Sales_K-means_clustering

November 8, 2023

[4]:
```
import numpy as np
import pandas as pd
```

[7]:
```
df = pd.read_csv(r"C:\Users\lalit\OneDrive\Desktop\practicals\machine␣
 ↪learning\practical4\sales_data_sample.csv")
```

[8]:
```
df.head()
```

[8]:
```
   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]
```

[9]:
```
df.head(50)
```

[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
5         10168               36      96.66                1   3479.76
6         10180               29      86.13                9   2497.77
7         10188               48     100.00                1   5512.32
8         10201               22      98.57                2   2168.54
9         10211               41     100.00               14   4708.44
10        10223               37     100.00                1   3965.66
11        10237               23     100.00                7   2333.12
12        10251               28     100.00                2   3188.64
13        10263               34     100.00                2   3676.76
14        10275               45      92.83                1   4177.35
15        10285               36     100.00                6   4099.68
16        10299               23     100.00                9   2597.39
17        10309               41     100.00                5   4394.38
18        10318               46      94.74                1   4358.04
19        10329               42     100.00                1   4396.14
20        10341               41     100.00                9   7737.93
21        10361               20      72.55               13   1451.00
22        10375               21      34.91               12    733.11
23        10388               42      76.36                4   3207.12
24        10403               24     100.00                7   2434.56
25        10417               66     100.00                2   7516.08
26        10103               26     100.00               11   5404.62
27        10112               29     100.00                1   7209.11
28        10126               38     100.00               11   7329.06
29        10140               37     100.00               11   7374.10
30        10150               45     100.00                8  10993.50
31        10163               21     100.00                1   4860.24
32        10174               34     100.00                4   8014.82
33        10183               23     100.00                8   5372.57
34        10194               42     100.00               11   7290.36
35        10206               47     100.00                6   9064.89
36        10215               35     100.00                3   6075.30
37        10228               29     100.00                2   6463.23
38        10245               34     100.00                9   6120.34
39        10258               32     100.00                6   7680.64
40        10270               21     100.00                9   4905.39
41        10280               34     100.00                2   8014.82
42        10291               37     100.00               11   7136.19
43        10304               47     100.00                6  10172.70
```

```
44        10312               48     100.00                     3  11623.70
45        10322               40     100.00                     1   6000.40
46        10333               26     100.00                     3   3003.00
47        10347               30     100.00                     1   3944.70
48        10357               32     100.00                    10   5691.84
49        10369               41     100.00                     2   4514.92

          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  …
5   10/28/2003 0:00   Shipped       4        10     2003  …
6   11/11/2003 0:00   Shipped       4        11     2003  …
7   11/18/2003 0:00   Shipped       4        11     2003  …
8    12/1/2003 0:00   Shipped       4        12     2003  …
9    1/15/2004 0:00   Shipped       1         1     2004  …
10   2/20/2004 0:00   Shipped       1         2     2004  …
11    4/5/2004 0:00   Shipped       2         4     2004  …
12   5/18/2004 0:00   Shipped       2         5     2004  …
13   6/28/2004 0:00   Shipped       2         6     2004  …
14   7/23/2004 0:00   Shipped       3         7     2004  …
15   8/27/2004 0:00   Shipped       3         8     2004  …
16   9/30/2004 0:00   Shipped       3         9     2004  …
17  10/15/2004 0:00   Shipped       4        10     2004  …
18   11/2/2004 0:00   Shipped       4        11     2004  …
19  11/15/2004 0:00   Shipped       4        11     2004  …
20  11/24/2004 0:00   Shipped       4        11     2004  …
21  12/17/2004 0:00   Shipped       4        12     2004  …
22    2/3/2005 0:00   Shipped       1         2     2005  …
23    3/3/2005 0:00   Shipped       1         3     2005  …
24    4/8/2005 0:00   Shipped       2         4     2005  …
25   5/13/2005 0:00  Disputed       2         5     2005  …
26   1/29/2003 0:00   Shipped       1         1     2003  …
27   3/24/2003 0:00   Shipped       1         3     2003  …
28   5/28/2003 0:00   Shipped       2         5     2003  …
29   7/24/2003 0:00   Shipped       3         7     2003  …
30   9/19/2003 0:00   Shipped       3         9     2003  …
31  10/20/2003 0:00   Shipped       4        10     2003  …
32   11/6/2003 0:00   Shipped       4        11     2003  …
33  11/13/2003 0:00   Shipped       4        11     2003  …
34  11/25/2003 0:00   Shipped       4        11     2003  …
35   12/5/2003 0:00   Shipped       4        12     2003  …
36   1/29/2004 0:00   Shipped       1         1     2004  …
37   3/10/2004 0:00   Shipped       1         3     2004  …
38    5/4/2004 0:00   Shipped       2         5     2004  …
```

```
39   6/15/2004 0:00   Shipped       2         6      2004  …
40   7/19/2004 0:00   Shipped       3         7      2004  …
41   8/17/2004 0:00   Shipped       3         8      2004  …
42    9/8/2004 0:00   Shipped       3         9      2004  …
43  10/11/2004 0:00   Shipped       4        10      2004  …
44  10/21/2004 0:00   Shipped       4        10      2004  …
45   11/4/2004 0:00   Shipped       4        11      2004  …
46  11/18/2004 0:00   Shipped       4        11      2004  …
47  11/29/2004 0:00   Shipped       4        11      2004  …
48  12/10/2004 0:00   Shipped       4        12      2004  …
49   1/20/2005 0:00   Shipped       1         1      2005  …

                               ADDRESSLINE1  ADDRESSLINE2            CITY
0                   897 Long Airport Avenue           NaN             NYC  \
1                        59 rue de l'Abbaye           NaN           Reims
2             27 rue du Colonel Pierre Avia           NaN           Paris
3                        78934 Hillside Dr.           NaN        Pasadena
4                           7734 Strong St.           NaN   San Francisco
5                         9408 Furth Circle           NaN      Burlingame
6                   184, chausse de Tournai           NaN           Lille
7               Drammen 121, PR 744 Sentrum           NaN          Bergen
8                 5557 North Pendale Street           NaN   San Francisco
9                         25, rue Lauriston           NaN           Paris
10                        636 St Kilda Road       Level 3       Melbourne
11                        2678 Kingston Rd.     Suite 101             NYC
12                            7476 Moss Rd.           NaN          Newark
13                      25593 South Bay Ln.           NaN     Bridgewater
14             67, rue des Cinquante Otages           NaN          Nantes
15                      39323 Spinnaker Dr.           NaN       Cambridge
16                            Keskuskatu 45           NaN        Helsinki
17                   Erling Skakkes gate 78           NaN         Stavern
18                         7586 Pompton St.           NaN       Allentown
19                  897 Long Airport Avenue           NaN             NYC
20                              Geislweg 14           NaN        Salzburg
21  Monitor Money Building, 815 Pacific Hwy       Level 6       Chatswood
22             67, rue des Cinquante Otages           NaN          Nantes
23                        1785 First Street           NaN     New Bedford
24             Berkeley Gardens 12  Brewery           NaN       Liverpool
25                       C/ Moralzarzal, 86           NaN          Madrid
26                   Erling Skakkes gate 78           NaN         Stavern
27                          Berguvsv„gen  8           NaN            Lule
28                           C/ Araquil, 67           NaN          Madrid
29                        9408 Furth Circle           NaN      Burlingame
30     Bronz Sok., Bronz Apt. 3/6 Tesvikiye           NaN       Singapore
31                         5905 Pompton St.     Suite 750             NYC
32                   31 Duncan St. West End           NaN  South Brisbane
33                         782 First Street           NaN    Philadelphia
```

```
34                       2, rue du Commerce           NaN            Lyon
35                             1900 Oak St.           NaN       Vancouver
36                        3675 Furth Circle           NaN         Burbank
37                           4658 Baden Av.           NaN       Cambridge
38                 567 North Pendale Street           NaN       New Haven
39                           2-2-8 Roppongi           NaN       Minato-ku
40  Monitor Money Building, 815 Pacific Hwy       Level 6       Chatswood
41                      Via Monte Bianco 34           NaN          Torino
42                             ?kergatan 24           NaN           Boras
43                   67, avenue de l'Europe           NaN      Versailles
44                          5677 Strong St.           NaN      San Rafael
45                 2304 Long Airport Avenue           NaN          Nashua
46                5557 North Pendale Street           NaN   San Francisco
47                        636 St Kilda Road       Level 3       Melbourne
48                          5677 Strong St.           NaN      San Rafael
49                         7825 Douglas Av.           NaN      Brickhaven

         STATE POSTALCODE    COUNTRY TERRITORY CONTACTLASTNAME
0           NY      10022        USA       NaN              Yu  \
1          NaN      51100     France      EMEA         Henriot
2          NaN      75508     France      EMEA        Da Cunha
3           CA      90003        USA       NaN           Young
4           CA        NaN        USA       NaN           Brown
5           CA      94217        USA       NaN          Hirano
6          NaN      59000     France      EMEA           Rance
7          NaN     N 5804     Norway      EMEA          Oeztan
8           CA        NaN        USA       NaN          Murphy
9          NaN      75016     France      EMEA         Perrier
10    Victoria       3004  Australia      APAC        Ferguson
11          NY      10022        USA       NaN           Frick
12          NJ      94019        USA       NaN           Brown
13          CT      97562        USA       NaN            King
14         NaN      44000     France      EMEA         Labrune
15          MA      51247        USA       NaN       Hernandez
16         NaN      21240    Finland      EMEA       Karttunen
17         NaN       4110     Norway      EMEA      Bergulfsen
18          PA      70267        USA       NaN              Yu
19          NY      10022        USA       NaN              Yu
20         NaN       5020    Austria      EMEA           Pipps
21         NSW       2067  Australia      APAC          Huxley
22         NaN      44000     France      EMEA         Labrune
23          MA      50553        USA       NaN         Benitez
24         NaN    WX1 6LT         UK      EMEA           Devon
25         NaN      28034      Spain      EMEA          Freyre
26         NaN       4110     Norway      EMEA      Bergulfsen
27         NaN   S-958 22     Sweden      EMEA        Berglund
28         NaN      28023      Spain      EMEA          Sommer
```

```
29          CA      94217        USA       NaN          Hirano
30         NaN      79903  Singapore     Japan       Natividad
31          NY      10022        USA       NaN       Hernandez
32  Queensland       4101  Australia      APAC        Calaghan
33          PA      71270        USA       NaN       Cervantes
34         NaN      69004     France      EMEA         Saveley
35          BC    V3F 2K1     Canada       NaN       Tannamuri
36          CA      94019        USA       NaN        Thompson
37          MA      51247        USA       NaN           Tseng
38          CT      97823        USA       NaN          Murphy
39       Tokyo   106-0032      Japan     Japan       Shimamura
40         NSW       2067  Australia      APAC          Huxley
41         NaN      10100      Italy      EMEA         Accorti
42         NaN   S-844 67     Sweden      EMEA         Larsson
43         NaN      78000     France      EMEA          Tonini
44          CA      97562        USA       NaN          Nelson
45          NH      62005        USA       NaN           Young
46          CA        NaN        USA       NaN          Murphy
47    Victoria       3004  Australia      APAC        Ferguson
48          CA      97562        USA       NaN          Nelson
49          MA      58339        USA       NaN          Nelson

   CONTACTFIRSTNAME DEALSIZE
0              Kwai    Small
1              Paul    Small
2            Daniel   Medium
3             Julie   Medium
4             Julie   Medium
5              Juri   Medium
6           Martine    Small
7            Veysel   Medium
8             Julie    Small
9         Dominique   Medium
10            Peter   Medium
11          Michael    Small
12          William   Medium
13            Julie   Medium
14           Janine   Medium
15            Marta   Medium
16            Matti    Small
17            Jonas   Medium
18            Kyung   Medium
19             Kwai   Medium
20            Georg    Large
21           Adrian    Small
22           Janine    Small
23          Violeta   Medium
```

```
24        Elizabeth    Small
25            Diego    Large
26            Jonas   Medium
27        Christina    Large
28           Mart¡n    Large
29             Juri    Large
30             Eric    Large
31            Maria   Medium
32             Tony    Large
33        Francisca   Medium
34             Mary    Large
35            Yoshi    Large
36            Steve   Medium
37            Kyung   Medium
38           Leslie   Medium
39            Akiko    Large
40           Adrian   Medium
41            Paolo    Large
42            Maria    Large
43           Daniel    Large
44          Valarie    Large
45          Valarie   Medium
46            Julie   Medium
47            Peter   Medium
48          Valarie   Medium
49            Allen   Medium

[50 rows x 25 columns]
```

[10]:
```
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
```

[11]:
```
df.shape
```

[11]:
```
(2823, 25)
```

[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]:

```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
```

[21]:

```
# Select relevant numerical columns for clustering
numerical_columns = df[['QUANTITYORDERED', 'PRICEEACH', 'SALES']]
```

[23]:

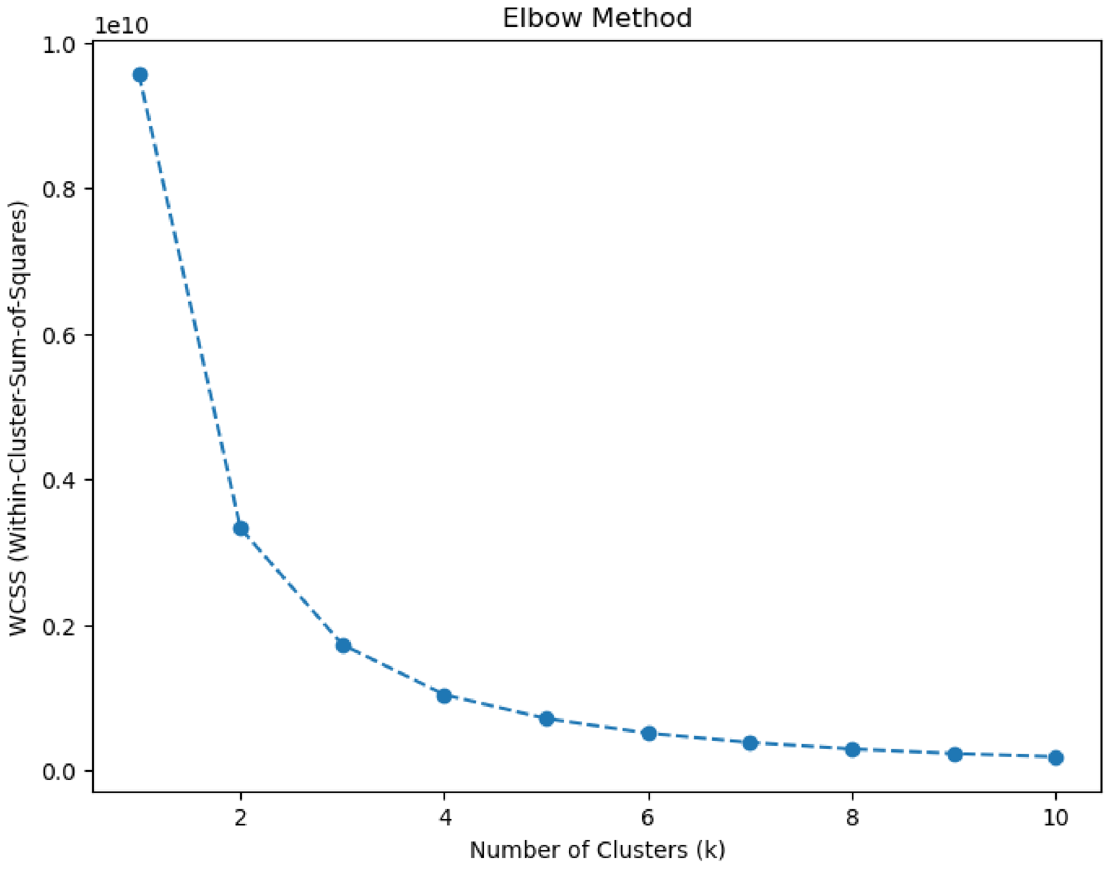
```
# Determine the optimal number of clusters using the elbow method
wcss = []  # Within-Cluster-Sum-of-Squares
```

[24]:

```
# Iterate over a range of values for k (number of clusters)
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(numerical_columns)  # Fit K-Means to the data
    wcss.append(kmeans.inertia_)  # Append the inertia value to the list
```

[25]:

```
# Plot the elbow graph
plt.figure(figsize=(8, 6))
plt.plot(range(1, 11), wcss, marker='o', linestyle='--')
plt.title('Elbow Method')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('WCSS (Within-Cluster-Sum-of-Squares)')
plt.show()
```



[26]:

```
# Based on the elbow method, let's say the optimal number of clusters is 3
optimal_k = 3
```

[29]:

```
# Perform K-Means clustering with the chosen number of clusters
kmeans = KMeans(n_clusters=optimal_k, random_state=42)
df['cluster'] = kmeans.fit_predict(numerical_columns)
```

[37]:

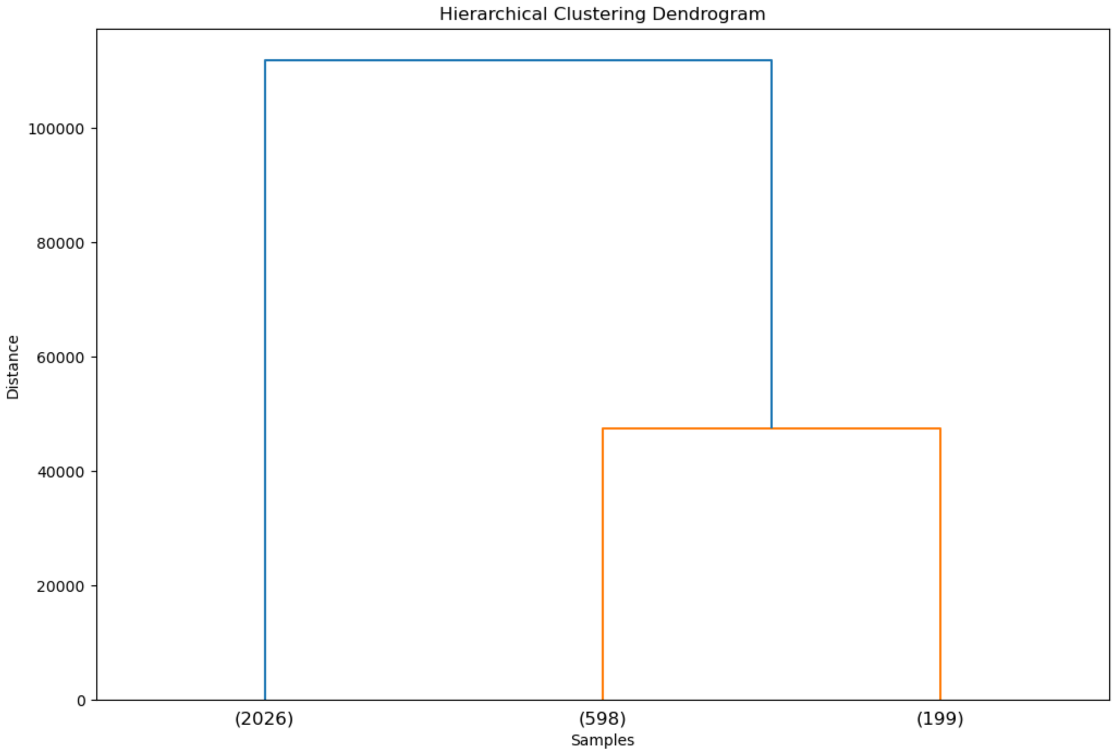
```
from scipy.cluster.hierarchy import dendrogram, linkage
import matplotlib.pyplot as plt

numerical_columns = df[['QUANTITYORDERED', 'PRICEEACH', 'SALES']]
```

[38]:

```
# Create a linkage matrix using the Ward method
Z = linkage(numerical_columns, method='ward')
```

[39]:

```
# Plot the hierarchical clustering dendrogram
plt.figure(figsize=(12, 8))
dendrogram(Z, p=optimal_k, truncate_mode='lastp')
plt.title('Hierarchical Clustering Dendrogram')
plt.xlabel('Samples')
plt.ylabel('Distance')
plt.show()
```



[40]:

```
from scipy.cluster.hierarchy import fcluster
```

[41]:

```
# Based on the dendrogram, let's say the optimal number of clusters is 3
optimal_k = 3
```

[42]:

```
# Cut the dendrogram to obtain cluster assignments
cluster_assignments = fcluster(Z, t=optimal_k, criterion='maxclust')
```

[44]:

```
# Add the cluster assignments to your original dataset
df['cluster'] = cluster_assignments
```

[47]:

```
df.head()
```

[47]:

```
   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  … ADDRESSLINE2
0   2/24/2003 0:00  Shipped       1         2     2003  …          NaN  \
1    5/7/2003 0:00  Shipped       2         5     2003  …          NaN
2    7/1/2003 0:00  Shipped       3         7     2003  …          NaN
3   8/25/2003 0:00  Shipped       3         8     2003  …          NaN
4  10/10/2003 0:00  Shipped       4        10     2003  …          NaN

            CITY STATE POSTALCODE COUNTRY TERRITORY CONTACTLASTNAME
0            NYC    NY      10022     USA       NaN              Yu  \
1          Reims   NaN      51100  France      EMEA         Henriot
2          Paris   NaN      75508  France      EMEA        Da Cunha
3       Pasadena    CA      90003     USA       NaN           Young
4  San Francisco    CA        NaN     USA       NaN           Brown

  CONTACTFIRSTNAME DEALSIZE cluster
0             Kwai    Small       1
1             Paul    Small       1
2           Daniel   Medium       1
3            Julie   Medium       1
4            Julie   Medium       2

[5 rows x 26 columns]
```

[ ]:

```

```