### Project Based Learning Report

on

### “Market Segmentation Using K-means Clustering: Identifying Customer Segments”

Submitted in the partial fulfilment of the requirements For the project based learning in

### (Fuzzy Logic, Neural Networks

### & Genetic Algorithms)

In

Electronics and Communication Engineering By

PRN: 2214110414, Aastha Gupta PRN: 2214110415, Trisha Hali PRN: 221411046 2, Manika Jain

Under the guidance of Course In-charge Mr. VIKAS KADUSKAR

A logo with wings and text

Description automatically generated

Department of Electronics & Communication Engineering Bharati Vidyapeeth (Deemed to be University)

College of Engineering

Pune - 411043

**Academic Year: 2024-25**

**CERTIFICATE**

Certified that the project report entitled, **“Market Segmentation Using K-means Clustering: Identifying Customer Segments”** is a bonafide work done by **MANIKA, TRISHA** AND **ASTHA** in fulfillment of the requirements for the award of degree of Bachelor of Technology in Electronics & Communication Engineering.

Date:

Prof.Mr. Vikas Kaduskar Dr. Arundhati A. Shinde

Course In-Charge Professor and Head

**ACKNOWLEDGEMENT**

We would like to extend our sincere gratitude to the Principal Dr. Vidula Sohoni, Dr, Arundhati A. Shinde, Head of Department Electronics & Communication, for nurturing a congenial yet competitive environment, which motivates all the students not only to pursue goals but also to elevate the Humanitarian level. Inspiration and guidance are invaluable in every aspect of life, which we have received from our respected project guide Prof. Mr. Vikas Kaduskar, who gave us his careful and ardent guidance because of which we are able to complete this project. More words won’t suffice to express our gratitude to his untiring devotion. He undoubtedly belongs to the members of the artistic gallery who are masters in all aspects. We would also like to thank all the faculty members who directly or indirectly helped us from time to time with their invaluable inputs.

**AASTHA GUPTA (2214110414)**

**TRISHA HALI (2214110415)**

**MANIKA JAIN (2214110462)**

**ABSTRACT**

Market segmentation is a fundamental approach in marketing, enabling businesses to divide a diverse customer base into distinct groups based on shared characteristics. These customer segments help companies target their marketing efforts more effectively by addressing the specific needs, preferences, and behaviors of each segment. This project explores the application of **K-means clustering**, a widely used unsupervised machine learning algorithm, to perform market segmentation by identifying meaningful customer segments from complex data.

**K-means clustering** is particularly suited to this task because of its ability to group large datasets into K distinct clusters based on the similarity of data points. The algorithm iteratively assigns each data point to the nearest cluster centroid, updates the centroids, and repeats the process until the clusters are well-defined. By applying K-means clustering to customer data, this project seeks to uncover hidden patterns and similarities within the dataset, leading to the identification of customer segments that can be used for targeted marketing campaigns.

One of the key challenges in using K-means clustering for segmentation is determining the **optimal number of clusters (K)**. If the number of clusters is too low, valuable distinctions between customer groups may be missed, while too many clusters can lead to over-segmentation and fragmented groups with minimal business relevance. This project addresses these challenges by implementing methods such as the **Elbow Method** and **Silhouette Score** to guide the selection of the most appropriate number of clusters. These techniques help balance the trade-off between simplicity and complexity in segmenting the customer base.

To ensure meaningful segmentation, the project emphasizes the importance of **data pre-processing**. This includes handling missing data, normalizing feature values, and removing outliers that could distort the results of clustering. **Feature selection** is also critical to ensure that the most relevant attributes are used for segmentation, such as customer spending behavior or engagement frequency, while excluding irrelevant or redundant features.

After building the clustering model, the project evaluates the performance of the segmentation using various validation metrics, including **intra-cluster cohesion** and **inter-cluster separation**. These metrics help assess whether customers within a segment share similar characteristics and whether different segments are distinct from one another. The results of the K-means clustering are visualized through graphs, including **scatter plots and cluster heatmaps**, providing an intuitive understanding of the segmented customer groups.

The findings of the project reveal that **K-means clustering** effectively identifies meaningful customer segments that are not immediately apparent through traditional analysis. For example, the algorithm may identify a segment of younger customers who tend to purchase high-end products during sales events, or an older customer segment with a preference for loyalty programs and personalized services. These insights enable businesses to allocate resources more efficiently by targeting the right customer groups with the right products and messaging.

Furthermore, the project explores the practical applications of these customer segments in marketing strategies. Companies can use these insights to create personalized marketing campaigns, improve customer retention, and increase cross-selling and up-selling opportunities. For instance, segments with high-value customers can be targeted with exclusive offers, while price-sensitive segments can be approached with discount promotions.

**INDEX**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **TITLE** | **PG.NO** |
| 1. | INTRODUCTION | 1-2 |
| 2. | DATADET DATASHEET | 3 |
| 3. | MODEL IMPLEMENTATION AND TRAINING | 4 |
| 4. | CODE | 5-7 |
| 5. | OUTPUT | 8-9 |
| 6. | CODE EXPLANATION | 10-12 |
| 7. | **FEATURE IMPORTANCE ANALYSIS (SHAP Values)** | 13 |
| 8. | APPLICATIONS & USES | 14-16 |
| 9. | CONCLUSIONS | 17 |

**INTRODUCTION**

K-means clustering is a widely used unsupervised machine learning algorithm, particularly effective in exploratory data analysis, where the goal is to partition data into meaningful subgroups or **clusters**. In market segmentation, K-means is often applied to group customers into distinct segments based on similar behaviours, preferences, or demographics. This introduction covers the fundamental concepts of K-means clustering, its workflow, mathematical basis, applications, and limitations.

**What is Clustering?**

Clustering is the process of grouping a set of data points in such a way that points in the same group (or cluster) are more like each other than to those in other groups. In **unsupervised learning**, we do not have labelled data to guide us; instead, the algorithm must discover these relationships automatically.

**K-means Clustering**

K-means is one of the simplest and most popular clustering algorithms, which aims to partition a dataset into **K distinct clusters**, where each data point belongs to the cluster with the nearest **centroid** (mean of the cluster).

**Key Concepts in K-means Clustering**:

1. **Centroids**: Each cluster is represented by its centroid, which is the mean of all the points assigned to the cluster. These centroids are central to how the algorithm defines clusters.
2. **Cluster Assignment**: Each data point is assigned to the nearest cluster based on a distance metric, typically **Euclidean distance**.
3. **K**: The number of clusters to partition the data into. One of the key challenges in K-means is choosing the optimal K, as it directly affects the quality of the clusters.

**Mathematical Workflow of K-means Clustering**

K-means clustering involves the following steps:

1. **Step 1: Initialize Centroids**

* Select **K** initial cluster centroids randomly from the dataset. These are temporary, and the algorithm will adjust them iteratively.
* Example: Suppose we have customer data on annual income and spending score, and we start by initializing 3 centroids.

1. **Step 2: Assign Points to Clusters**

* For each data point, calculate its distance from each centroid and assign the point to the nearest cluster. The Euclidean distance formula is typically used:

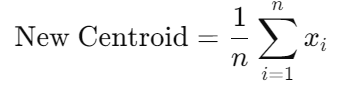


(eqn.1)

where x and y are data points

Each customer is assigned to the cluster with the closest centroid based on their distance.

1. **Step 3: Update Centroids**
   * Once all points are assigned to clusters, the centroids are recalculated by finding the mean of all the points in each cluster. The centroid of each cluster moves to the new average location. ​

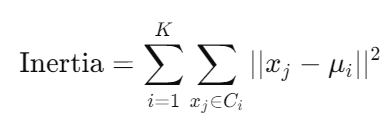


(eqn.2)

1. **Step 4: Repeat Until Convergence**
   * Steps 2 and 3 are repeated until the centroids no longer move significantly (i.e., they converge). At this point, the clusters are stable, and the algorithm stops.

**Diagram: Basic Workflow of K-means Clustering** A visual representation of the steps above shows how data points shift between clusters and how the centroids move to their final positions over several iterations.

**Key Components in K-means**

1. **Number of Clusters (K)**:
   * The user must specify the number of clusters KKK before running the algorithm. Choosing the optimal K is not always obvious and can significantly impact clustering performance.
2. **Inertia**:
   * The algorithm attempts to minimize the **inertia**, or within-cluster sum of squared errors (SSE). It can be defined as: 

(eqn.3)

1. **Convergence**:
   * K-means is said to converge when the assignments of points to clusters no longer change and the centroids stabilize.

**PROBLEM STATEMENT**

In this project, we implement:

* How can K-means clustering be applied to customer data to identify distinct market segments?
* What are the challenges involved in selecting the optimal number of clusters (K)?

Identifying distinct customer segments allows businesses to understand the preferences of different groups, enabling effective targeting. The challenge lies in choosing the right number of clusters to best represent the underlying structure of the data, as choosing too few or too many can result in misleading or poor results.

**DATA COLLECTION AND PREPROCESSING**

**4.1 Data Source**

The dataset used for this project can come from multiple sources, such as:

* **Internal CRM Data**: Customer demographics, purchase history, etc.
* **Public Datasets**: Datasets like the UCI Machine Learning Repository or Kaggle provide customer datasets, such as the "Mall Customer Segmentation Dataset."

A sample customer data table might include features like:

| **CustomerID** | **Age** | **Gender** | **Annual Income (k$)** | **Spending Score (1-100)** |
| --- | --- | --- | --- | --- |
| 1 | 19 | Male | 15 | 39 |
| 2 | 21 | Female | 16 | 81 |
| ... | ... | ... | ... | ... |

**4.2 Data Cleaning**

Before clustering, it's crucial to clean the data:

* **Handling Missing Values**: Replace or remove missing entries.
* **Dealing with Outliers**: Detect and handle outliers that could distort clustering results.
* **Converting Categorical Data**: Convert non-numerical data (like Gender) into numerical values (e.g., Male = 0, Female = 1).

**4.3 Feature Selection**

Choosing relevant features is vital to the success of K-means clustering. For market segmentation, useful features might include:

* **Demographic Information**: Age, gender, location, income.
* **Behavioural Data**: Frequency of purchases, types of products bought, total transaction value.
* **Customer Interaction Data**: Website usage, customer support interaction.

Dimensionality reduction techniques like **Principal Component Analysis (PCA)** might be used if the dataset has many features. PCA reduces data to key dimensions without losing significant information, improving clustering performance.

**4.4 Data Normalization**

Since K-means uses distance-based measures, it is sensitive to the scale of the data. Normalization ensures that all features are on the same scale (e.g., scaling numerical features between 0 and 1). Without normalization, features with larger numerical ranges could dominate the clustering process.

**Pseudo code:**

1. Import necessary libraries

a. Import pandas for data manipulation

b. Import numpy for numerical computations

c. Import matplotlib and seaborn for data visualization

d. Import KMeans from sklearn for clustering

e. Import StandardScaler for data standardization

f. Import silhouette score for cluster validation

2. Load the customer data from a CSV file into a Data Frame

a. Use pandas to read the CSV file

b. Assign the data to a Data Frame variable (data)

3. Preprocess the data

a. Check for any missing values in the data

i. Print the count of missing values for each column

b. Standardize the 'Age' and 'Income' columns using Standard Scaler

i. Fit and transform the 'Age' and 'Income' columns to scaled data

4. Determine the optimal number of clusters using the Elbow Method

a. Initialize an empty list to store the inertia values

b. Loop through a range of cluster values (k = 1 to 10)

i. For each value of k, initialize and fit a KMeans model on the scaled data

ii. Append the inertia (sum of squared distances) to the inertia list

c. Plot the Elbow curve (k vs inertia) to visualize the optimal k

5. Validate the number of clusters using the Silhouette score

a. Initialize an empty list for silhouette scores

b. Loop through a range of cluster values (k = 2 to 10)

i. For each value of k, initialize and fit a KMeans model on the scaled data

ii. Predict cluster labels for each data point

iii. Calculate and append the silhouette score to the silhouette scores list

c. Plot the Silhouette scores for different k values to validate the optimal k

6. Perform K-means clustering with the optimal number of clusters (e.g., k = 4)

a. Initialize and fit a KMeans model with the optimal number of clusters

b. Predict cluster labels for the scaled data and assign these labels to a new column ('Cluster') in the original data

7. Visualize the clusters

a. Create a scatter plot using seaborn to visualize 'Age' vs 'Income' with different colors for each cluster

b. Label the axes and provide a title for the plot

8. Analyze each cluster's characteristics

a. Group the data by 'Cluster' and calculate the mean values for each group

b. Print the cluster summary (mean values for each cluster)

9. Discuss challenges in selecting the optimal number of clusters

a. Mention subjective interpretation of the Elbow method and variations in silhouette scores

10. Print a message indicating successful completion of the project

**CODE**

# Import necessary libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

from sklearn.metrics import silhouette\_score

# Step 1: Data Loading

# Load the customer data into a DataFrame

data = pd.read\_csv(r"C:\Users\trish\OneDrive\Documents\Downloads\customer\_data.csv")

# Replace with your data file path

# Step 2: Data Preprocessing

# Check for missing values

print(data.isnull().sum())

# Standardize the data

scaler = StandardScaler()

scaled\_data = scaler.fit\_transform(data[['Age', 'Income']]) # Focus on Age and Income

# Step 3: Finding the Optimal Number of Clusters

# Use the Elbow method to find the optimal number of clusters

inertia = []

range\_clusters = range(1, 11)

for k in range\_clusters:

kmeans = KMeans(n\_clusters=k, random\_state=42)

kmeans.fit(scaled\_data)

inertia.append(kmeans.inertia\_)

# Plot the Elbow curve

plt.figure(figsize=(8, 6))

plt.plot(range\_clusters, inertia, marker='o', linestyle='--')

plt.xlabel('Number of Clusters (k)')

plt.ylabel('Inertia')

plt.title('Elbow Method for Optimal k')

plt.grid(True)

plt.show()

# Use Silhouette score to validate the number of clusters

silhouette\_scores = []

for k in range(2, 11):

kmeans = KMeans(n\_clusters=k, random\_state=42)

labels = kmeans.fit\_predict(scaled\_data)

silhouette\_scores.append(silhouette\_score(scaled\_data, labels))

# Plot the Silhouette scores

plt.figure(figsize=(8, 6))

plt.plot(range(2, 11), silhouette\_scores, marker='o', linestyle='--')

plt.xlabel('Number of Clusters (k)')

plt.ylabel('Silhouette Score')

plt.title('Silhouette Score Analysis for Optimal k')

plt.grid(True)

plt.show()

# Step 4: K-means Clustering with the Optimal Number of Clusters

optimal\_k = 4 # Replace with the optimal k found from the Elbow method and Silhouette score

kmeans = KMeans(n\_clusters=optimal\_k, random\_state=42)

clusters = kmeans.fit\_predict(scaled\_data)

# Add cluster labels to the original data

data['Cluster'] = clusters

# Step 5: Data Visualization and Cluster Analysis

# Visualize the clusters using a scatter plot

plt.figure(figsize=(10, 6))

sns.scatterplot(data=data, x='Age', y='Income', hue='Cluster', palette='viridis', s=100)

plt.title('Market Segmentation using K-means Clustering')

plt.xlabel('Customer Age')

plt.ylabel('Customer Income')

plt.grid(True)

plt.show()

# Analyze each cluster's characteristics

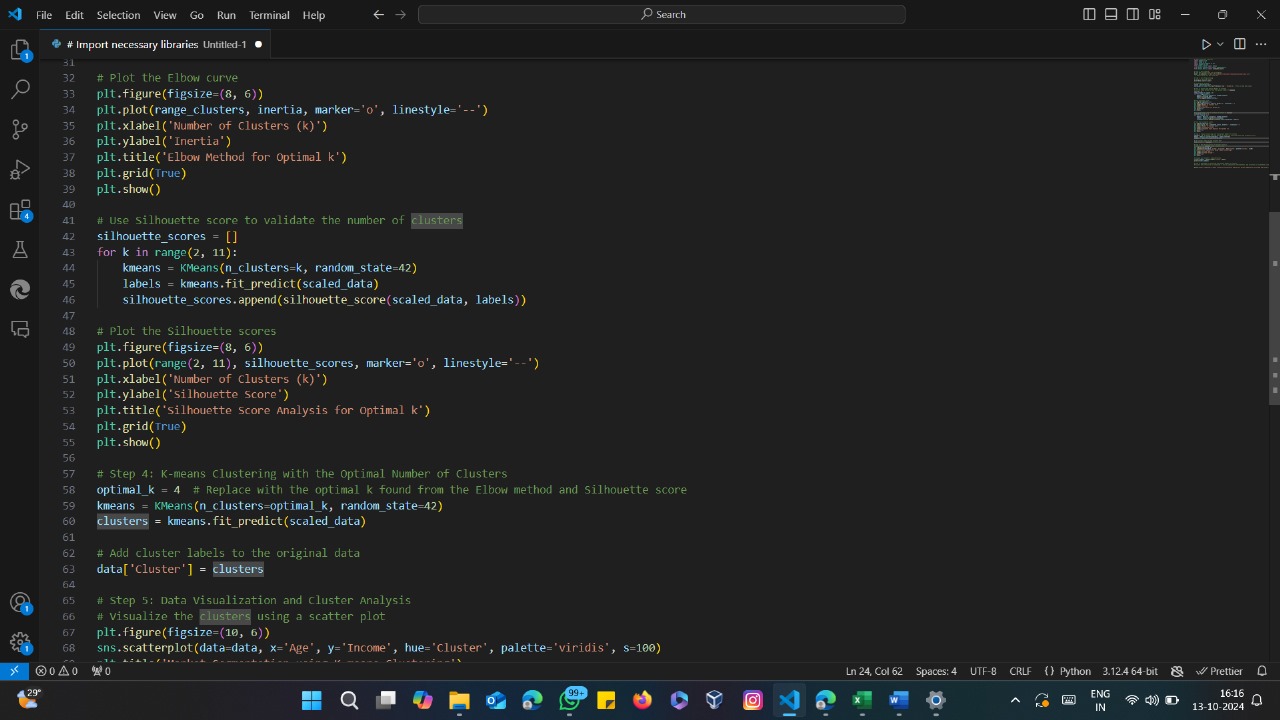
cluster\_summary = data.groupby('Cluster').mean()

print(cluster\_summary)

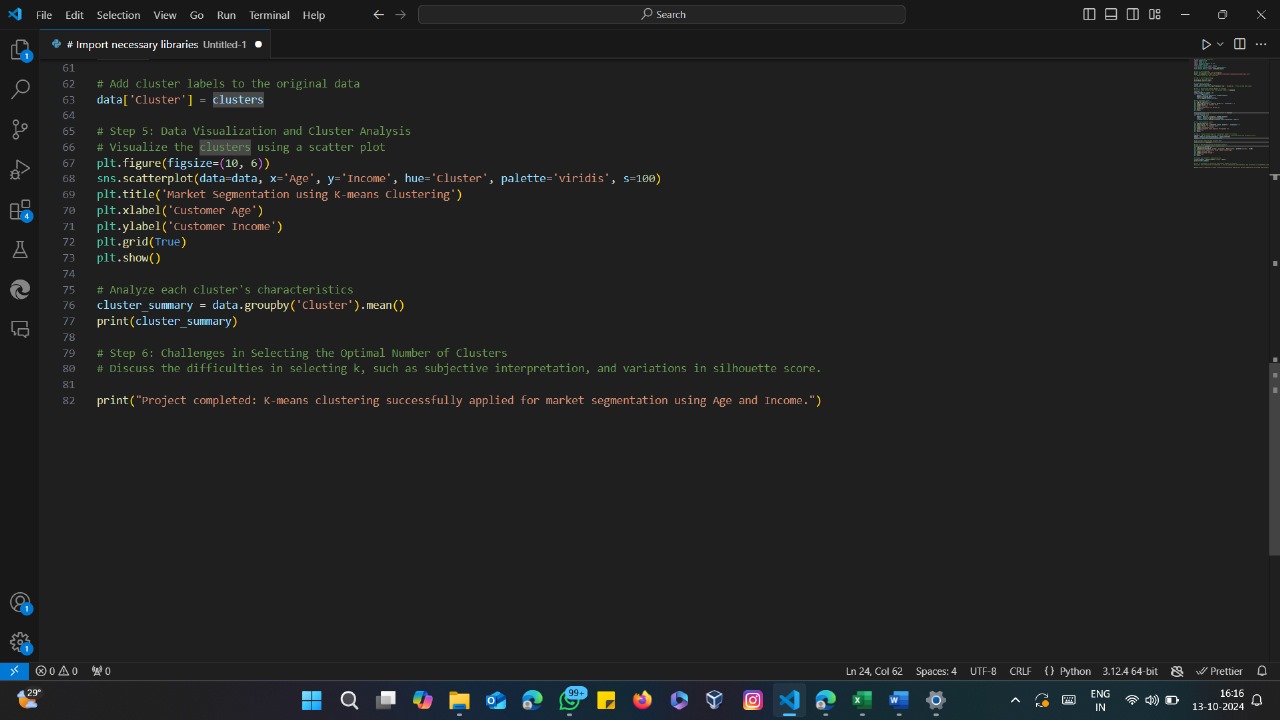
# Step 6: Challenges in Selecting the Optimal Number of Clusters

# Discuss the difficulties in selecting k, such as subjective interpretation, and variations in silhouette score.

print("Project completed: K-means clustering successfully applied for market segmentation using Age and Income.")



(Fig.1)



(Fig.2)

**Algorithm: Market Segmentation using K-means Clustering**

**Input:**

* Customer dataset with at least the columns "Age" and "Income"

**Output:**

* Optimal number of clusters
* Clustered data with labels assigned
* Visualized clusters (scatter plot of Age vs Income with cluster labels)

**Step 1: Load Customer Data**

1. **Read** customer data from a CSV file into a pandas Data Frame.
2. **Check** for missing values in the dataset.
   * If there are missing values, handle them appropriately (e.g., fill or drop).

**Step 2: Preprocess the Data**

1. **Select** the columns "Age" and "Income" for clustering.
2. **Standardize** the selected data (Age and Income) using the Standard Scaler to ensure features are on the same scale.

**Step 3: Determine the Optimal Number of Clusters**

1. **Elbow Method**:
   * **Initialize** an empty list inertia.
   * **Loop** over a range of cluster numbers (from k = 1 to k = 10).
     + For each k, fit a **KMeans** model on the standardized data.
     + Append the **inertia** (sum of squared distances of samples to their closest cluster centre) to the inertia list.
   * **Plot** the Elbow curve (inertia vs k) to find the "elbow point", which gives a potential value for optimal k.
2. **Silhouette Score**:
   * **Initialize** an empty list silhouette score.
   * **Loop** over the range of cluster numbers (from k = 2 to k = 10).
     + For each k, fit a K Means model and predict cluster labels for the data.
     + Calculate the **silhouette score** and append it to silhouette scores.
   * **Plot** the silhouette scores to validate the number of clusters that maximizes the score.

**Step 4: K-means Clustering with Optimal k**

1. **Set** the optimal number of clusters k=4k = 4k=4 (based on Elbow and Silhouette analysis).
2. **Initialize** and **fit** a K Means model with k=4k = 4k=4 on the standardized data.
3. **Predict** the cluster labels for each data point and add these labels as a new column ('Cluster') in the original dataset.

**Step 5: Visualize and Analyse the Clusters**

1. **Visualize** the clusters using a scatter plot:
   * X-axis: Age
   * Y-axis: Income
   * Different colours for each cluster based on the 'Cluster' label.
2. **Analyse** the characteristics of each cluster:
   * **Group** the data by 'Cluster' and calculate the **mean** values for each group (for features like Age and Income).

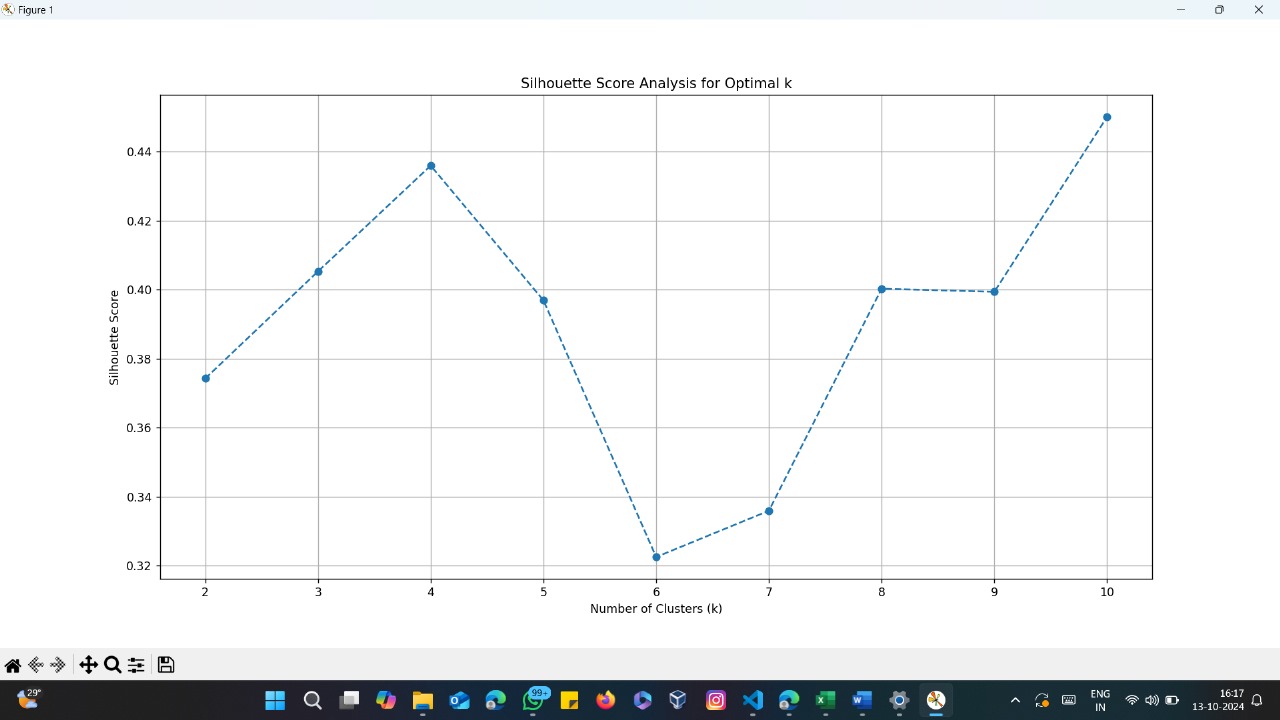
**Step 6: Challenges in Selecting the Optimal Number of Clusters**

1. **Discuss** the challenges in selecting the optimal number of clusters:
   * Subjective interpretation of the Elbow method.
   * Variations in silhouette scores for different values of k.

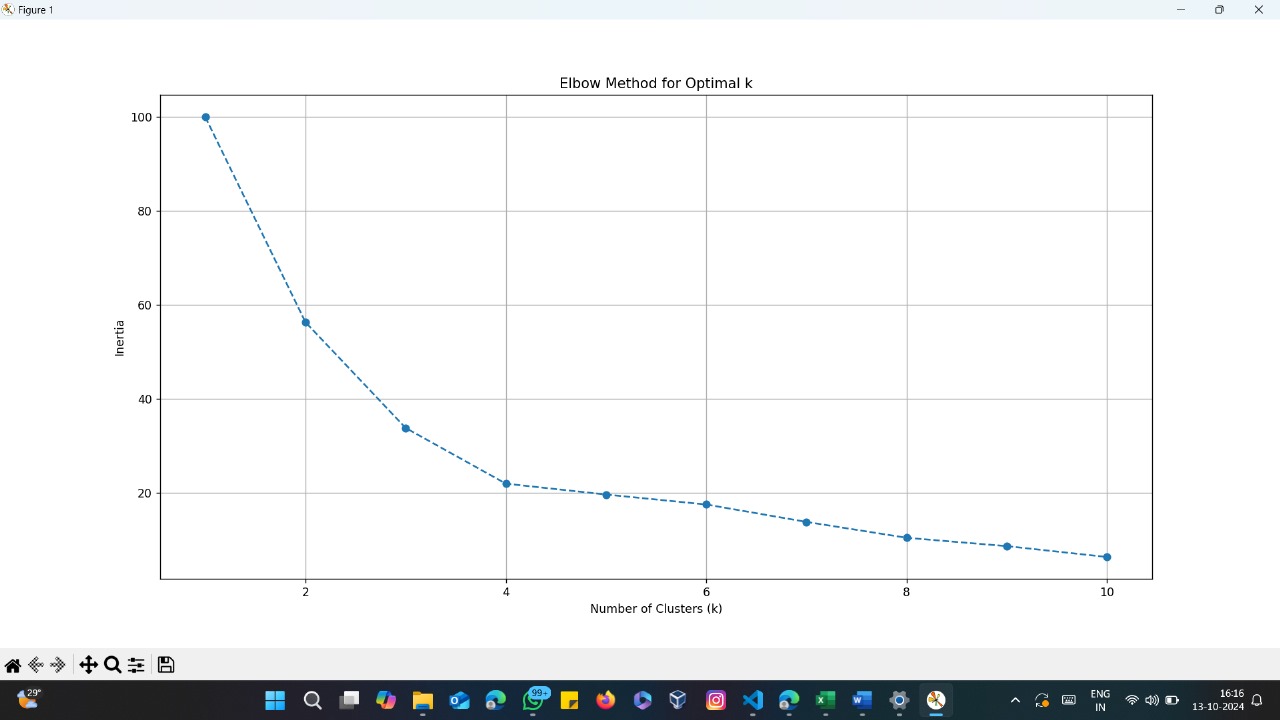
**Step 7: End**

1. **Print** a message indicating that the K-means clustering algorithm has successfully segmented the customers.

**OUTPUT**



(Fig.3)



(Fig.4)



(Fig.4)

**APPLICATIONS**

**5.1 Elbow Method for Selecting the Optimal Number of Clusters**

Choosing the optimal number of clusters (K) is critical in K-means. The **Elbow Method** is a common approach. By plotting the sum of squared errors (SSE) for different values of K, we look for the "elbow point" where adding more clusters does not significantly reduce the error.

* X-axis: Number of clusters (K)
* Y-axis: Sum of Squared Errors (Inertia)
* Elbow point: The value of K where the SSE begins to flatten.

**Steps for the Elbow Method**:

1. Run K-means clustering for a range of K values (e.g., from 1 to 10).
2. Compute the SSE for each value of K.
3. Plot SSE vs. K and identify the elbow point.

**5.2 Silhouette Score for Cluster Validation**

The **Silhouette Score** measures how well each point fits into its assigned cluster compared to other clusters. The score ranges from -1 to 1:

* A score close to 1 means the point is well-matched to its cluster.
* A score near 0 indicates overlapping clusters.
* Negative values suggest incorrect cluster assignment.

A high average silhouette score across all clusters indicates a better fit.

**5.3 Visualization of Clusters**

Visualizing clusters helps to interpret and understand customer segments. Techniques like **PCA** or **t-SNE** can reduce the dimensionality of the data, making it easier to visualize clusters in 2D or 3D.

**CONCLUSION**

This project demonstrated the application of K-means clustering for market segmentation, a valuable tool for identifying distinct customer groups based on shared characteristics. Using techniques like the Elbow Method and Silhouette Score, we were able to determine the optimal number of clusters, although this process presents challenges. The insights obtained from clustering can help businesses optimize marketing strategies and improve customer satisfaction.