

Project - 5 (DATASET: Online Retail) The transactions made by a UKbased, 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 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]: import pandas as pd  
        from matplotlib import pyplot as plt  
        %matplotlib inline
```

```
In [3]: df=pd.read_csv(r"C:\Users\yasodavaidyam\Downloads\online_retail.csv")
df
```

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

541909 rows × 8 columns

Data Cleaning and Preprocessing

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

```
Out[4]:
```

	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 [5]: `df.tail()`

Out[5]:

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

In [8]: `df['InvoiceNo'].value_counts()`

Out[8]: InvoiceNo

573585	1114
581219	749
581492	731
580729	721
558475	705
...	
554023	1
554022	1
554021	1
554020	1
C558901	1

Name: count, Length: 25900, dtype: int64

```
In [10]: df['CustomerID'].value_counts()
```

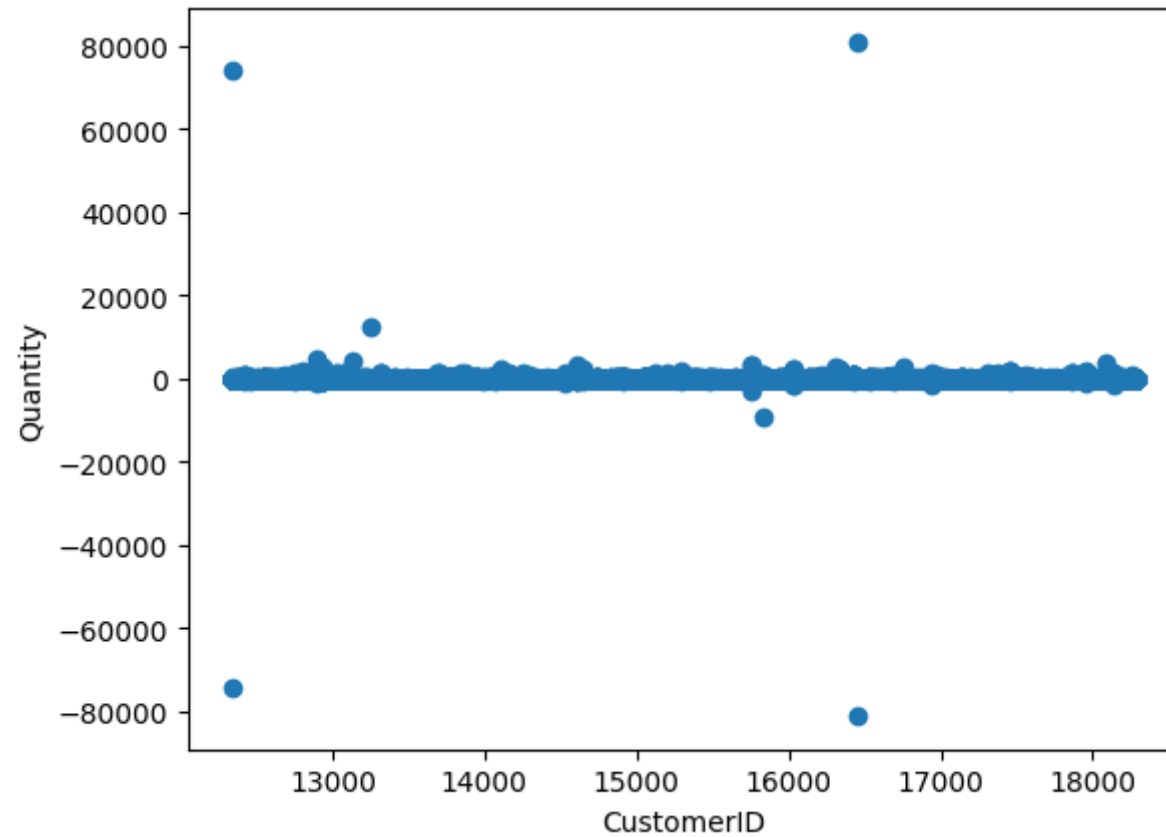
```
Out[10]: CustomerID
17841.0    7983
14911.0    5903
14096.0    5128
12748.0    4642
14606.0    2782
...
15070.0     1
15753.0     1
17065.0     1
16881.0     1
16995.0     1
Name: count, Length: 4372, dtype: int64
```

```
In [11]: df['Quantity'].value_counts()
```

```
Out[11]: 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 [12]: plt.scatter(df["CustomerID"],df["Quantity"])  
plt.xlabel("CustomerID")  
plt.ylabel("Quantity")
```

```
Out[12]: Text(0, 0.5, 'Quantity')
```



```
In [13]: 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 [14]: df.isnull().sum()
```

```
Out[14]: InvoiceNo          0
StockCode          0
Description        1454
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID        135080
Country            0
dtype: int64
```

```
In [15]: df.fillna(method='ffill',inplace=True)
```

```
In [16]: df.fillna(method='bfill',inplace=True)
```

```
In [17]: from sklearn.cluster import KMeans
km=KMeans()
km
```

Out[17]: 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 [18]: y_predicted=km.fit_predict(df[["CustomerID", "Quantity"]])
y_predicted
```

C:\Users\SASIDHAR ROYAL\AppData\Local\Programs\Python\Python311\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 warning
warnings.warn(

Out[18]: array([2, 2, 2, ..., 3, 3, 3])

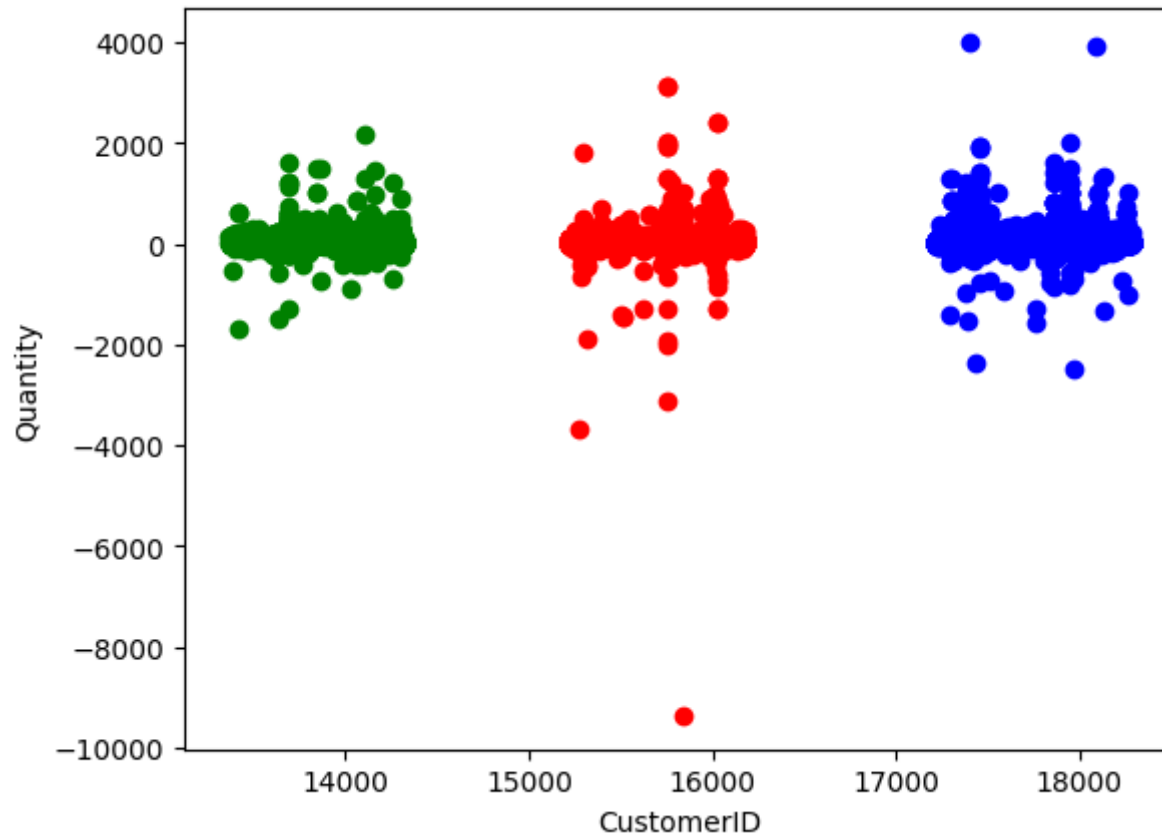
```
In [19]: df["cluster"]=y_predicted
df.head()
```

Out[19]:

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

```
In [20]: df1=df[df.cluster==0]
df2=df[df.cluster==1]
df3=df[df.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.xlabel("CustomerID")
plt.ylabel("Quantity")
```

```
Out[20]: Text(0, 0.5, 'Quantity')
```




```
In [21]: from sklearn.preprocessing import MinMaxScaler
scaler=MinMaxScaler()
scaler.fit(df[["Quantity"]])
df["Quantity"]=scaler.transform(df[["Quantity"]])
df.head()
```

Out[21]:

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

```
In [22]: scaler.fit(df[["CustomerID"]])
df["CustomerID"] = scaler.transform(df[["CustomerID"]])
df
```

```
Out[22]:
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	cluster
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	0.500037	01-12-2010 08:26	2.55	0.926443	United Kingdom	2
1	536365	71053	WHITE METAL LANTERN	0.500037	01-12-2010 08:26	3.39	0.926443	United Kingdom	2
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	0.500049	01-12-2010 08:26	2.75	0.926443	United Kingdom	2
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	0.500037	01-12-2010 08:26	3.39	0.926443	United Kingdom	2
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	0.500037	01-12-2010 08:26	3.39	0.926443	United Kingdom	2
...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	0.500074	09-12-2011 12:50	0.85	0.056219	France	3
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	0.500037	09-12-2011 12:50	2.10	0.056219	France	3
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	0.500025	09-12-2011 12:50	4.15	0.056219	France	3
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	0.500025	09-12-2011 12:50	4.15	0.056219	France	3
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	0.500019	09-12-2011 12:50	4.95	0.056219	France	3

541909 rows × 9 columns

K-Means Clustering

```
In [23]: km=KMeans()
```

```
In [24]: y_predicted=km.fit_predict(df[["CustomerID", "Quantity"]])
y_predicted
```

C:\Users\SASIDHAR ROYAL\AppData\Local\Programs\Python\Python311\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 warning
warnings.warn(

```
Out[24]: array([5, 5, 5, ..., 4, 4, 4])
```

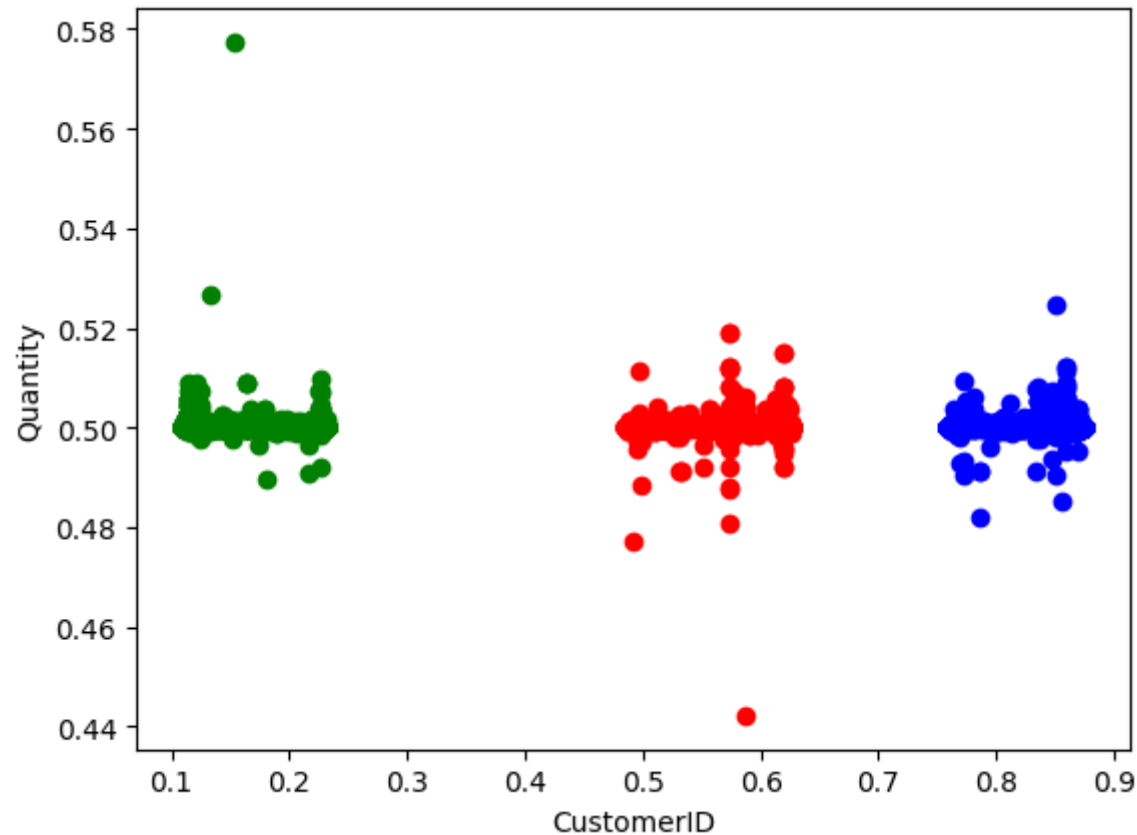
```
In [25]: df["New Cluster"]=y_predicted
df.head()
```

```
Out[25]:
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	cluster	New Cluster
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	0.500037	01-12-2010 08:26	2.55	0.926443	United Kingdom	2	5
1	536365	71053	WHITE METAL LANTERN	0.500037	01-12-2010 08:26	3.39	0.926443	United Kingdom	2	5
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	0.500049	01-12-2010 08:26	2.75	0.926443	United Kingdom	2	5
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	0.500037	01-12-2010 08:26	3.39	0.926443	United Kingdom	2	5
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	0.500037	01-12-2010 08:26	3.39	0.926443	United Kingdom	2	5

```
In [26]: 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.xlabel("CustomerID")
plt.ylabel("Quantity")
```

Out[26]: Text(0, 0.5, 'Quantity')

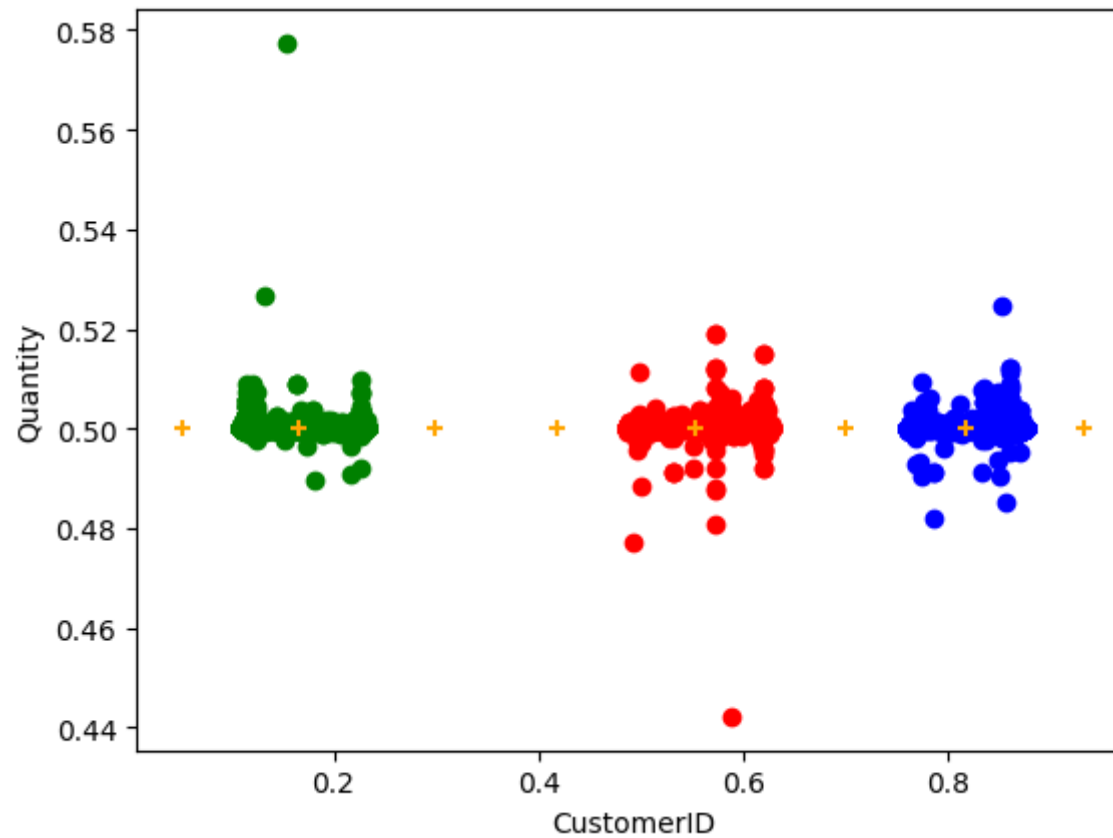


```
In [27]: km.cluster_centers_
```

```
Out[27]: array([[0.55316723, 0.50005385],  
                [0.16549358, 0.50006058],  
                [0.81756576, 0.50005988],  
                [0.41753404, 0.50006083],  
                [0.0515406 , 0.50006705],  
                [0.9328779 , 0.50005088],  
                [0.69962461, 0.50005827],  
                [0.29794168, 0.50006087]])
```

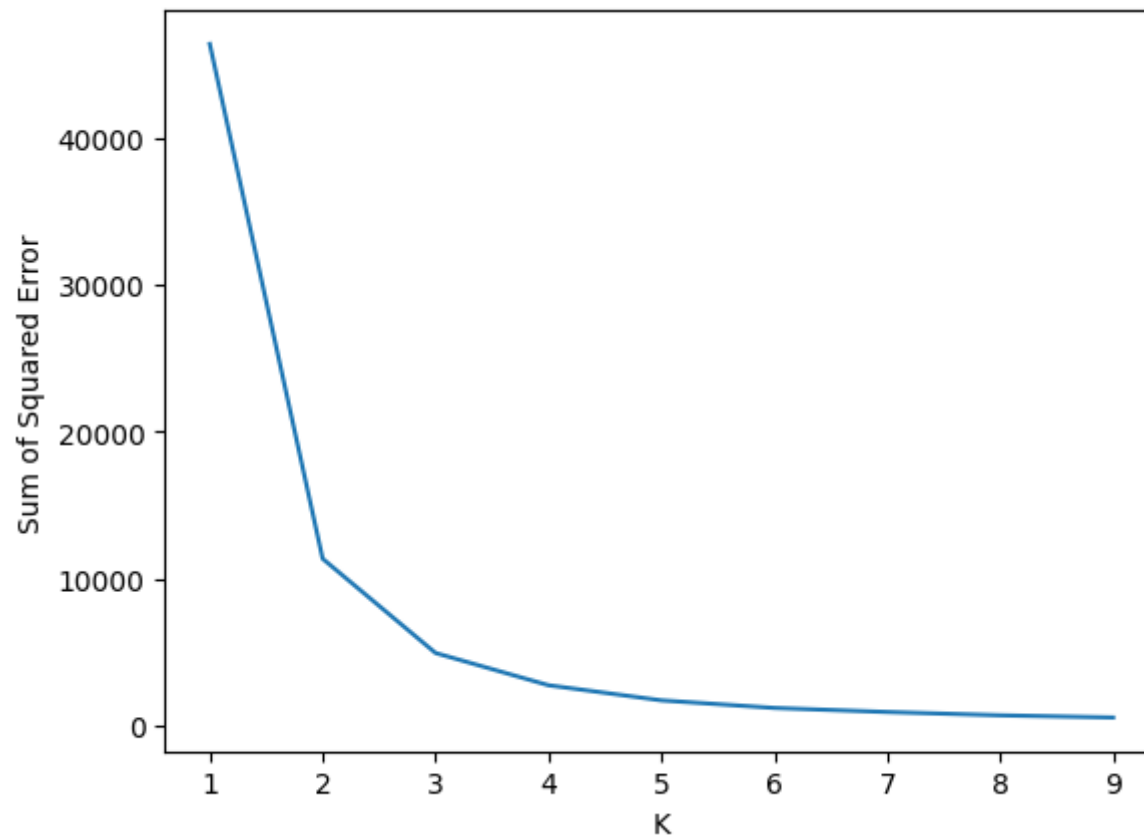
```
In [28]: df1=df[df["New Cluster"]==0]
df2=df[df["New Cluster"]==1]
df3=df[df["New Cluster"]==2]
plt.scatter(df1["CustomerID"],df1["Quantity"],color="red")
plt.scatter(df2["CustomerID"],df2["Quantity"],color="green")
plt.scatter(df3["CustomerID"],df3["Quantity"],color="blue")
plt.scatter(km.cluster_centers_[0],km.cluster_centers_[1],color="orange",marker="+")
plt.xlabel("CustomerID")
plt.ylabel("Quantity")
```

Out[28]: Text(0, 0.5, 'Quantity')



```
In [29]: k_rng=range(1,10)
sse=[]
```

```
In [30]: for k in k_rng:
          km=KMeans(n_clusters=k)
          km.fit(df[["CustomerID", "Quantity"]])
          sse.append(km.inertia_)
          print(sse)
          plt.plot(k_rng, sse)
          plt.xlabel("K")
          plt.ylabel("Sum of Squared Error")
```

CONCLUSION

For the given dataset we use K-means Clustering and done the grouping based on the given data. In the above dataset we will take customer id and quantity based on that we make the clusters. When the K-value is low error rate is more and the K-value is high error rate is very high. so, finally we can conclude the above dataset is best fit for K-Means.

In []: