Capstone Project – 4

Project Title : Online Retail Customer Segmentation

By

Pawar Sapana

Points for discussion

- Problem Statement
- Introduction
- RFM Analysis
- Data Summary
- EDA and Data visualization
- Data Preparation
- Model building
- Model evaluation
- Conclusion

Problem Statement

In this project, the task is to identify major customer segments on 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.

Introduction

- The online retail industry has changed the way customers shop as everything is available online. In order to build a loyal customer base, a company needs to deploy various marketing strategies focused on the diverse nature of its customers.

- A possible solution is to segment customers and make targeted marketing strategies for which historical data of customers is required.

Customer Profiling & Segmentation
- An Analytical Approach To
Business Strategy In Retail Banking

RFM Analysis

RFM (**Recency, Frequency, Monetary**) analysis is a proven marketing model for behavior based customer segmentation. It groups customers based on their transaction history – how recently, how often and how much did they buy.
RFM helps divide customers into various categories or clusters to identify customers who are more likely to respond to promotions and also for future personalization services.

Our goal is to cluster our customers to get insights in:

- Increasing **revenue** (Knowing customers who present most of our revenue)
- Increasing customer **retention**
- Discovering **Trends and patterns**
- Defining **customers at risk**
- We will do **RFM Analysis** as a first step and then **combine RFM with predictive algorithms (k-means)**.

Data Summary

Nominal

- ● **InvoiceNo:** Invoice number. 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. A 5-digit integral number uniquely assigned to each distinct product.
- ● **Description:** Product (item) name.
- ● **CustomerID:** Customer number. A 5-digit integral number uniquely assigned to each customer.
- ● **Country:** Country name. The name of the country where each customer resides.

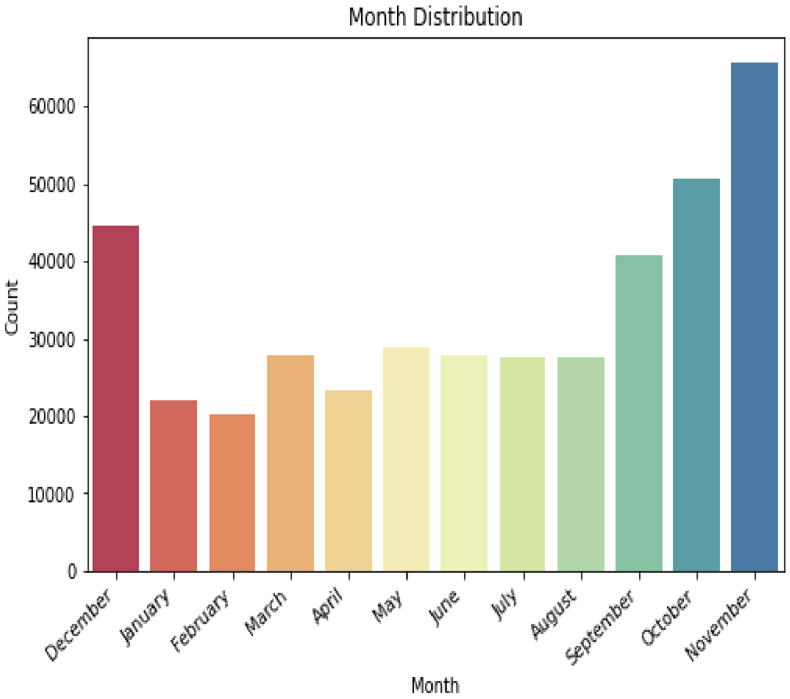
Numeric

- ● **Quantity:** The quantities of each product (item) per transaction.
- ● **InvoiceDate:** Invoice Date and time. The day and time when each transaction was generated.
- ● **UnitPrice:** Unit price. Product price per unit in sterling.

Exploratory Data Analysis

Month wise orders

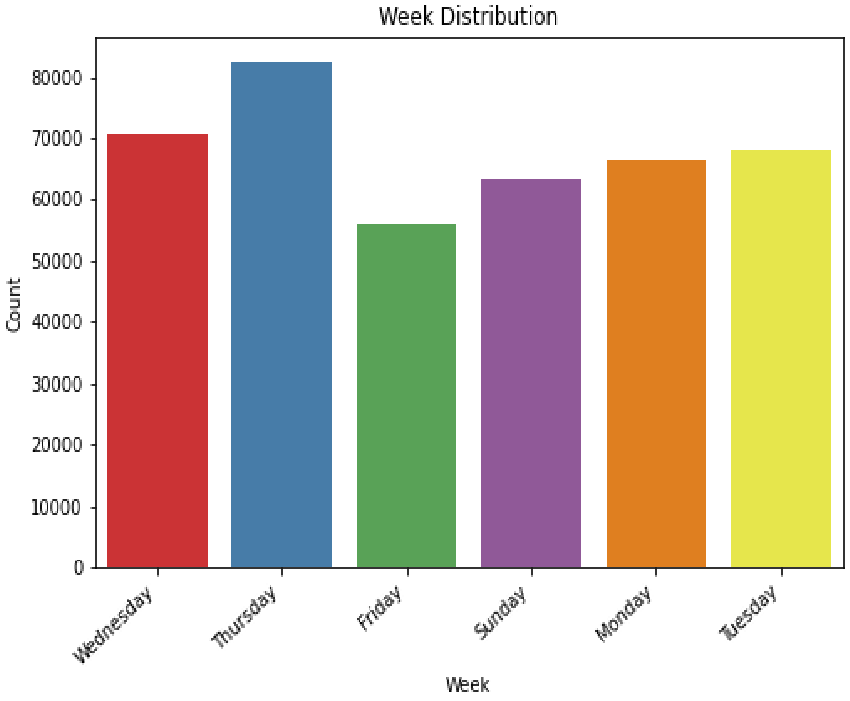
- Year ends show more transaction with November being the highest.



EDA Continued…

Week wise orders

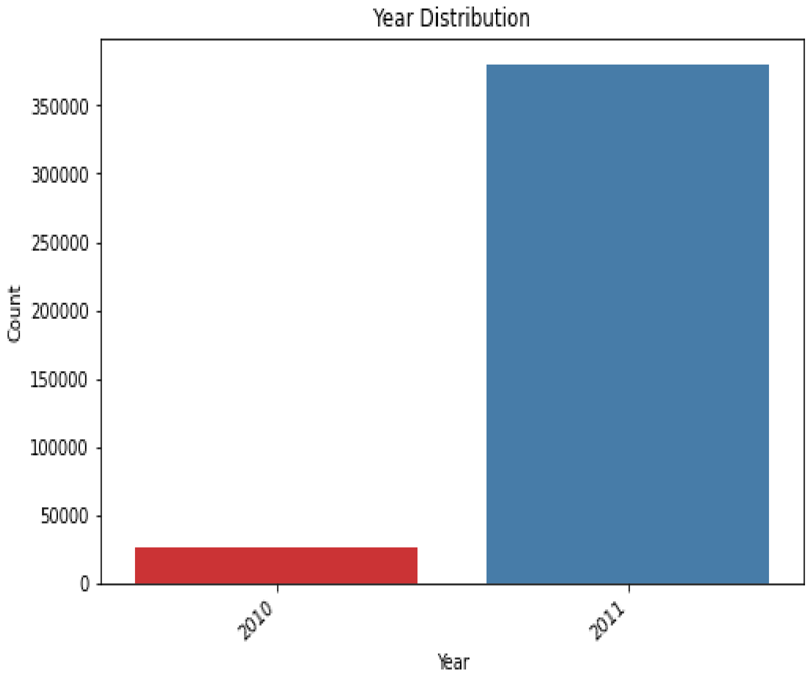
- Thursdays have shown more transactions and as compared to it Fridays shown the lowest transactions.
- No transactions on Saturdays.



EDA Continued…

Year wise distribution

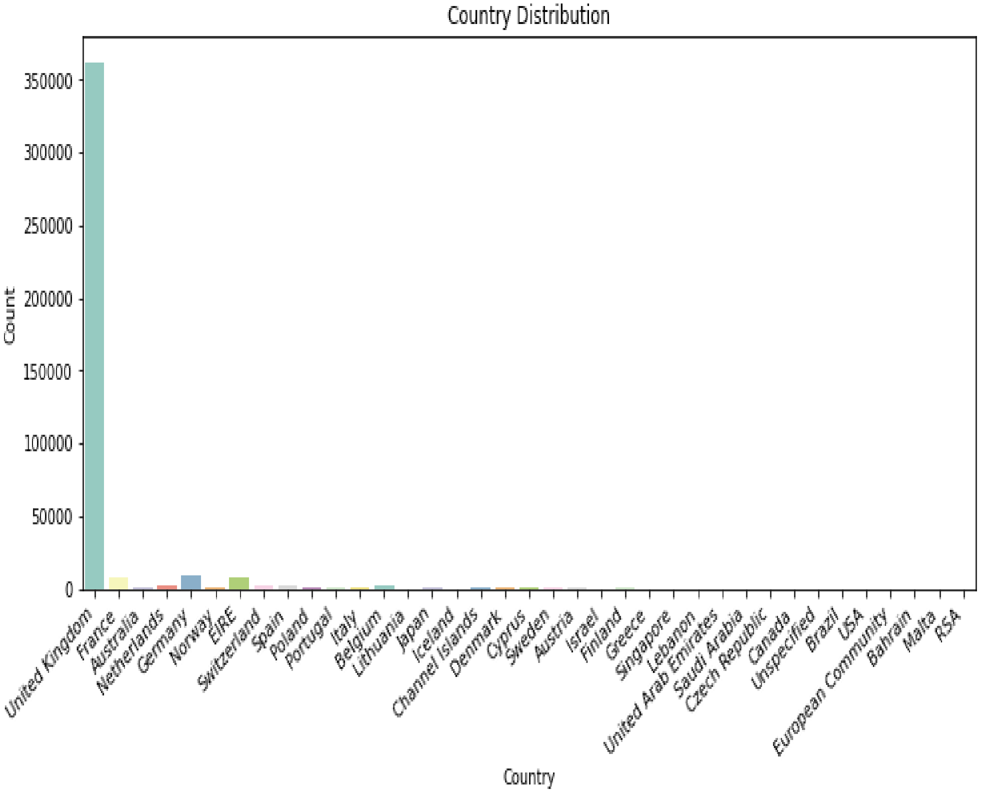
- As we can see more transaction is seen in 2011 as compared to 2010.



EDA Continued…

Country wise sales

- The company receives the highest number of orders in the UK since it is a UK based company.



Data Preparation

● As customer clusters may vary by geography, we'll restrict the data to only United Kingdom customers, which contains most of our customers historical data.

RFM Analysis:

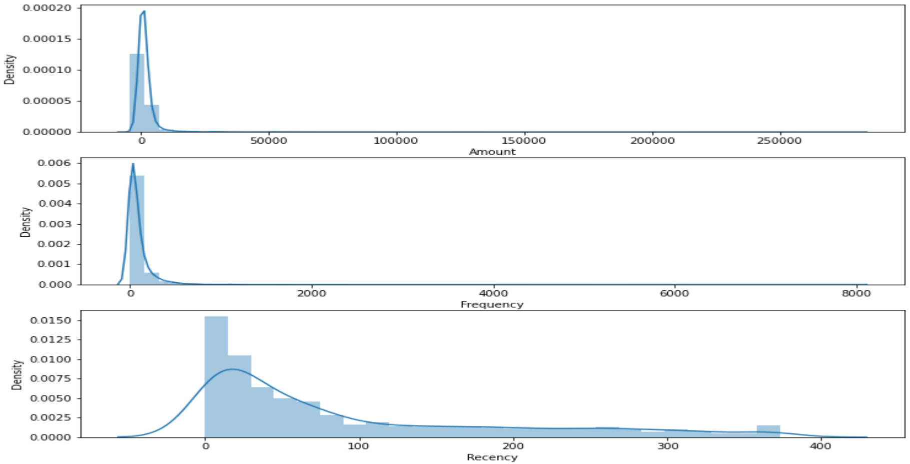
We are going to analyze the Customers based on below 3 factors:

● R (Recency): Number of days since last purchase

● F (Frequency): Number of transactions

● M (Monetary): Total amount of transactions (revenue contributed) So, we have created these three new variables.

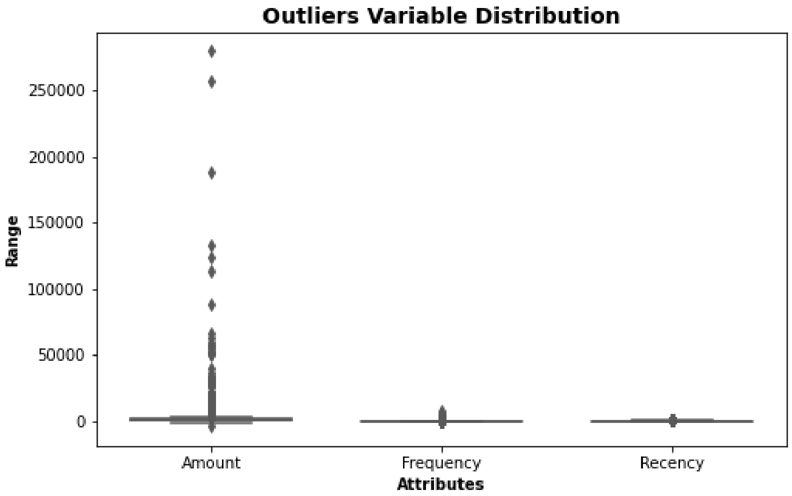
Distribution of Recency, Frequency and Monetary (Amount) variables.



- From the above figure, all the variables do not have a symmetrical distribution. All of them are skewed to the right.

Data Preparation Continued…

Outlier Analysis



We have removed outliers using interquartile range between (0.5,0.95).

Feature Scaling

- Rescaling the Attributes It is extremely important to rescale the variables so that they have a comparable scale. There are two common ways of rescaling:

1. Min-Max scaling
2. Standard (mean-0, sigma-1) Scalar Here, we will use Standard Scalar Scaling.

Model Building

1. K-Means Clustering:

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

The algorithm works as follows:

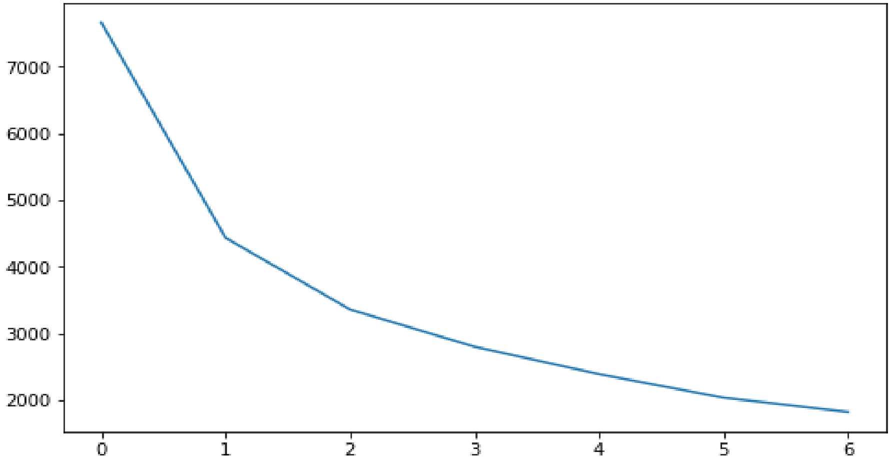
- First we initialize k points, called means, randomly.
- We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that mean so far.
- We repeat the process for a given number of iterations and at the end, we have our clusters.

Finding the Optimal Number of Clusters

Elbow Curve to get the right number of Clusters.

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered.

The Elbow Method is one of the most popular methods to determine this optimal value of k.



Model Building

2. Hierarchical clustering

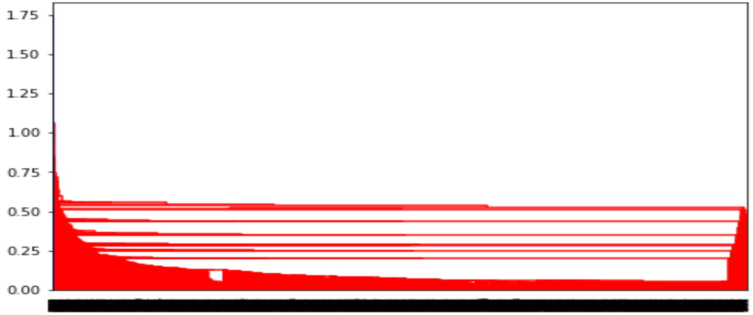
It involves creating clusters that have a predetermined ordering from top to bottom. In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the **dendrogram**.

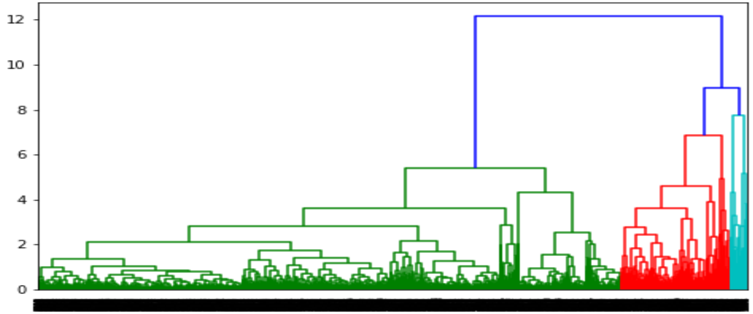
There are two types of hierarchical clustering,

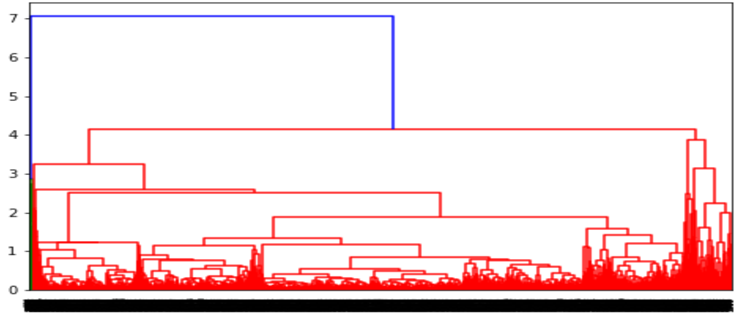
- Divisive
- Agglomerative.

The measure **closest distance** between the two clusters is crucial for the hierarchical clustering. These measures are called **Linkage methods**

- **Single Linkage**:The shortest distance between two points in each cluster.
- **Complete Linkage**:The longest distance between two points in each cluster.
- **Average Linkage:** The average distance between each point in one cluster to every point in the other cluster.







Model Evaluation

Model can be evaluated by metric such as:

Silhouette Score

To study the separation distance between the clusters formed by the algorithm silhouette analysis could be used.

```
For n_clusters=2, the silhouette score is 0.5415858652525395
For n_clusters=3, the silhouette score is 0.508489629614193 7
For n_clusters=4, the silhouette score is 0.48148099614734263
For n_clusters=5, the silhouette score is 0.46501354053484817
For n_clusters=6, the silhouette score is 0.4169515238218781
For n_clusters=7, the silhouette score is 0.4150058806779277
For n_clusters=8, the silhouette score is 0.40728763609819607
```

By observing all the silhouette scores for different values of k, I got the best results for k=3.

So the silhouette score of our best model is 0.50 which can be said to be good for this type of dataset.

.ve results we build the final model with 3 clusters.

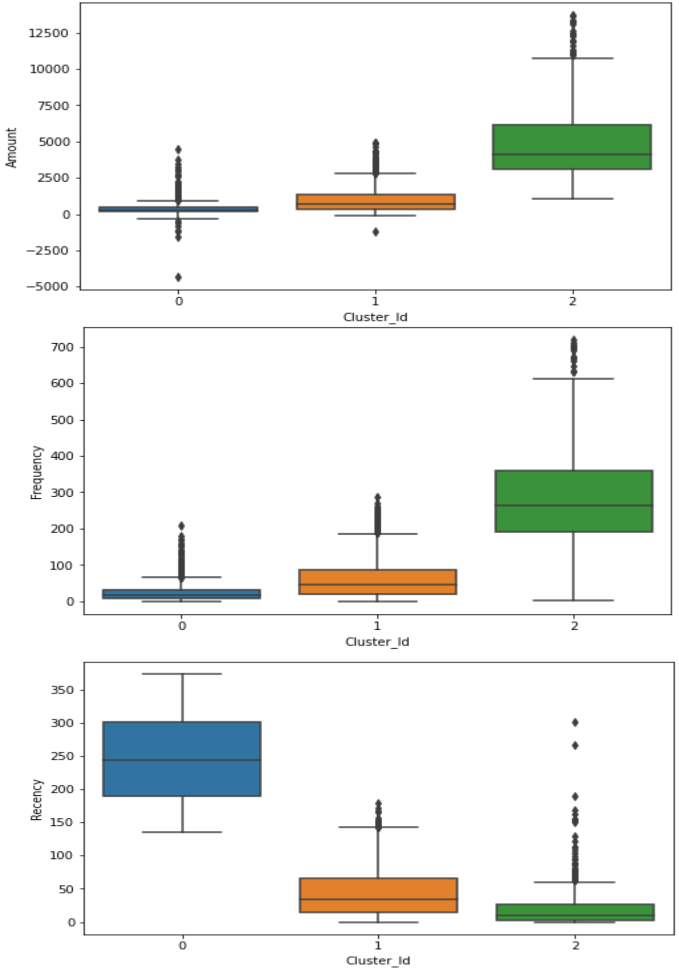
Conclusions

K-Means Clustering with 3 Cluster Ids:

- Customers with Cluster Id 2 are the customers with high amount of transactions as compared to other customers.

- Customers with Cluster Id 2 are frequent buyers.

- Customers with Cluster Id 1 are not recent buyers and hence least of importance from business point of view.



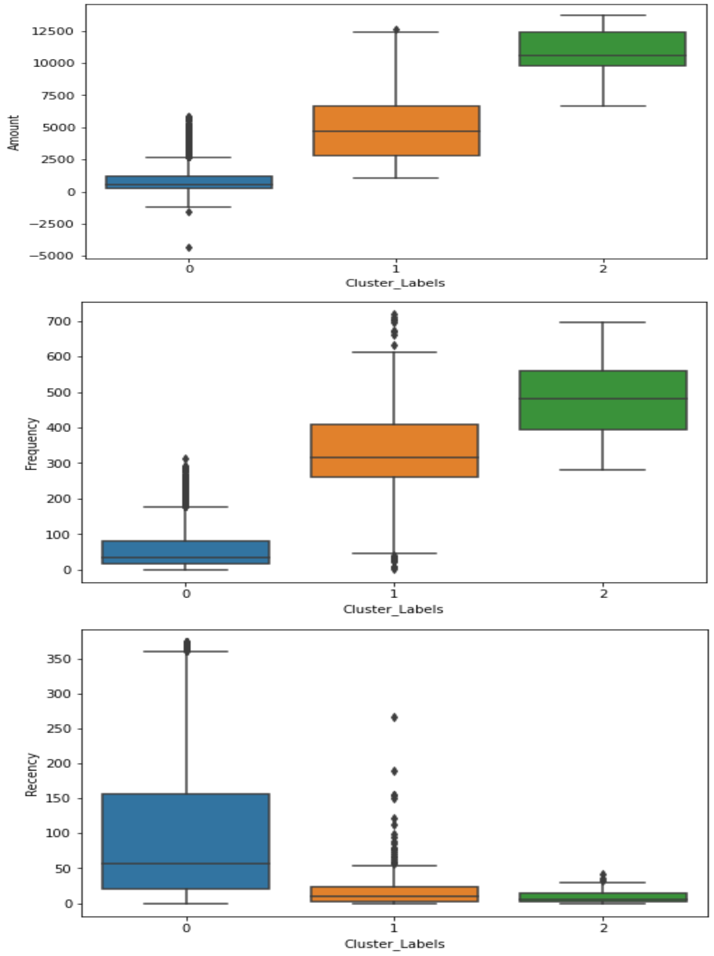
Conclusions Continued…

Hierarchical Clustering with 3 Cluster Labels:

- Customers with Cluster Labels 2 are the customers with high amount of transactions as compared to other customers.

- Customers with Cluster Labels 2 are frequent buyers.

- Customers with Cluster Labels 0 are not recent buyers and hence least of importance from business point of view.



Thank You