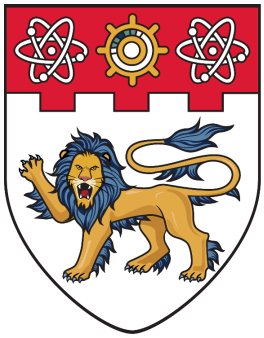


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

MH3511 Data Analysis with Computer
Group Project

Heart Disease by the Numbers: Critical Risk Factors

Name	Matriculation Number
Dana Yak	U2321776D
Goh Qing Wen	U2322556F
Jason Liu Wei Hang	U2321989B
Jolie Tan	U2322595G
Muhammad Aidil Firdaus Bin Rahmat	U2321295B

Abstract:
Cardiovascular disease remains a leading cause of mortality globally, necessitating effective diagnostic tools. This study investigates the Cleveland heart disease dataset, which comprises various patient attributes including age, sex, chest pain type, resting blood pressure, cholesterol levels, and electrocardiographic results. Utilizing these features, we aim to explore the underlying patterns and potential predictors associated with the presence or absence of heart disease. By applying data analysis techniques to this comprehensive dataset, we aim to identify key indicators and gain a better understanding of the factors influencing cardiac health. This analysis can potentially inform the development of more accurate and efficient diagnostic models for heart disease.

Table of Contents

1. Introduction

Cardiovascular diseases (CVD) such as heart disease remains a leading cause of mortality and morbidity globally (WHO, 2021). It places a significant burden on healthcare systems and standard-of-living. Accurate diagnosis and risk stratification are crucial for effective detection and timely intervention.

In our project, we use a dataset of clinical data encompassing a range of patient characteristics and diagnostic test results, which offers valuable insights into the factors associated with heart disease. Based on our dataset, we aim to investigate several key relationships and potential indicators of heart disease. Specifically, we will explore the following questions:

- Is the risk of heart disease of an individual dependent on their age and cholesterol levels?
- Can the risk of heart disease be effectively detected based on the maximum heart rate achieved during exercise?
- Is a higher level of resting blood sugar a significant indicator of heart disease?
- Can different types of chest pain experienced serve as an indicator of heart disease?

This report will detail the Cleveland dataset's characteristics and the R-based analytical methods used to address these research questions. For each question, we will employ appropriate statistical analyses and visualizations to derive meaningful conclusions with clear explanations.

2. Data Description

The database is from UCI Machine Learning Repository, and the original database contains 76 attributes, but all published experiments by ML researchers so far use just the Cleveland dataset (https://archive.ics.uci.edu/dataset/45/heart+disease). As such, this study uses the Cleveland heart disease dataset, a resource containing various clinical and physiological measurements from individuals with and without heart disease.

S.No	Attribute Name	Description	Range of Values
1	Age	Age of the person in years	29 to 79
2	Sex	Gender of the person [1: Male, 0: Female]	0, 1
3	Cp	Chest pain type [1-Typical Type 1 Angina 2- Atypical Type Angina 3-Non-angina pain 4-Asymptomatic)	1, 2, 3, 4
4	Trestbps	Resting Blood Pressure in mm Hg	94 to 200
5	Chol	Serum cholesterol in mg/dl	126 to 564
6	Fbs	Fasting Blood Sugar in mg/dl	0, 1
7	Restecg	Resting Electrocardiographic Results	0, 1, 2
8	Thalach	Maximum Heart Rate Achieved	71 to 202
9	Exang	Exercise Induced Angina	0, 1
10	OldPeak	ST depression induced by exercise relative to rest	1 to 3
11	Slope	Slope of the Peak Exercise ST segment	1, 2, 3
12	Ca	Number of major vessels colored by fluoroscopy	0 to 3
13	Thal	3 – Normal, 6 – Fixed Defect, 7 – Reversible Defect	3, 6, 7
14	Num	Class Attribute	0 or 1

[FIGURE 1: List of 14 variables we are using]

The dataset, heart.csv, has 303 observations and 14 attributes that describe individual patient characteristics relevant to cardiovascular health and heart disease diagnosis. This includes categorical and continuous variables that capture demographics, symptoms, medical history, diagnostic indicators such as "oldpeak" (exercise-induced ST depression), and other relevant factors like cholesterol levels, fasting blood sugar, and chest pain types. The outcome variable of interest is Num, a categorical variable with values from 0 (no presence) to 4.

3. Description and Cleaning of Dataset

In this section, we look into the data in more detail, investigating each variable individually to look for possible outliers, then performing cleaning to ensure the quality and integrity of the data before analysis.

3.1 Overview

We examined each variable for missing data, potential outliers, and then corrected inconsistencies and applied appropriate transformations to avoid highly skewed data. Summary statistics and visualisations such as histograms and boxplots were used to support the cleaning decisions.

Initial inspection using the summary() and str() functions revealed that the dataset was mostly clean, but a few steps were needed for preprocessing.

Although the dataset had told us which variables had missing data (ca and thal), we checked again and handled them accordingly.

Categorical variables need to be factorised or one-hot encoded first before data processing. After converting the categorical variables to factor data type, we can see that the missing values are stored as "?" instead of "na", so we cannot use the usual is.na function. Thus, we replace "?" values throughout the dataframe with "NA" so that we can use is.na to drop rows with NA values. As “?” is registered as a factor level, we drop this unused factor level after converting “?” to NA values. Next, we check for NA in each column after replacement and confirm that variables "ca" and "thal" have 4 and 2 missing values respectively.

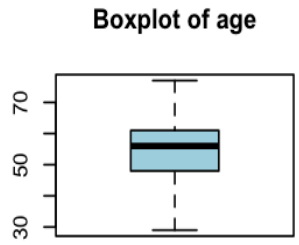
```
      age        sex     cp          trestbps          chol       fbs     restecg
 Min.   :29.00   0: 97   1: 23   Min.   : 94.0   Min.   :126.0   0:258   0:151
 1st Qu.:48.00   1:206   2: 50   1st Qu.:120.0   1st Qu.:211.0   1: 45   1:  4
 Median :56.00           3: 86   Median :130.0   Median :241.0           2:148
 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
    thalach       exang       oldpeak      slope      ca        thal      num
 Min.   : 71.0   0:204   Min.   :0.00   1:142   0.0 :176   3.0 :166   0:164
 1st Qu.:133.5   1: 99   1st Qu.:0.00   2:140   1.0 : 65   6.0 : 18   1: 55
 Median :153.0           Median :0.80   3: 21   2.0 : 38   7.0 :117   2: 36
 Mean   :149.6           Mean   :1.04           3.0 : 20   NA's:  2   3: 35
 3rd Qu.:166.0           3rd Qu.:1.60           NA's:  4              4: 13
 Max.   :202.0           Max.   :6.20
```

[FIGURE 2: Summary of attributes in heart_data dataset]

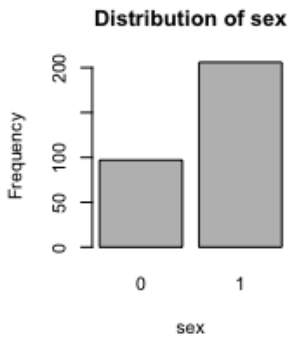
3.2 Summary and Cleaning of Variables

The histogram, the boxplot, any transformation applied or outliers removed from the variables are tabulated in the following subsections.

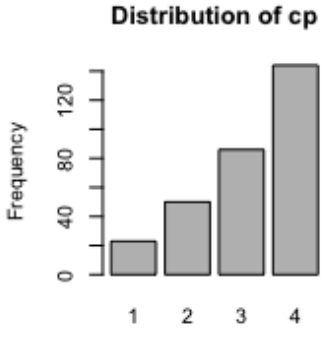
3.2.1 The main variable of interest: Age (numerical)

		• No missing or implausible values found. • The distribution is approximately normal. • Ages range between 28 to 77 years. • No cleaning needed.

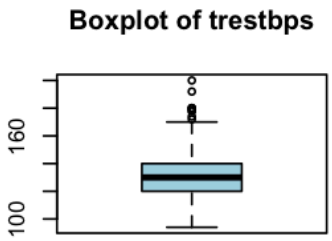
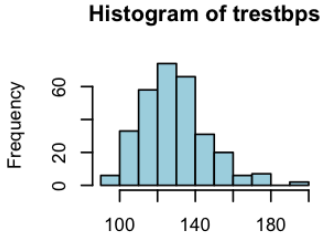
3.2.2 Sex (categorical)

	• Two levels: 0 - male, 1 - female • The dataset consists of more females than males • Converted to a factor data type before any data processing. • No missing entries, so no imputation or dropping of rows required.

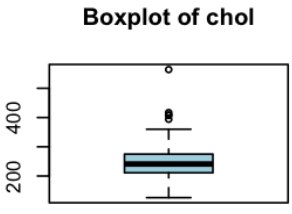
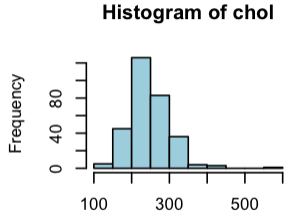
3.2.3 ChestPainType (cp) (categorical)

	• Four distinct categories:
	  ○ 1-Typical Type 1 Angina
	  ○ 2- Atypical Type Angina
	  ○ 3-Non-angina pain
	  ○ 4-Asymptomatic
	• Asymptomatic is the most common type.
	• Converted to a factor data type before any data processing.
	• No missing entries, so no imputation or dropping of rows required.

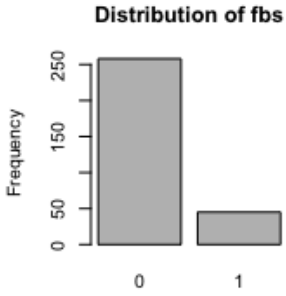
3.2.4 Resting Blood Pressure, mmHg (trestbps) (numerical)

 **before** removing outliers	 **after** removing outliers	• Values range between 94-200 mmHg are within expected physiological ranges (80–200 mmHg). • Outliers were imputed with the median value. • No transformation was applied.

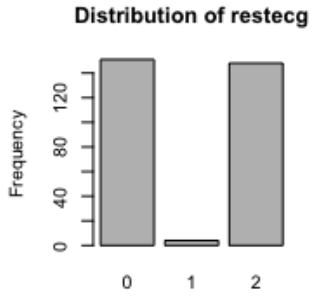
3.2.5 Serum cholesterol, mg/dl (chol) (numerical)

 **before** removing outliers	 **after** removing outliers	• Outliers were imputed with the median value. • A log transformation (log p) was considered but ultimately not applied as skewness was mild post-cleaning.

3.2.6 Fasting Blood Sugar > 120 mg/dl (fbs) (categorical)

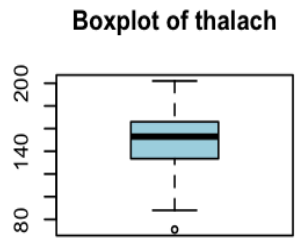
	• A binary variable with values 0 and 1, with 0 meaning fbs <120 mg/dl • The majority have a normal level of fasting blood sugar. • Converted to a factor data type before any data processing. • No missing entries, so no imputation or dropping of rows required.

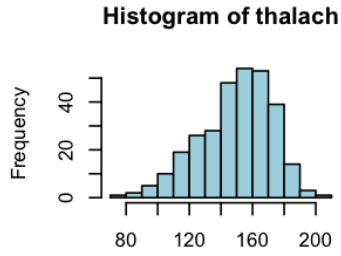
3.2.7 Resting Electrocardiographic Results (restecg) (categorical)



- Categorical with levels: 0 - Normal, 1 - ST, 2 - LVH
- Normal ST segment:
  - Represents the period of ventricular repolarization.
- ST-segment depression:
  - Can indicate ischemia (reduced blood flow to the heart muscle), LVH or other conditions.
- Left Ventricular Hypertrophy (LVH)
  - Often indicates a heart problem or increases the risk of developing one, and is an independent risk factor for cardiovascular disease.
- Majority had either a normal resting ECG or LVH.
- Converted to a factor data type before any data processing.
- No missing entries, so no imputation or dropping of rows required.

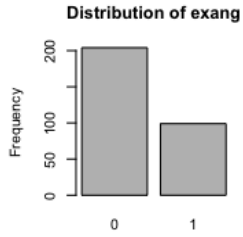
3.2.8 Maximum Heart Rate Achieved, bpm (thalach)





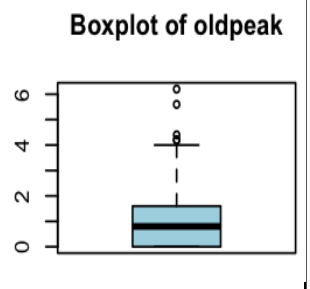
- Range: 60 - 202 bpm.
- Medical context suggests low-bpm (~70 bpm) are plausible in patients with severe heart disease.
- Hence, outliers are retained without imputation.

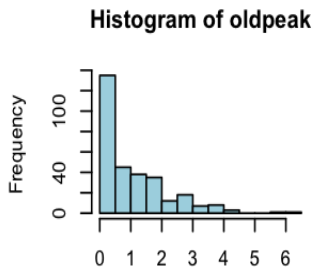
3.2.9 Exercise Induced Angina (exang) (categorical)



- Binary categorical variable: 0 - No, 1 - Yes
- More patients without exercise-induced angina than those with.
- Converted to a factor data type before any data processing.
- No missing entries, so no imputation or dropping of rows required.

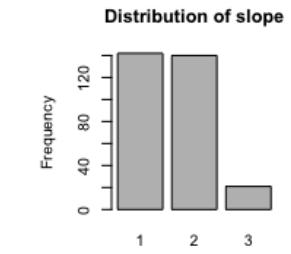
3.2.10 ST Depression Induced by Exercise (oldpeak) (numerical)





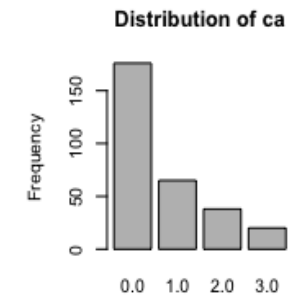
- ST depression refers to the position of the ST segment in a person's ECG results.
- A depressed or elevated ST segment can mean a person has an underlying health condition.
- Higher values of ST depression (the four outliers) likely indicate a more severe degree of myocardial ischemia induced by exercise.
- Right skew and outliers observed, but no removal or transformation done as it may distort the perceived distribution of the variable.

3.2.11 Slope of the Peak Exercise ST Segment (slope) (categorical)



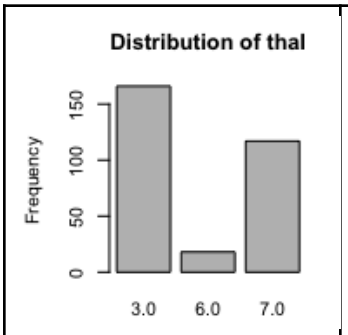
- Categorical variable with slope levels: 1 - Upwards, 2 - Flat, 3 - Downwards
- Most patients show an upwards or flat ST slope, with only a minority having a downwards slope.
- Converted to a factor data type before any data processing.
- No missing entries, so no imputation or dropping of rows required.

3.2.12 Number of major vessels colored by fluoroscopy (ca) (categorical)



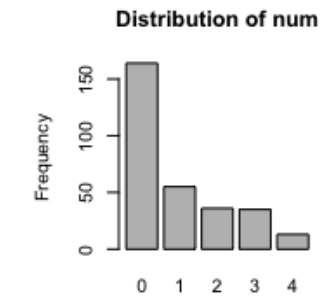
- Categorical variable with possible values 0, 1, 2, 3
- A higher value likely indicates more significant blockage or narrowing in those vessels.
- The majority have zero major vessels colored, meaning the majority had no major blood vessels showing significant narrowing or blockage during fluoroscopy.
- There is a decreasing frequency with an increasing number of vessels, which suggests that having a greater number of blockages in multiple major vessels is less common.
- Converted to a factor data type before any data processing.
- Convert "?" values to NA before dropping rows with NA (2 rows).

3.2.13 Thal (categorical)



- Categorical variable with 3 levels:
    - 3 – Normal, 6 – Fixed Defect, 7 – Reversible Defect
- 6: Fixed Defect, meaning an area of the heart muscle consistently shows reduced blood flow, both at rest and during stress. This could indicate a prior heart attack or significant scarring.
- 7: Reversible Defect, meaning an area of the heart muscle shows reduced blood flow during exercise (stress) but improves or appears normal at rest. This is a strong indicator of ischemia (temporary lack of blood flow).
- Majority showed normal blood flow to the heart muscle during the stress test (thal = 3.0)
- The second largest is exercise-induced ischemia, where blood flow is reduced during stress but improves at rest, which is a key indicator of coronary artery disease (thal = 7.0).
- Converted to a factor data type before any data processing.
- Convert "?" values to NA before dropping rows with NA (2 rows).

3.2.14 Diagnosis of Heart Disease (num) (categorical)



- Categorical variable with levels 0 to 4, 0 representing no heart disease, and increasing numbers representing increasing severity of heart disease.
- 54.1% have no heart disease (164 out of 303 observations with num = 0)
- There is a decreasing trend of the number of people with increasing severity of heart disease.
- Converted to a factor data type before any data processing.
- No missing entries, so no imputation or dropping of rows required.

3.3 Final Dataset for Analysis

Variables were appropriately encoded as factors or numeric as necessary. Imputation was done using the median for outliers in numerical variables. Rows with NA in categorical variables (ca and thal) were dropped using na.omit, as it comprised just ~2% of the data (6 out of 303). After rows with NA values were removed, we are left with 14 variables and 297 rows.

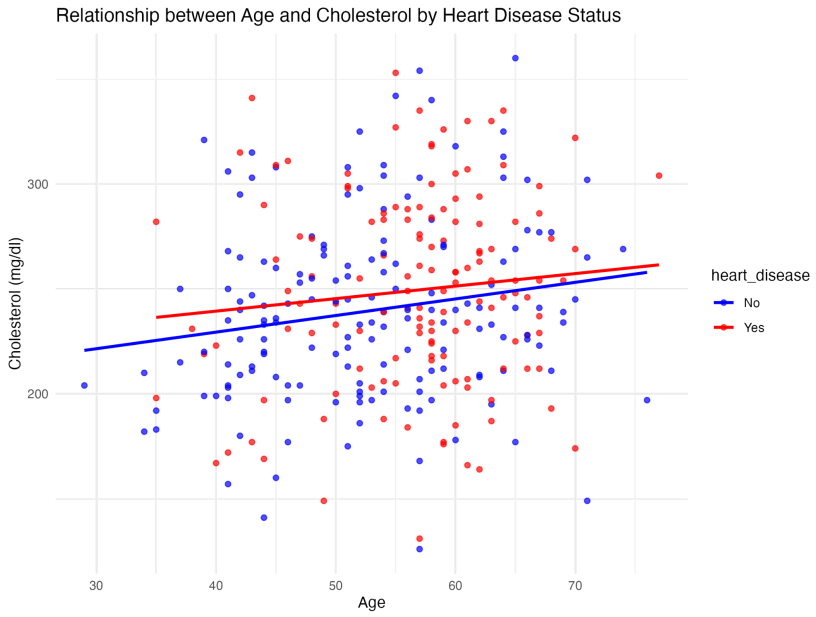
4. Statistical Analysis

With the dataset prepared, we can now conduct statistical testing to answer each of the questions previously outlined. For all our statistical tests, we tested combinations of different variables against the Diagnosis of Heart Disease attribute (num). We also converted the "num" data, from a categorical variable with levels 0 to 4 to a binary variable with values 0 and 1. "0" represents absence of heart disease, while "1" is a combination of levels 1, 2, 3 and 4, and represents that heart disease is present in individuals. The significance level for all tests will be standardised at $\alpha = 0.05$.

4.1 Statistical Tests

4.1.1 Relation between Heart Disease with Age and Cholesterol

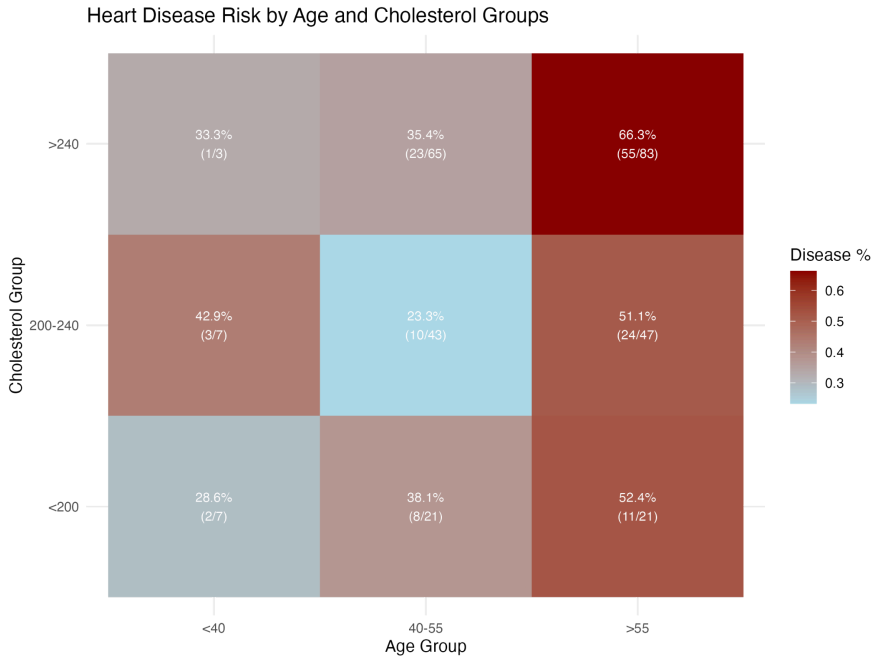
In this section we try to answer the question, "Is the risk of heart disease of an individual dependent on age and cholesterol?" Both age and cholesterol levels are commonly recognised risk factors for heart disease. Understanding their joint influence is crucial for effective risk assessment and prevention strategies.



[FIGURE 3: Relationship between Age and Cholesterol by Heart Disease Status]

The scatter plot of age against cholesterol levels in Figure 3 reveals an interesting pattern when points are coloured by heart disease status. Patients with heart disease (shown in red) tend to cluster in the upper-right region, showing that the combination of higher age and elevated cholesterol is linked with an increase in disease risk. The trend lines for each group demonstrate a

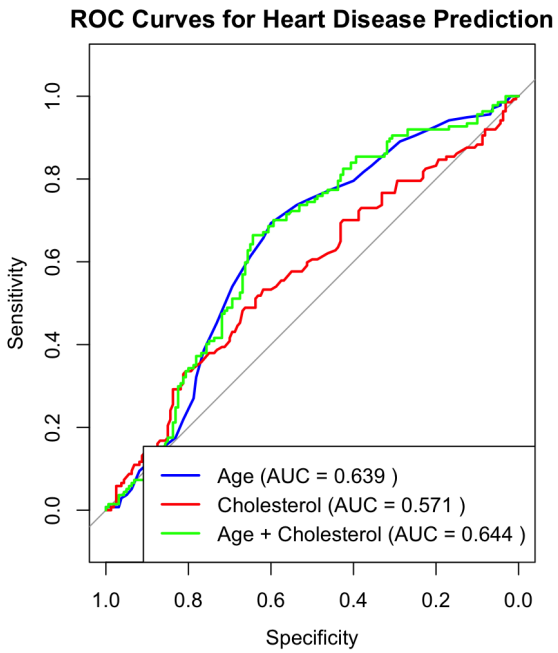
positive correlation between age and cholesterol, with the heart disease group maintaining consistently higher cholesterol levels across all age ranges.



[FIGURE 4: Heart Disease Risk by Age and Cholesterol Groups]

To better understand how these factors relate to heart disease risk, we created a heatmap visualising the disease prevalence across different age and cholesterol categories (Figure 4). The heatmap clearly demonstrates that the prevalence of heart disease increases progressively with both age and cholesterol levels. The highest risk group (age >55 with cholesterol >240 mg/dl) shows a 66.3% prevalence of heart disease, compared to just 28.6% in the lowest risk group (age <40 with cholesterol <200 mg/dl).

To determine the independent contribution of each factor to heart disease risk, we employed logistic regression analysis. Both age and cholesterol demonstrated statistically significant associations with heart disease ($p < 0.001$ for both). For each additional year of age, the odds of having heart disease increase by approximately 5.7% (95% CI: 3.1% to 8.3%), while each 10 mg/dl increase in cholesterol raises the odds by approximately 9.2% (95% CI: 5.3% to 13.2%).



[FIGURE 5: ROC curves for Heart Disease Prediction]

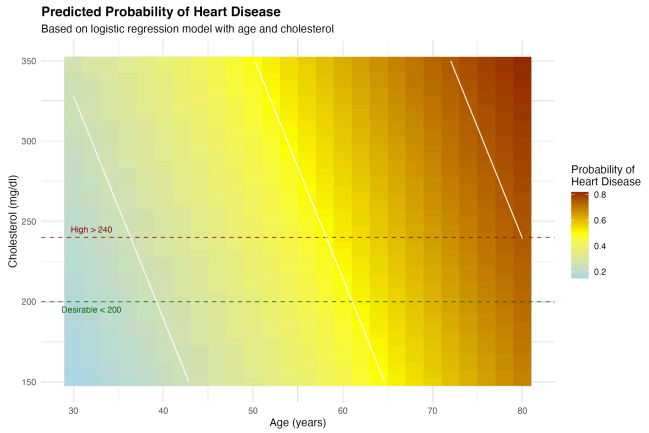
To compare the predictive capabilities of different models, we used Receiver Operating Characteristic (ROC) curve analysis.

- **Null Hypothesis (H_0)**: The model has no discriminative ability (AUC = 0.5).
- **Alternative Hypothesis (H_1)**: The model has discriminative ability (AUC > 0.5).

The ROC analysis (Figure 5) yielded the following areas under the curve (AUC):

- Age alone: AUC = 0.639
- Cholesterol alone: AUC = 0.571
- Age + Cholesterol: AUC = 0.644

All models showed AUC values significantly greater than 0.5 ($p < 0.001$), allowing us to reject the null hypothesis and conclude that all models have discriminative ability. These results indicate that age is a stronger predictor of heart disease than cholesterol in our dataset. While the combined model does show an improvement over using individual factors, the enhancement in predictive power is little. This suggests that age plays a more significant role in heart disease risk assessment, although cholesterol still contributes additional predictive information.



[FIGURE 6: Predicted Probability of Heart Disease by Age and Cholesterol]

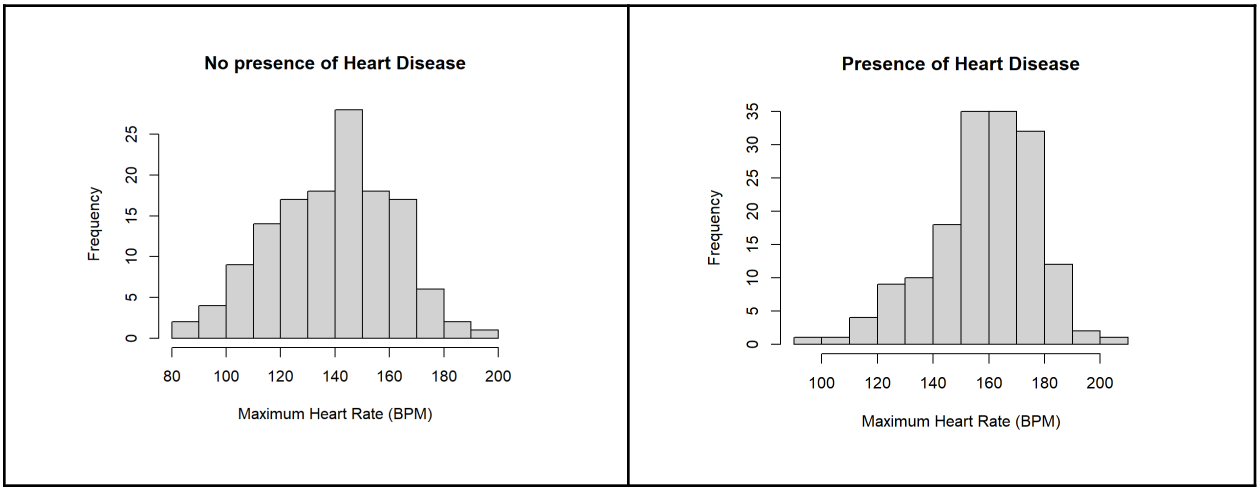
In order to further visualise the combined effects of age and cholesterol on heart disease risk, we created a probability heatmap (Figure 6) based on our logistic regression model. This visualisation demonstrates how the estimated probability of heart disease changes across different combinations of age and cholesterol levels. The colour gradient transitions from blue (low risk) through yellow (moderate risk) to dark red (high risk), with white contour lines marking key probability thresholds. Clinical reference lines for desirable (<200 mg/dl) and high (>240 mg/dl) cholesterol levels are included also. The visualisation demonstrates a clear gradient of increasing risk as both factors rise, confirming the additive effect of these two risk factors. This pattern reinforces our finding that the combination of advanced age and high cholesterol presents the highest risk scenario and highlights the importance of cholesterol management, particularly in older age groups.

Overall, we conclude that the risk of heart disease depends significantly on both age and cholesterol levels, with age appearing to be the more influential factor. However, the combined effect of both factors is particularly important, as demonstrated by the high disease prevalence in individuals with both advanced age and elevated cholesterol. These findings highlight the value of cholesterol management across all age groups, but especially in older individuals, where the cumulative risk is highest.

4.1.2 Relation between Heart Disease and Maximum Heart Rate

In this section, we try to answer the question, “Can the risk of heart disease be effectively detected based on the maximum heart rate achieved during exercise?”

Maximum Heart Rate is a numerical data that measures the Maximum Heart Rate of individuals achieved while exercising in Beats per Minute (BPM). Since this can be easily measured, it could be a simple way to test if an individual is at risk of heart disease.



A visual inspection of boxplot of the Maximum Heart Rate of each group shows that those with heart disease appear to have a higher Maximum Heart Rate than those without heart disease, indicating that it could be a predictor of whether or not an individual is at risk of heart disease. Since we want to predict a categorical variable (presence of heart disease, a binary variable) from an independent variable (Maximum Heart Rate), we make use of binary logistic regression. Logistic regression works similarly to linear regression, but for predicting categorical variables in general instead.

We then did Hypothesis Testing with definitions as follows:

- **Null hypothesis H_0**: Maximum Heart Rate has no effect in predicting Heart Disease (Regression Coefficient, r = 0)
- **Alternative hypothesis H_1**: Maximum Heart Rate has an effect in predicting Heart Disease (Regression Coefficient, r ≠ 0)

```
Call:
glm(formula = `Presence of Heart Disease` ~ `Maximum Heart Rate`,
    family = binomial, data = testdata)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          6.472142   1.002574   6.456 1.08e-10 ***
`Maximum Heart Rate` -0.044312   0.006627  -6.687 2.28e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 409.95  on 296  degrees of freedom
Residual deviance: 351.97  on 295  degrees of freedom
AIC: 355.97

Number of Fisher Scoring iterations: 4
```

After performing binary logistic regression, we get a p-value of 2.28e-11, which is significantly lower than the significance level of 0.05. Hence, we reject the null hypothesis and conclude that the Maximum Heart Rate achieved during exercise can be used to predict an individual's risk of having heart disease. Since the regression coefficient is negative (r = -0.044312), it shows that as the Maximum Heart Rate achievable by individuals increases, there may actually be a lower risk of individuals developing heart disease. This is in line with studies, where we know that being able to reach a higher maximum heart rate during exercise can indicate good cardiovascular fitness, and having greater cardiovascular fitness can decrease the risk of heart disease (Lang et al., 2024) (Radford et al., 2018).

4.1.3 Relation between Heart Disease and Fasting Blood Sugar

The Fasting Blood Sugar (FBS) in the dataset is binary (categorical) data that indicates whether a person has FBS above 120 mg/dl or not. An FBS above 120 mg/dl is generally considered higher than normal (Normal: <100 mg/dL; Indicator of prediabetes: 100-125 mg/dL; Indicator of diabetes: >=126 mg/dL) indicating the person has prediabetes and a high risk of Type 2 diabetes (Mayo Clinic., 2023). Recent studies have identified a significant link between diabetes and multiple aspects of cardiovascular health, such as coronary artery disease, heart failure, and stroke (Suman et al., 2023). Thus, we are interested in using our dataset to check if level of FBS is an indicator of heart disease.

A two-way contingency table between FBS and heart disease would then be as follows:

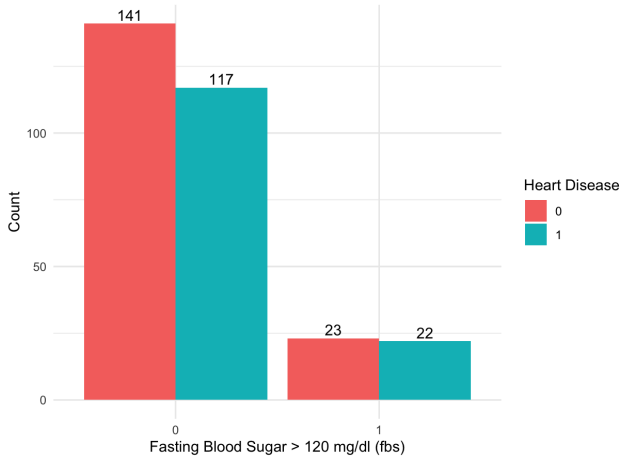
	0	1
0	141	117
1	23	22

Contingency Table of fbs vs Heart Disease

	0	1
0	0.55	0.45
1	0.51	0.49

Proportion table of Heart Disease within each fbs group

Next, we plot a bar graph of Heart Disease against FBS to better visualise the data:



[FIGURE 7: Heart Disease Count by Fasting Blood Sugar]

From the contingency table, proportion table and bar chart showing the distribution of heart disease status across FBS, we can tell the counts of patients with and without heart disease appear similar within each category, suggesting no strong association between elevated fasting blood sugar and heart disease in this dataset.

We then conducted a Chi-Square Test of Independence to examine the relationship between fbs and the presence of heart disease, first setting our hypothesis as follows:

- **Null Hypothesis (H_0)**: There is no relationship between fbs and Heart Disease.
- **Alternative Hypothesis (H_1)**: There is a relationship between fbs and Heart Disease.

```
[1] "Chi-Square Test Results:"
> print(chi_result)

        Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_table
X-squared = 0.077095, df = 1, p-value = 0.7813
```

As seen above, the test produced a p-value of 0.7813. Since the p-value is greater than the level of significance, 0.05, we fail to reject the null hypothesis. Hence, while elevated blood sugar may be a risk factor in other studies, this specific dataset does not provide strong evidence to support a significant relationship between fasting blood sugar (greater than 120 mg/dl) and heart disease diagnosis.

Next, we check the expected frequency table to see the number of observations we would anticipate if the variables were independent. To validate the use of the Chi-Square test and ensure the sampling distribution closely follows the Chi-Square distribution, we would expect all values to be at least 5.

```
          0         1
0 139.64356 118.35644
1  24.35644  20.64356
```

All expected frequencies exceeded 5, meeting the assumptions required to perform the Chi-Square Test of Independence as mentioned in the lecture.

For further evaluation, we carried out Cramér's V for an effect size measurement for the chi-square test of independence. Values closer to 1 indicate a stronger relationship, while values closer to 0 indicate a weaker relationship.

```
> cramers_v <- CramerV(contingency_table)
> print(paste("Cramer's V:", round(cramers_v, 3)))
[1] "Cramer's V: 0.025"
```

From our results, Cramér's V = 0.025, which indicates a very weak association between FBS and heart disease presence. Hence, higher fasting blood sugar does not appear to be an indicator of heart disease in this dataset, which would be contrary to most studies done previously. However, these results should be interpreted with caution as it is possible that other risk factors (e.g., age, cholesterol, chest pain type) play a more prominent role or that the dataset is not large and varied enough to detect a subtle effect of blood sugar.

4.1.4 Relation between Heart Disease and Different Types of Chest Pain

We now attempt to answer the question, "Can different types of chest pain be an indicator of heart disease?"

cp is a categorical variable while num is a binary variable. A two-way contingency table can be constructed between cp and num:

	num0	num1	num2	num3	num4
cp1	16	5	1	0	1
cp2	40	6	1	2	0
cp3	65	9	4	4	1
cp4	39	34	29	29	11

Contingency table where num has 5 factors

	num0 (Absent)	num1 (Present)
cp1	16	7
cp2	40	9
cp3	65	18
cp4	39	103

Contingency table with num as a binary variable

The tables indicate that patients with chest pain level 4 have a higher likelihood of heart disease compared to other levels of chest pain. Further statistical testing can confirm this prediction.

We use a Chi-square test to check if a statistically significant relationship between cp and num exists:

```
> chisq.test(cp_num_binary)
      Pearson's Chi-squared test
data:  cp_num_binary
X-squared = 77.276, df = 3, p-value < 2.2e-16
```

With an extremely low p-value (approximately 0), we reject the null hypothesis and conclude that such a relationship is indeed present.

An ANOVA test can be used to further analyze the data.
We did Hypothesis Testing with definitions as follows:

- **Null hypothesis H_0**: All levels of chest pain have statistically similar distributions of heart disease
- **Alternative hypothesis H_1**: There is a significant difference between at least 2 levels of chest pain

```
> aov_result = aov(num_binary ~ cp, data=heart_data1)
> summary(aov_result2)
             Df Sum Sq Mean Sq F value Pr(>F)
cp            3   19.2   6.401   34.35 <2e-16 ***
Residuals   293   54.6   0.186
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Rejecting the null hypothesis again, we have confirmed that differing distributions of heart disease exist between the values of cp. We finally use a pairwise comparison to establish which groups are significantly different:

```
> pairwise.t.test(heart_data1$num_binary, heart_data1$cp, p.adjust.method =
"none")
      Pairwise comparisons using t tests with pooled SD
data:  heart_data1$num_binary and heart_data1$cp
  1       2       3
2 0.27    -       -
3 0.39    0.67    -
4 2.0e-05 4.8e-13 8.2e-16
P value adjustment method: none
```

Notable differences in num are identified between the pairs 1-4, 2-4, and 3-4. All other pairs have p-values above 0.05. From the results of this test, we can conclude that chest pain level 4 (asymptomatic chest pain) is a significant indicator of heart disease in patients, while levels 1 through 3 are statistically similar and therefore much weaker predictors.

4.2 Multiple Regression

Previously, we investigated relationships between the presence of heart disease and a few key attributes. We will now build a regression model using all attributes given in the dataset and attempt to determine a combination of features that can most effectively predict heart disease.

We are using the binary variable of num in this model, so logistic regression is an ideal choice to predict the presence of heart disease. Additionally, a backwards stepwise approach will be taken, where variables that are not statistically significant will be iteratively removed with each step.

```
> summary(step_model)
Call:
glm(formula = num_binary ~ sex + cp + trestbps + chol + exang +
    oldpeak + slope + ca + thal, data = heart_data_reg)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.6310925  0.2237596  -2.820 0.005138 **
sex1         0.1441472  0.0473602   3.044 0.002559 **
cp2          0.0972941  0.0893308   1.089 0.277024
cp3          0.0319380  0.0826258   0.387 0.699391
cp4          0.2888124  0.0806982   3.579 0.000406 ***
trestbps     0.0020223  0.0013466   1.502 0.134285
chol         0.0008026  0.0004536   1.769 0.077930 .
exang1       0.0913776  0.0499479   1.829 0.068390 .
oldpeak      0.0567509  0.0241473   2.350 0.019455 *
slope2       0.1504732  0.0490495   3.068 0.002367 **
slope3       0.1140223  0.0881978   1.293 0.197142
ca1.0        0.2683866  0.0510646   5.256 2.92e-07 ***
ca2.0        0.3005453  0.0659553   4.557 7.75e-06 ***
ca3.0        0.3139264  0.0834021   3.764 0.000204 ***
thal6.0      0.0909427  0.0921068   0.987 0.324315
thal7.0      0.2210692  0.0498217   4.437 1.31e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The results from this regression model indicate that the strongest indicators of heart disease are sex=1 (male), cp=4, high values of oldpeak, slope=2 (flat), ca>0, and thal=7 (reversible defect).

5. Conclusion and Discussion

From our investigation, we conclude that:

- Age is a greater predictor of heart disease than cholesterol levels.
- As the maximum heart rate attainable increases, there is lower risk of developing heart disease.

- Higher fasting blood sugar does not seem to be an indicator of heart disease in this dataset.
    - As mentioned prior, this is contrary to most studies done previously. Hence, these results should be interpreted with caution. It is possible that other risk factors (e.g., age, cholesterol, chest pain type) play a more prominent role or that the dataset is not large and varied enough to detect a subtle effect of blood sugar.
- Chest pain level 4 (asymptomatic chest pain) is a significant indicator of heart disease in patients, as compared to levels 1, 2 and 3.
- Our multiple-regression model suggests that variables like oldpeak, slope, ca, and thal are potential indicators that warrant further analysis in future studies.

Thus, integrating such statistical models when making clinical decisions could enhance diagnostic efficiency and help healthcare systems to prioritise high-risk patients. Additionally, cross-validation with actual data from hospitals can also help to confirm the practical utility and robustness of the indicators in diverse populations.

However, this study does have some limitations: for one, the data was only sourced from a singular city in the US, which may not be an accurate representation of the world population. A much larger sample size of patients from different countries would improve the generalisation of the dataset to represent the world population. Additionally, various other factors that may be indicators of heart disease were not included in the dataset and thus may have been overlooked. This includes environmental factors, lifestyle habits or genetic predispositions (Heianza & Qi, 2019), (Münzel et al., 2022).

6. Appendix

Listing of code from R.

```
# Install and load necessary libraries
# Install packages if not already installed
if (!require(tidyverse)) install.packages("tidyverse")
if (!require(ggplot2)) install.packages("ggplot2")
if (!require(caret)) install.packages("caret")
if (!require(VIM)) install.packages("VIM")
if (!require(rpart)) install.packages("rpart")
if (!require(rpart.plot)) install.packages("rpart.plot")
if (!require(pROC)) install.packages("pROC")
if (!require(knitr)) install.packages("knitr")
if (!require(DescTools)) install.packages("DescTools")

# Load necessary libraries
library(tidyverse)
library(ggplot2)
library(caret)
library(VIM)
library(rpart)
library(rpart.plot)
library(pROC)
library(knitr)
library(DescTools)


# ------------------------------------------------------------
# Load dataset
heart_data <- read.table('/Users/dana/Downloads/NTU/Y2S2/MH3511
DataAnalysis/Project/heart+disease/processed.cleveland.data', sep=",")
# Note: You may need to adjust the file path to match your location

dim(heart_data)


# Rename headers for clarity
colnames(heart_data) <- c("age", "sex", "cp", "trestbps",
                          "chol", "fbs", "restecg", "thalach", "exang", "oldpeak",
                          "slope", "ca", "thal", "num")

head(heart_data)
```

```
## Check for variables with missing data and outliers
# Although the dataset had told us which variables had missing data (ca and thal), we check again.
# To handle the missing value we will check the columns of the datasets,
# if we found some missing data inside the columns then this generates the NA values as an
output, which can be not good for every model.

# ------------------------------------------------------------
## Encoding Categorical Variables
# Variables like cp (chest pain), thal, slope, sex, fbs, exang need to be factorised or one-hot
encoded first before data processing.

# Convert categorical variables to factors
categorical_vars <- c("sex", "cp", "fbs", "restecg", "exang", "slope", "ca", "thal", "num")
heart_data[categorical_vars] <- lapply(heart_data[categorical_vars], as.factor)
summary(heart_data)

# -   After converting the categorical variables to factor datatype, we can see that the missing
values are stored as "?" instead of "na", so we cannot use the usual is.na function
# -   Thus, we replace "?" values with "NA" so that we can use is.na
# -   From the results, we confirm that variables "ca" and "thal" have missing values

# Verify the number of "?" in each column (specifically "ca" and "thal")
missing_qmarks <- sapply(heart_data, function(x) sum(x == "?"))
print("Count of '?' in each column:")
print(missing_qmarks)

# Replace "?" with NA in the entire data frame
heart_data[heart_data == "?"] <- NA

# Drop unused factor levels
heart_data$ca <- droplevels(heart_data$ca)
heart_data$thal <- droplevels(heart_data$thal)

# Check for NA in each column after replacement
missing_nas <- sapply(heart_data, function(x) sum(is.na(x)))
print("Count of NA in each column:")
print(missing_nas)

# Now, summary() should not show "?"
summary(heart_data)

# ------------------------------------------------------------

## Method to check for distribution (in categorical variables)
par(mfrow = c(3, 3))
```

```
for (var in categorical_vars) {
  # Calculate the frequency of each category
  category_counts <- table(heart_data[[var]])

  # Create a bar chart
  barplot(category_counts,
          main = paste("Distribution of", var),
          xlab = var,
          ylab = "Frequency")
}

## Method to check for outliers (in numerical variables)
# Boxplots for key numerical variables
numeric_vars <- c("age", "trestbps", "chol", "thalach", "oldpeak")

par(mfrow = c(2, 3))
for (var in numeric_vars) {
  boxplot(heart_data[[var]], main = paste("Boxplot of", var), col = "lightblue")
}

# ------------------------------------------------------------

# Now, we imput NA values and outliers with the median for numerical variables
#
# * For thalach,
# * low-end values (~70 bpm) were initially considered as possible outliers, but further medical context suggests such values are plausible in patients with severe heart disease, so we exclude it when imputing outliers with the median.
# * For oldpeak,
# * Higher values of ST depression (the four outliers) likely indicate a more severe degree of myocardial ischemia induced by exercise.
# * Hence, we don't remove outliers, as we may be discarding potentially crucial information about individuals with a stronger ischemic response to exercise.

# Making a copy for data cleaning as in the RMD file
heart_data1 <- heart_data

# Function to impute with median for numerical variables
# Following the RMD approach, treating outliers for selected variables only
impute_with_median <- function(x) {
  Q1 <- quantile(x, 0.25, na.rm = TRUE)
  Q3 <- quantile(x, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  outliers <- x < (Q1 - 1.5 * IQR) | x > (Q3 + 1.5 * IQR)
  x[outliers | is.na(x)] <- median(x, na.rm = TRUE)
  return(x)
```

```
}

# As per the RMD file, only apply outlier treatment to these variables
# Excluding thalach and oldpeak as mentioned in the RMD
numeric_vars_for_imputation <- c("age", "trestbps", "chol")

# Apply imputation on selected numerical variables
heart_data1[numeric_vars_for_imputation] <- lapply(heart_data1[numeric_vars_for_imputation],
impute_with_median)

## Histogram (in numerical variables)
# Now, we check for skewed data

# Boxplots for key numerical variables
numeric_vars <- c("age", "trestbps", "chol", "thalach", "oldpeak")

par(mfrow = c(2, 3))
for (var in numeric_vars) {
  hist(heart_data1[[var]], main = paste("Histogram of", var), col = "lightblue")
}

# ------------------------------------------------------------

## Methods to handle missing values
# Dropping Rows (Listwise Deletion):
#
#   -   When to use: If the number of missing values is small and you believe the missingness is
random (MCAR), you can remove rows with NA values.
# -   Note: This can lead to significant data loss if missingness is widespread, but currently only 6
rows out of 303 are affected
#
#
# Imputation (Replacing Missing Values):
#
#   -   When to use: If dropping data would result in significant information loss, imputation is a
better option.
# -   Note: Imputation can introduce bias if not done carefully.
# -   Methods:
#   -   Mean/Median Imputation: Replace NA with the mean or median of the column. (for
numerical data)
# -   Mode Imputation: Replace NA with the most frequent category (mode). (for categorical data)
# -   Advanced Imputation: Use more sophisticated techniques like k-nearest neighbors (KNN)
imputation or model-based imputation.
# - We will not be doing this as it is too complicated for 6 missing values.

# As the number of rows affected by NA values are small, we drop rows affected by NA values
```

```
instead of imputing wiht the mode (possibility of introducing biases)


# Drop rows with NA values as done in the RMD file
# (instead of trying to impute categorical variables)
heart_data1 <- na.omit(heart_data1)

# Check the cleaned data
cat("Data preparation completed. Created heart_data1 with", nrow(heart_data1), "rows.\n")

#---------------------------------------------------------------
# END OF DATA CLEANING
#---------------------------------------------------------------

#---------------------------------------------------------------
# START OF PART 4.1.1: Is the risk of heart disease of an individual dependent on age and
cholesterol?
#---------------------------------------------------------------

#---------------------------------------------------------------
# Create binary heart disease variable for analysis
#---------------------------------------------------------------
# Convert num variable to binary (0 = No disease, 1-4 = Yes disease)
heart_df <- heart_data1 %>%
  mutate(heart_disease = ifelse(num == "0", "No", "Yes")) %>%
  mutate(heart_disease = factor(heart_disease, levels = c("No", "Yes")))

#---------------------------------------------------------------
# 1. Summary statistics for age and cholesterol by heart disease status
#---------------------------------------------------------------
age_chol_summary <- heart_df %>%
  group_by(heart_disease) %>%
  summarise(
    mean_age = mean(as.numeric(as.character(age))),
    median_age = median(as.numeric(as.character(age))),
    sd_age = sd(as.numeric(as.character(age))),
    min_age = min(as.numeric(as.character(age))),
    max_age = max(as.numeric(as.character(age))),
    mean_chol = mean(as.numeric(as.character(chol))),
    median_chol = median(as.numeric(as.character(chol))),
    sd_chol = sd(as.numeric(as.character(chol))),
    min_chol = min(as.numeric(as.character(chol))),
    max_chol = max(as.numeric(as.character(chol)))
  )

print(age_chol_summary)
```

```
#-------------------------------------------------------------
# 2. Visualize distributions - Exploratory Plots
#-------------------------------------------------------------
# Convert to numeric for plotting
heart_df$age <- as.numeric(as.character(heart_df$age))
heart_df$chol <- as.numeric(as.character(heart_df$chol))

# Age distribution by heart disease status
age_hist <- ggplot(heart_df, aes(x = age, fill = heart_disease)) +
  geom_histogram(position = "dodge", bins = 15, alpha = 0.7) +
  labs(title = "Age Distribution by Heart Disease Status",
       x = "Age",
       y = "Count") +
  theme_minimal()
print(age_hist)

# Cholesterol distribution by heart disease status
chol_hist <- ggplot(heart_df, aes(x = chol, fill = heart_disease)) +
  geom_histogram(position = "dodge", bins = 15, alpha = 0.7) +
  labs(title = "Cholesterol Distribution by Heart Disease Status",
       x = "Cholesterol (mg/dl)",
       y = "Count") +
  theme_minimal()
print(chol_hist)

#-------------------------------------------------------------
# 3. Box plots - These can be used for the report
#-------------------------------------------------------------

# FIGURE 1a (alternative): Box plot for Age by Heart Disease Status
age_box <- ggplot(heart_df, aes(x = heart_disease, y = age, fill = heart_disease)) +
  geom_boxplot() +
  labs(title = "Age by Heart Disease Status",
       x = "Heart Disease",
       y = "Age") +
  theme_minimal()
print(age_box)

# FIGURE 1b (alternative): Box plot for Cholesterol by Heart Disease Status
chol_box <- ggplot(heart_df, aes(x = heart_disease, y = chol, fill = heart_disease)) +
  geom_boxplot() +
  labs(title = "Cholesterol by Heart Disease Status",
       x = "Heart Disease",
       y = "Cholesterol (mg/dl)") +
  theme_minimal()
```

```
print(chol_box)

#---------------------------------------------------------------
# 4. Statistical tests
#---------------------------------------------------------------
# T-test for age between disease and no disease groups
age_ttest <- t.test(age ~ heart_disease, data = heart_df)
print(age_ttest)

# T-test for cholesterol between disease and no disease groups
chol_ttest <- t.test(chol ~ heart_disease, data = heart_df)
print(chol_ttest)

#---------------------------------------------------------------
# 5. FIGURE 1: Scatter plot for age vs cholesterol by heart disease status
#---------------------------------------------------------------
scatter_plot <- ggplot(heart_df, aes(x = age, y = chol, color = heart_disease)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE, aes(group = heart_disease)) +
  labs(title = "Relationship between Age and Cholesterol by Heart Disease Status",
       x = "Age",
       y = "Cholesterol (mg/dl)") +
  theme_minimal() +
  scale_color_manual(values = c("No" = "blue", "Yes" = "red"))
print(scatter_plot)

#---------------------------------------------------------------
# 6. Logistic Regression - Individual Predictors
#---------------------------------------------------------------
# Age as predictor
age_model <- glm(heart_disease ~ age, data = heart_df, family = "binomial")
summary(age_model)

# Cholesterol as predictor
chol_model <- glm(heart_disease ~ chol, data = heart_df, family = "binomial")
summary(chol_model)

# Interpreting the coefficients
age_odds_ratio <- exp(coef(age_model)["age"])
age_ci <- exp(confint(age_model)["age", ])
cat("Age odds ratio:", age_odds_ratio, "95% CI:", age_ci[1], "-", age_ci[2], "\n")

chol_odds_ratio <- exp(coef(chol_model)["chol"])
chol_ci <- exp(confint(chol_model)["chol", ])
cat("Cholesterol odds ratio (per 1 unit):", chol_odds_ratio, "95% CI:", chol_ci[1], "-", chol_ci[2], "\n")
cat("Cholesterol odds ratio (per 10 units):", chol_odds_ratio^10, "95% CI:", chol_ci[1]^10, "-",
```

```
chol_ci[2]^10, "\n")

#---------------------------------------------------------------
# 7. Logistic Regression - Combined Model
#---------------------------------------------------------------
# Age and cholesterol as predictors
age_chol_model <- glm(heart_disease ~ age + chol, data = heart_df, family = "binomial")
summary(age_chol_model)

# Combined model odds ratios
combined_odds <- exp(coef(age_chol_model))
combined_ci <- exp(confint(age_chol_model))
print(data.frame(
  Odds_Ratio = combined_odds,
  Lower_CI = combined_ci[, 1],
  Upper_CI = combined_ci[, 2]
))

#---------------------------------------------------------------
# 8. FIGURE 2: Create risk heatmap by age and cholesterol groups
#---------------------------------------------------------------
# Create age and cholesterol groups
heart_df <- heart_df %>%
  mutate(age_group = cut(age, breaks = c(0, 40, 55, 100),
                         labels = c("<40", "40-55", ">55")),
         chol_group = cut(chol, breaks = c(0, 200, 240, 600),
                          labels = c("<200", "200-240", ">240")))

# Create risk heatmap
risk_heatmap <- heart_df %>%
  group_by(age_group, chol_group) %>%
  summarise(
    total_count = n(),
    disease_count = sum(heart_disease == "Yes"),
    disease_proportion = mean(heart_disease == "Yes"),
    .groups = 'drop'
  )

# Display the risk by groups
print(risk_heatmap)

# Plot heatmap
heatmap_plot <- ggplot(risk_heatmap, aes(x = age_group, y = chol_group, fill = disease_proportion))
+
  geom_tile() +
  geom_text(aes(label = paste0(round(disease_proportion * 100, 1), "%\n(", disease_count, "/",
```

```
total_count, ")")),
          color = "white", size = 3) +
  scale_fill_gradient(low = "lightblue", high = "darkred") +
  labs(title = "Heart Disease Risk by Age and Cholesterol Groups",
       x = "Age Group",
       y = "Cholesterol Group",
       fill = "Disease %") +
  theme_minimal()
print(heatmap_plot)

#---------------------------------------------------------------
# 9. FIGURE 3: ROC Curves to evaluate predictive power
#---------------------------------------------------------------
# ROC curve for age model
age_probs <- predict(age_model, type = "response")
age_roc <- roc(heart_df$heart_disease, age_probs)
age_auc <- auc(age_roc)

# ROC curve for cholesterol model
chol_probs <- predict(chol_model, type = "response")
chol_roc <- roc(heart_df$heart_disease, chol_probs)
chol_auc <- auc(chol_roc)

# ROC curve for combined model
age_chol_probs <- predict(age_chol_model, type = "response")
age_chol_roc <- roc(heart_df$heart_disease, age_chol_probs)
age_chol_auc <- auc(age_chol_roc)

# Plot ROC curves
par(mfrow = c(1, 1))  # Reset plotting parameters
plot(age_roc, col = "blue", main = "ROC Curves for Heart Disease Prediction")
plot(chol_roc, col = "red", add = TRUE)
plot(age_chol_roc, col = "green", add = TRUE)
legend("bottomright", legend = c(paste("Age (AUC =", round(age_auc, 3), ")"),
                                 paste("Cholesterol (AUC =", round(chol_auc, 3), ")"),
                                 paste("Age + Cholesterol (AUC =", round(age_chol_auc, 3), ")")),
       col = c("blue", "red", "green"), lwd = 2)

#---------------------------------------------------------------
# 10. Compare models
#---------------------------------------------------------------
# Compare nested models
anova(age_model, age_chol_model, test = "Chisq")
anova(chol_model, age_chol_model, test = "Chisq")
```

```
#-------------------------------------------------------------
# 12. Standardized coefficients (to compare importance)
#-------------------------------------------------------------
# Scale the predictors
heart_df_scaled <- heart_df %>%
  mutate(age_scaled = scale(age),
         chol_scaled = scale(chol))

# Fit model with standardized predictors
model_scaled <- glm(heart_disease ~ age_scaled + chol_scaled,
                    data = heart_df_scaled, family = "binomial")
summary(model_scaled)


#-------------------------------------------------------------
# 13. FIGURE 4: Enhanced Predicted Probability Visualization
#-------------------------------------------------------------
# Create a more detailed grid for smoother visualization
# Use sequence instead of expand.grid to avoid duplicates
age_seq <- seq(30, 80, by = 2)
chol_seq <- seq(150, 350, by = 5)
detailed_grid <- expand.grid(age = age_seq, chol = chol_seq)

# Calculate predicted probabilities using our combined model
detailed_grid$predicted_prob <- predict(age_chol_model, newdata = detailed_grid, type =
"response")

# Create enhanced probability heatmap with simpler approach
enhanced_prob_plot <- ggplot(detailed_grid, aes(x = age, y = chol)) +
  geom_raster(aes(fill = predicted_prob)) +  # Use geom_raster instead of tile for better
performance
  scale_fill_gradient2(
    low = "lightblue",
    mid = "yellow",
    high = "darkred",
    midpoint = 0.5,
    name = "Probability of\nHeart Disease"
  ) +
  # Add clinical reference lines
  geom_hline(yintercept = 200, linetype = "dashed", color = "darkgreen", alpha = 0.7) +
  geom_hline(yintercept = 240, linetype = "dashed", color = "darkred", alpha = 0.7) +
  # Add annotations for cholesterol thresholds
  annotate("text", x = 32, y = 195, label = "Desirable < 200", color = "darkgreen", size = 3) +
  annotate("text", x = 32, y = 245, label = "High > 240", color = "darkred", size = 3) +
  # Add probability contours as separate overlays to avoid aesthetic conflicts
  stat_contour(aes(z = predicted_prob),
```

```
            breaks = c(0.25, 0.5, 0.75),
            color = "white",
            linewidth = 0.5) +
  # Improve labels and appearance
  labs(
    title = "Predicted Probability of Heart Disease",
    subtitle = "Based on logistic regression model with age and cholesterol",
    x = "Age (years)",
    y = "Cholesterol (mg/dl)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold"),
    legend.position = "right"
  )

print(enhanced_prob_plot)

# Save the figures for the report
# Save the scatter plot (Figure 1)
ggsave("figure1_age_chol_scatter.png", scatter_plot, width = 8, height = 6)

# Save the risk heatmap (Figure 2)
ggsave("figure2_risk_heatmap.png", heatmap_plot, width = 8, height = 6)

# Save the enhanced probability plot (Figure 5)
ggsave("figure5_enhanced_probability.png", enhanced_prob_plot, width = 9, height = 6)

# Save the decision tree (Figure 4)
# Note: For the ROC curve (Figure 3), you'll need to save it directly from the plot window
# or use a different approach as it's created with base R graphics, not ggplot2

#-------------------------------------------------------------
# END OF PART 4.1.1
#-------------------------------------------------------------




#-------------------------------------------------------------
# START OF PART 4.1.2: Can the risk of heart disease be detected based on the maximum heart
rate?
#-------------------------------------------------------------

myoverall=data.frame(heart_data1[8],heart_data1[14])
colnames(myoverall)=c("Maximum Heart Rate","Presence of Heart Disease")
```

```
boxplot(myoverall$`Maximum Heart Rate`,main="Maximum Heart Rate",ylab="Heart Rate (BPM)
",ylim=c(50,250))
hist(myoverall$`Maximum Heart Rate`,breaks=18,main="Maximum Heart Rate",xlab="Heart Rate
(BPM)")

testdata=myoverall
testdata$`Presence of Heart Disease`
length(testdata$`Presence of Heart Disease`)
for (i in 1:length(testdata$`Presence of Heart Disease`)){
  if( testdata$`Presence of Heart Disease`[i]!=0){
    testdata$`Presence of Heart Disease`[i]=1
  }
}

# testdata

par(mfrow=c(1,1))
# For Presence of Heart Disease = 0
zero1=subset(testdata,testdata$`Presence of Heart Disease`==0)
zero1=zero1$`Maximum Heart Rate`
hist(zero1,main="Presence of Heart Disease",xlab="Maximum Heart Rate (BPM)")


# For Presence of Heart Disease = 1
one1=subset(testdata,testdata$`Presence of Heart Disease`==1)
one1=one1$`Maximum Heart Rate`
hist(one1,main="No presence of Heart Disease",xlab="Maximum Heart Rate (BPM)")


# Fit binary logistic regression model
model <- glm(`Presence of Heart Disease` ~ `Maximum Heart Rate`, data = testdata, family =
binomial)

# View summary
summary(model)

#------------------------------------------------------------
# END OF PART 4.1.2
#------------------------------------------------------------




#------------------------------------------------------------
# START OF PART 4.1.3: Is a higher level of resting blood sugar an indicator of heart disease?
```

```
#------------------------------------------------------------

# Research Question: Is a higher level of resting blood sugar (fbs) an indicator of heart disease?

# Exploratory Data Analysis and Visualisation

# Convert heart disease to binary: 0 = no disease, 1 = presence of disease (1–4)
heart_data1$heart_disease <- ifelse(heart_data1$num == 0, 0, 1)
heart_data1$heart_disease <- as.factor(heart_data1$heart_disease)

# Check distribution of fbs and heart disease
table(heart_data1$fbs)
table(heart_data1$heart_disease)



# Frequency(contingency) table
fbs_hd_table <- table(heart_data1$fbs, heart_data1$heart_disease)
kable(fbs_hd_table, caption = "Contingency Table: Fasting Blood Sugar (fbs) vs Heart Disease")

# Proportion table
prop_table <- prop.table(fbs_hd_table, margin = 1)
kable(round(prop_table, 2), caption = "Proportion of Heart Disease Within Each fbs Group")



# Bar plot of heart disease by fbs level

ggplot(heart_data1, aes(x = fbs, fill = heart_disease)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Heart Disease Count by Fasting Blood Sugar",
    x = "Fasting Blood Sugar > 120 mg/dl (fbs)",
    y = "Count",
    fill = "Heart Disease"
  ) +
  geom_text(stat = "count", aes(label = ..count..), position = position_dodge(width = 0.9), vjust = -0.3)
+
  theme_minimal()

# Statistical Testing: Chi-Square Test of Independence

# Create a contingency table of fbs vs heart_disease
contingency_table <- table(heart_data1$fbs, heart_data1$heart_disease)
print("Contingency Table:")
print(contingency_table)
```

```
# Perform Chi-Square Test
chi_result <- chisq.test(contingency_table)
print("Chi-Square Test Results:")
print(chi_result)

# Check expected frequencies to validate assumptions
print("Expected Frequencies:")
print(chi_result$expected)

# Cramer's V for effect size measurement
cramers_v <- CramerV(contingency_table)
print(paste("Cramer's V:", round(cramers_v, 3)))

#------------------------------------------------------------
# END OF PART 4.1.3
#------------------------------------------------------------



#------------------------------------------------------------
# START OF PART 4.1.4: Can different types of chest pain types be an indicator of heart disease?
#------------------------------------------------------------

# Factoring num into binary
heart_data1$num_binary <- ifelse(heart_data1$num == 0, 0, 1)
heart_data1$num_binary

cp_num_binary <- table(heart_data1$cp, heart_data1$num_binary)
rownames(cp_num_binary) = c("cp1", "cp2", "cp3", "cp4")
colnames(cp_num_binary) = c("num0 (Absent)", "num1 (Present)")
cp_num_binary

chisq.test(cp_num_binary)

# Manual Chi-square (not in report)
colsum = matrix(colSums(cp_num_binary), ncol=2)
rowsum = matrix(rowSums(cp_num_binary), nrow=4)
expected = rowsum%*%colsum / sum(colsum)
expected
cp_num_m = matrix(cp_num_binary, ncol=2, byrow=TRUE)
chisq = sum((cp_num_m-expected)^2 / expected)
chisq
pvalue = 1-pchisq(chisq, df=12)
pvalue
```

```
# ANOVA and pairwise
aov_result = aov(num_binary ~ cp, data=heart_data1)
summary(aov_result)

pairwise.t.test(heart_data1$num_binary, heart_data1$cp, p.adjust.method = "none")

#---------------------------------------------------------------
# END OF PART 4.1.4:
#---------------------------------------------------------------



#---------------------------------------------------------------
# START  OF PART 4.2:
#---------------------------------------------------------------

# Regression

# Remove extraneous variables
heart_data_reg <- subset(heart_data1, select = -c(num, heart_disease))
# Create model
full_model = glm(num_binary ~ ., data=heart_data_reg)
step_model = step(full_model, direction = "backward")
summary(step_model)

#---------------------------------------------------------------
# END OF PART 4.2
#---------------------------------------------------------------
```

7. References

1. Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.
2. Lang, J. J., Prince, S. A., Merucci, K., Cadenas-Sanchez, C., Chaput, J., Fraser, B. J., Manyanga, T., McGrath, R., Ortega, F. B., Singh, B., & Tomkinson, G. R. (2024). Cardiorespiratory fitness is a strong and consistent predictor of morbidity and mortality among adults: an overview of meta-analyses representing over 20.9 million observations from 199 unique cohort studies. *British Journal of Sports Medicine*, 58(10), 556–566. https://doi.org/10.1136/bjsports-2023-107849
3. Mayo Clinic. (2023, August 17). Prediabetes: Diagnosis and treatment. https://www.mayoclinic.org/diseases-conditions/prediabetes/diagnosis-treatment/drc-20355284Radford, N. B., DeFina, L. F., Leonard, D., Barlow, C. E., Willis, B. L., Gibbons, L. W., Gilchrist, S. C., Khera, A., & Levine, B. D. (2018). Cardiorespiratory fitness, coronary artery calcium, and cardiovascular disease events in a cohort of generally healthy Middle-Age men. *Circulation*, 137(18), 1888–1895. https://doi.org/10.1161/circulationaha.117.032708
4. Suman, S., Biswas, A., Kohaf, N., Singh, C., Johns, R., Jakkula, P., & Hastings, N. (2023). The diabetes-heart disease connection: Recent discoveries and implications. Current Problems in Cardiology, 48(11), 101923. https://doi.org/10.1016/j.cpcardiol.2023.101923
5. World Health Organization. (2021, June 11). *Cardiovascular diseases (CVDs)*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
6. Heianza Y, Qi L. Impact of Genes and Environment on Obesity and Cardiovascular Disease. Endocrinology. 2019 Jan 1;160(1):81-100. https://doi.org/10.1210/en.2018-00591
7. Münzel T, Hahad O, Sørensen M, Lelieveld J, Duerr GD, Nieuwenhuijsen M, Daiber A. Environmental risk factors and cardiovascular diseases: a comprehensive expert review. Cardiovasc Res. 2022 Nov 10;118(14):2880-2902. https://doi.org/10.1093/cvr/cvab316