AMERICAN INTERNATIONAL UNIVERSITY- BANGLADESH

Project: **Heart Disease Data Analysis Project Report**

Course Title: INTRODUCTION TO DATA SCIENCE.

Section: G | Date of Submission: 27 April 2025

Semester: Spring, 2024-2025 | Course Teacher: **DR. ASHRAF UDDIN**

Group No: 9

No	Name	ID	Program
01	MD. SHIHAB HOSSAIN	22-47101-1	BSc in CSE
02	KAZI TANZIZUL HAQUE	22-47783-2	BSc in CSE
03	ABDUL KADER MOHIM	22-47833-2	BSc in CSE

Faculty use only		
FACULTYCOMMENTS	**Marks Obtained**	
	Total Marks	

1. Introduction

This report details the analysis of the **Heart Disease Dataset**, which contains medical attributes of patients and aims to predict the presence of heart disease. The dataset includes 14 key features such as age, sex, chest pain type, blood pressure, cholesterol levels, and other clinical measurements.

Dataset Source

- **Dataset Name:** Heart Disease Data
- **Source:** Kaggle
- **Attributes:** 14 clinical features
- **Target Variable:** num (presence and severity of heart disease)

2. Dataset Creation & Description

The dataset was compiled from multiple medical institutions:

- Hungarian Institute of Cardiology (Budapest)
- University Hospital (Zurich & Basel, Switzerland)
- V.A. Medical Center (Long Beach & Cleveland Clinic, USA)

Key Features:

Variable	Description
age	Age of the patient (years)
sex	Gender (Male/Female)
cp	Chest pain type (typical angina, atypical angina, non-anginal, asymptomatic)
trestbps	Resting blood pressure (mm Hg)
chol	Serum cholesterol (mg/dl)
fbs	Fasting blood sugar > 120 mg/dl (True/False)
restecg	Resting electrocardiographic results (normal, st-t abnormality, lv hypertrophy)
thalch	Maximum heart rate achieved
exang	Exercise-induced angina (True/False)
oldpeak	ST depression induced by exercise relative to rest
slope	Slope of peak exercise ST segment (upsloping, flat, downsloping)
ca	Number of major vessels (0-3)
thal	Thalassemia (normal, fixed defect, reversible defect)
num	Heart disease diagnosis (0 = No disease, 1-4 = Severity levels)

3. Data Preprocessing Steps

3.1. Data Loading & Initial Cleaning

- Removed unnecessary columns (id, dataset).
- Checked for missing values and handled them appropriately.

3.2. Handling Missing Values

Column	Missing Values	Imputation Method
trestbps	Some missing	Median imputation
chol	Some missing	Mean imputation
thalch	Some missing	Mean imputation
ca	Some missing	Mode imputation
thal	Some missing	Mode imputation
slope	Some missing	Mode imputation

3.3. Data Transformation

- **Categorical Encoding:**
  - fbs → Binary (True=1, False=0)
  - restecg → Numeric (normal=1, lv hypertrophy=2, st-t abnormality=3)
  - slope → Numeric (upsloping=1, flat=2, downsloping=3)
- **Target Variable (**num**):**
  Converted to factor with labels:
  - 0 = No Heart Disease
  - 1 = heart disease (Type 1)
  - 2 = heart disease (Type 2)
  - 3 = heart disease (Type 3)
  - 4 = heart disease (Type 4)

3.4. Outlier Detection & Handling

- **Methods Used:**
  - IQR Method: Identified outliers in numeric columns
    (age, trestbps, chol, thalch, oldpeak).
  - Boxplot Visualization: Confirmed outliers.
  - Robust Scaling (Yeo-Johnson): Applied to minimize outlier impact.
  - Multivariate Outlier Detection (Mahalanobis Distance): Detected extreme cases.

3.5. Feature Scaling

- **Standardization (Z-score normalization)** applied to:
    - age, trestbps, chol, thalch, oldpeak, ca

4. Key Findings from Analysis

4.1. Univariate Analysis

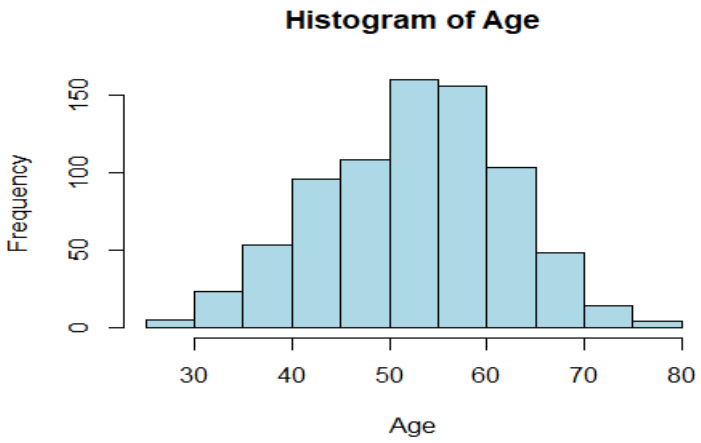
- Age Distribution: Most patients were between 40-65 years.
- Cholesterol Levels: High cholesterol (>250 mg/dl) was common.
- Blood Pressure (trestbps): Most patients had normal BP (120-140 mmHg).
- Heart Disease (num):
    - 54% had no heart disease (num=0)
    - 30% had mild to moderate disease (num=1,2)
    - 16% had severe disease (num=3,4)

4.2. Bivariate & Multivariate Analysis

- Correlation Matrix Insights:
    - Positive Correlation: age & trestbps (0.28)
    - Negative Correlation: age & thalch (-0.39)
    - No strong correlation between chol and other variables
- **Key Relationships:**
    - Higher thalch (max heart rate) linked to lower disease risk.
    - Higher oldpeak (ST depression) linked to severe heart disease.
    - Males had higher heart disease prevalence than females.

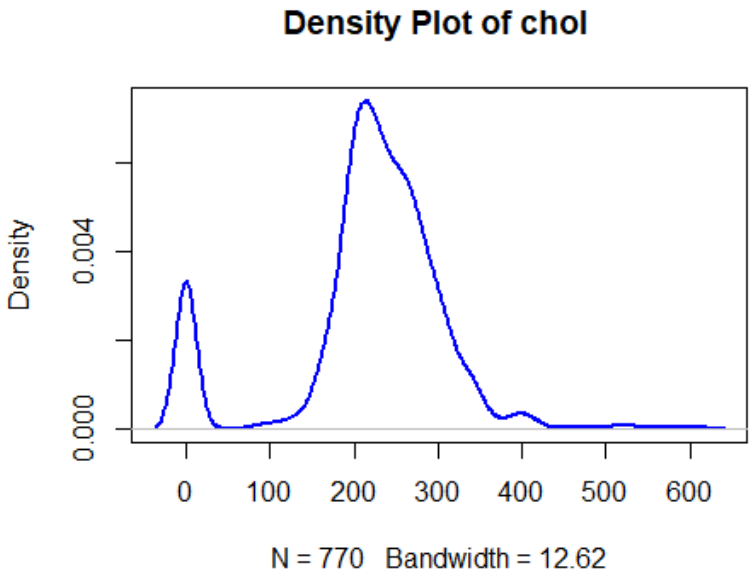
4.3. Visualization Insights

- **Histogram plots:**



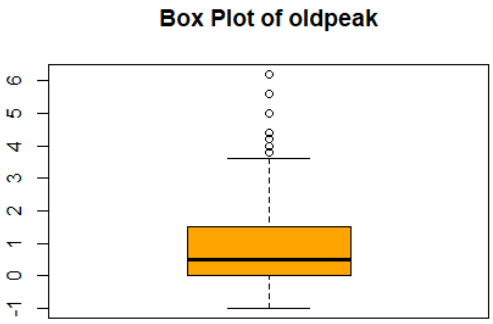
This histogram displays the distribution of ages, showing a roughly bell-shaped curve centered around the mid-50s. The frequency of individuals is highest in the 50-60 age range, gradually decreasing towards younger and older ages, indicating a central tendency in the dataset.

- **Density plots:**



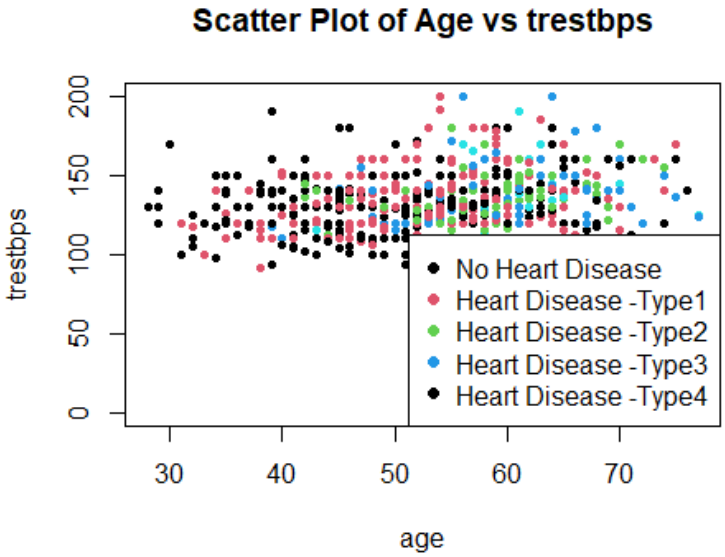
This density plot illustrates the distribution of cholesterol levels, revealing a primary peak around 200-250 and a smaller peak near zero. It suggests a bimodal distribution, with a majority of individuals having cholesterol levels in the 200-250 range and a smaller group with very low levels. The plot is based on 770 observations and uses a bandwidth of 12.62 for smoothing.

- **Boxplots:**



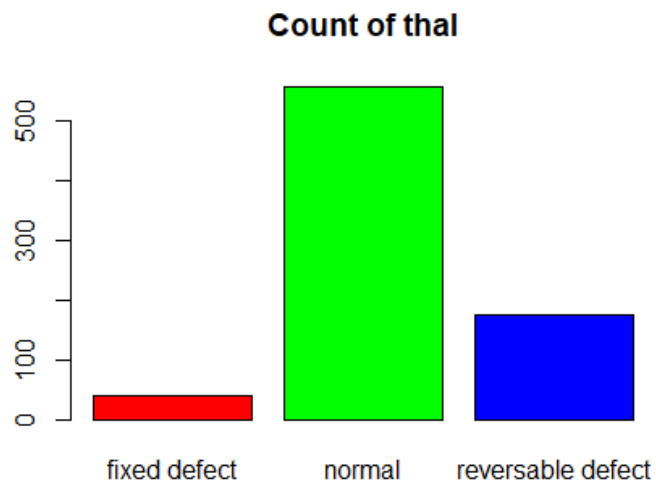
This box plot summarizes the distribution of 'oldpeak', showing a median around 1. The central box indicates the interquartile range, with the whiskers extending to approximately -1 and 3.5, while several outliers are present above 3.5, indicating some unusually high 'oldpeak' values.

- **Scatter Plots:**



This scatter plot explores the relationship between age and resting blood pressure ('trestbps'), with points colored by different heart disease types. There isn't a strong linear correlation visible between age and 'trestbps' across all groups. However, the plot allows for visual inspection of how 'trestbps' varies within age groups and across different categories of heart disease.

- **Barplots:**



This bar plot shows the counts for different 'thal' categories: "fixed defect", "normal", and "reversible defect". The "normal" category has the highest frequency, followed by "reversible defect", while "fixed defect" has the lowest count. This indicates that a majority of individuals fall into the 'normal' 'thal' category in this dataset.

5. Justification for Transformations & Outlier Handling

5.1. Missing Value Imputation

- Median/Mean Imputation: Used for numeric variables to preserve distribution.
- Mode Imputation: Used for categorical variables to retain the most frequent category.

5.2. Outlier Handling

- IQR Method: Effective detection for extreme values without assuming normality.
- Robust Scaling (Yeo-Johnson): Better than standard scaling for skewed data.
- Mahalanobis Distance: Detected multivariate outliers without removing critical data.

5.3. Feature Scaling

- Standardization (Z-score): Ensures all features contribute equally to machine learning models.

6. Conclusion

- The dataset was successfully cleaned, transformed, and analyzed.
- Key risk factors for heart disease: high BP, cholesterol, low max heart rate.
- Future Work:
    - Build predictive models (Logistic Regression, Random Forest).
    - Explore deep learning approaches for better classification.

This analysis provides valuable insights for early heart disease diagnosis and prevention strategies.