Alexandria University
Faculty of Computer and Data Science
Department : Data Science
Course Title: Data Science 2023-2024

Heart attack analysis

In

Introduction to Data Science

Course code: 02-24-00104

Members Names and Role

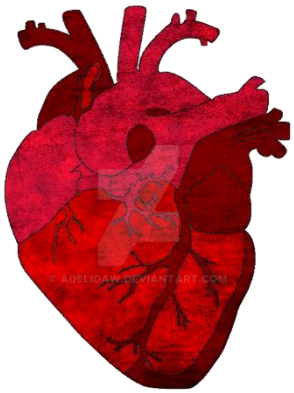
Name	ID	Role
Ghada Essam Khalifa	22010179	Writing the report and choosing suitable techniques
Ommeya Muhammad Abdelhady	22010323	Analysis and visualization Code on R
Jomana Esmat Anwar	22010327	UI design
Aya Muhammad Ahmad	22010065	UI design
Mennatullah moemen madani	22012051	Clustering and classification code and visualization
Shahd Ragab Saeed	22010352	UI design

1.Introduction:

In our project we chose the heart attack analysis and prediction data set.

It consists of the data of **303** patients containing **14** features:

- age - sex
- sum of their health measures : : trtbps (resting blood pressure), fbs(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false), chol (cholesterol)
- cp (chest pain type)
  Value 1: typical angina Value 2: atypical angina
  Value 3: non-anginal pain Value 4: asymptomatic
- restecg (resting electrocardiographic results)
  Value 0: normal
  Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- thalach : maximum heart rate achieved.
- thall: Thalassemia (a blood disorder)
  Value 0: NULL (dropped from the dataset in the cleaning step)
  Value 1: fixed defect (no blood flow in some part of the heart)
  Value 2: normal blood flow.
  Value 3: reversible defect (a blood flow is observed but it is not normal)
- oldpeak: ST depression induced by exercise relative to rest
- slp: the slope of the peak exercise ST segment
  0 = unsloping 1 = flat 2 = downsloping



and at the end is mentioned whether this patient has more chance of heart attack or less chance.

We analyzed this data to find out some patterns and correlations of the common symptoms of patients which have more chance of heart attack and extract information from this data to help people avoid getting such heart attacks.

2.Methodologies used:

At first we started by data preparation:

1- Data cleaning: we checked for the duplicated values and null values and removed them

```
> #checking for duplicates
> sum(duplicated(dataset))
[1] 1
> #cleaning our datasets from the duplicated value
> data <- distinct(dataset)
```

Here we found one duplicated row and removed it

```
> #checking for missing values
> sum(is.na(data))
[1] 0
```

Then we searched for the null values and we didn't find any

2- We checked the data types of all the columns

```
> #checking the data structure of our data
> all_columns_numeric <- sapply(data, function(col) all(sapply(col, i
s, "numeric")))
> all_columns_numeric
     age      sex       cp   trtbps     chol      fbs  restecg
    TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE
thalachh     exng  oldpeak      slp      caa    thall   output
    TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE
```

The data types of all the columns in the dataset is numeric

3- We separated the data of the patients (has more chance of a heart attack) and the non-patients(has less chance)

```
> patients <- subset(data, output == 1)
> non_patients <- subset(data, output == 0)
```

non_patients	138 obs. of 14 variables
patients	164 obs. of 14 variables

Now we have two subsets for patients and non- patients so we can find the characteristics for each.

Then we started to analyze each health measure for the patients and the non-patients using `summary()` method to find the mean , the maximum value, and the minimum value along with visualizing, using the suitable plotted graph according to the type of data we are analyzing to have a more obvious understanding .

After that we started to use the clustering and classification techniques to sum up the whole idea.

- We used **the k-means clustering** as an unsupervised clustering technique:
  As the k-means technique groups the unlabeled dataset into different clusters, it's also easy to implement, generalizes to clusters of different shapes and sizes so we used it to identify groups of patients with similar characteristics so we can find out the different cases in which the heart attack can happen.

- And we used **the decision tree** as a supervised classification technique:
  As it is relatively easy to interpret and understand, they are robust to outliers and can deal with missing values and it best suites our data set as it can help us to build a model on which we can classify future data and predict whether an individual is at risk of experiencing a heart attack within a certain time frame.

3.Challenges in the dataset:

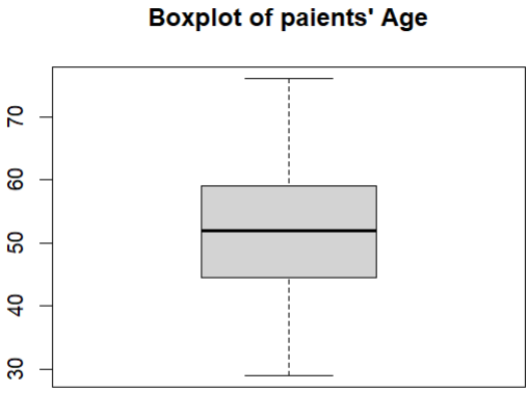
- There were many features for every patient in the dataset it was sometimes hard to decide which features we have to work on.
- The data set has small number of rows, so the results was quite not realistic.
- We weren't familiar with the domain of the dataset (medicine) and some expressions weren't easy to understand so we had to do some search about it.

4.Interpretations of the results:

So here let's see what we came up with in analyzing each feature:

1st: Age

```
> age_summary <- summary(data$age)
> patients_age_summary <- summary(patients$age)
> nonpatients_age_summary <- summary(non_patients$age)
> age_summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  29.00   48.00   55.50   54.42   61.00   77.00
> patients_age_summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  29.00   44.75   52.00   52.59   59.00   76.00
> nonpatients_age_summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   35.0    52.0    58.0    56.6    62.0    77.0
```



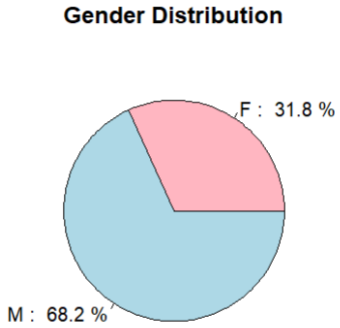
We used the boxplot in the age to find the range of most of the patients which is from nearly **45** to **59** the min. age of patients is **29** and the max is **76.**

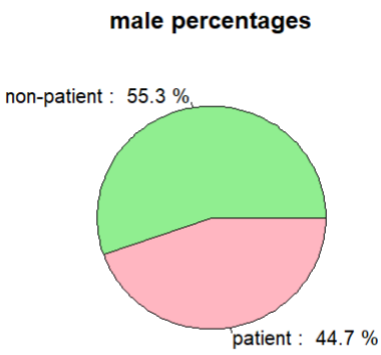
2nd : sex

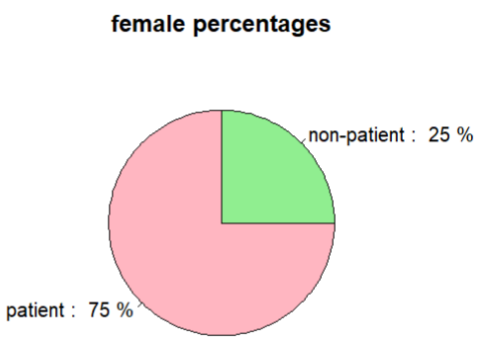
In the data we found that we have **96** female and **206** male

```
> sex_counts <- table(data$sex)
> sex_counts

  0   1
 96 206
```

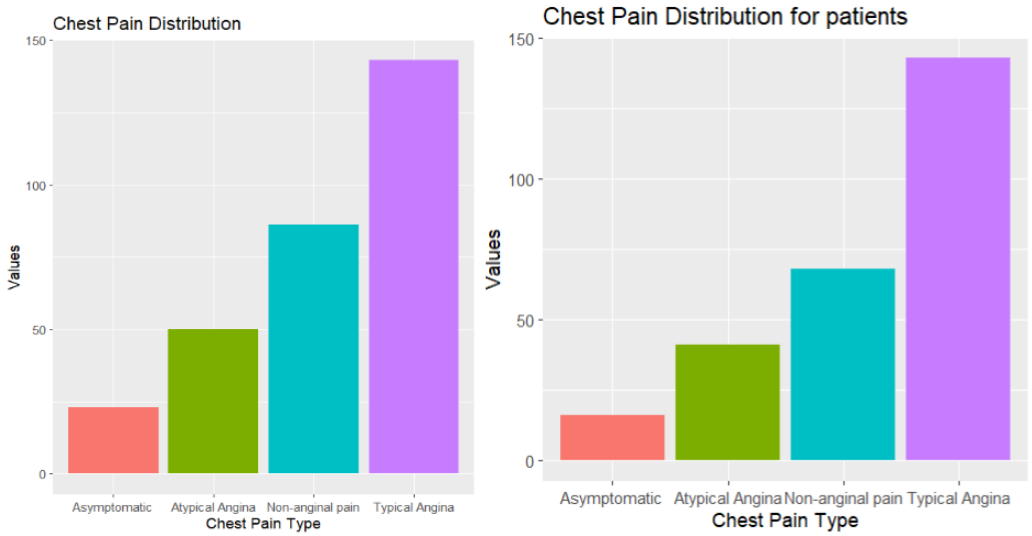






We used the the pie chart to get percentages and we found out that **75%** of the females in the data has more chance of heart attack and only **44.7%** of the males have more chance of a heart attack which means that females are more likely to get a heart attack than males

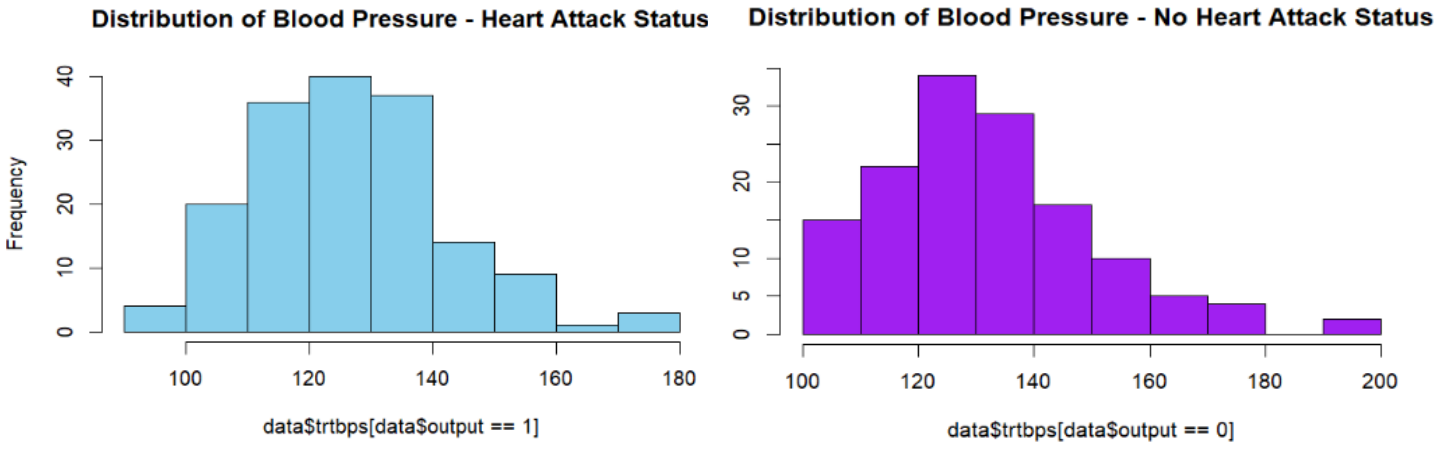
3rd:cp (chest pain type)



Here in the chest pain type we have four categories so we used the bar chart to compare between th

We noticed how the "typical angina" type of chest pain became more dominant among the patients and the other type decreased

4th:trtbps (resting blood pressure)



Here we used histogram to recognize the difference in values

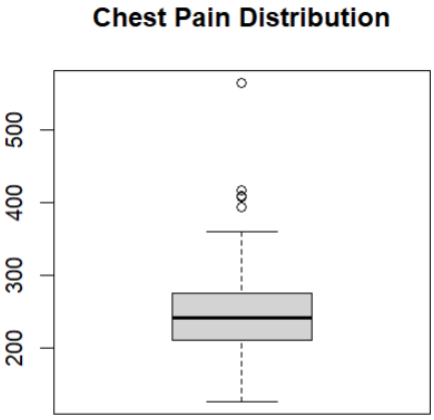
We can see how most of patients have relatively high blood pressure ranges from **110** to **140** mm Hg.

5th: chol (cholestrol)

First we used box plot to find outliers

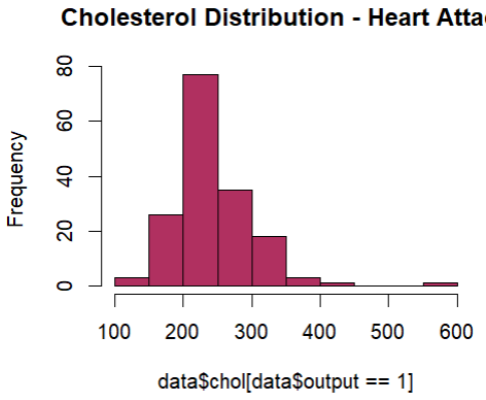
- We found five outliers

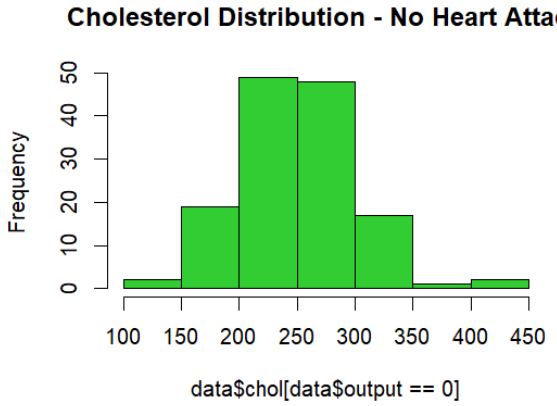
```
> chol_outlier <- boxplot(data$chol, main = "Chest Pain Distribution")$out
> chol_outlier
[1] 417 564 394 407 409
```



Then we also used histogram to compare The distribution of cholestrol values among patients and non-patients.

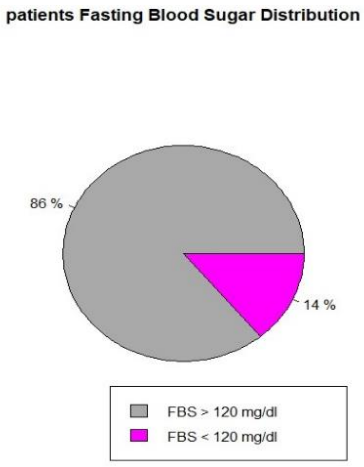
In the patients plot most patients (nearly 80 patient) have ranges from **200** to **250** mg/dl and in the non-paients plot they are mostly condensed from **200** to **300** mg/dl





6th: fbs (fasting blood sugar)

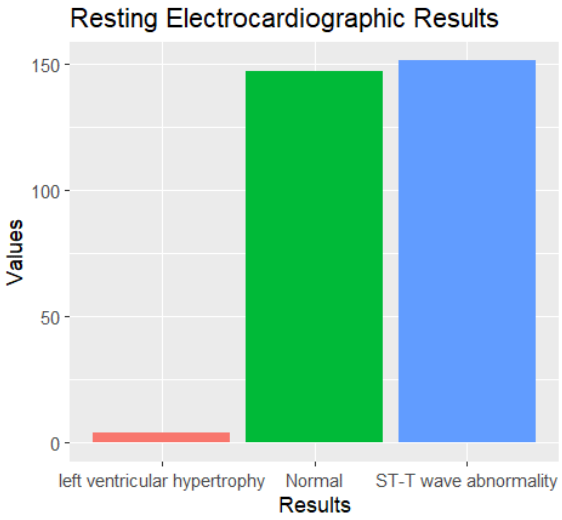
- We had two values in this column either 0 (fasting blood sugar<120) or 1 (fasting blood sugar>120) so we used the pie chart as they are only two vales they can be better viewed
- Here we see **86%** of the patients have FBS>120 mg\dl



7th : rstecg (Resting Electrocardiographic Results)

We used here the bar chart as we have multi categories

- There is only few who had “left venticular hypertrophy”
- The “ST-T wave abnormality” has the highest no. of patients and next comes the “normal” result there’s only alittle difference between them.

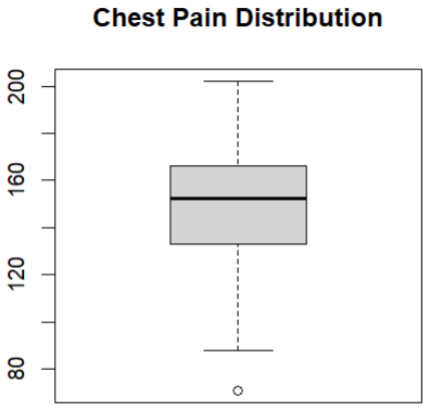


8th: thalach (max heart rate achieved)

We used boxplot to find outliers

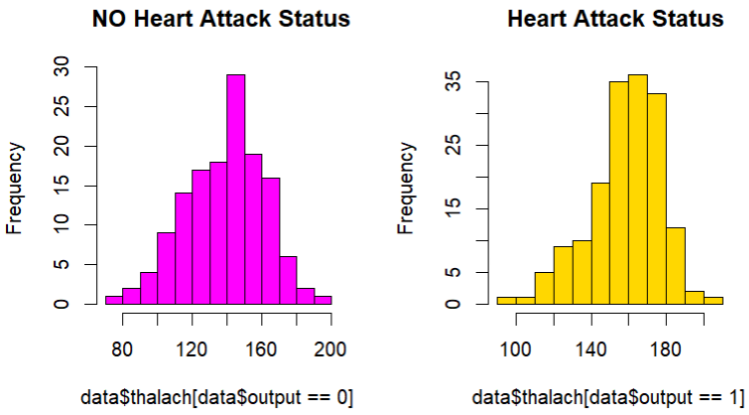
- we found one outlier equals 71

```
> thalachh_outlier <- boxplot(data$thalachh, main = "Maximum Heart
Rate Distribution")$out
> thalachh_outlier
[1] 71
```

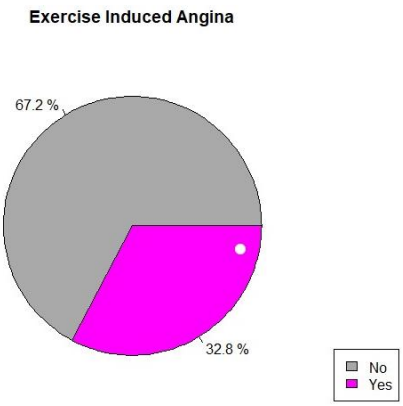


Then we used histogram to compare the distribution among the patients and the non- patients

- Most of the patients have max heart rate ranging from **150** to **180**
- Most of the non-patients have max heart rate ranging from **110** to **170** and about 30 of them have **150**



9th : exng (exercise induced angina)



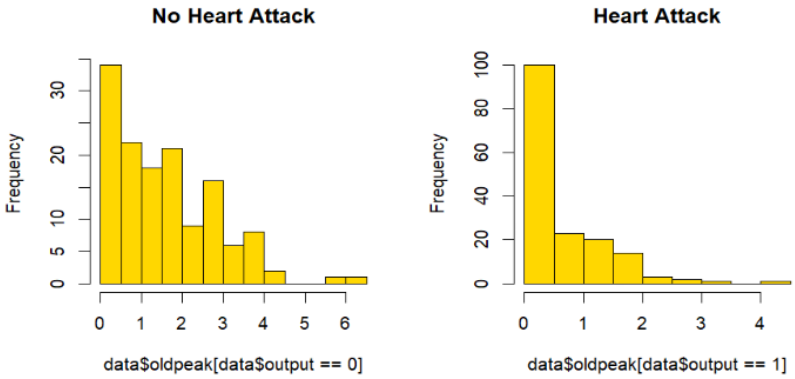
We only had two values for YES and NO so we used the pie chart.

- We fount that **67.2%** of individuals had exercise induced angina.

10th: old peak

Here we used histogram to compare the difference of the distribution of old peak among patients and non-patients.

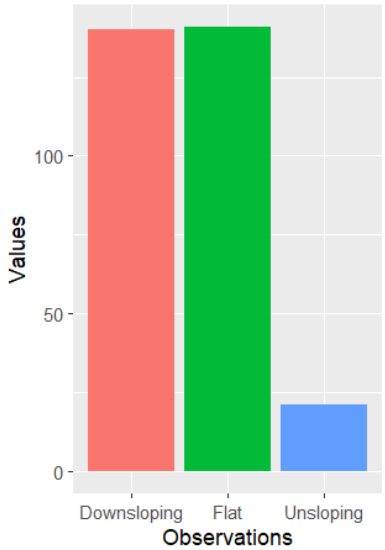
- We see that the old peak in case of heart attack is mostly zero and in the case of no heart attack it kinda varies from zero to 2 but still zero has the max value.



11th: slp (the slope of the peak exercise ST segment)

We used bar chart to compare the three types of slopes we had

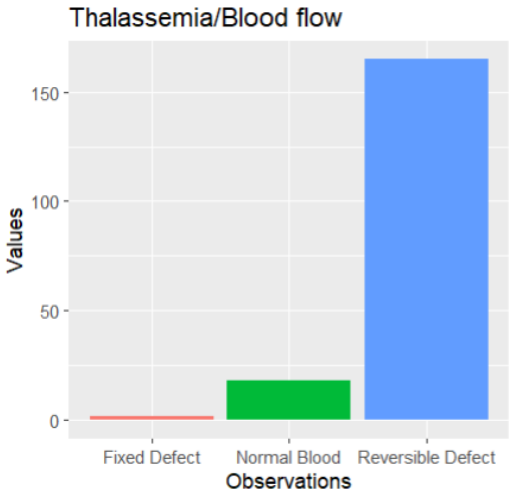
- We found that “down sloping” and “flat” have the maximum value, there’s a slight difference that “flat” is larger
- “unsloping ” has the min value



12th: thall (thalassemia: a blood disorder)

We have three categories here, so we choose the histogram .

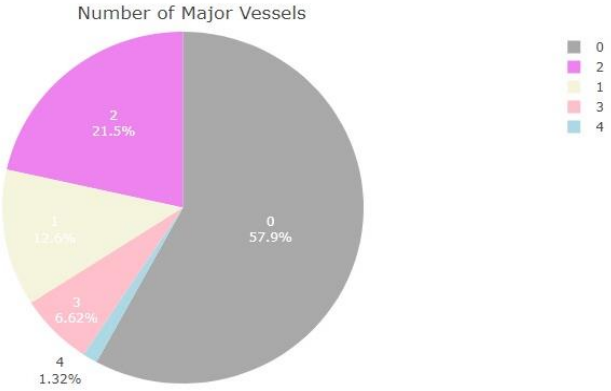
- Almost zero patients have fixed defect and only a few have normal blood most of them have reversible defect as seen .



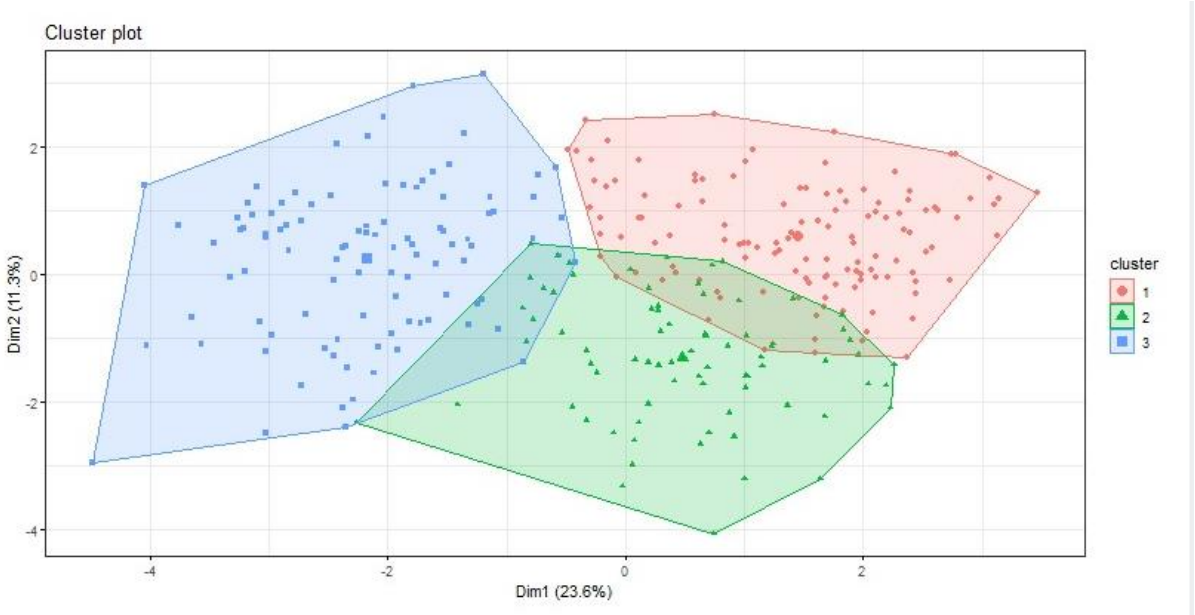
13th: caa (number of major vessels)

We used here pie chart to compare the percentages.

- The most common number of major vessels between patients is **0** while the least common number is 4.

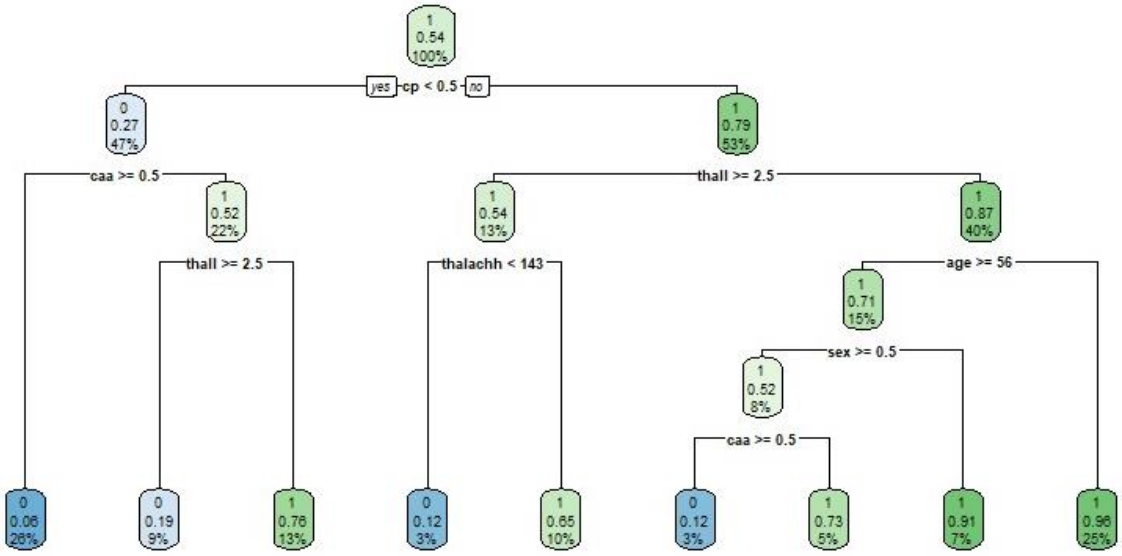


K-means clustering technique result:



- K-means helped us cluster the inviduals into three groups.
- Each group consists of individuals with similar characteristics and symptoms.
- Each point represents a patient, and the colors denote the assigned cluster. The segmentation into clusters aims to reveal patterns in symptomatology, allowing us to discern potential similarities or differences among patients. This analysis aids in identifying groups exhibiting similar symptom profiles, which could assist in understanding diverse risk factors or medical conditions contributing to heart attacks.

Decision tree classification result:



Here with the decision tree it's so easy to classify the state of any new individual

We can see that 26% of the patients has **less chance of heart attack** as they have cp = 0 (typical angina) and caa > 0 .

And 25% of the patients has **more chance of heart** attack as they have cp >0 (not typical angina) , thall < 3 (not reversible defect) and age < 56 .

- We tried to predict new data for 3 new individuals

According to the decision tree we found out that the first two patients has more chance of heart attack and the last one has less chance of a heart attack .

```
> new_data <- data.frame(
+   age = c(45, 55, 65),
+   sex = c(1, 0, 1),
+   cp = c(1, 2, 3),
+   trtbps = c(120, 140, 160),
+   chol = c(200, 220, 240),
+   fbs = c(0, 1, 0),
+   restecg = c(0, 1, 0),
+   thalachh = c(150, 160, 170),
+   exng = c(0, 1, 0),
+   oldpeak = c(0.5, 1.0, 1.5),
+   slp = c(1, 2, 3),
+   thall = c(2, 3, 1),
+   caa = c(0, 1, 2)
+ )
> predictions <- predict(tree_model, newdata = new_data,
type = "class")
> print(predictions)
1 2 3
1 1 0
Levels: 0 1
```

5.Conclusion:

After all let's see what we've came up with so far about this data set:

- Age of patients varies from **45** to **59.**
- females have more chance of a heart attack than males as **75%** of females are patients whilst **44.7%** of males are patients.
- **"Typical angina"** chest pain type is the most common among patients.
- Patients have relatively high resting blood pressure, mostly ranges from **110** to **140** mm Hg.
- Cholesterol of patients mostly ranges from **200** to **250** mg\dl whilst the non-patients are mostly condensed from **200** to **300** mg\dl.
- **86%** of patients have FBS > 120 mg\dl .
- In resting electrocardiographic results, most of patients have either "ST-T wave abnormality" or "normal" results.
- Patients have max heart rate ranging from **150** to **180**, while non-patients have max heart rate ranging from **110** to **170** and high percentage of them have **150.**
- **67.2%** of individuals had exercise induced angina
- Most patients have old peak equals to **Zero**.
- The slope of peak exercise ST segment of individuals is mostly either **"down sloping"** or **"flat"**.
- And for thalassemia most individuals have **"reversible defect"**.
- The most common no. of major vessels among individuals is **0** and the least is **4**.