Assignment No 9

Name: Ashutosh Shivthare
Roll No: C43364
Batch: B16

Program:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

# Load the dataset
data = pd.read_csv('sales_data_sample.csv', encoding='ISO-8859-1')


# Display the first few rows of the dataset
print(data.head())

# Check for missing values
print("Missing values in the dataset:")
print(data.isnull().sum())

# Select relevant features for clustering (adjust as necessary)
# Example: using numerical columns only
features = data.select_dtypes(include=[np.number])

# Handle missing values (if any)
features.fillna(features.mean(), inplace=True)

# Feature scaling
scaler = StandardScaler()
scaled_features = scaler.fit_transform(features)

# Elbow method to determine the optimal number of clusters
inertia = []
K = range(1, 11)  # Test for 1 to 10 clusters
for k in K:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(scaled_features)
    inertia.append(kmeans.inertia_)

# Plotting the elbow graph
plt.figure(figsize=(10, 6))
plt.plot(K, inertia, marker='o')
plt.title('Elbow Method for Optimal k')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Inertia')
plt.xticks(K)
plt.grid()
plt.show()
```

```
# From the elbow plot, choose the optimal k
optimal_k = 3  # Example, change this based on your elbow plot observation

# Apply K-Means clustering with the optimal number of clusters
kmeans = KMeans(n_clusters=optimal_k, random_state=42)
data['Cluster'] = kmeans.fit_predict(scaled_features)

# Visualizing the clusters (optional, for 2D visualization)
plt.figure(figsize=(10, 6))
plt.scatter(scaled_features[:, 0], scaled_features[:, 1], c=data['Cluster'], cmap='viridis')
plt.title('K-Means Clustering')
plt.xlabel('Feature 1')
plt.ylabel('Feature 2')
plt.colorbar(label='Cluster')
plt.grid()
plt.show()
```

Output:

```
= RESTART: C:\Users\ Akansha Sonar
OneDrive\Documents\Academics\BE\Lp3\Assignment No B4\Assignment No B4.py
   ORDERNUMBER  QUANTITYORDERED  ...  CONTACTFIRSTNAME  DEALSIZE
0        10107               30  ...             Kwai     Small
1        10121               34  ...             Paul     Small
2        10134               41  ...            Daniel    Medium
3        10145               45  ...            Julie     Medium
4        10159               49  ...            Julie     Medium

[5 rows x 25 columns]
Missing values in the dataset:
ORDERNUMBER           0
QUANTITYORDERED       0
PRICEEACH             0
ORDERLINENUMBER       0
SALES                 0
ORDERDATE             0
STATUS                0
QTR_ID                0
MONTH_ID              0
YEAR_ID               0
PRODUCTLINE           0
MSRP                  0
PRODUCTCODE           0
CUSTOMERNAME          0
PHONE                 0
ADDRESSLINE1          0
ADDRESSLINE2       2521
CITY                  0
STATE              1486
POSTALCODE           76
```

COUNTRY 0
TERRITORY 1074
CONTACTLASTNAME 0
CONTACTFIRSTNAME 0
DEALSIZE 0
dtype: int64