Learning outcomes:

After completing this exercise, you should be able to understand and perform below tasks.

- Applying K-means Clustering, Hierarchical clustering & PCA.
- Understand various cluster metrics generated in R.
- Evaluating the performance of clusters.
- Understanding the importance of standardizing data.
- Visualization and interpretation of results.

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.

 # PCA

We will perform the lab activity with two data sets. One with built-in data set in R and other one in csv format.

1. Load data 'attitude' available in R. To load the data.
   data(attitude)
   names(attitude)
   summary(attitude)
   str(attitude)
   attach(attitude)
2. Understand the summary of data and attributes available in data
3. Construct a linear regression model using the attribute 'rating' as dependent variable
4. Create new data set with independent variables separately and normalize the data.
5. Apply Principle Component Analysis on the data set created in the step 4
   pca_data <- princomp(<datasetname>)
6. Understand the outputs after applying PCA on data. Identify the components explaining more variance of data.
   summary(pca_data)
   plot(pca_data)
   print(pca_data)

7. Print the values of each component
   pca_data$scores

8. Create new data set with the dependent variable in 'attitude' data set and with the components explaining more variance.
   data2<-data.frame(rating,pca_data$scores[,1:4])

9. Build linear regression using dataset created in step 8.

Exercise-1: 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.

Exercise-2: Cereals data: Identify similar cereals using PCA and K-means

- Load the 'Cereals.csv' data into R
- Analyze the data and apply the required pre-processing and prepare data.
- Normalize/Scale the data.
- Apply principle component analysis on all numeric variables
- Understand the results and identify the components explaining more variability in the data K-means clustering
- Use the principle components data created on 'pca_data$scores', apply k-means clustering
- Build k-means clustering with the principle components which are important and capture the similar cereals.