CUSTOMER SEGMENTATION USING UNSUPERVISED LEARNING

A PROJECT REPORT

For

Introduction to AI(AI101B)

Session (2024-25)

Submitted by

Tanishka Gupta 202410116100218
Shipra Upadhyay 202410116100196
Vanshika Garg 202410116100235
Tripti Rajput 202410116100225

Submitted in partial fulfilment of the

Requirements for the Degree of

MASTER OF COMPUTER APPLICATION

Under the supervision of

Ms. KOMAL SALGOTRA
Assistant Professor

Submitted to

DEPARTMENT OF COMPUTER APPLICATIONS
KIET GROUP OF INSTITUTIONS, DELHI-NCR, GHAZIABAD-201206
(APRIL-2025)

CERTIFICATE

Certified that **Tanishka Gupta 202410116100218, Shipra Upadhyay 202410116100196**, **Vanshika Garg 202410116100235** and **Tripti Rajput 202410116100225** have carried out the project work having "**CUSTOMER SEGMENTATION USING UNSUPERVISED LEARNING**" (**INTRODUCTION TO AI (AI101B))** for **Master of Computer Application** from Dr. A.P.J. Abdul Kalam Technical University (AKTU**)** (formerly UPTU), Lucknow under my supervision. The project report embodies original work, and studies are carried out by the student himself/herself and the contents of the project report do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Ms. Komal Salgotra
Assistant Professor
Department of Computer Applications
KIET Group of Institutions, Ghaziabad

Dr. Akash Rajak
Dean
Department of Computer Applications
KIET Group of Institutions, Ghaziabad

CUSTOMER SEGMENTATION USING UNSUPERVISED LEARNING

Tanishka Gupta, Shipra Upadhyay, Vanshika Garg, Tripti Rajput

ABSTRACT

Customer behaviour is key to success in today's competitive business environment. The project emphasizes **customer segmentation via unsupervised learning**, and more specifically, the **K-Means clustering algorithm**, to categorize customers into clusters based on common characteristics like **age, gender, annual income, and spending score**. The goal is to establish patterns of customer data to assist businesses in making more effective marketing decisions, personalize services, and enhance customer experience.

The data set employed contains demographic and behavioural information of customers. Following data preprocessing and normalization, clustering is conducted to identify natural groupings of customers. Clusters are subsequently visualized via 2D and 3D plots in order to comprehend the nature of each segment more effectively.

The ultimate deliverable of the project is a segmented customer view in which every group possesses distinct characteristics. These segments can be used by marketing teams to design special promotions, formulate loyalty programs, or even introduce new products to particular groups.

Overall, this project illustrates how **machine learning** techniques can be used in actual business contexts to derive meaning from raw data. It illustrates the strength of unsupervised learning in discovering hidden patterns that can result in more intelligent business strategies and better customer engagement.

The project also investigates how various segments of customers conduct themselves in terms of spending and income, which will facilitate prioritization of customer outreach.

ACKNOWLEDGEMENTS

Success in life is never attained single-handedly. My deepest gratitude goes to my project supervisor, **Ms. Komal Salgotra** for her guidance, help, and encouragement throughout my project work. Their enlightening ideas, comments, and suggestions.

Words are not enough to express my gratitude to Dr. Akash Rajak, Professor and Dean, Department of Computer Applications, for his insightful comments and administrative help on various occasions.

Fortunately, I have many understanding friends, who have helped me a lot on many critical conditions.

Finally, my sincere thanks go to my family members and all those who have directly and indirectly provided me with moral support and other kind of help. Without their support, completion of this work would not have been possible in time. They keep my life filled with enjoyment and happiness.

Tanishka Gupta

Shipra Upadhyay

Vanshika Garg

Tripti Rajput

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Overview

In the last few years, companies have increasingly been relying on data-driven approaches to knowledge about customers' needs and behaviour. Customer segmentation — dividing customers into groups with similar characteristics — is one of the most powerful techniques in this direction. Through customer segmentation, businesses can provide customized services, conduct focused marketing campaigns, and increase customer satisfaction.

This project utilizes **unsupervised machine learning**, more specifically the **K-Means clustering algorithm**, to do customer segmentation. Unlike supervised learning, unsupervised learning doesn't need labelled outputs. It identifies patterns and structures in the data automatically and segments customers into useful clusters.

The project utilizes the "Mall Customers" dataset, which includes simple customer information like **age, gender, annual income, and spending score**. These are chosen features because they have a direct effect on purchasing and enable valid grouping.

Applying data preprocessing, visualization, and clustering, the project is able to identify different customer types like high-income high-spending customers, low-income high-spending customers, etc. These are represented through 2D and 3D plots for better visualization.

The long-term purpose is to show how artificial intelligence can assist companies in knowing their customers better and making better, more informed decisions.

1.2 Objectives

The primary goals of the project **"Customer Segmentation Using Unsupervised Learning"** are listed below:

1. **Analyse customer data** and derive meaningful features such as age, gender, income per year, and spending score.
2. **Implement the K-Means Clustering algorithm**, a widely used unsupervised learning algorithm, for grouping customers based on similarities.
3. **Reveal secret patterns** in customer activity that cannot be discerned through manual analysis.
4. **For visualizing customer segments** through 2D and 3D graphs for easier interpretation and comprehension of the resulting clusters.
5. **To help businesses identify** various categories of customers (e.g., high-spenders, low-income people, teenage consumers) for more effective marketing.
6. **To facilitate personalized marketing strategies** like loyalty schemes, discounts, or special products offered to certain groups.
7. **To offer an AI-driven solution** that is reusable, scalable, and capable of being adapted to various datasets or business domains.
8. **To show the strength of unsupervised** learning in practical applications such as market segmentation, customer analytics, and business intelligence.
9. **To facilitate data-driven decision making** by communicating customer insights effectively and simply to stakeholders.
10. **To provide a foundation for future enhancements**, e.g., adding recommendation systems or real-time customer analytics.

1.3 Significance

1. **Enriches Marketing Strategies**
    - Enables companies to identify various customer categories and orient marketing campaigns suitably.
2. **Facilitates Personalized Services**
    - Enables firms to provide customized promotions, discounts, and recommendations to targeted groups of customers.
3. **Enhances Customer Satisfaction**
    - By segmenting customers better, firms are able to enhance service quality and customer loyalty.
4. **Facilitates Data-Driven Decision Making**
    - Eliminates guessing and employs real customer data to inform business strategy and product development.
5. **Uncovering Business Opportunities**
    - Segmentation identifies uncovered or high-potential groups, allowing companies to access new markets.
6. **Eliminates Squandering of Resources**
    - Businesses save costs on useless promotions by targeting relevant customer groups alone.
7. **Creates Competitive Edge**
    - Businesses with good knowledge about their customers have a better opportunity to compete with others in the market.
8. **Ensures Customer Connection**
    - One-to-one engagement fosters trust and durability.
9. **Encourages Real-Life Application of AI**
    - Illustrates the strength of unsupervised learning in real-world business applications such as customer analytics.

1.4 Project Scope

1. **Customer Segmentation Focus**
   - The project will classify customers into separate segments according to their age, income, gender, and expenditure score.
2. **Utilization of Unsupervised Learning**
   - Uses the K-Means clustering algorithm to find patterns in the data without requiring labelled data.
3. **Utilizes Real-World Dataset**
   - Utilizes a publicly available mall customer dataset, which makes the project relevant to retail and e-commerce companies.
4. **2D and 3D Data Visualization**
   - Incorporates graphical cluster representation to facilitate visualization of group properties and interactions.
5. **Supports Business Intelligence**
   - Enables marketing and sales teams to gain insights into customer behavior for strategic decision-making.
6. **Scalable for Larger Datasets**
   - The method can be used with larger, more complicated datasets across various industries.
7. **Reusable Model Framework**
   - The design of the project makes it easily adaptable to other segmentation requirements or customer characteristics.
8. **Educational Value**
   - Offers a real-world illustration of how machine learning and AI can be implemented in the business world.
9. **Preparation for Future Extensions**
   - Can be further extended to support predictive analytics, recommendation engines, or customer lifetime value calculation.

2. METHODOLOGY

The methodology of the project is to make use of **unsupervised learning methods, especially K-Means clustering**, to classify customers based on their spend and behaviour. The process involves a number of well-defined steps:

1. **Data Collection and Understanding**

   The dataset for this project is the "Mall Customers" dataset. It has important customer data such as Customer ID, Gender, Age, Annual Income (in $), and Spending Score (1–100). The objective is to segment customers by income and spending behaviour.

2. **Data Preprocessing**

   The dataset is cleaned first to eliminate any missing or irrelevant data. Numerical features alone are chosen for clustering. The features chosen for primary clustering are **Annual Income** and **Spending Score**.

3. **Exploratory Data Analysis (EDA)**

   EDA is conducted using libraries such as Matplotlib and Seaborn. Scatter plots and distribution plots are made to observe the relationship between the features and identify any visible patterns of clustering.

4. **K-Means Clustering Algorithm**

   The Elbow Method is used to find the best number of clusters (K). K-Means is then used on the chosen features. The algorithm gives a cluster label to every customer.

5. **Cluster Visualization and Analysis**

   2D and 3D scatter plots are used to visualize the customer segments. Clusters are examined based on their traits such as high-income/low-spending or low-income/high-spending.

This approach creates a clear segmentation method for customer segmentation and behavior analysis through machine learning.

3. CODE AND OUTPUT

#Importing the necessary libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from mpl_toolkits.mplot3d import Axes3D
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")
data=pd.read_csv("/content/Mall_Customers.csv")
data.head()
```

	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

Fig. (1)

Select only numerical features for correlation analysis

```
numerical_features = data.select_dtypes(include=np.number)
# Calculate the correlation matrix
correlation_matrix = numerical_features.corr()
```

Display the correlation matrix

print(correlation_matrix)

```
                        CustomerID       Age  Annual Income (k$)  \
CustomerID                1.000000 -0.026763            0.977548
Age                      -0.026763  1.000000           -0.012398
Annual Income (k$)        0.977548 -0.012398            1.000000
Spending Score (1-100)    0.013835 -0.327227            0.009903

                        Spending Score (1-100)
CustomerID                            0.013835
Age                                  -0.327227
Annual Income (k$)                    0.009903
Spending Score (1-100)                1.000000
```

Fig. (2)

#Distribution of Annual Income

```
plt.figure(figsize=(10, 6))
sns.set(style = 'whitegrid')
sns.distplot(data['Annual Income (k$)'])
plt.title('Distribution of Annual Income (k$)', fontsize = 20)
plt.xlabel('Range of Annual Income (k$)')
plt.ylabel('Count')
```

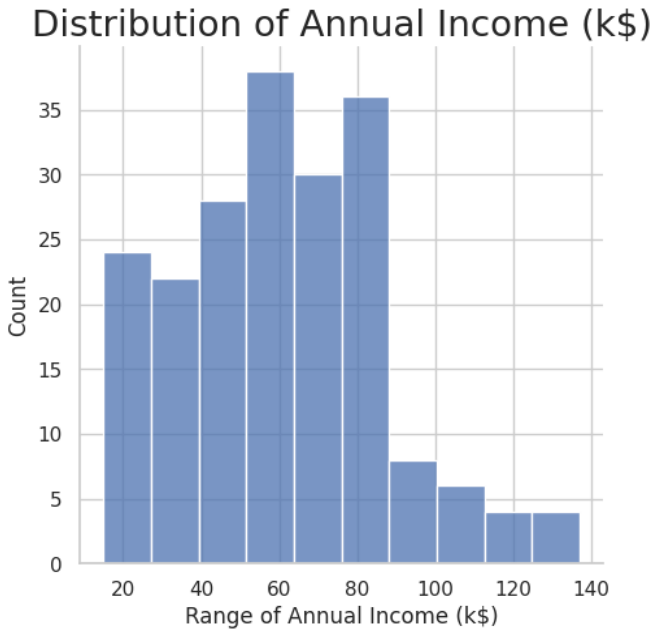

Fig. (3)

Figure (3) displays the distribution of annual income, where the x-axis represents income in k$ and the y-axis shows customer count. Most customers earn between 40k$ and 80k$.

#Distribution of age

```
plt.figure(figsize=(10, 6))
sns.set(style = 'whitegrid')
sns.distplot(data['Age'])
plt.title('Distribution of Age', fontsize = 20)
plt.xlabel('Range of Age')
plt.ylabel('Count')
```

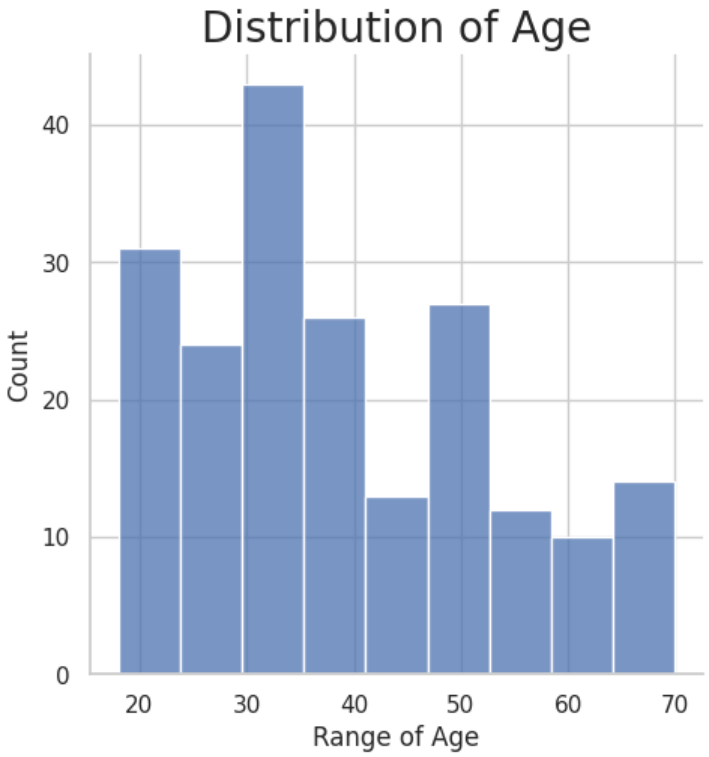

Fig. (4)

Figure (4) shows customer age distribution. Most customers are aged 30–40, with a peak around 35. The x-axis represents age range, and the y-axis shows customer count.

#Distribution of spending score

```
plt.figure(figsize=(10, 6))
sns.set(style = 'whitegrid')
sns.displot(data['Spending Score (1-100)'])
plt.title('Distribution of Spending Score (1-100)', fontsize = 20)
```

```
plt.xlabel('Range of Spending Score (1-100)')
plt.ylabel('Count')
```

```
Text(17.31944444444445, 0.5, 'Count')
<Figure size 1000x600 with 0 Axes>
```

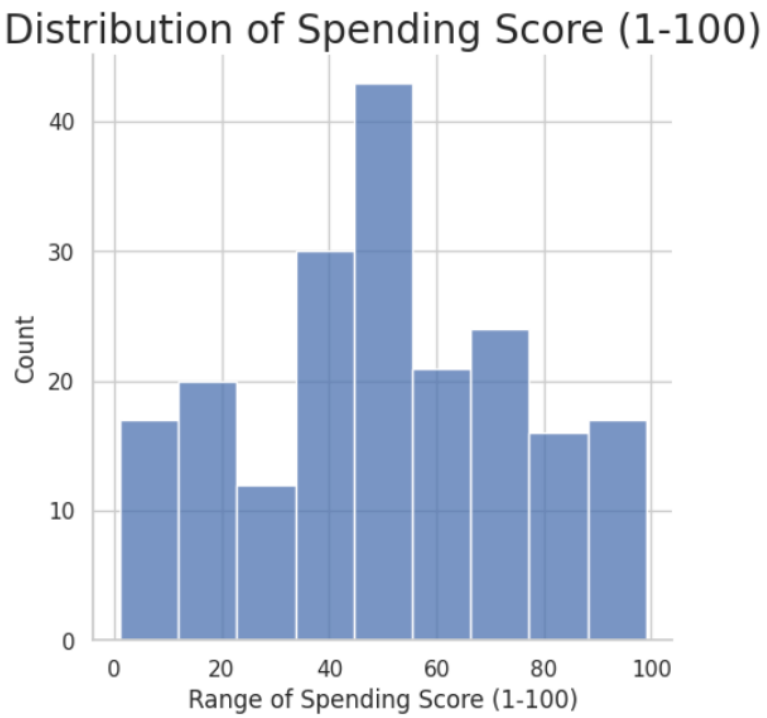

Fig. (5)

Displays spending score distribution; identifies how customers are spread across low to high spending behaviour, influencing segmentation.

```
genders = data.Gender.value_counts()
sns.set_style("darkgrid")
plt.figure(figsize=(10,4))
sns.barplot(x=genders.index, y=genders.values, palette=['violet', 'blue'])  # Specify colors
plt.show()
```

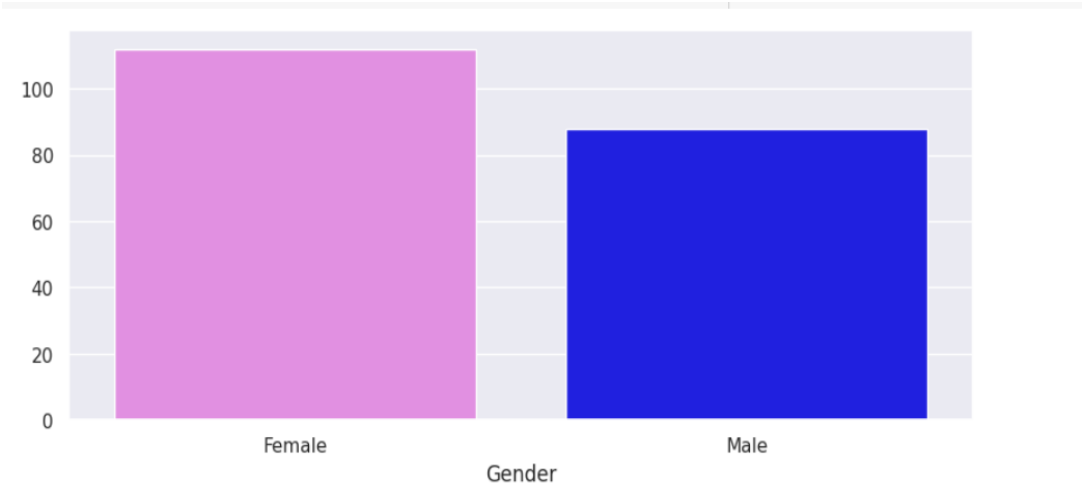

Fig. (6)

df1=data[["CustomerID","Gender","Age","Annual Income (k$)","Spending Score (1-100)"]]

X=df1[["Annual Income (k$)","Spending Score (1-100)"]]

X.head()

	Annual Income (k$)	Spending Score (1-100)
0	15	39
1	15	81
2	16	6
3	16	77
4	17	40

Fig. (7)

Figure shows a data table selecting annual income and spending score for clustering, simplifying input data

#Scatterplot of the input data

```
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()
```

Fig. (8)

Figure is a scatter plot of income versus spending, visually suggesting distinct grouping

#Importing KMeans from sklearn

```
from sklearn.cluster import KMeans
wcss=[]
for i in range(1,11):
    km=KMeans(n_clusters=i)
    km.fit(X)
    wcss.append(km.inertia_)
```

#The elbow curve

```
plt.figure(figsize=(12,6))
plt.plot(range(1,11),wcss)
plt.plot(range(1,11),wcss, linewidth=2, color="red", marker ="8")
plt.xlabel("K Value")
plt.xticks(np.arange(1,11,1))
plt.ylabel("WCSS")
plt.show()
```

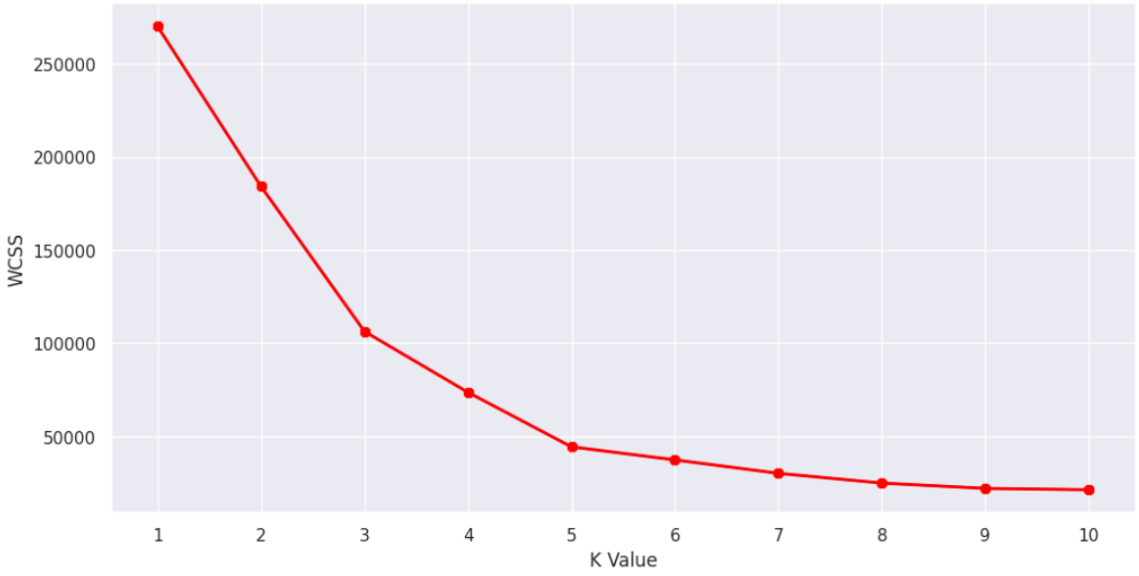

Fig. (9)

Figure uses the Elbow method to plot the number of clusters (K) against WCSS, indicating 5 as the optimal cluster count.

#Taking 5 clusters

```
km1=KMeans(n_clusters=5)
#Fitting the input data
km1.fit(X)
#predicting the labels of the input data
y=km1.predict(X)
```

#adding the labels to a column named label

df1["label"] = y

#The new dataframe with the clustering done

df1.head()

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

Fig. (10)

#Scatterplot of the clusters

```
plt.figure(figsize=(10,6))
sns.scatterplot(x = 'Annual Income (k$)',y = 'Spending Score (1-100)',hue="label",
                palette=['green','orange','brown','dodgerblue','red'], legend='full',data = df1  ,s = 60
)
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.title('Spending Score (1-100) vs Annual Income (k$)')
plt.show()
```

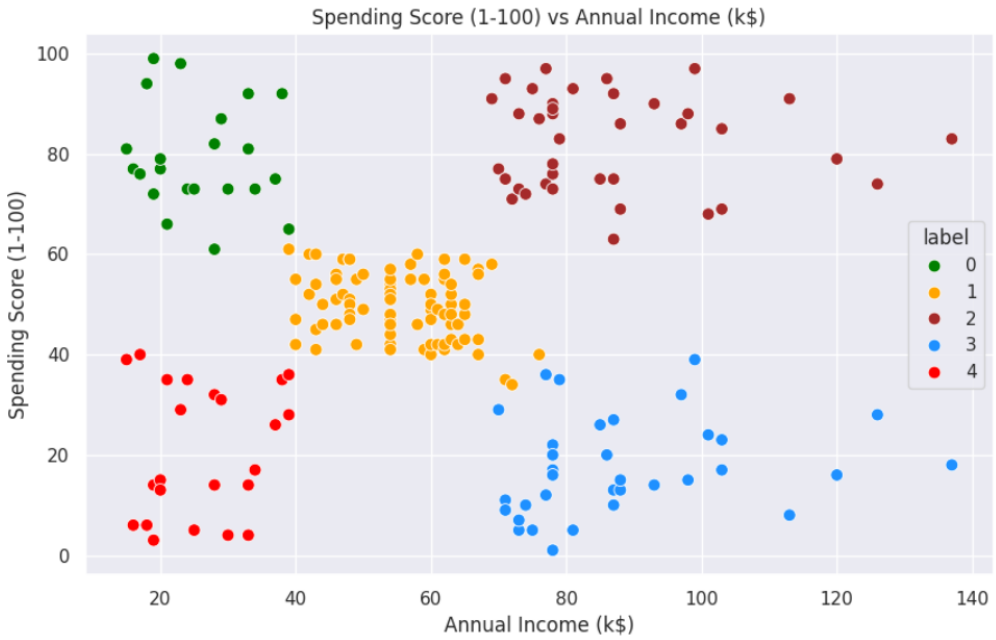

Fig. (11)

#Taking the features

```
X2=df1[["Age","Annual Income (k$)","Spending Score (1-100)"]]
#Now we calculate the Within Cluster Sum of Squared Errors (WSS) for different values of k.
wcss = []
for k in range(1,11):
    kmeans = KMeans(n_clusters=k, init="k-means++")
    kmeans.fit(X2)
    wcss.append(kmeans.inertia_)
plt.figure(figsize=(12,6))
plt.plot(range(1,11),wcss, linewidth=2, color="red", marker ="8")
plt.xlabel("K Value")
plt.xticks(np.arange(1,11,1))
plt.ylabel("WCSS")
plt.show()
```

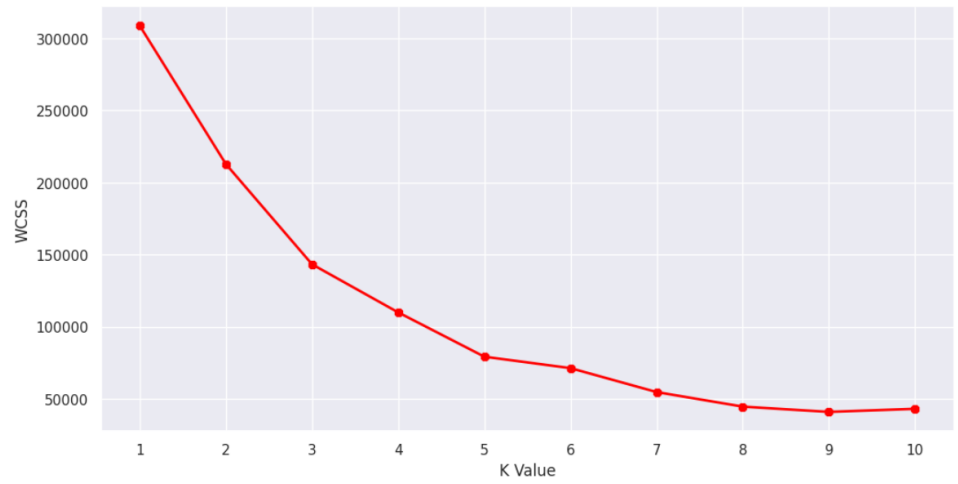

Fig. (12)

#We choose the k for which WSS starts to diminish

km2 = KMeans(n_clusters=5)

y2 = km.fit_predict(X2)

df1["label"] = y2

#The data with labels

df1.head()

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)	label
0	1	Male	19	15	39	4
1	2	Male	21	15	81	8
2	3	Female	20	16	6	4
3	4	Female	23	16	77	8
4	5	Female	31	17	40	4

Fig. (13)

#3D Plot as we did the clustering on the basis of 3 input features

```
fig = plt.figure(figsize=(20,10))
ax = fig.add_subplot(111, projection='3d')
ax.scatter(df1.Age[df1.label == 0], df1["Annual Income (k$)"][df1.label == 0],
df1["Spending Score (1-100)"][df1.label == 0], c='purple', s=60)
ax.scatter(df1.Age[df1.label == 1], df1["Annual Income (k$)"][df1.label == 1],
df1["Spending Score (1-100)"][df1.label == 1], c='red', s=60)
ax.scatter(df1.Age[df1.label == 2], df1["Annual Income (k$)"][df1.label == 2],
df1["Spending Score (1-100)"][df1.label == 2], c='blue', s=60)
ax.scatter(df1.Age[df1.label == 3], df1["Annual Income (k$)"][df1.label == 3],
df1["Spending Score (1-100)"][df1.label == 3], c='green', s=60)
ax.scatter(df1.Age[df1.label == 4], df1["Annual Income (k$)"][df1.label == 4],
df1["Spending Score (1-100)"][df1.label == 4], c='yellow', s=60)
ax.view_init(35, 185)
plt.xlabel("Age")
plt.ylabel("Annual Income (k$)")
ax.set_zlabel('Spending Score (1-100)')
plt.show()
```

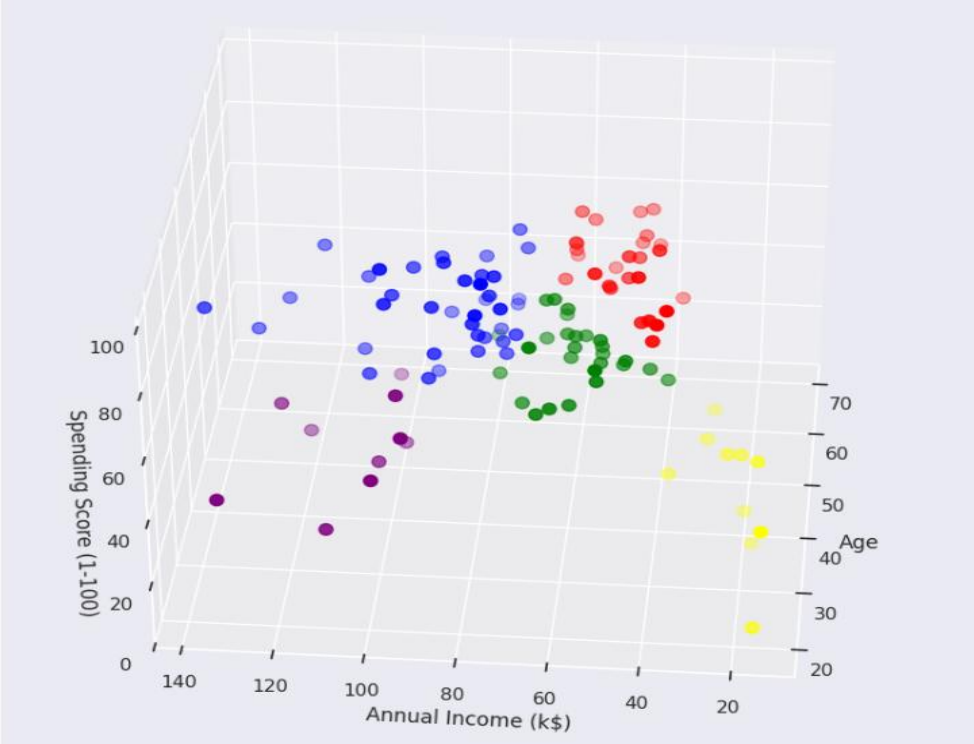

Fig. (14)

Figure 14 offers a 3D scatter plot with age, income, and spending on different axes, revealing clear groupings across all three features.

```
cust1=df1[df1["label"]==1]
print('Number of customer in 1st group=', len(cust1))
print('They are -', cust1["CustomerID"].values)
print("-----------------------------------------")
cust2=df1[df1["label"]==2]
print('Number of customer in 2nd group=', len(cust2))
print('They are -', cust2["CustomerID"].values)
print("-----------------------------------------")
cust3=df1[df1["label"]==0]
print('Number of customer in 3rd group=', len(cust3))
print('They are -', cust3["CustomerID"].values)
print("-----------------------------------------")
cust4=df1[df1["label"]==3]
print('Number of customer in 4th group=', len(cust4))
```

```
print('They are -', cust4["CustomerID"].values)
print("----------------------------------------------")
cust5=df1[df1["label"]==4]
print('Number of customer in 5th group=', len(cust5))
print('They are -', cust5["CustomerID"].values)
print("----------------------------------------------")
```

```
Number of customer in 1st group= 27
They are - [ 41  47  51  54  55  57  58  60  61  63  64  65  68  71  73  74  75  81
  83  87  91 103 107 109 110 111 117]
-------------------------------------------
Number of customer in 2nd group= 39
They are - [124 126 128 130 132 134 136 138 140 142 144 146 148 150 152 154 156 158
 160 162 164 166 168 170 172 174 176 178 180 182 184 186 188 190 192 194
 196 198 200]
-------------------------------------------
Number of customer in 3rd group= 10
They are - [181 183 185 187 189 191 193 195 197 199]
-------------------------------------------
Number of customer in 4th group= 27
They are - [ 43  56  67  72  77  78  80  82  84  86  90  93  94  97  99 102 105 108
 113 118 119 120 122 123 127 147 161]
-------------------------------------------
Number of customer in 5th group= 10
They are - [ 1  3  5  7 15 17 21 27 29 39]
-------------------------------------------
```

Fig. (15)

REFRENCES

- Scikit-learn Documentation – KMeans Clustering https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- Sudheer Nelakurthi, “Mall Customers Dataset,” GitHub: https://raw.githubusercontent.com/NelakurthiSudheer/Mall-Customers-Segmentation/refs/heads/main/Dataset/Mall_Customers.csv
- O. Maimon and L. Rokach, “Clustering methods”, in Data Mining and Knowledge Discovery Handbook. Boston: Springer US, 2005, pp. 321-352.
- Y. Chen, et al., “Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospitalbased assessment”, Computers in Biology and Medicine, vol. 42, no. 2, pp. 213-221, 2012.
- OpenAI, “GPT-4 Technical Report,” 2023. https://openai.com/research/gpt-4
- G. Lefait and T. Kechadi, “Customer segmentation architecture based on clustering techniques”, in Fourth International Conference on Digital Society, Sint Maarten, 2010, pp. 243-248.
- Visualization Libraries (Matplotlib & Seaborn)
  - https://matplotlib.org
  - https://seaborn.pydata.org
- M. Namvar, M. Gholamian and S. KhakAbi, “A two-phase clustering method for intelligent customer segmentation”, in International Conference on Intelligent Systems, Modelling and Simulation, Liverpool, 2010, pp. 215-219.