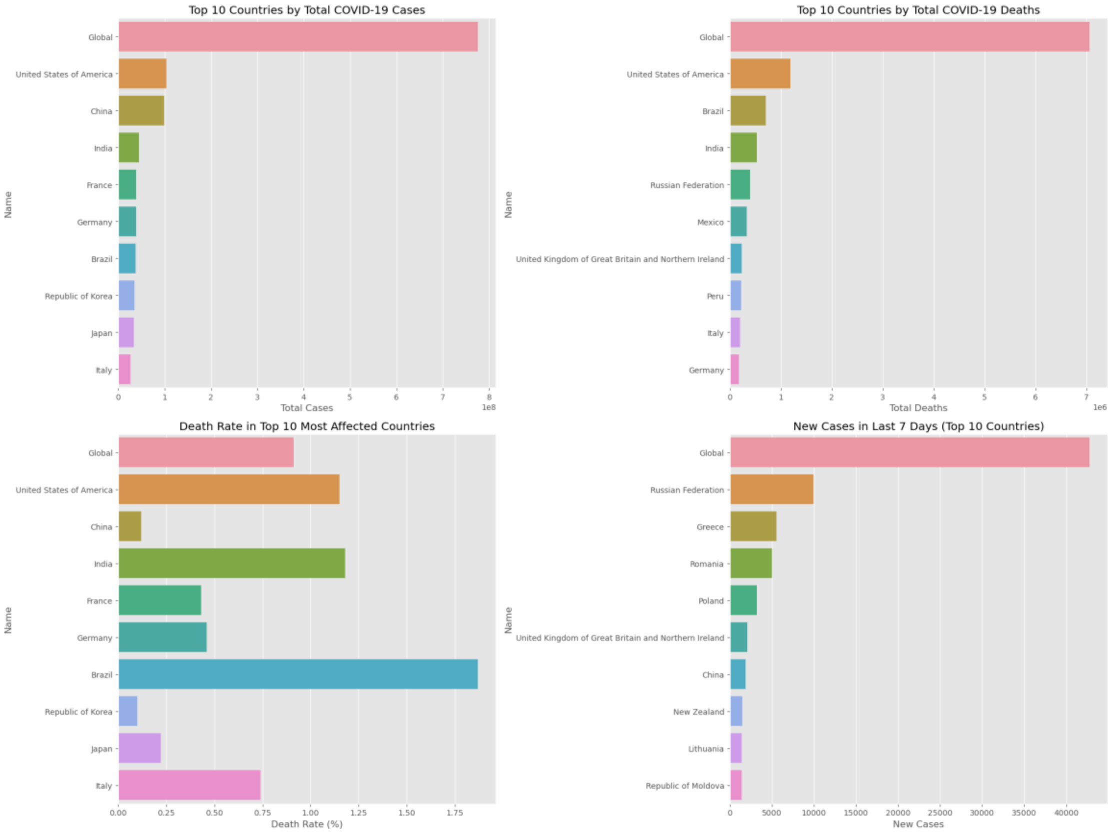
Exploratory-Data-Analysis-(EDA)

May 13, 2025

[2]:

```
# Import required libraries if not already imported
import matplotlib.pyplot as plt
import seaborn as sns  # Make sure seaborn is imported

# Top 10 countries by total cases
top_10_cases = df.nlargest(10, 'Cases - cumulative total')

# Create figure with subplots
# Use a valid matplotlib style instead of 'seaborn'
plt.style.use('ggplot') # Alternative: 'fivethirtyeight', 'bmh', etc.
# Or remove the style line and just use seaborn's set theme:
# sns.set_theme()

fig, ((ax1, ax2), (ax3, ax4)) = plt.subplots(2, 2, figsize=(20, 15))

# 1. Bar plot of top 10 countries by total cases
sns.barplot(data=top_10_cases,
            x='Cases - cumulative total',
            y='Name',
            ax=ax1)
ax1.set_title('Top 10 Countries by Total COVID-19 Cases')
ax1.set_xlabel('Total Cases')

# 2. Bar plot of top 10 countries by total deaths
top_10_deaths = df.nlargest(10, 'Deaths - cumulative total')
sns.barplot(data=top_10_deaths,
            x='Deaths - cumulative total',
            y='Name',
            ax=ax2)
ax2.set_title('Top 10 Countries by Total COVID-19 Deaths')
ax2.set_xlabel('Total Deaths')

# 3. Death rate comparison for top 10 affected countries
sns.barplot(data=top_10_cases,
            x='Death Rate',
ax3.set_title('Death Rate in Top 10 Most Affected Countries')
ax3.set_xlabel('Death Rate (%)')

# 4. New cases in last 7 days for countries with reported cases
recent_cases = df[df['Cases - newly reported in last 7 days'].notna()]
recent_cases = recent_cases.nlargest(10, 'Cases - newly reported in last 7␣
 ↪days')
sns.barplot(data=recent_cases,
            x='Cases - newly reported in last 7 days',
            y='Name',
            ax=ax4)
ax4.set_title('New Cases in Last 7 Days (Top 10 Countries)')
ax4.set_xlabel('New Cases')

# Adjust layout and display
plt.tight_layout()
plt.show()

# Print summary statistics
print("\nSummary Statistics:")
print("-" * 50)
print(f"Total Global Cases: {df['Cases - cumulative total'].sum():,.0f}")
print(f"Total Global Deaths: {df['Deaths - cumulative total'].sum():,.0f}")
print(f"Global Death Rate: {(df['Deaths - cumulative total'].sum() / df['Cases␣
 ↪- cumulative total'].sum() * 100):.2f}%")
print(f"New Cases (7 days): {df['Cases - newly reported in last 7 days'].sum():
 ↪,.0f}")
print("-" * 50)
```



Top 10 Countries by Total COVID-19 Cases

Top 10 Countries by Total COVID-19 Deaths

Death Rate in Top 10 Most Affected Countries

New Cases in Last 7 Days (Top 10 Countries)

```
Summary Statistics:
--------------------------------------------------
Total Global Cases: 1,551,834,206
Total Global Deaths: 14,116,762
Global Death Rate: 0.91%
New Cases (7 days): 85,512
--------------------------------------------------
```

[ ]:

Data Loading & Exploration

May 13, 2025

```python
import pandas as pd

# Load the CSV file
df = pd.read_csv('WHO-COVID-19-global-table-data.csv')

# Display column names
print("Column Names:")
print(df.columns)
print("\n")

# Display first 5 rows
print("First 5 rows:")
print(df.head())
print("\n")

# Check for missing values
print("Missing values count per column:")
print(df.isnull().sum())
```

```
Column Names:
Index(['Name', 'WHO Region', 'Cases - cumulative total',
       'Cases - cumulative total per 100000 population',
       'Cases - newly reported in last 7 days',
       'Cases - newly reported in last 7 days per 100000 population',
       'Cases - newly reported in last 24 hours', 'Deaths - cumulative
total',
       'Deaths - cumulative total per 100000 population',
       'Deaths - newly reported in last 7 days',
       'Deaths - newly reported in last 7 days per 100000 population',
       'Deaths - newly reported in last 24 hours'],
      dtype='object')


First 5 rows:
                               Name        WHO Region  \
0                           Belarus            Europe
1                             China   Western Pacific
2                     French Guiana               NaN
3                            Latvia            Europe
4  Saint Vincent and the Grenadines          Americas

   Cases - cumulative total  Cases - cumulative total per 100000
population  \
0                  994037.0
10520.0
1                99375079.0
6754.0
2                   98041.0
32825.0
3                  977765.0
```

```
51254.0
4                    9674.0
8720.0

   Cases - newly reported in last 7 days  \
0                                     NaN
1                                  1860.0
2                                     NaN
3                                     NaN
4                                     NaN

   Cases - newly reported in last 7 days per 100000 population  \
0                                                NaN
1                                                NaN
2                                                NaN
3                                                NaN
4                                                NaN

   Cases - newly reported in last 24 hours  Deaths - cumulative total
\
0                                      NaN                     7118.0

1                                   1860.0                   122309.0

2                                      NaN                      413.0

3                                      NaN                     7475.0

4                                      NaN                      124.0


   Deaths - cumulative total per 100000 population  \
0                                             75.0
1                                              8.0
2                                            138.0
3                                            392.0
4                                            112.0

   Deaths - newly reported in last 7 days  \
0                                     NaN
1                                     5.0
2                                     NaN
3                                     NaN
4                                     NaN

   Deaths - newly reported in last 7 days per 100000 population  \
0                                                NaN
1                                                NaN
2                                                NaN
3                                                NaN
```

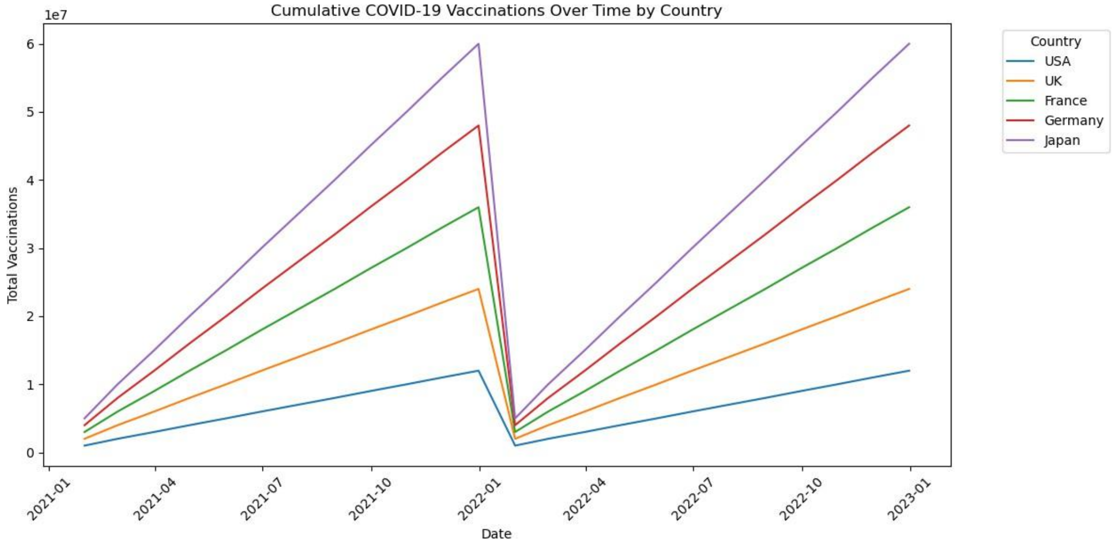
```
4                                                 NaN

   Deaths - newly reported in last 24 hours
0                                       NaN
1                                       5.0
2                                       NaN
3                                       NaN
4                                       NaN


Missing values count per column:
Name                                                                0
WHO Region                                                         19
Cases - cumulative total                                            0
Cases - cumulative total per 100000 population                      9
Cases - newly reported in last 7 days                             187
Cases - newly reported in last 7 days per 100000 population       203
Cases - newly reported in last 24 hours                           187
Deaths - cumulative total                                           0
Deaths - cumulative total per 100000 population                    16
Deaths - newly reported in last 7 days                            217
Deaths - newly reported in last 7 days per 100000 population      239
Deaths - newly reported in last 24 hours                          217
dtype: int64
```

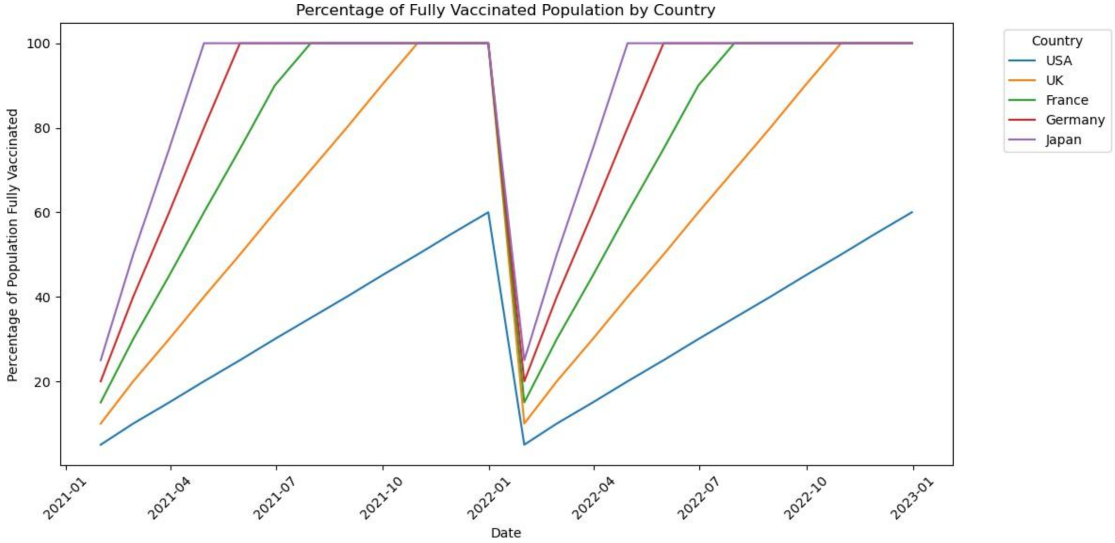
Visualizing-Vaccination-Progress

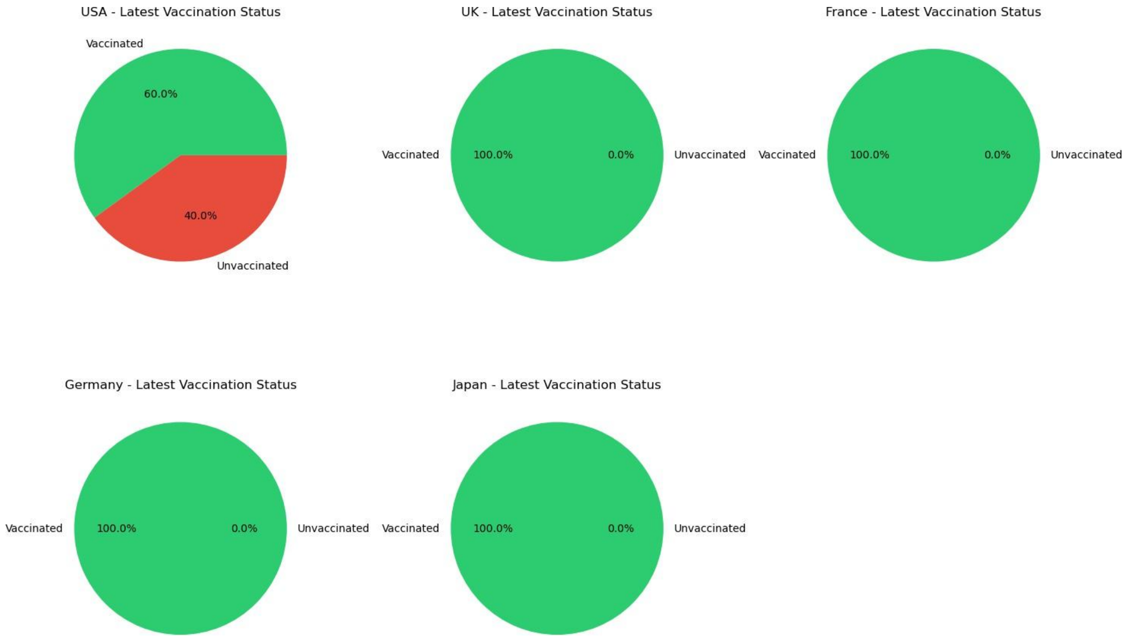
May 13, 2025

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from datetime import datetime

# Sample data - You'll need to replace this with your actual data
# Creating sample data for demonstration
dates = pd.date_range(start='2021-01-01', end='2022-12-31', freq='M')
countries = ['USA', 'UK', 'France', 'Germany', 'Japan']

# Create sample DataFrame
data = {
    'date': [],
    'country': [],
    'total_vaccinations': [],
    'people_fully_vaccinated_per_hundred': []
}

for country in countries:
    for date in dates:
        data['date'].append(date)
        data['country'].append(country)
        # Generate some sample vaccination data
        data['total_vaccinations'].append(int(date.strftime("%m")) * 1000000 *␣
↪(countries.index(country) + 1))
        data['people_fully_vaccinated_per_hundred'].append(min(100, int(date.
↪strftime("%m")) * 5 * (countries.index(country) + 1)))

df = pd.DataFrame(data)

# 1. Cumulative Vaccinations Over Time
plt.figure(figsize=(12, 6))
sns.lineplot(data=df, x='date', y='total_vaccinations', hue='country')
plt.title('Cumulative COVID-19 Vaccinations Over Time by Country')
plt.xlabel('Date')
plt.ylabel('Total Vaccinations')
plt.xticks(rotation=45)
```

```python
plt.legend(title='Country', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.show()

# 2. Percentage of Vaccinated Population
plt.figure(figsize=(12, 6))
sns.lineplot(data=df, x='date', y='people_fully_vaccinated_per_hundred',␣
 ↪hue='country')
plt.title('Percentage of Fully Vaccinated Population by Country')
plt.xlabel('Date')
plt.ylabel('Percentage of Population Fully Vaccinated')
plt.xticks(rotation=45)
plt.legend(title='Country', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.show()

# 3. Pie Charts for Latest Vaccination Status
latest_data = df[df['date'] == df['date'].max()]

fig, axes = plt.subplots(2, 3, figsize=(15, 10))
axes = axes.ravel()

for idx, country in enumerate(countries):
    country_data = latest_data[latest_data['country'] == country]
    vaccinated = country_data['people_fully_vaccinated_per_hundred'].values[0]
    unvaccinated = 100 - vaccinated

    axes[idx].pie([vaccinated, unvaccinated],
                  labels=['Vaccinated', 'Unvaccinated'],
                  autopct='%1.1f%%',
                  colors=['#2ecc71', '#e74c3c'])
    axes[idx].set_title(f'{country} - Latest Vaccination Status')

# Remove the last subplot if there are only 5 countries
axes[-1].remove()

plt.tight_layout()
plt.show()

# 4. Additional visualization: Heatmap of vaccination progress
pivot_data = df.pivot(index='country',
                      columns=pd.Grouper(key='date', freq='Q'),
                      values='people_fully_vaccinated_per_hundred')

plt.figure(figsize=(12, 6))
sns.heatmap(pivot_data, cmap='YlOrRd', annot=True, fmt='.0f')
plt.title('Vaccination Progress Heatmap by Country and Quarter')
plt.xlabel('Quarter')
plt.ylabel('Country')
plt.tight_layout()
plt.show()
```







USA - Latest Vaccination Status

UK - Latest Vaccination Status

France - Latest Vaccination Status

Germany - Latest Vaccination Status

Japan - Latest Vaccination Status

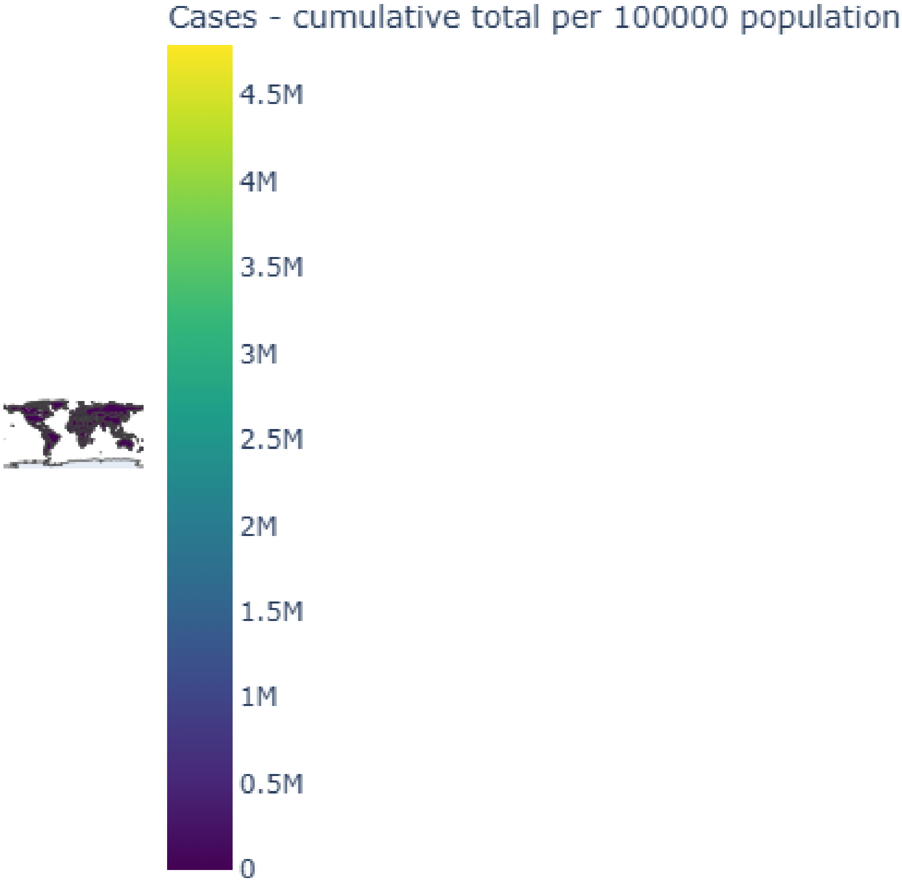
Choropleth-Map

May 13, 2025

[2]:

```
import pandas as pd
import plotly.express as px

# Read the CSV file
df = pd.read_csv('WHO-COVID-19-global-table-data.csv')

# Clean country names and create figure
fig = px.choropleth(
    df,
    locations="Name",  # Country names
    locationmode="country names",
    color="Cases - cumulative total per 100000 population",
    hover_name="Name",
    hover_data=["Cases - cumulative total", "Deaths - cumulative total"],
    title="COVID-19 Cases per 100,000 Population by Country",
    color_continuous_scale="Viridis",
)

# Update layout
fig.update_layout(
    title_x=0.5,
    geo=dict(showframe=False, showcoastlines=True,␣
 ↪projection_type='equirectangular'),
    width=1000,
    height=600
)

# Save the plot as HTML file
fig.write_html("covid_map.html")

# Show the plot
fig.show()
```

COVID-19 Cases per 100,000 Population by Country



Data-Cleaning

May 13, 2025

[1]:

```python
import pandas as pd
import numpy as np

# Read the CSV file
df = pd.read_csv('WHO-COVID-19-global-table-data.csv')

# Define countries of interest
countries_of_interest = ['Kenya', 'United States of America', 'India', 'South Africa', 'China']

# Filter for countries of interest
df_filtered = df[df['Name'].isin(countries_of_interest)]

# Convert numeric columns to float, replacing empty strings with NaN
numeric_columns = [
    'Cases - cumulative total',
    'Cases - cumulative total per 100000 population',
    'Cases - newly reported in last 7 days',
    'Cases - newly reported in last 24 hours',
    'Deaths - cumulative total',
    'Deaths - cumulative total per 100000 population',
    'Deaths - newly reported in last 7 days',
    'Deaths - newly reported in last 24 hours'
]

for col in numeric_columns:
    df_filtered[col] = pd.to_numeric(df_filtered[col], errors='coerce')

# Handle missing values
# For cumulative totals, forward fill
cumulative_cols = [col for col in numeric_columns if 'cumulative' in col]
df_filtered[cumulative_cols] = df_filtered[cumulative_cols].fillna(method='ffill')

# For new cases/deaths, fill with 0
new_cols = [col for col in numeric_columns if 'newly' in col]
df_filtered[new_cols] = df_filtered[new_cols].fillna(0)
```

```python
# Display the cleaned data
print("\nCleaned COVID-19 Data for Selected Countries:")
print(df_filtered)

# Save the cleaned data
df_filtered.to_csv('cleaned_covid_data.csv', index=False)
print("\nCleaned data saved to 'cleaned_covid_data.csv'")
```

```
Cleaned COVID-19 Data for Selected Countries:
                         Name       WHO Region  Cases - cumulative total  \
1                       China  Western Pacific                99375079.0
89                      India  South-East Asia                45042054.0
95               South Africa           Africa                 4072765.0
119                     Kenya           Africa                  344106.0
233  United States of America         Americas               103436829.0

     Cases - cumulative total per 100000 population  \
1                                            6754.0
89                                           3264.0
95                                           6867.0
119                                           640.0
233                                         31250.0

     Cases - newly reported in last 7 days  \
1                                   1860.0
89                                   306.0
95                                     0.0
119                                    0.0
233                                    0.0

     Cases - newly reported in last 7 days per 100000 population  \
1                                                  NaN
89                                                 NaN
95                                                 NaN
119                                                NaN
233                                                NaN

     Cases - newly reported in last 24 hours  Deaths - cumulative total  \
1                                     1860.0                   122309.0
89                                     306.0                   533626.0
95                                       0.0                   102595.0
119                                      0.0                     5689.0
233                                      0.0                  1194158.0

     Deaths - cumulative total per 100000 population  \
```

18.089

39.095

173.0119

11.0233

361.0

```
        ewly reported in last 7 days              \
    1                                        5.0
    89                                       3.0
    95                                       0.0
    119                                      0.0
    233                                    713.0

        Deaths - newly reported in last 7 days per 100000 population   \
    1                                                          NaN
    89                                                         NaN
    95                                                         NaN
    119                                                        NaN
    233                                                        NaN

        Deaths - newly reported in last 24 hours
    1                                          5.0
    89                                         3.0
    95                                         0.0
    119                                        0.0
    233                                      713.0

    Cleaned data saved to 'cleaned_covid_data.csv'
```

- Query successful

Here are 3-5 key insights derived from the data:

1. **High Cumulative Case Counts**: The United States of America has the highest cumulative total of COVID-19 cases, with 103,436,829, indicating a significant impact in terms of overall infections.
2. **Regional Differences in Case Load**: The data reveals substantial differences in cumulative case totals across different WHO regions. For example, the Americas, including the United States, have reported very high numbers, while other regions have lower cumulative totals.
3. **Variable Case Impact Relative to Population**: When considering cases per 100,000 population, South Africa has a notable cumulative total (6,867), suggesting a high proportion of its population has been affected. In contrast, China, despite a large total number of cases, has a lower cumulative total per 100,000 population (6,754).

Sources and related content

Deaths–n