Assessment Report

on

"E-Commerce Segmentation"

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

Artificial Intelligence

By

Shaurya Pratap Singh(202401100300229)

Under the supervision of

"Mr. Abhishek Shukla Sir"

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May, 2025

Problem Statement

E-commerce companies collect extensive data on customer transactions and behaviour. Understanding this data allows businesses to improve personalization, marketing strategies, and customer satisfaction.
This project aims to:

- Identify distinct customer groups (segmentation) based on behaviour and purchase patterns.
- Classify customers into "High" and "Low" spenders for targeted campaigns.

By combining **clustering (unsupervised learning)** and **classification (supervised learning)**, we can build a more complete understanding of customer profiles.

METHODOLOGY

The goal of this project is to segment e-commerce customers into distinct groups based on their purchasing habits and browsing behaviour. The methodology involved the following steps:

1. Data Loading and Preprocessing

The dataset was imported and initially inspected for any missing values. All rows with missing entries were dropped to ensure clean input for the clustering algorithm. Non-numeric columns, which are not directly usable by K Means, were also excluded.

2. Feature Scaling

Clustering algorithms like K Means are sensitive to the scale of input features. To normalize the data, **Standard Scaler** from scikit-learn was used to standardize features by removing the mean and scaling to unit variance.

3. Optimal Cluster Selection

To determine the most appropriate number of clusters (**k**), the **Elbow Method** was employed. The Within-Cluster Sum of Squares (WCSS) was plotted for a range of cluster values. The optimal number of clusters was identified based on the “elbow point” where the WCSS started to diminish at a slower rate.

4. Clustering with K Means

The **K Means algorithm** was applied with the selected number of clusters (k=4). Each customer in the dataset was assigned to a cluster based on similarity in behaviour and purchasing patterns.

5. Dimensionality Reduction with PCA

To visualize the high-dimensional customer data in 2D, **Principal Component Analysis (PCA)** was used to project the data into two principal components. This enabled effective visualization of customer clusters on a 2D plot.

6. Cluster Evaluation

To assess the quality of clustering, This metric measures how similar a data point is to its own cluster compared to other clusters, with values closer to 1 indicating better-defined clusters.

7. Visualization

Multiple visualizations were generated to better understand the clustering results:

- A **PCA scatter plot** coloured by cluster labels.

CODE

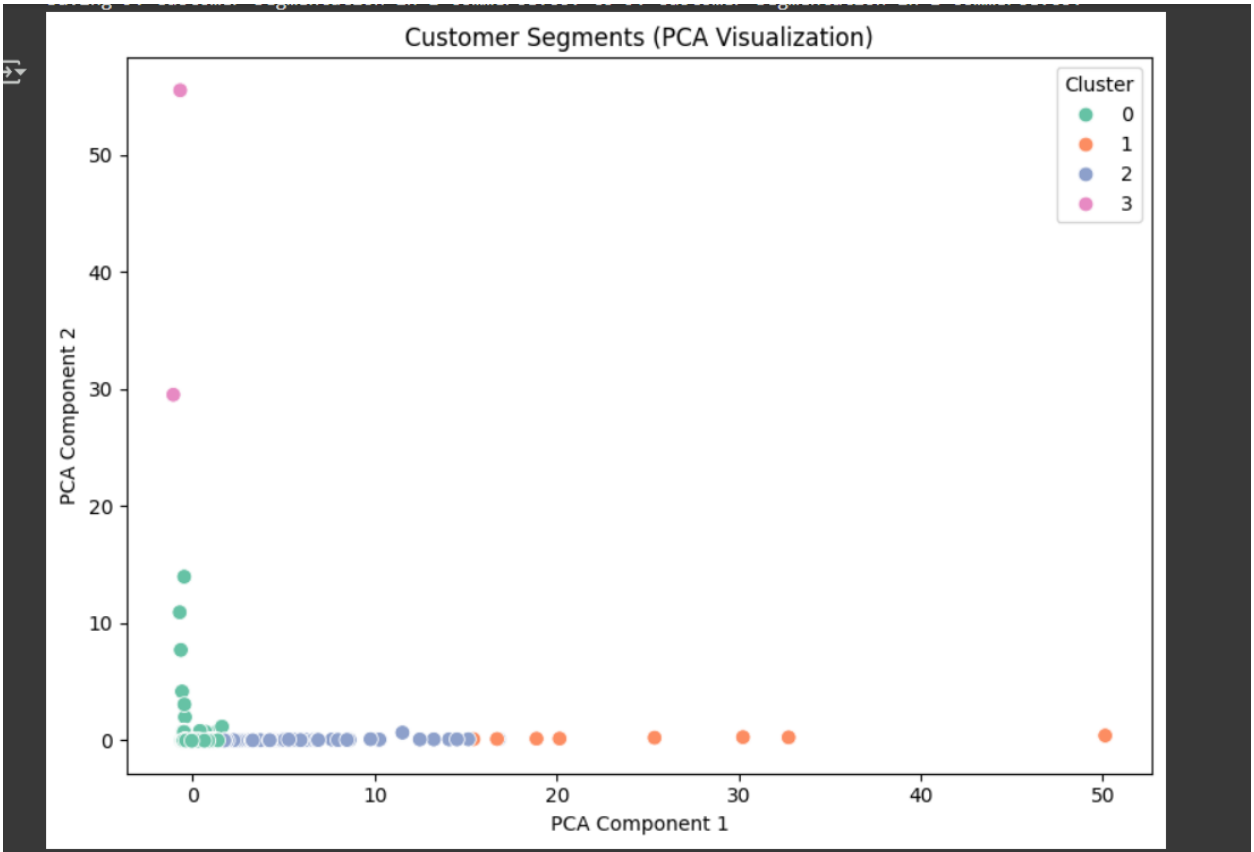
```
# STEP 1: Upload the dataset

from google.colab import files

uploaded = files.upload()


# STEP 2: Import required libraries

import pandas as pd

import numpy as np

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

from sklearn.decomposition import PCA

import matplotlib.pyplot as plt

import seaborn as sns


# STEP 3: Load and clean the data

# Replace with your actual file name after upload

filename = list(uploaded.keys())[0]

df = pd.read_csv(filename)

df = df.dropna(subset=["CustomerID"])


# Add total price per item

df["TotalPrice"] = df["Quantity"] * df["UnitPrice"]


# STEP 4: Create customer-level features

customer_df = df.groupby("CustomerID").agg({
```

```
    "InvoiceNo": "nunique",
    "Quantity": "sum",
    "UnitPrice": "mean",
    "TotalPrice": "sum"
}).reset_index()

customer_df.columns = ["CustomerID", "NumPurchases", "TotalQuantity", "AvgUnitPrice", "TotalSpent"]

# STEP 5: Scale features
features = customer_df[["NumPurchases", "TotalQuantity", "AvgUnitPrice", "TotalSpent"]]
scaler = StandardScaler()
scaled_features = scaler.fit_transform(features)

# STEP 6: Apply K-Means clustering
kmeans = KMeans(n_clusters=4, random_state=42)
customer_df["Cluster"] = kmeans.fit_predict(scaled_features)

# STEP 7: Reduce dimensions with PCA for visualization
pca = PCA(n_components=2)
pca_components = pca.fit_transform(scaled_features)
customer_df["PCA1"] = pca_components[:, 0]
customer_df["PCA2"] = pca_components[:, 1]

# STEP 8: Plot the clusters
plt.figure(figsize=(8, 6))
```

```
sns.scatterplot(data=customer_df, x="PCA1", y="PCA2", hue="Cluster", palette="Set2", s=60)

plt.title("Customer Segments (PCA Visualization)")

plt.xlabel("PCA Component 1")

plt.ylabel("PCA Component 2")

plt.legend(title="Cluster")

plt.tight_layout()

plt.show()


# STEP 9: Print cluster behavior summary

cluster_summary = customer_df.groupby("Cluster")[["NumPurchases", "TotalQuantity", "AvgUnitPrice",
"TotalSpent"]].mean()

print(cluster_summary)
```

OUTPUT / RESULT

The unsupervised clustering analysis successfully segmented the e-commerce customer base into distinct groups using K-Means clustering. The optimal number of clusters was determined to be 4, based on the Elbow Method, indicating a good level of cluster separation.



Cluster	NumPurchases	TotalQuantity	AvgUnitPrice	TotalSpent
0	4.030360	722.512827	5.187292	1196.116504
1	80.000000	83266.500000	5.182650	155483.796250
2	39.106195	10360.530973	4.042448	17500.305752
3	3.000000	29.500000	6171.705000	-1819.065000

REFERENCES / CREDITS

Tools and Libraries Used

- **Pandas**: Data manipulation and analysis
- **NumPy**: Numerical operations and array handling
- **Matplotlib & Seaborn**: Data visualization and plotting
- **Scikit-learn**:
    - Standard Scaler for feature scaling
    - K Means for clustering
    - PCA for dimensionality reduction

🔹 Academic and Technical References

1. **Jain, A. K. (2010)**. *Data clustering: 50 years beyond K-means*. Pattern Recognition Letters, 31(8), 651–666.
2. Scikit-learn documentation: https://scikit-learn.org/stable/
3. Towards Data Science articles and tutorials on customer segmentation and clustering.

🔹 Dataset Source

- The dataset titled **"9. Customer Segmentation in E-commerce.csv"** was provided for academic and analytical purposes. It includes anonymized customer behaviour attributes such as browsing activity, purchase frequency, and spending patterns.