VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY

UNIVERSITY OF ECONOMICS AND LAW

FACULTY OF INFORMATION SYSTEMS



REPORT: DATA ANALYSIS IN BUSINESS

TOPIC: CUSTOMER SEGMENTATION BY USING RFM MODEL AND CLUSTERING METHODS

By Nguyen Thi Hong Nhung

*Student ID* - *K184060744*

Supervisor PhD Ho Trung Thanh

*Ho Chi Minh, May 25th,2021*

Table of Contents

[1. Introduction 1](#_Toc72778588)

[1.1. Background of the study 1](#_Toc72778589)

[1.2. Purpose of the study 1](#_Toc72778590)

[2. Literature review 2](#_Toc72778591)

[2.1. Online retail 2](#_Toc72778592)

[2.2. Customer segmentation 2](#_Toc72778593)

[2.3. RFM analysis 3](#_Toc72778594)

[2.4. Cluster analysis 3](#_Toc72778595)

[2.5. K-means Algorithm 3](#_Toc72778596)

[3. Framework of the study 5](#_Toc72778597)

[4. Experimental Installation 6](#_Toc72778598)

[4.1. Experimental enviroment 6](#_Toc72778599)

[4.2. Experimental process 6](#_Toc72778600)

[4.3. Experimental results 8](#_Toc72778601)

[4.4. Evaluation 10](#_Toc72778602)

[5. Conclusion 11](#_Toc72778603)

List of Figures

[Fig 3‑1. Proposal framework 5](https://d.docs.live.net/4024dfb38109f96d/Documents/K184060744_NguyenThiHongNhung_DA_RFMAnalysis.docx#_Toc72782445)

[Fig 4‑1. Elbow Method 8](#_Toc72782446)

[Fig 4‑2. Customer segment distribution 10](#_Toc72782447)

List of Table

[Table 4‑1. Head of Cluster results 8](#_Toc72782476)

[Table 4‑2. Statistic cluster results 8](#_Toc72782477)

[Table 4‑3. Cluster rules 9](#_Toc72782478)

[Table 4‑4. Customer's key segments 9](#_Toc72782479)

[Table 4‑5. Customer segment statistic 10](#_Toc72782480)

# Introduction

## Background of the study

Customers are regarded as important strategic resources of an enterprise, and gaining and retention of customers has become the most critical factor of an enterprise’s success (Lai, 2009).

Howerver, customers now face a variety of suppliers with varying business plans in today's intensely competitive market. Managers in such a position must be mindful of consumer habits and customer situations in their segment. In such a market, mining consumer data is needed to achieve this aim. Therefore, the most significant factor in achieving profitability in a competitive environment is the creation of successful methods for interacting with various consumer classes.

In this project, I have extended the RFM model definition to a database of customer transactions and the type of customer of online retail industry. The methodology is applying RFM model with K-means clustering to determine customer type. Consequently, the customer type will be brought into co-consideration with the RFM principles to make it even practical customers and their segmentation.

## Purpose of the study

The purpose of this study is to gain a deeper understanding of customer segments in order to determine the value of focusing on them. It is now generally accepted that it costs about five times more to gain a new customer than to keep an existing one, and ten times more to get a dissatisfied customer back (Marcus, 1998).

In this study customer segmentation process is implemented to segment the customers and define some strategies for them using the RFM model and K-means clustering method.

# Literature review

## Online retail

The increasing of online sales indicates that the way consumers purchase for and use financial services has changed. There are unique characteristics of online shopping, such as: each customer ’s shopping process and activities can be tracked instantaneously and accurately, each customer’s order is associated with a billing address, and each customer has an online store account with essential contact and payment information. These enabled online retailers to treat customers personally with understanding of each customer and to build upon customer-centric business intelligence.

Regarding to customer-centric business model, Online retailers are usually concerned with the following common business concerns:

* Who are the most/least valuable customers to the business? What are the distinct characteristics of them?
* What are customers' purchase behavior patterns? Which products/items have customers purchased together often? In which sequence the products have been purchased? (M. A. Kadir, 2019)

## Customer segmentation

Customer segmentation is the practice of dividing a company’s customers into groups that reflect similarity among customers in each group. The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business.

Customer segmentation has the potential to allow marketers to address each customer in the most effective way. Using the large amount of data available on customers (and potential customers), a customer segmentation analysis allows marketers to identify discrete groups of customers with a high degree of accuracy based on demographic, behavioral and other indicators (Optimove, 2020).

## RFM analysis

Recency, Frequency and Monetary (RFM) analysis is a marketing technique in analyzing customer behavior such as how recently a customer has purchased, how often the customer purchases, and how much the customer spends. It could improve customer segmentation by dividing customers into various groups for future personalization services and to identify customers who are more likely to respond to promotions (Birant, 2011).

The advantage is that the customers’ behavior can be captured by using a relatively small number of features, which improves the transparency of the target selection models that are developed. The RFM variables are appropriate for capturing the specifics of the customer's purchase behavior (Kaymak, 2001).

The RFM model is based on three quantitative factors:

* Recency: How recently a customer has made a purchase
* Frequency: How often a customer makes a purchase
* Monetary Value: How much money a customer spends on purchases

## Cluster analysis

Cluster analysis is a traditional statistical method that was used for simple data mining. It is used for classifying things into segments whose have similar characteristics. Data inputs are treated similarly in order to obtain information for deciding associations or groups. Therefore, in clustering the characteristics according to which the objects are categorized into segments or classes are initially unknown. These techniques are often used for customer segmentation and can be applied to huge datasets.

## K-means Algorithm

There are several algorithms which are used in develop clustering. K-means is one of popular algorithm for cluster analysis which is a quite fast algorithm applicable in large datasets that requires predetermination of the number of clusters by the user (M. A. Kadir, 2019).

K-means algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster (Dabbura, 2018).

* Pseudocode of K-means clusstering algorithm

Input: Dataset D = {d1, d2,…, dn}

Number of cluster k

Output: A set of k clusters

Step:

1. Initialize k centroids randomly
2. Asscociate each data point in D with the nearest centroid. This will divide the data points into k clusters
3. Recalculate the position of centroids
4. Repeat step 2 & 3 until there are no more changes in the membership of the data point

# Framework of the study

This framework (Figure 3.1) presents an approach which uses RFM analysis and K-means clustering. The proposed model can assist managers in create effective marketing campaigns by making good use of the information gained from clustering and RFM analysis.

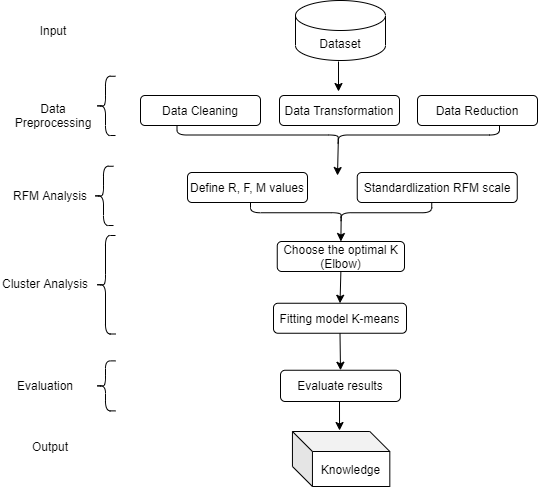


Fig 3‑1. Proposal framework

The framework includes 4 main stages. The first was related to pre-analysis which called Data preprocessing. Data preprocessing refer data cleaning, transformation and reduction. The purpose of this process is to enhance the clustering model accuracy, performance and scalability. Second, data were analyzed by using RFM analysis and being standardlized for the third stage. K-means clustering was used to cluster data in the third step. Finally, based on the evaluation of results of the optimal clustering model described above, the output data are the clusterd labels of customer segment.

# Experimental Installation

## Experimental enviroment

Experimental platform: Jupyter Notebook with Python language, use different libraries such as: Matplotlib, pandas, skit-learn, spicy, numpy,...

The dataset is a transnational data set which contains total 541909 records about the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

## Experimental process

### Dataset

* This study made use of the dataset named “Online Retail Data Set”, a database of comsumer transactions from 2010 to 2011.
* The data set includes 8 attributes and 541909 instances describe invoice number, customer number, country, ... of each customer.
* Sources: <https://archive.ics.uci.edu/ml/datasets/online+retail>
* The dataset includes information about:
* 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.

### Data Preprocessing

*Data Cleaning:* For the selected data set, data cleaning involves data processing with a missing value in the CustomerID property - there are 135080 values ​​in this property deficiency treatment. The processing technique is to ignore records with missing attributes due to percentage the records missing this value are small compared to the overall rows in the data set.

*Data Transformation:* Convert the type of InvoiceDate attribute from string to datetime.

*Data Reduction:* Remove attributes Description and StockCode because this is an unnecessary attribute in the analysis process. As customer clusters may vary by geography, we will restrict the data to only United Kingdom customers, which contains most of our customers historical data.

After preprocessing step, the dataset with 354321 records and 6 attributes is used.

### Define R, F M values

We start by calculating the recency (R) of 3920 customer to determine how much time has passed since his last purchase. Next, we determine the frequency (F) at which customers buy the products. At the end, we going to determine the Monetary (M) and then merge the R, F and M to get the final RFM dataset.

### Standardization RFM scale

K-means gives the best result under the following conditions:

- Data’s distribution is not skewed.

- Data is standardised (i.e. mean of 0 and standard deviation of 1) (Poh, 2020).

In this stage, I have check skew of RFM data and perform log transformation to reduce the skewness of each variable and then standardize the RFM table by using Standard Scaler.

### Choose optimal K – Elbow Method

In this step, I used the Elbow method in determining the number of cluster in a data set. The elbow method run k-means clustering on the dataset for a range of values of k (1 to 10 in my experimental), and for each value of k calculate the sum of squared errors (SSE).

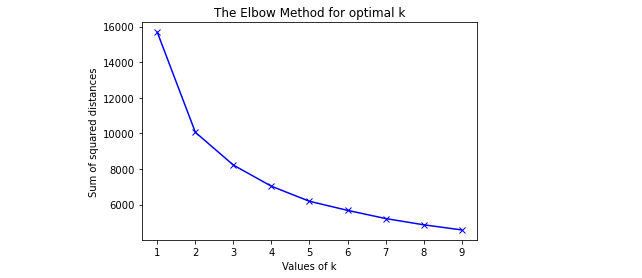


Fig 4‑1. Elbow Method

The drop in the sum of squared distance starts to slow down after k=4 so 4 is the optimal number of clusters for my analysis.

### Fiting model

In this process, customer types are clustered into 4 cluster 1,2, 3 and 4 as table below.

Table 4‑1. Head of Cluster results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No. | CustomerID | Recency | Frequency | Monetary | Cluster |
| 0 | 12346.0 | 325 | 1 | 77183.60 | 3 |
| 1 | 12747.0 | 1 | 103 | 4196.01 | 2 |
| 2 | 12748.0 | 0 | 4595 | 33719.73 | 2 |
| 3 | 12749.0 | 3 | 199 | 4090.88 | 2 |
| 4 | 12820.0 | 2 | 59 | 942.34 | 2 |

## Experimental results

Table 4‑2. Statistic cluster results

|  |  |  |
| --- | --- | --- |
| Cluster | Count | Proportion (%) |
| 1 | 935 | 23.85 |
| 2 | 798 | 20.36 |
| 3 | 1102 | 28.11 |
| 4 | 1085 | 27.68 |

The table above shows the distribution of each cluster in the dataset. The disparity between customer clusters is minor, with 23.85%, 20.36%, 28.11% and 27.68% for clusters 1,2, 3 and 4 respectively. Cluster 2 has the lowest proportion (20.36%) of the total number of customers in the dataset, representing 798 customers, while Cluster 3 has the highest proportion (28.11%) of the total number of customers in the dataset, representing 1102 customers.

Table 4‑3. Cluster rules

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster | Recency | Frequency | Monetary |
| mean | mean | mean |
| 1 | 176.0 | 15.0 | 278.0 |
| 2 | 14.0 | 264.0 | 6096.0 |
| 3 | 113.0 | 35.0 | 737.0 |
| 4 | 53.0 | 84.0 | 1264.0 |

Based on the above table, the common key segments that I will identify in the data set, are shown in the table below.

Table 4‑4. Customer's key segments

|  |  |  |
| --- | --- | --- |
| Cluster | Description | Label (Segment) |
| 1 | Purchase long ago, purchased few and spent little | Lost Cheap Customers |
| 2 | Bought most recently, most often and spend the most | Best Customers |
| 3 | Purchase long ago, purchase not frequently and spend pretty much | Almost Lost |
| 4 | Haven't purchase for some time, but purchased frequently and spend a lot | Loyal Customer |

## Evaluation

Table 4‑5. Customer segment statistic

|  |  |
| --- | --- |
| Segment | Number of customer |
| Almost Lost | 1102 |
| Best Customer | 798 |
| Lost Cheap Customer | 935 |
| Loyal Customer | 1085 |

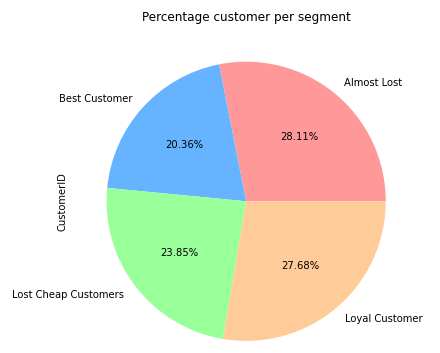


Fig 4‑2. Customer segment distribution

The tables statistic the number of customer while the pie charts above illustrates the proportion of four different segment in the dataset. Overall, the Almost Lost segment has the largest number of customer while the smallest goes to Best Customer segment.

As it can be seen in the chart, customers showing signs of leaving account for more than half of the business’s total customers, Almost Lost customer opcupied 28.11% corresponding to 1102 customers and Lost Cheap Customers accounts for 23.85% means 935 customers which is a huge number of customers. On the other hand, 27.68% of the customer is Loyal customer and Best Customer contributed only 20.36%.

Now that we knew our customers segments we can choose how to target or deal with each segment.

* Almost Lost: Send them personalized emails to encourage them to place a order.
* Best Customers: Rewards them for their multiples purchases. They can be early adopters to very new products. Suggest them “Refer a friend”. Also, they can be the most loyal customers that have the habit to order.
* Loyal Customers: Create loyalty cards in which they can gain points each time of purchasing and these points could transfer into a discount.

# Conclusion

This study suggest an extension of the concept of RFM model applied to customer purchases database and the customer form that includes 541909 transaction historical records. Through using the RFM model to determine customer activity based on customer category, clustering customer partitioning into 4 groups using the K-means algorithm and assigning categories of customers with 4 segment of Lost Cheap Customers, Best customers, Almost lost and Loyal Customer.

The output is the customer segmentation and recommendation strategies for each segment. Better customer segmentation is critical for the retail companies. Because grouping the customers with commonneeds, wants and behaviors allows companies to better identify their target market. Thus, companies may engage in practices such as customized marketing, price regulation, promotions, making more customer touchpoints, etc.

This proposed segmentation model actually makes customers feel happy while growing profits for the company and allowing them to acttact more targeted customers. Finally, it can make it easier to sustain and expand the system of Customer relationship management (CRM).

References

Bigdatauni. (n.d.). *K-means clustering va mo hinh RFM*. Retrieved May 22, 2021, from BigData Uni Web site: https://bigdatauni.com/tin-tuc/k-means-clustering-va-mo-hinh-rfm-p-2.html

Birant, D. (2011). Data Mining Using RFM Analysis. In D. Birant, *Data Mining Using RFM Analysis* (pp. 91-92). Turkey. doi:10.5772/13683

Chandan K Reddy, D. S. (2014, December). A survey on platforms for big data analytics. *Journal of Big Data*, 16-20. doi:10.1186/s40537-014-0008-6

Changyong Feng, W. H. (2014). Log-transformation and its implications for data analysis. *Shanghai Arrchives of Psychiatry*. doi:10.3969/j.issn.1002-0829.2014.02.009

Dabbura, I. (2018, September 17). *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks.* Retrieved May 24, 2021, from Towardsdatascience: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Kaymak, U. (2001). Fuzzy target selection using RFM variables. *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569)* (pp. 1038 - 43). Vancouver, Canada: IEEE. doi: 10.1109/NAFIPS.2001.944748

Kumar, M. (2019, November 11). *Online Retail K-Means & Hierarchical Clustering*. Retrieved May 19, 2021, from Kaggle: https://www.kaggle.com/hellbuoy/online-retail-k-means-hierarchical-clustering/notebook

Lai, X.-a. (2009). Segmentation Study on Enterprise Customers Based on Data Mining Technology. *First International Workshop on Database Technology and Applications.* Wuhan, China: IEEE.

M. A. Kadir, A. A. (2019). *Customer Segmentation on Online Retail using RFM Analysis: Big Data Case of Bukku.id.* Jakarta, Indonesia: EAI. doi:10.4108/eai.1-4-2019.2287279

Marcus, C. (1998). A practical yet meaningful approach to customer segmentation. *The Journal of Consumer Marketing, 15*(5), 494-504. doi:DOI:10.1108/07363769810235974

Optimove. (2020, May). *Customer Segmentation: Definition, Models*. Retrieved May 23, 2021, from Otimove: https://www.optimove.com/resources/learning-center/customer-segmentation

P.Anitha, Malini M.Patil. (2019). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*. Retrieved May 24, 2020, from https://www.sciencedirect.com/science/article/pii/S1319157819309802

Poh, L. T. (2020, April 8). *The Most Important Data Science Tool for Market and Customer Segmentation.* Retrieved May 20, 2020, from Aisingapore: https://aisingapore.org/2020/04/the-most-important-data-science-tool-for-market-and-customer-segmentation/