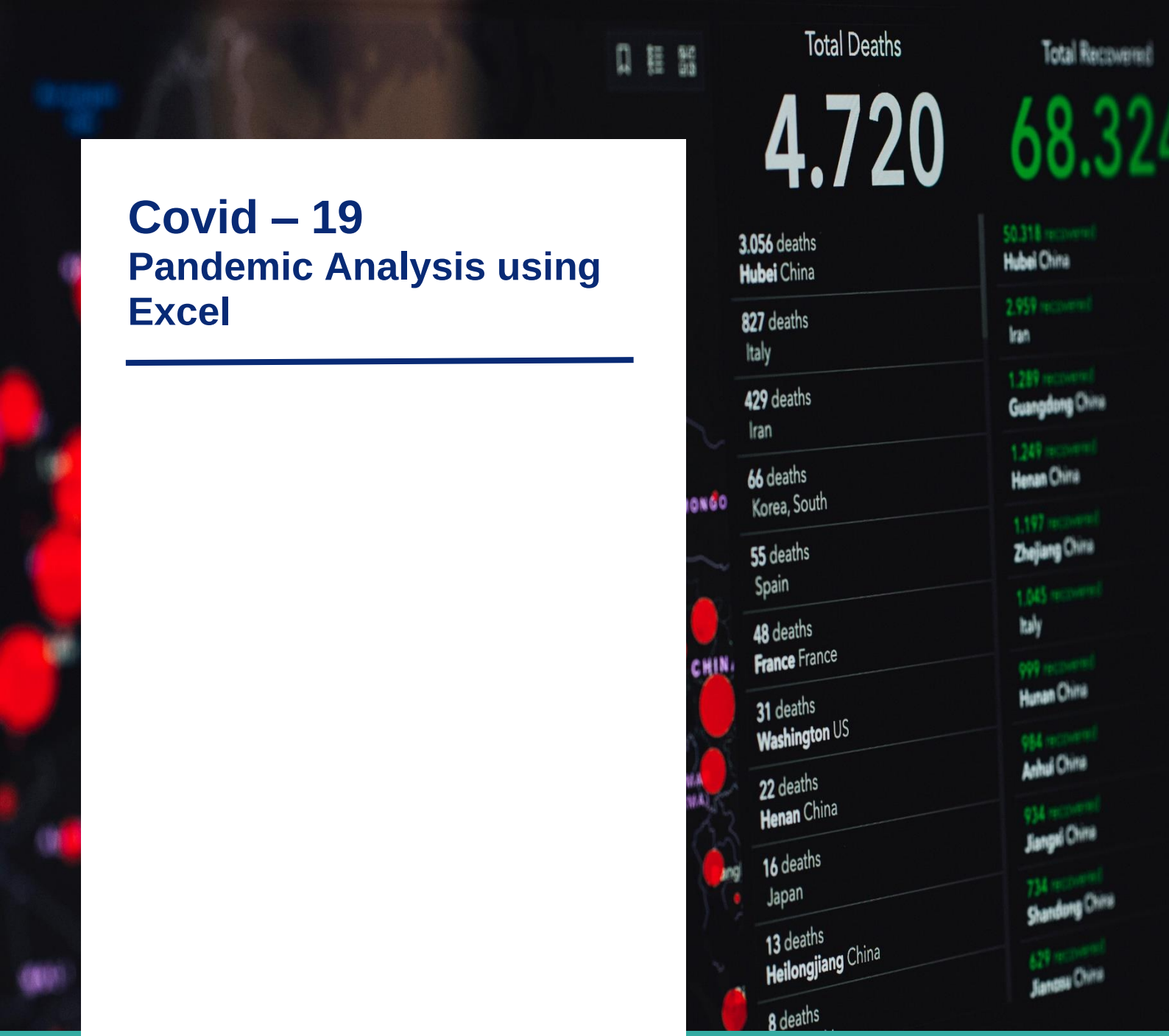


Covid – 19
Pandemic Analysis using Excel

NOVEMBER 29

Nikhil Jain
Roll no. 17
Research and Business Analytics (2021-23)
WeSchool Bengaluru

Guided By:
Prof. Swapnil Soni

Introduction

We all have been affected by the current COVID-19 pandemic. However, the impact of the pandemic and its consequences are felt differently depending on our status as individuals and as members of society. While some try to adapt to working online, homeschooling their children and ordering food via Instacart, others have no choice but to be exposed to the virus while keeping society functioning. Our different social identities and the social groups we belong to determine our inclusion within society and, by extension, our vulnerability to epidemics.

> ***"We are in this together - and we will get through this, together. "***
>
> ***UN Secretary-General Antonio Guterres United Nations***

The nation suffered a setback of COVID-19 pandemic and lost lives in the recent years. However, a few states in India managed well during the pandemic to save lives and livelihood. You are served with a comprehensive database of COVID-19 cases. You, as a data scientist, are asked to perform a detailed analysis to derive useful insights. Apart from basic descriptive (visual) analyses, try answering the following specific and pertinent questions.

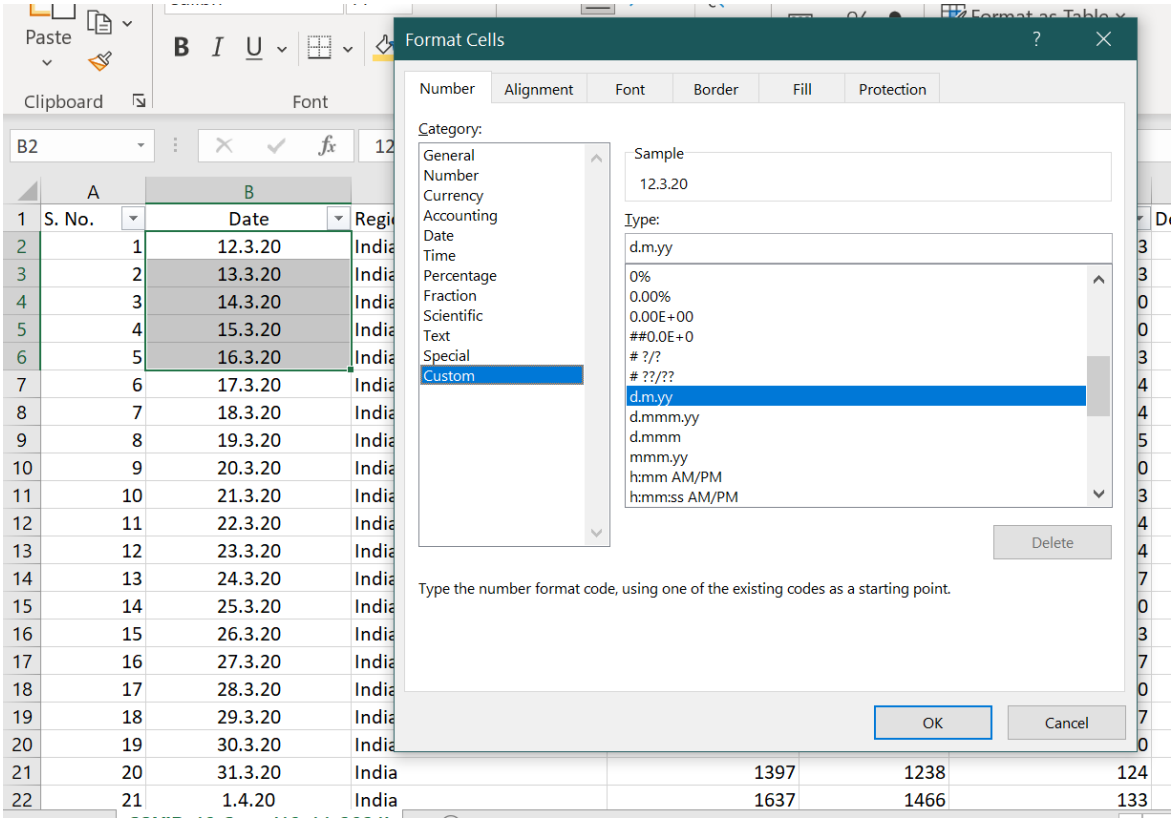
*Data Source: **https://prsindia.org/covid-19/cases***

Data Cleaning

- Deletion of "State assignment pending" data since this data is not relevant for analysis purpose.

	S. No.	Date	Region	Confirmed Cases	Active Cases	Cured/Discharged	Death
8016	18015	12.3.20	State assignment pending	0	0	0	0
8017	18016	13.3.20	State assignment pending	0	0	0	0
8018	18017	14.3.20	State assignment pending	0	0	0	0
8019	18018	15.3.20	State assignment pending	0	0	0	0
8020	18019	16.3.20	State assignment pending	0	0	0	0
8021	18020	17.3.20	State assignment pending	0	0	0	0

- Formatting Date column in “d.m.yy” consistent format for better understanding.



Assumptions:

1. **In Q2, we have filtered out the Union territory from the data because the question has asked only for states data and also comparing states with UT is not an appropriate comparison.**
2. **Data set is taken between from the dates of 12 march 2020 to 17 November 2021 on an average.**
3. **In Q3, we have filtered out the Union territory from the data because the question has asked only for states data and also comparing states with UT is not an appropriate comparison.**
4. **In Q4, because of limited availability of data, it has been assumed that the world depicts the same trend of cases as of India and hence the prediction of world data is done on the basis of India's data.**

Q1. Which states registered highest cases and deaths? Are there outliers?

Confirmed Cases

Row Labels	Max of Confirmed Cases
Andaman and Nicobar Islands	7674
Andhra Pradesh	2070286
Arunachal Pradesh	55230
Assam	614413
Bihar	726161
Chandigarh	65390
Chhattisgarh	1006406
Dadra and Nagar Haveli and Daman and D	10682
Delhi	1440484
Goa	178533
Gujarat	827014
Haryana	771463
Himachal Pradesh	226022
Jammu and Kashmir	334432
Jharkhand	349041
Karnataka	2992276
Kerala	5071135
Ladakh	21229
Lakshadweep	10365
Madhya Pradesh	792981
Maharashtra	6625872
Manipur	124588
Meghalaya	84080
Mizoram	129845
Nagaland	32018
Odisha	1045862
Puducherry	128495
Punjab	602833
Rajasthan	954539
Sikkim	32108
Tamil Nadu	2716421
Telangana	673889
Tripura	84691
Uttar Pradesh	1710288
Uttarakhand	344058
West Bengal	1605794

Reference	Type	What are we looking?	Inference
Figure A	Bar Graph	Highest no. of case	*Maharashtra* has the highest number of confirmed cases.
Figure B	Box Plot	Outlier	4 outliers in the box plot of confirmed cases and hence these are the data point that *differs significantly* from other observations.

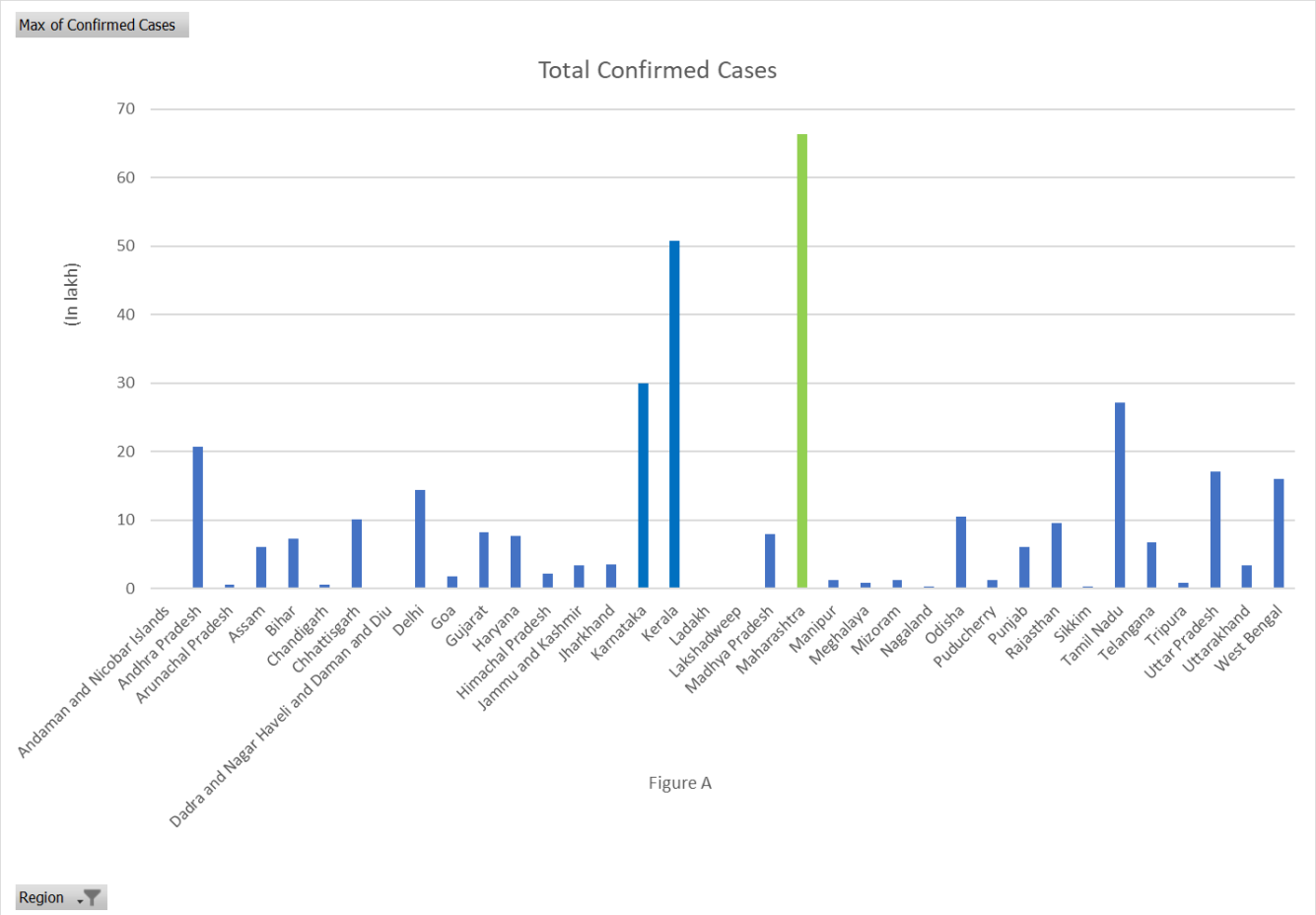


Figure A

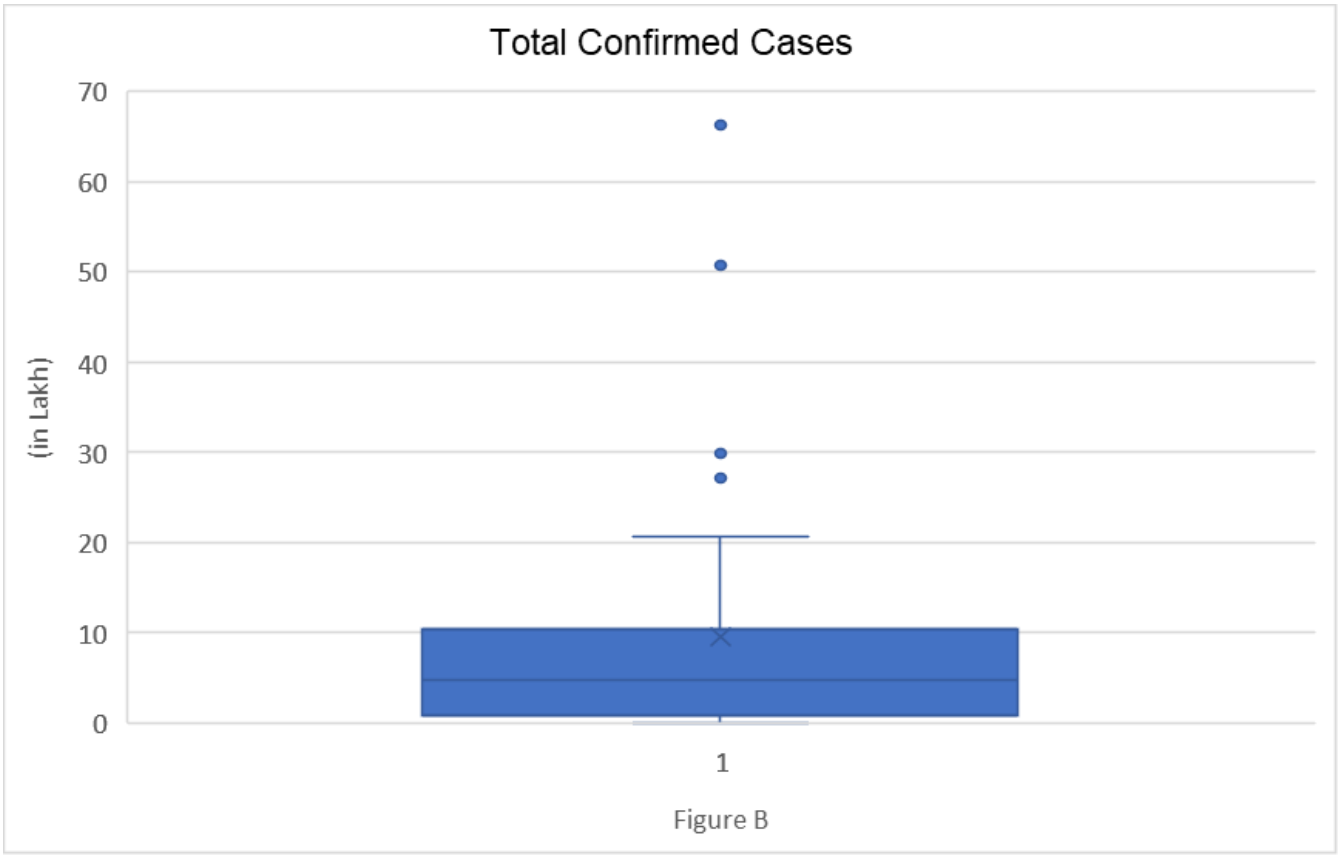


Figure B

Deaths

Row Labels	Max of Death	Outliers
Andaman and Nicobar Islands	129	129
Andhra Pradesh	14418	14418
Arunachal Pradesh	280	280
Assam	6056	6056
Bihar	9662	9662
Chandigarh	820	820
Chhattisgarh	13588	13588
Dadra and Nagar Haveli and Daman and D	4	4
Delhi	25095	25095
Goa	3376	3376
Gujarat	10090	10090
Haryana	10051	10051
Himachal Pradesh	3822	3822
Jammu and Kashmir	4455	4455
Jharkhand	5139	5139
Karnataka	38153	38153
Kerala	36087	36087
Ladakh	211	211
Lakshadweep	51	51
Madhya Pradesh	10525	10525
Maharashtra	140636	140636
Manipur	1953	1953
Meghalaya	1465	1465
Mizoram	467	467
Nagaland	695	695
Odisha	8381	8381
Puducherry	1866	1866
Punjab	16573	16573
Rajasthan	8954	8954
Sikkim	401	401
Tamil Nadu	36311	36311
Telangana	3976	3976
Tripura	817	817
Uttar Pradesh	22909	22909
Uttarakhand	7404	7404
West Bengal	19333	19333
Grand Total	**140636**	

Calculating Outliers with Quartile Method	
Q1	819.25
Q3	13795.5
IQR (Q3-Q1)	12976.25
IQR*1.5	19464.375
Lower Limit (Q1-1.5*IQR)	-18645.125
Upper Limit (Q3+1.5*IQR)	33259.875

Reference	Type	What are we looking?	Inference
Figure C	Bar Graph	Highest no. of death case.	*Maharashtra* has the highest number of death cases.
Outlier Table	Table	Outlier	The upper limit as per table of quartile is 33259.875 and hence there are 4 values above that (highlighted with orange color) which are outliers, hence these are our data point that *differs significantly* from other observations.

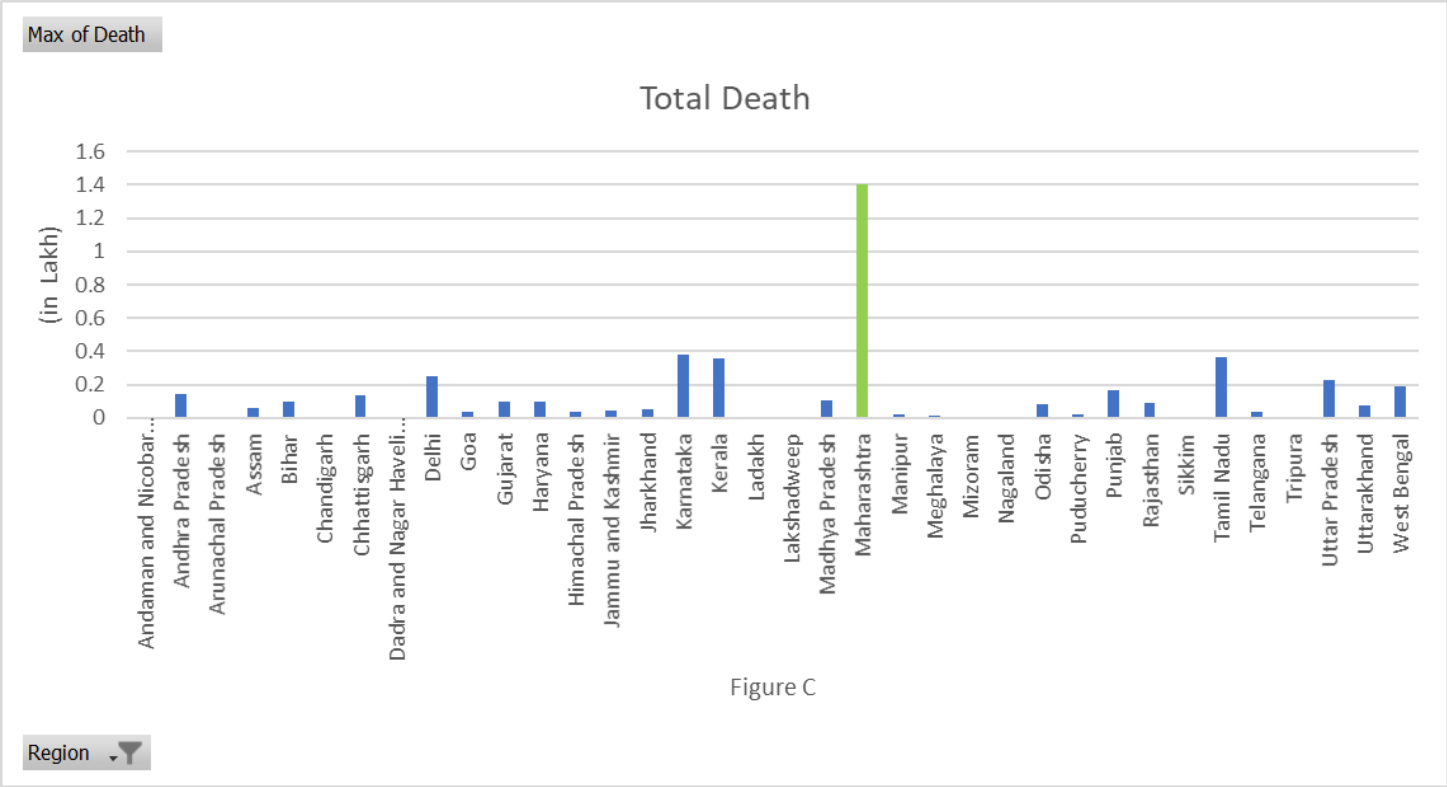

Max of Death
Total Death
(in Lakh)
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Andaman and Nicobar...
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
Dadra and Nagar Haveli...
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu and Kashmir
Jharkhand
Karnataka
Kerala
Ladakh
Lakshadweep
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Telangana
Tripura
Uttar Pradesh
Uttarakhand
West Bengal
Region

Figure C

Q2. Which states are best in recovery rates? Are there outliers?

Row Labels	Average of Recovery rate	Average RI
Andhra Pradesh	80%	80%
Arunachal Pradesh	81%	81%
Assam	83%	83%
Bihar	83%	83%
Chhattisgarh	81%	81%
Goa	82%	82%
Gujarat	81%	81%
Haryana	83%	83%
Himachal Pradesh	79%	79%
Jharkhand	81%	81%
Karnataka	77%	77%
Kerala	79%	79%
Madhya Pradesh	81%	81%
Maharashtra	76%	76%
Manipur	79%	79%
Meghalaya	78%	78%
Mizoram	73%	73%
Nagaland	80%	80%
Odisha	81%	81%
Punjab	79%	79%
Rajasthan	82%	82%
Sikkim	79%	79%
Tamil Nadu	81%	81%
Telangana	81%	81%
Tripura	83%	83%
Uttar Pradesh	80%	80%
Uttarakhand	79%	79%
West Bengal	80%	80%

RECOVERY RATE

Recovery rate=
Cured Cases/Confirmed Cases

Reference	Type	What are we looking?	Inference
Recovery Rate Table	Table	Highest value of recovery rate in %.	*Assam, Bihar, Haryana and Tripura* have the highest recovery rate of 83%.
Figure D	Box Plot	Outlier	There is one outlier of *Mizoram in the recovery rate*, hence this is our data point that *differs significantly* from other observations.

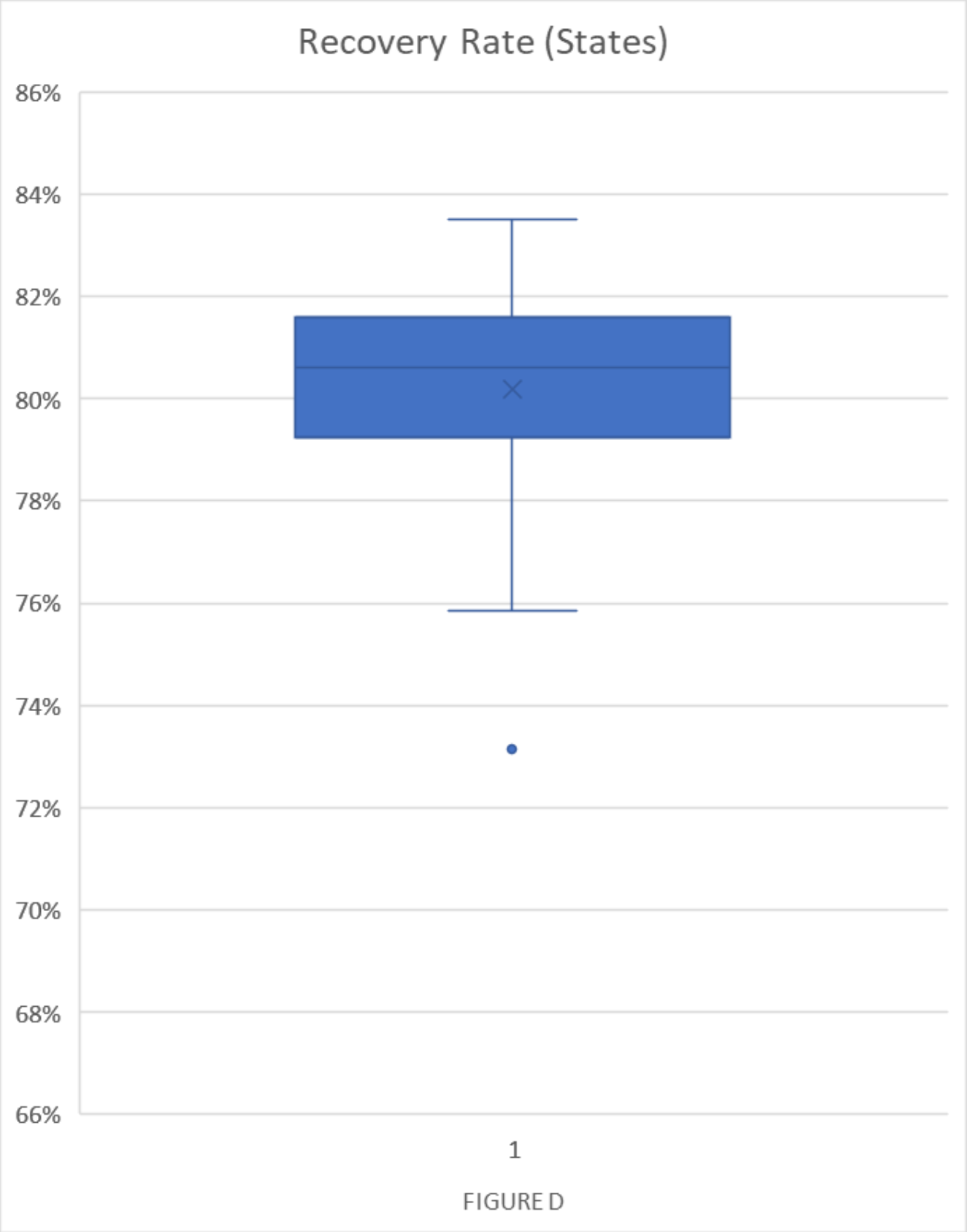


FIGURE D

Q3. Which states have done better in terms of having lower deaths despite higher cases as compared to other states?

DEATH RATE

Row Labels	Max of Dea	Max of Confirmed Cases	Death % on confirmed cas
Andhra Pradesh	14418	2070286	0.696%
Arunachal Pradesh	280	55230	0.507%
Assam	6056	614413	0.986%
Bihar	9662	726161	1.331%
Chhattisgarh	13588	1006406	1.350%
Goa	3376	178533	1.891%
Gujarat	10090	827014	1.220%
Haryana	10051	771463	1.303%
Himachal Pradesh	3822	226022	1.691%
Jharkhand	5139	349041	1.472%
Karnataka	38153	2992276	1.275%
Kerala	36087	5071135	0.712%
Madhya Pradesh	10525	792981	1.327%
Maharashtra	140636	6625872	2.123%
Manipur	1953	124588	1.568%
Meghalaya	1465	84080	1.742%
Mizoram	467	129845	0.360%
Nagaland	695	32018	2.171%
Odisha	8381	1045862	0.801%
Punjab	16573	602833	2.749%
Rajasthan	8954	954539	0.938%
Sikkim	401	32108	1.249%
Tamil Nadu	36311	2716421	1.337%
Telangana	3976	673889	0.590%
Tripura	817	84691	0.965%
Uttar Pradesh	22909	1710288	1.339%
Uttarakhand	7404	344058	2.152%
West Bengal	19333	1605794	1.204%

Reference	Type	What are we looking?	Inference
Death Rate Table	Table	Lowest value of death rate in %.	*Arunachal Pradesh, Mizoram, and Telangana* have the lowest death rate as compared to other states.

Q4. Is India's average recovery rate significantly greater than world average?

Problem:

In the dataset the world active cases and cured/discharged cases were missing and hence without that data calculation of recover rate is not possible because Recover rate = Cured Cases / Confirmed Cases.

Solution:

One possible solution is by using (Assumption 4), i.e because of limited availability of data, it has been assumed that the world depicts the same trend of cases as of India and hence the prediction of world data is done on the basis of India's data.

In order to find the missing data, we use the method of multiple regression model. In the global data, 2 parameters are given i.e., confirmed cases and no. of deaths.

To find the data using regression we considered the confirmed cases of Indian data as 'Y' and we took one unknown and one known variable as 'X', i.e., we took cured cases and deaths as 'X' value. In this case, the death parameter is our known value and cured parameter is our unknown value. We formulated an equation for confirmed cases, deaths and cured cases using multiple regression model. Now, going to data analysis, we used the regression model and gave the relevant inputs keeping zero constant and we got the model given below.

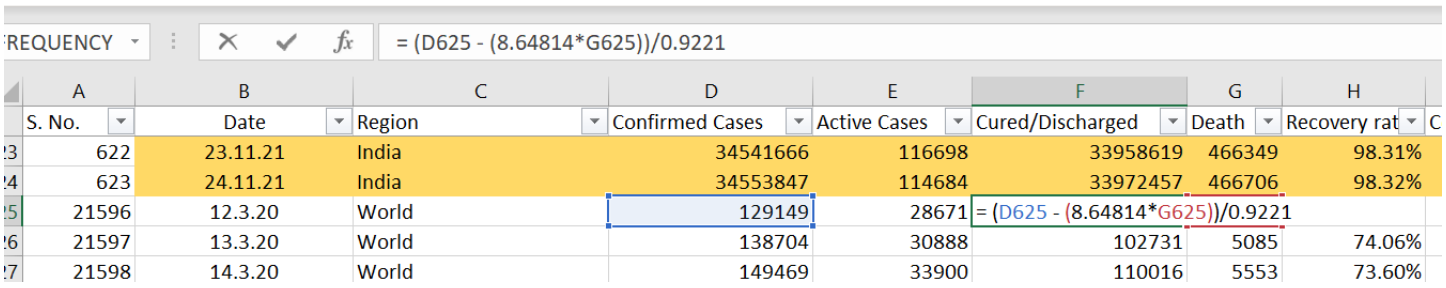
Prediction with Multiple Regression Model:

Multiple Regression model with cured and death as dependent variable and confirmed cases as independent variable.		
SUMMARY OUTPUT		
Regression Statistics		
Multiple R	99.91%	Positive correlation coff.
R Square	99.83%	The proportion of variation in confirmed cases explained by the death and cured cases
Adjusted R Square	0.996647	
Standard Error	802271.6	
Observations	623	

Regression Model

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Significance*
Intercept	0	#N/A	#N/A	#N/A	
Cured/Discharged	0.922101817	0.032984233	27.95583636	0.000000000%	Significant
Death	8.648141647	2.40379561	3.597702571	0.034659583%	Significant

Since the P value is less than 5% of both Cured and death, both values are significant



	A	B	C	D	E	F	G	H
	S. No.	Date	Region	Confirmed Cases	Active Cases	Cured/Discharged	Death	Recovery rat
23	622	23.11.21	India	34541666	116698	33958619	466349	98.31%
24	623	24.11.21	India	34553847	114684	33972457	466706	98.32%
25	21596	12.3.20	World	129149	28671	= (D625 - (8.64814*G625))/0.9221		
26	21597	13.3.20	World	138704	30888	102731	5085	74.06%
27	21598	14.3.20	World	149469	33900	110016	5553	73.60%

Inference of regression model:

From the multiple regression model, we can see the value of R Square is 99.83%. That means that the death and cured values combined together can explain 99.83% of variation in confirmed cases. Also, it is noticed that P-values is well below 5%, so variable is significant.

Regression Model Equation
Confirmed Cases = 0.9221*cured + 8.64814*death Thus cured = (Confirmed cases - 8.64814*death)/0.9221

Now we can put this equation in the global data and find the relevant values. In global data, we have been given the data for confirmed cases and deaths. We need to find the cured cases.

India Average Recovery Rate	World Average Recovery Rate
79.37%	79.30%

From the table, we can infer that, the recovery rate of India and World are very similar probably because we created the regression model using the India level data. And the equation was formulated using that. That is why, there is no significant difference in recovery rate of both India and World.

Q5. Is there any association between no. of cases and deaths? Can we predict no. of deaths based on the no. of cases across states? Does this relationship change over time (i.e., first wave vs second wave)?

Problem:

In the given data both the confirmed cases and no. of deaths are cumulative data i.e., they are increasing values. So, as the confirmed cases increase, the no. of death also increases.

Solution:

To find the relation, we used scatter plot of the whole data and then we noted the equation. But as the no. of cases and no. of deaths vary from state to state we need to find the equation of each individual state & UT. So, we used the filter function and found the equation for each individual state.

Let's take one such case of Delhi as an example

In the Figure H the equation is found to be 'y = 0.071x'.

This is the relation between no. of deaths and no. of confirmed cases for Delhi. Similarly, we found out the relation of no. of deaths and confirmed cases for all the states.

Deaths Vs Confirmed Caases for Delhi

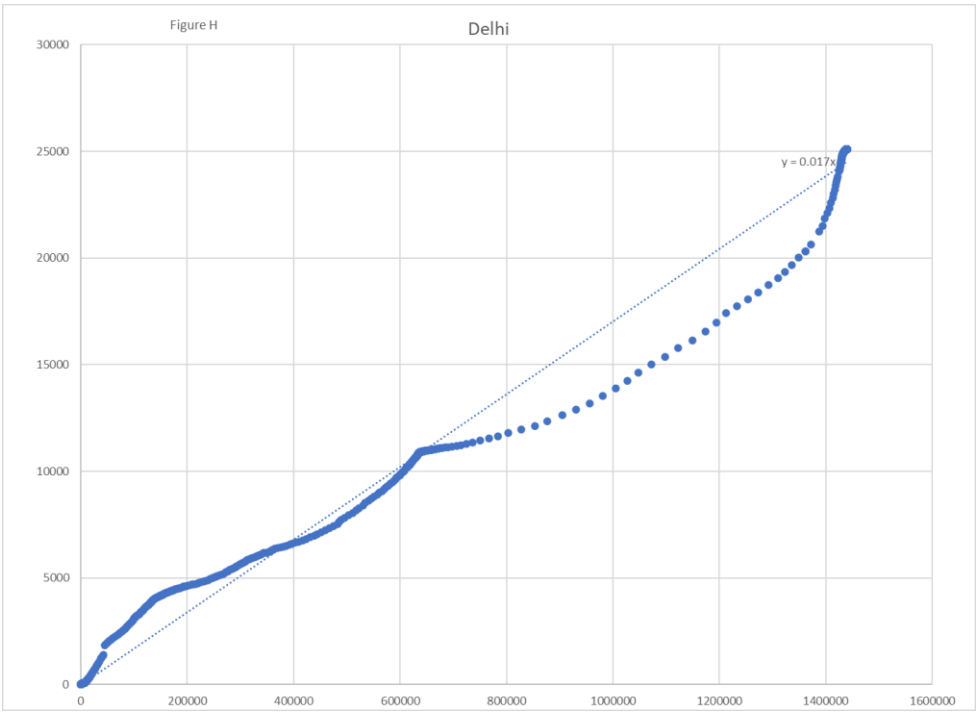


Table for relation between actual no. of cases and no. of deaths

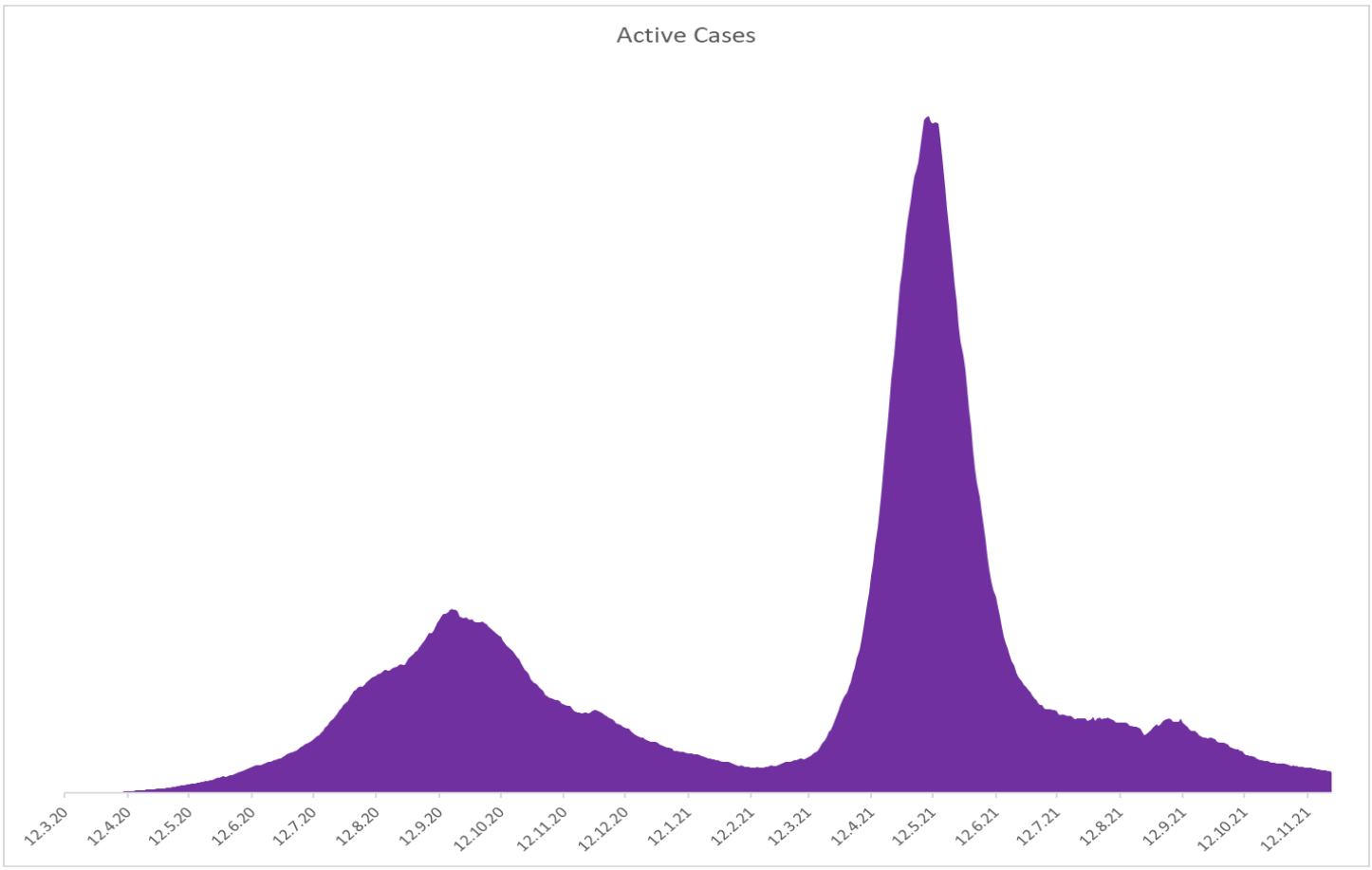
Region	Equation	Correlation Between actual no of cases and death	Correlation Between Actual and Calculated Deaths
Andhra Pradesh	y = 0.0071x	99.01%	99.44
Arunachal Pradesh	y = 0.0047x	99.87%	99.01
Assam	y = 0.0086x	99.86%	97.72
Bihar	y = 0.0113x	89.63%	94.5
Chhattisgarh	y = 0.0132x	99.53%	99.83
Delhi	y = 0.017x	97.78%	99.48
Goa	y = 0.0178x	99.28%	99.47
Gujarat	y = 0.0127x	99.92%	98.94
Haryana	y = 0.012x	96.71%	99
Himachal Pradesh	y = 0.0167x	99.60%	99.88
Jharkhand	y = 0.014x	98.74%	98.89
Karnataka	y = 0.0124x	99.91%	99.54
Kerala	y = 0.0052x	98.35%	98
Madhya Pradesh	y = 0.0125x	99.89%	98.39
Maharashtra	y = 0.0203x	94.05%	98.28
Manipur	y = 0.0154x	99.86%	99.75
Meghalaya	y = 0.0168x	99.91%	99.71
Mizoram	y = 0.0035x	99.53%	99.72
Nagaland	y = 0.0179x	99.46%	96.15
Odisha	y = 0.0064x	98.72%	94.7
Punjab	y = 0.0272x	99.09%	99.7
Rajasthan	y = 0.0091x	98.66%	99.54
Sikkim	y = 0.0133x	98.75%	97.98
Tamil Nadu	y = 0.0133x	98.67%	99.61
Telangana	y = 0.0058x	98.70%	99.74
Tripura	y = 0.0101x	98.69%	99.37
Uttar Pradesh	y = 0.013x	96.33%	99.35
Uttarakhand	y = 0.0206x	98.57%	99.4
West Bengal	y = 0.0126x	98.56%	98.19

Inference:

We have calculate the no. of death from the formulated equation. This is our predicted death. It is such because we calculated the death value from the formulated equation and because the actual data of no. of death is different from the predicted death, we need to calculate the correlation between the actual no. of death and the predicted death.

If the correlation is around 95% or more, we can say that the predicted no. of death is almost similar to the actual no. of death. If the correlation is less, we can say that the predicted death is varying a lot, as it is not similar to the actual no. of death.

Once we find the correlation of predicted death and actual no. of death of each individual state, we can put it in tabular format and examine the data. From the above table, we see that most of the states have a correlation of around 98% or 99% excluding some states like Bihar and Odisha where the correlation is between 94.5% to 94.7%. So, considering all the states, we can say that we can predict the no. of deaths based on the no. of cases across states.



1st wave vs 2nd wave:

Above area chart clearly shows us the effects of 2 waves in India, and by looking at chart it has been assumed that 1st wave ranges from 10th July 2020 to 12th nov 2020 and 2nd wave ranges from 11th Mar 2021 to 17th July 2021.

In 1st wave and 2nd wave, the relation between the no. of confirmed cases and no. of deaths will vary due to sudden change in the respective values. We calculated the correlation during the specific duration of 1st wave and 2nd wave and found the data given below (Figure G). From the data, we can infer that, there was a more sudden increase in confirmed cases and no. of deaths during 2nd wave as compared to 1st wave. Due to large variation during the 2nd wave, the correlation is less w.r.t the 1st wave of COVID-19.

Correlation between no. of confirmed case and no. of death

Region	1st Wave	2nd Wave
India	99.93%	97.91%

FIGURE G

Q6. Records for a few dates are missing. Predict them for India-level data.

S. No.	Date	Region	Confirmed Cases	Active Cases	Cured/Discharged	Death	India Date Ch
280	17.12.20	India	9956557	322366	9489740	144451	2
295	3.1.21	India	10323965	247220	9927310	149435	3
318	27.1.21	India	10689527	176498	10359305	153724	2
350	1.3.21	India	11112241	168627	10786457	157157	2
354	6.3.21	India	11192088	180304	10854128	157656	2
360	13.3.21	India	11333728	202022	10973260	158446	2
373	27.3.21	India	11908910	452647	11295023	161240	2
436	30.5.21	India	27894800	2114508	25454320	325972	2
461	25.6.21	India	30134445	612868	29128267	393310	2
486	21.7.21	India	31216337	407170	30390687	418480	2
488	26.7.21	India	31411262	411189	30579106	420967	4
541	18.9.21	India	33417390	340639	32632222	444529	2
587	4.11.21	India	34321025	148579	33712794	459652	2
596	14.11.21	India	34437307	135918	33837859	463530	2

Problem:

There are 598 data points of India Data and few dates are missing between the data. The problem is how to identify which dates are missing and how to calculate the missing data for those dates.

Solution:

1. Took a successive difference of the dates in a new column and the difference greater than 1 are our data points which are missing.

2. Filtered the India Date Checker column and marked the missing date columns with other color.

3. To calculate the missing values, used the average method with previous and successive dates where the value of India Date Checker column is 2 because besides active cases the remaining data is a cumulative data.

4. To calculate the missing values where the value is more than 2 as in the case of 3.1.21 – took the difference of 3.1.21 and 31.12.20 and divided that difference in 3 parts, then to predict the Confirmed cases of 1.1.21 (blue box), added the 31.12.20 value and the calculated difference (red box) and then successively calculated for 2.1.21. (Figure E)

5. Mathematical relation between all 4 variable is like this -
   Confirmed Cases = Active cases + Cured + Death. Thus, to calculate the active cases now, used this simple equation to calculate as we have all 3 variables now. (Figure F)

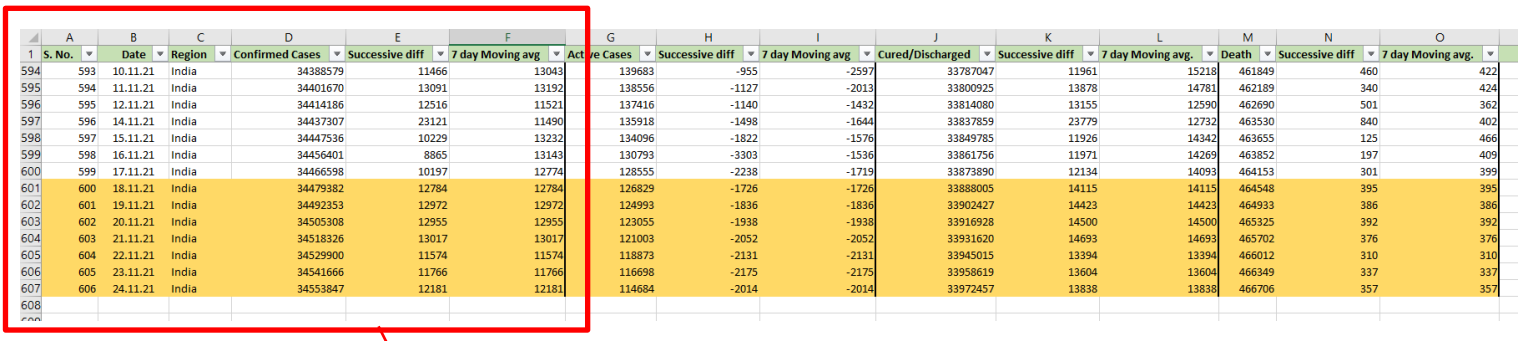
294	293	30.12.20	India	10244852	262272	9834141	148439	1		
295	294	31.12.20	India	10266674	257656	9860280	148738	1		
296		1.1.21	India	=D295+J296						19097
297		2.1.21	India	10304868						
298	295	3.1.21	India	10323965	247220	9927310	149435	3		
299	296	4.1.21	India	10340469	243953	9946867	149649	1		

Figure E

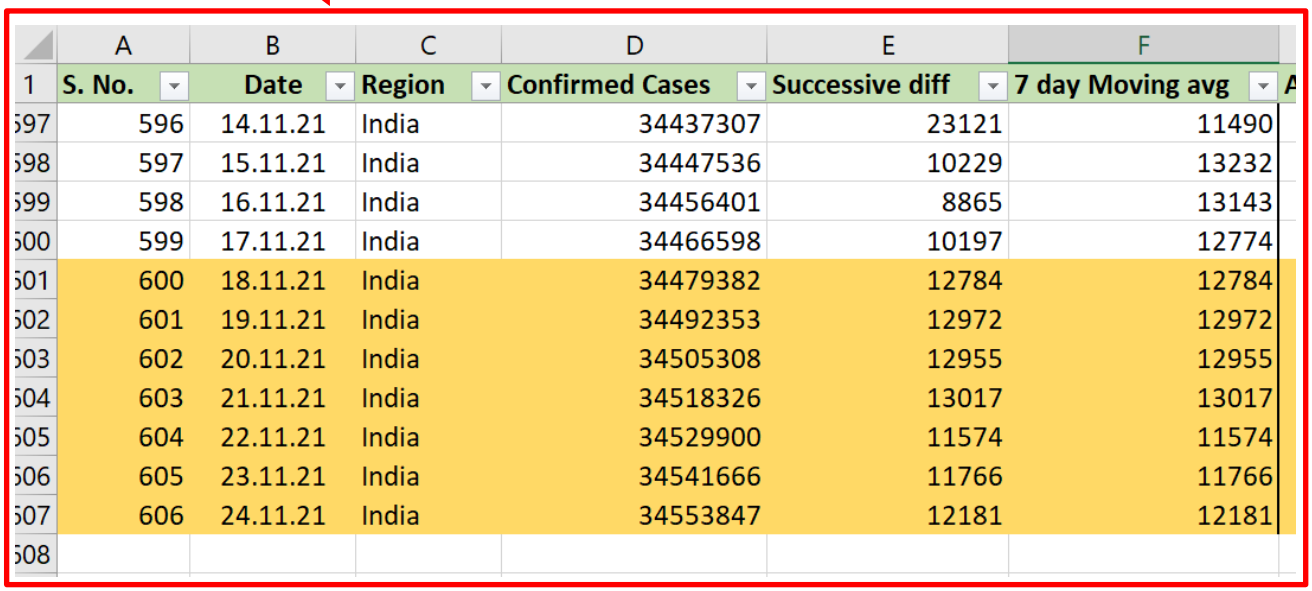
	A	B	C	D	E	F	G	H
1	S. No.	Date	Region	Confirmed Cases	Active Cases	Cured/Discharged	Death	India Date Cl ker
294	293	30.12.20	India	10244852	262272	9834141	148439	1
295	294	31.12.20	India	10266674	257656	9860280	148738	1
296		1.1.21	India	10285771	254177	9882623	148970	
297		2.1.21	India	10304868	=D297-(F297+G297)		149203	
298	295	3.1.21	India	10323965	247220	9927310	149435	3
299	296	4.1.21	India	10340469	243953	9946867	149649	1

Figure F

Q7. Forecast the number of cases, deaths and cures for India for next week.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	S. No.	Date	Region	Confirmed Cases	Successive diff	7 day Moving avg	Active Cases	Successive diff	7 day Moving avg	Cured/Discharged	Successive diff	7 day Moving avg.	Death	Successive diff	7 day Moving avg.
594	593	10.11.21	India	34388579	11466	13043	139683	-955	-2597	33787047	11961	15218	461849	460	422
595	594	11.11.21	India	34401670	13091	13192	138556	-1127	-2013	33800925	13878	14781	462189	340	424
596	595	12.11.21	India	34414186	12516	11521	137416	-1140	-1432	33814080	13155	12590	462690	501	362
597	596	14.11.21	India	34437307	23121	11490	135918	-1498	-1644	33837859	23779	12732	463530	840	402
598	597	15.11.21	India	34447536	10229	13232	134096	-1822	-1576	33849785	11926	14342	463655	125	466
599	598	16.11.21	India	34456401	8865	13143	130793	-3303	-1536	33861756	11971	14269	463852	197	409
600	599	17.11.21	India	34466598	10197	12774	128555	-2238	-1719	33873890	12134	14093	464153	301	399
601	600	18.11.21	India	34479382	12784	12784	126829	-1726	-1726	33888005	14115	14115	464548	395	395
602	601	19.11.21	India	34492353	12972	12972	124993	-1836	-1836	33902427	14423	14423	464933	386	386
603	602	20.11.21	India	34505308	12955	12955	123055	-1938	-1938	33916928	14500	14500	465325	392	392
604	603	21.11.21	India	34518326	13017	13017	121003	-2052	-2052	33931620	14693	14693	465702	376	376
605	604	22.11.21	India	34529900	11574	11574	118873	-2131	-2131	33945015	13394	13394	466012	310	310
606	605	23.11.21	India	34541666	11766	11766	116698	-2175	-2175	33958619	13604	13604	466349	337	337
607	606	24.11.21	India	34553847	12181	12181	114684	-2014	-2014	33972457	13838	13838	466706	357	357
608															



	A	B	C	D	E	F
1	S. No.	Date	Region	Confirmed Cases	Successive diff	7 day Moving avg
597	596	14.11.21	India	34437307	23121	11490
598	597	15.11.21	India	34447536	10229	13232
599	598	16.11.21	India	34456401	8865	13143
600	599	17.11.21	India	34466598	10197	12774
601	600	18.11.21	India	34479382	12784	12784
602	601	19.11.21	India	34492353	12972	12972
603	602	20.11.21	India	34505308	12955	12955
604	603	21.11.21	India	34518326	13017	13017
605	604	22.11.21	India	34529900	11574	11574
606	605	23.11.21	India	34541666	11766	11766
607	606	24.11.21	India	34553847	12181	12181
608						

Problem:

Have to forecast the next 7 days data but the core problem with the data is besides the active cases column, other columns have cumulative data and hence cannot be predicted with moving average on the same.

Solution:

Calculated the successive difference of confirmed cases, cured, death and active cases and then performed moving average on the successive difference column and predicted the next 7 days data by taking 7 days interval. Then performed again cumulation of data to calculate final values. In active cases column, the calculate moving average values are in negative which is correct because based on the trend, active cases are decreasing.

Inference:

Assuming all factors remain same, and the number of cases follow the same trend there will be an increase of **87,249** confirmed cases from 18.11.21 to 24.11.21 and at the same time active cases will decrease by **13,871**. **98,567** more people will be cured or discharged from the hospitals while unfortunately there will be **2,553** more deaths in India.

THANK YOU