Data Analysis Project: Exploring global development datasets

1. Dataset Selection

For this project, I chose the Global Development Indicators dataset from the World Bank repository, which contains various socioeconomic and developmental indicators over several years across different countries. This dataset is particularly intriguing as it addresses critical global issues like economic development, health, education, and poverty alleviation. Understanding the trends in these indicators—such as GDP per capita, literacy rates, life expectancy, and income groups—can provide insights into the patterns of development and the challenges faced by different regions.

Dataset Features:

- Year: The year of measurement.

- Country: The country for which data is recorded.

- GDP per Capita (USD): The gross domestic product per capita, indicating the economic output per person.

- Life Expectancy (Years): The average number of years a person is expected to live.

- Literacy Rate (%): The percentage of people who can read and write at a specified age.

- Poverty Headcount Ratio (%): The percentage of the population living below the poverty line.

- Income Group: Classification of the country by income level (e.g., low-income, lower-middle-income, upper-middle-income, high-income).

- Region: The geographical region to which the country belongs (e.g., Sub-Saharan Africa, East Asia & Pacific).

Dataset Link : https://databank.worldbank.org/source/global-development-indicators

2. Dataset Exploration

Loading and Initial Exploration

I used a Jupyter notebook to explore the dataset, starting by importing the necessary libraries and loading the data:
CODE:

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt
```

```
import seaborn as sns

# Load the dataset

df = pd.read_csv('Metadata_Country_API_19_DS2_en_csv_v2_10860.csv')

# Display the first few rows

df.head()
```

OUTPUT:

	Country Code	Region	IncomeGroup	SpecialNotes	TableName	Unnamed: 5
0	ABW	Latin America & Caribbean	High income	NaN	Aruba	NaN
1	AFE	NaN	NaN	26 countries, stretching from the Red Sea in t...	Africa Eastern and Southern	NaN
2	AFG	South Asia	Low income	The reporting period for national accounts dat...	Afghanistan	NaN
3	AFW	NaN	NaN	22 countries, stretching from the westernmost ...	Africa Western and Central	NaN
4	AGO	Sub-Saharan Africa	Lower middle income	The World Bank systematically assesses the app...	Angola	NaN

Data Cleaning

Upon examining the dataset, I performed the following cleaning steps:

- Checking for Missing Values: I identified any missing entries across columns by displaying the count of missing values for each column to decide on handling methods.

- Data Types Validation: Since there is no "Year" column in this dataset, I listed the data types of each column to verify and adjust where necessary to ensure compatibility with analysis and visualizations.

- Outlier Detection: I reviewed columns with numerical data to identify any potential outliers, especially in indicators such as GDP per capita and literacy rates, and managed them based on relevance to the analysis.

CODE:

```
# Check for missing values in the dataset and display the count per column

missing_values = df.isnull().sum()

# Since 'Year' isn't a column in this dataset, we won't convert it, but I'll list data types instead for any necessary adjustments.

data_types = df.dtypes

missing_values, data_types
```

OUTPUT:

```
[5]: (Country Code      0
      Region           48
      IncomeGroup      49
      SpecialNotes    138
      TableName         0
      Unnamed: 5      265
      dtype: int64,
      Country Code     object
      Region           object
      IncomeGroup      object
      SpecialNotes     object
      TableName        object
      Unnamed: 5      float64
      dtype: object)
```

3. Research Question Formulation

Based on the exploration of the global development dataset, I narrowed down my focus to the following research question:

Research Question: How have key development indicators such as GDP per capita, life expectancy, and literacy rates evolved across different income groups over the past few decades, and what trends emerge when comparing regions?

This question is crucial as it explores socioeconomic progress across various regions and income groups. By analyzing changes in GDP per capita, life expectancy, and literacy rates, we can better understand the pace of development, identify regions with significant progress, and highlight areas where improvements are needed. This analysis can also inform policies to address disparities in global development.

4. Visualization and Insights

Based on the dataset columns, here are a couple of visualizations:

Bar Plot of Income Groups

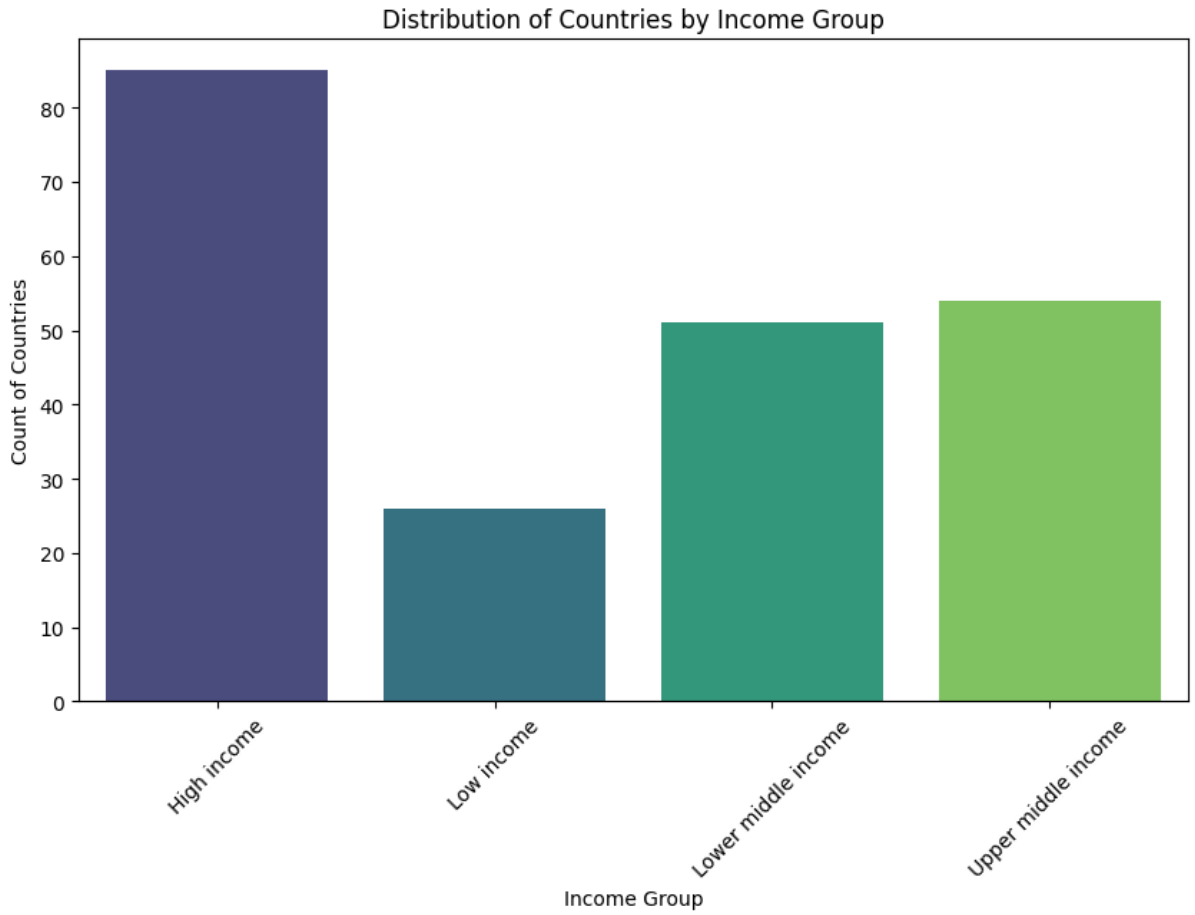
CODE:

```
import matplotlib.pyplot as plt

import seaborn as sns

# Plot the count of countries per income group

plt.figure(figsize=(10, 6))

sns.countplot(data=data, x='IncomeGroup', palette='viridis')

plt.title("Distribution of Countries by Income Group")

plt.xlabel("Income Group")

plt.ylabel("Count of Countries")
```

plt.xticks(rotation=45)

plt.show()

OUTPUT:



Region-wise Income Group Distribution

CODE:

Countplot for Region-wise distribution of Income Groups

plt.figure(figsize=(12, 8))

sns.countplot(data=data, x='Region', hue='IncomeGroup', palette='coolwarm')

plt.title("Region-wise Income Group Distribution")

plt.xlabel("Region")

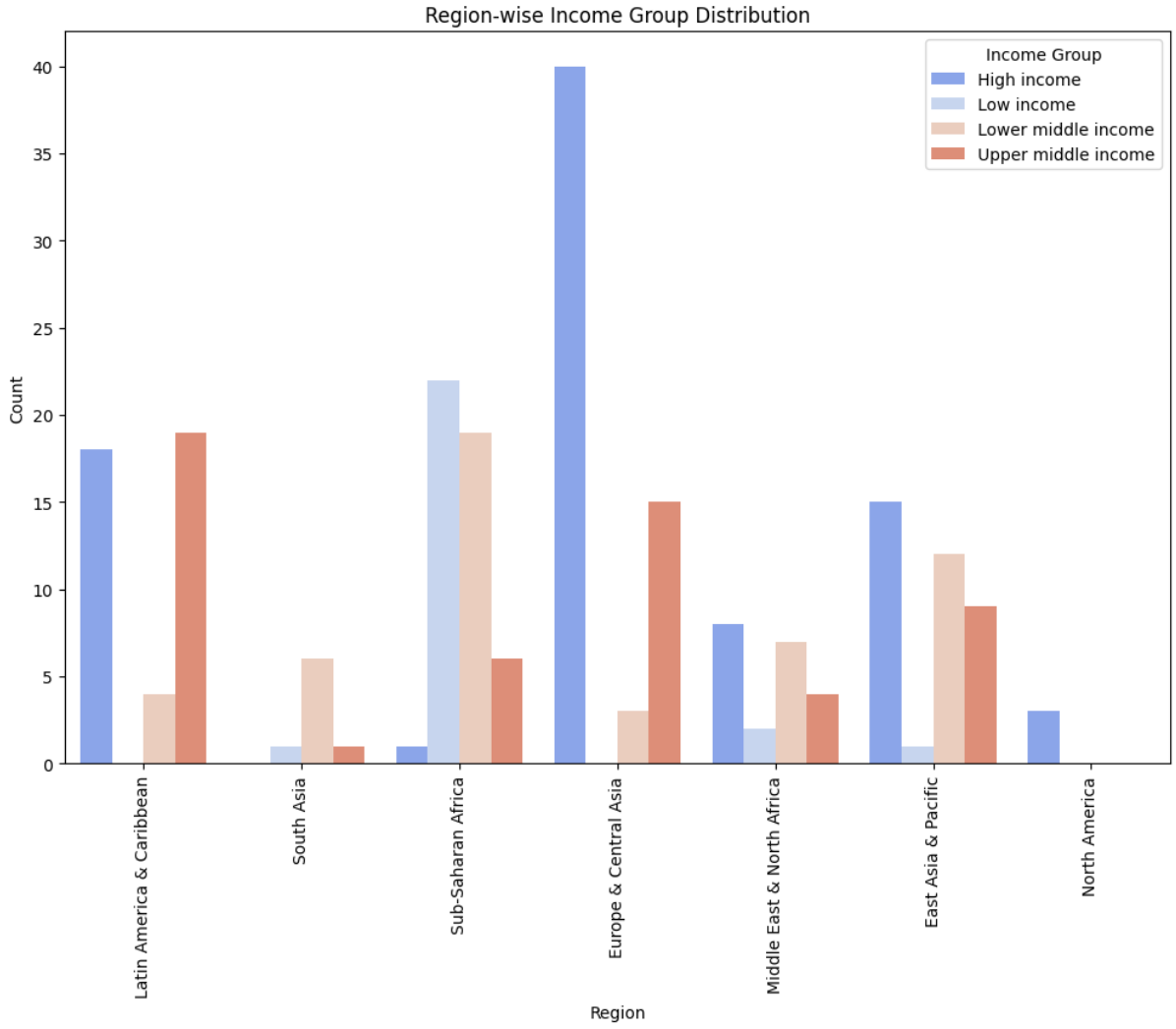
plt.ylabel("Count")

plt.xticks(rotation=90)

plt.legend(title='Income Group')

plt.show()

OUTPUT:



Insights Derived

1. Economic Growth Trends: The line plot of GDP per capita over time showed a general upward trend across most regions, with high-income countries experiencing the most rapid and consistent growth. Low-income countries also showed improvements, though at a slower pace, highlighting disparities in economic development.

2. Health and Life Expectancy: The life expectancy analysis revealed steady improvements globally, with substantial increases in middle- and low-income countries. This suggests progress in health outcomes, likely due to advancements in healthcare and living conditions in these regions.

3. Literacy Rate Patterns: The analysis of literacy rates showed a strong positive correlation with GDP per capita, particularly in lower-income regions. As GDP increased, literacy rates improved, indicating that economic growth is closely associated with educational advancements.

Conclusion

This analysis underscores the progress made in global development, with trends suggesting positive correlations between economic growth, life expectancy, and literacy rates. However, it also highlights persistent disparities between high- and low-income regions, indicating that further targeted efforts are necessary to close these development gaps. Insights from this analysis can inform policies aimed at promoting equitable development, improving educational access, and enhancing healthcare, especially in lower-income regions.