Experiment 7

Student Name: Aman Gokul	**UID:** 20BCS5449
Branch: CSE	**Section/Group:** 607/A
Semester: 6th	**Date of Performance:** 27/04/2023
Subject Name: Data Mining Lab	**Subject Code:** 20CSP-376

1. Aim:

To perform the cluster analysis by k-means method using R.

2. Apparatus / Simulation Used:

- Windows 7 or above
- R Studio

3. Objective:

- Demonstration of the k-means method using R.
- Performing the cluster analysis by k-means method using R.

4. Theory and Output:

K Means Clustering in R Programming is an Unsupervised Non-linear algorithm that cluster data based on similarity or similar groups. It seeks to partition the observations into a pre-specified number of clusters. Segmentation of data takes place to assign each training example to a segment called a cluster. In the unsupervised algorithm, high reliance on raw data is given with large expenditure on manual review for review of relevance is given. It is used in a variety of fields like Banking, healthcare, retail, Media, etc.

K-Means clustering groups the data on similar groups. The algorithm is as follows:

- Choose the number **K** clusters.
- Select at random K points, the centroids (Not necessarily from the given data).
- Assign each data point to closest centroid that forms K clusters.
- Compute and place the new centroid of each centroid.
- After final reassignment, name the cluster as Final cluster.

The Dataset

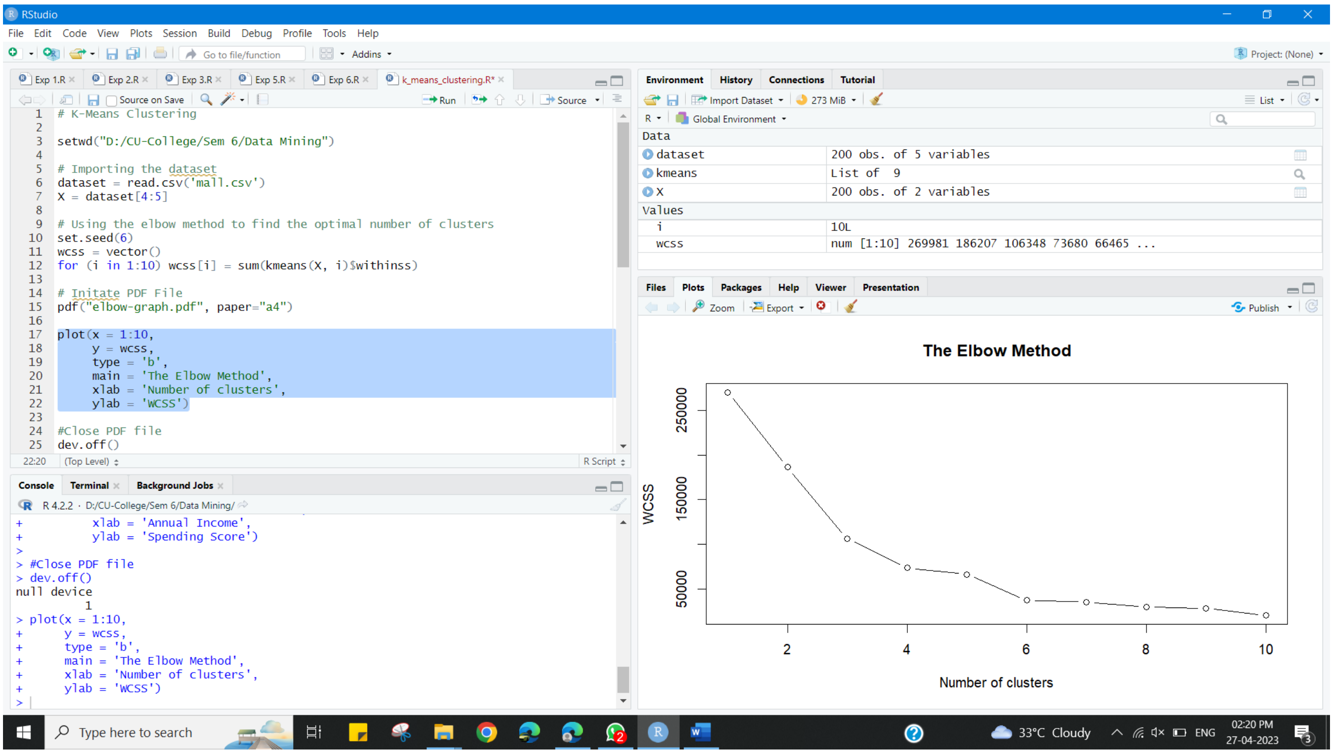
Iris dataset consists of 50 samples from each of 3 species of Iris(Iris setosa, Iris virginica, Iris versicolor) and a multivariate dataset introduced by British statistician and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems. Four features were measured from each sample i.e length and width of the sepals and petals and based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other.

```
# Loading data
data(iris)

# Structure
str(iris)
```

5. Code:

```
# K-Means Clustering

setwd("D:/CU-College/Sem 6/Data Mining")

# Importing the dataset
dataset = read.csv('mall.csv')
X = dataset[4:5]

# Using the elbow method to find the optimal number of clusters
set.seed(6)
wcss = vector()
for (i in 1:10) wcss[i] = sum(kmeans(X, i)$withinss)

# Initate PDF File
pdf("elbow-graph.pdf", paper="a4")

plot(x = 1:10,
     y = wcss,
     type = 'b',
     main = 'The Elbow Method',
     xlab = 'Number of clusters',
     ylab = 'WCSS')

#Close PDF file
dev.off()


# Fitting K-Means to the dataset
set.seed(29)
kmeans = kmeans(x = X,
                centers = 6,
                iter.max = 300,
                nstart = 10)

# Visualising the cluster
library(cluster)

# Initate PDF File
pdf("clusterplot.pdf", paper="a4")

clusplot(x = X,
         clus = kmeans$cluster,
         lines = 0,
         shade = TRUE,
         color = TRUE,
         labels = 4,
         plotchar = TRUE,
         span = TRUE,
         main = 'Clusters of customers',
        xlab = 'Annual Income',
        ylab = 'Spending Score')

#Close PDF file
dev.off()
```

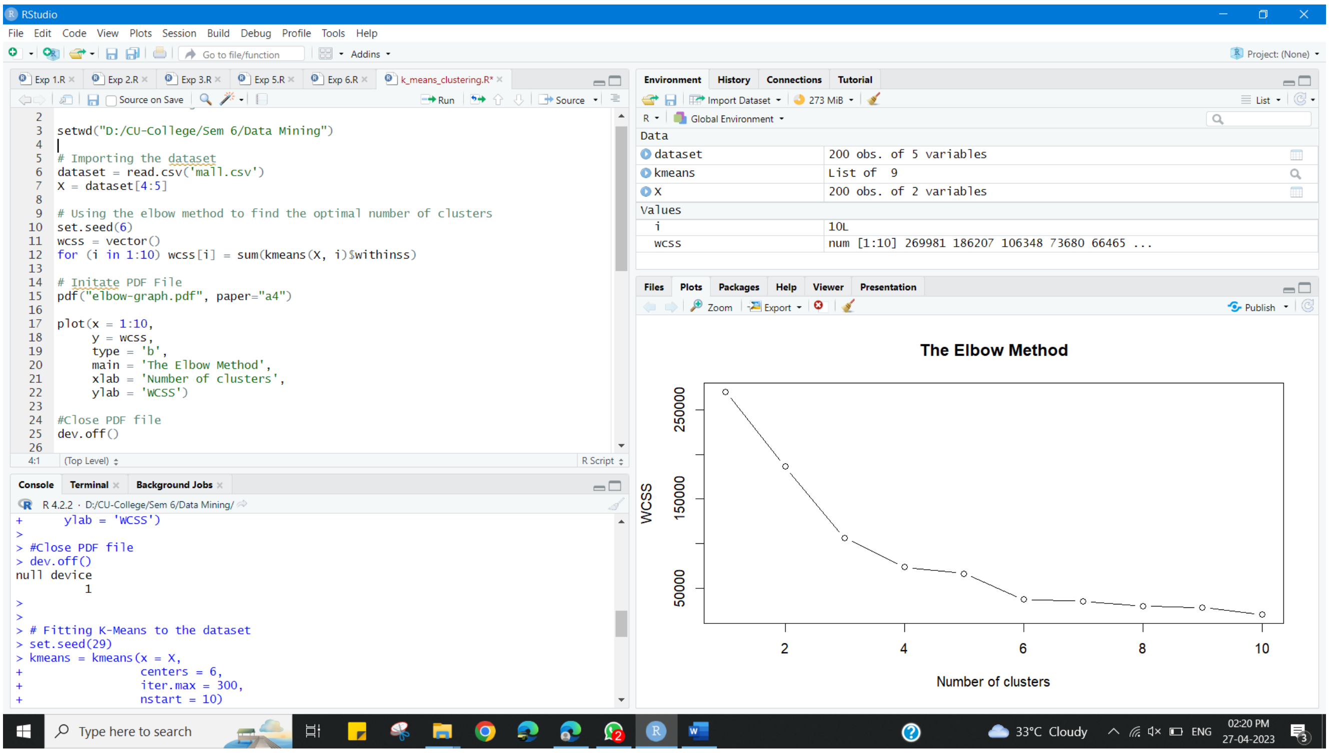
6. Output:



```
# K-Means Clustering

setwd("D:/CU-College/Sem 6/Data Mining")

# Importing the dataset
dataset = read.csv('mall.csv')
X = dataset[4:5]

# Using the elbow method to find the optimal number of clusters
set.seed(6)
wcss = vector()
for (i in 1:10) wcss[i] = sum(kmeans(X, i)$withinss)

# Initate PDF File
pdf("elbow-graph.pdf", paper="a4")

plot(x = 1:10,
     y = wcss,
     type = 'b',
     main = 'The Elbow Method',
     xlab = 'Number of clusters',
     ylab = 'WCSS')

#Close PDF file
dev.off()
```

Data	
dataset	200 obs. of 5 variables
kmeans	List of 9
X	200 obs. of 2 variables
Values	
i	10L
wcss	num [1:10] 269981 186207 106348 73680 66465 ...

```
R 4.2.2 · D:/CU-College/Sem 6/Data Mining/
+         xlab = 'Annual Income',
+         ylab = 'Spending Score')
>
> #Close PDF file
> dev.off()
null device
          1
> plot(x = 1:10,
+      y = wcss,
+      type = 'b',
+      main = 'The Elbow Method',
+      xlab = 'Number of clusters',
+      ylab = 'WCSS')
>
```



```r

setwd("D:/CU-College/Sem 6/Data Mining")

# Importing the dataset
dataset = read.csv('mall.csv')
X = dataset[4:5]

# Using the elbow method to find the optimal number of clusters
set.seed(6)
wcss = vector()
for (i in 1:10) wcss[i] = sum(kmeans(X, i)$withinss)

# Initate PDF File
pdf("elbow-graph.pdf", paper="a4")

plot(x = 1:10,
     y = wcss,
     type = 'b',
     main = 'The Elbow Method',
     xlab = 'Number of clusters',
     ylab = 'WCSS')

#Close PDF file
dev.off()

```

```
R 4.2.2 · D:/CU-College/Sem 6/Data Mining/
+      ylab = 'WCSS')
>
> #Close PDF file
> dev.off()
null device
          1
>
>
> # Fitting K-Means to the dataset
> set.seed(29)
> kmeans = kmeans(x = X,
+                 centers = 6,
+                 iter.max = 300,
+                 nstart = 10)
```

```
R
Data
dataset     200 obs. of 5 variables
kmeans      List of  9
X           200 obs. of 2 variables
Values
  i         10L
  wcss      num [1:10] 269981 186207 106348 73680 66465 ...
```

Learning outcomes (What I have learnt):

- Demonstration of the k-means method using R.
- Performing the cluster analysis by k-means method using R.