Assignment_2

October 9, 2024

0.1 Importing Libraries

[50]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

0.2 Reading Dataset

[51]:
```
df = pd.read_csv('sales_data_sample.csv', encoding='ISO-8859-1')
```

[52]:
```
df
```

[52]:

	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	
…	…	…	…	…	…	
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	…	\
0	2/24/2003 0:00	Shipped	1	2	2003	…	
1	05-07-2003 00:00	Shipped	2	5	2003	…	
2	07-01-2003 00:00	Shipped	3	7	2003	…	
3	8/25/2003 0:00	Shipped	3	8	2003	…	
4	10-10-2003 00:00	Shipped	4	10	2003	…	
…	…	…	…	…	…	…	
2818	12-02-2004 00:00	Shipped	4	12	2004	…	
2819	1/31/2005 0:00	Shipped	1	1	2005	…	
2820	03-01-2005 00:00	Resolved	1	3	2005	…	
2821	3/28/2005 0:00	Shipped	1	3	2005	…	
2822	05-06-2005 00:00	On Hold	2	5	2005	…	

```
                     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
…                               …           …            …   …
2818             C/ Moralzarzal, 86           NaN         Madrid   NaN
2819                   Torikatu 38           NaN           Oulu   NaN
2820             C/ Moralzarzal, 86           NaN         Madrid   NaN
2821          1 rue Alsace-Lorraine           NaN       Toulouse   NaN
2822             8616 Spinnaker Dr.           NaN         Boston    MA

     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
…         …      …       …              …                …       …
2818      28034    Spain      EMEA          Freyre            Diego    Small
2819      90110  Finland      EMEA       Koskitalo           Pirkko   Medium
2820      28034    Spain      EMEA          Freyre            Diego   Medium
2821      31000   France      EMEA          Roulet          Annette    Small
2822      51003      USA       NaN         Yoshido             Juri   Medium

[2823 rows x 25 columns]
```

[53]:
```
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
```

[54]:
```
df.isnull().sum()
```

[54]:
```
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
```

[55]:
```
df
```

[55]:

	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	
…	…	…	…	…	…	
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	…	\
0	2/24/2003 0:00	Shipped	1	2	2003	…	
1	05-07-2003 00:00	Shipped	2	5	2003	…	
2	07-01-2003 00:00	Shipped	3	7	2003	…	
3	8/25/2003 0:00	Shipped	3	8	2003	…	
4	10-10-2003 00:00	Shipped	4	10	2003	…	
…	…	…	…	…	…	…	
2818	12-02-2004 00:00	Shipped	4	12	2004	…	
2819	1/31/2005 0:00	Shipped	1	1	2005	…	
2820	03-01-2005 00:00	Resolved	1	3	2005	…	
2821	3/28/2005 0:00	Shipped	1	3	2005	…	
2822	05-06-2005 00:00	On Hold	2	5	2005	…	

	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	
…	…	…	…	…	
2818	C/ Moralzarzal, 86	NaN	Madrid	NaN	
2819	Torikatu 38	NaN	Oulu	NaN	
2820	C/ Moralzarzal, 86	NaN	Madrid	NaN	
2821	1 rue Alsace-Lorraine	NaN	Toulouse	NaN	
2822	8616 Spinnaker Dr.	NaN	Boston	MA	

	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
…	…	…	…	…	…	…
2818	28034	Spain	EMEA	Freyre	Diego	Small

```
2819      90110  Finland      EMEA       Koskitalo           Pirkko   Medium
2820      28034    Spain      EMEA          Freyre            Diego   Medium
2821      31000   France      EMEA          Roulet          Annette    Small
2822      51003      USA       NaN         Yoshido             Juri   Medium

[2823 rows x 25 columns]
```

0.3 UsIng Relevant Features

[58]:
```
data = df[['SALES', 'QUANTITYORDERED', 'PRICEEACH']].copy()
```

[59]:
```
data
```

[59]:
```
         SALES  QUANTITYORDERED  PRICEEACH
0     2871.00               30      95.70
1     2765.90               34      81.35
2     3884.34               41      94.74
3     3746.70               45      83.26
4     5205.27               49     100.00
…       …                …         …
2818  2244.40               20     100.00
2819  3978.51               29     100.00
2820  5417.57               43     100.00
2821  2116.16               34      62.24
2822  3079.44               47      65.52

[2823 rows x 3 columns]
```

[60]:
```
data.isnull().sum()
```

[60]:
```
SALES              0
QUANTITYORDERED    0
PRICEEACH          0
dtype: int64
```

0.4 Scaling the Data

[61]:
```
from sklearn.preprocessing import StandardScaler
```

[62]:
```
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data)
```

[63]:
```
data_scaled
```

[63]:
```
array([[-0.37082523, -0.52289086,  0.5969775 ],
       [-0.42789707, -0.11220131, -0.11445035],
       [ 0.17944282,  0.60650538,  0.54938372],
```

```
       …,
       [ 1.01202368,  0.81185016,  0.81015797],
       [-0.78072155, -0.11220131, -1.06186404],
       [-0.25763729,  1.2225397 , -0.89925195]])
```

[64]:
```
sse = []
k_range = range(1, 11)
```

0.5 Detemining the no. of clusters using Elbow Method

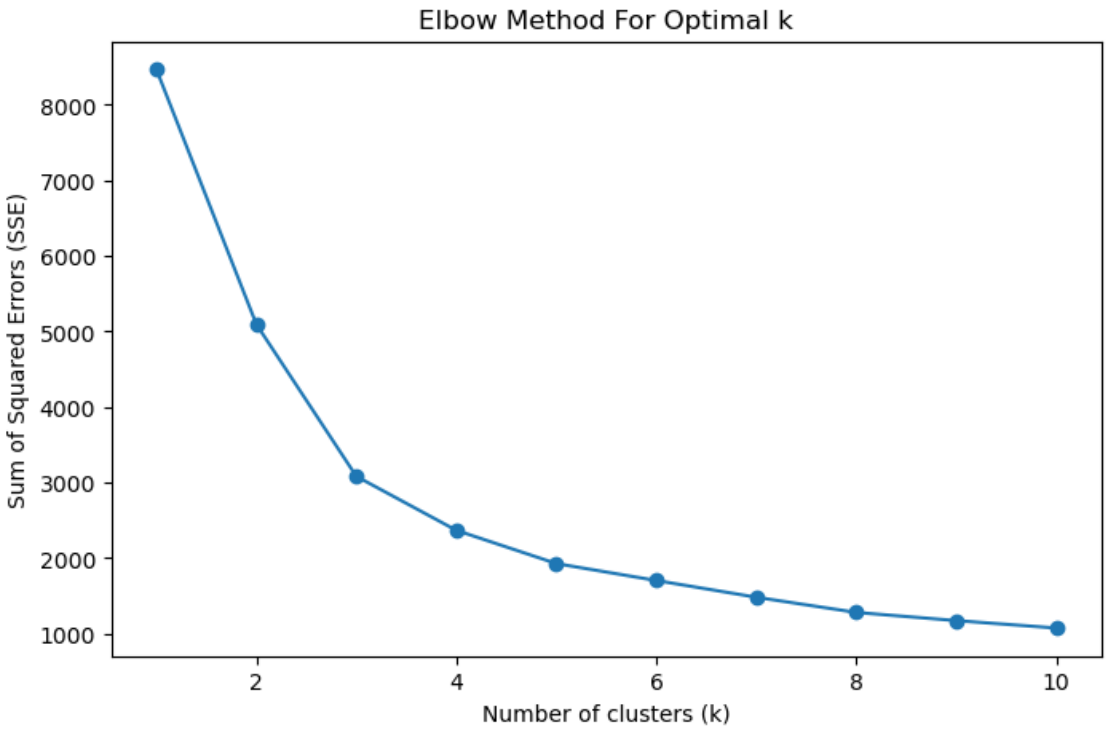
[65]:
```
from sklearn.cluster import KMeans
```

[66]:
```
for k in k_range:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(data_scaled)
    sse.append(kmeans.inertia_)
```

[67]:
```
sse
```

[67]:
```
[8469.0,
 5095.128140471077,
 3082.6505184877296,
 2368.229219196734,
 1928.5346658908288,
 1704.28784950298,
 1482.5707052959156,
 1281.6052111711733,
 1173.6424602864276,
 1075.423870472061]
```

[68]:
```
plt.figure(figsize=(8, 5))
plt.plot(k_range, sse, marker='o')
plt.title('Elbow Method For Optimal k')
plt.xlabel('Number of clusters (k)')
plt.ylabel('Sum of Squared Errors (SSE)')
plt.show()
```



0.6 Training the Model

[69]:
```
optimal_k = 3
kmeans = KMeans(n_clusters=optimal_k, random_state=42)
kmeans.fit(data_scaled)
```

[69]:
```
KMeans(n_clusters=3, random_state=42)
```

[71]:
```
df['Cluster'] = kmeans.labels_
```

[72]:
```
df
```

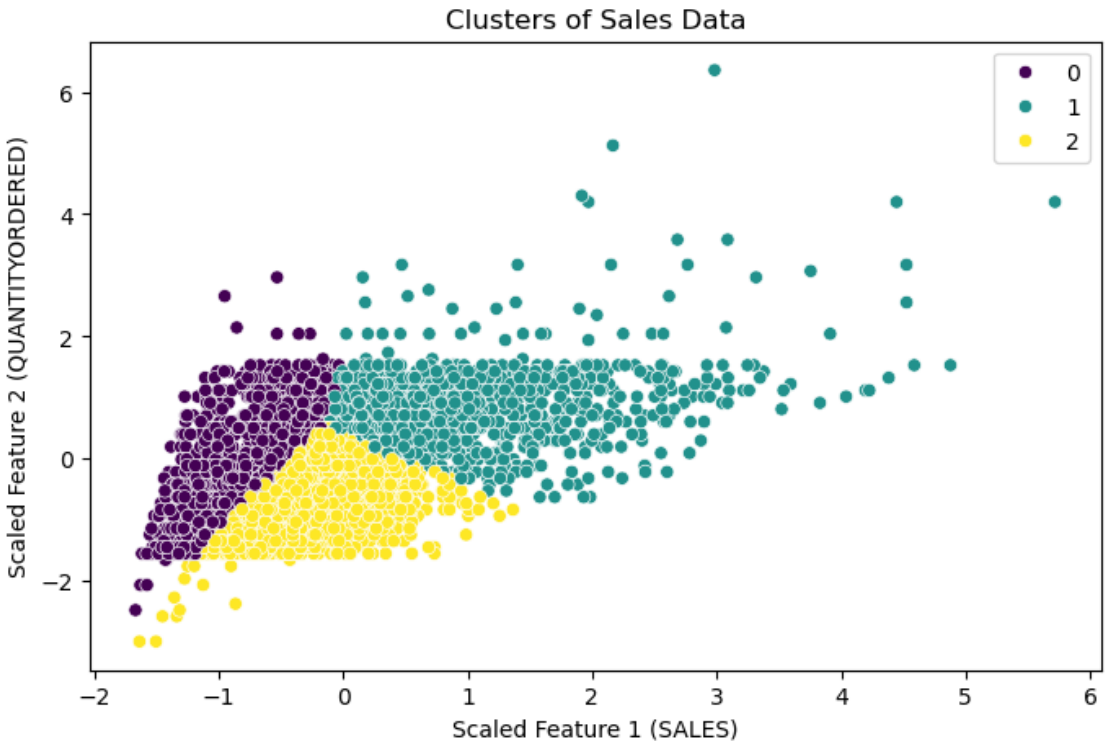
[72]:

	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	
…	…	…	…	…	…	
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  … ADDRESSLINE2  \
0       2/24/2003 0:00   Shipped       1         2     2003  …           NaN
1     05-07-2003 00:00   Shipped       2         5     2003  …           NaN
2     07-01-2003 00:00   Shipped       3         7     2003  …           NaN
3       8/25/2003 0:00   Shipped       3         8     2003  …           NaN
4     10-10-2003 00:00   Shipped       4        10     2003  …           NaN
…                  …       …     …       …       …  …          …
2818  12-02-2004 00:00   Shipped       4        12     2004  …           NaN
2819    1/31/2005 0:00   Shipped       1         1     2005  …           NaN
2820  03-01-2005 00:00  Resolved       1         3     2005  …           NaN
2821    3/28/2005 0:00   Shipped       1         3     2005  …           NaN
2822  05-06-2005 00:00   On Hold       2         5     2005  …           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
…               …   …         …       …       …              …
2818         Madrid   NaN      28034    Spain      EMEA          Freyre
2819           Oulu   NaN      90110  Finland      EMEA       Koskitalo
2820         Madrid   NaN      28034    Spain      EMEA          Freyre
2821       Toulouse   NaN      31000   France      EMEA          Roulet
2822         Boston    MA      51003      USA       NaN         Yoshido

     CONTACTFIRSTNAME DEALSIZE Cluster
0                Kwai    Small       2
1                Paul    Small       2
2              Daniel   Medium       1
3               Julie   Medium       1
4               Julie   Medium       1
…                   …       …     …
2818            Diego    Small       2
2819           Pirkko   Medium       2
2820            Diego   Medium       1
2821          Annette    Small       0
2822             Juri   Medium       0

[2823 rows x 26 columns]
```

[73]:
```
plt.figure(figsize=(8, 5))
sns.scatterplot(x=data_scaled[:, 0], y=data_scaled[:, 1], hue=kmeans.labels_,␣
 ↪palette='viridis')
plt.title('Clusters of Sales Data')
```

```
plt.xlabel('Scaled Feature 1 (SALES)')
plt.ylabel('Scaled Feature 2 (QUANTITYORDERED)')
plt.show()
```



[ ]: