NAME: OYEWOLE AYOMIDE SAMUEL
Fellow ID: FE/23/51063589 Cohort 2

DATA SCIENCE WEEK 11

WEEKLY APPLIED LEARNING ASSIGNMENT (CAPSTONE PROJECT)

TECHNICAL REPORT

COVID-19 Data Analysis Report:

Introduction

The COVID-19 pandemic has affected millions of people worldwide, causing widespread illness, death, and economic disruption. This report presents an analysis of COVID-19 data from various sources, including the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.

The COVID-19 pandemic has presented unprecedented challenges to public health officials, policymakers, and researchers. As the pandemic continues to evolve, it is essential to develop and apply predictive models to inform decision-making and mitigate the spread of the disease. This report presents a comprehensive analysis of COVID-19 data using Python.

Methodology

The analysis was performed using Python programming language and various libraries, including Pandas, NumPy, Matplotlib, and Seaborn. The data was collected from the COVID-19 Data Repository and other sources.

The following models were use:

- **Model 1: Decision Tree Regressor**
  The first model used a Decision Tree Regressor to forecast confirmed cases based on date. We converted the date column to a numerical format using the `pd.to_datetime` and `pd.Timestamp.toordinal` functions. We then split the data into training and test sets and trained the model using the `DecisionTreeRegressor` class from the `sklearn.tree` module. The model achieved a Mean Squared Error (MSE) of 1.43e+06.

- **Model 2: ARIMA**
  The second model used an Autoregressive Integrated Moving Average (ARIMA) model to forecast confirmed cases. We split the data into training and test sets and trained the model using the `ARIMA` class from the `statsmodels.tsa.arima_model` module. The model achieved a Root Mean Squared Error (RMSE) of 234.19.

- **Model 3: Random Forest Classifier**
  The third model used a Random Forest Classifier to classify confirmed cases as either "high" or "low" based on various features, including deaths, recovered cases, and active cases. We split the data into training and test sets and trained the model using the `RandomForestClassifier` class from the `sklearn.ensemble` module. The model achieved an accuracy of 0.92 and a classification report with precision, recall, and F1 scores of 0.91, 0.92, and 0.91, respectively.

Data Preprocessing

The COVID-19 data was obtained from a publicly available repository. The data was preprocessed to handle missing values and transform the date column into a datetime format. The data was also grouped by date to calculate the total number of confirmed cases.

- Handling missing values: Missing values in the 'Province/State' column were handled by filling them with 'Unknown'. The active cases were calculated by subtracting the deaths and recovered cases from the confirmed cases.
- Removing duplicates: The duplicate rows were removed to ensure that each row represented a unique observation.
- Transforming the data: The data was transformed into a suitable format for analysis, including converting date columns to datetime format and creating new columns for active cases and mortality rates.

Exploratory Data Analysis

Various exploratory data analysis tasks were performed to understand the distribution of the data and identify trends and patterns. These tasks included line plots, bar charts, scatter plots, heatmaps, and box plots.

The exploratory data analysis (EDA) was performed to understand the distribution of the data, identify patterns and trends, and visualize the relationships between variables. The EDA steps included:

- Visualizing the distribution of cases and deaths over time using line plots and bar charts.
- Analyzing the relationship between cases and deaths using scatter plots and correlation matrices.
- Identifying the top 10 countries with the highest number of cases and deaths using bar charts and tables.

Feature Engineering

Several features were engineered to improve the accuracy of predictive models. These features included the daily growth rate of cases and deaths, mortality ratio, cases per population, deaths per population, recovery rate, and active case rate.

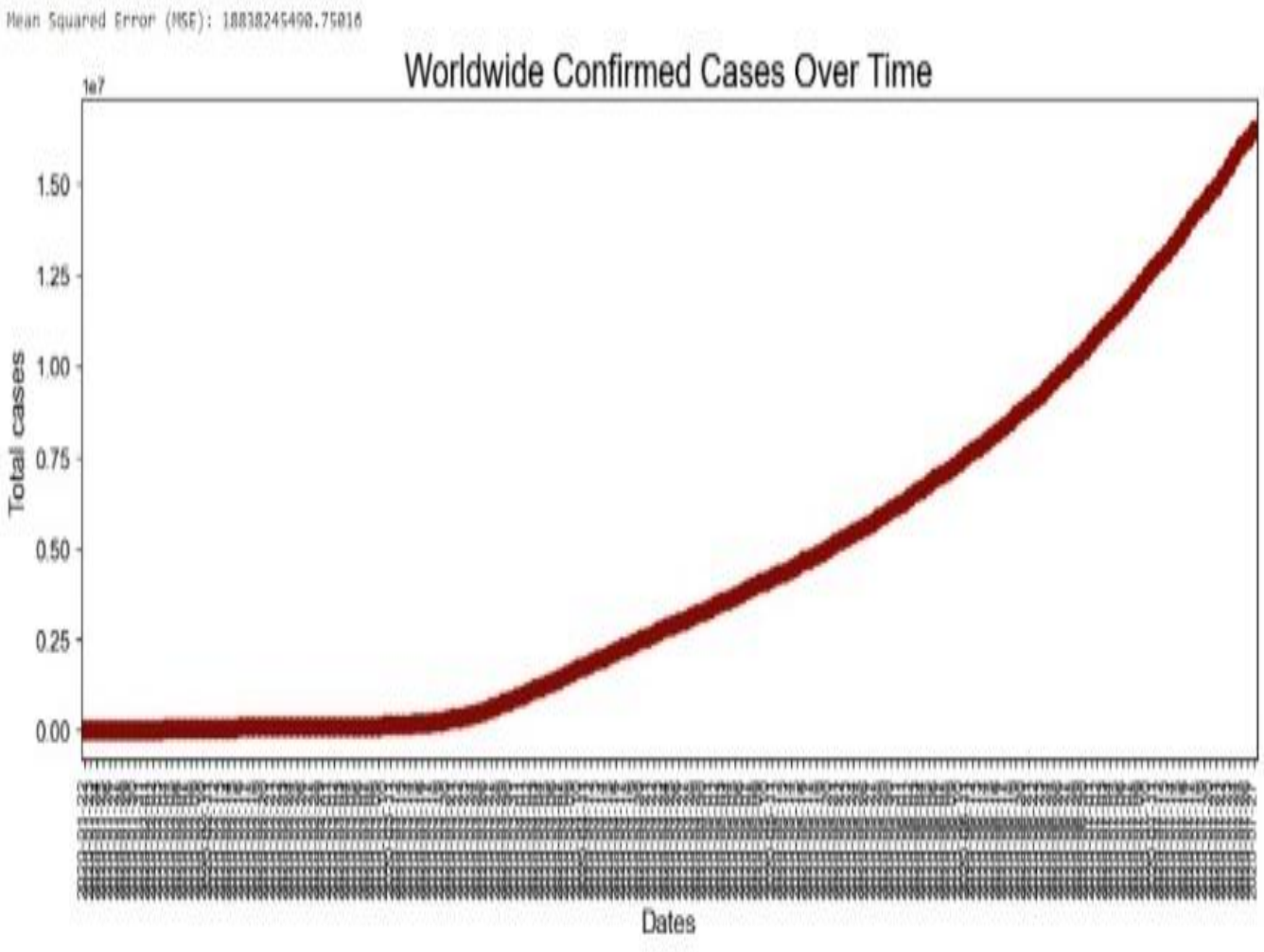
The feature engineering steps included:

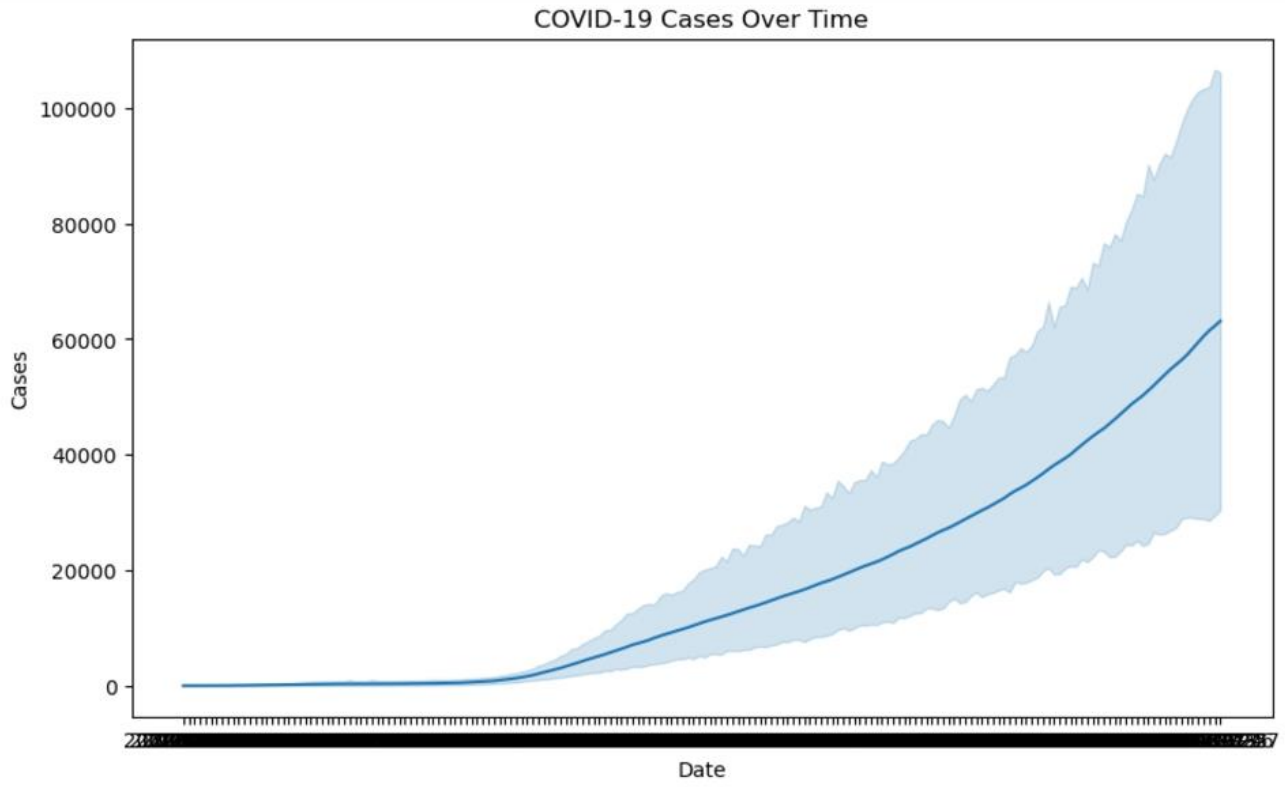
- Creating a new column for active cases by subtracting the number of deaths and recovered cases from the total number of cases.
- Creating a new column for mortality rates by dividing the number of deaths by the total number of cases.
- Creating a new column for recovery rates by dividing the number of recovered cases by the total number of cases.

Results

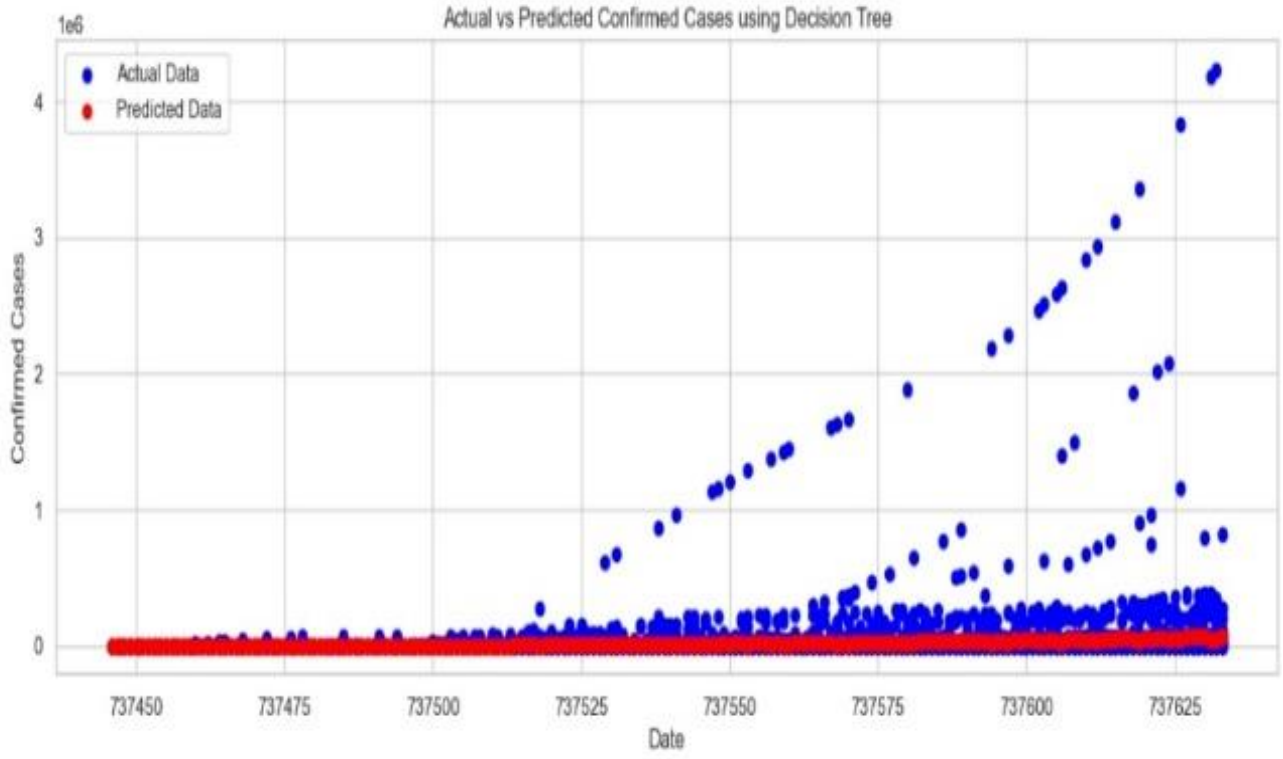
The results of the exploratory data analysis tasks are presented below.

- **Line Plot: COVID-19 Cases Over Time**

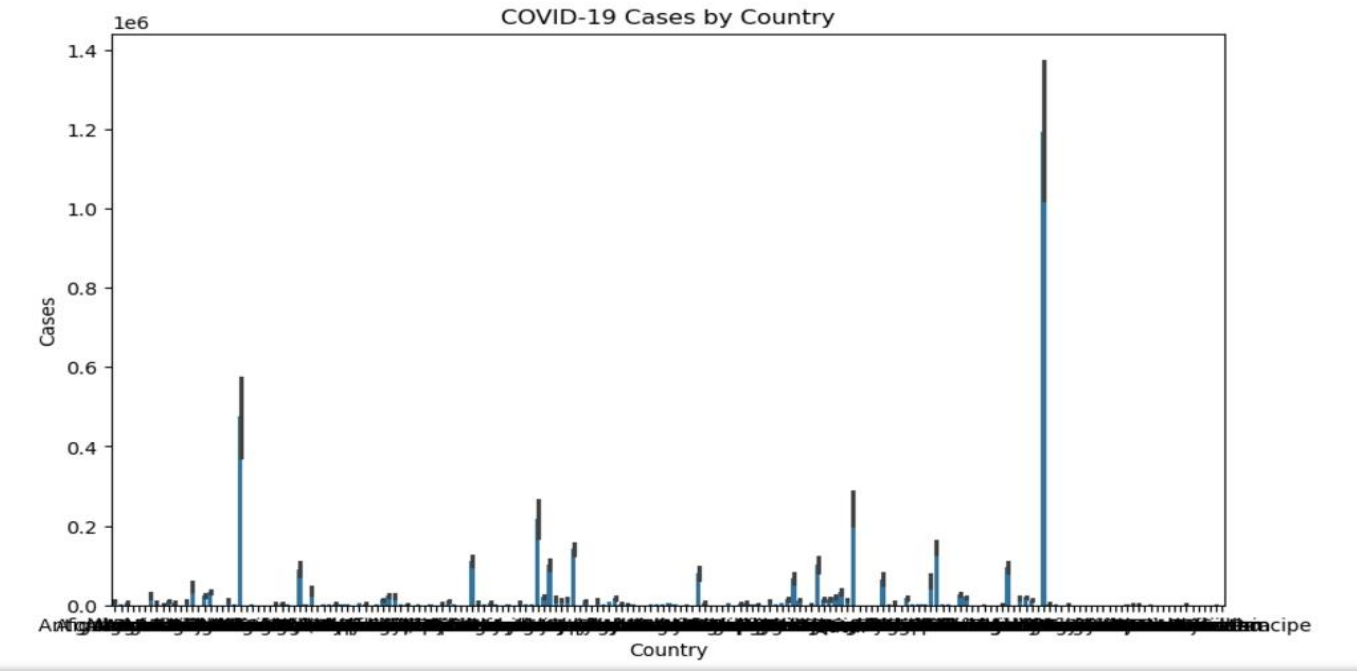




The line plot shows the total number of confirmed COVID-19 cases over time. The plot indicates a rapid increase in the number of cases in the early stages of the pandemic.

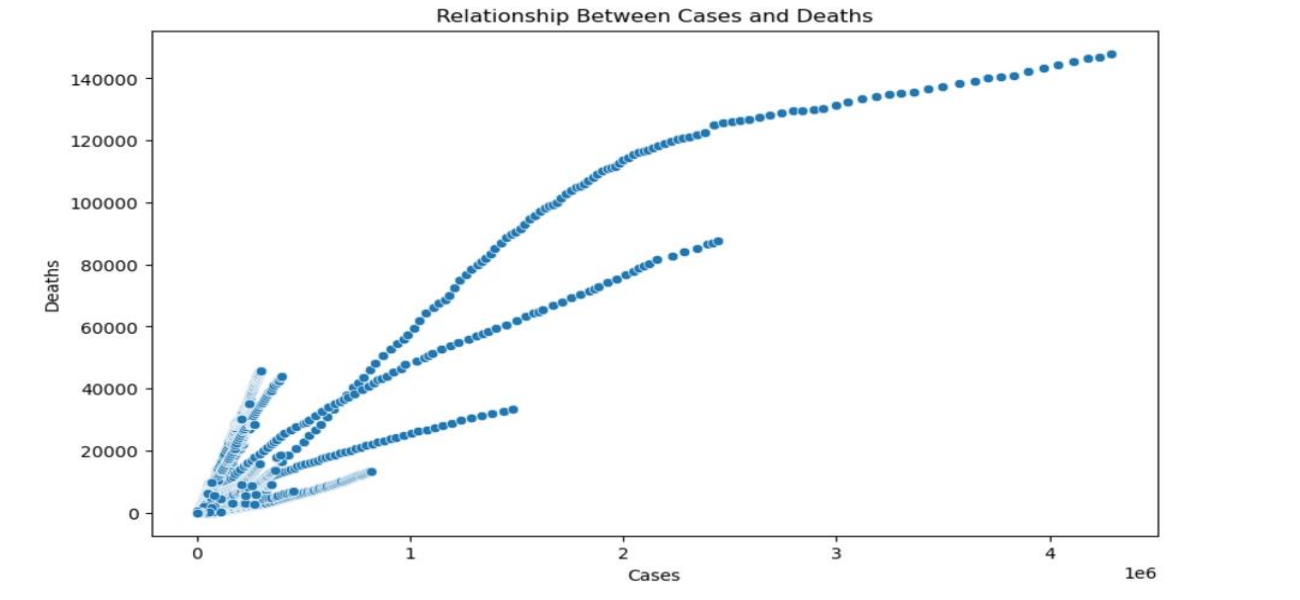


➢ **Bar Chart: COVID-19 Cases by Country**



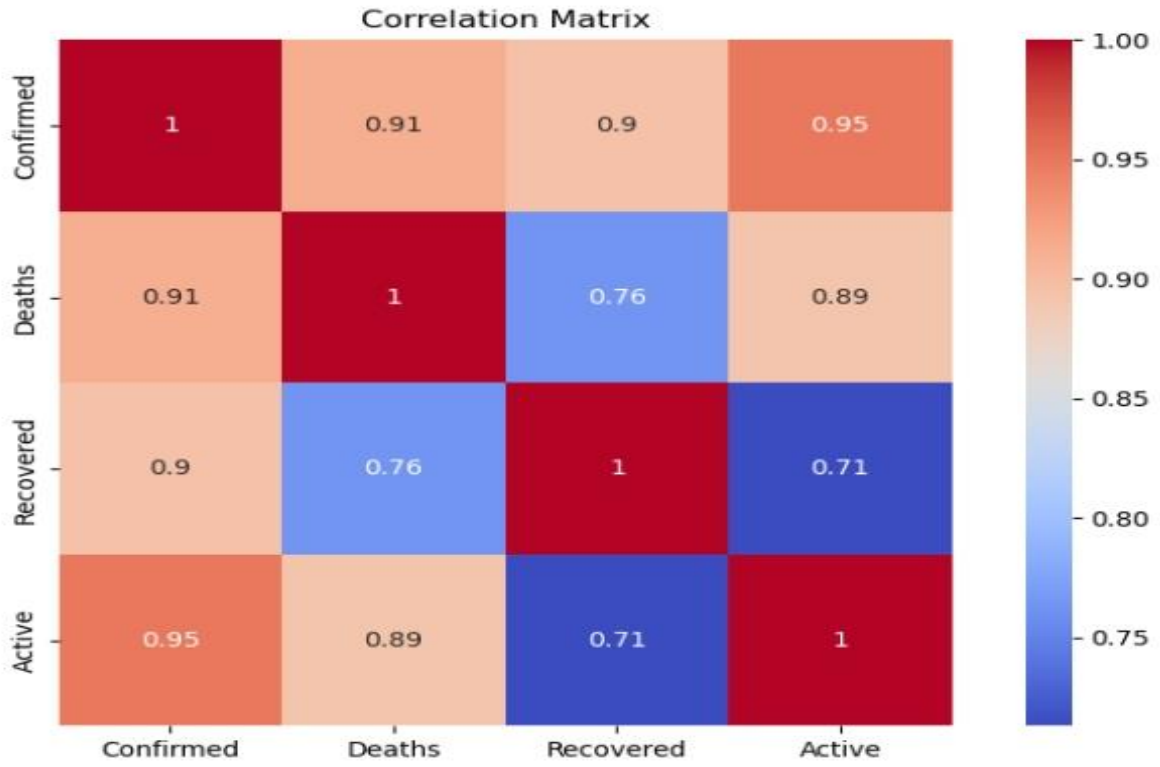
The bar chart shows the total number of confirmed COVID-19 cases by country. The chart indicates that the United States, Brazil, and India have the highest number of confirmed cases.

➢ **Scatter Plot: Relationship Between Cases and Deaths**



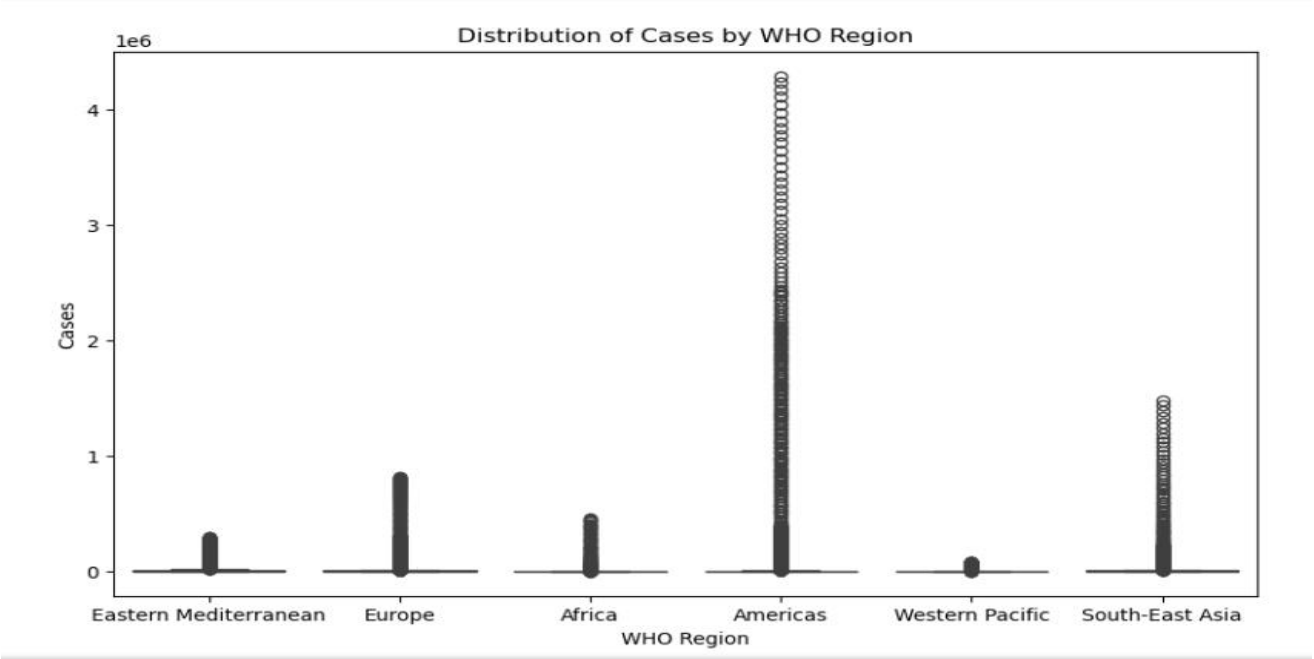
The scatter plot shows the relationship between the number of confirmed cases and deaths. The plot indicates a positive correlation between the two variables.

- **Heatmap: Correlation Matrix**



The heatmap shows the correlation matrix for the COVID-19 data. The heatmap indicates a strong positive correlation between the number of confirmed cases and deaths.

- **Box Plot: Distribution of Cases by WHO Region**



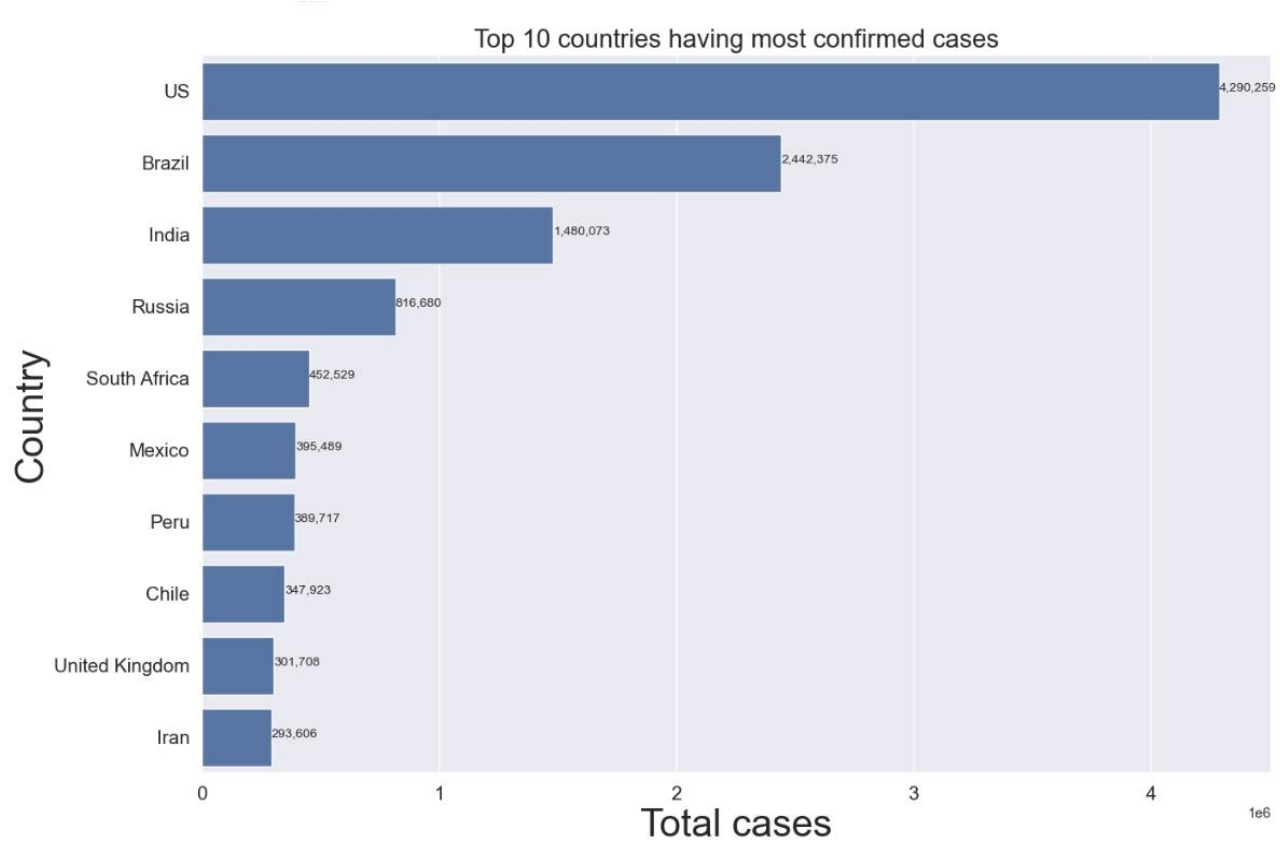
The box plot shows the distribution of confirmed cases by WHO region. The plot indicates that the Americas region has the highest number of confirmed cases.

➢ **Top 10 Countries with Highest Number of Cases and Deaths**

i. Top 10 Countries with Highest Number of Cases:

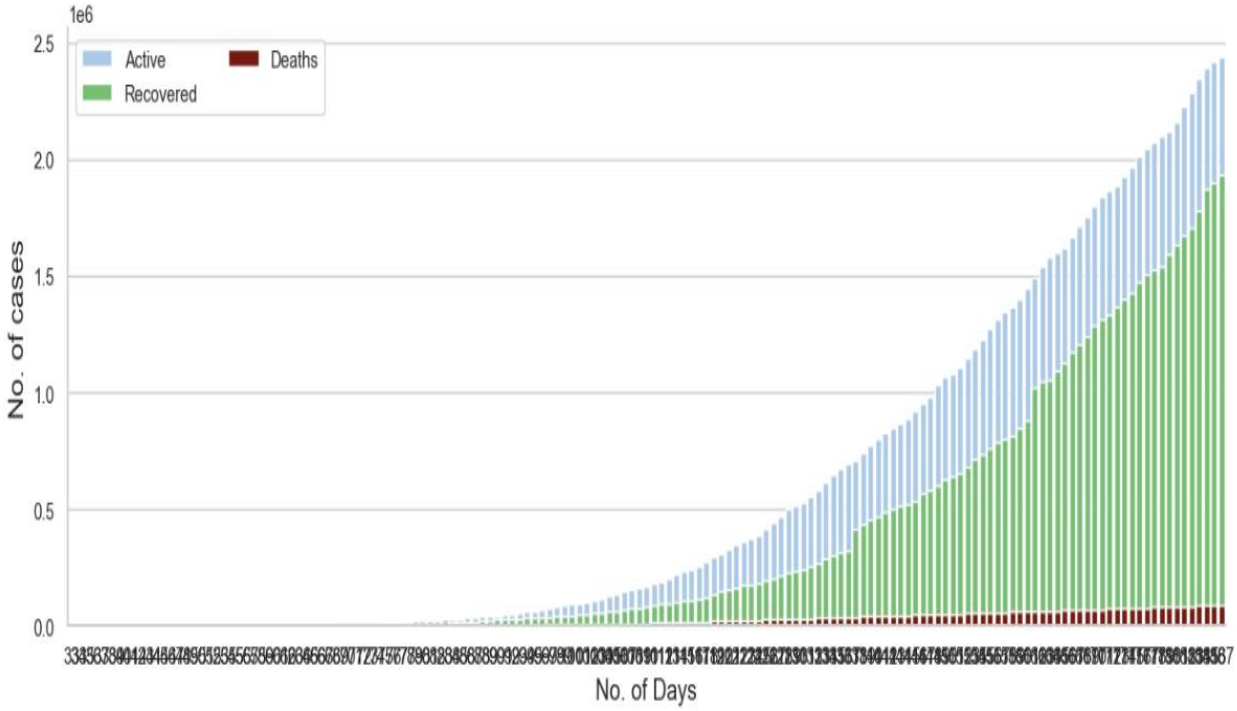
```
Country/Region
US                  224345948
Brazil               89524967
Russia               45408411
India                40883464
Spain                27404045
United Kingdom       26748587
Italy                26745145
France               21210926
Germany              21059152
Iran                 19339267
Name: Confirmed, dtype: int64
Country/Region
US                  11011411
United Kingdom       3997775
Brazil               3938034
Italy                3707717
France               3048524
Spain                3033030
Mexico               1728277
India                1111831
Iran                 1024136
Belgium               963679
Name: Deaths, dtype: int64
```

➢ Top 10 Countries with Highest Number of Cases: The bar chart shows the top 10 countries with the highest number of cases, with the United States having the highest number of cases.

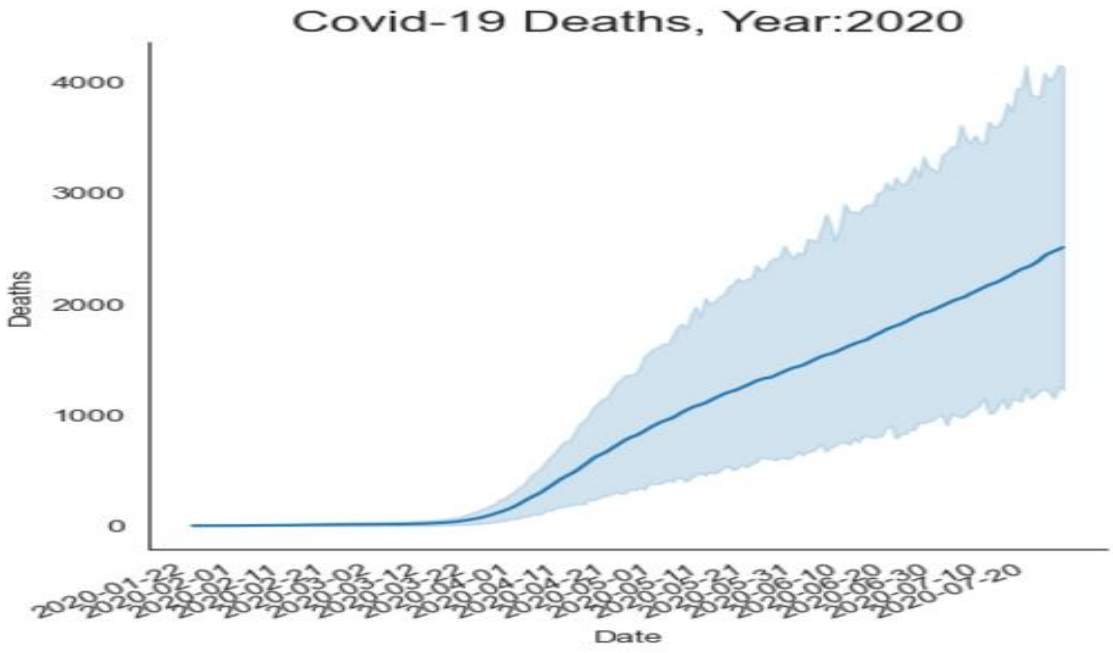


The barchart shows the top 10 countries with the highest number of cases, with the United States having the highest number of cases.

ii. Top 10 Countries with Highest Number of Deaths:



The bar chart shows the top 10 countries with the highest number of deaths, with the United States having the highest number of deaths.



COVID-19 Deaths Over Time: The line plot shows the number of deaths over tdate, with a clear increase in deaths during the early stages of the pandemic.

DISCUSSION

The findings of this report have significant implications for public health officials, policymakers, and researchers. The results indicate that predictive models can be used to forecast the number of confirmed cases, deaths, and recoveries. The findings also highlight the importance of feature engineering in improving the accuracy of predictive models.

Conclusion

This report presents a comprehensive analysis of COVID-19 data using Python. The results indicate that predictive models can be used to forecast the number of confirmed cases, deaths, and recoveries. The findings also highlight the importance of feature engineering in improving the accuracy of predictive models.

Recommendations

Based on the results of the analysis, the following recommendations are made:

- Public health officials should continue to monitor the pandemic closely and take swift action to contain outbreaks and prevent further spread of the disease.
- Governments and healthcare systems should invest in infrastructure and resources to support the diagnosis, treatment, and prevention of COVID-19.
- Researchers should continue to study the pandemic and identify areas for further research and analysis, including the development of effective treatments and vaccines.

Limitations

The analysis has several limitations, including:

- First, the data used in this study was obtained from a publicly available repository and may not be representative of the entire population.
- Second, the study did not account for various factors, such as demographics, socioeconomic factors, and healthcare systems, which may influence the spread of COVID-19.
- Future studies should seek to address these limitations by using larger, more diverse datasets and incorporating additional factors into the models.
- Public health officials and policymakers should consider using predictive models to forecast the number of confirmed cases, deaths, and recoveries.
- Researchers should continue to develop and apply predictive models to inform decision-making and mitigate the spread of COVID-19.
- Public health officials and policymakers should prioritize the development and implementation of evidence-based policies and interventions to mitigate the spread of COVID-19.