**Advance Analytical Methods Assignment**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Prepared By** | **SID** | **TAID** |
| **Name** | Rajib Mandal | 118310 | TA17002 |
| **Submitted Date** | 12-Mar-2018 | | |

Table of Contents

[1. Data Set Information 3](#_Toc508570485)

[1.1 Attribute Information 3](#_Toc508570486)

[2. Project Objective 3](#_Toc508570487)

[3. Data Pre-Processing 3](#_Toc508570488)

[4. k mean clustering 4](#_Toc508570489)

[4.1 Data Visualizaion 5](#_Toc508570490)

[4.2 Interpretations of the clusters 6](#_Toc508570491)

[5. Hierarchical Clustering 11](#_Toc508570492)

[5.1 Interpretations OF THE clusters 12](#_Toc508570493)

[6. ANNEXTURE 14](#_Toc508570494)

# Data Set Information

The dataset contains 406,830 customer transaction data for a retail ecommerce company for the period 1-DEC-2010 to 9 -DEC-2011.

Here is the summary.

* There were 22, 190 valid transactions in total for the above period.
* There were 4,372 unique customers for the above period.
* There were unique 3,684 product associated with 22,190 transactions.
* There were 37 countries associated with transactions.

## Attribute Information

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Type | Data Type | Data Format |
| InvoiceNo | Invoice number; a 6-digit Alpha numeric number uniquely assigned to each  transaction | Nominal | Alpha numeric  e. g541431  C541433- credit note |
| StockCode | Product (item) code; a 5-digit number uniquely assigned to each distinct product | Nominal | Alpha numeric  e.g 23166, 85167B |
| Description | Product (item) name; | Nominal | Text  e. COLOUR GLASS. STAR T-LIGHT HOLDER |
| Quantity | The quantities of each product (item) per transaction | Numeric | Numeric e.g 12 |
| InvoiceDate | The day and time when each transaction was generated; | Nominal | m/d/yyyy h:mm  e.g  10/12/2011 10:23:00 AM |
| UnitPrice | Product price per unit in sterling; | Numeric | Numeric e.g  1.45 |
| CustomerID | Customer Id | Nominal | Integer |
| Country | Delivery address -country | Nominal | Text e.g Iceland |

The customer transaction dataset has 8 variables. The detail is given below along with data format.

# Project Objective

The main purpose of this analysis is to help the business better understand its customers and therefore conduct customer-centric marketing more effectively. Based on the Recency, Frequency, and Monetary model, customers of the business have been segmented into various meaningful groups using the k -means & hierarchical clustering algorithm.

# Data Pre-Processing

To conduct the required RFM model-based clustering analysis, the original dataset needs to be pre-processed.

The steps and relevant tasks involved in the data preparation is given below:

1. We have selected following six variables : InvoiceNo, StockCode , Quantity , UnitPrice ,InvoiceDate and CustomerID
2. Create an aggregated variable named monetery, by multiplying Quantity with UnitPrice , which gives the total amount of money spent per product / item in each transaction.
3. Create two variables Date and Time from InvoiceDate . This allows different transactions created by the same consumer on the same day but at different times to be treated separately.
4. Calculate the number of purchases by the customer in the last 12 months .This will give frequency
5. The number of months that have passed since the customer last purchased will give the recency. Consider DEC 2011(current purchase month) as 0 and previous month as 1 and continue till Dec 2010 as 12.It will give the recency .

The corresponding r code is given below

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

### Separate the variable InvoiceDate into

# two variables Date and Time

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

rfmdata$Date <- as.Date(rfmdata$InvoiceDate , format="%m/%d/%Y")

rfmdata$Time <- format(as.POSIXct(strptime(rfmdata$InvoiceDate,"%m/%d/%Y %H:%M",tz="")) ,format = "%H:%M")

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

### Calcultate amount for each transaction ####################################################

rfmdata$monetery<-as.numeric(rfmdata$Quantity \* rfmdata$UnitPrice)

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

### calculate recency ,frequency & monetary

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

retailtranformdata <- rfmdata %>%

group\_by(CustomerID) %>%

summarise(recency=12 \* as.numeric((as.yearmon("2011-12-09") - as.yearmon(max(Date)))),

frequency=n\_distinct(InvoiceNo), monetery= sum(monetery))

summary(retailtranformdata)

.

# k mean clustering

We wanted to identify whether consumers can be segmented meaningfully in the view of recency, frequency and monetary values. The k -means clustering algorithm was used for this purpose.

 k-means clustering aims to [partition](https://en.wikipedia.org/wiki/Partition_of_a_set) n observations into k clusters in which each observation belongs to the [cluster](https://en.wikipedia.org/wiki/Cluster_(statistics)) with the nearest [mean](https://en.wikipedia.org/wiki/Mean).

We have removed all the negative amount(monetary) from the dataset before applying min-max normalization. The data with negative amount could be the bad data.

The corresponding R code is given below

rfm <- rfm[rfm$monetery>0,]

minmaxstandard <- function(x)

{

(x-min(x))/(max(x)-min(x))

}

rfm$recency<-as.numeric(minmaxstandard(rfm$recency))

rfm$monetery <-as.numeric(minmaxstandard(rfm$monetery))

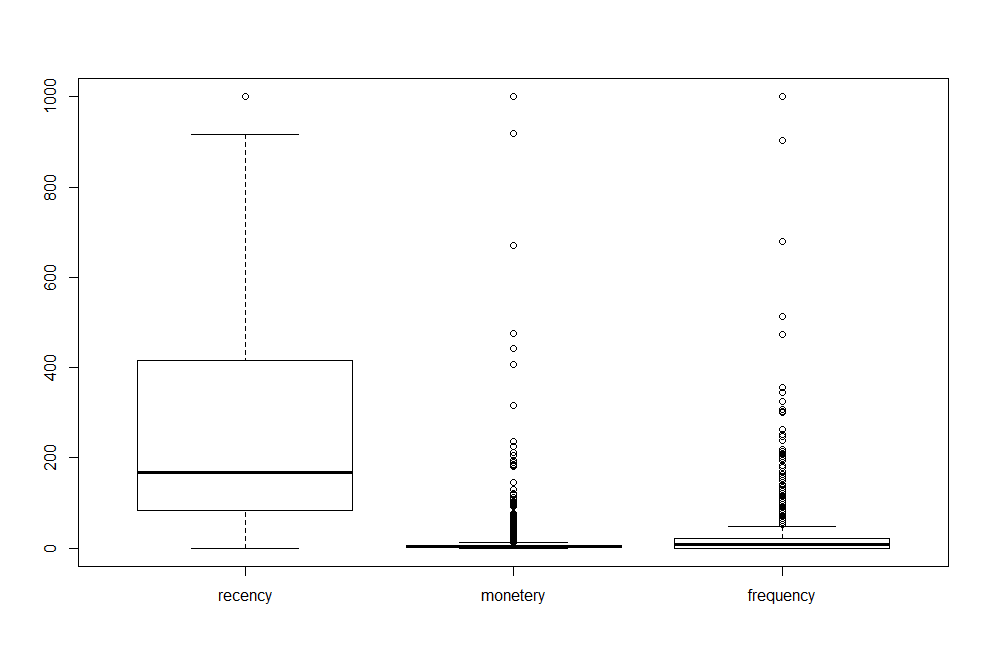
rfm$frequency <-as.numeric(minmaxstandard(rfm$frequency))

We have also removed CustomerID from the dataset

rfm$CustomerID<-NULL

## Data Visualizaion

Below is the box plot to find the outlier.

.

There are a few instances having quite different monetary ,recency and frequency values compared to most of the instances in the dataset. These instances are valid from the business point

of view as they are valid transaction records. However, they are outliers from the

data analysis point of view.

**Determine value for K(No of Cluster )**

To determine an appropriate value for k, the k-means algorithm is used to identify clusters For each value of k(2 to 25), the WSS is calculated.

The following R code loops through several k-means analyses for the number of centroids, k, varying from 2 to 25. For each k, the option nstart=25 specifies that the k-means algorithm will be repeated 25 times, each starting with k random initial centroids.

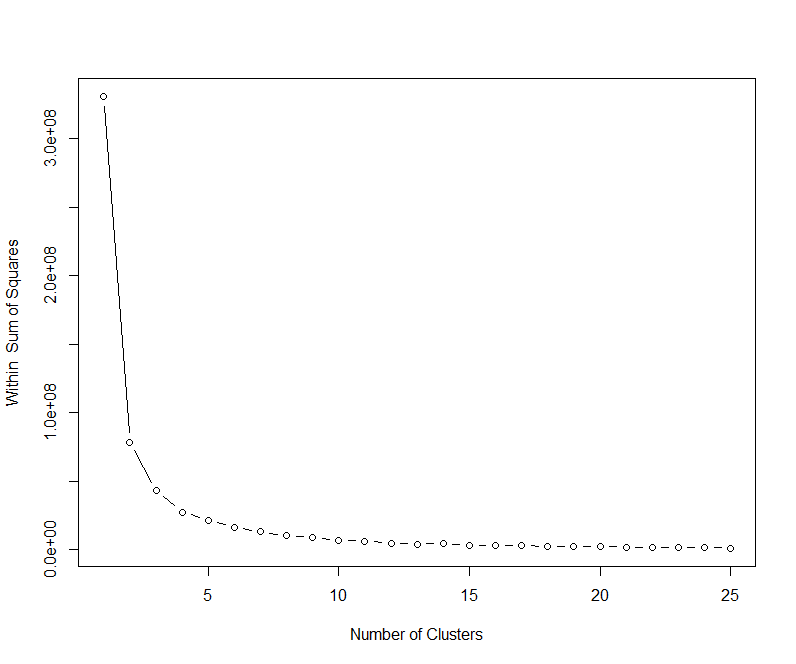
for ( i in 2:25)

{

wss[i]<-kmeans(rfm,i,iter.max =40,nstart=60)$tot.withinss

}

Each WSS is plotted against the respective number of centroids, 1 through 25. This plot is given below.



From the above plot we can see that WSS is greatly reduced when k increases from 6 onwards two.. Therefore, the k-means analysis will be conducted for K=5

The code for K-means clustering with 5 cluster is given below

km<-kmeans(rfm,5,iter.max =40,nstart = 50)

km$size



The size of the 5 clusters are 460,304,1949,1032,572.

## Interpretations of the clusters

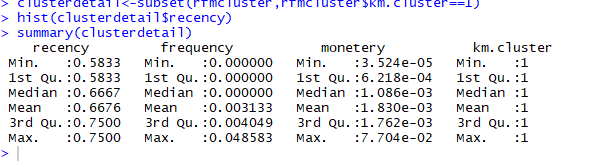
Each of the five clusters contain a group of consumers that have distinct features.

The clustering and segment results with five clusters are shown.

The distribution of customer across the cluster is given below.

|  |  |  |
| --- | --- | --- |
| Cluster | Total records in the cluster | % 0f overall population |
| 1 | 460 | 10.66 |
| 2 | 304 | 7.04 |
| 3 | 1949 | 45.15 |
| 4 | 1032 | 23.91 |
| 5 | 572 | 13.25 |

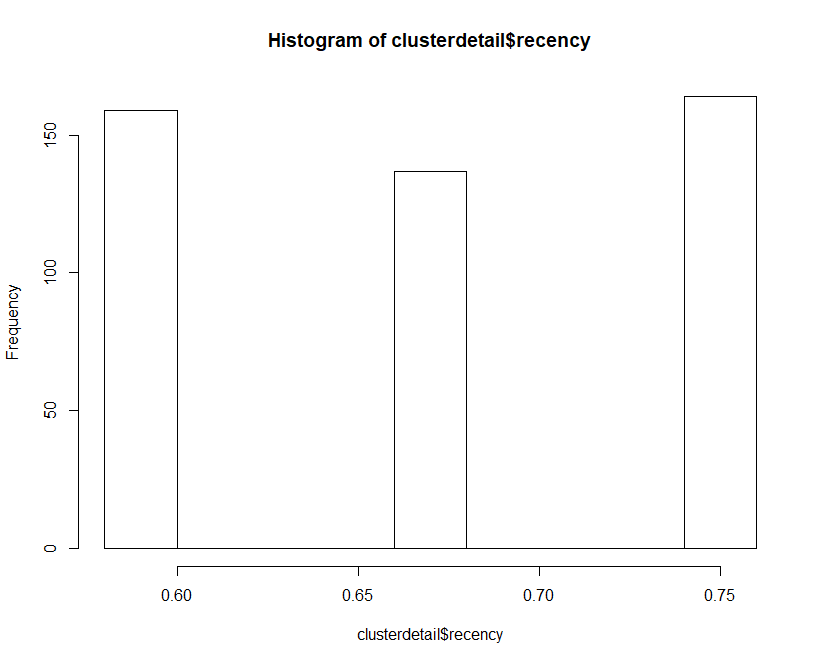
**Cluster 1**



There are 460 customers in cluster 1 which constitute 10.55 % of overall consumer.

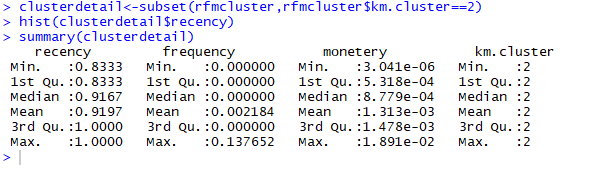
This group can be categorized as average recency, low frequency and low monetary with a low spending per consumer. The compny should focus this cluster

.



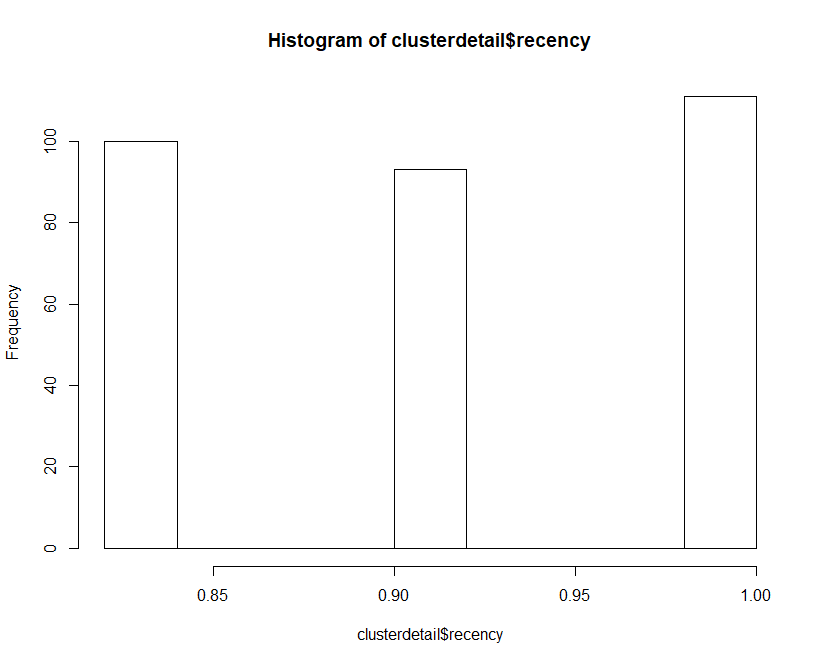
From the above histogram for recency VS frequency we can find that, the customer of this cluster purchased throughout the year.

**Cluster 2**



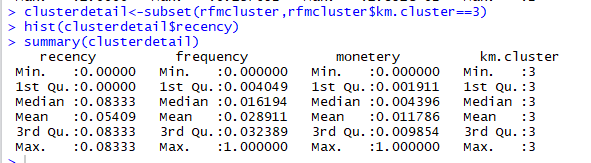
There are 304 customer in cluster 2 which constitute 7.04 % of overall consumer.

This group of consumers can be categorized as very high recency, average frequency and low monetary with a low spending per consumer. This seems to be least profitable group.



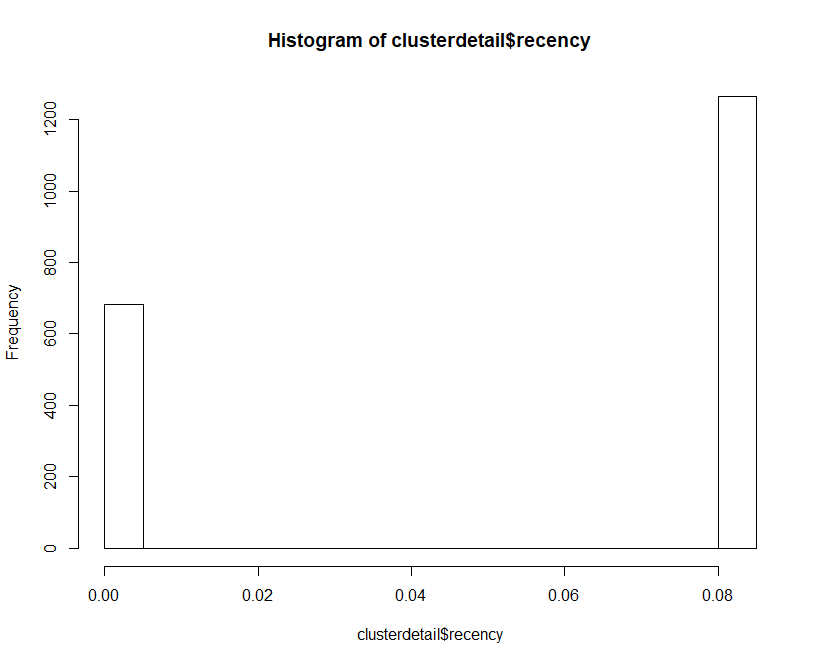
From the above histogram for recency we can find that, the customer of this cluster purchased through the year. There was spike in purchase in end of the year

**Cluster 3**



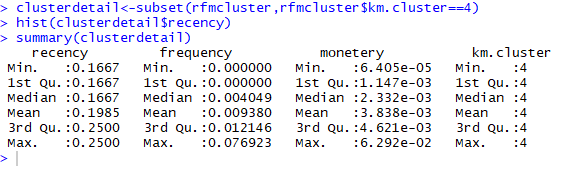
The cluster 3 is the largest size group with 1949 customers. It contains 45.15 % of overall consumer.

The consumer purchased quite often and as a result, spent a quite high amount of money. This group of consumers can be categorized as low recency (recent purchase), very high frequency and very high monetary with a high spending per consumer. This group seems to be very high profitable group.



From the above histogram for recency we can find that ,the customer did not purchase on quarter 2 and quarter 3. Also the customer purchased heavily on quarter 4 but on quarter 1 frequency of purchase reduced compare to quarter 4.

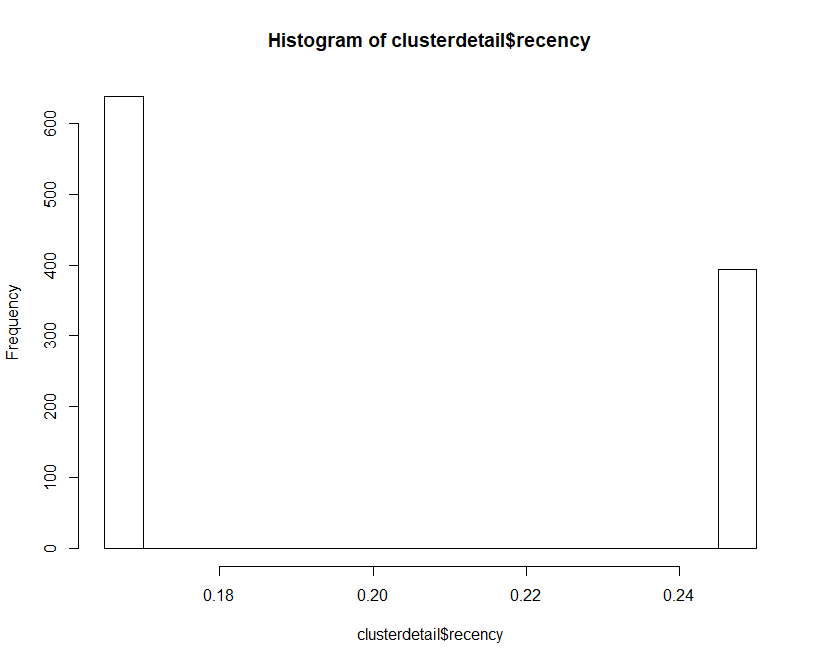
**Cluster 4**



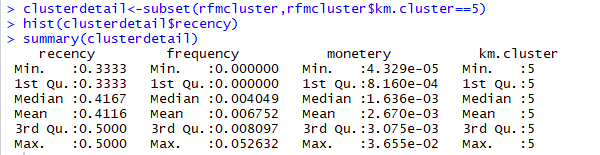
The cluster 1 have 1032 consumers which constitute 23.91% of overall consumer.

The average frequency of buying the product is 0.0093 and the average recency is 0.198

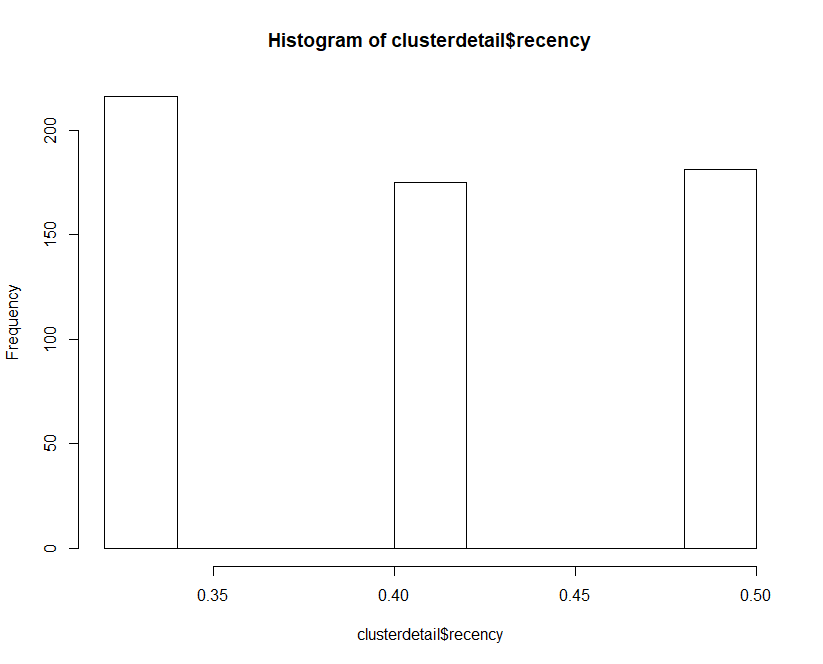
The consumer spent average amount of money. This group seems to be the second high profit group.



**Cluster 5**



Cluster 5 contains some 572 consumers which constitutes 13.25 % of overall population with medium recency ,average frequency and average monetary with average spending customer. This group seems to have ordinary consumer.



From the above histogram for recency we can find that, the customer of this cluster purchased through the year. There was spike in purchase in Quarter 1.

# Hierarchical Clustering

This hierarchical clustering method defines the cluster distance between two clusters to be the maximum distance between their individual components. At every stage of the clustering process, the two nearest clusters are merged into a new cluster. The process is repeated until the whole data set is agglomerated into one single cluster.

The sample code with complete linkage hierarchical clustering is given below.

eudist <- dist(as.matrix(rfm), method = "euclidean")

hcluster <- hclust(eudist, method = "complete" )

# Cut tree into 5 groups

subgrp<- cutree(hcluster, k =5)

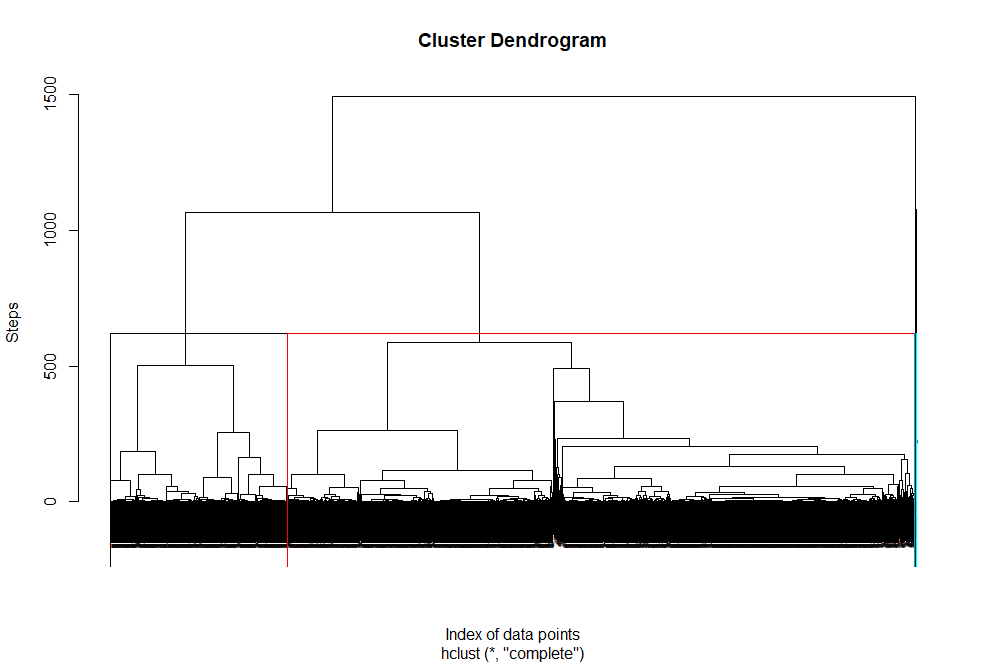
# Number of members in each cluster

table(subgrp)

plot(hcluster, cex = 0.1,xlab="Index of data points", ylab="Steps",main = "Cluster Dendrogram")

rect.hclust(hcluster, k = 5, border = 1:5)

The corresponding dendrogram is given below.



* Since we have considered 5 cluster in K-mean analysis(section 4.1), we have considered same number of cluster for hierarchical clustering as well.
* The horizontal axis of the dendrogram represents the distance or dissimilarity between clusters. The vertical axis represents the objects and clusters

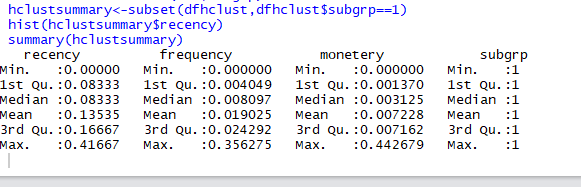
We can see that the clustering pattern for complete linkage distance tends to create compact clusters

## Interpretations OF THE clusters

The distribution of customer across the cluster is given below.

|  |  |  |
| --- | --- | --- |
| Cluster | Total records in the cluster | % 0f overall population |
| 1 | 3363 | 77.9 |
| 2 | 945 | 21.89 |
| 3 | 2 | 0.05 |
| 4 | 4 | 0.09 |
| 5 | 3 | 0.07 |

**Cluster 1**

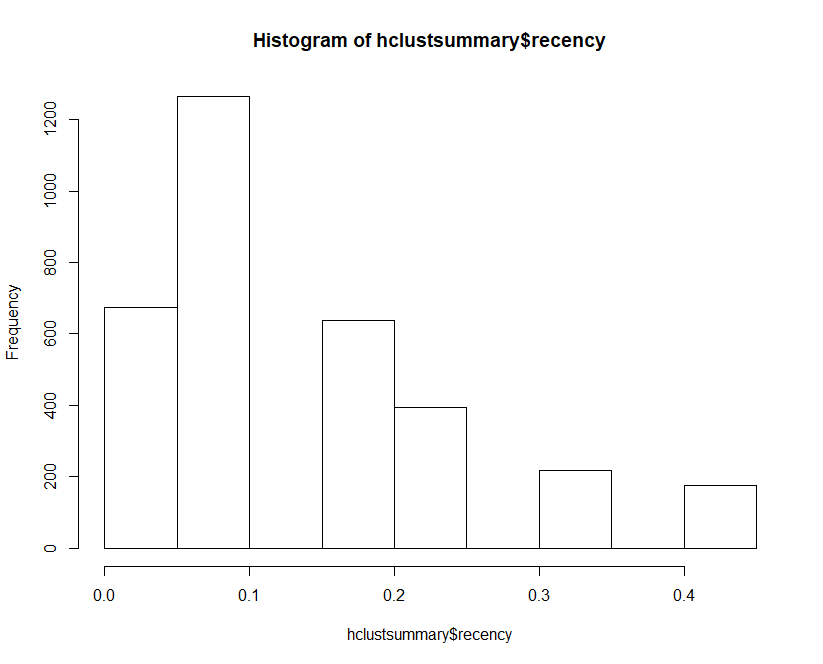


The cluster 1 have 3363 consumers which constitute 77.92% of overall consumer.

The average frequency of buying the product is 0.019 and the low recency is 0.13535

This is the largest size customer. Consumers in this group have a reasonable value of frequency. This group of consumers can be categorized as very low recency, and medium monetary with a low spending per consumer.

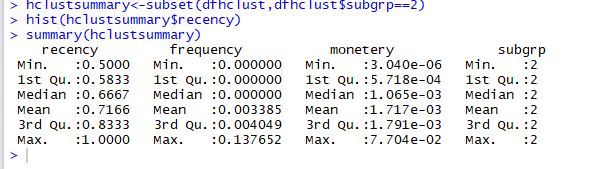
Since the cluster contains most 77.92% consumer, the cluster 1 is earns more revenue for the company compare to the other cluster .The company should focus this cluster .



From the above histogram for recency we can find that, the customer of this cluster purchased through the year. There was spike in purchase in Quarter 1.

The purchase gradually decreased from Quarter 4 to Quarter 1.

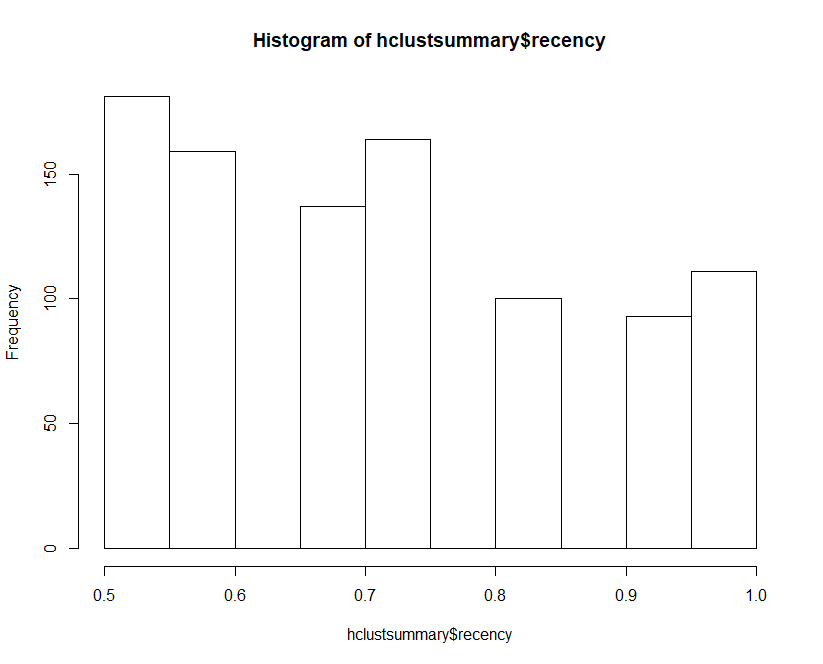
**Cluster 2**



The cluster 2 have 945 consumers which constitute 21.89% of overall consumer.

The average frequency of buying the product is 0.0033 and the high recency is 0.7166 and low monetary.

Since frequency is very low and monetary is also low ,this cluster seems to low profit customer.



From the above histogram for frequency Vs recency plot we can find that the customer almost purchased throughout the year. The first-two quarter the frequency of buying is good but in the next two quarters it gets reduced.

Since the cluster 1 &2 constitute more than 99 % customers ,we exclude rest of the customers from our study.

# ANNEXTURE

1. AAM\_ASSIGNMENT\_TA17002.R **Source code**
2. **retailrfm.csv Data**