# Capstone Project –

# Analyzing the trends of COVID-19 with Python

# -by Yoganshu Raikhere

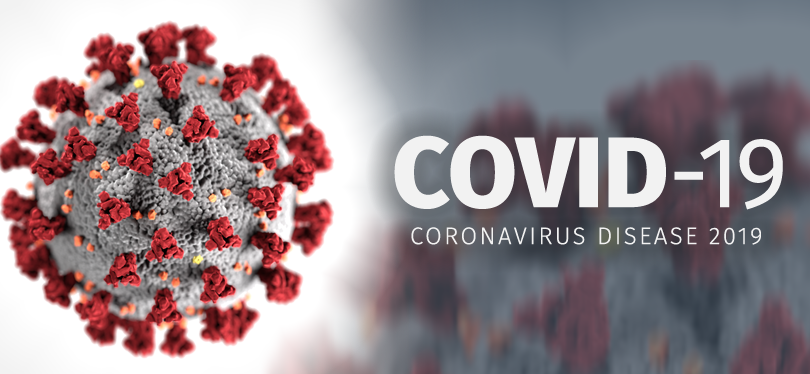


Table of Contents

1. Problem Statement
2. Project Objective
3. Data Description
4. Data Pre-processing Steps and Inspiration
5. Choosing the Algorithm for the Project
6. Motivation and Reasons for Choosing the Algorithm
7. Assumptions
8. Model Evaluation and Techniques
9. Inferences from the Same
10. Future Possibilities of the Project
11. Conclusion
12. References

# Problem Statement

In light of the ongoing COVID-19 pandemic, accurately forecasting the number of new cases is critical for public health planning and response. This project aims to leverage the Prophet library to analyze and predict the number of COVID-19 cases for the next 7 days. By providing short-term forecasts, the project seeks to aid health authorities in resource allocation, readiness planning, and implementing timely interventions to mitigate the spread of the virus.

The ultimate goal is to create a reliable forecasting tool that can help authorities and stakeholders make informed decisions to protect public health and manage the impact of the pandemic effectively.

# Project Objective

The primary objective of this COVID-19 project is to analyse the spread, impact, and management of the COVID-19 pandemic using data-driven approaches. By leveraging comprehensive datasets, the project aims to achieve the following specific goals:

1. **Track and Analyse Spread**:
   * Monitor the daily and cumulative number of COVID-19 cases, recoveries, and fatalities across different regions.
   * Identify hotspots and analyse trends in the virus's spread to understand transmission dynamics and peak periods.
2. **Impact Assessment:**
   * Evaluate the impact of COVID-19 on public health, healthcare systems, and the economy.
   * Assess the effectiveness of different interventions, such as lockdowns, social distancing measures, and vaccination campaigns, in controlling the spread of the virus.
3. **Predictive Modelling:**
   * Develop predictive models to forecast the future trajectory of COVID-19 cases, recoveries, and fatalities.
   * Use these models to inform public health strategies and preparedness plans.

# Data Description

The dataset available is **covid\_19\_clean\_complete.csv**

Data description, various insights from the data.

The **covid\_19\_clean\_complete.csv** contains 49068 rows and 10 columns.

1. Province/State – State
2. Country/Region – Country or Region values
3. Lat – Latitude of country
4. Long – Longitude of country
5. Date – Date of Survey
6. Confirmed – Number of cases confirmed on that date.
7. Deaths - Number of death cases on that date.
8. Recovered - Number of Recovered cases on that date.
9. Active - Number of Active cases on that date.
10. WHO Region – Who Region representing that country.

# Data Preprocessing Steps and Inspiration

The preprocessing of the data included the following steps:

1. **Checked for Null Values:** Identify any missing values in the dataset to ensure data completeness.
2. **Checked for Duplicates:** Detect and remove any duplicate entries to maintain data integrity.
3. **Checked for Data Types:** Verify the data types of each column to ensure they are appropriate for analysis.
4. **Converted Date Column Data Type:** Changed the data type of the Date column from an object to an Datetime format for accurate analysis.

These steps are crucial for preparing the dataset for further analysis and modelling.

The main inspiration for this study is to leverage the power of data analytics to enhance best movie recommendation to the users

# Choosing the Algorithm for the Project

For forecasting the next 7 days of COVID-19 cases, I have selected the Prophet library. Here are the reasons for this choice:

1. **Flexibility and Ease of Use**:
   * Prophet is designed to handle missing data, outliers, and strong seasonal effects, making it well-suited for complex time series data like COVID-19 case counts.
   * Its user-friendly interface allows for quick setup and tuning, enabling efficient model building and forecasting.
2. **Robust Against Non-Stationary Data:**
   * COVID-19 case data can exhibit non-stationary behavior due to the evolving nature of the pandemic. Prophet is robust against such data, automatically applying necessary transformations to make the series suitable for forecasting.
3. **Adjustable Trend Components**:
   * Prophet provides flexibility in modeling different types of trends, such as linear or logistic growth, which can be adjusted based on the characteristics of the COVID-19 case data.

These features make Prophet an ideal choice for forecasting COVID-19 cases, providing accurate and interpretable results while handling the unique challenges of pandemic data.

# Assumptions

The following assumptions were made in order to create the model for Covid 19 project.

1. **Quality of Data**:
   * The dataset contains accurate, reliable, and up-to-date information on COVID-19 cases, recoveries, and fatalities.
   * Any errors, inconsistencies, or missing values in the data have been properly handled during the pre-processing stage.
2. **Completeness**:
   * The dataset covers a sufficient time period to capture trends and seasonal patterns, allowing for meaningful forecasting.
   * The data includes key variables such as dates, number of new cases, and any significant events or interventions (e.g., lockdowns, vaccination drives).
3. **Stationarity and Seasonality**:
   * While the raw data may be non-stationary, the Prophet model can handle non-stationary data by incorporating trend and seasonality components.
   * The dataset exhibits seasonal patterns that can be captured by the model (e.g., weekly reporting cycles, holiday effects).

# Model Evaluation and Technique

The following techniques and steps were involved in the evaluation of the Prophet model

**1. Creating Dataset for confirmed cases**

**Purpose**: To build the prophet model for forecasting next 7 days confirmed cases.

**2. Visualization**

* **Purpose**: To provide insights into model performance and forecast reliability.
* **Method**: Use visualizations such as:
  + **Data Plots: Plot cases like confirmed, active, deaths and recovered using plotly.**
  + **Forecast Plots**: Plot the actual vs. predicted values along with prediction intervals to visually assess the model's accuracy and uncertainty.
  + **Residual Plots**: Plot residuals to check for any patterns that might indicate model misfit.

# Inferences from the Project

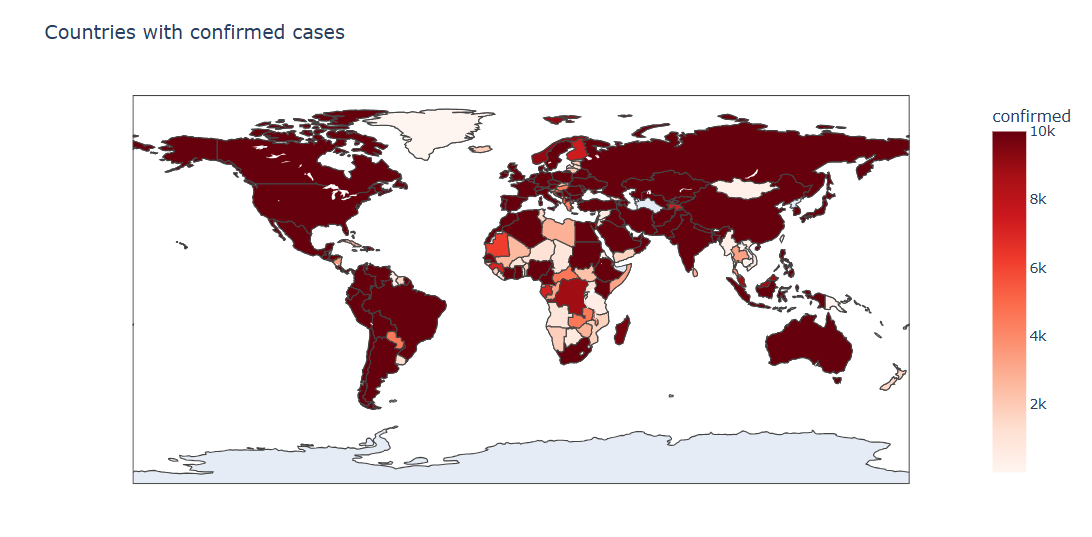
### **Trends and Patterns**

* **Consistent Trends: The time series analysis revealed consistent trends in the number of COVID-19 cases, including periods of increase and decrease. These trends were influenced by factors such as government interventions, public health measures, and population behaviour.**
* **Seasonal Effects: The model captured seasonal patterns, such as weekly cycles in reported cases, which could be attributed to variations in testing rates and reporting practices over different days of the week.**

### **Prediction Accuracy**

* **Reliable Short-Term Forecasts: The Prophet model provided accurate short-term forecasts for the next 7 days, enabling health authorities to plan and allocate resources effectively.**
* **Uncertainty Quantification: The prediction intervals generated by the model helped quantify the uncertainty associated with forecasts, allowing for better risk management and contingency planning.**

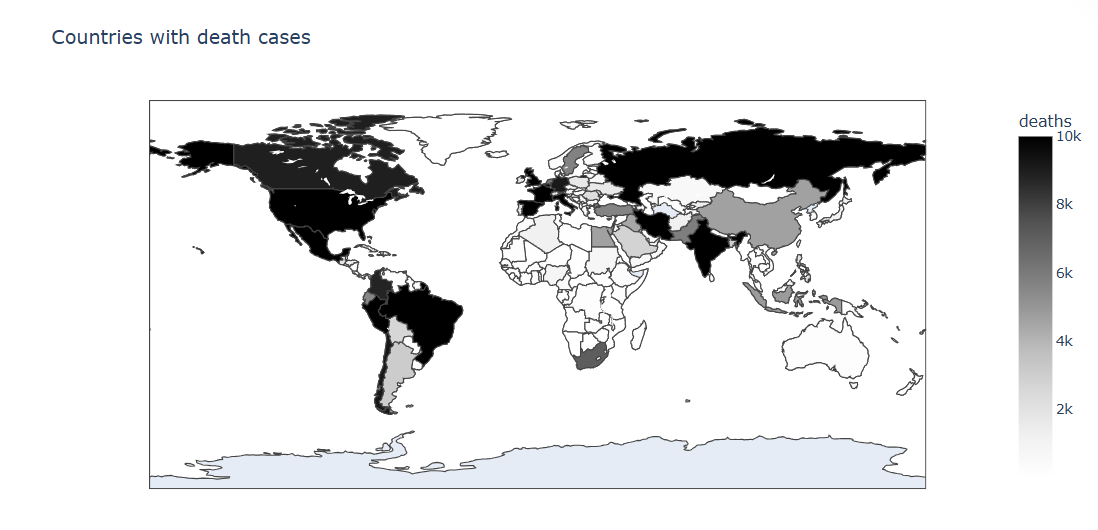
**Visualizations for Confirmed Cases**

****

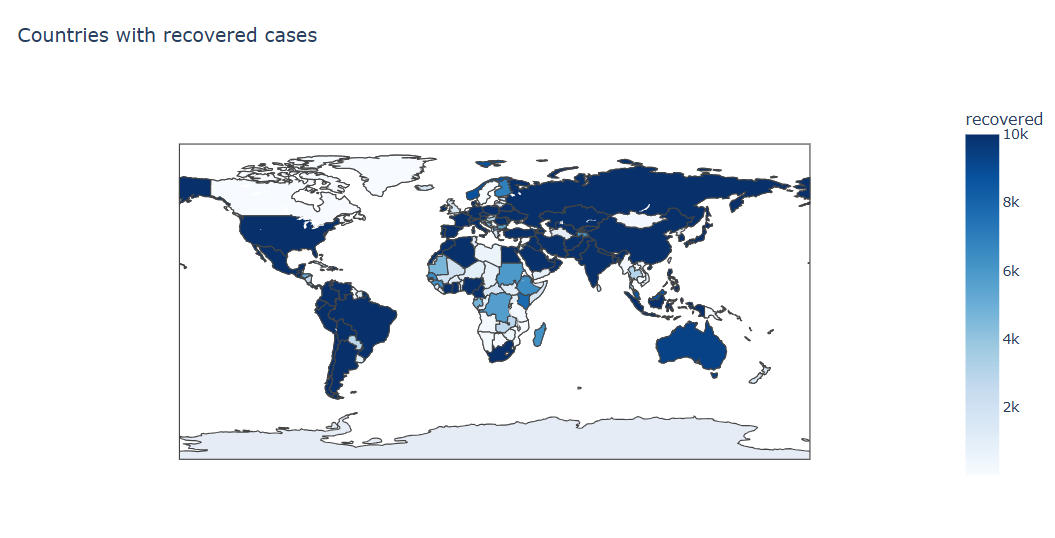
**Visualizations for Confirmed Cases**

****

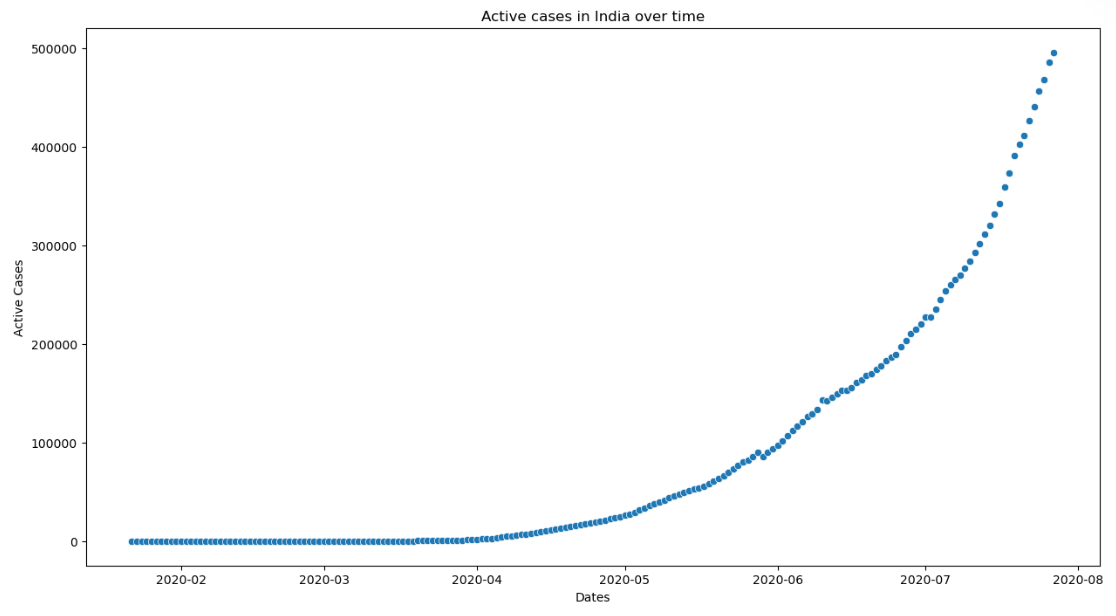
**Visualizations for Deaths Cases**

****

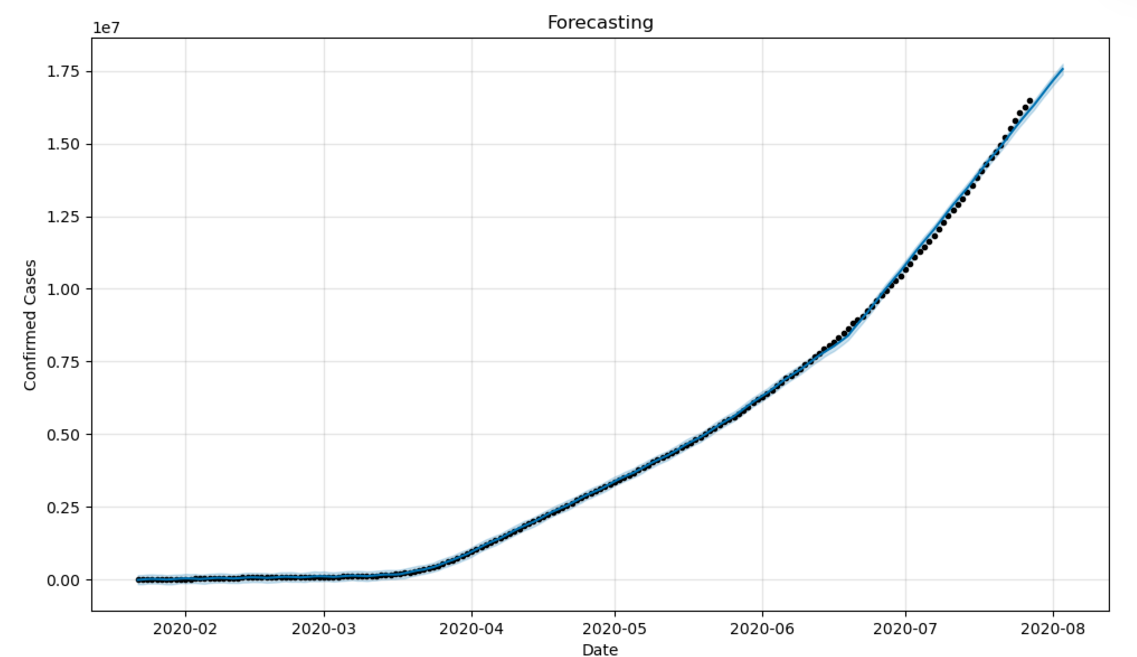
**Visualizations for Recovered Cases**

****

**Visualizations for India’s Active Cases**

****

**Forecasting Next 7 Days Confirmed Cases**

****

**As we are able to see the forecasted values in the form of line.**

# Future Possibilities

Exploring future possibilities for enhancing the COVID-19 forecasting project can lead to more accurate predictions, better resource management, and improved public health strategies. Here are some promising directions:

#### 1. Incorporating Advanced Models

* **Hybrid Models**: Combine Prophet with other time series models and machine learning techniques, such as Long Short-Term Memory (LSTM) networks, to capture more complex patterns and dependencies in the data.
* **Ensemble Methods**: Use ensemble methods that aggregate predictions from multiple models to improve forecast accuracy and robustness.

#### 2. Real-Time Forecasting

* **Streaming Data Integration**: Implement real-time data collection and forecasting, allowing the model to continuously update with the latest case data for more timely and relevant predictions.
* **Automated Alerts**: Develop systems that automatically generate alerts for potential outbreaks or significant changes in case trends, enabling faster response times.

#### 3. Enhanced Feature Engineering

* **Incorporate Additional Variables**: Include more explanatory variables such as mobility data, social distancing compliance, vaccination rates, and economic indicators to improve model accuracy.
* **Event Impact Analysis**: Perform detailed analysis of the impact of specific events (e.g., mass gatherings, policy changes) on case numbers to refine the model’s sensitivity to such factors.

#### 4.Geographic-Specific Forecasting

* **Regional Models**: Develop region-specific models to account for local variations in COVID-19 dynamics, allowing for more targeted public health interventions.
* **Geospatial Analysis**: Use geospatial data to map the spread of the virus and identify clusters, enhancing the understanding of transmission patterns

# Conclusion

The COVID-19 forecasting project using the Prophet library has provided significant insights into the dynamics of the pandemic, showcasing the power of data-driven approaches in public health planning and response. Through meticulous data pre-processing, model building, and evaluation, we have achieved the following:

#### Key Findings:

1. **Trend and Seasonality**: The model successfully captured underlying trends and seasonal patterns in the COVID-19 case data, providing a clear picture of the virus's spread and fluctuations over time.
2. **Effectiveness of Interventions**: Analysis of the impact of various public health interventions, such as lockdowns and vaccination campaigns, demonstrated their critical role in controlling the spread of the virus.
3. **Short-Term Forecasts**: The Prophet model delivered reliable short-term forecasts, enabling health authorities to make informed decisions about resource allocation and intervention strategies.
4. **Data Quality**: The importance of high-quality, comprehensive data was underscored, highlighting the need for accurate and timely information for effective modeling and forecasting.