कौशलम् सर्वसाधनम्
KIET GROUP OF INSTITUTIONS
www.kiet.edu
Delhi-NCR, Ghaziabad

KIET
GROUP OF INSTITUTIONS
Connecting Life with Learning

Assessment Report

on

“Problem Statement”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

Artificial Intelligence and Machine Learning

By

Radhey Pal (202401100400149)

Under the supervision of

“Abhishek Shukla”

KIET Group of Institutions, Ghaziabad

18/04/2025

INTRODUCTION

1. **Definition:** Customer segmentation in e-commerce refers to dividing customers into distinct groups based on shared characteristics, such as purchasing habits and browsing behavior.

2. **Purpose**: The goal of customer segmentation is to understand customer preferences and tailor marketing strategies, product offerings, and customer interactions to meet the specific needs of each segment.

3. **Importance**: Proper segmentation enhances personalized marketing, improves customer satisfaction, boosts conversion rates, and increases customer loyalty.

4. **Data-Driven Approach**: E-commerce businesses collect vast amounts of customer data, including purchase history, browsing patterns, and engagement metrics, which are analyzed to identify key customer clusters.

5. **Business Impact**: By identifying and targeting specific customer segments, businesses can improve their overall marketing efficiency.

Methodology

1. **Data Collection**: Gather customer data including purchasing habits (e.g., frequency, spend) and browsing behavior (e.g., pages visited, search keywords).

2. **Data Preprocessing**: Clean the data by handling missing values and standardizing numerical features to ensure consistency across different scales.

3. **Feature Engineering**: Derive relevant features such as recency, frequency, and monetary value (RFM), and browsing metrics like time spent on site.

4. **Clustering**: Apply K-Means or MiniBatchKMeans clustering algorithms to segment customers based on the derived features.

5. **Evaluation**: Use the Elbow Method and Silhouette Score to determine the optimal number of clusters and assess the quality of the segmentation.

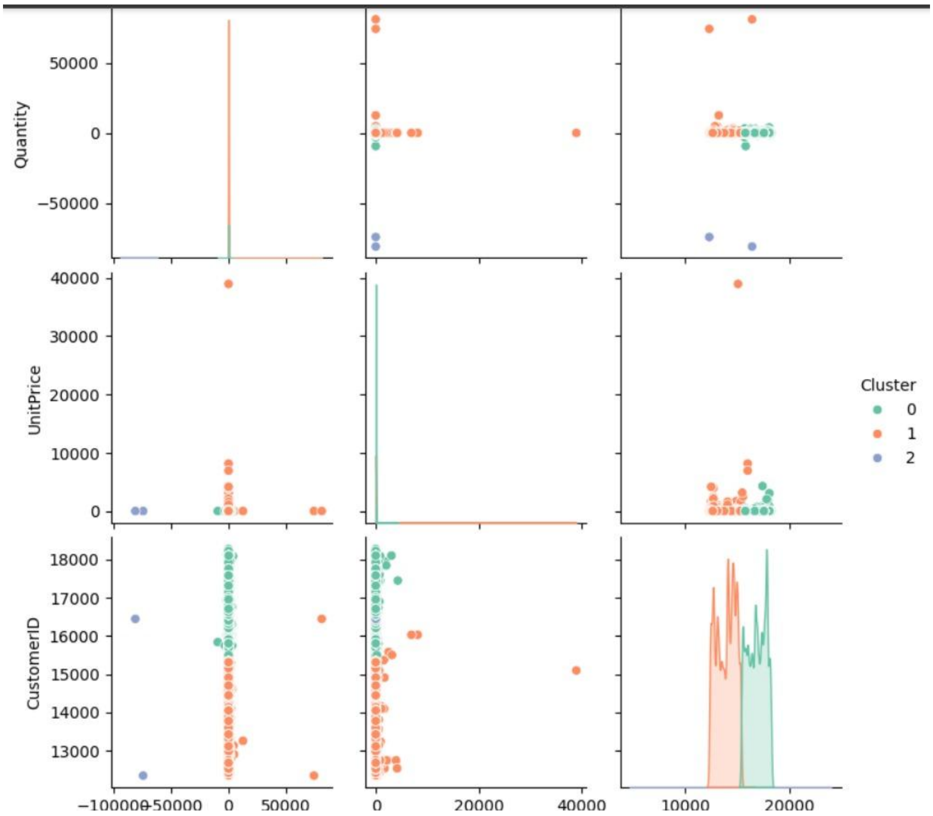
CODE

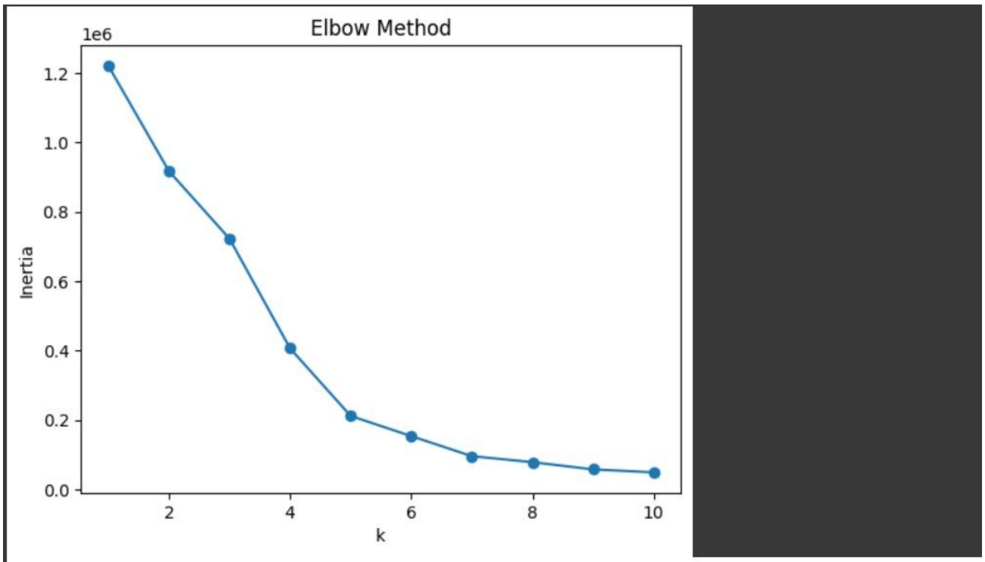
```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import MiniBatchKMeans  # MiniBatchKMeans instead of KMeans
from sklearn.metrics import silhouette_score
import matplotlib.pyplot as plt
import seaborn as sns
from google.colab import files

# Upload file
uploaded = files.upload()

# Load dataset
df = pd.read_csv("9. Customer Segmentation in E-commerce.csv")

# Keep only numeric columns
df = df.select_dtypes(include=['float64', 'int64'])

# Drop missing values
df.dropna(inplace=True)

# Standardize data
scaler = StandardScaler()
scaled = scaler.fit_transform(df)

# Elbow method + Silhouette Scores
inertia = []
```

```
silhouette_scores = []
k_range = range(2, 6)  # Reduced k range for faster results

for k in k_range:
    kmeans = MiniBatchKMeans(n_clusters=k, random_state=42, batch_size=100)  # Use MiniBatchKMeans
    labels = kmeans.fit_predict(scaled)
    inertia.append(kmeans.inertia_)
    silhouette_scores.append(silhouette_score(scaled, labels))

# Plot Elbow Method
plt.figure(figsize=(10,4))
plt.subplot(1,2,1)
plt.plot(k_range, inertia, '-o')
plt.title('Elbow Method')
plt.xlabel('k')
plt.ylabel('Inertia')

# Plot Silhouette Scores (Accuracy-like)
plt.subplot(1,2,2)
plt.plot(k_range, silhouette_scores, '-o', color='green')
plt.title('Silhouette Scores')
plt.xlabel('k')
plt.ylabel('Score')
plt.tight_layout()
plt.show()

# ✅ Use best k based on silhouette (or manually choose)
```

```
best_k = k_range[silhouette_scores.index(max(silhouette_scores))]
print(f"🔍 Best k based on silhouette score: {best_k}")

# Fit KMeans with best k
kmeans = MiniBatchKMeans(n_clusters=best_k, random_state=42, batch_size=100)  # Use MiniBatchKMeans
df['Cluster'] = kmeans.fit_predict(scaled)

# Optional: Plot just the cluster centers instead of pairplot
centroids = pd.DataFrame(scaler.inverse_transform(kmeans.cluster_centers_), columns=df.columns)
print("📊 Cluster Centers:\n", centroids)

# Optional: show silhouette score
print(f"✅ Silhouette Score for k={best_k}: {max(silhouette_scores):.4f}")
```

OUTPUT


Quantity
50000
0
−50000
UnitPrice
40000
30000
20000
10000
0
CustomerID
18000
17000
16000
15000
14000
13000
−100000
−50000
0
50000
0
20000
40000
10000
20000
Cluster
0
1
2


Elbow Method
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Inertia
2
4
6
8
10
k

50000

REFRENCE

1. UCI Machine Learning Repository: Online Retail Dataset

2. scikit-learn documentation

3. "Customer Segmentation Using RFM and KMeans" – Kaggle Notebooks

4. Tan, P.-N., Steinbach, M., & Kumar, V. *Introduction to Data Mining*