CAPSTONE PROJECT REPORT

Capstone Project: Predictive Modelling for COVID-19 in Public Health

The COVID-19 pandemic was one of the worse pandemic that shook the whole world, redefining our lives. Its effects resonated to every part of the world. As a data scientist, I will be exploring the data gathered to make meaningful insights and recommendations that will help better understand facts from the data.
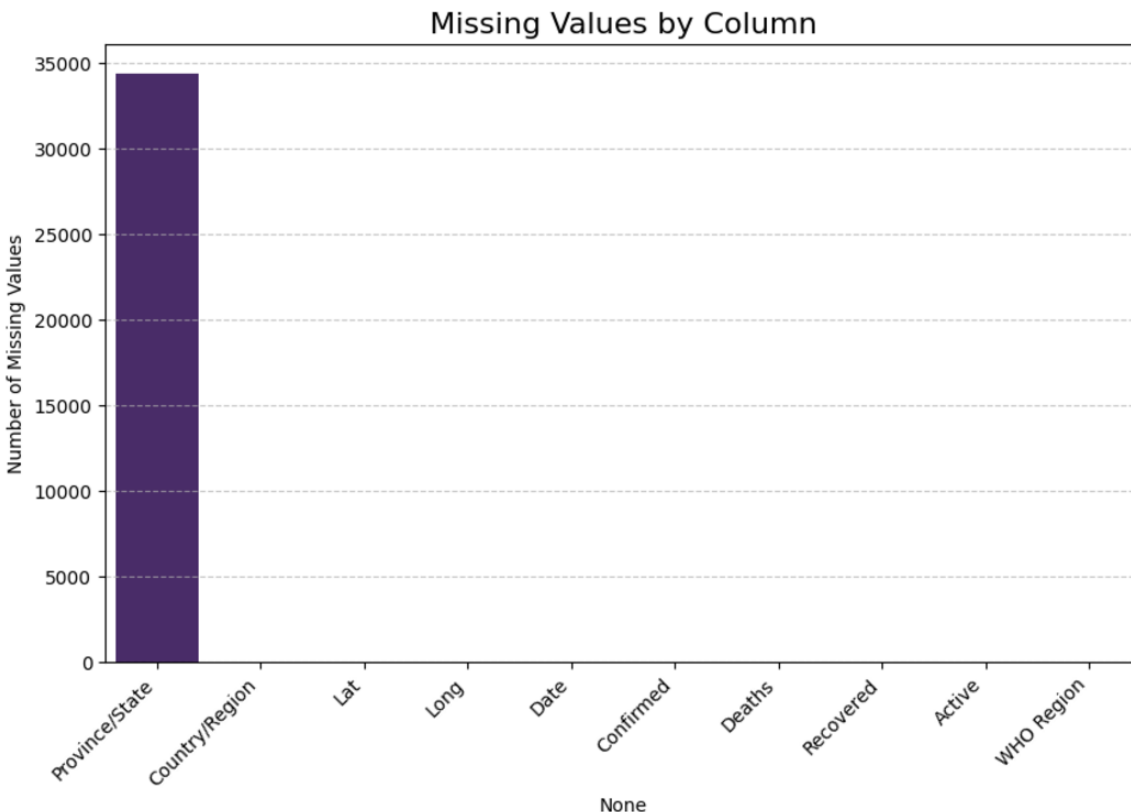
Methodology

Data Collection

● Dataset: Access the COVID-19 Open Research Dataset (CORD-19) on

Kaggle, which includes COVID-19 case counts, demographic data, and

various health metrics.

● Dataset Source: CORD-19 on Kaggle

Data Preprocessing:

The data had missing data on the "Province/State" column as shown in the chart below.
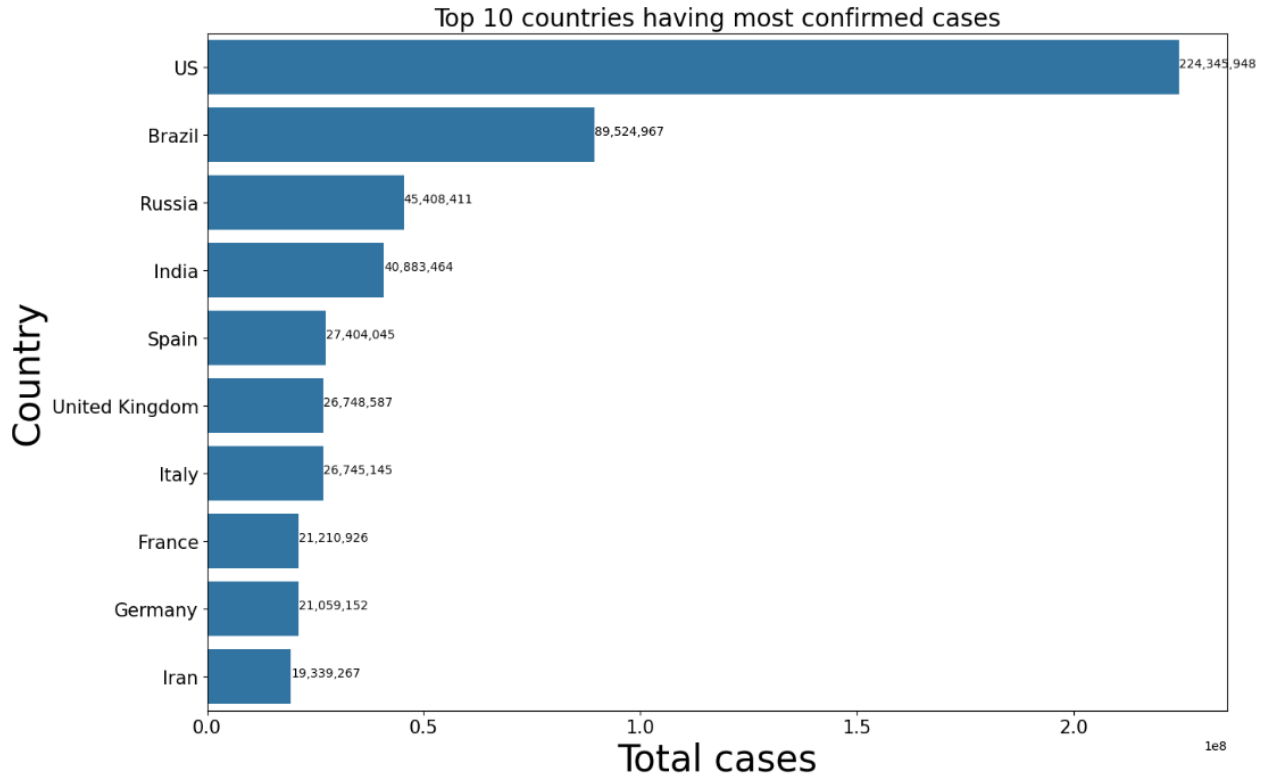

I resolve the issue of missing values by filling it. I standardize the dates to have a uniform date format.

Exploratory Data Analysis (EDA)

The data revealed that there is a progressive increase in the number of confirmed cases globally. The top country as regards to the confirmed cases is US.
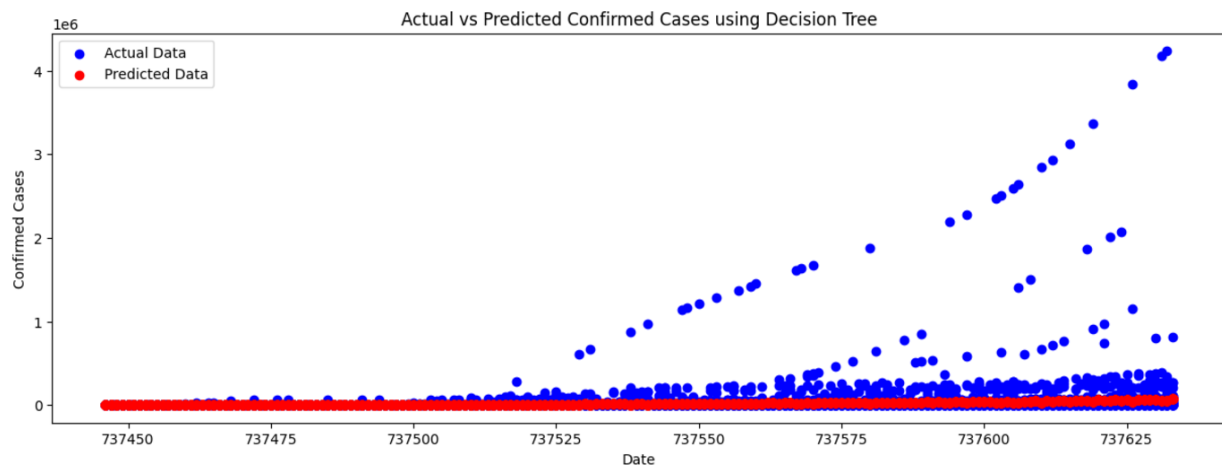
Text(0.5, 1.0, 'Top 10 countries having most confirmed cases')
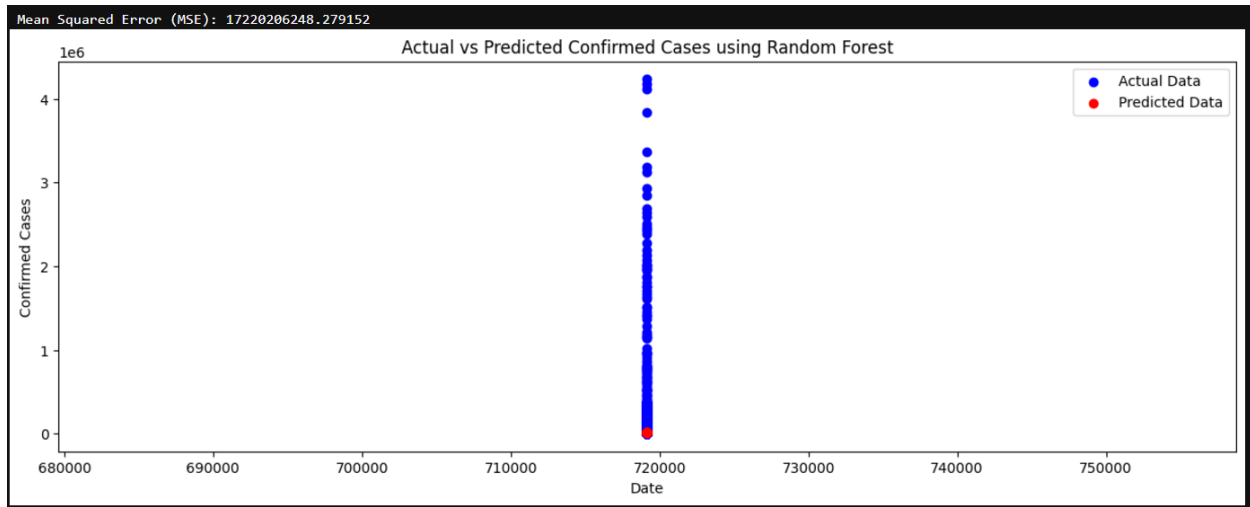

Model Development

I also proceed to predict the model. I used the Decision Tree Regressor and the Random Forest Regressor to predict the model. This is shown by the charts below respectively.

Mean Squared Error (MSE): 18838245490.75016

The actual value follows the events in the past as we have seen by the data. That is, an increasing number of cases. However, the predicted data show the opposite. According to the predicted data we see that the increase is under control. This shows that there is a significant improvement in the combat against the COVID-19 pandemic.