TERM PROJECT

Question 1: Image Recognition using SVD:

SVD is a powerful technique employed for noice reduction, image compression, image recognition etc. In this question we are asked to find the representative image of the 10 images given for 15 different people. With that image we are supposed to recognise the 10 images and report the accuracy.

Procedure:

1. We were given 150 images (15 people, 10 images each). These images were uploaded and stored as n*m*150, where n and m are the dimensions of the image.

2. Data from each image was converted to from an integer format to a double format (stored as im_data).

3. Image data was mean shifted by subtracting the mean of all 150 images. Each image was stored as an image vector of dimension (n*m)*1 (where n and m are 64 each making n*m = 4096) where each pixel represents one data point. We create an image vector X of the size 4096*150 for all images.

4. We have to perform SVD on matrix X($X = U\Sigma V^T$) to find the U matrix. To find U we need to find the eigen vectors of A^TA which is difficult owing to its size 4096*4096.

5. Easier way is to first find V matrix and then evaluate U from that. V matrix is the eigen vector matrix of A^TA(150*150 much smaller). U can be derived as:

$$A^TAv_i = \lambda_i v_i$$

$$AA^T(Av_i) = \lambda_i Av_i$$

$$Let u_i = Av_i$$

$$AA^T u_i = \lambda_i u_i$$

6. The above equations indicate that if v_i is an eigenvector of A^TA with an eigenvalue of λ_i , then $u_i = Av_i$ will be the eigenvector of AA^T with the same eigenvalues.

7. We have arranged the eigenvalues of A^TA in the order of decreasing eigenvalues. As we know that AA^T has 4096*4096 components while ATA has only 150*150 components, we will obtain 4096*150 components by the above said method. These components will correspond to the highest eigenvalues. The components corresponding to the highest eigenvalues will be the most significant ones; this lets us eliminate the remaining components that were not included in 4096*150.

8. These eigenvectors are normalized such that each image vector has a magnitude of 1 and are called eigenfaces.

9. Representative image for each face is obtained by taking the mean of all 10 images for a person (after mean shifting). These representative images are projected on eigenface matrix that will be used for identification purposes.

Prediction:

For prediction an image is taken and converted to a (n*m)*1 vector. The image is projected onto eigenface matrix after mean shifting. It is recognized by finding the representative image that is located at the least Euclidean distance from it.

Result:

We were successfully able to identify 149 out of 150 images (accuracy of **99.33%**).

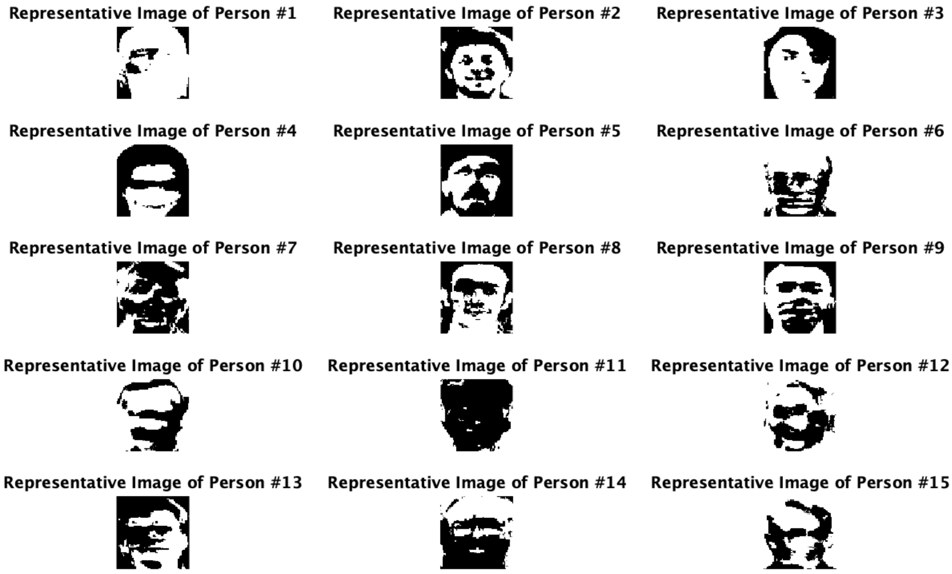

Representative Image of Person #1
Representative Image of Person #2
Representative Image of Person #3
Representative Image of Person #4
Representative Image of Person #5
Representative Image of Person #6
Representative Image of Person #7
Representative Image of Person #8
Representative Image of Person #9
Representative Image of Person #10
Representative Image of Person #11
Representative Image of Person #12
Representative Image of Person #13
Representative Image of Person #14
Representative Image of Person #15

Figure 1: Representative Image for Every Person

Question 2:

1) Describe the statistics of the data.

The dataset given in this question contains information about different parameters of a reactor and their corresponding operating conditions. Here we'll discuss statistics of each parameter individually.

i) Temperature: The temperature of the reactor in kelvin is described here. There are a total of 1000 data points.

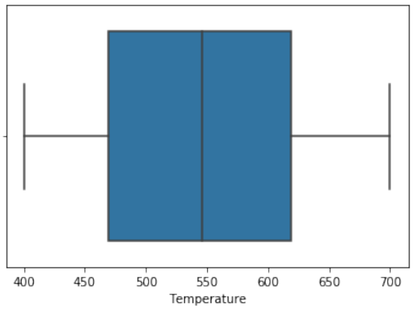


Figure 2: Boxplot of temperature

Mean	546.766
Standard deviation	86.858
Minimum value	400
25 %	469.735
50 %	545.800
75 %	618.877
Maximum value	699.87

Table 1: Statistics of temperature

ii) Pressure: The pressure(bar) in the reactor is described here.

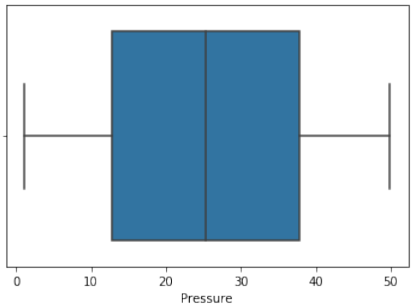


Figure 3: Boxplot of pressure

Mean	25.493
Standard deviation	14.252
Minimum value	1.06
25 %	12.725
50 %	25.375
75 %	37.82
Maximum value	49.89

Table 2: Statistics of Pressure

iii) Feed Flow Rate: The feed flow rate(k-mol/hr) of the reactor is described here.

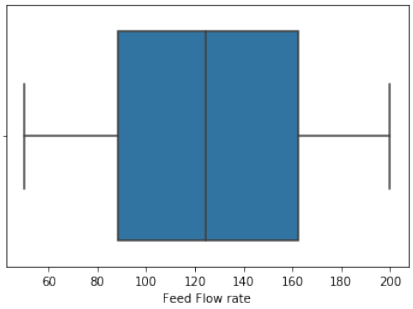


Figure 4: Boxplot of Feed flow rate

Mean	125.029
Standard deviation	43.508
Minimum value	50.03
25 %	88.587
50 %	124.59
75 %	162.562
Maximum value	199.96

Table 3: Statistics of Feed flow rate

iv) Coolant Flow rate: The flow rate of the coolant(L/hr) in reactor is described here.

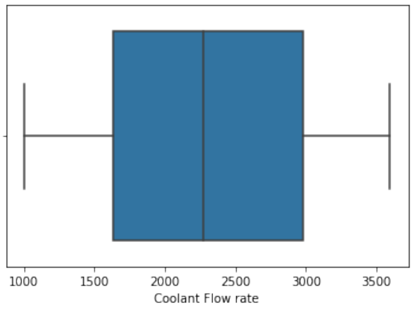


Figure 5: Boxplot of Coolant flow rate

Mean	2295.797
Standard deviation	763.680
Minimum value	1002.53
25 %	1635.682
50 %	2268.710
75 %	2983.692
Maximum value	3595.620

Table 4: Statistics of coolant flow rate

v) Inlet reactant concentration : The concentration of the reactant (mole fraction) at inlet of the reactor is described here.

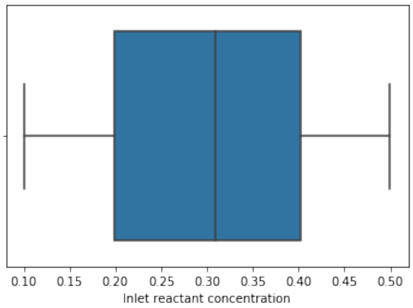


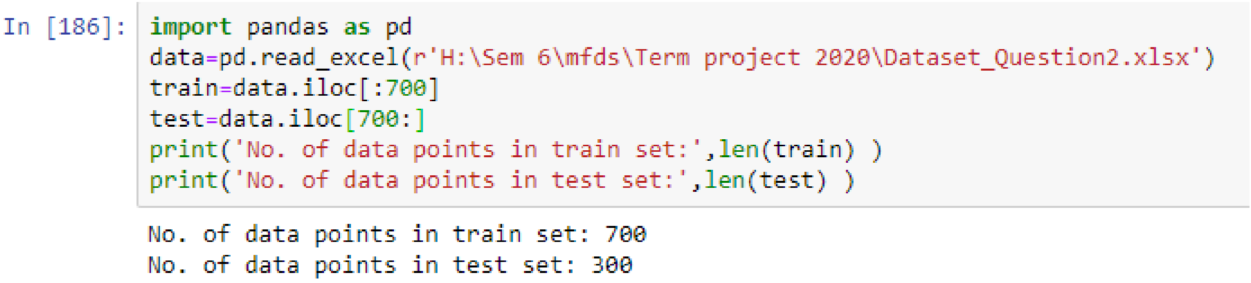
Figure 6: Box plot of Inlet Reactant Concentration

Mean	0.302
Standard deviation	0.116
Minimum value	0.100
25 %	0.199
50 %	0.308
75 %	0.401
Maximum value	0.499

Table 5: Statistics of Inlet reactant concentration

2) Partition your data into a training set and a test set. Keep 70% of your data for training and set aside the remaining 30% for testing.

First we change the categorical variable into numeric. The **Test** variable has two conditions **Pass** and **Fail**. We map **Pass : 0** and **Fail : 1**. Now we partition the data into train and test sets. This can be easily done using **Pandas** package and **iloc** function. The code used to partition data is given below.



```
In [186]: import pandas as pd
          data=pd.read_excel(r'H:\Sem 6\mfds\Term project 2020\Dataset_Question2.xlsx')
          train=data.iloc[:700]
          test=data.iloc[700:]
          print('No. of data points in train set:',len(train) )
          print('No. of data points in test set:',len(test) )

          No. of data points in train set: 700
          No. of data points in test set: 300
```

Figure 7: Python code for splitting data

3) Logistic Regression:

The objective function is the **Sigmoid function**

$$S(z) = \frac{1}{1+e^{-z}}$$

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5$$

The 6 thetas are the parameters to be estimated. θ_o is the bias term. x1 to x5 are the 5 variables terms. If the output value is $>$ 0.5 the estimated value will be 1, and if the output is $<$ 0.5 then the estimated value will be 0.

Error Function:

$$J(\theta) = -\frac{1}{m}\Sigma[y_{(i)}log(h_\theta(x_{(i)})) + (1-y_{(i)})log(1-h_\theta(x_{(i)}))]$$

$$\mathrm{h}_\theta(x) = \frac{1}{1+e^{(\theta_0+\theta_1 x_1+\theta_2 x_2+\theta_3 x_3+\theta_4 x_4+\theta_5 x_5)}}$$

Gradient Descent:

Our aim is to minimize the error function. This can be done by using gradient descent. First we must calculate the gradient of the error function, and then move the point in the opposite direction of the gradient.

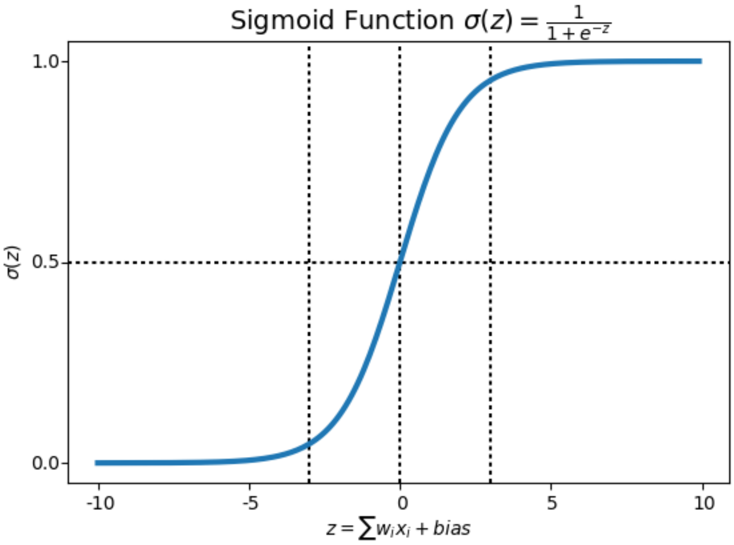


Figure 8: Sigmoid function

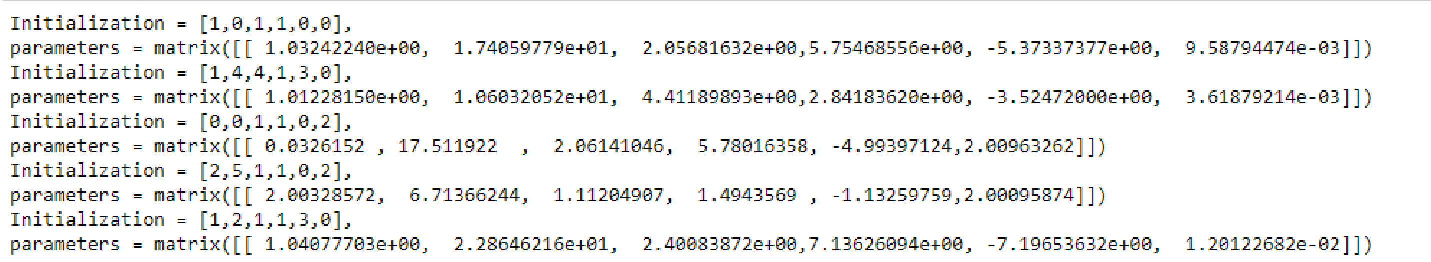
Gradient of the error function:

$$\frac{\partial J(\theta)}{\partial \theta_j} = \Sigma(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

The updated theta values after moving in the negative gradient direction are:

$$\theta_j(updated) = \theta_j(previous) - \alpha\frac{\partial J(\theta)}{\partial \theta_j}$$

Using the above method we have coded for gradient descent(Code attached). By taking different initialization points we get different parameters. Which can be seen below.

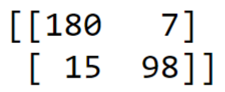


```
Initialization = [1,0,1,1,0,0],
parameters = matrix([[ 1.03242240e+00,  1.74059779e+01,  2.05681632e+00,5.75468556e+00, -5.37337377e+00,  9.58794474e-03]])
Initialization = [1,4,4,1,3,0],
parameters = matrix([[ 1.01228150e+00,  1.06032052e+01,  4.41189893e+00,2.84183620e+00, -3.52472000e+00,  3.61879214e-03]])
Initialization = [0,0,1,1,0,2],
parameters = matrix([[ 0.0326152 , 17.511922  ,  2.06141046,  5.78016358, -4.99397124,2.00963262]])
Initialization = [2,5,1,1,0,2],
parameters = matrix([[ 2.00328572,  6.71366244,  1.11204907,  1.4943569 , -1.13259759,2.00095874]])
Initialization = [1,2,1,1,3,0],
parameters = matrix([[ 1.04077703e+00,  2.28646216e+01,  2.40083872e+00,7.13626094e+00, -7.19653632e+00,  1.20122682e-02]])
```

Figure 9: Initialisation and corresponding parameters

4) Evaluation:

i)Confusion matrix:



```
[[180   7]
 [ 15  98]]
```

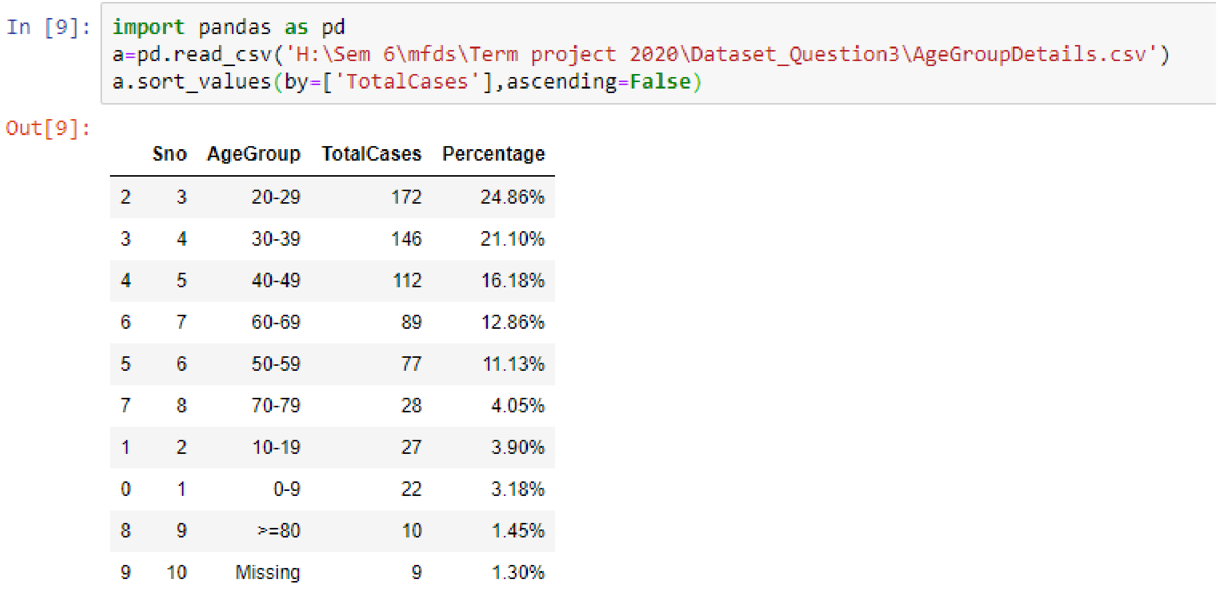
Figure 10: Confusion matrix

ii) F1 Score = 0.90

Question 3:

1) Which age group is the most infected?

From the following piece of python code, we arranged the data in descending order(Big to small) of total cases. So the highest number of cases are displayed in the first row.



```
In [9]: import pandas as pd
        a=pd.read_csv('H:\Sem 6\mfds\Term project 2020\Dataset_Question3\AgeGroupDetails.csv')
        a.sort_values(by=['TotalCases'],ascending=False)
```

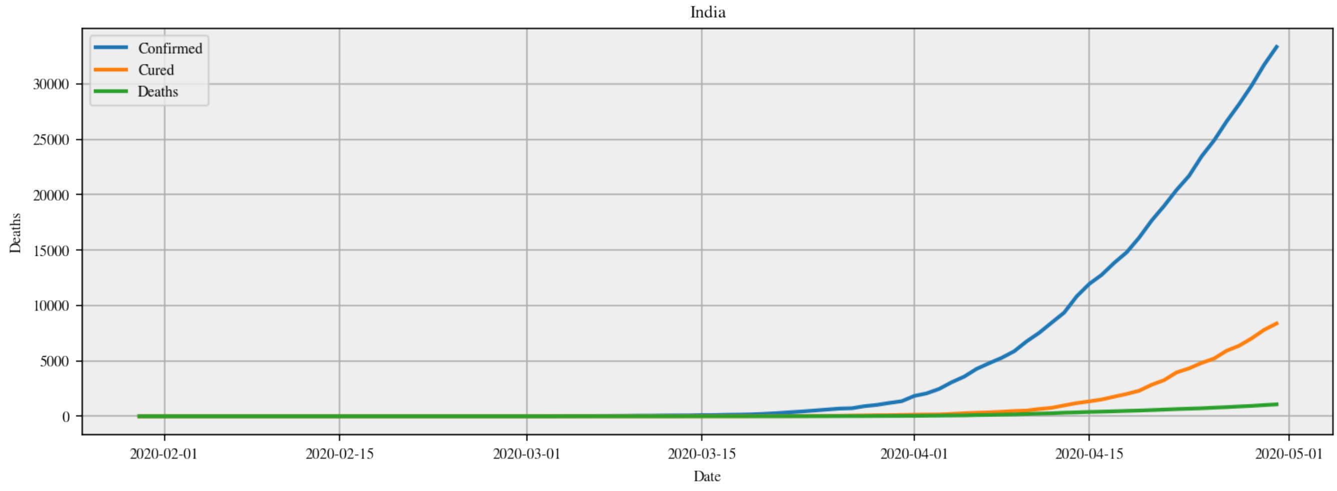
Out[9]:

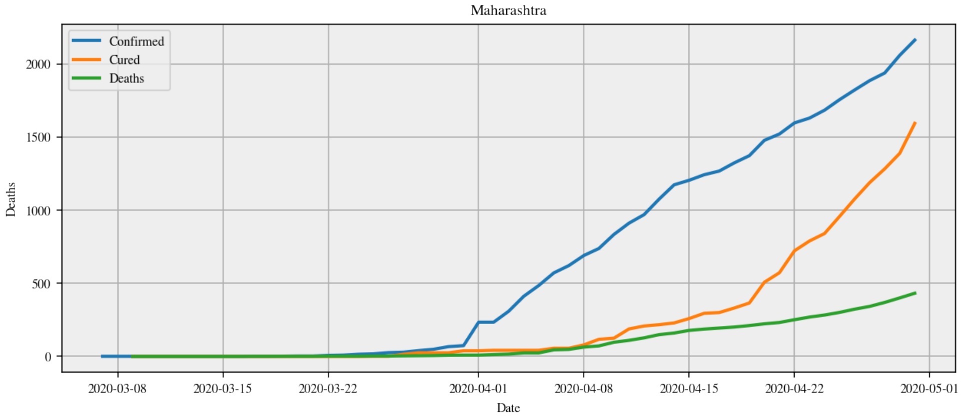
	Sno	AgeGroup	TotalCases	Percentage
2	3	20-29	172	24.86%
3	4	30-39	146	21.10%
4	5	40-49	112	16.18%
6	7	60-69	89	12.86%
5	6	50-59	77	11.13%
7	8	70-79	28	4.05%
1	2	10-19	27	3.90%
0	1	0-9	22	3.18%
8	9	>=80	10	1.45%
9	10	Missing	9	1.30%

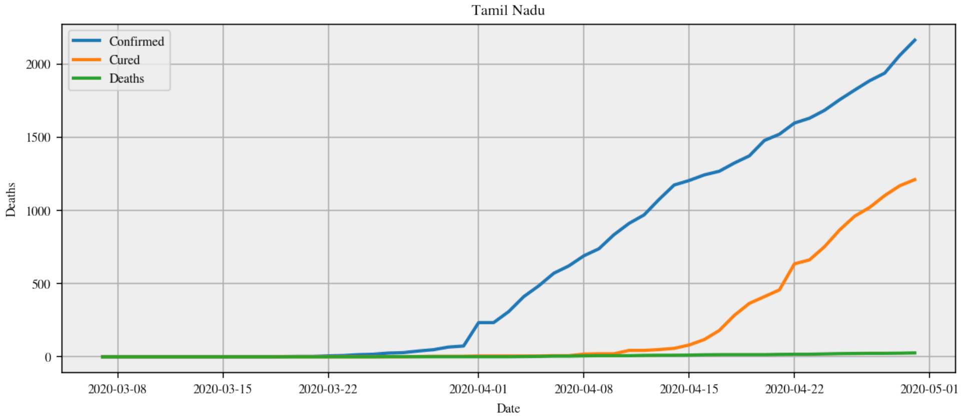
From this we can conclude that the highest affected age group is **20-29**, which is 24.86% of the entire cases.

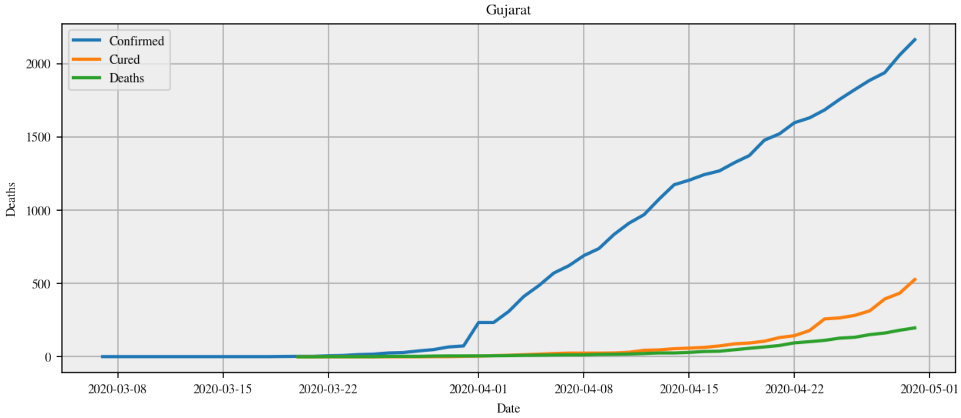
2) Plot graphs of the cases observed, recovered, deaths per day country-wise and statewise.

On analyzing the given data in the covid_19 dataset and converting it into a suitable form, the following graphs of the cases observed, recoveries and deaths per day state-wise have been plotted.







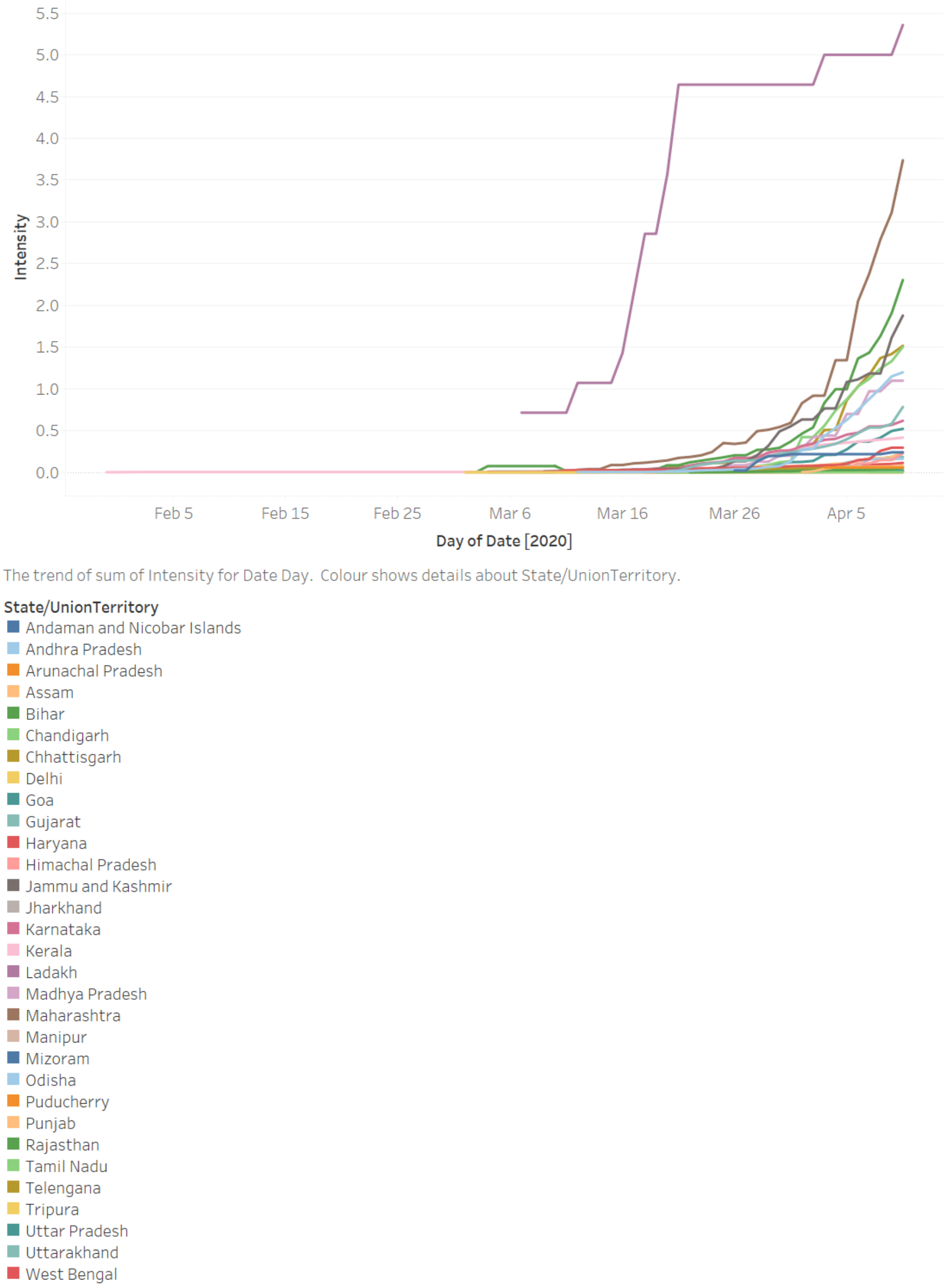


Plots for only three states are given here. Please check the attached python notebook for all states.(**Question2.ipynb**)

3) Identify the positive cases on a state level. Quantify the intensity of virus spread for each state.

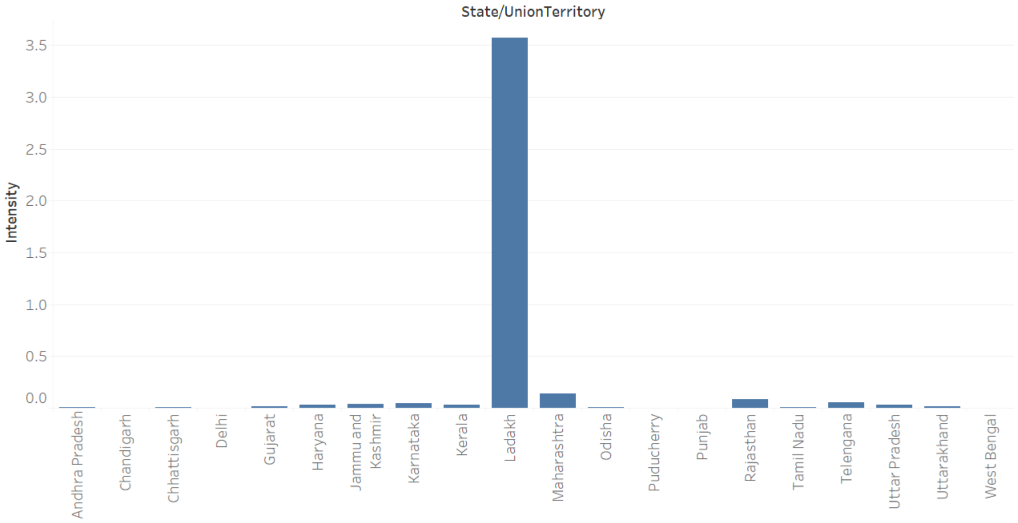
Intensity here means No.of positive cases divided by Population density. By analyzing Population_india_census2011.csv and covid_19 datasets the following visualizations are constructed. Here we have considered cases from 30th Jan to 10th April.

Evolution of Intensity with Time for all the States



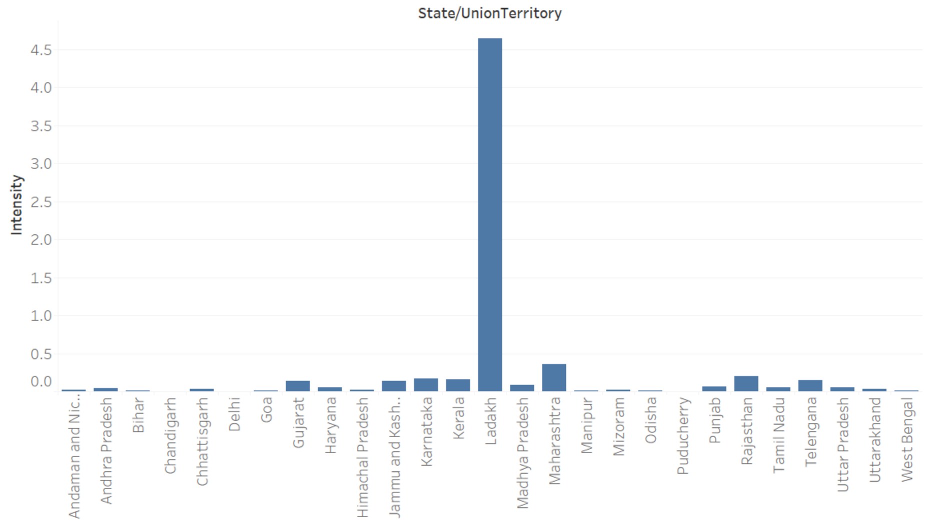
The trend of sum of Intensity for Date Day. Colour shows details about State/UnionTerritory.

Intensity of spread on 20th March



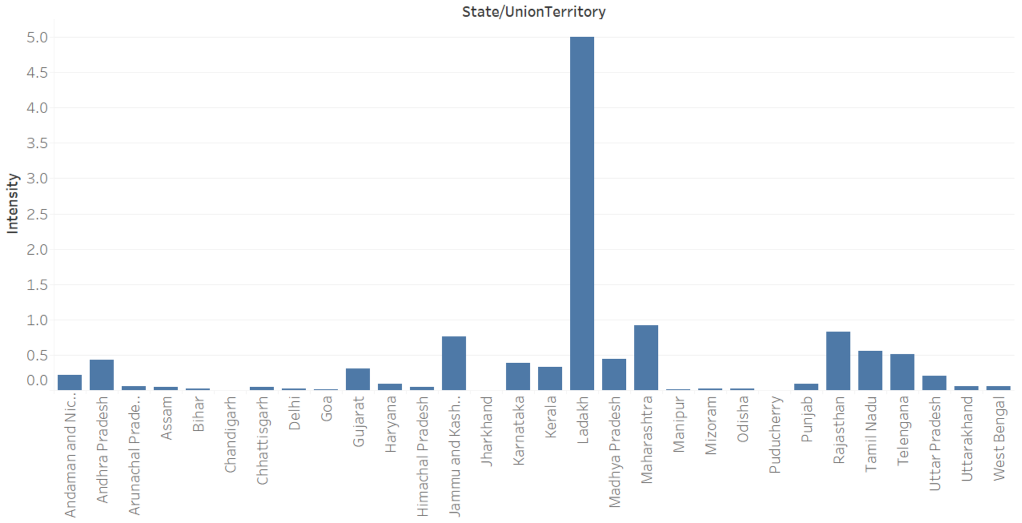
Sum of Intensity for each State/UnionTerritory.

Intensity of spread on 27th March



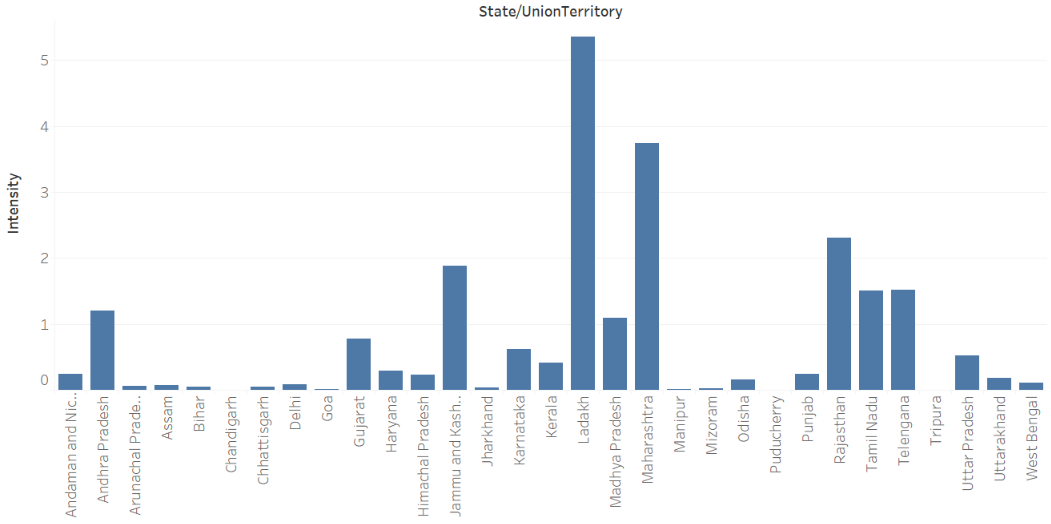
Sum of Intensity for each State/UnionTerritory.

Intensity of spread on 3rd April



Sum of Intensity for each State/UnionTerritory.

Intensity of spread on 10th April



Sum of Intensity for each State/UnionTerritory.

4) List places in the country which are active hotspots/clusters as on 10.04.2020

Assumption : Since there were not enough entries in detected_city column of IndividualDetails dataset, a district having 10 or more active cases has been taken as a hotspot.

The following districts are hotspots:
Adilabad; Agra; Ahmadabad; Ahmadnagar; Akola; Amritsar; Anantapur; Aurangabad; Badgam; Bandipore; Banswara; Baramula; Barwani; Belagavi; Bengaluru; Bhavnagar; Bhilwara; Bhopal; Bidar; Bikaner; Buldana ;Chandigarh ;Chengalpattu ;Chennai ;Chittoor ;Churu ;Coimbatore ;Cuddalore ;Dehradun ;Dindigul ;East Godavari ;Ernakulam ;Erode ;Evacuees* ;Faridabad ;Firozabad ;Gandhinagar ;Gautam Buddha Nagar ;Ghaziabad ;Guntur ;Gurugram ;Hyderabad ;Indore ;Jaipur ;Jaisalmer ;Jalandhar ;Jammu ;Jhalawar ;Jhunjhunu ;Jodhpur ;Jogulamba Gadwal ;Kamareddy ;Kanniyakumari ;Kannur ;Karimnagar ;Karur ;Kasaragod ;Khargone ;Khordha ;Korba ;Kota ;Kozhikode ;Krishna ;Kupwara ;Kurnool ;Leh ;Lucknow ;Ludhiana ;Madurai ;Mahabubnagar ;Malappuram ;Mansa ;Meerut ;Morena ;Mumbai ;Mysuru ;Nagapattinam ;Nagpur ;Nalgonda ;Namakkal ;Nirmal ;Nizamabad ;Nuh ;Palghar ;Palwal ;Patan ;Pathankot ;Prakasam ;Pune ;Rajkot ;Ranga Reddy ;Ranipet ;S.A.S. Nagar ;S.P.S. Nellore ;Saharanpur ;Salem ;Sangli ;Shahid Bhagat Singh Nagar ;Shamli ;Shupiyan ;Sitapur ;Siwan ;South Delhi ;Srinagar ;Surat ;Thane ;Thanjavur ;Theni ;Thiruvallur ;Thiruvarur ;Thoothukkudi ;Thrissur ;Tiruchirappalli ;Tirunelveli ;Tirupathur ;Tiruppur ;Tonk ;Udhampur ;Ujjain ;Una ;Vadodara ;Vellore ;Viluppuram ;Virudhunagar ;Visakhapatnam ;Warangal Urban ;West Godavari Y.S.R.

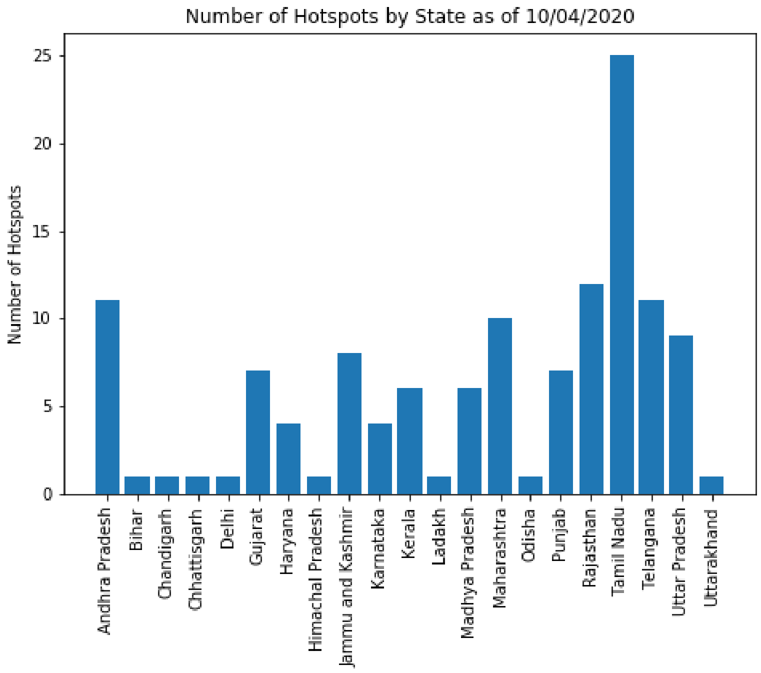
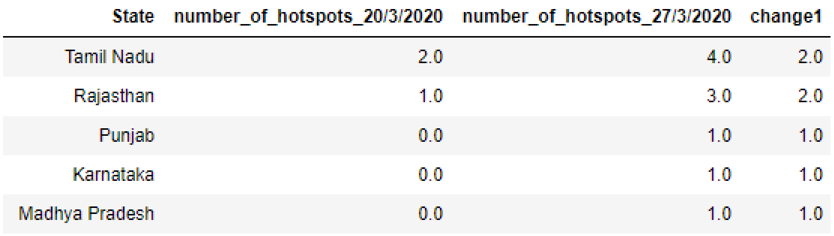


Figure 11: Visualization

5) Change in number of hotspots:

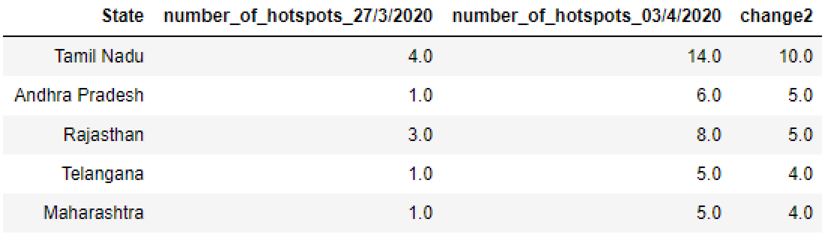
i) Increase in 1st week: Maximum increase in number of hotspots is found in **Tamil Nadu** and **Rajasthan**. Top five states are listed below.



State	number_of_hotspots_20/3/2020	number_of_hotspots_27/3/2020	change1
Tamil Nadu	2.0	4.0	2.0
Rajasthan	1.0	3.0	2.0
Punjab	0.0	1.0	1.0
Karnataka	0.0	1.0	1.0
Madhya Pradesh	0.0	1.0	1.0

Figure 12: Change in 1st week

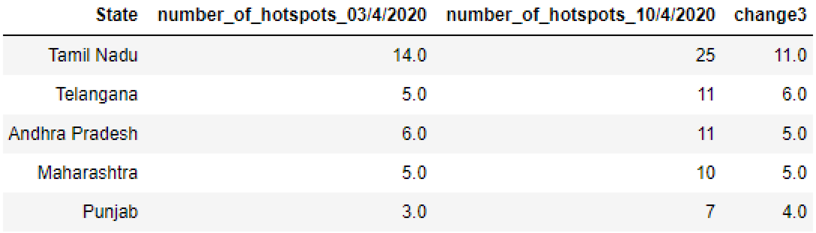
ii) Increase in 2nd week: Maximum increase in number of hotspots is found in **Tamil Nadu**(change is 10), which is followed by **Andhra Pradesh**



State	number_of_hotspots_27/3/2020	number_of_hotspots_03/4/2020	change2
Tamil Nadu	4.0	14.0	10.0
Andhra Pradesh	1.0	6.0	5.0
Rajasthan	3.0	8.0	5.0
Telangana	1.0	5.0	4.0
Maharashtra	1.0	5.0	4.0

Figure 13: Change in 2nd week

iii) Increase in 3rd week: Maximum increase in number of hotspots is found in **Tamil Nadu**(change is 11), followed by Telangana .



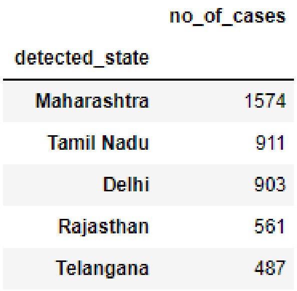
State	number_of_hotspots_03/4/2020	number_of_hotspots_10/4/2020	change3
Tamil Nadu	14.0	25	11.0
Telangana	5.0	11	6.0
Andhra Pradesh	6.0	11	5.0
Maharashtra	5.0	10	5.0
Punjab	3.0	7	4.0

Figure 14: Change in 3rd week

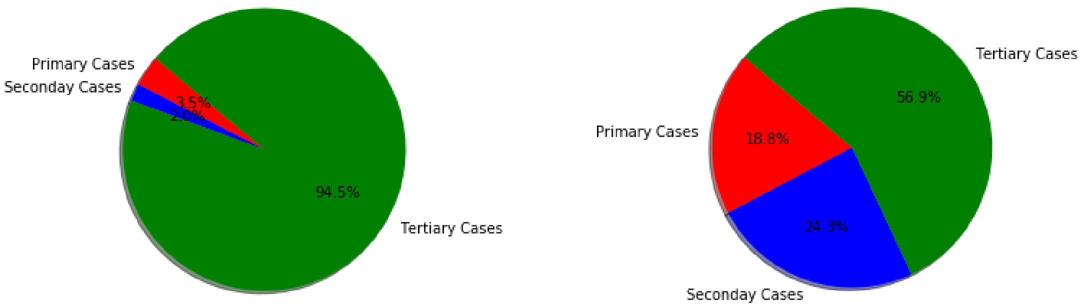
There has been no decrease in hotspot count, but many states have maintained the count constant.

6) For the given data, identify cases with international travel history (primary case), personal contact with primary case (secondary case). Cases which do not fall in the primary and secondary fall into tertiary case. Quantify them based on the percentage for the top 5 states with maximum cases till 10.04.2020.

Inspected the IndividualDetails dataset to get the number of primary, secondary and tertiary cases as well as the top 5 states with maximum cases till 10/04/2020.
Top 5 states with maximum cases:

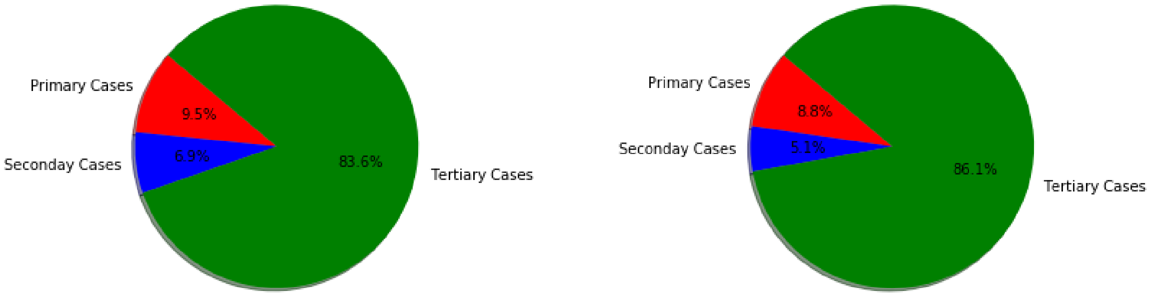


	no_of_cases
detected_state	
Maharashtra	1574
Tamil Nadu	911
Delhi	903
Rajasthan	561
Telangana	487



(a) Delhi

(b) Rajasthan



(c) Maharashtra

(d) Tamil Nadu

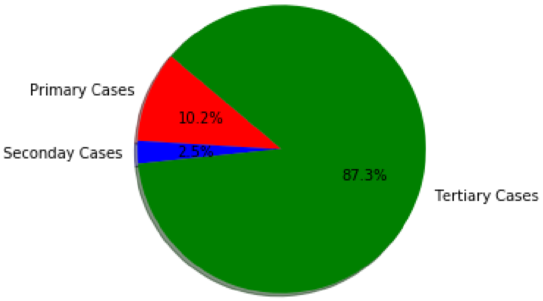


Figure 15: Telangana

7) Find out the number of additional labs needed from the current existing labs (assume 100 tests per day per lab) with an increase rate of 10% cases per day from 11.04.2020 - 20.04.2020. List out any further assumptions considered.

Assumptions:

- From ICMRTestingDetails dataset, rate of testing on 10/04/2020 is approximately 4.6%. Assume that the testing rate stays the same till 20/04/2020. The number of labs required to carry out tests such that all positive cases are detected by 20/04/2020 is calculated.
- The number of currently active labs was obtained by subtracting number of tests on 09/04/2020 from that on 10/04/2020 and dividing by the number of tests per lab per day.
- Rate of testing is the percentage of tests which were covid positive out of the total number of tests conducted.
- Assume that all infected individuals up till 10/04/2020 are detected and tested positive.

Cases on 10/04/2020, 20/04/2020, number of active labs on 10/04/2020 and number of additional labs required on 20/04/2020 so that all positive tests are discovered:

Cases till 10/04/20	Cases till 20/04/20 at given rate	No. of Active Labs on 10/04/2020	No. of Additional Labs Required on 20/04/20
6872	17824	164	74

8) Plot the number of cases starting from 1st March - 10th April. Based on this plot can you comment on the popular notion of ‘flattening the curve’.

Flattening of curve: “Flattening the curve” refers to the above curve flattening at a particular value for “Number of Cases”. This would mean that the given country/state that the curve represents has been successful in containing the virus. However, as shown by the curve for India above, the number of cases is increasing exponentially and we might have to wait for a substantial amount of time before we see the curve flattening.

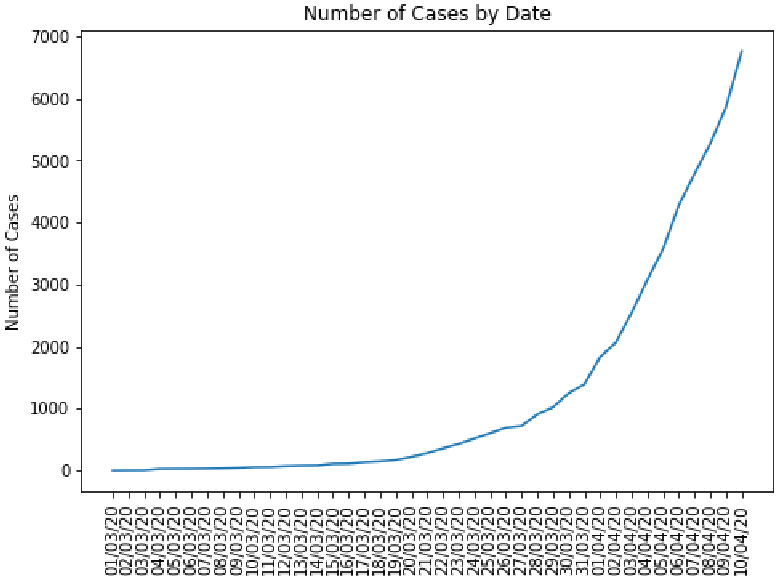


Figure 16: Total no. of positive cases by date

9) As we know, social distancing is the best option to avoid the spread. Based on the time series data (covid_19_india.csv), can you suggest how successful the 21 days lockdown has been?

In order to visualize the effect of the lockdown we considered the data from 31st January 2020 to 24th March 2020 i.e when the lockdown was enforced. We fit a regression model to this data. Once we obtained the fit, we extrapolated the graph to see the number of cases that would have occurred had the lockdown not been enforced.

Regression:

On plotting the scatter plot of the data till 24th March we can clearly see that the data does not follow a linear relationship.

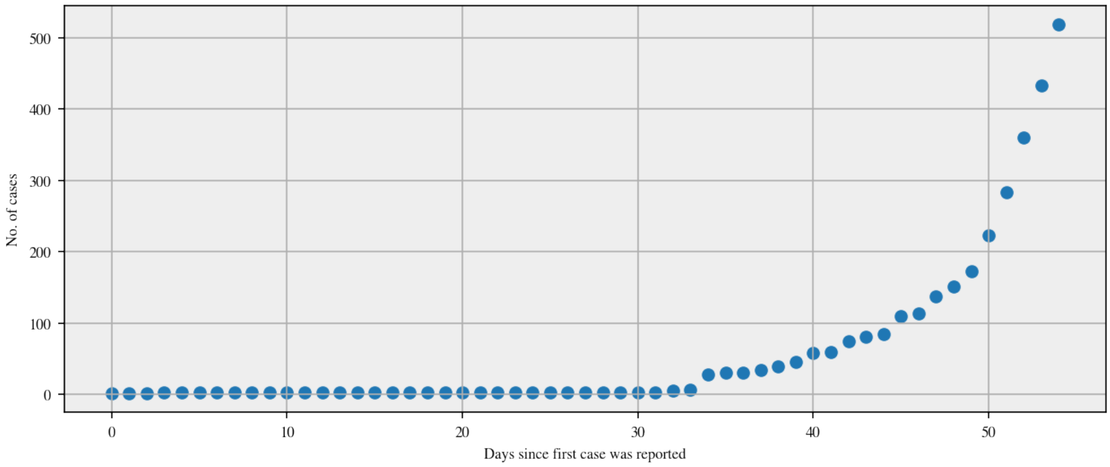


Figure 17: Scatterplot till 24th march

We inferred that the data could follow the following distributions:

- Exponential(log(y)=ax+b)
- Polynomial and exponential($log(y) = ax + bx^2 ....$)
- Polynomial($y = ax + bx^2 ...$)

Exponential:

On plotting the graph for exponential relationship on the data obtained till 24th March 2020 we noticed that the graph does not coincide as well as in the polynomial case. We therefore prefer extrapolating the data using polynomial regression. The RMSE=71.75430801258224 is much higher than what we get with simple polynomial regression This is also why the predicted values till 30th April show such a high deviation from the actual curve.

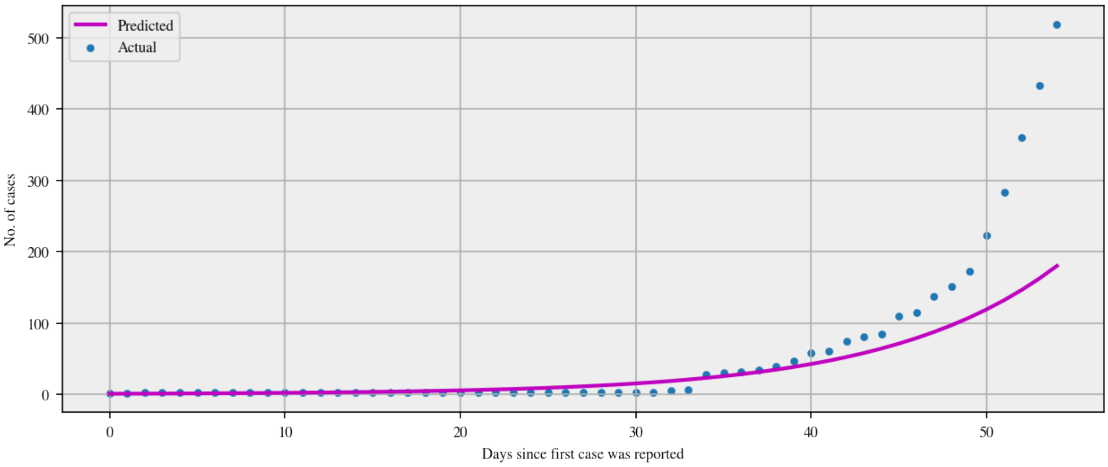


Figure 18: Exponential fit

Polynomial Exponential:
On plotting the graph for polynomial exponential relationship using a degree 2 on the data obtained till 24th March 2020 we noticed that the graph does not coincide as well as in the polynomial case. We therefore prefer extrapolating the data using polynomial regression. The RMSE=30.8852725706685 is much higher than what we get with simple polynomial regression This is also why the predicted values till 30th April show such a high deviation from the actual curve.

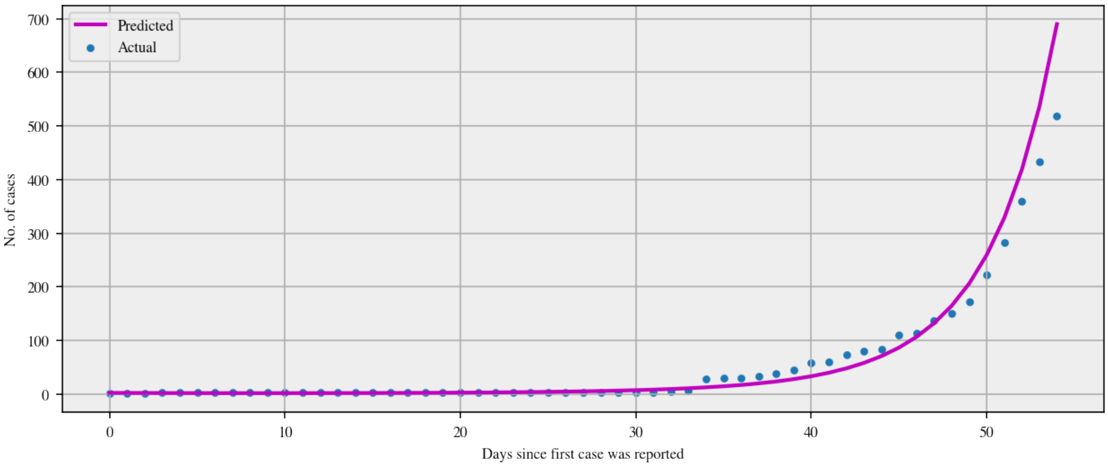


Figure 19: Polynomial exponential fit

Polynomial: On plotting the graphs for polynomial regression with different degrees, we can see that the RMSE is minimized at degree 7.
Hence, we used polynomial regression with a degree of 7 to determine the spread if there was no lockdown. Using this we can see that the lockdown prevented roughly 167,000 cases(200000-33000).

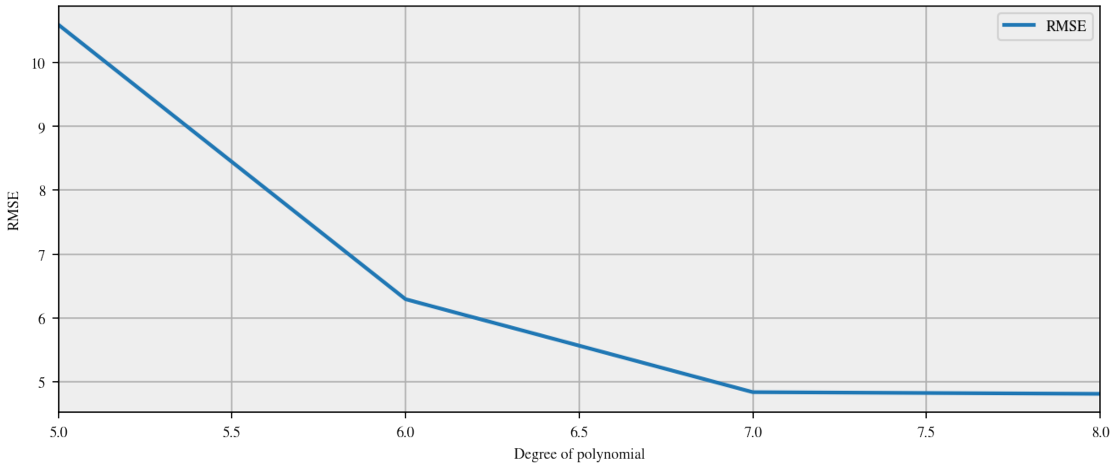


Figure 20: Degree vs RMSE

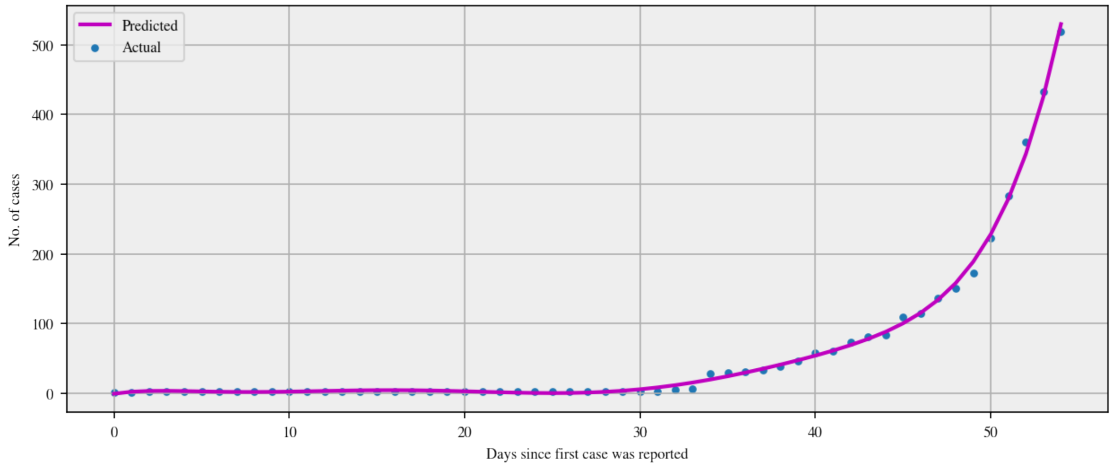


Figure 21: Polynomial fit with degree 7

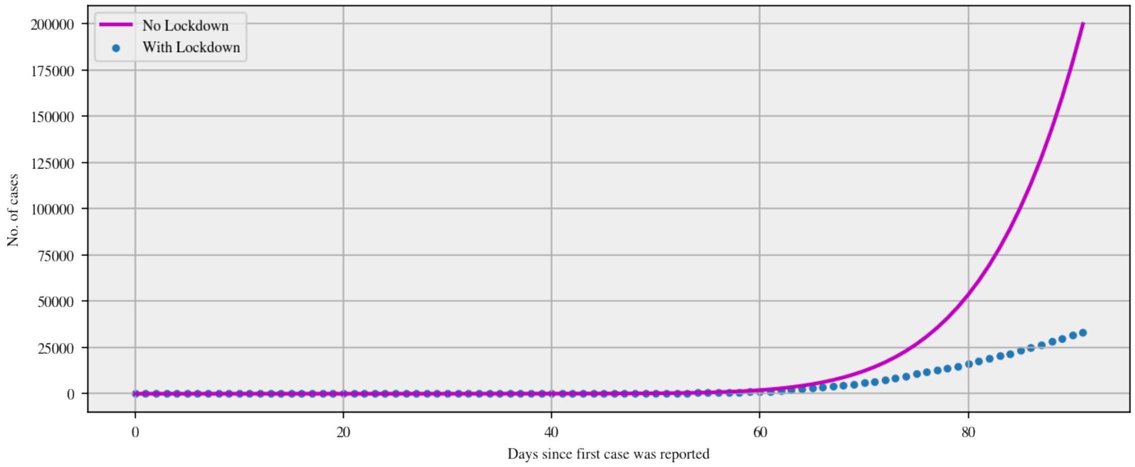


Figure 22: No. of cases with and without lockdown