ass5ml

[1]:
```
# Step 1: Import necessary 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 scipy.cluster.hierarchy import dendrogram, linkage
```

[8]:
```
# Attempt to read the dataset with a different encoding
df = pd.read_csv('sales.csv', encoding='ISO-8859-1')
```

[9]:
```
df.head()
```

[9]:

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	\
0	10107	30	95.70	2	2871.00	
1	10121	34	81.35	5	2765.90	
2	10134	41	94.74	2	3884.34	
3	10145	45	83.26	6	3746.70	
4	10159	49	100.00	14	5205.27	

	ORDERDATE	STATUS	QTR_ID	MONTH_ID	YEAR_ID	…	\
0	2/24/2003 0:00	Shipped	1	2	2003	…	
1	5/7/2003 0:00	Shipped	2	5	2003	…	
2	7/1/2003 0:00	Shipped	3	7	2003	…	
3	8/25/2003 0:00	Shipped	3	8	2003	…	
4	10/10/2003 0:00	Shipped	4	10	2003	…	

	ADDRESSLINE1	ADDRESSLINE2	CITY	STATE	\
0	897 Long Airport Avenue	NaN	NYC	NY	
1	59 rue de l'Abbaye	NaN	Reims	NaN	
2	27 rue du Colonel Pierre Avia	NaN	Paris	NaN	
3	78934 Hillside Dr.	NaN	Pasadena	CA	
4	7734 Strong St.	NaN	San Francisco	CA	

	POSTALCODE	COUNTRY	TERRITORY	CONTACTLASTNAME	CONTACTFIRSTNAME	DEALSIZE
0	10022	USA	NaN	Yu	Kwai	Small

```
1      51100  France      EMEA         Henriot             Paul    Small
2      75508  France      EMEA       Da Cunha           Daniel   Medium
3      90003     USA       NaN          Young            Julie   Medium
4        NaN     USA       NaN          Brown            Julie   Medium

[5 rows x 25 columns]
```

[10]:
```
df.tail()
```

[10]:
```
      ORDERNUMBER  QUANTITYORDERED  PRICEEACH  ORDERLINENUMBER    SALES  \
2818        10350               20     100.00               15  2244.40
2819        10373               29     100.00                1  3978.51
2820        10386               43     100.00                4  5417.57
2821        10397               34      62.24                1  2116.16
2822        10414               47      65.52                9  3079.44

           ORDERDATE    STATUS  QTR_ID  MONTH_ID  YEAR_ID  …  \
2818  12/2/2004 0:00   Shipped       4        12     2004  …
2819  1/31/2005 0:00   Shipped       1         1     2005  …
2820   3/1/2005 0:00  Resolved       1         3     2005  …
2821  3/28/2005 0:00   Shipped       1         3     2005  …
2822   5/6/2005 0:00   On Hold       2         5     2005  …

               ADDRESSLINE1  ADDRESSLINE2      CITY STATE POSTALCODE  COUNTRY  \
2818     C/ Moralzarzal, 86           NaN    Madrid   NaN      28034    Spain
2819            Torikatu 38           NaN      Oulu   NaN      90110  Finland
2820     C/ Moralzarzal, 86           NaN    Madrid   NaN      28034    Spain
2821  1 rue Alsace-Lorraine           NaN  Toulouse   NaN      31000   France
2822     8616 Spinnaker Dr.           NaN    Boston    MA      51003      USA

     TERRITORY CONTACTLASTNAME CONTACTFIRSTNAME DEALSIZE
2818      EMEA          Freyre            Diego    Small
2819      EMEA       Koskitalo           Pirkko   Medium
2820      EMEA          Freyre            Diego   Medium
2821      EMEA          Roulet          Annette    Small
2822       NaN         Yoshido             Juri   Medium

[5 rows x 25 columns]
```

[11]:
```
df.isnull()
```

[11]:
```
      ORDERNUMBER  QUANTITYORDERED  PRICEEACH  ORDERLINENUMBER  SALES  \
0           False            False      False            False  False
1           False            False      False            False  False
2           False            False      False            False  False
3           False            False      False            False  False
4           False            False      False            False  False
```

```
…          …              …        …              …    …
2818       False          False    False          False  False
2819       False          False    False          False  False
2820       False          False    False          False  False
2821       False          False    False          False  False
2822       False          False    False          False  False

      ORDERDATE  STATUS  QTR_ID  MONTH_ID  YEAR_ID  …  ADDRESSLINE1  \
0         False   False   False     False    False  …         False
1         False   False   False     False    False  …         False
2         False   False   False     False    False  …         False
3         False   False   False     False    False  …         False
4         False   False   False     False    False  …         False
…         …     …     …       …       …  …           …
2818      False   False   False     False    False  …         False
2819      False   False   False     False    False  …         False
2820      False   False   False     False    False  …         False
2821      False   False   False     False    False  …         False
2822      False   False   False     False    False  …         False

      ADDRESSLINE2   CITY  STATE  POSTALCODE  COUNTRY  TERRITORY  \
0             True  False  False       False    False       True
1             True  False   True       False    False      False
2             True  False   True       False    False      False
3             True  False  False       False    False       True
4             True  False  False        True    False       True
…            …    …    …         …       …        …
2818          True  False   True       False    False      False
2819          True  False   True       False    False      False
2820          True  False   True       False    False      False
2821          True  False   True       False    False      False
2822          True  False  False       False    False       True

      CONTACTLASTNAME  CONTACTFIRSTNAME  DEALSIZE
0               False             False     False
1               False             False     False
2               False             False     False
3               False             False     False
4               False             False     False
…               …                …       …
2818            False             False     False
2819            False             False     False
2820            False             False     False
2821            False             False     False
2822            False             False     False

[2823 rows x 25 columns]
```

[12]:
```
df.isnull().sum()
```

[12]:
```
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
```

[13]:
```
df.notnull()
```

[13]:
```
      ORDERNUMBER  QUANTITYORDERED  PRICEEACH  ORDERLINENUMBER  SALES  \
0            True             True       True             True   True
1            True             True       True             True   True
2            True             True       True             True   True
3            True             True       True             True   True
4            True             True       True             True   True
…             …                …          …                …      …
2818         True             True       True             True   True
2819         True             True       True             True   True
2820         True             True       True             True   True
2821         True             True       True             True   True
2822         True             True       True             True   True

      ORDERDATE  STATUS  QTR_ID  MONTH_ID  YEAR_ID  …  ADDRESSLINE1  \
0          True    True    True      True     True  …          True
1          True    True    True      True     True  …          True
```

```
2          True    True    True      True     True  …          True
3          True    True    True      True     True  …          True
4          True    True    True      True     True  …          True
…           …       …       …         …        …  …           …
2818       True    True    True      True     True  …          True
2819       True    True    True      True     True  …          True
2820       True    True    True      True     True  …          True
2821       True    True    True      True     True  …          True
2822       True    True    True      True     True  …          True

      ADDRESSLINE2  CITY  STATE  POSTALCODE  COUNTRY  TERRITORY  \
0            False  True   True        True     True      False
1            False  True  False        True     True       True
2            False  True  False        True     True       True
3            False  True   True        True     True      False
4            False  True   True       False     True      False
…              …   …     …          …        …         …
2818         False  True  False        True     True       True
2819         False  True  False        True     True       True
2820         False  True  False        True     True       True
2821         False  True  False        True     True       True
2822         False  True   True        True     True      False

      CONTACTLASTNAME  CONTACTFIRSTNAME  DEALSIZE
0                True              True      True
1                True              True      True
2                True              True      True
3                True              True      True
4                True              True      True
…                …                 …         …
2818             True              True      True
2819             True              True      True
2820             True              True      True
2821             True              True      True
2822             True              True      True

[2823 rows x 25 columns]
```

[14]:
```
df.notnull().sum()
```

[14]:
```
ORDERNUMBER         2823
QUANTITYORDERED     2823
PRICEEACH           2823
ORDERLINENUMBER     2823
SALES               2823
ORDERDATE           2823
STATUS              2823
```

```
QTR_ID              2823
MONTH_ID            2823
YEAR_ID             2823
PRODUCTLINE         2823
MSRP                2823
PRODUCTCODE         2823
CUSTOMERNAME        2823
PHONE               2823
ADDRESSLINE1        2823
ADDRESSLINE2         302
CITY                2823
STATE               1337
POSTALCODE          2747
COUNTRY             2823
TERRITORY           1749
CONTACTLASTNAME     2823
CONTACTFIRSTNAME    2823
DEALSIZE            2823
dtype: int64
```

[15]:
```
df.describe()
```

[15]:

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	\
count	2823.000000	2823.000000	2823.000000	2823.000000	
mean	10258.725115	35.092809	83.658544	6.466171	
std	92.085478	9.741443	20.174277	4.225841	
min	10100.000000	6.000000	26.880000	1.000000	
25%	10180.000000	27.000000	68.860000	3.000000	
50%	10262.000000	35.000000	95.700000	6.000000	
75%	10333.500000	43.000000	100.000000	9.000000	
max	10425.000000	97.000000	100.000000	18.000000	

	SALES	QTR_ID	MONTH_ID	YEAR_ID	MSRP
count	2823.000000	2823.000000	2823.000000	2823.00000	2823.000000
mean	3553.889072	2.717676	7.092455	2003.81509	100.715551
std	1841.865106	1.203878	3.656633	0.69967	40.187912
min	482.130000	1.000000	1.000000	2003.00000	33.000000
25%	2203.430000	2.000000	4.000000	2003.00000	68.000000
50%	3184.800000	3.000000	8.000000	2004.00000	99.000000
75%	4508.000000	4.000000	11.000000	2004.00000	124.000000
max	14082.800000	4.000000	12.000000	2005.00000	214.000000

[16]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2823 entries, 0 to 2822
Data columns (total 25 columns):
```

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ORDERNUMBER       2823 non-null   int64
 1   QUANTITYORDERED   2823 non-null   int64
 2   PRICEEACH         2823 non-null   float64
 3   ORDERLINENUMBER   2823 non-null   int64
 4   SALES             2823 non-null   float64
 5   ORDERDATE         2823 non-null   object
 6   STATUS            2823 non-null   object
 7   QTR_ID            2823 non-null   int64
 8   MONTH_ID          2823 non-null   int64
 9   YEAR_ID           2823 non-null   int64
 10  PRODUCTLINE       2823 non-null   object
 11  MSRP              2823 non-null   int64
 12  PRODUCTCODE       2823 non-null   object
 13  CUSTOMERNAME      2823 non-null   object
 14  PHONE             2823 non-null   object
 15  ADDRESSLINE1      2823 non-null   object
 16  ADDRESSLINE2      302 non-null    object
 17  CITY              2823 non-null   object
 18  STATE             1337 non-null   object
 19  POSTALCODE        2747 non-null   object
 20  COUNTRY           2823 non-null   object
 21  TERRITORY         1749 non-null   object
 22  CONTACTLASTNAME   2823 non-null   object
 23  CONTACTFIRSTNAME  2823 non-null   object
 24  DEALSIZE          2823 non-null   object
dtypes: float64(2), int64(7), object(16)
memory usage: 551.5+ KB
```

[19]:
```
# Remove the column 'addressline 2'
df = df.drop(columns=['ADDRESSLINE2'], errors='ignore')
```

[20]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2823 entries, 0 to 2822
Data columns (total 24 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ORDERNUMBER       2823 non-null   int64
 1   QUANTITYORDERED   2823 non-null   int64
 2   PRICEEACH         2823 non-null   float64
 3   ORDERLINENUMBER   2823 non-null   int64
 4   SALES             2823 non-null   float64
 5   ORDERDATE         2823 non-null   object
 6   STATUS            2823 non-null   object
 7   QTR_ID            2823 non-null   int64
```

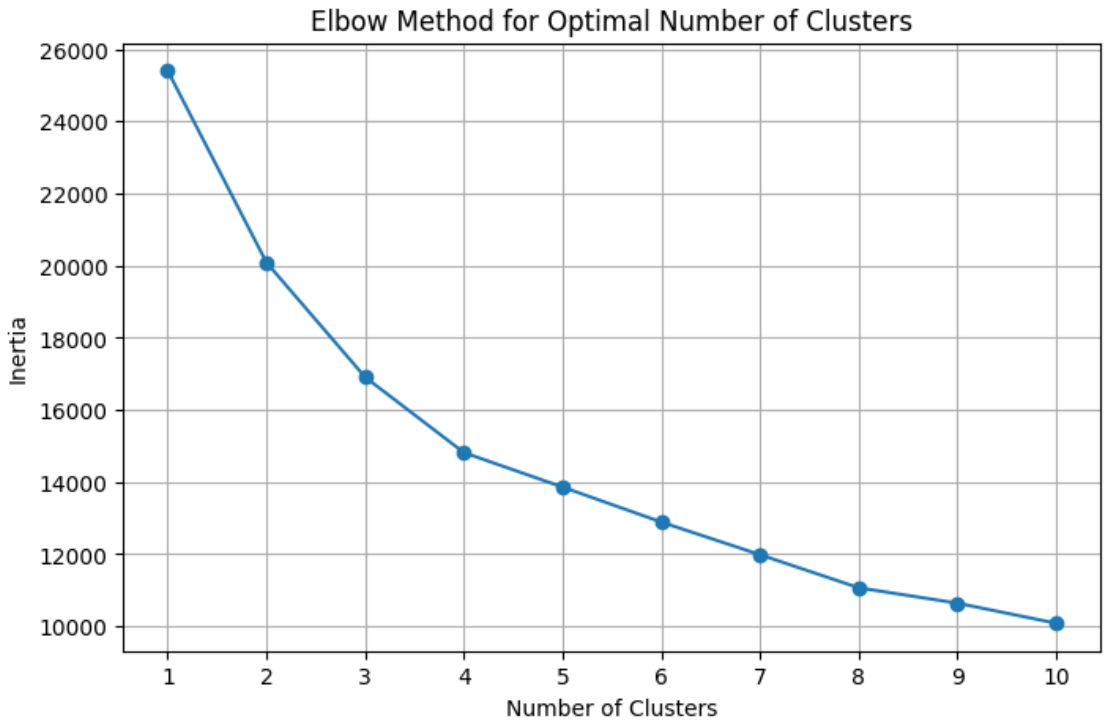
```
 8   MONTH_ID          2823 non-null   int64
 9   YEAR_ID           2823 non-null   int64
 10  PRODUCTLINE       2823 non-null   object
 11  MSRP              2823 non-null   int64
 12  PRODUCTCODE       2823 non-null   object
 13  CUSTOMERNAME      2823 non-null   object
 14  PHONE             2823 non-null   object
 15  ADDRESSLINE1      2823 non-null   object
 16  CITY              2823 non-null   object
 17  STATE             1337 non-null   object
 18  POSTALCODE        2747 non-null   object
 19  COUNTRY           2823 non-null   object
 20  TERRITORY         1749 non-null   object
 21  CONTACTLASTNAME   2823 non-null   object
 22  CONTACTFIRSTNAME  2823 non-null   object
 23  DEALSIZE          2823 non-null   object
dtypes: float64(2), int64(7), object(15)
memory usage: 529.4+ KB
```

[21]:
```python
# Step 4: Preprocess the data
X = df.select_dtypes(include=[np.number])  # Selecting only numeric columns
```

[22]:
```python
# Standardizing the data
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

[23]:
```python
# Step 5: Determine the optimal number of clusters using the elbow method
inertia = []
K = range(1, 11)

for k in K:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(X_scaled)
    inertia.append(kmeans.inertia_)
```

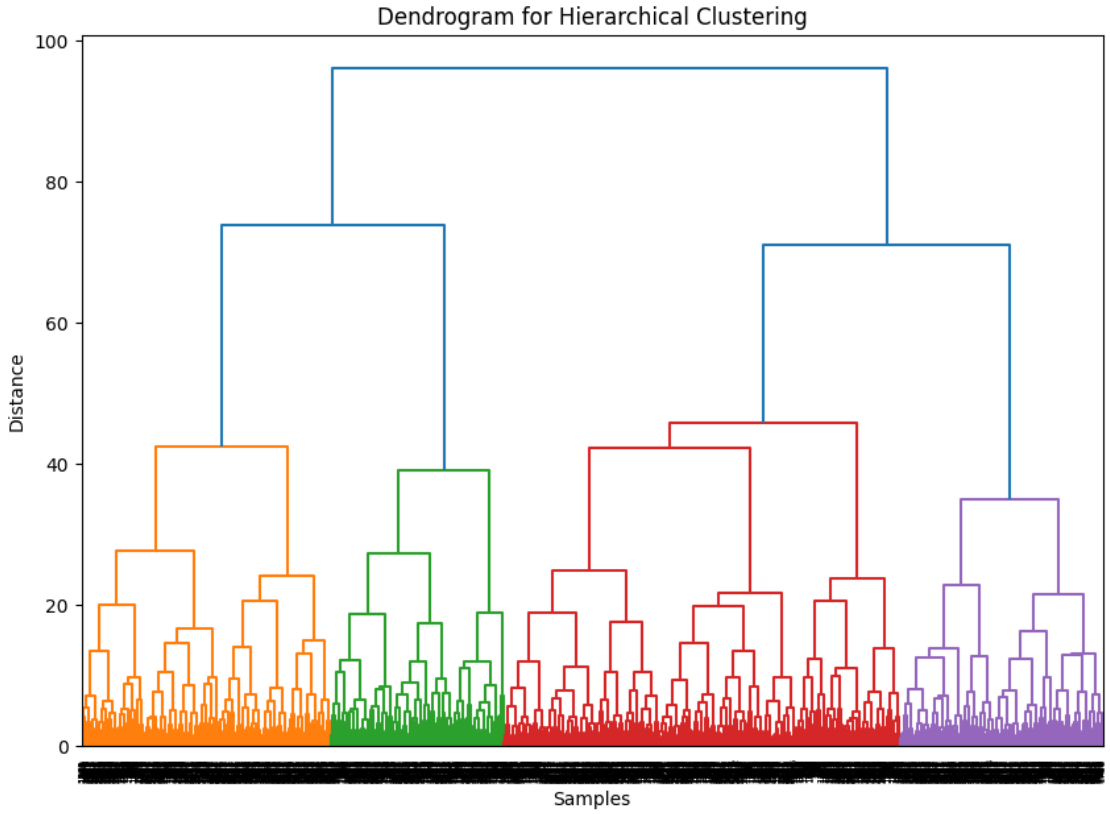
[24]:
```python
# Plotting the elbow curve
plt.figure(figsize=(8, 5))
plt.plot(K, inertia, marker='o')
plt.title('Elbow Method for Optimal Number of Clusters')
plt.xlabel('Number of Clusters')
plt.ylabel('Inertia')
plt.xticks(K)
plt.grid()
plt.show()
```



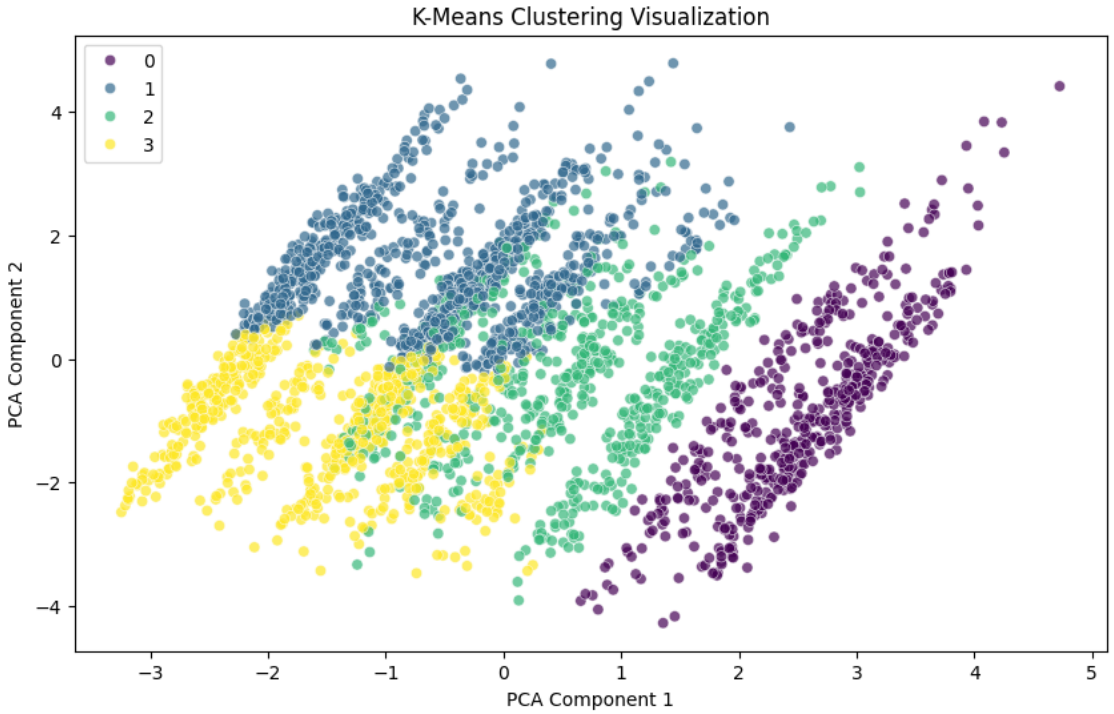
[25]:
```
# Step 6: Apply K-Means clustering with the optimal number of clusters
optimal_k = 4  # Replace this with the optimal number you find from the elbow␣
 ↪method
kmeans = KMeans(n_clusters=optimal_k, random_state=42)
clusters = kmeans.fit_predict(X_scaled)
```

[26]:
```
# Add the cluster labels to the original DataFrame
df['Cluster'] = clusters
```

[27]:
```
# Step 7: Apply Hierarchical clustering
linked = linkage(X_scaled, 'ward')

# Step 8: Plotting the dendrogram for Hierarchical clustering
plt.figure(figsize=(10, 7))
dendrogram(linked, orientation='top', distance_sort='descending',␣
 ↪show_leaf_counts=True)
plt.title('Dendrogram for Hierarchical Clustering')
plt.xlabel('Samples')
plt.ylabel('Distance')
plt.show()
```



[28]:
```
# Optional: Visualizing K-Means Clusters
from sklearn.decomposition import PCA

pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

plt.figure(figsize=(10, 6))
sns.scatterplot(x=X_pca[:, 0], y=X_pca[:, 1], hue=clusters, palette='viridis',␣
↪alpha=0.7)
plt.title('K-Means Clustering Visualization')
plt.xlabel('PCA Component 1')
plt.ylabel('PCA Component 2')
plt.legend()
plt.show()
```


K-Means Clustering Visualization
0
1
2
3
PCA Component 2
PCA Component 1
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
4
5