DOI: 10.17323/2587-814X.2022.1.72.82

# Customer segmentation using *k*-means clustering for developing sustainable marketing strategies

**Nidhi Gautam** [a]
E-mail: nidhi.uiams@pu.ac.in

**Nitin Kumar** [b]
E-mail: nk2268913@gmail.com

[a] University Institute of Applied Management Sciences, Panjab University
Address: South Campus, Behind P.U. Alumni House Udyog Path, Punjab University Rd, Sector 25, Chandigarh 160014, India

[b] Housing Development Finance Corporation Limited
Address: Dalhousie Road, near Shani Dev Mandir, Pathankot, Punjab 145001, India

**Abstract**

Sales and marketing is the indispensable department of an organization which leads to the generation of revenue and building customer relationship. Marketing is the process of finding the potential customers and sales is the process of converting those potential customers into real customers. Hence, it is imperative that marketing and sales go hand in hand. Developing marketing strategies needs proper market research which can cover the relevant pointers like demographics, culture, spending power, income and many more. The process of segmentation, targeting and positioning (STP) is carried out to develop marketing and sales strategies. STP is done by collection of the marketing intelligence. For this process, surveys are also used but data mining has far more effective and better results so far. Organizations tend to take risk because of the importance and relevance of the marketing and sales department. Most of the budget in the organizations is allocated for marketing and promotional activities. For making data-driven and accurate decisions, data mining is used in various fields to extract valuable information and patterns. This paper discusses the use of the data mining concept on marketing. This paper aims to analyze marketing data with *k*-means data mining clustering techniques and to find the relationship between marketing and *k*-means data mining clustering techniques.

**Keywords:** marketing, *k*-means, clustering, data mining tools, data mining techniques

**Citation:** Gautam N., Kumar N. (2022) Customer segmentation using *k*-means clustering for developing sustainable marketing strategies. *Business Informatics*, vol. 16, no. 1, pp. 72–82. DOI: 10.17323/2587-814X.2022.1.72.82

## Introduction

Marketing is a value creation process according to Philip Kotler [1]. It is all about what you deliver to your customer. People consider marketing and sales as the same but both are different and have their own relevance and impact in the organization. Marketing is the process of finding the potential customer and sales is the process of converting those potential customers into real customers [2]. The concept of marketing can be explained using different marketing concepts. First, there is a production concept, which explains that supply will create the demand. Second, the sales concept emphasis on selling the product by hook or by crook. Third, the product concept gives importance to product innovation and differentiation. Fourth, the marketing concept, which believes in catering to the needs of the customer as customer, is king for them. Lastly, the societal marketing concept focuses on the customer as well as society and the environment. It totally depends upon the organization to decide which concept it wants to follow. Whenever anyone considers marketing strategies, marketing mix is used. According to Philip Kotler, marketing mix comprises four components i.e., product, price, place and promotion [1].

Data mining is a process of extracting knowledge from large amounts of data. It is a technique to find trends, patterns, correlations, anomalies in databases which can be helpful to make accurate future decisions. It is also known as knowledge discovery in databases (KDD) results. Data mining helps experts to understand the data and leads to better and data driven decisions. Data mining is an intersection of three fields: databases, artificial intelligence and machine learning. The steps of data mining includes data cleansing (for the removal of noisy and inconsistent data), data integration (to combine data from multiple sources for efficient data processing), data selection (to select and retrieve the relevant data for analysis), data transformation (to transform the collected data into a desirable form for further data processing), data mining (a technique applied on data for various pattern matching methods), pattern evolution (to determine important patterns which represent knowledge and data insights) and knowledge presentation (this is done by various visualization methods for knowledge representation pictorially or graphically). The various applications of data mining include market basket analysis (association mining). Market basket analysis is the method to discover relations or correlations among the set of data items. Classification analyzes a training set of objects with known labels and tries to form a model for each class based on the features in the data. Regression is to predict values of some missing data or to build a model for some attributes using other attributes of the data. Time Series Analysis analyzes time series data to find certain regularities and interestingness in data. Clustering is used to identify clusters embedded in the data. The task of clustering is to find clusters for which intra-cluster similarity is high and inter-cluster similarity is low. Outlier analysis is used to find outliers in the data, namely detect data which are very far away from average behavior of the data [3].

This paper focuses on the importance of data mining approaches, specifically clustering, for formalizing the marketing strategies by understanding customer needs and spending behavior. The paper highlights the importance of visualization tools to understand the important relationships between various parameters in the dataset. It shows that how a pair-plot can be helpful in identifying clusters and the silhouette score for deciding the k value for *k*-means clustering method.

The rest of the paper is organized as follows: Section 1 presents the motivation and research rationale of the study. Section 2 describes numerous existing models and related work using various clustering methods. Section 3 presents the research design and methodology followed for this work. Section 4 describes the results and discussions. The last Section concludes the paper with future recommendations.

## 1. Motivation and research rationale

This study aims to work on the transaction dataset of a store which is taken from an internet source [4] for making clusters of customers depending upon their income and spending score. This will lead to segmentation, targeting and positioning (STP) of the customers and their behavior patterns will help us to develop sustainable marketing strategies. Nowadays, everything is connected to the customer with the help of the Internet of things. You just need space to store data to perform analysis to keep track of the customers. This study can help organizations to gain new customers and maintain loyal customers. The main contributions of this paper are summarized as follows:

- to perform market segmentation of the customers depending upon their spending using unsupervised machine learning;
- to perform STP;
- to know the target customer and develop marketing strategies.

## 2. Existing models with literature review

Association rule mining is a data-mining concept which is used to optimize the patterns associated with dynamic behaviors of transactions made by customers when purchasing some specific products. The insights generated from this technique can be used by the retailing business for making data-driven decision-making. Using this algorithm, the frequent transactions made by the customers have been analyzed using the support and confidence of the customers in buying associated items. The analysis conducted by Association rule mining model can best be used in managing product placement on the shelves in the supermarket [5–7].The study was conducted in order to make a market basket analysis by using association rules. The data used in the study was the sales data of a supermarket from the Vancouver Island University website. Data was analyzed in the Weka tool where the dataset contained 225 different products for analysis [8].

The study focused on small and medium enterprises (SMEs), where customer behavior was analyzed from the perspective of SMEs. The model proposed the integration of customer relationship management (CRM) and the data mining techniques to provide effective rules and new patterns for better decision-making. The model suggested that as the era of big data is unwinding itself, the data mining application may enhance accuracy of rules and patterns, all of which can further help the enterprises to improve customers' satisfaction and loyalty, reduce customer churn and so on. The suggested models can also help SMEs to classify the priority customer groups and offer them better facilities and ranges to retain them. The suggested models can help the enterprises to further improve their market share, position in the market and maintain a positive development process [9–11].

The model incorporated a new graphical display for clustering techniques. In this model, each cluster is represented by a silhouette. The silhouette is based on the comparison of its tightness and separation. This silhouette showed which objects lie well within their cluster and which ones are merely somewhere in between clusters. The whole clustering is displayed by combining the silhouettes into a single plot. The average silhouette width provides an evaluation of clustering validity and may help to select an 'appropriate' number of clusters [12–13].

Classification of aquifer vulnerability using *k*-means cluster analysis uses the application of the cluster analysis in ground vulnerability assessment using the *k*-means technique. In this study, a clustering technique is used because it removes some of the subjectivity associated with the indexing method. It creates a vulnerability map that does not rely on fixed weights and ratings and provides a more objective representation of the system's physical characteristics. The model was applied to an aquifer in Iran and compared with the standard DRASTIC approach using the water quality parameters nitrate, chloride and total dissolved solids (TDS) as surrogate indicators of aquifer vulnerability. The model having clustering techniques outperformed the other methods [14–16].

The paper discovered the segments of organic food consumers in Lebanon by using a market segmentation based on lifestyle and attitude variables to generate appropriate marketing strategies for each market segment [17]. Market basket analysis (MBA) is a very powerful data mining technique which provides various types of information, like buying behavior of the customer, likes, dislikes, etc. to the retailer, all of which can help the retailer perform correct decision-making. It can be used in various fields, such as marketing, management, bioinformatics, the education field and many more. MBA is a very useful technique to find out interesting patterns from a large amount of data which can automatically track any type of changes in facts from previous data [18–20].

The authors used a *k*-means clustering algorithm to identify the customer segmentation in a supervised manner. The methodology could understand the complex relationships existing in the data attributes [21]. The authors analyzed the data of an e-commerce portal to understand the requirements of the customers so as to provide them better services in the future. A *k*-means clustering algorithm was used to analyze the data, considering customer segmentation as an important aspect of it [22–23]. The authors analyzed data of three online food chains and applied various clustering algorithms on the same. Though there is no fixed model of a particular algorithm which could show best results, *k*-means clustering has shown promising results on the data [24]. The authors suggested how customer segmentation can be implemented and can be useful to understand the customers. Customer segmentation can be a stepping-stone for identifying future prospective customers and specific marketing strategies can be formulated considering the customer segment [25]. The authors have stressed the use of machine learning algorithms for customer segmentation, since it can enhance productivity and profitability of an organization. *K*-means clustering was used for the study and it has shown very promising results for customer segmentation [26].

## 3. Research design and methodology (data collection method)

**A. Sampling method.** Secondary data, mall customer segmentation dataset from an internet source [4].

**B. Sample size.** The dataset which is used for the analysis contains people's purchasing attributes in the Malls. This dataset has five features – customerID, age, gender, credit score and income. There are data for about 200 transactions used for data analysis.

**C. Research rationale.** To devise a comprehensive model that can be used to classify customers based on spending score and annual income for developing appropriate marketing strategies.

**D. Tools used.** R Studio [27], Weka [28], MS Excel [29].

R Studio and Weka are freeware which are used for data analytics. These tools are mostly used for data analytics in the field of industry as well as academia. MS Excel is a sub-tool of Microsoft Office, which is generally used for data preparation and preprocessing.

**E. Clustering algorithms used.** The *k*-means clustering algorithm [24] is based on the Euclidian distance to figure out k clusters in the data. The clusters are homogeneous within them and represent similar types of data. *K*-means clustering is most suitable to handle big and hyper spherical data. It is best suited for market segmentation, social network analytics, image segmentation and so on.

## 4. Results and discussion

The dataset is preprocessed and cleansed in Excel [29] by using descriptive statistics. The dataset is balanced as the distribution of males and females are almost the same. The distribution of data is checked in Weka [28] and pair plots are generated to show the same.

From the pair plot in *Fig. 1*, we found that the last row is insightful since it gives an indication of hidden clusters in the data. There is cluster formation between the Spending score (1–100) vs. CustomerID, the Spending score vs. Age and the Spending score (1–100) vs. Annual income (in thousands of US dollars).

The joint-plots, distributions, correlation matrix, silhouette coefficient and *k*-means clusters are generated in R-Studio [27] by using its libraries. The joint-plot of spending score and age as shown in *Fig. 2a* shows that there are two bright core areas where density is very high.


*Fig. 1.* Pair–plot of variables representing a pairwise relationship among CustomerID, Age, Annual Income (in thousands of US dollars) and Spending Score (1–100).

There are two different spending habits in different age groups represented in the plot.

The joint-plot of the spending score and annual income as shown in *Fig. 2b* shows that there is one area where density is very high i.e., in the middle. The other four areas show different patterns for the user. This may be possible because of the different purchasing power of the customers and their different spending habits. The five groups from the observations include Low income & High spending habits, Low income & Low spending habits, Moderate income & Moderate spending habits,

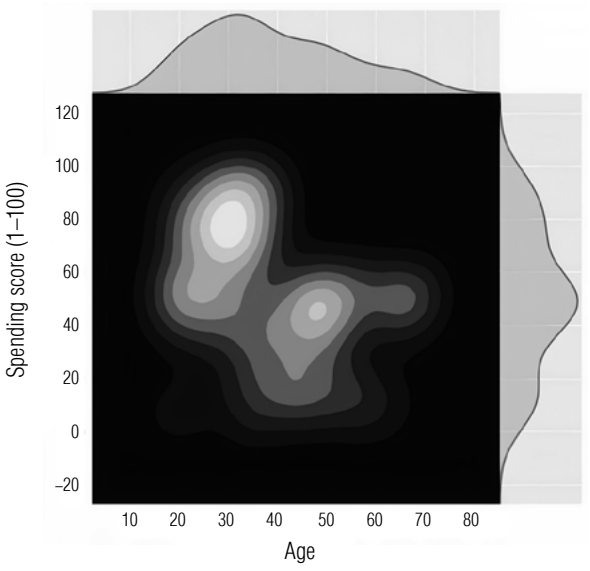


*Fig. 2a*. Joint plots for describing
(Spending score and Age) distributions on the same plot.

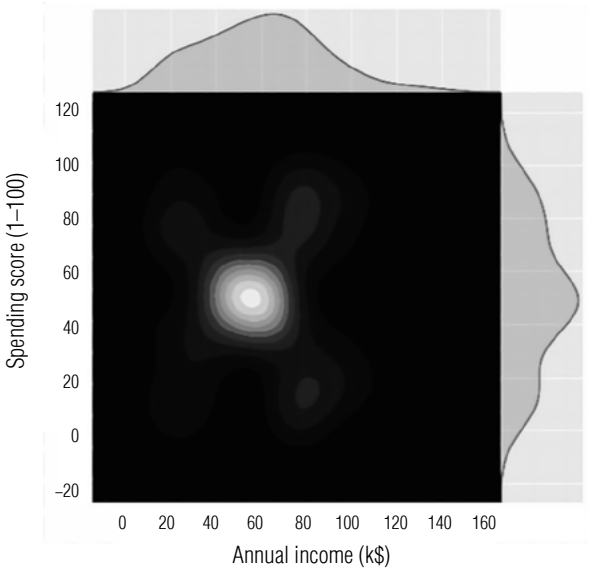


*Fig 2b*. Joint plots for describing
(Spending score and Annual income) distributions on the same plot.

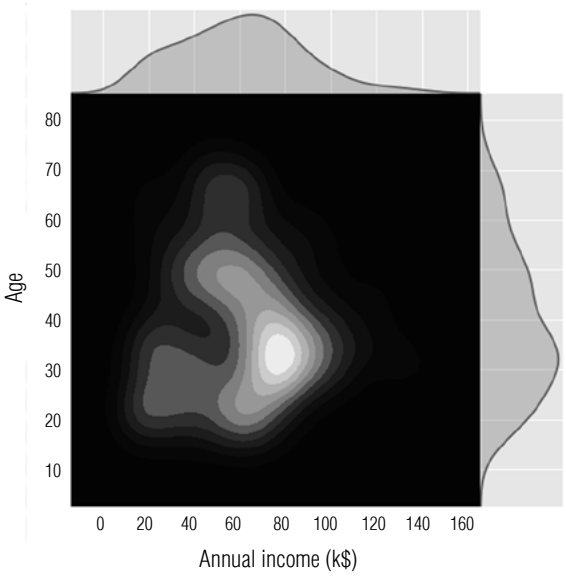


*Fig 2c*. Joint plots for describing
(Annual income and Age) distributions on the same plot.

High income & High spending habits and High income & Low spending habits. Joint-plot of Annual income and Age as shown in *Fig. 2c* and *Fig. 3* shows that the people in mid-30s have roughly a mean income of $80 000. If we compare the yearly income of more than $100 000, we find that males have more than $100 000 income in their early 30s and in the case of females it's around mid-40s. Perhaps, this is due to the disparity in pay.

The average value of the spending score of females is slightly more than that of the males. Notice the bulge of the graph in *Fig. 3*, which shows the mean value. *Figure 4* shows the Annual income of Males and Females. Thus, the average income of females (Female = 0) is less than that of males (Male = 1).

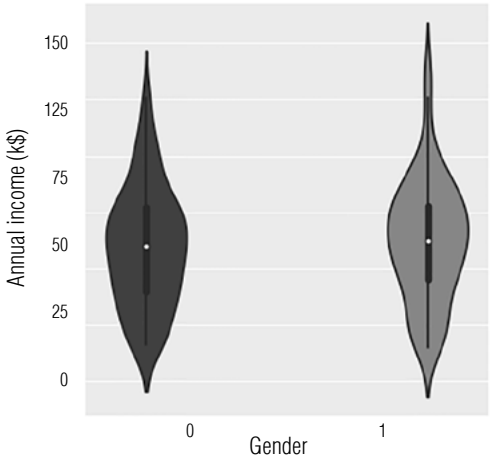


*Fig. 3*. Annual income (in thousands of USD) vs. Gender.

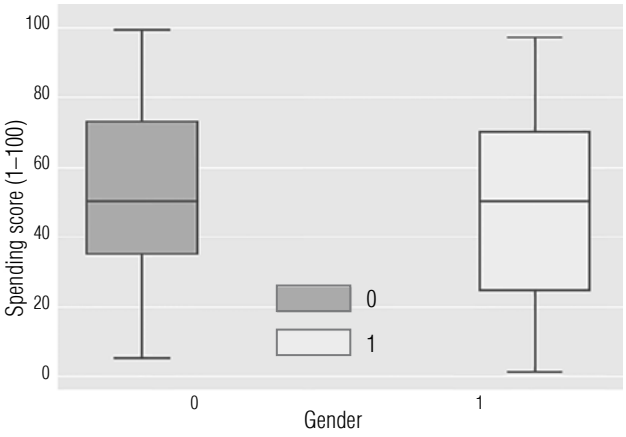


*Fig. 4.* Box–plot for Spending score (1–100) vs. Gender.

The distribution of the data showed that most of the people are less than 45 years of age. For most of the customers, the spending score centers between 40 and 60. Annual income and gender have a positive correlation. There is a positive correlation between annual income and the spending score as shown in *Fig. 5*.

**Customer Segmentation with *k*-means using the Silhouette Score.** The Silhouette coefficient in *k*-means clustering is calculated using the mean intra-cluster distance "*a*" and the mean nearest-cluster distance "*b*" for each sample. The Silhouette coefficient for a sample is $(b - a) / \max(a, b)$ where $b$ is the distance between a sample and the nearest cluster that the sample is not a part of. Note that the Silhouette coefficient is only defined if the number of labels is $2 \leq n\_labels \leq (n\_samples - 1)$. Trying to find clusters based on features like Annual income & Spending score. Since the Silhouette score is maximum for $k = 5$, it is a good idea to cluster the data into five subgroups.

The Silhouette coefficient has helped us to decide on the number of clusters to be taken for the study. In *k*-means clustering, selecting the value of $k$ is of utmost importance. By using the Silhouette coefficient, selecting the value $k$ has become very clear and simple. From the plot in *Fig. 6* there are five clusters: *cluster A, cluster B, cluster C, cluster D* and *cluster E*. Customers of the *cluster E* are misers as they have more purchasing power but they have a lower spending score. Customers of the *cluster A, cluster B* and *cluster C* are easy to handle. The customers of *cluster D* groups are a threat to the organization because they have less income but a larger spending score. They can be defaulters in the future. Therefore, by using *k*-means clustering and the Silhouette coefficient, customer classifications can be easily visualized. Hence, the marketing teams can easily identify potential customers.

## Conclusion

Nowadays, it is very crucial to identify your potential customers in order to have a more data driven strategy to target customers. From the above data analysis, it is concluded that the distribution of males and females

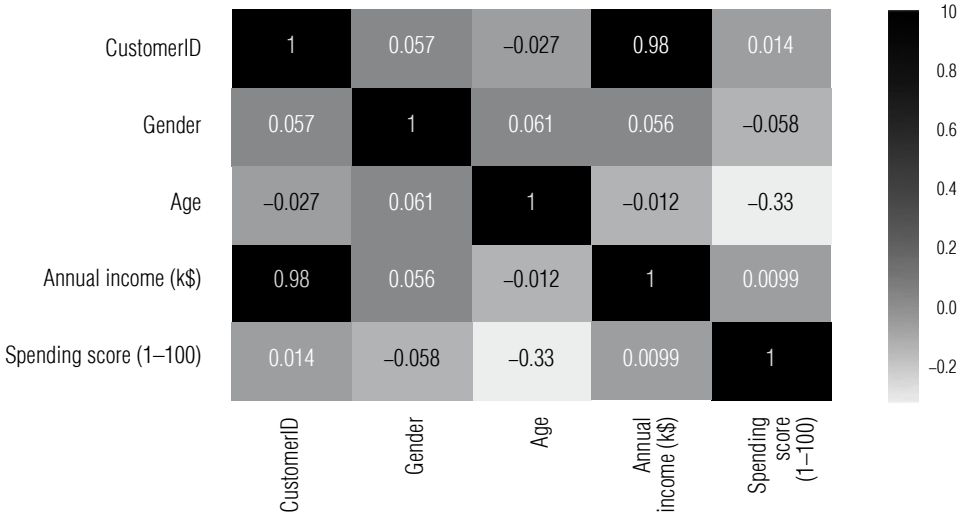


*Fig. 5.* Correlation matrix to show the correlation among variables such as customerID, Gender, Age, Annual income, Spending score.

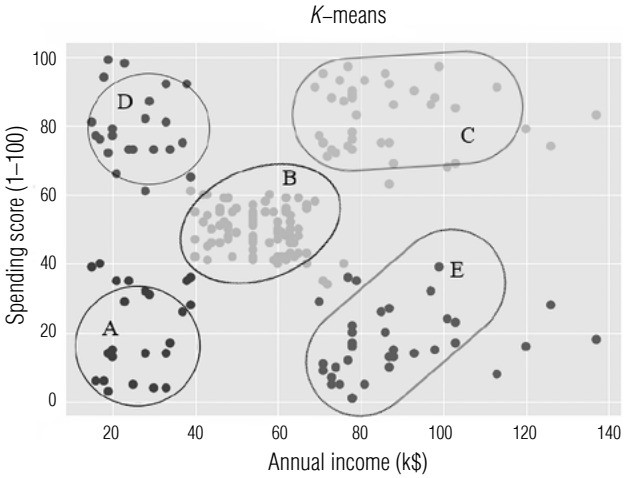


*Fig. 6. K–means clusters for customer classification.*

is nearly same. The pair-plot helps us to explain the permutation of the attributes, which helps in pattern, or cluster identification. With the help of the joint-plot between the spending score and annual income, purchasing capacity as well as spending habits of the customers can be analyzed. It is observed from the *k*-means clustering that customers can be classified into five groups such as "Low income & High spending habits," "Low income & Low spending habits," "Moderate income & Moderate spending habits," "High income & High spending habits" and "High income & Low spending habits." The box-plot for Spending score and Gender shows that the mean Spending scores of females are slightly more than that of the males whereas the violin plot shows that the average income of females is less than that of males. In this data, the majority of the people are less than 45 years of age and most of the customers, spending score centers between 40 and 60. Annual income and gender have positive correlation. There is a positive correlation annual income and spending score. The segmentation of the customers is done by a *k*-means algorithm. The value of the *k* is decided by the Silhouette score. We tried to find clusters based on features like Annual income & Spending score. The algorithm divided the data into five subgroups from which one could formulate marketing strategies in order to sell their products to the target audience.

In this study, the following problems have been resolved:

- to perform STP;
- to know the target customer and develop marketing strategies;
- to solve this problem, the dataset of a retail store was taken and we performed exploratory data analysis by using data visualization using various plots starting from the pair plot to the *k*-means plot.

The pair plot explains the relationships and patterns, which act as a stepping-stone for further analysis. There is cluster formation between Spending score (1-100) vs. CustomerID, Spending score vs. Age and Spending score (1-100) vs. Annual income (in thousands of USD) and these attributes were taken up for the study using joint plots. The joint-plot of Spending score and Annual income shows that there is one area where density is very high i.e., in the middle. The other four areas show different patterns for customers. This may be because of the different purchasing capacity of the customers and their different spending habits. The five groups from the observations includes Low income & High spending habits, Low income & Low spending habits, Moderate income & Moderate spending habits, High income & High spending habits and High income & Low spending habits.

The data visualization between Spending score and Annual income was presented. To know the accurate number of clusters, the Silhouette coefficient was used and its value came out to be five in this dataset, i.e. five clusters were considered for this study. Customers in *cluster E* have an annual income but they are spending very less from their income because their spending score is very low. This is a very important segment because they have the money to pay. A lot of marketing and promotional activities are required to influence this segment (group). A market basket analysis should be done to know the pattern associated with the purchase of goods. This can help us to know the better product placement of the product on the aisle. Marketing should be done in such a way that the consumers relate the product with their lifestyle; this will boost sales. A discount and offer can help the store to get more revenue. Customers who are in *cluster A, cluster B* and *cluster C* clusters have nearly the same annual income and spending score. These customers do not need more marketing and promotional activities. They can be tackled by the salesmen directly. Customers in *cluster D* are the ones who rely on the credit card because their annual income is very much less than their spending score. These customers have higher chances of becoming defaulters because their income is low but they are spendthrifts. The difference of outcomes in comparison with other competitive approaches and solutions is that our solutions are accurate because they have scientific and mathematical backup. The solution proposed in this paper is based on logic and data, not on

gut feeling and experiences. The solutions are safer and more reliable than traditional methods of STP. These solutions can be applicable to SMEs and small business persons with low investment and help them to use their hard-earned money in a justified way. In this paper, an algorithm divided the data into five subgroups which could be used to formulate marketing strategies in order to sell products to the target audience. IT systems can be developed for the SMEs and small business for the management of the sales and marketing of the business so that they can allocate their limited resources and budget for maximum benefits. Data scientists have a crucial role in this field because it is very important to understand the data of a particular organization. The exponential generation of the data and the growth of artificial intelligence has given an opportunity to data scientists and marketing tycoons so they can come together and build an IT system at affordable rates which will enhance the business process. ■

## References


1. Kotler P., Cunningham M.H., Keller K.L. (2008) *A framework for marketing management.* Toronto: Pearson Prentice Hall.
2. Rust R.T. (2020) Outside-in marketing: Why, when and how? *Industrial Marketing Management*, vol. 89, pp. 102–104. https://doi.org/10.1016/j.indmarman.2019.12.003
3. Gupta M.K., Chandra P. (2020) A comprehensive survey of data mining. *International Journal of Information Technology,* vol. 12, pp. 1243–1257. https://doi.org/10.1007/s41870-020-00427-7
4. Chaudhary V. (2018) *Mall customer segmentation data: Market basket analysis.* Data set. Available at: https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python (accessed 2 March 2021).
5. Guney S., Peker S., Turhan C. (2020) A combined approach for customer profiling in video on demand services using clustering and association rule mining. *IEEE Access*, vol. 8, pp. 84326–84335. https://doi.org/10.1109/ACCESS.2020.2992064
6. Wang S., Zhang H., Chen H., Shi Q., Li Y. (2020) Association rule mining for precision marketing of power companies with user features extraction. *2020 IEEE Power & Energy Society General Meeting (PESGM)*, pp. 1–5, https://doi.org/10.1109/PESGM41954.2020.9281644
7. Ananda I., Salamah U. (2020) The application of product marketing strategy using an association rule mining apriori method. *International Journal of Information System and Computer Science*, vol. 4, no. 2, pp. 80–88. Available at: https://ojs.stmikpringsewu.ac.id/index.php/ijiscs/article/download/899/pdf (accessed 21 February 2021).
8. Ünvan Y.A. (2020) Market basket analysis with association rules. *Communications in Statistics – Theory and Methods*, vol. 50, no. 7, pp. 1615–1628. https://doi.org/10.1080/03610926.2020.1716255
9. Ranjan J., Bhatnagar V. (2008) Critical success factors for implementing CRM using data mining. *Interscience Management Review*, vol. 1, no. 1, article 7. https://doi.org/10.47893/IMR.2008.1006
10. Ngai E.W.T., Xiu L., Chau D.C.K. (2009) Application of data mining techniques in customerrelationship management: A literature review and classification. *Expert Systems with Applications*, vol. 36, no. 2, pp. 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021
11. Hosseini S.M.H., Maleki A., Gholamian M.R. (2010) Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, vol. 37, no. 7, pp. 5259–5264. https://doi.org/10.1016/j.eswa.2009.12.070
12. Silva S., Cortez P., Mendes R., Pereira P.J., Matos L.M., Garcia L. (2018) A categorical clustering of publishers for mobile performance marketing. In: *The 13th International Conference on Soft Computing Models in Industrial and Environmental Applications*. Springer, Cham. https://doi.org/10.1007/978-3-319-94120-2_14
13. Beheshtian-Ardakani A., Fathian M., Gholamian M. (2018) A novel model for product bundling and direct marketing in e-commerce based on market segmentation. *Decision Science Letters*, vol. 7, no. 1, pp. 39–54. https://doi.org/10.5267/j.dsl.2017.4.005
14. Javadi S., Hashemy S.M., Mohammadi K., Howard K.W.F., Neshat A. (2017) Classification of aquifer vulnerability using *k*-means cluster analysis. *Journal of Hydrology*, vol. 549, pp. 27–37. https://doi.org/10.1016/j.jhydrol.2017.03.060
15. Rahmani B., Javadi S., Shahdany S.M.H. (2021) Evaluation of aquifer vulnerability using PCA technique and various clustering methods. *Geocarto International*, vol. 36, no. 18, pp. 2117–2140. https://doi.org/10.1080/10106049.2019.1690057
16. Jahwar A.F., Abdulazeez A.M. (2020) Meta-heuristic algorithms for k-means clustering: A review. *PalArch's Journal of Archaeology of Egypt / Egyptology*, vol. 17, no. 7, pp. 12002–12020. Available at: https://archives.palarch.nl/index.php/jae/article/view/4630 (accessed 21 February 2021).
17. Tleis M., Callieris R., Roma, R. (2017) Segmenting the organic food market in Lebanon: An application of k-means cluster analysis. *British Food Journal*, vol. 119, no. 7, pp. 1423–1441. https://doi.org/10.1108/BFJ-08-2016-0354
18. Kaur M., Kang S. (2016) Market basket analysis: Identify the changing trends of market data using association rule mining. *Procedia Computer Science*, vol. 85, pp. 78–85. https://doi.org/10.1016/j.procs.2016.05.180
19. Gupta S., Mamtora R. (2014) A survey on association rule mining in market basket analysis. *International Journal of Information and Computation Technology*, vol. 4, no. 4, pp. 409–414. Available at: http://ripublication.com/irph/ijict_spl/ijictv4n4spl_11.pdf (accessed 21 February 2021).
20. Aguinis H., Forcum L.E., Joo H. (2013) Using market basket analysis in management research. *Journal of Management*, vol. 39, no. 7, pp. 1799–1824. https://doi.org/10.1177/0149206312466147

21. Muhal H., Jain H. (2021) Two-stage customer segmentation using k-means clustering and artificial. *International Research Journal of Engineering and Technology*, vol. 8, no. 3, pp. 485–490.

22. Punhani R., Arora V.P.S., Sabitha S., Kumar Shukla V. (2021) Application of clustering algorithm for effective customer segmentation in E-Commerce. In: *Proceedings of the 2021 International Conference on Computational Intelligence and KnowledgeEconomy (ICCIKE), 17-18 March 2021, Dubai, United Arab Emirates*, pp. 149–154. https://doi.org/10.1109/ICCIKE51210.2021.9410713

23. Popović N., Savić A., Bjelobaba G., Veselinović R., Stefanović H., Ilić P.M. (2021) The implementation of hierarchical and nonhierarchical clustering for customer segmentation in one luxury goods company. In: *Proceedings of the 37th International Business Information Management Association (IBIMA), 30–31 May 2021, Cordoba, Spain*, pp. 8370–8381.

24. Aktaş A.A., Tunalı O., Bayrak A.T. (2021) Comparative unsupervised clustering approaches for customer segmentation. In: *Proceedings of the 2021 2nd International Conference on Computing and Data Science (CDS), 28-29 Jan. 2021, Stanford, CA, USA*, pp. 530–535. https://doi.org/10.1109/CDS52072.2021.00097

25. Suresh Y., Senthilkumar J., Mohanraj V., Kesavan S. (2021) Customer segmentation using machine learning in python. *Turkish Journal of Physiotherapy and Rehabilitation*, vol. 32, no. 3, pp. 4338–4342. Available at: https://turkjphysiotherrehabil.org/pub/pdf/321/32-1-521.pdf (accessed 21 February 2021).

26. Pradana M., Ha H. (2021) Maximizing strategy improvement in mall customer segmentation using *k*-means clustering. *Journal of Applied Data Sciences*, vol. 2, no. 1, pp. 19–25. https://doi.org/10.47738/jads.v2i1.18

27. RStudio Team (2020) *RStudio: Integrated Development for R*. RStudio, PBC, Boston, MA. Available at: http://www.rstudio.com/ (accessed 21 February 2021).

28. Ngo T. (2011) Data mining: practical machine learning tools and technique, Third Edition by Ian H. Witten, Eibe Frank, Mark A. Hell. *ACM SIGSOFT Software Engineering Notes*, vol. 36, no. 5, pp. 51–52. https://doi.org/10.1145/2020976.2021004

29. Microsoft Corporation (2018) *Microsoft Excel*. Available at: https://office.microsoft.com/excel (accessed 21 February 2021).

## About the authors

**Nidhi Gautam**

PhD (Computer Science & Engineering);

Assistant Professor, Fellow Senate Panjab University, University Institute of Applied Management Sciences, Panjab University, Chandigarh, India;

E-mail: nidhi.uiams@pu.ac.in

ORCID: 0000-0002-9454-3625

**Nitin Kumar**

MBA (IT & Telecommunications);

Management Trainee, Housing Development Finance Corporation Limited, Pathankot, Punjab, India;

E-mail: nk2268913@gmail.com

ORCID: 0000-0002-8887-2350

# Сегментация клиентов с использованием кластеризации на основе метода *k*-средних для разработки устойчивых маркетинговых стратегий

**Н. Гаутам** [a]

E-mail: nidhi.uiams@pu.ac.in

**Н. Кумар** [b]

E-mail: nk2268913@gmail.com

[a] University Institute of Applied Management Sciences, Panjab University
Адрес: South Campus, Behind P.U. Alumni House Udyog Path, Punjab University Rd, Sector 25, Chandigarh 160014, India

[b] Housing Development Finance Corporation Limited
Адрес: Dalhousie Road, near Shani Dev Mandir, Pathankot, Punjab 145001, India

**Аннотация**

Департаменты продаж и маркетинга являются незаменимыми подразделениями организации, обеспечивающими получение дохода и поддержку взаимоотношений с клиентами. Маркетинг представляет собой процесс поиска потенциальных клиентов, а продажи – это процесс превращения потенциальных клиентов в реальных. Поэтому взаимодействие подразделений маркетинга и продаж представляется крайне важным. Разработка маркетинговых стратегий требует соответствующего исследования рынка, которое может охватывать такие факторы как демография, культура, покупательная способность, доход и многое другое. Разработка маркетинговых стратегий и стратегий продаж связана с процессом сегментации, таргетинга и позиционирования (segmentation, targeting and positioning, STP). Процесс STP реализуется путем сбора маркетинговой информации. Для этого также используются опросы, но интеллектуальный анализ данных в настоящее время дает гораздо более эффективные и лучшие результаты. Организации склонны принимать определенные риски ввиду важности и актуальности подразделений маркетинга и продаж. Значительная часть бюджета в организациях выделяется на маркетинговые и рекламные мероприятия. Для принятия точных решений, основанных на данных, интеллектуальный анализ данных используется в различных областях для извлечения ценной информации и поиска закономерностей. В данной статье обсуждается использование концепции интеллектуального анализа данных в маркетинге. Цель статьи – проанализировать маркетинговые данные с применением метода *k*-средних для кластеризации, а также найти взаимосвязь между маркетингом и методами кластеризации на основе *k*-средних.

**Ключевые слова:** маркетинг, метод *k*-средних, кластеризация, средства интеллектуального анализа данных, методы интеллектуального анализа данных

**Цитирование:** Gautam N., Kumar N. Customer segmentation using *k*-means clustering for developing sustainable marketing strategies // Business Informatics. 2022. Vol. 16. No. 1. P. 72–82. DOI: 10.17323/2587-814X.2022.1.72.82

## Литература

1. Kotler P., Cunningham M.H., Keller K.L. A framework for marketing management. Toronto: Pearson Prentice Hall, 2008.
2. Rust R.T. Outside-in marketing: Why, when and how? // Industrial Marketing Management, 2020. Vol. 89. P. 102–104. https://doi.org/10.1016/j.indmarman.2019.12.003
3. Gupta M.K., Chandra P. A comprehensive survey of data mining // International Journal of Information Technology. 2020. Vol. 12. P. 1243–1257. https://doi.org/10.1007/s41870-020-00427-7
4. Chaudhary V. Mall customer segmentation data: Market basket analysis. Data set. Available at: https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python (accessed 2 March 2021).
5. Guney S., Peker S., Turhan C. A combined approach for customer profiling in video on demand services using clustering and association rule mining // IEEE Access. 2020. Vol. 8. P. 84326–84335. https://doi.org/10.1109/ACCESS.2020.2992064
6. Wang S., Zhang H., Chen H., Shi Q., Li Y. Association rule mining for precision marketing of power companies with user features extraction // 2020 IEEE Power & Energy Society General Meeting (PESGM). 2020. P. 1–5. https://doi.org/10.1109/PESGM41954.2020.9281644
7. Ananda I., Salamah U. The application of product marketing strategy using an association rule mining apriori method // International Journal of Information System and Computer Science. 2020. Vol. 4. No. 2. P. 80–88. Available at: https://ojs.stmikpringsewu.ac.id/index.php/ijiscs/article/download/899/pdf (accessed 21 February 2021).
8. Ünvan Y.A. Market basket analysis with association rules // Communications in Statistics – Theory and Methods. 2020. Vol. 50. No. 7. P. 1615–1628. https://doi.org/10.1080/03610926.2020.1716255
9. Ranjan J., Bhatnagar V. Critical success factors for implementing CRM using data mining // Interscience Management Review. 2008. Vol. 1. No. 1, article 7. https://doi.org/10.47893/IMR.2008.1006
10. Ngai E.W.T., Xiu L., Chau D.C.K. Application of data mining techniques in customerrelationship management: A literature review and classification // Expert Systems with Applications. 2009. Vol. 36. No. 2. P. 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

11. Hosseini S.M.H., Maleki A., Gholamian M.R. Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty // Expert Systems with Applications. 2010. Vol. 37. No. 7. P. 5259–5264. https://doi.org/10.1016/j.eswa.2009.12.070

12. Silva S., Cortez P., Mendes R., Pereira P.J., Matos L.M., Garcia L. A categorical clustering of publishers for mobile performance marketing // Proceedings of the 13th International Conference on Soft Computing Models in Industrial and Environmental Applications. Springer, Cham, 2018. https://doi.org/10.1007/978-3-319-94120-2_14

13. Beheshtian-Ardakani A., Fathian M., Gholamian M. A novel model for product bundling and direct marketing in e-commerce based on market segmentation // Decision Science Letters. 2018. Vol. 7. No. 1. P. 39–54. https://doi.org/10.5267/j.dsl.2017.4.005

14. Javadi S., Hashemy S.M., Mohammadi K., Howard K.W.F., Neshat A. Classification of aquifer vulnerability using *k*-means cluster analysis // Journal of Hydrology. 2017. Vol. 549. P. 27–37. https://doi.org/10.1016/j.jhydrol.2017.03.060

15. Rahmani B., Javadi S., Shahdany S.M.H. Evaluation of aquifer vulnerability using PCA technique and various clustering methods // Geocarto International. 2021. Vol. 36. No. 18. P. 2117–2140. https://doi.org/10.1080/10106049.2019.1690057

16. Jahwar A.F., Abdulazeez A.M. Meta-heuristic algorithms for *k*-means clustering: A review // PalArch's Journal of Archaeology of Egypt / Egyptology. 2020. Vol. 17. No. 7. P. 12002–12020. Available at: https://archives.palarch.nl/index.php/jae/article/view/4630 (accessed 21 February 2021).

17. Tleis M., Callieris R., Roma R. Segmenting the organic food market in Lebanon: An application of k-means cluster analysis // British Food Journal. 2017. Vol. 119. No. 7. P. 1423–1441. https://doi.org/10.1108/BFJ-08-2016-0354

18. Kaur M., Kang S. Market basket analysis: Identify the changing trends of market data using association rule mining // Procedia Computer Science. 2016. Vol. 85. P. 78–85. https://doi.org/10.1016/j.procs.2016.05.180

19. Gupta S., Mamtora R. A survey on association rule mining in market basket analysis // International Journal of Information and Computation Technology. 2014. Vol. 4. No. 4. P. 409–414. Available at: http://ripublication.com/irph/ijict_spl/ijictv4n4spl_11.pdf (accessed 21 February 2021).

20. Aguinis H., Forcum L.E., Joo H. Using market basket analysis in management research // Journal of Management. 2013. Vol. 39. No. 7. P. 1799–1824. https://doi.org/10.1177/0149206312466147

21. Muhal H., Jain H. Two-stage customer segmentation using k-means clustering and artificial // International Research Journal of Engineering and Technology. 2021. Vol. 8. No. 3. P. 485–490.

22. Punhani R., Arora V.P.S., Sabitha S., Kumar Shukla V. Application of clustering algorithm for effective customer segmentation in E-Commerce // Proceedings of the 2021 International Conference on Computational Intelligence and KnowledgeEconomy (ICCIKE), 17-18 March 2021, Dubai, United Arab Emirates. P. 149–154. https://doi.org/10.1109/ICCIKE51210.2021.9410713

23. Popović N., Savić A., Bjelobaba G., Veselinović R., Stefanović H., Ilić P.M. The implementation of hierarchical and nonhierarchical clustering for customer segmentation in one luxury goods company // Proceedings of the 37th International Business Information Management Association (IBIMA), 30–31 May 2021, Cordoba, Spain. P. 8370–8381.

24. Aktaş A.A., Tunalı O., Bayrak A.T. Comparative unsupervised clustering approaches for customer segmentation 2021 // Proceedings of the 2021 2nd International Conference on Computing and Data Science (CDS), 28-29 Jan. 2021, Stanford, CA, USA. P. 530–535. https://doi.org/10.1109/CDS52072.2021.00097

25. Suresh Y., Senthilkumar J., Mohanraj V., Kesavan S. Customer segmentation using machine learning in python // Turkish Journal of Physiotherapy and Rehabilitation. 2021. Vol. 32. No. 3. P. 4338–4342. Available at: https://turkjphysiotherrehabil.org/pub/pdf/321/32-1-521.pdf (accessed 21 February 2021).

26. Pradana M., Ha H. Maximizing strategy improvement in mall customer segmentation using *k*-means clustering // Journal of Applied Data Sciences. 2021. Vol. 2. No. 1. P. 19–25. https://doi.org/10.47738/jads.v2i1.18

27. RStudio Team. RStudio: Integrated Development for R / RStudio, PBC, Boston, MA, 2020. Available at: http://www.rstudio.com/ (accessed 21 February 2021).

28. Ngo T. Data mining: practical machine learning tools and technique, Third Edition by Ian H. Witten, Eibe Frank, Mark A. Hell // ACM SIGSOFT Software Engineering Notes. 2011. Vol. 36. No. 5. P. 51–52. https://doi.org/10.1145/2020976.2021004

29. Microsoft Corporation. Microsoft Excel. Available at: https://office.microsoft.com/excel (accessed 21 February 2021).

## Об авторах

**Нидхи Гаутам**

PhD (Computer Science & Engineering);

Assistant Professor, Fellow Senate Panjab University, University Institute of Applied Management Sciences, Panjab University, Chandigarh, India;

E-mail: nidhi.uiams@pu.ac.in

ORCID: 0000-0002-9454-3625

**Нитин Кумар**

MBA (IT & Telecommunications);

Management Trainee, Housing Development Finance Corporation Limited, Pathankot, Punjab, India;

E-mail: nk2268913@gmail.com

ORCID: 0000-0002-8887-2350