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 offer one-of-a-kind gifts for every occasion. The company has a large number of wholesalers as clients.Company Objective Using the global online retail dataset, we will design a clustering model and select the ideal group of clients for the business to target.

In [1]:

```python
#import libraries
import pandas as pd
from matplotlib import pyplot as plt
%matplotlib inline
```

In [2]:

```
df=pd.read_csv(r"C:\Users\raja\Downloads\OnlineRetail2.csv")
df
```

Out[2]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	H
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	H
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	H
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	H
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	H
...	...	...	...	...	...	...	...	
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	09-12-2011 12:50	0.85	12680.0	
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	09-12-2011 12:50	2.10	12680.0	
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	09-12-2011 12:50	4.15	12680.0	
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	09-12-2011 12:50	4.15	12680.0	
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	09-12-2011 12:50	4.95	12680.0	

541909 rows × 8 columns

In [3]:

```
df.head()
```

Out[3]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	United Kingdom

In [4]:

```
df.tail()
```

Out[4]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	09-12-2011 12:50	0.85	12680.0	
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	09-12-2011 12:50	2.10	12680.0	
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	09-12-2011 12:50	4.15	12680.0	
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	09-12-2011 12:50	4.15	12680.0	
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	09-12-2011 12:50	4.95	12680.0	

In [5]:

```python
df['CustomerID'].value_counts()
```

Out[5]:

```
CustomerID
17841.0    7983
14911.0    5903
14096.0    5128
12748.0    4642
14606.0    2782
           ...
15070.0       1
15753.0       1
17065.0       1
16881.0       1
16995.0       1
Name: count, Length: 4372, dtype: int64
```

In [6]:

```python
df['Quantity'].value_counts()
```

Out[6]:

```
Quantity
 1      148227
 2       81829
 12      61063
 6       40868
 4       38484
          ...
-472         1
-161         1
-1206        1
-272         1
-80995       1
Name: count, Length: 722, dtype: int64
```

In [7]:

```python
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
```

In [8]:

```python
df.isnull().sum()
```

Out[8]:

```
InvoiceNo          0
StockCode          0
Description      1454
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID     135080
Country            0
dtype: int64
```

In [9]:

```python
df.fillna(method='ffill',inplace=True)
```

In [10]:

```python
df.isnull().sum()
```

Out[10]:

```
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
```

In [16]:

```python
plt.scatter(df["Quantity"],df["CustomerID"])
plt.xlabel("Quantity")
plt.ylabel("CustomerID")
```

Out[16]:

```
Text(0, 0.5, 'CustomerID')
```
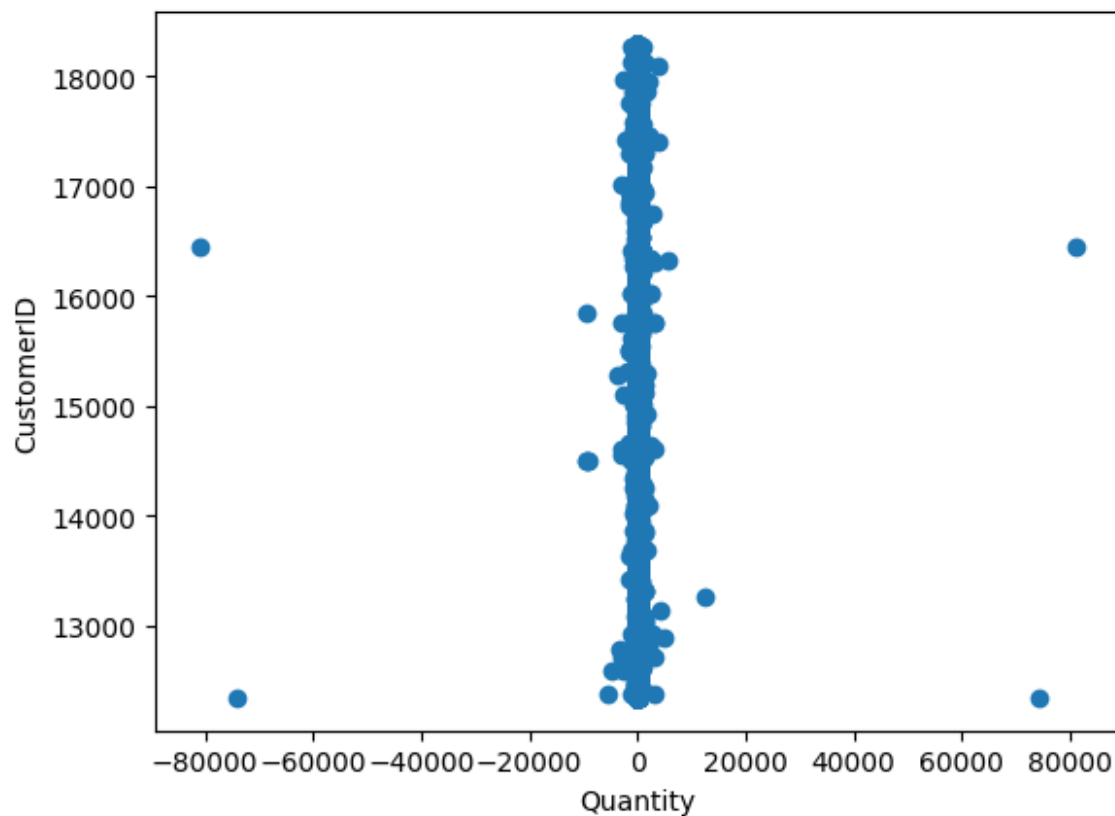

KMeans Clustering

In [12]:

```python
from sklearn.cluster import KMeans
```

In [13]:

```python
km=KMeans()
km
```

Out[13]:

```
KMeans()
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [14]:

```python
y_predicted=km.fit_predict(df[["Quantity","CustomerID"]])
y_predicted
```

```
C:\Users\raja\AppData\Local\Programs\Python\Python310\lib\site-packages\sk
learn\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[14]:

```
array([4, 4, 4, ..., 7, 7, 7])
```

In [15]:

```python
df["Cluster"]=y_predicted
df.head()
```
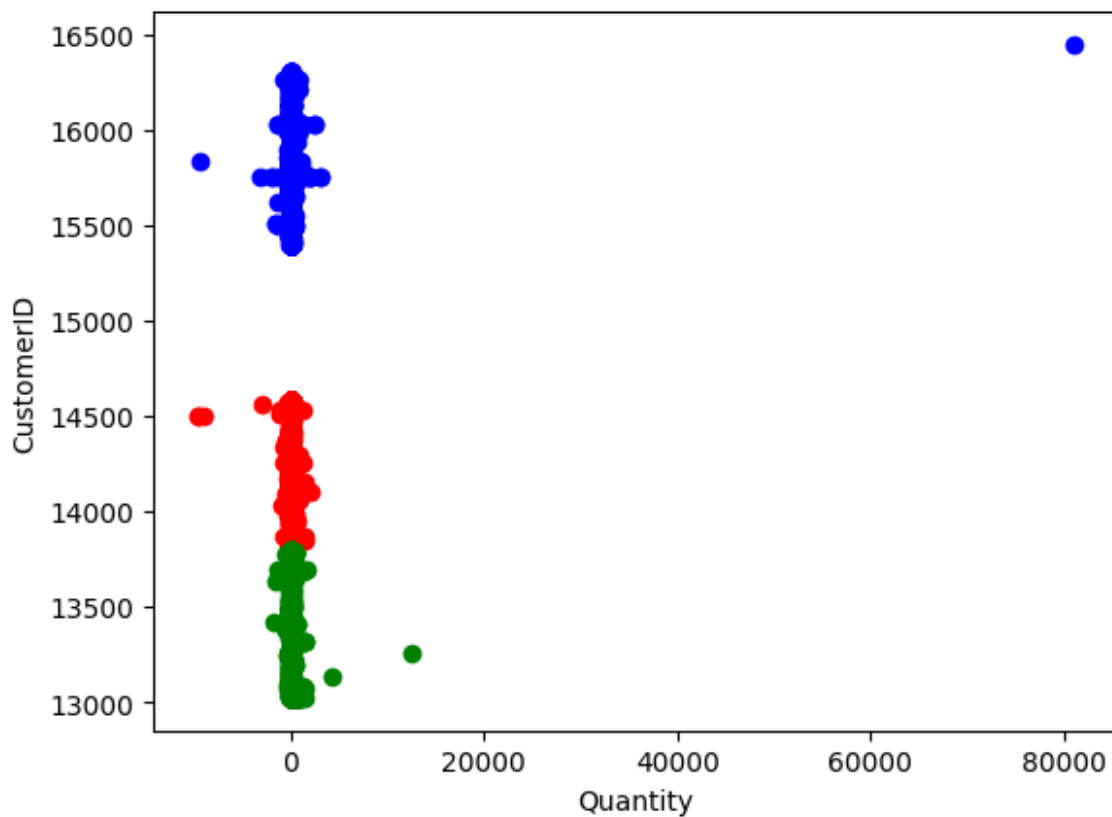
Out[15]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	United Kingdom

In [17]:

```python
df1=df[df.Cluster==0]
df2=df[df.Cluster==2]
df3=df[df.Cluster==3]
plt.scatter(df1["Quantity"],df1["CustomerID"],color="red")
plt.scatter(df2["Quantity"],df2["CustomerID"],color="green")
plt.scatter(df3["Quantity"],df3["CustomerID"],color="blue")
plt.xlabel("Quantity")
plt.ylabel("CustomerID")
```

Out[17]:

```
Text(0, 0.5, 'CustomerID')
```


In [18]:

```python
from sklearn.preprocessing import MinMaxScaler
```

In [19]:

```python
scaler=MinMaxScaler()
```

In [20]:

```python
scaler.fit(df[["CustomerID"]])
df["CustomerID"]=scaler.transform(df[["CustomerID"]])
df.head()
```

Out[20]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	0.926443	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	0.926443	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	0.926443	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	0.926443	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	0.926443	United Kingdom

In [21]:

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

Out[21]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	0.500037	01-12-2010 08:26	2.55	0.926443	United Kingdom
1	536365	71053	WHITE METAL LANTERN	0.500037	01-12-2010 08:26	3.39	0.926443	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	0.500049	01-12-2010 08:26	2.75	0.926443	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	0.500037	01-12-2010 08:26	3.39	0.926443	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	0.500037	01-12-2010 08:26	3.39	0.926443	United Kingdom

In [22]:

```
km=KMeans()
```

In [23]:

```
y_predicted=km.fit_predict(df[["Quantity","CustomerID"]])
y_predicted
```

```
C:\Users\raja\AppData\Local\Programs\Python\Python310\lib\site-packages\sk
learn\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[23]:

```
array([7, 7, 7, ..., 3, 3, 3])
```

In [24]:

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

Out[24]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	0.500037	01-12-2010 08:26	2.55	0.926443	United Kingdom
1	536365	71053	WHITE METAL LANTERN	0.500037	01-12-2010 08:26	3.39	0.926443	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	0.500049	01-12-2010 08:26	2.75	0.926443	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	0.500037	01-12-2010 08:26	3.39	0.926443	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	0.500037	01-12-2010 08:26	3.39	0.926443	United Kingdom

In [25]:

```python
df1=df[df["New cluster"]==0]
df2=df[df["New cluster"]==1]
df3=df[df["New cluster"]==2]
plt.scatter(df1["Quantity"],df1["CustomerID"],color="red")
plt.scatter(df2["Quantity"],df2["CustomerID"],color="green")
plt.scatter(df3["Quantity"],df3["CustomerID"],color="blue")
plt.xlabel("Quantity")
plt.ylabel("CustomerID")
```

Out[25]:

```
Text(0, 0.5, 'CustomerID')
```
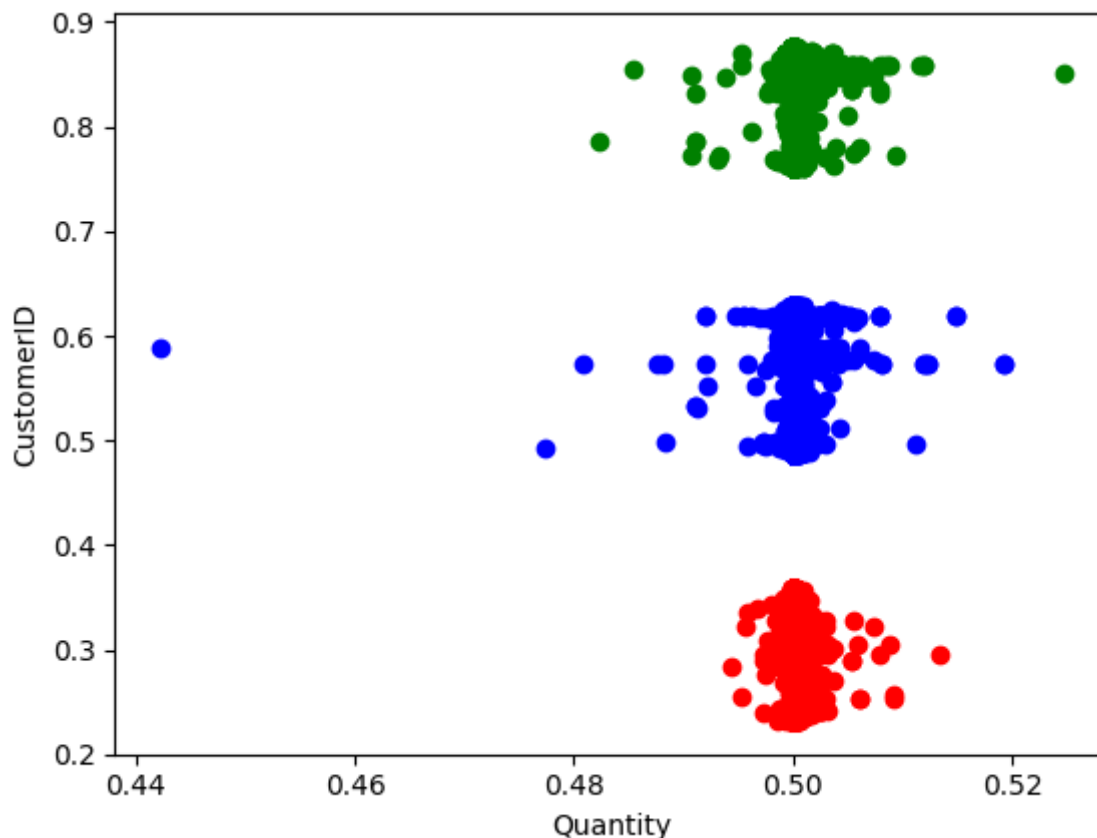

In [26]:

```python
km.cluster_centers_
```

Out[26]:

```
array([[0.50006065, 0.29837516],
       [0.50006047, 0.81912277],
       [0.50005342, 0.55551005],
       [0.50006705, 0.05156814],
       [0.50005804, 0.70196132],
       [0.50006106, 0.4185919 ],
       [0.50006062, 0.16561407],
       [0.50005085, 0.93332193]])
```

In [27]:

```
df1=df[df["New cluster"]==0]
df2=df[df["New cluster"]==1]
df3=df[df["New cluster"]==2]
plt.scatter(df1["Quantity"],df1["CustomerID"],color="red")
plt.scatter(df2["Quantity"],df2["CustomerID"],color="green")
plt.scatter(df3["Quantity"],df3["CustomerID"],color="blue")
plt.scatter(km.cluster_centers_[:,0],km.cluster_centers_[:,1],color="orange",marker="+")
plt.xlabel("Quantity")
plt.ylabel("CustomerID")
```

Out[27]:

Text(0, 0.5, 'CustomerID')
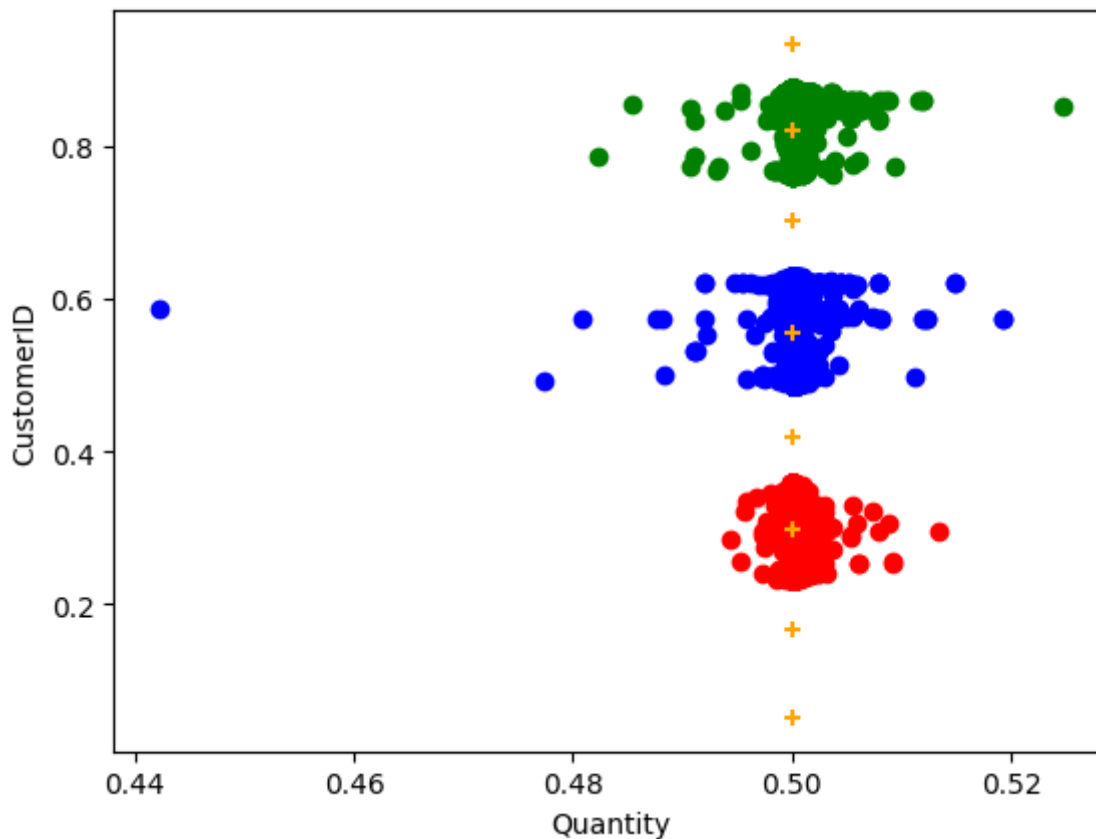
In [31]:

```
k_rng=range(1,10)
sse=[]
```

In [32]:

```python
for k in k_rng:
    km=KMeans(n_clusters=k)
    km.fit(df[["Quantity","CustomerID"]])
    sse.append(km.inertia_)
#km.inertia_ will give you the value of sum of square errorprint(sse)
plt.plot(k_rng,sse)
plt.xlabel("K")
plt.ylabel("Sum of Squared Error")
```

C:\Users\raja\AppData\Local\Programs\Python\Python310\lib\site-packages\sk
learn\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(
C:\Users\raja\AppData\Local\Programs\Python\Python310\lib\site-packages\sk
learn\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(
C:\Users\raja\AppData\Local\Programs\Python\Python310\lib\site-packages\sk
learn\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(
C:\Users\raja\AppData\Local\Programs\Python\Python310\lib\site-packages\sk
learn\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(
C:\Users\raja\AppData\Local\Programs\Python\Python310\lib\site-packages\sk
learn\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(
C:\Users\raja\AppData\Local\Programs\Python\Python310\lib\site-packages\sk
learn\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(
C:\Users\raja\AppData\Local\Programs\Python\Python310\lib\site-packages\sk
learn\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(
C:\Users\raja\AppData\Local\Programs\Python\Python310\lib\site-packages\sk
learn\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(
C:\Users\raja\AppData\Local\Programs\Python\Python310\lib\site-packages\sk
learn\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
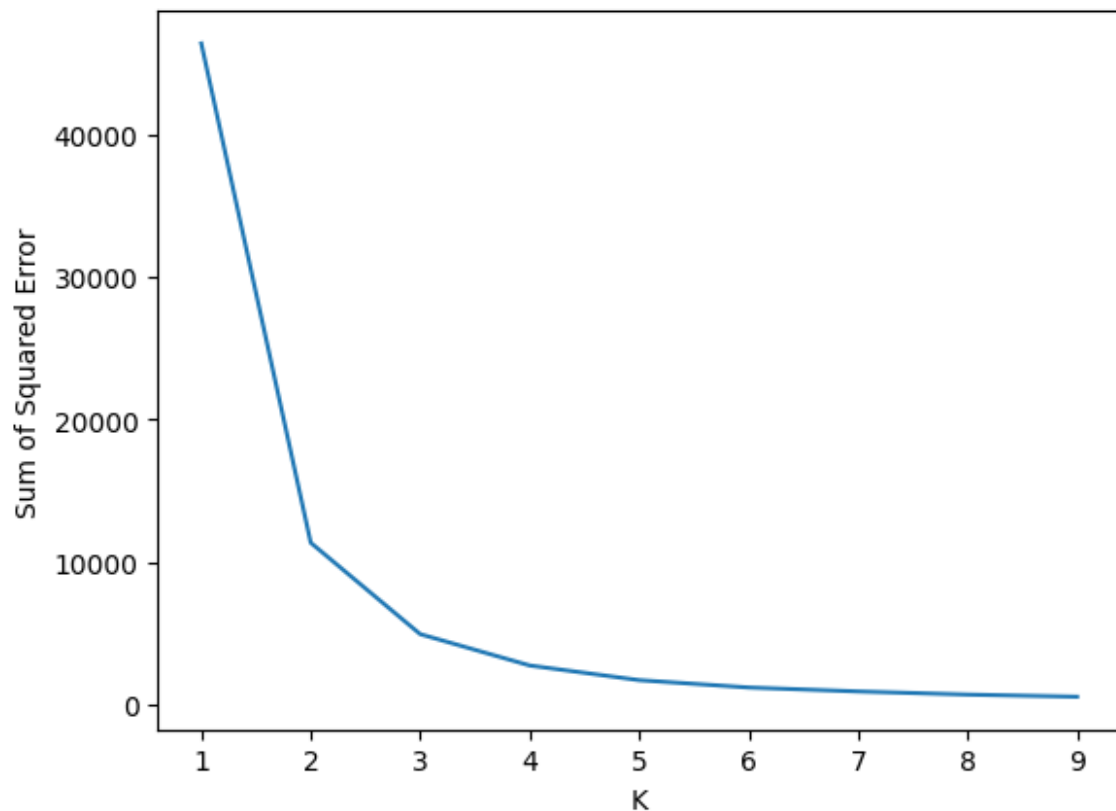to suppress the warning
  warnings.warn(
C:\Users\raja\AppData\Local\Programs\Python\Python310\lib\site-packages\sk
learn\cluster_kmeans.py:870: FutureWarning: The default value of `n_init`
will change from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly
to suppress the warning
  warnings.warn(

Out[32]:

Text(0, 0.5, 'Sum of Squared Error')

Conclusion:

 In This dataset we are performing clustering on Quantity and CustomerID.By usi
ng kMeans Algorithm.so we conclude that KMeans alogorithm is best fot this Data
set.