|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AIUB | | **American International University- Bangladesh (AIUB)**  **Faculty of Engineering (EEE)** | | |
| **Course Name:** | Introduction to data science | |
| **Semester:** | Spring 2022-23 | | **Submission date:** | 30-4-2023 | |
| **Topic:** | Finalterm project report | | | |
| **Student Name:** | **Khan,MD zubair hossain** | | **Student ID:** | **20-42733-1** | |
| **Department:** | **CSE** | | **Section:** | **D** | |

**CODE :**



* **library(factoextra)**

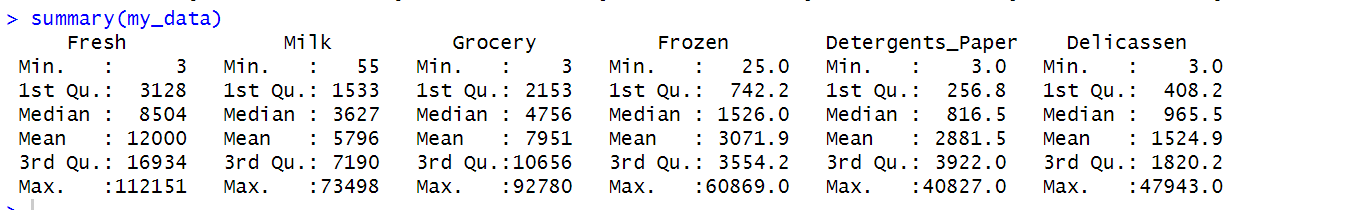
This code loads the factoextra package, which provides functions for visualizing clustering results.

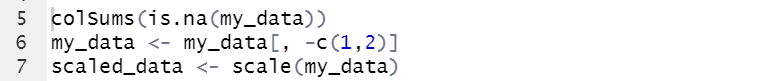
* **my\_data <- read.csv("C:/Users/Asus/Desktop/ds/wholesale\_customers\_data.csv",header = TRUE, sep = ",")**

This code reads in the wholesale\_customers\_data.csv file from the specified file path and stores it in a variable called "my\_data".

* **summary(my\_data)**

This code provides a summary of the "my\_data" dataset, which includes basic descriptive statistics for each variable.





* **colSums(is.na(my\_data))**

This code checks if there are any missing values in the "my\_data" dataset and returns a vector containing the number of missing values in each variable.

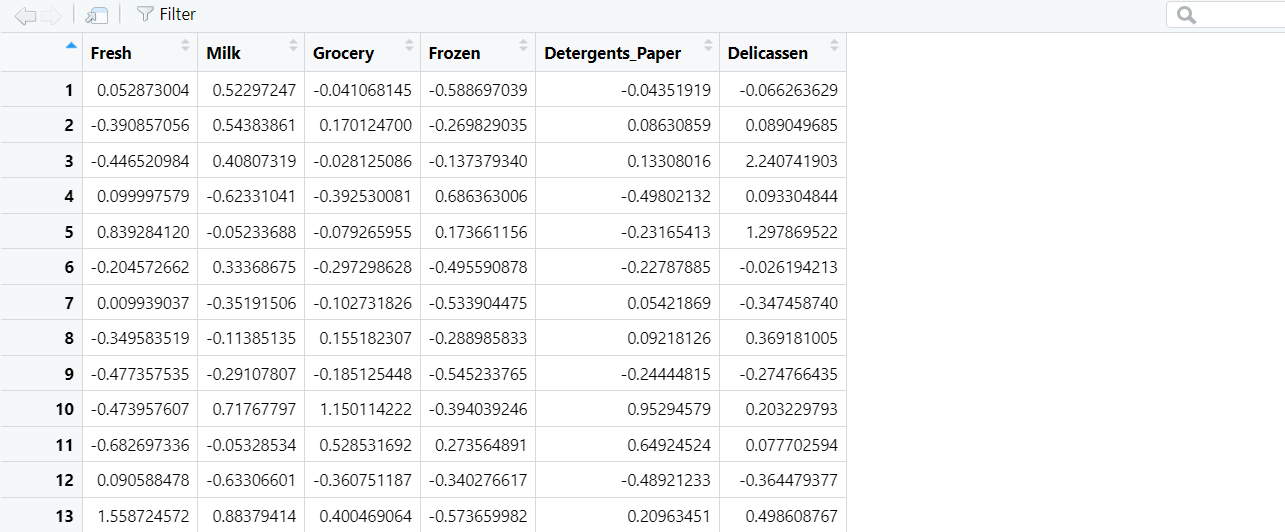


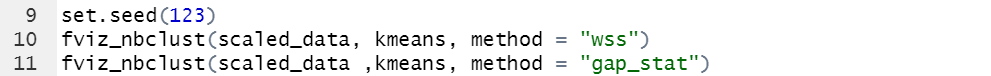
* **my\_data <- my\_data[, -c(1,2)]**

This code removes the first two columns from the "my\_data" dataset using negative indexing.

* **scaled\_data <- scale(my\_data)**

This code scales the "my\_data" dataset by centering and scaling each variable to have a mean of 0 and standard deviation of 1. This is important for ensuring that variables with different scales do not dominate the clustering results.

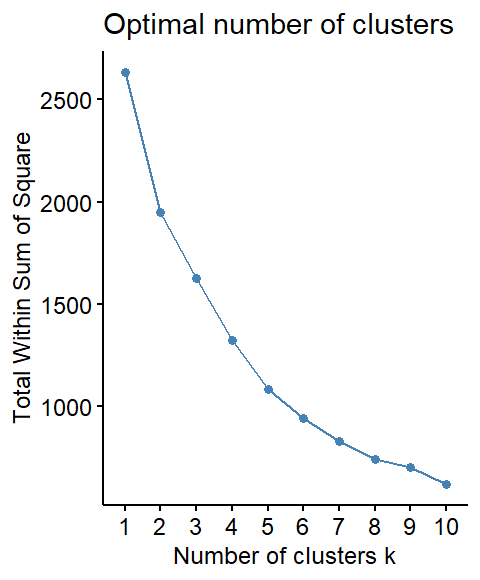




* **set.seed(123)**

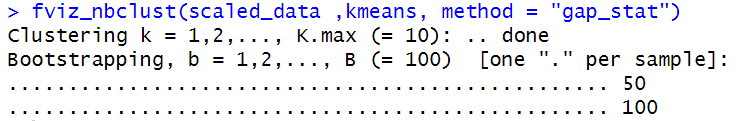
**fviz\_nbclust(scaled\_data, kmeans, method = "wss")**

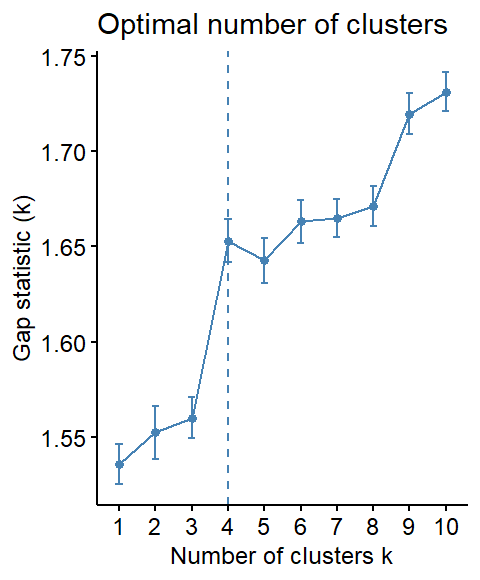
This code sets the random seed to ensure reproducibility and uses the fviz\_nbclust function from the factoextra package to visualize the within-cluster sum of squares (WSS) for different values of k (number of clusters) using the kmeans algorithm.

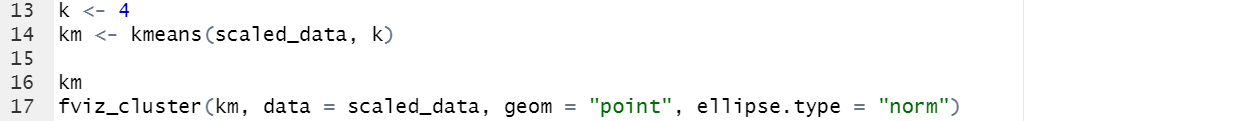


* **fviz\_nbclust(scaled\_data ,kmeans, method = "gap\_stat")**

This code uses the fviz\_nbclust function to visualize the gap statistic for different values of k using the kmeans algorithm. The gap statistic measures the difference between the within-cluster dispersion and a null reference distribution of the data, and is another method for choosing the optimal number of clusters in K-means clustering.







* **k <- 4**

**km <- kmeans(scaled\_data, k)**

This code sets the number of clusters to 4 and uses the kmeans function to perform clustering on the scaled data.

* **km**

**fviz\_cluster(km, data = scaled\_data, geom = "point", ellipse.type = "norm")**

km, is an R object that contains the results of the K-means clustering analysis performed on the scaled data. package to visualize the results of the K-means clustering analysis.

