Tri Lam (1916079)

November 23, 2024

COSC 3337 Group Project Report: EDA on COVID-19 Dataset

1. Introduction

The COVID-19 dataset used in this project is provided by Our World in Data (OWID). While the dataset is updated daily on OWID’s GitHub repository, we recommend using the version included with this report for consistent results.

This project explores three tasks to analyze COVID-19's global impact: identifying the most affected countries, evaluating testing rates among the most populated nations, and comparing case growth between two countries. Each task includes descriptive statistics, visualizations, and machine learning applications like regression and outlier detection.

The analysis uses pandas, matplotlib.pyplot, and seaborn. To focus on countries, we filtered out OWID-defined regions (e.g., continents) using the str.contains function. Task-specific variables were created for analysis, which may differ from the dataset's original variables.

2. Task 1: Impact of COVID-19 on Countries (Cases and Deaths)

<u>Objective:</u>

While it is undeniable that most, if not all, countries are dealing with this pandemic, it is also true that some countries are more impacted than others. By analyzing total cases and deaths, we aim to identify the 10 countries most affected by COVID-19.

<u>Analysis: (Tri)</u>

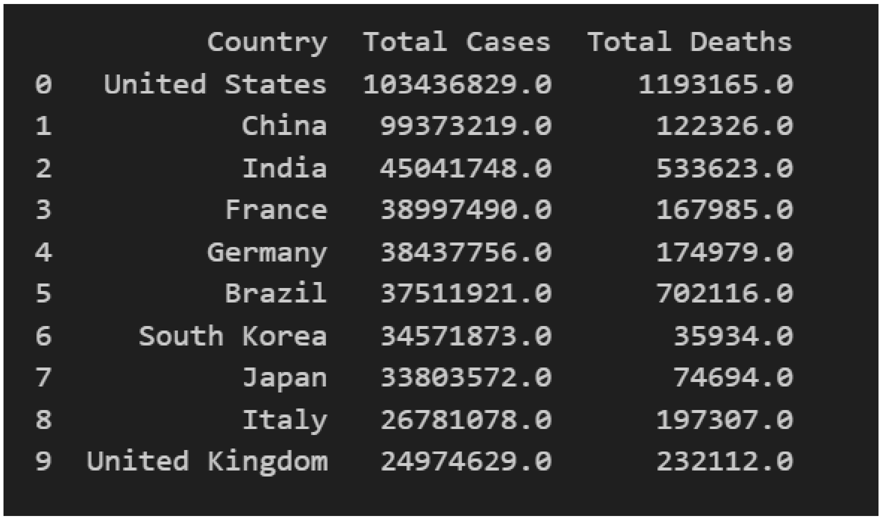
We analyzed the total cases and deaths for each country by aggregating daily reported values using the groupby function. The results were sorted in descending order to identify the top 10 most affected countries. These metrics were stored in a new DataFrame for easier visualization.

To effectively compare both metrics, the data was restructured using the melt function, allowing the creation of a side-by-side bar plot. The visualization, generated using the Seaborn library, highlights significant disparities in the pandemic's impact across countries.

- **Key Insights**:
  - The United States ranks first in both total cases (over 100 million) and deaths (exceeding 1 million).
  - China, despite having the second-highest total cases, has a relatively low death toll compared to other countries.
  - Brazil presents a contrasting trend with the second-highest death count but ranks sixth in total cases.

These findings reflect how differences in healthcare systems, government policies, and population dynamics shaped the pandemic's impact across nations.

Below are the results and visualization:

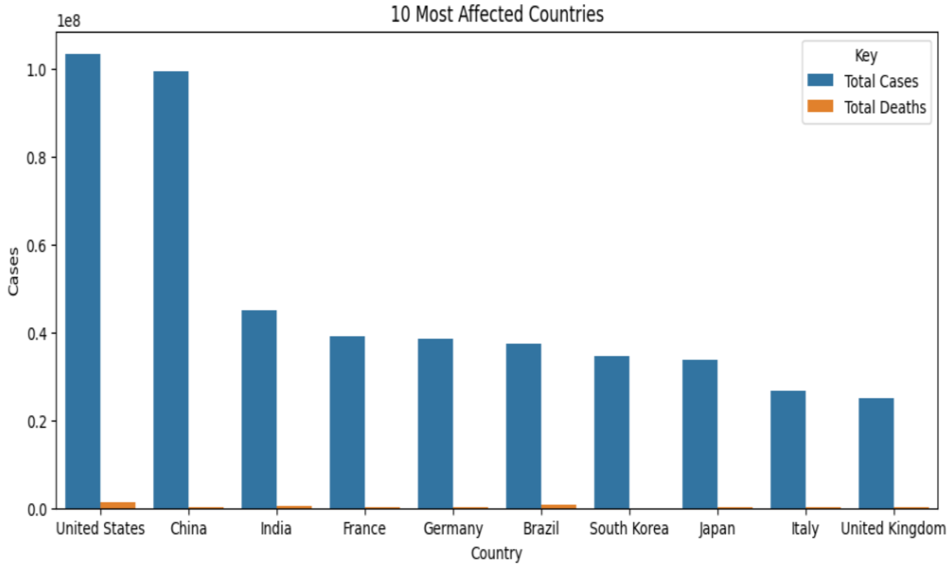


```
          Country   Total Cases   Total Deaths
0   United States  103436829.0      1193165.0
1           China   99373219.0       122326.0
2           India   45041748.0       533623.0
3          France   38997490.0       167985.0
4         Germany   38437756.0       174979.0
5          Brazil   37511921.0       702116.0
6     South Korea   34571873.0        35934.0
7           Japan   33803572.0        74694.0
8           Italy   26781078.0       197307.0
9  United Kingdom   24974629.0       232112.0
```

Visualization (Bar Plot):

The bar plot below visualizes the total cases and deaths for each of the top 10most affected countries:

The plot shows the stark contrasts between total cases and deaths across the selected countries, emphasizing the varying severity of the pandemic's impact globally.



3. Task 2: Testing Rates Across the Most Populated Countries

Objective:

To evaluate the testing vigilance of the 10 most populated countries, we calculated the testing rate as the ratio of total tests to population. Since OWID stopped updating testing data on June 23, 2022, the dataset was filtered to include data before this date.

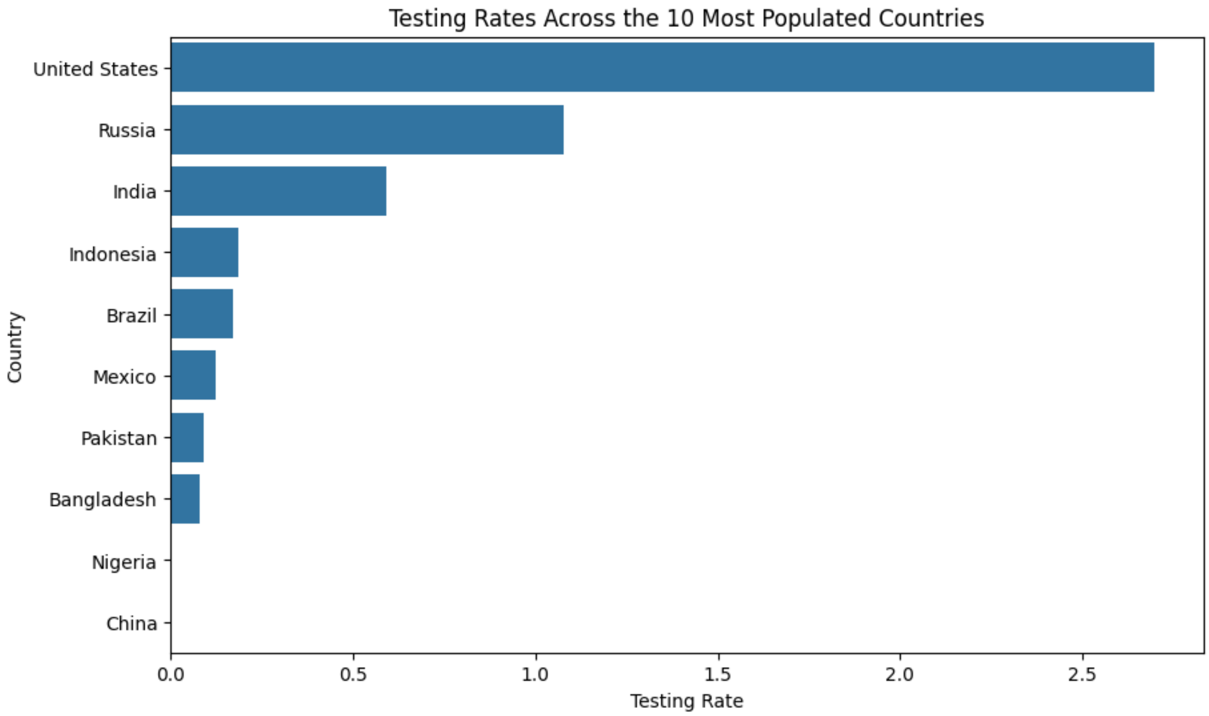
Analysis: (Tri)

Using the groupby function, the latest population and total tests for each country were aggregated. The top 10 most populated countries were identified based on their population. Testing rates for these countries were computed using the formula: Testing Rate=Total TestsPopulation\text{Testing Rate} = \frac{\text{Total Tests}}{\text{Population}}Testing Rate=PopulationTotal Tests

The analysis revealed notable disparities:

- The United States had the highest testing rate of 2.70, followed by Russia at 1.08.
- China’s testing rate was recorded as 0, likely due to missing data in OWID’s dataset.
- Nigeria recorded a very low testing rate of 0.0032, reflecting significant disparities in testing capacity.

These discrepancies underscore the challenges in maintaining consistent testing data, as highlighted by OWID’s discussions on shifting priorities and reporting practices during the pandemic. Variations in testing rates directly impact case detection rates and public health interventions.



Visualization:

The bar plot below displays the testing rates for the 10 most populated countries, emphasizing the uneven distribution of testing efforts:

This visualization highlights key disparities in testing efforts. While the United States and Russia show high vigilance, countries like China and Nigeria demonstrate lower rates, influenced by factors such as data availability, resources, and shifting pandemic priorities.

4. Task 3: Comparing Rates of Increase Between Two Countries

Objective:

The two countries selected for comparison are the United States and China, identified in Task 1 as the most affected by COVID-19. This task aims to analyze their rates of case increases throughout 2023 to understand how both countries have managed the pandemic during this period.

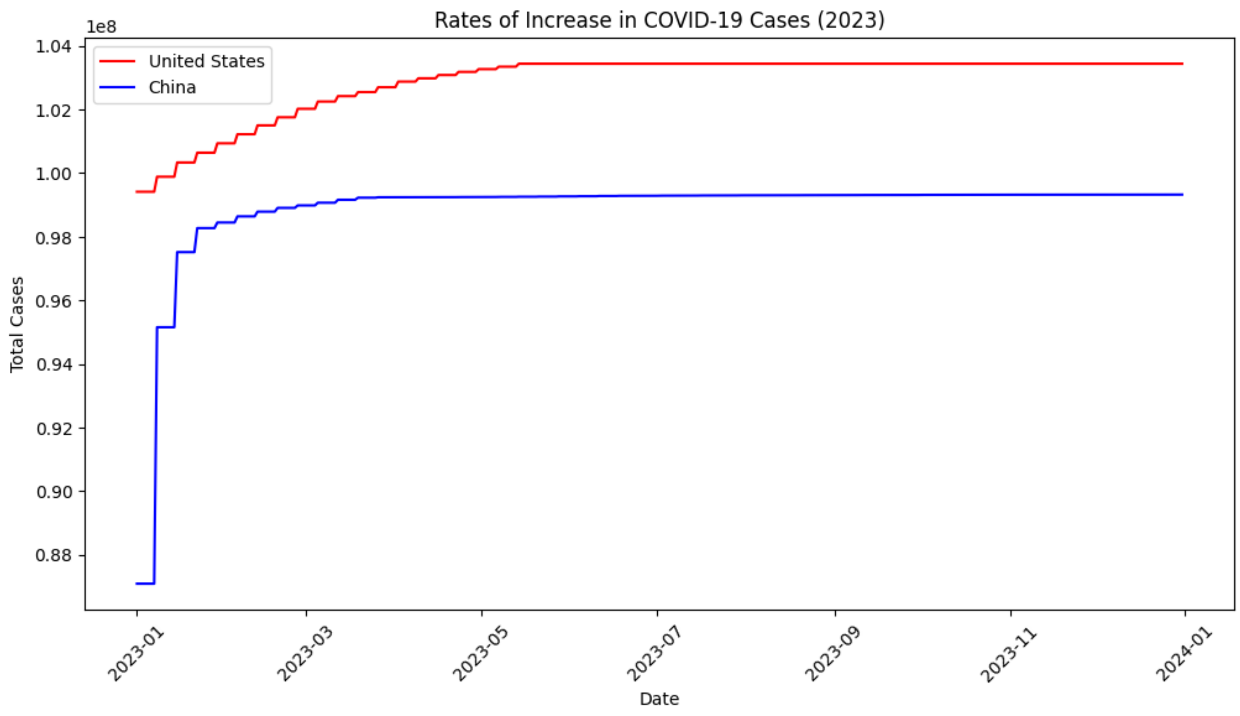
Analysis: (Tri)

Using the loc function, data for the United States and China was extracted to isolate records specific to each country. The dataset was filtered to include only entries from 2023, retaining columns for dates and cumulative case counts. These filtered datasets were merged into a single DataFrame using an inner join on the date column, allowing for direct comparison of daily case counts.

The merged data was visualized using a line plot, with distinct colors representing each country for clarity. The results showed:

- At the beginning of 2023, the United States had approximately 99.5 million cases, while China had around 86 million cases.
- By mid-March, China's case count surged to roughly 99 million, while the United States increased to approximately 103 million by early May.
- Following these rapid increases, both countries exhibited relatively stable case counts for the remainder of the year.

This stabilization suggests a plateau in COVID-19 spread in both regions during the latter half of 2023, with China adding only about 600 additional cases by the year's end.



Visualization:

The line plot below illustrates the cumulative case counts for the United States and China in 2023, highlighting their trajectories:

This comparison underscores the differences in the trajectory of COVID-19 between the two countries, shaped by factors such as population size, testing policies, and public health interventions.

5. Task 4 Outliers (Tri)

Objective:

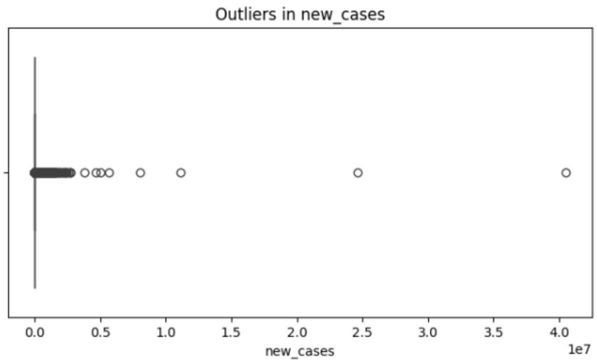
This task involves identifying and analyzing outliers within the dataset to uncover potential data irregularities or extreme values that could impact the results of our analysis.

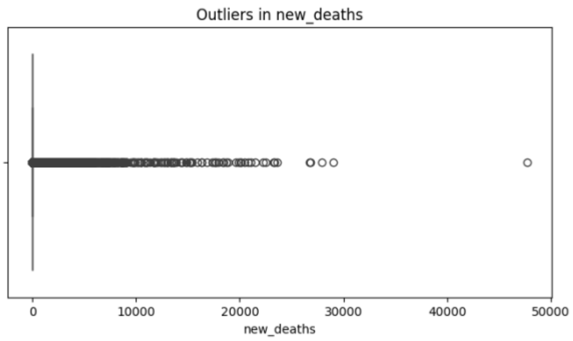
Analysis:

Statistical methods, including the interquartile range (IQR) and Z-scores, were used to detect outliers in key metrics such as daily cases (new_cases) and daily deaths (new_deaths).

- **Outliers in new_cases**:
    - Represent days with abnormally high reported case counts.
    - Often caused by data reporting delays where multiple days' cases are aggregated into one.
- **Outliers in new_deaths**:
    - Highlight extreme death counts, potentially due to irregularities in healthcare systems or localized outbreaks.

The presence of these outliers emphasizes the importance of addressing inconsistencies in data reporting. Left unchecked, such anomalies can distort analyses, misinform public health decisions, and lead to inefficient resource allocation.





Visualizations:

The box plots below illustrate the presence of outliers for new_cases and new_deaths, with individual points beyond the whiskers representing extreme values

Insights:

- **Impact on Public Health Decisions**:
    - Overestimated cases or deaths can lead to unnecessary resource allocation or panic.
    - Underreported values may result in delayed interventions, exacerbating the crisis.
- **Real-World Context**:
    - Some outliers reflect actual surges, such as during localized outbreaks or holidays when reporting schedules change.
    - Others highlight systemic issues, such as delayed aggregation or inconsistencies in testing and reporting methods.

By identifying and analyzing these anomalies, public health agencies can better interpret the data, refine their strategies, and allocate resources effectively. Addressing outliers ensures the reliability of insights derived from the analysis.

6. Task 5 Regression((Tri)

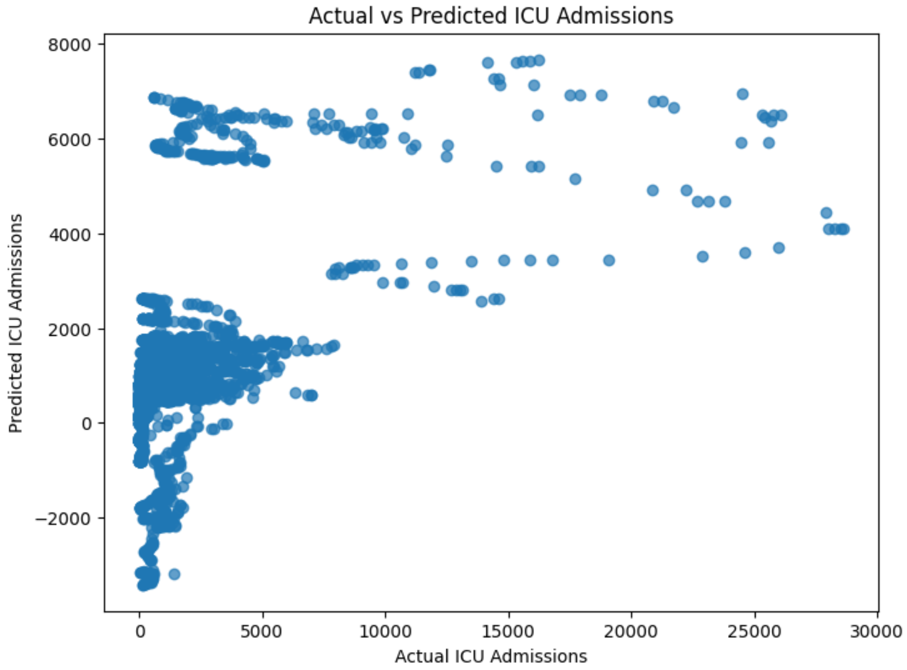
<u>Objective:</u>

This task focuses on using regression analysis to predict healthcare needs, specifically ICU admissions, based on historical COVID-19 data. The goal is to model relationships between key features and outcomes to identify trends and improve decision-making.

Analysis:

Linear regression models were applied to predict ICU admissions (icu_patients) using features such as total_cases and total_deaths. The dataset was prepared by filtering rows with non-null ICU admission values and splitting the data into training and testing sets.

- **Evaluation Metrics**:
    - R-squared: Assessed the proportion of variance explained by the model, measuring how well the independent variables predict the dependent variable.
    - Mean Squared Error (MSE): Measured the average squared difference between predicted and actual ICU admissions.
- **Findings**:
    - The scatter plot below compares actual vs. predicted ICU admissions. While the model captures general trends, noticeable deviations suggest areas for further improvement, such as incorporating additional features like testing rates or healthcare infrastructure data



○ .

Visualization:

The scatter plot demonstrates the model's performance, with actual ICU admissions plotted on the x-axis and predicted ICU admissions on the y-axis:

Insights:

The regression model provided a foundation for understanding ICU admission trends. However, the observed variability indicates several areas for improvement:

- Model Complexity: Incorporating advanced machine learning models (e.g., Random Forests or Gradient Boosting) could better capture non-linear relationships.
- Additional Variables: Including features such as vaccination rates, hospital capacity, and testing rates may enhance the model's accuracy.
- Data Quality: Addressing outliers and missing values could reduce prediction errors.

By refining the model and expanding its feature set, this analysis could provide more actionable insights for public health planning and resource allocation.

7. Task 6: Hospitalization Metrics

Objective:

This task focuses on calculating and categorizing hospitalization needs based on hospitalization rates to understand the burden of COVID-19 on healthcare systems.

Analysis:

The hospitalization rate was calculated as the percentage of total cases requiring hospitalization:

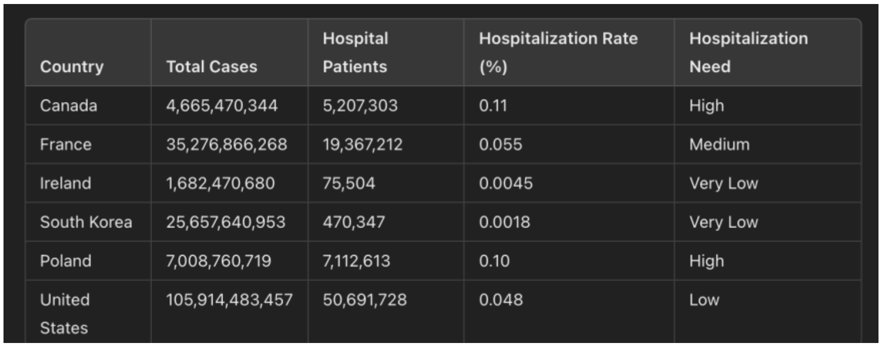
Hospitalization Rate =(Hospital Patients/Total Cases)×100

Countries were then categorized into "Very Low," "Low," "Medium," and "High" hospitalization needs based on statistical thresholds.

Key Findings:

- Most countries fell into the "Low" or "Very Low" categories, indicating relatively low hospitalization rates compared to total cases.
- Countries such as **Canada**, **Poland**, and **South Africa** exhibited "High" hospitalization needs, with rates exceeding 0.1%.
- In contrast, countries like **Ireland**, **South Korea**, and **Norway** demonstrated "Very Low" hospitalization rates, reflecting better containment or lower reported hospitalizations relative to cases.

These results suggest substantial disparities in hospitalization rates, potentially influenced by variations in healthcare systems, data reporting practices, and pandemic management strategies.



Country	Total Cases	Hospital Patients	Hospitalization Rate (%)	Hospitalization Need
Canada	4,665,470,344	5,207,303	0.11	High
France	35,276,866,268	19,367,212	0.055	Medium
Ireland	1,682,470,680	75,504	0.0045	Very Low
South Korea	25,657,640,953	470,347	0.0018	Very Low
Poland	7,008,760,719	7,112,613	0.10	High
United States	105,914,483,457	50,691,728	0.048	Low

Insights:

This categorization provides critical insights into the strain placed on healthcare systems during the pandemic. Countries with "High" hospitalization needs may have experienced more severe outbreaks or limited healthcare capacity, while those with "Very Low" needs likely benefited from robust containment measures or underreporting.

8. Conclusion (Tri)

In this project, we analyzed the impact of COVID-19 using data from Our World in Data. Through exploratory data analysis and machine learning techniques, we achieved the following:

- Identified the countries most affected by the pandemic based on total cases and deaths.
- Evaluated testing rates among the most populated nations, revealing significant disparities in testing efforts.

- Compared the rates of case increases in 2023 between the United States and China, uncovering unique trajectories influenced by public health policies.

The outlier analysis highlighted irregularities in reported daily cases and deaths, emphasizing the need for consistent and accurate data reporting. Regression modeling demonstrated the potential for predicting healthcare demands, providing insights that can aid in public health planning.

Our findings emphasize the critical role of robust data collection and analysis in managing global health crises. These insights offer valuable lessons for improving responses to future pandemics, ensuring better preparedness and decision-making.

9. References

1. OWID. "Covid-19-Data/Public/Data at Master · Owid/Covid-19-DATA." *GitHub*, https://github.com/owid/covid-19-data/tree/master/public/data. Accessed 23 Nov. 2024.

2. OWID. "Ending Our Covid-19 Testing Data Updates · Owid COVID-19-Data · Discussion #2667." *GitHub*, 31 May 2022, https://github.com/owid/covid-19-data/discussions/2667. Accessed 23 Nov. 2024.