# K-means clustering

#import data

heartdata= read.csv("dataset/heart.csv")

#Exploring data with 0,1 values as they won't be of much help in making the clusters

heartdata\_c = heartdata[,-which(colnames(heartdata) %in% c("sex","fbs","exng","output"))]

# Standardize the dataset and apply K-means clustering

heartdata\_scale = scale(heartdata\_c)

summary(heartdata\_scale)

#calculate for 2 clusters eventually (1 : a significant heart Disease, 0 : not a significant heart disease)

km <- kmeans(heartdata\_scale, centers = 2, nstart = 10)

km

heartdata\_clustered <- as\_tibble(heartdata\_scale) %>% add\_column(cluster = factor(km$cluster))

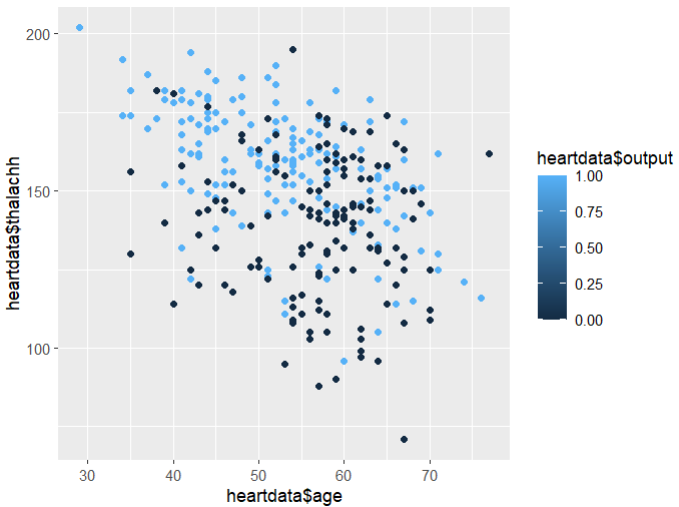
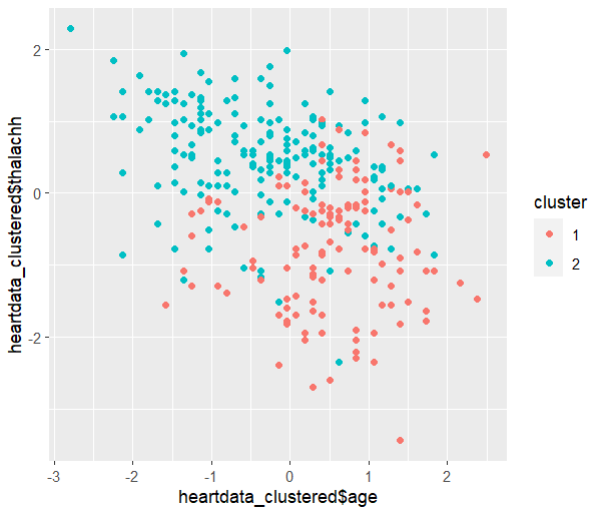
heartdata\_clustered

#plot the clustering result

ggplot(heartdata\_clustered, aes(x = heartdata\_clustered$age, y = heartdata\_clustered$thalachh, color = cluster)) + geom\_point()

#actual output

ggplot(heartdata, aes(x = heartdata$age, y = heartdata$thalachh, color = heartdata$output)) + geom\_point()



#calculates the accuracy of the prediction

accuracy = ifelse(heartdata\_clustered$cluster == heartdata$output+1, 'Y','N')

acc\_rate = sum(accuracy == 'Y')/length(accuracy)

acc\_rate 0.7821