Online Retail

In [40]:

```
import pandas as pd
from matplotlib import pyplot as plt
%matplotlib inline
import seaborn as sns
import matplotlib.pyplot as plt
```

In [41]:

```
df=pd.read_csv(r"C:\Users\Svijayalakshmi\Downloads\OnlineRetail new.csv")
df
```

Out[41]:

	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	6	01-12-2010 08:26	3.39	17850.0	United Kingdom

In [42]:

```
df.head()
```

Out[42]:

	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	Unitec Kingdom
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	Unitec Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	Unitec Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	Unitec Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	Unitec Kingdom

In [43]:

```
df.tail()
```

Out[43]:

	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 [44]:

```
df.describe
```

Out[44]:

```
<bound method NDFrame.describe of        InvoiceNo StockCode
Description  Quantity
0         536365    85123A   WHITE HANGING HEART T-LIGHT HOLDER         6
\
1         536365     71053                  WHITE METAL LANTERN         6
2         536365    84406B       CREAM CUPID HEARTS COAT HANGER         8
3         536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE         6
4         536365    84029E       RED WOOLLY HOTTIE WHITE HEART.         6
...          ...       ...                                  ...       ...
541904    581587     22613          PACK OF 20 SPACEBOY NAPKINS        12
541905    581587     22899         CHILDREN'S APRON DOLLY GIRL          6
541906    581587     23254        CHILDRENS CUTLERY DOLLY GIRL          4
541907    581587     23255      CHILDRENS CUTLERY CIRCUS PARADE         4
541908    581587     22138        BAKING SET 9 PIECE RETROSPOT          3

             InvoiceDate  UnitPrice  CustomerID         Country
0       01-12-2010 08:26       2.55     17850.0  United Kingdom
1       01-12-2010 08:26       3.39     17850.0  United Kingdom
2       01-12-2010 08:26       2.75     17850.0  United Kingdom
3       01-12-2010 08:26       3.39     17850.0  United Kingdom
4       01-12-2010 08:26       3.39     17850.0  United Kingdom
...                  ...        ...         ...             ...
541904  09-12-2011 12:50       0.85     12680.0          France
541905  09-12-2011 12:50       2.10     12680.0          France
541906  09-12-2011 12:50       4.15     12680.0          France
541907  09-12-2011 12:50       4.15     12680.0          France
541908  09-12-2011 12:50       4.95     12680.0          France

[541909 rows x 8 columns]>
```

In [45]:

```
df.isna().any()
```

Out[45]:

```
InvoiceNo      False
StockCode      False
Description     True
Quantity       False
InvoiceDate    False
UnitPrice      False
CustomerID      True
Country        False
dtype: bool
```

In [46]:

```
df.shape
```

Out[46]:

```
(541909, 8)
```

In [47]:

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

In [48]:

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

Out[48]:

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

In [49]:

```
df=df[['Quantity','UnitPrice','CustomerID']]
df
```

Out[49]:

	Quantity	UnitPrice	CustomerID
0	6	2.55	17850.0
1	6	3.39	17850.0
2	8	2.75	17850.0
3	6	3.39	17850.0
4	6	3.39	17850.0
...	...	...	...
541904	12	0.85	12680.0
541905	6	2.10	12680.0
541906	4	4.15	12680.0
541907	4	4.15	12680.0
541908	3	4.95	12680.0

541909 rows × 3 columns

In [50]:

```
df.shape
```

Out[50]:

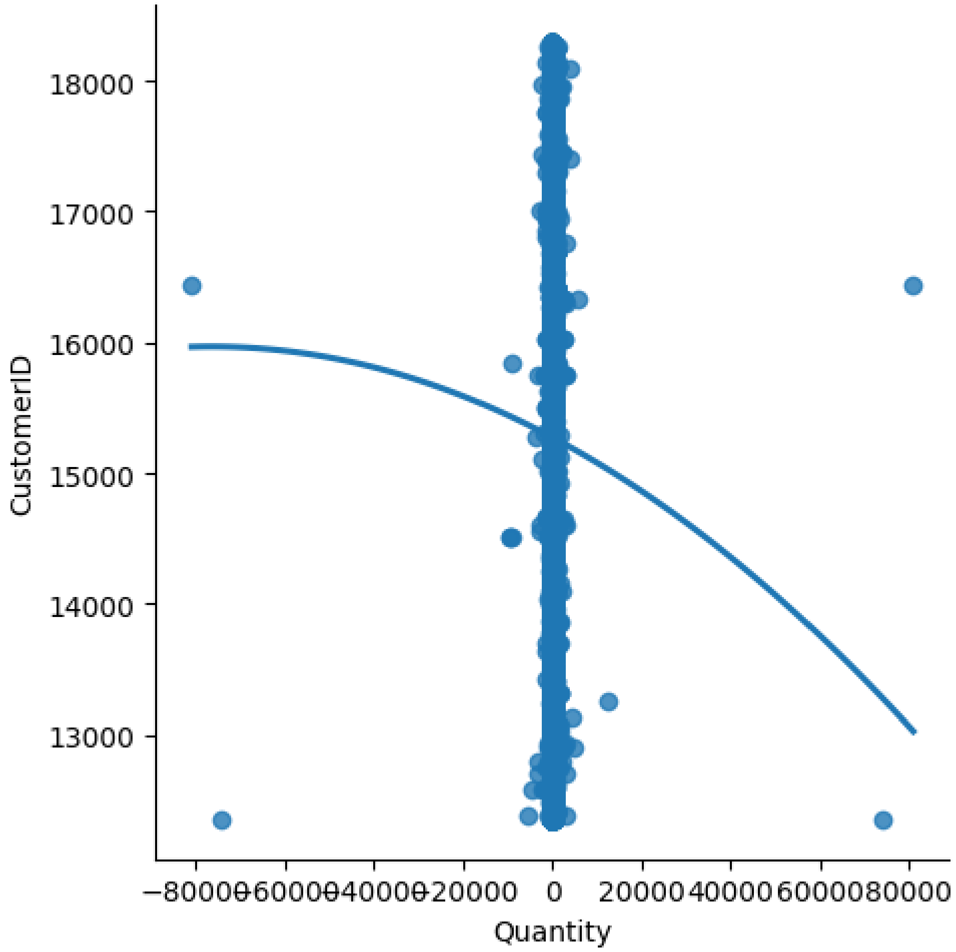
```
(541909, 3)
```

In [54]:

```
sns.lmplot(x='Quantity',y='CustomerID',data=df,order=2,ci=None)
```

Out[54]:

```
<seaborn.axisgrid.FacetGrid at 0x1a05bcb8590>
```



In [55]:

```
from sklearn.cluster import KMeans
km=KMeans()
km
```

Out[55]:

```
▾ KMeans
KMeans()
```

In [58]:

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

```
C:\Users\Svijayalakshmi\AppData\Local\Programs\Python\Python311\Lib\site-p
ackages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value o
f `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[58]:

```
array([3, 3, 3, ..., 4, 4, 4])
```

In [59]:

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

```
C:\Users\Svijayalakshmi\AppData\Local\Temp\ipykernel_23292\1084992799.py:
1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  df["cluster"]=y_predicted
```

Out[59]:

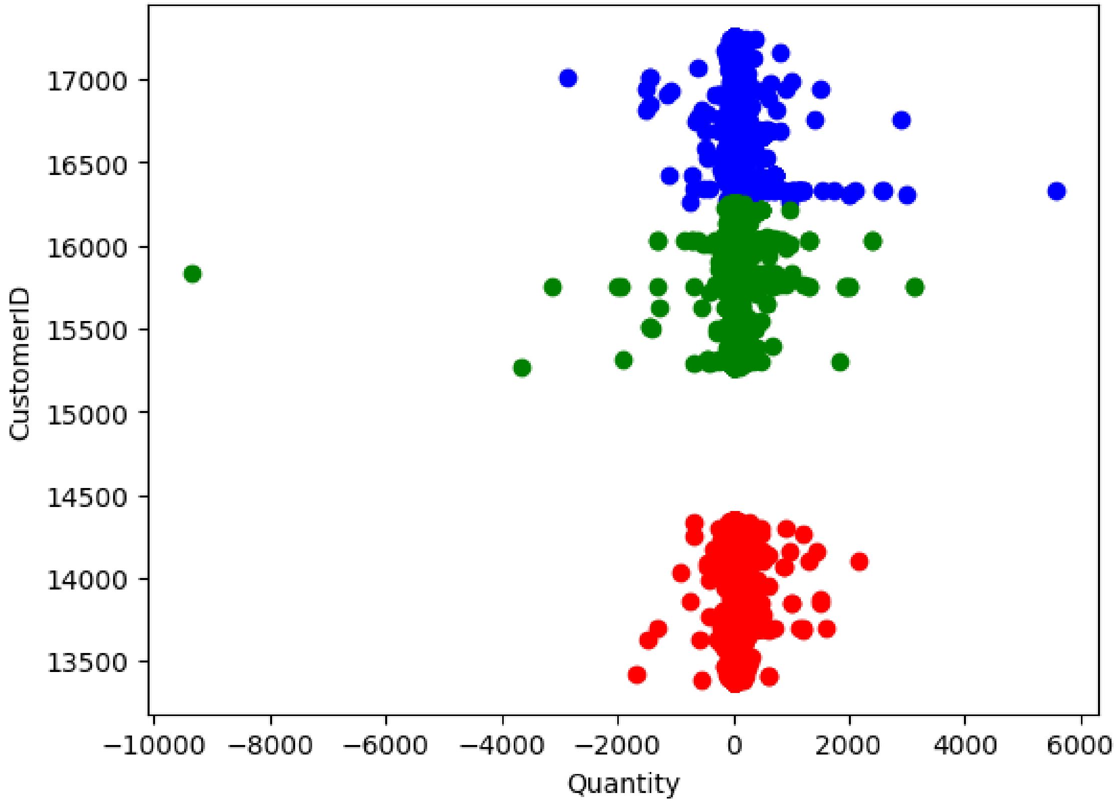
	Quantity	UnitPrice	CustomerID	cluster
0	6	2.55	17850.0	3
1	6	3.39	17850.0	3
2	8	2.75	17850.0	3
3	6	3.39	17850.0	3
4	6	3.39	17850.0	3

In [61]:

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

Out[61]:

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



In [64]:

```
from sklearn.preprocessing import MinMaxScaler
scaler=MinMaxScaler()
scaler.fit(df[["CustomerID"]])
df["CustomerID"]=scaler.transform(df[["CustomerID"]])
df.head()
```

```
C:\Users\Svijayalakshmi\AppData\Local\Temp\ipykernel_23292\2760179153.py:
4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  df["CustomerID"]=scaler.transform(df[["CustomerID"]])
```

Out[64]:

	Quantity	UnitPrice	CustomerID	cluster
0	6	2.55	0.926443	3
1	6	3.39	0.926443	3
2	8	2.75	0.926443	3
3	6	3.39	0.926443	3
4	6	3.39	0.926443	3

In [65]:

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

```
C:\Users\Svijayalakshmi\AppData\Local\Programs\Python\Python311\Lib\site-p
ackages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value o
f `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[65]:

```
array([0, 0, 0, ..., 0, 0, 0])
```

In [66]:

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

```
C:\Users\Svijayalakshmi\AppData\Local\Temp\ipykernel_23292\2515908307.py:
1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  df["New Cluster"]=y_predicted
```

Out[66]:

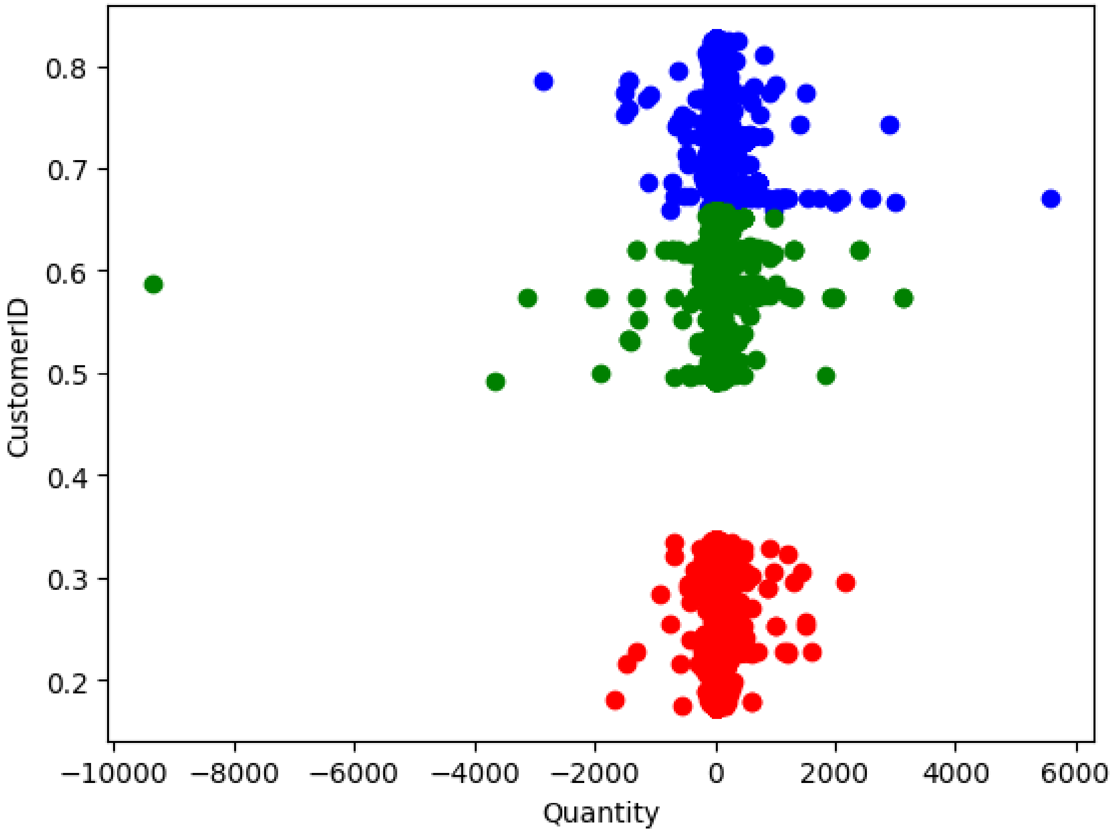
	Quantity	UnitPrice	CustomerID	cluster	New Cluster
0	6	2.55	0.926443	3	0
1	6	3.39	0.926443	3	0
2	8	2.75	0.926443	3	0
3	6	3.39	0.926443	3	0
4	6	3.39	0.926443	3	0

In [67]:

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

Out[67]:

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



In [68]:

```
km.cluster_centers_
```

Out[68]:

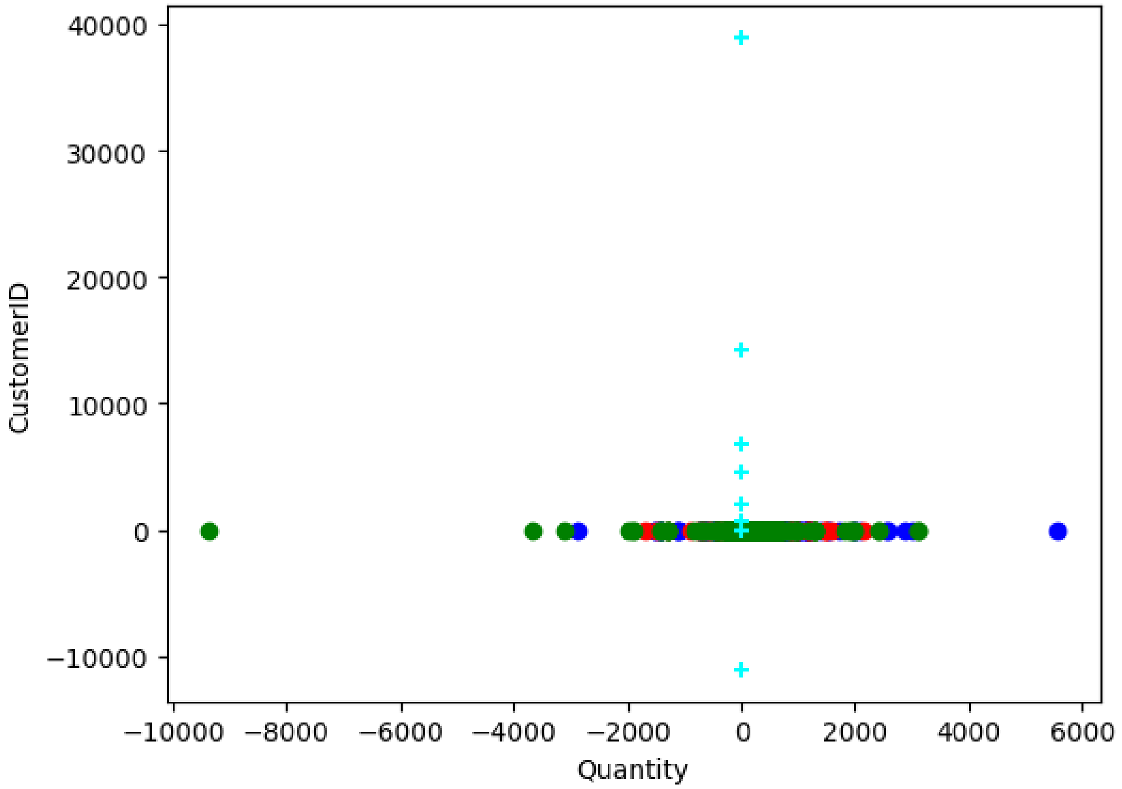
```
array([[ 4.92676188e-01,  3.54270543e+00],
       [ 7.11982644e-01,  1.42139478e+04],
       [ 5.13339505e-01,  6.72797062e+03],
       [ 4.63221680e-01,  3.89700000e+04],
       [ 8.70223868e-01, -1.10620600e+04],
       [ 3.99012512e-01,  2.00749111e+03],
       [ 4.19964420e-01,  6.84400259e+02],
       [ 5.13767736e-01,  4.62258059e+03]])
```

In [71]:

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

Out[71]:

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



In [72]:

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

ELBOW METHOD

In [74]:

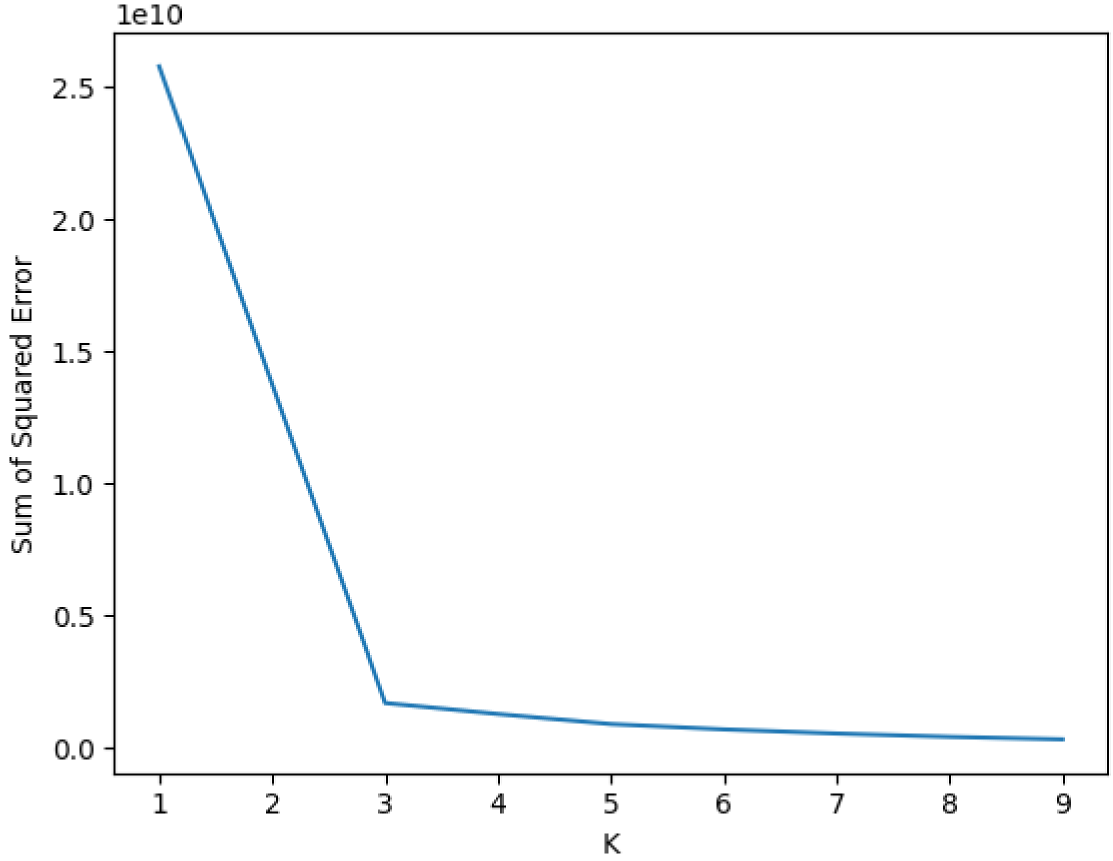
```
for k in k_rng:
    km=KMeans(n_clusters=k)
    km.fit(df[["Quantity","CustomerID"]])
    sse.append(km.inertia_)
print(sse)
plt.plot(k_rng,sse)
plt.xlabel("K")
plt.ylabel("Sum of Squared Error")
```

```
C:\Users\Svijayalakshmi\AppData\Local\Programs\Python\Python311\Lib\site-p
ackages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value o
f `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\Svijayalakshmi\AppData\Local\Programs\Python\Python311\Lib\site-p
ackages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value o
f `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\Svijayalakshmi\AppData\Local\Programs\Python\Python311\Lib\site-p
ackages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value o
f `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\Svijayalakshmi\AppData\Local\Programs\Python\Python311\Lib\site-p
ackages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value o
f `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\Svijayalakshmi\AppData\Local\Programs\Python\Python311\Lib\site-p
ackages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value o
f `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\Svijayalakshmi\AppData\Local\Programs\Python\Python311\Lib\site-p
ackages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value o
f `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\Svijayalakshmi\AppData\Local\Programs\Python\Python311\Lib\site-p
ackages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value o
f `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\Svijayalakshmi\AppData\Local\Programs\Python\Python311\Lib\site-p
ackages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value o
f `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\Svijayalakshmi\AppData\Local\Programs\Python\Python311\Lib\site-p
ackages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value o
f `n_init` will change from 10 to 'auto' in 1.4. Set the value of `n_init`
explicitly to suppress the warning
  warnings.warn(

[25772861053.693607, 13724779146.521996, 1682716588.6219149, 1266025677.06
92048, 884145825.7167574, 683157159.2944075, 526587267.9317037, 400854397.
2494503, 305748537.3068115]
```

```
Out[74]:

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



```
In [ ]:
```