PRACTICAL NO.:06

Name: Imam Mahamad Shaikh

Batch:C2

Roll no.:B23025

Problem Statement:

Implement K-Means clustering/ hierarchical clustering on sales_data_sample.csv dataset. Determine the number of clusters using the elbow method.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df = pd.read_csv("sales_data_sample.csv",encoding='latin1')

df
```

```
      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       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  ...
...                ...       ...     ...       ...      ...  ...
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
\
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]

df.head()

   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]
```

```
df.tail()
```

```
      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]
```

```
df.describe()
```

```
        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
```

```
df.dtypes
```

```
ORDERNUMBER           int64
QUANTITYORDERED       int64
PRICEEACH           float64
ORDERLINENUMBER       int64
SALES               float64
ORDERDATE            object
STATUS               object
QTR_ID                int64
MONTH_ID              int64
YEAR_ID               int64
PRODUCTLINE          object
```

```
MSRP                int64
PRODUCTCODE        object
CUSTOMERNAME       object
PHONE              object
ADDRESSLINE1       object
ADDRESSLINE2       object
CITY               object
STATE              object
POSTALCODE         object
COUNTRY            object
TERRITORY          object
CONTACTLASTNAME    object
CONTACTFIRSTNAME   object
DEALSIZE           object
dtype: object
```

```
df.isnull()
```

```
      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]

df.isnull().sum()

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
```

```
df2 =
df[['ORDERNUMBER','QUANTITYORDERED','ORDERLINENUMBER','QTR_ID','MONTH_
ID','YEAR_ID','MSRP']]
```

```
df2
```

```
      ORDERNUMBER  QUANTITYORDERED  ORDERLINENUMBER  QTR_ID  MONTH_ID
\
0           10107               30                2       1         2

1           10121               34                5       2         5

2           10134               41                2       3         7

3           10145               45                6       3         8

4           10159               49               14       4        10

...           ...              ...              ...     ...       ...

2818        10350               20               15       4        12

2819        10373               29                1       1         1

2820        10386               43                4       1         3

2821        10397               34                1       1         3
```

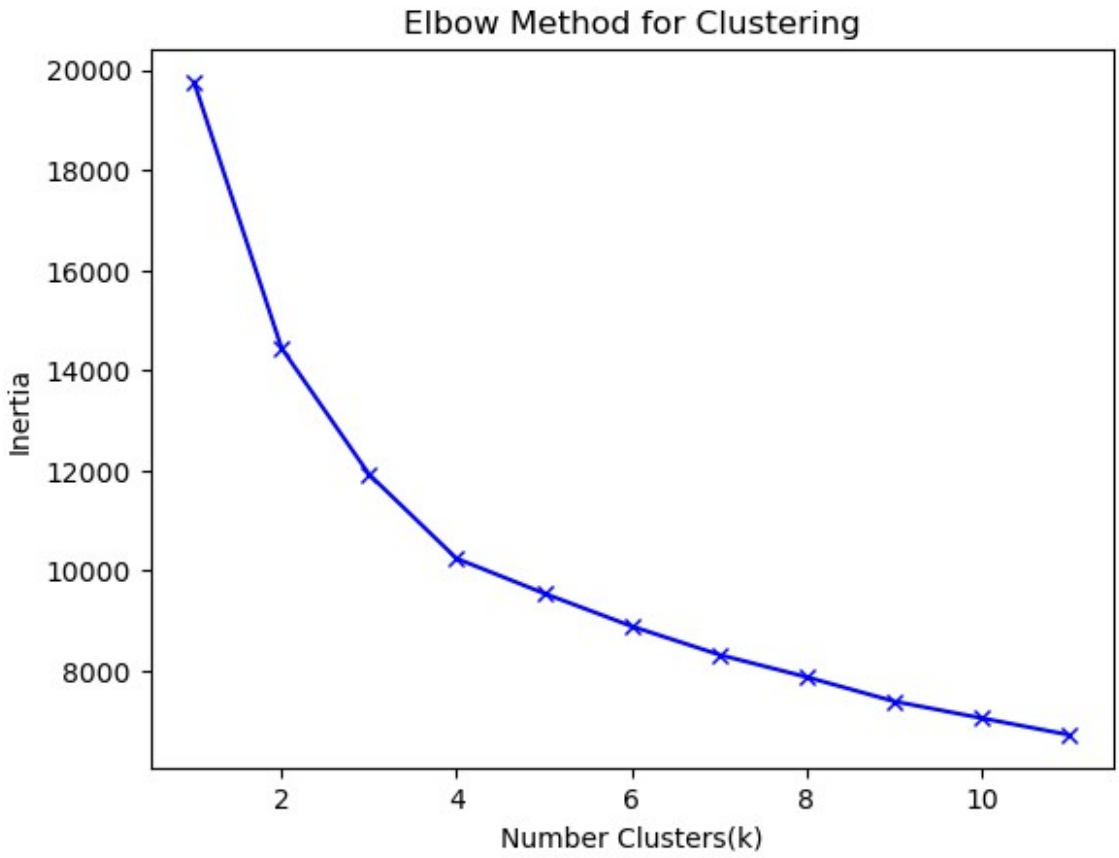
```
2822        10414                47                 9       2          5


      YEAR_ID  MSRP
0        2003    95
1        2003    95
2        2003    95
3        2003    95
4        2003    95
...       ...   ...
2818     2004    54
2819     2005    54
2820     2005    54
2821     2005    54
2822     2005    54

[2823 rows x 7 columns]

df2.dropna()

      ORDERNUMBER  QUANTITYORDERED  ORDERLINENUMBER  QTR_ID  MONTH_ID
\
0           10107               30                2       1         2

1           10121               34                5       2         5

2           10134               41                2       3         7

3           10145               45                6       3         8

4           10159               49               14       4        10

...           ...              ...              ...     ...       ...

2818        10350               20               15       4        12

2819        10373               29                1       1         1

2820        10386               43                4       1         3

2821        10397               34                1       1         3

2822        10414               47                9       2         5


      YEAR_ID  MSRP
0        2003    95
1        2003    95
2        2003    95
3        2003    95
4        2003    95
```

```
...       ...   ...
2818     2004    54
2819     2005    54
2820     2005    54
2821     2005    54
2822     2005    54

[2823 rows x 7 columns]
```

```
df2.isnull().sum()
```

```
ORDERNUMBER         0
QUANTITYORDERED     0
ORDERLINENUMBER     0
QTR_ID              0
MONTH_ID            0
YEAR_ID             0
MSRP                0
dtype: int64
```

```
from sklearn.preprocessing import StandardScaler

scaling = StandardScaler()

scaled_data = scaling.fit_transform(df2)

from sklearn.cluster import KMeans

distortions=[]
k_range =range(1,12)
for n in k_range:
    kmean= KMeans(n_clusters=n, random_state=45, n_init=11)
    kmean.fit(scaled_data)
    distortions.append(kmean.inertia_)

plt.plot(k_range, distortions, 'bx-')
plt.title("Elbow Method for Clustering")
plt.xlabel("Number Clusters(k)")
plt.ylabel("Inertia")

Text(0, 0.5, 'Inertia')
```


Elbow Method for Clustering
Inertia
20000
18000
16000
14000
12000
10000
8000
2
4
6
8
10
Number Clusters(k)