BIO101 Epidemiology Assignment - COVID SIR Model

South Carolina, USA

Abstract

In late December 2019, Chinese health authorities reported an outbreak of pneumonia of unknown origin in Wuhan, Hubei Province. A few days later, the genome of a novel coronavirus was released and made publicly available to the scientific community. This novel coronavirus was provisionally named 2019-nCoV, now SARS-CoV-2 according to the Coronavirus Study Group of the International Committee on Taxonomy of Viruses. SARS-CoV-2 belongs to the **Coronaviridae** family, **Betacoronavirus** genus, subgenus **Sarbecovirus**. Since its discovery, the virus has spread globally, causing thousands of deaths and having an enormous impact on our health systems and economies.

Through our report, we aim to predict various parameters and aim to fit our predicted hypothesis on the famous SIR Model. Our report contains the initial Exploratory Data Analysis we did to familiarize ourselves with the current scenario of infected and death cases. We then defined the SIR model based on its governing differential equations and solved it to get values for Alpha and phi, followed by Beta and Gamma. We also used Logistic Regression to fit the infected data to predict the maximum number of infectives and other important details.

Exploratory Data Analysis (EDA)

Exploratory data analysis is an approach to analysing data sets by summarizing their main characteristics with visualizations. The EDA process is a crucial step prior to building a model in order to unravel various insights that later become important in developing a robust algorithmic model.

Applications of EDA

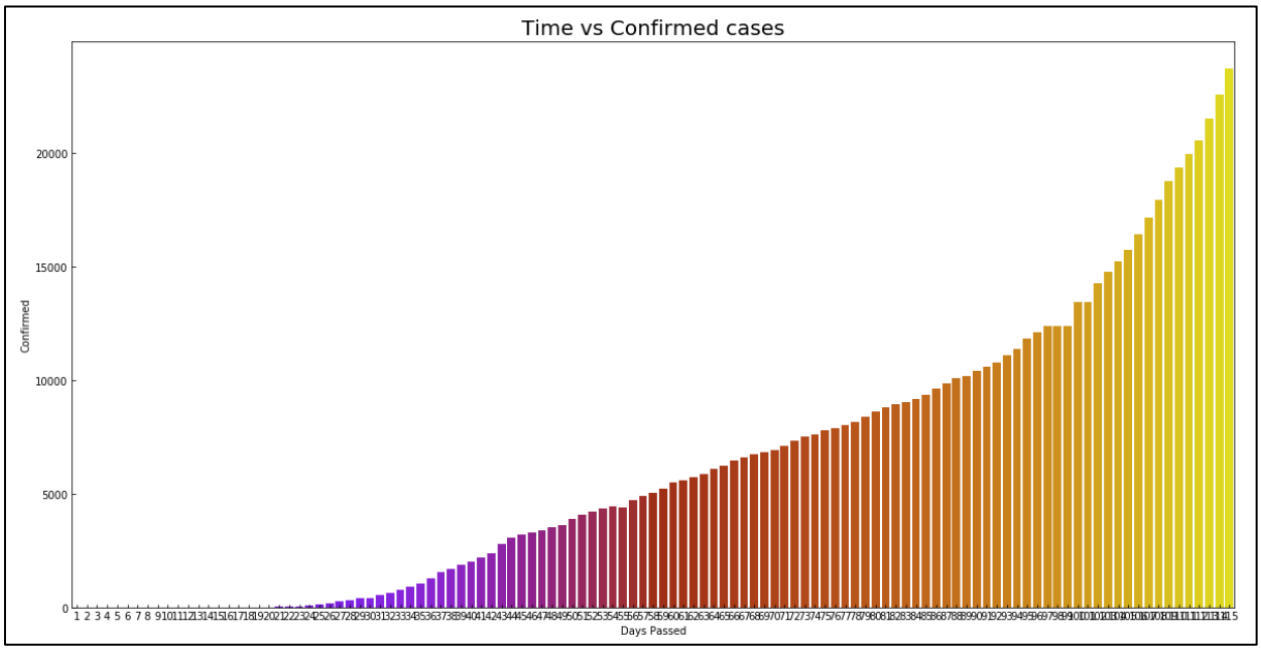
- provides relevant insights which help analysts make key business decisions
- Handle categorical variables with numerically coded values
- Identify and treat missing values and remove dataset outliers

I used the following libraries in my assignment

- **NumPy** – Arrays and calculations
- **Matplotlib** – For graphs and plotting data
- **Seaborn** – For plotting data
- **Math** – For predefined functions
- **Pandas** – Data Frame structure and tables
- **SKLearn** – ML Regression
- **SciPy** – Integration and Optimizing

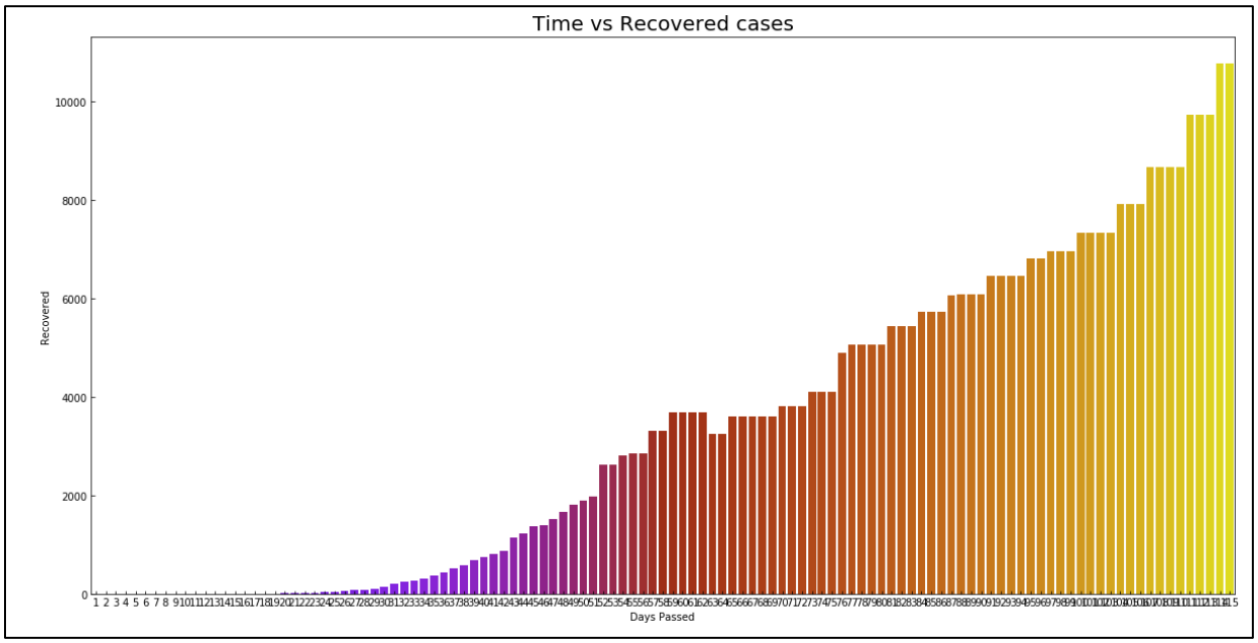
EDA on the dataset

- **Confirmed cases vs Time**



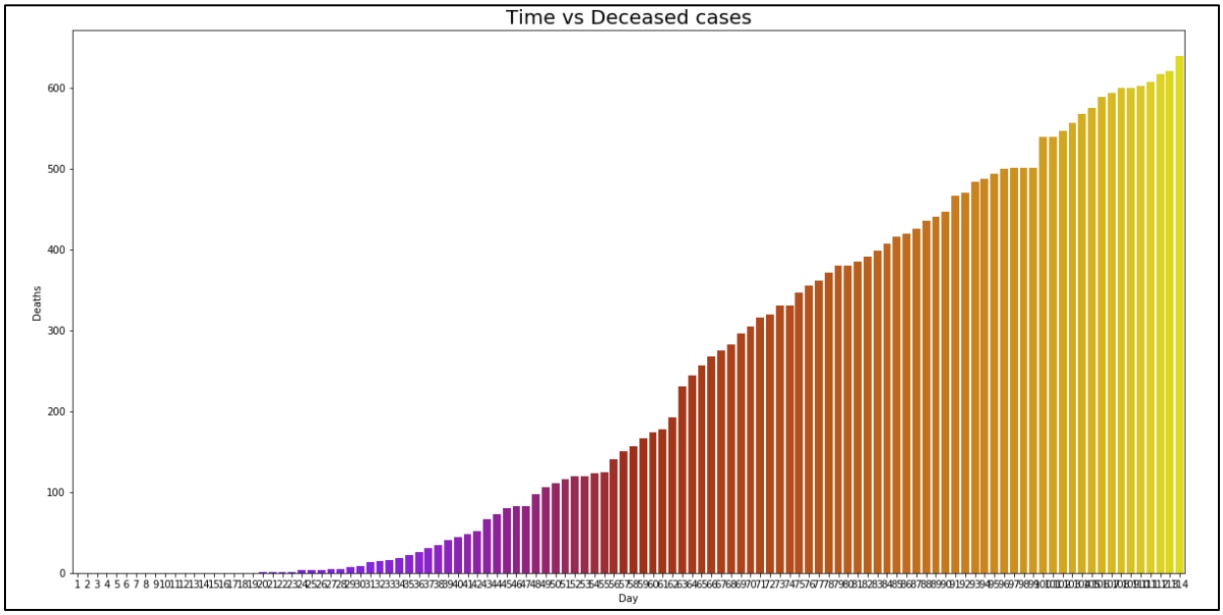
Observation: We observe that the number of infected cases grows at a high rate with time (almost 3 months). This shows that the graph has yet to peak.

- **Recovered cases vs Time**



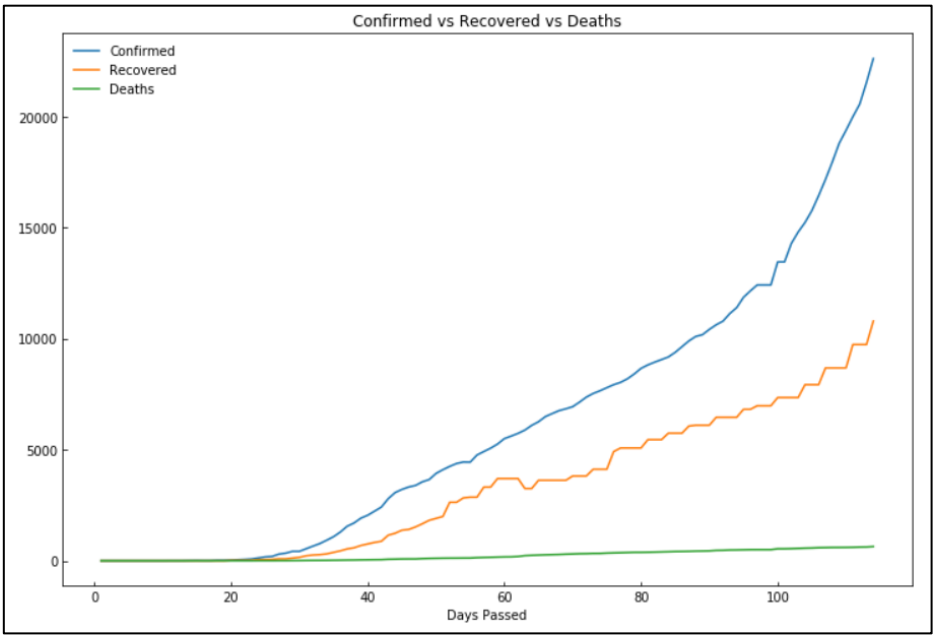
Observation: Number of patients who recover is also steadily increasing. In certain parts, the number of recovered patients decreased from the previous day. This can be due to improper filling of data or some of the patients who were reported to be recovered did not recover fully.

- **Deceased cases vs Time**



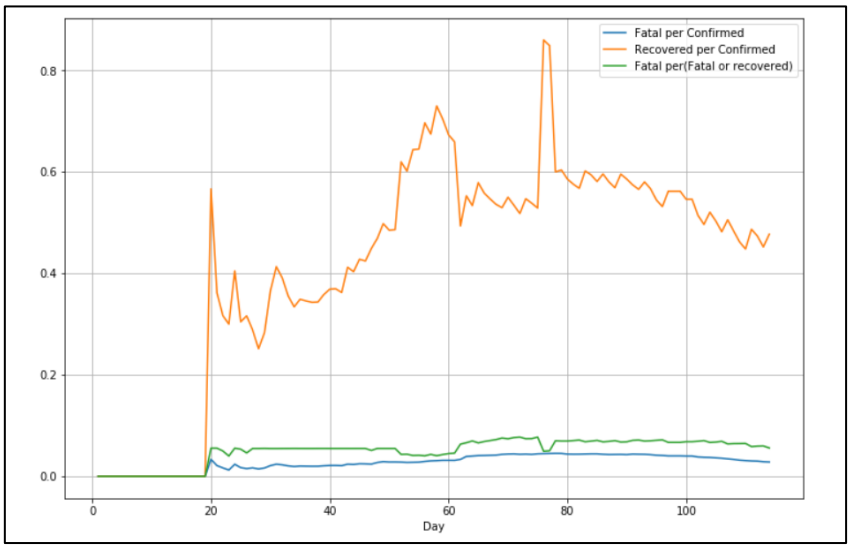
Observation: Although the number of deceased patients is increasing with time, it is a more linear graph than the graphs of Confirmed and Recovered patients, which suggests that the patients are receiving healthcare and treatment.

- **Visualising all three curves together as a LINEPLOT**



Observation: The graph indicates that the situation in South Carolina is yet to be controlled as the number of infected patients is far more than the ones who have recovered.

- **Plotting the rates of recovered and death cases with respect to number of Infected (Confirmed cases)**



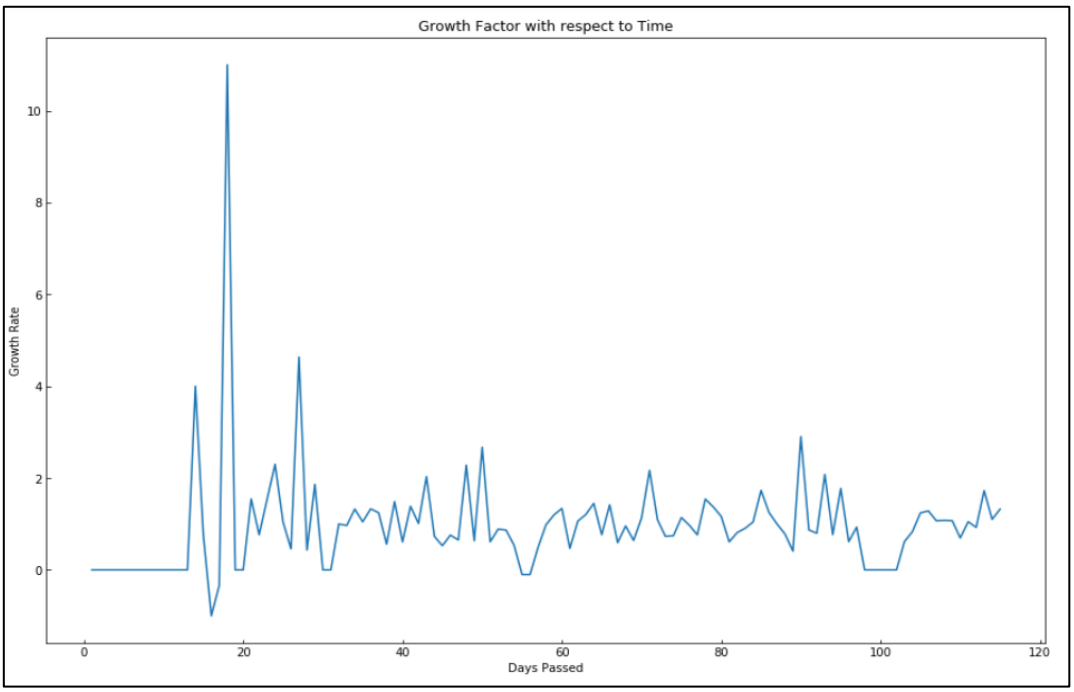
Observation: We can observe from the above graph that the recovery rate of patients in South Carolina peaked at around day 70 reaching as high as 95%. Overall, the rate of fatality was quite low (Approx. 5%)

Growth Factor

Growth factor is defined as the addend by which a quantity increases/decreases over time. For COVID, growth factor is defined by the following formula

$$\text{Growth Factor} = \frac{\Delta C_n}{\Delta C_{n-1}}$$

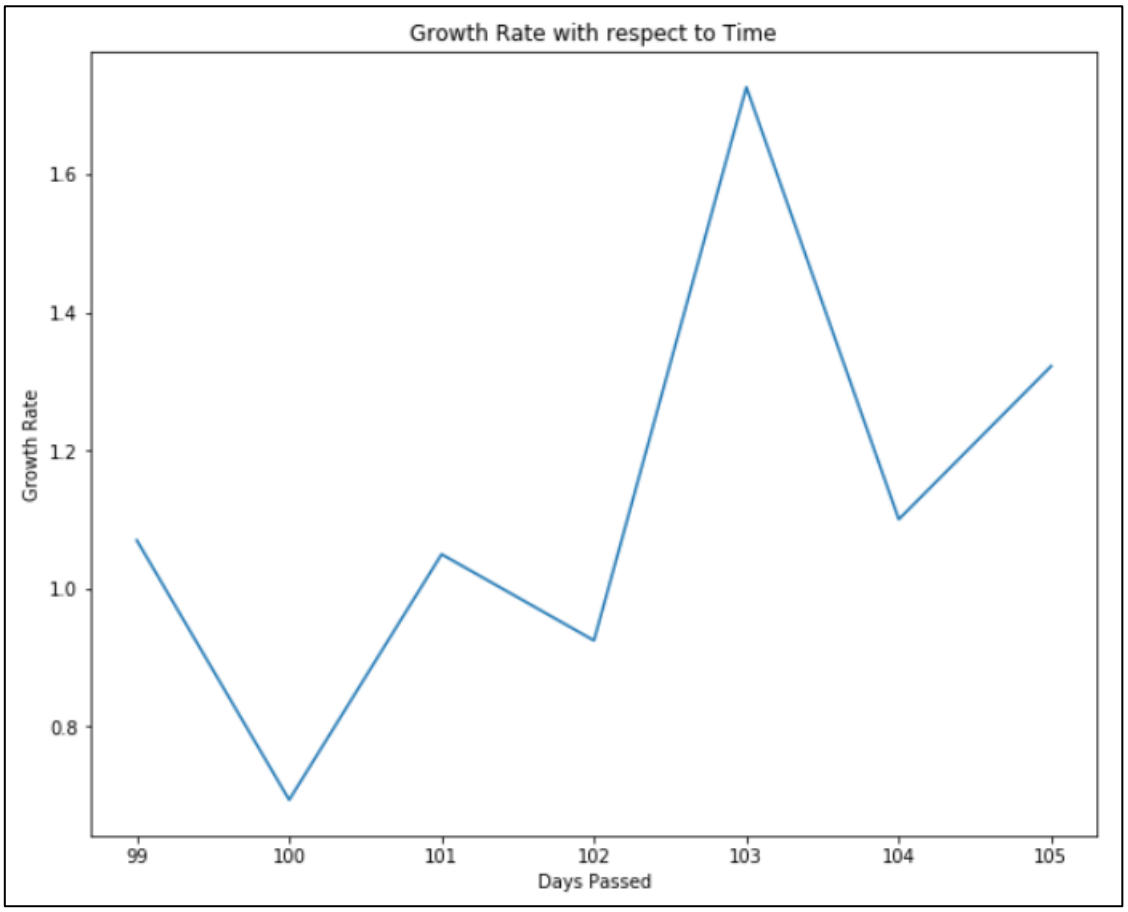
where C is the number of confirmed cases



NOTE:

- **Outbreaking Scenario** : growth factor > 1 for the last 7 days
- **Stopping Scenario**: growth factor < 1 for the last 7 days
- **At a crossroad** : Growth factor approx. 1 for the last 7 days

Growth factor for the last 7 days (till 22/06/2020)



Observation: We can see that the growth factor was close to 1 for several days and peaked at 1.7 on 20th June 2020 . Hence, the condition in South Carolina is going to worsen till it peaks.

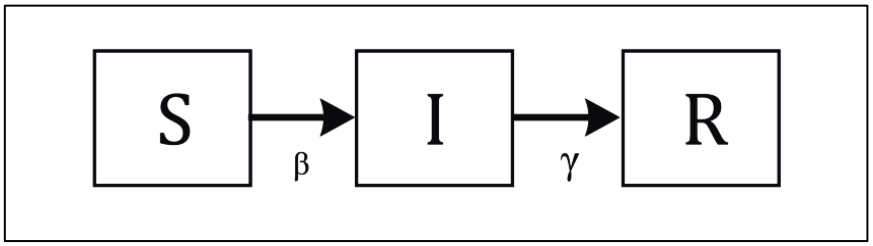
_____EDA ends_____

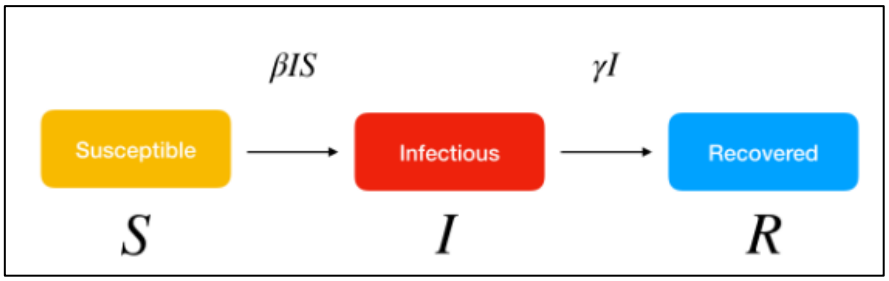
SIR Model and Inferences

About SIR

SIR is a simple model that considers a population that belongs to one of the following states:

1. **Susceptible (S)**. The individual hasn't contracted the disease, but she can be infected due to transmission from infected people
2. **Infected (I)**. This person has contracted the disease
3. **Recovered/Deceased (R)**. The disease may lead to one of two destinies: either the person survives, hence developing immunity to the disease, or the person is deceased.





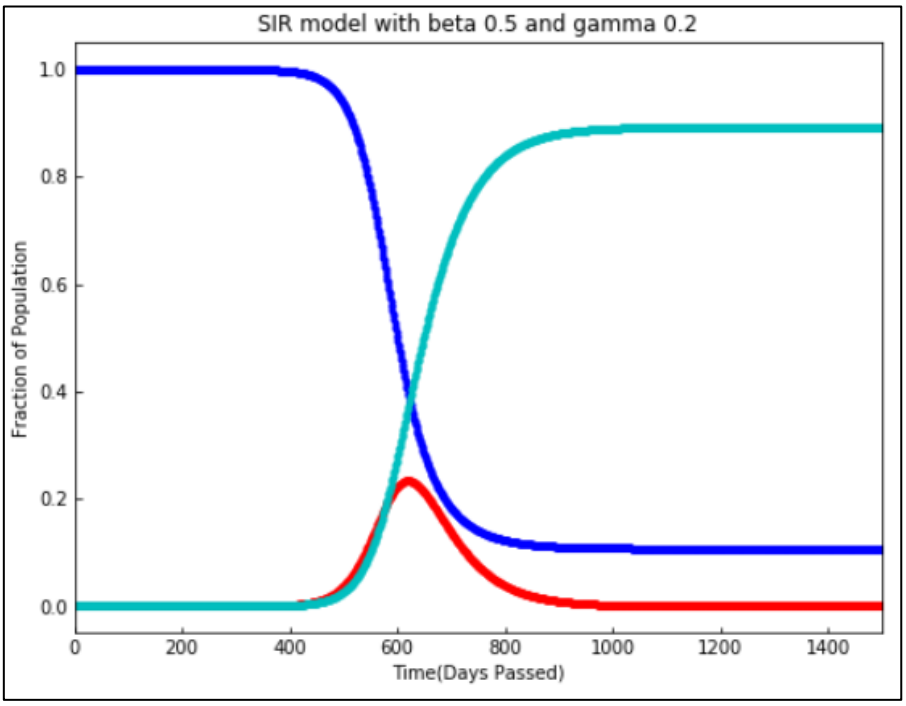
SIR Governing Differential Equations

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$
$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

Sample SIR Model

Consider the following values

- N = 7800*(10**6) # Population
- b0 = 0 # Initial fraction of population infected (In factors of N)
- beta = 0.5 # Rate of transition from Susceptible to Infected
- gamma = 0.2 # Rate of transition from Infected to Recovered/Deceased



Observations :

- The number of infected cases increases for a certain time period, and then eventually decreases given that individuals recover/decease from the disease
- The susceptible fraction of population decreases as the virus is transmitted, to eventually drop to the absorbent state 0
- The opposite happens for the recovered/deceased case

NOTE: Different initial conditions and parameter values will lead to different scenarios

Fitting SIR Parameters to South Carolina dataset

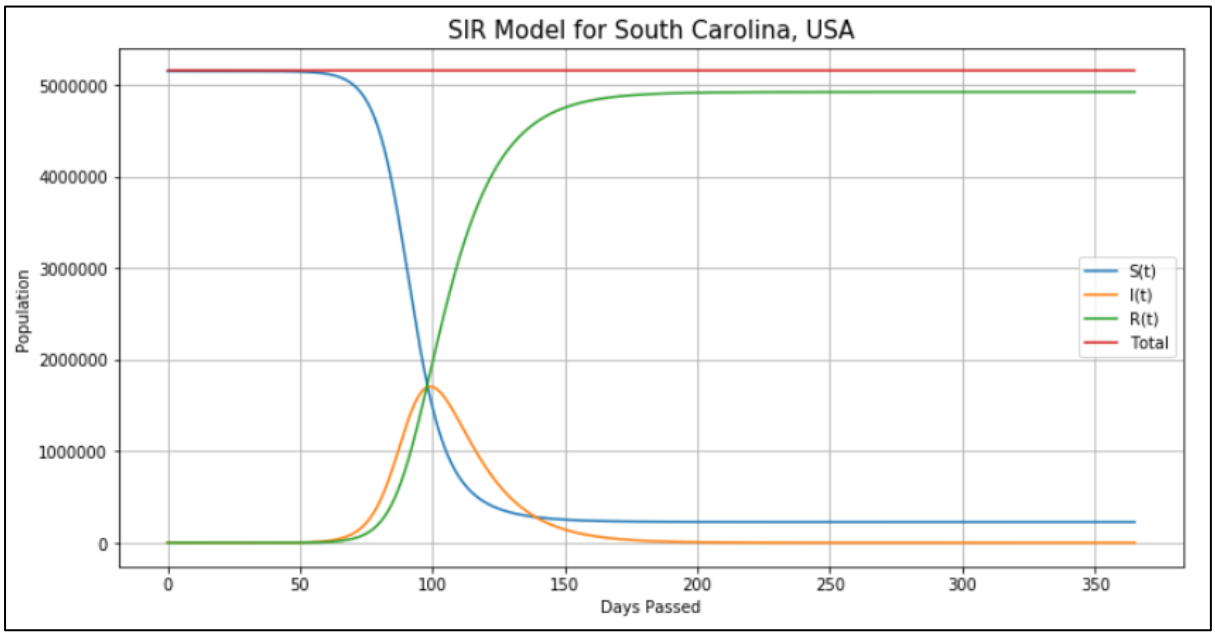
The SIR model is purely theoretical, and we are interested into a real approximation of the COVID-19 (specifically in South Carolina, USA) expansion in order to extract insights and understand the transmission of the virus. Hence, we need to extract the β and γ parameters for each case if we hope to be able to predict the evolution of the system.

The population of South Carolina is 51.5 lakhs (51,50,000)

Applying the formulas involved, I got the following optimal values for the parameters

- **Rho** : 1577169.2522
- **Alpha** : 2.2653
- **Phi** : 7.7082
- **Beta** : $4.619737569513828 \times 10^{-8}$
- **Gamma** : 0.0728610804787037

The resulting graph was as follows



Observation:

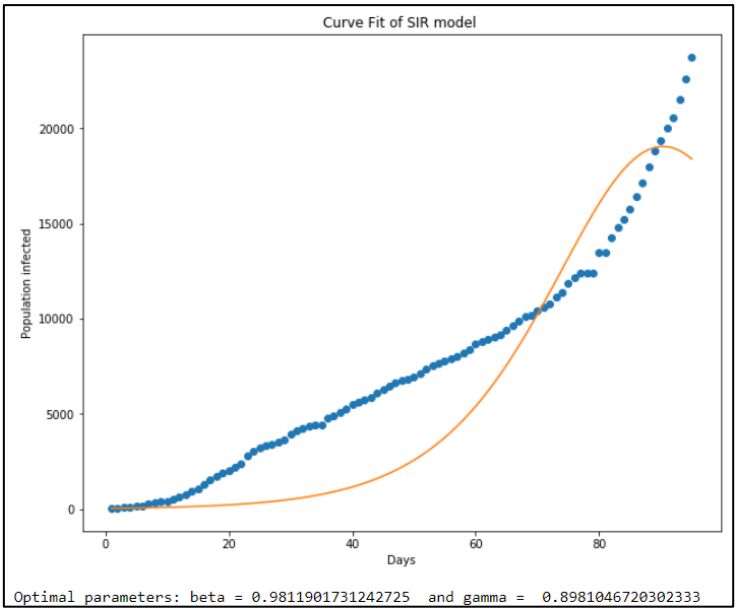
- Most of the people recover after **200 days** have passed
- Recovered individuals were approximately **49.3 lakhs (95.7 % of the population)**
- Duration of the epidemic was approximately **250 days**
- Maximum number of infected individuals was approximately **17.1 lakhs (33% of the population)**
- The R_0 value was approximately **3.27**

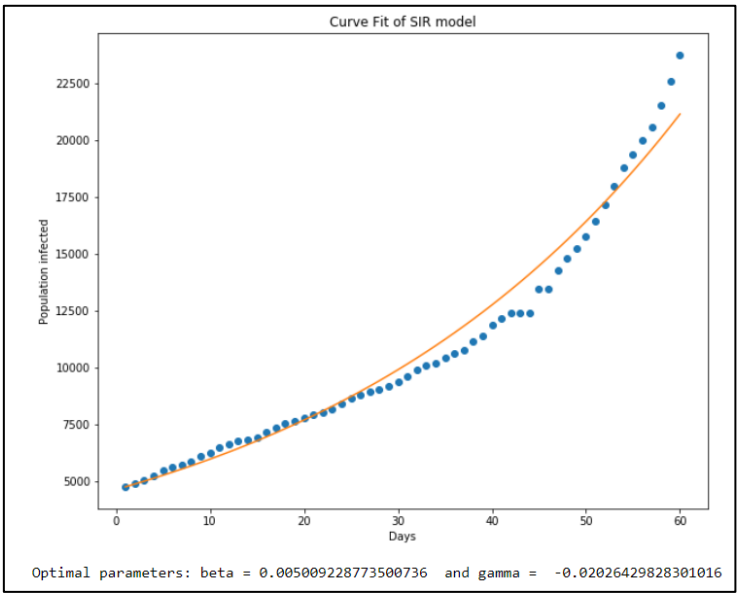
Limitations and Difficulties

As the number of infected cases had not peaked, it was very difficult for me to apply advanced algorithm like Curve Fitting and Logistic Regression. However, I was able to implement Logistic Regression and it turned out to be consistent with the SIR model found by solving for the optimal parameters.

As the number of infected (confirmed) cases were increasing day by day, applying Curve fitting **resulted in negative values for Gamma and Beta** which is impossible, as a **negative gamma and beta value will decrease the number of recovered patients and increase the number of susceptible individuals**! This is practically impossible.

The below images are the graph plotted when applying curve fitting

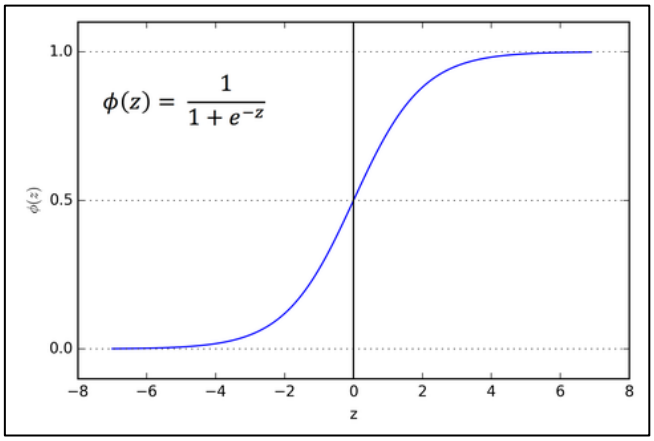




I have uploaded the Jupyter notebook in which I have tried to apply Curve fitting and the code is also available on my GitHub profile. **Link is in the references section**.

Logistic Regression

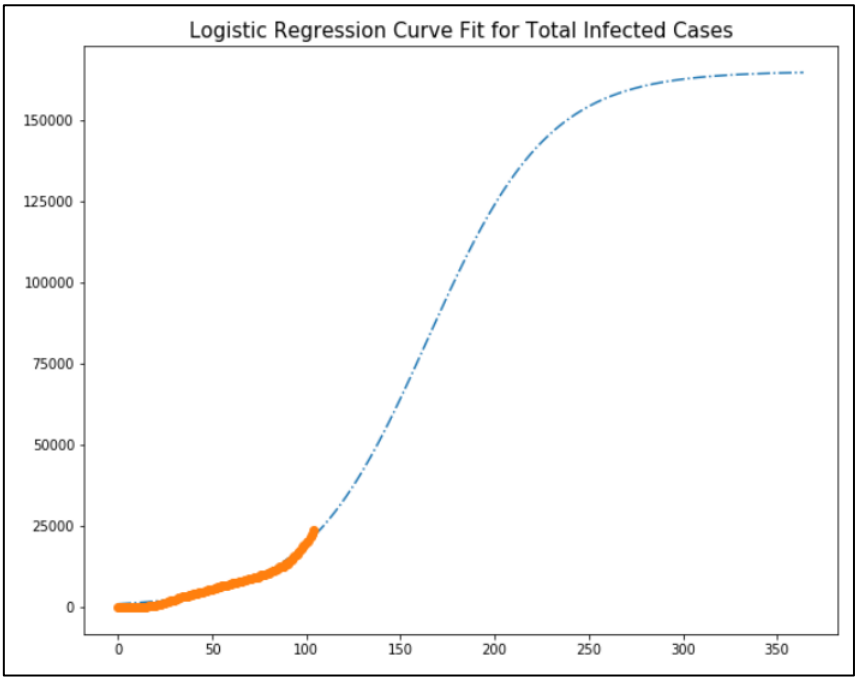
Logistic Regression works on a function known as a **sigmoid function**, which is given below.



Logistic regression's cost function is such that it predicts **0 if the hypothesis is less than 0.5** and **1 if it is greater than (or equal to) 0.5**.

I used the sigmoid function to calculate and predict a good hypothesis. The sigmoid function **fit the confirmed cases data well** and predicted the number of days and maximum number of infectives with a good accuracy. However, since the data had not peaked, finding optimal gamma and beta values proved to be difficult.

The following is the Logistic Regression fit which has fit the data quite well



This concludes my results and observations. This assignment helped me **gain insight behind the biological 'Infected model' which made me appreciate the forces at work and how statistics and machine learning can play a huge role in plotting and predicting the future of the Covid'19 virus.**

References

- Google Classroom Videos : For theory and understanding of the working formulas
- Logistic Regression : https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
- SIR Model : https://scipython.com/book/chapter-8-scipy/additional-examples/the-sir-epidemic-model/
- COVID : https://www.karger.com/Article/Abstract/507423
- South Carolina Raw Dataset : https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports_us
- South Carolina Refined Dataset- https://github.com/KartikeySharma/Covid_SIR_Model/blob/master/COVID_SIR_Model/Dataset/GithubData.csv
- South Carolina Population: https://bit.ly/3er5bnx
- **GitHub Repository - https://github.com/KartikeySharma/Covid_SIR_Model**
- Images – Google Images

Created by : Kartikey Sharma (Roll No : 1904112)

Programming Language and Technologies Used