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:	**Vaishnavi Borkar**
UID:	**2021300016**
Batch:	**COMPS A (Batch G)**
Exp:	**2**

Aim:

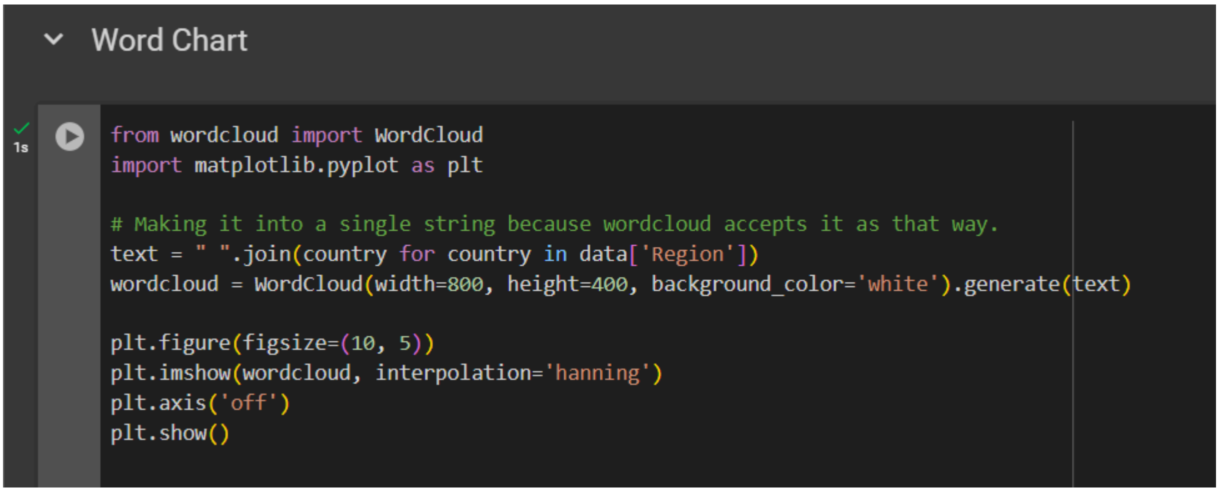
- Create advanced charts using Tableau / Power BI / R / Python / Plotly or Chart or D3.js to be performed on the dataset - Socio economic data
- Advanced - Word chart, Box and whisker plot, Violin plot, Regression plot (linear and nonlinear), 3D chart, Jitter, Line, Area, Waterfall, Donut, Treemap, Funnel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

Dataset Description:
This dataset contains about 62 statistical indicators of the 66 countries. It covers a broad spectrum of areas including General Information, Broader Economic Indicators, Social Indicators, Environmental & Infrastructure Indicators, Military Spending, Healthcare Indicators, Trade Related Indicators e.t.c.

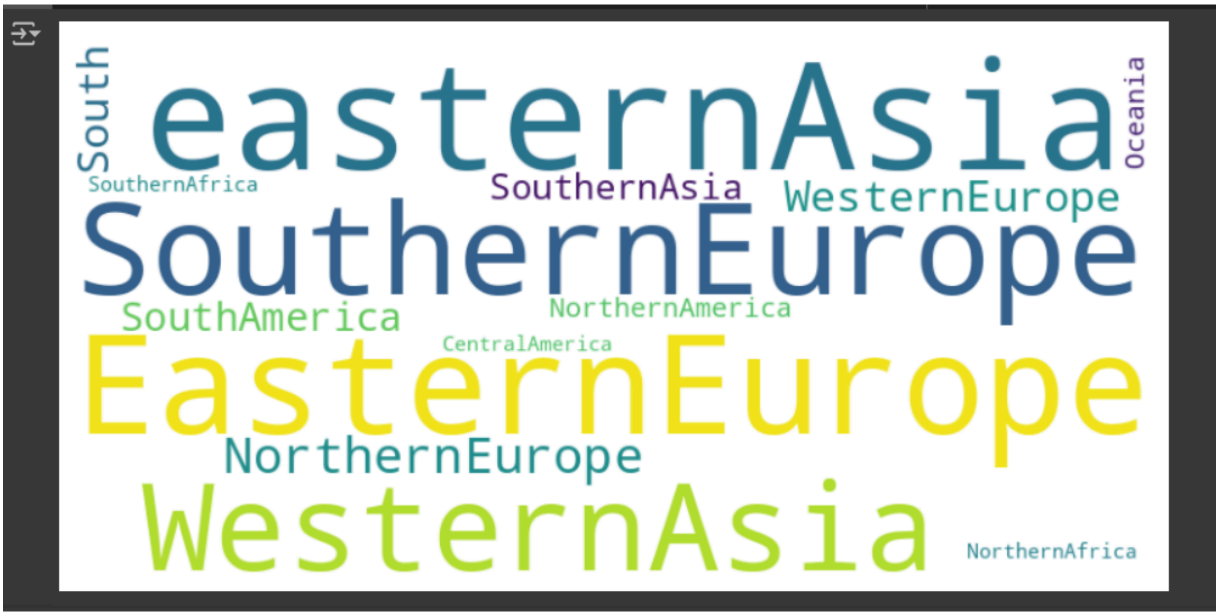
Dataset Link:
https://www.kaggle.com/datasets/nishanthsalian/socioeconomic-country-profiles?resource=download

Plots and Inference:

Word Chart

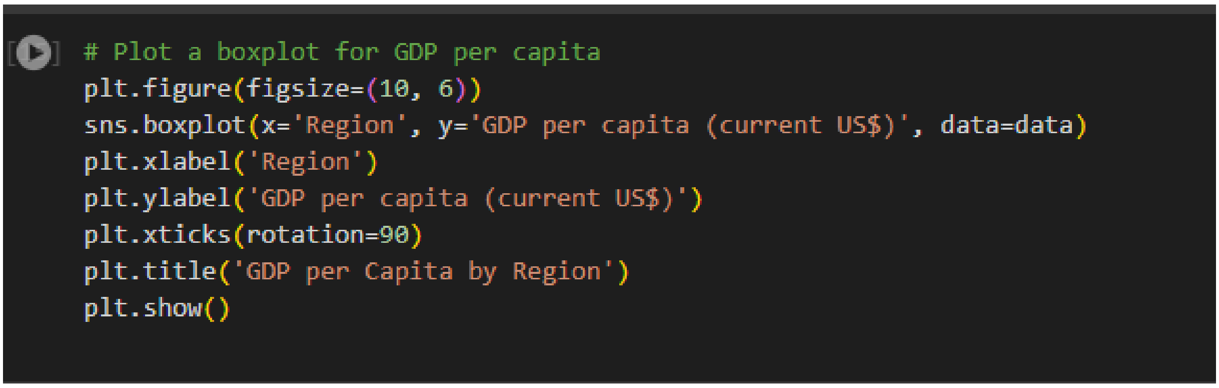


```
from wordcloud import WordCloud
import matplotlib.pyplot as plt

# Making it into a single string because wordcloud accepts it as that way.
text = " ".join(country for country in data['Region'])
wordcloud = WordCloud(width=800, height=400, background_color='white').generate(text)

plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='hanning')
plt.axis('off')
plt.show()
```

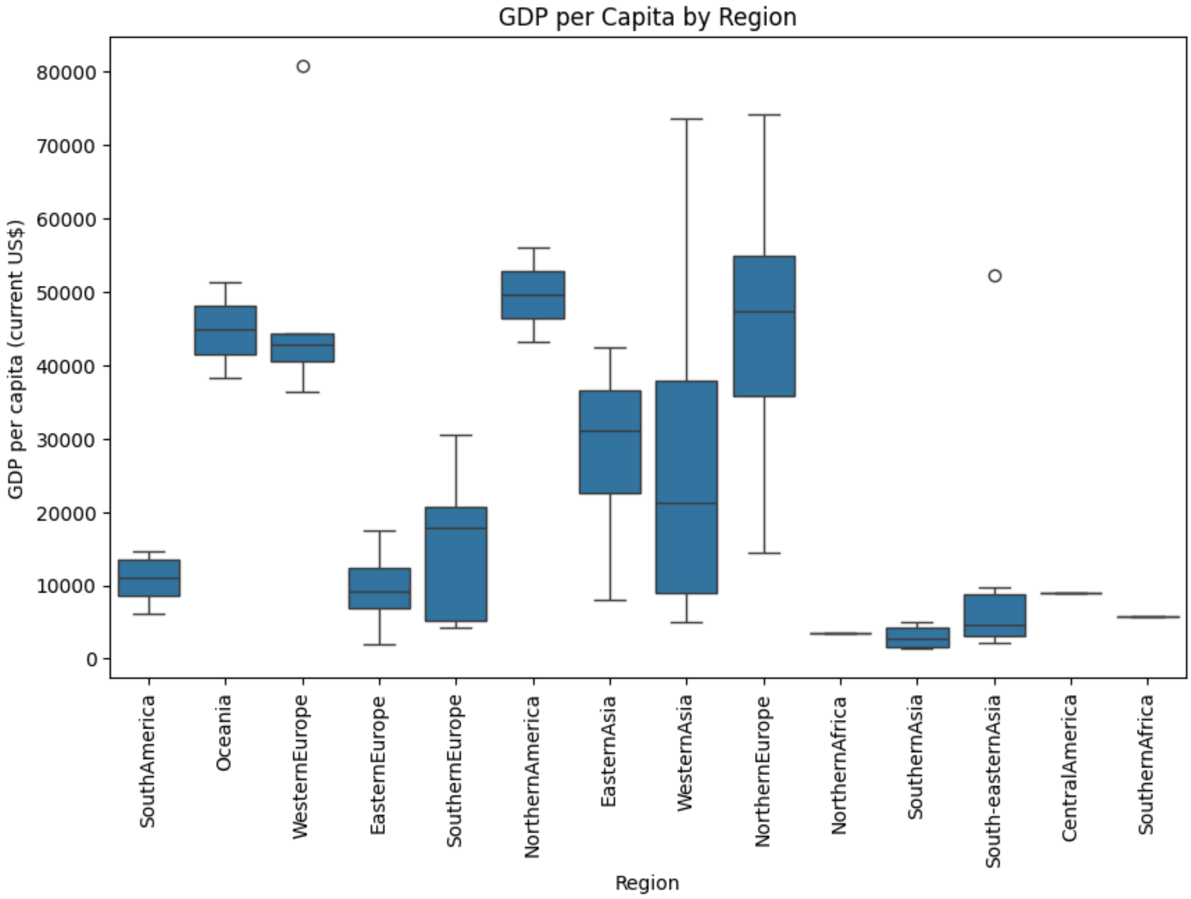


This word cloud tells us which region of the world appears frequently i.e maximum countries of which region are included in the dataset. We can find that more countries of Eastern Europe are there.

Box and Whisker Plot

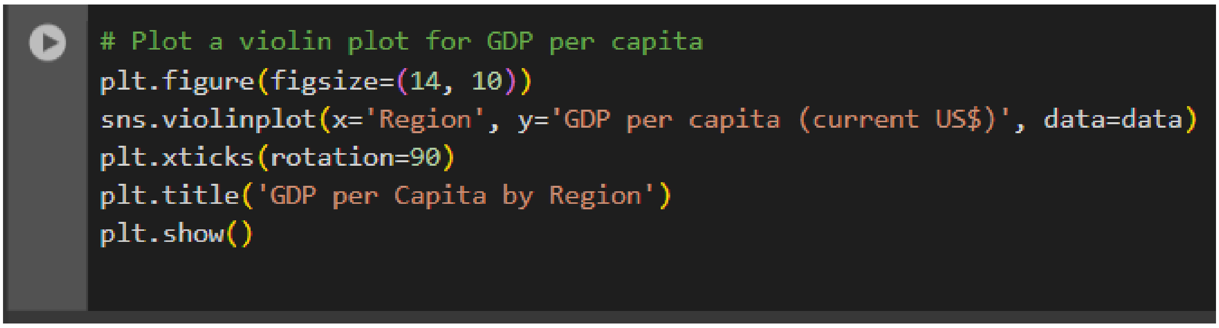


```
# Plot a boxplot for GDP per capita
plt.figure(figsize=(10, 6))
sns.boxplot(x='Region', y='GDP per capita (current US$)', data=data)
plt.xlabel('Region')
plt.ylabel('GDP per capita (current US$)')
plt.xticks(rotation=90)
plt.title('GDP per Capita by Region')
plt.show()
```

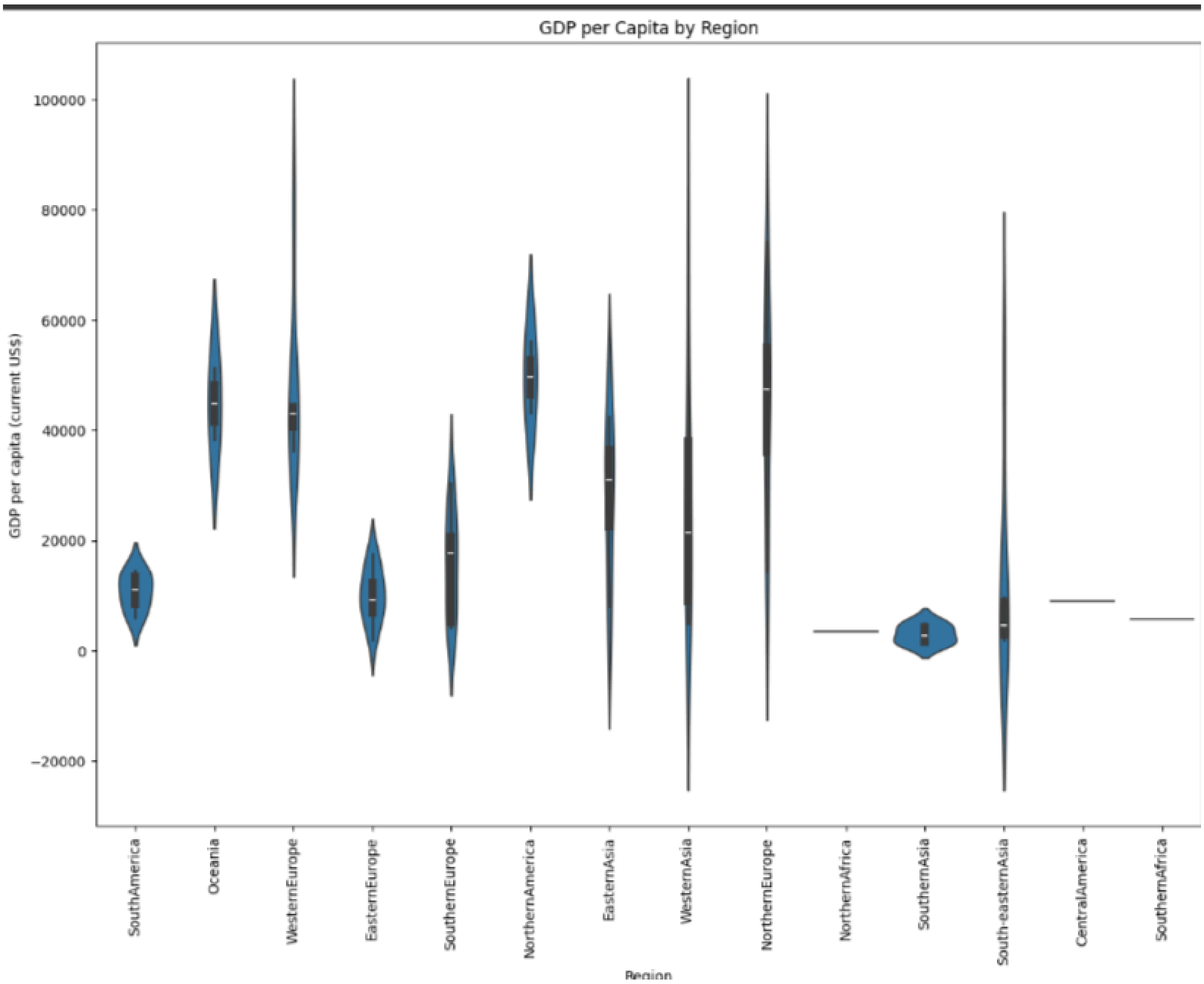


Northern America and Northern Europe have higher median GDP per capita, indicating a relatively higher standard of living and economic prosperity. Western and Northern Europe have much higher GDP per capita compared to regions like Southern Asia and Central America suggesting economic disparity.

Violin Plot

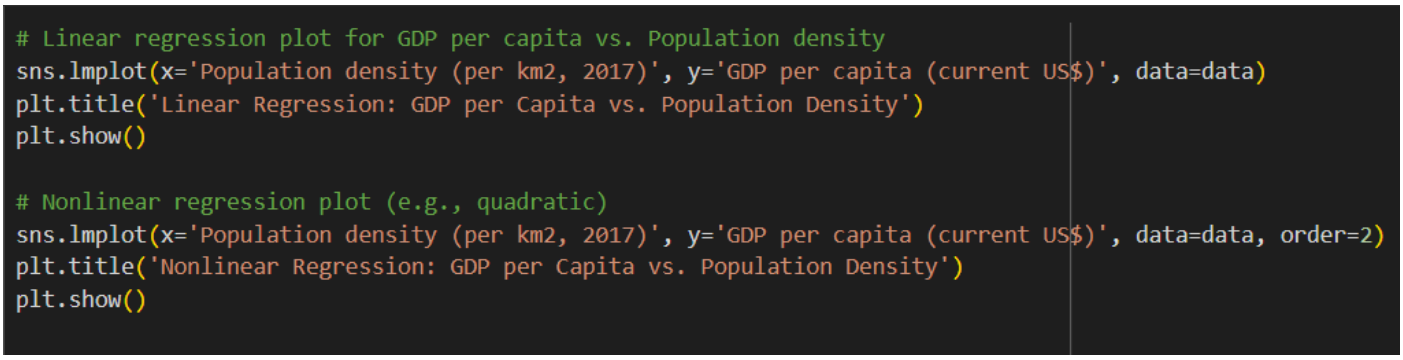


```
# Plot a violin plot for GDP per capita
plt.figure(figsize=(14, 10))
sns.violinplot(x='Region', y='GDP per capita (current US$)', data=data)
plt.xticks(rotation=90)
plt.title('GDP per Capita by Region')
plt.show()
```

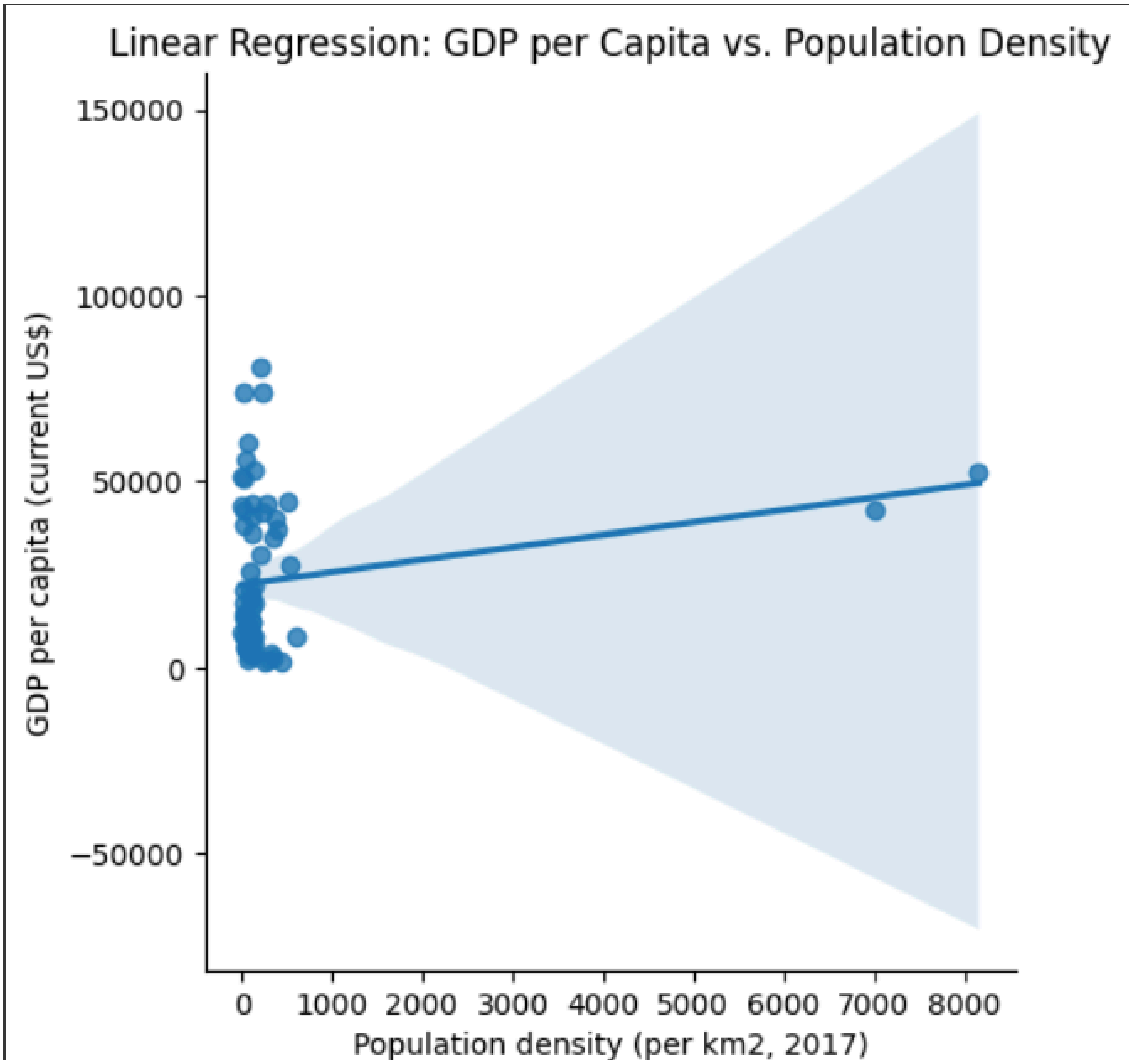


This is the same as the box-whisker plot but it also shows the density of the values plotted.

Regression Plot(linear and non-linear)

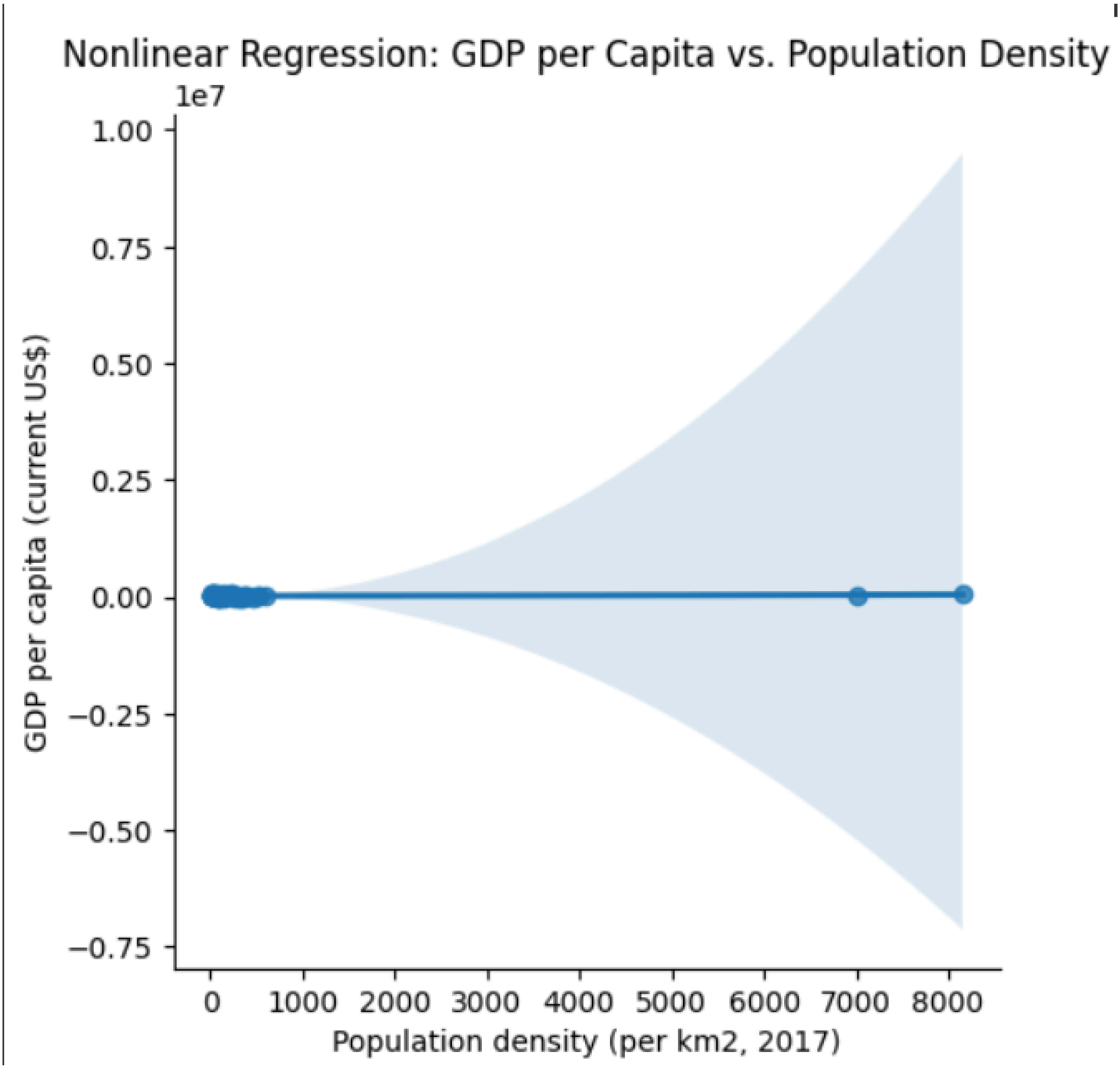


```
# Linear regression plot for GDP per capita vs. Population density
sns.lmplot(x='Population density (per km2, 2017)', y='GDP per capita (current US$)', data=data)
plt.title('Linear Regression: GDP per Capita vs. Population Density')
plt.show()

# Nonlinear regression plot (e.g., quadratic)
sns.lmplot(x='Population density (per km2, 2017)', y='GDP per capita (current US$)', data=data, order=2)
plt.title('Nonlinear Regression: GDP per Capita vs. Population Density')
plt.show()
```



Positive Correlation: The regression line has a positive slope, suggesting a slight positive relationship between population density and GDP per capita. As population density increases, GDP per capita tends to increase, but the relationship appears weak.

Wide Confidence Interval: The shaded area around the regression line represents the confidence interval, which widens significantly as population density increases. This indicates greater uncertainty or variability in the predicted GDP per capita at higher population densities.

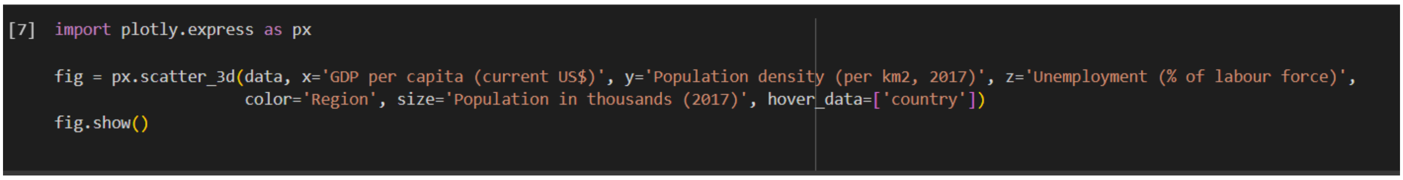


Flat Nonlinear Relationship: The regression line is essentially flat, indicating no apparent relationship between population density and GDP per capita in this nonlinear model. The predicted GDP per capita remains constant regardless of changes in population density.

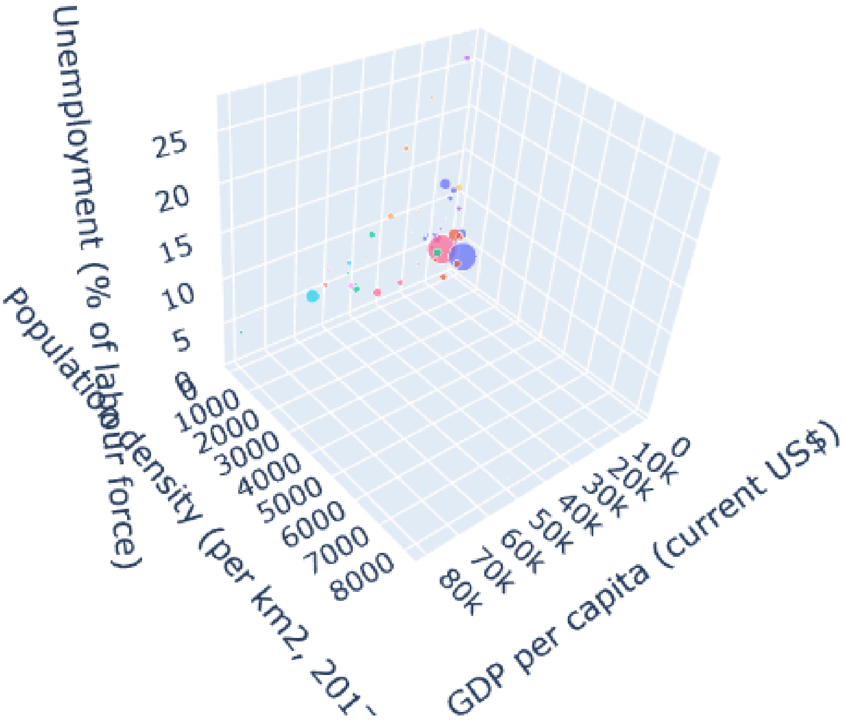
Expanding Confidence Interval: The confidence interval (shaded region) widens dramatically as population density increases. This suggests higher uncertainty in the predictions for higher population densities. The model may be overfitting to the limited data at high population densities, leading to larger confidence intervals.

Cluster of Data: Similar to the linear regression plot, the majority of the data points are tightly clustered at low population densities (under 2000 km²), with no significant variation in GDP per capita. This suggests the model has little information at higher population densities to make accurate predictions.

3D Chart

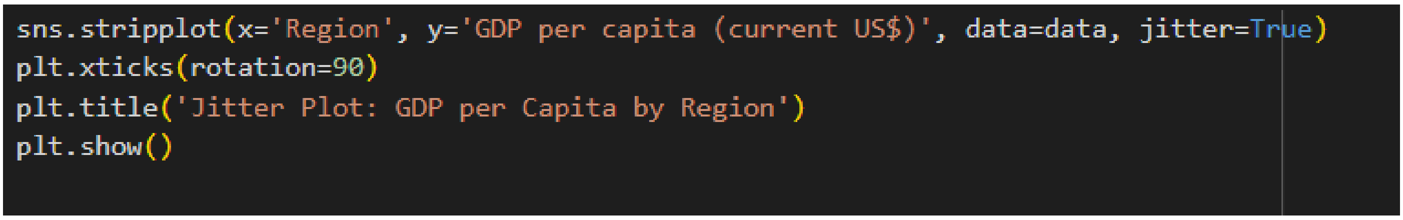


```
[7] import plotly.express as px

    fig = px.scatter_3d(data, x='GDP per capita (current US$)', y='Population density (per km2, 2017)', z='Unemployment (% of labour force)',
                        color='Region', size='Population in thousands (2017)', hover_data=['country'])
    fig.show()
```

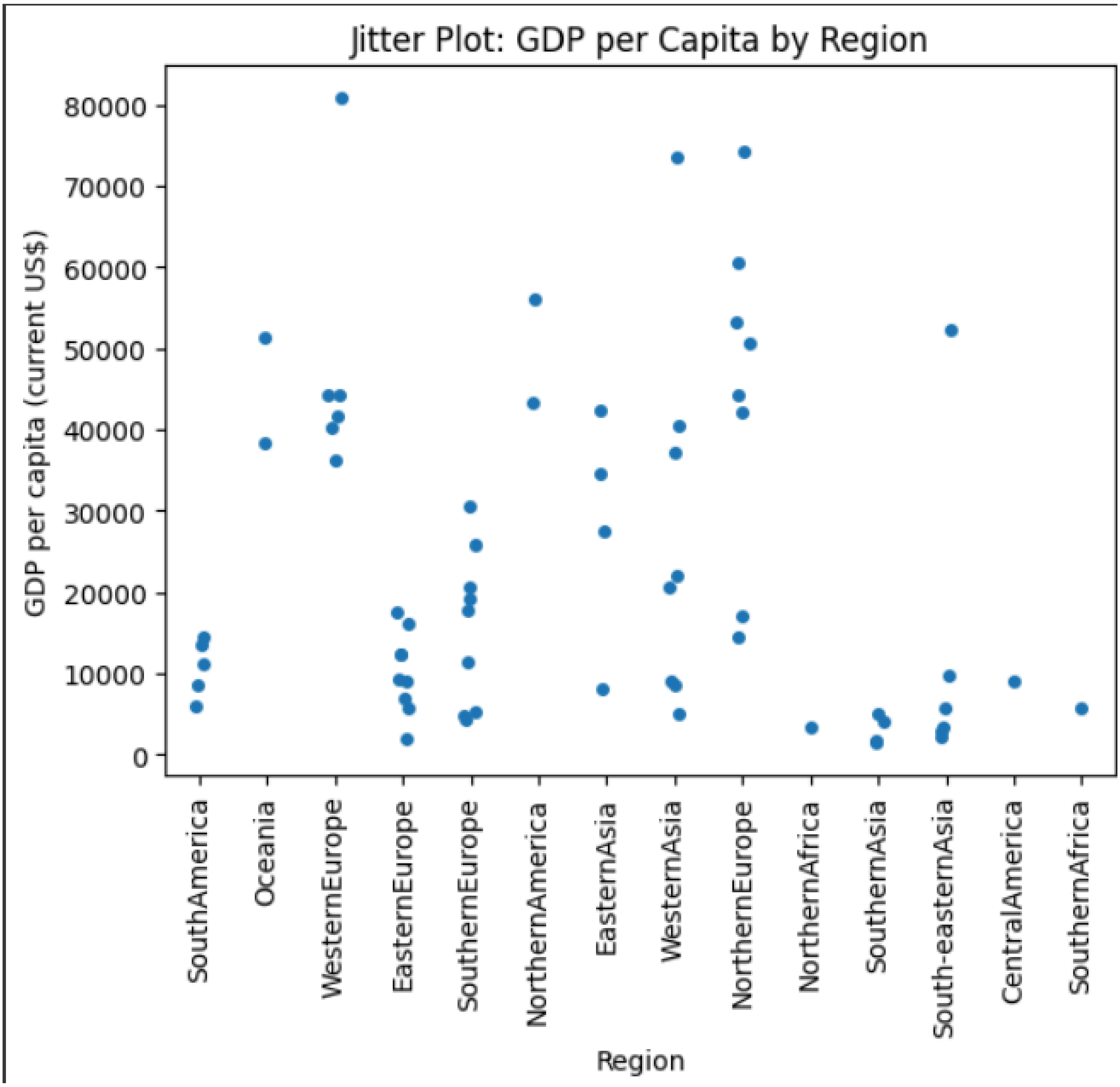


Most of the data points are clustered at lower population densities, unemployment rates, and GDP per capita values. This indicates that the majority of observations fall within a narrow range for these variables

Jitter Chart

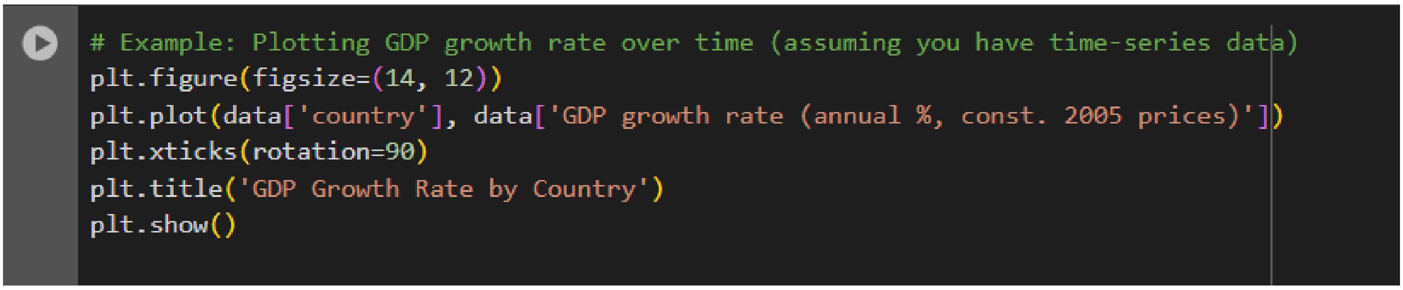


```
sns.stripplot(x='Region', y='GDP per capita (current US$)', data=data, jitter=True)
plt.xticks(rotation=90)
plt.title('Jitter Plot: GDP per Capita by Region')
plt.show()
```

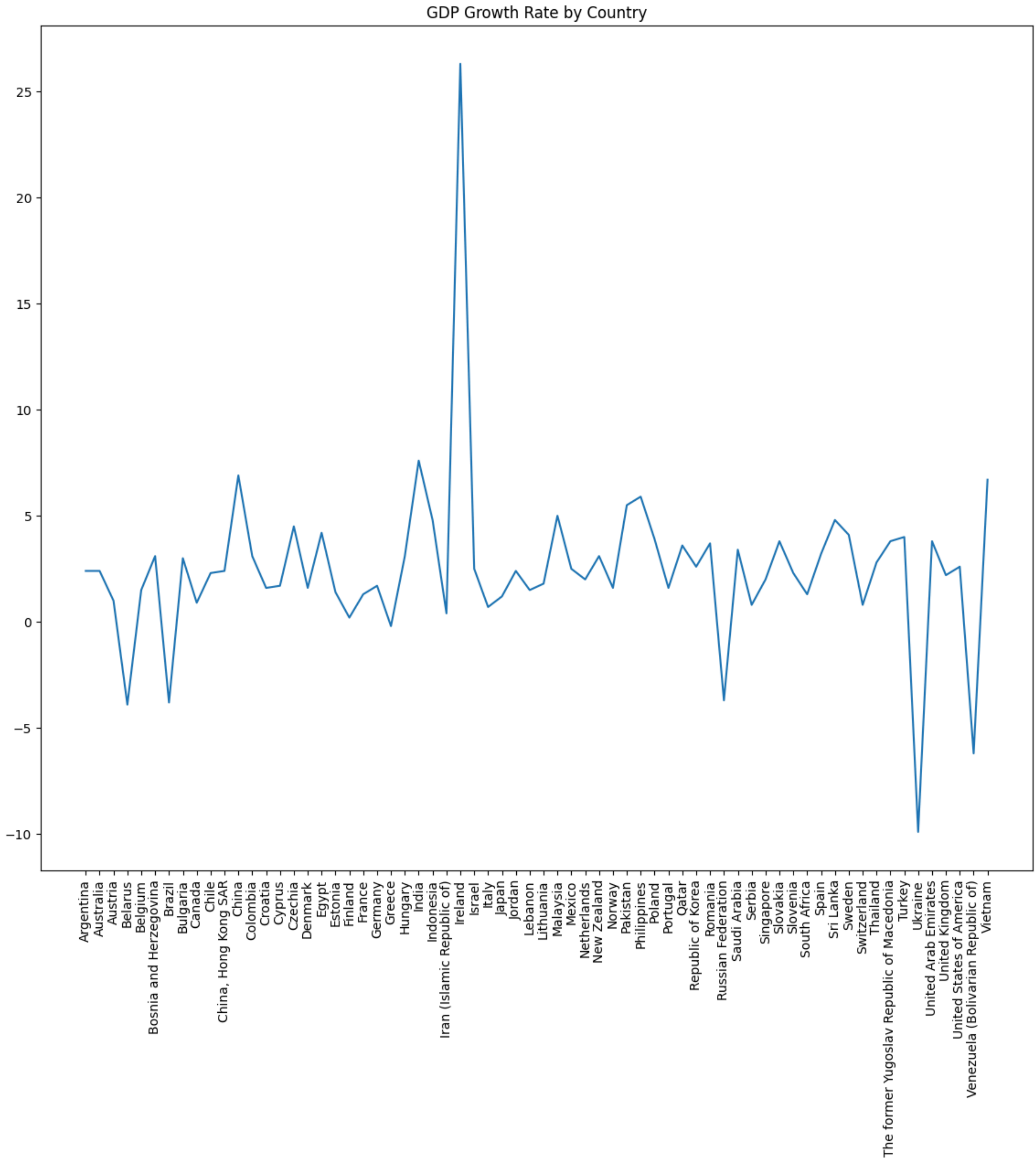


This plot shows the wide disparities in GDP per capita across different global regions, with regions like **North America** and **Western Europe** enjoying significantly higher income levels compared to regions like **South America** or **Africa**. The plot effectively highlights the unequal distribution of wealth on a global scale.

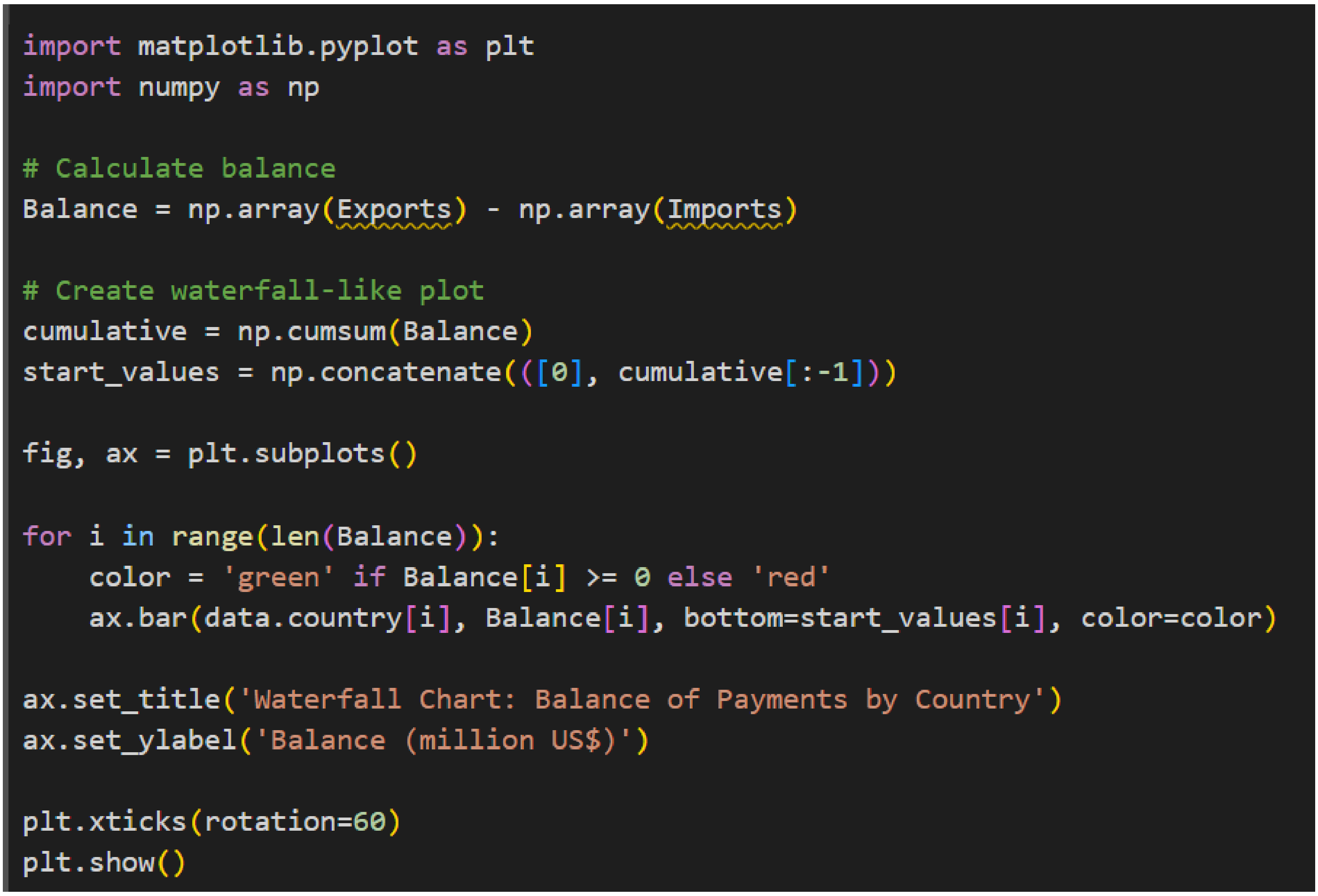
Line Chart



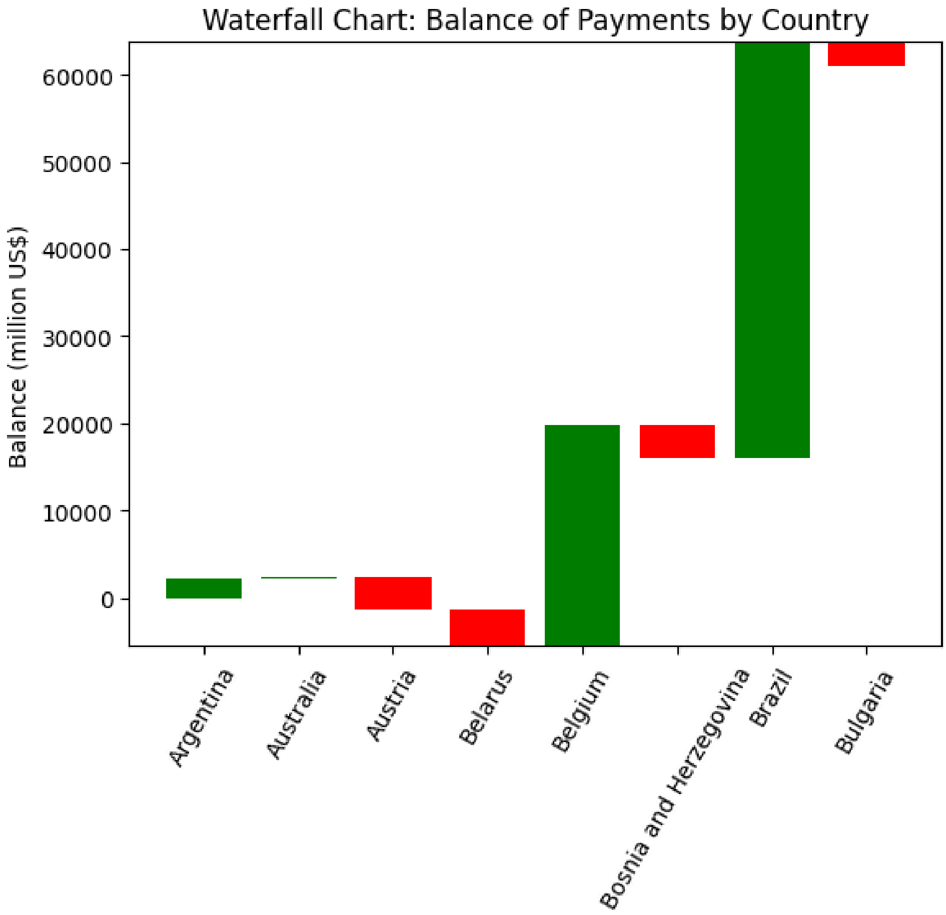
```
# Example: Plotting GDP growth rate over time (assuming you have time-series data)
plt.figure(figsize=(14, 12))
plt.plot(data['country'], data['GDP growth rate (annual %, const. 2005 prices)'])
plt.xticks(rotation=90)
plt.title('GDP Growth Rate by Country')
plt.show()
```


GDP Growth Rate by Country
25
20
15
10
5
0
−5
−10
Argentina
Australia
Austria
Belarus
Belgium
Bosnia and Herzegovina
Brazil
Bulgaria
Canada
Chile
China, Hong Kong SAR
China
Colombia
Croatia
Cyprus
Czechia
Denmark
Egypt
Estonia
Finland
France
Germany
Greece
Hungary
India
Indonesia
Iran (Islamic Republic of)
Ireland
Israel
Italy
Japan
Jordan
Lebanon
Lithuania
Malaysia
Mexico
Netherlands
New Zealand
Norway
Pakistan
Philippines
Poland
Portugal
Qatar
Republic of Korea
Romania
Russian Federation
Saudi Arabia
Serbia
Singapore
Slovakia
Slovenia
South Africa
Spain
Sri Lanka
Sweden
Switzerland
Thailand
The former Yugoslav Republic of Macedonia
Turkey
Ukraine
United Arab Emirates
United Kingdom
United States of America
Venezuela (Bolivarian Republic of)
Vietnam

Waterfall Chart



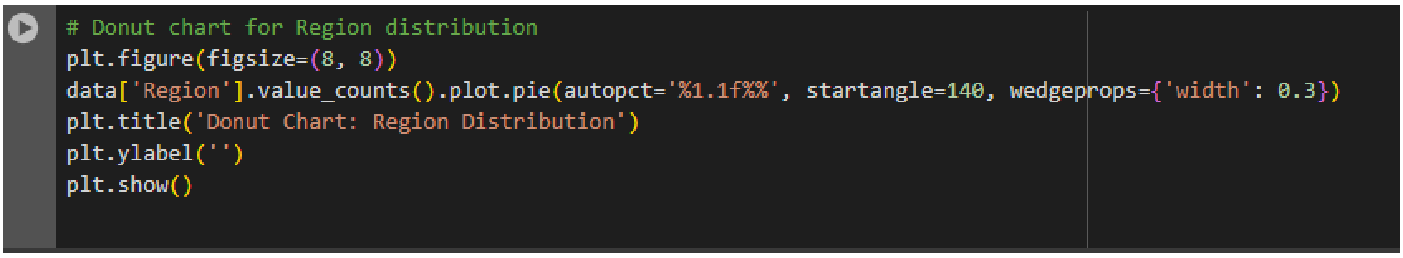
```python
import matplotlib.pyplot as plt
import numpy as np

# Calculate balance
Balance = np.array(Exports) - np.array(Imports)

# Create waterfall-like plot
cumulative = np.cumsum(Balance)
start_values = np.concatenate(([0], cumulative[:-1]))

fig, ax = plt.subplots()

for i in range(len(Balance)):
    color = 'green' if Balance[i] >= 0 else 'red'
    ax.bar(data.country[i], Balance[i], bottom=start_values[i], color=color)

ax.set_title('Waterfall Chart: Balance of Payments by Country')
ax.set_ylabel('Balance (million US$)')

plt.xticks(rotation=60)
plt.show()
```



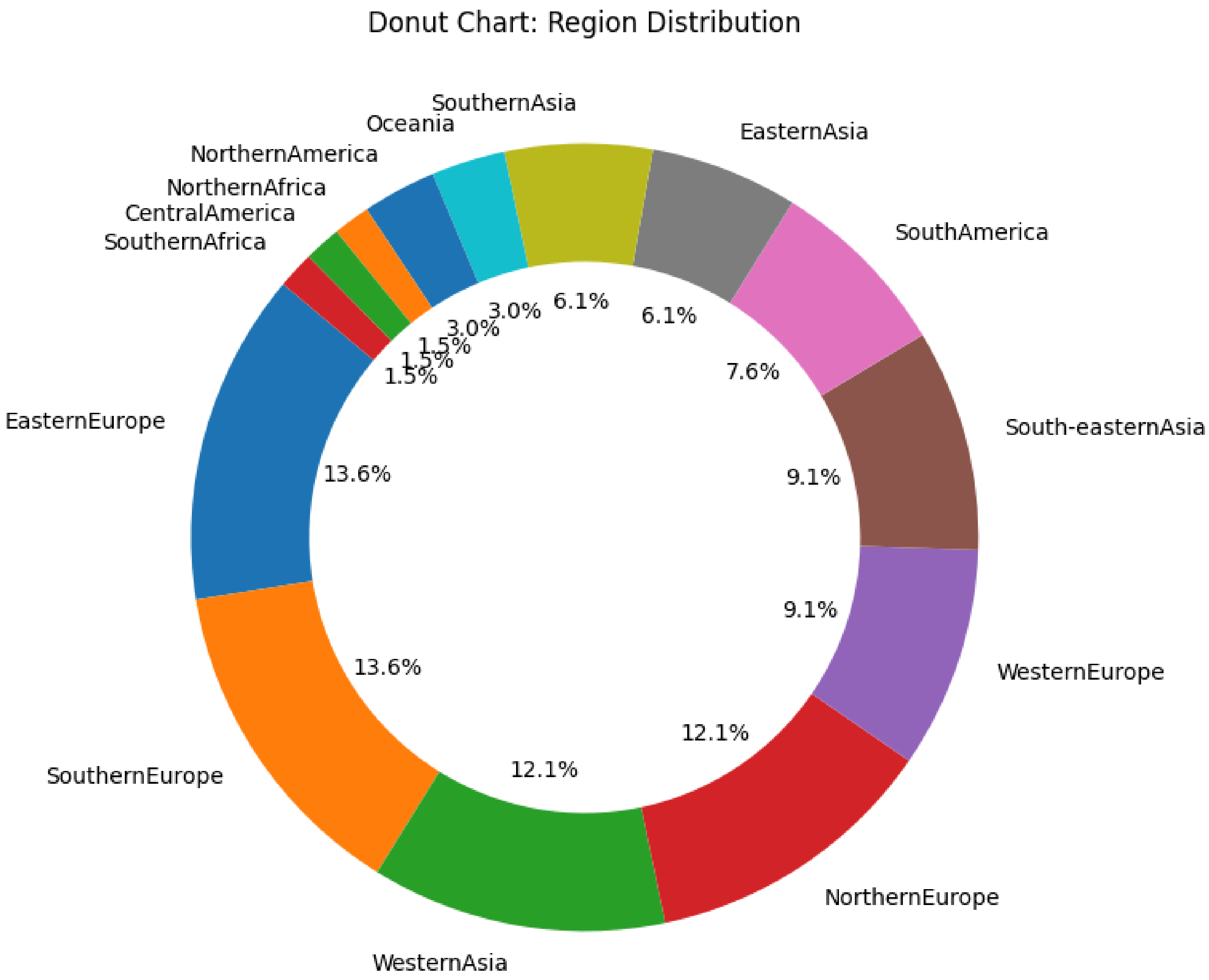
Brazil and Belgium are the strongest performers in terms of balance of payments, with significant surpluses.

Belarus and Austria struggle with deficits, indicating a negative balance of payments for these countries.

Donut Chart



```
# Donut chart for Region distribution
plt.figure(figsize=(8, 8))
data['Region'].value_counts().plot.pie(autopct='%1.1f%%', startangle=140, wedgeprops={'width': 0.3})
plt.title('Donut Chart: Region Distribution')
plt.ylabel('')
plt.show()
```



13.6% data is of the Southern Europe region.