Objective:
In this session, multiple distance computation methods, implementation of k-means clustering, identifying number of clusters and interpreting the results.

Key takeaways:

- Multiple Distance Computation Methods
- Data understanding and preparation for cluster analysis
- Implementing cluster analysis
    - Steps to form clusters from the given data
- Interpreting the results

<u>Problem Statement:</u>

“Cereals.csv” contains dietary characteristics of 77 products. Segment the products based on their characteristics.

1. Import the data into R
2. Data Exploration and preparation
    a. Understand the attributes and the data
    b. Drop the attributes which are not required
    c. Check for missing values
    d. Impute the missing values (if any) using KNN imputation
3. Standardize the data using ‘z-score’
4. Implementing k-means clustering
    a. Execute algorithm with random ‘k’

```
#K-means clustering
fit<-kmeans(mydata,centers=5)
```

    b. Extracting output of the model and interpretation

```
fit
#With-in sum of squares in each cluster
fit$withinss
sum(fit$withinss)
#Cluster Centers
fit$centers
#To check cluster number of each row in data
fit$cluster
```

    c. Determine number of clusters

```
# K-means:  Determine number of clusters
wss <- 0
for (i in 1:15) {
  set.seed(1234)
  wss[i] <- sum(kmeans(data,centers=i)$withinss)
}

# Plot the cluster number and withinness error
plot(1:15, wss,
     type="b",
     xlab="Number of Clusters",
     ylab="Total within-cluster sum of squares")
```

d. Fit the model with selected cluster number

```
set.seed(1234)
fit <- kmeans(data, centers = 6)
#With-in sum of squares in each cluster
fit$withinss
sum(fit$withinss)
```