**Clustering Activity:**

On the inbuild ‘mtcars’ data set, we will be clustering the similar cars based on different features using K-means and Hierarchial clustering.

**R Code:**

1. Load inbuild ‘mtcars’ data available in R

2. Understand the data and apply the necessary pre-processing steps.

3. Normalize/Scale the data.

Note: Identify the cluster performance with and without normalizing/scaling the data and identify the importance of the scaling the data.

**#Hierarchical Clustering Activity:**

1. Calculate the distance between different cars using “dist” function using different distance methods.

d <- dist(mydata, method = "euclidean") # distance matrix

d

Note: Experiment with different distance methods.

2. Build the hierarchical clustering using “hclust” function using agglomerative method ward.D2

fit <- hclust(d, method="ward.D2")

Note: You can explore different methods single, complete, average

3. Visualize the clusters. Tree like structure is called as dendrogram.

plot(fit)

# dendrogram displays all possible clusters from the data in bottom up approach

4. Creating 5 clusters using cutree function, “K” specifies number of cluster to create.

groups <- cutree(fit, k=5) # cut tree into 5 clusters

groups

# draw dendogram with red borders around the 5 clusters

rect.hclust(fit, k=5, border="red")

5. Append cluster labels to the actual data frame

Mydata\_cluster <- data.frame(mydata, groups)

**K-means clustering:**

6. Build the cluster using kmeans function by mentioning the number of clusters.

# K-means clustering

fit<- kmeans(mydata,centers=2)

fit

7. Check sum of Inter cluster distance(betweenness) and Intra cluster distances(With-in sum of squares).

fit$withinss

sum(fit$withinss)

#Cluster Centers

fit$centers

#To check cluster number of each row in data

fit$cluster

8. Identifying the ideal number of cluster:

• Write a for loop which should start with 2 clusters and build k-means model up to 15 clusters.

• Capture the within-sum of squares for different number of cluster, save sum(fit$withinss) for each model.

• Plot sum(fit$withinss) generated in all models

• Find the best cluster based on the curve.

**Exercise: Cereals data: Identify similar cereals using K-means clustering**

**Cereals data:** Data consists of the information of proteins, calories, vitamins, carbohydrates, minerals etc. for different cereals. Using K-means technique identify/cluster the similar cereals.

• Load the cereals data into R.

• Analyze the data and apply the required pre-processing steps and prepare data for clustering.

• Use a distance metric to compute distance matrix.

• Apply k-means clustering technique, identify the ideal number of cluster.

• Identify the similar cereals based on the clusters.