Customer Segmentation using Kmeans

submitted in fulfilment of the requirements for

Major-Project Sem-VII by

# Mr . Omkar Gavade (29) Mr. Yash Gawli (30) Mr. Vrushali Jagtap (37) Mr. Ashish Jangam (39)

Supervisor:

# Mr. Sandesh Patil



#### Department of Artificial Intelligence and Machine

**Learning**

#### VIDYA VIKAS EDUCATION TRUST’S

**UNIVERSAL COLLEGE OF ENGINEERING KAMAN, VASAI - 401208**

#### UNIVERSITY OF MUMBAI 2023-2024

**Vidya Vikas Education Trust’s Universal College of Engineering, Vasai (E)**

Department of Artificial Intelligence and Machine

Learning



**CERTIFICATE**

This is to certify that, the Major Project entitled **“Customer Segmentation using K-Means”** is the bonafide work of **Mr. Omkar Gavade (29) Mr. Yash Gawli (30), Ms**. **Vrushali Jagtap (37) and Mr. Ashish Jangam (39)** submitted to the University of Mumbai in fulfilment of the requirement for the Major Project-I Semester VII project work of B.E. Artificial Intelligence and Machine Learning at Universal College of Engineering, Vasai, Mumbai at the Department of Artificial Intelligence and Machine Learning, in the academic year 2023-2024,Semester – VII

Mr. Sandesh Patil

##### Supervisor

Mr. John Kenny

##### Head of Department

Dr. J.B. Patil

**Principal**

# Major Project-I Report Approval for B. E.

This project report entitled “**Customer Segmentation using K-Means**” by Mr. Omkar Gavade, Mr. Yash Gawli, Ms. Vrushali Jagtap and Mr. Ashish Jangam is approved for the Major Project-I Semester VII project work of B.E Artificial Intelligence and Machine Learning at Universal College of Engineering, Vasai, in the academic year 2023-2024.

##### Internal Examiner External Examiner

Date:

# Declaration

I declare that this written submission represents my ideas in my own words and where others’ ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

**Mr. Omkar Gavade (29)**

##### 

##### Mr. Yash Gawli (30)

**Ms. Vrushali Jagtap (37)**

##### Mr. Ashish Jangam (39)

Date:

# 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 report, 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. Then the segmentation is done to get 3D clusters. We have applied the K-means clustering algorithm to segment mall customers into distinct groups. By analyzing customer data, we identified the "Target Customers" who are most likely to easily converge with marketing efforts. These insights have empowered the marketing team to plan and implement tailored strategies, enhancing customer engagement and optimizing marketing campaigns for increased mall performance and competitiveness in the retail sector.

***Keywords- data segmenting; machine learning; numpy; customer segment; k-Means algorithm; sklearn.***

|  |  |
| --- | --- |
| **Contents** |  |
| Abstract  List of figures List of tables  List of abbreviations | i iv v vi |
| **1 INTRODUCTION** | **1** |

1.1 Project Overview

1. **REVIEW OF LITERATURE 3**
   1. Existing System
   2. Literature Survey
   3. Problem Statement and Objective
   4. Scope

#### PROPOSED SYSTEM 10

* 1. Analysis/Framework/Algorithm
  2. System Requirements
     1. Hardware Requirements
     2. Software Requirements
  3. Design Details
     1. System Architecture
     2. System Modules
  4. Data Model and Description
     1. Entity Relationship Model
  5. Fundamental Model
     1. Data Flow Model
  6. Unified Modelling Language Diagram
     1. Use Case Diagram
     2. Activity Diagram
     3. Sequence Diagram
  7. Methodology

1. RESULT 25
   1. [Proposed System Result](#_bookmark0)
   2. Existing and proposed system

[Conclusion 33](#_bookmark1)

[Appendix 34](#_bookmark2)

[References 35](#_bookmark3)

[Acknowledgement 37](#_bookmark4)

#### List of Figures

* 1. System Architecture 13
  2. Entity Relationship Diagram 17
  3. Data Flow Diagram 19
  4. Use Case Diagram 20
  5. Activity Diagram 21
  6. Sequence Diagram 22
  7. Data Exploration 24
  8. Data Types 24
  9. Data Visualization 25
  10. Count Plot 25
  11. Relation between Variables 26
  12. Scatter Plot 26
  13. Boxplot and Swarmplot 27
  14. Clusters based on Inertia 27
  15. Scatterplot of Cluster 28
  16. Segmented 3D Cluster 28

#### List of Tables

2.1 Literature Survey 4

#### List of abbreviations

1. sk-learn – Scikit Learn
2. UI – User Interface
3. IEEE - Institute of Electrical and Electronics Engineers
4. RAM – Random Access Memory
5. IDE – Integrated Development Environment
6. DFD – Data Flow Diagram
7. UML – Unified Modelling Language
8. ER – Entity Relationship

# Chapter 1 Introduction

In this digital age, businesses are inundated with vast amounts of data about their customers. Customer segmentation provides a structured framework for transforming this data into actionable intelligence. Whether you're a small e-commerce startup or a global conglomerate, understanding the unique attributes, preferences, and behaviours of your customers is paramount for success. This process allows you to move beyond generic mass marketing and adopt a more customer-centric approach, ultimately fostering stronger, more enduring relationships.

In this journey of customer segmentation, we will explore the various methods, variables, and strategies that enable businesses to divide their customer base into meaningful segments. We'll delve into the art and science of creating customer profiles, tailoring marketing strategies, and, most importantly, the ongoing process of adaptation to meet the ever-changing dynamics of the market and customer preferences.

## Project Idea

Customer segmentation serves as a powerful motivator for businesses, primarily because it enables a level of personalization that enhances customer engagement and satisfaction. By categorizing customers into distinct groups based on shared characteristics, businesses can tailor their products, services, and marketing efforts to precisely meet the unique needs and preferences of each segment. This not only fosters a deeper connection with customers but also drives increased relevance, attracting and retaining a more loyal customer base. Moreover, segmentation enables businesses to allocate their marketing resources more efficiently, concentrating their efforts on the segments most likely to convert, thus optimizing their return on investment.

Additionally, customer segmentation acts as a catalyst for better customer retention. Understanding and addressing the specific pain points and desires of different customer groups leads to higher customer satisfaction and loyalty. As a result, businesses can sustain their relationships with existing customers, ultimately increasing customer lifetime value. This strategic approach also empowers companies to gain a competitive advantage. When they provide a more personalized and relevant experience compared to competitors, they are better positioned to acquire and retain customers, th1us staying ahead in the market.

# Chapter 2 Review of Literature

A literature survey was carried out to find various papers published in international journals such as IEEE etc. related to tracing missing people using facial recognition to get the best algorithm for the same.

## 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 analyze 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

## Literature Survey

We have examined various research papers in the domain of Customer Segmentation for our project to delve deeper into the details of the various researches conducted in the field of Machine Learning. Table 2.1 shows survey of the research paper done for the project.

Table 2.1 – Literature Survey table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sr. no | Paper Name | Year of Publication | Author | Publication | ProposedWork | ResearchGap |
| 1. | Customer Segmentat ion using Machine Learning. | 2021 | AMAN BANDHUNI,  Prof ILAVENDHAN | IJCRT | In this research K-means algorithm is applied to customer data, and a scalar library for K-Means has been developed. This work aims to provide actionable insights for businesses seeking to better understand and serve their diverse customer base through effective data mining  and segmentation techniques. | The research gap in the provided context lies in the absence of a clear methodology or evaluation of the internal clustering validation techniques used to segment customers. |
| 2. | Finding Best Possible Number of Clusters using K- Means Algorithm | 2019 | K. Maheswari | IJEAT | In this research, the focus is on applying clustering, an unsupervised learning algorithm, to analyze unlabeled data and find similarities among data points. The primary emphasis is on measuring the distance between data points and the significance of clustering to unveil hidden relationships between them. | This research highlights the need for further exploration of advanced algorithms to improve the accuracy and efficiency of clustering in online retail settings, ensuring customers are better catered to through data-driven strategies. |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 3. | Customer Segmentation for Life Insurance in Iran Using K- means Clustering | 2021 | Farzan Khamesia n, Farbod Khanizade ha, Alireza Bahiraieb | IJNAA | The proposed work centers on enhancing the field of life insurance through effective marketing strategies and customer segmentation. | This evolves in more in depth exploration of segmentation because the penetration rate is one of the most strategic plans for insurance companies |
| 4. | Customer Segmentation using RFM Model and K- Means Clustering | 2021 | Rahul Shirole,La xmiputra Salokhe, Saraswati Jadhav | IJSRST | This paper is to segment customer based on transaction from a supermarket which include 200 data entries. Here the clustering was done tohelp retail industry to develop new market strategy | It evolves of more exploration of other techniques that could benefit from comparative analysis with other segmentation methods. |
| 5. | A review on customer segmentation methods for personalized customer targeting in e‑commerce use cases. | 2023 | Miguel Alves Gomes1  ·Tobias Meisen1 | ICE | This research mainly focuses on increasing the precision and efficiency of customer analysis. | This gap highlights the need for standardized evaluation metrics in customer segmentation and emphasizes the exploration of advanced techniques for more effective customer segmentation. |
| 6. | Customer Segmentation Based on RFM Model Using K-Means, K- Medoids, and DBSCAN  Methods | 2021 | Rahma Wati Br Sembiring Berahman, Fahd Agodzo Mohamme  .,  Kankamol Chairuang | LKJITI | Here customer segmentation is done by in putting the annual transaction, the clustering is done by using k-means , K- medoids , etc. The results of clustering is a grouped data based on five customer labels, namely Superstar, Golden, Every Day, Occasional and Dormant. | The research gap emphasizes the importance of conducting multiple experiments with varying cluster numbers to explore the full range of customer classes and identify the best- suited clustering technique. |
| 7. | RFM model for customer purchase behavior using K-Means algorithm. | 2019 | P. Anitha, Malini M.Patil | ISE | This primarily focuses on customer clustering and segmentation in the retail industry using the K-Means algorithm based on RFM(Recency,Frequency,Monet ary)values. The key objectives are to track historical purchasing behavior, maximize sales in specific areas, and design effective sales and marketing  strategies. | The research gap in this context revolves around the need for more specialized and detailed studies within specific product categories, like Mobile Phones and Accessories. |
| 8. | A Hybrid Method for Customer  Segmentation | 2022 | Abdullah Alghamd | ICRT | This research aims to develop a new data-driven approach to reveal customers’ satisfaction in  restaurants. In this research, the | Various text mining approaches can be incorporated for customer  satisfaction analysis and |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | in Saudi Arabia Restaurants Using Clustering, Neural Networks  and Optimization Learning  Techniques |  |  |  | data of customers on the service quality of restaurants are collected from the TripAdvisor platform and the result evaluates the prediction model through a set of evaluation metric. | segmentation. |
| 9. | An Extended Regularized K- Means Clustering Approach for High- Dimensional Customer Segmentation With Correlated  Variables | 2021 | HONG- HAO ZHAO, XI-CHUN LUO, RUI MA , AND XI LU | IEEE | In this the proposed regularized technique is introduced instead of the traditional clustering method.The proposed methods aim to provide sensible clustering results in high- dimensional applications with correlated variables. | This research fails to discuss the tuning parameter selection for α in the LEN k- means method. However,the flexibility of the LEN k- means method is obtained using such parameters. |
| 10. | An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K-  Means Algorithm | 2023 | Jun Wu, Li Shi, Wen-Pin Lin ,Sang- Bing Tsai, Yuanyuan Li,Liping Yang, and  Guangshu X | IJRCT | In this paper, customer purchase behaviors are analyzed systematically based on the online transaction data of a company by using RFM and K- means clustering  algorithm and different CRM strategies are proposed to gain a high level of customer  satisfaction. | More suitable algorithms are needed to match the new datasets. |

In the paper presented by AMAN BANDUNI, Prof ILAVENDHAN, clustering is performed by applying k means algorithm to customer data, and a scalar library for K-Means has been developed. In this training was conducted using a silhouette score with two feature sets, resulting in the identification of four distinct customer segments based on factors like the number of items purchased and the average number of customers per month. These segments are labeled as cluster\_metrics\_1, cluster\_metrics\_2, cluster\_metrics\_3, and cluster\_metrics\_4. This work aims to provide actionable insights for businesses seeking to better understand and serve their diverse customer base through effective data mining and segmentation techniques.[1].

In the paper presented by K. Maheshwari, the focus is on applying clustering which is an unsupervised learning algorithm, to analyze unlabeled data and to find similarities among data points. The primary emphasis is on measuring the distance between data points and the significance of clustering to unveil hidden relationships between them. The data used in this study is sourced from an internet repository, with an understanding that clean and high-quality data facilitates model construction, ultimately improving performance. To extract meaningful value from unstructured data, both classification and clustering techniques are employed.[2].

In the paper presented by Farzan Khamesian, Farbod Khanizadeha, Alireza Bahiraieb, centers on enhancing the field of life insurance through effective marketing strategies and customer segmentation. By applying data mining techniques and clustering algorithms, the research aims to gain a deeper understanding of customer needs and behaviors. This insight will enable the creation of tailored products and services, aligning with various budgets and eliminating the need for physical exams in some cases. [3].

In the l paper presented by Rahul Shirole, Laxmiputra Salokhe, Saraswati Jadhav ,segments the customer based on transaction from a supermarket which include 200 data entries. In this, the clustering was done in order to help the retail industry to develop new market strategy . This paper uses Hierarchical clustering algorithm which does not prerequire information about no of clusters required. [4].

In the paper presented by Miguel Alves Gomes1 · Tobias Meisen1, the research methodology is focused on developing robust metrics for evaluating the quality of segmentation algorithms, striving to establish consensus in this area. The aim is to improve the precision and efficiency of customer analysis and targeting strategies. The research can build upon the literature overview and insights provided, contributing to a more comprehensive understanding of customer behavior analysis and segmentation techniques. [5].

In the paper presented by Rahma Wati Br Sembiring Berahman, Fahd Agodzo Mohamme. Kankamol Chairuang, segmentation is done by inputting the annual transaction. Here the clustering is done by using k-means, K-medoids, etc. The group validation process is done using the Davies-Bouldin Index and the Silhouette Index Method. Then the data modeling process is based on the results obtained from the data modeling process. The results of clustering will group data based on five customer labels, namely Superstar, Golden, Every Day, Occasional and Dormant.[6].

In the paper presented by P. Anitha, Malini M. Patil primarily focuses on customer clustering and segmentation in the retail industry using the K-Means algorithm based on RFM (Recency, Frequency, Monetary) values. The key objectives are to track historical purchasing behavior, maximize sales in specific areas, and design effective sales and marketing strategies. [7].

In the paper presented by Abdullah Alghamd, aims to develop a new data-driven approach to reveal customers’ satisfaction in restaurants. Specifically, k-means and Artificial Neural Network (ANN) with the aid of the Particle Swarm Optimization (PSO) technique are, respectively, used in data clustering and prediction tasks. In this research, the data of customers on the service quality of restaurants are collected from the TripAdvisor platform. The results of the data analysis are provided. We evaluate the prediction model through a set of evaluation metric.[8].

In the paper presented by HONG-HAO ZHAO, XI-CHUN LUO, RUI MA, and XI LU a proposed regularized technique is introduced instead of the traditional clustering method. The proposed methods aim to provide sensible clustering results in high-dimensional applications with correlated variables. [9].

In the paper presented by Jun Wu, Li Shi, Wen-Pi n Lin , Sang-Bing Tsai, Yuanyuan Li,Liping Yang, and Guangshu X customer purchase behaviors are analyzed systematically based on the online transaction data of a company by using RFM and K means clustering algorithm and different CRM strategies are proposed to gain a high level of customer satisfaction. [10].

## Problem Statement and Objective

To provide classification for the customer preferences and their respective intents. Businesses often face the challenge of trying to connect with diverse customer groups, each with its unique needs, preferences, and behaviors. Without a clear customer segmentation strategy, companies find it challenging to deliver personalized experiences, efficient marketing efforts, and tailored products and services.

## Project Scope

The future scope for integrating K-means clustering with neural networks holds significant promise in the field of unsupervised learning and data analysis. This hybrid approach, often referred to as neural K-means or deep clustering, is expected to become more prevalent in various applications. It leverages the strengths of K-means, a traditional clustering algorithm, and neural networks' capacity to handle complex data representations. As neural network architectures continue to advance, they will offer more efficient and effective feature extraction and dimensionality reduction, complementing K-means in partitioning data into meaningful clusters. This synergy will find applications in diverse domains, from image and speech processing to recommendation systems and anomaly detection.

# Chapter 3 Proposed System

This chapter includes a brief description of the proposed system and explores the different modules involved along with the various models through which this system is understood and represented.

## Analysis/Framework/ Algorithm

KMEANS is one of the most basic classification algorithms in machine learning. It belongs to the supervised learning category of machine learning. k-means is often used in search applications where you are looking for “similar” items. The way we measure similarity is by creating a vector representation of the items, and then compare the vectors using an appropriate distancemetric (like the Euclidean distance, for example).

It is generally used in data mining, pattern recognition, recommender systems and intrusion detection. K-MEANS classifier is best suited for classifying persons based on their images due to its lesser execution time and better accuracy than other commonly used methods which include Hidden Markov Model and Kernel method. Although methods like SVM and Adaboost algorithms are proved to be more accurate than K-MEANS classifier, K-MEANS classifier has a faster execution time and is dominant than SVM. The simplest classification scheme is a nearest neighbour classification in the image space. Under this scheme an image in the test set is recognized by assigning to it the label of the closest point in the learning set, where distance is measured in image space. The Euclidean distance metric is often chosen todetermine the closeness between the data points in K-MEANS. A distance is assigned between all pixels in a dataset. Distance is defined as the Euclidean distance between two pixels.

K-means is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. The k-Means algorithm is amongst the simplest of all machine learning algorithms: an object is classified by

a majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbours (k is a positive integer, typically small). If k

= 1, then the object is simply assigned to the class of its nearest neighbour.

Step 1: Calculate d(x, xi) where i = 1 to n and d = Euclidean distance between the points. Step 2: Arrange the calculated n Euclidean distances in non-decreasing order.

Step 3: Let k be a +ve integer, take the first k distances from this sorted list. Step 4: Find those k-points corresponding to these k-distances.

Step 5: Let ki = the number of points belonging to the ith class among k points i.e. k ≥ 0 Step 6: If ki >kj ∀ i ≠ j then put x in class i

## System Requirements

This section will provide the user the required specification of the hardware and software components on which the proposed system is to be implemented.

#### Hardware Requirements

This subsection will provide the minimum requirements that must be fulfilled by the hardware components. The hardware requirements are as follows: -

* + - A desktop with
      1. RAM – minimum 4 gigabytes
      2. Storage – minimum 100 gigabytes
      3. Processor – minimum quadcore or hexacore

#### Software Requirements

This subsection will provide the versions of software applications that must be installed.The software requirements are as follows: -

* + - Python 3.6
    - Python IDE
    - Google Collab
    - VS studio
    - Internet connection

## Design Details

In design details, we analyze the System Architecture and System Modules in detail. We study the flow and process of the entire project in order to develop the project in an orderly and systematic manner. There are 3 modules in Customer Segmentation using K-means - Loading Dataset, Processing and Segmentation ,Clustering and Data Visualization

## 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 calculated through randomly chosen centers and repeat it until there is no change in cluster centers. Thereafter we will analyze the data through data visualization. Finally, we will get the outcome.

A. Algorithm

1. K-Means Clustering
   1. 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.
   2. Here K is the total no of clusters.
   3. Every point belongs to only one cluster.
   4. Clusters cannot overlap.
2. Steps of Algorithm
3. Arbitrarily choose k objects from D as the initial cluster centers.
4. Repeat.
5. 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.
6. Update the cluster means, i.e. calculate the mean value of the objects for each cluster.
7. Until no change.

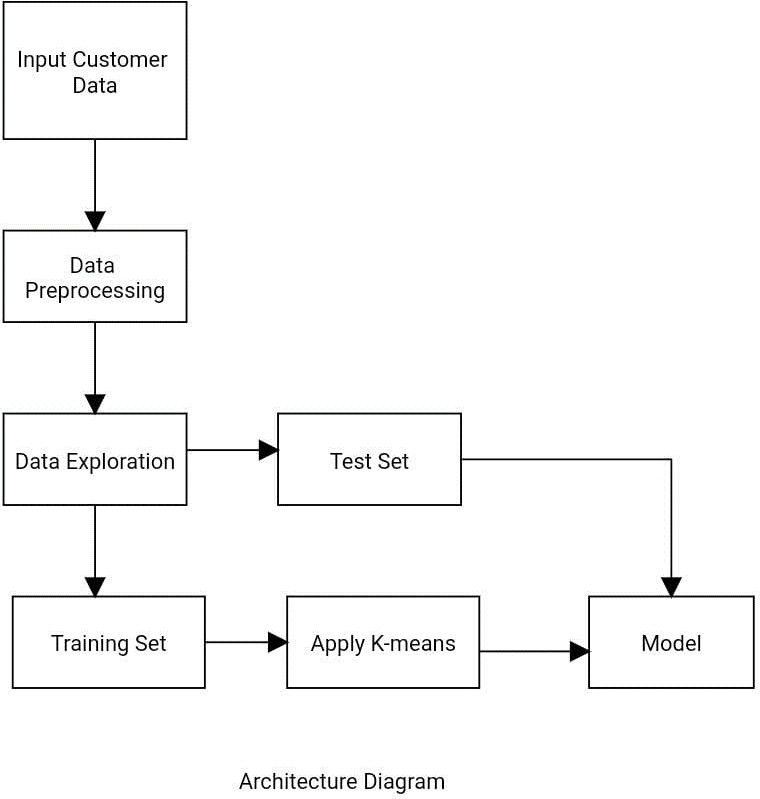


Figure. 3.1 – System Architecture

## Details of Modules

The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business

The modules are:

1. Loading Dataset
2. Processing and Segmentation
3. Clustering
4. Data Visualization

##### Loading Dataset

Loading a dataset for customer segmentation using the K-means algorithm involves the initial steps of acquiring, preparing, and structuring the data to facilitate the clustering process. Data is gathered from various sources, including customer records, transaction data, and relevant attributes. This data is then cleaned to remove inconsistencies, duplicates, and missing values, ensuring data quality. Afterward, the dataset is transformed, which might involve standardizing numerical values and encoding categorical variables.

##### Processing and Segmentation

In the realm of customer segmentation using the K-means algorithm, the process involves two key phases: data processing and segmentation. Data processing is the initial step where raw customer data is collected, cleaned, and preprocessed. This involves gathering information from multiple sources, cleaning the data to eliminate errors and inconsistencies, and transforming it to make it suitable for analysis. Categorical variables may be encoded, and numerical values may be standardized to ensure data quality and consistency.

Once the data is prepared, the segmentation phase begins. In this phase, the K-means algorithm is applied to group customers into clusters based on their similarities. The number of clusters (K) is determined based on the business objectives and can be established using techniques like the elbow method. K-means works by iteratively assigning customers to clusters and recalculating cluster centroids until convergence is reached. The result is a set of distinct customer segments, each characterized by shared attributes or behaviors.

##### Clustering

K-means iteratively assigns each customer to the cluster whose centroid is closest to them, and it recalculates the centroids until convergence. The result is a set of customer segments, where each group shares similar characteristics or behaviors. These segments can provide valuable insights for businesses, helping them tailor their marketing efforts, product offerings, and customer experiences to different customer preferences. However, it's crucial to continually monitor and update the clustering model as customer behaviors and preferences change over time, ensuring that the segmentation remains relevant and effective for strategic decision-making and customer engagement.

##### Data Visualization

Data visualization plays a crucial role in the process of customer segmentation using the K-means algorithm. It serves as a powerful tool for understanding and presenting the results of clustering analysis. After preprocessing and clustering the customer data, visualizing the segments helps convey the insights derived from the K-means algorithm in a comprehensible manner. Techniques such as scatter plots, heatmaps, and bar charts can be employed to represent the clusters and their characteristics. These visualizations enable businesses to see how customers are grouped and identify the distinguishing features or behaviors of each segment.

## Data Model and Description

Data Model describes the relationship and association among data which includes Entity Relationship Model.

#### Entity Relationship Model

Figure 3.4 shows the Entity Relationship Diagram of the proposed system. Entity Relationship diagram is a data modelling technique that graphically illustrates an information system’s entities and the relationships between those entities. Here, the entities are: - Customer, Data Analyst, Data Warehouse Manager and Marketing Analyst. The diagram shows the different attributes of these entities and also shows the relationship among these different entities.

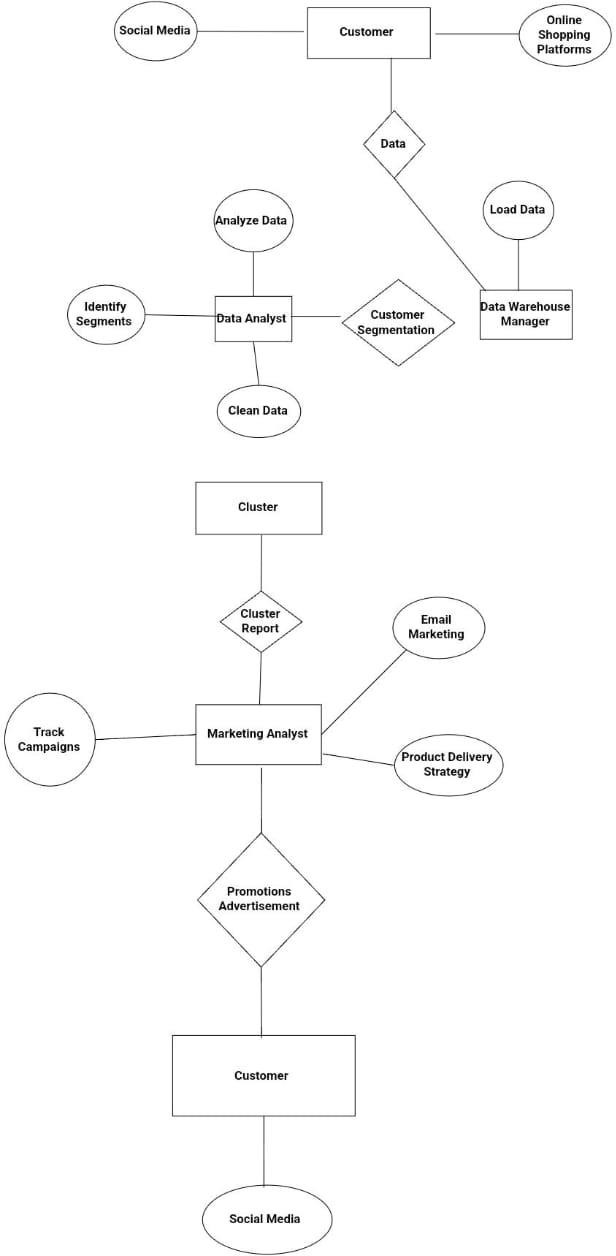


Figure 3.2- Entity Relationship Diagram

## Fundamental Model

Fundamental model of the project gives overall idea about the project. How the entities are related to each other, what are the attributes of the entities, how the data flows between the entities is shown by the fundamental model.

#### Data Flow Model

Data Flow Diagram (DFD) shows graphical representation of the” flow” of data through an information system, modelling its process aspects. It includes data inputs and outputs, data stores, and the various subprocesses the data moves through. DFDs are built using standardized symbols and notation to describe various entities and their relationships.

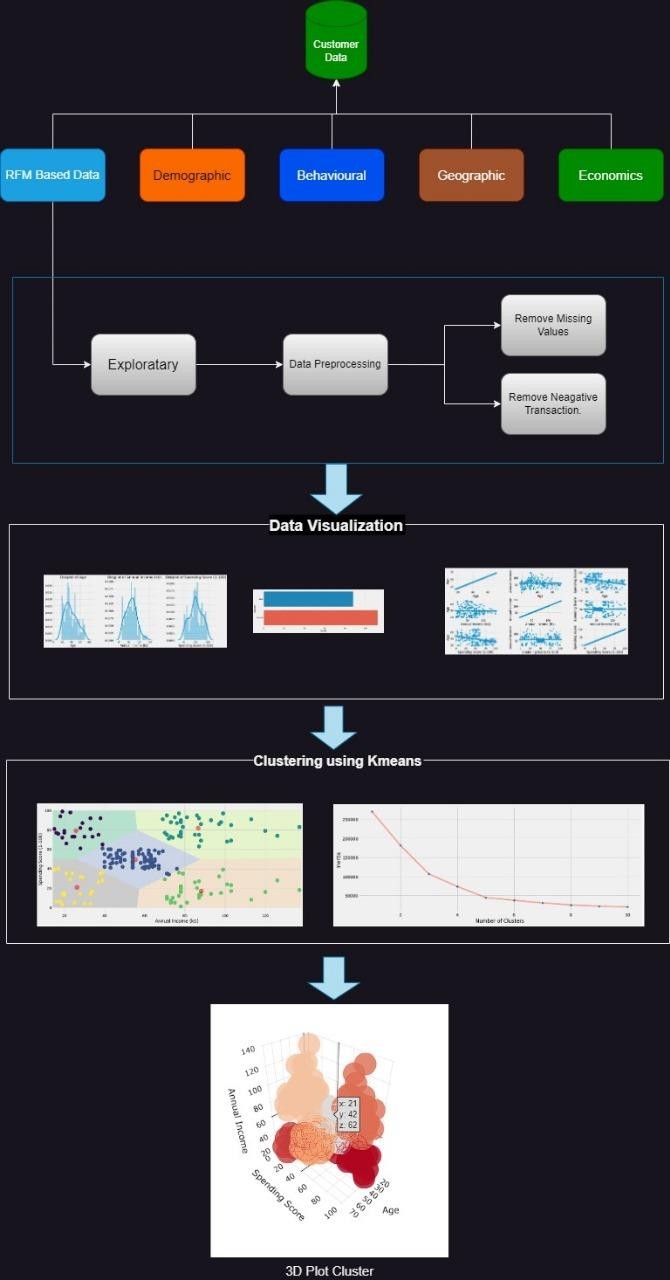


Figure 3.3 – Dataflow Diagram

## UML (Unified Modelling Language) Diagram

The Unified Modelling Language is a general-purpose, developmental, modelling language in the field of software engineering that is intended to provide a standard way to visualizethe design of a system. We have prepared and designed the UML diagrams of – Use Case, Activity, Component, Deployment and Sequence Diagrams.

## Use Case Diagram

Figure 3.8 denotes the Use Case Diagram of the proposed system. It shows the user’sinteraction with the systems. The purpose of a use case diagram in Unified ModellingLanguage (UML) is to demonstrate the different ways that a user might interact with a system. In this use case diagram, there are three actors involved, the first actor is user, the second actor is police officer and the third actor is facial recognition system. It depicts the interactions between the various actors used in this system. All these interactions between actors and system is done in the Machine Learning environment. The various use cases involved in this system such as upload photo and view status whose association is between user and facial recognition system and police officer respectively.

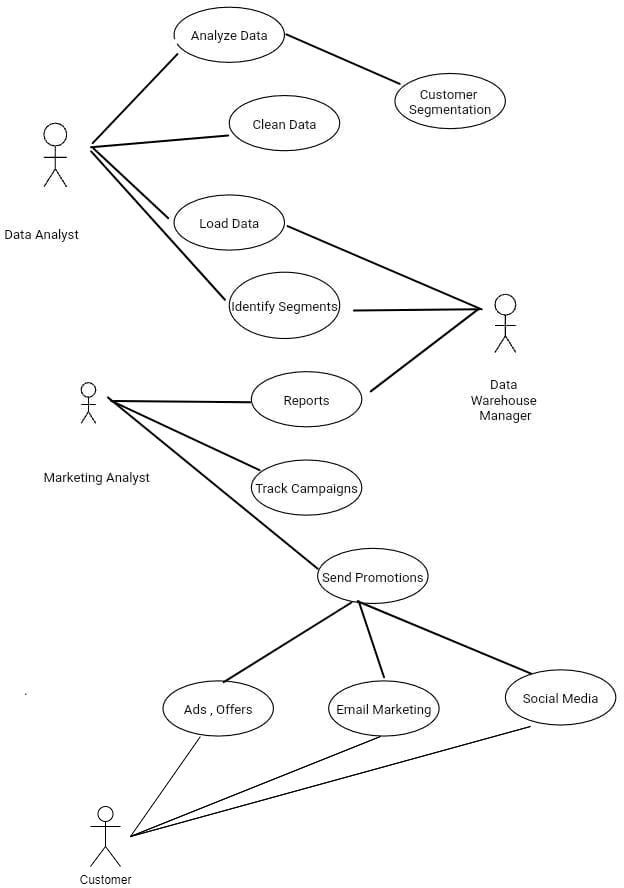


Figure 3.4 – Use Case Diagram

## Activity Diagram

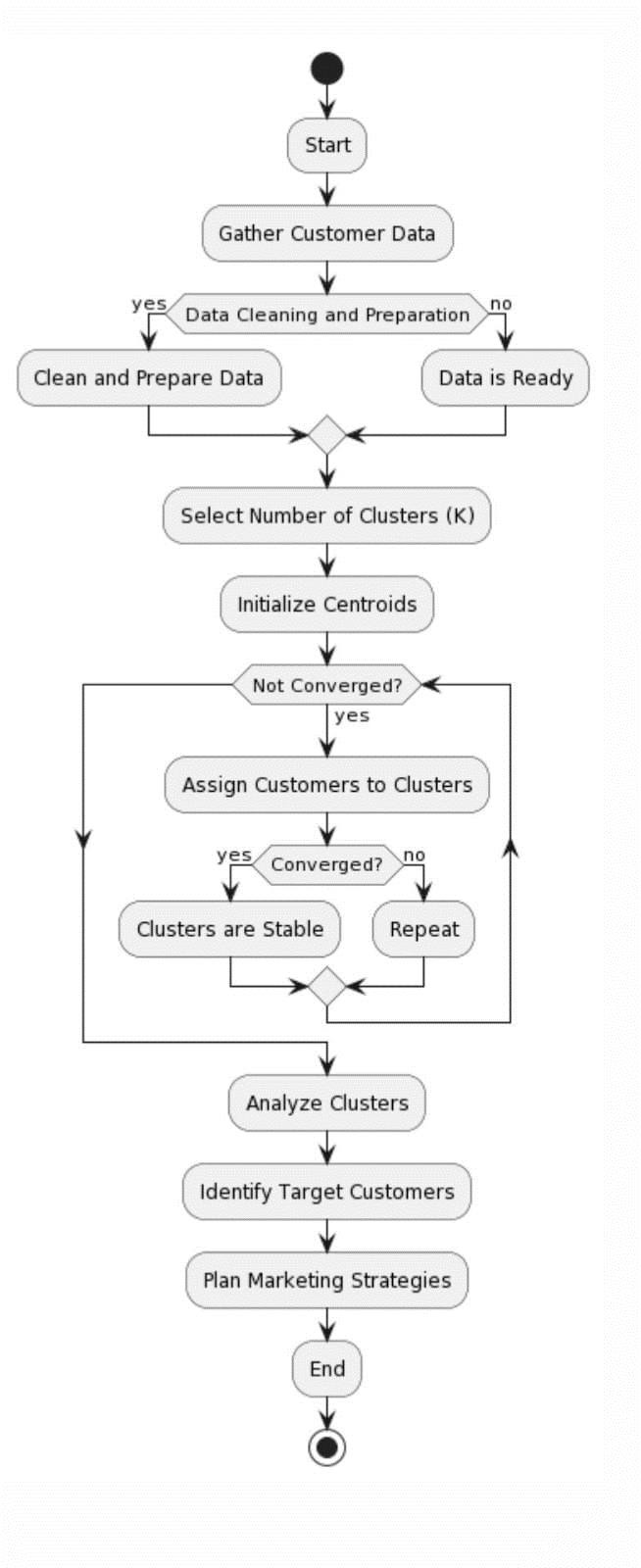


Figure 3.5 – Activity Diagram

This process involves several key steps to better understand and target different customer segments. Initially, data is collected and prepared, encompassing details such as demographic information, purchase history, and other relevant features. Subsequently, features are selected, highlighting the most significant ones for the segmentation. The pivotal decision of determining the number of clusters, often referred to as "K," is made, typically with the assistance of statistical methods like the Elbow Method or Silhouette Score. The K-means clustering algorithm is then applied, dividing customers into these clusters, enabling the identification of common traits and behaviors within each group. This understanding allows for the creation of distinct marketing strategies, promotions, and recommendations tailored to each segment. The results are continuously monitored, and strategies are adjusted as needed. The process is iterative, ensuring that customer segmentation remains relevant as customer preferences and business goals evolve. Overall, the activity of customer segmentation is a strategic approach to improving marketing efforts and enhancing customer satisfaction by targeting and understanding distinct customer groups.

## Sequence Diagram

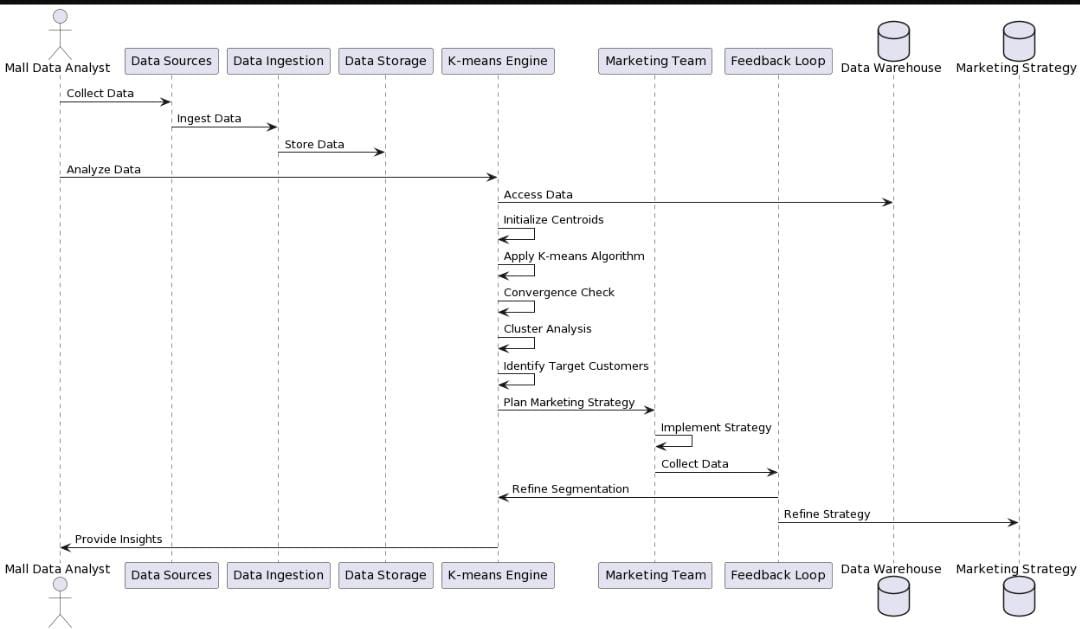


Figure 3.6 – Sequence Diagram

A sequence diagram is a visual representation used in software engineering and systems analysis to illustrate the dynamic behavior of a system or a software application. It provides a chronological and detailed view of how various objects or components interact and exchange messages over time. In a sequence diagram, objects are represented as lifelines, and their interactions are depicted through messages, which can be synchronous or asynchronous. These diagrams are particularly useful for modeling and understanding the sequence of events and the order in which different parts of a system collaborate to achieve a specific task. Sequence diagrams play a crucial role in the design, documentation, and communication of complex systems, helping developers and stakeholders visualize the flow of control and data within a software application and ensuring that the system functions as intended.

## Methodology

Customer segmentation using K-means clustering is a common approach in marketing and data analysis to group customers with similar characteristics together. K-means is an unsupervised machine learning algorithm that tries to partition a dataset into K distinct, non-overlapping subgroups (clusters). Here is a step-by-step methodology for customer segmentation using K- means:

* Dataset used:

We used mall dataset customer.csv that is well accessible online and additionally you'll be able to download it.

* Seaborn and sklearn:

Seaborn is a library for making statistical graphics in Python.

Scikit-Learn, also known as sklearn is a python library to implement machine learning models and statistical modelling. Through scikit-learn, we can implement various machine learning models for regression, classification, clustering, and statistical tools for analyzing these models.

* Clustering:

This includes 3 Python files where the primary one is employed to store the ‘.csv’ file format, second is employed to process and segment the data and the third one is employed visualized the data.

* Methodology / Approach

1. First of all we will import all the necessary libraries or modules (pandas, numpy, seaborn).
2. Then we will read dataset and analyze 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
   * Technologies Used

It is generally used in visualization. Following are list of libraries used:

* 1. Seaborn
  2. Pandas
  3. NumPy
  4. Scikit-Learn

# Chapter 4

**Result and Discussion**

### Proposed System Result

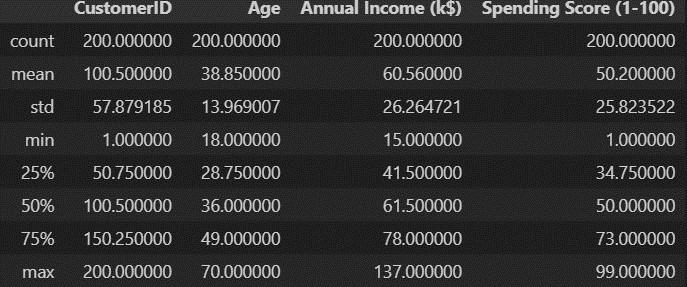


Fig: 4.1 Data Exploration

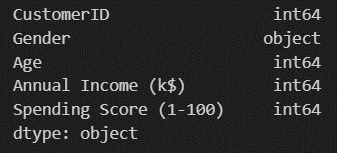


Fig:4.2 Data Types

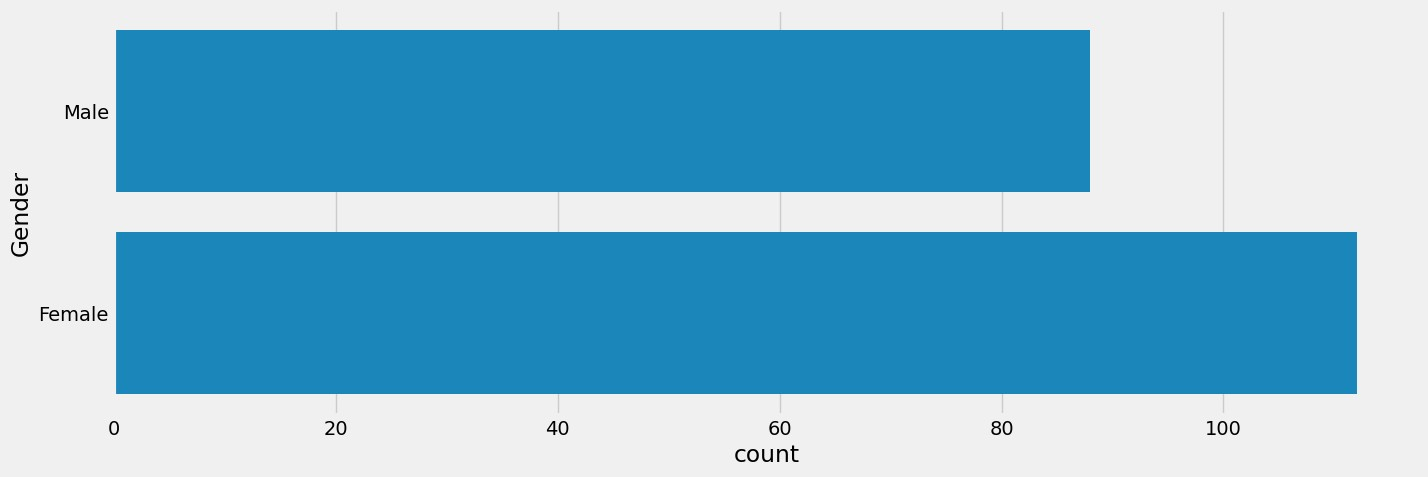
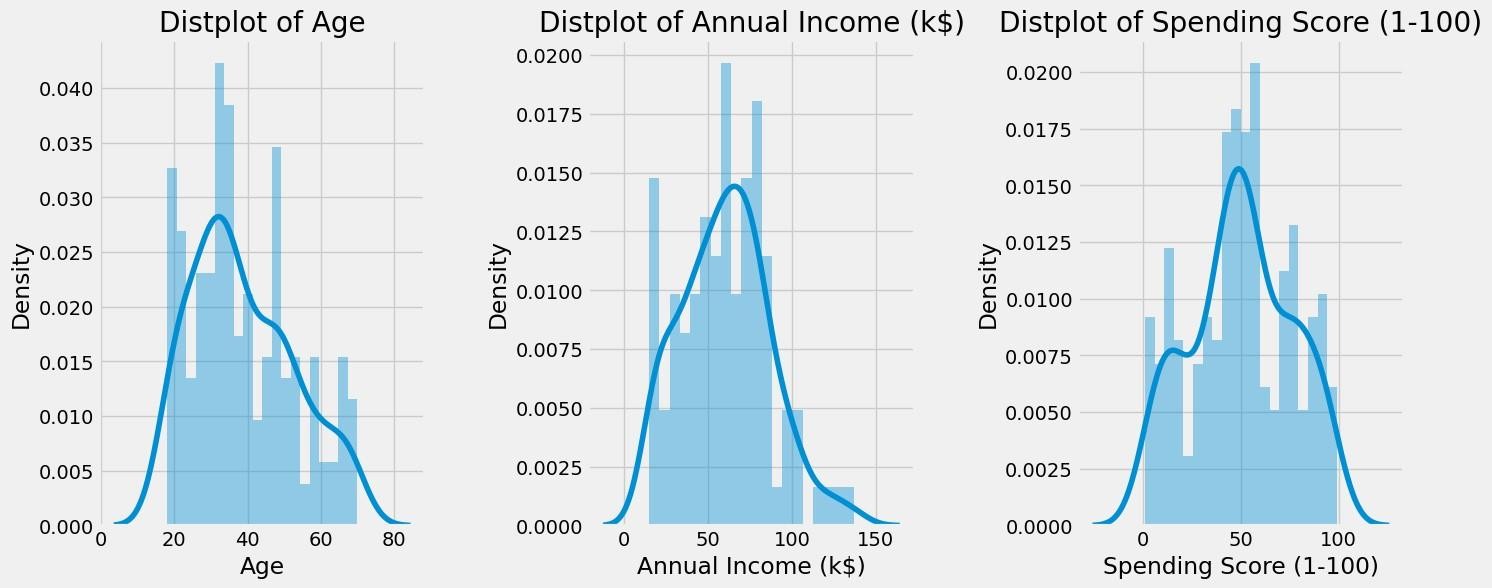


Fig: 4.3 Data Visualization

Fig: 4.4 Count Plot

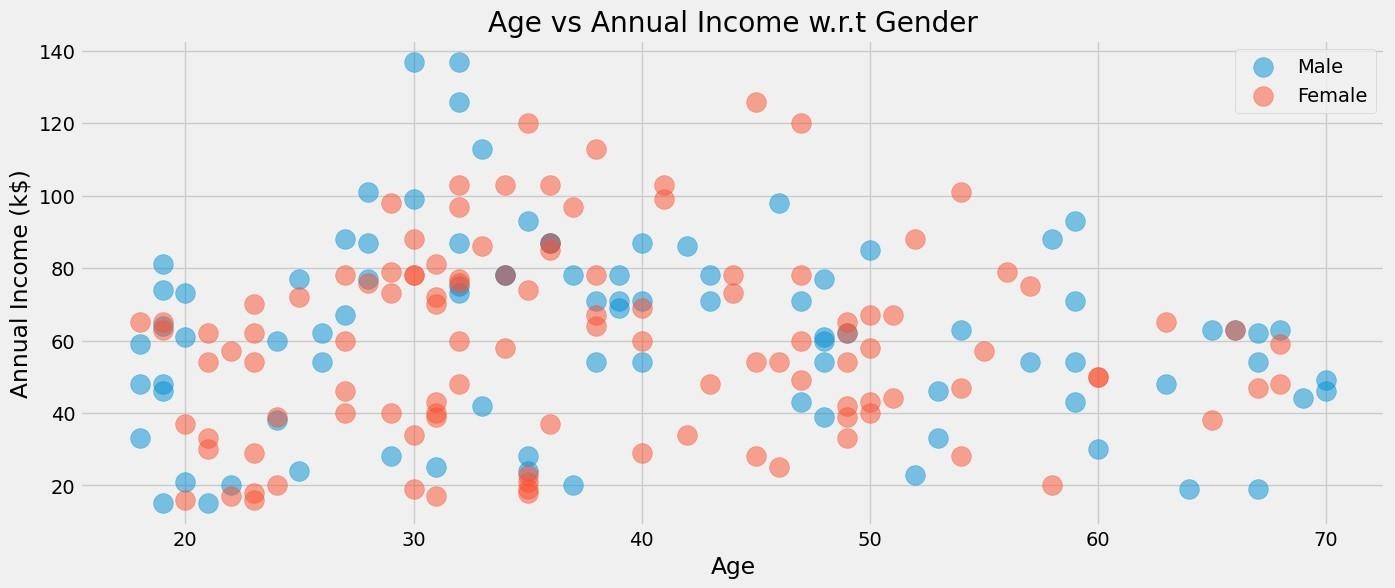
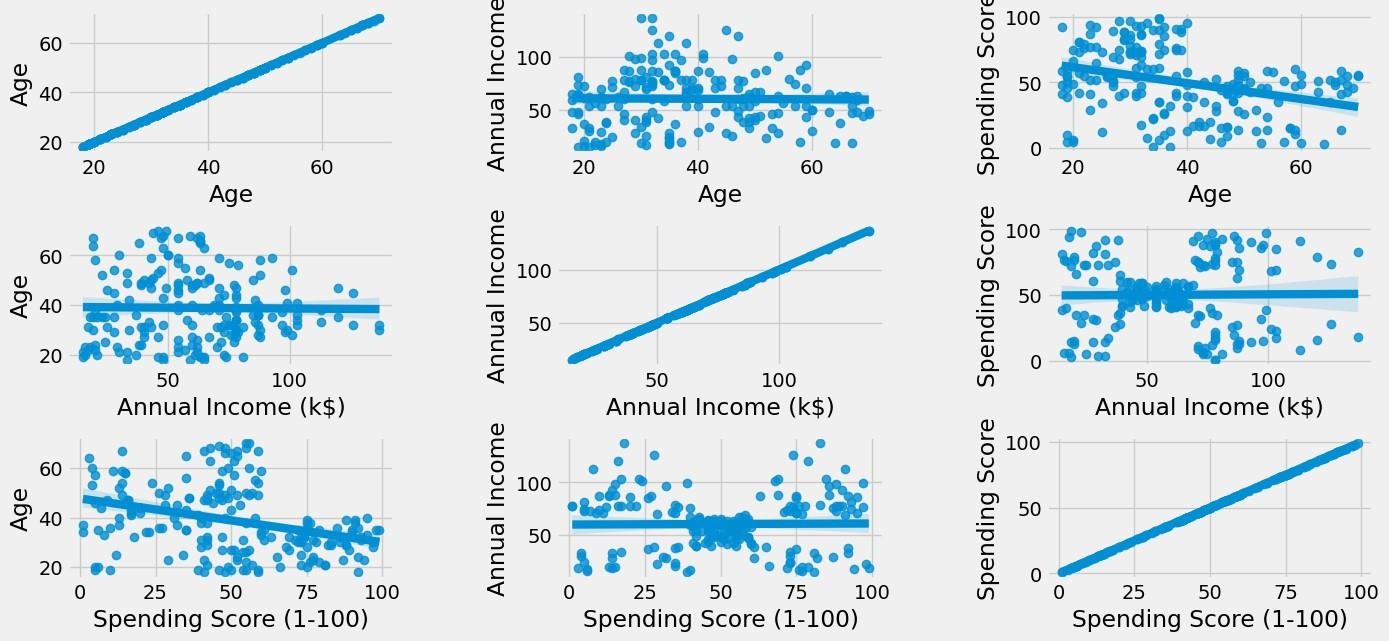


Fig: 4.5 Relation Between Variables

Fig 4.6 Scatter Plot

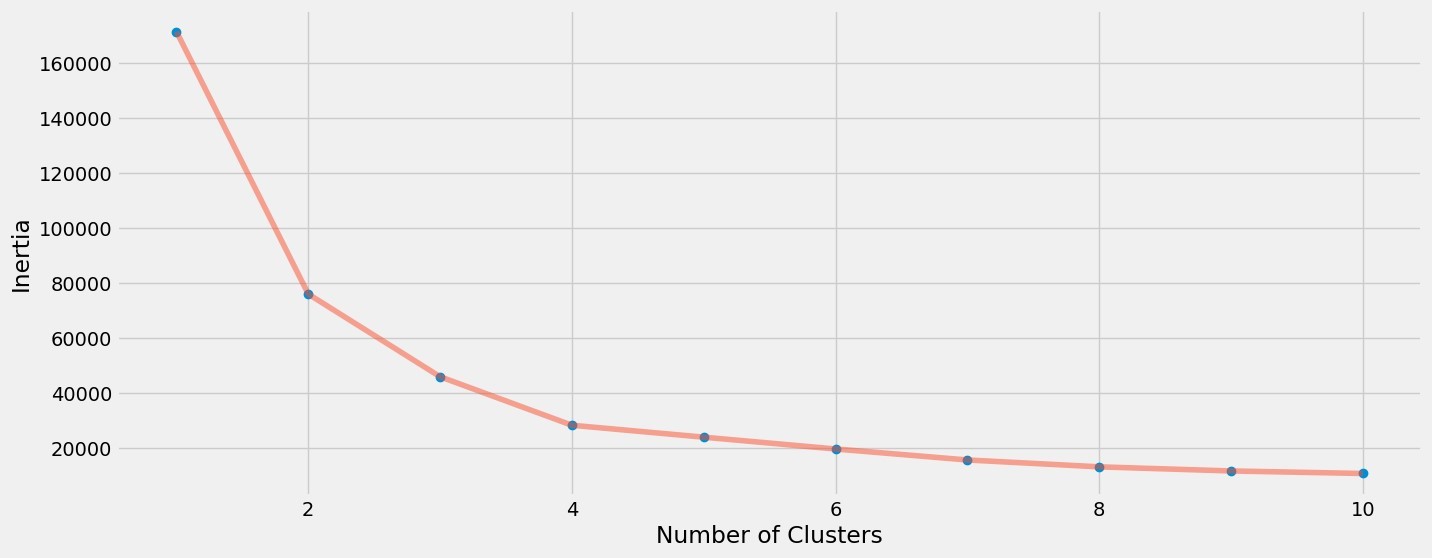
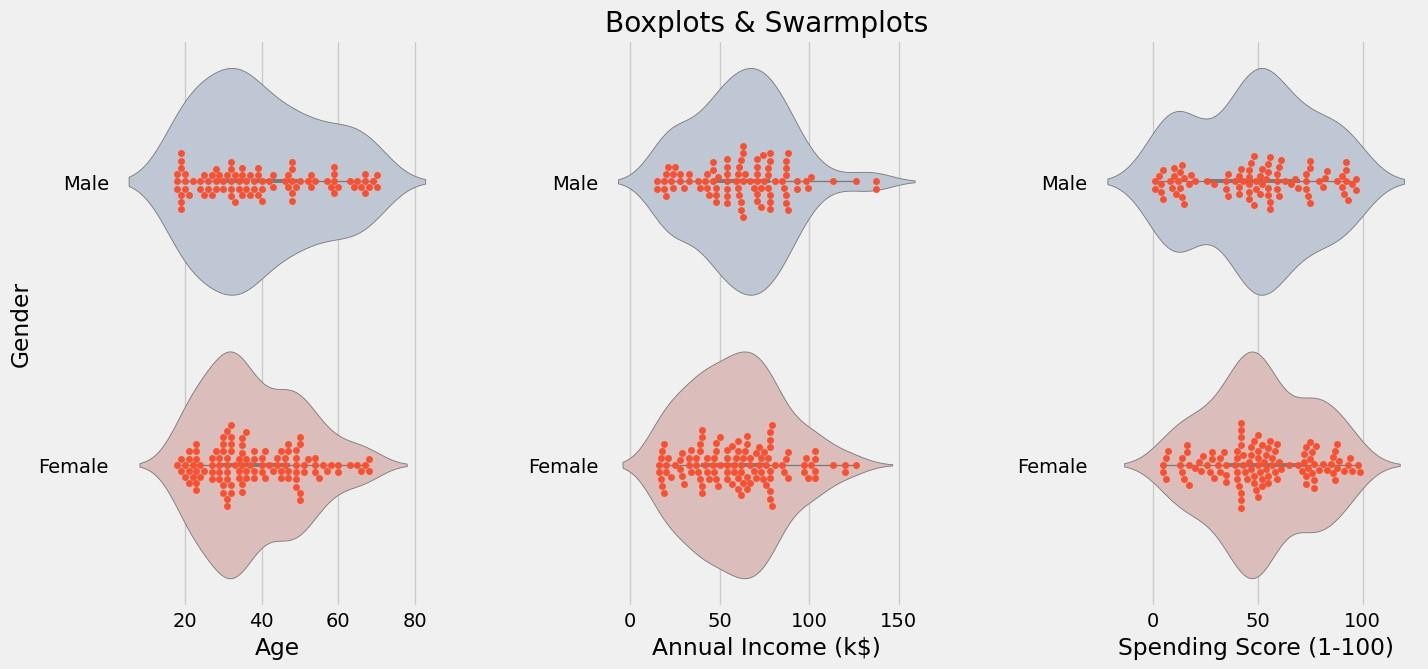


Fig: 4.7 Boxplot & Swarmplot

Fig: 4.8 Clusters based on Inertia

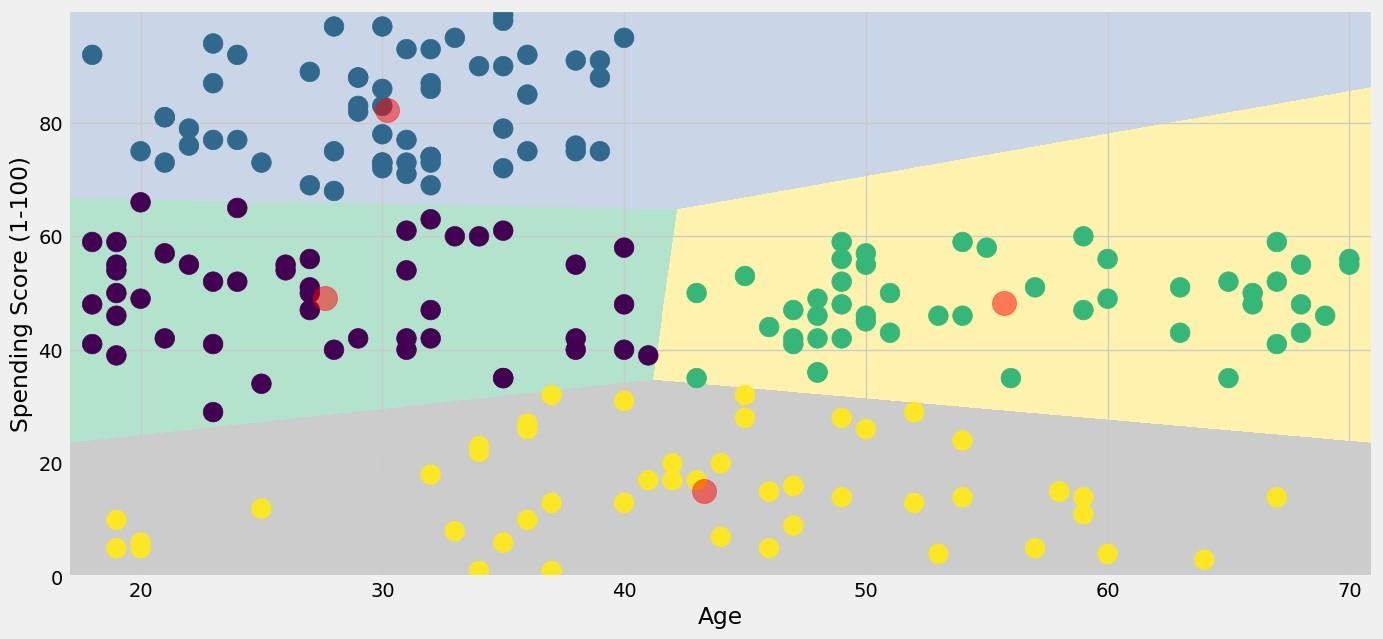


Fig: 4.9 Scatterplot Of Clusters

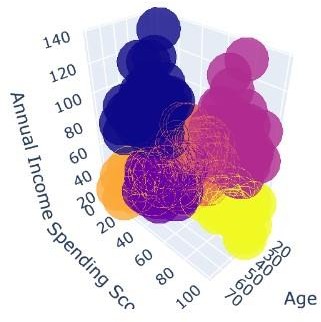


Fig: 4.10 Segmented 3D Cluster

### Proposed system versus existing system

When comparing a proposed system for customer segmentation using K-means to an existing system, it's crucial to evaluate the potential advantages and improvements that the new system brings to the table. The choice between the two depends on a variety of factors, including the specific needs and goals of the business.

The proposed system often offers enhancements in terms of accuracy and effectiveness. It may incorporate more sophisticated algorithms, refined feature selection, and better data preprocessing techniques. As a result, it has the potential to produce more precise customer segmentation results. In contrast, the existing system may have limitations, potentially relying on outdated methods or lacking robust data processing capabilities, leading to less accurate results.

Scalability is another critical consideration. The proposed system may be designed to handle larger datasets and accommodate the scalability requirements of a growing business. This is especially important in industries with rapidly expanding customer bases. In contrast, the existing system may struggle to handle an increasing number of customers and could reach its operational limits.

# Conclusion

So, we concluded that the,

1. The Annual income, Spending score can be target these type of customers as they earn more money and spend as much as they want.
2. Annual score, Spending Score can be target these types of customers by asking feedback and advertising the product in a better way.
3. Annual income, Spending score may or may not be beneficial to the mall owners of this type of customers.
4. Low income, High spending can be targeting these types of customers by providing them with low-cost EMI’s etc.
5. Low income, Low spending doesn’t target these types 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.

In conclusion, customer segmentation through K-means clustering in a mall environment is a valuable approach for gaining insights into the diverse customer base. By categorizing shoppers into distinct segments, we can identify the "Target Customers" - those who are most likely to easily converge, and tailor marketing strategies to meet their specific needs and preferences. This data-driven approach not only optimizes resource allocation but also enhances the overall shopping experience, ensuring that the marketing team can create more effective strategies to engage and retain customers, ultimately leading to increased sales and customer satisfaction

# Appendix

1. https://app.creately.com/

Creately tremendously helped is making the UML diagrams in the project. The various UML diagrams made in the project are – Data Flow Diagrams, Use Case Diagrams and the Entity Relationship Diagrams.

1. PyCharm IDE or VS code

PyCharm IDE was used for setting up the entire project.

1. Python 3.11.5

Python programming language with the version of 3.6 to sync our libraries and make our app compatible on all systems.

1. Seaborn

Seaborn is a library for making statistical graphics in Python.

1. sklearn

Scikit-Learn, also known as sklearn is a python library to implement machine learning models and statistical modelling. Through scikit-learn, we can implement various machine learning models for regression, classification, clustering, and statistical tools for analyzing these models.

# References

1. Ong, A.K.S.; Prasetyo, Y.T.; De Leon, L.A.S.; Ayuwati, I.D.; Nadlifatin, R.; Persada,

S.F. Plantitas/Plantitos Preference Analysis on Succulents Attributes and Its Market Segmentation: Integrating Conjoint Analysis and K-Means Clustering for Gardening Marketing Strategy. Sustainability 2022, 14, 16718. https://doi.org/ 10.3390/su142416718

1. Chen, Daqing, Sain, Sai Liang, Guo, Kun, 2012. Data mining for the online retail industry: a case study of RFM model-based customer segmentation using data mining. J. Database Mark. Customer Strategy Manage. 19 (3), 197–208. https:// doi.org/10.1057/dbm.2012.17.
2. J. Wu, L. Shi, W. Lin et al., “An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and KMeans Algorithm,” Mathematical Problems in Engineering, vol. 2020, Article ID 8884227, 7 pages, 2020
3. Balakrishnan, P.V., Cooper, M.C., Jacob, V.S., Lewis, P.A., 1994. A study of the classi®cation capabilities of neural networks using unsupervised learning: A comparison with k-means clustering. Psychometrika 59, 509±525
4. K. M. Manero, R. Rimiru, and C. Otieno, "Customer Behaviour Segmentation among Mobile Service Providers in Kenya using K-Means Algorithm," International Journal of Computer Science, vol. 15, no. 5, pp. 67–76, 2018.
5. Int. J. Nonlinear Anal. Appl. Volume 12, Special Issue, Winter and Spring 2021, 633- 642 ISSN: 2008-6822 (electronic) <http://dx.doi.org/10.22075/ijnaa.2021.22324.2350>
6. Vladimír Holý, Ondřej Sokol, Michal Černý,”Clustering Retail Products Based on Customer Behaviour”, Applied Soft Computing, Elsevier Vol 60, PP: 752-762, 2017.
7. Customer Segmentation using RFM Model and K-Means Clustering International Journal of Scientific Research in Science and Technology Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com) doi: <https://doi.org/10.32628/IJSRST2183118>
8. Hwang, Y. H. (2019). Hands-on Advertising Science Data: Develop your machine learning marketing strategies… using python and r. S.l: Packt printing is limited
9. A review on customer segmentation methods for personalized customer targeting in e-commerce use cases Information Systems and e-Business Management https://doi.org/10.1007/s10257-023-00640-4

# Acknowledgement

We take this opportunity to express our deep sense of gratitude to our project guide and project coordinator**, Mr. John Kenny**, for her continuous guidance and encouragement throughout the duration of our major project work. It is because of her experience and wonderful knowledge; we can fulfil the requirement of completing the major project within the stipulated time. We would also like to thank **Mr. John Kenny**, Head of Artificial Intelligence and Machine department for his encouragement, whole-hearted cooperation and support.

We would also like to thank our Principal, **Dr. J. B. Patil** and the management of Universal College of Engineering, Vasai, Mumbai for providing us all the facilities and the work friendly environment. We acknowledge with thanks, the assistance provided by departmental staff, library and lab attendants.

**Mr . Omkar Shankar Gavade (29)**

##### Mr. Yash Kishor Gawli (30)

**Ms. Vrushali Haridas Jagtap (37)**

##### Mr. Ashish Anand Jangam (39)