**American International University-Bangladesh**



**Course:** INTRODUCTION TO DATA SCIENCE

**Assignment Title:** Final-term Project

**Submitted by:**

Name: Sinthia Maliha Mumu  
Id: 20-42242-1  
Section: D

**Submitted to:**

TOHEDUL ISLAM

Assistant Professor, Computer Science  
American International University-Bangladesh

**Submission Date:** April 30, 2023

**Project Overview:**

**Chosen Dataset:**

[**https://www.kaggle.com/datasets/kingabzpro/heart-disease-patients**](https://www.kaggle.com/datasets/kingabzpro/heart-disease-patients)

For this project I have chosen a dataset hosted on Kaggle which contains information about patients with heart disease. The dataset includes various medical and demographic features such as age, sex, blood pressure, cholesterol levels, and whether the patient has heart disease.

This dataset table have the following columns:

* age - Age of patient
* sex - Gender of patient
* cp - chest pain type
* trestbps - Resting blood pressure (in mm Hg on admission to the hospital)
* chol - Serum cholesterol in mg/dl
* fbs - Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
* restecg - Resting electrocardiographic results.
* thalach - Maximum heart rate achieved.
* exang - Exercise induced angina (1 = yes; 0 = no)
* oldpeak - ST depression induced by exercise relative to rest.
* slope - The slope of the peak exercise ST segment.

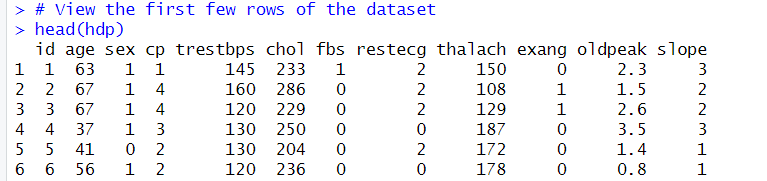
Before applying K-means clustering algorithm to the selected unsupervised data set we need to follow these steps.

1. **Loading the dataset**

hdp <- read.csv("C:/Users/HP/Downloads/heart\_disease\_patients.csv", header = TRUE)

1. **Exploring the dataset**

head(hdp)



summary(hdp)

A picture containing table

Description automatically generated

dim(hdp)



str(hdp)

Text

Description automatically generated

1. **Preprocessing the dataset**

heart\_data <- na.omit(hdp)

heart\_data <- hdp[, -1]

heart\_data

Table

Description automatically generated

hdp\_subset <- hdp[,c(2,5,6,9)]

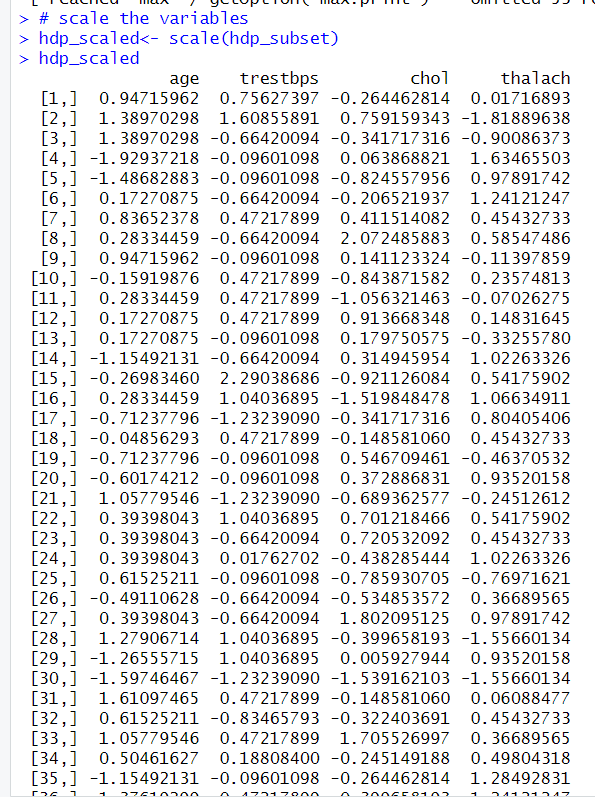
hdp\_subset

Table

Description automatically generated

hdp\_scaled<- scale(hdp\_subset)

hdp\_scaled



Text, letter

Description automatically generated

1. **Estimating optimal number of clusters**

library(factoextra)

fviz\_nbclust(hdp\_scaled, kmeans, method = "wss")

Chart, line chart

Description automatically generated

1. **K-means clustering algorithm to the data**

kmeans\_model <- kmeans(hdp\_scaled, centers = 3, nstart = 25)

kmeans\_model

A picture containing calendar

Description automatically generated

1. **Visualizing the clusters**

fviz\_cluster(kmeans\_model, data = hdp\_scaled )

Chart, map, scatter chart

Description automatically generated

From the visualization, we can see that there are 3 distinct clusters. We will analyze each cluster further to understand their characteristics.

1. **Analyzing the clusters**

library(ggplot2)

library(gridExtra)

p1 <- ggplot(heart\_clusters, aes\_string(x = "age", y = "trestbps", color = "cluster")) +

geom\_point() +

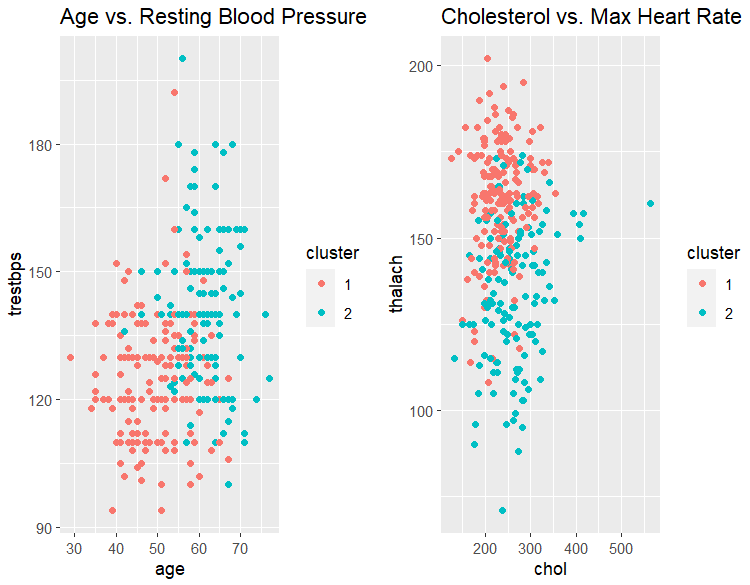
labs(title = "Age vs. Resting Blood Pressure")

p2 <- ggplot(heart\_clusters, aes\_string(x = "chol", y = "thalach", color = "cluster")) +

geom\_point() +

labs(title = "Cholesterol vs. Max Heart Rate")

grid.arrange(p1, p2, ncol = 2)



1. **Visualizing the output of K-means Clustering Algorithm**

aggregate(hdp\_scaled, by = list(cluster = kmeans\_result$cluster), mean)

Text

Description automatically generated

From the analysis, we can conclude that the provided dataset can be effectively clustered into 3 distinct groups using K-means clustering. Each cluster has its own distinct characteristics, with Cluster 2 having relatively high values for all the features, Cluster 1 having relatively low values for almost all the features except thalach, and Cluster 3 having relatively high values for age and trestbps, and relatively low values for chol and thalach.