Problem Statement

The Transactions made by a UK-based,registered,non-store online retailer betweem December 1,2010,and December 9,2011, are all included in the transactional 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.compaly objective using the global online retail dataset, we will design a clustering model and select the ideal group of clients for the business to terget.

```
IMPORTING LIBRARIES
```

In [7]:

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

In [8]:

```
data=pd.read_csv(r"C:\Users\shaha\OneDrive\Desktop\Excel\Online Retail.csv")
data
```

Out[8]:

	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
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0
...	...	...	...	...	...	...	...
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 [9]:

```
data.head()
```

Out[9]:

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

In [10]:

```
data.tail()
```

Out[10]:

	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 [11]:

```
data.shape
```

Out[11]:

```
(541909, 8)
```

In [12]:

```
data.describe()
```

Out[12]:

	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 [13]:

```
data.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 [14]:

```
data.isnull().sum()
```

Out[14]:

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

In [15]:

```
data.dropna(inplace=True)
```

In [16]:

```
data.isnull().sum()
```

Out[16]:

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

In [17]:

```
data.columns
```

Out[17]:

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

In [18]:

```
data['CustomerID'].value_counts()
```

Out[18]:

```
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 [19]:

```
data['UnitPrice'].value_counts()
```

Out[19]:

```
UnitPrice
1.25      46555
1.65      37503
2.95      27211
0.85      26396
0.42      22032
          ... 
3.56          1
4.37          1
6.89          1
0.98          1
224.69        1
Name: count, Length: 620, dtype: int64
```

In [20]:

```
data['Quantity'].value_counts()
```

Out[20]:

```
Quantity
 1        73314
 12       60033
 2        58003
 6        37688
 4        32183
          ... 
 828          1
 560          1
-408          1
 512          1
-80995        1
Name: count, Length: 436, dtype: int64
```

In [21]:

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

Out[21]:

```
▾ KMeans
KMeans()
```

In [22]:

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

```
C:\Users\shaha\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
```

Out[22]:

```
array([4, 4, 4, ..., 2, 2, 2])
```

In [23]:

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

Out[23]:

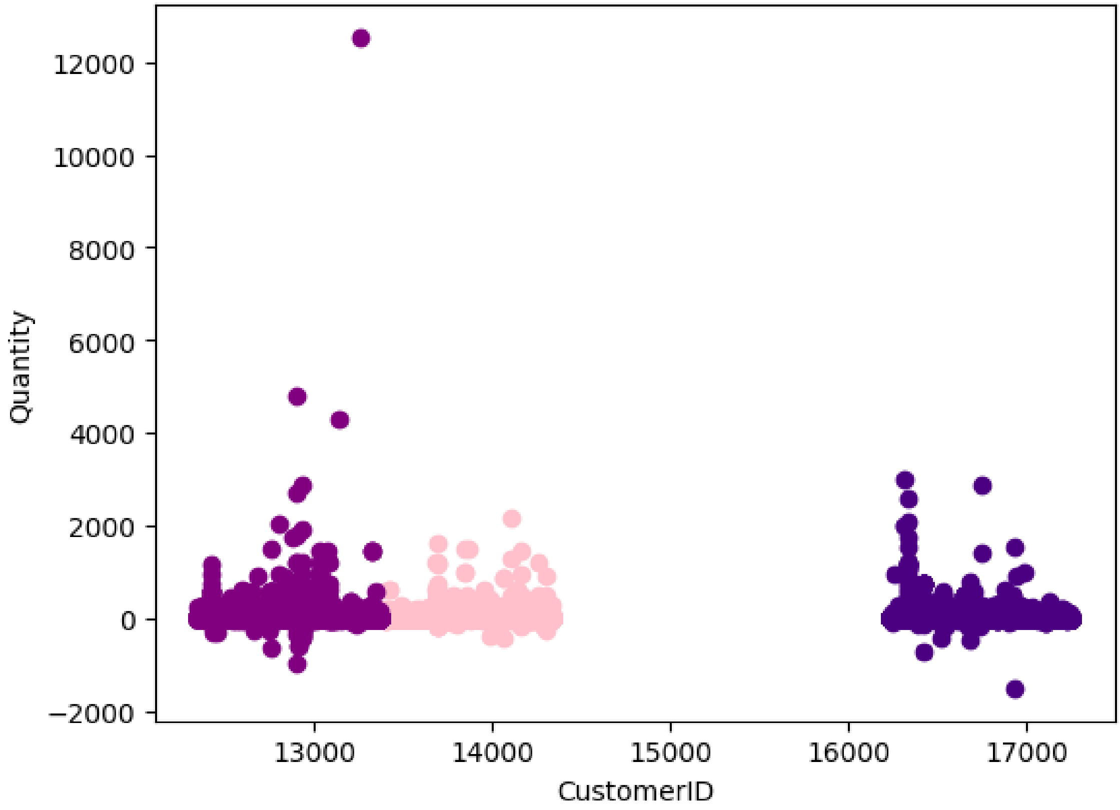
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	Unitec Kingdom
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	Unitec Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT 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 [24]:

```
data1=data[data.cluster==0]
data2=data[data.cluster==1]
data3=data[data.cluster==2]
plt.scatter(data1["CustomerID"],data1["Quantity"],color="pink")
plt.scatter(data2["CustomerID"],data2["Quantity"],color="indigo")
plt.scatter(data3["CustomerID"],data3["Quantity"],color="purple")
plt.xlabel("CustomerID")
plt.ylabel("Quantity")
```

Out[24]:

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



In [25]:

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

Out[25]:

	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	17850.0	Unitec Kingdon
1	536365	71053	WHITE METAL LANTERN	0.500037	01-12-2010 08:26	3.39	17850.0	Unitec Kingdon
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	0.500049	01-12-2010 08:26	2.75	17850.0	Unitec Kingdon
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	0.500037	01-12-2010 08:26	3.39	17850.0	Unitec Kingdon
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	0.500037	01-12-2010 08:26	3.39	17850.0	Unitec Kingdon

In [26]:

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

Out[26]:

	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	Unitec Kingdon
1	536365	71053	WHITE METAL LANTERN	0.500037	01-12-2010 08:26	3.39	0.926443	Unitec Kingdon
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	0.500049	01-12-2010 08:26	2.75	0.926443	Unitec Kingdon
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	0.500037	01-12-2010 08:26	3.39	0.926443	Unitec Kingdon
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	0.500037	01-12-2010 08:26	3.39	0.926443	Unitec Kingdon

In [27]:

```
km=KMeans()
```

In [28]:

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

```
C:\Users\shaha\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
```

Out[28]:

```
array([6, 6, 6, ..., 4, 4, 4])
```

In [29]:

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

Out[29]:

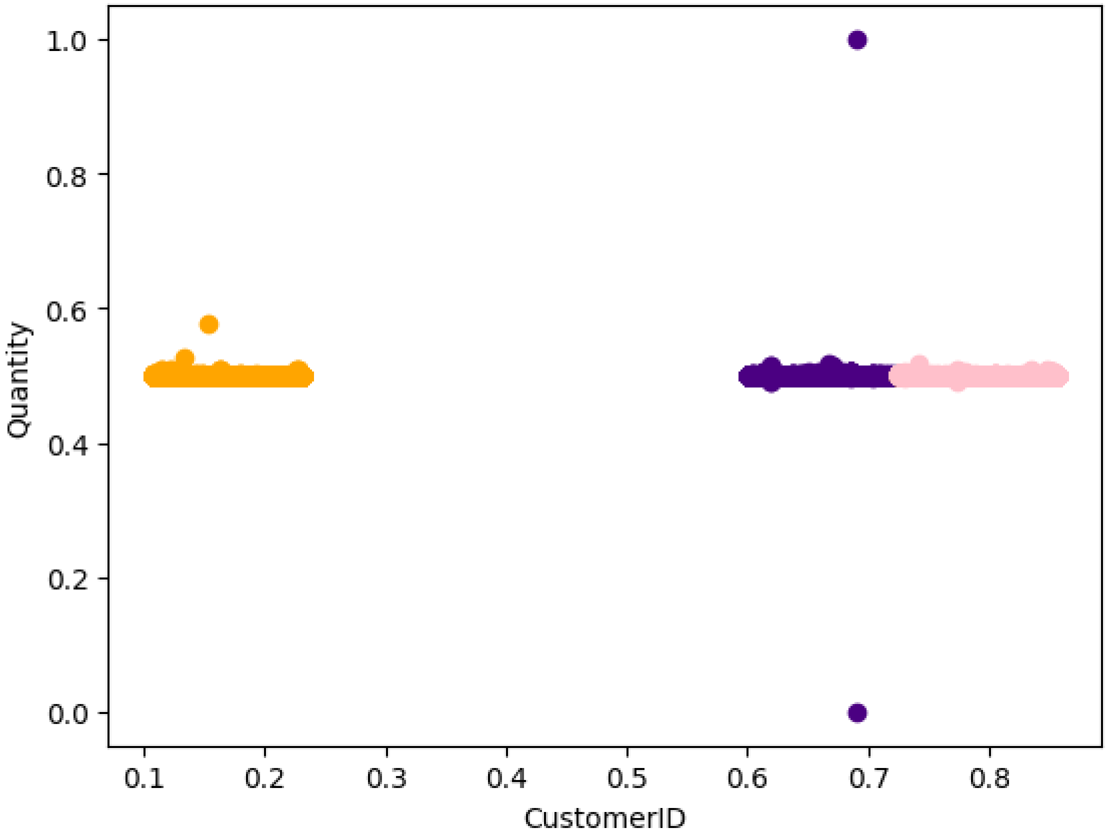
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	0.500037	01-12-2010 08:26	2.55	0.926443	Unitec Kingdon
1	536365	71053	WHITE METAL LANTERN	0.500037	01-12-2010 08:26	3.39	0.926443	Unitec Kingdon
2	536365	84406B	CREAM CUPID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

In [30]:

```
data1=data[data["New Cluster"]==0]
data2=data[data["New Cluster"]==1]
data3=data[data["New Cluster"]==2]
plt.scatter(data1["CustomerID"],data1["Quantity"],color="indigo")
plt.scatter(data2["CustomerID"],data2["Quantity"],color="orange")
plt.scatter(data3["CustomerID"],data3["Quantity"],color="pink")
plt.xlabel("CustomerID")
plt.ylabel("Quantity")
```

Out[30]:

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



In [31]:

```
km.cluster_centers_
```

Out[31]:

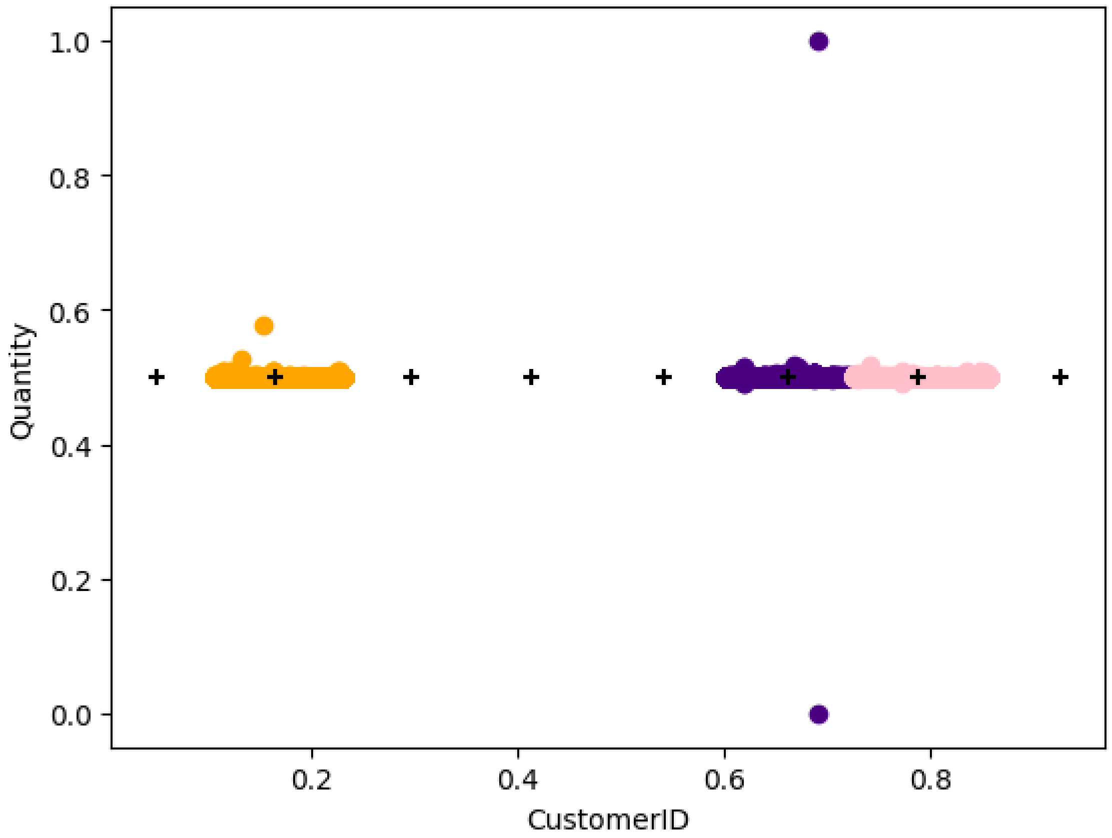
```
array([[0.6618957 , 0.50007355],
       [0.16473511, 0.50008211],
       [0.78776778, 0.50006619],
       [0.41321305, 0.50007252],
       [0.05090876, 0.50009106],
       [0.29757081, 0.50007491],
       [0.92462477, 0.5000745 ],
       [0.54075001, 0.50006319]])
```

In [32]:

```
data1=data[data["New Cluster"]==0]
data2=data[data["New Cluster"]==1]
data3=data[data["New Cluster"]==2]
plt.scatter(data1["CustomerID"],data1["Quantity"],color="indigo")
plt.scatter(data2["CustomerID"],data2["Quantity"],color="orange")
plt.scatter(data3["CustomerID"],data3["Quantity"],color="pink")
plt.scatter(km.cluster_centers_[:,0],km.cluster_centers_[:,1],color="black",marker="+")
plt.xlabel("CustomerID")
plt.ylabel("Quantity")
```

Out[32]:

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



In [33]:

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

In [34]:

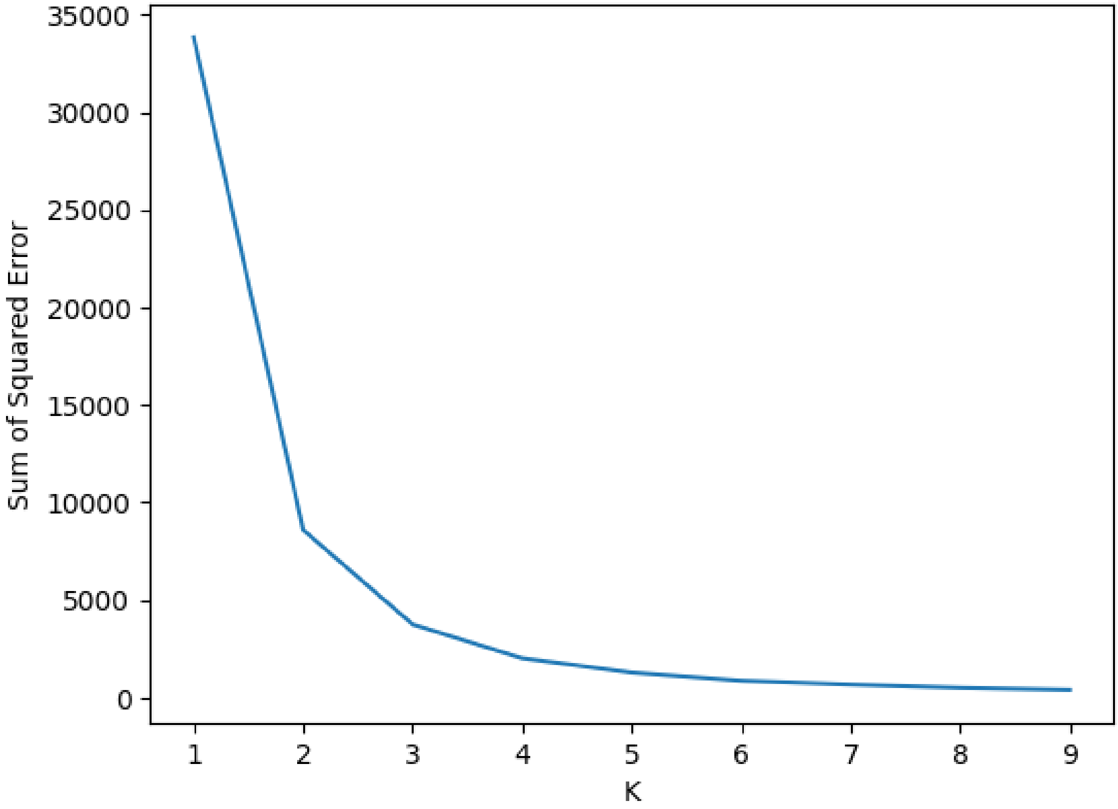
```
for k in k_rng:
    km=KMeans(n_clusters=k)
    km.fit(data[["CustomerID","Quantity"]])
    sse.append(km.inertia_)
print(sse)
plt.plot(k_rng,sse)
plt.xlabel("K")
plt.ylabel("Sum of Squared Error")
```

```
C:\Users\shaha\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\shaha\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\shaha\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\shaha\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\shaha\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\shaha\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\shaha\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\shaha\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\shaha\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(

[33847.22708730174, 8593.142723177269, 3751.717856227737, 2018.33180601552
7, 1286.5232589622271, 868.984760290506, 672.3848200238324, 503.9464808143
947, 398.10153578666166]
```

```
Out[34]:

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



CONCLUSION

From the given Online Retail dataset,Here we have created our final model with 3 clusters and added our cluster labels obtained from kmeans to our Dataframe consisting of Unique customers.With the help of Scatterplots we can visualize the clusters formed on different features

```
In [ ]:
```