Mae Fah Luang University
หมู่ที่ 1 333, Tha Sut, Mueang,
Chiang Rai 57100, Thailand

1501413 : Big Data Management & Analytics

Mini Project Report

---------- Title ----------

Customer Segmentation for E-Commerce using k-Means Clustering

Student ID: 6531501208 **Student NAME: Sai Hae Naing Lay**

Student ID: 6531501202 **Student NAME: Aung Ko Hein**

Student ID: 6531501237 **Student NAME: Zwe Lulin Maung**

Under the guidance of

Dr. Prof. Shanmugam Nandagopalan

School of Applied Digital Technology

Objectives

To segment customers based on their purchasing behavior using **Recency, Frequency, and Monetary (RFM)** features and apply **K-Means Clustering** to group similar customers for targeted marketing strategies.

1. Data Loading and Cleaning

- **Source:** `data.csv`
- **Encoding:** `ISO-8859-1`
- **Cleaning Steps:**
  - Dropped rows with missing `CustomerID` or `Description`.
  - Filtered only positive values for `Quantity` and `UnitPrice`.
  - Converted `InvoiceDate` to datetime format.
  - Created a new feature `TotalAmount = Quantity × UnitPrice`.

2. RFM Feature Engineering

- **Reference Date:** One day after the latest `InvoiceDate` in the dataset.
- **RFM Metrics:**
  - **Recency:** Days since the customer's most recent purchase.
  - **Frequency:** Number of unique invoices (transactions).
  - **Monetary:** Total amount spent.

These metrics were grouped by `CustomerID` to profile each customer.

3. Feature Scaling

- **Tool Used:** `StandardScaler` from `sklearn`
- **Reason:** To normalize the RFM features so that each contributes equally to distance calculations in clustering.

4. Finding Optimal Number of Clusters (k)

- **Method Used:** Elbow Method
- **Metric:** SSE (Sum of Squared Errors/Inertia)
- **Range Tested:** 1 to 10 clusters
- **Observation:** A bend (elbow) around `k=4`, suggesting 4 is the optimal number of clusters.

5. K-Means Clustering

- **Chosen k:** 4
- **Model:** KMeans with `random_state=42` for reproducibility
- **Assignment:** Each customer was assigned to one of the 4 clusters.

6. Cluster Analysis

Summary Table:

```
Cluster Summary:
         Recency  Frequency   Monetary  NumCustomers
Cluster
0          43.70       3.68    1359.05          3054
1         248.08       1.55     480.62          1067
2           7.38      82.54  127338.31            13
3          15.50      22.33   12709.09           204
```

Interpretation Tips:

- **Lower Recency** = More recent customers.
- **Higher Frequency and Monetary** = More valuable and loyal customers.
- This analysis helps in understanding customer segments like:
    - Loyal customers
    - High spenders
    - At-risk customers
    - New customers

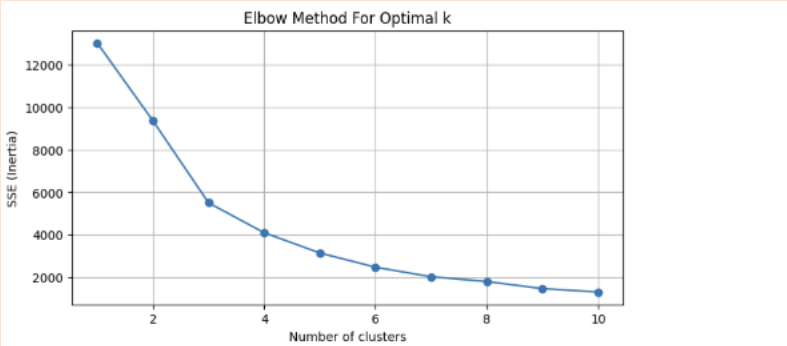
7. Visualization

- **Tool Used:** `Seaborn PairPlot`
- **Dimensions:** Recency, Frequency, and Monetary
- **Colored by:** Cluster
- **Purpose:** To visually explore the separation and distribution of customers across clusters.

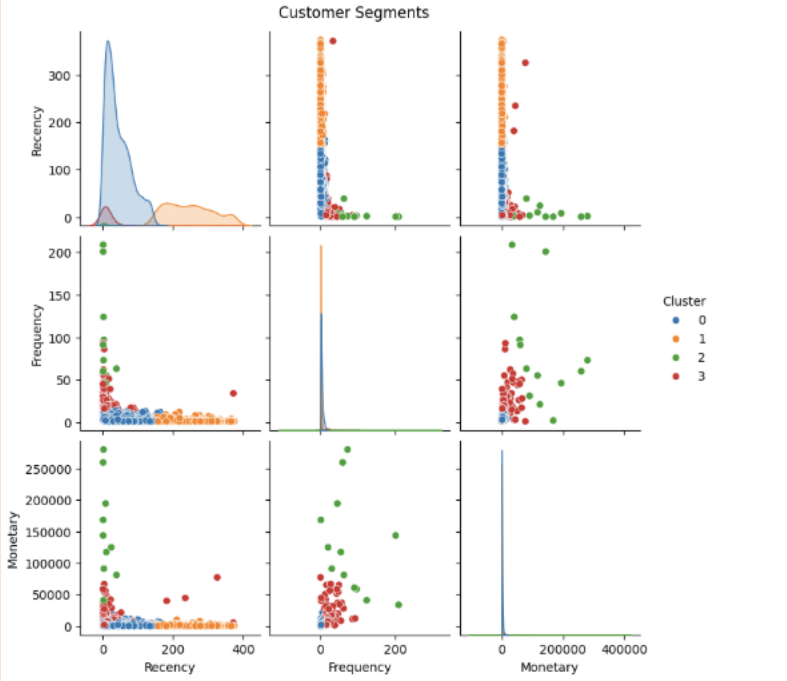
Complete Code / Program

```python
import pandas as pd
import matplotlib.pyplot as plt
from datetime import timedelta
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
import seaborn as sns

df = pd.read_csv("data.csv", encoding='ISO-8859-1')

df = df.dropna(subset=["CustomerID", "Description"])
df = df[(df["Quantity"] > 0) & (df["UnitPrice"] > 0)]
df["InvoiceDate"] = pd.to_datetime(df["InvoiceDate"])
df["TotalAmount"] = df["Quantity"] * df["UnitPrice"]

reference_date = df["InvoiceDate"].max() + timedelta(days=1)
rfm = df.groupby("CustomerID").agg({
    "InvoiceDate": lambda x: (reference_date - x.max()).days,
    "InvoiceNo": "nunique",
    "TotalAmount": "sum"
}).reset_index()

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

scaler = StandardScaler()
rfm_scaled = scaler.fit_transform(rfm[["Recency", "Frequency", "Monetary"]])

sse = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(rfm_scaled)
    sse.append(kmeans.inertia_)

plt.figure(figsize=(8, 4))
plt.plot(range(1, 11), sse, marker='o')
plt.title("Elbow Method For Optimal k")
plt.xlabel("Number of clusters")
plt.ylabel("SSE (Inertia)")
plt.grid(True)
plt.show()

k = 4
kmeans = KMeans(n_clusters=k, random_state=42)
rfm["Cluster"] = kmeans.fit_predict(rfm_scaled)

cluster_summary = rfm.groupby("Cluster").agg({
    "Recency": "mean",
    "Frequency": "mean",
    "Monetary": "mean",
    "CustomerID": "count"
}).rename(columns={"CustomerID": "NumCustomers"}).round(2)

print("\nCluster Summary:")
print(cluster_summary)

sns.pairplot(rfm, hue="Cluster", palette="tab10", vars=["Recency", "Frequency", "Monetary"])
plt.suptitle("Customer Segments", y=1.02)
plt.show()
```

Result




```
Cluster Summary:
         Recency  Frequency   Monetary  NumCustomers
Cluster
0          43.70       3.68    1359.05          3054
1         248.08       1.55     480.62          1067
2           7.38      82.54  127338.31            13
3          15.50      22.33   12709.09           204
```



Conclusion

This project successfully performed customer segmentation using the RFM model and k-means clustering. The insights derived from the cluster profiles can help businesses:

- Tailor marketing strategies
- Prioritize high-value customers
- Re-engage inactive customers
- Optimize customer service efforts

Reference

https://www.kaggle.com/datasets/carrie1/ecommerce-data?resource=download