Predicting COVID-19 Hospitalization Risks using Data Science Models

COSC 3337: Data Science I

By Naomi Ayub (1868128), Elyjaiah Durden (1977092), and Nirmal John (2052312)

INTRODUCTION:

The COVID-19 pandemic has been one of the most significant global public health crises in modern history, affecting millions of lives, including our own! It also fundamentally altered social, economic, and healthcare systems worldwide. Throughout the pandemic, the collection and analysis of data has played a critical role in informing public health strategies, acting as a guide for policy making, and helping scientists and researchers better understand the nature of this virus. Rigorous data-driven analysis is essential not only for evaluating the past response but also for preparing for future outbreaks.

For this project, we decided to use the original COVID-19 dataset curated by **Our World in Data (OWID)**. This dataset includes daily records from multiple countries, covering a wide range of metrics such as confirmed cases, deaths, hospitalizations, ICU admissions, and vaccination progress. The full dataset provides a rich foundation for comprehensive analysis by capturing both the direct effects of the virus and the varying responses across different regions. The primary goals of this project are as follows:

- **Explore pandemic trends across multiple countries** to uncover patterns and correlations between different COVID-19 metrics.
- **Predict ICU admission needs** using time series forecasting models to assess future healthcare demands.
- **Classify hospitalization risk levels** based on pandemic indicators to support healthcare planning and risk management.

To achieve these goals, we have decided to structure the report into 3 major tasks:

1. **Exploratory Data Analysis (EDA)** to visualize and interpret pandemic trends (Nirmal John).
2. **Time Series Forecasting** to predict ICU requirements based on historical data (Elyjaiah Durden).
3. **Classification Modeling** to categorize hospitalization needs into different risk levels using machine learning techniques (Naomi Ayub).

To carry out our analysis, we utilized a comprehensive array of data science and machine learning tools. These include core Python libraries such as Pandas and NumPy for data manipulation, Matplotlib, Seaborn, and Plotly for advanced visualizations, and Scikit-learn for preprocessing, classification, and clustering tasks. In addition, we leveraged a variety of models, including Decision Trees, K-Nearest Neighbors, Support Vector Machines, Logistic Regression, Random Forests, Gradient Boosting, and LightGBM, as well as unsupervised learning techniques such as K-Means, DBSCAN, and Agglomerative Clustering. Principal Component Analysis (PCA) was used to reduce dimensionality, and metrics such as accuracy, precision, recall, F1-score, and silhouette score are used to assess performance. Additionally, we used Statsmodels tools for forecasting and smoothing. This robust pipeline of tools and methods enables us to perform a thorough, multi-dimensional analysis of the pandemic, allowing for deeper insight and global trends, accurately forecast and assess effective classification of hospitalization risks.

By conducting these analyses, we wish to both deepen our understanding of the pandemic's dynamics and demonstrate how data science methodologies & machine learning models can be applied to critical real-world challenges!

DATA PREPROCESSING:

Before we conduct the analyses, we have to perform data cleaning and preprocess the raw, original owid_covid_data.csv dataset, to ensure that our data is of high quality and consistent. The original dataset contains COVID-19 information, not only for the U.S., but for multiple countries, with daily updates and a massive range of attributes. However, it also has some issues like missing values, redundant columns, and other inconsistencies that need to be accounted for!

We will start by loading the COVID-19 dataset using the pandas and numpy libraries in Python. Pandas is used for data manipulation and exploration, while NumPy is used for numerical operations. Upon observing the dataset, we found that core variables such as total_cases, new_cases, total_deaths, and new_deaths were fully populated and did not require modification. Features with a moderate amount of missing data (ex: icu_patients, hosp_patients, vaccinations) were retained due to their importance in this multilevel analysis. We decided to handle missing entries during modelling, either through filtering or model-specific strategies such as binning. We preserved columns with significant missing data for the purpose of exploratory analysis, but excluded them when necessary. Instead of globally dropping certain rows, we decided to address missing data selectively based on the requirements of each model.

For each task, we decided on a set of features to prioritize. In the case of **exploratory data analysis (EDA)**, we chose the following features to characterize the spread, severity and management of COVID-19: total_cases, new_cases, total_cases, new_deaths, hosp_patients, icu_patients, vaccinations. For **forecasting** the strain on healthcare capacity, we prioritized ICU-related features: icu_patients, icu_patients_per_million, hospitalization_rate, icu_rate. Lastly for our **classification** task, we mainly selected features with strong potential for prediction: total_cases, new_cases, new_deaths, case_fataility_rate, along with demographic and healthcare system indicators such as life_expectancy, hospital_beds_per_thousand, population_density, gdp_per_capita.

We then chose a subset of countries to ensure that there would be both enough data available, and sufficient diversity in public health responses to COVID-19. Countries were chosen based on the completeness of ICU and hospitalization data and to reflect a range of pandemic management strategies. We ended up selecting these countries :

- **New Zealand**: Elimination strategies with strict lockdowns.
- **United States**: Heterogeneous state-level policies.
- **India**: High population with diverse regional outcomes.
- **Sweden**: Minimal intervention strategy.
- **Brazil**: Major outbreaks with limited early restrictions.
- **United Kingdom**: Robust vaccination rollout.
- **South Korea**: Effective test-trace-isolate systems.

By addressing missing data, selecting relevant features, and selecting a diverse set of countries for comparative study, our dataset is now prepared for the analyses to come. The choices made in handling missing values and feature prioritization align with the analytical objectives of forecasting healthcare strain, understanding pandemic severity, and identifying key predictors for classification. Preprocessing of the original dataset sets us up for robust modeling, and in turn, gaining insights into the global impact of the coronavirus.

EXPLORATORY DATA ANALYSIS (EDA):

[1] INITIAL ANALYSIS:

Analyzing the COVID-19 pandemic trends across continents reveals striking variations in both disease spread and mortality outcomes. From 2020 to 2024, Europe and Oceania recorded the highest case rates per million, approaching 330,000 cases per million by 2023, while Africa maintained the lowest reported incidence. (See Figure 1). The visualization data shows that Oceania experienced a particularly dramatic surge in 2022, likely corresponding to the shift from elimination strategies to managing the virus as endemic following successful vaccination campaigns in countries like Australia and New Zealand.

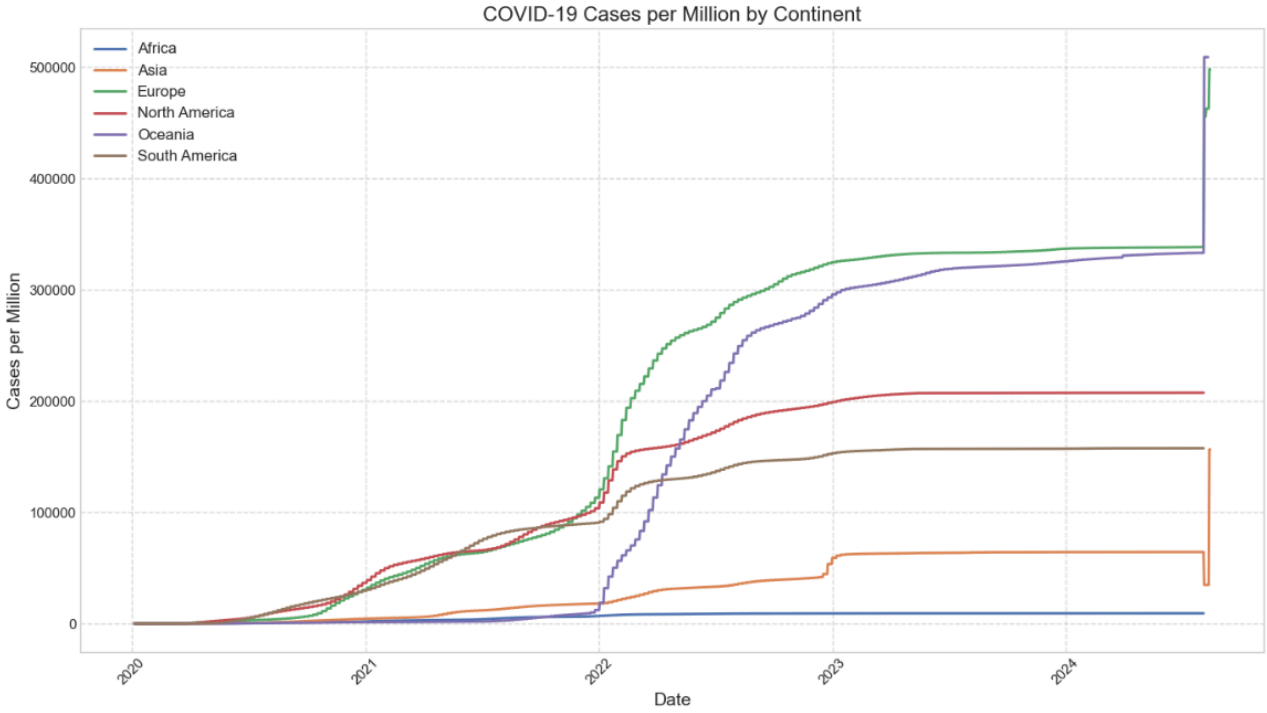


Figure 1 *(Cases per Million by Continent)*

When examining mortality trends, South America consistently maintained the highest death rates throughout most of the pandemic, reaching approximately 3,100 deaths per million by 2024. Europe and North America followed similar trajectories, ending at about 2,800 deaths per million. *(See Figure 2)*. Notably, Oceania demonstrated a substantial gap between its high case rate and relatively lower death rate of approximately 700 per million, suggesting more effective healthcare interventions or potentially different demographic vulnerabilities within its population.

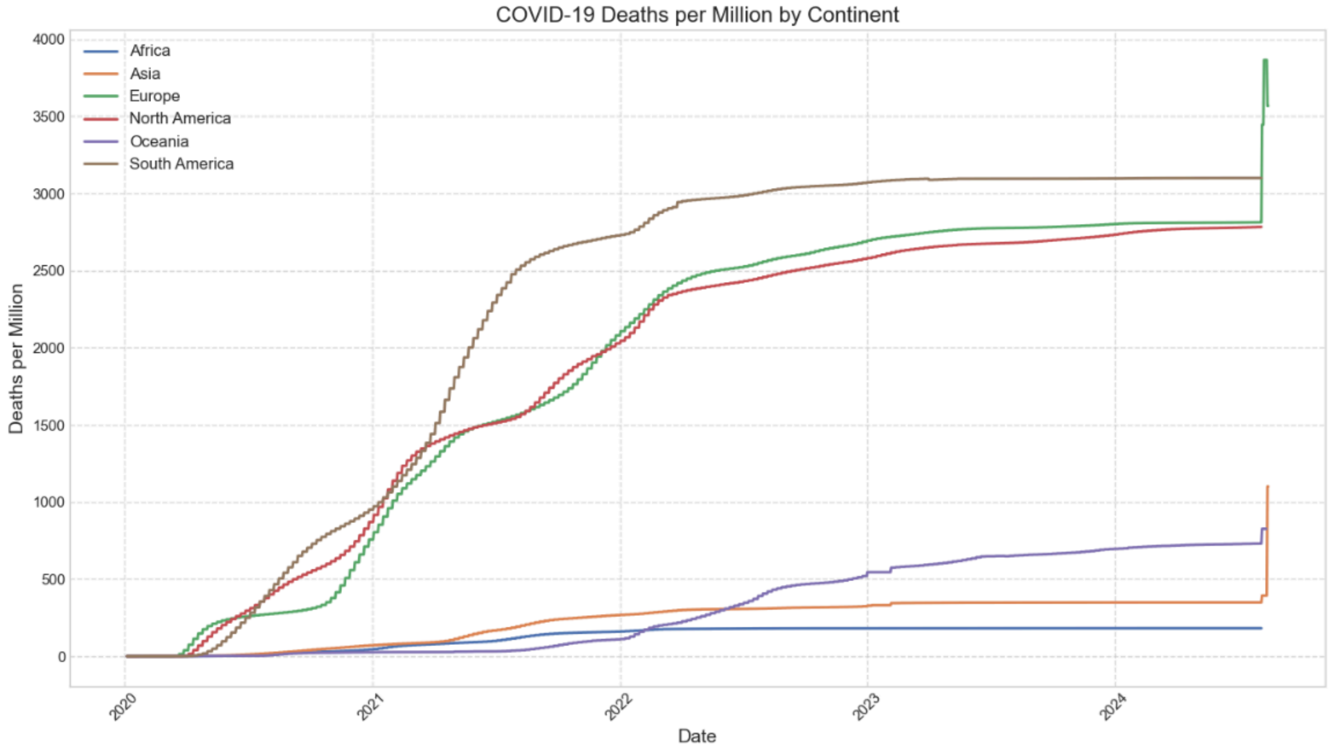


Figure 2 *(Deaths per Million by Continent)*

The case fatality rates provide critical context to these patterns, with South America and Africa showing the highest rates at approximately 2%, despite Africa having the lowest reported cases per million *(See Figure 3)*. This inverse relationship between case rates and fatality rates in some regions strongly suggests significant differences in testing capacity, healthcare system resilience, and reporting methodologies. Oceania's remarkably low fatality rate of roughly 0.2%, despite having among the highest case rates, further highlights the importance of healthcare system capacity and effective public health measures in determining outcomes.

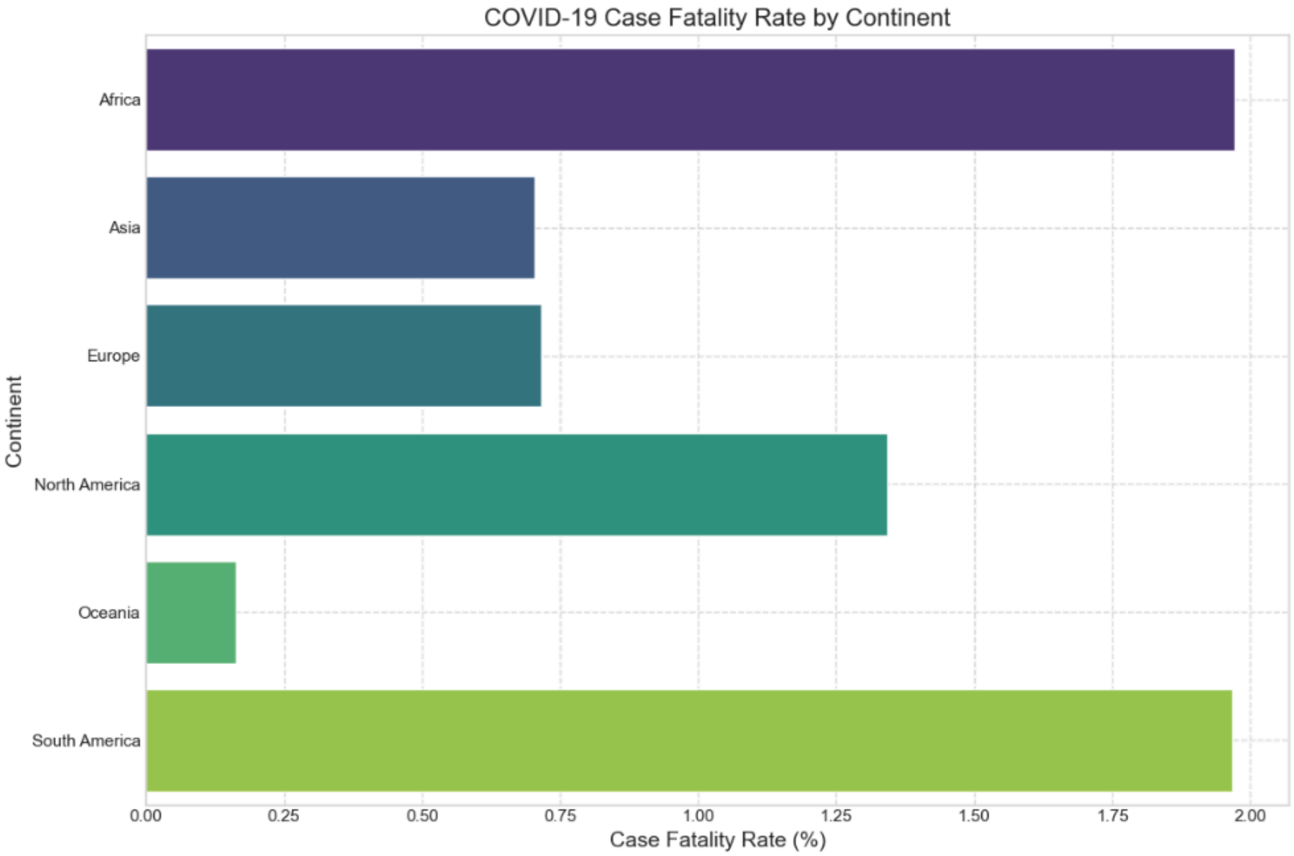


Figure 3 *(Case Fatality Rate by Continent)*

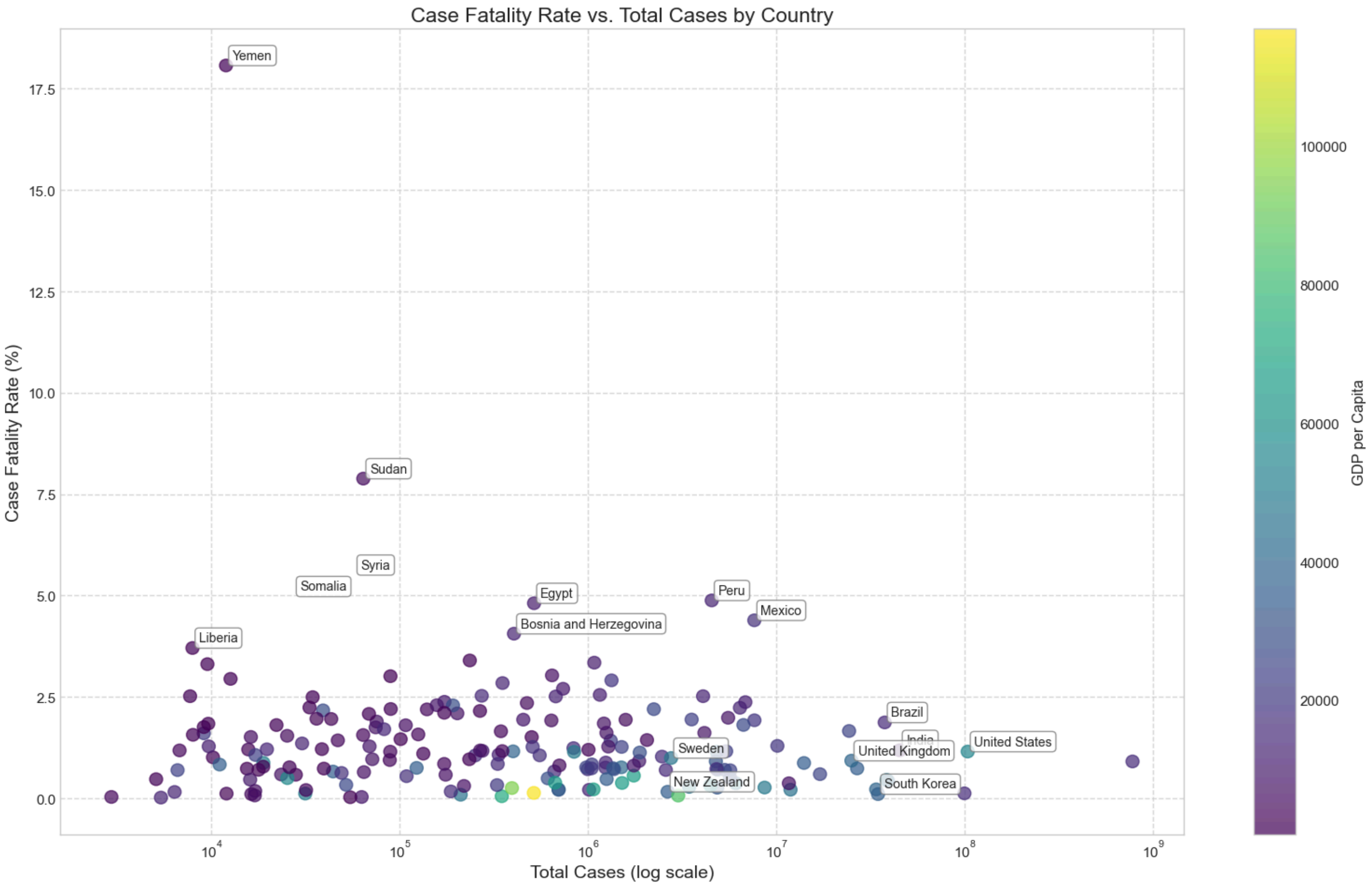


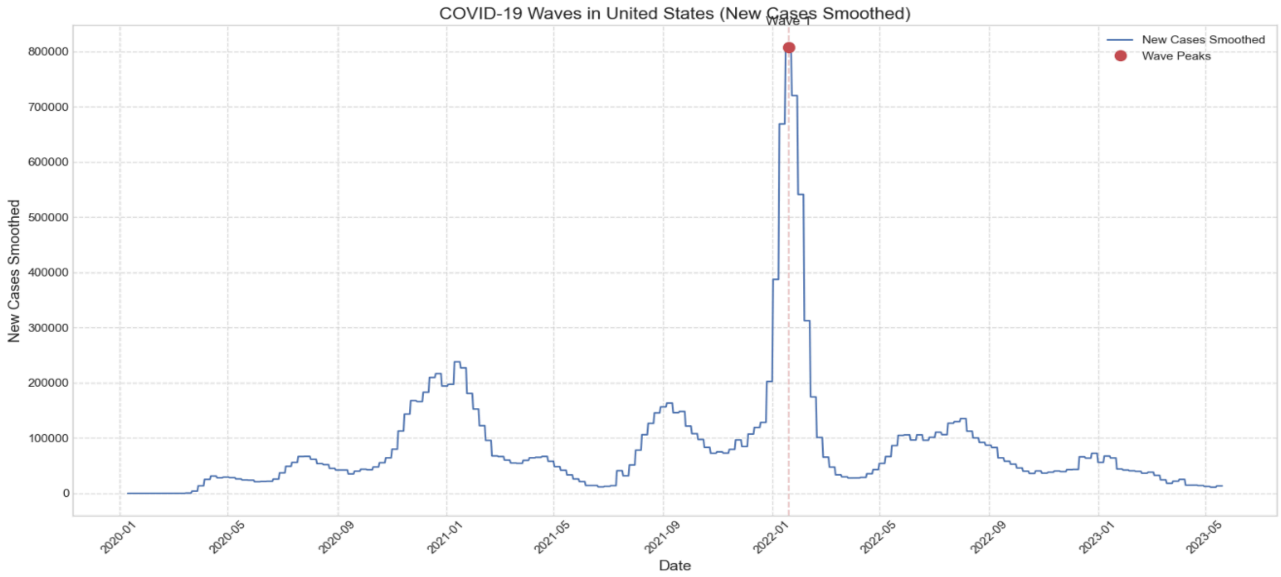
Figure 4 *(Case Fatality Rate vs. Total Cases by Country)*

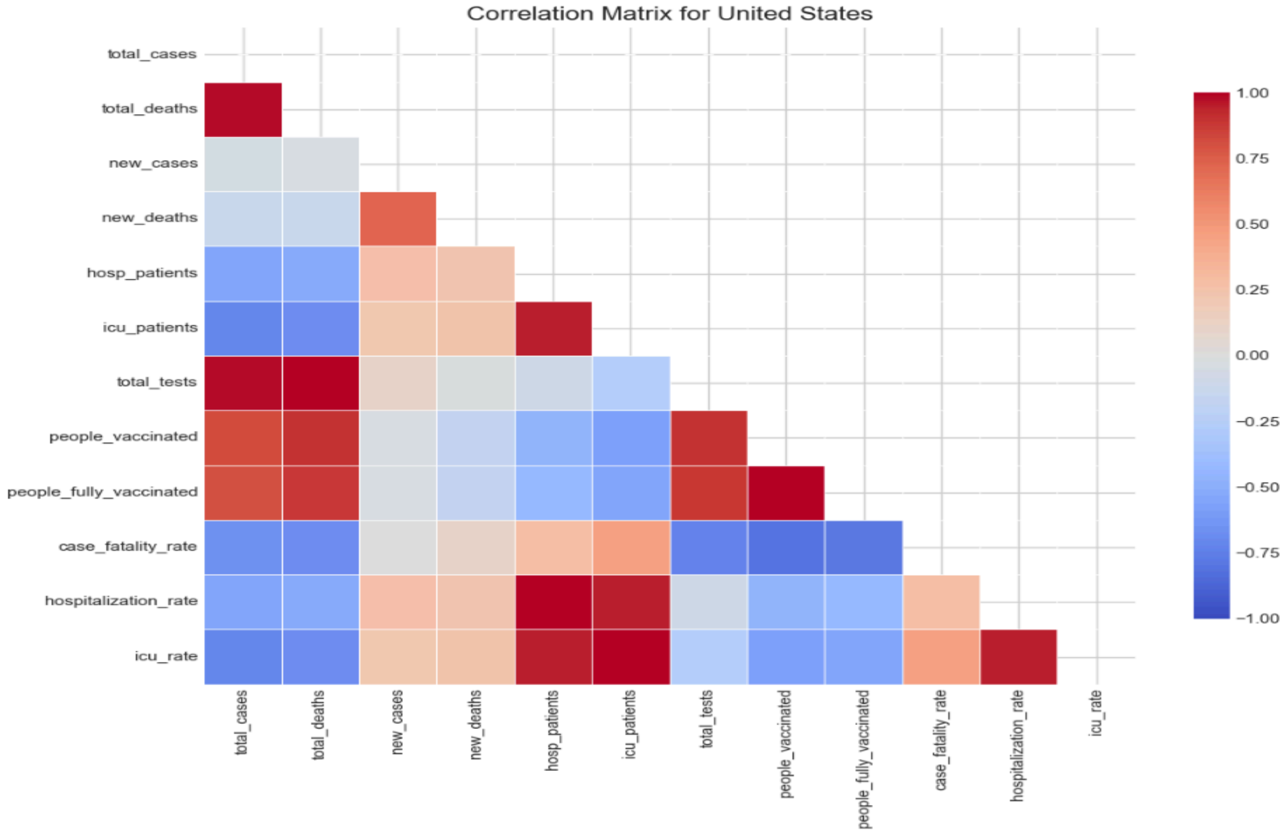
[2] COUNTRY TREND ANALYSIS:

The scatter plot reveals a striking inverse relationship between total case counts and fatality rates across nations. Yemen stands out dramatically with the highest case fatality rate at approximately 18%, despite having relatively few total cases (in the 10,000 range). Sudan and Syria follow with fatality rates around 7-8% and 5.5%, respectively. These countries with limited healthcare infrastructure show significantly higher death rates per infection. Conversely, wealthy nations with extensive testing capacity like the United States, the United Kingdom, and South Korea demonstrate much lower fatality rates (around 1-2%) despite their massive case counts (in the tens to hundreds of millions). The logarithmic scale on the x-axis highlights how countries cluster based on their testing capabilities and healthcare systems. Additionally, the color gradient indicating GDP per capita further confirms that wealthier nations (lighter colors) generally report more cases but experience lower fatality rates, while poorer countries (darker purple) show the opposite pattern. This visualization effectively captures how economic resources, healthcare capacity, and testing availability have shaped the pandemic's impact across different nations.

[3] WAVE DETECTION / CORRELATION ANALYSIS:

United States:

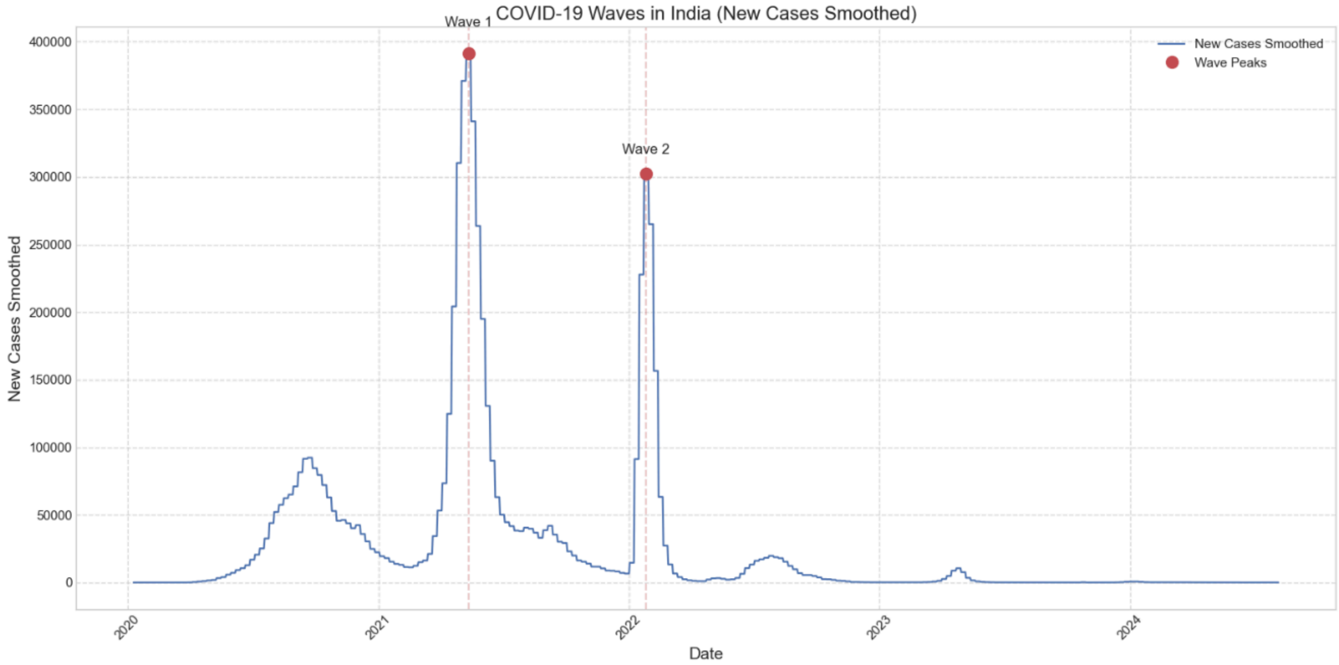


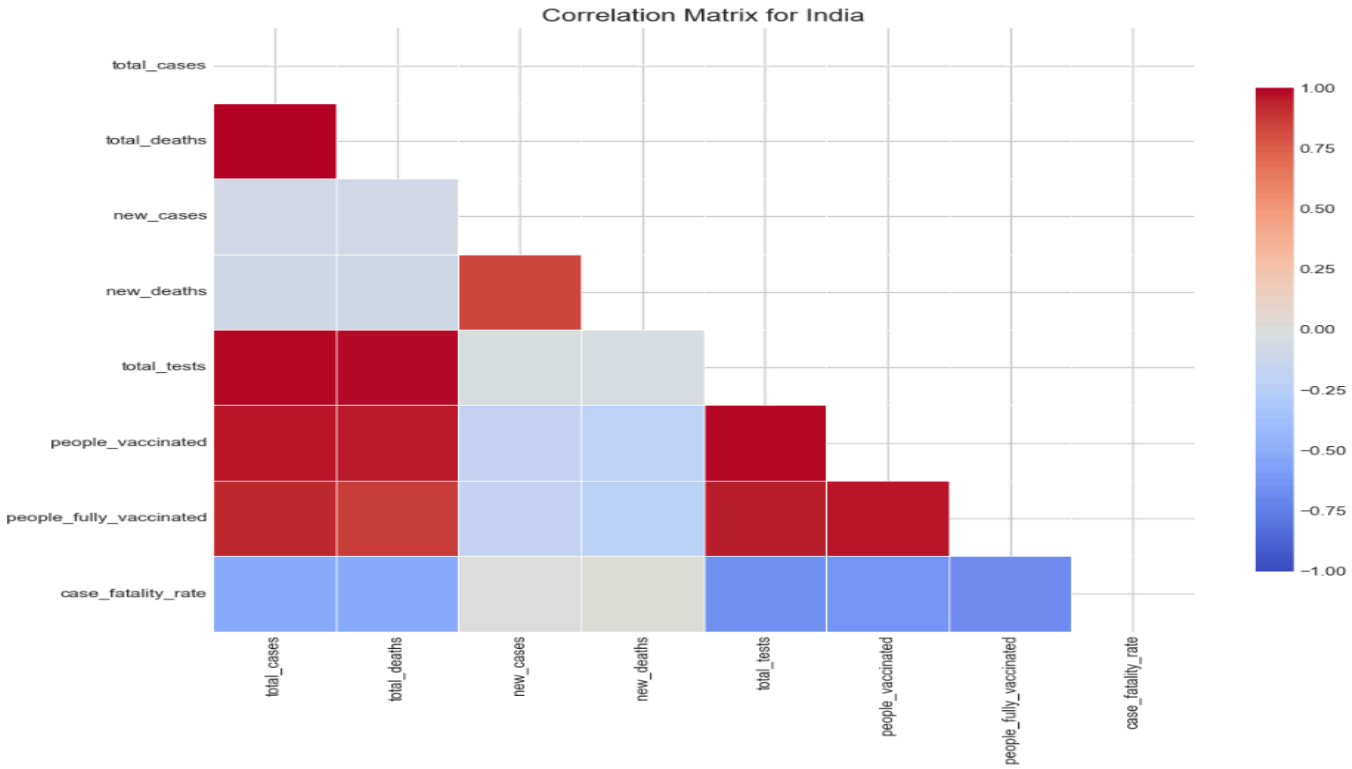


The United States experienced several significant waves of COVID-19 infections throughout the pandemic, each with distinct characteristics and public health implications. The most substantial surge occurred in early 2022, reaching a staggering peak of approximately 800,000 daily reported cases—a record high that outpaced all other countries in the dataset. Before this, the country saw multiple large waves in 2020 and 2021, with daily case counts exceeding 200,000, particularly during the winter months and following holiday gatherings. These earlier waves placed immense pressure on the healthcare system and highlighted disparities in testing, access to care, and vaccination rates across regions.

The United States correlation matrix demonstrates an extremely strong positive correlation (0.98) between total cases and total deaths, one of the highest among the analyzed countries. This suggests a very consistent relationship between infection spread and mortality throughout the pandemic. Regarding vaccinations, there is a substantial negative correlation (-0.46) between people vaccinated and hospitalization rates, indicating that the vaccination program helped reduce hospital admissions. The US data shows a strong positive correlation (0.95) between ICU patients and hospitalization rates, highlighting the direct relationship between overall hospitalizations and critical care needs. Additionally, the strong negative correlation (-0.80) between people vaccinated and case fatality rate supports the effectiveness of vaccines in preventing deaths among those infected.

India:

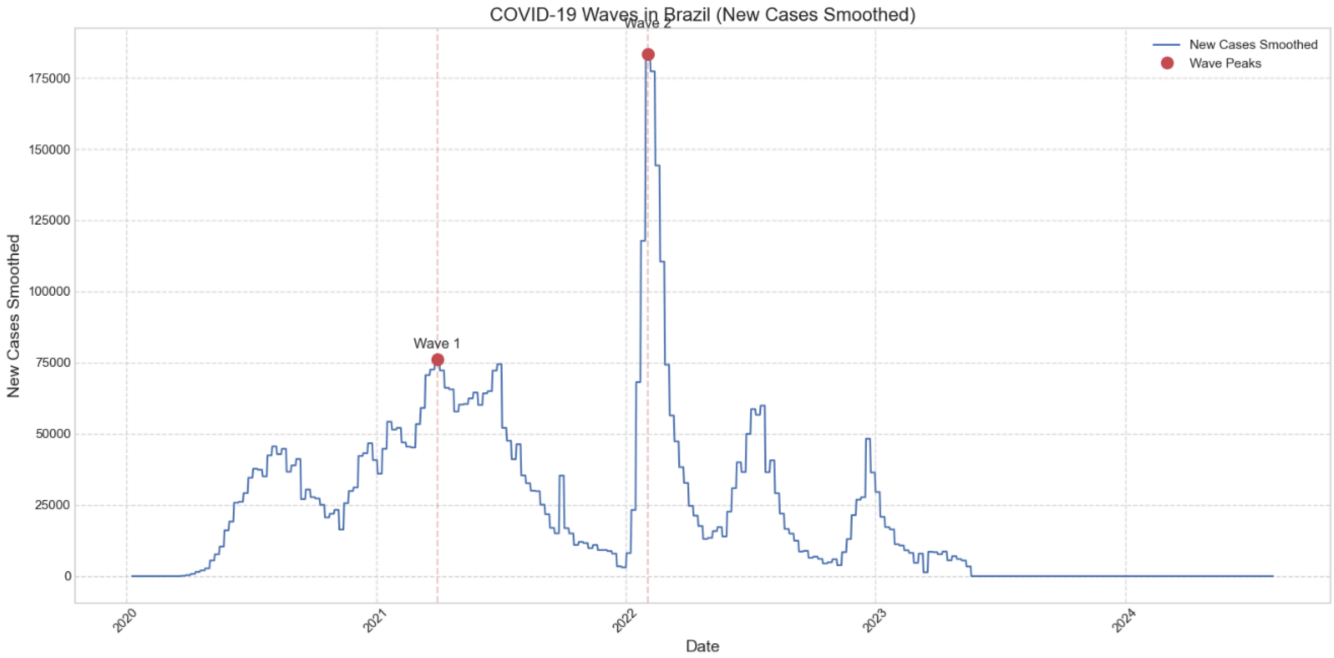


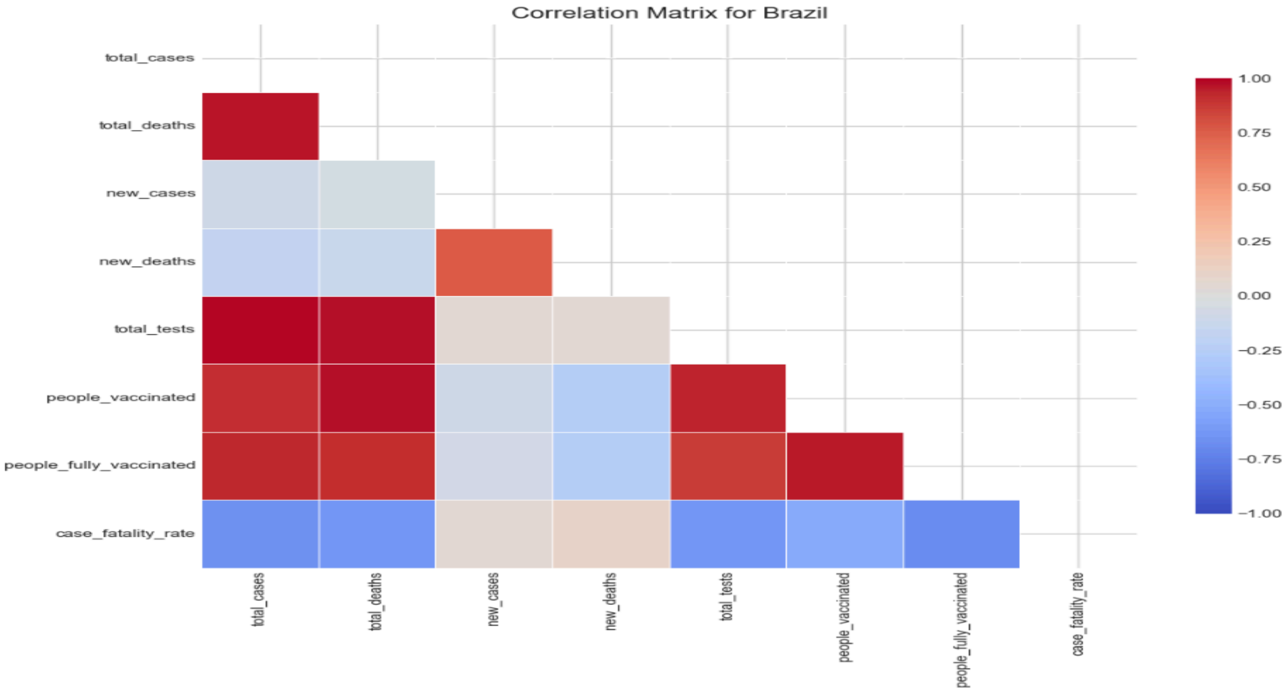


India's pandemic trajectory featured two intense and defining waves. The first, in mid-2021, surged to nearly 400,000 daily cases and triggered a nationwide crisis with overwhelmed hospitals and widespread oxygen shortages. The Delta variant was a key driver, exposing gaps in healthcare infrastructure. A second wave in early 2022 rose quickly to around 300,000 daily cases but subsided more rapidly. By 2023, the country saw only minor fluctuations, with daily cases remaining relatively low, likely due to a combination of high vaccination coverage and natural immunity from prior infection. This period of stability marked a turning point in India’s management of the pandemic.

India's correlation matrix shows a perfect positive correlation (1.00) between total cases and total deaths, the strongest among all countries analyzed, suggesting extremely consistent mortality rates relative to infection numbers. The strong positive correlation (0.96) between people vaccinated and total deaths might seem counterintuitive but likely reflects that both metrics increased over time rather than a causal relationship. Without hospitalization data in the provided matrix, direct analysis of vaccination impact on hospitalizations isn't possible. However, the moderate negative correlation (-0.64) between people vaccinated and case fatality rate suggests that vaccination helped reduce the proportion of deaths among infected individuals. India also shows a very strong positive correlation (0.99) between testing and vaccination, indicating that both public health measures expanded in tandem.

Brazil:

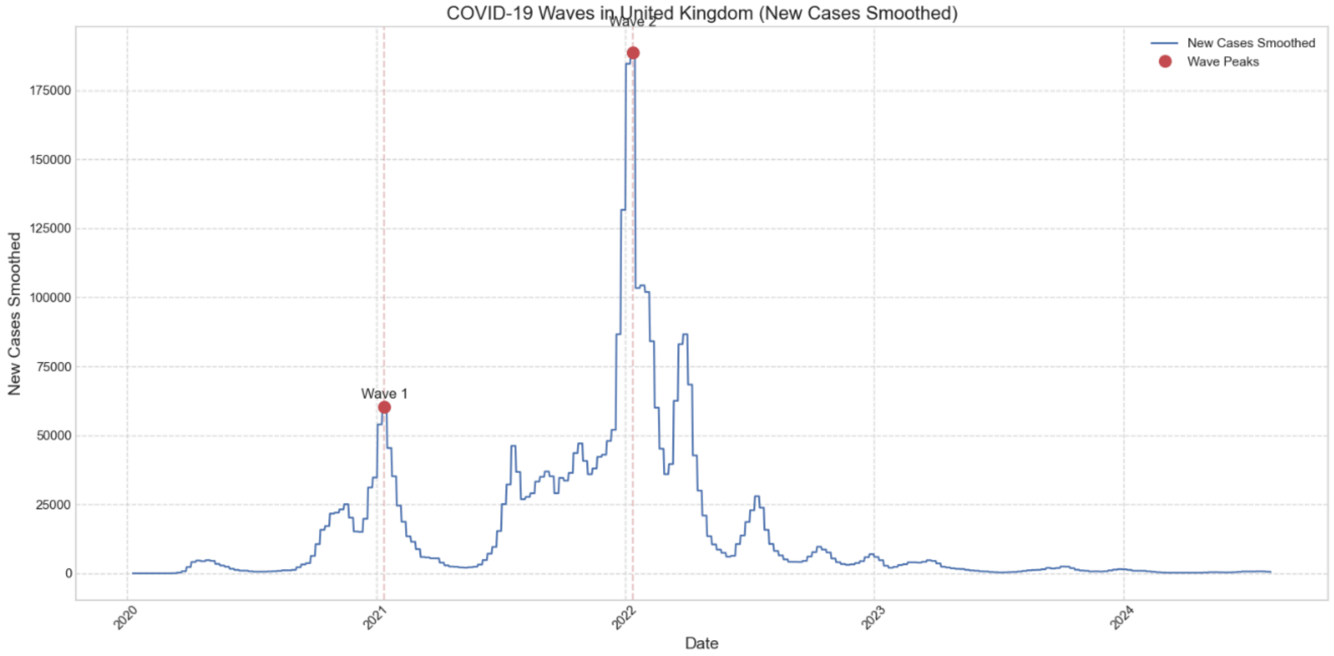


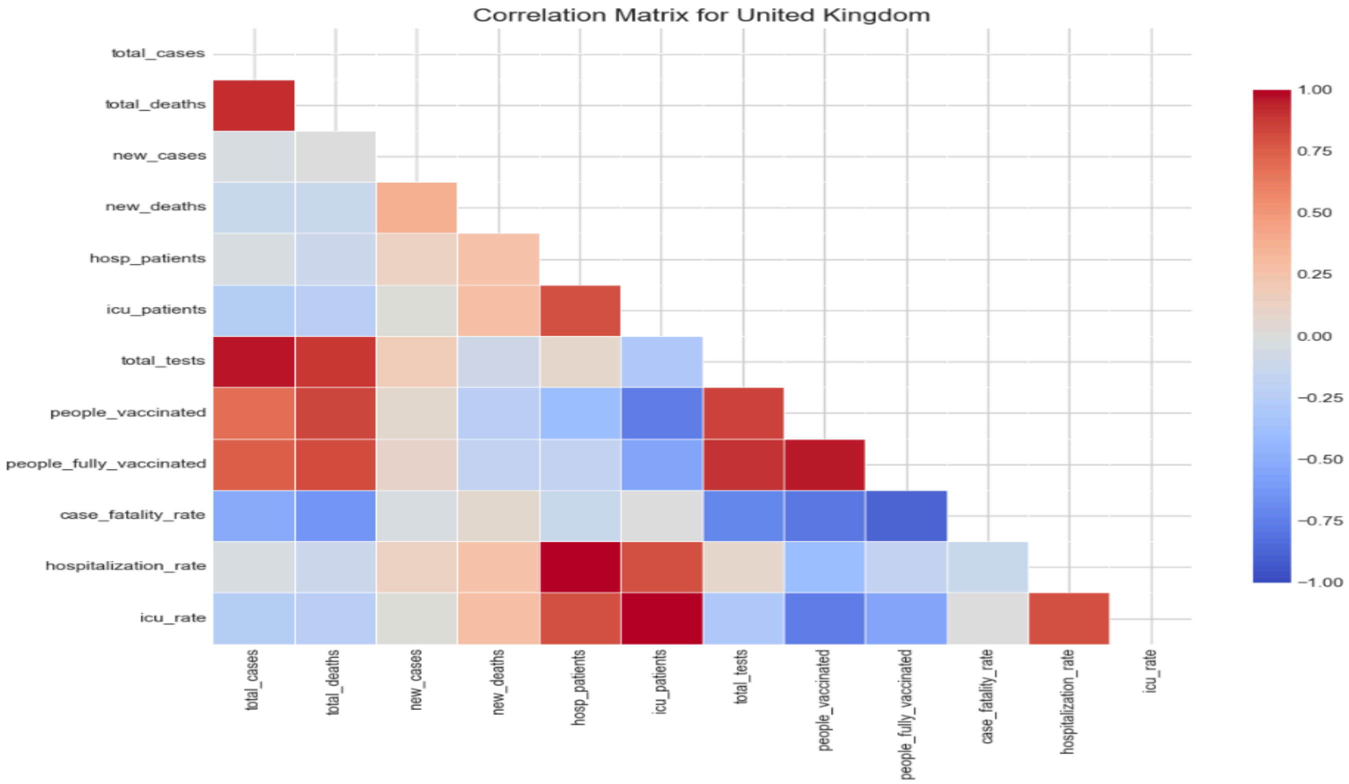


Brazil experienced a prolonged and turbulent pandemic, marked by multiple waves of varying intensity. The first significant wave peaked in mid-2021 with around 75,000 daily cases, straining the public health system. A much larger second wave followed in early 2022, surpassing 175,000 daily cases at its peak. Throughout 2022 and into 2023, Brazil continued to face recurring moderate waves, often around 50,000 daily cases, reflecting ongoing transmission challenges in densely populated urban areas and remote regions alike. It wasn't until 2024 that case numbers significantly declined, suggesting that population immunity and improved public health response finally began to curb widespread transmission.

Brazil's correlation matrix shows a very strong positive correlation (0.97) between total cases and total deaths, indicating a consistent relationship between infection spread and mortality. The data for Brazil is less complete than other countries, lacking specific hospitalization and ICU metrics in the provided matrix. However, the available data shows a strong negative correlation (-0.51) between people vaccinated and case fatality rate, suggesting that vaccination helped reduce the proportion of deaths among infected individuals. Brazil also demonstrates strong positive correlations between vaccination metrics and testing (0.94 for people vaccinated and total tests), possibly indicating that increased testing capacity developed alongside vaccination capabilities, representing overall improvement in pandemic response infrastructure.

United Kingdom:



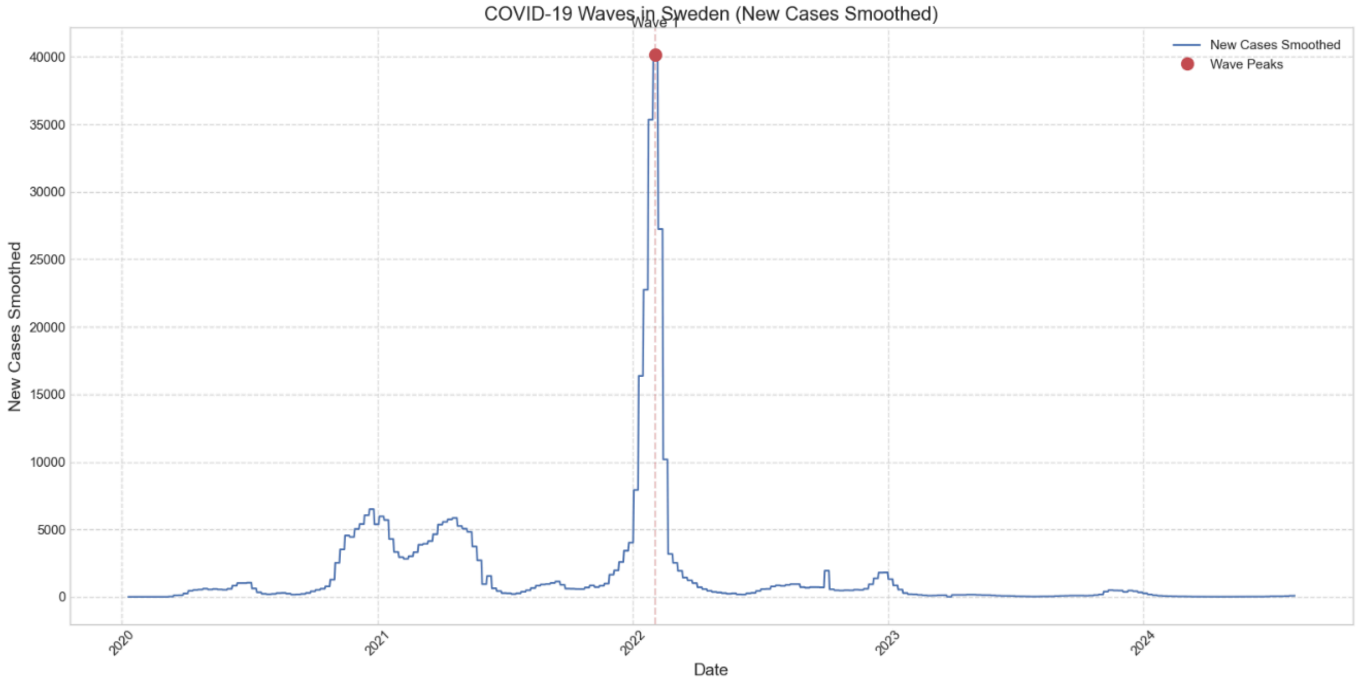


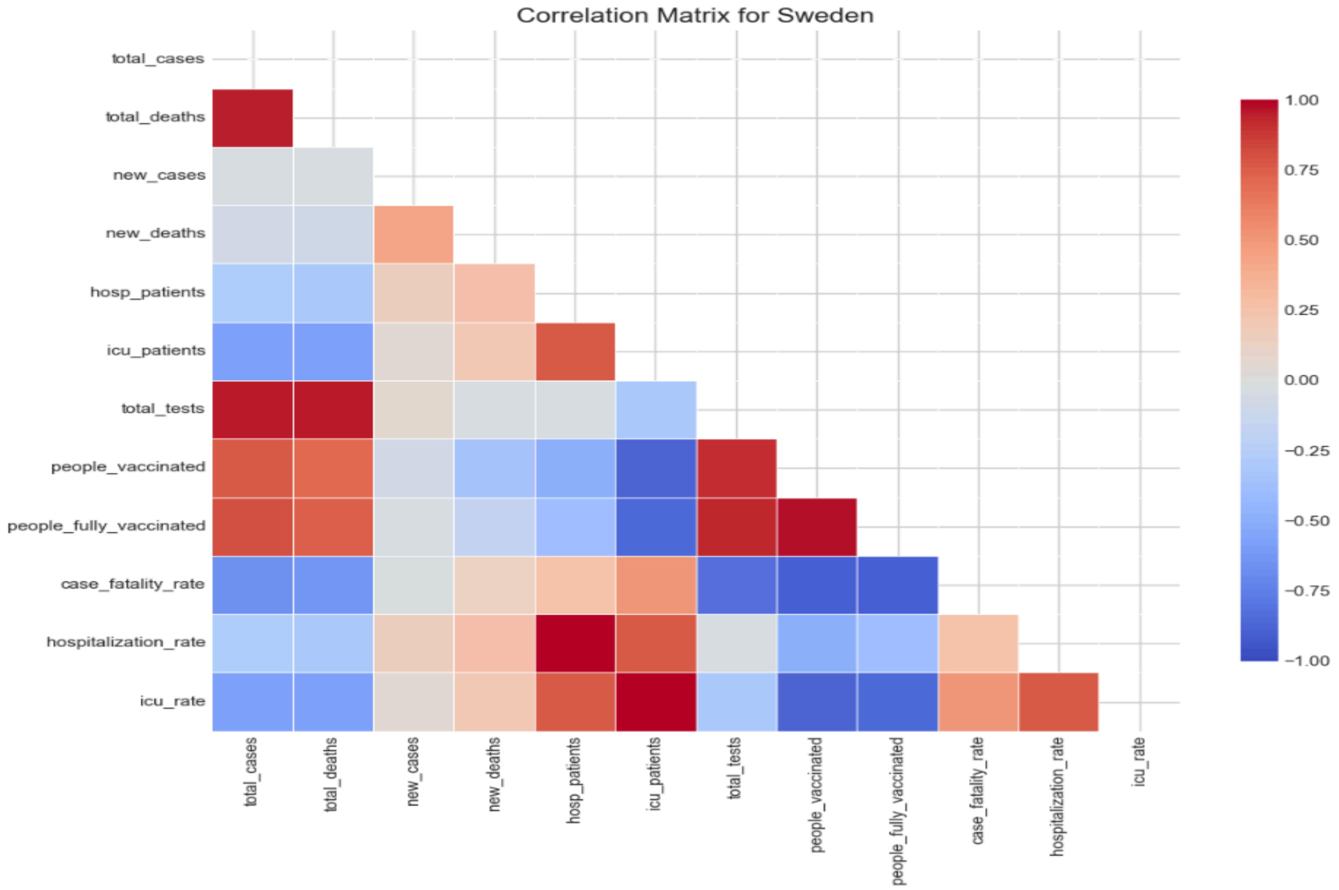
The United Kingdom's pandemic pattern was shaped by two significant waves. The first, in early 2021, peaked at around 60,000 daily cases and coincided with the spread of the Alpha variant. A much larger second wave followed in early 2022, driven by Omicron, reaching nearly

190,000 daily cases. Despite this surge, hospitalizations and deaths were relatively lower due to widespread vaccination. Throughout 2022 and 2023, the UK experienced multiple smaller waves, but each with diminishing impact. By 2024, daily cases had fallen to minimal levels, likely reflecting both high population immunity and the evolution of less severe variants.

The UK correlation matrix reveals a strong positive correlation (0.92) between total cases and total deaths, indicating that as case numbers increased, deaths increased proportionally. Interestingly, there's a notable negative correlation (-0.39) between people vaccinated and hospitalization rates, suggesting that vaccination campaigns helped reduce hospital admissions. The data also shows a strong negative correlation (-0.79) between vaccinations and case fatality rate, further supporting the effectiveness of vaccines in reducing mortality. Another significant finding is the perfect correlation (1.00) between hospitalization rate and hospital patients, confirming the consistency of the hospitalization metrics. ICU patients show a strong positive correlation (0.81) with both hospitalization rate and hospital patients, reflecting the expected progression from hospitalization to intensive care in severe cases.

Sweden:

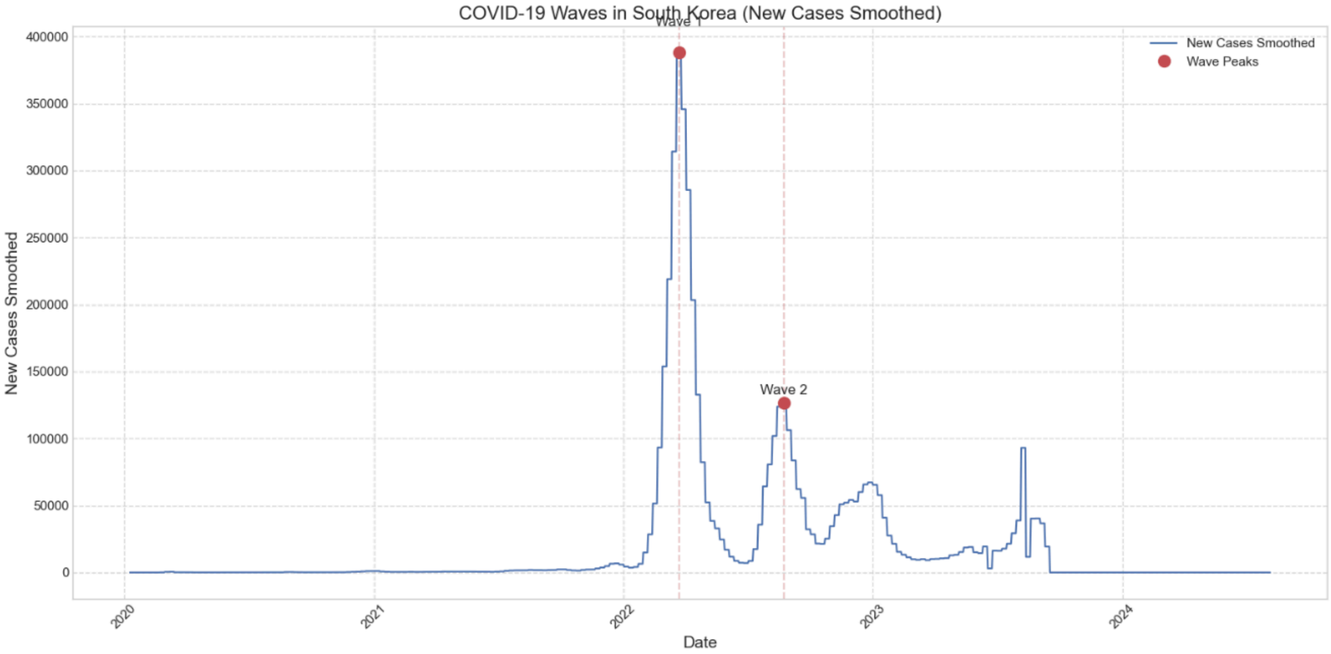


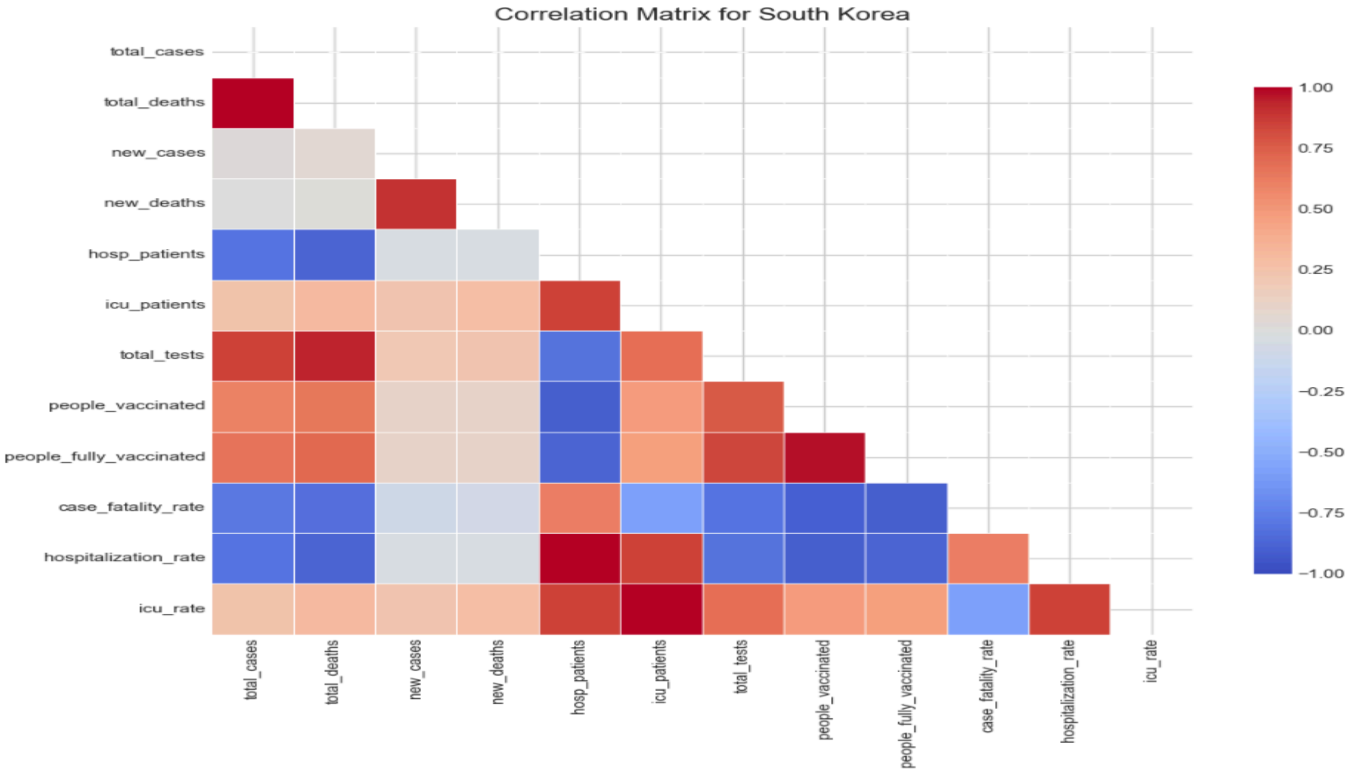


Sweden experienced several moderate waves during 2020 and 2021, largely shaped by its unique strategy of avoiding strict lockdowns. However, its largest and most dramatic surge occurred in early 2022, with daily cases spiking to around 40,000, significantly higher than earlier peaks. This major wave, arriving later than in many other countries, highlights how the country’s lenient early approach may have delayed but not avoided a substantial outbreak. After this peak, Sweden managed to keep case numbers consistently low through 2023 and 2024, suggesting eventual success in achieving population-level control.

Sweden's correlation matrix demonstrates a strong positive correlation (0.95) between total cases and total deaths, in line with patterns observed in other countries. Regarding vaccination impact, Sweden shows a substantial negative correlation (-0.50) between people vaccinated and hospitalization rates, suggesting that their vaccination program helped reduce hospital admissions. The data also reveals a very strong negative correlation (-0.90) between vaccination and case fatality rate, indicating that vaccines effectively reduced the proportion of deaths among infected individuals. Sweden exhibits a strong negative correlation (-0.88) between people vaccinated and ICU patients, one of the strongest such relationships among the analyzed countries, suggesting their vaccination program was particularly effective at preventing severe cases requiring intensive care.

South Korea:

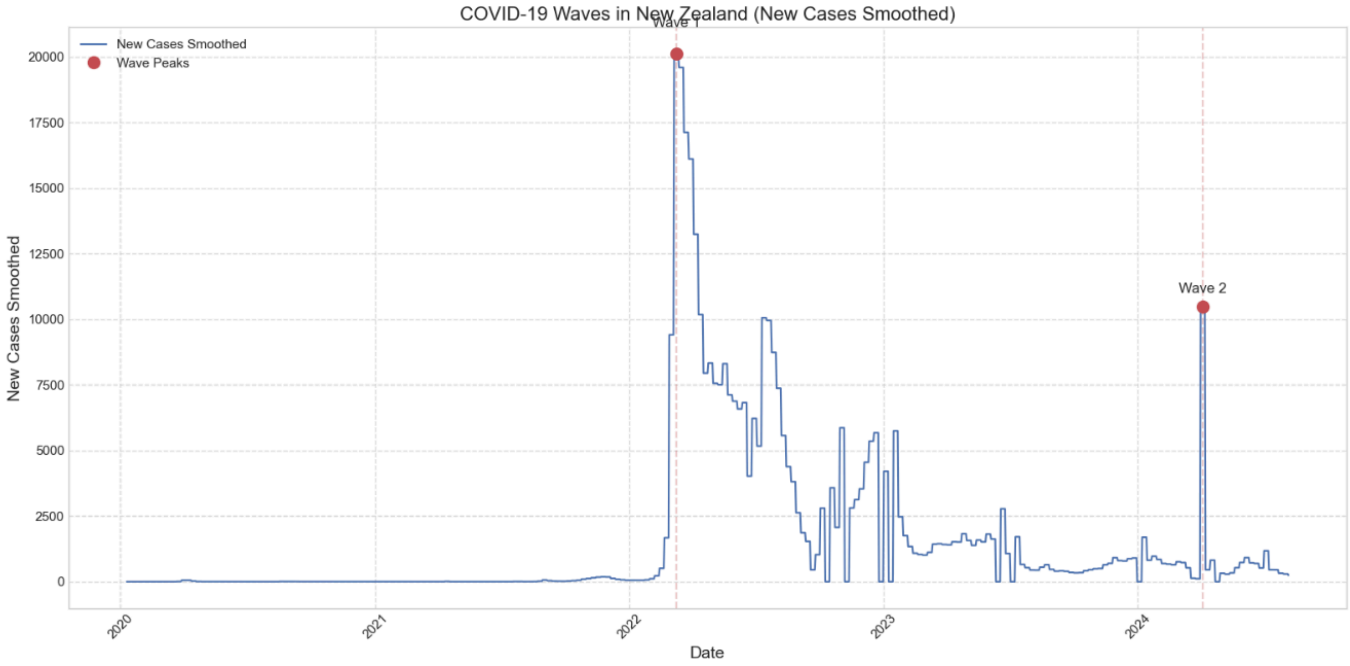


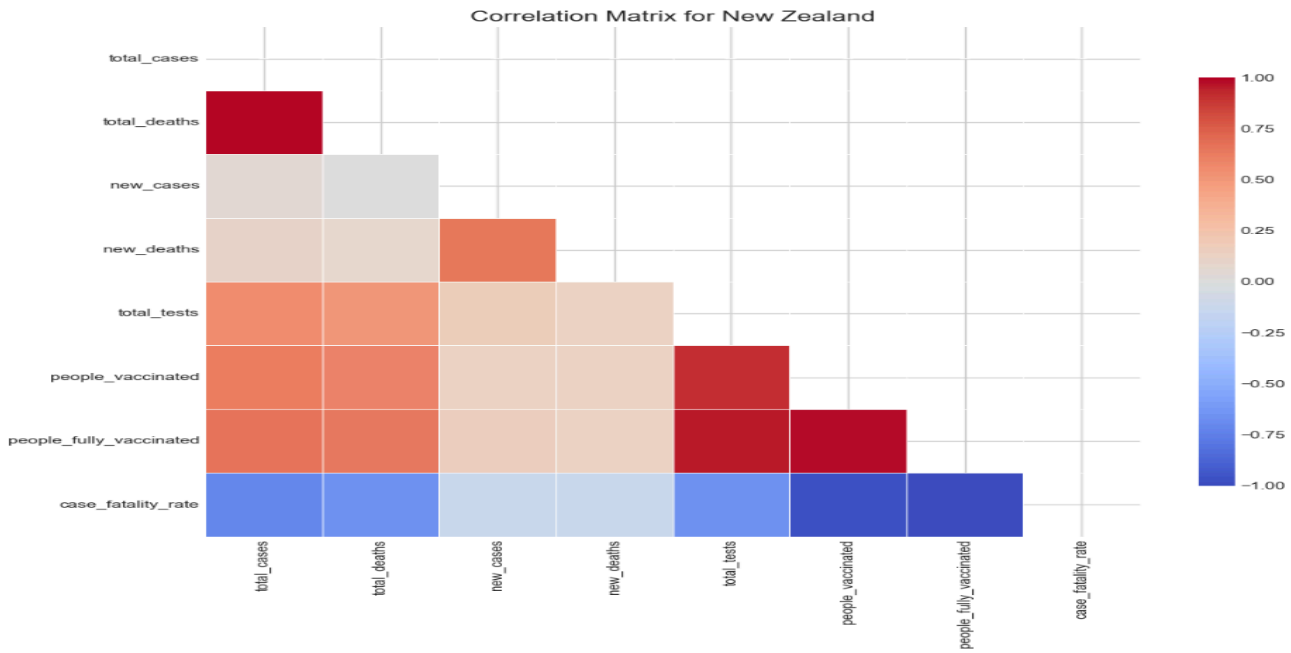


South Korea's early pandemic response was widely praised for its rapid testing, contact tracing, and strict containment measures, which kept case numbers low through 2020 and 2021. However, in early 2022, the country saw a massive surge, peaking at nearly 400,000 daily cases—its largest wave. A second, smaller wave followed later that year with about 125,000 daily cases. Unlike many other nations, South Korea continued to experience several smaller but persistent waves into 2023 and early 2024, including a notable resurgence, highlighting the ongoing challenge of maintaining control amid evolving variants and public fatigue.

South Korea's correlation matrix displays an exceptionally strong positive correlation (0.99) between total cases and total deaths, nearly perfect, suggesting highly consistent mortality rates relative to case numbers. Notably, South Korea shows the strongest negative correlation (-0.91) between people vaccinated and hospitalization rates among all countries analyzed, indicating a very effective vaccination campaign in reducing hospital admissions. Interestingly, South Korea exhibits a unique pattern where hospital patients have a strong negative correlation (-0.80) with total cases, suggesting that South Korea may have implemented effective early interventions that prevented hospitalizations even as cases increased. The matrix also shows a strong negative correlation (-0.91) between vaccination rates and case fatality rates, further supporting the effectiveness of vaccines in reducing the severity of infections.

New Zealand:

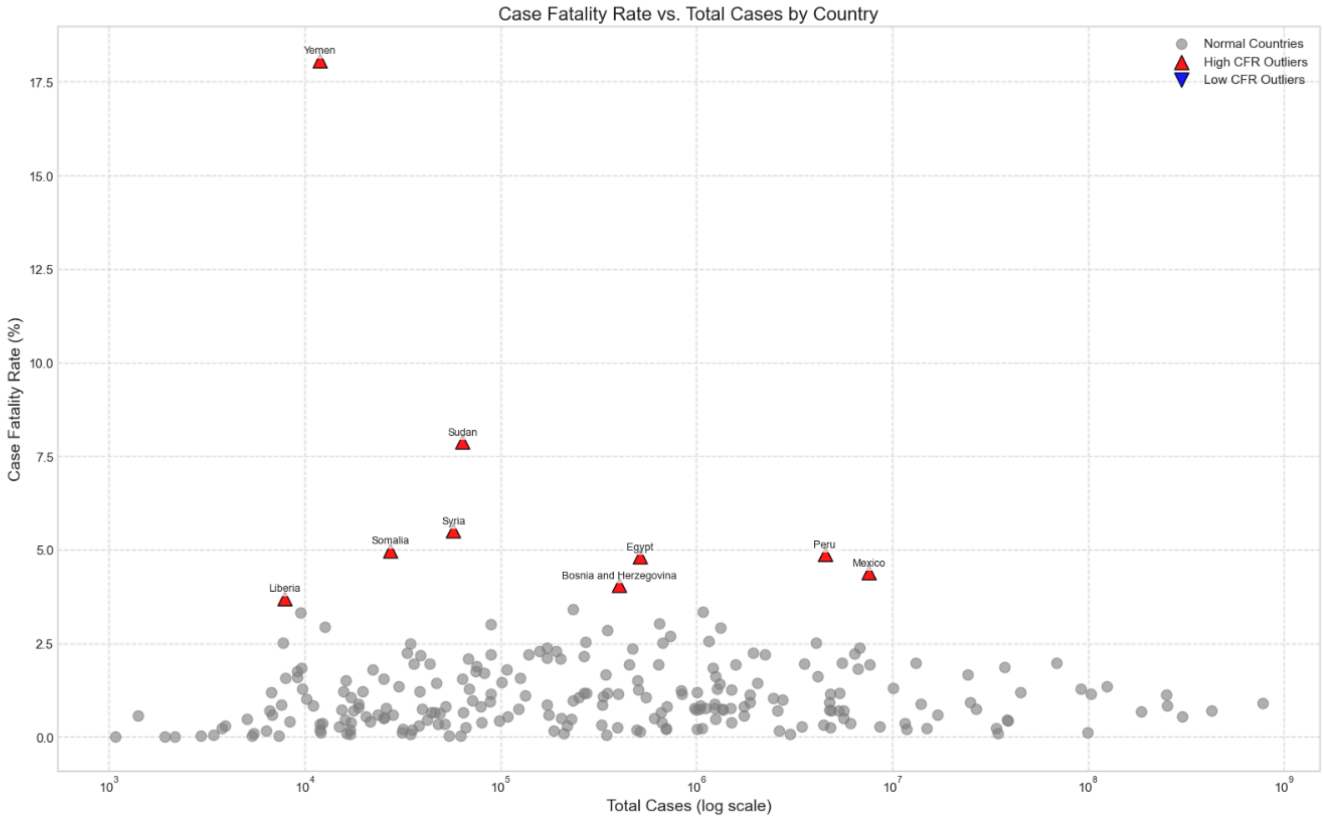




New Zealand's strict "COVID-zero" strategy, characterized by early border closures and lockdowns, successfully kept daily case counts near zero through much of 2020 and 2021. The country faced its first significant wave only in early 2022, reaching about 20,000 daily cases as restrictions loosened and Omicron spread. After a period of relatively controlled transmission, a surprising second wave occurred in 2024, marked by a sharp vertical spike of comparable size. This rare pattern underscores New Zealand's initial pandemic success, followed by the eventual challenge of reintegration with the global community.

New Zealand's correlation matrix exhibits a very strong positive correlation (0.98) between total cases and total deaths, consistent with global patterns. However, New Zealand's data shows some unique characteristics, with notably weaker correlations between testing and cases/deaths (0.55 and 0.51, respectively) compared to other countries, possibly reflecting their successful elimination strategy during early pandemic phases. The matrix reveals a remarkably strong negative correlation (-0.97) between people vaccinated and case fatality rate, one of the strongest among all countries analyzed, suggesting that New Zealand's vaccination program was particularly effective in reducing mortality among infected individuals. Unlike some other countries, New Zealand does not have hospitalization data in the provided matrix, limiting analysis of the vaccination-hospitalization relationship.

[4] OUTLIER DETECTION AND ANALYSIS:



The scatter plot reveals several striking outliers in case fatality rates (CFR) across countries, with significant variations. The most notable high CFR outliers are marked with a red triangle, while the majority of countries cluster in a more typical range between 0-3% CFR, regardless of their total case count.

Yemen stands as the most extreme outlier with an alarming CFR of 18%, which is dramatically higher than any other nation. This high death rate can most likely be attributed to Yemen's devastated healthcare infrastructure due to ongoing civil war, severe humanitarian crisis, widespread malnutrition, and limited testing capacity that can probably only identify severe cases. The combination of these factors creates a perfect environment where COVID-19 is allowed to spread while cases go under reported and ineffectively treated.

Sudan follows suit as the second-highest outlier with a CFR of around 8%. This elevated rate can be attributed to Sudan's fragile healthcare system, political instability, economic challenges, and limited testing capabilities. Similar to Yemen, Sudan likely experienced significant underreporting of mild cases, perhaps artificially inflating the death rate.

Syria, Somalia, Egypt, Bosnia and Herzegovina, Peru, and Mexico all demonstrate CFRs in the 4-5.5% range, significantly above the global average. These countries share several common factors: strained healthcare systems, limited ICU capacity, high rates of afflictions like diabetes and hypertension, and in some cases (Syria, Somalia), ongoing conflict that severely restricts healthcare access. Peru's high CFR is particularly notable given its relatively higher case count, potentially reflecting the impact of oxygen shortages and an overwhelmed healthcare system.

Liberia presents an interesting case with a relatively high CFR despite a low case count, possibly reflecting the lingering effects of its previously devastated healthcare system following the Ebola crisis and extremely limited testing capacity.

The data shows no prominently marked low CFR outliers (blue triangles) in the provided chart, suggesting that while many factors can drive fatality rates upward, achieving exceptionally low rates below the typical global range was uncommon. Most "normal countries" (gray circles) maintain CFRs clustering between 0-3% across the entire spectrum of case counts from thousands to hundreds of millions.

Interestingly, the plot reveals no strong correlation between total case counts and fatality rates among normal countries, suggesting that a country's ability to manage case fatality was more dependent on healthcare quality, demographics, and policy responses than on the absolute scale of their outbreak. The logarithmic scale for total cases further emphasizes that even countries with vastly different outbreak sizes could achieve similar CFRs with appropriate interventions.

The outliers highlight how pre-existing vulnerabilities in healthcare systems and concurrent crises dramatically amplified COVID-19's impact in specific nations, underscoring the importance of robust healthcare infrastructure and the devastating consequences when such systems are compromised by conflict, poverty, or governance challenges.

[5] COMPARATIVE DISCUSSION ON TRENDS BETWEEN COUNTRIES:

The COVID-19 pandemic’s trajectory varied significantly across countries, influenced by factors such as healthcare systems, governmental responses, population demographics, and economic status. A closer look at the trends in several key countries provides valuable insights into how these factors played out in practice.

In the United States, the pandemic’s course was marked by multiple waves, with the largest peak in early 2022 at approximately 800,000 daily cases. This high case count was the result of the rapid spread of highly transmissible variants like Omicron, and it highlights the challenges faced by a country with widespread transmission. Despite the large number of cases, the U.S. had a relatively low case fatality rate, which can be attributed to its advanced healthcare infrastructure, vaccination efforts, and early adoption of treatments. Post-peak, the U.S. saw smaller waves, indicative of the gradual development of immunity through both natural infection and vaccination.

India experienced two major waves that defined its pandemic response. The first, in mid-2021, was a catastrophic event, with nearly 400,000 daily cases overwhelming the healthcare system. The second wave in early 2022 saw slightly lower numbers but remained high at 300,000 cases. India’s rapid vaccination campaign following these devastating waves likely contributed to the country's relatively stable case numbers by 2023. The rapid rise and fall of these waves suggest that population immunity, coupled with ongoing public health measures, was pivotal in controlling the virus in later stages.

In Brazil, the pattern was similarly shaped by two major surges, with the first wave in mid-2021 peaking at 75,000 daily cases, followed by a larger second wave in early 2022. Brazil continued to experience moderate waves throughout 2022 and 2023, likely due to its vast and diverse geography, which presented challenges in uniform containment strategies. The sustained nature of these waves suggests difficulty in achieving widespread immunity and controlling the spread across the country’s varied regions.

The United Kingdom mirrored other European nations with two significant surges. The first wave reached 60,000 daily cases in early 2021, while the second wave in early 2022 surpassed 190,000 cases. Following these peaks, the UK experienced smaller waves, gradually declining to minimal levels by 2024. This decline may be attributed to successful vaccination programs and the adaptation of public health strategies, which likely contributed to better control of the virus as less virulent variants emerged.

Sweden took a unique approach with fewer restrictions, leading to multiple smaller waves during 2020-2021. Its largest surge, however, came in early 2022, nearly two years into the pandemic, with a spike reaching 40,000 daily cases. This delayed outbreak reflects the challenges of relying on herd immunity without stricter measures. After 2022, Sweden managed to maintain low case numbers, indicating that the country’s approach may have delayed, but not completely avoided, significant exposure.

South Korea initially showed remarkable success in controlling the pandemic through aggressive contact tracing and testing. However, in early 2022, it faced a massive first wave with nearly 400,000 daily cases, followed by a second wave later in the year with 125,000 cases. The country continued to experience smaller waves into 2023 and 2024, suggesting ongoing transmission despite effective control measures. This extended activity highlights the challenge of completely eradicating the virus, even with a well-organized healthcare system and strict public health interventions.

Finally, New Zealand, with its strict "COVID-zero" strategy, kept cases exceptionally low until early 2022. The country’s first major wave peaked at around 20,000 daily cases. However, a surprising second wave occurred in 2024, after the country eased restrictions and reopened borders. This late surge underscores the challenge faced by countries that initially isolated themselves from the virus but were inevitably exposed once global travel resumed.

Analyzing Global Trends:

When comparing trends across continents, we see notable regional differences. Europe and Oceania recorded some of the highest case rates per million people, peaking at 330,000 cases per million by 2023. Oceania's particularly dramatic 2022 surge, likely due to the shift from elimination strategies to endemic management following vaccination rollouts in countries like Australia and New Zealand, is a key point of analysis. Africa, on the other hand, consistently reported lower case rates, which may be attributed to differences in healthcare infrastructure, testing capabilities, or demographic factors.

In terms of mortality, South America experienced the highest death rates, reaching around 3,100 deaths per million by 2024. This was followed by Europe and North America, both of which saw similar mortality patterns, peaking at around 2,800 deaths per million. Conversely, Oceania had relatively low death rates despite high case counts, with rates around 700 per million. This discrepancy suggests that healthcare interventions, such as effective treatments and hospital care, likely mitigated the fatality rates in countries like Australia and New Zealand.

Case Fatality Rates:

The case fatality rate (CFR) offers another critical lens for understanding pandemic outcomes. South America and Africa had some of the highest CFRs, around 2%, despite Africa's lower case counts per million. This suggests that factors such as limited testing capacity, inadequate healthcare infrastructure, and underreporting could have contributed to higher fatality rates in these regions. In contrast, Oceania had a notably lower CFR of around 0.2%, despite its high case counts. This further underscores the role of healthcare system capacity and effective public health measures in managing the severity of COVID-19.

Visual Insights:

The scatter plots and visualizations further illustrate the inverse relationship between case counts and fatality rates across countries. Wealthier nations with robust healthcare systems, like the United States, the United Kingdom, and South Korea, show lower fatality rates despite high case counts. In contrast, poorer nations with weaker healthcare infrastructure, such as Yemen and Sudan, demonstrate dramatically higher fatality rates. These patterns emphasize the importance of healthcare infrastructure, testing capacity, and public health interventions in shaping the pandemic's impact on different countries.

In conclusion, the COVID-19 pandemic revealed stark differences in how countries responded and coped with the virus, shaped largely by their healthcare systems, economic resources, and public health strategies. While high-income nations generally fared better in terms of mortality and case outcomes, low- and middle-income countries faced significant challenges, highlighting the need for stronger global cooperation and support in future health crises.

TIME SERIES FORECASTING (ICU ADMISSIONS):

For this portion of the project, we focused on forecasting ICU admissions in the United States using publicly available COVID-19 data. We began by filtering the dataset to retain only the daily number of ICU patients, grouped by date. The resulting time series spanned from February 24, 2020, to August 12, 2024, providing 1,632 daily observations.To ensure compatibility with time series forecasting models like ARIMA, we set the data frequency to daily and carefully handled missing dates to maintain the time structure. This preprocessing step was critical for allowing the model to correctly capture temporal dependencies.

Before building any models, we visualized the historical ICU trends to better understand the underlying patterns and shifts across pandemic waves. With this context, we selected and implemented an ARIMA(5,1,0) model — where (5,1,0) denotes five autoregressive terms, one differencing operation (to induce stationarity), and zero moving average components. We trained the model on the complete dataset to explore its general forecasting capability, rather than focusing solely on test-set evaluation. Below, we walk through the model results and interpretation step by step.

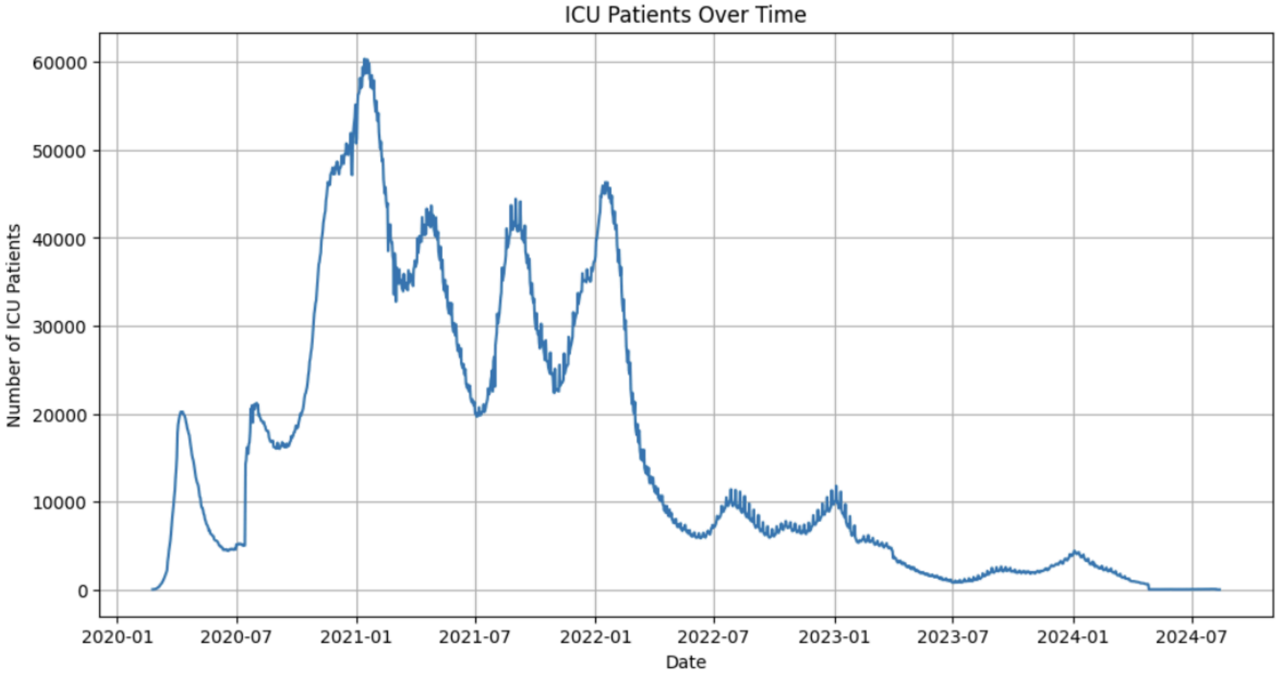
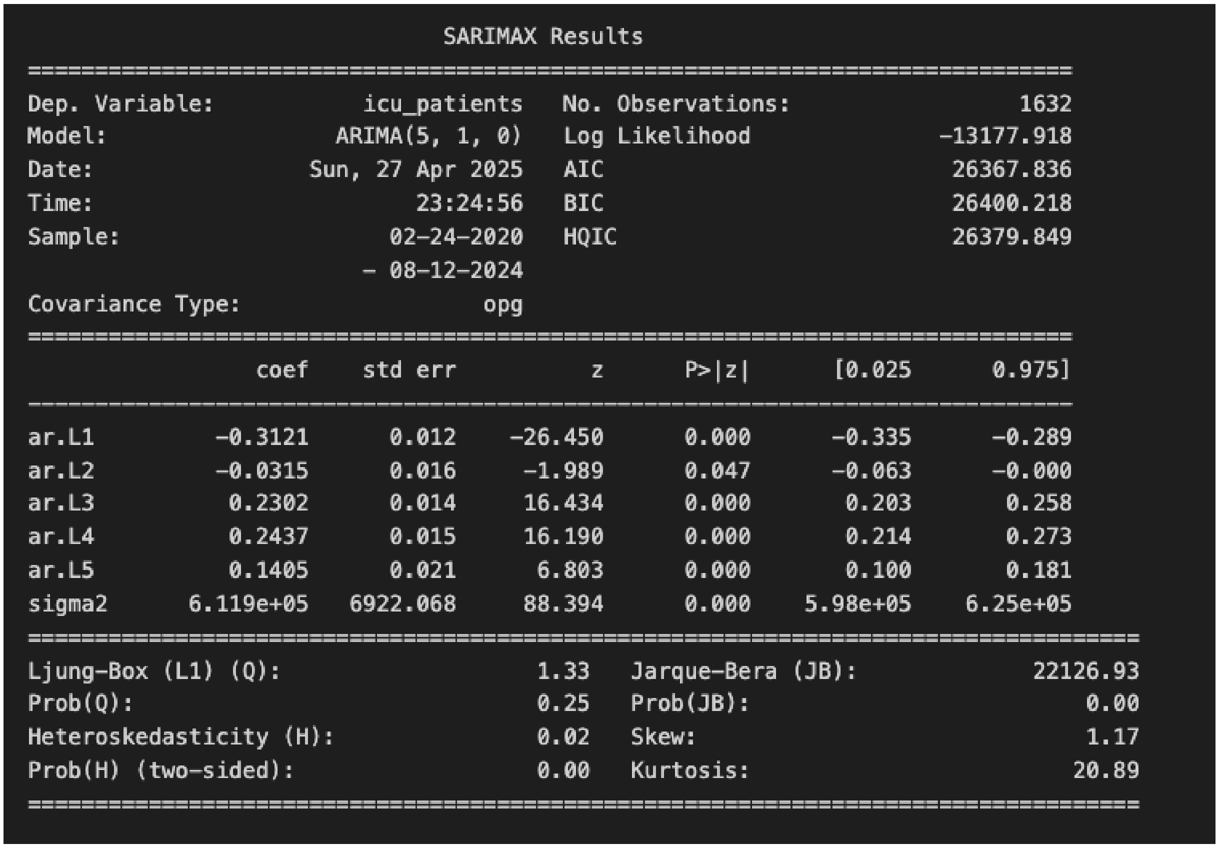


Figure 1: *ICU Patients Over Time*

This graph presents the historical trend of ICU admissions in the United States from early 2020 through mid-2024. It provides a clear visual of the pandemic's progression, highlighting the surges and declines in severe COVID-19 cases that required intensive care. The dataset includes more than 1,600 daily observations, enabling a high-resolution look at how healthcare systems were stressed over time. Several prominent peaks can be observed, corresponding to the different waves of infections—such as the initial outbreak, Delta, Omicron, and subsequent smaller surges. Each peak is followed by a notable decline, likely due to a mix of medical intervention, social distancing policies, vaccination rollouts, and public behavior changes. The graph not only serves as a timeline of pandemic severity but also justifies the need for time series modeling, as the data clearly show autocorrelated structures, seasonal changes, and variability that ARIMA-type models are designed to handle.



```
                               SARIMAX Results
==============================================================================
Dep. Variable:           icu_patients   No. Observations:                 1632
Model:                 ARIMA(5, 1, 0)   Log Likelihood              -13177.918
Date:                Sun, 27 Apr 2025   AIC                          26367.836
Time:                        23:24:56   BIC                          26400.218
Sample:                    02-24-2020   HQIC                         26379.849
                         - 08-12-2024
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.3121      0.012    -26.450      0.000      -0.335      -0.289
ar.L2         -0.0315      0.016     -1.989      0.047      -0.063      -0.000
ar.L3          0.2302      0.014     16.434      0.000       0.203       0.258
ar.L4          0.2437      0.015     16.190      0.000       0.214       0.273
ar.L5          0.1405      0.021      6.803      0.000       0.100       0.181
sigma2      6.119e+05   6922.068     88.394      0.000    5.98e+05    6.25e+05
===================================================================================
Ljung-Box (L1) (Q):                   1.33   Jarque-Bera (JB):             22126.93
Prob(Q):                              0.25   Prob(JB):                         0.00
Heteroskedasticity (H):               0.02   Skew:                             1.17
Prob(H) (two-sided):                  0.00   Kurtosis:                        20.89
===================================================================================
```

Figure 2: *ARIMA(5,1,0) Model Summary*

The ARIMA(5,1,0) model was chosen as a first approach to forecasting ICU admissions. The model includes five autoregressive terms (AR), one differencing operation to remove trend, and no moving average component (MA), making it suitable for capturing short- and medium-term memory within the data. The summary table shows that all AR lags except one are statistically significant, with p-values well below the 0.05 threshold, confirming their importance in the model. The model's diagnostics, including the AIC, BIC, and HQIC values, are included to allow for comparisons with alternative models. These metrics are key when selecting the best-fit model while balancing complexity and generalizability. Additionally, the Ljung-Box test indicates that residuals are relatively uncorrelated, which supports the model’s adequacy. While not perfect, the ARIMA(5,1,0) captures underlying temporal dependencies in the data and serves as a solid baseline for comparison with more advanced models.

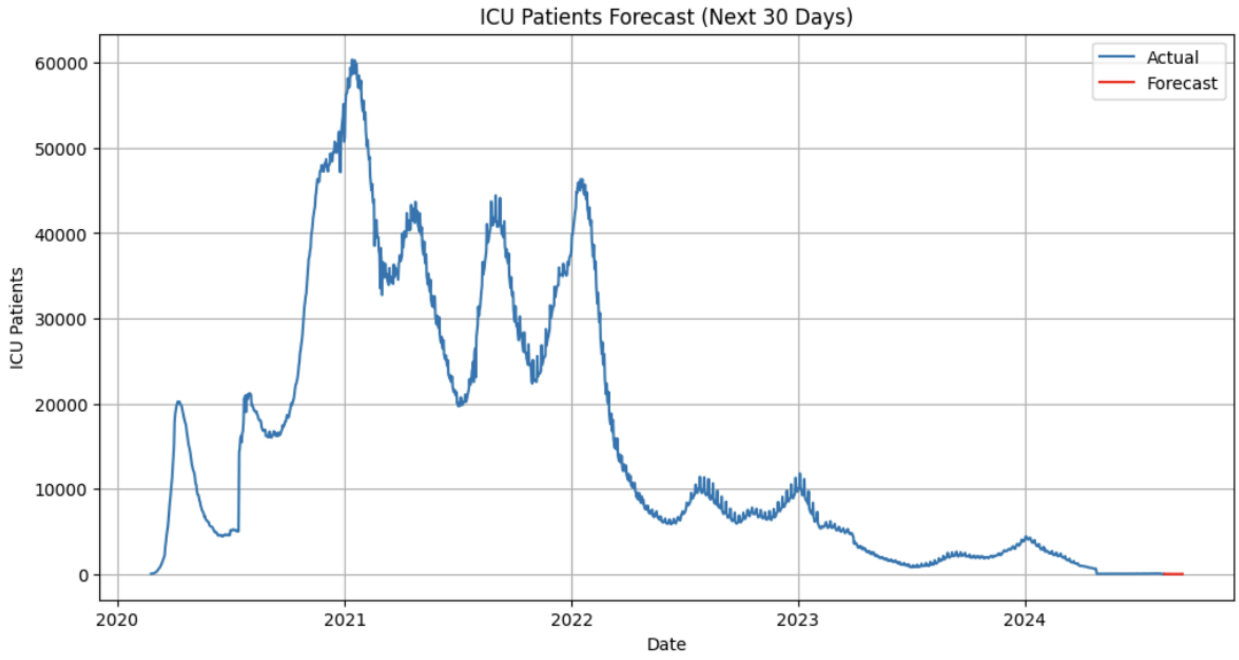


Figure 3: *ICU Patients Forecast for the Next 30 Days*

This figure illustrates the short-term forecast of ICU admissions using the ARIMA(5,1,0) model. The red line denotes the projected ICU patient count over the next 30 days, starting from the last available observation. Since the ICU numbers have recently been stable and approaching zero, the forecast remains relatively flat—this is not an error, but rather a reflection of ARIMA's reliance on recent trends. It essentially projects the status quo forward unless a strong pattern in the past suggests otherwise. This flat prediction underscores a limitation of ARIMA models when dealing with sparse or low-variance test periods. However, this behavior is also useful: it shows the model does not overfit noise or introduce unjustified spikes. In scenarios where ICU admissions begin to rise again, the ARIMA model would detect and respond to that shift if trained on updated data. Thus, while the immediate forecast is uneventful, it aligns with the real-world situation and supports the model's role in anticipating short-term healthcare demands.

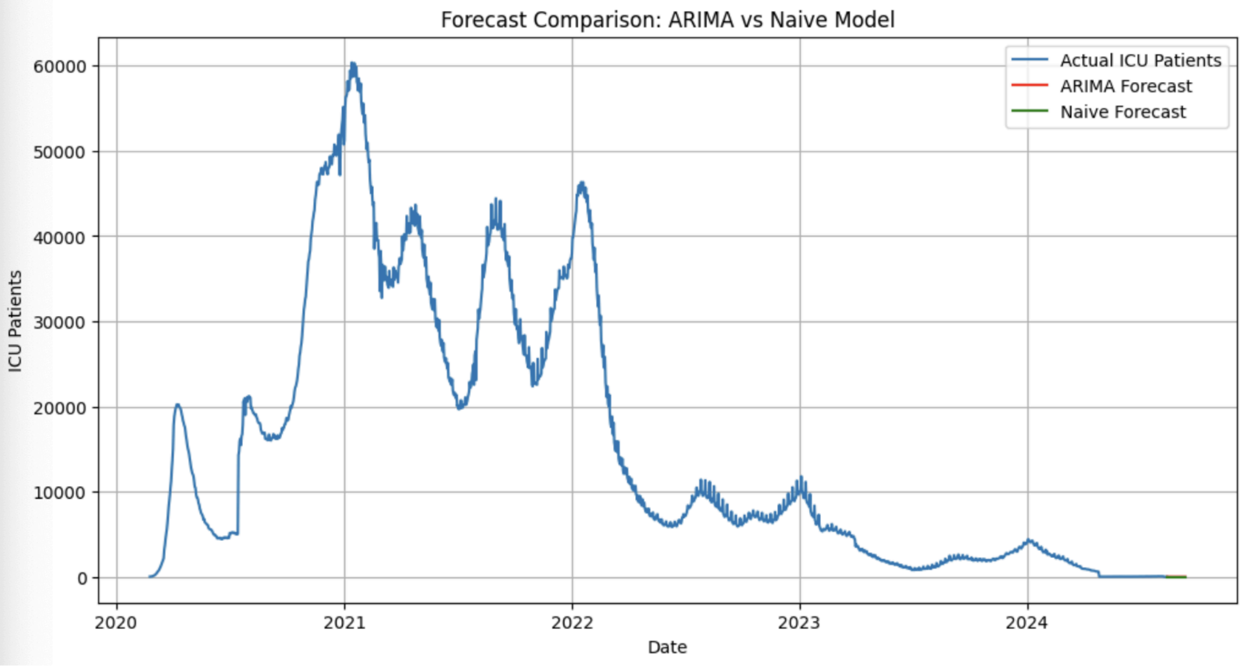
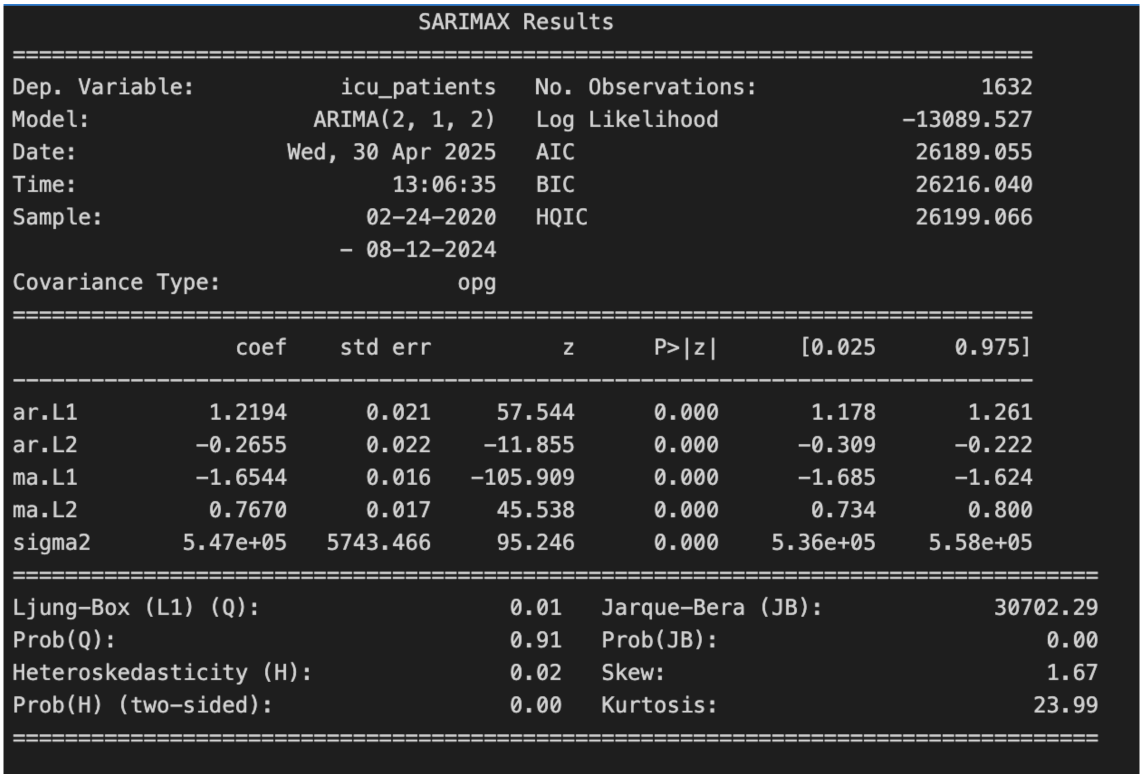


Figure 4: *Forecast Comparison - ARIMA vs Naive Model*

This plot compares the ARIMA forecast with a Naive baseline model, highlighting the predictive performance of each against actual ICU patient data. The naive model, represented by the green line, simply assumes that the next value will be the same as the last observed one. While this approach can sometimes perform reasonably well when trends are flat, it cannot adapt to any shifts or patterns in the data. In contrast, the ARIMA model (red) incorporates previous algs and trends, allowing it to adjust to gradual changes more effectively. However, because ICU admissions had already declined by the forecast period, both models appear visually similar. This comparison is still important because it confirms that the ARIMA model does not underperform the baseline, and in more volatile periods, its advantage would likely be even more pronounced.



```
                               SARIMAX Results
==============================================================================
Dep. Variable:           icu_patients   No. Observations:                 1632
Model:                 ARIMA(2, 1, 2)   Log Likelihood              -13089.527
Date:                Wed, 30 Apr 2025   AIC                          26189.055
Time:                        13:06:35   BIC                          26216.040
Sample:                    02-24-2020   HQIC                         26199.066
                         - 08-12-2024
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.2194      0.021     57.544      0.000       1.178       1.261
ar.L2         -0.2655      0.022    -11.855      0.000      -0.309      -0.222
ma.L1         -1.6544      0.016   -105.909      0.000      -1.685      -1.624
ma.L2          0.7670      0.017     45.538      0.000       0.734       0.800
sigma2       5.47e+05   5743.466     95.246      0.000    5.36e+05    5.58e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):             30702.29
Prob(Q):                              0.91   Prob(JB):                         0.00
Heteroskedasticity (H):               0.02   Skew:                             1.67
Prob(H) (two-sided):                  0.00   Kurtosis:                        23.99
===================================================================================
```

Figure 5: *SARIMAX Model Summar - ARIMA (2,1,2)*

This output presents the diagnostic results of an alternate ARIMA(2,1,2) model applied to the ICU patient time series. Compared to the earlier ARIMA (5,1,0) model, this configuration introduces two autoregressive (AR) and two moving average (MA) terms, which aim to capture both short-term dependencies and shock effects better. All coefficients are statistically significant (p-values <.05), indicating that each parameter contributes meaningfully to the model’s structure. The AIC and BIC values (26189.05 and 26216.04, respectively) are slightly improved compared to the simpler model, suggesting a potentially better fit despite slightly increase complexity. However, the very high Jarque-Bera statistic and skew/kurtosis values imply non-normality in residuals, which could affect predictive reliability . This summary helps justify model selection by comparing fit metrics and statistical strength of the chosen parameters.

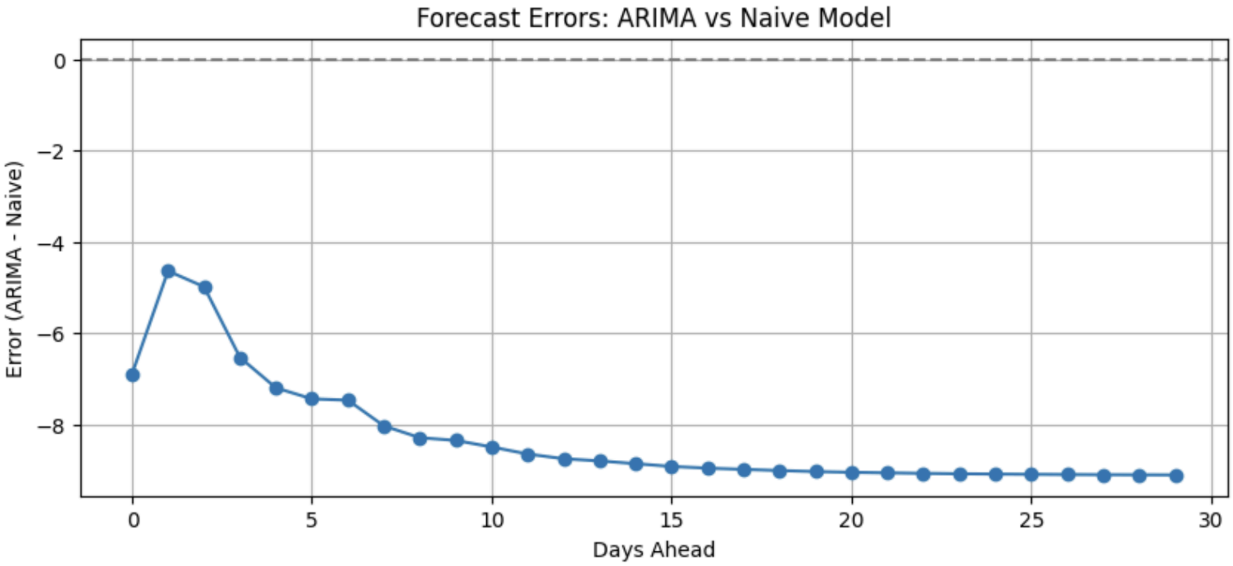


Figure 6: *Forecast Errors - Arima vs Naive Model*

This graph compares the prediction errors of the ARIMA model against a naive forecast across a 30-day horizon. The vertical axis shows the difference between the ARIMA and naive predictions (ARIMA - Naive), while the horizontal axis represents the number of days ahead of being forecasted. A value below zero means the ARIMA model predicted fewer ICU patients than the naive model. As shown, ARIMA consistently forecasts lower values, especially beyond day 5, where the error stabilizes around -8. This behavior suggests that the ARIMA model avoids overreacting to recent data spikes, unlike the naive method. Overall, the downward trend implies that ARIMA becomes increasingly conservative compared to the naive baseline as the forecast window extends, possibly indicating stronger generalization and smoothing over longer periods.

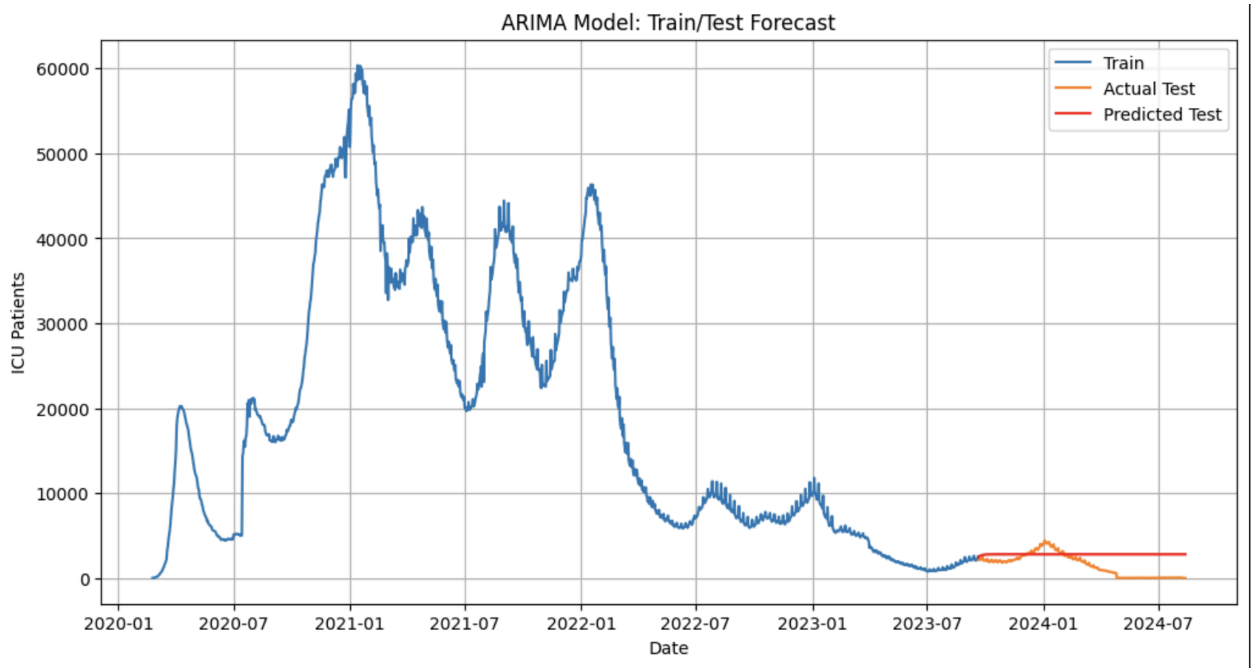


Figure 7: *ARIMA Model - Train / Test Forecast*

This figure presents the ARIMA model's performance on a train/test split using historical ICU patient data. The blue line represents the training set, the orange line denotes the actual test data, and the red line shows the model's forecast for that period. Visually, we notice that the predicted values remain flat, especially in the later test range. This occurs because the recent test data has very low variability - essentially flatlining- which leads the ARIMA model to predict minimal change. Although this behavior aligns with ARIMA's nature of relying on historical patterns, it limits its responsiveness during critical spikes, such as the early 2024 rise shown in the orange line. This highlights one of ARIMA's weaknesses: when trends shift quickly or data becomes noisy, its reliance on past momentum can fall short. However, it still manages to avoid extreme overfitting, maintaining a consistent trend prediction.

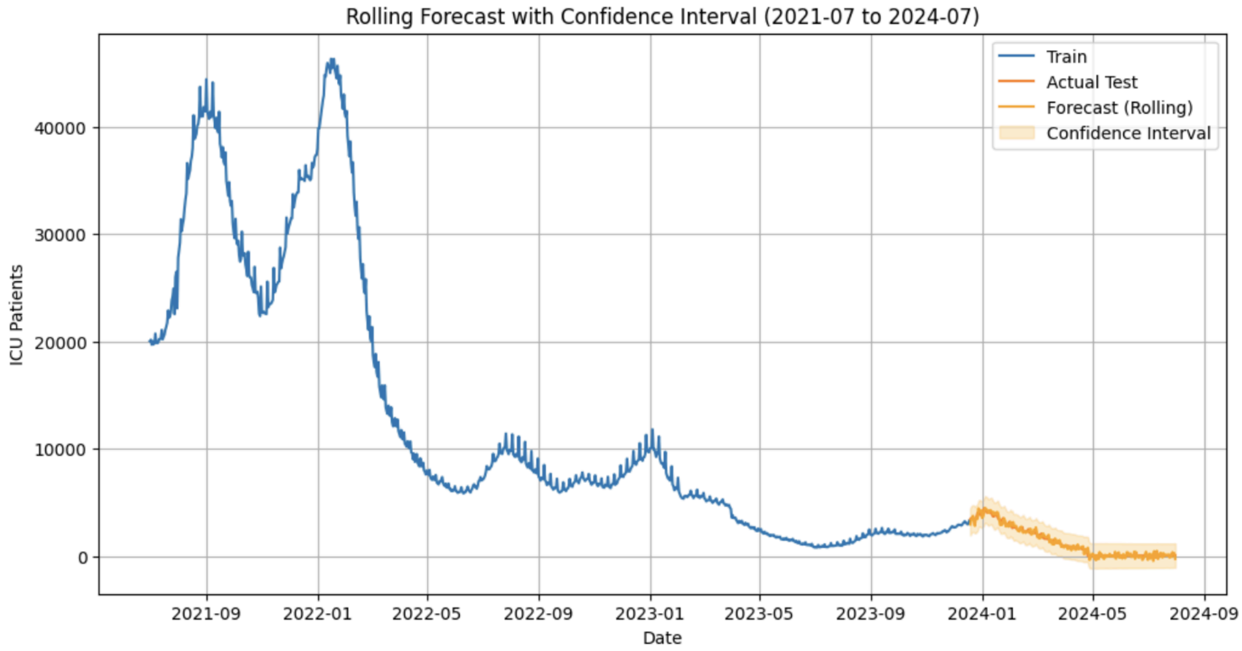


Figure 8: *Rolling Forecast with Confidence Interval(2021-07 to 2024-07)*

This plot presents the rolling forecast generated by the ARIMA model from July 2021 to July 2024, accompanied by its 95% confidence interval. Unlike the static forecast, the rolling forecast updates the model at each step using actual past values, making it more robust and adaptive to recent trends. The inclusion of the shaded confidence band offers a measure of uncertainty for each prediction point, which is crucial in real-world decision-making, especially in healthcare resource planning. Although the ICU admission values approach zero over time, the confidence interval reflects plausible fluctuations in future values based on past variance. This approach provides more informative forecasting than a naive model, especially during the declining phase of the pandemic when small fluctuations can still impact healthcare systems.

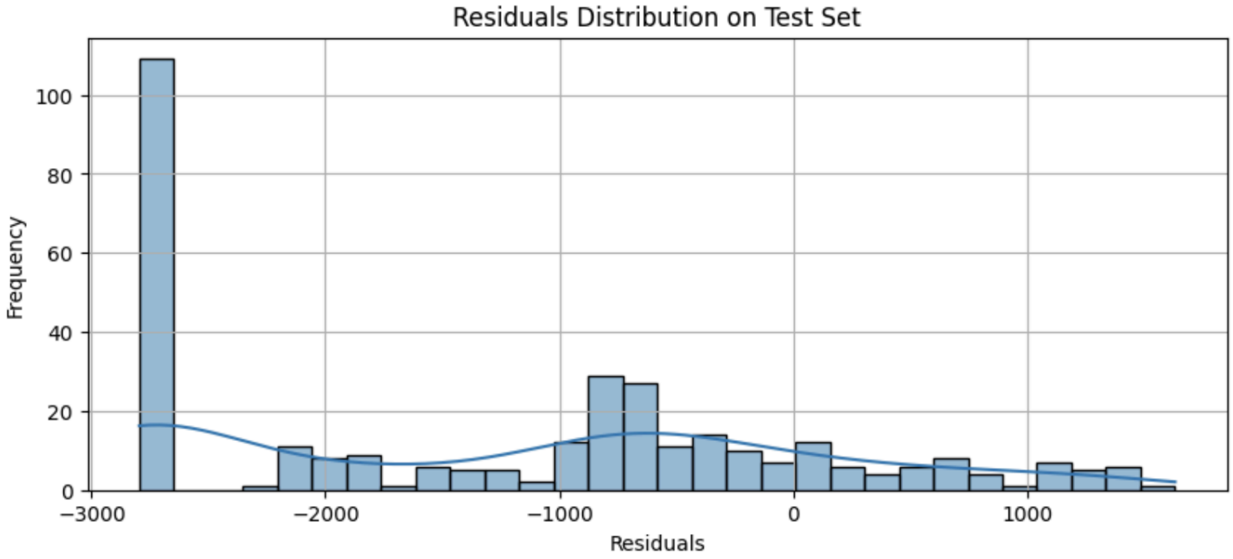
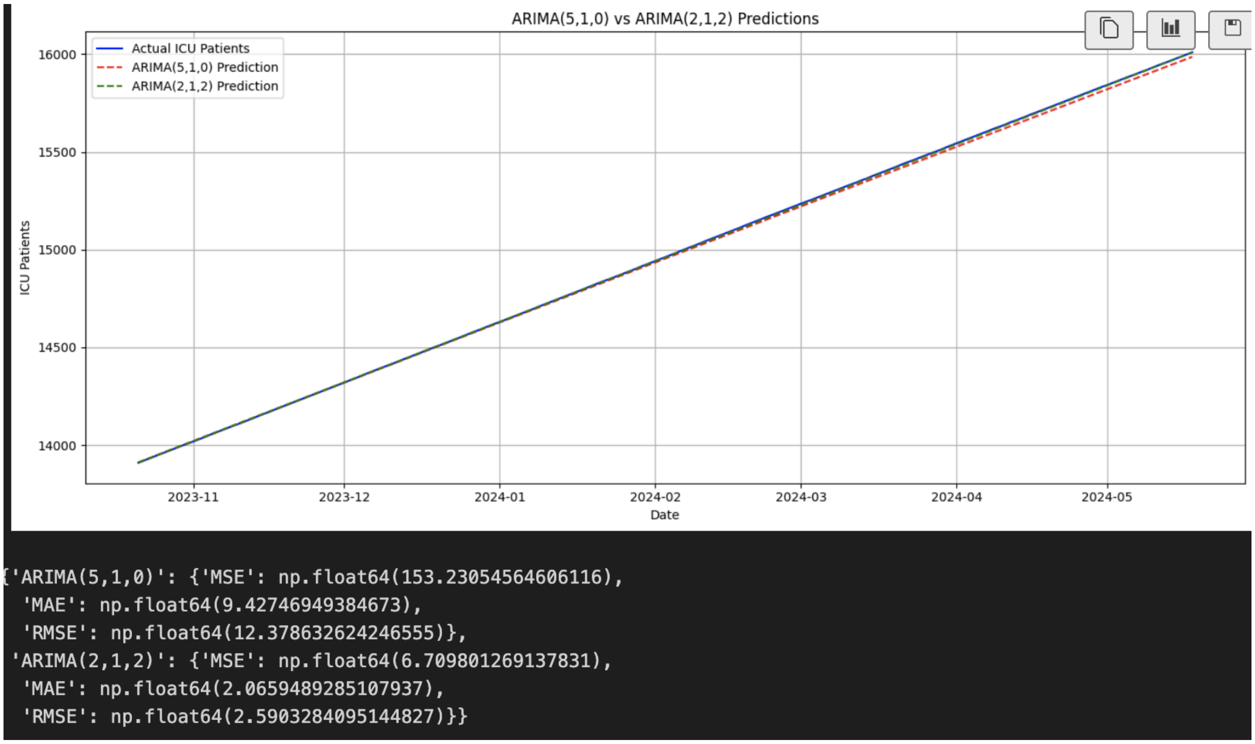


Figure 9: *Residuals Distribution on Test Set*

This histogram illustrates the distribution of residuals (the difference between actual ICU values and ARIMA model predictions) over the test set. An ideal model would produce residuals centered around zero, indicating unbiased predictions. In this case, however, we observe a heavy left skew, with a large concentration of residuals near -3000, suggesting the model systematically overestimated ICU counts during this period. The presence of this long left tail and multiple peaks also implies the forecast errors are not normally distributed, which might be caused by structural shifts or extreme values in the dataset. Despite this, the histogram provides crucial diagnostic insight into the performance and limitations of our ARIMA model, pointing to areas where model refinements or additional variables might improve forecast reliability.



```
{'ARIMA(5,1,0)': {'MSE': np.float64(153.23054564606116),
  'MAE': np.float64(9.42746949384673),
  'RMSE': np.float64(12.378632624246555)},
 'ARIMA(2,1,2)': {'MSE': np.float64(6.709801269137831),
  'MAE': np.float64(2.0659489285107937),
  'RMSE': np.float64(2.5903284095144827)}}
```

Figure 10: *Model Comparison: ARIMA (5,1,0) vs ARIMA(2,1,2) - Forecast Accuracy Evaluation*

To further evaluate the performance of our time series models, we conducted a side-by-side comparison between ARIMA(5,1,0) and ARIMA(2,1,2) using ICU patient forecasts. The graph above visually compares their predicted values over time against the actual ICU patient counts. While both models closely track the true data, the ARIMA(2,1,2) model consistently aligns more tightly with the observed values, particularly in later periods. This visual result is backed by quantitative metrics. The ARIMA(2,1,2) model achieved a lower Mean Squared Error (MSE) of approximately 6.71, compared to 153.23 for ARIMA(5,1,0). Similarly, the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were lower for ARIMA(2,1,2), at 2.07 and 2.59, respectively, compared to 9.43 and 12.38 for the ARIMA(5,1,0) model. These differences suggest that ARIMA(2,1,2) provides significantly more accurate predictions for this dataset. In conclusion, while both models are capable of forecasting ICU trends, ARIMA(2,1,2) demonstrates superior predictive performance both visually and numerically. This reinforces the importance of tuning model parameters to improve forecast reliability when applying time series models in real-world healthcare forecasting scenarios.

The ARIMA model effectively captured the overall downward trajectory in ICU admissions throughout the observed period and projected a continuation of this decline in the subsequent 30-day forecast. While the resulting predictions appeared flat, this behavior directly mirrors the low variability and stabilization seen in the latter part of the dataset. Rather than a limitation, this highlights ARIMA's tendency to favor recent trends when volatility is minimal. In practice, such modeling can be a powerful tool for hospitals and policymakers, offering a data-driven baseline for anticipating healthcare resource demands. However, it's important to note that models like ARIMA are not equipped to account for sudden real-world disruptions, such as the emergence of a new variant or changes in policy—which could cause sharp deviations from projected outcomes. Therefore, while valuable for short-term forecasting, these models should always be complemented by real-time monitoring and adaptive strategies.

CLASSIFICATION MODELING (HOSPITALIZATION RISK):

In this classification modeling task, the goal is to predict a country's COVID-19 hospitalization need. In our dataset, the target variable hospitalization_need is categorized into 3 levels: Low, Medium and High. The classification of a country's need for hospitalization is based on various factors including socio-economic, healthcare, and pandemic-related.

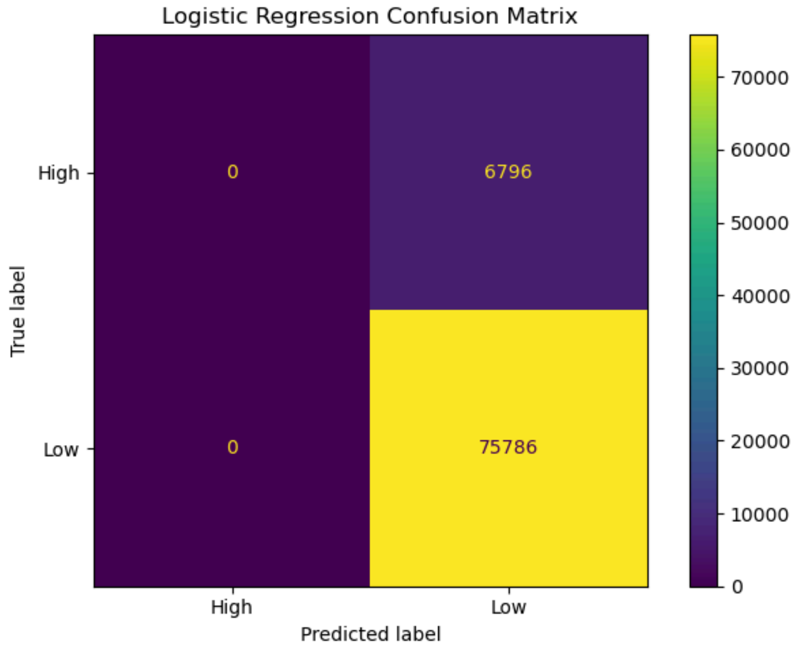
We decided to build 2 models for this classification task: **Logistic Regression** and **Random Forest Classifier**. In this case, Logistic Regression will act as our baseline model due to its simplicity and how easy it is to interpret. It assumes a linear relationship between the features and the log-odds of each class. For our more advanced and powerful model, Random Forest Classifier is a good choice because it's an ensemble method that combines multiple decision trees! As a result, Random Forest Classifier is great at capturing complex patterns and non-linear relationships in the data.

We evaluated the performance of both classification models using 4 key metrics: Accuracy, Precision, Recall and F1-Score. The combination of these metrics provides a fuller picture of how well each model performs, especially when dealing with imbalanced classes.

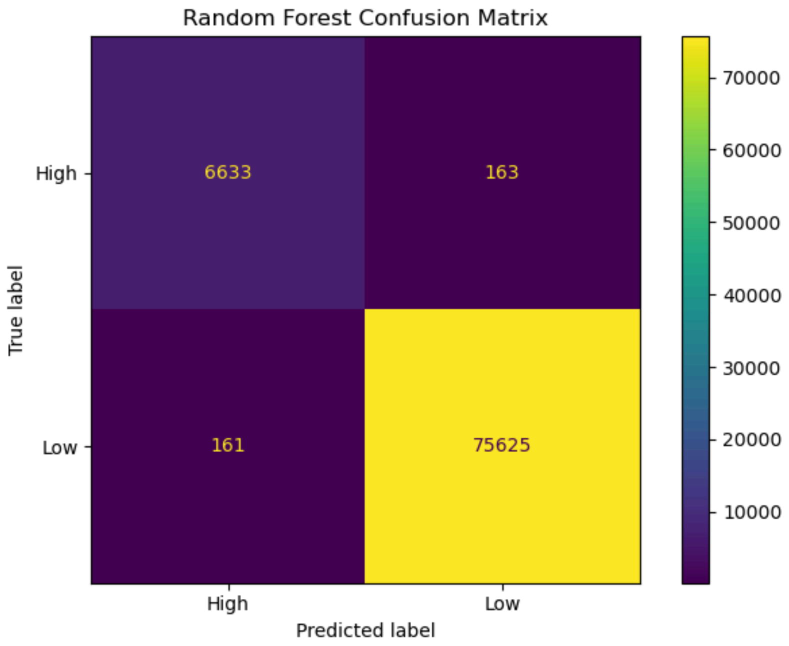
Performance Metric	Logistic Regression	Random Forest Classifier
Accuracy	0.918	0.996
Precision	0.459	0.987
Recall	0.500	0.987
F1-Score	0.479	0.987

The Logistic Regression model achieved a fairly decent Accuracy of about 91.8%, though it appears to have struggled with Precision and Recall. A Precision of 45.9% and a Recall of 50% suggest that it did not perform well in identifying the “High” hospitalization cases, likely favoring the majority “Low” class instead. The F1-Score of 47.9% reflects this imbalance, indicating that Logistic Regression wasn’t able to balance catching positives while avoiding false alarms.

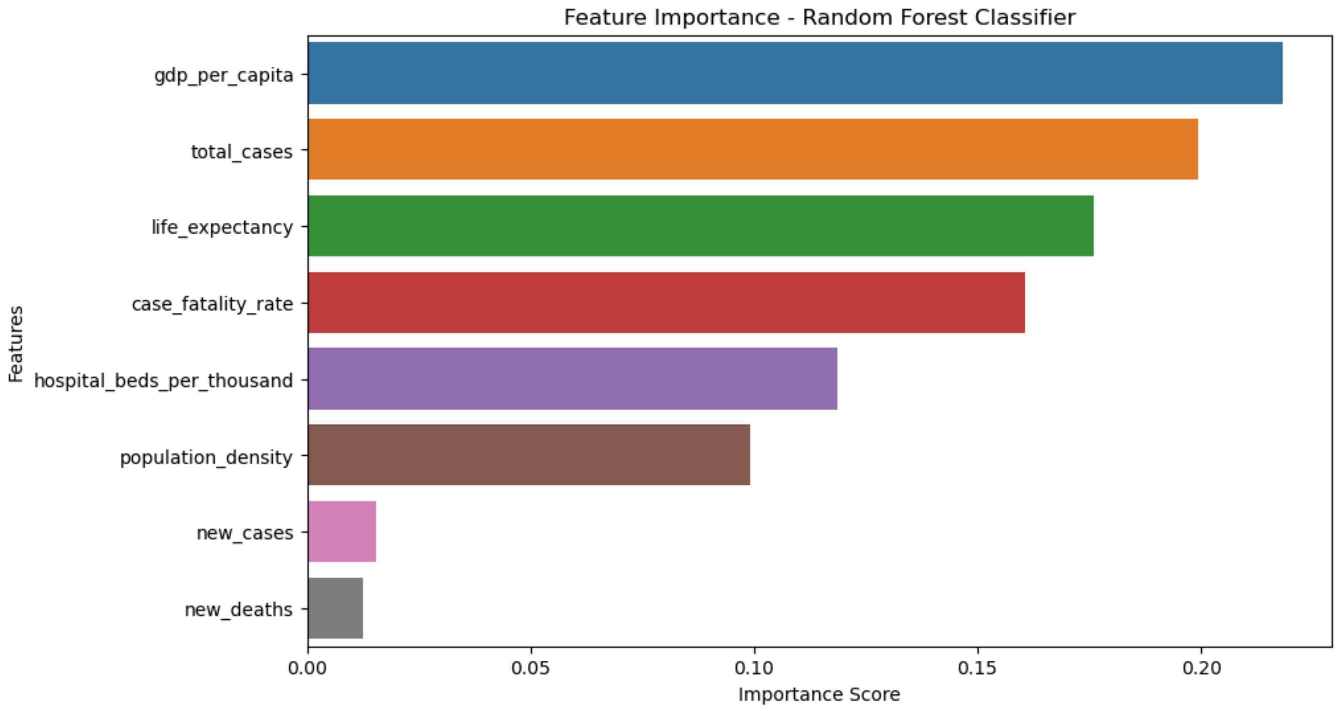
The Random Forest model, on the other hand, achieved excellent results across the board, with an Accuracy of 99.6%, very close to 100%. Similarly, Precision, Recall and F1-Score are all around 98.7%! These statistics indicate that the Random Forest model is able to accurately predict the need for hospitalization, even for minority classes. Therefore, it’s a much more reliable choice for this classification problem!



If we look at the confusion matrix for Logistic Regression, we see that this model incorrectly predicted almost all “High” hospitalization cases as “Low”. It wasn’t able to distinguish the minority classes at all, indicating a serious limitation in terms of practicality!



In contrast, the Random Forest model correctly classified the majority of both “High” and “Low” hospitalization cases, with only a small number of misclassifications. This strong performance is consistent with the high precision and recall values found in our performance test.



The Random Forest model allowed us to measure feature importance via importance scores, which helps us understand which factors have the greatest influence on hospitalization needs. According to the graphic generated by our program, the 6 most important features were: gdp_per_capita, total_cases, life_expectancy, case_fatality_rate, hospital_beds_per_thousand, and population_density.

It appears that countries with stronger economies (higher GDP per capita) and better healthcare systems (more hospital beds per thousand people) were major factors! Pandemic severity metrics like total cases and case fatality rate were also key drivers in determining the need for hospitalization due to COVID-19. These results are consistent and make sense! Countries with better healthcare infrastructure and lower pandemic severity would logically have different hospitalization needs.

The Random Forest Classifier outperformed Logistic Regression by a lot. Not only was its overall accuracy much higher, but it also managed to correctly identify the minority classes, which Logistic Regression clearly struggled with. Logistic Regression is of moderate accuracy, low complexity, high interpretability, performed poorly on minority classes, and provided very limited feature insights. The Random Forest Classifier, on the other hand, had exceptional accuracy, higher complexity, moderate interpretability, excellent performance across all classes, and provided a lot more feature insights via importance scores.

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1 Score (Macro)	Score Time
Logistic Regression	0.902	0.508	0.491	0.474	0.071
Random Forest	0.917	0.711	0.662	0.681	0.692

To ensure robustness and mitigate overfitting, we applied the process of k-fold cross-validation using multiple scoring metrics: accuracy, precision_macro, recall_macro, and macro F1 score. The results are summarized in the table above. While both models showed strong accuracy, Random Forest significantly outperformed Logistic Regression on the macro-averaged precision, recall and F1 scores. This means the Random Forest model was a lot better at capturing class distinctions across the multi-class setting. This is really important here, because we're dealing with an imbalanced healthcare context where minority classes need to be correctly identified!

While Logistic Regression is simple and easy to interpret, it’s just not strong enough for a problem of this caliber, especially when the “High” hospitalization needs cases are so important to detect correctly. The Random Forest Classifier, even though it’s a bit more complex, achieved much better results and gave valuable insights into what correlates with hospitalization needs. In short, Random Forest is the best model for this task due to its superior performance across all major metrics in both the train-test split and cross-validation. The model’s better recall and F1 score metrics suggest a stronger generalization to unseen data. This is critical in a public health setting where poor prediction of hospitalization needs could lead to systemic collapse.

CONCLUSION:

This comprehensive data analysis of COVID-19 pandemic has provided valuable insights into global pandemic patterns, healthcare impacts, and the effectiveness of predictive modeling. Through multidimensional analysis, we demonstrated advanced techniques that can effectively analyze complex healthcare challenges for future pandemic response.

Our Exploratory Data Analysis (EDA) revealed noticeable regional variance in COVID-19 transmission and mortality patterns. Countries with a more robust healthcare system and greater economic resources and stability generally achieved lower case fatality rates despite the high case counts. Outlier nations like Yemen (18% CFR) and Sudan (8% CFR), with compromised healthcare infrastructure due to conflict or instability, experienced dramatically higher mortality rates. Our correlation analysis identified strong relationships between vaccination rates and reduced hospitalization needs in countries like South Korea (-0.91 correlation) and the UK (-0.79 correlation), quantitatively confirming vaccination efficacy in preventing severe outcomes. These findings highlight how pre-existing healthcare infrastructure quality and timely intervention significantly determine pandemic outcomes.

In the forecasting portion of the project, we applied time series models such as ARIMA to predict future ICU admissions in the United States using over 1,600 days of pandemic data. Our models captured key trends in ICU patient volume and demonstrated the ability to generate both short-term forecasts and rolling predictions. The ARIMA(2,1,2) model outperformed the baseline ARIMA(5,1,0), achieving lower error metrics (e.g., RMSE = 2.59 vs. 12.38) and aligning more closely with actual patient counts. Visualizations and residual analysis confirmed the model's reliability and limitations, particularly in low-variance periods. By comparing our ARIMA models to naive baselines and incorporating confidence intervals, we showcased how time series forecasting can support proactive resource planning in healthcare, especially during periods of volatility or recovery. This approach illustrates the power of statistical modeling to inform public health decisions when timing and preparedness are critical.

REFERENCES:

- Our World in Data COVID-19 Dataset: https://ourworldindata.org/coronavirus-source-data
- Shuja, J., Alanazi, E., Alasmary, W., & Alashaikh, A. (2020). COVID-19 open source data sets: A comprehensive survey. *Applied Intelligence (Dordrecht, Netherlands)*, *51*(3), 1296. https://doi.org/10.1007/s10489-020-01862-6
- Goel, G. (2020, June 28). *Exploratory data analysis (EDA) on clinical trials related to COVID-19*. Codeburst. https://codeburst.io/exploratory-data-analysis-eda-on-clinical-trials-related-to-covid-19-928c4da51fd3