STUDY OF CONFIRM CASES AND AGE DISTRIBUTION OF FATAL CASES AND A COMPERATIVE STUDY ON THE EFFECT OF TESTING OF COVID-19 PATIENTS IN USA, ITALY AND SOUTH KOREA TILL 28TH MARCH,2020 ALONG WITH PREDICTION OF CONFIRM CASES IN USA TILL 04th APRIL,2020.

PREPARED BY

Amit Agarwal – C20003

Debajyoti Podder – C20011

Mohit Kumar Saini – C20015

Shashank Shekhar – C20024

Sourav Maji – C20026

UNDER THE GUIDANCE OF

Dr. Sayantani Roy Choudhury

Praxis Business School, Kolkata

HIGHLIGHTS:

- COVID-19 caused 104,705 cases and 1707 deaths in United States of America as of March 28, 2020.
- 56 State boundary wise geographical representation of confirm cases has been done.
- The first case of positive COVID-19 was identified on January 21, 2020 in Everett, Washington.
- Till March 28, 2020 New York state had the maximum positive cases of COVID-19 with 52,318 number of patients.
- The fatality case is higher among the aged persons with mean age of 64 years in USA.
- The smaller number of COVID-19 tests caused an alarming situation in USA like ITALY, which is just the opposite scenario for South Korea.
- Prediction of COVID-19 confirm case for USA for one week from March 29, 2020 to April 4, 2020.

OBJECTIVES:

1. Establishing the descriptive statistics of age distribution with respect to number of fatal cases and the correlation between the age and number of deaths.
2. A comparative study between three countries USA, ITALY and SOUTH KOREA on the basis of total number of positive cases with respect to number of COVID-19 test.
3. To predict the number of confirm case for one week with respect to the current trend.

INTRODUCTION:

The novel coronavirus (SARS-CoV-2) outbreak, that emerged out of the city of Wuhan, China in December 2019 has expanded to touch nearly every corner of the globe. It has demonstrated the whole world how fast it can spread following human mobility patterns. The World Health Organization has declared the virus a global health emergency and rated COVID-19's global risk of spread and impact as "very high," the most serious designation the organization gives. The first confirm case to be identified in the World was from China on January 10, 2020.

The symptoms of COVID-19 are very similar to that of flu-like symptoms, ranged from people with mild symptoms to people being severely ill and dying. The symptoms may appear 2-14 days after exposure but among the patients who have died, the time from symptom onset to outcome ranges from 2-8 weeks. Based on all 72,314 cases of COVID-19 confirmed, suspected, and asymptomatic cases in China as of February 11, a paper by the Chinese CCDC released on February 17 and published in the Chinese Journal of Epidemiology has found that:

- **80.9% of infections are mild** (with flu-like symptoms) and **can recover at home**.
- **13.8% are severe**, developing severe diseases including **pneumonia** and **shortness of breath**.
- **4.7% as critical** and can include: **respiratory failure**, **septic shock**, and **multi-organ failure**.
- In about 2% of reported cases the virus is fatal.

Patients who reported no pre-existing ("comorbid") medical conditions had a case fatality rate of 0.9%. Pre-existing illnesses put patients at higher risk of dying from a COVID-19. Nevertheless, a clear picture of the epidemiology of this novel coronavirus is still being elucidated.

The number of cases of COVID-19 in the province of Hubei, the disease epicentre, quickly climbed up following an exponential growth trend. The total number of positive COVID-19 cases is at 81,439 including 3300 deaths in China as of March 28, 2020. Fortunately, by February 15, 2020 the daily number of new reported cases in China started to decline across the country although Hubei Province reported 128 cases on average per day in the week of 2nd March to 8th March, 2020. But by that time the virus had spread across the borders and infected almost 150 countries.

United States of America witnessed their first confirmed COVID-19 case on January 21, 2020 in the state of Everett, Washington. In the subsequent one month only 16 cases of positive COVID-19 were registered, but in the following weeks the number of positive cases started to rise exponentially and on March 28, 2020 registered 18,697 cases alone. The Figure-1 below shows the number of confirmed cases for each state of USA as of March 28, 2020.

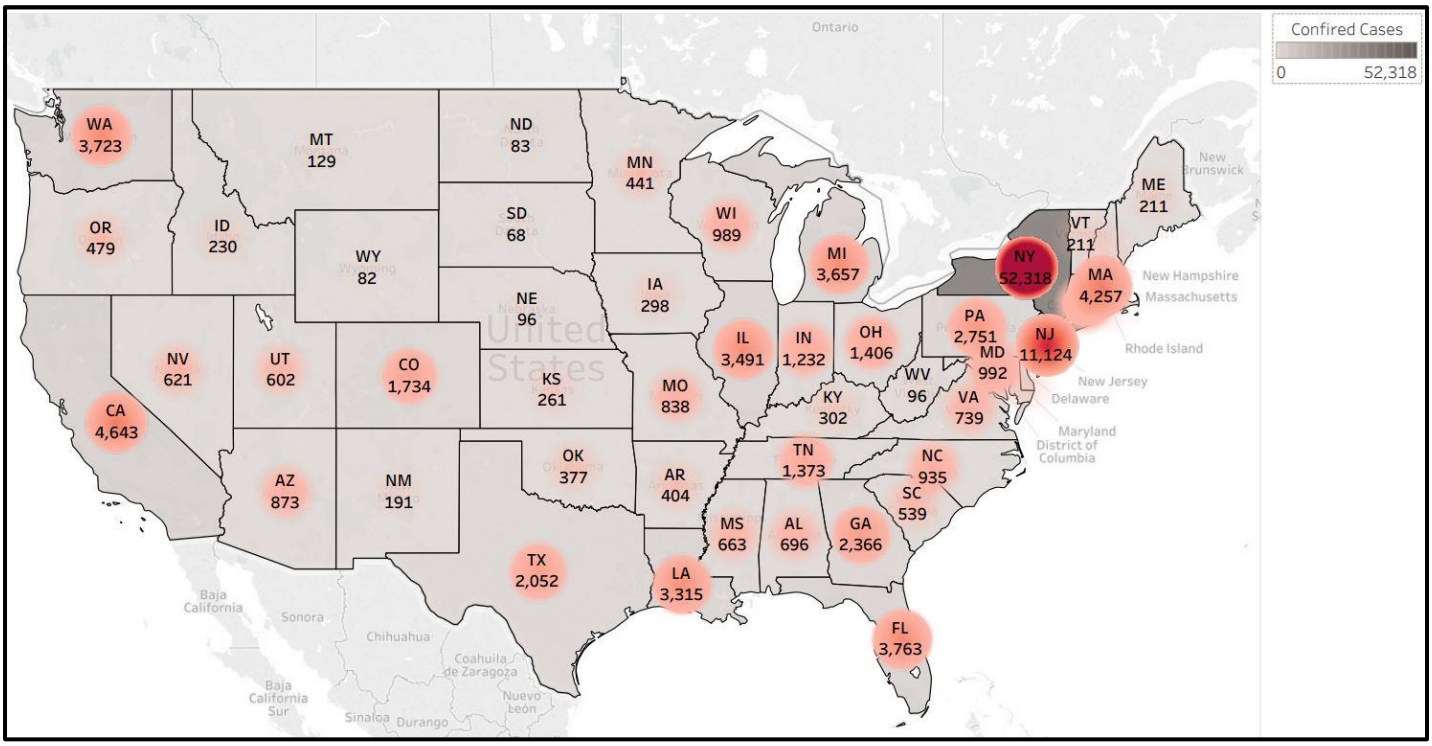


Figure-1

The first case that was being registered, was a man who returned from Wuhan, China. Two days later, most major airlines suspended flights between the US and China and government declared a public health emergency and announced restrictions on travellers arriving from China. The United States got off to a slow start in COVID-19 testing because of some protocol, the kit manufacturer need to follow. USA government also didn't declare complete lockdown to slow down the community spreading. As a result, USA also saw a sudden spike in number of deaths in COVID-19 patients. First death was registered on March 1, 2020 and by March 28, 2020 USA had a total 1707 fatal cases. The Centers for Disease Control and Prevention (CDC) warned that widespread transmission of the disease may force large numbers of people to seek hospitalization and other healthcare, which may overload healthcare systems.

Figure-2 shows the number of confirmed COVID-19 cases day wise.

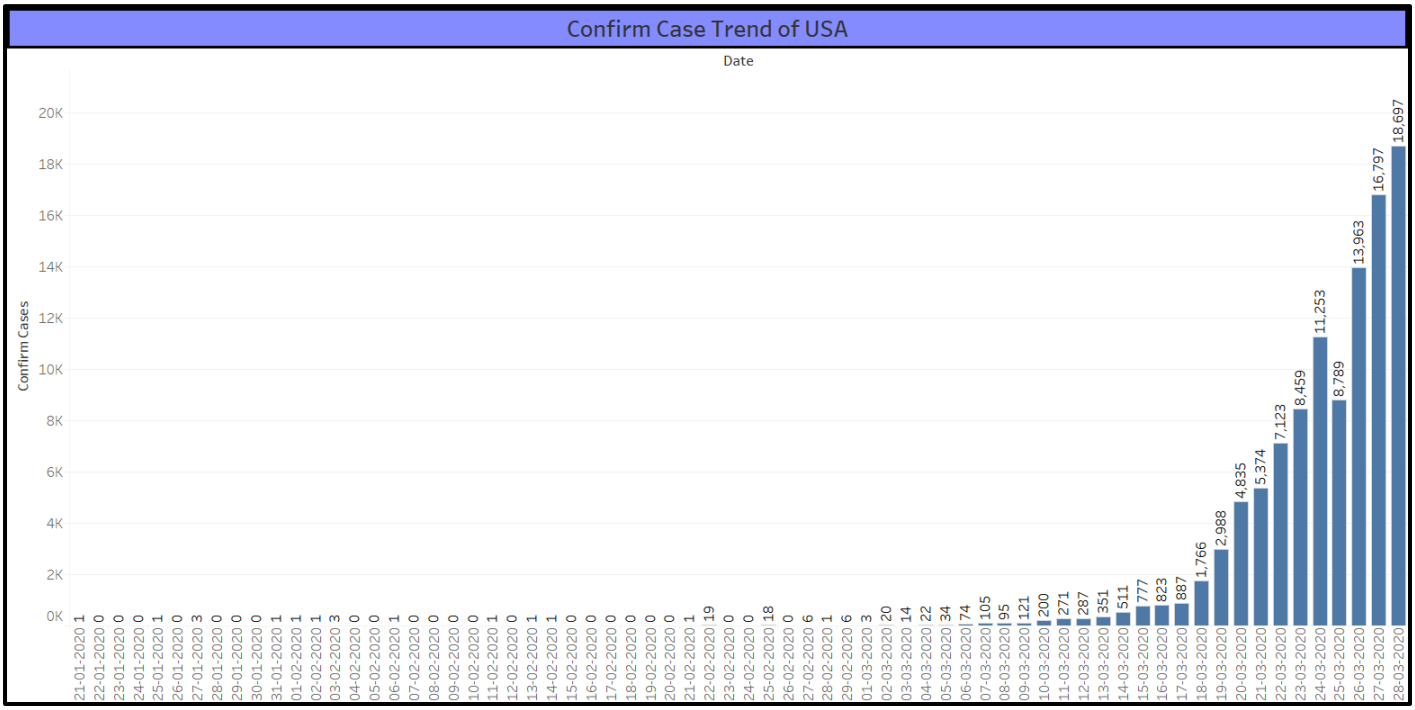


Figure-2

Figure-3 shows the number of deaths from COVID-19 cases day wise.

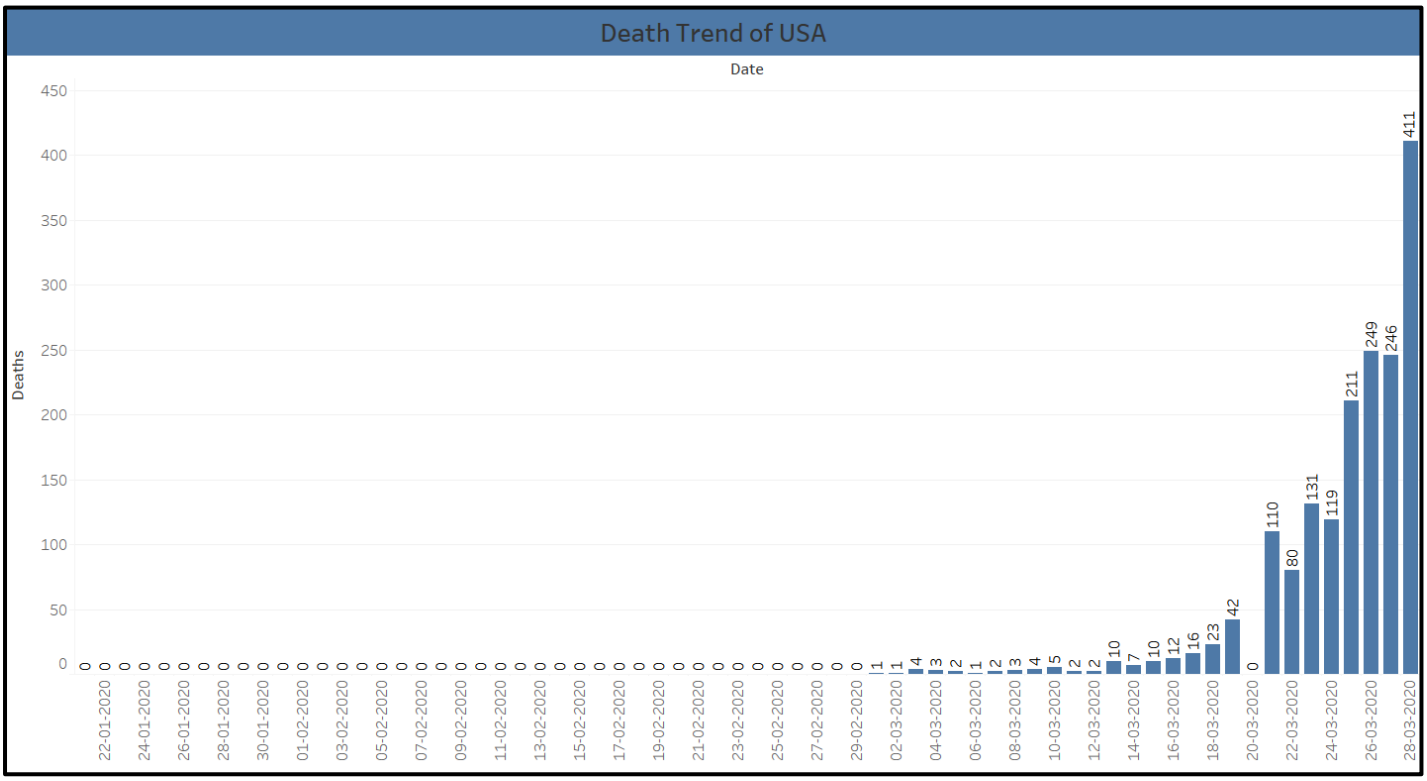


Figure-3

In this report, using a mathematical model parameterized with case series of COVID-19 outbreak in USA, we investigate the age distribution with number of deaths, and the effect of early COVID-19 testing on the growth of number of positive cases.

METHODS:

IN BRIEF

The day wise statistics of COVID-19 positive cases, death cases with age distribution and coronavirus test cases for USA, ITALY and SOUTH KOREA were collected from publicly available sources. By the use of Microsoft Excel, we have calculated the central tendency i.e. Mean, Median and Mode, measure of dispersion Standard Deviation, Quartile Deviation and the Skewness of the age distribution with respect to number of deaths. Used Tableau for the analysis and visualization of the data sets. With the help of Python, we are trying to design a prediction model to predict the number of positive cases over the next one week.

DATA

We have obtained the daily series of confirmed and death cases of COVID-19 along with age bifurcation in USA from the first day of identification i.e. January 21, 2020 to March 28, 2020 that are publicly available in the internet. We have calculated all the central tendency

and correlation with those data sets. We have gathered datasets on the confirm and death cases of both Italy and South Korea for the same period as USA to draw the comparison. But the dataset on number of tests of COVID-19 collected from ourworldindata.org is as of March 20, 2020. All the reference of the datasets used are mentioned at the end of the report.

COMPUTATION OF CENTRAL TENDENCY

The central tendency has been calculated for the continuous age group from 0 to 100 with class interval of 20 years with respect to number of deaths from COVID-19 over the period of January 21, 2020 to March 28, 2020. We are doing so to establish what the mean, median and modal age of patients that are dying of COVID-19. In the dataset we have, there are total of 1707 records of death. But for 13 records the age group is unknown. So, we have left out those 13 records while doing the calculations.

Figure-4 shows here the number of deaths for each age group.

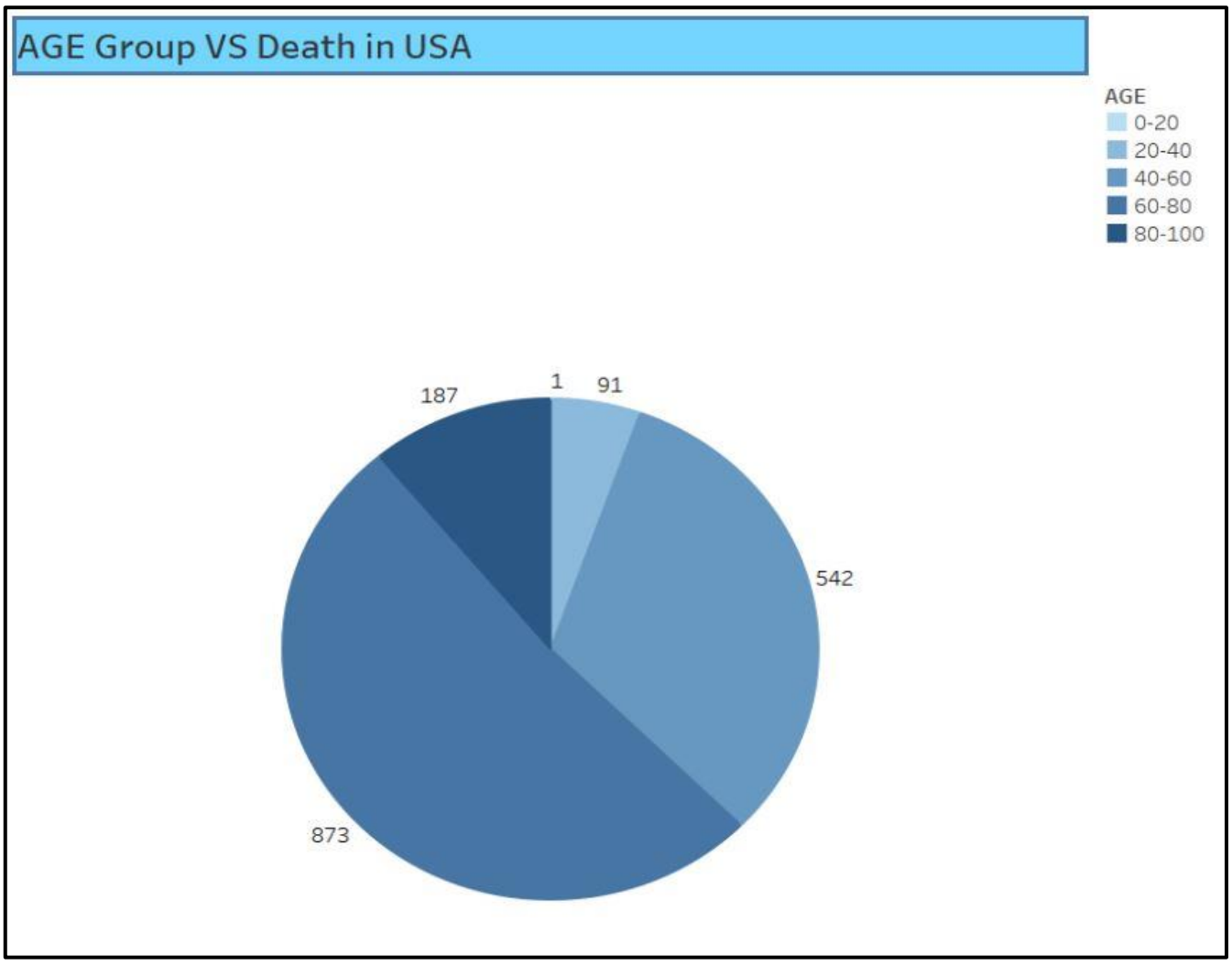


Figure-4

Table-1 shows the frequency distribution of number of deaths with age group.

AGE in Years (Class Interval)	No. of Deaths (Frequency)
0-20	1
20-40	91
40-60	542
60-80	873
80-100	187
Totals	1694

Table-1

Figure-5 shows the Frequency Polygon Graph for Table-1

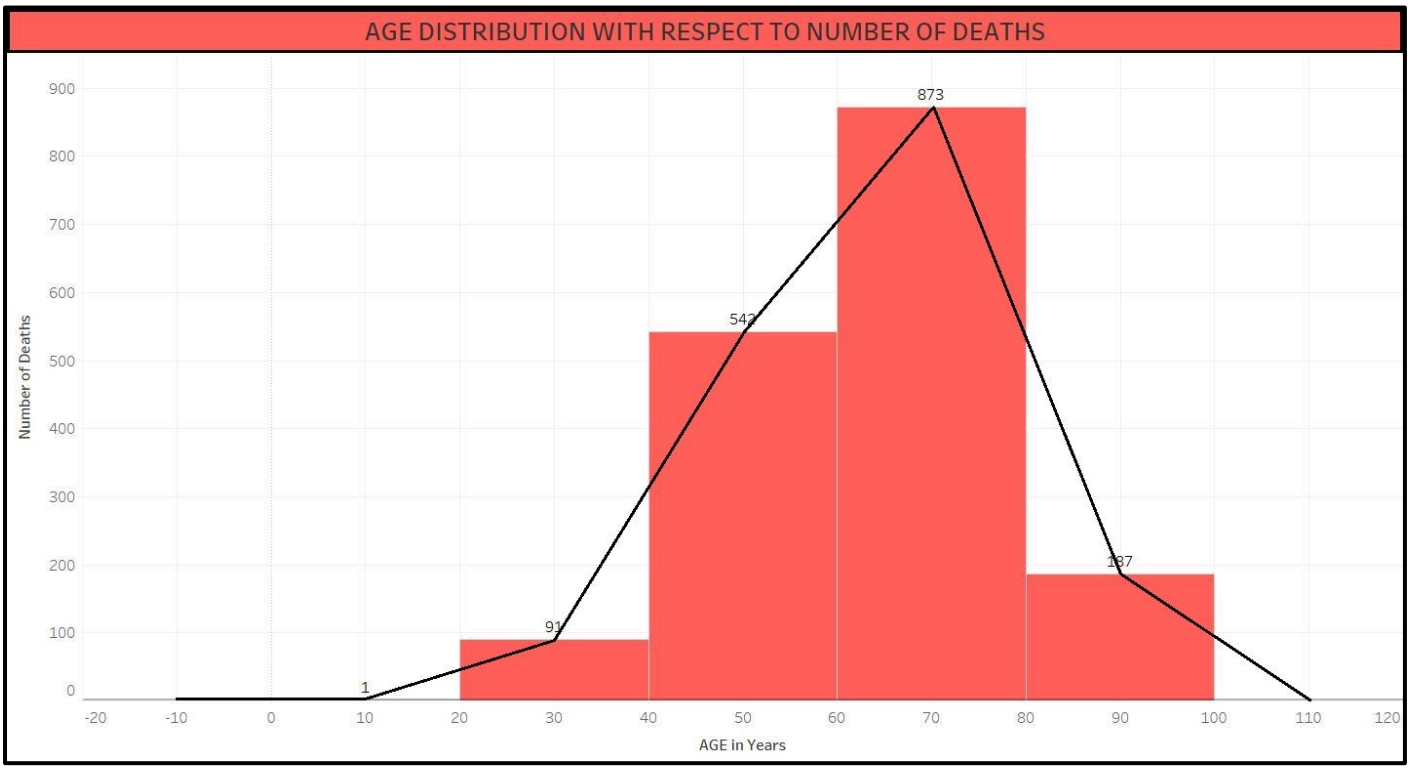


Figure-5

Table-2 shows the calculation table for Mean, Median, Mode, SD, IQR and QD

Class Boundary	Frequency(f)	Mid value(x)	Cumulative Frequency(F)	fx	x^2	fx^2	$(fx^2)/n$	$(fx)/n$
0-20	1	10	1	10	100	100	0.059031877	0.005903188
20-40	91	30	92	2730	900	81900	48.34710744	1.611570248
40-60	542	50	634	27100	2500	1355000	799.8819362	15.99763872
60-80	873	70	1507	61110	4900	4277700	2525.206612	36.07438017
80-100	187	90	1694	16830	8100	1514700	894.1558442	9.935064935
SUM	**1694**			**107780**		**7229400**	**4267.650531**	**63.62455726**

TABLE-2

Mean ($\overline{x}$) age of the patients dying of COVID-19:

$\overline{x} = (\sum f_i x_i) / N = 107780/1694 = 63.62 = 64$ Years (approx.)

Where,

$N = 1694 = \sum f_i$ = Total number of frequencies

Median age of the patients dying of COVID-19:

$L_1 + ((((N/2) - F) * i) / f) = 60 + ((847 - 634) * 20) / 873 = 64.88 = 65$ Years (approx.)

where,

Median Class is the class with Class Boundary 60-80, since N/2 = 847

$L_1 = 60$ = Lower boundary of the median class

F = 634 = Cumulative frequency of the preceding class of the median class

f = 873 = Frequency of the Median Class

i = 20 = Width of the class boundary

Modal age of the patients dying of COVID-19:

$L_1 + (((F_0 - F_1) * i) / (2F_0 - F_1 - F_2)) = 60 + (((873 - 542) * 20) / ((2*873) - 542 - 187)) = 66.51 = 67$ Years (approx.)

where,

Modal Class is the class with Class Boundary 60-80.

$L_1 = 60$ = Lower boundary of the modal class

$F_0 = 873$ = Frequency of the modal class

$F_1 = 542$ = Frequency of the preceding class of the modal class

$F_2 = 187$ = Frequency of the succeeding class of the median class

i = 20 = Width of the class boundary

SKEWNESS OF THE DISTRIBUTION

Figure-6 shows the age distribution plot

Figure-6

By inferring to the results obtained from the computation of central tendency we can say that Mean < Median < Mode. So, the distribution is not a normal one. It is definitely a Negatively Skewed distribution. From the negatively skewed distribution we can conclude that, the distribution has got a greater number of frequencies, in this case number of deaths, on the right of the middle most value of the class intervals i.e. Age. So, it can be observed that death rate is high in aged patients (>65 Years) than younger patients (<65 Years).

MEASURE OF DISPERSION

Standard Deviation (σ)

$$\sigma = \sqrt{\left(\left(\sum f_i x_i^2 / N\right) - \left(\sum f_i x_i / N\right)^2\right)} = \sqrt{(4267.651 - (63.62456)^2)} = 14.82$$

Quartile Deviation (QD)

Positional value of Q_1 = N / 4 = 1694 / 4 = 423.5

Positional value of Q_3 = 3N / 4 = 5082 / 4 = 1270.5

To find Q_1 and Q_3

$(Q_1 – 40) / (60 – 40) = (423.5 – 92) / (634 – 92)$

So, $Q_1 = 52.23$

$(Q_3 – 60) / (80 – 60) = (1270.5 – 634) / (1507 – 634)$

So, $Q_3 = 74.58$

Inter Quartile Range (IQR) = $Q_3 – Q_1$ = 74.58 - 52.23 = 22.35

QD = IQR / 2 = 22.35 / 2 = 11.175

The values of Q3 and Q1 can conclusively say that 50% of the death cases are inside the age group of 75 Years (approx.) and 52 Years (approx.). This statistic also shows us that the number of deaths from COVID-19 are higher in the aged people.

BOX PLOT AND OUTLIERS

Figure-7 shows the box plot.

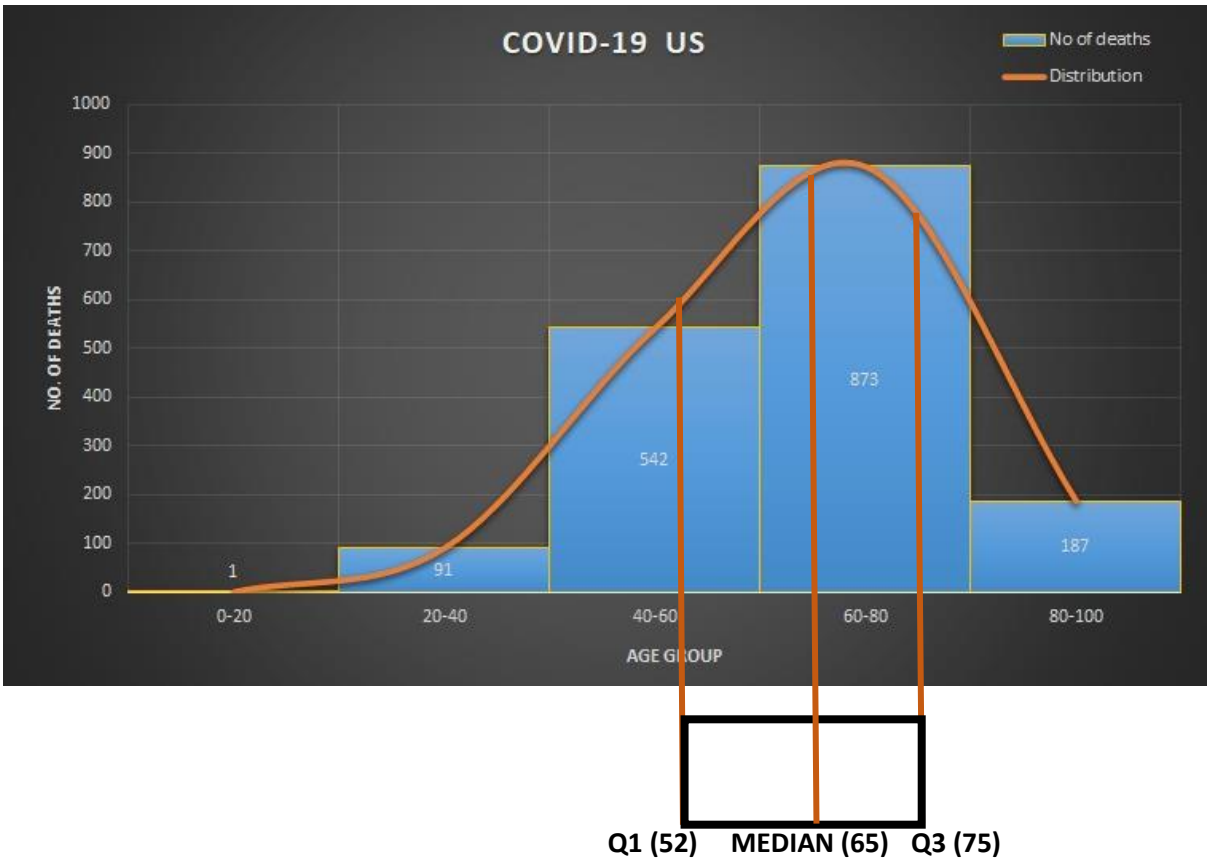


Figure-7

Outliers:

Lowest Value = Q1 - 1.5 * IQR = 52.23 – 1.5 * 11.175 = 35.467 = 36 Years (approx.)

Highest Value = Q3 + 1.5 * IQR = 74.58 + 1.5 * 11.175 = 91.34 = 91 Years (approx.)

Extreme Outliers:

Lowest Value = Q1 - 3 * IQR = 52.23 – 3 * 11.175 = 18.705 = 19 Years (approx.)

Highest Value = Q3 + 1.5 * IQR = 74.58 + 3 * 11.175 = 108.105 = 108 Years (approx.)

So, any data point lying between the Q1 and 36 and between Q3 and 91 will be an outlier. And any data point lying between 36 and 19 and between 91 and 108 will be an extreme outlier.

CORRELATION COEFFICIENT (r)

Table-3 shows the chart between age and number of deaths to calculate the correlation coefficient

Patient Age (Years)	Mid value(x)	No. of Death(y)	xy	y^2	x^2
0-20	10	1	10	1	100
20-40	30	91	2730	8281	900
40-60	50	542	27100	293764	2500
60-80	70	873	61110	762129	4900
80-100	90	187	16830	34969	8100
Total	**250**	**1694**	**107780**	**1099144**	**16500**

Table-3

$$r = Cov(x, y) / (\sigma x * \sigma y)$$

$$= ((\sum xy / n) - ((\sum x * \sum y)/n^2)) / (\sqrt{((\sum x^2/n) - (\sum x/n)^2)} * \sqrt{((\sum y^2/n) - (\sum y/n)^2)})$$

where,

$$Cov(x, y) = Covariance = ((\sum xy / n) - ((\sum x * \sum y)/n^2))$$

$$r = 0.504$$

Since, the value of r is $0.3 < |r| \leq 0.75$ i.e. r = 0.504 and r is positive so, number of death (y) has a positively moderate association with the age of the patient (x). This means that with the increase in age the probability of dying from COVID-19 will also increase.

Comparative study between USA, ITALY and SOUTH KOREA on the basis of total number of positive cases with respect to number of COVID-19 test

Figure-8 shows the confirmed and death case with each day for USA as of March 28, 2020

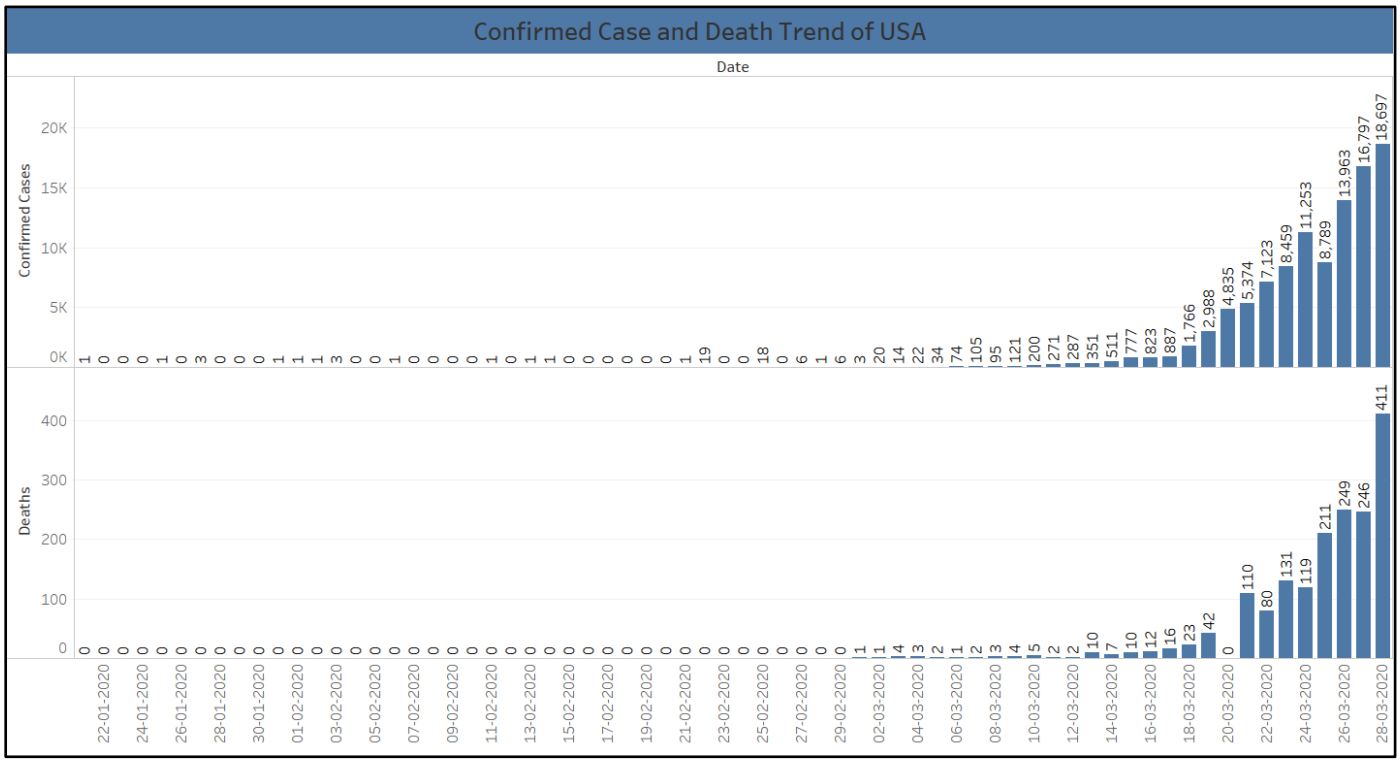


FIGURE-8

Figure-9 shows the confirmed and death case with each day for ITALY as of March 28, 2020

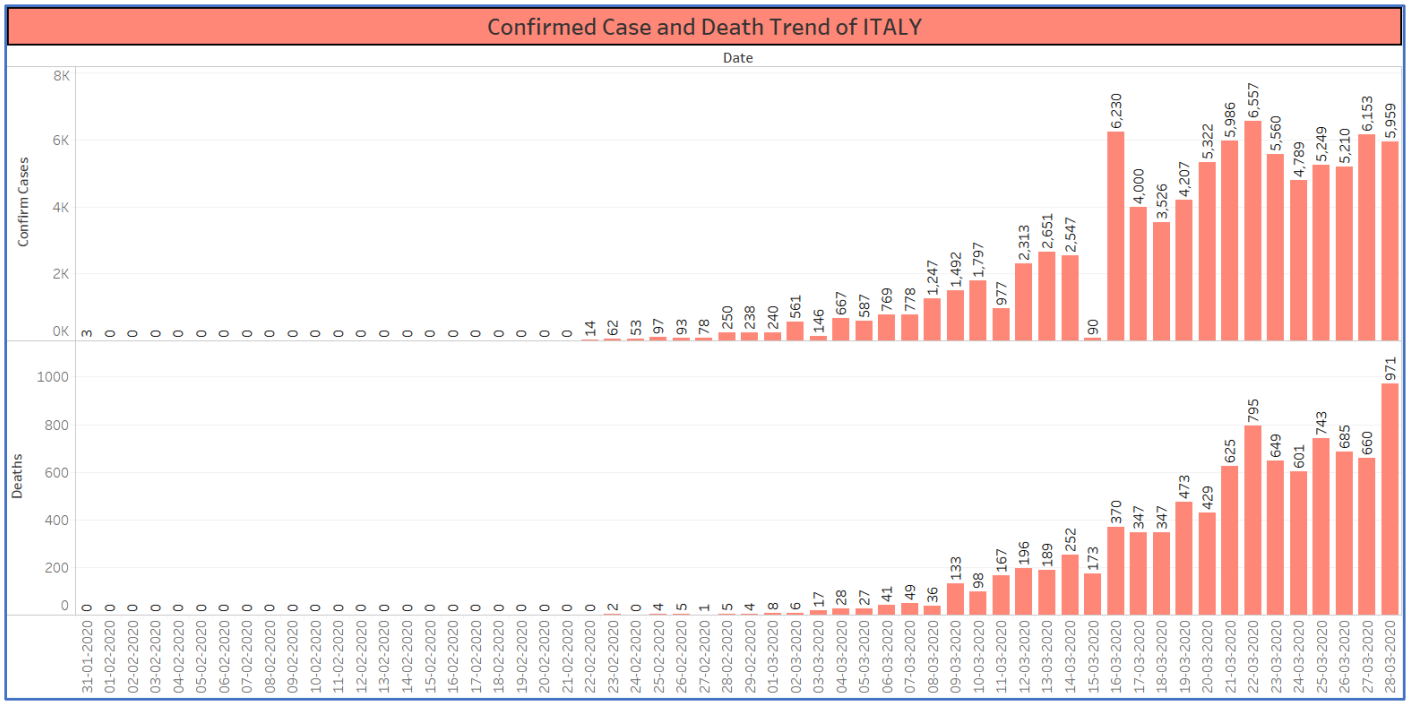


FIGURE-9

Figure-10 shows the confirmed and death case with each day for SOUTH KOREA as of March 28, 2020

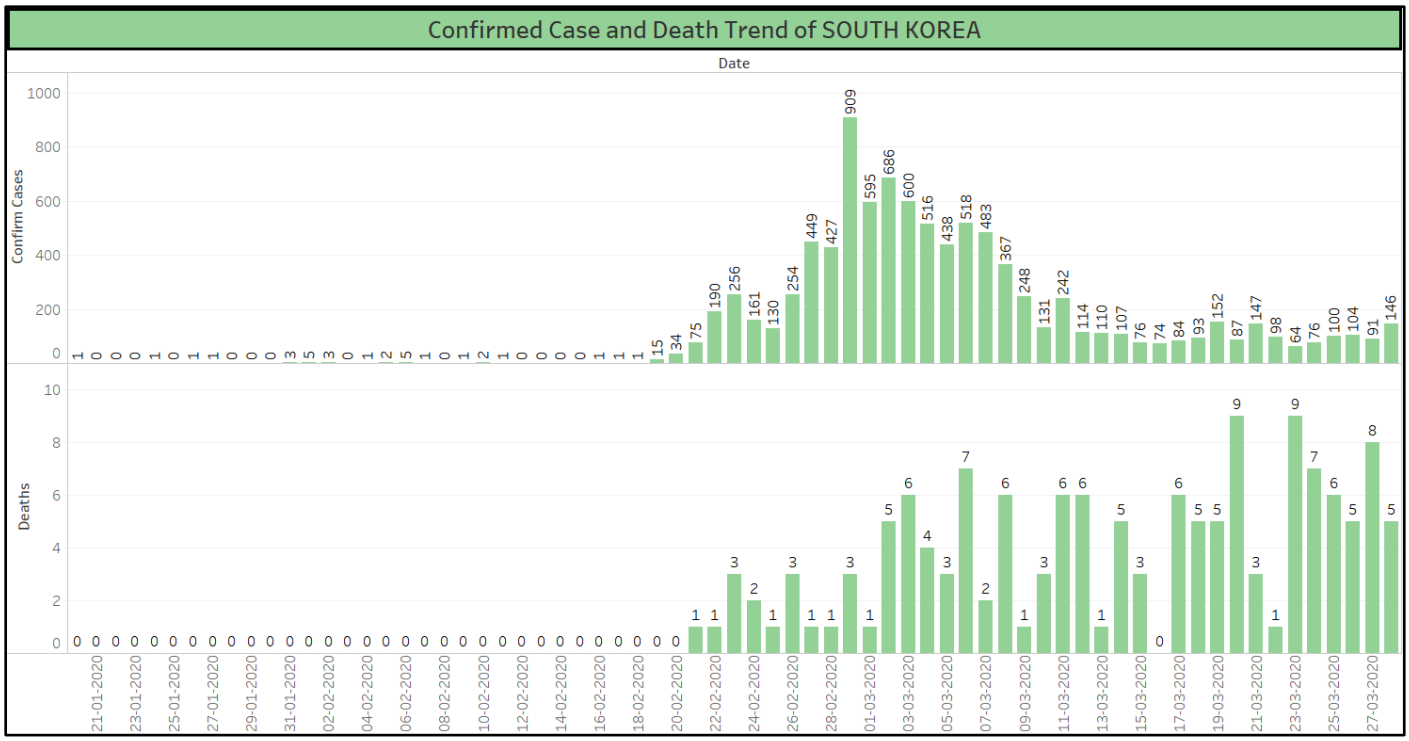


FIGURE-10

The number of confirm cases in USA is increasing in an exponential rate with each day. Just within a span of a month from February 28, 2020 to March 28, 2020 the number of confirmed cases has increased from 60 to 104705.

Another country which is having the same trend is Italy. The number of positive cases in Italy are also increasing almost in the same rate as USA. Italy also in the same span of one-month February 23, 2020 to March 28, 2020 has seen almost the same growth curve for the number of positive cases i.e. from 79 to 86,498.

South Korea is the third country with whom we will be comparing these two afore mentioned countries having just the opposite growth rate of number of positive COVID-19 cases. South Korea is one such country who has successfully contained the number of positive cases. They have set a role model for other countries, for the way they have not only controlled the growth rate but made it to decrease. The day wise confirm cases are gradually decreasing day by day. It was at peak on February 29, 2020 since then it is decreasing with almost each day. The same is true for death cases also for all the three countries.

From the above figures we can verify the same arguments. The figures show us how both the confirmed and death case are getting changed with each day.

There are a lot of factors for which the rate of growth is different in different countries. But one of them is the number of COVID-19 Test conducted by each country.

Figure-11 shows the comparison between USA, ITALY and SOUTH KOREA on the basis of confirmed case, death case as of March 28, 2020 and COVID-19 test conducted as of March 20, 2020.

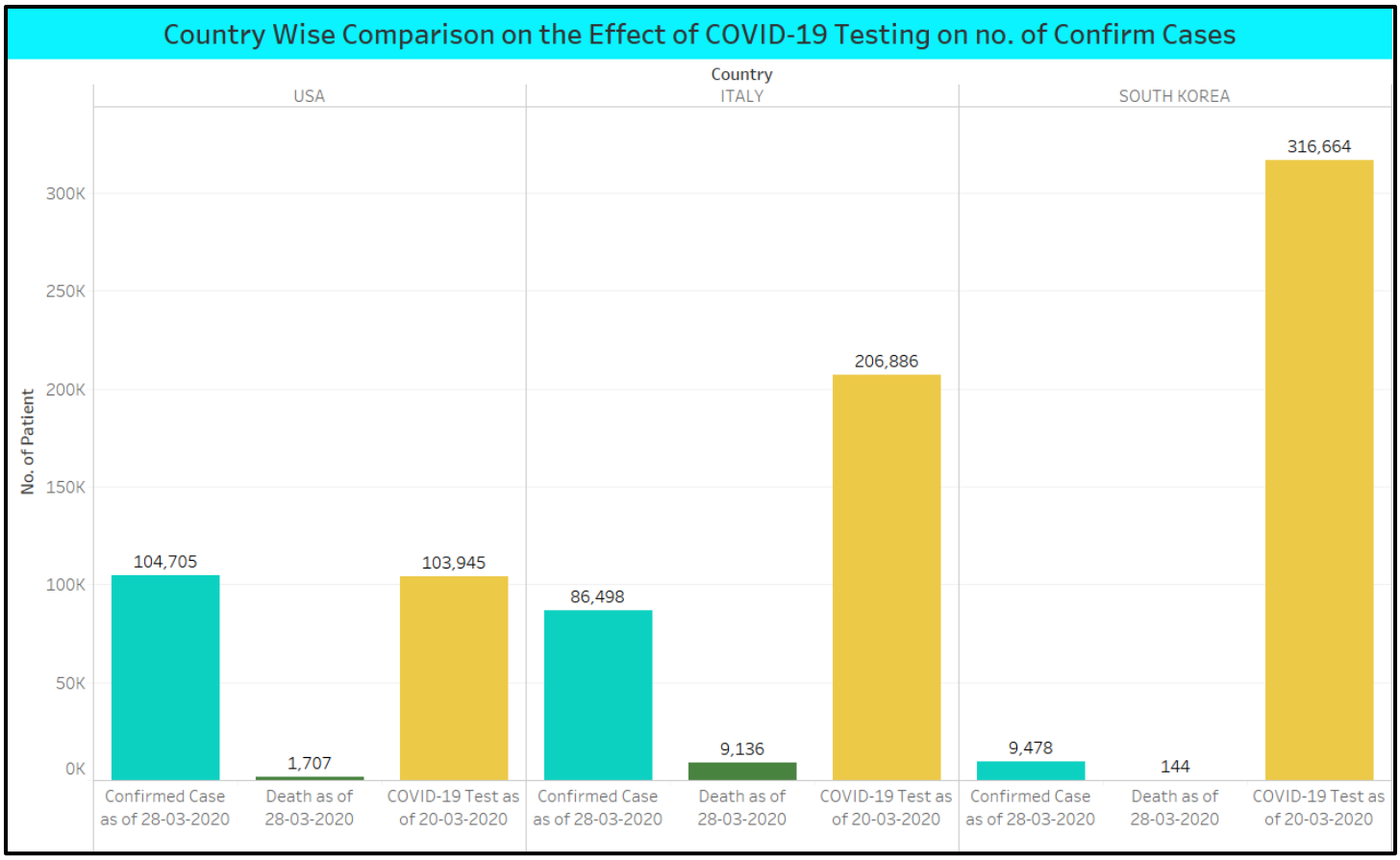


Figure-11

The above chart is clearly representing the relation between the growth of case with number of COVID-19 test. South Korea from almost the very first day has started doing COVID-19 test in random basis. South Korean government even hasn't ordered complete lockdown till date to control the spreading of novel coronavirus, but still they are successfully containing the spread.

Whereas USA started their COVID-19 test pretty late for some protocol issue from the government end. After resolving the issue also, USA government didn't do vigorous COVID-19 testing like South Korea, which in turn became one of the primary reasons for the spreading of novel coronavirus in such magnitude. Italy also lagged behind by much from other countries in respect to number of COVID-19 test, which triggered the novel coronavirus spreading. In case of South Korea, they have quarantined all the patients that were tested positive in the initial days only, which restricted the community spread. But it was not scenario for USA or Italy, so both of them had to face the same fate.

In many cases of novel coronavirus, the patients remain asymptomatic or might have mild symptoms just like common flu. So, without knowing, the patient might transmit the virus in its community. If those patients are tested before hand by the government and tested positive, corrective measures could be taken to quarantine them. In South Korea only there were four such clusters which has transmitted within the community. The largest cluster is

the Shincheonji Church of Jesus. As of March 8, 2020, 4,482 confirmed cases of COVID-19 are linked to this cluster according to KCDC, with the first case (the 31st patient in the country) confirmed on February 18, 2020. This type of transmission can be stopped if early identification can be done.

PREDICTION OF CONFIRM CASES OVER A PERIOD OF ONE WEEK

The trend of daily confirm cases in USA is not a linear one. We have two variables in this model one is the number of confirm cases and the other is the date. There is a non-linear growth in the number of cases with time. The curve is increasing at an increasing rate. So, we don't have the option of predicting the values using a linear regression model, since all the data points will not fit the curve i.e. a straight line.

Figure-12 shows the prediction model using a Linear Regression Model by taking Logarithmic value of the number of cases.

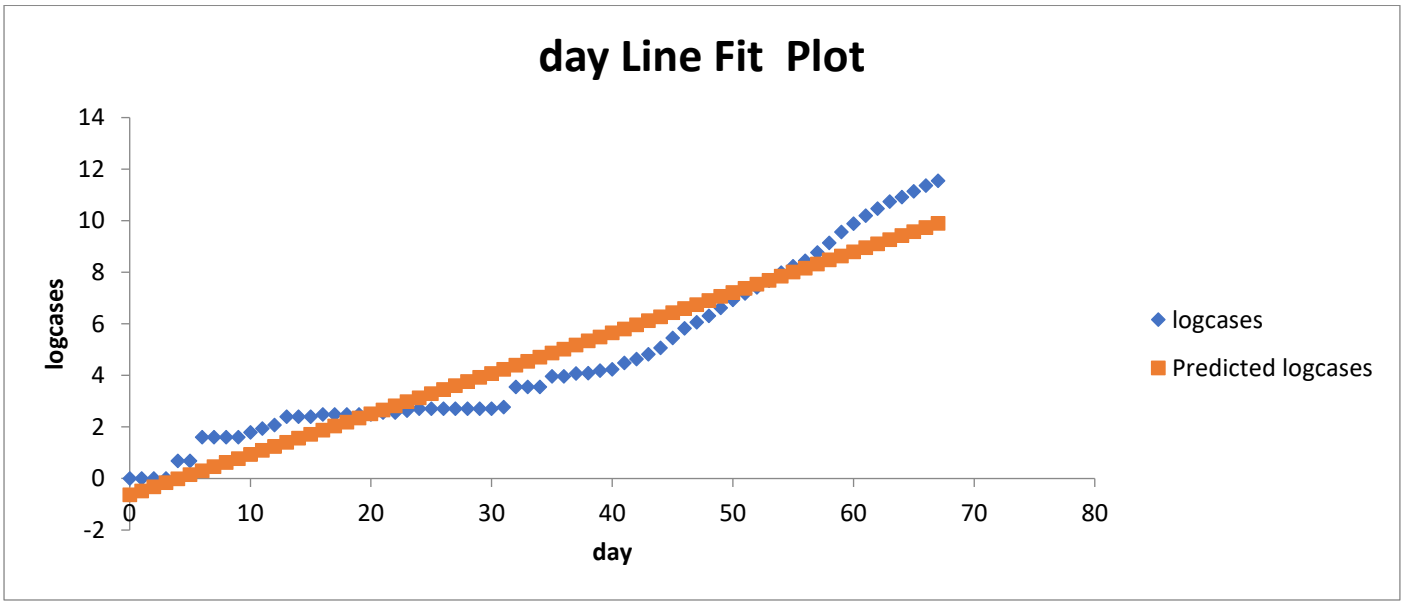


Figure-12

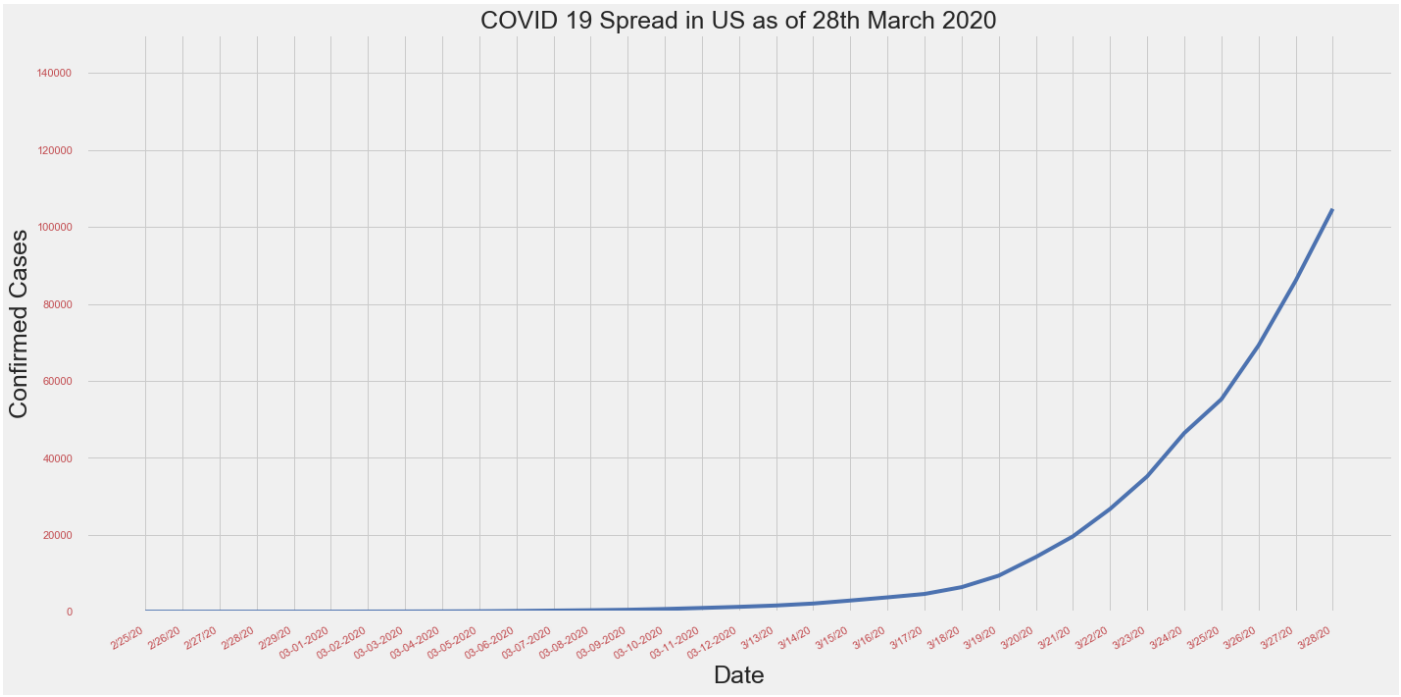
So, depending on the data, we used polynomial regression and try to fit the polynomial equation.

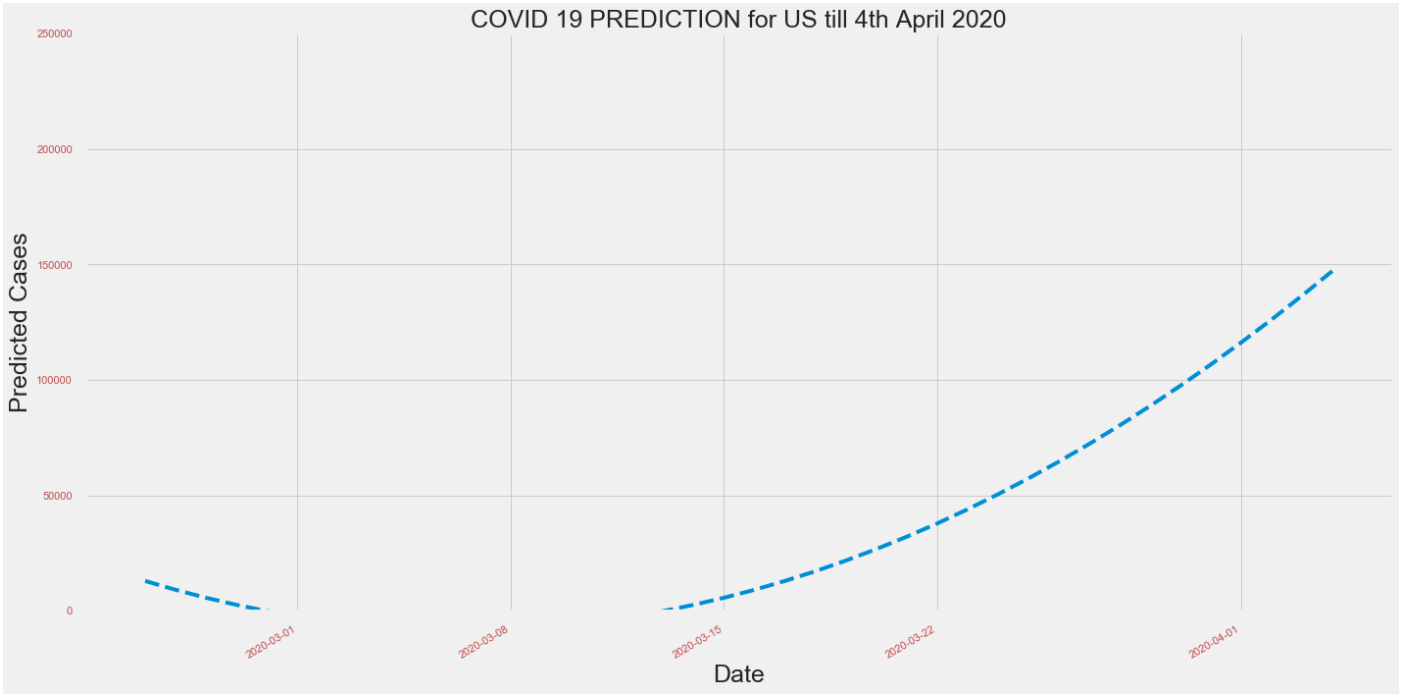
$y = \vartheta_0 + \vartheta_1 x_1 + \vartheta_2 x_1^2 + \vartheta_3 x_1^3$

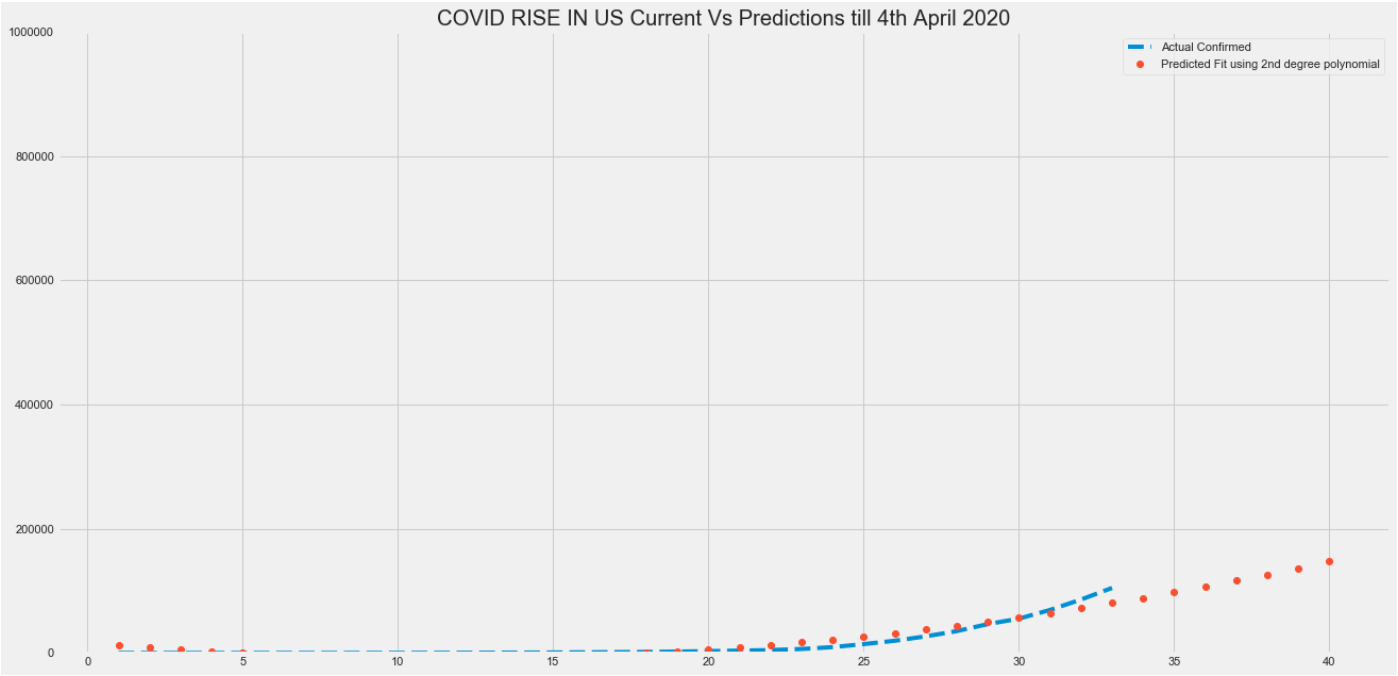
We have used the polynomial regression to fit a polynomial line so that we can achieve a minimum error or minimum cost function. We have tried to build the model with three types of polynomial fit. 2nd Degree, 3rd Degree and 4th Degree Polynomial Regression. But got the best possible fit or prediction with the 3rd Degree Polynomial Regression.

The error in the 2nd Degree and the 4th Degree was more than that we got in 3rd Degree Polynomial Fit. 2nd Degree polynomial model was not showing that much growth and is much more linear than the actual scenario.

Figure-13 shows the Prediction model with 2nd Degree Polynomial Fit





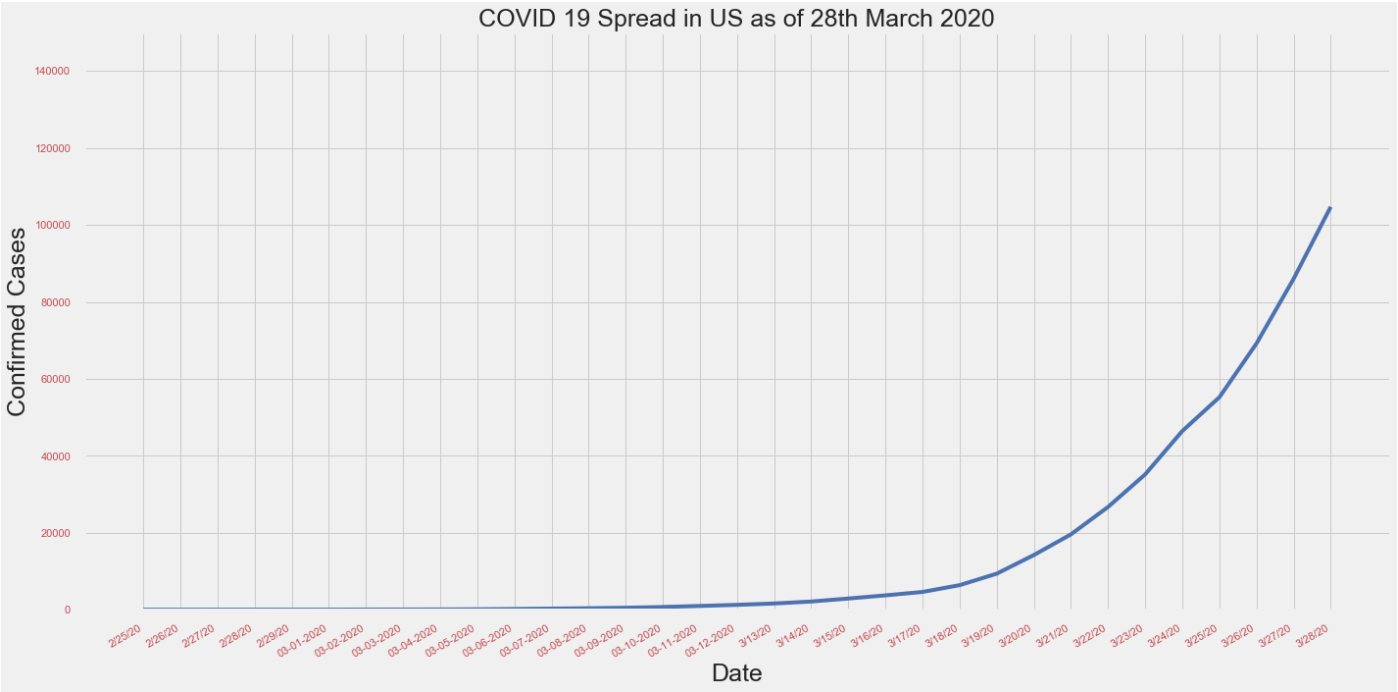


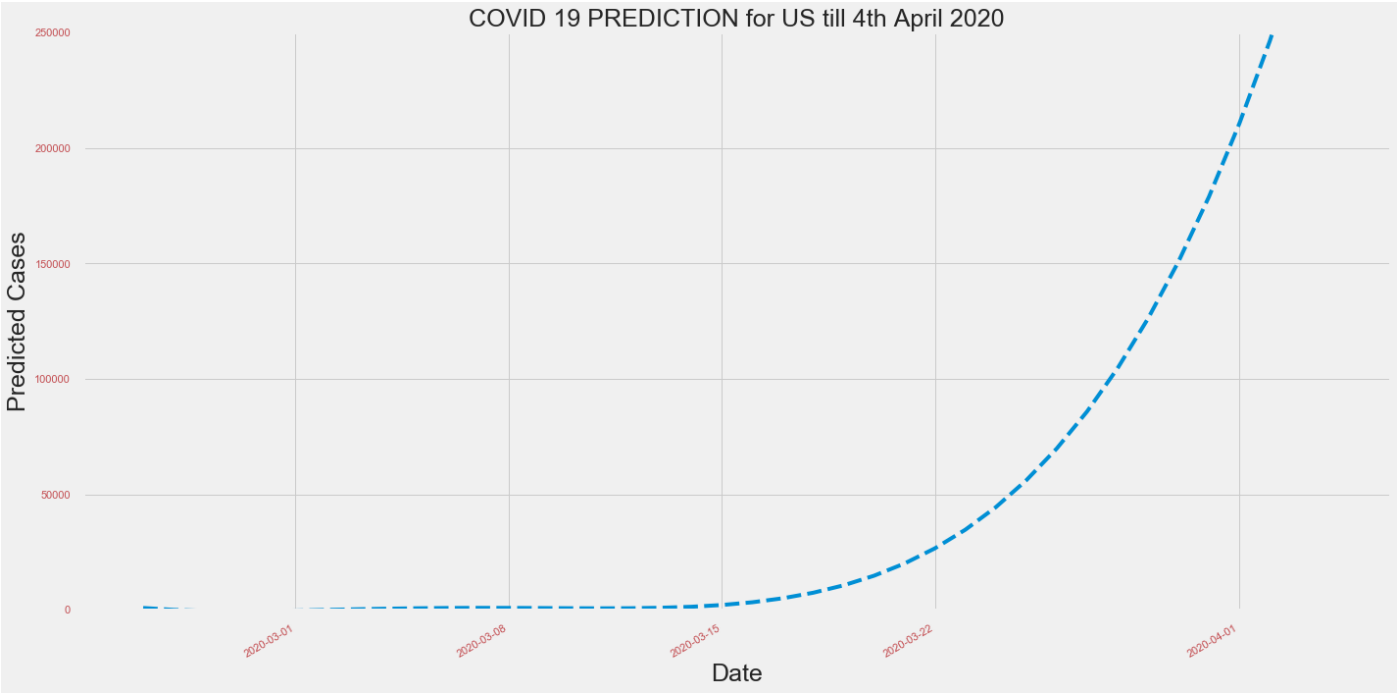
DATE	Predicted	Original	Difference=Predicted-Original
29-03-2020	88708	123578	-34870
30-03-2020	97513	143491	-45978
31-03-2020	106700	163788	-57088
01-04-2020	116271	188530	-72259
02-04-2020	126224	215003	-88779
03-04-2020	136560	244877	-108317
04-04-2020	147279	277161	-129882

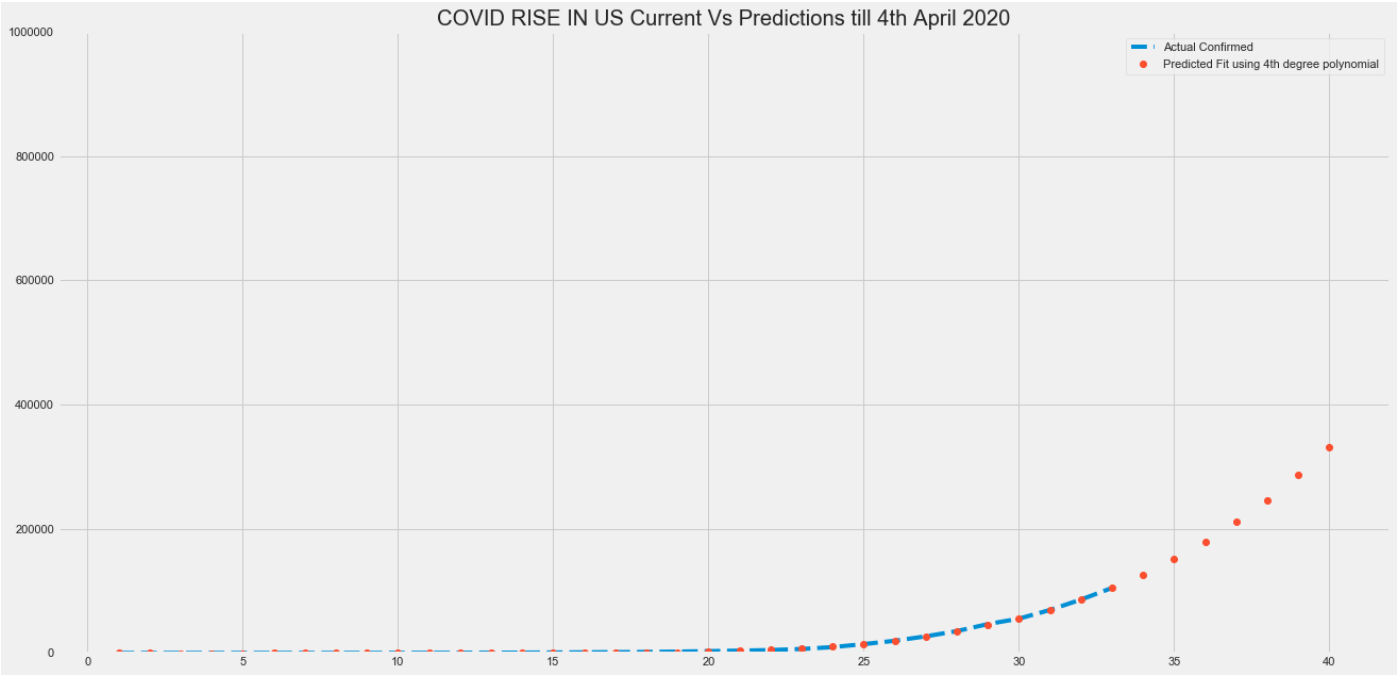
Figure-13

4th Degree polynomial model was having high fluctuation or the increasing rate is too high.

Figure-14 shows the Prediction model with 4th Degree Polynomial Fit





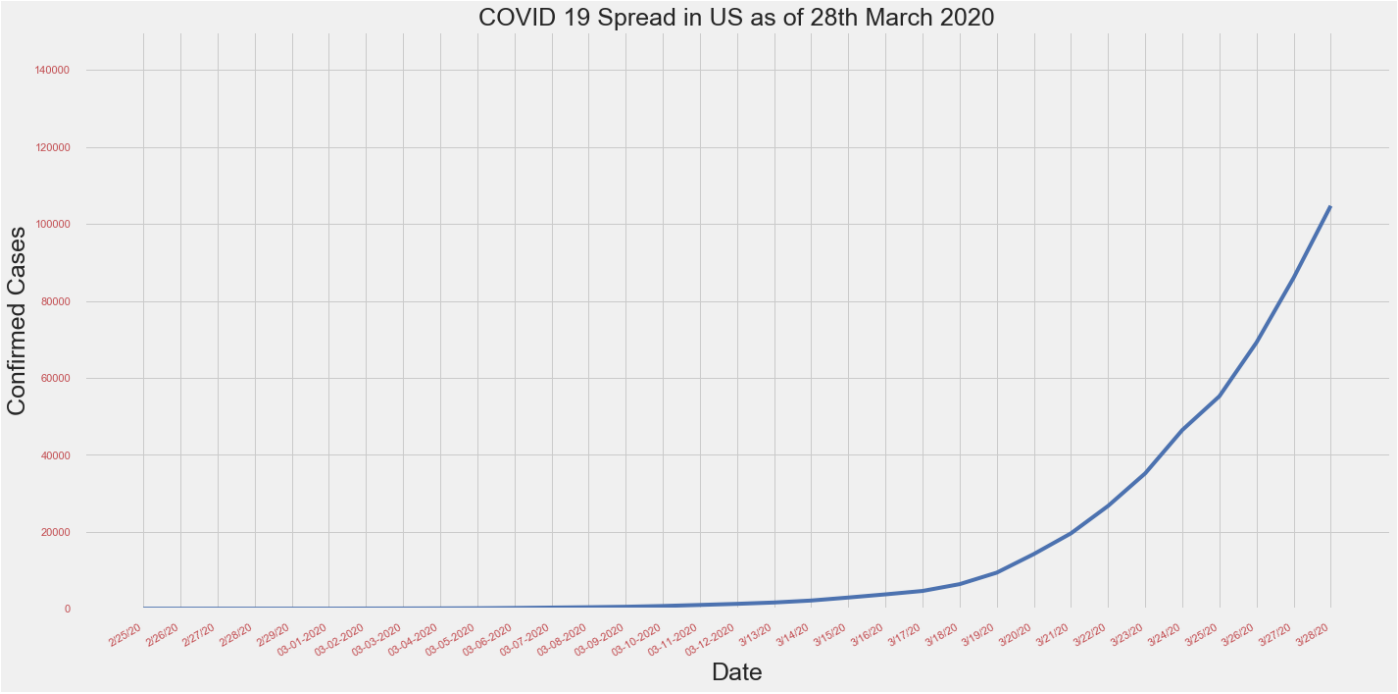


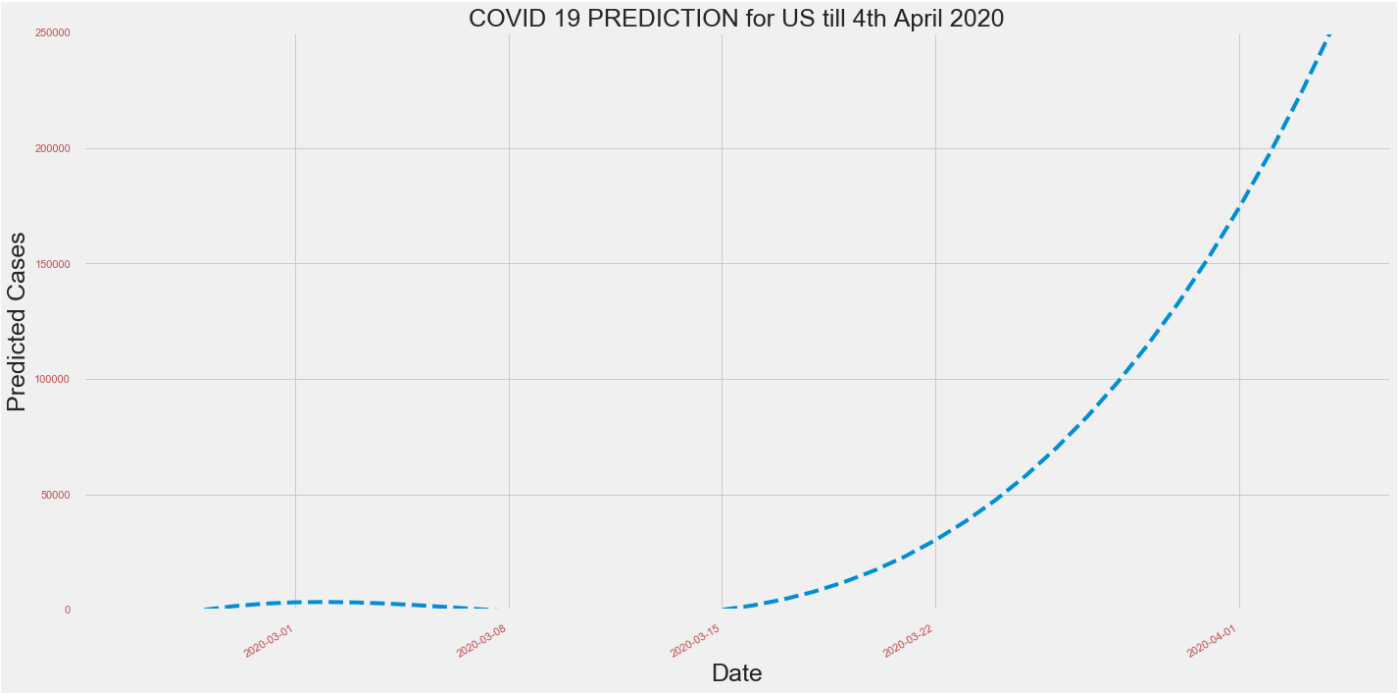
DATE	Predicted	Original	Difference=Predicted-Original
29-03-2020	126136	123578	2558
30-03-2020	150785	143491	7294
31-03-2020	178833	163788	15045
01-04-2020	210573	188530	22043
02-04-2020	246311	215003	31308
03-04-2020	286362	244877	41485
04-04-2020	331056	277161	53895

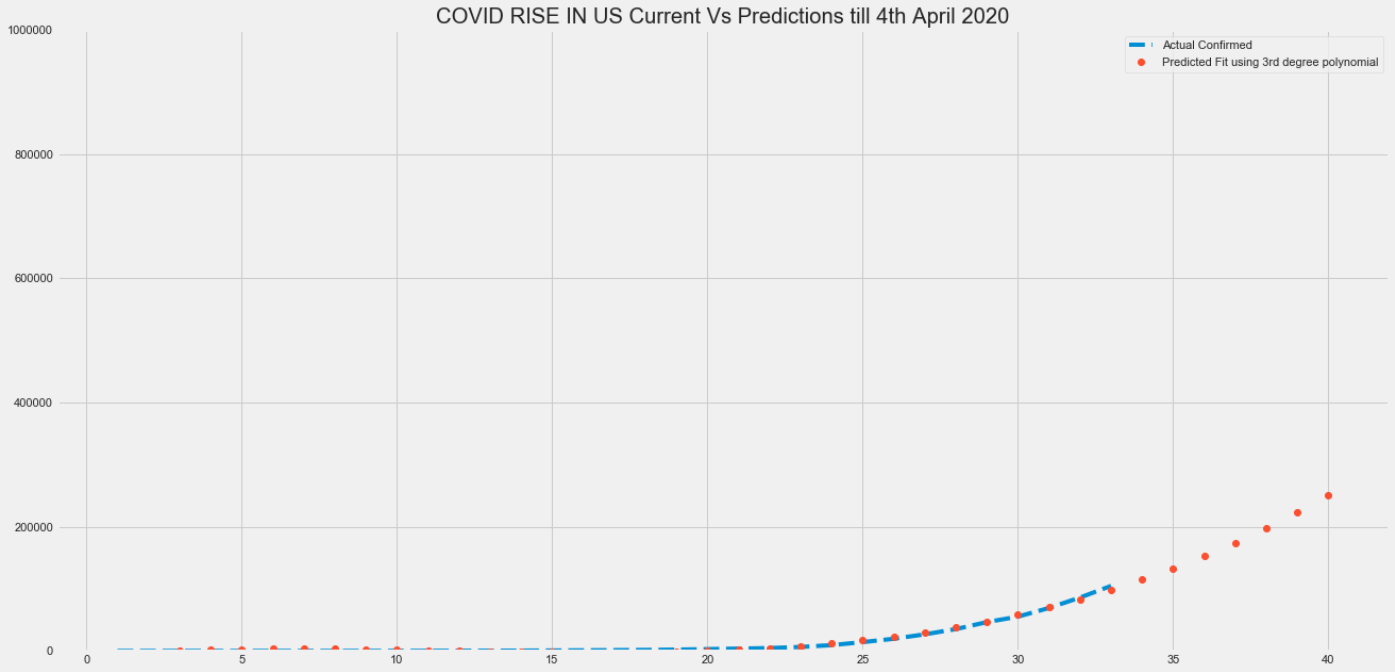
Figure-14

But a 3rd Degree polynomial model learned and adjusted the coefficients to give the best fit for the data points.

Figure-15 shows the Prediction model with 3rd Degree Polynomial Fit







DATE	Predicted	Original	Difference=Predicted-Original
29-03-2020	114938	123578	8640
30-03-2020	133000	143491	10491
31-03-2020	152768	163788	11020
01-04-2020	174314	188530	14216
02-04-2020	197713	215003	17290
03-04-2020	223038	244877	21839
04-04-2020	250361	277161	26800

Figure-15

TOOLS USED:

1.TABLEAU

2.PYTHON 3.8

3.MICROSOFT EXCEL

4.MICROSOFT WORD

CONCLUSION:

Since the outbreak of Novel Coronavirus (COVID-19) in USA on January 21, 2020, the number of cases is increasing rapidly with each day, resulting in 104,705 positive cases including 1707 fatal cases as of March 28, 2020. We have observed that particularly the aged people are more vulnerable to this disease as the mean age for death cases are 64, median age is 65 and the modal age is 67. The age distribution is negatively skewed with a

Standard Deviation of 14.82 and Quartile Deviation of 11.17. And 50% of the death cases patients are in the age range between 52 Years and 75 Years. Death from COVID-19 has a positively moderate association with age of a patient with a Correlation Coefficient of 0.503. And from a comparative study, we can draw an opinion on the fact that with the increase in COVID-19 test, it is possible or easy to control the spreading of novel coronavirus. As many such patients can be quarantined if they are tested positive, not letting them to transmit the virus.

The prediction of the confirm cases for one week has been done with the 3rd Degree Polynomial Regression Model, as it best fits the curve. The prediction value and original value is very close, but the model is not the exact or best model to predict the number of confirm cases. Since in our model, we are not considering any other factors on which the growth rate of confirm case may depend. We are just taking time into account. But in actual model, other features also need to be taken in consideration which will vary the growth rate of confirm cases. Such as, preventive measures taken by a country, reproductive rate i.e. R_0 of the virus etc.

But in a nutshell the COVID-19 situation must be contained very soon under the guidelines from WHO and Government, otherwise it may take a much worse shape with coming days.

REFERENCE:

1. https://covidtracking.com/data/
2. https://www.cdc.gov/mmwr/volumes/69/wr/mm6912e2.htm
3. https://www.worldometers.info/coronavirus/
4. https://www.wikipedia.org/
5. https://www.sciencemag.org/news/2020/03/coronavirus-cases-have-dropped-sharply-south-korea-whats-secret-its-success
6. https://ourworldindata.org/covid-testing
7. https://www.statista.com/statistics/1103023/coronavirus-cases-distribution-by-age-group-italy/
8. https://www.statista.com/statistics/1105061/coronavirus-deaths-by-region-in-italy/
9. https://www.statista.com/statistics/1102730/south-korea-coronavirus-cases-by-age/
10. https://www.statista.com/statistics/241488/population-of-the-us-by-sex-and-age/
11. https://www.indexmundi.com/italy/demographics_profile.html
12. https://www.indexmundi.com/south_korea/demographics_profile.html
13. https://www.ijidonline.com/article/S1201-9712(20)30150-8/fulltext
14. https://www.ijidonline.com/article/S1201-9712(20)30150-8/fulltext
15. https://coronavirus.1point3acres.com/en/test
16. https://art-bd.shinyapps.io/nCov_control/