UNIFIED MENTOR

UNID: UMIP278557

Name: Tarushi Umapathi

Project 1:

Title : "Analyzing Factors Affecting OCD Severity and Medication Trends"(Intermediate)

Project 2:

Title: “Life Expectancy Analysis”

Analyzing Factors Affecting OCD Severity and Medication Trends

INTRODUCTION:

Obsessive-Compulsive Disorder (OCD) is a chronic mental health condition characterized by recurring obsessions and compulsions. The dataset contains demographic, diagnostic, and treatment-related information about OCD patients. The goal is to analyze this dataset to understand the factors affecting OCD severity, identify patterns in medication usage, and address missing information to improve data completeness for better decision-making.

Objective:

To identify patterns, correlations, and potential predictors of OCD severity using patient demographics, symptoms, and treatment data. Additionally, explore how medication use varies across different patient groups.

Problem Statement:

What factors influence OCD severity (measured by Y-BOCS scores)?

Are there any patterns in medication usage based on demographics, symptoms, or OCD severity?

CODE:

Step 1: Import Libraries and read the data

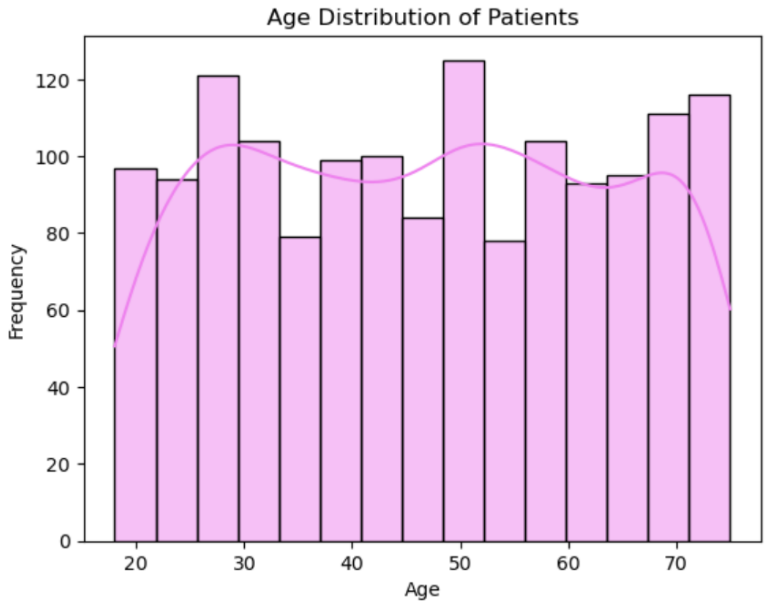
```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

data = pd.read_csv("OCD_Patient_Dataset.csv")

data

data.columns

data.shape

data.describe()

data.info()
```

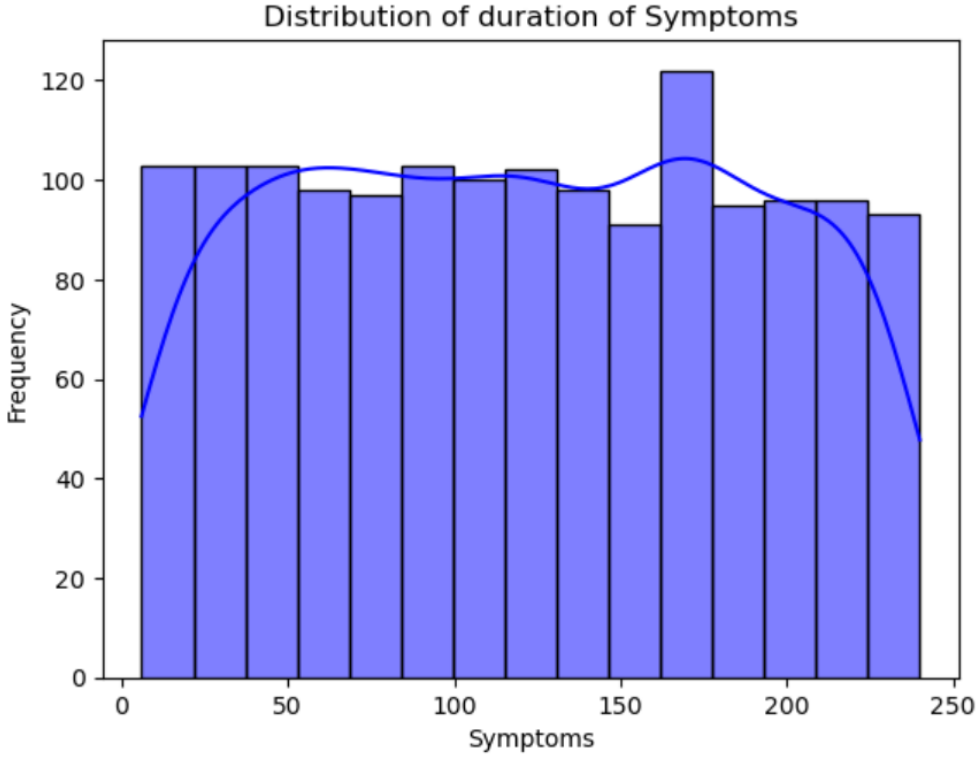
Step2: Handling missing values

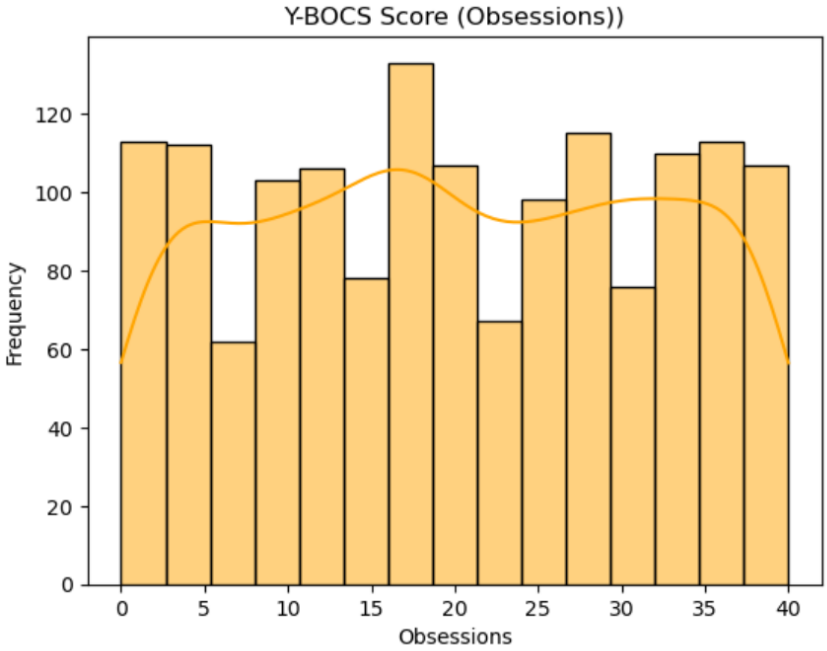
```
data.isnull().sum()
```

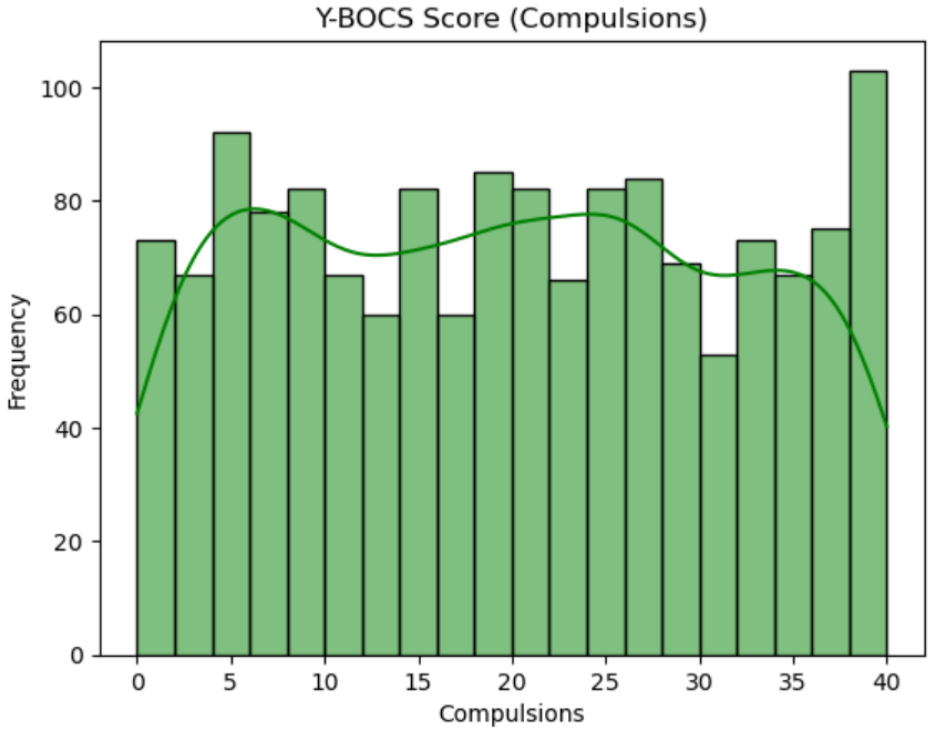
```
# Separate rows with missing values in specific columns
missing_values_rows = data[data[['Previous Diagnoses', 'Medications']].isna().any(axis=1)]
non_missing_values_rows = data.dropna(subset=['Previous Diagnoses', 'Medications'])

# Save missing rows separately (if needed)
missing_values_rows.to_csv('missing_values_rows.csv', index=False)
non_missing_values_rows.to_csv('non_missing_values_rows.csv', index=False)

# Display the counts
print(f"Rows with missing values: {missing_values_rows.shape[0]}")
print(f"Rows without missing values: {non_missing_values_rows.shape[0]}")
#Handling missing values
#The dataset dosent  seem much like a pattern , let us take it as unknown
data['Previous Diagnoses'].fillna('Unknown', inplace=True)
data['Medications'].fillna('Unknown', inplace = True)

data.head()
Step 3: Histogram
sns.histplot(data['Age'], bins=15, kde=True, color = 'violet')
plt.title('Age Distribution of Patients')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
sns.histplot(data['Duration of Symptoms (months)'], bins=15, kde=True, color = 'blue')
plt.title('Distribution of duration of Symptoms')
plt.xlabel('Symptoms')
plt.ylabel('Frequency')
plt.show()
sns.histplot(data['Y-BOCS Score (Obsessions)'], bins=15, kde=True, color = 'orange')
plt.title('Y-BOCS Score (Obsessions))')
plt.xlabel('Obsessions')
```

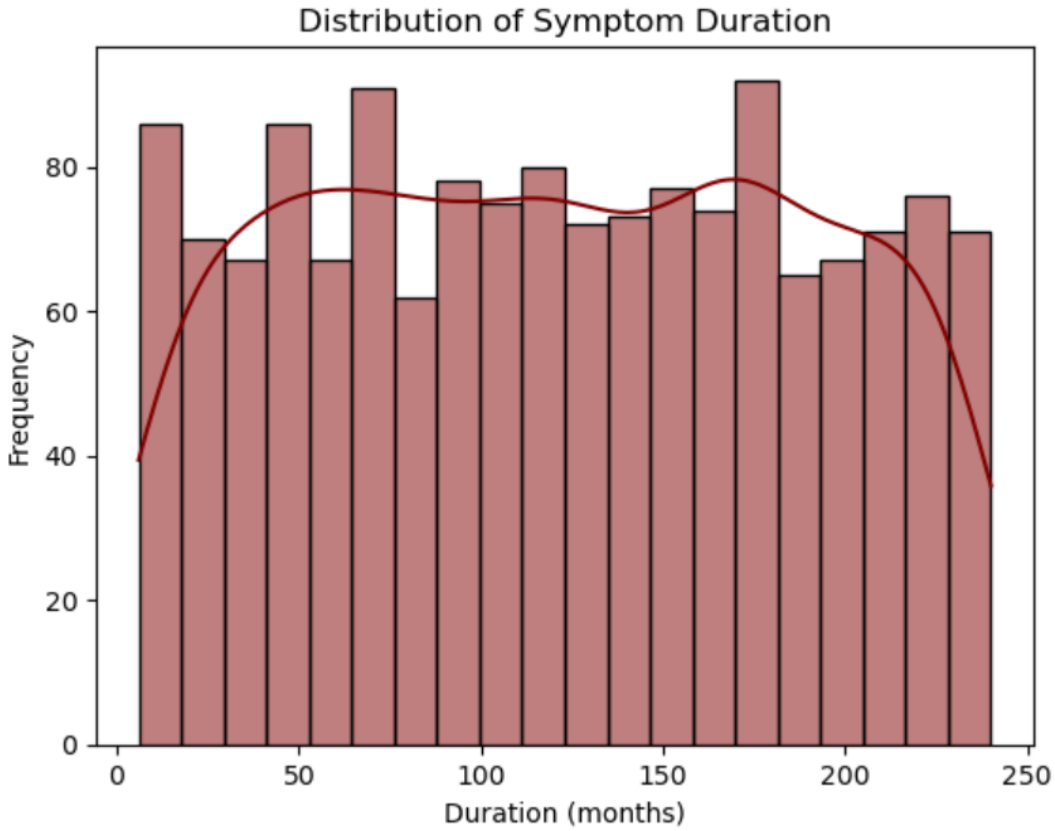
```
plt.ylabel('Frequency')
plt.show()
sns.histplot(data['Y-BOCS Score (Compulsions)'], bins=20, kde=True, color = 'green')
plt.title('Y-BOCS Score (Compulsions)')
plt.xlabel('Compulsions')
plt.ylabel('Frequency')
plt.show()
filtered_data = data[data['Medications'].str.lower() != 'none']

# Plot histogram for the 'Age' column
filtered_data['Age'].hist(bins=15, edgecolor='black', color='skyblue')
plt.title('Age Distribution for Patients on Medications')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
#Gender distribution
sns.countplot(x='Gender', data=data)
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
#Distribution of symptom duration
sns.histplot(data['Duration of Symptoms (months)'], bins=20, kde=True, color = "maroon")
plt.title('Distribution of Symptom Duration')
plt.xlabel('Duration (months)')
plt.ylabel('Frequency')
plt.show()
```

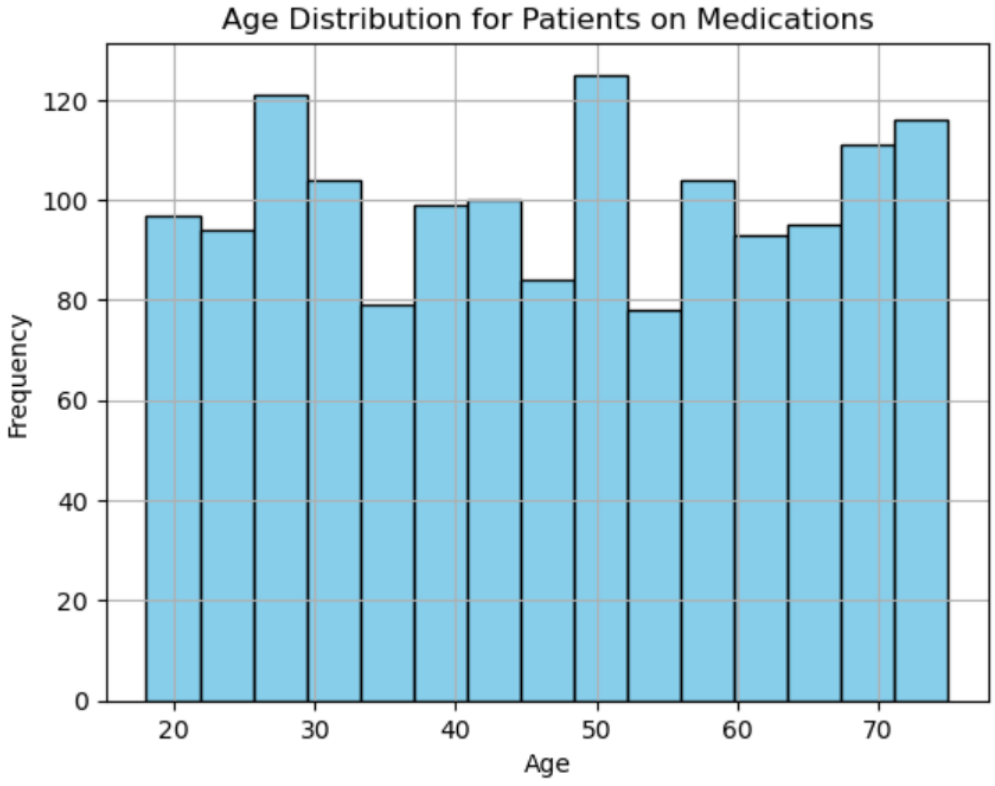

Age Distribution of Patients
Frequency
Age
0
20
40
60
80
100
120
20
30
40
50
60
70


Distribution of duration of Symptoms
Frequency
Symptoms
0
20
40
60
80
100
120
0
50
100
150
200
250


Y-BOCS Score (Obsessions))
Frequency
Obsessions
0
20
40
60
80
100
120
0
5
10
15
20
25
30
35
40


Y-BOCS Score (Compulsions)
Frequency
Compulsions
0
20
40
60
80
100
0
5
10
15
20
25
30
35
40

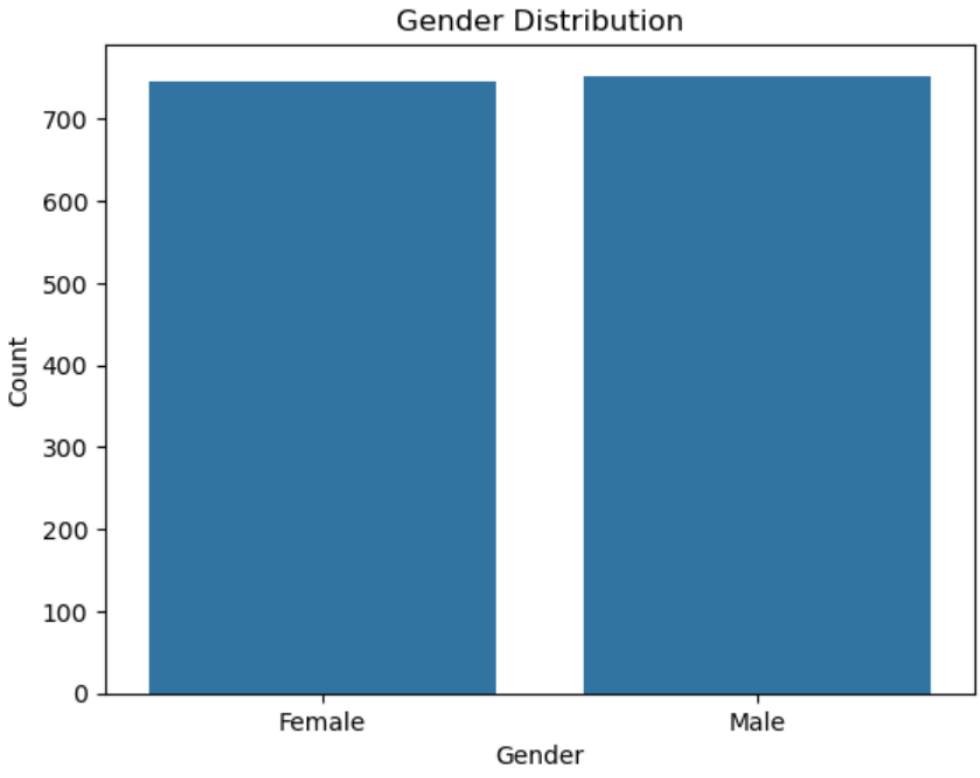


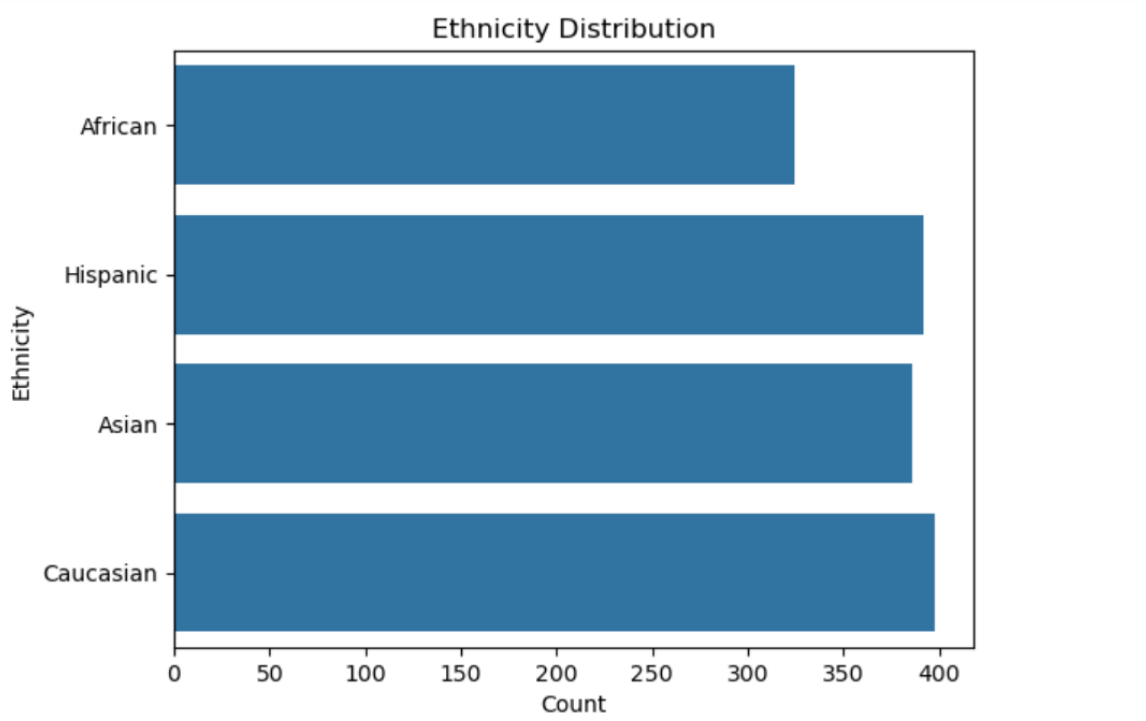


Observation: -- Age distribution for the patients who are on medications -- It is observed that that the highest are around age of 49 - 51 -- Next highest are around age of 27 - 29 -- Third highest are are 71 – 73

COUNTPLOT:

```
#Gender distribution
sns.countplot(x='Gender', data=data)
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
# Ethnicity distribution
sns.countplot(y='Ethnicity', data=data)
plt.title('Ethnicity Distribution')
plt.xlabel('Count')
plt.ylabel('Ethnicity')
plt.show()
```





BOXPLOT:

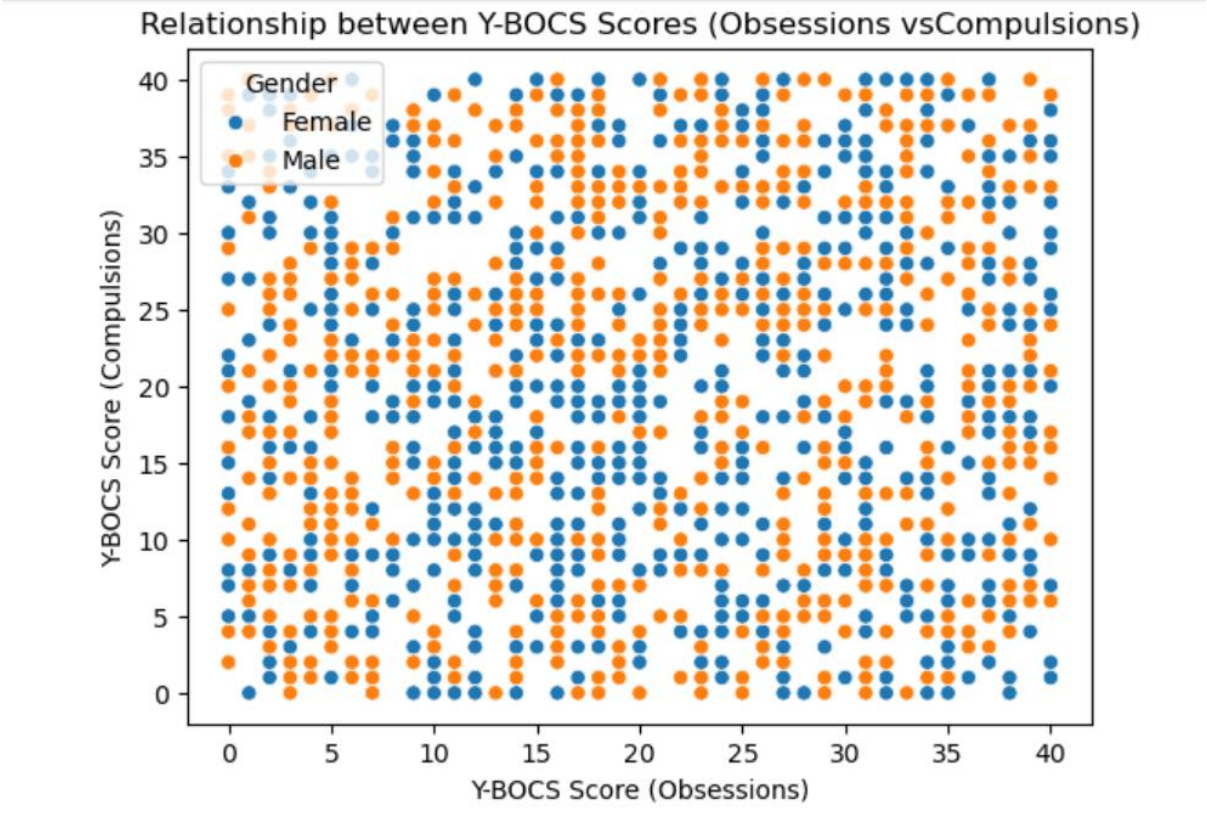
```
# Boxplot of Y-BOCS Scores by Gender
 sns.boxplot(x='Gender', y='Y-BOCS Score (Obsessions)', data=data)
 plt.title('Y-BOCS Obsession Scores by Gender')
 plt.xlabel('Gender')
 plt.ylabel('Y-BOCS Score (Obsessions)')
 plt.show()
```

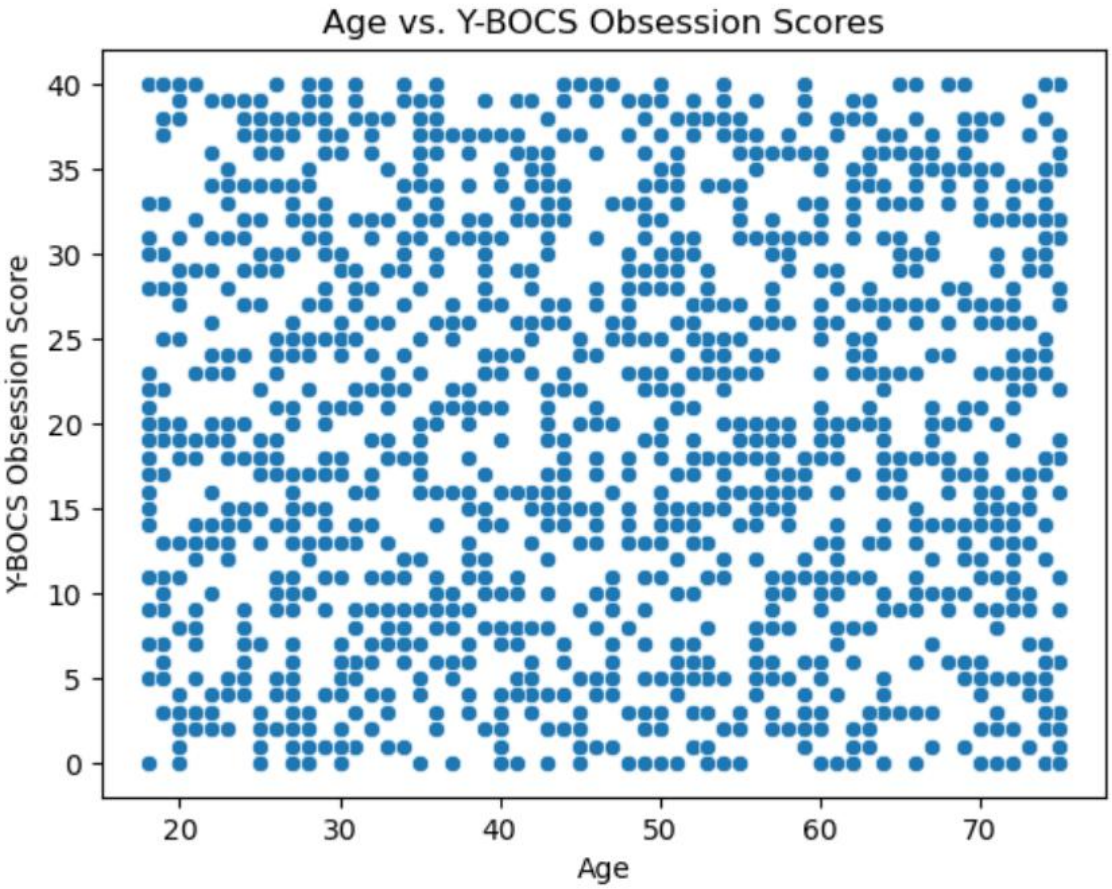
```
data['Severity Group'] = pd.cut(data['Y-BOCS Score (Obsessions)'] + data['Y-BOCS Score (Compulsions)'],
bins=[0, 14, 23, 40], labels=['Mild', 'Moderate', 'Severe'])
```

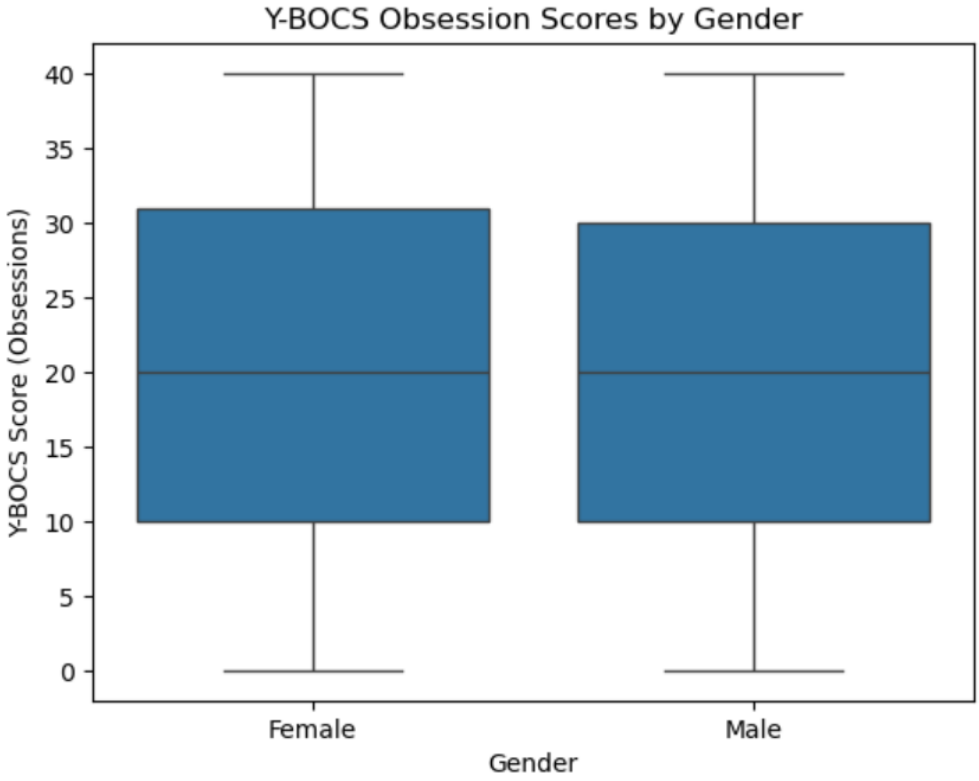
```
sns.boxplot(x='Severity Group', y='Age', data=data)
plt.title('Age Distribution Across OCD Severity Groups')
plt.xlabel('Severity Group')
plt.ylabel('Age')
plt.show()
```

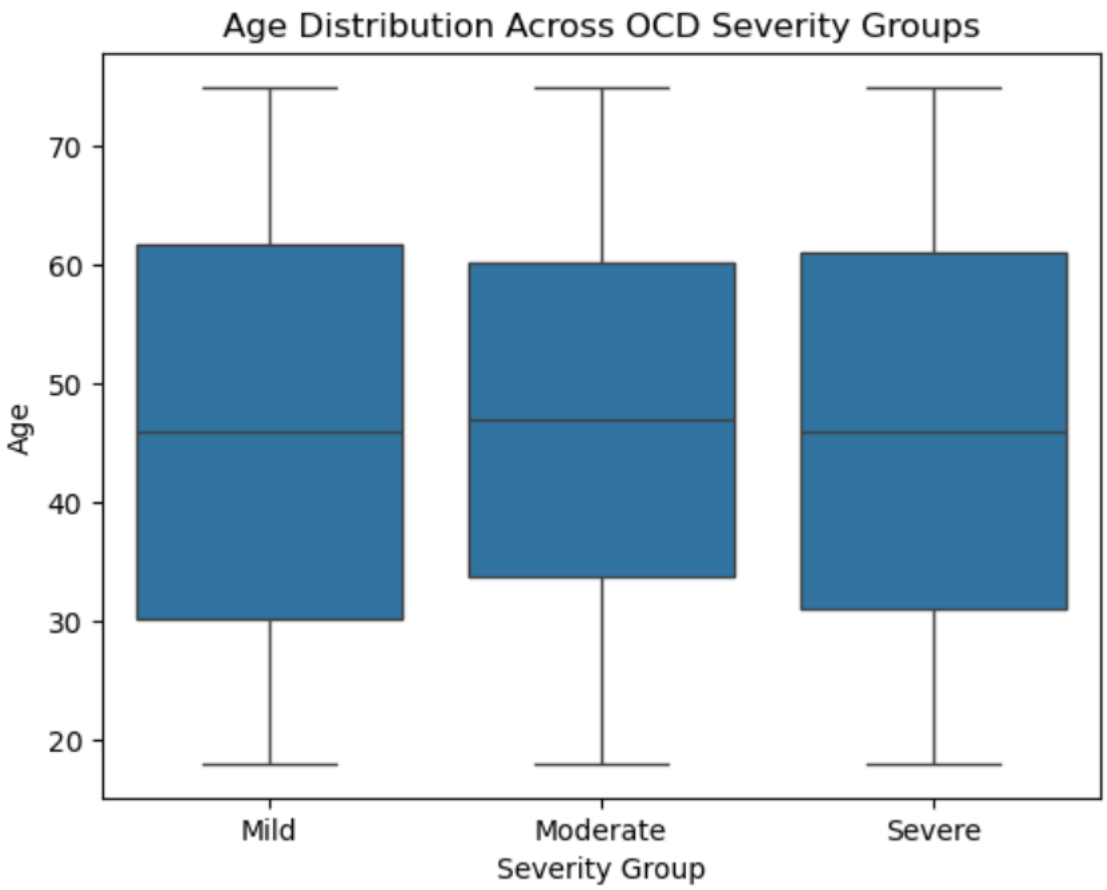
SCATTERPLOT:

```
# Relationship between Obsession and Compulsion Y-BOCS Scores

sns.scatterplot(x='Y-BOCS Score (Obsessions)', y='Y-BOCS Score (Compulsions)', hue='Gender', data=data)

plt.title('Relationship between Y-BOCS Scores (Obsessions vsCompulsions)')

plt.xlabel('Y-BOCS Score (Obsessions)')

plt.ylabel('Y-BOCS Score (Compulsions)')

plt.show()

sns.scatterplot(x='Age', y='Y-BOCS Score (Obsessions)', data=data)

plt.title('Age vs. Y-BOCS Obsession Scores')

plt.xlabel('Age')

plt.ylabel('Y-BOCS Obsession Score')

plt.show()
```



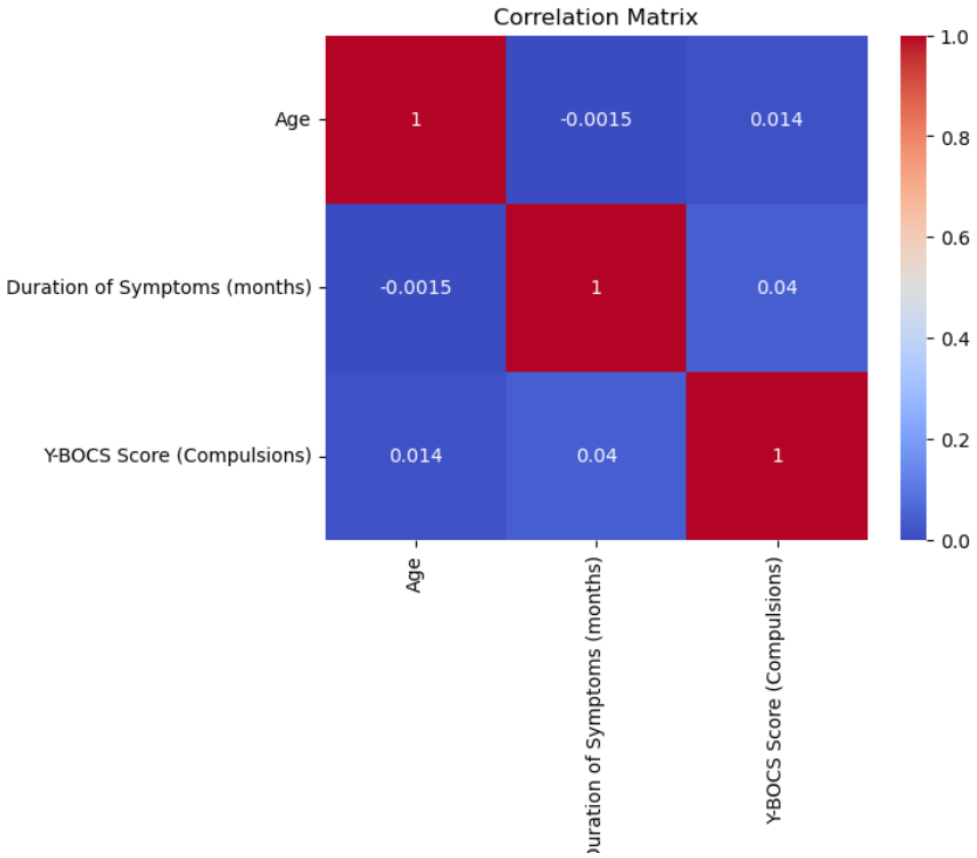

Age vs. Y-BOCS Obsession Scores
Y-BOCS Obsession Score
Age
0
5
10
15
20
25
30
35
40
20
30
40
50
60
70


Y-BOCS Obsession Scores by Gender
Y-BOCS Score (Obsessions)
Gender
Female
Male
0
5
10
15
20
25
30
35
40



HEATMAP:

```
corr_matrix = data[['Age', 'Duration of Symptoms (months)','Y-BOCS Score (Compulsions)']].corr()


sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```



Observation:

The scatterplot shows a trend where younger patients have higher Y-BOCS scores: Younger individuals may experience more severe OCD symptoms. This could indicate that early-life stressors or developmental factors might influence symptom severity.

Patients with a family history of OCD have higher median Y-BOCS scores compared to those without. A genetic predisposition or familial environment might play a role in symptom severity.

Symptom Type (Obsession/Compulsion) vs. Severity:

Patients with specific obsession types (e.g., contamination fears) have higher Y-BOCS scores. Interpretation: Certain obsession types might be harder to manage, leading to more severe symptoms.

Duration of Symptoms vs. OCD Severity:

A positive correlation between the duration of symptoms and Y-BOCS scores. Longer symptom duration may worsen severity due to a lack of treatment or ineffective management.

Conclusion:

By analyzing the data and interpreting these observations, we can provide actionable insights to stakeholders, whether they are healthcare providers, researchers, or patients.

Focus on younger patients with high Y-BOCS scores, as they may need early intervention.

Tailor medication based on symptom type and severity.

Investigate why certain obsession types lead to higher severity.

Explore if medication patterns differ due to external factors like availability or socioeconomic status.

Highlight the importance of reporting symptoms early and accurately to get appropriate treatment.

Life Expectancy Analysis:

Problem Statement

Life expectancy is a critical measure of a country's overall health and well-being. However, significant disparities exist across the globe, influenced by socio-economic, environmental, and healthcare-related factors. Understanding these disparities and the factors driving them is crucial for policymakers, healthcare providers, and researchers

Identify the key determinants of life expectancy.

Highlight global patterns and trends in life expectancy.

Provide actionable recommendations to improve health outcomes and reduce disparities.

Introduction

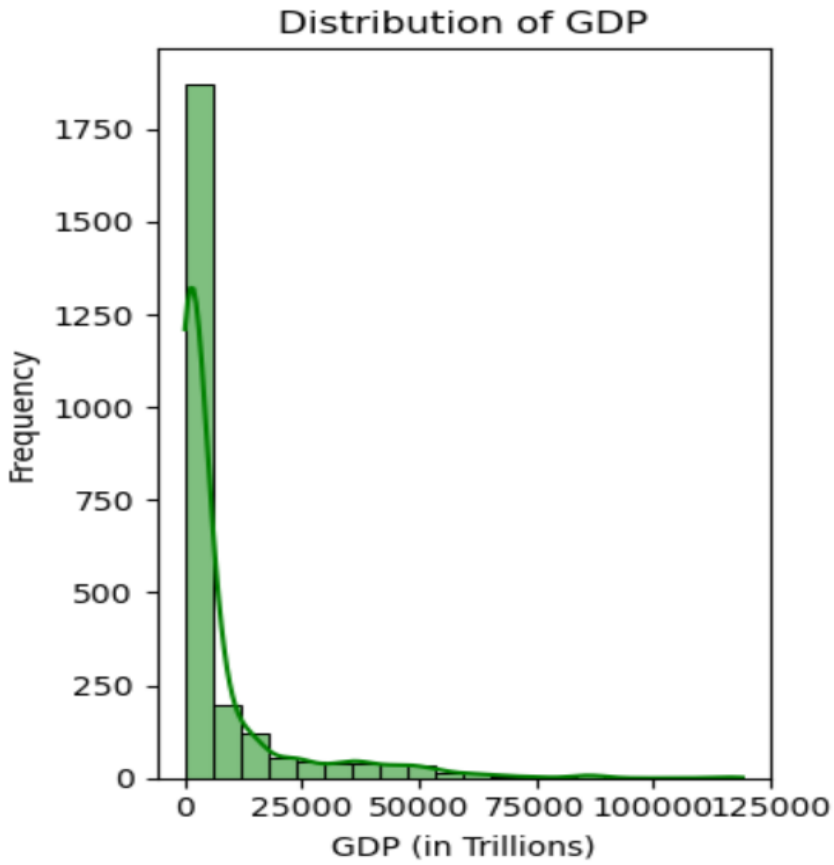
This report provides an analysis of the "Life Expectancy Data" dataset, focusing on key variables that impact life expectancy across countries. The analysis includes various visualizations and insights derived from histograms, countplots, boxplots, pie charts, heatmaps, and bar plots, this analysis highlights critical patterns and disparities globally. The findings are intended to inform decision-making processes and policy design to improve health outcomes and reduce inequalities.

Code:

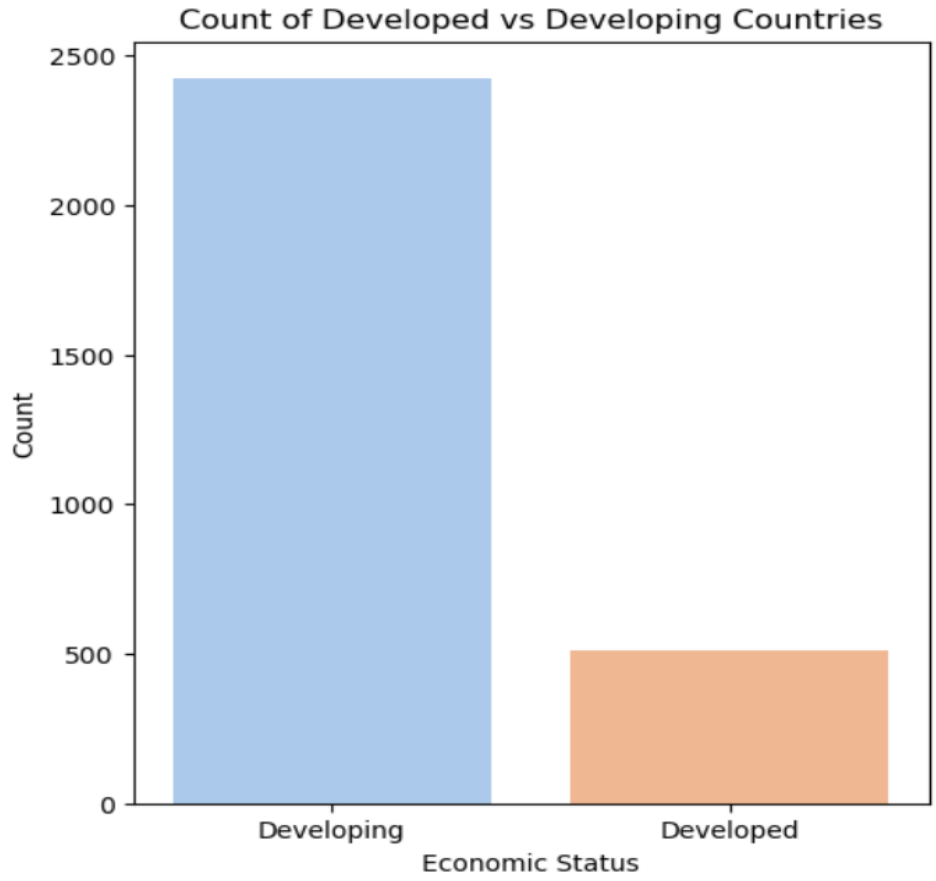
```
#import libraries

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from scipy import stats

 # Load the dataset

df = pd.read_csv("C:/Users/utaru/OneDrive/Desktop/internships2025/Unified_Mentor/Life Expectancy Data.csv")

df.head()
```

```
#cleaning the data and check for missing values
df.describe()
df.columns
df.shapes
df.info()
df.isnull().sum()

#handling the columns that have missing values
df['life_expectancy'] = df['life_expectancy'].fillna(df['life_expectancy'].median())
df['adult_mortality'] = df['adult_mortality'].fillna(df['adult_mortality'].median())
df['alcohol'] = df['alcohol'].fillna(df['alcohol'].median())
df['total_expenditure'] = df['total_expenditure'].fillna(df['total_expenditure'].median())
df['hepatitis_b'] = df.groupby('status')['hepatitis_b'].transform(lambda x: x.fillna(x.median()))
df['polio'] = df.groupby('status')['polio'].transform(lambda x: x.fillna(x.median()))
df['diphtheria'] = df.groupby('status')['diphtheria'].transform(lambda x: x.fillna(x.median()))

#plotting
plt.figure(figsize=(12, 6))

# Life Expectancy Histogram
plt.subplot(1, 2, 1)
sns.histplot(df['life_expectancy'], kde=True, bins=20, color='skyblue')
plt.title("Distribution of Life Expectancy")
plt.xlabel("Life Expectancy (years)")
plt.ylabel("Frequency")
plt.show()
```

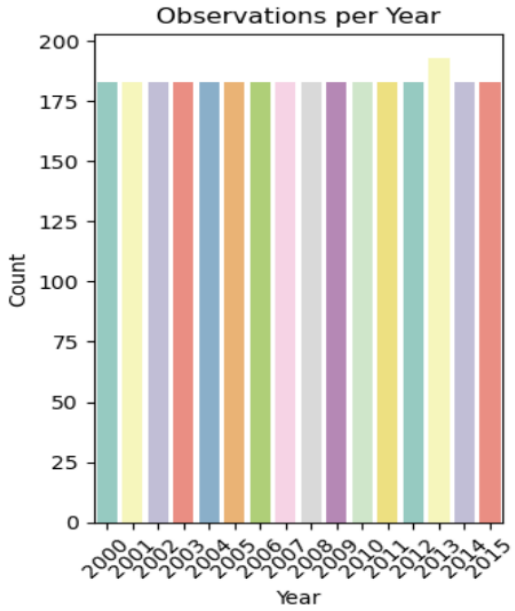

```
# GDP Histogram
plt.subplot(1, 2, 2)
sns.histplot(df['gdp'], kde=True, bins=20, color='green')
plt.title("Distribution of GDP")
plt.xlabel("GDP (in Trillions)")
plt.ylabel("Frequency")

plt.tight_layout()
plt.show()
```




```
plt.figure(figsize=(12, 6))


# Status Countplot
plt.subplot(1, 2, 1)
sns.countplot(data=df, x='status', palette='pastel')
plt.title("Count of Developed vs Developing Countries")
plt.xlabel("Economic Status")
plt.ylabel("Count")
plr.show()
```

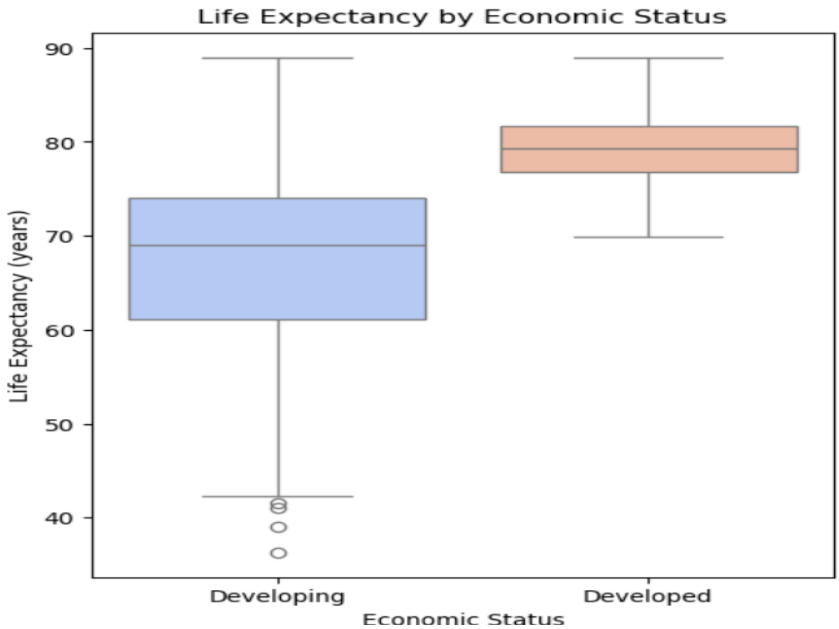



```
plt.subplot(1, 2, 2)
sns.countplot(data=df, x='year', palette='Set3')
plt.title("Observations per Year")
plt.xlabel("Year")
plt.ylabel("Count")
plt.xticks(rotation=45)

plt.tight_layout()
plt.show()
```

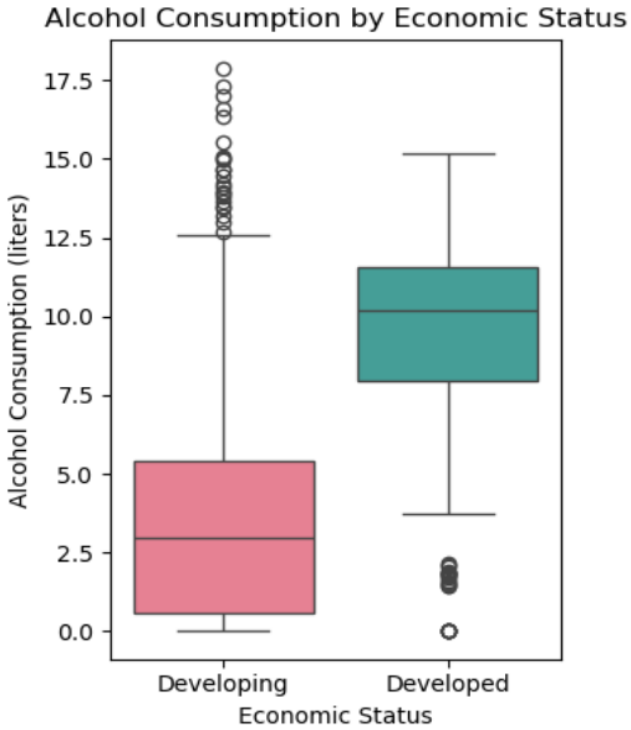


Data is evenly distributed across years, suggesting consistency in data collection over time.

```
plt.figure(figsize=(12, 6))


# Boxplot: Life Expectancy by Status
plt.subplot(1, 2, 1)
sns.boxplot(data=df, x="status", y="life_expectancy", palette="coolwarm")
plt.title("Life Expectancy by Economic Status")
plt.xlabel("Economic Status")
plt.ylabel("Life Expectancy (years)")
plt.show()
```

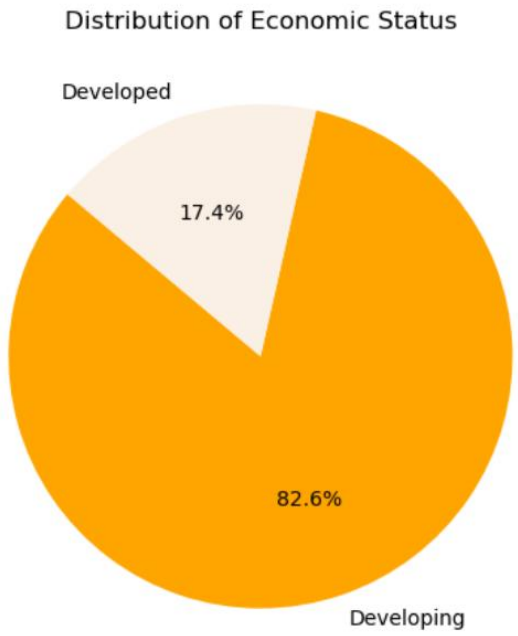


```
plt.subplot(1, 2, 2)
sns.boxplot(data=df, x="status", y="alcohol", palette="husl")
plt.title("Alcohol Consumption by Economic Status")
plt.xlabel("Economic Status")
plt.ylabel("Alcohol Consumption (liters)")

plt.tight_layout()
plt.show()
```



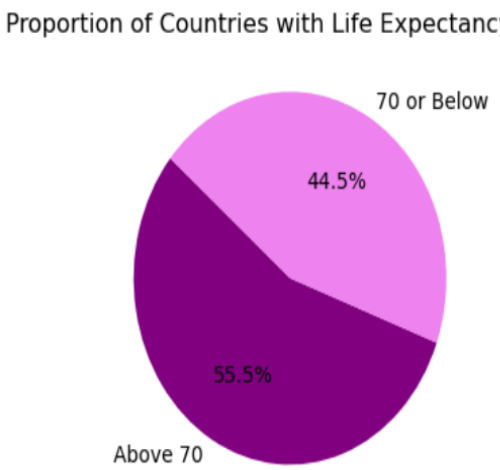
Developed countries consume more alcohol on average than Developing countries.

```
plt.figure(figsize=(12, 6))

# Pie Chart: Distribution of Economic Status
plt.subplot(1, 2, 1)
status_counts = df['status'].value_counts()
plt.pie(status_counts, labels=status_counts.index, autopct='%1.1f%%', colors=['orange', 'linen'], startangle=140)
plt.title("Distribution of Economic Status")
plt.show()
```



Around 75% of the data represents Developing countries, while 25% is from Developed countries.

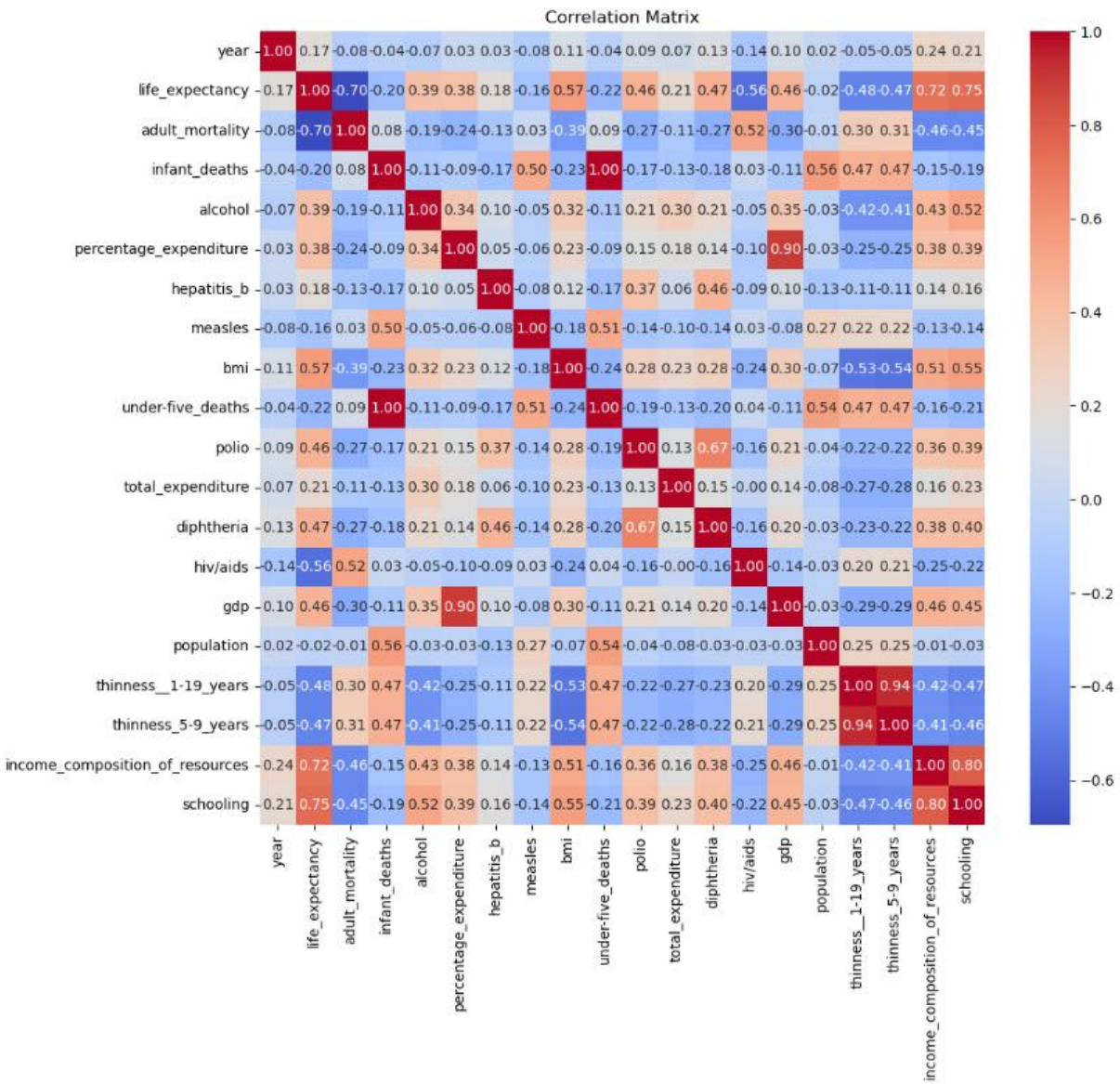
```
above_70 = df['life_expectancy'] > 70

life_exp_counts = above_70.value_counts()

plt.subplot(1, 2, 2)

plt.pie(life_exp_counts, labels=['Above 70', '70 or Below'], autopct='%1.1f%%', colors=['purple', 'violet'], startangle=140)

plt.title("Proportion of Countries with Life Expectancy > 70")


plt.tight_layout()

plt.show()
```



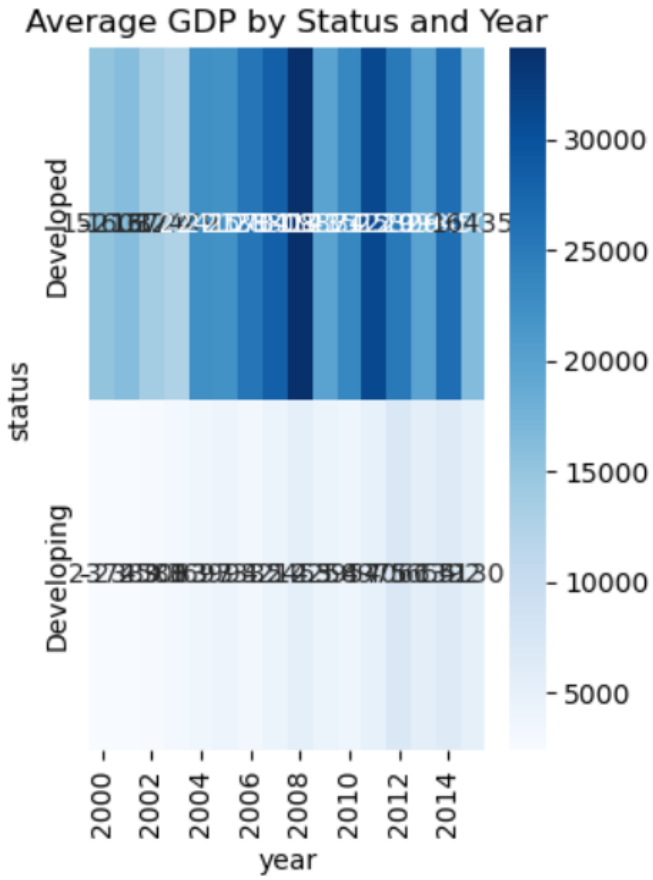
Approximately 50% of countries have a life expectancy greater than 70 years, reflecting global health disparities

```
plt.figure(figsize=(30, 10))

# Heatmap: Correlation Matrix
plt.subplot(1, 2, 1)
sns.heatmap(df.corr(numeric_only=True), annot=True, fmt=".2f", cmap="coolwarm", cbar=True)
plt.title("Correlation Matrix")
```

```
Text(0.5, 1.0, 'Correlation Matrix')
```



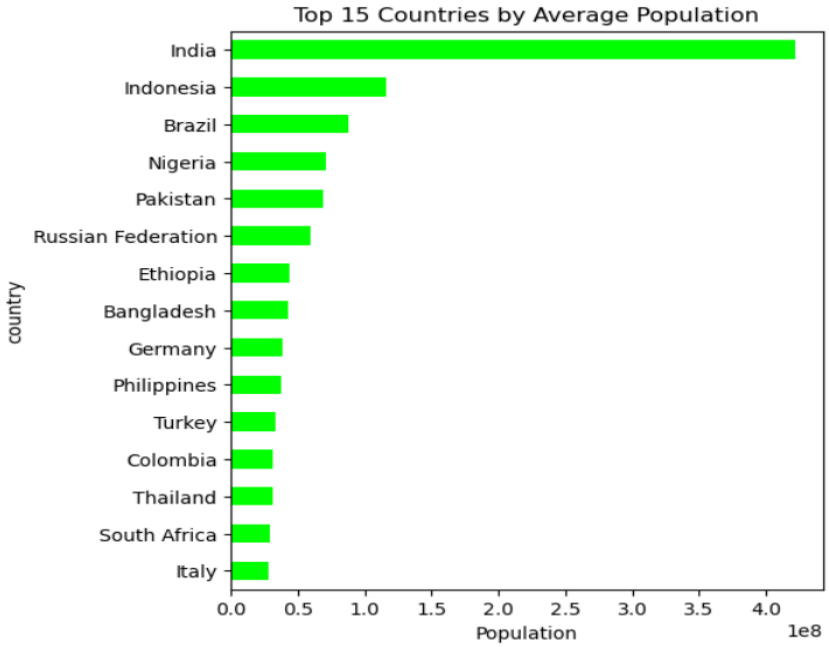

```
plt.subplot(1, 2, 2)
gdp_status = df.pivot_table(index="status", columns="year", values="gdp", aggfunc="mean")
sns.heatmap(gdp_status, cmap="Blues", annot=True, fmt=".0f")
plt.title("Average GDP by Status and Year")

plt.tight_layout()
plt.show()
```



Developed countries consistently exhibit higher GDP than Developing countries across all years.

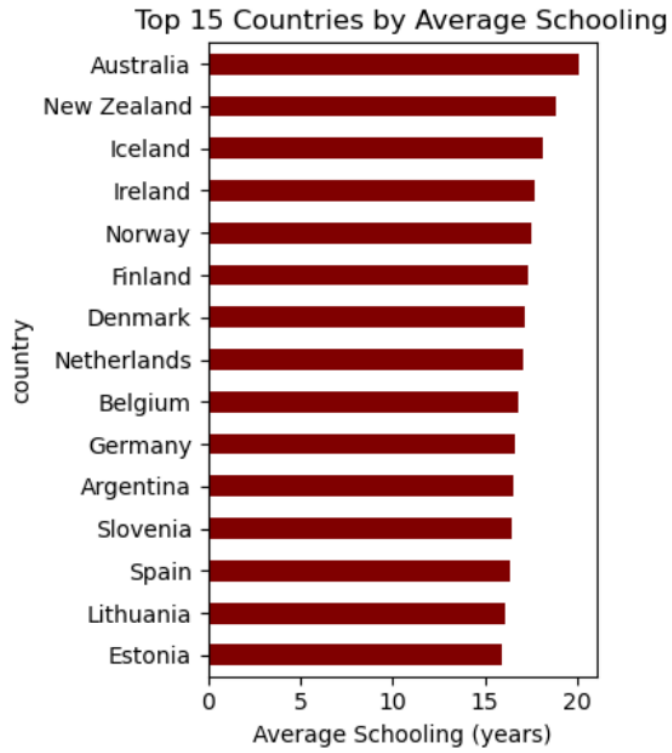
```
plt.figure(figsize=(12, 6))


# Bar Plot: Top 15 Countries by Population
plt.subplot(1, 2, 1)
country_pop = df.groupby("country")["population"].mean().nlargest(15)
country_pop.sort_values().plot(kind='barh', color='lime')
plt.title("Top 15 Countries by Average Population")
plt.xlabel("Population")
```

```
Text(0.5, 0, 'Population')
```



Countries like India and China dominate due to their significantly larger populations. The disparity in population sizes across countries is evident.

```
# Bar Plot: Top 15 Countries by Schooling

plt.subplot(1, 2, 2)

country_school = df.groupby("country")["schooling"].mean().nlargest(15)

country_school.sort_values().plot(kind='barh', color='maroon')

plt.title("Top 15 Countries by Average Schooling")

plt.xlabel("Average Schooling (years)")


plt.tight_layout()

plt.show()
```



Developed nations like Switzerland and Norway feature prominently due to their advanced education systems.

Analysis and Observations

Histograms

1. Life Expectancy Distribution:
    - Most countries have life expectancy values between 60 and 80 years.
    - The distribution is right-skewed, with a smaller number of countries having very high or very low life expectancy.
2. GDP Distribution:
    - GDP values are highly skewed to the right, with most countries having low GDP.
    - A few outliers exist with very high GDP values.

Countplots

1. Economic Status:
    - Developing countries dominate the dataset, constituting approximately 75% of observations.

2. Observations per Year:
    - The data is evenly distributed across years, reflecting consistency in data collection.

Boxplots

1. Life Expectancy by Economic Status:
    - Developed countries have a higher and more consistent life expectancy compared to Developing countries.
    - Developing countries show a wider range, with several outliers indicating very low life expectancy.
2. Alcohol Consumption by Economic Status:
    - Alcohol consumption is generally higher in Developed countries, with less variability compared to Developing countries.

Pie Chart

1. Distribution of Economic Status:
    - Approximately 75% of the data represents Developing countries, while 25% represents Developed countries.
2. Countries with Life Expectancy > 70:
    - Around 50% of the observations pertain to countries where life expectancy exceeds 70 years.

Heatmaps

1. Correlation Matrix:
    - Positive correlations:
        - Life expectancy strongly correlates with schooling and income composition of resources.
    - Negative correlations:
        - Life expectancy is negatively correlated with adult mortality and thinness variables.
2. GDP by Status and Year:
    - Developed countries consistently exhibit higher GDP than Developing countries across all years.
    - A gradual increase in GDP is observed for both Developed and Developing countries over time.

Bar Plots

1. Top 15 Countries by Population:
    - Countries like India and China dominate due to their significantly larger populations.

    - The disparity in population sizes across countries is evident.
2. Top 15 Countries by Schooling:
    - Developed nations such as Switzerland and Norway lead in average years of schooling, reflecting strong education systems.
    - Educational disparities between countries are highlighted.

Life expectancy is closely linked to socio-economic factors such as schooling, income composition, and healthcare indicators.

- Developed countries consistently perform better in terms of life expectancy, GDP, and education, while Developing countries show higher variability and outliers.
- The dataset reflects global health disparities and emphasizes the importance of economic and educational development in improving life expectancy.

Conclusion

This analysis provides a comprehensive understanding of factors influencing life expectancy. The findings highlight significant global disparities and underscore the critical role of education, healthcare, and economic growth in bridging these gaps. Policymakers and stakeholders can leverage these insights to develop targeted strategies for improving health outcomes, particularly in Developing countries. Continued focus on these key indicators will be instrumental in achieving equitable health advancements and fostering sustainable development worldwide.