# CUSTOMER SEGMENTATION

**A MINI PROJECT REPORT**

**18CSC305J - ARTIFICIAL INTELLIGENCE**

***Submitted by***

**P Yaswanth Sai [RA2011003010991]**
**S Keshav [RA2011003010997]**
**P Mukesh Reddy [RA2011003010970]**

*Under the guidance of*

**Dr.L. Kavishanker**

Assistant Professor, Department of Data science and business systems

***in partial fulfillment for the award of the degree***

***of***

**BACHELOR OF TECHNOLOGY**

in

**COMPUTER SCIENCE & ENGINEERING**

of

**FACULTY OF ENGINEERING AND TECHNOLOGY**


S.R.M. Nagar, Kattankulathur, Chengalpattu District

**MAY 2023**

# SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Under Section 3 of UGC Act, 1956)

**BONAFIDE CERTIFICATE**

Certified that Mini project report titled **"CUSTOMER SEGEMENTATION"** is the bona fide work of **P Yaswanth Sai (RA2011003010991), Sabbavarapu Keshav (RA2011003010997) , P Mukesh Reddy (RA2011003010970)** who carried out the minor project under my supervision. Certified further, that to the best of my knowledge, the work reported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

**SIGNATURE**

DR. L.Kavishanker

**GUIDE**

Assistant Professor

Department of data science and business system

**SIGNATURE**

Dr. pushpalatha

**HEAD OF THE DEPARTMENT**

Professor & Head

Department of data science and business system

# ABSTRACT

We live in a world where large and vast amount of data is collected daily. Analyzing such data is an important need. In the modern era of innovation, where there is a large competition to be better then everyone, the business strategy needs to be according to the modern conditions. The business done today runs on the basis of innovative ideas as there are large number of potential customers who are confounded to what to buy and what not to buy. The companies doing the business are also not able to diagnose the target potential customers. This is where the machine learning comes into picture, the various algorithms are applied to identify the hidden patterns in the data for better decision making. The concept of which customer segment to target is done using the customer segmentation process using the clustering technique. In this paper, the clustering algorithm used is K-means algorithm which is the partitioning algorithm, to segment the customers according to the similar characteristics. To determine the optimal clusters, elbow method is used.

# TABLE OF CONTENTS

# LIST OF FIGURES

# ABBREVIATIONS

| | |
|---|---|
| **ML** | Machine learning |
| **BI** | Business Intelligence |
| **K-NN** | K nearest neighbor |
| **KDE** | Kernel density estimation |
| **OLS** | Ordinary Least squares |
| **WCCS** | with in the cluster sum of square |

# CHAPTER 1

# INTRODUCTION

Customers are buyers and users of products by the company. Customers have the requirements and the finaldeterminant whether or not the company product is given accordingly. Customers [1] are the company's market share that creates sales and profits for the company.

Loyal behavior [2] is customer behavior towards a product or service which of course is profitable because customers will continue to look for the product they want. Just as in obtaining customer satisfaction, everybusiness owner must improve the quality of their existing services. If the number of business company grew fasterand resulted in a lot of data growth but there was no data utilization system to find out the type of customer and to measure customer value. Thus, the business owner cannot determine which customer gives a large profit and which customer does not provide benefits.

The use of data mining techniques [3] is one solution to the problem of customer segmentation in the development of marketing strategies. Identification of customer criteria is based on the data obtained and grouped.These groups are called clustering. Clustering [4] [5] [6] is a data processing technique for grouping customers based on interactions that occur between customers and business. Customer data is customer transaction data that has many attributes [7]. The existence of this attribute results in poor data processing. There needs to be a good selection of attributes to get more optimal results. To overcome the problem of selecting customer attributes in the segmentation process, it is proposed to use the K-Means algorithm [8]. Determining customer segmentation requires a reference alternative as a benchmark in its assessment, the reference in question can be in the form of data and information obtained from several sources such as recency, frequency and monetary values of RFM. Information obtained from related sources is obtained through data observation methods. The data that has been collected can be the basis of assessment for segmentation of customers [9] [10].

Over the years, the competition amongst businesses is increased and the large historical data that is available has resulted in the widespread use of data mining techniques in extracting the meaningful and strategic information from the database of the organization.

Data mining is the process where methods are applied to extract data patterns in order to present it in the human readable format which can be used for the purpose of decision support. According to,[4] Clustering techniques consider data tuples as objects. 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 theprocess 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 mange the demand andsupply 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. 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, determining new market opportunities, making better brand strategy, identifying customers retention.

This research aims to find out the type of customer and measure the value of the customer so that the businessowner can determine which customer gives a large profit and which customer does not provide benefits.

# CHAPTER 2

# LITERATURE SURVEY

Over the years, as there is very strong competition in the business world, the organizations have to enhancetheir profits and business by satisfying the demands of their customers and attract new customers according totheir needs. The identification of customers and satisfying the demands of each customer is a very complex and tedious task. This is because customers may be different according to their demands, tastes, preferences and so on. Instead of "one-size-fits-all" approach, customer segmentation clusters the customers into groups sharing the same properties or behavioral characteristics. According to,[5] customer segmentation is a strategy of dividing themarket into homogenous groups. The data used in customer segmentation technique that divides the customers into groups depends on various factors like, data geographical conditions, economic conditions, demographical conditions as well as behavioral 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, determining new market opportunities, making better brand strategy, identifying customers retention.

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. 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 was applied to build customer segmentation in mega business retail outfit for targeted customer services [4]. The background was because customerscan be segmented based ondemography and behaviors. Customer segmentation was conducted using K-Means for milk industry to develop customer profile to explore their profitable [11] and market segmentation to learn customer's preferences.

These researches [5, 6] employed K-Means and decision tree as data mining approaches. K- Means was used to group bank customer based on their similarity in credit payment history [5] and RFM score [6]. The classification rules were developed using decision tree as classification technique to predict customer eligibility in credit approval [5] and categorize customer rank [6]. K-Means and RFM methods were used for customer segmentation [9, 10]. The researches proposed customer segmentation using RFM, K-Means and LEM2 methods in electronic company [8]and fertilizer manufacturer [13].

These researches [5, 6] employed K-Means and decision tree as data mining approaches. K- Means was used to group bank customer based on their similarity in credit payment history [5] and RFM score [6]. The classification rules were developed using decision tree as classification technique to predict customer eligibility in credit approval [5] and categorize customer rank [6].

# CHAPTER 3

# SYSTEM ARCHITECTURE AND DESIGN

Customer segmentation using k-means clustering is a popular technique used in marketing analytics to group customers into distinct segments based on their similarities in behavior or preferences. Here's an overview of the system architecture and information required for customer segmentation using k-means clustering:

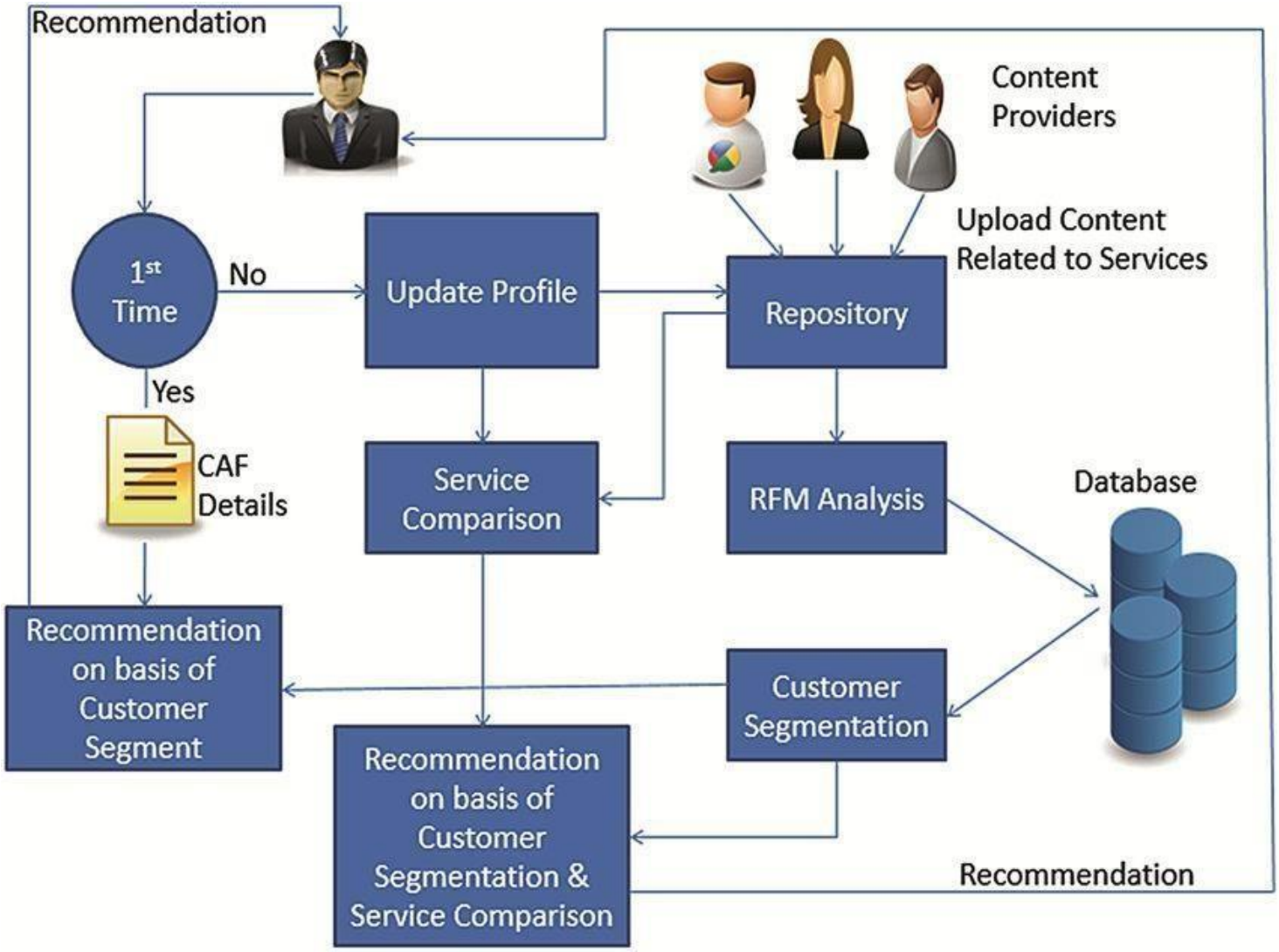



Fig 3.1 system architecture diagram

1) Data Collection: Collect customer data from various sources such as customer transactions, social media,surveys, etc.

2) Data Preprocessing: Clean the data by removing duplicates, missing values, outliers, etc. Transform thedata to a standardized format for analysis.

3) Feature Extraction: Extract relevant features from the data, such as customer demographics, purchasehistory, interests, etc.

4) Clustering: Apply k-means clustering algorithm to group customers based on their similarities in theextracted features.

5) Visualization: Visualize the clusters using graphs, charts, or other visualization tools to better understandthe segments.

6) Analysis: Analyze the segments to identify their unique characteristics and formulate marketing strategiesto target each segment.

Implementation: Implement the marketing strategies for each segment and monitor the results.

## 3.1 DATA MINING

Data mining is the process of sorting through large data sets to identify patterns and relationships that canhelp solve business problems through data analysis. Data mining techniques and tools enable enterprises to predictfuture trends and make more-informed business decisions. Data mining is a key part of data analytics overall and one of the core disciplines in data science, which uses advanced analytics techniques to find useful information in data sets. At a more granular level, data mining is a step in the knowledge discovery in databases (KDD) process,a data science methodology for gathering, processing and analyzing data. Data mining and KDD are sometimes referred to interchangeably, but they're more commonly seen as distinct things.

Data mining is a crucial component of successful analytics initiatives in organizations. The information it generates can be used in business intelligence (BI) and advanced analytics applications that involve analysis of historical data, as well as real-time analytics applications that examine streaming data as it's created or collected.

After the data selection process we will mine the data and look for the important attributes in the data and their values, the spending score of the customer that each customer how much he ca spend upon, the annual income how each Individual Customer earns and how much he will be capable to spend and how each actually spends, so based upon all of the values we will mine the data and use it for clustering.

```
#Import the necessary libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
```

Fig 3.1.1 DEPEDENCIES

Numpy :- NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, Fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely NumPy stands for Numerical Python.

In Python we have lists that serve the purpose of arrays, but they are slow to process. NumPy aims to provide an array object that is up to 50x faster than traditional Python lists. The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy. Arrays are very frequently used in data science, where speed and resources are very important.

Pandas:- Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008. Pandas allows us to analyze big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant. Relevantdata is very important in data science

Matplotlib:- Matplotlib: Visualization Matplotlib is a comprehensive library for creating static, animated, andinteractive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

1. Create publication quality plots.
2. Make interactive figures that can zoom,

pan, update.3.Customize visual style and

layout.

4. Export to many file formats.
5. Embed in Jupyter Lab and Graphical User

Interfaces. 6.Use a rich array of third-party

packages built on Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.Matplotlib makes easy things easy and hard things possible.

Seaborn:- Importing libraries in Jupyter Notebook

->Loading dataset

->Python Seaborn Plotting Functions

->Bar plot

->Count plot

->Distribution plot

->Heatmap

->Scatter plot

->Pair plot

->Linear Regression plot

->Box plot

Python Seaborn library is a widely popular data visualization library that is commonly used for data science and machine learning tasks. You build it on top of the matplotlib data visualization library and can perform exploratoryanalysis. You can create interactive plots to answer questions about your data.

To understand the Seaborn library and the different plotting functions in detail, you'll need to use a few datasets tocreate the visualizations.

Seaborn is more comfortable with Pandas data frames. It utilizes simple sets of techniques to produce lovely images in Python. Matplotlib is highly customized and robust. With the help of its default themes, Seaborn prevents overlapping plots.

Scikit Learn:- Scikit-learn is an open source data analysis library, and the gold standard for Machine Learning (ML) in the Python ecosystem. Key concepts and features include:

Algorithmic decision-making methods, including:

Classification: identifying and categorizing data based on patterns.

Regression: predicting or projecting data values based on the average mean of existing and planned data.Clustering: automatic grouping of similar data into datasets.

Algorithms that support predictive analysis ranging from simple linear regression to neural network patternrecognition.

Interoperability with NumPy, pandas, and matplotlib libraries.

## 3.2 DATA VISUALISATION

### 3.2.1 GENDER VISUALIZATION

To visualize the count of genders we take bar plot as a visualization tool. we takegenders on X-axis and count on Y-axis.

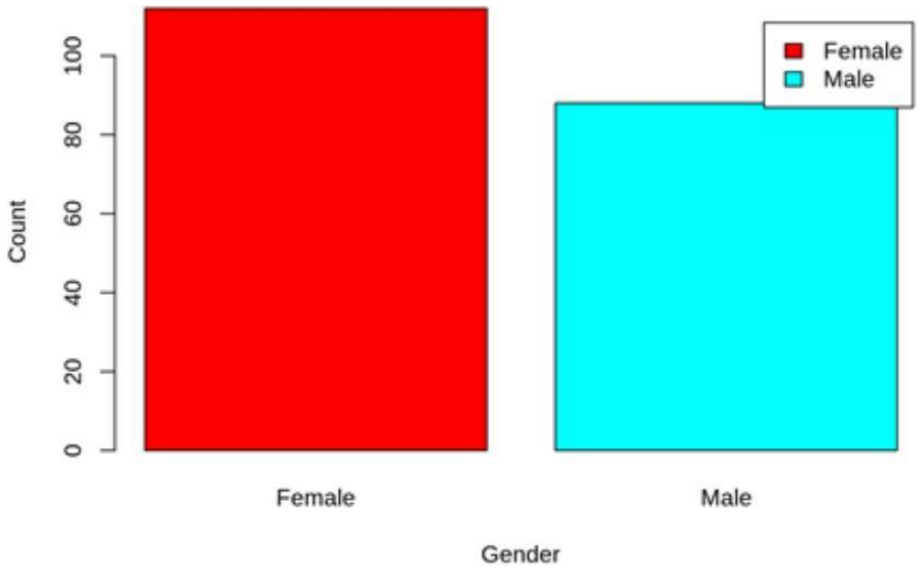



Fig 3.2.1 Gender visualization

### 3.2.2 AGE VISUALISATION

We take Histogram to visualize the age of the customers. We take age class andfrequency as an input to visualize the dataset.

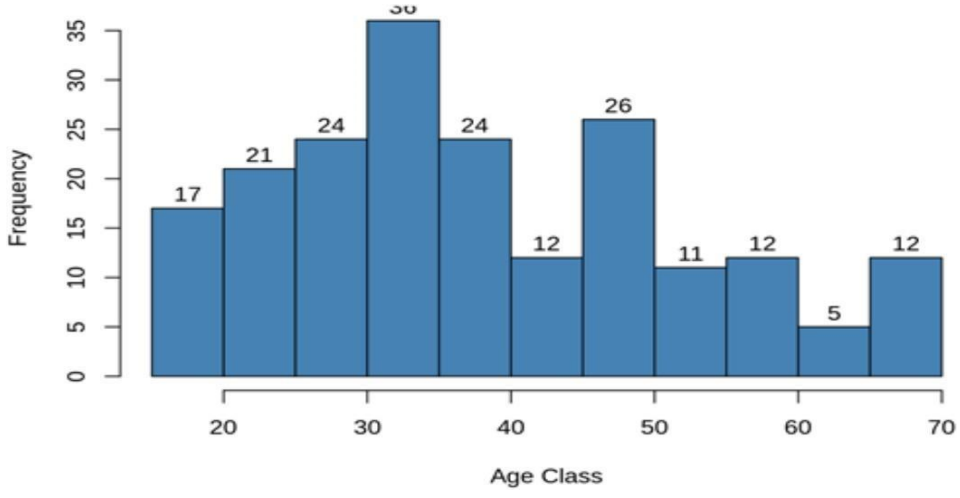



Fig 3.2.2 Age visualization

### 3.2.3 VISUALIZATION OF SEGMENT OF CUSTOMERS

To visualize segments of the customer we draw a graph between spending score and annual income.

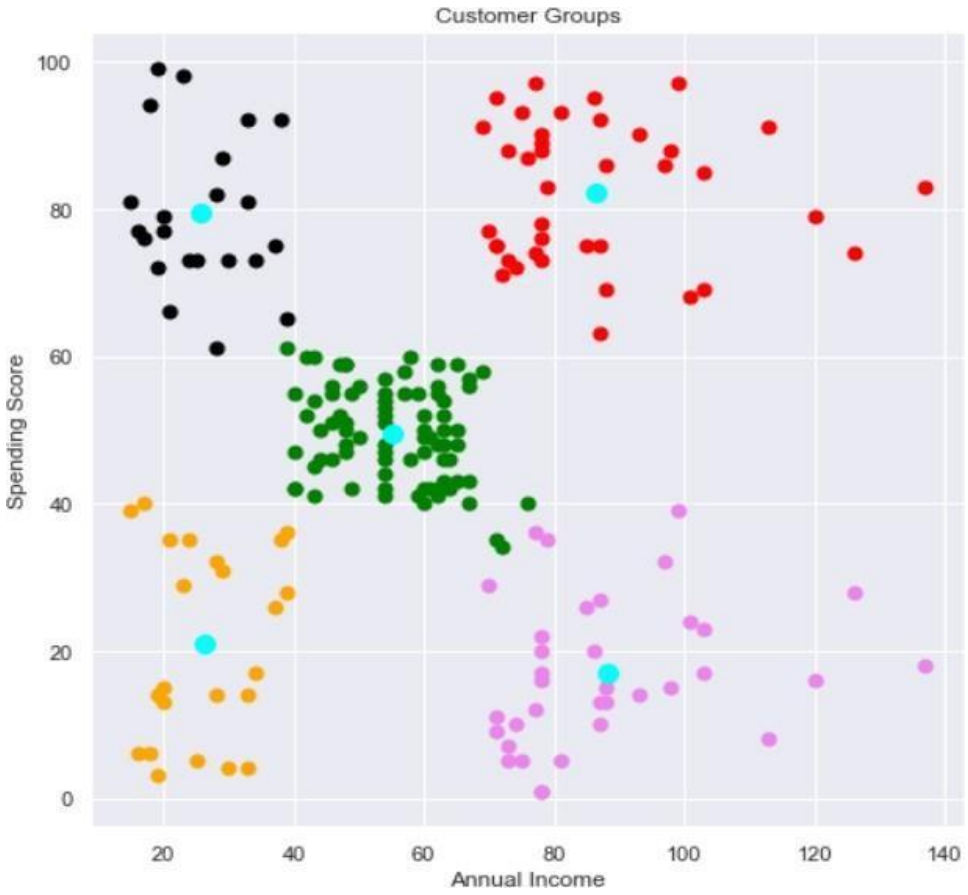



Fig 3.2.3 segments of customer

# CHAPTER 4

# METHODOLOGY

## 4.1 PROBLEM DEFINITION

The supermarkets, malls are crowded with people every time but unable to convert them into potentialbuyers, this makes difficult to the mall, supermarket shopping malls owners to have a track of the customer behavior. So customer segmentation is a marketing strategy that involves dividing a customer base into groups of individuals who have similar characteristics of behaviors. K means clustering is a popular machine learning algorithm used for customer segmentation which gives potential information of the customer behaviors.

## 4.2 SCOPE

Online retailers can use customer segmentation to group customers based on their purchasing behavior and preferences. This can help to personalize marketing campaigns and improve the customer experience. Banks can use customer segmentation to group customers based on their transaction history, account balances, and credit scores. This can help to tailor product offerings and customer service to each segment. Hospitals and health insurance providers can use customer segmentation to group patients based on their medical history, demographics, and lifestyle factors. This can help to personalize healthcare plans and improve patient outcomes. Hotels, airlines, and other travel companies can use customer segmentation to group customers based on their travel history, preferences, and loyalty status. This can help to tailor marketing campaigns and improve the customer experience. Brick-and-mortar retailers can use customer segmentation to group customers based on their shopping behavior and demographics. This can help to personalize in-store experiences and improve customer retention.

## 4.3 PROPOSED SOLUTION

The proposed solution of Customer Segmentation is the subdivision of a market into discrete customer groups that share similar characteristics. Customer Segmentation can be a powerful means to identify unsatisfied customer needs. Using the above data companies can then outperform the competition by developing uniquely appealing products and services.

The most common ways in which businesses segment their customer base are:

1. Demographic information, such as gender, age, familial and marital status, income, education, and occupation.

2. Geographical information, which differs depending on the scope of the company. For localized businesses, this info might pertain to specific towns or counties. For larger companies, it might mean a customer's city, state, or even country of residence.

3. Psychographics, such as social class, lifestyle, and personality traits.

4. Behavioral data, such as spending and consumption habits, product/service usage, and desired benefits.

### 4.3.1 ADVANTAGES OF CUSTOMER SEGMENTATION

1.Determine appropriate product

pricing. 2.Develop customized

marketing campaigns.3.Design an

optimal distribution strategy.

4.Choose specific product features for

deployment.5,Prioritize new product

development efforts.

## 4.4 DATA CLEANING

Data cleaning is the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in data. It is an important step in data analysis and machine learning because the quality of the data can greatly impact the accuracy of the results.

### 4.4.1 DATA CLEANING PROCESS

The process of data cleaning can involve a variety of tasks, including:

Removing duplicate records: Identifying and removing records that contain duplicate data.

1) Handling missing values: Determining how to handle missing data, such as imputing missing values orremoving records with missing data.

2) Correcting data inconsistencies: Identifying and correcting data that is inconsistent or does not make sense.

3) Converting data types: Ensuring that data is in the correct format and type for analysis.

4) Standardizing data: Ensuring that data is consistent across the dataset, such as standardizing date formats orconverting text to lowercase.

5) Removing outliers: Identifying and removing data points that are significantly different from the rest of thedata.

Overall, data cleaning is an essential step in preparing data for analysis and ensuring the accuracy of the results. Itcan be a time-consuming process, but it is important to invest the time and effort to ensure the quality of the data.

Table II shows attributes of Customers visiting the Mall. Meanwhile, data cleaning processes applied isshown in table III.

TABLE II. ATTRIBUTES OF CUSTOMERS VISITING MALL

| **Attribute Name** |
|---|
| Customer-id |
| Gender |
| Age |
| Annual Income |
| Spending Score |

Table 4.4.1 attribute names

TABLE III. DATA CLEANING

| **Problem** | **Solution** |
|---|---|
| There are null values in theattributes of the spending score and the annual income | Take the sum of the all values of these attributes and average them and add them to the missingvalues |

## 4.5 DATA SELECTION

Data selection is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data. Data selection precedes the actual practice of data collection. This definition distinguishes data selection from selective data reporting (selectively excluding data that is not supportive of a research hypothesis) and interactive/active data selection (using collected data for monitoring activities/events, or conducting secondary data analyses). The process of selecting suitable data for a research project can impact data integrity.

The primary objective of data selection is the determination of appropriate data type, source, and instrument(s) that allow investigators to adequately answer research questions. This determination is often discipline-specific and is primarily driven by the nature of the investigation, existing literature, and accessibility to necessary data sources.

Integrity issues can arise when the decisions to select 'appropriate' data to collect are based primarily on cost and convenience considerations rather than the ability of data to adequately answer research questions. Certainly, cost and convenience are valid factors in the decision-making process. However, researchers should assess to what degree these factors might compromises the integrity of the research endeavor.

According the data set and the problem we are going to solve, the most effective attributes would be selecting the spending score and the annual income of an individual customer to segregate the customers into different clusters using k means clustering.

# 4.6 K-MEANS CLUSTURING

### 4.6.1 ELBOW METHOD

It is the simplest and most commonly used iterative type of unsupervised learning algorithm. Unlike supervised learning, we don't have labeled data in K-Means. Some other unsupervised learning algorithms arePCA (Principle Component analysis), K-Medoid, etc. In K-Means, we randomly initialize the K number of cluster centroids in the data (the number of k found using the Elbow Method will be discussed later in this tutorial) and iterates these centroids until no change happens to the position of the centroid. Let's go through the steps involved in K-means clustering for a better understanding.

Select the number of clusters for the dataset (K) Select the K number of centroids randomly from the dataset. Now we will use Euclidean distance or Manhattan distance as the metric to calculate the distance of the points from the nearest centroid and assign the points to that nearest cluster centroid, thus creating K clusters.Now we find the new centroid of the clusters thus formed. Again reassign the whole data point based on this new centroid, then repeat step 4. We will continue this for a given number of iterations until the position of the centroiddoesn't change, i.e., there is no more convergence. Finding the optimal number of clusters is an important part of this algorithm. A commonly used method for finding the optimum K value is Elbow Method.

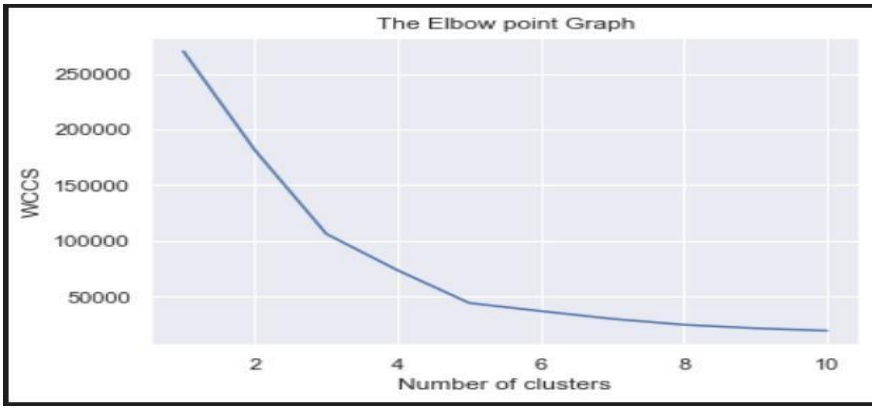



Fig 4.6.1 ELBOW METHOD

So from the above methods we can say that choosing of k value would be more significant and more convenient inperforming the clustering algorithm and to get better results, accuracy and precision.

### 4.6.2 K-MEANS ALGORITHM

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering. K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way thateach dataset belongs only one group that has similar properties. It allows us to cluster the data into different groupsand a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

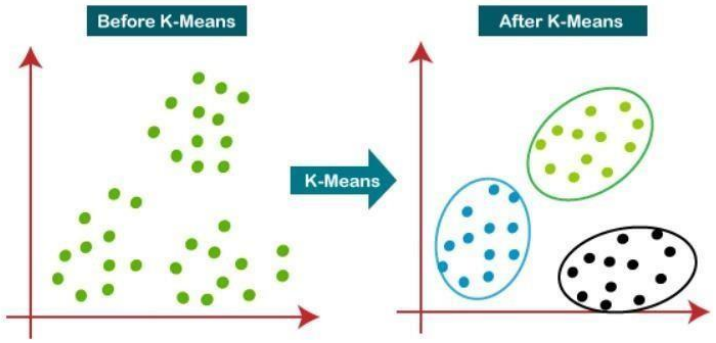



**Fig 4.6.2 K-means clustering**

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm. k-means clustering algorithm mainly performs two tasks:

Determines the best value for K center points or centroids by an iterative process.

Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, createa cluster.

### 4.6.3 STEPS IN K MEANS ALGORITHM

1) Select the number K to decide the number of clusters.

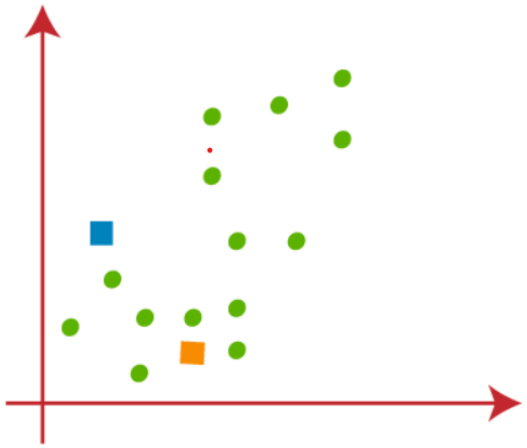


Fig 4.6.3 selecting the value of k

2) Now we will assign each data point of the scatter plot to its closest K-point or centroid. We will compute it byapplying some mathematics that we have studied to calculate the distance between two points. So, we will draw a median between both the centroids. Consider the below image:

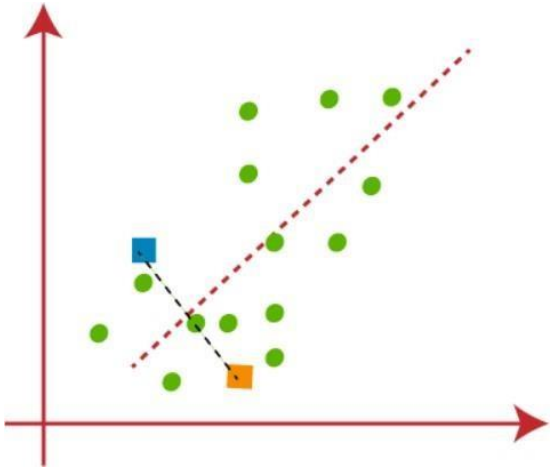


Fig 4.6.4 Assigning the data points to the scatter plot

3) As we need to find the closest cluster, so we will repeat the process by choosing a new centroid**.** To choose thenew centroids, we will compute the center of gravity of these centroids, and will find new centroids as below:

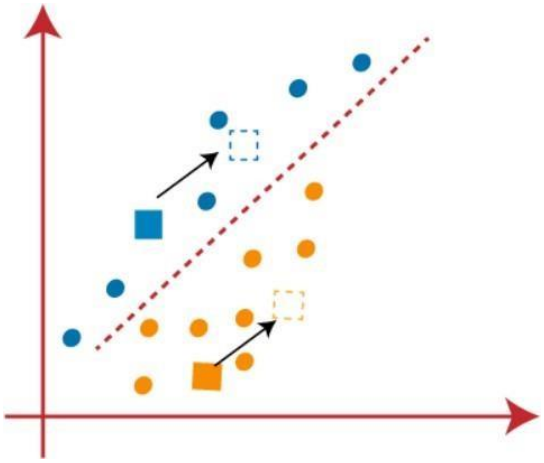


Fig 4.6.5 choosing a new centroid

4) By keep on repeating this process all the data points of the cluster will get assign to the different centroids andthis make all the data points to form two different categories and form two different clusters as shown below

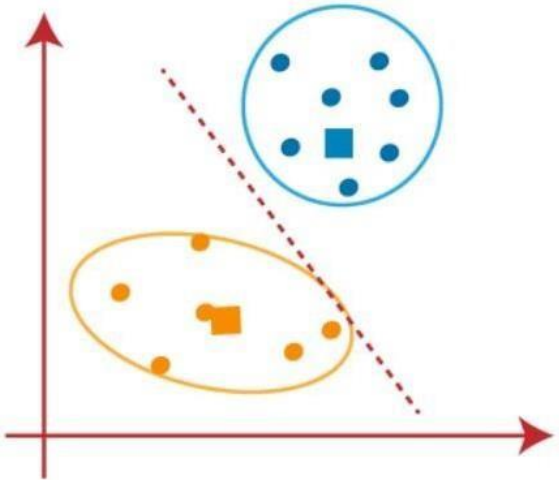


Fig 4.6.6 final clusters

5) As our model is ready, so we can now remove the assumed centroids and we can say that two final clusters areformed and thus the main objective of the algorithm is solved.

# CHAPTER 5

# CODING AND TESTING

## 5.1 PREPROCESSING OF THE DATA

The below functions are used to load the dataset

```
df = pd.read_csv('data/churn modelling.csv',
index_col=0)df.head()
customer_data.tail()
```

### 5.1.1 FINDING NULL VALUES IN THE DATA

With the help of "df.isna().sum()" we can find the null values in the data set.

```
CustomerID              0
Gender                  0
Age                     0
Annual Income (k$)      0
Spending Score (1-100)  0
dtype: int64
```

## 5.2 SELECTING THE VALUES

Iloc function is used to select the attributes in the data set and we will store it in the variable x and we willmake use of that variable to print the data.

```
X=customer_data.iloc[:,[
 3,4]].valuesprint(X)
```

## 5.3 CHOOSING THE NUMBER OF CLUSTERS IN THE DATASET

It is very important to choose the number of clusters for performing the kmeans clustering algorithm asdepending upon the number of clusters only we will segregate the clusters.

```
wccs=[]
for i in range(1,11):
  kmeans=KMeans(n_clusters=i,init='k-
  means++',random_state=42)kmeans.fit(X)

  wccs.append(kmeans.inertia_)
```

## 5.4 ELBOW METHOD

. Finding the optimal number of clusters is an important part of this algorithm. A commonly used method forfinding the optimum K value is Elbow Method. So in this one we got the k value as 5.

```
#Plot and
elbow graph
sns.set()
plt.plot(range(
1,11),wccs)
plt.title("The Elbow
point Graph")
plt.xlabel('Number of
clusters')
plt.ylabel('WCCS')
plt.show()
```

## 5.5 APPLYING THE PREDICTION ALGORITHM TO THE DATA

In this one we will fit the data to the model and we will find the outcome through it and by using printfunction we will print the data.

```
#Training the k means clustering model
kmeans=KMeans(n_clusters=5,init='k-
means++',random_state=0)#Return a label for each
data point based on their cluster
Y=kmeans.fit_predict(X)
```

## 5.6 VISUALIZATION OF THE DATA

In this we will show the output in the form of graphs which we predicted in the prediction state. And the outputwill be the number of clusters that formed after the k means clustering

```
plt.figure(figsize=(8,8))
plt.scatter(X[Y==0,0],X[Y==0,1],s=50,c='violet',label='Cluster  1')
plt.scatter(X[Y==1,0],X[Y==1,1],s=50,c='green',label='Cluster  2')
plt.scatter(X[Y==2,0],X[Y==2,1],s=50,c='red',label='Cluster 3')
plt.scatter(X[Y==3,0],X[Y==3,1],s=50,c='black',label='Cluster 4')
plt.scatter(X[Y==4,0],X[Y==4,1],s=50,c='orange',label
='Cluster 5')#Plot the centroids
plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],s=100,c='cyan',label='centroids')
plt.title('Customer Groups') plt.xlabel('Annual Income') plt.ylabel('Spending Score') plt.show()
```

# CHAPTER 6

# SCREENSHOTS AND RESULTS

## 6.1 ADDING DEPENDENCIES

In [3]:
```
#Import the necessary libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
```

## 6.2 UPLOADING THE DATASET

In [4]:
```
#Data analysis
```

In [5]:
```
#loading the data from csv file to a pandas DataFrame
customer_data=pd.read_csv("C:\\Users\\kotha\\OneDrive\\Desktop\\Mall_Customers.csv")
```

## 6.3 PRINTING OF DATASET

In [8]:
```
#printing first five rows of the dataframe
customer_data.head()
```

Out[8]:

| | 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 |

In [9]:
```
#printing last five rows of the dataframe
customer_data.tail()
```

Out[9]:

| | CustomerID | Gender | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|---|
| **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 |

## 6.4 CHECKING THE DATASET SIZE

```
In [10]: #Finding the number of rows nad columns
         customer_data.shape
```

```
Out[10]: (200, 5)
```

## 6.5 PRINTING THE INFORMATION OF THE DATASET

```
In [13]: #Getting some information about the dataset
         customer_data.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
```

## 6.6 CHECKING THE NULL VALUES OF THE DATASET

```
In [14]: #Checking the missing values
         customer_data.isnull().sum()
```

```
Out[14]: CustomerID              0
         Gender                  0
         Age                     0
         Annual Income (k$)      0
         Spending Score (1-100)  0
         dtype: int64
```

```
In [15]: #In the given dataset the customer id is useless because we are doing depending upon the score
         #The group of customers will be classified based on the annual income and spending score.
```

```
In [16]: #Choosing the annual income column and spending score column
```

## 6.6 SELECTING THE ATTRIBUTES OF THE DATASET

```
In [20]: print(X) #In the array first value represnts annual income and second value represents spending score
```

```
[[ 15  39]
 [ 15  81]
 [ 16   6]
 [ 16  77]
 [ 17  40]
 [ 17  76]
 [ 18   6]
 [ 18  94]
 [ 19   3]
 [ 19  72]
 [ 19  14]
 [ 19  99]
 [ 20  15]
 [ 20  77]
 [ 20  13]
 [ 20  79]
 [ 21  35]
 [ 21  66]
 [ 23  29]
```

```
 [ 71  75]
 [ 71   9]
 [ 71  75]
 [ 72  34]
 [ 72  71]
 [ 73   5]
 [ 73  88]
 [ 73   7]
 [ 73  73]
 [ 74  10]
 [ 74  72]
 [ 75   5]
 [ 75  93]
 [ 76  40]
 [ 76  87]
 [ 77  12]
 [ 77  97]
 [ 77  36]
 [ 77  74]
```

```
In [21]: #Choosing the number of the clusters

         #WCCS-> WITHIN CLUSTERS SUM OF SQUARES
```

```
In [22]: #Finding wccs values for different number of clusters
```

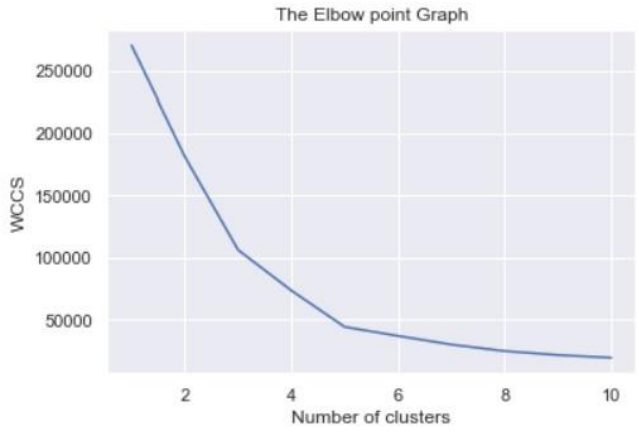
## 6.7 CHOOSING THE NUMBER OF CLUSTERS

```
In [30]: wccs=[]
         for i in range(1,11):
             kmeans=KMeans(n_clusters=i,init='k-means++',random_state=42)
             kmeans.fit(X)

             wccs.append(kmeans.inertia_)
```

## 6.8 ELBOW METHOD

In [31]:
```
#Plot and elbow graph
sns.set()
plt.plot(range(1,11),wccs)
plt.title("The Elbow point Graph")
plt.xlabel('Number of clusters')
plt.ylabel('WCCS')
plt.show()
```




In [32]:
```
#It is also called cutoff point graph
```

In [33]:
```
#optimum number of clusters is 5
```

## 6.9 TRAINING THE DATASET

In [36]:
```
#Training the k means clustering model
kmeans=KMeans(n_clusters=5,init='k-means++',random_state=0)

#Return a label for each data point based on their cluster
Y=kmeans.fit_predict(X)
```

In [37]:
```
print(Y)
```

```
[4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4
 3 4 3 4 3 4 1 4 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 2 0 2 1 2 0 2 0 2 1 2 0 2 0 2 0 2 0 2 1 2 0 2 0 2
 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0
 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2]
```

In [38]:
```
#Data visualization
#Visualizing all the clusters
```

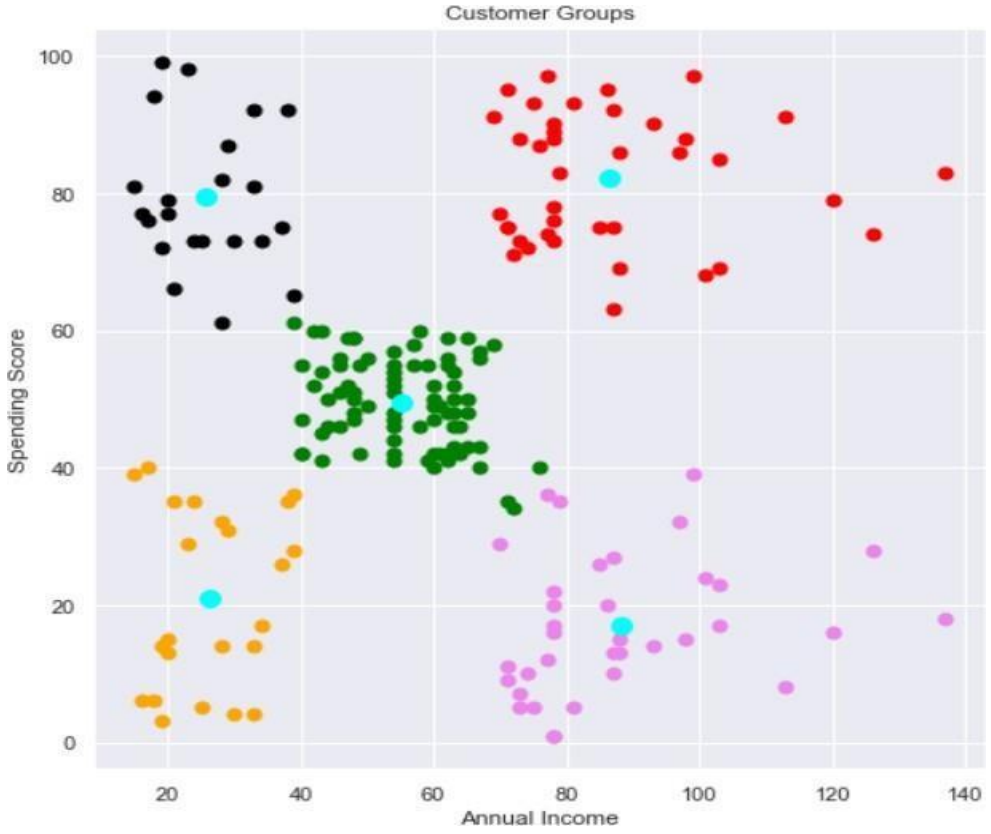
In [40]:
```
#Plotting all the clusters and their centroids
#5 clusters=0,1,2,3,4
```

## 6.10 VISUALIZATION OF THE DATASET

In [49]:
```
plt.figure(figsize=(8,8))
plt.scatter(X[Y==0,0],X[Y==0,1],s=50,c='violet',label='Cluster 1')
plt.scatter(X[Y==1,0],X[Y==1,1],s=50,c='green',label='Cluster 2')
plt.scatter(X[Y==2,0],X[Y==2,1],s=50,c='red',label='Cluster 3')
plt.scatter(X[Y==3,0],X[Y==3,1],s=50,c='black',label='Cluster 4')
plt.scatter(X[Y==4,0],X[Y==4,1],s=50,c='orange',label='Cluster 5')

#Plot the centroids
plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],s=100,c='cyan',label='centroids')

plt.title('Customer Groups')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.show()
```




In [4]:
```
#This is how the malls improve their markieting and give the great discount on the group of customers
```

In [3]:
```
#In the above example there are some group of customers where the customers are buying and leaving
```

# CHAPTER 7

# CONCLUSION AND FUTUREENHANCEMENTS

## CONCLUSION

From the above visualization it can be observed that Cluster 1 denotes the customer who has high annual income as well as high yearly spend. Cluster 2 represents the cluster having high annual income and low annual spend. Cluster 3 represents customer with low annual income and low annual spend. Cluster 5 denotes the low annual income but high yearly spend. Cluster 4 and cluster 6 denotes the customer with medium income and medium spending score. The results of the tests performed show that in the K-Means algorithm test, cluster 3 isthe best cluster that the company can use as a promotional medium to loyal customers. Further research can be added with several criteria so that the system can be better and right on target. Development of supporting applications that use other tools and methods can be used as comparisons to the system that have been developed.

# FUTURE SCOPE

The presented recommendation engines can be used to help the telecom companies in recommending appropriate services to the customer. Hence, gaining competitive edge in the market. Some open issues that need special attention in future research are,

- Approach to integrate above recommendation engine with the systems already used without increasing workload of them.
- To have more input parameters on basis of which recommendation could be given.
- To give real time recommendations to the customer on the basis of location, time etc

K-means clustering can be used to segment customers based on their behavior and preferences, enabling businesses to personalize their marketing and sales strategies for each group. For example, segmenting customers based on their purchase history, browsing behavior, or demographic information can help businesses target specific groups with tailored messages and offers.

Customer segmentation using k-means clustering can help businesses understand the needs and preferences of different customer segments. This information can be used to develop products that meet thespecific needs of each segment, resulting in more successful product launches and higher customer satisfaction.

K-means clustering can be used to segment customers based on their geographic location, enabling businesses to conduct targeted market research in specific regions. This information can be used to develop localized marketing campaigns or to tailor products to meet the needs of customers in specific regions.

K-means clustering can be used to identify customers who are likely to purchase complementary products or upgrade to higher-tier services. This information can be used to develop targeted cross-selling and upselling campaigns, resulting in increased revenue and customer satisfaction.

# REFERENCES


1. Ericsson, L.M.: More than 50 billion connected devices. White Paper (2011)

2. Ricci, F., Rokach, L., Shapira, B.: Introduction to Recommender Systems Handbook.Springer, US (2011)

3. Ge, Y., Xiong, H., Tuzhilin, A., Xiao, K., Gruteser, M., Pazzani, M.J.: An energy-efficient mobile recommender system (PDF). In: Proceedings of the 16th ACM SIGKDD InternationalConference on Knowledge Discovery and Data Mining, pp. 899–908. ACM, New York (2010). Accessed 17 Nov 2011 Big Data Analytics Based Recommender System for Value Added Services (VAS) 149

4. Bouneffouf, D.: Following the customer's interests in mobile context-aware recommendersystems: the hybrid-e-greedy algorithm. In: Proceedings of the 2012 26th International Conference on Advanced Information Networking and Applications Workshops (PDF). LNCS, pp. 657–662. IEEE Computer Society (2012). ISBN: 978-0-7695-4652-0 [dead link]

5. Yeung, K.F., Yang, Y.: A proactive personalized mobile news recommendation system. In: 2010 Developments in E-systems Engineering (DESE), pp. 207–212. IEEE (2010)

6. Danalet, A., Farooq, B., Bierlaire, M.: A Bayesian approach to detect pedestrian destinationsequences fromWiFi signatures. Transp. Res. Part C Emerg. Technol. 44, 146–170 (2014). doi:10.1016/j.trc.2014.03.015

7. Fang, B., Liao, S., Xu, K., Cheng, H., Zhu, C., Chen, H.: A novel mobile recommender systemfor indoor shopping. Expert Syst. Appl. 39(15), 11992–12000 (2012)

8. Colombo-Mendoza, L.O., Valencia-García, R., Rodríguez-González, A., Alor-Hernández, G.,Samper-Zapater, J.J.: RecomMetz: a context-aware knowledge-based mobile recommender system for movie showtimes. Expert Syst. Appl. 42(3), 1202–1222 (2015)

9. Chiu, P.-H., Kao, G.Y.-M., Lo, C.-C.: Personalized blog content recommender system formobile phone customers. Int. J. Hum. Comput. Stud. 68(8), 496–507 (2010)

10. Buettner, R.: A framework for recommender systems in online social network recruiting: an interdisciplinary call to arms. In: 47th Annual Hawaii International Conference on System Sciences, Big Island, Hawaii, pp. 1415–1424. IEEE (2014). doi:10.13140/RG.2.1.2127.3048

11. Chen, H., Gou, L., Zhang, X., Giles, C.: Collabseer: a search engine for collaborationdiscovery. In: ACM/IEEE Joint Conference on Digital Libraries (JCDL) (2011)

12. Felfernig, A., Isak, K., Szabo, K., Zachar, P.: The VITA financial services sales supportenvironment. In: AAAI/IAAI 2007, Vancouver, Canada, pp. 1692–1699 (2007)

13. Goel, A., Gupta, P., Sirois, J., Wang, D., Sharma, A., Gurumurthy, S.: The who-to-followsystem at Twitter: strategy, algorithms, and revenue impact. Interfaces 45(1), 98–107 (2015)

14. Kwon, H.-J., Hong, K.-S.: Personalized real-time location-tagged contents recommender

system based on mobile social networks. In: IEEE International Conference on ConsumerElectronics (ICCE), pp. 558–559. Las Vegas (2012)

15. Singh, S., Chana, I.: EARTH: energy-aware autonomic resource scheduling in cloudcomputing. J. Intell. Fuzzy Syst. 30(3), 1581–1600 (2016)

16. Singh, S., Chana, I.: Resource provisioning and scheduling in clouds: QoS perspective. J. Supercomput. 72(3), 926–960 (2016)

17. Singh, S., Chana, I.: QoS-aware autonomic resource management in cloud computing: asystematic review. ACM Comput. Surv. (CSUR) 48(3), 42 (2016)

18. Singh, S., Chana, I.: QRSF: QoS-aware resource scheduling framework in cloud computing.

J. Supercomput. 71(1), 241–292 (2015). Springer

19. Singh, S., Chana, I., Singh, M., Buyya, R.: SOCCER: self-optimization of energy-efficientcloud resources. Cluster Comput. 19, 1–14 (2016). doi:10.1007/s10586-016-0623-4. Springer

20. Jaccard, P.: Etude comparative de la distribution florale dans une portion des Alpes et du Jura.Impr. Corbaz (1901)

21. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. 17(6), 734–749 (2005). doi:10.1109/TKDE.2005.99

22. Sayyida, S.; Hartini, S.; Gunawan, S.; Husin, S.N. The Impact of the Covid-19 Pandemic on Retail ConsumerBehavior. Aptisi Trans. Manag. (ATM) 2021, 5, 79–88. [CrossRef]

2.3 Bhaskara, G.I.; Filimonau, V. The COVID-19 pandemic and organisational learning for disaster planning andmanagement: A perspective of tourism businesses from a destination prone to consecutive disasters. J. Hosp. Tour. Manag. 2021,46, 364–375. [CrossRef]

24. Nie, F.; Li, Z.; Wang, R.; Li, X. An Effective and Efficient Algorithm for K-means Clustering with NewFormulation. IEEE Trans. Knowl. Data Eng. 2022. Available online: https://ieeexplore.ieee.org/abstract/document/9723527/ (accessed on 2May 2022).

25. Brandtner, P.; Darbanian, F.; Falatouri, T.; Udokwu, C. Impact of COVID-19 on the customer end of retailsupply chains: A big data analysis of consumer satisfaction. Sustainability 2021, 13, 1464. [CrossRef]

26. Khong, D.W.K. Rents: How Marketing Causes Inequality by Gerrit De Geest. Asian J. Law Policy

27. Janardhanan, S.; Muthalagu, R. Market segmentation for profit maximization using machine learningalgorithms. J. Phys. Conf.
Ser. 2020, 1706, 012160. [CrossRef]

28. Dawane, V.; Waghodekar, P.; Pagare, J. RFM Analysis Using K-Means Clustering to Improve Revenue andCustomer Retention.
In Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021), Online, 29–30 April2021. [CrossRef]

29. Rachmawati, I.K. Collaboration Technology Acceptance Model, Subjective Norms and Personal Innovationson Buying Interest
Online. Int. J. Innov. Sci. Res. Technol. 2020, 5, 115–122.

30. Shirole, R.; Salokhe, L.; Jadhav, S. Customer Segmentation using RFM Model and K-Means Clustering. Int. J.Sci. Res. Sci. Technol.
2021, 8, 591–597. [CrossRef]

31. Ekren, B.Y.; Mangla, S.K.; Turhanlar, E.E.; Kazancoglu, Y.; Li, G. Lateral inventory share-based models forIoT-enabled E-commerce
sustainable food supply networks. Comput. Oper. Res. 2021, 130, 105237. [CrossRef]

32. Sinaga, K.P.; Yang, M.S. Unsupervised K-means clustering algorithm. IEEE Access 2020, 8, 80716–80727.[CrossRef]

33. Kuruba Manjunath, Y.S.; Kashef, R.F. Distributed clustering using multi-tier hierarchical overlay super-peerpeer-to-peer network
architecture for efficient customer segmentation. Electron. Commer. Res. Appl. 2021, 47, 101040. [CrossRef]Sustainability 2022, 14, 7243 15 of 15

34. Deng, Y.; Gao, Q. A study on e-commerce customer segmentation management based on improved K-meansalgorithm. Inf. Syst.
Bus. Manag. 2020, 18, 497–510. [CrossRef]

35. Suryadi, D.; Kim, H.M. A data-driven methodology to construct customer choice sets using online data andcustomer reviews.
J. Mech. Des. Trans. ASME 2019, 141, 111103. [CrossRef]

36. Miloudi, S.; Wang, Y.; Ding, W. A Gradient-Based Clustering for Multi-Database Mining. IEEE Access 2021,9, 11144–11172.
[CrossRef]

37. Mishra, S.K.; Dwivedi, V.; Sarvanan, C.K.; Pathak, K. Pattern Discovery in Hydrological Time Series DataMining during the
Monsoon Period of the High Flood Years in Brahmaputra River Basin. Int. J. Comput. Appl. 2013,