ABSTRACT

Looking around and finding that most companies are now data-driven. They make strategic decisions based on data analysis, enabling them to examine and organize their data for better service. There has always been a lot of competition in the market as to who can provide the best customer experience, attract new customers based on their needs, and satisfy their demands, enhancing their profit and growth. However, this is not very easy and calls for various data mining techniques and algorithms.

Machine learning can help them target potential customers. The algorithms deep dive into the data pool to extract hidden treasures and patterns that can bring wonderous profits to many organizations and better decision making. Customer segmentation is one such beautiful concept. Customer segmentation finds its use in many sectors. For example, in Netflix, it can be used as are commendation system to find a group of similar users and use it for filtering, categorizing, or recommending movies. Banks or insurance companies use it for fraud detection or to evaluate certain insurance risks to segmented customers.

Will be using Customer Segmentation in the retail industry, a Mall, to segment customers into various groups and target potential. The industry can then work towards attractive ideas to sell products and services inclined towards these specific customers.

Keywords: Clustering, Elbow Method, K-Means Algorithm, Customer Segmentation, Visualization.

INDEX

LIST OF FIGURES

Chapter 1: Introduction

Nowadays the competition is vast and lot of technologies came into account for effective growth and revenue generation. For every business the most important component is data. With the help of grouped or ungrouped data, we can perform some operations to find customer interests.

Data mining helpful to extract data from the database in a human readable format. But, we may not known the actual beneficiaries in the whole dataset. Customer Segmentation is useful to divide the large data from dataset into several groups based on their age, demographics, spent, income, gender, etc. These groups are also known as clusters. By this, we can get to know that, which product got huge number of sales and which age group are purchasing etc. And, we can supply that product much for better revenue generation.

Initially we are going to take the old data. As we know that old is gold so, by using the old data we are going to apply K-means clustering algorithm and we have to find the number of clusters first. So, at lastly, we have to visualize the data. One can easily find the potential group of data while observing that visualization.

The goal 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

1.1 Problem Statement

Customer Segmentation is the best application of unsupervised learning. Using clustering, identify segments of customers in the dataset to target the potential user base. They divide customers into various groups according to common characteristics like gender, age, interest, and spending habits so they can market to each group effectively. Use K-Means Clustering and also visualize the gender and age distributions. Then analyze their annual income and spending scores. As it describes about how we can divide the customers based on their similar characteristics according to their needs by using k-means clustering which is a classification of unsupervised machine learning.

Chapter 2: Existing System

The existing method is storing customer data through paperwork and computer software (digital data) is increasing day by day. At end of the day they will analyse their data as how many things are sold or actual customer count etc. By analysing the collected data they got to know who is beneficial to their business and increase their sales. It requires more time and more paperwork. Also, it is not much effective solution to find the desired customers data.

2.1 Disadvantages of the Existing System

- Data Points are overlapped
- The data does not hold some patterns.
- Data clusters were Inappropriate for accurate considerations

Chapter 3: Proposed System

To overcome the traditional method i.e paper work and computerized digital data this new method will play vital role. As we collect a vast data day by day which requires more paperwork and time to do. As new technologies were emerging in today's world. Machine Learning which is powerful innovation which is used to predict the final outcome which has many algorithms. So for our problem statement we will use K-Means Clustering which groups the data into different clusters based on their similar characteristics. And then we will visualize the data.

3.1 Advantages of the Proposed System

- Determine appropriate product pricing
- Develop customized marketing campaigns
- Design an optimal distribution strategy
- Choose specific product features for deployment
- Prioritize new product development efforts.

Chapter 4: System Architecture

Initially we will see the dataset and then we will perform exploratory data analysis which deals with the missing data, duplicates values and null values. And then we will deploy our algorithm k-means clustering which is unsupervised learning in machine learning.

As in order to find the no of clusters we use elbow method where distance will be calculate through randomly chosen centres and repeat it until there is no change in cluster centres. Thereafter we will analyse the data through data visualization. Finally we will get the outcome.

4.1 Data Flow Diagram

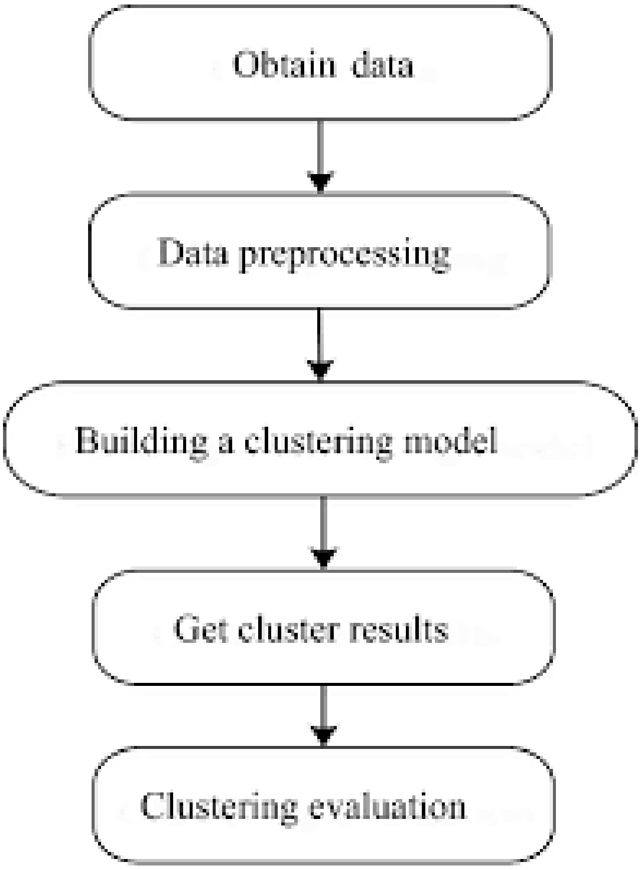


Figure 4.1 Data Flow Diagram

4.2 Algorithm

4.2.1 K-Means Clustering

- K Means algorithm in an iterative algorithm that tries to partition the dataset into K predefined distinct non overlapping sub groups which are called as cluster.
- Here K is the total no of clusters.
- Every point belongs to only one cluster.
- Clusters cannot overlap.

4.2.2 Steps of Algorithm

- Arbitrarily choose k objects from D as the initial cluster centers.
- Repeat.
- 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.
- Update the cluster means, i.e. calculate the mean value of the objects for each cluster.
- Until no change.

Chapter 5: Methodology

1. First of all we will import all the necessary libraries or modules (pandas, numpy, seaborn).
2. Then we will read dataset and anyalse whether it contains any null values, missing values and duplicate values. So we will fix them by dropping or fixing the value with their means, medians etc which is technically named as Data Preprocessing.
3. We will deploy our model algorithm K-Means Clustering, which divides the data into group of clusters based on similar characteristics. To find no.of clusters we will use elbow method.
4. Finally, we will visualize our data using matplot, which concludes the customers divided into groups who are similar to each other on their group.

Chapter 6: Implementation And Analysis

6.1 Customer segmentation

Customer segmentation is partitioning a customer database into group of people with similar characteristics. It is an application of unsupervised learning. It is a business strategy that allows targeting a specific group of customers and effectively allocate marketing resources. For such large datasets, we need an analytical approach like clustering to make customer segments.

There are four major ways of segmentation, i.e., geographical, economic, demographic, and behavioural patterns. In this project, we divide a Mall customer's dataset based on gender, age, income, spending habits, etc. We also visualize gender and age distributions and analyse their annual incomes and spending scores to target the potential user base. The method used is K-means clustering.

Language: Python

Supervised and Unsupervised Learning

Supervised learning is training a model with labelled data. There are two types regression and classification. Regression is the process of predicting a continuous value as opposed to predicting an absolute value in classification. In classification, the class is predefined and predict categorial classed labels. Classification approaches include decision trees, logistic regression to predict the default value of the new entry.In unsupervised learning, the model discovers information on its own. There is no prior information on the data or the outcomes of the analysis. Dimension reduction, density estimation, market basket analysis, and clustering are the most widely used unsupervised machine learning techniques. Generally, clustering is used for exploratory data analysis, summarisation, dimension reduction, outlier detection, and other such data mining tasks. In comparison to supervised learning, unsupervised learning has fewer models and fewer evaluation methods that can be used to ensure that the outcome of the model is accurate. As such, unsupervised learning creates a less controllable environment as the machine is creating outcomes for us

Clustering

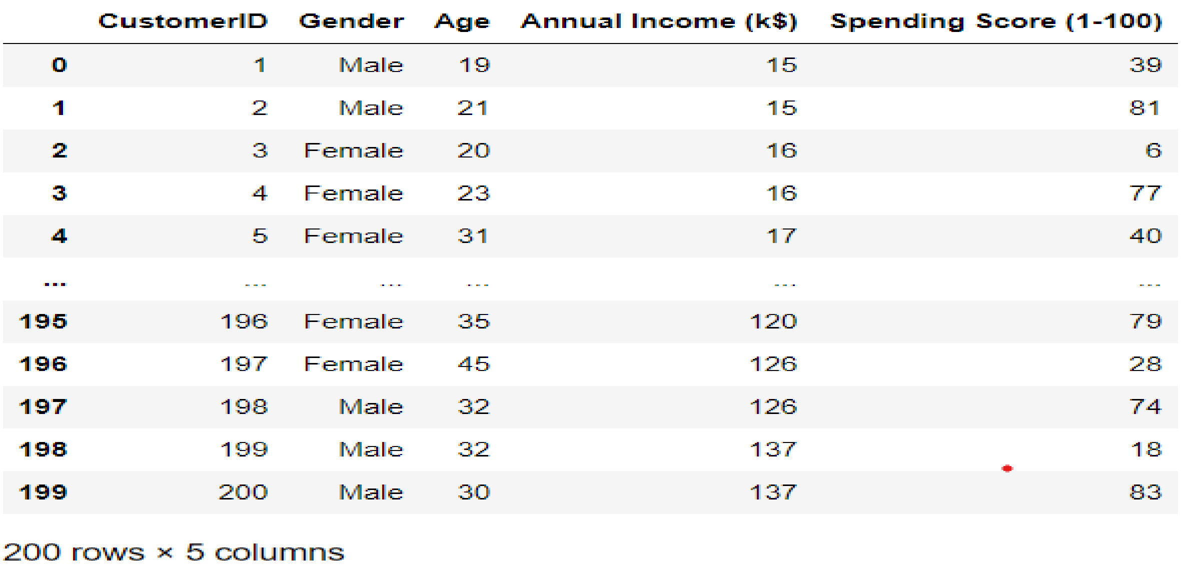
Clustering can group data unsupervised solely based on similarities to each other. It will partition customers into mutually exclusive groups aka clusters. Having the result would help understand and predict customer preferences and differences, thus making the company deliver personalised experiences for each group of customers.

Partition-based clustering is a group of clustering algorithms that produces sphere-like clusters, such as; K-Means, K-Medians or Fuzzy c-Means. These algorithms are relatively efficient and are used for medium to large- sized databases. Hierarchical clustering algorithms produce trees of clusters, such as agglomerative and divisive algorithms. This group of algorithms are very intuitive and are generally suitable for use with small-size datasets. Density-based clustering algorithms produce arbitrary-shaped clusters. They are outstanding when dealing with spatial clusters or noise in the data set, for example, the DB scan algorithm.

Figure 6.2 Overview of a Dataset

6.2 Overview of a Dataset

This is a mall customer segmentation data which contains 5 columns and 200 rows.



	CustomerID	Gender	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
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

200 rows × 5 columns

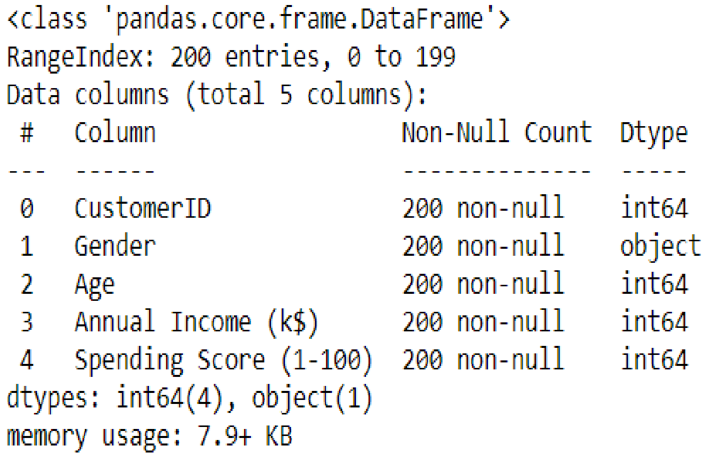
Figure 6.2 Overview of a Dataset

6.3 Exploratory Data Analysis

It deals with the data preprocessing, whether it contains any missing values or null values. There after we will see the information and description of the dataset.

6.3..1 Information of the dataset

#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   Gender                  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
```

Figure 6.3.1 Information of Dataset

As here it overview the information of the data. And it gives it doesn’t contain any null values.

6.3.2 visualisation

It consists of graphs visualisation between no.of consumers or people and Age,Annual income and spending scores which describes the distribution plots from various segments.

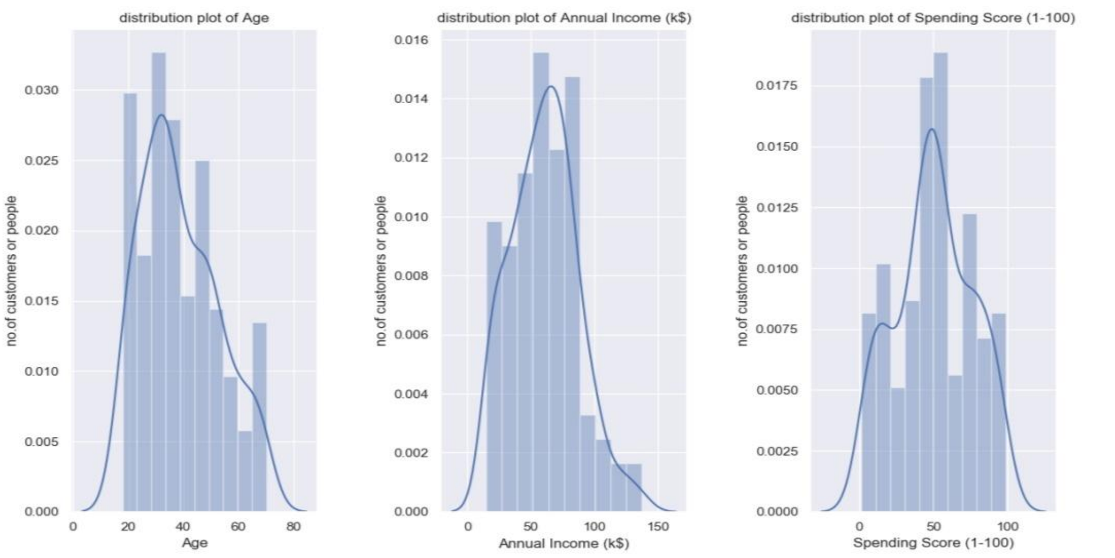
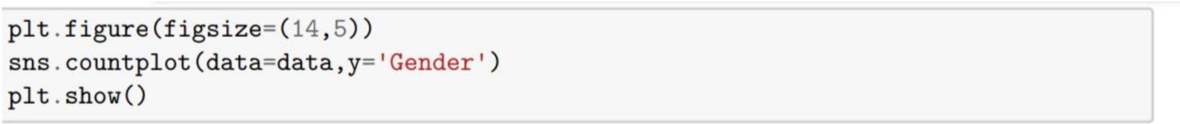


Figure 6.3.2 visualisation plot

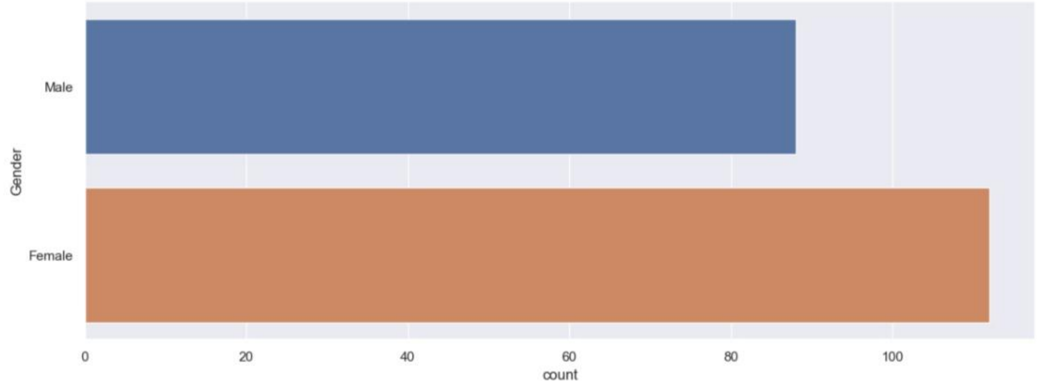
6.3.3 Gender plot analysis and comparissions between genders

Here it overview the gender analysis

we label the x-axis as Gender and y-axix as Count and we plot it by using barplot.From the plot we will con clued that the there are more female customers than the male customers i.e female customers are more than 100 whereas male customers are nearly 80.



```
plt.figure(figsize=(14,5))
sns.countplot(data=data,y='Gender')
plt.show()
```



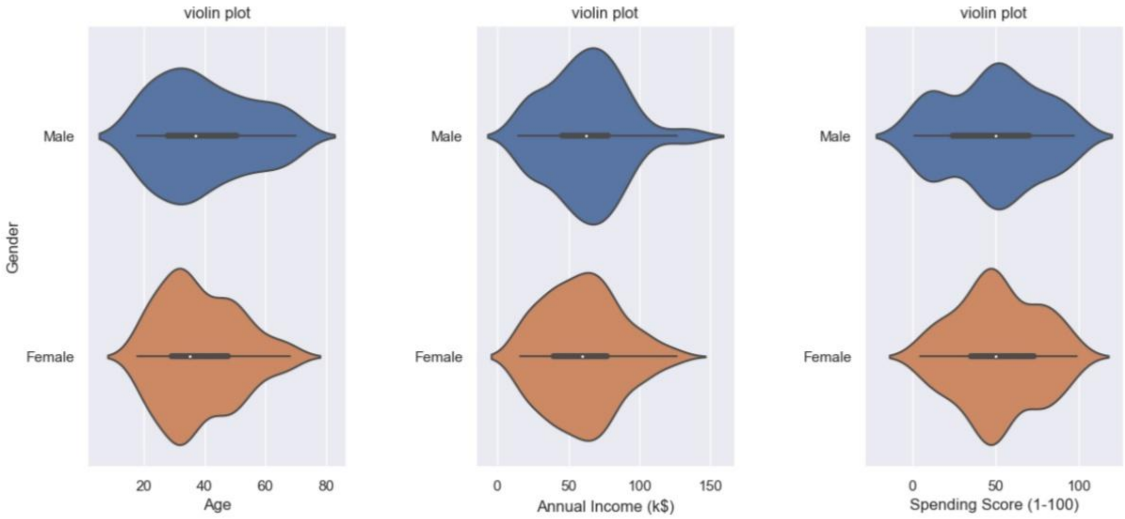
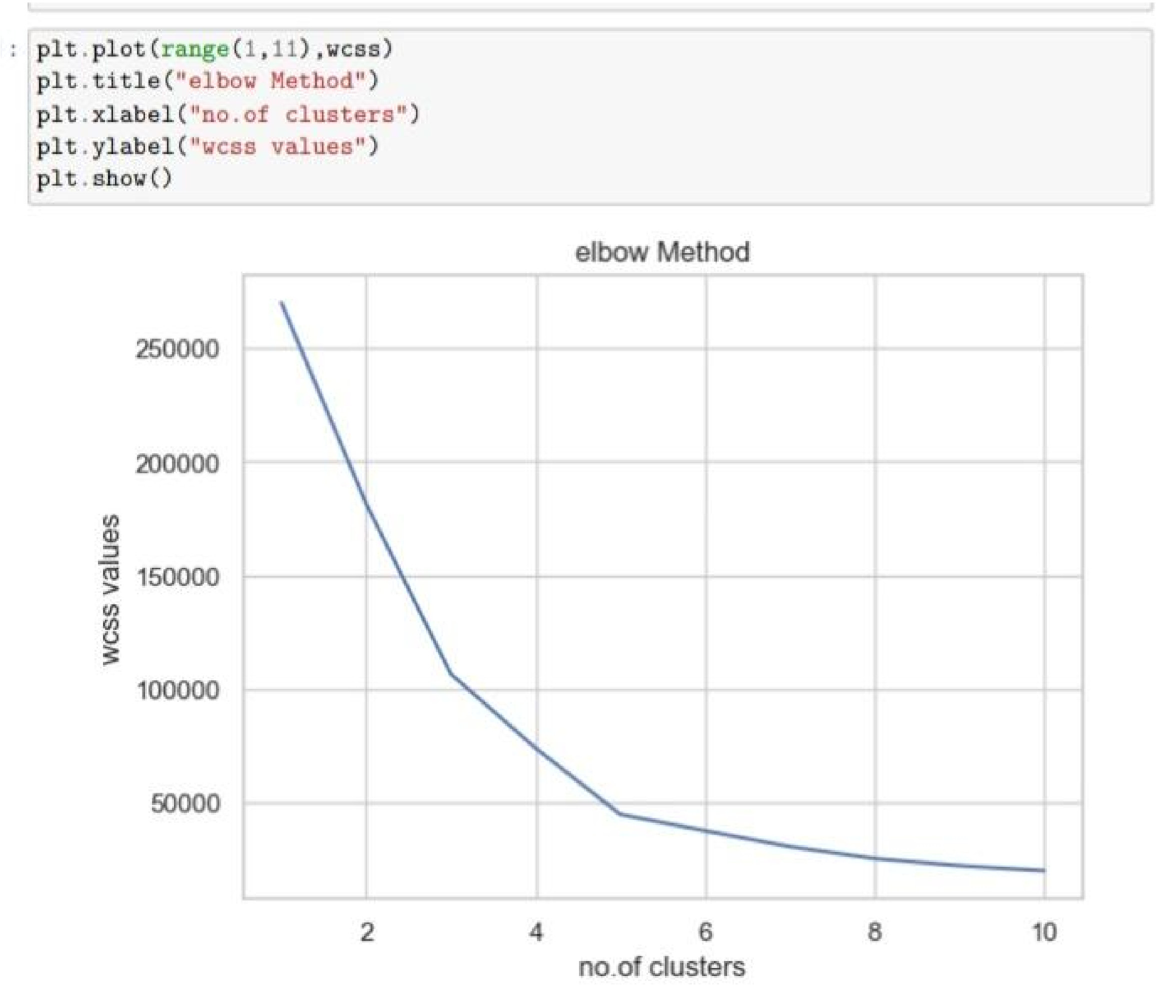


Figure 6.3.3 Gender and comparission plots

6.4 Elbow Method

The elbow method is based on the observation that increasing the number of clusters can help to reduce the sum of within-cluster variance of each cluster. This is because having more clusters allows one to capture finer groups of data objects that are more similar to each other. To define the optimal clusters, Firstly, we use the clustering algorithm for various values of k. This is done by ranging k from 1 to 10 clusters. Then we calculate the total intra-cluster sum of square. Then, we proceed to plot intra-cluster sum of square based on the number of clusters. The plot denotes the approximate number of clusters required in our model. The optimum clusters can be found from the graph where there is a bend in the graph.



```
plt.plot(range(1,11),wcss)
plt.title("elbow Method")
plt.xlabel("no.of clusters")
plt.ylabel("wcss values")
plt.show()
```

Figure 6.4 Elbow method

6.5 Fitting the Algorithm

As here we initialized the kmeans with 10 clusters by using for loop and we will fit it. There after we will predict the data and store it in wcss.

6.6 Visualization of clusters

Visualizing the clusters based on Annual Income and Spending Score of the customers. As here we plot a graph named as Clusters of Customers to visualize the data in terms of groups or cluster.

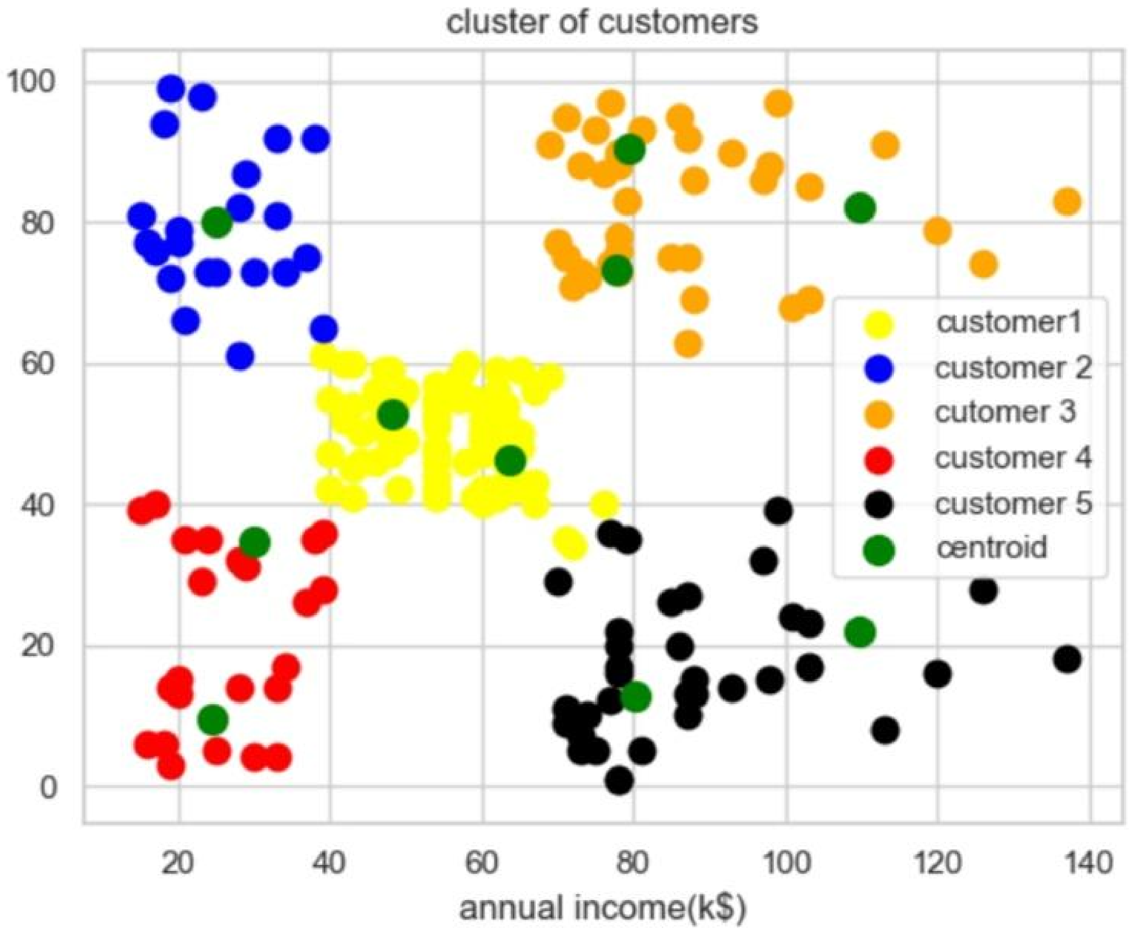


Figure 6.6 Visualization of clusters

From the above one we observed that the there are 5 clusters which are named as 1, 2, 3, 4and 5.to form this cluster a centroid is initiated from every cluster that evaluates and form the data point it into particular cluster.

The main formula is used to form clusters is

Euclid distance = $\sqrt{(x2 - x1) + (y2 - y1)}$

- Cluster 1 which is at centre, average annual income with average spending score.
- Cluster 2 which is at top left, lowest annual income with highest spending score.
- Cluster 3 which is at top right, highest annual income with highest spending score.
- Cluster 4 which is at bottom left, lowest annual income with lowest spending score.
- Cluster 5 which is at bottom right, high annual income with low spending score.

Chapter 7:. Conclusion

So we concluded that the ,

- The Highest income , high spending can be target these type of customers as they earn more money and spend as much as they want.
- Highest income, low spending can be target these type of customers by asking feedback and advertising the product in a better way.
- Average income, Average spending may or may not be beneficial to the mall owners of this type of customers.
- Low income, High spending can be target these type of customers by providing them with low-cost EMI's etc.
- Low income, Low spending don't target these type of customers because they earn a bit and spend some amount of money.
- So high income, high spending are the most beneficial ones to the mall owners which increases the owner's business. (Cluster 3)

Chapter 8: Reference

[1] Al-Qaed F, Sutcliffe A. Adaptive Decision Support System (ADSS) for B2C E Commerce. 2006 ICEC Eighth Int Conf Electron Commer Proc NEW E-COMMERCE Innov Conqu Curr BARRIERS, Obs LIMITATIONS TO Conduct Success Bus INTERNET. 2006:492-503.

[2] Mobasher B, Cooley R, Srivastava J. Automatic Personalization Based on Web Usage Mining. Commun ACM. 2000;43(8).

[3] Cherna Y, Tzenga G. Measuring Consumer Loyalty of B2C e-Retailing Service by Fuzzy Integral: a FANP-Based Synthetic Model. In: International Conference on Fuzzy Theory and Its Applications iFUZZY.; 2012:48-56.

[4] Magento. An Introduction to Customer Segmentation. 2014. info2.magento.com/.../ An_Introduction_to_Customer_Segmentation.

APPENDIX

Appendix A: Abbreviation

AI: Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving.

MI: Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.