Predicting the spread of Covid-19 in India

Nediff Amala Nixon

May 10, 2020

1.Introduction

1.1 Background

Our world is dramatically changing as our global community deals with covid-19. As the ripple effect of Covid-19 careens around the world, it's forcing humankind to innovate and change the way we work and live. The coronavirus has spread significantly worldwide, since it first emerged in China in December, 2019, making it a global pandemic. There are about 4 million confirmed cases around the world.

Though we follow, social distancing, self-isolation policies, avoiding mass gatherings and postponing major events, many of us had thought already when will this pandemic actually end.

1.2 Problem

However, it is impossible to say when and how this coronavirus will die, and the world will finally be free from it. The report mainly focuses on the next few days spread based on the trend it had followed in the past days.

1.3 Interest

in the unprecedented times like these, the power of prediction helps people in understanding the trend, and keep them aware and helps foresee things. However, we can’t model exactly how long or bad this can go, with the available data we can definitely see the trend and behave accordingly.

Data

2.1 Data Sources

The data was extracted from John Hopkins University's dashboard on Corona Virus, and the data are extracted from Google sheets that were made available. The datasets used were,

1.Covid – Worldwide Stats
2.Covid – India Stats
3.Age Distribution
4.ICMR Testing data
5.Statewise tested numbers
6.Individual Details
7.Hospital Beds in India

2.2 Data Cleaning

Data scraped from google spreadsheets doesn't required much of a cleansing. Very basic cleansing has been carried out like, date type conversion and converting float to integer value to support the analysis.

Where ever the Province was null, it got replaced with Country. And, Mainland China and China were combined to a single entity

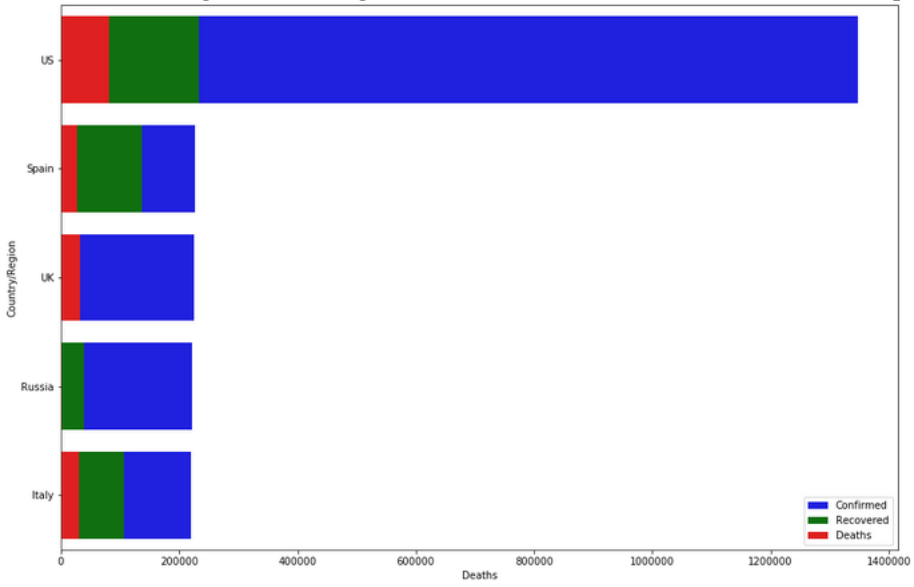
2.3 Feature Selection

Different analysis that were planned to carry out were World wide spread stats, Recovery versus death rate. Rate of spread in India, number of confirmed and death cases over time, gender and age distribution of infected people, number of cases across states, test stats and recovery rate across states.

3.Exploratory Data Analysis

3.1 Worldfests

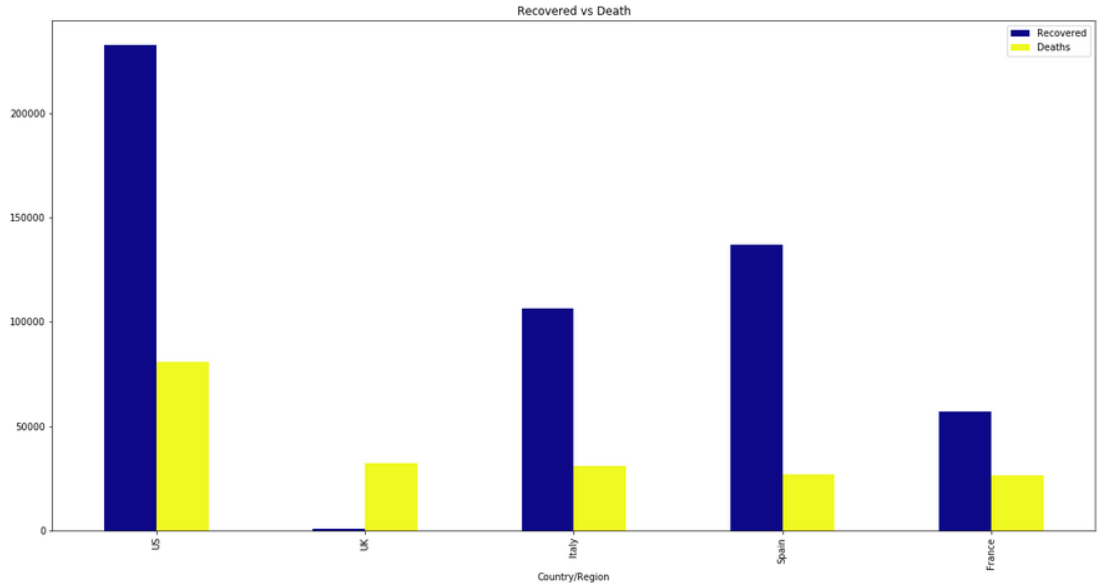
After accessing and cleaning the data obtained, the overview of data was plotted.



With United States being the highest number of confiremed cases, United Kingdom showed very low recovery rate compared to any other countries.

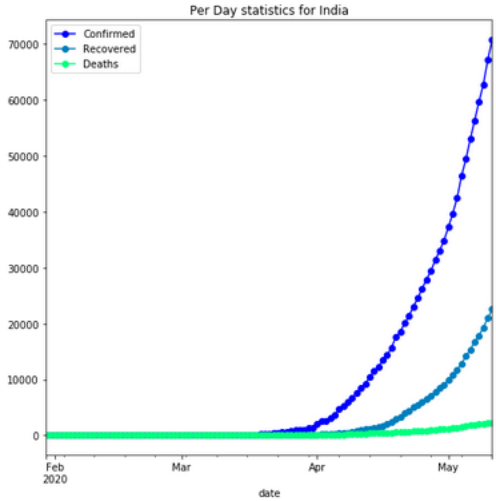
3.2 Recovery versus Death rate

After analyzing the world wide states, in order to know the Recovery rate, recovery versus death rate had been plotted,



3.3 Rate of Spread in India

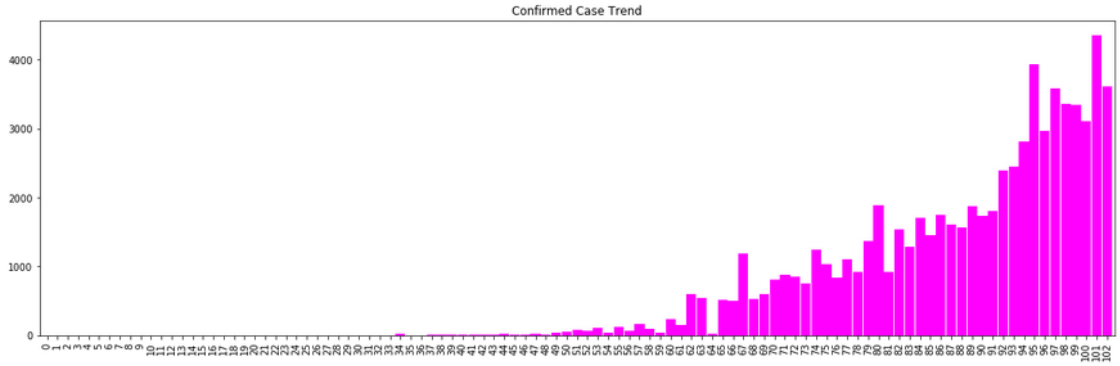
After analyzing enough of worldwide stats, while checking on India stats, it was idle to check the rate of spread over time.



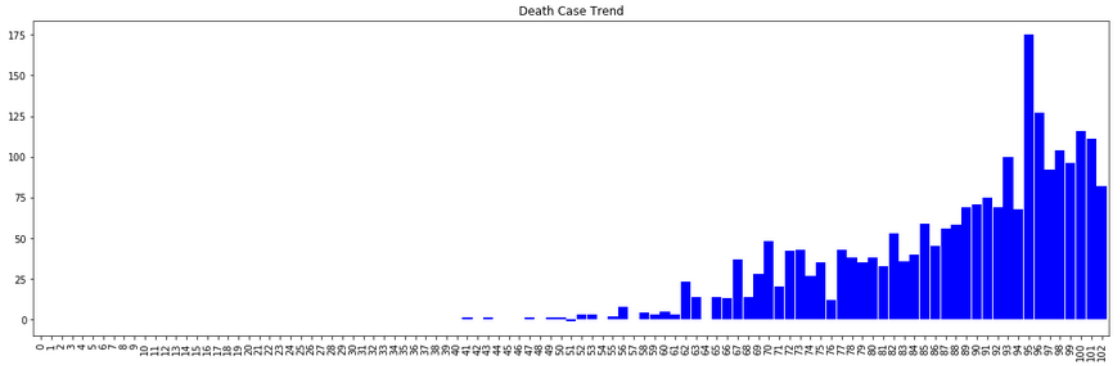
It is clear that, the spread started to rise after the beginning of April, and it had been growing since then. Recovery rate shows a very slow increase while on the good part, death rate is near to flat.

3.4 Number of confirmed and death cases

After analysis the spread across time death and confirmed cases over time had been plotted to see if any pattern had been followed.




```
num_plot_india['Death Case Increase'].plot(kind='bar',width=0.95,colormap='winter',figsize=(20,6),title='Death Case Trend')
plt.show()
```



Adding other datasources for further analysis of India on State-level

From the graph it was clear that it doesn't follow a trend and the spike and the downfall occurred on random days

3.5 Age and Gender Distribution

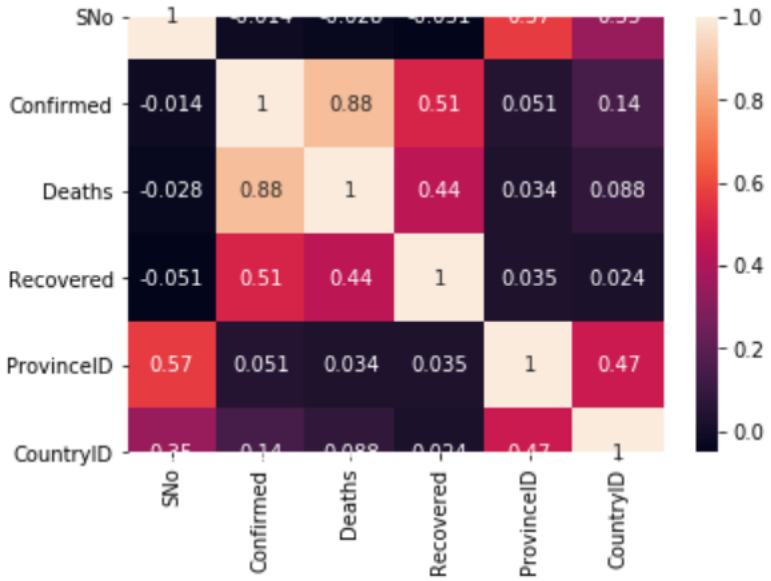
Age and gender distribution are plotted to see if any correlation is between variables', it seems male got infected more and the coronavirus is directly proportional to the age, as the age increases, the infected people percentage got increased too.

3.6 Number of cases and the cases tested

Also, number of cases tested is promotional to the number of positive cases. The states with higher number of tests, results in higher positive cases. Although the percentage of positive cases among the tested is low, it is directly proportional to the tests carried out

3.7 Correlation Matrix

In order to see, if there is any correlation between variables, correlation matrix is calculated.



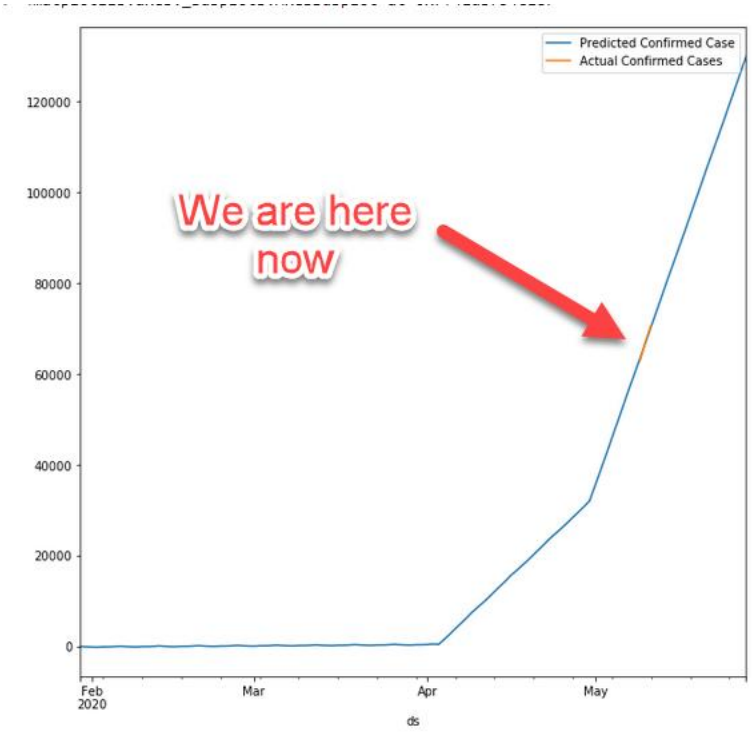
None of the variables seems to have any correlation, except for the Confirmed, recovery and the death cases

4.Predictive Modelling

The main objective of the project is to predict the spread, based on the trend. I had used prophet to predict the values

4.1Prection till May end

Upon training the model, it is predicted that by 17th May (End of Lockdown 3) more than 91.4k confirmed cases are predicted as per this model with upper limit of around 93.2k



4.2 Prediction on entire population

If the same trend continues, by the end of Jan 2021, 1.4 million population will be infected.

5.Conclusion

I know, I did paint the picture too dark. Nevertheless, with proper precautionary measures like social distancing, avoiding mass gatherings, self-isolation will definitely help us over this global pandemic. Like, China, South Korea where the situation is very much under control, we can be there too. Within India, Kerala, being one of the earliest states to find COVID, now with the recovery rate as 94%, the situation is very much under control

6.Future Direction

Nothing helps more than self -isolation and social distancing. To be a step ahead of virus, testing more cases will help us overcome the situation. Rather than chasing the contacts of infected people. More testing has its own pro and cons, it requires more man power, more testing kits, and not to mention the monetary aspect.