Statistical Analysis of Cases Related to Novel Coronavirus (COVID-19)

Submitted By:

Saksham Chauhan

Vishesh Malik

Sachin Bhandari

Samriddhi Soni

Vrinda Rawal

A. INTRODUCTION

Novel coronavirus disease (COVID 19) outbreak which originated from Wuhan, China, spread to the entire world in no time. On 11 march 2020, the outbreak was declared as a pandemic by WHO, who urged the countries to take urgent actions and increase response to diagnose, track and minimize transmission to save people's lives. India was quick to take actions against the outbreak ever since it faced its first confirmed case on 30th January. Since then the covid 19 cases have reached the 25,000 mark. The government's rapid actions in these desperate circumstances have so far prevented India from entering the third stage of covid 19 (i.e. community transmission) as is claimed by the authorities. In addition, the cases per million are fewer than the rest of the world and the recovery rate of infected patients is also higher. All this just makes us a little more hopeful that India might be able to pull through this outbreak, given the fact that the situation could have been much worse with India's huge population.

Coronavirus has four stages of transmission. The first stage is the first appearance of the disease. India had hit the first stage by 4th February, with its first three cases, all of which were students from Wuhan, China. By 4th march 22 new cases came to light, all with a history of travelling. As a result, on march 12th the government cancelled all tourist visas and had already started thermal screening and quarantining the incoming passengers way before in mid-January.

India reached the second stage which is local transmission in mid-March with an instant spike in the number of cases. To curb the virus' rapid spread and to buy more time to prepare the health system, the government of India under Prime Minister Narendra Modi announced a nationwide lockdown for 21 days on March 24th . This decision was followed after a 14 hour voluntary public curfew. Observers have reported that the lockdown has slowed the growth rate of the pandemic by 6 April. For India to shut down its entire economy was a tough call whose repercussions it will certainly face in the coming years. However, for the world's second most populous country, where social distancing is not easy, lockdown proves to be one of the few ways to fight the pandemic. But no matter what, there is little question that the decision lacked a little planning and came into effect without much advance warning. The lockdown has made clear the heavy prices paid by different sections of society especially the poor. The heart-rending situation of daily workers walking hundreds of kilometres to reach their homes at the start of the lockdown provided an insight into the suffering of millions of people unable to afford regular meals or even shelter.

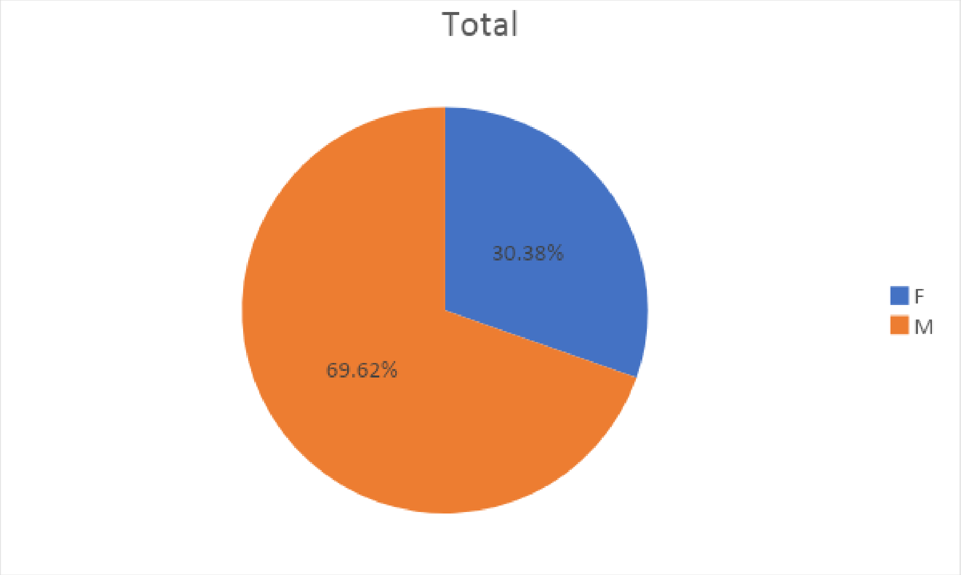
On 31st March, a Tablighi Jamaat religious congregation event that took place in Delhi in early March came into light. It emerged as a new virus hotspot after multiple cases across the country were linked back to the event. The incident has contributed to a rise in the number of cases and increased the risks of a community transmission. More than 10,000 people are believed to have attended the event including people from foreign countries. The cases continued to rise in India which led various states to ponder over the extension of the lockdown. On 14th April, PM Modi extended the nationwide lockdown until 3rd May, with a conditional relaxation after 20th April for the regions where the spread has been contained.

Despite all the precautions taken by the government, India is still likely to face a widespread outbreak. The mere fact that many cases are found to be asymptomatic only adds to all the worries. Moreover, the testing rate in India is extremely low owing to lack of testing kits and there is also a relative shortage of PPE kits. The actual number of cases might be more than there is to the surface. Also, the reluctance of people to comply with the lockdown and their indifference makes the situation worrisome. This period is not only hard on the citizens but on the country itself.

The whole economic slowdown India is going to face will drag it back in time. Each and every sector of the country is getting affected by this. We can only hope and play our part as responsible citizens that India will pull through all this and is going to stand victorious.

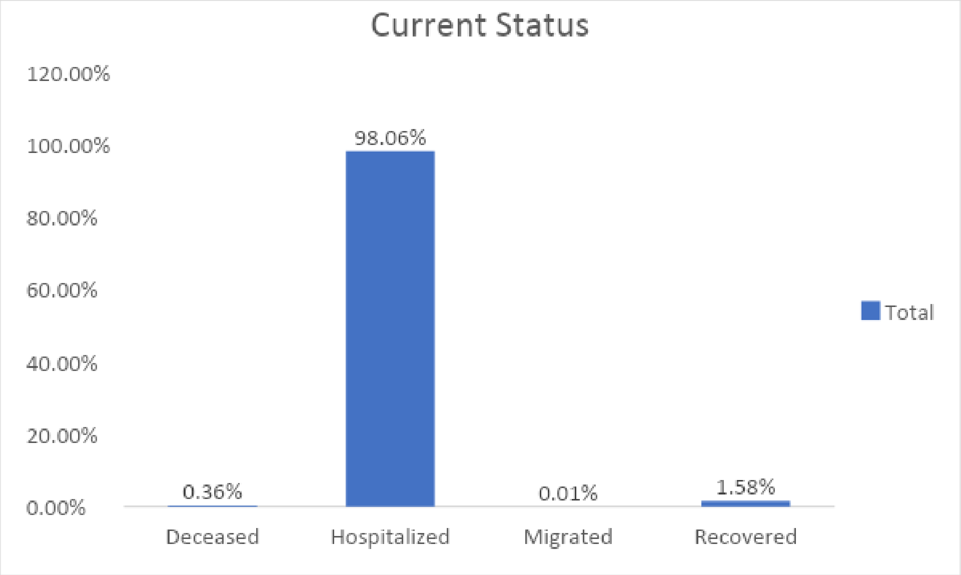
B. BASIC VISUALISATION OF DATA

- GENDER WISE GRAPH OF INFECTED PEOPLE



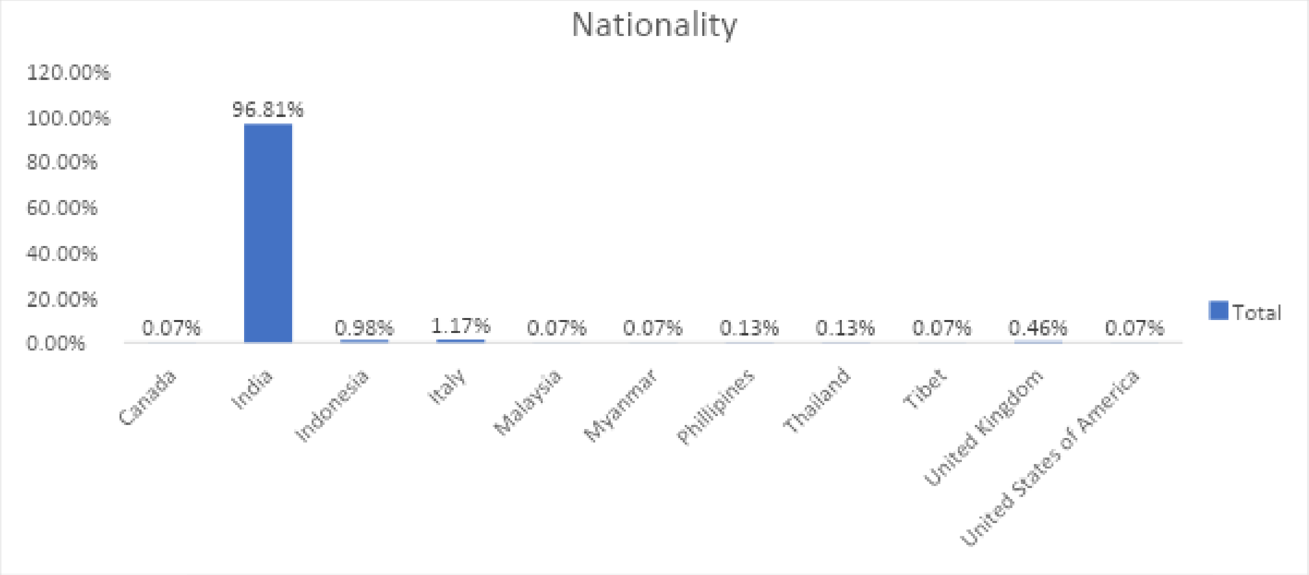
The percentage of males is 69.62 % indicating a higher chances of a male contracting the disease compared to females. The lower female percentage maybe because of lower interaction chances or the virus is more harmful to men of which the latter can be useful.

- CURRENT STATUS



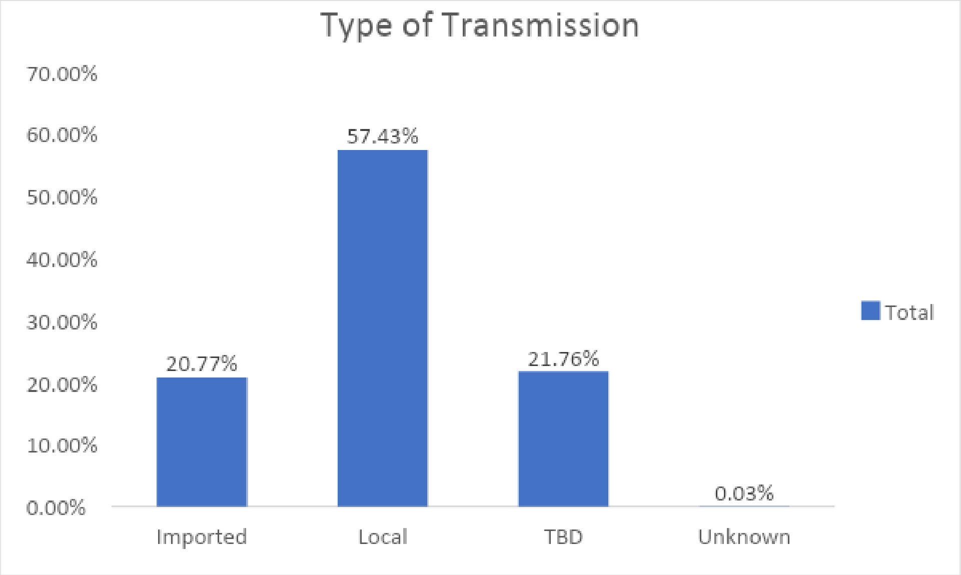
The current death rate could be considered to be 0.36% and the recovery rate to be 1.58%. This indicates the issues in the India’s medical facilities as there should be a bigger difference between the two than this.

- <u>NATIONALITY</u>



The data collected showed a few foreign nationals that had contracted the disease while still in India. Because the data is collected is only from India that is why it shows so much variation.

- <u>TYPE OF TRANSMISSION</u>



The type of transmission has mostly been local but still a large part of it has yet to be defined and the rest is imported. This shows that maybe the government was not as swift in its action to implement lockdown to prevent local transmission because otherwise it would have reduced the total cases by a lot.

- <u>AGE VS GENDER OF COVID 19 RECOVERED PATIENTS</u>

Count of Gender	Column Labels		

Row Labels	F	M	Grand Total
<1 or (blank)	8	13	21
1-20	3	8	11
21-40	9	40	49
41-60	11	13	24
61-80	7	8	15
81-100	1	2	3
Grand Total	39	84	123

- AGE VS GENDER OF COVID 19 CONFIRMED PATIENTS

Count of Gender	Column Labels		
Row Labels	F	M	Grand Total
<1 or (blank)	368	763	1131
1-20	72	101	173
21-40	151	475	626
41-60	115	310	425
61-80	67	123	190
81-100	3	6	9
Grand Total	776	1778	2554

- AGE VS GENDER OF COVID 19 DECEASED PATIENTS

Count of Gender	Column Labels		
Row Labels	F	M	Grand Total
1-20	0	1	1
21-40	1	2	3
41-60	3	6	9
61-80	6	19	25
81-100	1	0	1
Grand Total	11	28	39

C. CONSTRUCTION OF 3 RISK ZONES

COVID 19-A brief summary

The virus associated with the outbreak originating in Wuhan,China, has been designated Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV2).The disease caused by that virus is now officially called COVID-19. The virus has now spread gobally, and is officially declared as a Global Pandemic by WHO.

It has also spread in India with the total active cases being 23,452(as of 24 April 2020). According to a large section of experts, India's risks of catching the virus are disproportionately high because of its high population density, creaky healthcare mechanism and high internal migration.

High Risk States

There are many reasons which led to the increase of the confirmed cases in the high risk zones, but the most recent one being the community event through which transmission took place was the Tabhlighi Jamaat Congression in Delhi, with the total number of infected from this event rapidly increasing , across 14 states in just two days.Although Maharashtra has the highest number of cases, but Delhi ,Tamil Nadu, Andhra Pradesh and Telangana were the states which were directly impacted due to this congression.

India's case could be more worrisome than most other countries in the event of this outbrak because of its high population density; many of India's cities such as Mumbai, Kolkata and Delhi are extremely dense. These metropolitan cities are home to Asia's largest slum, hence ensuring dense population in various parts of the city; each square kilometre in India accounts for as many as 420 people, way higher than in most countries of the world.

In many other countries the virus spread because of the immigrants coming from the hotspot of this virus.But in India the early cases actually involved Itallians.So it arrived after it was fully seeded in Italy, and the screening of the passengers from this zone was on a small level.This is the main reason as to why the metropolitan cities became the hotspots of this pandemic in India.

Medium Risk States

India is a country of more than a billion people.It has extreme poverty in many areas; and many of its cities such as Mumbai and Kolkata are extraordinarily dense.Since the people in India travel through state borders for work, the deadly virus spread from the high risk zones to other parts of the country.

While India took quick steps to curb travel and put the country on lockdown, it has been criticized since the beginning of the outbreak for testing too few patients. Low rates of testing increases the risk communal spread of this virus.

There are many cases of people breaking the lockdown and not taking the preventive measures which are necessary to prevent the further spread of this virus.Also there are many cases of violence by people towards the healthcare staff treating them.This ignorant behaviour does nothing but aggreavates and worsen the present condition.

Low Risk States

Spread of the virus in this zone is also a result of interstate travel, but the spread of the virus in this region has been contained by adopting the preventive measures.

High risk states-(500 and above confirmed cases)

S.No.	Name of State/UT	Confirmed Cases	Cured/ Discharged/ Migrated	Death
1	Andhra Pradesh	955	145	29
2	Delhi	2376	808	50
3	Gujarat	2624	258	112
4	Madhya Pradesh	1852	203	83
5	Maharashtra	6430	840	283
6	Rajasthan	1964	230	27

7	Tamil Nadu	1683	752	20
8	Telangana	984	253	26
9	Uttar Pradesh	1604	206	24
10	West Bengal	514	103	15
	TOTAL	**20986**	**3798**	**669**

Medium Risk States-(100-500 confirmed cases)

S.No.	Name of State/UT	Total Confirmed Cases	Cured/ Discharged/ Migrated	Death
1	Bihar	176	46	2
2	Haryana	272	156	3
3	Jammu Kashmir	427	92	5
4	Karnataka	463	150	18
5	Kerala	448	324	3
6	Punjab	277	65	16
	TOTAL	**2063**	**833**	**47**

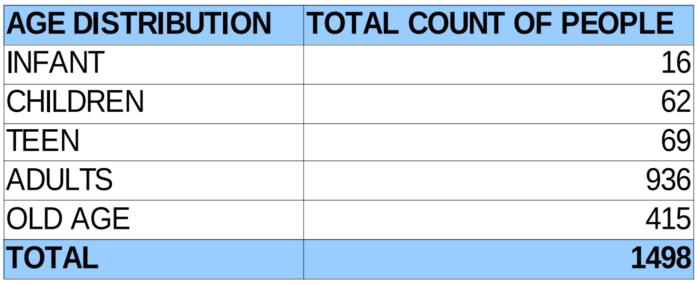
Low Risk States-(0-100 Confirmed Cases)

S.No.	Name of State/UT	Total Confirmed Cases	Cured/Discharged/ Migrated	Death
1	Andaman and Nicobar Islands	22	11	0
2	Arunachal Pradesh	1	1	0
3	Assam	36	19	1
4	Chandigarh	27	14	0
5	Chattisgarh	36	28	0
6	Goa	7	7	0
7	Himachal Pradesh	40	18	1
8	Jharkhand	55	8	3
9	Ladakh	18	14	0
10	Manipur	2	2	0
11	Meghalaya	12	0	1
12	Mizoram	1	0	0
13	Odisha	90	33	1
14	Puducherry	7	3	0
15	Tripura	2	1	0

16	Uttarakhand	47	24	0
	TOTAL	**403**	**183**	**7**

TABLES AND BASIC VISUALIZATION OF 3 RISK ZONES

1. AGE WISE DISTRIBUTION

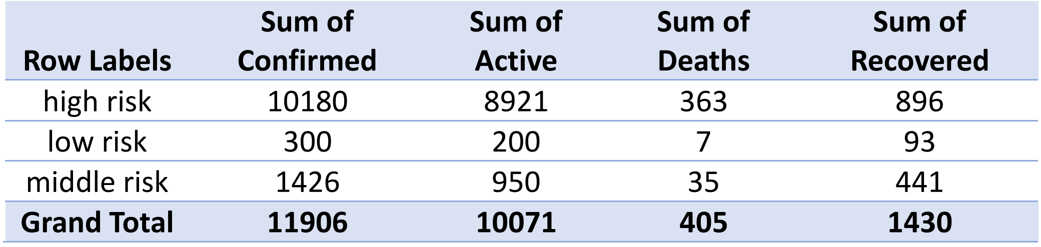


AGE DISTRIBUTION	TOTAL COUNT OF PEOPLE
INFANT	16
CHILDREN	62
TEEN	69
ADULTS	936
OLD AGE	415
TOTAL	**1498**

INFANT-0-3 YEARS
CHILDREN-3-13 YEARS
TEEN-13-19 YEARS
ADULTS-19-50 YEARS
OLD AGE-BEYOND 50 YEARS

INTERPRETATION-Maximum number of people who were infected belonged to the age group of 19-50 years (adults).The people who were least infected belonged to the category of 0-3 years (infant).

2. STATE WISE DISTRIBUTION



Row Labels	Sum of Confirmed	Sum of Active	Sum of Deaths	Sum of Recovered
high risk	10180	8921	363	896
low risk	300	200	7	93
middle risk	1426	950	35	441
Grand Total	**11906**	**10071**	**405**	**1430**

HIGH RISK STATES-Confirmed cases above 500
MEDIUM RISK STATES-Confirmed case between 100-500
LOW RISK STATES-Confirmed cases below 100

INTERPRETATION-
1. The High Risk States have the maximum number of confirmed , active and recovered cases and maximum number of deaths.
2. The Low Risk States have the minimum number of confirmed , active and recovered cases and minimum number of deaths.

- HIGH RISK STATES

States	Confirmed	Recovered	Deaths	Active
Maharashtra	2801	259	178	2364
Delhi	1561	31	30	1500
Tamil Nadu	1242	118	14	1110
Rajasthan	1046	147	11	888
Madhya Pradesh	741	64	53	624
Uttar Pradesh	735	55	11	669
Gujarat	695	59	30	606
Telangana	644	110	18	516
Andhra Pradesh	502	16	11	475
West Bengal	213	37	7	169

INTERPRETATION-
1. Maharashtra has the highest number of confirmed,recovered and active cases and maximum number of deaths among the High Risk States.
2. West Bengal has the lowest number of confirmed and active cases and lowest number of deaths.Andhra Pradesh has the lowest number of recovered cases among the high risk states.

- MEDIUM RISK STATES-

States	Confirmed	Recovered	Deaths	Active
Kerala	387	218	2	167
Jammu and Kashmir	300	30	4	266
Karnataka	279	80	12	187
Haryana	204	57	3	144
Punjab	186	27	13	146
Bihar	70	29	1	40

INTERPRETATION-
1.Kerala has maximum number of confirmed and recovered cases among the medium risk states.Punjab has maximum deaths among the medium risk states.Jammu and Kashmir has maximum number of active cases among the medium risk states.
2.Bihar has minimum number of confirmed, recovered and active cases and minimum number of deaths among the medium risk states.

- LOW RISK STATES-

States	Confirmed	Recovered	Deaths	Active
Odisha	60	18	1	41
Uttarakhand	37	9	0	28
Chhattisgarh	33	13	0	20
Himachal Pradesh	33	13	2	18
Assam	32	0	1	31
Jharkhand	27	0	2	25
Chandigarh	21	9	0	12

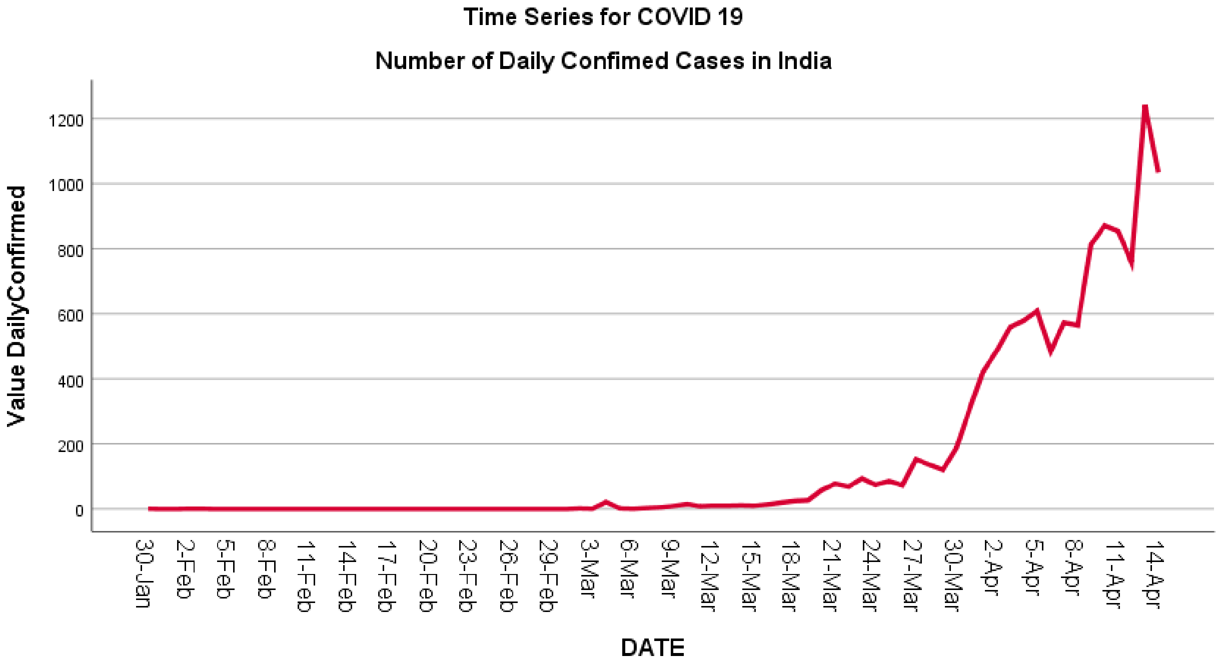
Ladakh	17	12	0	5
Andaman and Nicobar Islands	11	10	0	1
Goa	7	5	0	2
Puducherry	7	1	0	6
Manipur	2	1	0	1
Tripura	2	1	0	1
Mizoram	1	0	0	1
Arunachal Pradesh	1	1	0	0
Dara, Daman, Diu and Nagar	1	0	0	1
Nagaland	1	0	0	1
Meghalaya	7	0	1	6
Lakshadweep	0	0	0	0
Sikkim	0	0	0	0

INTERPRETATION-
1.Odisha has maximum number of confirmed,recovered and active cases among the Low Risk States.Himachal Pradesh and Jharkhand have the maximum number of casualties among the Low Risk States.
2.Lakshadweep and Sikkim has no cases.

D. TIME SERIES MODELLING AND FORECASTING OF NUMBER OF DAILY CONFIRMED COVID 19 CASES IN INDIA

1. OBSERVED TIME SERIES

Below is the Observed Time Series for the Number of Daily Confirmed Cases in India:

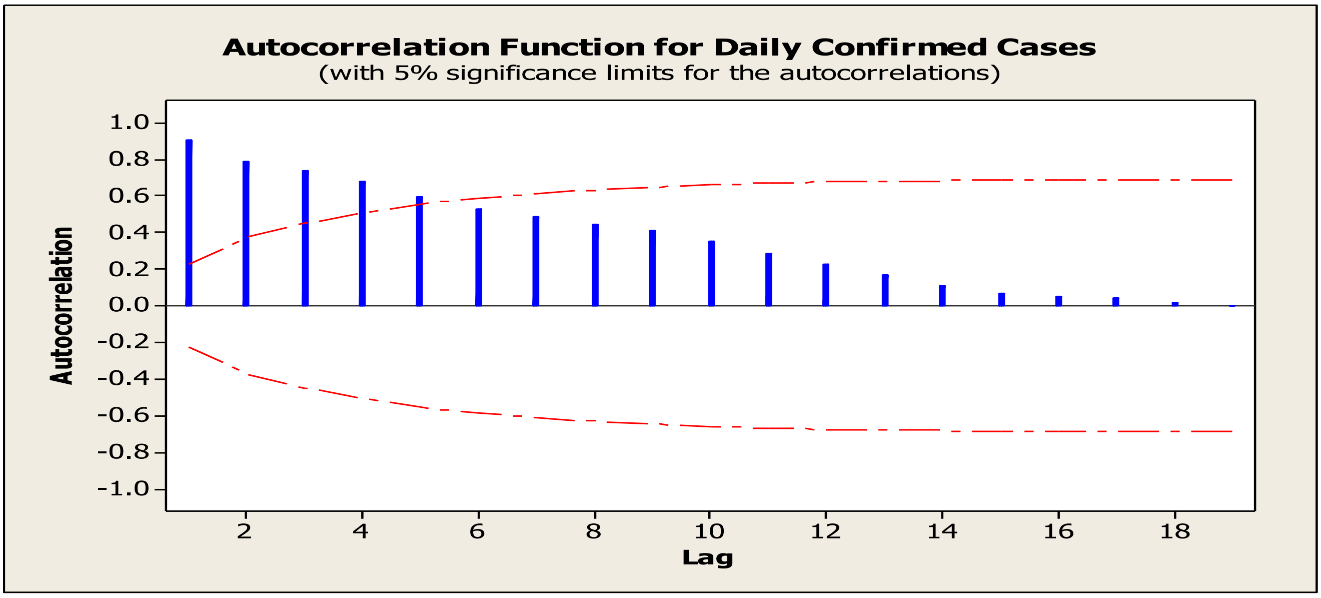


The Time Series is Non-Stationary representing Upward Trend (Exponential Increase)

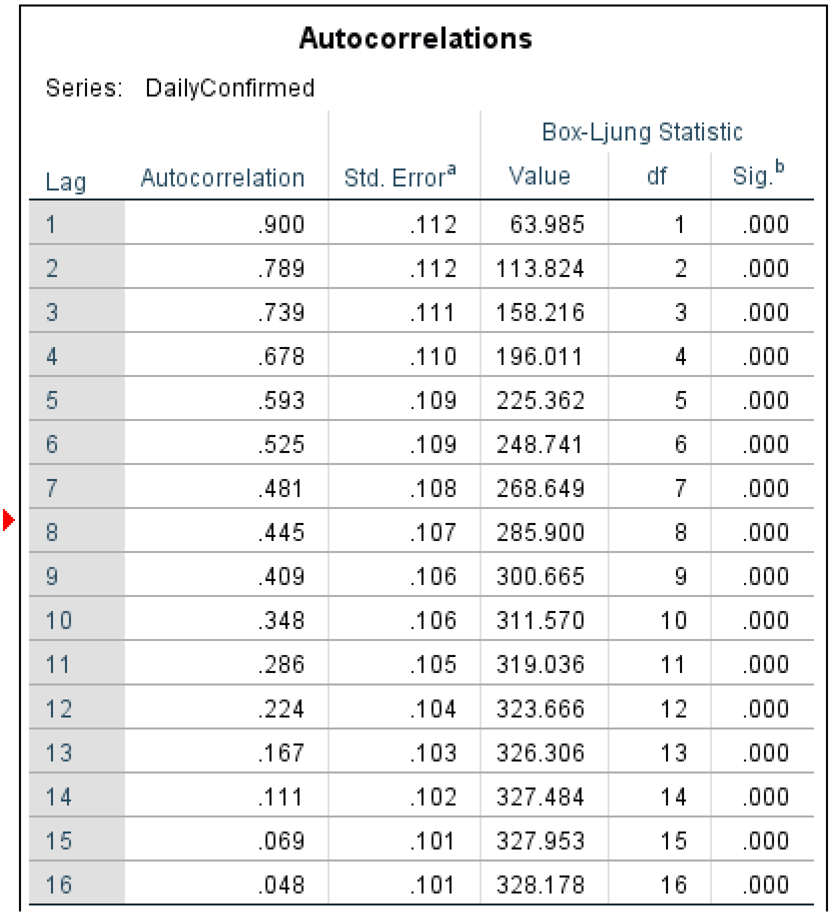
As evident from the given Time Series, the curve is rising which essentially represents Non-Stationarity due to presence of an upward trend in the data. Moreover, as known from past case studies and historical evidences, the data for such a variable related to an Epidemic seems to grow exponentially, which is the case here.

Below are the Sample Autocorrelation and Sample Partial Autocorrelation Function for the given data:

- AUTOCORRELATION FUNCTION



The presence of a long exponential decay in the ACF function confirms the fact that there are signs of Non-Stationarity in the observed data.



Autocorrelations

Series: DailyConfirmed

			Box-Ljung Statistic		
Lag	Autocorrelation	Std. Error[a]	Value	df	Sig.[b]
1	.900	.112	63.985	1	.000
2	.789	.112	113.824	2	.000
3	.739	.111	158.216	3	.000
4	.678	.110	196.011	4	.000
5	.593	.109	225.362	5	.000
6	.525	.109	248.741	6	.000
7	.481	.108	268.649	7	.000
8	.445	.107	285.900	8	.000
9	.409	.106	300.665	9	.000
10	.348	.106	311.570	10	.000
11	.286	.105	319.036	11	.000
12	.224	.104	323.666	12	.000
13	.167	.103	326.306	13	.000
14	.111	.102	327.484	14	.000
15	.069	.101	327.953	15	.000
16	.048	.101	328.178	16	.000

BOX LJUNG Q TEST

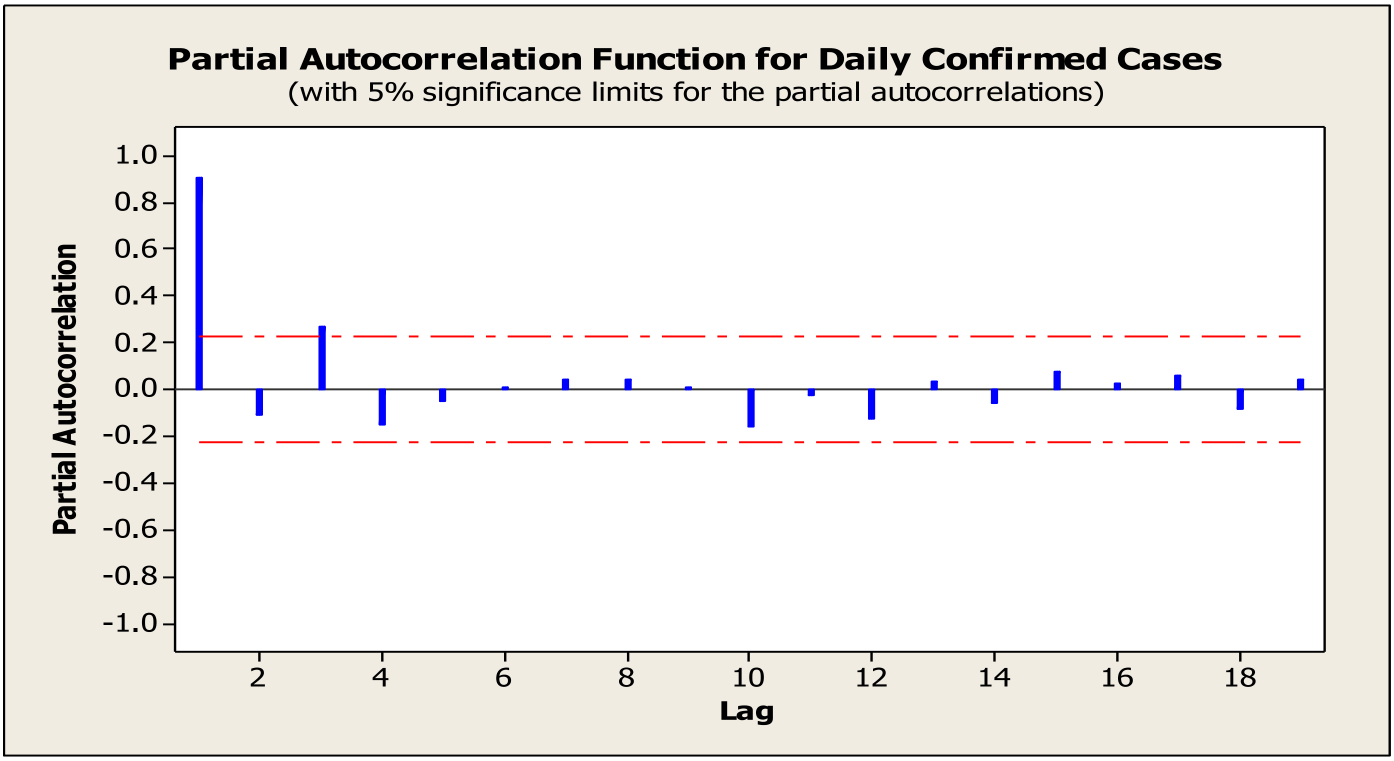
We have used the Box Ljung Q Test to confirm whether the given series is a White Noise Process or an Autoregressive One.

Null Hypothesis, H_o: The given series is Random or White Noise process, or the given ACF function for all lags is 0.

Alternative Hypothesis, H_1: The ACF function for lags is not 0.

Result: Given the p-value 0.000 for all the lags which is less than 0.05 which is Level of Significance, means that the given observations provide enough evidence in favor of rejection of Null Hypothesis. Hence, we conclude that the given series is not White Noise process.

- PARTIAL AUTOCORRELATION FUNCTION



The PACF represents Non-Stationarity in the observed model. At lag 1, there is a spike which is statistically significant, while the other spikes after lag 1 are relatively quite short. This is typical of the presence of Non-Stationarity in the Model.

- TEST FOR STATIONARITY

The Augmented Dickey-Fuller Test is used to test whether there is any unit root present in the Time Series. A series with a trend line (Non-Stationarity) will have a unit root.

Null Hypothesis, H_o: The given series has a unit root and a trend line, hence non-stationary.

Alternative Hypothesis, H_1: The given series is a stationary one.

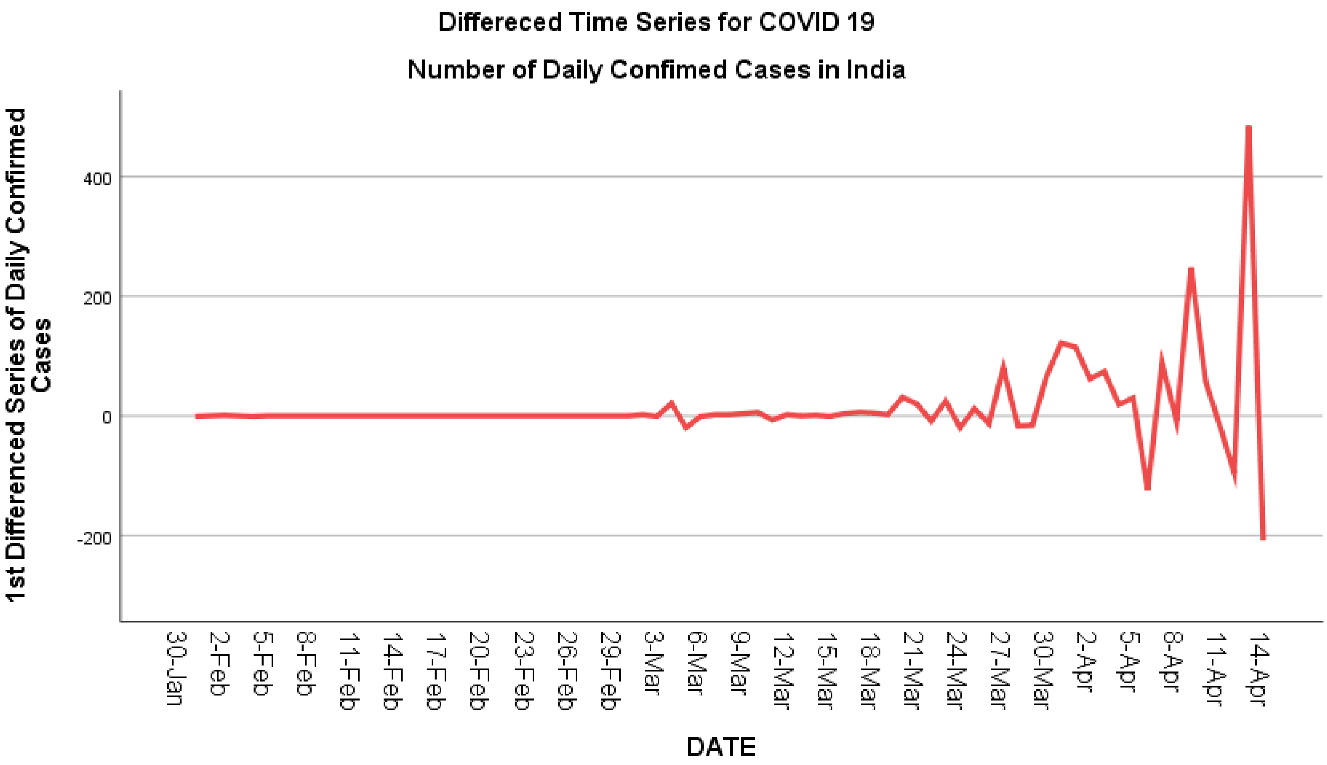
```
Augmented Dickey-Fuller Test

data:  dailyconfirmedcases
Dickey-Fuller = 1.255, Lag order = 4, p-value = 0.99
alternative hypothesis: stationary
```

Since p value is 0.99 which is much greater than 0.05 (Level of Significance), then on the basis of given observed series, we cannot reject the Null Hypothesis. Thus, we conclude that **the given series is Non-Stationary.**

2. DIFFERENCED TIME SERIES

As the observed series is non-stationary, we need to make it stationary using Differencing of Observed Time Series Model.
We use Differencing at Order 1.

Below is the Differenced Series of Observed Time Series for the Number of Daily Confirmed Cases in India:



The given figure depicts the First Differenced Series to be stationary, as it is concentrated near a constant mean value above 0.

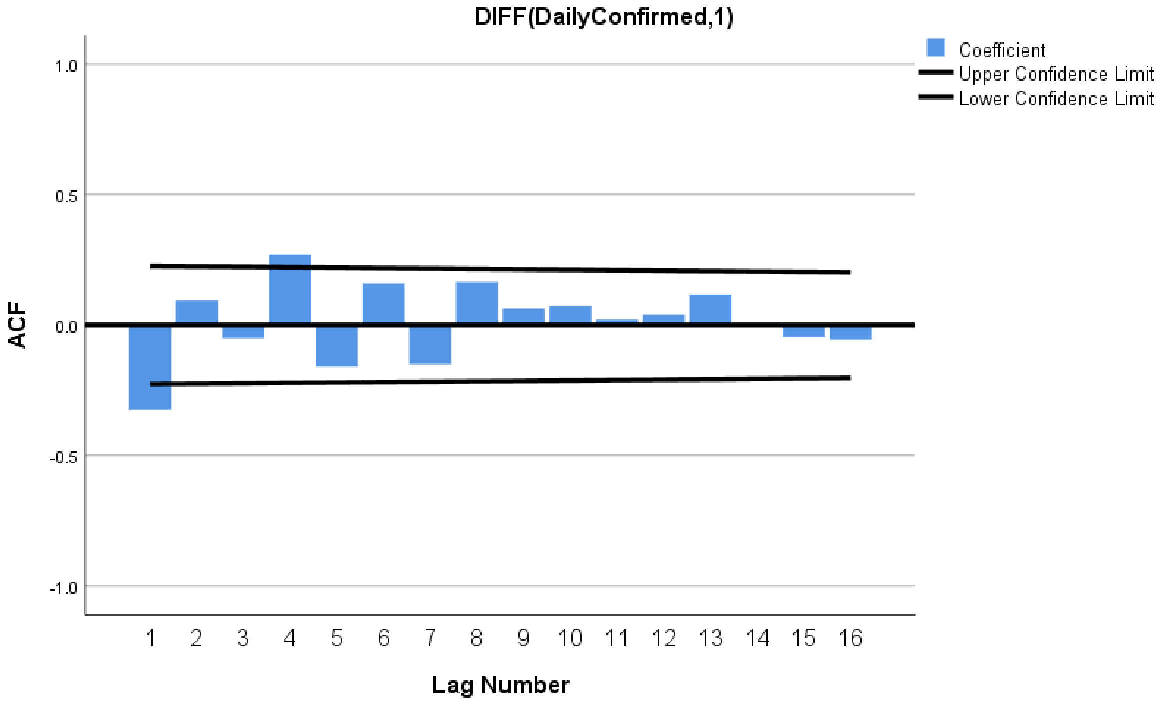
We again conduct an Augmented Dickey Fuller Test to check the Stationarity of the Differenced Model.

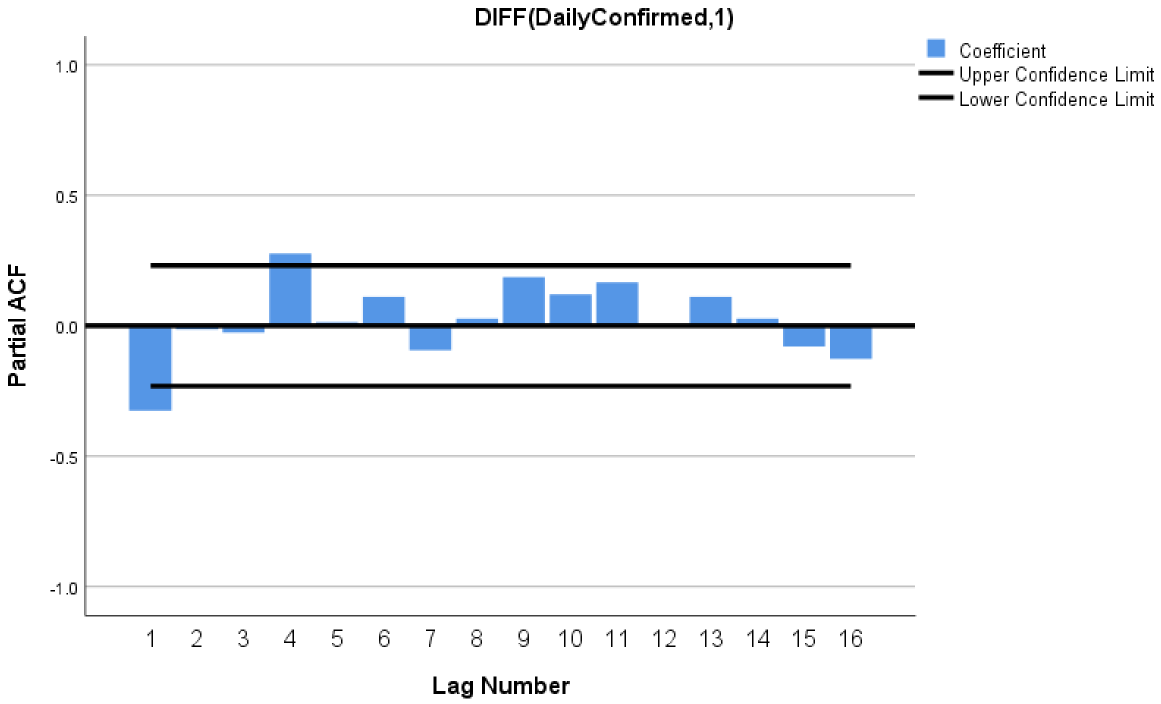
- <u>TEST FOR STATIONARITY</u>

```
Augmented Dickey-Fuller Test

data:  diffdailyconfirmedcases
Dickey-Fuller = -4.4553, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

Since p value is 0.01 which is less than 0.05 (Level of Significance), then on the basis of given observed series, we have to reject the Null Hypothesis. Thus, we conclude that **the given series is Stationary.**

- <u>AUTOCORRELATION FUNCTION AND PARTIAL AUTOCORRELATION FUNCTION</u>





Since, the lag 1 ACF is neither too spiked nor less than -0.5, it is concluded that the series is neither under-differenced nor over-differenced.

Moreover, there are patternless spikes at lags in both ACF and PACF of the differenced series, indicating achievement of stationarity merely with differencing of order 1.

Moreover, there are spikes which are outside the confidence limits indicating that the process is not White Noise. This can also be confirmed with the application of Box Ljung Q Test for Randomness of Time Series.

Hence, we obtain a Stationary Series.

3. MODEL SELECTION AND DIAGNOSTIC

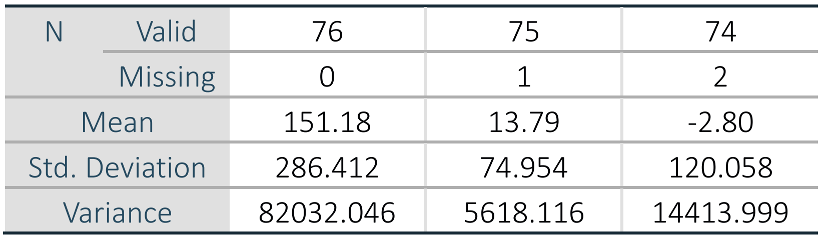
a) Selection of Necessary Order of Differencing

With the first order differencing, we attain stationarity.

However, one would like to check for a second order differencing. The differencing order should be so selected that the obtained series should be variance stabilized. Hence, we differenced the original time series with an order 2, to obtain another series and below we provide the Descriptive Statistics of the differenced, double differenced and original time series.

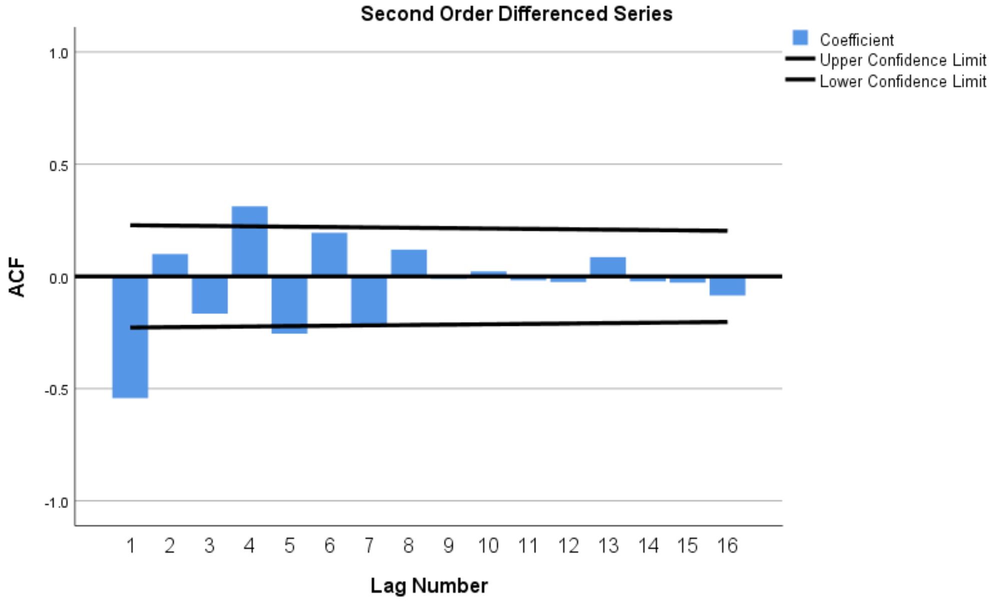
Statistics

	DailyConfirmed	DIFF(DailyConfirmed,1)	DIFF(DiffDailyconfirmed,1)



N	Valid	76	75	74
	Missing	0	1	2
Mean		151.18	13.79	-2.80
Std. Deviation		286.412	74.954	120.058
Variance		82032.046	5618.116	14413.999

As witnessed from the given descriptive statistics table, first differenced series offers the least Standard Deviation, and hence, if we move to order 2 differencing, then it would lead to over-differencing,



Another indicator of over-differencing is the Spike at Lag 1, which has a value > (-0.5). Differencing eliminates or reverses the high autocorrelation between terms and tend to negate it. Therefore, over differencing leads to lag 1 value going close to or greater than (-0.5).

Note: Addition of an AR term represents partially adding a differencing in the model.

While, Addition of an MA term represents partially cancelling out a differencing in the model.

Therefore, depending upon the orders of AR and MA in model, we can have several befitting models without considering differencing of the series.

b) Selection of AR(p) and MA(q)

With the help of PACF, the correct order of MA can be determined as it cuts off after the q lag. With the help of ACF, the correct order of AR can be determined as it cuts off after the p lag. So keeping in view the ACF and PACF of Differenced Series, it can be seen that no graph cuts off after any lag, but is exponentially decaying in either direction (depending upon the sign of estimates).

This is because, we do have an Autoregressive Integrated Moving Average (ARIMA) model application or Integrated Moving Average (IMA) model application in which the behavior of ACF and PACF is quite difficult to understand.

However, we take the following combinations which could be understood from the given PACF and ACF.

i. p=1 , q=2
ii. p=0 , q=2
iii. p=4 , q=4
iv. p=4 , q=3

c) Model Adequacy

To check for model adequacy, we have several options, among which Normalised BIC value is the best one. The best model features with least BIC value.

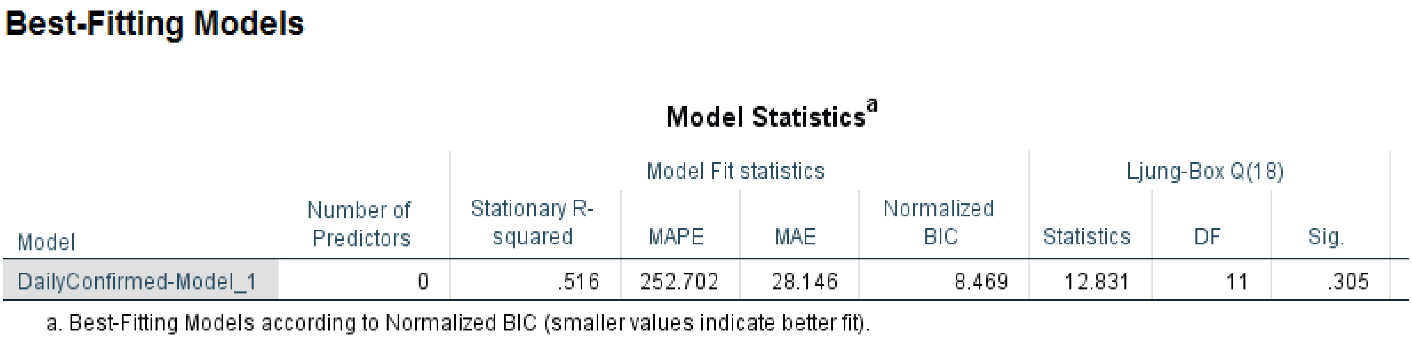
However, Stationary R Squared Values are also indicator for Goodness of Fit to the model

d) Diagnostic Checking

We conduct the Box Ljung Q test again in order to test the randomness of the residual left after the model fitting.

Below are the models which have been put into due consideration and selection has been done on the basis of diagnostics method observed with Residual Plot Analysis.

MODEL # 01 : ARIMA (4,1,3)

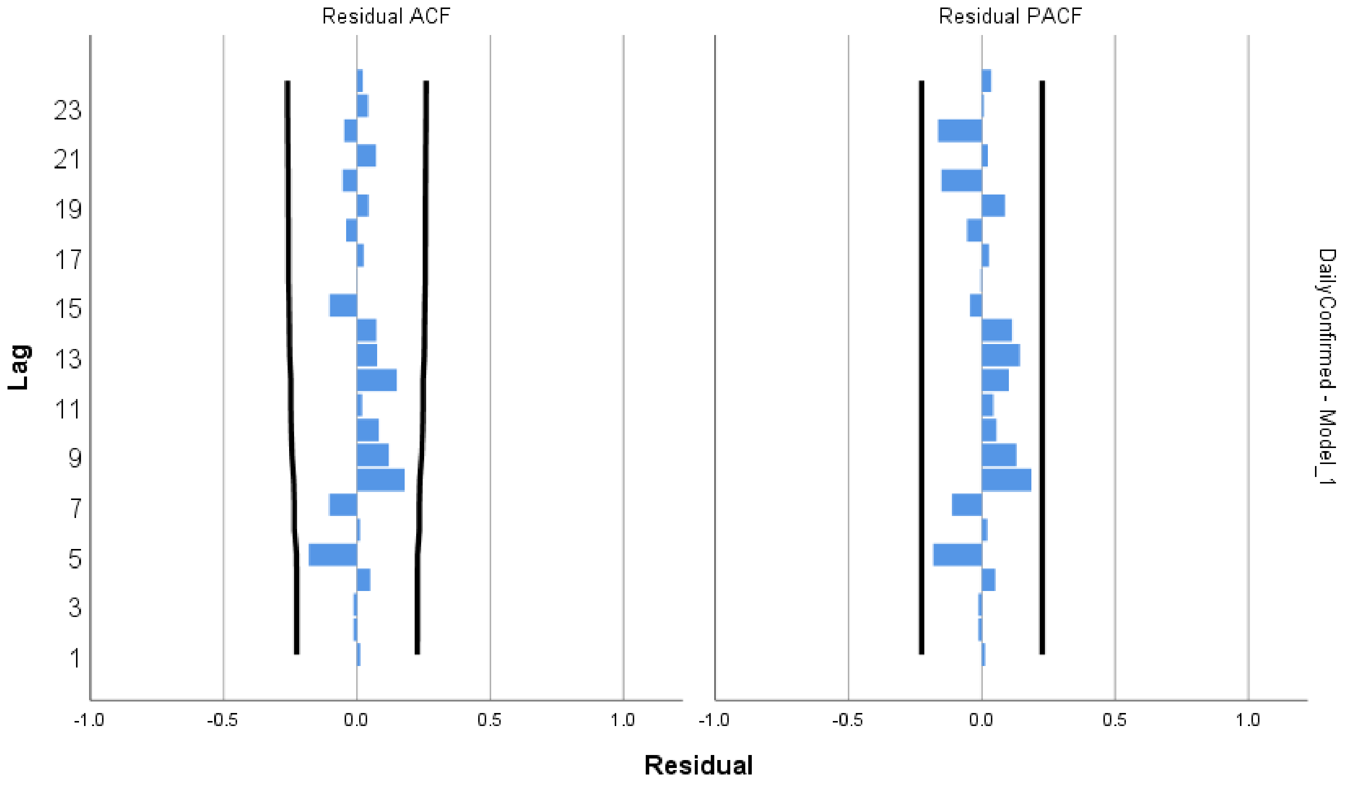


Best-Fitting Models

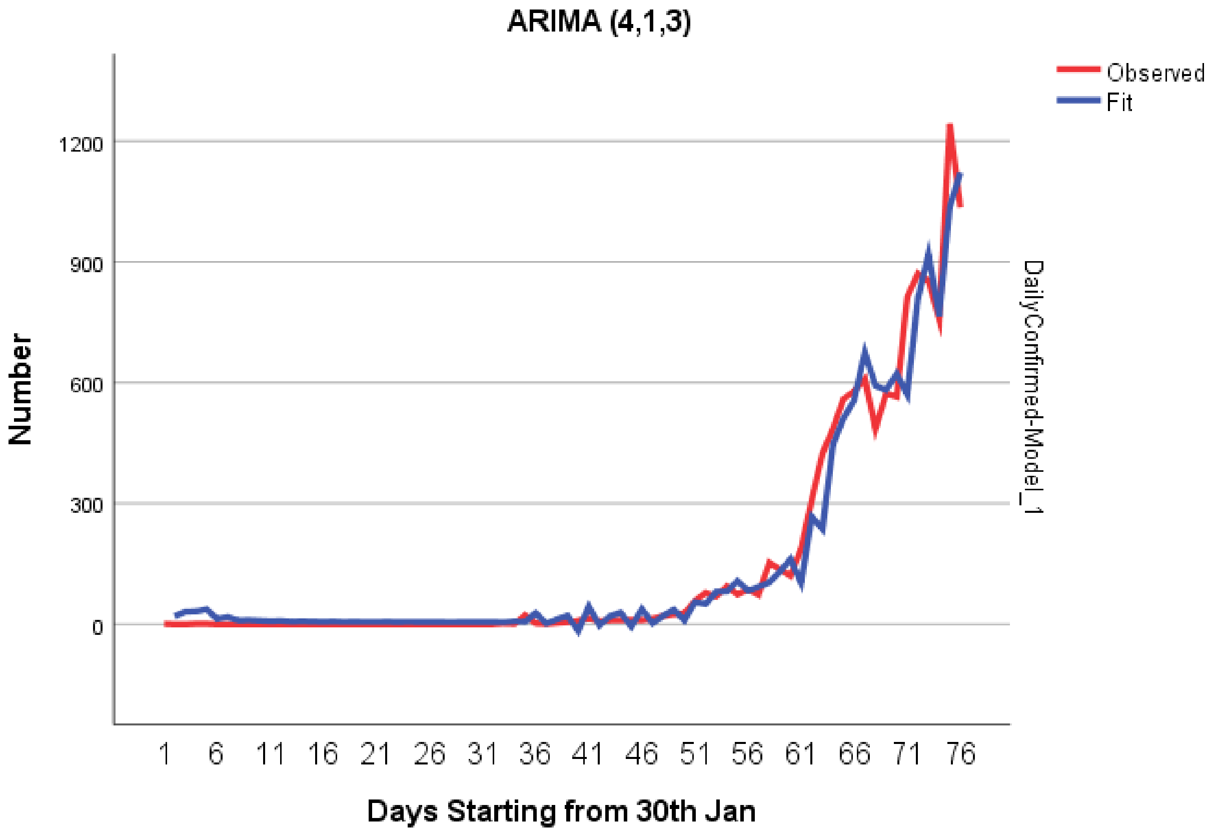
Model Statistics[a]

		Model Fit statistics				Ljung-Box Q(18)		
Model	Number of Predictors	Stationary R-squared	MAPE	MAE	Normalized BIC	Statistics	DF	Sig.
DailyConfirmed-Model_1	0	.516	252.702	28.146	8.469	12.831	11	.305

a. Best-Fitting Models according to Normalized BIC (smaller values indicate better fit).

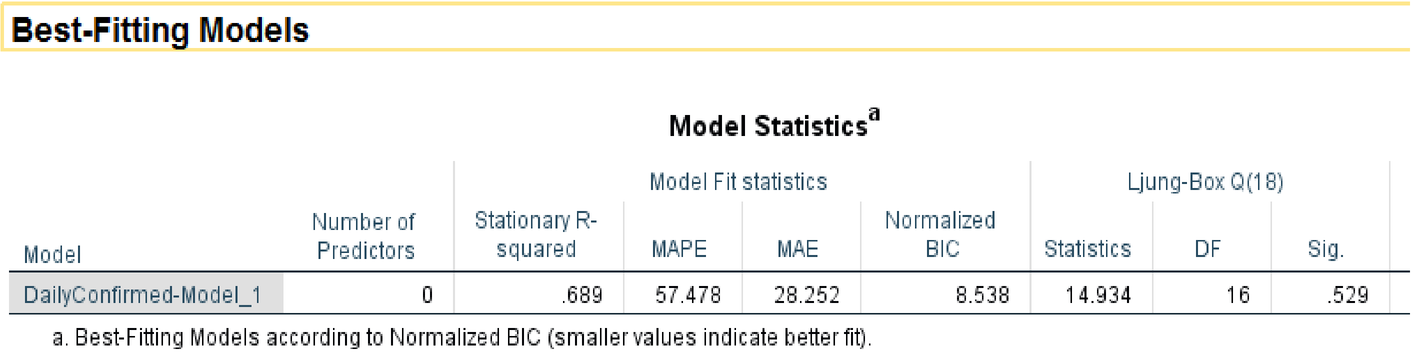


Residual ACF
Residual PACF
Lag
23
21
19
17
15
13
11
9
7
5
3
1
-1.0
-0.5
0.0
0.5
1.0
Residual
DailyConfirmed - Model_1



ARIMA (4,1,3)
Observed
Fit
Number
1200
900
600
300
0
1
6
11
16
21
26
31
36
41
46
51
56
61
66
71
76
Days Starting from 30th Jan
DailyConfirmed-Model_1

MODEL # 02 : ARIMA (0,2,2)

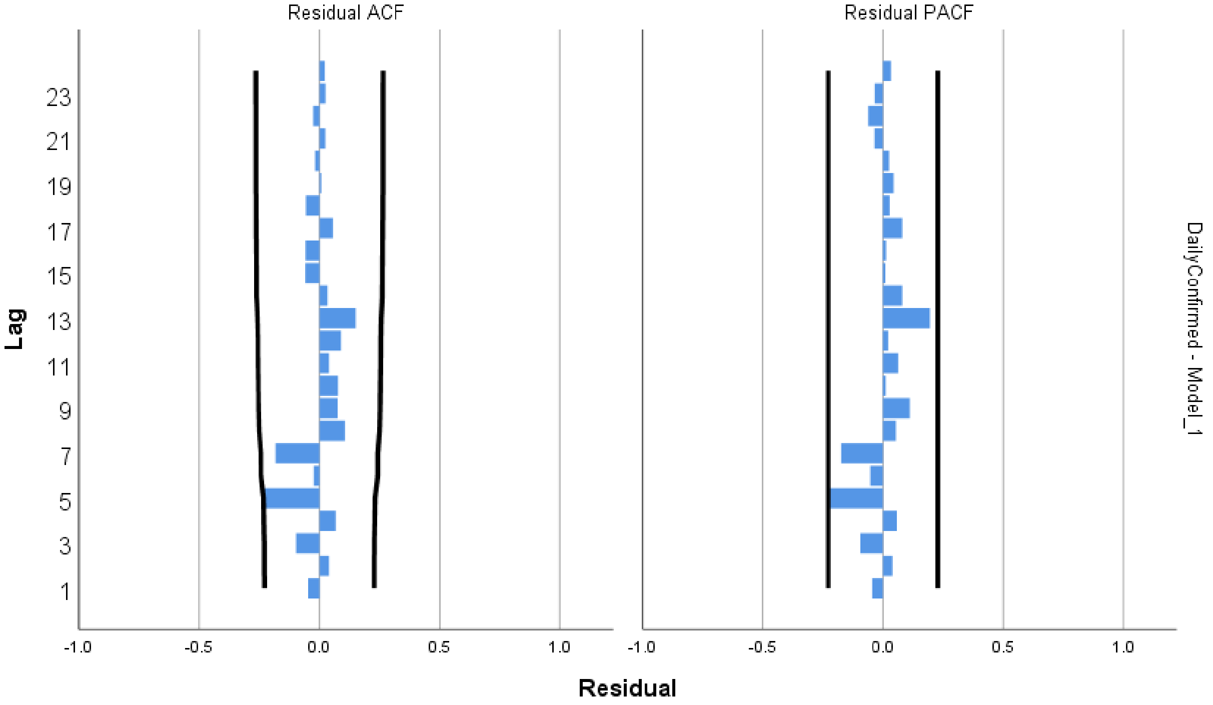


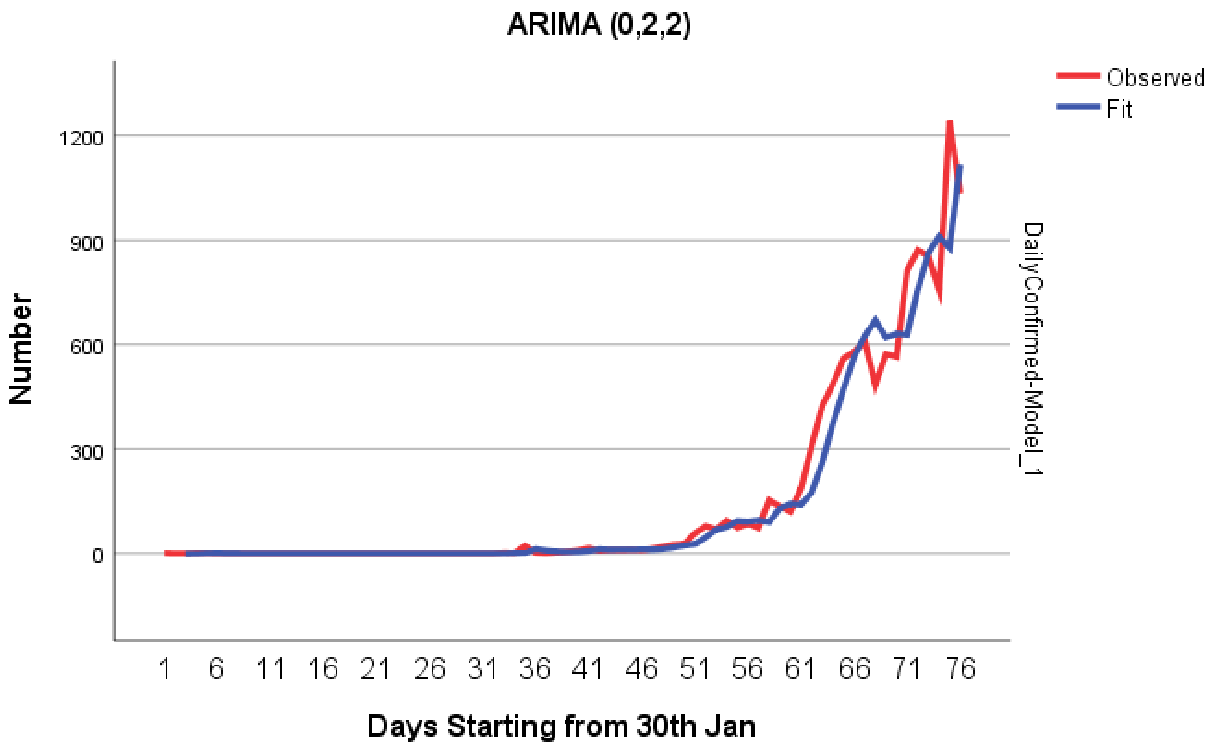
Best-Fitting Models

Model Statistics[a]

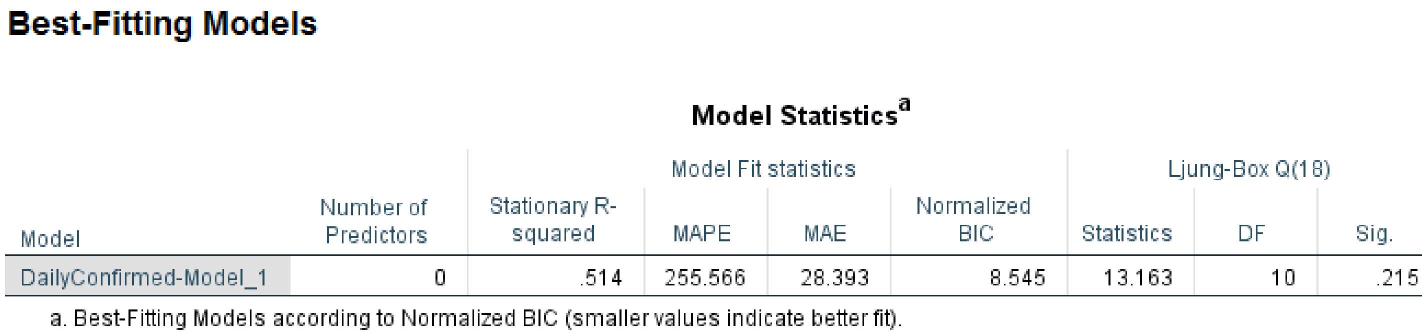
		Model Fit statistics				Ljung-Box Q(18)		
Model	Number of Predictors	Stationary R-squared	MAPE	MAE	Normalized BIC	Statistics	DF	Sig.
DailyConfirmed-Model_1	0	.689	57.478	28.252	8.538	14.934	16	.529

a. Best-Fitting Models according to Normalized BIC (smaller values indicate better fit).





MODEL # 03 : ARIMA (4,1,4)

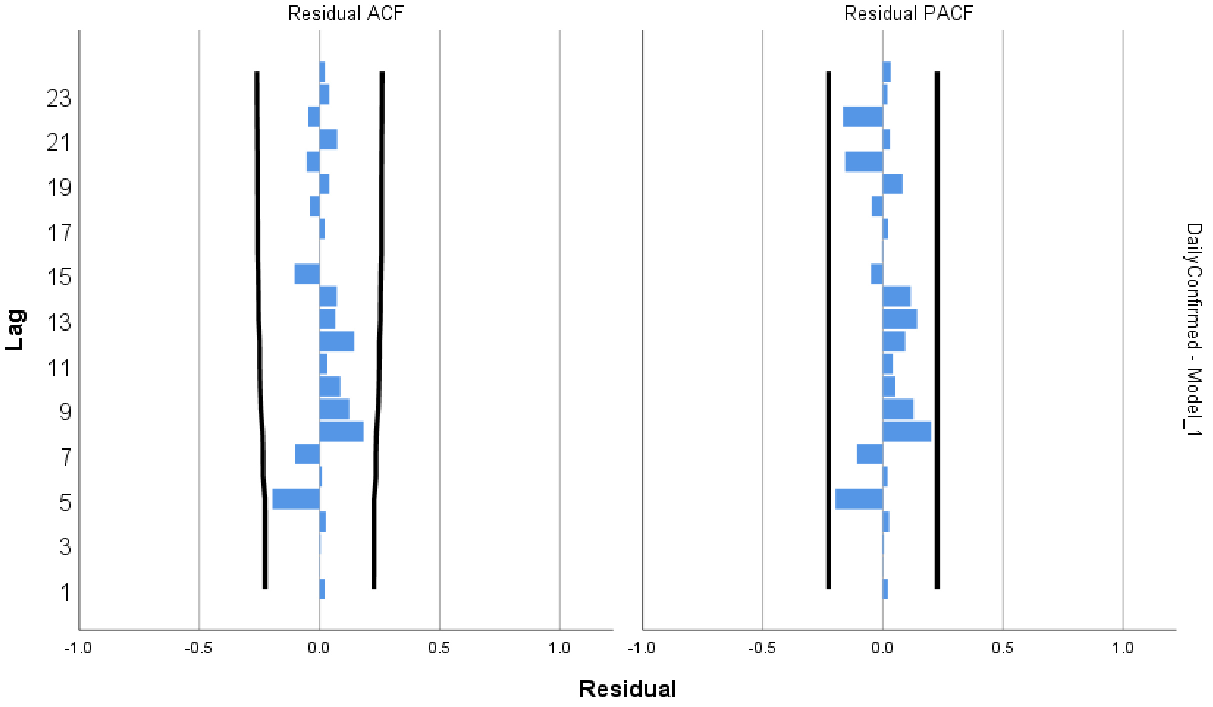


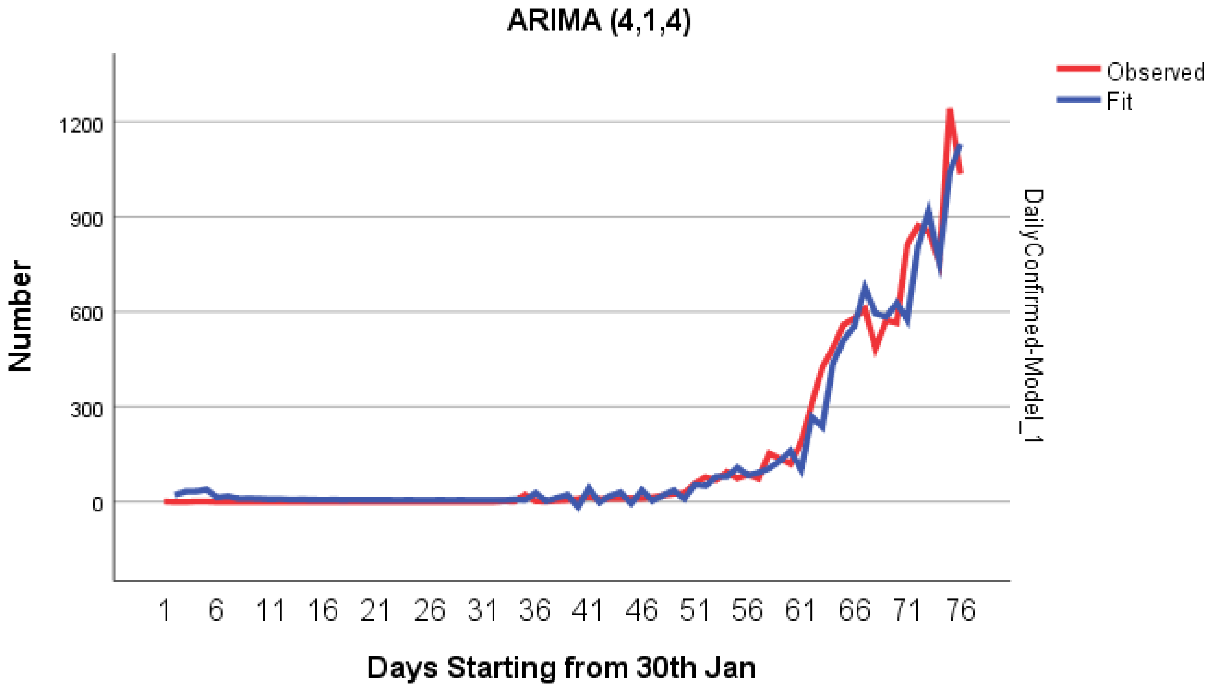
Best-Fitting Models

Model Statistics[a]

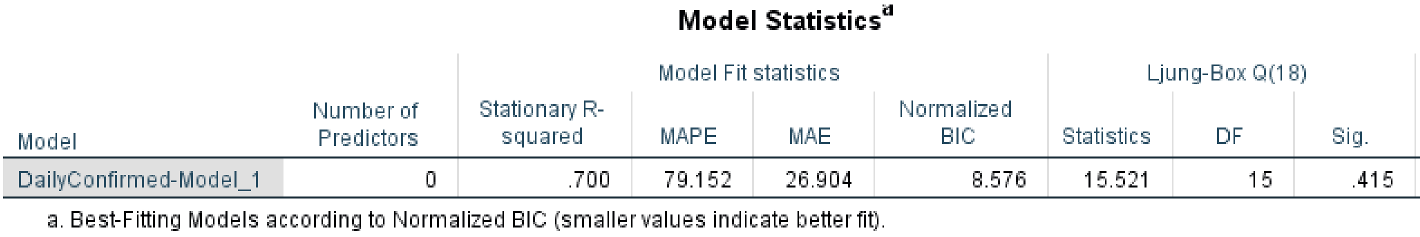
Model	Number of Predictors	Model Fit statistics				Ljung-Box Q(18)		
		Stationary R-squared	MAPE	MAE	Normalized BIC	Statistics	DF	Sig.
DailyConfirmed-Model_1	0	.514	255.566	28.393	8.545	13.163	10	.215

a. Best-Fitting Models according to Normalized BIC (smaller values indicate better fit).





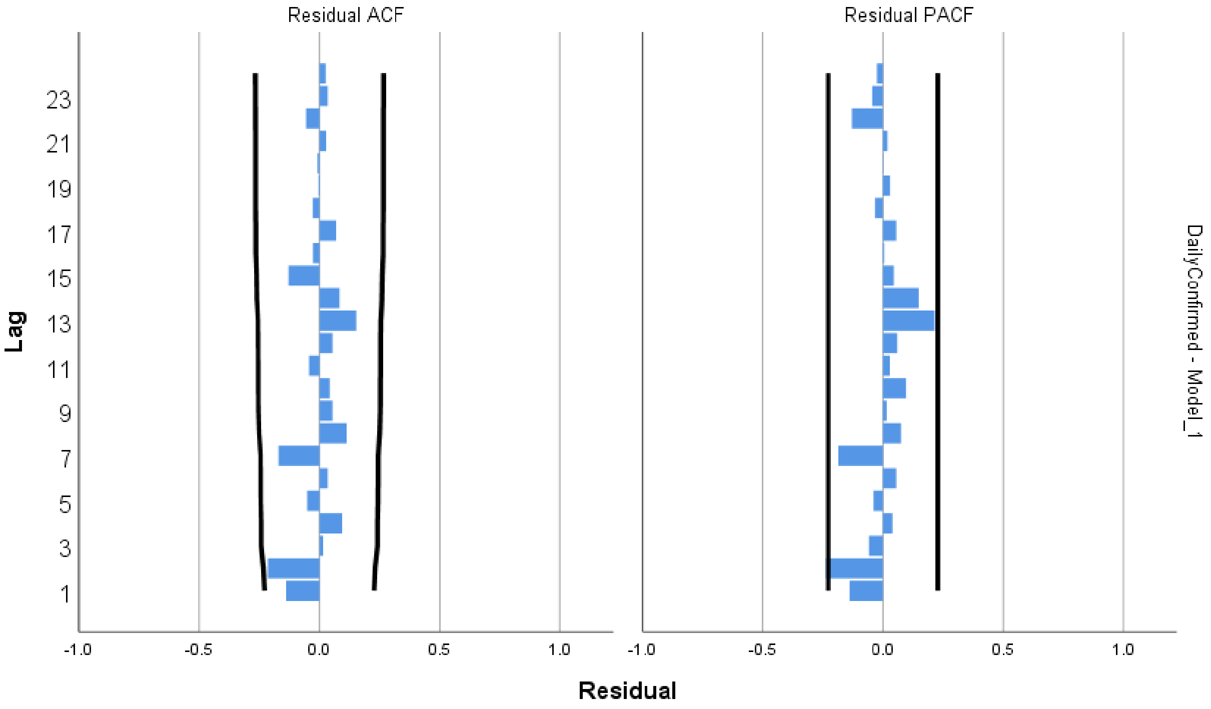
<u>MODEL # 03 : ARIMA (1,2,2)</u>

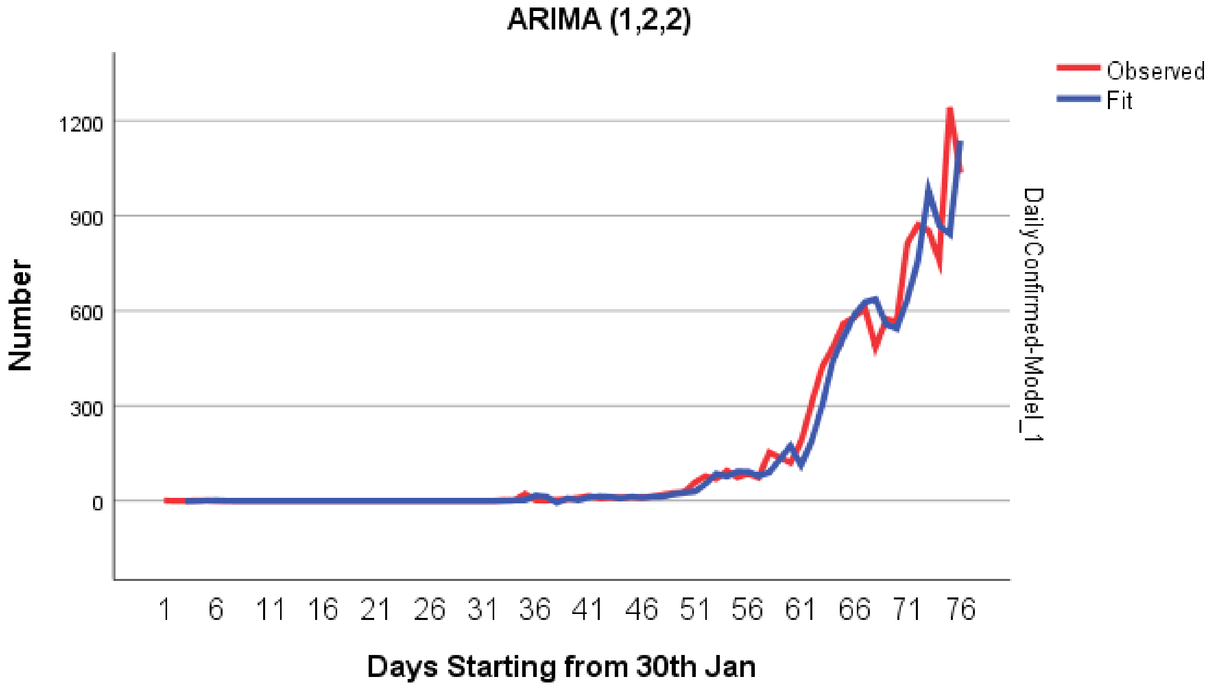


Model Statistics[a]

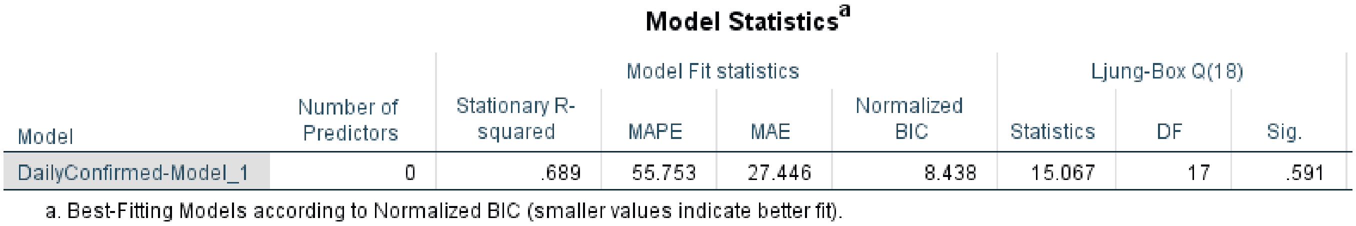
Model	Number of Predictors	Model Fit statistics				Ljung-Box Q(18)		
		Stationary R-squared	MAPE	MAE	Normalized BIC	Statistics	DF	Sig.
DailyConfirmed-Model_1	0	.700	79.152	26.904	8.576	15.521	15	.415

a. Best-Fitting Models according to Normalized BIC (smaller values indicate better fit).





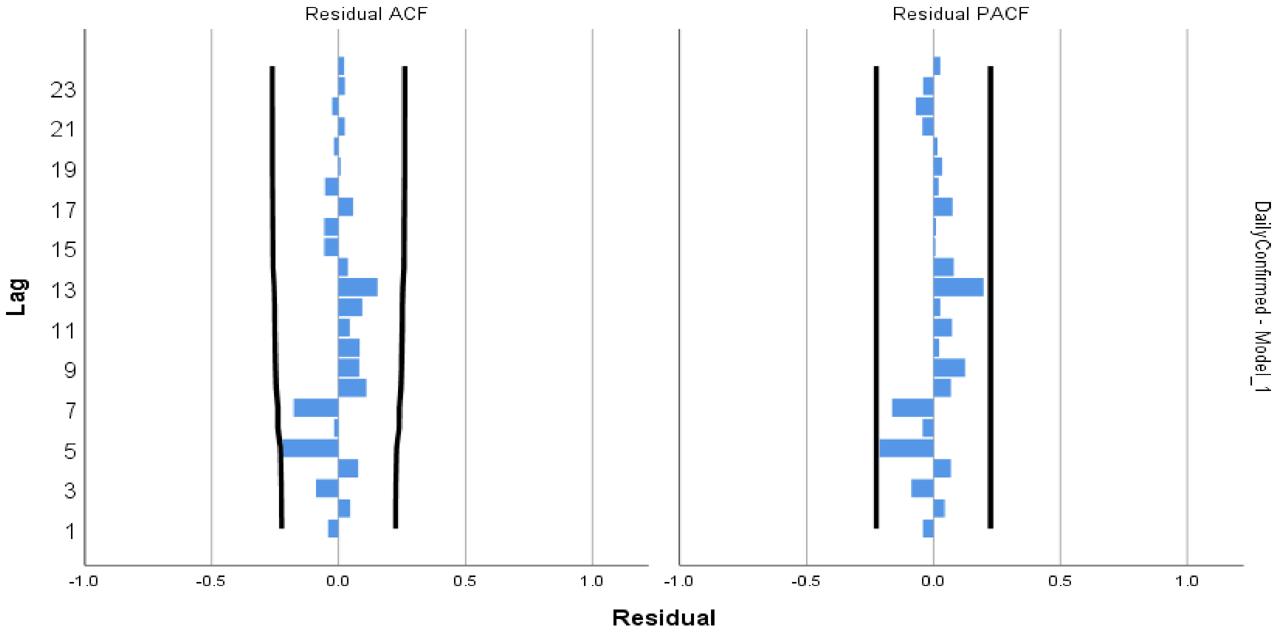
MODEL # 03 : Brownian Exponential Smoothing with Linear Trend Assumption



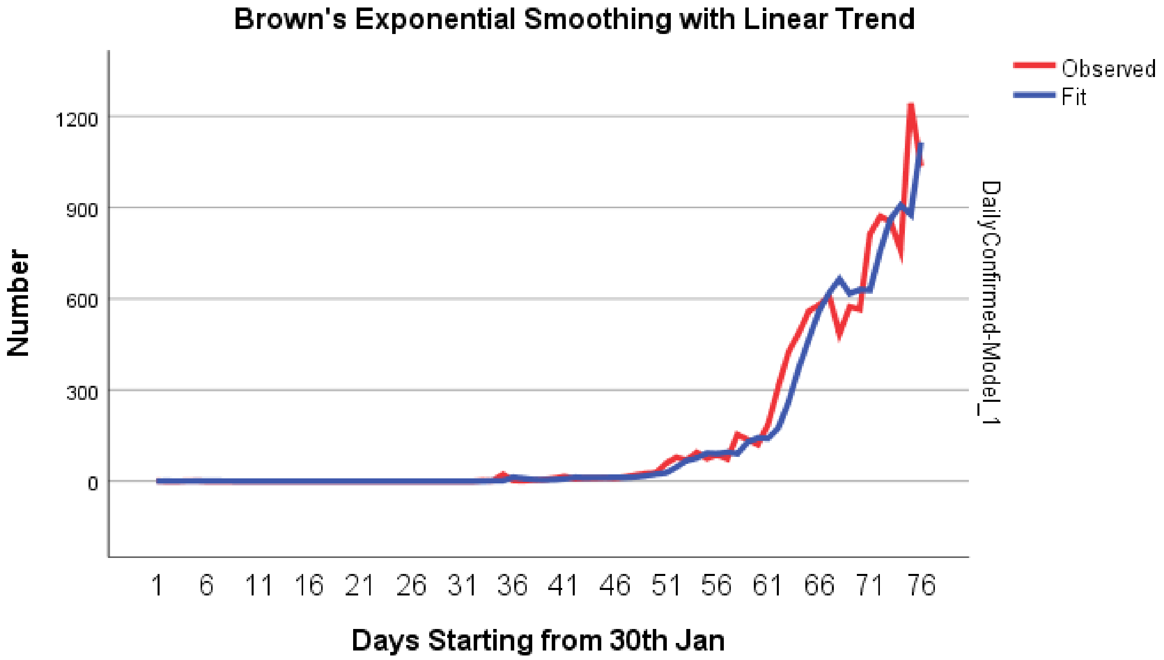
Model Statistics[a]

Model	Number of Predictors	Model Fit statistics				Ljung-Box Q(18)		
		Stationary R-squared	MAPE	MAE	Normalized BIC	Statistics	DF	Sig.
DailyConfirmed-Model_1	0	.689	55.753	27.446	8.438	15.067	17	.591

a. Best-Fitting Models according to Normalized BIC (smaller values indicate better fit).

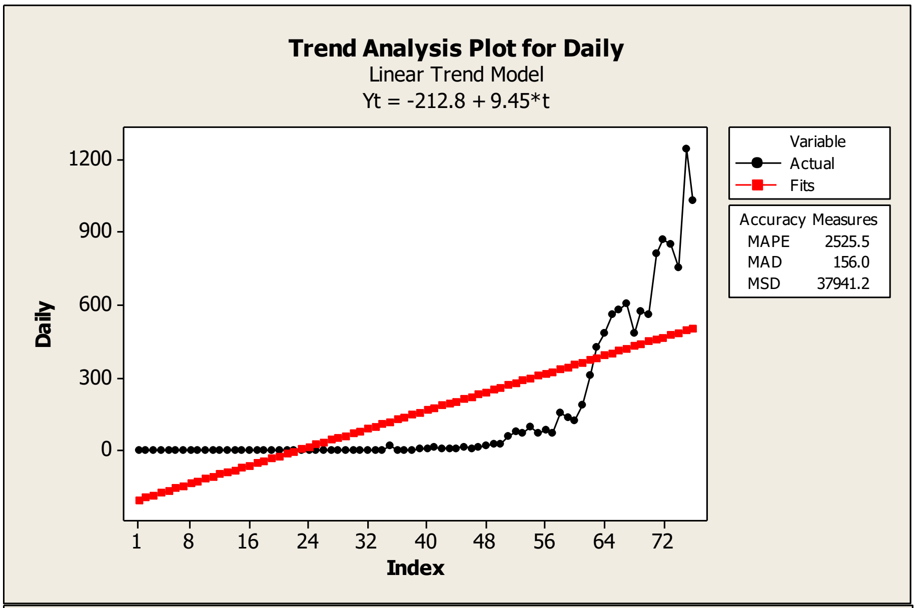


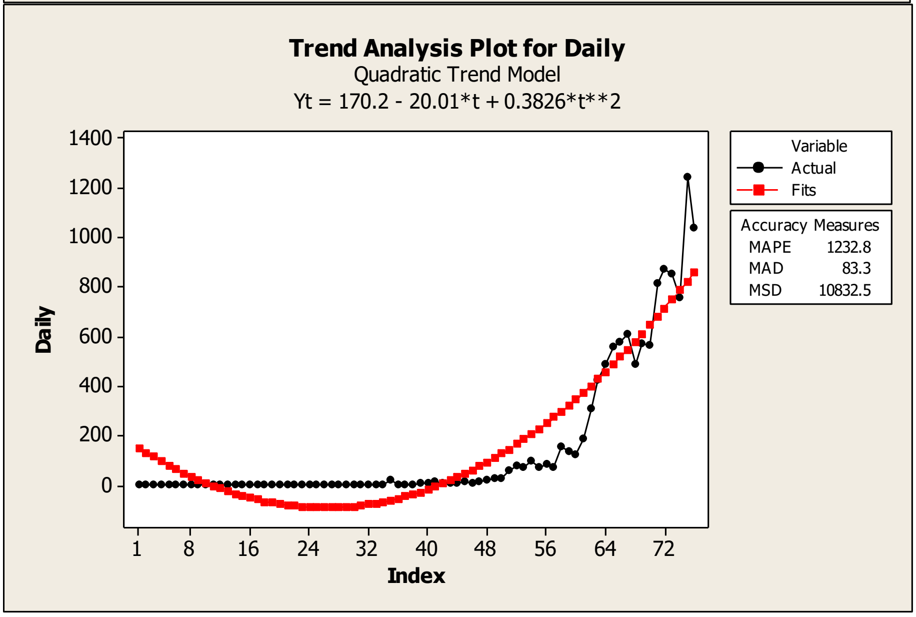
Brownian Exponential Smoothing Method in the presence of a Linear Trend assumes giving exponentially decreasing weights to the past observations, and thereby gives bigger weights to the next to current observations.



But since it assumes presence of a linear trend, it won't be feasible to apply this model even if it has the least BIC value, as the given trend is more quadratic. Moreover, Brown's Exponential Smoothing is equivalent to ARIMA (0,2,2), which means we can always plot it as an effective alternative.

The details of Linear Trend Fit in the Series are below:

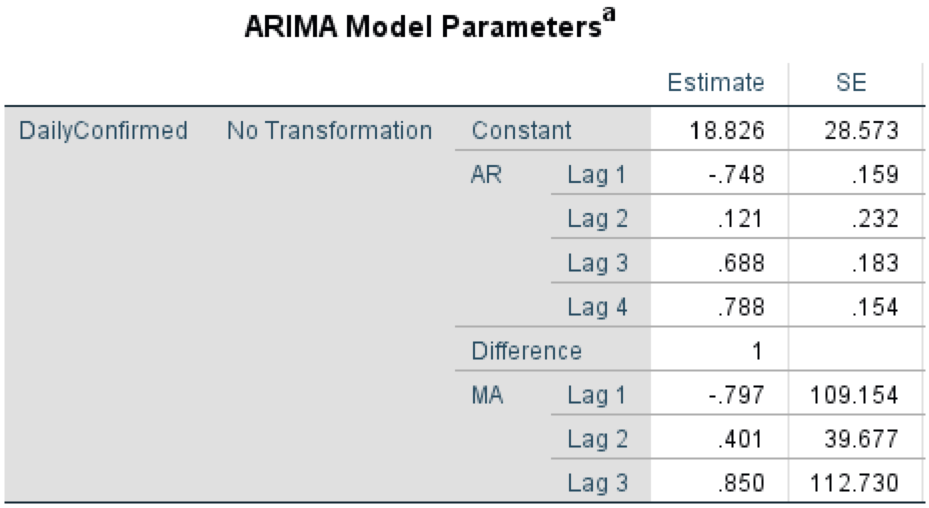




BEST MODEL SELECTION AND MODEL COMPARISON

MODEL	STATIONARY R SQUARED	NORMALISED BIC
ARIMA (4,1,3)	0.516	8.469
ARIMA (0,2,2)	0.689	8.538
ARIMA (4,1,4)	0.514	8.5545
ARIMA (1,2,2)	0.700	8.576

Hence, with the selection of Least BIC value, we can select the best fit model.
Best Fit Model is ARIMA (4,1,3)



ARIMA Model Parameters[a]

				Estimate	SE
DailyConfirmed	No Transformation	Constant		18.826	28.573
		AR	Lag 1	-.748	.159
			Lag 2	.121	.232
			Lag 3	.688	.183
			Lag 4	.788	.154
		Difference		1	
		MA	Lag 1	-.797	109.154
			Lag 2	.401	39.677
			Lag 3	.850	112.730

The equation of the model is:

$X_t = 18.826 + \{ 0.252* X_{(t-1)} + 0.869* X_{(t-2)} + 0.547* X_{(t-3)} + 0.1*X_{(t-4)} - 0.788*X_{(t-4)} \} + \{ e_t + 0.797*e_{t-1} - 0.401*e_{t-2} - 0.850*e_{t-3} \}$

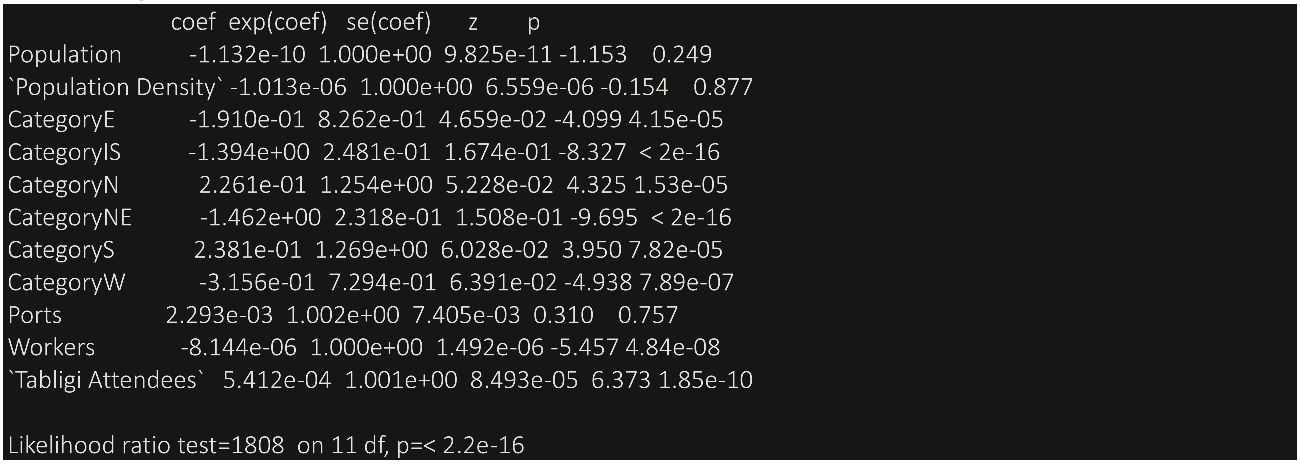
E. EFFECT OF DIFFERENT FACTORS ON THE HAZARD OF CONFIRMED INFECTIONS

COVARIATES

1. **Population** – population of the corresponding state as per the Census 2011 India
2. **Population Density** – population density of the corresponding state as per the Census 2011 India
3. **Category** – The whole country is divided into 7 parts on the basis of geographical location
   a. North India
   b. South India
   c. Central India
   d. East India
   e. West India
   f. North-East India
   g. Islands
4. **Ports** – Number of international airports & seaports in the corresponding states
5. **Workers** – Number of people working overseas in Arab countries from the corresponding state
   (2017 figures)
6. **Tablighi Attendees** – Number of people from the corresponding states who attended the Tablighi Jamaat event in Delhi (based on initial investigation)

Considering the assumptions under the proportional hazard model theory, following cox proportional models are generated;

MODEL – 1 [Population + Population Density + Category + Ports + Workers + Tabligi Attendees]



```
                      coef  exp(coef)   se(coef)      z       p
Population         -1.132e-10  1.000e+00  9.825e-11 -1.153    0.249
`Population Density` -1.013e-06  1.000e+00  6.559e-06 -0.154    0.877
CategoryE          -1.910e-01  8.262e-01  4.659e-02 -4.099 4.15e-05
CategoryIS         -1.394e+00  2.481e-01  1.674e-01 -8.327  < 2e-16
CategoryN           2.261e-01  1.254e+00  5.228e-02  4.325 1.53e-05
CategoryNE         -1.462e+00  2.318e-01  1.508e-01 -9.695  < 2e-16
CategoryS           2.381e-01  1.269e+00  6.028e-02  3.950 7.82e-05
CategoryW          -3.156e-01  7.294e-01  6.391e-02 -4.938 7.89e-07
Ports             2.293e-03  1.002e+00  7.405e-03  0.310    0.757
Workers           -8.144e-06  1.000e+00  1.492e-06 -5.457 4.84e-08
`Tabligi Attendees`   5.412e-04  1.001e+00  8.493e-05  6.373 1.85e-10

Likelihood ratio test=1808  on 11 df, p=< 2.2e-16
```

This model includes all the covariates covered in the data. The above summary of the model provides us the p-values corresponding to each covariate, out of the considered covariates we observes that three covariates viz. Ports, Population, Population density are insignificant to the model at 5% level of significance since the p values are greater than 0.05.

MODEL – 2 [Category + Workers + Tabligi Attendees]

```
                     coef  exp(coef)   se(coef)       z        p
CategoryE         -1.625e-01  8.500e-01  3.626e-02  -4.483 7.36e-06
CategoryIS        -1.346e+00  2.604e-01  1.574e-01  -8.549  < 2e-16
CategoryN          2.659e-01  1.305e+00  2.429e-02  10.948  < 2e-16
CategoryNE        -1.414e+00  2.431e-01  1.367e-01 -10.348  < 2e-16
CategoryS          2.685e-01  1.308e+00  2.022e-02  13.280  < 2e-16
CategoryW         -2.760e-01  7.588e-01  1.732e-02 -15.931  < 2e-16
Workers          -7.661e-06  1.000e+00  7.661e-07 -10.001  < 2e-16
`Tabligi Attendees`  5.944e-04  1.001e+00  6.314e-05   9.414  < 2e-16

Likelihood ratio test=1803  on 8 df, p=< 2.2e-16
```

This model drops out the three covariates found as insignificant in the initial model and estimates the cox regression parameters for the remaining three.

MODEL – 3 [Population + Population Density + Category + Workers + Tabligi Attendees + Population * Population Density]

```
                       coef  exp(coef)   se(coef)      z        p
Population           -2.039e-10  1.000e+00  1.487e-10 -1.371 0.170293
`Population Density`   -9.451e-06  1.000e+00  1.216e-05 -0.777 0.436983
CategoryE            -1.572e-01  8.545e-01  5.835e-02 -2.694 0.007050
CategoryIS           -1.325e+00  2.657e-01  1.824e-01 -7.267 3.68e-13
CategoryN             2.966e-01  1.345e+00  9.149e-02  3.242 0.001189
CategoryNE           -1.403e+00  2.458e-01  1.623e-01 -8.647  < 2e-16
CategoryS             3.034e-01  1.354e+00  8.733e-02  3.474 0.000512
CategoryW            -2.604e-01  7.707e-01  7.436e-02 -3.503 0.000461
Workers             -9.148e-06  1.000e+00  1.947e-06 -4.699 2.62e-06
`Tabligi Attendees`     5.469e-04  1.001e+00  7.534e-05  7.259 3.90e-13
Population            7.291e-13  1.000e+00  8.515e-13  0.856 0.391837
*`Population Density`

Likelihood ratio test=1809  on 11 df, p=< 2.2e-16
```

There exists a high degree of prejudice about the effect to population demographic on the spread of infections hence this model is produced in order to test the significance of interaction between Population & Population Density. Here the p-value of the interaction effect is 0.391837 > 0.05 which provides us enough evidence to state the interaction of the concerned two covariates is also insignificant.

Likelihood Ratio Test

1. MODEL – 1 v/s MODEL – 2

```
Analysis of Deviance Table
 Cox model: response is  Surv(data$Days, data$`Daily Count Rise`)
 Model 1: ~ Population + `Population Density` + Category + Workers + `Tabligi          Attendees`
 Model 2: ~ Category + Workers + `Tabligi Attendees`
   loglik Chisq Df P(>|Chi|)
1 -240712
2 -240714 5.171  2   0.07536
```

The LRT test generates a p value of 0.07536 > 0.05 which gives us enough evidence that dropping the covariates viz. Population, Population Density & Ports will not lead to a significant loss of information from the model.

2. MODEL – 2 v/s MODEL – 3

```
Analysis of Deviance Table
 Cox model: response is  Surv(data$Days, data$`Daily Count Rise`)
 Model 1: ~ Category + Workers + `Tabligi Attendees`
 Model 2: ~ Population * `Population Density` + Category + Workers + `Tabligi          Attendees`
   loglik  Chisq Df P(>|Chi|)
1 -240714
2 -240711 5.9084  3    0.1162
```

The LRT test generates a p value of 0.1162 > 0.05 which gives us enough evidence that dropping the covariate corresponding to the interaction effect between Population & Population Density will not lead to a significant loss of information from the model.

FINAL MODEL

```
                    exp(coef) exp(-coef) lower .95 upper .95
CategoryE            0.8500     1.1765    0.7917    0.9126
CategoryIS           0.2604     3.8405    0.1913    0.3545
CategoryN            1.3047     0.7665    1.2440    1.3683
CategoryNE           0.2431     4.1137    0.1860    0.3178
CategoryS            1.3080     0.7645    1.2572    1.3609
CategoryW            0.7588     1.3178    0.7335    0.7851
Workers              1.0000     1.0000    1.0000    1.0000
`Tabligi Attendees`  1.0006     0.9994    1.0005    1.0007

Concordance= 0.603  (se = 0.002 )
Likelihood ratio test= 1803  on 8 df,   p=<2e-16
Wald test            = 1773  on 8 df,   p=<2e-16
Score (logrank) test = 1843  on 8 df,   p=<2e-16
```

The following are the results provided by the cox proportional model fitting:
(The interpretations given assume other covariates at baseline hazard providing the independent effect of the concerned covariate on the hazard)

The Central India is considered as the base line hazard here and the exponential coefficients provide that East India, West India, North-East India, Islands see 15%, 24%, 76%, 74% less chances of infection form COVID-19 respectively, while North India and South India observe a rise in chances of infection by 30% each.

The number of Tablighi attendees also has a positive impact on the hazard of infection with each unit increase in the covariate raises the chances by 0.06%.

While opposite to the expected results the number of workers has a negative impact in the hazard with each 100 units increase leading to a fall of 0.08% in the chances of infection.

E. PREDICTED SPREAD OF COVID 19 IN INDIA USING DIFFERENTIAL EQUATION DETERMINISTIC MODELS

Theory:

The idea behind compartmental models is to divide the host population into a set of distinct classes, according to its epidemiological status. One simple such model is the *SIR* formalism which classifies individuals as *S*usceptible to the disease (*S*), currently *I*nfectious (*I*), and *R*ecovered (*R*). The total size of the host population is then ***N = S+I+R***.

In the classic SIR model, it is usually assumed that the individuals leave the infectious class at a constant rate. Even if this assumption seems the most intuitive, it is not always the most realistic in terms of the duration individuals stay infective.

Here we use the SEIR where the society is broken into 4 compartments, people belonging to the susceptible group have not yet contracted the disease. Once people get infected, they are moved to the pre-infectious group, at this stage the infected person may not have any symptoms and cannot infect others. After some time passes the infected person himself becomes the infectious anyone who contacts this person may get infected and after some more time he may recover and gain immunity so he cannot become infected again. Time it takes to recover may depend on each person but we may take an average amount take it for everyone else.
We use '**r**' to refer to the rate at which a person may be infected and use '**f**' for the rate at which infected person may become infectious.

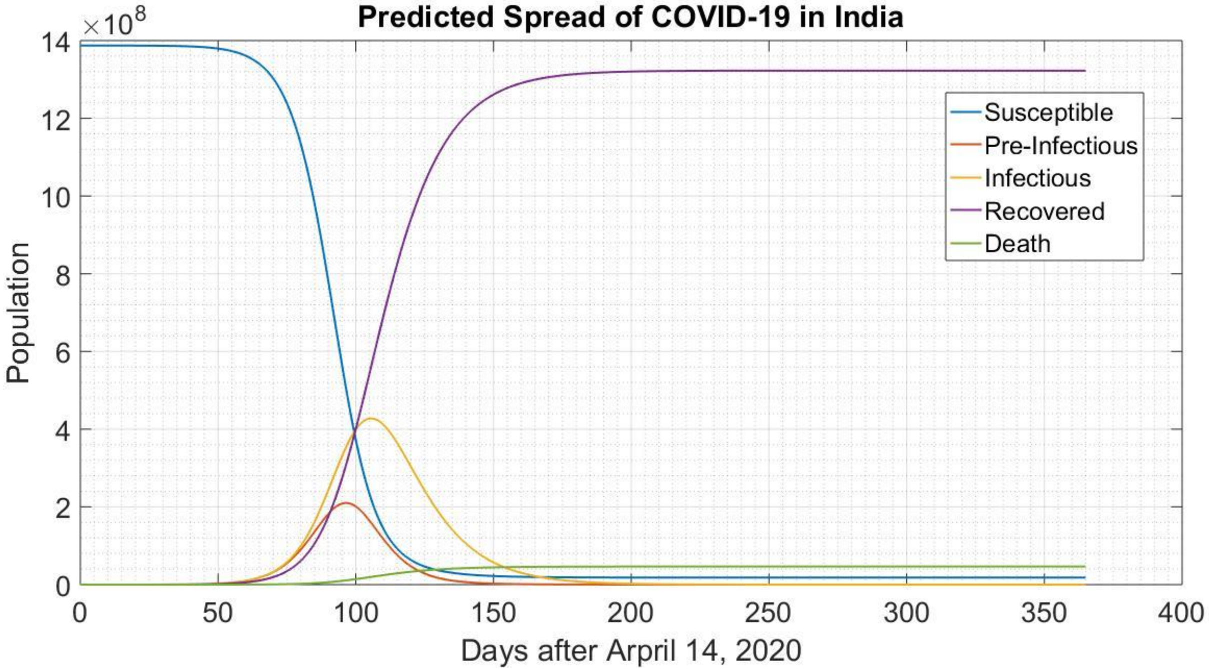
Now we use differential equations to show how each compartment changes over time.

$dS(t)/dt= -\lambda(t)*S(t)$
$dE(t)/dt= \lambda(t)*S(t) – f*E(t)$
$dI(t)/dt= f*(1\text{-Death}) *I(t)$
$dR(t)/dt= (\text{Death})*r*I(t)$
where $\lambda_t=\beta*I_t$
and Death is the death rate of Covid-19

Assumptions:

1. Deterministic
2. Short term
3. Random Mixing

Model:



Prediction of the variables From April 15,2020 to April 21,2020:

Date	Susceptible	Exposed	Infectious	Recovered	Dead
15/04/20	1387282983	2708.178	9320.675	2019.004863	421.0188
16/04/20	1387280053	4900.254	9392.593	2662.255087	443.6591
17/04/20	1387277040	6785.915	9835.032	3323.721187	466.9405
18/04/20	1387273839	8514.249	10580.11	4026.476655	491.6752
19/04/20	1387270363	10191.97	11588.99	4789.803769	518.5417
20/04/20	1387266529	11902.61	12840.46	5631.41599	548.1637
21/04/20	1387262266	13704.72	14332.46	6567.358505	581.1058

Interpretation:

Here we can see a prediction of the situation that would happen if there was no government lockdown to prevent the spread of disease. Now since there are no government policies to prevent interaction and there is a sudden fall in the susceptible people with a rise in the recovered as well as infectious people indicating there would be large increase in infectious population till almost all the population has become immune to the disease. When the susceptible population is about to reach zero the infectious is at maximum and the recovered start to rise and the deaths start to stabilize this indicates that the death rate has become constant and soon the infectious population also becomes zero and the recovered become constant signifying that the whole population has become immune to the disease.