FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 6 No. 1, March, 2022, pp 387 - 393
DOI: https://doi.org/10.33003/fjs-2022-0601-910

ANALYSIS AND VISUALIZATION OF MARKET SEGEMENTATION IN BANKING SECTOR USING KMEANS MACHINE LEARNING ALGORITHM

Durojaye D. I. and *Obunadike G.N.

Department of Computer Science and IT, Federal University Dutsin-Ma, Katsina State

*Corresponding authors' email: **gobunadike@fudutsinma.edu.ng** Phone: +2348095232233

ABSTRACT

Segmentation is a way of assigning each dataset to a segment called cluster. It is widely applied in different area of human endeavor such as banking sector, health sector, retail, media etc. Many organizations are faced with problems of ineffective customer care services and intelligent management decisions because of inability to effectively analyze customer data that will give insight to the nature of customers to help in effective customer services and intelligent management decision. Kmeans algorithm is the widely used algorithm for market segmentation, normally the k value of Kmeans algorithm are randomly picked. Picking the optimal k value is usually a challenge in application of Kmeans algorithm and this usually affects the performance of Kmeans algorithm. This work applies elbow method to obtain the optimal k value that was applied to analyze dataset from banking sector (in this case United Bank of Africa) for better insight, business management and marketing strategy. The customer cluster created was evaluated using visual plots and cluster centers. The optimal k value of six (6) was obtained using the elbow function. The dataset was thus segmented based on the optimal k value of 6 obtained. The clustering results obtained showed high intra cluster similarity (data within a cluster are similar) and low inter cluster similarity (data from different clusters are dissimilar). The result also showed that customers in cluster 3 and 4 has similar marketing needs and can be served together.

Keywords: Clustering, Data visualization, K-means Algorithm, market segmentation

INTRODUCTION

Market segmentation can be defined as dividing a market into distinct groups of customers, with different needs, characteristics or behavior, who might require separate products or who may respond differently to various combinations of marketing efforts (Armstrong and Kotler, 2005). Some bases of segmentation that may be used include geographic, demographic, psychographic and behavioral. Other variables that may be used for segmentation include situational (e.g., purchase/use occasions), and customer preferences for products or specific product attribute levels.

Many organizations are faced with problems of ineffective customer care services and intelligent management decisions because of inability to effectively analyze customer data that will give insight to the nature of customers to help in effective customer services and intelligent management decision. Segmentation is critical because a company has limited resources, and must focus on how to best identify and serve its customers.

Clustering is one of the commonly used market segmentation techniques (Wedel and Kamakura, 2000). Kmeans algorithm is widely used for clustering, normally the k value of Kmeans algorithm are randomly picked. Picking the optimal k value is usually a challenge in application of Kmeans algorithm and this usually affects the performance of Kmeans algorithm.

Clustering techniques minimize intra cluster distance and maximize inter cluster distance for data segmentation. Clustering methods can be broadly classified into four categories: Partitioning, Hierarchical, Density-based and Grid-based methods. Clustering can be applied in information retrieval, web pages grouping, image and market segmentation etc. (Pham and Afify, 2006). For segment identification, customer clustering is applied (Saarenvirta, 1998) using customer characteristics (demographics, socioeconomic factors, and geographic location), product-related behavioral characteristics (purchase behavior, consumption behavior, preference for attractions, experiences, and services). K-means clustering are widely used for market segmentation (Hung and Tsai, 2008). Clustering can also be applied for evaluating supermarket shopping paths or deriving employers' branding strategies apart from customer segmentation (Moroko and Uncle, 2009).

Marketing Strategies

Market segmentation using clustering technique such as Kmeans algorithm will help to group customers with similar needs together. This will enhance the marketing strategies of the organization. The three areas of marketing which are usually taken into consideration when marketing a product are:

The first area is mass marketing. It covers the area of mass production, mass distribution and mass promotion of products to all buyers (Gunter and Furnham, 1992). However, marketers have realized the great varieties in each individual customer and therefore the market segmentation is a helpful tool for the marketers to customize their marketing programmes for each individual customer (Dibb and Simkin, 1999).

The second area is product differentiated marketing. The marketer produces two or more products that display different features, styles, quality, sizes etc.

The third, and dominating, area is target marketing where the marketer distinguishes among a variety of market segments, chooses one or more of the segments and then develops products and marketing mixes customized to each segment (Gunter and Furnham, 1992).

In business it is a matter of being able to communicate your message in a persuasive way. Companies therefore need to be able to adapt to their target audience's needs, wants and values (Kotler and Keller, 2009). In order for companies to do so, they may ask themselves questions like; who the customers are? What do they buy? And where can they be found? It is not possible for the companies to reach out to all customers in large, broad, or diverse markets and therefore by dividing the customers into groups or segment(s), the company can choose which group they wish to target (Kotler and Keller, 2009).

Segmentation Process

The three stages of the segmentation process are; segmentation, targeting, and positioning. ***Segmentation variables:*** The first stage of the segmentation process involves the selection of suitable variables for grouping customers. These are also referred to as base variables or the segmentation basis. There is rarely one best way of segmenting a market and more than one variable can be used. Segmentation analysis: Research also plays a very vital role in segmentation as segmentation analysis requires a range of data from a wide variety of sources on markets, customers‟ attitudes, motives and behaviour as well as competitor information.

Targeting: Targeting is the next step in the sequential process and involves a business making choices about segment(s) on which resources are to be focused. There are three major targeting strategies: undifferentiated, concentrated, and differentiated. During this process the business must balance its resources and capabilities against the attractiveness of different segments.

Positioning: This follows on logically from the segmentation and targeting stages. Customer perceptions are central to the product position especially in relation to the competition's offering. The product or service has to satisfy key customer requirements and this has to be clearly communicated to customers.

From the above, one can draw the conclusion that, the sequential logic used in most marketing management literature are the following: identify your target segment; describe the characteristics of the segment members; determine their needs as to the product that you are selling, adapt the marketing mix components according to the segment's needs, sell the products, get increased product profitability and thus increased profitability of the firm. Criticism from one school of thought is that it is increasingly difficult to build marketing activities on the notion of a market. The "markets" are fragmenting rapidly and we are moving towards a time when the only relevant segment is the individual customer (Wedel and Kamakura, 2000).

Clustering

Clustering is a popular tool for customer segmentation which aims to identify groups of similar observations. Most clustering algorithms are designed to work on a fixed dataset which does not change over time. When it does, the entire model needs to be recomputed, an extension to this are stream clustering algorithms which aim to find and maintain clusters over time in an endless stream of new observations. These algorithms process data points in real time and avoid expensive computations by incrementally updating the model. To do so, clusters need to contain enough information to update them with new information. For example, merely storing the centres of clusters allows merging clusters. However, it does not allow splitting clusters again if necessary. To overcome this, stream clustering algorithms usually rely on an online and offline phase (Rajagopal, 2011). The online component evaluates the stream in real time and captures relevant summary statistics. As an example, it is possible to split the data space into grid-cells and simply count the number of observations per cell. This results in a large number of micro-clusters that summarize the data stream. When desired, an offline component 'reclusters' the micro-clusters to derive a final set of macro-clusters. This step often uses a variant of traditional clustering algorithm.

K-Means Algorithm

K-Means clustering is an efficient machine learning algorithm to solve data clustering problems. It's an unsupervised algorithm that is quite suitable for solving market segmentation problems. Unsupervised machine learning is quite different from supervised machine learning. It's a special kind of machine learning algorithm that discovers patterns in the dataset from unlabeled data. Unsupervised machine learning algorithms can group data points based on similar attributes in the dataset. One of the main types of unsupervised models is clustering models. A clustering machine learning algorithm is an unsupervised machine learning algorithm. It is used for discovering natural groupings or patterns in the dataset. It's worth noting that clustering algorithms just interpret the input data and find natural clusters in it (Rajagopal, 2011).

Unlike supervised learning algorithms, K-means clustering is an unsupervised machine learning algorithm. This algorithm is used when a dataset has an unlabelled data. Unlabelled data means input data without categories or groups provided. Our market segmentation data is like this for this problem. The algorithm discovered groups (cluster) in the data, where the number of clusters is represented by the K value. The algorithm acts iteratively to assign each input data to one of K clusters, as per the features provided. All of this makes k-means quite suitable for the market segmentation problem.

While using the k-means clustering algorithm, the first step is to indicate the number of clusters (k) that can be used to produce the final output. The algorithm starts by selecting k objects from dataset randomly that will serve as the initial centers for our clusters. These selected objects are the cluster means, also known as centroids. Then, the remaining objects have to be assigned to the closest centroid. This centroid is defined by the Euclidean Distance present between the object and the cluster mean. This step refers as "cluster assignment". When the assignment is completed, the algorithm proceeds to calculate new mean value of each cluster present in the data. After the recalculation of the centers, the observations are checked if they are closer to a different cluster. Using the updated cluster mean, the objects undergo reassignment. This goes on repeatedly through several iterations until the cluster assignments stop altering. The clusters that are present in the current iteration are the same as the ones obtained in the previous iteration (Witten and Frank, 2000).

Let $DS = \{x_1,\ x_2,\ \dots x_n\}$ be a set of n dataset, where $A_1, A_2, \dots . A_1$ is the set of M attributes. Kmeans applies minimization objective functions P with unknown variables U and Z as shown in Equation 1.

$$P(U,Z) = \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} U_{i,j}\, d(x_{i,j},\ Z_{i,j}) \qquad (1)$$

Subject to

$$\sum_{l=1}^{k} U_{l,j} = 1 \qquad 1 \le i \le n \qquad (2)$$

$$U_{i,j} \in \{0,1\} \quad 1 \le i \le n, \quad 1 \le l \le k \qquad (3)$$

Where

U is a $n \times k$ matrix $U_{l,j}$ is binary variable and $U_{l,j} = 1$ indicates that objects x_i is allocated to cluster c_i.

$Z_l = [z_1, z_2,\ \dots z_k\ ]$ is a set representing $k - centroids.$

$d(x_{ij},\ z_{ij})$ is the dissimilarity measures between data objects x_i and the centroid c_i on attributes A_j. The distance between two identical values is 0 while the distance between two distinct values is 1.

$$d(x_{ij},\ z_{ij}) = f(x) = \begin{cases} 0, (x_{ij} = z_{ij}) \\ 1, (x_{ij} \neq z_{ij}) \end{cases} \qquad (4)$$

The algorithm of Kmeans is as shown in Figure 1.

Kmeans Algorithm

1. Choose the k value
2. Choose a random K point centroids
3. Assign each data point to closest centroid of the k points
4. Compute and replace the new centroids of each data points
5. Reassign each data point to new cluster
6. Stop when the centroids stop changing or number of iteration reached

Figure 1: Kmeans Algorithm

METHODOLOGY

The dataset used for this research was collected from UBA, Katsina. It has 200 records with five attributes including the class attribute. A sample of the data set and the attributes is as shown in Table 1. Before the application of the clustering algorithm, an exploratory analysis was done to get a better insight into the dataset before segmentation. This research was implemented using software called R Studio. It is a free complete machine learning software that contains all necessary tools for data analysis ranging from data pre-processing to analysis. Figure 3 showed how the dataset was read into the R studio environment. The application of the kmeans algorithm was executed using R studio programming language. The work flow methodology is as shown in Figure 2.

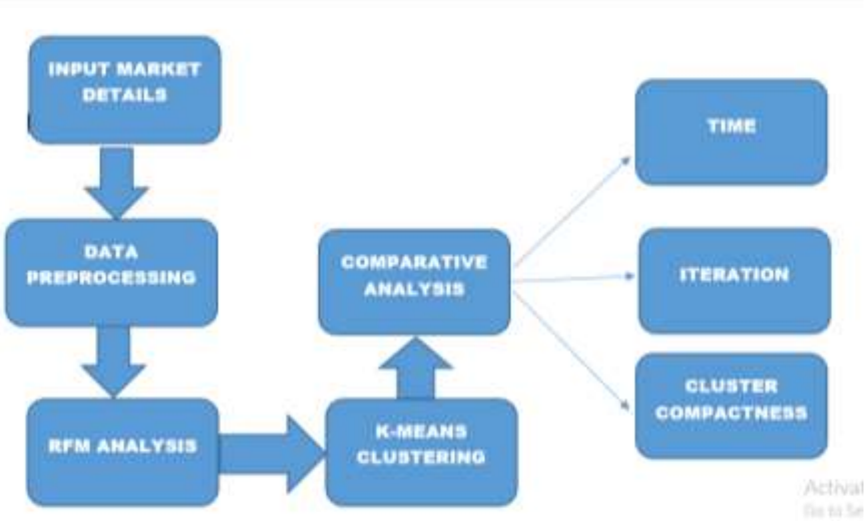
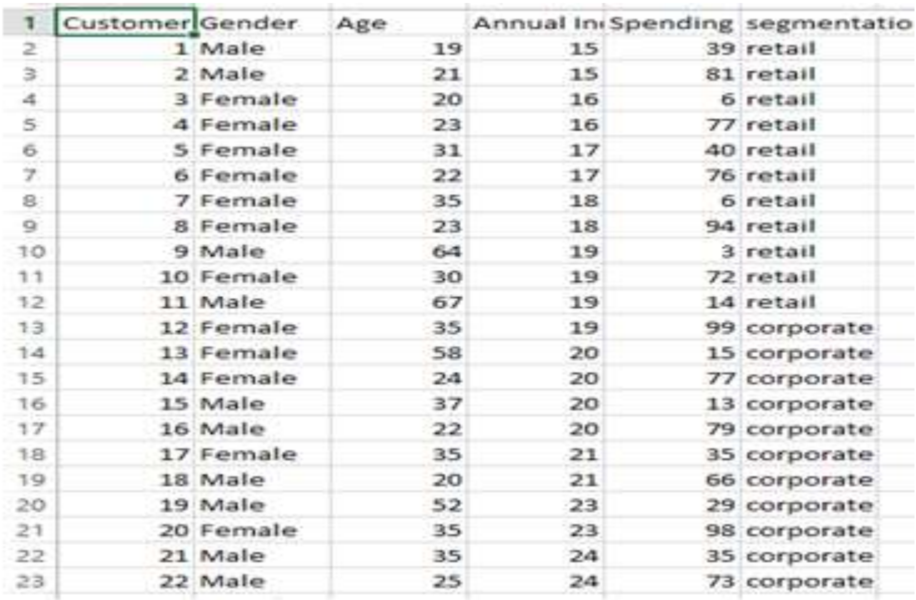


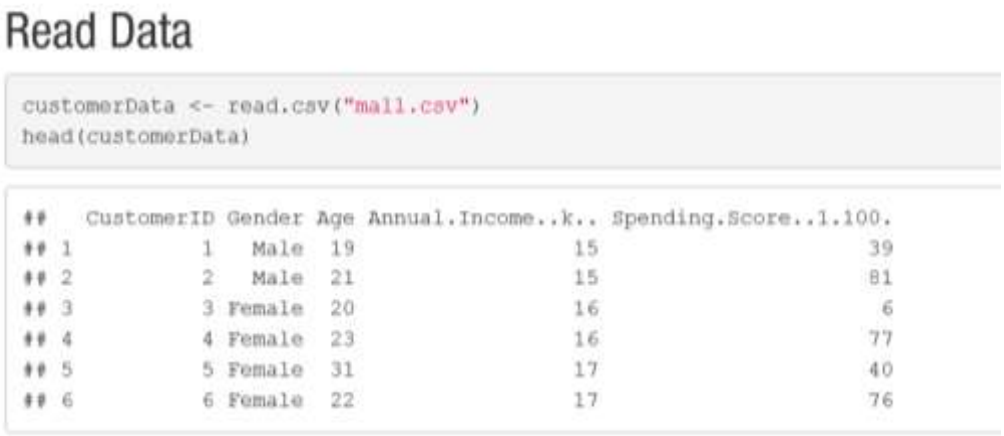
Figure 2: Methodology Diagram

Table 1 shows the structure of the training data set supplied to K-Means algorithm for building the cluster model.

Table 1: Dataset Structure



1	Customer	Gender	Age	Annual In	Spending	segmentatio
2	1	Male	19	15	39	retail
3	2	Male	21	15	81	retail
4	3	Female	20	16	6	retail
5	4	Female	23	16	77	retail
6	5	Female	31	17	40	retail
7	6	Female	22	17	76	retail
8	7	Female	35	18	6	retail
9	8	Female	23	18	94	retail
10	9	Male	64	19	3	retail
11	10	Female	30	19	72	retail
12	11	Male	67	19	14	retail
13	12	Female	35	19	99	corporate
14	13	Female	58	20	15	corporate
15	14	Female	24	20	77	corporate
16	15	Male	37	20	13	corporate
17	16	Male	22	20	79	corporate
18	17	Female	35	21	35	corporate
19	18	Male	20	21	66	corporate
20	19	Male	52	23	29	corporate
21	20	Female	35	23	98	corporate
22	21	Male	35	24	35	corporate
23	22	Male	25	24	73	corporate



Read Data

```
customerData <- read.csv("mall.csv")
head(customerData)
```

```
##   CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100.
## 1          1   Male  19                 15                     39
## 2          2   Male  21                 15                     81
## 3          3 Female  20                 16                      6
## 4          4 Female  23                 16                     77
## 5          5 Female  31                 17                     40
## 6          6 Female  22                 17                     76
```

Figure 3**:** Read Dataset

Exploratory Analysis

Exploratory analysis was carried out to give better insight into the nature of the data before clustering. There are three identifiable customer segments namely; retail banking, corporate banking, Savings and Credit Co-operative Society (SACCO Banking)

The relative densities of the various attributes in the data set are as shown in figure 4. The visualization of relationships between important attributes is as shown in Figure 5, Figure 6 and Figure 7.

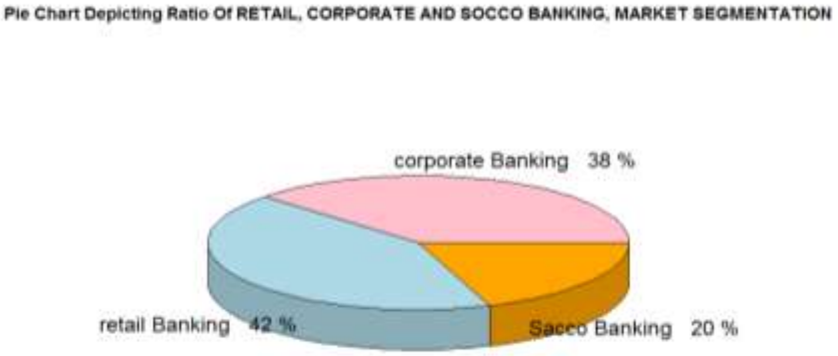


Figure 4**:** Basic Market Segmentation

The exploratory analysis revealed that there are three identifiable customer segments namely; retail banking, corporate banking, Savings and Credit Co-operative Society (SACCO Banking)

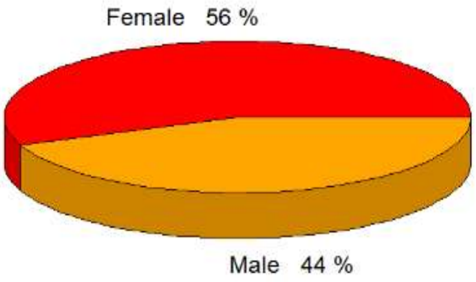


Figure 5: Gender Analysis of the customer

The exploratory analysis revealed that the 56% of the loan seekers are female while about 44% are male.

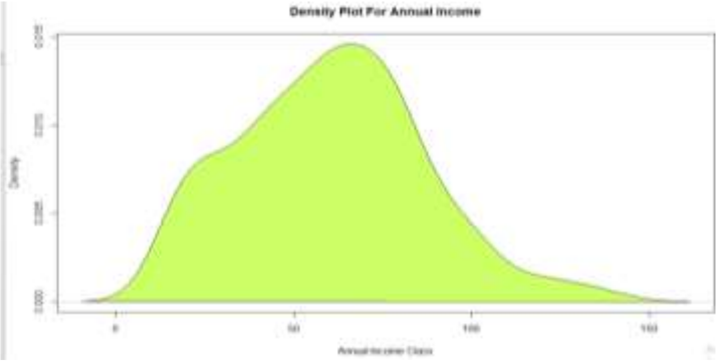


Figure 6: Density of annual income of the customer

The visualization of the relationship between the annual income reveals that majority of the higher income are within the amount of 60-80 thousands.

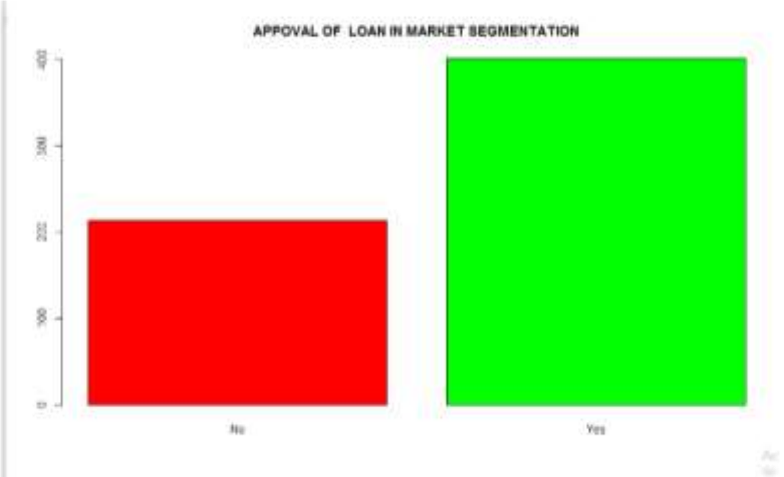


Figure 7: Loan approval analysis of the customer

The exploratory analysis revealed that majority of the loans applied for were approved only small number was not approved.

Elbow Method for Optimal K Value The basic 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 applied to determine this optimal k value. The elbow method runs kmeans clustering algorithm on dataset for a range of values for k and computes an average scores for all clusters for each k value. The elbow method involves varying the number of cluster k from 1 to 10, for each value of k WCSS (Within Cluster Sum of Square) is calculated. The WCSS is the sum of squared distance between each point and the centroid in a cluster. When WCSS is plotted against the k values the plot looks like an elbow as shown in Figure 8. As number of clusters increases, the WCSS value will start to decrease. WCSS is largest when k = 1, when looked at the graph in Figure 8, it is observed that the graph rapidly changed at a point and thus creating an elbow shape. From this point the graph starts to move parallel to the x-axis. The k value corresponding to this point is the optimal k value or optimal number of clusters (Tajunisha, 2010; Rousseeuw, 1984)

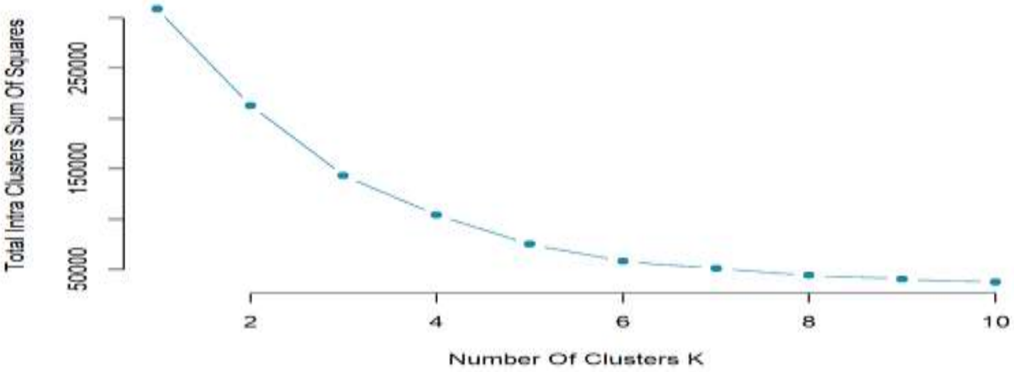


Figure 8: Optimal Number of K Values using Elbow method

Data Clustering and Evaluation

The application of Kmeans algorithm was carried out in R studio platform using optimal k value of k = 6 which was obtained using the elbow function. The effectiveness of the clustering was evaluated using cluster plots and cluster centers as evaluation metrics. The main goal of clustering is to attain high intra cluster similarity (data within a cluster are similar) and low inter cluster similarity (data from different clusters are dissimilar). This is an internal criterion for measuring the quality of clustering. Once clustering has been performed, to measure the effectiveness of the clusters produced can be achieved through some metrics. Ideal clustering is characterized by minimal intra cluster and maximal inter cluster distance. In this work the two metrics employed to measure the effectiveness of clustering are cluster plots and cluster centers. The cluster plots and cluster centers are used to measure the separation distance between clusters. It displays a measure of how close each point in a cluster is to points in the neighbouring clusters. Table 2 shows the characteristics of customer segments created using the Rstudio and Figure 9 is the pictorial representation of the segments created.

Table 2: Characteristics of the Customer Segments Created

Cluster Number	Age	Annual Income (Thousands)	Spending Score..1.100	Cluster Sizes	Within cluster sum of squares by cluster:
1	56.15556	53.37778	49.08889	45	8062.133
2	44.14286	25.14286	19.52381	21	7732.381
3	41.68571	88.22857	17.28571	35	16690.857
4	32.69231	86.53846	82.12821	39	13972.359
5	27.00000	56.65789	49.13158	38	7742.895
6	25.27273	25.72727	79.36364	22	4099.818

Table 2 shows the characteristics of customer segments created , it shows for instance that the customers in segment one are those within the age of 56 with annual income of 53 thousand, the cluster size is 45 and within the cluster sum of square is 8062133 which shows the clusters re well separated

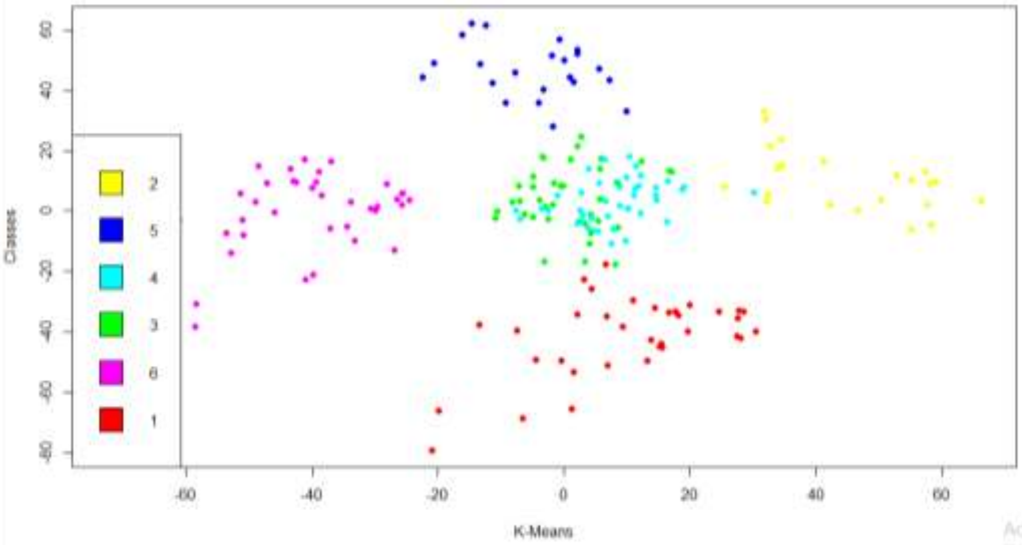


Figure 9: Pictorial Representation of customer segmentation

Figure 9: is the pictorial representation of the customer segments created using the K-means algorithm. The diagram shows that customers in segment three and four has almost the same characteristics and can be handled together.

CONCLUSION

The results of various analyses on the dataset showed that proper segmentation of customers is fundamental for the development of a successful business. K-Means clustering is an efficient machine learning algorithm to solve data clustering problems but required k value to be picked at random, picking optimal k value is usually a problem and this usually affects the performance of Kmeans algorithm. This work applied elbow technique to obtain optimal k value that was used on the dataset for optimal result. The application of Kmeans algorithm was carried out in R studio platform using optimal k value of k = 6 which was obtained using the elbow function. The effectiveness of the clustering was evaluated using cluster plots and cluster centers as evaluation metrics. The main goal of clustering is to attain high intra cluster similarity (data within a cluster are similar) and low inter cluster similarity (data from different clusters are dissimilar). This is an internal criterion for measuring the quality of clustering. Once clustering has been performed, to measure the effectiveness of the clusters produced can be achieved through some metrics. Ideal clustering is characterized by minimal intra cluster and maximal inter cluster distance. In this work the two other metrics employed to measure the effectiveness of clustering are cluster plots and cluster centers. The cluster plots and cluster centers are used to measure the separation distance between clusters. It displays a measure of how close each point in a cluster is to points in the neighbouring clusters. The clustering results obtained showed high intra cluster similarity (data within a cluster are similar) and low inter cluster similarity (data from different clusters are dissimilar).

REFERENCES

Armstrong, G. and Kotler, P. (2005). Marketing: An Introduction, Upper Saddle River, N.J. 7th Ed. Prentice Hall.

Dibbs, S. and Simkin, L. (2009) Bridging the segmentation theory/practice divide. Journal of marketing Management, 25(3), pp219 – 225

Guter, B. nd Funrnham, A. (1992) consumer Profiles: An Introduction to Psychographics, Routledge, London.

Haley R I(1985) Developing effective communications strategy–a benefit segmentation approach. Wiley, New York

Hung, C. and Tsai, C. (2008) "Market Segmentation based on Hierarchical Self Organizing Map for Markets of Multimedia on Demand", Expert Systems with Applications, vol. 34, pp.780–787, http://dx.doi.org/10.1016/j.eswa.2006.10.012

Kotler, P. nd Keller, L. (2009) “Marketing Management”, 13th Edition, Person prentice Hall, Upper Saddle River.

Moroko, L and Uncles, M. D. (2009) "Employer Branding and Market Segmentation", Journal of Brand Management, vol. 17, no. 3, pp. 181–196, http://dx.doi.org/10.1057/bm.2009.10

Pham, D. T. and Afify, A. A. (2006)"Clustering Techniques and their Applications in Engineering", Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, pp. 103–119. http://dx.doi.org/10.1016/B978-008045157-2/50060-2

Rajagopal, S. (2011) "Customer Data Clustering using Data Mining Technique", International Journal of Database Management Systems (IJDMS), vol. 3, no. 4, pp. 1–11, http://dx.doi.org/10.5121/ijdms.2011.3401 https://arxiv.org/ftp/arxiv/papers/1112/1112.2663.pdf

Rousseeuw, P. J. (1984) "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis", Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65.

Saarenvirta, G. (1998) "Mining Customer Data", DB2 Magazine, vol. 3, no. 3, pp. 10–20.

Tajunisha, S. (2010) "Performance Analysis of K-means with Different Initialization Methods for High Dimensional Data", International Journal of Artificial Intelligence and Applications, vol. 1, no. 4, pp. 44–52, http://dx.doi.org/10.5121/ijaia.2010.1404

Wedel, M. and Kamakura, W. A. (2000) "Market Segmentation Conceptual and Methodological Foundations", Kluwer Academic Publishers.

Witten, I. H. and Frank, E. (2000). Data mining: Practical machine learning tools and techniques. Morgan Kaufmann series in data management systems

©2022 This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International license viewed via https://creativecommons.org/licenses/by/4.0/ which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is cited appropriately.