**Business Optimization Using Data Clustering**

**Introduction/Problem definition**

These days, business is being competitive in the ways that have never been witnessed. The race to grab the greater market share by this business has led to introduction of several technologies and strategies in the economy. One of the key revolutions in understanding of consumer patterns is machine learning. Using the e-commerce platform, it is very easy and efficient to track and analyse the purchase pattern of customers. This will help the business to optimise the platform for the potential customers as well as for the existing ones.

Because they can analyse and comprehend patterns of consumer behaviour, machine learning algorithms are being used more and more in e-commerce. Customer demographic data, including age, gender, salary, and spending score based on prior purchases, is provided by the mall data set. Using this data, it is intended to group customers according to their characteristics in order to create an advertising strategy that targets demographics.

To group together comparable data points in machine learning, clustering methods are frequently utilised. Based on the demographics and purchasing habits of the customers.

We have a ‘Mall’ data set containing the customer’s demographic data (i.e. . Spending Score is based on their previous purchases from 1-100. Now our project will make clusters of these customers based on their details. Eventually we make the advertising strategy accordingly.

**Existing system: -** Companie**s** are increasingly using machine learning algorithms to analyse client data and spot patterns in an effort to address these issues. These algorithms enable businesses to cluster their clients according to their preferences, buying patterns, demographics, and other pertinent characteristics. This increases the likelihood of success by enabling businesses to develop focused marketing campaigns and offerings that are suited to client groups.

Clustering methods may be used to divide clients into several categories in the Mall dataset stated previously based on their demographics and purchasing patterns. These categories may then be utilised to create distinctive advertising campaigns that will resonate with various client segments.

consumers with high salaries and high spending scores, for instance, can be grouped together and targeted with premium offers and goods, whilst consumers with low salaries and low spending scores might be targeted with more budget-friendly goods and offers. This targeted strategy may result in more successful marketing initiatives and improved customer satisfaction, both of which may boost sales and profitability for the company.

**Background / Related work**

The paragraphs explain the idea of clustering and how it is used in business, particularly the e-commerce sector. Exploratory data analysis frequently employs the clustering approach, which involves assembling related data points into clusters or subgroups. Businesses may leverage these insights on customer behaviour and purchase trends to better serve both current and future customers through platform optimisation. By concentrating on several target audience segments, the overall objective is to increase the effectiveness of the advertising campaign.

Due to heightened rivalry and the necessity for businesses to maximise returns on investment, clustering has taken on major relevance in recent years. Clustering is an effective method for analysing the massive volumes of data generated by e-commerce platforms and extracting valuable insights. Businesses may increase customer retention, more precisely target new consumers, and enhance their advertising campaigns by better understanding client behaviour.

Several industries, including finance, healthcare, and marketing, can benefit from clustering. Clustering can be used in healthcare to classify people with similar medical problems and in finance to spot fraud trends. Customers may be grouped using clustering in marketing based on their characteristics and purchase patterns, which can then be utilised to create tailored marketing campaigns.

Overall, clustering is a crucial tool for exploratory data analysis, and its use in business has completely changed how businesses approach advertising and customer targeting.

# Methodology

# Multiple steps make up the methodology used for this project. The dataset must first be loaded and go through the necessary cleaning and transformations as part of the pre-processing procedure. To get insights and spot trends, the second phase is studying and visualising the data. Applying clustering algorithms to the data is the third stage in identifying consumer categories based on demographics and spending patterns. Because of the developed tailored advertising strategies for each of the identified client categories, there should be a better use of advertising resources and return on investment.

K-means, a well-liked and often used clustering method, was the particular clustering algorithm employed in this research. Each data point is repeatedly allocated to the nearest centroid (representative point) via K-means, which then updates the centroids in accordance with the mean of the assigned points. The process goes on until convergence, or when the centroids stop moving. Before starting the procedure, the hyperparameter k must be supplied. It may be found using a number of techniques, including the elbow approach or the silhouette score.

# Proposed System - The main algorithm used in the project are:

Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as Euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific.

Clustering analysis can be done on the basis of features where we try to find subgroups of samples based on features or on the basis of samples where we try to find subgroups of features based on samples. We’ll cover here clustering based on features. Clustering is used in market segmentation; where we try to find customers that are similar to each other whether in terms of behaviours or attributes, image segmentation/compression; where we try to group similar regions together, document clustering based on topics, etc.

Unlike supervised learning, clustering is considered an unsupervised learning method since we don’t have the ground truth to compare the output of the clustering algorithm to the true labels to evaluate its performance. We only want to try to investigate the structure of the data by grouping the data points into distinct subgroups.

We shall now understand **K-means algorithm,** which is considered as one of the most used clustering algorithms due to its simplicity.

**K-means Algorithm**

K-means algorithm is an iterative algorithm that tries to partition the dataset into *K-*pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The lesser variation we have within these clusters, the more homogeneous (similar) the data points are within the same cluster.

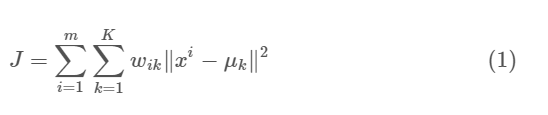
The way k-means algorithm works is as follows:

1. Specify number of clusters *K*.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K*data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’t changing.

* Compute the sum of the squared distance between data points and all centroids.
* Assign each data point to the closest cluster (centroid).
* Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

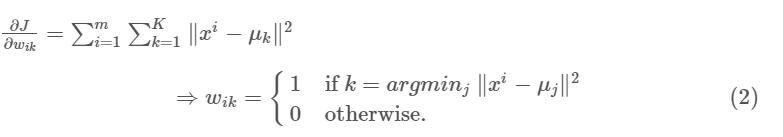
The approach k-means follows to solve the problem is called **Expectation-Maximization**. The E-step is assigning the data points to the closest cluster. The M-step is computing the centroid of each cluster. Below is a breakdown of how we can solve it mathematically.

The objective function is:



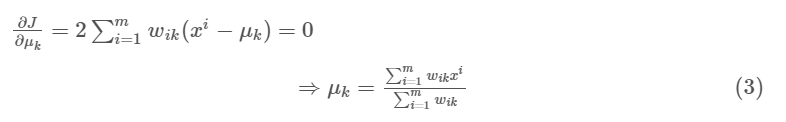
where wik=1 for data point xi if it belongs to cluster *k*; otherwise, wik=0. Also, μk is the centroid of xi’s cluster.

It’s a minimization problem of two parts. We first minimize J w.r.t. wik and treat μk fixed. Then we minimize J w.r.t. μk and treat wik fixed. Technically speaking, we differentiate J w.r.t. wik first and update cluster assignments (*E-step*). Then we differentiate J w.r.t. μk and re-compute the centroids after the cluster assignments from the previous step (*M-step*). Therefore, E-step is:



In other words, assign the data point xi to the closest cluster judged by its sum of squared distance from the cluster's centroid.

And M-step is:

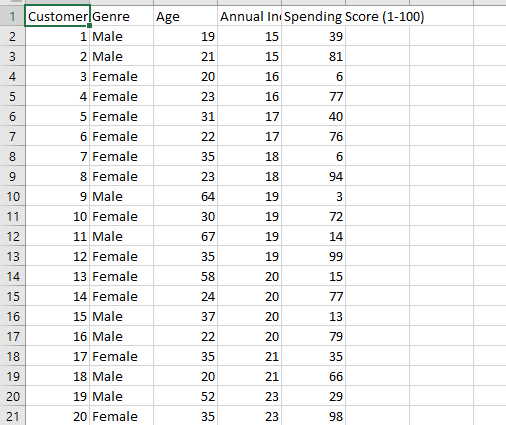


Which translates to re-computing the centroid of each cluster to reflect the new assignments.

Few things to note here:

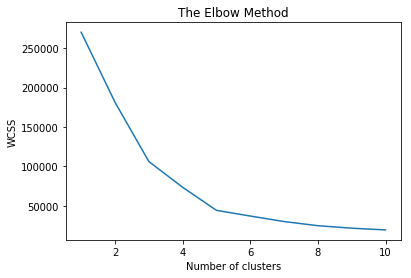
* Since clustering algorithms including k means use distance-based measurements to determine the similarity between data points, it’s recommended to standardize the data to have a mean of zero and a standard deviation of one since almost always the features in any dataset would have different units of measurements such as age vs income.
* Given k-means iterative nature and the random initialization of centroids at the start of the algorithm, different initializations may lead to different clusters since k-means algorithm may be stuck *in a local optimum and may not converge to global optimum*. Therefore, it’s recommended to run the algorithm using different initializations of centroids and pick the results of the run that yielded the lower sum of squared distance.
* Assignment of examples isn’t changing is the same thing as no change in within-cluster variation:



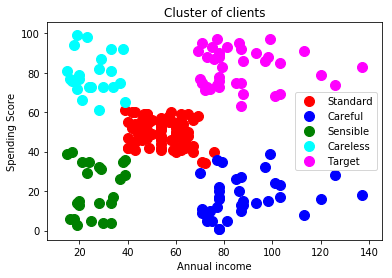


This is the Sample of the dataset we are using for this project.

(MALL DATASET)



Elbow Method : As it is elaborated above a method in K-Means algorithm to find the appropriate number of clusters to be formed.



Finally, the end result showing that there are total five clusters formed from this dataset.

**According to K-Means the clusters formed are**:

**1.Careful**-These type of customers generally spends lower amount while shopping . So, normal offers should be available to them.

**2.Standard**-Standard customers spends a moderate amount so attractive offers should be provided to these customers as their spending rate oftenly depends on the offers available too.

**3.Target-** The important category of customers as they are the major buyers and if offers is available they spends a tremendous amount in the mall. So, in the main the mall should focus on them.

**4.Careless**- Some normal offers should be available to them as they are not the permanent ones and the spending score can decrease at any time.

**5.Sensible** – Regular offers should be available for them.

**Project Scope:** Using this system, we design an optimized advertising strategy to increase the overall turnover of the company.