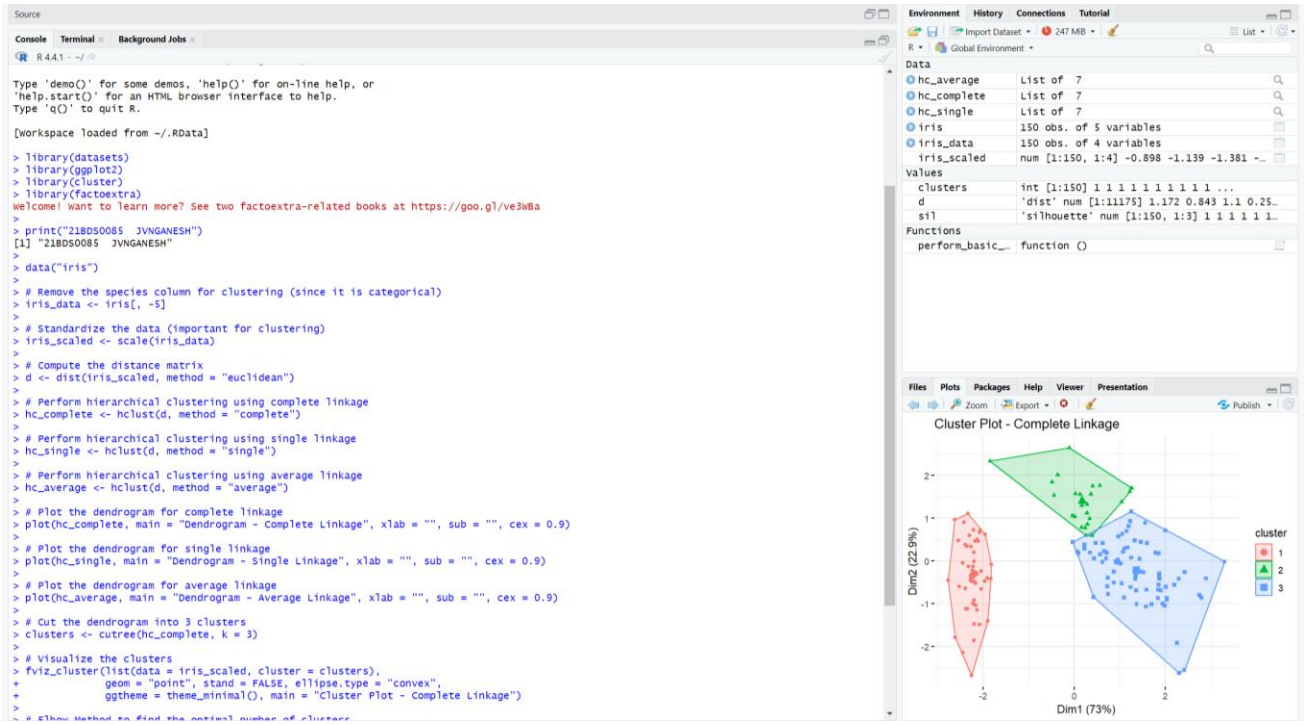
EDA Experiment – 10

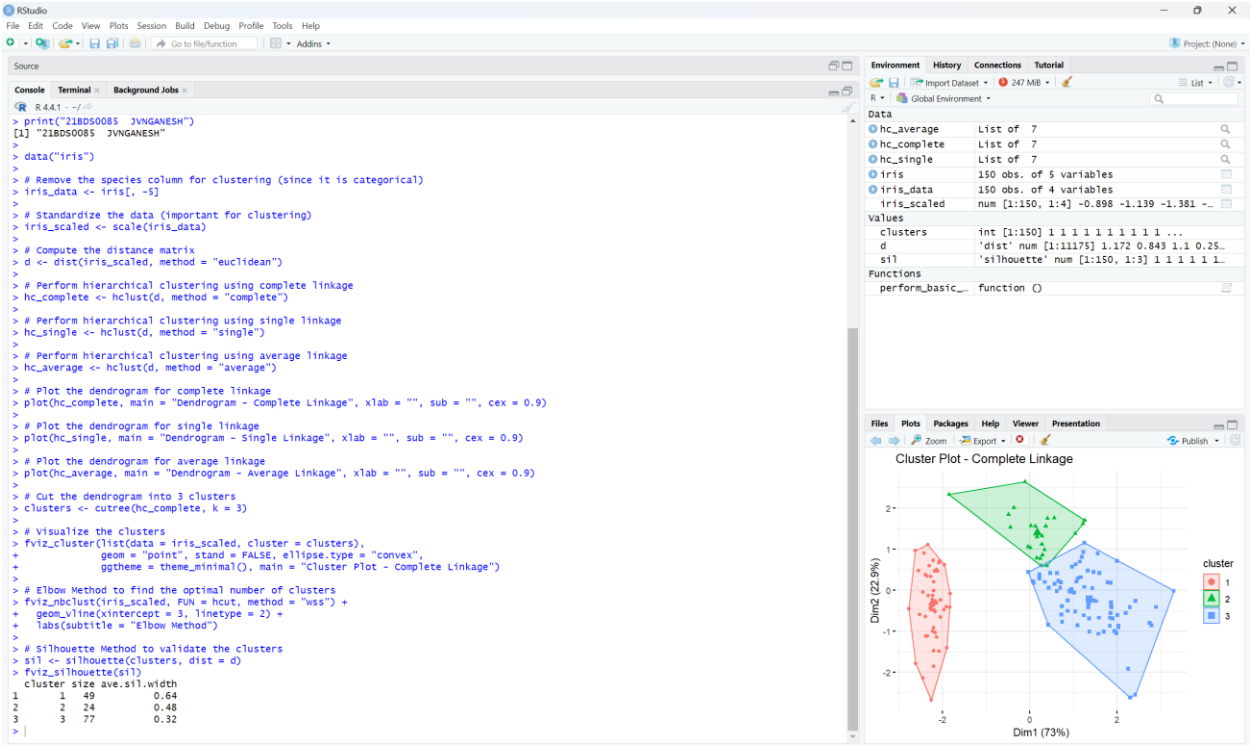
Digital Assignment

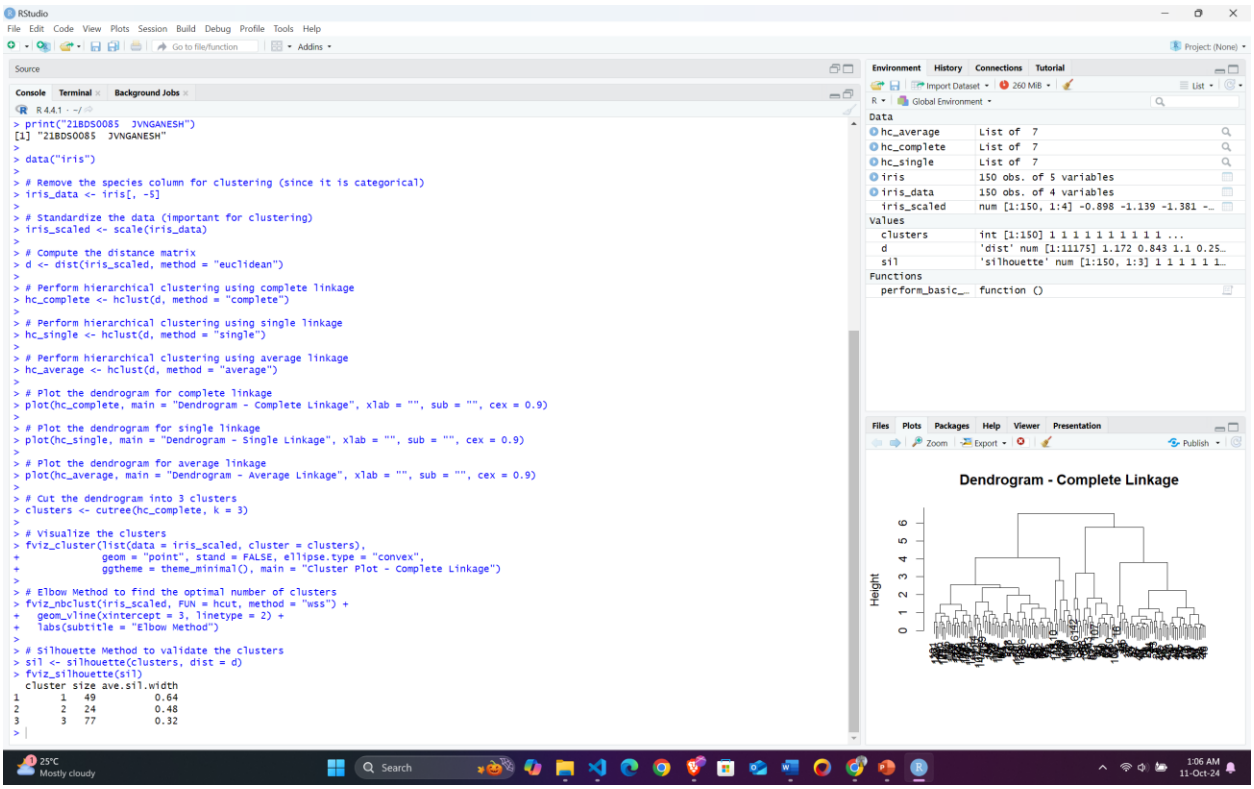
Name : Jvn Ganesh

Roll No : 21BDS0085

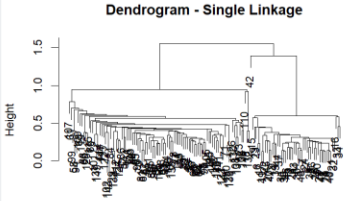
Screenshots

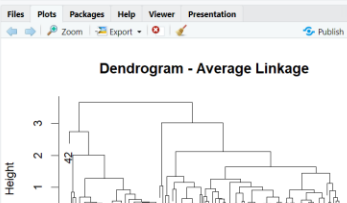




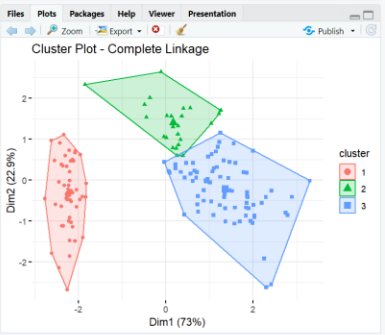

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal Background Jobs
R 4.4.1 · ~/
> print("21BDS0085 JVNGANESH")
[1] "21BDS0085 JVNGANESH"
> data("iris")
> # Remove the species column for clustering (since it is categorical)
> iris_data <- iris[, -5]
> # Standardize the data (important for clustering)
> iris_scaled <- scale(iris_data)
> # Compute the distance matrix
> d <- dist(iris_scaled, method = "euclidean")
> # Perform hierarchical clustering using complete linkage
> hc_complete <- hclust(d, method = "complete")
> # Perform hierarchical clustering using single linkage
> hc_single <- hclust(d, method = "single")
> # Perform hierarchical clustering using average linkage
> hc_average <- hclust(d, method = "average")
> # Plot the dendrogram for complete linkage
> plot(hc_complete, main = "Dendrogram - Complete Linkage", xlab = "", sub = "", cex = 0.9)
> # Plot the dendrogram for single linkage
> plot(hc_single, main = "Dendrogram - Single Linkage", xlab = "", sub = "", cex = 0.9)
> # Plot the dendrogram for average linkage
> plot(hc_average, main = "Dendrogram - Average Linkage", xlab = "", sub = "", cex = 0.9)
> # Cut the dendrogram into 3 clusters
> clusters <- cutree(hc_complete, k = 3)
> # Visualize the clusters
> fviz_cluster(list(data = iris_scaled, cluster = clusters),
+ geom = "point", stand = FALSE, ellipse.type = "convex",
+ ggtheme = theme_minimal(), main = "Cluster Plot - Complete Linkage")
> # Elbow Method to find the optimal number of clusters
> fviz_nbclust(iris_scaled, FUN = hcut, method = "wss") +
+ geom_vline(xintercept = 3, linetype = 2) +
+ labs(subtitle = "Elbow Method")
> # Silhouette Method to validate the clusters
> sil <- silhouette(clusters, dist = d)
> fviz_silhouette(sil)
cluster size ave.sil.width
1 1 49 0.64
2 2 24 0.48
3 3 77 0.32
Environment History Connections Tutorial
Import Dataset 260 MiB
Global Environment
Data
hc_average List of 7
hc_complete List of 7
hc_single List of 7
iris 150 obs. of 5 variables
iris_data 150 obs. of 4 variables
iris_scaled num [1:150, 1:4] -0.898 -1.139 -1.381 -...
Values
clusters int [1:150] 1 1 1 1 1 1 1 1 1 1 ...
d 'dist' num [1:11175] 1.172 0.843 1.1 0.25...
sil 'silhouette' num [1:150, 1:3] 1 1 1 1 1 1...
Functions
perform_basic_... function ()
Files Plots Packages Help Viewer Presentation
Zoom Export Publish
Dendrogram - Complete Linkage
Height
0 1 2 3 4 5 6
25°C
Mostly cloudy
Search
1:06 AM
11-Oct-24

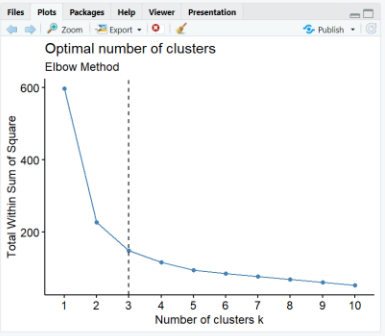
```
> print("21BDS0085  JVNGANESH")
[1] "21BDS0085  JVNGANESH"
>
> data("iris")
>
> # Remove the species column for clustering (since it is categorical)
> iris_data <- iris[, -5]
>
> # Standardize the data (important for clustering)
> iris_scaled <- scale(iris_data)
>
> # Compute the distance matrix
> d <- dist(iris_scaled, method = "euclidean")
>
> # Perform hierarchical clustering using complete linkage
> hc_complete <- hclust(d, method = "complete")
>
> # Perform hierarchical clustering using single linkage
> hc_single <- hclust(d, method = "single")
>
> # Perform hierarchical clustering using average linkage
> hc_average <- hclust(d, method = "average")
>
> # Plot the dendrogram for complete linkage
> plot(hc_complete, main = "Dendrogram - Complete Linkage", xlab = "", sub = "", cex = 0.9)
>
> # Plot the dendrogram for single linkage
> plot(hc_single, main = "Dendrogram - Single Linkage", xlab = "", sub = "", cex = 0.9)
>
> # Plot the dendrogram for average linkage
> plot(hc_average, main = "Dendrogram - Average Linkage", xlab = "", sub = "", cex = 0.9)
>
> # Cut the dendrogram into 3 clusters
> clusters <- cutree(hc_complete, k = 3)
>
> # Visualize the clusters
> fviz_cluster(list(data = iris_scaled, cluster = clusters),
+              geom = "point", stand = FALSE, ellipse.type = "convex",
+              ggtheme = theme_minimal(), main = "Cluster Plot - Complete Linkage")
>
> # Elbow Method to find the optimal number of clusters
> fviz_nbclust(iris_scaled, FUN = hcut, method = "wss") +
+   geom_vline(xintercept = 3, linetype = 2) +
+   labs(subtitle = "Elbow Method")
>
> # Silhouette Method to validate the clusters
> sil <- silhouette(clusters, dist = d)
> fviz_silhouette(sil)
  cluster size ave.sil.width
1       1   49          0.64
2       2   24          0.48
3       3   77          0.32
>
```

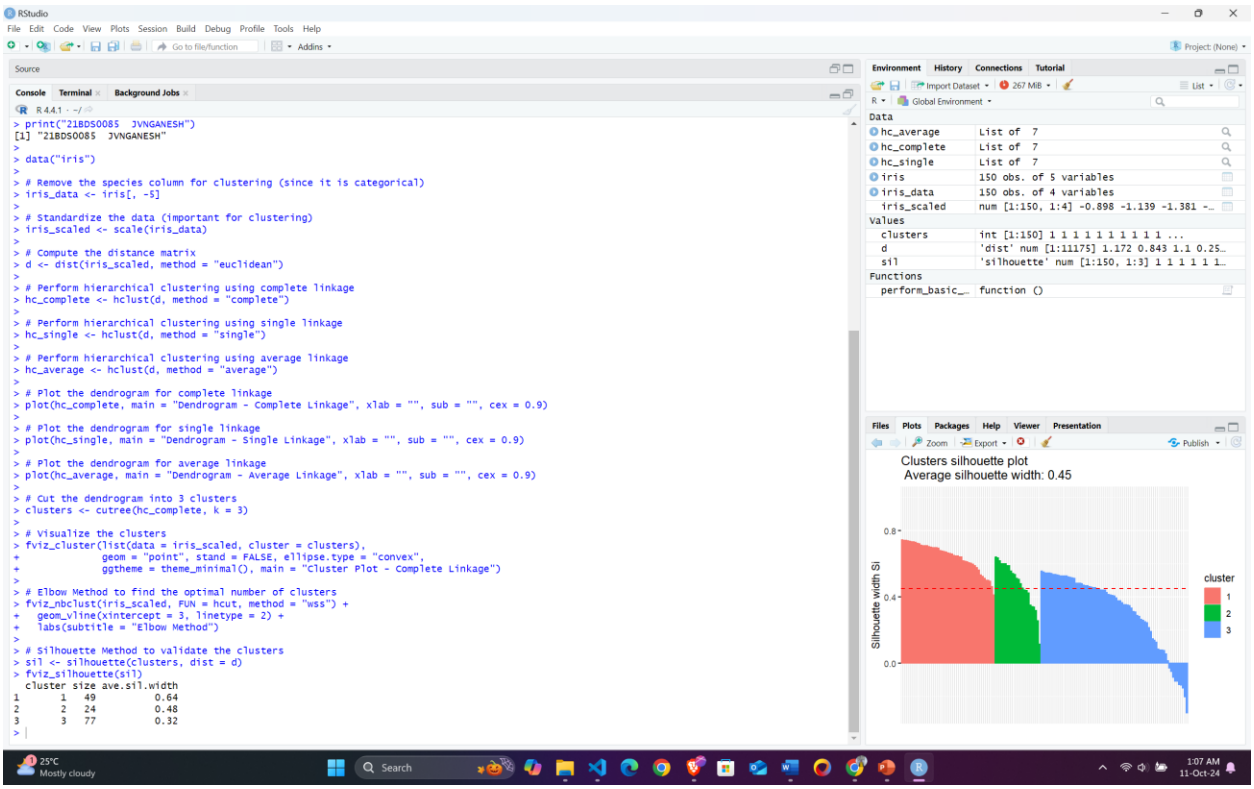


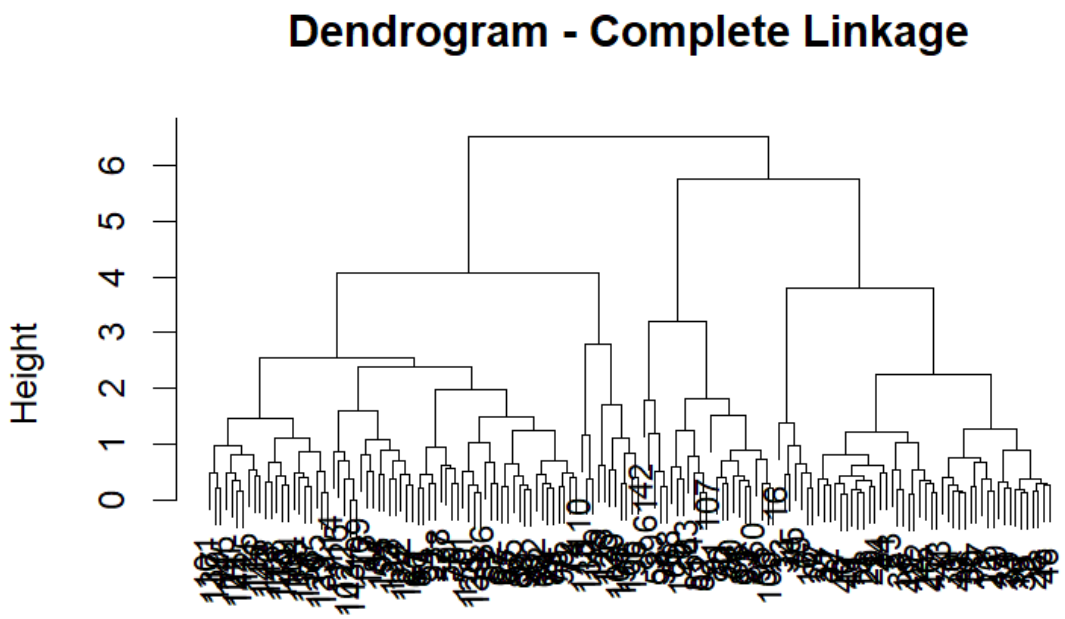


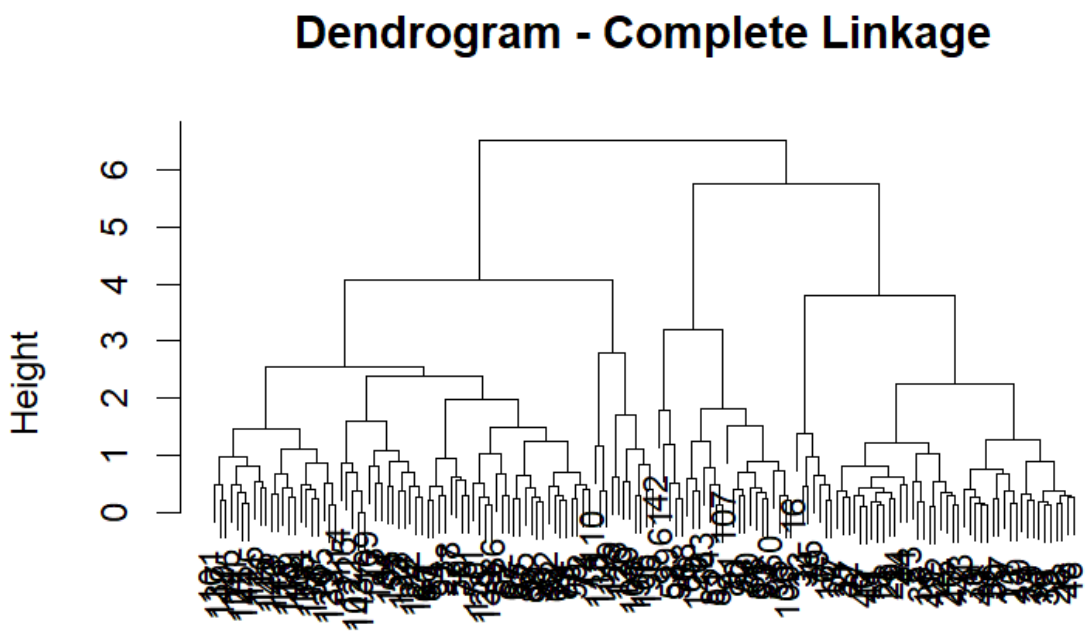
```
> print("21BDS0085  JVNGANESH")
[1] "21BDS0085  JVNGANESH"
>
> data("iris")
>
> # Remove the species column for clustering (since it is categorical)
> iris_data <- iris[, -5]
>
> # Standardize the data (important for clustering)
> iris_scaled <- scale(iris_data)
>
> # Compute the distance matrix
> d <- dist(iris_scaled, method = "euclidean")
>
> # Perform hierarchical clustering using complete linkage
> hc_complete <- hclust(d, method = "complete")
>
> # Perform hierarchical clustering using single linkage
> hc_single <- hclust(d, method = "single")
>
> # Perform hierarchical clustering using average linkage
> hc_average <- hclust(d, method = "average")
>
> # Plot the dendrogram for complete linkage
> plot(hc_complete, main = "Dendrogram - Complete Linkage", xlab = "", sub = "", cex = 0.9)
>
> # Plot the dendrogram for single linkage
> plot(hc_single, main = "Dendrogram - Single Linkage", xlab = "", sub = "", cex = 0.9)
>
> # Plot the dendrogram for average linkage
> plot(hc_average, main = "Dendrogram - Average Linkage", xlab = "", sub = "", cex = 0.9)
>
> # Cut the dendrogram into 3 clusters
> clusters <- cutree(hc_complete, k = 3)
>
> # Visualize the clusters
> fviz_cluster(list(data = iris_scaled, cluster = clusters),
+              geom = "point", stand = FALSE, ellipse.type = "convex",
+              ggtheme = theme_minimal(), main = "Cluster Plot - Complete Linkage")
>
> # Elbow Method to find the optimal number of clusters
> fviz_nbclust(iris_scaled, FUN = hcut, method = "wss") +
+   geom_vline(xintercept = 3, linetype = 2) +
+   labs(subtitle = "Elbow Method")
>
> # Silhouette Method to validate the clusters
> sil <- silhouette(clusters, dist = d)
> fviz_silhouette(sil)
  cluster size ave.sil.width
1       1   49          0.64
2       2   24          0.48
3       3   77          0.32
>
```

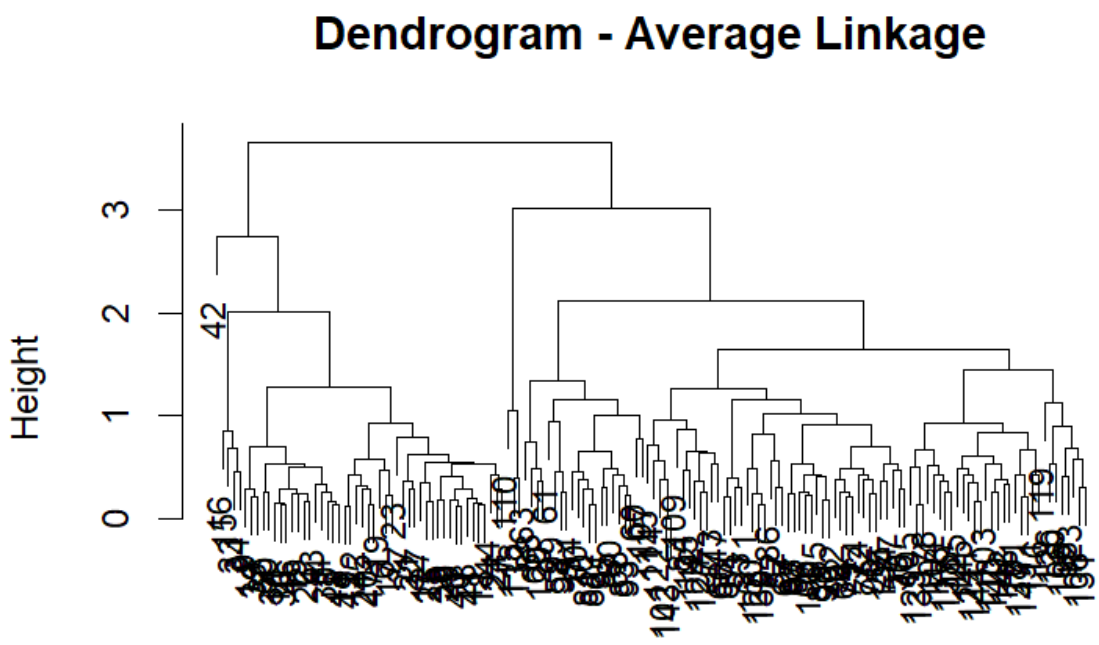


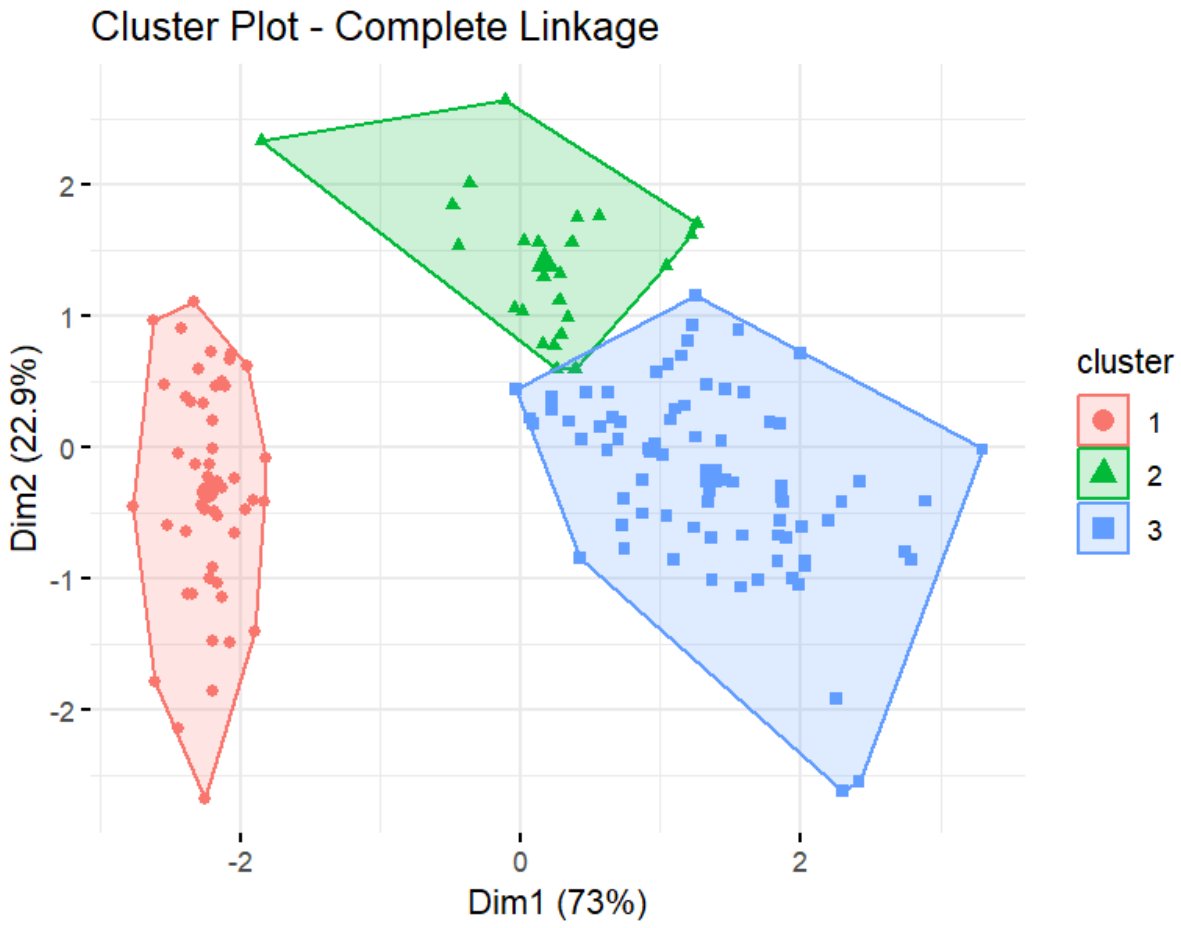


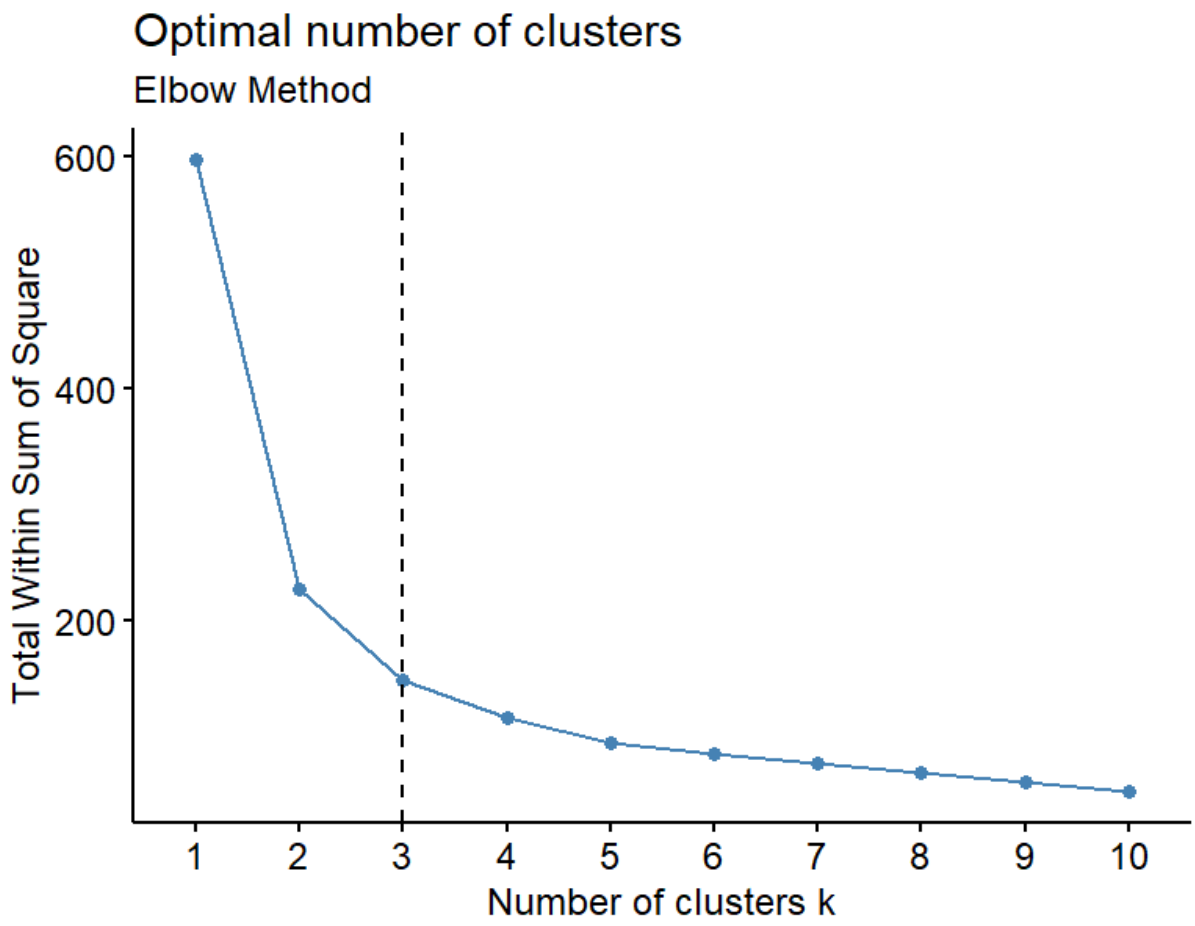

RStudio
> print("21BDS0085 JVNGANESH")
[1] "21BDS0085 JVNGANESH"
> data("iris")
> # Remove the species column for clustering (since it is categorical)
> iris_data <- iris[, -5]
> # Standardize the data (important for clustering)
> iris_scaled <- scale(iris_data)
> # Compute the distance matrix
> d <- dist(iris_scaled, method = "euclidean")
> # Perform hierarchical clustering using complete linkage
> hc_complete <- hclust(d, method = "complete")
> # Perform hierarchical clustering using single linkage
> hc_single <- hclust(d, method = "single")
> # Perform hierarchical clustering using average linkage
> hc_average <- hclust(d, method = "average")
> # Plot the dendrogram for complete linkage
> plot(hc_complete, main = "Dendrogram - Complete Linkage", xlab = "", sub = "", cex = 0.9)
> # Plot the dendrogram for single linkage
> plot(hc_single, main = "Dendrogram - Single Linkage", xlab = "", sub = "", cex = 0.9)
> # Plot the dendrogram for average linkage
> plot(hc_average, main = "Dendrogram - Average Linkage", xlab = "", sub = "", cex = 0.9)
> # Cut the dendrogram into 3 clusters
> clusters <- cutree(hc_complete, k = 3)
> # Visualize the clusters
> fviz_cluster(list(data = iris_scaled, cluster = clusters),
+ geom = "point", stand = FALSE, ellipse.type = "convex",
+ ggtheme = theme_minimal(), main = "Cluster Plot - Complete Linkage")
> # Elbow Method to find the optimal number of clusters
> fviz_nbclust(iris_scaled, FUN = hcut, method = "wss") +
+ geom_vline(xintercept = 3, linetype = 2) +
+ labs(subtitle = "Elbow Method")
> # Silhouette Method to validate the clusters
> sil <- silhouette(clusters, dist = d)
> fviz_silhouette(sil)
cluster size ave.sil.width
1 1 49 0.64
2 2 24 0.48
3 3 77 0.32
Clusters silhouette plot
Average silhouette width: 0.45
Silhouette width Si
cluster
1
2
3

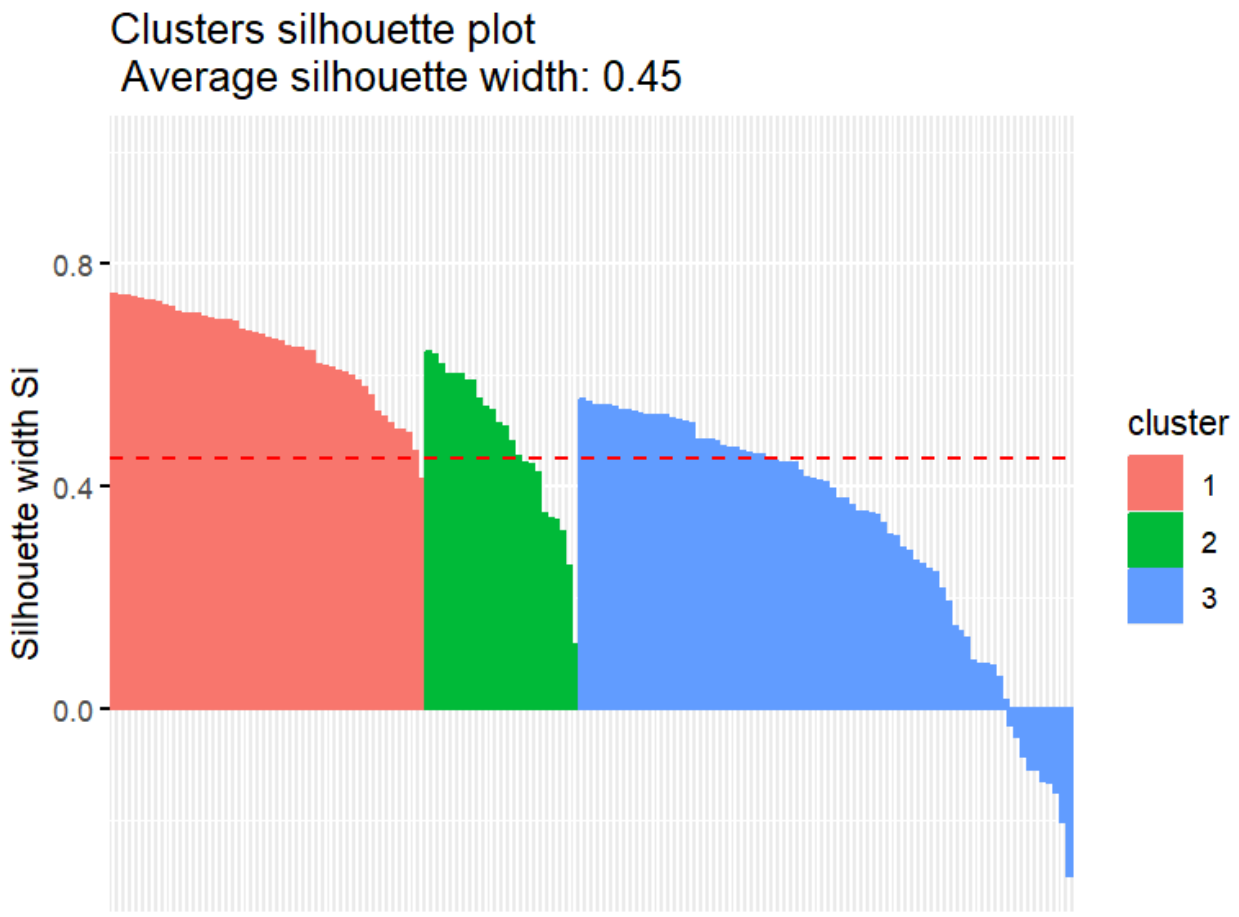

Dendrogram - Complete Linkage
Height
0 1 2 3 4 5 6


Dendrogram - Complete Linkage
Height
0
1
2
3
4
5
6


Dendrogram - Average Linkage
Height
0
1
2
3


Cluster Plot - Complete Linkage
Dim2 (22.9%)
Dim1 (73%)
2
1
0
-1
-2
-2
0
2
cluster
1
2
3


Optimal number of clusters
Elbow Method
Total Within Sum of Square
600
400
200
1
2
3
4
5
6
7
8
9
10
Number of clusters k


Clusters silhouette plot
Average silhouette width: 0.45
Silhouette width Si
0.8
0.4
0.0
cluster
1
2
3

Code:

```
library(datasets)
library(ggplot2)
library(cluster)
library(factoextra)

print("21BDS0085  JVNGANESH")

data("iris")

# Remove the species column for clustering (since it is categorical)
iris_data <- iris[, -5]

# Standardize the data (important for clustering)
```

```
iris_scaled <- scale(iris_data)


# Compute the distance matrix
d <- dist(iris_scaled, method = "euclidean")


# Perform hierarchical clustering using
complete linkage
hc_complete <- hclust(d, method =
"complete")


# Perform hierarchical clustering using single
linkage
hc_single <- hclust(d, method = "single")


# Perform hierarchical clustering using
average linkage
```

```
hc_average <- hclust(d, method = "average")

# Plot the dendrogram for complete linkage
plot(hc_complete, main = "Dendrogram - Complete Linkage", xlab = "", sub = "", cex = 0.9)

# Plot the dendrogram for single linkage
plot(hc_single, main = "Dendrogram - Single Linkage", xlab = "", sub = "", cex = 0.9)

# Plot the dendrogram for average linkage
plot(hc_average, main = "Dendrogram - Average Linkage", xlab = "", sub = "", cex = 0.9)

# Cut the dendrogram into 3 clusters
```

```
clusters <- cutree(hc_complete, k = 3)


# Visualize the clusters
fviz_cluster(list(data = iris_scaled, cluster =
clusters),
          geom = "point", stand = FALSE,
ellipse.type = "convex",
          ggtheme = theme_minimal(), main =
"Cluster Plot - Complete Linkage")


# Elbow Method to find the optimal number
of clusters
fviz_nbclust(iris_scaled, FUN = hcut, method
= "wss") +
  geom_vline(xintercept = 3, linetype = 2) +
  labs(subtitle = "Elbow Method")
```

```
# Silhouette Method to validate the clusters
sil <- silhouette(clusters, dist = d)
fviz_silhouette(sil)
```

OUTPUT

```
> library(datasets)
> library(ggplot2)
> library(cluster)
> library(factoextra)
Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
>
> print("21BDS0085  JVNGANESH")
[1] "21BDS0085  JVNGANESH"
```

```
>
> data("iris")
>
> # Remove the species column for clustering (since it is categorical)
> iris_data <- iris[, -5]
>
> # Standardize the data (important for clustering)
> iris_scaled <- scale(iris_data)
>
> # Compute the distance matrix
> d <- dist(iris_scaled, method = "euclidean")
>
> # Perform hierarchical clustering using complete linkage
```

```
> hc_complete <- hclust(d, method =
"complete")
>
> # Perform hierarchical clustering using
single linkage
> hc_single <- hclust(d, method = "single")
>
> # Perform hierarchical clustering using
average linkage
> hc_average <- hclust(d, method =
"average")
>
> # Plot the dendrogram for complete linkage
> plot(hc_complete, main = "Dendrogram -
Complete Linkage", xlab = "", sub = "", cex =
0.9)
```

```
>
> # Plot the dendrogram for single linkage
> plot(hc_single, main = "Dendrogram - Single Linkage", xlab = "", sub = "", cex = 0.9)
>
> # Plot the dendrogram for average linkage
> plot(hc_average, main = "Dendrogram - Average Linkage", xlab = "", sub = "", cex = 0.9)
>
> # Cut the dendrogram into 3 clusters
> clusters <- cutree(hc_complete, k = 3)
>
> # Visualize the clusters
> fviz_cluster(list(data = iris_scaled, cluster = clusters),
```

```
+            geom = "point", stand = FALSE,
ellipse.type = "convex",
+            ggtheme = theme_minimal(), main =
"Cluster Plot - Complete Linkage")
>
> # Elbow Method to find the optimal number
of clusters
> fviz_nbclust(iris_scaled, FUN = hcut,
method = "wss") +
+   geom_vline(xintercept = 3, linetype = 2) +
+   labs(subtitle = "Elbow Method")
>
> # Silhouette Method to validate the clusters
> sil <- silhouette(clusters, dist = d)
> fviz_silhouette(sil)
  cluster size ave.sil.width
```

1	1	49	0.64
2	2	24	0.48
3	3	77	0.32