[1]:

**import pandas as pd import numpy as np**

[2]:

df = pd.read\_csv('./sales\_data\_sample.csv', encoding='unicode\_escape')

[3]:

df.head

1. : <bound method NDFrame.head of 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  … 2818 | 10159  … 10350 | … | | 49  20 | … | 100.00  100.00 | … | | … | | 14  15 | 5205.27  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  … 2818 | 10/10/2003  … 12/2/2004 | 0:00  0:00 | Shipped  … …  Shipped | 4  … 4 | | | 10  … … 12 | 2003  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  …  2818 |  |  |  | 7734 Strong St.  …  C/ Moralzarzal, 86 | | … | NaN  NaN | San | Francisco  … …  Madrid | CA  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]>

[4]:

df.info

1. : <bound method DataFrame.info of 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  …  2818 |  |  | 7734 Strong St.  …  C/ Moralzarzal, 86 | | | NaN  …  NaN | | San | Francisco  … …  Madrid | | CA  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]>

[5]:

*#Columns to Remove*

to\_drop = ['ADDRESSLINE1', 'ADDRESSLINE2', 'STATE', 'POSTALCODE', 'PHONE']

df = df.drop(to\_drop, axis=1)

[6]:

*#Check for null values*

df.isnull().sum()

[6]: 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 |
| CITY | 0 |
| COUNTRY | 0 |
| TERRITORY | 1074 |
| CONTACTLASTNAME | 0 |
| CONTACTFIRSTNAME | 0 |
| DEALSIZE | 0 |
| dtype: int64 |  |

[7]:

*# Basic commands*

[8]:

df.dtypes

[8]: 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

CITY object

COUNTRY object

TERRITORY object

CONTACTLASTNAME object CONTACTFIRSTNAME object

DEALSIZE object dtype: object

[9]:

*#ORDERDATE Should be in date time*

df['ORDERDATE'] = pd.to\_datetime(df['ORDERDATE'])

[10]:

*#We need to create some features in order to create cluseters*

*#Recency: Number of days between customer's latest order and today's date #Frequency : Number of purchases by the customers*

*#MonetaryValue : Revenue generated by the customers*

**import datetime as dt**

snapshot\_date = df['ORDERDATE'].max() + dt.timedelta(days = 1) df\_RFM = df.groupby(['CUSTOMERNAME']).agg({

'ORDERDATE' : **lambda** x : (snapshot\_date - x.max()).days, 'ORDERNUMBER' : 'count',

'SALES' : 'sum'

})

*#Rename the columns*

df\_RFM.rename(columns = { 'ORDERDATE' : 'Recency',

'ORDERNUMBER' : 'Frequency', 'SALES' : 'MonetaryValue'

}, inplace=**True**)

[11]:

df\_RFM.head()

[11]: Recency Frequency MonetaryValue CUSTOMERNAME

AV Stores, Co. 196 51 157807.81

Alpha Cognac 65 20 70488.44

Amica Models & Co. 265 26 94117.26

Anna's Decorations, Ltd 84 46 153996.13

Atelier graphique 188 7 24179.96

[12]:

*# Divide into segments*

*# We create 4 quartile ranges*

df\_RFM['M'] = pd.qcut(df\_RFM['MonetaryValue'], q = 4, labels = range(1,5)) df\_RFM['R'] = pd.qcut(df\_RFM['Recency'], q = 4, labels = list(range(4,0,-1))) df\_RFM['F'] = pd.qcut(df\_RFM['Frequency'], q = 4, labels = range(1,5))

df\_RFM.head()

[12]:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Recency Frequency MonetaryValue | | | | | M | R | F |
| CUSTOMERNAME  AV Stores, Co. |  | 196 | 51 | 157807.81 | 4 | 2 | 4 |
| Alpha Cognac |  | 65 | 20 | 70488.44 | 2 | 4 | 2 |
| Amica Models & | Co. | 265 | 26 | 94117.26 | 3 | 1 | 2 |
| Anna's Decorations, Ltd | | 84 | 46 | 153996.13 | 4 | 3 | 4 |
| Atelier graphique | | 188 | 7 | 24179.96 | 1 | 2 | 1 |

[13]:

*#Create another column for RFM score*

df\_RFM['RFM\_Score'] = df\_RFM[['R', 'M', 'F']].sum(axis=1)

df\_RFM.head()

[13]:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Recency Frequency MonetaryValue | | | | M | R | F | RFM\_Score |
| CUSTOMERNAME |  |  |  |  |  |  |  |
| AV Stores, Co. | 196 | 51 | 157807.81 | 4 | 2 | 4 | 10 |
| Alpha Cognac | 65 | 20 | 70488.44 | 2 | 4 | 2 | 8 |
| Amica Models & Co. | 265 | 26 | 94117.26 | 3 | 1 | 2 | 6 |
| Anna's Decorations, Ltd | 84 | 46 | 153996.13 | 4 | 3 | 4 | 11 |
| Atelier graphique | 188 | 7 | 24179.96 | 1 | 2 | 1 | 4 |

[14]:

**def** rfm\_level(df):

**if** bool(df['RFM\_Score'] >= 10):

**return** 'High Value Customer'

**elif** bool(df['RFM\_Score'] < 10) **and** bool(df['RFM\_Score'] >= 6):

**return** 'Mid Value Customer'

**else**:

**return** 'Low Value Customer'

df\_RFM['RFM\_Level'] = df\_RFM.apply(rfm\_level, axis = 1) df\_RFM.head()

[14]:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Recency Frequency MonetaryValue | | | | | M | R | F | \ |
| CUSTOMERNAME  AV Stores, Co. |  | 196 | 51 | 157807.81 | 4 | 2 | 4 | |
| Alpha Cognac |  | 65 | 20 | 70488.44 | 2 | 4 | 2 | |
| Amica Models & Co. |  | 265 | 26 | 94117.26 | 3 | 1 | 2 | |
| Anna's Decorations, | Ltd | 84 | 46 | 153996.13 | 4 | 3 | 4 | |
| Atelier graphique |  | 188 | 7 | 24179.96 | 1 | 2 | 1 | |
| CUSTOMERNAME |  | RFM\_Score | RFM\_Level | | | | | |
| AV Stores, Co. |  | 10 | High Value Customer | | | | | |
| Alpha Cognac |  | 8 | Mid Value Customer | | | | | |
| Amica Models & Co. |  | 6 | Mid Value Customer | | | | | |
| Anna's Decorations, | Ltd | 11 | High Value Customer | | | | | |
| Atelier graphique |  | 4 | Low Value Customer | | | | | |

[15]:

*# Time to perform KMeans*

data = df\_RFM[['Recency', 'Frequency', 'MonetaryValue']] data.head()

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| [15]: | CUSTOMERNAME | Recency | Frequency | MonetaryValue |
|  | AV Stores, Co. | 196 | 51 | 157807.81 |
|  | Alpha Cognac | 65 | 20 | 70488.44 |
|  | Amica Models & Co. | 265 | 26 | 94117.26 |
|  | Anna's Decorations, Ltd | 84 | 46 | 153996.13 |

Atelier graphique 188 7 24179.96

[16]:

*# Our data is skewed we must remove it by performing log transformation*

data\_log = np.log(data) data\_log.head()

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| [16]: | CUSTOMERNAME | Recency | Frequency | MonetaryValue |
|  | AV Stores, Co. | 5.278115 | 3.931826 | 11.969133 |
|  | Alpha Cognac | 4.174387 | 2.995732 | 11.163204 |
|  | Amica Models & Co. | 5.579730 | 3.258097 | 11.452297 |
|  | Anna's Decorations, Ltd | 4.430817 | 3.828641 | 11.944683 |
|  | Atelier graphique | 5.236442 | 1.945910 | 10.093279 |

[17]:

*#Standardization*

**from sklearn.preprocessing import** StandardScaler scaler = StandardScaler()

scaler.fit(data\_log)

data\_normalized = scaler.transform(data\_log)

data\_normalized = pd.DataFrame(data\_normalized, index = data\_log.index,␣

↪columns=data\_log.columns)

data\_normalized.describe().round(2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| [17]: |  | Recency | Frequency | MonetaryValue |
|  | count | 92.00 | 92.00 | 92.00 |
|  | mean | 0.00 | -0.00 | 0.00 |
|  | std | 1.01 | 1.01 | 1.01 |
|  | min | -3.51 | -3.67 | -3.82 |
|  | 25% | -0.24 | -0.41 | -0.39 |
|  | 50% | 0.37 | 0.06 | -0.04 |
|  | 75% | 0.53 | 0.45 | 0.52 |
|  | max | 1.12 | 4.03 | 3.92 |

[18]:

*#Fit KMeans and use elbow method to choose the number of clusters*

**import matplotlib.pyplot as plt import seaborn as sns**

**from sklearn.cluster import** KMeans

sse = {}

**for** k **in** range(1, 21):

kmeans = KMeans(n\_clusters = k, random\_state = 1) kmeans.fit(data\_normalized)

sse[k] = kmeans.inertia\_

[19]:

plt.figure(figsize=(10,6)) plt.title('The Elbow Method')

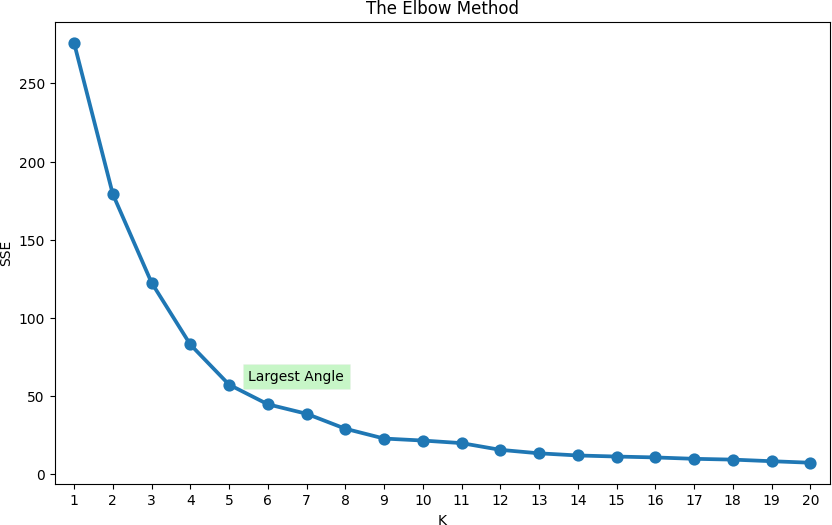
plt.xlabel('K') plt.ylabel('SSE') plt.style.use('ggplot')

sns.pointplot(x=list(sse.keys()), y = list(sse.values()))

plt.text(4.5, 60, "Largest Angle", bbox = dict(facecolor = 'lightgreen', alpha␣

↪= 0.5))

plt.show()



[20]:

*# 5 number of clusters seems good*

kmeans = KMeans(n\_clusters=5, random\_state=1) kmeans.fit(data\_normalized)

cluster\_labels = kmeans.labels\_

data\_rfm = data.assign(Cluster = cluster\_labels) data\_rfm.head()

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| [20]: | CUSTOMERNAME | Recency | Frequency | MonetaryValue | Cluster |
|  | AV Stores, Co. | 196 | 51 | 157807.81 | 4 |
|  | Alpha Cognac | 65 | 20 | 70488.44 | 2 |
|  | Amica Models & Co. | 265 | 26 | 94117.26 | 2 |
|  | Anna's Decorations, Ltd | 84 | 46 | 153996.13 | 4 |
|  | Atelier graphique | 188 | 7 | 24179.96 | 1 |