DATA MINING AND WEB ALGORITHMS LAB PROJECT REPORT

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY

jiit

विद्या तत्व ज्योतिसम:

TOPIC: CUSTOMER SEGMENTATION USING K-MEANS

Group Member Details:

1. Naina Gupta	20104006, B14
2. Ojus Chugh	20104046, B15
3. Shoolin Tyagi	20104052, B15

Submitted to:

1. Dr. Aditi Sharma
2. Dr. Alka Singhal

TABLE OF CONTENTS

ABSTRACT

When you need to find your best customer, customer segmentation is the ideal method. This project uses the K-Means Clustering Algorithm in Python to perform one of the most essential applications of machine learning, Customer Segmentation. We have done the descriptive analysis of data to implement several versions of K means. Additionally, we can use the collected data to gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that yield maximum profits. This way, we can strategize the marketing techniques more efficiently and minimize the possibility of risk to the investment.

INTRODUCTION

As a result of customer segmentation, the customer base is divided into several groups of people who share a similarity that is relevant to marketing, such as gender, age, interests, and spending behaviour. In the first step of this data science project, we explored the data. We imported the essential packages required for this role and read the data. As a final step, we have gone through the input data to gain insight into it.

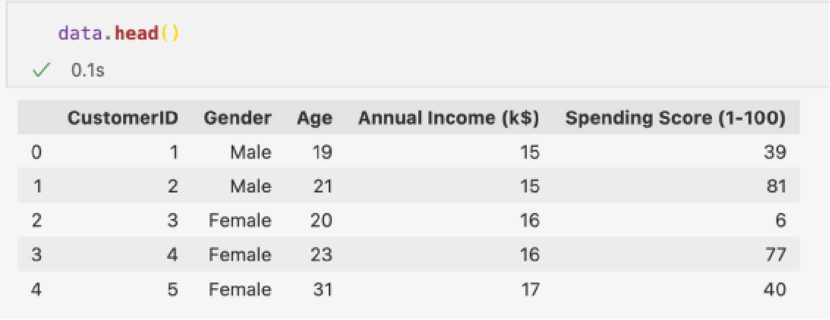
The objective of the project are as follows:

1. Identify the potential customer base for selling the product.
2. Implement Clustering Algorithms to group the customer base.

Dataset description

Mall Customer Segmentation Data

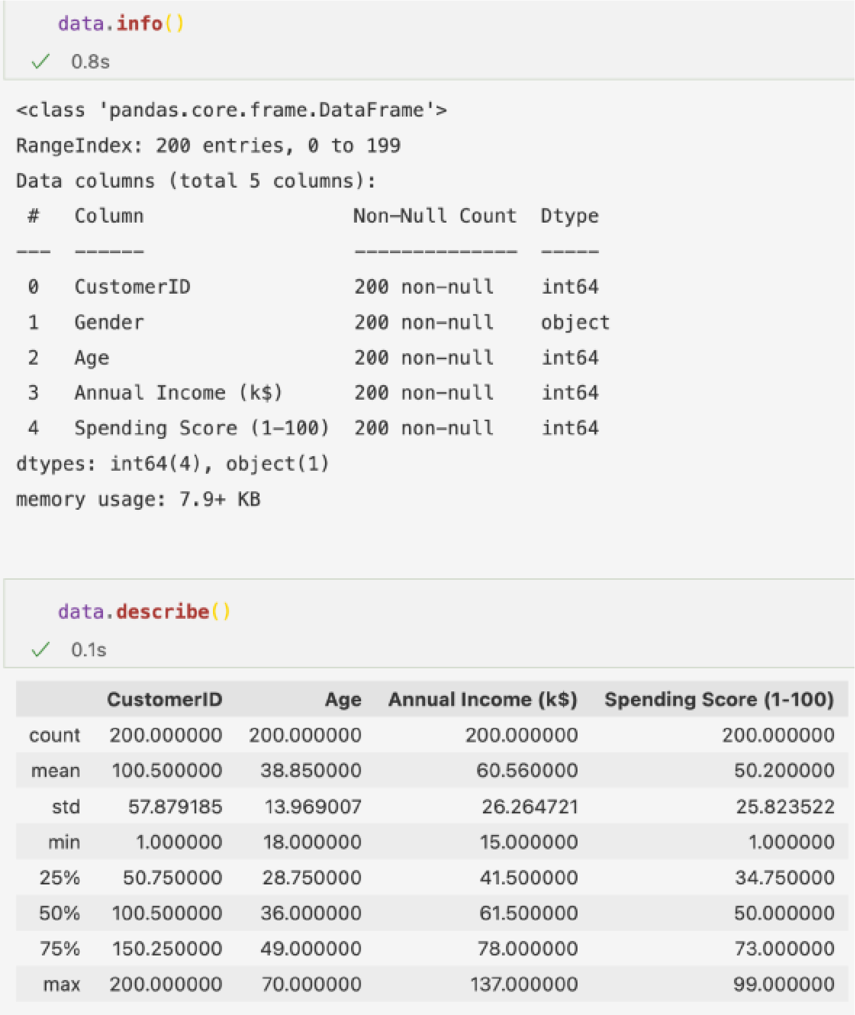
The data is given by Exposys Data Labs. It has individual unique customer IDs, A categorical variable in the form of Gender and three columns of Age, Annual Income and Spending Score which will be our main targets to identify the patterns in the customers shopping and spending spree.



```
data.head()
```

✓ 0.1s

	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



```
data.info()
```

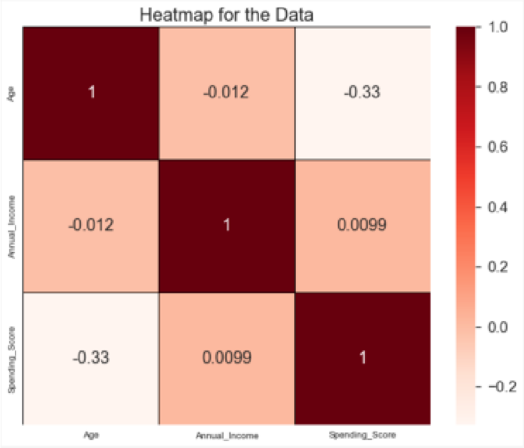
✓ 0.8s

```
<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
```

```
data.describe()
```

✓ 0.1s

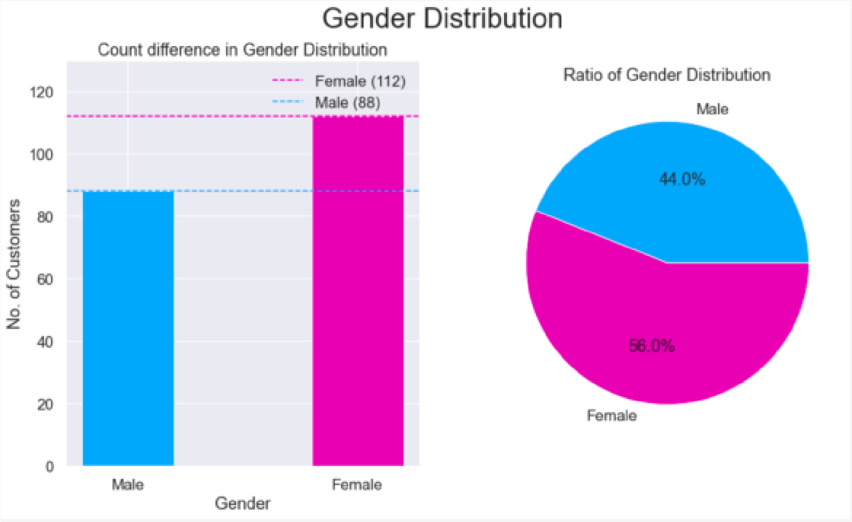
	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000



EXPLORATORY DATA ANALYSIS

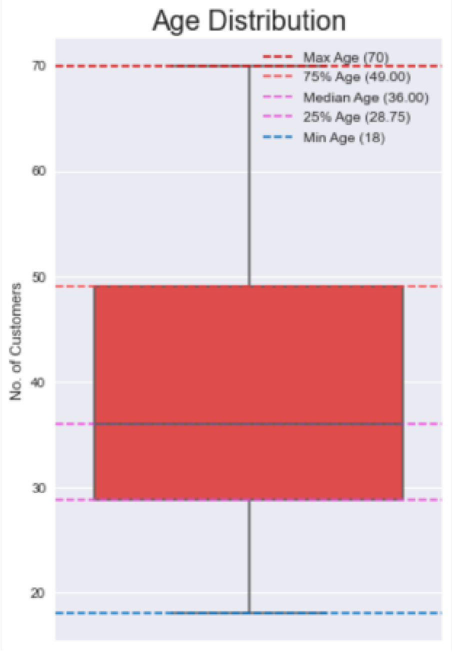
1. Visualization of Distribution of Males and Females

From the above graphs, we observe that the number of females (112) is higher than the males (88). The Ratio of Gender population is 56% Females and 44% Males. By this we can say that the majority of the customers that visit the mall are Females.



2. Age Analysis of Customers

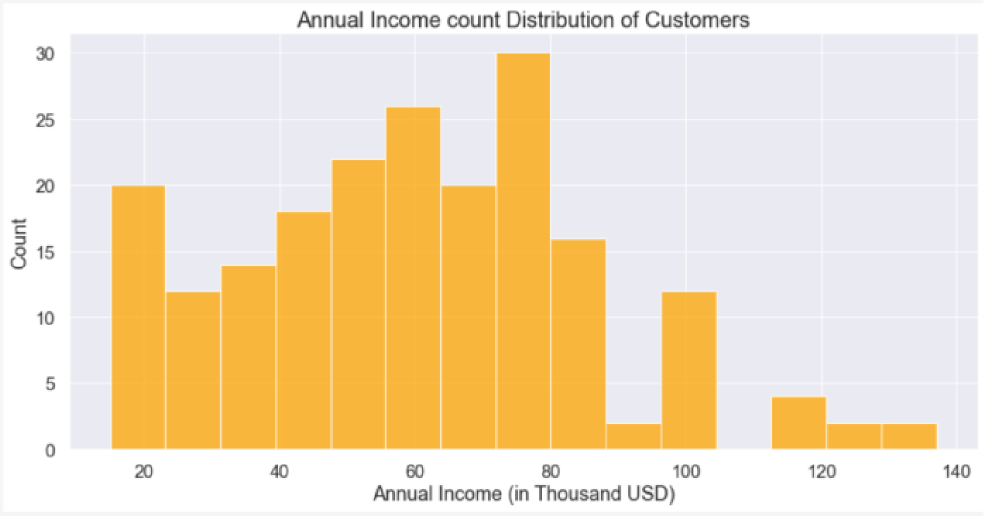
From the above boxplot, we can conclude that a large number of ages are between 30 and 35. Min Age is 18, Max Age is 70. By comparing the age distribution of the customers, we can conclude that most of the customers were within the band between 30 to 50, where the mean is around 35 years old.



3. Annual Income and Spending Score Analysis

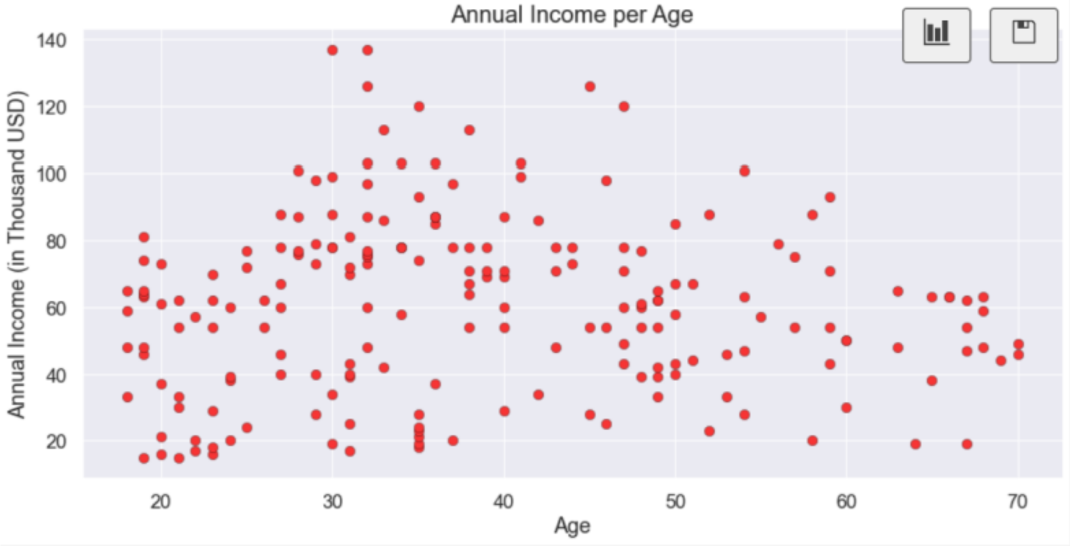
The distribution of Annual Income and Spending Score exhibited an approximation of normal distribution, with highest density around the mean of the variables. The maximum and minimum of Annual Income are 137 and 15 respectively, with the mean at 60.56. From the plot, we can see that the peak of the distribution fell in the region of 60 to 75.

For the Spending score, the maximum and minimum are 99 and 1, while the histplot 10 indicated that the highest number of customers have the spending score ranging from 40 to 60.

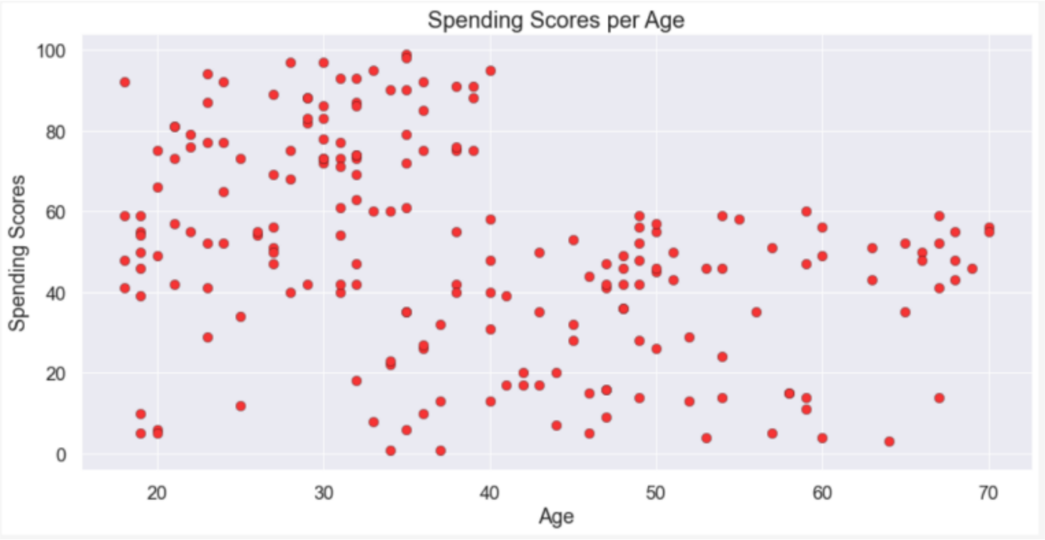


CHARACTERISTIC RELATIONS

1. Annual Income vs Age analysis



2. **Spending Score vs Age analysis**



METHODOLOGY

In our project we used following packages:

1. **Pandas**

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

2. **Numpy**

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

3. **Matplotlib**

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.

4. **Scikit Learn**

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python.

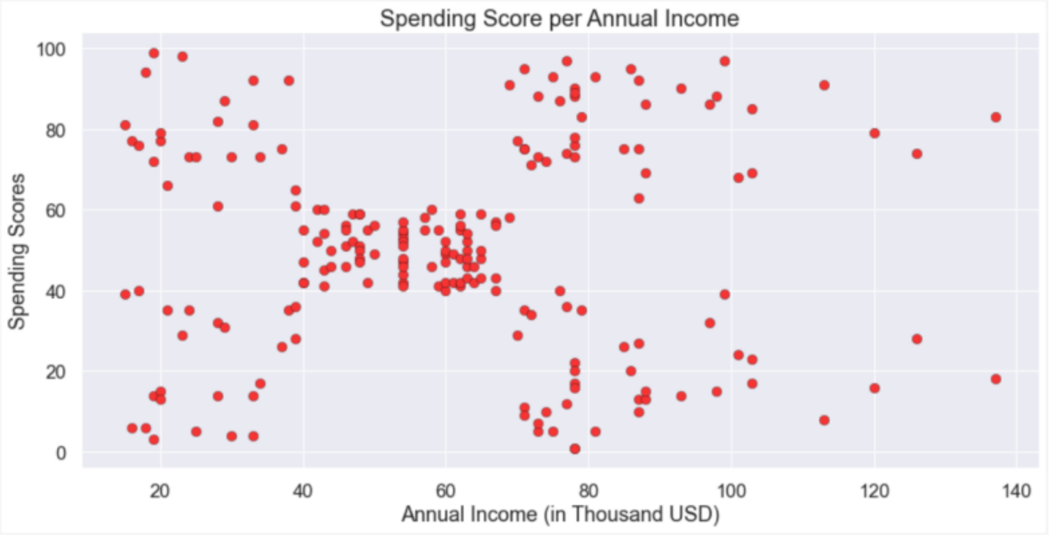
5. Seaborn

Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas' data structures. Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.

IMPLEMENTATION

What is clustering?

The process of segregating objects into groups based on their respective characteristics is called clustering. In clusters, the features of objects in a group are similar to other objects present in the same group. Clustering is used in various fields like image recognition, pattern analysis, medical informatics, genomics, data compression etc. It is part of the unsupervised learning algorithm in machine learning. This is because the data-points present are not labelled and there is no explicit mapping of input and outputs. As such, based on the patterns present inside, clustering takes place.



What is K- means clustering?

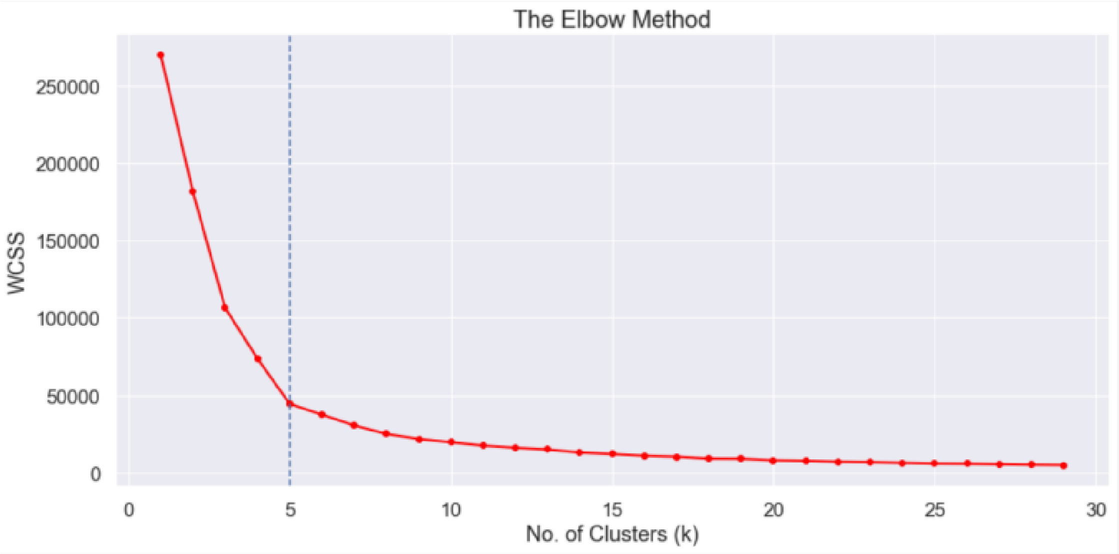
K-Means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (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 is at the minimum. The less variation we have within clusters, the more homogeneous the data points are within the same cluster.

ELBOW METHOD

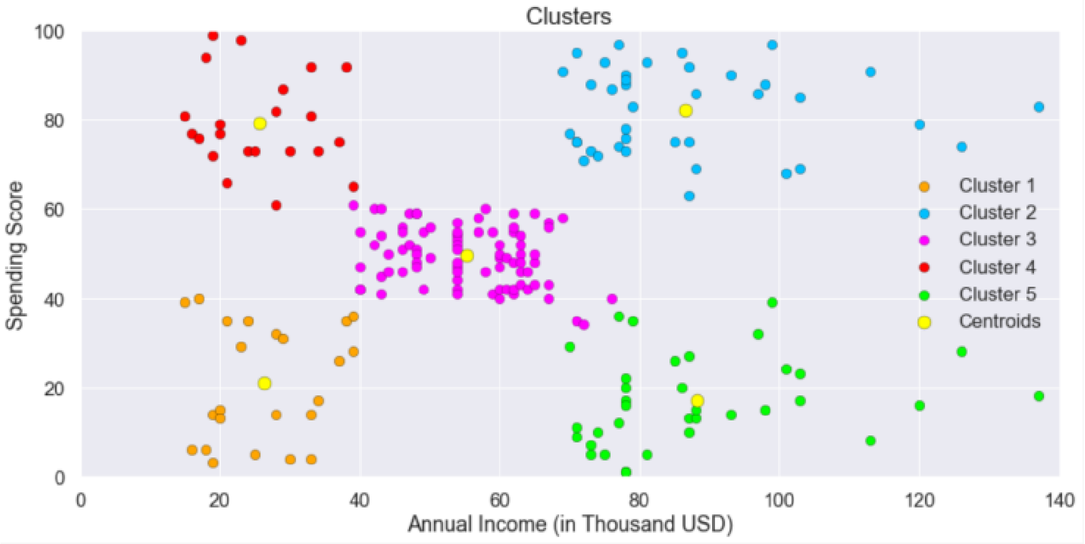
The Elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center. We use the Elbow Method which uses Within Cluster Sum Of Squares (WCSS) against the number of clusters (K Value) to figure out the optimal number of clusters value. WCSS measures the sum of distances of observations from their cluster centroids which is given by the below formula.

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$

where Y_i is centroid for observation X_i. The main goal is to maximize the number of clusters and in the limiting case each data point becomes its own cluster centroid.



It is clear, that the optimal number of clusters for our data are 5, as the slope of the curve is not steep enough after it. When we observe this curve, we see that last elbow comes at k = 5, it would be difficult to visualize the elbow if we choose the higher range.



ANALYSIS

The following clusters are created by the model:

1. Cluster Orange
2. Cluster Blue
3. Cluster Purple
4. Cluster Red
5. Cluster Green

1. Cluster Orange - Balanced Customers

They earn less and spend less. We can see people have low annual income and low spending scores, this is quite reasonable as people having low salaries prefer to buy less, in fact, these are the wise people who know how to spend and save money. The shops/mall will be least interested in people belonging to this cluster.

2. Cluster Green – Misers

Earning high and spending less. We see that people have high income but low spending scores, this is interesting. Maybe these are the people who are unsatisfied or unhappy by the mall's services. These can be the prime targets of the mall, as they have the potential to spend money. So, the mall authorities will try to add new facilities so that they can attract these people and can meet their needs.

3. Cluster Purple - Normal Customer

Customers are average in terms of earning and spending An Average consumer in terms of spending and Annual Income we see that people have average income and an average spending score, these people again will not be the prime targets of the shops or mall, but again they will be considered and other data analysis techniques may be used to increase their spending score.

4. Cluster Red – Spenders

This type of customer earns less but spends more. Annual Income is less but spending is high, so can also be treated as potential target customers. We can see that people have low income but higher spending scores, these are those people who for some reason love to buy products more often even though they have a low income. Maybe it’s because these people are more than satisfied with the mall services. The shops/malls might not target these people that effectively but still will not lose them.

5. Cluster Blue - Target Customers

Earning high and also spending high are Target Customers. Annual Income High as well as Spending Score is high, so a target consumer. We see that people have high income and high spending scores, this is the ideal case for the mall or shops as these people are the prime sources of profit. These people might be the regular customers of the mall and are convinced by the mall’s facilities.

RESULTS

We examined five segments based on customers' annual income and spending score, which are reportedly the best factors or attributes to determine the segments of customers within a mall. Customers include penny pinchers, balanced customers, target customers, spenders, and normal customers. The Target Customers can be put into some kind of alerting system where SMS and emails can be sent to them on a daily basis regarding offers and discounts that they can get at the mall and the rest we can send blast SMSs to remind them about our products once per week in a month. In addition, we can now determine customer behaviour based on their Annual Income and Spending Score. These Cluster Analysis can be used to determine many marketing strategies. High income and High spending score customers are our target customers and we would always want to retain them as they give the most profit margin to our organization. High Income and Less spending score customers can be attracted with a wide range of products in their lifestyle demands and it might attract them towards the Mall Supermarket. Less Income Less Spending Scores can be given extra offers, and constantly sending them discounts and offers will encourage them to spend. We can also have a cluster analysis done on what kind of products customers tend to buy and can make other marketing strategies accordingly. The data set did not have enough data to carry out more analytics on the same.

CONCLUSIONS

Companies, Malls, supermarkets on Small Business Enterprises should carry out Market Basket Analysis for their business. This will enable companies to target specific groups of customers, a customer segmentation model allows for the effective allocation of marketing resources and the maximization of cross- and up-selling opportunities. When a group of customers is sent personalized messages as part of a marketing mix that is designed around their needs, it's easier for companies to send those customers special offers meant to encourage them to buy more products. Customer segmentation can also improve customer service and assist in customer loyalty and retention. As a by-product of its personalized nature, marketing materials sent out using customer segmentation tend to be more valued and appreciated by the customer who receives them as opposed to impersonal brand messaging that doesn't acknowledge purchase history or any kind of customer relationship Finally with customer segmentation Companies will stay a step ahead of competitors in specific sections of the market and identify new products that 21 exist or potential customers could be interested in or improving products to meet customer expectations.