**Name:** Shubham Dahane  
**Class:** CSE(AI) B div  
**Roll no.:** 282015  
**PRN No.:** 22310315  
**Subject:** Machine Learning

**Assignment 3 – Data Visualization**

**Objective**

The objective of this assignment is to explore and interpret the heart disease dataset used in Assignments 1 and 2 using various data visualization techniques. The visualizations aim to uncover patterns, correlations, and distributions within the dataset using graphs such as Scatter Plot, Bar Chart, Box Plot, Pie Chart, Line Chart, and more.

**Dataset Overview**

The dataset contains medical information on patients, such as age, sex, chest pain type, cholesterol levels, resting blood pressure, thalassemia test results, and whether or not the patient has heart disease (target variable AHD).

**Preprocessing Steps**

* Dropped irrelevant column: Unnamed: 0.
* Filled missing values in Ca with median and in Thal with mode.
* Categorical variables like ChestPain, Thal, and AHD were label encoded.
* Numerical features were standardized using StandardScaler.
* A Random Forest Classifier was trained to validate data quality.
* Performance metrics like accuracy, confusion matrix, and classification report were evaluated.

**Visualizations**

**1. Scatter Plot**

sns.scatterplot(data=df, x='Age', y='Chol', hue='AHD')

**Purpose:**  
Displays the relationship between Age and Cholesterol with respect to Heart Disease status. It helps identify how cholesterol levels vary across different age groups and how that relates to the presence or absence of disease.

**2. Bar Chart**

sns.countplot(x='ChestPain', data=df, hue='AHD')

**Purpose:**  
Illustrates the frequency of different chest pain types and how they are distributed between heart disease positive and negative patients. Helps in spotting which chest pain types are more likely linked with heart issues.

**3. Box Plot**

sns.boxplot(x='AHD', y='Oldpeak', data=df)

**Purpose:**  
Shows the distribution and spread of ST depression (Oldpeak) across heart disease statuses, highlighting the presence of outliers and differences in medians.

**4. Pie Chart**

plt.pie(df['Sex'].value\_counts(), labels=['Male', 'Female'], autopct='%1.1f%%')

**Purpose:**  
Visualizes the gender distribution in the dataset. Useful for understanding demographic balance.

**5. Line Chart**

sns.lineplot(x='Age', y='MaxHR', data=age\_hr)

**Purpose:**  
Shows how maximum heart rate (MaxHR) changes with age on average. This helps in spotting performance decline patterns across age brackets.

**Additional Visualizations (Bonus)**

**6. Heatmap**

sns.heatmap(df.corr(), annot=True)

**Purpose:**  
Reveals correlations among numerical features, assisting in feature selection and understanding multi-collinearity.

**7. Pair Plot**

sns.pairplot(df[['Age', 'Chol', 'MaxHR', 'Oldpeak', 'AHD']], hue='AHD')

**Purpose:**  
Explores pairwise relationships between selected features and their distribution split by heart disease.

**8. Violin Plot**

sns.violinplot(x='AHD', y='MaxHR', data=df)

**Purpose:**  
Combines boxplot and KDE to show the distribution of maximum heart rate across AHD classes.

**9. Swarm Plot**

sns.swarmplot(x='AHD', y='Chol', data=df)

**Purpose:**  
Provides a clear view of individual cholesterol values and their relationship with disease presence.

**10. Histogram + KDE**

sns.histplot(df['RestBP'], kde=True)

**Purpose:**  
Visualizes the distribution of resting blood pressure values with a density curve for better understanding of spread and central tendency.

**11. Count Plots for Multiple Categorical Features**

for col in ['ChestPain', 'Fbs', 'RestECG', 'ExAng', 'Thal']:

sns.countplot(data=df, x=col, hue='AHD')

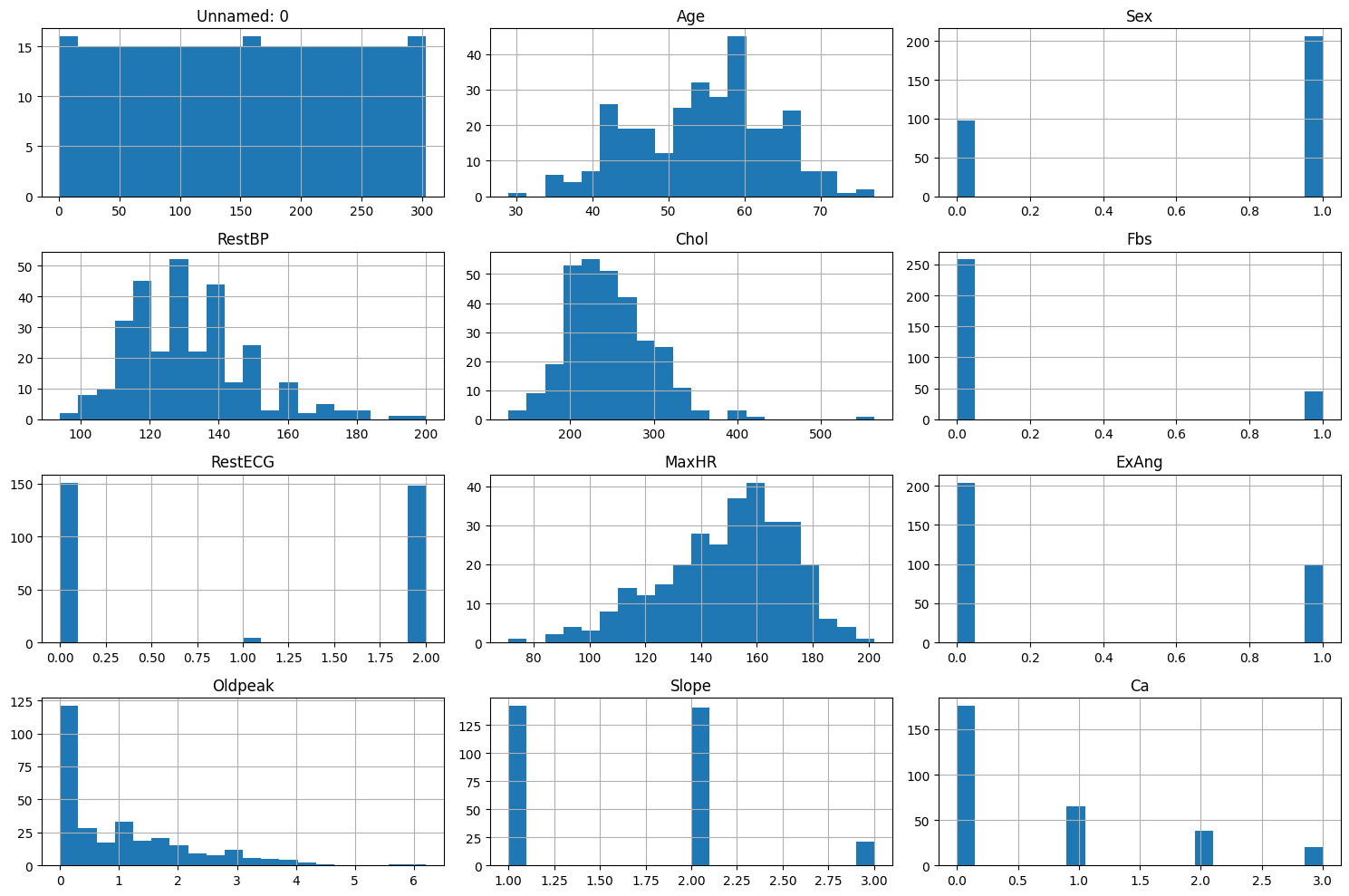
**Purpose:**  
Allows categorical feature-wise breakdown across heart disease conditions for detailed analysis.

**Conclusion**

This assignment emphasizes the crucial role of data visualization in discovering hidden insights within datasets.

* **Scatter plots** reveal relationships.
* **Bar and pie charts** show categorical distribution.
* **Box, violin, and swarm plots** highlight statistical spread and density.
* **Line and histogram plots** detect trends and patterns over continuous variables.
* **Heatmaps and pair plots** support correlation analysis.

These visual techniques not only enhance understanding but also help guide preprocessing and model development efforts effectively.



A blue and orange pie chart

AI-generated content may be incorrect.

A graph showing a number of heart rate

AI-generated content may be incorrect.

A diagram of a number of numbers

AI-generated content may be incorrect.

