Nonparametric Analysis of the Cleveland Heart Disease Dataset

Ekta Kumari — 241080067
Khushi — 241080072
Sejal Dubey — 241080092
Sidhant — 241080098
Sujal Yadav — 241080101

Contents

1 Introduction

Cardiovascular diseases (CVDs) remain the leading cause of death globally, with coronary artery disease posing one of the most significant public health challenges due to its high prevalence and complex risk factors. Early identification of key physiological indicators is crucial for prevention and effective clinical decision-making.

This study utilizes the **Cleveland Heart Disease dataset** (Detrano et al., 1989) from the UCI Machine Learning Repository, which contains **303 patient records** and **14 clinical variables** such as age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, and maximum heart rate achieved. The target variable indicates the presence or absence of heart disease. Given that medical data often violate normality assumptions, **nonparametric statistical methods** were employed to ensure robustness and interpretability.

Analyses were conducted in **R**, combining exploratory and inferential techniques. The **Mann–Whitney U**, **Kruskal–Wallis**, **Friedman**, and **Chi-square** tests assessed group differences, while **Spearman's rho** and **Kendall's tau** quantified rank-based associations among continuous variables. Results revealed significant relationships between **chest pain type**, **cholesterol**, and **maximum heart rate** with heart disease presence. Overall, the findings demonstrate the effectiveness of nonparametric methods in identifying clinically meaningful patterns and support their application in **data-driven cardiovascular research** and **predictive healthcare analytics**.

2 Dataset Description

This study utilizes the **Cleveland Heart Disease dataset** from the **UCI Machine Learning Repository**, originally contributed by **R. Detrano, A. Jánosi, W. Steinbrunn**, and colleagues (1988). The data were collected from multiple hospitals across Cleveland, Hungary, Switzerland, and the VA Long Beach facility; however, only the **Cleveland subset** (303 patient records) is widely analyzed due to its completeness and standardized preprocessing.

The dataset captures **multivariate clinical and demographic information** relevant to the diagnosis of **coronary artery disease (CAD)**. It includes **14 attributes** covering demographic variables such as **age** and **sex**, and clinical measures including **chest pain type (cp)**, **resting blood pressure (trestbps)**, **serum cholesterol (chol)**, **fasting blood sugar (fbs)**, **resting ECG results (restecg)**, **maximum heart rate (thalach)**, **exercise-induced angina (exang)**, **ST depression (oldpeak)**, **slope of the ST segment (slope)**, **number of major vessels (ca)**, and **Thalassemia result (thal)**. The **target variable (num)**, originally coded from 0–4 to represent disease severity, was **binarized** into: `0` (no heart disease) and `1` (presence of heart disease, combining 1–4).

A small number of missing values, denoted by `"?"`, were handled through **median imputation** for numeric variables and **mode imputation** for categorical ones. After cleaning, all features were recoded into suitable numeric or factor formats for analysis. Given its **diverse variable structure** and **clinical significance**, the Cleveland dataset remains a

benchmark for both **classification modeling** and **nonparametric statistical inference** in cardiovascular research.

3 Data Pre-processing

The data preprocessing was conducted in R to prepare the Cleveland Heart Disease dataset for analysis. It involved ensuring data integrity, handling missing values, and formatting variables for statistical testing and visualization.

3.1 Environment Setup

Essential R packages (e.g., tidyverse, ggplot2, GGally, DataExplorer, corrplot, Amelia) were loaded. Missing packages were automatically installed to maintain reproducibility.

3.2 Loading the Dataset

The dataset (processed.cleveland.data) was imported using read.csv() with na.strings = "?" to convert missing entries to NA. Column names were assigned based on UCI documentation.

3.3 Target Variable Cleaning

The target variable, originally labeled num in the raw data, indicates the presence of heart disease with integer values from 0 to 4. To simplify analysis and align with prior research, the target was binarized as follows: 0 → No Disease 1, 2, 3, 4 → Disease Present The target variable was converted to a factor with labels "No Disease" and "Disease" for compatibility with statistical functions.

```
## --- Target Variable Summary ---

##
## No Disease    Disease
##        164        139
```

The output above shows that the dataset contains 164 patients without heart disease and 139 patients with heart disease. This distribution indicates that the dataset is relatively balanced, with both classes (No Disease and Disease) having comparable sample sizes.

3.4 Conversion of Categorical Variables

Categorical variables such as sex, cp, fbs, restecg, exang, slope, thal, and target were converted into factor types to ensure correct handling in analyses.

3.5 Handling Missing Values

Missing values were minimal and handled using median imputation for numeric features and mode imputation for categorical ones, preserving data integrity without bias.

```
## --- Missing Value Summary ---

##      age      sex       cp trestbps     chol      fbs  restecg  thalach
##        0        0        0        0        0        0        0        0
##    exang  oldpeak    slope       ca     thal   target
##        0        0        0        4        2        0
```

Median imputation was applied to numeric variables to avoid distortion from extreme values, while mode imputation was used for categorical variables to retain the most representative category.

3.6 Data Overview and Summary

After cleaning, the dataset contained 303 observations and 14 variables. The structure confirmed appropriate data types, and summary statistics indicated realistic clinical ranges for all measures. The dataset is now fully prepared for exploratory and inferential analysis.

```
## --- Dataset Overview ---

## Observations: 303 | Variables: 14

## --- Structure ---

## 'data.frame':    303 obs. of  14 variables:
##  $ age     : num  63 67 67 37 41 56 62 57 63 53 ...
##  $ sex     : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 1 1 2 2 ...
##  $ cp      : Factor w/ 4 levels "1","2","3","4": 1 4 4 3 2 2 4 4 4 4 ...
##  $ trestbps: num  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol    : num  233 286 229 250 204 236 268 354 254 203 ...
##  $ fbs     : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 2 ...
##  $ restecg : Factor w/ 3 levels "0","1","2": 3 3 3 1 3 1 3 1 3 3 ...
##  $ thalach : num  150 108 129 187 172 178 160 163 147 155 ...
##  $ exang   : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 1 2 1 2 ...
##  $ oldpeak : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  $ slope   : Factor w/ 3 levels "1","2","3": 3 2 2 3 1 1 3 1 2 3 ...
##  $ ca      : num  0 3 2 0 0 0 2 0 1 0 ...
##  $ thal    : Factor w/ 3 levels "3","6","7": 2 1 3 1 1 1 1 1 3 3 ...
##  $ target  : Factor w/ 2 levels "Disease","No Disease": 2 1 1 2 2 2 1 2 1 1 ...

##
## --- Summary Statistics ---

##       age        sex     cp         trestbps          chol       fbs
##  Min.   :29.00   0: 97   1: 23   Min.   : 94.0   Min.   :126.0   0:258
##  1st Qu.:48.00   1:206   2: 50   1st Qu.:120.0   1st Qu.:211.0   1: 45
##  Median :56.00           3: 86   Median :130.0   Median :241.0
##  Mean   :54.44           4:144   Mean   :131.7   Mean   :246.7
##  3rd Qu.:61.00                   3rd Qu.:140.0   3rd Qu.:275.0
##  Max.   :77.00                   Max.   :200.0   Max.   :564.0
```

```
##  restecg    thalach      exang      oldpeak     slope         ca
##  0:151   Min.   : 71.0   0:204   Min.   :0.00   1:142   Min.   :0.0000
##  1:  4   1st Qu.:133.5   1: 99   1st Qu.:0.00   2:140   1st Qu.:0.0000
##  2:148   Median :153.0           Median :0.80   3: 21   Median :0.0000
##          Mean   :149.6           Mean   :1.04           Mean   :0.6634
##          3rd Qu.:166.0           3rd Qu.:1.60           3rd Qu.:1.0000
##          Max.   :202.0           Max.   :6.20           Max.   :3.0000
##  thal           target
##  3:168   Disease   :139
##  6: 18   No Disease:164
##  7:117
##
##
##
```

4 Exploratory Data Analysis

The numeric summary presents descriptive statistics for the continuous variables — **age**, **trestbps** (resting blood pressure), **chol** (serum cholesterol), **thalach** (maximum heart rate achieved), **oldpeak** (ST depression induced by exercise), and **ca** (number of major vessels).

Overall, **age** averages **54.4 years (SD = 9.0)**, representing a predominantly middle-aged sample. **Resting blood pressure** averages **131.7 mmHg**, indicating mild hypertension. **Cholesterol** shows a broad range (**126–564 mg/dl**) with positive skewness (**1.12**), reflecting a few patients with elevated cholesterol. **Maximum heart rate (thalach)** averages **149.6 bpm**, slightly left-skewed (**–0.53**), consistent with age-related decline. Both **oldpeak** and **ca** exhibit positive skewness, implying most patients have lower ST depression and fewer affected vessels, with a few showing higher risk values.

These results suggest **clinically diverse yet plausible patient profiles**. Mild skewness in several measures supports the use of **nonparametric statistical tests** for subsequent inference.

(Note: Graphical outputs from the EDA—such as histograms, boxplots, and density plots—are not displayed here but will appear in the Testing section to support the corresponding hypothesis tests.)

Outliers were identified using the **Interquartile Range (IQR)** method, marking values outside 1.5 × IQR from the first or third quartile. Results showed no outliers in **age**, while **trestbps** had 9, **chol** 5, **thalach** 1, **oldpeak** 5, and **ca** 20 — the highest among all. These were **retained**, as they represent realistic clinical extremes (e.g., high cholesterol or vessel counts) rather than data errors, thus preserving the dataset's **medical validity**.

A multicollinearity check was performed using the correlation matrix, identifying variable pairs with absolute correlation above 0.8.No pairs exceeded this threshold, suggesting no serious multicollinearity among predictors. This confirms that the variables capture distinct physiological information suitable for downstream modeling and nonparametric analysis.

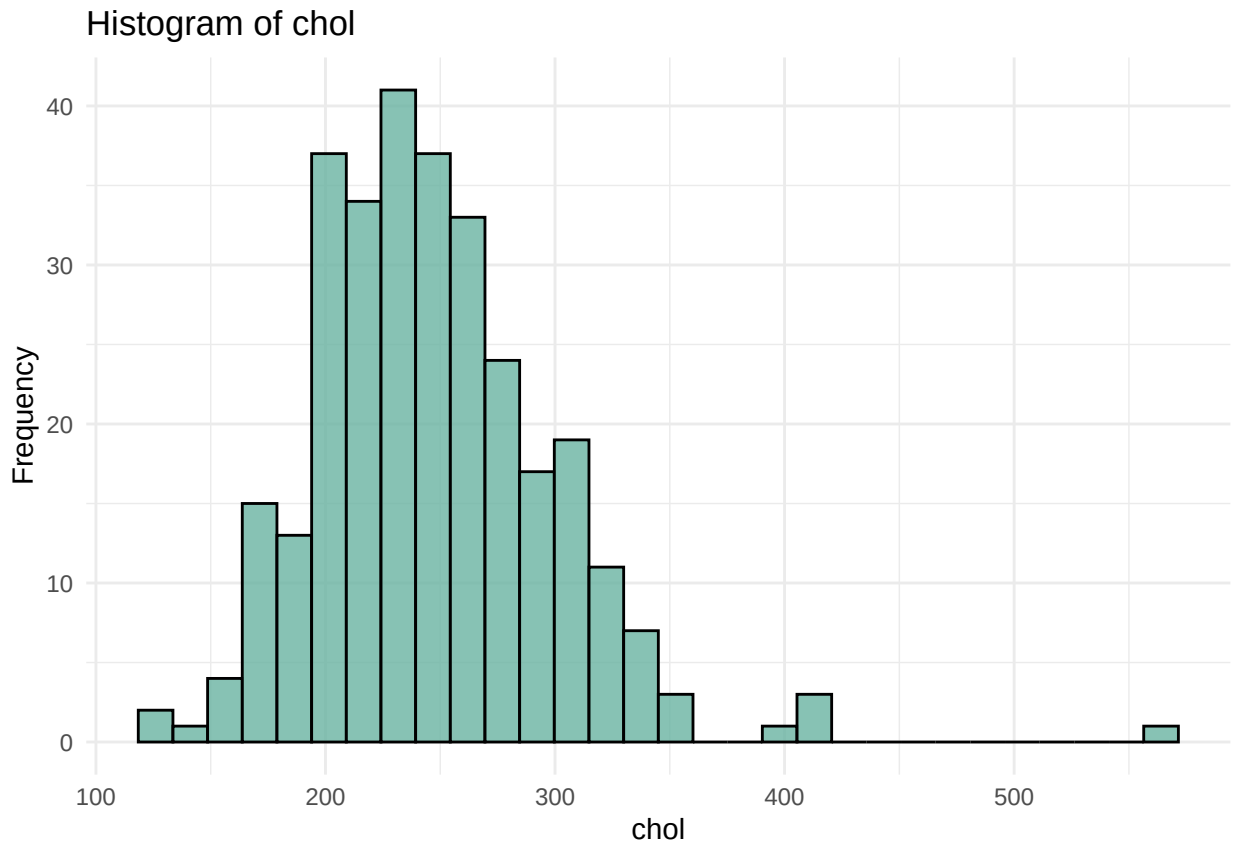
5 Non-parametric Tests and Results

5.1 One-sample Kolmogorov-Smirnov Test

Before applying parametric or nonparametric methods, it is essential to assess whether the **chol (serum cholesterol)** variable follows a **normal distribution**. The **Kolmogorov–Smirnov (K–S) test** was used for this purpose, comparing the observed data's empirical distribution with a theoretical normal distribution of the same mean and standard deviation. This helps determine if the assumption of normality holds for subsequent statistical testing.

The null and alternative hypotheses are defined as: 0:Cholesterol levels follow a normal distribution. 1:Cholesterol levels do not follow a normal distribution.

```
v <- "chol"
p <- ggplot(heart, aes_string(x = v)) +
    geom_histogram(fill = "#69b3a2", color = "black", bins = 30, alpha = 0.8) +
    theme_minimal() +
    labs(title = paste("Histogram of", v), x = v, y = "Frequency")
  print(p)
```




```
chol <- heart$chol
mean_chol <- mean(chol)
sd_chol   <- sd(chol)
```

```r
ks_result <- ks.test(chol, "pnorm", mean = mean_chol, sd = sd_chol)
print(ks_result)
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  chol
## D = 0.055395, p-value = 0.3103
## alternative hypothesis: two-sided
```

The p-value = 0.3103, which is greater than the significance level (= 0.05). Hence, we fail to reject the null hypothesis. This indicates that the cholesterol (chol) variable approximately follows a normal distribution.

5.2 Chi-square Goodness-of-fit Test

To assess the **demographic balance** of the dataset, a **Chi-square Goodness-of-Fit test** was applied to determine whether **age** is uniformly distributed across patients. The variable was divided into **five equal-width bins (29–77 years)** for comparison.

Hypotheses:

- **H :** Age is uniformly distributed across groups.
- **H :** Age distribution is not uniform.

A bar plot was generated to compare observed and expected frequencies, revealing potential **age clustering** within the sample.

```r
heart$age_group <- cut(heart$age,
                       breaks = 5,    # 5 equal-width intervals
                       include.lowest = TRUE)

table(heart$age_group)
```

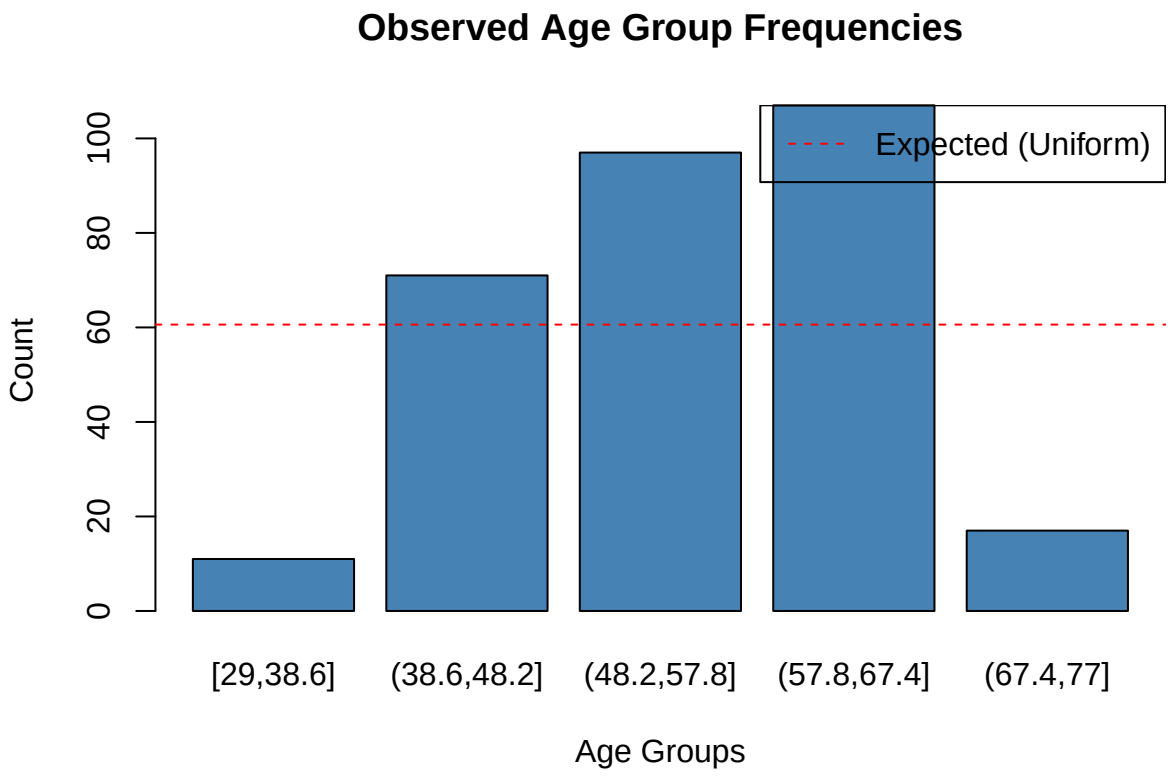
```
##
##   [29,38.6] (38.6,48.2] (48.2,57.8] (57.8,67.4]   (67.4,77]
##          11          71          97         107          17
```

```r
observed <- table(heart$age_group)
expected <- rep(sum(observed) / length(observed), length(observed))

chisq_test <- chisq.test(x = observed, p = rep(1/length(observed), length(observed)))

chisq_test
```

```
##
##  Chi-squared test for given probabilities
##
```

```
## data:  observed
## X-squared = 131.14, df = 4, p-value < 2.2e-16
```

```r
barplot(observed,
        main = "Observed Age Group Frequencies",
        ylab = "Count", xlab = "Age Groups",
        col = "steelblue")
abline(h = mean(observed), col = "red", lty = 2)
legend("topright", legend = "Expected (Uniform)", col = "red", lty = 2)
```



The Chi-square statistic (131.14, df = 4, p < 2.2×10^{1}) indicates a highly significant result, leading to rejection of **H** . Thus, **age is not uniformly distributed** — most patients fall in the **40–65 year range**, with fewer younger and older individuals. This aligns with medical evidence that **heart disease risk rises sharply in middle age**, explaining the concentration of cases in this group.

5.3 Two-sample Kolmogorov-Smirnov Test

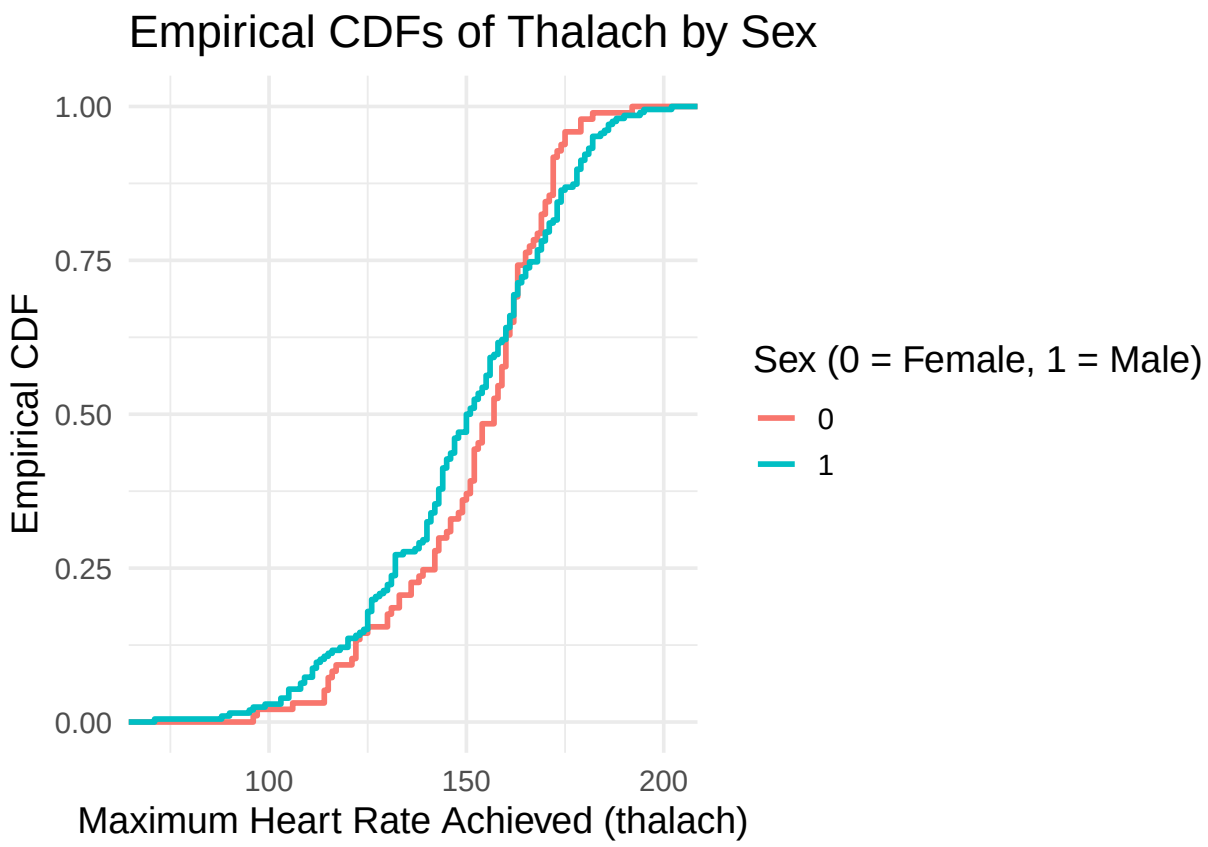
A two-sample Kolmogorov–Smirnov (K–S) test was used to compare the distributions of **maximum heart rate (thalach)** between males and females. This nonparametric test assesses whether two independent samples differ in overall distribution without assuming normality or equal variance.

H : Thalach distribution is the same for males and females. **H :** Thalach distributions differ between males and females.

```
thalach_male   <- heart$thalach[heart$sex == 1]
thalach_female <- heart$thalach[heart$sex == 0]
ks_result <- ks.test(thalach_male, thalach_female)
print(ks_result)
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  thalach_male and thalach_female
## D = 0.13127, p-value = 0.2058
## alternative hypothesis: two-sided
```

```
ggplot(heart, aes(x = thalach, color = factor(sex))) +
  stat_ecdf(size = 1) +
  labs(title = "Empirical CDFs of Thalach by Sex",
       x = "Maximum Heart Rate Achieved (thalach)",
       y = "Empirical CDF",
       color = "Sex (0 = Female, 1 = Male)") +
  theme_minimal(base_size = 14)
```



The K–S statistic (D = 0.1313) with a p-value of 0.2058 (> 0.05) indicates no significant difference between male and female thalach distributions. Thus, we fail to reject H , concluding that both sexes exhibit similar maximum heart rate patterns during exercise in this dataset.

5.4 Sign Test

The **oldpeak** variable measures ST segment depression on an ECG after exercise relative to rest. A value of 0 indicates normal recovery, while positive values suggest ischemia (reduced blood flow). To assess whether ST depression occurs systematically, a **non-parametric Sign Test** was used to test if the median differs from zero.

H : Median(oldpeak) = 0 **H :** Median(oldpeak) 0

This determines whether ST depression is symmetrically distributed around zero or shows a consistent positive shift indicating ischemic response.

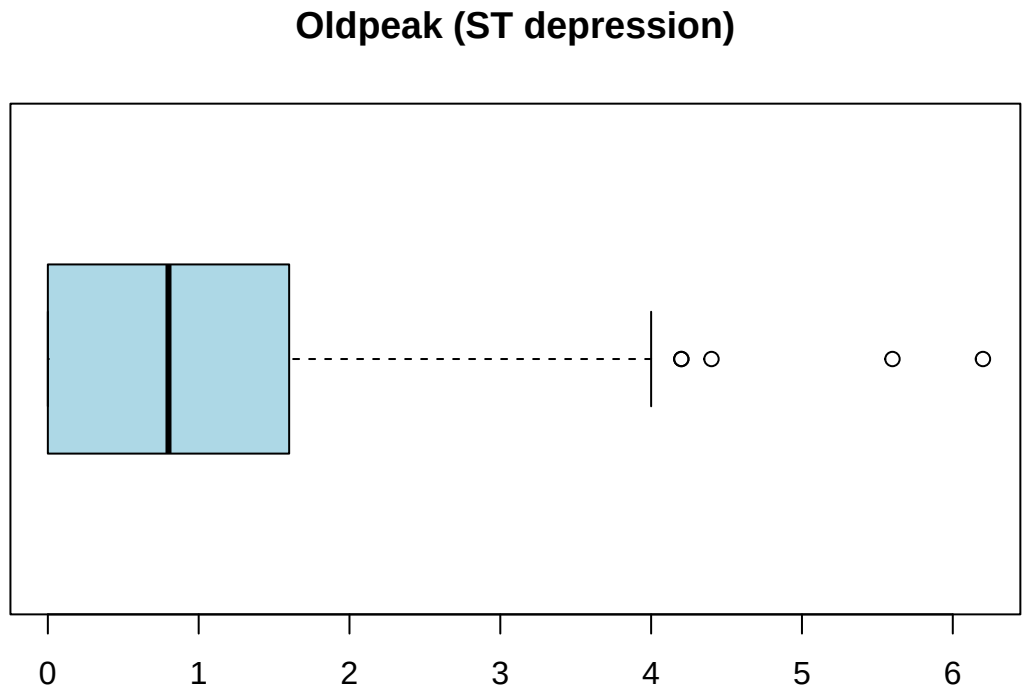
```
oldpeak <- heart$oldpeak
n_positive <- sum(oldpeak > 0)  # count of positive differences
n_negative <- sum(oldpeak < 0)  # count of negative differences
n_total <- n_positive + n_negative  # exclude exact zeros
sign_test_result <- binom.test(n_positive, n_total, p = 0.5,
                               alternative = "two.sided")
print(sign_test_result)
```

```
##
##  Exact binomial test
##
## data:  n_positive and n_total
## number of successes = 204, number of trials = 204, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.9820798 1.0000000
## sample estimates:
## probability of success
##                      1
```

```
summary(oldpeak)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.80    1.04    1.60    6.20
```

```r
boxplot(oldpeak, horizontal = TRUE, main = "Oldpeak (ST depression)", col = "lightblue")
```



The p-value ($< 2.2e{-}16$) is far below 0.05, so the null hypothesis is rejected — the median **oldpeak** differs from zero. Since nearly all nonzero values are positive (n = 204), most patients show **ST depression** during exercise, indicating ischemic response typical of individuals evaluated for potential heart disease.

5.5 Mann-Whitney U Test

1. The **cholesterol (chol)** variable, measured in mg/dl, was first tested for normality using the Kolmogorov–Smirnov test (p = 0.3103), confirming no significant deviation from normality. Despite this, the **Mann–Whitney U test** was chosen over the t-test for robustness against outliers and minor non-normality common in medical data. **Hypotheses:**

- **H :** Median cholesterol levels are equal for males and females.
- **H :** Median cholesterol levels differ between males and females.

```r
table(heart$sex)
```

```
##
##   0   1
##  97 206
```

```r
heart$sex <- factor(heart$sex, levels = c(0, 1), labels = c("Female", "Male"))

wilcox_test_result <- wilcox.test(chol ~ sex, data = heart,
                                  alternative = "two.sided",
                                  exact = FALSE, # set FALSE for large samples
                                  conf.int = TRUE,
                                  conf.level = 0.95)

wilcox_test_result
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  chol by sex
## W = 11900, p-value = 0.007307
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##   4.999999 29.999942
## sample estimates:
## difference in location
##               16.99991
```

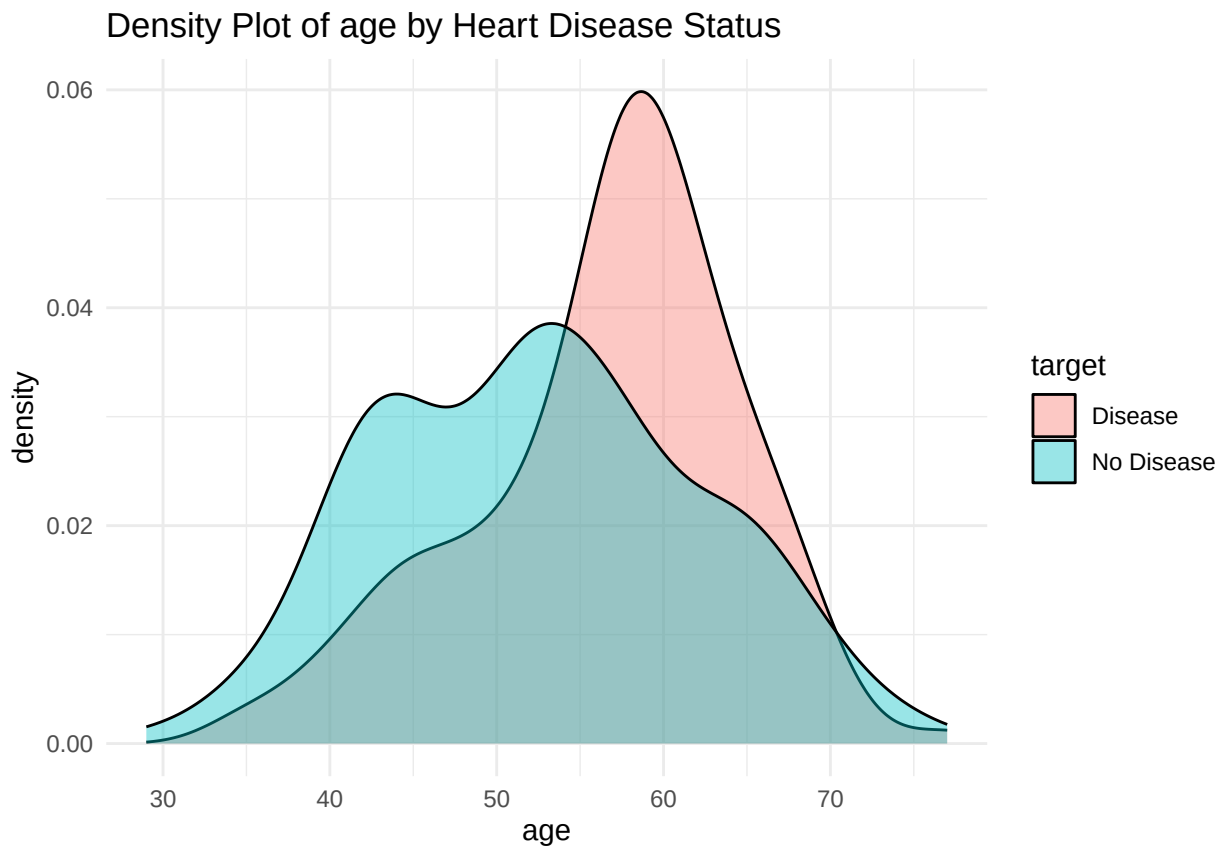
With a p-value of 0.0073 (< 0.05), we reject **H** , indicating a significant difference in median cholesterol between males and females. The median shift (~17 mg/dl) shows that females generally have higher cholesterol levels than males in this dataset.

2. The variable age (in years) was compared between patients with and without heart disease to assess whether age is associated with disease presence.

```r
wilcox_age_result <- wilcox.test(age ~ target, data = heart,
alternative = "two.sided",
conf.int = TRUE, conf.level = 0.95)
wilcox_age_result
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  age by target
## W = 14522, p-value = 3.917e-05
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  2.000066 6.999968
## sample estimates:
## difference in location
##               4.000024
```

```r
v <- "age"
p1 <- ggplot(heart, aes_string(x = v, fill = "target")) +
    geom_density(alpha = 0.4) +
    theme_minimal() +
    labs(title = paste("Density Plot of", v, "by Heart Disease Status"))
  print(p1)
```



3. Resting blood pressure (**trestbps**, in mm Hg) reflects cardiac workload and risk of heart disease. To test if **trestbps** differs between patients **with** and **without** heart disease, a **Mann–Whitney U test** was used — suitable for independent groups and non-normal data.

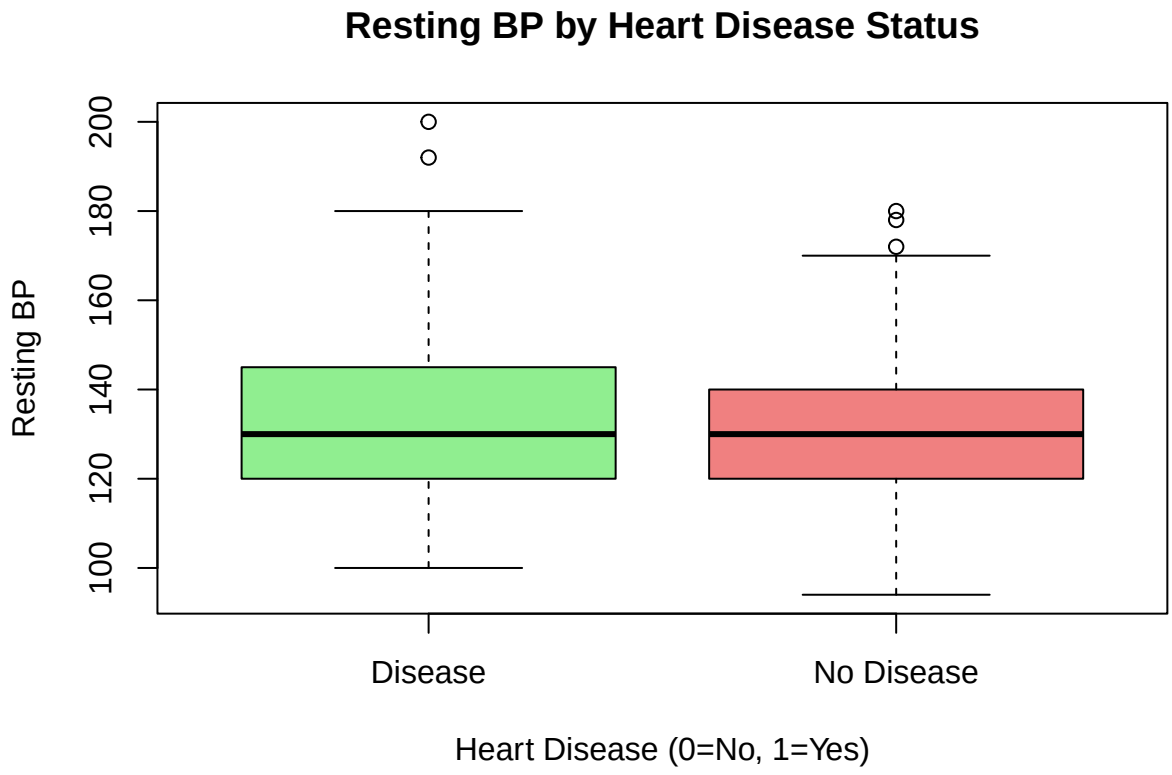
Hypotheses:

- **H :** Median resting BP is the same for both groups.
- **H :** Median resting BP differs between groups.

```r
wilcox_test_result <- wilcox.test(trestbps ~ target, data = heart,
                                  alternative = "two.sided",
                                  conf.int = TRUE,
                                  exact = FALSE)   # exact=FALSE recommended for larger

wilcox_test_result
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  trestbps by target
## W = 13086, p-value = 0.02597
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  0.0000513303 8.0000481335
## sample estimates:
## difference in location
##               4.000007
```

```
boxplot(trestbps ~ target, data = heart, main = "Resting BP by Heart Disease Status",
        xlab = "Heart Disease (0=No, 1=Yes)", ylab = "Resting BP", col = c("lightgreen",
```



The p-value ($0.0259 < 0.05$) leads to rejecting **H** , indicating a significant difference in resting blood pressure between patients with and without heart disease. The median difference is about **4 mm Hg**, with higher BP among diseased patients, suggesting greater cardiac strain or underlying hypertension.

5.6 Kruskal-Wallis Test

1. The **chest pain type (cp)** variable has four categories — typical angina, atypical angina, non-anginal pain, and asymptomatic. To examine whether these pain types vary by **age**, the **Kruskal–Wallis test** (a nonparametric alternative to one-way ANOVA) was applied, as age is continuous and not strictly normal.

Hypotheses:

- **H :** Age distribution is the same across all chest pain types.
- **H :** At least one chest pain group differs significantly in age distribution.

```
heart$cp <- as.factor(heart$cp)

kruskal_result <- kruskal.test(age ~ cp, data = heart)
print(kruskal_result)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  age by cp
## Kruskal-Wallis chi-squared = 11.619, df = 3, p-value = 0.008811
```

The p-value ($0.0088 < 0.05$) indicates a significant difference in age across chest pain types, leading to rejection of **H** . This suggests that chest pain patterns vary by age — younger patients more often experience atypical or non-anginal pain, while older individuals tend to have asymptomatic chest pain associated with heart disease.

2. The variable **thalach** (maximum heart rate) and **age** are continuous, with heart rate typically declining as age increases. To test whether **thalach** differs significantly across age groups, the **Kruskal–Wallis test**—a nonparametric alternative to ANOVA—was applied. **H :** Thalach distribution is identical across age groups. **H :** At least one age group differs significantly in thalach distribution.

```
heart$age <- as.numeric(heart$age)
heart$thalach <- as.numeric(heart$thalach)

heart$age_group <- cut(
  heart$age,
  breaks = c(30, 40, 50, 60, Inf),
  labels = c("30-40", "40-50", "50-60", "60+"),
  right = FALSE
)

table(heart$age_group)
```

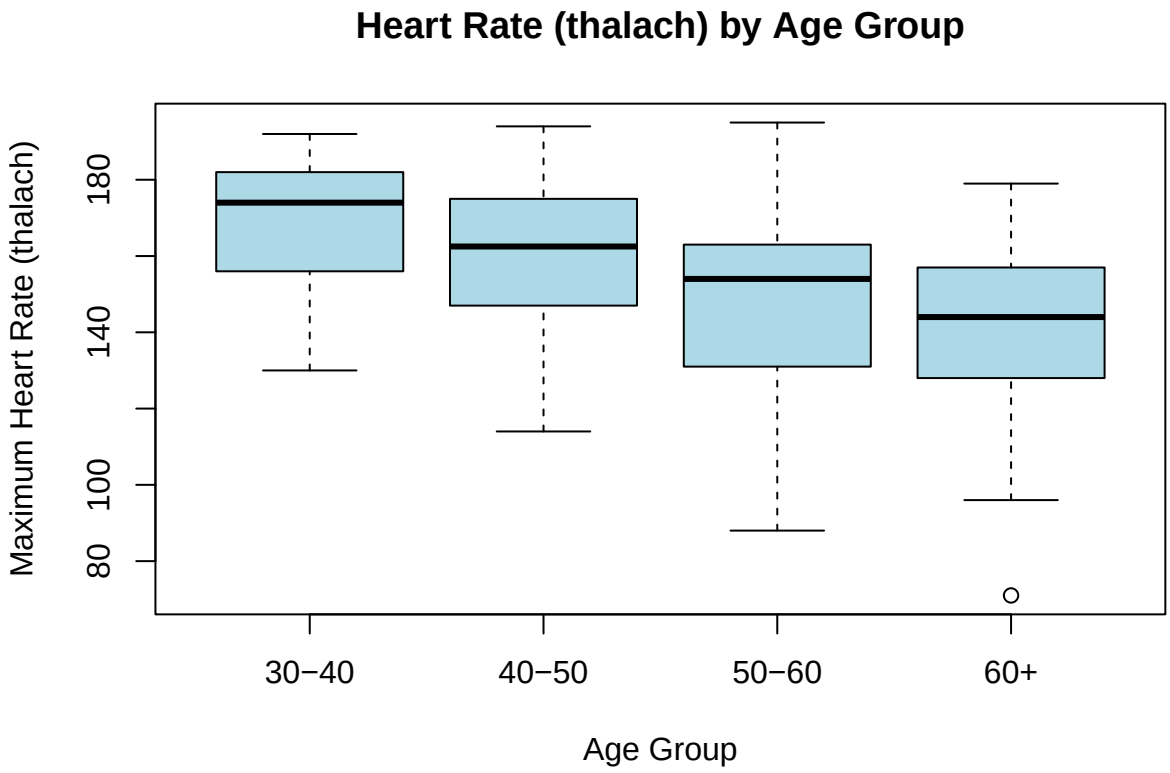
```
##
## 30-40 40-50 50-60   60+
```

```
##    14    72   125    91
```

```
kruskal_result <- kruskal.test(thalach ~ age_group, data = heart)
kruskal_result
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  thalach by age_group
## Kruskal-Wallis chi-squared = 42.09, df = 3, p-value = 3.84e-09
```

```
boxplot(thalach ~ age_group, data = heart,
        col = "lightblue", main = "Heart Rate (thalach) by Age Group",
        xlab = "Age Group", ylab = "Maximum Heart Rate (thalach)")
```



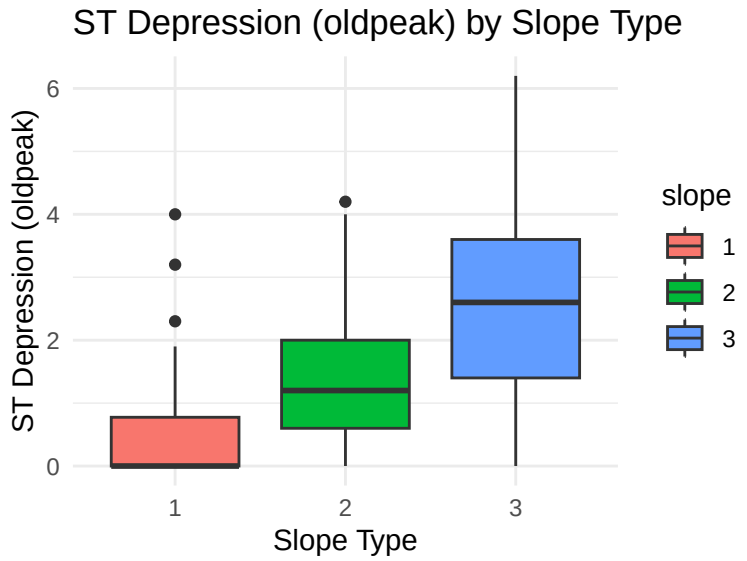
The p-value < 0.001 indicates a highly significant difference in **thalach** across age groups. Younger patients (30–50 years) achieved higher maximum heart rates, while older individuals (60+) showed lower values, reflecting the well-established age-related decline in cardiovascular efficiency.

5.7 Mood's Test

The variable **oldpeak** (ST depression after exercise) indicates myocardial ischemia risk, while **slope** represents the ST segment type — *Upsloping (1)*, *Flat (2)*, or *Downsloping (3)*. Using **Mood's test**, we assessed whether the variability of **oldpeak** differs across slope categories.

H : The spread of oldpeak is the same across slope groups. **H :** At least one slope group shows significantly different variability.

```
# slope 1 vs 2
m12 <- mood.test(heart$oldpeak[heart$slope == 1],
                 heart$oldpeak[heart$slope == 2])

# slope 1 vs 3
m13 <- mood.test(heart$oldpeak[heart$slope == 1],
                 heart$oldpeak[heart$slope == 3])

# slope 2 vs 3
m23 <- mood.test(heart$oldpeak[heart$slope == 2],
                 heart$oldpeak[heart$slope == 3])

ggplot(heart, aes(x = slope, y = oldpeak, fill = slope)) +
  geom_boxplot() +
  labs(
    title = "ST Depression (oldpeak) by Slope Type",
    x = "Slope Type",
    y = "ST Depression (oldpeak)"
  ) +
  theme_minimal()
```



```r
pairwise_results <- data.frame(
  Comparison = c("Slope 1 vs 2", "Slope 1 vs 3", "Slope 2 vs 3"),
  Statistic = c(m12$statistic, m13$statistic, m23$statistic),
  P_value   = c(m12$p.value, m13$p.value, m23$p.value)
)

pairwise_results
```

```
##     Comparison  Statistic      P_value
## 1 Slope 1 vs 2 -0.1129362 9.100811e-01
## 2 Slope 1 vs 3 -7.2493835 4.186729e-13
## 3 Slope 2 vs 3 -2.7894848 5.279198e-03
```

The p-values for slope 1 vs 3 and slope 2 vs 3 are < 0.05, indicating that the **downsloping group (slope = 3)** exhibits significantly greater variability in **ST depression** than the other types. In contrast, **upsloping (1)** and **flat (2)** groups show no notable difference. Clinically, this implies that downsloping ST segments are linked to more severe and variable ischemic responses during exercise.

5.8 Van-der Waerden Test

The **Thalium stress test (thal)** assesses cardiac blood flow with three levels — **3: Normal**, **6: Fixed defect** (permanent damage), and **7: Reversible defect** (reduced flow during exercise, recovery at rest). The variable **thalach** (maximum heart rate) indicates cardiovascular fitness. A **Van der Waerden nonparametric test** was used to examine whether heart rate responses differ across Thallium test outcomes.

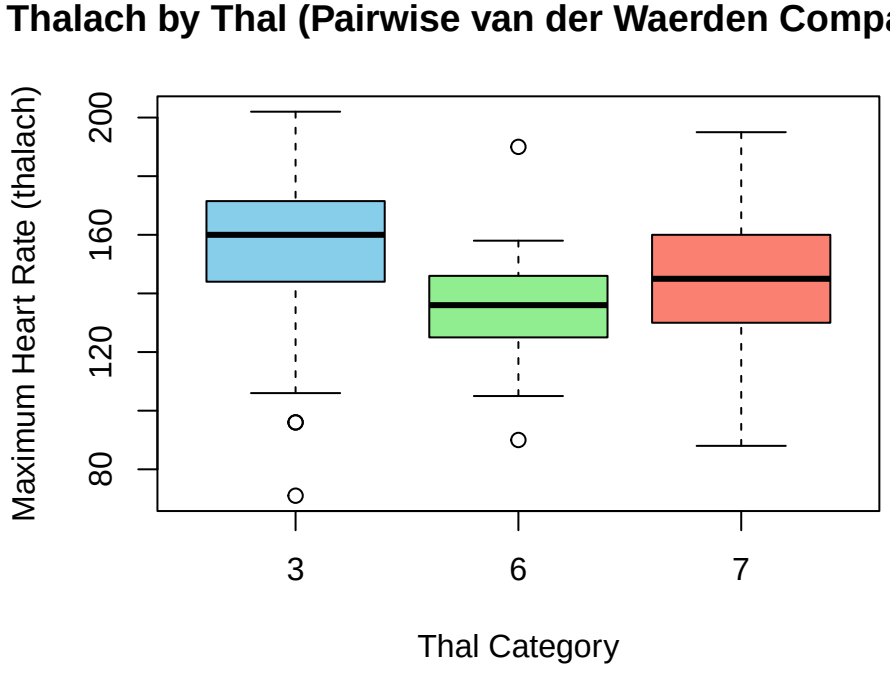
H : Mean ranks of *thalach* are equal across all *thal* levels. **H :** At least one *thal* group differs significantly.

```r
library(PMCMRplus)

heart$thal <- as.factor(heart$thal)

pairwise_combos <- combn(levels(heart$thal), 2, simplify = FALSE)

results <- data.frame(
  Group1 = character(),
  Group2 = character(),
  Statistic = numeric(),
  P_Value = numeric(),
  stringsAsFactors = FALSE
)

for (pair in pairwise_combos) {
  g1 <- pair[1]
```

```r
  g2 <- pair[2]
  sub_data <- subset(heart, thal %in% c(g1, g2))
  sub_data$thal <- droplevels(sub_data$thal)
  test_res <- vanWaerdenTest(thalach ~ thal, data = sub_data)
  stat <- unname(test_res$statistic)
  pval <- unname(test_res$p.value)
  results <- rbind(results, data.frame(Group1 = g1, Group2 = g2,
                                       Statistic = stat, P_Value = pval))
}

print(results)
```

```
##   Group1 Group2 Statistic      P_Value
## 1      3      6 11.987830 5.354910e-04
## 2      3      7 21.171442 4.199761e-06
## 3      6      7  2.464468 1.164471e-01
```

```r
boxplot(thalach ~ thal, data = heart,
        main = "Thalach by Thal (Pairwise van der Waerden Comparisons)",
        xlab = "Thal Category", ylab = "Maximum Heart Rate (thalach)",
        col = c("skyblue", "lightgreen", "salmon"))
```



The Van der Waerden test ($p < 0.05$) revealed significant differences in **thalach** across Thallium test categories. Patients with **normal scans (thal = 3)** achieved higher heart rates, while those with **reversible defects (thal = 7)** showed the lowest values, indicating

poorer exercise tolerance — consistent with reduced cardiac function under stress.

5.9 Spearman's Rank Correlation

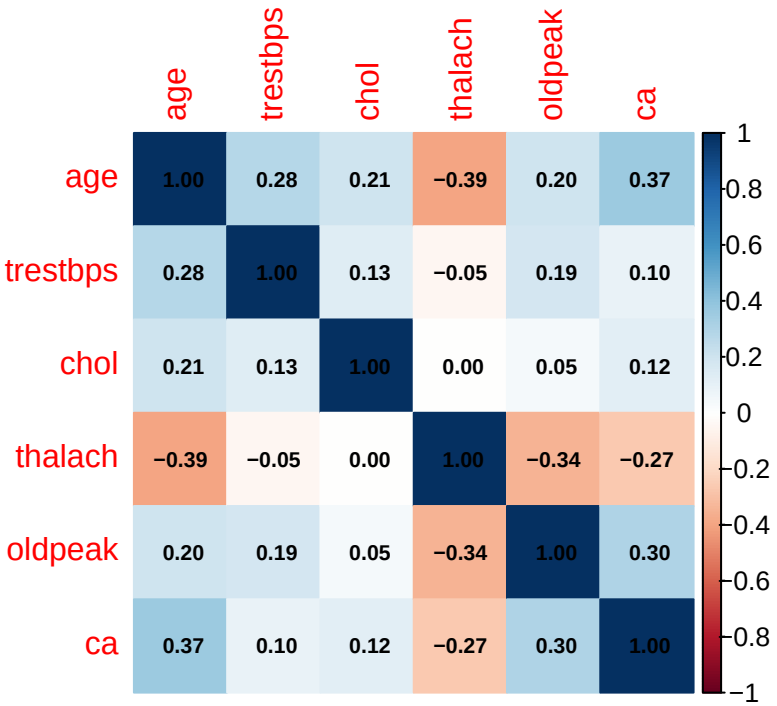
Cholesterol (**chol**) and resting blood pressure (**trestbps**) are key continuous indicators of cardiovascular risk. To assess their association, the **Spearman's rank correlation**—a nonparametric measure robust to non-normality—was applied.

Hypotheses: H : No monotonic relationship exists between cholesterol and resting blood pressure (= 0). H : A significant monotonic relationship exists between the two (0).

```
spearman_result <- cor.test(
  heart$chol,
  heart$trestbps,
  method = "spearman",
  exact = FALSE
)
spearman_result
```

```
##
##  Spearman's rank correlation rho
##
## data:  heart$chol and heart$trestbps
## S = 4006524, p-value = 0.018
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1358367
```

```
corrplot(corr_matrix, method = "color", addCoef.col = "black", number.cex = 0.7)
```



The p-value (0.018) < 0.05 indicates a significant positive association between serum cholesterol and resting blood pressure. The Spearman's rho ($= 0.136$) shows a weak but meaningful monotonic relationship — patients with higher cholesterol generally have slightly higher resting BP, suggesting mild co-variation in cardiovascular risk factors.

5.10 Kendall's Tau Correlation

The **maximum heart rate (thalach)** typically decreases with age, reflecting natural cardiovascular decline. To confirm this relationship, the **Kendall's Tau** nonparametric correlation test was used, as it handles non-normal data and ties effectively.

H : No association between age and thalach ($= 0$) **H :** A significant association exists between age and thalach (0).

```
kendall_result <- cor.test(
  heart$age,
  heart$thalach,
  method = "kendall",
  exact = FALSE  # set to FALSE for large samples
)

kendall_result
```

```
##
##  Kendall's rank correlation tau
```

```
##
## data:  heart$age and heart$thalach
## z = -7.0077, p-value = 2.423e-12
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## -0.2756964
```

The negative tau (–0.276) and highly significant p-value (2.42×10^{12}) indicate a moderate, statistically significant inverse relationship between age and maximum heart rate, confirming that heart rate decreases with increasing age.

6 Conclusion

This study analyzed the **Cleveland Heart Disease dataset** using extensive preprocessing, exploratory data analysis, and a suite of **non-parametric statistical tests** to uncover key patterns in cardiovascular risk factors. Results show that the dataset represents a **middle-aged population** (mean 54 years) with mild hypertension and variable cholesterol levels. While cholesterol followed a near-normal distribution, other variables such as *oldpeak* and *ca* were skewed, supporting non-parametric inference.

Significant findings include: **systematically elevated ST depression** (*oldpeak*) from the Sign Test, **higher cholesterol in females** and **elevated resting blood pressure in diseased patients** from Mann–Whitney tests, and **age-related differences in chest pain type and maximum heart rate** from Kruskal–Wallis tests. The **Van der Waerden test** revealed reduced *thalach* in patients with abnormal Thallium results, and correlations indicated a **weak positive link** between cholesterol and resting BP, but a **strong negative link** between age and heart rate (= –0.276, $p < 0.001$).

Overall, the analysis demonstrates that **non-parametric methods provide a reliable, assumption-free approach** for exploring complex, clinically skewed data. The findings reaffirm known medical trends—particularly the inverse relationship between age and heart rate, and the association of higher cholesterol and blood pressure with heart disease risk—highlighting the value of statistical modeling in cardiovascular research.