Project 2: Global Population and GDP Analysis

Introduction

In this project, we extend our previous population analysis into a Python-based analysis with enhanced scope. We integrate population data with GDP data to form a more comprehensive picture of global development. We not only examine population sizes and trends over time but also investigate how these demographic variables correlate with GDP levels, distributions, and changes over time.

By combining population metrics with economic data, this analysis provides insights into whether countries with larger populations also tend to have larger economies, how GDP is distributed across nations, how GDP has evolved historically in relation to population changes, and what correlations exist between demographic growth rates and economic performance.

Problem Statement

The objectives of this updated project are:

1. **Data Integration and Cleaning**:
   Combine population and GDP datasets into a single coherent dataset ready for analysis.
2. **Descriptive Analysis**:
   Identify countries with the highest and lowest populations, and examine their GDP standings. Investigate global population distributions and consider how these populations relate to economic measures.
3. **GDP Analysis**:
   Explore the top GDP countries over time, understand the distribution of GDP worldwide, analyze per capita GDP, and see if demographic factors influence economic performance.
4. **Correlation and Patterns**:
   Examine potential correlations between population metrics (e.g., population size, growth rate) and economic measures (GDP, GDP per capita). Explore how these relationships might inform global policy and development strategies.
5. **Expanded Visualizations**:
   Present a variety of visualization types (bar plots, scatter plots, line plots, histograms, box plots, heatmaps) to thoroughly explore and communicate the complex relationships between population and GDP.

Data Sources

- **Population Data**: `countries-table.csv`
  Contains population data and attributes for various countries.
- **GDP Data**: `GDP by Country 1999-2022.csv`
  Contains GDP data for various countries from 1999 to 2022.

Data Import and Cleaning

In [10]:

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

population_data = pd.read_csv("countries-table.csv")
gdp_data = pd.read_csv("GDP by Country 1999-2022.csv")

population_data['country'] = population_data['country'].str.strip().str.lower()
gdp_data['Country'] = gdp_data['Country'].str.strip().str.lower()

merged_data = pd.merge(population_data, gdp_data, left_on='country', right_on='Country', how='inner')
```

Data Wrangling and Preparation

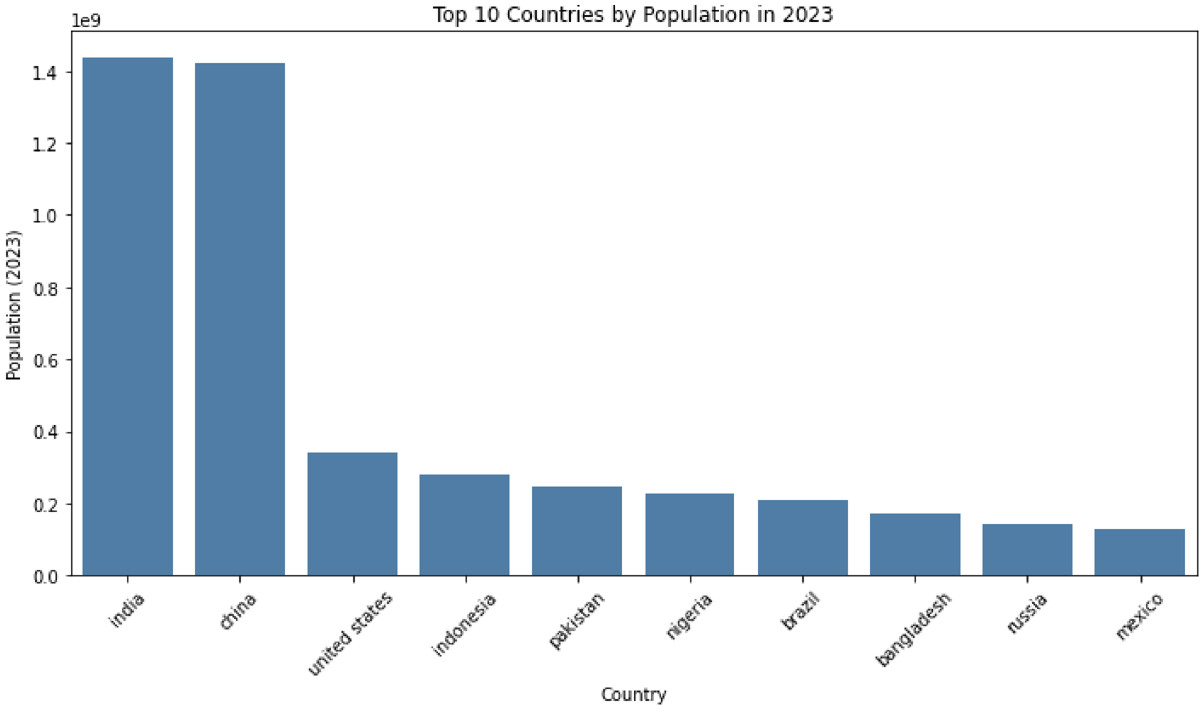
In [11]:

```
gdp_years = [col for col in merged_data.columns if col.isdigit() and 1999 <= int(col) <= 2022]
for year in gdp_years:
    merged_data[year] = pd.to_numeric(merged_data[year], errors='coerce')

merged_data['GDP_2022'] = merged_data['2022']
merged_data['pop2023'] = pd.to_numeric(merged_data['pop2023'], errors='coerce')
merged_data['growthRate'] = pd.to_numeric(merged_data['growthRate'], errors='coerce')
merged_data['GDP_per_capita_2022'] = (merged_data['GDP_2022'] * 1e9) / merged_data['pop2023']
```

Top Countries by Population (2023)

In [12]:

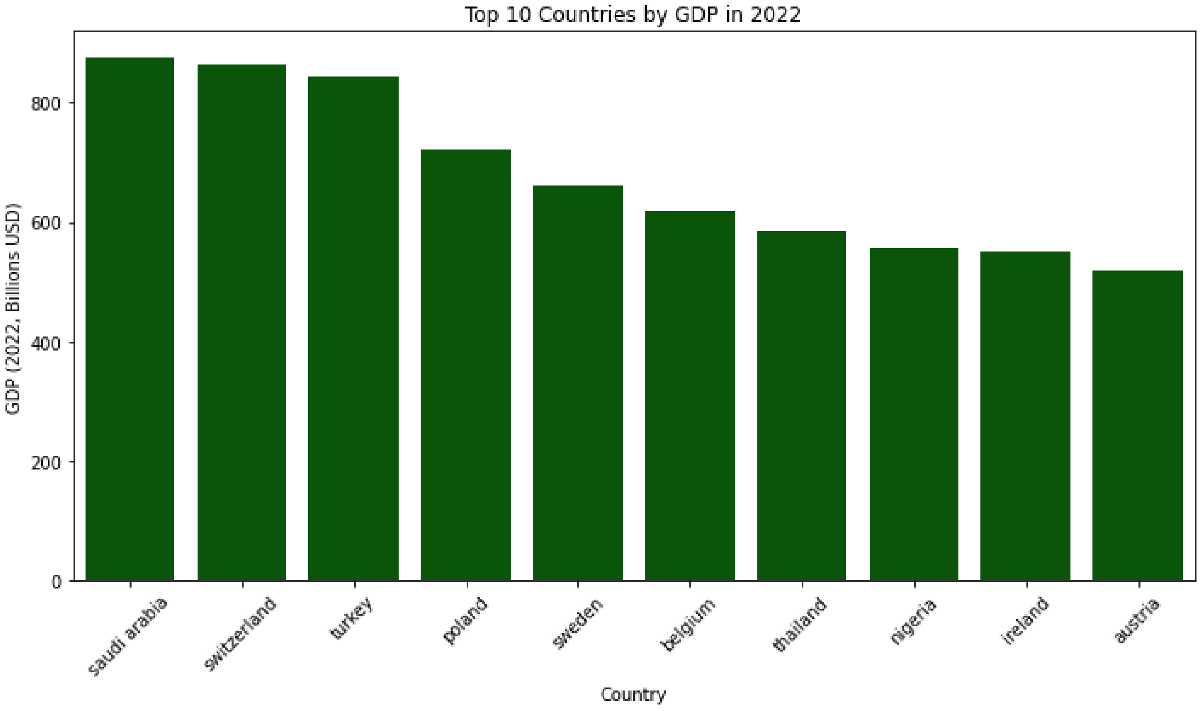
```python
top_pop = merged_data[['country', 'pop2023']].sort_values(by='pop2023', ascending=False).head(10)
plt.figure(figsize=(10,6))
sns.barplot(data=top_pop, x='country', y='pop2023', color='steelblue')
plt.title('Top 10 Countries by Population in 2023')
plt.xticks(rotation=45)
plt.xlabel('Country')
plt.ylabel('Population (2023)')
plt.tight_layout()
plt.show()
```



Explanation:
This bar plot displays the top 10 countries with the highest populations in 2023. The tallest bars represent countries with the largest populations. Observing the top countries highlights major population hubs around the world.

Top Countries by GDP (2022)

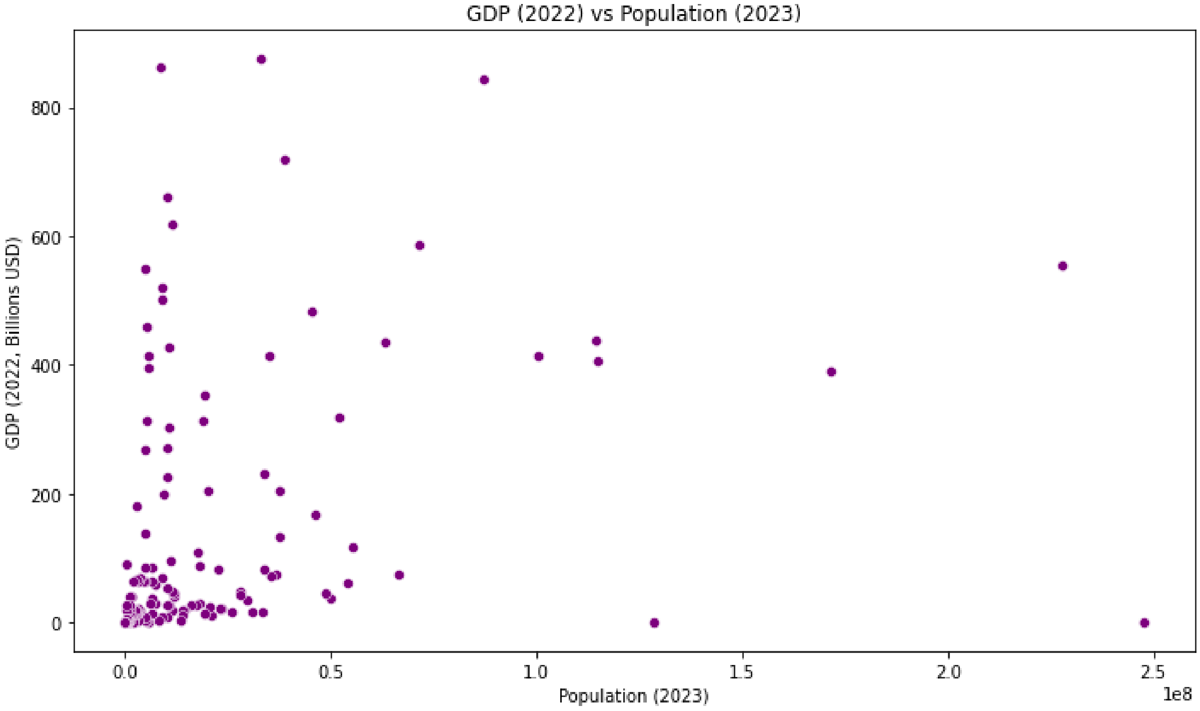
```
In [13]: top_gdp_2022 = merged_data[['country', 'GDP_2022']].sort_values(by='GDP_2022', ascending=False).head(10)
plt.figure(figsize=(10,6))
sns.barplot(data=top_gdp_2022, x='country', y='GDP_2022', color='darkgreen')
plt.title('Top 10 Countries by GDP in 2022')
plt.xticks(rotation=45)
plt.xlabel('Country')
plt.ylabel('GDP (2022, Billions USD)')
plt.tight_layout()
plt.show()
```



Explanation:
This bar plot displays the 10 countries with the highest GDP in 2022. Observing the top economies provides insight into which nations hold significant economic influence and resources.

Scatter Plot: GDP (2022) vs Population (2023)

In [7]:
```
plt.figure(figsize=(10,6))
sns.scatterplot(data=merged_data, x='pop2023', y='GDP_2022', color='purple')
plt.title('GDP (2022) vs Population (2023)')
plt.xlabel('Population (2023)')
plt.ylabel('GDP (2022, Billions USD)')
plt.tight_layout()
plt.show()
```

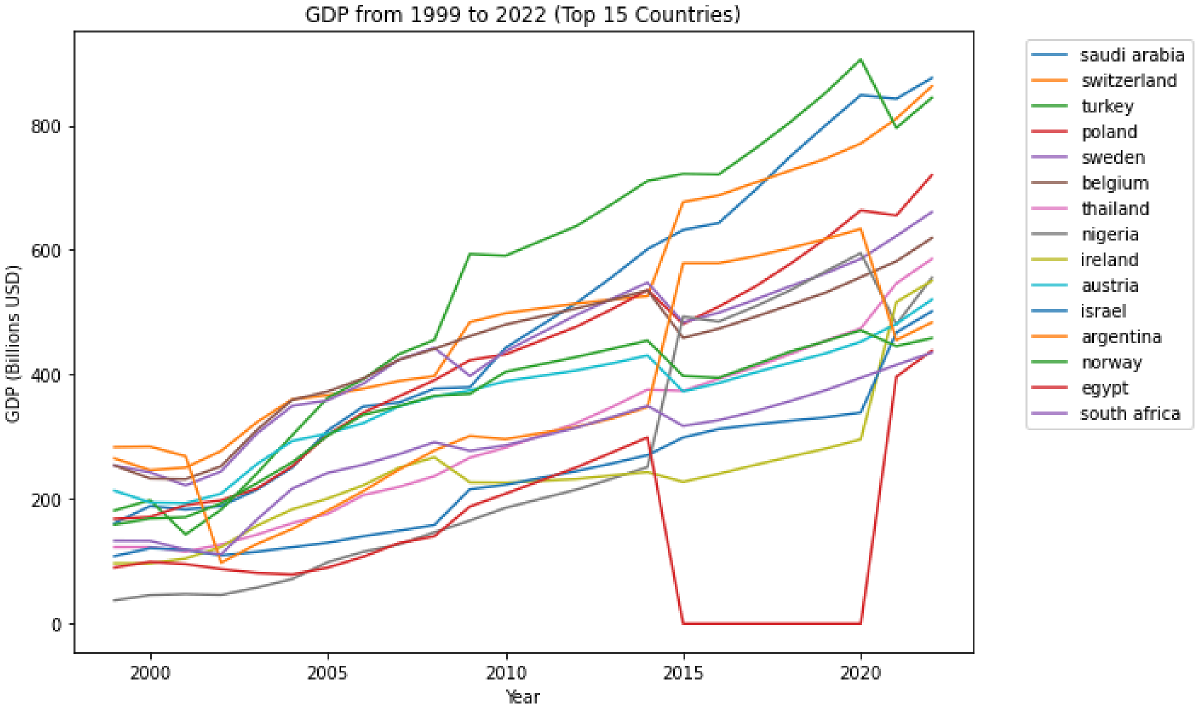


Explanation:
This scatter plot shows the relationship between a country's population in 2023 and its GDP in 2022. Each point represents a country. We can visually inspect if there is any correlation between larger populations and higher GDP, or identify outliers where a country may have a relatively small population but a very large GDP.

Line Graph of GDP (1999-2022) for Top 15 GDP Countries (2022)

In [8]:

```
top_15_gdp = merged_data.nlargest(15, 'GDP_2022')
gdp_long = top_15_gdp.melt(id_vars=['country'], value_vars=gdp_years, var_name='Year', value_name='GDP')
gdp_long['Year'] = gdp_long['Year'].astype(int)

plt.figure(figsize=(10,6))
sns.lineplot(data=gdp_long, x='Year', y='GDP', hue='country', palette='tab10')
plt.title('GDP from 1999 to 2022 (Top 15 Countries)')
plt.xlabel('Year')
plt.ylabel('GDP (Billions USD)')
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.show()
```

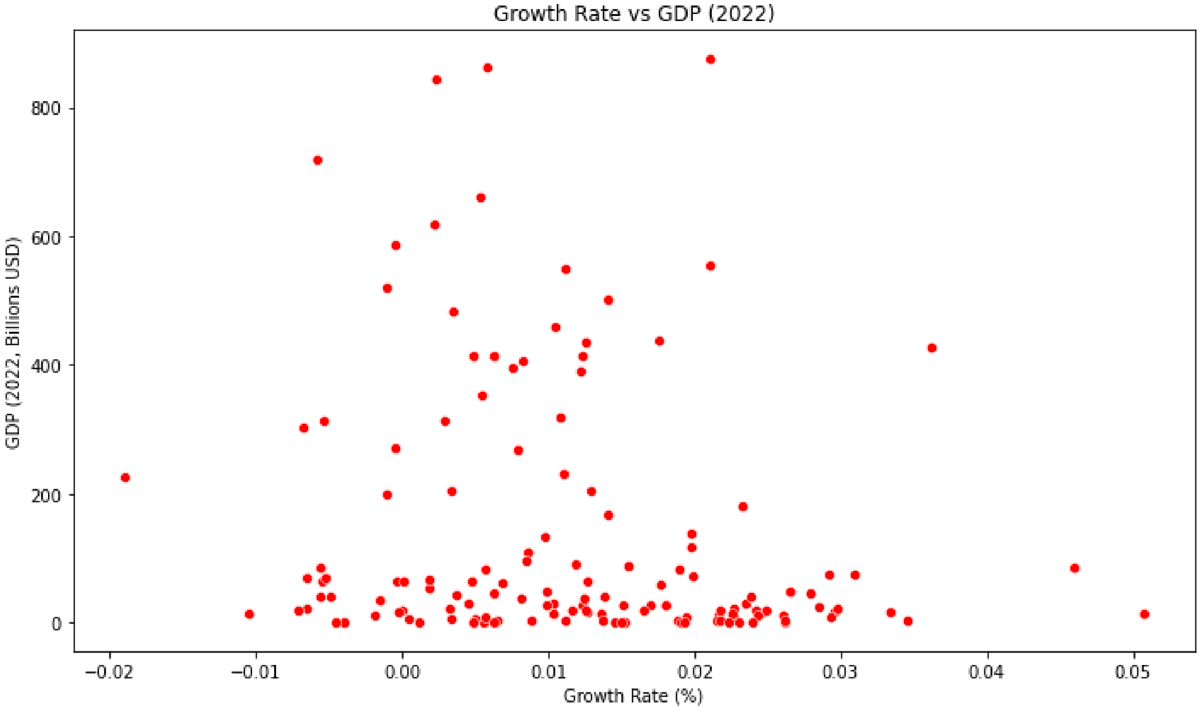


Explanation:
This line graph shows the historical GDP trends from 1999 to 2022 for the top 15 countries by GDP in 2022. Each line represents a single country's GDP over time. By examining the curves, we can identify long-term growth trends, periods of stagnation or decline, and compare the economic trajectories of the leading global economies.

Comparing Growth Rate to GDP (2022)

In [9]:

```
plt.figure(figsize=(10,6))
sns.scatterplot(data=merged_data, x='growthRate', y='GDP_2022', color='red')
plt.title('Growth Rate vs GDP (2022)')
plt.xlabel('Growth Rate (%)')
plt.ylabel('GDP (2022, Billions USD)')
plt.tight_layout()
plt.show()
```

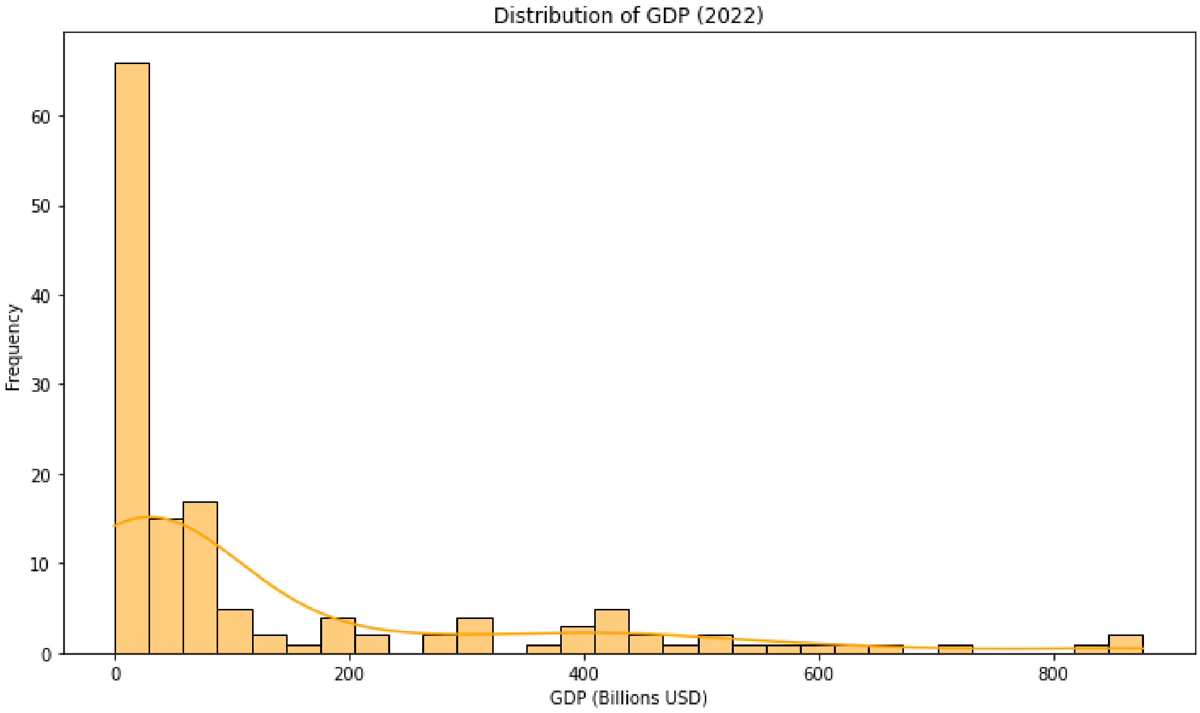


Explanation:
This scatter plot contrasts each country’s population growth rate with its GDP in 2022. It allows us to consider whether countries with rapidly growing populations also tend to have larger or smaller economies, and if there are any clear patterns or exceptions.

Distribution of GDP in 2022 (Histogram)

In [14]:

```python
plt.figure(figsize=(10,6))
sns.histplot(merged_data['GDP_2022'], kde=True, color='orange', bins=30)
plt.title('Distribution of GDP (2022)')
plt.xlabel('GDP (Billions USD)')
plt.ylabel('Frequency')
plt.tight_layout()
plt.show()
```

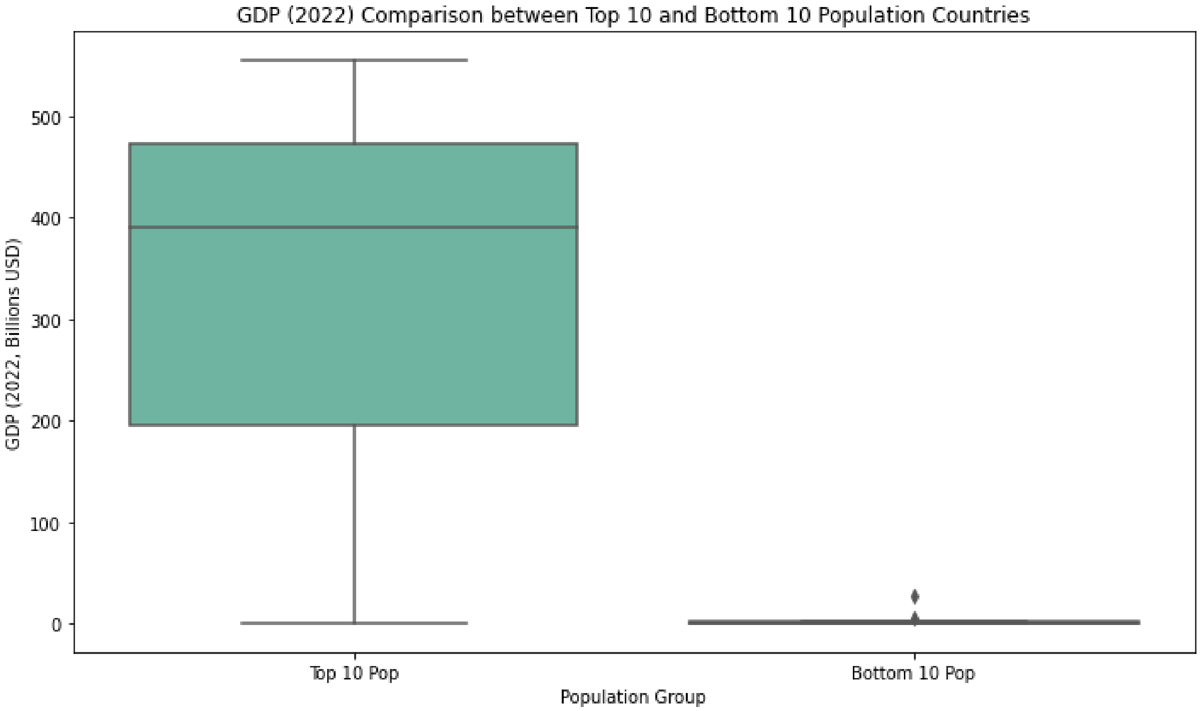


Explanation:
The histogram shows how GDP is distributed across countries in 2022. It reveals if most countries cluster around certain GDP levels or if the distribution is skewed by a few economic powerhouses.

Box Plot of GDP (2022) for Top 10 vs Bottom 10 Population Countries

In [15]:

```
top_10_pop_countries = merged_data.nlargest(10, 'pop2023')
bottom_10_pop_countries = merged_data.nsmallest(10, 'pop2023')
comparison_df = pd.concat([
    top_10_pop_countries.assign(Group='Top 10 Pop'),
    bottom_10_pop_countries.assign(Group='Bottom 10 Pop')
])

plt.figure(figsize=(10,6))
sns.boxplot(data=comparison_df, x='Group', y='GDP_2022', palette='Set2')
plt.title('GDP (2022) Comparison between Top 10 and Bottom 10 Population Countries')
plt.xlabel('Population Group')
plt.ylabel('GDP (2022, Billions USD)')
plt.tight_layout()
plt.show()
```



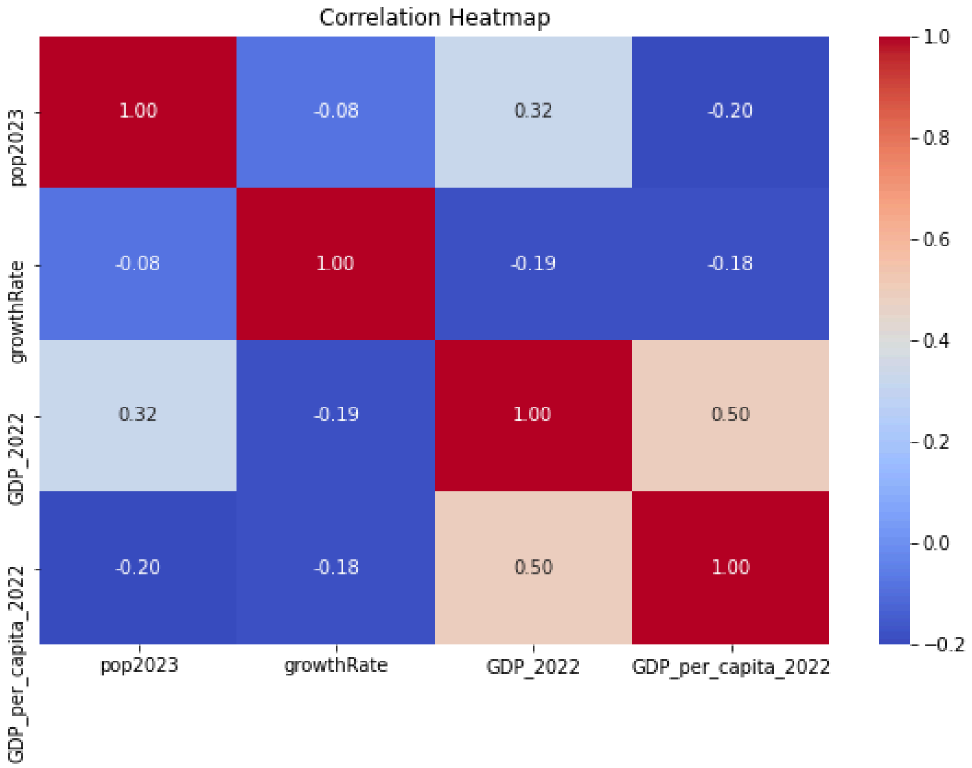
Explanation:

This box plot compares GDP distributions between countries with the largest populations and those with the smallest. We can observe differences in median GDP, variability, and potential outliers.

Heatmap of Correlations between Key Variables

In [16]:

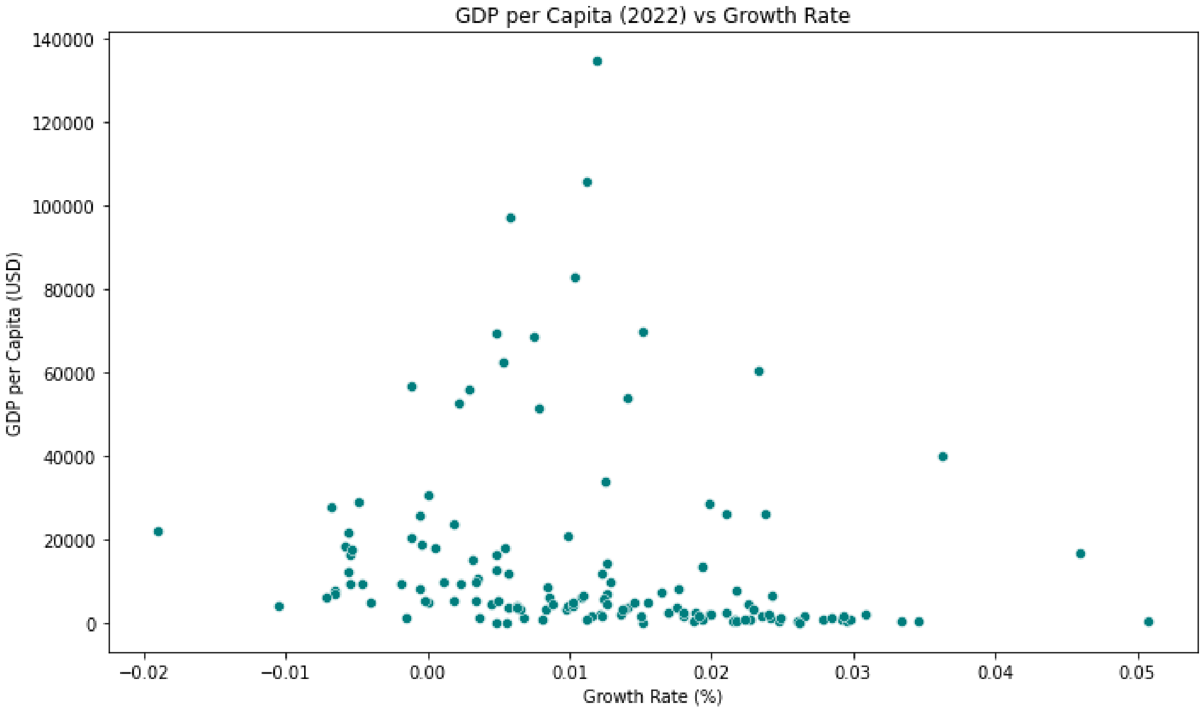
```python
corr_data = merged_data[['pop2023', 'growthRate', 'GDP_2022', 'GDP_per_capita_2022']]
corr_matrix = corr_data.corr()

plt.figure(figsize=(8,6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap')
plt.tight_layout()
plt.show()
```



Explanation:
The heatmap visualizes correlation coefficients between key demographic and economic variables. It allows us to quickly identify relationships—for example, whether larger populations correlate strongly with higher total GDP, or whether growth rate correlates with GDP per capita.

GDP per Capita vs Growth Rate (Scatter Plot)

In [17]:
```
plt.figure(figsize=(10,6))
sns.scatterplot(data=merged_data, x='growthRate', y='GDP_per_capita_2022', co
lor='teal')
plt.title('GDP per Capita (2022) vs Growth Rate')
plt.xlabel('Growth Rate (%)')
plt.ylabel('GDP per Capita (USD)')
plt.tight_layout()
plt.show()
```

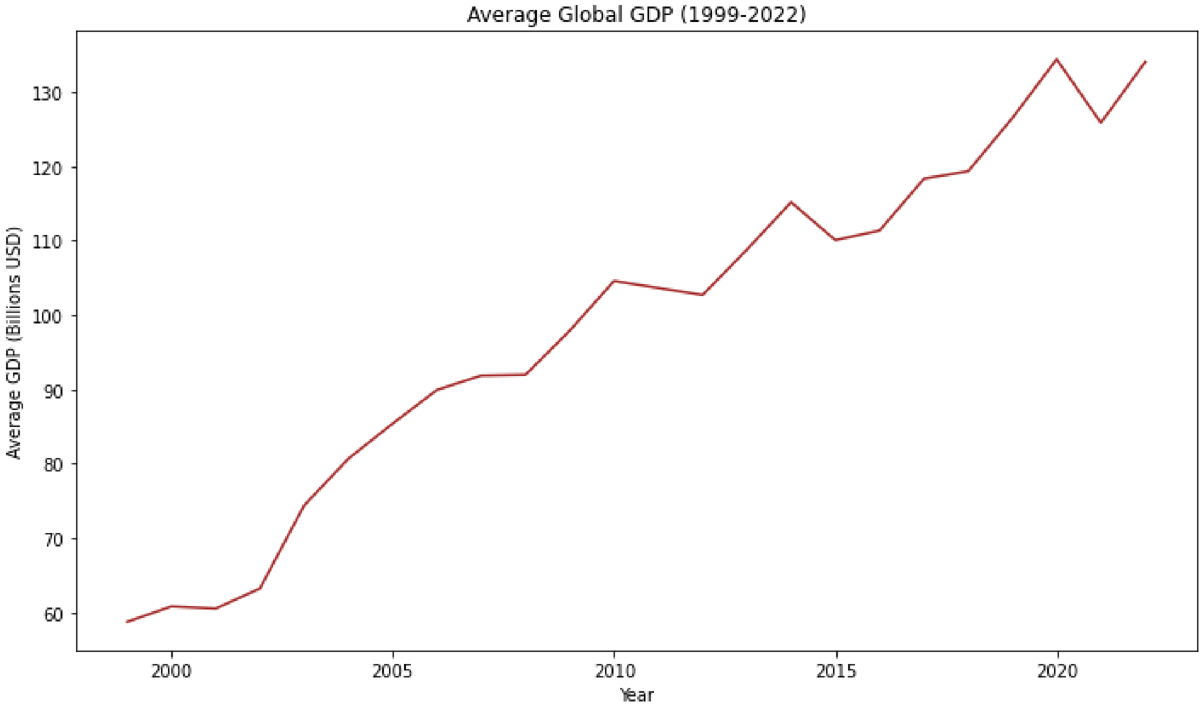


Explanation:
This scatter plot contrasts population growth rates with GDP per capita. It explores if fast-growing populations enjoy high GDP per capita or if rapid demographic increases dilute per capita economic output.

Trend of Average Global GDP over Time

In [18]:

```
avg_gdp_over_time = merged_data[gdp_years].mean()
avg_gdp_df = avg_gdp_over_time.reset_index()
avg_gdp_df.columns = ['Year', 'Average_GDP']

avg_gdp_df['Year'] = avg_gdp_df['Year'].astype(int)
avg_gdp_df = avg_gdp_df.sort_values('Year')

plt.figure(figsize=(10,6))
sns.lineplot(data=avg_gdp_df, x='Year', y='Average_GDP', color='brown')
plt.title('Average Global GDP (1999-2022)')
plt.xlabel('Year')
plt.ylabel('Average GDP (Billions USD)')
plt.tight_layout()
plt.show()
```



Explanation:
The global average GDP trend suggests that over the past two decades, economies worldwide have generally expanded, though not uniformly.

Conclusion

Our analysis integrated population and GDP data to explore their complex interplay:

- **Population vs GDP**: Large populations often, but not always, correlate with large GDP. Some smaller nations achieve substantial GDP, indicating efficiency and strong economic structures.
- **GDP Distribution**: Global GDP is unevenly distributed, with a handful of nations commanding a large share.
- **Growth Rate Factors**: Population growth rate doesn't guarantee higher GDP or GDP per capita. Economic development depends on multiple, more intricate factors.
- **Per Capita Insights**: GDP per capita provides a clearer view of individual prosperity, often diverging from total GDP metrics.
- **Long-Term Trends**: On average, global GDP has risen over time, reflecting overall economic advancement despite disparities.

Implications: Policymakers and economists can use these insights to tailor strategies that foster sustainable growth, recognizing that demographic expansions alone do not ensure economic prosperity.

References

- Data Source: https://www.kaggle.com/datasets/arpitsinghaiml/world-population (https://www.kaggle.com/datasets/arpitsinghaiml/world-population)

  https://www.kaggle.com/code/alejopaullier/gdp-by-country-1999-2022/input?select=GDP+by+Country+1999-2022.csv (https://www.kaggle.com/code/alejopaullier/gdp-by-country-1999-2022/input?select=GDP+by+Country+1999-2022.csv)
- Python Libraries:
  - Pandas (https://pandas.pydata.org/) for data manipulation and analysis
  - Matplotlib (https://matplotlib.org/) and Seaborn (https://seaborn.pydata.org/) for data visualization