PROJECT TITLE

DATA CLEANING FOR A REAL-WORLD DATASET

NAME

SYED HABIB HAIDER

DATE OF SUBMISSION

12 MARCH 2025

AFFILIATION

BISTARTX

Data Cleaning Report: WHO COVID-19 Dataset

📌 1. Introduction

The **WHO COVID-19 Global Daily Data** dataset provides daily reports on COVID-19 cases and deaths across different countries. The dataset is essential for tracking the spread of the virus, analyzing trends, and making data-driven decisions for public health. However, real-world datasets often contain missing values, inconsistencies, and duplicates, which can lead to incorrect analysis.

This report outlines the **data cleaning process** applied to ensure the dataset is complete, structured, and analysis-ready.

📊 2. Exploratory Data Analysis (EDA)

Before cleaning the data, an initial examination was conducted to identify issues.

✅ 2.1 Summary of Missing Values

- `Country_code`: **1,878** missing values
- `New_cases`: **242,847** missing values
- `New_deaths`: **297,906** missing values

All other columns had **no missing values**.

✅ 2.2 Duplicate Records

- **0 duplicate rows** were found.

✅ 2.3 Data Types of Columns

Column Name	Data Type
`Date_reported`	Object (Date)
`Country_code`	Object
`Country`	Object
`WHO_region`	Object
`New_cases`	Float64
`Cumulative_cases`	Int64
`New_deaths`	Float64
`Cumulative_deaths`	Int64

🛠 3. Data Cleaning Process

◆ 3.1 Handling Missing Values

- **`Country_code`**: Missing values replaced with `"Unknown"`.
- **`New_cases` and `New_deaths`**: Missing values replaced with **0** to ensure accurate calculations.

◆ 3.2 Data Type Corrections

- The `Date_reported` column was converted to **datetime format** for proper analysis.

◆ 3.3 Final Cleaned Dataset

After the cleaning process, the dataset had **0 missing values and no inconsistencies**.

📊 4. Visualizations & Insights

To validate the cleaning process, the following visualizations were generated:

- **Missing Values Heatmap (Before Cleaning)**: Showed significant gaps in `New_cases` and `New_deaths`.
- **Missing Values Heatmap (After Cleaning)**: Confirmed that all missing values were successfully handled.
- **Distribution of COVID-19 Cases**: A histogram visualizing the spread of daily cases.

✅ 5. Conclusion

The data cleaning process ensured that the dataset is **complete, accurate, and ready for further analysis**. The cleaned data can now be used for:

- **Trend analysis of COVID-19 cases and deaths.**
- **Regional impact assessment based on WHO regions.**
- **Predictive modeling for COVID-19 case forecasting.**

The final cleaned dataset is stored as **`WHO-COVID-19-cleaned.csv`** and is available in the project repository.