Eleana Cabello
DSA 5103

Homework 8

Description of Heart Disease Dataset

The heart disease dataset was collected from patients seen at Cleveland Clinic Foundation. It consists of 14 attributes and 303 observations. The attributes are related to several physiological conditions of patients suffering from heart disease and healthy patients. The observations are classified into four classes based on the extent of their heart disease condition. All attributes are numeric.

- age – age of patient in years
- sex – biological sex of patient (0 = female, 1 = male)
- cp – chest pain type (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
- trestbps – resting blood pressure in mmHg when admitted to hospital
- chol – serum cholesterol in mg/dl
- fbs – fasting blood sugar above 120 mg/dl (0 = false, 1 = true)
- restecg – resting electrocardiographic results (0= normal, 1=having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), 2= showing probable definite left ventricular hypertrophy by Estes' criteria)
- thalach – max heart rate achieved
- exang – exercise induced angina (0 = no, 1= yes)
- oldpeak – ST depression induced by exercise relative to rest
- slope - the slope of the peak exercise ST segment (1=upsloping, 2=flat, 3=downsloping)
- ca - number of major vessels (0-3) colored by flourosopy
- thal - 3 = normal, 6 = fixed defect, 7 = reversable defect
- num - diagnosis of heart disease (angiographic disease status) (0=< 50% diameter narrowing, 1= > 50% diameter narrowing (in any major vessel: attributes 59 through 68 are vessels)

Website Name: UC Irvine Machine Learning Repository
URL: https://archive-beta.ics.uci.edu/dataset/45/heart+disease

K Value Selection for K-Means

No obvious elbow was found in the within sum square error elbow plot below.

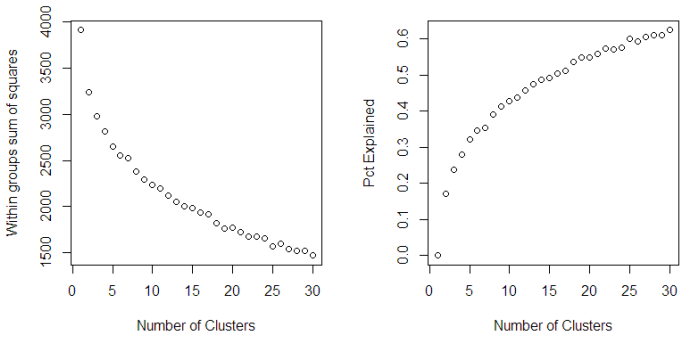

Figure 1 - Within cluster sum of square error elbow plot.

From Hartigan's rule results below, k = 6 seems to be the optimal choice.

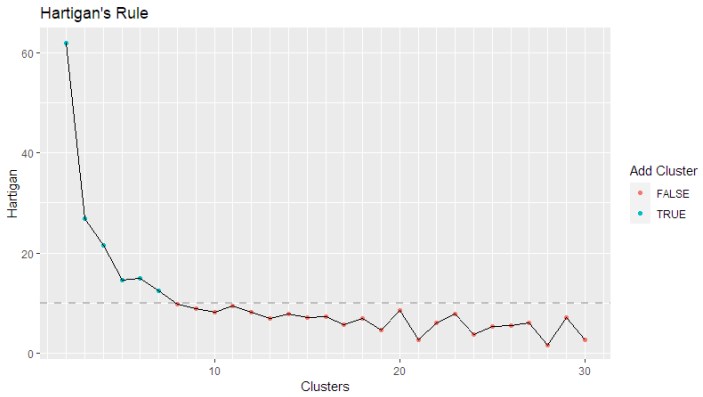

Figure 2 - Hartigan's rule results plot.

From nbclust results below, k = 5 could be a good choice as well.

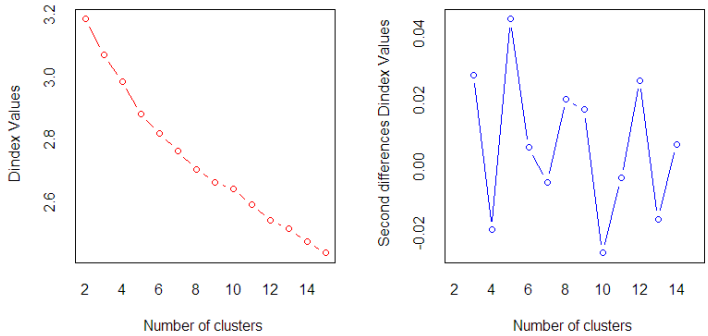

Figure 3 - NbClust results.

Clustering Results

Method #1 – K-Means

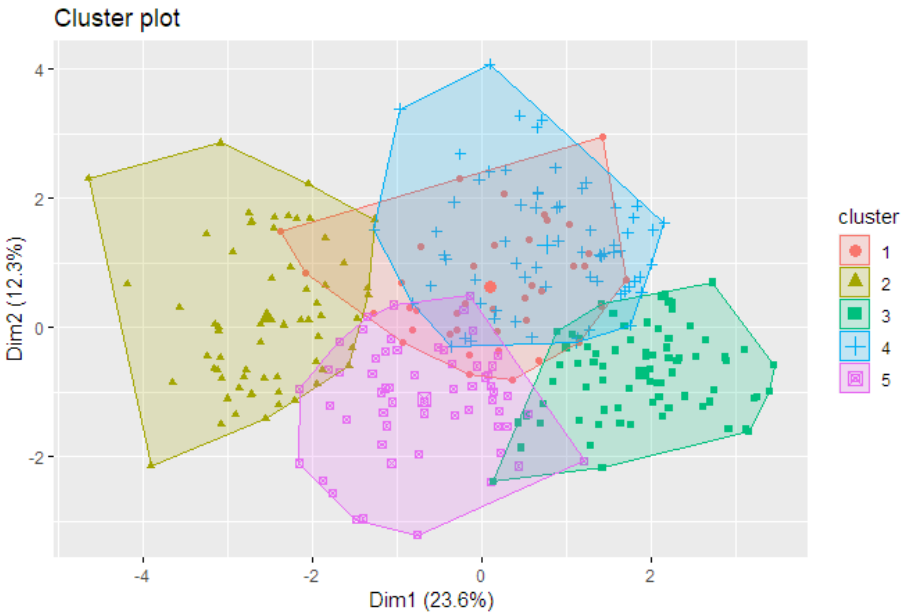

Figure 4 - Kmeans results for k = 5.

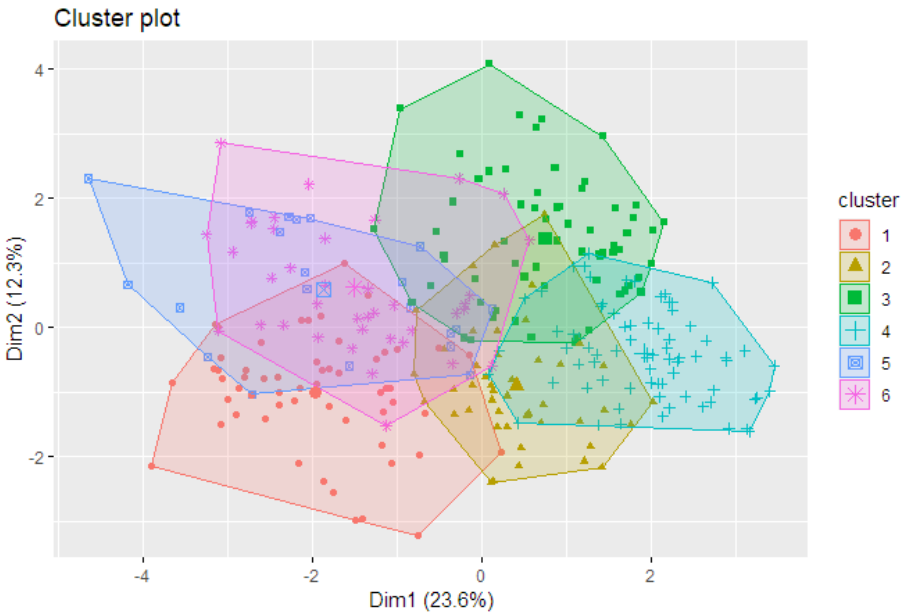

Figure 5 - Kmeans results for k = 6.

Method #2 – Hierarchical Clustering

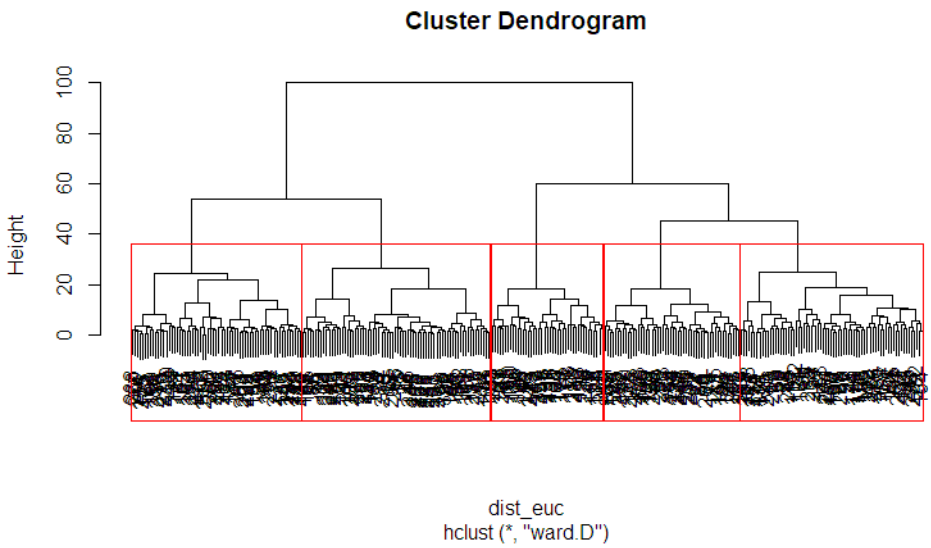

Figure 6 - Cluster Dendrogram using euclidean distances.

Method #3 – dbscan

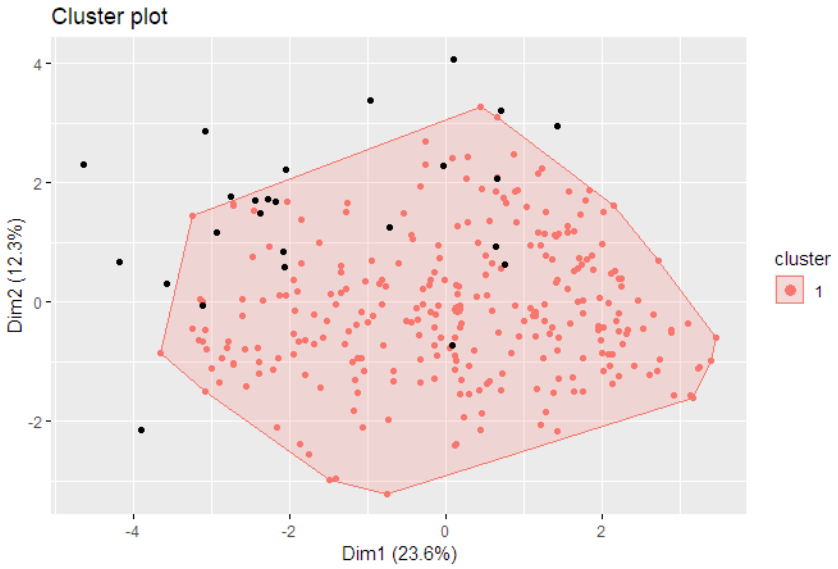

Figure 7 - Dbscan results.

Based on the description of the data set, patient records can be classified under five possible classes of differing degrees of heart disease. The visual clustering results above showed that the k-means and hierarchical clustering methods performed far better than the dbscan method. When using methods k-means with a k = 5 and hierarchical clustering, five clustering groups could be seen which aligns with the five classes possible. The dbscan method could only identify one cluster.

K-Means Interpretation

Table 1 - Centroids of clusters.

	AGE	SEX	CP	TRESTBPS	CHOL	FBS	RESTECG	THALACH	EXANG	OLDPEAK	SLOPE	CA	THAL	KMEANS_RESULTS.SIZE
1	0.088	0.419	0.744	-0.062	-0.018	-0.059	0.013	-0.983	1.263	0.567	0.660	-0.062	0.617	64
2	0.116	0.647	-1.009	0.315	-0.248	0.217	0.199	0.275	-0.579	0.100	0.388	-0.335	0.201	54
3	0.452	-1.387	-0.078	0.009	0.483	-0.112	0.029	0.090	-0.375	-0.433	-0.305	-0.272	-0.805	66
4	-0.972	0.282	-0.230	-0.458	-0.355	-0.298	-0.503	0.788	-0.467	-0.658	-0.819	-0.529	-0.405	74
5	0.686	0.201	0.661	0.461	0.201	0.488	0.539	-0.369	0.221	0.808	0.400	1.799	0.745	44

- **Cluster 1** – fasting blood sugar far below 120, low max heart rate, downsloping in ST segment
- **Cluster 2** – likely not exercise induced angina, low ST depression, high BP
- **Cluster 3** – high cholesterol, fasting blood sugar below 120
- **Cluster 4** – youngest group, lowest BP, low cholesterol, resting ecg normal, high max heart rate, upsloping in ST segment, lowest number of major blood vessels
- **Cluster 5** – oldest group, high BP, fasting blood sugar above 120, resting ecg far from normal showing possible signs of left ventricular hypertrophy, high ST depression , high number of major blood vessels

Based on the conclusions drawn above from the centroids of the clusters, cluster 4 is likely the healthy patient records, or class 0, while clusters 1,2,3 and 5 are those suffering from heart disease. Cluster 5 is those suffering from severe heart disease, or class 4. Cluster 2 results indicate higher blood pressure (BP) and more abnormal ECG results than clusters 1 and 3, meaning it is likely class 3. Based on similar trends, cluster 3 had higher bp results and more abnormalities in ECGs than cluster 1, meaning clusters 3 and 1 are likely classes 2 and 1. However, it's important to note there may be severe overlapping between clusters 4 and 1 and their corresponding classes because early stages of heart disease are harder to recognize and can appear normal. This overlapping between clusters 4 and 1 is also evident in Fig. 4.