KIET GROUP OF INSTITUTIONS

Connecting Life with Learning

www.kiet.edu

Delhi-NCR, Ghaziabad

A

Assesment Report

on

"Customer Segmentation in E-Commerce."

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

Name of discipline

By

Name- UDIT KASTWAR (Roll Numbar-2024011100400199)

Under the supervision of

"Abhishek Shukla Sir"

KIET Group of Institutions, Ghaziabad

Affiliated to

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

(Formerly UPTU)

May, 2025

b. Introduction:

Customer segmentation is a crucial aspect of e-commerce business strategy. It involves dividing customers into distinct groups based on their buying behaviors, preferences, and financial worth. This helps businesses personalize their marketing, improve customer service, and ultimately boost profits. In this project, we used the Online Retail dataset from UCI ML Repository to perform customer segmentation using unsupervised machine learning techniques.

c. Methodology:

1. Loaded the 'Online Retail' dataset into Google Colab.
2. Cleaned the dataset by removing null values and invalid quantities/prices.
3. Created Recency, Frequency, and Monetary (RFM) features for each customer.
4. Standardized the data using StandardScaler.
5. Applied KMeans clustering and determined the optimal number of clusters using silhouette score.
6. Visualized the clusters using an interactive 3D plot.
7. Summarized each cluster's RFM characteristics to interpret the segments.

d. Code:

Due to length, the full code is provided in the accompanying notebook/script file.

It includes:

- Data upload and preprocessing
- Feature engineering (RFM)
- Clustering using KMeans
- Evaluation and visualization

You can copy and run the entire code block provided in the Colab notebook.

d. Code:

Below is the complete Python code used for the customer segmentation project.
This code includes all steps from data upload, preprocessing, RFM feature creation,
clustering using KMeans, silhouette score evaluation, 3D visualization, and summary generation.
You can run this code in Google Colab using the Online Retail dataset.

```
# ■ Required 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
import plotly.express as px
from google.colab import files

# ■ Upload the dataset file
uploaded = files.upload()  # Select 'Online Retail.xlsx' after running this

# ■ Load the Excel File
df = pd.read_excel('Online Retail.xlsx')

# ■ Data Cleaning
df.dropna(inplace=True)
df = df[(df['Quantity'] > 0) & (df['UnitPrice'] > 0)]
df['TotalPrice'] = df['Quantity'] * df['UnitPrice']

# ■ RFM Feature Creation
snapshot_date = df['InvoiceDate'].max() + pd.Timedelta(days=1)
rfm = df.groupby('CustomerID').agg({
    'InvoiceDate': lambda x: (snapshot_date - x.max()).days,
    'InvoiceNo': 'nunique',
    'TotalPrice': 'sum'
}).reset_index()

rfm.columns = ['CustomerID', 'Recency', 'Frequency', 'Monetary']

# ■ Scaling the Features
scaler = StandardScaler()
rfm_scaled = scaler.fit_transform(rfm[['Recency', 'Frequency', 'Monetary']])

# ■ Finding Best Number of Clusters Using Silhouette Score
best_k = 0
best_score = -1
best_model = None

for k in range(2, 11):
    model = KMeans(n_clusters=k, random_state=42)
    labels = model.fit_predict(rfm_scaled)
    score = silhouette_score(rfm_scaled, labels)
    print(f"K={k} --> Silhouette Score: {score:.4f}")

    if score > best_score:
        best_k = k
        best_score = score
        best_model = model

# ■ Final Clustering
rfm['Cluster'] = best_model.labels_
print(f"\n■ Best K: {best_k} with Silhouette Score: {best_score:.4f}")

# ■ 3D Cluster Visualization
fig = px.scatter_3d(rfm, x='Recency', y='Frequency', z='Monetary', color='Cluster',
                    title=f'Customer Segmentation (K={best_k})',
                    labels={'Recency': 'Recency', 'Frequency': 'Frequency', 'Monetary': 'Monetary'})
fig.show()

# ■ Cluster Summary Table
summary = rfm.groupby('Cluster').agg({
    'Recency': 'mean',
    'Frequency': 'mean',
    'Monetary': ['mean', 'count']
```

```
}).round(2)

print("\n■ Cluster Summary:\n")
print(summary)
```

e. Output/Result:

The model achieved a silhouette score above 0.80, indicating well-defined clusters.

Customers were successfully segmented based on Recency, Frequency, and Monetary metrics.

Please paste your result screenshot here (from Google Colab output).

f. References/Credits:

- Dataset: Online Retail Dataset from UCI ML Repository

- Libraries: pandas, numpy, matplotlib, seaborn, sklearn, plotly

- IDE: Google Colab