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
                                                  95.70
                                                                            2871.00
      0
                   10107
                                         30
      1
                                         34
                                                  81.35
                                                                        5
                                                                            2765.90
                   10121
      2
                                                                         2
                                                                            3884.34
                   10134
                                         41
                                                  94.74
      3
                   10145
                                         45
                                                  83.26
                                                                        6
                                                                            3746.70
      4
                                                                            5205.27
                   10159
                                         49
                                                 100.00
                                                                        14
      2818
                   10350
                                         20
                                                 100.00
                                                                       15
                                                                            2244.40
      2819
                   10373
                                         29
                                                 100.00
                                                                         1
                                                                            3978.51
      2820
                   10386
                                         43
                                                 100.00
                                                                            5417.57
      2821
                   10397
                                         34
                                                  62.24
                                                                         1
                                                                            2116.16
      2822
                                         47
                                                  65.52
                                                                            3079.44
                   10414
                                  STATUS
                                           QTR ID
                                                    MONTH ID
                                                               YEAR ID
                    ORDERDATE
      0
                                 Shipped
               2/24/2003 0:00
                                                 1
                                                           2
                                                                  2003
                                                2
      1
             05-07-2003 00:00
                                 Shipped
                                                           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
                                                                  2005
               1/31/2005 0:00
                                 Shipped
                                                1
                                                           1
      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
```

		AI	DDRESSLINE1	ADDRESSLINE2	CITY	STATE	\	
0	897	Long Air	ort Avenue	naN	NaN NYC			
1		59 rue d	de l'Abbaye	e NaN	NaN Reims			
2	27 rue du	Colonel H	Pierre Avia	. NaN	Paris	NaN		
3		78934 H	illside Dr.	NaN	Pasadena	CA		
4		7734	Strong St.	NaN	San Francisco	CA		
•••				***	***			
2818	;	C/ Moral	lzarzal, 86	NaN	Madrid	NaN		
2819	ı		Γorikatu 38	NaN	Oulu	NaN		
2820	ı	C/ Moral	lzarzal, 86	NaN	Madrid	NaN		
2821	1		ce-Lorraine		Toulouse	NaN		
2822		8616 Sp:	innaker Dr.	NaN	Boston	MA		
		•						
	POSTALCODE	COUNTRY	TERRITORY	CONTACTLASTNAME	CONTACTFIRSTN	AME DE	ALSIZE	
0	10022	USA	NaN	Yu	K	wai	Small	
1								
1	51100	France	EMEA	Henriot	Pa	aul	Small	
2	51100 75508	France France	EMEA EMEA	Henriot Da Cunha	Pa Dan		Small Medium	
					Dan	iel 1		
2	75508	France	EMEA	Da Cunha	Dan Ju	iel l lie l	Medium	
2 3	75508 90003	France USA	EMEA NaN	Da Cunha Young	Dan Ju	iel l lie l	Medium Medium	
2 3 4	75508 90003 NaN 	France USA	EMEA NaN	Da Cunha Young Brown	Dan Ju Ju 	iel l lie l	Medium Medium	
2 3 4 	75508 90003 NaN 28034	France USA USA 	EMEA NaN NaN 	Da Cunha Young Brown 	Dan Ju Ju 	iel I lie I lie I	Medium Medium Medium	
2 3 4 2818	75508 90003 NaN 28034 90110	France USA USA Spain	EMEA NaN NaN EMEA	Da Cunha Young Brown Freyre	Dan Ju Ju Di Pir	iel I lie I lie I ego kko I	Medium Medium Medium Small	
2 3 4 2818 2819	75508 90003 NaN 28034 90110 28034	France USA USA Spain Finland	EMEA NaN NaN EMEA EMEA	Da Cunha Young Brown Freyre Koskitalo	Dan Ju Ju Di Pir	iel I lie I lie I ego kko I ego I	Medium Medium Medium Small Medium	
2 3 4 2818 2819 2820	75508 90003 NaN 28034 90110 28034 31000	France USA USA Spain Finland Spain	EMEA NaN NaN EMEA EMEA EMEA	Da Cunha Young Brown Freyre Koskitalo Freyre	Dan Ju Ju Die Pir Die Anne	iel I lie I lie I ego kko I ego I tte	Medium Medium Medium Small Medium 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

```
object
      12 PRODUCTCODE
                             2823 non-null
      13
          CUSTOMERNAME
                             2823 non-null
                                              object
                                              object
      14
          PHONE
                             2823 non-null
      15
          ADDRESSLINE1
                             2823 non-null
                                              object
      16
                                              object
          ADDRESSLINE2
                             302 non-null
          CITY
      17
                             2823 non-null
                                              object
      18
          STATE
                             1337 non-null
                                              object
      19 POSTALCODE
                             2747 non-null
                                              object
          COUNTRY
                             2823 non-null
                                              object
                             1749 non-null
      21
          TERRITORY
                                              object
      22
          CONTACTLASTNAME
                             2823 non-null
                                              object
          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
                              0
      QUANTITYORDERED
      PRICEEACH
                              0
      ORDERLINENUMBER
                              0
                              0
      SALES
                              0
      ORDERDATE
                              0
      STATUS
      QTR_ID
                              0
                              0
      MONTH_ID
                              0
      YEAR_ID
      PRODUCTLINE
                              0
                              0
      MSRP
                              0
      PRODUCTCODE
      CUSTOMERNAME
                              0
      PHONE
                              0
      ADDRESSLINE1
                              0
                           2521
      ADDRESSLINE2
      CITY
                              0
                           1486
      STATE
      POSTALCODE
                             76
      COUNTRY
                              0
      TERRITORY
                           1074
      CONTACTLASTNAME
                              0
                              0
      CONTACTFIRSTNAME
```

[55]: df

DEALSIZE

dtype: int64

0

[55]:		ORDERNUMBER	R QUANT	ITYORDEREI	PRICE	EEACH OF	RDERLINENU	MBER	SAL	ES	\
	0	10107	7	30) 9	5.70		2	2871.	00	
	1	10121	[34	1 8	31.35		5	2765.	90	
	2	10134	1	41	. 9	94.74		2	3884.	34	
	3	10145	5	45	5 8	3.26		6	3746.	70	
	4	10159	9	49) 10	00.00		14	5205.	27	
				•••	•••		•••	•••			
	2818	10350)	20) 10	00.00		15	2244.	40	
	2819	10373	3	29	10	00.00		1	3978.	51	
	2820	10386	3	43	3 10	00.00		4	5417.	57	
	2821	10397	7	34	1 6	32.24		1	2116.		
	2822	10414	l	47	7 6	55.52		9	3079.	44	
		0000	70 D A MIN	GEL A EULO	000 10	MONETI T			,		
	0		ERDATE	STATUS	QTR_ID	MONTH_I			\		
	0	2/24/2003		Shipped	1		2 200				
	1	05-07-2003		Shipped	2		5 200				
	2	07-01-2003		Shipped	3		7 200				
	3	8/25/2003		Shipped	3		8 200				
	4	10-10-2003	00:00	Shipped	4	1	.0 200	3			
		10 00 0001									
	2818	12-02-2004		Shipped	4	1	.2 200				
	2819	1/31/2005		Shipped	1		1 200				
	2820	03-01-2005		Resolved	1		3 200				
	2821	3/28/2005		Shipped	1		3 200				
	2822	05-06-2005	00:00	On Hold	2		5 200	b			
			AI	DDRESSLINE	E1 ADDF	RESSLINE2	2	CITY	STAT	E.	\
	0	897 I	Long Air	port Avenu	ıe	NaN	I	NYC	C N	Υ	
	1		-	de l'Abbay		NaN	I	Reims	s Na	ιN	
	2	27 rue du 0	Colonel I	Pierre Avi	la	NaN	Ī	Paris	s Na	ιN	
	3		78934 H:	illside Dr		NaN	I Pa	sadena	a C	:A	
	4		7734	Strong St	5.	NaN	I San Fra	ncisco) C	:A	
	•••			•••		•••	•••	•••			
	2818		C/ Mora	lzarzal, 8	36	NaN	Ī	Madrid	l Na	ιN	
	2819			Torikatu 3	38	NaN	I	Oulu	ı Na	ιN	
	2820		C/ Mora	lzarzal, 8	36	NaN	I	Madrid	l Na	ιN	
	2821	1 1	rue Alsa	ce-Lorrain	ıe	NaN	I To	ulouse	e Na	ιN	
	2822		8616 Sp:	innaker Dr	· .	NaN	I	Bostor	ı M	ΙA	
		DOGTAL CODE	COUNTRY	TEDDITTODI	Z CONTAC	יידו אפידיאו או	TE CONTRACT	ידים מידיו	IAME E	TAT	O T O D
		POSTALCODE 10022	USA	TERRITORY							
	0	51100		NaN EME <i>A</i>		Henric	u .+		Kwai Paul		mall
	1 2	75508	France France	EME <i>F</i> EME		Da Cunh			niel		mall dium
	3	90003	USA			Your			ılie		dium
	3 4	NaN	USA	NaN Na		Brow	•		ılie ılie		dium
		IValv	AGU	NaN		DIOM	111	Jl	TTTE	rie	u i ulli
	 2010	 28034	 Cnoin	 EME /		Emarra			ogo	Ç.	mall
	2818	20034	Spain	EME <i>A</i>	1	Freyr	. 🖯	נע	Lego	D	mall

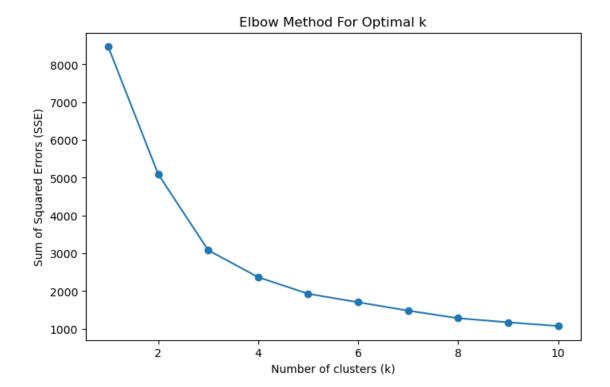
2819	90110	Finland	EMEA	Koskitalo	Pirkko	Medium
2820	28034	Spain	EMEA	Freyre	Diego	Medium
2821	31000	France	EMEA	Roulet	Annette	Small
2822	51003	USA	NaN	Yoshido	Juri	Medium

[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
                                           94.74
            3884.34
                                  41
      3
            3746.70
                                  45
                                           83.26
      4
            5205.27
                                  49
                                          100.00
      2818 2244.40
                                  20
                                         100.00
                                  29
                                          100.00
      2819 3978.51
      2820 5417.57
                                  43
                                          100.00
      2821 2116.16
                                  34
                                          62.24
      2822 3079.44
                                           65.52
                                  47
      [2823 rows x 3 columns]
[60]: data.isnull().sum()
[60]: SALES
                         0
      QUANTITYORDERED
                         0
      PRICEEACH
                         0
      dtype: int64
          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 Determining 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()
```

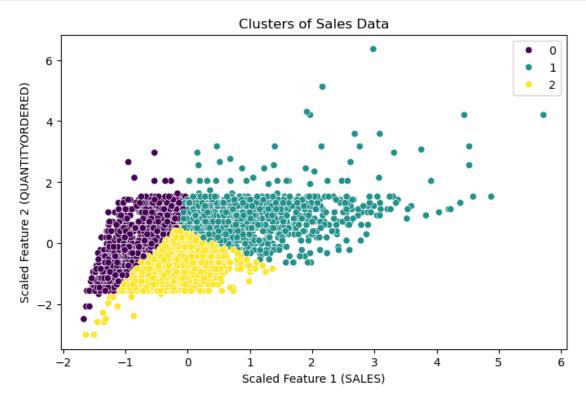


0.6 Training the Model

```
[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)
      df['Cluster'] = kmeans.labels_
[71]:
     df
[72]:
[72]:
            ORDERNUMBER
                          QUANTITYORDERED
                                             PRICEEACH
                                                         ORDERLINENUMBER
                                                                             SALES
                   10107
                                                 95.70
                                                                           2871.00
      0
                                         30
                                                                        2
      1
                   10121
                                         34
                                                 81.35
                                                                        5
                                                                           2765.90
      2
                                         41
                                                 94.74
                                                                        2
                                                                           3884.34
                   10134
      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
                                                                           3978.51
                                                                        1
      2820
                   10386
                                         43
                                                100.00
                                                                           5417.57
      2821
                   10397
                                         34
                                                 62.24
                                                                           2116.16
```

:	2822	10414		47	6	5.52			9 30	79.44	
		ORDERDAT	E STA	TUS QT	TR_ID	MONT	H_ID	YEAR_ID	ADDI	RESSLINE2	\
(0	2/24/2003 0:0	0 Ship	ped	1		2	2003	•••	NaN	
	1	05-07-2003 00:0	0 Ship	ped	2		5	2003	•••	NaN	
	2	07-01-2003 00:0	00 Ship	ped	3		7	2003	•••	NaN	
;	3	8/25/2003 0:0	00 Ship	ped	3		8	2003		NaN	
	4	10-10-2003 00:0	00 Ship	_	4		10	2003	•••	NaN	
	 2818	12-02-2004 00:0	 00 Ship	 ped	4		12	2004	•••	NaN	
:	2819	1/31/2005 0:0	00 Ship	ped	1		1	2005		NaN	
:	2820	03-01-2005 00:0	00 Resol	ved	1		3	2005	•••	NaN	
:	2821	3/28/2005 0:0	0 Ship	ped	1		3	2005	•••	NaN	
:	2822	05-06-2005 00:0	00 On H	old	2		5	2005	•••	NaN	
		CITY S	TATE POS	TALCODE	E COU	NTRY	TERRIT	ORY CONT	'ACTLAS'	TNAME \	
(0	NYC	NY	10022	2	USA		NaN		Yu	
	1	Reims	NaN	51100) Fr	ance	E	EMEA	He	nriot	
•	2	Paris	NaN	75508	3 Fr	ance	E	EMEA	Da (Cunha	
;	3	Pasadena	CA	90003	3	USA		NaN	,	Young	
•	4	San Francisco	CA	NaN	J	USA		NaN	1	Brown	
						····	-		-		
	2818	Madrid	NaN	28034		pain		EMEA		reyre	
	2819	Oulu	NaN	90110		land		EMEA		italo	
	2820	Madrid	NaN	28034		pain		EMEA		reyre	
	2821	Toulouse	NaN	31000		ance		EMEA		oulet	
	2822	Boston	MA	51003	3	USA		NaN	Yo	shido	
	(CONTACTFIRSTNAME	DEALSIZ	E Clust	er						
(0	Kwai	Smal	1	2						
	1	Paul	Smal	1	2						
:	2	Daniel	Mediu	m	1						
;	3	Julie	Mediu	m	1						
•	4	Julie	Mediu	m	1						
	 2818	 Diego	 Smal	 1	2						
:	2819	Pirkko		m	2						
	2820	Diego			1						
	2821	Annette			0						
	2822	Juri			0						
	[2823	rows x 26 colum	ns]								
[73]:	plt.f:	igure(figsize=(8	3, 5))								
	sns.so	catterplot(x=dat ette='viridis')		[: , 0],	, y=da	ta_sc	caled[:	, 1], hu	ie=kmea:	ns.labels_	_,⊔
	plt.t:	itle('Clusters o	of Sales	Data')							

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



[]: