Project Report: COVID-19 Data Analysis

Team Members

Name	E-mail id.
Sachin Pandewad	Sachinpandewad10@gmail.com
Heena Tabassum	tabassumheena965@gmail.com
Thejaswini Prakash	thejaswini2105@gmail.com

Project Title: COVID-19 Data Analysis

Objective

The primary objective of this project is to **analyze the global impact and progression of the COVID-19 pandemic** using real-world data. The analysis focuses on identifying monthly trends in confirmed cases, deaths, and recoveries across different countries. The goal is to extract insights, visualize key patterns, and support public health understanding and response strategies using data science techniques such as:

- Time-series trend analysis

- Aggregation of daily data into monthly summaries

- Interactive data visualization using Power BI / Tableau

This project ultimately aims to support informed decision-making and help communicate the evolution of the pandemic in an accessible visual format.

Tools Used

- Programming Languages: Python, NumPy, Pandas, Matplotlib, Seaborn
- Tools: Jupyter Notebook, Power BI (for visualization).

Dataset

The dataset was collected from a global COVID-19 tracking source. It includes daily records of confirmed cases, deaths, and recoveries across various countries. The dataset was cleaned and transformed to aggregate values on a monthly basis for improved analysis and visualization.

Raw Dataset (COVID-19_Dataset1.csv) Link◎

- **Rows**: 54,522 daily records

- **Columns**: 6

- **Source**: Aggregated COVID-19 global from Kaggle & WHO

Process Followed

- **Problem understanding and objective framing:**
  The objective was to study the global spread and impact of COVID-19. We aimed to analyze the evolution of confirmed cases, deaths, and recoveries over time across different countries. The goal was to provide insights through data trends and visual dashboards.

- **Data collection and preprocessing:**
  The raw dataset contained 54,522 daily entries with columns like Country, Date, Confirmed, Death, newConfirmed, and newDeath. Data preprocessing involved removing unnecessary columns, converting date formats, handling missing values, and aggregating the data on a monthly basis per country. The cleaned dataset contains 1,974 entries, each summarizing a month for a specific country with total confirmed cases, deaths, and recoveries.

- **Exploratory Data Analysis (EDA):**
  The raw dataset contained 54,522 daily entries with columns like Country, Date, Confirmed, Death, newConfirmed, and newDeath. Data preprocessing involved removing unnecessary columns, converting date formats, handling missing values, and aggregating the data on a monthly basis per country. The cleaned dataset contains 1,974 entries, each summarizing a month for a specific country with total confirmed cases, deaths, and recoveries.

1. Raw Dataset (COVID-19_Dataset1.csv) Link
- **Rows**: 54,522 daily records
- **Columns**:
  - Country
  - Date (MM/DD/YYYY)
  - Confirmed
  - Death
  - newConfirmed
  - newDeath

2. Cleaned Dataset (file2.xlsx) <u>Link</u>
- **Rows**: 1,974 monthly records (aggregated per country)
- **Columns**:
  - Country
  - Month (YYYY-MM)
  - days_inMonth
  - total_confirmed
  - total_deaths
  - total_recoverd

- **Transformation**:
  - Daily records aggregated to monthly values
  - Unnecessary columns dropped
  - Cleaned and normalized for dashboard visualization

- # Dashboard creation and visualization:
  The cleaned dataset was used to create interactive dashboards using Power BI / Tableau. The dashboard visualizes monthly COVID-19 data by country, including confirmed, recovered, and death counts. Filters and drill-down features were added to allow dynamic exploration of country-specific trends.

- # Results and insight generation:
  The analysis revealed distinct waves of infection across different time periods. Some countries experienced early peaks, while others had delayed surges. The dashboards provided an accessible and informative way to monitor and compare the progression of the pandemic across regions.

Results

The final analysis provided monthly trends of COVID-19 cases, deaths, and recoveries for each country. It showed how different countries experienced varying intensities of the pandemic, highlighting waves and recovery phases. The dashboards enabled users to interactively explore data by month and country, revealing the worst-hit periods and the effectiveness of recovery measures.

Challenges Faced

Several challenges were encountered during the project:

- Missing or incomplete data for some countries and time periods.

- Inconsistent date formats which required transformation for aggregation.

- Data imbalance: Some countries had dense daily records while others had sparse data.

- Aggregating daily data to monthly values while maintaining data integrity.

- Managing performance issues with large datasets during visualization.

Final Deliverables

The following deliverables were prepared as part of this project:

- A cleaned and aggregated dataset with monthly statistics.

- A comprehensive project report describing the objective, methodology, and outcomes.

- An interactive dashboard built using Power BI/Tableau.

All Links:

Raw Dataset: COVID-19_Dataset1.csv

Cleaned & Aggregated Dataset: file2.xlsx

Jupyter notebook: Covid19_DatasetEDA

Power BI Dashboard: COVID-19 Dashboard