# Shopping Mall Customer Data Segmentation Using Machine-Learning Algorithm

Jerin P Jose
MCA Department
Amal Jyothi College of Engineering,
Kanjirappally,Kerala, India
jerinpjose@mca.ajce.in

Ms.Rini Kurien
MCA Department
Amal Jyothi College of Engineering,
Kanjirappally,Kerala, India
rinikurian@amaljyothi.ac.in

***Abstract*— The Customer segmentation is the process of segmenting customers based on example of their purchase behavior over the previous time (e.g.: years). The data of Mall Customers are used for the predictions. It is unlabeled data that comprises information about mall customers (features include Annual income (k$), Spending score, Genre and Age. Our goal is to group consumers together based on their annual income and expenditure score. This method can be used to find each customer's salary and expenses. The output can be seen using a graphical way.**

## I. INTRODUCTION

As new businesses spring up on a regular basis, it's becoming more important than ever for current businesses to adopt marketing strategies in order to stay competitive. The simple marketing rule in today's world is "change or perish." It gets more difficult for businesses to address the needs of each consumer as the number of customers grows. Data mining can help identify hidden trends in a company's database in this situation. Customer segmentation, also known as client segmentation, is a data mining technique that allows organisations to group together customers that have similar behaviours, making it easier to manage a big customer base. Because segmentation opens up many new paths to discover, such as which segment the product will be good for, customising marketing plans for each segment, providing discounts for a specific segment, and deciphering the customer and object relationship, which was previously unknown to the company, it can influence the marketing strategy directly or indirectly. Customer segmentation allows companies to see what their customers are buying, allowing them to better serve them and improve customer happiness. It also allows companies to identify their target customers and improve their marketing strategies in order to increase income from them.

Client relationship management and maintenance have always played an important role in providing business knowledge to firms in order to construct and develop a beneficial long-term customer relationship. Organizations have a invested to interest in developing customer acquisition, to maintain the development in strategies. Business intelligence plays a critical role in allowing companies to leverage their technological expertise and obtain a deeper understanding of their customers through outreaches. By using the clustering techniques likes the k-means algorithm to customers with similar means are clustered together. Client segmentation helps the marketing team in identifying the exposure of several customer segments that think in different ways and use distinct buying tactics. Customer segmentation aids in the identification of customers with varying interests, attributes, desires and expectations. The fundamental goal of customer segmentation is to group people with similar interests so that the marketing team can come up with an effective marketing strategy. Clustering is an iterative process for extracting knowledge from large volumes of unstructured data. Clustering is a data mining exploration approach that is used in a variety of applications such as machine learning, classification, and pattern recognition..

## II. LITERATURE VIEWS

Because the business world is so competitive, organisations must match their customers' wants and attract new consumers based on their aspirations in order to grow their business and raise revenues. Identifying and meeting the needs of each customer is a difficult and time-consuming task. This is due to the fact that clients have different demands, quality, and tastes, among other things. In contrast to a "one-size-fits-all" strategy, customer segmentation splits customers into groups based on similar attributes or behaviours. Customer segmentation is a marketing approach that divides the market into homogeneous groups depending on the customers. The information used in the customer segmentation technique, which divides customers into groups, was based on a variety of elements, including data geographic circumstances, economic conditions, demographical conditions, and behavioural tendencies. A customer segmentation strategy can help a company make better use of its marketing budget, gain a competitive edge over competitors, and display a deeper understanding of customer expectations. It can also help a business improve marketing efficiency, identify new market opportunities, establish a stronger brand strategy, and evaluate client retention.

### *A. Clustering*

Clustering is one of the most used data exploration techniques for gaining a comprehensive grasp of the data structure. It is an unsupervised learning technique. It can be defined as the task of locating subtitles and subgroups within a large dataset. Many subcategories contain data that is similar. A cluster is a grouping of aggregated data pieces that share common properties. Clustering is used in market basket

*DOI: 10.5281/zenodo.6367141*
*ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam*

research to classify customers based on their income and transactions.

Algorithm Clustering techniques create clusters that are similar within themselves depending on specific features. The distance between the items in space is used to define similarity. One of the most often used centroid-based algorithms is the K-means method. Assume D is a data collection with n items in space. Methods of partitioning separate objects in D into k clusters, C1,C2..,Ck, i.e., Ci D and Ci Cj = for (1 I j k). In a centroid-based dividing strategy, the centroid of the cluster is (Ci), then it is utilised to represents that all the cluster. In terms of idea, the centroid of a cluster is its centre point. Dist(p,ci) calculates the difference between an item p Ci and the cluster representative ci, where dist(x,y) is the Euclidean distance between two points on the x and y axes.

The most frequent strategy for answering this question with the k-means clustering method is the so-called elbow method. It involves looping the algorithm with an increasing number of cluster choices and plotting a clustering score as a function of cluster number.

We change the values (numbers) of clusters (K) from 1 to etc. using the Elbow approach. The WCSS is calculated for each value of K. (Within Clusters, Square Sum of Squares) The sum of squared distances between each site and the cluster's centroid is the WCSS. When plotted, the WCSS with the K value looks like an Elbow. The WCSS value will drop as the number of clusters increases. When K = 1, the WCSS value is highest. We can see from the graph that it will shift at a certain point, forming an elbow shape. The graph begins to shift practically parallel to the X-axis at this point. The optimum (K) value, or the optimal number of clusters, is found at this stage.

In this studies these machine learning algorithm are (Elbow Method and K-Means Clustering ) implemented using python programming. It is used for customer segmentation.

## III. RESULTS AND DISCUSSION

### 1. *A. K-Means Clustering algorithm*

It is the most fundamental and extensively used unsupervised iterative learning algorithm. In this strategy, it can be randomly initialized K centroids in the data (the number of k is calculated using the Elbow method, which will be detailed later in this articles) and iterate these centroids until the position of the centroid does not any changes. Let's go over the steps involved in K means clustering for a better and clearer understanding.

1) Determine the number of clusters in the dataset(K)

2) Determine the number of centroids to be used(K).

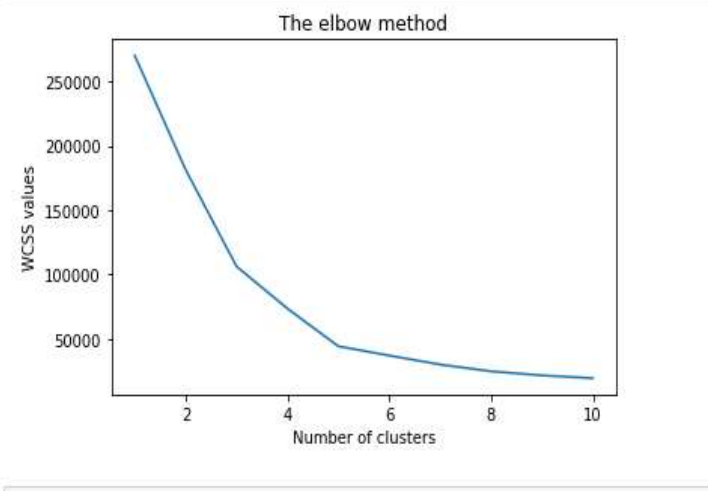
3) Assign the points to the nearest centroid using the Euclidean or Manhattan distance, resulting in K groups.

4) Now, in each group, locate the original centroid.

5) Reassign the whole data point depending on this new centroid, then resume the step.

### 2. *B. Elbow Method*

In the Elbow approaches, we may easily change the number of clusters (K) from 1 to 10. We're crunching the figures right now. For each value of K, calculate the WCSS. (Square Sum of Squares Within Clusters) The sum of WCSS is the squared distances and its between each point in a cluster and its centroid. When we plot the WCSS with the K value, we obtain an Elbow. The WCSS value decreases as the number of clusters grows. When K = 1, the WCSS value is highest. Looking at the graph below, we can observe how it quickly shifts at one point, forming an elbow shape. The graph begins to run roughly parallel to the X-axis at this point. The ideal K value, or the optimal number of clusters, is equal to the K value at this point.




### 3. Then, using Python, we can build K-Means Clustering

### 4. *C. Dataset*

The Mall Customers or the visitors dataset is being used in this case. It is unlabeled data that includes information about mall customers (such as annual income (k$), genre, age, and spending score). Our aime is to group consumers based on key characteristics such as annual income and spending score. The dataset can be downloaded from the web. The dataset is:

*DOI: 10.5281/zenodo.6367141*
*ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam*

| | CustomerID | Genre | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|---|
| 0 | 1 | Male | 19 | 15 | 39 |
| 1 | 2 | Male | 21 | 15 | 81 |
| 2 | 3 | Female | 20 | 16 | 6 |
| 3 | 4 | Female | 23 | 16 | 77 |
| 4 | 5 | Female | 31 | 17 | 40 |
| 5 | 6 | Female | 22 | 17 | 76 |
| 6 | 7 | Female | 35 | 18 | 6 |
| 7 | 8 | Female | 23 | 18 | 94 |
| 8 | 9 | Male | 64 | 19 | 3 |
| 9 | 10 | Female | 30 | 19 | 72 |

Here we can print the total data size of the data from the given dataset. From the above mentioned, 199 is the column and 5 is the rows of the dataset, that is

```
In [4]: df.shape
Out[4]: (200, 5)
```

We can find the structure of the dataset is and memory usage from the below mentioned keyword df.info().

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Genre                   200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

From the dataset we can take the specific rows that is (annual income and Spending score)

```
In [6]: x = df.iloc[:, [3,4]].values
In [7]: x
Out[7]: array([[ 15,  39],
               [ 15,  81],
               [ 16,   6],
               [ 16,  77],
               [ 17,  40],
               [ 17,  76],
               [ 18,   6],
               [ 18,  94],
               [ 19,   3],
               [ 19,  72],
               [ 19,  14],
               [ 19,  99],
               [ 20,  15],
               [ 20,  77],
               [ 20,  13],
               [ 20,  79],
               [ 21,  35],
               [ 21,  66],
               [ 23,  29],
```

Then we have to find the optimal value K for clustering the data. And now we are using the Elbow method to find the optimal of K value, that is.

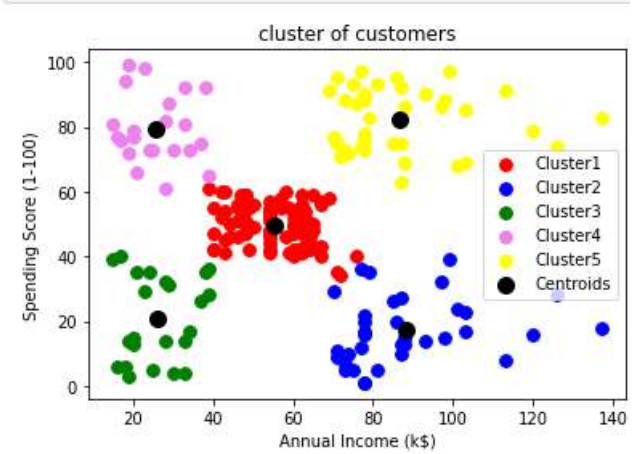
```
optimal of K value.
from sklearn.cluster import KMeans
wcss = [] for i in range(1, 11):
for i in range(1, 11):
kmeans = KMeans(n_clusters = i, init = 'k-means++',
random_state = 42)
kmeans.fit(X)
wcss.append(kmeans.inertia_)
```

The point at which the elbow shaped is formed is five(5), corresponding to its the K value or optimal number of clusters. Let's now train the model on the dataset using a range of clusters 5 that is,

```
kmeans = KMeans(n_clusters = 5, init = "k-means++", random_state = 42)
y_kmeans = kmeans.fit_predict(x)
```

Now the plot of all the clusters is using the matplotlib the annual income and spending score is the below graph that is

*DOI: 10.5281/zenodo.6367141*
*ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam*




Based on the results from these analyses, The Blue color customer have relatively less salary and their rate is high. Similarly, the black-colored customers have quite high salary but their spending rate is pretty low. While comparing these two, the red colored customers are well different from those. They have a decent salary and they are well spending in nature. At the output, they became the target audience of the companies. The company's focus more on red customers to promote their product so that there is a huge chance to get it sold!

## IV. CONCLUSION

Most businesses can benefit from segmentation in order to better identify and categorise customers based on their purchasing behaviour, income level, age, and other criteria. Clustering is a technique for splitting or subdividing photographs based on colour pattern, grey level, contract, and other criteria. In companies like e-commerce, understanding consumer segmentation is critical for upselling and cross-selling. When it comes to client mall spending, however, this is just the tip of the iceberg. Buyer or customer segmentation is a broad word in and of itself.

## V. REFERENCES

[1]https://www.geeksforgeeks.org/k-means-clustering-introduction/

[2]https://techwakerai.blogspot.com/2020/10/k-means-clustering-machine-learning.html/

[3] https://www.geeksforgeeks.org/machine-learning/

[4] https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/

[5] https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/

*DOI: 10.5281/zenodo.6367141*
*ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam*