ONLINE RETAIL

k-mean cluster

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.

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

DATA COLLECTION

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

Out[2]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	Custome
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	1785
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	1785
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	1785
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	1785
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	1785
...	...	...	...	...	...	...	
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	09-12-2011 12:50	0.85	1268
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	09-12-2011 12:50	2.10	1268
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	09-12-2011 12:50	4.15	1268
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	09-12-2011 12:50	4.15	1268
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	09-12-2011 12:50	4.95	1268

541909 rows × 8 columns

DATA CLEANING

In [3]:
```
df.head(10)
```

Out[3]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	C
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	K
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	K
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	K
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	K
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	K
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	01-12-2010 08:26	7.65	17850.0	K
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	01-12-2010 08:26	4.25	17850.0	K
7	536366	22633	HAND WARMER UNION JACK	6	01-12-2010 08:28	1.85	17850.0	K
8	536366	22632	HAND WARMER RED POLKA DOT	6	01-12-2010 08:28	1.85	17850.0	K
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	01-12-2010 08:34	1.69	13047.0	K

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

Out[4]:

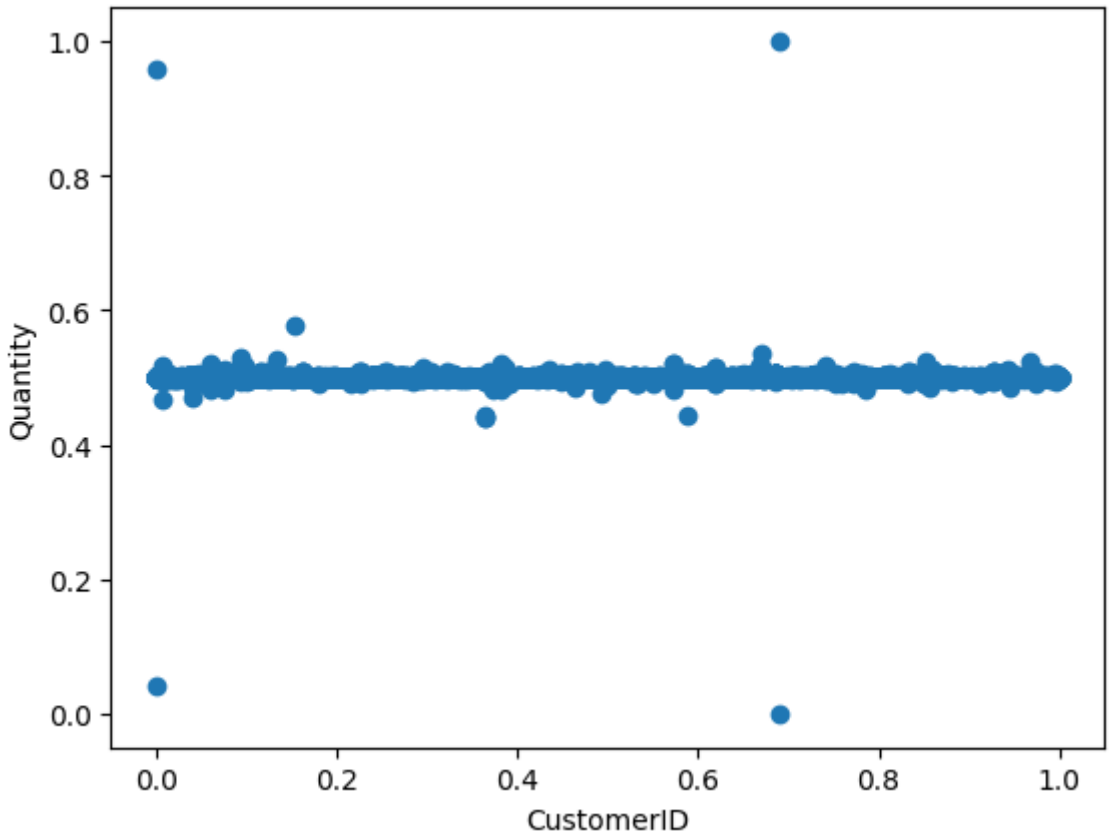
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	Custome
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	09-12-2011 12:50	0.85	1268
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	09-12-2011 12:50	2.10	1268
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	09-12-2011 12:50	4.15	1268
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	09-12-2011 12:50	4.15	1268
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	09-12-2011 12:50	4.95	1268

In [5]:
```
df['Description'].value_counts()
```

Out[5]:
```
Description
WHITE HANGING HEART T-LIGHT HOLDER     2369
REGENCY CAKESTAND 3 TIER               2200
JUMBO BAG RED RETROSPOT                2159
PARTY BUNTING                          1727
LUNCH BAG RED RETROSPOT                1638
                                       ...
Missing                                   1
historic computer difference?....se       1
DUSTY PINK CHRISTMAS TREE 30CM            1
WRAP BLUE RUSSIAN FOLKART                 1
PINK BERTIE MOBILE PHONE CHARM            1
Name: count, Length: 4223, dtype: int64
```

In [34]:
```
plt.scatter(df["CustomerID"],df["Quantity"])
plt.xlabel("CustomerID")
plt.ylabel("Quantity")
```

Out[34]:
```
Text(0, 0.5, 'Quantity')
```



In [35]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 10 columns):
 #   Column       Non-Null Count   Dtype  
---  ------       --------------   -----  
 0   InvoiceNo    541909 non-null  object 
 1   StockCode    541909 non-null  object 
 2   Description  541909 non-null  object 
 3   Quantity     541909 non-null  float64
 4   InvoiceDate  541909 non-null  object 
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   541909 non-null  float64
 7   Country      541909 non-null  object 
 8   cluster      541909 non-null  int32  
 9   New Cluster  541909 non-null  int32  
dtypes: float64(3), int32(2), object(5)
memory usage: 37.2+ MB
```

In [36]:
```
df.isnull().sum()
```

Out[36]:
```
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
cluster        0
New Cluster    0
dtype: int64
```

In [37]:
```
df.fillna(method='ffill',inplace=True)
```

In [38]:
```
df.isnull().sum()
```

Out[38]:
```
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
cluster        0
New Cluster    0
dtype: int64
```

In [39]:
```
from sklearn.cluster import KMeans
```

In [40]:
```
km=KMeans()
km
```

Out[40]:
```
KMeans()
```

In [41]:
```
y_predicted=km.fit_predict(df[["CustomerID","Quantity"]])
y_predicted
```

```
C:\Users\chait\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 wa
rning
  warnings.warn(
```

Out[41]:
```
array([0, 0, 0, ..., 2, 2, 2])
```

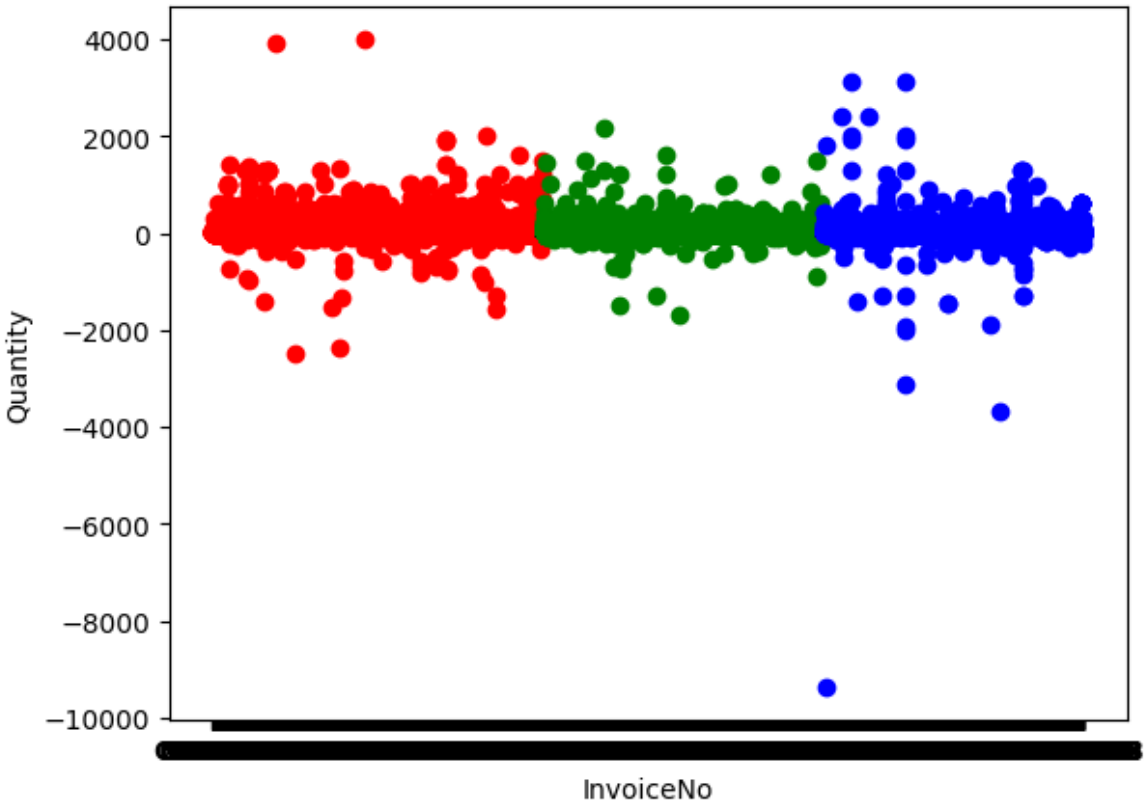
In [42]:
```
df["cluster"]=y_predicted
df.head()
```

Out[42]:

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

In [17]:
```
df1=df[df.cluster==0]
df2=df[df.cluster==1]
df3=df[df.cluster==2]
plt.scatter(df1["InvoiceNo"],df1["Quantity"],color="red")
plt.scatter(df2["InvoiceNo"],df2["Quantity"],color="green")
plt.scatter(df3["InvoiceNo"],df3["Quantity"],color="blue")
plt.xlabel("InvoiceNo")
plt.ylabel("Quantity")
```

Out[17]:
```
Text(0, 0.5, 'Quantity')
```



In [19]:
```
from sklearn.preprocessing import MinMaxScaler
```

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

Out[20]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	C
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	0.500037	01-12-2010 08:26	2.55	17850.0	K
1	536365	71053	WHITE METAL LANTERN	0.500037	01-12-2010 08:26	3.39	17850.0	K
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	0.500049	01-12-2010 08:26	2.75	17850.0	K
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	0.500037	01-12-2010 08:26	3.39	17850.0	K
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	0.500037	01-12-2010 08:26	3.39	17850.0	K

In [21]:
```
scaler.fit(df[["CustomerID"]])
df["CustomerID"]=scaler.transform(df[["CustomerID"]])
df.head()
```

Out[21]:

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

K-MEAN CLUSTER

In [22]:
```
km=KMeans()
```

In [23]:
```
y_predicted=km.fit_predict(df[["CustomerID","Quantity"]])
y_predicted
```

```
C:\Users\chait\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 wa
rning
  warnings.warn(
```

Out[23]:
```
array([5, 5, 5, ..., 3, 3, 3])
```

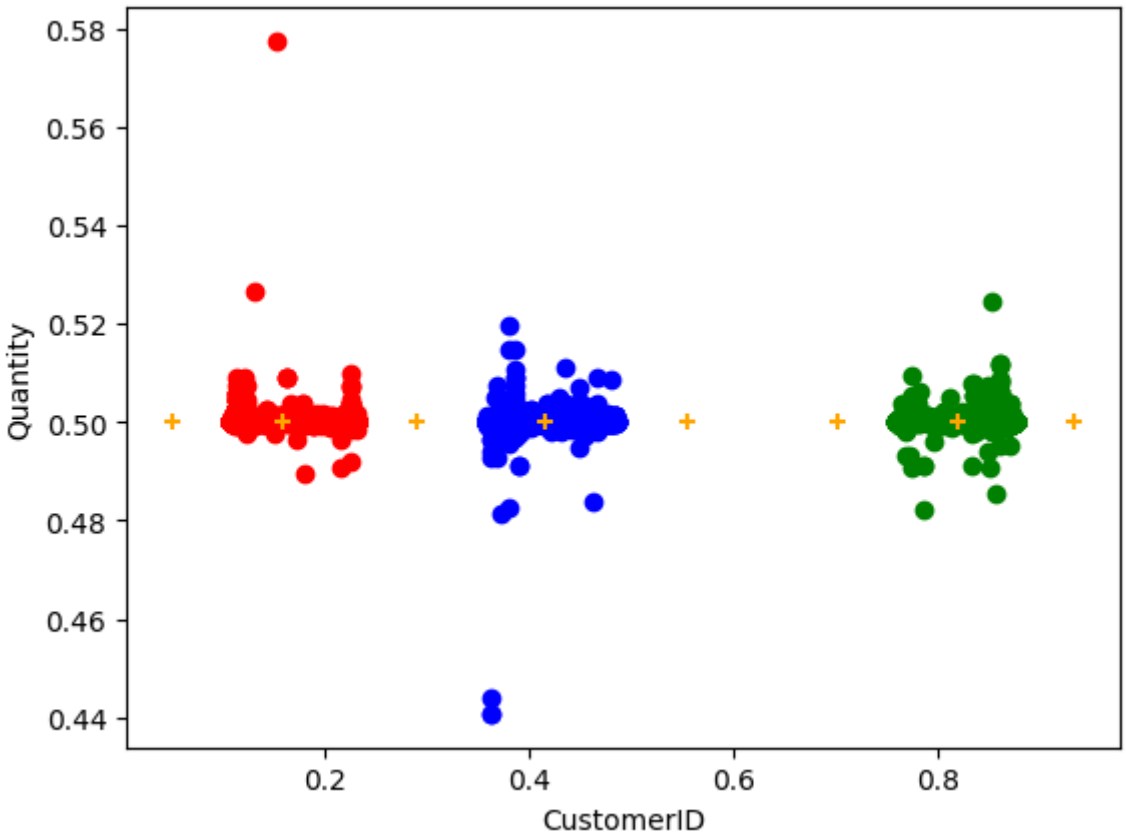
In [24]:
```
df["New Cluster"]=y_predicted
df.head()
```

Out[24]:

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

In [43]:
```
df1=df[df["New Cluster"]==0]
df2=df[df["New Cluster"]==1]
df3=df[df["New Cluster"]==2]
plt.scatter(df1["CustomerID"],df1["Quantity"],color="red")
plt.scatter(df2["CustomerID"],df2["Quantity"],color="green")
plt.scatter(df3["CustomerID"],df3["Quantity"],color="blue")
plt.scatter(km.cluster_centers_[:,0],km.cluster_centers_[:,1],color="orange",mar
plt.xlabel("CustomerID")
plt.ylabel("Quantity")
```

Out[43]:
```
Text(0, 0.5, 'Quantity')
```

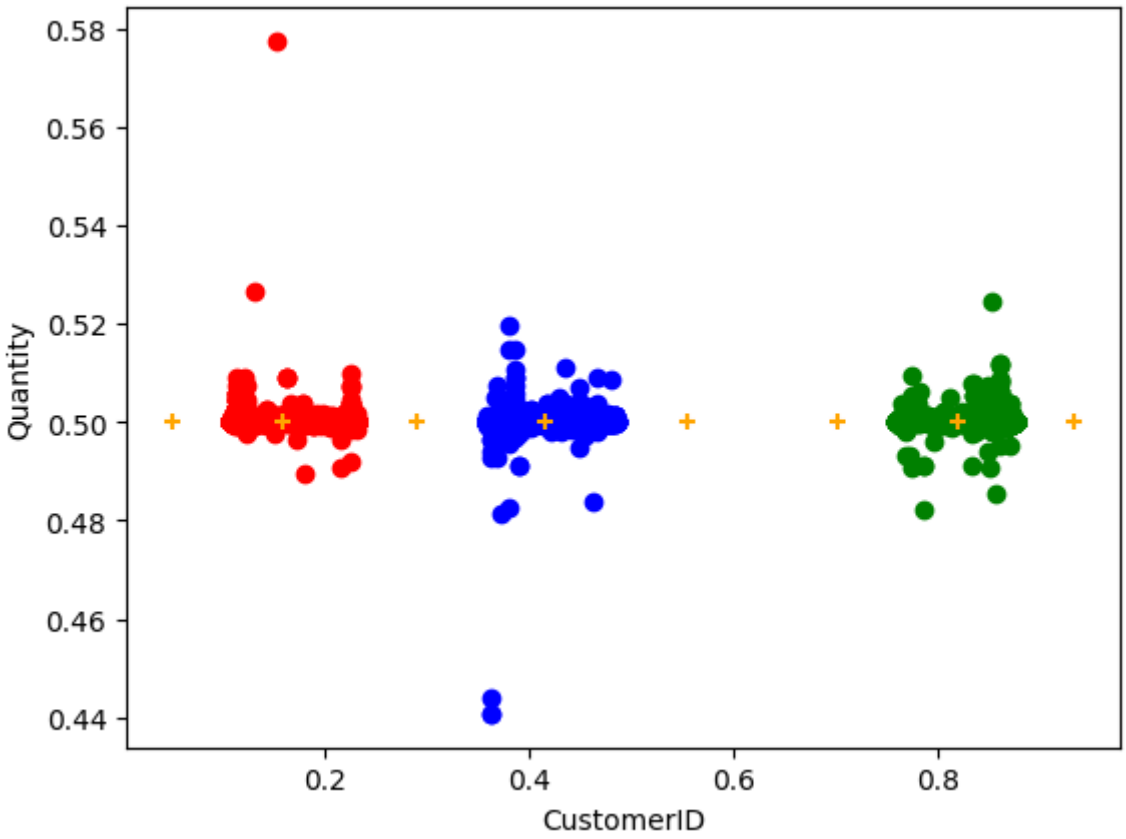


In [44]:
```
km.cluster_centers_
```

Out[44]:
```
array([[0.93301334, 0.50005098],
       [0.41601901, 0.50005993],
       [0.05119252, 0.50006679],
       [0.70125271, 0.50005786],
       [0.1603687 , 0.50005698],
       [0.55455788, 0.50005364],
       [0.81846395, 0.50006031],
       [0.29135039, 0.50006544]])
```

In [45]:
```
df1=df[df["New Cluster"]==0]
df2=df[df["New Cluster"]==1]
df3=df[df["New Cluster"]==2]
plt.scatter(df1["CustomerID"],df1["Quantity"],color="red")
plt.scatter(df2["CustomerID"],df2["Quantity"],color="green")
plt.scatter(df3["CustomerID"],df3["Quantity"],color="blue")
plt.scatter(km.cluster_centers_[:,0],km.cluster_centers_[:,1],color="orange",mar
plt.xlabel("CustomerID")
plt.ylabel("Quantity")
```

Out[45]:
```
Text(0, 0.5, 'Quantity')
```



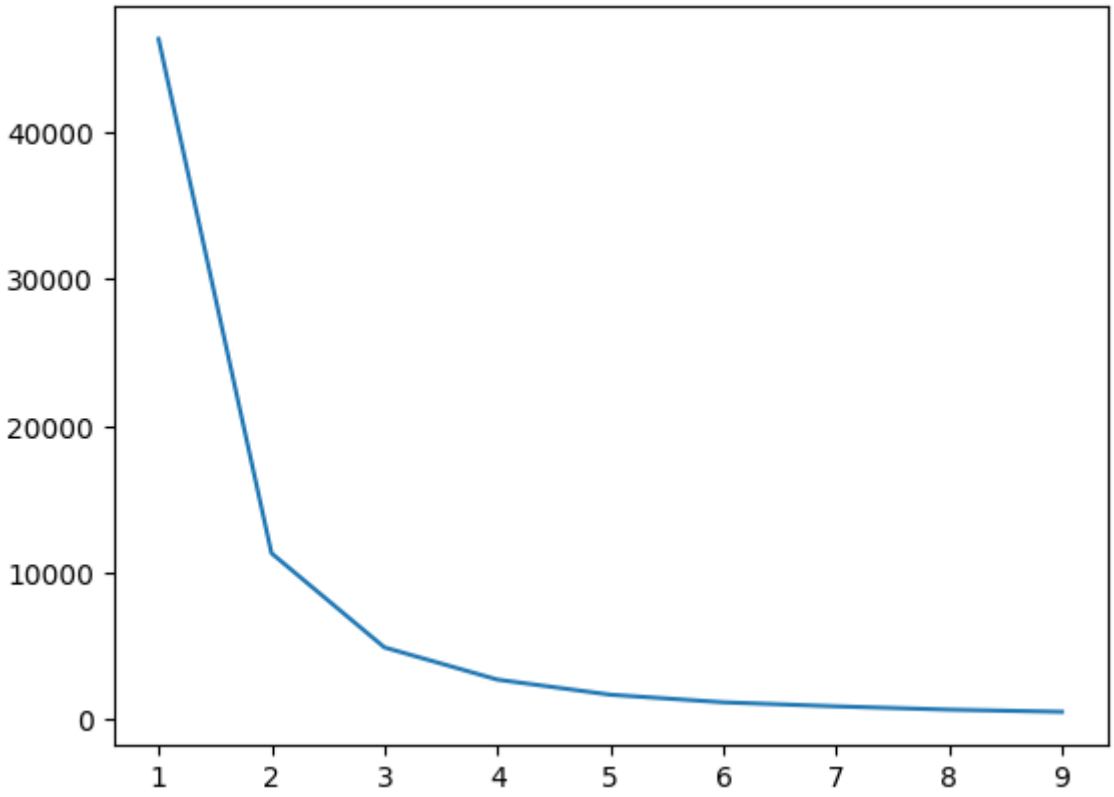
In [46]:
```
k_rng=range(1,10)
se=[]
```

In [47]:
```
for k in k_rng:
    km=KMeans(n_clusters=k)
    km.fit(df[["CustomerID","Quantity"]])
    se.append(km.inertia_)

print(se)
plt.plot(k_rng,se)
```

```
C:\Users\chait\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 wa
rning
  warnings.warn(
C:\Users\chait\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 wa
rning
  warnings.warn(
C:\Users\chait\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 wa
rning
  warnings.warn(
C:\Users\chait\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 wa
rning
  warnings.warn(
C:\Users\chait\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 wa
rning
  warnings.warn(
C:\Users\chait\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 wa
rning
  warnings.warn(
C:\Users\chait\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 wa
rning
  warnings.warn(
C:\Users\chait\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 wa
rning
  warnings.warn(
C:\Users\chait\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 wa
rning
  warnings.warn(
[46374.84553398485, 11336.065820168866, 4918.434064877548, 2723.5191051894612, 16
95.0392229312763, 1178.4630922671322, 902.5668563345514, 676.5249170799557, 528.8
251995247876]
```

```
Out[47]: [<matplotlib.lines.Line2D at 0x1df488840a0>]
```



CONCLUSION

From the above dataset, Online Retail of the data used to take K-Mean cluster method to find the correct form of DataFrame.