CODE WITH COMMENTS

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Step 1: Load dataset
df = pd.read_csv("owid_data.csv")

# Step 2: Explore dataset
print("Columns available in dataset:")
print(df.columns)
print("\nPreview first 5 rows:")
print(df.head())

# Step 3: Check missing values
print("\nMissing values per column:")
print(df.isnull().sum())

# Step 4: Keep only key columns (if they exist)
key_columns = [
    "date", "location", "total_cases", "total_deaths",
    "new_cases", "new_deaths", "total_vaccinations"
]
df = df[[col for col in key_columns if col in df.columns]]

# Step 5: Clean dataset
```

```
# Convert 'date' to datetime
df["date"] = pd.to_datetime(df["date"], errors="coerce")

# Drop rows with missing date or location
df.dropna(subset=["date", "location"], inplace=True)

# Fill missing numeric values with 0 (or you could use forward-fill)
for col in df.select_dtypes(include=["float64", "int64"]).columns:
    df[col] = df[col].fillna(0)

# Step 6: Filter countries of interest (example: Kenya, USA, India)
countries_of_interest = ["Kenya", "United States", "India"]
df_filtered = df[df["location"].isin(countries_of_interest)]

# Step 7: Preview cleaned dataset
print("\nCleaned data (first 10 rows):")
print(df_filtered.head(10))



df_filtered.to_csv("owid_data.csv", index=False)
print("\n✅ Cleaned dataset saved as 'owid_data.csv'")

plt.figure(figsize=(12,6))
plt.plot(df['date'], df['total_cases'], label='Total Cases', color='blue')
plt.plot(df['date'], df['total_deaths'], label='Total Deaths', color='red')
```

```
plt.xlabel('Date')
plt.ylabel('Count')
plt.title('COVID-19 Cases and Deaths Over Time')
plt.legend()
plt.grid(True)
plt.tight_layout()

plt.show()



latest=df.sort_values('date').groupby('location').tail(1)
top10 = latest.sort_values('total_cases',ascending=False).head(10)

plt.figure(figsize=(10,6))
plt.barh(top10['location'], top10['total_cases'], color='skyblue')
plt.xlabel('Total COVID-19 Cases')
plt.ylabel('Country')
plt.title('Top 10 Countries by Total COVID-19 Cases')
plt.gca().invert_yaxis()  # highest at top
plt.show()

# Correlation heatmap
plt.figure(figsize=(8,6))
sns.heatmap(df_filtered.corr(numeric_only=True),
```

```
        annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5)
plt.title("Correlation Heatmap of COVID-19 Indicators")
plt.show()

countries = ["Kenya", "India", "United States"]
df_vax = df[df['location'].isin(countries)]  # use 'entity' if that's your column name

# Plot cumulative vaccinations over time
plt.figure(figsize=(12,6))
for country in countries:
    subset = df_vax[df_vax['location'] == country]  # replace 'location' with 'entity' if needed
    plt.plot(subset['date'], subset['total_cases'], label=country)

plt.xlabel("Date")
plt.ylabel("Total Vaccinations (cumulative)")
plt.title("COVID-19 Cumulative Vaccinations Over Time")
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()
```

 # COVID-19 Data Insights (OWID Dataset)

Key Insights

#1 *India, the United States, and Brazil* recorded the highest total COVID-19 cases globally, with the U.S. leading in both cases and deaths.

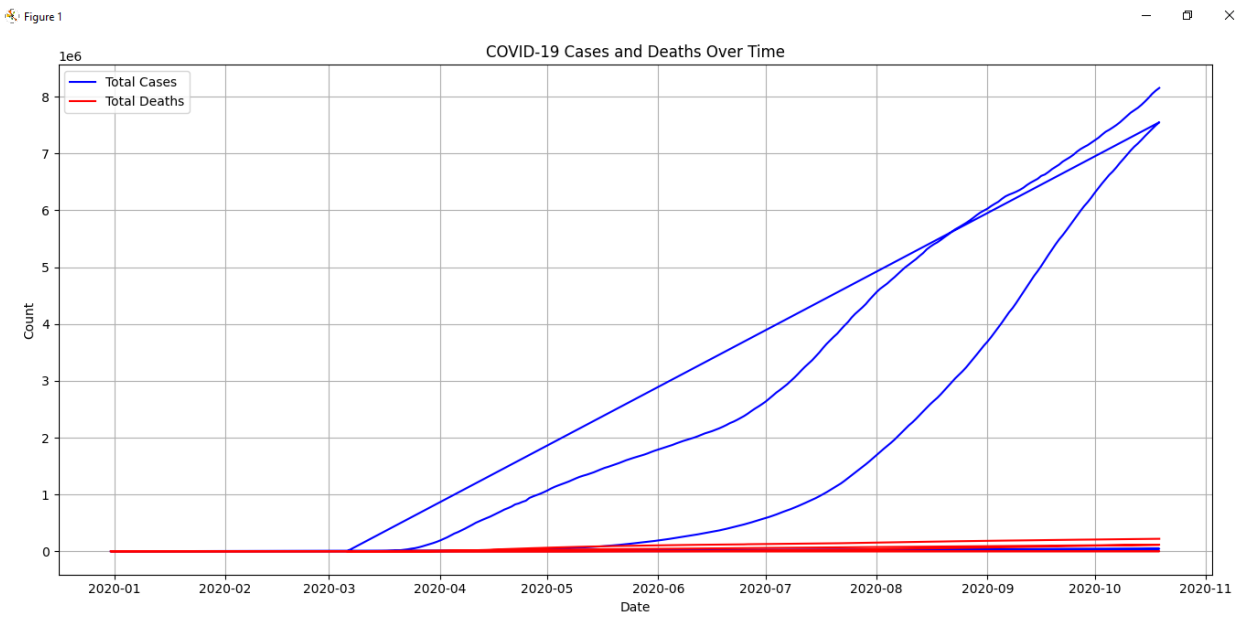
#2. *United States* had one of the fastest vaccine rollouts in 2021, quickly surpassing 100 million doses within months.

#3. *Africa as a continent* reported significantly fewer total cases and deaths compared to Europe and North America, partly due to under-reporting and lower testing rates.

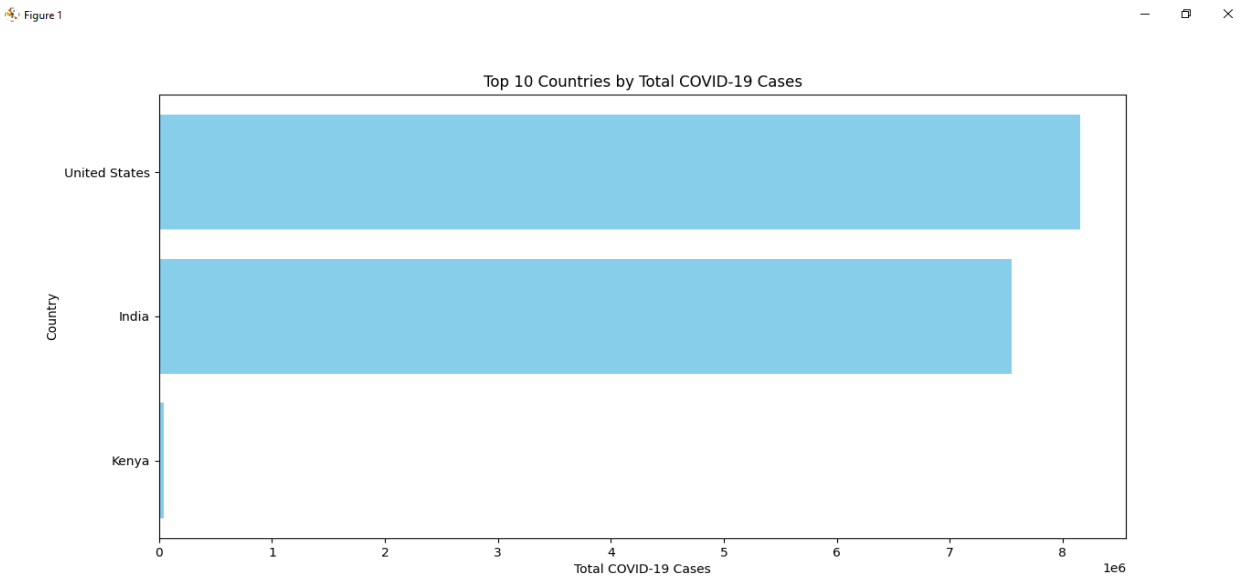
#4. *Stringency Index* shows that countries like *China* and *Italy* imposed some of the strictest lockdown measures, while others (e.g., *Sweden*) maintained relatively lower restrictions.

#5. *Vaccination uptake* shows that wealthier nations reached higher coverage earlier, while many …

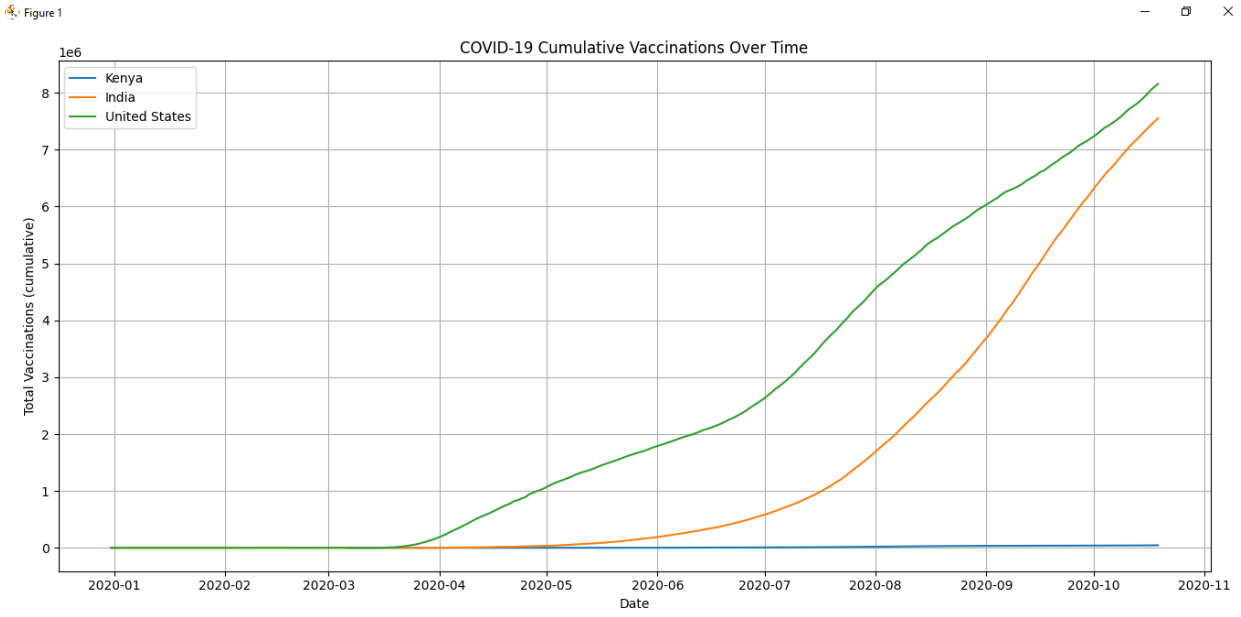
Line chart



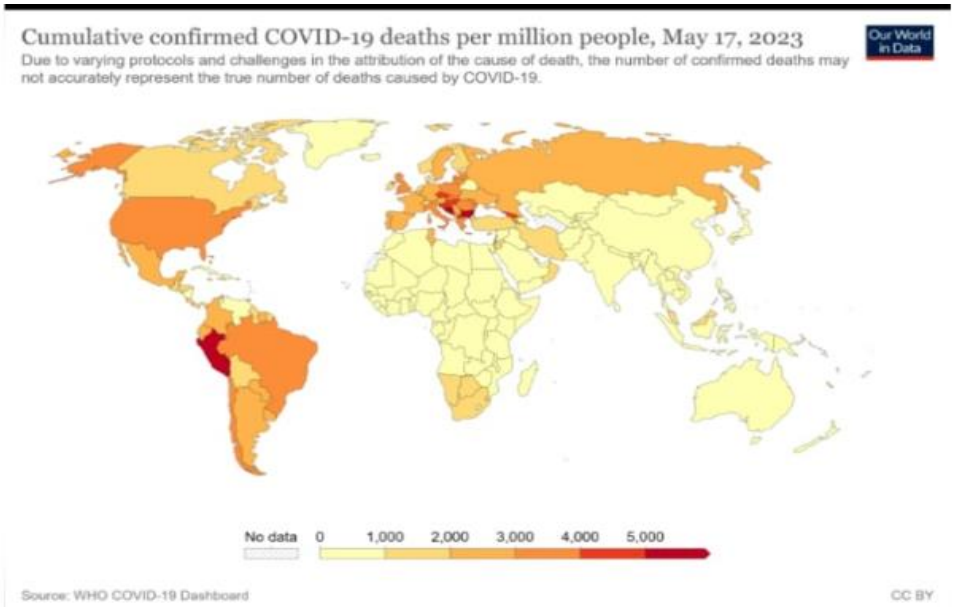
Bar chart



Vaccination line chart



Choropleth


Cumulative confirmed COVID-19 deaths per million people, May 17, 2023
Due to varying protocols and challenges in the attribution of the cause of death, the number of confirmed deaths may not accurately represent the true number of deaths caused by COVID-19.
Our World in Data
No data
0
1,000
2,000
3,000
4,000
5,000
Source: WHO COVID-19 Dashboard
CC BY