Key takeaways:

Implementing clustering methods

- Hierarchical clustering
- K- medoids clustering
- Spectral clustering

Hierarchical clustering

1. Load data 'mtcars' available in R
   mydata <- mtcars
2. Understand the summary of data
3. Remove the records with missing values in data
4. Standardize the data using 'scale' function
5. Hierarchical Clustering

```
# Ward Hierarchical Clustering
#First we need to compute the distance matrix
d<-dist(data2,method = "euclidean")
#To see the distances, we needto convert it to matrix
distances<-as.matrix(d)

#Now cluster
fit<-hclust(d,method="single")
plot(fit)

fit <- hclust(d, method="ward")
plot(fit) # display dendogram
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")
```

K-medoids clustering

Please use the below R code as reference and then implement on 'Cereals' data set

```
## generate 25 objects, divided into 2 clusters.
x <- rbind(cbind(rnorm(10,0,0.5), rnorm(10,0,0.5)),
        cbind(rnorm(15,5,0.5), rnorm(15,5,0.5)))

library(cluster)
pamx <- pam(x, 2)
pamx # Medoids: '7' and '24' ...
summary(pamx)
plot(pamx)
#mean(y[7, c(11:25)])-mean(y[7,c(1:6,8:10)]))/mean(y[7, c(11:25)])- Silhoutte widths
#Nearer to zero points between clusters, and nearer to 1 well cluster
```

Spectral Clustering

Please use the below R code as reference and then implement on 'Cereals' data set

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


##Manual Calculation of Spectral clustering
distances = as.matrix(dist(x))
W = exp(-distances^2)
G = diag(rowSums(W))
L = G - W
eig = eigen(L)

km2 = kmeans(eig$vectors[,24],
          centers=2)

plot(x,xlab="",ylab="",
   col=c("red","black","blue")[km2$cluster],
   main="spectral clustering")
```

Jaccard's Distance

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

Identify the symmetric and asymmetric attributes- (Hint: Asymmetric attributes are those categorical attributes in which one level is more important than the other)

Identify which two of the three are closer
Jaccard's coefficient:

$$Dist = \frac{number\ of\ dissimiar\ attributes\ between\ the\ records}{number\ of\ dissimilar\ attributes + number\ of\ similar\ attributes\ (excluding\ records\ with\ 0,0)}$$

	Data point j 1	Data point j 0
Data point i 1	a	b
0	c	d