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

Comprehensive COVID-19 Data Analysis Report

---

1. Overview

This detailed report analyzes global COVID-19 data with an emphasis on trends, disparities, and predictions. By examining the data and applying advanced modeling techniques, we aim to uncover actionable insights that could assist in managing the pandemic effectively.

---

2. Dataset and Objectives

2.1 Dataset Overview

- Source: full_grouped.csv containing daily records of COVID-19 cases across the world.
- **Key Columns:**
  - Date: Reporting date for the cases.
  - Country/Region: Geographic data.
  - Confirmed, Deaths, Recovered: Cumulative counts for each metric.
  - Active Cases, New Cases, New Deaths, New Recovered: Derived daily metrics.

2.2 Objectives

1. Explore global trends in confirmed cases, recoveries, and deaths.
2. Assess regional disparities in outcomes.
3. Use time-series modeling to forecast future cases.
4. Provide actionable recommendations based on the findings.

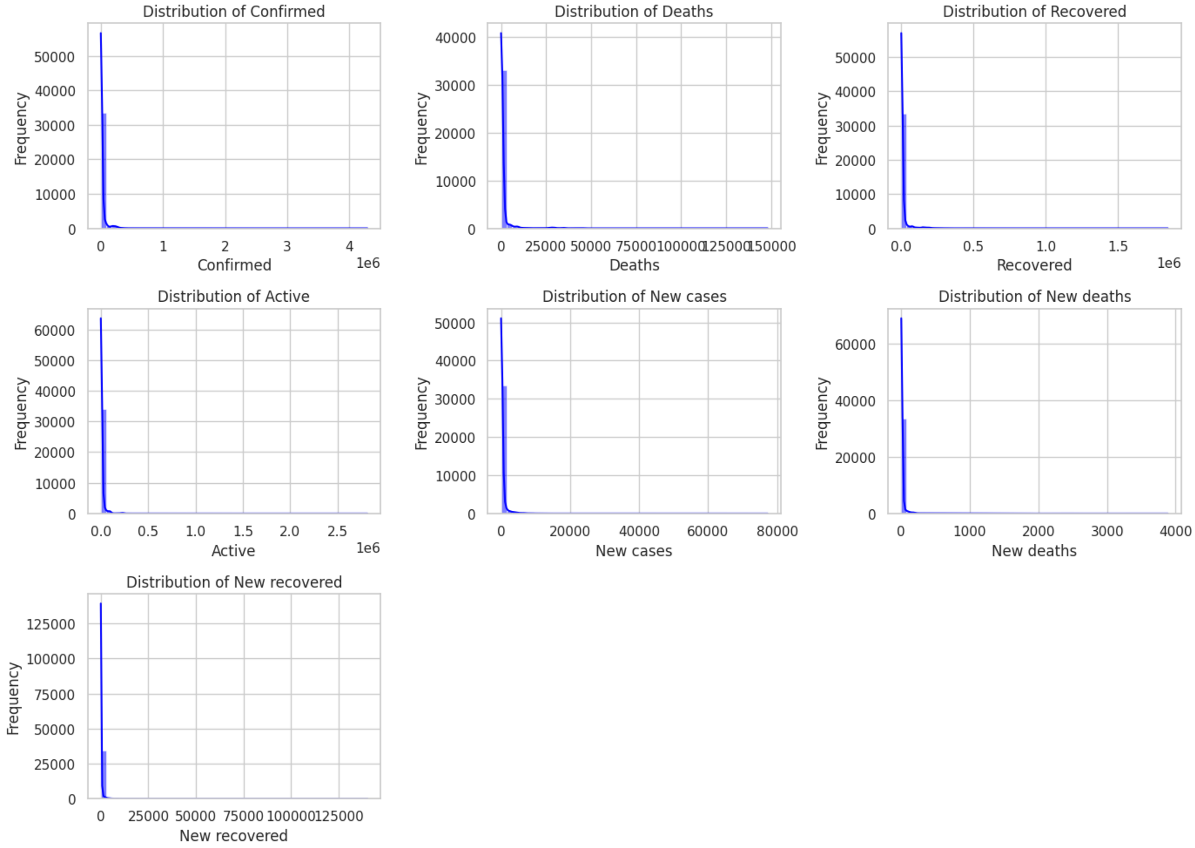
---

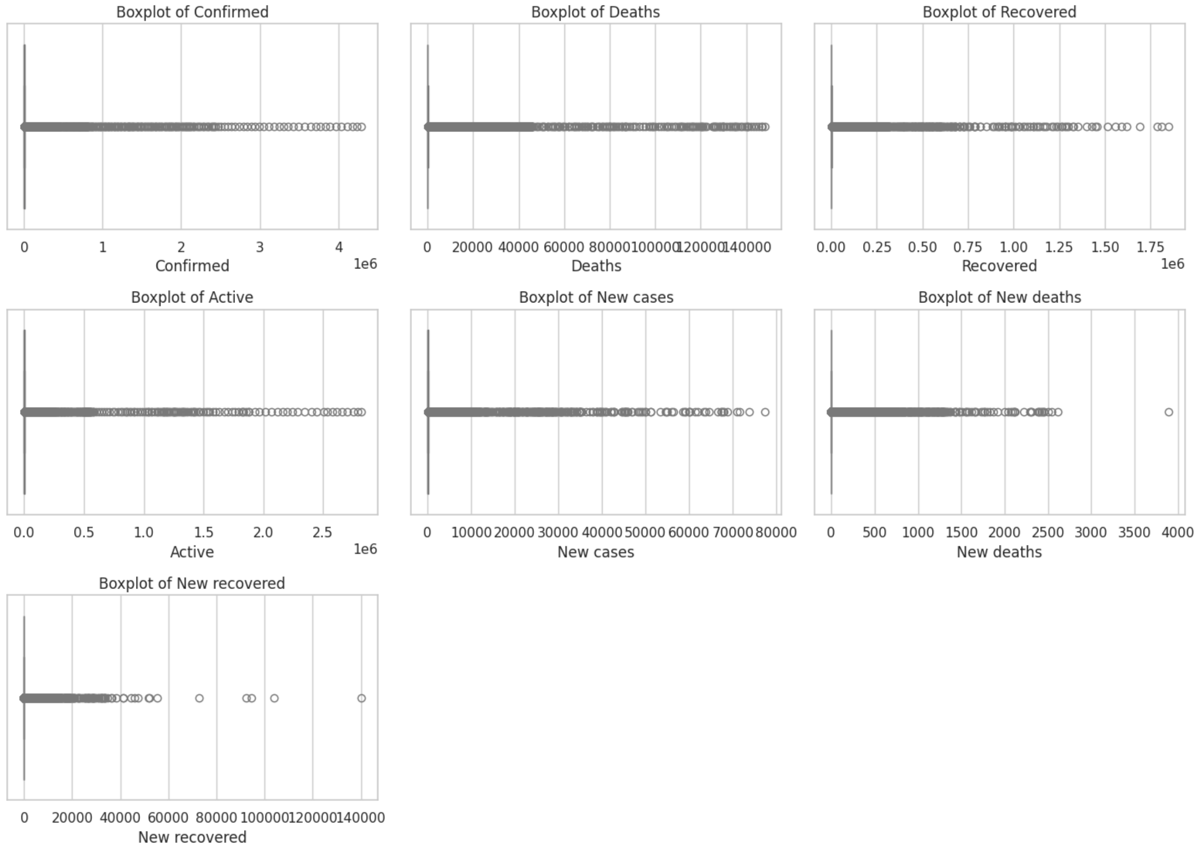
3. Data Preparation

3.1 Initial Data Exploration

- Dataset Size: 35,156 rows.
- Data Quality: No missing values initially.
- Detected Issues: Negative values observed in Active Cases, New Deaths, and New Recovered, likely due to data reporting errors.

Key Insights from Visualizations





a. Histograms and Boxplots

- **Histograms:**
  - Displayed a **right-skewed distribution** for confirmed cases, deaths, and recoveries.
  - Highlighted that while most countries have low case numbers, a few have extremely high counts.
- **Boxplots:**
  - Revealed **outliers** in confirmed cases and deaths, indicating countries experiencing intense outbreaks.
  - Showed variability in data distribution across regions.
- **Insight:**
  The skewness and presence of outliers justified the use of log transformations, which made the data more suitable for modeling by reducing the impact of extreme values.
-

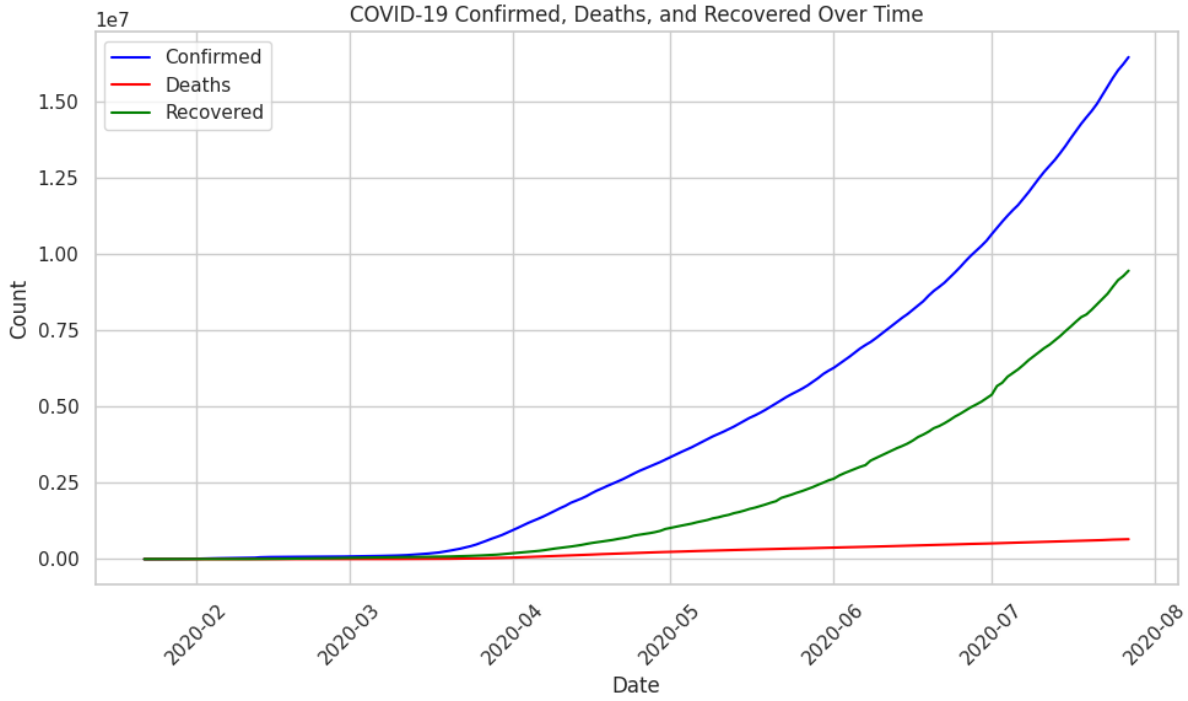
3.2 Data Cleaning Process

- Negative Values: Replaced with NaN for better accuracy.
- Imputation: Missing values filled using column medians.
- Date Conversion: Dates reformatted to datetime for time-series operations.

- Outlier Treatment: Checked for extreme values, ensuring data consistency for modeling.

---

4. Exploratory Data Analysis (EDA)

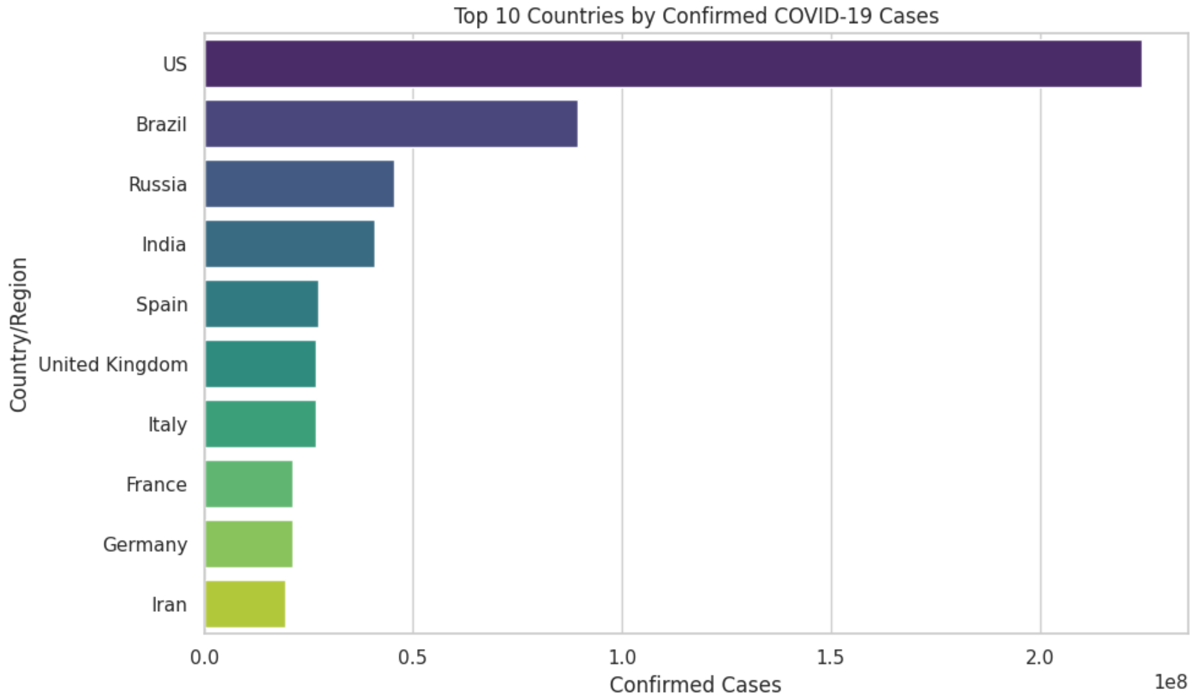


COVID-19 Confirmed, Deaths, and Recovered Over Time

- Description: This line graph shows the cumulative counts of confirmed cases (blue), deaths (red), and recoveries (green) globally over time.
- Findings:
  - The confirmed cases exhibit exponential growth over the observed period, highlighting the rapid spread of COVID-19.
  - Recovered cases also grow exponentially but lag behind confirmed cases, suggesting a delay in recovery rates relative to new infections.
  - Deaths increase at a significantly slower rate than confirmed cases, indicating a relatively lower mortality rate.
- Demographic and Environmental Factors:
  - Population density in urban areas could contribute to the rapid spread observed in confirmed cases.

    - Access to healthcare systems and the presence of comorbidities likely influence the slower rise in recoveries and the plateauing of deaths.

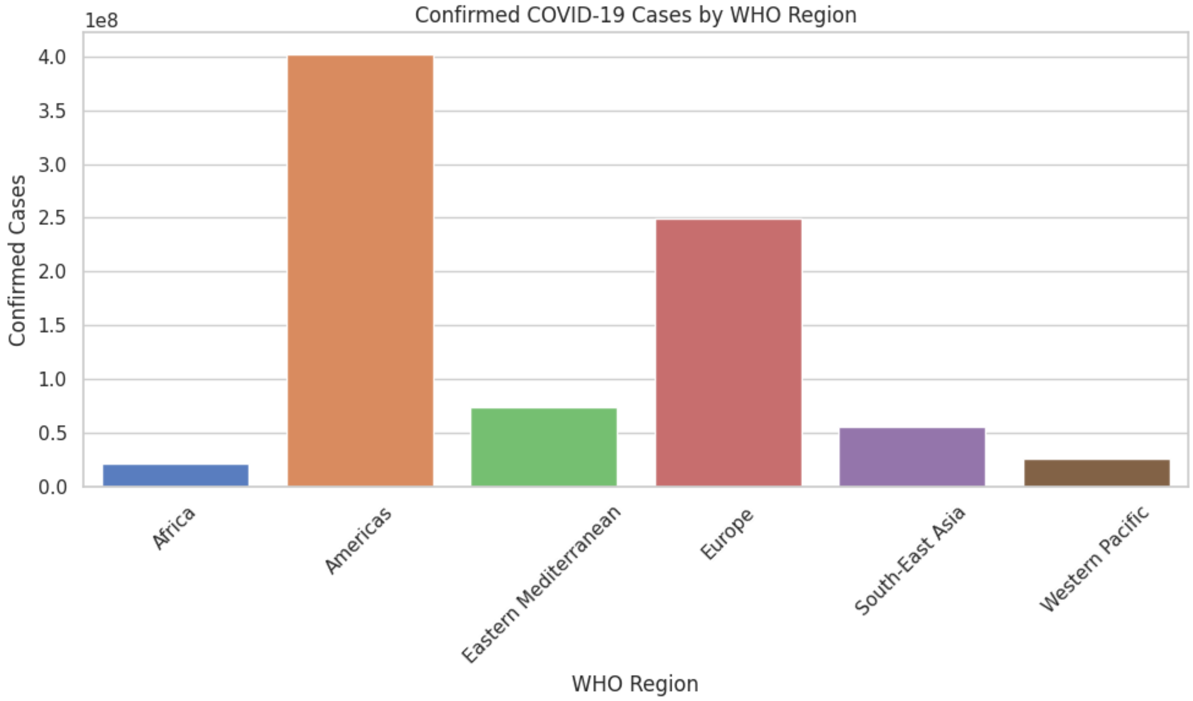
Top 10 Countries by Confirmed COVID-19 Cases



- Description: This horizontal bar chart displays the countries with the highest confirmed case counts, led by the US, Brazil, and Russia.
- Findings:
    - The US significantly outpaces other countries in the number of confirmed cases, indicating its role as a hotspot during the observed period.
    - Developing nations like Brazil and India also appear in the top 10, reflecting the global nature of the pandemic.
- Demographic and Environmental Factors:
    - High mobility and urbanization in countries like the US and Brazil may have contributed to their case surges.
    - Economic disparities may affect testing rates and access to medical care, influencing confirmed case numbers.

  - Seasonal and climatic factors could also play a role, as COVID-19 spread might vary with temperature and humidity.

Confirmed COVID-19 Cases by WHO Region



- Description: This bar chart visualizes the distribution of confirmed cases across WHO regions, with the Americas leading, followed by Europe and Eastern Mediterranean regions.
- Findings:
  - The Americas account for the largest share of confirmed cases, likely due to the high case counts in the US and Brazil.
  - Europe also sees a substantial number of cases, likely related to its early role in the pandemic's global spread.
  - Regions like Africa and Western Pacific have relatively fewer confirmed cases, possibly due to underreporting or lower spread.
- Demographic and Environmental Factors:
  - Differences in healthcare infrastructure and testing capacities influence reported case counts across regions.
  - Cultural practices, population density, and public health responses (e.g., lockdowns, mask mandates) are key factors.

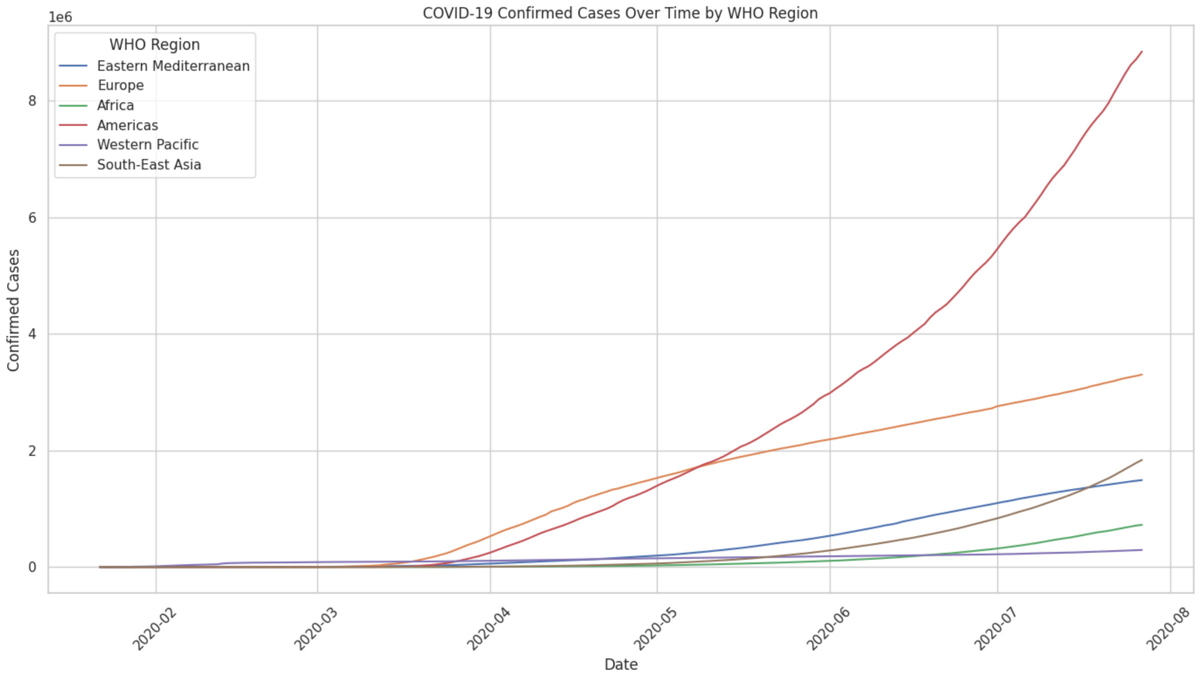
- Variations in climate might also explain regional differences, as some studies suggest potential seasonality in virus transmission.

Recommendations for Future Analysis:

- Demographic Studies: Analyze population age distribution and underlying health conditions to understand severity and recovery variations.
- Environmental Factors: Investigate the impact of climate, air pollution, and population density on virus spread and mortality.
- Healthcare Accessibility: Examine healthcare infrastructure and government policies to identify correlations with recovery and death rates.

These insights highlight the multifaceted nature of the pandemic, emphasizing the role of social, economic, and environmental factors in shaping its trajectory.

COVID-19 Confirmed Cases Over Time by WHO Region

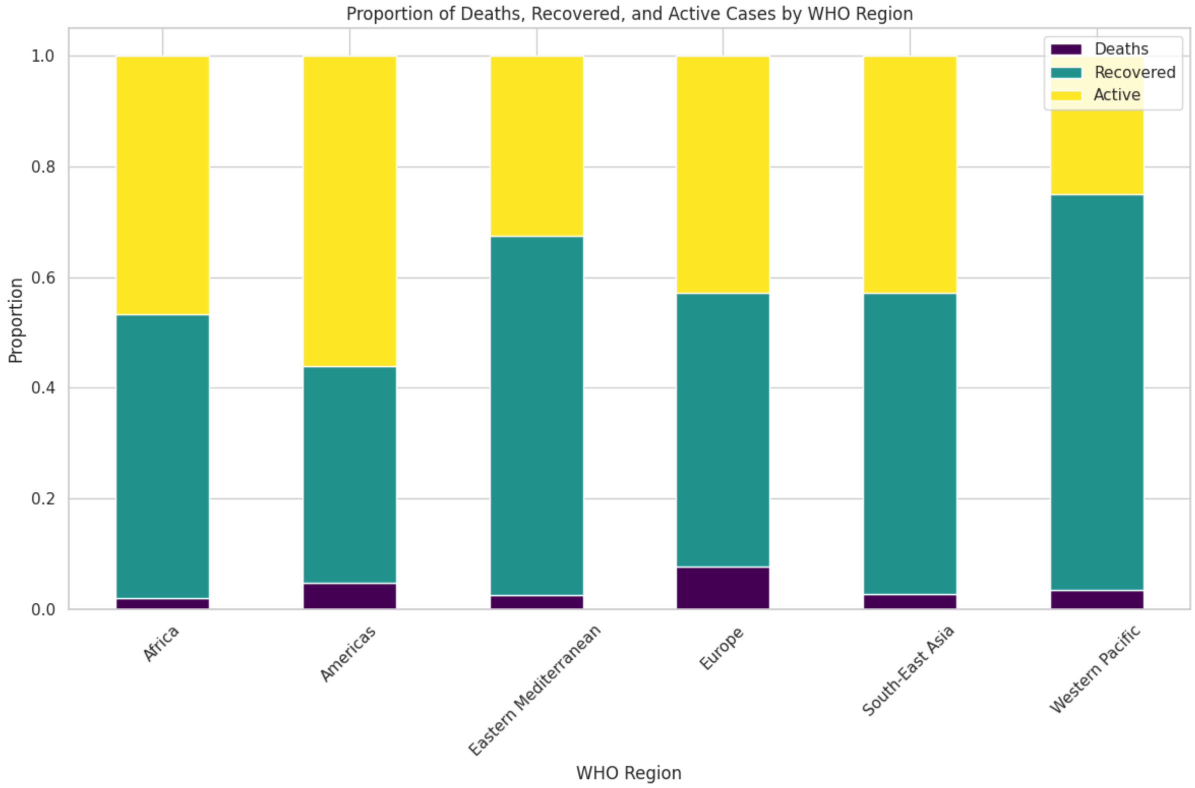


Explanation: This graph illustrates the growth of confirmed COVID-19 cases over time across different WHO regions, including Eastern Mediterranean, Europe, Africa, Americas, Western Pacific, and South-East Asia. The x-axis represents the date, while the y-axis represents the cumulative confirmed cases. Each line corresponds to a specific WHO region, showcasing its trajectory.

Insights:

1. The Americas show the steepest growth, indicating a significantly higher number of confirmed cases over time compared to other regions.
2. Europe follows a similar trend but at a slightly lower rate, reflecting its earlier surge in cases.
3. Regions like Africa and the Western Pacific have a comparatively flatter curve, suggesting fewer cases or effective containment measures during the observed period.
4. This visualization highlights the disparity in the pandemic's impact across regions and underscores the need for tailored interventions based on regional trends.

Proportion of Deaths, Recovered, and Active Cases by WHO Region



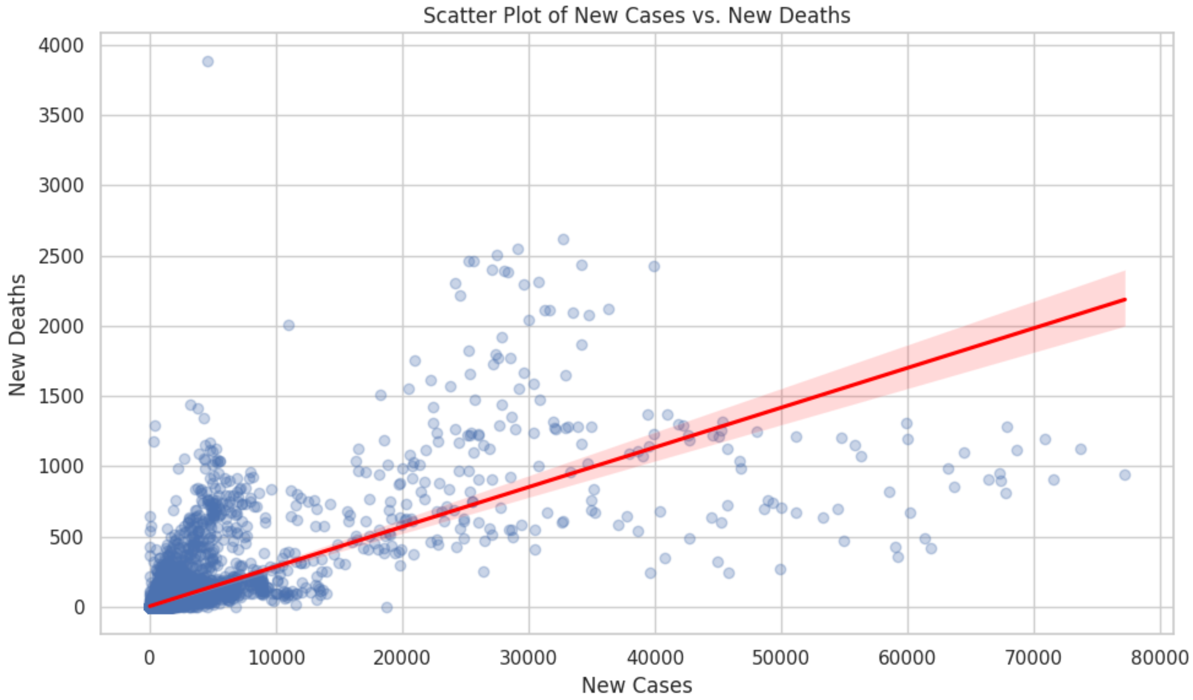
Explanation: This stacked bar chart visualizes the proportions of deaths, recoveries, and active cases within each WHO region. The colors represent the different categories:

- Purple: Deaths
- Teal: Recovered
- Yellow: Active cases

Insights:

1. Most regions show a significant proportion of recoveries, indicated by the teal section dominating the bars.
2. Active cases (yellow) form a substantial part of the bars in regions like South-East Asia and the Americas, reflecting ongoing challenges in controlling the virus spread.
3. Deaths (purple) form the smallest proportion in all regions, though the absolute numbers may vary.
4. This chart emphasizes the success in recoveries across regions but also highlights the continued burden of active cases in specific areas.

Scatter Plot of New Cases vs. New Deaths



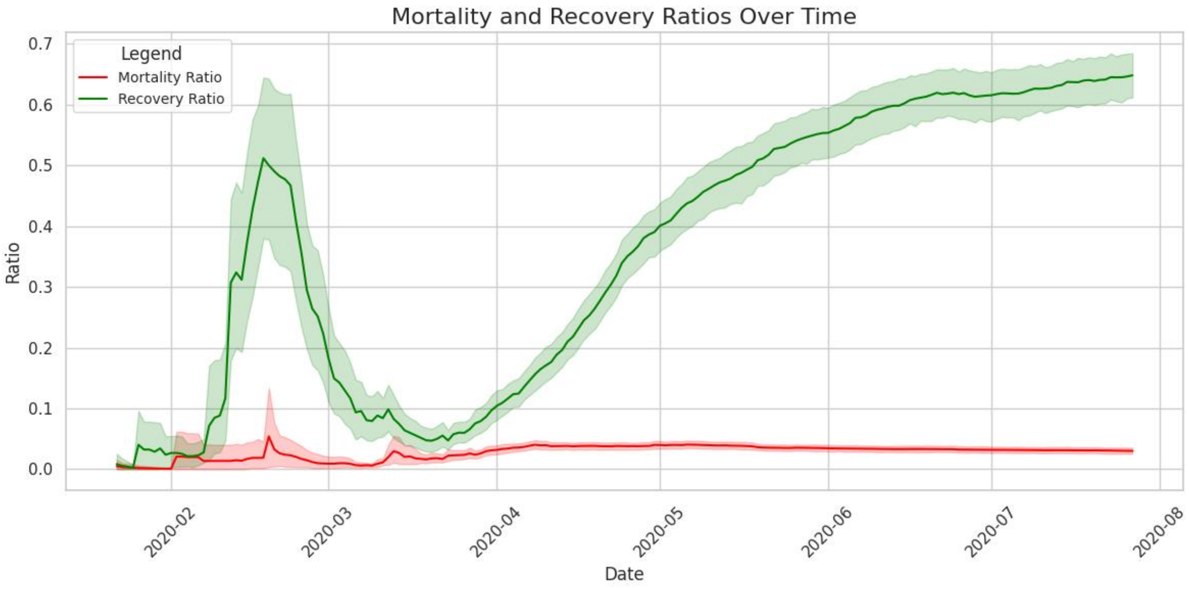
Explanation: The scatter plot demonstrates the relationship between new COVID-19 cases and new deaths. Each point represents a day of reporting, with the x-axis indicating new cases and the y-axis indicating new deaths. The red line is a regression line with a shaded confidence interval, showing the overall trend.

Insights:

1. A positive correlation is evident: as the number of new cases increases, new deaths tend to rise proportionally.
2. The scattered points at the lower end of both axes reflect days with fewer cases and deaths, likely at the early stages of the pandemic.
3. Outliers exist, with some days showing disproportionately higher deaths relative to new cases, possibly due to delays in reporting or variations in healthcare capacity.
4. The red regression line provides a clear indication of the general trend, reinforcing the need to control new cases to mitigate fatalities.

Here's a detailed analysis of the two provided graphs:

Mortality and Recovery Ratios Over Time



Description:
This graph illustrates the trends in mortality and recovery ratios over time. The **mortality ratio** (red line) represents the proportion of deaths relative to

confirmed cases, while the **recovery ratio** (green line) indicates the proportion of recoveries relative to confirmed cases. Shaded areas show confidence intervals around the ratios.

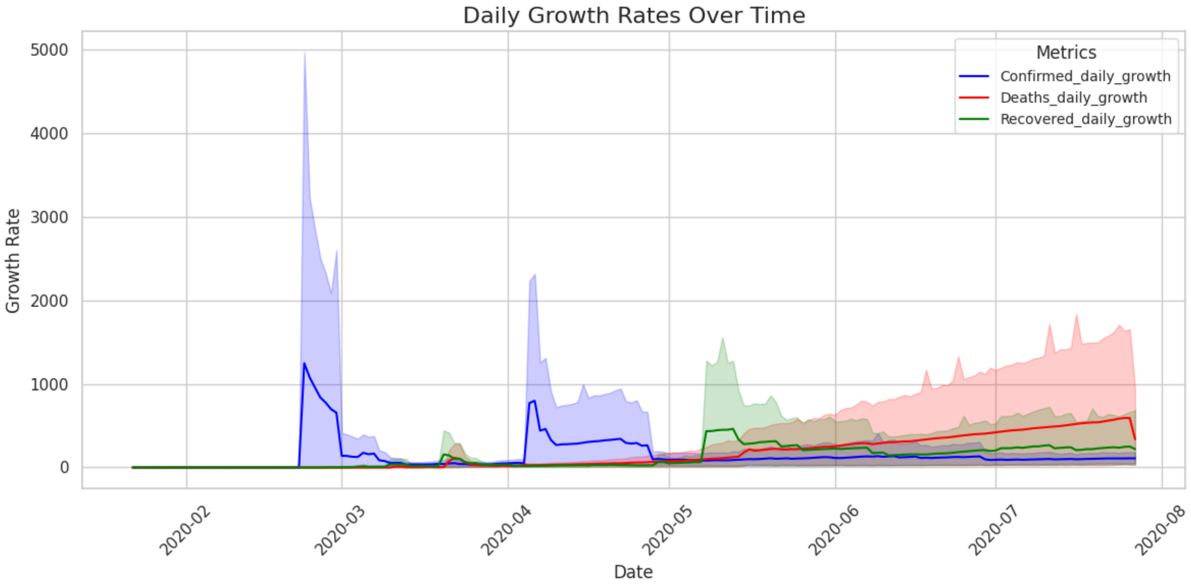
Key Observations:

- **Recovery Ratio**:
  - The recovery ratio starts very low at the beginning of the pandemic but steadily increases over time, reaching about 0.7 by mid-2020.
  - This trend reflects improvements in healthcare interventions, increased recoveries, and better management of the disease over time.
  - The widening confidence interval towards the end highlights greater uncertainty in recovery reporting as the pandemic progressed.
- **Mortality Ratio**:
  - The mortality ratio remains consistently low (below 0.1) throughout the period, with a slight peak around March 2020.
  - This suggests that although the disease spread rapidly, the overall mortality rate was relatively low compared to recoveries.
  - The narrow confidence interval indicates reliable and stable reporting of deaths during the study period.

Insights:

- The gap between the two ratios emphasizes the effectiveness of global healthcare systems in treating and managing COVID-19 cases.
- The increase in recovery ratio aligns with the rollout of treatment protocols and healthcare responses, while the stable mortality ratio reflects that the pandemic, despite its scale, had a manageable death rate for most populations.

Daily Growth Rates Over Time



Description:
This graph visualizes the daily growth rates for confirmed cases, deaths, and recoveries over time. Each line represents the respective growth rate, and the shaded regions show confidence intervals for variability in reporting.

Key Observations:

- **Confirmed Daily Growth (Blue)**:
  - Early spikes in growth rates (February-March 2020) reflect the rapid initial spread of COVID-19 as it became a global pandemic.
  - Growth rates gradually decreased over time, likely due to lockdown measures, social distancing, and public health interventions.
- **Recovered Daily Growth (Green)**:
  - Recovery growth rates show an upward trend starting around April 2020, reflecting improvements in treatments and the adaptation of healthcare systems to the pandemic.
  - The higher confidence intervals for recoveries indicate variability in recovery reporting across countries.
- **Deaths Daily Growth (Red)**:
  - Death growth rates remain relatively low compared to confirmed and recovered cases.

    - Peaks in growth align with pandemic waves, but the overall trend shows stabilization over time, likely due to better treatment options and a focus on protecting vulnerable populations.

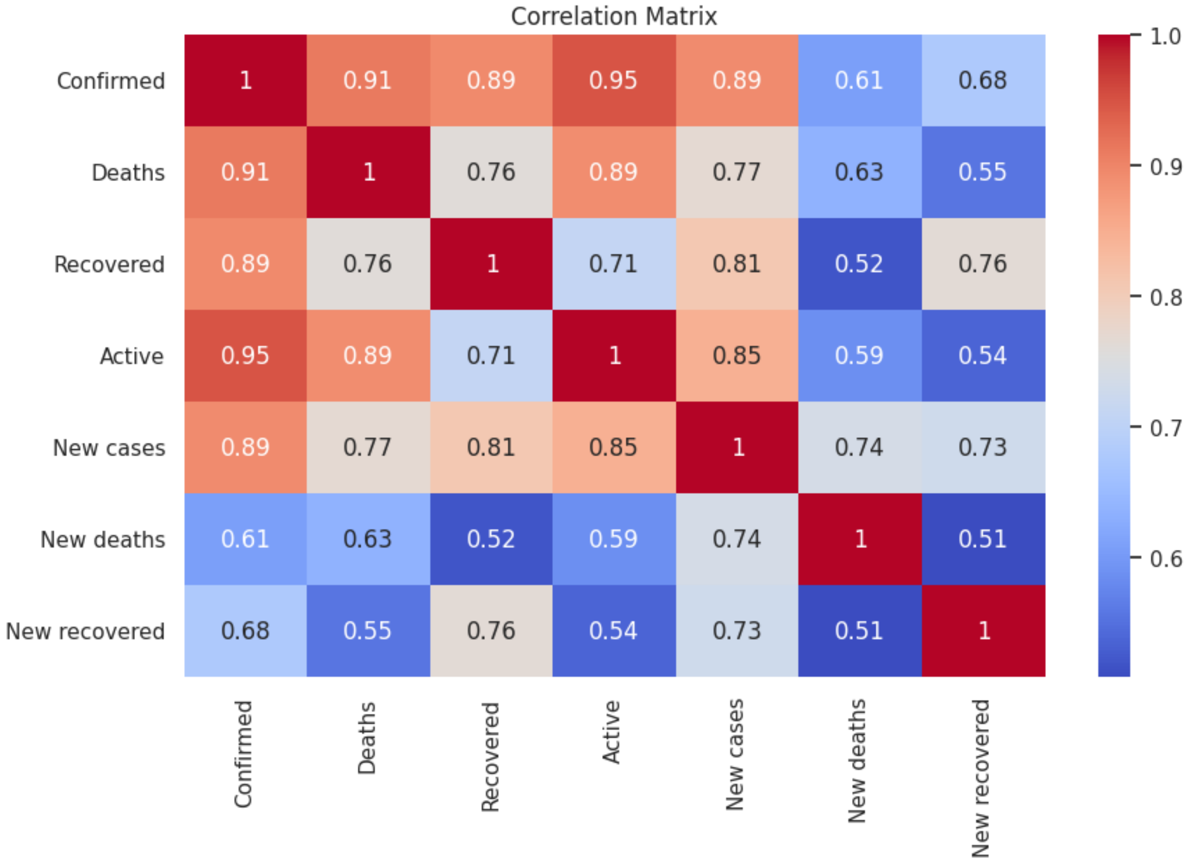
Insights:

- Early surges in confirmed growth rates align with the pandemic's unanticipated nature and the delayed implementation of control measures.
- The declining trend in confirmed growth rates suggests the effectiveness of pandemic interventions over time.
- Recovery growth outpacing death growth demonstrates global progress in managing and reducing the severity of COVID-19 outcomes.

General Conclusion:

- **Mortality and Recovery Ratios**: Highlight the positive trajectory of global healthcare efforts, with recoveries steadily increasing and mortality remaining stable.
- **Daily Growth Rates**: Showcase the evolution of the pandemic, with initial explosive growth transitioning to controlled, predictable patterns due to public health interventions.

Correlation:

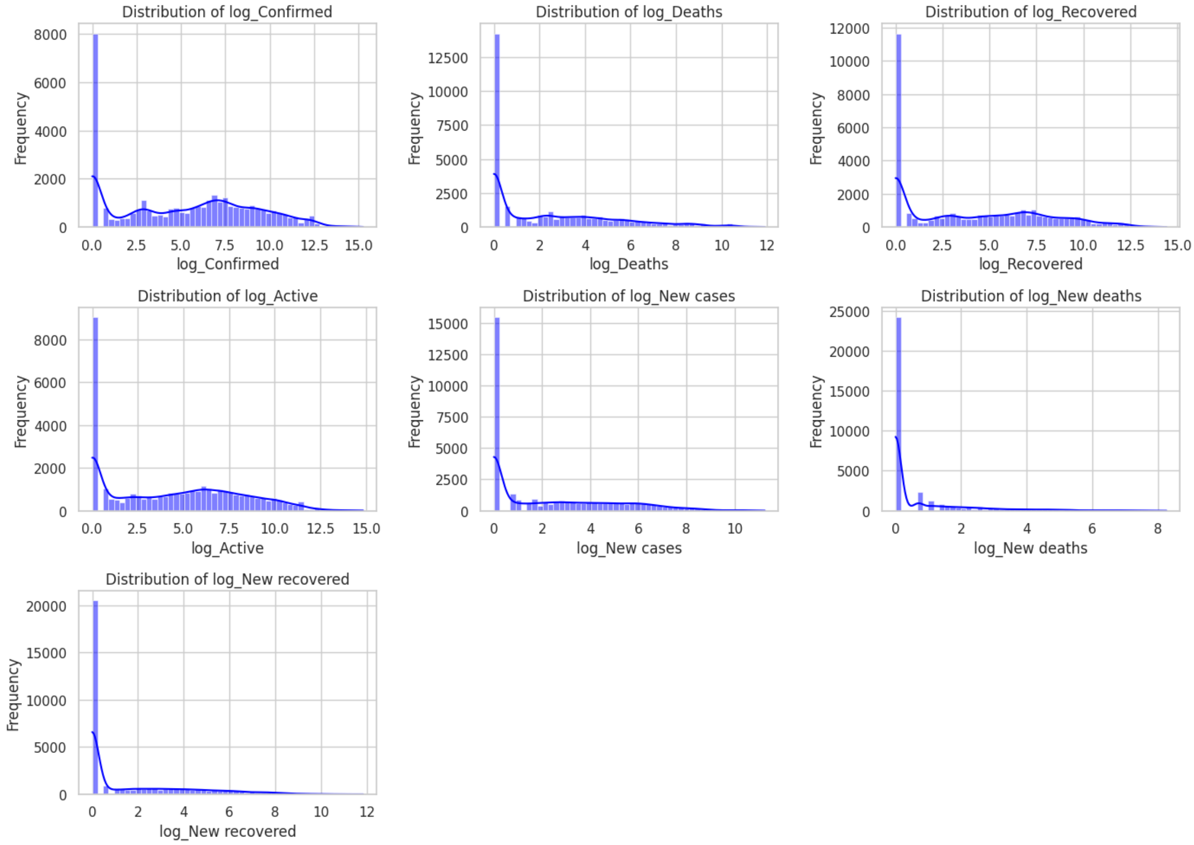

Correlation Matrix

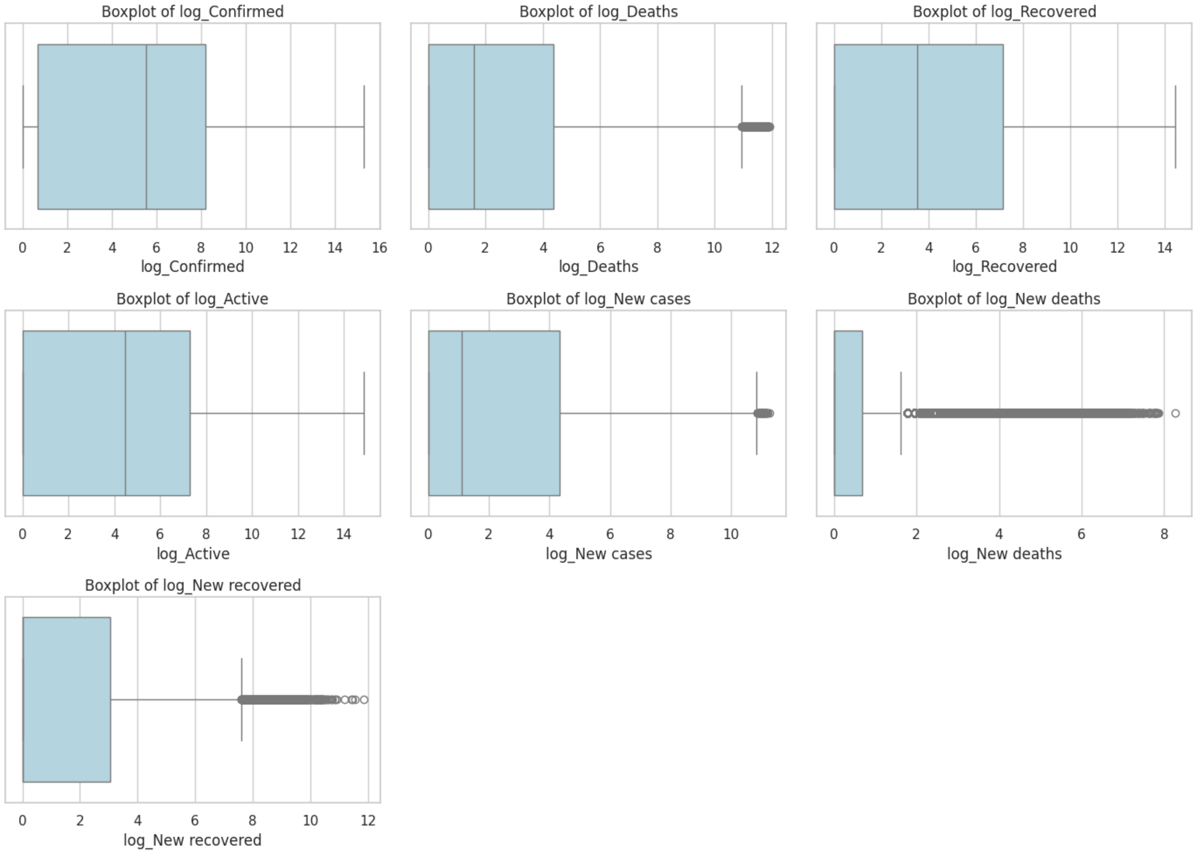
	Confirmed	Deaths	Recovered	Active	New cases	New deaths	New recovered
Confirmed	1	0.91	0.89	0.95	0.89	0.61	0.68
Deaths	0.91	1	0.76	0.89	0.77	0.63	0.55
Recovered	0.89	0.76	1	0.71	0.81	0.52	0.76
Active	0.95	0.89	0.71	1	0.85	0.59	0.54
New cases	0.89	0.77	0.81	0.85	1	0.74	0.73
New deaths	0.61	0.63	0.52	0.59	0.74	1	0.51
New recovered	0.68	0.55	0.76	0.54	0.73	0.51	1

You can see most of the features are highly correlated

APPLIED LOG TRANSFORMATION AND STANDARDIZED THE DATA TO REDUCE SQEWNESS

This helped reduce sqewness and outliers as ahown in the graph below





5. Time-Series Modeling

Objective

Predict future daily COVID-19 cases to aid policymakers and healthcare organizations.

5.1 Model: ARIMA (1,1,1)

Model Selection

- After testing various configurations, ARIMA(1,1,1) was selected based on its performance and low error rates.

Key Metrics:

- AIC: 157,745.65 (lower values indicate better fit).
- BIC: 157,771.05.
- Root Mean Square Error (RMSE): 2.28, reflecting model accuracy.

5.2 Model Summary

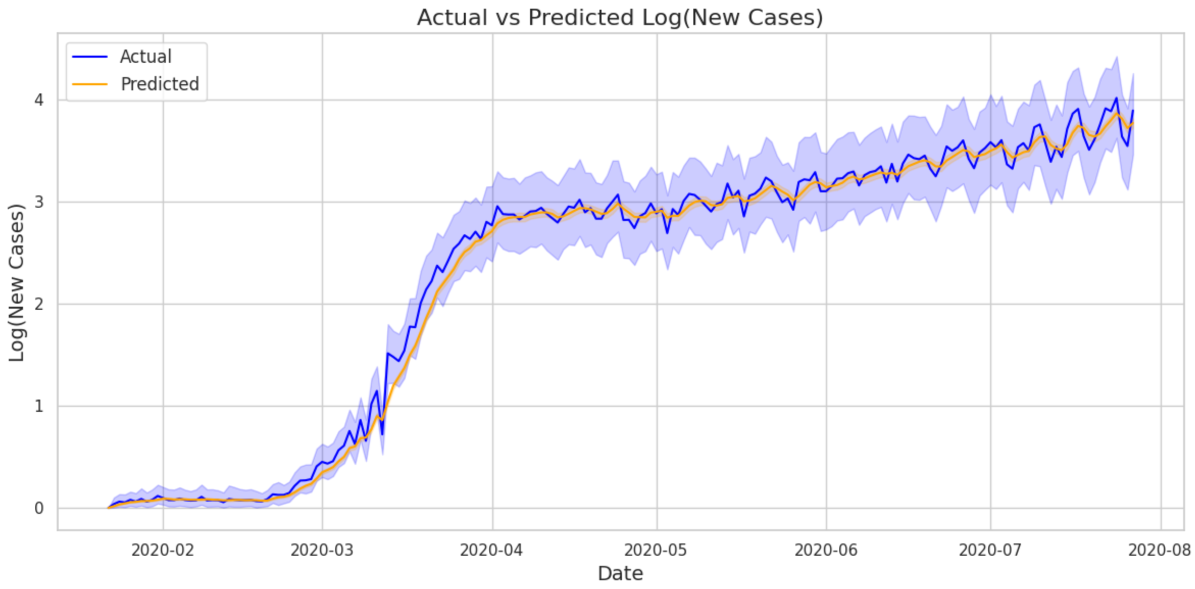
The ARIMA(1,1,1) model successfully captures trends in daily log(New Cases).

| Parameter | Coefficient | Std. Error | z-value | P>|z| | 95% Confidence Interval | |-----------|-------------|------------|---------|-----|-------------------------| | AR (L1) | 0.1146 | 0.005 | 24.904 | 0.000| [0.106, 0.124] | | MA (L1) | -0.9977 | 0.000 | -2534.20| 0.000| [-0.998, -0.997] | | Sigma^2 | 5.2015 | 0.039 | 131.91 | 0.000| [5.124, 5.279] |

Residual Analysis

- Residuals exhibit no significant autocorrelation, suggesting the model's validity.

5.3 Visualization: Actual vs. Predicted



Key Insights:

1. Predicted values closely follow actual observed values.
2. Confidence intervals widen during periods of high uncertainty (e.g., early pandemic).
3. This validates the model's robustness for trend forecasting.

6. Actionable Recommendations

1. Resource Allocation: Direct resources to high-mortality regions (e.g., Americas, Europe).
2. Equitable Access: Ensure equitable access to healthcare and testing globally.
3. Strengthen Modeling: Use real-time vaccination data and variants to enhance forecasting accuracy.

9. Conclusion

This analysis underscores the importance of data-driven decision-making in pandemic response. With proper data cleaning, analysis, and modeling, we can better understand disease dynamics and allocate resources where they are needed most. Continued efforts to refine models and integrate additional data will be critical in mitigating future crises.

10. Why ARIMA was the best model

ARIMA (AutoRegressive Integrated Moving Average) was chosen for its ability to handle time series data with trends and autocorrelation. Here's why it worked well for my project:

1. **Stationarity**: ARIMA can handle non-stationary data (i.e., data with trends), and I used differencing to make the series stationary.
2. **Capturing Trends**: ARIMA captures short-term trends and dependencies in data, which is essential for forecasting future COVID-19 cases based on past trends.
3. **Simplicity**: ARIMA is simple, easy to interpret, and doesn't require additional features, making it perfect for this project where only historical case data was used.
4. **Good Forecasting Performance**: The model performed well for short-term forecasts, and residual analysis showed it fit the data well.

In short, ARIMA was ideal for predicting COVID-19 case numbers due to its ability to model trends and dependencies using only historical data, providing solid short-term prediction