*#Name : Anuprita Girme #Roll No: COBA56*

import pandas as pd

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

data = pd.read\_csv('sales\_data\_sample.csv', sep = ',', encoding = 'Latin-1')

data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 |  |  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | ORDERDATE 2/24/2003 0:00 | | STATUS QTR\_ID MONTH\_ID YEAR\_ID  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

1. 10022 USA NaN Yu Kwai Small
2. 51100 France EMEA Henriot Paul Small
3. 75508 France EMEA Da Cunha Daniel Medium
4. 90003 USA NaN Young Julie Medium
5. 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]

selected\_features = data[['QUANTITYORDERED', 'PRICEEACH']] selected\_features

|  |  |  |
| --- | --- | --- |
|  | QUANTITYORDERED | PRICEEACH |
| 0 | 30 | 95.70 |
| 1 | 34 | 81.35 |
| 2 | 41 | 94.74 |
| 3 | 45 | 83.26 |
| 4 | 49 | 100.00 |
| ... | ... | ... |
| 2818 | 20 | 100.00 |
| 2819 | 29 | 100.00 |
| 2820 | 43 | 100.00 |
| 2821 | 34 | 62.24 |
| 2822 | 47 | 65.52 |

[2823 rows x 2 columns]

from sklearn.preprocessing import StandardScaler scaler = StandardScaler()

normalized\_features = scaler.fit\_transform(selected\_features) normalized\_features

|  |  |
| --- | --- |
| array([[-0.52289086, | 0.5969775 ], |
| [-0.11220131, | -0.11445035], |
| [ 0.60650538, | 0.54938372], |

...,

[ 0.81185016, 0.81015797],

[-0.11220131, -1.06186404],

[ 1.2225397 , -0.89925195]])

wcss = []

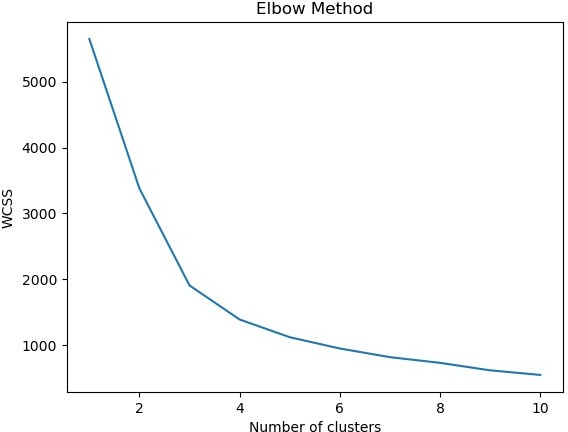
for i in range(1, 11):

kmeans = KMeans(n\_clusters=i, init='k-means++', max\_iter=300, n\_init=10, random\_state=0)

kmeans.fit(normalized\_features) wcss.append(kmeans.inertia\_)

plt.plot(range(1, 11), wcss) plt.title('Elbow Method') plt.xlabel('Number of clusters') plt.ylabel('WCSS')

plt.show()



kmeans = KMeans(n\_clusters=optimal\_clusters, init='k-means++', max\_iter=300, n\_init=10, random\_state=0)

cluster\_labels = kmeans.fit\_predict(normalized\_features)

plt.scatter(normalized\_features[:, 0], normalized\_features[:, 1], c=cluster\_labels, cmap='viridis')

plt.xlabel('QUANTITYORDERED') plt.ylabel('PRICEEACH') plt.title('K-Means Clustering') plt.show()

