Bharatiya Vidya Bhavan's
Sardar Patel Institute of Technology
(Autonomous Institute Affiliated to University of Mumbai)
Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai-400058-India

NAME: Ayush Kedare

ROLL NUMBER: 2021300058

AIM:

Create advanced charts using Tableau / Power BI / R / Python / Plotly or Chart or D3.js to be performed on the dataset - Socio economic data
● Advanced - Word chart, Box and whisker plot, Violin plot, Regression plot (linear and nonlinear), 3D chart, Jitter, Line, Area, Waterfall, Donut, Treemap, Funnel Chart
● Write observations from each chart

DATASET:

Socio-Economic Country Profiles Dataset
The dataset is a socio-economic and demographic profile of various countries. It contains multiple attributes that describe different aspects of each country, including population, economic indicators, military expenditure, life expectancy, and more. Here's a breakdown of the key features in the dataset:
Key Features:

1. Country and Region:
    - country: The name of the country.
    - Region: The geographical region the country belongs to (e.g., Western Europe, South America).
2. Geographical Information:
    - Surface area (km2): The total surface area of the country in square kilometers.
3. Demographic Information:
    - Population in thousands (2017): The total population of the country in thousands as of 2017.
    - Population density (per km2, 2017): The number of people per square kilometer.
    - Sex ratio (m per 100 f, 2017): The ratio of males to females in the population.
    - Life expectancy at birth, male (years): The average life expectancy of males at birth.
    - Life expectancy at birth, total (years): The overall life expectancy at birth for the total population.
4. Economic Indicators:
    - GDP: Gross domestic product (million current US$): The total market value of all goods and services produced by the country in current US dollars.
    - GDP growth rate (annual %, const. 2005 prices): The annual growth rate of GDP adjusted to constant 2005 prices.
    - GDP per capita (current US$): The GDP divided by the population, representing the average economic output per person.
    - Tax revenue (% of GDP): The percentage of GDP collected as tax revenue.
    - Taxes on income, profits and capital gains (% of revenue): The proportion of tax

revenue derived from income, profits, and capital gains.

5. Social Indicators:
    - Urban population (% of total population): The percentage of the population living in urban areas.
    - Military expenditure (% of GDP): The percentage of GDP spent on military expenses.
    - Inflation, consumer prices (annual %): The annual percentage change in consumer prices (inflation rate).
6. Additional Indicators:
    - Population, female: The total female population.
    - Population, male: The total male population.

PROCEDURE DESCRIPTION:

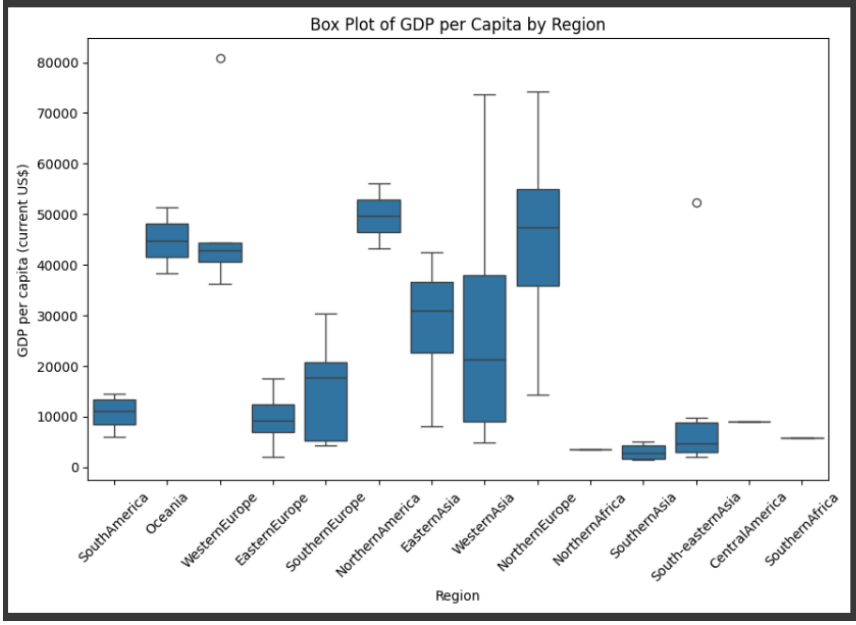
```
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud
from mpl_toolkits.mplot3d import Axes3D
import numpy as np
import plotly.express as px
import pandas as pd
!pip install squarify -qqq
import squarify

df = pd.read_csv('soci_econ_country_profiles.csv')
```

1)Word Cloud

```
plt.figure(figsize=(10, 6))
wordcloud = WordCloud(width=800, height=400, background_color='white').generate(' '.join(df['country']))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title('Word Cloud of Country Names')
plt.show()
```

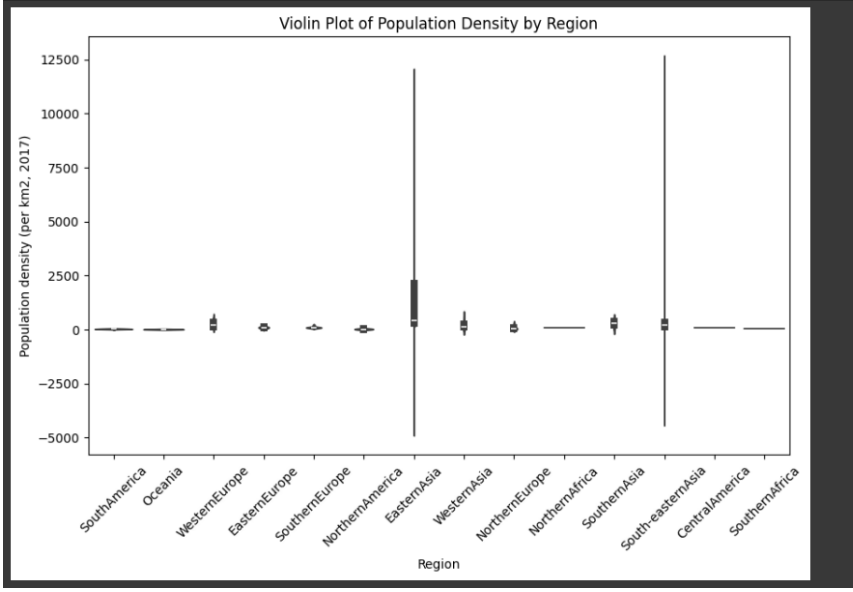
2. Box and Whisker Plot

```
plt.figure(figsize=(10, 6))
sns.boxplot(x='Region', y='GDP per capita (current US$)', data=df)
plt.xticks(rotation=45)
plt.title('Box Plot of GDP per Capita by Region')
plt.show()
```



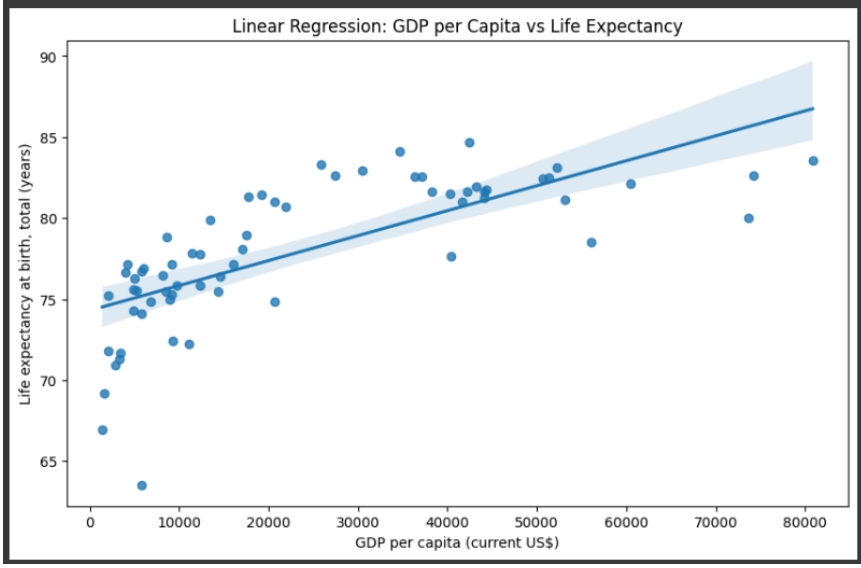
3. Violin Plot

```
plt.figure(figsize=(10, 6))
sns.violinplot(x='Region', y='Population density (per km2, 2017)', data=df)
plt.xticks(rotation=45)
plt.title('Violin Plot of Population Density by Region')
plt.show()
```



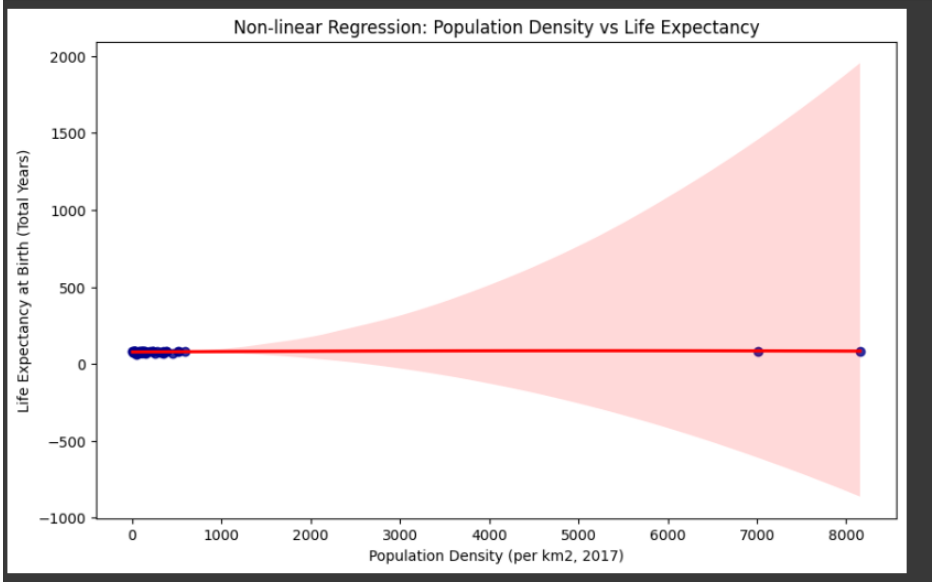
4. Linear Regression Plot

```
plt.figure(figsize=(10, 6))
sns.regplot(x='GDP per capita (current US$)', y='Life expectancy at birth, total (years)', data=df)
plt.title('Linear Regression: GDP per Capita vs Life Expectancy')
plt.show()
```



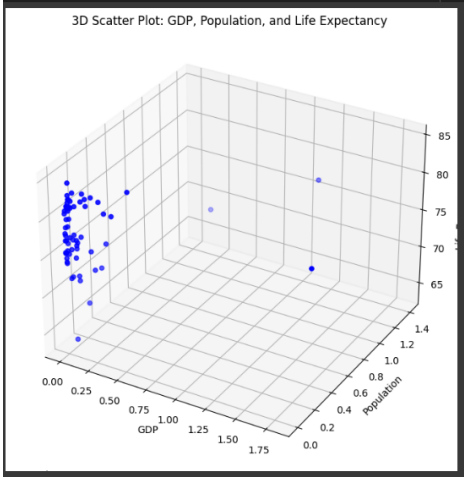
5. Non-linear Regression Plot

```
plt.figure(figsize=(10, 6))
sns.regplot(x='Population density (per km2, 2017)', y='Life expectancy at birth, total (years)',
            data=df, scatter_kws={"color": "darkblue"}, line_kws={"color": "red"}, order=2)
plt.title('Non-linear Regression: Population Density vs Life Expectancy')
plt.xlabel('Population Density (per km2, 2017)')
plt.ylabel('Life Expectancy at Birth (Total Years)')
plt.show()
```



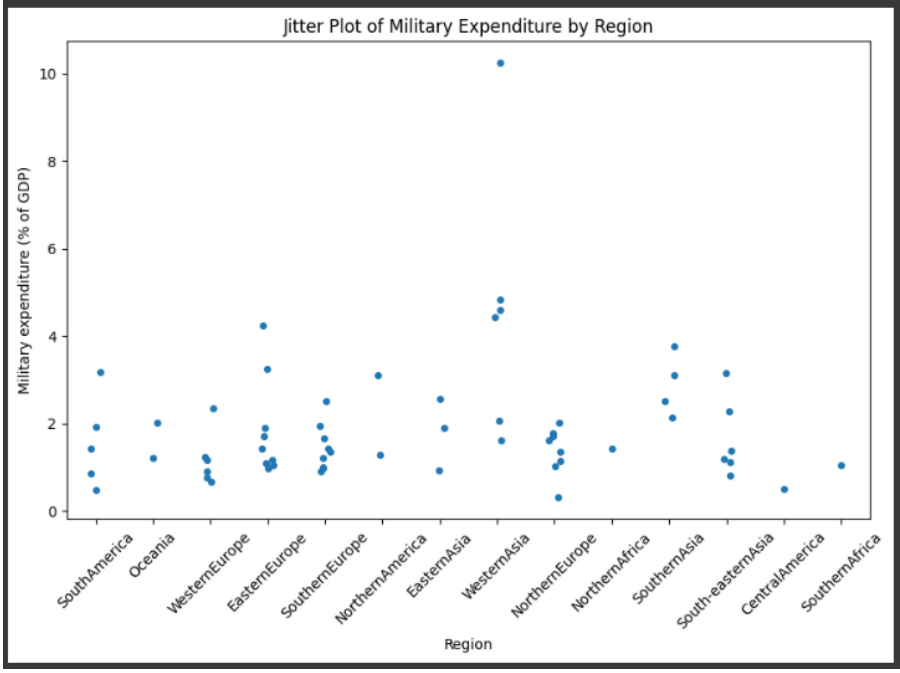
6. 3D Scatter Plot

```
fig = plt.figure(figsize=(10, 8))
ax = fig.add_subplot(111, projection='3d')
ax.scatter(df['GDP: Gross domestic product (million current US$)'], df['Population in thousands (2017)'], df['Life expectancy at birth, total (years)'], c='b', marker='o')
ax.set_xlabel('GDP')
ax.set_ylabel('Population')
ax.set_zlabel('Life Expectancy')
plt.title('3D Scatter Plot: GDP, Population, and Life Expectancy')
plt.show()
```



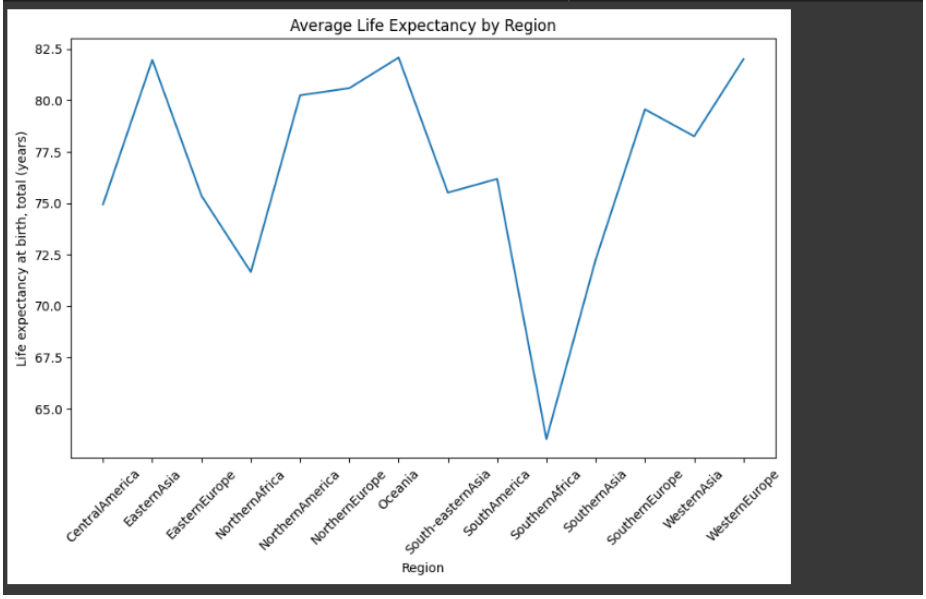
7. Jitter Plot

```
plt.figure(figsize=(10, 6))
sns.stripplot(x='Region', y='Military expenditure (% of GDP)', data=df, jitter=True)
plt.xticks(rotation=45)
plt.title('Jitter Plot of Military Expenditure by Region')
plt.show()
```



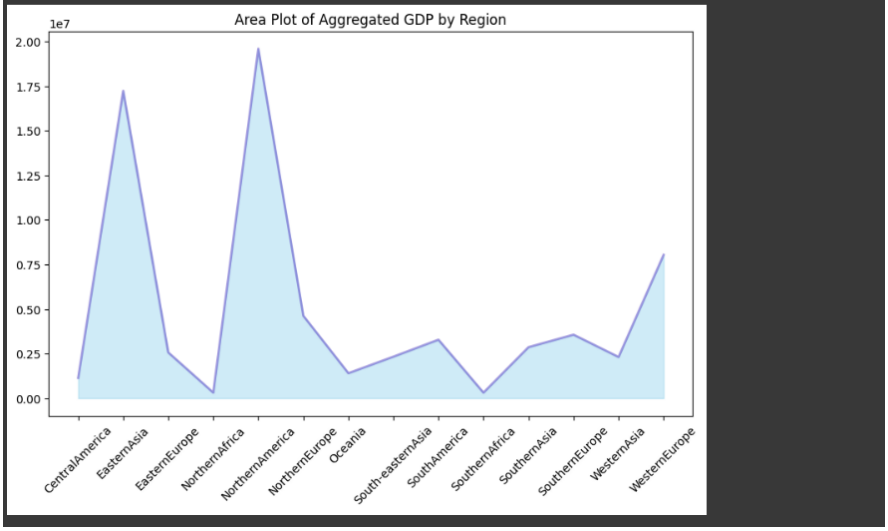
8. Line Plot

```
df_life_exp = df.groupby('Region')['Life expectancy at birth, total (years)'].mean().reset_index()
plt.figure(figsize=(10, 6))
sns.lineplot(x='Region', y='Life expectancy at birth, total (years)', data=df_life_exp)
plt.xticks(rotation=45)
plt.title('Average Life Expectancy by Region')
plt.show()
```



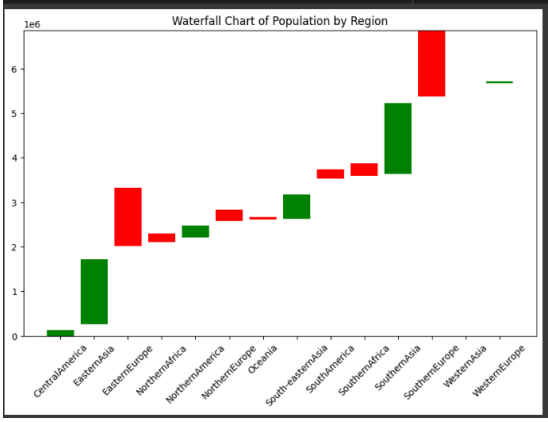
8. Area Plot

```
df_gdp = df.groupby('Region')['GDP: Gross domestic product (million current US$)'].sum().reset_index()
plt.figure(figsize=(10, 6))
plt.fill_between(df_gdp['Region'], df_gdp['GDP: Gross domestic product (million current US$)'], color="skyblue", alpha=0.4)
plt.plot(df_gdp['Region'], df_gdp['GDP: Gross domestic product (million current US$)'], color="Slateblue", alpha=0.6, linewidth=2)
plt.xticks(rotation=45)
plt.title('Area Plot of Aggregated GDP by Region')
plt.show()
```



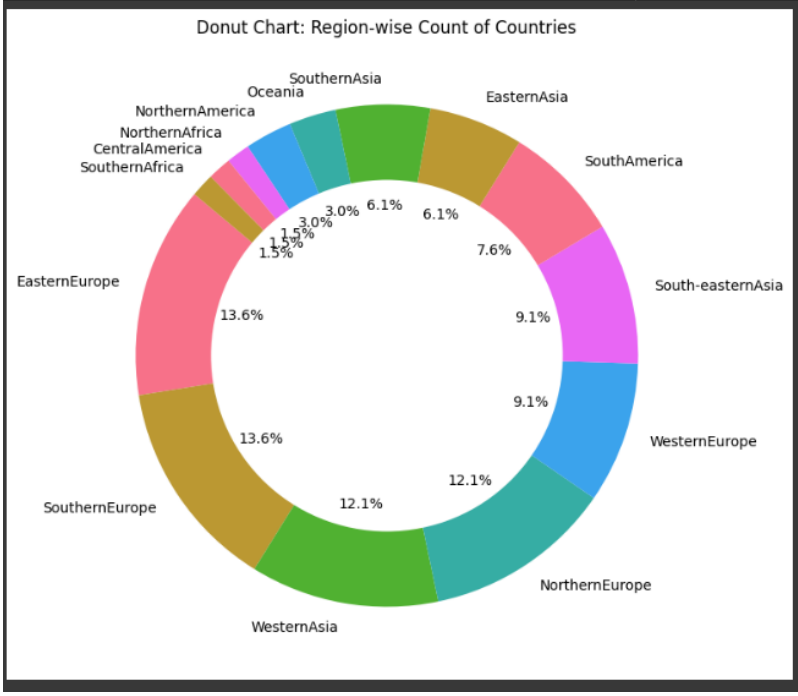
10. Waterfall Chart

```
population_by_region = df.groupby('Region')['Population in thousands (2017)'].sum().reset_index()
population_by_region['previous'] = population_by_region['Population in thousands (2017)'].shift(1, fill_value=0)
population_by_region['difference'] = population_by_region['Population in thousands (2017)'] - population_by_region['previous']
population_by_region['total'] = population_by_region['Population in thousands (2017)'].cumsum()

fig, ax = plt.subplots(figsize=(10, 6))
ax.bar(population_by_region['Region'], population_by_region['difference'], bottom=population_by_region['total'] - population_by_region['difference'], color=(population_by_region['difference'] > 0).map({True: 'green', False: 'red'}))
plt.title('Waterfall Chart of Population by Region')
plt.xticks(rotation=45)
plt.show()
```



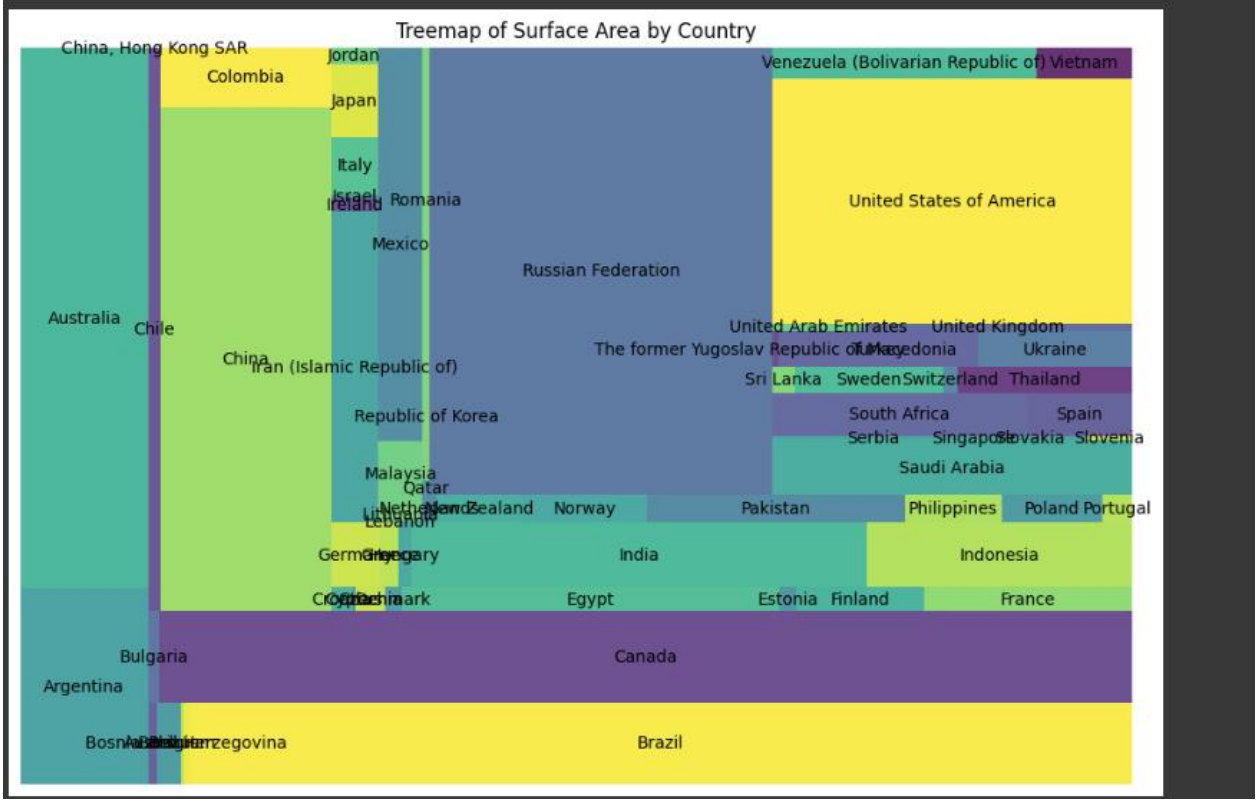
11. Donut Chart

```
region_count = df['Region'].value_counts()
plt.figure(figsize=(8, 8))
plt.pie(region_count, labels=region_count.index, wedgeprops=dict(width=0.4), startangle=140, autopct='%1.1f%%', colors=sns.color_palette('husl'))
plt.gca().add_artist(plt.Circle((0, 0), 0.70, fc='white'))
plt.title('Donut Chart: Region-wise Count of Countries')
plt.show()
```



12. Treemap Chart

```
sizes = df['Surface area (km2)']
labels = df['country']
plt.figure(figsize=(12, 8))
squarify.plot(sizes=sizes, label=labels, alpha=0.8)
plt.axis('off')
plt.title('Treemap of Surface Area by Country')
plt.show()
```



13. Funnel Chart

Conclusion:

The analysis of the socio-economic and demographic dataset using advanced visualizations has yielded valuable insights into global trends, regional disparities, and key relationships among indicators such as GDP, population, life expectancy, and military expenditure. Below are the key conclusions drawn:

1. **Global Prominence and Frequency**:
    - Countries like the United States, China, and India dominate the dataset, underscoring their significant influence in global discussions and analysis.
2. **Economic Disparities**:
    - The Box and Whisker plot along with the Area plot highlight stark economic disparities across regions. Western Europe and North America lead in GDP per capita and overall GDP contributions, concentrating global wealth, while regions like Africa and parts of Asia contribute far less economically.
3. **Population Dynamics**:
    - The Violin plot and Waterfall chart demonstrate considerable variation in population density and total population across regions. Asia emerges as the most populous region, with South America showing notable diversity in population density.
4. **Health and Economic Correlations**:
    - Both linear and non-linear regression plots show a strong positive correlation between GDP per capita and life expectancy, suggesting that wealthier nations tend to have better health outcomes. However, the non-linear regression also reveals that extremely high population densities may negatively affect life expectancy.
5. **Military Expenditure Variability**:
    - The Jitter plot reveals significant differences in military spending across regions, particularly in the Middle East, where several countries allocate a substantial portion of their GDP to defense spending.
6. **Regional Health Disparities**:
    - The Line plot highlights that Western Europe enjoys the highest average life expectancy, while other regions, particularly Eastern Europe and parts of Africa, show lower figures, indicating ongoing health disparities.
7. **Geographical Size Comparison**:
    - The Treemap visualizes the substantial differences in land area among countries, with Russia, Canada, and China being the largest, while many European countries occupy relatively small land areas.
8. **Population Distribution and Concentration**:
    - The Funnel chart emphasizes the overwhelming concentration of the global population in Asia, followed by Africa, showcasing the demographic weight of these regions.

These insights provide a comprehensive view of global socio-economic and demographic patterns, shedding light on both challenges and opportunities in different regions.