Unveiling the Invisible Enemy: A Big Data Journey Through COVID-19 Mortality

Team 8, GMU ECE552 Big Data Technologies, Spring 2024

Vrushali Vijay Patil
vpatil2@gmu.edu

Ethan Cheng
ycheng26@gmu.edu

Abstract: This study delves into the analysis of COVID-19 death ratios and related trends, drawing insights from data spanning various countries and states. Leveraging a dataset stored in a Parquet file, named 'all_weekly_excess_deaths.parquet', we meticulously calculated death ratios for each country, defined as the proportion of COVID-19 deaths to total deaths, thereby offering a comprehensive view of the pandemic's impact globally. Temporal variations in COVID-19 mortality rates, elucidated through detailed tables and visualizations, highlighted the dynamic nature of the pandemic. Moreover, correlations between population size, death rates, and other influencing factors were explored, shedding light on the intricate dynamics shaping pandemic outcomes.State-level analyses within the United States further enriched our understanding, revealing disparities in case fatality rates and temporal patterns across different regions. Noteworthy observations included the disproportionate impact of densely populated areas and the importance of healthcare capacity in mitigating mortality rates.

Keywords—PySpark, Covid-19, Big data

I. INTRODUCTION

The COVID-19 pandemic, an invisible enemy, has swept across the globe, causing unprecedented disruption and mortality. The project aims to delve into the depths of this global crisis using the power of big data. The dataset used for this project is sourced from Kaggle, providing a comprehensive record of COVID-19 deaths worldwide.Before the advent of big data, understanding the intricacies of a global pandemic was a daunting task, often requiring the manual collection and analysis of data. The idea behind this project is to simplify this process by leveraging big data technologies, specifically Apache Spark, to analyze, visualize, and interpret the vast amounts of data related to COVID-19 mortality.The COVID-19 mortality dataset is not unique to this project; it’s been used in many studies across the world. It’s a great example of how technology can be used to streamline access to vital information and improve our understanding of this global crisis. COVID-19 mortality data can be accessed and analyzed in many ways, such as through web-based platforms, APIs, and other data services. This project will demonstrate the power and potential of big data in unraveling the complexities of COVID-19 mortality, ultimately contributing to the global fight against this invisible enemy.

II. OBJECTIVE

This report will provide insights into global and US-specific trends in COVID-19 mortality, potentially informing public health interventions and resource allocation strategies.

1. Analyze and compare the COVID-19 death ratio (total covid deaths/total non covid deaths) across 46 countries for the period December 28th, 2019 to July 17th, 2022 (weekly data).

2. Identify the month with the highest COVID-19 mortality (excess deaths) for each of the 46 countries.

3. Explore the relationship between population size and COVID-19 mortality (excess deaths) across the 46 countries.

4. Determine the month with the highest COVID-19 cases for each state within the United States (data covers January 21st, 2020 to September 29th, 2021).

5. Calculate the COVID-19 mortality rate (number of deaths divided by the number of cases) for each state within the US.

6. Analyze the distribution of the Case Fatality Rate (CFR) (proportion of deaths among confirmed cases) across different counties in the United States.

III. LITERATURE REVIEW

The COVID-19 pandemic has presented unprecedented public health challenges worldwide. Several studies have been conducted to understand the factors contributing to COVID-19 mortality. One such study used big data technologies to analyze the spread and impact of the virus[1]. This study provides valuable insights into the spread and impact of the virus. Another study highlighted the application of big data and artificial intelligence in COVID-19 prevention and control. The st udy emphasized the role of big data analysis in understanding the spread of the virus and implementing effective control measures[2]. These studies demonstrate the potential of big data in understanding and predicting COVID-19 mortality. They provide a solid foundation for your project, 'Unveiling the Invisible Enemy: A Big Data Journey Through COVID-19 Mortality'. Your project, which will use the dataset provided on Kaggle, will contribute to this growing body of research by providing further insights into COVID-19 mortality at the county level in the US and weekly excess deaths worldwide. The use of big data technologies, specifically Apache Spark, will enable you to analyze, visualize, and interpret the vast amounts of data related to COVID-19 mortality, ultimately contributing to the global fight against this invisible enemy.

IV. DATA REVIEW

The dataset "all_weekly_excess_deaths .parquet" contains weekly data for 46 countries from December 28, 2019, to July 17, 2022. The data includes the following fields:

- country: The name of the country.
- region: The specific region within the country.
- region_code: The code representing the region.
- start_date and end_date: The start and end dates of the week.
- days: The number of days in the week.
- year and week: The year and week number.
- population: The population of the region.
- total_deaths: The total number of deaths in the region during the week.
- covid_deaths: The number of deaths due to COVID-19.
- expected_deaths: The expected number of deaths based on historical data.
- excess_deaths: The number of deaths above the expected number.
- non_covid_deaths: The number of deaths not related to COVID-19.
- covid_deaths_per_100k: The number of COVID-19 deaths per 100,000 people.
- excess_deaths_per_100k: The number of excess deaths per 100,000 people.
- excess_deaths_pct_change: The percentage change in excess deaths.

The dataset "us-counties.parquet" contains daily data for each county in the United States from January 21, 2020, to September 29, 2021. The data includes the following fields:

- date: Date of the data entry.
- geoid: Geographic identifier for the county.
- county: Name of the county.
- state: Name of the state.
- cases: Total number of COVID-19 cases reported.
- cases_avg: Average number of cases over a period.
- cases_avg_per_100k: Average number of cases per 100,000 population over a period.
- deaths: Total number of COVID-19 deaths reported.
- deaths_avg: Average number of deaths over a period.
- deaths_avg_per_100k: Average number of deaths per 100,000 population over a period.

V. SYSTEM ARCHITECTURE AND METHODOLOGY

The process begins with the downloading of the COVID-19 Deaths Dataset from Kaggle. This dataset comprises two files: us-counties.csv and all_weekly_excess_deaths.csv. These CSV files are then processed using Jupyter Notebook, where Python programming is employed to convert the CSV files into a more efficient Parquet file format. Parquet is a columnar storage file format that is optimized for use with big data processing frameworks like Apache Spark. Once the data is in Parquet format, PySpark is utilized to perform data integration tasks. PySpark is the Python library for Spark and is used for handling big data processing and analytics. After the data integration, the data is visualized using a library like Matplotlib. This step allows for the graphical representation of the data, making it easier to identify patterns, trends, and outliers. Parallelly, the integrated data is stored in MongoDB, a popular NoSQL database. MongoDB is known for its flexibility and is used for storing data in a format that includes fields that can vary between documents and data structures. The MongoDB instance is containerized using Docker, which allows for easier deployment and scaling.

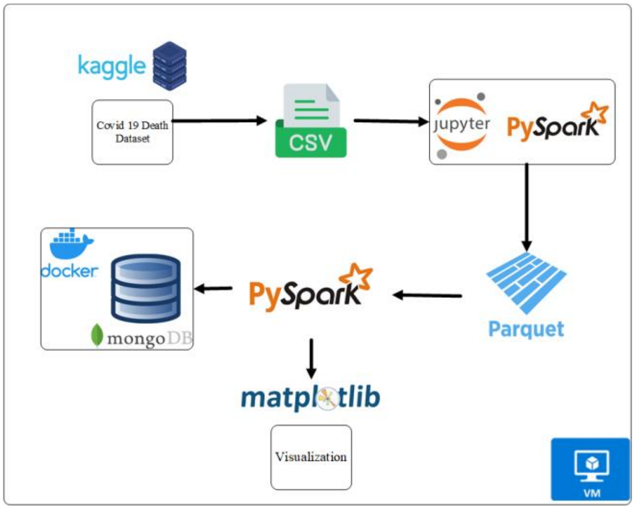


Fig. 1. System Architecture

The dataflow begins with the input of a CSV file, which contains the COVID-19 Deaths Dataset. The CSV file is then processed and converted into a more efficient Parquet file format. This conversion facilitates more efficient data processing and analysis in the subsequent steps. Once the data is in Parquet format, PySpark is utilized to perform data integration tasks. This step involves cleaning, transforming, and otherwise preparing the data for analysis. After the data integration, the data is visualized. This step allows for the graphical representation of the data, making it easier to identify patterns, trends, and outliers. Finally, the visualized data is stored in MongoDB for future use or reference.

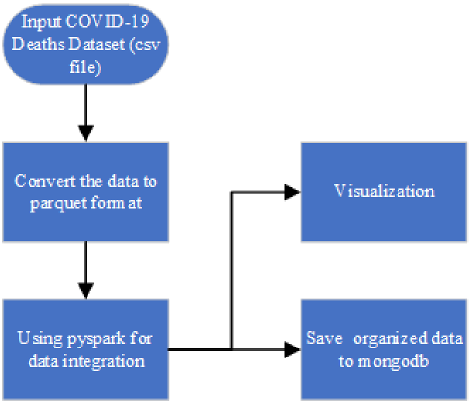

Input COVID-19 Deaths Dataset (csv file)
Convert the data to parquet format
Using pyspark for data integration
Visualization
Save organized data to mongodb

Fig. 2. Data Flow diagrams

VI. RESULT

In our study, we analyzed the death ratio for each country, which is defined as the proportion of COVID-19 deaths to the total number of deaths (both COVID-19 and non-COVID-19). We used a dataset stored in a Parquet file named ‘all_weekly_excess_deaths.parquet’, which contains weekly excess death data for various countries. The dataset was loaded into a DataFrame, and we performed several operations to calculate the total number of non-COVID-19 deaths and COVID-19 deaths for each country. We then calculated the death ratio and sorted the countries based on this ratio in descending order. The resulting DataFrame was then converted to a Pandas DataFrame for visualization purposes. We plotted a horizontal bar chart showing the death ratio for each country, sorted in ascending order. This visualization provides a clear and concise view of the death ratio across different countries.

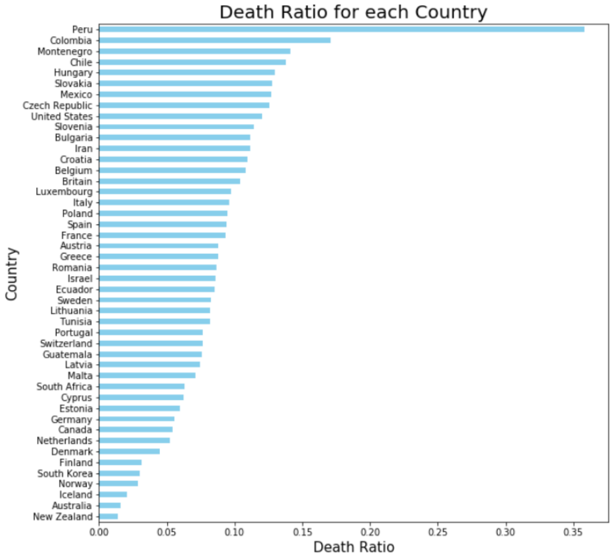


Fig. 3. Death Ratio for each County

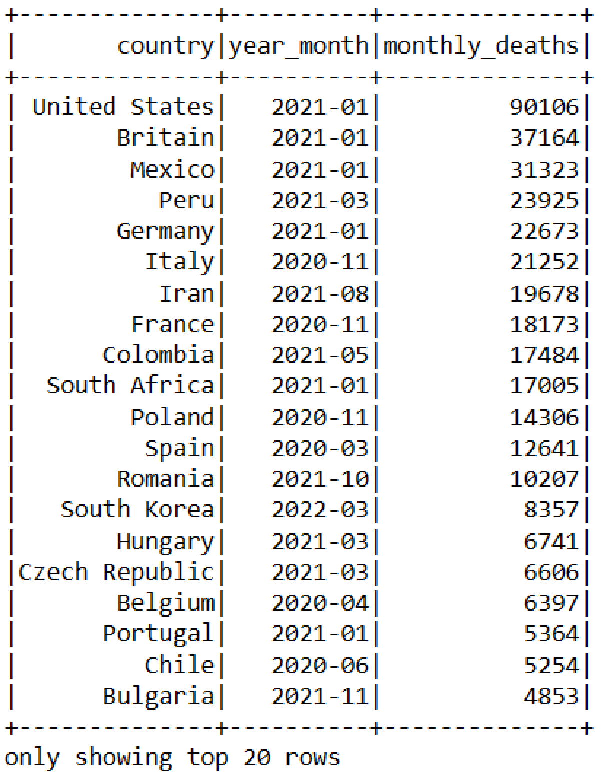
From Figure 3, we can observe that Peru has the highest death ratio, indicating a high proportion of COVID-19 deaths relative to the total number of deaths. On the other hand, New Zealand has the lowest death ratio among the countries listed.

Peru's COVID-19 death toll has more than doubled, making it the country with the highest COVID-19 death rate relative to population in the world, according to revised data following a government review. Peru’s health system is ill-prepared and lacks adequate funding. According to a BBC report on June 2, 2021[4], Peru only has about 1,600 intensive care beds across the country, far fewer than some neighboring countries. Peru's vaccination rollout has been slow, with less than 4% of the country's population fully vaccinated.

Countries with lower death rates are more likely to have their governments and people adopt and comply with measures to control the spread, according to a new study[5]. These countries are more sensitive to the threat of COVID-19, respond more quickly, and are more willing to follow recommended safety measures. The study found that differences between countries' true case fatality rates were largely due to differences in societies' susceptibility to the COVID-19 threat, how quickly they responded, and the extent to which they followed recommended safety measures.

Table 1 provides a snapshot of the month in which each country experienced the highest number of COVID-19 related deaths.

According to the table, we can see that it provides a clear picture of the severity of the COVID-19 pandemic across different countries and at different times. For instance, the United States experienced its highest number of COVID-19 deaths in January 2021, with a total of 90,106 deaths.



```
+--------------+----------+--------------+
|       country|year_month|monthly_deaths|
+--------------+----------+--------------+
| United States|   2021-01|         90106|
|       Britain|   2021-01|         37164|
|        Mexico|   2021-01|         31323|
|          Peru|   2021-03|         23925|
|       Germany|   2021-01|         22673|
|         Italy|   2020-11|         21252|
|          Iran|   2021-08|         19678|
|        France|   2020-11|         18173|
|      Colombia|   2021-05|         17484|
|  South Africa|   2021-01|         17005|
|        Poland|   2020-11|         14306|
|         Spain|   2020-03|         12641|
|       Romania|   2021-10|         10207|
|   South Korea|   2022-03|          8357|
|       Hungary|   2021-03|          6741|
|Czech Republic|   2021-03|          6606|
|       Belgium|   2020-04|          6397|
|      Portugal|   2021-01|          5364|
|         Chile|   2020-06|          5254|
|      Bulgaria|   2021-11|          4853|
+--------------+----------+--------------+
only showing top 20 rows
```

Table 1. Monthly Peak COVID-19 Deaths by Country

This table is particularly useful for understanding the temporal distribution of COVID-19 deaths. It shows that the impact of the pandemic varied significantly not only from one country to another but also from one month to another within the same country. This could be due to a variety of factors, including the timing and effectiveness of public health measure s, the spread of different variants of the virus, and the progress of vaccination campaigns.

It would be interesting to correlate these findings with other data, such as the timing of lockdowns, the percentage of the population vaccinated at different times, and the prevalence of different variants in each country. This could provide further insights into the factors that influenced the monthly death toll in each country.

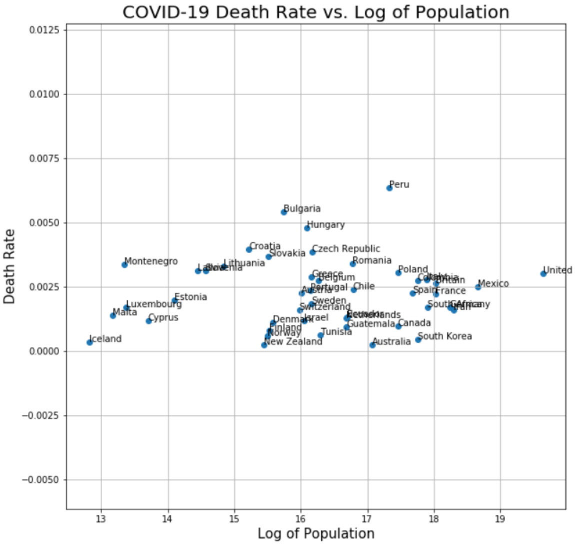


Fig. 4.COVID-19 Death Rate vs. Log of Population

The scatter plot, Figure 4 plots the death rate against the log of the population for each country. The death rate is calculated as the total number of COVID-19 deaths divided by the population of the country. The plot does not show a clear correlation between population size and death rate, suggesting that other factors may play a more significant role in determining the death rate.

For example, Peru, despite having a smaller population than the United States, has a high death rate of approximately 0.0064. This could be due to differences in healthcare infrastructure, government response, or other factors. On the other hand, countries like South Korea and Australia, despite having large populations, have managed to keep their death rates relatively low.

The bar chart, figure 5 provides a visual representation of the COVID-19 case fatality rate across different states in the United States. The case fatality rate is calculated as the total number of deaths divided by the total number of cases for each state.

From the chart, we can observe that New Jersey has the highest case fatality rate, followed by Massachusetts and New York. These states have case fatality rates of approximately 0.0238, 0.0230, and 0.0229 respectively.

New Jersey's high case fatality rate could be attributed to several factors. New Jersey is one of the most densely populated states in the United States. Higher population density can lead to increased transmission of infectious diseases like COVID-19, putting more people at risk of severe illness and death. During the peak of the pandemic, New Jersey's healthcare system faced significant strain due to the surge in COVID-19 cases. Hospitals may have struggled to provide adequate care to all patients, leading to higher mortality rates.[6]

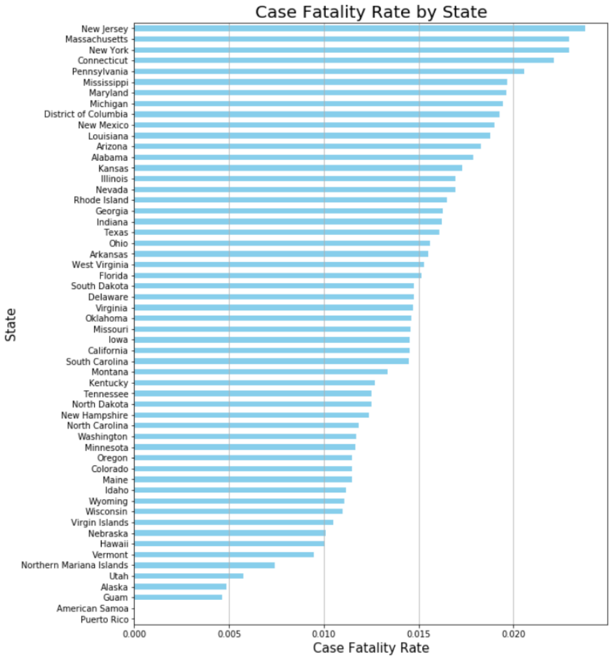
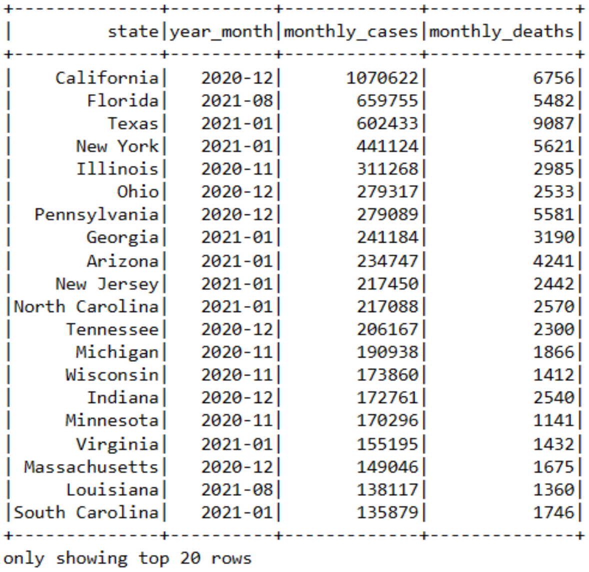


Fig. 5. Case fatality rate by State



```
+--------------+----------+-------------+--------------+
|         state|year_month|monthly_cases|monthly_deaths|
+--------------+----------+-------------+--------------+
|    California|   2020-12|      1070622|          6756|
|       Florida|   2021-08|       659755|          5482|
|         Texas|   2021-01|       602433|          9087|
|      New York|   2021-01|       441124|          5621|
|      Illinois|   2020-11|       311268|          2985|
|          Ohio|   2020-12|       279317|          2533|
|  Pennsylvania|   2020-12|       279089|          5581|
|       Georgia|   2021-01|       241184|          3190|
|       Arizona|   2021-01|       234747|          4241|
|    New Jersey|   2021-01|       217450|          2442|
|North Carolina|   2021-01|       217088|          2570|
|     Tennessee|   2020-12|       206167|          2300|
|      Michigan|   2020-11|       190938|          1866|
|     Wisconsin|   2020-11|       173860|          1412|
|       Indiana|   2020-12|       172761|          2540|
|     Minnesota|   2020-11|       170296|          1141|
|      Virginia|   2021-01|       155195|          1432|
| Massachusetts|   2020-12|       149046|          1675|
|     Louisiana|   2021-08|       138117|          1360|
|South Carolina|   2021-01|       135879|          1746|
+--------------+----------+-------------+--------------+
only showing top 20 rows
```

Table 2.Peak Monthly COVID-19 Cases and Deaths by State

Table 2 provides insights into the variation in COVID-19 mortality rates across different states and highlights the periods when certain states face higher fatalities. Analyzing the table, we can see that it provides a clear picture of the severity of the COVID-19 pandemic across different states and at different times. For instance, California experienced its highest number of COVID-19 cases in December 2020, with a total of 1,070,622 cases and 6,756 deaths. It shows that the impact of the pandemic varied significantly not only from one state to another but also from one month to another within the same state. Understanding these trends can aid policymakers and healthcare professionals in implementing targeted interventions to mitigate the impact of the pandemic.

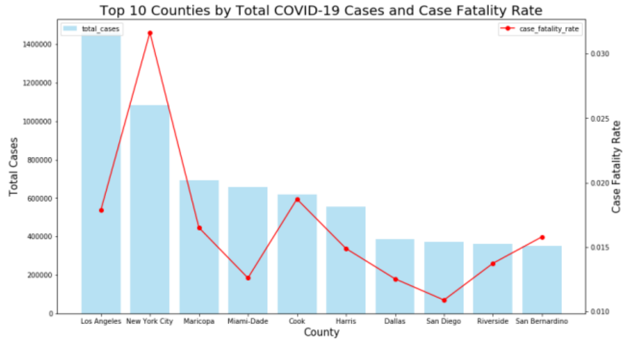


Fig. 6.The 10 counties with the highest number of cases

According to figure 6, Los Angeles County in California tops the list with a total of 1,458,381 cases, significantly more than any other county. This could be related to the population density and size of Los Angeles County. Despite having fewer total cases (1,082,099) than Los Angeles County, New York City has the highest case fatality rate (0.0316). This could be related to factors such as the city’s healthcare resources, population age structure, or virus variants.

Figure 6 shows no direct correlation between the number of total cases and the case fatality rate. For example, Maricopa County and Miami-Dade County have similar total case numbers but significantly different fatality rates. This could be related to factors such as healthcare resources, disease control measures, and population health status in each county.

Among the top 10 counties with the most cases, four are in California: Los Angeles County, San Diego County, Riverside County, and San Bernardino County. This could be related to the population density and size of California.

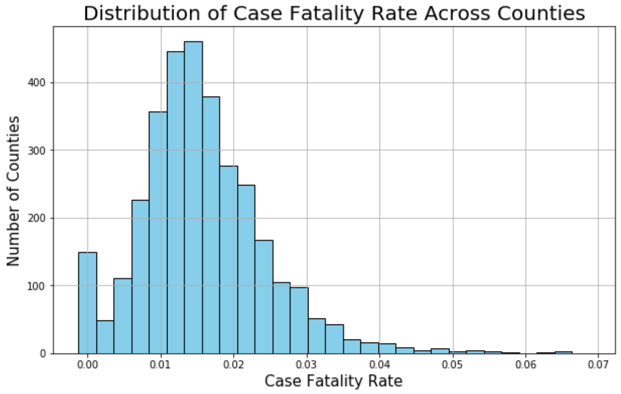


Fig. 7. Distribution of Case fatality rate across counties

The histogram figure 7 provides a distribution of the case fatality rate across all 1930 counties. The x-axis represents the case fatality rate, ranging from 0 to 0.07, while the y-axis indicates the number of counties falling within each bin of case fatality rates. The most common range for case fatality rates is between approximately 0.015 and 0.025, with over 400 counties falling within this range. The distribution is slightly right-skewed, indicating that there are a few counties with higher case fatality rates. Very few counties have a case fatality rate above 0.05 or below 0.01. In the dataset, some counties have very few cases, which causes their case fatality rate to be higher than other counties.

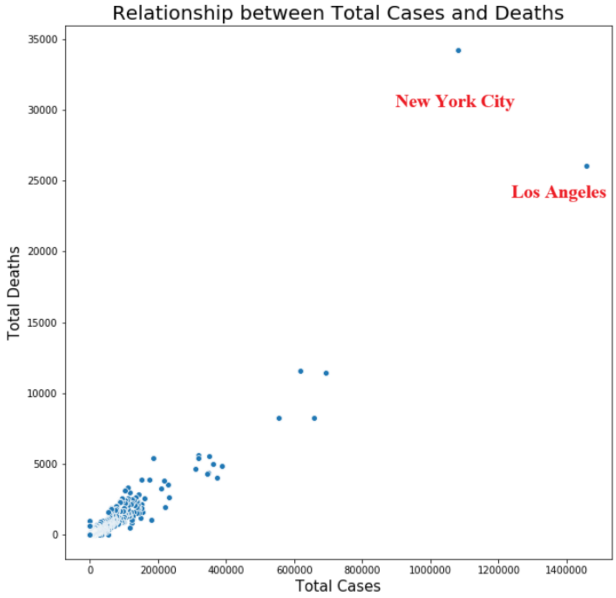


Fig. 8.Relationship between Total Cases and Deaths

The scatter plot figure 8 visualizes the relationship between total COVID-19 cases and deaths across various counties in the United States. The x-axis represents the total number of cases, while the y-axis represents the total number of deaths.

A positive correlation is observed in the scatter plot, indicating that as the number of total cases increases, the total deaths tend to increase as well. This is expected as areas with more cases are likely to have more deaths due to the virus.However, the relationship is not perfectly linear, suggesting that other factors may be influencing the number of deaths. For instance, the quality of healthcare, the demographic profile (such as age and pre-existing conditions), and the timeliness of the response to the outbreak could all play a role.

Two distinct data points labeled “Los Angeles” and “New York City” are outliers with significantly higher numbers of both cases and deaths. This could be due to these areas’ high population density, which can facilitate the virus’s spread. Most counties have lower total cases and deaths, clustering towards the origin of the axes. This suggests that a large number of counties have managed to keep both metrics relatively low, possibly due to effective public health measures.

VII. CONCLUSION

The study analyzed a dataset of COVID-19 deaths and cases across various countries and counties within the United States. The analysis yielded several key findings.

Peru had the highest death ratio, indicating a high proportion of COVID-19 deaths relative to total deaths. This could be attributed to factors like weak healthcare infrastructure and slow vaccination rollout. In contrast, New Zealand had the lowest death ratio, suggesting a more effective pandemic response.

The monthly death toll varied significantly across countries. This highlights the influence of factors like public health measures, variant spread, and vaccination progress. Further analysis correlating these factors with death tolls would be insightful. No clear correlation emerged between population size and death rate. Countries like Peru with a smaller population exhibited high death rates, suggesting the influence of other factors like healthcare systems.

New Jersey had the highest case fatality rate among US states, possibly due to high population density and strain on the healthcare system during the peak. Los Angeles County has the highest total number of cases, likely due to its higher population density.However, New York City had a higher case fatality rate, suggesting a role for factors beyond total cases. A positive correlation was observed between total cases and deaths across counties, as expected. However, the presence of outliers like Los Angeles and New York City, and the non-linearity of the relationship, suggest the influence of additional factors like healthcare quality and demographics.

In summary, our project provides valuable insights into the complex dynamics of the COVID-19 pandemic, emphasizing the need for coordinated efforts at local, national, and global levels to combat the virus effectively. Further research and analysis are warranted to delve deeper into the multifaceted nature of the pandemic and inform evidence-based decision-making for future public health crises.

VIII. REFERENCES

[1] Estiri, H., Strasser, Z.H., Klann, J.G. *et al.* Predicting COVID-19 mortality with electronic medical records. *npj Digit. Med.* 4, 15 (2021). https://doi.org/10.1038/s41746-021-00383-x

[2] Dong, J., Wu, H., Zhou, D., Li, K., Zhang, Y., Ji, H., ⋯ & Liu, Z. (2021). Application of big data and artificial intelligence in covid-19 prevention, diagnosis, treatment and management decisions in china. Journal of Medical Systems, 45(9). https://doi.org/10.1007/s10916-021-01757-0

[3] Dhruvil Dave. (2021). COVID-19 Deaths Dataset [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DS/1358431

[4] Covid: Why has Peru been so badly hit? (bbc.com)

[5] Lim, T. Y., Xu, R., Ruktanonchai, N., Saucedo, O., Childs, L. M., Jalali, M. S., Rahmandad, H., & Ghaffarzadegan, N. (2023). Why Similar Policies Resulted In Different COVID-19 Outcomes: How Responsiveness And Culture Influenced Mortality Rates. Health Affairs, 43(1), 91-97. https://doi.org/10.1377/hlthaff.2023.00713

[6] COVID-19 Cases, Staff Shortages Put Strain On NJ Hospitals | Cinnaminson, NJ Patch