Ecommerce project with R: Project 3

By: Bhupendra Nirgun

Contents

[1. Read the csv file : 3](#_Toc527387793)

[2. Visualize the data structure 3](#_Toc527387794)

[3. Complete source code : 9](#_Toc527387795)

[4. Complete console output : 10](#_Toc527387796)

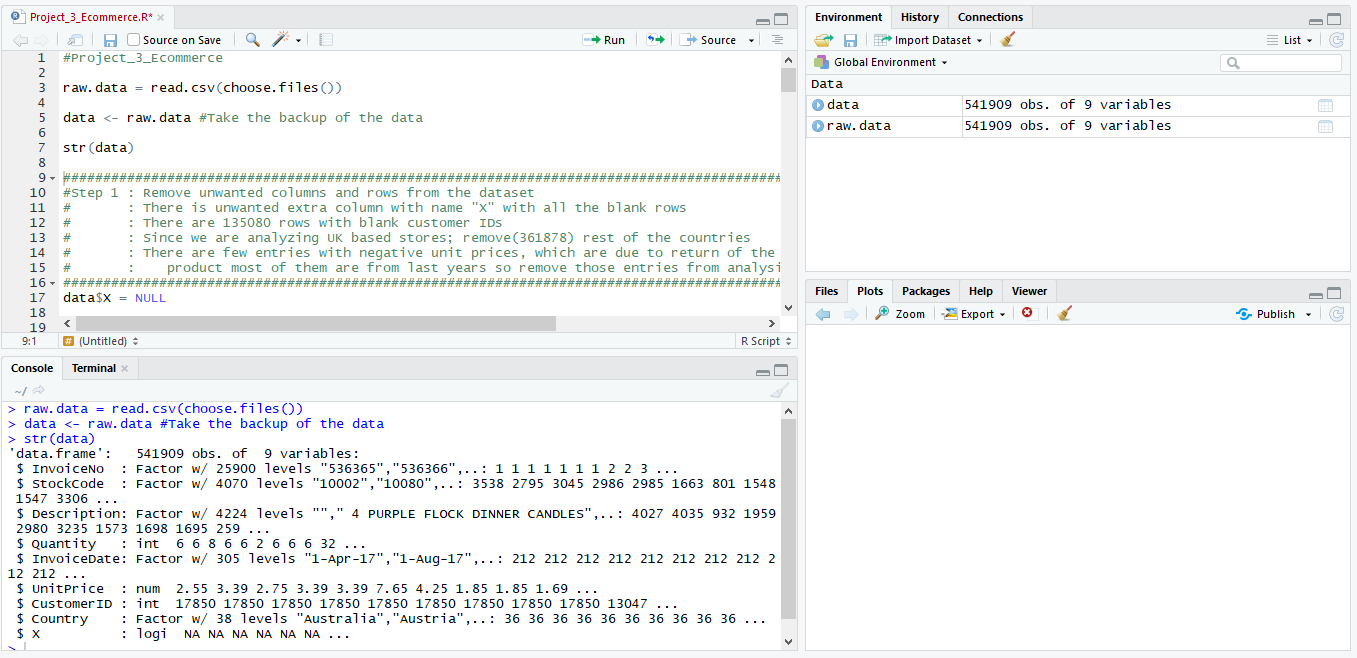
## Read the csv file :

Read the Project 2\_Dataset.csv

## Visualize the data structure

This can be visualized using various forms some of them are

Structure, summary, glimpse, class, head; I have used tibble and summary and below are the screenshots of the same.



##############################################################################

#Step 1 : Remove unwanted columns and rows from the dataset

# : There is unwanted extra column with name "X" with all the blank rows

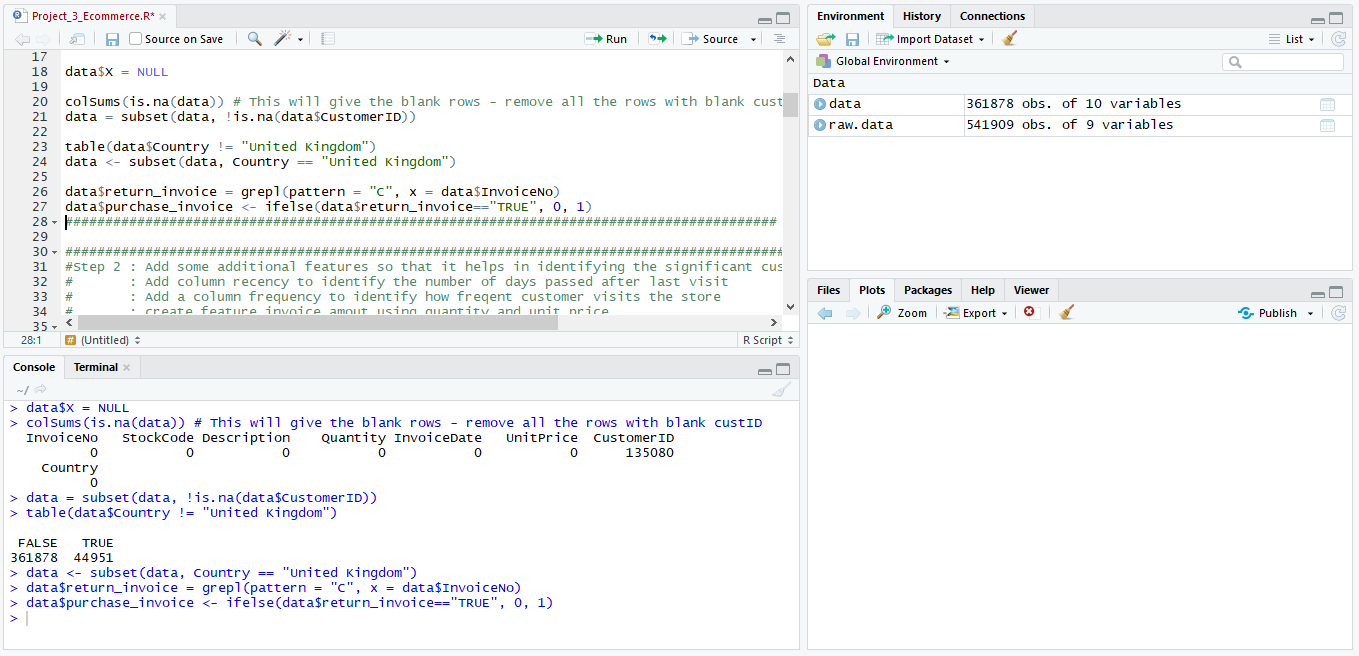
# : There are 135080 rows with blank customer IDs

# : Since we are analysing UK based stores; remove(361878) rest of the countries

# : There are few entries with negative unit prices, which are due to return of the

# : product most of them are from last year’s so remove those entries from analysis

##############################################################################



##############################################################################

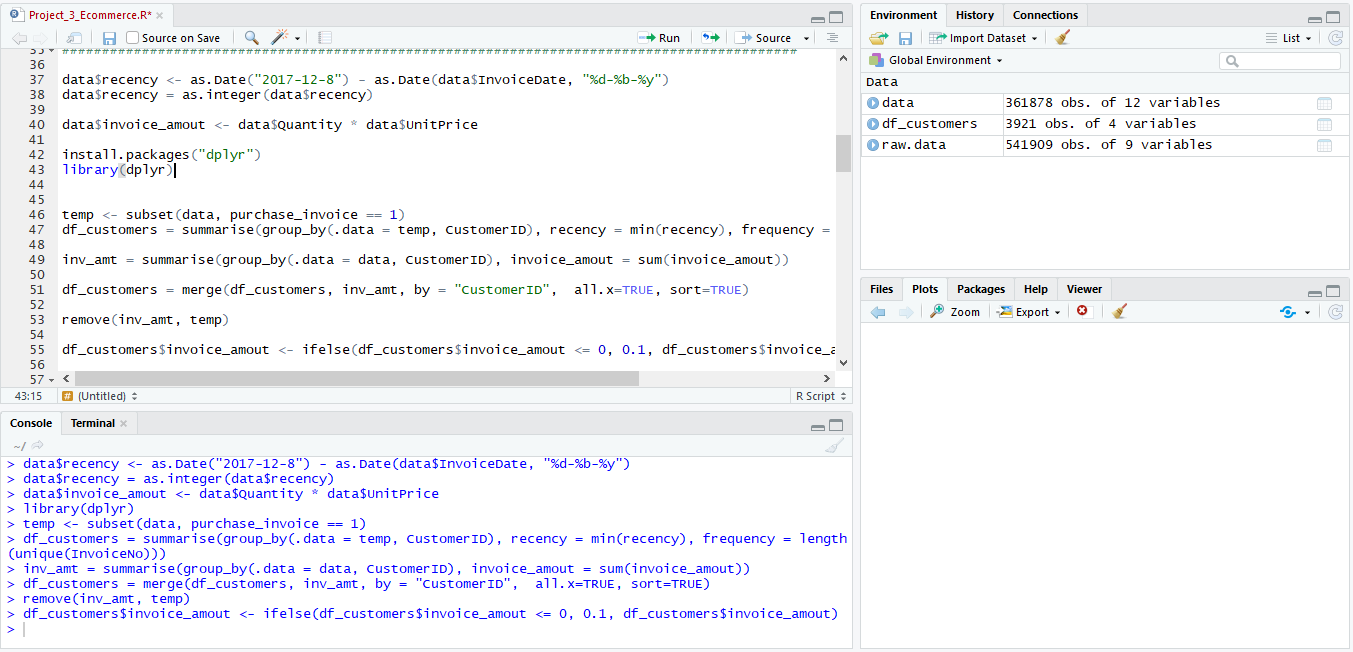
#Step 2 : Add some additional features so that it helps in identifying the significant customers

# : Add column recency to identify the number of days passed after last visit

# : Add a column frequency to identify how freqent customer visits the store

# : create feature invoice\_amout using quantity and unit price

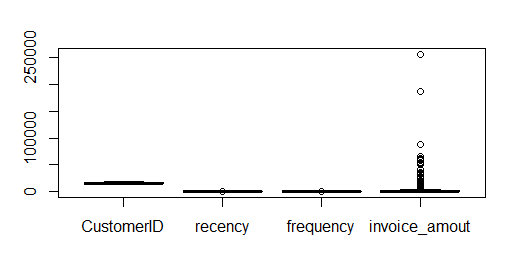
##############################################################################



##############################################################################

#Step 3 : customer data frame looks to be right skewed, scale the data first using logs

##############################################################################



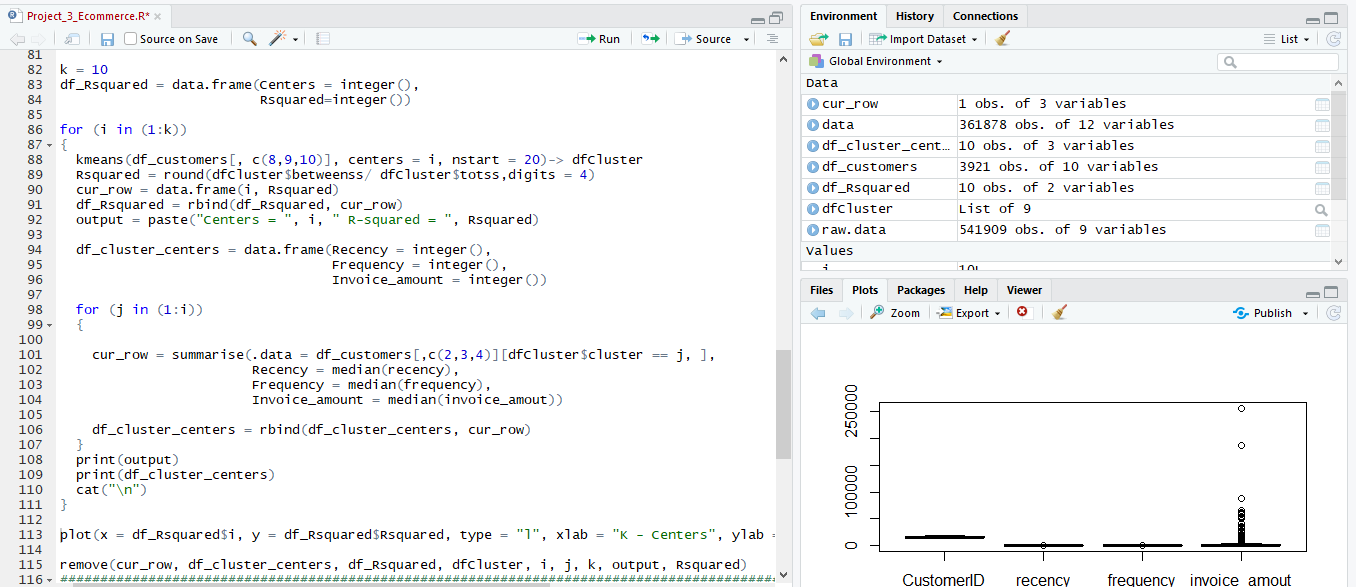


##############################################################################

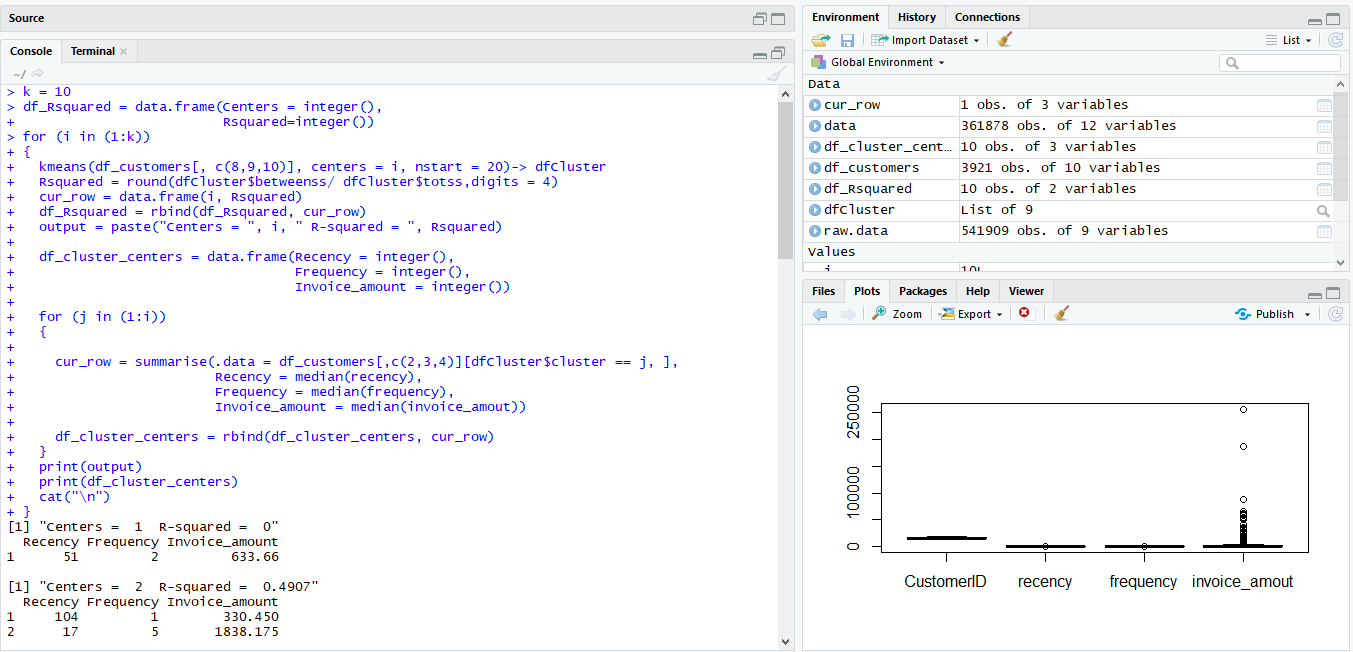
#Step 4 : Execute k-means cluster algorithm in loop for k=10 and identify the best

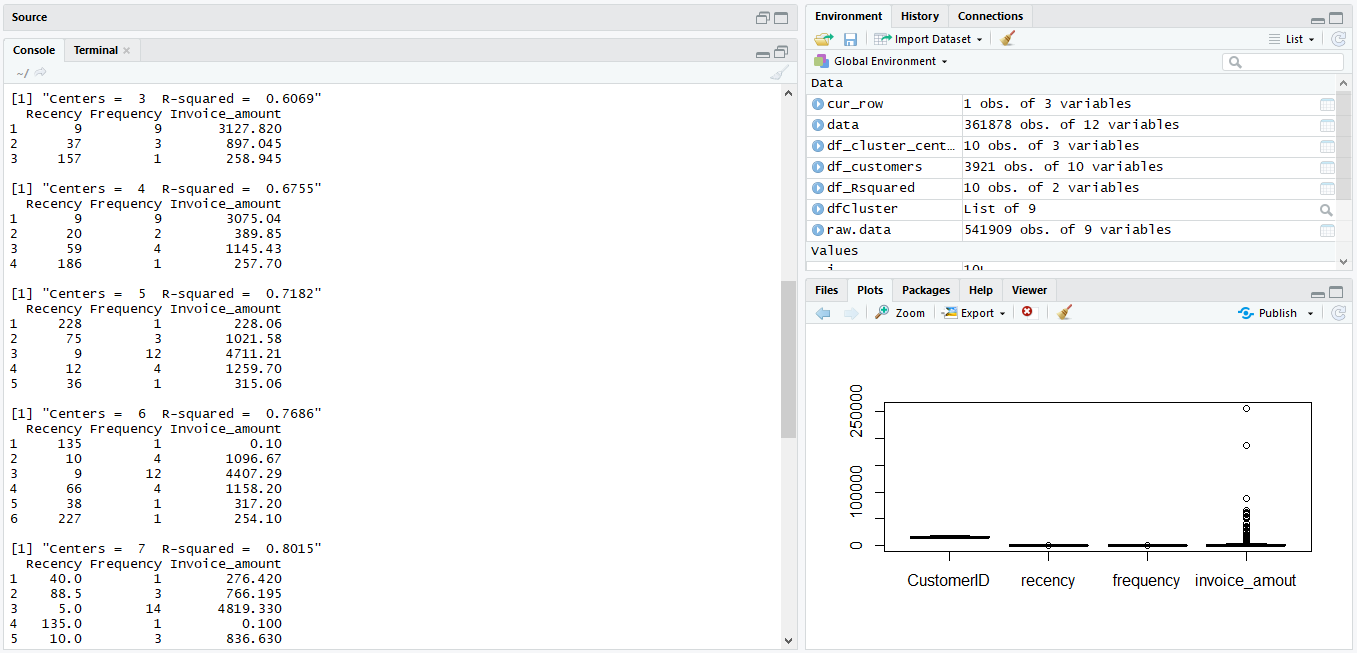
# cluster size

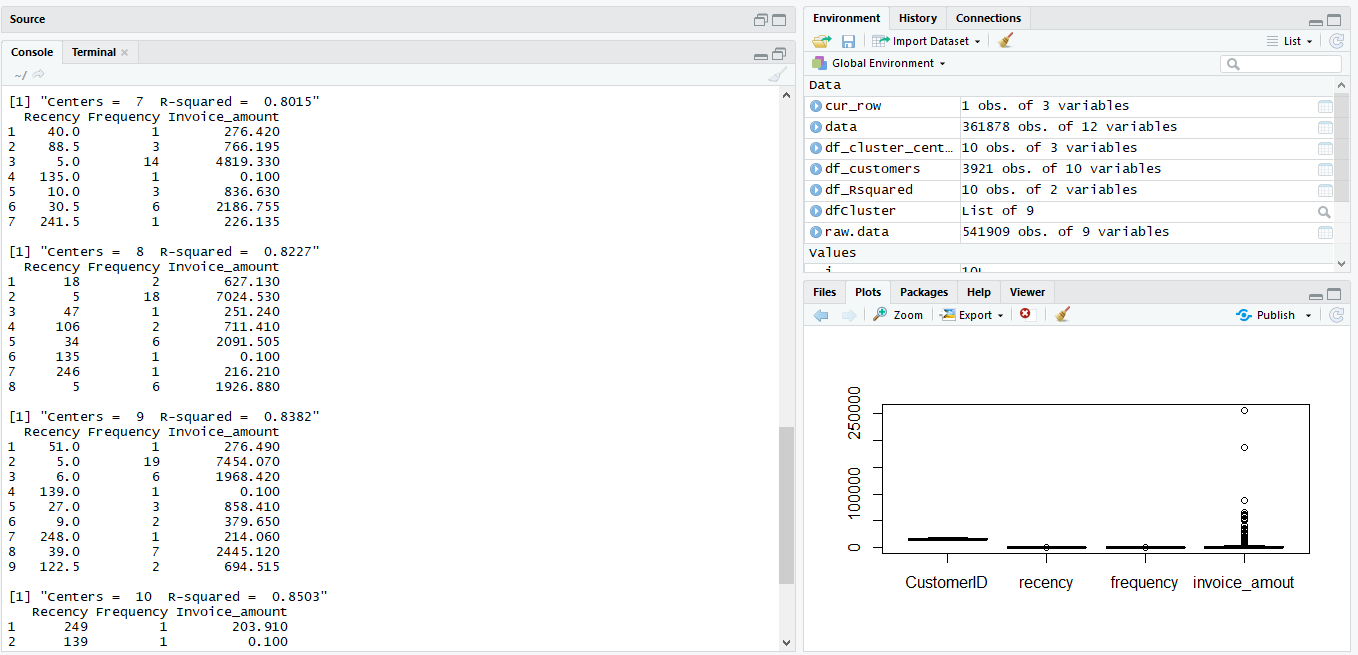
##############################################################################

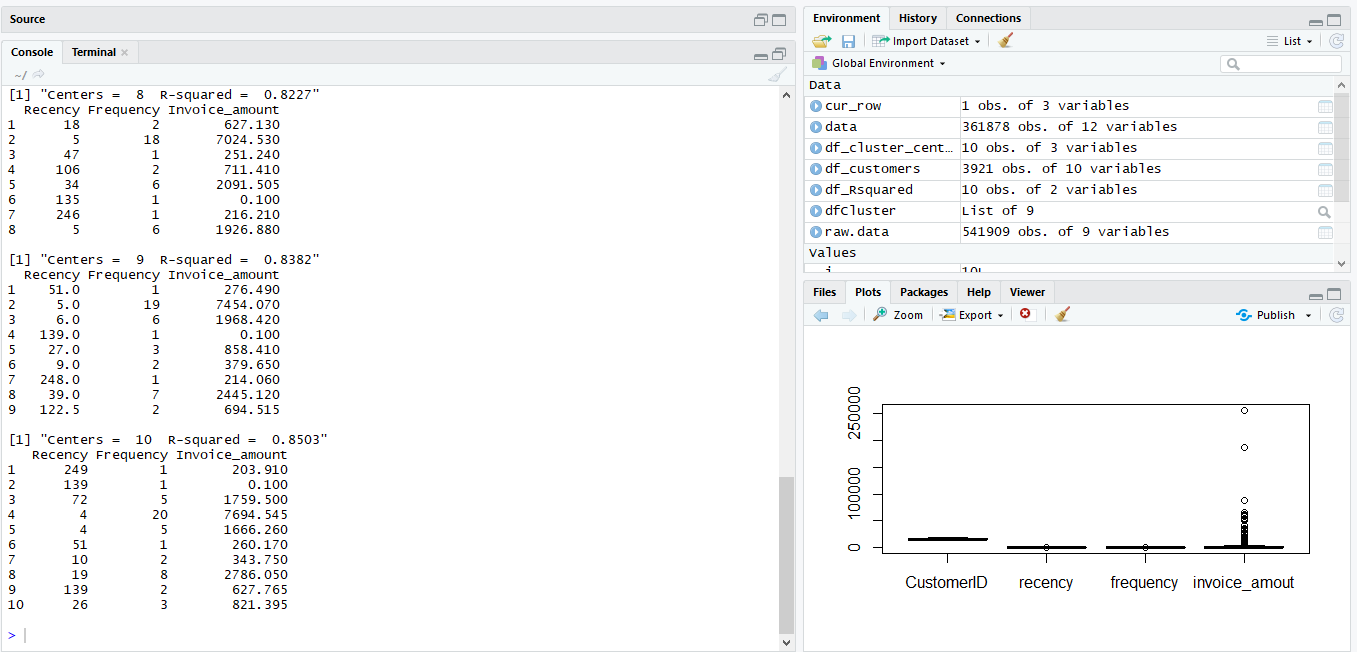


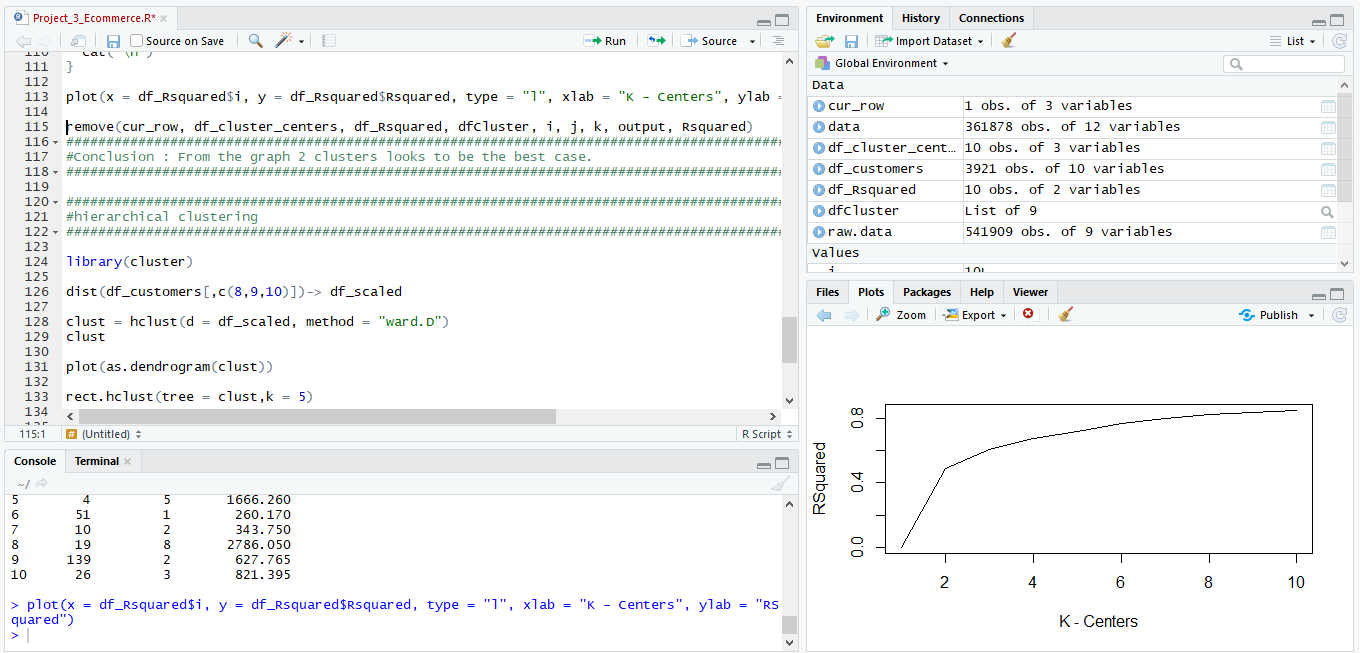
Console











##############################################################################

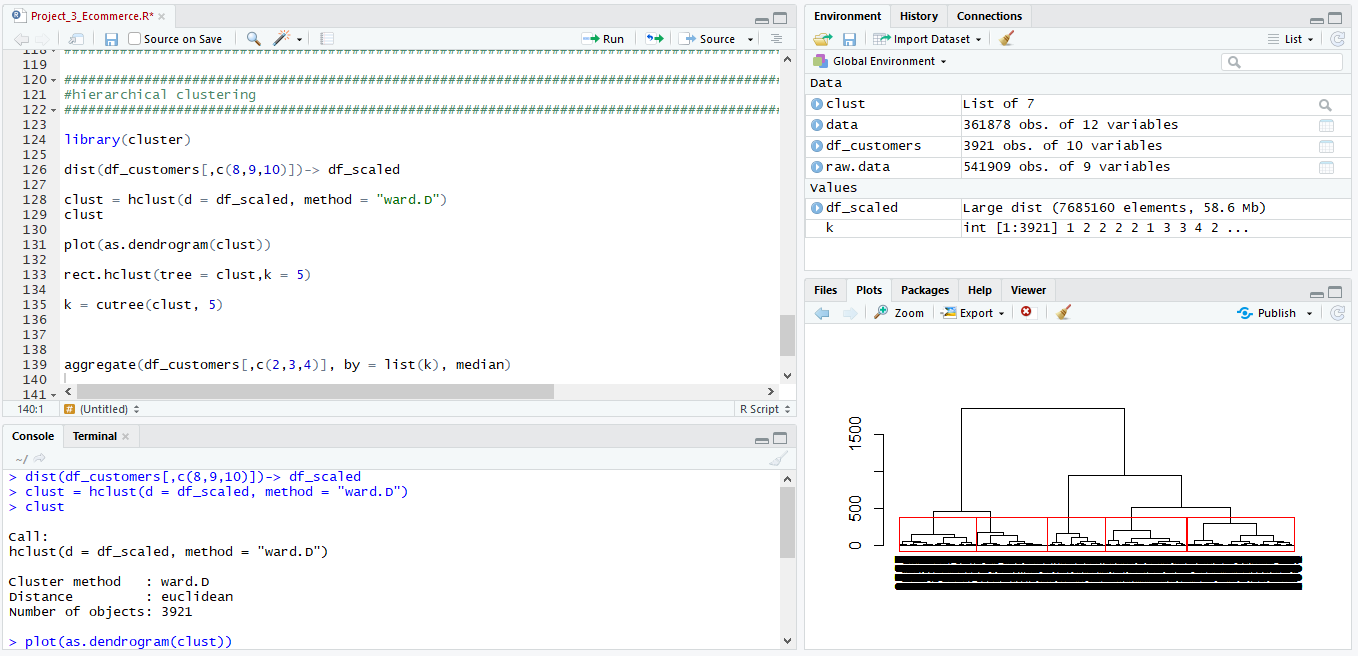
#Conclusion : From the RSquared value graph 2 clusters looks to be the best case.

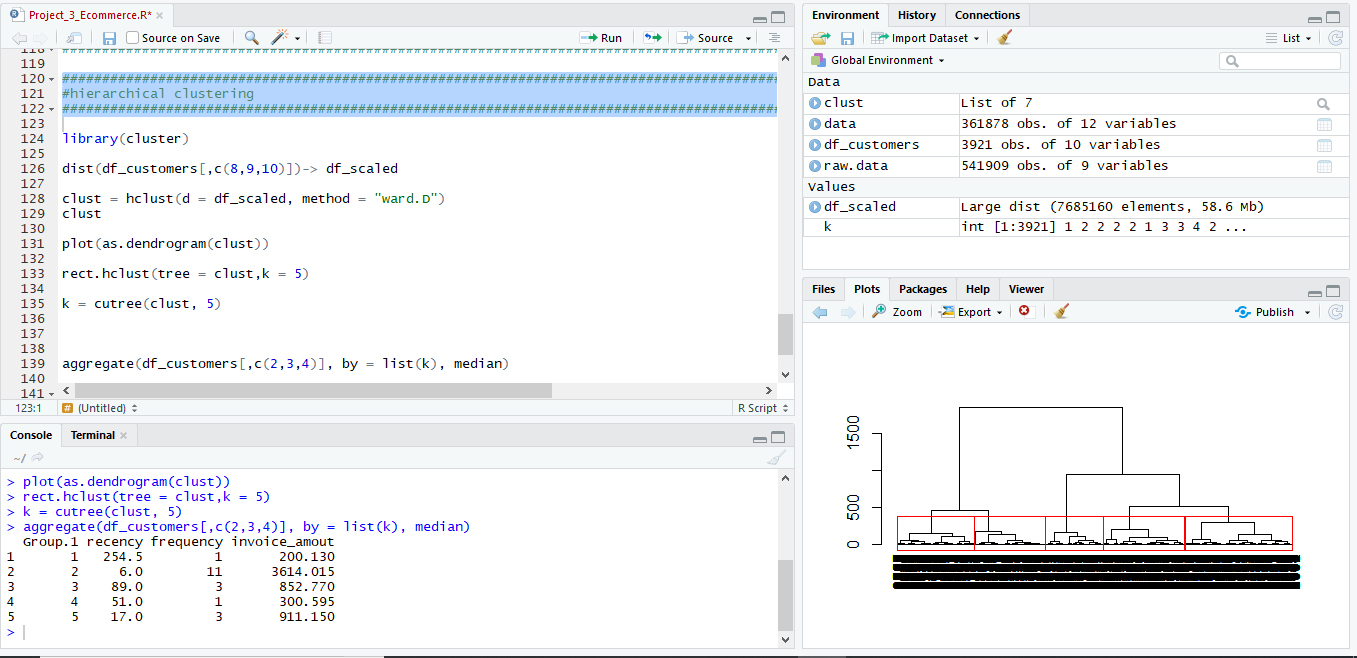
##############################################################################

##################################################################################

#hierarchical clustering

##################################################################################





##################################################################################

#Looking at the graph 5 clusters seems to be best case.

##################################################################################

## Complete source code :



## Complete console output :

