BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

**Macro Management & Public Policies**
https://ojs.bilpublishing.com/index.php/mmpp

ARTICLE

# Data Driven Customer Segmentation for Vietnamese SMEs in Big Data Era

**Pham Thi Tam**[1] **Duong Minh Son**[2] **Trinh Le Tan**[3*] **Hoang Ha**[4]

1. Lien Chieu Ditstrict, Danang city, Vietnam
2. Dong A University, Danang city, Vietnam
3. FPT University, Danang city, Vietnam
4. University of Economics, The University of Danang, Da Nang, 550000, Vietnam

ARTICLE INFO

*Article history*
Received: 13 August 2021
Accepted: 17 August 2021
Published Online: 19 August 2021

*Keywords*:
Data driven
Customer segmentation
Behavioural segmentation
Clustering
Agglomerative

ABSTRACT

Almost Vietnamese big businesses often use outsourcing services to do marketing researches such as analysing and evaluating consumer intention and behaviour, customers' satisfaction, customers' loyalty, market share, market segmentation and some similar marketing studies. One of the most favourite marketing research business in Vietnam is ACNielsen and Vietnam big businesses usually plan and adjust marketing activities based on ACNielsen's report. Belong to the limitation of budget, Vietnamese small and medium enterprises (SMEs) often do marketing researches by themselves. Among the marketing researches activities in SMEs, customer segmentation is conducted by tools such as Excel, Facebook analytics or only by simple design thinking approach to help save costs. However, these tools are no longer suitable for the age of data information explosion today. This article uses case analysing of the United Kingdom online retailer through clustering algorithm on R package. The result proves clustering method's superiority in customer segmentation compared to the traditional method (SPSS, Excel, Facebook analytics, design thinking) which Vietnamese SMEs are using. More important, this article helps Vietnamese SMEs understand and apply clustering algorithm on R in customer segmenting on their given data set efficiently. On that basis, Vietnamese SMEs can plan marketing programs and drive their actions as contextualizing and/or personalizing their message to their customers suitably.

## 1. Introduction

Customer segmentation is a very important step in marketing research. Through customer segmentation allows businesses to understand customers' behavior and preferences, acquire knowledge about different customer groups for the allocation of resources and adaptation of product mixes, the development of new product/market approaches [31] and for satisfying customer segments according to their specific preferences and needs [10], the retention to loyalty customers and capturing new customers [19].

In Vietnam, businesses usually segment their customers based on four main approaches or mixed of them including Demographic, Behaviour, Geographic and Psychographic. However, different tools which Vietnamese's businesses use to analyze data, different size and complication of

**Corresponding Author:*
*Trinh Le Tan,*
*FPT University, Danang city, Vietnam;*
*Email: letandtu@gmail.com*

Distributed under creative commons license 4.0 DOI: https://doi.org/10.30564/mmpp.v3i2.3553

data set, different results they get. These differences are bigger between big enterprises and SEMs. In Vietnamese SMEs, data sets are usually small size and simple, so they can not employ a Data Driven approach, it means it use data to drive their marketing actions; and three tools that Vietnamese SMEs popularly used are Excel, Facebook analytics and Design Thinking.

With Excel application, it has been using for analyzing sales data of Vietnamese SMEs through describe customer features. It helps in identifying different customer groups with specific business requirements. However, using Excel has disadvantages such as the Excel data file itself only holds 1,080,000 rows without support and integration with other tools. Therefore, it is not suitable for analyzing big data. Besides, Excel bases on descriptive statistics to divide into customer clusters that are simple and not reality.

In recent years, some of Vietnamese medium enterprises have been applying some methods for customer segmenting, with emerging of sale online. These medium businesses track consumer behavior through optimize the entire customer experience on mobile devices, web, bots, in actual life, and more. Marketer can group into customer clusters through Facebook analytics. However, there is some defect of this tool. In order to take advantage of Facebook analytics truly, you must install more tools into your marketing' toolset which almost of Vietnamese SMEs' limitation. And that very toolkit may have exploded in its associations with Google Analytics, Google AdWords, and so on. Besides that, it shows disadvantageous of Facebook not recording the transactions of direct selling at the store like other selling software.

In addition, micro-businesses, especially startups implement traditional market segmentation such as Design thinking. This approach includes some steps such as going to the major stores, observing real groups of customer and tracking how factors of brands creating value (6P & 7P analysis). Outline portraits (5 segmentation criteria) of groups of customers. The most important is assuming the needs (Need), why they come and buy that brand (mapping analysis 6P/7P with assumptions of demand). Group all groups of observed customers into large customer sets in Consumer Segmentation model, eliminating duplicate groups. It only applies this method when the business is small and the customer data are not really much. This leads to enterprises facing difficulties in analyzing big data in order to draw a comprehensive picture of business performance, in which, customer segmentation is associated with an effective marketing strategy that requires a more effective tool to explore and analyze data.

In past ten years, the World's big enterprises as Facebook, Google, Amazon, Netflix, Alibaba, eBay,… can increase scale quickly by creating new business models. These companies are using the Internet to reach a geographically unconstrained mass market and unlimited customers. Those capabilities provide the barriers to entry that give these giants many advantages to outperform their competitors. In order to follow up this strategy, these companies often build and operate their business models based on Business Intelligence [26], especially upon customer-centric business intelligence. In this approach, big online retailers can understand customer's behavior through what they performed on platforms such as website, social media. It solves some concerned questions through customer centric business intelligence approach: How long has a customer stayed with each web page, and in which sequence has a customer visited a set of product webpages? Who are the most/least valuable customers to the business? Who are the most/least loyal customers? What are customers' purchasing behavior patterns in terms of various perspectives such as products/ items, regions and time and so on? [8]

With customer segmentation activity in big data, businesses usually use a clustering algorithm which is analysed in R or Python software. However, when make a comparison between R and Python, R has some advantages such as with complicated and big data, R is more productivity, visual and easy to use complicated function, testing and statistics models than Python.

This article wants to guide Vietnamese SMEs in customers segmentation through case analysing of the United Kingdom (UK) online retailer by using one of clustering algorithm on R package to analyse data. This case analysing is Data Driven in customer segmentation of the UK online retailer on a data set with 330,379 transactions. The primary purpose of this approach is to help businesses use data to drive their actions as contextualizing and/or personalizing its message to its customers, pricing policies, and another marketing programs through understanding its customers. The Agglomerative Clustering algorithm was used to segment customers into four groups. Some fundamental characteristics and recommendations of each segment were proposed. The analysis was performed with R 3.5.3.

The remainder of the paper is organized: In section 2, the background about Data Driven in customer segmentation is provided. In section 3, methodology is presented about the present situation about customer segmentation of Vietnamese SMEs and the performance on the target data set by Agglomerative Clustering analysis. The next section of this article is concluding remarks which review some good points which the customer segmentation approach in case study solved. The

Distributed under creative commons license 4.0

DOI: https://doi.org/10.30564/mmpp.v3i2.3553

last section is a recommendation in guiding framework to Vietnamese SMEs in customer segmentation.

## 2. Literature Review

### 2.1 Customer Segmentation

Customer segmentation is not only "one of the major way to operationalizing the marketing concept" but also helps companies in planning marketing strategy [39]. It is the key step of a marketing plan and at the center of successful marketing [16,28].

The importance of segmentation for marketing strategies is undeniable [21], in both academic and practical marketing [16]. There were more 100 research articles about this topic [39] and the number of papers have been increasing. Segmentation gives a good understanding of the need of the customers and helps in identifying the potential customers of the company. Dividing the customers into segments also increases the revenue of the company. It is believed that retaining the customers is more important than finding new customers. For instance, the company can deploy marketing strategies that are specific to an individual segment to retain the customer [9].

Customer segmentation is subgroups of consumers who will respond similarly to a market mix [30], a method of grouping customers in the market bases on specific criteria to identify and respond to the need of those customer groups [18]. At its core, it is about aggregating individual consumer behavior into a manageable number of groups that are mutually exclusive and share well-defined characteristics [40]. Alternatively, it can be understood "as the process of splitting customers, or potential customers, within a market into different groups, or segments within which customers have the same, or similar requirements satisfied by a distinct marketing mix" [29].

Segmentation provides businesses the opportunity to develop, target and fine tune products and market rationally and precisely on the demand side of the market [34]. As a result of customer segmentation, a company will have a portfolio of segmentation and then depending on the strategy of company, one or more segmentations will be chosen to target and differentiated marketing mix programs will be developed for each segmentation.

### 2.2 Advantages of Customer Segmentation

The customer segmentation could provide four vital benefits to marketing departments and management of a business:

First, customer segmentation allows the effective identification of the important customer groups that include the most profitable and loyal customers [10,24,39]. This method has also been used to determine the potential profit of customers by research efforts [6,7,8,14,22,32,38]. With these results, important decisions will be made toward high-value groups of customers with direct and concentrated efforts.

Second, customer segmentation allows businesses to understand customers' behavior and characteristics, then gain knowledge about diverse customer groups. By this way, it is possible to satisfy customer segments according to their specific preferences and needs [10]. Such knowledge provides many opportunities to more accurately tailor marketing actions and materials to individual customer expectations [11,27]. In this respect, several studies have applied customer segmentation to create discriminative customer management and effective marketing strategies for different customer groups [6,25,37].

Third, segmentation assists in customer retention and capturing new markets, because without a specific focus, customers may turn to market niche players who operate for specific segments [19]. Therefore, if they fail to identify and apply market segmentation businesses can lose competitive advantages when they compete with their competitors and in particular niche competition.

Finally, segmentation definitions become more relevant for understanding customers, for the allocation of resources and adaptation of product mixes, and for the development of new product/market approach [31].

### 2.3 Classification of Segmentation

There are many types of segmentation methods [17], but every way of segmenting market has merits and limitations which depend on the considered product market and the managerial objectives in segmenting [16].

Basically, three methods are acknowledged: Geographic segmentation, demographic segmentation, volume segmentation or "heavy_haft" theory [17,36]. However, a key problem of these methods is descriptive, relying on "an ex-post facto analysis of the kinds of people" rather than based on identifying and analyzing causal elements [17]. Such methods are now regarded as being not really effective in predicting of the future consumer behaviour [2].

Another point of view, there are three categories of segmentation such as behavioural, psychographic and profile segmentation [15]. Even though which kind of segmentation is chosen, also need to focus on the benefits which customers interest. The simple reason is "customers buy benefits, not products; customers buy an answer, a solution, to their specific set of existing needs and demands" [18], so customers will seek some differential benefits when deciding in buying/using a product/service [17]. Therefore, there is an existence of true

Distributed under creative commons license 4.0
DOI: https://doi.org/10.30564/mmpp.v3i2.3553

customer segments which are incurred from differential benefit soughts [2]. In three types of segmentations above, behavioral segmentation will base on some variables, such as benefits sought from the product (benefit segment), buying patterns such as frequency and volume of purchase, perceptions and beliefs to group the customers [2].

With the features of the original data set, this article approached on the behavioral segmentation with the chosen variable is buying patterns: frequency and volume of purchase.

### 2.4 Data Driven in Customer Segmentation

In customer segmentation, there are two ways used: (1) Typological segmentation and (2) Data Driven or post hoc in segmentation [12,4].

The typological segmentation is similar to "a priori segmentation designates groups of consumers who are similar in terms of some factor or factors that are known or felt in advance to be related to product/service consumer" [30] and generally multidimensional and conceptual [3]. Contrarily, a Data Driven segmentation, or post hoc segmentation, is empirical. A company employs a Data Driven approach, and it means it uses data to drive their actions as promotion programs, pricing policies and another Marketing strategy to its customers. The goal of this method is classifying cases according to their measured similarity on observed variables [3]. With this approach, the first point is an empirical data set, and quantitative techniques of data analysis are used for this data set to derive a grouping [12]. Generally, the basic difference between typology and Data Driven is conceptual and empirical [3].

### 2.5 Clustering Analysis Method

## Understanding Clustering Analysis

Most researches, using Data Driven segmentation to group individuals in the market, use a statistical method based on one of the family of cluster analysis [13]. Cluster analysis is a method for the analysis and organizing an enormous bulk of multivariate or scientific data to decrease information overload and to discover relationship and classes in unorganized data sets. It can assist in discovery the structure or causality in complex bodies of data [1,13].

Dolnicar (2002) indicates that cluster analysis is "a toolbox of highly interdisciplinary techniques of multivariate data analysis" by dividing number of individuals into subgroups based on "a pre-specified criterion (e.g. minimal variance within each resulting cluster) which is assumed to reflect the similarity of individuals within the subgroups and the dissimilarity between them" [12].

Deepak (2019) proposes clustering analysis is a prominent technique to segment market based on benefit sought [20]. Several studies by Soutar McNeil (1991), Minhas and Jacobs (1996), Brunner and Siegrist (2011) and Li et al. (2011) have used benefit segmentation with factor and cluster analysis method to segment market [35,33,5,22].

Importantly, cluster analysis conducts classifications from initially uncategorical data, but not a type of identification (also called dissection), it is sorting entities based on certain classes after identifying, in many ways, it is "the reverse-engineering of classification taxonomies" [13,1].

### 2.6 Clustering Method

In the last four decades, the computational power has increased, so there are many studies using similar techniques in data mining and pattern recognition [13,1]. The starting point for clustering analysis is a multidimensional data set. The process of clustering includes deciding which algorithm should be used to analyze the data, which measure of association is the most appropriate, how many groups of respondents should emerge. This complex first step is followed by the actual data analytic step which results in a partition of the respondents (every respondent is assigned to one of subgroups), and forms the basis for interpretation by studying differences in group responses [12].

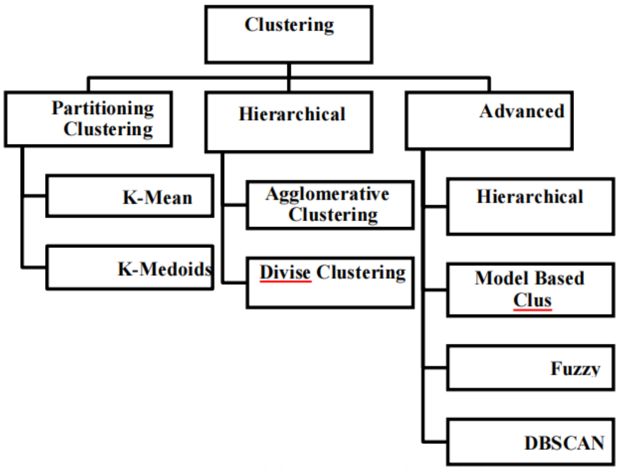


**Figure 1.** Diagram of clustering analysing techniques.

Figure 1 showed that there are three types of clustering analysis: Partitioning Clustering, Hierarchical Clustering, Advanced Clustering [23]. Specifically, partitioning clustering is used to classify observations within a data set into multiple groups based on their similarity. Hierarchical clustering is an alternative approach to partitioning clustering for grouping objects based on their similarity in contrast to partitioning clustering, hierarchical clustering

Distributed under creative commons license 4.0

DOI: https://doi.org/10.30564/mmpp.v3i2.3553

does not require to pre-specify the number of clusters to be produced [23].

Partitioning Clustering includes three methods: K-Mean, K-Medoids, Clara; Hierarchical Clustering has two methods such as Agglomerative Clustering, Divisive Clustering; and Advanced Clustering includes four methods as Hierarchical K-mean, Model Based Clus, Fuzzy, SOM, DBSCAN [23].

### 2.7 Hierarchical Clustering, Agglomerative Clustering

Hierarchical clustering can be subdivided into two types: Agglomerative clustering and Divise clustering. With the data types of this study, numeric data, the Agglomerative clustering is used for analyzing data. Agglomerative clustering in which, each observation is initially considered as a cluster of its leaf. Then, the most similar clusters are successively merged until there is just one single big cluster (root). The agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. It's also known as AGNES (Agglomerative Nesting). The algorithm starts by treating each object as a singleton cluster. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects. The result is a tree-based representation of the objects named dendrogram [23].

## 3. Methodology

We conducted the study in two steps. The first step is to assess the current situation in customer segmentation activities of Vietnamese SMEs. The survey sample includes 50 Vietnamese SMEs which operates in different industries such as tourism and hospitality, offline and online retailers (clothes, bags and accessories, beauty product, shoes), food & beverage (coffee shop, food store), seafood, real estate. Chief Executive Officers of small businesses and Head of Marketing & Sale Department of medium enterprises are interviewed. Depth-interviewing method by qualitative questionnaires is used. Survey time is from October 2019 to May 2020. The second step is analyzing our dataset mined from a UK online retailer, which had 541,910 transactions of customers across most parts of Europe in years 2010 and 2011.

## 4. Result

### 4.1 Reality of Customer Segmentation Activities in Vietnamese SMEs

The survey's result shows that only 30% of companies surveyed do market research, mainly tourism & hospitality businesses and these business's market research activities mainly comprise some basic steps: identify the problem to be studied, design the content to be surveyed, select the appropriate survey method, collect and analyse data. The remaining 70% of enterprises do not do market research, but mainly analyse available data or follow industrial information to make decisions.

Regarding customer segmentation activities, 42% of businesses segmented customers according to Demographic, Behavior, Geographic and Psychographic; in these 42% businesses, 78% of enterprises use Excel software, the rest use Excel. The remaining 46% of businesses segment mainly based on Demographic, Geographic and the technical tool which used is a combination of Design Thinking and Excel. The remaining 12% of businesses have absolutely no customer segmentation.

Among enterprises that use software to process data, they show that 100% Marketing & Sale staffs are responsible for collecting and analysing data, the convenience in using the current softwares is not too complicated and the Marketing staffs are familiar and have skills to implement on software.

In survey sample, 70% of businesses said that they have heard and known about terms "Big Data applied in Marketing" or something like that, 30% have never heard of it. The reason 70% of these companies not applied Big Data to market research is not knowing how to start, not having ability in engineering, software, cost and capable employees.

Summarily, Vietnamese SMEs mainly use Exel, Facebook Analytics, and Design Thinking approach to segment their customers. The disadvantage is these softwares are not responsive when businesses need deeper and more complex information about markets and customers and big data sets. So, their marketing decision seldom based on data driven, which really shows what their customers' characteristics. This is the biggest barrier to help them maintain their segments or jump into any new segments.

### 4.2 Clustering in Our Case Study with Dataset from a UK Online Retailer

#### Data pre-processing Original Data Set

The original data set was from one online retailer, had 541,910 transactions of customers across most parts of Europe in years 2010 and 2011. This data set has 8 variables as shown in Table 1.

From this data set, this article chose transactions to

Distributed under creative commons license 4.0

DOI: https://doi.org/10.30564/mmpp.v3i2.3553

**Table 1.** Variables customer transaction dataset (541,910 instances)

| Variable name | Data type | Description; typical values & meanings |
|---|---|---|
| Customer ID | Nominal | Customer Identifier; a 5-digit integral number uniquely assigned to each customer account |
| InvoiceNo | Nominal | Invoice Number; a 6-digit integral number uniquely assigned to each transaction |
| StockCode | Nominal | Product (item) code; assigned to each distinct product |
| Description | Nominal | Product (item) name;<br>CREAM CUPID HEARTS COAT HANGER |
| Quantity | Numeric | The quantity of each product (item) per transaction |
| InvoiceDate | Numeric | The day and time when each transaction was generated; 6/8/2011 15:26 |
| UnitPrice | Numeric | Product price per unit in sterling; £2.55 |
| Country | Nominal | Delivery address country; England |

become the sample. These transactions are from 1 January 2011 to 31 December 2011, in UK, besides in which these transactions were chosen, some observations that had missing value in one variable or unreasonable value (the price variable is equal 0), were to be excluded. So, size of sample is 3,813 customer identify corresponding to 330,379 transactions, is seen as the original data set.

## Data Pre-Processing

In order to conduct the clustering analysis, the original data set needs to be pre-processed.

The major steps involved in this process are:

- Select appropriate variables from the original data set: For this article, there are five variables have been chosen: *Customer ID, InvoiceNo, InvoiceDate, Quantity, UnitPrice.*
- Create an aggregated variable named *Amount*, by multiplying *Quantity* with *Price*, which gives the total amount of money spent per product/item in each transaction.

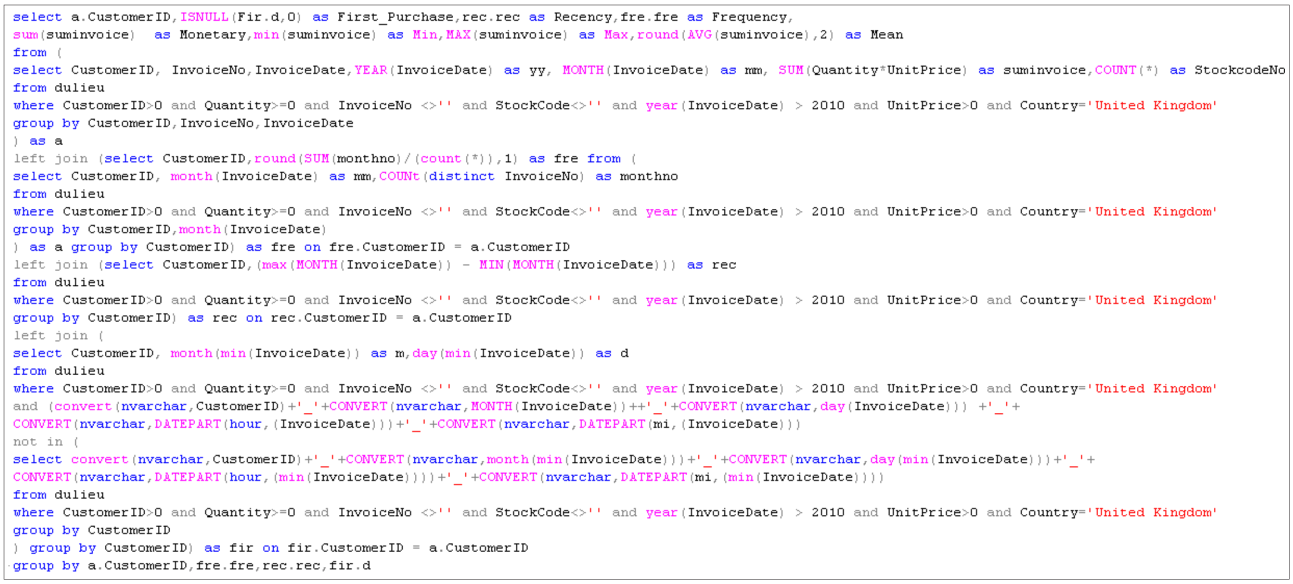
Separate the variable *InvoiceDate* into two variables *Date* and *Time*. This allows different transactions created by the same consumer on the same day, but at different times to be treated separately. Also, a simpler case is different transactions created by the same consumed in the same month but at different dates to be treated separately.

Sort out the data set by *Customer ID* and create four essential aggregated variables: *First_Purchase, Recency*, *Frequency* and *Monetary*. Calculate the values of these variables per Customer ID.

After conducting these steps, a target data set has been generated. The original data set was in MS Excel format and was transformed into the final target data set in SQL Server 2012. Code of the transformative process from the original data set to the target data set are shown in Table 2.

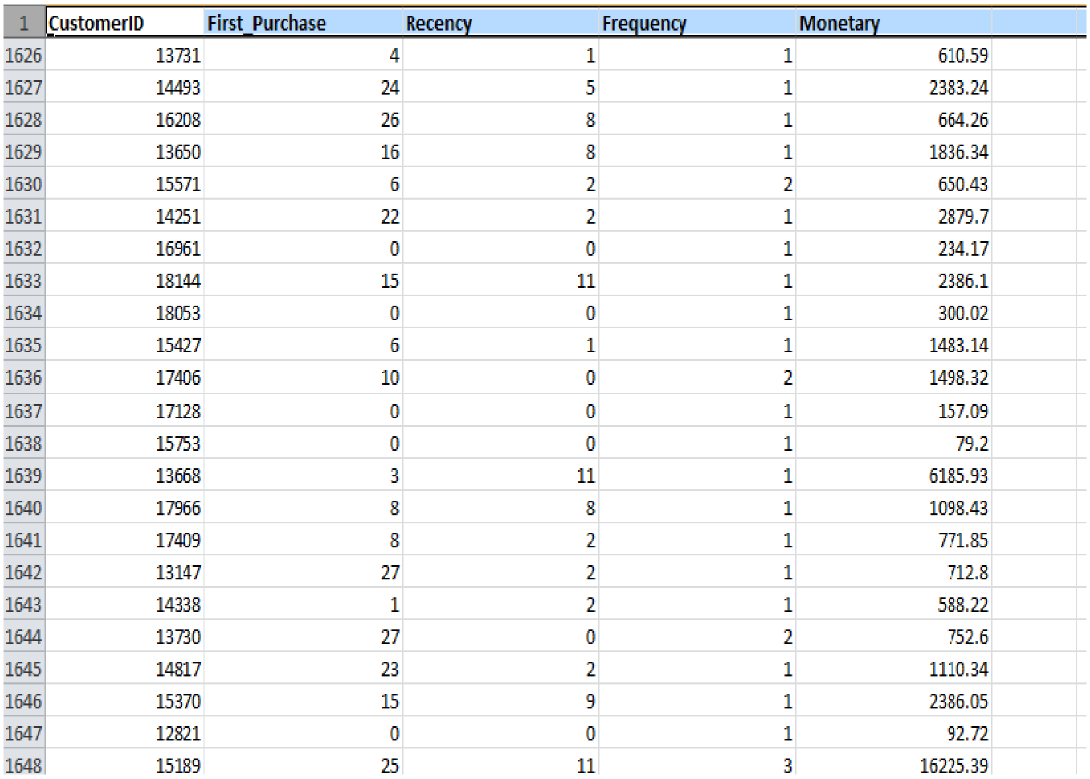
And part of the target data set is shown in Figure 2,

**Table 2.** SQL Server codes for transformative process from original data set to target data set



```
select a.CustomerID,ISNULL(Fir.d,0) as First_Purchase,rec.rec as Recency,fre.fre as Frequency,
sum(suminvoice)  as Monetary,min(suminvoice) as Min,MAX(suminvoice) as Max,round(AVG(suminvoice),2) as Mean
from (
select CustomerID, InvoiceNo,InvoiceDate,YEAR(InvoiceDate) as yy, MONTH(InvoiceDate) as mm, SUM(Quantity*UnitPrice) as suminvoice,COUNT(*) as StockcodeNo
from dulieu
where CustomerID>0 and Quantity>=0 and InvoiceNo <>'' and StockCode<>'' and year(InvoiceDate) > 2010 and UnitPrice>0 and Country='United Kingdom'
group by CustomerID,InvoiceNo,InvoiceDate
) as a
left join (select CustomerID,round(SUM(monthno)/(count(*)),1) as fre from (
select CustomerID, month(InvoiceDate) as mm,COUNt(distinct InvoiceNo) as monthno
from dulieu
where CustomerID>0 and Quantity>=0 and InvoiceNo <>'' and StockCode<>'' and year(InvoiceDate) > 2010 and UnitPrice>0 and Country='United Kingdom'
group by CustomerID,month(InvoiceDate)
) as a group by CustomerID) as fre on fre.CustomerID = a.CustomerID
left join (select CustomerID,(max(MONTH(InvoiceDate)) - MIN(MONTH(InvoiceDate))) as rec
from dulieu
where CustomerID>0 and Quantity>=0 and InvoiceNo <>'' and StockCode<>'' and year(InvoiceDate) > 2010 and UnitPrice>0 and Country='United Kingdom'
group by CustomerID) as rec on rec.CustomerID = a.CustomerID
left join (
select CustomerID, month(min(InvoiceDate)) as m,day(min(InvoiceDate)) as d
from dulieu
where CustomerID>0 and Quantity>=0 and InvoiceNo <>'' and StockCode<>'' and year(InvoiceDate) > 2010 and UnitPrice>0 and Country='United Kingdom'
and (convert(nvarchar,CustomerID)+'_'+CONVERT(nvarchar,MONTH(InvoiceDate))++'_'+CONVERT(nvarchar,day(InvoiceDate))) +'_'+
CONVERT(nvarchar,DATEPART(hour,(InvoiceDate)))+'_'+CONVERT(nvarchar,DATEPART(mi,(InvoiceDate)))
not in (
select convert(nvarchar,CustomerID)+'_'+CONVERT(nvarchar,month(min(InvoiceDate)))+'_'+CONVERT(nvarchar,day(min(InvoiceDate)))+'_'+
CONVERT(nvarchar,DATEPART(hour,(min(InvoiceDate))))+'_'+CONVERT(nvarchar,DATEPART(mi,(min(InvoiceDate))))
from dulieu
where CustomerID>0 and Quantity>=0 and InvoiceNo <>'' and StockCode<>'' and year(InvoiceDate) > 2010 and UnitPrice>0 and Country='United Kingdom'
group by CustomerID
) group by CustomerID) as fir on fir.CustomerID = a.CustomerID
group by a.CustomerID,fre.fre,rec.rec,fir.d
```

Distributed under creative commons license 4.0 DOI: https://doi.org/10.30564/mmpp.v3i2.3553



| 1 | CustomerID | First_Purchase | Recency | Frequency | Monetary |
|---|---|---|---|---|---|
| 1626 | 13731 | 4 | 1 | 1 | 610.59 |
| 1627 | 14493 | 24 | 5 | 1 | 2383.24 |
| 1628 | 16208 | 26 | 8 | 1 | 664.26 |
| 1629 | 13650 | 16 | 8 | 1 | 1836.34 |
| 1630 | 15571 | 6 | 2 | 2 | 650.43 |
| 1631 | 14251 | 22 | 2 | 1 | 2879.7 |
| 1632 | 16961 | 0 | 0 | 1 | 234.17 |
| 1633 | 18144 | 15 | 11 | 1 | 2386.1 |
| 1634 | 18053 | 0 | 0 | 1 | 300.02 |
| 1635 | 15427 | 6 | 1 | 1 | 1483.14 |
| 1636 | 17406 | 10 | 0 | 2 | 1498.32 |
| 1637 | 17128 | 0 | 0 | 1 | 157.09 |
| 1638 | 15753 | 0 | 0 | 1 | 79.2 |
| 1639 | 13668 | 3 | 11 | 1 | 6185.93 |
| 1640 | 17966 | 8 | 8 | 1 | 1098.43 |
| 1641 | 17409 | 8 | 2 | 1 | 771.85 |
| 1642 | 13147 | 27 | 2 | 1 | 712.8 |
| 1643 | 14338 | 1 | 2 | 1 | 588.22 |
| 1644 | 13730 | 27 | 0 | 2 | 752.6 |
| 1645 | 14817 | 23 | 2 | 1 | 1110.34 |
| 1646 | 15370 | 15 | 9 | 1 | 2386.05 |
| 1647 | 12821 | 0 | 0 | 1 | 92.72 |
| 1648 | 15189 | 25 | 11 | 3 | 16225.39 |

**Figure 2.** Samples of the target data set

the variables in the target data set and their descriptive statistics are shown in Tables 3 and 4. Finally, the target data set was uploaded into R 3.5.3 for analysis.

**Table 3.** Variables in the target data set

| Variable name | Data type | Description |
|---|---|---|
| Customer ID | Nominal | Customer Identifier |
| First_Purchase | Numeric | Time (Date) after the first purchase per Customer ID |
| Recency | Numeric | Equal to time (month) in the last purchase - Time (month) in the first purchase per Customer ID |
| Frequency | Numeric | Frequency of purchase per Customer ID |
| Monetary | Numeric | Total of money in sterling per Customer ID |

**Table 4.** Summary of the target data set (3813 instances)

| Variable name | Minimum | Median | Maximum |
|---|---|---|---|
| First_Purchase | 0.0 | 7.0 | 31.0 |
| Recency | 0.0 | 3.0 | 11.0 |
| Frequency | 1.0 | 1.0 | 14.0 |
| Monetary | 3.8 | 637.5 | 231822.7 |

## Agglomerative clustering Analysis

From the target data set, this study was conducted by Agglomerative clustering analysis to divide customers into some meaningful segments in the view of First_Purchase, Recency, Frequency and Monetary values. With the data types of this study, numeric data, the Agglomerative clustering was chosen to perform with the support of R 3.5.3. Some steps in R with Agglomerative clustering are below:

First, in the target Data set, four variables *First_Purchase, Recency*, *Frequency* and *Monetary* were chosen as input for the clustering analysis.

Second, the target data set was converted to a numeric matrix with standardised variables (UK7 Data set).

Third, Euclidean distance and Ward Linkage were used in Agglomerative clustering method.

The results with four clusters are shown in Tables 5 and 6, and Dendogram of four cluster are indicated in Figure 3.

**Table 5.** Instances in each cluster

| Cluster. No | Frequency | Frequency Percentage |
|---|---|---|
| 1 | 1726 | 45 |
| 2 | 1765 | 46 |
| 3 | 319 | 8 |
| 4 | 3 | 1 |

Distributed under creative commons license 4.0
DOI: https://doi.org/10.30564/mmpp.v3i2.3553

**Table 6.** Statistics of clusters

| Cluster. No | Minimum | Mean | Median | Maximum | Cluster. No | Minimum | Mean | Median | Maximum |
|---|---|---|---|---|---|---|---|---|---|
| Cluster 1 | | | | | Cluster 3 | | | | |
| First_Purchase | 0.00 | 1.21 | 0.0 | 15.00 | First_Purchase | 0.0 | 16.55 | 17.00 | 31.0 |
| Recency | 0.00 | 0.40 | 0.0 | 4.00 | Recency | 0.0 | 4.26 | 2.00 | 11.0 |
| Frequency | 1.00 | 1.00 | 1.0 | 1.00 | Frequency | 1.0 | 2.34 | 2.00 | 14.0 |
| Monetary | 3.75 | 423.66 | 302.1 | 4481.35 | Monetary | 6.9 | 6954.69 | 1447.14 | 84351.3 |
| Cluster 2 | | | | | Cluster 4 | | | | |
| First_Purchase | 1.00 | 16.91 | 17.0 | 31.00 | First_Purchase | 7.0 | 9.00 | 9 | 11.0 |
| Recency | 0.00 | 6.74 | 7.0 | 11.00 | Recency | 7.0 | 9.33 | 10 | 11.0 |
| Frequency | 1.00 | 1.00 | 1.0 | 1.0 | Frequency | 1.0 | 3.33 | 4 | 5.0 |
| Monetary | 20.35 | 1795.12 | 1223.9 | 19015.38 | Monetary | 168472.5 | 197605.7 | 192522 | 231822.7 |

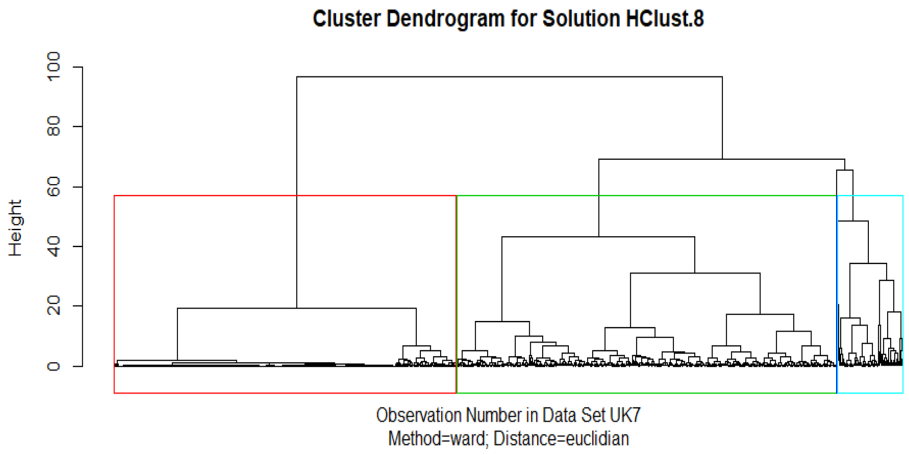


**Figure 3.** Dendogram of Aggolomerative Clustering Analysis result

**Table 7.** Comparative characteristics between 4 clusters

| Cluster. No | Total data (%) | Monetary (Mean) | Frequency (Mean) | Recency (Mean) | First_Purchase (Mean) |
|---|---|---|---|---|---|
| Cluster 1 | 10% | 423.66 | 1 | 0.4 | 1.21 |
| Cluster 2 | 48% | 1795.12 | 1 | 6.74 | 16.91 |
| Cluster 3 | 33% | 6954.69 | 2.34 | 4.24 | 16.55 |
| Cluster 4 | 9% | 197605.7 | 3.33 | 9.33 | 9.0 |

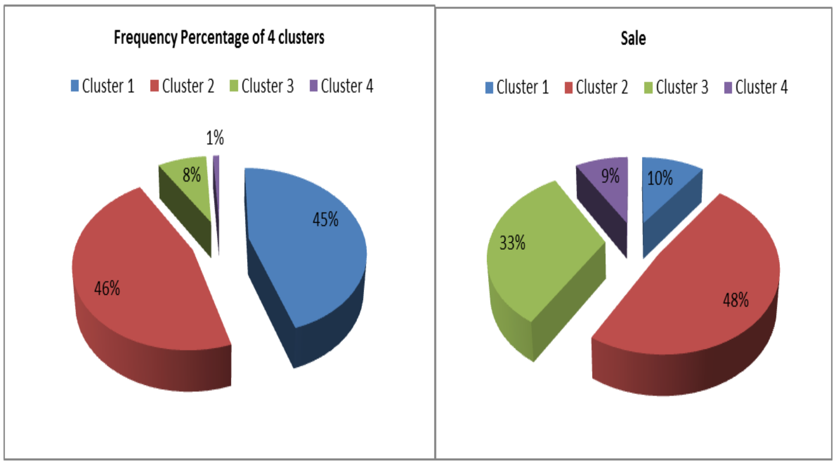


**Figure 4.** Customer segmentation (left) and associated sales (right) by cluster

### 4.3 Understanding the Clusters and Recommendations

Examining Table 6, 7 and Figures 3 and 4, it is curious to see that each cluster indeed contains a group of consumers that have meaningful and certain distinct characteristics as detailed below:

Cluster 1 related to 1726 consumers, composed of 45 percent of the target data set. Half of this cluster didn't repurchase, and haft of existence repurchased after the first time before the 15th. Besides, the total money of cluster 1 was 700726.5, composed of 10 percent of the target data set, the average value of money per one customer was 423.66, and the average value of frequency was only 1.0. With this information, we could see this group distributed second low total money, lowest value of money per customer, frequency and recency. Therefore, this group seems to be the least profitable group. This group includes many customers buying the first time but the size of this group was quite high (2rd high), so, in the long-term view, some consumers might be potentially very highly profitable or not, and the business need to have the suitable marketing policies to attract them.

Cluster 2 related to 1765 consumers, composed of 46 percent of the target data set. Half of cluster 2 repurchased after the first time before the 17th, and haft after the 17th. Besides, the total money of cluster 1 was 3297639, composed of 48 percent of the target data set, the average value of money per one customer was 1795.12, and the average value of frequency was only 1.0. So, we can categorize this group as the highest total money, medium value of money per customer, lowest frequency, but the second high recency. From this analysis, this group could be seen as the second profitable group, contribute a large part of regular revenue of enterprises, but the frequency of purchasing wasn't high, so some instantly necessary activities in products" strategies, pricing policies, optimizing customers" experience through shopping platforms, ect to help customers get more value from buying of this online retailer, therefore, they can increase number of purchasing.

Cluster 3 related to 319 consumers, composed of 8 percent of the target data set. Like cluster 2, median of the First Purchase variable in this cluster was 17, it means the dates after the first time (purchase) spread in all dates of month. The total money of this cluster was 2218547, composed of 33 percent of the target data set, the average value of money per one customer was 6954.69, and the average value of frequency was only 2.34. So, we can see this group had the second high total money, second high money per customer, second high frequency, and the lowest recency. Definitively, this group seems to be the highest profitable group. Therefore, the enterprise should have suitable marketing strategy to increase benefit for customers not only in money but also in users' experience through products strategies, customer care programs to build this segment become a stable loyalty segment.

Cluster 4 related to 3 consumers, composed of 1 percent of the entire population. This segment usually repurchased after the first time from the 7th-11th in month. The total money of cluster 4 was 592817.1, composed of 9 percent of the target data set, the average value of money per one customer was 197605.7, and the average value of frequency was only 3.33. This group had the lowest total money (but the distance not too far from the second low group), the highest money per customer, the highest frequency, but also the highest recency in the context size of this group is the smallest, 3 customers. So, we can see this group related to organizational customers. Considering all benefit aspects of this segmentation, such as distribution of this segment and not too much complicated to serve the small segment with 3 customers, we can see this group as the third profitable segment.

### 4.4 Implications for Vietnamese SMEs

Given the lack of financial resources of SMEs in market research and customer segmentation, big data-based analytical processing tools will be extremely important. This case analysing gives a framework to guide Vietnamese SMEs in segmenting their customers in a given big data set efficiently as step-by-step below:

Reading data set to define the characteristics which can be used for segmenting in next steps.

Preprocessing original data set to final data set, which is suitable to segmenting characteristics in first step (Use any software suitably: Excel, SPSS, STATA, SQL…).

Run Clustering algorithm on R (as this research did).

Reading, analysing output to get meanings under output.

## 5. Conclusions

This article's customer segmentation approach gives a guideline for Vietnamese SMEs through case analysing. A case study has been presented in this article to demonstrate how to group customers of online retailer into different segments by Data Driven approach and clustering algorithm on R package. From the characteristics of each group, the business better understands its customers and then adopt appropriate marketing strategies for different segments. Data analyzing process included 2 steps: data pre-processing and data analyzing method conducting.

Vietnamese SMEs should take advantage of rich data sources and have the appropriate strategy to collect and develop their own database to better understand their customers. Thereby creating sustainable competitive advantages over competitors.

## References

[1] ANDERBERG, M. R. (1973). 6. Hierarchical clustering methods. Cluster Analysis for Applications, 132-156.

[2] Arunachalam, D., & Kumar, N. (2018). Benefit-based consumer segmentation and performance evaluation of clustering approaches: An evidence of data-driven decision-making. Expert Systems with Applications, 111, 11-34.

[3] Bailey, K. D. (1994). Typologies and taxonomies: An introduction to classification techniques (Issue 102). Sage.

[4] Balasubramanian, S., Gupta, S., Kamakura, W., & Wedel, M. (1998). Modeling large data sets in marketing. Statistica Neerlandica, 52(3), 303-323.

[5] Brunner, T. A., & Siegrist, M. (2011). A consumer-oriented segmentation study in the Swiss wine market. British Food Journal, 113(3), 353-373. https://doi.org/10.1108/00070701111116437.

[6] Chan, C. C. H. (2008). Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. Expert Systems with Applications, 34(4), 2754-2762.

[7] Chen, J., & Bell, P. C. (2012). Implementing market segmentation using full-refund and no-refund customer returns policies in a dual-channel supply chain structure. International Journal of Production Economics, 136(1), 56-66.

[8] Chen, M.-Y., Huang, M.-J., & Cheng, Y.-C. (2009). Measuring knowledge management performance using a competitive perspective: An empirical study. Expert Systems with Applications, 36(4), 8449-8459. https://doi.org/10.1016/j.eswa.2008.10.067.

[9] Christy, A.J., Umamakeswari, A., Priyatharsini, L. and Neyaa, A., 2018. RFM ranking-An effective approach to customer segmentation. Journal of King Saud University-Computer and Information Sciences.

[10] Dibb, S. (1998). Market segmentation: Strategies for success. Marketing Intelligence & Planning, 16(7), 394-406. https://doi.org/10.1108/02634509810244390.

[11] Dibb, S., & Simkin, L. (1997). A program for implementing market segmentation. Journal of Business & Industrial Marketing.

[12] Dolnicar, S. (2002). A review of unquestioned standards in using cluster analysis for data-driven market segmentation.

[13] Dunn, G., Everitt, B. S., & Pickles, A. (1993). Modelling Covariances and Latent Variables Using EQS. CRC Press.

[14] Dursun, A., & Caber, M. (2016). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. Tourism Management Perspectives, 18, 153-160.

[15] Fahy, J., & Jobber, D. (2015). Foundations of marketing.

[16] Grover, R., & Srinivasan, V. (1987). A Simultaneous Approach to Market Segmentation and Market Structuring. Journal of Marketing Research, 24(2), 139-153. https://doi.org/10.1177/002224378702400201.

[17] Haley, R. I. (1968). Benefit segmentation: A decision-oriented research tool. Journal of Marketing, 32(3), 30-35.

[18] Harrigan, K. R. (1985). An application of clustering for strategic group analysis. Strategic Management Journal, 6(1), 55-73. https://doi.org/10.1002/smj.4250060105.

[19] Jenkins, M., & McDonald, M. (1997). Market segmentation: Organizational archetypes and research agendas. European Journal of Marketing, 31(1), 17-32. https://doi.org/10.1108/03090569710157016.

[20] Jurek-Loughrey, A., & P, D. (Eds.). (2019). Linking and Mining Heterogeneous and Multi-view Data (1st ed. 2019). Springer International Publishing: Imprint: Springer. https://doi.org/10.1007/978-3-030-01872-6.

[21] Kamakura, W. A., & Wedel, M. (1997). Statistical Data Fusion for Cross-Tabulation. Journal of Marketing Research, 34(4), 485-498. https://doi.org/10.1177/002224379703400406

[22] Kao, Y.-T., Wu, H.-H., Chen, H.-K., & Chang, E.-C. (2011). A case study of applying LRFM model and clustering techniques to evaluate customer values. Journal of Statistics and Management Systems, 14(2), 267-276.

[23] Kassambara, A. (2018). Machine learning essentials (Edition 1). STHDA.

[24] Kotler, P. (1997). Marketing management: Analysis, planning, implementation and control.

[25] Li, D.-C., Dai, W.-L., & Tseng, W.-T. (2011). A two-stage clustering method to analyze customer characteristics to build discriminative customer management: A case of textile manufacturing business. Expert Systems with Applications, 38(6), 7186-7191.

[26] Luhn, H. P. (1958). A Business Intelligence System. IBM Journal of Research and Development, 2(4), 314-319. https://doi.org/10.1147/rd.24.0314.

Distributed under creative commons license 4.0

DOI: https://doi.org/10.30564/mmpp.v3i2.3553

[27] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1(14), 281-297.

[28] McDonald, M. (2010). A brief review of marketing accountability, and a research agenda. Journal of Business & Industrial Marketing, 25(5), 383-394. https://doi.org/10.1108/08858621011058142.

[29] MCDONALD, M., & DUNBAR, I. (1998). Segmentation. How to Do It, I-low to Profit from It. London: MacMillan Press.

[30] Myers, J. H., & Tauber, E. M. (2011). Market structure analysis. Marketing Classics Press.

[31] Palmer, R. A., & Millier, P. (2004). Segmentation: Identification, intuition, and implementation. Industrial Marketing Management, 33(8), 779-785. https://doi.org/10.1016/j.indmarman.2003.10.007.

[32] Safari, F., Safari, N., & Montazer, G. A. (2016). Customer lifetime value determination based on RFM model. Marketing Intelligence & Planning.

[33] Singh Minhas, R., & Jacobs, E. M. (1996). Benefit segmentation by factor analysis: An improved method of targeting customers for financial services. International Journal of Bank Marketing, 14(3), 3-13. https://doi.org/10.1108/02652329610113126.

[34] Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. Journal of Marketing, 21(1), 3-8.

[35] Soutar, G. N., & McNeil, M. M. (1991). A Benefit Segmentation of the Financial Planning Market. International Journal of Bank Marketing, 9(2), 25-29. https://doi.org/10.1108/02652329110140194.

[36] Twedt, D. W. (1964). How Important to Marketing Strategy Is the" Heavy User"? Journal of Marketing (Pre-1986), 28(000001), 71.

[37] Wei, J.-T., Lin, S.-Y., Weng, C.-C., & Wu, H.-H. (2012). A case study of applying LRFM model in market segmentation of a children's dental clinic. Expert Systems with Applications, 39(5), 5529-5533.

[38] Wiedmann, K.-P., Hennigs, N., & Siebels, A. (2009). Value-based segmentation of luxury consumption behavior. Psychology & Marketing, 26(7), 625-651.

[39] Wind, Y. (1978). Issues and Advances in Segmentation Research. Journal of Marketing Research, 15(3), 317-337. https://doi.org/10.1177/002224377801500302.

[40] Yankelovich, D., & Meer, D. (2006). Rediscovering market segmentation. Harvard Business Review, 84(2), 122.