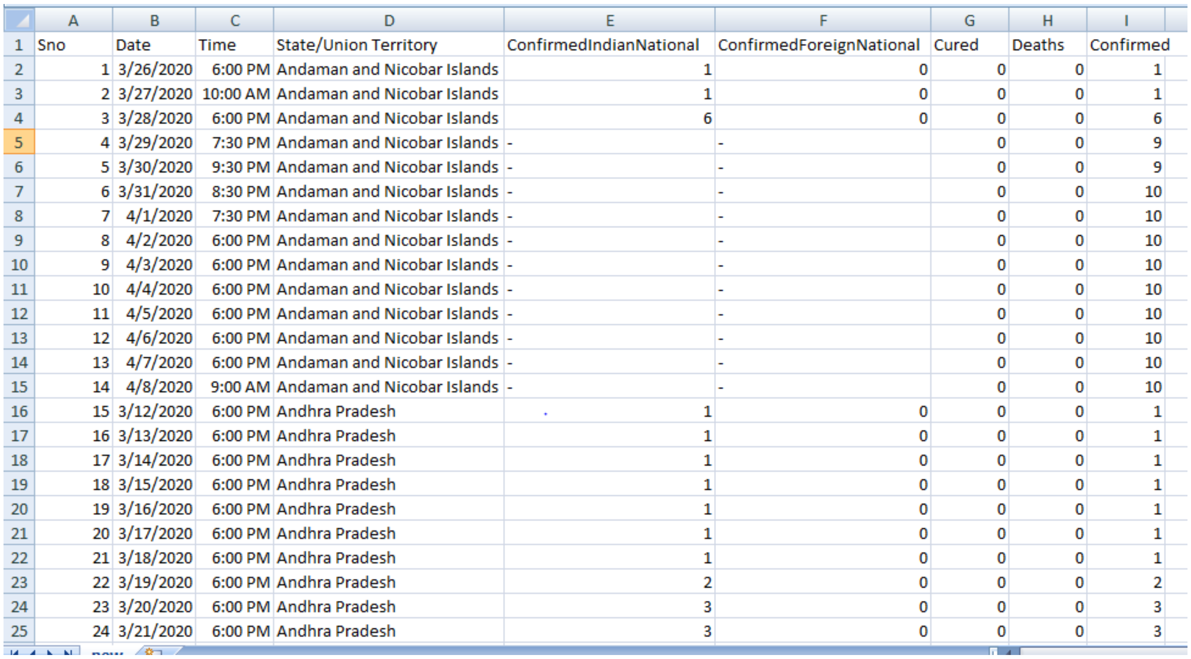
Anonymous Usage Statistics

Do you wish to opt-in to sharing statistics about how you use Orange? More info...

INDIA

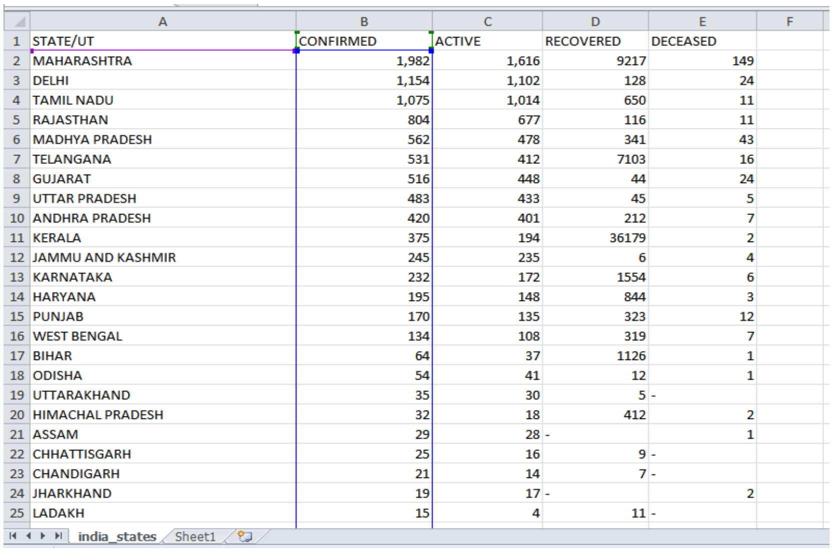
The dataset taken in consideration here shows the impact of COVID 19 on the various states and union territories of India with time.

DATASET



	A	B	C	D	E	F	G	H	I
1	Sno	Date	Time	State/Union Territory	ConfirmedIndianNational	ConfirmedForeignNational	Cured	Deaths	Confirmed
2	1	3/26/2020	6:00 PM	Andaman and Nicobar Islands	1	0	0	0	1
3	2	3/27/2020	10:00 AM	Andaman and Nicobar Islands	1	0	0	0	1
4	3	3/28/2020	6:00 PM	Andaman and Nicobar Islands	6	0	0	0	6
5	4	3/29/2020	7:30 PM	Andaman and Nicobar Islands	-	-	0	0	9
6	5	3/30/2020	9:30 PM	Andaman and Nicobar Islands	-	-	0	0	9
7	6	3/31/2020	8:30 PM	Andaman and Nicobar Islands	-	-	0	0	10
8	7	4/1/2020	7:30 PM	Andaman and Nicobar Islands	-	-	0	0	10
9	8	4/2/2020	6:00 PM	Andaman and Nicobar Islands	-	-	0	0	10
10	9	4/3/2020	6:00 PM	Andaman and Nicobar Islands	-	-	0	0	10
11	10	4/4/2020	6:00 PM	Andaman and Nicobar Islands	-	-	0	0	10
12	11	4/5/2020	6:00 PM	Andaman and Nicobar Islands	-	-	0	0	10
13	12	4/6/2020	6:00 PM	Andaman and Nicobar Islands	-	-	0	0	10
14	13	4/7/2020	6:00 PM	Andaman and Nicobar Islands	-	-	0	0	10
15	14	4/8/2020	9:00 AM	Andaman and Nicobar Islands	-	-	0	0	10
16	15	3/12/2020	6:00 PM	Andhra Pradesh	1	0	0	0	1
17	16	3/13/2020	6:00 PM	Andhra Pradesh	1	0	0	0	1
18	17	3/14/2020	6:00 PM	Andhra Pradesh	1	0	0	0	1
19	18	3/15/2020	6:00 PM	Andhra Pradesh	1	0	0	0	1
20	19	3/16/2020	6:00 PM	Andhra Pradesh	1	0	0	0	1
21	20	3/17/2020	6:00 PM	Andhra Pradesh	1	0	0	0	1
22	21	3/18/2020	6:00 PM	Andhra Pradesh	1	0	0	0	1
23	22	3/19/2020	6:00 PM	Andhra Pradesh	2	0	0	0	2
24	23	3/20/2020	6:00 PM	Andhra Pradesh	3	0	0	0	3
25	24	3/21/2020	6:00 PM	Andhra Pradesh	3	0	0	0	3

In the dataset, various cases reported in different states or union territories have been recorded with the respective dates. The total number confirmed cases, deaths and recovered patients are present as attributes.



	A	B	C	D	E	F
1	STATE/UT	CONFIRMED	ACTIVE	RECOVERED	DECEASED	
2	MAHARASHTRA	1,982	1,616	9217	149	
3	DELHI	1,154	1,102	128	24	
4	TAMIL NADU	1,075	1,014	650	11	
5	RAJASTHAN	804	677	116	11	
6	MADHYA PRADESH	562	478	341	43	
7	TELANGANA	531	412	7103	16	
8	GUJARAT	516	448	44	24	
9	UTTAR PRADESH	483	433	45	5	
10	ANDHRA PRADESH	420	401	212	7	
11	KERALA	375	194	36179	2	
12	JAMMU AND KASHMIR	245	235	6	4	
13	KARNATAKA	232	172	1554	6	
14	HARYANA	195	148	844	3	
15	PUNJAB	170	135	323	12	
16	WEST BENGAL	134	108	319	7	
17	BIHAR	64	37	1126	1	
18	ODISHA	54	41	12	1	
19	UTTARAKHAND	35	30	5	-	
20	HIMACHAL PRADESH	32	18	412	2	
21	ASSAM	29	28	-	1	
22	CHHATTISGARH	25	16	9	-	
23	CHANDIGARH	21	14	7	-	
24	JHARKHAND	19	17	-	2	
25	LADAKH	15	4	11	-	

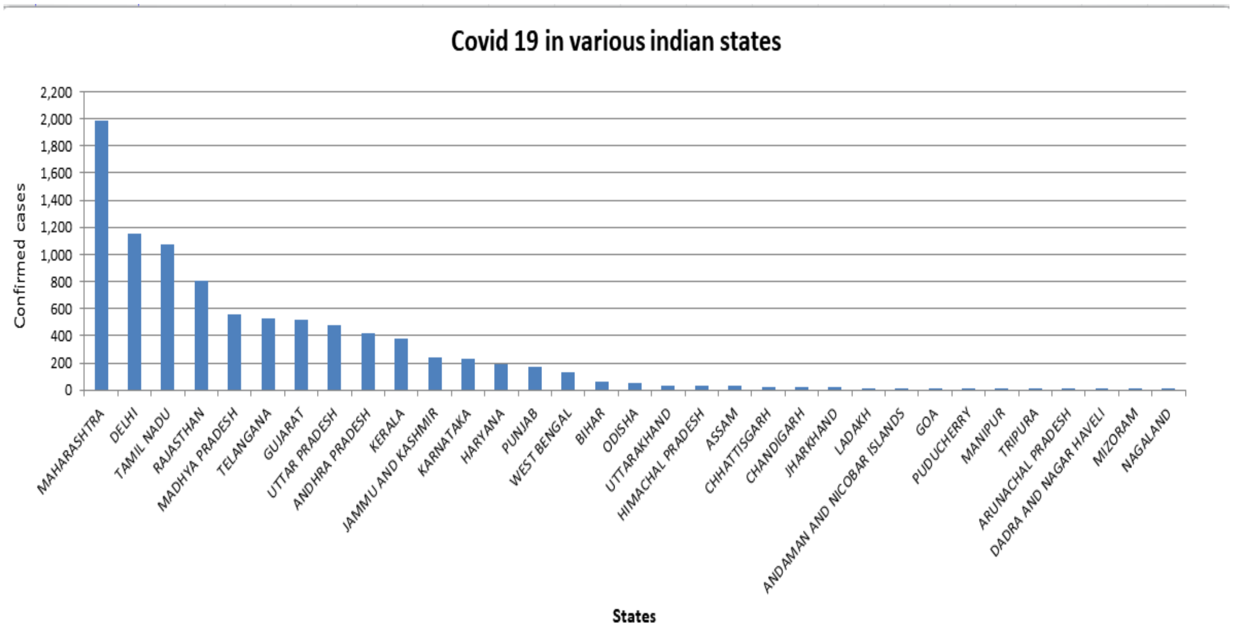
ANALYSIS

ANALYSING THE GIVEN STATISTICS

1. Impact of COVID 19 on different States/Union Territories of India

A bar graph has been plotted, depicting the number of confirmed cases in the various different states.
The data here is over the time period of 30/01/2020 to 12/04/2020.

Covid 19 Cases Vs States



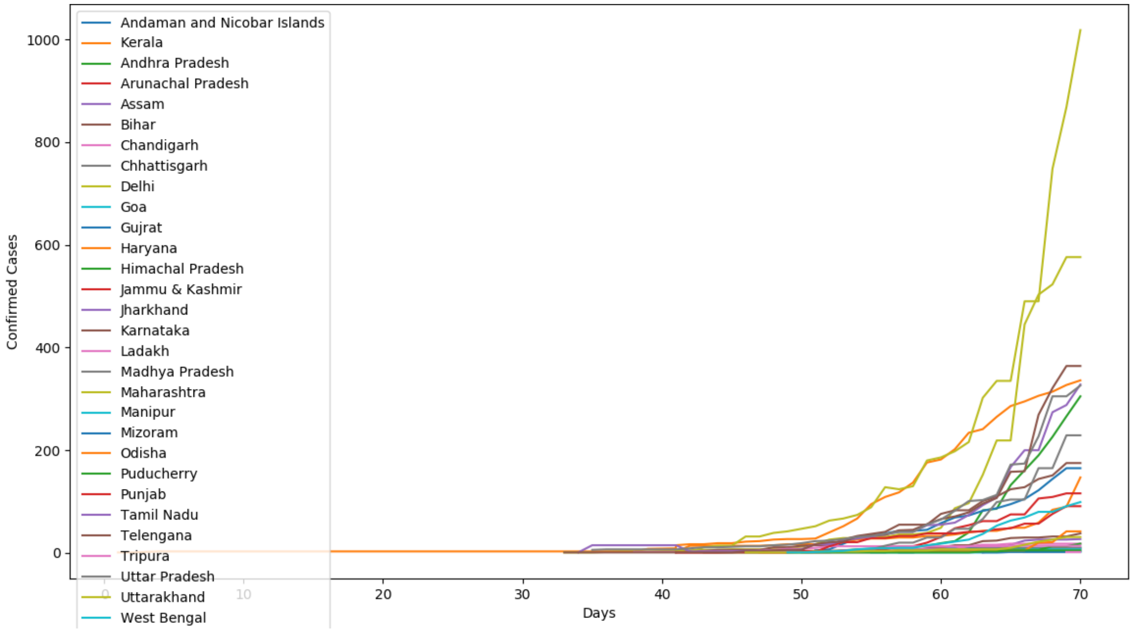
The following observations and analysis could be drawn from the visualisation above:

1. It gives an overview of which state is facing the most severe impact of the covid 19 and are hotspots for further faster transmission.
   Here,we can see that Maharastra is the most affected.
2. It is a comparative analysis among the states based on the total confirmed cases.

Scope:

1. More strict lockdown strategies could be imposed on the regions facing a greater impact of covid 19.
2. The health facilities could be increased or managed accordingly and efficiently in the states that are more affected.

3. The pandemic is surely going to affect the economy due to lockdown. Certain relaxations could be given in less transmitted areas to resume work but with precautionary measures.

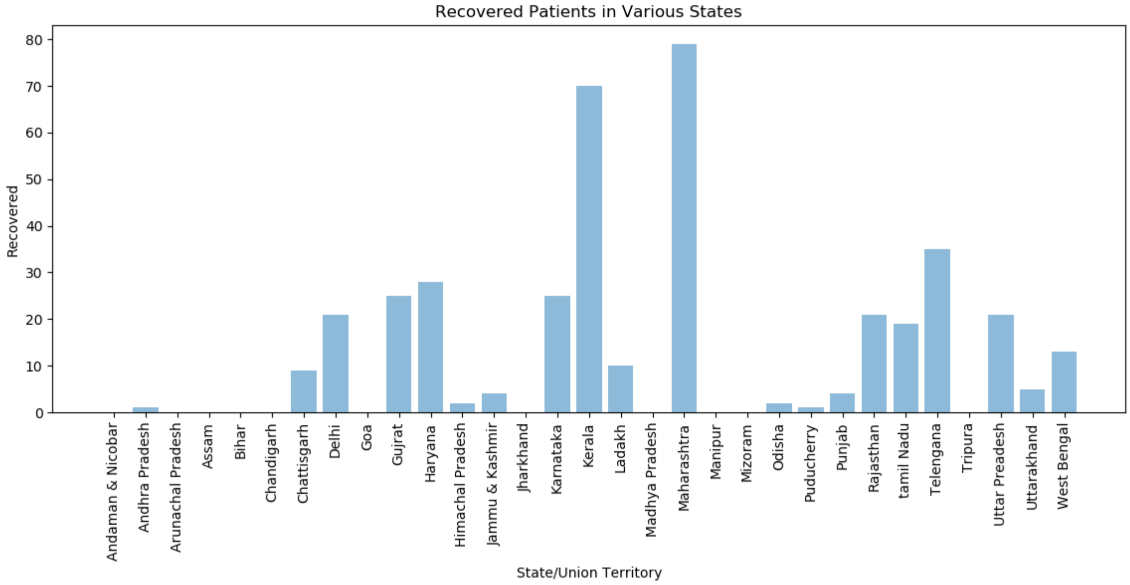


Analysis and Observations:

1. The above line plot shows the exponential growth(or hike) in the number of cases. Greater the slope more is the transmission rate in that rate.
2. The initial point of each curve is the day the first case was reported in that state.The first case was reported on 30/01/2020 in Kerala.

2.1 Recovered Patients in different States

Some proportion of the covid patients are said to have recovered from the same.

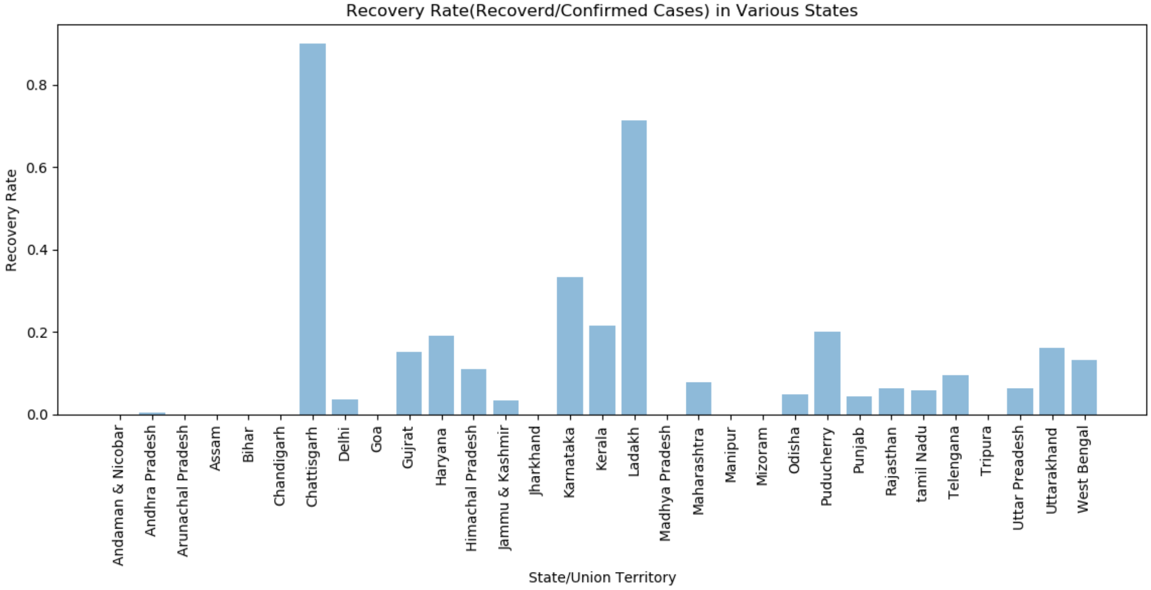


The bar graph depicts the number of recovered patients in various states.

Analysis and Observations:

States like Maharashtra and Kerala as of now have the highest number of recovered patients.

2.2 Studying the Recovered/Confirmed Ratio in different States

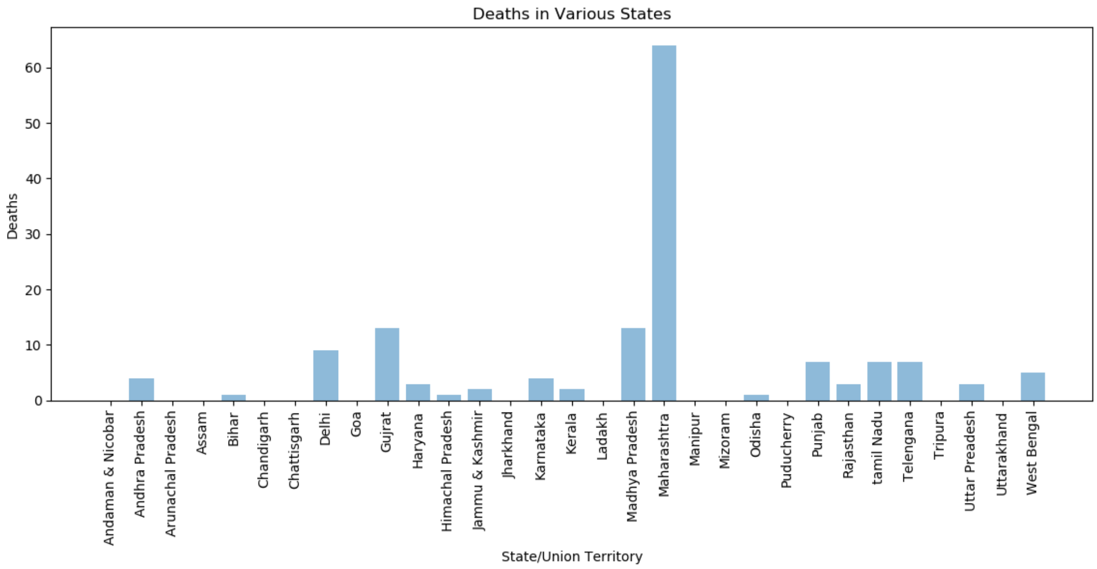


The above plot is between Recovered/Confirmed Vs States.
Recovered/Confirmed cases is a ratio and ranges [0,1]

Analysis and Observations:

As of now, the ratio is highest in Chattisgarh,i.e., the most recovered patients when compared to the total confirmed cases in that area/region.

3. Deaths due to COVID 19

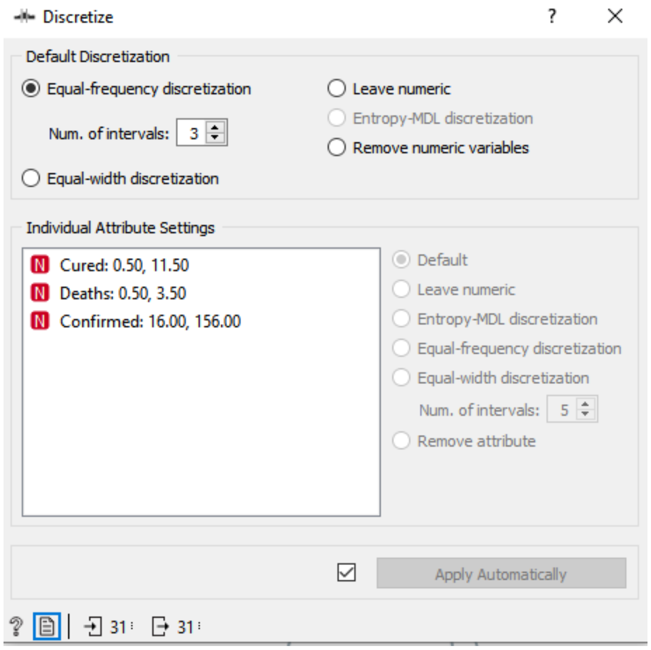


Analysis and Observations :
As of now, Maharashtra has recorded the highest number of deaths. Maharashtra is facing the greatest impact of the pandemic. The confirmed cases and deaths are highest in the state.

DISCRETIZATION

Applying discretization on the attributes confirmed cases, recovered and deaths according to frequency distribution in the form of three intervals and then classifying them as HIGH,MEDIUM and LOW.

On performing discretization on the said attributes, Orange divides them into three equal frequency district as follows:



Discretize

Default Discretization

● Equal-frequency discretization

Num. of intervals: 3

○ Equal-width discretization

○ Leave numeric

○ Entropy-MDL discretization

○ Remove numeric variables

Individual Attribute Settings

N Cured: 0.50, 11.50

N Deaths: 0.50, 3.50

N Confirmed: 16.00, 156.00

○ Default

○ Leave numeric

○ Entropy-MDL discretization

○ Equal-frequency discretization

○ Equal-width discretization

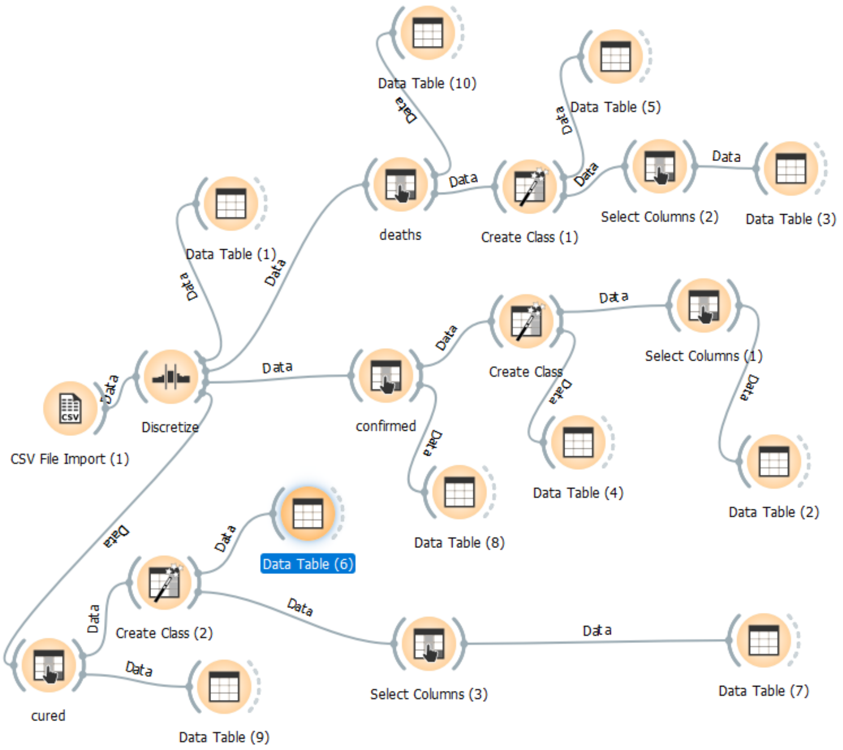
Num. of intervals: 5

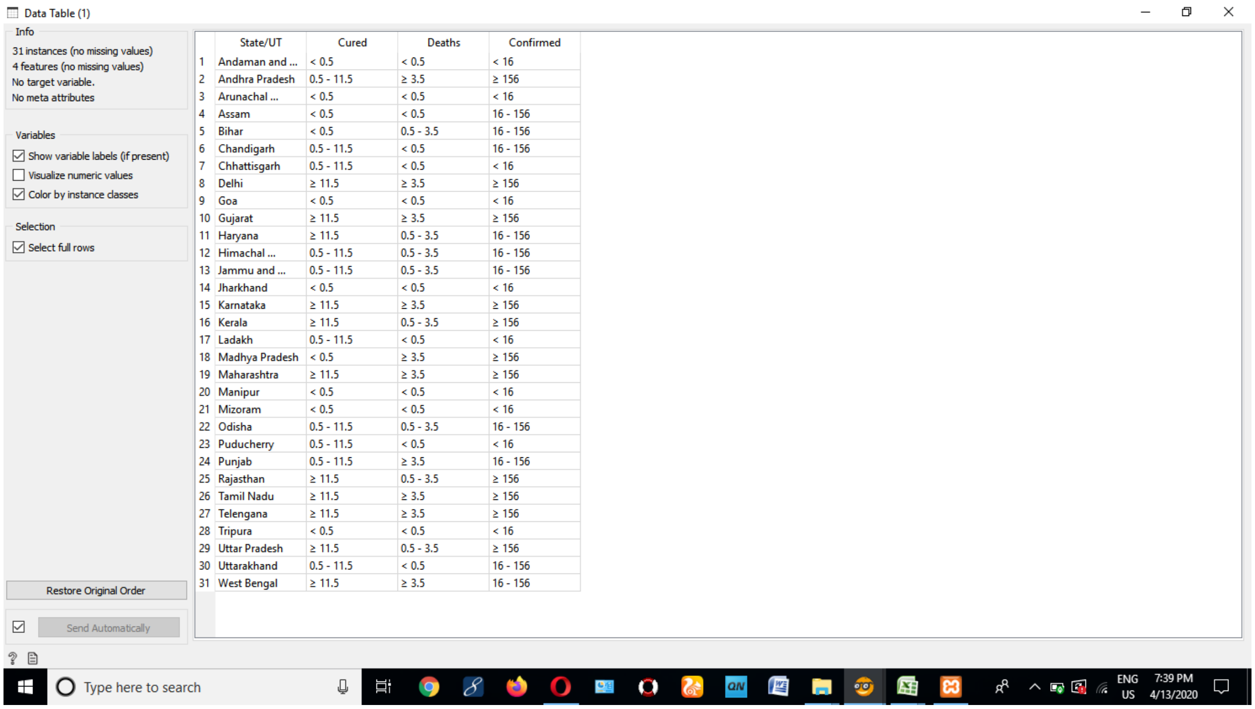
○ Remove attribute

☑ Apply Automatically

31 31

Here, discretization gives us a general idea regarding the range or interval where a tuple lies with respect to a specific attribute.
After discretization,





Data Table (1)

Info

31 instances (no missing values)
4 features (no missing values)
No target variable.
No meta attributes

Variables

☑ Show variable labels (if present)
☐ Visualize numeric values
☑ Color by instance classes

Selection

☑ Select full rows

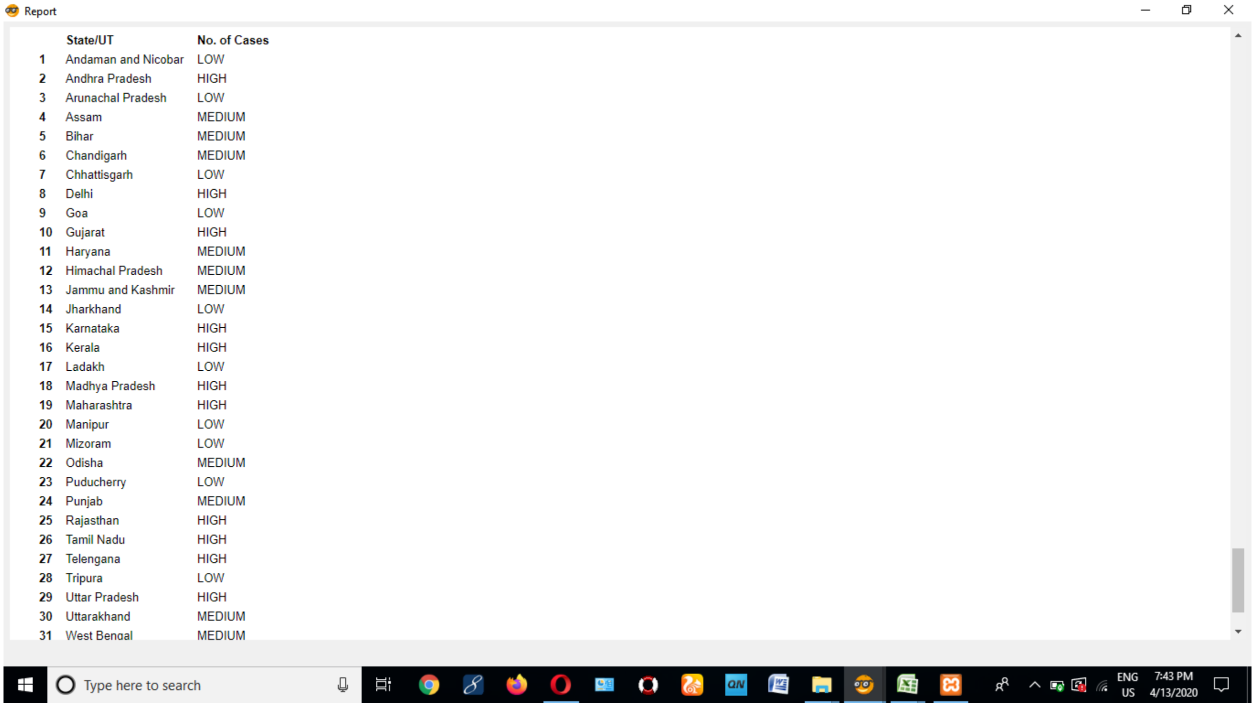
	State/UT	Cured	Deaths	Confirmed
1	Andaman and ...	< 0.5	< 0.5	< 16
2	Andhra Pradesh	0.5 - 11.5	≥ 3.5	≥ 156
3	Arunachal ...	< 0.5	< 0.5	< 16
4	Assam	< 0.5	< 0.5	16 - 156
5	Bihar	< 0.5	0.5 - 3.5	16 - 156
6	Chandigarh	0.5 - 11.5	< 0.5	16 - 156
7	Chhattisgarh	0.5 - 11.5	< 0.5	< 16
8	Delhi	≥ 11.5	≥ 3.5	≥ 156
9	Goa	< 0.5	< 0.5	< 16
10	Gujarat	≥ 11.5	≥ 3.5	≥ 156
11	Haryana	≥ 11.5	0.5 - 3.5	16 - 156
12	Himachal ...	0.5 - 11.5	0.5 - 3.5	16 - 156
13	Jammu and ...	0.5 - 11.5	0.5 - 3.5	16 - 156
14	Jharkhand	< 0.5	< 0.5	< 16
15	Karnataka	≥ 11.5	≥ 3.5	≥ 156
16	Kerala	≥ 11.5	0.5 - 3.5	≥ 156
17	Ladakh	0.5 - 11.5	< 0.5	< 16
18	Madhya Pradesh	< 0.5	≥ 3.5	≥ 156
19	Maharashtra	≥ 11.5	≥ 3.5	≥ 156
20	Manipur	< 0.5	< 0.5	< 16
21	Mizoram	< 0.5	< 0.5	< 16
22	Odisha	0.5 - 11.5	0.5 - 3.5	16 - 156
23	Puducherry	0.5 - 11.5	< 0.5	< 16
24	Punjab	0.5 - 11.5	≥ 3.5	16 - 156
25	Rajasthan	≥ 11.5	0.5 - 3.5	≥ 156
26	Tamil Nadu	≥ 11.5	≥ 3.5	≥ 156
27	Telengana	≥ 11.5	≥ 3.5	≥ 156
28	Tripura	< 0.5	< 0.5	< 16
29	Uttar Pradesh	≥ 11.5	0.5 - 3.5	≥ 156
30	Uttarakhand	0.5 - 11.5	< 0.5	16 - 156
31	West Bengal	≥ 11.5	≥ 3.5	16 - 156

Restore Original Order

☑ Send Automatically

Now, classifying the three equal frequency intervals as LOW,MEDIUM and HIGH according to the interval a specific tuple lies in.
Considering the attributes individually corresponding to the states the following could be derived:

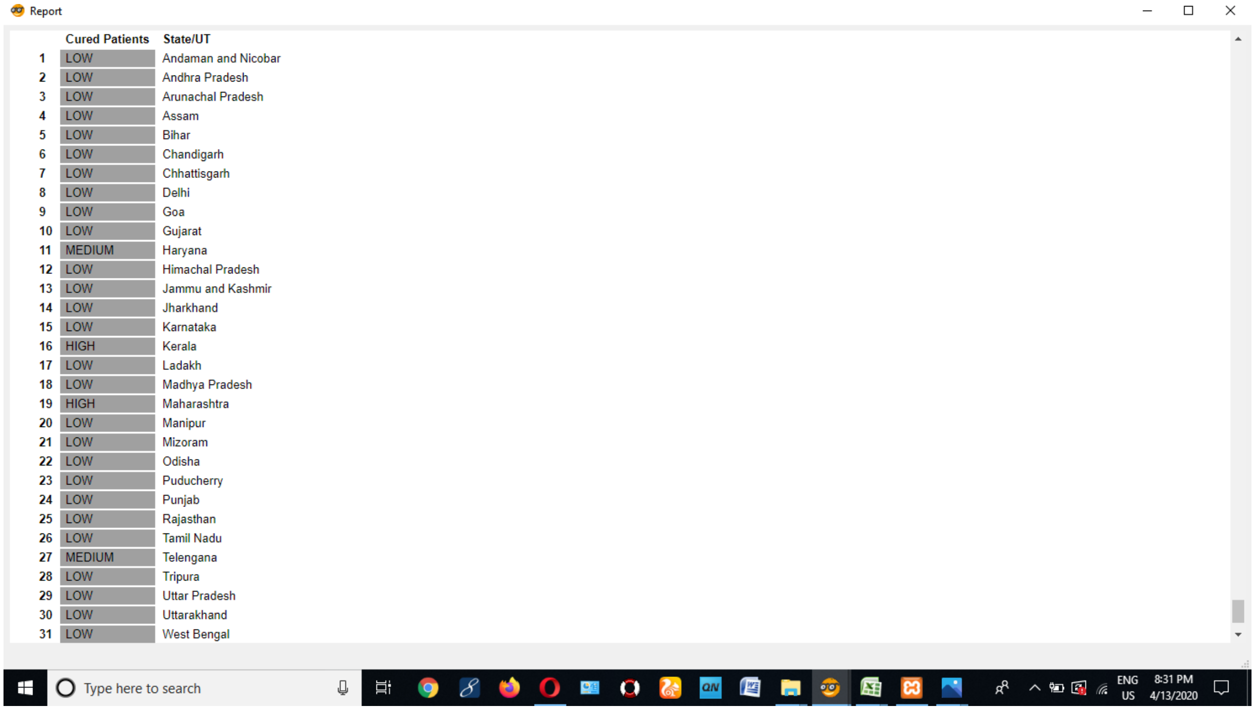
- **Confirmed Cases**



Report

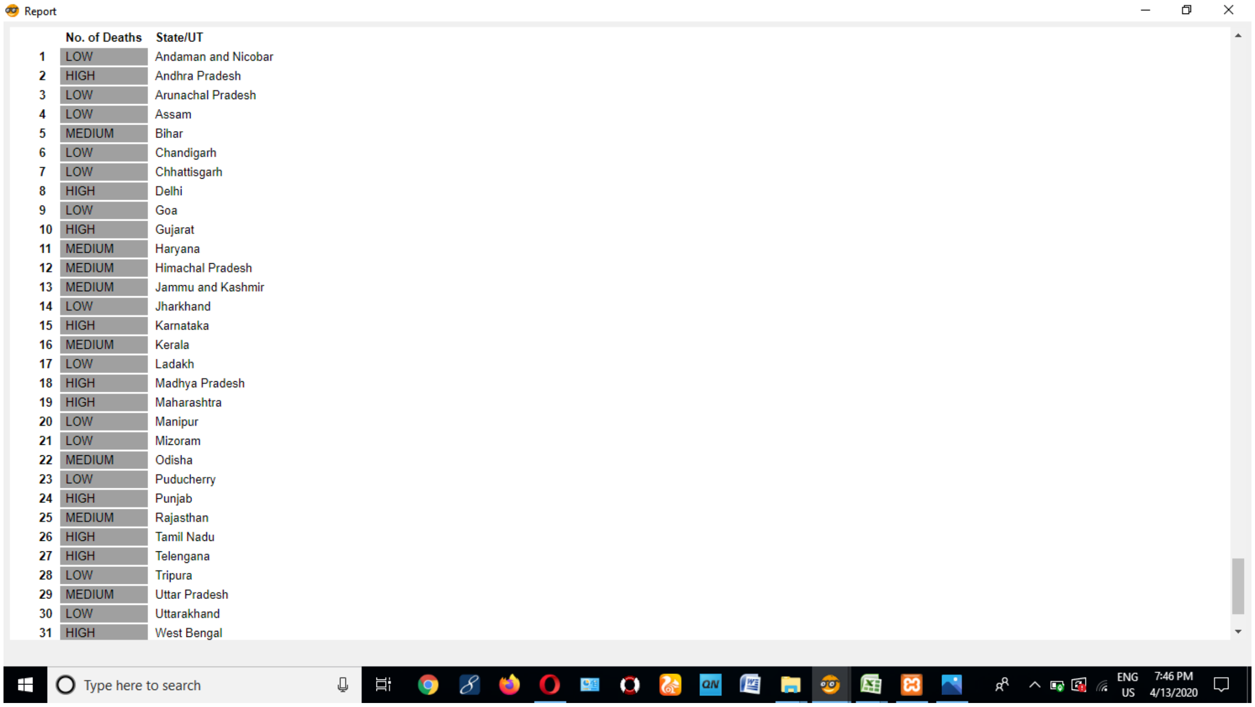
	State/UT	No. of Cases
1	Andaman and Nicobar	LOW
2	Andhra Pradesh	HIGH
3	Arunachal Pradesh	LOW
4	Assam	MEDIUM
5	Bihar	MEDIUM
6	Chandigarh	MEDIUM
7	Chhattisgarh	LOW
8	Delhi	HIGH
9	Goa	LOW
10	Gujarat	HIGH
11	Haryana	MEDIUM
12	Himachal Pradesh	MEDIUM
13	Jammu and Kashmir	MEDIUM
14	Jharkhand	LOW
15	Karnataka	HIGH
16	Kerala	HIGH
17	Ladakh	LOW
18	Madhya Pradesh	HIGH
19	Maharashtra	HIGH
20	Manipur	LOW
21	Mizoram	LOW
22	Odisha	MEDIUM
23	Puducherry	LOW
24	Punjab	MEDIUM
25	Rajasthan	HIGH
26	Tamil Nadu	HIGH
27	Telengana	HIGH
28	Tripura	LOW
29	Uttar Pradesh	HIGH
30	Uttarakhand	MEDIUM
31	West Bengal	MEDIUM

- ## Cured Patients



	Cured Patients	State/UT
1	LOW	Andaman and Nicobar
2	LOW	Andhra Pradesh
3	LOW	Arunachal Pradesh
4	LOW	Assam
5	LOW	Bihar
6	LOW	Chandigarh
7	LOW	Chhattisgarh
8	LOW	Delhi
9	LOW	Goa
10	LOW	Gujarat
11	MEDIUM	Haryana
12	LOW	Himachal Pradesh
13	LOW	Jammu and Kashmir
14	LOW	Jharkhand
15	LOW	Karnataka
16	HIGH	Kerala
17	LOW	Ladakh
18	LOW	Madhya Pradesh
19	HIGH	Maharashtra
20	LOW	Manipur
21	LOW	Mizoram
22	LOW	Odisha
23	LOW	Puducherry
24	LOW	Punjab
25	LOW	Rajasthan
26	LOW	Tamil Nadu
27	MEDIUM	Telengana
28	LOW	Tripura
29	LOW	Uttar Pradesh
30	LOW	Uttarakhand
31	LOW	West Bengal

- ## Deaths



	No. of Deaths	State/UT
1	LOW	Andaman and Nicobar
2	HIGH	Andhra Pradesh
3	LOW	Arunachal Pradesh
4	LOW	Assam
5	MEDIUM	Bihar
6	LOW	Chandigarh
7	LOW	Chhattisgarh
8	HIGH	Delhi
9	LOW	Goa
10	HIGH	Gujarat
11	MEDIUM	Haryana
12	MEDIUM	Himachal Pradesh
13	MEDIUM	Jammu and Kashmir
14	LOW	Jharkhand
15	HIGH	Karnataka
16	MEDIUM	Kerala
17	LOW	Ladakh
18	HIGH	Madhya Pradesh
19	HIGH	Maharashtra
20	LOW	Manipur
21	LOW	Mizoram
22	MEDIUM	Odisha
23	LOW	Puducherry
24	HIGH	Punjab
25	MEDIUM	Rajasthan
26	HIGH	Tamil Nadu
27	HIGH	Telengana
28	LOW	Tripura
29	MEDIUM	Uttar Pradesh
30	LOW	Uttarakhand
31	HIGH	West Bengal

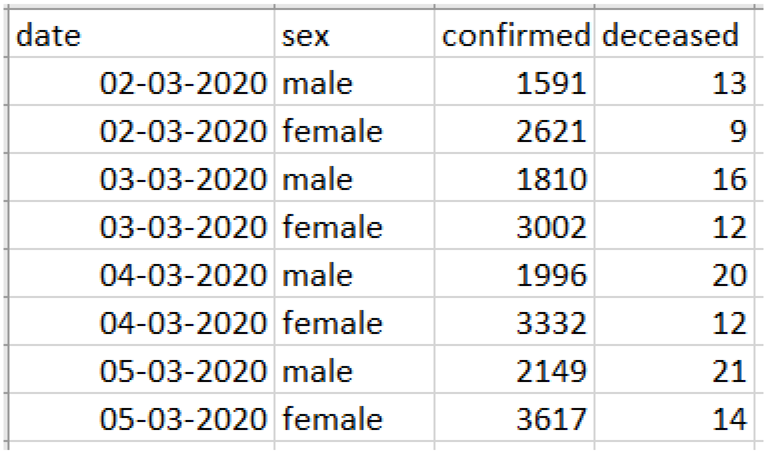
SOUTH KOREA

DATASET

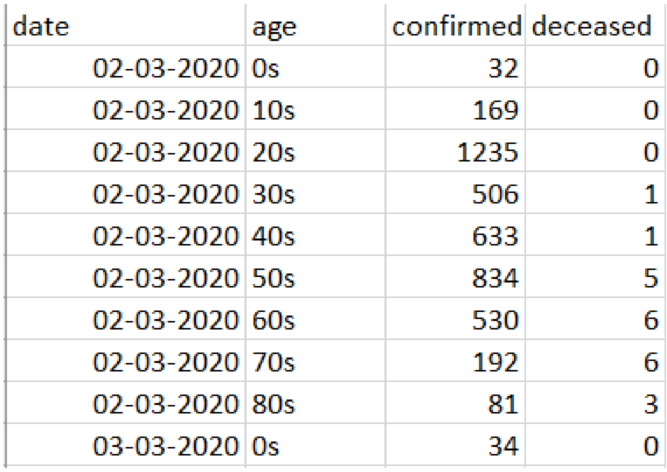
The following datasets have been used in the analysis:

- Time series data of COVID-19 status in terms of the age in South Korea.
- Time series data of COVID-19 status in terms of gender in South Korea
- Time series data of COVID-19 status in terms of the Province in South Korea

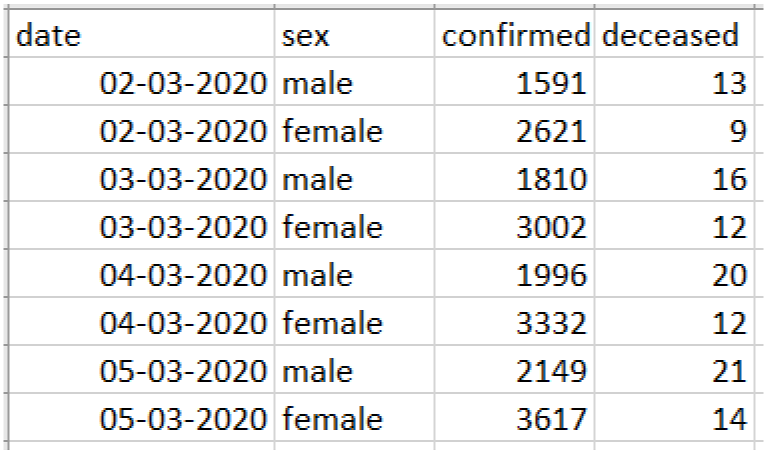
All the datasets provided have recorded the information from 2nd march to 7th april.



date	sex	confirmed	deceased
02-03-2020	male	1591	13
02-03-2020	female	2621	9
03-03-2020	male	1810	16
03-03-2020	female	3002	12
04-03-2020	male	1996	20
04-03-2020	female	3332	12
05-03-2020	male	2149	21
05-03-2020	female	3617	14



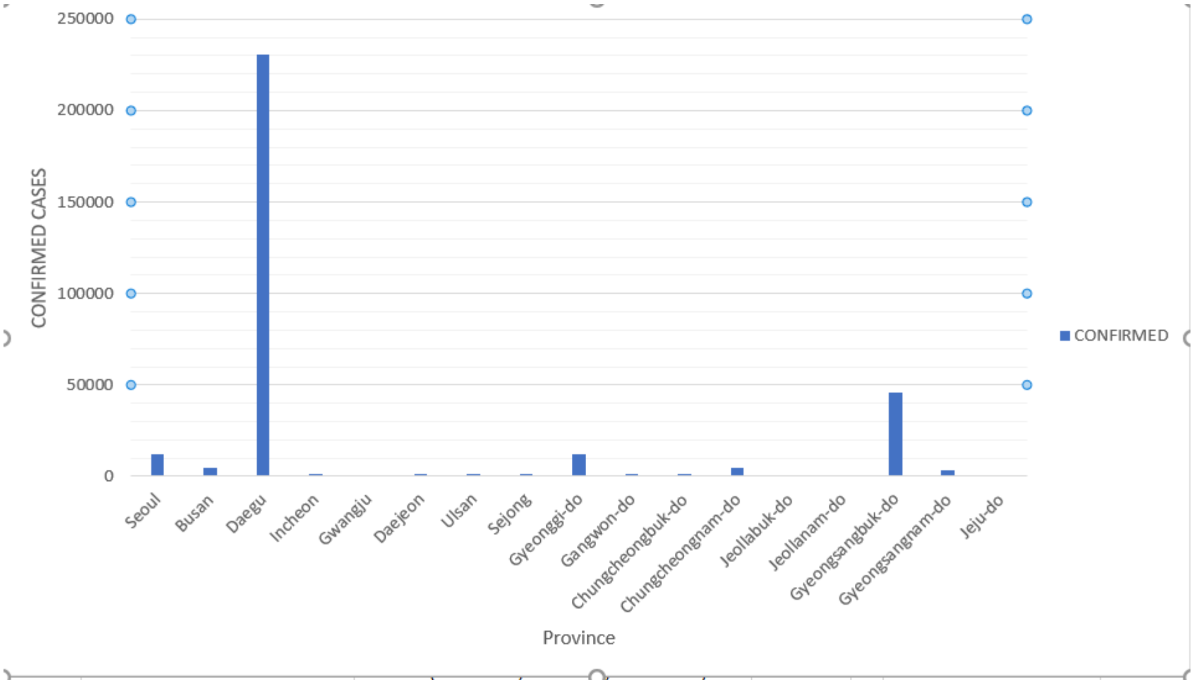
date	age	confirmed	deceased
02-03-2020	0s	32	0
02-03-2020	10s	169	0
02-03-2020	20s	1235	0
02-03-2020	30s	506	1
02-03-2020	40s	633	1
02-03-2020	50s	834	5
02-03-2020	60s	530	6
02-03-2020	70s	192	6
02-03-2020	80s	81	3
03-03-2020	0s	34	0



date	sex	confirmed	deceased
02-03-2020	male	1591	13
02-03-2020	female	2621	9
03-03-2020	male	1810	16
03-03-2020	female	3002	12
04-03-2020	male	1996	20
04-03-2020	female	3332	12
05-03-2020	male	2149	21
05-03-2020	female	3617	14

ANALYSIS

1. **Number of Cases recorded in various provinces of South Korea**

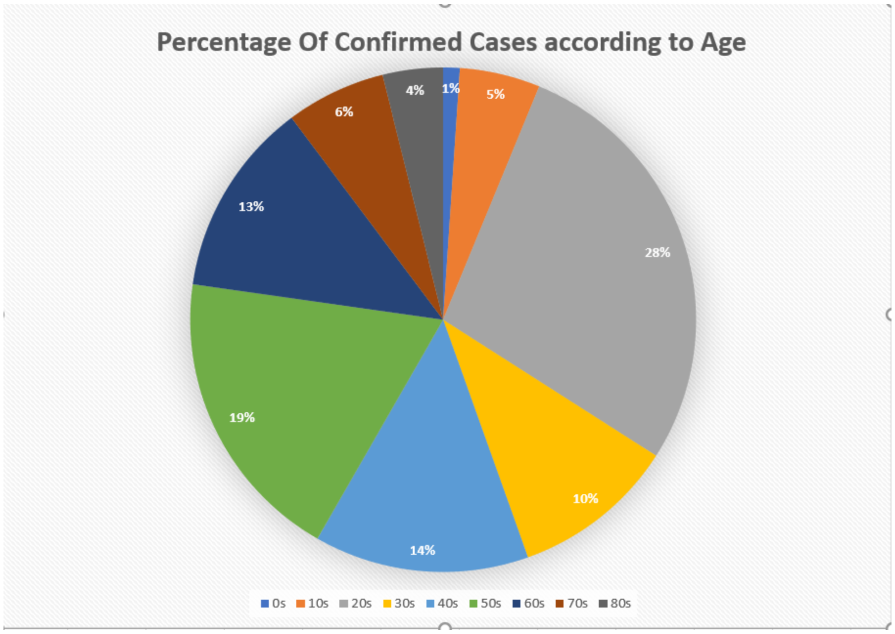


DESCRIPTION: The above bar graph shows the number of confirmed cases in different provinces of south korea.

INFERENCE:

- From the above graph it can be inferred that the highest number of confirmed cases were found in DAEGU province followed by Gyeongsangbuk-d and Seoul.
- In all other provinces, cases were in thousands and hundreds.

2. Distribution of cases among various age groups

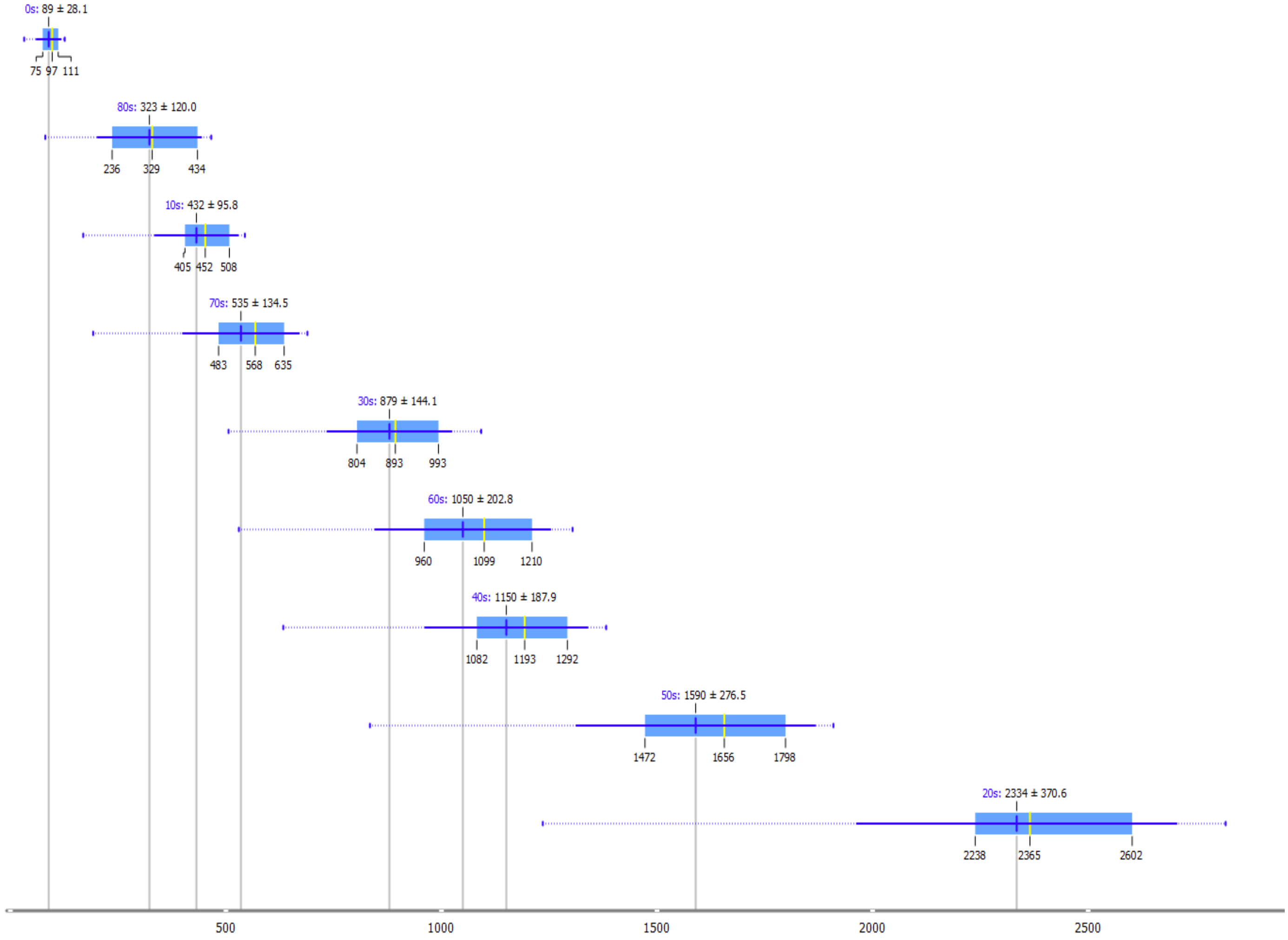


DESCRIPTION: The above pie chart represents the percentage of confirmed cases for each age group.

INFERENCE:

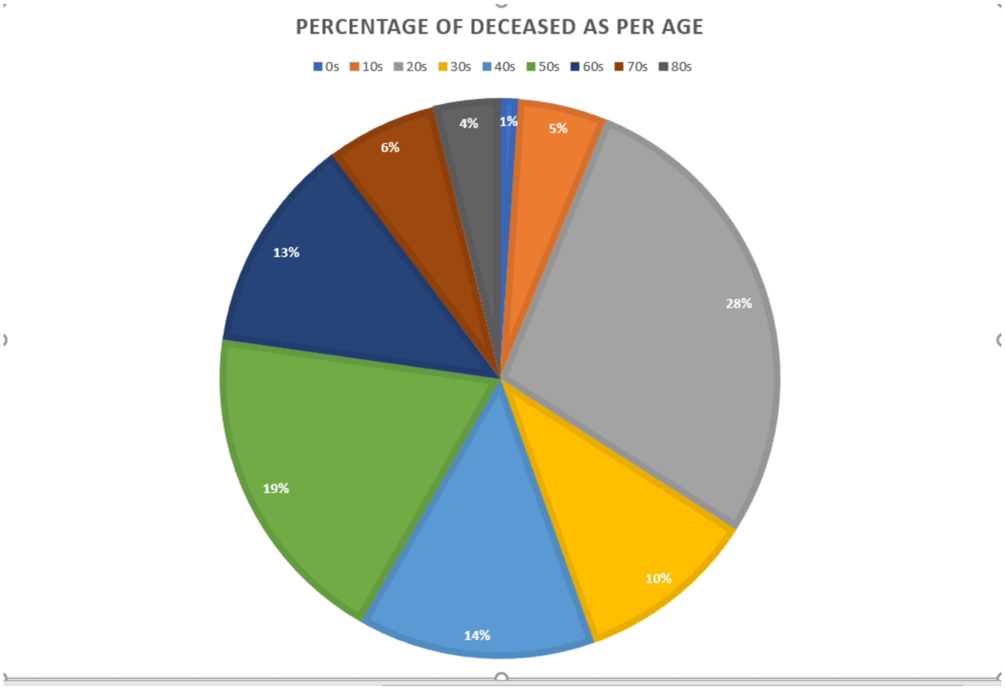
- It can be inferred that the age group which is most affected by this virus are in the range of (20 - 30) followed by people that fall in the age group (50-60) and (40-50).
- People which are least affected are children that are in range (0-20).

Box Plot representation:



DESCRIPTION: The above box plot represents the percentage of confirmed cases for each age group.
INFERENCE: It can be inferred that the age group which is most affected by this virus are in the range of (20 - 30) followed by people that fall in the age group (50-60) and (40-50). People which are least affected are children that are in range (0-20).

3. Percentage of Deceased patients in accordance to their age group

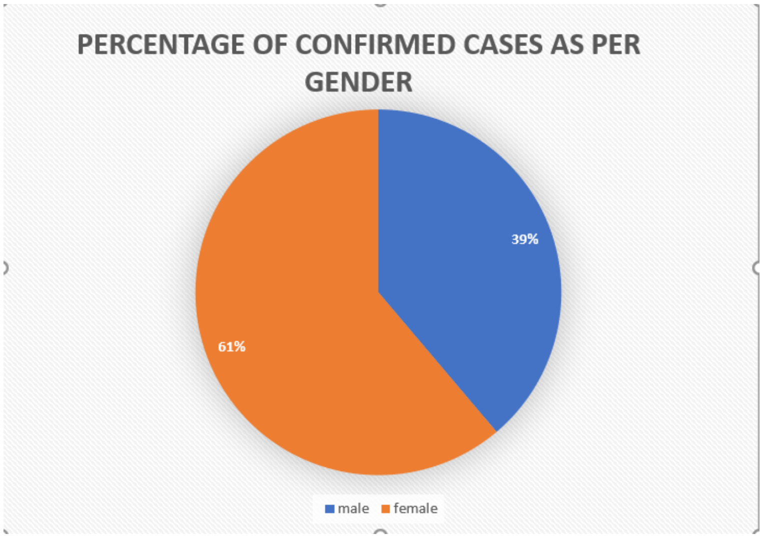


DESCRIPTION: The above pie chart represents percentage of Deceased for each age group

INFERENCE:

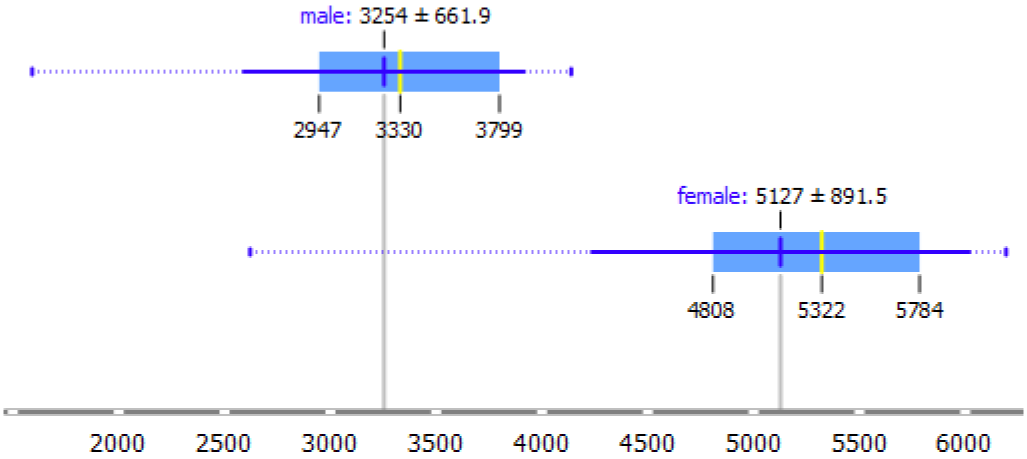
- It can be inferred that the virus is affecting the most to people whose age falls in the interval of (20-30) followed by the people that fall in the age group (50-60).
- People which are least affected are children that are in range (0-20).

4. Gender wise distribution of the Confirmed Cases



DESCRIPTION: The above pie chart represents the percentage of confirmed cases for Male and Female.
INFERENCE: We infer from this that Gender which is more affected are female although there is not much difference between the percentages but we can say Females are little bit affected than males.

Box Plot Representation

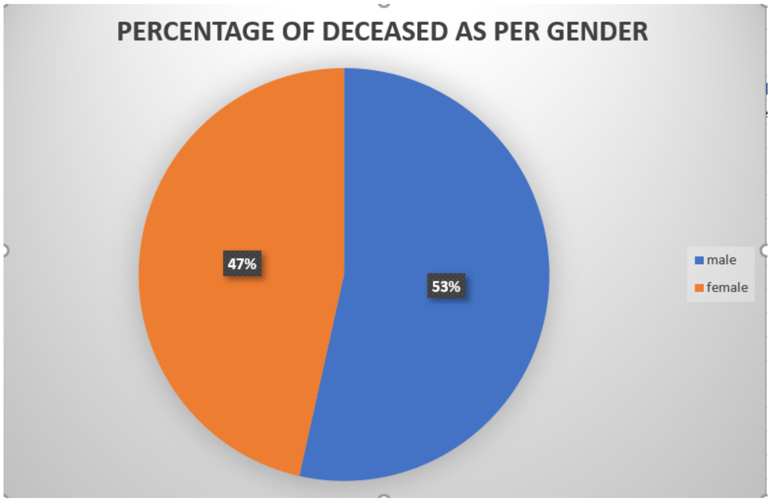


The Box plot represent the spread of data and is represented by five values

1) The blue end points are the minimum and maximum values
2) The end points of blue shaded region are first quartile and third quartile respectively.
3) The middle yellow line represents the median.

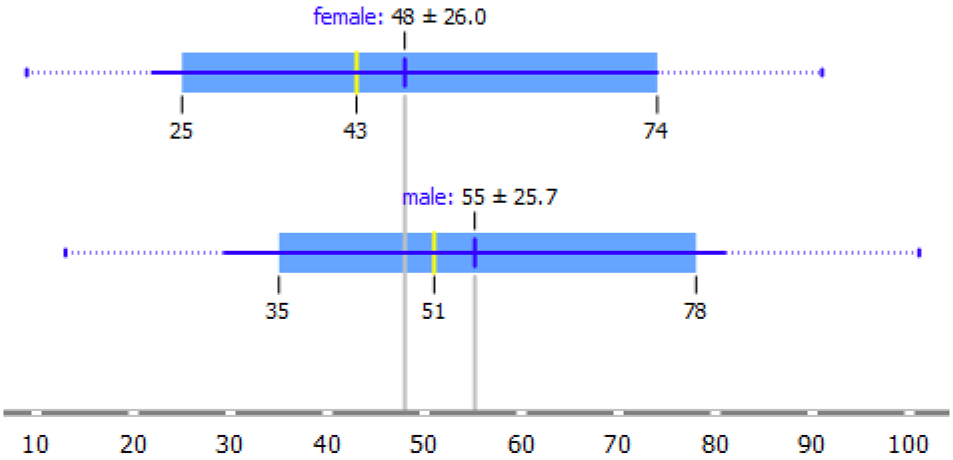
DESCRIPTION: The above box plots are about the number of confirmed cases for each gender.
INFERENCE: From the median of each gender , we can infer that Gender which is more affected are female.

5. Gender Distribution of Deceased Patients



DESCRIPTION: The above pie chart represents the percentage of Deceased for Male and Female.

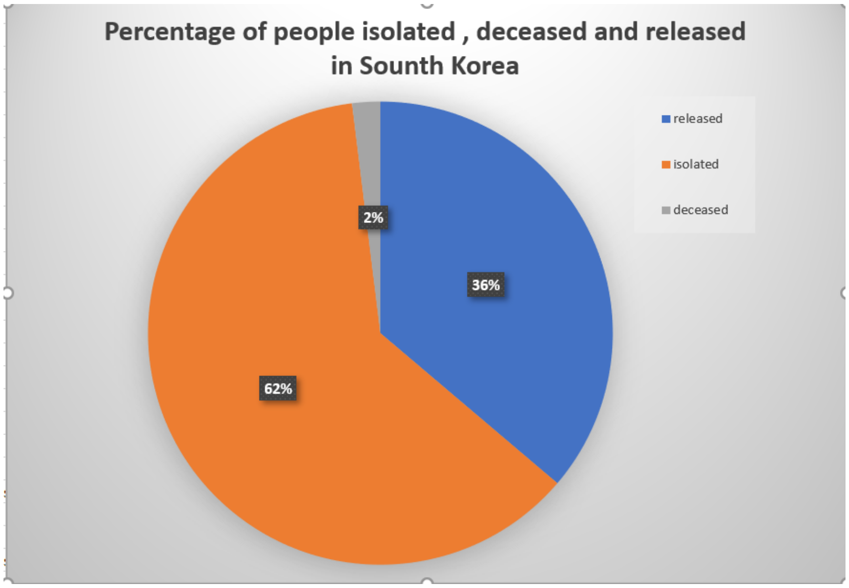
INFERENCE: We infer from this that it is equally affected both male and female Gender



DESCRIPTION: The above box plots are about the number of Deceased cases for each gender.

INFERENCE: From the median of each gender , we can infer that Both genders are equally affected and has spread almost same in both.

6. Isolated,Released after Recovery and Deaths Distribution



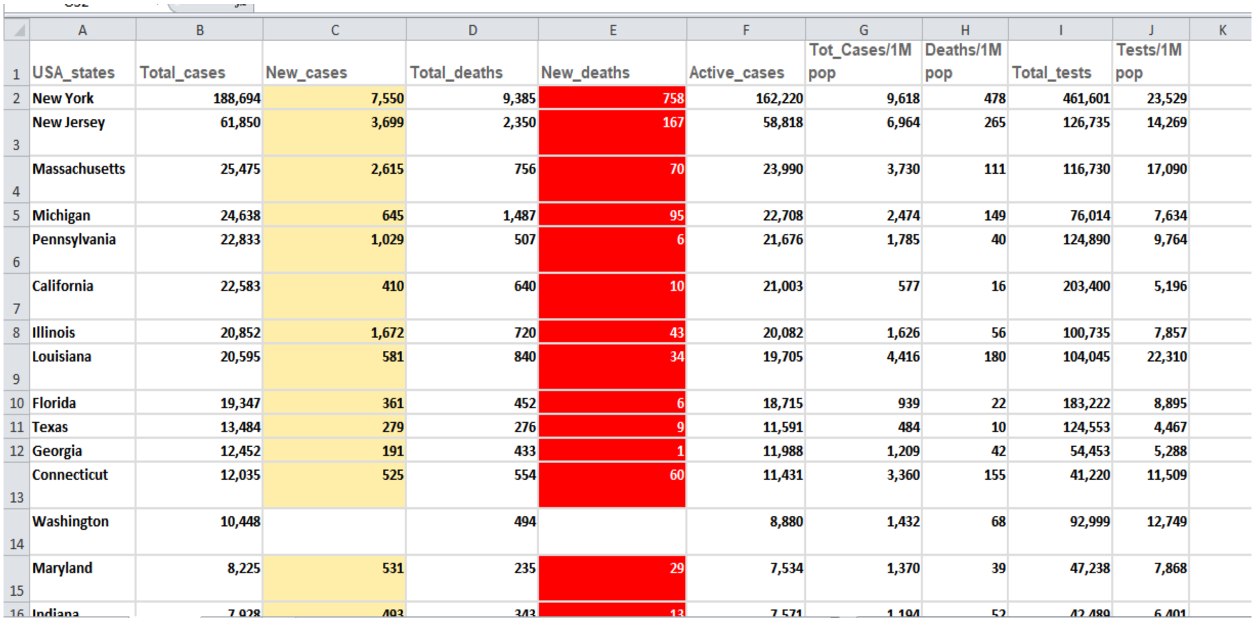
DESCRIPTION: The above pie chart represents the percentage of people who are isolated , deceased and released as per 7th april.

INFERENCE: It can be inferred that more than half of the people have been isolated and less than half of the people are released and a minor segment of those people have died as per data till 7th april.

USA

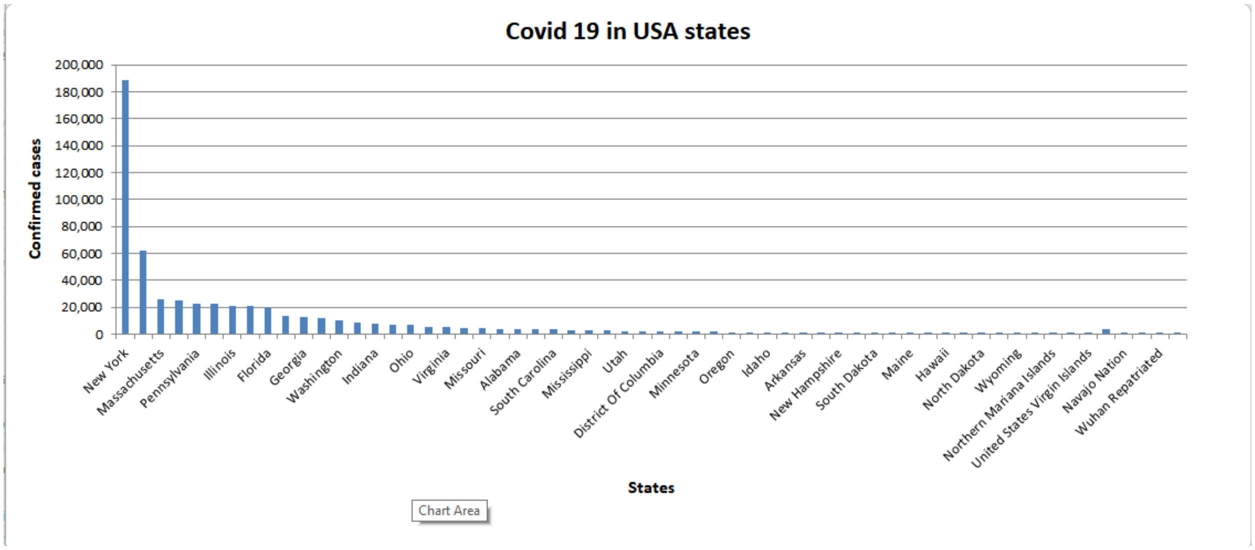
As of now, USA is facing the most severe impact of COVID 19.
It has recorded the maximum number of cases over the world and cases are growing at an exponential rate.

Dataset:



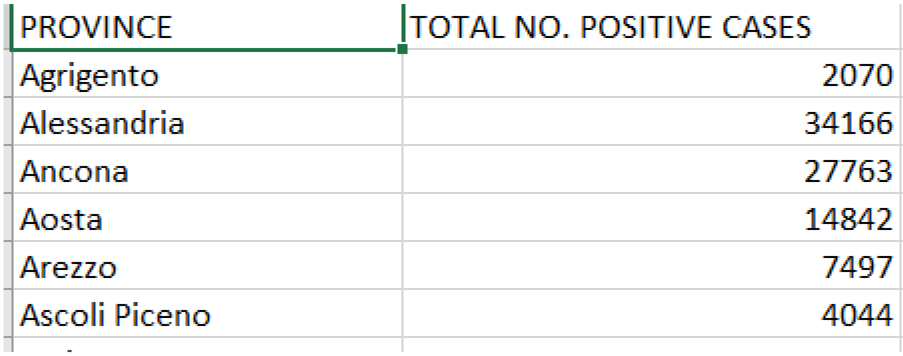
	A	B	C	D	E	F	G	H	I	J	K
1	USA_states	Total_cases	New_cases	Total_deaths	New_deaths	Active_cases	Tot_Cases/1M pop	Deaths/1M pop	Total_tests	Tests/1M pop	
2	New York	188,694	7,550	9,385	758	162,220	9,618	478	461,601	23,529	
3	New Jersey	61,850	3,699	2,350	167	58,818	6,964	265	126,735	14,269	
4	Massachusetts	25,475	2,615	756	70	23,990	3,730	111	116,730	17,090	
5	Michigan	24,638	645	1,487	95	22,708	2,474	149	76,014	7,634	
6	Pennsylvania	22,833	1,029	507	6	21,676	1,785	40	124,890	9,764	
7	California	22,583	410	640	10	21,003	577	16	203,400	5,196	
8	Illinois	20,852	1,672	720	43	20,082	1,626	56	100,735	7,857	
9	Louisiana	20,595	581	840	34	19,705	4,416	180	104,045	22,310	
10	Florida	19,347	361	452	6	18,715	939	22	183,222	8,895	
11	Texas	13,484	279	276	9	11,591	484	10	124,553	4,467	
12	Georgia	12,452	191	433	1	11,988	1,209	42	54,453	5,288	
13	Connecticut	12,035	525	554	60	11,431	3,360	155	41,220	11,509	
14	Washington	10,448		494		8,880	1,432	68	92,999	12,749	
15	Maryland	8,225	531	235	29	7,534	1,370	39	47,238	7,868	
16	Indiana	7,928	493	343	13	7,571	1,194	52	42,489	6,401	

Distribution of COVID cases among the US States



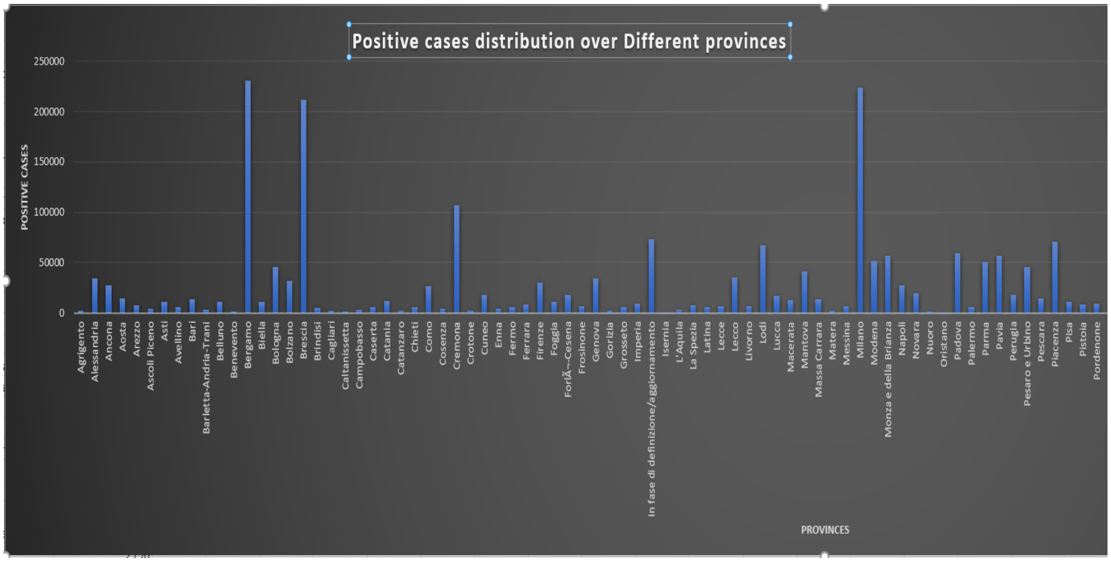
ITALY

DATASET



PROVINCE	TOTAL NO. POSITIVE CASES
Agrigento	2070
Alessandria	34166
Ancona	27763
Aosta	14842
Arezzo	7497
Ascoli Piceno	4044

ANALYSIS



DESCRIPTION: This bar graph shows the number of positive cases for each province.
INFERENCE: We can observe most of the positive cases are detected in the provinces of Bergamo , Brescia, Milano.

REFERENCES

https://drive.google.com/open?id=11RPI6U70MqjGPjxNLNk1MXznojOknn2W