Executive Program in Advance Business Analysis 2017-18

Batch-01

Online Retail Analysis

Submitted to

Prof. Kavitha Ranganathan

Submitted by

Shwetank Pandey

Mridul Sharma

Introduction

As a recent Deloitte reports states that customers are more empowered than state, online retail has become more competitive than ever. Today's customers have vastly different and more sophisticated expectations about the product and its value. The future of online retail depends on the number of customer connections and therefore, retail giants rely more on information about their customers to make necessary strategic decisions.

In this project, we are implementing few things at different stages. First, the data from the desired source is imported. Second, data cleaning or segmentation is performed based on the imported dataset. Third, an RFM analysis is performed based on recency, frequency and monetary scores of all the customers after segmentation. Fourth, implementation of various algorithms (like K-means) to find the optimal value. Fifth come the clustering based on the value being scored in the previous step. And lastly, analyzing various outputs based on the clustering and visualizing the RFM data.

After carefully performing the above-mentioned stages, different graphs will be plotted to determine impeccable analysis for the online retailer and recommendations will be done based on that. So not only the retailer will be able to recommend useful products to the customer, it can also help to focus on those customers who are important for increasing his/her sales. This will contribute to take necessary actions in order to increase retailer's profit.

Problem Statement

A retail seller wants to increase his profit based on historical transactions of customers. We need to segment/cluster his customers based on their behavior.

The main objective of this proposal is to do in depth online retail analysis. For example, if a retailer wants to increase its profit based on historical transactions or maybe to provide recommendation based on its previous purchases.

It includes various after the segmentation, clustering the values and finally analyzing the same to perform necessary actions.

Data

We leveraged the data for analysis from following source:

Link: http://archive.ics.uci.edu/ml/machine-learning-databases/00352/

Data Set Information:

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Data Set Characteristics:	Multivariate, Sequential, Time-Series	**Number of Instances:**	541909	**Area:**	Business
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	8	**Date Donated**	2015-11-06
Associated Tasks:	Classification, Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	119466

Source:

Dr Daqing Chen, Director: Public Analytics group. chend '@' lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK.

Attribute Information:

InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.

StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.

Description: Product (item) name. Nominal.

Quantity: The quantities of each product (item) per transaction. Numeric.

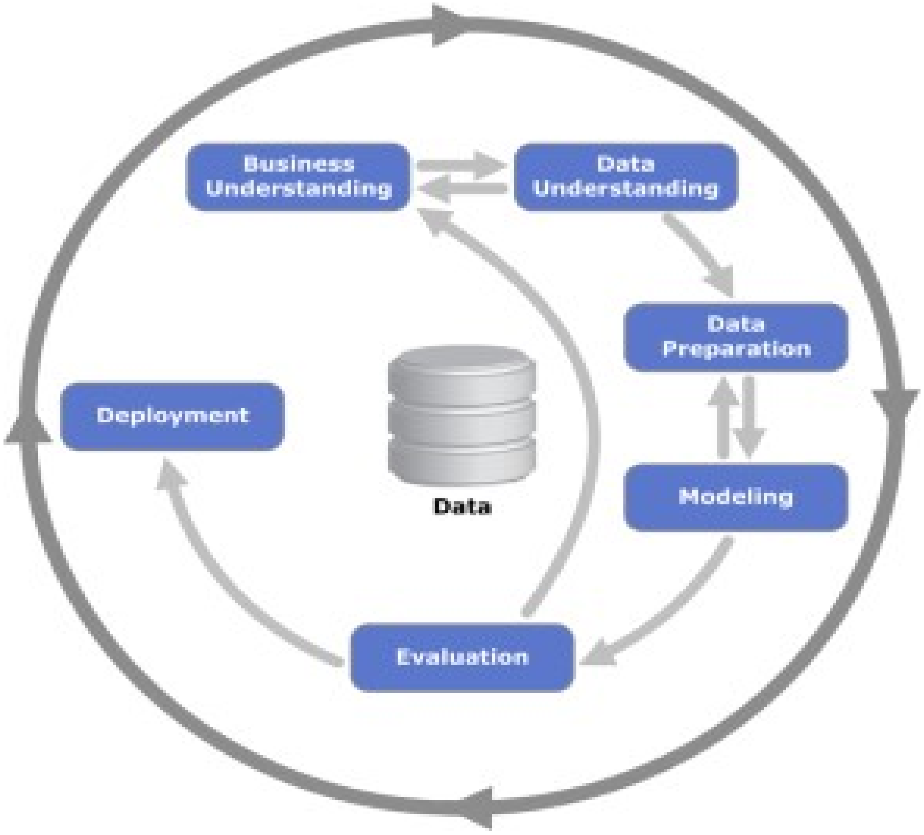
InvoiceDate: Invice Date and time. Numeric, the day and time when each transaction was generated.

UnitPrice: Unit price. Numeric, Product price per unit in sterling.

CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.

Country: Country name. Nominal, the name of the country where each customer resides.

Problem Solving Methodology: CRISP-DM Framework



Analytics problem solving involves multiple steps like data cleaning, preparation, modelling, model evaluation etc. Completing a typical analytics project may take several months, and thus it is important to have a structure for it. The structure for analytics problem solving is called the CRISP-DM framework - Cross Industry Standard Process for Data Mining.

As a business analyst, we will face a multitude of challenges ranging from understanding various business problems to choosing the best techniques to solve them. To avoid getting lost, data scientists have developed a robust process to solve virtually any analytics problem in any industry–appropriately called the Cross Industry Standard Process for Data Mining (CRISP–DM) framework.

It involves a series of steps which you will soon find quite intuitive:

1. **Business understanding**

   For a business analyst, understanding the business and its specific problems is of utmost importance. Here it's about online retail

2. **Data understanding**

   The data is downloaded from the mentioned sources in CSV format and loaded with necessary R packages

3. **Data Preparation**

   Data is cleaned, processed in the format that makes it ready for the purpose of analysis. The descriptive analysis is performed over the data to identify the patterns, trends and outliers, and draw the data insights that will helped to treat outliers and fit the right model or algorithm

4. **Data Modelling**

   Modelling is the heart of data analytics. One can think of a model as a black box which takes relevant data as input and gives an output you are interested in. Here we used two algorithms to perform cluster analysis i.e. K-Means and Hierarchical Clustering Algorithms

5. **Model Evaluation**

   Evaluating the results in different tools, reviewing the process and summarizing the results keeping the business success constraints in mind

6. **Model Deployment**

   In data analytics, evaluation is when we put everything we have done to litmus tests. If the results obtained from model evaluation are not satisfactory, we reiterate the whole process. If the model performs well and gives us accurate results, congratulations. We can move on to implementation of the model.

Here, imported the dataset and performed data cleaning followed by exploratory data analysis. Did RFM analysis and calculated RFM Scores. In RFM analysis, look at the recency, frequency and the monetary scores of all the customers for segmentation:

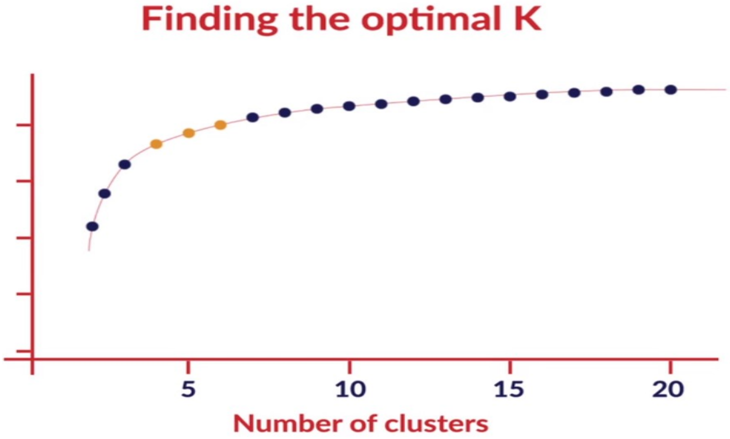
Recency
How recently did the customer purchase?

Frequency
How often do they purchase?

Monetary Value
How much do they spend?

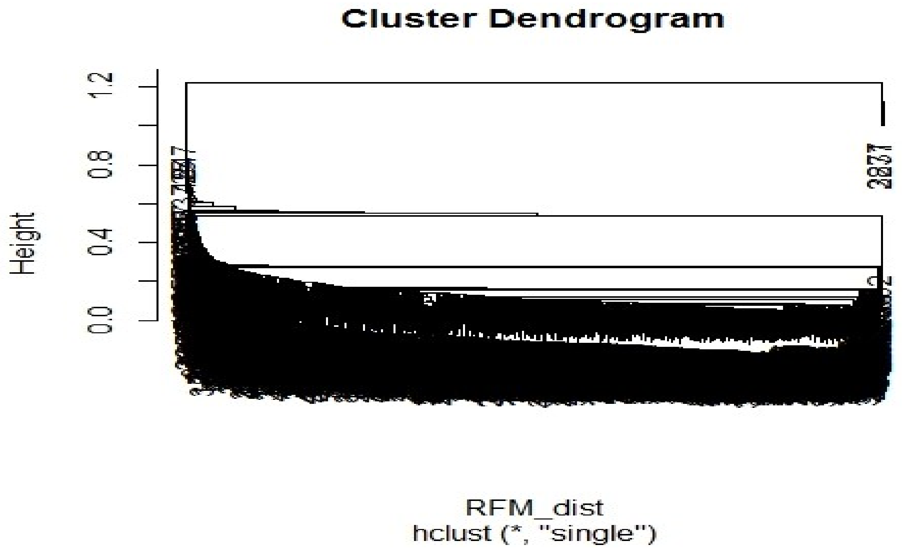
K-Means Algorithm

Further, implementation of K-means algorithm has been done with finding the optimal value of K (here K=5) by using the elbow method to arrive at a range of the values of K for which can get optimal clusters, i.e. clusters with the maximum inter-cluster variance and minimum intra-cluster variance and further appended the ClusterID to RFM data.

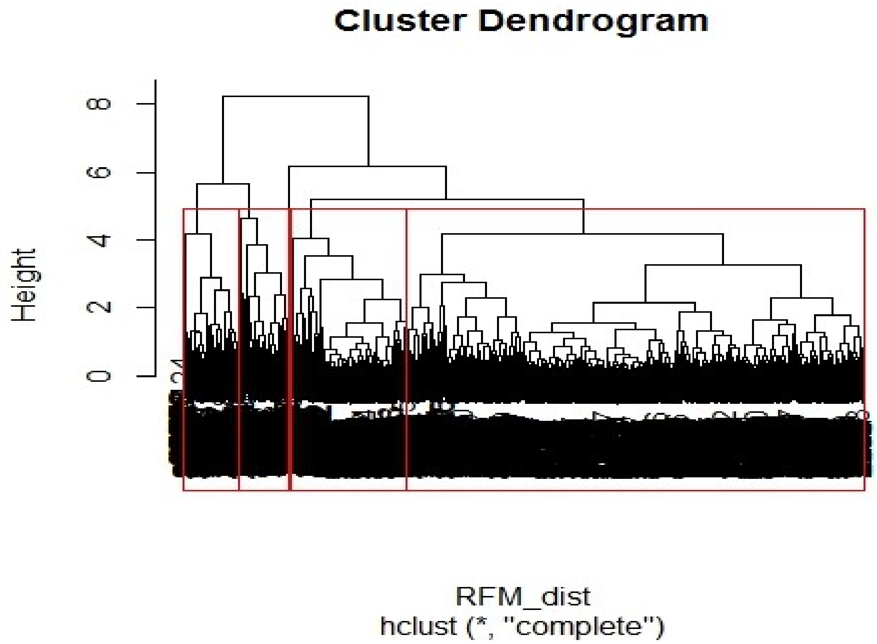


Hierarchical Clustering Algorithm

Further, implementation of hierarchical clustering has been done to get optimal value of clusters by the help of cutting the dendrograms.



Here, the dendrogram we are getting is very narrow and the range of the height is very small. So, we used another linkage i.e. complete linkage.



Here, the height of the dendrogram has actually increased considerably and also the clusters are very distinct.

Results and Conclusion Discussions

K-Means Algorithm

Here, each graph has the cluster number on the x axis, whereas the value of Recency, Frequency and Monetary is on the y axis.

1. **K-Means Analysis 1:**

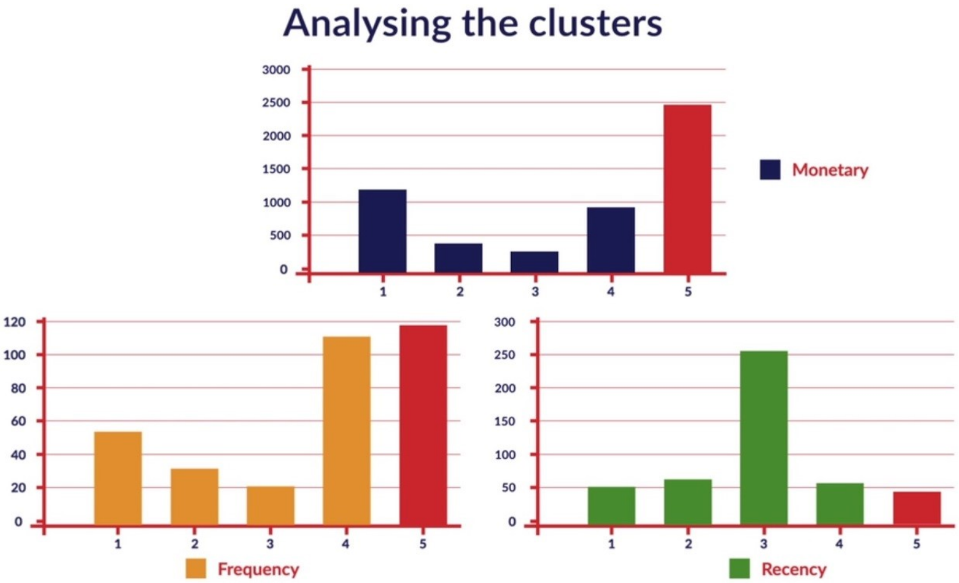


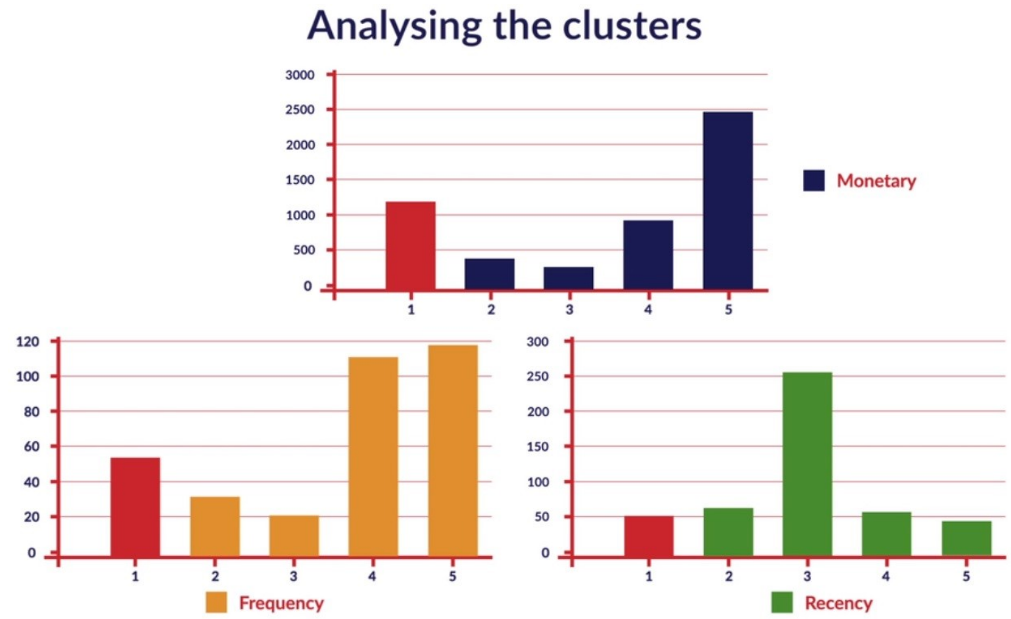
Figure 6: K-Means Analysis 1

Cluster 5 is the best customer segment from the store's point of view:

- These customers make a purchase for a higher amount
- More frequently
- These customers had visited the site recently

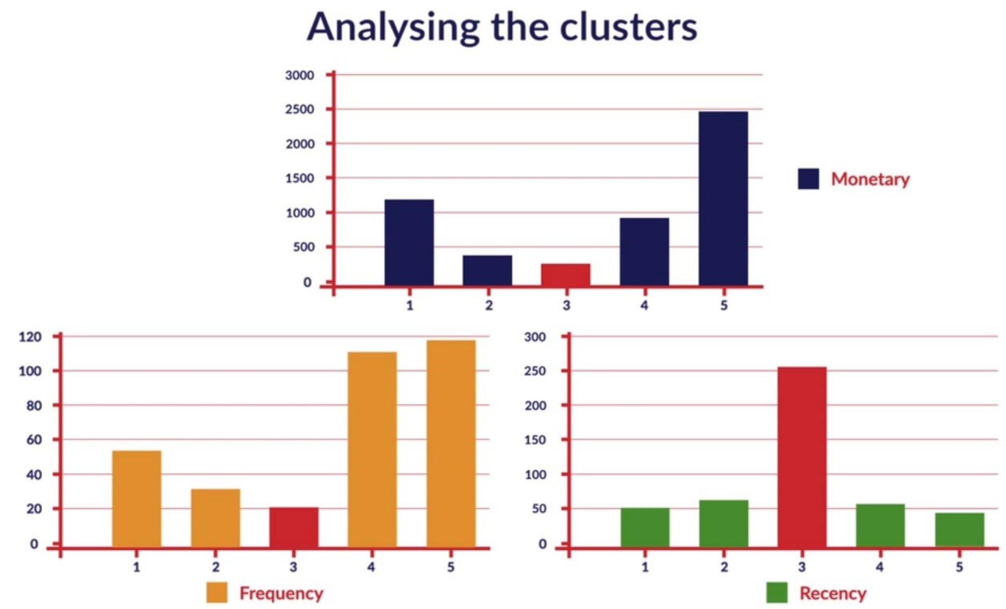
Thus, the store may offer them a reward or loyalty points or some privileged status, to keep them attracted and coming back to the store

2. **K-Means Analysis 2:**



Cluster 1, the customers has favorable features in terms of the purchase amount and recency; however, these have low frequency. But, if the store can re-design its incentive strategy and entice these customers into making a purchase more frequently, they could turn profitable for the store.

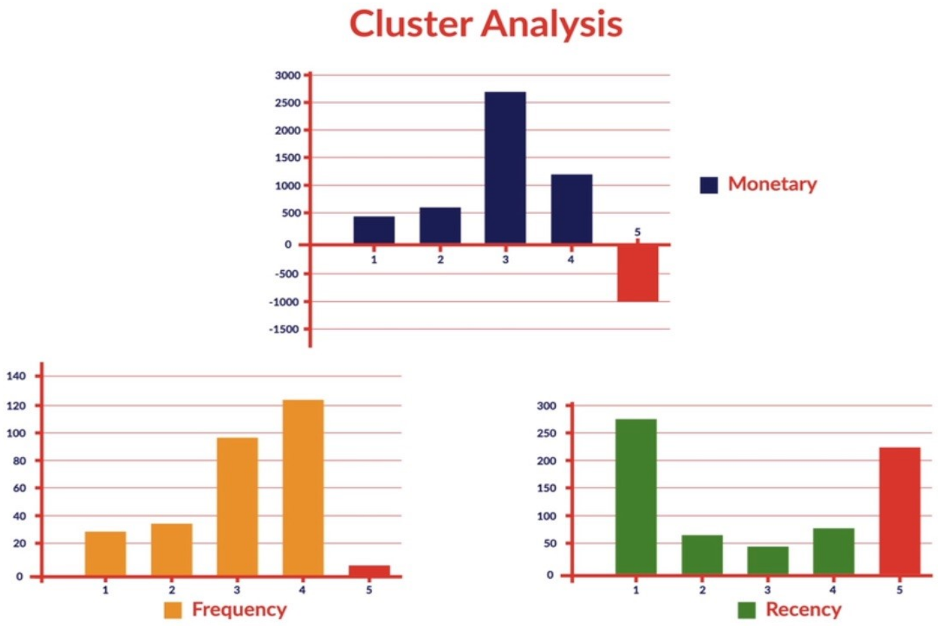
3. **K-Means Analysis 3:**



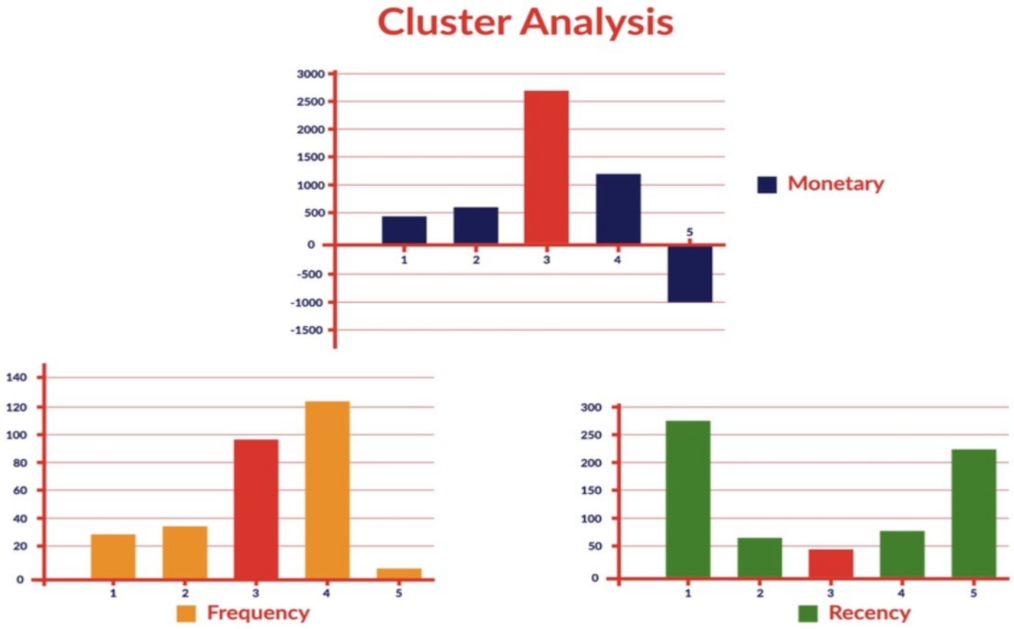
Cluster 3 has the worst customers from the store's point of view. Thus, the store may decide to focus more on this group

Hierarchical Clustering Algorithm

Appended the obtained ClusterIDs to the RFM data set, and analyzed the characteristics of each cluster to derive the business insights from the different customer segments or clusters



Cluster 5: is worst group from the retailer’s point of view. Here, the customers have maximum tendency to return the product apart from very less frequency and recency.



Cluster 3: is the most valuable costumer’s group from the retailer’s point of view and can be awarded loyalty points for further benefits.

Scope of Further Work

India, a country widely seen as protectionist in its retail sector, initiated consultations with online retailers and industry lobby groups on allowing FDI in e-commerce. There are currently no limits to FDI in business-to-business (B2B) e-commerce ventures, or online marketplaces, but foreign direct investment in B2C online ventures is still not permitted. Though there are no firm indicators yet on whether the government is likely to permit foreign ownership, and to what degree, developments are worth monitoring because India's e-commerce market is expected to grow from ₹184.43($2.9) billion in 2013 to more than ₹6359.50 ($100) billion by 2020. The growth in online retail is upsetting traditional brick-and-mortar sellers in India who are demanding heightened protection from the government. Further, as e-commerce is a new sector, the existing laws governing taxation of businesses are not fully suited for online retail companies. This has made it difficult for companies to correctly comply with the laws of the land. The introduction of the Goods and Services Tax is widely expected to ease such issues.

Also, today, the economic reality is well established. The research firm Forrester estimates that e-commerce is now approaching $200 billion in revenue in the United States alone and accounts for 9% of total retail sales, up from 5% five years ago. The corresponding figure is about 10% in the United Kingdom, 3% in Asia-Pacific, and 2% in Latin America. Globally, digital retailing is probably headed toward 15% to 20% of total sales, though the proportion will vary significantly by sector. Moreover, much digital retailing is now highly profitable. It is essential for stores to digitize in order to meet the increased customer expectations now a reality in an always-on, whatever-you-want world. More than 60 percent of world population has a smartphone and 80 percent of these consumers are "smartphone shoppers" – they use their phones to help them shop while in a store, most often to research product reviews, specifications and compare prices.

The various interactions consumers have with digital media and digital platforms have rewritten the arc of the consumer decision journey, causing shoppers to become accustomed to a much greater level of convenience, choice and accessibility. The use of a variety of online-only features – such as personal recommendations, product reviews from other customers, huge product assortments and availability, and 1-click everything – has afforded shoppers the power to make purchasing decisions much more on their own terms.

Acknowledgement

With immense pleasure, I would like to present this project proposal on "***Online Retail Analysis*". It has been our enriching experience to undergo with Executive Business Analytics from IIM-Ahmedabad. The requisite knowledge and information would not have been possible without this course. As a student of Indian Institute of Management- Ahmedabad, we would like to extend our sincere thanks and gratitude to all the professors and faculty members for shaping our understandings.

Finally, thanks and appreciation are also extended to every member of Hughes Centre. Herewith, we assure that the present thesis has been written independently along with all the thoughts and execution proposals.

Reference

Data source: Dr Daqing Chen, Director: Public Analytics group. chend '@' lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK.

Appendix

R-Codes:

```
# Install and Load the required packages
#install.packages("ggplot2")
library(ggplot2)



## Importing the dataset

Online.Retail <- read.csv("Online Retail.csv", stringsAsFactors=FALSE)

## NA value treatment

order_wise <- na.omit(Online.Retail)

## Making RFM data

Amount <- order_wise$Quantity * order_wise$UnitPrice
order_wise <- cbind(order_wise,Amount)

order_wise <- order_wise[order(order_wise$CustomerID),]

monetary <- aggregate(Amount~CustomerID, order_wise, sum)
```

```
frequency <- order_wise[,c(7,1)]

k<-table(as.factor(frequency$CustomerID))

k<-data.frame(k)

colnames(k)[1]<-c("CustomerID")

master <-merge(monetary,k,by="CustomerID")

recency <- order_wise[,c(7,5)]

recency$InvoiceDate<-as.Date(recency$InvoiceDate,"%d-%m-%Y %H:%M")

maximum<-max(recency$InvoiceDate)

maximum<-maximum+1

maximum$diff <-maximum-recency$InvoiceDate

recency$diff<-maximum$diff

df<-aggregate(recency$diff,by=list(recency$CustomerID),FUN="min")

colnames(df)[1]<- "CustomerID"

colnames(df)[2]<- "Recency"

RFM <- merge(monetary, k, by = ("CustomerID"))
RFM <- merge(RFM, df, by = ("CustomerID"))

RFM$Recency <- as.numeric(RFM$Recency)

## Outlier treatment

box <- boxplot.stats(RFM$Amount)
out <- box$out

RFM1 <- RFM[ !RFM$Amount %in% out, ]

RFM <- RFM1

box <- boxplot.stats(RFM$Freq)
out <- box$out

RFM1 <- RFM[ !RFM$Freq %in% out, ]

RFM <- RFM1

box <- boxplot.stats(RFM$Recency)
out <- box$out

RFM1 <- RFM[ !RFM$Recency %in% out, ]

RFM <- RFM1

## Standardisation of data
```

```
RFM_norm1<- RFM[,-1]

RFM_norm1$Amount <- scale(RFM_norm1$Amount)
RFM_norm1$Freq <- scale(RFM_norm1$Freq)
RFM_norm1$Recency <- scale(RFM_norm1$Recency)

## Implementing K-Means algorithm

clus3 <- kmeans(RFM_norm1, centers = 3, iter.max = 50, nstart = 50)
str(clus)
## Finding the optimal value of K

r_sq<- rnorm(20)

for (number in 1:20){clus <- kmeans(RFM_norm1, centers = number, nstart = 50)
r_sq[number]<- clus$betweenss/clus$totss
}

plot(r_sq)

## Running the K-Means algorithm for K =4,5,6

clus4 <- kmeans(RFM_norm1, centers = 4, iter.max = 50, nstart = 50)

clus5 <- kmeans(RFM_norm1, centers = 5, iter.max = 50, nstart = 50)

clus6 <- kmeans(RFM_norm1, centers = 6, iter.max = 50, nstart = 50)
```

```
## Appending the ClusterIDs to RFM data

RFM_km <-cbind(RFM,clus5$cluster)

colnames(RFM_km)[5]<- "ClusterID"

## Cluster Analysis

library(dplyr)

km_clusters<- group_by(RFM_km, ClusterID)

tab1<- summarise(km_clusters, Mean_amount=mean(Amount), Mean_freq=mean(Freq),
Mean_recency=mean(Recency))

ggplot(tab1, aes(x= factor(ClusterID), y=Mean_amount)) + geom_bar(stat = "identity")
ggplot(tab1, aes(x= factor(ClusterID), y=Mean_freq)) + geom_bar(stat = "identity")
ggplot(tab1, aes(x= factor(ClusterID), y=Mean_recency)) + geom_bar(stat = "identity")

## Hierarchical clustering

## Calcualting the distance matrix

RFM_dist<- dist(RFM_norm1)

## Constructing the dendrogram using single linkage

RFM_hclust1<- hclust(RFM_dist, method="single")
plot(RFM_hclust1)
```

```
## Constructing the dendrogram using complete linkage

RFM_hclust2<- hclust(RFM_dist, method="complete")
plot(RFM_hclust2)

## Visualising the cut in the dendrogram

rect.hclust(RFM_hclust2, k=5, border="red")

## Making the cut in the dendrogram

clusterCut <- cutree(RFM_hclust2, k=5)

## Appending the ClusterIDs to RFM data

RFM_hc <-cbind(RFM,clusterCut)

colnames(RFM_hc)[5]<- "ClusterID"

## Cluster Analysis

hc_clusters<- group_by(RFM_hc, ClusterID)

tab2<- summarise(hc_clusters, Mean_amount=mean(Amount), Mean_freq=mean(Freq),
Mean_recency=mean(Recency))
ggplot(tab2, aes(x= factor(ClusterID), y=Mean_recency)) + geom_bar(stat = "identity")
ggplot(tab2, aes(x= factor(ClusterID), y=Mean_amount)) + geom_bar(stat = "identity")
ggplot(tab2, aes(x= factor(ClusterID), y=Mean_freq)) + geom_bar(stat = "identity")
```