

10/8/23, 10:32 PMml5

In [1]:

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

In [3]:

```
df = pd.read_csv("C:\\Users\\Sujeet\\OneDrive\\Desktop\\sales_data_sample.csv")
df.head()
```

Out[3]:

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE
0	10107	30	95.70	2	2871.00	2/24/2003 0:00
1	10121	34	81.35	5	2765.90	05-07-2003 00:00
2	10134	41	94.74	2	3884.34	07-01-2003 00:00
3	10145	45	83.26	6	3746.70	8/25/2003 0:00
4	10159	49	100.00	14	5205.27	10-10-2003 00:00

5 rows × 25 columns

In [4]:

```
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
```

```
In [5]: df.describe()
```

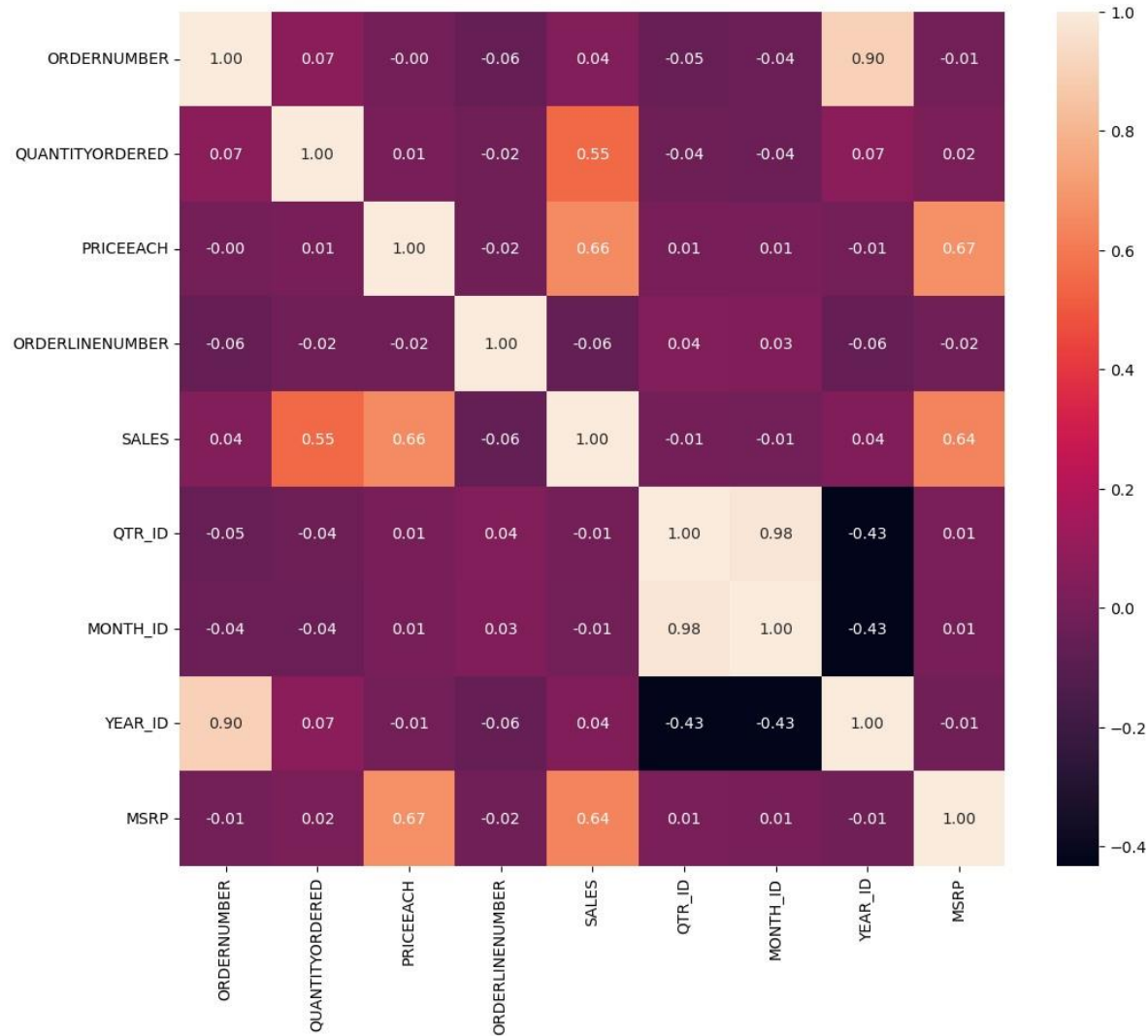
Out[5]:

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	
count	2823.000000	2823.000000	2823.000000	2823.000000	2823.000000	282
mean	10258.725115	35.092809	83.658544	6.466171	3553.889072	
std	92.085478	9.741443	20.174277	4.225841	1841.865106	
min	10100.000000	6.000000	26.880000	1.000000	482.130000	
25%	10180.000000	27.000000	68.860000	3.000000	2203.430000	
50%	10262.000000	35.000000	95.700000	6.000000	3184.800000	
75%	10333.500000	43.000000	100.000000	9.000000	4508.000000	
max	10425.000000	97.000000	100.000000	18.000000	14082.800000	

```
In [6]: fig = plt.figure(figsize=(12,10))
sns.heatmap(df.corr(), annot=True, fmt='.2f')
plt.show()
```

C:\Users\Ashish\AppData\Local\Temp\ipykernel\_20220\1537228670.py:2: FutureWarning: The default value of numeric\_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric\_only to silence this warning.

```
sns.heatmap(df.corr(), annot=True, fmt='.2f')
```



```
In [7]: df= df[['PRICEEACH', 'MSRP']]
df.head()
```

Out[7]:

	PRICEEACH	MSRP
0	95.70	95
1	81.35	95
2	94.74	95
3	83.26	95
4	100.00	95

```
In [8]: df.isna().any()
Out[8]: PRICEEACH    False
        MSRP        False
        dtype: bool
In [9]: df.describe().T
```

Out[9]:

	count	mean	std	min	25%	50%	75%	max
PRICEEACH	2823.0	83.658544	20.174277	26.88	68.86	95.7	100.0	100.0

MSRP	2823.0	100.715551	40.187912	33.00	68.00	99.0	124.0	214.0
------	--------	------------	-----------	-------	-------	------	-------	-------

In [10]: df.shape

Out[10]: (2823, 2)

```
from sklearn.cluster import KMeans

inertia = []

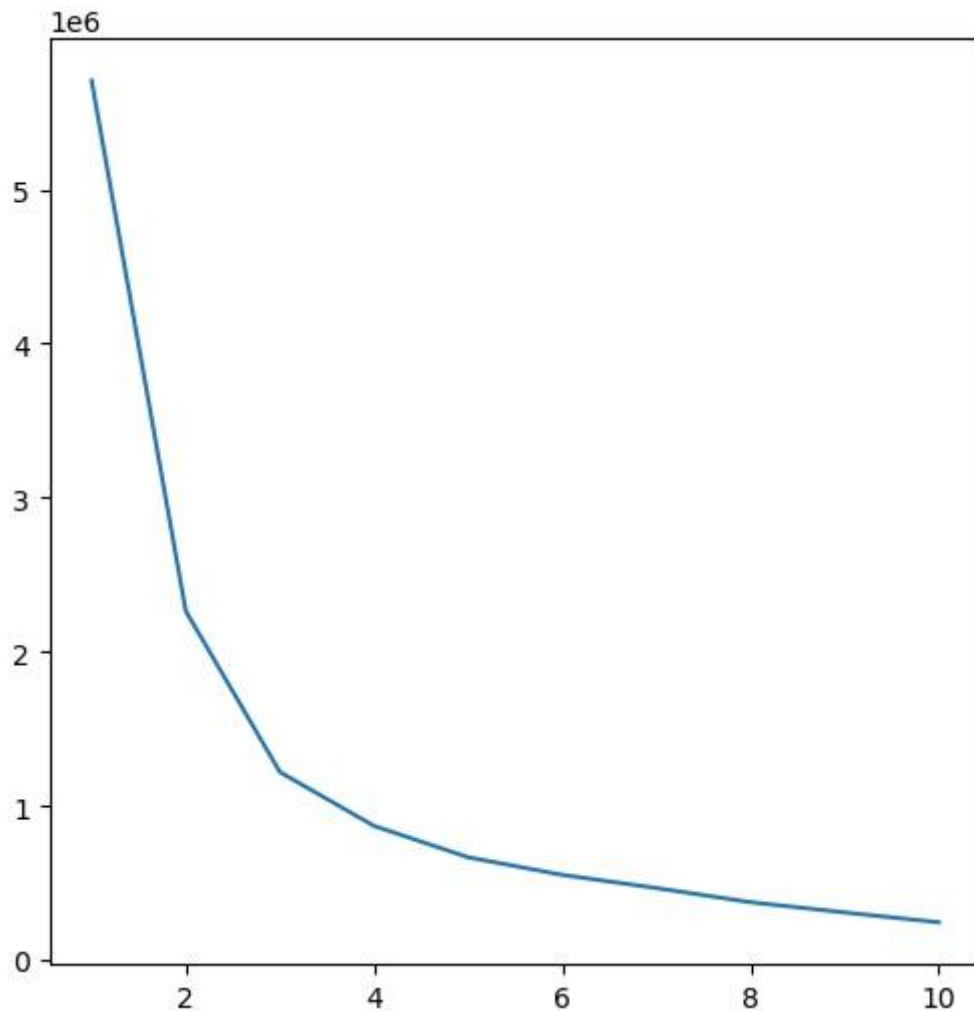
for i in range(1, 11):    clusters = KMeans(n_clusters=i, init='k-
means++', random_state=42)    clusters.fit(df)
    inertia.append(clusters.inertia_)

plt.figure(figsize=(6, 6)) sns.lineplot(x = [1, 2, 3, 4, 5,
6, 7, 8, 9, 10], y = inertia)
```

In [12]:

```
C:\Users\Ashish\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 10 to 'auto' in 1.4.
Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
<Axes: >
```

Out[12]:



```
In [13]: kmeans = KMeans(n_clusters = 3, random_state =
42) y_kmeans = kmeans.fit_predict(df) y_kmeans
```

C:\Users\Ashish\anaconda3\lib\site-packages\sklearn\cluster\\_kmeans.py:870: FutureWarning: The default value of `n\_init` will change from 10 to 'auto' in 1.4. Set the value of `n\_init` explicitly to suppress the warning warnings.warn(

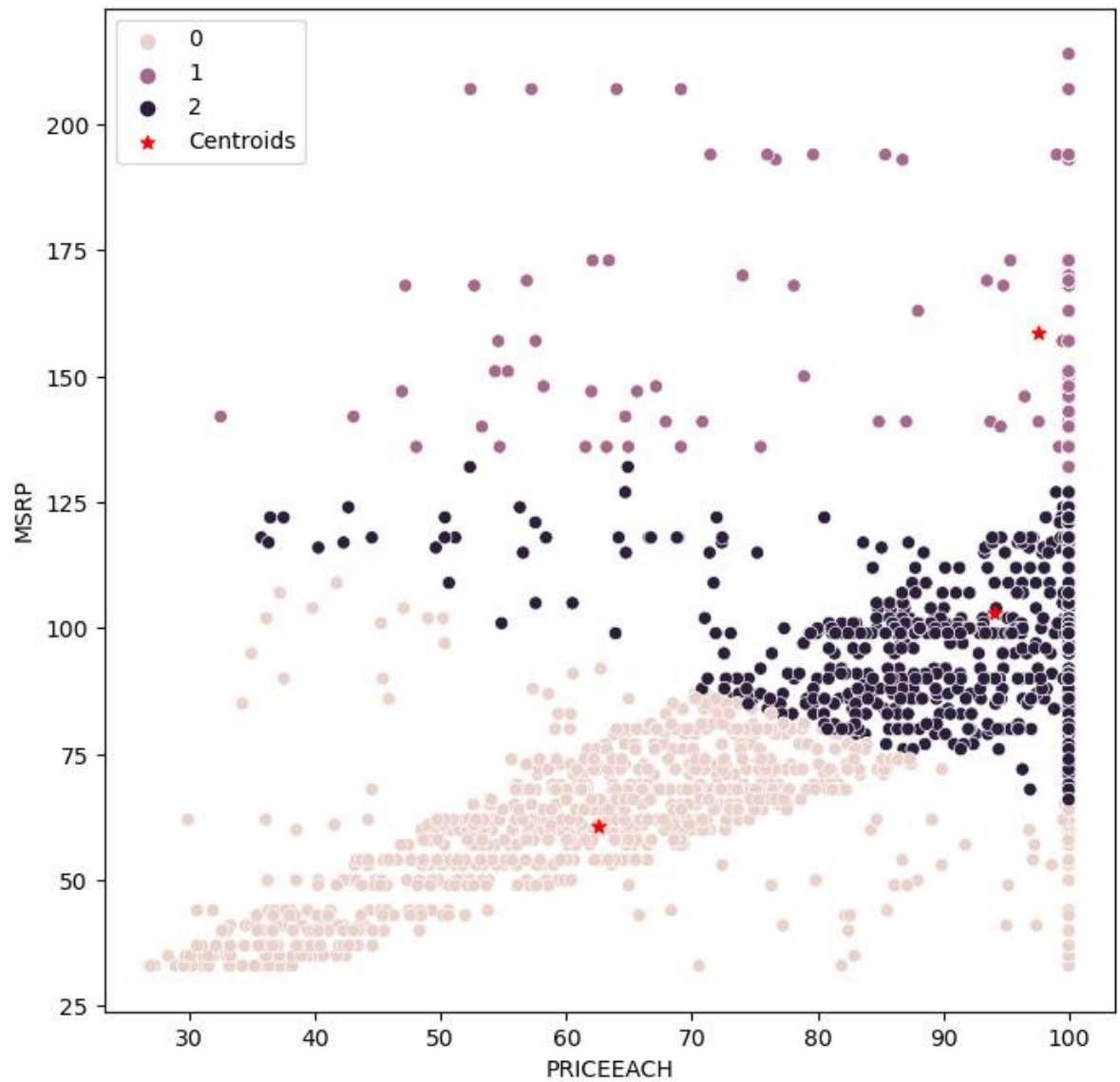
```
Out[13]: array([2, 2, 2, ..., 0, 0, 0])
```

```
plt.figure(figsize=(8,8)) sns.scatterplot(x=df['PRICEEACH'],
y=df['MSRP'], hue=y_kmeans)
plt.scatter(kmeans.cluster_centers_[0], kmeans.cluster_centers_[1], c =
'red') plt.legend()
```

```
In [14]:
```

```
<matplotlib.legend.Legend at 0x16cebb36f80>
```

```
Out[14]:
```



In [15]: `kmeans.cluster_centers_`

```
array([[ 62.49548902,
        60.71556886],
```

Out[15]:

```
[ 97.59890263, 158.7202473 ],
 [ 94.03841567, 102.88841567]])
```

In [ ]: