Time Series Analysis of COVID-19 Data

DRONAMRAJU ANUSHA

Foundational Internship/Section-1 - KL University,Hyderabad

Period of Internship: 21ST January 2026 – 17TH February 2026

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

1. Abstract

This project involves the analysis of a real-world global data set to derive useful insights through data analysis. This includes the importation and processing of the raw data for analysis. The project highlights data cleaning, processing, and formatting for accuracy. Aggregation and summarization techniques are employed to gain insights into the data set. Time-series analysis is conducted to analyze data variations over time. Comparative analysis is employed to compare variations in the data set. The project also involves the arrangement of data in organized tables for easier analysis. Statistical summaries are employed to identify important data points and maximum values.

2.Introduction

This project is centered on the analysis of the WHO COVID-19 Global Dataset to examine patterns and trends associated with the COVID-19 pandemic. The dataset, which is published by the World Health Organization, is a reliable source of information on reported cases and deaths from the COVID-19 pandemic across various countries and regions. The project emphasizes the role of data analysis in understanding public health conditions and making informed decisions. The project is done using Python and the Pandas library to efficiently clean, transform, and analyze data. The process involves loading the dataset, specifying data types, and structuring relevant information for analysis. Different aggregation and time-related methods are used to investigate trends at various times. The project assists in recognizing important patterns and regional differences in reported data. In general, it enhances data analysis skills while offering valuable insights from the WHO COVID-19 Global Dataset. Some of the topics that were helpful in doing this project were covered during the training period of the internship are listed below :

1. Python programming basics
2. Machine Learning-Regression
3. Machine Learning-Classification
4. LLM Fundamentals

3.Project Objective

1. To analyze the WHO COVID-19 Global Dataset in order to understand trends in reported cases and deaths across different countries and regions.

2. To demonstrate the process of cleaning, transforming, and organizing large real-world datasets using Python and data analysis libraries.

3. To perform time-based analysis (daily, monthly, and quarterly) to illustrate how pandemic patterns evolved over different periods.

4. To compare regional case distributions and identify variations in impact across WHO regions.

5. To illustrate the use of aggregation, grouping, and pivot table techniques for extracting meaningful statistical summaries.

4.Methodology

The data used in this project was collected from the WHO COVID-19 Global Dataset, which is publicly available and maintained by the World Health Organization (WHO). The dataset contains officially reported information on COVID-19 cases and deaths from countries and regions worldwide.

Tools and Technologies Used

- **Python Programming Language** – Core implementation language

- **Pandas Library** – Data cleaning, transformation, grouping, and aggregation

- **Google Colab / Jupyter Notebook** – Development and execution environment

- **GitHub** – Code version control and repository management

The dataset was processed entirely using Python-based data analysis techniques.

Steps Followed During the Project

Below are the major steps carried out:

Step 1: Data Import

- Imported necessary libraries (Pandas).
- Loaded the WHO COVID-19 dataset using `pd.read_csv()`.
- Displayed initial rows using `.head()` to understand structure.

Step 2: Data Understanding

Inspected columns and its data types using df.info().

Identified relevant variables such as Date_reported, Country, WHO_region, New_cases, and New_deaths.

Step 3: Data Cleaning

- Checked for missing values (if any).
- Converted Date_reported column into proper datetime format using `pd.to_datetime()`.
- Ensured numerical columns were in correct format.

Step 4: Data Pre-processing

- Created a subset of relevant columns.
- Filtered dataset within a specific time range (2020–2023).
- Created new time-based columns such as Quarter and Month_Year using `.dt.to_period()`.

Step 5: Data Aggregation and Analysis

- Grouped data by Country to identify top affected countries.
- Calculated daily global cases using `.groupby('Date_reported')`.
- Aggregated quarterly cases and deaths using `.agg()`.
- Calculated total cases by WHO region.
- Identified peak day using `.idxmax()` and `.max()`.
- Created pivot tables using `pd.pivot_table()` for monthly regional comparison.

Step 6: Result Interpretation

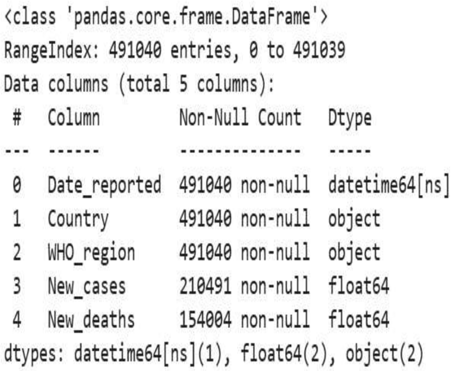
- Interpreted patterns in case distribution.
- Compared regional impacts and observed time-based changes.

5. Data Analysis and Results

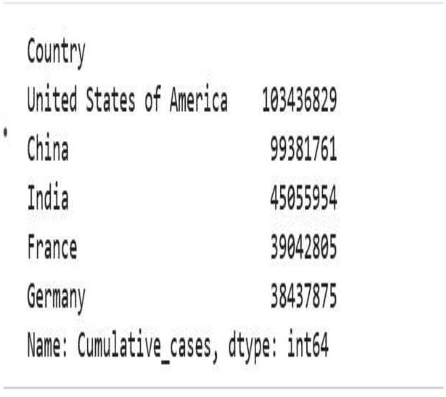
Dataset Overview

- Total Countries Covered: 240
- Date Range: 4 January 2020 – 10 August 2025
- Total Global Cases Reported: 778,491,932
- Total Global Deaths Reported: 7,099,904

This confirms that the dataset spans the entire pandemic period from its early outbreak to recent years.

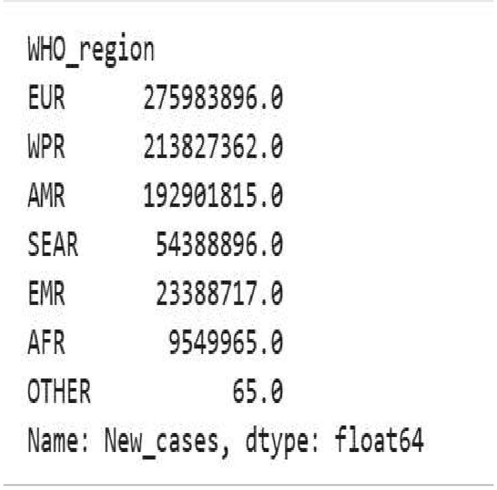


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 491040 entries, 0 to 491039
Data columns (total 5 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   Date_reported  491040 non-null  datetime64[ns]
 1   Country        491040 non-null  object
 2   WHO_region     491040 non-null  object
 3   New_cases      210491 non-null  float64
 4   New_deaths     154004 non-null  float64
dtypes: datetime64[ns](1), float64(2), object(2)
```



```
Country
United States of America    103436829
China                        99381761
India                        45055954
France                       39042805
Germany                      38437875
Name: Cumulative_cases, dtype: int64
```

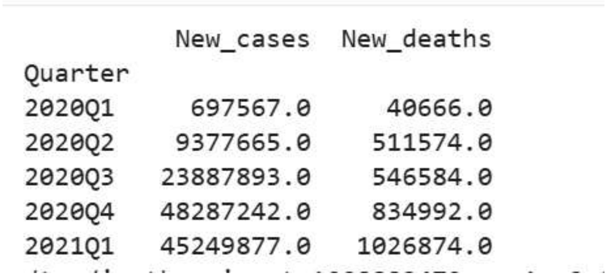
The United States recorded the highest cumulative cases globally, followed by China and India.



```
WHO_region
EUR      275983896.0
WPR      213827362.0
AMR      192901815.0
SEAR      54388896.0
EMR       23388717.0
AFR        9549965.0
OTHER           65.0
Name: New_cases, dtype: float64
```

The **European Region (EUR)** recorded the highest total cases, followed by the Western Pacific (WPR) and the Americas (AMR).

The African region (AFR) reported comparatively lower cases.



```
          New_cases  New_deaths
Quarter
2020Q1     697567.0     40666.0
2020Q2    9377665.0    511574.0
2020Q3   23887893.0    546584.0
2020Q4   48287242.0    834992.0
2021Q1   45249877.0   1026874.0
```

COVID-19 cases increased dramatically from Q1 to Q4 of 2020, indicating rapid global transmission. The highest number of cases during this period occurred in **2020 Q4**. The highest number of deaths occurred in **2021 Q1**, despite slightly lower case counts compared to Q4 2020.

6. Conclusion

Analysis of the WHO COVID-19 Global Dataset shows that the pandemic had a wave-like progression with several surges between 2020 and 2023. The most dramatic global surge occurred on 30 January 2022, when more than 8.4 million cases were reported in a single day. Europe was identified as the most impacted WHO region in terms of the total number of reported cases, while the African region had the lowest total number of cases.

The United States had the highest cumulative number of cases globally, followed by China and India. The data clearly shows the disparity in the number of reported cases across different regions, which could be affected by factors such as the size of the population, public health policies, testing facilities, and reporting mechanisms.

There is a clear drop in the number of reported cases after 2023, which indicates that the immunity levels, vaccination effects, or reporting intensity have improved. However, it is confirmed from the dataset that the global impact of COVID-19 was widespread and uneven.

The project has been able to successfully show the application of data cleaning, aggregation, and time-series analysis concepts using Python with valuable insights into the reporting trends of COVID-19.

7. APPENDICES

1. https://github.com/anushad-23/IDEASTIHSpringInternship
2. https://data.humdata.org
3. https://data.who.int
4. www.kaggle.com
5. Pandas official documentation
6. Research articles on Covid 19 Data