Mini-project 5:

PROBLEM STATEMENT:The transactions made by a UK-based, registered, non-store online retailer between December 1, 2010, and December 9, 2011, are all included in the transnational data set known as online retail. The company primarily offers one-of-a-kind gifts for every occasion. The company has a large number of wholesalers as clients.Company ObjectiveUsing the global online retail dataset, we will design a clustering model and select the ideal group of clients for the business to target.

Import Libraries:

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

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

Out[2]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Coun
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	1/12/2010 8:26	2.55	17850.0	Uni Kingd
1	536365	71053	WHITE METAL LANTERN	6	1/12/2010 8:26	3.39	17850.0	Uni Kingd
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	1/12/2010 8:26	2.75	17850.0	Uni Kingd
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	1/12/2010 8:26	3.39	17850.0	Uni Kingd
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	1/12/2010 8:26	3.39	17850.0	Uni Kingd
...	...	...	...	...	...	...	...	
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	9/12/2011 12:50	0.85	12680.0	Frar
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	9/12/2011 12:50	2.10	12680.0	Frar
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	9/12/2011 12:50	4.15	12680.0	Frar
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	9/12/2011 12:50	4.15	12680.0	Frar
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	9/12/2011 12:50	4.95	12680.0	Frar

541909 rows × 8 columns

Data cleaning & preprocessing:

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

In [5]:
```
df.shape
```

Out[5]:
```
(541909, 8)
```

In [6]:
```
df.describe()
```

Out[6]:

	Quantity	UnitPrice	CustomerID
count	541909.000000	541909.000000	406829.000000
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
max	80995.000000	38970.000000	18287.000000

In [7]:
```
df.columns
```

Out[7]:
```
Index(['InvoiceNo', 'StockCode', 'Description', 'Quantity', 'InvoiceDate',
       'UnitPrice', 'CustomerID', 'Country'],
      dtype='object')
```

In [8]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype  
---  ------       --------------   -----  
 0   InvoiceNo    541909 non-null  object 
 1   StockCode    541909 non-null  object 
 2   Description  540455 non-null  object 
 3   Quantity     541909 non-null  int64  
 4   InvoiceDate  541909 non-null  object 
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object 
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
```

Checking any null values:

In [9]:
```
df.isnull().sum()
```

Out[9]:
```
InvoiceNo          0
StockCode          0
Description     1454
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID     135080
Country            0
dtype: int64
```

In [10]:
```
df.dropna(inplace=True)
```

In [11]:
```
df.isnull().sum()
```

Out[11]:
```
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
```

In [12]:
```
df["Description"].value_counts()
```

Out[12]:
```
Description
WHITE HANGING HEART T-LIGHT HOLDER    2070
REGENCY CAKESTAND 3 TIER              1905
JUMBO BAG RED RETROSPOT               1662
ASSORTED COLOUR BIRD ORNAMENT         1418
PARTY BUNTING                         1416
                                      ...
ANTIQUE RASPBERRY FLOWER EARRINGS        1
WALL ART,ONLY ONE PERSON                 1
GOLD/AMBER DROP EARRINGS W LEAF          1
INCENSE BAZAAR PEACH                     1
PINK BAROQUE FLOCK CANDLE HOLDER         1
Name: count, Length: 3896, dtype: int64
```

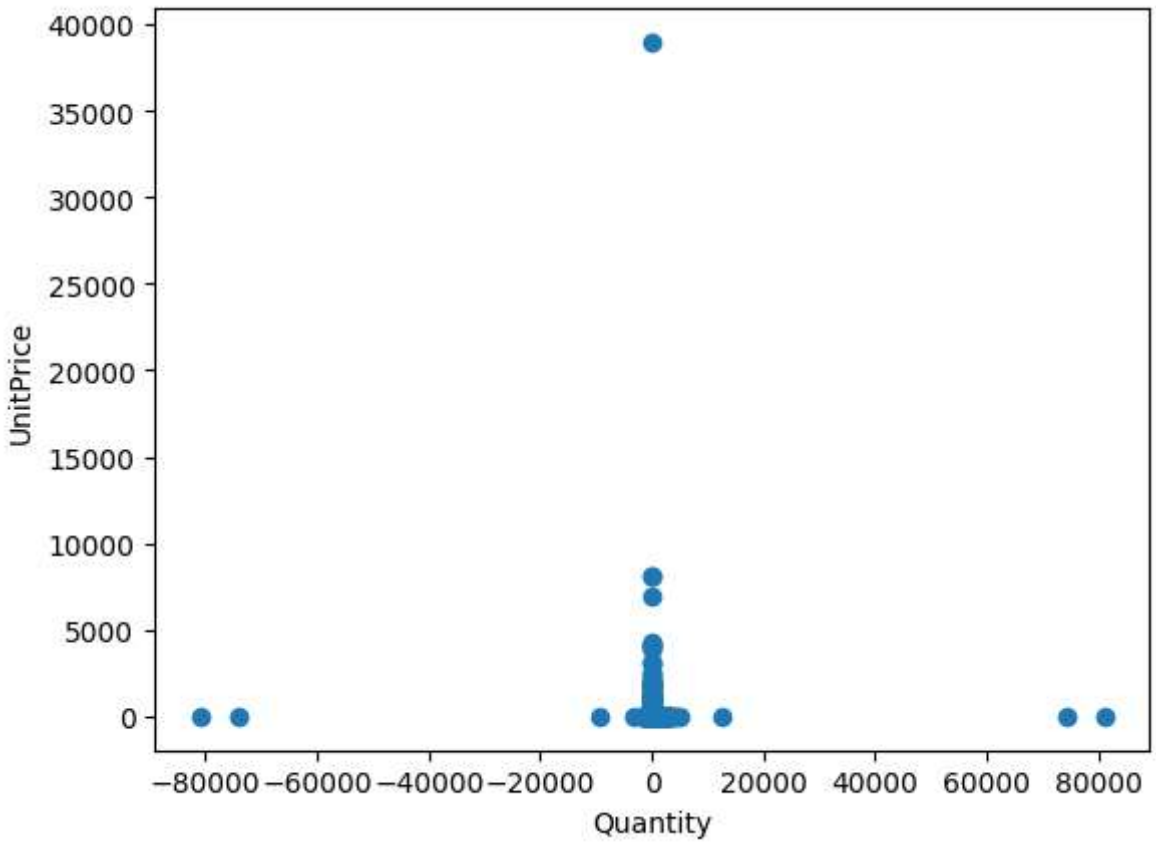
In [13]:
```
df["Country"].value_counts()
```

Out[13]:
```
Country
United Kingdom          361878
Germany                   9495
France                    8491
EIRE                      7485
Spain                     2533
Netherlands               2371
Belgium                   2069
Switzerland               1877
Portugal                  1480
Australia                 1259
Norway                    1086
Italy                      803
Channel Islands            758
Finland                    695
Cyprus                     622
Sweden                     462
Austria                    401
Denmark                    389
Japan                      358
Poland                     341
USA                        291
Israel                     250
Unspecified                244
Singapore                  229
Iceland                    182
Canada                     151
Greece                     146
Malta                      127
United Arab Emirates        68
European Community          61
RSA                         58
Lebanon                     45
Lithuania                   35
Brazil                      32
Czech Republic              30
Bahrain                     17
Saudi Arabia                10
Name: count, dtype: int64
```

KMeans clustering:

In [14]:
```
plt.scatter(df["Quantity"],df["UnitPrice"])
plt.xlabel("Quantity")
plt.ylabel("UnitPrice")
```

Out[14]:
```
Text(0, 0.5, 'UnitPrice')
```



In [15]:
```
from sklearn.cluster import KMeans
km=KMeans()
km
```

Out[15]:
```
KMeans()
```

In [16]:
```
y_predicted=km.fit_predict(df[["Quantity","UnitPrice"]])
y_predicted
```

```
C:\Users\DELL\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklea
rn\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[16]:
```
array([0, 0, 0, ..., 0, 0, 0])
```

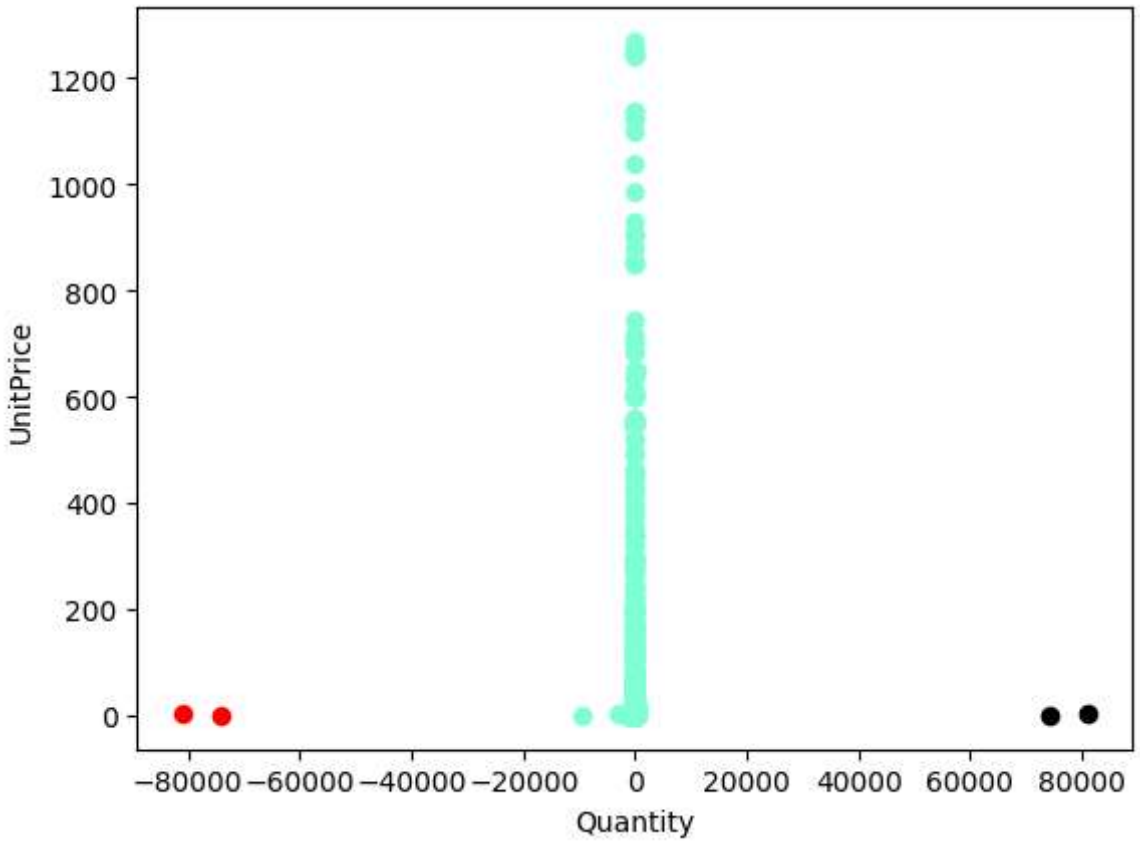
In [17]:
```
df["cluster"]=y_predicted
df.head()
```

Out[17]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	clu
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 [18]:
```
df1=df[df.cluster==0]
df2=df[df.cluster==1]
df3=df[df.cluster==2]
plt.scatter(df1["Quantity"],df1["UnitPrice"],color="aquamarine")
plt.scatter(df2["Quantity"],df2["UnitPrice"],color="red")
plt.scatter(df3["Quantity"],df3["UnitPrice"],color="black")
plt.xlabel("Quantity")
plt.ylabel("UnitPrice")
```

Out[18]:
```
Text(0, 0.5, 'UnitPrice')
```



In [19]:
```
from sklearn.preprocessing import MinMaxScaler
scaler=MinMaxScaler()
scaler.fit(df[["UnitPrice"]])
df["UnitPrice"]=scaler.transform(df[["UnitPrice"]])
df.head()
```

Out[19]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	clu
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	1/12/2010 8:26	0.000065	17850.0	United Kingdom	
1	536365	71053	WHITE METAL LANTERN	6	1/12/2010 8:26	0.000087	17850.0	United Kingdom	
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	1/12/2010 8:26	0.000071	17850.0	United Kingdom	
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	1/12/2010 8:26	0.000087	17850.0	United Kingdom	
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	1/12/2010 8:26	0.000087	17850.0	United Kingdom	

In [20]:
```
scaler.fit(df[["Quantity"]])
df["Quantity"]=scaler.transform(df[["Quantity"]])
df.head()
```

Out[20]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	cl
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	0.500037	1/12/2010 8:26	0.000065	17850.0	United Kingdom	
1	536365	71053	WHITE METAL LANTERN	0.500037	1/12/2010 8:26	0.000087	17850.0	United Kingdom	
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	0.500049	1/12/2010 8:26	0.000071	17850.0	United Kingdom	
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	0.500037	1/12/2010 8:26	0.000087	17850.0	United Kingdom	
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	0.500037	1/12/2010 8:26	0.000087	17850.0	United Kingdom	

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

```
C:\Users\DELL\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklea
rn\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[21]:
```
array([0, 0, 0, ..., 0, 0, 0])
```

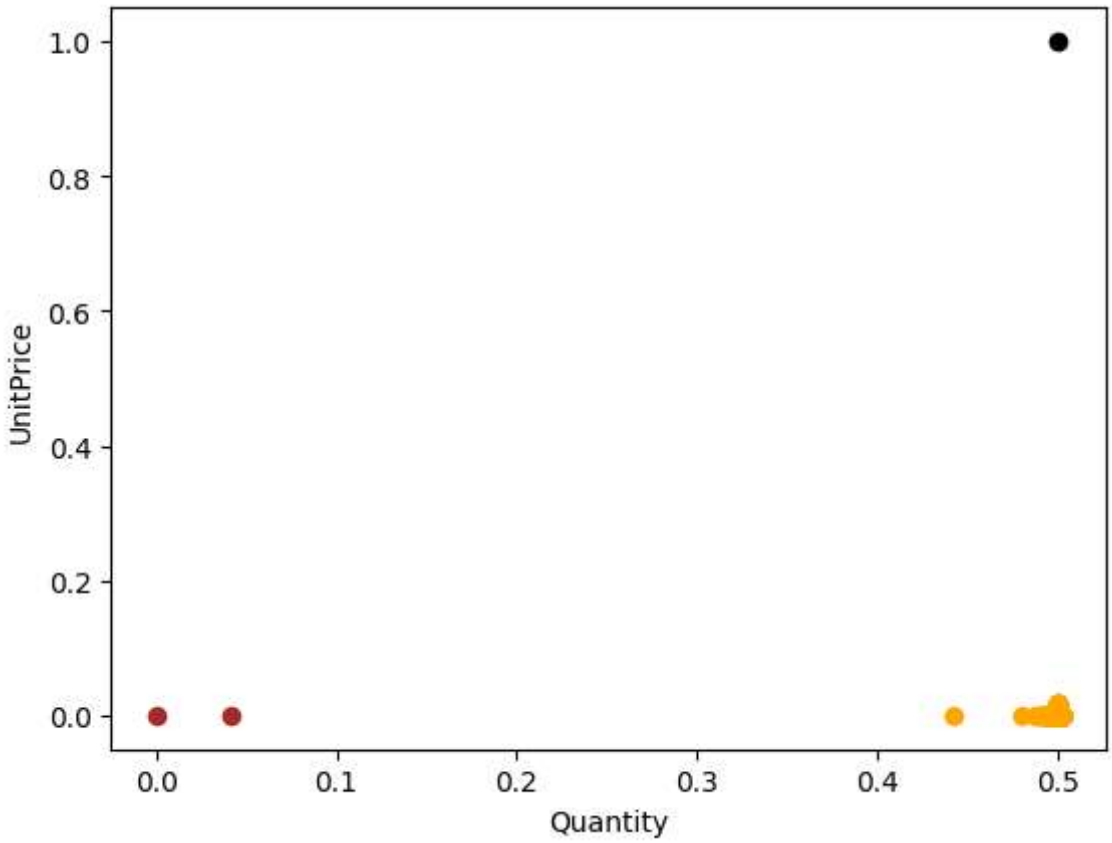
In [22]:
```
df["New cluster"]=y_predicted
df.head()
```

Out[22]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	cl
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	0.500037	1/12/2010 8:26	0.000065	17850.0	United Kingdom	
1	536365	71053	WHITE METAL LANTERN	0.500037	1/12/2010 8:26	0.000087	17850.0	United Kingdom	
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	0.500049	1/12/2010 8:26	0.000071	17850.0	United Kingdom	
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	0.500037	1/12/2010 8:26	0.000087	17850.0	United Kingdom	
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	0.500037	1/12/2010 8:26	0.000087	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="orange")
plt.scatter(df2["Quantity"],df2["UnitPrice"],color="brown")
plt.scatter(df3["Quantity"],df3["UnitPrice"],color="black")
plt.xlabel("Quantity")
plt.ylabel("UnitPrice")
```

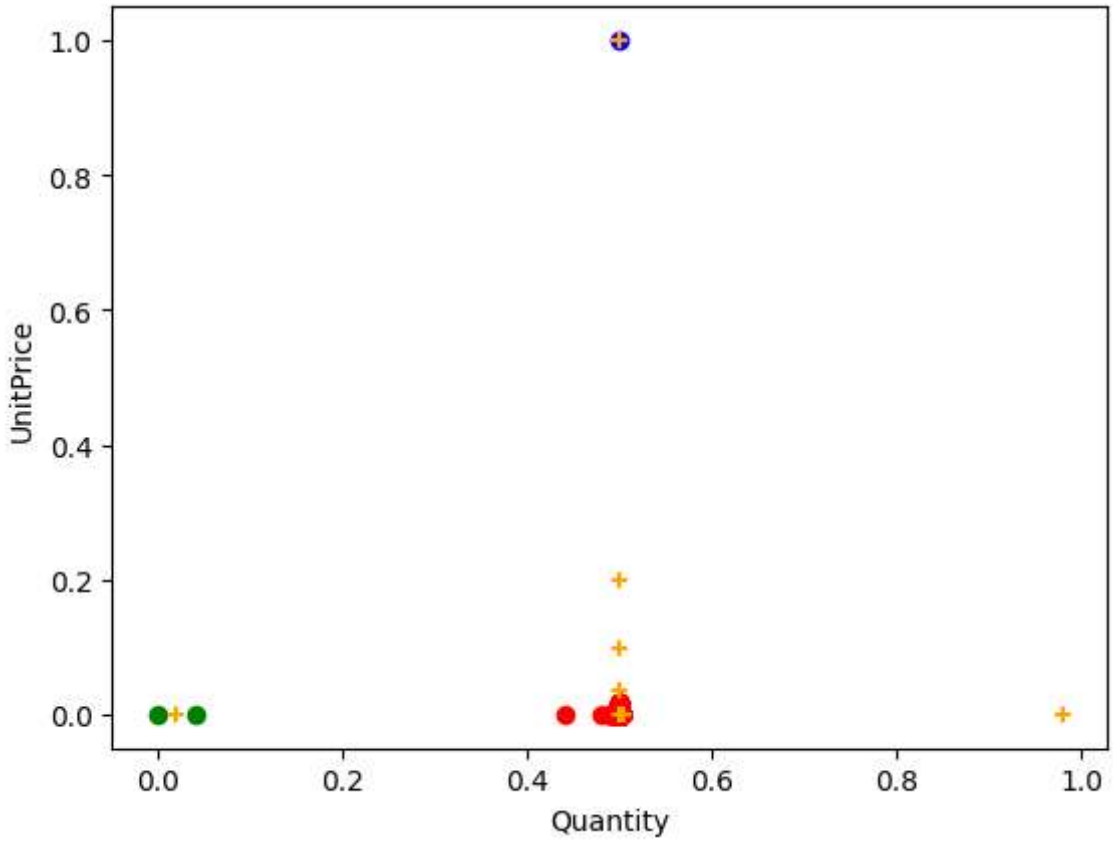
Out[23]:
```
Text(0, 0.5, 'UnitPrice')
```



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

Out[24]:
```
array([[5.00066521e-01, 7.88109807e-05],
       [2.09272177e-02, 4.00307929e-05],
       [4.99993827e-01, 1.00000000e+00],
       [4.99998628e-01, 9.70602743e-02],
       [9.79072782e-01, 4.00307929e-05],
       [4.99999863e-01, 3.63112708e-02],
       [4.99997942e-01, 1.98575828e-01],
       [5.04476646e-01, 3.05072535e-05]])
```

In [26]:
```
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="red")
plt.scatter(df2["Quantity"],df2["UnitPrice"],color="green")
plt.scatter(df3["Quantity"],df3["UnitPrice"],color="blue")
plt.scatter(km.cluster_centers_[:,0],km.cluster_centers_[:,1],color="orange",ma
plt.xlabel("Quantity")
plt.ylabel("UnitPrice")
plt.show()
```



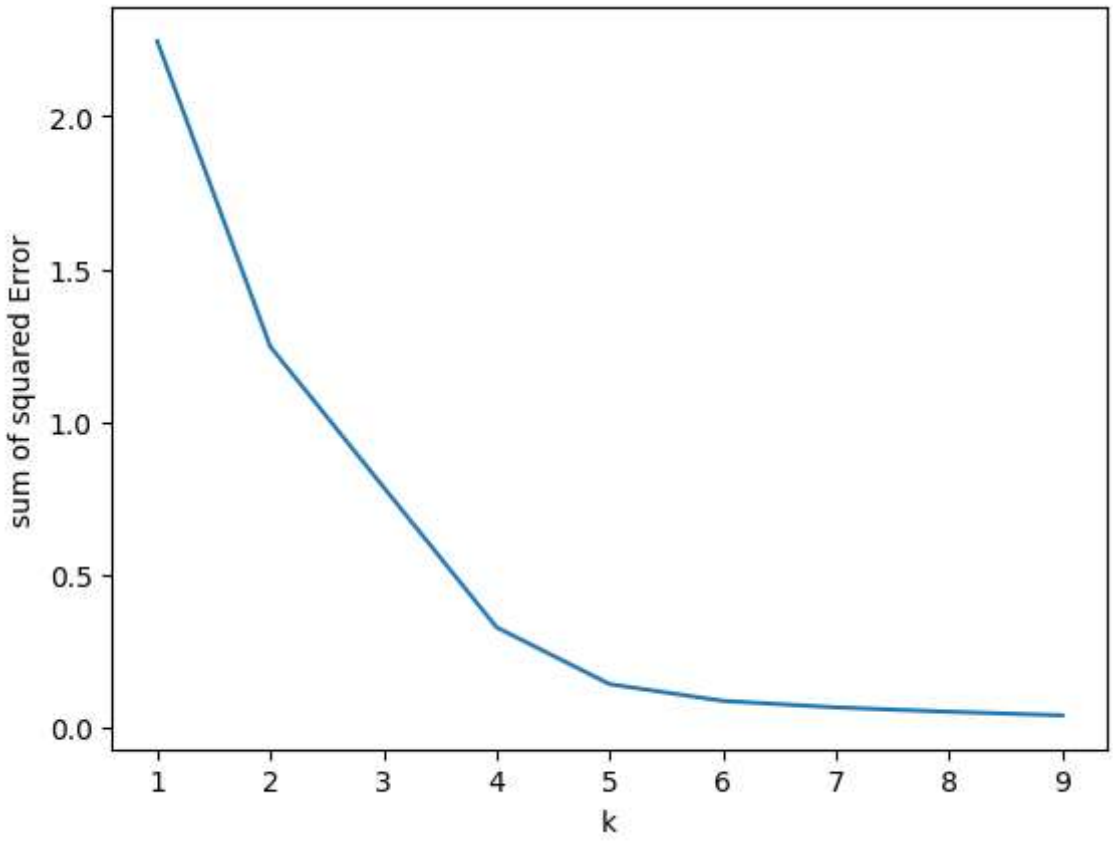
In [28]:
```
k_rng = range(1,10)
sse = []
for k in k_rng:
 km = KMeans(n_clusters = k)
 km.fit(df[["Quantity","UnitPrice"]])
 sse.append(km.inertia_)
sse
```

```
C:\Users\DELL\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklea
rn\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\DELL\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklea
rn\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\DELL\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklea
rn\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\DELL\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklea
rn\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\DELL\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklea
rn\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\DELL\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklea
rn\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\DELL\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklea
rn\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\DELL\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklea
rn\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\DELL\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklea
rn\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[28]:
```
[2.245959720214337,
 1.2461348448882017,
 0.7869684289528796,
 0.32809189128436744,
 0.14220510497766678,
 0.0874728096488116,
 0.06618719940451731,
 0.05195491893577105,
 0.0401880944778316]
```

In [29]:
```
plt.plot(k_rng,sse)
plt.xlabel("k")
plt.ylabel("sum of squared Error")
```

Out[29]:
```
Text(0, 0.5, 'sum of squared Error')
```



CONCLUSION:This Online Retail.csv DataFrame is done by using KMeans Clustering.