**Capstone Project Technical Report:**

**Predictive Modelling for COVID-19 in Public Health**

**Table of Contents**

[**Aim:** 2](#_Toc183594112)

[**Objectives:** 2](#_Toc183594113)

[**Data Preparation:** 2](#_Toc183594114)

[**Exploratory Data Analysis** 3](#_Toc183594115)

[**1. Key trends were analysed and visualised using line and bar charts:** 3](#_Toc183594116)

[**2. Correlation Analysis** 7](#_Toc183594117)

[**Feature Engineering** 10](#_Toc183594118)

[**Predictive Modelling** 11](#_Toc183594119)

[**Linear Regression:** 11](#_Toc183594120)

[**Decision Tree Regressor:** 11](#_Toc183594121)

[**Random Forest Regressor:** 11](#_Toc183594122)

[**Public Health Implications of this Project** 12](#_Toc183594123)

[**Conclusion** 13](#_Toc183594124)

# **Aim:**

To use historical COVID-19 data to develop predictive models to forecast COVID-19 trends in the United States. This will provide “Health Guard Analytics” with actionable insights to inform policies, anticipate future outbreaks, and improve health resource allocation.

# **Objectives:**

* To perform data preprocessing to ensure that the data is clean and normalized for the ML model
* To perform EDA on the data to uncover relevant trends and correlations/relationships
* To build and train a model that can forecast future COVID-19 trends

# **Data Preparation:**

The source of data for this capstone project contained 6 datasets;

* Worldometer\_data
* Country\_wise\_latest
* Full\_grouped
* USA\_country\_wise
* Day\_wise
* COVID-19\_clean\_complete

Out of the six datasets, two (OVID-19\_clean\_complete and Full\_grouped) were selected because they were relevant to helping me achieve my project aims and objectives.

Data cleaning and preprocessing including removing null and duplicated values and ensuring a consistent data type were carried out using Python programming language on Google Collab.

# **Exploratory Data Analysis**

## **1. Key trends were analysed and visualised using line and bar charts:**

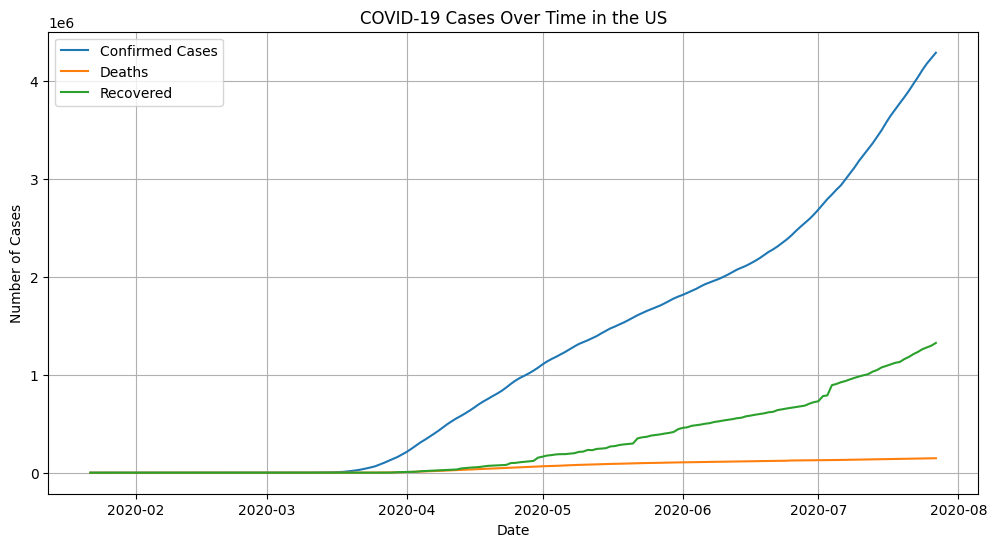


Figure 1: COVID-19 trends over time in the US

Figure 1 is a line graph with a gridline showing the total confirmed cases, recovered cases and death cases in the US from 22nd January to 27th July. From the chart above, the total confirmed cases were levelled from late January to late March, but after that it saw a spike and upward trend, depicting a significant increase in the total number of confirmed cases. Total recovery also followed a similar trend, however, total death increased only slightly.

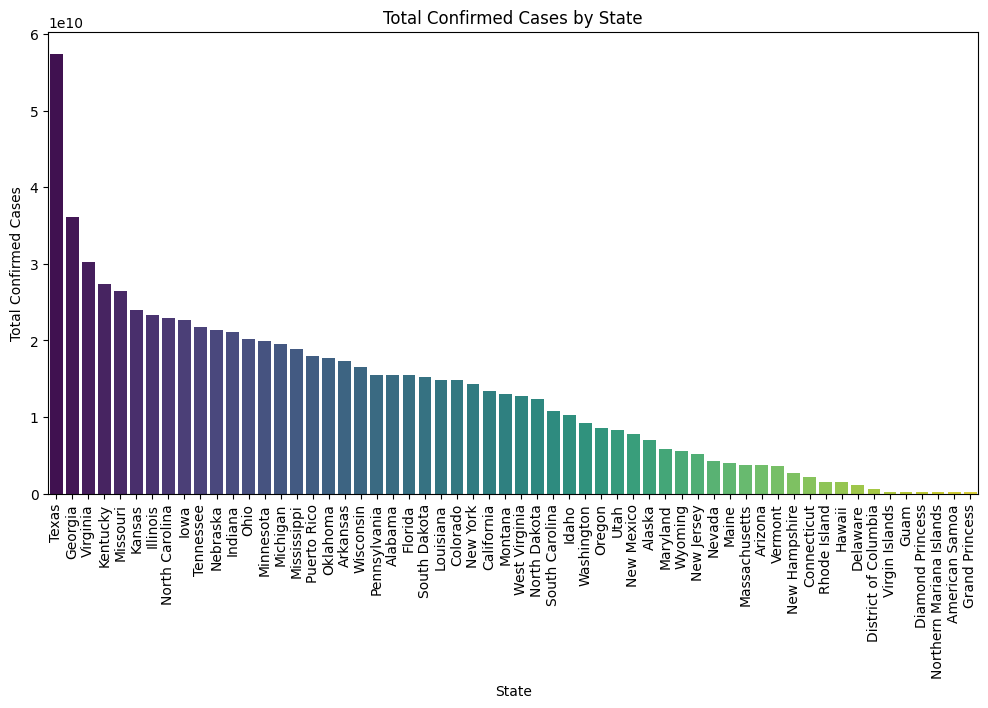


Figure 2: Total COVID-19 confirmed cases in the US by States

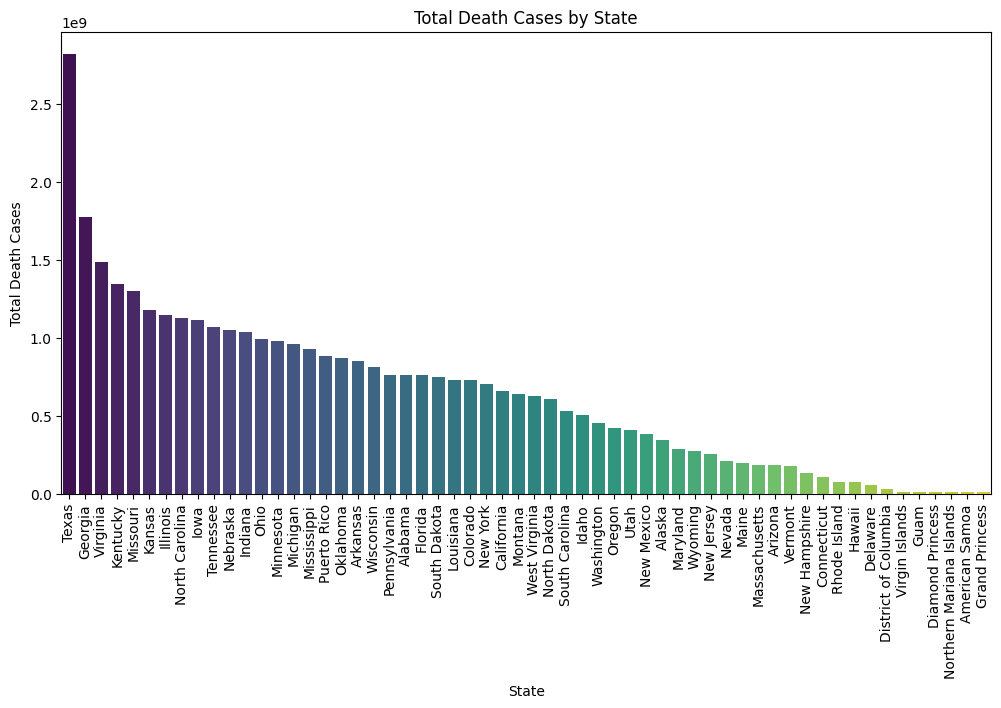


Figure 3: Total COVID-19 death cases in the US by States

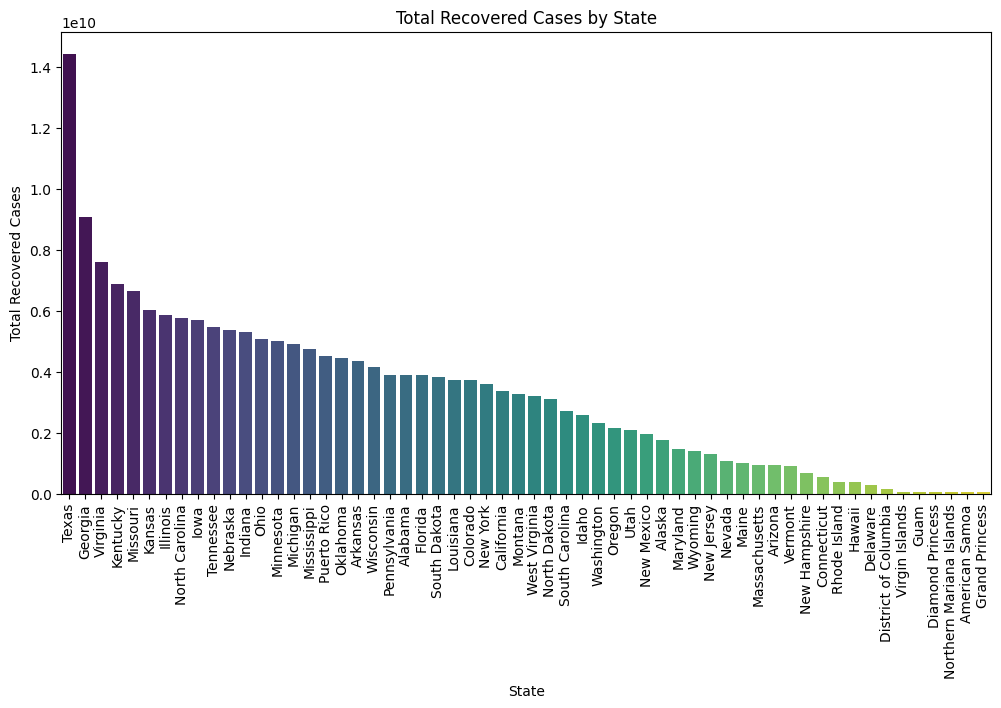


Figure 4: Total COVID-19 recovered cases in the US by States

Figures 2, 3, and 4, are bar charts that show the total confirmed, death and recovered COVID-19 cases in the US from 22nd January to 27th July 2020 respectively.

From those 3 charts, out of the 50 states in the USA, Texas had the highest COVID-19 confirmed, recovered and death cases followed by Georgia, Virginia and Kentucky. States with the least confirmed and recovered cases and deaths were; Northern Mariana Islands, American Samoa and Grand Princess. High cases in Texas for instance, may be due to its large population (about 29 Million) and high mobility, while lower cases in territories could be due to smaller populations and geographic isolations.

## **2. Correlation Analysis**

This analysis was carried out to understand the relationships between variables in the data.

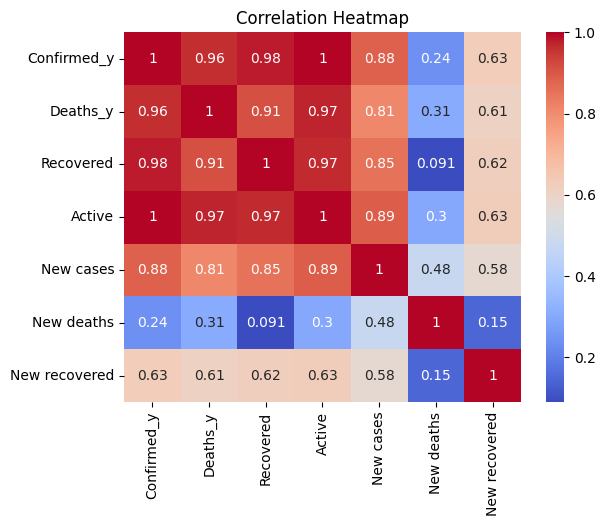


Figure 6: Correlation heatmap of key variables

The correlation heatmap above explains how one variable in the data is affected by another.

From the heatmap, the red colour indicates that a variable has a positive correlation, and the strength of this correlation increases as the colour intensity increases and as the value is closer to 1. Conversely, the blue colour indicates that a variable has a negative correlation. The strength of the negative correlation increases as the colour intensity increases and the value is farther from -1.

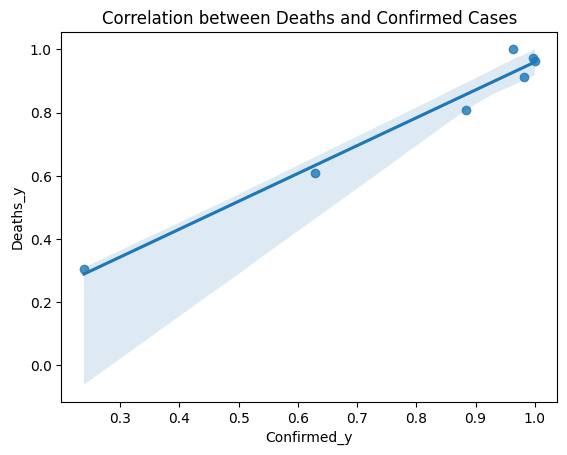


Figure 7. Correlation between death and confirmed cases

This regression plot illustrates that there is a positive correlation between confirmed cases and death cases, meaning that an increase in confirmed cases can lead to an increase in death cases.

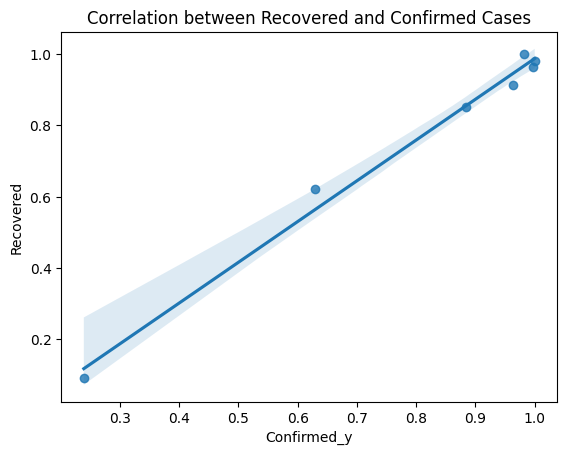


Figure 8. Correlation between confirmed cases and recovered cases

Figure 8 is a regression plot illustrates that there is a positive correlation between confirmed cases and recovered cases, meaning that an increase in confirmed cases can lead to an increase in recovered cases.

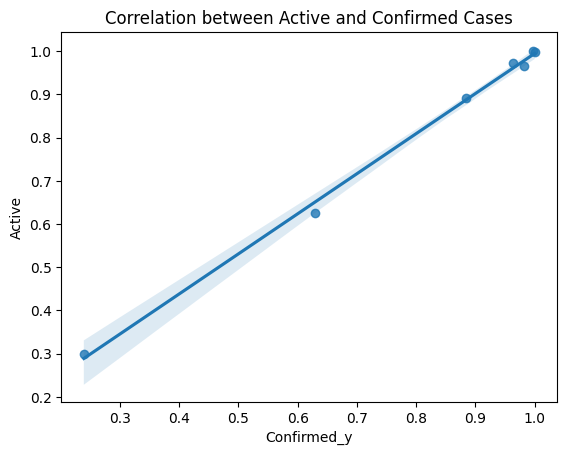


Figure 9. Correlation between active and confirmed cases

Figure 9 is a regression plot illustrates that there is a positive correlation between confirmed cases and active cases, meaning that an increase in confirmed cases can lead to an increase in active cases.

# **Feature Engineering**

To enhance the quality of the data and strengthen the predictive model insights, feature engineering was conducted. The following derived variables were created:

Growth Rate = Confirmed cases/ Percentage change between the current and prior cases

Mortality Rate = Death/Confirmed cases

Recovery Rate = Recovered/Confirmed cases

Active Case Ratio = Active cases/Confirmed cases

New Case Per 100k of the Population = New cases/Population

Test Per Confirmed Cases = Total Test/Confirmed cases

# **Predictive Modelling**

The aim was to build a model that can forecast future active COVID-19 cases

Three Algorithms were used for building the model;

* Linear Regression
* Decision Tree Regressor
* Random Forest Regressor

## **Linear Regression:**

80% of the dataset as training data

20% of the dataset as testing data

**Evaluation;**

* Mean Squared Error (MSE) = 3.095999362488507e-18
* R-squared (R²) = 1.0
* Root Mean Squared Error (RMSE) = 1.7595452146757999e-09

## **Decision Tree Regressor:**

80% of the dataset as training data

20% of the dataset as testing data

**Evaluation;**

* Mean Squared Error: 0.0
* R-squared: 1.0
* Root Mean Squared Error: 0.0
* Accuracy: 1.0

## **Random Forest Regressor:**

n-estimators = 100

random\_state = 42

**Evaluation;**

* Mean Squared Error: 0.0
* R-squared: 1.0
* Root Mean Squared Error: 0.0

MSE tells us how far off the model's predictions are from the actual values, on average. A smaller MSE (such as 3.0) means that the prediction are closer to actual values, thus a better model.

R² shows how well the model explains the variability in the data. An R² of 1 means that the model perfectly predicts the data (explains 100% of the variation). Most times, a higher R2 could be due to overfitting of the model.

RMSE is the square root of the MSE. A lower RMSE (such as 1.7) means that it’s a better prediction as it is close to the actual value.

# **Public Health Implications of this Project**

By accurately forecasting active COVID-19 cases, my model can help health authorities in the US predict surges in COVID-19 cases, which will allow for better allocation of critical resources, such as hospital beds, ventilators, and medical staff, ensuring that healthcare systems are not overwhelmed.

The model provides quantitative insights into the progression of the pandemic, enabling governments to make evidence-based decisions about public health measures.

# **Conclusion**

In this capstone project, I demonstrated my data science and analytical skills by analyzing a COVID-19 dataset and building predictive models to forecast COVID-19 trends in the United States from January 22 to July 27, 2020. The project had three primary objectives: data preprocessing to clean and normalize the dataset, exploratory data analysis (EDA) to uncover trends and relationships, and building machine learning models to predict future active COVID-19 cases.

The trends in confirmed, death, and recovered cases were visualized using line charts. Initially, there were almost no confirmed cases in the US, as the line remained flat, but a sharp spike occurred in late March, continuing in an upward trend. A similar pattern was observed for recovered cases, while death cases increased at a slower rate (refer to Figure 1). State-level analysis revealed that Texas had the highest confirmed, recovered, and death cases, followed by Georgia and Virginia, likely due to its large population (over 29 million) and central location. Conversely, territories such as American Samoa and Northern Mariana Islands had the lowest case numbers, likely due to their small populations and geographic isolation.

The relationships between these COVID-19 trends were analyzed using a correlation heatmap and regression plots. Positive correlations were identified between confirmed cases and recovered cases, death cases, and active cases, suggesting that an increase in confirmed cases is associated with corresponding increases in other metrics.

Finally, predictive models were developed using three algorithms to forecast future active cases. These models have significant public health implications, as they can assist agencies like the CDC in predicting surges in COVID-19 cases. This enables better resource allocation—such as hospital beds, ventilators, and medical staff—helping ensure healthcare systems remain resilient during outbreaks.