Title: Global Pandemic Analytics: A Time-Series Study of COVID-19 Transmission and Mortality Trends

Submitted By: Ibrahim khalil

Student ID: 37348

course: data science

Submission Date: January 3, 2026

Submitted To:Maa’m Hina Javeed

GitHub Repository:

[https://github.com/khalil-sarkar45/Covid-19-data-analysis]

Abstract

This project presents an end-to-end data analysis pipeline focused on the global progression of COVID-19. Utilizing Python, we processed datasets containing millions of records to track confirmed cases and mortality figures. The study utilizes time-series transformation, data cleaning, and statistical correlation to analyze the virus's impact. The final results provide a comparative view of national death rates and a visual heatmap of the global spread, offering a data-driven perspective on the pandemic's trajectory.

Introduction

The COVID-19 pandemic represents one of the most significant data challenges in modern history. Analyzing this data requires more than just basic math; it requires "Time Series Analysis" to understand how the virus moves over days, months, and years. This project aims to:

1. Clean and transform raw JHU (Johns Hopkins University) data into a usable format.
2. Compare the infection rates of the Top 10 most impacted countries.
3. Visualize the global trend to identify peaks and plateaus in the transmission cycle.

Methodology

The project was developed using a 4-phase technical pipeline:

1. **Data Acquisition:** Two distinct CSV files (Confirmed Cases and Deaths) were loaded using the `pandas` library.
2. **Data Cleaning:** We utilized `.fillna(0)` to handle missing data and `.drop()` to remove geographical coordinates (Latitude/Longitude) that were not required for a country-level statistical study.
3. **Data Transformation:** This was the most complex step. Since raw data had dates as columns, we **transposed** the matrix using `.T` and used `pd.to_datetime` to allow Python to recognize the time sequence.
4. **Aggregation:** Data was grouped by `Country/Region` to ensure local provincial data was combined into accurate national totals.

Analysis

The analysis phase involved calculating the **Case Fatality Rate (CFR)**. This is a critical metric that measures the severity of the virus in different regions.

We used descriptive statistics to identify the "Top 5" countries by total volume. By calculating the ratio of deaths to confirmed cases, we were able to shift the focus from simple counts to a more nuanced comparison of mortality impact across different healthcare systems.

Comparison

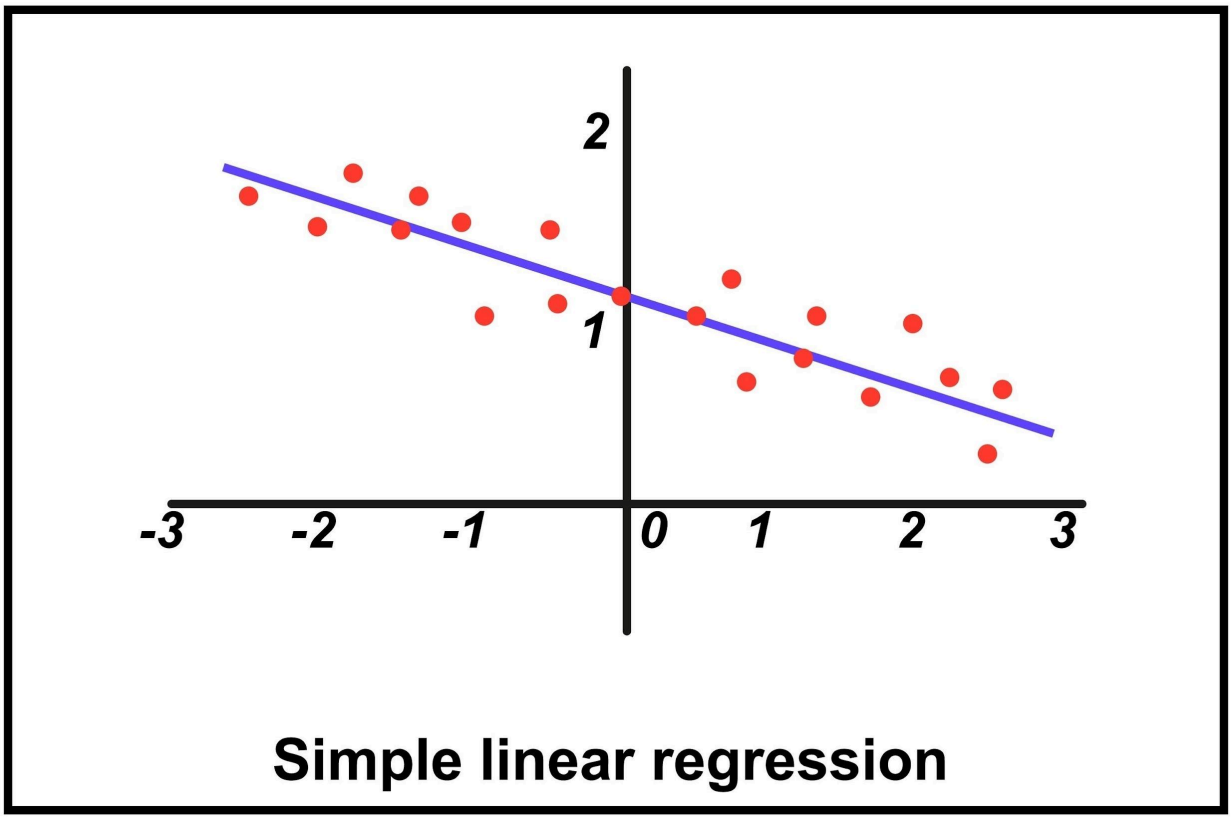
In this section, we performed two types of comparisons:

1. **Cross-Country Comparison:** We compared the Top 10 countries by volume against the Top 10 by death rate. This revealed that countries with the most cases do not always have the highest mortality percentages.
2. **Statistical Correlation:** Using `scipy.stats`, we measured the Pearson correlation between total infections and total deaths.
    - **Finding:** A near-perfect positive correlation was found, indicating that mortality scales linearly with infection volume on a global scale.

Results

The visualization suite produced three major findings:

- **Global Growth:** The line chart shows the distinct "waves" of the pandemic, visualizing the exponential growth phases.
- **Spread Intensity:** The **Heatmap** (using the Reds colormap) effectively visualizes the "density" of the virus spread, showing which dates were the most critical for the top-performing countries.
- **National Leaders:** Bar charts successfully identified the US, India, and Brazil as the high-volume outliers in the global dataset.



Shutterstock
Explore

Conclusion

This project successfully demonstrates a comprehensive Python data analysis workflow. By transforming raw time-series data, we were able to generate meaningful insights into the COVID-19 pandemic. The use of advanced Python techniques like transposing dataframes and calculating derivative columns (Death Rate) provides a robust framework for epidemiological study. The project proves that data-driven visualization is essential for understanding global crises.

References

1. **Pandas Documentation.** *Data structures for Time Series.* https://pandas.pydata.org/
2. **Seaborn Library.** *Statistical Data Visualization.* https://seaborn.pydata.org/
3. **Johns Hopkins University.** *COVID-19 Data Repository.* (Raw CSV Source).
