

**Department of Computer Science**

**American International University-Bangladesh**

**Final Term Report**



 Course Name:  **INTRODUCTION TO DATA SCIENCE**

“A Report on Data Science Final Project”

**Supervised By:**

**Tohedul Islam**

**Submitted By:**

**Sajjad Hossein**

ID: 20-43639-2

Section: [E]

.

**Project Title: Applying K-means Clustering Algorithm to The Selected Unsupervised Data Set.**

**Dataset Overview:**

This project's dataset, which describes nine socioeconomic and health factors from a list of 167 countries.

Here,

Country: Name of the Country.

Child mort: Deaths of children under the age of five per 1,000 live births country: Name of the country.

Exports: The amount of goods and services exported per person. As a percentage of the GDP per person.

Health: Spending on health care overall and per person.

Imports: Imports of goods and services expressed as a percentage of GDP per capita. As a percentage of the GDP per person.

Income: Net income per individual.

Life Expectancy: The average number of years a newborn would live if current mortality patterns continued.

Inflation: The measurement of the annual growth rate of the Total Gross Domestic Product.

Total\_fer: The total number of children each woman would have if the current age-fertility rates persisted

Gdpp: is the per-capita GDP. Calculated by dividing the total population by the total GDP.

**The Software and Language utilized for this project:**

The language we are going to use to conduct the project is R and the software we’re going to use to process data and shape data is R Studio. R language is designed specifically for statistical computing and analysis. R has powerful tools for data visualization, which helps to analyze and interpret complex data sets more easily. It allows us to create high-quality graphs, charts, and other visual representations of data. And R Studio is a powerful and easy way to interact with R programming. R Studio has built-in data visualization tools that allow users to create high-quality graphs, charts, and other visual representations of data.

**Importing Data from CSV File**

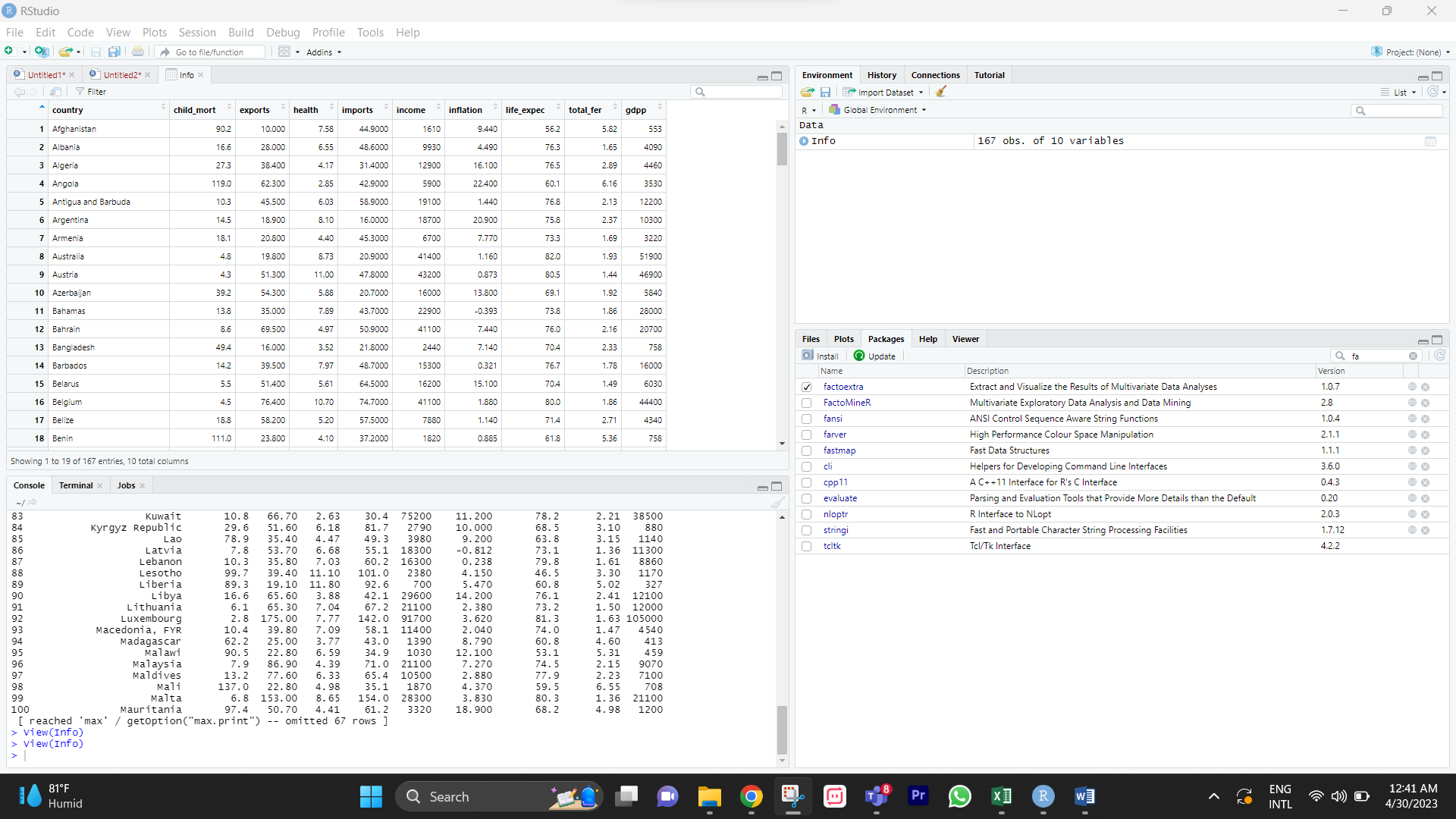
**CODE:**

# Calling CSV File

**Info<- read.csv("D:/C\_Dataset\_Final.csv",header = TRUE,sep=",")**

**Info**

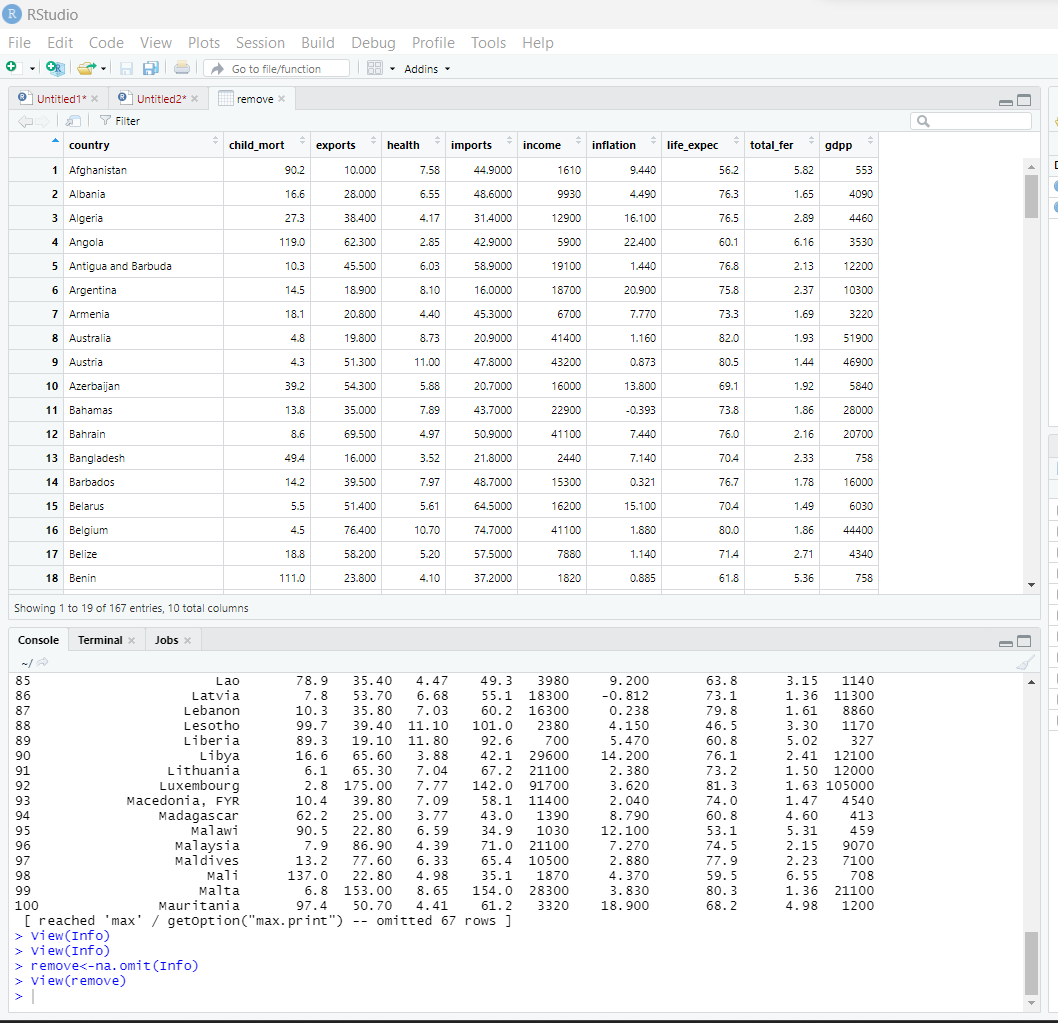
**Output:**



**Removing Null Value:**

**Code:**

**remove<-na.omit(Info)**



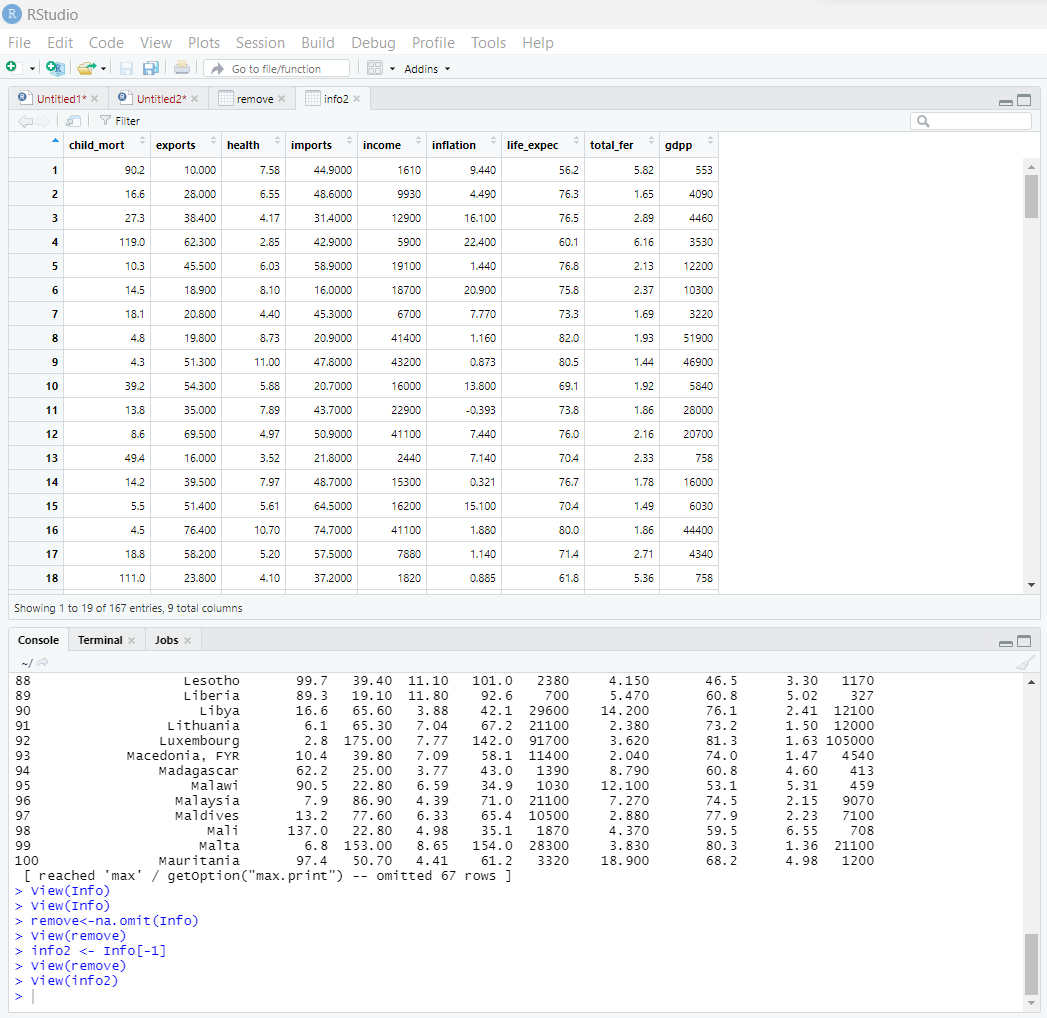
As The Data Set Has No Null Value So there’s No Need of Replacing Any Missing Value.

**Removing Column No 1**

**Code:**

**info2 <- Info[-1]**

**Output:**

****

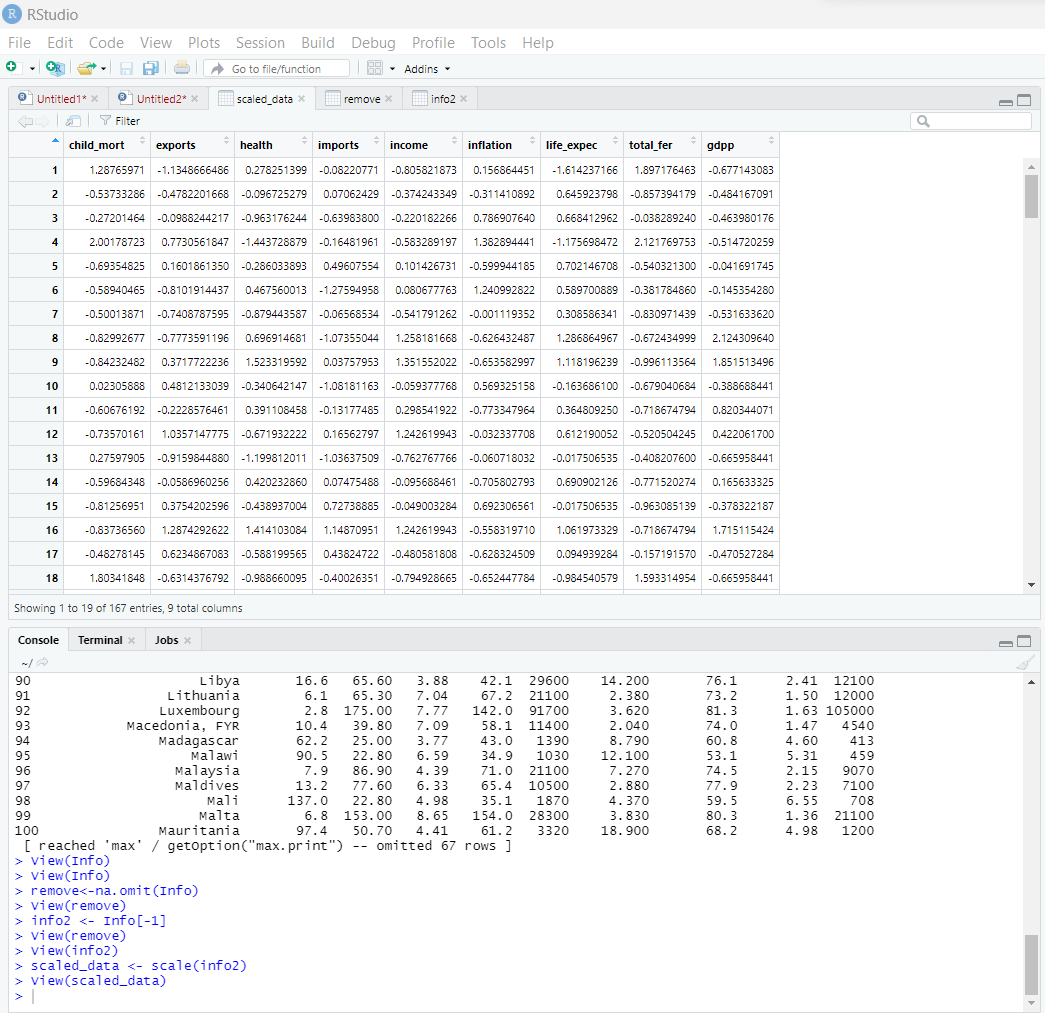
Here, we can see the column name (country) which refer to the name of the countries has been removed because of not being numerical.

**Scaling Data From The Following Dataset**

**Code:**

**scaled\_data <- scale(info2)**

**Output:**



We can see that all the values in the dataset has been scaled within 0-1

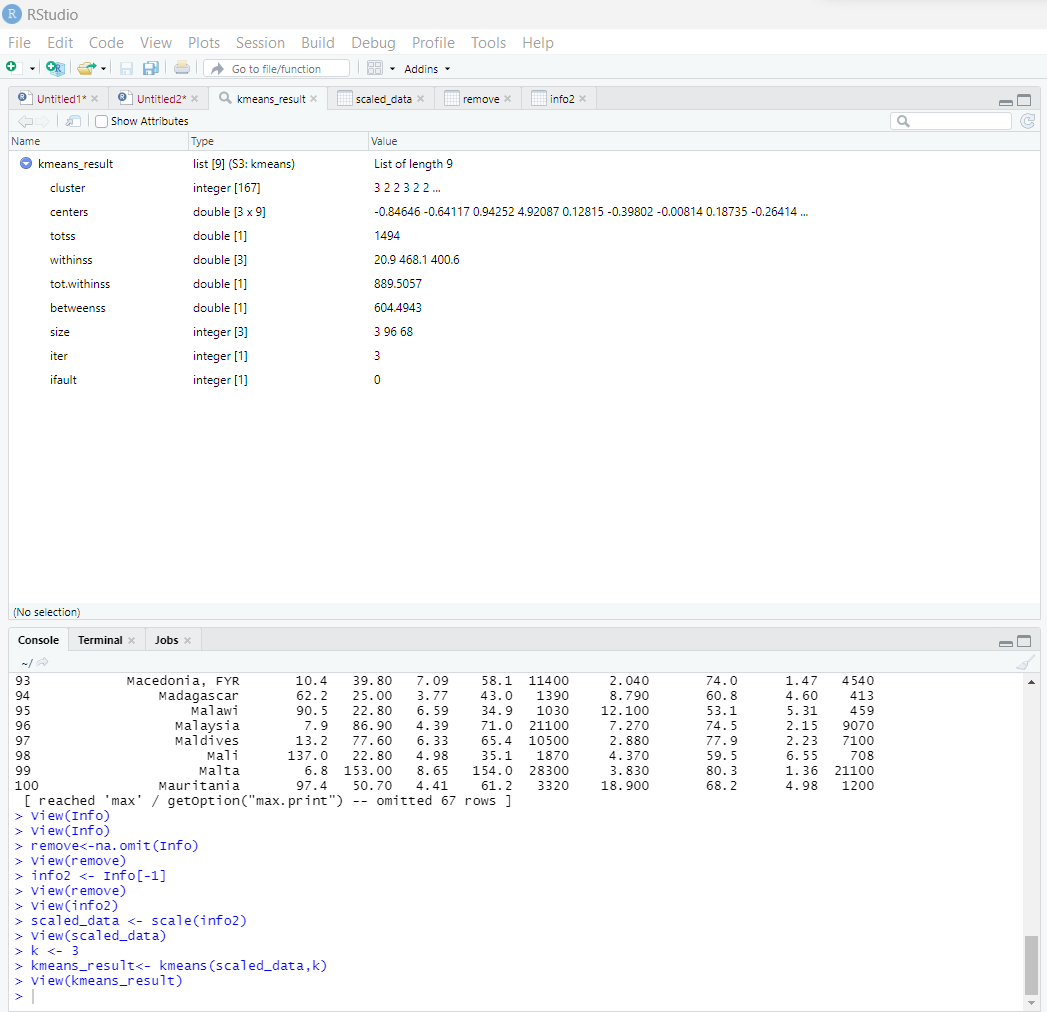
**K means Clustering Algorithm**

**Code:**

**k <- 3**

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

**Output:**



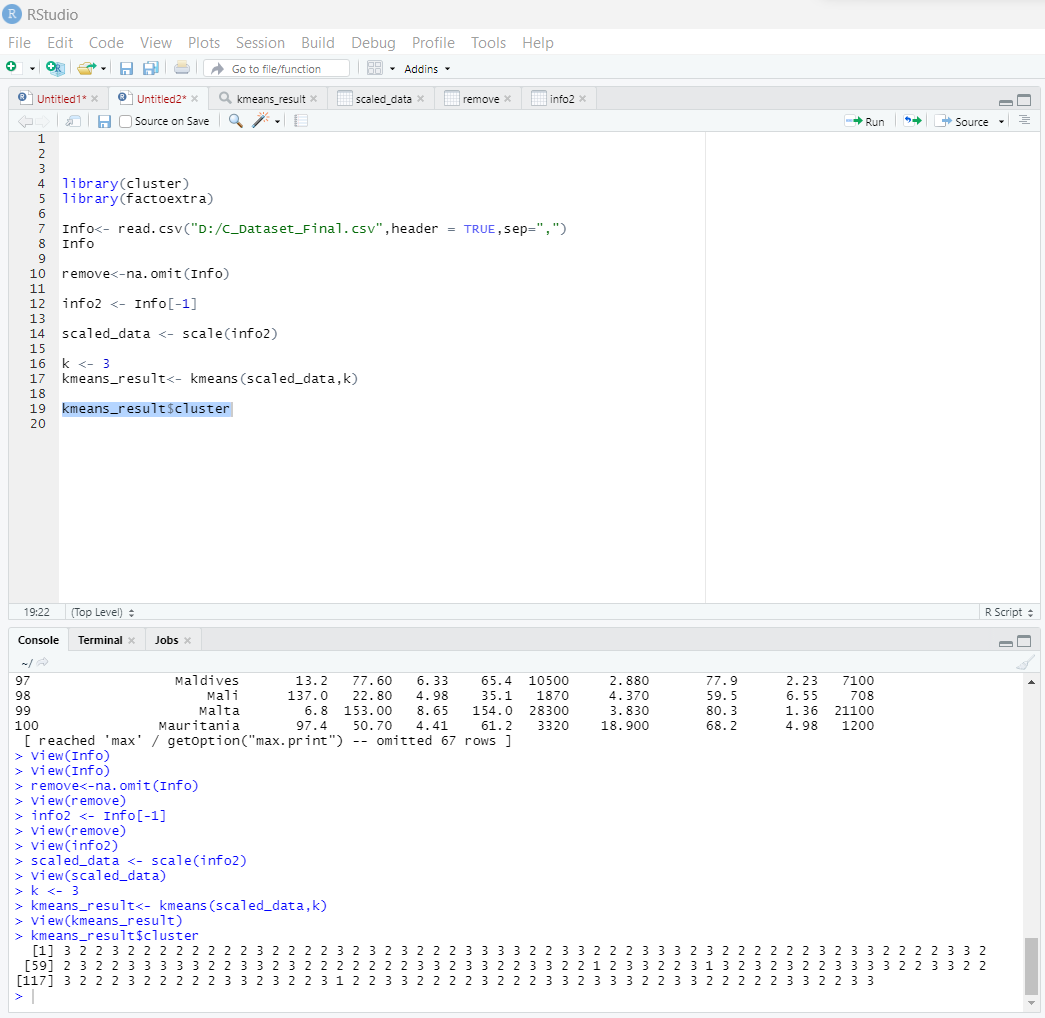
Here, k = 3, that means there’ll be 3 cluster

**View the Cluster Assignments for each observation**

**Code:**

**kmeans\_result$cluster**

**Output:**

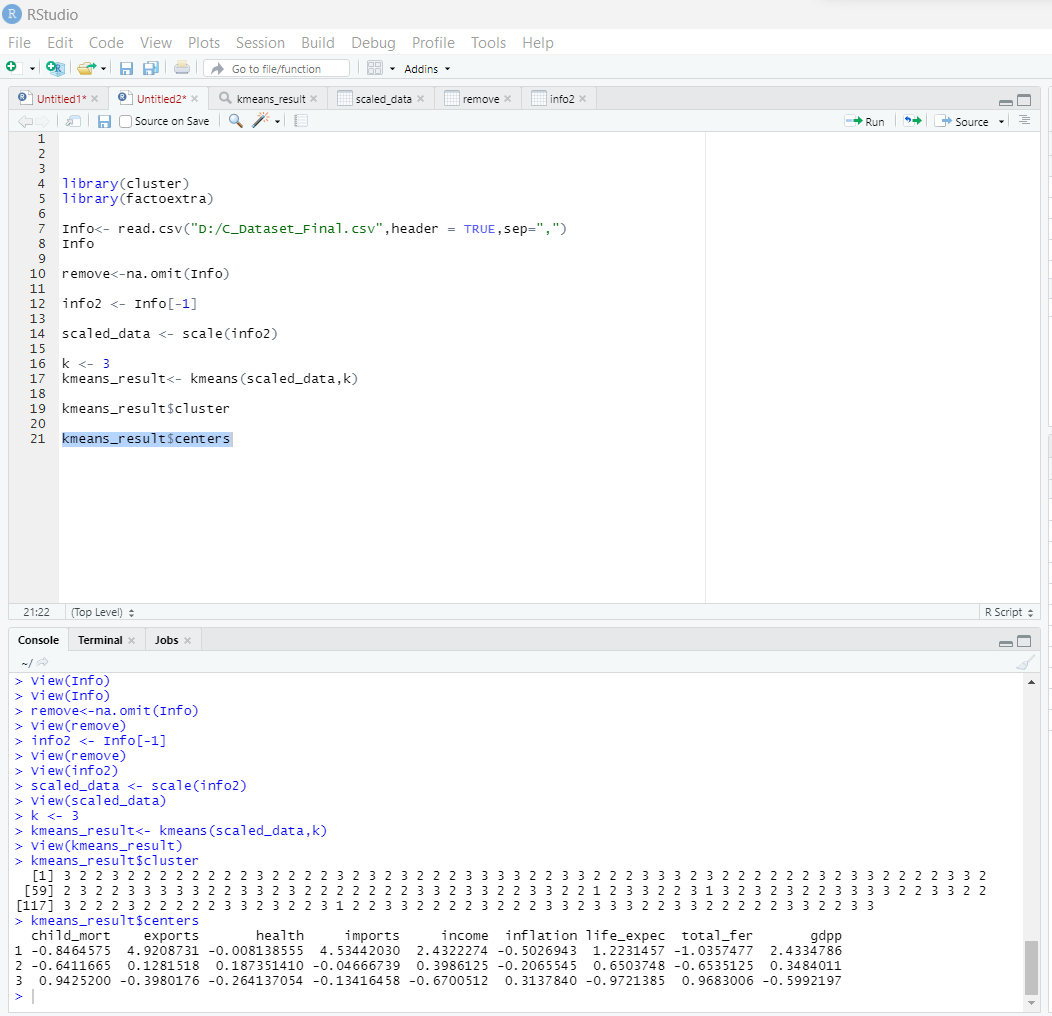


**View the Center Coordinates for each Cluster**

**Code:**

**kmeans\_result$centers**

**Output:**



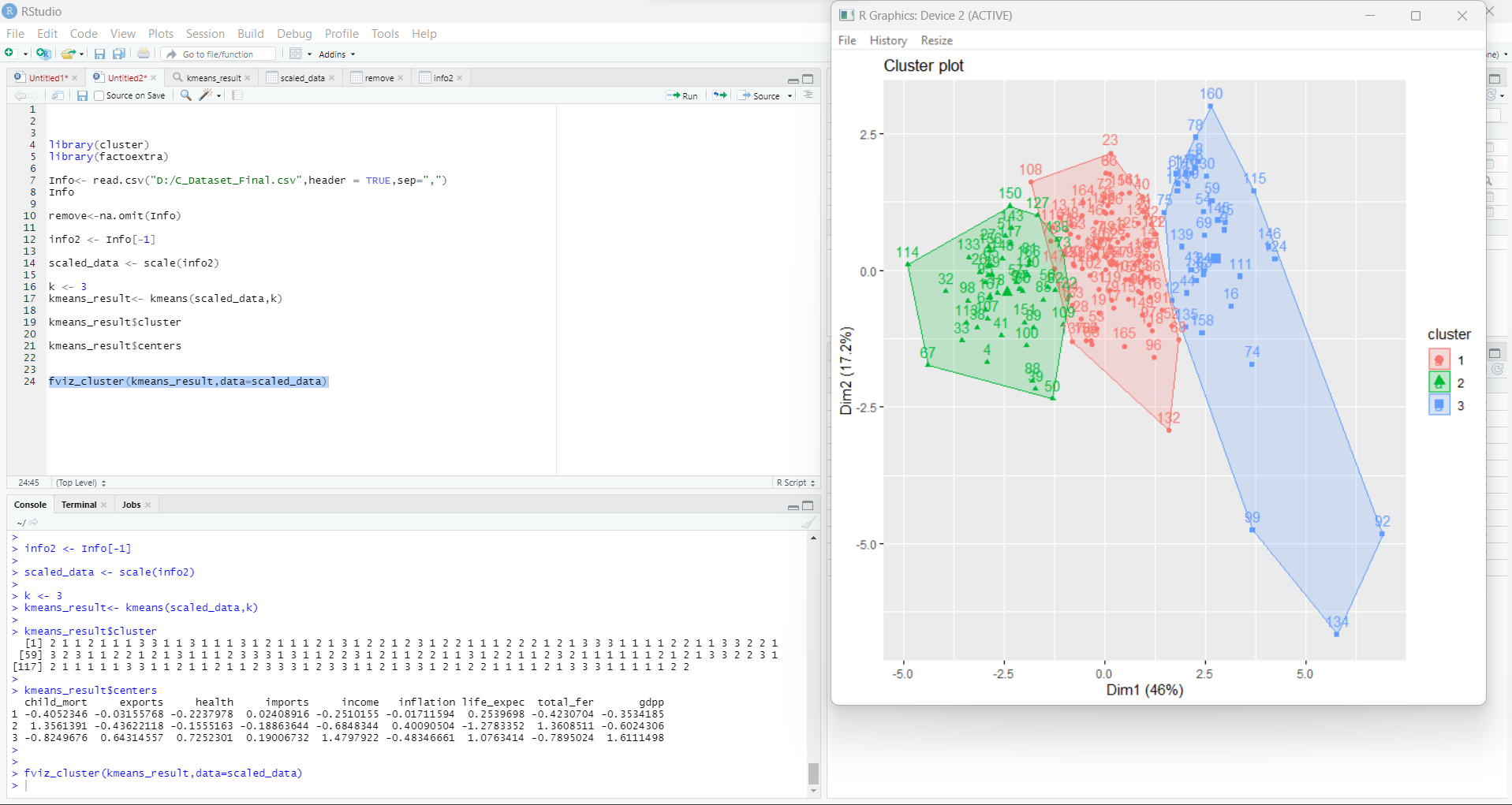
Its likely refers to the centroids of the clusters that were computed by the algorithm.

**Showing the Plots**

**Code:**

**fviz\_cluster(kmeans\_result,data=scaled\_data)**

**Output:**



Here we can see the three cluster plots

**Conclusion:**

In conclusion, this dataset provides a valuable resource for exploring patterns and relationships within socio-economic indicators across different countries. By applying unsupervised learning techniques, we can gain insights into which countries have similar characteristics and potentially identify factors that contribute to differences in socio-economic development.