Project 1: GDP and Life Expectancy Analysis

Exploring the Interconnectedness of GDP and Key Socioeconomic Factors

Introduction

GDP and life expectancy are important measures of a country's health and economy. But we can explore deeper and find hidden relationships between these and other affecting variable. This project dives into a data set containing detailed global metrics. By analyzing and visualizing these data points, I aim to explore relationships that show how different factors influence a country's growth and quality of life. This data set consists of 204 columns with 38 rows, covering diverse economic and social metrics across countries worldwide. The data set provides an expansive view of global trends with indicators such as education enrollment, unemployment, homicide rates, CO_2 emissions, and tourism statistics. In this analysis, I will visualize how different indicators relate to one another to find insights into their interactions.

Key features of the data set

GDP: Gross Domestic Product (in current US dollars), representing the total economic output of a country.

Sex Ratio: The ratio of males to females in the population, highlighting demographic trends.

Life Expectancy: Average lifespan for males and females, an essential indicator of healthcare quality.

Education Enrollment Rates: Data on primary, secondary, and post-secondary education enrollment for males and females, reflecting educational attainment.

Unemployment Rate: Percentage of the labor force that is unemployed, indicating economic health.

Homicide Rate: Number of homicides per 100,000 population, providing insight into safety and crime levels.

Urban Population Growth: Rate of growth in urban populations, illustrating migration trends.

CO2 Emissions: Carbon dioxide emissions per capita, an important measure of environmental impact.

Forested Area: Percentage of land covered by forests, indicating biodiversity and environmental health.

Tourist Numbers: Total number of international visitors, which can reflect a country's tourism potential.

Exploratory data analysis (EDA)

1. Import Libraries and Load the Dataset

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


# Raw URL of the CSV file on GitHub
url = 'https://raw.githubusercontent.com/pherathm/gdp_data/main/country_data.csv'

# Since the data set was not readble;
#Read the CSV file into a Pandas data frame with a specified encoding
df = pd.read_csv(url, encoding='ISO-8859-1')  # Use 'ISO-8859-1' or 'latin1'

# Display the first few rows of the data frame
print(df.head())
```

```
        gdp  sex_ratio  surface_area  life_expectancy_male  unemployment  \
0   20514.0      105.4      652864.0                  62.8          11.2
1   15059.0      103.7       28748.0                  76.7          12.8
2  173757.0      102.1     2381741.0                  75.4          11.5
```

```
3    3238.0      102.3        468.0                  NaN           NaN
4  105902.0       97.9    1246700.0                 57.8           6.8

   imports  homicide_rate                              currency iso2  \
0   8370.0            6.7         {'code': 'AFN', 'name': 'Afghani'}   AF
1   5908.0            2.3             {'code': 'ALL', 'name': 'Lek'}   AL
2  45140.0            1.4  {'code': 'DZD', 'name': 'Algerian Dinar'}   DZ
3   1538.0            0.0            {'code': 'EUR', 'name': 'Euro'}   AD
4  21340.0            4.8          {'code': 'AOA', 'name': 'Kwanza'}   AO

   employment_services  ...  pop_growth           region  pop_density  \
0                 39.4  ...         2.5    Southern Asia         59.6
1                 43.7  ...        -0.1  Southern Europe        105.0
2                 59.6  ...         2.0  Northern Africa         18.4
3                  NaN  ...        -0.2  Southern Europe        164.2
4                 41.7  ...         3.3    Middle Africa         26.4

   internet_users gdp_per_capita  fertility  refugees  \
0            13.5          551.9        4.6    2826.4
1            71.8         5223.8        1.6       4.3
2            49.0         4114.7        3.0      99.5
3            91.6        42051.6        1.2       NaN
4            14.3         3437.3        5.6      70.1

   primary_school_enrollment_male  co2_emissions  tourists
0                           124.2            NaN       NaN
1                           105.2            4.3    5340.0
2                           112.4          130.5    2657.0
3                             NaN            NaN    3042.0
4                           121.1           18.0     218.0

[5 rows x 38 columns]
```

2. Data Cleaning

```
# Check for missing values
missing_values = df.isnull().sum()
print( missing_values)
```

```
gdp                                   1
sex_ratio                             0
surface_area                          1
life_expectancy_male                  6
unemployment                         10
imports                               5
homicide_rate                        23
currency                              0
iso2                                  1
employment_services                  11
employment_industry                  11
urban_population_growth               0
secondary_school_enrollment_female   14
employment_agriculture               11
capital                               0
forested_area                         4
exports                               5
life_expectancy_female                6
post_secondary_enrollment_female     33
post_secondary_enrollment_male       33
primary_school_enrollment_female      8
infant_mortality                      8
gdp_growth                            1
threatened_species                    0
population                            0
urban_population                      0
secondary_school_enrollment_male     14
name                                  0
pop_growth                            0
region                                0
pop_density                           0
internet_users                        2
gdp_per_capita                        1
fertility                             5
refugees                             14
primary_school_enrollment_male        8
co2_emissions                        59
tourists                             10
dtype: int64
```

```
#drop missing values
data_cleaned = df.dropna()
```

```
data_cleaned = df.dropna()

# Check for missing values after dropping
missing_values_after = data_cleaned.isnull().sum()
print(missing_values_after)
```

```
gdp                                 0
sex_ratio                           0
surface_area                        0
life_expectancy_male                0
unemployment                        0
imports                             0
homicide_rate                       0
currency                            0
iso2                                0
employment_services                 0
employment_industry                 0
urban_population_growth             0
secondary_school_enrollment_female  0
employment_agriculture              0
capital                             0
forested_area                       0
exports                             0
life_expectancy_female              0
post_secondary_enrollment_female    0
post_secondary_enrollment_male      0
primary_school_enrollment_female    0
infant_mortality                    0
gdp_growth                          0
threatened_species                  0
population                          0
urban_population                    0
secondary_school_enrollment_male    0
name                                0
pop_growth                          0
region                              0
pop_density                         0
internet_users                      0
gdp_per_capita                      0
fertility                           0
refugees                            0
primary_school_enrollment_male      0
co2_emissions                       0
tourists                            0
dtype: int64
```

```
# # Display the head of cleaned data
print(data_cleaned.head())
```

```
        gdp  sex_ratio  surface_area  life_expectancy_male  unemployment  \
1   15059.0      103.7       28748.0                  76.7          12.8
2  173757.0      102.1     2381741.0                  75.4          11.5
4  105902.0       97.9     1246700.0                  57.8           6.8
6  518475.0       95.3     2780400.0                  73.0          10.4
7   12433.0       88.8       29743.0                  71.1          16.6

   imports  homicide_rate                                  currency iso2  \
1   5908.0            2.3             {'code': 'ALL', 'name': 'Lek'}   AL
2  45140.0            1.4  {'code': 'DZD', 'name': 'Algerian Dinar'}   DZ
4  21340.0            4.8          {'code': 'AOA', 'name': 'Kwanza'}   AO
6  49125.0            5.3  {'code': 'ARS', 'name': 'Argentine Peso'}   AR
7   5053.0            1.7   {'code': 'AMD', 'name': 'Armenian Dram'}   AM

   employment_services  ...  pop_growth           region  pop_density  \
1                 43.7  ...        -0.1  Southern Europe        105.0
2                 59.6  ...         2.0  Northern Africa         18.4
4                 41.7  ...         3.3    Middle Africa         26.4
6                 78.9  ...         1.0    South America         16.5
7                 53.6  ...         0.3     Western Asia        104.1

   internet_users gdp_per_capita  fertility  refugees  \
1            71.8         5223.8        1.6       4.3
2            49.0         4114.7        3.0      99.5
4            14.3         3437.3        5.6      70.1
6            74.3        11687.6        2.3     165.6
7            64.7         4212.1        1.8      19.0

   primary_school_enrollment_male  co2_emissions  tourists
1                           105.2            4.3    5340.0
2                           112.4          130.5    2657.0
4                           121.1           18.0     218.0
6                           109.9          183.4    6942.0
```

```
7                            92.7           5.2    1652.0

[5 rows x 38 columns]
```

Data Profiling

```
# 1. Understanding Data Types
data_types = data_cleaned.dtypes
print(data_types)
```

```
gdp                                  float64
sex_ratio                            float64
surface_area                         float64
life_expectancy_male                 float64
unemployment                         float64
imports                              float64
homicide_rate                        float64
currency                              object
iso2                                  object
employment_services                  float64
employment_industry                  float64
urban_population_growth              float64
secondary_school_enrollment_female   float64
employment_agriculture               float64
capital                               object
forested_area                        float64
exports                              float64
life_expectancy_female               float64
post_secondary_enrollment_female     float64
post_secondary_enrollment_male       float64
primary_school_enrollment_female     float64
infant_mortality                     float64
gdp_growth                           float64
threatened_species                     int64
population                             int64
urban_population                     float64
secondary_school_enrollment_male     float64
name                                  object
pop_growth                           float64
region                                object
pop_density                          float64
internet_users                       float64
gdp_per_capita                       float64
fertility                            float64
refugees                             float64
primary_school_enrollment_male       float64
co2_emissions                        float64
tourists                             float64
dtype: object
```

```
# 2. Summary Statistics
summary_statistics = data_cleaned.describe(include='all')
print(summary_statistics)
```

```
                 gdp   sex_ratio  surface_area  life_expectancy_male  \
count   1.210000e+02  121.000000  1.210000e+02            121.000000
unique           NaN         NaN           NaN                   NaN
top              NaN         NaN           NaN                   NaN
freq             NaN         NaN           NaN                   NaN
mean    6.810738e+05  101.552066  1.051827e+06             72.363636
std     2.294058e+06   23.192937  2.453332e+06              6.413787
min     5.507000e+03   84.500000  3.150000e+02             53.300000
25%     3.787600e+04   95.300000  6.561000e+04             68.400000
50%     1.059560e+05   98.500000  2.383910e+05             73.300000
75%     3.826740e+05  100.600000  7.960950e+05             77.400000
max     2.058022e+07  302.400000  1.709825e+07             81.600000

        unemployment       imports  homicide_rate  \
count     121.000000  1.210000e+02     121.000000
unique           NaN           NaN            NaN
top              NaN           NaN            NaN
freq             NaN           NaN            NaN
```

```
mean        6.696694  1.446775e+05       5.561157
std         4.562326  3.391843e+05       9.526239
min         0.100000  9.800000e+02       0.200000
25%         3.800000  9.470000e+03       1.100000
50%         5.300000  3.137200e+04       2.100000
75%         8.900000  1.165560e+05       5.000000
max        28.500000  2.567490e+06      52.000000

                               currency iso2  employment_services  ...  \
count                               121  121           121.000000  ...
unique                               90  113                  NaN  ...
top     {'code': 'EUR', 'name': 'Euro'}   AL                  NaN  ...
freq                                 22    2                  NaN  ...
mean                                NaN  NaN            60.328926  ...
std                                 NaN  NaN            15.731706  ...
min                                 NaN  NaN            18.000000  ...
25%                                 NaN  NaN            49.400000  ...
50%                                 NaN  NaN            62.000000  ...
75%                                 NaN  NaN            72.900000  ...
max                                 NaN  NaN            87.500000  ...

        pop_growth        region  pop_density  internet_users gdp_per_capita  \
count   121.000000           121   121.000000      121.000000     121.000000
unique         NaN            19          NaN             NaN            NaN
top            NaN  Western Asia          NaN             NaN            NaN
freq           NaN            15          NaN             NaN            NaN
mean      1.045455           NaN   215.838843       64.596694   18721.280165
std       1.093542           NaN   793.099110       24.555301   22330.468596
min      -1.500000           NaN     2.100000        5.300000     498.900000
25%       0.200000           NaN    36.200000       47.200000    3449.600000
50%       1.000000           NaN    88.500000       71.800000    8258.200000
75%       1.700000           NaN   136.500000       82.300000   26005.100000
max       4.300000           NaN  8357.600000       99.700000  117369.500000

         fertility     refugees  primary_school_enrollment_male  \
count   121.000000   121.000000                      121.000000
unique         NaN          NaN                             NaN
top            NaN          NaN                             NaN
freq           NaN          NaN                             NaN
```

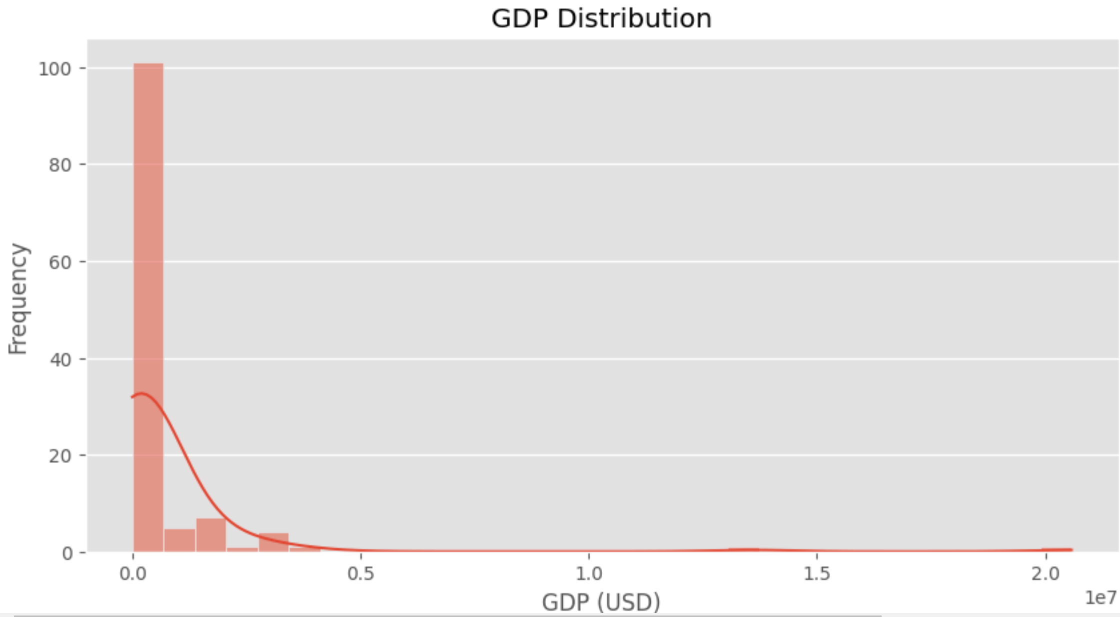
Data Visualization

1. Histogram for **GDP**

```
#use ggplot
plt.style.use('ggplot')
plt.figure(figsize=(10, 5))
#histogram for the plot
sns.histplot(data_cleaned['gdp'], bins=30, kde=True)

# Titles and labels
plt.title('GDP Distribution')
plt.xlabel('GDP (USD)')
plt.ylabel('Frequency')

plt.grid(axis='x')
plt.show()
```

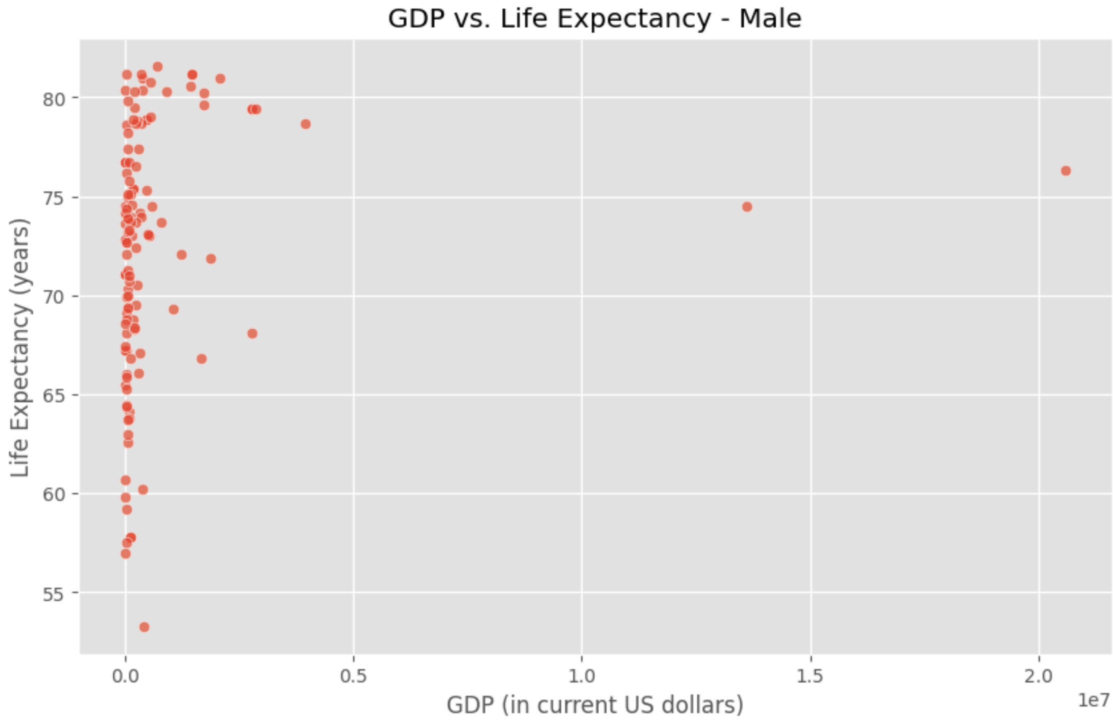


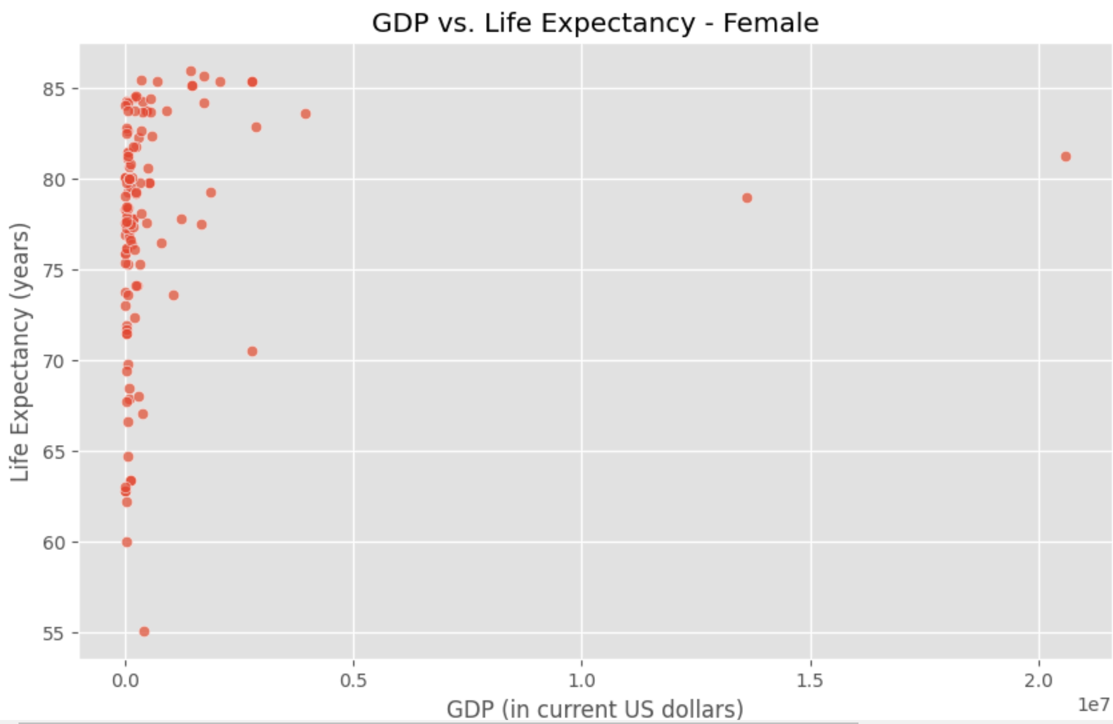
The right-skewed histogram for GDP shows that most countries in the dataset have low to moderate GDPs, meaning they are mostly developing countries facing challenges like limited resources and high poverty rates. In contrast, a few countries, like the United States and China, have very high GDPs, which raises the average GDP. It highlights the global economic inequality.

2. Scatter Plot for GDP vs Life Expectancy for male and female

```
plt.style.use('ggplot')
plt.figure(figsize=(10, 6))
sns.scatterplot(x='gdp', y='life_expectancy_male', data=data_cleaned, alpha=0.7)
plt.title('GDP vs. Life Expectancy - Male')
plt.xlabel('GDP (in current US dollars)')
plt.ylabel('Life Expectancy (years)')
plt.grid(True)
plt.show()

plt.style.use('ggplot')
plt.figure(figsize=(10, 6))
sns.scatterplot(x='gdp', y='life_expectancy_female', data=data_cleaned, alpha=0.7)
plt.title('GDP vs. Life Expectancy - Female')
plt.xlabel('GDP (in current US dollars)')
plt.ylabel('Life Expectancy (years)')
plt.grid(True)
plt.show()
```



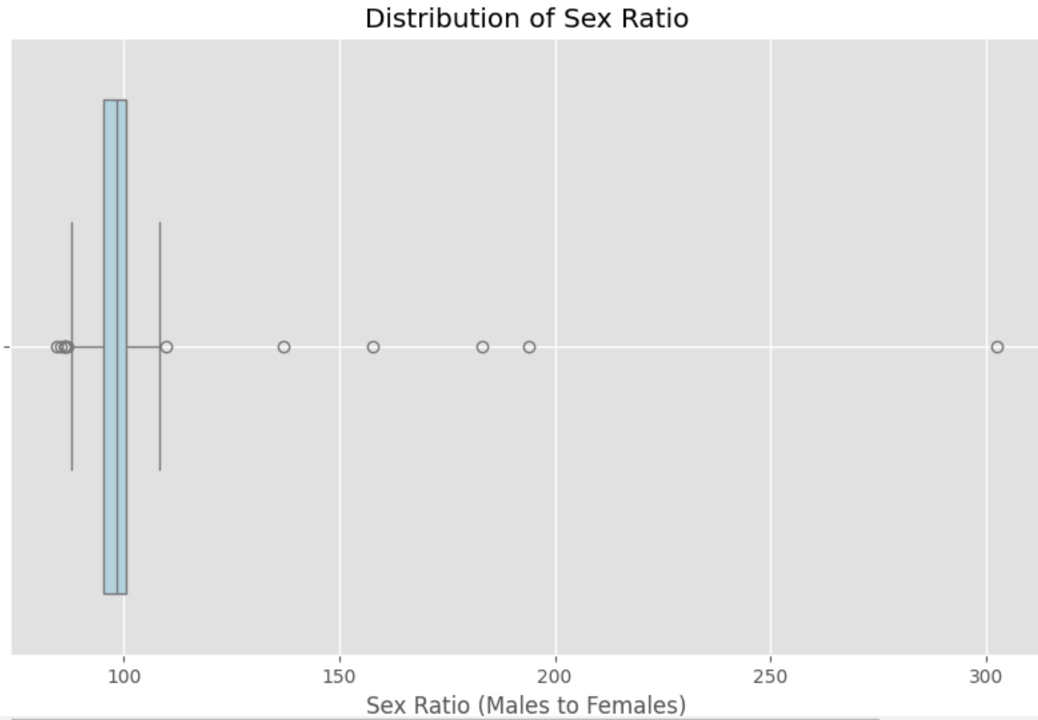


The scatter plot of GDP and both male and female life expectancy shows that most countries are grouped on the left side, which means they have low GDPs and may be facing economic difficulties. This can lead to lower life expectancy rates. A few countries on the right side have much higher GDPs and tend to have better healthcare and living standards, resulting in longer life expectancies. Overall, while there is a trend showing that higher GDP usually means better life expectancy, the clustering of countries on the left highlights the economic challenges many face and emphasizes the need for economic growth to improve healthcare and life quality.

3. Sex Ratio Distribution

```
plt.style.use('ggplot')
plt.figure(figsize=(10, 6))
sns.boxplot(x='sex_ratio', data=data_cleaned, color='lightblue')
plt.title('Distribution of Sex Ratio')
plt.xlabel('Sex Ratio (Males to Females)')
```

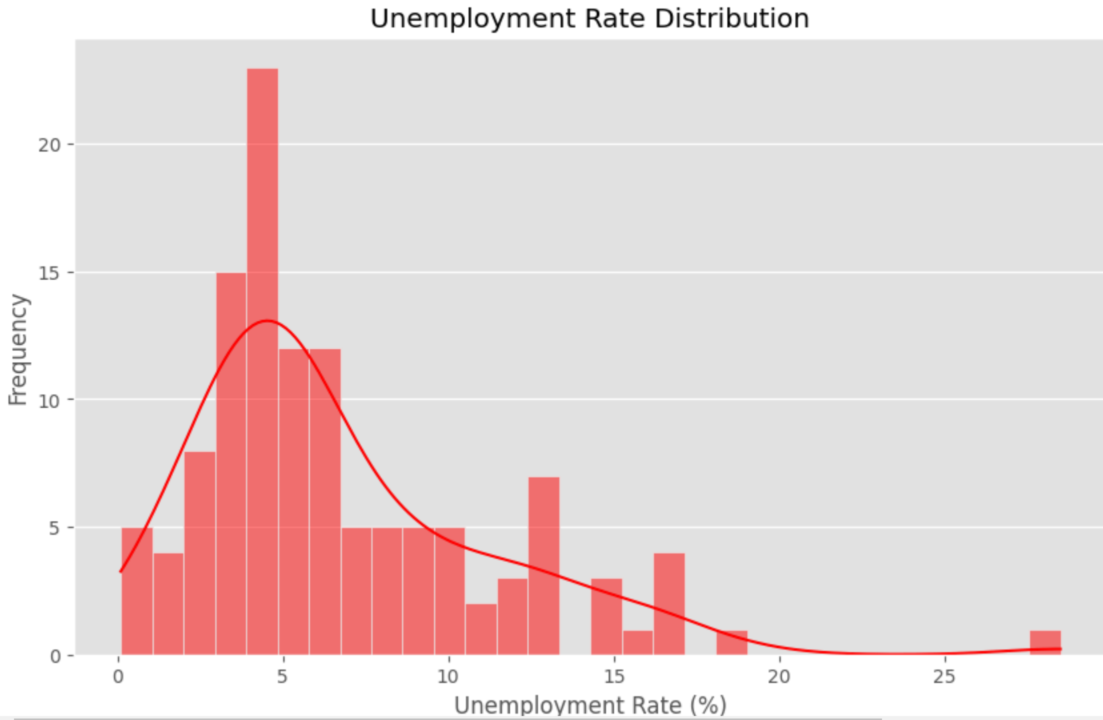
```
plt.grid('True')
plt.show()
```



A right-skewed box-plot indicates that there are usually more females than males or a balanced ratio in many populations. Many sex ratios might be around 90 males for every 100 females, showing a female-heavy population.

4. Unemployment Rate

```
plt.style.use('ggplot')
plt.figure(figsize=(10, 6))
sns.histplot(data_cleaned['unemployment'], bins=30, kde=True, color='red')
plt.title('Unemployment Rate Distribution')
plt.xlabel('Unemployment Rate (%)')
plt.ylabel('Frequency')
plt.grid(axis='x')
plt.show()
```

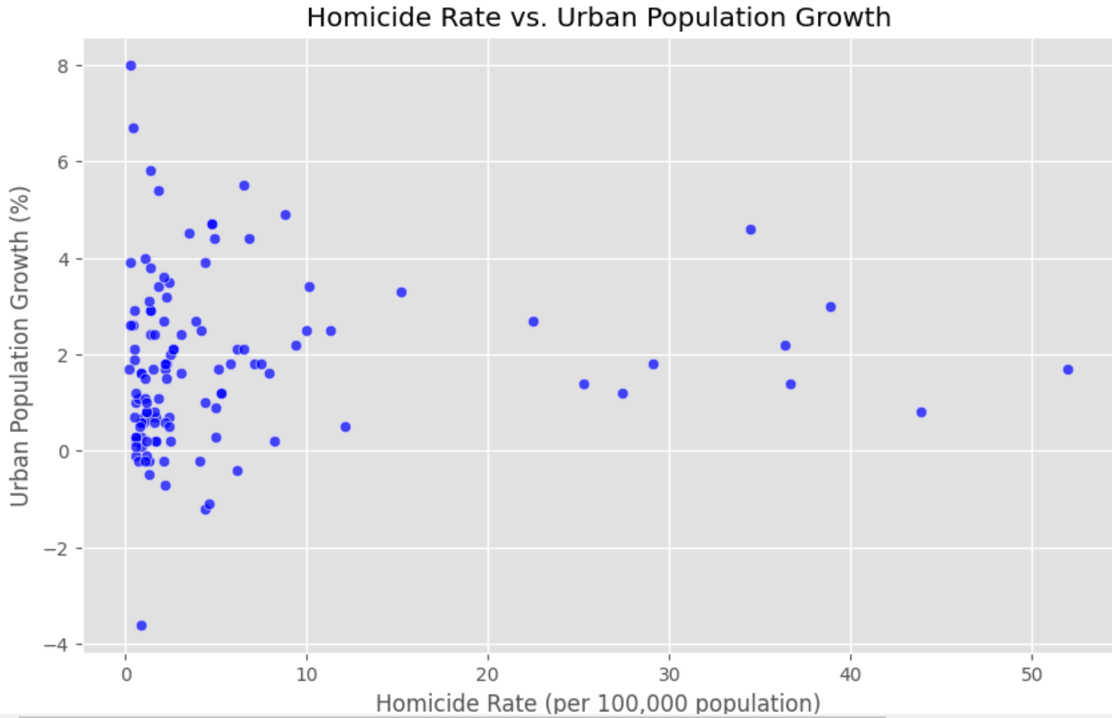


The histogram shows that most countries have low unemployment rates, with many clustering around 4%. This indicates that many countries experience stable job markets. However, there are also outliers with much higher unemployment rates, suggesting economic difficulties in those specific areas.

5. Homicide Rate vs Urban Population Growth

```
plt.style.use('ggplot')
plt.figure(figsize=(10, 6))
sns.scatterplot(x='homicide_rate', y='urban_population_growth', data=data_cleaned, palette='viridis', alpha=0.7, color='blue')
plt.title('Homicide Rate vs. Urban Population Growth')
plt.xlabel('Homicide Rate (per 100,000 population)')
plt.ylabel('Urban Population Growth (%)')
plt.grid(True)
plt.show()
```

```
<ipython-input-73-621ac18c47cd>:3: UserWarning: Ignoring `palette` because no `hue` variable has been assigned.
  sns.scatterplot(x='homicide_rate', y='urban_population_growth', data=data_cleaned, palette='viridis', alpha=0.7, color='blue')
```



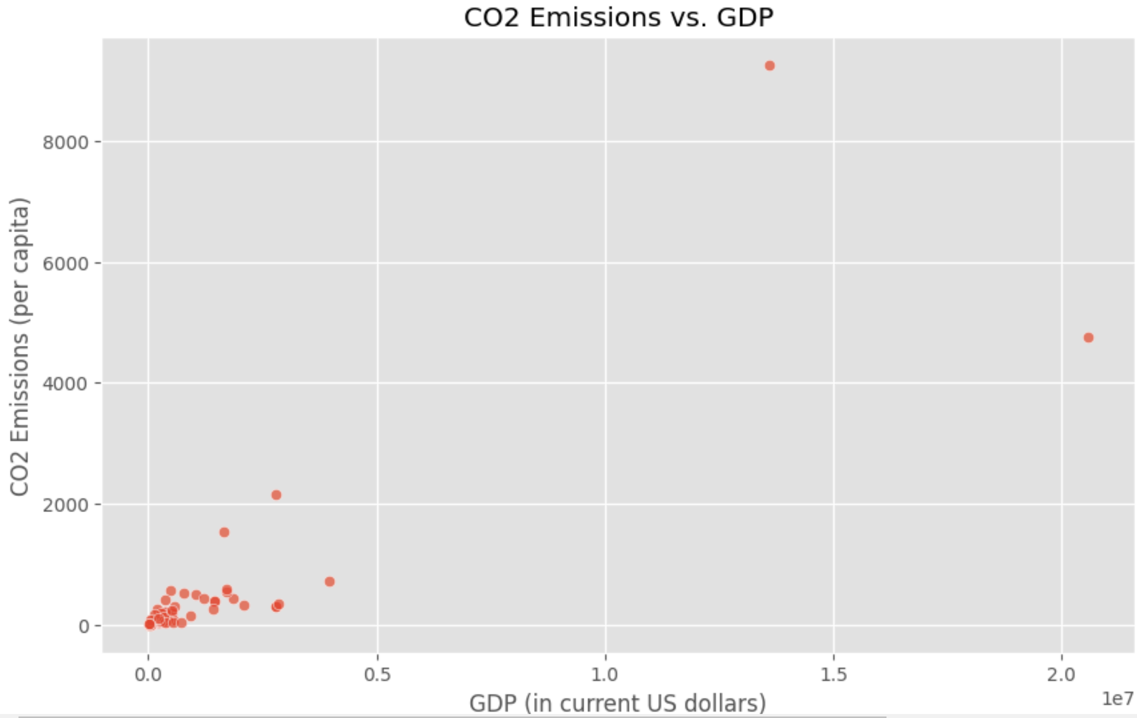
The scatter plot of homicide rates versus urban population growth indicates a positive correlation, with most countries exhibiting low homicide rates and corresponding urban growth. This suggests that safer environments are linked to healthier urban development. On the right side, however, where homicide rates are higher, there seems to be little to no clear relationship with urban population growth. This implies that in regions with increased violence, the urban growth rates vary widely and do not follow a consistent trend.

Start coding or generate with AI.

6. CO2 Emissions vs GDP

```
plt.style.use('ggplot')
plt.figure(figsize=(10, 6))
sns.scatterplot(x='gdp', y='co2_emissions', data=data_cleaned, palette='viridis', alpha=0.7)
plt.title('CO2 Emissions vs. GDP')
plt.xlabel('GDP (in current US dollars)')
plt.ylabel('CO2 Emissions (per capita)')
plt.grid(True)
plt.show()
```

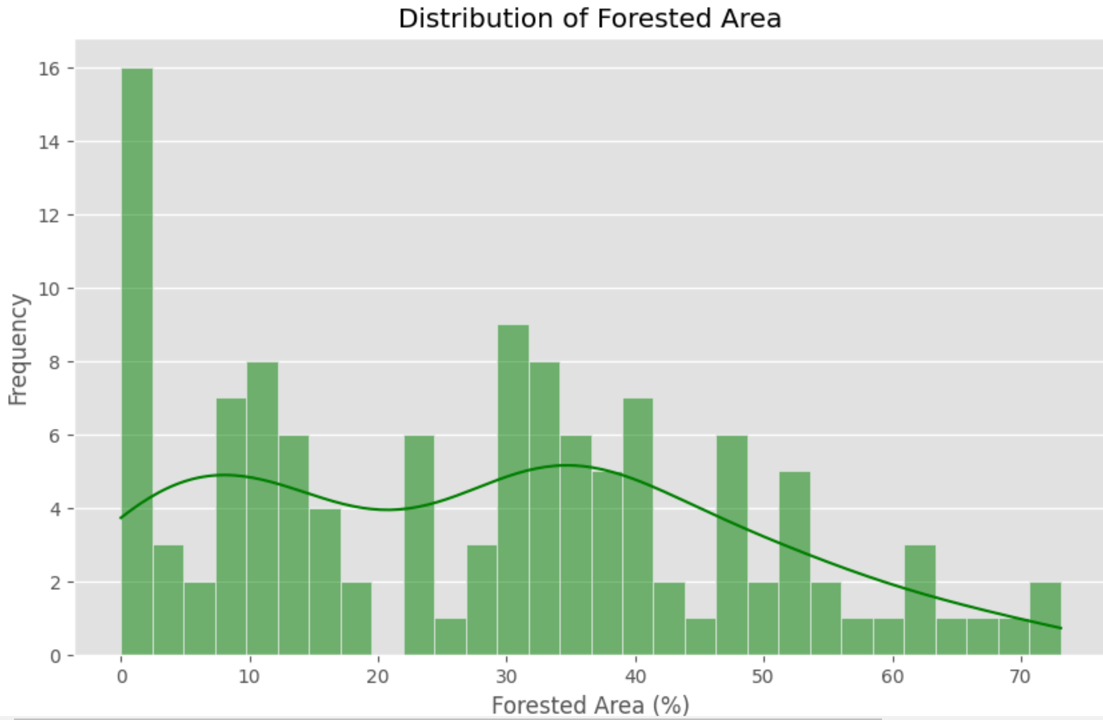
```
<ipython-input-76-fde6b6df64a1>:3: UserWarning: Ignoring `palette` because no `hue` variable has been assigned.
  sns.scatterplot(x='gdp', y='co2_emissions', data=data_cleaned, palette='viridis', alpha=0.7)
```



The scatter plot of CO2 emissions versus GDP shows a positive correlation, meaning that higher GDP generally corresponds to higher per capita CO2 emissions. This suggests that wealthier countries tend to have more economic activity, leading to increased emissions. However, some outliers indicate countries with high CO2 emissions relative to their GDP, possibly due to factors like heavy fossil fuel use or specific industrial practices. Overall, while there is a trend linking economic growth to emissions, these outliers point to the influence of other factors on CO2 emissions.

7. Forested Area

```
plt.style.use('ggplot')
plt.figure(figsize=(10, 6))
sns.histplot(data_cleaned['forested_area'], bins=30, kde=True, color = 'green')
plt.title('Distribution of Forested Area')
plt.xlabel('Forested Area (%)')
plt.ylabel('Frequency')
plt.grid(axis='x')
plt.show()
```

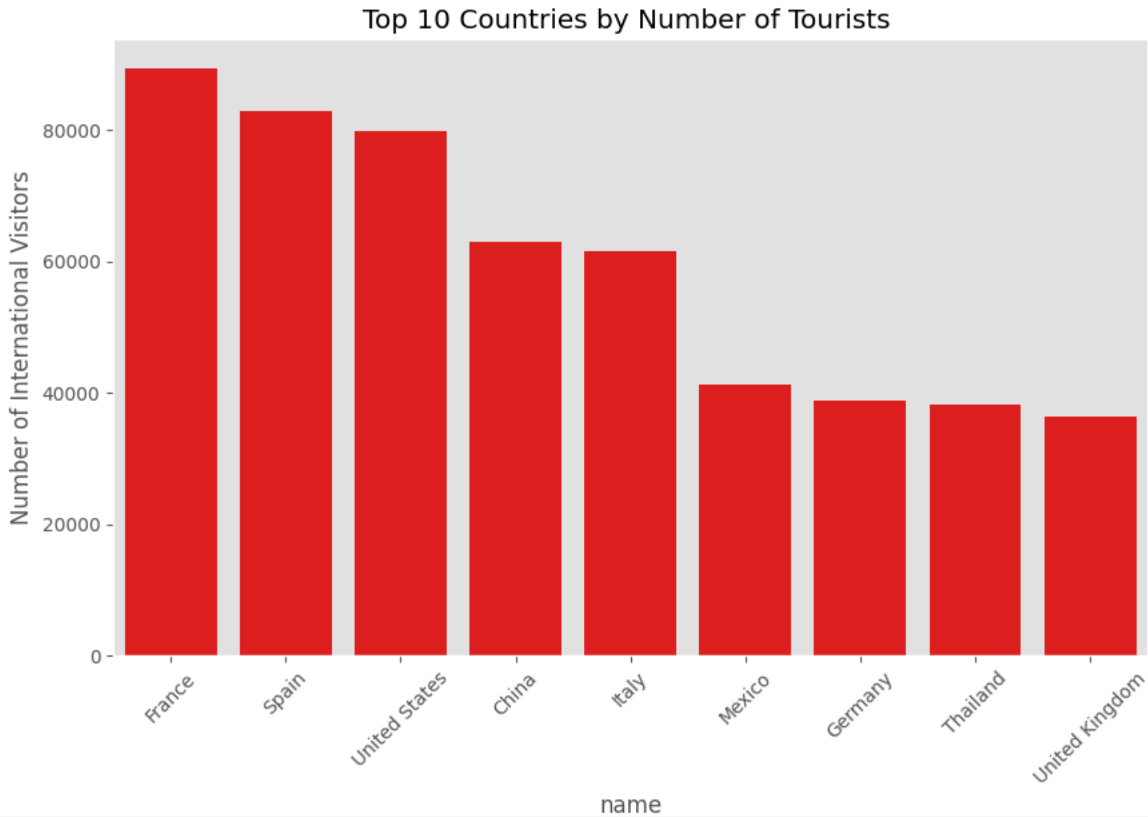


The histogram of forested area displays two bell-shaped curves, indicating two distinct groups of countries regarding forest cover. Most countries have low forested areas, likely due to urbanization, agriculture, or deforestation, while fewer countries show higher percentages of forest cover.

8. Tourist Numbers

```
plt.figure(figsize=(10, 6))
plt.figure(figsize=(10, 6))
sns.barplot(x='name', y='tourists', data=data_cleaned.sort_values('tourists', ascending=False).head(10), color = 'red')
plt.title('Top 10 Countries by Number of Tourists ')
plt.xticks(rotation=45)
plt.ylabel('Number of International Visitors')
plt.grid(axis='y')
plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```



The bar plot displaying the top 10 countries by the number of international tourists reveals that France, Spain, the United States lead in most attracting visitors.

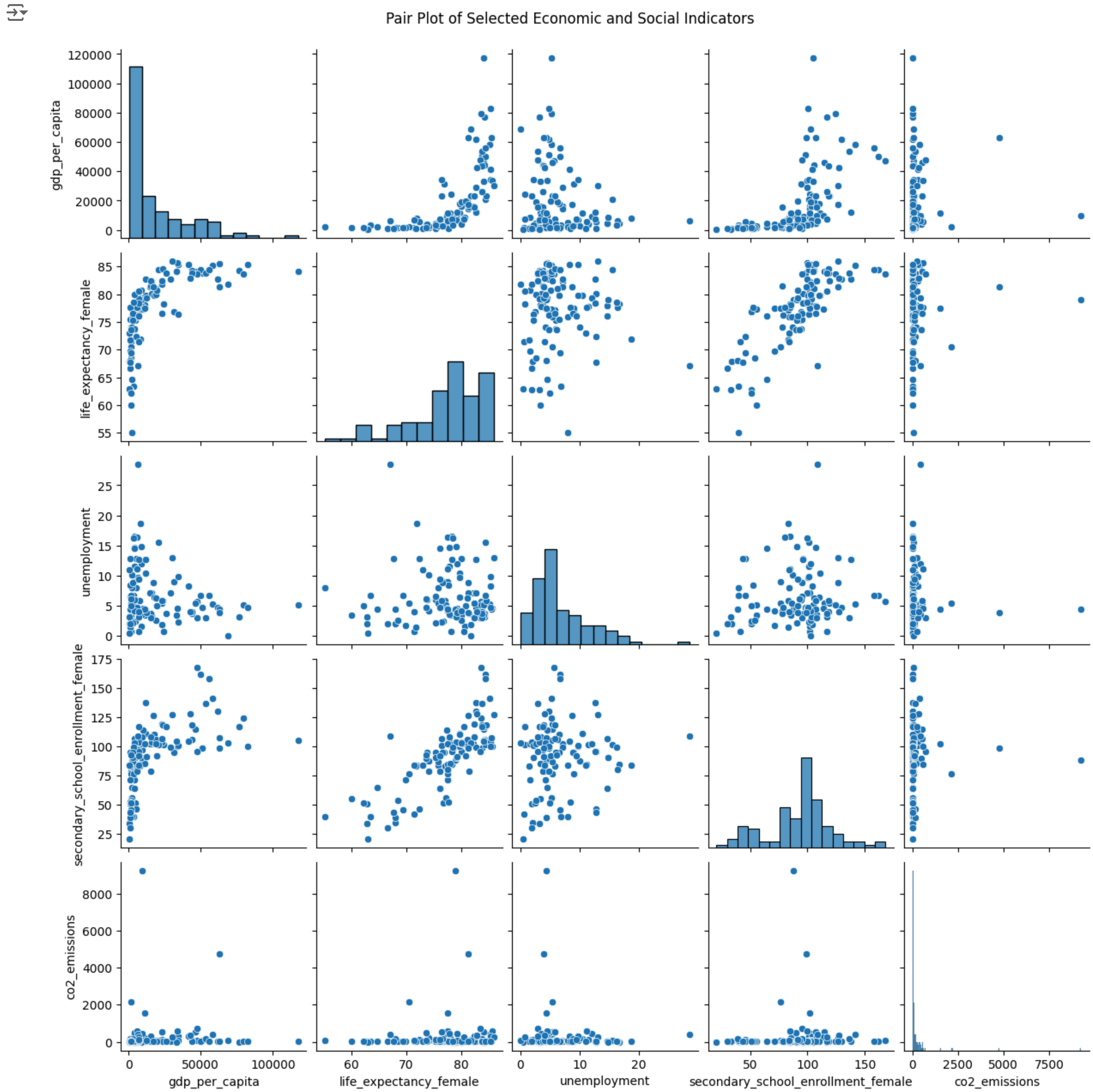
9. Pair plot

```
# Select suitable variables for the pair plot
variables = ['gdp_per_capita', 'life_expectancy_female', 'unemployment', 'secondary_school_enrollment_female', 'co2_emissions']

# Create a pair plot using the selected variables
pair_plot = sns.pairplot(data_cleaned, vars=variables)

#title
pair_plot.fig.suptitle("Pair Plot of Selected Economic and Social Indicators", y=1.02)

plt.show()
```



The pair plot shows several relationships between economic and social factors. Countries with higher GDP per capita tend to have longer life expectancy for women and higher rates of girls in secondary school. This likely means that richer countries have better healthcare, living conditions, and education. There is also a slight link between high GDP and high CO2 emissions, which may be due to more industry and energy use. Unemployment doesn't seem to affect life expectancy for women in a clear way. However, there is a positive link between girls' school enrollment and life expectancy. It suggests that countries where more girls go to school also provide better healthcare.

10. Correlation Heat Map

```
numeric_df = data_cleaned.select_dtypes(include=[np.number])
plt.figure(figsize=(15, 10))
sns.heatmap(numeric_df.corr(), annot=True, fmt='.2f', cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```