Exploratory Data Analysis (EDA)

GLOBAL DATASET:

Introduction to the Global Dataset

The **Global Dataset** contains health-related attributes collected from various regions worldwide. The dataset aims to analyze risk factors contributing to **heart disease** and related conditions by exploring demographic, lifestyle, and medical variables.

The key objectives of performing **Exploratory Data Analysis (EDA)** on this dataset are:

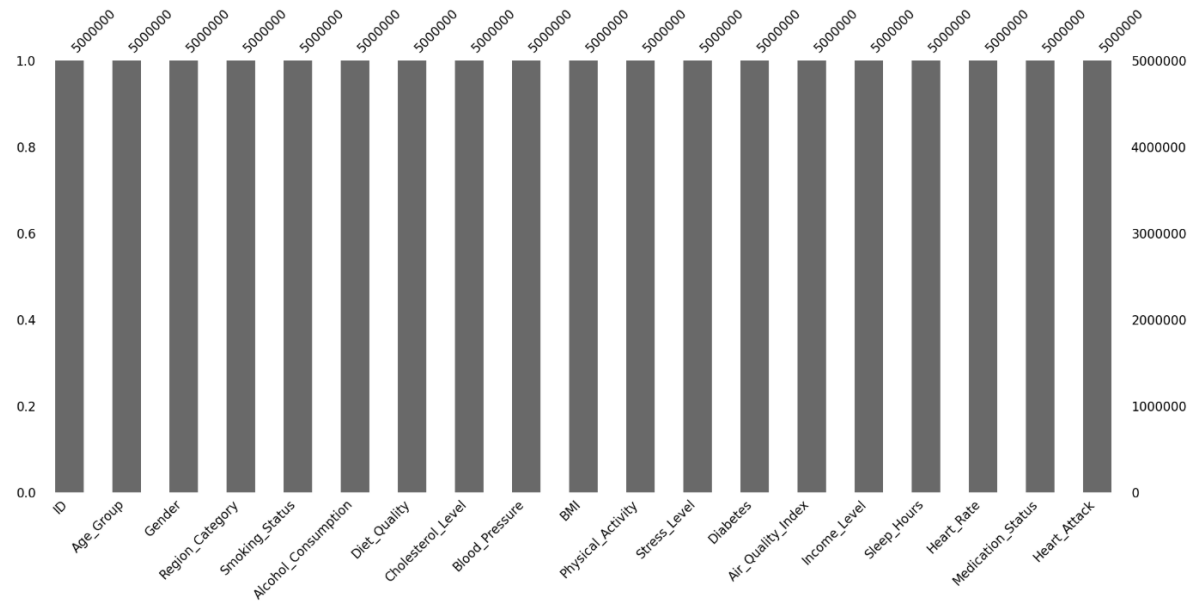
✅ **Understand the distribution of health parameters** (age, cholesterol, blood pressure, etc.)

✅ **Identify patterns & relationships** between features and heart disease

✅ **Detect anomalies, missing values, or outliers** that may affect analysis

✅ **Gain insights that will help in predictive modeling & decision-making**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
global_df = pd.read_csv('C:/Users/jsuri/Downloads/H_A/cleaned_global_dataset.csv')
global_df.head()  # Display the first few rows

global_df.info()
global_df.isnull().sum()
global_df.describe(include ='all')

import missingno as msno
msno.bar(global_df) plt.show()
```



📊 Univariate, Bivariate, and Multivariate Analysis for Global Dataset

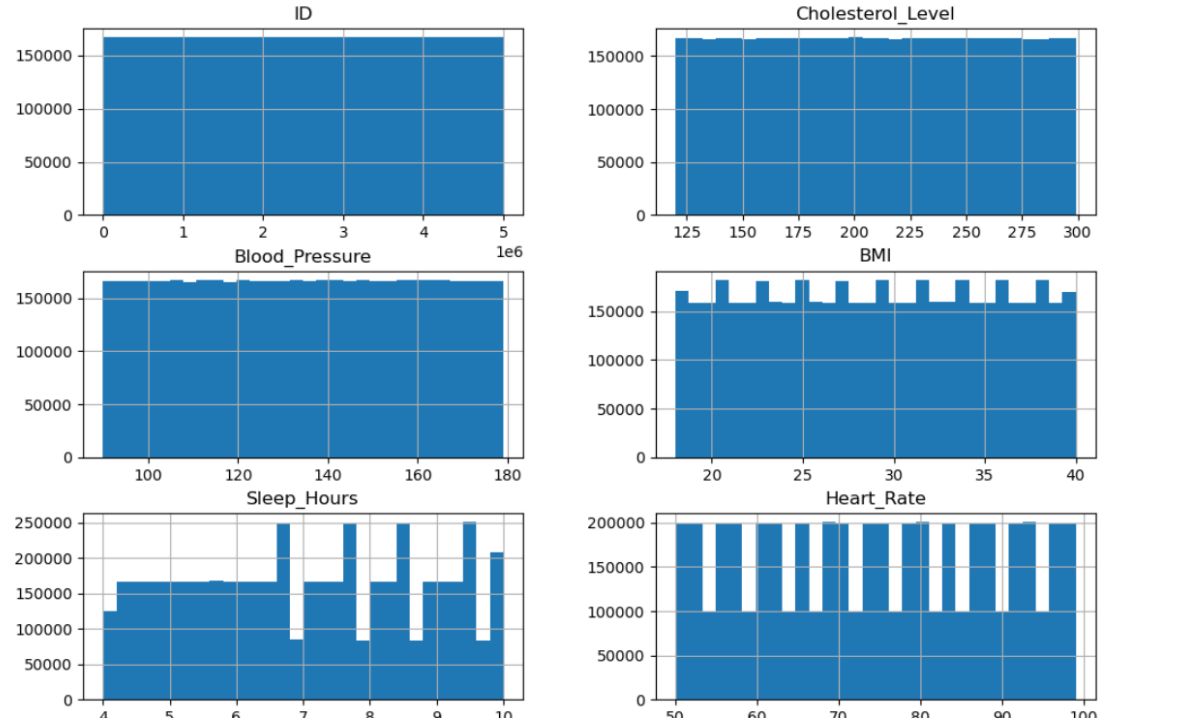
After importing the dataset, we now analyze its structure using **Univariate, Bivariate, and Multivariate** techniques. These methods help us identify patterns, trends, and correlations in the dataset. Below is the structured breakdown of the EDA process:

Univariate Analysis

Univariate analysis examines individual features separately to understand their distribution, outliers, and central tendency.

📌 Histogram of All Numeric Features

```
global_df.hist(figsize=(12, 8), bins=30)
plt.show()
```



📝 Insight:

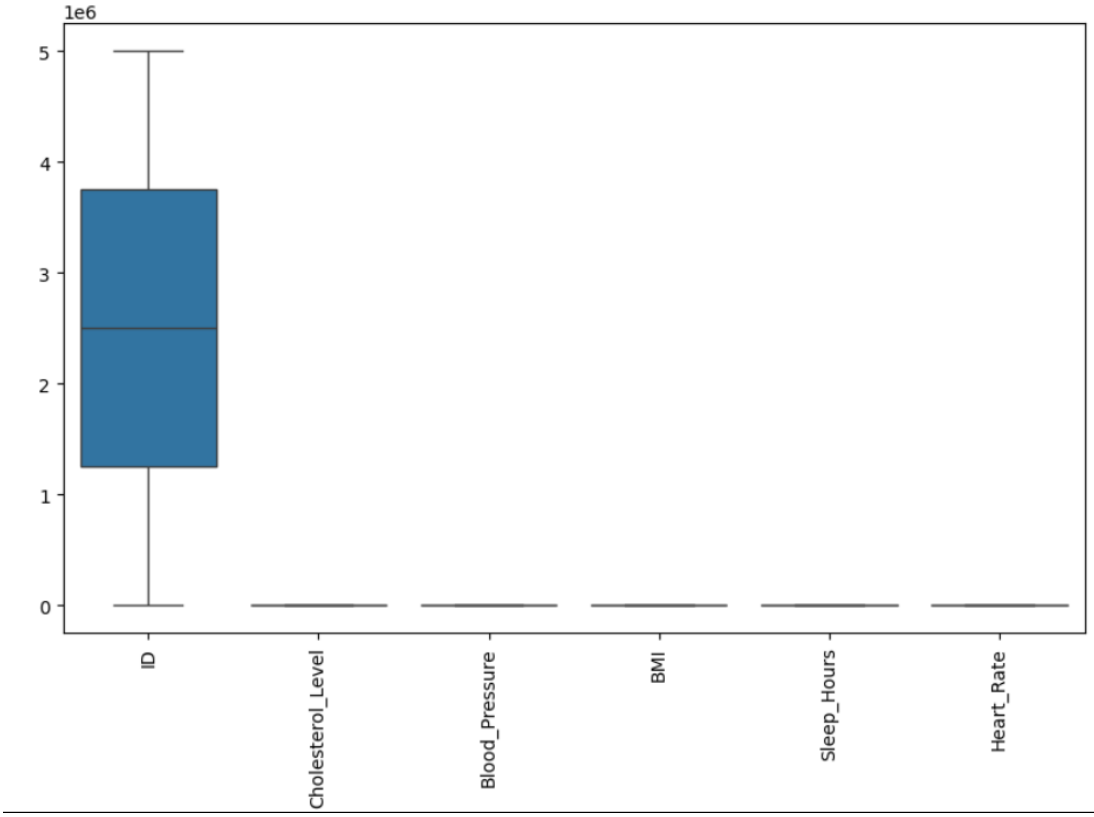
✅ The histograms show the distribution of numerical variables.

✅ Some variables may be skewed, requiring transformation.

✅ The spread of data helps in understanding normality or the presence of outliers.

📌 Boxplot for Numeric Features

```
plt.figure(figsize=(10,6))
sns.boxplot(data=global_df)
plt.xticks(rotation=90)
plt.show()
```



📝 **Insight:**

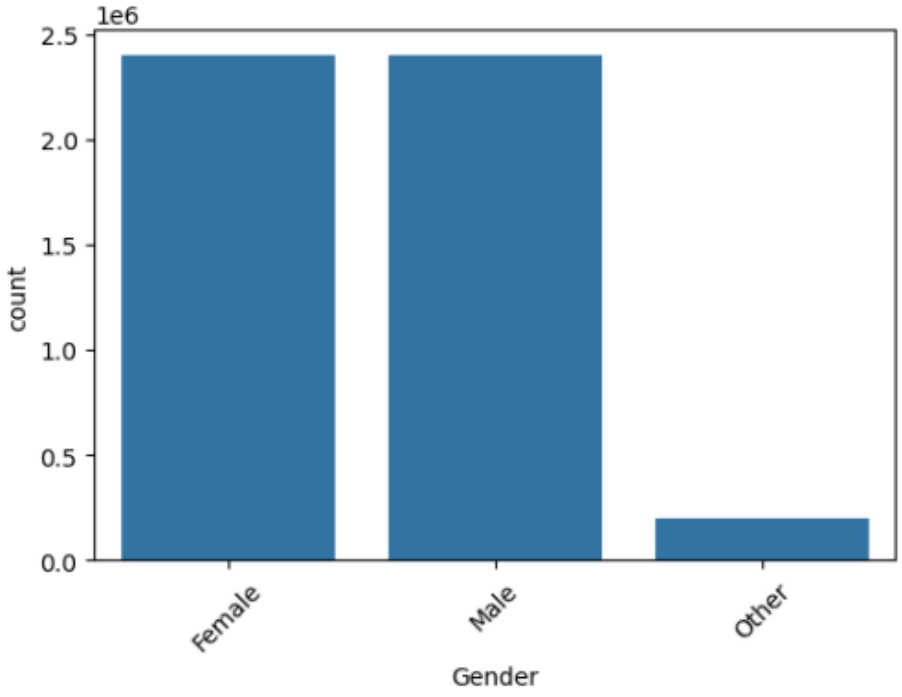
✅ Boxplots help detect **outliers** and **spread of data** in numeric features.

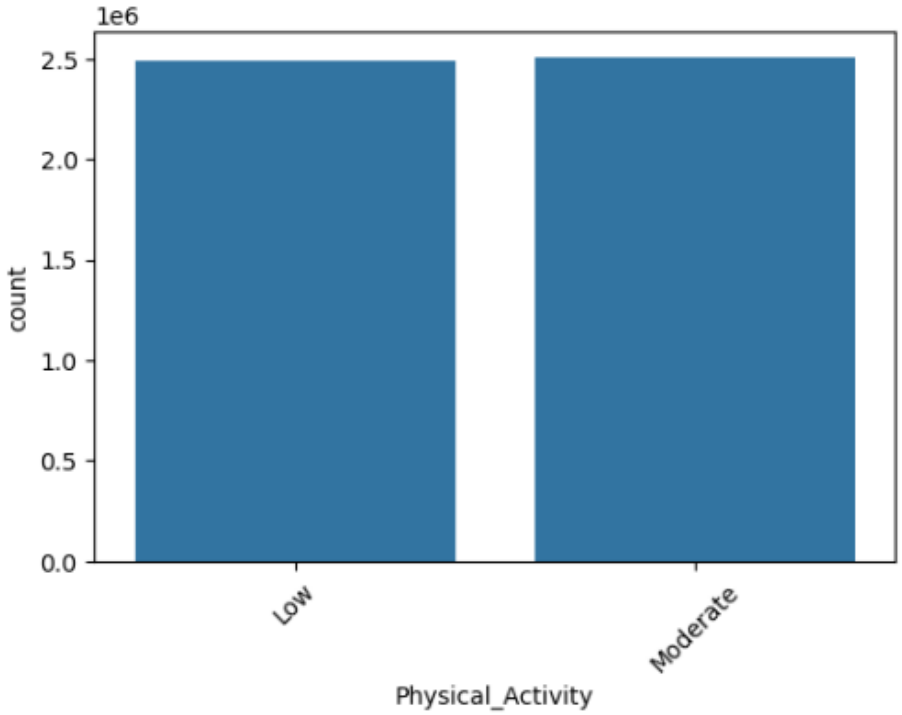
✅ Features with extreme values may need **scaling or transformation**.

✅ Wider interquartile ranges (IQR) indicate **higher variability** in certain features.

📌 **Countplot for Categorical Variables**

```
categorical_cols = ['Gender', 'Physical_Activity']
for col in categorical_cols:
    plt.figure(figsize=(6, 4))
    sns.countplot(x=global_df[col])
    plt.xticks(rotation=45)
 plt.show()
```





📝 **Insight:**

✅ **Gender distribution** helps analyze male vs. female representation.

✅ **Physical activity levels** indicate how many individuals follow an active lifestyle.

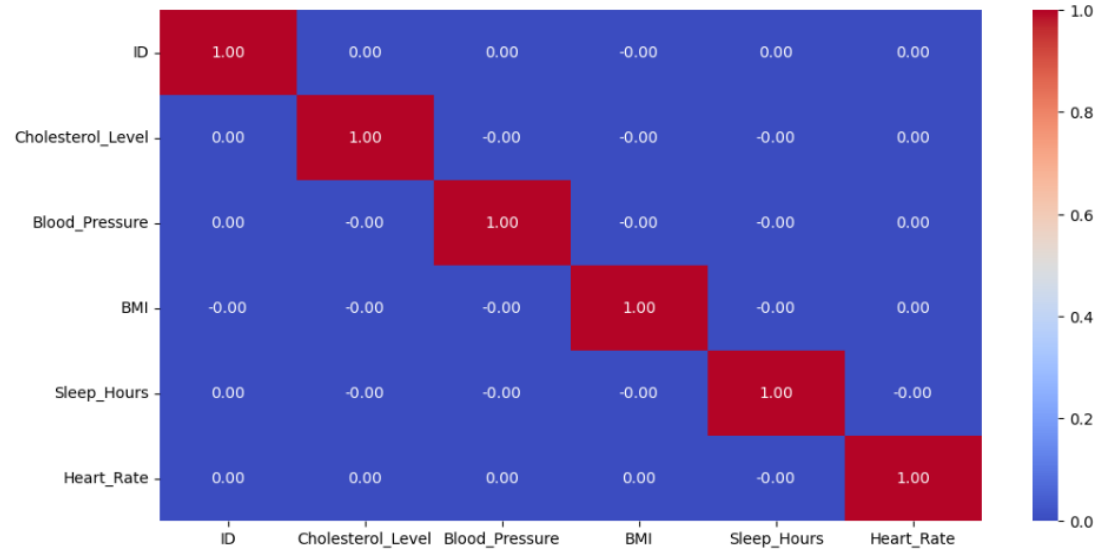
✅ Skewed distributions may affect model predictions and analysis.

Bivariate Analysis

Bivariate analysis examines **relationships between two variables** to identify dependencies and trends.

📌 **Correlation Heatmap**

```
plt.figure(figsize=(12, 6))
sns.heatmap(global_df.corr(numeric_only=True), annot=True, cmap="coolwarm", fmt=".2f")
plt.show()
```



📝 **Insight:**

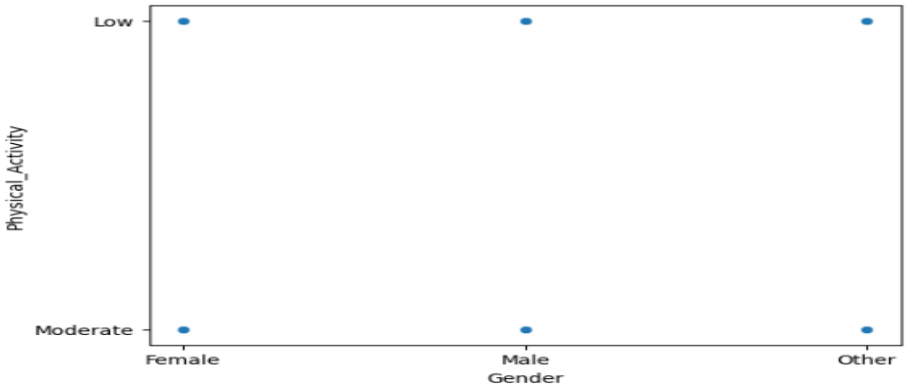
✅ The heatmap shows **how numerical variables are related** to each other.

✅ Strong correlations (values near +1 or -1) indicate a linear relationship.

✅ Features with weak correlations may not contribute much to predictive models.

Scatterplot: Gender vs. Physical Activity

```
sns.scatterplot(x='Gender', y='Physical_Activity', data=global_df)
plt.show()
```

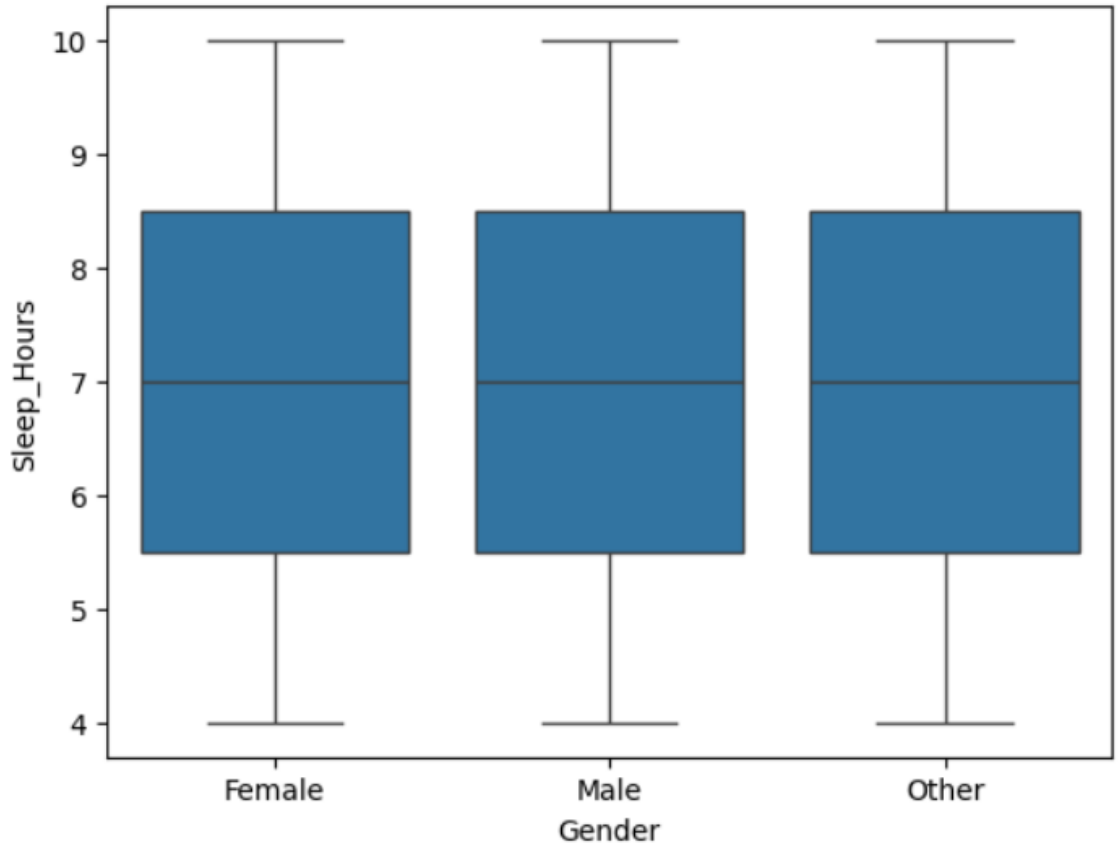


Multivariate Analysis

Multivariate analysis examines **relationships among three or more variables** to uncover complex patterns.

📌 **Boxplot: Gender vs. Sleep Hours**

```
sns.boxplot(x='Gender', y='Sleep_Hours', data=global_df)
plt.show()
```



✅ **Conclusion**

🔹 Univariate analysis helped us understand **distribution & outliers** in numerical and categorical data.

🔹 Bivariate analysis identified **relationships between variables** (e.g., Gender vs. Physical Activity).

🔹 Multivariate analysis uncovered **complex interactions** (e.g., Gender vs. Sleep Hours).

INDIA DATASET:

📌 Introduction

The India dataset contains heart disease-related data specific to the Indian population. The goal of this analysis is to understand key health indicators, lifestyle habits, and their impact on heart disease. The EDA process includes:

- Univariate Analysis: Examining individual features.
- Bivariate Analysis: Exploring relationships between two variables.
- Multivariate Analysis: Analyzing interactions among multiple variables.

Importing Libraries & Loading the Dataset

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

ind_df = pd.read_csv("C:/Users/jsuri/Downloads/H_A/cleaned_India_dataset.csv")
```

Dataset Overview

```
ind_df.info()
ind_df.describe()
ind_df.isnull().sum()
```

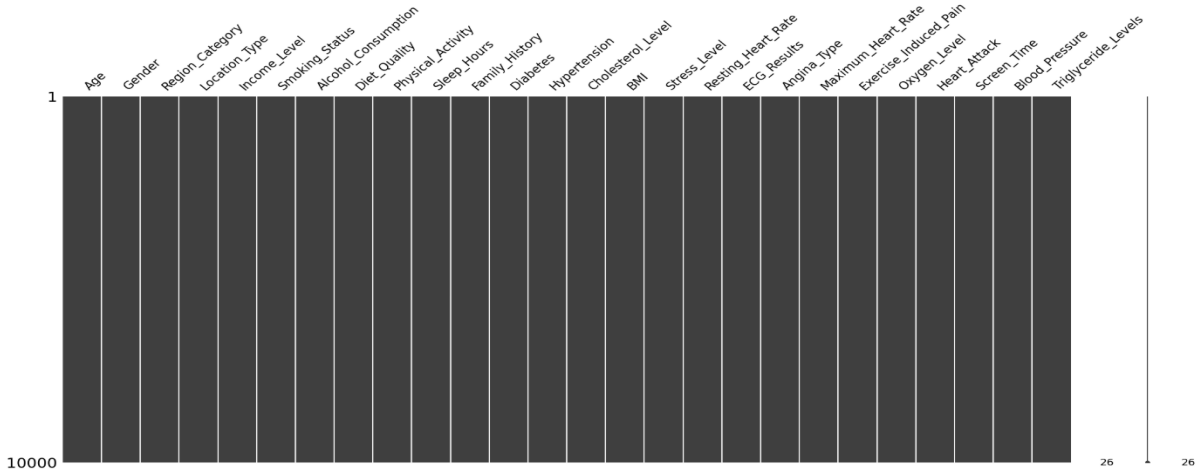
🔹 Insights:

- **info()** shows column names, data types, and missing values.
- **describe()** provides summary statistics (mean, min, max, etc.).
- **isnull().sum()** checks for missing values in each column.

Univariate Analysis

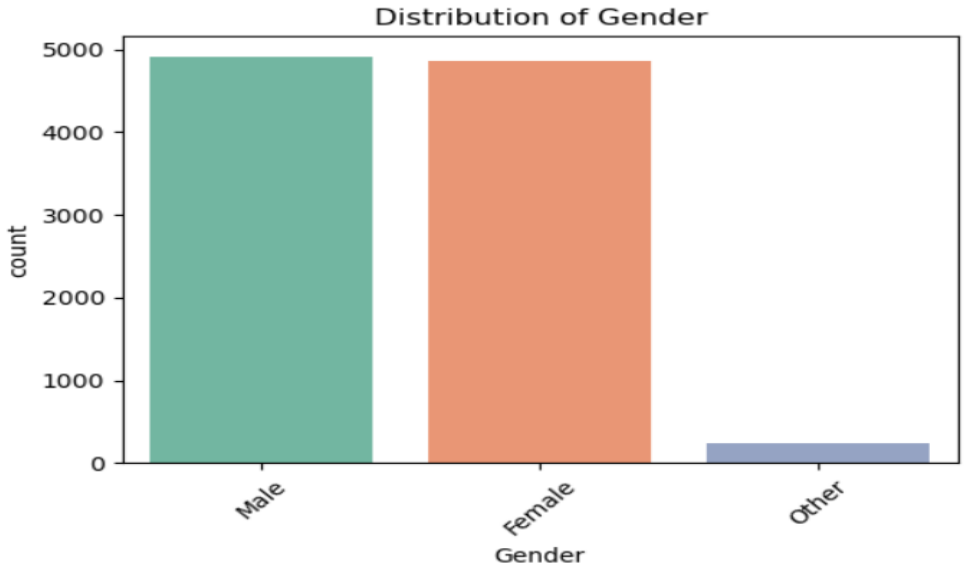
Missing Data Visualization

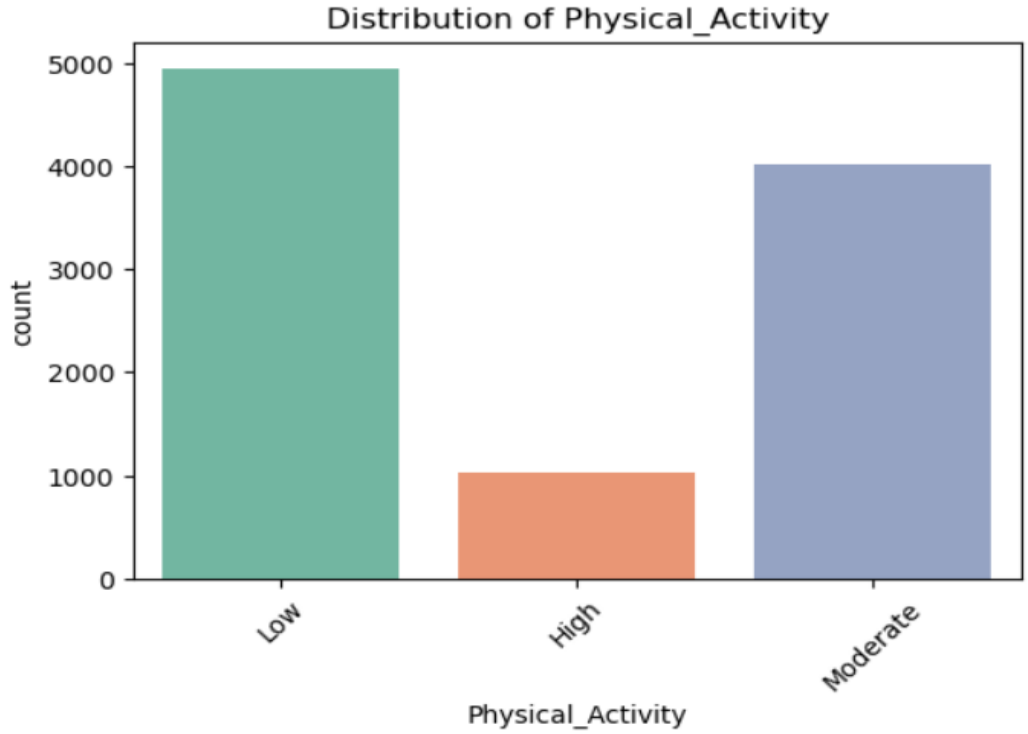
```
msno.matrix(ind_df)
plt.show()
```

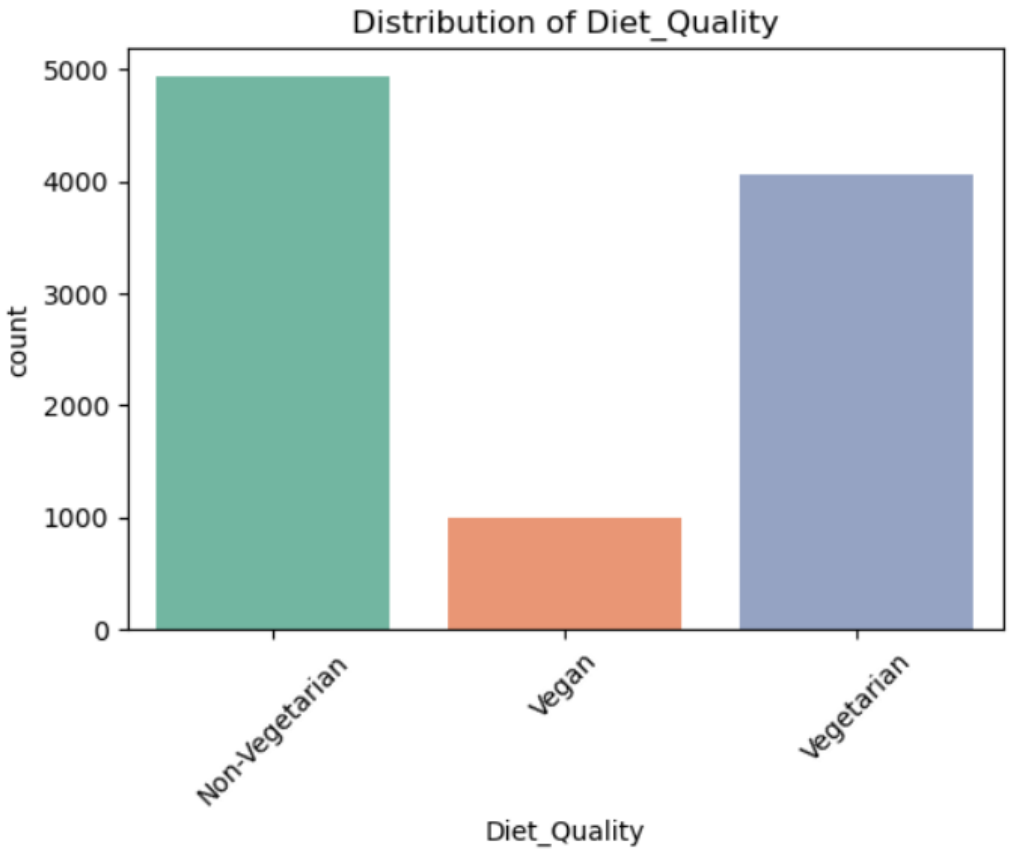


Countplots for Categorical Variables

```
categorical_cols = ['Gender', 'Physical_Activity', 'Diet_Quality']
for col in categorical_cols:
    plt.figure(figsize=(6, 4))
    sns.countplot(x=ind_df[col], palette="Set2")
    plt.xticks(rotation=45)
    plt.title(f"Distribution of {col}")
    plt.show()
```







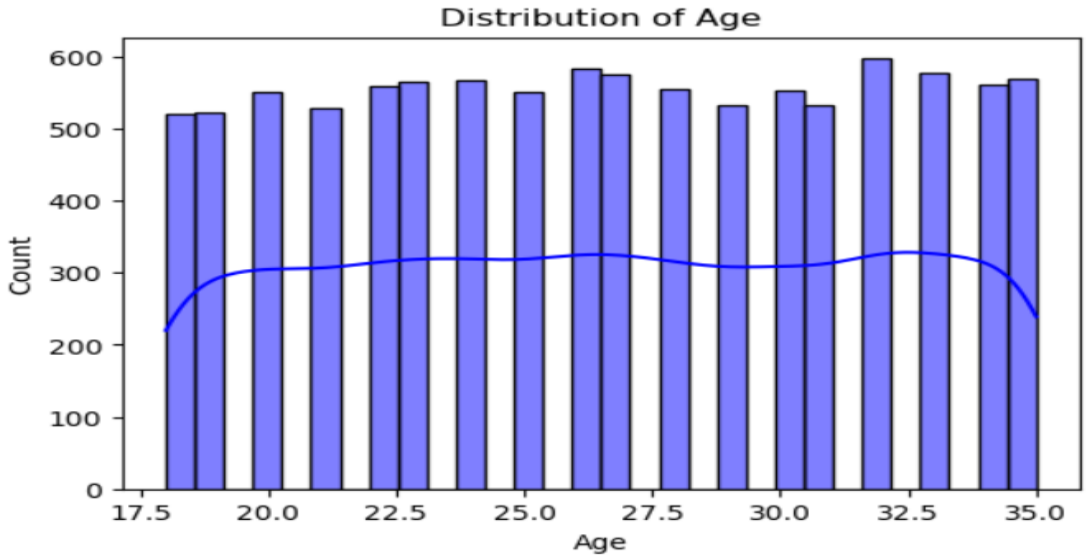
🔹 **Insight:**

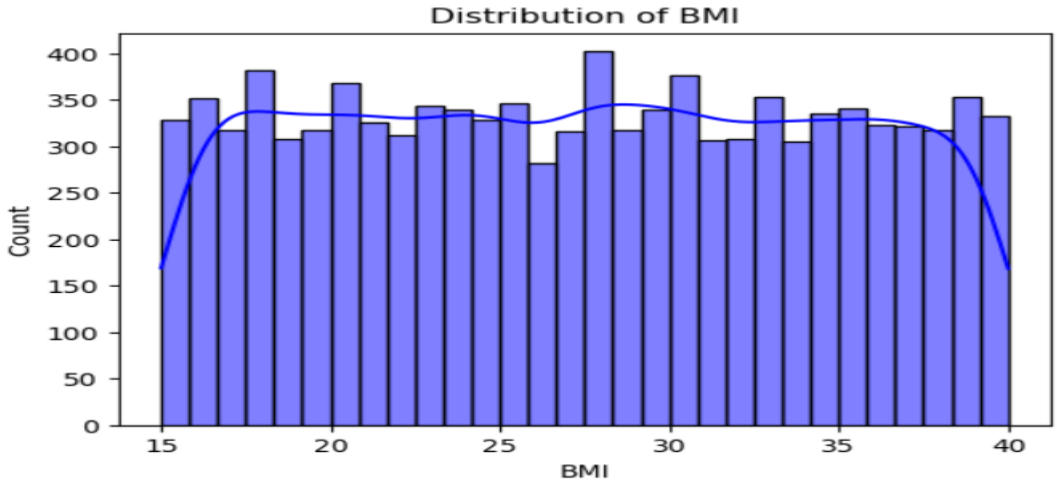
✅ Shows the **distribution of categorical features** (e.g., Gender, Physical Activity, Diet Quality).

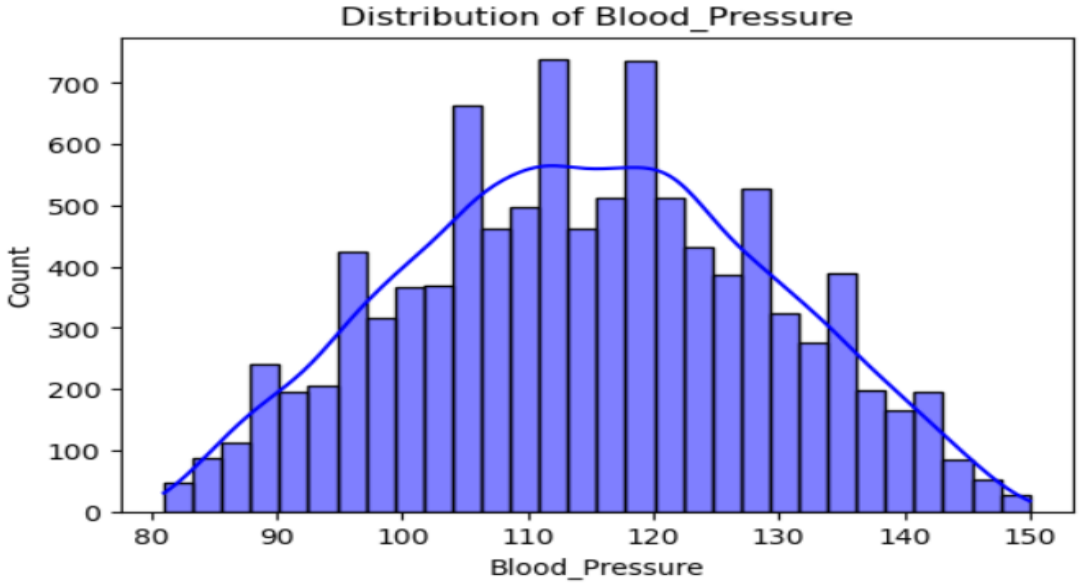
✅ Helps identify **imbalances in categories**.

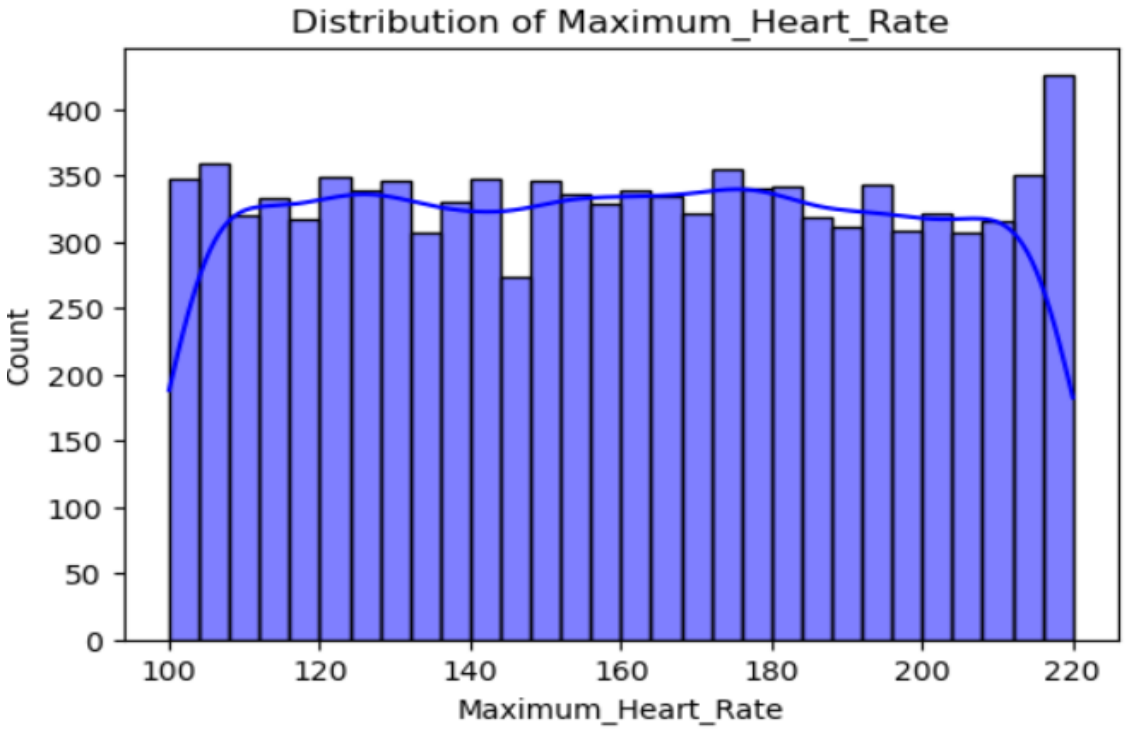
Histograms for Numerical Variables

```
numerical_cols = ['Age', 'BMI', 'Blood_Pressure', 'Maximum_Heart_Rate']
for col in numerical_cols:
    plt.figure(figsize=(6, 4))
    sns.histplot(ind_df[col], bins=30, kde=True, color='blue')
    plt.title(f"Distribution of {col}")
    plt.show()
```


Distribution of Age
Count
Age
0
100
200
300
400
500
600
17.5
20.0
22.5
25.0
27.5
30.0
32.5
35.0


Distribution of BMI
Count
BMI
0
50
100
150
200
250
300
350
400
15
20
25
30
35
40


Distribution of Blood_Pressure
Count
Blood_Pressure
0
100
200
300
400
500
600
700
80
90
100
110
120
130
140
150



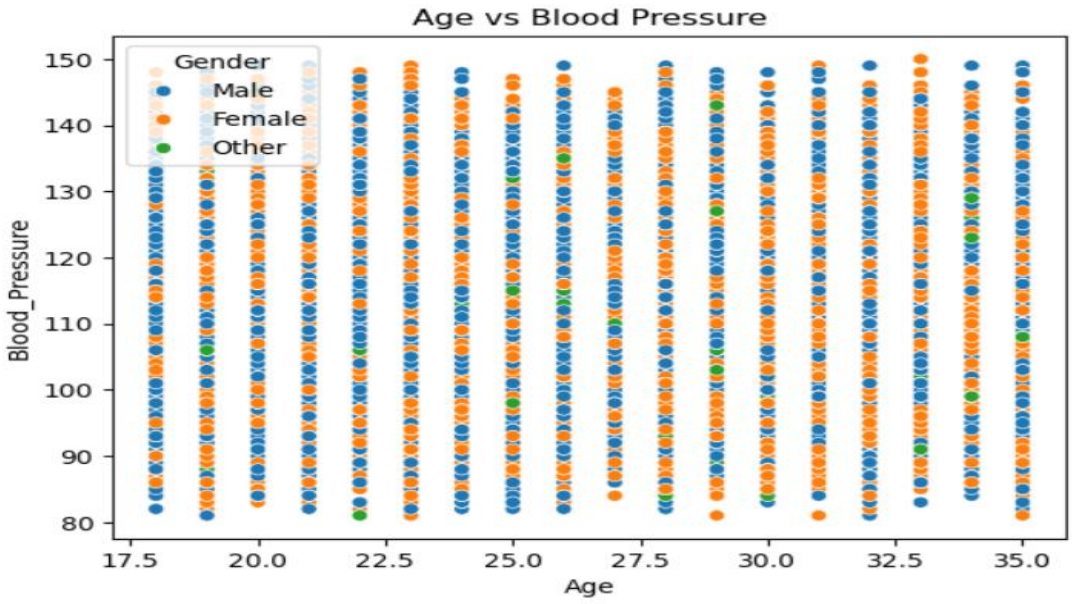
- **Insight:**

Helps understand **data distribution, skewness, and outliers** in numerical features.

Bivariate Analysis

Scatterplot: Age vs. Blood Pressure (Colored by Gender)

```
sns.scatterplot(x=ind_df['Age'], y=ind_df['Blood_Pressure'], hue=ind_df['Gender'])
plt.title("Age vs Blood Pressure")
plt.show()
```



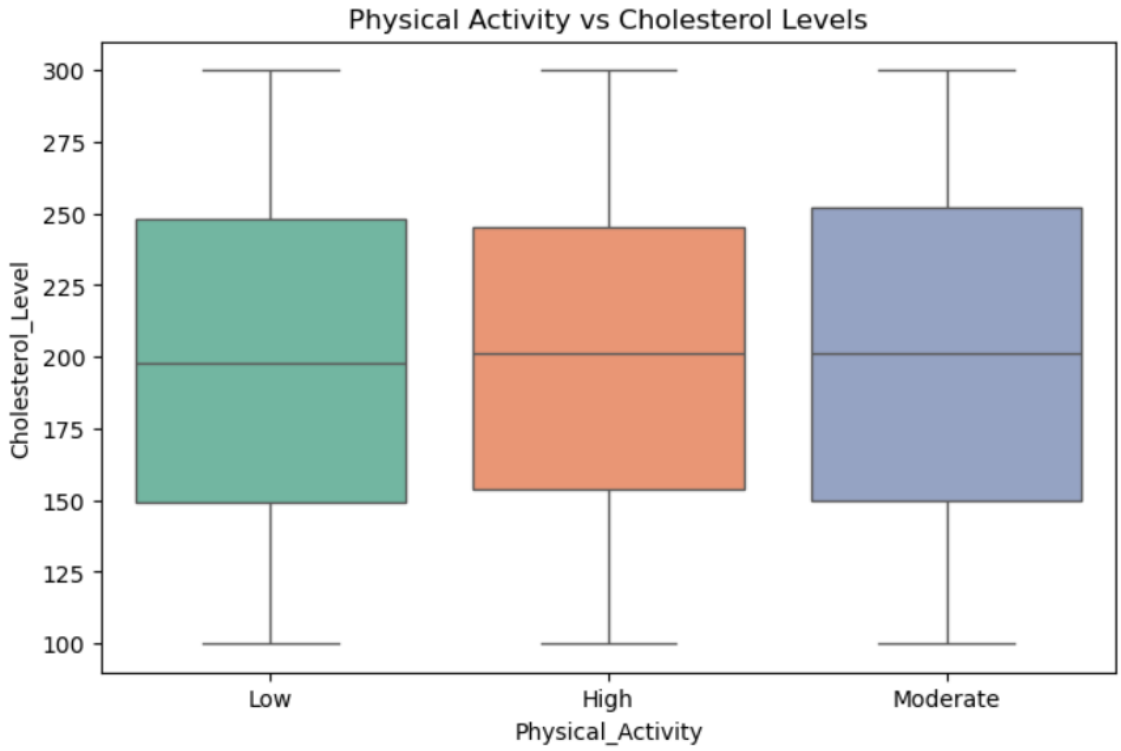
🔹 **Insight:**

✅ Helps understand if **Blood Pressure increases with Age**.

✅ Identifies **gender-based differences** in Blood Pressure levels.

Boxplot: Physical Activity vs. Cholesterol Level

```
plt.figure(figsize=(8, 5))
sns.boxplot(x=ind_df['Physical_Activity'], y=ind_df['Cholesterol_Level'], palette="Set2")
plt.title("Physical Activity vs Cholesterol Levels")
plt.show()
```



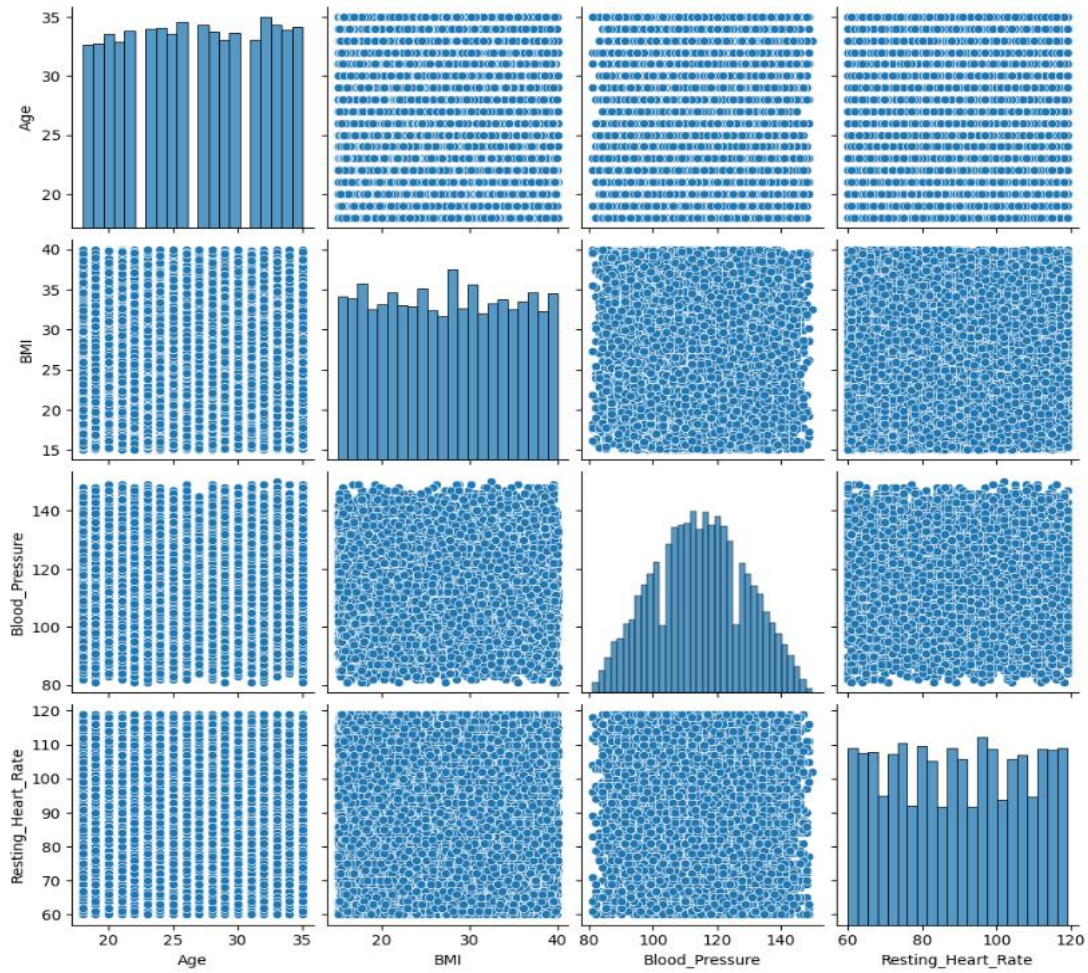
🔹 **Insight:**

✅ **Cholesterol levels** may be lower in individuals with **higher physical activity**.

Multivariate Analysis

Pairplot: Relationships Between Key Features

```
sns.pairplot(ind_df[['Age', 'BMI', 'Blood_Pressure', 'Resting_Heart_Rate']])
plt.show()
```

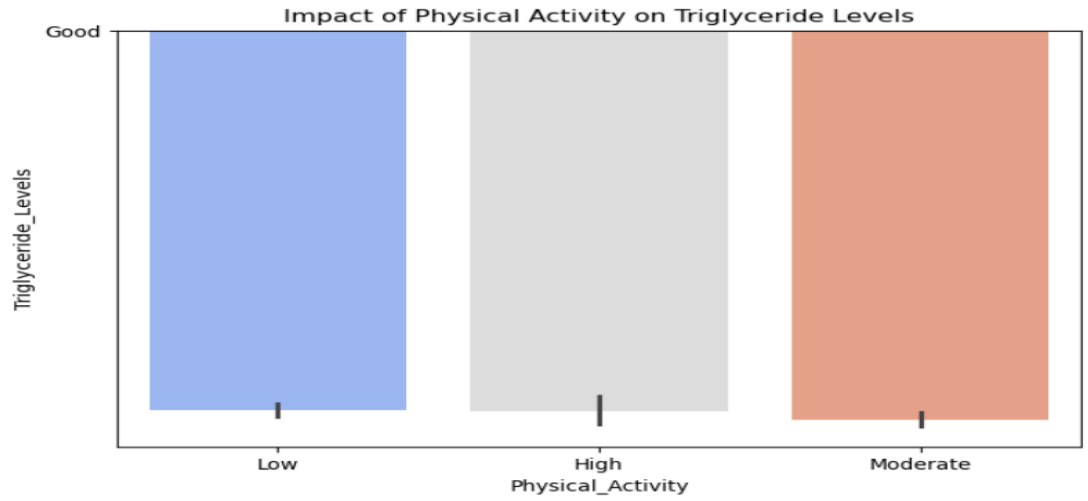


🔹 **Insight:**

✅ Helps identify **correlations** and **patterns** between multiple numerical variables.

Barplot: Physical Activity vs. Triglyceride Levels

```
plt.figure(figsize=(8, 5))
sns.barplot(x='Physical_Activity', y='Triglyceride_Levels', data=ind_df, palette="coolwarm")
plt.title("Impact of Physical Activity on Triglyceride Levels")
plt.show()
```



🔹 **Insight:**

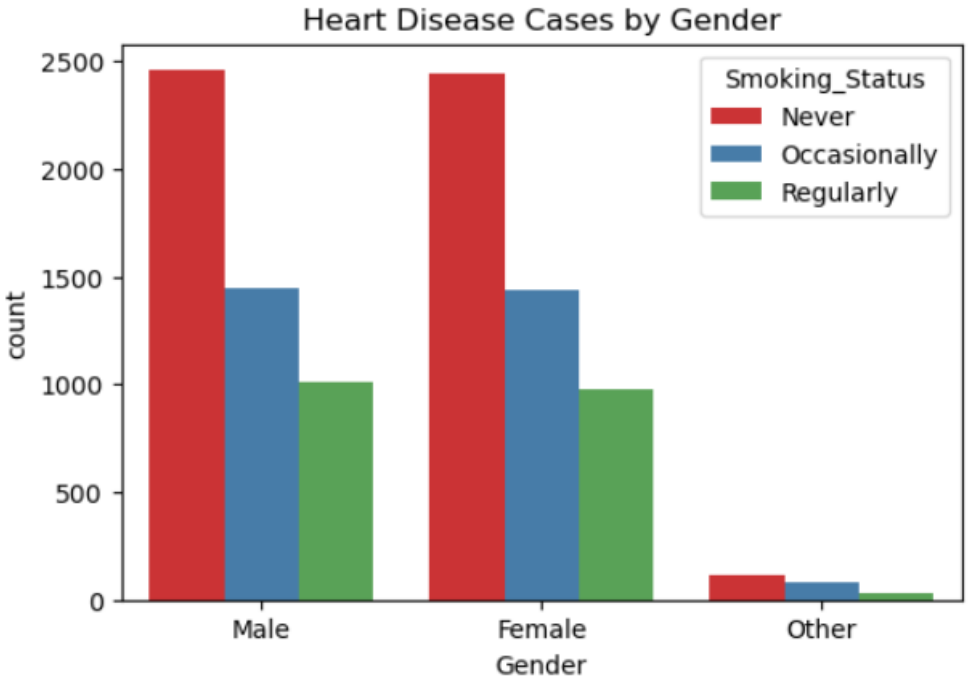
✅ Analyzes **how triglyceride levels vary with physical activity**.

Countplot: Gender vs. Smoking Status

```
plt.figure(figsize=(6, 4))
sns.countplot(x='Gender', hue='Smoking_Status', data=ind_df, palette="Set1")
plt.title("Heart Disease Cases by Gender")
plt.show()
```

🔹 **Insight:**

✅ Compares **smoking prevalence among genders** and its impact on heart health.



🔹 **Insight:**

✅ Compares **smoking prevalence among genders** and its impact on heart health.

✅ **Conclusion**

🔹 **Univariate analysis** provided insights into data distribution and missing values.

🔹 **Bivariate analysis** helped identify relationships between features.

🔹 **Multivariate analysis** uncovered **complex interactions between variables**.

INDONESIA DATASET

Introduction

The Indonesia dataset contains heart disease-related data specific to the Indonesian population. The goal of this analysis is to examine key health indicators, lifestyle habits, and their impact on heart disease. The EDA process includes:

- Univariate Analysis: Examining individual features.
- Bivariate Analysis: Exploring relationships between two variables.
- Multivariate Analysis: Analyzing interactions among multiple variables.

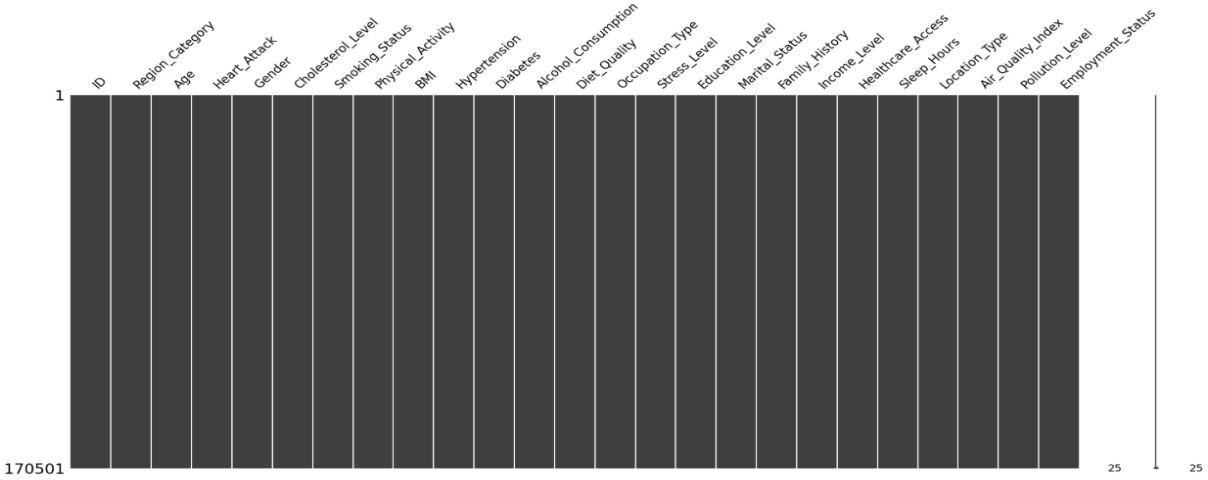
Importing Libraries & Loading the Dataset

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

indo_df = pd.read_csv("C:/Users/jsuri/Downloads/H_A/cleaned_Indonesia_dataset.csv")

indo_df.info()
indo_df.isnull().sum()
indo_df.describe()
indo_df.describe(include='object')
```

- **Insights:**
  - **info()** displays data types and missing values.
  - **isnull().sum()** helps identify missing values.
  - **describe()** provides statistical summaries.
  - **describe(include='object')** gives details on categorical features.

Missing Data Visualization

```
msno.matrix(indo_df)
plt.show()
```



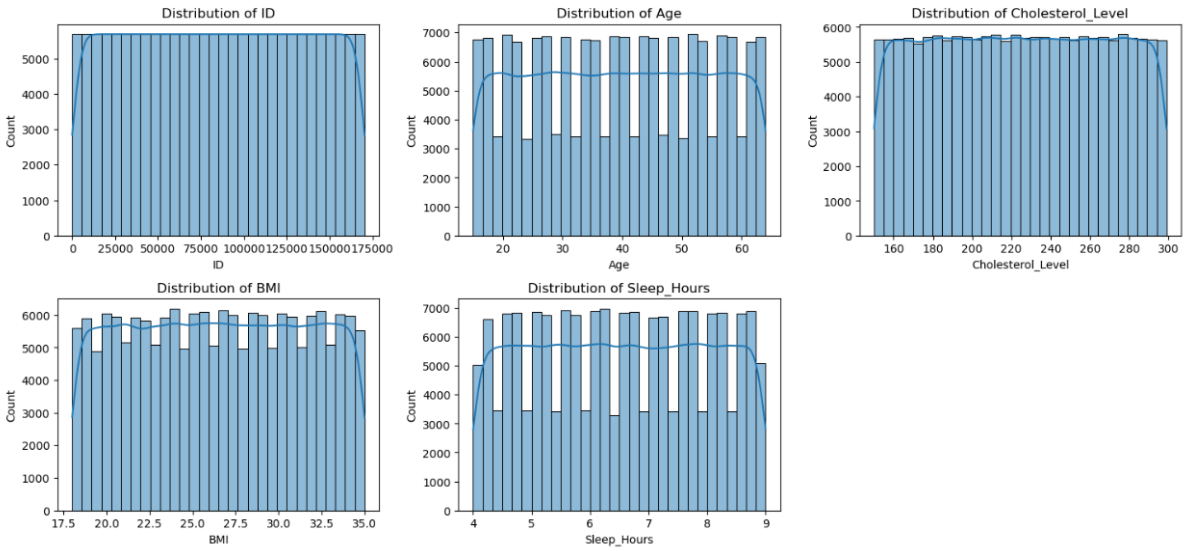
🔹 **Insight:**

✅ Helps visualize missing values in the dataset.

Univariate Analysis

Histograms for Numerical Variables

```
numerical_cols = indo_df.select_dtypes(include=['int64', 'float64']).columns

plt.figure(figsize=(15, 10))
for i, col in enumerate(numerical_cols, 1):
    plt.subplot(3, 3, i)
    sns.histplot(indo_df[col], bins=30, kde=True)
    plt.title(f"Distribution of {col}")
plt.tight_layout()
plt.show()
```

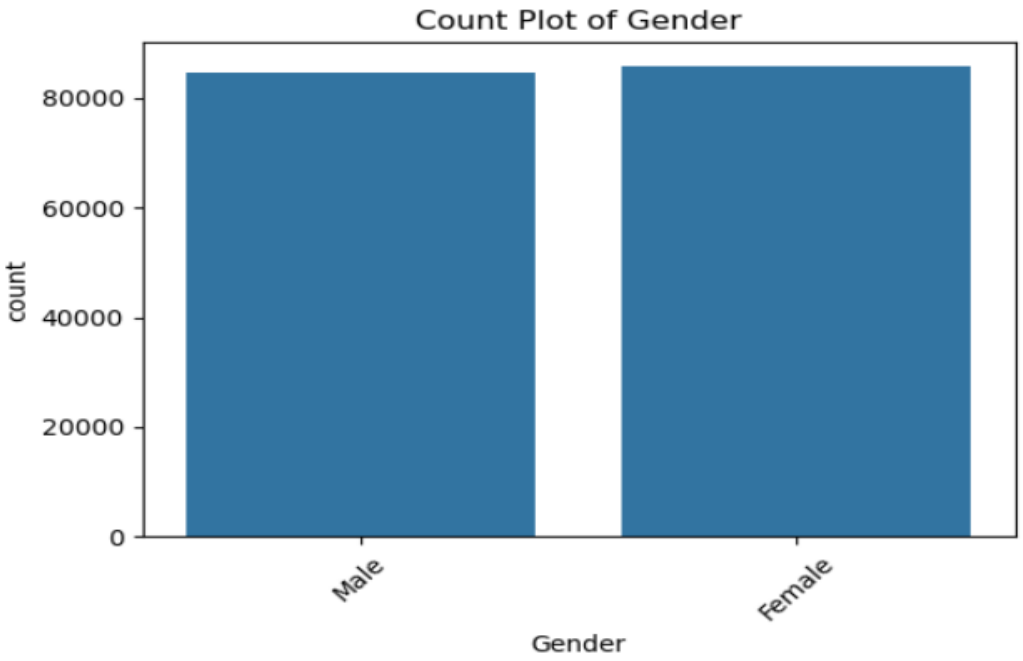


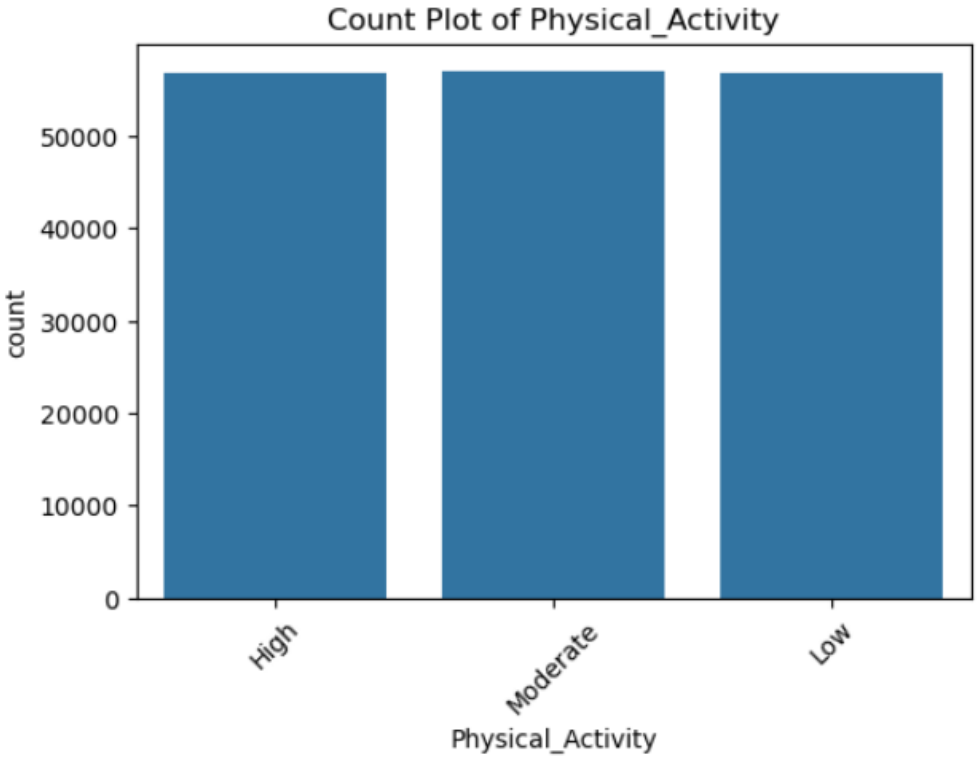
🔹 **Insight:**

✅ Helps understand data distribution, skewness, and potential outliers in numerical features.

Countplots for Categorical Variables

```
categorical_cols = ['Gender', 'Physical_Activity', 'Diet_Quality']
for col in categorical_cols:
    plt.figure(figsize=(6, 4))
    sns.countplot(x=indo_df[col])
    plt.xticks(rotation=45)
    plt.title(f"Count Plot of {col}")
    plt.show()
```

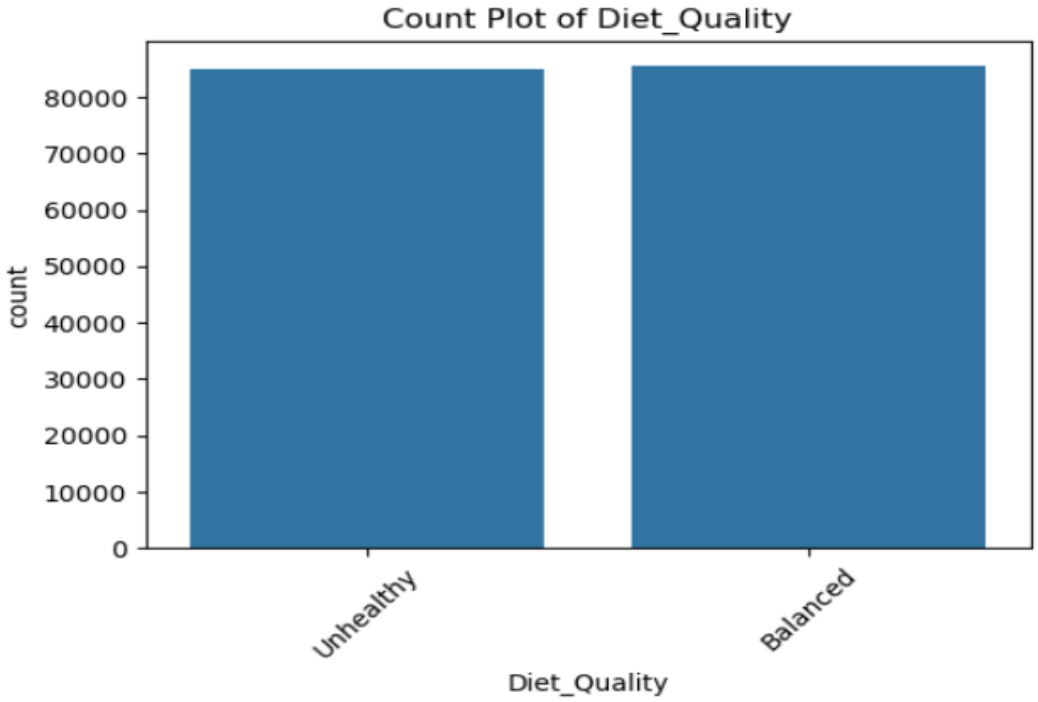




🔹 **Insight:**

✅ Shows the **distribution of categorical features** (e.g., Gender, Physical Activity, Diet Quality).

✅ Identifies **imbalances in category representation**.



🔹 **Insight:**

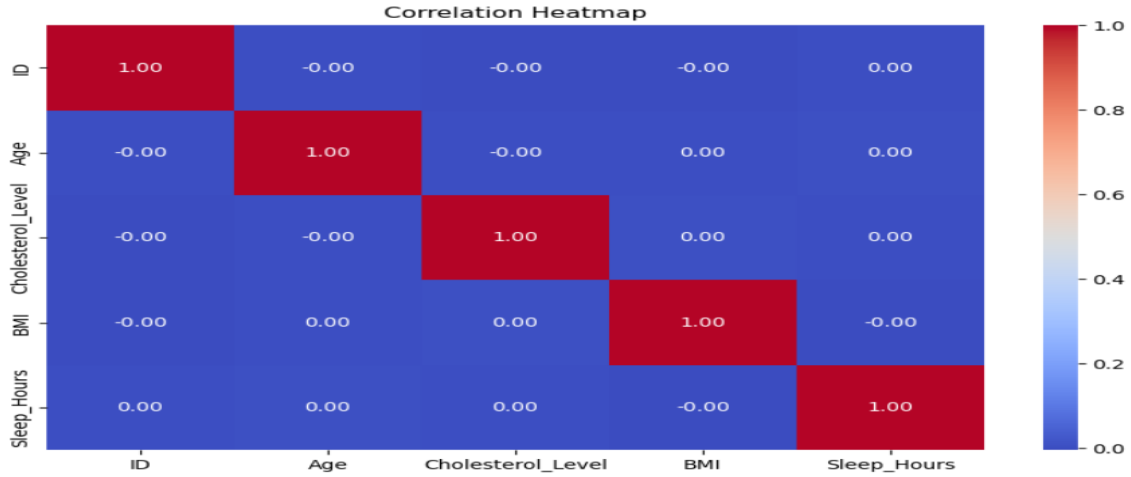
✅ Shows the **distribution of categorical features** (e.g., Gender, Physical Activity, Diet Quality).

✅ Identifies **imbalances in category representation**.

Bivariate Analysis

Correlation Heatmap

```
plt.figure(figsize=(10, 6))
sns.heatmap(indo_df.corr(numeric_only=True), annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Correlation Heatmap")
plt.show()
```



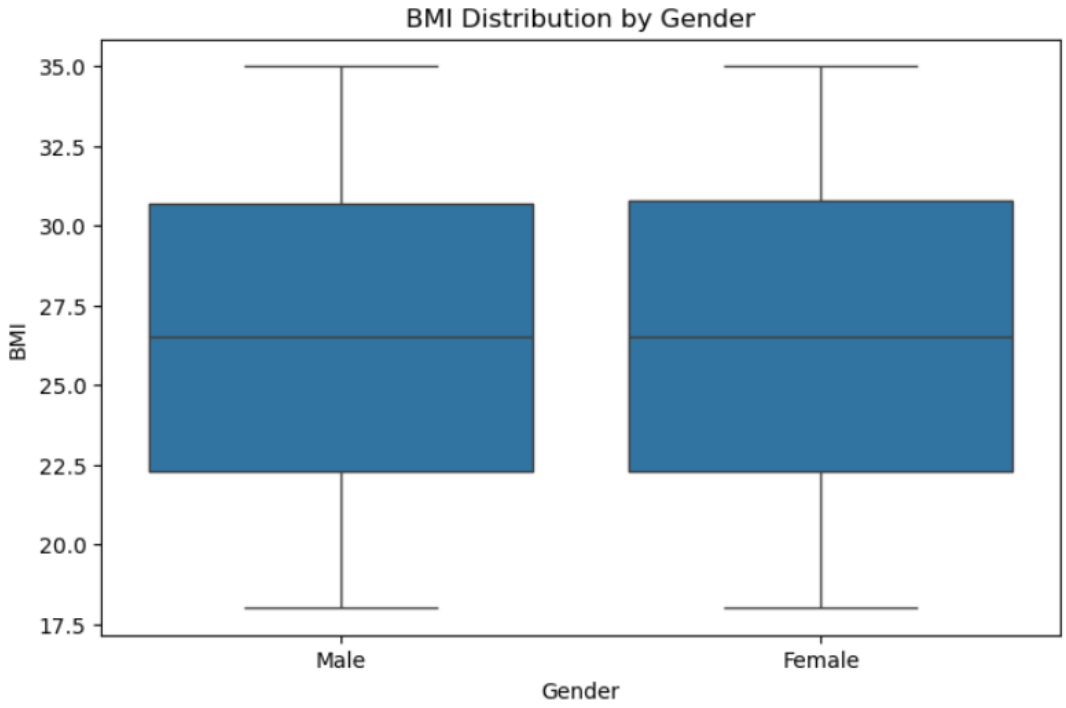
🔹 **Insight:**

✅ Identifies **relationships between numerical variables**.

✅ Highlights **strongly correlated features** that might impact heart disease.

Boxplot: BMI Distribution by Gender

```
plt.figure(figsize=(8, 5))
sns.boxplot(x=indo_df['Gender'], y=indo_df['BMI'])
plt.title("BMI Distribution by Gender")
plt.show()
```



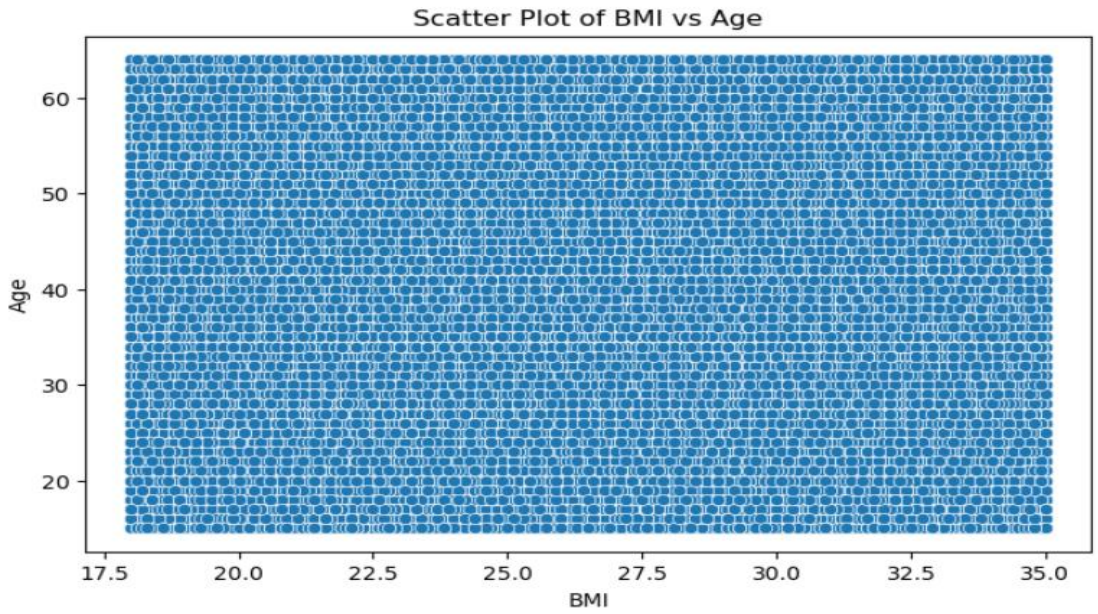
🔹 **Insight:**

✅ Analyzes **differences in BMI across genders**.

✅ Detects **potential outliers** in BMI distribution.

Scatter Plot: BMI vs. Age

```
plt.figure(figsize=(8, 5))
sns.scatterplot(x=indo_df['BMI'], y=indo_df['Age'])
plt.title("Scatter Plot of BMI vs Age")
plt.show()
```



🔹 **Insight:**

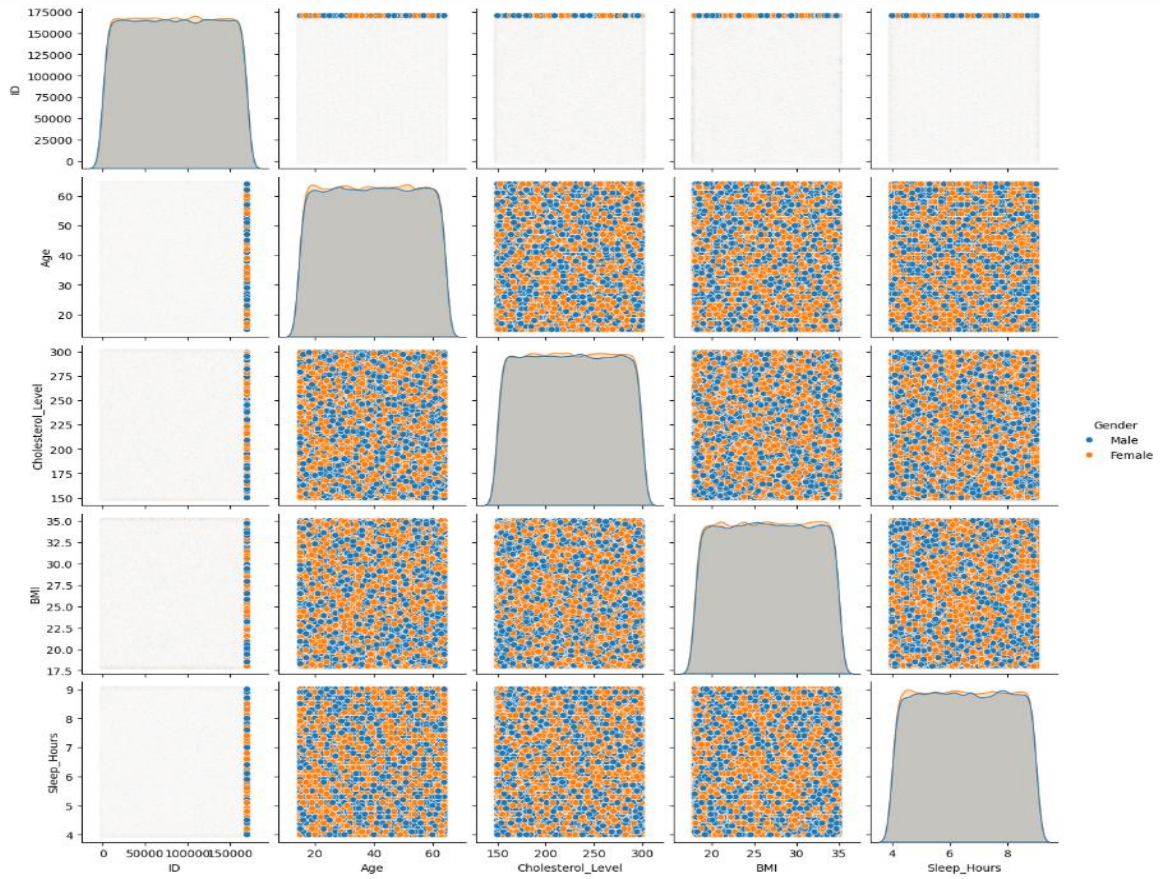
✅ Helps analyze if **BMI changes with age**.

Multivariate Analysis

Pairplot for Multiple Relationships

sns.pairplot(indo_df, hue='Gender')

plt.show()



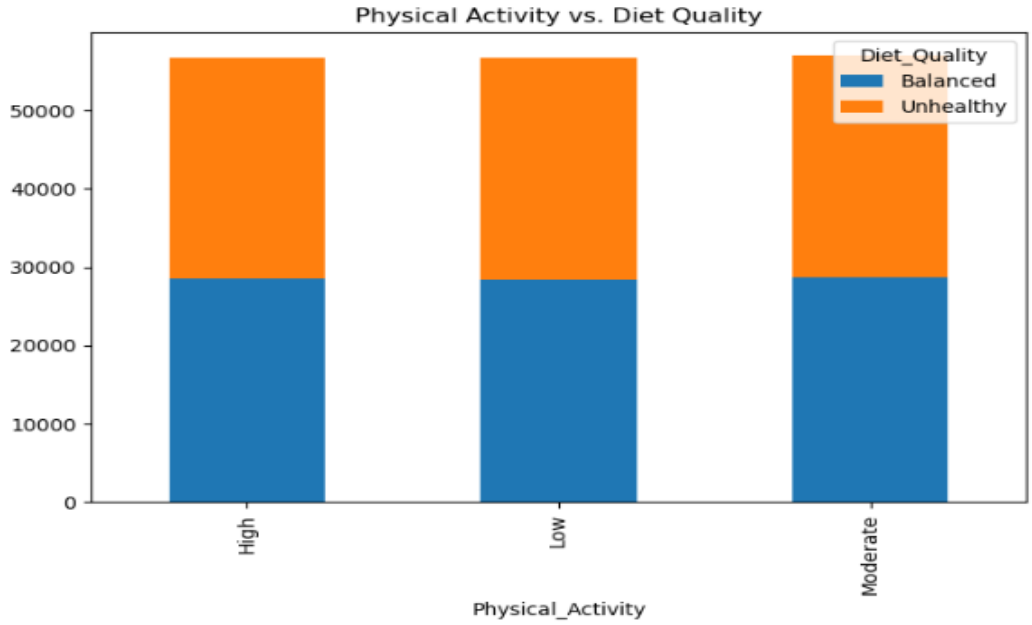
🔹 **Insight:**

✅ Helps identify **correlations and clusters** among features.

✅ Provides a **holistic view of relationships** between multiple variables.

Stacked Bar Plot: Physical Activity vs. Diet Quality

```
pd.crosstab(indo_df['Physical_Activity'], indo_df['Diet_Quality']).plot(kind="bar", stacked=True, figsize=(8, 5))
plt.title("Physical Activity vs. Diet Quality")
plt.show()
```



CHINA DATASET

Introduction

The China dataset contains heart disease-related health metrics from the Chinese population. The goal of this analysis is to uncover patterns, trends, and relationships between key health indicators, such as BMI, blood pressure, cholesterol levels, and lifestyle factors.

✅ **Key Steps in the Analysis:**

1. Univariate Analysis – Understanding the distribution of individual features.
2. Bivariate Analysis – Exploring relationships between two variables.
3. Multivariate Analysis – Analyzing multiple variables together for deeper insights.

Importing Libraries & Loading the Dataset

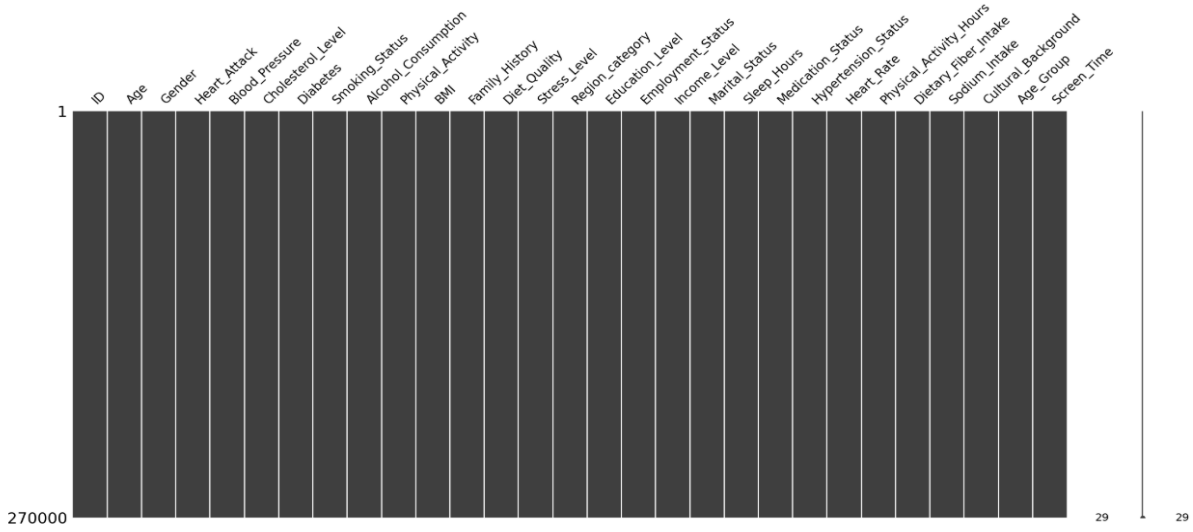
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

china_df  = pd.read_csv("C:/Users/jsuri/Downloads/H_A/cleaned_china_dataset.csv")

china_df.info()
china_df.isnull().sum()
china_df.describe()
```

🔹 **Insights:**

- **info()** helps understand data types and missing values.
- **isnull().sum()** identifies missing data.
- **describe()** provides key statistical summaries.

Missing Data Visualization

```
msno.matrix(china_df)
plt.show()
```



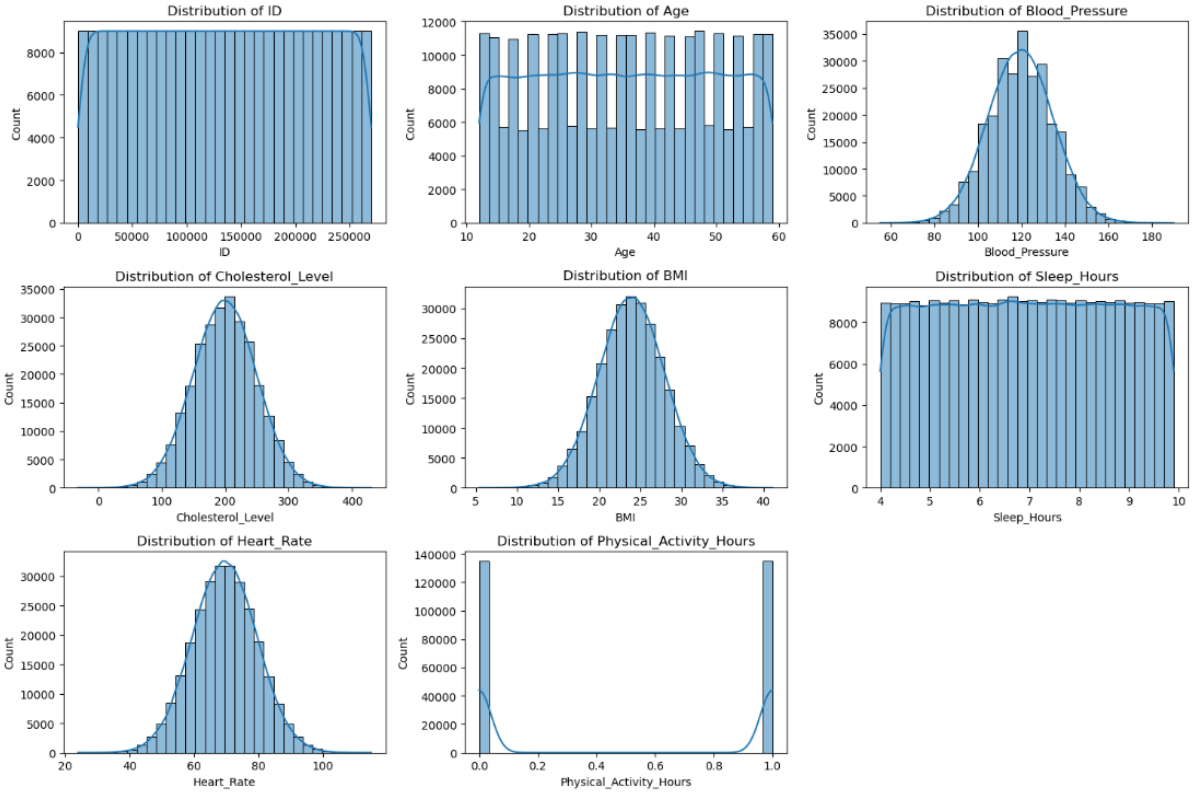
🔹 **Insight:**

✅ **Identifies missing values** across columns.

Univariate Analysis

Histograms for Numerical Variables

```
numerical_cols = china_df.select_dtypes(include=['int64', 'float64']).columns

plt.figure(figsize=(15, 10))
for i, col in enumerate(numerical_cols, 1):
    plt.subplot(3, 3, i)
    sns.histplot(china_df[col], bins=30, kde=True)
    plt.title(f"Distribution of {col}")
plt.tight_layout()
plt.show()
```



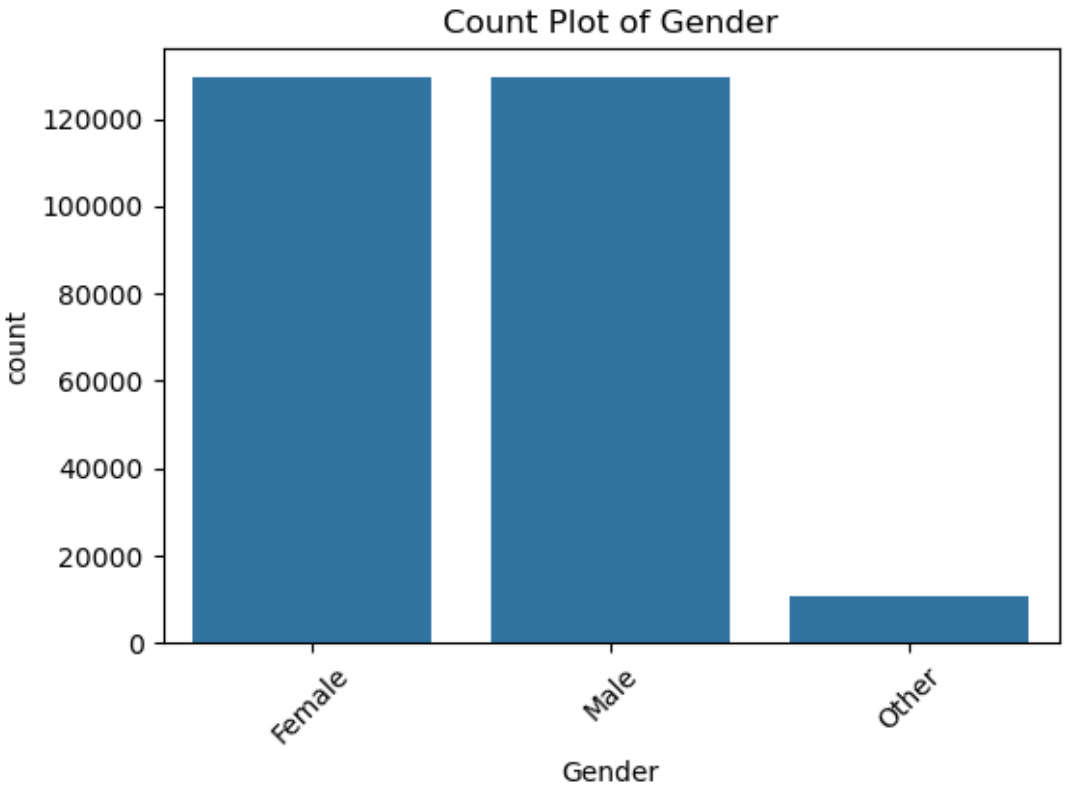
🔹 **Insight:**

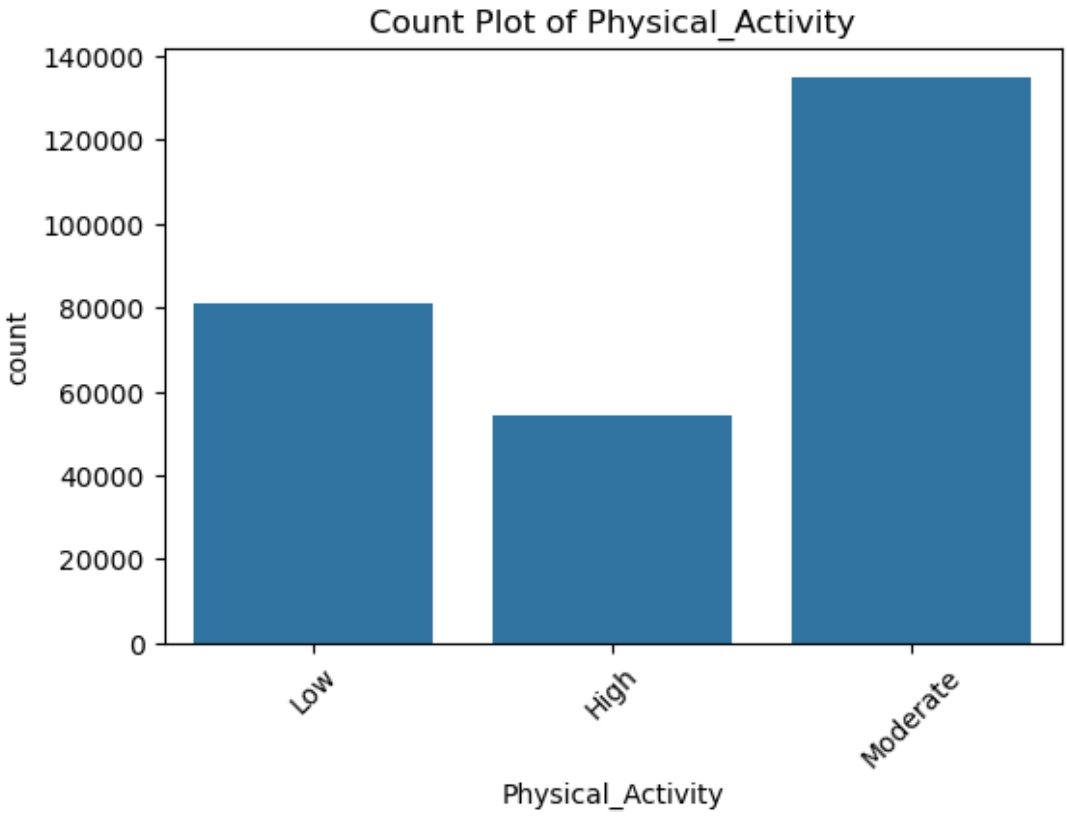
✅ Helps identify **skewness, distribution shape, and outliers** in numerical features.

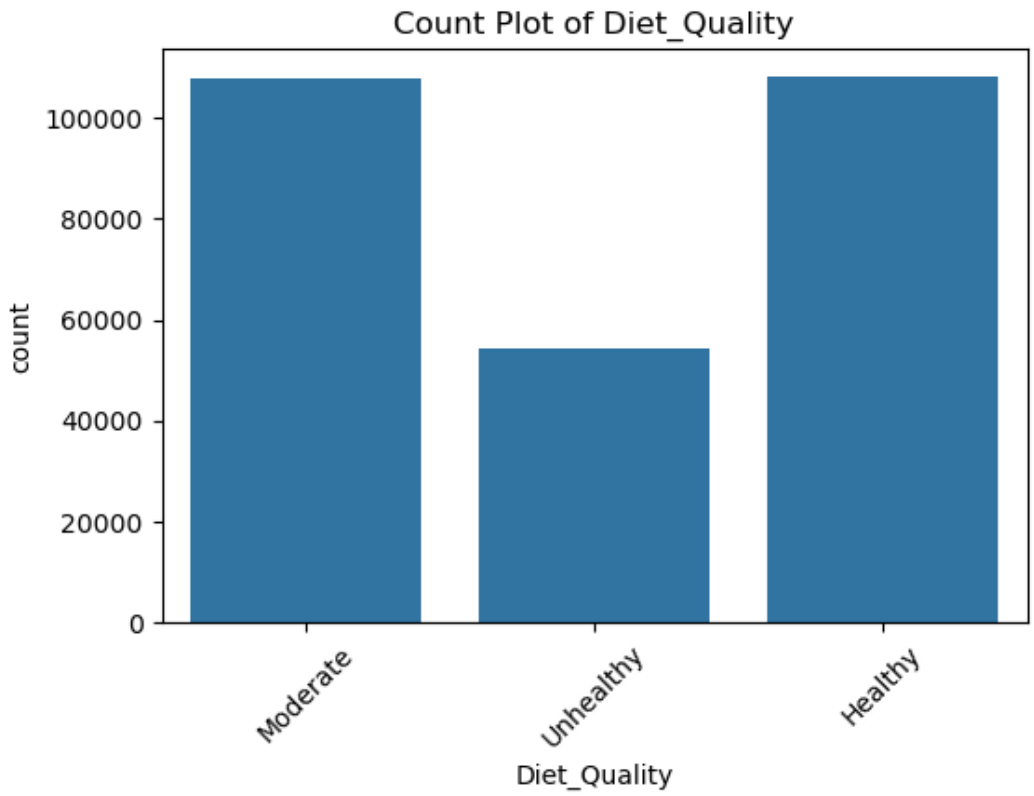
Countplots for Categorical Variables

```
categorical_cols = ['Gender', 'Physical_Activity', 'Diet_Quality']
for col in categorical_cols:
    plt.figure(figsize=(6, 4))
    sns.countplot(x=china_df[col])
```

```
plt.xticks(rotation=45)
plt.title(f"Count Plot of {col}")
plt.show()
```







🔹 **Insight:**

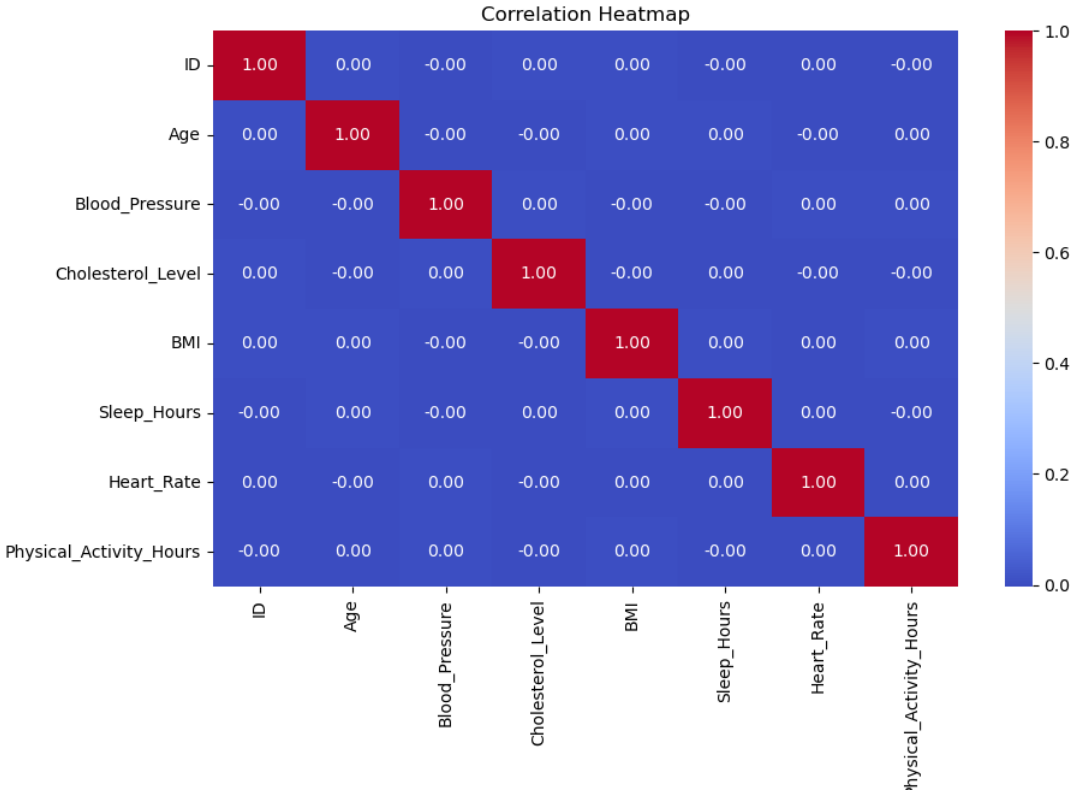
✅ Shows the **distribution of categorical features** (e.g., Gender, Physical Activity, Diet Quality).

✅ Highlights any **imbalanced categories**.

Bivariate Analysis

Correlation Heatmap

```
plt.figure(figsize=(10, 6))
sns.heatmap(china_df.corr(numeric_only=True), annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Correlation Heatmap")
plt.show()
```



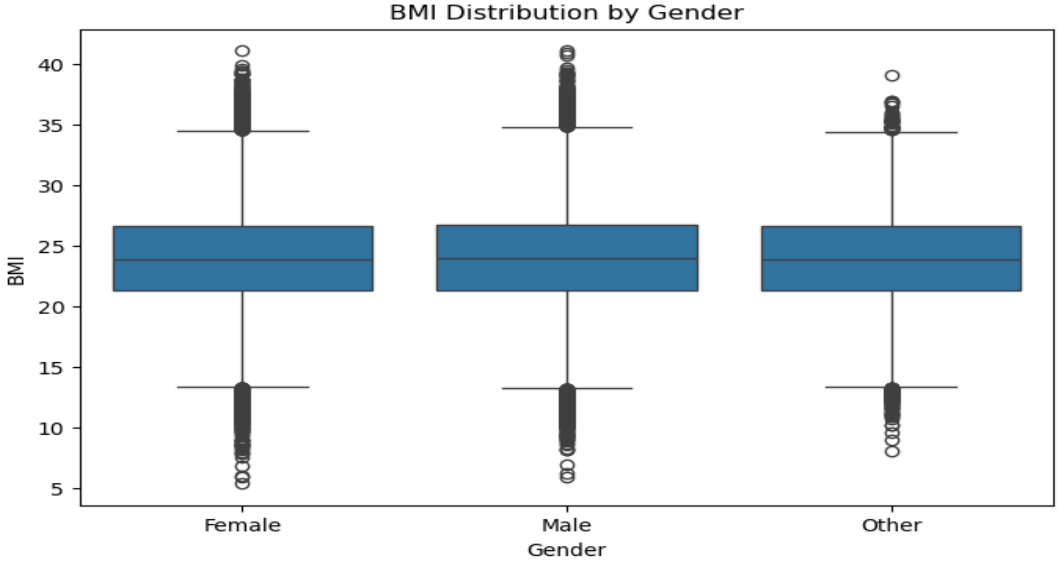
🔹 **Insight:**

✅ Identifies **relationships between numerical variables**.

✅ Helps detect **strong correlations** between health indicators.

Boxplot: BMI Distribution by Gender

```
plt.figure(figsize=(8, 5))
sns.boxplot(x=china_df['Gender'], y=china_df['BMI'])
plt.title("BMI Distribution by Gender")
plt.show()
```



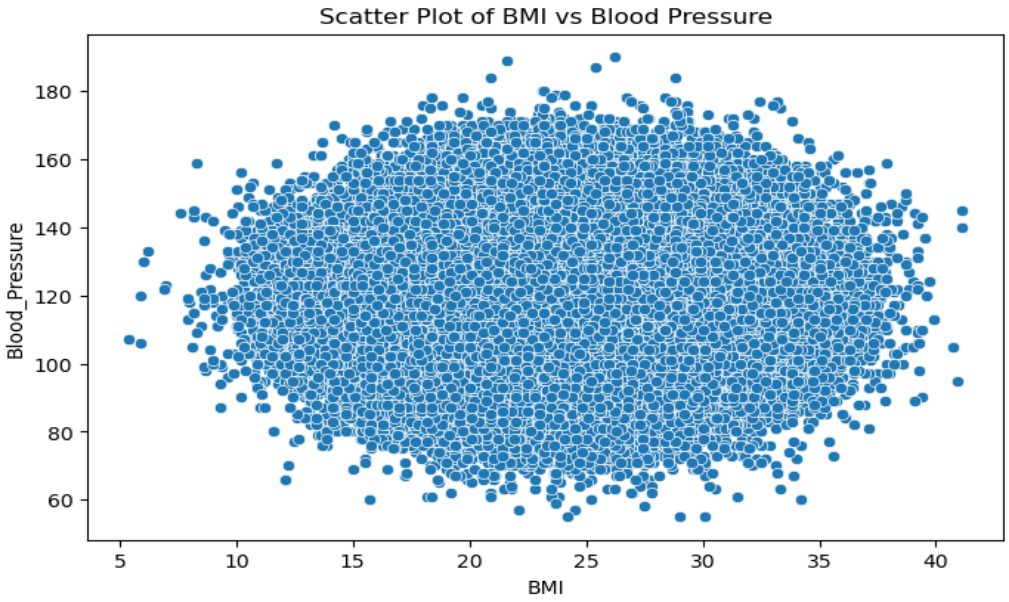
🔹 **Insight:**

✅ Compares **BMI distribution across genders**.

✅ Identifies **outliers in BMI levels**.

Scatter Plot: BMI vs. Blood Pressure

```
plt.figure(figsize=(8, 5))
sns.scatterplot(x=china_df['BMI'], y=china_df['Blood_Pressure'])
plt.title("Scatter Plot of BMI vs Blood Pressure")
plt.show()
```



NORTH AMERICA DATASET

Introduction

The **China dataset** contains heart disease-related health metrics from the Chinese population. The goal of this analysis is to uncover **patterns, trends, and relationships** between key health indicators, such as **BMI, blood pressure, cholesterol levels, and lifestyle factors**.

✅ **Key Steps in the Analysis:**

1. **Univariate Analysis** – Understanding the distribution of individual features.
2. **Bivariate Analysis** – Exploring relationships between two variables.
3. **Multivariate Analysis** – Analyzing multiple variables together for deeper insights.

Importing Libraries & Loading the Dataset

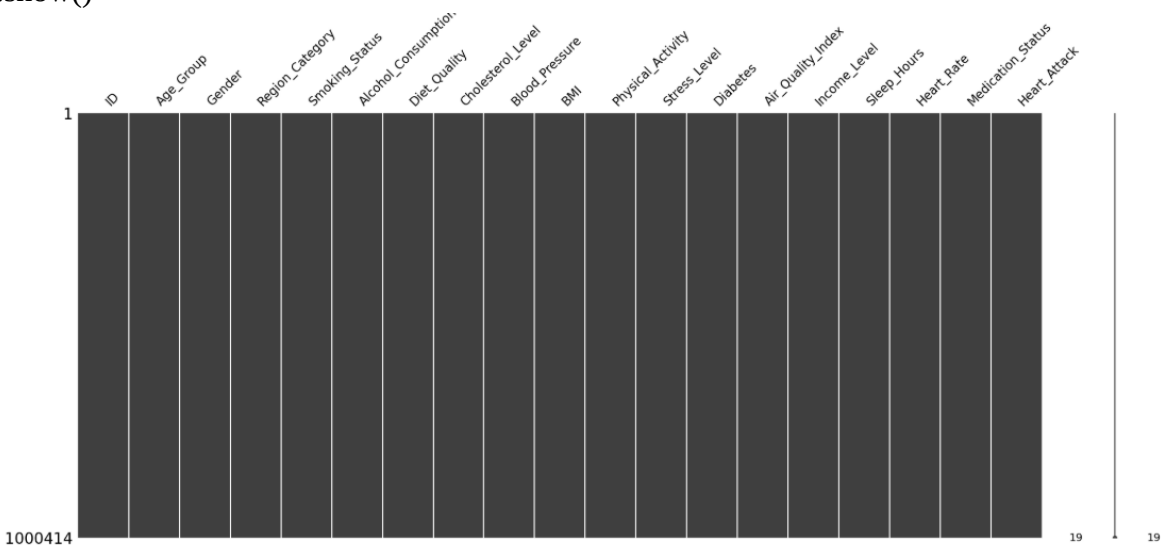
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

china_df  = pd.read_csv("C:/Users/jsuri/Downloads/H_A/cleaned_china_dataset.csv")
china_df.info()
china_df.isnull().sum()
china_df.describe()
```

🔹 **Insights:**

- **info()** helps understand data types and missing values.
- **isnull().sum()** identifies missing data.
- **describe()** provides key statistical summaries.

Missing Data Visualization

```
msno.matrix(china_df)
plt.show()
```



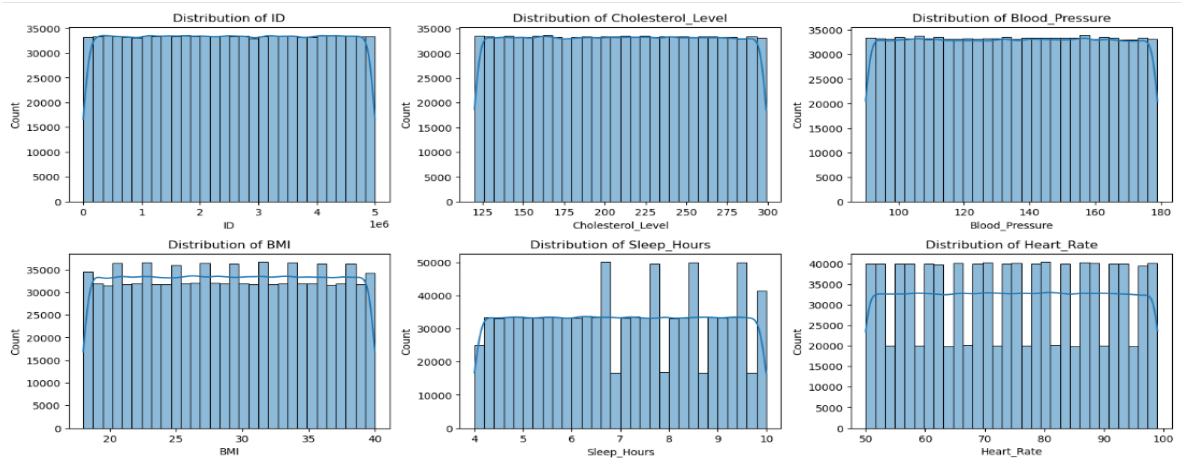
🔹 **Insight:**

✅ **Identifies missing values** across columns.

Univariate Analysis

Histograms for Numerical Variables

```
numerical_cols = china_df.select_dtypes(include=['int64', 'float64']).columns
plt.figure(figsize=(15, 10))
for i, col in enumerate(numerical_cols, 1):
    plt.subplot(3, 3, i)
    sns.histplot(china_df[col], bins=30, kde=True)
    plt.title(f"Distribution of {col}")
plt.tight_layout()
plt.show()
```

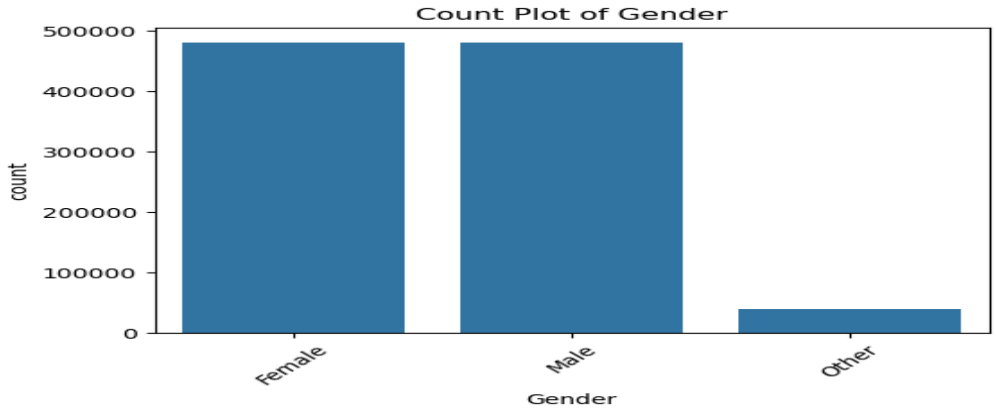


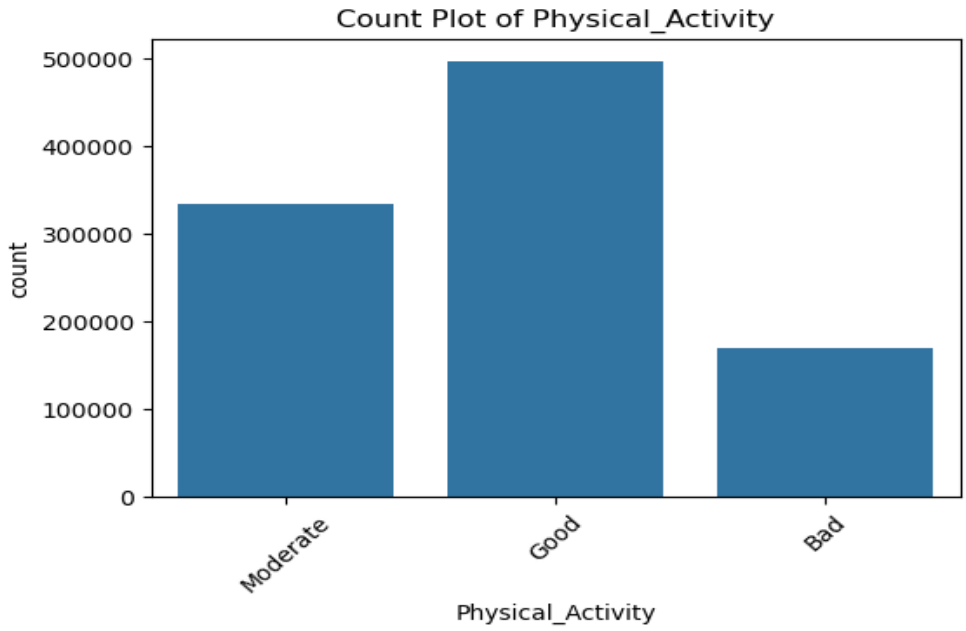
🔹 **Insight:**

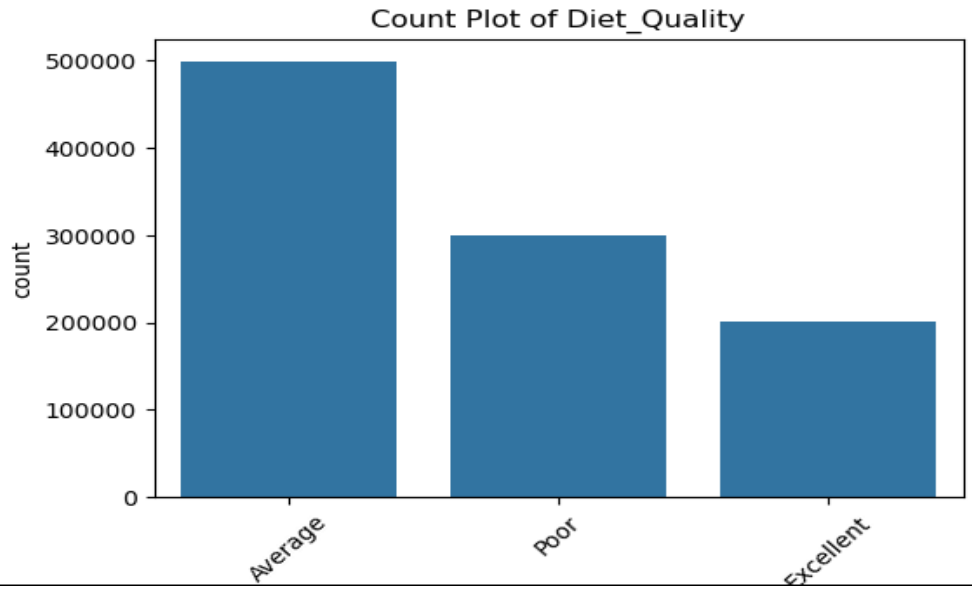
✅ Helps identify **skewness, distribution shape, and outliers** in numerical features.

Countplots for Categorical Variables

```
categorical_cols = ['Gender', 'Physical_Activity', 'Diet_Quality']
for col in categorical_cols:
    plt.figure(figsize=(6, 4))
    sns.countplot(x=china_df[col])
    plt.xticks(rotation=45)
    plt.title(f"Count Plot of {col}")
    plt.show()
```







🔹 **Insight:**

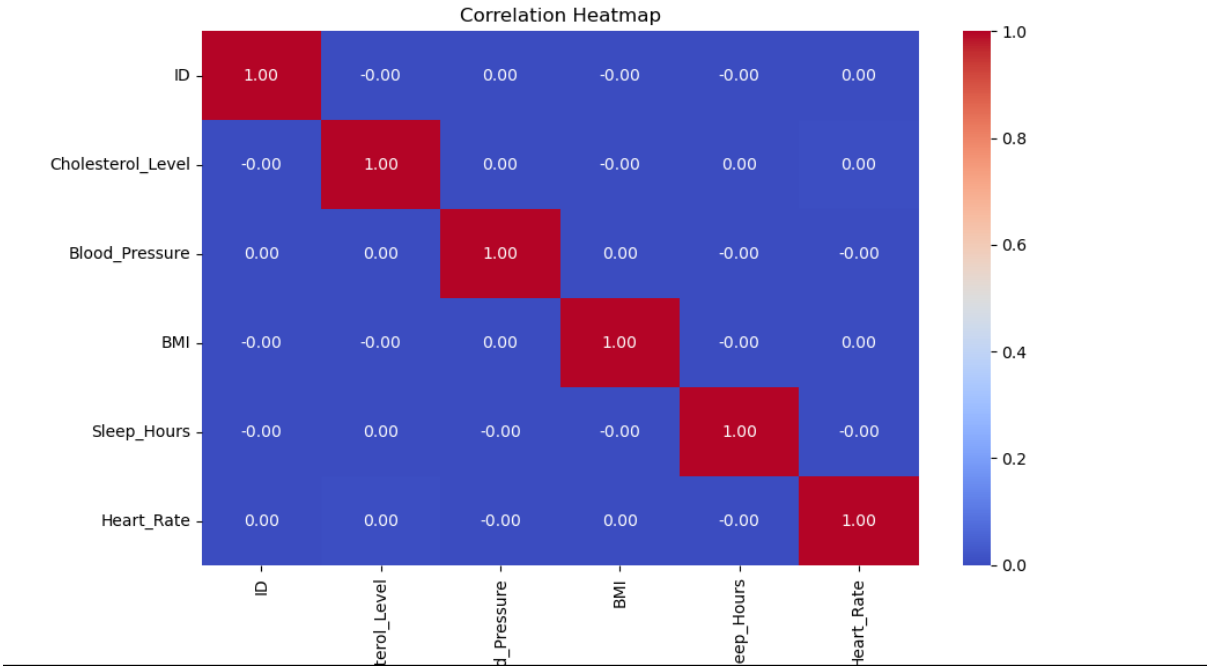
✅ Shows the **distribution of categorical features** (e.g., Gender, Physical Activity, Diet Quality).

✅ Highlights any **imbalanced categories**.

Bivariate Analysis

Correlation Heatmap

```
plt.figure(figsize=(10, 6))
sns.heatmap(china_df.corr(numeric_only=True), annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Correlation Heatmap")
plt.show()
```



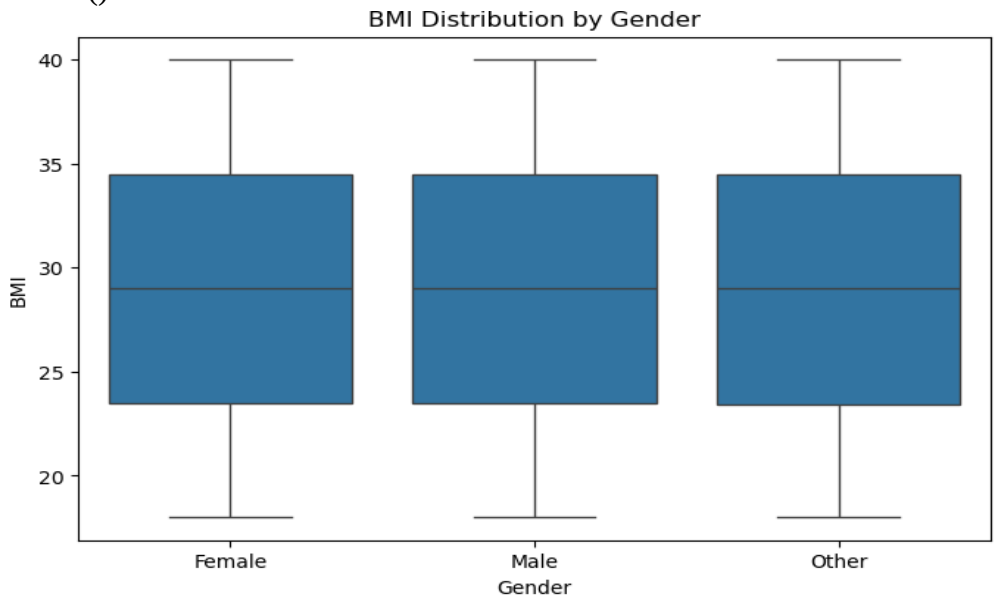
🔹 **Insight:**

✅ Identifies **relationships between numerical variables**.

✅ Helps detect **strong correlations** between health indicators.

Boxplot: BMI Distribution by Gender

```
plt.figure(figsize=(8, 5))
sns.boxplot(x=china_df['Gender'], y=china_df['BMI'])
plt.title("BMI Distribution by Gender")
plt.show()
```



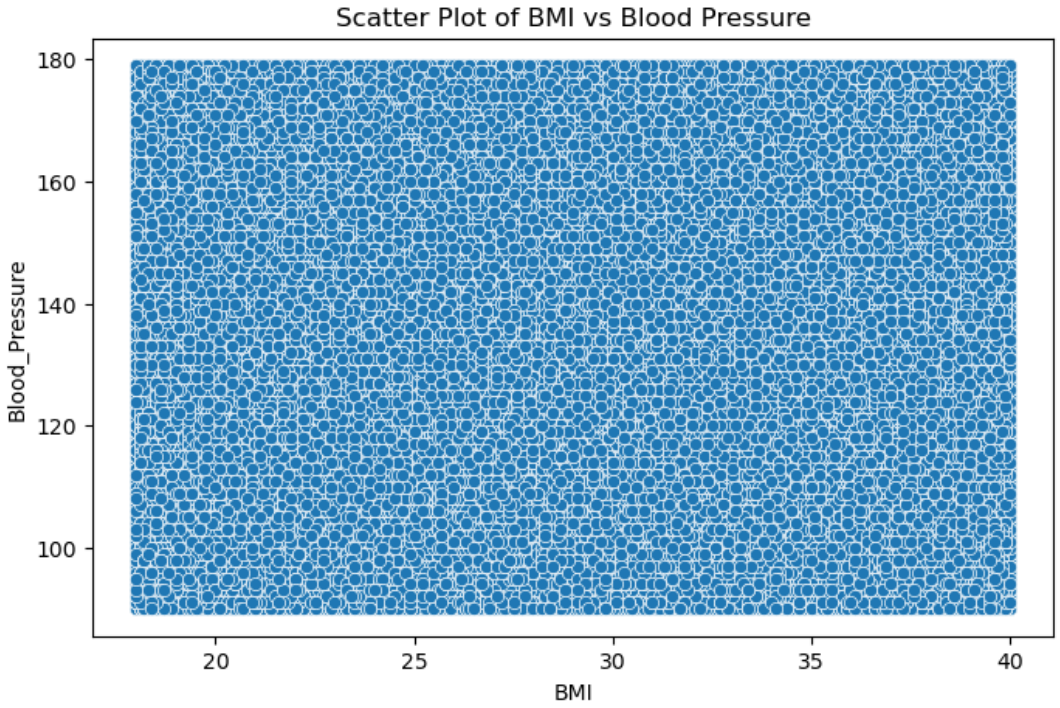
🔹 **Insight:**

✅ Compares **BMI distribution across genders**.

✅ Identifies **outliers in BMI levels**.

Scatter Plot: BMI vs. Blood Pressure

```
plt.figure(figsize=(8, 5))
sns.scatterplot(x=china_df['BMI'], y=china_df['Blood_Pressure'])
plt.title("Scatter Plot of BMI vs Blood Pressure")
plt.show()
```



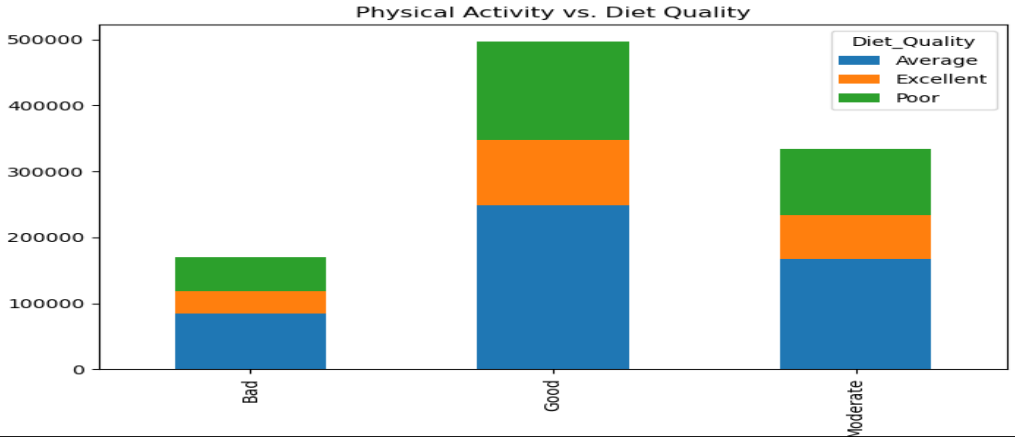
- **Insight:**

✅ Analyzes whether **higher BMI is associated with higher blood pressure**.

Multivariate analysis

Stacked Bar Plot: Physical Activity vs. Diet Quality

```
pd.crosstab(na_df['Physical_Activity'], na_df['Diet_Quality']).plot(kind="bar", stacked=True, figsize=(8, 5))
plt.title("Physical Activity vs. Diet Quality")
plt.show()
```



✅ **Conclusion**

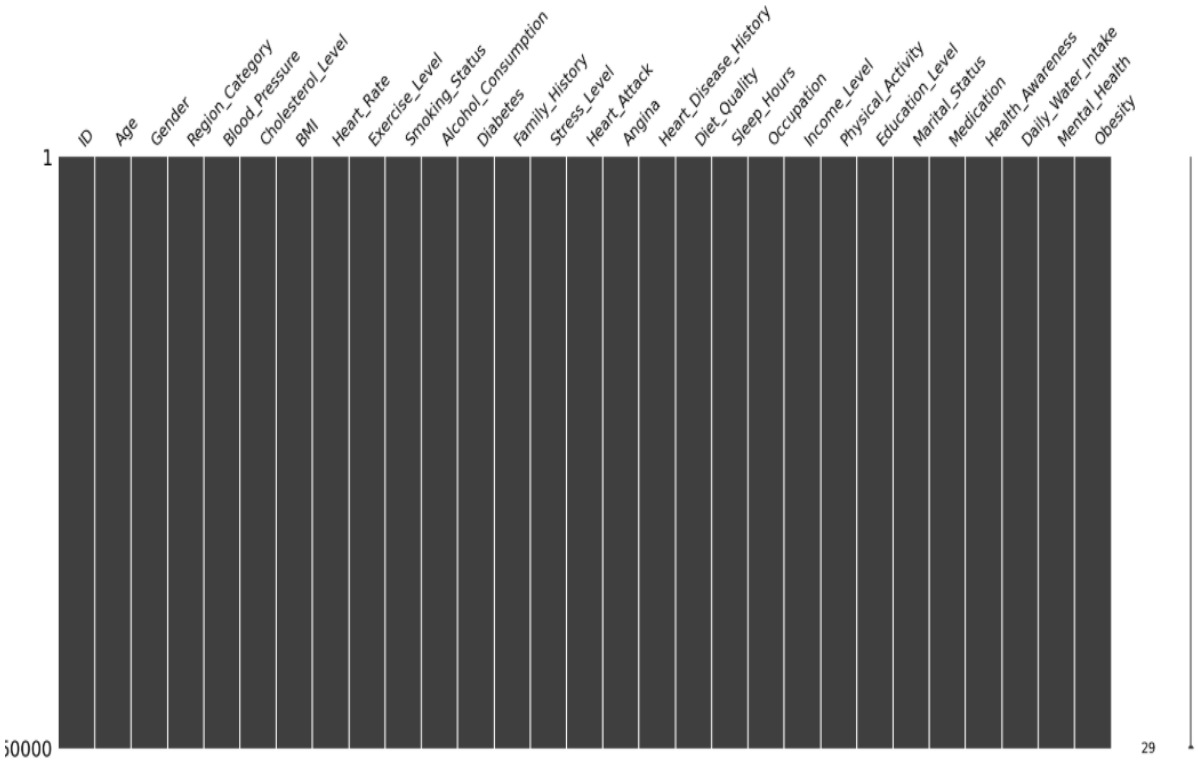
- 🔹 **Univariate analysis** provided insights into the distribution of individual features.
- 🔹 **Bivariate analysis** identified relationships between key variables such as BMI and blood pressure.
- 🔹 **Multivariate analysis** highlighted trends between physical activity and diet quality.

RUSSIA DATASET

1. Introduction

The Russia dataset consists of various health-related attributes that help analyze heart attack risk factors among individuals in the region. This analysis aims to uncover trends in age, BMI, blood pressure, physical activity, diet quality, and other factors affecting heart health.

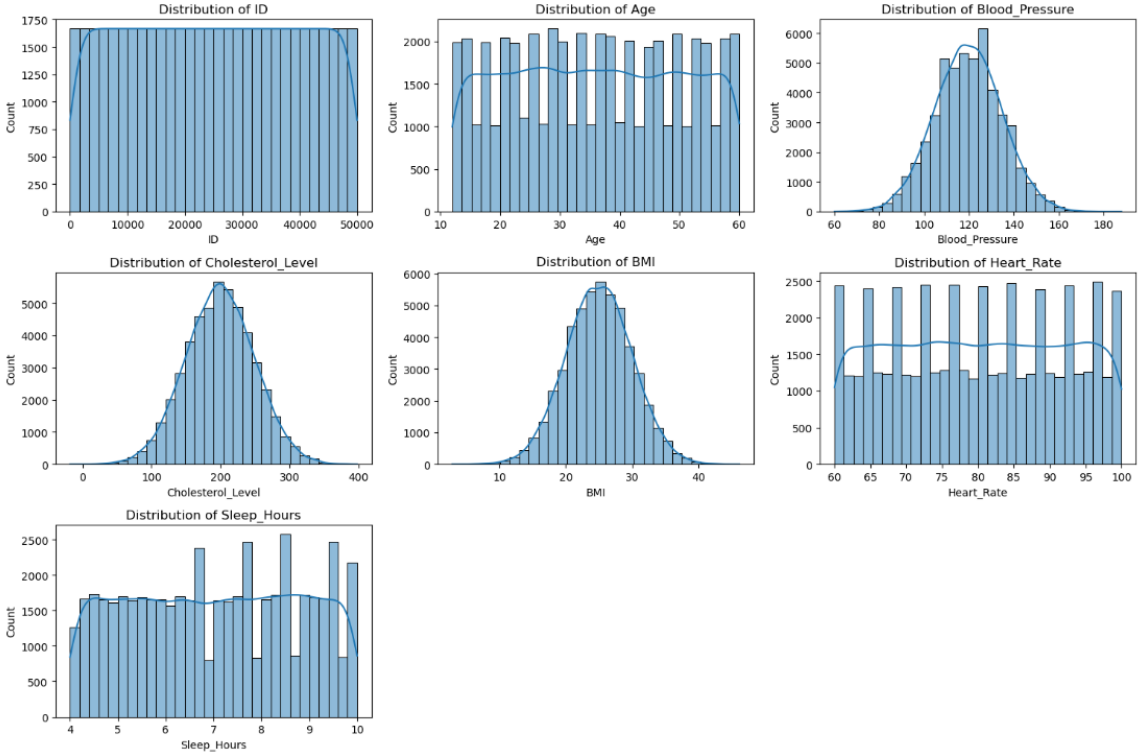
```
import pandas as pd
 import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
ru_df= pd.read_csv("C:/Users/jsuri/Downloads/H_A/cleaned_russia_dataset.csv")
ru_df
 ru_df.info()
ru_df.describe()
msno.matrix(ru_df)
 plt.show()
```



Histogram for Numerical Columns

```
numerical_cols = ru_df.select_dtypes(include=['int64', 'float64']).columns
plt.figure(figsize=(15, 10))
for i, col in enumerate(numerical_cols, 1):
    plt.subplot(3, 3, i)
```

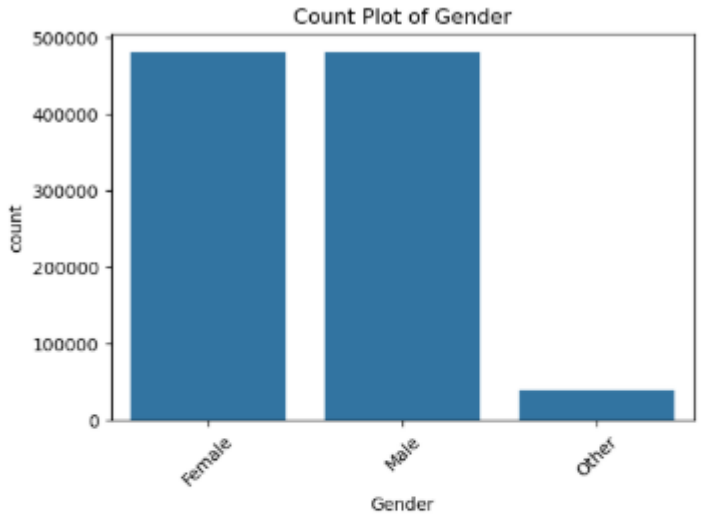
```
    sns.histplot(ru_df[col], bins=30, kde=True)
    plt.title(f"Distribution of {col}")
plt.tight_layout()
plt.show()
```

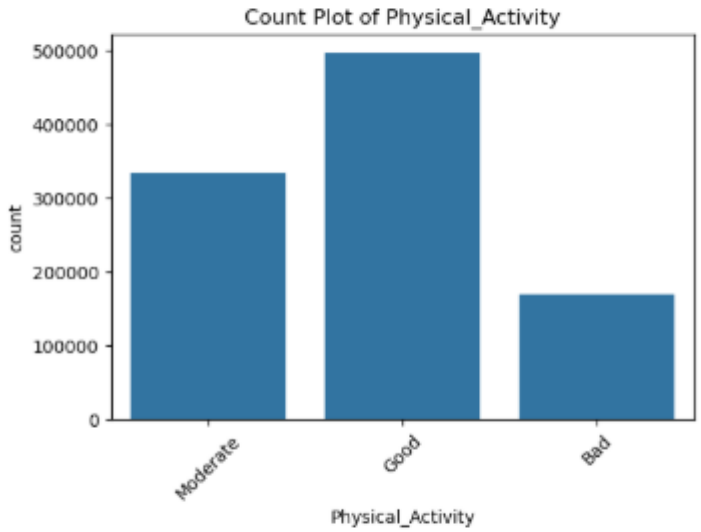


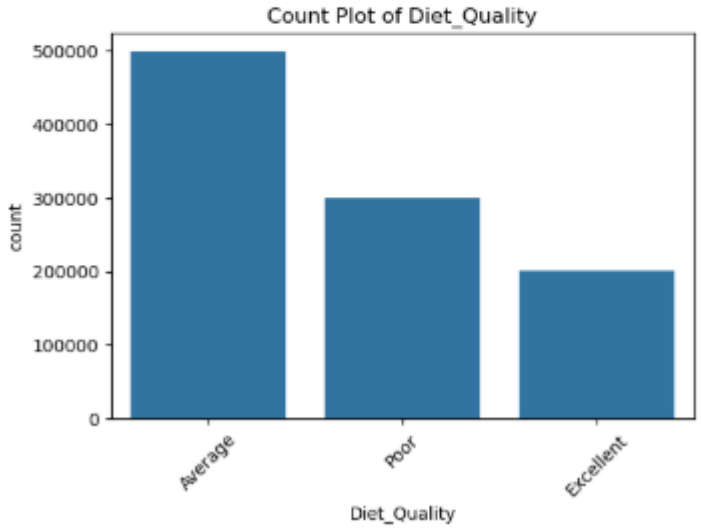
Count Plots for Categorical Variables

```
categorical_cols = ['Gender', 'Physical_Activity', 'Diet_Quality']  # Update with correct column names
for col in categorical_cols:
    plt.figure(figsize=(6, 4))
    sns.countplot(x=ru_df[col])
    plt.xticks(rotation=45)
    plt.title(f"Count Plot of {col}")
```

plt.show()



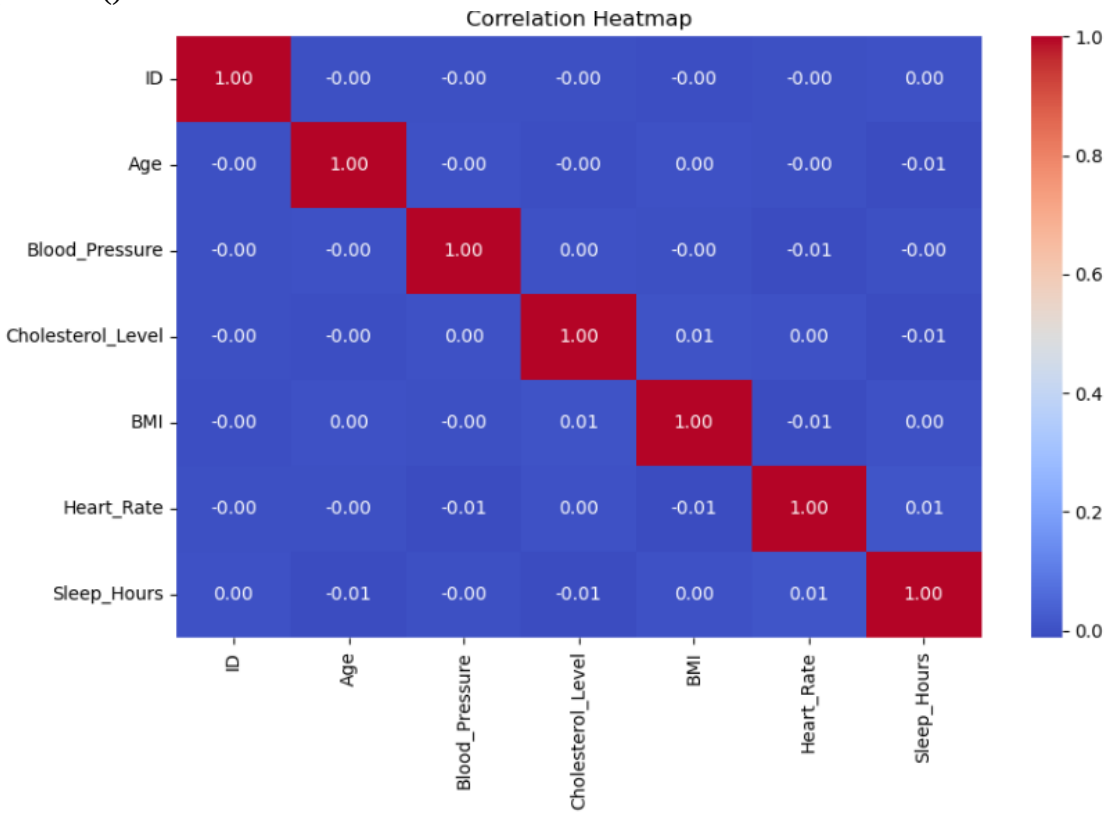




Bivariate Analysis (Two Variables Analysis)
Correlation Heatmap:

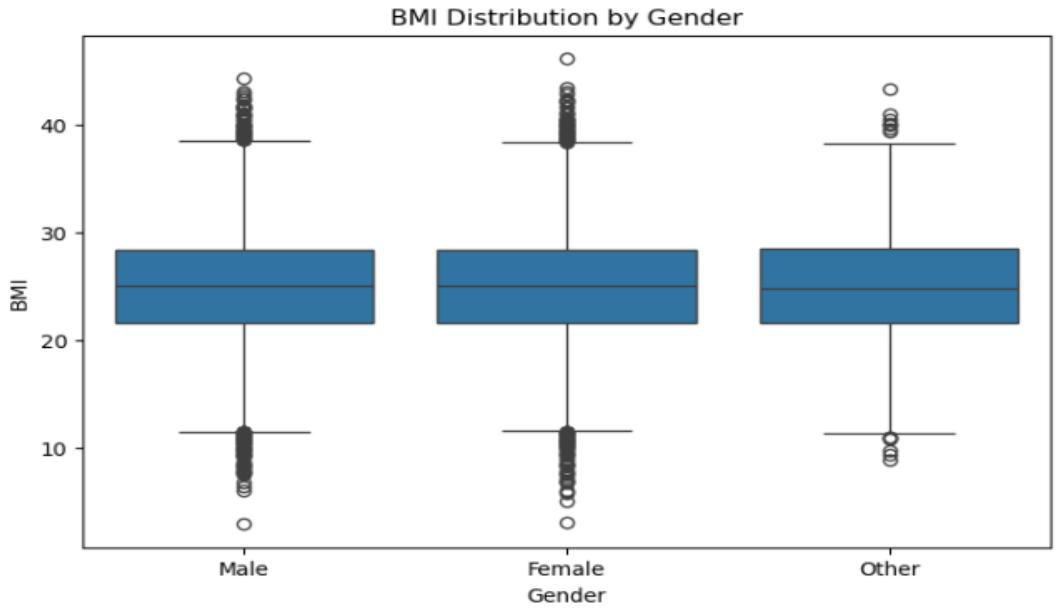
plt.figure(figsize=(10, 6))

```
sns.heatmap(ru_df.corr(numeric_only=True), annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Correlation Heatmap")
plt.show()
```



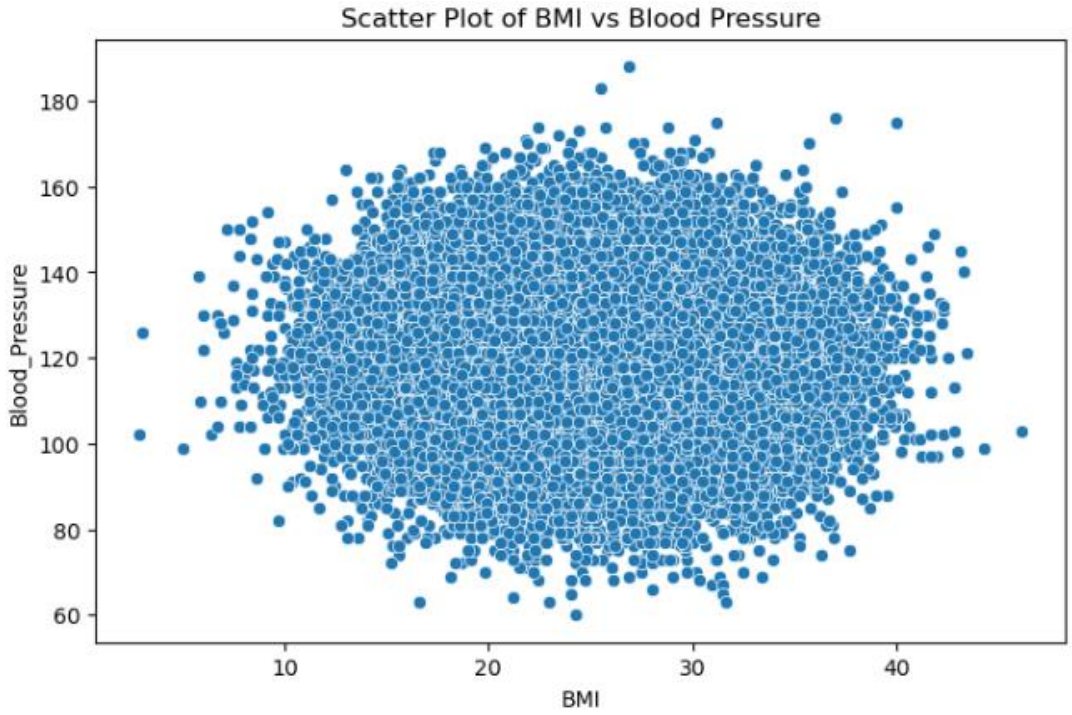
Boxplot: Gender vs BMI

```
plt.figure(figsize=(8, 5))
sns.boxplot(x=ru_df['Gender'], y=ru_df['BMI'])
plt.title("BMI Distribution by Gender")
plt.show()
```



Scatter Plot: BMI vs Blood Pressure

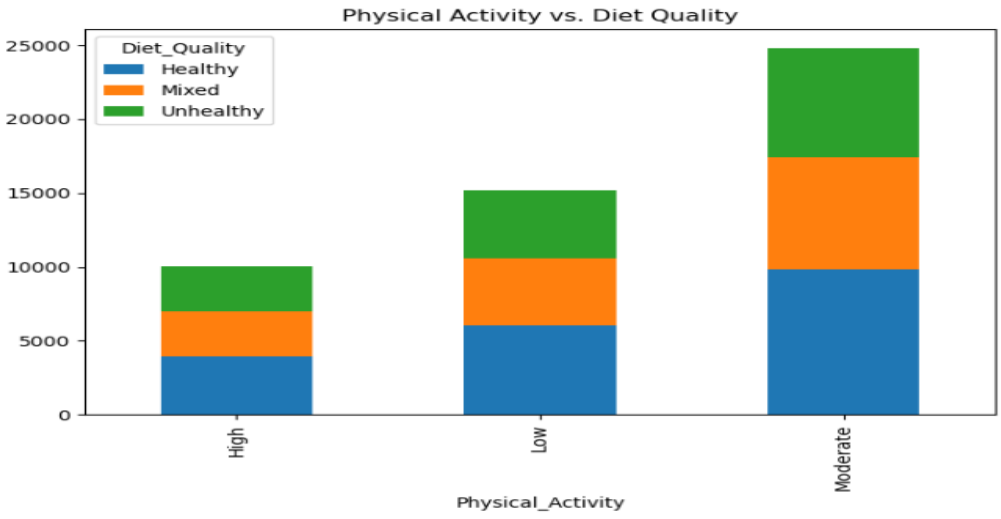
```
plt.figure(figsize=(8, 5))
sns.scatterplot(x=ru_df['BMI'], y=ru_df['Blood_Pressure'])
plt.title("Scatter Plot of BMI vs Blood Pressure")
plt.show()
```



Multivariate Analysis (Multiple Variables Analysis)

Stacked Bar Plot: Physical Activity vs. Diet Quality

```
pd.crosstab(ru_df['Physical_Activity'], ru_df['Diet_Quality']).plot(kind="bar", stacked=True, figsize=(8, 5))
plt.title("Physical Activity vs. Diet Quality")
plt.show()
```



FRANCE DATASET

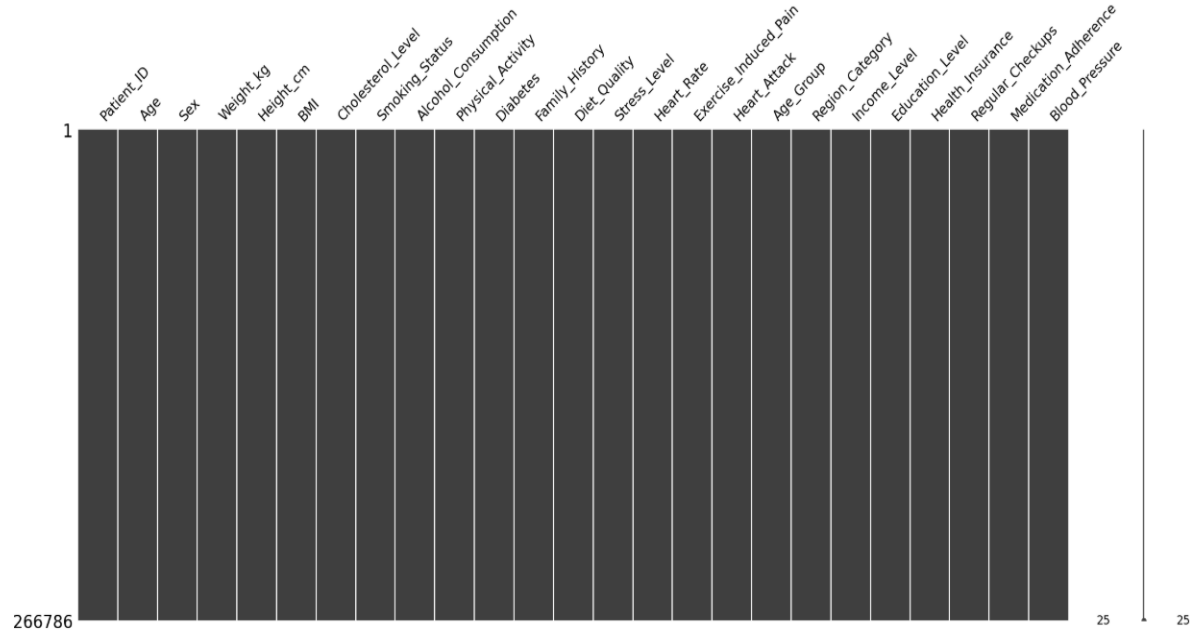
Introduction:

The France dataset provides valuable insights into health-related factors such as BMI, Blood Pressure, Physical Activity, Diet Quality, Cholesterol Levels, and more, which play a crucial role in assessing heart disease risk. By performing Exploratory Data Analysis (EDA), we aim to identify patterns, trends, and relationships that can help in understanding the prevalence of heart-related issues in the French population.

This analysis will cover:

- Univariate Analysis - Understanding the distribution of individual variables.
- Bivariate Analysis - Examining relationships between two variables.
- Multivariate Analysis - Exploring the combined effects of multiple factors.

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

fr_df = pd.read_csv("C:/Users/jsuri/Downloads/H_A/cleaned_france_dataset.csv")
fr_df
fr_df.info()
fr_df.describe()


msno.matrix(fr_df)
plt.show()
```

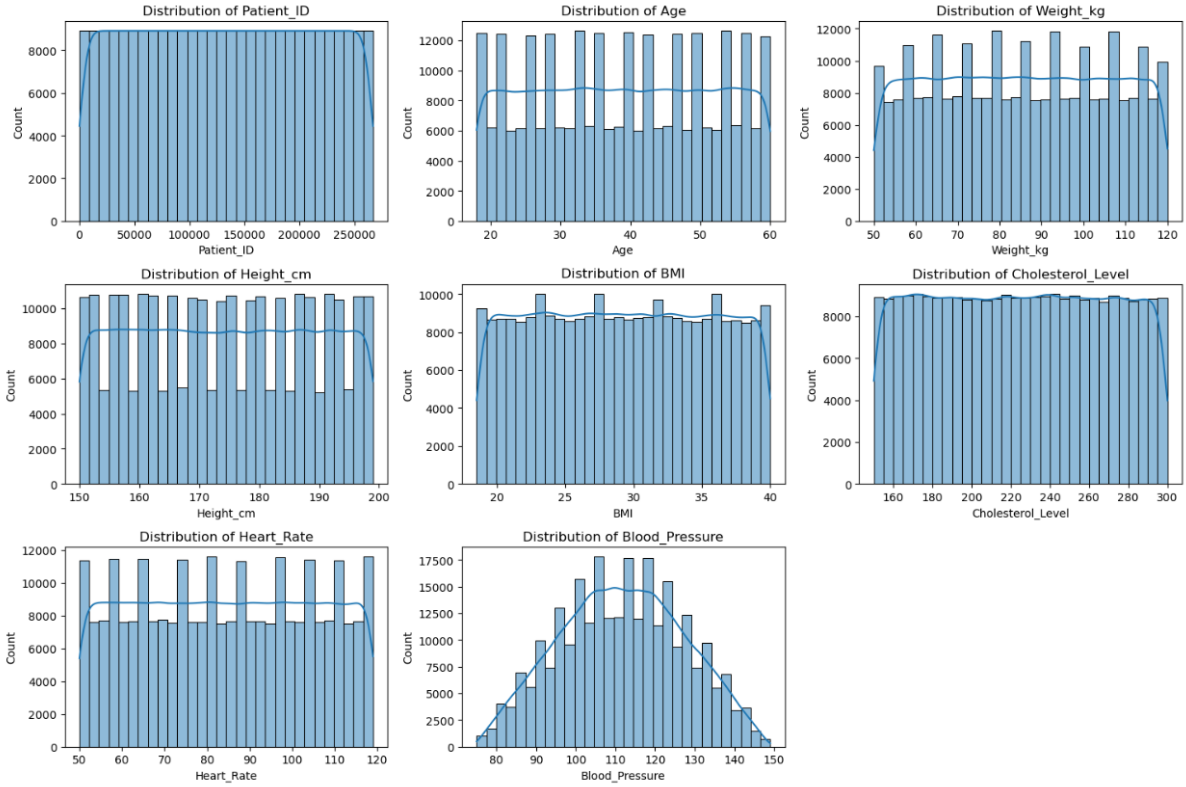


Univariate analysis

Histogram for numerical columns

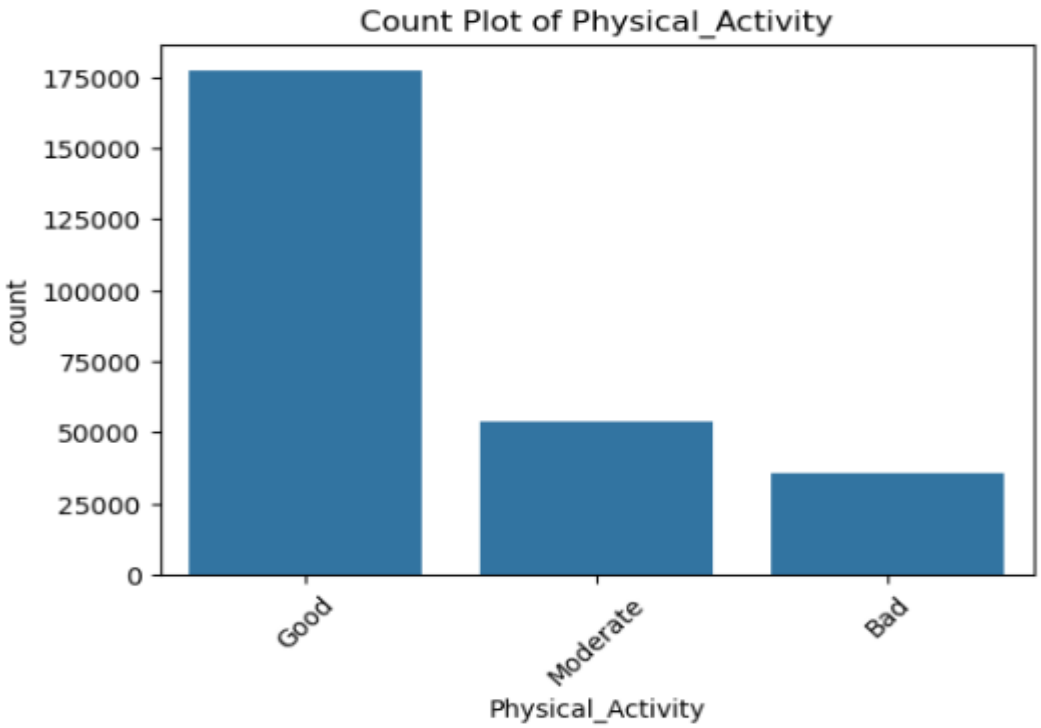
```
numerical_cols = fr_df.select_dtypes(include=['int64', 'float64']).columns
plt.figure(figsize=(15, 10))
```

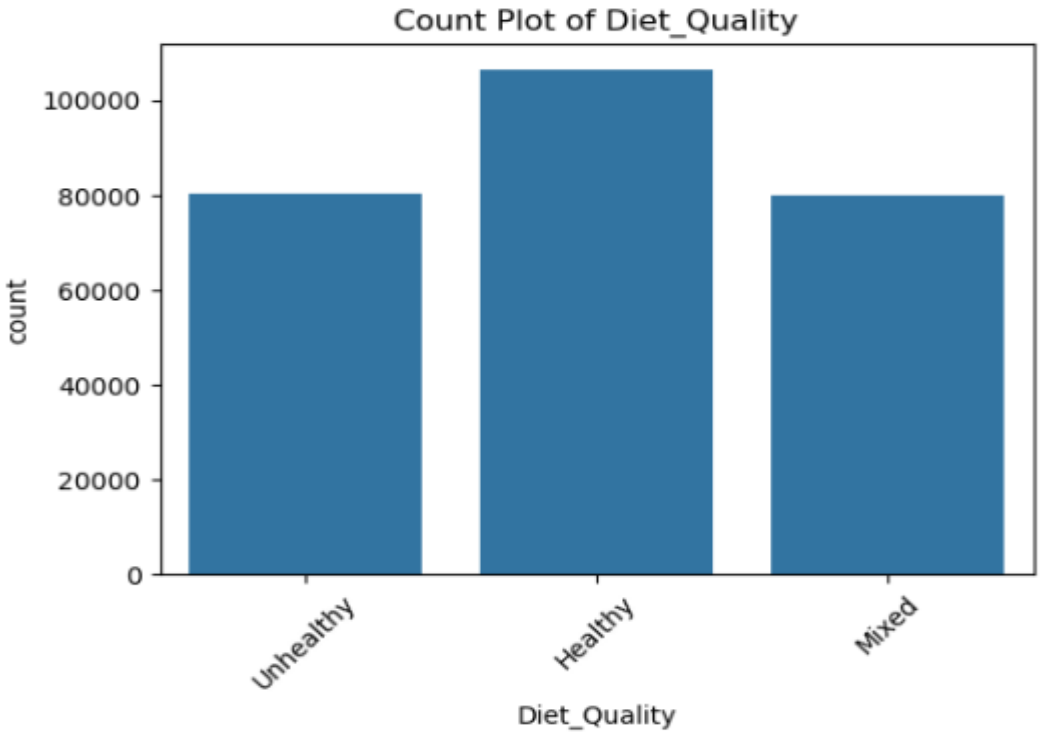
```
for i, col in enumerate(numerical_cols, 1):
    plt.subplot(3, 3, i)
    sns.histplot(fr_df[col], bins=30, kde=True)
    plt.title(f"Distribution of {col}")
plt.tight_layout()
plt.show()
```



Count plots for categorical variables

```
categorical_cols = ['Physical_Activity', 'Diet_Quality']
for col in categorical_cols:
    plt.figure(figsize=(6, 4))
    sns.countplot(x=fr_df[col])
    plt.xticks(rotation=45)
    plt.title(f"Count Plot of {col}")
    plt.show()
```

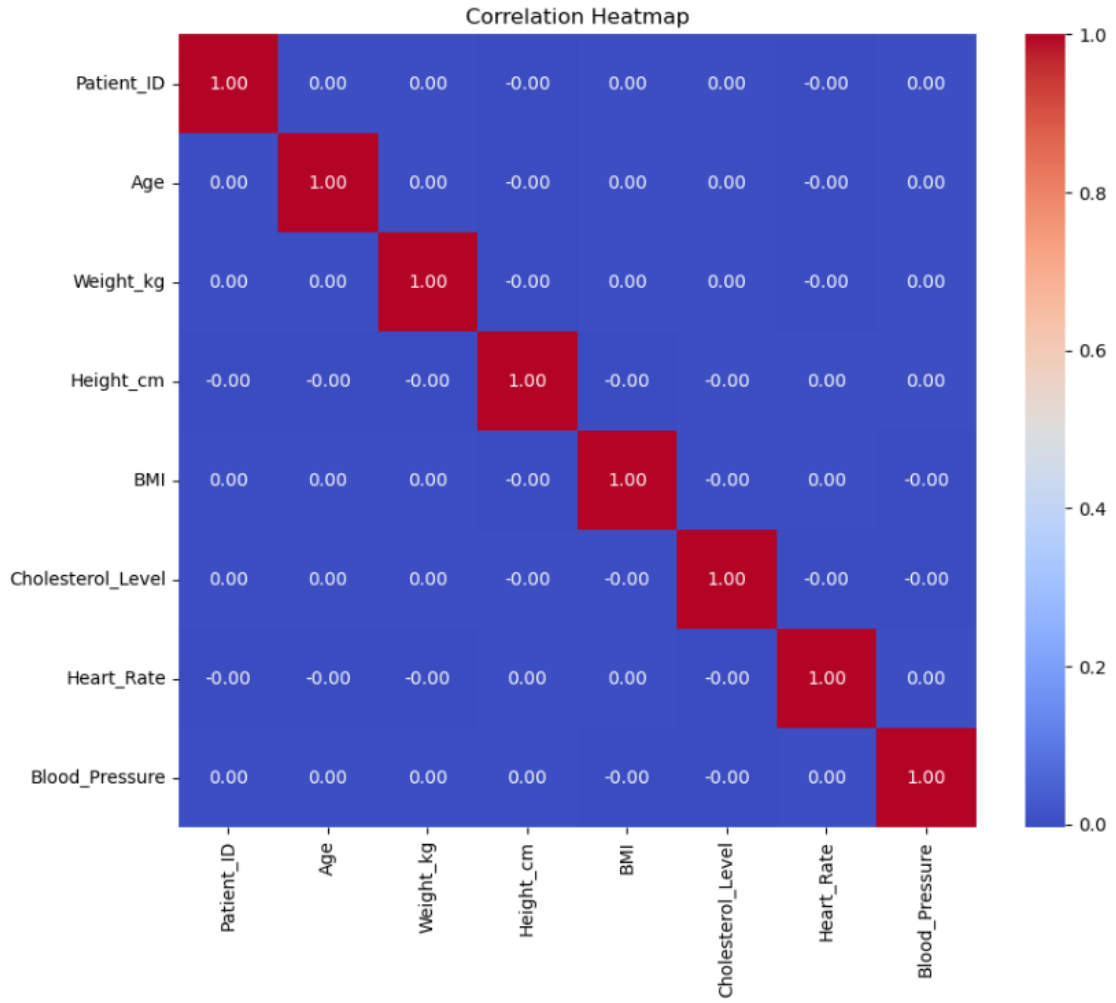




Bivariate analysis

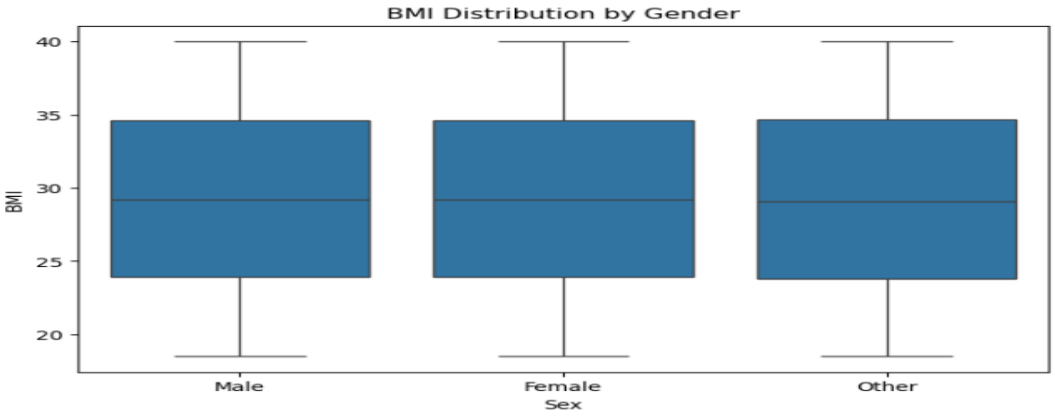
Correlation matrix

```
plt.figure(figsize=(10, 8))
sns.heatmap(fr_df.corr(numeric_only=True), annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Correlation Heatmap")
plt.show()
```



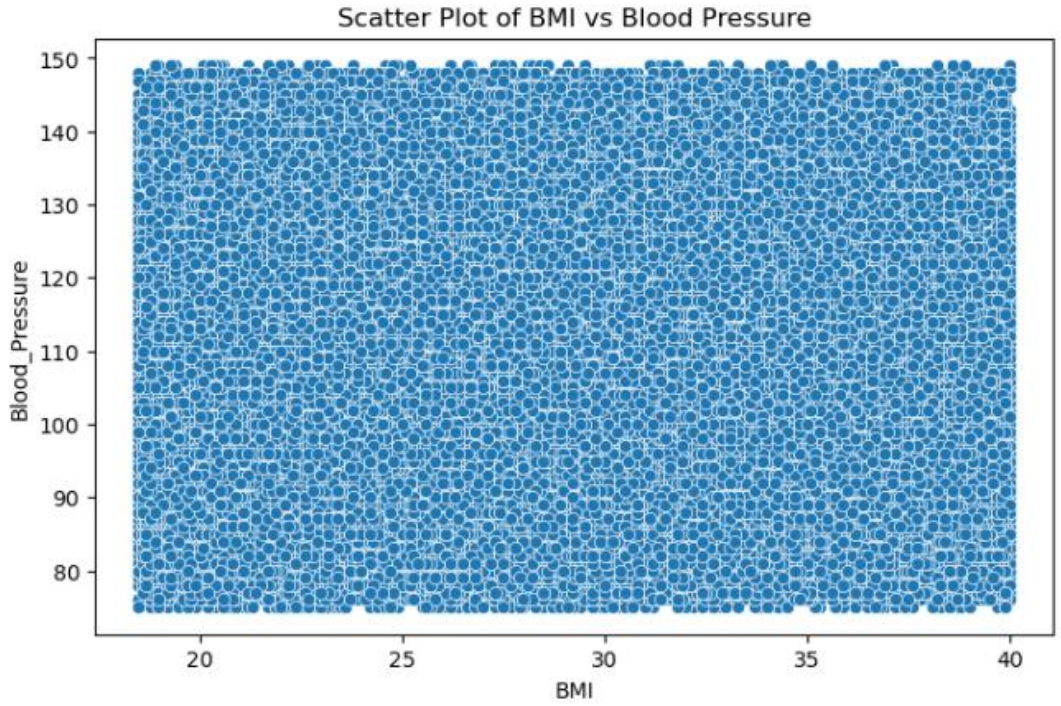
Boxplot of Age vs. BMI

```
plt.figure(figsize=(8, 5))
sns.boxplot(x=fr_df['Sex'], y=fr_df['BMI'])
plt.title("BMI Distribution by Gender")
plt.show()
```



Scatter plot : BMI vs. Blood Pressure

```
plt.figure(figsize=(8, 5))
sns.scatterplot(x=fr_df['BMI'], y=fr_df['Blood_Pressure'])
plt.title("Scatter Plot of BMI vs Blood Pressure")
plt.show()
```



Multivariate analysis

Stacked bar plot: Physical Activity vs. Diet Quality

```
pd.crosstab(fr_df['Physical_Activity'], fr_df['Diet_Quality']).plot(kind="bar", stacked=True, figsize=(8, 5))
plt.title("Physical Activity vs. Diet Quality")
plt.show()
```