|  |  |
| --- | --- |
| A picture of a winding road and trees  The Application of Predictive Analytics on COVID 19 Cases and Deaths USA and India | Ragunath Gunasekaran |

## 

Contents

[Introduction 2](#_Toc73800701)

[Problem Statement 3](#_Toc73800702)

[Research Questions 3](#_Toc73800703)

[Document Overview 4](#_Toc73800704)

[Scope 4](#_Toc73800705)

[Technical Approach 5](#_Toc73800706)

[Analysis 5](#_Toc73800707)

[Higher Level Data model for the files 6](#_Toc73800708)

[Summary Report : Confirmed Cases, Death Count at Date Level 7](#_Toc73800709)

[Interactive Dashboard : Summary Confirmed and Death counts at State and Date Level 8](#_Toc73800710)

[Summary Confirmed and Death counts at Date Level. 8](#_Toc73800711)

[*Indian Top 10 Covid States* 9](#_Toc73800712)

[*US Death Rate Analysis* 9](#_Toc73800713)

[Model Development 10](#_Toc73800714)

[Summary of the ARIMA Model 11](#_Toc73800715)

[SARIMAX model for Deaths 12](#_Toc73800716)

[Final Prediction of Confirmed Cases - ARIMA Model - Time Series Forecasting Confirmed Cases ( Next 9 Months ) 13](#_Toc73800717)

[Final Prediction of Deaths - ARIMA Model - Time Series Forecasting Deaths  ( Next 9 Months ) 14](#_Toc73800718)

[India – one State Preidiciton for Confirmed Cases 15](#_Toc73800719)

[*Forecasting Death count for every from today to next 6 months* 15](#_Toc73800720)

[Testing and Evaluation 16](#_Toc73800721)

[HypothesisTest 16](#_Toc73800722)

[Conclusion 17](#_Toc73800723)

[Enhancements 18](#_Toc73800724)

[References: 19](#_Toc73800725)

## Introduction

As of April 19, CDC national forecasts predict 249,000–724,000 new COVID19 cases will likely be reported during the week ending May 15 and CDC national forecast predict 2,100–8,800 new COVID19 deaths will be reported during the week ending June 5th. That would bring the total coronavirus cases are 33 million and deaths are 580K in USA. This forecast and prediction approaches are based for my project to explore COVID19 Cases and Deaths.

COVID-19 was identified in Wuhan, China in December 2019. COVID-19 is caused by the virus severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a new virus in humans causing respiratory illness which can be spread from person-to-person. On February 11, 2020, the World Health Organization [announced](https://twitter.com/DrTedros/status/1227297754499764230) an official name for the disease that is causing the 2019 novel coronavirus outbreak, first identified in Wuhan China.

The new name of this disease is coronavirus disease 2019, abbreviated as COVID-19. In COVID-19, “CO” stands for corona, “VI” for virus, and ”D” for disease. Formerly, this disease was referred to as “2019 novel coronavirus” or “2019-nCoV.”

COVID-19 is a recent new disease, caused by a coronavirus that has not previously been seen in humans. Because it is a new virus, scientists and this world are learning more each day. Although most people who have COVID-19 have mild symptoms, COVID-19 can also cause severe illness and even death. COVID19 spread and death due to coronavirus is increasing day by day.

Predictive analytics is mathematical and statistical analysis that uses data mining, machine learning algorithms based on historical data, behavior, trend to project the future of our business or outcome.

Python as a programming language has numerous uses such as Machine Learning, AI, mobile applications, statistical , etc. As we know, AI, machine learning, and data analysis/analytics is where it has amassed most of its popularity. Python contains more libraries and packages is continually being employed which helped to solve complex industry issues, data-driven strategies, risk analysis etc.

We need to use four main performance metrics used to evaluate the effectiveness of classification models:

* Accuracy: test’s ability to correctly predict both classes
* Precision: test’s ability to correctly detect positive classes from all predicted positive classes
* Recall (Sensitivity): test’s ability to correctly detect positive classes from all actual positive classes.
* F1 Score: harmonic mean of precision and recall

( Source – Ref 1 )

## Problem Statement

In this pandemic situation, we need to predict pandemic’s behavior within an acceptable degree of uncertainty by establishing when and under which conditions countries can expect increases, peaks, and reductions in new Cases and Deaths (mortality). The Prediction of new cases helps to demand for acute medical services and the prediction help the planning for required health technologies (PPE, ventilators, etc.) and ensure adequate end-to-end supply chain and distribution and managing human resources for an appropriate and timely response. The sad part is death prediction helps to plan for the cremation services and arrangements.

The prediction of the pandemic behavior and outcome alert the society towards prevention and fear to stop the spread.

In this Project, the main goal of the project is to build a predictive model to predict and forecast the number of cases and deaths over the period by using the time series and regression algorithms.

As part of COVID Analysis, we are going to find the solution which can predict and answer the below questions.

### Research Questions

1. Forecasting Confirmed Cases, Deaths & Recovered Cases and Computing Mean Absolute Error (MAE)
2. Predict the Case Fatality Rates and Infection Fatality Rates
3. Prediction for the possible end of new cases of COVID-19 epidemics
4. Support Vector Regression model for prediction of COVID19 cases
5. Evolution of Active Cases, Confirmed Cases, Death Cases and Recovered Cases
6. Calculate Recovery and Death Rates, Deaths per 100k
7. Count of patients : Infected by Virus and Deaths, Distribution

## Document Overview

For this project, I modelled various predictive process data related to COVID 19 datasets. The Datasets comes from WHO and Kaggle, which contains 13 columns for Global data and 400 columns for date related data.

1. WHO Dashboard: <https://covid19.who.int/table>
2. NY Times :  <https://aws.amazon.com/marketplace/pp/prodview-jmb464qw2yg74>
3. WHO Time Series CSV: <https://covid19.who.int/WHO-COVID-19-global-data.csv>
4. <https://www.cdc.gov/coronavirus/2019-ncov/faq.html#Basics>
5. <https://www.kaggle.com/sudalairajkumar/covid19-in-india>

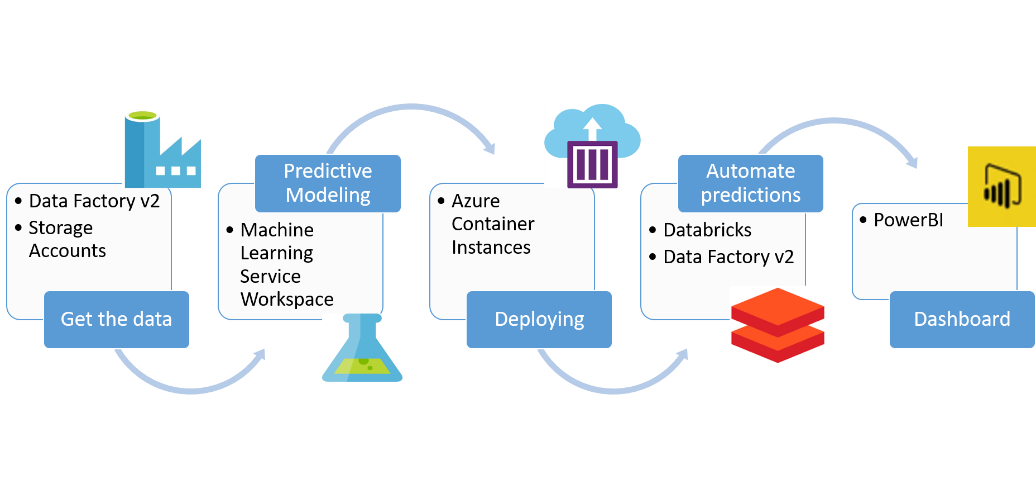
## Scope

The scope of this project will be to analyze the COVID19 data in WHO and Kaggle to identify and predict, forecast the cases and deaths. As part of the COVID 19 analysis, we need to predict cases and deaths by end of August and so on. Also need include vaccination analytics.

1. Forecasting Confirmed Cases, Deaths & Recovered Cases and Computing Mean Absolute Error (MAE)
2. Predict the Case Fatality Rates and Infection Fatality Rates
3. Prediction for the possible end of new cases of COVID-19 epidemics
4. Daily Confirmed, new Confirmed and Death cases Analysis by Country, State, County
5. Predict the Corona Cases and Death
6. State Level Counts of Corona virus, Comparison between States
7. Calculate Recovery and Death Rates, Deaths per 100k
8. Number of Corona Cases comparison : Positive vs Negative
9. Testing Count Details by Country, State, County
10. Count of patients : Infected by Virus and Deaths

## Technical Approach

As part of Technical Approach, I wanted to use Microsoft Azure ML and Power BI Technologies to automate end to end flow of Azure ML model deployment. The Predicted outcome, we need to use Power BI Dashboard to showcase the results which may help Business leaders to identify easily.



## Analysis

The Source Data comes from various sources, I performed the EDA ( exploratory Data Analysis) to find any outliers, missing values, handling n/a values, removing duplicates and patterns. Also, clean and preprocess the data before modelling. Since we are dealing with three different domains in the Predictive application study.

Datasets from NY Times and CDC USA Government website

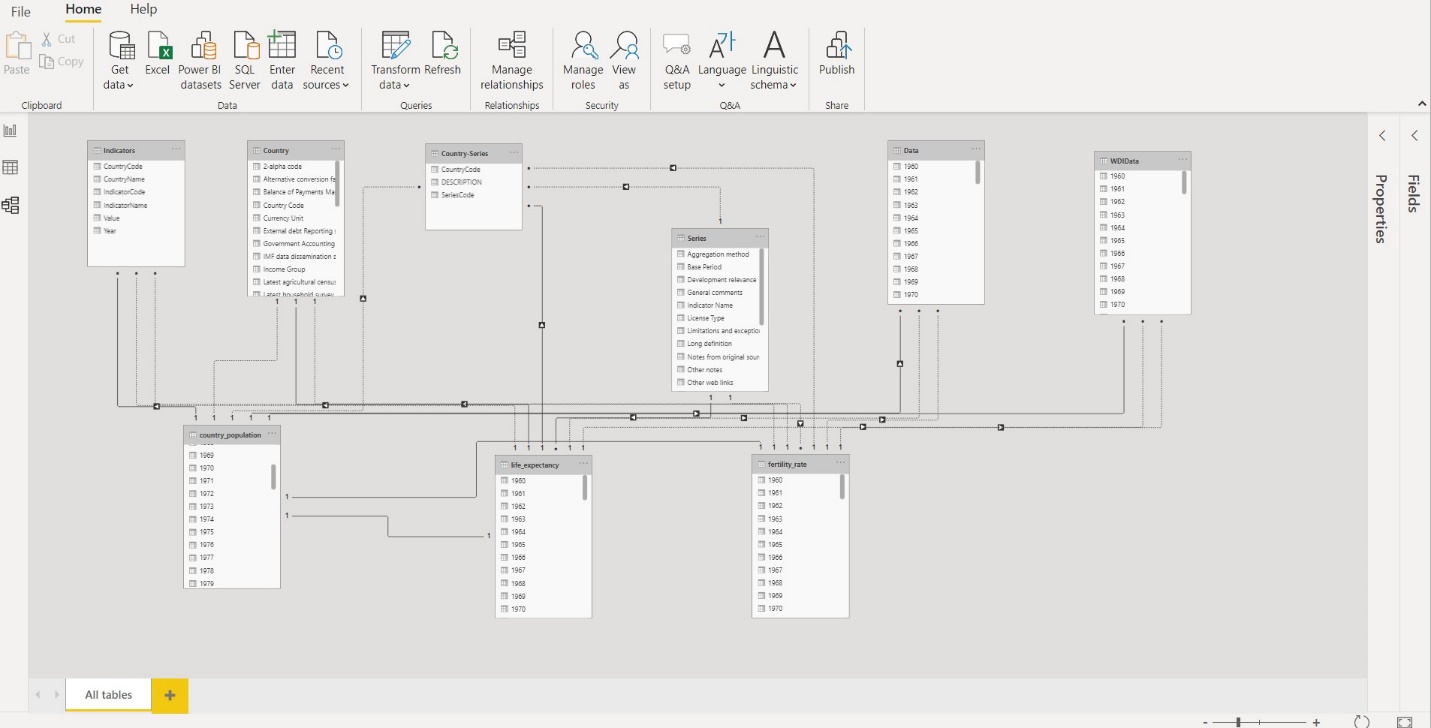
<https://aws.amazon.com/marketplace/pp/prodview-jmb464qw2yg74>

<https://www.cdc.gov/nchs/covid19/covid-19-mortality-data-files.htm>

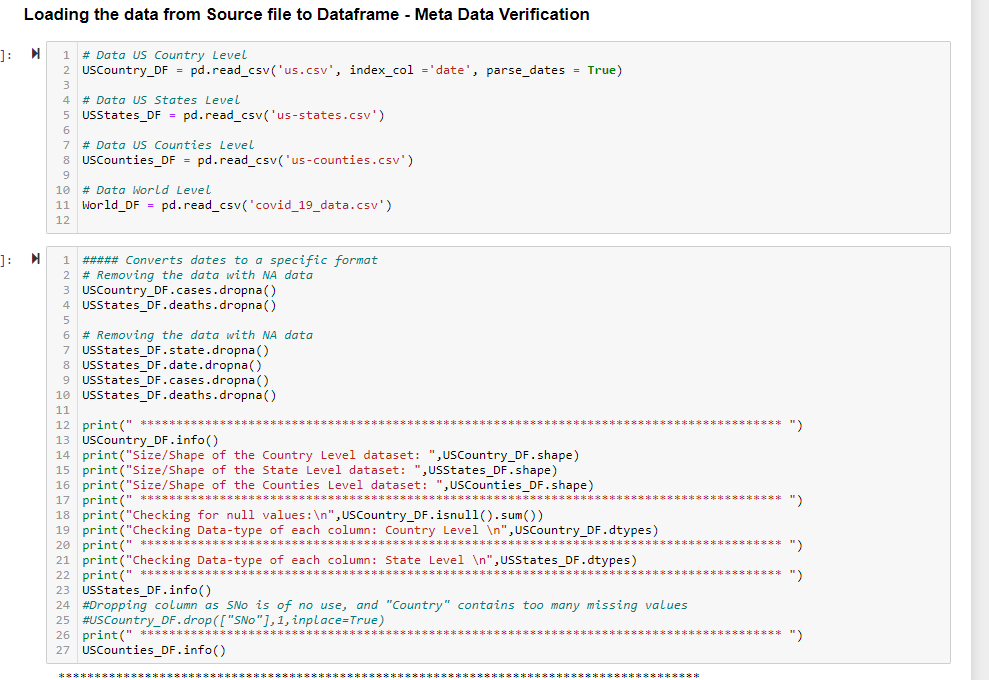
Datasets from Kaggle and Novel Corona Virus Disease 2019 in India

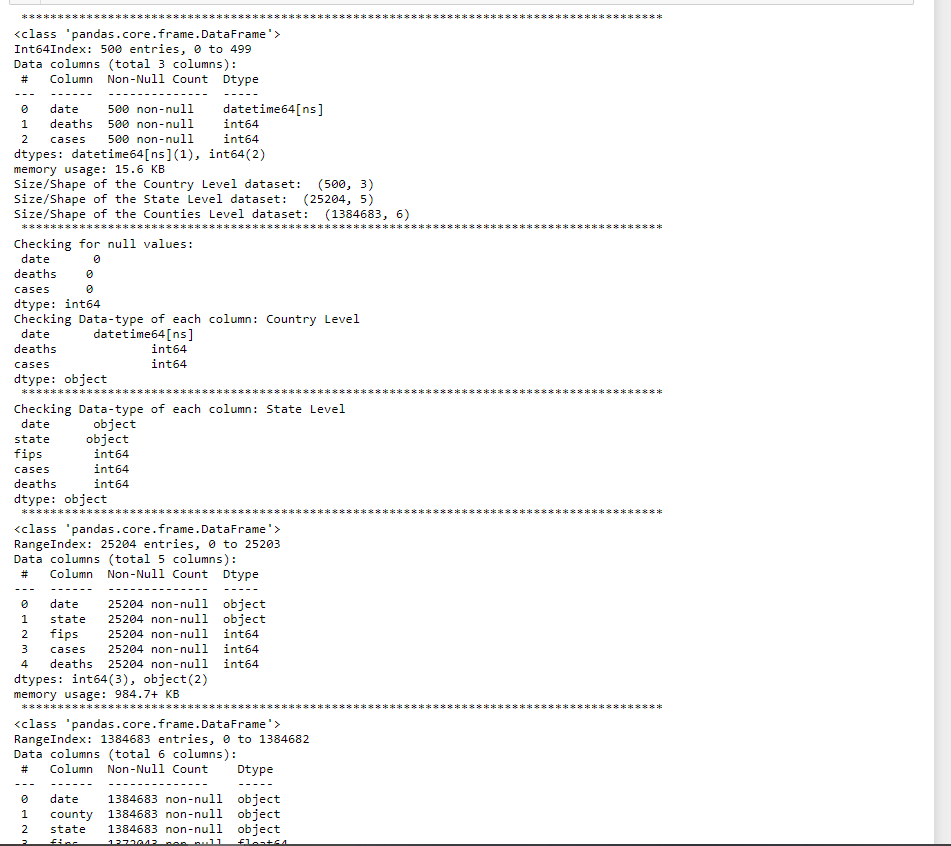
<https://www.kaggle.com/sudalairajkumar/covid19-in-india>

### Higher Level Data model for the files

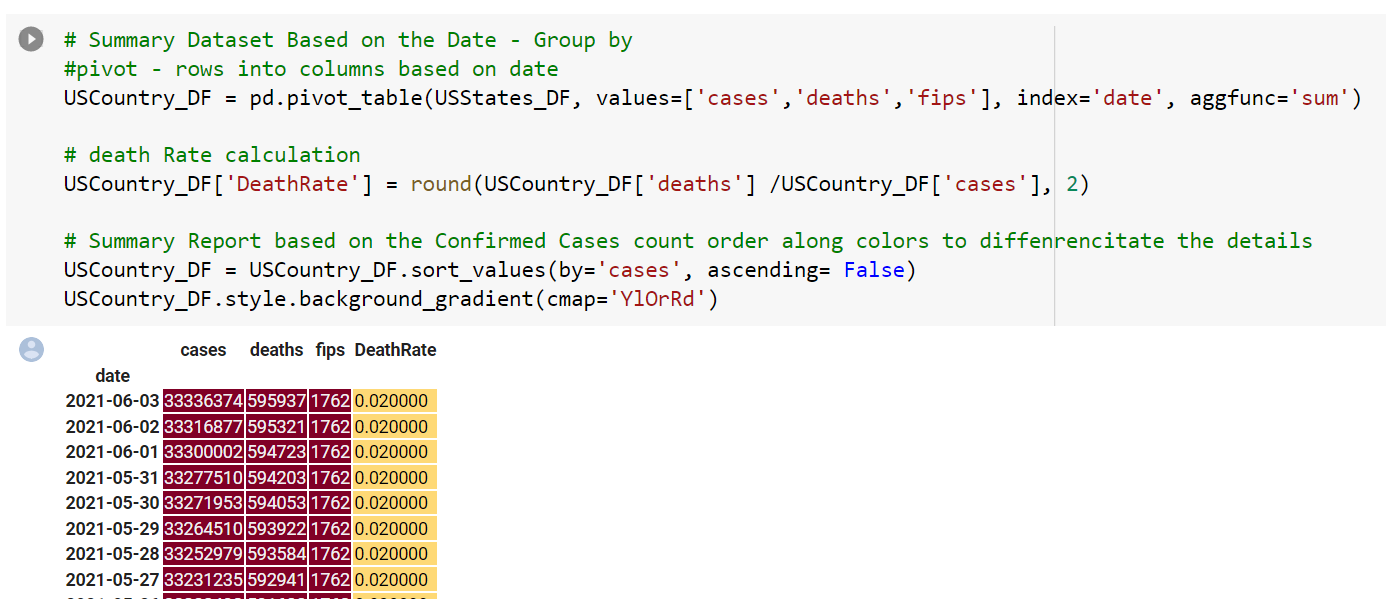


Based on the reading, “data scientists usually spend 70% of the project time here, preprocessing and exploring the data”. So, the Data Preparation is important for predictive analytics project. We need to make sure the data is suitable and well prepared for model. While preparing the Data, we need to verify the metadata of the given dataset and need to understand very clearly about data patterns, relationships etc.

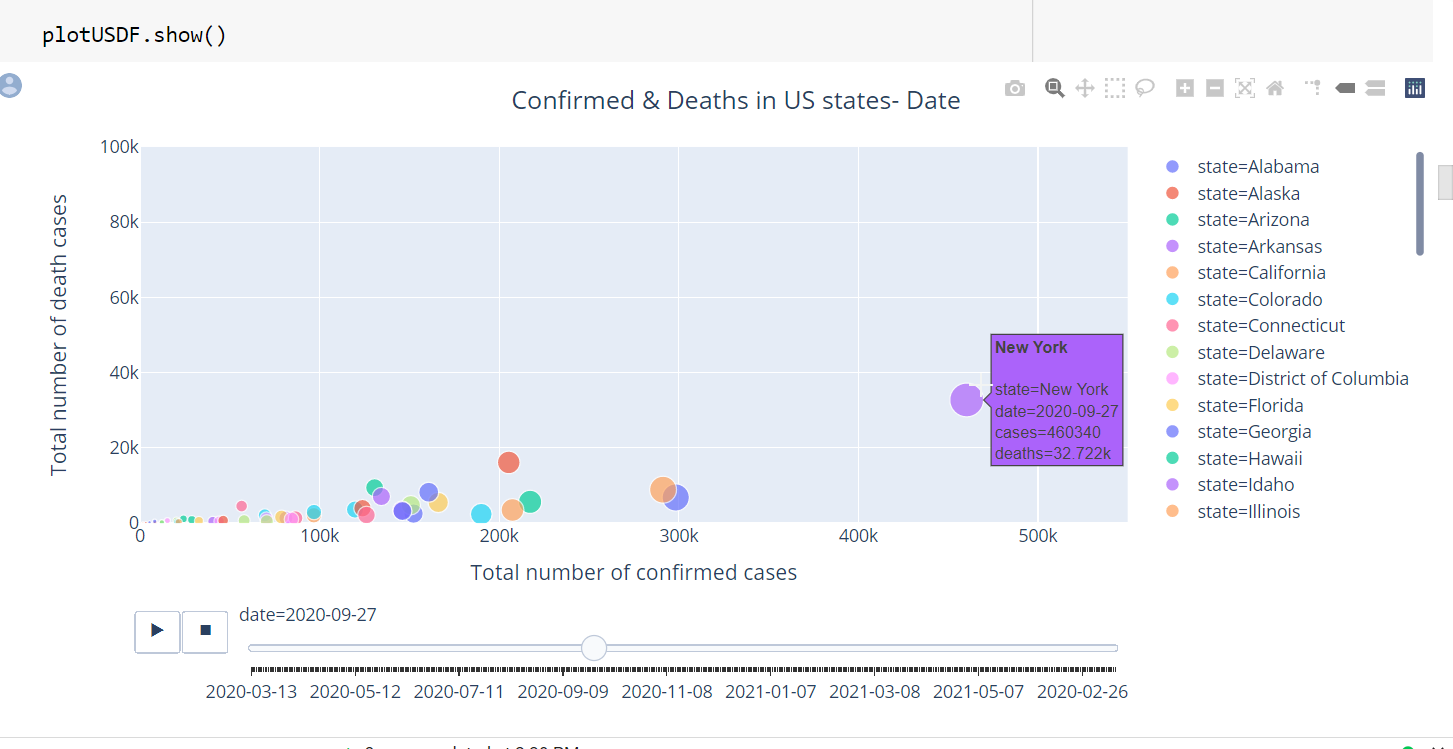
* Need to describe the dataset describe()
* Need to understand the number of columns and rows shape.
* Need to understand the summary of the data ( info ) 



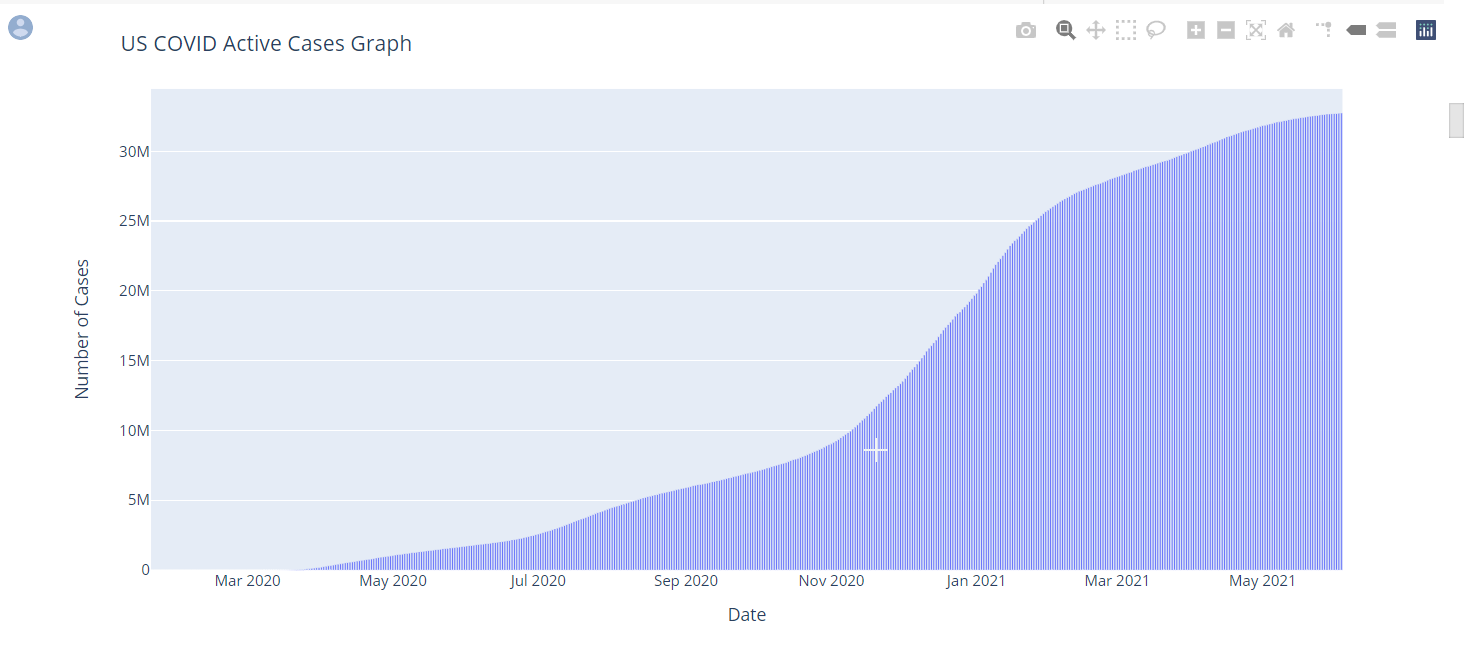
### Summary Report : Confirmed Cases, Death Count at Date Level



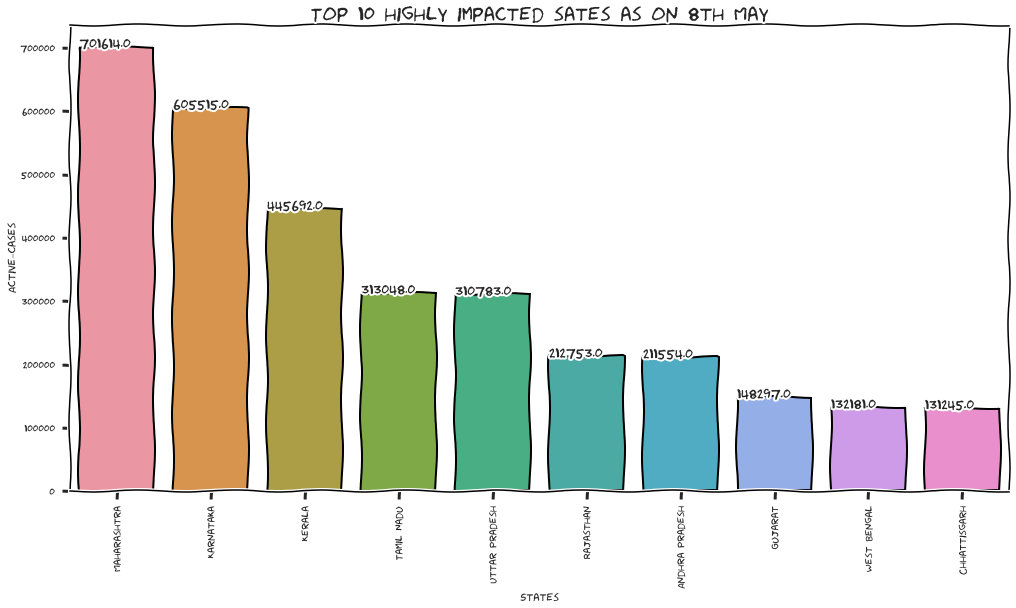
### Interactive Dashboard : Summary Confirmed and Death counts at State and Date Level



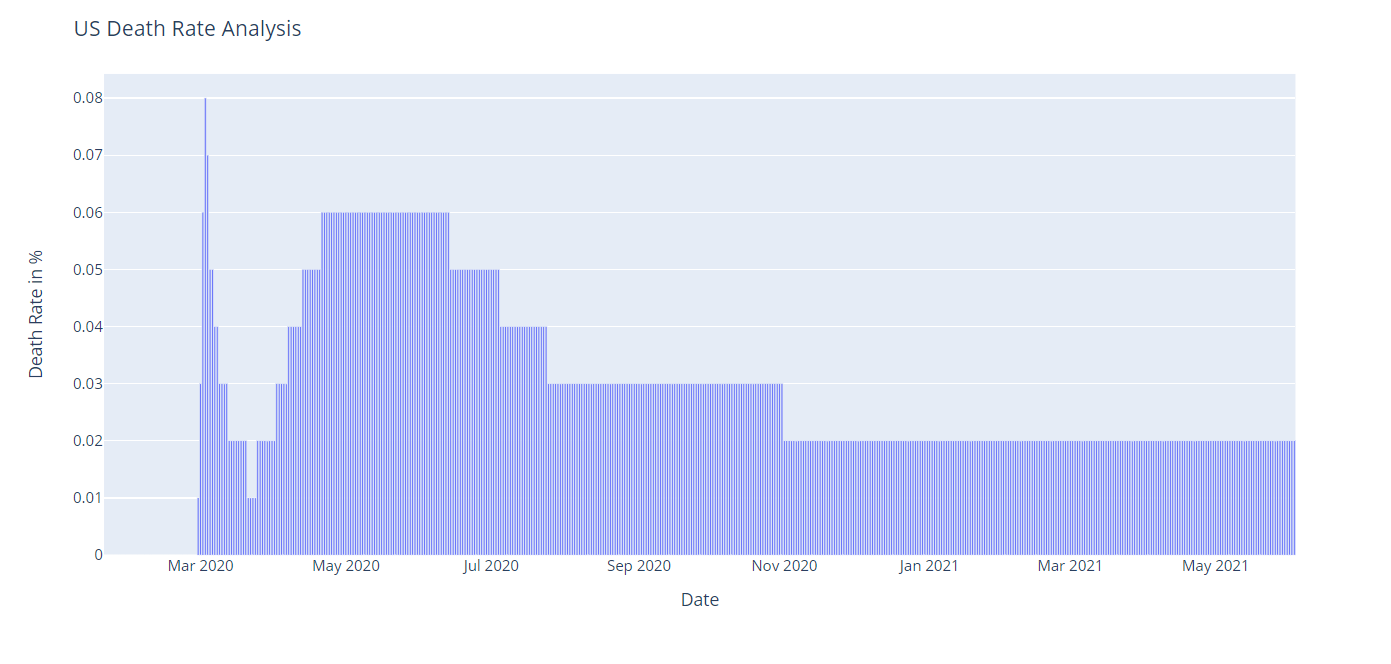
### Summary Confirmed and Death counts at Date Level.



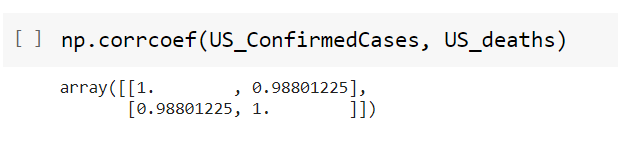
### *Indian Top 10 Covid States*



### *US Death Rate Analysis*



Here are the mean and standard deviation of Variables in the Dataset (12426345.542, 12108431.117040819)



## Model Development

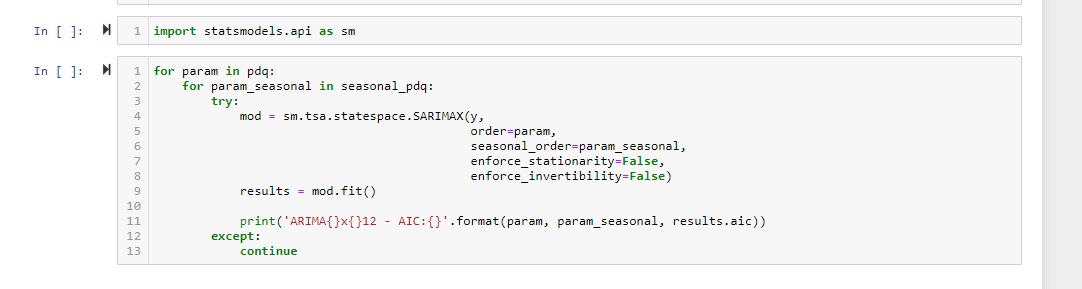
I used Time Series and ARIMA and SARIMAX models used to predict and forecast the number of deaths and confirmed cases in this project.

ARIMA stands for auto regressive integrated moving average.

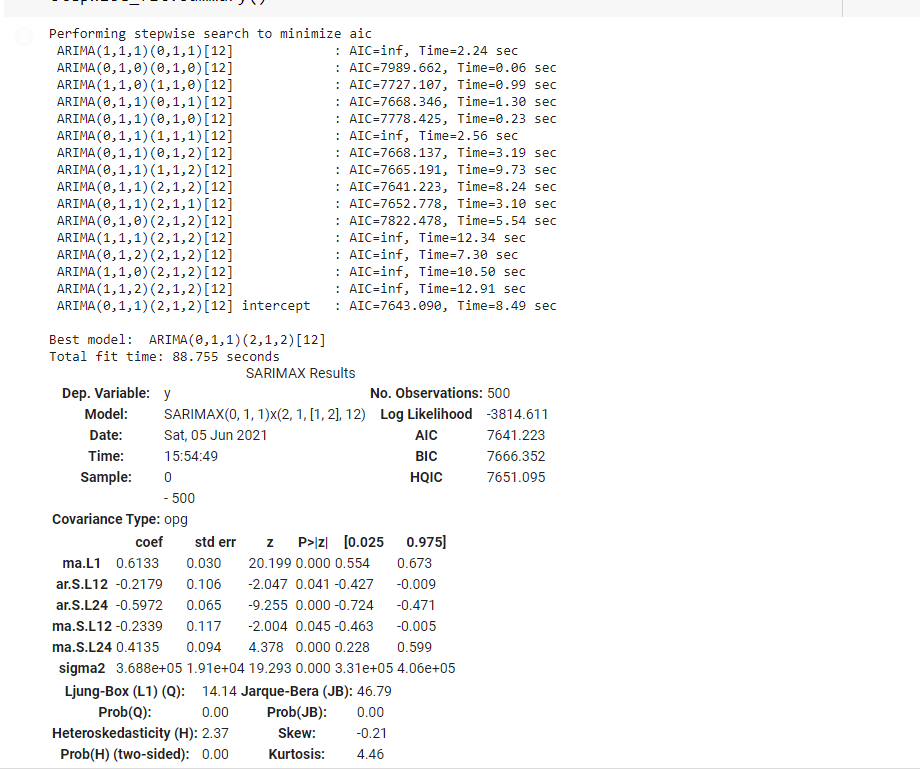
SARIMAX is similar and stands for seasonal auto regressive integrated moving average with exogenous factors.

As we know, auto regressive (AR) model the model predicts the future or next data point by looking at previous data points and using a mathematical formula like linear regression.

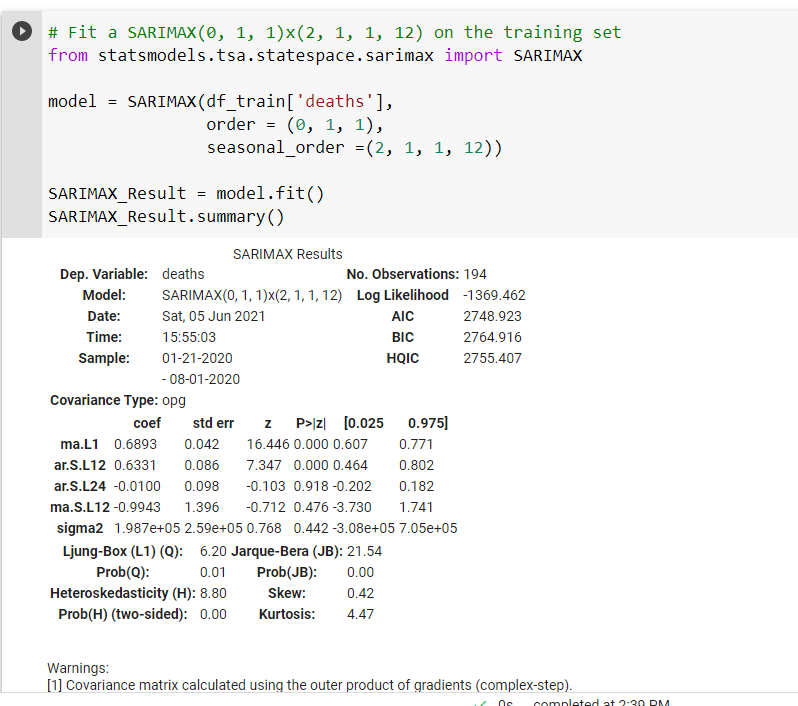
As part of the forecasting approach, I initiated the model by using Statsmodels.api library.



### Summary of the ARIMA Model

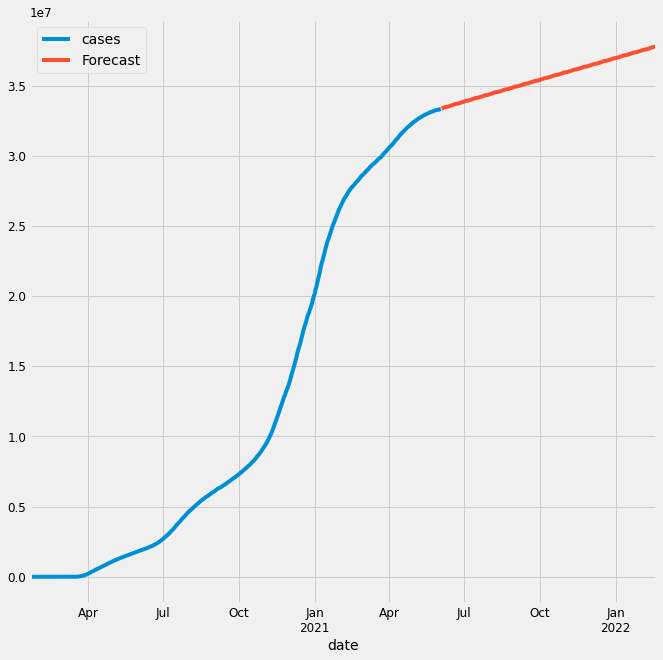


### SARIMAX model for Deaths



### Final Prediction of Confirmed Cases - ARIMA Model - Time Series Forecasting Confirmed Cases ( Next 9 Months )

As part of the below prediction shows that by next year January 2022, the Confirmed Cases count may reach to around 3.7 M Confirmed Case Counts.



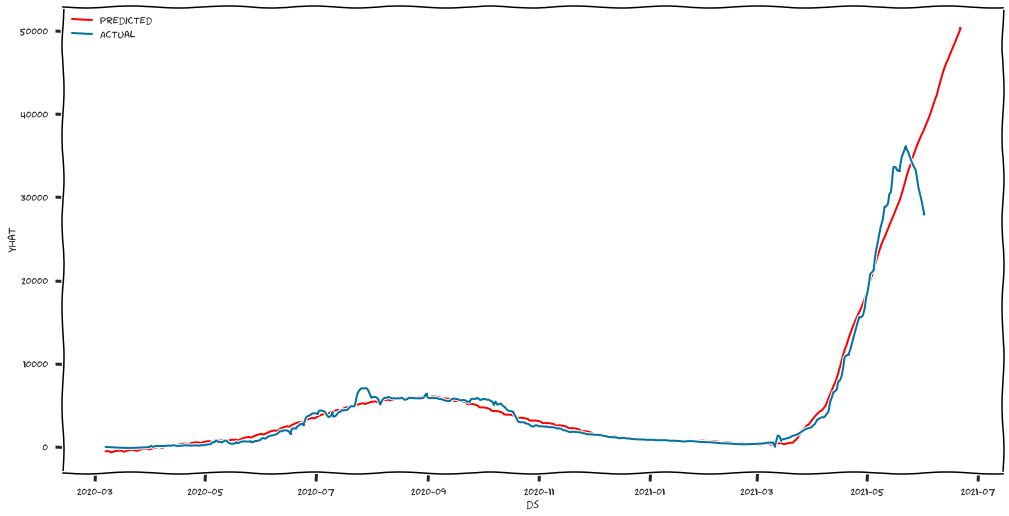
### Final Prediction of Deaths - ARIMA Model - Time Series Forecasting Deaths  ( Next 9 Months )

As part of the above prediction shows that by next year January 2022, the death count may reach to around 710 K deaths.

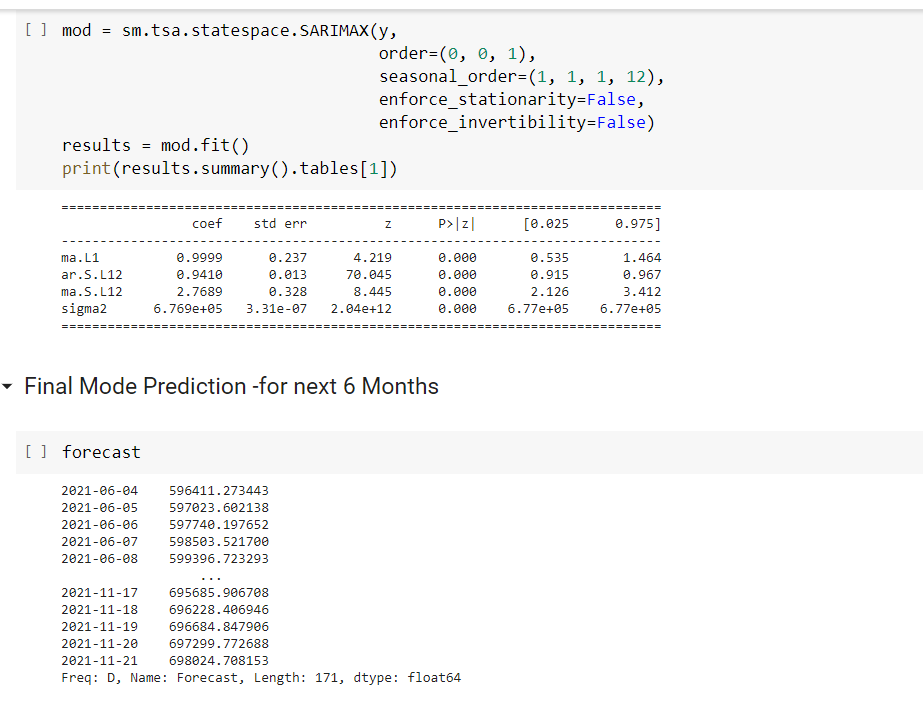
##### 

### India – one State Preidiciton for Confirmed Cases

As part of the below prediction shows that by next year January 2022, the Confirmed Cases count may reach to around 50 K Confirmed cases on daily basis.

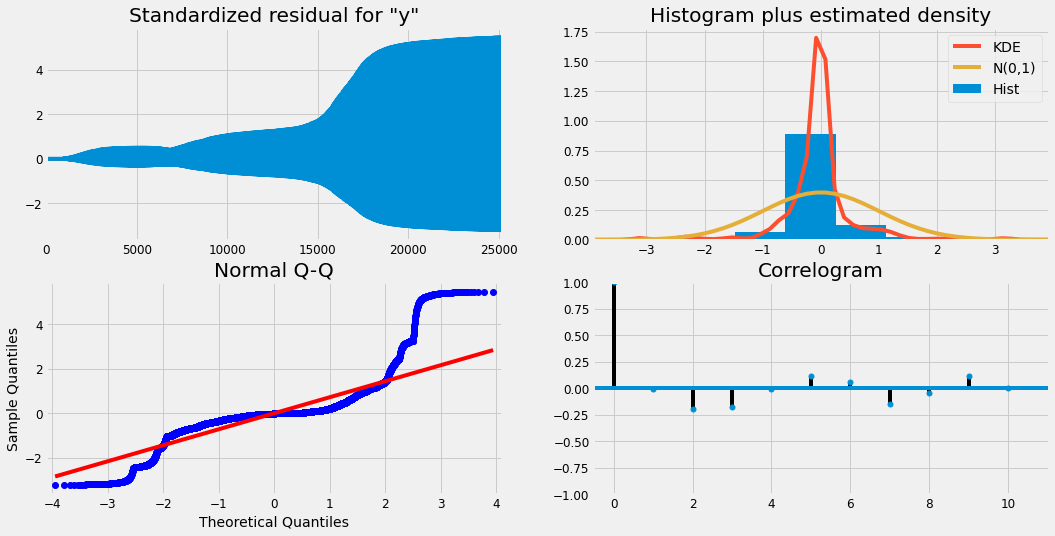


#### Forecasting Death count for every from today to next 6 months.

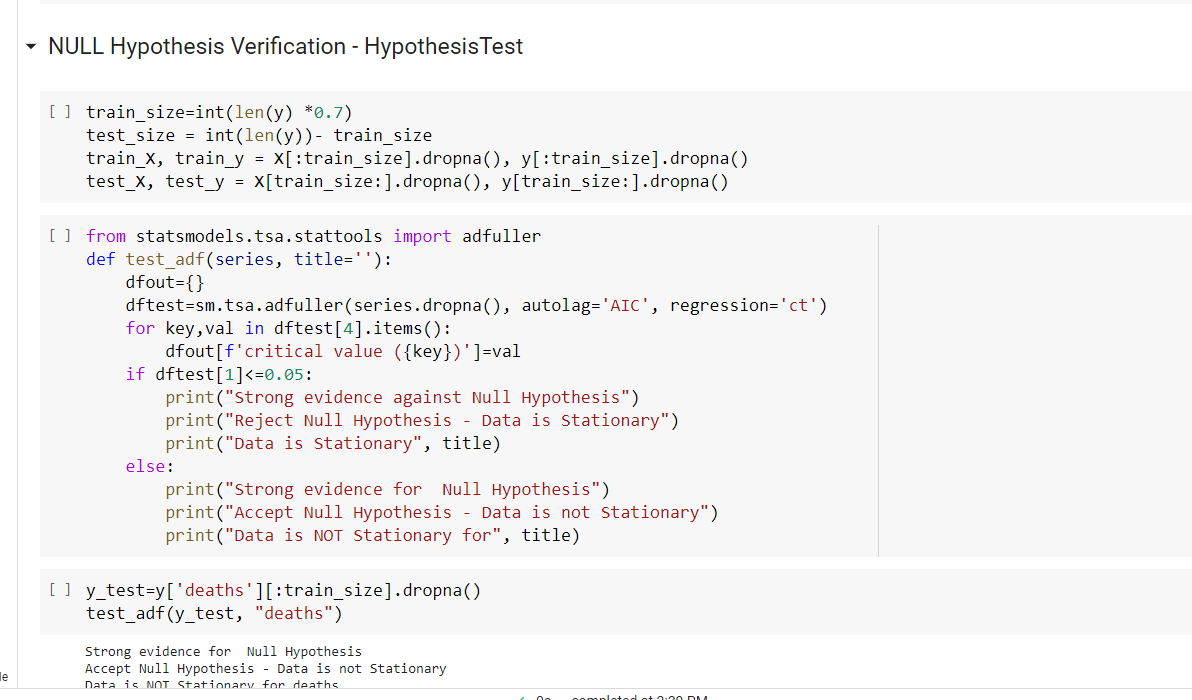


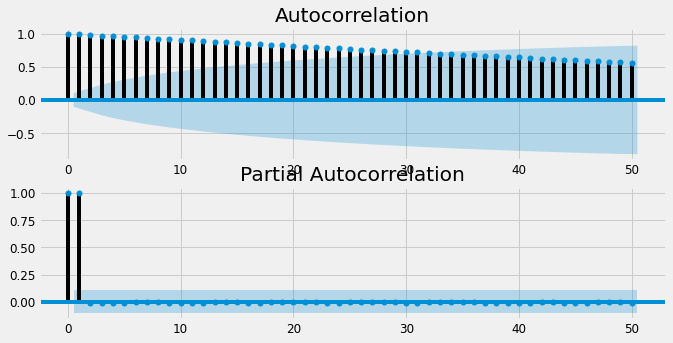
## Testing and Evaluation

I tested and evaluated the models by using training and Testing datasets ( Considering data across and Random data for Training ). By running those models, we can evaluate and identify the anomalies. For model and probability evaluation, we will use various tests and functions.



### HypothesisTest





## Conclusion

As part of this project, I have analyzed various techniques to perform the EDA of COVID19 Trends and Outbreak Prediction of Spread in USA and India.

The below are the outcomes of my Analytics.

1. Calculated DeathRate Ratio - From Feb 29,2020 to June 5,2021, overall Death Count is 598 K. Initially Death Ratio was increased, and it started gradually decreasing from July,2020.

2. Number of Death : Number of deaths is increasing day by day ( as of June 5 )

3. Confirmed Cases : Number of positive Count is increasing day by day ( as of June 5 ) - 33.61 M

4. State Level Cases : Created Animation plot for State Level counts on daily basis. ( Both Confirmed and Death count )

   observed NY State count had highest counts.

5. Based on the Data as of June 5,2021, the prediction of Death count on January 31,2022 is 710K

6. Based on the Data as of June 5,2021, the prediction of Confirmed Cases count on January 31,2022 is 33 million. Based on the Data as of June 5,2021, the prediction of Confirmed Cases in India counts on January next month going to be 50 K

7. Verified various Analysis through various Regression, Charts and Probability function to verify the accuracy

8. Compared the model with prediction and actual values to make sure the steps we followed

9. Performed Various Tred Analysis and Hypothesis Test, P-Value findings

10. Created Animation for State Level count with Date Level

### Enhancements

As of June 5,2021, I am hearing that 41% of US adult population fully vaccinated and I hope this will help to stop the COVID Spread, and deaths and our Prediction may go down once vaccination increases.

We may need to include vaccine analytics into this Project. In that way, Our Analytics will capture all the details – Confirmed Cases, Recovery, Deaths and Vaccinated details.

As part of this project, I learnt the predict and forecasting future in deep.

My sincere Thanks to Professor Dr.Brett Werner for all his guidance and support on this semester which helped me to perfume this detailed and predictive analysis of COVID Spread in USA and India.

## References:

1. We’re Sharing Coronavirus Case Data for Every U.S. County by NY Times

https://www.nytimes.com/article/coronavirus-county-data-us.html

2. Coronavirus Disease 2019 (COVID-19)

https://covid.cdc.gov/covid-data-tracker/?CDC\_AA\_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fcases- updates%2Fcases-in-us.html#cases\_casesinlast7days

3. Analyze NY Times Covid-19 Dataset, Medium

https://towardsdatascience.com/analyze-ny-times-covid-19-dataset-86c802164210

4. HOW TO USE DATA ANALYSIS FOR MACHINE LEARNING by Sharp Sight

https://www.sharpsightlabs.com/blog/data-analysis-machine-learning-example-1/

5. Python | ARIMA Model for Time Series Forecasting by geeksforgeeks https://www.geeksforgeeks.org/python-arima-model-for-time-series-forecasting/

6. Modeling COVID-19 epidemic with Python Medium

https://towardsdatascience.com/modeling-covid-19-epidemic-with-python-bed21b8f6baf

https://www.kaggle.com/dhrupal1289/covid-analysis-prediction

<https://www.kaggle.com/khoongweihao/covid-19-novel-coronavirus-eda-forecasting-cases#Forecasting-Total-Number-of-Cases-Worldwide>