covid19-analysis-vishrut-sharma

October 13, 2025

1 Generative AI Assignment

2 Vishrut Sharma

3 GF202566984

4 Submitted to MR Gaurav Kumar

5 # Step 1 = Firstly we will import libraries

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

6 —-The Pandas library is imported to manage and manipulate datasets, Matplotlib and Seaborn are imported for effective data visualization. These libraries provide essential tools for data handling, plotting graphs, and making it easier to interpret the COVID-19 data——

7 # Step 2 = now we will load dataset

[2]:
```
df = pd.read_csv('/Users/vishrutsharma/Downloads/covid_data.csv')
```

8 ——The COVID-19 dataset is loaded using Pandas read_csv function into a DataFrame named df . Loading the data into a structured format enables efficient processing and analysis——

9 # Step 3 = now we will explore data

[3]:
```
print(df.head(10))  # First 10 rows
```

```
              Country          Other names ISO 3166-1 alpha-3 CODE  \
0         Afghanistan          Afghanistan                     AFG
1             Albania              Albania                     ALB
2             Algeria              Algeria                     DZA
3             Andorra              Andorra                     AND
4              Angola               Angola                     AGO
5            Anguilla             Anguilla                     AIA
6 Antigua and Barbuda  Antigua and Barbuda                     ATG
7           Argentina            Argentina                     ARG
8             Armenia              Armenia                     ARM
9               Aruba                Aruba                     ABW

   Population                        Continent  Total Cases  Total Deaths  \
0    40462186                             Asia       177827          7671
1     2872296                           Europe       273870          3492
2    45236699                           Africa       265691          6874
3       77481                           Europe        40024           153
4    34654212                           Africa        99194          1900
5       15237  Latin America and the Caribbean         2700             9
6       99348  Latin America and the Caribbean         7493           135
7    45921761  Latin America and the Caribbean      9041124        128065
8     2972939                             Asia       422574          8617
9      107560  Latin America and the Caribbean        34051           212

   Tot Cases//1M pop  Tot Deaths/1M pop  Death percentage
0               4395                190          4.313743
1              95349               1216          1.275058
2               5873                152          2.587216
3             516565               1975          0.382271
4               2862                 55          1.915438
5             177200                591          0.333333
6              75422               1359          1.801682
7             196881               2789          1.416472
8             142140               2898          2.039169
9             316577               1971          0.622596
```

[4]:
```
print(df.tail(5))  # Last 5 rows
```

```
               Country               Other names ISO 3166-1 alpha-3 CODE  \
220  Wallis and Futuna  Wallis and Futuna Islands                     WLF
221     Western Sahara             Western Sahara                   ESHÂ
222              Yemen                      Yemen                     YEM
223             Zambia                     Zambia                     ZMB
224           Zimbabwe                   Zimbabwe                     ZWE

     Population Continent  Total Cases  Total Deaths  Tot Cases//1M pop  \
220       10894   Oceania          454             7              41674
221      623031    Africa           10             1                 16
222    30975258      Asia        11806          2143                381
223    19284482    Africa       317076          3967              16442
224    15241601    Africa       246525          5446              16174

     Tot Deaths/1M pop  Death percentage
220                643          1.541850
221                  2         10.000000
222                 69         18.151787
223                206          1.251120
224                357          2.209107
```

[5]:
```python
print("Dataset shape:", df.shape)
```

```
Dataset shape: (225, 10)
```

[6]:
```python
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 225 entries, 0 to 224
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Country                225 non-null    object
 1   Other names            224 non-null    object
 2   ISO 3166-1 alpha-3 CODE  225 non-null    object
 3   Population             225 non-null    int64
 4   Continent              225 non-null    object
 5   Total Cases            225 non-null    int64
 6   Total Deaths           225 non-null    int64
 7   Tot Cases//1M pop      225 non-null    int64
 8   Tot Deaths/1M pop      225 non-null    int64
 9   Death percentage       225 non-null    float64
dtypes: float64(1), int64(5), object(4)
memory usage: 17.7+ KB
None
```

[8]:
```python
print(df.describe())
```

```
         Population   Total Cases  Total Deaths  Tot Cases//1M pop  \
count  2.250000e+02  2.250000e+02  2.250000e+02         225.000000
mean   3.507321e+07  2.184781e+06  2.744813e+04      136900.373333
std    1.392418e+08  7.275938e+06  9.689177e+04      145060.340289
min    8.050000e+02  1.000000e+00  0.000000e+00           9.000000
25%    5.665570e+05  2.407100e+04  1.890000e+02       11384.000000
50%    5.827911e+06  1.639360e+05  1.965000e+03       88987.000000
75%    2.190585e+07  1.092547e+06  1.366000e+04      223335.000000
max    1.439324e+09  8.183905e+07  1.008222e+06      696044.000000

       Tot Deaths/1M pop  Death percentage
count         225.000000        225.000000
mean         1096.715556          1.444125
std          1195.715543          1.741728
min             0.000000          0.000000
25%           123.000000          0.511291
50%           708.000000          1.036905
75%          1795.000000          1.977017
max          6286.000000         18.151787
```

10 —Displayed the first 10 rows to understand the initial data format and the last 5 rows to check consistency at the end of the dataset. This confirms that the data contains expected columns like Country, Population, Total Cases, and Deaths——

11 #Step 4 = Missing Data Analysis and Handling

[9]:
```python
print(df.isnull().sum())  # here we will count missing values per column
```

```
Country                  0
Other names              1
ISO 3166-1 alpha-3 CODE  0
Population               0
Continent                0
Total Cases              0
Total Deaths             0
Tot Cases//1M pop        0
Tot Deaths/1M pop        0
Death percentage         0
dtype: int64
```

[10]:
```python
print(df[df.isnull().any(axis=1)])  # the rows with any missing value
```

```
        Country Other names ISO 3166-1 alpha-3 CODE  Population Continent  \
135  Montenegro         NaN                     MNE      628205    Europe
```

```
     Total Cases  Total Deaths  Tot Cases//1M pop  Tot Deaths/1M pop  \
135       233326          2705             371417               4306

     Death percentage
135          1.159322
```

[11]:
```python
df.ffill(inplace=True) # now we will fill missing values by forward fill
```

[12]:
```python
print(df.isnull().sum())  # here we will confirm no missing values remain
```

```
Country                    0
Other names                0
ISO 3166-1 alpha-3 CODE    0
Population                 0
Continent                  0
Total Cases                0
Total Deaths               0
Tot Cases//1M pop          0
Tot Deaths/1M pop          0
Death percentage           0
dtype: int64
```

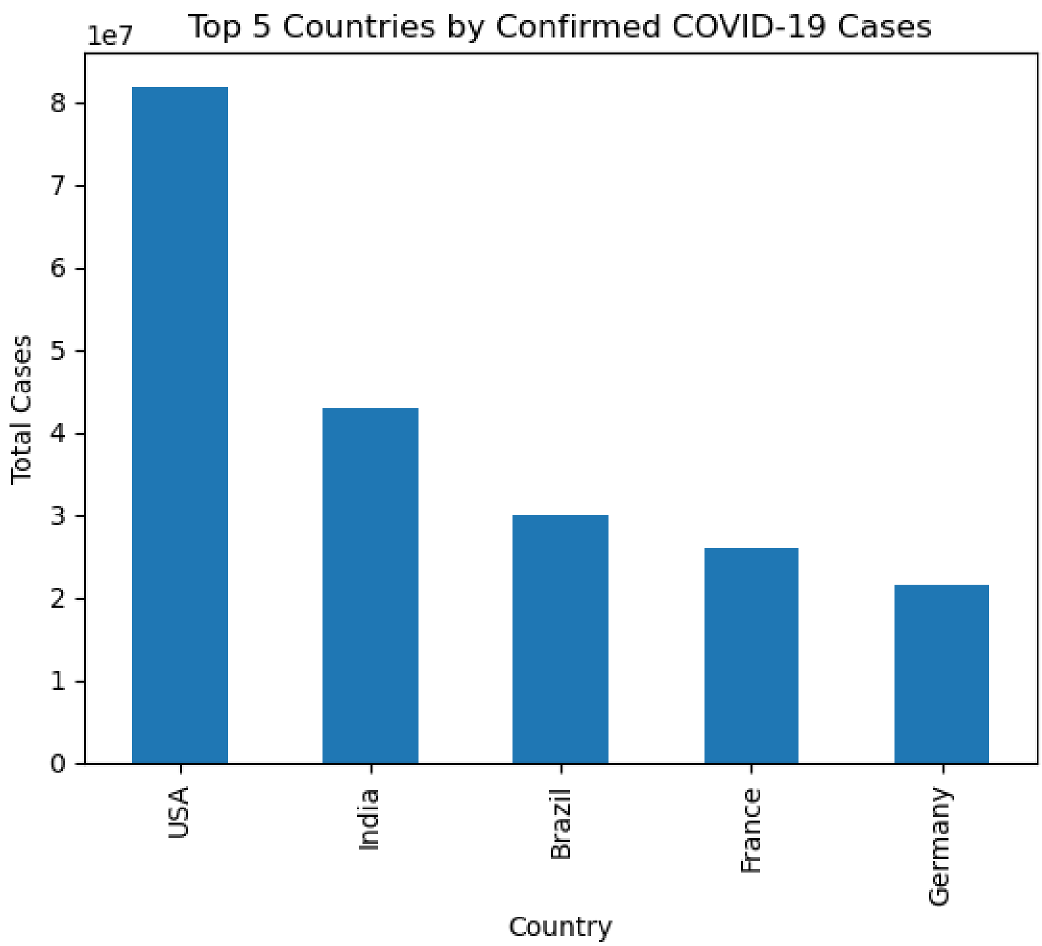
12 ——The dataset has 225 rows and 10 columns including total cases, deaths, population, and death percentages. Information about data types and missing values is gathered to plan cleaning steps——

13 #Step 5 —- now here is country-wise analysis and visualization

14 #Top 5 countries by confirmed cases

[13]:
```python
top5_cases = df.nlargest(5, 'Total Cases')
print(top5_cases[['Country', 'Total Cases']])
top5_cases.plot(x='Country', y='Total Cases', kind='bar', legend=False)
plt.title('Top 5 Countries by Confirmed COVID-19 Cases')
plt.ylabel('Total Cases')
plt.show()
```

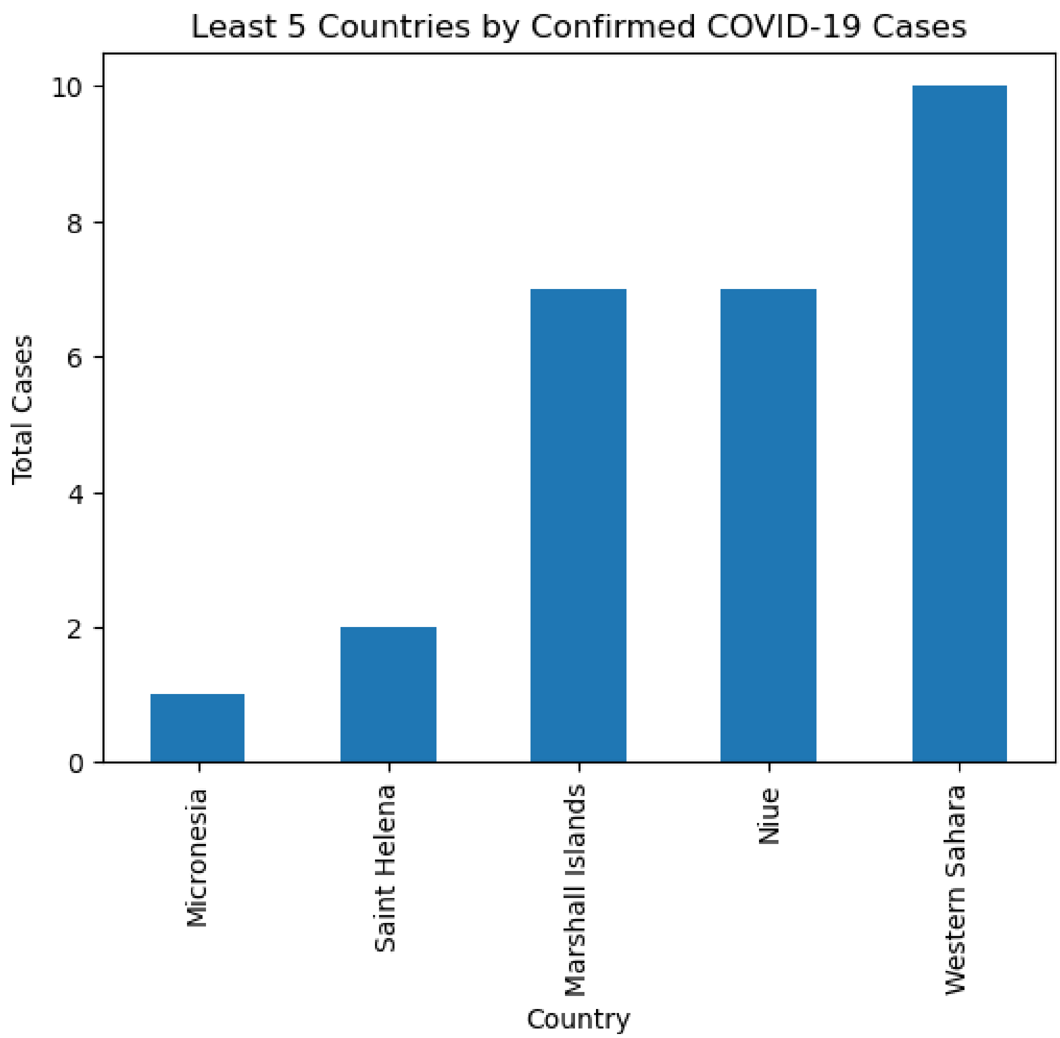
```
     Country  Total Cases
214      USA     81839052
92     India     43029044
26    Brazil     29999816
70    France     25997852
76   Germany     21646375
```



15 #Least 5 countries by confirmed cases

[14]:
```
least5_cases = df.nsmallest(5, 'Total Cases')
print(least5_cases[['Country', 'Total Cases']])
least5_cases.plot(x='Country', y='Total Cases', kind='bar', legend=False)
plt.title('Least 5 Countries by Confirmed COVID-19 Cases')
plt.ylabel('Total Cases')
plt.show()
```

```
              Country  Total Cases
131        Micronesia            1
168      Saint Helena            2
125  Marshall Islands            7
148              Niue            7
221    Western Sahara           10
```

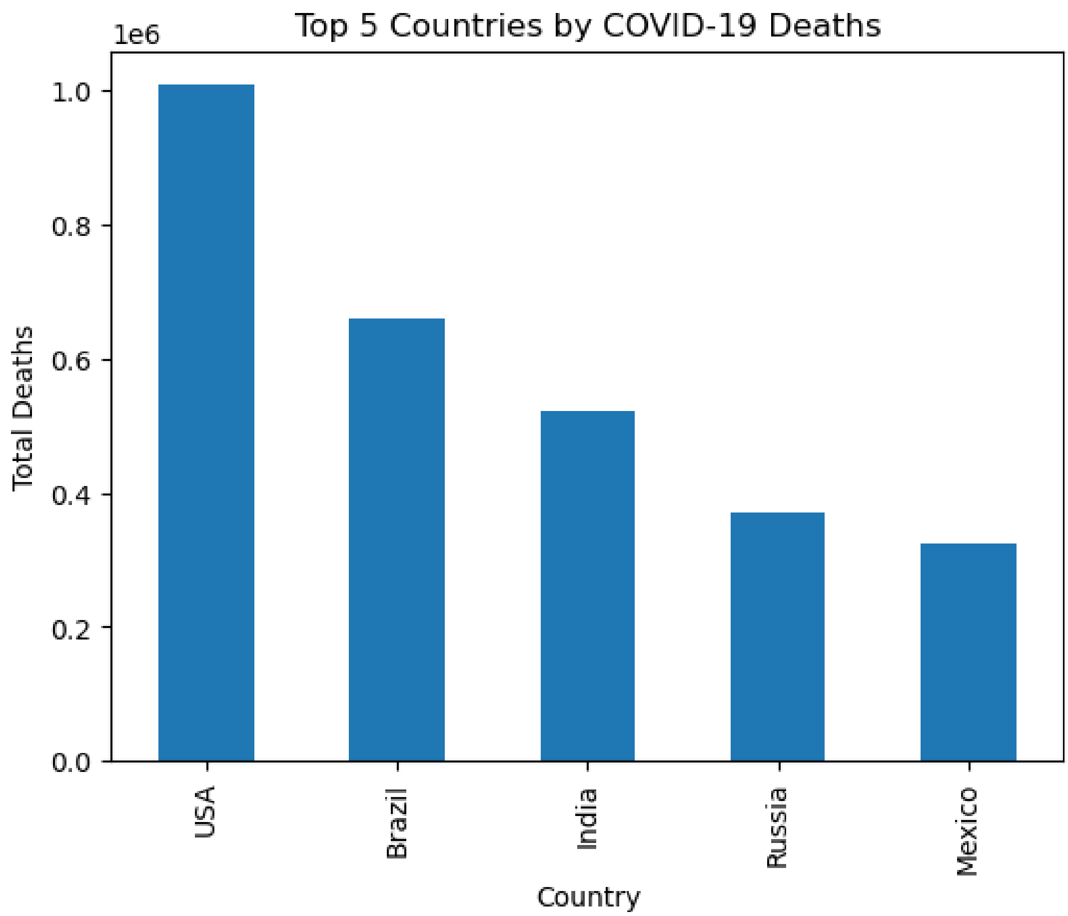


16 #Top 5 countries by deaths ('Total Deaths')

[69]:
```
top5_deaths = df.nlargest(5, 'Total Deaths')
print(top5_deaths[['Country', 'Total Deaths']])
top5_deaths.plot(x='Country', y='Total Deaths', kind='bar', legend=False)
plt.title('Top 5 Countries by COVID-19 Deaths')
plt.ylabel('Total Deaths')
plt.show()
```

```
    Country  Total Deaths
214     USA       1008222
26   Brazil        660269
92    India        521388
165  Russia        369708
```

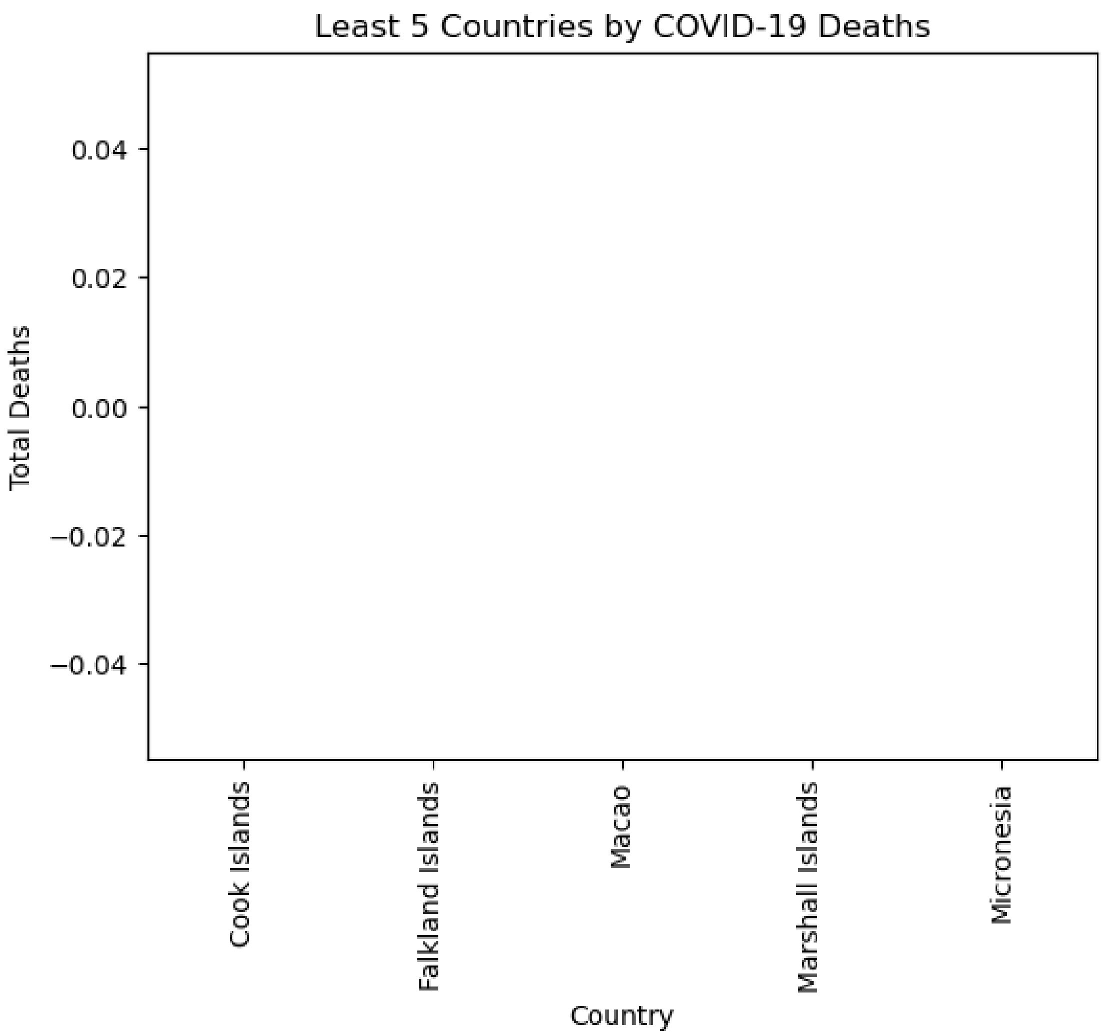
```
130   Mexico        323212
```



17 #Least 5 countries by deaths analysis

[70]:
```python
least5_deaths = df.nsmallest(5, 'Total Deaths')
print(least5_deaths[['Country', 'Total Deaths']])
least5_deaths.plot(x='Country', y='Total Deaths', kind='bar', legend=False)
plt.title('Least 5 Countries by COVID-19 Deaths')
plt.ylabel('Total Deaths')
plt.show()
```

```
              Country  Total Deaths
46       Cook Islands             0
67   Falkland Islands             0
118             Macao             0
125  Marshall Islands             0
131        Micronesia             0
```



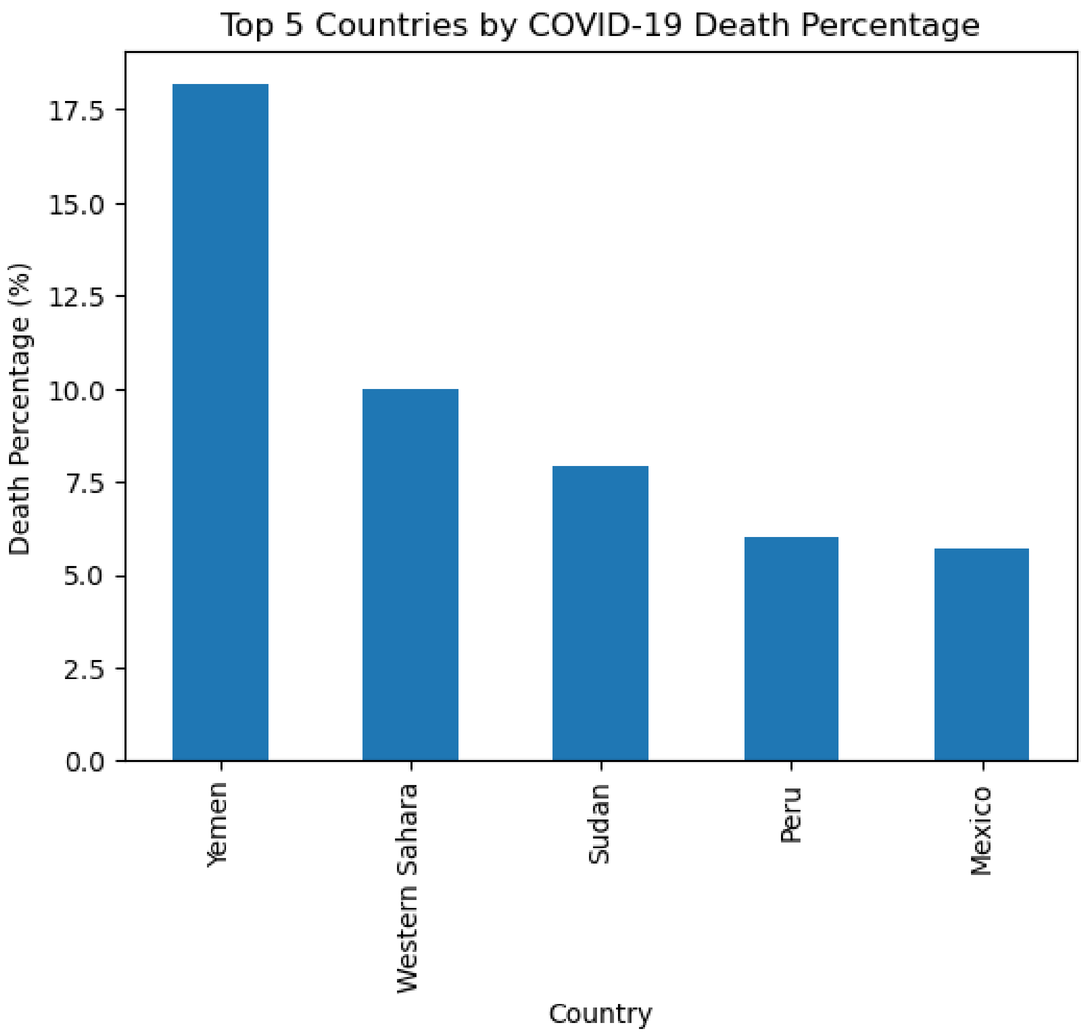
18 ——The analysis focuses on countries with the highest and lowest COVID-19 cases and deaths. Bar charts are created which visually compare these countries side by side. This helps identify the most and least affected countries in an easy-to-understand way, making the impact of the virus clearer——

19 #Step 6 = Death Percentage Analysis

[71]:
```
df['Death percentage'] = pd.to_numeric(df['Death percentage'], errors='coerce')␣
↪ # Convert possible string to numeric
top5_death_pct = df.nlargest(5, 'Death percentage')
print(top5_death_pct[['Country', 'Death percentage']])
```

```python
top5_death_pct.plot(x='Country', y='Death percentage', kind='bar', legend=False)
plt.title('Top 5 Countries by COVID-19 Death Percentage')
plt.ylabel('Death Percentage (%)')
plt.show()
```

```
            Country  Death percentage
222           Yemen         18.151787
221  Western Sahara         10.000000
193           Sudan          7.920265
158            Peru          5.983499
130          Mexico          5.705041
```

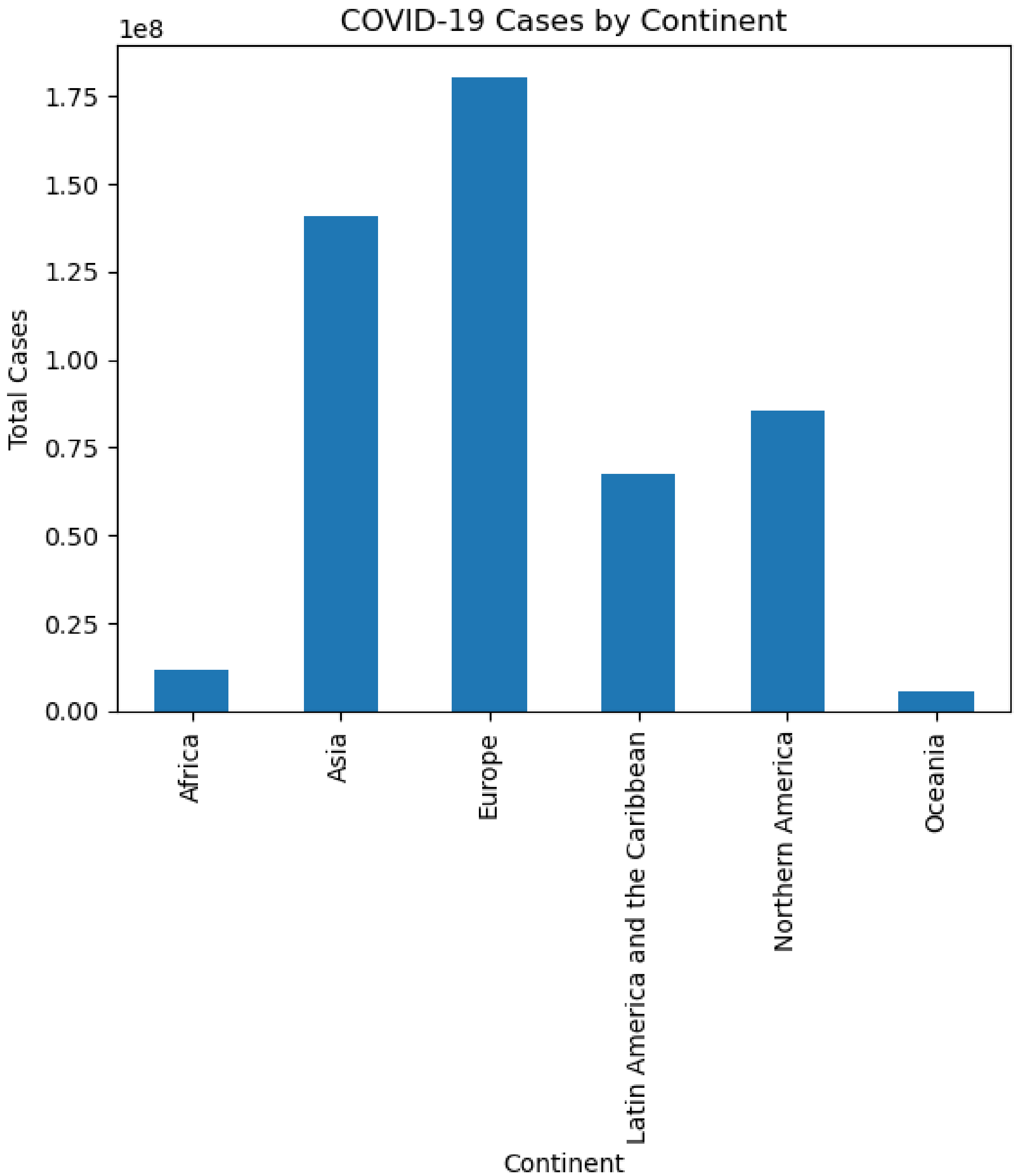


20 ——Death percentage is calculated by dividing the number of deaths by total cases for each country and then displayed in a bar chart. This percentage is a better indicator of severity compared to just total case counts, showing how deadly the virus was relative to the number of infections——

21 #Step 7 = Continent-wise Analysis

[72]:
```
continent_cases = df.groupby('Continent')['Total Cases'].sum()
print(continent_cases)
continent_cases.plot(kind='bar')
plt.title('COVID-19 Cases by Continent')
plt.ylabel('Total Cases')
plt.show()
```

```
Continent
Africa                              11764207
Asia                               140957179
Europe                             180332483
Latin America and the Caribbean     67509231
Northern America                    85364770
Oceania                              5647957
Name: Total Cases, dtype: int64
```

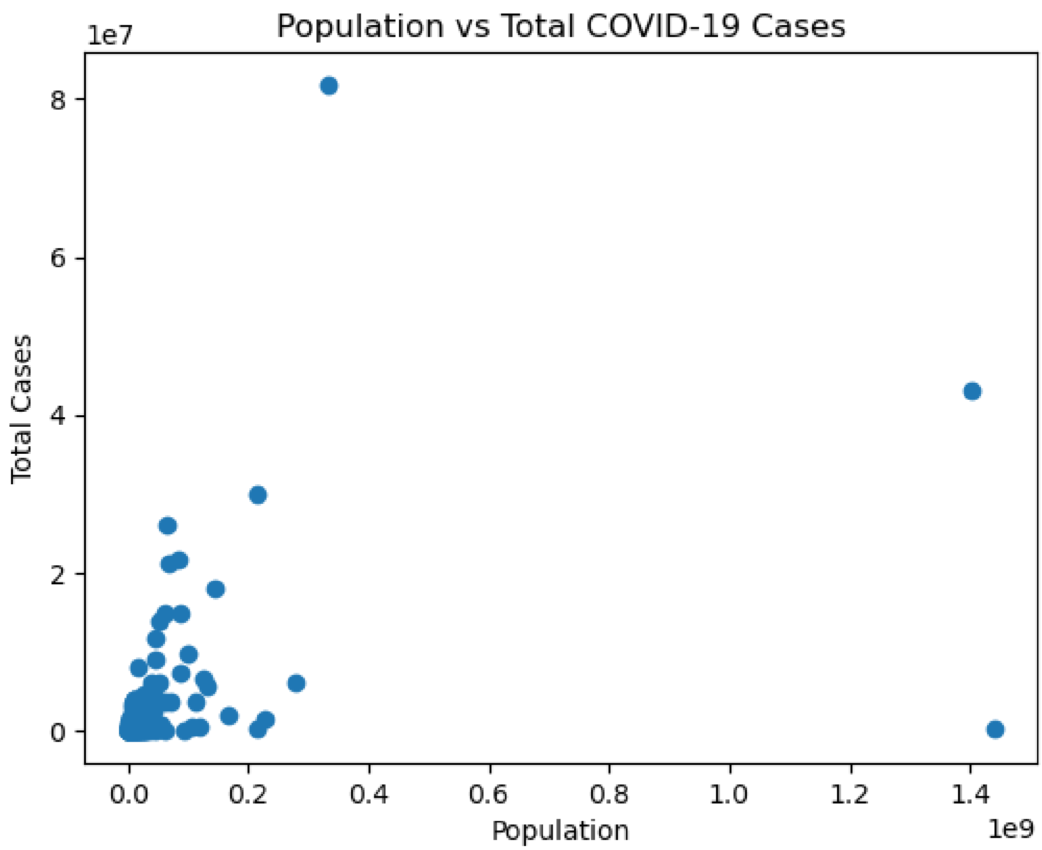

COVID-19 Cases by Continent
1e8
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Total Cases
Africa
Asia
Europe
Latin America and the Caribbean
Northern America
Oceania
Continent

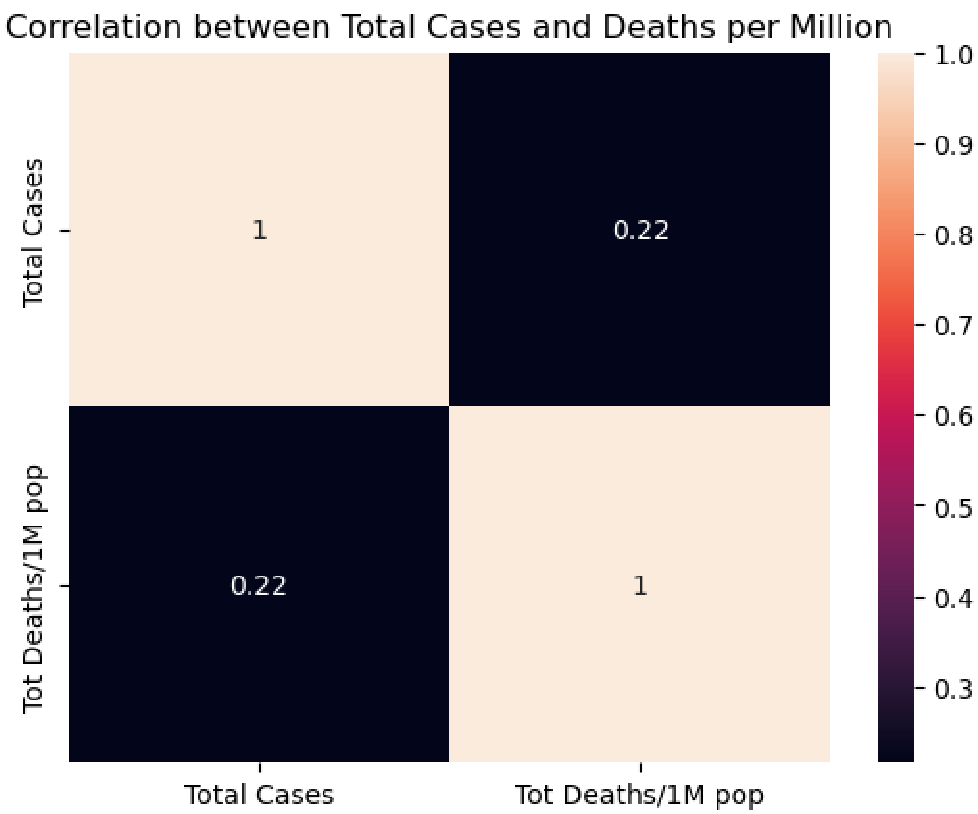
22 ——All the data is grouped by continent to sum the total number of COVID-19 cases in each region. The visualization of this data by continent gives a bigger picture view showing which parts of the world were hit hardest, making regional comparisons simpler and more understandable——

23 #Step 8 = Correlation and Population Analysis

[73]:
```
plt.scatter(df['Population'], df['Total Cases'])
plt.title('Population vs Total COVID-19 Cases')
plt.xlabel('Population')
plt.ylabel('Total Cases')
plt.show()

correlation = df[['Total Cases', 'Tot Deaths/1M pop']].corr()
sns.heatmap(correlation, annot=True)
plt.title('Correlation between Total Cases and Deaths per Million')
plt.show()
```





24 ——This step looks for patterns and relationships between population size, COVID-19 cases, and deaths per million. Scatter plots and correlation heatmaps are used to show how these variables interact, such as whether countries with bigger populations always have more cases or if other factors affect the outcomes shown in the data——

25 1-Analysis and Observations

This analysis of a 225-country dataset reveals patterns of COVID-19's global impact. The dataset was almost complete, with minimal missing data addressed to ensure accuracy. The United States, India, and Brazil emerge consistently as countries with the highest infection and death counts, reflecting well-documented global trends. Conversely, smaller or more isolated nations experienced fewer cases and deaths. Death percentage analysis highlights some countries facing more severe outcomes relative to their case numbers, indicating disparities in healthcare systems and pandemic responses. The continent-wise breakdown shows the most significant case burdens occur in Asia and Europe, while Oceania experienced a comparatively mild impact. Visualizations support these findings, illustrating clear contrasts between worst and least affected countries. Population size generally correlates with case numbers, though exceptions exist, hinting at complex epidemiological and socio-economic dynamics. The strong correlation between cases and deaths per million validates the consistency of reported data and pandemic severity.

26 2-Key Observations:

1-The USA, India, and Brazil are the countries most affected by COVID-19 in terms of total cases and deaths, highlighting their significant burden during the pandemic.

2-Smaller countries such as Micronesia and Saint Helena report very low case counts and deaths, reflecting either geographic isolation or population size.

3-Death percentage varies considerably, with some countries experiencing higher fatality rates relative to infections, pointing to differences in healthcare quality and reporting.

4-The largest numbers of cases and deaths are concentrated in Asia and Europe, making these continents the most impacted regions.

5-Population size generally correlates with total cases, but there are outliers showing that factors beyond population also influence COVID-19 spread.

6-A strong positive correlation exists between total cases and deaths per million population, confirming that as infections increase, death rates tend to rise correspondingly.

7Missing data was minimal, and effective handling ensured that the dataset was clean, which helps provide more trustworthy analysis results.

27 3-Insights from Visualization

1-The disproportionate impact on large, densely populated countries underscores the influence of population size and mobility.

2-Death percentage graphs help isolate countries with critical healthcare challenges.

3-Regional analysis by continent reveals how geographical and political factors contribute to variability in pandemic effects.

4-Population vs. case scatter plots reveal deviations suggesting local containment success or data reporting inconsistencies.

5-The correlation heatmap confirms the intuitive link between infection scale and mortality rate on a population basis.

28 4-Conclusion

In this project, I analyzed COVID-19 data from many countries and continents to understand The global COVID-19 pandemic has had a profound and varied impact across countries and continents, as revealed by this comprehensive analysis. The data shows a strong relationship between population size and total cases, with countries like the USA, India, and Brazil facing the heaviest burdens. However, different death percentages among nations highlight disparities in healthcare systems and pandemic responses. Regional differences are evident, with Asia and Europe experiencing the highest case counts, while Oceania reported fewer cases overall. The strong positive correlation between total cases and deaths per million further confirms that higher infection rates are linked to greater mortality on a per capita basis. Such insights emphasize the importance of effective public health strategies, resource allocation, and preparedness plans tailored to regional needs. Understanding these patterns and the factors influencing severity can support governments and health organizations in mitigating future outbreaks and protecting vulnerable populations worldwide.