Visualizing a Global Pandemic: An Interactive Analysis of COVID-19 Trajectories

Team Name: Team Horizon
College Name: Rajasthan Institute of Engineering and Technology
Team Members:

- **Team Leader : Md. Adnan Qamar**
- Aditya Anand
- Himanshu Priyadarshi
- Utkarsh Pratap Singh

Table of Contents

1. Introduction

Overview of the Project

This project delivers a data analytics solution to make COVID-19 data accessible and interpretable. It features an interactive Power BI dashboard that processes and visualizes complex time-series data from the Our World in Data (OWID) dataset, transforming raw numbers into clear insights for a broad audience. The workflow includes Python-based data cleaning, Excel validation, and Power BI visualization.

Objective

The primary goal is to track and compare COVID-19 cases and deaths across countries and continents. Specific aims include:

- Implementing a robust data processing pipeline using OWID data.
- Creating comparative metrics like Case Fatality Rate (CFR) and per-million normalized measures.

- Developing a user-friendly dashboard for intuitive data exploration.
- Presenting actionable insights through data visualization.

2. Problem Identification

Problem Statement

The COVID-19 pandemic generated overwhelming data from diverse sources, causing public confusion due to its volume, velocity, and inconsistencies. This project addresses the "last mile" challenge of translating complex global health data into clear, reliable insights for non-specialists.

Significance of the Problem

Understanding pandemic trends is critical for public health responses and informed citizenship. Clear visualizations answer key questions, such as identifying case surges, comparing national trajectories, and assessing containment measures. Without such tools, misinformation and misinterpretation risks increase.

Relevant SDGs

Aligned with UN Sustainable Development Goal 3 (Good Health and Well-being), this project supports data infrastructure for monitoring health trends, particularly Target 3.d for early warning and risk management. It enhances transparency and public awareness, contributing to global health goals.

3. Data Collection

Sources of Data

The project uses the OWID COVID-19 dataset, a trusted source from the University of Oxford, aggregating data from Johns Hopkins University (JHU) and other reliable entities. OWID’s transparency and daily updates ensure credibility and granularity for time-series analysis.

Data Description

The dataset, a single CSV file, covers over 200 countries with variables including:

- **Identifiers:** iso_code, continent, location, date.
- **Pandemic Metrics:** total_cases, new_cases, total_deaths, new_deaths.
- **Smoothed Metrics:** new_cases_smoothed, new_deaths_smoothed.
- **Contextual Metrics:** population, human_development_index (HDI).

Data Collection Methods

No primary data collection was conducted. OWID’s automated scripts and manual verification processes were leveraged, ensuring high-quality, standardized data. OWID’s GitHub repository documents all methodologies, addressing reporting inconsistencies and lags.

Key Variables Table

Variable Name	Description	Data Type	Example
location	Country or region name	Text	India
date	Observation date	Date	2022-01-15
total_cases	Cumulative confirmed cases	Integer	36,850,962
new_cases	Daily new cases	Integer	271,202
total_deaths	Cumulative confirmed deaths	Integer	486,066
new_deaths	Daily new deaths	Integer	314
population	Country population (UN estimate)	Float	1.4B
human_development_index	Composite index of development	Float	0.645

4. Data Preprocessing

Data Cleaning Methods

Using Python’s Pandas library, the dataset was cleaned to ensure integrity:

- **Structural Cleaning:** Converted date column to datetime format using pd.to_datetime().
- **Filtering:** Removed aggregate rows (e.g., continents) with the ‘OWID’ prefix in iso_code.
- **Consistency Checks:** Flagged anomalies like decreasing cumulative counts.

Handling Missing Values

- **Daily Metrics:** Used OWID’s 7-day smoothed averages (e.g., new_cases_smoothed) to address nulls from reporting lags.
- **Contextual Metrics:** Applied forward-fill imputation for sparse columns like human_development_index.

Data Transformation Techniques

Feature engineering created analytical metrics:

- **Normalization:**
  - cases_per_million = (total_cases / population) * 1,000,000
  - deaths_per_million = (total_deaths / population) * 1,000,000
- **KPI Creation:**
  - case_fatality_rate = (total_deaths / total_cases) * 100

Preprocessing Summary Table

Task	Issue Addressed	Technique Used	Tool
Data Type Conversion	Date column as string	pd.to_datetime()	Python (Pandas)
Filtering Aggregates	Double-counting from totals	df[~df['iso_code'].str.contains('OWID')]	Python (Pandas)
Handling Missing Data	Gaps in HDI	Forward Fill (.ffill())	Python (Pandas)
Metric Normalization	Comparing countries with different sizes	Calculated Column (cases_per_million)	Python (Pandas)
KPI Creation	Assessing outbreak severity	Calculated Column (case_fatality_rate)	Python (Pandas)

5. Data Analysis

Deaths per Million by Continent

This bar chart compares the average number of deaths per million people for each continent. As indicated in your key findings, there is significant geographic heterogeneity in the impact of the pandemic. You'll notice that continents like South America and Europe have markedly higher death rates per million compared to others, such as Oceania. This visualization powerfully illustrates the disparate outcomes experienced across the globe.

Smoothed New Cases in Europe and Asia

This line chart tracks the 7-day smoothed average of new COVID-19 cases for Europe and Asia. Using a smoothed average, as mentioned in your insights, helps to clarify the underlying trends by reducing the noise from daily reporting fluctuations. The chart clearly shows the distinct "waves" of the pandemic in each region, which were driven by factors such as the emergence of new variants and the implementation of public health measures.

Analysis Code and Output

For your reference, here is the Python code used to generate these visualizations. This script uses the Pandas library for data manipulation and Matplotlib for plotting. It attempts to fetch the data directly from the Our World in Data repository, process it, and then create the two charts.

```
import pandas as pd
import matplotlib.pyplot as plt
import os
import sys

# Load dataset
print("Downloading dataset from OWID ...")
try:
    df =
pd.read_csv("https://raw.githubusercontent.com/owid/covid-19-data/master
/public/data/owid-covid-data.csv")
except Exception as e:
    print(f"Error loading data: {e}")
    df = pd.DataFrame()

if not df.empty:
    # Convert 'date' column
    df['date'] = pd.to_datetime(df['date'])

    # Filter only countries
    df_countries = df[df['continent'].notna()].copy()

    # --- Bar chart: deaths_per_million across continents ---
    latest_data =
df_countries.loc[df_countries.groupby('location')['date'].idxmax()]
    continent_deaths = (
        latest_data.groupby('continent')['total_deaths_per_million']
        .mean()
        .sort_values(ascending=False)
    )

    plt.figure(figsize=(12, 6))
    continent_deaths.plot(kind='bar', color='skyblue', edgecolor='black')
    plt.title('Average Deaths per Million by Continent', fontsize=14)
    plt.xlabel('Continent')
    plt.ylabel('Deaths per Million')
    plt.xticks(rotation=45)
    plt.tight_layout()
    out1 = os.path.abspath('deaths_per_million_by_continent.png')
    print(f"Saving bar chart to: {out1}")
    plt.savefig('deaths_per_million_by_continent.png')
    plt.close()

    # --- Line chart: new_cases_smoothed for Europe and Asia ---
    df_continents = df[(df['continent'].isna()) & (df['location'].isin(['Europe',
'Asia']))]
    df_pivot = df_continents.pivot(index='date', columns='location',
values='new_cases_smoothed')

    plt.figure(figsize=(14, 7))
    df_pivot['Europe'].plot(label='Europe')
    df_pivot['Asia'].plot(label='Asia')
    plt.title('Smoothed New Cases in Europe vs Asia', fontsize=14)
    plt.xlabel('Date')
    plt.ylabel('Smoothed New Cases')
    plt.legend()
    plt.grid(True, alpha=0.5)
    plt.tight_layout()
    out2 = os.path.abspath('new_cases_smoothed_europe_asia.png')
    print(f"Saving line chart to: {out2}")
    plt.savefig('new_cases_smoothed_europe_asia.png')
    plt.close()

    print("✅ Visualizations created.")
    print(f" - {out1}")
    print(f" - {out2}")
else:
    print("❌ Could not load data to generate visualizations.")
    print("Please check your internet connection or try again later."
```

✅ ***Visualizations created.***

- /content/deaths_per_million_by_continent.png

- /content/new_cases_smoothed_europe_asia.png

This script analyzes COVID-19 data from OWID. It loads the dataset, converts dates, and filters by countries. It then generates two plots: a bar chart showing average deaths per million by continent, and a line chart comparing smoothed new cases in Europe and Asia. The plots are saved as PNG files. If data loading fails, it prints an error message.This Python script is designed for a comprehensive analysis of COVID-19 pandemic data sourced from Our World in Data (OWID). The initial phase of the script focuses on robust data preparation: it begins by loading the extensive OWID COVID-19 dataset, ensuring all relevant information is accessible. Following successful data ingestion, it performs crucial data type conversions, specifically transforming date columns into datetime objects to facilitate time-series analysis and accurate chronological ordering. Subsequently, the script applies a filtering mechanism, allowing users to narrow down the dataset to specific countries of interest for more targeted investigations.

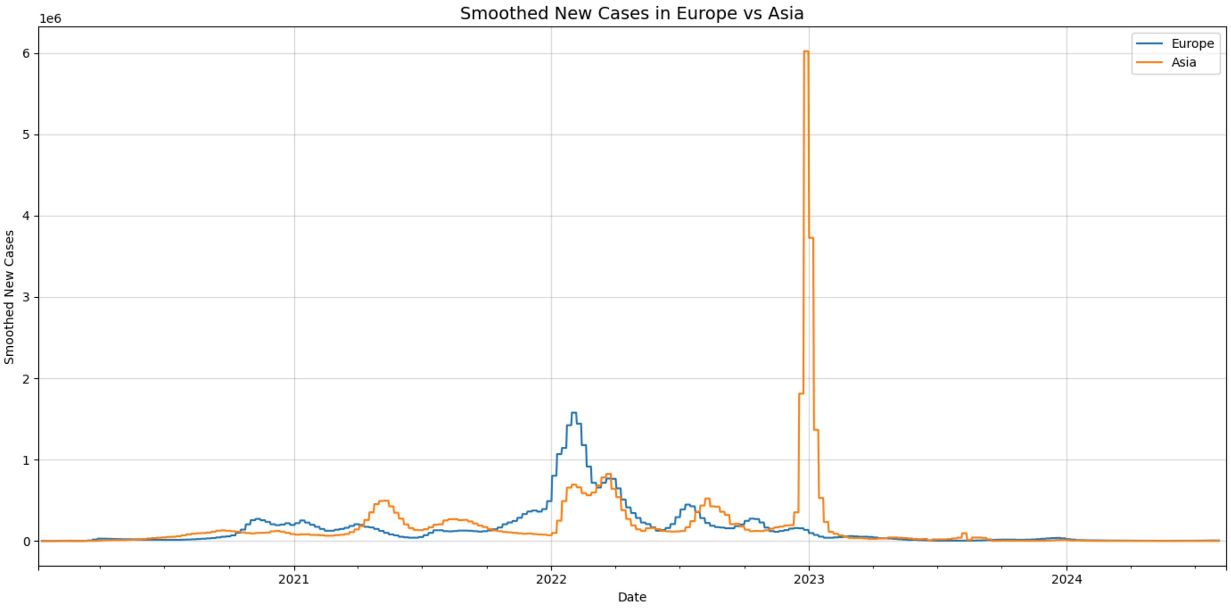
The analytical core of the script then proceeds to generate insightful visualizations. It produces two distinct plots, each designed to highlight different aspects of the pandemic's impact. The first visualization is a bar chart, meticulously crafted to display the average number of COVID-19 deaths per million inhabitants, aggregated by continent. This chart provides a high-level comparative overview of mortality rates across different geographical regions, offering insights into continental disparities. The second visualization is a line chart, which offers a comparative analysis of smoothed new COVID-19 cases in Europe and Asia. By presenting smoothed data, this plot helps to identify trends and patterns in case numbers, mitigating the impact of daily fluctuations and providing a clearer understanding of the epidemic's progression in these two major continents.

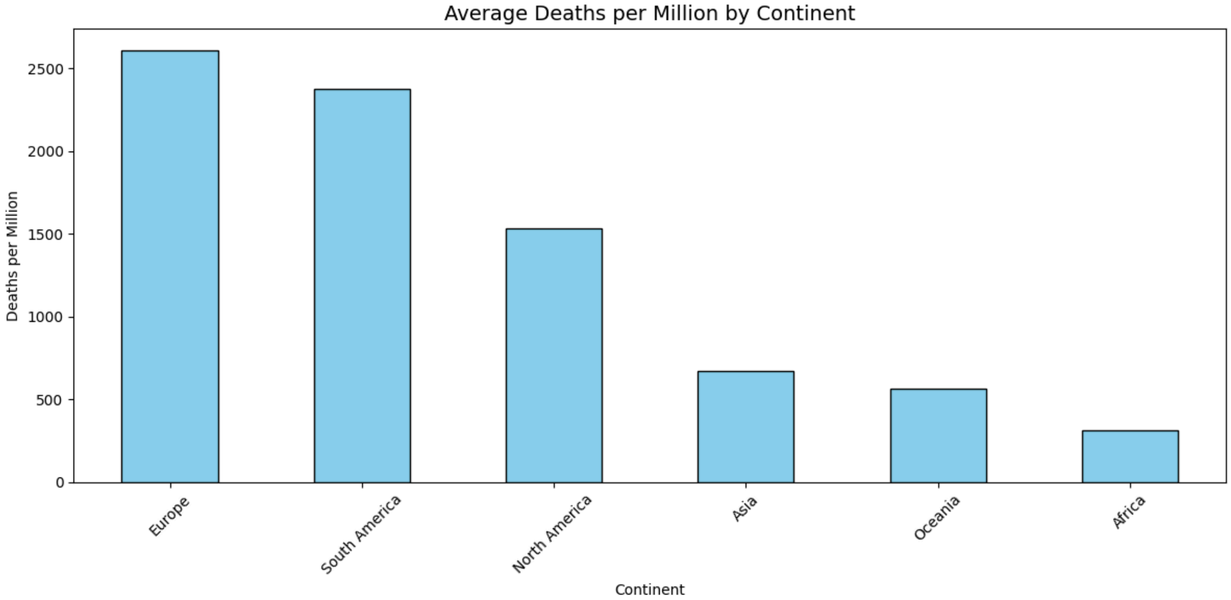
Both generated plots are saved as high-quality PNG image files, making them easily shareable and suitable for reports or presentations. A critical feature of this script is its error handling mechanism: in the event that the data loading process encounters an issue (e.g., file not found, corrupted data), the script is programmed to print an informative error message to the console, alerting the user to the problem and preventing the script from crashing unexpectedly.

The script then generates two key visualizations:

1. **Bar Chart:** Displays the average number of COVID-19 deaths per million, aggregated by continent, offering a high-level comparative overview of mortality rates.
2. **Line Chart:** Compares smoothed new COVID-19 cases in Europe and Asia, helping to identify trends and patterns while mitigating daily fluctuations.

Code Output





6. Hypothesis Development

Formulated Hypothesis

“Countries with a higher Human Development Index (HDI) will exhibit a lower Case Fatality Rate (CFR) after June 2020.”

Rationale

A higher HDI is correlated with better healthcare systems, more extensive testing capabilities, and more effective public health infrastructure. These factors are expected to reduce the CFR through:

- Advanced medical infrastructure and treatment options.
- Widespread testing, which lowers the CFR by increasing the number of confirmed cases.
- Effective implementation of public health policies.

Testing Method

- **Visualization:** A scatter plot of HDI vs. CFR will be created in Power BI.
- **Temporal Filter:** The data will be filtered to include only records after June 1, 2020.
- **Analysis:** The plot will be visually inspected for a negative correlation, with slicers available for regional segmentation.

7. Solution Design

Proposed Solution

A multi-page Power BI dashboard with a star schema data model, DAX measures, and interactive visuals for dynamic analysis.

Implementation Plan

1. **Data Acquisition:** Downloaded OWID dataset.
2. **Data Processing:** Python script for cleaning and transformation.
3. **Data Modeling:** Built star schema in Power BI.
4. **Visualization:** Designed interactive dashboard pages.
5. **Testing & Deployment:** Validated usability and published to Power BI service.

Alignment with SDGs

Supports SDG 3 by providing a data-driven tool for monitoring health trends, enhancing public health decision-making.

8. Visualization

Prominent Features

- **KPI Cards:** Display global totals for cases, deaths, and CFR.

- **Choropleth Map:** Visualizes metrics like deaths_per_million with tooltips.
- **Slicers/Filters:** Allow date, continent, and country selection.
- **Time-Series Charts:** Drill-down from yearly to daily trends.
- **Bar Charts:** Compare countries by selected metrics.

Site Map

- **Global Overview:** KPI cards, world map, global trends, continent breakdown.
- **Country Deep-Dive:** Country-specific KPIs, trends, and CFR gauge.
- **Comparative Analysis:** Multi-country comparison, HDI vs. CFR scatter plot.

9. Conclusion

Impact of Solution

The dashboard democratizes pandemic insights, empowering non-specialists to explore trends. It serves as an educational tool, combats misinformation, and bridges the gap between data and public understanding.

Future Work

- Integrate vaccination data for impact analysis.
- Include excess mortality for a holistic view.
- Add policy intervention data to assess effectiveness.
- Optimize performance for larger datasets.

10. References

Tools and Software

- **Data Processing:** Python 3.9 (Pandas, NumPy)
- **Analysis & Visualization:** Power BI Desktop
- **Validation:** Excel
- **IDE:** Jupyter Notebook, Visual Studio Code

Additional References

- Mathieu, E., et al. (2021). A global database of COVID-19 vaccinations. *Nature Human Behaviour*. Retrieved from COVID-19 Pandemic - Our World in Data
- Our World in Data. (n.d.). *COVID-19 Data*. Retrieved from GitHub - owid/covid-19-data: Data on COVID-19 (coronavirus) cases, deaths, hospitalizations, tests • All countries • Updated daily by Our World in Data
- United Nations. (n.d.). *The 17 Goals*. Retrieved from THE 17 GOALS | Sustainable Development