ML_5

November 11, 2025

[1]:
```
import pandas as pd
```

[3]:
```
df = pd.read_csv(r"C:
↪\Users\suraj\OneDrive\Desktop\LP3-master\ML\datasets\sales_data_sample_utf8.
↪csv")
```

[5]:
```
df.head()
```

[5]:
```
   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]
```

[7]:
```
data = df.select_dtypes(include=['float64','int64'])
data = data.fillna(data.mean())
data.head()
```

[7]:
```
   ORDERNUMBER  QUANTITYORDERED  PRICEEACH  ORDERLINENUMBER    SALES  QTR_ID  \
0        10107               30      95.70                2  2871.00       1
1        10121               34      81.35                5  2765.90       2
2        10134               41      94.74                2  3884.34       3
3        10145               45      83.26                6  3746.70       3
4        10159               49     100.00               14  5205.27       4

   MONTH_ID  YEAR_ID  MSRP
0         2     2003    95
1         5     2003    95
2         7     2003    95
3         8     2003    95
4        10     2003    95
```

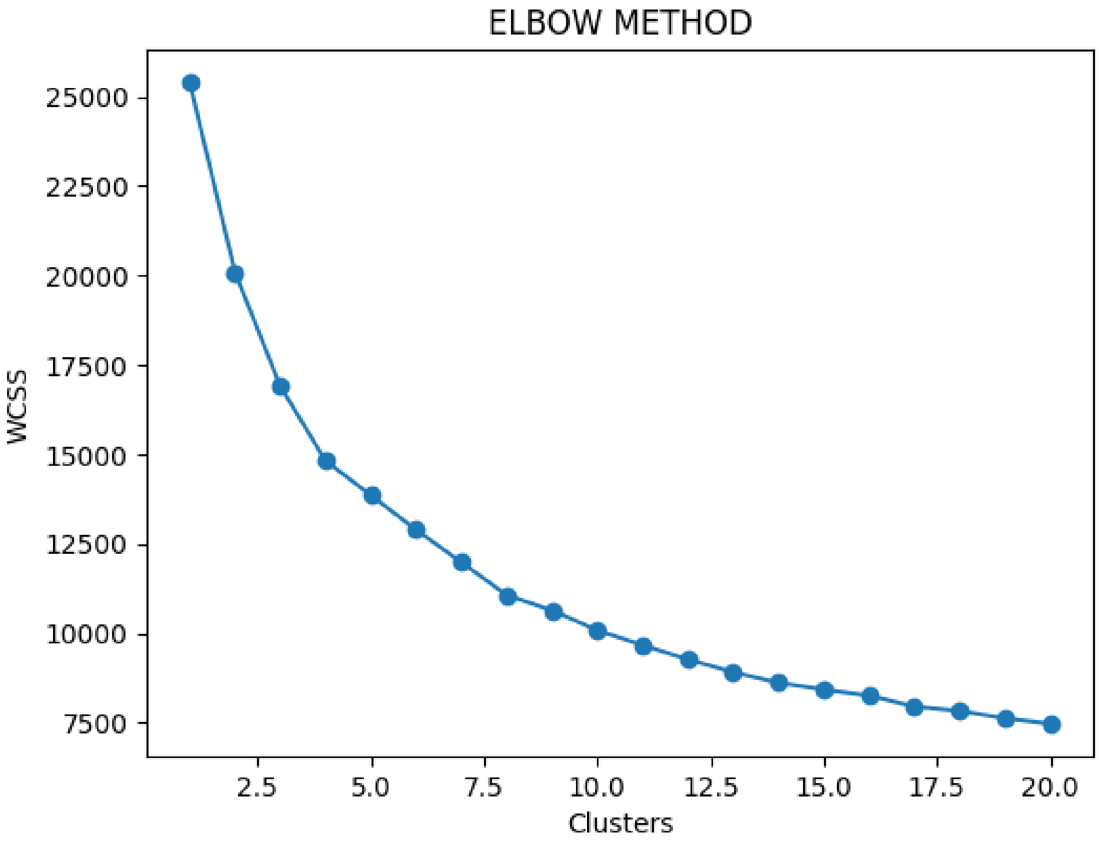
[9]:
```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
scaleddata = sc.fit_transform(data)
```

[17]:
```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
```

[18]:
```
wcss = []

for i in range(1, 21):
    kmeans = KMeans(n_clusters=i,random_state=42)
    kmeans.fit(scaleddata)
    wcss.append(kmeans.inertia_)
```

[21]:
```
plt.plot(range(1,21),wcss,marker='o')
plt.xlabel("Clusters")
plt.ylabel("WCSS")
plt.title("ELBOW METHOD")
plt.show()
```



[23]:

```
kmeans = KMeans(n_clusters = 5,random_state=42)
df['Clusters']=kmeans.fit_predict(scaleddata)
```

[25]:

```
df['Clusters'].head()
```

[25]:

```
0    2
1    2
2    4
3    4
4    1
Name: Clusters, dtype: int32
```

[27]:

```
df['Clusters'].value_counts()
```

[27]:

```
Clusters
3    647
2    631
1    575
4    508
0    462
```

```
Name: count, dtype: int64
```

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

```
[29]:    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 Clusters  
      0             Kwai    Small        2  
      1             Paul    Small        2  
      2           Daniel   Medium        4  
      3            Julie   Medium        4  
      4            Julie   Medium        1  

      [5 rows x 26 columns]
```

```
[ ]:
```