Covid-19 Analysis

Project Group Members: Kenny Mai, Yutong Xie, Jason Yan, Leo Ren, Brandon Bu
Datathon

Table of Contents

Introduction:

With the current situation of COVID-19 pandemics, our group decided to focus on predicting the date in which the virus would have reached a peak in India and the number of totals confirmed cases in June 30th with the given datasets to serve as a guideline to help people have an understanding of when will society start recovering from the virus.

Description of the Model:

For the first machine learning model, our group used the Linear Regression Polynomial model to predict the date in which the increase in confirmed cases per day from COVID-19 has reached a peak in the whole of India.

The reason why we use the linear regression's polynomial model is that we want to predict the trend where the COVID-19 infection increase rate slows down, which in particular is a downward parabola. However, merely relying on the infection model of confirming rate is far from enough as it isn't necessarily the actual "slowdown" when the increase rate of confirmed cases slows down and comes to a decrease. We also need to take the recovery rate into account and consider the overall effect of both in order to find the peak, because the overall cases of COVID-19 decrease after the recovery rate meets the confirm-rate.

In terms of the recovery rate, as interventions and lockdown policies are put into practice, we are expected to see more and more people recover from the virus. So we expect the recovery rate model to be somewhat increased. With the interaction point of the two models, we can get the idea of where the peak is**.** Based on the provided data from April 21st to May 16th, we predict for 14 days afterward in order to get the intersection point.

Code Snippet:

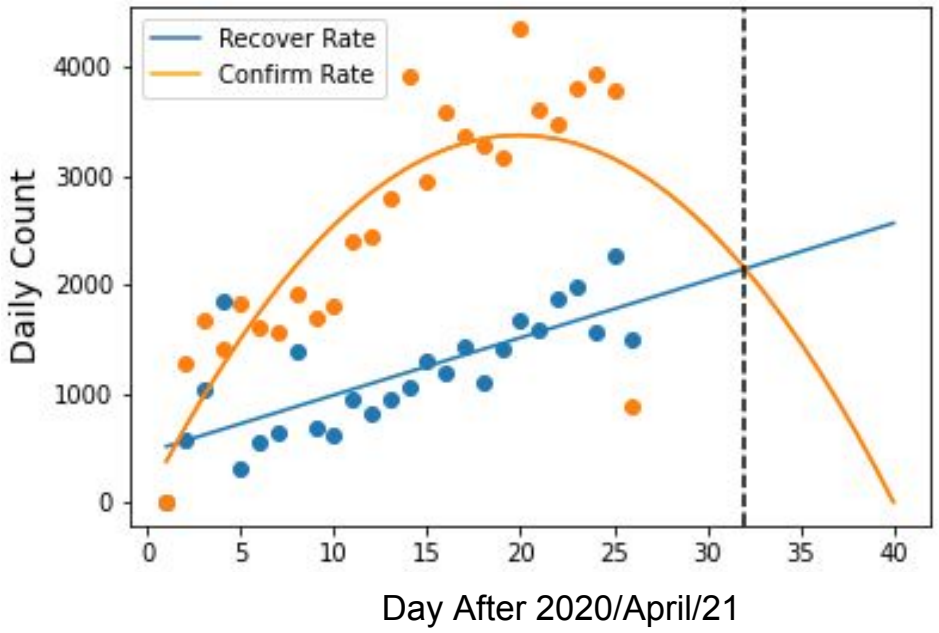
```
# For the linear regression polynomial model of confirmed rate
polynomial_features = PolynomialFeatures(degree=2)
x_poly = polynomial_features.fit_transform(X)
m = LinearRegression().fit(x_poly, y)
x_extend = np.arange(1, 41).reshape(-1, 1)
x_poly_extend = polynomial_features.fit_transform(x_extend)
y_poly_pred = m.predict(x_poly)
y_poly_pred_extend = m.predict(x_poly_extend)

# Sort the data of linear regression polynomial model
sort_axis = operator.itemgetter(0)
sorted_zip = sorted(zip(X, y_poly_pred), key=sort_axis)
x, y_poly_pred = zip(*sorted_zip)
plt.plot(X, y_poly_pred, color='m')

# For the linear model of recovery rate
bestScore = 0
bestPredict = 0
bestModel = 0
for i in np.arange(10).tolist():
 Xtrain, Xvalid, ytrain, yvalid = train_test_split(X, y, test_size=0.2)
 m = linear_model.Ridge(alpha=1).fit(Xtrain, ytrain)
 if ((bestScore < m.score(Xtrain, ytrain)) and (bestPredict <
m.score(Xvalid, yvalid))):
   bestScore = m.score(Xtrain, ytrain)
   bestPredict = m.score(Xvalid, yvalid)
   bestModel = m
predict_X = np.arange(1, 41).reshape(-1, 1)
y_pred = bestModel.predict(predict_X)
```

Prediction of the peak number of COVID-19

Graph (India):

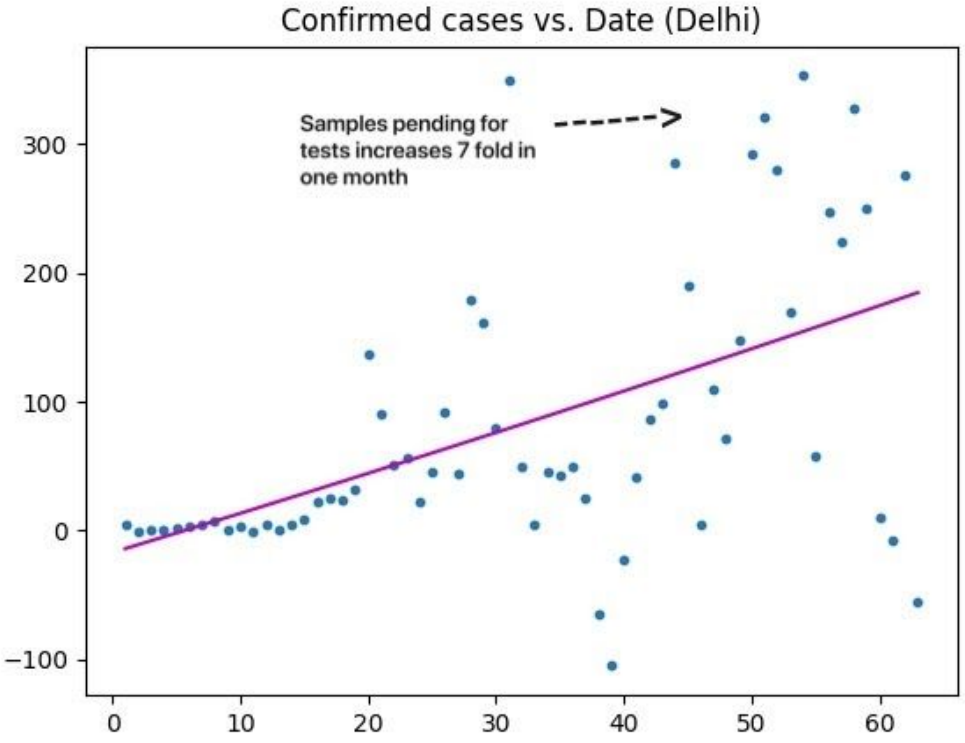


Analysis:

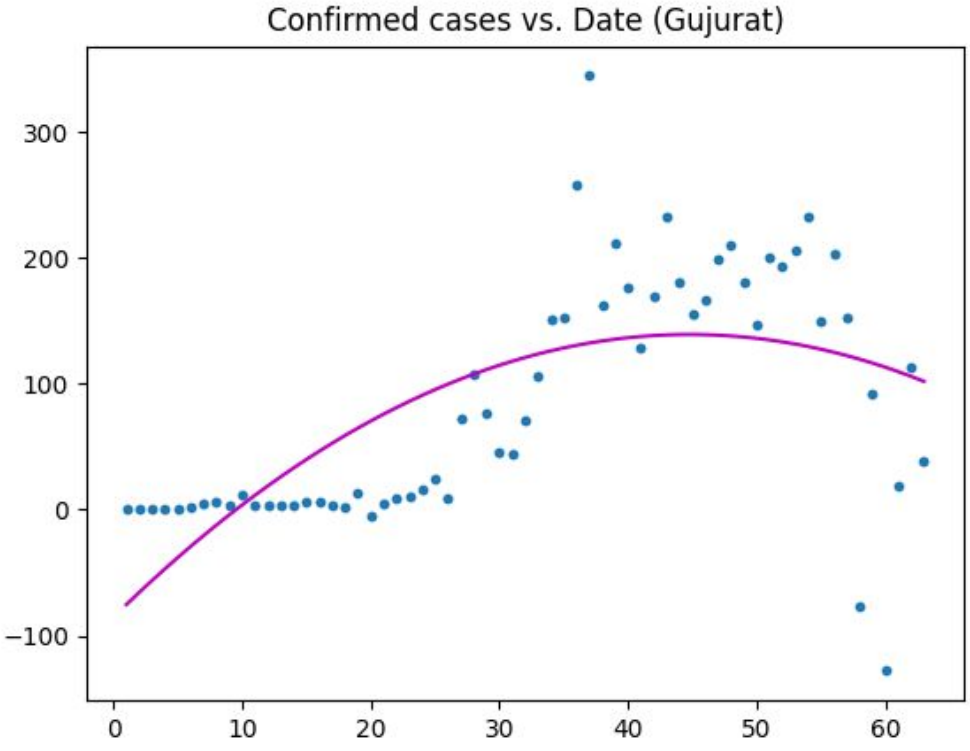
The graph above uses the daily_may_15 datasets and with India as whole to plot the graph explaining the relationship between the rate of change in the increase of confirmed cases per day (shown in orange) and the rate of change in the increase of recovered cases per day (shown in blue) with date starting from April 21st to May 16th, the data points after is predicted by the model. The model predicted that the shape of the graph for the increase of confirmed cases is a downward parabola and the increased rate of confirmed cases reaches the max at about 3300 per day around May 10th. After that, the increase rate starts to decrease so that the total cases will increase at a lower rate. The machine learning model shows that the peak in the number of COVID-19 has passed.

Below are graphs generated by the same model used above but instead of using the daily_may_15 datasets we used the state_wise_daily datasets to obtain more data points to predict the peak in some of the more population intense states and districts.

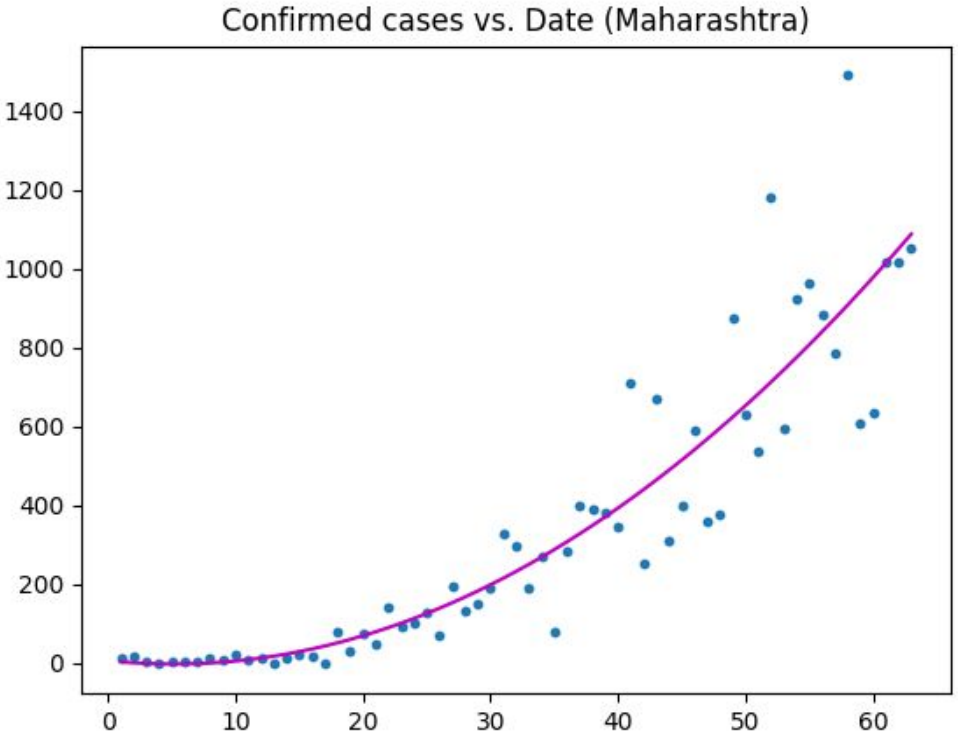
Graph (State Delhi):



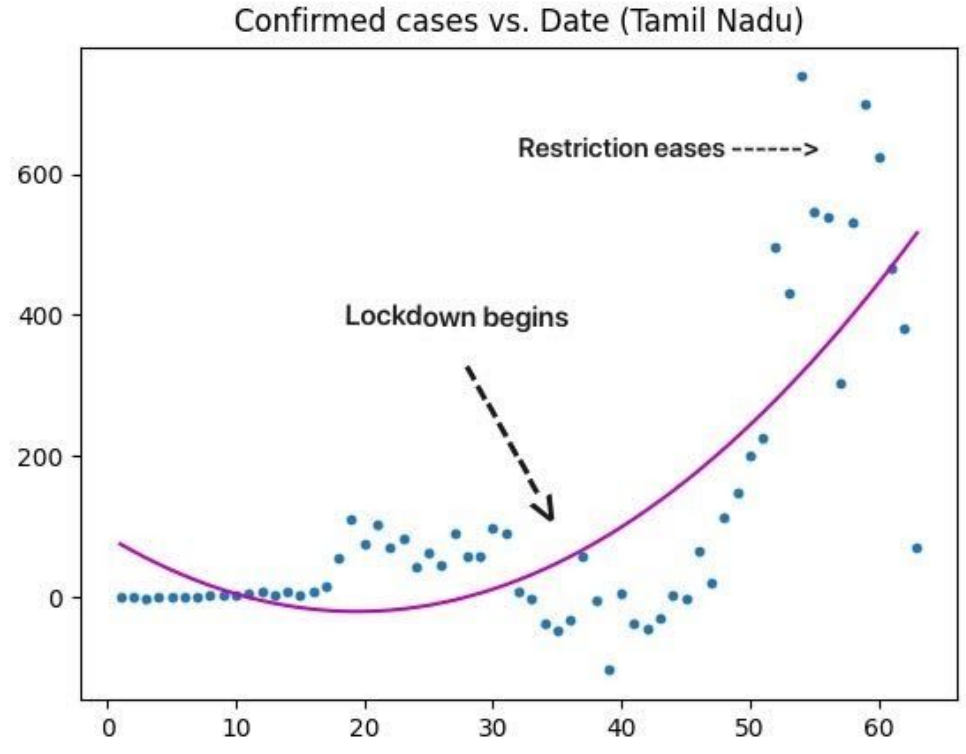
Graph (State Gujurat):



Graph (State Maharashtra):



Graph (State Nadu):



Analysis:

The four graphs above use the state_wise_daily datasets to plot the graph explaining the relationship between the rate of change in the increase of confirmed cases per day (shown in purple) date starting from March 14th.

In the graphs above, we plotted the increased rate of daily confirmed cases against the date for the four major states Maharashtra, Delhi, Tamil Nadu, Gujarat as a comparison to India as a whole. We can see that the curve has flattened for India as a whole and for the state of Gujurat. However, for other states like Delhi, Maharashtra, and Nadu, the number of daily confirmed cases increased is not slowing down.

This might be caused by the irregularities in the dataset. In the case of Delhi and Gujurat, we suspect the irregularities to be a result of the lab's limited capacities to process test results. In fact, the number of samples pending for tests in Delhi increased by 7 times in April.

Things for Nadu are a bit different. We see a sharp decrease when lockdown began and an increase when the lockdown restrictions eased on May 11. Though we only have a limited amount of data, we can see from the dataset itself that a decrease in confirmed cases might have started. As Nadu tightens its lockdown, we expect the number of daily confirmed cases to slow down.

Conclusion:

The irregularities in the dataset might have been caused by the inconsistencies in data collection. As a result, we don't have the "true" data available to us. Though we know labs in India are working vigorously every day, the inconsistencies in our dataset have made it difficult to analyze and make predictions. And our predictions might not be very accurate at all. Although we do expect to see a decrease in active cases in India as a whole, the data from Delhi, Gujurat, and Nadu say otherwise. Though we might be able to see a downward trend in the days to come, India still needs to be very cautious about a potential second wave coming.