Perdicting Heart Disease: A Classification Approach Using Patient Health Metrics

Syed H

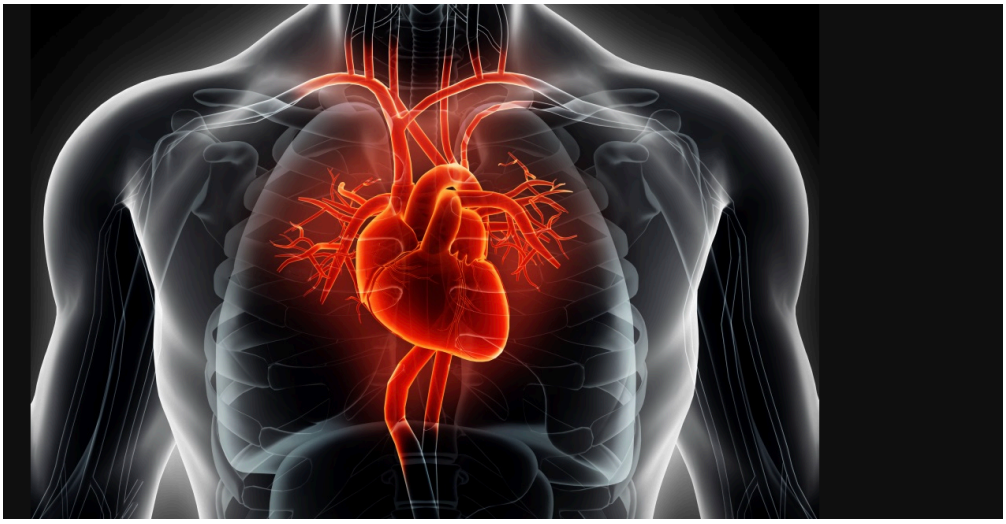


Figure 1: AI-generated image of the human heart. Created using OpenAI's ChatGPT.

Introduction

Cardiovascular diseases are one of leading causes of high mortality in the world, particularly heart disease. These diseases develop as we age and are influenced by risk factors such as high blood pressure, smoking, alcohol use, poor diet, and air pollution. According to "Our World in Data", 18.5 million deaths in 2019 were caused by cardiovascular diseases, that is 50,850 deaths daily. Therefore, early detection is important for preventing life threatening complications and early medical intervention.

Objectives

The primary goal of this project is to apply classification algorithms to determine whether a patient is likely to have a heart disease based on their health metrics. To achieve this task, I will analyze factors such as age, cholesterol level, blood pressure, and chest pain type to predict presence of heart disease. This significance of this research lies in its potential to support medical professionals in making faster and more accurate diagnosis. The dataset used in this project includes multiple patient attributes and a binary target variable indicating the presence or absence of heart disease status which makes it an appropriate choice for building and testing classification models.

Steps envolved in analysis:

- Exploring heart disease data using EDA
- Applying different classification algorithms for perdiction.

Dataset

Link: Heart.csv

The data set for this project was found on kaggle. It includes patient health metrics and a target variable that shows the presence or absence of heart disease. The data was downloaded as a CSV file and imported into the R Markdown file for analysis. The data set contains **303 observations** and **14 features**. Each row represents a patient and each column contains a specific health measurement. The feature names are abbreviated, so I will rename the columns to make them easier to understand.

According to the publisher of this data set there are **7 faulty entries** that should be removed. These entries were addressed in the original version, but the Kaggle version still includes them. For that reason, I will drop these entries during the data pre-processing step.

Section 1 - Exploratory Data Analysis (EDA)

This section provides an overview of the data set through summary statistics and a data dictionary. Moreover, we use descriptive statistics such as mean, median, standard deviation, and range to understand the central tendencies and variability of key variables. A class distribution plot is also included to show the balance between patients with and without heart disease. This initial analysis helps with building a foundation for later sections.

1.1 - Imports

```
library(knitr)
library(dplyr)
library(rstudioapi)
library(ggplot2)
library(glmnet)
library(GGally)
library(psych)
library(patchwork)
library(ggcorrplot)
library(randomForest)
library(scales)
library(caret)
library(pROC)
library(class)
library(kableExtra)
library(tidyr)
library(rstatix)
library(DescTools)
library(formatR)
```

1.2 - Set Working Directory

```
current_path <- rstudioapi::getActiveDocumentContext()$path
setwd(dirname(current_path))
print(getwd())
```

```
[1] "/Users/syedwasi/Desktop/FinalProjectX"
```

1.3 - Load Dataset

```
health_data <- read.csv("heart.csv")
```

1.4 - Dataset Observation

```
knitr::kable(head(health_data, 5), caption = "First 5 Observations from Dataset")
```

Table 1: First 5 Observations from Dataset

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

1.5 - Data Dictionary

```
# Creating data frame for data dictionary
data_dictionary <- data.frame(
  Feature = c(
    "Age", "Sex", "cp (Chest Pain Type)", "trestbps", "chol",
    "fbs", "restecg", "thalach", "exang", "oldpeak",
    "slope", "ca", "thal", "target"
  ),

  Description = c(

    "Age of the patient",

    "Sex (1 = male; 0 = female)",

    "Chest pain type (0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic)"
```

```
    "Resting blood pressure (mm Hg)",

    "Serum cholesterol (mg/dl)",

    "Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)",

    "Resting electrocardiographic results (0 = normal, 1 = ST-T wave abnormality, 2 = left ventricular h

    "Maximum heart rate achieved",

    "Exercise-induced angina (1 = yes; 0 = no)",

    "ST depression induced by exercise",

    "Slope of the peak exercise ST segment (0 = upsloping, 1 = flat, 2 = downsloping)",

    "Number of major vessels colored by fluoroscopy (0-3)",

    "Thalassemia (0 = error, 1 = fixed defect, 2 = normal, 3 = reversible defect)",

    "Target (1 = heart disease present; 0 = no disease)"
  )
)


kable(data_dictionary, caption = "Data Dictionary of Heart Disease Dataset")
```

Table 2: Data Dictionary of Heart Disease Dataset

Feature	Description
Age	Age of the patient
Sex	Sex (1 = male; 0 = female)
cp (Chest Pain Type)	Chest pain type (0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic)
trestbps	Resting blood pressure (mm Hg)
chol	Serum cholesterol (mg/dl)
fbs	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
restecg	Resting electrocardiographic results (0 = normal, 1 = ST-T wave abnormality, 2 = left ventricular hypertrophy)
thalach	Maximum heart rate achieved
exang	Exercise-induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise
slope	Slope of the peak exercise ST segment (0 = upsloping, 1 = flat, 2 = downsloping)
ca	Number of major vessels colored by fluoroscopy (0–3)
thal	Thalassemia (0 = error, 1 = fixed defect, 2 = normal, 3 = reversible defect)
target	Target (1 = heart disease present; 0 = no disease)

1.6 - Removing Faulty data

```
# Filtering out incorrect 'ca' values
health_data <- health_data[health_data$ca < 4, ]

# Filtering out incorrect 'thal' values
health_data <- health_data[health_data$thal > 0, ]

# Remaining number of rows after correction
cat("The length of the data now is", nrow(health_data), "instead of 303!\n")
```

```
The length of the data now is 296 instead of 303!
```

1.7 - Renaming Columns

```
# Renaming Columns for Clarity
names(health_data)[names(health_data) == "cp"] <- "chest_pain_type"
names(health_data)[names(health_data) == "trestbps"] <- "resting_blood_pressure"
names(health_data)[names(health_data) == "chol"] <- "cholesterol"
names(health_data)[names(health_data) == "fbs"] <- "fasting_blood_sugar"
names(health_data)[names(health_data) == "restecg"] <- "resting_electrocardiogram"
names(health_data)[names(health_data) == "thalach"] <- "max_heart_rate_achieved"
names(health_data)[names(health_data) == "exang"] <- "exercise_induced_angina"
names(health_data)[names(health_data) == "oldpeak"] <- "st_depression"
names(health_data)[names(health_data) == "slope"] <- "st_slope"
names(health_data)[names(health_data) == "ca"] <- "num_major_vessels"
names(health_data)[names(health_data) == "thal"] <- "thalassemia"
```

1.8 - Recoding Categorical Values

```
# Re-coding categorical variables to readable labels

# Sex
health_data$sex <- ifelse(health_data$sex == 0, "Female", "Male")

# Chest Pain Type
health_data$chest_pain_type <- factor(health_data$chest_pain_type,
                                      levels = c(0, 1, 2, 3),
                                      labels = c("Typical Angina", "Atypical Angina",
                                                 "Non-Anginal Pain", "Asymptomatic"))

# Fasting Blood Sugar
health_data$fasting_blood_sugar <- ifelse(health_data$fasting_blood_sugar == 0,
                                          "Lower than 120 mg/dl", "Greater than 120 mg/dl")

# Resting Electrocardiogram
health_data$resting_electrocardiogram <- factor(health_data$resting_electrocardiogram,
                                                levels = c(0, 1, 2),
                                                labels = c("Normal", "ST-T Wave Abnormality",
```

```
                                                    "Left Ventricular Hypertrophy"))

# Exercise Induced Angina
health_data$exercise_induced_angina <- ifelse(health_data$exercise_induced_angina == 0,
                                              "No", "Yes")

# ST Slope
health_data$st_slope <- factor(health_data$st_slope,
                               levels = c(0, 1, 2),
                               labels = c("Upsloping", "Flat", "Downsloping"))

# Thalassemia
health_data$thalassemia <- factor(health_data$thalassemia,
                                  levels = c(1, 2, 3),
                                  labels = c("Fixed Defect", "Normal", "Reversible Defect"))
```

1.9 - Dataset Summary (Continuous Variables)

```
# Continuous features
continuous_data <- health_data %>%
  select(age, resting_blood_pressure, cholesterol, max_heart_rate_achieved, st_depression, num_major_ves

# Summary statistics
summary_stats <- continuous_data %>%
  summarise(across(everything(), list(
    Mean = ~mean(., na.rm = TRUE),
    Median = ~median(., na.rm = TRUE),
    SD = ~sd(., na.rm = TRUE),
    Min = ~min(., na.rm = TRUE),
    `25th` = ~quantile(., 0.25, na.rm = TRUE),
    `75th` = ~quantile(., 0.75, na.rm = TRUE),
    Max = ~max(., na.rm = TRUE)
  ), .names = "{.col}_{.fn}")) %>%
  pivot_longer(cols = everything(), names_to = c("Feature", "Metric"), names_sep = "_(?=[^_]+$)") %>%
  pivot_wider(names_from = Metric, values_from = value)

# Rounding values
summary_stats <- summary_stats %>%
  mutate(across(where(is.numeric), ~round(., 2)))

# Displaying Table
kable(summary_stats, caption = "Summary Statistics", booktabs = TRUE, longtable = TRUE) %>%
  kable_styling(latex_options = c("hold_position", "repeat_header"), font_size = 10) %>%
  column_spec(1, width = "3.5cm") %>%
  column_spec(2:ncol(summary_stats), width = "2cm")
```

Table 3: Summary Statistics

Feature	Mean	Median	SD	Min	25th	75th
age	54.52	56.0	9.06	29	48	61.00

Table 3: Summary Statistics *(continued)*

Feature	Mean	Median	SD	Min	25th	75th
resting_blood_pressure	131.60	130.0	17.73	94	120	140.00
cholesterol	247.16	242.5	51.98	126	211	275.25
max_heart_rate_achieved	149.56	152.5	22.97	71	133	166.00
st_depression	1.06	0.8	1.17	0	0	1.65
num_major_vessels	0.68	0.0	0.94	0	0	1.00

1.9.1 - Dataset Summary Interpretation

```
interpretation_df <- data.frame(
  Variable = c(
    "Age",
    "Resting Blood Pressure",
    "Cholesterol",
    "Max Heart Rate Achieved",
    "ST Depression",
    "Number of Major Vessels"
  ),
  Summary = c(
    "Mean: 54.52, Median: 56, Min: 29, Max: 77",
    "Mean: 131.60, Median: 130, Min: 94, Max: 200",
    "Mean: 247.16, Median: 242.5, Min: 126, Max: 564",
    "Mean: 149.56, Median: 152.5, Min: 71, Max: 202",
    "Mean: 1.06, Median: 0.8, Min: 0, Max: 6.2",
    "Mean: 0.68, Median: 0, Min: 0, Max: 3"
  ),
  Interpretation = c(
    "Older age increases heart disease risk. Age is commonly used in risk prediction scores.",
    "Elevated blood pressure is a known contributor to heart strain and cardiovascular disease.",
    "Higher cholesterol can clog arteries and significantly increases cardiovascular risk.",
    "Lower max heart rate may suggest poor fitness or underlying heart issues during stress testing.",
    "ST depression reflects heart stress. Higher values are strong indicators of myocardial ischemia.",
    "Fewer major vessels visible may indicate arterial blockage or reduced coronary circulation."
  )
)

kable(interpretation_df,
      caption = "Summary Interpretation",
      booktabs = TRUE, longtable = TRUE) %>%
  kable_styling(latex_options = c("hold_position", "scale_down"), font_size = 9) %>%
  column_spec(1, width = "3.5cm") %>%
  column_spec(2, width = "5.5cm") %>%
  column_spec(3, width = "7.5cm")
```

Table 4: Summary Interpretation

Variable	Summary	Interpretation
Age	Mean: 54.52, Median: 56, Min: 29, Max: 77	Older age increases heart disease risk. Age is commonly used in risk prediction scores.
Resting Blood Pressure	Mean: 131.60, Median: 130, Min: 94, Max: 200	Elevated blood pressure is a known contributor to heart strain and cardiovascular disease.
Cholesterol	Mean: 247.16, Median: 242.5, Min: 126, Max: 564	Higher cholesterol can clog arteries and significantly increases cardiovascular risk.
Max Heart Rate Achieved	Mean: 149.56, Median: 152.5, Min: 71, Max: 202	Lower max heart rate may suggest poor fitness or underlying heart issues during stress testing.
ST Depression	Mean: 1.06, Median: 0.8, Min: 0, Max: 6.2	ST depression reflects heart stress. Higher values are strong indicators of myocardial ischemia.
Number of Major Vessels	Mean: 0.68, Median: 0, Min: 0, Max: 3	Fewer major vessels visible may indicate arterial blockage or reduced coronary circulation.

Interpretation

The features presented in Table 4 are important because they are continuous, and relevant for statistical summary like mean, median, std, etc. Moreover, they are clinically important for affecting heart disease outcomes. Lastly, these features are commonly used in risk scoring in ML models and medical research.

1.10 - Variable Classification by Data Type

```
continuous_vars <- c("age", "resting_blood_pressure", "cholesterol",
                     "max_heart_rate_achieved", "st_depression")
```

```
categorical_vars <- c("sex", "chest_pain_type", "fasting_blood_sugar",
                      "resting_electrocardiogram", "exercise_induced_angina",
                      "st_slope", "num_major_vessels", "thalassemia")
```

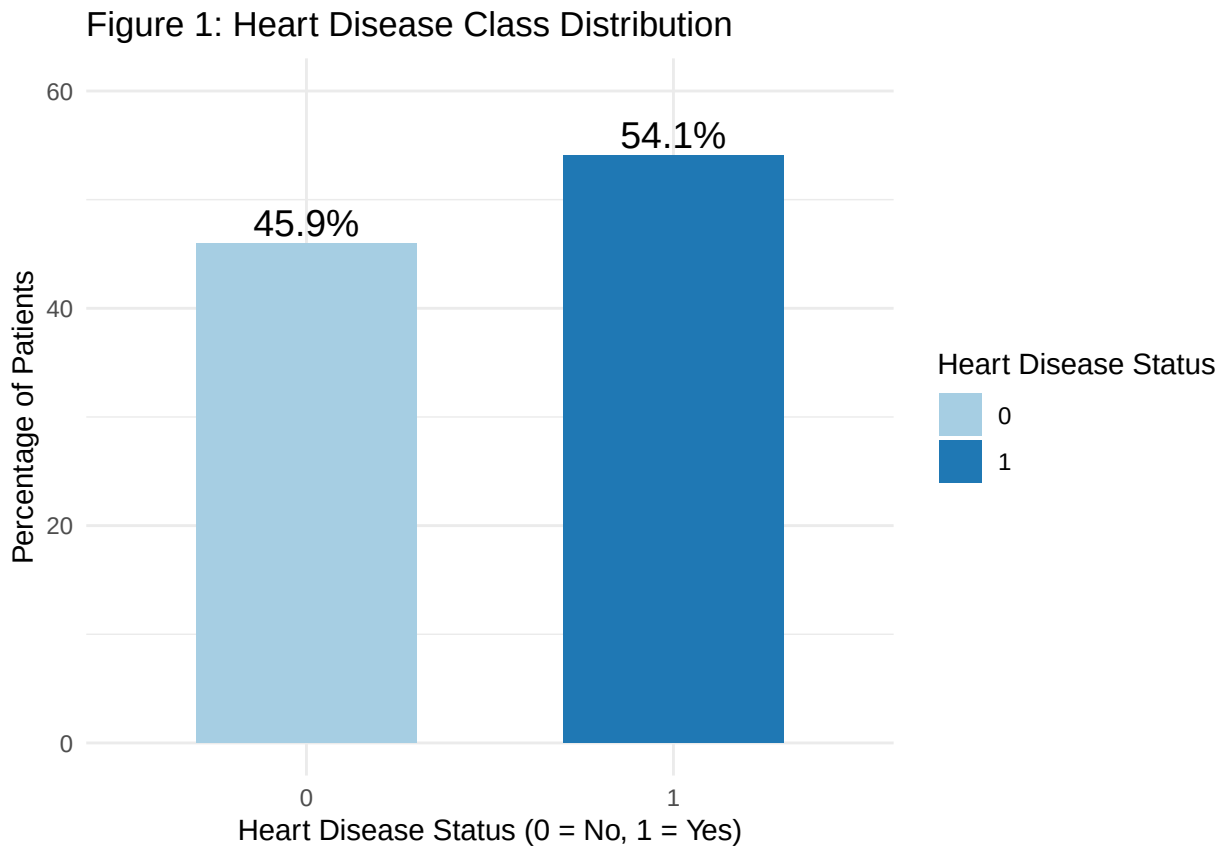
Section 2 - Visual Exploration for Key Variables

This section uses visualizations to explore how different variables relate to heart disease. Box plots and density plots are used for continuous variables to show patterns, differences in spread, and key statistics like medians and outliers. For categorical variables, a Cramér's V heatmap is used to measure the strength of their association with heart disease. These visual tools help identify which features might be useful for prediction.

2.1 - Bar Chart: Class Balance

```
target_dist <- health_data %>%
  count(target) %>%
  mutate(percent = n / sum(n) * 100)
```

```r
ggplot(target_dist, aes(x = as.factor(target), y = percent, fill = as.factor(target))) +
  geom_bar(stat = "identity", width = 0.6) +
  geom_text(aes(label = paste0(round(percent, 1), "%")),
            vjust = -0.3, size = 5) +
  labs(
    title = "Figure 1: Heart Disease Class Distribution",
    x = "Heart Disease Status (0 = No, 1 = Yes)",
    y = "Percentage of Patients",
    fill = "Heart Disease Status"
  ) +
  scale_fill_manual(values = c("#a6cee3", "#1f78b4")) +
  ylim(0, 60) +   # Extend y-axis so the label fits nicely
  theme_minimal()
```



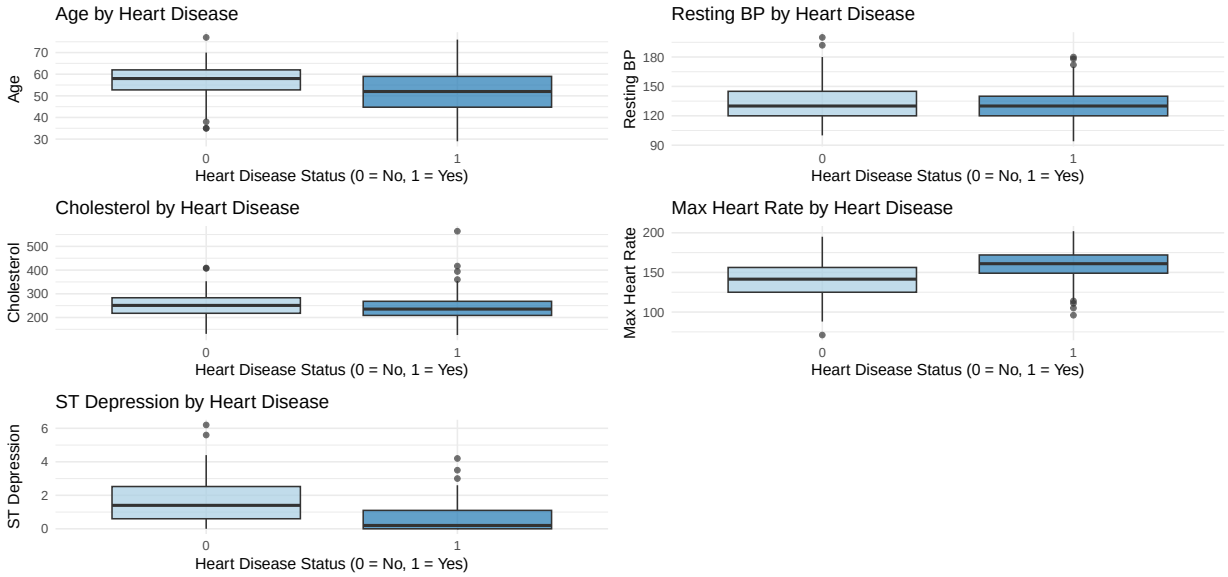
Interpretation:

Figure (1): Heart Disease Class Distribution is a Bar Chart that shows the percentage of patients with and without heart diseases in the data set. Approximately **54.1%** of the patients have **heart disease (target = 1)** whereas, **45.9%** do not **(target = 0)**. Knowing the distribution is important because it helps prevent bias toward one class during training of models. The distribution in this data set is fairly balanced which ensures that when I design the model, it will have enough example to learn patterns from both classes which will improve its overall predictive performance.

2.2 - Box Plot: Variable Distribution by Heart Disease

```r
# Defining display names for each plot

var_labels <- c(
  "age" = "Age",
  "resting_blood_pressure" = "Resting BP",
  "cholesterol" = "Cholesterol",
  "max_heart_rate_achieved" = "Max Heart Rate",
  "st_depression" = "ST Depression"
)

# Creating plots
plot_list <- lapply(names(var_labels), function(var) {
  ggplot(health_data, aes(x = as.factor(target), y = .data[[var]], fill = as.factor(target))) +
    geom_boxplot(alpha = 0.7) +
    labs(
      title = paste(var_labels[[var]], "by Heart Disease"),
      x = "Heart Disease Status (0 = No, 1 = Yes)",
      y = var_labels[[var]]
    ) +
    scale_fill_manual(values = c("#a6cee3", "#1f78b4")) +
    theme_minimal() +
    theme(legend.position = "none")
})

# Combining into 2-column layout and adding a title for the plot
combined_plot <- wrap_plots(plotlist = plot_list, ncol = 2) +
  plot_annotation(
    title = "Figure 2: Variables Distribution by Heart Disease Status",
    theme = theme(plot.title = element_text(size = 14, face = "bold", hjust = 0.5))
  )

combined_plot
```

Figure 2: Variables Distribution by Heart Disease Status



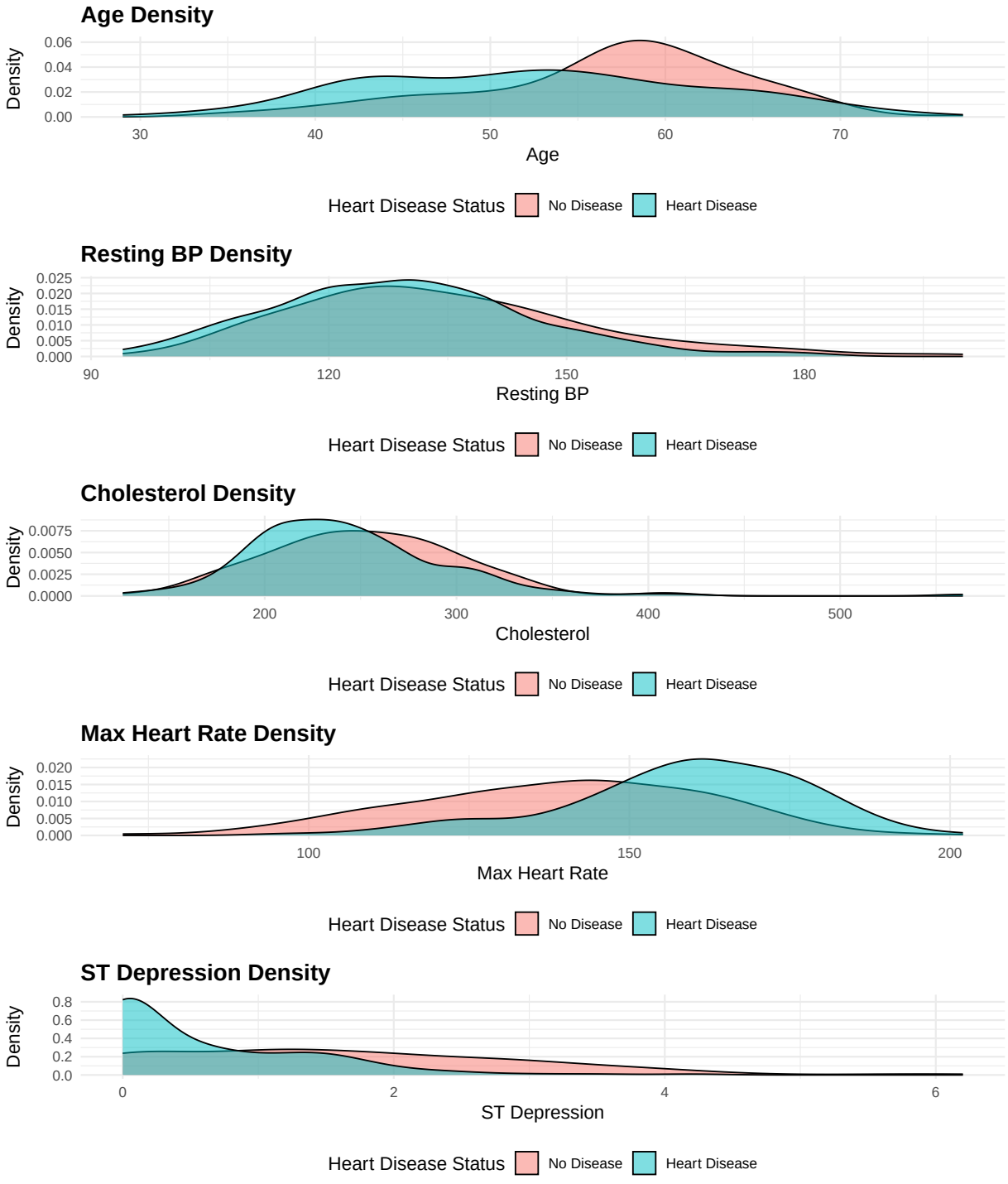
Interpretation of Box plots

Figure 2 compares the distribution of key variables for individuals with and without heart disease. People with heart disease are shown to have higher max heart rate, and lower st_depression levels when compared to individuals with no heart disease. Furthermore, cholesterol levels appear to be similar among both groups. Lastly, We can see that people with heart disease are younger and their resting bp is lower than the other group.

2.3 - Density Plot: Density Distribution by Heart Disease

```
# Mapping the labels

var_labels <- c(age = "Age", resting_blood_pressure = "Resting BP", cholesterol = "Cholesterol",
    max_heart_rate_achieved = "Max Heart Rate", st_depression = "ST Depression")

# Generating plots

density_plots <- lapply(names(var_labels), function(var) {
    ggplot(health_data, aes_string(x = var, fill = "factor(target)")) + geom_density(alpha = 0.5) +
        labs(title = paste(var_labels[[var]], "Density"), x = var_labels[[var]],
            y = "Density", fill = "Heart Disease Status") + scale_fill_manual(values = c("#F8766D",
        "#00BFC4"), labels = c("No Disease", "Heart Disease")) + theme_minimal(base_size = 13) +
        theme(plot.title = element_text(face = "bold"), legend.position = "bottom")
})

# Combining vertically with clear title

wrap_plots(density_plots, ncol = 1) + plot_annotation(title = "Figure 3: Density Distribution of Variabl
    theme = theme(plot.title = element_text(size = 14, face = "bold", hjust = 0.5)))
```

Figure 3: Density Distribution of Variables by Heart Disease Status



Interpretation:

In Figure 3 can observe how each continuous variable is distributed for people with and without heart disease. People with heart disease tend to be younger, while those without are a bit older. Resting blood pressure and cholesterol look mostly the same for both groups. However, people with heart disease reach higher maximum heart rates. Moreover, st_depression is lower in those with heart disease, which might be an indication of possible heart issues during stress. These patterns help us better understand how the features differ between

the two groups.

2.4 Importance of Using Both Box Plots and Density Plots

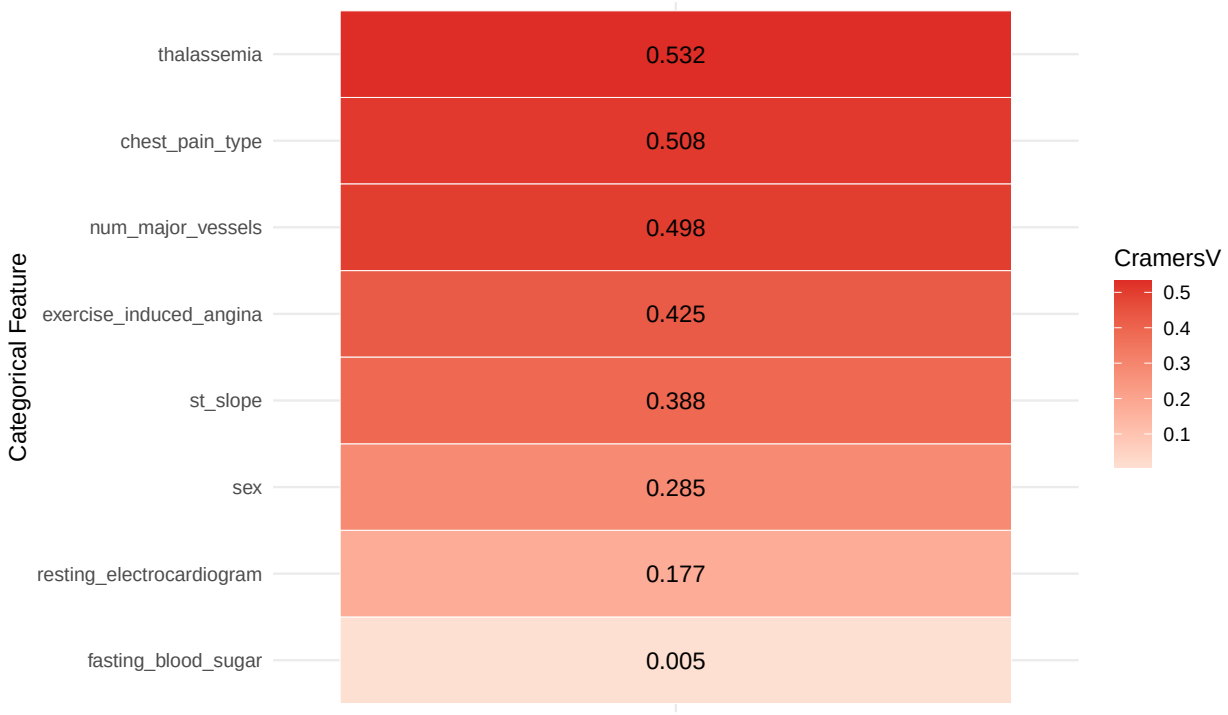
The use of box plots and density plots tools in exploratory data analysis is important because of the purposes both tools serve. Density plots show the overall shape of the data, helping us understand how each feature is distributed and whether it is skewed or symmetrical. This is important for spotting patterns or unusual spreads. On the other hand, box plots provide a summary of key statistics like the median, quartiles, and outliers, which makes comparing central tendencies and data spread between groups. Using both plots together gives a clearer, more complete view of how variables behave across different heart disease outcomes.

2.5 - Categorical Feature Association with Heart Disease

(Cramér's V Heatmap)

```
# Calculating Cramer's V for each categorical variable vs target

cramer_df <- data.frame(Variable = character(), CramersV = numeric())

for (var in categorical_vars) {
    tbl <- table(health_data[[var]], health_data$target)
    cramer_value <- DescTools::CramerV(tbl)
    cramer_df <- rbind(cramer_df, data.frame(Variable = var, CramersV = round(cramer_value,
        3)))
}

# Plotting the heatmap

ggplot(cramer_df, aes(x = "", y = reorder(Variable, CramersV), fill = CramersV)) +
    geom_tile(color = "white") + geom_text(aes(label = CramersV), size = 4) + scale_fill_gradient(low =
    high = "#DE2D26") + labs(title = "Figure 4: Categorical Features Association with Heart Disease (Cra
    x = NULL, y = "Categorical Feature") + theme_minimal() + theme(axis.text.x = element_blank(),
    axis.ticks.x = element_blank(), legend.position = "right", plot.title = element_text(face = "bold",
        size = 14, hjust = 0.5))
```

Figure 4: Categorical Features Association with Heart Disease (Cramér´s V)



Interpretation:

Figure 4 shows how strongly each categorical feature is related to heart disease using Cramér's V. Thalassemia (0.532), chest pain type (0.508), and number of major vessels (0.498) have the strongest links. This means they might be important features to help predict heart disease.

Section 3 - Statisitcal Testing

This section presents (3) statistical tests to evaluate the relationship between selected features and the presence of heart disease. We use (2) types of statistical tests: Point-Biserial Correlation & T-Test for continuous variables and the Chi-Square Test for categorical variables.

3.1 - Point-Biserial Correlation Analysis

Purpose: Point-biserial correlation tells how strongly each continuous variable is related to the binary target (0/1).

- Null Hypothesis (H0): There is no linear relationship between the continuous feature and heart disease status.
- Alternative Hypothesis (H1): There is a linear relationship between the continuous feature and heart disease status.

```r
# Converting Target to Factor

health_data$target <- factor(health_data$target, labels = c("No Disease", "Heart Disease"))


# Creating correlation summary table

correlation_df <- data.frame(Feature = continuous_vars, Correlation = sapply(continuous_vars,
    function(var) {
        round(cor(health_data[[var]], as.numeric(health_data$target == "Heart Disease")),
            3)
    }), stringsAsFactors = FALSE)

# output the table

kable(correlation_df, caption = "Correlation Between Continuous Features and Heart Disease",
    booktabs = TRUE, longtable = TRUE) %>%
    kable_styling(latex_options = c("hold_position", "scale_down"), font_size = 9)
```

Table 5: Correlation Between Continuous Features and Heart Disease

	Feature	Correlation
age	age	-0.225
resting_blood_pressure	resting_blood_pressure	-0.149
cholesterol	cholesterol	-0.077
max_heart_rate_achieved	max_heart_rate_achieved	0.427
st_depression	st_depression	-0.429

Interpretation:

Table (5) shows how strongly each continuous feature is related to heart disease using point-biserial correlation. Max heart rate (0.427) and ST depression (-0.429) have the strongest associations, suggesting they may be useful predictors. Additionaly, Age (-0.225) and blood pressure (-0.149) show weaker relationships, while cholesterol (-0.077) shows almost no link. These results suggest that only some continuous features have a meaningful linear relationship with heart disease.

3.2 T-Test: Comparing Means Across Heart Disease Status

Purpose: A t-test checks if the mean of the variable differs significantly between those with and without heart disease.

- Null Hypothesis (H0): The means of the two groups (with and without heart disease) are equal.
- Alternative Hypothesis (H1): The means of the two groups are different.

```r
# Creating data frame to store t-test
t_test_df <- data.frame(Feature = continuous_vars, P_Value = sapply(continuous_vars,
    function(var) {
        ttest <- t.test(health_data[[var]] ~ health_data$target)
        formatC(ttest$p.value, format = "e", digits = 4)
```

```
    }), stringsAsFactors = FALSE)

# Output the table
kable(t_test_df, caption = "T-Test Results for Continuous Features by Heart Disease Status",
    booktabs = TRUE, longtable = TRUE) %>%
    kable_styling(latex_options = c("hold_position", "scale_down"), font_size = 9)
```

Table 6: T-Test Results for Continuous Features by Heart Disease Status

	Feature	P_Value
age	age	7.1701e-05
resting_blood_pressure	resting_blood_pressure	1.1233e-02
cholesterol	cholesterol	1.8628e-01
max_heart_rate_achieved	max_heart_rate_achieved	4.6279e-14
st_depression	st_depression	2.1392e-13

Interpretation:

Table (6) shows the results of t-tests comparing continuous features across heart disease status. Age (p = 7.17e-05), resting blood pressure (p = 0.0112), max heart rate achieved (p = 4.63e-14), and ST depression (p = 2.14e-13) show statistically significant differences between the two groups. Therefore, we can conclude that these features may help in identifying heart disease. Cholesterol (p = 0.1863) does not show a significant difference, meaning it may not be useful on its own for prediction.

3.3 Chi-Square Test: Association b/w Categorical Variables & Heart Disease

Purpose: The chi-square test evaluates whether there is a significant association between each categorical feature and heart disease status.

- Null Hypothesis (H): The categorical feature and heart disease status are independent.
- Alternative Hypothesis (H): The categorical feature and heart disease status are dependent.

```
# Creating data frame to store chi-square test results
chi_sq_df <- data.frame(Feature = categorical_vars, P_Value = sapply(categorical_vars,
    function(var) {
        tbl <- table(health_data[[var]], health_data$target)
        chi_test <- chisq.test(tbl)
        formatC(chi_test$p.value, format = "e", digits = 4)
    }), stringsAsFactors = FALSE)

# Output the Table

kable(chi_sq_df, caption = "Chi-Square Test Results for Categorical Features by Heart Disease Status",
    booktabs = TRUE, longtable = TRUE) %>%
    kable_styling(latex_options = c("hold_position", "scale_down"), font_size = 9)
```

Table 7: Chi-Square Test Results for Categorical Features by Heart Disease Status

	Feature	P_Value
sex	sex	1.7189e-06
chest_pain_type	chest_pain_type	1.7678e-16
fasting_blood_sugar	fasting_blood_sugar	1.0000e+00
resting_electrocardiogram	resting_electrocardiogram	9.7433e-03
exercise_induced_angina	exercise_induced_angina	6.5171e-13
st_slope	st_slope	2.1158e-10
num_major_vessels	num_major_vessels	7.9956e-16
thalassemia	thalassemia	6.4666e-19

Interpretation: Table (7) shows the p-values from chi-square tests between categorical features and heart disease. Thalassemia (p = 6.47e-19), chest pain type (p = 1.77e-16), number of major vessels (p = 7.99e-16), ST slope (p = 2.12e-10), and exercise-induced angina (p = 6.52e-13) all show strong associations with heart disease. However, Sex (p = 1.72e-06) and resting electrocardiogram (p = 9.74e-03) are also significant, but not as others. Lastly, Fasting blood sugar (p = 1.00) shows no association, suggesting it may not be a useful predictor in this data set.

EDA Summary

In this project, I performed exploratory data analysis (EDA) to better understand which factors are linked to heart disease. I analyzed both continuous and categorical variables to find patterns and relationships that could support building a predictive model.

For continuous features, I used point-biserial correlation and t-tests to compare differences between patients with and without heart disease. The results showed that maximum heart rate achieved and ST depression were the most strongly associated with heart disease. Age also showed a meaningful link. On the other hand, cholesterol and resting blood pressure had weaker or non-significant associations.

For categorical features, I used bar plots and chi-square tests to assess how each variable is related to heart disease status. The analysis revealed that chest pain type, ST slope, number of major vessels, thalassemia, and exercise-induced angina had strong associations with heart disease. Sex and resting electrocardiogram also showed some significance. However, fasting blood sugar did not appear to be a useful feature.

Conclusively, the EDA helped identify the most relevant features for predicting heart disease, which will guide the feature selection process in the modeling stage.

Next Steps:

The next step involves building classification models based on the features identified in this EDA. The modeling process is documented in a separate file titled HeartDisease_Modeling_swh0085.Rmd.