Learning outcomes:

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

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

Objective:

Can we find cars that are similar on various aspects?

Clustering Activity:

On the inbuild 'mtcars' data set, we will be clustering the similar cars based on different features using k-means and hierarchical clustering.

R Code:

- Load inbuild 'mtcars' data available in R
- Understand the data and apply the necessary pre-processing steps.
- 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:

- 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.
- 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
- Visualize the clusters. Tree like structure is called as dendrogram.
  ```
  plot(fit)
  ```
  # dendrogram displays all possible clusters from the data in bottom up approach
- 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")
  ```
- Append cluster labels to the actual data frame

```
Mydata_cluster <- data.frame(mydata, groups)
```

K-means clustering:

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

```
# K-means clustering
fit<-kmeans(mydata,centers=2)
fit
```

- 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
```

Identifying the ideal number of clusters

- 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.
- Using factoextra library

On Unseen data

For new data, how do you know which cluster does it belong to?

Let's select one record randomly from the data. Given the cluster centres we can find the distance between the new point and cluster centre. The one with small distance will be the cluster number assigned to the new data.

```
test_datapoint <- mtcars[sample(1:nrow(mtcars),1),]
closest.cluster <- function(x) {
    cluster.dist <- apply(fit$centers, 1, function(y) sqrt(sum((x-y)^2)))
    print(cluster.dist)
    return(which.min(cluster.dist)[1])
}

closest.cluster(test_datapoint)
```

Performance

- Quality check: For each observation i, the *silhouette width* s(i) is computed; Observations with a large s(i) (almost 1) are very well clustered, a small s(i) (around 0) means that the observation lies between two clusters, and observations with a negative s(i) are probably placed in the wrong cluster.

```
library(cluster)
#gower_dist = daisy(x = mydata, metric = "gower")
distance_matrix = daisy(x = mydata, metric = "euclidean")
clust_assignment = mydata$fit.cluster
sil_value_hc_mixed = silhouette(clust_assignment, dist = distance_matrix)
plot(sil_value_hc_mixed)
```

- Stability of clusters is important. Ideally, we do not want the clusters to change when new data is added. The way to evaluate is take a 90 or 95% of the records and apply k-means algorithm with the same number of clusters as decided for complete data. Using the 'adj.rand.index' function we can see how much is the deviation in clustering.

```
## stability check
set.seed(12)
index <- (sample(nrow(mydata),.90*nrow(mydata)))
dataTest <- mydata[index,]
StabClus <- kmeans(dataTest,5)
dataTest$clusters <- StabClus$cluster

group1 <- mydata[index,12]
group2 <- dataTest$clusters
group <- cbind(group1, group2)
write.csv(group, "clusgroup.csv")

#install.packages("fossil")
library(fossil)
stabilitycheck <- adj.rand.index(group1, group2)
stabilitycheck
```

For more: https://davetang.org/muse/2017/09/21/adjusted-rand-index/

```
#install.packages("clusteval")
library(clusteval)
Stabindex <- cluster_similarity(group1, group2, similarity = "jaccard",
method="independence")
Stabindex
```

Dealing with mixture of attribute types (Good to know)

In all the above steps we have assumed that the data is numeric and have scaled them. This is mainly because we were applying 'Eucledian' distance measure. However, there is a distance metric 'gower' which can deal with mixture of attribute types.

```
mydata2<-mtcars
names(mydata2)
#categorical data
data1<-as.data.frame(apply(mydata2[,c(2,8,9,10)],2,as.factor))
#Numeric data - standardization
data2<-scale(mydata2[,-c(2,8,9,10)],scale=T,center = T)
data_allnum<-scale(mydata2,scale=T,center=T)

# combine
dat_gower_numcat<-cbind(data2,data1)
# distance using gower measure
distMat <- daisy(dat_gower_numcat, metric = "gower")
# hierarchical clustering
fitGower<-hclust(distMat,method="single")
dev.off()
plot(fitGower)
```

What is the alternate way of deal with mixture of attribute types?

Convert all the categorical attributes into numeric and later scale before computing distance matrix

Analysing the clusters

- Now the challenge is, how to explain the clusters to business and use them. Can we find some cluster characteristics?

  Let's inspect the cluster centres from K-means output and try to name them. This step has to be done in collaboration with the business teams and agree.

  fit$centers

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
1	-0.5404546	0.6415922	0.2819522	1.6235100	0.3680169	-0.0482903	-1.6688182	-0.8680278			
2	-0.0565170	-0.1049878	-0.5399595	-0.4402893	0.5862165	-0.1262191	-0.1000312	0.124004			
3	-0.8363478	1.0148821	1.0238513	0.6924910	-0.8897477	0.9063586	-0.3952280	-0.8680278			
4	1.3247791	-1.2248578	-1.1062677	-0.9453003	1.0982062	-1.2008698	0.3364684	0.8680278			
5	0.2570746	-0.7769098	-0.4244185	-0.7713723	-0.3115188	-0.1239195	1.4915135	1.1160357			

```
1  1.1899014  1.7789276  1.9734398
2  0.1878792  0.4235542  0.7352031
3 -0.8141431 -0.9318192  0.1676779
4  1.1899014  0.7623975 -0.8125929
5 -0.8141431 -0.3896699 -0.8745047
```

Cluster 1 – cars with high ‘HP’, ‘AM’, ‘Gear’, ‘Carb’
Cluster 3 – cars with high ‘Disp’ and ‘Cyl’
Cluster 4 – cars with high ‘MPG’ and ‘Drat’
Cluster 5 – cars with high ‘Qsec’ and ‘VS’
Cluster 2 – otherwise

Kernel clustering

```
rm(x)
group.one = cbind(runif(100,0,15),  runif(100,0,1))
group.two = cbind(rnorm(100,15,1),  rnorm(100,5,1))
uniform.pts = runif(100,4,15)
group.three = cbind(uniform.pts+
                     rnorm(100,0,.5),
                    uniform.pts+
                     rnorm(100,0,.5))
x = rbind(group.one,group.two,group.three)
plot(x, xlab="",ylab="")

km = kmeans(x, centers=3,
          iter.max=20)

plot(x,xlab="",ylab="",
    col=c("red","black","blue")[km$cluster],
    main="k-means")

require(kernlab)
specclu = specc(x, kernel = "laplacedot",centers=3)
plot(x, col=specclu)

library(kernlab)
specclu = specc(x, centers=2)
plot(x, col=specclu)
```

Assignment:

Cereals data: This data set contains nutritional information for 77 different breakfast cereals. It was used for the 1993 Statistical Graphics Exposition as a challenge data set. We retrieved this data from StatLib at CMU. The data is from the nutritional labels and is in CSV format.

Objective: the schools wants to pick a cereal each day of the week from similar cereals.

The variables are:

- Cereal name;
- manufacturer (e.g.,Kellogg's);
- type (cold/hot);
- calories (number);
- protein (g);
- fat (g);
- sodium (mg);
- dietary fiber (g);
- complex carbohydrates (g);
- sugars (g);
- display shelf (1, 2, or 3, counting from the floor);
- potassium (mg);
- vitamins and minerals (0, 25, or 100, respectively);
- weight (in ounces) of one serving (serving size);
- cups per serving.

Manufacturers are represented by their first initial: A=American Home Food Products, G=General Mills, K=Kelloggs, N=Nabisco, P=Post, Q=Quaker Oats, R=Ralston Purina.

(Download)

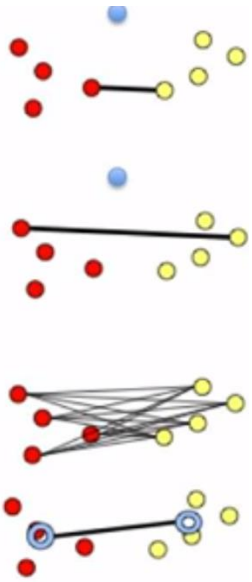
Using K-means technique identify/cluster the similar cereals.

- 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.(optional, unless trying hierarchical clustering)
- Apply k-means clustering technique, identify the ideal number of cluster.
- Identify the similar cereals based on the clusters.

Additional references:

Snapshot of different clustering methods

- Single link: $D(c_1,c_2) = \min_{x_1 \in c_1, x_2 \in c_2} D(x_1,x_2)$
  - distance between closest elements in clusters
  - produces long chains a→b→c→…→z
- Complete link: $D(c_1,c_2) = \max_{x_1 \in c_1, x_2 \in c_2} D(x_1,x_2)$
  - distance between farthest elements in clusters
  - forces "spherical" clusters with consistent "diameter"
- Average link: $D(c_1,c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\sum_{x_1 \in c_1}\sum_{x_2 \in c_2} D(x_1,x_2)$
  - average of all pairwise distances
  - less affected by outliers
- Centroids: $D(c_1,c_2) = D\left(\left(\frac{1}{|c_1|}\sum_{x \in c_1}\vec{x}\right),\left(\frac{1}{|c_2|}\sum_{x \in c_2}\vec{x}\right)\right)$
  - distance between centroids (means) of two clusters



Reference:

https://websites.pmc.ucsc.edu/~mclapham/Rtips/cluster.htm

http://rstudio-pubs-static.s3.amazonaws.com/5378_6b71c972feb04fd19776daa9063df5cd.html

https://www.youtube.com/watch?v=VMyXc3SiEqs