REPORT FOR DATA ANALYSIS ON CUSTOMER SEGMENTATION

ABSTRACT

A lot of customers buy products from the mall and to generate more revenue for the mall, the authorities need to attract these customers and for this large amount of capital is required. After the advertisement, the output is only around 30-40%. Hence customer segmentation comes into the picture.

Customer Segmentation is a popular application of unsupervised learning and by using this technique we'll only focus on the potential customers or target customer's (customers whose probability of buying the product is very high) which is also termed as Market Basket Analysis. With this technique, the output will drastically increase to 90-95%.

Our project aims to build clusters of customers based on their Spending Score and Annual Income. The algorithm used in this project is K-means and Clustering.

Keywords: Target Customers, Clusters, Unsupervised Learning, K-Means, Hierarchical Clustering Segmentation, Market Basket Analysis

CONTENT

SL NO	TOPICS	PAGE NO
1	FIGURES	4 - 9
2	TABLES	10
3	INTRODUCTION	11 - 12
4	LITERATURE SURVEY	13 - 14
5	DATA DESCRIPTION	15 - 16
6	MODELLING	17 - 18
7	DISCUSSION	19
8	CONCLUSION	20 - 21

FIGURES

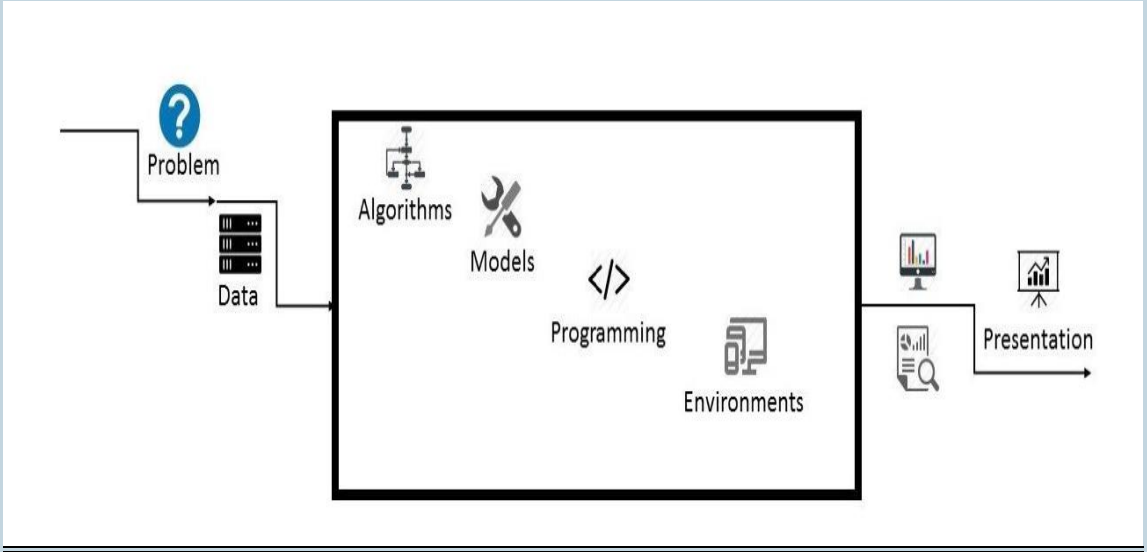


Fig-1: Flowchart of Proposed System

Heatmap is used for determining the correlation between dataset attributes. The data of a given dataset can be easily represented graphically by using a Heatmap. It uses a colour system to represent the correlation among different attributes. It is a data visualization library (Seaborn) element.

Heatmap colour encoded matrix can be described as lower the intensity of the colour of an attribute related to the target variable, higher is the dependency of target and attribute variables.

Based on the Mall_Customers.csv Dataset the heatmap obtained gives output as Figure-2. The observation based on the heatmap is the attributes and their correlation.

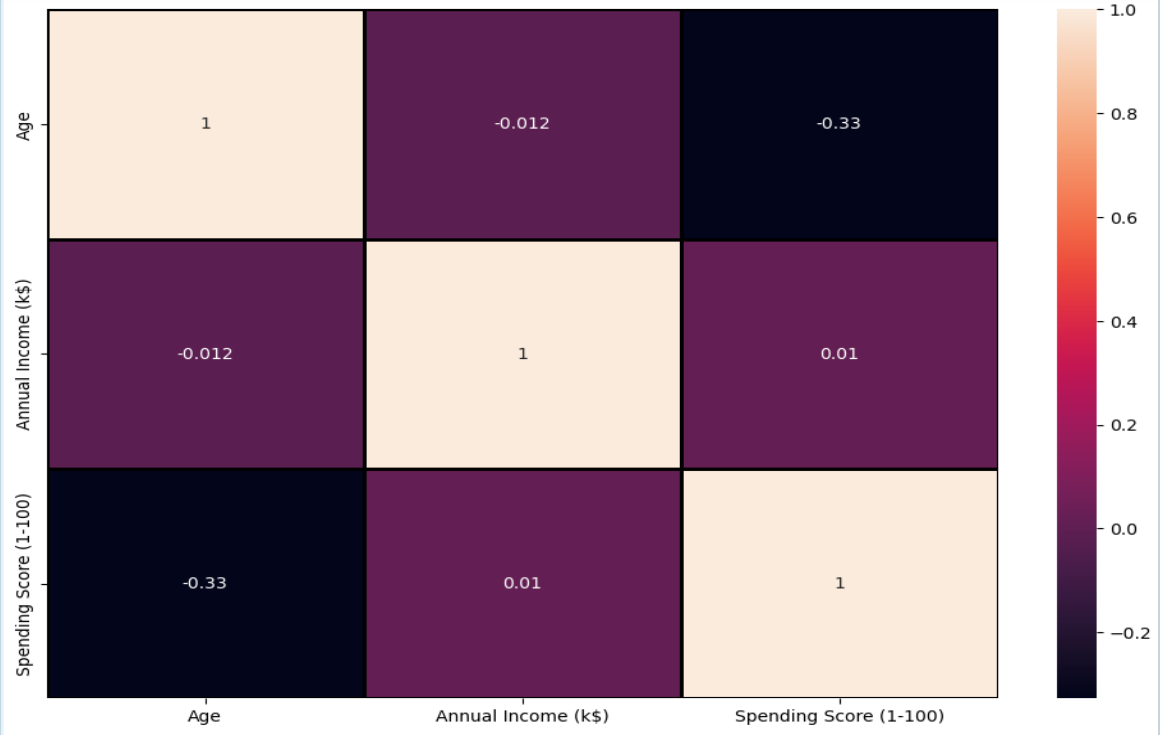


Fig-2: Heatmap for correlation between attributes

Gender Plot Analysis

From the Count plot, it is observed that the number of Female customers is more than the total number of Male customers.

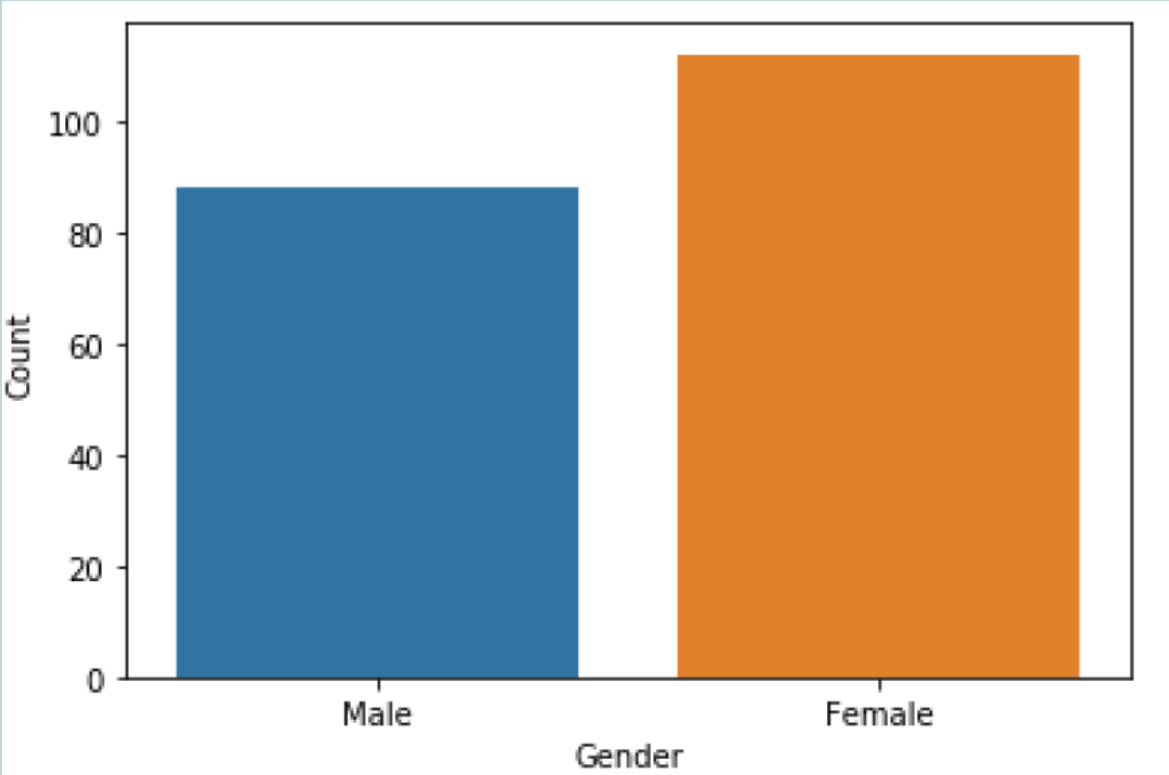


Fig-3: Count Plot for Gender

Age Plot Analysis

From the Histogram it is evident that there are 3 age groups that are more frequently shop at the mall, they are: 15-22 years, 30-40 years, and 45-50 years.

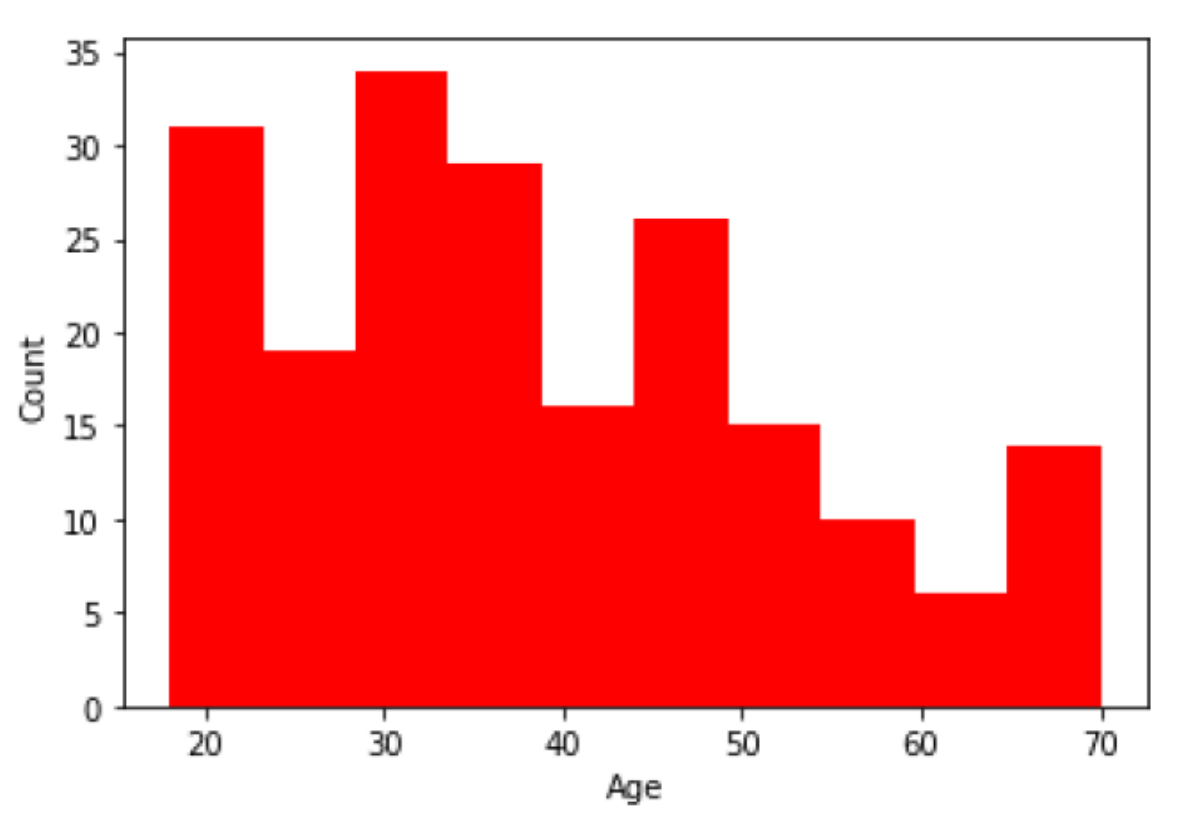


Fig-4: Count Plot for Age

Scatter Plot For Age Vs Spending Score Analysis

From the Age Vs Spending Score plot we observe that customers whose spending score is more than 65 have their Age in the range of 15-42 years. Also from the Scatter plot it is observed that customers whose spending score is more than 65 consists of more Females than Males.

The customers having average spending score ie: in the range of 40-60 consists of the age group of the range 15-75 years and the count of males and females in this age group is also approximately the same.

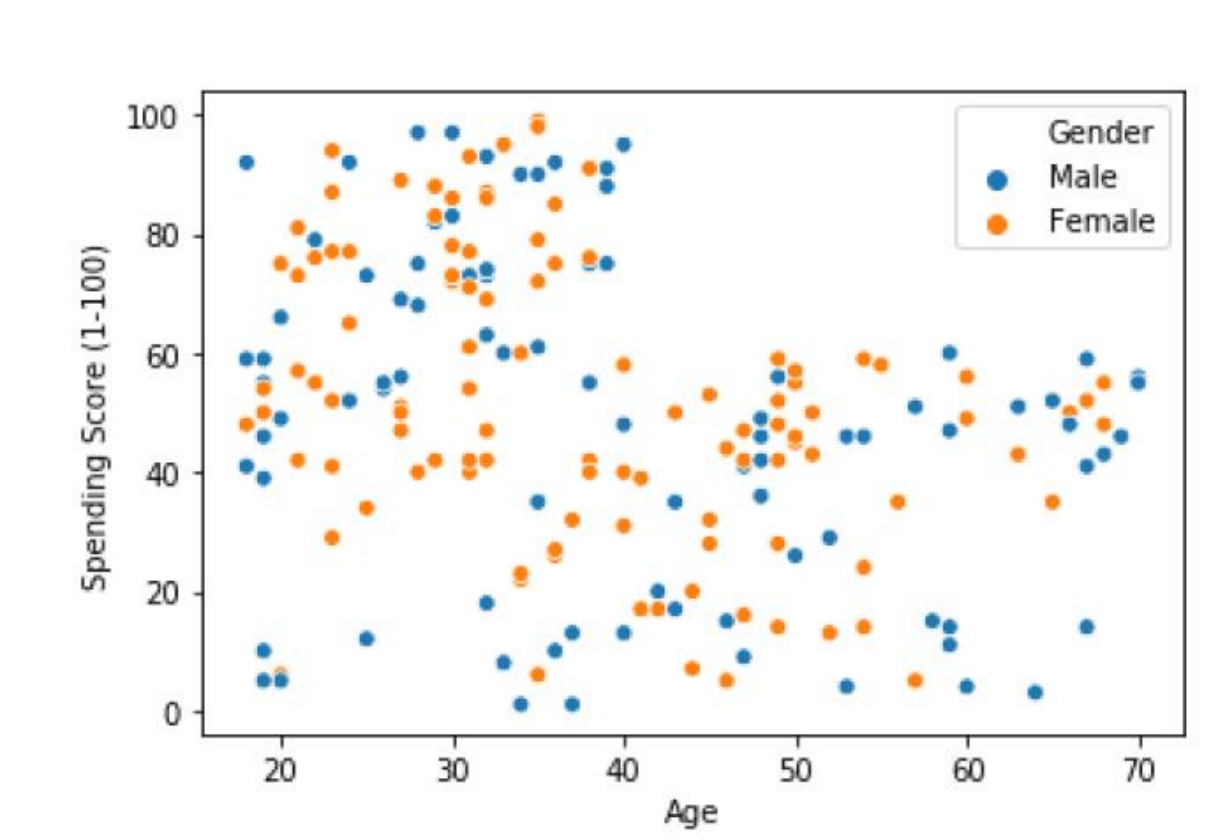


Fig-5: Scatter Plot For Age Vs Spending Score

Annual Income Vs Spending Score Analysis

We observe that there are 5 clusters and can be categorized as:

a. High Income, High Spending Score (Top Right Cluster)

b. High Income, Low Spending Score (Bottom Right Cluster)

c. Average Income, Average Spending Score (Centre Cluster)

d. Low Income, High Spending Score (Top Left Cluster)

e. Low Income, Low Spending Score (Bottom Left Cluster)

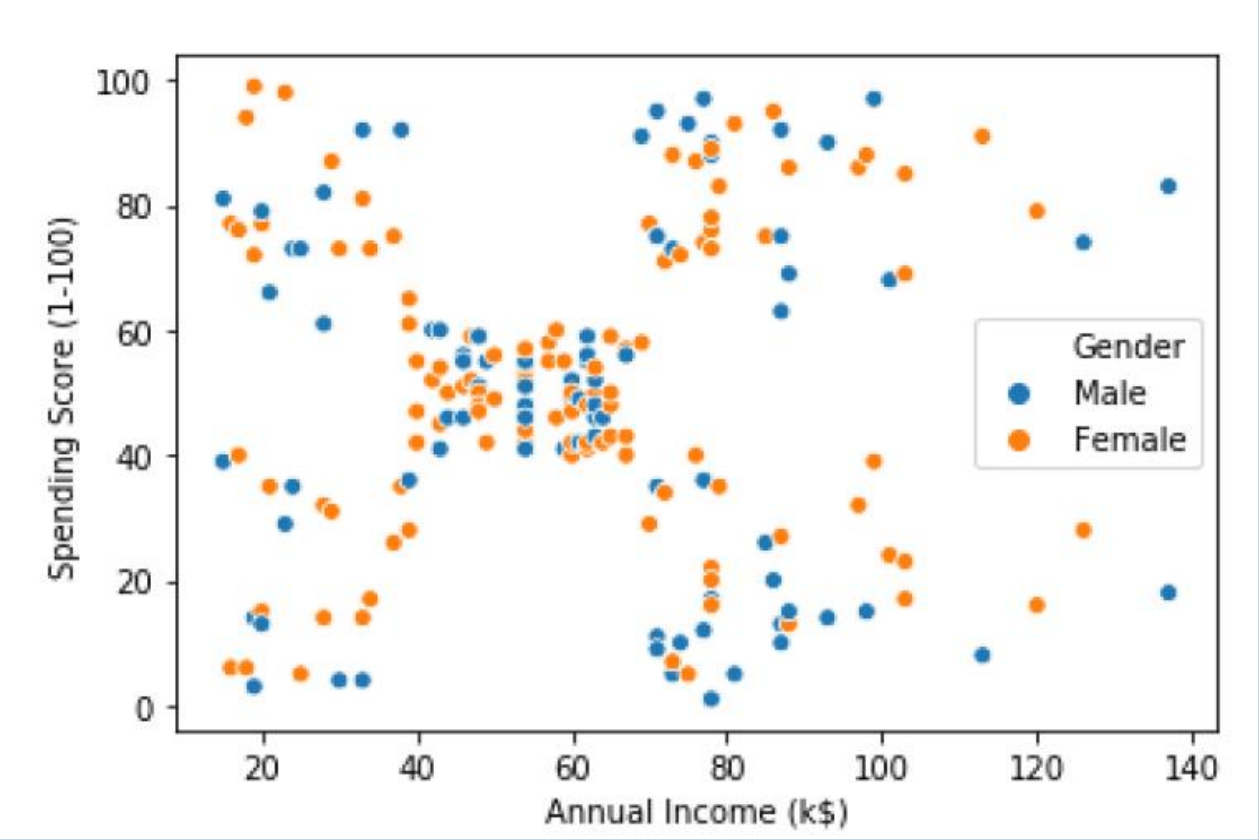


Fig-6: Scatter Plot For Annual Income Vs Spending Score

TABLES

TABLE-1: DATASET DEFINITION

SR NO	VARIABLE	DEFINITION
1	CUSTOMER ID	UNIQUE CUSTOMER ID
2	GENDER	SEX OF CUSTOMER
3	AGE	CUSTOMER AGE
4	ANNUAL INCOME (k$)	ANNUAL INCOME OF THE CUSTOMER
5	SPENDING SCORE (1-100)	AMOUNT SPENT BY THE CUSTOMER

INTRODUCTION

To make predictions and find the clusters of potential customers of the mall and thus find appropriate measures to increase the revenue of the mall is one of the prevailing applications of unsupervised learning.

For example, a group of customers have high income but their spending score (amount spent in the mall) is low so from the analysis we can convert such type of customers into potential customers (whose spending score is high) by using strategies like better advertising, accepting feedback and improving the quality of products.

To identify such customers, this project analyses and forms clusters based on different criteria which are discussed in the further sections.

They partition the data objects into groups or clusters, so that objects within a cluster are similar to one another and dissimilar to objects in other clusters.

Customer Segmentation is the process of division of customer base into several groups called as customer segments such that each customer segment consists of customers who have similar characteristics. The segmentation is based on the similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits.

The customer segmentation has the importance as it includes, the ability to modify the programs of market so that it is suitable to each of the customer segment, support in business decision; identification of products associated with each customer segment and to manage the demand and supply of that product; identifying and targeting the potential customer base,

and predicting customer defection, providing directions in finding the solutions.

The thrust of this paper is to identify customer segments using the data mining approach, using the partitioning algorithm called as K-means clustering algorithm. The elbow method determines the optimal clusters.

LITERATURE SURVEY

Customer Segmentation

Over the years, as there is very strong competition in the business world, the organizations have to enhance their profits and business by satisfying the demands of their customers and attract new customers according to their needs. The identification of customers and satisfying the demands of each customer is a very complex task. This is because customers may be different according to their demands, desires, preferences and so on. Instead of “one-size-fits-all” approach, customer segmentation clusters the customers into groups sharing the same properties or behavioural characteristics. According to, [1] customer segmentation is a strategy of dividing the market into homogenous groups. The data used in customer segmentation technique that divides the customers into groups depends on various factors like, demographical conditions, data geographical conditions and economic conditions as well as behavioural patterns. The customer segmentation technique allows the business to make better use of their marketing budgets, gain a competitive edge over their rival companies, demonstrating the better knowledge of the needs of the customer. It also helps an organization in, increasing their marketing efficiency, plan the marketing budget, determining new market opportunities, making better brand strategy, identifying customer’s retention.

According to

[1]Decision makers use many variables to segment customers. Demographic variables such as age, gender, family, education level and income are the easiest and common variables for

segmentation. Socio- cultural, geographic, psychographic and behavioural variables are the other major variables that are used for segmentation.

[2] Presented various clustering algorithms taking into account the characteristics of Big Data such as size, noise, dimensionality, algorithm calculations, cluster shape and presented a brief overview of the various clustering algorithms grouped under partitioning, hierarchical, density, grid-based and model-based algorithms.

[3] Explored the necessity of segmentation of the customers using clustering algorithms as the core functionality of CRM. The mostly used K-Means and Hierarchical Clustering were studied and the advantages and disadvantages of these techniques were highlighted. At last, the idea of creating a hybrid approach is addressed by integrating the above two strategies with the potential to surpass the individual designs.

DATA DESCRIPTION

The dataset name is ‘Mall_Customers.csv’ consists of 5 columns which are Customer ID , Gender , Age , Annual Income (k$) , Spending Score (1-100) where Gender is a categorical value and rest all features are numeric.

Customer segmentation is routinely utilized for dividing clients into bunches in light of orientation, age, geographic area and spending examples to give some examples. Notwithstanding, in this report a more surprising customer clustering approach in view of client conduct in the Pick E-commercial center will be assessed. This cycle did not depend on any previous relations or rules. Rather, the actual information uncovers potential similitudes between clients. This issue is regularly alluded to as data over-burden and can be addressed by giving particular article streams to every client contingent upon his/her inclinations.

Customer segmentation is right now performed by handling client data set, for example segment information or buy history. A few scientists talk about the client division strategy on their papers, who utilized a few factors to perform client division, specifically exchange variable, item factor, geographic variable, side interests variable and page saw variable to examine client division techniques for Business Rule, Supervised Clustering, Unsupervised Clustering, Customer Profiling, RFM Cell Classification Grouping, Customer Likeness Clustering and Purchase Affinity Clustering. A portion of these strategies have closeness. Different scientists examine the execution of client division. This paper will characterize client division techniques in view of information handling.

This dataset is created only for the learning purpose of the customer segmentation concepts also known as **Market Basket Analysis**. I will demonstrate this by using unsupervised ML Techniques i.e. **Clustering** and **K-Means Clustering Algorithm** in the simplest form. You are owing a supermarket mall and through membership cards you have some basic data about your customers like customer ID, age, gender, annual income and spending score.

Spending score is something you assign to the customer based on your defined parameter like customer behaviour and purchasing data.

MODELLING

A. Clustering.

Clustering is one of the most common methods used in exploring data to obtain a clear understanding of the data structure. It can be characterized as the task of finding the subtitles and subgroups in the complete dataset. Similar data is clustered in many subgroups. A cluster refers to a collection of aggregated data points due to some similarities. Clustering is used in Market basket analysis used to segment the customers based on their behaviours and transactions.

Clustering algorithms generates clusters such that within the clusters are similar based on some characteristics. Similarity is defined in terms of how close the objects are in space.

B. K-means Clustering Algorithm.

K-means algorithm in one of the most popular centroid based algorithm. Suppose data set, D, contains n objects in space. Partitioning methods distribute the objects in D into k clusters, C1,...,Ck , that is, $Ci \subset D$ and $Ci \cap Cj = \emptyset$ for $(1 \leq i, j \leq k)$. A centroid-based partitioning technique uses the centroid of a cluster, Ci, to represent that cluster. Conceptually, the centroid of a cluster is its center point. The difference between an object $p \in Ci$ and ci, the representative of the cluster, is measured by dist (p, ci), where dist (x, y) is the Euclidean distance between two points x and y.

K Means Clustering is the most common and simplest Machine learning algorithm and it follows an iterative approach which attempts to partition the dataset into different “k” number of predefined and non-overlapping subgroups where each data point belongs to only one subgroup according to their similar qualities.

The k-means algorithm for partitioning, where each cluster’s center is represented by the mean value of the objects in the cluster.

Input: k: the number of clusters, D: a data set containing n objects.

Output: A set of k clusters. Method: (1) arbitrarily choose k objects from D as the initial cluster canters; (2) repeat (3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster; (4) update the cluster means, that is, calculate the mean value of the objects for each cluster; (5) until no change.

K-means algorithm is used in this project to analyse and form clusters of customers based on their income and spending score features.

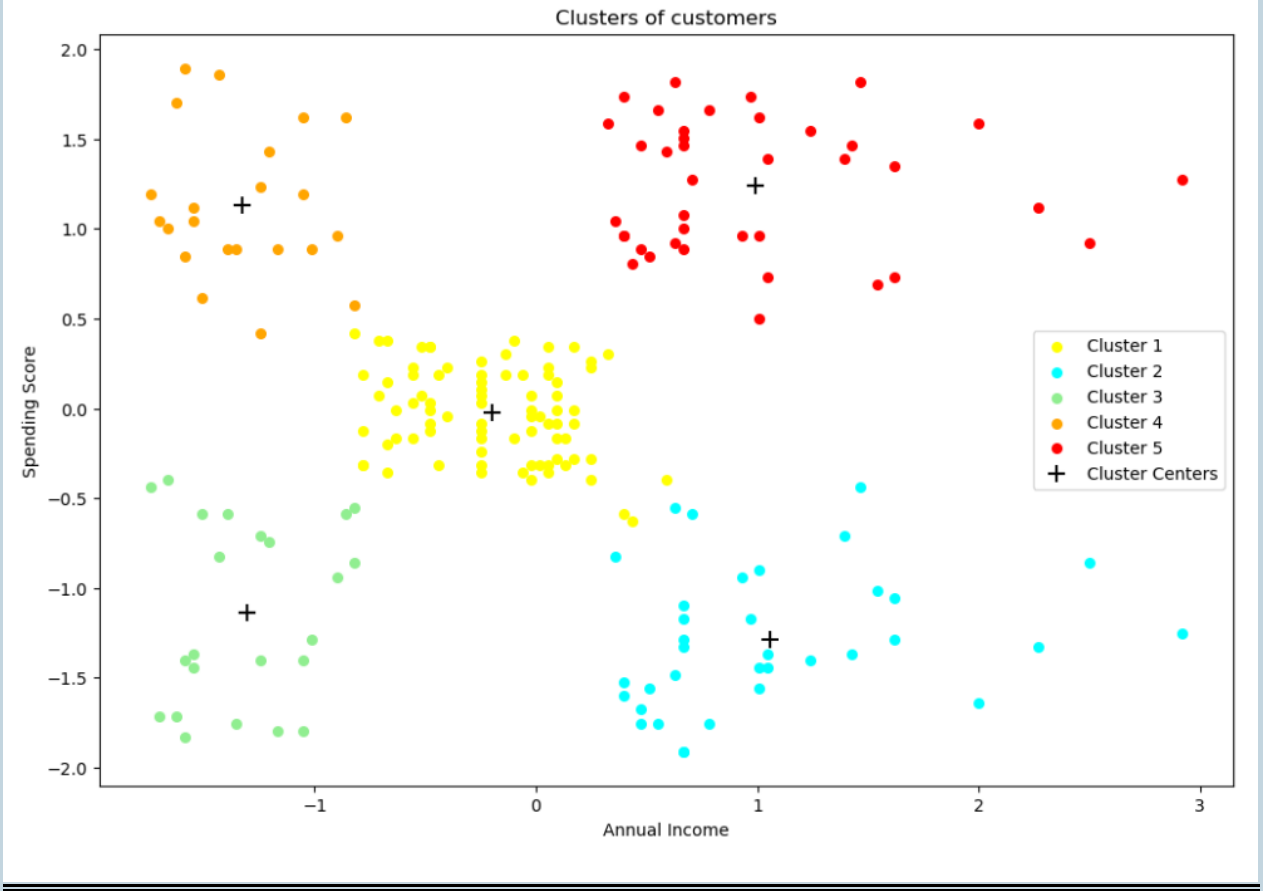
K-means model is used and is hyper tuned parameters like *n_clusters=5* using elbow method to find the optimal number of clusters also *init=’k-means++’* to avoid random initialization traps.

DISCUSSION

By the customer segmentation method the project is evaluated successfully. The accessible informative elements in the data set is thing perspectives, preferences and discussions, but the underlying rendition of the framework assessed in this report just utilizes sees. During the final phases of the execution a small variant issues were risen, but no critical enhancements were noticed. All things considered, the spotlight during this task was on the bunching investigation, the pre-processing phase of this undertaking could be improved by joining preferences and discussion to the appraisals computations utilizing some weighting of different elements.

CONCLUSION

For this project, the K-means algorithm is used and performs the best (with n_clusters = 5 and init = 'kmeans++'). After the clustering algorithm is applied to the dataset, this is the output.



Clustering Analysis

- **High Income, High Spending Score (Cluster 5) - Target these customers by sending new product alerts which would lead to an increase in the revenue collected by the mall as they are loyal customers.**
- **High Income, Low Spending Score (Cluster 2) - Target these customers by asking the feedback and**

advertising the product in a better way to convert them into Cluster 5 customers.

- Average Income, Average Spending Score (Cluster 1) - May or may not target these groups of customers based on the policy of the mall.
- Low Income, High Spending Score (Cluster 4) - Can target these set of customers by providing them with Low-cost EMI's, etc.
- Low Income, Low Spending Score (Cluster 3) - Don't target these customers since they have less income and need to save money.