**E commerce project in R:**

DESCRIPTION

**Background of Problem Statement:**  
 A UK-based online retail store has captured the sales data for different products for the period of one year (Nov 2016 to Dec 2017). The organization sells gifts primarily on the online platform. The customers who make a purchase consume directly for themselves. There are small businesses that buy in bulk and sell to other customers through the retail outlet channel.

**Project Objective:**  
Find significant customers for the business who make high purchases of their favourite products. The organization wants to roll out a loyalty program to the high-value customers after identification of segments. Use the clustering methodology to segment customers into groups:

**Domain:** E-commerce

**Dataset Description:**  
This is a transnational dataset that contains all the transactions occurring between Nov-2016 to Dec-2017 for a UK-based online retail store.

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| InvoiceNo | Invoice number (A 6-digit integral number uniquely assigned to each transaction) |
| StockCode | Product (item) code |
| Description | Product (item) name |
| Quantity | The quantities of each product (item) per transaction |
| InvoiceDate | The day when each transaction was generated |
| UnitPrice | Unit price (Product price per unit) |
| CustomerID | Customer number (Unique ID assigned to each customer) |
| Country | Country name (The name of the country where each customer resides) |

**Analysis tasks to be performed:**  
 Use the clustering methodology to segment customers into groups:  
 Use the following clustering algorithms:

1. K means
2. Hierarchical

• Identify the right number of customer segments.  
• Provide the number of customers who are highly valued.  
• Identify the clustering algorithm that gives maximum accuracy and explains robust clusters.  
• If the number of observations is loaded in one of the clusters, break down that cluster further using the clustering algorithm. [ hint: Here loaded means if any cluster has more number of data points as compared to other clusters then split that clusters by increasing the number of clusters and observe, compare the results with previous results.]

***Business scenario:*** *To find out the number of customers who made high purchases so that the organization can roll out an offer to high-value customers.*

***Expectation/Goals:*** *Properly segment the customers based on high purchases and identify high-value customers.*

***Code:*** *No.of observations=534165 (max number of observations in 1 cluster out of 8*

*K Means algorithm: Hierarchical algorithm:*

library(readr)

Project\_2\_Dataset1<-read.csv("Practice datasets/Project 2\_Dataset.csv",header=TRUE)

View(Project\_2\_Dataset1)

Project\_2\_Dataset1$X=NULL

Project\_2\_Dataset1$Purchases=Project\_2\_Dataset1[,4]\*Project\_2\_Dataset1[,6]

View(Project\_2\_Dataset1)

dim(Project\_2\_Dataset1)

summary(Project\_2\_Dataset1)

Project\_2\_Dataset2<-data.matrix(Project\_2\_Dataset1, rownames.force = NA)

Project\_2\_Dataset2<-scale(Project\_2\_Dataset2)

Sample\_data\_short<-Project\_2\_Dataset2[,c(9)]

View(Sample\_data\_short)

Sample\_data\_short\_matrix<-data.matrix(Sample\_data\_short)

wss<-(nrow(Sample\_data\_short\_matrix)-1)\*sum(apply(Sample\_data\_short\_matrix,2,var))

for(i in 2:15)wss[i]<-sum(kmeans(Sample\_data\_short\_matrix,centers = i)$withinss)

plot(1:15,wss,type="b",xlab="Number of Clusters",ylab="within sum of squares")

cl<-kmeans(Sample\_data\_short\_matrix,8,nstart=100)

cl

Sample\_data\_short\_matrix<-cbind(Sample\_data\_short\_matrix,cl$cluster)

View(Sample\_data\_short\_matrix)

install.packages("mclust")

library(mclust)

fit<-Mclust(Sample\_data\_short\_matrix)

plot(fit)

library(readr)

project\_2\_dataset1<-read.csv("Practice datasets/Project 2\_Dataset.csv",header=TRUE)

View(project\_2\_dataset1)

project\_2\_dataset1$X=NULL

project\_2\_dataset1$purchases=project\_2\_dataset1[,4]\*project\_2\_dataset1[,6]

View(project\_2\_dataset1)

project\_2\_dataset2<-data.matrix(project\_2\_dataset1, rownames.force = NA)

Sample\_data<-project\_2\_dataset2[sample(nrow(project\_2\_dataset1),500,replace=TRUE),]

View(Sample\_data)

Sample\_data1<-hclust(dist(Sample\_data, method = "euclidian"))

plot(Sample\_data1)

plot(Sample\_data1, labels = Sample\_data1$purchases)

groups<-cutree(Sample\_data1,k=4)

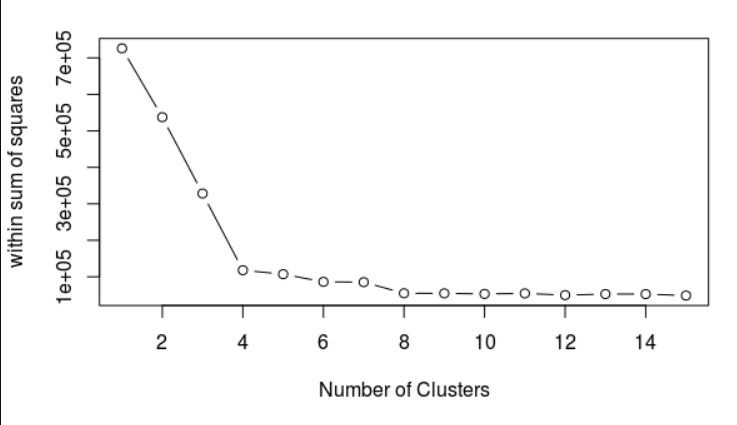
groups

***Analysis:***

1. If the whole set of data is taken and performed K means the number of clusters are 8 and if only sample data is taken then the no.of clusters are 4.
2. When the whole data set is traken and performed Hierarchical clustering the no.of clusters are 4.
3. I feel K means algorithm gives max accuracy and explains robust clusters however it requires a lot of pre-processing work to be made before using Kmeans() function.

Not enough space to put in Hierarchical dendrogram output screenshot.

***Output Screenshot:***



*\*\*The above boxes contain the whole code. Please drag the boxes to view it.\*\**