Introduction

Heart diseases which is a lay term for Cardiovascular Disease(CVD) is an important cause of morbidity and mortality in Malaysia based on (Kementerian Kesihatan Malaysia, 2017) and it's also the leading cause of death in the United States (Felson, 2021). Statistics from National Institutes of Health and other government sources made by American Heart Institutions in 2019 stated that 48% of adults that are 20 years old and above have some kind of cardiovascular disease. With the rising numbers everywhere in the world, various studies are being held to find out factors that are linked to heart disease. There are many type of CVD and commonly are Arrhythmias or abnormal heart rhythms, Coronary Artery Disease (CAD), stroke, heart failure and aneurysm to name a few. Hence, detailed investigation and research are needed to give accurate variables for CVD (Dutta, & Shroff, 2021).

In this report, the data that will be studied is based on a few heart variables that contain information of CVD patients in the CSV file given. Chest pain type has four categories which are Atypical Angina (ATA) which is caused by coronary artery diseases, Non-Anginal pain (NAP) which is a type of chest pain that is not caused by coronary artery disease, Asymptomatic (ASY) which means there is no chest pain symptoms and TA that stands for typical angina that is caused by reduced blood flow to the heart muscle. The Resting Blood Pressure is to check on hypertension existence in the patient, Cholesterol column for cholesterol count, Fasting Blood Sugar to screen for diabetes condition if any and Resting ECG or electrocardiographic is used to detect heart condition in a painless and faster way (Zhu et al., 2020). Furthermore, the Maximum Heart rate is to check how high patient heart rate can achieved, Exercise angina is to detect whether the patient has some kind of chest pain during other activities that makes your heart work harder, Oldpeak that stands for "ST depression induced by exercise relative to rest" is used to measure amount heart muscle damaged due to coronary artery disease, while ST-Slope which is the slope of the peak exercise ST segment is used to detect sign of coronary artery disease or serious heart disease(R S et al., 1986). And finally, the heart Disease column concluded whether the patient has a heart disease or not.

Column	Data
Age	Age
Sex	1 = male; 0 = female
Chest pain type	ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic, TA: Typical Angina
Resting Blood Pressure	Resting Blood pressure In mmHg
Cholesterol	Cholesterol level in mg/dl
Fasting Blood Sugar	1 = true; 0 = false
RestingECG	Normal, Left Ventricular Hypertrophy (LVH) , ST
Maximum Heart Rate	Maximum Heart Rate achieved
ExerciseAngina	1 = yes; 0 = no
OLDPEAK	Amount heart muscle damaged
ST_SLOPE	Up = Has CAD, Flat = No CAD
Heart Disease	1 = yes; 0 = no

Table 1: Heart.csv dataset column overview

This report consists of two parts analysis that will explain the details of the data, the relationship of the data and conclusion can be made based on the analysis made. The Exploratory Data Analysis (EDA) part is to understand and gain insights of the underlying structure of the dataset. The main focus of EDA is to understand the overview of the dataset and do some preprocessing like checking for missing values and removing outliers, then another part will discuss the insights of the dataset using Descriptive Analysis where graphical information is shown for the numerical and categorical part of the analysis, in order to help in the analysis of understanding the patterns, the relationship between variables and also to find what factors contribute most to heart disease presence.

The rising number of heart diseases is worrying, and various research is being held to recognise the factors that cause this health problem. The understanding of cardiovascular health and the identification of significant risk factors play a crucial role in diagnosing and preventing heart diseases. This report presents a comprehensive exploratory and descriptive analysis of a dataset encompassing various variables related to cardiovascular health. With a focus on key factors such as age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, resting ECG results, maximum heart rate achieved, exercise angina, oldpeak, ST slope, and the presence of heart disease, this analysis aims to uncover meaningful insights and patterns that can contribute to improved diagnosis of cardiovascular conditions.

By examining the relationships between these variables and heart disease incidence, we hope to gain valuable knowledge that can assist healthcare professionals in making informed decisions and implementing effective preventive measures.

Exploratory Data Analysis (EDA)

```
      Age            Sex            ChestPainType      Resting Blood Pressure
 Min.   :28.00   Length:918         Length:918         Min.   :  0.0
 1st Qu.:47.00   Class :character   Class :character   1st Qu.:120.0
 Median :54.00   Mode  :character   Mode  :character   Median :130.0
 Mean   :53.51                                         Mean   :132.4
 3rd Qu.:60.00                                         3rd Qu.:140.0
 Max.   :77.00                                         Max.   :200.0
  Cholesterol    Fasting Blood Sugar  RestingECG         Maximum Heart Rate
 Min.   :  0.0   Length:918          Length:918         Min.   : 60.0
 1st Qu.:173.2   Class :character    Class :character   1st Qu.:120.0
 Median :223.0   Mode  :character    Mode  :character   Median :138.0
 Mean   :198.8                                         Mean   :136.8
 3rd Qu.:267.0                                         3rd Qu.:156.0
 Max.   :603.0                                         Max.   :202.0
 ExerciseAngina       Oldpeak          ST_Slope          HeartDisease
 Length:918         Min.   :-2.6000   Length:918         Length:918
 Class :character   1st Qu.: 0.0000   Class :character   Class :character
 Mode  :character   Median : 0.6000   Mode  :character   Mode  :character
                    Mean   : 0.8874
                    3rd Qu.: 1.5000
                    Max.   : 6.2000
```

Diagram 1 : Summary of Data set

```
'data.frame':   918 obs. of  12 variables:
 $ Age                   : int  40 49 37 48 54 39 45 54 37 48 ...
 $ Sex                   : chr  "M" "F" "M" "F" ...
 $ ChestPainType         : chr  "ATA" "NAP" "ATA" "ASY" ...
 $ Resting Blood Pressure: int  140 160 130 138 150 120 130 110 140 120 ...
 $ Cholesterol           : int  289 180 283 214 195 339 237 208 207 284 ...
 $ Fasting Blood Sugar   : chr  "No" "No" "No" "No" ...
 $ RestingECG            : chr  "Normal" "Normal" "ST" "Normal" ...
 $ Maximum Heart Rate    : int  172 156 98 108 122 170 170 142 130 120 ...
 $ ExerciseAngina        : chr  "N" "N" "N" "Y" ...
 $ Oldpeak               : num  0 1 0 1.5 0 0 0 0 1.5 0 ...
 $ ST_Slope              : chr  "Up" "Flat" "Up" "Flat" ...
 $ HeartDisease          : chr  "No" "Yes" "No" "Yes" ...
```

Diagram 2: Structure of variables

The summary() function was applied to the given dataset, providing an overview of the variables included in the dataset and their descriptive statistics. The summary reveals information about key variables related to heart disease. For the Age, Cholesterol, Oldpeak, Resting Blood Pressure, and Maximum Heart Rate variable, the summary displays the minimum and maximum ages recorded, along with the first quartile, median, and third quartile values. Similarly, for Sex, ChestPainType, Fasting Blood Sugar, RestingECG, ExerciseAngina, ST_Slope, and HeartDisease, the summary provides the frequencies or counts of different categories or levels. These descriptive statistics offer valuable insights into the distribution and characteristics of the variables in the dataset, providing a foundation for further exploratory data analysis and modelling tasks related to heart disease prediction and analysis.

The str() function was applied to the "heart" dataset, revealing the structure and information about the variables contained within. The output of str() provides details such as the data type and format of each variable. For the given dataset, it indicates that the dataset consists of 918 observations and 12 variables. The variables include Age, Cholesterol, Resting Blood Pressure, Maximum Heart Rate represented as an integer, Sex, Chest Pain Type, Fasting Blood Sugar, RestingECG, Maximum Heart Rate, ExerciseAngina, ST_Slope, and HeartDisease, all represented as character strings. Only Oldpeak is represented as numeric. The str() function provides a concise summary of the dataset's structure, which aids in understanding the types of variables present and informs subsequent data manipulation and analysis tasks.

	Age	Sex	ChestPainType	Resting Blood Pressure	Cholesterol	Fasting Blood Sugar
1	40	M	ATA	140	289	No
2	49	F	NAP	160	180	No
3	37	M	ATA	130	283	No
4	48	F	ASY	138	214	No
5	54	M	NAP	150	195	No
6	39	M	NAP	120	339	No

	RestingECG	Maximum Heart Rate	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
1	Normal	172	N	0.0	Up	No
2	Normal	156	N	1.0	Flat	Yes
3	ST	98	N	0.0	Up	No
4	Normal	108	Y	1.5	Flat	Yes
5	Normal	122	N	0.0	Up	No
6	Normal	170	N	0.0	Up	No

Diagram 3 : First 6 rows of dataset

The head() function was applied to the given dataset, providing a glimpse into the initial rows of the data. By using head(heart), we are presented with a subset of the dataset that showcases the first few observations. This allows us to quickly assess the structure and content of the data.

```
> sum(is.na(heart))
[1] 0
```

Diagram 4 : Checking missing data

The expression sum(is.na(heart)) was applied to the given dataset to determine the total count of missing values across the variables. Upon evaluation, it was found that the dataset contains no missing values, thus no further steps need to be taken to handle the missing values.

Age	Sex	ChestPainType
36	0	0
Resting Blood Pressure	Cholesterol	Fasting Blood Sugar
51	8	0
RestingECG	Maximum Heart Rate	ExerciseAngina
0	28	0
Oldpeak	ST_Slope	HeartDisease
31	0	0

Diagram 5 : Outliers count

```
      Age             Sex             ChestPainType      Resting.Blood.Pressure
 Min.   :35.00   Length:784         Length:784         Min.   : 96.0
 1st Qu.:47.00   Class :character   Class :character   1st Qu.:120.0
 Median :54.00   Mode  :character   Mode  :character   Median :130.0
 Mean   :53.41                                         Mean   :130.9
 3rd Qu.:60.00                                         3rd Qu.:140.0
 Max.   :72.00                                         Max.   :165.0
  Cholesterol    Fasting.Blood.Sugar  RestingECG         Maximum.Heart.Rate
 Min.   :  0.0   Length:784          Length:784         Min.   : 86.0
 1st Qu.:177.0   Class :character    Class :character   1st Qu.:120.0
 Median :223.0   Mode  :character    Mode  :character   Median :138.0
 Mean   :199.4                                         Mean   :137.5
 3rd Qu.:266.0                                         3rd Qu.:156.0
 Max.   :417.0                                         Max.   :186.0
 ExerciseAngina       Oldpeak          ST_Slope          HeartDisease
 Length:784         Min.   :-1.1000   Length:784         Length:784
 Class :character   1st Qu.: 0.0000   Class :character   Class :character
 Mode  :character   Median : 0.5000   Mode  :character   Mode  :character
                    Mean   : 0.8133
                    3rd Qu.: 1.5000
                    Max.   : 3.0000
```

Diagram 6: Summary of dataset after removing outliers

The number of outliers present in each variable of the given dataset was determined through a systematic analysis. By utilising various statistical measures, such as mean and standard deviation, it was possible to identify observations that significantly deviated from the expected patterns. The threshold for identifying outliers was set at a certain number of standard deviations away from the mean. By applying this criterion to each variable, the number of outliers for each variable was computed hence resulting as shown in the picture above.

The Figure 9 shows summary of data after removing outliers, providing an updated overview of the variables. Removing outliers caused the changes in the minimum and maximum values recorded, along with the first quartile, median, and third quartile values. One of the changes can be seen in the reduced number of rows from original 918 to 784 rows of data where only around 0.05% of the data is removed. There are also few changes that can be seen such as minimum age, minimum OldPeak minimum Resting Blood Pressure and others. These updated descriptive statistics provide a refined understanding of the dataset, assisting in more accurate data analysis and interpretation, particularly in the context of heart disease.

The main idea of a EDA is to get the summary of the data before doing a further analysis. We can figure out the range, mean, mode or median of the data or even whether the data was distributed normally or skewing. There are two kinds of data in this data set which are numerical dataset and categorical dataset. The following analysis will identify the data shape, pattern and character.

- **Numerical Variables**

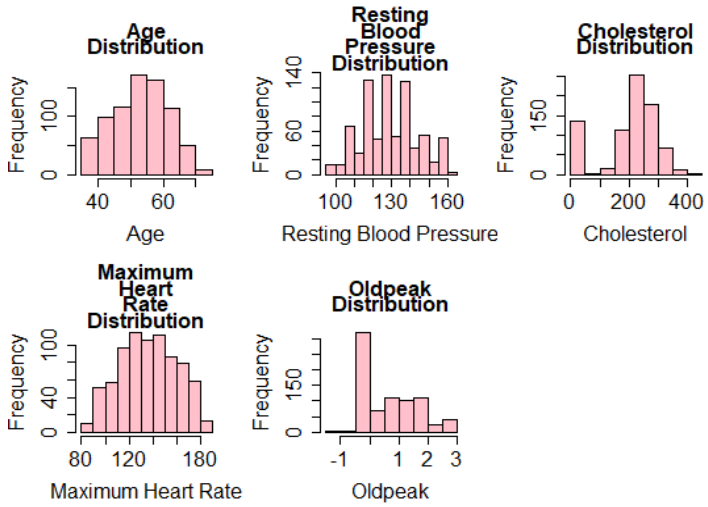

Diagram 7 : Distribution with outliers.

Diagram 7 above is created to show the distribution for 5 variables which are Age, Resting Blood Pressure, Cholesterol, maximum Heart Rate and Oldpeak. Outliers is the data that is too different from the rest of the data and can distort the central tendency of the dataset. From the Resting Blood Pressure Diagram 7 above, some of the blood pressure is very low, almost to 100 while most of the patients have blood pressure between 120 to 140 but there are some patients that have blood pressure between 120 to 160 in between. This graph is a bit hard to read and understand due to not being normally distributed yet. Another example we can see is in the cholesterol variable where there are around almost 150 individuals who have 0 cholesterol. It created an uneven distribution to identify the data pattern.

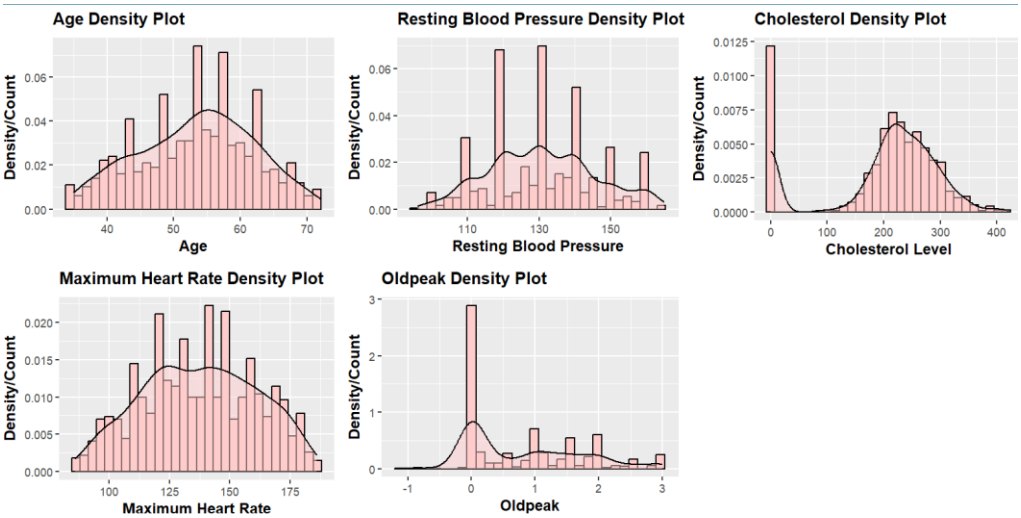

Diagram 8 : Density plot with histogram for numerical variables

A density plot shows the distribution of the data by using a smooth curve to connect the data points while a histogram shows the distribution of the data by using bars to represent the number of data points in each range of values. The density plot can show the peaks of where the values are concentrated. For the age density plot, most of the patients' age range concentrated from 50 to 60 years old and the least amount of individuals' ages are below 40 and near 70. Resting blood pressure has an uneven distribution , but the density plot focused towards the middle with resting blood pressure range from 120 to 140 mmhg. The diagram also shows that most cholesterol levels are concentrated around 200 mg/dl and 0 mg/dl despite outliers having been discarded. We can see that the dataset is little different between people who have cholesterols and none. Most maximum heart rates range from 125 to 150 and the oldpeak concentrated to 0 value means most individuals do not have any damage to the heart muscle due to any heart disease before. With this distribution, we can see the data leaned toward a certain age range, certain resting blood pressure and maximum heart rate with no oldpeak and cholesterol level around 200 mg/dl.

- **Categorical Variables**

Diagram 9: Histogram plot for percentage of Sex, ST-Slope type and Exercise Angina

For measuring the frequency count and percentage of each variable, histogram is used with different colours of the bar to differentiate the category. Diagram 9 shows the pink coloured bar for sex category, around 78.7% of the dataset is males and the balance is females. For the ST-Slope type, the flat or no slope is the highest count with around 50.38% which means 50% of the patients had none to not serious heart disease but 44. 77 % has Up St-Slope and 4.85% has Down ST-Slope which may indicate serious heart disease based on (R S et al., 1986). The right bar chart shows that more than 60% of the patients have Exercise Angina and the rest don't. In order to know this, they need to do some exercise to detect chest pain. So, from the bar chart above, we can identify that this dataset leans towards males gender compared to female, the ST-Slope for flat and up type dominated the dataset and 60% of the individuals experienced exercise angina from the dataset.

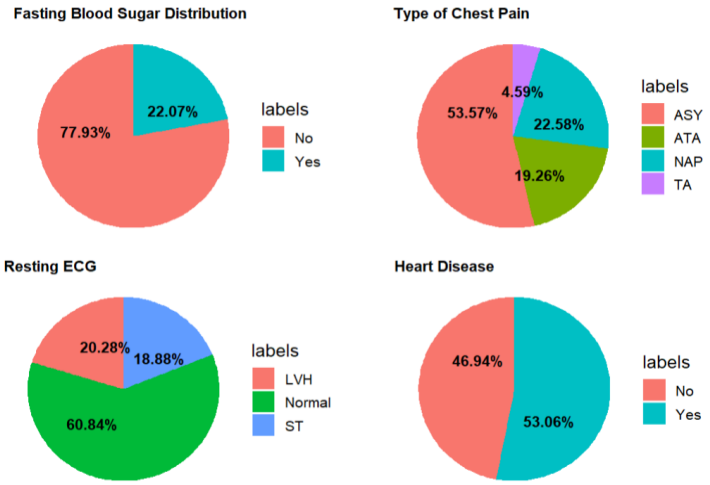

Diagram 10 : Frequency count and percentage for each variable

Diagram 10 shows the percentage for variable fasting blood sugar (FBS), chest pain type, resting ECG and percentage of how many individuals have heart diseases. From the FBS pie chart, around 78% of individuals have no diabetes since blood sugar is taken to measure the screen, diagnose, and monitor diabetes mellitus (Ghazanfari Z, et al., 2010). For the type of pain chest chart, ASY (no chest pain) dominates the dataset with 53.57%, while NAP (non-anginal pain) is experienced by 22.58% of individuals. With little difference of 2%, ATA chest pain comes to the third type of chest pain experienced by individuals which is 19.26% and TA has the least amount of individuals which is only 4.59%. The resting ECG pie chart shows that around 60% of individuals have normal resting ECG and 20.28% have LVH while the rest has ST. Lastly, we can see that 53.06% of the individuals have heart diseases and 46.94% don't from the heart diseases pie chart.

Descriptive Analysis

- **Categorical Variables**

The pie chart and bar chart below are used to show measurement of non-numerical data from the dataset. In this part, we analyse the correlation between gender which are female and male, and also the chest pain type which are Asymptomatic (ASY), Atypical Angina (ATA), Typical Angina (TA) and Non-Anginal Pain (NAP).

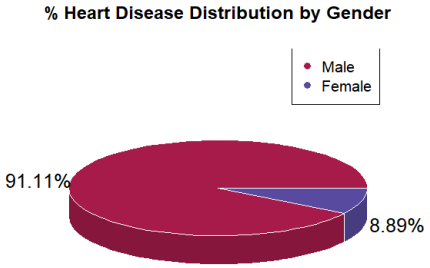

Diagram 11 : Percentage of heart disease between male and females

Pie charts are used to show the percentage of heart disease between male and females. Since the data category is only two, pie charts are used for easier interpretation of the gender variable percentage. Pie chart Diagram 11 is showing that a large number of males have heart disease which is 91.11 % compared to women which only has 8.89%. Hence, more than 90% of the analysis to understand the correlation of the factors for heart disease are mostly based on the male gender. We can assume that most males may have a higher potential of getting heart disease compared to women among the patients in this chart.

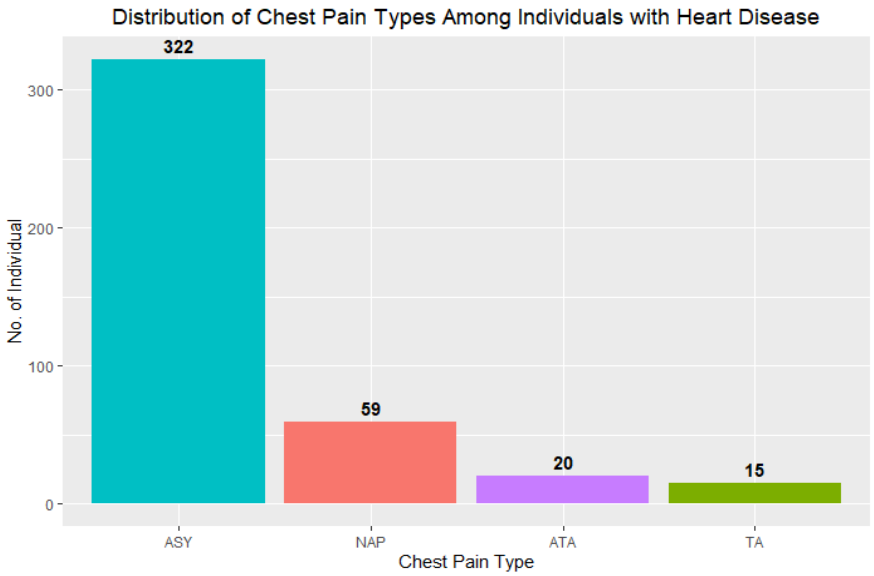

Diagram 12: Distribution of chest pain types among individuals with heart disease.

Then we examine which kind of chest pain is highest among the patients.We use bar plot since we only have 4 categories of the variable to be analysed. Based on Diagram 12 above, we can see that most patients (322 of individuals) do not have chest pain symptoms (ASY) and only 94 individuals experienced chest pain in total. This raises an alarming message that most heart disease can be undetectable due to lack of pain in the chest and thus can cause delayed treatment before it gets worse. Only 59 individuals experienced chest pain that is not cause to heart disease (NAP), 20 individuals experienced chest pain that is related to the heart (ATA) and only 15 experienced chest pain of TA.

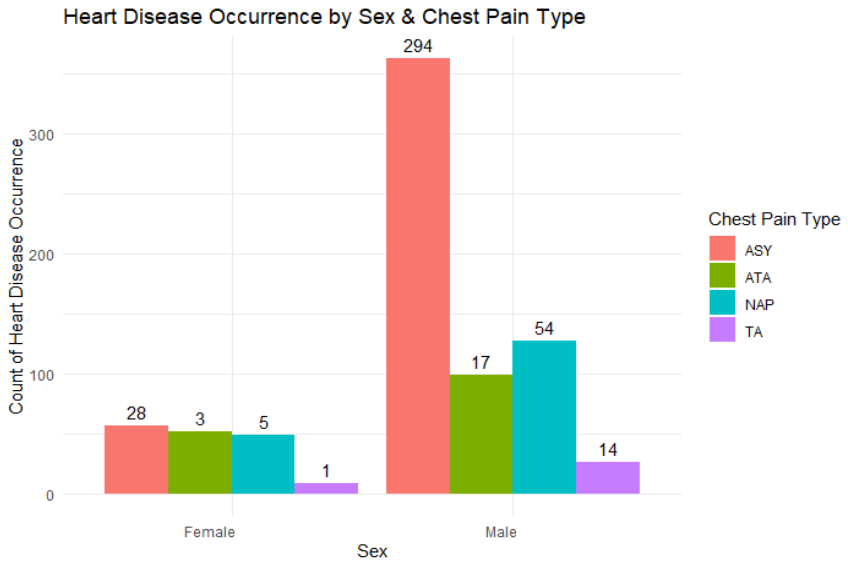

Diagram 13 : Heart disease occurrence by sex and chest pain type.

From the diagram, the most category for the male and female are ASY which is they do not experienced any chest pain, but this is understandable because from Diagram 12, ASY which stands for Asymptotic has the highest number of occurrence around 77% from the rest, so, the total number of ASY type occurrence among gender will also be high due

to the volume. The female gender has minimal representation in the ATA, NAP, and TA categories, with only 3, 5, and 1 individual, respectively, compared to the ASY, which has 28 individuals. For male gender, NAP is the second most common type of chest pain, experienced by 54 individuals out of a total of 59 individuals in Diagram 12. This means that while 54 males experienced chest pain, it is not chest pain caused by coronary artery disease. The least common type of chest pain for males is TA, experienced by only 14 individuals, while 17 males experienced ATA. Diagram 13 shows the high number of ASY in males is due to 90% of the patients data being males from Diagram 11 and also, the high number of ASY is because 77% of the total individuals from Diagram 12 stated no chest pain.

- **Numerical Variables**

The dataset contains five numerical variables: age, resting blood pressure, maximum heart rate, and old peak. This section will examine the relationship between those numerical variables as well as their relationship to the target variable, heart disease.

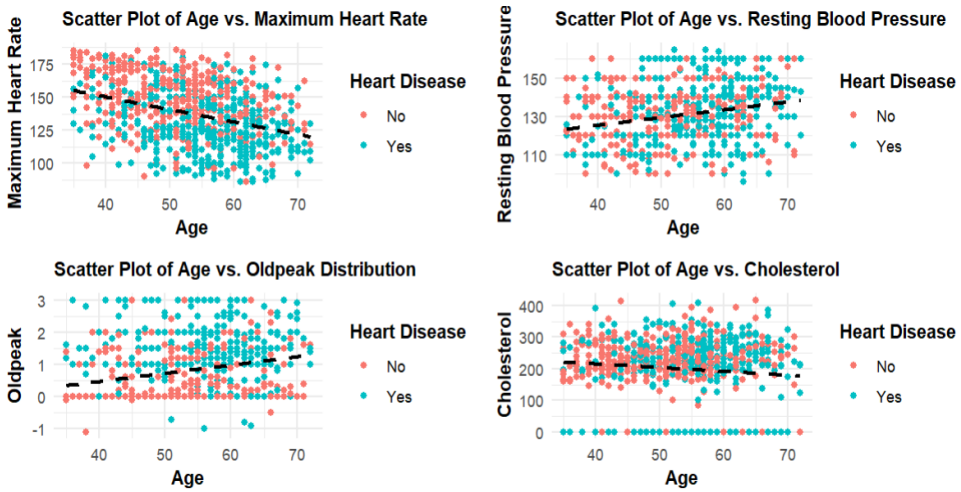

Diagram 14 :Scatter plot between age and other variables

A scatter plot is a data visualisation technique that shows the relationship between two numerical variables. The scatter plots Diagram 14 show how age impacts other parameters, with a regression line created to help understand the relationship. There is a negative relationship between age and maximal heart rate, indicating that as age grows, heart rate decreases or vice versa. Age and blood pressure, as well as oldpeak, appear to have a positive relationship, indicating that as age grows, so may blood pressure and oldpeak, or vice versa. For age and cholesterol, we can see that the regression line is practically horizontal, indicating that they have a very weak relationship, and that ageing has no effect on one's cholesterol level.

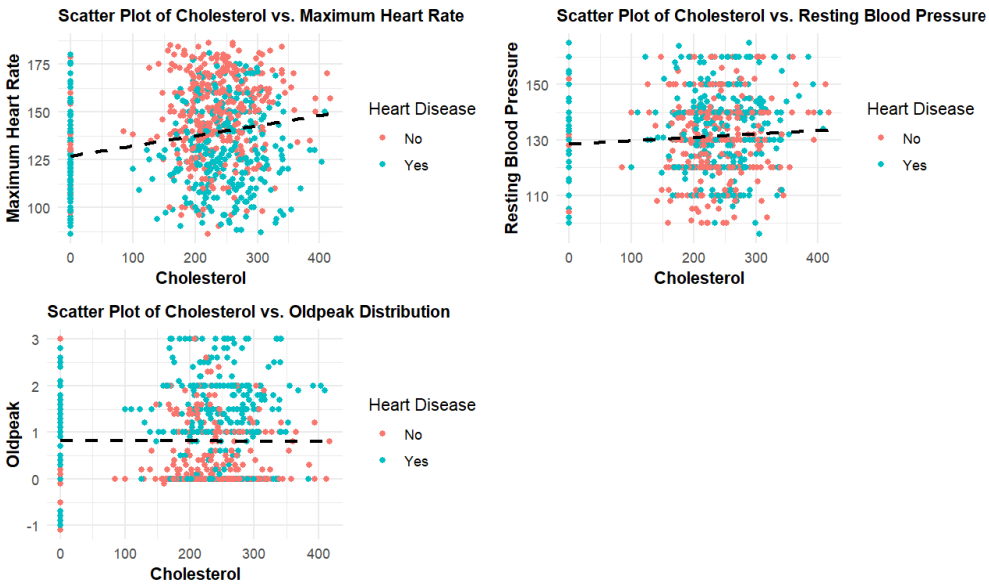

Diagram 15 : Scatter plot between cholesterol and other variables

The scatter plots Diagram 15 are plotted to show how cholesterol affects the other variables. cholesterol and heart rate have a positive relationship, which means that if cholesterol levels rise, so will heart rate. When we compare cholesterol to oldpeak and blood pressure, we can see that the correlation relationship is practically horizontal, indicating that there is either no or a weak correlation between them, implying that cholesterol has no effect on blood pressure or oldpeak.

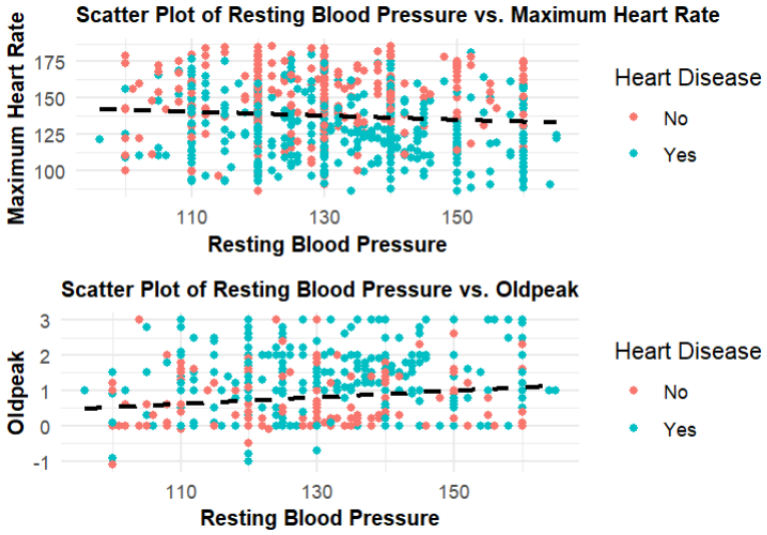

Diagram 16: Scatter plot of Resting Blood Pressure vs Maximum Heart rate with Regression Line.

The scatter plots shown above were created to show how blood pressure affected the remaining factors of heart rate and old peak. The scatter plots show that blood pressure and old peak then have a greater correlation than blood pressure and heart rate. Furthermore, blood pressure and age have a positive relationship, whereas blood pressure and heart rate have a slight negative relationship.

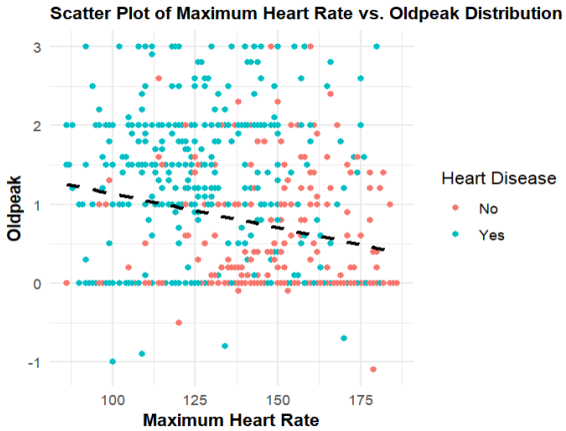

Diagram 17 : Scatter plot between heart rate and old peak

The scatter plot above demonstrates how heart rate affects old peak, and we can see that they appear to have a significant negative relationship, with heart rate increasing causing old peak to fall and vice versa.

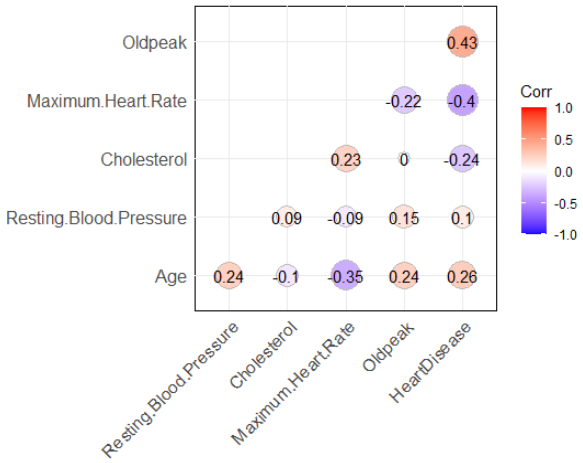

Diagram 18: Correlation plot between variables

After using scatter plots to acquire an understanding of the relationship between the variables, a correlation plot is drawn to see the exact correlation between them as well as the target variable heart disease. A correlation plot depicts the relationships between variables in a dataset visually. It allows us to comprehend the interactions and interdependence between several variables. The correlation between two variables is represented by each cell in the matrix, which ranges from -1 to 1. The intensity and direction of the correlation are represented by the colour of the cell or a heatmap. A high positive correlation is frequently represented by a bright colour, and a high negative correlation is frequently represented by a dark colour.

In this dataset, we can see that age and heart rate have the highest correlation among the numeric variables with -0.35, indicating that they are substantially negatively connected. This could be because the body is healthier at younger ages than it is at older ages. On the other hand, we can conclude that there is no association between oldpeak and cholesterol because the correlation is close to zero. If we look at age and blood pressure, they have a positive correlation with 0.24, showing that as people become older, their blood pressure may rise since the arteries become stiffer as we age, causing the blood pressure to rise. While for cholesterol and heart rate, there is a correlation between the two of 0.23, indicating that high cholesterol will increase heart rate. This could be due to cholesterol clogging blood vessels, causing the heart to pump faster to get enough oxygen.

Furthermore, to see which variables correlated with the target variables, we can see that Oldpeak and Heart Disease have the highest correlation, 0.43, indicating that they are strongly positively correlated, followed by Maximum Heart Rate and Heart Disease with -0.4 correlation, indicating that they have a strongly negative relationship. This could be since the lower the heart rate will cause oxygen deficit, causing the greater the risk of heart disease. Then age and heart

disease have a 0.24 relationship, indicating that age does increase the risk of developing heart disease, which could be caused by an unhealthy lifestyle, high blood pressure, or any other condition.

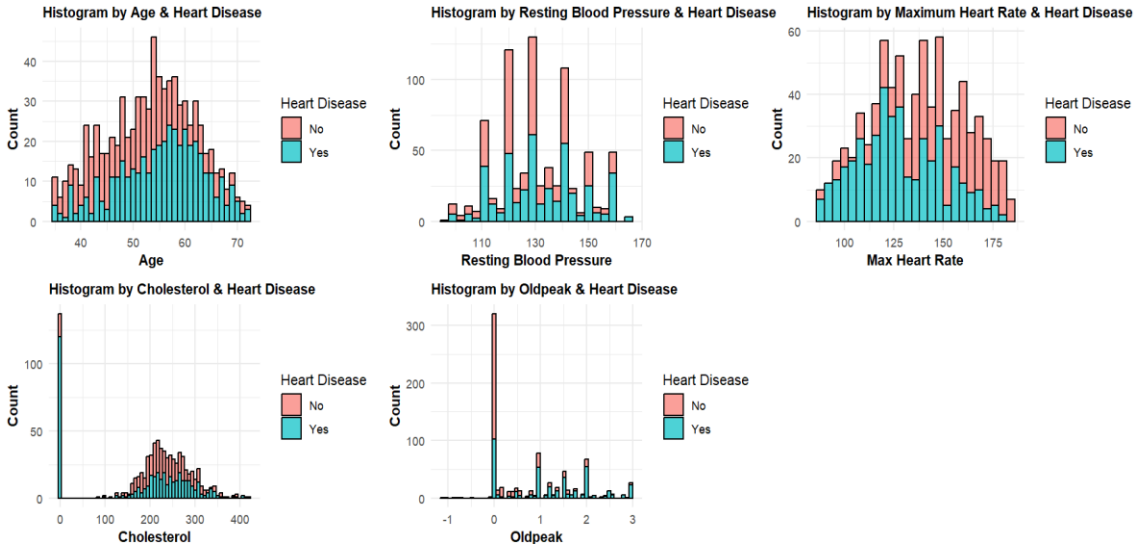

Diagram 19 : Relation of numerical variables to heart diseases

The above histogram was displayed to show how the numerical variables related with the target variable, heart disease, with the possibility of yes or no. If we look at the age histogram, we can see that persons between the ages of 50 and 60 have a lower risk of developing heart disease than people in other age groups. Furthermore, if we look at a blood pressure histogram, we can notice that high no categories are in the range of 120 to 140, which could be attributable to the fact that 120 to 140 is considered a normal pressure range, implying a lower risk of heart disease. With a heart rate histogram, we can see that when the maximum heart rate increases, the more the no category, and when the rate is low, the more the yes category, implying that fewer people suffer heart disease with a proper heartbeat rate. As per the cholesterol level histogram, less people are having heart disease at level 200, as 200 mg/dL is considered a good cholesterol level. When we look at the old peak histogram, we can see that when the value is zero, there is a lower risk of developing heart disease, and when it is greater than zero, the risk of developing heart disease increases.

Conclusion

Based on the analysis made, there are more males in these studies compared to women with the majority age between 50 to 60 years old. Hence, the normal high blood pressure for this age group is between 120 to 140 mmhg since the higher the age, the higher the resting blood pressure from the analysis. Furthermore, if the patient has damage to heart muscles which can be indicated by high numbers of oldpeak, the blood pressure will also increase. The normal cholesterol level is around 200mm/dl and while it has nothing to do with age, higher cholesterol results in higher heart rate. Also, cholesterol has been found to have no effect on blood pressure and old peaks. Despite heart rate being high, it has a lower possibility of getting heart diseases since the heart easily pumps blood to the body. Also, due to increasing age, older people are prone to have heart disease hence, we can see the age group in this dataset is mostfully between 50 to 60 years old. Therefore, factors that affect heart disease are older age, high blood pressure, high number of old peaks and lower heart rate. The conclusion can be shown in the simpler table below to show the correlation of all factors that are being monitored for heart diseases after analysis has been made. The above findings can be used by individuals or health sectors to take preventive steps by taking into account the related factors to heart disease in order to reduce the risk of getting heart disease.

Age ↑	Heart rate ↓	Resting B.P ↑	OldPeak (OP) ↑	Cholesterol ↓ (No correlation between BP and OP)	Heart Diseases ↑

Table 2 : Factors affecting heart diseases based on analysis made

References

American Heart Association. (n.d.). *Understanding Blood Pressure Readings*. American Heart Association. Retrieved May 22, 2023, from https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

American Heart Association, Inc. (2019). Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association. *Circulation*, *139*(10). https://doi.org/10.1161/CIR.0000000000000659

American Heart Association News. (2019, January 31). *Cardiovascular diseases affect nearly half of American adults, statistics show*. American Heart Association. Retrieved May 22, 2023, from https://www.heart.org/en/news/2019/01/31/cardiovascular-diseases-affect-nearly-half-of-american-adults-statistics-show

Dutta,, D. S., & Shroff, D. S. (2021, July 14). *Non-Communicable Diseases*. Medindia. Retrieved May 22, 2023, from https://www.medindia.net/patientinfo/non-communicable-diseases.htm#what-are-non-communicable-diseases

Felson, S. (2021, November 2). *Cardiovascular (Heart) Diseases: Types and Treatments*. WebMD. Retrieved May 22, 2023, from https://www.webmd.com/heart-disease/guide/diseases-cardiovascular

Ghazanfari Z, Haghdoost AA, Alizadeh SM, Atapour J, Zolala F. A Comparison of HbA1c and Fasting Blood Sugar Tests in General Population. Int J Prev Med. 2010 Summer;1(3):187-94. PMID: 21566790; PMCID: PMC3075530.

Kementerian Kesihatan Malaysia. (2017). *Primary & Secondary Prevention of CVD 2017*. Portal Rasmi Kementerian Kesihatan Malaysia. https://www.moh.gov.my/moh/resources/Penerbitan/CPG/CARDIOVASCULAR/3.pdf

R S, F., K E, N., T R, V., S D, M., & R C, B. (1986). ST segment/heart rate slope as a predictor of coronary artery disease: comparison with quantitative thallium imaging and conventional ST segment criteria. *Am. Heart J.; (United States)*, *2*. https://doi.org/10.1016/0002-8703(86)90265-6

Zhu, H., Cheng, C., Yin, H., Li, X., & Zuo, P. (2020, June 4). Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study. *The Lancet Digital Health*, *2*(7), 348 - e357. https://doi.org/10.1016/S2589-7500(20)30107-2

APPENDIX A : R-code

```
library(ggplot2)
library(dplyr)
library(gridExtra)

heart<- read.csv("C:/Users/syein/Downloads/heart.csv")
summary(heart)

# Convert 0 to "No" and 1 to "Yes" in the specified column
heart$FastingBS <- ifelse(heart$FastingBS == 0, "No", "Yes")
heart$HeartDisease <- ifelse(heart$HeartDisease == 0, "No", "Yes")

summary(heart)
str(heart)

colnames(heart)[colnames(heart) == "RestingBP"] <- "Resting Blood Pressure"
colnames(heart)[colnames(heart) == "FastingBS"] <- "Fasting Blood Sugar"
colnames(heart)[colnames(heart) == "MaxHR"] <- "Maximum Heart Rate"

str(heart)
head(heart)
sum(is.na(heart)) #missing values in a data frame or column

#boxplot with outliers
boxplot(heart$Age)
boxplot(heart$`Resting Blood Pressure`)
boxplot(heart$Cholesterol)
boxplot(heart$`Maximum Heart Rate`)
boxplot(heart$Oldpeak)

#number of outliers for all variables

threshold <- 2 #for adjusting outliers
num_outliers <- sapply(heart, function(x) {
  mean_value <- mean(x, na.rm = TRUE)
  sd_value <- sd(x, na.rm = TRUE)
  num_outliers <- sum(x < mean_value - threshold * sd_value | x > mean_value + threshold * sd_value, na.rm = TRUE)
  num_outliers
})
num_outliers

#clean all of outliers
df_clean <- as.data.frame(lapply(heart, function(x) {
  mean_value <- mean(x, na.rm = TRUE)
  sd_value <- sd(x, na.rm = TRUE)
  x_clean <- x
  x_clean[x_clean < mean_value - threshold * sd_value | x_clean > mean_value + threshold * sd_value] <- NA
  x_clean
}))


df_clean <- na.omit(df_clean) #deleting missing values after removing outliers
total_rows <- nrow(heart) #number of row before removing outliers
total_rows
total_rows <- nrow(df_clean) ##number of row after removing outliers
total_rows
```

```
#histogram with outliers
par(mfrow = c(2, 3))  # Set the layout to 2 rows and 3 columns

# Histogram of Age
hist(df_clean$Age, main = "Age Distribution", xlab = "Age", ylab = "Frequency", col = "pink", cex.main = 0.8)

# Histogram of Resting Blood Pressure
hist(df_clean$Resting.Blood.Pressure, main = "Resting Blood Pressure Distribution", xlab = "Resting Blood Pressure", ylab
= "Frequency", col = "pink", cex.main = 0.8)

# Histogram of Cholesterol
hist(df_clean$Cholesterol, main = "Cholesterol Distribution", xlab = "Cholesterol", ylab = "Frequency", col = "pink",
cex.main = 0.8)

# Histogram of Maximum Heart Rate
hist(df_clean$Maximum.Heart.Rate, main = "Maximum Heart Rate Distribution", xlab = "Maximum Heart Rate", ylab =
"Frequency", col = "pink", cex.main = 0.8)

# Histogram of Oldpeak
hist(df_clean$Oldpeak, main = "Oldpeak Distribution", xlab = "Oldpeak", ylab = "Frequency", col = "pink", cex.main = 0.8)


#density + Histogram plot without outliers

# Set the fill color for histograms and density plots
fill_color <- "#FFCCCC"  # Light pink color

# Define a function to create bold labels and titles
bold_labels <- function(label) {
  bquote(bold(.(label)))
}

# Density plots for each variable
density_plots <- list(
  ggplot(df_clean, aes(x = Age)) +
    geom_histogram(aes(y = ..density..), fill = fill_color, color = "black") +
    geom_density(alpha = 0.5, fill = fill_color, color = "black") +
    labs(title = bold_labels("Age Density Plot"), x = bold_labels("Age"), y = bold_labels("Density/Count")) +
    theme(plot.title = element_text(size = rel(1.2), face = "bold"),
          axis.title = element_text(size = rel(1.1), face = "bold")),

  ggplot(df_clean, aes(x = Resting.Blood.Pressure)) +
    geom_histogram(aes(y = ..density..), fill = fill_color, color = "black") +
    geom_density(alpha = 0.5, fill = fill_color, color = "black") +
    labs(title = bold_labels("Resting Blood Pressure Density Plot"), x = bold_labels("Resting Blood Pressure"), y =
bold_labels("Density/Count")) +
    theme(plot.title = element_text(size = rel(1.2), face = "bold"),
          axis.title = element_text(size = rel(1.1), face = "bold")),

  ggplot(df_clean, aes(x = Cholesterol)) +
    geom_histogram(aes(y = ..density..), fill = fill_color, color = "black") +
    geom_density(alpha = 0.5, fill = fill_color, color = "black") +
    labs(title = bold_labels("Cholesterol Density Plot"), x = bold_labels("Cholesterol Level"), y =
bold_labels("Density/Count")) +
    theme(plot.title = element_text(size = rel(1.2), face = "bold"),
          axis.title = element_text(size = rel(1.1), face = "bold")),
```

```
  ggplot(df_clean, aes(x = Maximum.Heart.Rate)) +
    geom_histogram(aes(y = ..density..), fill = fill_color, color = "black") +
    geom_density(alpha = 0.5, fill = fill_color, color = "black") +
    labs(title = bold_labels("Maximum Heart Rate Density Plot"), x = bold_labels("Maximum Heart Rate"), y =
bold_labels("Density/Count")) +
    theme(plot.title = element_text(size = rel(1.2), face = "bold"),
          axis.title = element_text(size = rel(1.1), face = "bold")),

  ggplot(df_clean, aes(x = Oldpeak)) +
    geom_histogram(aes(y = ..density..), fill = fill_color, color = "black") +
    geom_density(alpha = 0.5, fill = fill_color, color = "black") +
    labs(title = bold_labels("Oldpeak Density Plot"), x = bold_labels("Oldpeak"), y = bold_labels("Density/Count")) +
    theme(plot.title = element_text(size = rel(1.2), face = "bold"),
          axis.title = element_text(size = rel(1.1), face = "bold"))
)

# Combine density plots into a grid
grid.arrange(grobs = density_plots, nrow = 2, ncol = 3)


#bar chart

df_summary_sex <- df_clean %>%
  count(Sex) %>%
  mutate(percent = prop.table(n) * 100)

df_summary_slope <- df_clean %>%
  count(ST_Slope) %>%
  mutate(percent = prop.table(n) * 100)

df_summary_angina <- df_clean %>%
  count(ExerciseAngina) %>%
  mutate(percent = prop.table(n) * 100)

#bar chart updated

plot_sex <- ggplot(df_clean, aes(x = Sex)) +
  geom_bar(fill = "pink") +
  geom_text(data = df_summary_sex, aes(label = n, y = n, vjust = -0.5), color = "black") +
  geom_text(data = df_summary_sex, aes(label = paste0("(", round(percent, 2), "%)"), y = n, vjust = 1.5, fontface = "bold"),
color = "black") +
  labs(title = "Sex Distribution", x = "Sex", y = "Frequency") +
  scale_x_discrete(labels = c("Female", "Male")) +  # Change the axis labels
  theme(plot.title = element_text(size = 14, face = "bold"), axis.text = element_text(size = 12))  # Adjust the size and
boldness of the title and axis text

plot_slope <- ggplot(df_clean, aes(x = ST_Slope)) +
  geom_bar(fill = "lightgreen") +
  geom_text(data = df_summary_slope, aes(label = n, y = n, vjust = -0.5, fontface = "bold"), color = "black") +
  geom_text(data = df_summary_slope, aes(label = paste0("(", round(percent, 2), "%)"), y = n, vjust = 1.5, fontface =
"bold"), color = "black") +
  labs(title = "ST Slope Distribution", x = "Type of ST Slope", y = "Frequency") +
  theme(plot.title = element_text(size = 14, face = "bold"), axis.text = element_text(size = 12))  # Adjust the size and
boldness of the title and axis text

plot_angina <- ggplot(df_clean, aes(x = ExerciseAngina)) +
  geom_bar(fill = "grey") +
  geom_text(data = df_summary_angina, aes(label = n, y = n, vjust = -0.5, fontface = "bold"), color = "black") +
```

```
  geom_text(data = df_summary_angina, aes(label = paste0("(", round(percent, 2), "%)"), y = n, vjust = 1.5, fontface =
"bold"), color = "black") +
  labs(title = "Exercise Angina Distribution", x = "Exercise Angina", y = "Frequency") +
  scale_x_discrete(labels = c("No", "Yes")) +  # Change the axis labels
  theme(plot.title = element_text(size = 14, face = "bold"), axis.text = element_text(size = 12))  # Adjust the size and
boldness of the title and axis text


# Combine plots into a grid
grid.arrange(plot_sex, plot_slope, plot_angina, nrow = 1)


#pie chart

http://127.0.0.1:46529/graphics/plot_zoom_png?width=510&height=436
# Function to create polar bar plot with labels
create_polar_plot <- function(df, title) {
  ggplot(df, aes(x = "", y = values, fill = labels)) +
    geom_bar(stat = "identity") +
    coord_polar("y", start = 0) +
    geom_text(aes(label = paste0(round(percent, 2), "%")),
              position = position_stack(vjust = 0.7),
              fontface = "bold") +
    labs(title = title) +
    theme_void(base_size = 13)+
    theme(plot.title = element_text(size = 11, face = "bold"))
}

# Calculate frequency count and percentage for each variable
Fasting.Blood.Sugar_counts <- table(df_clean$Fasting.Blood.Sugar)
df1 <- data.frame(labels = names(Fasting.Blood.Sugar_counts), values = as.numeric(Fasting.Blood.Sugar_counts))
df1$percent <- df1$values / sum(df1$values) * 100

ChestPainType_counts <- table(df_clean$ChestPainType)
df2 <- data.frame(labels = names(ChestPainType_counts), values = as.numeric(ChestPainType_counts))
df2$percent <- df2$values / sum(df2$values) * 100

RestingECG_counts <- table(df_clean$RestingECG)
df3 <- data.frame(labels = names(RestingECG_counts), values = as.numeric(RestingECG_counts))
df3$percent <- df3$values / sum(df3$values) * 100

HeartDisease_counts <- table(df_clean$HeartDisease)
df4 <- data.frame(labels = names(HeartDisease_counts), values = as.numeric(HeartDisease_counts))
df4$percent <- df4$values / sum(df4$values) * 100

# Create plots
plot1 <- create_polar_plot(df1, "Fasting Blood Sugar Distribution")
plot2 <- create_polar_plot(df2, "Type of Chest Pain")
plot3 <- create_polar_plot(df3, "Resting ECG")
plot4 <- create_polar_plot(df4, "Heart Disease")

# Combine plots into a grid
grid.arrange(plot1, plot2, plot3, plot4, nrow = 2)

library(ggplot2)
library(dplyr)
library(RColorBrewer) #barplot color
library(plotrix) #3Dpiechart
```

```
#Descriptive analysis

##1. Create a bar plot to analyze the count of males and females
#with heart disease to understand if there are any gender-related patterns.

#to rename the chr variable in column
df_clean$Sex = gsub("M", "Male", df_clean$Sex)
df_clean$Sex = gsub("F", "Female", df_clean$Sex)

df_clean

#to sum up total of heart disease by gender
df2 = df_clean[, c(2, 12)] #to choose column
df3 = aggregated_data=df_clean%>%
  group_by(df_clean$Sex, df_clean$HeartDisease) %>%
  summarise(Count = n())

##1. create a pie chart to analyze the count of males and females
#with heart disease to understand if there are any gender-related patterns.

#filter the data for individuals with heart disease
with_disease = df_clean %>%
  filter(df_clean$HeartDisease == "Yes")

#calculate the count of males and females with heart disease
sex_count = with_disease %>%
  count(Sex)

#create 3D pie chart
#dataframe
df.sex_count = c(379, 37)
sex_labels = c("Male", "Female")

#percentage calculation
sex_percentage = paste0(round(100*df.sex_count/sum(df.sex_count), 2), "%")
sex_percentage
pie3D = pie3D(df.sex_count,
      col = hcl.colors(length(sex_percentage), "Spectral"),
      border = "white",
      main = "% Heart Disease Distribution by Gender",
      labels = sex_percentage,
      labelcex = 1.3,
      labelcol = "black")
legend(x = "topright", legend = sex_labels,
       col = hcl.colors(length(sex_percentage), "Spectral"), pch = 16) #add legends
pie3D = pie3D +  theme(plot.title = element_text(hjust = 0.5)) #to insert title at center

#2. Create bar plot to examine the distribution of chest pain types among individuals with heart disease.
#This can help to identify the most prevalent chest pain type associated with heart disease.

#Calculate the count of each chest pain type
count_type = with_disease %>%
  count(with_disease$ChestPainType)

#Sort the number in descending order
count_type = count_type %>%
  arrange(desc(n))
```

```
#Create the bar plot using ggplot2
# Specify the color palette from RColorBrewer
color_palette = brewer.pal(4, "Paired")  # Adjust the number and palette name as needed

Pain_type = c("ASY", "NAP", "ATA", "TA")
N = c(322, 59, 20, 15)
count_type = data.frame(
  Pain_type = c("ASY", "NAP", "ATA", "TA"),
  n = c(322, 59, 20, 15))

# Set the order of Pain_type levels
count_type$Pain_type = factor(count_type$Pain_type, levels = c("ASY", "NAP", "ATA", "TA"))

bar_plot = ggplot(count_type, aes(x = Pain_type, y = n,  fill = color_palette)) +
  geom_bar(stat = "identity") +
  labs(title = "Distribution of Chest Pain Types Among Individuals with Heart Disease",
       x = "Chest Pain Type",
       y = "No. of Individual",) +
  theme(legend.position = "none") + theme(plot.title = element_text(hjust = 0.5))  #to remove theme legend position and
adjust the title position to center.
  theme_minimal()
bar_plot = bar_plot + geom_text(aes(label = N), vjust = -0.5, fontface = "bold")
bar_plot

#3. Calculate the count of heart disease occurrences by gender, chest pain type

with_disease = df_clean %>% filter(df_clean$HeartDisease == "Yes")
disease_counts = with_disease %>%
  group_by(Sex, ChestPainType) %>%
  summarise(count = sum(HeartDisease == "Yes"))
disease_counts
count_label = c(28, 3, 5, 1, 294, 17, 54, 14)

#Create group bar plot
groupbar_plot = ggplot(disease_counts, aes(x = Sex, y = count, fill = ChestPainType)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Heart Disease Occurrence by Sex & Chest Pain Type",
       x = "Sex",
       y = "Count of Heart Disease Occurrence",
       fill = "Chest Pain Type") +
  theme_minimal()
groupbar_plot = groupbar_plot + geom_text(aes(label = count_label), position = position_dodge(width = 0.9), vjust = -0.5)
groupbar_plot

library(ggplot2)
library(dplyr)
library(cowplot)
library(GGally)

df_clean$HeartDisease<-ifelse(df_clean$HeartDisease == "No", 0, 1)

##Correlation plot

# Select the variables of interest
variables <- c("Age", "Resting.Blood.Pressure", "Cholesterol", "Maximum.Heart.Rate", "Oldpeak", "HeartDisease")
subset_heart <- df_clean[variables]

# Compute the correlation matrix
```

```
cor_matrix <- cor(subset_heart)

# Load the ggcorrplot package
library(ggcorrplot)

# Create the correlation plot with numbers
ggcorrplot(cor_matrix, method = "circle", type = "lower", lab = TRUE)+
  theme(panel.background = element_rect(fill = "white"))

df_clean$HeartDisease <- ifelse(df_clean$HeartDisease == 0, "No", "Yes")

#AGE SCATTER PLOTS
scatter_plot1 <- ggplot(df_clean, aes(x = Age, y = Maximum.Heart.Rate, color = as.factor(HeartDisease))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "black", linetype = "dashed") +
  labs(x = bquote(bold("Age")), y = bquote(bold("Maximum Heart Rate")), color = bquote(bold("Heart Disease"))) +
  ggtitle(bquote(bold("Scatter Plot of Age vs. Maximum Heart Rate"))) +
  theme_minimal() +
  theme(plot.title = element_text(size = 11, face = "bold"),
        axis.title = element_text(size = 11, face = "bold"))

scatter_plot2 <- ggplot(df_clean, aes(x = Age, y = Resting.Blood.Pressure, color = as.factor(HeartDisease))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "black", linetype = "dashed") +
  labs(x = bquote(bold("Age")), y = bquote(bold("Resting Blood Pressure")), color = bquote(bold("Heart Disease"))) +
  ggtitle(bquote(bold("Scatter Plot of Age vs. Resting Blood Pressure"))) +
  theme_minimal() +
  theme(plot.title = element_text(size = 11, face = "bold"),
        axis.title = element_text(size = 11, face = "bold"))

scatter_plot3 <- ggplot(df_clean, aes(x = Age, y = Oldpeak, color = as.factor(HeartDisease))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "black", linetype = "dashed") +
  labs(x = bquote(bold("Age")), y = bquote(bold("Oldpeak")), color = bquote(bold("Heart Disease"))) +
  ggtitle(bquote(bold("Scatter Plot of Age vs. Oldpeak Distribution"))) +
  theme_minimal() +
  theme(plot.title = element_text(size = 11, face = "bold"),
        axis.title = element_text(size = 11, face = "bold"))

scatter_plot4 <- ggplot(df_clean, aes(x = Age, y = Cholesterol, color = as.factor(HeartDisease))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "black", linetype = "dashed") +
  labs(x = bquote(bold("Age")), y = bquote(bold("Cholesterol")), color = bquote(bold("Heart Disease"))) +
  ggtitle(bquote(bold("Scatter Plot of Age vs. Cholesterol"))) +
  theme_minimal() +
  theme(plot.title = element_text(size = 11, face = "bold"),
        axis.title = element_text(size = 11, face = "bold"))

multi_plot <- cowplot::plot_grid(scatter_plot1, scatter_plot2, scatter_plot3, scatter_plot4, nrow = 2)

print(multi_plot)

#CHOLESTROL SCATTER PLOTS

# Scatter plot for Cholesterol vs. Maximum Heart Rate
scatter_plot1 <- ggplot(df_clean, aes(x = Cholesterol, y = Maximum.Heart.Rate, color = as.factor(HeartDisease))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "black", linetype = "dashed") +
```

```
  labs(x = "Cholesterol", y = "Maximum Heart Rate", color = "Heart Disease") +
  ggtitle("Scatter Plot of Cholesterol vs. Maximum Heart Rate") +
  theme_minimal()+
  theme(plot.title = element_text(size = 11, face = "bold"),
        axis.title = element_text(size = 11, face = "bold"))

# Scatter plot for Cholesterol vs. Resting Blood Pressure
scatter_plot2 <- ggplot(df_clean, aes(x = Cholesterol, y = Resting.Blood.Pressure, color = as.factor(HeartDisease))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "black", linetype = "dashed") +
  labs(x = "Cholesterol", y = "Resting Blood Pressure", color = "Heart Disease") +
  ggtitle("Scatter Plot of Cholesterol vs. Resting Blood Pressure") +
  theme_minimal()+
  theme(plot.title = element_text(size = 11, face = "bold"),
        axis.title = element_text(size = 11, face = "bold"))

# Scatter plot for Cholesterol vs. Oldpeak
scatter_plot3 <- ggplot(df_clean, aes(x = Cholesterol, y = Oldpeak, color = as.factor(HeartDisease))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "black", linetype = "dashed") +
  labs(x = "Cholesterol", y = "Oldpeak", color = "Heart Disease") +
  ggtitle("Scatter Plot of Cholesterol vs. Oldpeak Distribution") +
  theme_minimal()+
  theme(plot.title = element_text(size = 11, face = "bold"),
        axis.title = element_text(size = 11, face = "bold"))

# Create a grid of the four scatter plots
multi_plot <- cowplot::plot_grid(scatter_plot1, scatter_plot2, scatter_plot3, nrow = 2)

# Display the grid of scatter plots
print(multi_plot)


#BP SCATTER PLOTS

# Scatter plot for Resting Blood Pressure vs. Maximum Heart Rate
scatter_plot1 <- ggplot(df_clean, aes(x = Resting.Blood.Pressure, y = Maximum.Heart.Rate, color =
as.factor(HeartDisease))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "black", linetype = "dashed") +
  labs(x = "Resting Blood Pressure", y = "Maximum Heart Rate", color = "Heart Disease") +
  ggtitle("Scatter Plot of Resting Blood Pressure vs. Maximum Heart Rate") +
  theme_minimal()+
  theme(plot.title = element_text(size = 10, face = "bold"),
        axis.title = element_text(size = 10, face = "bold"))

# Scatter plot for Resting Blood Pressure vs. Oldpeak
scatter_plot2 <- ggplot(df_clean, aes(x = Resting.Blood.Pressure, y = Oldpeak, color = as.factor(HeartDisease))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "black", linetype = "dashed") +
  labs(x = "Resting Blood Pressure", y = "Oldpeak", color = "Heart Disease") +
  ggtitle("Scatter Plot of Resting Blood Pressure vs. Oldpeak") +
  theme_minimal()+
  theme(plot.title = element_text(size = 10, face = "bold"),
        axis.title = element_text(size = 10, face = "bold"))

# Create a grid of the four scatter plots
multi_plot <- cowplot::plot_grid(scatter_plot1, scatter_plot2, nrow = 2)
```

```
# Display the grid of scatter plots
print(multi_plot)


#HEART RATE SCATTER PLOTS


# Scatter plot for Maximum Heart Rate vs. Oldpeak
scatter_plot2 <- ggplot(df_clean, aes(x = Maximum.Heart.Rate, y = Oldpeak, color = as.factor(HeartDisease))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "black", linetype = "dashed") +
  labs(x = "Maximum Heart Rate", y = "Oldpeak", color = "Heart Disease") +
  ggtitle("Scatter Plot of Maximum Heart Rate vs. Oldpeak Distribution") +
  theme_minimal()+
  theme(plot.title = element_text(size = 11, face = "bold"),
        axis.title = element_text(size = 11, face = "bold"))

scatter_plot2

# Histogram for Age and Heart Disease
hist_age <- ggplot(df_clean, aes(x = Age, fill = as.factor(HeartDisease))) +
  geom_histogram(binwidth = 1, color = "black", alpha = 0.7) +
  labs(x = "Age", y = "Count", title = "Histogram by Age & Heart Disease") +
  scale_fill_discrete(name = "Heart Disease", labels = c("No", "Yes")) +
  theme_minimal()+
  theme(plot.title = element_text(size = 11, face = "bold"),
        axis.title = element_text(size = 11, face = "bold"))

# Histogram for Resting Blood Pressure and Heart Disease
hist_bp <- ggplot(df_clean, aes(x = Resting.Blood.Pressure, fill = as.factor(HeartDisease))) +
  geom_histogram(binwidth = 3, color = "black", alpha = 0.7) +
  labs(x = "Resting Blood Pressure", y = "Count", title = "Histogram by Resting Blood Pressure & Heart Disease") +
  scale_fill_discrete(name = "Heart Disease", labels = c("No", "Yes")) +
  theme_minimal()+
  theme(plot.title = element_text(size = 11, face = "bold"),
        axis.title = element_text(size = 11, face = "bold"))

# Histogram for Maximum Heart Rate and Heart Disease
hist_hr <- ggplot(df_clean, aes(x = Maximum.Heart.Rate, fill = as.factor(HeartDisease))) +
  geom_histogram(binwidth = 4, color = "black", alpha = 0.7) +
  labs(x = "Max Heart Rate", y = "Count", title = "Histogram by Maximum Heart Rate & Heart Disease") +
  scale_fill_discrete(name = "Heart Disease", labels = c("No", "Yes")) +
  theme_minimal()+
  theme(plot.title = element_text(size = 11, face = "bold"),
        axis.title = element_text(size = 11, face = "bold"))

# Histogram for Cholesterol and Heart Disease
hist_cholesterol <- ggplot(df_clean, aes(x = Cholesterol, fill = as.factor(HeartDisease))) +
  geom_histogram(binwidth = 7, color = "black", alpha = 0.7) +
  labs(x = "Cholesterol", y = "Count", title = "Histogram by Cholesterol & Heart Disease") +
  scale_fill_discrete(name = "Heart Disease", labels = c("No", "Yes")) +
  theme_minimal()+
  theme(plot.title = element_text(size = 11, face = "bold"),
        axis.title = element_text(size = 11, face = "bold"))

# Histogram for Oldpeak and Heart Disease
hist_oldpeak <- ggplot(df_clean, aes(x = Oldpeak, fill = as.factor(HeartDisease))) +
```

```
  geom_histogram(binwidth = 0.08, color = "black", alpha = 0.7) +
  labs(x = "Oldpeak", y = "Count", title = "Histogram by Oldpeak & Heart Disease") +
  scale_fill_discrete(name = "Heart Disease", labels = c("No", "Yes")) +
  theme_minimal()+
  theme(plot.title = element_text(size = 11, face = "bold"),
        axis.title = element_text(size = 11, face = "bold"))

# Create a grid of all histograms
histogram_grid <- plot_grid(hist_age, hist_bp, hist_hr, hist_cholesterol, hist_oldpeak, nrow = 2)

# Display the grid of histograms
print(histogram_grid)
```