Project Title: COVID-19 Data Analysis and Visualization

Prepared by: khaja pasha - 2310039459

Geeresh - 2310039455

Charith - 2310030218

Sai Charan -2310030

Lateef - 2310030069

Abstract / Executive Summary:

- A short overview of the project:
    - This project analyzes COVID-19 datasets to identify case trends, recovery patterns, death rates, and vaccination progress. The analysis uses Python for data cleaning, processing, and visualization, highlighting key insights about the pandemic.

Introduction:

- This project**,** COVID-19 Data Analysis and Visualization, focuses on cleaning raw COVID-19 datasets, processing the data using Python libraries, and applying visualization techniques to uncover meaningful insights. A major part of the analysis involves calculating rolling averages (e.g., 7-day averages) to smooth daily fluctuations and highlight overall trends.
- Furthermore, the project compares key variables, such as confirmed cases vs. deaths**,** recoveries vs. deaths**,** and vaccination rates, to assess the progression of the pandemic and the positive impact of vaccines. Through clear graphs and visualizations, the project aims to make complex data more understandable for policymakers, healthcare professionals, and the general public.
- By conducting this analysis, the project not only provides a retrospective view of COVID-19 but also demonstrates how data science can be applied to public health challenges. These insights can help shape preparedness strategies for future outbreaks and emphasize the importance of continuous monitoring and visualization in managing global health crises.
- Objectives

- The primary aim of this project is to perform an in-depth analysis of COVID-19 data and visualize key trends that highlight the progression and impact of the pandemic. The specific objectives include:

Data Collection & Cleaning

1. Gather COVID-19 datasets from reliable sources such as Johns Hopkins University, WHO, or Kaggle.
2. Perform preprocessing tasks such as handling missing values, removing duplicates, converting dates into proper formats, and preparing the data for analysis.

Trend Analysis of COVID-19 Cases

3. Analyze the trends in **confirmed cases, deaths, and recoveries** over time.
4. Use **7-day rolling averages** to smooth short-term fluctuations and highlight the actual progression of the pandemic.

Comparative Study of Recovery and Mortality

5. Compare **recovery rates** with **death rates** to understand how the pandemic has evolved.
6. Identify peak points, decline phases, and changes across different waves.

Vaccination Progress Monitoring

7. Analyze vaccination data to study how vaccination campaigns have impacted the spread of COVID-19.
8. Evaluate correlations between vaccination progress and case reduction trends.

Visualization of Insights

9. Create clear, easy-to-interpret visualizations (line graphs, bar charts, etc.) that communicate the findings effectively.
10. Highlight key events or changes using visual representations for better understanding.

Deriving Insights for Decision-Making

11. Provide insights into the effectiveness of measures like vaccination drives.
12. Support policymakers, healthcare workers, and researchers by presenting data-driven findings.

Documentation & Reporting

13. Prepare structured documentation with analysis results, insights, and visualizations.
14. Suggest potential **future scope**, such as predictive modeling and machine learning applications for forecasting cases.

Dataset Description

- For this project, COVID-19 datasets were obtained from publicly available and reliable sources such as:
- Johns Hopkins University (JHU) Center for Systems Science and Engineering (CSSE) – provides global COVID-19 time series data.
- World Health Organization (WHO) – publishes official COVID-19 situation reports and vaccination progress.
- Kaggle COVID-19 Datasets – community-maintained datasets for confirmed cases, deaths, recoveries, and vaccinations.

Dataset Features

- The dataset typically includes the following columns:
- Date **-** The date on which the data was recorded.
- Country/Region – The name of the country or region.
- Confirmed Cases – The cumulative number of people who tested positive for COVID-19.
- Deaths – The cumulative number of deaths caused by COVID-19.
- Recovered – The cumulative number of people who recovered from COVID-19.
- Active Cases – The number of currently infected patients (calculated as Confirmed – Deaths – Recovered).
- Vaccinations – The number of doses administered (first dose, second dose, or total doses depending on the dataset).
- Population – The population of the country/region to calculate percentages such as vaccination coverage.

Data Cleaning Steps

- Before performing analysis, the dataset was cleaned to ensure accuracy and consistency:
- Handling Missing Values**:** Missing entries in confirmed cases, deaths, or recoveries were filled using forward-fill methods or removed when appropriate.
- Date Formatting**:** Converted date columns into proper datetime format for time-series analysis.
- Duplicate Removal**:** Eliminated duplicate rows caused by repeated reporting.
- **Derived Columns:**
    1. Daily Confirmed = difference in confirmed cases between consecutive days.
    2. Daily Deaths = difference in death counts between consecutive days.
    3. Daily Recoveries = difference in recovery counts between consecutive days.
    4. 7-Day Rolling Average for cases, deaths, and recoveries to smooth fluctuations.

- Normalization**:** In some cases, values were normalized (e.g., cases per 100,000 population) to enable fair comparison between countries with different population sizes.
- Tools and Technologies
- Languages: Python
- Libraries: Pandas, NumPy, Matplotlib, Seaborn
- Platform: Jupyter Notebook / Google Colab
- Visualization**:** PowerPoint charts, plots
- Methodology

Data Collection

COVID-19 datasets were collected from reliable sources like Johns Hopkins University, WHO, and Kaggle.
The data included confirmed cases, deaths, recoveries, and vaccination information for multiple countries.

Data Cleaning

Missing values, duplicates, and inconsistent entries were handled to ensure data accuracy.
Dates were standardized, and additional columns like daily cases and 7-day rolling averages were created.

Exploratory Data Analysis (EDA)

Descriptive statistics were calculated to understand trends and distribution of COVID-19 metrics.
Visualizations such as line charts, bar graphs, and histograms were used to summarize the data.

Visualization

Time series plots were created to show confirmed, recovered, and death trends over time.
Vaccination progress and comparative analysis between countries were visualized for insights.

Trend Analysis

7-day rolling averages were computed to smooth out short-term fluctuations in daily cases.
Peaks, waves, and declining phases were identified to better understand pandemic progression.

Insights & Reporting

Key patterns and relationships between variables were summarized in charts and tables.
Findings were documented to support decision-making and highlight future research directions.

Analysis & Visualization

The analysis focuses on understanding the progression of COVID-19 cases, recoveries, deaths, and vaccination trends through data visualization techniques.

Confirmed Cases Trend:

- A line graph was plotted for daily confirmed cases over time.
- A 7-day rolling average was applied to smooth daily fluctuations, highlighting overall trends and identifying peak waves.

Deaths vs Recovered Cases:

- Comparative line charts were created to visualize cumulative deaths and recoveries.
- This helps assess recovery rates relative to deaths and understand the severity of the pandemic over time.

Vaccination Progress:

- Vaccination data was plotted as a line chart to show the number of doses administered over time.
- Trends indicate the impact of vaccination on controlling case numbers and mitigating the spread of the virus.

Daily Cases and Rolling Average:

- Daily confirmed cases were calculated to track short-term trends.
- The 7-day rolling average reduces volatility, making it easier to identify the true direction of the pandemic.

Country-wise Comparison :

- Cases, deaths, recoveries, and vaccination rates were compared across selected countries.
- Normalizing data per 100,000 population allows fair comparisons between countries with different population sizes.

Visual Tools Used

- Python Libraries: Matplotlib and Seaborn were used for creating all plots and charts.
- Charts: Line charts, bar graphs, and comparative visualizations were used to present trends clearly.
- Key Observations from Visualizations

- Multiple waves of COVID-19 are clearly visible in the 7-day rolling average trends.
- Recovery rates generally surpass death rates, showing improved treatment outcomes over time.
- Vaccination campaigns show a positive impact, with case growth slowing as vaccination coverage increases.
- Results & Key Insights

- 7-day rolling averages reduce noise in data.
- Vaccination increased steadily, reducing case severity.
- Recovery cases surpass deaths significantly.
- Multiple waves can be detected in case curves.

Conclusion:

- COVID-19 data analysis reveals important patterns.
- Vaccination is a major factor in reducing deaths.
- Continuous monitoring and predictive models can improve future preparedness.