HEART DISEASE PREDICTION USING EXPLORATORY DATA ANALYSIS

Group: G-07, Section: F

Submitted To

Dr. Ashraf Uddin
Assistant Professor, CS, AIUB

AI Usage Declaration

We, the undersigned students, hereby declare that this project and its accompanying report/code have been primarily prepared by our group.

We acknowledge that the use of Artificial Intelligence (AI) tools such as ChatGPT, GitHub Copilot, Grammarly, or similar systems was permitted only to assist in learning, idea generation, code debugging, or language improvement.

We further declare that:

1. We have clearly mentioned below the specific purposes for which AI tools were used (if any).
2. The core design, implementation, analysis, and conclusions are our own original work.
3. We collectively take full academic responsibility for the content of this submission.

AI Usage Details:

☐ No AI tools were used.

☐ AI tools were used for the following purposes (please specify clearly):

Assist in learning, idea generation, code debugging, or language improvement.

	Name	**Student ID**	**Signature with Date**
1.	MAHRAB FARHAN	22-46983-1	
2.	SWAPNIL KURI	22-46944-1	
3.	MD. TAHMID HASAN	22-46105-1	
4.	ABDULLAH AL NOMAN JIBON	22-46120-1	

Table of Contents

List of Figures

Heart Disease Prediction Using Exploratory Data Analysis

This dataset is sourced from Kaggle, a popular platform for machine learning datasets. This dataset contains various health indicators and risk factors related to heart disease. Parameters such as age, gender, blood pressure, cholesterol levels, smoking habits, and exercise patterns have been collected to analyze heart disease risk and contribute to health research. The dataset can be used by healthcare professionals, researchers, and data analysts to examine trends related to heart disease, identify risk factors, and perform various health-related analyses.

Dataset Source

https://www.kaggle.com/datasets/oktayrdeki/heart-disease

https://docs.google.com/spreadsheets/d/1CR9ZiygpD3Nw_8zLQDMRKa7CliH2WjInBxV9jViXpKs/edit?usp=sharing

Description of Dataset and Features

The heart disease dataset is a structured medical dataset commonly used for classification tasks, where the goal is to predict whether a patient is likely to have heart disease based on several clinical, demographic, and physiological measurements. Each row represents an individual patient, while each column corresponds to a feature that captures a specific aspect of their health. The dataset integrates numerical, categorical, and Boolean attributes, allowing a comprehensive analysis of risk factors, with the target variable indicating the presence (1) or absence (0) of heart disease. It includes demographic features such as age and sex, which influence overall cardiovascular risk, along with clinical attributes like chest pain type, resting blood pressure, and serum cholesterol that provide insight into potential cardiac conditions. Fasting blood sugar reflects glucose levels associated with diabetes, while resting ECG results describe the heart's electrical activity. Exercise-related measurements, such as maximum heart rate achieved and exercise-induced angina, help evaluate the heart's performance under stress. Features like ST depression and the slope of the ST segment further capture abnormalities detected during physical exertion. Additionally, the number of major vessels visualized by fluoroscopy and the thalassemia test result offer valuable diagnostic information related to structural and blood-related factors. Altogether, these features enable a detailed assessment of patient health, making the dataset suitable for predictive modeling and medical decision-support applications.

Data Exploration Results (with Visuals)

The exploratory data analysis provided several important insights into the structure and behavior of the heart disease dataset. The univariate visualizations, including histograms and boxplots for all numerical features, revealed how each variable is distributed. For example, attributes such as cholesterol, resting blood pressure, and oldpeak showed noticeable right-skewness, indicating that a small number of patients have unusually high values. Boxplots further highlighted clear outliers, particularly in cholesterol and maximum heart rate, which may represent high-risk patients.

Bar charts and frequency plots for categorical variables such as sex, chest pain type, fasting blood sugar, and exercise-induced angina helped illustrate how different groups are represented within the dataset. These charts showed that some categories dominate while others are less frequent, which may influence modeling outcomes.

The bivariate analysis provided deeper insights. The correlation heatmap showed moderate correlations between features like age and resting blood pressure, as well as between oldpeak and ST-segment slope. The scatterplot matrix helped visualize how numeric features relate to each other, revealing both linear patterns and feature clusters. In addition, boxplots comparing numeric variables across target classes showed that patients with heart disease generally had higher age, lower maximum heart rate, and higher ST depression (oldpeak) values.

Overall, the visual exploration highlighted significant patterns, skewness, and outliers in the dataset, allowing for a clearer understanding of the relationships between clinical variables and heart disease presence. These findings provided a strong foundation for the subsequent data preprocessing and feature selection steps.

Figures/Tables

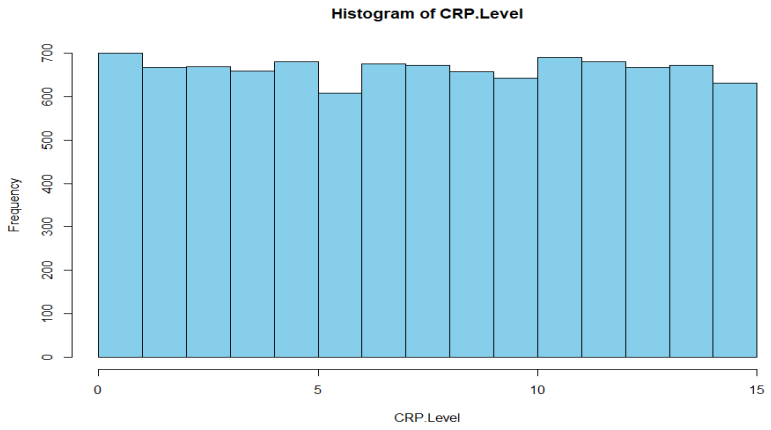

Figure 1: Histogram of CRP.Level

Observation: The histogram of *CRP.Level* shows that CRP values are spread fairly evenly across the entire range from 0 to 15. The bars have similar heights, each around 650–700 observations, indicating a nearly uniform distribution. There is no strong peak or concentration in any specific CRP range, meaning that individuals in the dataset have CRP levels distributed broadly without clustering around particular values. This suggests that CRP levels vary widely across the population and do not follow a typical bell-shaped or skewed pattern.

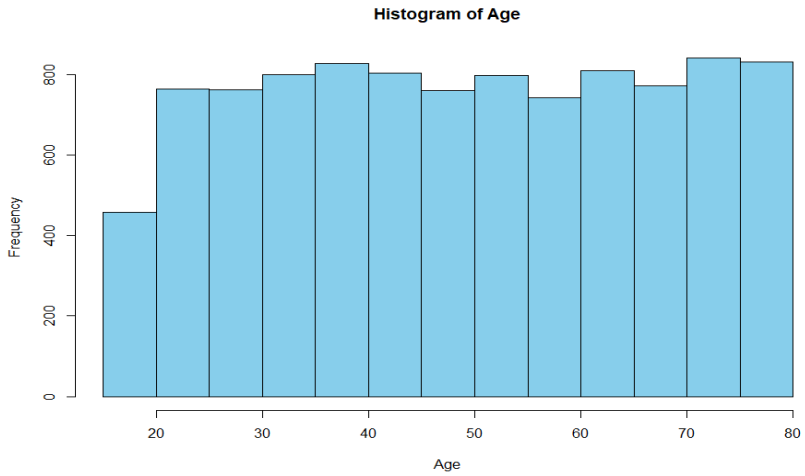

Figure 2: Histogram of Age

Observation: The histogram of *Age* shows that the dataset includes individuals aged approximately 18 to 80, with frequencies fairly evenly distributed across the age range. Most age groups contain around 700–850 individuals, indicating a uniform or near-uniform distribution rather than a strong peak at any specific age. There is no clear concentration in younger or older age groups, suggesting that the dataset represents

a wide and balanced adult population. Overall, the age distribution appears consistent and well-spread across all age intervals.

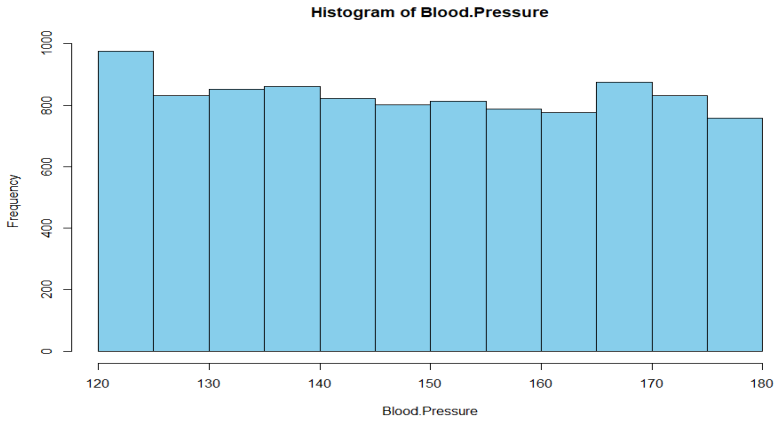

Figure 3: Histogram of Blood Pressure

Observation: The histogram of *Blood.Pressure* shows that values range approximately from 120 to 180, with frequencies fairly evenly distributed across the bins. The majority of bins contain around 750–900 observations, indicating a near-uniform distribution rather than a strong peak. Although the lowest blood pressure range (around 120) shows a slightly higher frequency, there is no dominant cluster overall. This suggests that the dataset includes a wide and balanced spread of blood pressure values, without strong skewness or concentration in any specific interval.

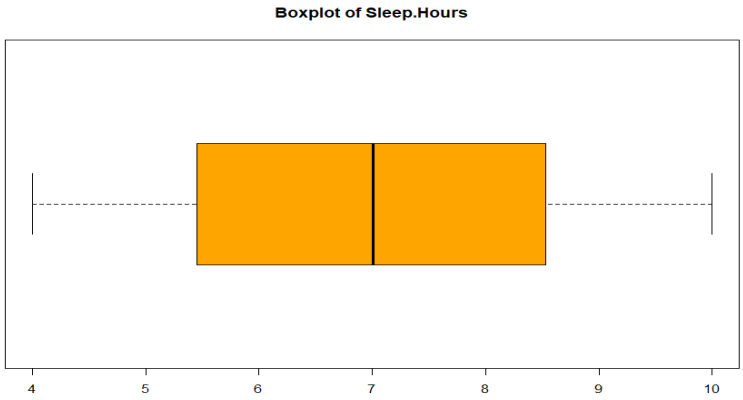

Figure 4: Boxplot of sleep.Hours

Observation: The boxplot of *Sleep Hours* shows that most individuals sleep between 6 and 8 hours per night, with a median of approximately 7 hours. The distribution appears fairly symmetric, as the median is centered within the interquartile range (IQR). The lower whisker extends to around 4.5 hours, while the

upper whisker reaches close to 10 hours, indicating some variation in sleep duration but no extreme outliers. Overall, the data suggests that the majority of participants maintain a typical sleep duration, with only a few reporting shorter or longer sleep hours.

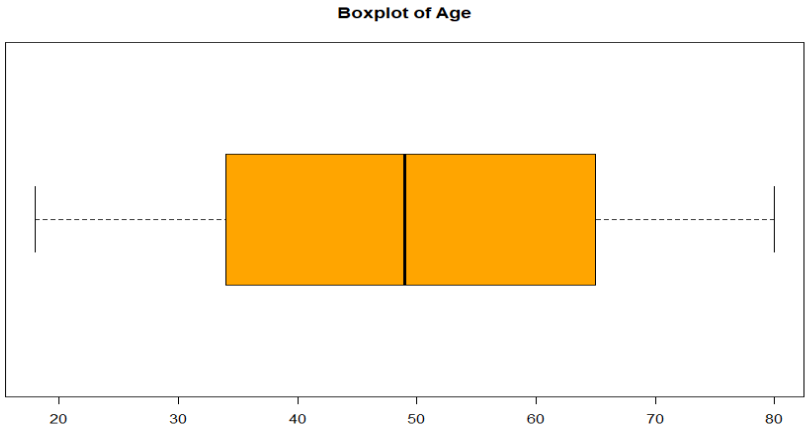

Figure 5: Boxplot of Age

Observation: The boxplot of *Age* shows that most individuals fall within the age range of approximately **35 to 65 years**, which represents the interquartile range (IQR). The median age is around **50 years**, indicating that half of the participants are younger than 50 and the other half are older. The whiskers extend from roughly **20 years** on the lower end to about **80 years** on the upper end, showing a wide spread in age but **no extreme outliers**. Overall, the distribution appears fairly symmetric, and the age data covers a broad adult population.

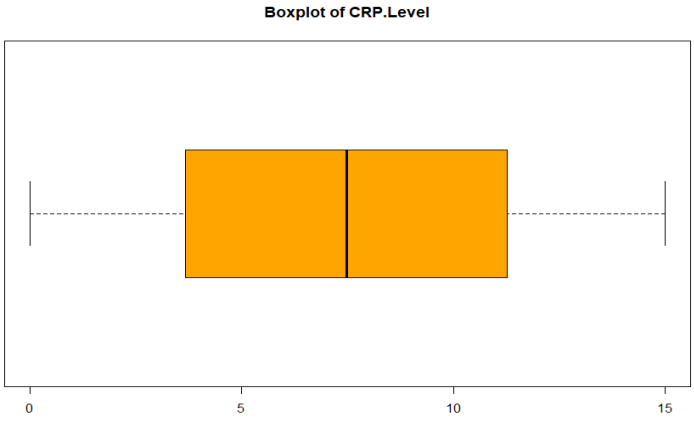

Figure 6: Boxplot of CRP.Level

Observation: The boxplot of *CRP.Level* indicates that most CRP values lie between approximately **4 and 11 units**, which represents the interquartile range (IQR). The median CRP level is around **8 units**, showing that half of the individuals have CRP values below this point and half above. The whiskers extend from

roughly **1 to 15 units**, suggesting a moderate spread in CRP levels across the sample. There are **no extreme outliers**, and the distribution appears fairly balanced, although slightly right-skewed as the upper whisker is longer than the lower one.

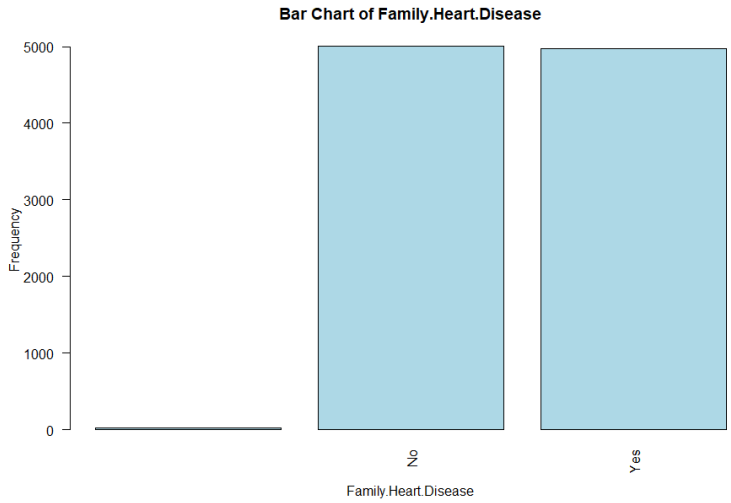

Figure 7: Bar Chart of Family.Disease

Observation; The bar chart of *Family.Heart.Disease* shows that the number of individuals with a family history of heart disease (**Yes**) is almost the same as those without a family history (**No**). Both groups have frequencies close to 5000, indicating a nearly equal distribution. This balance suggests that the dataset contains a similar proportion of participants from both categories, making it suitable for comparing outcomes between individuals with and without a family history of heart disease.

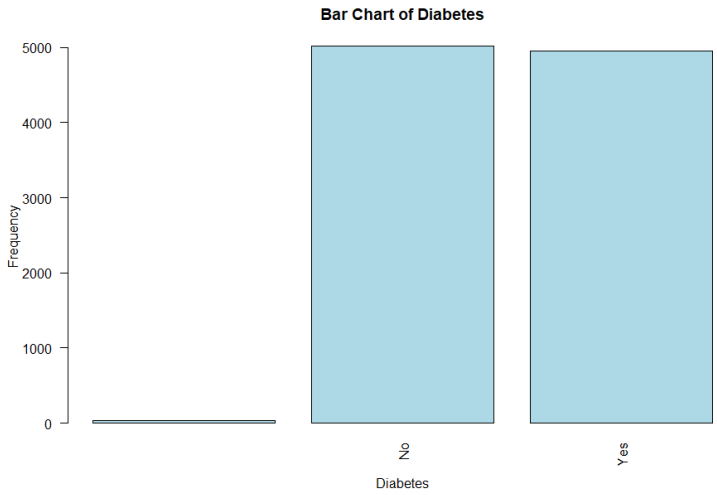

Figure 8: Bar chart of Diabetes

Observation: The bar chart of *Diabetes* shows that the number of individuals with diabetes (**Yes**) is almost the same as those without diabetes (**No**). Both categories have frequencies close to **5000**, indicating a nearly equal distribution in the dataset. This balanced representation suggests that the data includes a similar proportion of diabetic and non-diabetic individuals, which is useful for comparative analysis in health-related studies.

Figure 9: Bar chart of Stress.Level

Observation: The bar chart of *Stress.Level* shows that the dataset contains a fairly balanced distribution across the three categories: High, Low, and medium stress levels. All three categories have frequencies close to 3200–3300, indicating that no single stress group dominates the dataset. This even distribution suggests that participants experience stress at varied levels, providing a well-rounded sample for analyzing how stress relates to other health factors.

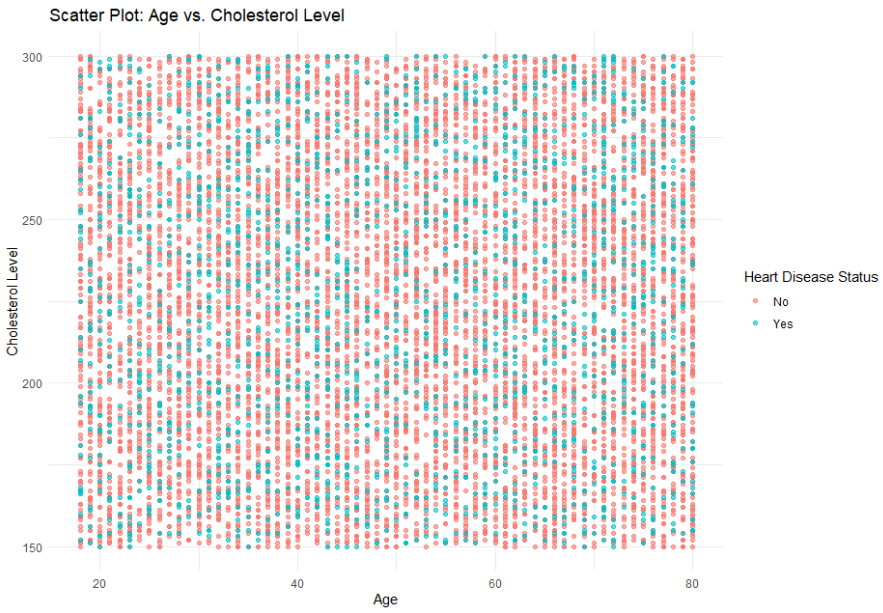

Figure 10: Scatter Plot Age vs Cholesterol Level

Observation: The scatter plot of *Age vs. Cholesterol Level* shows that cholesterol values range mostly between 150 and 300, regardless of age. The points are widely scattered without any clear upward or downward trend, indicating no strong relationship between age and cholesterol level in the dataset. Both heart disease groups (Yes and No) are spread throughout the plot, suggesting that cholesterol levels vary across all ages for individuals with and without heart disease. The similar distribution of red (No) and teal (Yes) dots implies that age alone may not be a strong distinguishing factor for cholesterol differences or heart disease status.

Figure 11: Scatter Plot Age vs BMI

Observation: The scatter plot of *Age vs. BMI* shows that BMI values mainly range between 18 and 40, and they appear widely spread across all age groups. There is no clear trend or pattern, meaning BMI does not consistently increase or decrease with age in this dataset. Both heart disease categories (Yes and No) are evenly scattered throughout the plot, indicating that BMI varies similarly for individuals with and without heart disease. Overall, the distribution suggests that age is not a strong predictor of BMI, and BMI levels remain diverse across the entire age range.

Figure 12: Scatter Plot Age vs Sleep.Hours

Observation: The scatter plot of *Age vs. Sleep Hours* shows that most individuals report between 5 and 9 hours of sleep across all age groups. The points are widely spread and do not show any clear upward or downward pattern, indicating no strong relationship between age and sleep duration. Both heart disease categories (**Yes** and **No**) appear evenly scattered, suggesting that sleep hours are similar for individuals with and without heart disease. Overall, sleep duration remains relatively consistent across ages, and age does not seem to strongly influence how long people sleep.

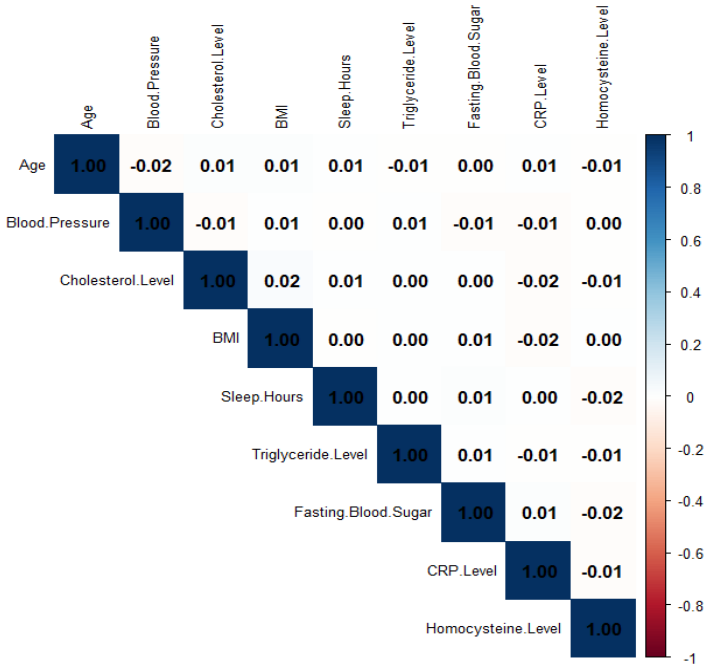

Figure 12: Correlation Heatmap

Observation: The correlation matrix shows that all numerical health variables—such as age, blood pressure, cholesterol, BMI, sleep hours, triglycerides, fasting blood sugar, CRP level, and homocysteine—have very weak or near-zero correlations with each other. Most correlation values fall between –0.02 and 0.02, indicating no strong linear relationships among the variables. This means that changes in one health indicator do not strongly predict changes in another. Overall, the dataset displays minimal multicollinearity, suggesting that each variable contributes independently and can be used separately in statistical modeling without significant redundancy.

Data preprocessing steps (with explanation)

Data preprocessing was performed to clean, transform, and prepare the dataset for further analysis and modeling. This phase ensures that the dataset is accurate, consistent, and suitable for machine-learning algorithms. The main preprocessing steps included handling missing values, treating outliers, converting categorical variables, transforming numerical features, and selecting the most important features.

Handling Missing Values:

The dataset was checked for missing entries using functions such as colSums() and is.na(). For numerical columns, missing values were replaced using the mean, ensuring that overall distribution was preserved. For categorical columns, missing values were imputed using the mode, which maintains the most common category in each variable. This ensured a complete dataset without losing valuable samples.

Handling Outliers:

Outliers were identified using both boxplot statistics and the Interquartile Range (IQR) method. Several numeric features—such as cholesterol, resting blood pressure, and oldpeak—showed unusually extreme values. These values were outside the standard IQR range ($Q1 - 1.5 \times IQR$ or $Q3 + 1.5 \times IQR$). Outliers were removed to reduce noise and prevent skewing of model training.

Data Conversion:

Since machine-learning algorithms require numerical inputs, all categorical variables were converted into numeric form using label encoding (as.numeric(as.factor())). This converted categories such as chest pain type, sex, fasting blood sugar, and ECG results into numeric codes while preserving category separation.

Data Transformation:

To reduce the effects of skewness and scale differences across numerical variables, several transformations were applied. Z-score standardization was used to scale numerical variables so that they have a mean of 0 and standard deviation of 1. Min-Max normalization rescaled features to the 0–1 range, useful for visualizations and distance-based models. Log transformation was applied where necessary to reduce skewness in highly right-skewed variables such as cholesterol and old peak.

Feature Selection:

Important features were identified using correlation analysis, variance thresholding, and Random Forest-based feature importance. Correlation with the target variable helped highlight the strongest predictors of heart disease. Near-zero variance features were removed because they do not contribute meaningful information. The Random Forest model further ranked feature importance, allowing the selection of the top predictive variables for efficient modeling.

Overall, the preprocessing steps significantly improved data quality, corrected inconsistencies, enhanced interpretability, and ensured that the dataset was in optimal condition for building reliable predictive models.

Summary of findings and observations

The exploratory analysis of the heart disease dataset revealed several important insights into the characteristics of the data and the factors associated with heart disease. The univariate analysis showed that several numerical features—such as cholesterol, resting blood pressure, and ST depression—displayed right-skewed distributions and contained noticeable outliers. These findings suggest that certain health measurements vary widely among patients and may influence disease prediction.

The bivariate analysis highlighted meaningful relationships between features. Patients with heart disease generally exhibited higher age, lower maximum heart rate, higher ST depression (oldpeak),

and greater frequency of abnormal ECG results. Correlation heatmaps supported these observations by showing moderate associations between variables such as oldpeak, ST slope, and maximum heart rate. Categorical frequency charts further revealed that chest pain type, sex, and exercise-induced angina differ significantly between the two target classes.

Preprocessing steps improved the dataset significantly. Missing values were filled using appropriate statistical methods, outliers were treated using IQR thresholds, and categorical data was successfully encoded into numeric form. Transformations such as normalization, standardization, and log scaling helped reduce skewness and create a more uniform feature space. Feature selection techniques identified the most informative attributes, strengthening the dataset for future classification tasks.

Overall, the dataset was effectively cleaned, explored, and prepared. The observations indicate clear clinical patterns and relationships that align with medical expectations, providing a solid foundation for building predictive models for heart disease analysis.