**Experiment 8**

**Student Name:** Sahil Kaundal **UID:** 21BCS8197

**Branch:** CSE (Lateral Entry)  **Section/Group:** 616/A

**Semester:** 6th **Date of Performance:** 04/05/2023

**Subject Name:** Data Mining Lab **Subject Code:** 20CSP-376

1. **Aim:**

To perform the hierarchical clustering using R programming.

1. **Apparatus / Simulation Used:**

* Windows 7 or above
* R Studio

1. **Objective:**

* Demonstration of the hierarchical clustering method using R.
* Performing the hierarchical clustering using R.

1. **Theory and Output:**

**Cluster analysis** or [clustering](https://www.geeksforgeeks.org/clustering-in-machine-learning/) is a technique to find subgroups of data points within a data set. The data points belonging to the same subgroup have similar features or properties. Clustering is an unsupervised machine learning approach and has a wide variety of applications such as market research, pattern recognition, recommendation systems, and so on. The most common algorithms used for clustering are [K-means](https://www.geeksforgeeks.org/k-means-clustering-introduction/) clustering and Hierarchical cluster analysis. In this article, we will learn about hierarchical cluster analysis and its implementation in [R programming](https://www.geeksforgeeks.org/introduction-to-r-programming-language/).

**Hierarchical cluster analysis** (also known as hierarchical clustering) is a clustering technique where clusters have a hierarchy or a predetermined order. Hierarchical clustering can be represented by a tree-like structure called a **Dendrogram**. There are two types of hierarchical clustering:

·         **Agglomerative hierarchical clustering**: This is a bottom-up approach where each data point starts in its own cluster and as one moves up the hierarchy, similar pairs of clusters are merged.

·         **Divisive hierarchical clustering**: This is a top-down approach where all data points start in one cluster and as one moves down the hierarchy, clusters are split recursively.

Performing**Hierarchical Cluster Analysis using R**

For computing hierarchical clustering in R, the commonly used functions are as follows:

·         **hclust** in the stats package and **agnes** in the cluster package for agglomerative hierarchical clustering.

·         **diana** in the cluster package for divisive hierarchical clustering.

We will use the Iris flower data set from the datasets package in our implementation. We will use sepal width, sepal length, petal width, and petal length column as our data points. First, we load and normalize the data. Then the dissimilarity values are computed with *dist* function and these values are fed to clustering functions for performing hierarchical clustering.

1. **Code:**

# Load required packages

library(datasets) # contains iris dataset

library(cluster) # clustering algorithms

library(factoextra) # visualization

library(purrr) # to use map\_dbl() function

# Load and preprocess the dataset

df <- iris[, 1:4]

df <- na.omit(df)

df <- scale(df)

# Dissimilarity matrix

d <- dist(df, method="euclidean")

d

hc1 <- hclust(d, method = "complete")

plot(hc1, cex = 0.6, hang=-1)

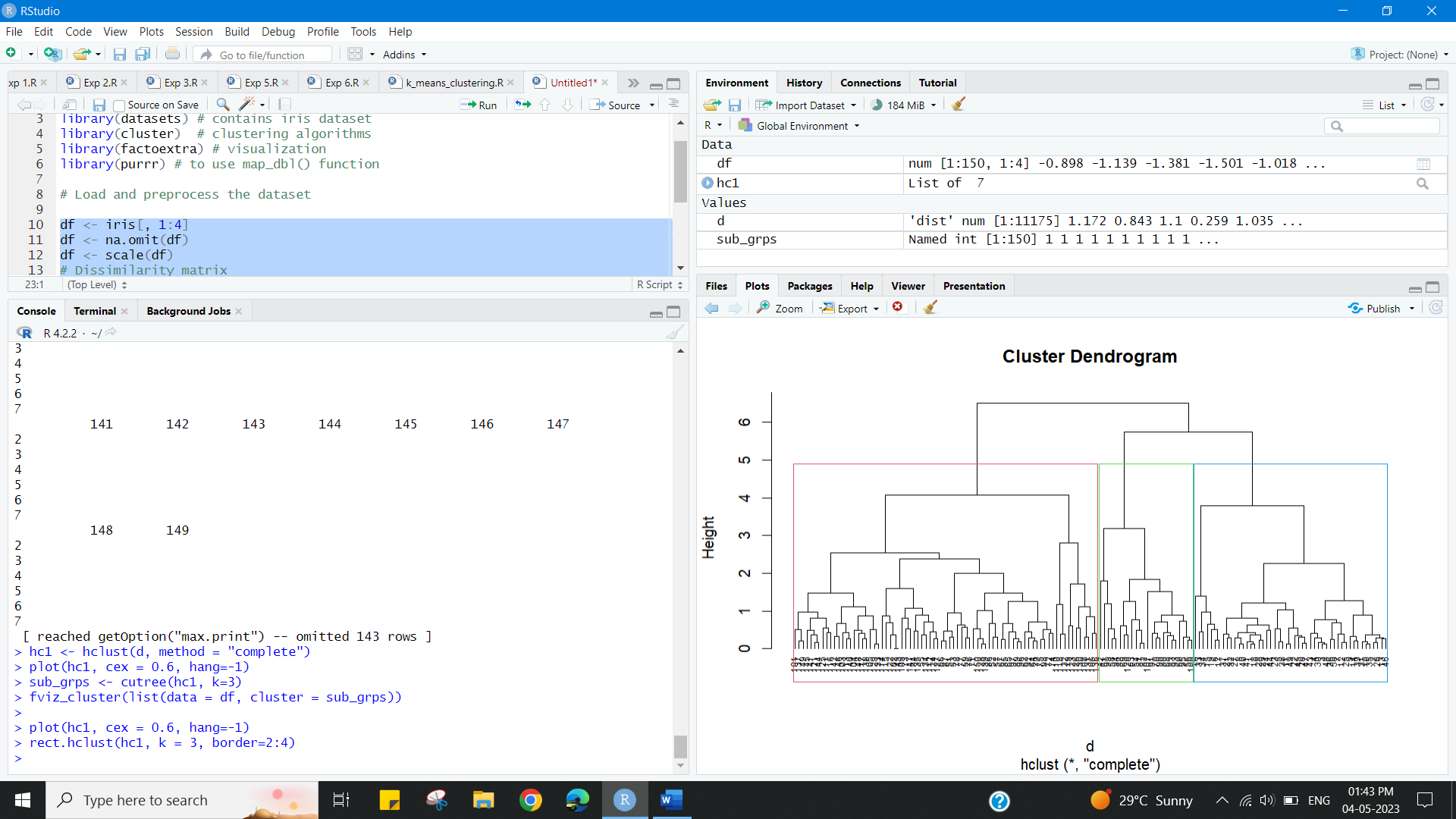
sub\_grps <- cutree(hc1, k=3)

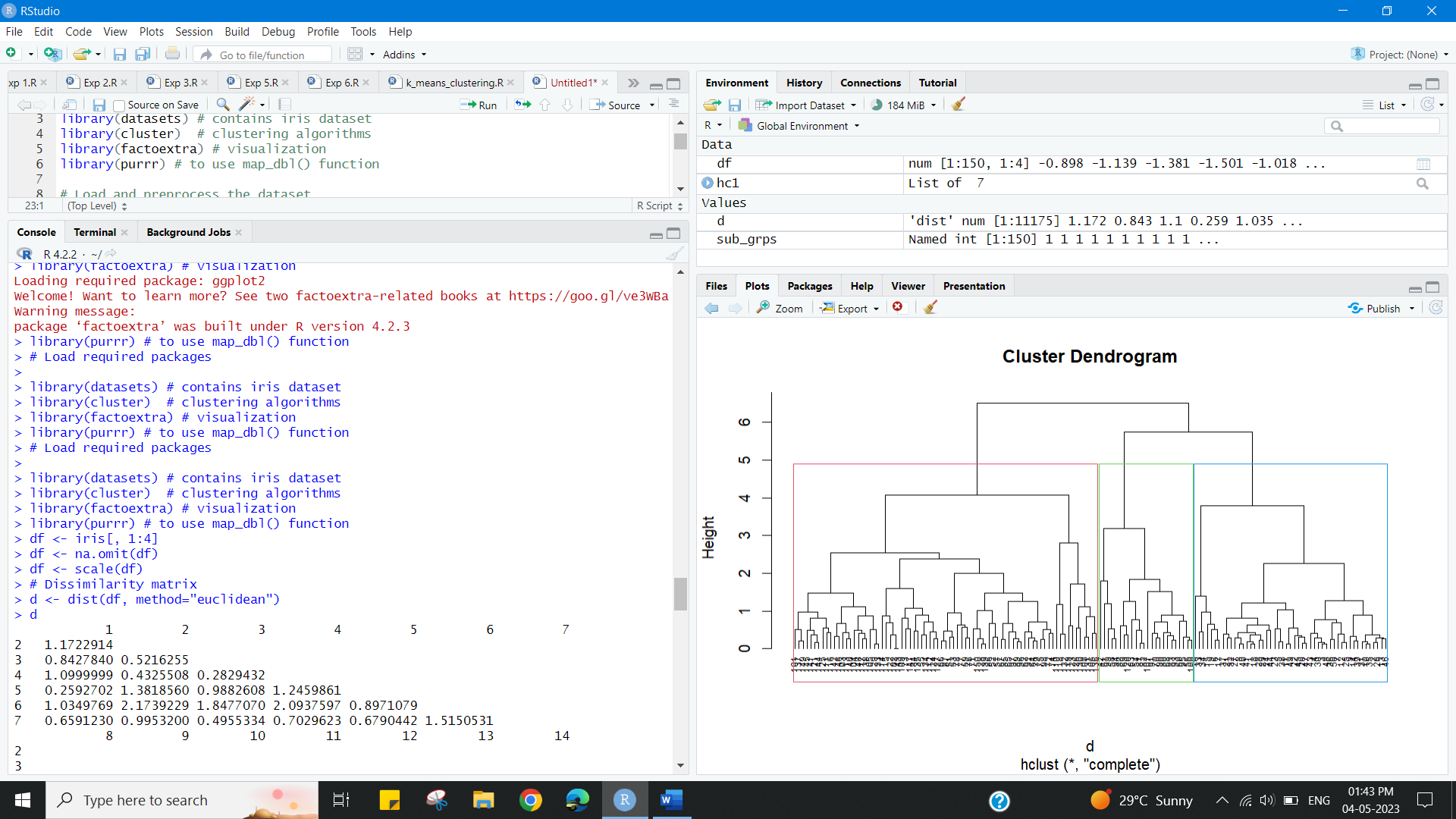
fviz\_cluster(list(data = df, cluster = sub\_grps))

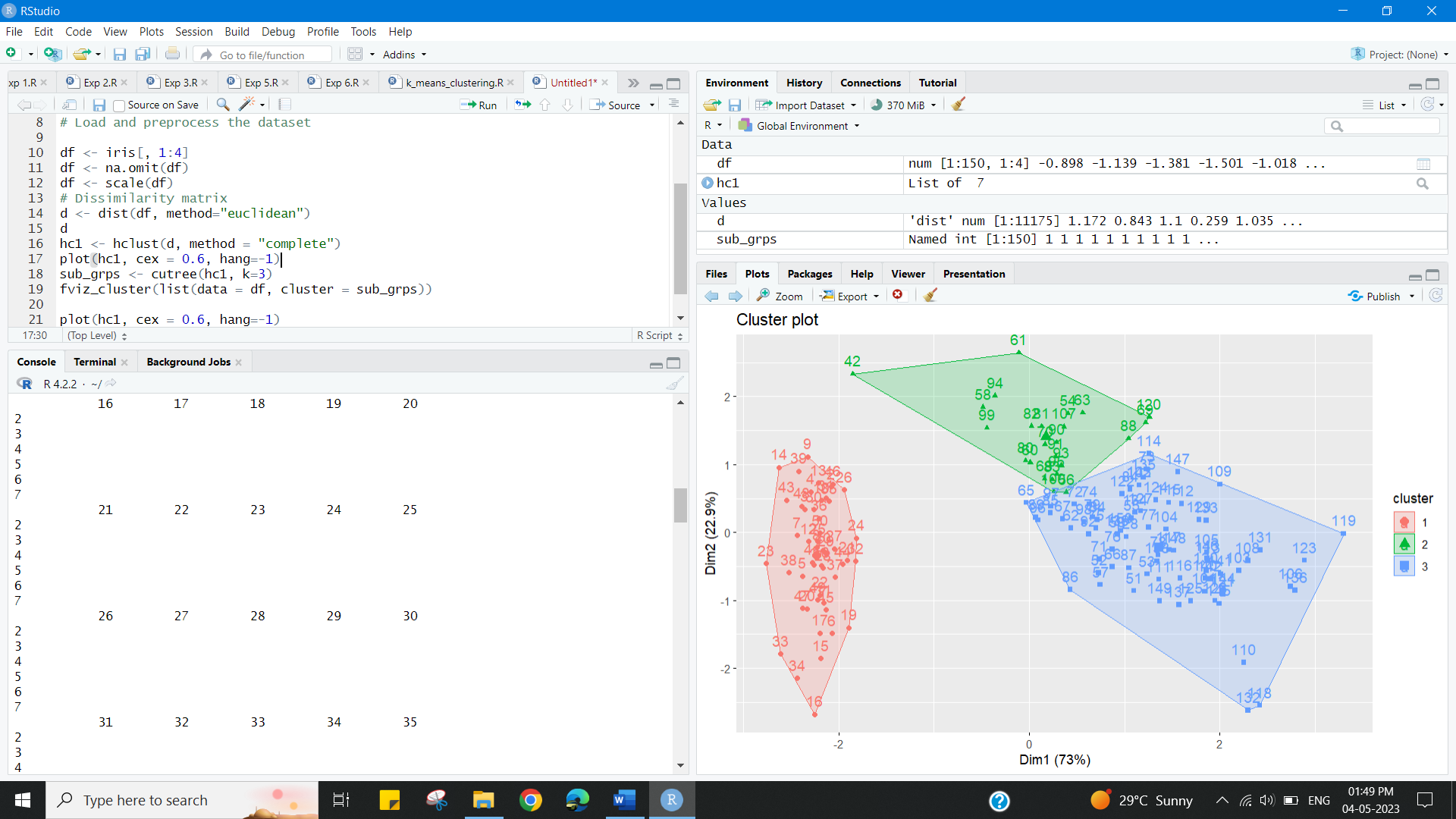
plot(hc1, cex = 0.6, hang=-1)

rect.hclust(hc1, k = 3, border=2:4)

1. **Output:**

****

****

****

**Learning outcomes (What I have learnt):**

* Demonstration of the hierarchical clustering method using R.
* Performing the hierarchical clustering using R.