Customer Segmentation Analysis using K-Means Clustering

Krishnendu Jana
22MA60R29

Manishankar Bag
22MA60R13

Rishav Karmahapatra
22MA60R09

Subhankar Pal
22MA60R08

Swarup Bej
22MA60R17

April 9, 2023

AI & ML Project Report
Subject No. MA60274

Computer Science and Data Processing
Dept. of Mathematics, IIT Kharagpur

Contents

Abstract

This project report focuses on the application of K-Means Clustering Algorithm to segment customers for targeted marketing campaigns. The aim of the study is to identify distinct customer groups based on their purchasing behavior and demographic information, which would enable businesses to tailor their marketing strategies to specific customer segments. The study uses a dataset of customer transactions and demographic information, which was pre-processed and cleaned before applying K-Means clustering. The results of the clustering analysis are presented through visualizations and statistical measures, such as the silhouette coefficient and within-cluster sum of squares . The study found that the optimal number of clusters was five, which corresponded to distinct customer groups based on their purchasing patterns and demographic characteristics. The report concludes by highlighting the potential benefits of using customer segmentation for marketing campaigns and the limitations of the K-Means clustering algorithm.

Keywards : K-Means Cluster, Customer Segmentation, Unsupervised Learning.

1 Introduction

Customer segmentation analysis is a powerful tool that businesses can use to better understand their customers and tailor their marketing strategies accordingly. One popular method for conducting customer segmentation analysis is through the use of unsupervised machine learning algorithms, such as k-means clustering.

K-means clustering is a technique that involves grouping similar data points into clusters based on their similarities. In the context of customer segmentation analysis, this means grouping customers together based on shared characteristics such as demographics, purchasing behavior, and preferences. By doing so, businesses can identify different customer segments with unique needs and tailor their marketing efforts to each segment in a more targeted and effective manner.

In this analysis, we will explore how k-means clustering can be used for customer segmentation analysis, including the steps involved in the process and how to interpret the results. We will also discuss some common challenges and limitations associated with this approach and explore some potential solutions. Overall, this analysis will provide a comprehensive overview of how businesses can use k-means clustering to better understand their customers and improve their marketing strategies.

1.1 Cluster

In data science, a cluster refers to a group of data points that share similar characteristics or properties. Clustering is the process of grouping these data points together based on their similarities, such as their attributes or behaviors. The goal of clustering is to identify groups of similar data points that can be analyzed or used for further processing.

Example 1.1 (**Retail Marketing**)**.** Retail companies often use clustering to identify groups of households that are similar to each other.

For example, a retail company may collect the following information on households:

- Household income
- Household size
- Head of household Occupation
- Distance from nearest urban area

They can then feed these variables into a clustering algorithm to perhaps identify the following clusters:

- Cluster 1: Small family, high spenders
- Cluster 2: Larger family, high spenders
- Cluster 3: Small family, low spenders
- Cluster 4: Large family, low spenders

The company can then send personalized advertisements or sales letters to each household based on how likely they are to respond to specific types of advertisements.

Example 1.2 (**Health Insurance**)**.** Actuaries at health insurance companies often used cluster analysis to identify "clusters" of consumers that use their health insurance in specific ways.

For example, an actuary may collect the following information about households:

- Total number of doctor visits per year
- Total household size
- Total number of chronic conditions per household

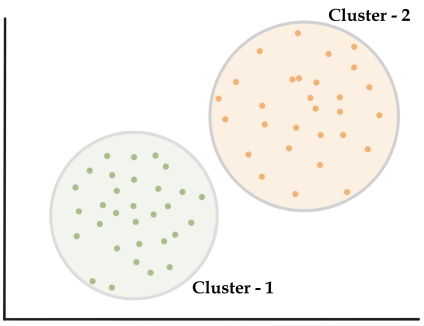

Figure 1: Two Clusters

- Average age of household members

An actuary can then feed these variables into a clustering algorithm to identify households that are similar. The health insurance company can then set monthly premiums based on how often they expect households in specific clusters to use their insurance.

1.2 Properties of Cluster

Property 1 (Cohesion)**.** Clusters are characterized by a high degree of cohesion, which means that the objects within a cluster are more similar to each other than to objects in other clusters. See Figure (1).

Property 2 (Separation)**.** Clusters are also characterized by a high degree of separation, which means that the objects in different clusters are dissimilar from each other

1.3 Applications of Clustering in Real-World Scenarios

Clustering is a widely used technique in the industry. It is actually being used in almost every domain, like -

1. Customer Segmentation
2. Document Clustering
3. Image Segmentation
4. Recommendation Engines

2 Methodology

From the first property of clusters – **it states that the points within a cluster should be similar to each other**. So, our aim here is to minimize the distance between the points within a cluster.

2.1 Algorithm : K-Means Cluster

The algorithm for K-means clustering can be summarized in the following steps:

1. First choose the number of clusters K.
2. Randomly select K points from the dataset to serve as the initial centroids for the clusters.

3. Assign each data point to the nearest centroid based on the Euclidean distance between the point and the centroid.

4. Recalculate the centroids for each cluster by taking the mean of all the points assigned to that cluster.

5. Repeat steps 3 and 4 until the centroids no longer move significantly, or a maximum number of iterations is reached.

2.2 Stopping Criteria for K-Means Clustering

There are essentially three stopping criteria that can be adopted to stop the K-means algorithm:

1. Centroids of newly formed clusters do not change

2. Points remain in the same cluster

3. Maximum number of iterations are reached

2.3 Data Collection

The data set for this project has been taken from Kaggle,. This dataset contains customer information such as *CustomerID, Gender, Age, Annual Income (k $) , Spending Score.*

3 Data Visualization

Here we have analyze the data using Python and implemented K-Means Clustering. Then we get the these results:

CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40

Figure 2: First five rows of the data

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

Figure 3: Data Description

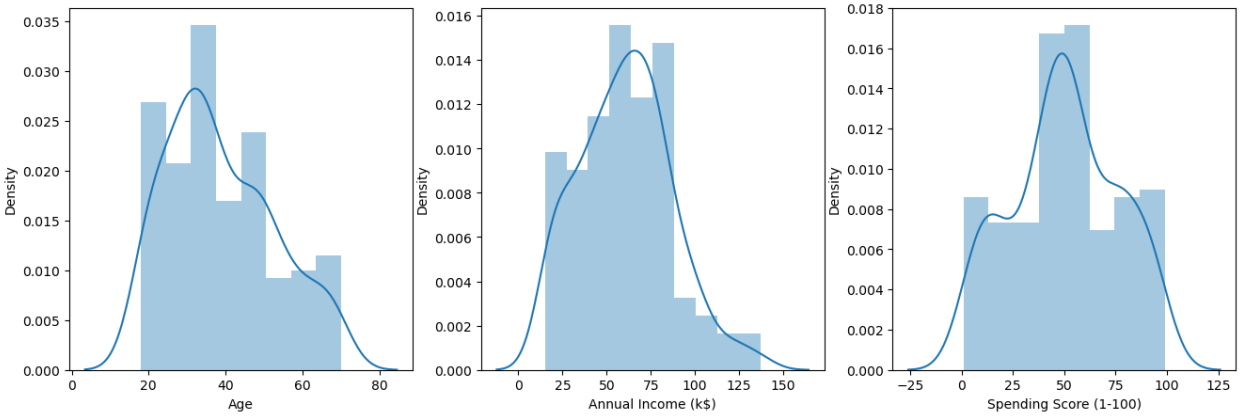

Figure 4: Data Visualization

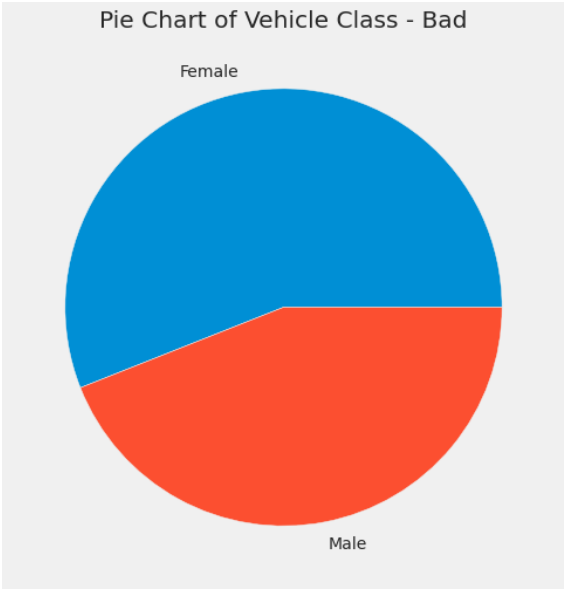

Figure 5: Pie diagram - Male vs Female

4 Results

Now after analysing the data we see get some results. The results are as follows:

1. From the Figure (3) we see that in this mall there are 25% people has average age = 28 years, 50% people has average age = 36 years, 50% people are of average age = 50 years.
2. From the Figure (4) we see that :
   - In Age columns most people belong to 18 to 50 Age,
   - Maximum Annual Income 45k to 90k.
   - Maximum Spending Sore is 50
3. From the Figure (5) we see that females are going to that mall more than males.
4. From the Figure (7) we see that there are $K = 5$ clusters for this model(Annual Income vs Spending Score). This Clustering Analysis gives us a very clear insight about the different segments of the customers in the Mall. There are clearly Five segments of Customer Clusters namely [See Figure (8)]:
   - Cluster 1 (Red Color) : Earning high but spending less.

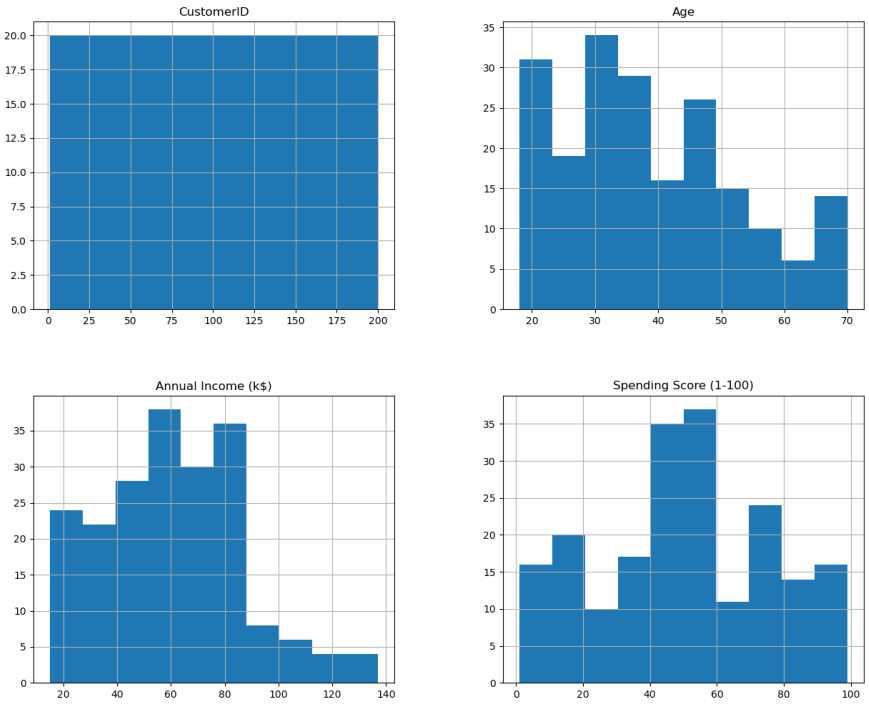

Figure 6: Histogram of the data

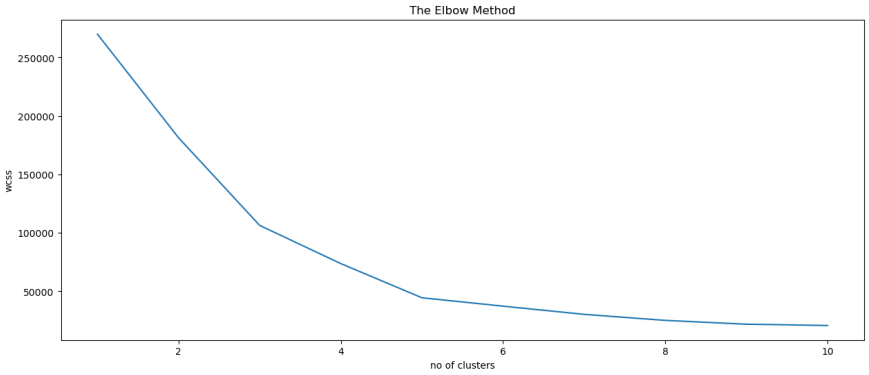

Figure 7: Elbow Curve for number of clusters

- Cluster 2 (Blue Color) : Average in terms of earning and spending.
- Cluster 3 (Green Color) : Earning high and also spending high [**TARGET SET**].
- Cluster 4 (Cyan Color) : Earning less but spending more.
- Cluster 5 (Magenta Color) : Earning less , spending less.

5 Conclusion

Targeting both high-earning and high-spending customers in a shopping mall can be a strategic move to increase revenue and profitability. Here are some possible discussions on how to attract and retain these two types of customers:

1. **Understanding the needs and preferences of high-earning customer** : High-earning customers typically prioritize convenience, quality, and exclusivity in their shopping experience. Therefore, shopping malls can cater to their needs by providing valet parking, personalized services, upscale brands, and VIP lounges. Moreover, high-earning customers tend to be time-poor, so offering extended hours, online shopping, and delivery services can increase their loyalty and satisfaction.

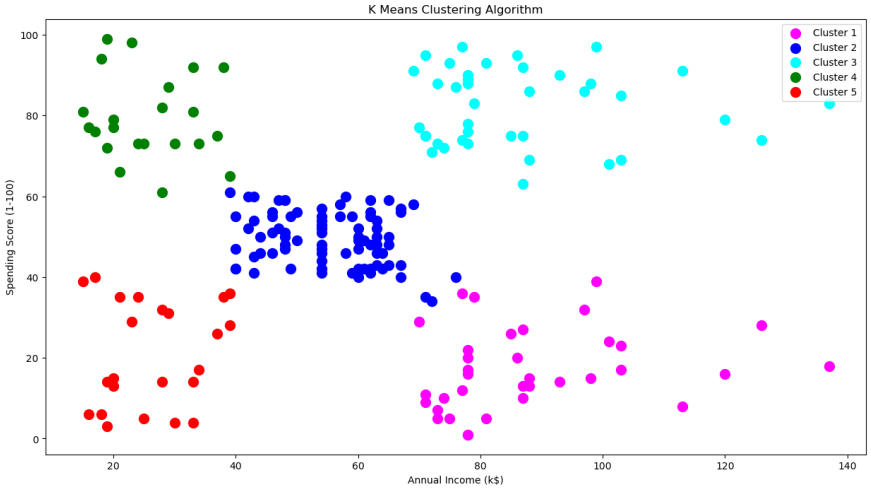

Figure 8: Clusters Visualization

2. **Offering value propositions for high-spending customers** : High-spending customers may not necessarily be high-earning, but they are willing to splurge on luxury items or experiences. Therefore, shopping malls can create value propositions for them by offering discounts, loyalty programs, gift cards, free samples, and other incentives that make them feel appreciated and rewarded for their spending. Moreover, organizing events, promotions, and partnerships with premium brands can enhance their shopping experience and entice them to spend more.

3. **Enhancing the ambience and atmosphere of the shopping mall** : Both high-earning and high-spending customers value the ambiance and atmosphere of the shopping mall. A well-designed and maintained environment with pleasant lighting, music, and decor can create a positive mood and stimulate their senses. Moreover, having amenities such as restaurants, cafes, cinemas, spas, and playgrounds can offer a holistic experience that encourages customers to stay longer and spend more.

4. **Leveraging technology to personalize the shopping experience** : Technology can play a crucial role in attracting and retaining high-earning and high-spending customers. For instance, implementing a loyalty app that offers customized recommendations, discounts, and promotions based on their shopping history and preferences can increase their engagement and loyalty. Moreover, using artificial intelligence and data analytics to analyze their behavior and anticipate their needs can offer a seamless and personalized shopping experience that enhances their satisfaction and loyalty.

In conclusion, targeting high-earning and high-spending customers in a shopping mall requires a deep understanding of their needs, preferences, and behavior. By offering tailored value propositions, enhancing the ambiance and atmosphere, leveraging technology, and providing personalized services, shopping malls can attract and retain these valuable customers and boost their revenue and profitability.

References

[1] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.