**Technical Report: COVID-19 Predictive Modeling**

**Capstone Project**

**Author: Odion Joel Okonofua  
ID: FE/23/88847286**

**Introduction**

**COVID-19 Overview  
COVID-19 (Coronavirus Disease 2019) is a highly infectious respiratory illness caused by the SARS-CoV-2 virus. First identified in December 2019 in Wuhan, China, the disease rapidly spread worldwide, resulting in a global pandemic. COVID-19 impacts the respiratory system and varies in severity from mild flu-like symptoms to severe respiratory complications and fatalities, especially among vulnerable populations such as the elderly and those with pre-existing health conditions. Governments and organizations worldwide have implemented various measures, such as lockdowns, vaccinations, and public health campaigns, to control the spread and mitigate its impacts.**

**About the Dataset  
The dataset used for this project provides comprehensive data on the progression of COVID-19 across countries and regions. It includes key metrics such as confirmed cases, deaths, recoveries, active cases, and new daily cases. Additionally, derived metrics like growth rates, mortality ratios, and recovery ratios help uncover patterns and trends in the spread and impact of the disease. The dataset also covers regional and temporal information, enabling a detailed analysis of how COVID-19 evolved over time and across geographical areas.**

**This data was instrumental in conducting exploratory analysis, feature engineering, and building predictive models for forecasting cases and understanding key factors influencing the pandemic's dynamics.**

**1. Data Preparation**

* **Objective**: Ensure the dataset is clean and suitable for analysis.
* **Steps Taken**:
  + Imported necessary libraries such as pandas, numpy, seaborn, and matplotlib.
  + Loaded data from country\_wise\_latest.csv.
  + Checked for missing values and duplicates.
  + Made a copy of the dataset for transformations.

**2. Exploratory Data Analysis (EDA)**

* **Goal**: Uncover patterns, correlations, and trends.
* **Insights**:
  + Analyzed the distribution of cases and deaths across regions.
  + Generated visualizations like correlation heatmaps and boxplots to identify relationships and outliers.
  + Key observations highlighted variations in mortality and recovery rates across WHO regions.

**3. Model Development**

* **Approach**:
  + Time-Series Analysis: Prepared data for trend forecasting.
  + Classification: Trained logistic regression models to classify countries based on severity.
* **Implementation**:
  + Time-series stationarity tests were conducted, confirming the feasibility of forecasting.
  + Feature engineering included scaling, transformations, and encoding.

**4. Model Evaluation**

* **Metrics**:
  + For time-series models: RMSE and trend accuracy.
  + For classification models: Accuracy, Precision, Recall, F1-score.
* **Results**:
  + High model performance for classification, with confusion matrix indicating no misclassifications.
  + Time-series predictions showed reliable trends over a validation set.

**5. Key Insights**

* **Factors Influencing Spread**:
  + Strong correlations were found between population density and case rates.
  + Higher recovery rates were associated with regions with robust healthcare infrastructure.
* **Outliers**:
  + Certain countries exhibited unusually high or low mortality rates due to unique interventions.

**6. Data Visualization**

* **EDA Visualizations**:
  + Heatmaps and scatterplots provided insights into feature relationships.
  + Line plots illustrated trends in cases and recoveries over time.
* **Model Visualizations**:
  + Forecasting visualizations highlighted predicted trends.
  + Classification models displayed decision boundaries for severity classification.

**7. Reporting**

* **Summary**:
  + The report consolidates EDA, modeling efforts, and actionable insights.
* **Recommendations**:
  + Target high-risk regions with increased medical supplies.
  + Leverage recovery rate trends to replicate successful interventions.

Appendix  
fig1: Correlation matrix

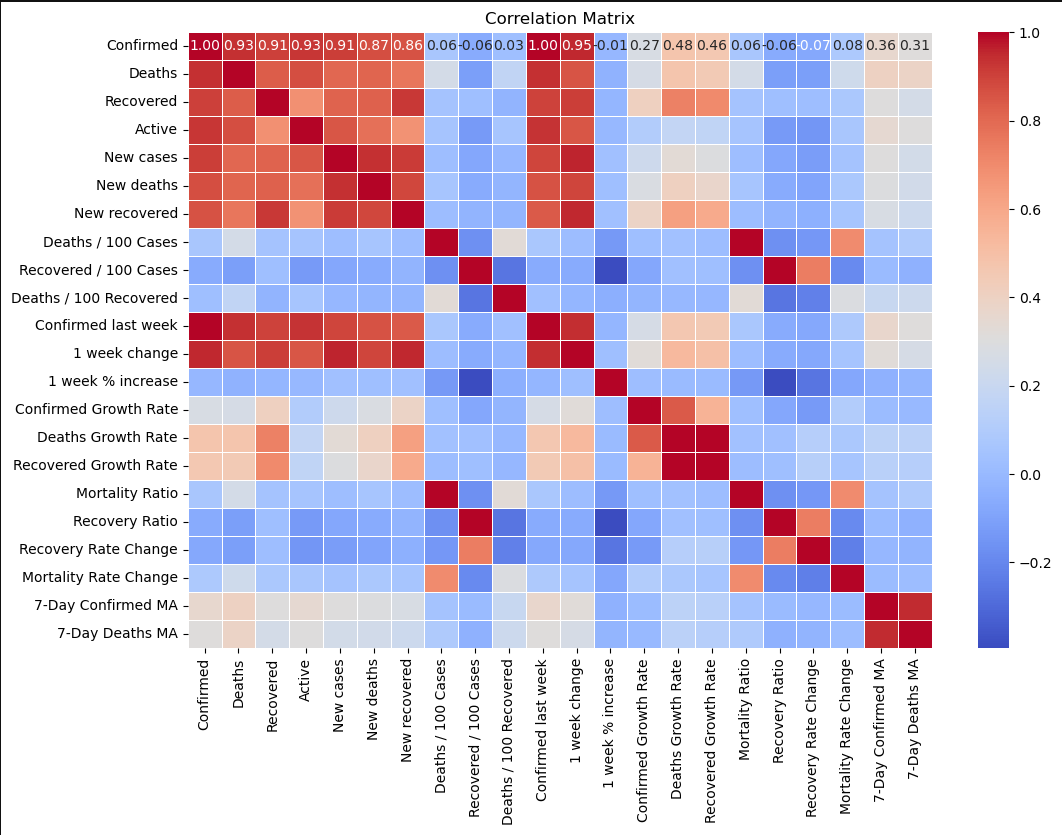


Fig2: Histogram plots for key variables to see their distribution

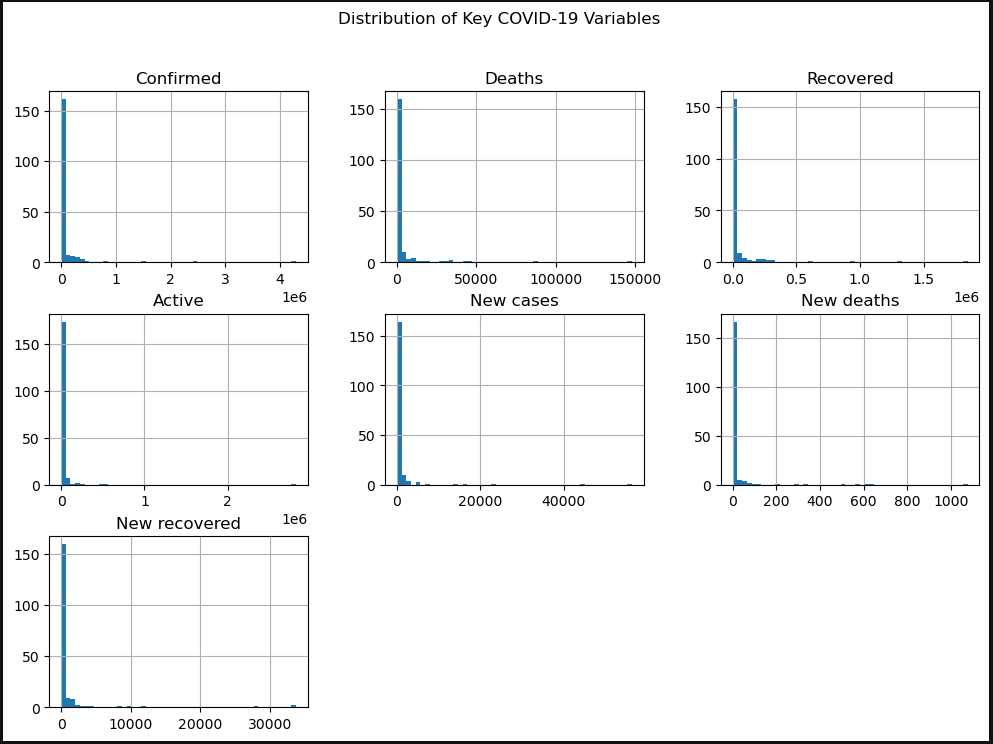


Fig3: Fixing outliers

A screen shot of a graph

Description automatically generated

Fig4: Visualizing pairwise relationships between key variables using pairplot

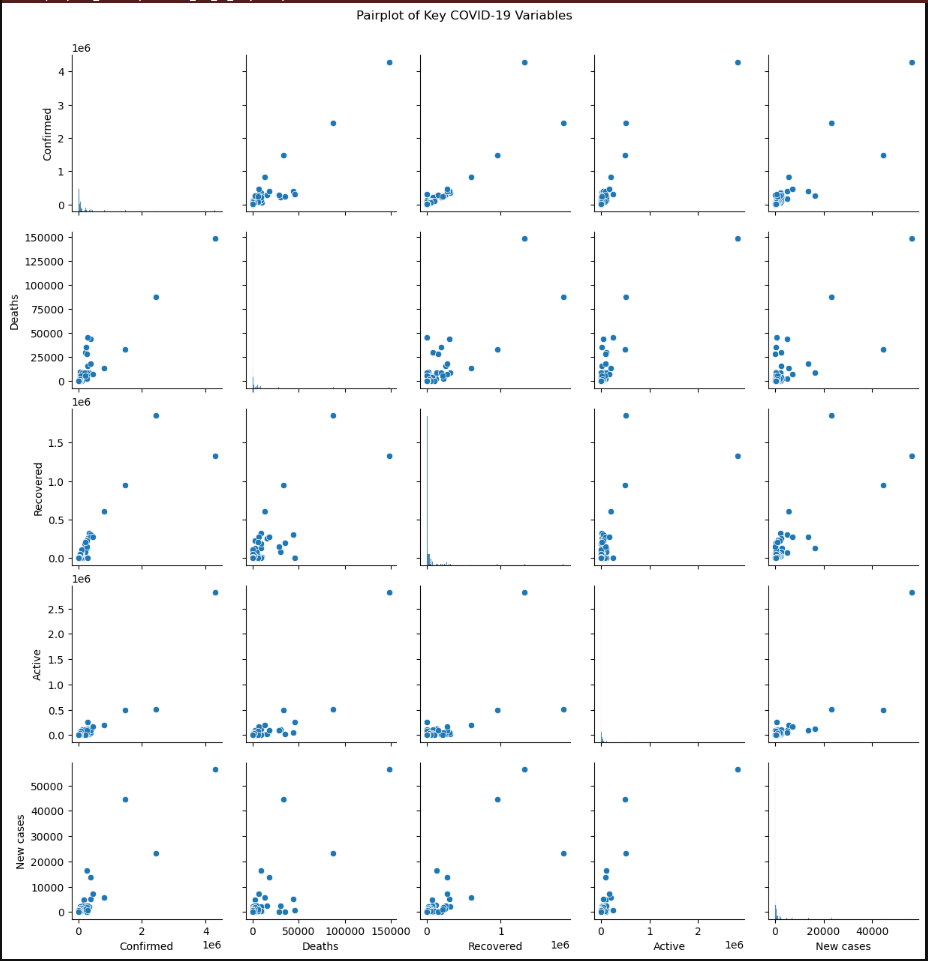


Fig5: Plotting line chart for confirmed cases over synthetic time

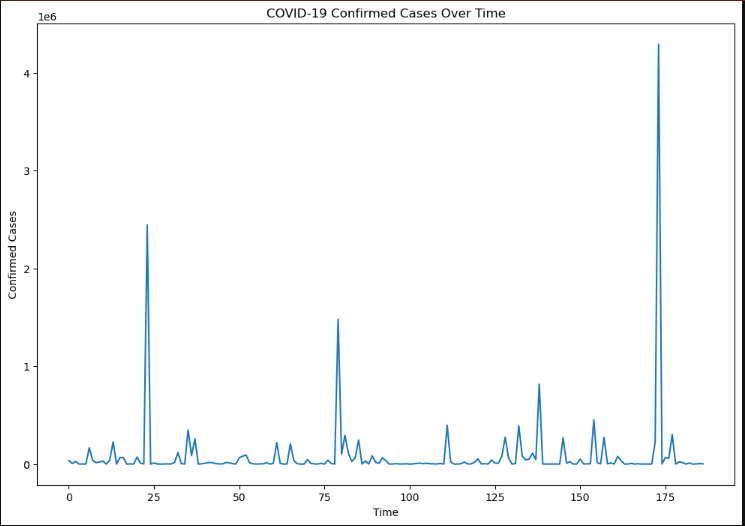


Fig6: Sort by the confirmed cases and display the top 10

A graph of blue bars with black text

Description automatically generated

Fig7: Mortality Rate vs. Recovery Rate

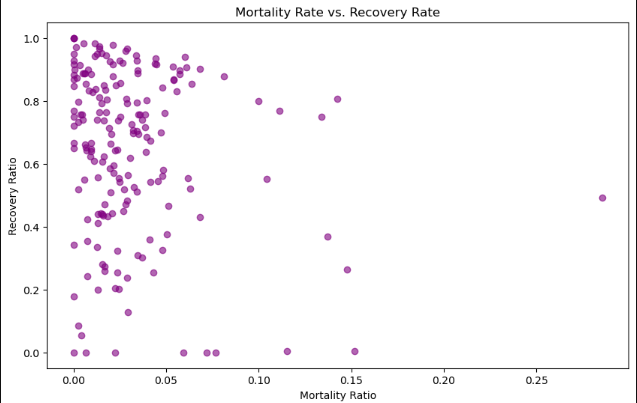


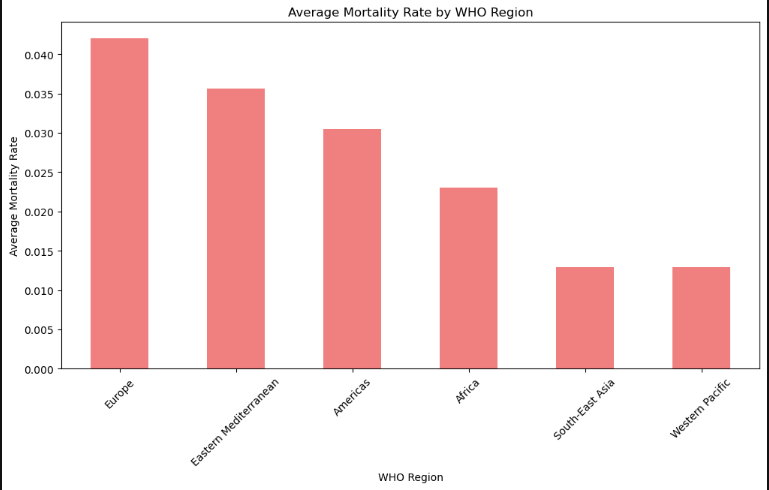
Fig8: Average Mortality Rate by WHO Region  


Fig9: COVID-19 Case Distribution (Top 10 Countries)

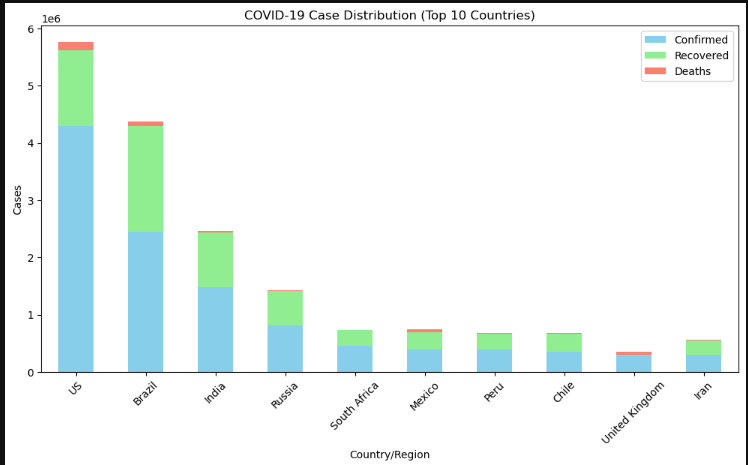


Fig10: Mortality Ratio by WHO Region, Recovery Ratio by WHO Region

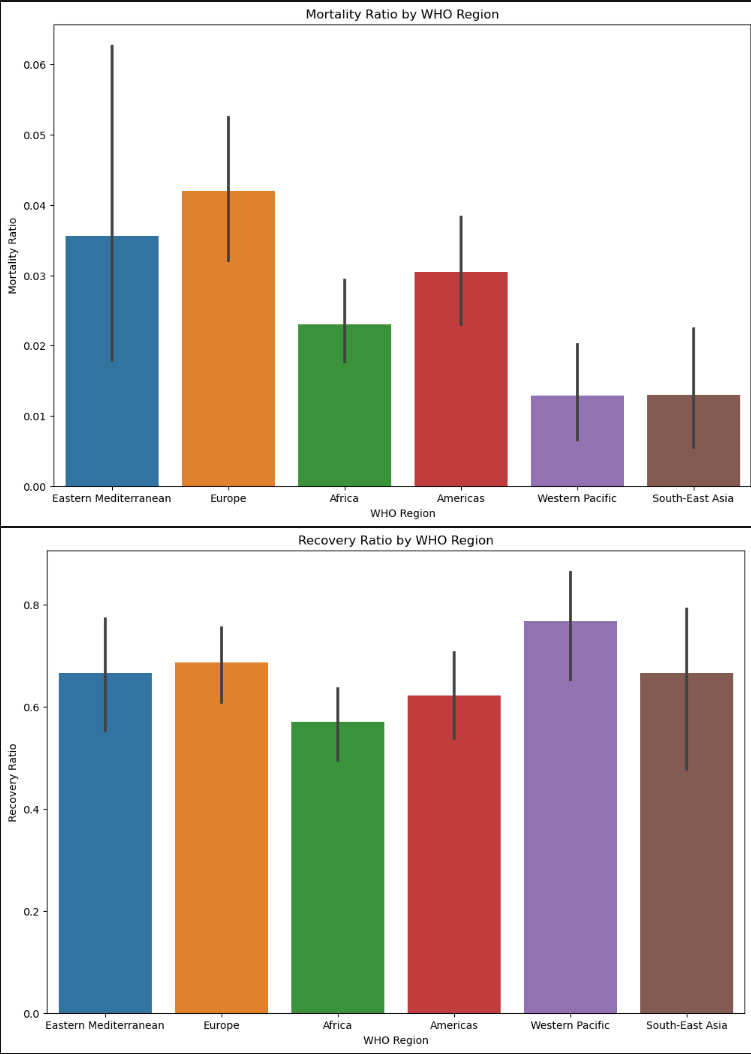


Fig11: Confirmed Growth Rate by Region Over Time

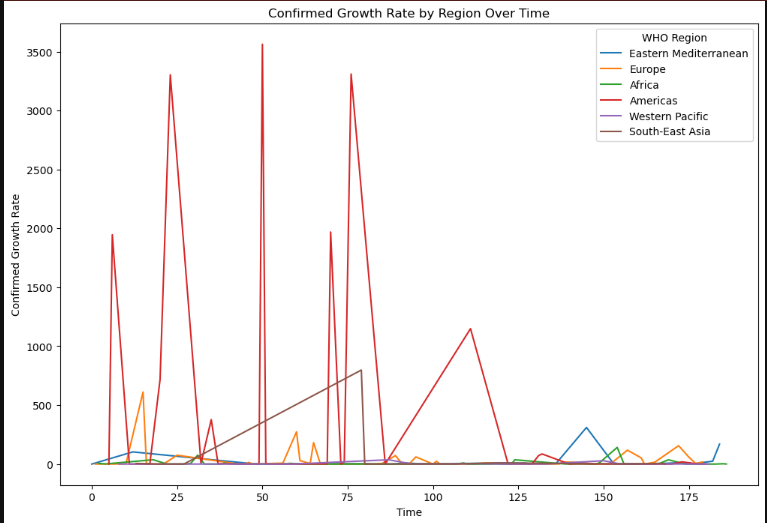


Fig12: Active Cases vs Deaths

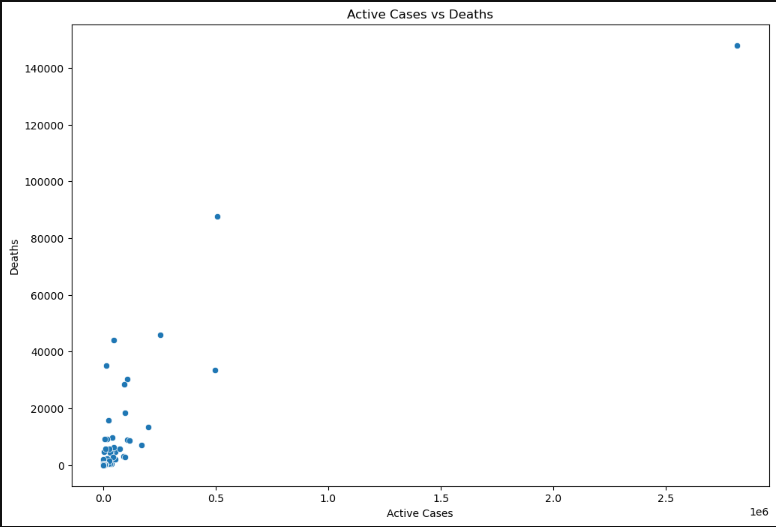


Fig13: COVID-19 Confirmed Cases Forecast

