Data Science Project

Submitted by:

Nour Mohamed Eletreby 5200917

Haneen Ahmed Elgabry 5200199

Major:

Statistics

Submitted to:

Dr Sara Ossama

Due to:

1/1/2024

Introduction

Our dataset analyzes the presence of heart disease in the patient; this analysis has major contributions in this field, having a bunch of information about heart disease is really important. It's like having a big collection of details about different people's health and what kind of heart problems they might have. When scientists look closely at this information, they can find patterns and things that make people more likely to get heart issues. This helps doctors spot problems early and figure out better ways to treat them. It's like having a smart computer that learns from all this information to predict and personalize treatments for each person. So, having and studying this data is a big deal because it helps us understand and take care of our hearts better. Our data analysis will be done using R software.

Data Description

The source of our data is Kaggle " https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?select=heart.csv"

The target population is patients from Cleveland Foundation "for suspected coronary artery disease".

The dataset contains 1025 observations and 14 variables but we will analyze 9 of our interest.

We will state the variables in our study with a brief explanation (We are interested in analyzing these 9 variables from the study):

1- Age: continuous variable measured in "years" units.
2- Sex: categorical variable (1 = male; 0 = female)
3- Cp: categorical variable; "chest pain type" (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
4- Trestbps: continuous variable "resting blood pressure (on admission to the hospital)" measured in "mm Hg" units.
5- Chol: continuous variable "serum cholestoral" measured in "mg/dl" units
6- Thalach: continuous variable "maximum heart rate achieved"
7- Exang: categorical variable "exercise induced angina" (1 = yes; 0 = no)

8- Slope: categorical variable “the slope of the peak exercise ST segment” (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
9- Target(num) : categorical variable “diagnosis of heart disease (angiographic disease status)” (Value 0: < 50% diameter narrowing, Value 1: > 50% diameter narrowing)

Data Cleaning and Preparation

We removed the variables that we are not interested in our analysis now we will work on 9 variables (4 continuous and 5 categorical)

We checked with several codes on R we found no missing values in our dataset.

Then we checked for outliers by applying boxplot for our 4 continuous variables:

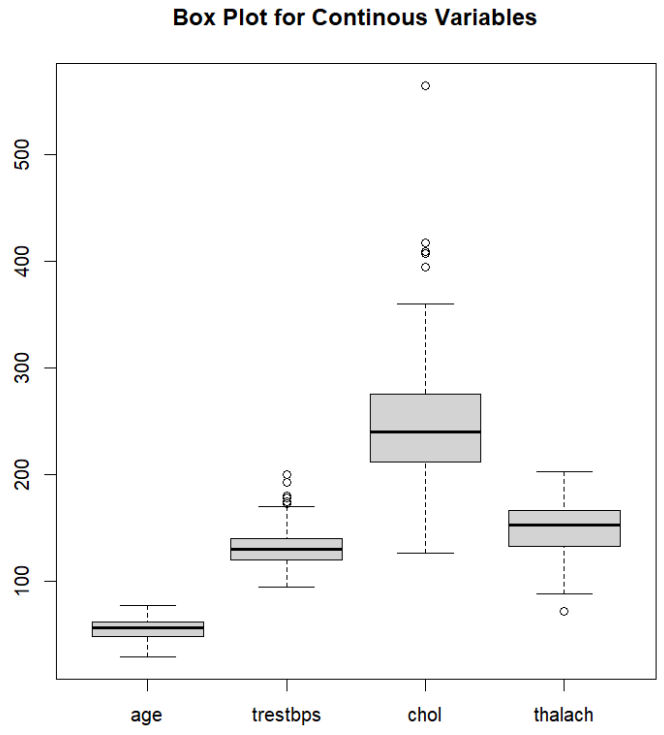


Figure (1): Box Plot for our continuous variables

Comment: we found that trestbps, chol and thalach have outliers.

We will deal with the outliers with the subset function; we identified the outliers by (Q1 – 1.5 * IQR or Q3 + 1.5 * IQR) then we used subset function. Hence, we got rid of the outliers in our dataset.

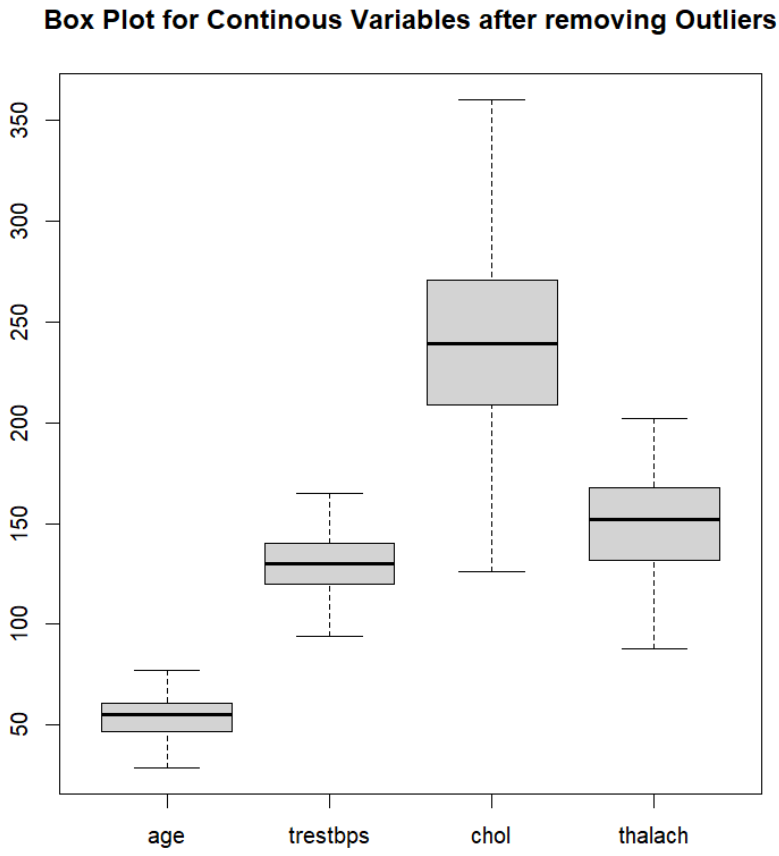


Figure (2): Box plot for continuous variables after removing outliers

Comment: after removing outliers by subset function we can see the new boxplot for the variables with no outliers

Now our dataset has no missing values and no outliers with 960 observations to start our analysis

Descriptive Analysis

First: Continuous Variables

4 variables: age, trestbps , chol and thalach

Descriptive measures:

Table (1): Descriptive measures for the continuous variables

Variable	Mean	Median	IQR	25% Quantile	75% Quantile	95% Quantile	Skewness	Kurtosis
Age	54.01	55	14	47	61	68	-0.18	-0.56
Trestbps	129.40	130	20	120	140	155	0.13	-0.40
Chol	241.78	239.00	62	209	271	319	0.22	-0.36
Thalach	149.27	152	36	132	168	182	-0.43	-0.47

Comment: -the age has mean equal to 54.01, median equal to 55, IQR equal to 14, 25% quantile equal to 47, 75% Quantile equal to 61, 95% Quantile equal to 68, skewness equal to -0.18 and kurtosis equal to -0.56

-The trestbps has mean equal to 129.40, median equal to 130, IQR equal to 20, 25% quantile equal to 120, 75% Quantile equal to 140, 95% Quantile equal to 155, skewness equal to 0.13 and kurtosis equal to -0.40

-The chol has mean equal to 241.78, median equal to 239, IQR equal to 62, 25% quantile equal to 209, 75% Quantile equal to 271, 95% Quantile equal to 319, skewness equal to 0.22 and kurtosis equal to -0.36

-The thalach has mean equal to 149.27, median equal to 152, IQR equal to 36, 25% quantile equal to 132, 75% Quantile equal to 271, 95% Quantile equal to 182, skewness equal to -0.43 and kurtosis equal to -0.47

Table (2): More descriptive measures for the continuous variables

Variable	Minimum	Maximum	Standard Deviation
Age	29	77	9.14
Trestbps	94	165	14.65
Chol	126	360	45.33
Thalach	88	202	22.89

Comment: -for the age the minimum value is equal to 29 while the maximum value is equal to 77 and the standard deviation is equal to 9.14

-For the trestbps the minimum value is equal to 94 while the maximum value is equal to 165 and the standard deviation is equal to 14.65

-For the chol the minimum value is equal to 126 while the maximum value 360 is equal to and the standard deviation is equal to 45.33

-For the thalach the minimum value is equal to 88 while the maximum value is equal to 202 and the standard deviation is equal to 22.89

Histograms:

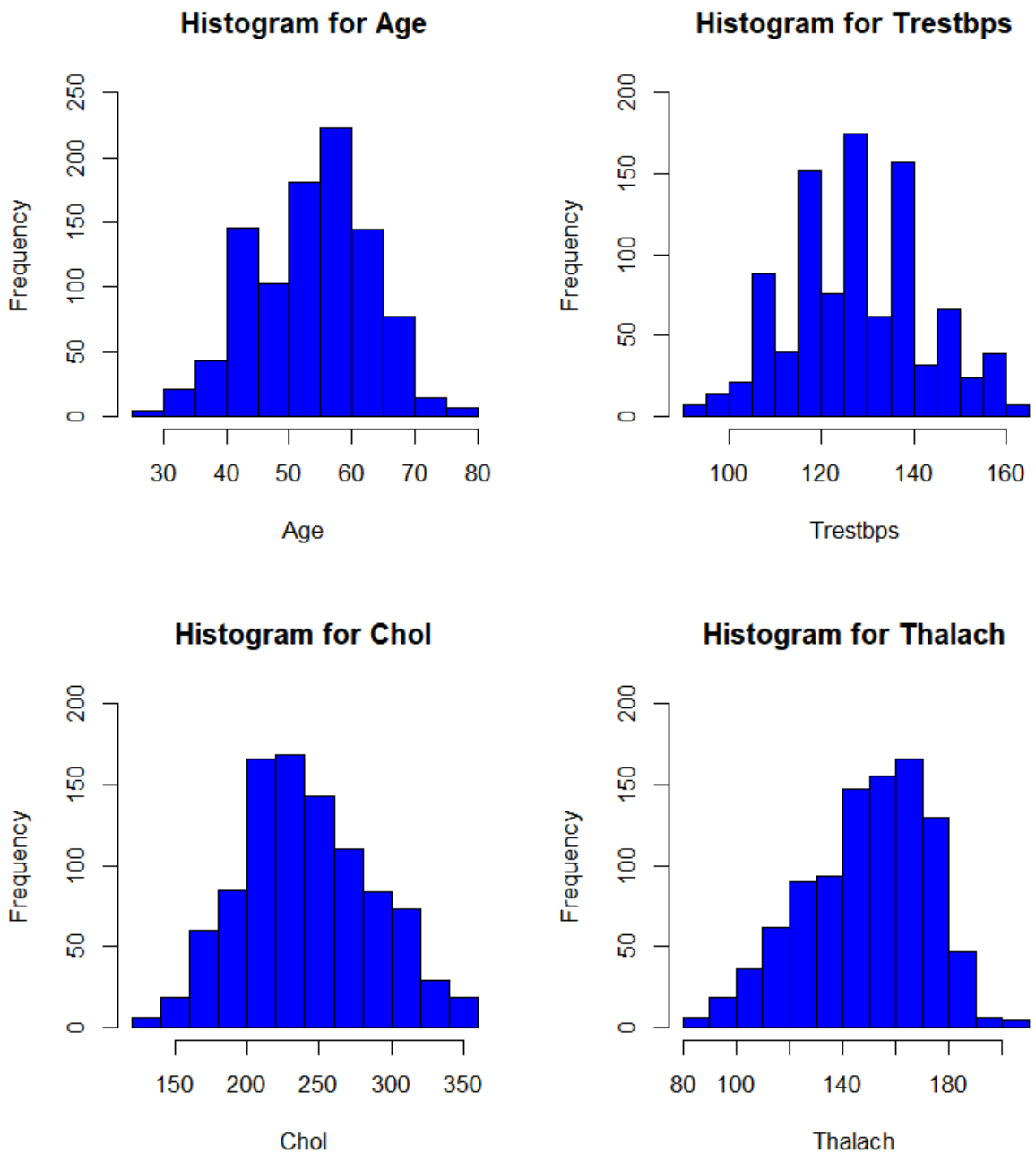


Figure (3): Histograms for Continuous variables

Comment: There are around 225 patients with age of 55 to 60 years which is the highest frequency

There are around 175 patients with resting blood pressure (trestbps) from 125 to 130 mm Hg which is the highest frequency

There are around 170 patients with serum cholestoral(chol) from 225 to 250 mg/dl which is the highest frequency.

There are around 170 patients with maximum heart rate(thalach) achieved from 160 to 170 units which is the highest frequency

Correlations between continuous variables:

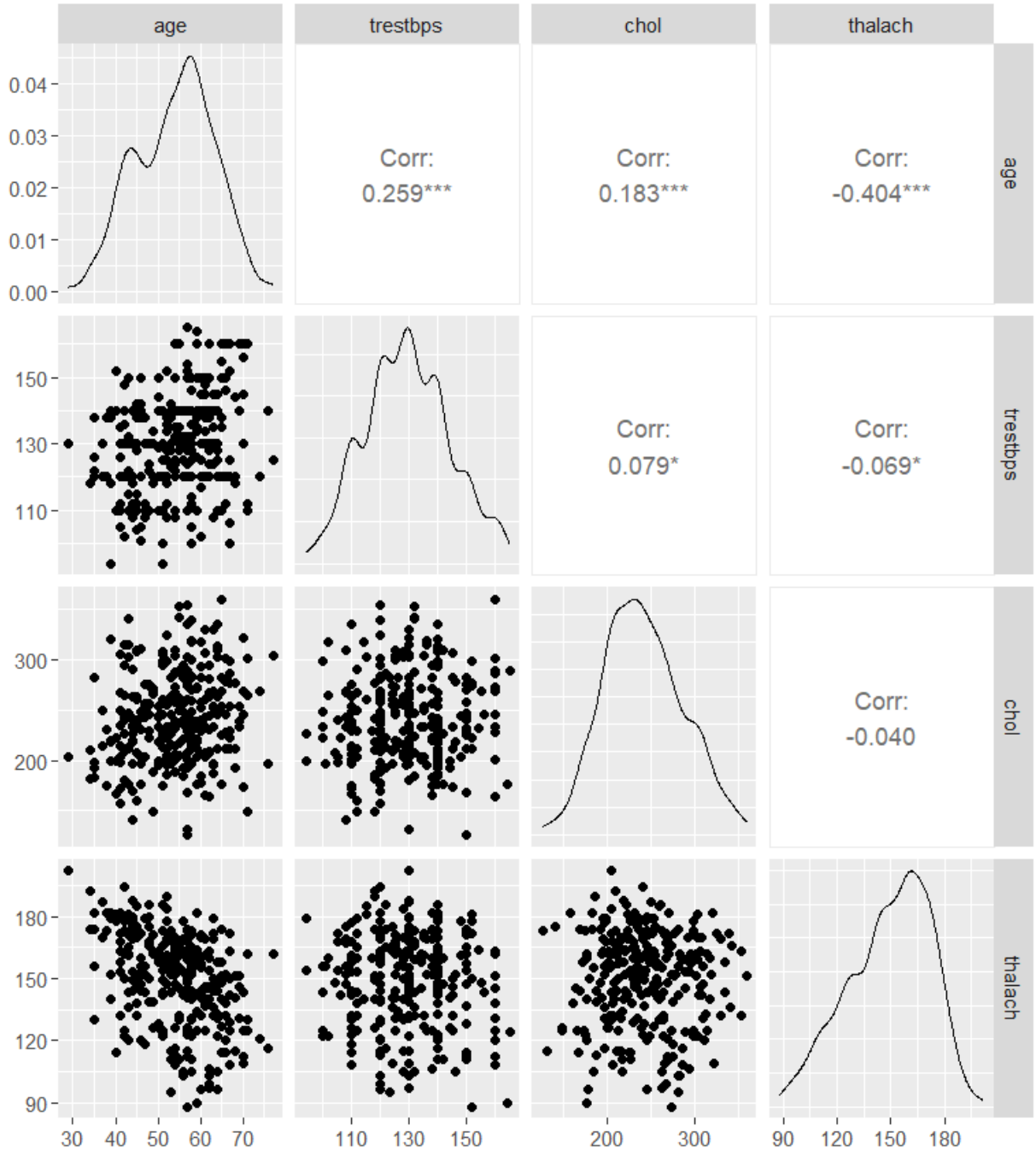


Figure (4): Correlations between continuous variables

Comment: -it is clear that the correlation between age and resting blood pressure (trestbps) is 0.259 which indicates a positive weak relationship between them

-it is clear that the correlation between age and serum cholestoral(chol) is 0.183 which indicates a positive weak relationship between them

-it is clear that the correlation between age and maximum heart rate(thalach) is -0.404 which indicates a negative moderate relationship between them

-it is clear that the correlation between resting blood pressure (trestbps) and serum cholestoral(chol) is 0.079 which indicates a positive weak relationship between them

-it is clear that the correlation between resting blood pressure (trestbps) and maximum heart rate(thalach) is -0.069 which indicates a negative weak relationship between them

-it is clear that the correlation between serum cholestoral(chol) and maximum heart rate(thalach) is -0.040 which indicates a negative weak relationship between them

Second: Qualitative Variables

5 variables: sex,cp,exang,slope,target

Frequency tables:

Sex

Table (3): Frequency table for Sex

Variables	Male	Female
Sex	686	274

It is clear that the higher frequency is the frequency of males 686 males and the frequency of females is 274 females

Cp

Table (4): Frequency table for Cp

Variable	Typical Angina	Atypical Angina	Non Anginal Pain	Asymptomatic
Cp	458	164	272	66

It is clear that the highest frequency goes to typical angina with 458 units and lowest frequency goes to asymptomatic with 66 units

Exang

Table (5): Frequency table for Exang

Variable	No	Yes
Exang	646	314

It is clear that the higher frequency with 646 units had no “exercise induced angina” while 314 units had “exercise induced angina”

Slope

Figure (6): Frequency table for Slope

Variable	Unsloping	Flat	Downsloping
Slope	64	439	457

It is clear that the highest frequency with 457 units are “Downsloping” while the least frequency with 64 units are “unsloping”

Target

Table (7): Frequency table for Target

Variable	< 50% diameter narrowing	> 50% diameter narrowing
Target	456	504

It is clear that the higher frequency with 504 units had “>50% diameter narrowing” and 456 units had “< 50% diameter narrowing”

Modes

Mode of Sex → Males

Mode of Cp → Typical Angina

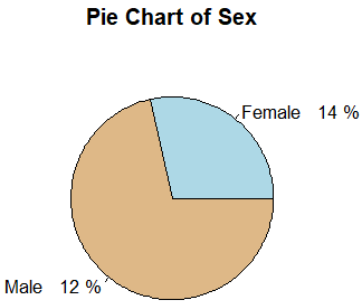
Mode of Exang → No

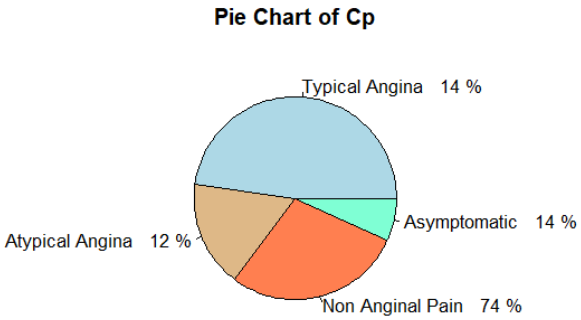
Mode of Slope → Downsloping

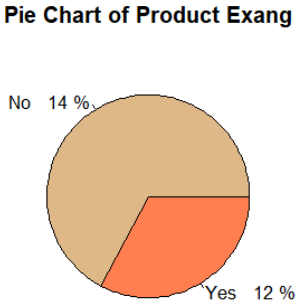
Mode of target → > 50% diameter narrowing

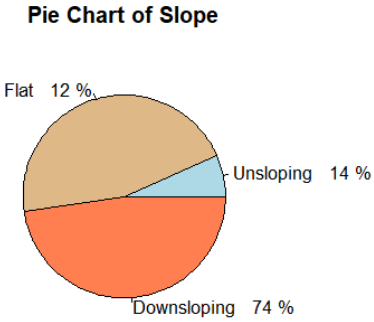
Visual presentations for frequencies of the variables:

Pie Charts:









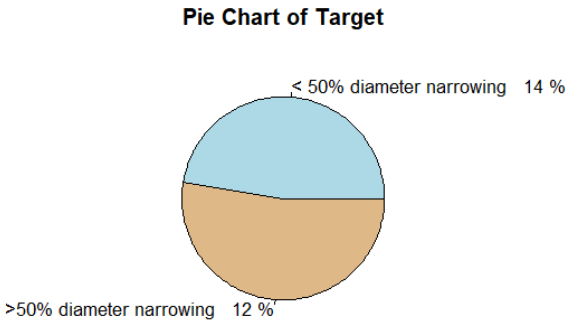


Figure (5): Pie charts for all categorical variables

Comment: these pie charts are visual presentations by percentages for the conducted frequency tables (from table 4 to table 7)

Histograms:

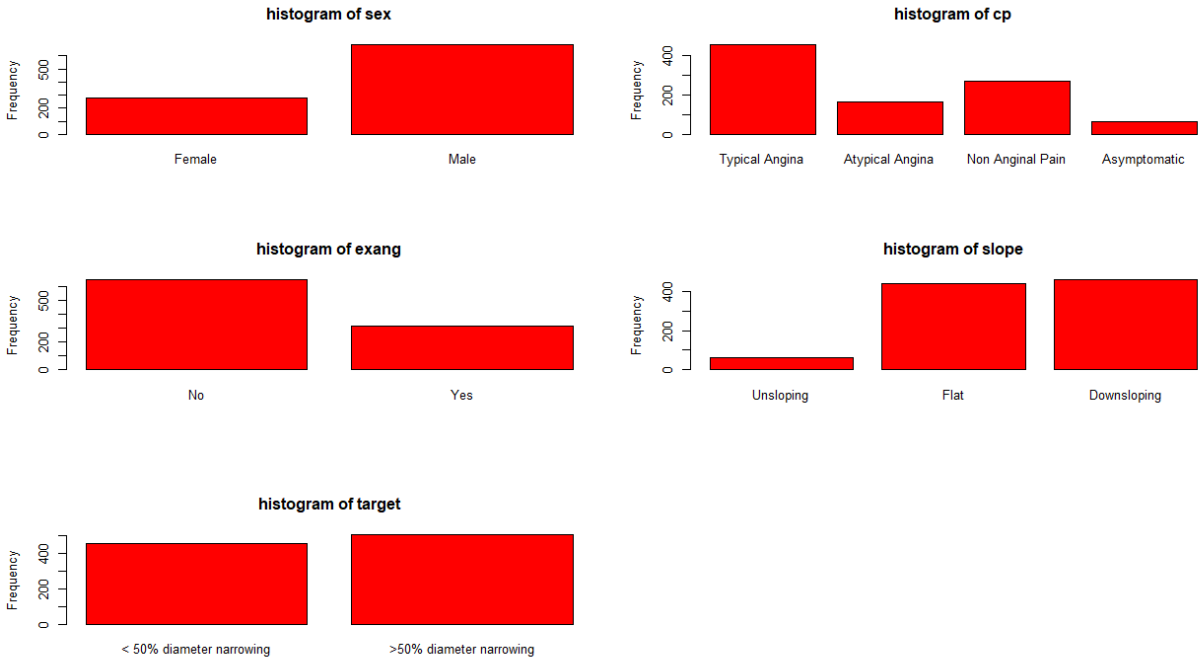


Figure 6: histogram for all categorical variables

Comment: the histograms are visual presentations for the frequencies (that can be seen in the tables from table 4 till 7) of the categorical variables.

Bivariate plots of sex variables vs all other categorical variables:

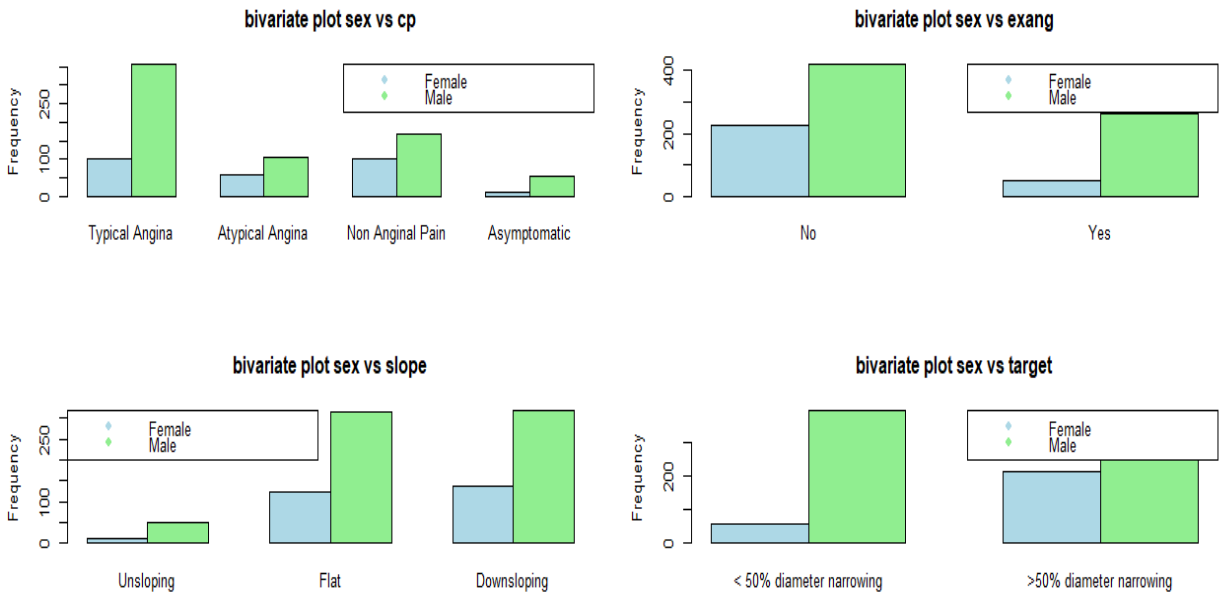


Figure 7: bivariate plot of age vs other var

1. **Chest Pain Types by Gender:**

-The first plot (top left) compares the frequency of different chest pain types (Typical Angina, Atypical Angina, Non-Anginal Pain, Asymptomatic) between males and females.

-Males have a higher frequency of Typical Angina, while females and males have approximately equal frequency in the other categories.

2. **Exercise-Induced Angina by Gender**:

-The second plot (top right) shows the frequency of exercise-induced angina (Yes/No) for each gender.

-we can see that males exhibit higher frequencies for both 'Yes' and 'No'.

3. **Slope of Peak Exercise ST Segment by Gender**:

-The third plot (bottom left) illustrates the frequencies associated with different slopes of the peak exercise ST segment (Upsloping, Flat, Downsloping).

- males are more frequent in all of the 3 categories.

4. **Heart Disease Presence by Gender**:

-The fourth plot (bottom right) displays frequencies related to heart disease presence based on >50% diameter narrowing as an indicator.

-Females have a high frequency approximately equal to males for "> 50% diameter narrowing", while males are more frequent in the "<50% diameter narrowing" category.

Contingency tables:

	Typical Angina	**Atypical Angina**	**Non-Anginal Pain**	**Asymptomatic**	**Total**
Female	101	57	103	13	274
Male	357	107	169	53	686
Total	458	164	272	66	960

Table 8: contingency table for age vs cp

This contingency table is about the health conditions by gender:

Comment: we can see that, Typical Angina: Females experience this type less frequently (101 cases) compared to males (357 cases). Atypical Angina: Again, females have fewer instances (57 cases) compared to males (107 cases). Non-Anginal Pain: Females exhibit lower frequency (103 cases) than males (169 cases). Asymptomatic: Females have a significantly lower occurrence (13 cases) compared to males (53 cases).

Target = < 50% diameter narrowing		**exang = no**	**exang = yes**	**Sum**
	slope = upsloping	15	24	39
	slope = flat	116	175	291
	slope = downsloping	79	47	126
	Sum	210	246	456
Target = > 50% diameter narrowing	**Slope** = upsloping	22	3	25
	slope = flat	120	28	148
	slope = downsloping	294	37	331
	Sum	436	68	504
	Slope = upsloping	37	27	64
Sum	**slope** = flat	236	203	439
	slope = downsloping	373	84	457
	Sum	646	314	960

Table 9: 3 way contingency table for target, slope , exang

This table represents the relationship between three categorical variables: Target**,** Exang**,** and Slope. Each cell contains the count of observations that fall into a specific combination of levels for these variables.

- The table provides information on the distribution of patients based on their target (diameter narrowing), exercise-induced angina (Exang), and the slope of the ST segment during exercise (Slope).
- For example, there are 15 patients with "< 50% diameter narrowing," no angina, and an upsloping ST segment.

pair plot of 4 continuous variables by region category:

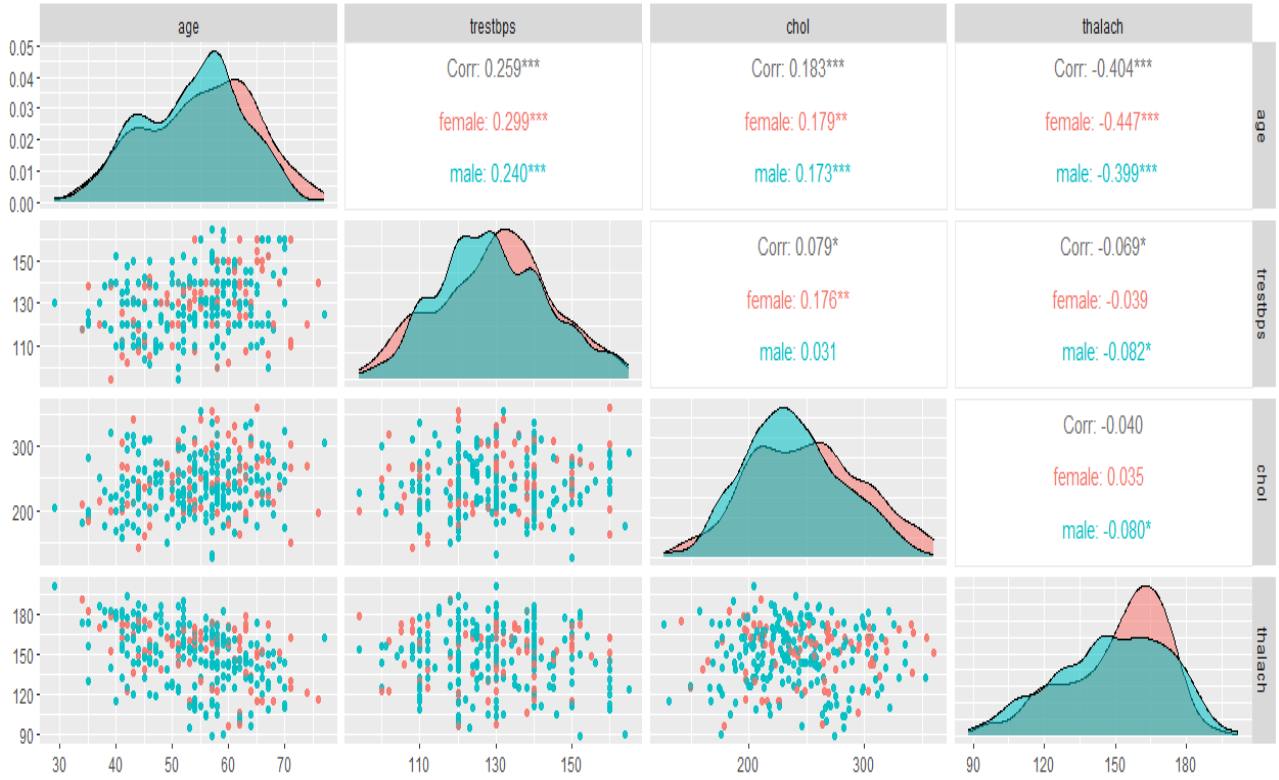


Figure 8: scatter plot matrix of 4 variables by sex

This is a scatter plot matrix of four variables: age, trestpbs, Chol, thalach. The data is grouped by sex (male, female), each represented by a different color.

Correlations differ somewhat between sex type of patients (colored correlations), with female patients generally having stronger relationships with the variables than males.

Regression/Logistic model

Logistic regression is a statistical method used for binary classification problems, where the outcome variable is categorical with two possible values. In the context of our project, a logistic regression model was employed to predict the presence or absence of heart disease (Target) based on features such as age (Age), sex (Sex), chest pain type (Cp), resting blood pressure (Trestbps), serum cholesterol (Chol), exercise-induced angina (Exang), and the slope of the peak exercise ST segment (Slope). The logistic regression model estimates the probability that an individual has heart disease.

The model:

The logistic regression model was fitted using the glm function in R, with the family parameter set to "binomial" to indicate the binomial distribution of the response variable. The model formula was specified as:

$$\text{logit}(P(\text{target}=1)) = 1.725283 - 0.030976 \cdot \text{age} - 2.158085 \cdot \text{sex} + 0.768983 \cdot \text{cp} - 0.017548 \cdot \text{trestbps} - 0.008479 \cdot \text{chol} - 0.911089 \cdot \text{exang} + 0.889897 \cdot \text{slope} + 0.028383 \cdot \text{thalach}$$

interpretation:

These coefficients of the model represent the estimated effects of each predictor on the log-odds of the target event occurring. For example, a one-unit increase in age is associated with a decrease of approximately 0.031 in the log-odds of having heart disease, while being male (sex = 1) is associated with a decrease of approximately 2.158 in the log-odds.

We also found out that all of the variables in the model are statistically significant (all of the p-values <0.05).

Statistical Assessment of the Model:

1. Likelihood Ratio Test (lrtest):

Null Hypothesis: The predictors (age, sex, cp, trestbps, chol, exang, slope, thalach) have no effect, i.e., the model reduces to intercept-only.

Alternative Hypothesis: The predictors in Model has a significant effect.

Interpretation: The p-value is very small ($< 2.2e-16$), suggesting strong evidence to reject the null hypothesis. Therefore, including the predictors significantly improves the model fit compared to an intercept-only model.

2. Hosmer-Lemeshow Goodness-of-Fit Test:

Null Hypothesis: The model fits the data well.

Alternative Hypothesis: The model does not fit the data well.

Interpretation: The p-value is 0.09578, which is greater than of 0.05. Therefore, do not have enough evidence to reject the null hypothesis. The model appears to fit the data well based on the Hosmer-Lemeshow test.

3. McFadden's Pseudo-R-squared Value:

Value of 0.39 indicates a moderately good fit, meaning that the model captures a significant portion of the variability in the target variable based on the predictors included in the model.

Checking multicollinearity assumption:

These are the vif's of the variables in the model:

```
  age      sex       cp     chol    exang    slope  thalach
1.213655 1.211122 1.114825 1.134302 1.099750 1.154821 1.324078
```

Since that they're all less than 10 then there's no multicollinearity in the model.

Backward and forward selection:

1. Backward selection:

The backward selection aims to build a regression model that includes only statistically significant predictors related to the response variable. In this case, it seems that including all the variables provides the best fit according to the AIC.

The model with no predictors removed has an AIC of 810.61, which is the lowest. Therefore, the final model includes all the predictor variables.

2. Forward selection

The stepwise selection process evaluated different combinations of predictors and selected the model that resulted in the lowest AIC value. Therefore, the selection_2 model represents the best-fitting model according to the stepwise selection criteria.

The selected model, selection_2, is the same as the original model with the same predictors.

By conducting these 2 selection methods we found out that our original model is a good fit.

New Data and Predicted Probabilities for Heart Disease Prediction

We sought to predict the probability of heart disease for two hypothetical individuals using our logistic regression model. The characteristics of these individuals are summarized in the new data table below:

	Age	Sex	Cp	Trestbps	Chol	Thalach	Exang	Slope
1	40	1	1	120	200	150	0	2
2	35	0	2	130	220	160	1	1

Observation 1: A 40-year-old male with typical angina, a resting blood pressure of 120 mm Hg, a serum cholesterol level of 200 mg/dl, a maximum heart rate of 150 bpm, no exercise-induced angina, and a peak exercise ST segment slope of 2.

Observation 2: A 35-year-old female with atypical angina, a resting blood pressure of 130 mm Hg, a serum cholesterol level of 220 mg/dl, a maximum heart rate of 160 bpm, experiencing exercise-induced angina, and a peak exercise ST segment slope of 1.

The logistic regression model was applied to these observations, resulting in the following predicted probabilities of heart disease:

1)For the first observation (40-year-old male), the model predicts an 79.12% probability of having heart disease.

2)For the second observation (35-year-old female), the model predicts a higher probability of 92.77% for heart disease.

Model Performance Evaluation:

the predicted_probs variable contains the predicted probabilities of having heart disease generated by the logistic regression model. The head(predicted_probs) shows the predicted probabilities for the first six observations in the dataset:

1: 0.62562103

2: 0.05928595

3: 0.01826779

4: 0.42769079

5: 0.22948808

6: 0.60436317

These probabilities represent the model's estimated likelihood of each observation having heart disease (a value of 1) based on the given predictors.

Comparing the predicted_probs to the data_no_outliers$target values:

1: 0

2: 0

3: 0

4: 0

5: 0

6: 1

We can see that the observations from 2nd to 5th have predicted probabilities below 0.5, suggesting a low likelihood of heart disease. These probabilities align with the corresponding target values of 0, indicating the absence of heart disease.

However, the first and sixth observations has a predicted probability greater than 0.5. This higher probability aligns with the target value of 1, indicating the presence of heart disease.

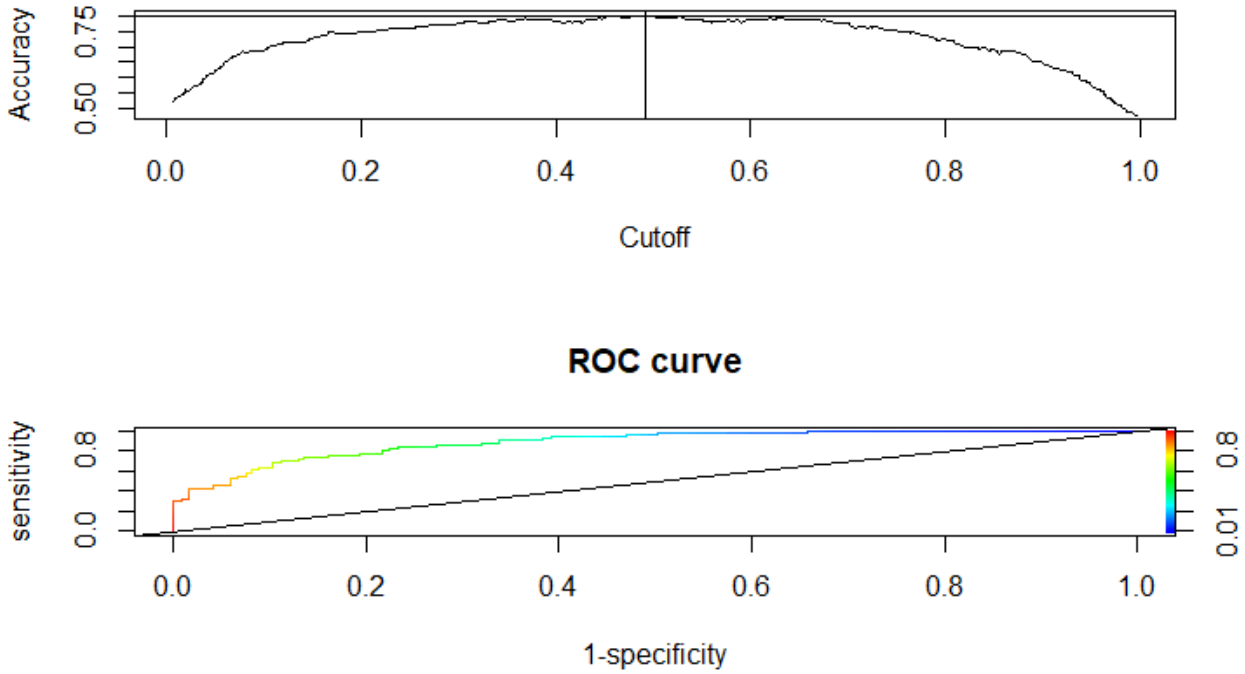


Figure 9

1. ROC Curve Plot:

The curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) for various threshold values.

The abline(0,1) adds a diagonal line, representing the ROC curve for a random classifier.

2. Accuracy and Cutoff Point:

The code calculates the accuracy and the corresponding cutoff point on the ROC curve.

The Accuracy value (approximately 80.3%) represents the accuracy of the model at the selected cutoff point.

The Cutoff value (approximately 0.491) indicates the threshold on the ROC curve where the accuracy is maximized.

Confusion Matrix and Statistics:

		reference	
Prediction		0	1
	0	350	83
	1	106	421

True Positives (TP): 421 - The number of instances correctly predicted as positive (individuals with heart disease).

True Negatives (TN): 350 - The number of instances correctly predicted as negative (individuals without heart disease).

False Positives (FP): 83 - The number of instances incorrectly predicted as positive.

False Negatives (FN): 106 - The number of instances incorrectly predicted as negative.

-we calculated the accuracy, precision and sensitivity of the model and found out that:

Accuracy: The overall accuracy of the model is 80.31%, indicating that 80.31% of instances are correctly classified.

Precision: The positive predictive value is 80.83%. This means that when the model predicts a positive outcome (heart disease), it is correct 80.83% of the time.

Recall (Sensitivity): The sensitivity of the model is 76.75%. This indicates that out of all actual positive instances (individuals with heart disease), the model correctly identifies 76.75% of them.

ML algorithms:

1)knn

The dataset was divided into training and testing sets, with a 70-30 split ratio. The training set was used for model training, and the testing set for evaluation.

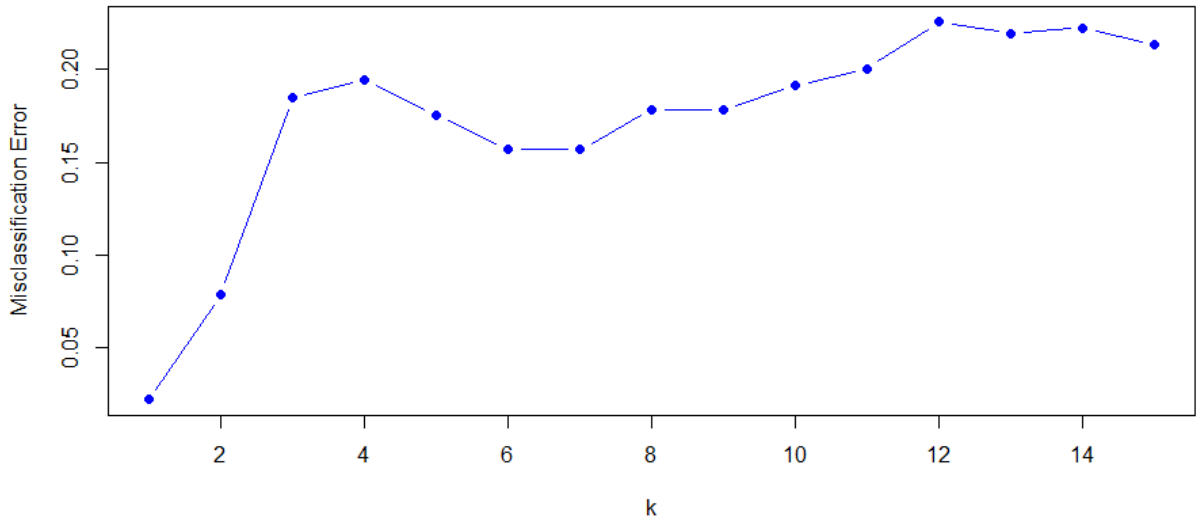
The predictor variables in both the training and testing sets were scaled to normalize their values.

KNN Model Training:

A k-Nearest Neighbors (KNN) model was trained using a loop that iterated over different values of k (from 1 to 15). For each k, the model was evaluated on the testing set, and the misclassification error was calculated.

Plotting Misclassification Error:

A plot was generated to visualize the relationship between k values and misclassification error. The x-axis represented different values of k, while the y-axis depicted the corresponding misclassification error.



The optimal value of k was determined based on the lowest misclassification error observed in the plot which is k=1 .

Confusion matrix:

This confusion matrix and associated statistics provide a comprehensive evaluation of the performance of a classification model.

	Reference		
		0	1
Prediction	0	148	3
	1	4	164

True Positives (TP): 164

True Negatives (TN): 148

False Positives (FP): 3

False Negatives (FN): 4

Accuracy and Confidence Interval:

Accuracy: 97.81% - This indicates the overall correctness of the model predictions.

95% Confidence Interval: (95.53%, 99.11%) - A range within which we are 95% confident that the true accuracy lies.

Other Metrics:

Sensitivity (True Positive Rate): 97.37% - The ability of the model to correctly identify positive instances.

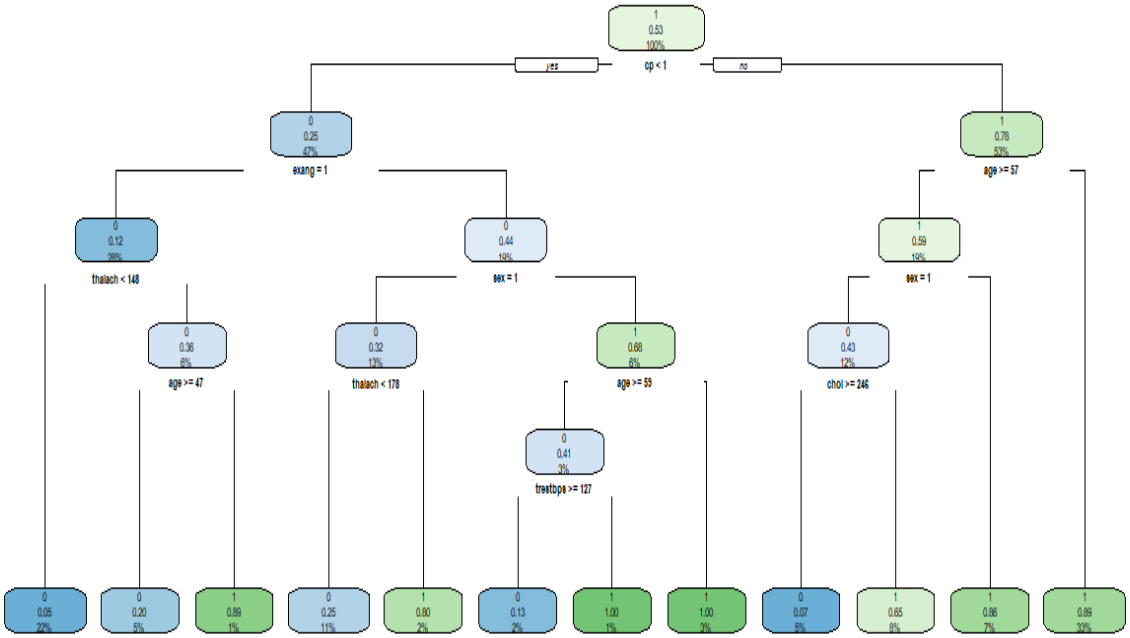
Specificity (True Negative Rate): 98.20% - The ability of the model to correctly identify negative instances.

Positive Predictive Value (Precision): 98.01% - The proportion of predicted positive instances that are correctly classified.

Negative Predictive Value: 97.62% - The proportion of predicted negative instances that are correctly classified.

Comment: The model shows high accuracy, sensitivity, and specificity, indicating its effectiveness in both positive and negative class predictions.

2)decision tree:



The top node of the tree uses the criterion “cp < 1”.

- If this condition is true, it leads to another criterion “exang=1” if true “thalach < 148”, which further breaks down based on “age >= 47”.
- If the condition is false, it leads to “sex=1” and it separate into two branches with criteria “thalach<178” and “age >= 59”, respectively.

Feature Importance: The feature importance, as indicated by the variable importance plot, is ranked in descending order based on the overall contribution of each variable to the decision tree model. The values represent the importance score.

thalach (Overall: 108.14): The highest-ranked variable, indicating a significant contribution to the model.

cp (Overall: 90.04): Another highly important variable.

exang (Overall: 81.76)

slope (Overall: 78.54)

sex (Overall: 69.05)

chol (Overall: 50.94)

age (Overall: 39.35)

trestbps (Overall: 28.68)

We can see from these score that the variables (thalach ,cp , exang , slope ,sex) contribute the most to the decision-making process of the model.

Confusion matrix of decision tree:

	Reference		
Target		0	1
	0	117	35
	1	20	147

-Accuracy: The overall accuracy of the model is 82.76%, indicating the proportion of correctly classified instances among all instances.

-Sensitivity (True Positive Rate): The model correctly identifies 85.40% of the instances belonging to the positive class.

-Specificity (True Negative Rate): The model correctly identifies 80.77% of the instances belonging to the negative class.

-Positive Predictive Value (Precision): Of the instances predicted as positive, 76.97% are correctly classified.

-Negative Predictive Value: Of the instances predicted as negative, 88.02% are correctly classified.

Comments:

The decision tree is prioritizing features such as 'thalach,' 'cp,' 'exang,' , ‘age’,and 'slope' as key contributors to predicting the target variable.

Overall, the model appears to have good performance, with notable accuracy, sensitivity, and specificity.

Creating our own function (for () and while ()):

We have created a function with for () loop called convert_to_factor to facilitate the conversion of specified columns to factor variables. The function takes two parameters, the data frame and a vector of column names. For each specified column, the function checks if the data type is integer, and if so, converts it to a factor.

And by that we applied this function to our dataset and converted the variables (sex, cp, target, slope and exange) into factors in order to be able to proceed with the while loop.

We have also implemented a function using while loop named calculate_continuous_means to compute the mean values for continuous variables in the dataset. The function iterates through each column in the data frame, identifies numeric columns, and calculates their respective means.

The output displays the mean values for continuous variables in the train_data dataset as follows:

Mean of age: 53.89173

Mean of trestbps: 129.4917

Mean of chol: 243.0647

Mean of thalach: 149.0075

Mean of heart_disease_prob: 0.5067669

Main results and conclusions:

Results:

- Descriptive analysis revealed patterns in variable distributions and relationships between features.

- Logistic regression model showed all predictors were statistically significant with $p<0.05$.

- Model had good fit based on Hosmer-Lemeshow and pseudo-R-squared tests.

- KNN and decision tree models achieved high accuracy (>80%).

- Decision tree prioritized thalach, cp, exang, age, and slope as important predictors.

Conclusions:

- Features like age, sex, trestbps, chol, thalach showed some correlation with heart disease risk.

- Logistic regression effectively predicted probability of disease presence based on predictors.

- KNN and decision tree classification models demonstrated good performance on this dataset.

- Variables thalach, cp, exang, and slope contributed most to decision tree predictions.

- Descriptive stats and visualizations provided insight into variable distributions and relationships.

- Our analysis confirms the influence of factors like age, sex, blood pressure, cholesterol on heart health.

- With further model optimization, these tools can aid doctors in personalized risk assessment.

- Larger and more diverse datasets may improve model generalizability for new populations.

In summary, through statistical learning we developed models to predict heart disease using key clinical and lifestyle factors. The results support the importance of these variables and show the opportunity for data-driven tools to support clinicians.

Classification of work:

We nearly made all of the project together, as we gathered together and made all of the codes together and then at each point we divided it on us to make sure that each one of us contributed in the project equally and that we understand each point.

Appendix

```
data<-read.csv("heart.csv",header=T)
data[,c(6,7,10,12,13)]<-NULL
View(data)
summary(data)
class(data)
str(data)
if (any(is.na(data))) {
  print("There are missing values in the data.")
} else {
  print("There are no missing values in the data.")
}
sum(!is.na(data))
sum(is.na(data))
dim(data) ##we checked by these codes and we found our data is clean with no missing values
##Box Plot for 4 variables: age, trestbps , chol , thalach
boxplot(data[c(1,4,5,6)],main="Box Plot for Continous Variables") ##Outliers in all except age
# Identify the outliers: (Q1 – 1.5 * IQR or Q3 + 1.5 * IQR)
out_w <- quantile(data$trestbps,c(0.25, 0.75), na.rm = T) + c(-1.5, 1.5) *
  IQR(data$trestbps, na.rm =T)
out_h <- quantile(data$chol,c(0.25, 0.75), na.rm = T) + c(-1.5, 1.5) *
  IQR(data$chol, na.rm =T)
out_t<-quantile(data$thalach,c(0.25, 0.75), na.rm = T) + c(-1.5, 1.5) *
  IQR(data$thalach, na.rm =T)
# Remove the outliers
data_no_outliers <- subset(data,
                           (data$trestbps > out_w[1] & data$trestbps < out_w[2])
                           &
```

```
                (data$chol > out_h[1] & data$chol< out_h[2])
               &
                (data$thalach > out_t[1] & data$thalach< out_t[2])

)
dim(data_no_outliers)
boxplot(data_no_outliers[c(1,4,5,6)],main="Box Plot for Continous Variables after removing Outliers")
summary(data_no_outliers)
##Analysis for Cont. age, trestbps , chol , thalach, oldpeak and ca
library(DescTools)
Desc(data_no_outliers$age)
Desc(data_no_outliers$trestbps)
Desc(data_no_outliers$chol)
Desc(data_no_outliers$thalach)
min(data_no_outliers$age) ; max(data_no_outliers$age)
min(data_no_outliers$trestbps) ; max(data_no_outliers$trestbps)
min(data_no_outliers$chol) ; max(data_no_outliers$chol)
min(data_no_outliers$thalach) ; max(data_no_outliers$thalach)
##Histograms
par(mfrow = c(2, 2))
hist(data_no_outliers$age,xlab ="Age",main=paste("Histogram for Age"),col="blue",ylim=range(0,250))
hist(data_no_outliers$trestbps,xlab="Trestbps",main=paste("Histogram for Trestbps"),col="blue",ylim=range(0,200))
hist(data_no_outliers$chol,xlab="Chol",main=paste("Histogram for Chol"),col="blue",ylim=range(0,200))
hist(data_no_outliers$thalach,xlab ="Thalach",main=paste("Histogram for Thalach"),col="blue",ylim = range(0,200))
##Correlations
```

```
library(dplyr)
library(GGally)
data_no_outliers %>%
  select(age, trestbps, chol,thalach) %>%
  ggpairs(alpha = 0.5)
##Qual.
##sex,cp,exang,slope,target
label1 <- c("Female","Male") #Sex
label2 <- c("Typical Angina","Atypical Angina","Non Anginal Pain","Asymptomatic") #cp
label3 <- c("No","Yes") #exang
label4 <- c("Unsloping","Flat","Downsloping") #slope
label5 <- c("< 50% diameter narrowing",">50% diameter narrowing") #Target
#Freq tables
table1<-table(data_no_outliers$sex)
table2<-table(data_no_outliers$cp)
table3<-table(data_no_outliers$exang)
table4<-table(data_no_outliers$slope)
table5<-table(data_no_outliers$target)
#plots for each variable
par(mfrow = c(2,2))
pct1 <- round(table1/sum(table1)*100)
lbls1 <- paste(label1, " ", pct, "%")
pie(table1,
    main = "Pie Chart of Sex",
    labels = lbls1,
    col = c("lightblue", "burlywood"))

pct2 <- round(table2/sum(table2)*100)
```

```
lbls2 <- paste(label2, " ", pct, "%")
pie(table2,
   main = "Pie Chart of Cp",
   labels = lbls2,
   col = c("lightblue", "burlywood", "coral", "aquamarine"))
pct3 <- round(table3/sum(table3)*100)
lbls3 <- paste(label3, " ", pct, "%")
pie(table3,
   main = "Pie Chart of Product Exang",
   labels = lbls3,
   col = c( "burlywood", "coral"))
pct4 <- round(table4/sum(table1)*100)
lbls4 <- paste(label4, " ", pct, "%")
pie(table4,
   main = "Pie Chart of Slope",
   labels = lbls4,
   col = c("lightblue", "burlywood", "coral"))

pct5<- round(table5/sum(table5)*100)
lbls5 <- paste(label5, " ", pct, "%")
pie(table5,
   main = "Pie Chart of Target",
   labels = lbls5,
   col = c("lightblue", "burlywood"))
##Histograms
par(mfrow = c(2,2))
freq_product <- table1
barplot(freq_product,
```

```
        names.arg =label1,
        ylab = "Frequency",
        main = "histogram of sex",
        col = "red",
        border = "black"
)

freq_product <- table2
barplot(freq_product,
        names.arg =label2,
        ylab = "Frequency",
        main = "histogram of cp",
        col = "red",
        border = "black"
)

freq_product <- table3
barplot(freq_product,
        names.arg =label3,
        ylab = "Frequency",
        main = "histogram of exang",
        col = "red",
        border = "black"
)

freq_product <- table4
barplot(freq_product,
        names.arg =label4,
```

```
        ylab = "Frequency",
        main = "histogram of slope",
        col = "red",
        border = "black"
)
freq_product <- table5
barplot(freq_product,
        names.arg =label5,
        ylab = "Frequency",
        main = "histogram of target",
        col = "red",
        border = "black"
)

##bivariate plots
par(mfrow=c(2,2))
bi_freq <- table(data_no_outliers$sex,data_no_outliers$cp)
p1 <-   barplot(bi_freq,
        beside = TRUE,
        names.arg = label2,
        col = c("lightblue","lightgreen"),
        ylab = "Frequency",
        main = "bivariate plot sex vs cp"
)
legend("topright", legend = label1, pch = 16, col = c("lightblue", "lightgreen"))

bi_freq <- table(data_no_outliers$sex,data_no_outliers$exang)
p2 <- barplot(bi_freq,
```

```
      beside = TRUE,
       names.arg = label3,
       col = c("lightblue","lightgreen"),
       ylab = "Frequency",
       main = "bivariate plot sex vs exang"
)
legend("topright", legend = label1, pch = 16, col = c("lightblue", "lightgreen"))

bi_freq <- table(data_no_outliers$sex,data_no_outliers$slope)
p3 <-  barplot(bi_freq,
       beside = TRUE,
       names.arg = label4,
       col = c("lightblue","lightgreen"),
       ylab = "Frequency",
       main = "bivariate plot sex vs slope"
)
legend("topleft", legend = label1, pch = 16, col = c("lightblue", "lightgreen"))

bi_freq <- table(data_no_outliers$sex,data_no_outliers$target)
p4 <-  barplot(bi_freq,
       beside = TRUE,
       names.arg = label5,
       col = c("lightblue","lightgreen"),
       ylab = "Frequency",
       main = "bivariate plot sex vs target"
)
legend("topright", legend = label1, pch = 16, col = c("lightblue", "lightgreen"))
```

```
##contengency tables

table1 <- xtabs(~ data_no_outliers$sex+data_no_outliers$cp) |> addmargins()

tab1 <- matrix(table1, nrow=3,ncol=5, dimnames = list(c(label1,"Total"),c(label2,"Total"))) #2-way

tab1


table2 <- xtabs(~ data_no_outliers$exang+data_no_outliers$slope+data_no_outliers$target) |> addmargins()

table2 # 3 -way




##pair plot of 4 cont variables by sex category:

library(dplyr)

library(ggplot2)

library(GGally)


data_no_outliers <- within(data_no_outliers, {

  sexcateg <- NA

  sexcateg[sex == 1] <- "male"

  sexcateg[sex == 0] <- "female"

})


data_no_outliers%>%

  select(age,trestbps,chol,thalach)%>%

  ggpairs(aes(color = data_no_outliers$sexcateg , alpha = 0.3))



##logistic model
```

```
library(car)
library(caret)
library(pROC)
library(plotROC)
library(ROCR)
library(pscl)
library(lmtest)
library(ResourceSelection)
# Create a logistic regression model
attach(data_no_outliers)
model <- glm(target ~ age + sex + cp + trestbps + chol + exang + slope +thalach, data = data_no_outliers, family = "binomial")
summary(model)
model
lrtest(model) #statistical assesment of the model
pscl::pR2(model)["McFadden"]
hoslem.test(data_no_outliers$target, fitted(model))
vif(model) #checking multicollienarity


#backward , forward model
selection_backward <- step(model, direction = "backward")
selection_forward <- step(model, scope = formula(~ age + sex + cp + trestbps + chol + exang + slope + thalach), direction = "forward")


# Creating a data frame for new observations
new_data <- data.frame(age = c(40, 35),
                       sex = c(1, 0),
                       cp = c(1, 2),
                       trestbps = c(120, 130),
```

```
                chol = c(200, 220),
                thalach = c(150, 160),
                exang = c(0, 1),
                slope = c(2, 1))

predictions <- predict(model, newdata = new_data, type = "response")
print(predictions)

#model performance evaluation #
predicted_probs <- predict(model, data_no_outliers,type = 'response')
summary(predicted_probs)
data_no_outliers$predicted_probs<-fitted(model)
predict<-predict(model,data_no_outliers)
summary(predict)
head(predicted_probs)
head(data_no_outliers$target)
#roc curve\
par(mfrow=c(2,2))
pred<-prediction(predicted_probs,data_no_outliers$target)
roc<-performance(pred,"acc")
plot(roc)
abline(h=0.9778523,v=0.5724582)
max<-which.max(slot(roc,'y.values')[[1]])
acc<-slot(roc,'y.values')[[1]][max]
cut<-slot(roc,'x.values')[[1]][max]
print(c(Accuracy = acc , Cutoff= cut))
roc_curve<-performance(pred,'tpr','fpr')
plot(roc_curve,colorize=T,xlab='1-specificity',ylab="sensitivity",main='ROC curve')
```

```
abline(0,1)
#conf matrix and statistics
predicted_labels <- ifelse(predicted_probs > 0.491406, 1, 0)  # Use the selected cutoff
conf_matrix <- confusionMatrix(data = factor(predicted_labels), reference =
factor(data_no_outliers$target))
print(conf_matrix)
accuracy <- conf_matrix$overall["Accuracy"]
precision <- conf_matrix$byClass["Pos Pred Value"]
recall <- conf_matrix$byClass["Sensitivity"]
cat("Accuracy:", accuracy, "\n")
cat("Precision:", precision, "\n")
cat("Recall (Sensitivity):", recall, "\n")

##ML ALGORITHMS
#KNN#
library(e1071)
library(caTools)
library(class)
# Split the data into training and testing sets
split <- sample.split(data_no_outliers, SplitRatio = 0.7)
train_set <- subset(data_no_outliers, split == "TRUE")
test_set <- subset(data_no_outliers, split == "FALSE")
# Scale the predictor variables
train_scale <- scale(train_set[, 1:8]) |> as.data.frame()
test_scale <- scale(test_set[, 1:8])       |> as.data.frame()
# Implement KNN with multiple values of k
misClassError <- c()
for (k in 1:15) {
  classifier_knn <- knn(train = train_scale,
```

```
                    test = test_scale,
                    cl = train_set$target,
                    k = k)
  misClassError[k] <- mean(classifier_knn != test_set$target)
}

# Plotting misclassification error against values of k
plot(1:15, misClassError, type = 'b', pch = 19, col = 'blue', xlab = 'k', ylab = 'Misclassification Error')

# Identify the optimal value of k based on the lowest misclassification error
optimal_k <- which.min(misClassError)
print(paste('Optimal k value:', optimal_k))

classifier_knn <- knn(train = train_scale,
                      test = test_scale,
                      cl = train_set$target,
                      k = 1)
classifier_knn

#confusion matrix
predicted_labels <- knn(train = train_scale, test = test_scale, cl = train_set$target, k = optimal_k)
conf_matrix <- confusionMatrix(data = factor(predicted_labels), reference = factor(test_set$target))
print(conf_matrix)

#decison tree#
library(caTools)
library(rpart)
```

```
library(rpart.plot)
library(caret)
library(dplyr)

model <- rpart(target ~ ., data = train_set, method = "class")
model

rpart.plot(model)

#Feature importance
importances <- varImp(model)
importances %>%
  arrange(desc(Overall))

#Making predictions
preds <- predict(model, newdata = test_set, type = "class")
preds
#Confusion matrix
actual_labels <- factor(test_set$target, levels = c("0", "1"))  # Adjust levels accordingly
predicted_labels <- factor(preds, levels = c("0", "1"))  # Adjust levels accordingly
min_length <- min(length(actual_labels), length(predicted_labels))
actual_labels <- actual_labels[1:min_length]
predicted_labels <- predicted_labels[1:min_length]

conf_matrix <- confusionMatrix(actual_labels, predicted_labels)

print(conf_matrix)
```

```
##creating functions
#Create a function with for to convert the specified categorical columns to factor variables.
convert_to_factor <- function(data, columns) {
  for (col in columns) {
    if (is.integer(data[[col]])) {
      data[[col]] <- as.factor(data[[col]])
    }
  }
  return(data)
}

train_data <- convert_to_factor(train_set, c("sex", "cp", "exang", "slope", "target"))
str(train_data)

#Function by while to calculate the mean of continuous variables
calculate_continuous_means <- function(data) {
  # Get the names of all columns in the dataset
  all_columns <- colnames(data)


  i <- 1 # Initialize an index


  while (i <= length(all_columns)) {          # Start a while loop

    current_column <- all_columns[i]          # Extract the current column
```

```
    if (is.integer(data[[current_column]])) {

      mean_value <- mean(data[[current_column]])


      cat("Mean of", current_column, ":", mean_value, "\n")
    }


    i <- i + 1
  }
}

calculate_continuous_means(data)
```