Name - Janhvi Saste
PRN - 22311825
Roll No - 282065
Year - SY B

Assignment No - 3

Data Analysis and Visualization of Heart Disease Dataset

Problem Statement :

Heart disease is a major global health concern, and data analysis can help with early detection and prevention. This assignment uses a dataset of 303 patients with key features like Age, Sex, Cholesterol, and AHD (heart disease presence). Since raw data isn't easily interpretable, visualization is essential to uncover patterns and relationships. The goal is to preprocess the data and use various plots to extract insights and support predictive modeling.

Objective :

The assignment aims to preprocess the heart disease dataset and visualize key feature patterns using various plots (scatter, bar, box, pie, and line charts) to uncover insights and support heart disease analysis and modeling.

Software Used :

- Python 3.x
- Google Colab

Libraries and Packages Used :

The following Python libraries were utilized for data preprocessing and analysis:

- Pandas (For loading, manipulating, and analyzing tabular data)
- Numpy (For numerical computations and array operations)
- Seaborn (For generating plots and visualizations)

- Matplotlib (For advanced data visualization, such as heatmaps or distribution plots)

Theory :

Methodology :

This assignment follows a structured pipeline to load, preprocess, and visualize the heart disease dataset with the goal of uncovering meaningful insights. The process begins by importing the dataset (`Heart (1).csv`) into a Pandas DataFrame, followed by data preprocessing steps such as handling missing values, removing irrelevant columns like `Unnamed: 0`, and ensuring overall consistency. Visualization plays a key role in the analysis, with scatter plots used to explore the relationship between Age and Cholesterol based on heart disease status (AHD), bar plots to compare heart disease prevalence across genders, and box plots to examine Cholesterol distribution and outliers by AHD. Pie charts illustrate the proportion of patients with and without heart disease, while line charts track average MaxHR trends across different age groups. Each of these plots is interpreted to extract valuable insights into the underlying data patterns and potential risk factors associated with heart disease.

Main Function :

The assignment utilizes a combination of Pandas, Seaborn, and Matplotlib to perform data visualization and analysis. Key functions and operations include:

- scatterplot() : Creates a scatter plot with hue and color palette to examine variable relationships.

- countplot() : Visualizes the count of categorical variables, useful for comparing class distributions.

- boxplot() : Illustrates numerical feature distributions and highlights potential outliers.

- pie() : Renders a pie chart to show proportions of heart disease occurrence.

- groupby() and plot() : Computes and visualizes average MaxHR by Age as a line chart.

- show() : Displays all the generated plots.

Advantages :

- Comprehensive Insights: Various plot types offer a complete view—Scatter (relationships), Bar (counts), Box (distributions), Pie (proportions), and Line (trends).
- Ease of Use: Libraries like Seaborn make it easy to create complex visualizations with minimal code.
- Interpretability: Visuals help non-technical users (e.g., clinicians) understand data patterns clearly.
- Scalability: The same approach can be reused for other datasets with similar structures.

Disadvantages :

- Limited Preprocessing: The provided code lacks extensive cleaning (e.g., handling missing Ca values), which could affect visualization accuracy.
- Static Plots: Interactive features (e.g., tooltips) are absent, limiting exploration.
- Single Dataset: Visualization is constrained to one dataset, reducing generalizability.
- Over-Simplification: Aggregations (e.g., Line Chart) may mask individual variability.

Applications with Example :

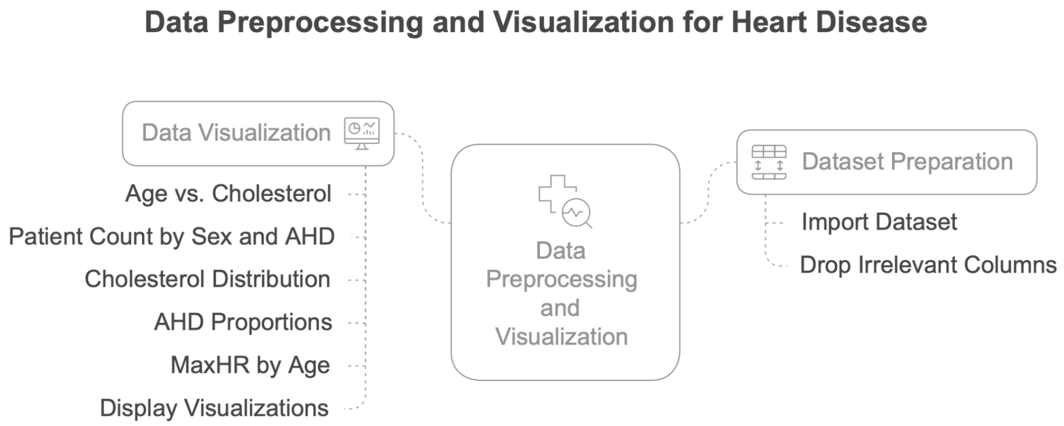
In healthcare, the Scatter Plot (Age vs. Chol) highlights that older patients with high cholesterol are more likely to have heart disease, supporting early risk profiling—e.g., a 65-year-old with a cholesterol level of 300 could be flagged for screening. In marketing, Bar Plots can compare purchase patterns across gender and product types for better customer segmentation. In finance, Box Plots are useful for spotting outlier transactions, which may indicate fraudulent activity. In education, Line Charts can visualize average test score trends across different age groups, helping identify learning patterns and gaps.

Working/Algorithm :

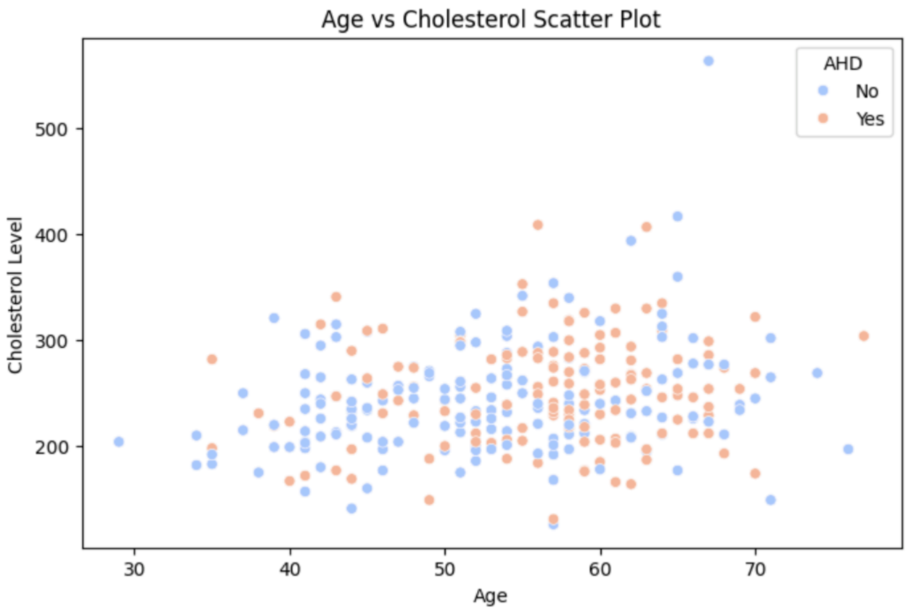
- Input: Heart disease dataset (Heart (1).csv) with 303 rows and 15 columns.
- Load Data: Use pd.read_csv() to read the CSV file.

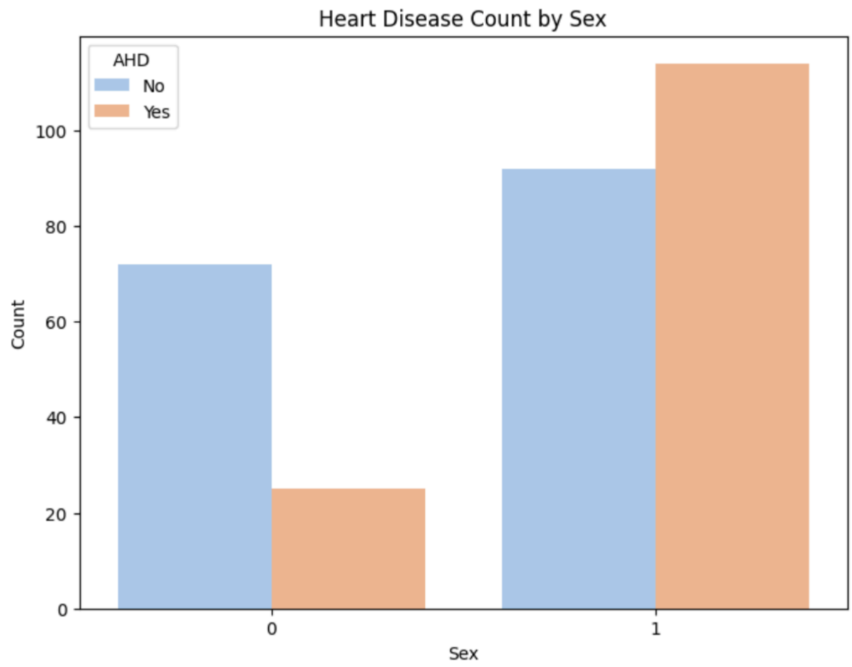
- Clean Data: Remove unnecessary column Unnamed: 0 using df.drop().
- Scatter Plot: Plot Age vs. Chol with heart disease status using sns.scatterplot().
- Bar Plot: Show counts of Sex vs. AHD using sns.countplot().
- Box Plot: Visualize Chol distribution by AHD using sns.boxplot().
- Pie Chart: Display AHD proportions with value_counts() and plt.pie().
- Line Chart: Plot average MaxHR by Age using groupby() and plt.plot().
- Output: All five plots displayed using plt.show().

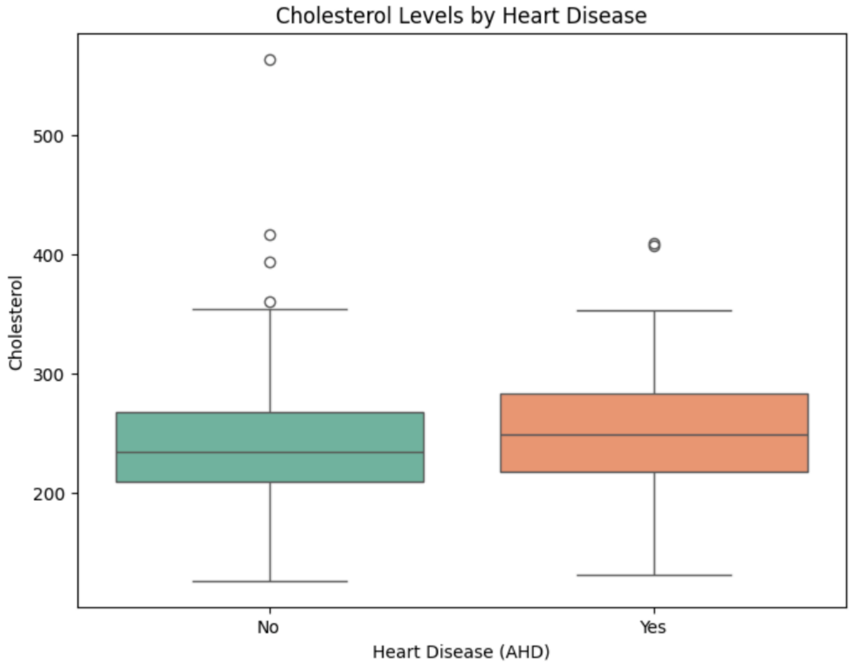
Diagram :

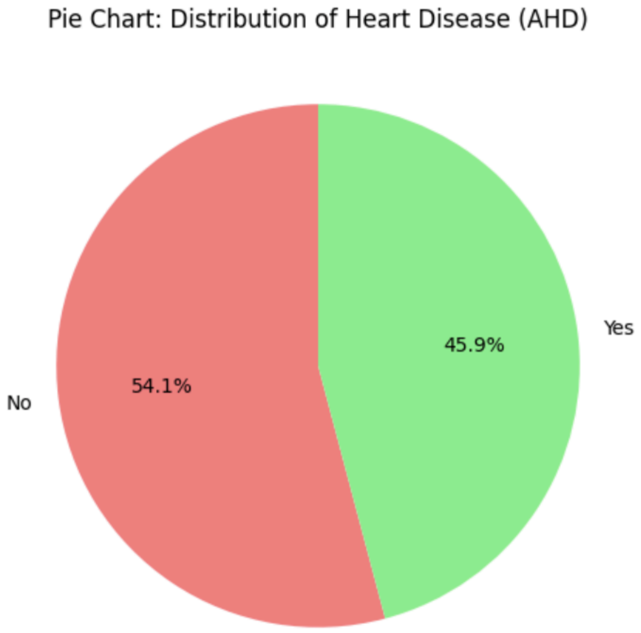


Result :




Heart Disease Count by Sex
AHD
No
Yes
Count
0
20
40
60
80
100
0
1
Sex


Cholesterol Levels by Heart Disease
Cholesterol
200
300
400
500
No
Yes
Heart Disease (AHD)


Pie Chart: Distribution of Heart Disease (AHD)
Yes
45.9%
No
54.1%

Conclusion :

This assignment effectively visualizes the heart disease dataset through five key plot types, fulfilling the objectives of Assignment 1 and 2. The Scatter Plot shows a link between higher age, cholesterol, and heart disease, while the Bar Plot highlights a greater prevalence among males. The Box Plot reveals elevated cholesterol in affected patients, with some outliers, and the Pie Chart confirms a nearly balanced class distribution. The Line Chart captures a declining trend in MaxHR with age. These visual insights support early diagnosis and model building. Limitations include basic preprocessing and static visuals; future work could enhance interactivity and data quality.