CUSTOMER SEGMENTATION
IN E-COMMERCE

A MINI PROJECT REPORT

Submitted by

V.PAVITHRA (312217205065)
V.PRATHEEKSHA(312217205067)
P.PRIYA DHARSHINI(312217205068)

BACHELOR OF TECHNOLOGY

IN

INFORMATION TECHNOLOGY

SSN COLLEGE OF ENGINEERING

APRIL 2020

SSN COLLEGE OF ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report **"CUSTOMER SEGMENTATION IN E-COMMERCE"** is the bonafide work of **PAVITHRA.V(312217205065), PRATHEEKSHA.V(312217205067) AND PRIYADHARSHINI.P(312217205068)** who carried out the mini project work under my supervision.

SIGNATURE	**SIGNATURE**
Dr. T.NAGARAJAN	Dr.S.MOHANAVALLI
HEAD OF THE DEPARTMENT	**SUPERVISOR**
Professor	Associate Professor
Department of Information Technology,	Department of Information Technology,
SSN College of Engineering	SSN College of Engineering
SSN Nagar, Kalavakkam - 603 110	SSN Nagar, Kalavakkam - 603 110

Submitted to the Viva voce Examination held on ______________________

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

I thank **ALMIGHTY GOD** who gave me the wisdom to complete this project. My sincere thanks to our beloved founder **Mr. Shiv Nadar, Chairman, HCL Technologies**. I also express my sincere thanks to our Principal **Dr. S. Salivahanan**, for all the help he has rendered during this course of study.

My heartfelt gratitude goes to **Dr. T. Nagarajan, Professor and Head of the Department, IT** for his words of advice and encouragement and I also express hearty gratitude to project Coordinator **Dr. V. Thanikachalam** and professors of the department for their scholarly guidance.

I profusely thank my supervisor **Dr. S. Mohanavalli**, Associate Professor, Department of Information Technology, for her innovative ideas, keen interest, suggestions, guidance and co-operation that paved the way for the successful completion of the project work.

I also thank all the faculty of the IT department for their kind advice, support and encouragement and last but not the least I thank my parents and my friends for their moral support and valuable help.

ABSTRACT

This project presents a segmentation of e-commerce customers. Competition between the e-commerce websites is increasing more and more. Change or die have become the simple rule of marketing in today's world. As more and more businesses are coming up every day, it has become significantly important for the old businesses to apply marketing strategies to sustain in the market.

Objective for customer segmentation is determining what price different groups of consumers are willing to pay for a product. When the customers have been divided into segments based on what they can afford to pay, segments that pay the lowest or highest prices can be focussed. For those segments that can afford more, additional features or a higher level of services can be offered.

Data mining (DM) and Machine Learning (ML) techniques can be used to solve marketing problems in e-commerce. The most common statistical clustering technique K-means is used to classify the customers into various groups. After identification of targeted customers and their associative buying pattern, the business managers can take the strategic profitable decisions accordingly.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

In the emerging global economy, e-commerce and e-business have increasingly become a necessary component of business strategy and a strong catalyst for economic development. E-commerce brings convenience for customers as they do not have to leave home and only need to browse websites online. It could help customers buy a wider range of products and save customers' time. But the buying behaviours of e-commerce consumers are not homogeneous.

According to [1], due to the diverse range of products and services available in the market as well as the intense competition among organizations, customer relationship management has come to play a significant role in the identification and analysis of a company's best customers and the adoption of best marketing strategies to achieve and sustain competitive advantage.

Customer segmentation is used for dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits.

1.2 MOTIVATION

Customer base is increasing exponentially day by day. Challenge for the companies is to cater the needs of customers. The problem that comes up is the overloading of information. This problem can be overcome by an implementation of personalization in ecommerce services such as providing product recommendation, links recommendation, ads or text and graphics that correspond to the users' characteristics and needs[9]. In addition to solving the problem of overloaded information, personalized services in ecommerce can maintain customer loyalty of existing customers, getting new customers by providing service to customers in accordance with their needs and characteristics.

There are many reasons why Ecommerce store owners fail to target the desired customers, and one of the causes is there is no adequate segmentation of existing customers in their business. Most of the merchants live with the presumption that mass marketing will bring them more sales and customers. But this is a costly, time consuming and unproductive approach to deal with a large customer base at a time. Instead, use the concepts of customer segmentation and target the relevant customer in the digital sphere. Digital marketers can make the best use of this concept to implement it while targeting the audience. Data mining techniques are used to do the bifurcation of customers.

1.3 CLUSTERING

It is a type of unsupervised learning method. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labelled responses.

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them[11].

Clustering techniques reveal internally homogeneous and externally heterogeneous groups[2]. Customers vary in terms of behaviour, needs, wants and characteristics and the main goal of clustering techniques is to identify different customer types and segment the customer base into clusters of similar profiles so that the process of target marketing can be executed more efficiently.

For getting favourable results, a specific clustering algorithm along with certain parameters may be best suited in a market domain[3].

There are various clustering techniques out of which we have used K-means clustering for segmenting customers on an e-commerce website.

1.3.1 K-means Clustering:

K-Means is one of the most widely used clustering Algorithms , and is simple and efficient. The aim of the K-Means algorithm is to divide M points in N dimensions into K clusters (assume k centroids) fixed a priori. The K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible[4]. This process is a repeatable and iterative task where vast amounts of raw data are scanned for similarities and patterns[5]. The unorganized data is searched for knowledge that is important and then data points are assigned.

K-means clustering is used as it is the simple and efficient way of clustering the large e-commerce dataset.

K-means is a fast method because it does not have many computations. The space complexity of k-means is O((m+k)n) and its time complexity is O(I*k*m*n) where m is the number of points, n is the number of attributes and I is the number of iterations required for convergence[6].

This study aims to explore the avenues of using customer segmentation, as a business intelligence tool as well as the use of clustering techniques for helping organizations redeem a clearer picture of the valuable customer base.

CHAPTER 2

SYSTEM DESIGN

The data chosen for customer segmentation is the transcation history of customers. Magento also performed an analysis of purchase history to get the best customer, unprofitable customers, potential customer profit[10]. This transaction data is preprocessed. The RFM score is calculated based on the date of purchase, no of transactions and the total amount spent. The products are grouped into different categories .Then the customers are segmented based on above derived attributes. An analysis over the segments is made to determine the customer's characteristics.

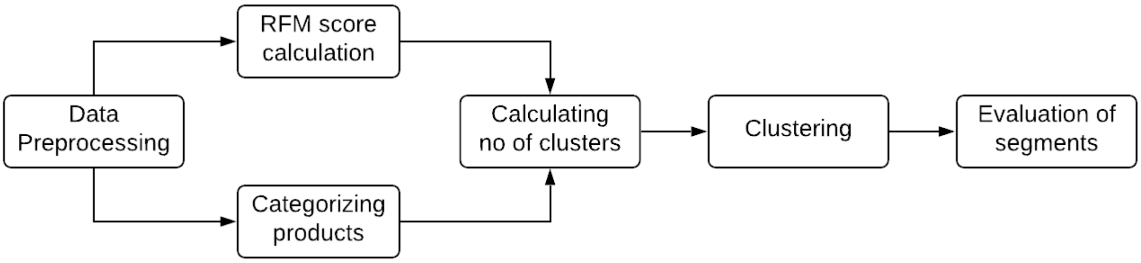
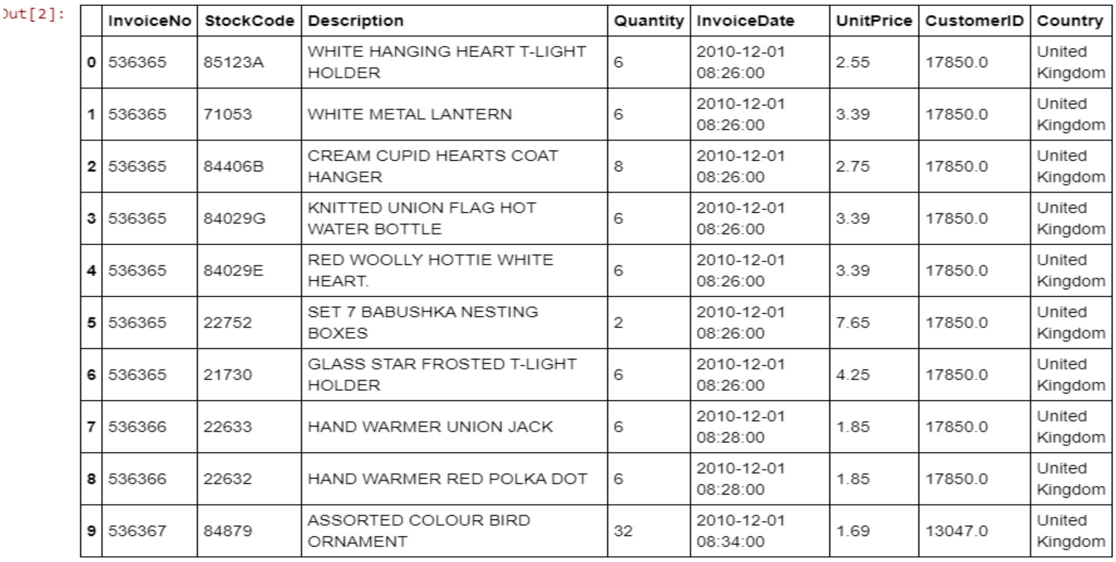


Figure 2.1 Methodology

2.1 DESCRIPTION OF DATASET

The data for our analysis is a transnational data which contains all the transactions occurring between 1/12/2010 and 9/12/2011 for a UK-based online retail company. Invoice number, Stock code, Description, Quantity, Invoice Date, Unit price, Customer ID and Country are the attributes present in the dataset as shown in Figure 2.1. The dataset initially contains 5,41,909 records out of which 4,01,604are the valid records obtained after data pre-processing.



Out[2]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	7.65	17850.0	United Kingdom
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	4.25	17850.0	United Kingdom
7	536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	1.85	17850.0	United Kingdom
8	536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00	1.85	17850.0	United Kingdom
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:34:00	1.69	13047.0	United Kingdom

Figure 2.2 Initial Dataset

2.2 DATA PREPROCESSING

From the initial dataset, the records containing missing entries(1,35,080) are removed and then the records containing duplicate entries(5,225) are removed.

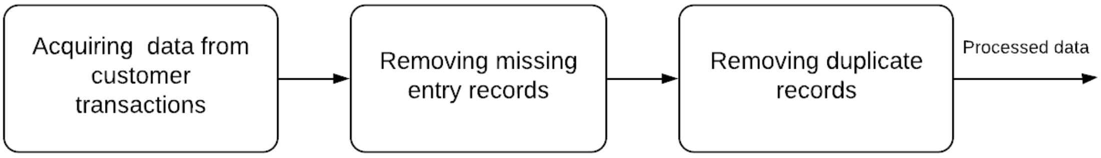

Acquiring data from customer transactions
Removing missing entry records
Removing duplicate records
Processed data

Figure 2.3 Data Preprocessing

2.3 OVERVIEW OF SYSTEM DESIGN

Segmentation of customers is done based on two features:

1)RFM analysis 2) Categories of products purchased

2.3.1 Overview of K-means:

It is the simplest algorithm of clustering based on partitioning principle. The algorithm is sensitive to the initialization of the centroids position, the number of K (centroids) is calculated by elbow method, after calculation of K centroids by the terms of Euclidean distance data points are assigned to the closest centroid forming the cluster, after the cluster formation the barycentre's are once again calculated by the means of the cluster and this process is repeated until there is no change in centroid position[7].

2.3.2 Determining number of clusters:

Elbow method is used for finding optimal value of K for K-means clustering algorithm. This method works by finding the SSE of each data point with its nearest centroid with different values of K. As value of K increases the SSE will decrease and at a particular value of K where there is most decline in the SSE is the elbow, the point at which we should stop dividing data further[8].

Algorithm:

Elbow method is applied to calculate value of K for the dataset.

Step-1: Run the algorithm for various values of k i.e. making the k vary from 1 to 10.

Step-2: Calculate the within cluster squared error.

Step-3: Plot the calculated error, where a bent elbow like structure will form, will give the optimal value of clusters.

2.3.3 Calculation of RFM Score :

RFM stands for **Recency, Frequency, and Monetary** value, each corresponding to some key customer trait. These RFM metrics are important indicators of a customer's behaviour because frequency and monetary value affects a customer's lifetime value, and recency affects retention, a measure of engagement[12]. RFM factors illustrate the facts such as the more recent the purchase(recency), the more responsive the customer is to promotions the more frequently the customer buys(frequency), the more engaged and satisfied they are monetary value differentiates heavy spenders from low-value purchasers(monetary)[9].

2.3.4 Categorizing products:

Text mining technique is used to categorise the various products in our dataset. It is the process of deriving high quality information from the text. The most frequently occurring 50 words in the product description are retrieved. The products are then grouped into different clusters as shown in Figure 2.2. Each cluster is a product category.

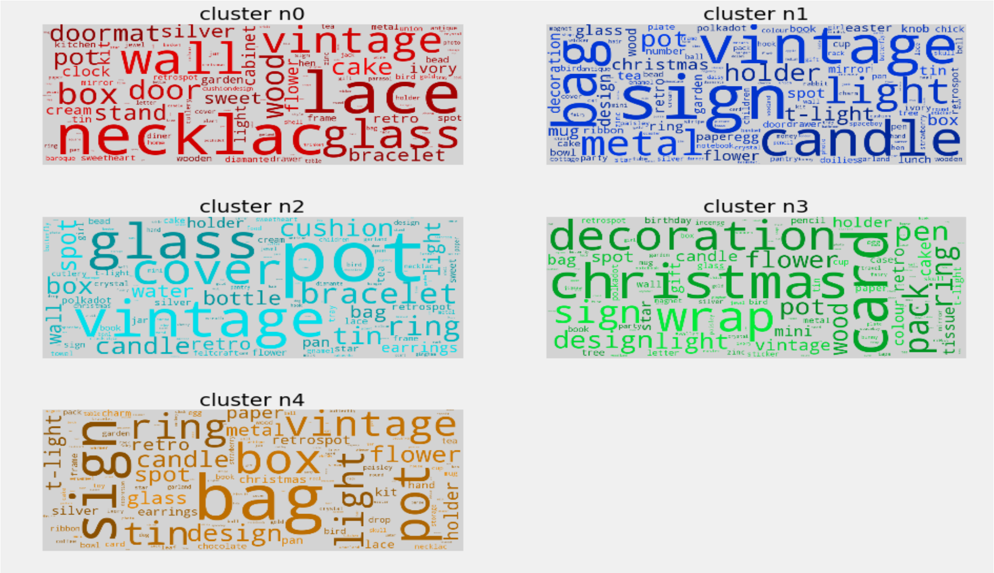


Figure 2.4 Word Cloud for Product Categories

CHAPTER 3

IMPLEMENTATION

3.1 SEGMENTATION

3.1.1 RFM SCORE

The RFM score is calculated for every customer. Then the Elbow method is applied for finding the optimal number of clusters. From the graph in Figure 3.1, it is inferred that optimal no of clusters is 5.

```
# Using the elbow method to find the optimal number of clusters
from sklearn.cluster import KMeans
wcss = []
for i in range(2,15):
 kmeans = KMeans(n_clusters = i, init = 'k-means++', max_iter = 300,
 n_init = 10, random_state = 0)
 kmeans.fit(scaled_matrix)
 wcss.append(kmeans.inertia_)
```

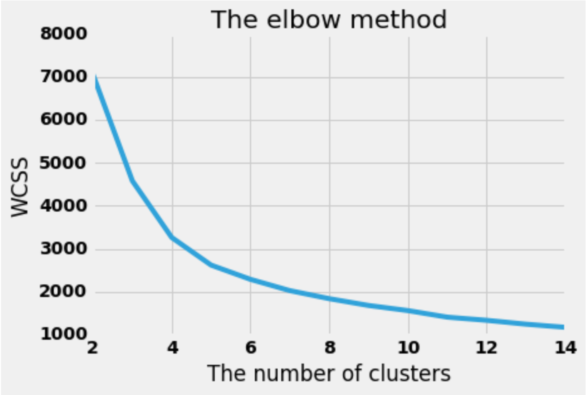
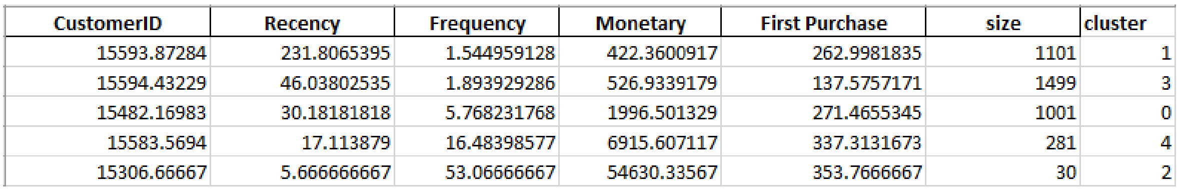


Figure 3.1 Elbow method graph for RFM data

K-means Clustering is performed for the obtained no. of clusters i.e. 5 from Figure 3.1. Then Recency, Frequency, Monetary, Size and First Purchase for each cluster is obtained as shown in Figure 3.2.

```
n_clusters = 5
kmeans = KMeans(init='k-means++', n_clusters = n_clusters, n_init=100)
kmeans.fit(scaled_matrix)
clusters_clients = kmeans.predict(scaled_matrix)
```



CustomerID	Recency	Frequency	Monetary	First Purchase	size	cluster
15593.87284	231.8065395	1.544959128	422.3600917	262.9981835	1101	1
15594.43229	46.03802535	1.893929286	526.9339179	137.5757171	1499	3
15482.16983	30.18181818	5.768231768	1996.501329	271.4655345	1001	0
15583.5694	17.113879	16.48398577	6915.607117	337.3131673	281	4
15306.66667	5.666666667	53.06666667	54630.33567	353.7666667	30	2

Figure 3.2 Segments obtained from RFM data

3.1.2 PRODUCT CATEGORIES

In this section, the customers are segmented based on the category of products they had purchased. These categories are obtained using text mining and clustering techniques. Then the Elbow method is applied for finding the optimal number of clusters. From the graph in Figure 3.3, it is inferred that optimal no of clusters is 6.

```
# Using the elbow method to find the optimal number of clusters
from sklearn.cluster import KMeans
wcss = []
for i in range(2,15):
   kmeans = KMeans(n_clusters = i, init = 'k-means++', max_iter = 300,
   n_init = 10, random_state = 0)
kmeans.fit(scaled_matrix)
wcss.append(kmeans.inertia_)
```

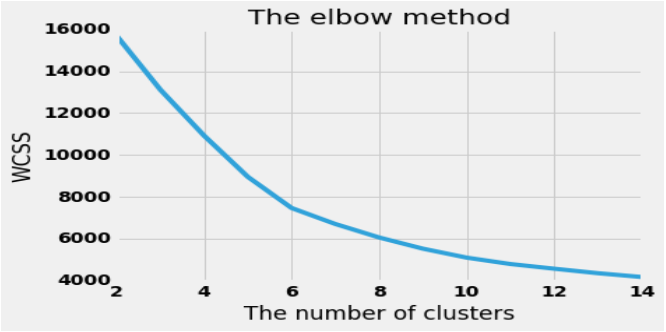


Figure 3.3 Elbow method graph for Product Categories

K-means clustering is performed using the obtained number of clusters i.e. 6 from the Figure 3.3. Amount spent on each category and segment to which each customer belongs is shown in Figure 3.4.

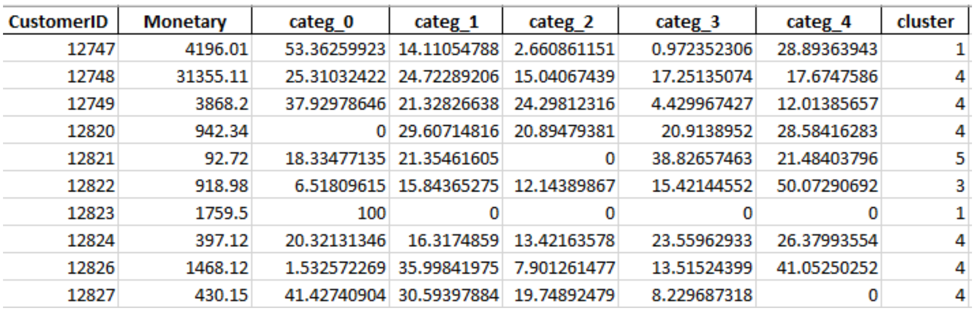
```
n_clusters = 6

kmeans = KMeans(init='k-means++', n_clusters = n_clusters, n_init=100)

kmeans.fit(scaled_matrix)

clusters_clients = kmeans.predict(scaled_matrix)

product_data.loc[:, 'cluster'] = clusters_clients
```



CustomerID	Monetary	categ_0	categ_1	categ_2	categ_3	categ_4	cluster
12747	4196.01	53.36259923	14.11054788	2.660861151	0.972352306	28.89363943	1
12748	31355.11	25.31032422	24.72289206	15.04067439	17.25135074	17.6747586	4
12749	3868.2	37.92978646	21.32826638	24.29812316	4.429967427	12.01385657	4
12820	942.34	0	29.60714816	20.89479381	20.9138952	28.58416283	4
12821	92.72	18.33477135	21.35461605	0	38.82657463	21.48403796	5
12822	918.98	6.51809615	15.84365275	12.14389867	15.42144552	50.07290692	3
12823	1759.5	100	0	0	0	0	1
12824	397.12	20.32131346	16.3174859	13.42163578	23.55962933	26.37993554	4
12826	1468.12	1.532572269	35.99841975	7.901261477	13.51524399	41.05250252	4
12827	430.15	41.42740904	30.59397884	19.74892479	8.229687318	0	4

Figure 3.4 Segments obtained from Product Categories data

CHAPTER 4

RESULTS AND OBSERVATIONS

4.1 DESCRIPTION OF RADAR CHART

Radar Charts are a way of comparing multiple quantitative variables. This makes them useful for seeing which variables have similar values or if there are any outliers amongst each variable. Radar Charts are also useful for seeing which variables are scoring high or low within a dataset, making them ideal for displaying performance. It gives a clear visualization of the outputs when large datasets are used.

4.2 ANALYSIS OF SEGMENTS

4.2.1 RFM SCORE

The recency, frequency and monetary score for each customer segment is pictorially represented using radar chart in the Figure 4.1.

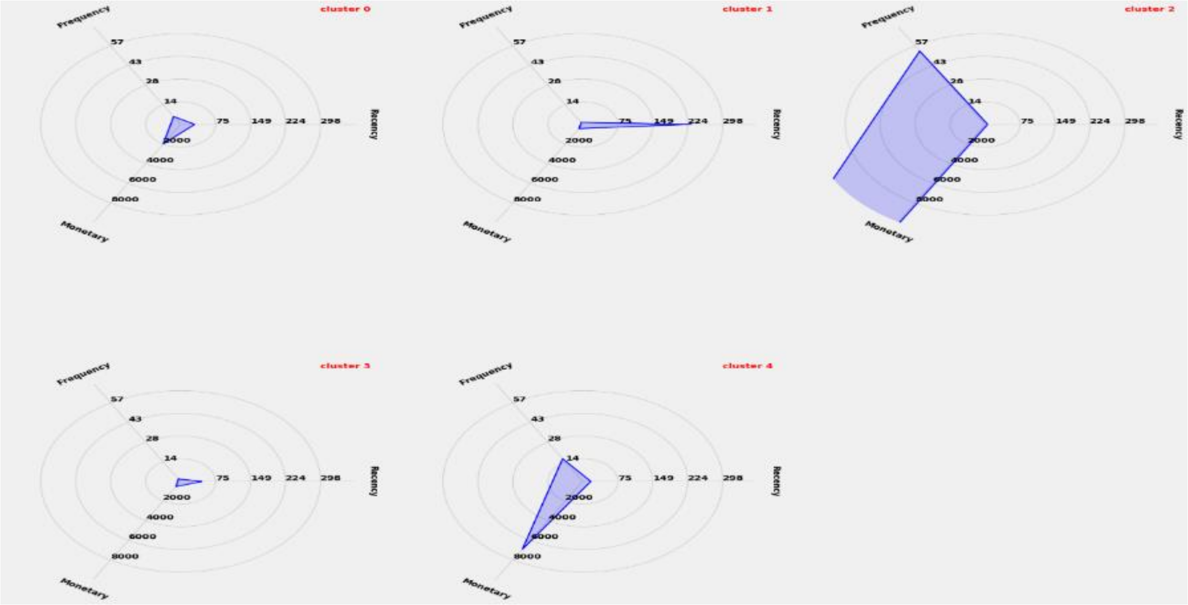


Figure 4.1 Customer segments obtained using RFM data

- **Cluster 0** - These users can be termed as active users who has equal recency, frequency and monetary scores.
- **Cluster 1** - These users can be termed as new users whose recency score is very high.
- **Cluster 2** - These users are inactive in the recent times but they have been regular customers in the past. These customers needs attention.
- **Cluster 3** - These customers are inactive right from the beginning and needs lesser consideration.
- **Cluster 4** - These customers are valuable users. They have a good monetary score in small number of purchases.

4.2.2 PRODUCT CATEGORIES

The spending score of customer segments in each product category is pictorially represented in radar chart which is shown in Figure 4.2.

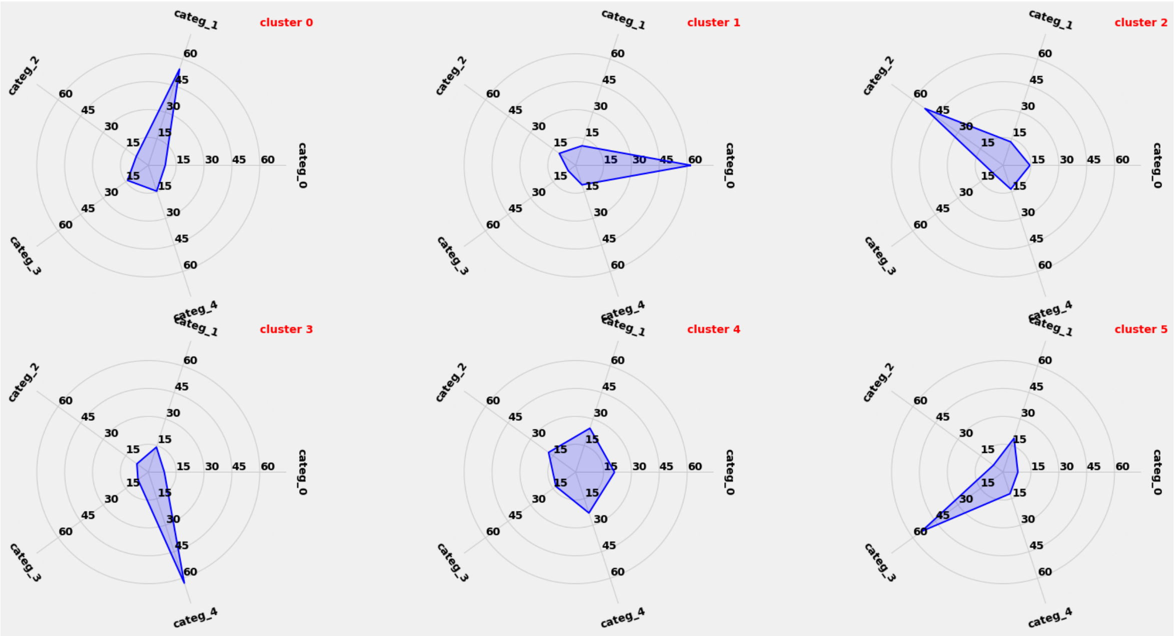


Figure 4.2 Customer Segments obtained using purchased products

- **Cluster 0 -**These customers are interested more on vintage products and home utilities and least bothered about the rest.
- **Cluster 1 -** These customers are interested more on jewellery and accessories.
- **Cluster 2-** These customers often purchase collectibles like glass items, pottery, wall stickers, lights etc.
- **Cluster 3-** They purchase day-to-day essential products and thus saves money.
- **Cluster 4 -** These customers are not interested on specific products. They purchase in a random fashion.
- **Cluster 5 -** These customers are interested on gift articles like birthday cards, Christmas lights, decoration items, wrappers, etc. They make more purchases on party items.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 CONCLUSION

Customer segmentation is a way to improve communication with the customer, to know the wishes of the customer, customer activity so that appropriate communication can be built .This project aims to segment customers to establish a healthy relationship with them . The initial data is the transaction history of the customers. This data is pre-processed to remove redundancy and inconsistency. The customers are segmented based on the RFM(Recency Frequency Monetary) score and the category of products purchased using clustering technique. Analysis of customer characteristics in each segment is carried out. Customer segments are represented using radar chart for better analysis.

5.2 FUTURE WORK

Customer segmentation has a good scope in future. The project can be extended by segmenting customers based on their geography and demography. It will enhance the specificity of the model, thus increasing the scope of use. Customer segmentation can be used in tele-communication sectors, online streaming services, Banking and other platforms where managing customer relationship is essential.

REFERENCES

[1] A. Ansari and A. Riasi, “Taxonomy of marketing strategies using bank customers clustering”, International Journal of Business and Management, vol. 11, no. 7, pp. 106-119, 2016.

[2] Tripathi, S., A. Bhardwaj, and E. Poovammal. "Approaches to clustering in customer segmentation." International Journal of Engineering & Technology 7.3.12 (2018): 802-807.

[3] .D. MacKay, “An example inference task: Clustering”, in Information theory, inference, and learning algorithms. Cambridge, UK: Cambridge University Press, 2003, pp. 284-292.

[4] J. MacQueen, “Some methods for classification and analysis of multivariate observations”, in Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, 1967, pp. 281-297.

[5] T. Kanungo, et al., “An efficient k-means clustering algorithm: analysis and implementation”,IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-892, 2002.

[6] .http://wwwusers.cs.umn.edu/ ~kumar/dmbook/ch8.pdf

[7] Tanupriya Choudhury, Vivek Kumar, Darshika Nigam, Intelligent Classification & Clustering Of Lung & Oral Cancer through Decision Tree & Genetic Algorithm, International Journal of Advanced Research in Computer Science and Software Engineering,2015

[8] Kansal, Tushar, et al. "Customer Segmentation using K-means Clustering." *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*. IEEE, 2018.

[9] Sari, J. N., Nugroho, L. E., Ferdiana, R., & Santosa, P. I. (2016). Review on customer segmentation technique on ecommerce. Advanced Science Letters, 22(10), 3018-3022.

[10] Magento. An Introduction to Customer Segmentation. 2014. info2.magento.com/.../.

[11] Firdaus, Sabhia, and Md Ashraf Uddin. "A survey on clustering algorithms and complexity analysis." International Journal of Computer Science Issues (IJCSI) 12.2 (2015): 62.

[12] Birant, Derya. "Data mining using RFM analysis." Knowledge-oriented applications in data mining. IntechOpen, 2011.