

**School of Computer Science**

**Masters in Applied Computing (M.A.C)**

**Subject Code: COMP8157**

**Subject Name: Advanced Database Topics**

**Professor Dr. Shafaq Khan**

**Lab Assignment**

**by**

**Hamza Baig (110089314)**

Table of Contents

[Data Exploration: 3](#_Toc125660447)

[1. Import Dataset: 3](#_Toc125660448)

[2. Summary of dataset: 3](#_Toc125660449)

[3. Structure and Dimension of dataset: 4](#_Toc125660450)

[Structure: 4](#_Toc125660451)

[Dimension: 4](#_Toc125660452)

[4. First 8 rows and last 5 rows: 4](#_Toc125660453)

[Head: 4](#_Toc125660454)

[Tail: 5](#_Toc125660455)

[5. Column names: 5](#_Toc125660456)

[Data Pre-Processing: 5](#_Toc125660457)

[6. Class variable: 5](#_Toc125660458)

[7. Data type of class variable: 5](#_Toc125660459)

[8. Change the class type of class variable: 5](#_Toc125660460)

[9. Sum of missing values: 5](#_Toc125660461)

[10. Show columns with missing values: 5](#_Toc125660462)

[11. Replace missing values with 0: 6](#_Toc125660463)

[12. Rename sex attributes: 6](#_Toc125660464)

[Data Visualization: 6](#_Toc125660465)

[13. Scatter plot: 6](#_Toc125660466)

[A: 6](#_Toc125660467)

[B: 7](#_Toc125660468)

[14. Ggplot: 7](#_Toc125660469)

[15. Bar plot for age: 8](#_Toc125660470)

[16. Histogram for chest pain: 8](#_Toc125660471)

[17. Boxplot for age: 9](#_Toc125660472)

[18. Correlation plot: 9](#_Toc125660473)

[Code: 11](#_Toc125660474)

This file contains the outputs as well as the code for the R file.

# Data Exploration:

## Import Dataset:

Logo

Description automatically generated

## Summary of dataset:

It gives information about all the columns/variables such as min. value of each column, max. value, mean, median, etc.

Calendar

Description automatically generated with medium confidence

## Structure and Dimension of dataset:

### Structure:

It explains the structure of data; the number of columns/variables in data and the number of observations with data types with possible values.

Text

Description automatically generated

### Dimension:

Gives the dimension of data; number of rows/observations and columns/variables.



## First 8 rows and last 5 rows:

### Head:

Table

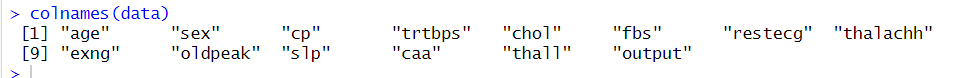
Description automatically generated

### Tail:

Word

Description automatically generated with medium confidence

## Column names:



# Data Pre-Processing:

## Class variable:

The class variable in the dataset will be the output variable/column, it shows whether the patient can have heart attack or not.

class\_variable <- data$output

## Data type of class variable:



## Change the class type of class variable:

A picture containing table

Description automatically generated

## Sum of missing values:

Text

Description automatically generated

## Show columns with missing values:

Graphical user interface, application, Word

Description automatically generated

## Replace missing values with 0:

Text

Description automatically generated with medium confidence

## Rename sex attributes:

Text

Description automatically generated with low confidence

# Data Visualization:

## Scatter plot:

### A:

Chart, scatter chart

Description automatically generated

### B:

Chart, scatter chart

Description automatically generated

## Ggplot:

Chart, scatter chart

Description automatically generated

## Bar plot for age:

Chart, histogram

Description automatically generated

## Histogram for chest pain:

Chart, histogram

Description automatically generated

## Boxplot for age:

Chart, box and whisker chart

Description automatically generated

It shows the median age being dispersed between 50-60 years in the dataset.

## Correlation plot:

Table

Description automatically generated

It shows the relationships between the variables/columns, which columns are closely related and contributing to the label/class variable.

# Code:

# get working directory

#getwd()

#import libraries

library(ggplot2)

# Data Exploration

# 1. import data

data <- read.csv("Heart\_Attack.csv")

# 2. summary of data

# It gives information about all the columns/variables such as min. value of each column, max. value, mean, median, etc.

summary(data)

# 3. structure of data

# It explains the structure of data; the number of columns/variables in data and the number of observations with data types with possible values.

str(data)

# 3. dimension of data

# Gives the dimension of data; number of rows/observations and columns/variables

dim(data)

# 4. First 8 rows

head(data, 8)

# 4. Last 5 rows

tail(data, 5)

# 5. Column names of dataset.

colnames(data)

# Data Pre-Processing

# There is no class variable/column in the provided dataset

# class\_var <- data$class

# Assuming you're asking about the variable that represents the label in the dataset, then it will be the output variable, and it represents whether a patient can have heart attack or not.

# 6. Class Variable is output.

class\_variable <- data$output

# 7. Datatype of output variable - integer

class(class\_variable)

# 8. Change data type of output to factor

data$class <- as.factor(data$output)

# 8. Show output

data$class

# 9. Sum of missing values

print("Sum of missing values:")

sum(is.na(data))

# 10. Sum of missing values for each column/variable

colSums(is.na(data))

# 11. Replace missing values

data[is.na(data)] <- 0

# 11. Check if values replaced - should return 0

sum(is.na(data))

# 12. Rename the sex attribute

data$sex <- ifelse(data$sex == 0, "Male", "Female")

# 12. Check if renamed (0 to Male and 1 to Female)

data$sex

# Data Visualization

# 13. scatter plot between cholestrol and age.

plot(data$chol, data$age, xlab = "Cholesterol", ylab = "Age", main = "Scatter Plot of Cholesterol and Age")

# 13. a

points(data$chol, data$age, col = "blue")

# 13. b

points(data$chol, data$age, col = "red",pch=6)

# 14. ggplot

# library imported above

# Plot between age and cholestrol, based on output.

ggplot(data, aes(x=age, y=chol,color=output)) + geom\_point(shape=15)

# 15. Barplot

barplot(table(data$age), main = "Age distribution", xlab = "Age", ylab = "Frequency", col = "blue")

# 16. Histogram for chest pain

# a. Minimum and maximum

minimum\_cp <- min(data$cp)

maximum\_cp <- max(data$cp)

# b and c - histogram

hist(data$cp, breaks = seq(minimum\_cp, maximum\_cp, 1), xlab="Chest pain",ylab="Frequency",main = "Chest Pain Type",col="red")

# 17. Boxplot for age

boxplot(data$age,main="Age distribution", xlab="Age",ylab="Frequency")

# 18. Correlation plot between all the outputs.

plot(data)