PROBLEM STATEMENT:- TO DIVIDE THE DATA INTO CLUSTERS BASED ON THE SIMILARITY

In [1]:

```
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression
```

In [2]:

```
df=pd.read_csv(r"C:\Users\hp\Documents\OnlineRetail.csv")
df
```

Out[2]:

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

541909 rows × 7 columns

In [3]:

```
df.head()
```

Out[3]:

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

In [4]:

```
df.tail()
```

Out[4]:

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

In [5]:

```
df.describe()
```

Out[5]:

	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 [6]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 7 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   UnitPrice    541909 non-null  float64
 5   CustomerID   406829 non-null  float64
 6   Country      541909 non-null  object 
dtypes: float64(2), int64(1), object(4)
memory usage: 28.9+ MB
```

In [7]:
```
df.isnull().any()
```

Out[7]:
```
InvoiceNo      False
StockCode      False
Description     True
Quantity       False
UnitPrice      False
CustomerID      True
Country        False
dtype: bool
```

In [8]:
```
df.shape
```

Out[8]:
```
(541909, 7)
```

In [9]:
```
df.fillna(method='ffill',inplace=True)
```

In [10]:
```
df.isnull().sum()
```

Out[10]:
```
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
```

In [11]:
```
del df['InvoiceNo']
```

```
In [12]: df
```

Out[12]:

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

541909 rows × 6 columns

In [13]:

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

Out[13]:

```
        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 [14]:

```
df.shape
```

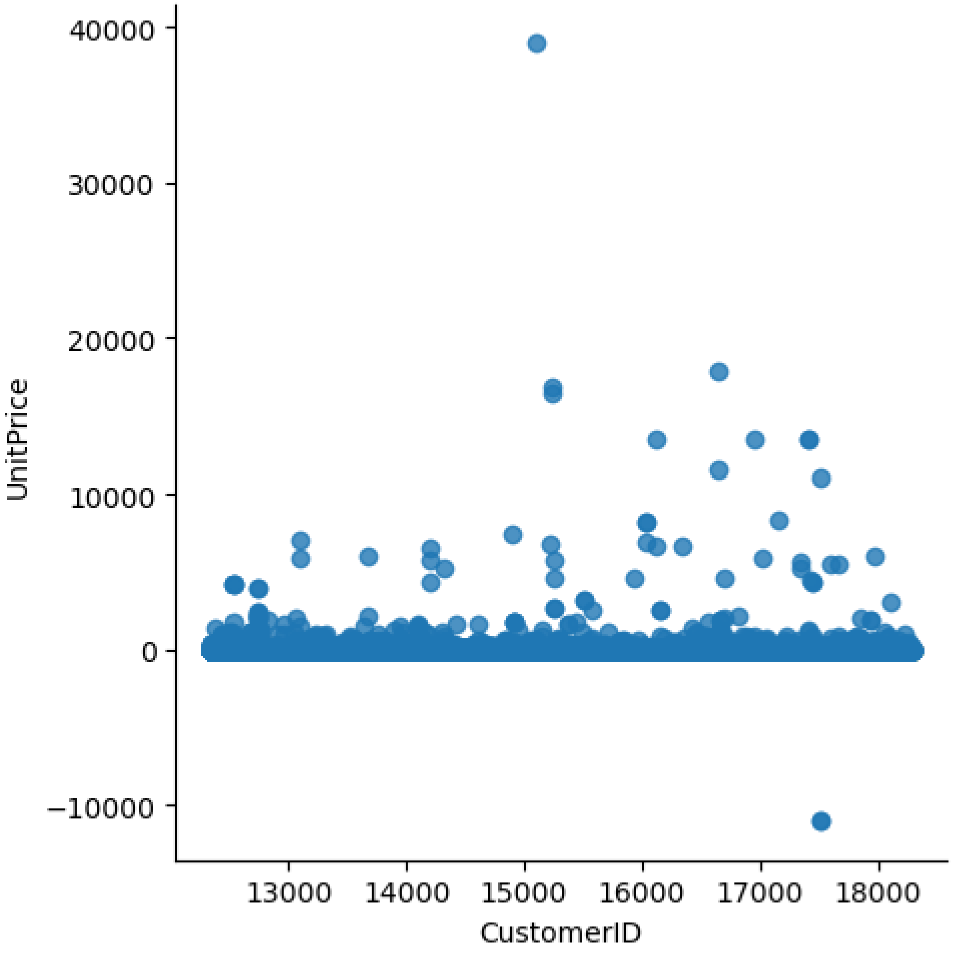
Out[14]:

```
(541909, 3)
```

In [15]:

```
import seaborn as sns
import matplotlib.pyplot as plt
```

```
In [16]: sns.lmplot(x='CustomerID',y='UnitPrice',data=df,order=2,ci=None)

Out[16]: <seaborn.axisgrid.FacetGrid at 0x1a12ba377c0>
```



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

Out[17]:
```
▾ KMeans

KMeans()
```

In [18]:
```
y_predicted=km.fit_predict(df[["CustomerID","UnitPrice"]])
y_predicted
```

```
C:\Users\hp\anaconda3\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([3, 3, 3, ..., 2, 2, 2])
```

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

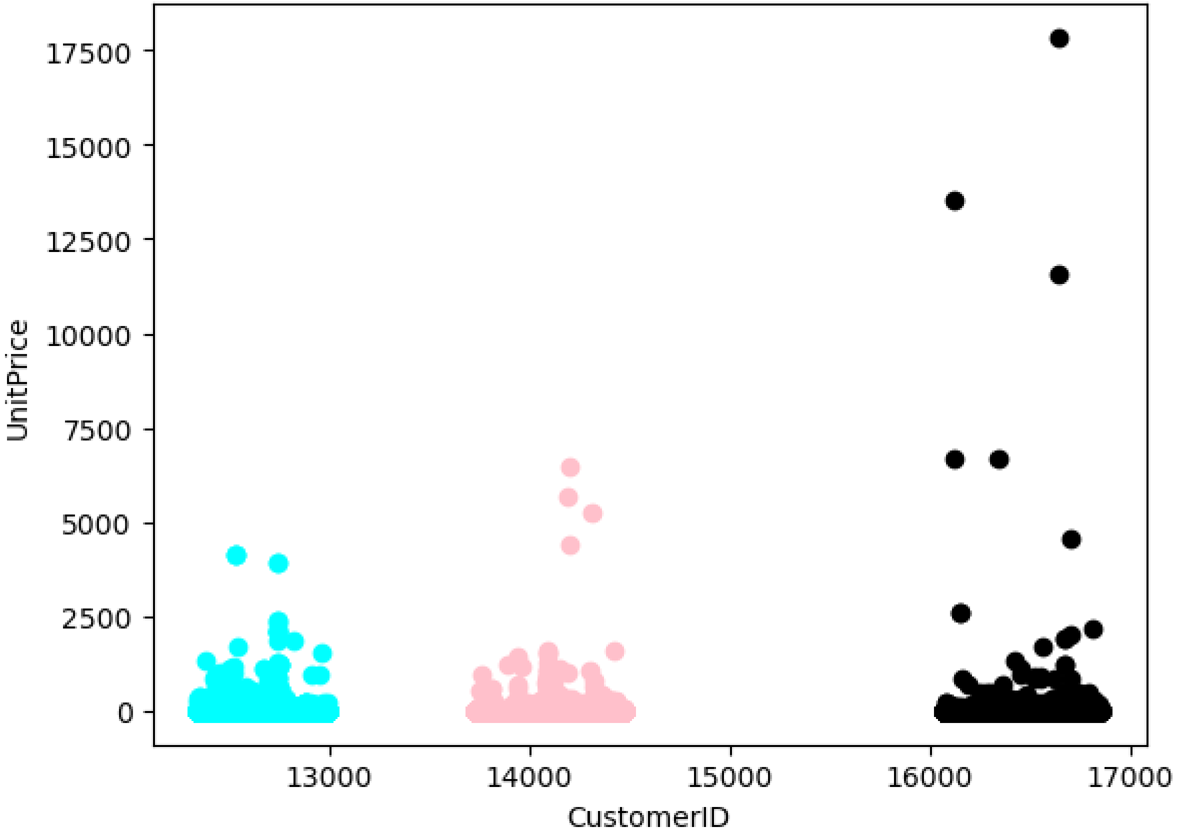
```
C:\Users\hp\AppData\Local\Temp\ipykernel_6424\1084992799.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

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

Out[19]:

	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 [20]: df1=df[df.cluster==0]
         df2=df[df.cluster==1]
         df3=df[df.cluster==2]
         plt.scatter(df1["CustomerID"],df1["UnitPrice"],color="black")
         plt.scatter(df2["CustomerID"],df2["UnitPrice"],color="pink")
         plt.scatter(df3["CustomerID"],df3["UnitPrice"],color="cyan")
         plt.xlabel("CustomerID")
         plt.ylabel("UnitPrice")
```

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



In [21]:

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

```
C:\Users\hp\AppData\Local\Temp\ipykernel_6424\4223297019.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-docs/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["UnitPrice"]=scaler.transform(df[["UnitPrice"]])
```

Out[21]:

	Quantity	UnitPrice	CustomerID	cluster
0	6	0.221150	17850.0	3
1	6	0.221167	17850.0	3
2	8	0.221154	17850.0	3
3	6	0.221167	17850.0	3
4	6	0.221167	17850.0	3

In [22]:

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

```
C:\Users\hp\anaconda3\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[22]:

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

In [23]:

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

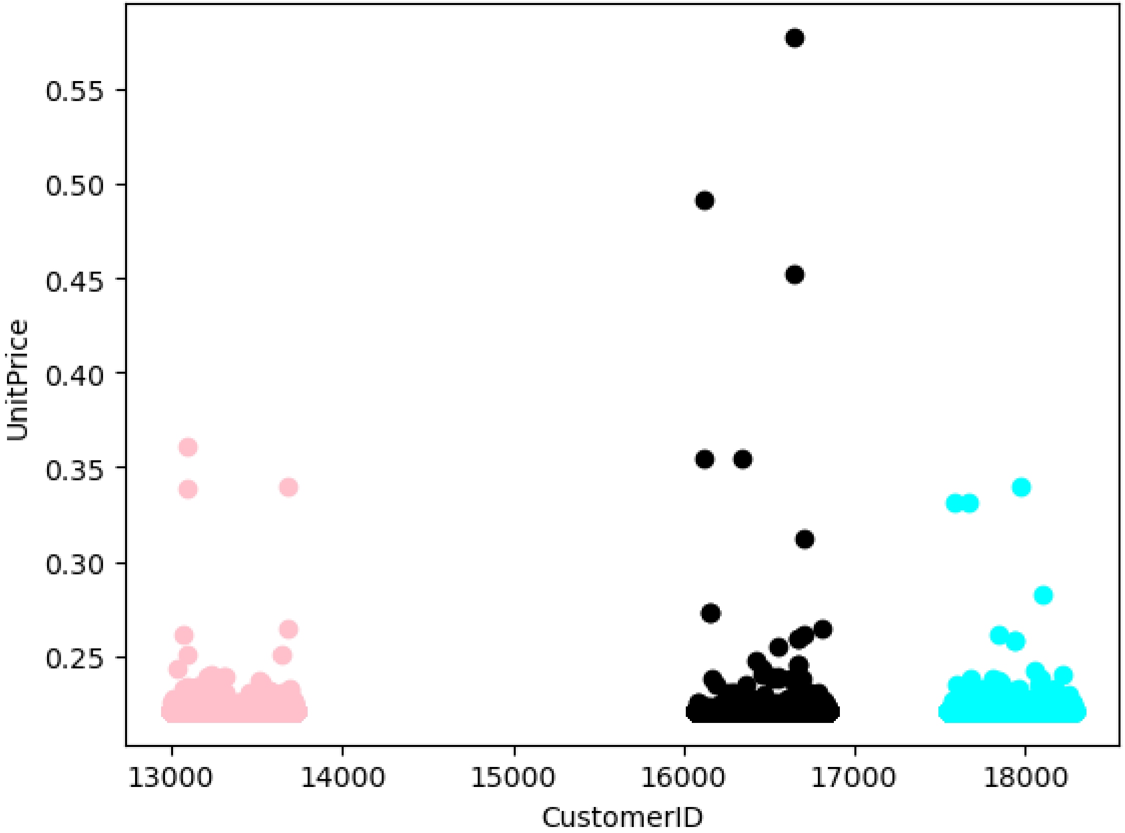
```
C:\Users\hp\AppData\Local\Temp\ipykernel_6424\2515908307.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/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[23]:

	Quantity	UnitPrice	CustomerID	cluster	New Cluster
0	6	0.221150	17850.0	3	2
1	6	0.221167	17850.0	3	2
2	8	0.221154	17850.0	3	2
3	6	0.221167	17850.0	3	2
4	6	0.221167	17850.0	3	2

In [24]:
```
df1=df[df["New Cluster"]==0]
df2=df[df["New Cluster"]==1]
df3=df[df["New Cluster"]==2]
plt.scatter(df1["CustomerID"],df1["UnitPrice"],color="black")
plt.scatter(df2["CustomerID"],df2["UnitPrice"],color="pink")
plt.scatter(df3["CustomerID"],df3["UnitPrice"],color="cyan")
plt.xlabel("CustomerID")
plt.ylabel("UnitPrice")
```

Out[24]:
```
Text(0, 0.5, 'UnitPrice')
```

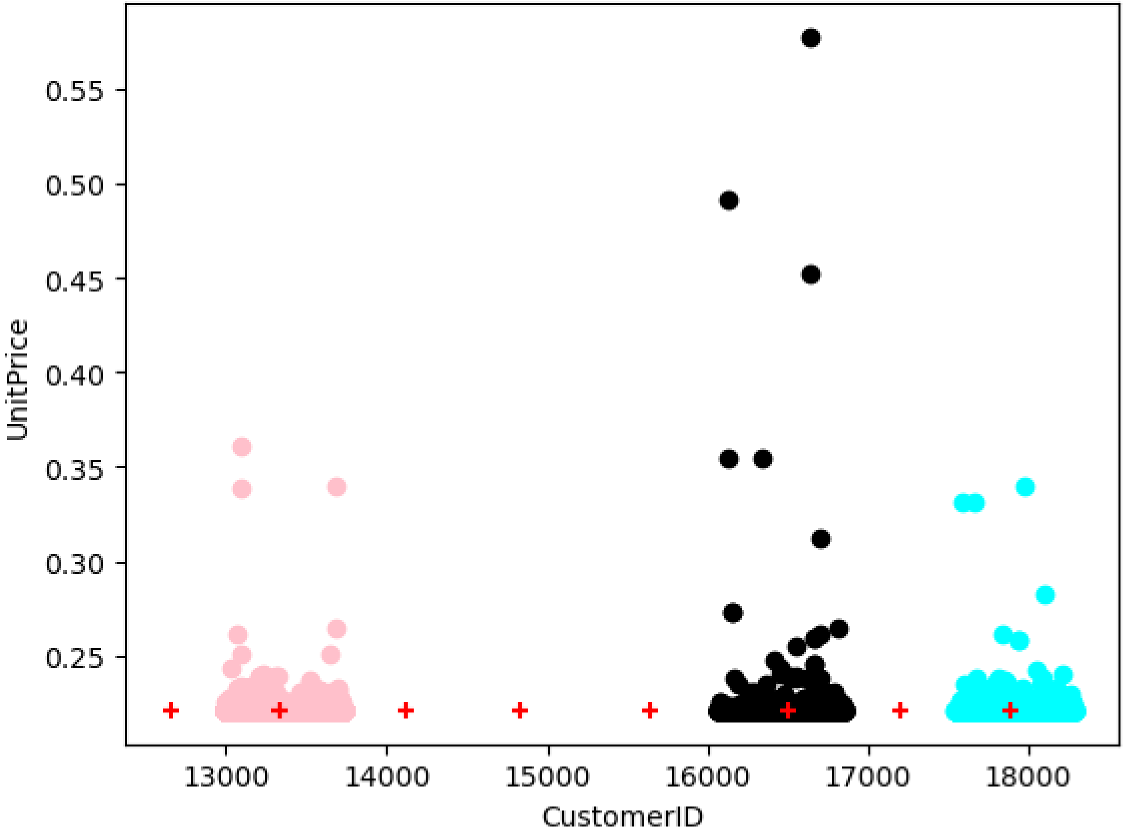


```
In [25]: km.cluster_centers_
```

```
Out[25]: array([[1.65044245e+04, 2.21198131e-01],
                [1.33381329e+04, 2.21184378e-01],
                [1.78889352e+04, 2.21178304e-01],
                [1.48313908e+04, 2.21195936e-01],
                [1.26557810e+04, 2.21202510e-01],
                [1.72043031e+04, 2.21199485e-01],
                [1.41244730e+04, 2.21187467e-01],
                [1.56365669e+04, 2.21187222e-01]])
```

```
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["UnitPrice"],color="black")
         plt.scatter(df2["CustomerID"],df2["UnitPrice"],color="pink")
         plt.scatter(df3["CustomerID"],df3["UnitPrice"],color="cyan")
         plt.scatter(km.cluster_centers_[:,0],km.cluster_centers_[:,1],color="red",marker="+")
         plt.xlabel("CustomerID")
         plt.ylabel("UnitPrice")
```

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



In [27]:

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

ELBOW METHOD:-

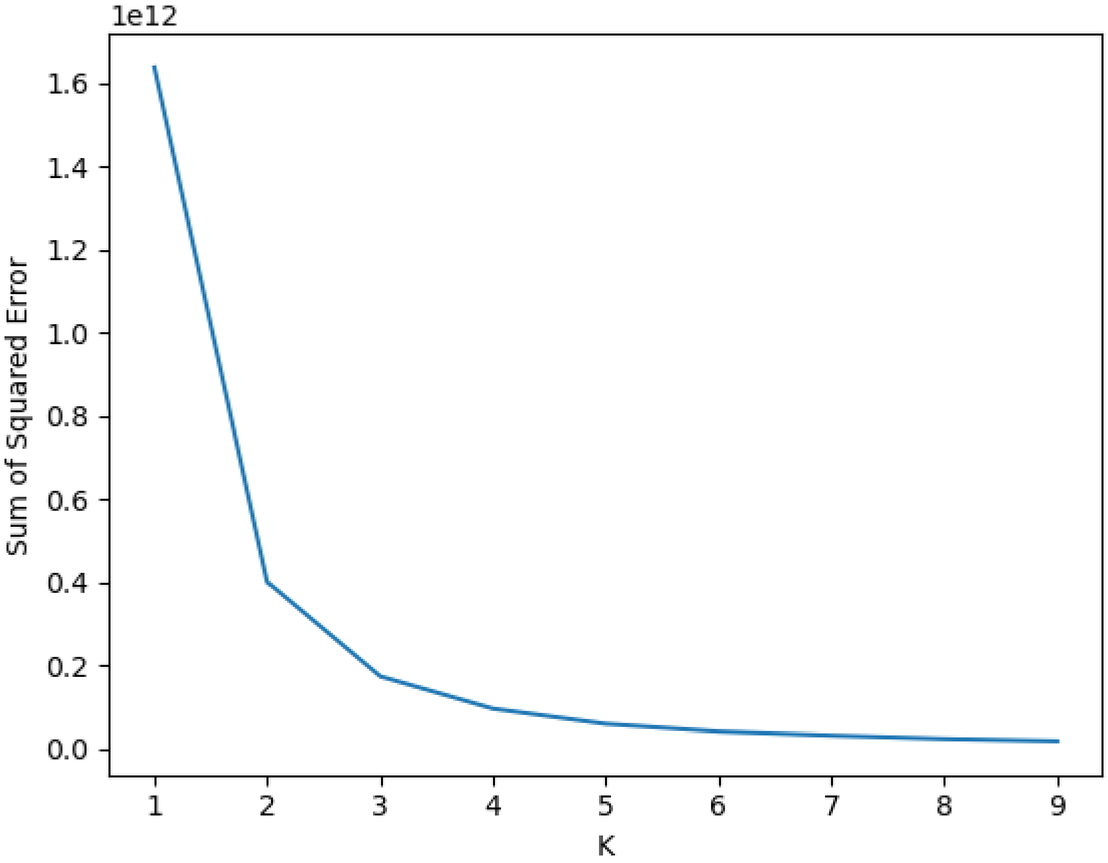
In [28]:

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

```
C:\Users\hp\anaconda3\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(
C:\Users\hp\anaconda3\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(
C:\Users\hp\anaconda3\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(
C:\Users\hp\anaconda3\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(
C:\Users\hp\anaconda3\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(
C:\Users\hp\anaconda3\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(
C:\Users\hp\anaconda3\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(
C:\Users\hp\anaconda3\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(
C:\Users\hp\anaconda3\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(

[1636787813359.9795, 400077229910.0807, 173655295069.7994, 96093251027.8487, 59792896184.73555, 41558169453.14693, 3
1835461273.526344, 23845563976.120872, 18614268499.003963]
```

Out[28]:

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



CONCLUSION:- BASED ON THE ABOVE PROGRAM DATA HAS BEEN DIVIDED INTO SEVARAL CLUSTERS