**CUSTOMER SEGMENTATION BY USING K-MEAN CLUSTER ANALYSIS**

Customer segmentation is a practice dividing company customer into groups that reflect similarity between each other. The goal of fragmenting customers is to decide how to relate to customers in each segment to maximize the value of each customer to the business. This procedure makes it easier to target specific customer groups with customized products, services, and marketing strategies.

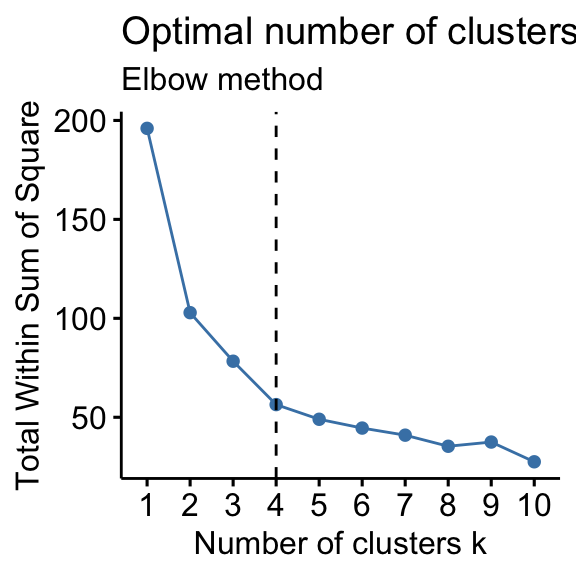
Advantages of customer segmentation:

1. **Price optimization:**  The customer and financial status of your customers will help you to pay some with the price optimization accordingly data helps in better resource allocation, which in turn helps you to gain economies of scale.
2. **Enhances competitiveness:** If you segment your market, you are well known to your customers, and you can even come up with new products according to the changing preferences of your customers if you are vigilant enough you may also have the first movers’ advantage in a product segment customer segmentation will also broaden your channel of distribution by increasing competition.
3. **Brand awareness:** You can make your brand known to your customers by segmenting them. Identifying your brand will allow your customers to directly engage with your product, increasing your market goodwill and the brand value that has been established among your competitors.
4. **Acquisition and retention:** A personalized connection with your customers will help you win satisfied customers. Approximately 75% of satisfied customers are more likely to stay with the organization that meets their needs on a regular basis. Better customer segmentation will result in a more positive relationship with your prospective customers. Customer segmentation enables you to learn a lot about your customers and better cater to their needs.

**K- MEANS ALGORITHM:**

It is an iterative algorithm that seeks to divide the data set into k predefined, distinct, non-overlapping subgroups, where each data point only belongs to one group. The algorithm seeks to maximise the similarity between data points within a cluster while maintaining the maximum degree of differentiation between clusters. The more homogeneous the data points are within a cluster, the less variation there is between clusters as measured by the sum of the squared distances between the data points and the cluster centroid.

**Elbow Method:**

This method is used to choose the optimum value of K.

Here we draw a curve between WSS (within sum of squares) and the number of clusters.

It is called the elbow method because the curve looks like a human arm and the elbow points gives us the optimum number of clusters as we can see that after the elbow points there are very slow changes in the value of WSS so we should take the elbow point value as the final number of clusters.

<https://www.datanovia.com/en/wp-content/uploads/dn-tutorials/004-cluster-validation/figures/015-determining-the-optimal-number-of-clusters-k-means-optimal-clusters-wss-silhouette-1.png>

**Content of project:**

The owner of supermarket mall gathering the data through membership card where owner collect some basic data such as customer ID, age, gender, annual income, and spending score.

Spending score: Assign to the customer based on defined parameter like customer behavior and purchasing data.

**Problem statement:**

The owner of the mall wants to understand the customer like who can be easily converge (target customer) so that the sense can be given to marketing team and plan the strategy accordingly.

**Methodology:**

The above objective is achieved by using K-means clustering to customer segmentation.

**Environment and Tools for the project:**

Language: Python

Libraries: pandas, NumPy, matplotlib, scikitlearn, seaborn

**Analysis:**

1. Mean value of age, annual income and spending score is 38.85, 60.56 and 50.20 respectively. The average income of 200 members is 60.56k dollars, with a maximum income of 137k dollars and a minimum income of 15k dollars. In addition, the customer's majority income is around 78k dollars.
2. Customer ID, Age, Annual Income and Spending Score are integer types. The Gender is of object datatypes consists of male and female data.
3. There is no null data in dataset.
4. Age group around 20 to 40 has maximum density means data set of age consists of more that between 20 to 40 and peak at 30. Similarly Annual income has maximum density between 50 to 100 k. The large amount of customer has annual income between 50 to 70 k. The most people have spending score of 50.
5. The customer of age between 26-35 has a greater number of customers followed by age group of 46 to 55 and then age group between 36 to 45.
6. There is not much relation between Annual income and spending score but there is some relation i.e., more concentration of data point between annual income in between 40 to 60 K$ which lies between spending score of 40 to 60.
7. There is total five cluster are formed between three dimensions i.e., Age, Annual Income and Spending Score. The 3d plot are formed with five colors of red, green, orange, blue and purple.
8. The customer in different cluster in blue, red, green, orange and purple are 39, 23, 79, 36 and 23 respectively.
9. The owner of mall needs to focus on orange cluster which has high annual income and low spending score, to those customers need to give some offer so that they incline to increase more average order value.
10. The cluster with purple color which as 23 customers for which owner need to give more discount as they have less annual income but for spending score so that they can increase in overall revenue.