Analisis Pelanggan Toko Menggunakan Metode Klustering K-Means english.docx

by Raisa Turnitin

Submission date: 06-Jun-2024 09:06PM (UTC+0530)
Submission ID: 2396955447
File name: Analisis_Pelanggan_Toko_Menggunakan_Metode_Klustering_K-Means_english.docx (453.74K)
Word count: 1889
Character count: 10679

Store Customer Analysis Using K-Means ClusteringMethod

1st Najwa Afifi Situmorang
Computer Science Study Program, University of North Sumatra Medan, Indonesia
najwaafifi121@gmail.com

2st Jovinka Aphellia Salva
Computer Science Study Program, University of North Sumatra Medan, Indonesia
jovinkav@gmail.com

3st Dewi Sartika Br Ginting S.Kom., M.Kom
Computer Science Study Program, University of North Sumatra
Medan, Indonesia
dewidintingdg90@gmail.com

Abstract-In this article, the K-Means algorithm is used to analyze customer groups in a store to identify different customer segments based on parameters such as age, annual income, and estimated spending. The goal of this research is to create a more targeted marketing strategy using the clustering results. The ideal number of sets can be determined by using elbow and ssilhouette techniques. The analysis results show that there are four main groups of customers with different characteristics. These characteristics are then used to create the most suitable marketing strategy.

Keywords-Clustering, K-Means, Marketing, Customer Analysis, Elbow Method, Silhouette Method.

I. INTRODUCTION

In the rapidly evolving digital era, companies face increasingly complex challenges in understanding consumer behavior and preferences. An important solution to this problem is customer segmentation. It allows companies to target customer segments specifically and effectively. Through effective segmentation, companies can not only identify consumer needs and wants, but also create more targeted and effective marketing campaigns.

Grouping is one of the one of the most common most common customer segmentation methods. The k- means algorithm is a popular method for clustering data based on their similarities. Data is classified into specific groups with comparable characteristics within each group. This process can identify patterns and trends that cannot be seen through traditional data analysis. comparable within each group. This process can identify patterns and trends that cannot be seen through traditional data analysis.

The most commonly used data in this study are age, annual income, and consumer income. These factors were chosen because of their importance in determining consumer behavior. The data was prepared for cluster analysis using appropriate pre-processing techniques.

Finding suitable clusters is an important step in the K-Means process. In this study, we used the elbow and silhouette methods to estimate and determine the optimal number of clusters. The elbow method helps decide when a group joins another group. On the other hand, the silhouette function estimates how similar one group is to another, which reflects the amount of information shared.

The results show that there are four main groups with different characteristics. Each cultural group has cultural needs and preferences. Businesses can use these powerful features to develop more targeted and effective marketing strategies. For example, retailers that sell large quantities of expensive products may be confident in their quality, while low-cost, high-performance brands tend to lower prices.

Overall, this study shows that cluster analysis using K-Means not only helps better segment customers, but also helps companies develop better marketing strategies to meet the needs of each customer segment.

II. METHODOLOGY

A. *Data Collection and Understanding*

The key traits are age, income level and annual expenditure. This data represents the demographics and spending habits of the store's core customers. Use charts to explore the underlying data and better understand data distribution and component relationships.

B. *Data Preprocessing*

Data preprocessing includes several key steps:

Data cleaning means handling missing values and removing duplicates.

Instructions: Ensure each part is equally important. Therefore, the minimum-maximum standard deviation method is used to determine the data. This is especially important for the K-means method, which is sensitive to statistical parameters.

Data Conversion: Find secret data in numerical order whenever possible.

C. *Selection of Number of Clusters*

One of the important steps in K-means clustering is determining the optimal number of clusters. The process is as follows;

Elbow method: The inertia value, which is the sum of the squares of the distances to each point, is shown for different numbers of clusters. The point at which redundancy starts to decrease is when the optimal number of clusters has been selected.

Silhouette grading: Silhouette grading is used to improve compositing efficiency. If the exponent

close to 1, then the data is well classified.

D. K-Means Implementation

Step by Step Implementation of K- Means Algorithm:

Centroid Initialization: You need to have a focal point K to start the focal point.
Cluster Assignment: Assign data points to clusters based on Euclidean distance.
Recalculate Centroid Point: Updates the position of the center point using the average of all points in the group.
Iteration: Calculating midpoints and assigning clusters is repeated until convergence or at least similar midpoints are reached.

E. Model Evaluation

Evaluation results using the K-means model:

Cluster Visualization: Separate and isolate clusters using 2D or 3D space.
Cluster profile analysis: To understand the profile of each customer segment, analyze each group based on key characteristics such as average age, income, and spending.

III. Discussion

This dataset contains 16,000 rows of data, or 2000 rows x 8 columns. Customers are represented as strings containing various demographic and behavioral information. This dataset is used for cluster analysis, which helps determine customer segments based on their characteristics and behavior. This helps you create effective marketing strategies and personalized customer service.

```
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 8 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              2000 non-null   int64
 1   Gender                  2000 non-null   object
 2   Age                     2000 non-null   int64
 3   Annual Income ($)       2000 non-null   int64
 4   Spending Score (1-100)  2000 non-null   int64
 5   Profession              1965 non-null   object
 6   Work Experience         2000 non-null   int64
 7   Family Size             2000 non-null   int64
dtypes: int64(6), object(2)
```

A. Data Description

- The number of unique identification numbers for each customer is stored in CustomerID (2000 non-null, int64).
- Customer gender: The gender of the customer (2000 non-null, object), such as "Male" or "Female".
- Customer age: The age of the customer in 2000 (non-null, int64).
- Annual Revenue ($): Annual customer revenue in dollars (2000 non-null, int64).
- Spending Score (1-100): Spending score based on customer spending behavior, with values ranging from 1 to 100 (2000 non-null, int64).
- Profession: Customer occupation, with some missing data (1965 non-null, object).
- Work Experience: Customer's work experience in 2000 (non-null, int64)
- Family size: Client family size (2000 non-null, int64).

	CustomerID	Gender	Age	Annual Income ($)	Spending Score (1-100)	Profession	Work Experience	Family Size
0	1	Male	19	15000	39	Healthcare	1	4
1	2	Male	21	35000	81	Engineer	3	3
2	3	Female	20	86000	6	Engineer	1	1
3	4	Female	23	59000	77	Lawyer	0	2
4	5	Female	31	38000	40	Entertainment	2	6

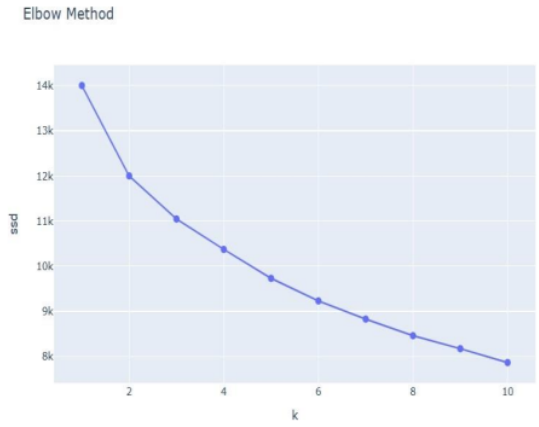
B. K-means

"k-means clustering" is used to group data into groups called "clusters" to ensure that data in one cluster is highly similar and significantly different f r o m data in other clusters. This algorithm works by calculating the centroid or cluster centroid and then organizing each data point into a cluster.

1. Elbow Chart

The Elbow method uses the K-means clustering algorithm. This method requires different steps and depends on the part number, the part number used, and the angle of the screen.
Standard square or cluster
sico-square (VCSS) is the statistical difference between the data and the data.

The Elbow method is used to maximize the cluster and minimize the VCSS in each segment. However, it is time to think about VCSS which is the causative agent of bees. The number of units used is indicated by the "elbow" or the number of units used.



2. Silhouette Method

The Silhouette method is a way to evaluate the performance of sequencing algorithms. It is used to calculate the degree of similarity of a data set within its own cluster and between other clusters. Each record is given a score from -1 to 1:

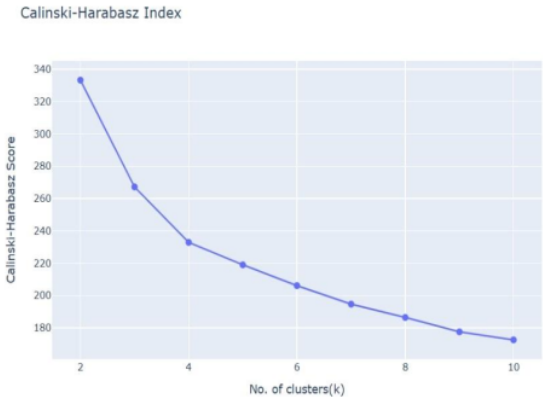
If a data group scores 1, then it is in its own group. But if it is 0, then the edge is between the two groups.

Data entry can be in the error column if given a value of -1. The score for a cluster solution is determined by summing the score points for each record in the data set. Improve your documentation solution with Silhouette GPA.

For each segment, the matching method calculates a score, k. The optimal number of segments is selected based on the high score of the silhouette. The silhouette method can be used in conjunction with the manual method to determine the optimal number of segments for a particular market.

For reference, the score can be calculated as:

To measure the distance between i and all other points in the cluster, calculate the value of (i). Also measure the distance between that point and all other points in the boundary set. The displayed value is b(i). Compare the image scores with s(i) = (b(i) - a(i))/max(a(i),b(i)).
For correlation results, the total score is the average of all scores in each dataset.



3. Calinski-Harabasz Index

The relative density index, or Calinski-Harabasz index, is a statistical filtering measure that calculates the ratio of density in a given area to that of the dispersed area.

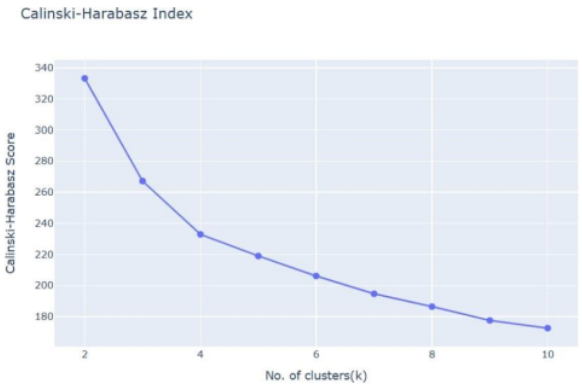
The distance between the individual points in the group and the center is called the standard deviation within the group. The distance between the centers of each group is called the standard deviation.

The formula:

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}$$

where n is the total number of data, B(k) is the difference between groups, W(k) is the difference between groups, and CH(k) is the Kalinsky-Harabasz index for a given value of k (number of groups)).

The ratio of within-cluster to within-cluster dispersion is expected to increase the Calinski-Harabasz index. Calinski-Harabasz index values are added to obtain larger clusters

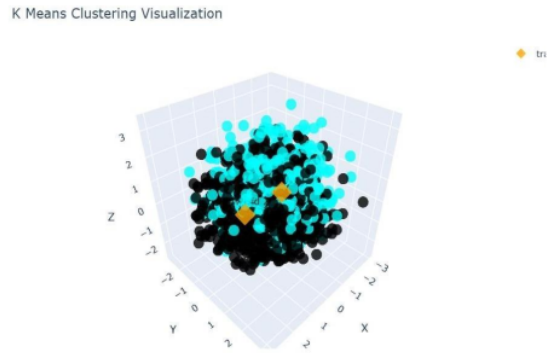


C. Visualization

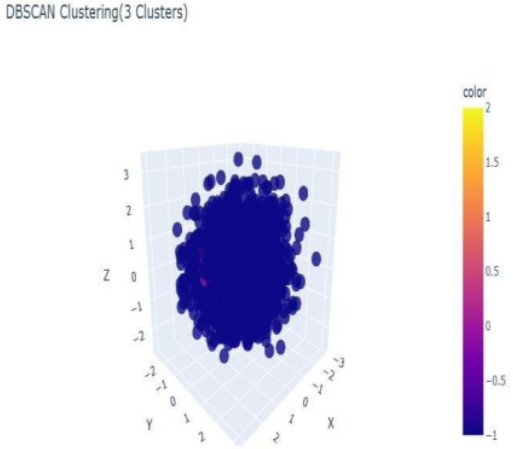
1. 3D VISUALIZATION:

With the help of a three-dimensional presentation, we depict a three-dimensional distribution of data. This creates three distinct differences. The survey presents annual revenue and cost figures as well as customer experiences in 3D imaging. Three-dimensional images reveal shapes and clusters that are not visible in a two-dimensional view. It facilitates the definition of three-dimensional information distribution and the construction of k- means clusters.

- K-means Visualization



- DBCSAN Clustering

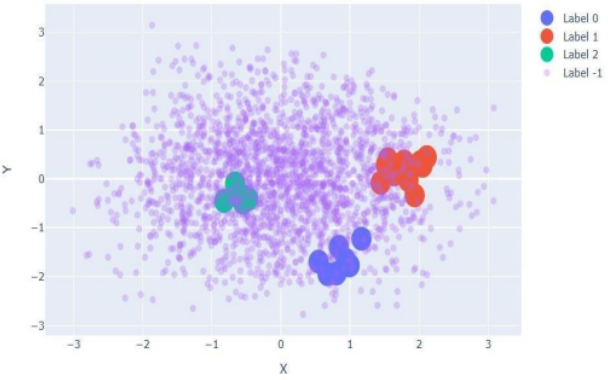


2. 2D VISUALIZATION:

This study used large 2D visuals to examine the distribution of data. Several variables were compared to compare the relationship between variables and to identify which groups were supported. For example, revenue recognition is related to the number of products sold. The 2D plan is simple and easy to match with the 3D map, but the details are clearly visible.

- K-means Visualization

- DBCSAN Clustering



IV. CONCLUSION

This K-means journal is used to estimate customers based on age, annual income, and estimated expenditure. The process begins with data collection and preprocessing. This process involves cleaning, storing, and converting non-numeric data into numeric form. The researchers used the elbow and silhouette method to identify the optimal number of clusters. The trends and trending of customer data were determined by the K-Means algorithm.

The findings show that each of the four primary consumer clusters has different qualities. The consumption patterns and demographics of each cluster are different. Younger clients with moderate income and spending, for example, are not the same as older clients with similar income and spending. These findings help businesses to better understand

The needs and preferences of each consumer segment, allowing them to develop more focused and successful marketing plans.

Overall, this study shows that the K-Means method for customer segmentation can provide useful information about customer preferences and behaviors. Using this information, companies can create marketing strategies that are better suited to each customer segment, thereby increasing customer satisfaction and, ultimately, revenue. This method demonstrates the importance of data analysis in creating smarter and more effective business plans.

REFERENCES

[1] Hua H Y, Zhao H C. Application of Clustering Algorithms in Bank Customer Segmentation [J].

[2] Syakur, M. A., et al. Integration of k-means clustering method and elbow method for identification of the

[3] Shop Customer Clustering. https://www.kaggle.com/code/utkarshsaxenadn/shop-customer-clustering/notebook#Data-Visualization

Analisis Pelanggan Toko Menggunakan Metode Klustering K-Means english.docx

ORIGINALITY REPORT

8%	6%	2%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	1%
2	Submitted to University of Chichester Student Paper	1%
3	cris.brighton.ac.uk Internet Source	1%
4	ojs.unud.ac.id Internet Source	1%
5	doi.org Internet Source	1%
6	Submitted to University of Sheffield Student Paper	1%
7	repository.tudelft.nl Internet Source	1%
8	conference.trunojoyo.ac.id Internet Source	1%
9	Thomas Käster. "Comparing Clustering Methods for Database Categorization in	1%

Image Retrieval", Lecture Notes in Computer Science, 2003
Publication

10	fasilkom-ti.usu.ac.id Internet Source	<1%
11	www.researchgate.net Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		