Project Report on

CUSTOMER SEGMENTATION USING CLUSTERING ALGORITHM

(EC881)

In partial fulfilment of the requirements

For the of Bachelor of Technology in Electronics and Communication Engineering

SUBMITTED BY:

Ayan Singha Roy Chowdhury(16900321182)

Amitava Ghosh(16900321185)

Debanjan Muhuri(16900321149)

Sagnik Ray(16900321178)

Deep Nandi(16900321163)

UNDER THE GUIDANCE

OF:

PROF. CHIRANJIT GUCHAIT

ACADEMY OF TECHNOLOGY G.T. ROAD,

Adisaptagram, P.O.: Aedconagar

Hooghly-712502,West Bengal

MAULANA ABUL KALAM AZAD
UNIVERSITY OF TECHNOLOGY,
WEST BENGAL

ACADEMY OF TECHNOLOGY G.T. ROAD,

Adisaptagram, P.O.: Aedconagar

Hooghly-712502,West Bengal

MAULANA ABUL KALAM AZAD
UNIVERSITY OF TECHNOLOGY,
WEST BENGAL

This is to certify that the Project Report Titled “CUSTOMER SEGMENTATION USING CLUSTERING ALGORITHM” submitted by Ayan Singha Roy Chowdhury (16900321182), Amitava Ghosh(16900321185), Debanjan Muhuri(16900321149), Sagnik Ray(16900321178) and Deep Nandi(16900321163) in partial fulfilment of the requirements for the award of the degree Bachelor of Technology in Electronics and communication Engineering of Maulana Abdul Kalam Azad University of Technology, commonly known as MAKAUT or WBUT, Kolkata , West Bengal ,is a record of Bonafede work carried out under my guidance and supervision.

Project Guide,	Head of the Department (H.O.D),
PROF. CHIRANJIT GUCHAIT	DR. ABHIJIT BANERJEE
Department of M.E.	Department of E.C.E
A.O.T	A.O.T

STATEMENT

We hereby state that we have prepared this project report as a record of our ongoing Final-year project. This project report is being submitted for partial fulfilment of the requirements of our final year project of the curriculum at the Academy of Technology, Hooghly, West Bengal.

Ayan Singha Roy Chowdhury(16900321182) …………………………………

Amitava Ghosh(16900321185) …………………………………

Debanjan Muhuri(16900321149) …………………………………

Sagnik Ray(16900321178) …………………………………

Deep Nandi(16900321163) …………………………………..

ACKNOWLEDGMENT

This is a great pleasure for us to avail ourselves the opportunity to express our gratefulness to all those who were immensely helping us in our project. We firstly thank our head of Department, Dr. Abhijit Banerjee for being the constant source of inspiration in the completion of the project.

We would like to express our sincere gratitude to Prof. Chiranjit Guchait for his valuable guidance, cooperation, and enthusiasm during the progress of this work.

While making this project we have taken references from research papers and different books that are available in our library.

We would also like to thank all others who have helped in the progress of this project work.

PROJECT STUDENTS

Ayan Singha Roy Chowdhury(16900321182)

Amitava Ghosh(16900321185)

Debanjan Muhuri(16900321149)

Sagnik Ray(16900321178)

Deep Nandi(16900321163)

Abstract

In this modern era, everything and everyone is innovative, where everyone competes with being better than others. The emergence of many entrepreneurs, competitors, and business interested people has created a lot of insecurities and tension among competing businesses to find new customers and hold the old customers. Because of this one should need and maintain exceptional customer service and it becomes very appropriate irrespective of the business scale. And also, it is equally important to understand the needs of customers specifically to provide greater customer support and to advertise them with the most appropriate products. In the pool of these online products customers are confused about what to buy and what not to and also the company or the business people are confused about which section of customers to be targeted for selling their particular type of products. This confusion will probably be possible by the process called customer segmentation. The process of segmenting the customers with similar interests and similar shopping behaviour into the same segment and with different interests and different shopping patterns into different segments is called customer segmentation. Customer segmentation and pattern extraction are the major aspects of a business decision support system. Each segment has the same set of customers who most probably has the same kind of interests and shopping patterns. In this paper, we planned to do this customer segmentation using three different clustering algorithms namely K means clustering algorithm, Mini batch means, and hierarchical clustering algorithms and also going to compare all these clustering algorithms based on their efficiency and root mean squared errors.

Table of Content

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

ML - Machine Learning

CSV - Comma-Separated Values (for dataset format)

EDA - Exploratory Data Analysis

KM - K-Means (clustering algorithm)

CC - Cluster Centroid

CHAPTER 1: INTRODUCTION

1.1 OVERVIEW

Data is very precious in today’s ever-competitive world. Every day organizations and people are encountered with a large amount of data. A most efficient way to handle this data is to classify or categorize the data into Clusters, set of groups, or partitions. “Usually, the classification methods are either supervised or unsupervised, depending on whether they have labelled datasets or not”. Unsupervised classification is the exploratory data analysis where there won’t be any training data set and having to extract hidden patterns in the data set with no labelled responses is achieved whereas classification of supervised learning model is machine learning task of deducing a function from training data set. The main focus is to enhance the propinquity or closeness in data points belonging to the same group and increase the variance among various groups and all this is achieved through some measure of similarity. Exploratory- by data analysis is all about dealing with a wide range of applications such as engineering, text mining, pattern recognition, bioinformatics, spatial data analysis, - mechanical engineering, voice mining, textual document collection, artificial intelligence, image segmentation, ”. This diversity explains the importance of clustering in scientific research but this diversity can lead to contradictions due to different purposes and nomenclature.

Customer segmentation has importance as it includes, the ability to modify the pro- grams of the market so that it is suitable to each of the segments, support in a business decision, identification of products associated with each customer segment, and managing the demand and supply of that product, and predicting customer defection, identifying and targeting the potential customer base, providing directions in finding the solutions. Clustering is an iterative process of knowledge discovery from unorganized and huge amounts of data that is raw. Clustering is one of the kinds of exploration of data mining that is used in several applications, those are classification, machine learning, and recognition of patterns.

1.2 MACHINE LEARNING

Machine learning (ML) is a field of artificial intelligence focused on creating algorithms that automatically improve with experience and data, without explicit programming. ML algorithms build models from training data to make predictions or decisions, applied widely in fields like medicine, email filtering, speech recognition, and computer vision. While related to computational statistics and data mining, ML emphasizes predictive modeling and pattern recognition. Some ML techniques use neural networks to simulate the brain's processing, and in business, ML is often known as predictive analytics. The primary aim of ML is to extract and model data patterns for human understanding and practical use, setting it apart from traditional programming methods.

1.3 MACHINE LEARNING APPROACHES

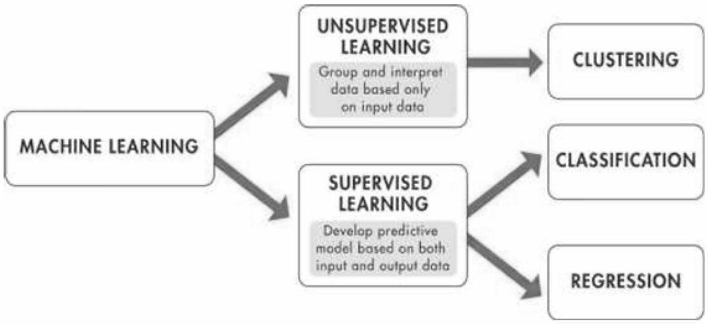
In machine learning, tasks square measure is typically classified into broad classes. These classes square measure supported however learning is received or however, feedback on the education is given to the system developed. Two of the foremost wide adopted machine learning strategies are square measure supervised learning that trains algorithms supported example input and output information that's tagged by humans, and unattended learning that provides the algorithmic program with no tagged information to permit it to search out structure at intervals its computer file. Machine learning approaches are traditionally divided into three broad categories, depending on the nature of the "signal" or "feedback" available to the learning system: Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs. Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning). Reinforcement learning: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or playing a game against an opponent). As it navigates its problem space, the program is provided feedback that's analogous to rewards, which it tries to maximize.

1.3.1 Supervised Learning

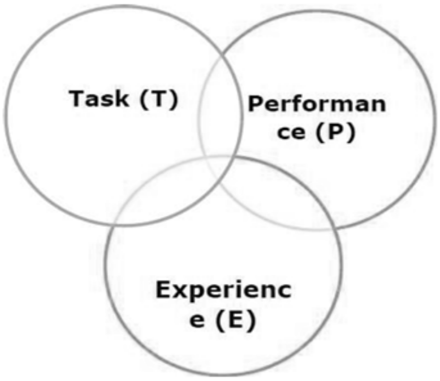
Supervised learning involves training algorithms on labelled input-output pairs so they can learn to predict outputs for new, unlabelled data. By comparing predictions with known outcomes, the model adjusts to minimize errors. It’s used in applications like image recognition (e.g., identifying sharks as fish and oceans as water), stock market prediction, spam filtering, and image classification. Key supervised learning methods include classification, regression, and similarity learning. Classification is used when outputs are discrete (e.g., email categorization), while regression handles continuous outputs. Similarity learning focuses on comparing objects for applications in ranking, recommendations, and identity verification.

1.3.2 Unsupervised Learning

Unsupervised learning uses unlabelled data to identify patterns and features within datasets, often to discover hidden structures or groupings without predefined categories. It’s commonly applied to transactional data, such as analysing customer purchases to reveal insights that may be challenging to identify manually. For example, unsupervised algorithms might find that women of a certain age buying unscented soaps are likely pregnant, allowing for targeted marketing campaigns related to maternity and baby products.


MACHINE LEARNING
UNSUPERVISED LEARNING
Group and interpret data based only on input data
SUPERVISED LEARNING
Develop predictive model based on both input and output data
CLUSTERING
CLASSIFICATION
REGRESSION

1.1 Machine Learning Classification



1.2 Machine Learning Task

1.4 CLUSTERING

Clustering is the task of dividing the data points into definite groups such that the data points in the same group have similar characteristics or similar behaviour. In short, segregating the data points into different clusters based on their similar traits. Clustering is a type of unsupervised learning method of machine learning. In the unsupervised learning method, the inferences are drawn from the data sets which do not contain labelled output variables. It is an exploratory data analysis technique that allows us to analyse the multivariate data sets. It depends on the type of algorithm we use which decides how the clusters will be created. The inferences that need to be drawn from the data sets also depend upon the user as there is no criterion for good clustering.

1.4.1 Types of Clustering

Clustering itself can be categorized into two types viz. Hard Clustering and Soft Clustering. In hard clustering, one data point can belong to one cluster only. But in soft clustering, the output provided is a probability likelihood of a data point belonging to each of the pre-defined numbers of clusters. The task of clustering is subjective which means there are many ways of achieving the goal of clustering. Each methodology has its own set of rules to segregate data points into different clusters. There is n number of clustering algorithms in which these are few mostly used algorithms such as K means clustering algorithm, Hierarchical clustering algorithms, and Mini-batch K means clustering algorithm, etc.

1.4.2 Density-Based Clustering

In this method, the clusters are created based upon the density of the data points which are represented in the data space. The regions that become dense due to the huge number of data points residing in that region are considered clusters. The data points in the sparse region (the region where the data points are very few) are considered as noise or outliers. The clusters created in these methods can be of arbitrary shape.

1.4.3 Hierarchical Clustering

Hierarchical Clustering groups (Agglomerative or also called Bottom-Up Approach) or divides (Divisive or also called Top-Down Approach) the clusters based on the distance metrics. In Agglomerative clustering, each data point acts as a cluster initially, and then it groups the clusters one by one. Divisive is the opposite of Agglomerative, it starts with all the points into one cluster and divides them to create more clusters. These algorithms create a distance matrix of all the existing clusters and perform the linkage between the clusters depending on the criteria of the linkage. The clustering of the data points is represented by using a dendrogram.

1.4.4 Centroid-based

Centroid-based clustering is the one you probably hear about the most. It's a little sensitive to the initial parameters you give it, but it's fast and efficient. These types of algorithms separate data points based on multiple centroids in the data. Each data point is assigned to a cluster based on its squared distance from the centroid. This is the most commonly used type of clustering. K-Means clustering is one of the most widely used algorithms. It partitions the data points into k clusters based upon the distance metric used for the clustering. The value of ‘k’ is to be defined by the user. The distance is calculated between the data points and the centroids of the clusters. The data point which is closest to the centroid of the cluster gets assigned to that cluster. After an iteration, it computes the centroids of those clusters again and the process continues until a pre-defined number of iterations are completed or when the centroids of the clusters do not change after an iteration. It is a very computationally expensive algorithm as it computes the distance of every data point with the centroids of all the clusters at each iteration. This makes it difficult for implementing the same for huge data sets.

1.4.5 Applications of Clustering

Clustering is used in our daily lives such as in data mining, in academics, in web cluster engines, in bioinformatics, in image processing, and many more. There are a few common applications where clustering is used as a tool are Recommendation engines, Market segmentation, Customer segmentation, Social Network Analysis(SNA), Search result Clustering, Identification of cancer cells, biological data analysis, and medical imaging analysis.

- **Image Segmentation** − Clustering helps partition images into segments for object and pattern recognition.
- **Customer Segmentation** − Clustering groups customers based on behaviour and preferences for targeted marketing.
- **Anomaly Detection** − Clustering identifies unusual patterns in data, useful in fraud detection and network security.
- **Document Clustering** − Clustering organizes documents into topics for easier information retrieval and topic modelling.
- **Genomic Data Analysis** − Clustering groups genes or proteins with similar expression patterns to understand biological functions.
- **Social Network Analysis** − Clustering finds communities within networks, enhancing insights into social interactions.
- **Market Basket Analysis** − Clustering identifies products frequently bought together to optimize product placement and recommendations.
- **Image Compression** − Clustering reduces the number of colours in images for efficient storage and compression.

CHAPTER 2: LITERATURE SURVEY

- **Kishana R. Kashwan et al.[1]** proposed a customer segmentation model titled *"Customer Segmentation Using Clustering and Data Mining Techniques."* This model uses K-means clustering to analyse supermarket sales data and supports real-time, online operations to predict seasonal sales trends. Designed as an intelligent tool, it receives daily sales inputs and updates segmentation statistics automatically at the end of each business day. Tested over three months, the model analysed data from 2,138 customers, segmenting them into four groups based on purchasing behaviour. Results showed high accuracy in segmenting customers, demonstrating the model's effectiveness in a retail environment.

- **Kayalvily Tabianan et al.[2]** conducted a study on e-commerce customer segmentation, focusing on behavioral factors. By using K-means clustering, they aimed to analyze purchase behaviors in e-commerce systems, optimizing similarity within clusters and maximizing dissimilarity between them. The study identified relationships among three clusters—event type, products, and categories—to help vendors target profitable customer segments. By grouping customers with similar behaviors, vendors can improve long-term customer retention and increase profitability. The K-means clustering results were satisfactory, supporting the approach's effectiveness in segmenting customers for enhanced business strategies.

- **Pranjali Joshi et al.[3]** conducted a study on profiling retail banking customers using a combination of deterministic rules and unsupervised machine learning. The study analyzed customer transaction data with clustering techniques to identify patterns in past behaviors. The model categorized customers into distinct clusters based on transaction behavior, enabling the bank to tailor product offerings to different customer segments effectively. This approach helps banks personalize their marketing strategies, improving customer engagement with relevant products.

- **Kai Peng (Member, IEEE), Victor C. M. Leung, (Fellow, IEEE), and Qinghai Huang in [4]** get to know in detail about mini-batch K-means clustering algorithm. Get to know about the advantages and disadvantages of the algorithm and also about the implementation.

- **Fionn Murtagh and Pedro Contreras in [5]** studied hierarchical clustering algorithms. In this paper get to know more about this clustering algorithm and also observe how clusters formed and also about advantages and disadvantages and compare it with the other different clustering algorithms.

- **D. P. Yash Kushwaha, Deepak Prajapati in [6]** studied customer segmentation in detail and also studied in detail about k-means clustering algorithm and performed customer segmentation using K-means clustering algorithm and observed the clusters formed and compared the results with the other clustering algorithms.

CHAPTER 3: THEORITICAL BACKGROUND

Programming Language:

In our customer segmentation project, we used Python as our primary programming language due to its versatility and extensive libraries that support data science and machine learning tasks. Python’s simplicity and readability make it an excellent choice for implementing complex algorithms and handling large datasets. For this project, libraries like NumPy and Pandas enabled efficient data manipulation and preprocessing, while Scikit-learn provided robust tools for clustering algorithms. Additionally, we used Matplotlib and Seaborn to visualize customer clusters and interpret the segmentation results more effectively. Overall, Python’s comprehensive ecosystem facilitated a streamlined and effective approach to customer segmentation.

Jupyter Notebook:

In this Project we have used Jupyter Notebook as a platform for coding.The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

In our Project we have used following Libraries:

- NumPy (Version: 1.25.2)
- Pandas (Version: 2.1.4)
- Seaborn (Version: 0.13.2)
- Matplotlib (Version: 3.8.2)
- Scikit Learn (Version: 1.3.2)

NumPy:

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, 1/0, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

Pandas:

The Pandas library in Python is a powerful tool for data manipulation and analysis, widely used in data science and machine learning projects. It provides flexible data structures, mainly Data Frames and Series, that allow users to handle and analyse structured data efficiently. Pandas simplifies tasks like data cleaning, merging, reshaping, and transformation, making it easy to handle missing data, perform group operations, and conduct statistical analyses. With its extensive functions for data handling and compatibility with other libraries like NumPy and Matplotlib, Pandas is essential for preparing data for analysis and model training.

Seaborn:

The Seaborn library is a Python data visualization library built on top of Matplotlib, designed specifically for creating visually appealing and informative statistical graphics. Seaborn simplifies the process of creating complex visualizations, such as heatmaps, violin plots, and pair plots, that help in understanding patterns and relationships in data. It comes with a variety of color palettes and themes, making it easy to customize the appearance of graphs. With built-in support for aggregating data, visualizing distributions, and showing statistical relationships, Seaborn is particularly useful for exploratory data analysis in data science projects.

Matplotlib:

Matplotlib is a foundational plotting library in Python used to create static, interactive, and animated visualizations. It provides a flexible framework for creating a wide range of charts, including line graphs, bar charts, histograms, scatter plots, and more. Known for its versatility, Matplotlib allows detailed control over plot elements such as colors, labels, axes, and legends, making it suitable for both basic and advanced visualizations. While it may require more customization for complex visualizations, Matplotlib integrates well with libraries like Pandas and Seaborn, making it a core tool for data visualization in scientific and analytical applications.

Scikit-Learn:

Scikit-learn is a popular machine learning library in Python that provides a wide range of tools for data analysis and predictive modeling. It offers efficient implementations of many machine learning algorithms, including classification, regression, clustering, and dimensionality reduction. The library also includes utilities for data preprocessing, model selection, and evaluation. With its extensive documentation and integration with libraries like Pandas, NumPy, and Matplotlib, Scikit-learn is a fundamental tool for building and testing machine learning models.

2.Univariate Analysis: -

2.1 Statistical Summary: -

Statistics are essential in machine learning for:

1. **Data Understanding**: Summarize data with mean, median, and standard deviation to identify patterns and outliers.
2. **Data Preprocessing**: Handle missing values and normalize data for better model performance.
3. **Feature Selection**: Identify important features by analysing correlations, reducing redundancy.
4. **Model Evaluation**: Use metrics like accuracy, precision, and recall to assess model performance.

1. Mean

The mean, or average, is calculated by summing all values in a dataset and dividing by the number of values. It provides a central value that represents the dataset.

Formula:

$$\text{Mean} = \frac{\sum X}{N}$$

where X represents each value, and N is the number of values.

Example:
For a dataset [4,8,6,5,3][4,8,6,5,3]:

$$\text{Mean} = \frac{4+8+6+5+3}{5} = \frac{26}{5} = 5.2$$

2. Median

The median is the middle value in an ordered dataset, providing a measure of central tendency that isn't affected by extreme values. For an odd number of values, it's the middle one; for an even number, it's the average of the two middle values.

Example:
For the dataset [3,5,4,8,6][3,5,4,8,6], first arrange it in ascending order: [3,4,5,6,8][3,4,5,6,8]. The median is the middle value, which is **5**.

If the dataset has an even number of values, e.g., [3,4,5,6][3,4,5,6], the median is:

$$\text{Median} = \frac{4+5}{2} = 4.5$$

3. Mode

The mode is the value that appears most frequently in a dataset. It’s useful for understanding which value occurs most often.

Example:
For the dataset [3,7,3,5,3,8,5][3,7,3,5,3,8,5], the mode is **3**, as it appears most frequently.

4. Percentage

Percentage is a way to express a value as a fraction of 100, providing a normalized view. It’s calculated by dividing the part by the whole and multiplying by 100.

Formula:

$$\text{Percentage} = \left(\frac{\text{Part}}{\text{Whole}}\right) \times 100$$

Example:
If a class has 20 students and 8 of them scored above 80%, then the percentage of students scoring above 80% is:

$$\left(\frac{8}{20}\right) \times 100 = 40\%$$

5. Standard Deviation Standard deviation (SD) measures how spread out the values in a dataset are around the mean. A higher SD indicates greater variability.

Formula:

$$\text{SD} = \sqrt{\frac{\sum(X - \text{Mean})^2}{N}}$$

where X represents each value, Mean is the dataset’s average, and N is the number of values.

Example:
For the dataset [2,4,4,4,5,5,7,9]:

1.Calculate the Mean:

$$\text{Mean} = \frac{2+4+4+4+5+5+7+9}{8} = 5$$

2.Calculate Each Value’s Difference from the Mean, Square It, and Find the Average:

$$\text{SD} = \sqrt{\frac{(2-5)^2+(4-5)^2+(4-5)^2+(4-5)^2+(5-5)^2+(5-5)^2+(7-5)^2+(9-5)^2}{8}} = \sqrt{\frac{34}{8}} \approx 2.07$$

2.2 Categorical features: -

Categorical features represent distinct categories or groups (e.g., gender, region). Bar plots are commonly used to visualize their distribution.

Bar Plot:

Displays the frequency or count of each category, helping to understand the spread and dominance of categories.

2.3 Numerical Features: -

Numerical features have continuous or discrete numerical values (e.g., age, income). Several plots are used to analyze their distribution and variability.

Histogram: Shows the frequency distribution of a numerical feature by grouping values into bins, providing insights into its range, shape, and spread.

Box Plot: Visualizes the spread, central tendency, and outliers in a numerical feature. It displays the median, quartiles, and potential outliers, making it useful for identifying extreme values.

3. Bivariate Analysis: -

3.1 Pairwise Relationships: - A pairwise relationship refers to examining the relationship between two individual features (variables) in a dataset. In data analysis, looking at pairwise relationships helps to understand how one feature might relate to another, and can reveal trends, dependencies, or patterns within the data.

How Pairwise Relationships Are Used:

- **Scatter Plots:** Scatter plots are commonly used to visualize pairwise relationships between two numerical features. This can help to observe correlations, trends, or clusters.

This scatter plot is showing a positive relationship between the Age and Annual Income i.e as Age increases, the Annual Income also tends to increase

3.2 Correlation Matrix: - A correlation matrix is used to measure and display the relationships between numerical features in a dataset. It shows how strongly or weakly each feature correlates with the others, helps us to identify patterns or redundancies in data.

In a correlation matrix:

- **Correlation values** range from -1 to +1.
  - **+1** indicates a perfect positive correlation (as one feature increases, the other increases proportionally).
  - **-1** indicates a perfect negative correlation (as one feature increases, the other decreases proportionally).
  - **0** suggests no correlation (the features are independent of each other).
- **Purpose**:
  - **Identify Relationships**: Features with high positive or negative correlations might have strong relationships that could impact analysis.
  - **Feature Selection**: Highly correlated features may be redundant, so analysts might choose to remove or combine them.
  - **Modelling Insight**: For clustering or regression, correlation insights can help choose or modify features to improve model performance.
  - **Heatmap: -** A heatmap is a visualization tool that represents the values of a matrix (like a correlation matrix) using color gradients. In the context of a correlation matrix, a heatmap color-codes each correlation coefficient, making it easier to see patterns and quickly identify strong or weak correlations.

CHAPTER 4: PROBLEM SPECIFICATION & DATASET

❖ PROBLEM SPECIFICATION: -

In today's competitive market, understanding customer needs and behaviours is crucial for any business aiming to maintain its market share and grow. Customer segmentation allows businesses to categorize their customers into different groups, enabling them to target each group more effectively with customized marketing strategies and offers.

In the context of this project, we have been provided with a dataset comprising mall customers' information, presented in the form of an Excel sheet. This dataset encompasses approximately 200 entries, with each row representing an individual customer. The dataset includes the following columns:

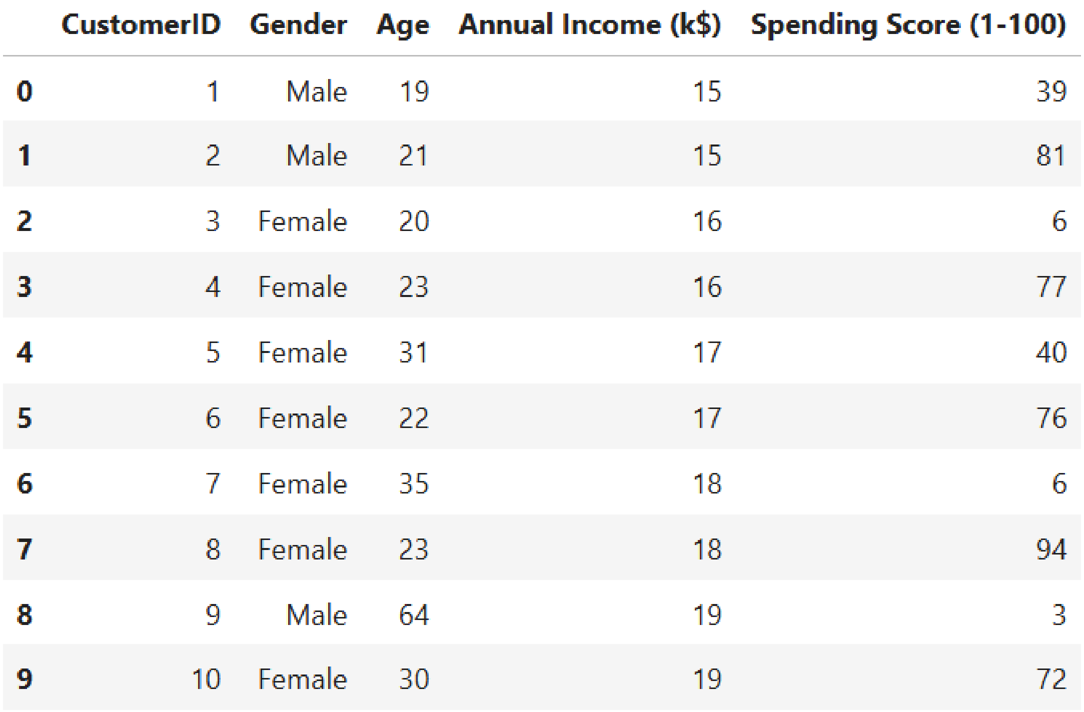
- **Customer ID**
- **Age**
- **Gender**
- **Annual Income**
- **Spending Score**

The objective is to develop a robust mathematical model capable of predicting a customer's *Spending Score* based on their **Annual Income**, **Age**, and **Gender**. Achieving the highest possible accuracy for this predictive model is of utmost importance.

This predictive model will empower mall retailers to make data-driven decisions by identifying trending purchasing patterns. Consequently, they will be better equipped to stock popular products, enhancing customer satisfaction and boosting sales efficiency.

❖ DATASET

- **Customer ID**: A unique identifier assigned to each customer.
- **Age**: The age of the customer.
- **Gender**: The gender of the customer (e.g., male or female).
- **Annual Income**: The yearly income of the customer, typically in monetary units.
- **Spending Score**: A measure of the customer's spending behaviour and purchasing patterns.



	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
5	6	Female	22	17	76
6	7	Female	35	18	6
7	8	Female	23	18	94
8	9	Male	64	19	3
9	10	Female	30	19	72

Table:- 4.1 Sample Dataset

CHAPTER 5: PROBLEM SOLVING APPROACH

To perform customer segmentation through clustering analysis, we will utilize the **K-Means Clustering** algorithm, incorporating five key attributes: **Customer ID**, **Gender**, **Age**, **Annual Income (in k$)**, and **Spending Score (1-100)**. These variables have been chosen for their significant contribution to understanding customer behaviour and enabling the grouping of customers into similar segments.

The process begins with importing the customer dataset, typically in the form of a .csv file, and converting it into a **Pandas Data Frame**. The dataset will undergo preprocessing steps, including handling missing values, encoding categorical data (such as the **Gender** column), and scaling numerical attributes to enhance clustering performance.

To identify the optimal number of clusters, the program will employ the **Elbow Method**, a widely used technique in clustering analysis. Once the optimal cluster count is determined, the **K-Means algorithm** will be applied to segment the customers into distinct groups based on their attributes.

The output will feature comprehensive visualizations, aiding in the interpretation of these clusters and offering actionable insights into customer behaviour. These insights will empower businesses to develop targeted marketing strategies and foster better customer engagement.

To execute this analysis, we will leverage the power of Python's data science libraries, including **Pandas**, **NumPy**, **Seaborn**, **Matplotlib**, and **Sklearn**. These tools will facilitate data preprocessing, the implementation of the K-Means algorithm, and the creation of insightful visualizations for effective interpretation of the results.

CHAPTER 6: COMPUTER SIMULATION

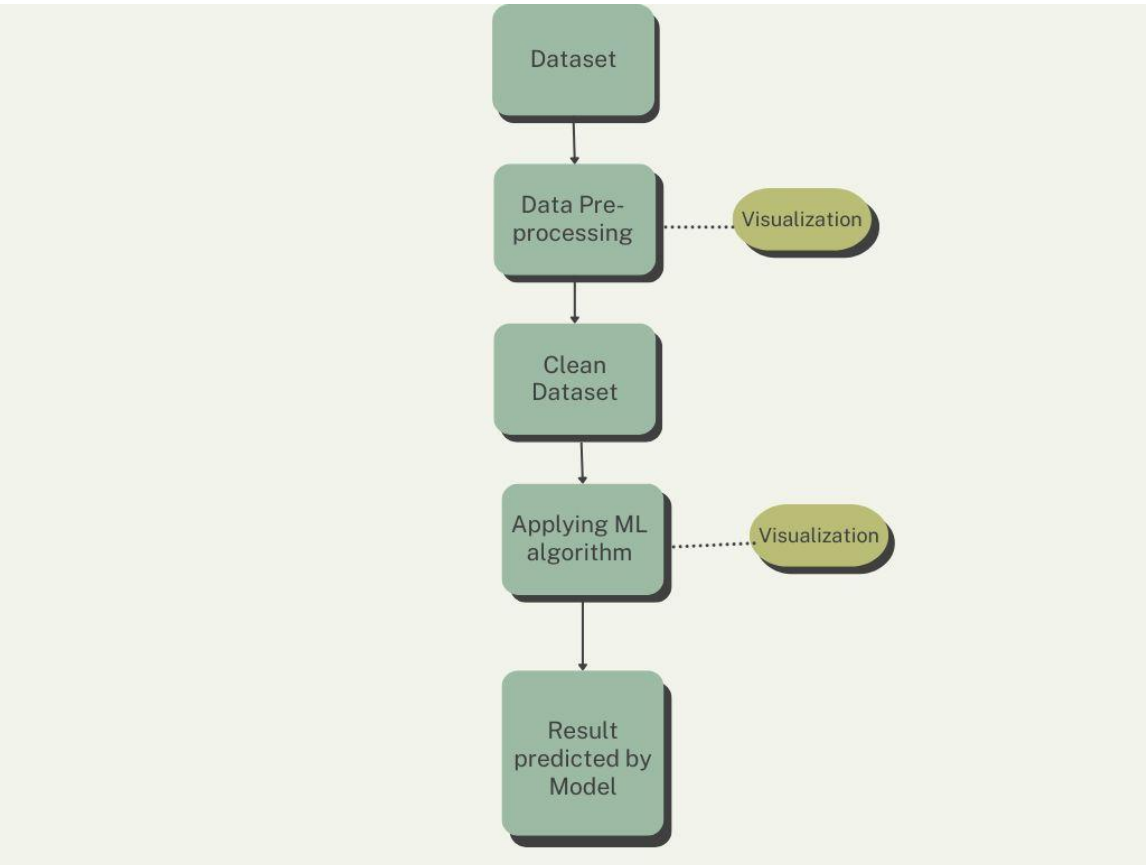


Fig-6.1 Flow Chart

- **Computer Simulation of our dataset**

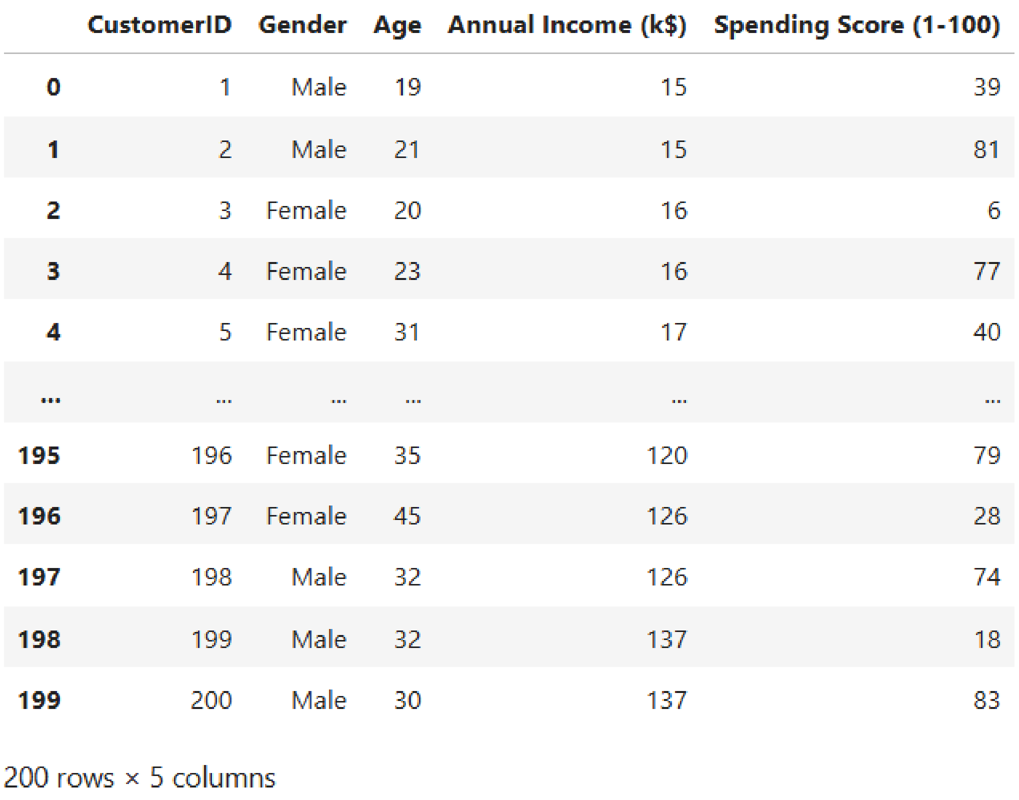
Code:-

```
import numpy as np

import pandas as pd

import seaborn as sns
import matplotlib.pyplot as plt
import os

df=pd.read_csv(r"C:\Users\wwwsa\Desktop\Mall_Customers.csv")
```

```
        df
#To check the information
        df.info()
# To check duplicate values
        df.duplicated().sum
#identifying garbage values
        for i in df.select_dtypes(include="object").columns:
            print(df[i].value_counts())
            print("***"*10)

#describe data
        df.describe().T
# To check the distribution of data
        for i in df.select_dtypes(include="number").columns:
            sns.histplot(data=df,x=i)
            plt.show()

# To check outliers
        for i in df.select_dtypes(include="number").columns:
            sns.boxplot(data=df,x=i)
            plt.show()


#Scatterplot of Age vs. Spending Score
        Clustering1 = KMeans(n_clusters=5)
        Clustering1.fit(df[['Age','Spending Score (1-100)']])
        df['Spending and Age Cluster'] =clustering1.labels_
        df.head()
        intertia_scores1=[]
        for i in range(1,11):
            kmeans1=KMeans(n_clusters=i)
            kmeans1.fit(df[['Age','Spending Score (1-100)']])
            intertia_scores1.append(kmeans1.inertia_)
        plt.plot(range(1,11),intertia_scores1)
centers =pd.DataFrame(clustering1.cluster_centers_)
centers.columns = ['x','y']
plt.figure(figsize=(10,8))

plt.scatter(x=centers['x'],y=centers['y'],s=100,c='black',marker='*')
sns.scatterplot(data=df,x='Age',y='Spending Score(1-100)',hue='Spending and Age Cluster',palette='tab10')
plt.savefig('clustering_bivaraiate1.png')


#Scatterplot of Annual Income vs. Spending Score
        X=df[["Annual Income (k$)","Spending Score (1-100)"]]
        plt.figure(figsize=(10,6))
        sns.scatterplot(x = 'Annual Income (k$)',y = 'Spending Score (1-100)',  data = X,s = 60 )
        plt.xlabel('Annual Income (k$)')
        plt.ylabel('Spending Score (1-100)')
        plt.title('Spending Score (1-100) vs Annual Income (k$)')
        plt.show()


Clustering2= KMeans(n_clusters=5)
Clustering2.fit(df[['Annual Income (k$)','Spending Score (1-100)']])
df['Spending and Income Cluster'] =clustering2.labels_
df.head()


intertia_scores2=[]
for i in range(1,11):
    kmeans2=KMeans(n_clusters=i)
    kmeans2.fit(df[['Annual Income (k$)','Spending Score (1-100)']])
    intertia_scores2.append(kmeans2.inertia_)
plt.plot(range(1,11),intertia_scores2)
```

```
centers =pd.DataFrame(clustering2.cluster_centers_)

centers.columns = ['x','y']

plt.figure(figsize=(10,8))

plt.scatter(x=centers['x'],y=centers['y'],s=100,c='black',marker='*')

sns.scatterplot(data=df, x ='Annual Income (k$)',y='Spending Score (1-100)',hue='Spending and Income Cluster',palette='tab10')

plt.savefig('clustering_bivaraiate.png')
```

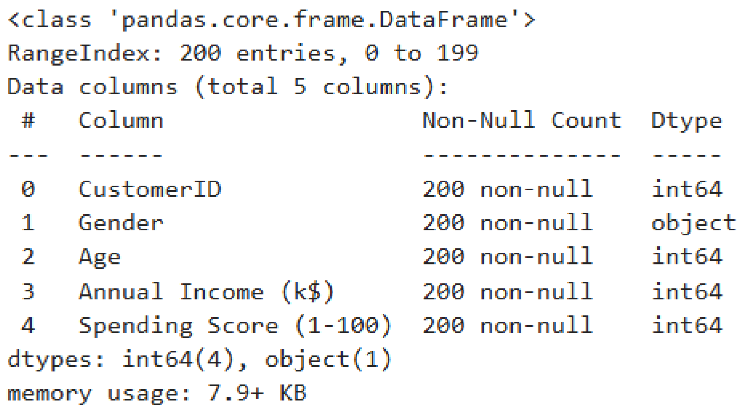
CHAPTER 7 : RESULT & DISCUSSION



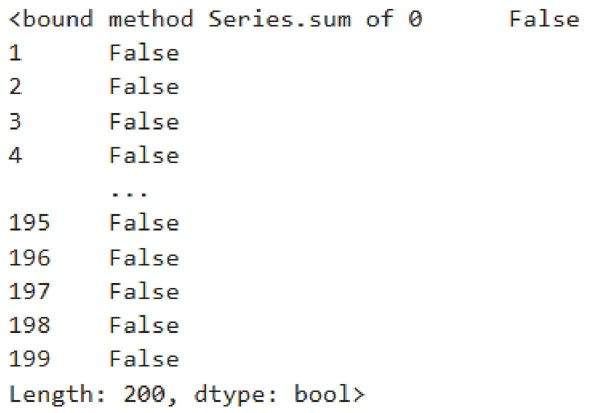
	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

200 rows × 5 columns

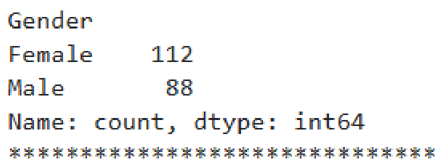
Table 7.1: Customer Demographics and Spending score Dataset



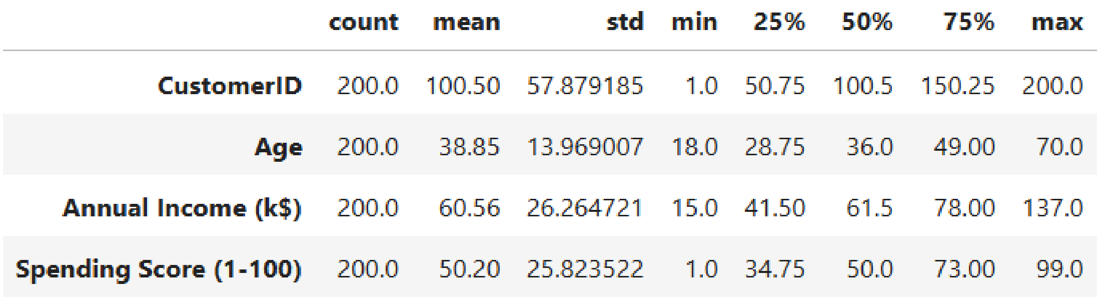
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   CustomerID              200 non-null    int64 
 1   Gender                  200 non-null    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
```



```
<bound method Series.sum of 0      False
1      False
2      False
3      False
4      False
       ...  
195    False
196    False
197    False
198    False
199    False
Length: 200, dtype: bool>
```



```
Gender
Female    112
Male       88
Name: count, dtype: int64
******************************
```



	count	mean	std	min	25%	50%	75%	max
CustomerID	200.0	100.50	57.879185	1.0	50.75	100.5	150.25	200.0
Age	200.0	38.85	13.969007	18.0	28.75	36.0	49.00	70.0
Annual Income (k$)	200.0	60.56	26.264721	15.0	41.50	61.5	78.00	137.0
Spending Score (1-100)	200.0	50.20	25.823522	1.0	34.75	50.0	73.00	99.0

Table 7.2: Statistical Summary of Customer Attributes (Age, Income, Spending Score)

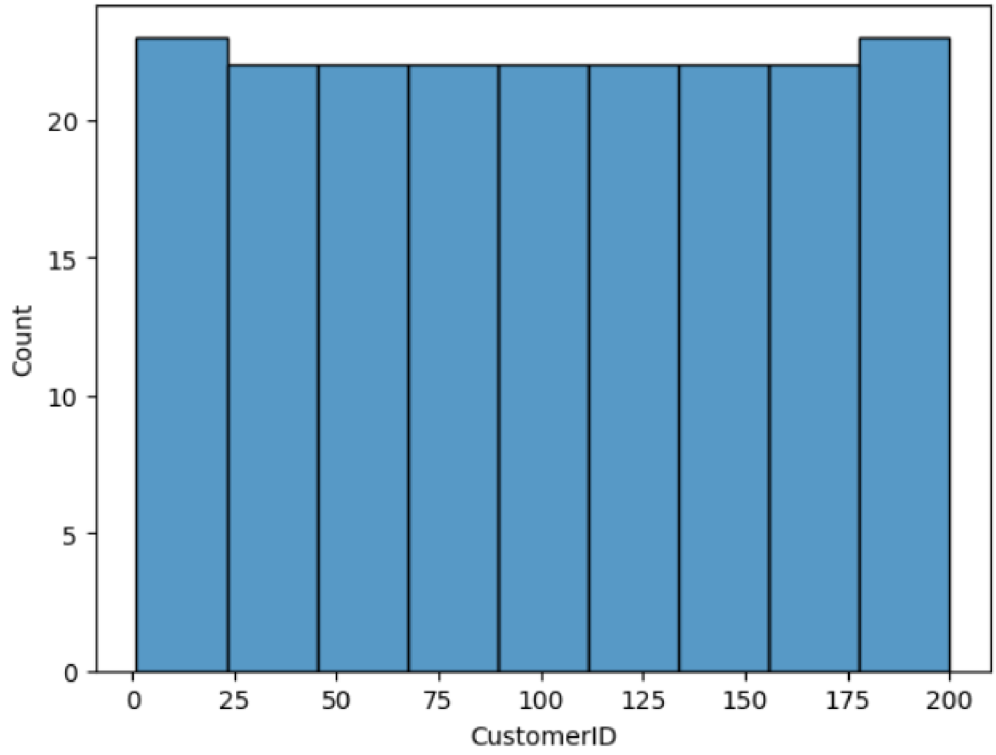


Figure 7.1: Histogram Showing Uniform Distribution of Customer IDs from 1-200

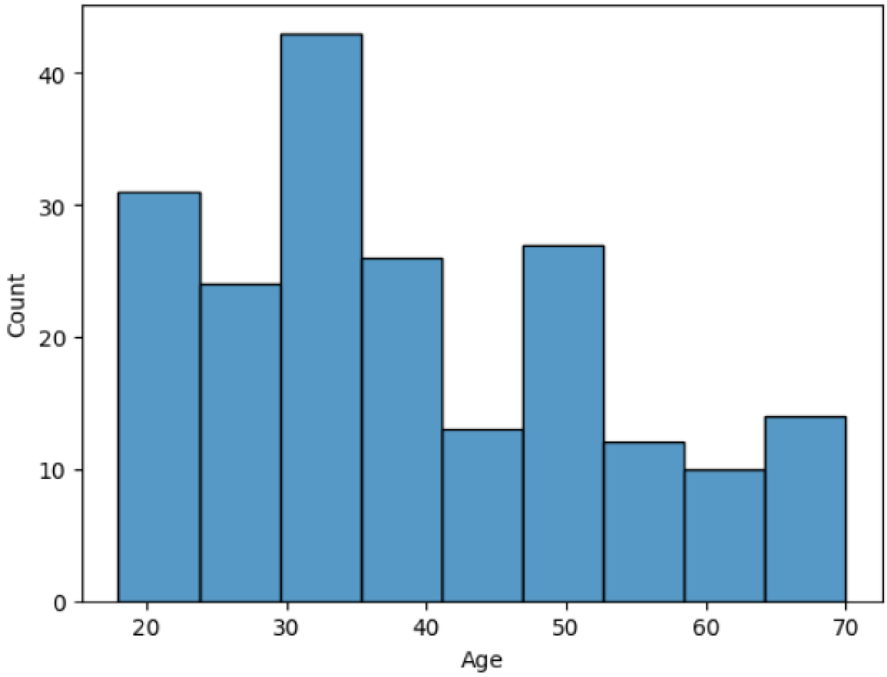


Figure 7.2: Histogram Showing Customer Age Distribution with Higher Frequency in the 30-Year Age Group

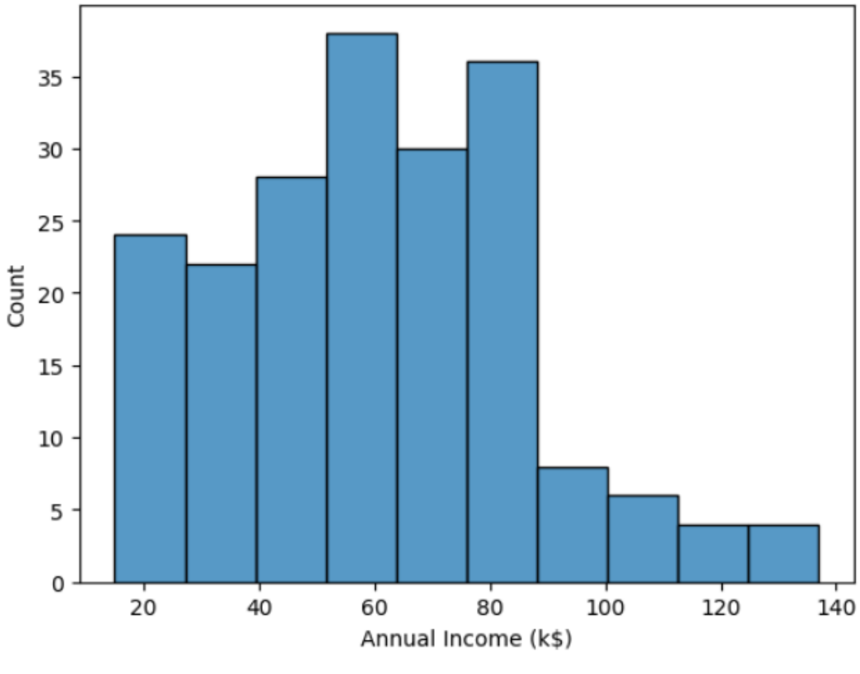


Figure 7.3: Histogram Showing Annual Income Distribution with Most Customers Earning Between 40k to 80k

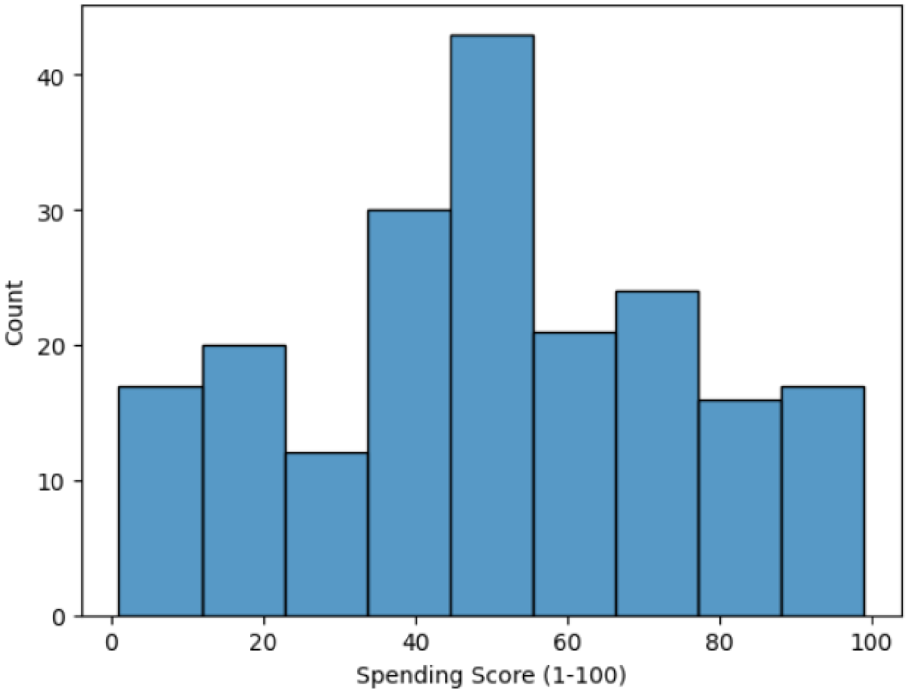


Figure 7.4: Histogram Showing Spending Score Distribution with More Customers Scoring Around 40 to 60

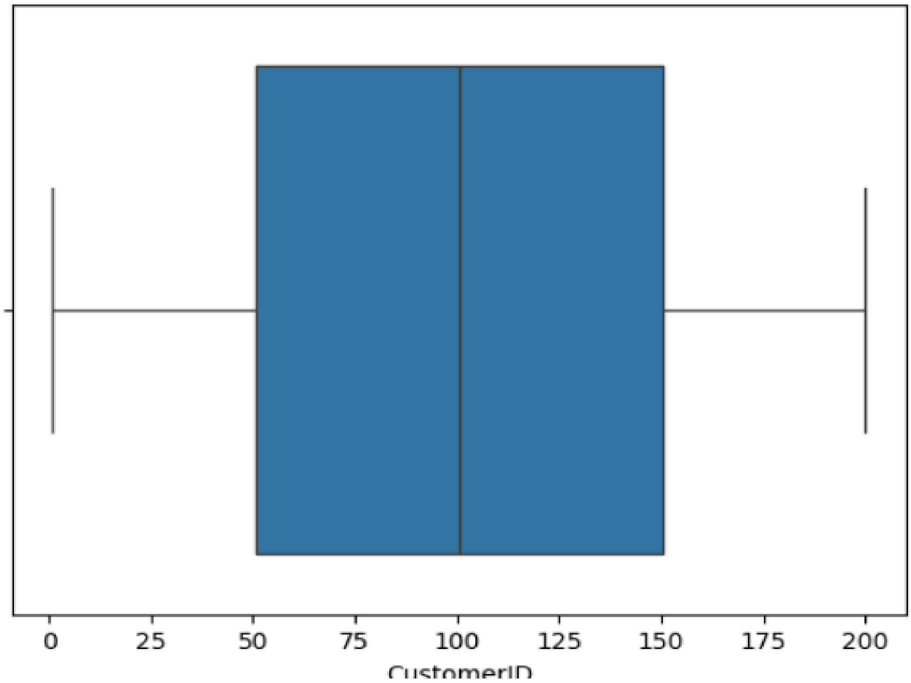


Figure 7.5: Box Plot of Customer IDs Showing Uniform Distribution with No Outliers

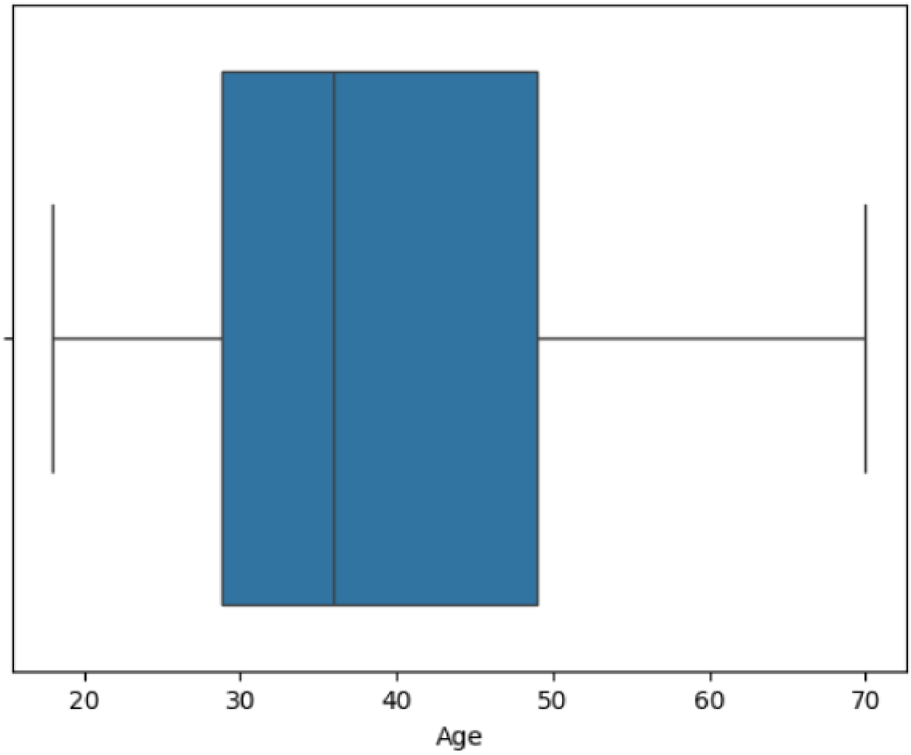


Figure 7.6: Box Plot of Customer Age Showing Symmetrical Distribution with no outliers

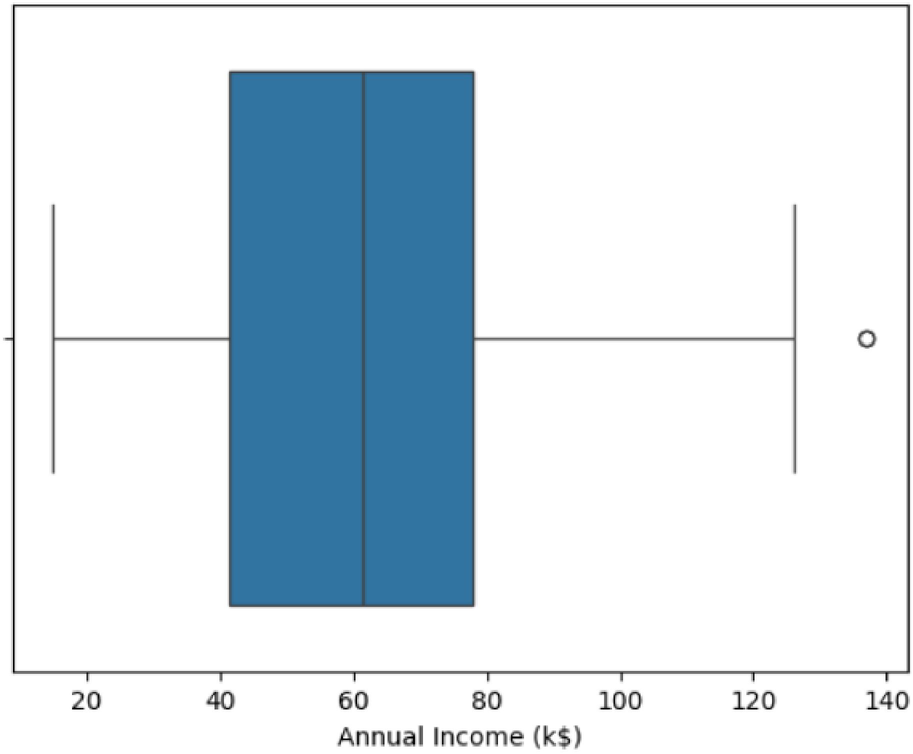


Figure 7.7: Box Plot of Annual Income Highlighting a Slight Outlier Above the Normal Range

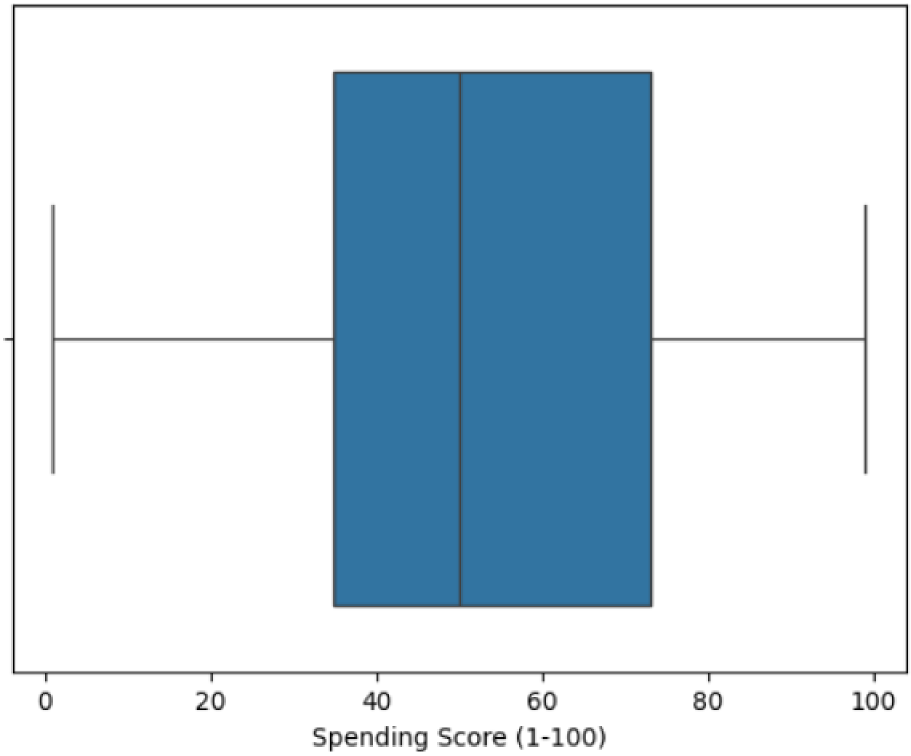
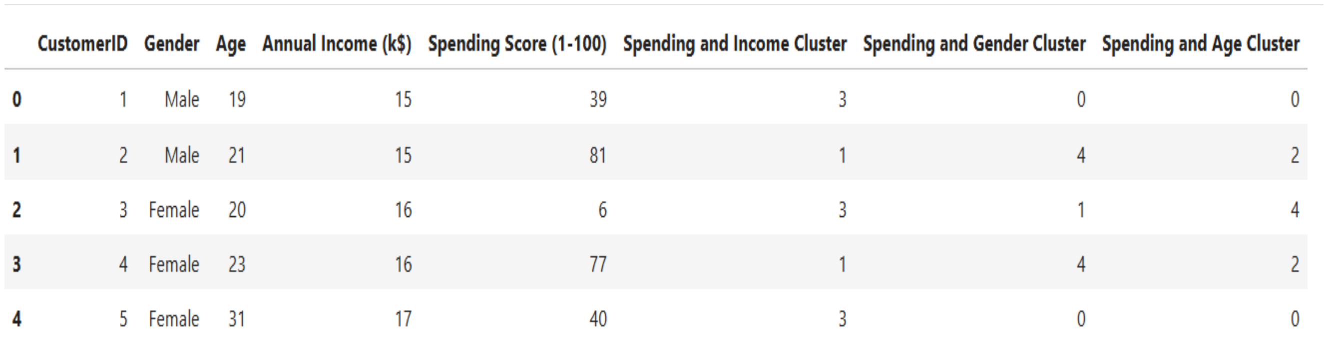


Figure 7.8: Box Plot of Spending Score Showing Even Distribution Without Outliers

- **Brief description of the above plots:**

  The following plots provide visual insights into the distribution of customer data. The histogram of Customer ID shows a uniform distribution from 1 to 200, confirming equal sampling. The age distribution histogram highlights a higher frequency in the younger age group, particularly around 30 years. The histogram of annual income reveals that most customers earn between 40k$ to 80k$, while the spending score histogram indicates a concentration of customers scoring around 40 to 60. Additionally, box plots show that the Customer ID and Age variables have symmetrical distributions with no outliers, supporting the consistency and cleanliness of the dataset.



	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)	Spending and Income Cluster	Spending and Gender Cluster	Spending and Age Cluster
0	1	Male	19	15	39	3	0	0
1	2	Male	21	15	81	1	4	2
2	3	Female	20	16	6	3	1	4
3	4	Female	23	16	77	1	4	2
4	5	Female	31	17	40	3	0	0

Table 7.3: Customer Data with Assigned Cluster Labels Based on Age and Spending Score Features

```
[<matplotlib.lines.Line2D at 0x25ffc2e51d0>]
```

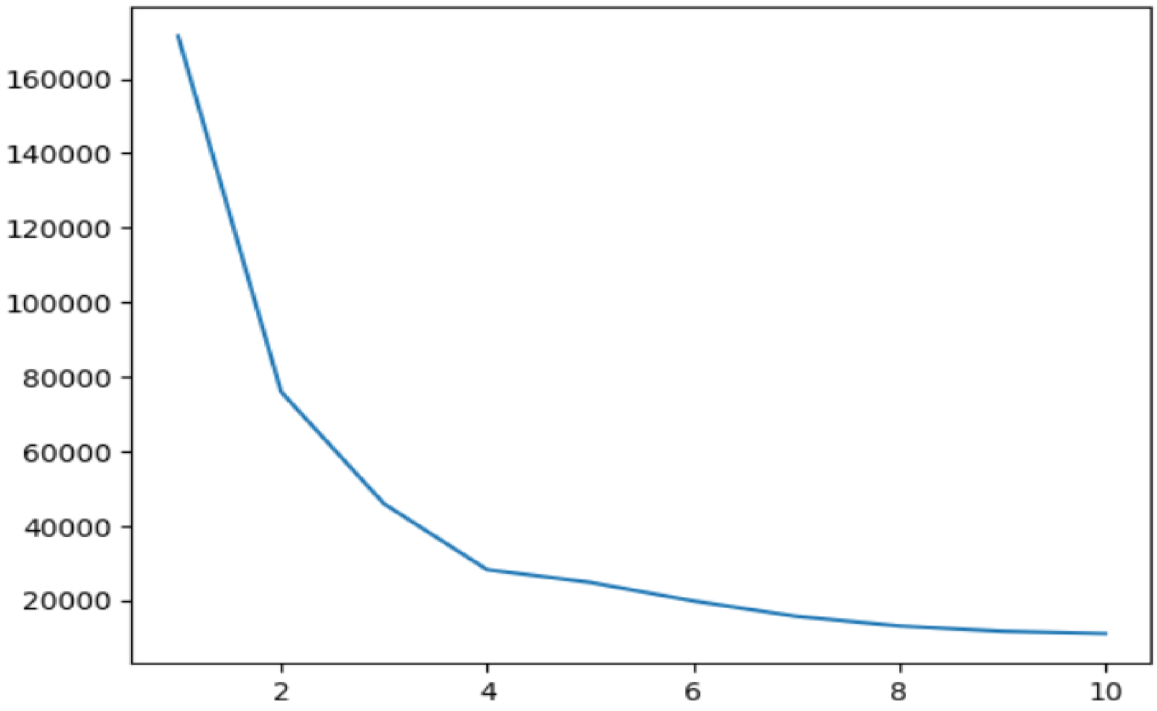


Figure 7.9: Elbow Curve Showing the Inertia Score Decrease to Identify Optimal Cluster Count

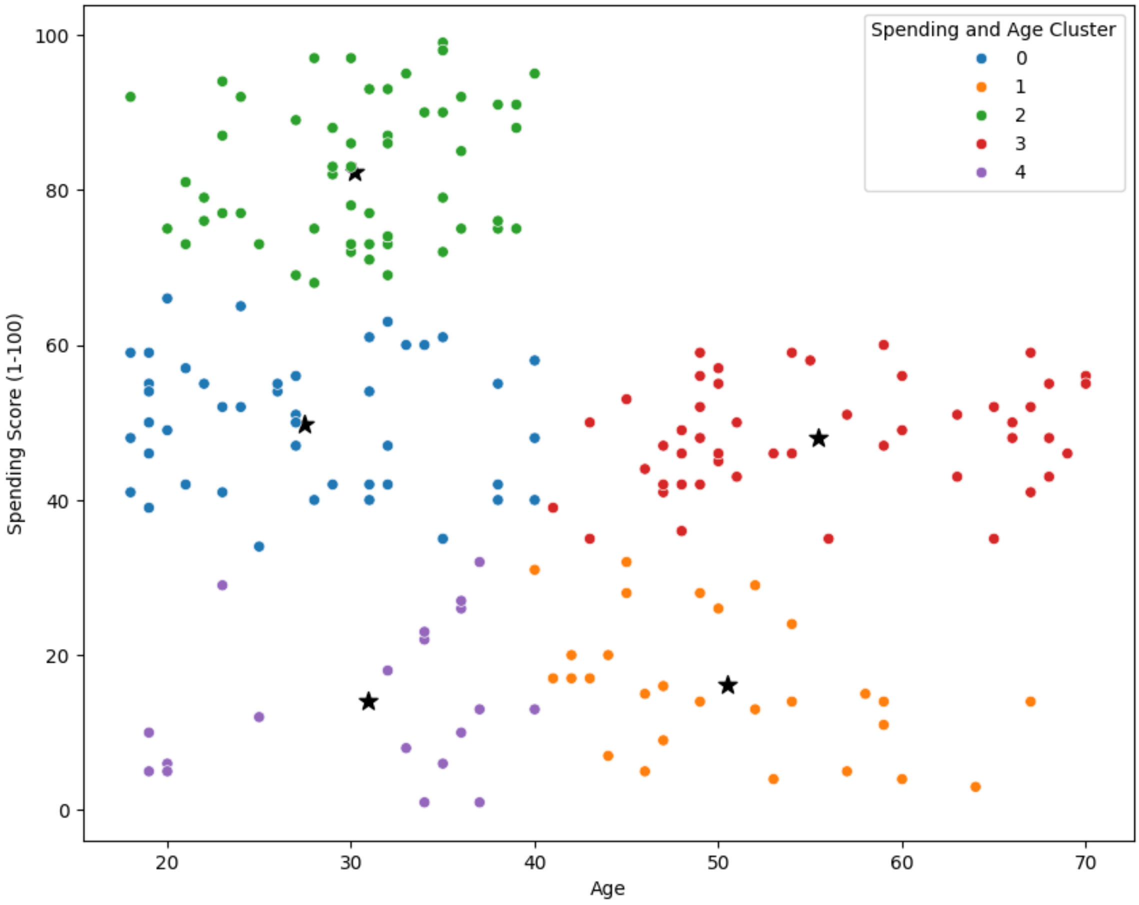


Figure 7.10: Scatter Plot Showing K-Means Clustering of Customers Based on Age And Spending Score with Cluster Centres Marked

- **Brief description of above Scatter Plot**

The scatter plot visualizes the results of K-Means clustering applied to customer data based on Age and Spending Score. Each point represents a customer, and the colours indicate the five distinct clusters identified by the algorithm. The stars denote the centroids of each cluster, indicating the average position of customers within that group. The plot reveals distinct customer segments: for instance, younger customers with high spending scores (green cluster), middle-aged customers with average spending (blue cluster), and older customers with lower spending scores (orange or purple clusters). This clustering helps in understanding customer behaviour and can guide targeted marketing strategies.

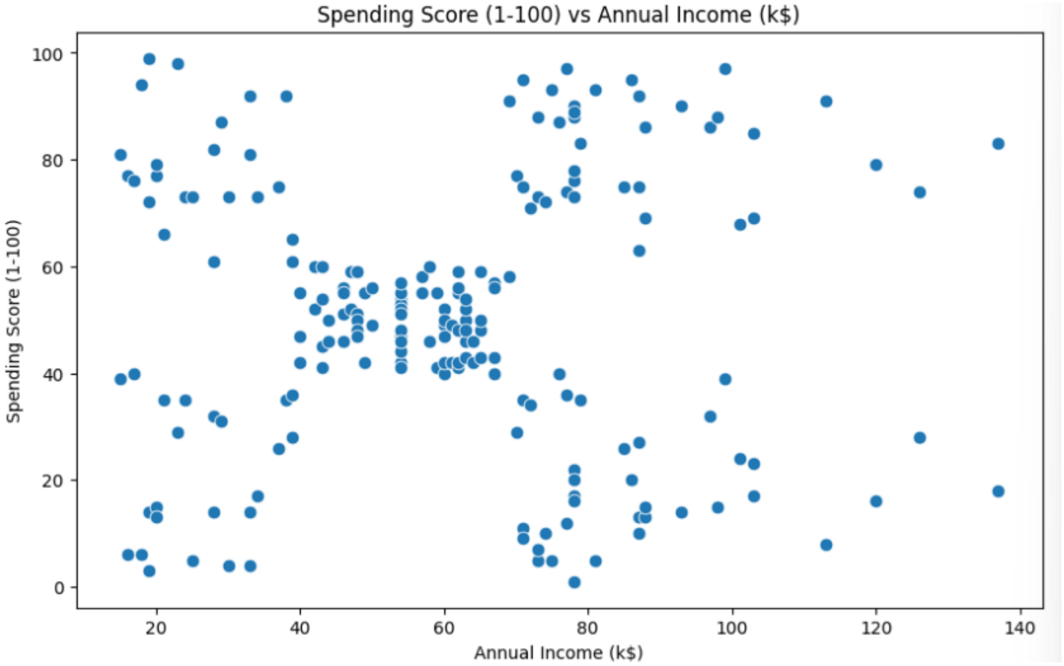
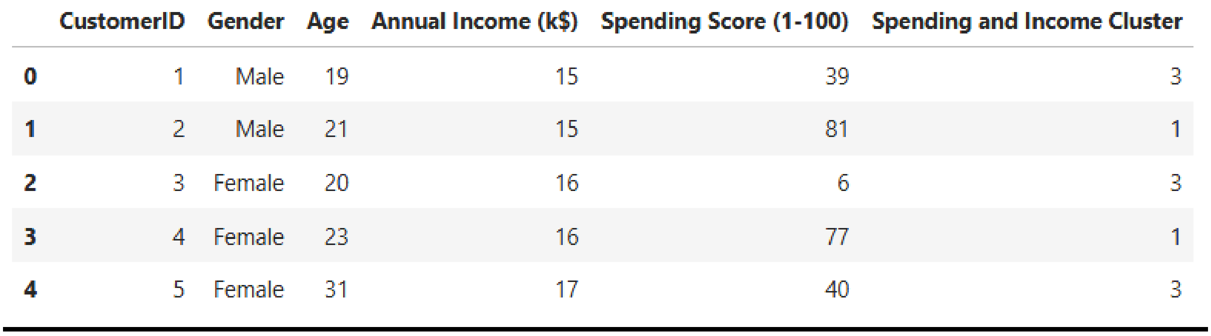


Figure 7.11: Scatter Plot Showing the Relationship Between Annual Income And Spending Score to Identify Customer Grouping Patterns



	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)	Spending and Income Cluster
0	1	Male	19	15	39	3
1	2	Male	21	15	81	1
2	3	Female	20	16	6	3
3	4	Female	23	16	77	1
4	5	Female	31	17	40	3

Table 7.4: Displaying Customer Details Along with Cluster Labels Based on Annual Income and Spending Score Using K-Mean

```
[<matplotlib.lines.Line2D at 0x25f80138550>]
```

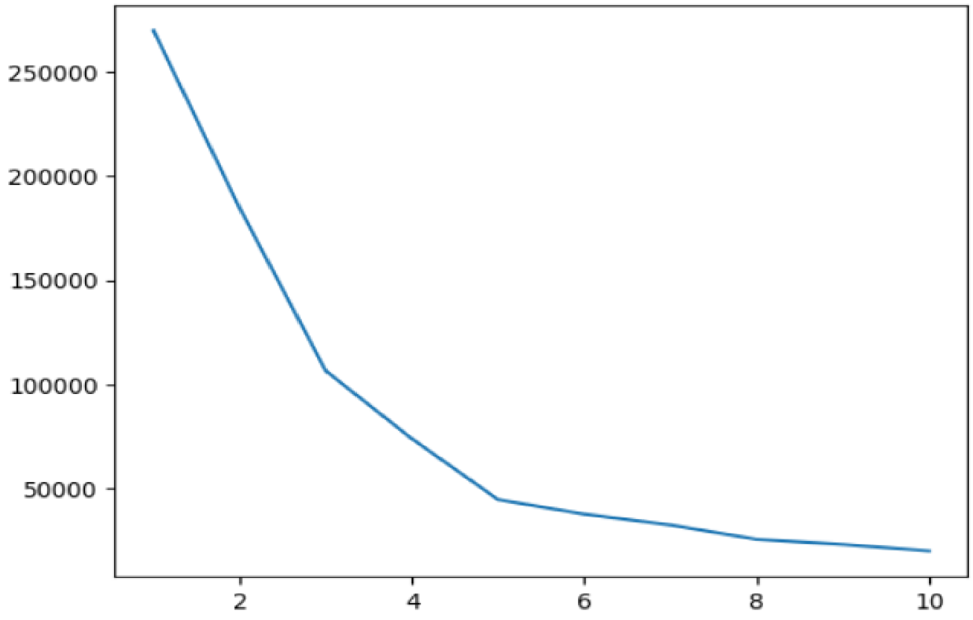


Figure 7.12: Elbow Curve Representing Inertia Values to Determine Optional Number of Clusters Based on Annual Income and Spending Score Features

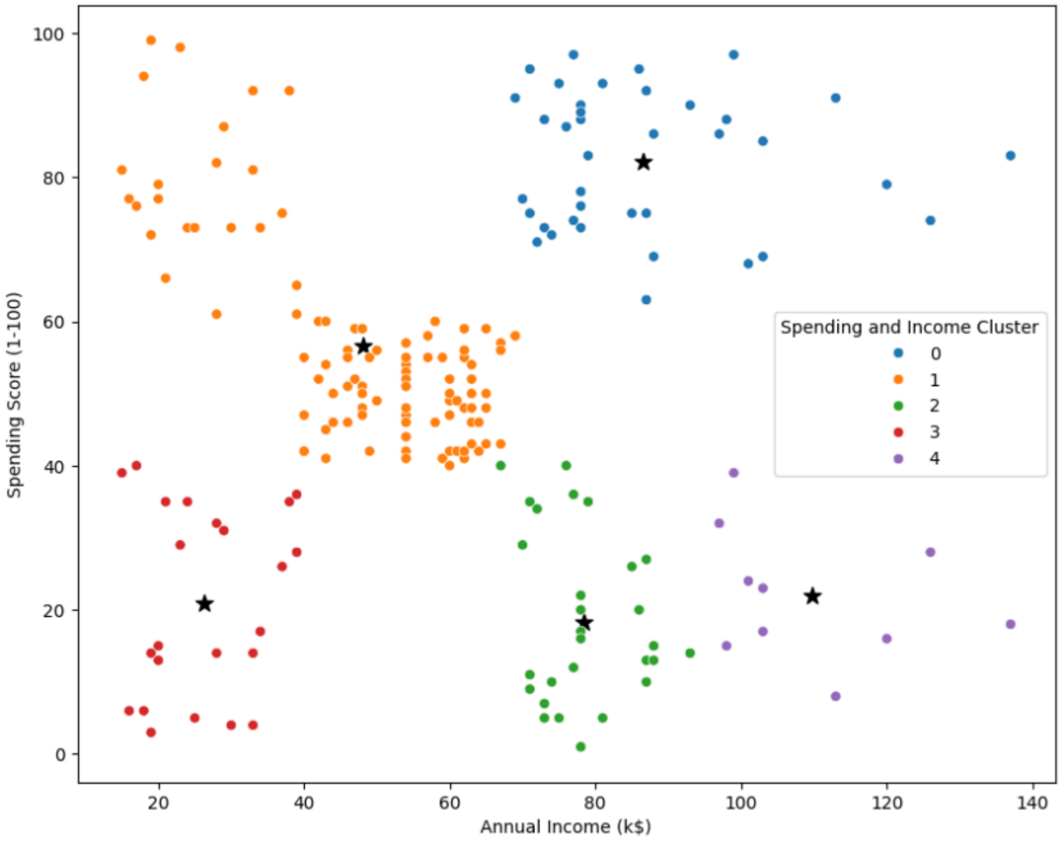


Figure 7.13: K-Means Clustering of Customers Based on Annual Income and Spending Score with Cluster Centres Marked in Black

- **Brief description of above Scatter Plot**

  The scatter plot illustrates the K-Means clustering of customers based on their Annual Income and Spending Score. Each point represents a customer and is coloured according to the cluster it belongs to, while the black stars indicate the centroids of each cluster. The plot effectively segments the customers into five distinct behavioural groups. For example, some clusters show high spending scores regardless of income, while others represent low spenders with either low or high income. Notably, one cluster (light blue) contains high-income customers with high spending, suggesting a premium customer segment. This clustering helps identify and target different consumer groups for strategic marketing and personalized services.

CHAPTER 8: CONCLUSION

To understand customer segmentation, clustering was performed using two feature combinations: Age vs. Spending Score and Annual Income vs. Spending Score. Upon analysing the scatter plots of both clustering, it was observed that the segmentation based on Annual Income and Spending Score provided clearer and more meaningful groupings of customers.

In the Age vs. Spending Score plot(Fig-6.11)clusters were more overlapping and less distinct, especially in the mid-age ranges, leading to ambiguity in separating customer behaviour patterns. On the other hand, the Spending vs. Income cluster plot displayed well-separated and denser groups, making it easier to identify distinct customer categories such as high income–low spending, low income–high spending, and moderate behaviour.

Therefore, the clustering based on Spending Score and Annual Income(Fig-6.14) was chosen as the final result due to its superior segmentation quality and better interpretability for targeted marketing strategies.

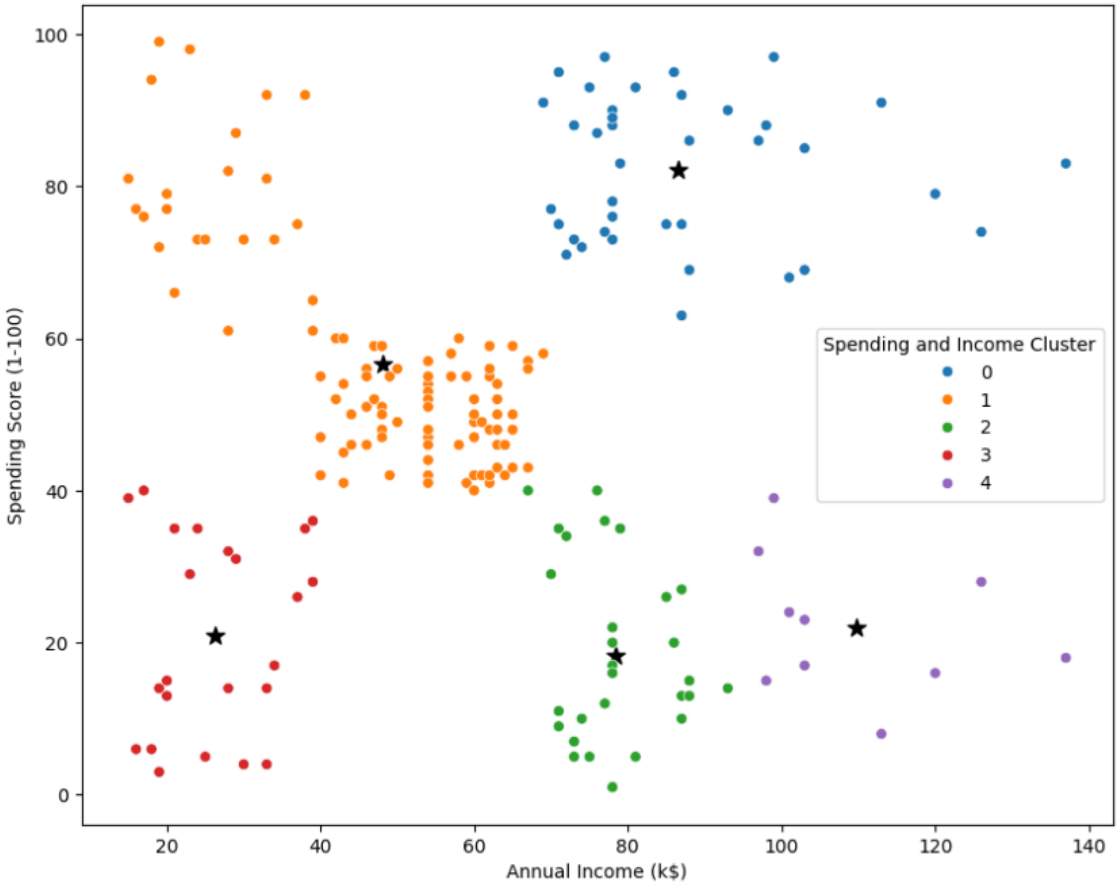


Figure 8.1: K-Means Clustering of Customers Based on Annual Income and Spending Score with Cluster Centres Marked in Black

After preprocessing the dataset and performing exploratory data analysis (EDA), clustering was applied using the K-Means algorithm. The Elbow Method indicated that the optimal number of clusters was 5, based on the point where the within-cluster sum of squares (WCSS) began to plateau. This ensured a good trade-off between model complexity and segmentation quality.

The final clusters showed distinct characteristics:

Blue (Cluster 0): High annual income, high spending score -Wealthy and high spenders.

Orange (Cluster 1): Average annual income, average to high spending score- Mid-income and moderate to high spenders.

Green (Cluster 2): High annual income, low spending score - Wealthy but careful spenders.

Red (Cluster 3): Low annual income, low spending score - Low income and low spenders.

Violet (Cluster 4): Very high annual income, low spending score - Very wealthy but minimal spenders.

Visualizations like 2D scatter plots showed clear separation between clusters, confirming the algorithm’s ability to capture meaningful patterns. Compared to Mini-Batch K-Means, the standard K-Means produced slightly more stable clusters but at higher computational cost. Hierarchical clustering provided a dendrogram representation but lacked scalability for larger datasets.

The segmentation outcomes highlight how unsupervised learning can support personalized marketing strategies. Retailers can now align product placement, advertising, and promotional strategies with specific customer groups, thereby optimizing resource allocation and improving ROI.

REFERENCES

BOOKS:-

1. **Jeeva Jose**, Machine Learning with Python, Khanna Publication

2. **S. Dutt, S. Chandramouli, A.K. Das**, Machine Learning, Pearson Publication

3. Python for Data Analysis" by Wes McKinney

4. "Data Science from Scratch: First Principles with Python" by Joel Grus

JOURNALS : -

1. **Kashwan, Kishana R., and C. M. Velu**. "Customer segmentation using clustering and data mining techniques." *International Journal of Computer Theory and Engineering* 5.6 (2013): 856.

2. **Tabianan, K.; Velu, S.; Ravi, V. K-**Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. Sustainability 2022, 14, 7243. https://doi.org/10.3390/ su14127243

3. **Pranjali Joshi, Anuj Mutha, Nidhi Patil, Chaitralee Datar, Sarang Agrawal**-MACHINE LEARNING ASSISTED PROFILING OF RETAIL BANKING CUSTOMERS. Juni Khyat ISSN: 2278-4632 (UGC Care Group I Listed Journal) Vol-13, Issue-05, No.03, May : 2023

4. KAI PENG(Member, IEEE), VICTOR C. M. LEUNG, (Fellow, IEEE), AND QINGJIA HUANG, "Clustering Approach based on Mini batch K-Means ", In 2018, College of Engineering, Huaqiao University, Quanzhou 362021, China.

5. Fionn Murtagh and Pedro Contreras, "Methods of Hierarchical Clustering", In 2018, Science Foundation Ireland, Wilton Place, Dublin, Ireland Department of Computer Science, Royal Holloway, University of London

6. D. P. Yash Kushwaha, Deepak Prajapati, “Customer Segmentation using K- Means Algorithm,” 8th Semester Student of Beach in Computer Science and Engineering, Galgotias University, India.