**Specimen 'A'**

**Covid-19 Data Analysis and Sentiment Analysis**

***Project report submitted to***

***Indian Institute of Information Technology, Nagpur, in partial fulfillment of the requirements for the award of the degree***

**Bachelor of Technology**

**In**

**Department of Computer Science and Engineering**

***by***

**Arvind Kumar Sahu Manthan Chaourasia Shubham Munale**

**BT17CSE087 BT17CSE105 BT17CSE088**

Under the guidance of

**Dr. Mayuri Digalwar**



**Department of Computer Science and Engineering**

***Indian Institute of Information Technology, Nagpur* 440 006(India)**

**Year 2020-21**

**Specimen 'A'**

**Covid-19 Data Analysis and Sentiment Analysis**

***Project report submitted to***

***Indian Institute of Information Technology, Nagpur, in partial fulfillment of the requirements for the award of the degree***

**Bachelor of Technology**

**In**

**Department of Computer Science and Engineering**

***by***

**Arvind Kumar Sahu Manthan Chaourasia Shubham Munale**

**BT17CSE087 BT17CSE105 BT17CSE088**

Under the guidance of

**Dr. Mayuri Digalwar**



**Department of Computer Science and Engineering**

***Indian Institute of Information Technology, Nagpur* 440 006(India)**

**Year 2020-21**

© *Indian Institute of Information Technology, Nagpur* (IIIT) 2020 (Times New Roman 10)

 **Specimen- B**

**Department of**

**Computer Science and Engineering**

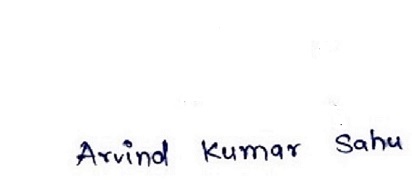
Indian Institute of Information Technology, Nagpur

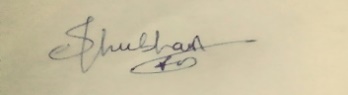
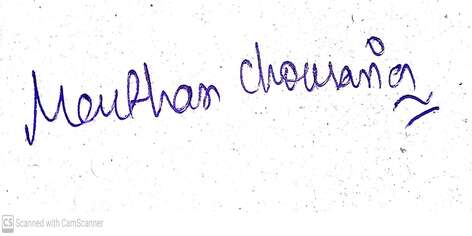
**Declaration**

We, **Arvind Kumar Sahu, Manthan Chourasia and Shubham Munale**, hereby declare that this project work titled “**Covid-19 Data Analysis and Sentiment Analysis**” is carried out by me in the Department of **Computer Science and Engineering** of **Indian Institute of Information Technology, Nagpur**. The work is original and has not been submitted earlier whole or in part for the award of any degree/diploma at this or any other Institution /University.

**Date: 05th Nov, 2020**

**Arvind Kumar Sahu Manthan Chaourasia Shubham Munale**

 **BT17CSE087 BT17CSE105 BT17CSE088**



** Specimen- C**

**Department of**

**Computer Science and Engineering**

Indian Institute of Information Technology, Nagpur

**Declaration**

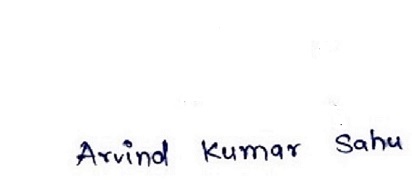
We, **Arvind Kumar Sahu (BT17CSE087), Manthan Chourasia (BT17CSE105) and Shubham Munale (BT17CSE088),** understand that plagiarism is defined as any one or the combination of the following:

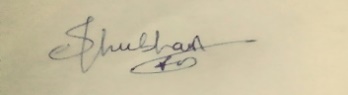
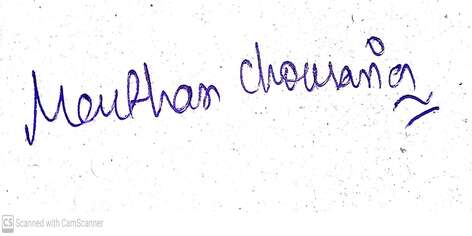
1. Uncredited verbatim copying of individual sentences, paragraphs or illustrations (such as graphs, diagrams, etc.) from any source, published or unpublished, including the internet.
2. Uncredited improper paraphrasing of pages or paragraphs (changing a few words or phrases, or rearranging the original sentence order).
3. Credited verbatim copying of a major portion of a paper (or thesis chapter) without clear delineation of who did or wrote what. (Source: IEEE, the institute, Dec.2004)I have made sure that all the ideas, expressions, graphs, diagrams, etc. that are not a result of my own work, are properly credited. Long phrases or sentences that had to be used verbatim from published literature have been clearly identified using quotation marks.

I affirm that no portion of my work can be considered as plagiarism and I take full responsibility if such complaint occurs. I understand fully well the guide of the thesis may not be in a position to check for possibility of such incidences of plagiarism in this body of work.

**Date : 05th Nov, 2020**

**Arvind Kumar Sahu Manthan Chaourasia Shubham Munale**

 **BT17CSE087 BT17CSE105 BT17CSE088**



**Specimen- D**

**Certificate**

This is to certify that the project titled “**Covid-19 Data Analysis and Sentiment Analysis**”, submitted by **Arvind Kumar Sahu, Manthan Chourasia and Shubham Munale** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering**, IIIT Nagpur. The work is comprehensive, complete and fit for final evaluation.

**Dr. Mayuri Digalwar**

**( Supervisor )**

(Expert Name and Sign) (Expert Name and Sign)

**Dr. Jitendra Tembhurane**

Name and sign of HoD

IIIT, Nagpur

# Acknowledgements

I would like to thank the one who made this dissertation possible, my sincere appreciation and gratitude to my academic supervisor, **Dr. Mayuri Digalwar**. I thank him for believing in me and guiding with patience. Her invaluable advice, unwavering trust, and unconditional support helped immensely in the timely and successful completion of the project.

I am grateful to **Dr. O.G. Kakde** Director, all the faculties, Computer Science and Engineering Department, Indian Institute of Information Technology, Nagpur for extending the departmental facilities for my research work.

I would also like to thank Manthan Chaurasia, Pranay Fating, Pradnyshil Ghajbiye, Aman Priyadaarshi and all those who have helped me directly or indirectly during the completion of this project.

Lastly, I would like to thank my family membersspecially my momfor helping me realize my own potential and their continuous moral support.

# Abstract

COVID-19 has been recognized as pandemic. The number of corona virus positive patient among all countries is more than 63 million and deaths is 1.475 million. In India only more than 9.9 million people infected with this virus and more than 140k led to death. Many govt. agencies and research organization are involved to prevent this pandemic and its impact. The main aim of this project is to draw a statistical model for better understanding of COVID-19 in India and through the world by thoroughly studying the reported cases in the countries till Nov 2020. An Exploratory Data Analysis (EDA) concept has been used for analyzing the COVID cases. The dataset are collected from various institutions like John Hopkins University, WHO, ICMR (Indian Council of Medical Research, India), Twitter Dataset, Git hub respiratory and other sources too. The result of the analysis divulge the impact of COVID-19 in India on daily and weekly manner, analogize India with abutting countries as well as with the countries who are badly affected using machine learning (ML). Also we have analyzed the sentiment of people of India as well as all states about COVID-19 using sentimental analysis, in which we have predicted sentiment of people are positive, negative or neutral.

**Keywords:** COVID-19, Exploratory Data Analysis (EDA), sentimental analysis, abutting countries analysis, Healthcare sector analysis, Machine Learning (ML).

# Contents

**List of Figures iii**

**List of Tables Iv**

1. **Introduction** 1
   1. Motivation 2
   2. Objective 2

**2 Literature Review** 5

**3 Data and Model** 10

* 1. Data Source 7
  2. Data Description 7
  3. Statistical Models 9
  4. Predictive Algorithm/Model Selection 10

3.4.1 Regression Analysis 10

3.4.2 SIR Predictive Modeling 11

3.4.3 SIR-D predictive Modeling 12

3.4.4 SIR-F predictive Modeling 13

1. **Data Visualization 15**
   1. 15
   2. Proposed Algorithm 15
   3. Modelling as an optimization model 15
   4. Modelling as a control system 19
   5. Architecture 20
2. **Results 23**
   1. Experimental setup 23
   2. Experimental results 24

ii

1. **Conclusion and Future work** 27

**Bibliography** 28

iii

List of Figures

4.1 Figure 5.1 COVID cases over time 15

4.2 Figure 5.2 COVID deaths over time 16

4.3 Figure 5.3 COVID recovery over time 16

4.4 Figure 5.4 COVID active cases over time 17

4.5 Figure 5.5 Top five countries in Confirmed cases 18

4.6 Figure 5.6 Top five countries in confirmed deaths 18

4.7 Figure 5.7 COVID-19 confirmed recoveries 19

4.8 Figure 5.8 US confirmed cases over time 19

4.9 Figure 5.9 US daily increase in Confirmed cases 20

4.10 Figure 5.10 US daily increase in Deaths 20

4.11 Figure 5.11 India Confirmed cases 21

4.12 Figure 5.12 India Daily Increases in confirmed cases 21

4.13 Figure 5.13 India Daily Increases in Deaths 22

5.1 Figure 5.1 SVM Model prediction 24

5.2 Figure 5.2 Polynomial model prediction 25

5.3 Figure 5.3 SIR prediction in April without lockdown 27

5.4 Figure 5.4 SIR prediction in May without lockdown 27

5.5 Figure 5.5 SIR prediction in best case 28

5.6 Figure 5.6 SIR prediction in worst case 29

5.7 Figure 5.7 SIR prediction 29

5.8 Figure 5.8 Effect of rho ( 

|  |  |
| --- | --- |
| iv  List of Tables |  |
| Table 3.1 Confirmed positive cases country wise | 7 |
| Table 3.2 Confirmed death cases country wise | 7 |
| Table 3.3 Recovered cases country wise | 8 |
| Table 3.4 Covid-19 India state wise | 8 |
| Table 3.5 COVID-19\_Cases\_Summarized\_by\_Age\_Group | 8 |
| Table 3.6 State wise Testing Details | 9 |
| Table 4.1 State wise analysis | 22 |
| Table 5.2 Polynomial prediction | 26 |
| Table 5.3 Bayesian Ridge Regression Prediction | 26 |
|  |  |
|  |  |
|  |  |
|  |  |

# 1 Introduction

The coronavirus COVID-19 pandemic is the defining global health crisis of our time and the greatest challenge we have faced since World War II. COVID-19 is pandemic which have been already impacted each and every kind of people of more than 220 countries across the world and continent except Antarctica. Basically this virus is severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2; formerly called 2019-nCoV), which was first identified amid an outbreak of respiratory illness cases in Wuhan City, Hubei Province, China on 31st December 2019. The first case outside of China was reported in Thailand on 13 January 2020. WHO declares COVID-19 outbreak as a Public Health Emergency of International Concern (PHEIC) by WHO on 30 January 2020. Due to this virus we have now reached the tragic milestone of more than 1.5 million deaths, and the human family is suffering under an almost intolerable burden of loss. Not only men but animals has also effect of this.

. Governments and other legislative bodies rely on insights from prediction models to suggest new policies and to assess the effectiveness of the enforced policies. But the pandemic is much more than a health crisis, it's also an unprecedented socio-economic crisis. Stressing every one of the countries it touches, it has the potential to create devastating social, economic and political effects that will leave deep and longstanding scars. The novel Coronavirus disease (COVID-19) has been reported to infect more than 64,948,823 people, with more than 1,501,535 confirmed deaths worldwide. This pandemic has impacted people not only physical appearance but also on mental strength, so there is to be analyzed the sentiment of people.

Every day, people are losing jobs and income, with no way of knowing when normality will return. Small island nations, heavily dependent on tourism, have empty hotels and deserted beaches. [The International Labour Organization](https://www.ilo.org/global/about-the-ilo/newsroom/news/WCMS_749398/lang--en/index.htm) estimates that 400 million jobs could be lost. The World Bank projects a US$110 billion decline in remittances this year, which could mean 800 million people will not be able to meet their basic needs.

The structure of this model is as follows; Section – 1 introduces COVID-19 and explains the significance of this work. Section – 2 describes about previous work done, Section 3 describes about data taken for analysis and mathematical models for analysis. Section 4 describes about Data visualization and EDA for all the data used in this project. Section 5 tells about result of analysis and prediction. Section 6 describes about Conclusion and future work.

## 1.1 Motivation

It is during adversities that real heroes emerge. They are heroes because they see adversity as an opportunity to solve a problem, or to overcome a problem.  It’s the courage, strength and wisdom to think clearly in the times of crisis that makes an individual emerge as a hero – to a family, to a community, to a nation and to the world.

On the other hand, the one who looks at adversity as a problem will never emerge as a hero. Because this individual has allowed a problem to enter into his head and is busy either figuring out how to remove the problem from his head or anticipating restlessly when the problem would end so that he can again live peacefully.  These individuals look at the problem as a roadblock to their goals. They never realise that this problem is indeed their golden opportunity to reach their goal faster and also emerge as a hero.

Our society is in the era of unbelievable attempts to struggle upon the spread of this life-threatening condition in terms of infrastructure, finance, business, manufacturing, and several other resources. Artificial Intelligence (AI) researchers strengthen their proficiency in developing mathematical paradigms for investigating this pandemic using nationwide distributed data. This article intends to apply the machine learning models simultaneously with the forecast of expected reachability of the COVID-19 over the nations by using the real-time data from the Johns Hopkins dashboard.

 The recent global COVID-19 pandemic has exhibited a nonlinear and complex nature. In addition, the outbreak has differences with other recent outbreaks, which brings into question the ability of standard models to deliver accurate results. Besides the numerous known and unknown variables involved in the spread, the complexity of population-wide behavior in various geopolitical areas and differences in containment strategies had dramatically increased model uncertainty. Consequently, standard epidemiological models face new challenges to deliver more reliable results.

To overcome this challenge, many novel models have emerged which introduce several assumptions to modeling (e.g., adding social distancing in the form of curfews, quarantines etc.). This pandemic has impacted people not only physical appearance but also on mental strength, so there is to be analyzed the sentiment of people so that we could help the people who are suffering or might suffer in the future, man has to fit mentally and physically both.

Do you want to sit and still check the number of covid-19 cases that are adding up, or prepare for this unique pandemic with more vigour and energy to prevent such cases from ever happening again or to minimize the effect of it?

## 1.2 Objective

Currently, solving problem data plays a very important role. In industry to reduce cost and maximize profit, data analysis is very useful. Now we come to the point, COVID-19 cases are increasing rapidly on daily basis. The purpose of this project is to quantitatively analyze the impact of the COVID-19 pandemic on our societies in the form of people’s mobility, health, countries economy and sentiment of people. How the policy makers are struggling with their opinions and doctors are busy with their related work. So with the help of these analysis, they can better understand the situation and can react accordingly.

The main aim of this project is to study and analyze the COVID-19 spread in the world and India the day of spreading virus. We shall understand how the situation changed from epidemic to pandemic. We will analyze the effect of Government rules Lockdown, Partial Lockdown and No Lockdown.

At first we collected data from various resources kaggle.com github.com WHO, worldometer.com, John Hopkins University, from various other sources. The datasets are varies for different-different analysis. In this project detailed visualization have been done, graphing number of different-different visualization for active cases, deaths, recoveries, mortality rate (CFR) and recovery rate, country specific graph. Prediction for confirmed cases world-wide. For specific India, where a detailed analysis for states and city wise. Mental health of people of India have been analyzed using twitter dataset. We have done analysis for each and every states in India and for India giving positive or negative sentiment of people.

# 2 Literature Review

The transmission trend of COVID-19 from China to other countries, confirmed cases on daily basis, surveillance strategy of India, China, America, South Korea, Japan, Italy, Brazil, Iran Spain and many other countries from the first day of outbreak. Along with the effect of government policies of the above countries in controlling the COVID-19 outbreak by finding the linear relation between outbreak condition and “case fatality rate (CFR)”. And these were analyzed by taking global statistics such as confirmed, death and recovered cases and making prediction with respect to China using “Linear Regression”. However these days confirmed cases are increasing in some other manner like exponentially, polynomial or the other way.

Many researchers and other organization collected data from Ding Xiang Yuan, John Hopkins University and WHO, then they did analysis and prediction for COVID-19 cases all over the world. The data sets are uploaded on corona tracker website and many other organization and authorities too.  For prediction, the Susceptible-Exposed-Infected-Removed (SEIR) model was used. They also provide sentiment analysis of news on Covid-19, and they found 561 positive articles and 2548 negative articles.

In this analysis, researchers provided an effect of comorbidity on Covid-19 patients. They analyzed 1590 confirmed cases in China hospitalized in different hospitals. A total of 686 female patients, 399 patients had comorbidities. In this research, they found that comorbidity plays a crucial role in clinical treatment, and patients with comorbidities have poor clinical outcomes. Another study showed that the most common symptoms of covid-19 were fever, cough, expectoration, headache and myalgia or fatigue.

In a performed a clinical prediction of mortality of covid-19 based on 150 patients in Wuhan city, China. Of these 150 cases, 68 and 82 were deaths and discharges, respectively. In this study, they found that there is a significant difference between age in death cases and discharge cases. Forty-three out of 68 deaths had comorbidities, and in discharge cases, 34 out of 82 had comorbidities. Sixty-three patients died due to respiratory failure or myocardial damage. Only 5 patients died without any known cause.

# 3 Data and Model

3.1 Data Source

For COVID-19 data analysis, data is collected from verified source such as WHO, Ding Xiang Yuan (https://dataconomy.com/2020/04/apis-to-track-coronavirus-covid-19/), John Hopkins University [20], ICMR (Indian Council of Medical Research, https://www.icmr.gov.in/) [21], Kaggle.com [22].For sentiment analysis data collected from IBM competition.

3.2 Data Description

|  |  |  |  |
| --- | --- | --- | --- |
| Name of Country country | 22/01/2020 | 23/01/2020 | 24/01/2020 |
| Afghanistan | 0 | 0 | 0 |
| China | 444 | 444 | 549 |

(A) Time series covid-19 confirmed case:- In this data set, there are 316 columns containing date starting from 22nd Jan 2020 and in row there are names of countries in which corona positive patient found.

*Table 3.1 Confirmed positive cases country wise*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name of country | 22/01/2020 | 23/01/2020 | 24/01/2020 | 25/01/2020 |
| Afghanistan | 0 | 0 | 0 | 0 |
| China | 17 | 17 | 24 | 40 |
| India | 0 | 0 | 0 | 0 |
| US | 0 | 0 | 2 | 1 |

(B) Time series covid-19 death case:- In this data set, there are 316 columns containing date starting from 22nd Jan 2020 and in row there are names of countries in which corona dead patient found.

*Table 3.2 Confirmed death cases country wise*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name of Country country | 22/01/2020 | 23/01/2020 | 24/01/2020 | 25/01/2020 |
| Afghanistan | 0 | 0 | 0 | 0 |
| China | 28 | 28 | 31 | 32 |

(C) Time series covid-19 recovered case:- In this data set, there are 316 columns containing date starting from 22nd Jan 2020 and in row there are names of countries in which corona dead patient found.

*Table 3.3 Recovered cases country wise*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Date* | *State/Union* | *Confirmed(I)* | *Confirmed(F)* | *Cured* | *Deaths* | *Confirmed* |
| *31/01/2020* | *Kerala* | *1* | *0* | *0* | *0* | *1* |
| *01/02/2020* | *Kerala* | *1* | *0* | *0* | *0* | *1* |
| *02/02/2020* | *Kerala* | *2* | *0* | *0* | *0* | *2* |

For visualization in India (D) covid\_19\_india dataset: It contains all states with their states and how much corona positive (Indian, Foreign), deaths, recovered and confirmed (both Indian and Foreign positive patient)

*Table 3.4 Covid-19 India state wise*  (E) COVID-19\_Cases\_Summarized\_by\_Age\_Group: This data contains age group in one column and other columns consist of confirmed New and Cumulative

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Age-group | Confirmed New | Confirmed Cumulative |
| 12/03/2020 | 51-60 | 2 | 6 |
| 13/03/2020 | 51-60 | 3 | 9 |
| 14/03/2020 | 51-60 | 1 | 10 |

*Table 3.5 COVID-19\_Cases\_Summarized\_by\_Age\_Group*

### (F) State wise Testing Details: This table of data contains date on which testing performed and columns consisting of Total Sample tested, Negative and positive found patient.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| DatE | State | | Total Sample | Negative | Positive |
| 02/04/2020 | AP | 1800 | | 1175 | 132 |
| 10/04/2020 | AP | | 6374 | 6009 | 365 |
| 11/04/2020 | AP | | 6958 | 6577 | 381 |

*Table 3.6 State wise Testing Details*

# 3.3 Statistical Models

## Machine Learning

We all know, this decades and next coming decades are of Machine Learning (ML) and Artificial Intelligence (AI) which have capability of automatically learn and improve from experience without being explicitly programmed. [10] This is tool to focus on the development of computer programs that can access data and use it learn for themselves. Thus, it is widely used for data analysis in various domains like financial sector, business sector, education sector, medical engineering and these days even in sports. It comes under Artificial Intelligence which teaches machines from training datasets. Through machine learning, we can identify patterns, analyze data, and make correct decisions with no human intervention or less human intervention. Machine learning is broadly categorized into three parts which are given below:

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

Superior learning means that a machine or model teaches the teacher, or in other words, we can say that the machine or model learns through a training dataset. In supervised learning, class-level information is available in the training datasets.

Whereas unsupervised learning means-learning without a teacher or in other words learning algorithms learn dynamically with help partitioning or clustering algorithm. Most of the clustering algorithms are available in literature such as K-Means, Fuzzy C-Means, hierarchical clustering methods, and so on.

Reinforcement learning is a combination of supervised and unsupervised learning methods. It is about taking suitable action to maximize reward in a particular situation. In the absence of a training dataset, it is bound to learn from its experience. Example: The problem is as follows: We have an agent and a reward, with many hurdles in between.

## 3.4 Predictive Algorithm/Model Selection

## 3.4.1 Regression Analysis

Regression analysis is a part of machine learning or in other words, regression analysis is a subset of machine learning algorithms. It is the first machine learning algorithm. Regression analysis inventor says that “Regression analysis consists of a set of machine learning methods that allow us to predict a continuous outcome variable (Y) based on the value of one or multiple predictor variables (X). It assumes a linear relationship between the outcome and the predictor variables”. Let us consider equation straight line connecting any two variables X and Y can be stated algebraically as:

*Y = a X + b* - - - (1) (Linear Regression)

Where b is called the intercept on the y-axis and ‘a’ is called the slope of the line. Here ‘a’ and ‘b’ are also called the parameters of regression analysis. These parameters should learn through proper learning methods. This is called Linear Regression Analysis.

In this proposed, we have proposed six regression analysis based models known as exponential, quadratic, third degree fourth degree, fifth degree polynomial. The description of these models is given below:

***Y = aebx + b*** - - - (2) (Exponential Reg.)

*Y = aX2 + bX + c* - - - (3) (Polynomial Reg.)

*Y = a X 3+ bX2 + cX + d* - - - (4)

## 3.4.2 SIR Predictive Modeling

In this section we will discuss about another simplest predictive modeling which is Susceptible-Infected-Removed (SIR) that will describe the COVID-19 outbreak.

**What is in it?**

* S : Susceptible = (Total – Confirmed)
* I : Infected = (Confirmed – Recovered – Deaths)
* R : Recovered or fatal = (Recovered + Deaths)

Note: SIR model is not the general model

Here Recovered is sum of recovered and fatal, means the people having immunity either they died or get rid of decease. And so mortality rate can’t be ignored in real data.

Procedure of conversion:

**S** **I** **I** **** **R**

Where ****: is effective contact rate [1/min] and ****: Recovery (+mortality) rate [1/min]

Ordinary Differential Equation (ODE) will be like

**- I**

**I**

Where N = S **+** I **+** R is total population of data set and T is elapsed time from beginning of decease.

**3.4.3 SIR-D predictive Modeling**

In this modeling we consider fatality rate and recovered cases as different situation. Like in previous mathematical modeling where recovered was sum of recovered and death.

**What is in it?**

* S : Susceptible = (Total – Confirmed)
* I : Infected = (Confirmed – Recovered – Deaths)
* R : Recovered (Only Recovered and no Deaths)
* D: Deaths

Procedure of conversion:

**S** **I** **I** **** **R**

**I  D**

Where is Mortality Rate (Fatality Rate) is Effective contact rate is Recovery Rate

and Ordinary Differential Equation (ODE) will be like

**– (**I

**I**

****

Where N = S **+** I **+** R **+** D is total population of data set and T is elapsed time from beginning of decease.

**3.4.4 SIR-F predictive Modeling**

In the case of any decease some patient die before clinical diagnosis and that too happened in the case of COVID-19 pandemic. So we will consider this issue as

“S + I → Fatal + I” and this will be summed in this model than previous one.

**What is in it?**

* S : Susceptible
* S\*: Confirmed and Uncategorized
* I : Confirmed and Categorized as I
* R : Recovered
* F: Fatal with the confirmation

**So** here equation will be like Confirmed = I + R + F Recovered = R Deaths = F

**Procedure of conversion:**

Where 1: Mortality rate of S\* cases 2: is mortality rate of I cases : effective contact rate : is recovery rate

ere Point to consider is: If (1= 0) then this SIR-F model is same as SIR-D model

Ordinary Differential Equation (ODE) will be like

**– (**I

**I**

**N-1 S I + 2 I**

Where N = S + I + R + F is total population of data set and T is elapsed time from beginning of decease.

After we have stored the news inside the Corona Tracker database, we extract news description as it contains a summary of the news that is neither too short nor too long, which can be bad for the model we are going to use otherwise. We only select descriptions that are at least more than 8 words, and discard non-English descriptions because the pre-trained model we use have been trained on SST-2 [25], which is a dataset for sentiment analysis for English language. We use a library called transformers by hugging face [26]. The input sentences will be separated by their respective polarity for further analysis like topic modelling and generating word cloud for each

# 4 Data Visualization

5.1 World data visualization

Figure 4.1 shows the Increasing number of COVID-19 positive cases over time starting from the 22nd Jan 2020 to 10th Dec 2020. Where blue line show continuous increase by every day and the doted yellow line shows the number of positive patient increased in 7 days. And figure 5.2, 5.3 and 5.4 shows the same on deaths, recovery and active cases respectively.

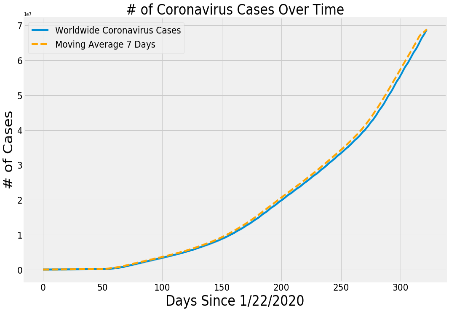


Figure 4.1 COVID-19 cases over time

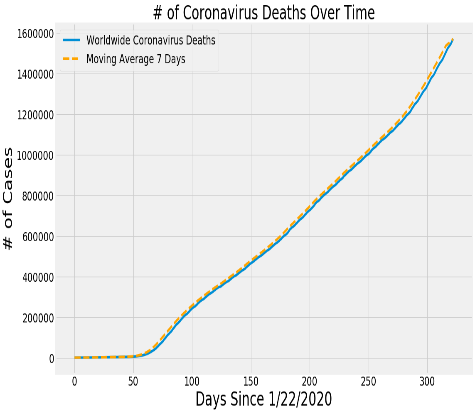


Figure 4.2 COVID deaths over time

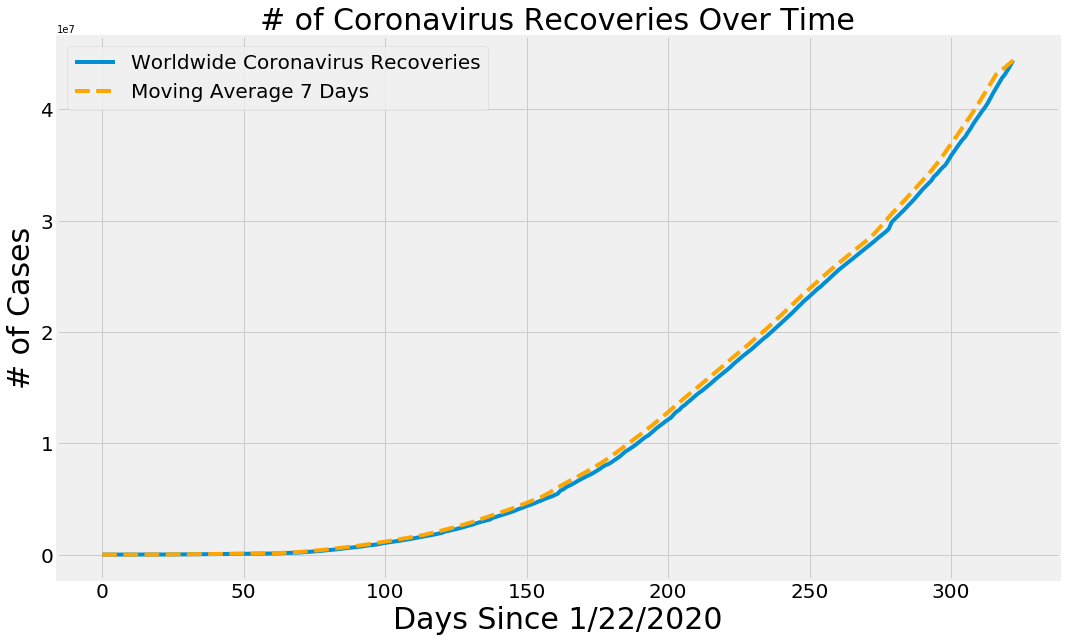


Figure 4.3 COVID-19 recovery over time

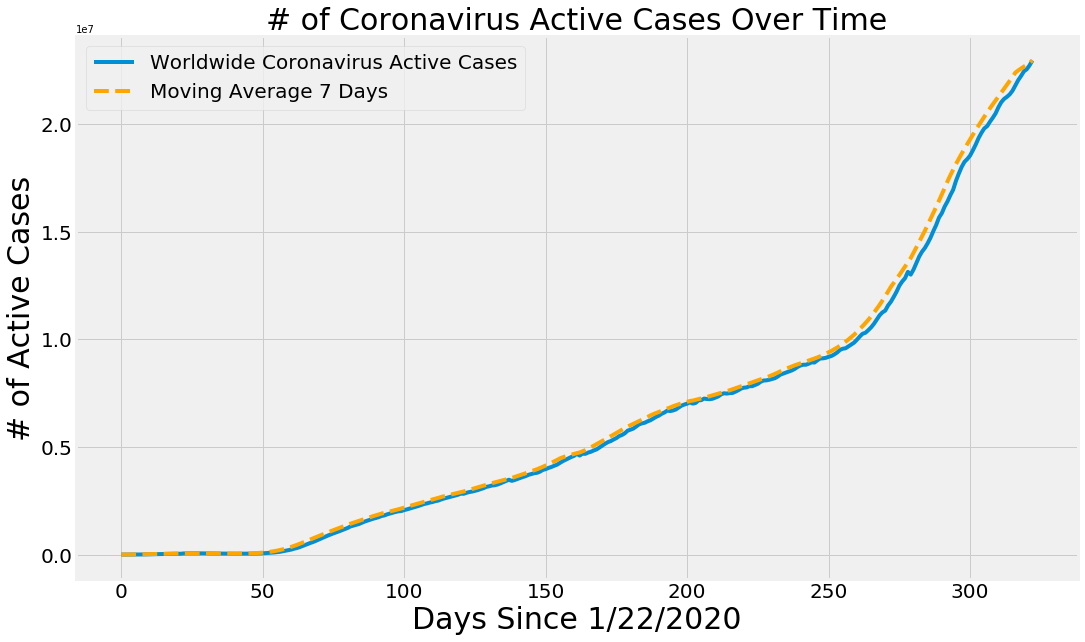


Figure 4.4 COVID-19 active cases

5.2 Top five countries in World

In the world United States of America stands with highest number of COVID-19 cases in the world with more than 16 million confirmed cases and on the 2nd position, India is with approximately 10 million cases then Brazil (6880,595), Russia (2625848) and France (2365319) are on the 3rd, 4th and 5th position respectively.

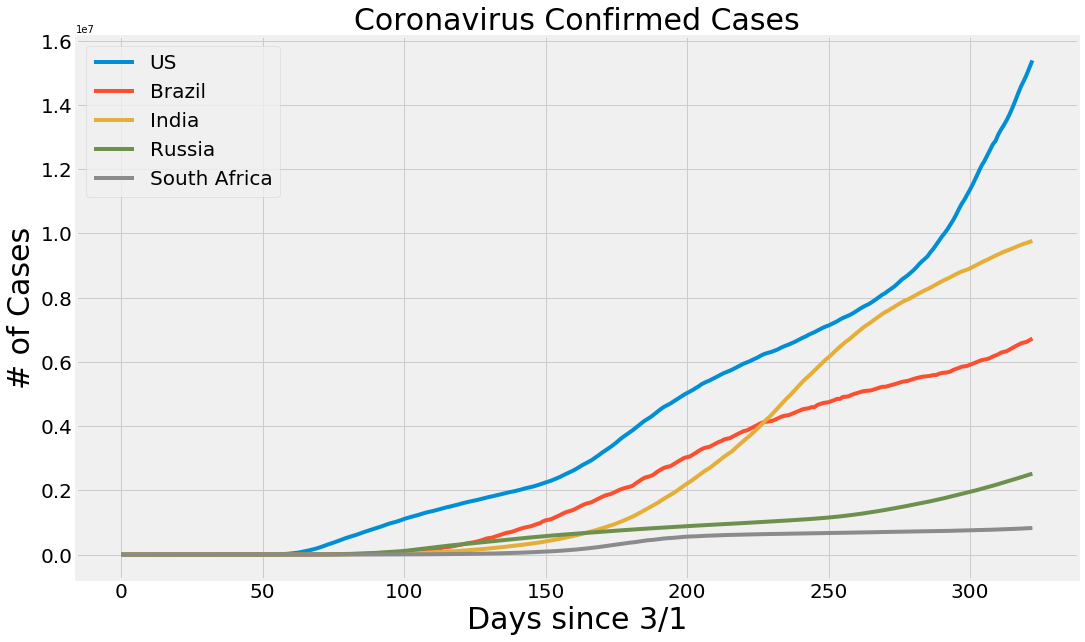


Figure 4.5 Top five countries in Confirmed cases

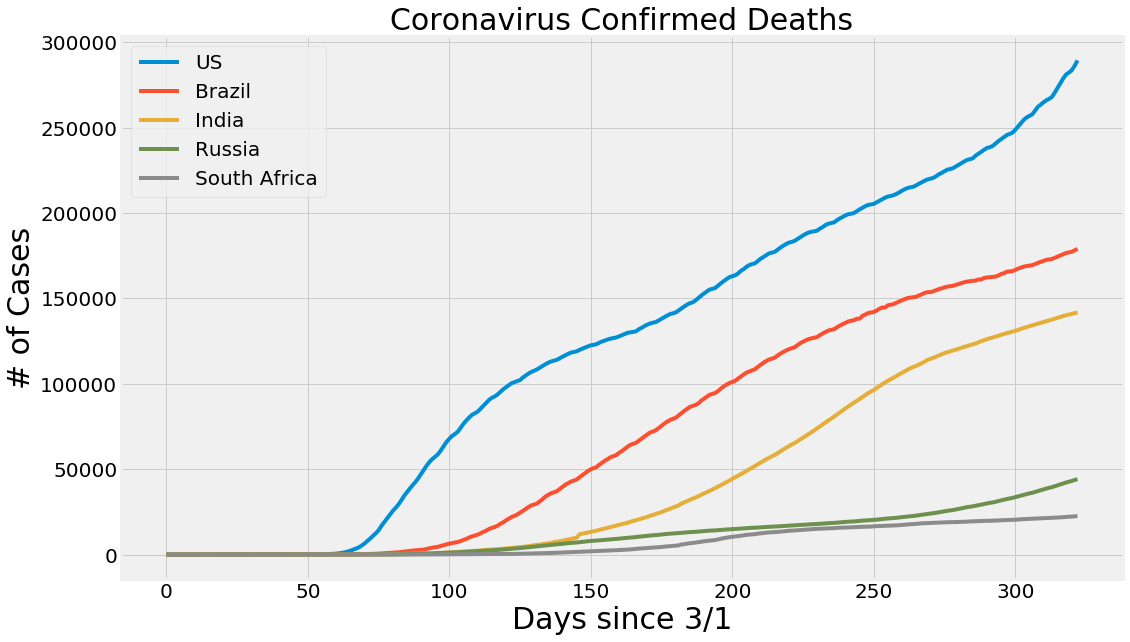


Figure 4.6 Top five countries in confirmed deaths

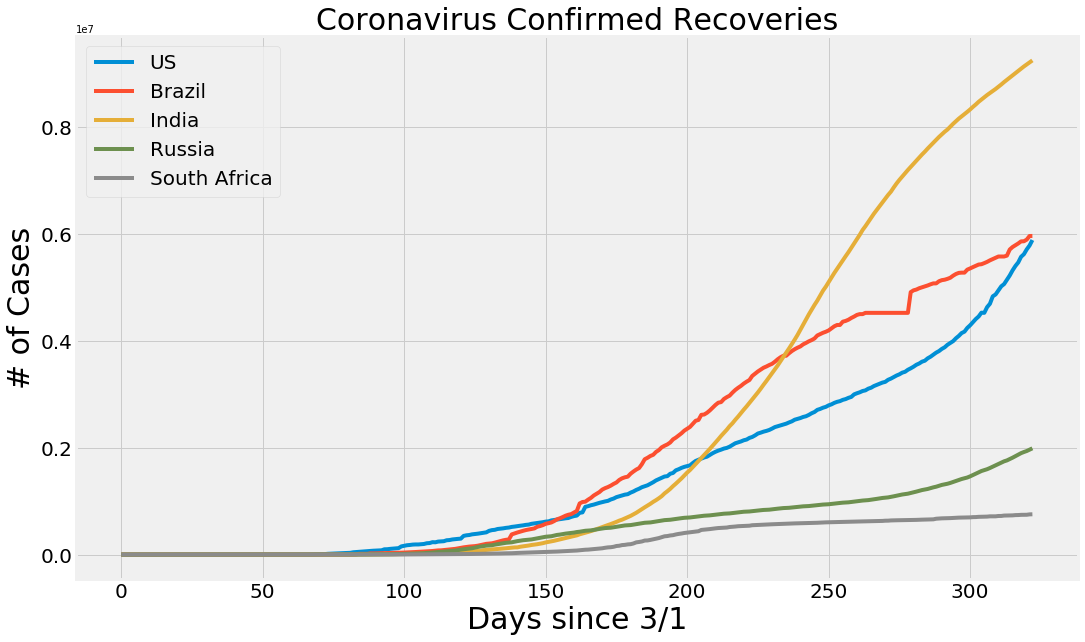


Figure 4.7 COVID-19 confirmed recoveries

5.3 US data visualization

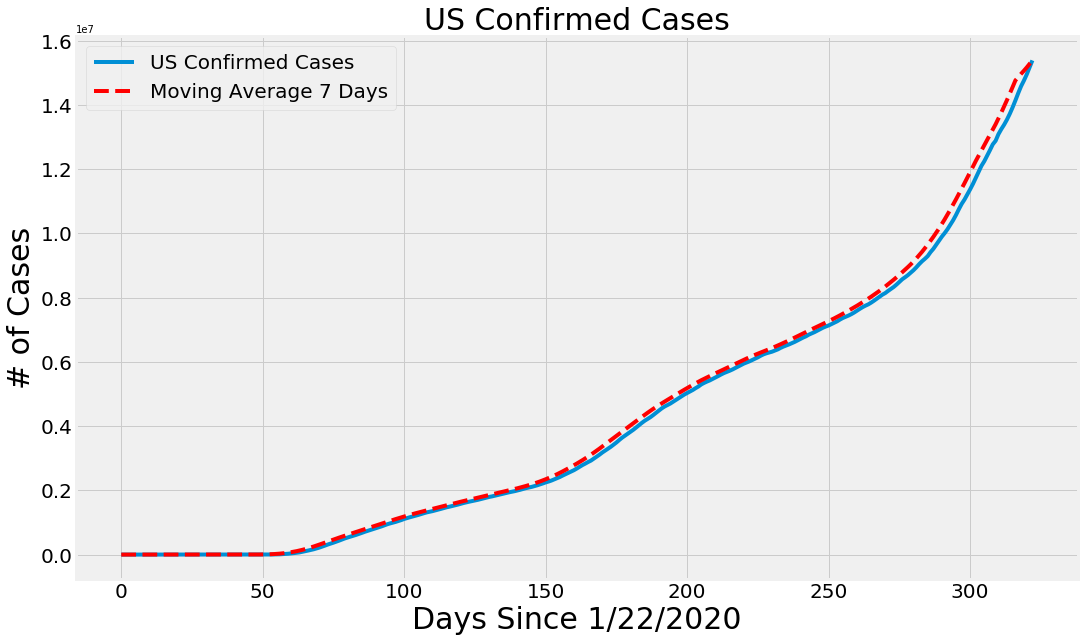


Figure 4.8 US confirmed cases over time

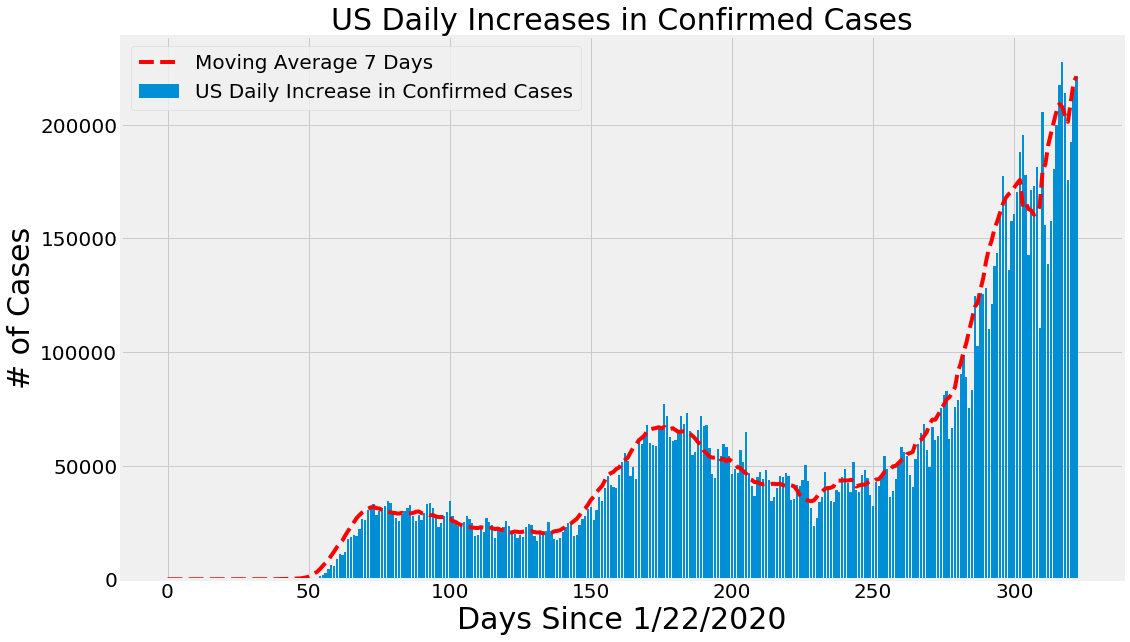


Figure 4.9 US daily increase in Confirmed cases

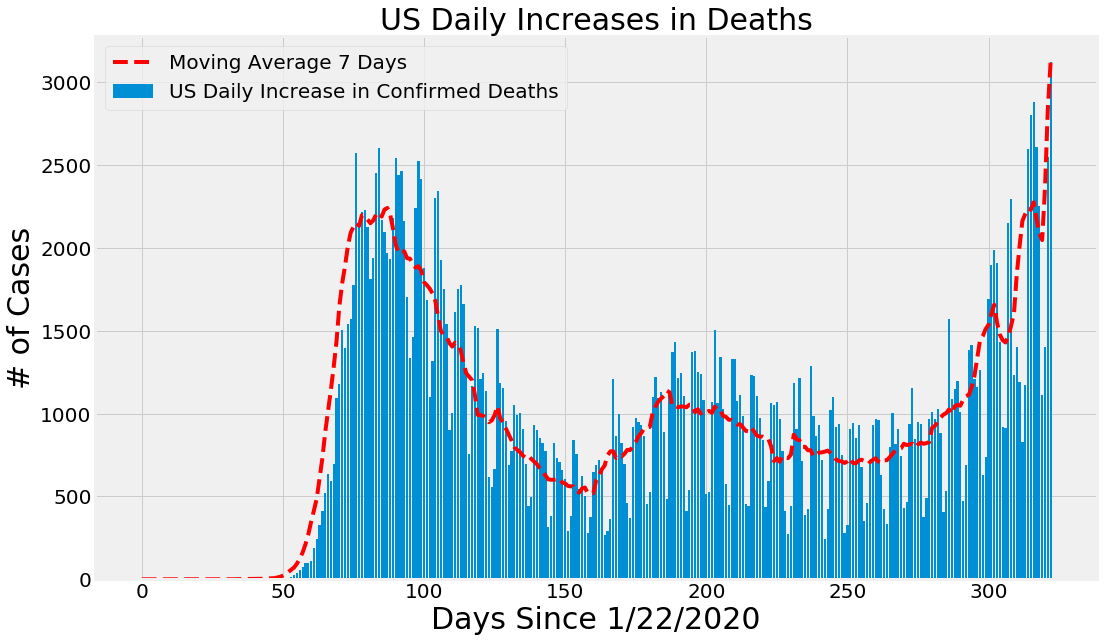


Figure 4.10 US daily increase in Deaths

5.3 India Data Visualization

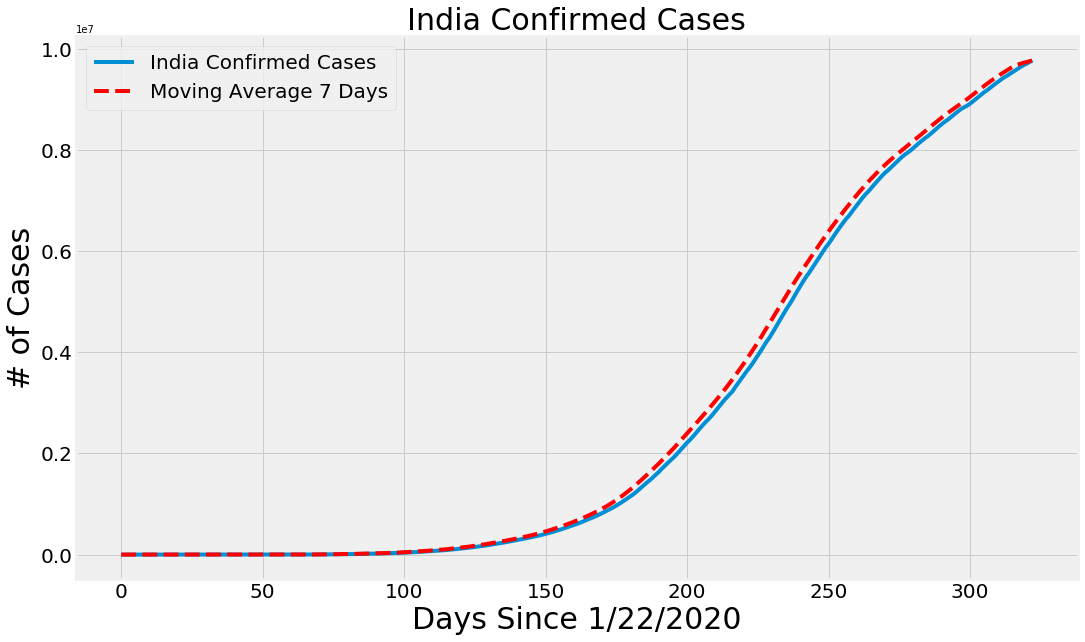


Figure 4.11 India Confirmed cases

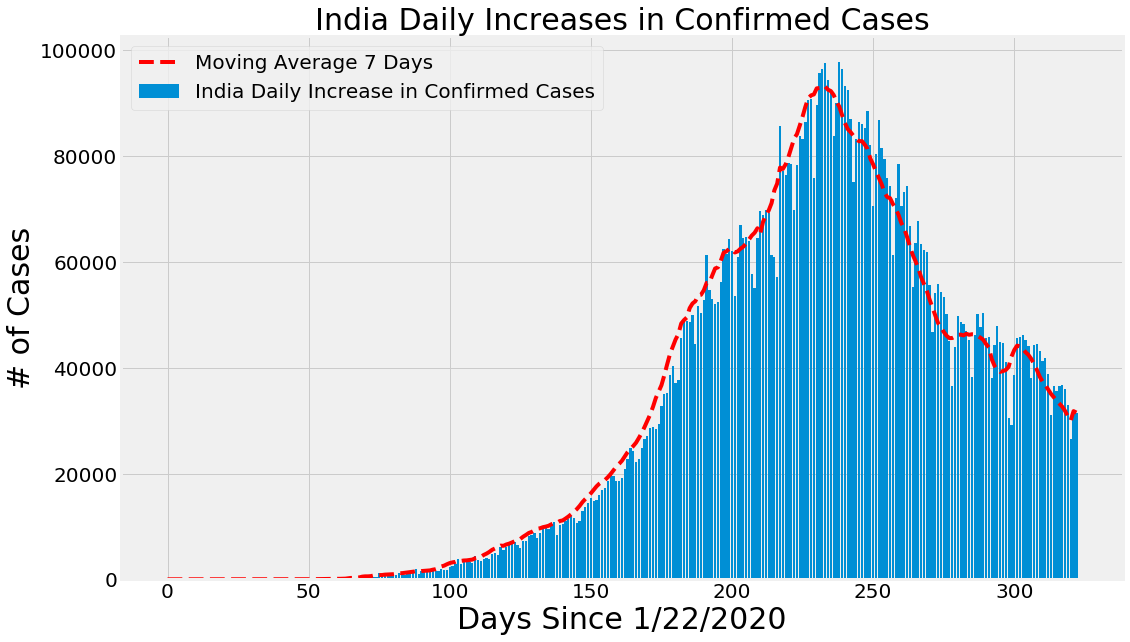


Figure 4.12 India Daily Increases in confirmed cases

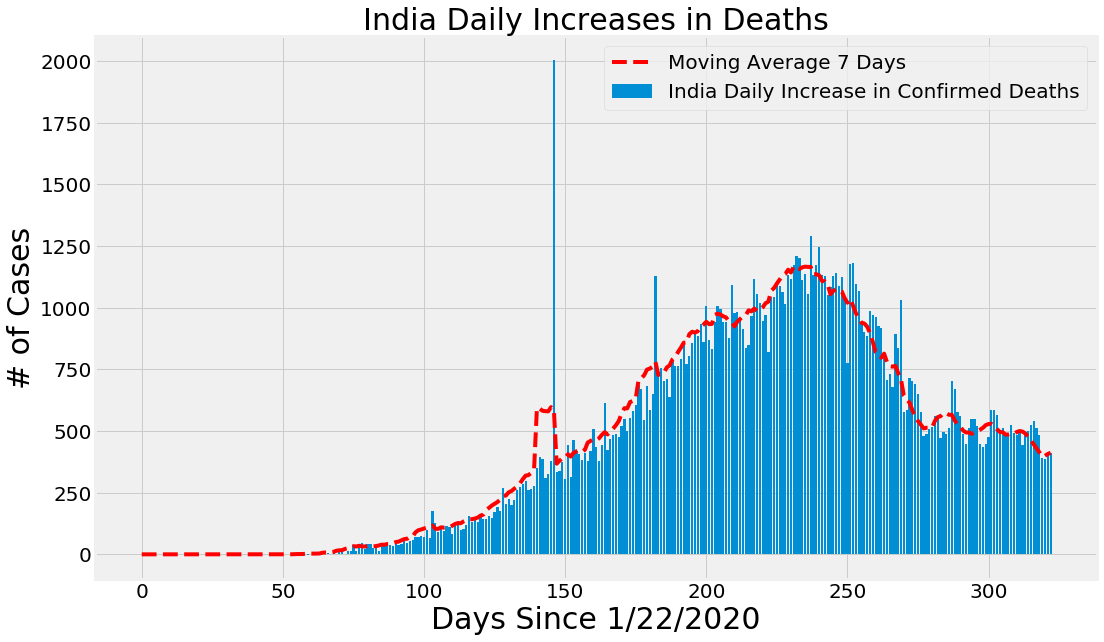


Figure 5.12 India Daily Increase in Deaths

Figure 4.13 India Daily Increases in Deaths

# 5.4 India: A State Level Analysis

Here we can visualize that Maharashtra tops among all States/UT and then Andhra Pradesh but we can clearly see that incidence rate of Delhi, AP and Maharashtra is still high and that has to be lowered and preventive measure should be raise immediately.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **State Name** | **Number of Confirmed Cases** | **Number of Deaths** | **Number of Active Cases** | Incidence rate | Mortality rate |
| **0** | Maharashtra | 1617658 | 42633 | 159346.000000 | 1313.628817 | 0.026355 |
| **1** | Andhra Pradesh | 793299 | 6508 | 32376.000000 | 1471.705130 | 0.008204 |
| **2** | Karnataka | 782773 | 10696 | 100459.000000 | 1158.587745 | 0.013664 |
| **3** | Tamil Nadu | 697116 | 10780 | 35480.000000 | 895.560963 | 0.015464 |
| **4** | Uttar Pradesh | 461475 | 6755 | 29364.000000 | 193.992649 | 0.014638 |
| **5** | Kerala | 361841 | 1232 | 93527.000000 | 1013.576038 | 0.003405 |
| **6** | Delhi | 340436 | 6128 | 24117.000000 | 1819.450693 | 0.018000 |
| **7** | West Bengal | 333126 | 6244 | 35579.000000 | 334.432618 | 0.018744 |
| **8** | Odisha | 274181 | 1181 | 18087.000000 | 591.463941 | 0.004307 |
| **9** | Punjab | 231195 | 6383 | 7303.000000 | 521.089163 | 0.027609 |
| **10** | Telangana | 227580 | 1292 | 20183.000000 | 578.161089 | 0.005677 |
| **11** | Bihar | 208377 | 1019 | 11118.000000 | 166.968849 | 0.004890 |
| **12** | Assam | 202774 | 889 | 25807.000000 | 569.477288 | 0.004384 |
| **13** | Rajasthan | 178933 | 1788 | 19185.000000 | 220.815824 | 0.009993 |
| **14** | Chhattisgarh | 167639 | 1628 | 25795.000000 | 569.498860 | 0.009711 |
| **15** | Madhya Pradesh | 163296 | 2828 | 12386.000000 | 191.305037 | 0.017318 |
| **16** | Gujarat | 162823 | 3660 | 14193.000000 | 254.919187 | 0.022478 |
| **17** | Haryana | 153367 | 1674 | 10187.000000 | 543.764137 | 0.010915 |
| **18** | Jharkhand | 98061 | 851 | 6206.000000 | 254.083879 | 0.008678 |
| **19** | Jammu and Kashmir | 89582 | 1402 | 8088.000000 | 658.385221 | 0.015650 |
| **20** | Uttara Khand | 59106 | 960 | 5085.000000 | 525.346600 | 0.016242 |
| **21** | Goa | 41339 | 557 | 3099.000000 | 2606.083530 | 0.013474 |
| **22** | Puducherry | 33622 | 580 | 4026.000000 | 2378.563920 | 0.017251 |
| **23** | Tripura | 29925 | 334 | 2339.000000 | 717.661352 | 0.011161 |
| **24** | Himachal Pradesh | 19621 | 279 | 2636.000000 | 263.300033 | 0.014219 |
| **25** | Manipur | 16267 | 124 | 3846.000000 | 526.177041 | 0.007623 |
| **26** | Arunachal Pradesh | 13912 | 31 | 2682.000000 | 885.856228 | 0.002228 |
| **27** | Chandigarh | 13795 | 209 | 744.000000 | 1190.791671 | 0.015150 |
| **28** | Meghalaya | 8621 | 77 | 1870.000000 | 256.066011 | 0.008932 |
| **29** | Nagaland | 8139 | 28 | 1683.000000 | 361.782375 | 0.003440 |
| **30** | Unknown | 7541 | 22 | 781032.000000 | 0.000000 | 0.002917 |
| **31** | Laddakh | 5781 | 68 | 848.000000 | 2107.631002 | 0.0117 |

Table 4.1 State Level Analysis

# 5 Results

5.1 Outbreak Prediction for world

COVID-19 cases are increasing continuously instead of many rules and regulation of government, also different modes of lockdown and social distancing measure world-wide by government of countries. Many guidelines have been given by World Health Organization (WHO) to reduce the impact of COVID-19 and irrespective of these rules more than 1.61 million people died due to this pandemic. If we have earlier information that how many people are going to be affected by this pandemic, we and our government will have specific plans to deal with this pandemic.

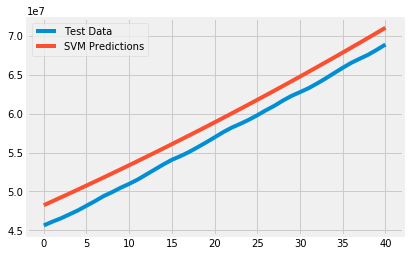
5.1.1 World Prediction using SVM

Figure 5.1 SVM Model prediction

And prediction for next 10 days:

|  |  |
| --- | --- |
| Date | SVM Predicted # of Confirmed Cases Worldwide |
| 12/10/2020 | 71698804.000000 |
| 12/11/2020 | 72352435.000000 |
| 12/12/2020 | 73010114.000000 |
| 12/13/2020 | 73671853.000000 |
| 12/14/2020 | 74337664.000000 |
| 12/15/2020 | 75007560.000000 |
| 12/16/2020 | 75681553.000000 |
| 12/17/2020 | 76359655.000000 |
| 12/18/2020 | 77041880.000000 |
| 12/19/2020 | 77728239.000000 |

Table 5.1 SVM Prediction for world data

We can see that by this model slightly over fits the expected over predicted.

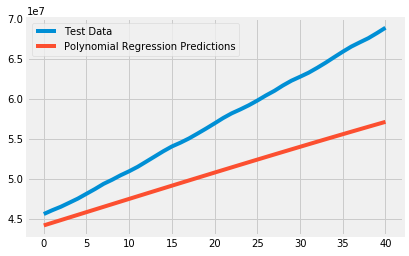
5.1.2 World Prediction Using Polynomial Regression

Figure 5.2 Polynomial model prediction

|  |  |
| --- | --- |
| Date | Polynomial Predicted # of Confirmed Cases Worldwide |
| 12/10/2020 | 57444875.000000 |
| 12/11/2020 | 57747538.000000 |
| 12/12/2020 | 58048668.000000 |
| 12/13/2020 | 58348223.000000 |
| 12/14/2020 | 58646159.000000 |
| 12/15/2020 | 58942434.000000 |
| 12/16/2020 | 59237002.000000 |
| 12/17/2020 | 59529821.000000 |
| 12/18/2020 | 59820847.000000 |
| 12/19/2020 | 60110034.000000 |

Table 5.2 Polynomial prediction

We can see that prediction curve over fits the expected and difference is increasing with time, so not so good algorithm.

5.1.3 World Prediction Using Bayesian Ridge Regression

|  |  |
| --- | --- |
| Date | SVM Predicted # of Confirmed Cases Worldwide |
| 12/10/2020 | 59214049.000000 |
| 12/11/2020 | 59582370.000000 |
| 12/12/2020 | 59950760.000000 |
| 12/13/2020 | 60319198.000000 |
| 12/14/2020 | 60687665.000000 |
| 12/15/2020 | 61056143.000000 |
| 12/16/2020 | 61424610.000000 |
| 12/17/2020 | 61793047.000000 |
| 12/18/2020 | 62161434.000000 |
| 12/19/2020 | 62529752.000000 |

Table 5.3 Bayesian Ridge Regression Prediction

5.2 Outbreak Prediction for India

5.2.1 Prediction Using SIR Model

This is a potential SIR model, if lockdown had not been imposed from 14th March to 14th April, *i.e.* for 30 days then

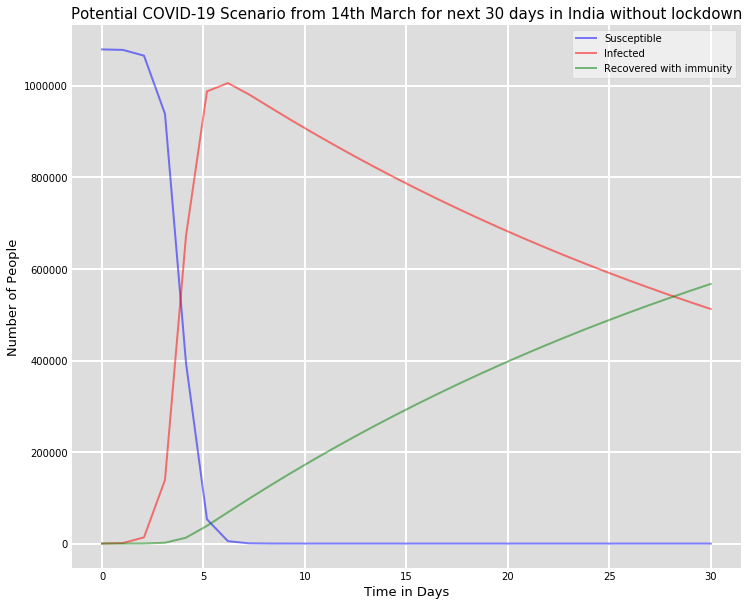


Figure 5.3 SIR prediction in April without lockdown

Here beta () = 0.22807272 and gamma () = 0.01422848

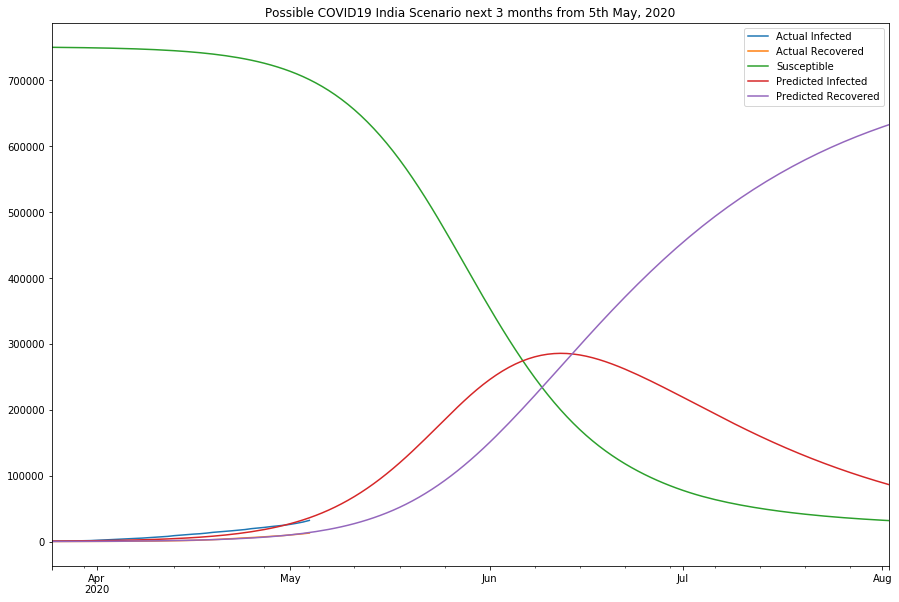
Now estimating beta () and gamma () during lockdown period and using it to predict the figures for the next three months from 5th may 2020 onwards

Figure 5.4 SIR prediction in May without lockdown

Here beta () = 0.14268499 and gamma () = 0.03824572

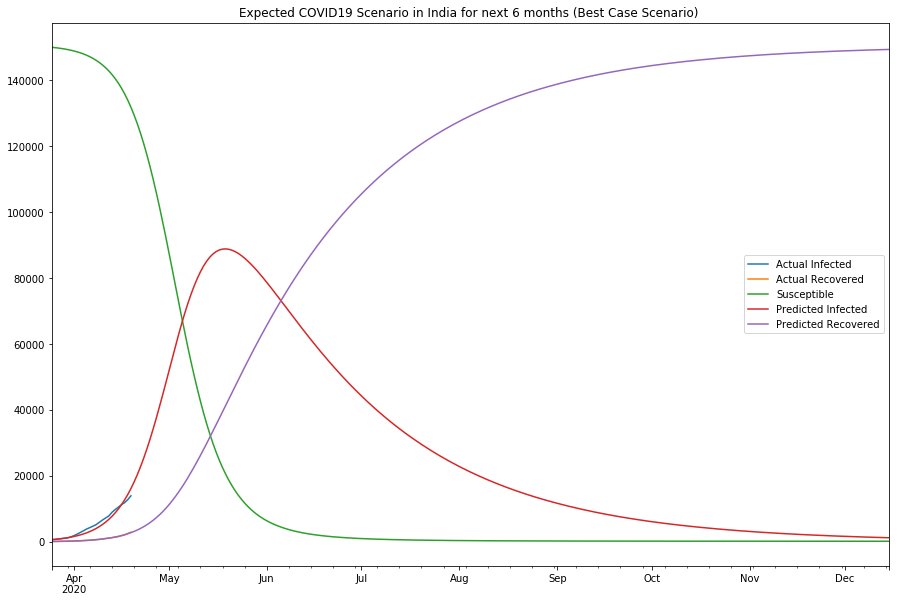
Now estimating beta () and gamma () during lockdown period and using it to predict the figures for the next three months from 5th may 2020 onwards in best case scenario

Figure 5.5 SIR prediction in best case

Here beta () = 0.15858943 and gamma () = 0.02195618

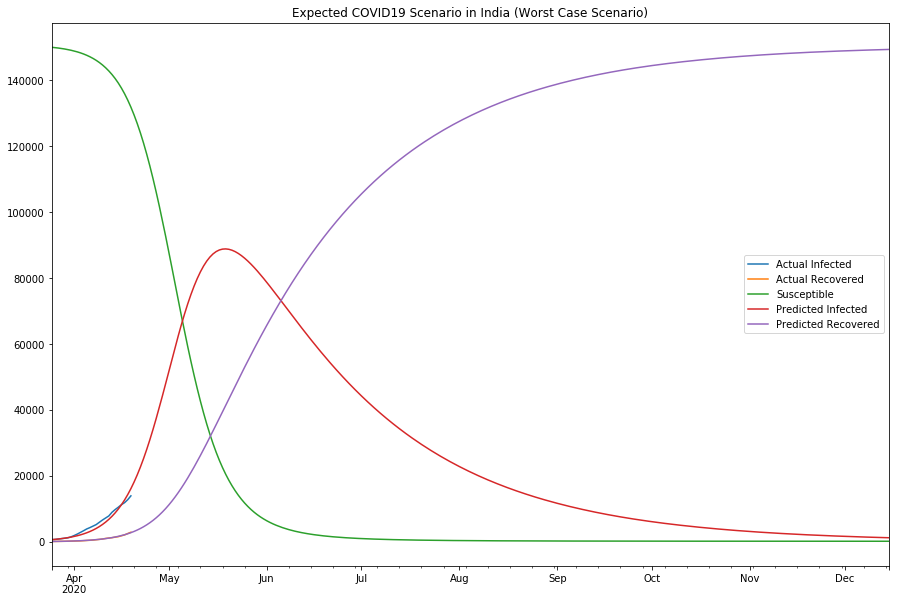
Same as above in worst case scenario:

Figure 5.6 SIR prediction in worst case

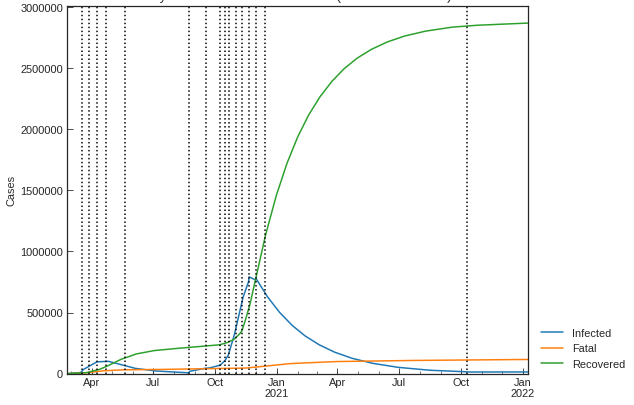
 India prediction using SIR

Figure 5.7 SIR prediction

|  |  |  |
| --- | --- | --- |
| Date | Infected | Recovered |
| 12/10/2020 | 30796 | 30080 |
| 12/11/2020 | 28640 | 32802 |
| 12/12/2020 | 26987 | 27640 |
| 12/13/2020 | 30394 | 32549 |
| 12/14/2020 | 28653 | 30643 |
| 12/15/2020 | 30122 | 29891 |
| 12/16/2020 | 29874 | 30498 |
| 12/17/2020 | 30123 | 32501 |
| 12/18/2020 | 28640 | 29845 |
| 12/19/2020 | 27899 | 30145 |

With beta () = 0.21774528 and gamma () = 0.17736864

5.2.2 Prediction using SIR-D model for India

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Infected | Death | Recovered |
| 12/10/2020 | 28996 | 361 | 29880 |
| 12/11/2020 | 28640 | 328 | 31862 |
| 12/12/2020 | 27987 | 354 | 33120 |
| 12/13/2020 | 29399 | 389 | 32549 |
| 12/14/2020 | 30658 | 403 | 3231 |
| 12/15/2020 | 30123 | 413 | 31231 |
| 12/16/2020 | 29871 | 379 | 32498 |
| 12/17/2020 | 29148 | 358 | 33501 |
| 12/18/2020 | 28657 | 381 | 31845 |
| 12/19/2020 | 27819 | 401 | 30145 |

Here beta () = 0.15858943 and gamma () = 0.21765473 and alpha () = 0.02195618

5.2.3 System with social distancing:

Social distancing includes avoiding large gatherings, physical contact, and other efforts to mitigate the spread of infectious disease. According to our model, the term this is going to impact is our contact rate, β.

Let’s introduce a new value, ρ, to capture our social distancing effect. This is going to be a constant term between 0–1, where 0 indicates everyone is locked down and quarantined while 1 is equivalent to our base case above. To introduce this into our model, we’ll modify Equations of SIR-D and SIR-F above by multiplying this with our β.

If we set ρ to 1, 0.8, and 0.5, we can visualize the flattening effect as we increase our efforts to contain the disease through simple, every day actions.

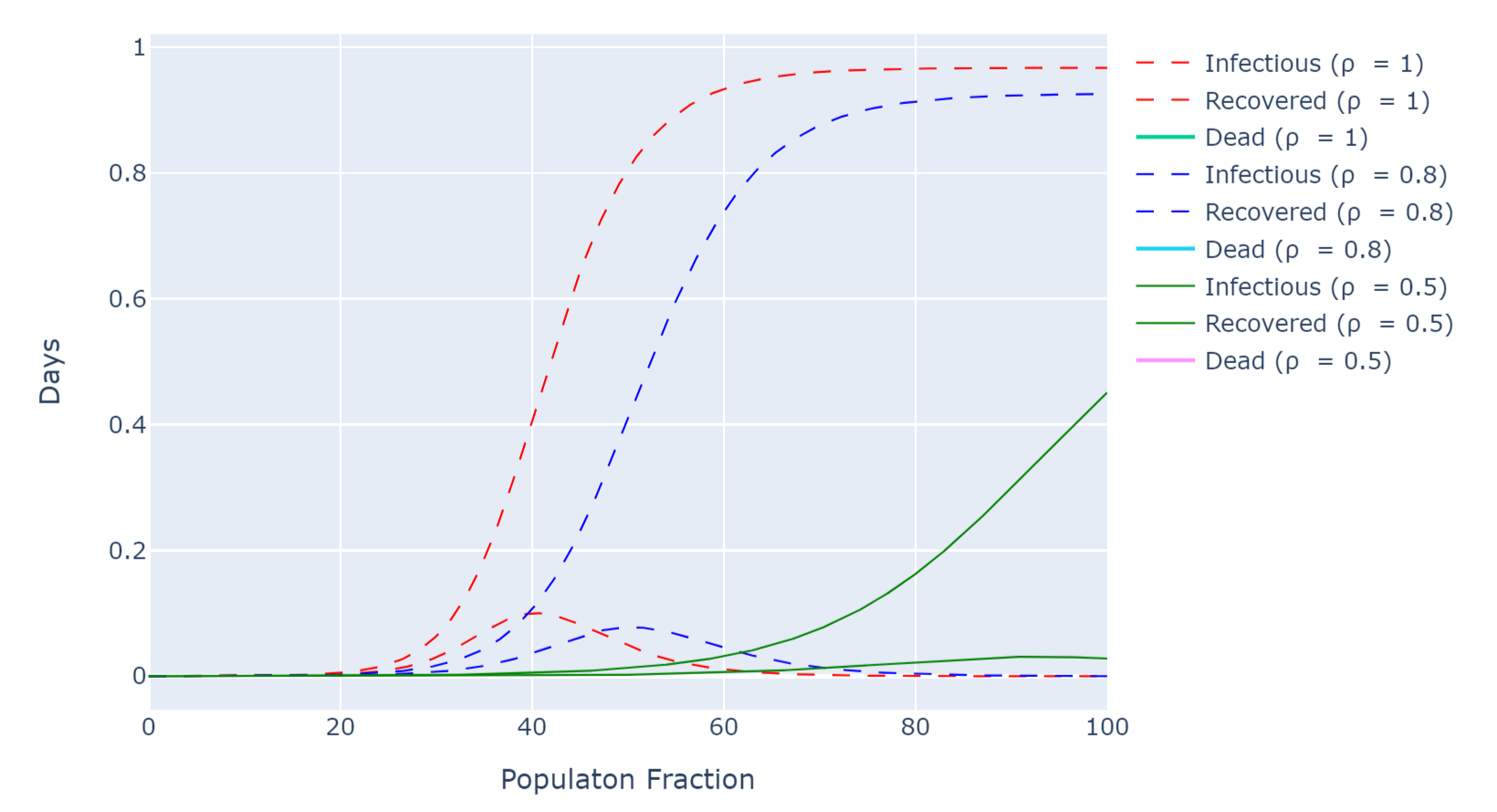


Figure 5.8 Effect of rho (

From above graph we can clearly see that if rho () is 1 then approximately whole population tend to getting infected but still more than 90% people will recovered reason behind it is, most of population will have Herd Immunity and COVID-19 will not effect. But still number of deaths will increase for getting Herd Immunity.

# 6 Conclusion and Future Work

This paper presented outbreak visualization and trend of COVID-19 from 22nd Jan to Oct-Nov 2020 due to high impact on people and all government and organization. Mathematical model based on Regression (Linear, polynomial and exponential) and Support Vector Machine (SVM) were proposed to predict for world, and polynomial regression performed better than all other.

Mathematical model SIR, SIR-D and SIR-F were proposed to predict the India data where SIR model showed a significant difference between both model and SIR-D with (beta () = 0.15858943 and gamma () = 0.21765473 and alpha () = 0.02195618) performed better than SIR with (beta () = 0.21774528 and gamma () = 0.17736864). Optimization of parameter of both SIR and SIR-D models improved the prediction accuracy significantly.

# 6 Bibliography

1. WHO Coronavirus Disease (COVID-19) Dashboard | WHO Coronavirus Disease (COVID-19) Dashboard, https://covid19.who.int/, last accessed 2020/07/24.
2. MoHFW | Government of India Home, https://www.mohfw.gov.in/, last accessed 2020/11/15.
3. World Health Organization (WHO), "Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCov)," WHO, 2020
4. World Health Organization, "Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19)," World Health Organization, 2020.
5. World Health Organization, "Coronavirus disease 2019 (COVID-19) Situation Report- 13," World Health Organization, 2020.
6. Corona Tracker Community, "Corona Tracker," Corona Tracker, 2020. [Online]. Available: https://www.coronatracker.com/. [Accessed 28 September 2020].
7. D. Toppenberg-Pejcic, J. Noyes, T. Allen, N. Alexander, M. Vanderford and G. Gamhewage, "Emergency Risk Communication: Lessons Learned from a Rapid Review of Recent Gray Literature on Ebola, Zika, and Yellow Fever," Health Communication, vol. 34, no. 4, pp. 437- 455, 2018..
8. F. Samreen, Y. Elkhatib, M. Rowe, and G. S. Blair, “Daleel: Simplifying cloud instance selection using machine learning,” in *NOMS 2016-2016 IEEE/IFIP Network Operations and Management Symposium*, pp. 557–563, IEEE, 2016.
9. N. J. Yadwadkar, B. Hariharan, J. E. Gonzalez, B. Smith, and R. H. Katz, “Selecting the best vm across multiple public clouds: A data-driven performance modeling approach,” in *Proceedings of the 2017 Symposium on Cloud Computing*, pp. 452–465, ACM, 2017.
10. O. Alipourfard, H. H. Liu, J. Chen, S. Venkataraman, M. Yu, and M. Zhang, “Cherrypick: Adaptively unearthing the best cloud configurations for big data analytics,” in *14th* {*USENIX*} *Symposium on Networked Systems Design and Implementation (*{*NSDI*} *17)*, pp. 469–482, 2017.
11. M. Ghobaei-Arani, S. Jabbehdari, and M. A. Pourmina, “An autonomic resource provisioning approach for service-based cloud applications: A hybrid approach,” *Future Generation Computer Systems*, vol. 78, pp. 191–210, 2018.
12. A. Rachah and D. F. M. Torres, "Analysis, simulation and optimal control of a SEIR model for Ebola virus with demographic effects," Common. Fac. Sac. Univ. Ank. Series A1, vol. 67, no. 1, pp. 179-197, 2018.
13. A. T. Porter, "A path-specific approach to SEIR modeling," Ph. D Thesis University of Iowa, 2012
14. John Hopkins University, "Coronavirus Map," John Hopkins University, 17 March 2020. [Online]. Available: https://coronavirus.jhu.edu/map.html. [Accessed 17 March 2020].
15. Q. Li, Q. Hao, L. Xiao, and Z. Li, “Adaptive management of virtualized resources in cloud computing using feedback control,” in *2009 First International Conference on Information Science and Engineering*, pp. 99–102, IEEE, 2009.
16. The Institute for Disease Modelling, "SEIR and SEIRs models," Institute for Disease Modelling, 2019. [Online]. Available: https://institutefordiseasemodeling.github.io/Documentation/general/model-seir.html. [Accessed 03 Oct 2020].
17. Kim, Y. Jeong, Y. Kim, K. Kang and M. Song, "Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news," Journal of Information Science, vol. 42, no. 6, pp. 763-781, 2016.
18. P. Xiong, Z. Wang, S. Malkowski, Q. Wang, D. Jayasinghe, and C. Pu, “Economical and robust provisioning of n-tier cloud workloads: A multi-level control approach,” in *2011 31st International Conference on Distributed Computing Systems*, pp. 571–580, IEEE, 2011.
19. H. C. Lim, S. Babu, J. S. Chase, and S. S. Parekh, “Automated control in cloud computing: challenges and opportunities,” in *Proceedings of the 1st workshop on Automated control for datacenters and clouds*, pp. 13–18, 2009.
20. John Hopkins University, "Coronavirus Map," John Hopkins University, 02 Dec. 2020. [Online]. Available: https://coronavirus.jhu.edu/map.html. [Accessed 17 March 2020].
21. ICMR (Indian Council of Medical Research, <https://www.icmr.gov.in/> for Data related to India as state wise all data, age-group wise etc. On
22. .Time series data from here which is collected from John Hopkins through web scrapping, “<https://www.kaggle.com/sudalairajkumar>” on daily basis. Data accessed on Sep 2020 and last modified on 2 Dec 2020.
23. S. K. Khor, "The Politics of the Coronavirus Outbreak," Think Global Health, 24 January 2020. [Online]. Available: https://www.thinkglobalhealth.org/article/politics-coronavirusoutbreak. [Accessed 04 Oct 2020].
24. S. K. Khor, "Malaysia does not have a good record of transparency," The Star, 15 January 2020. [Online]. Available: https://www.thestar.com.my/opinion/columnists/vitalsigns/2020/01/15/malaysia-does- not-have-a-good-record-of-transparency. [Accessed 04 Oct 2020]