Project on Online Retail

K-Means Clustering

In [1]:
```
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [2]:
```
df=pd.read_csv(r"C:\Users\rubin\Documents\OnlineRetail.csv")
df.head()
```

Out[2]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	1/12/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	1/12/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	1/12/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	1/12/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	1/12/2010 8:26	3.39	17850.0	United Kingdom

In [3]:
```
df.head()
```

Out[3]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	1/12/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	1/12/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	1/12/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	1/12/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	1/12/2010 8:26	3.39	17850.0	United Kingdom

In [4]:
```
df.tail()
```

Out[4]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Cou
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	9/12/2011 12:50	0.85	12680.0	Fra
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	9/12/2011 12:50	2.10	12680.0	Fra
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	9/12/2011 12:50	4.15	12680.0	Fra
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	9/12/2011 12:50	4.15	12680.0	Fra
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	9/12/2011 12:50	4.95	12680.0	Fra

In [5]:
```
df.fillna(method='ffill',inplace=True)
```

In [6]:
```
df.isnull().sum()
```

Out[6]:
```
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
```

In [7]:
```
# There are no null values in the given dataset.
```

In [8]:
```
df.columns
```

Out[8]:
```
Index(['InvoiceNo', 'StockCode', 'Description', 'Quantity', 'InvoiceDate',
       'UnitPrice', 'CustomerID', 'Country'],
      dtype='object')
```

In [29]:
```
df['CustomerID'].value_counts
```

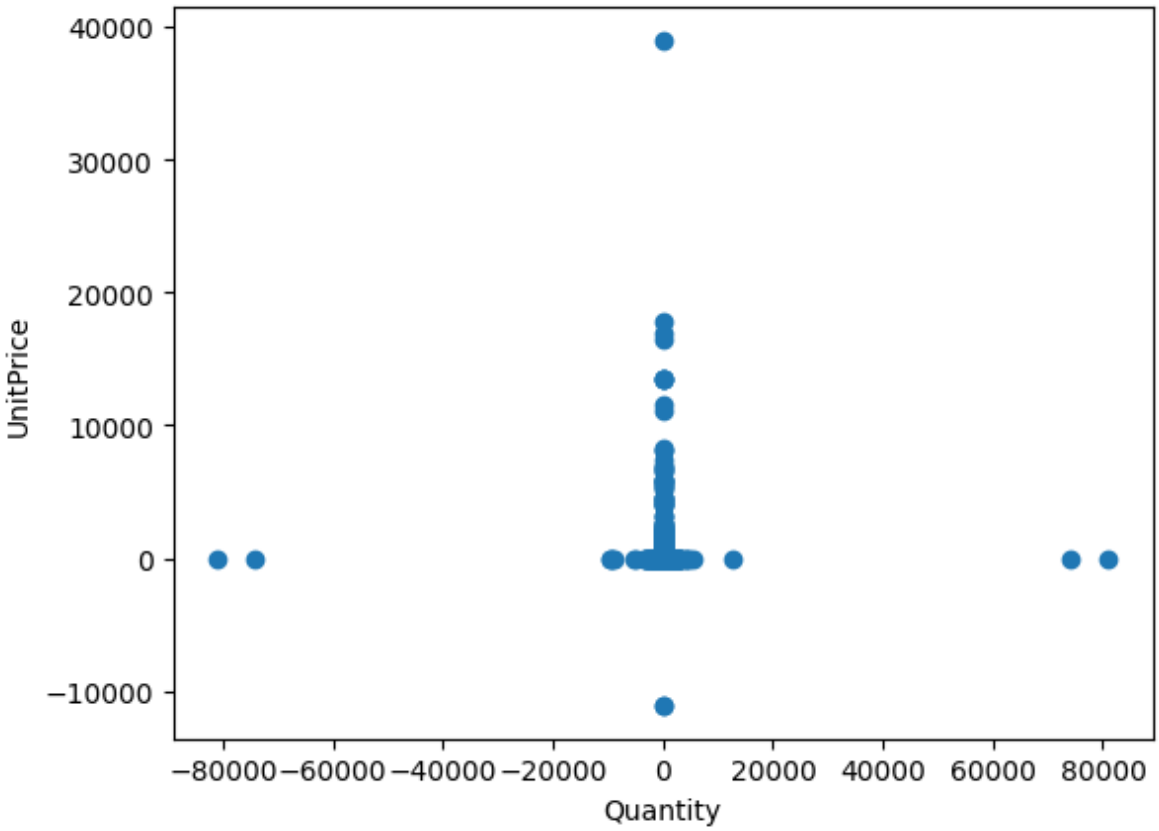
Out[29]:
```
<bound method IndexOpsMixin.value_counts of 0         17850.0
1         17850.0
2         17850.0
3         17850.0
4         17850.0
           ...
541904    12680.0
541905    12680.0
541906    12680.0
541907    12680.0
541908    12680.0
Name: CustomerID, Length: 541909, dtype: float64>
```

In [30]:
```
df['Country'].value_counts
```

Out[30]:
```
<bound method IndexOpsMixin.value_counts of 0         United Kingdom
1         United Kingdom
2         United Kingdom
3         United Kingdom
4         United Kingdom
                ...
541904           France
541905           France
541906           France
541907           France
541908           France
Name: Country, Length: 541909, dtype: object>
```

In [10]:
```
plt.scatter(df["Quantity"],df["UnitPrice"])
plt.xlabel("Quantity")
plt.ylabel("UnitPrice")
```

Out[10]:
```
Text(0, 0.5, 'UnitPrice')
```



In [11]:
```
from sklearn.cluster import KMeans
```

In [12]:
```
km=KMeans()
km
```

Out[12]:
```
KMeans()
```

In [13]:
```
y_predicted=km.fit_predict(df[["Quantity","UnitPrice"]])
y_predicted
```

```
C:\Users\rubin\AppData\Local\Programs\Python\Python310\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([0, 0, 0, ..., 0, 0, 0])
```

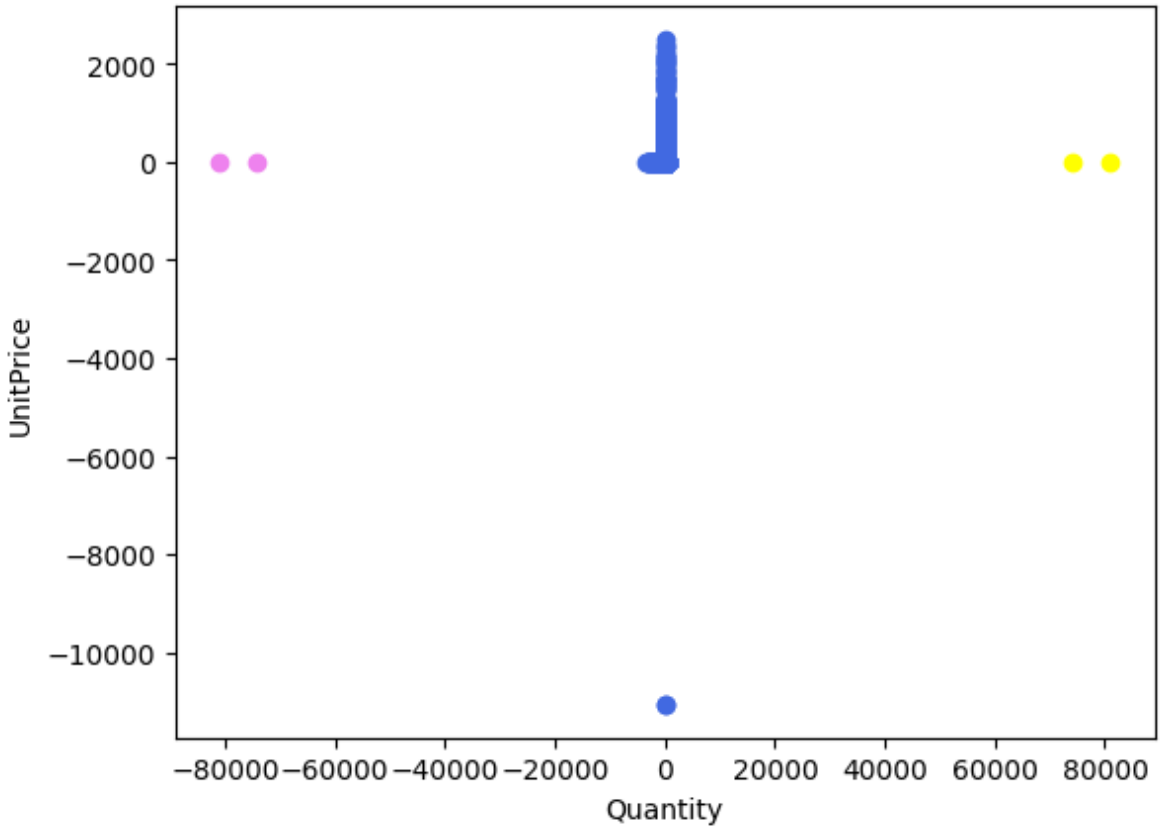
In [14]:
```
df["cluster"]=y_predicted
df.head()
```

Out[14]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	cl
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	1/12/2010 8:26	2.55	17850.0	United Kingdom	
1	536365	71053	WHITE METAL LANTERN	6	1/12/2010 8:26	3.39	17850.0	United Kingdom	
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	1/12/2010 8:26	2.75	17850.0	United Kingdom	
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	1/12/2010 8:26	3.39	17850.0	United Kingdom	
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	1/12/2010 8:26	3.39	17850.0	United Kingdom	

In [15]:
```
df1=df[df.cluster==0]
df2=df[df.cluster==1]
df3=df[df.cluster==2]
plt.scatter(df1["Quantity"],df1["UnitPrice"],color="royalblue")
plt.scatter(df2["Quantity"],df2["UnitPrice"],color="violet")
plt.scatter(df3["Quantity"],df3["UnitPrice"],color="yellow")
plt.xlabel("Quantity")
plt.ylabel("UnitPrice")
```

Out[15]:
```
Text(0, 0.5, 'UnitPrice')
```



In [16]:
```
from sklearn.preprocessing import MinMaxScaler
```

In [17]:
```
Scaler=MinMaxScaler()
```

In [18]:
```
Scaler.fit(df[["Quantity"]])
df["Quantity"]=Scaler.transform(df[["Quantity"]])
df.head()
```

Out[18]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	c
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	0.500037	1/12/2010 8:26	2.55	17850.0	United Kingdom	
1	536365	71053	WHITE METAL LANTERN	0.500037	1/12/2010 8:26	3.39	17850.0	United Kingdom	
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	0.500049	1/12/2010 8:26	2.75	17850.0	United Kingdom	
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	0.500037	1/12/2010 8:26	3.39	17850.0	United Kingdom	
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	0.500037	1/12/2010 8:26	3.39	17850.0	United Kingdom	

In [19]:
```
Scaler.fit(df[["UnitPrice"]])
df["UnitPrice"]=Scaler.transform(df[["UnitPrice"]])
df.head()
```

Out[19]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	c
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	0.500037	1/12/2010 8:26	0.221150	17850.0	United Kingdom	
1	536365	71053	WHITE METAL LANTERN	0.500037	1/12/2010 8:26	0.221167	17850.0	United Kingdom	
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	0.500049	1/12/2010 8:26	0.221154	17850.0	United Kingdom	
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	0.500037	1/12/2010 8:26	0.221167	17850.0	United Kingdom	
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	0.500037	1/12/2010 8:26	0.221167	17850.0	United Kingdom	

In [20]:
```
km=KMeans()
km
```

Out[20]:
```
KMeans
KMeans()
```

In [21]:
```
y_predicted=km.fit_predict(df[["Quantity","UnitPrice"]])
y_predicted
```

```
C:\Users\rubin\AppData\Local\Programs\Python\Python310\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[21]:
```
array([0, 0, 0, ..., 0, 0, 0])
```

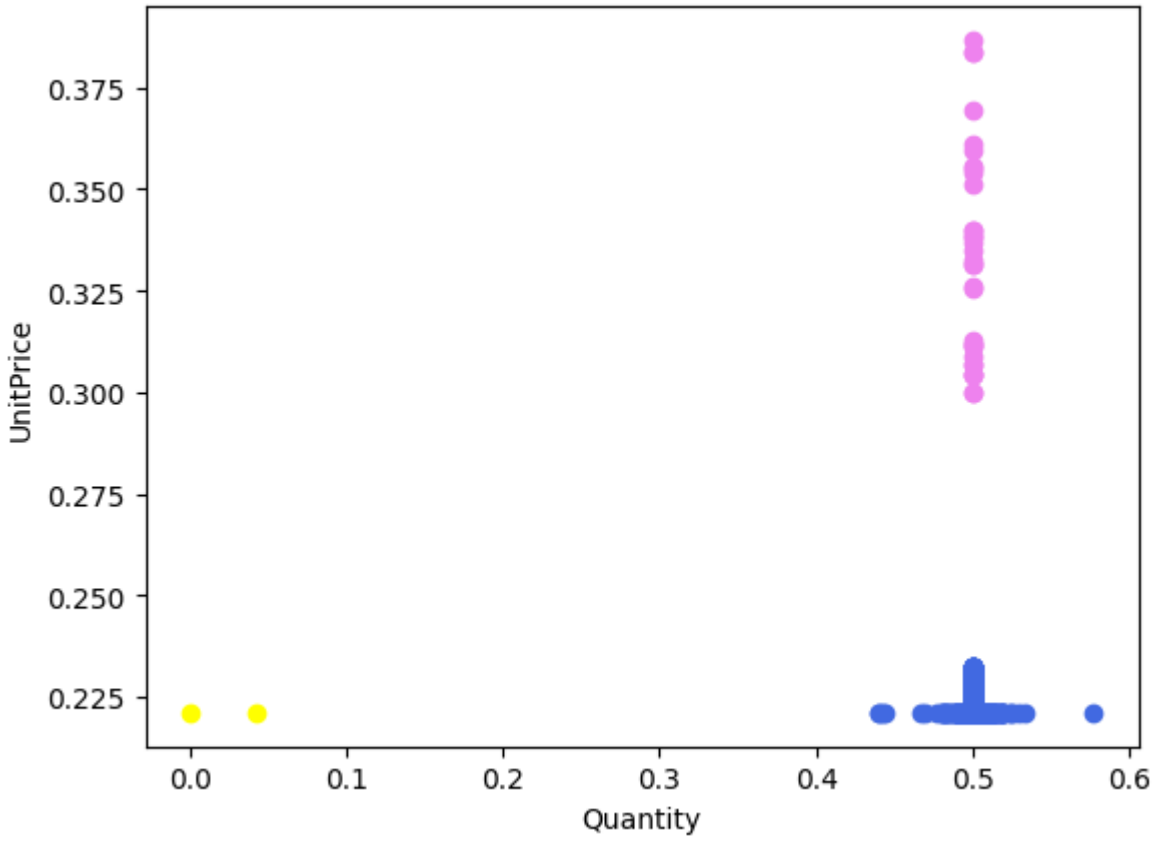
In [22]:
```
df["New cluster"]=y_predicted
df.head()
```

Out[22]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	c
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	0.500037	1/12/2010 8:26	0.221150	17850.0	United Kingdom	
1	536365	71053	WHITE METAL LANTERN	0.500037	1/12/2010 8:26	0.221167	17850.0	United Kingdom	
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	0.500049	1/12/2010 8:26	0.221154	17850.0	United Kingdom	
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	0.500037	1/12/2010 8:26	0.221167	17850.0	United Kingdom	
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	0.500037	1/12/2010 8:26	0.221167	17850.0	United Kingdom	

In [23]:
```
df1=df[df["New cluster"]==0]
df2=df[df["New cluster"]==1]
df3=df[df["New cluster"]==2]
plt.scatter(df1["Quantity"],df1["UnitPrice"],color="royalblue")
plt.scatter(df2["Quantity"],df2["UnitPrice"],color="violet")
plt.scatter(df3["Quantity"],df3["UnitPrice"],color="yellow")
plt.xlabel("Quantity")
plt.ylabel("UnitPrice")
```

Out[23]:
```
Text(0, 0.5, 'UnitPrice')
```

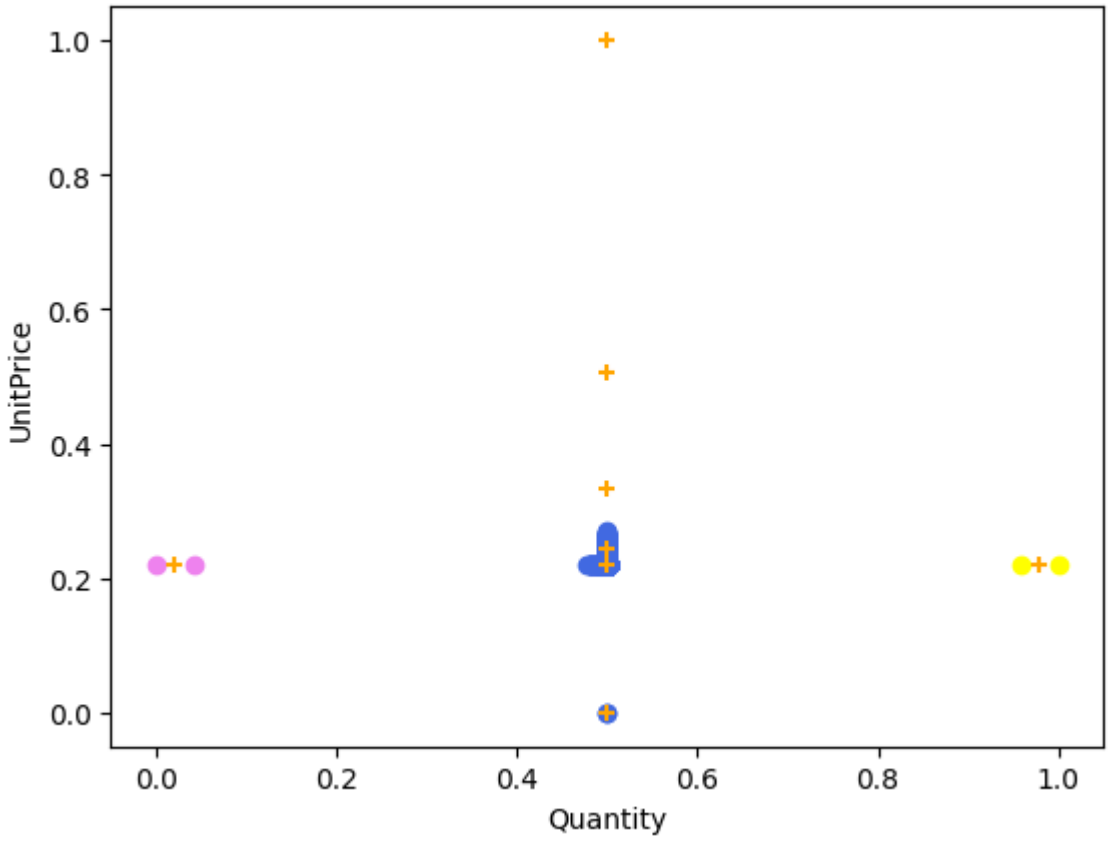


In [24]:
```
km.cluster_centers_
```

Out[24]:
```
array([[0.50005899, 0.22117195],
       [0.49999588, 0.33389462],
       [0.02092722, 0.22113061],
       [0.49999383, 1.        ],
       [0.97907278, 0.22113061],
       [0.49999657, 0.50519622],
       [0.50000617, 0.        ],
       [0.50000432, 0.24394336]])
```

In [25]:
```
df1=df[df.cluster==0]
df2=df[df.cluster==1]
df3=df[df.cluster==2]
plt.scatter(df1["Quantity"],df1["UnitPrice"],color="royalblue")
plt.scatter(df2["Quantity"],df2["UnitPrice"],color="violet")
plt.scatter(df3["Quantity"],df3["UnitPrice"],color="yellow")
plt.scatter(km.cluster_centers_[:,0],km.cluster_centers_[:,1],color="orange",m
plt.xlabel("Quantity")
plt.ylabel("UnitPrice")
```

Out[25]:
```
Text(0, 0.5, 'UnitPrice')
```



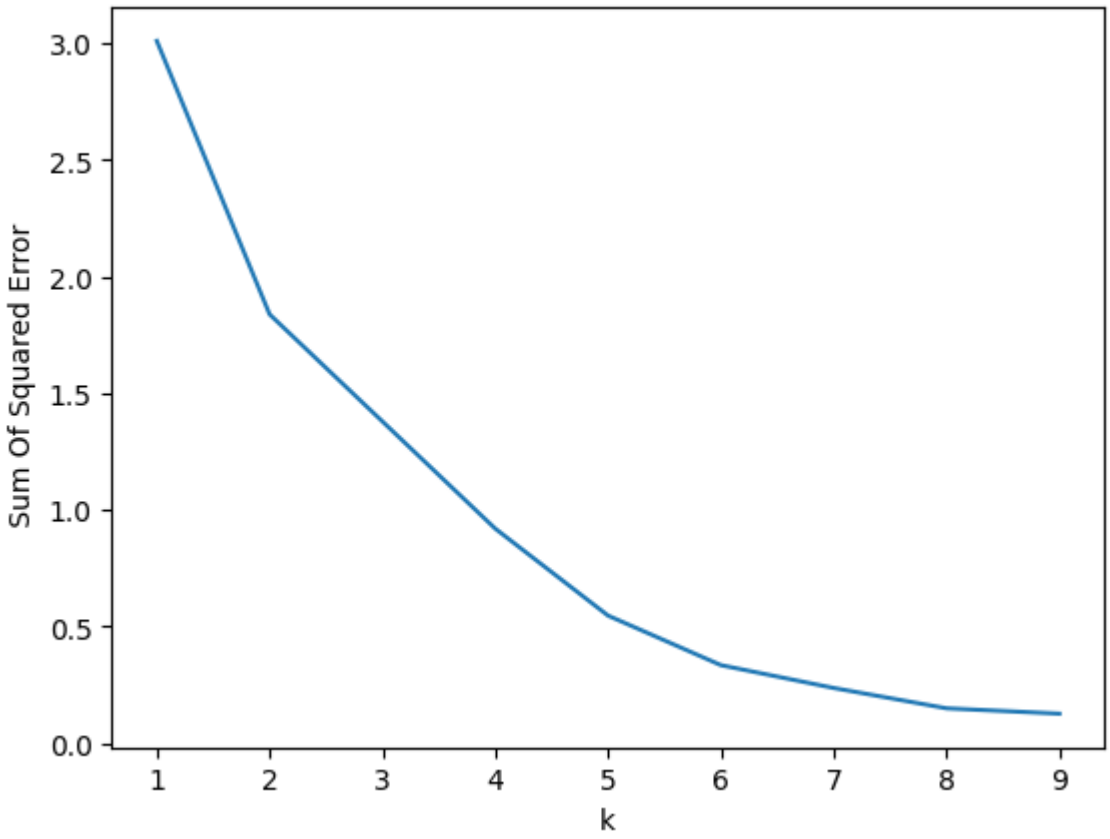
In [27]:
```
k_rng=range(1,10)
sse=[]
for k in k_rng:
    km=KMeans(n_clusters=k)
    km.fit(df[["Quantity","UnitPrice"]])
    sse.append(km.inertia_)
sse
```

```
C:\Users\rubin\AppData\Local\Programs\Python\Python310\lib\site-packages\skle
arn\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 supp
ress the warning
  warnings.warn(
C:\Users\rubin\AppData\Local\Programs\Python\Python310\lib\site-packages\skle
arn\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 supp
ress the warning
  warnings.warn(
C:\Users\rubin\AppData\Local\Programs\Python\Python310\lib\site-packages\skle
arn\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 supp
ress the warning
  warnings.warn(
C:\Users\rubin\AppData\Local\Programs\Python\Python310\lib\site-packages\skle
arn\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 supp
ress the warning
  warnings.warn(
C:\Users\rubin\AppData\Local\Programs\Python\Python310\lib\site-packages\skle
arn\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 supp
ress the warning
  warnings.warn(
C:\Users\rubin\AppData\Local\Programs\Python\Python310\lib\site-packages\skle
arn\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 supp
ress the warning
  warnings.warn(
C:\Users\rubin\AppData\Local\Programs\Python\Python310\lib\site-packages\skle
arn\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 supp
ress the warning
  warnings.warn(
C:\Users\rubin\AppData\Local\Programs\Python\Python310\lib\site-packages\skle
arn\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 supp
ress the warning
  warnings.warn(
C:\Users\rubin\AppData\Local\Programs\Python\Python310\lib\site-packages\skle
arn\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 supp
ress the warning
  warnings.warn(
```

Out[27]:
```
[3.009005955427561,
 1.8375047279575472,
 1.378594567968576,
 0.9194617812520164,
 0.546741407168444,
 0.3345398935085626,
 0.23669990686851558,
 0.14964381100038765,
 0.12673474424186904]
```

In [31]:
```
plt.plot(k_rng,sse)
plt.xlabel("k")
plt.ylabel(" Sum Of Squared Error")
```

Out[31]:
```
Text(0, 0.5, ' Sum Of Squared Error')
```



Conclusion:

For the given dataset we implemented K-Means Clustering and have done for the given data.we have taken two columns and divided into clusters.

In [ ]: