Original article
UDC 519.87/005.74 https://doi.org/10.25198/2077-7175-2023-3-72

# APPLICATION OF CLUSTER ANALYSIS OF BANK CUSTOMERS

**N. Hüseynov**
Azerbaijan State University of Economics, Baku, Azerbaijan
e-mail: Natig_huseynov@unec.edu.az

***Abstract.*** *Since companies can have thousands and more customers, effective management of this customer base is one of the most important conditions for business success. In order to know the customers, it is possible to categorize them by dividing them into small groups according to their different similarities, and then specify the type of services to be offered to them. Customer segmentation has the potential to make a difference in different businesses. The fact that the variety of products and services offered in the banking sector is increasing day by day and the transition to the digital environment is faster in this sector shows that the correct segmentation of the customers of the banks will save them more profit and time in this competitive market.*

*In this study the main intention is to divide customers into small manageable groups using clustering algorithms and to find the relative importance of these groups using multi-criteria decision-making technique. In this regard, the customer segmentation approach was implemented in one of the banks operating in Azerbaijan. Currently, the bank is one of the financial institutions with the largest service network in Azerbaijan. The bank in question provides services to more than 5 million individuals and more than 22 thousand legal entities. In addition to these, it closely participates in a number of social software applications developed by the state and applies a several of the programs for the improvement of the real sector.*

***Key words:*** *customer segmentation, K-means, cluster analysis.*

***Cite as:*** Hüseynov, N. (2023) [Application of cluster analysis of bank customers]. *Intellekt. Innovacii. Investicii* [Intellect. Innovations. Investments]. Vol. 3, pp. 72–82, https://doi.org/10.25198/2077-7175-2023-3-72.

Научная статья

# ПРИМЕНЕНИЕ КЛАСТЕРНОГО АНАЛИЗА КЛИЕНТОВ БАНКА

**Н. Гусейнов**
Азербайджанский государственный экономический университет, Баку, Азербайджан
e-mail: Natig_huseynov@unec.edu.az

***Аннотация.*** *Поскольку у компаний могут быть тысячи и более клиентов, эффективное управление этой клиентской базой является одним из важнейших условий успеха в бизнесе. Чтобы узнать клиентов, можно классифицировать их, разделив на небольшие группы в соответствии с их различным сходством, а затем указать тип услуг, которые им будут предлагаться. Сегментация клиентов может изменить ситуацию в разных сферах бизнеса. Тот факт, что разнообразие продуктов и услуг, предлагаемых в банковском секторе, увеличивается день ото дня, а переход в цифровую среду в этом секторе происходит быстрее, показывает, что правильная сегментация клиентов банков сэкономит им больше прибыли и времени в этот конкурентный рынок.*

*Целью данного исследования является разделение клиентов на небольшие управляемые группы с использованием алгоритмов кластеризации, а также определение относительной важности этих групп с использованием многокритериальной техники принятия решений. В связи с этим в одном из банков, действующих в Азербайджане, был реализован подход сегментации клиентов. В настоящее время банк является одним из финансовых учреждений с самой большой сетью обслуживания в Азербайджане. Рассматриваемый банк обслуживает более 5 млн физических и более 22 тыс. юридических лиц. Помимо этого, он активно участвует в ряде социальных программ, реализуемых государством, и воплощает ряд программ развития реального сектора.*

***Ключевые слова:*** *сегментация клиентов, К-среднее, кластерный анализ.*

***Для цитирования:*** Hüseynov, N. (2023) [Application of cluster analysis of bank customers]. *Intellekt. Innovac ii. Investicii* [Intellect. Innovations. Investments]. Vol. 3, pp. 72–82, https://doi.org/10.25198/2077-7175-2023-3-72.

## Introduction

Customer segmentation helps business to provide individual services and obtain more specific needs. The communication between consumers and the business is more connected to the customer relation management systems than before. New technologies as well as artificial intelligence brings huge competitive advantages and useful techniques for the deep understanding of target groups. In this regard, different AI algorithms are being used but in terms of the customer relations management several of them are most useful. Customer segmentation methods are being widely used in the business sectors like financial institutions that usually interact with the enormous customer size. Customer segmentation can help banks in many ways including more effective marketing and sales targeting, customer loyalty and satisfaction, product, and service improvement.

Customer segmentation research enables the bank to optimize its marketing and advertising strategies. By identifying the unique characteristics of each customer segment, the bank can create targeted marketing campaigns that resonate with each group [13]. This approach is more effective than generic advertising, as it enables the bank to communicate with customers on a more personal level, increasing the likelihood of conversion and customer loyalty. It can help the company to gain more information about priority and needs of consumers, have different policies for selected segment to amend consumer satisfaction, and increase income [9].

In total, 81 percent of global marketers report that they mainly compete on the basis of customer experience [10]. To obtain insight into target audience, banks may use big data analytics, machine learning, and other procedures. Banks can establish more effective client contact points with this data. AI removes most of the guesswork involved in client interactions, whether a bank is doing email marketing or giving customer service [5]. Modern banks harness the power of artificial intelligence and machine learning to segment client data and get a better knowledge of their data [16]. Segmenting client data allows banks to tailor customer experiences while also improving and defining goods, allowing them to swiftly respond to the demands, habits, and interests of their consumers. By analyzing the needs and preferences of each customer segment, the bank can identify gaps in its existing product line and develop new offerings that meet those needs [8]. Segmenting the market is also a crucial component of marketing, as it entails splitting customers and consumers into clusters or sections based on their specific desires and demands [7]. Cluster analysis is a powerful technique for discovering hidden patterns and structures in data. However, the effectiveness of this technique depends heavily on the quality of the data used [4].

## Developing a Customer Segmentation Framework for Personalized Services

In order to determine the customer segmentation, first of all, the processes of defining the business environment and preparing the data set were carried out. Then the customer's requirements or needs for using banking products were identified and defined because of the process, and a calculation was made for the customers. In our research, customer characteristics were determined according to the frequency of use and behavior of existing bank products, they are "Full Cash Customer"; "Pos Transaction Client"; "Installment Customer"; "Cash Preference Assorted Customer". Determining the names of these segments took place as a result of the analyzes carried out in the 4th part.

Using cluster analyses of section sets up four types of customers. These four customer groups give special priority to information services and technology. Five different bank customer commitment profiles were identified by Fullerton (2019) with various types based on factors such as size, behavior, and intentions [3]. In a study by Piercy, Campbell, and Heinrich (2011) based on demographic-based segmentation as a means of targeting customers of financial services identified 10 clusters [14].

Customer segmentation is divided into 2 parts, single variable based (SD) segmentation and multi variable based (MD) segmentation. The first approach is an approach that segments customers based on their combined characteristics, called a total score, based on a calculation of each of them according to predetermined characteristics. The second approach, on the other hand, segments customers according to the characteristics they have in common. As mentioned, customer characteristics are segmented under 4 headings in this study.

## Dividing Customers: An Approach to Customer Segmentation

Cluster analysis is an autonomous technique for machine learning that separates the input dataset into clusters in a manner that the objects within a cluster are more akin to each other than to those in other clusters. There are various methods available, for clustering analysis. However, a fundamental problem in utilizing many of the current clustering methods is that the number of clusters (k parameter) necessitates pre-specification before clustering is conducted. The parameter k is either recognized by users based on prior data or determined in a particular way. Clustering results may largely depend on the number of clusters designated.

It is necessary to provide educated guidance for determining the number of clusters to achieve appropriate clustering results. Since the number of clusters is infrequently previously recognized, the typical approach is to operate the clustering algorithm multiple times with a distinct k value for each operation [2].

Cluster analysis is a segmentation method that categorizes samples, such as individual customers, customer groups, companies, or entire countries, into homogeneous groups called clusters. The goal is to group samples within a cluster to be as similar as possible to each other while being as different as possible from samples in other clusters.

Segmenting customers is a crucial business strategy that helps companies better understand and target their customers. The first step in this process is to select the characteristics by which to group customers, which can range from A to Z. For example, a company may segment the market based on customers' price sensitivity and brand loyalty, which can be measured on a scale of 0 to 100. Customers' values for these variables can be illustrated in a graph or table.

Cluster analysis is a method used to group customers based on their similar levels of brand loyalty and price sensitivity. The aim is to assign customers to clusters where the members within each group are as alike as possible while being as different as possible from members of other clusters. Once the variables to be used for clustering have been identified, the choice of clustering procedure is essential since different techniques may need different pre-analysis preparations. Therefore, selecting an appropriate clustering method is crucial for the success of the analysis, and it requires careful consideration of the data and the research goals.

Table 1. Displaying customer data values in a table

| Customers | A | B | C | D | E | F | G | H | İ |
|---|---|---|---|---|---|---|---|---|---|
| X | 24 | 36 | 28 | 42 | 30 | 54 | 48 | 33 | 26 |
| Y | 30 | 44 | 38 | 29 | 49 | 57 | 66 | 28 | 36 |

*Source: developed by the author*

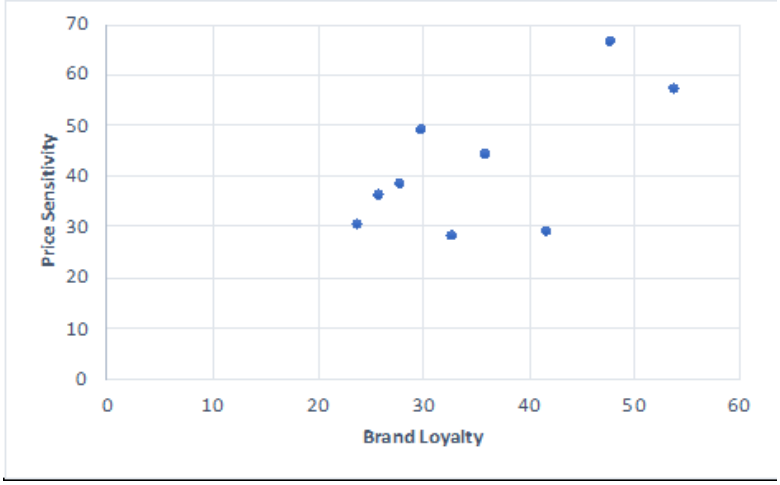


Figure 1. Display of data on a graph
*Source: developed by the author*

Although cluster analysis aims to group similar samples into clusters, the procedures used for this purpose vary and involve different stages, which will be explained in this chapter. One of the critical considerations before starting the clustering process is how to measure similarity. Most methods use similarity measures that estimate the distance between pairs of samples. The patterns that have smaller distances between them are considered more similar, while those with larger distances are seen as more different. Deciding on the number of clusters to extract from the data is another crucial aspect of cluster analysis. This final step also requires the evaluation of the stability and reliability of the clustering solution. Understanding the different

stages of cluster analysis, including measuring similarity and determining the number of clusters, is essential to the successful implementation of this technique.

**Deciding on a Segmentation Procedure**

When it comes to dividing a dataset into distinct groups, the method of segmentation chosen will dictate how clusters are formed. Creating clusters involves weighing up a range of factors, such as reducing the amount of variation within clusters (known as within-cluster variance) or increasing the separation between clusters. Another important consideration is how to gauge the similarity or dissimilarity of samples in the newly formed clusters compared to the remaining data points.

There are numerous clustering techniques available, each with their own unique characteristics and classifications. For example, clustering methods can be classified as either overlapping or non-overlapping, unimodal or multimodal, detailed or incomplete. A significant differentiation exists between hierarchical and partitioning techniques, with k-means being a popular partitioning method. In this particular study, the k-means method has been chosen as the approach for clustering.

**Partitioning Techniques: Exploring K-means Algorithm**

Separate clustering techniques are an essential group of methods used for clustering analysis. Partitioning clustering offers a more extensive selection of algorithms in comparison to hierarchical clustering. Among these algorithms, the k-means is one of the most frequently used method in market research. The k-means method follows a different approach from hierarchical clustering. One of the major differences is in the initiation of the analysis. In k-means clustering, the analyst pre-determines the number of clusters to create before initiating the analysis. The algorithm then assigns each sample to its respective cluster based on this predetermined number of clusters.

Several of the ways can be used to initiate the k-means algorithm, such as randomly selecting k samples as starting centers, using the first or last k samples as starting centers, dividing all samples randomly to the k groups, and computing the centroid of each group as initial centers, or defining an initial grouping variable to determine the groups among the samples, and utilizing the averages or medians of these groups as initial centers.

Following the initialization step, k-means clustering algorithm iteratively assigns data points to clusters in order to minimize the within-cluster variance. This variance is calculated as the squared distance between each observation and the centroid of the corresponding cluster. If reassigning a data point to a different cluster leads to a reduction in the within-cluster variance, the data point is moved to that cluster.

K-means does not create a hierarchy like hierarchical clustering because the cluster relationships can change during the analysis. Thus, k-means clustering is categorized as a non-hierarchical clustering technique

To better understand this approach, let's take a look at how it works in practice. Figures 2, 3, 4 illustrate the four steps of the k-means clustering process - studies identify several variants of the original algorithm.

Firstly, the analyst using k-means clustering must decide on the number of clusters to extract from the dataset. Once this is determined, the algorithm will select the initial centroids of each cluster randomly based on this input. For instance, in the present case, two cluster centers, BK (first cluster) and İK (second cluster), were chosen randomly and are shown in figure 2.

Subsequently, the k-means algorithm computes the Euclidean distances between each cluster center and every data point, after which it assigns each data point to the closest cluster center. As illustrated in figure 3, samples A, B, and C are assigned to the first cluster, while samples D, E, and F are assigned to the second cluster, thereby dividing the data into two distinct groups.

In step 3, the k-means algorithm calculates the geometric center of each cluster based on the initial split obtained in step 2. This is done by computing the mean values of the data points within each cluster, such as A, B, and C in the first cluster, for each of the variables (brand attachment and price dependence). After that, the cluster centers are shifted to new locations, BK2 in the first cluster and İK2 in the second cluster.

Ultimately, in stage 4, the intervals amidst each specimen and the freshly confirmed group nuclei are evaluated, and the specimens are designated to a specific cluster grounded on their minimum interval to other group nuclei (BK2 and İK2). As the site of the group nuclei modifies from the initial phase in the first stage, this could give rise to a distinct cluster settlement. This is also valid in the instance, as specimen E is presently nearer to the first group nucleus, BK, than to the second nucleus, İK2, dissimilar to the initial part. Consequently, this specimen is at present delegated to the first cluster (fig. 4).

The k-means algorithm follows a process of iteration until either a preset number of iterations is achieved or convergence is attained. Convergence is said to have occurred when there are no further changes in the communication amidst clusters.

It must be emphasized that k-means is designed to work with metric measurements, using the distance between each pair of points measured using the Euclidean

method to determine the squared Euclidean distance from the centroid. Therefore, only Euclidean distances should be used in conjunction with k-means.

It is also crucial to recognize that the results produced by k-means are influenced by the starting point which the researcher or the software choose. This implies that the algorithm could attain convergence towards a local optimum, resulting in a solution that is optimal only in comparison to analogous solutions, but not in a global context. To overcome this limitation, the k-means method should be run multiple times using different starting points.

Compared to hierarchical clustering methods, k-means requires less computational effort, making it a popular choice for datasets with large sample sizes, particularly those exceeding 500.

However, before running k-means, it is necessary to predefine the number of clusters to be established. We will address the identification of the most suitable number of clusters in the next section.

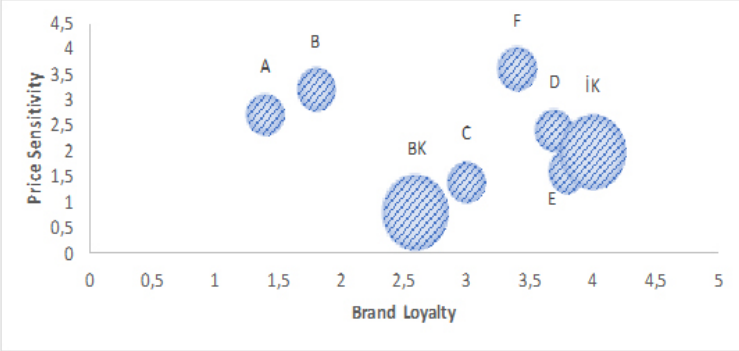


Figure 2. K-means procedure (step 1: placement of random cluster centers)
*Source: developed by the author*

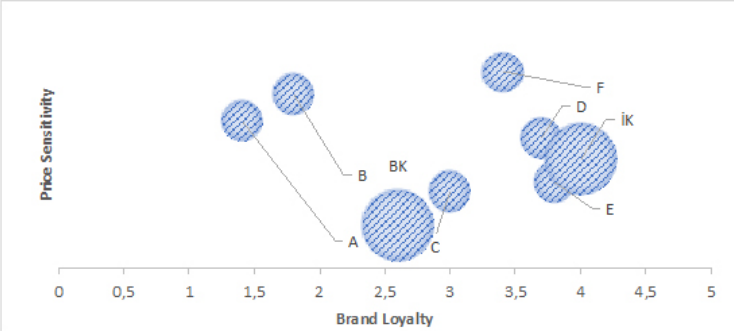


Figure 3. K-means procedure (step 2: assigning samples to the nearest cluster center)
*Source: developed by the author*

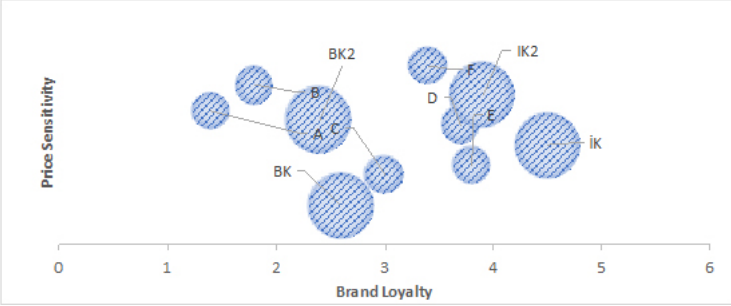


Figure 4. K-means procedure (steps 3 and 4: recalculating cluster centers and reassigning samples to cluster centers)
*Source: developed by the author*

**An Overview of Similarity Measures**

The previous sections covered the k-means procedure and the different linkage algorithms utilized in agglomerative hierarchical clustering. All of these clustering techniques rely on measurements that describe the similarity or dissimilarity between pairs of samples. This section will explore various measures that are suitable for metric variables.

Drawing a straight line between two samples is a straightforward method of evaluating the proximity between them. For instance, when examining the scatterplot depicted in Figure 1, We can easily observe that the distance between observations B and C is significantly shorter than that between B and D, which is known as Euclidean distance or straight line distance, and is the most widely used measure for analyzing variables.

The Euclidean distance, also known as the L2 distance metric, is a commonly used measure in machine learning and data science [1]. Stata, a statistical software widely used in the field, uses the term L2 to refer to the Euclidean distance as well [15]. This metric is especially advantageous in clustering algorithms as it is a useful tool for evaluating similarities between data points. Scientists may also utilize the Euclidean distance squared, which Stata designates as the square of L2. For our research technique, k-means, it is more fitting to use the Euclidean distance squared since that is how the technique computes the distances from the samples to the centroids.

To employ the hierarchical clustering technique, we need to mathematically define these distances. Using the data in Table 1, we can compute the Euclidean distance between client B and client C (referred to as d(B,C)) based on their x and y variables by applying the subsequent equation:

$$d_{Euclidean}(B,C) = \sqrt{(x_B - x_C)^2 + (y_B - y_C)^2}$$

As observed, the Euclidean distance corresponds to the square root of the total squares of the differences of the changing parameters. Employing the values in Table 1, we arrive at the following computation:

$$d_{Euclidean}(B,C) = \sqrt{(36-28)^2 + (44-38)^2} = \sqrt{100} = 10$$

This metric indicates the magnitude of the line segment that joins objects B and C. In this scenario, we only employed two variables, but we can determine distances between objects by incorporating more variables inside the square root symbol of the formula. Nevertheless, each supplementary variable will append another aspect to our research issue (e.g., if we have six grouping variables, we would have to handle six dimensions), which makes it impractical to represent the outcome graphically.

**Determining the Optimal Number of Segments**

Segmenting more or less than needed, when we are deciding on the quantity of the clusters will be a destructive factor on the business considerations like customer targeting. Various clustering techniques demand distinct methods to determine the cluster count. Thus, we discuss the partitioning methods separately according to the clustering method to be used in our study.

Studies have identified various measures for determining the number of clusters in a database. Variance ratio criterion is one of the most prominent criteria. For a solution consisting of n number of samples and k number of clusters, the variance ratio is defined as follows [6]:

$$VRC_k = (SS_B/(k-1))/SS_W/(n-k))$$

The sum of squares between clusters is denoted by $SS_B$, while the sum of squares within clusters is referred to as $SS_w$. The optimal number of clusters can be determined by selecting the value that maximizes the variance ratio. However, since the variance ratio usually decreases with more clusters, The formula below should be used to calculate the difference (ωk) between the variance values of each cluster solution:

$$\omega k = (VRC_{k+1} - VRC_k) - (VRC_k - VRC_{k-1})$$

The optimal cluster solution is achieved by selecting the value of k that minimizes ωk. Milligan and Coo-

per (1985) demonstrated that the variance ratio value is a reliable method for determining the correct number of clusters across various datasets. However, due to the term $VRC_{k-1}$ , which is not intended for a single cluster, the minimum number of selectable clusters should be three, which is regarded as a drawback of using the ωk statistic [11].

When utilizing partitioning algorithms such as K-means, it is imperative to ascertain the optimal quantity of clusters to extract from the dataset. There exist different techniques for determining the optimal cluster quantity:

One way is to calculate the degree of variance for a range of cluster numbers and select a solution that maximizes the degree of variance or minimizes ωk. This approach requires comparing the degree of variance across multiple cluster solutions to find the optimal number of clusters.

An alternative is to utilize a hierarchical algorithm to identify the optimal quantity of clusters by constructing a dendrogram followed by executing the k-means technique. This technique also enables researchers to identify initial values for cluster centers, resolving the issue of the method's sensitivity to initial classification.

Alternatively, we can make a decision on the quantity if the clusters by relying on prior knowledge or previous studies. For example, they may rely on the findings of comparable prior research to ascertain the suitable quantity of the cluster for their data set.

**Application of Cluster Analysis to the Banking Sector**

Determining customer characteristics in the banking sector according to their spending behavior and creating customer groups of customers with similar characteristics based on the resulting data is the basis of cluster analysis. For this purpose, in order to ensure the clustering of the bank's customers, the analysis was carried out by applying the k-means algorithm to the customer data of the bank to be analyzed.

**Data Collection and Preparation for Analysis**

In this study, the transaction data of credit card customers was obtained using SQL from the bank's internal database. As a result, a dataset analysis of 321,305 customers with transaction data was developed. In our table with general information, there are variables that show the customer id, the total amount of spending covering the period from the day of activation of credit cards to 31.03.2022 and the percentage of this amount spent online, cash, installments, POS. After data collection, all analyzes are performed in R software.

**Descriptive Statistics of Data**

Before creating certain clusters using the K-means algorithm, calculating descriptive statistics of the data at hand will be useful to understand the database at hand. To this end, Table 1 contains descriptive statistics of the data in the previous subsection.

Table 2. Descriptive statistics of variables

| The name of the variable | Average | Standard Deviation | Minimum | Maximum |
|---|---|---|---|---|
| **Total Spending** | 1953.260 | 9145.700 | 50 | 3420394.000 |
| **Online Spending** | 151.544 | 7970.194 | 0 | 3303477.000 |
| **Cashing out** | 1446.946 | 3411.750 | 0 | 866541.900 |
| **Installment Transaction** | 162.879 | 559.450 | 0 | 20216.940 |
| **Operation Pos** | 191.869 | 1159.913 | 0 | 193037.700 |

*Source: developed by the author*

When paying attention to Table 2, from the date of activation of credit cards to 03.01.2022, when the data was collected, the average of the amount of Total Spending by customers is 1953.26 AZN, the standard deviation is 9145.7, the minimum is 50 AZN, and the maximum is 3420294 AZN. The reason why the minimum amount is 50 AZN here is that the customers included in the analysis had less than 50 AZN usage until the date of the analysis and they are included in the inactive customer group.

Looking at the individual channels of credit card usage within the total amount of spending, the average for total online spending is 151,544 AZN, the standard deviation is 7,970,194, the minimum is 0 AZN, and the maximum is 3,303,477 AZN. The reason the maximum amount is higher here is because some customers use their credit cards as business cards and make high monthly transactions. The average for the Total Cashing operation is 1446.946 AZN, the standard deviation is 3411.75 AZN, the minimum is 0 AZN, and the maximum is 866541.9 AZN. When the important installment operations are considered, the average for

these operations is 162,879 AZN, the standard deviation is 559.45 AZN, the minimum is 0 AZN, and the maximum is 20,216.94 AZN. Finally, the average for the POS operation is 191.869 AZN, the standard deviation is 1159.913 AZN, the minimum is 0 AZN, and the maximum is 193037.7 AZN.

### Selection and Standardization of Variables for Cluster Analysis

The selection of variables to incorporate in the cluster analysis is a crucial step in the process. Considering the topic analyzed at this stage, it was decided that the most suitable variables for cluster analysis are the channels of using credit cards of customers. For this reason, online spending, cashing, installment and POS operations of its customers were selected as important variables. Nevertheless, it is noteworthy that these variables differ from one customer to another. That is, some customers spent a total of 1000 AZN and spent 500 AZN (50%) of it in an installment transaction, while another customer spent a total of 100 AZN and spent all of it (100%) in the installment channel. When doing a comparative analysis in this way, it can be seen that the first customer is a higher ranking installment customer. However, the second customer elevates to a higher-tier installment customer by utilizing the entire transaction amount, which enhances the quality of the clustering analysis. Therefore, in the subsequent phase, harmonizing the variables to the same category and unit amplifies the effectiveness of the analysis. In order to bring the variables to the same gender, the variables were standardized by dividing the amount of each customer's spending on separate channels by the total amount of spending of the customer and obtaining the percentages. In the next stage, the analysis is continued using the standardized data calculated as a percentage.

### Choosing the optimal number of Clusters

Cluster analysis is an essential tool for identifying natural groups of objects, but determining the appropriate number of clusters is a critical challenge. The number of clusters is a significant parameter that can affect the quality of the results or complicate the algorithm. Therefore, it is necessary to select the optimal number of clusters for a given dataset and research question. Several techniques are available for determining the appropriate number of clusters, such as silhouette analysis, the elbow method, and gap statistic. Researchers should consider using these techniques to ensure the accuracy and reliability of the results [12].

At this stage, it is desired to apply the k-means algorithm using our standardized variables. For this purpose, the stage of choosing the optimal number of clusters discussed before should be applied first and it is necessary to determine the suitable number of clusters. In determining the optimal number of clusters, VRC values are computed for each potential number of clusters and the number of clusters where changes in VRC value are minimal (or not very important) is decided, and this method is known in the literature as the Elbow Rule used for k-means. According to the Elbow Rule, which aids in selecting the optimal number of clusters, Figure 5 depicts the decreasing pattern of the VRC value differences.

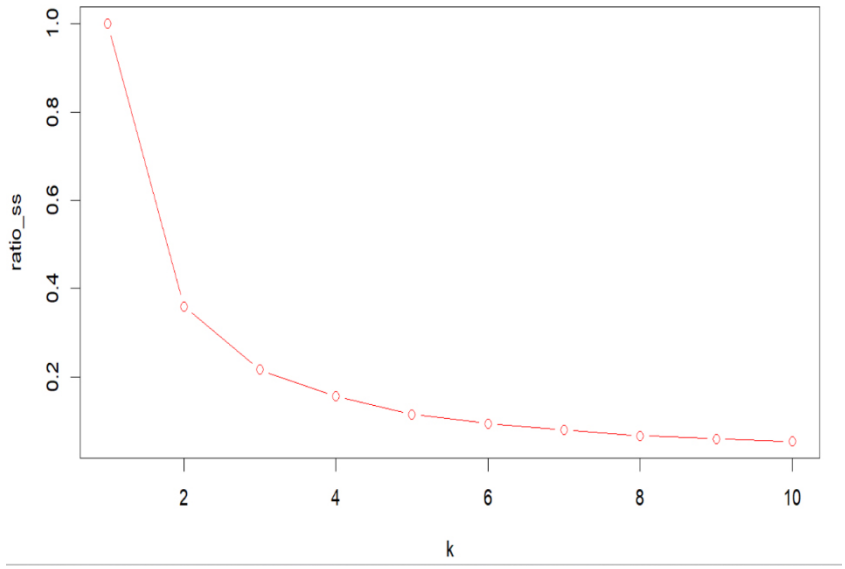


Figure 5. The Elbow Rule is utilized to select the optimal number of clusters
*Source: developed by the author*

The Elbow Rule, illustrated in Figure 5, is employed to determine the optimal number of clusters. The marginal usefulness of each cluster created, which is its ability to explain the total variance, is considered in selecting the optimal number of clusters. For example, whenever the quantity of the clusters is decided on 2, the explanatory power of the total variance increases by 0.4 (40%), while when the quantity of the clusters is decided on 3, it increases by about 0.2 (20%). After the 4th cluster, the increase in the proportion of the entire variance. disclosure stabilizes and it becomes meaningless to create more clusters per branch. As a result of the fact that the number corresponding to the full elbow part when we bend our arm is 4 (hence the name Elbow Rule), the optimal number of clusters in this study is decided to be 4.

**Generation of segments using k-means algorithm**

In the previous subsection, it was decided that the optimal number of segments is 4. At this stage, the decided segments are obtained using the k-means algorithm. Table 3 shows the general characteristics of the obtained segments and their corresponding names.

Table 3. General characteristics of the obtained segments

| No | The name of the segment | Number of customers | Percent of Segment | Online Spending | POS transaction | Installment transaction | Cash Out Transaction |
|---|---|---|---|---|---|---|---|
| **1** | Fully Cashout Customer | 204667 | 48.55% | 0.58% | 0.95% | 0.30% | 98.17% |
| **2** | POS transaction client | 15980 | 7.58% | 18.17% | 57.19% | 12.01% | 12.62% |
| **3** | Installment Customer | 18326 | 13.04% | 2.67% | 7.45% | 82.44% | 7.43% |
| **4** | A Variety of Customers with Cash Preference | 32494 | 30.83% | 9.62% | 15.79% | 12.25% | 62.33% |
| | | 271467 | | | | | |

*Source: developed by the author*

When paying attention to the obtained segments in Table 3, the first segment includes a total of 204,667 customers, and this segment covers 48.55% of the total number of customers and is the largest segment. When looking at this segment, it can be seen that 98.17% of the customers in the segment make cash transactions. For this reason, it is possible to call the customers in this segment as customers who make a cash transaction.

When we pay attention to the second segment, a total of 15,980 customers are included in this segment, and this segment covers 7.58% of the total number of customers and is the smallest segment. Customers belonging to this segment spend more of the total amount of spending, i.e. 57.19% on POS transaction, 18.17% on online spending, 12.01% on installment transactions and 12.62% on cashing transaction. Since the percentage of POS transactions in this segment has a large place, we can call this segment as POS customers.

Upon examining the third segment, it becomes apparent that there are 18,326 customers included within it and this number is 13.04% of the total customer base of the investigated bank. The reason for deciding the name of this segment as the installment customer segment is that spending habits using bank products are mostly related to installment transactions, i.e. 82.44%, as is clear from the table. Clients included in this segment perform online spending at the rate of 2.67%, POS operations at 7.45%, and cashing operations at the rate of 7.43% (among total payment operations).

The fourth and last segment of the bank's customer base is the cash-preferred diversified customer group, which is the most mixed group compared to other segments. The total number of customers in this segment was 32,494, and it has a 30.83% share in the customer base. Since cash transactions have a share of 62.33% in total transactions and there is no sharp difference between transactions, it was decided that the name of the segment should be cash-preferred variety customer segment. Online spending by customers in this segment is 9.62%, POS transactions are 15.79%, and installment transactions are 12.25%.

In order to show the created segments more clearly, the distribution graph of the segments is given in Figure 6.

If we look at the distribution graph of the segments, it is clearly visible here the distribution of different payment types of customers belonging to each segment according to their share in their total operations. 62.33% of the customers in the customer segment with preference for cashing perform cash, 12.25% installment, 15.79% POS, 9.62% online transactions. 82.44%

of customers in the installment customer segment performed installment transactions. The customers of this segment spent 7.43% cash, 7.45% POS, and 2.67% online and gave the most priority to installment payments in their total spending. 57.19% of customers in the POS transaction segment performed POS, 12.62% cash, 12.01% installment, and 18.17% online transactions. We can see the distribution of spending behaviors with sharper differences in the cash-only segment than in other segments. The share of cash transactions covered 98.17% of total spending transactions, leaving a very small portion for other transactions.

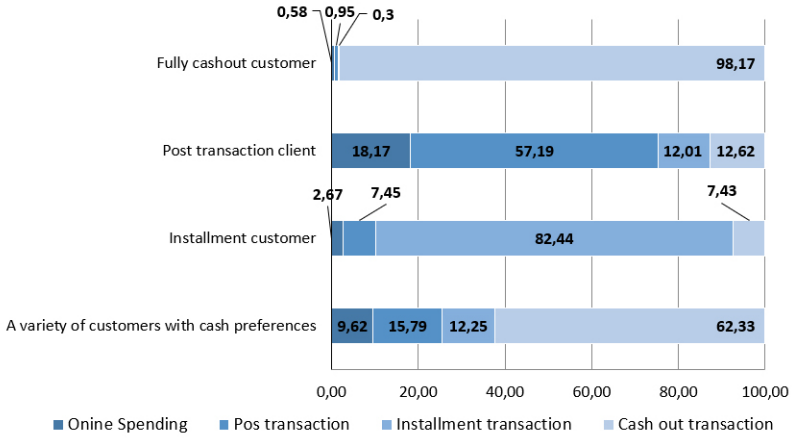


Figure 6. Distribution Chart of Segments
*Source: developed by the author*

## Results

The segments obtained as a result of the segmentation of bank customers open the door to many benefits and directions when defining the marketing and sales strategies of the bank, and make it possible to offer banking products to the most suitable target audience with a more personalized approach. In our research, the k-means segmentation algorithm, implemented on real customer data, determined the optimal number of clusters and ensured the distribution of customers according to those clusters. Most of the research conducted in the direction of increasing the efficiency of artificial intelligence in businesses deals with the perspectives in this field and the more available artificial intelligence tools. As a result of the literature review, it was determined that many studies are conducted on the regulation and integration of artificial intelligence into businesses, rather than empirical studies that are of interest to businesses, especially the banking sector. The large number of customers in the banking sector of the business and the continuous implementation of the collection of customer behavior data from the direction of spending habits are among the facilitating elements in the direction of obtaining the basic data necessary for our research.

Using the results that emerged at the end of the research, various analyzes will be conducted in future researches to evaluate the improvement of the bank's economic efficiency by applying the artificial intelligence algorithm.

## References

1. Alpaydin, E. (2020) Introduction to Machine Learning (3rd ed.). *MIT Press,* pp. 712.
2. Calinski, T., Harabasz, J. (1974) A dendrite method for cluster analysis. *Communications in Statistics*. Vol. 3. Is. 1, pp. 1–27, https://doi.org/10.1080/03610927408827101 (In Engl.).
3. Fullerton, G. (2019) Using latent commitment profile analysis to segment bank customers. *International Journal of Bank Marketing.* Vol. 38. Is. 3, pp. 627–641, https://doi.org/10.1108/IJBM-04-2019-0135 (In Engl.).
4. Hastie, T., Tibshirani, R., & Friedman, J. (2009) The elements of statistical learning: data mining, inference, and prediction. *Springer Science & Business* Media, https://doi.org/10.1007/978-0-387-84858-7 (In Engl.).
5. Jarek, K. and Mazurek, G. (2019) Marketing and Artificial Intelligence. *Central European Business Review*. Vol. 8(2), pp. 46–55, https://doi.org/10.18267/j.cebr.213 (In Engl.).
6. Kotler, P. et al. (2019) Marketing management. *Pearson Education,* pp. 840.

7. Kumar, V., & Reinartz, W. (2018) Customer relationship management: Concept, strategy, and tools. *Springer.* https://doi.org/10.1007/978-3-662-55381-7.

8. Liu, J. et al. (2019) Market segmentation: A multiple criteria approach combining preference analysis and segmentation decision. *Omega*. Vol. 83, pp. 1–13, https://doi.org/10.1016/j.omega.2018.01.008.

9. Mahr, D., Stead, S., Odekerken-Schröder, G. (2019) Making sense of customer service experiences: a text mining review. *Journal of Services Marketing*. Vol. 33. Is.1, pp. 88–103, https://doi.org/10.1108/JSM-10-2018-0295 (In Engl.).

10. Milligan, G. W., Cooper, M. (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrics*. Vol. 50(2), pp. 159–179, https://doi.org/10.1080/0022250X.1985.971362 (In Engl.).

11. Neslin, S. A., Shankar, V. (2009) Key issues in multichannel customer management: Current knowledge and future directions. *Journal of Interactive Marketing*. Vol. 23. Is. 1, pp. 70–81, https://doi.org/10.1016/j.intmar.2008.10.005 (In Engl.).

12. Patil, C., Baidari, I. (2019) Estimating the Optimal Number of Clusters *k* in a Dataset Using Data Depth. *Data Science and Engineering*, Vol. 4, p. 132–140, https://doi.org/10.1007/s41019-019-0091-y (In Engl.).

13. Piercy, N., Campbell, C., & Heinrich, D. (2011) Suboptimal segmentation: Assessing the use of demographics in financial services advertising. *Journal of Financial Services Marketing*, Vol. 16, pp. 173–182, https://doi.org/10.1057/fsm.2011.21 (In Engl.).

14. StataCorp. (2021) Stata 17 Base Reference Manual. Stata Press, https://www.stata-press.com/manuals/documentation-set.

15. Wirth, N. (2018) Hello marketing, what can artificial intelligence help you with? *International Journal of Market Research*, Vol. 60, Is. 5, pp. 435–438, https://doi.org/10.1177/14707853182776841 (In Engl.).

16. Zgurovsky, M. Z., Zaychenko, Y. P. (2020) The Cluster Analysis in Big Data Mining. *Big Data: Conceptual Analysis and Applications* pp. 1–42, https://doi.org/10.1007/978-3-030-14298-8_1 (In Engl.).

**Information about the author:**

**Natiq Hüseynov,** Candidate of Economic Sciences, Azerbaijan State University of Economics, Baku, Azerbaijan

**ORCID ID:** 0000-0003-1635-5511

email: Natig_huseynov@unec.edu.az

The paper was submitted: 16.03.2023.

Accepted for publication:01.06.2023.

The author has read and approved the final manuscript.

**Информация об авторе:**

**Натик Гусейнов**, PhD, Азербайджанский государственный экономический университет, Баку, Азербайджан

**ORCID ID:** 0000-0003-1635-5511

e-mail: Natig_huseynov@unec.edu.az

Статья поступила в редакцию: 16.03.2023; принята в печать: 01.06.2023.

Автор прочитал и одобрил окончательный вариант рукописи.