Data Set- Heart disease

The dataset used in this assignment is the **heart disease** dataset available in ***heart-c.csv*** from the **Blackboard**. This dataset describes 13 risk factors for heart disease. The attribute ***num*** represents the (binary) class attribute: class <50 means no disease; class >50_1 indicates increased level of heart disease.

age = *age in years*
sex = *sex (1 = male; 0 = female)*
cp = *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 mm Hg on admission to the hospital)*
chol = *serum cholestoral in mg/dl*
fbs = *(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)*
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*
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 :*
- ✓ *Value 1: upsloping*
- ✓ *Value 2: flat*
- ✓ *Value 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) :*
- ✓ *Value 0: < 50% no disease*
- ✓ *Value 1: > 50% increased level of heart disease*

1. Run K-means clustering on the above heart disease dataset and answer the following questions

1) Why should the attribute "*class*" in ***heart-c.csv*** ("***num***") **not** be included for clustering?
2) Run K-means algorithm by choosing different numbers of clusters, *numCluster* = 2, 3, 4,5, then observe the differences of clusters generated:
   a. How are the *Within Cluster Sum of Squared Errors*[1] changed for different numbers of clusters?
   b. What can you conclude?
   c. How can you explain this conclusion from clustering analysis point of view?

[1] "***Intra-cluster sum of squared errors***", or "***intra-cluster distance***"

2. Run the hierarchical clustering on above heart disease dataset, and answer the following questions

1) Show the clustering results in tree structure;
2) Describe the link method you used;
3) What are the strengths and limitations of this link method in hierarchical clustering?