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

In [32]:

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

In [4]:

```
df=pd.read_csv(r"C:\Users\samit\OneDrive\Desktop\jupyter\Online Retail csv.csv")
df
```

Out[4]:

	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 [5]:

```
df.head()
```

Out[5]:

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

In [6]:

```
df.tail()
```

Out[6]:

	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 [7]:

```
df.describe()
```

Out[7]:

	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 [8]:

```
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 [9]:

```
df.isnull().any()
```

Out[9]:

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

In [10]:

```
df.shape
```

Out[10]:

```
(541909, 8)
```

In [11]:

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

In [12]:

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

Out[12]:

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

In [13]:

```
del df['InvoiceNo']
```

In [14]:

```
df
```

Out[14]:

	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	United Kingdom
1	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
2	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	United Kingdom
3	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
4	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...
541904	22613	PACK OF 20 SPACEBOY NAPKINS	12	09-12-2011 12:50	0.85	12680.0	France
541905	22899	CHILDREN'S APRON DOLLY GIRL	6	09-12-2011 12:50	2.10	12680.0	France
541906	23254	CHILDRENS CUTLERY DOLLY GIRL	4	09-12-2011 12:50	4.15	12680.0	France
541907	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	09-12-2011 12:50	4.15	12680.0	France
541908	22138	BAKING SET 9 PIECE RETROSPOT	3	09-12-2011 12:50	4.95	12680.0	France

541909 rows × 7 columns

In [15]:

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

Out[15]:

	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 [16]:

```
df.shape
```

Out[16]:

```
(541909, 3)
```

In [17]:

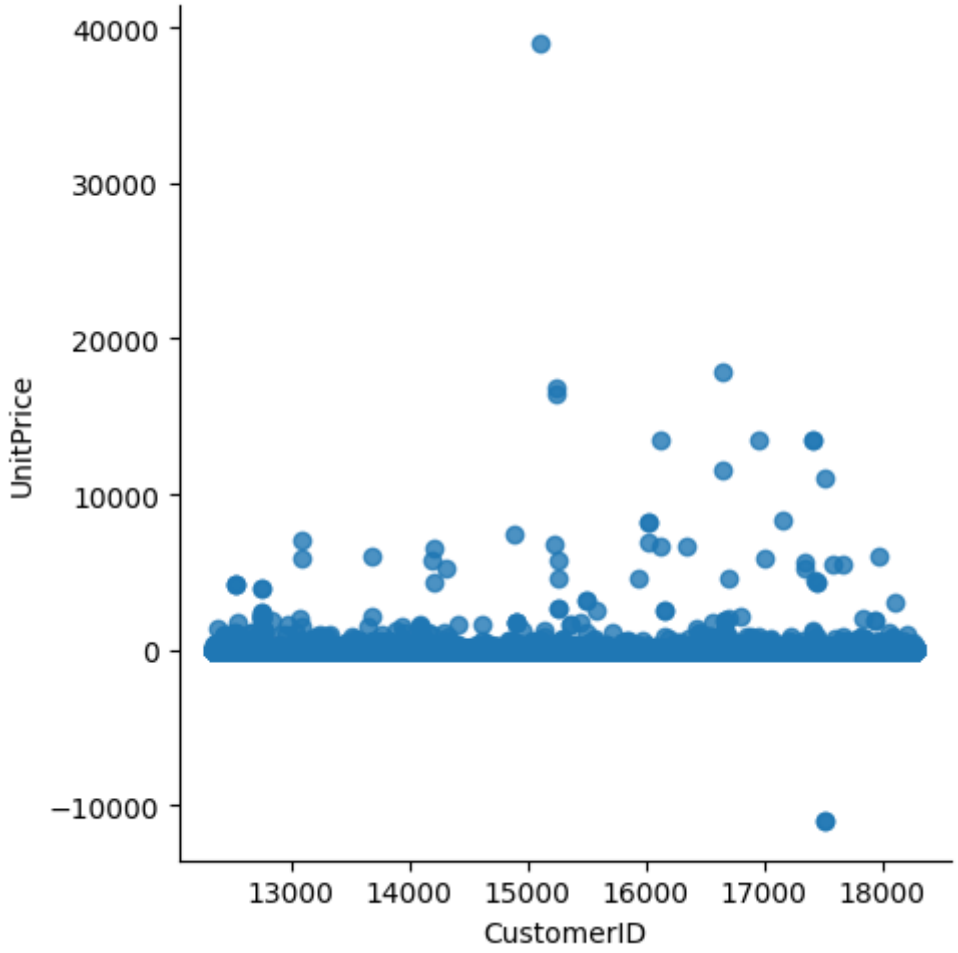
```
import seaborn as sns
import matplotlib.pyplot as plt
```

In [18]:

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

Out[18]:

```
<seaborn.axisgrid.FacetGrid at 0x14c6e099610>
```



In [19]:

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

Out[19]:

```
KMeans()
```

In [20]:

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

```
C:\Users\samit\AppData\Local\Programs\Python\Python311\Lib\site-packages
\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_i
nit` will change from 10 to 'auto' in 1.4. Set the value of `n_init` expl
icitly to suppress the warning
  warnings.warn(
```

Out[20]:

```
array([1, 1, 1, ..., 2, 2, 2])
```

In [21]:

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

```
C:\Users\samit\AppData\Local\Temp\ipykernel_18152\2282443312.py:1: Settin
gWithCopyWarning:
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-do
cs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (http
s://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returni
ng-a-view-versus-a-copy)
  df["cluster"]=y_predicted
```

Out[21]:

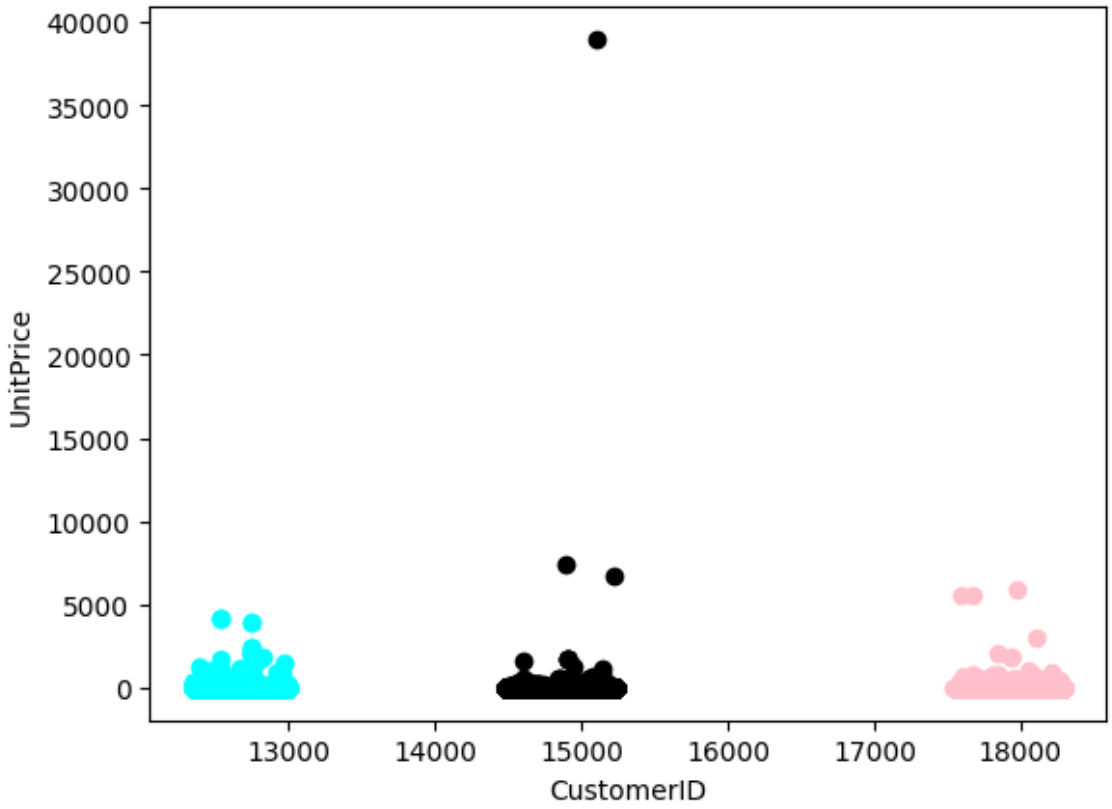
	Quantity	UnitPrice	CustomerID	cluster
0	6	2.55	17850.0	1
1	6	3.39	17850.0	1
2	8	2.75	17850.0	1
3	6	3.39	17850.0	1
4	6	3.39	17850.0	1

In [22]:

```
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[22]:

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



In [23]:

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

```
C:\Users\samit\AppData\Local\Temp\ipykernel_18152\4223297019.py:4: Settin
gWithCopyWarning:
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-do
cs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (http
s://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returni
ng-a-view-versus-a-copy)
  df["UnitPrice"]=scaler.transform(df[["UnitPrice"]])
```

Out[23]:

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

In [24]:

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

```
C:\Users\samit\AppData\Local\Programs\Python\Python311\Lib\site-packages
\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_i
nit` will change from 10 to 'auto' in 1.4. Set the value of `n_init` expl
icitly to suppress the warning
  warnings.warn(
```

Out[24]:

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

In [25]:

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

```
C:\Users\samit\AppData\Local\Temp\ipykernel_18152\2865533764.py:1: Settin
gWithCopyWarning:
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-do
cs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (http
s://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returni
ng-a-view-versus-a-copy)
  df["New Cluster"]=y_predicted
```

Out[25]:

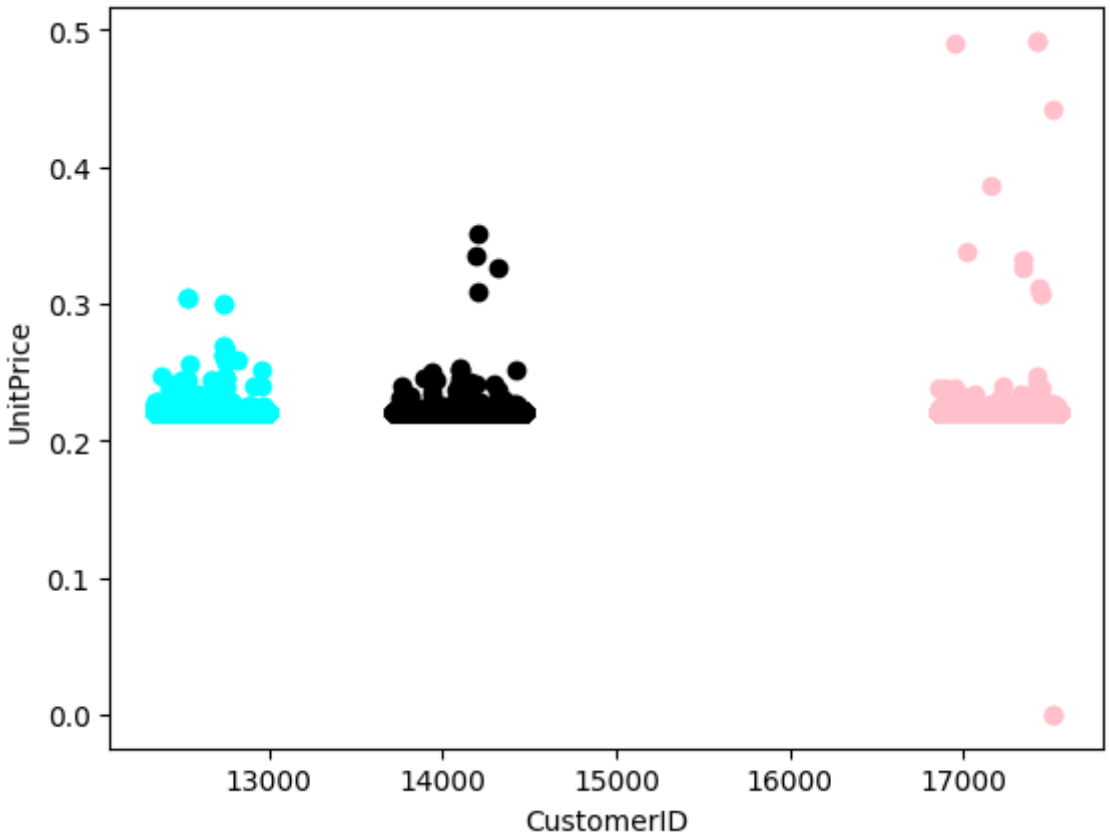
	Quantity	UnitPrice	CustomerID	cluster	New Cluster
0	6	0.221150	17850.0	1	4
1	6	0.221167	17850.0	1	4
2	8	0.221154	17850.0	1	4
3	6	0.221167	17850.0	1	4
4	6	0.221167	17850.0	1	4

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.xlabel("CustomerID")
plt.ylabel("UnitPrice")
```

Out[26]:

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



In [27]:

```
km.cluster_centers_
```

Out[27]:

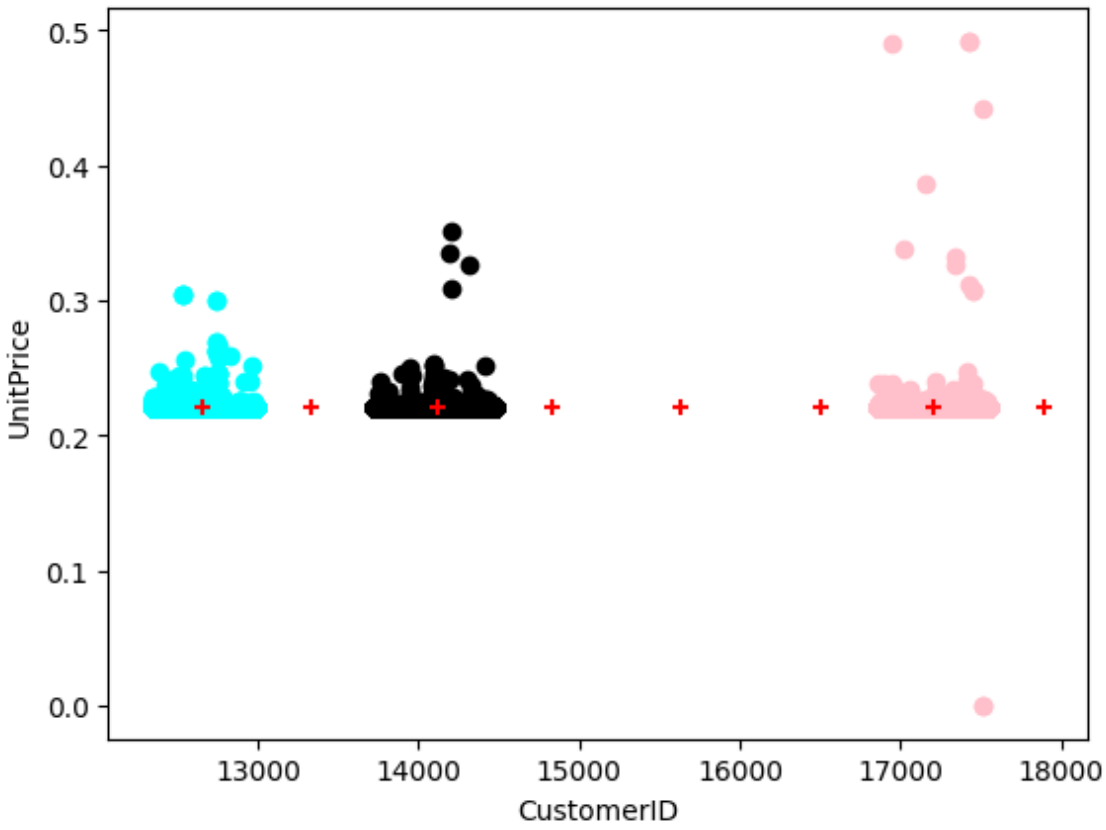
```
array([[1.41181124e+04, 2.21187350e-01],
       [1.72033145e+04, 2.21199591e-01],
       [1.26523663e+04, 2.21202876e-01],
       [1.56349467e+04, 2.21197183e-01],
       [1.78882276e+04, 2.21178244e-01],
       [1.33299132e+04, 2.21184417e-01],
       [1.48289027e+04, 2.21187553e-01],
       [1.65039144e+04, 2.21198101e-01]])
```

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["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[28]:

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



In [29]:

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

ELBOW METHOD:-

In [31]:

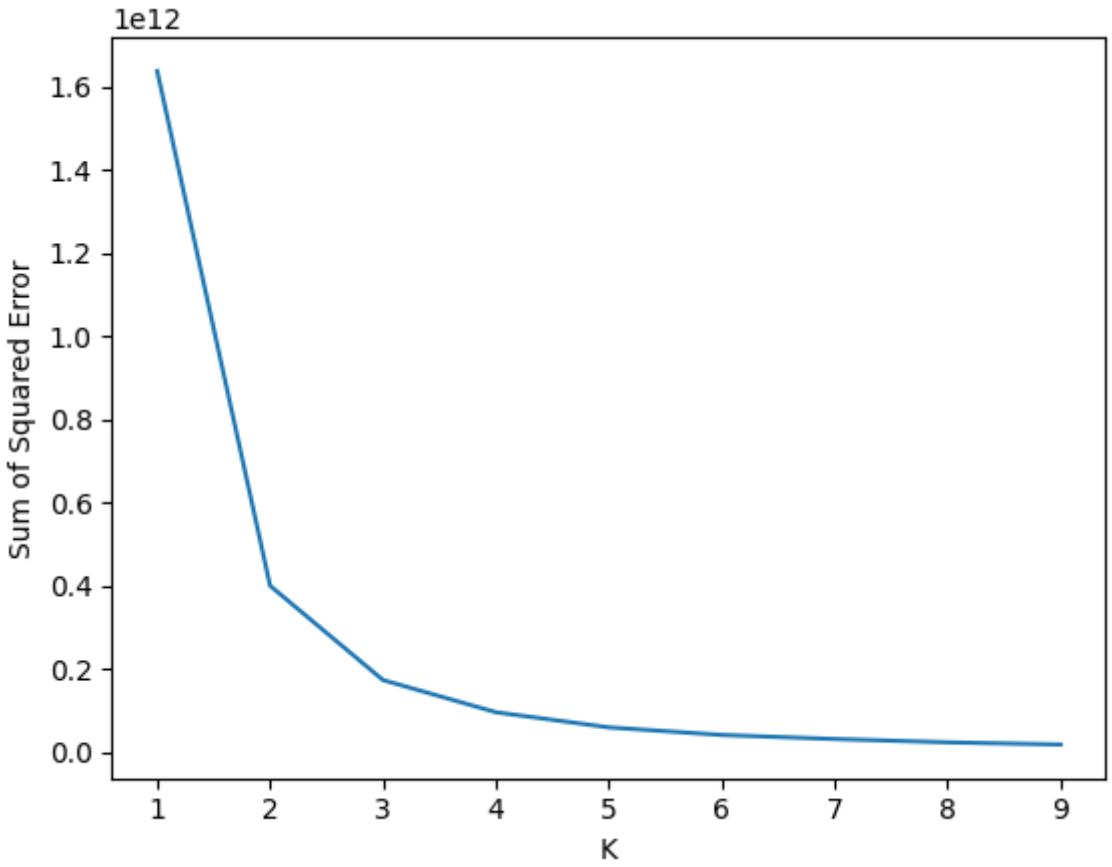
```
for k in k_rng:
    km=KMeans(n_clusters=k)
    km.fit(df[["CustomerID","UnitPrice"]])
    sse.append(km.inertia_)
print(sse)
plt.plot(k_rng,sse)
plt.xlabel("K")
plt.ylabel("Sum of Squared Error")
```

```
C:\Users\samit\AppData\Local\Programs\Python\Python311\Lib\site-packages
\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_i
nit` will change from 10 to 'auto' in 1.4. Set the value of `n_init` expl
icitly to suppress the warning
  warnings.warn(
C:\Users\samit\AppData\Local\Programs\Python\Python311\Lib\site-packages
\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_i
nit` will change from 10 to 'auto' in 1.4. Set the value of `n_init` expl
icitly to suppress the warning
  warnings.warn(
C:\Users\samit\AppData\Local\Programs\Python\Python311\Lib\site-packages
\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_i
nit` will change from 10 to 'auto' in 1.4. Set the value of `n_init` expl
icitly to suppress the warning
  warnings.warn(
C:\Users\samit\AppData\Local\Programs\Python\Python311\Lib\site-packages
\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_i
nit` will change from 10 to 'auto' in 1.4. Set the value of `n_init` expl
icitly to suppress the warning
  warnings.warn(
C:\Users\samit\AppData\Local\Programs\Python\Python311\Lib\site-packages
\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_i
nit` will change from 10 to 'auto' in 1.4. Set the value of `n_init` expl
icitly to suppress the warning
  warnings.warn(
C:\Users\samit\AppData\Local\Programs\Python\Python311\Lib\site-packages
\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_i
nit` will change from 10 to 'auto' in 1.4. Set the value of `n_init` expl
icitly to suppress the warning
  warnings.warn(
C:\Users\samit\AppData\Local\Programs\Python\Python311\Lib\site-packages
\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_i
nit` will change from 10 to 'auto' in 1.4. Set the value of `n_init` expl
icitly to suppress the warning
  warnings.warn(
C:\Users\samit\AppData\Local\Programs\Python\Python311\Lib\site-packages
\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_i
nit` will change from 10 to 'auto' in 1.4. Set the value of `n_init` expl
icitly to suppress the warning
  warnings.warn(
C:\Users\samit\AppData\Local\Programs\Python\Python311\Lib\site-packages
\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_i
nit` will change from 10 to 'auto' in 1.4. Set the value of `n_init` expl
icitly to suppress the warning
  warnings.warn(
```

```
[1636787813359.9038, 400077229910.0796, 173473424696.1372, 96093174431.13
193, 59794644206.69411, 41558097996.873886, 31829293596.37155, 2384890443
0.7886, 18619563731.525654]
```

Out[31]:

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



Based on the above prograom data has been divided into several clusters