Project Title:

Is COVID-19 Impact Uniform Across Countries?

Abstract:

The COVID-19 pandemic impacted countries in vastly different ways, despite similarities in active case numbers. This project investigates the relationship between active COVID-19 cases and death rates to explore why outcomes varied. The goal was to answer: *What factors helped some countries manage the pandemic better than others?*

We used a 1.6 GB dataset from Johns Hopkins University to analyze global COVID-19 statistics. The analysis focused on active cases, death rates, and their patterns across countries. Basic visualizations like a bar chart and scatter plot revealed preliminary trends, while a bubble chart served as the main explanatory visualization. This chart combined active cases, death rates, and case magnitude to show how different countries managed the crisis.

The findings highlight that healthcare systems, early government actions, and population factors played significant roles in outcomes. Countries like New Zealand and Australia maintained low death rates through robust measures, while nations like Peru and Mexico experienced higher mortality despite fewer cases. The conclusion underscores that managing a pandemic requires more than just controlling case numbers—it demands tailored strategies, strong healthcare infrastructure, and timely action.

1.Dataset(s):

Source and Retrieval:

The dataset was sourced from the Johns Hopkins University COVID-19 Data Repository, a comprehensive and frequently updated collection of global COVID-19 statistics.

Dataset Description:

- Size: 1.6 GB
- Number of Rows: Millions of records spanning three years (2020–2023).
- Attributes:
- Country: The geographical region where data was recorded.
  - Confirmed Cases: Total reported COVID-19 cases.
  - Active Cases: Current infections, excluding recoveries and deaths.
  - Deaths: Total COVID-19-related fatalities.
  - Death Rate: A calculated percentage of deaths relative to confirmed cases.

Big Data Characteristics:

- Volume: Large dataset containing millions of records across multiple years.
- Velocity: Updated daily, reflecting real-time changes in pandemic trends.
- Variety: Includes a wide range of attributes like geographical data, case numbers, and calculated metrics.

This dataset exemplifies big data by integrating volume, velocity, and variety to enable a comprehensive analysis of the pandemic's global impact.

2.Data Exploration, Processing, Cleaning, and Integration:

Preparation Steps:

To ensure the dataset was ready for visualization, the following steps were taken:

1. Filtering Data: Only countries with complete and reliable records were included. Missing or incomplete entries were excluded.
2. Handling Missing Values: Interpolation methods were used to estimate missing values in critical fields like deaths, where feasible.
3. Creating New Columns: A "Death Rate" column was calculated to standardize comparisons across countries:

Death Rate (%) = (Deaths / Confirmed Cases)×100

4. Normalizing Data: Logarithmic scaling was applied to active case numbers to account for population size differences.
5. Reducing Dataset Size: Data from 2023 onwards was retained, and unnecessary columns were dropped to streamline analysis.

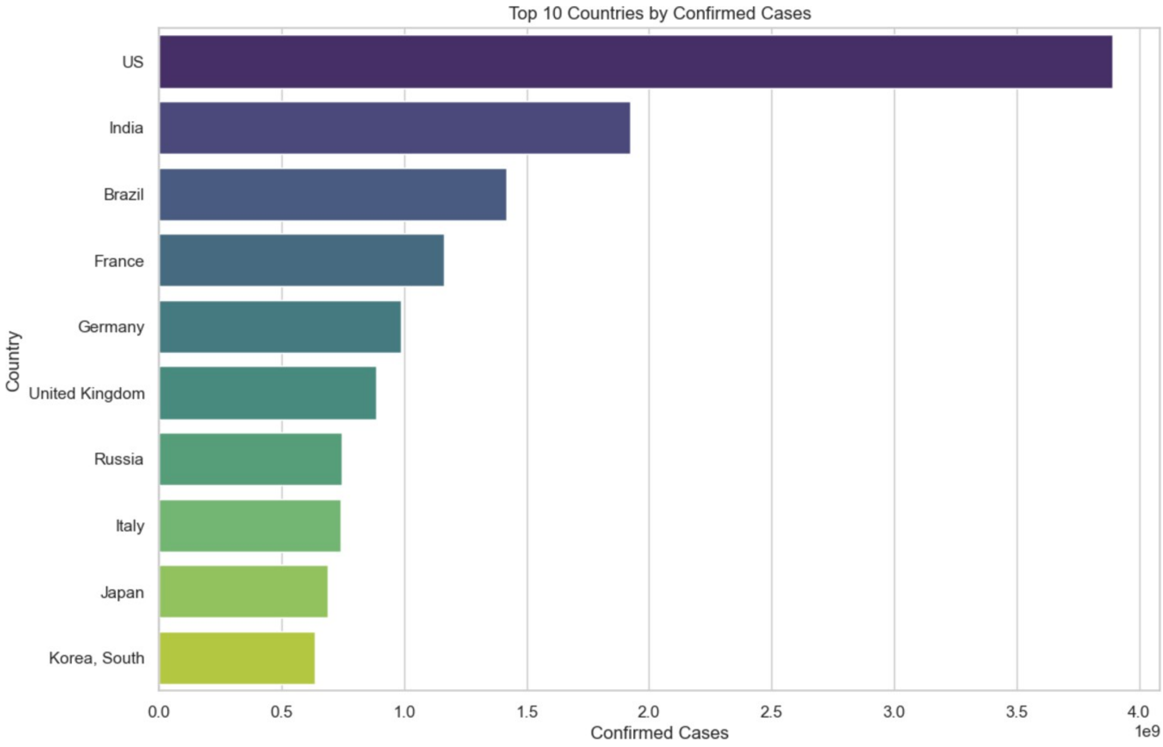
Attribute Selection:

The following attributes were selected for visualization:

- Active Cases: Represents the strain on healthcare systems.
- Death Rate: Indicates the severity of the pandemic in terms of mortality.
- Country: To compare trends across different nations.

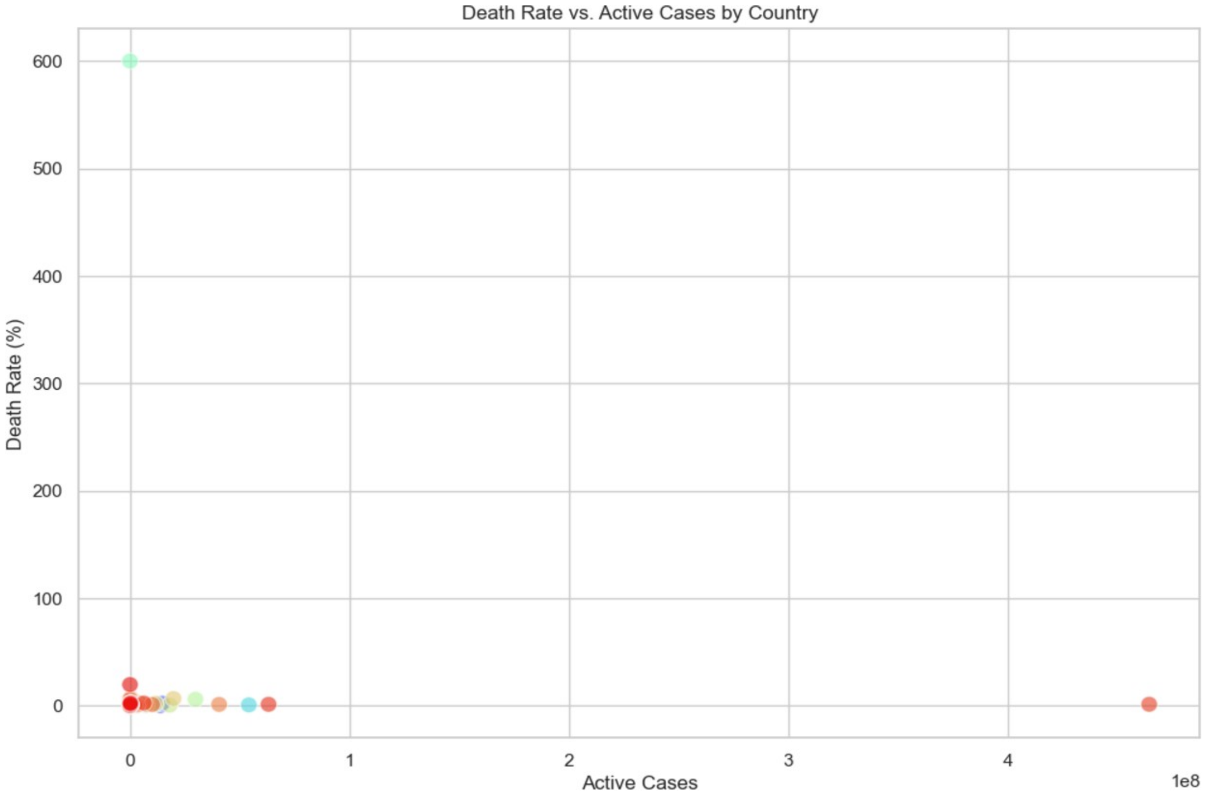
These attributes were chosen to focus on the relationship between healthcare burden and mortality outcomes, aligning with the research question.

- Figure 1: Bar Chart - Confirmed Cases by Country



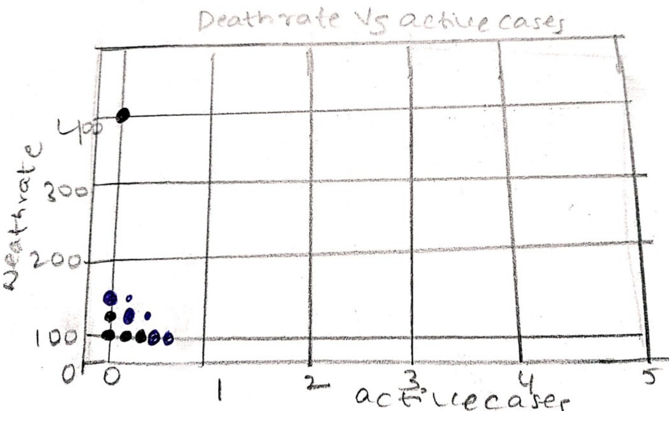
"Figure 1 shows the bar chart that compares confirmed COVID-19 cases across countries, highlighting nations such as India and the United States, which had the highest case counts."

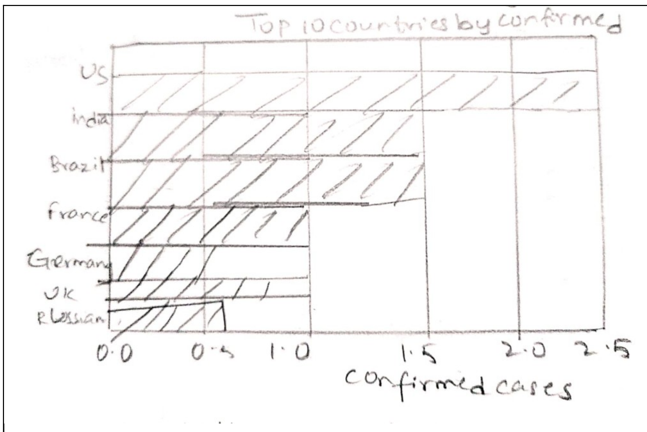
- Figure2:ScatterPlot-DeathRatesvsActiveCases



- "Figure 2 illustrates the scatter plot depicting the relationship between active cases and death rates. This helped us identify countries like Peru and Mexico, which had higher death rates despite moderate active cases."

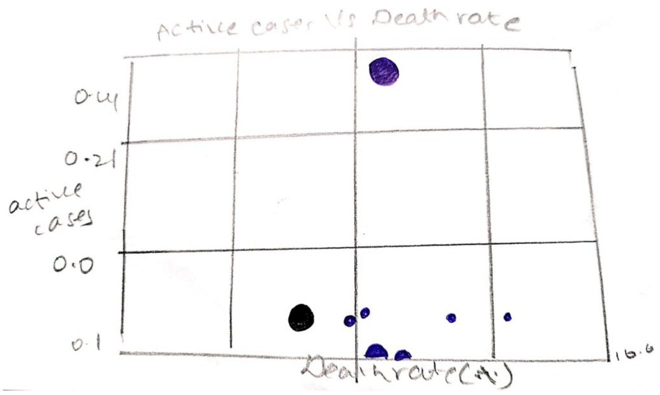
3.Visualisation:

 	**Sketch 1: scatter plot** *This helped us identify countries like Peru and Mexico, which had higher death rates despite moderate active cases.*



Sketch 2: Bar graph

compares confirmed COVID-19 cases across countries, highlighting nations such as India and the United States, which had the highest case counts.



Sketch 3:bubble chart

Compared confirmed cases across countries. Analyzed the relationship between active cases and death rates. *Combined active cases, death rates, and case magnitude for the main analysis.*

Design Process:

The visualization process began with sketches and exploratory graphs:

1. Bar Chart: Used to compare confirmed cases across countries, identifying those with the highest numbers.
2. Scatter Plot: Explored the relationship between active cases and death rates, highlighting trends and outliers.

The bubble chart was selected as the main explanatory visualization due to its ability to represent three dimensions of data.

Design Choices:

1. Chart Type:
    - The bubble chart was chosen to show the relationship between active cases (Y-axis), death rates (X-axis), and case magnitude (bubble size). This allows for clear comparisons across countries.

2. Color Scheme:
    - Bright, contrasting colors were used to differentiate data clusters and make trends visually distinct.
3. Labels and Annotations:
    - Country names were added to bubbles for clarity, while tooltips provided detailed information on hover.
4. Interactivity:
    - The chart was made interactive using Power BI, allowing users to explore details by hovering over bubbles.

Sketch of Design:

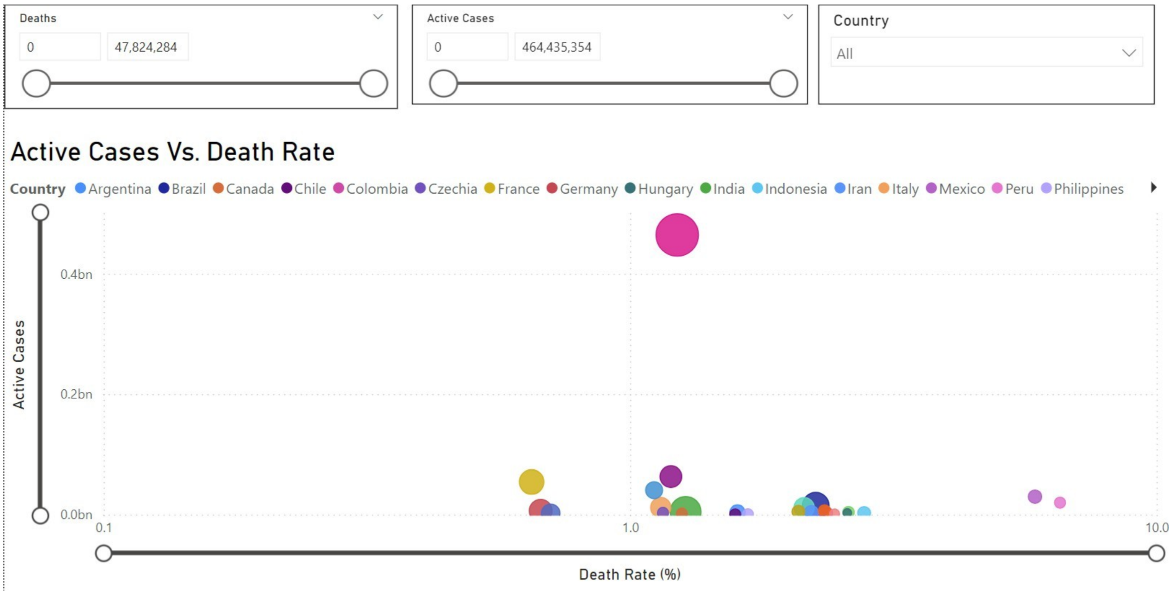
Sketch included to illustrate the initial planning for the bubble chart, showing the arrangement of axes, bubble sizes, and labels.

Final Visualization:

The final visualization was created in Power BI and included:

1. Bar Chart *(Exploratory)*: Compared confirmed cases across countries.
2. Scatter Plot *(Exploratory)*: Analyzed the relationship between active cases and death rates.
3. Bubble Chart *(Explanatory)*: Combined active cases, death rates, and case magnitude for the main analysis.

The bubble chart (Figure 3) is highlighted as the primary visual, answering the research question effectively.


Deaths: 0 – 47,824,284

Active Cases: 0 – 464,435,354

Country: All

4.Conclusion:

Key Findings:

1. Unequal Impact: COVID-19 affected countries differently, even with similar active case numbers.
2. Critical Role of Interventions: Early measures like lockdowns, testing, and vaccinations helped lower death rates.
3. Healthcare Systems Matter: Strong healthcare systems were vital for managing mortality outcomes.

Tools and Libraries Used:

- Data Cleaning: Jupyter Notebook and Google Colab (Pandas, NumPy).
- Visualization: Power BI for interactive charting and Matplotlib for exploratory visualizations.

Reflection:

- Strengths: The visualizations clearly highlight key patterns and disparities, effectively answering the research question.
- Improvements: Adding variables like vaccination rates or population density could provide deeper insights.
- Challenges: Due to the dataset's size, some advanced animations and dynamic visualizations were not feasible.

Collaboration:

This project was completed as a group effort:

- Nikhil Kumar: Focused on data cleaning, preparing exploratory graphs, and drafting the report.
- Venga Vamsi Krishna Maanam: Designed and implemented the final bubble chart in Power BI and provided key insights for the analysis.

References:

1. Johns Hopkins University COVID-19 Data Repository: https://github.com/CSSEGISandData/COVID-19
2. Power BI Documentation: Microsoft Link
3. Python Libraries: Pandas, NumPy, Matplotlib (Documentation at Python.org).