Population Data Analysis Report

Summary of Key Insights:

• India and China have the highest populations.
• Population distribution is heavily skewed toward Asia.
• Some countries like Monaco show up as outliers with high population density.
• Africa has many countries with lower population but high growth trends.

Code: population_scraper.py

```
from selenium import webdriver
from selenium.webdriver.chrome.options import Options
from bs4 import BeautifulSoup
import pandas as pd
import time

# Optional: Prevent browser from opening visibly
options = Options()
options.add_argument("--headless")  # Run in background
driver = webdriver.Chrome(options=options)  # You must have ChromeDriver in PATH

# Open the website
url = "https://www.worldometers.info/world-population/population-by-country/"
driver.get(url)
time.sleep(5)  # Wait for page to load

# Parse page using BeautifulSoup
soup = BeautifulSoup(driver.page_source, "html.parser")
driver.quit()

# Scrape the correct table
table = soup.find("table")  # First table on the page (no need for ID)

# Extract headers
headers = [th.text.strip() for th in table.find_all("th")]

# Extract rows
rows = []
for tr in table.tbody.find_all("tr"):
    cols = [td.text.strip().replace(",", "") for td in tr.find_all("td")]
    rows.append(cols)

# Convert to DataFrame
df = pd.DataFrame(rows, columns=headers)

# Save to CSV
df.to_csv("population_data.csv", index=False)
print("âœ… Data saved to population_data.csv")
```

Code: eda_population.py

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

# Load dataset
df = pd.read_csv("population_data.csv")

# Print column names to verify
print("Available columns:")
print(df.columns)

# Optionally: Strip all column names of extra spaces
df.columns = df.columns.str.strip()

# Check again after strip
print("Stripped column names:")
```

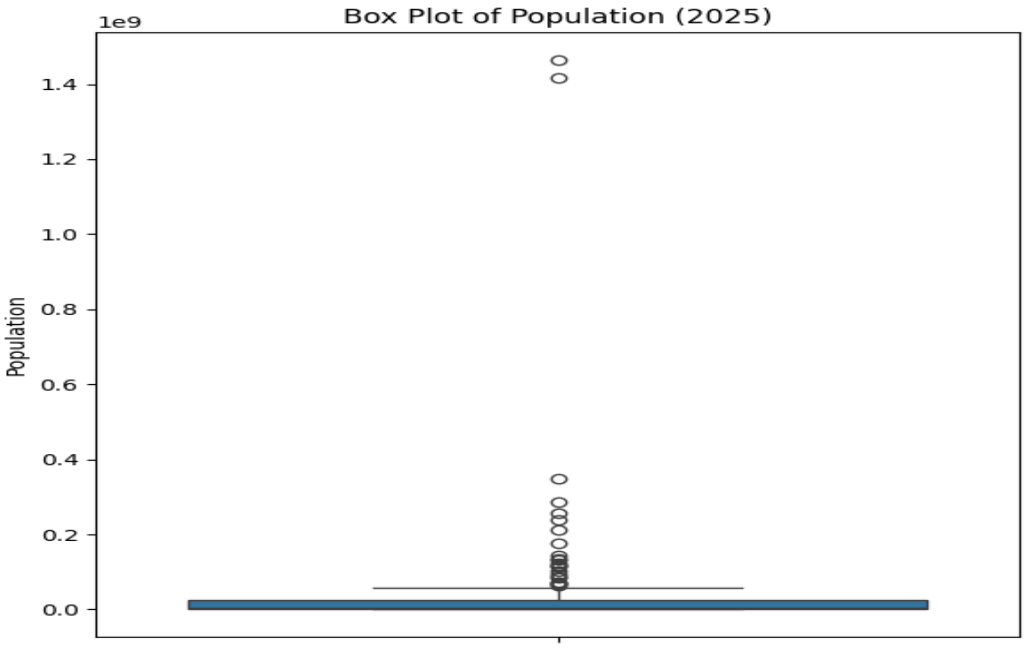
```
print(df.columns)

# Replace these column names with the correct ones from your CSV
# You may need to adjust these if your file has different headers
cols_to_convert = [
    'Population (2020)',
    'Yearly Change',
    'Net Change',
    'Density (P/KmÂ²)',
    'Land Area (KmÂ²)'
]

# Safely convert columns to numeric (with protection)
for col in cols_to_convert:
    if col in df.columns:
        df[col] = pd.to_numeric(df[col].astype(str).str.replace('%', '').str.replace
(',', ''), errors='coerce')
    else:
        print(f"Warning: Column '{col}' not found in data!")

# Display dataset info
print(df.info())
print(df.describe())

# --- Visualization 1: Top 10 most populated countries ---
if 'Population (2020)' in df.columns:
    top10 = df.sort_values('Population (2020)', ascending=False).head(10)
    plt.figure(figsize=(10, 6))
    sns.barplot(x='Population (2020)', y='Country (or dependency)', data=top10)
    plt.title("Top 10 Most Populated Countries (2020)")
    plt.tight_layout()
    plt.savefig("top10_population.png")

    plt.show()
else:
    print("Skipping population plot â€“ 'Population (2020)' not found")

# --- Visualization 2: Population Density distribution ---
if 'Density (P/KmÂ²)' in df.columns:
    plt.figure(figsize=(10, 5))
    sns.histplot(df['Density (P/KmÂ²)'].dropna(), bins=30, kde=True)
    plt.title("Population Density Distribution")
    plt.xlabel("Density (P/KmÂ²)")
    plt.tight_layout()
    plt.savefig("density_distribution.png")
    plt.show()
else:
    print("Skipping density plot â€“ 'Density (P/KmÂ²)' not found")
```

Code: eda_and_visualization.py

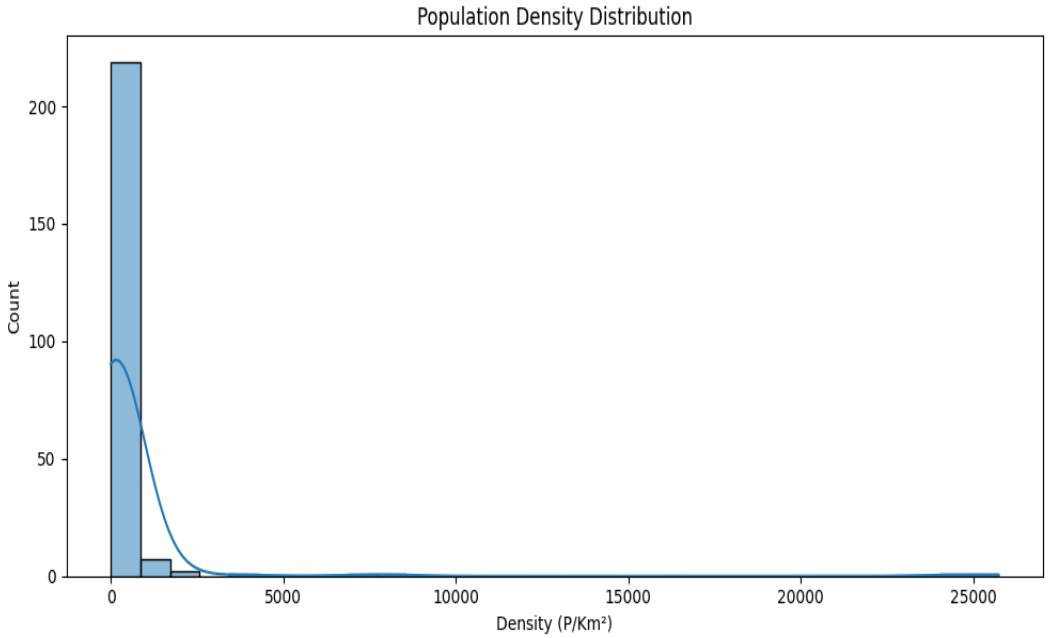
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the CSV
df = pd.read_csv("population_data.csv")
print("Columns in the dataset:")
print(df.columns)

# --------------------------------
# 1. Bar Chart â€“ Top 10 Countries
# --------------------------------
top10 = df.sort_values(by='Population 2025', ascending=False).head(10)

plt.figure(figsize=(12, 6))
sns.barplot(data=top10, x='Country (or dependency)', y='Population 2025')
plt.title('Top 10 Countries by Population (2025)')
plt.xticks(rotation=45)
plt.tight_layout()
plt.savefig("top10_population.png")
plt.show()

# ---------------------------------------
```

```
# 2. Pie Chart â€“ Countries per Continent
# ---------------------------------------
continent_counts = df['Continent'].value_counts()

plt.figure(figsize=(8, 8))
plt.pie(continent_counts, labels=continent_counts.index, autopct='%1.1f%%', startang
le=140)
plt.title('Countries per Continent')
plt.axis('equal')
plt.tight_layout()
plt.savefig('countries_per_continent.png')
plt.show()

# ----------------------------------------
# 3. Histogram â€“ Population Distribution
# ----------------------------------------
plt.figure(figsize=(10, 6))
sns.histplot(df['Population 2025'], bins=30, kde=True)
plt.title('Population Distribution (2025)')
plt.xlabel('Population')
plt.ylabel('Number of Countries')
plt.tight_layout()
plt.savefig('population_distribution.png')
plt.show()

# -------------------------------
# 4. Box Plot â€“ Population by Continent
# -------------------------------
plt.figure(figsize=(12, 6))
sns.boxplot(x='Continent', y='Population 2025', data=df)
plt.title('Population Spread by Continent (2025)')
plt.xticks(rotation=45)
plt.tight_layout()
plt.savefig('population_boxplot.png')
plt.show()
```

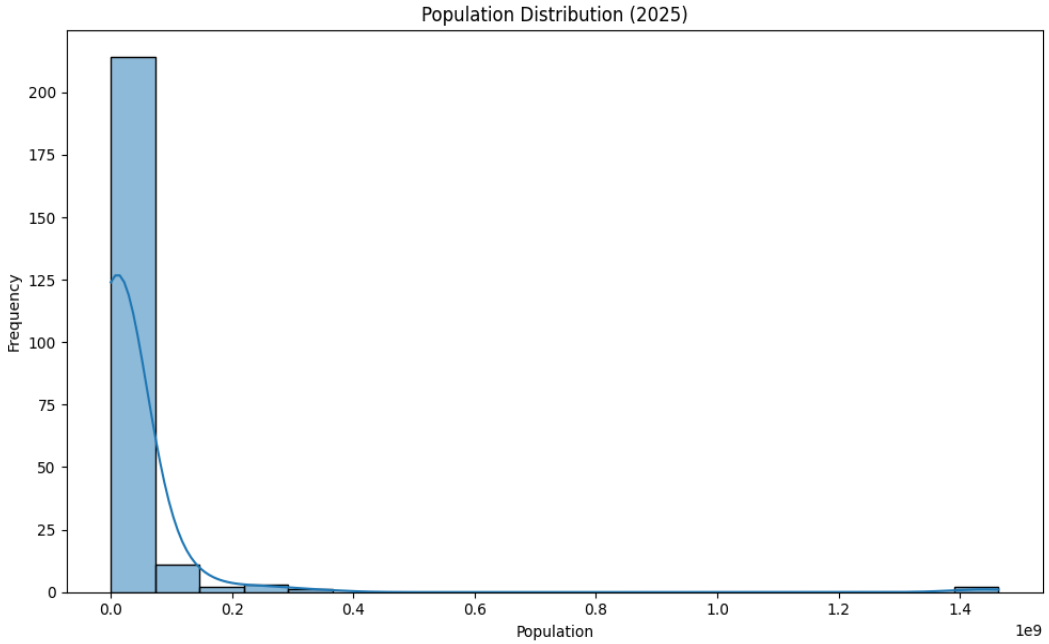
Graph: box_plot_population.png



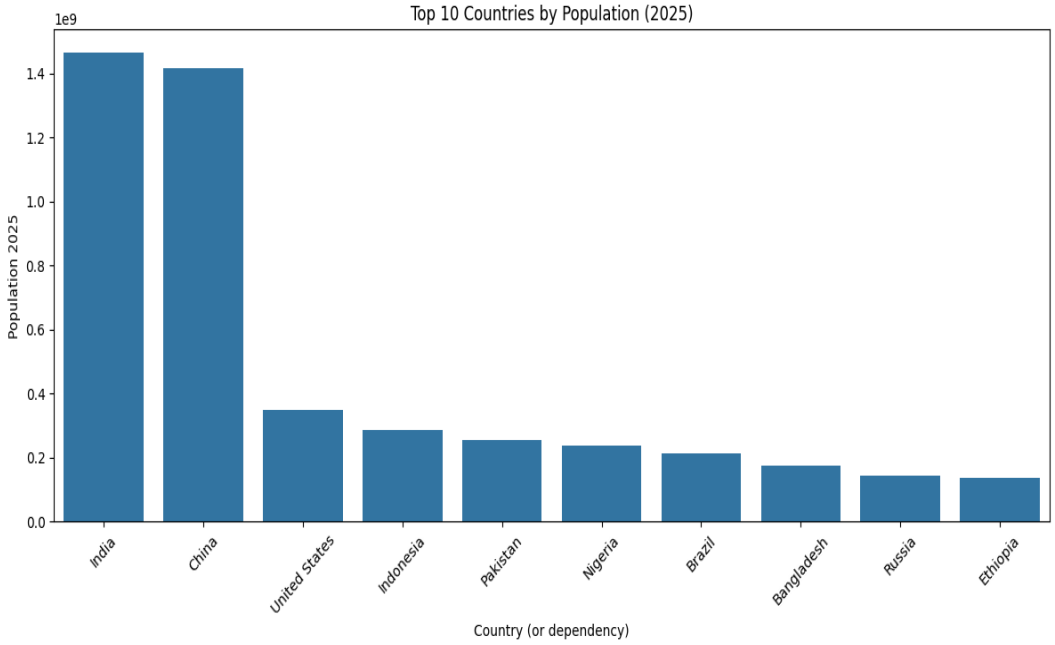
Graph: density_distribution.png



Graph: histogram_population.png



Graph: top10_population.png


Top 10 Countries by Population (2025)
1e9
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Population 2025
India
China
United States
Indonesia
Pakistan
Nigeria
Brazil
Bangladesh
Russia
Ethiopia
Country (or dependency)