# PROJECT

# 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.

## Importing libraries

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

## Reading the data

```
In [3]: df=pd.read_csv(r"C:\Users\Lenovo\OneDrive\Desktop\Data Sets\OnlineRetail.csv")
        df
```

Out[3]:

| | InvoiceNo | StockCode | Description | Quantity | InvoiceDate | UnitPrice | CustomerID | Coun |
|---|---|---|---|---|---|---|---|---|
| **0** | 536365 | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | 6 | 1/12/2010 8:26 | 2.55 | 17850.0 | Uni Kingd |
| **1** | 536365 | 71053 | WHITE METAL LANTERN | 6 | 1/12/2010 8:26 | 3.39 | 17850.0 | Uni Kingd |
| **2** | 536365 | 84406B | CREAM CUPID HEARTS COAT HANGER | 8 | 1/12/2010 8:26 | 2.75 | 17850.0 | Uni Kingd |
| **3** | 536365 | 84029G | KNITTED UNION FLAG HOT WATER BOTTLE | 6 | 1/12/2010 8:26 | 3.39 | 17850.0 | Uni Kingd |
| **4** | 536365 | 84029E | RED WOOLLY HOTTIE WHITE HEART. | 6 | 1/12/2010 8:26 | 3.39 | 17850.0 | Uni Kingd |
| **...** | ... | ... | ... | ... | ... | ... | ... | |
| **541904** | 581587 | 22613 | PACK OF 20 SPACEBOY NAPKINS | 12 | 9/12/2011 12:50 | 0.85 | 12680.0 | Frar |
| **541905** | 581587 | 22899 | CHILDREN'S APRON DOLLY GIRL | 6 | 9/12/2011 12:50 | 2.10 | 12680.0 | Frar |
| **541906** | 581587 | 23254 | CHILDRENS CUTLERY DOLLY GIRL | 4 | 9/12/2011 12:50 | 4.15 | 12680.0 | Frar |
| **541907** | 581587 | 23255 | CHILDRENS CUTLERY CIRCUS PARADE | 4 | 9/12/2011 12:50 | 4.15 | 12680.0 | Frar |
| **541908** | 581587 | 22138 | BAKING SET 9 PIECE RETROSPOT | 3 | 9/12/2011 12:50 | 4.95 | 12680.0 | Frar |

541909 rows × 8 columns

# Data cleaning and preprocessing

```
In [4]: df.columns
```

```
Out[4]: Index(['InvoiceNo', 'StockCode', 'Description', 'Quantity', 'InvoiceDate',
               'UnitPrice', 'CustomerID', 'Country'],
              dtype='object')
```

```
In [5]: df.isnull().sum()
```

```
Out[5]: InvoiceNo           0
        StockCode           0
        Description      1454
        Quantity            0
        InvoiceDate         0
        UnitPrice           0
        CustomerID     135080
        Country             0
        dtype: int64
```

```
In [6]: df.fillna(method="ffill",inplace=True)
```

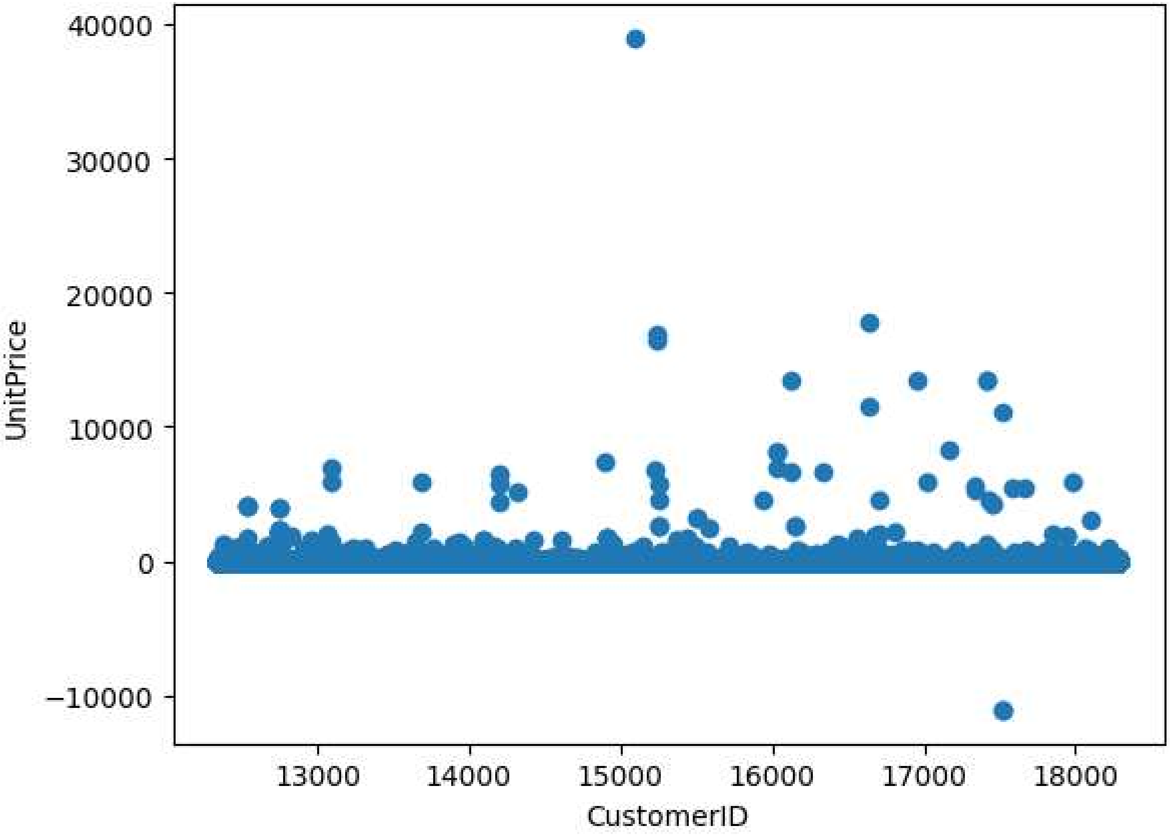
```
In [7]: df.isnull().sum()
```

```
Out[7]: InvoiceNo      0
        StockCode      0
        Description    0
        Quantity       0
        InvoiceDate    0
        UnitPrice      0
        CustomerID     0
        Country        0
        dtype: int64
```

# Applying KMeans

In [8]:
```
plt.scatter(df["CustomerID"],df["UnitPrice"])
plt.xlabel("CustomerID")
plt.ylabel("UnitPrice")
```

Out[8]:
```
Text(0, 0.5, 'UnitPrice')
```



In [9]:
```
from sklearn.cluster import KMeans
```

In [10]:
```
km=KMeans()
km
```

Out[10]:
```
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 [12]:
```
y_predicted=km.fit_predict(df[["CustomerID","UnitPrice"]])
y_predicted
```

```
C:\Users\Lenovo\AppData\Local\Programs\Python\Python311\Lib\site-packages\skl
earn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_init` wil
l change from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly to su
ppress the warning
  warnings.warn(
```

Out[12]:
```
array([2, 2, 2, ..., 5, 5, 5])
```

In [13]:

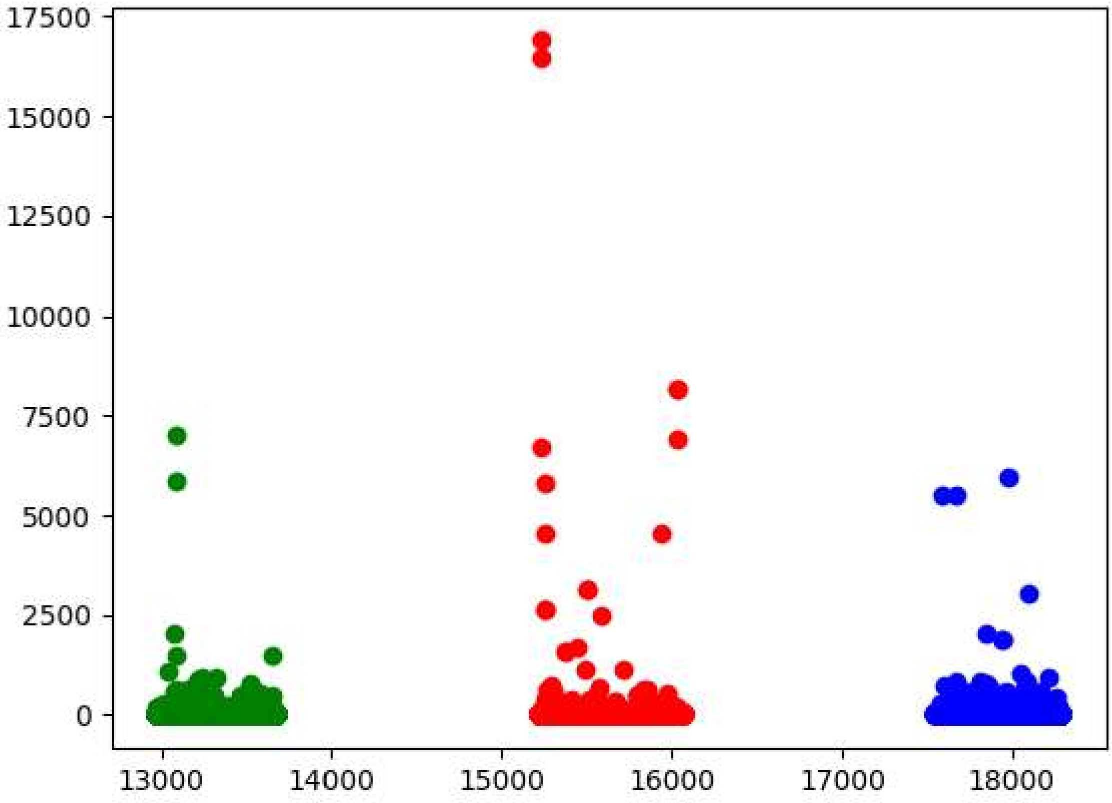
```
df["Cluster"]=y_predicted
df.head()
```

Out[13]:

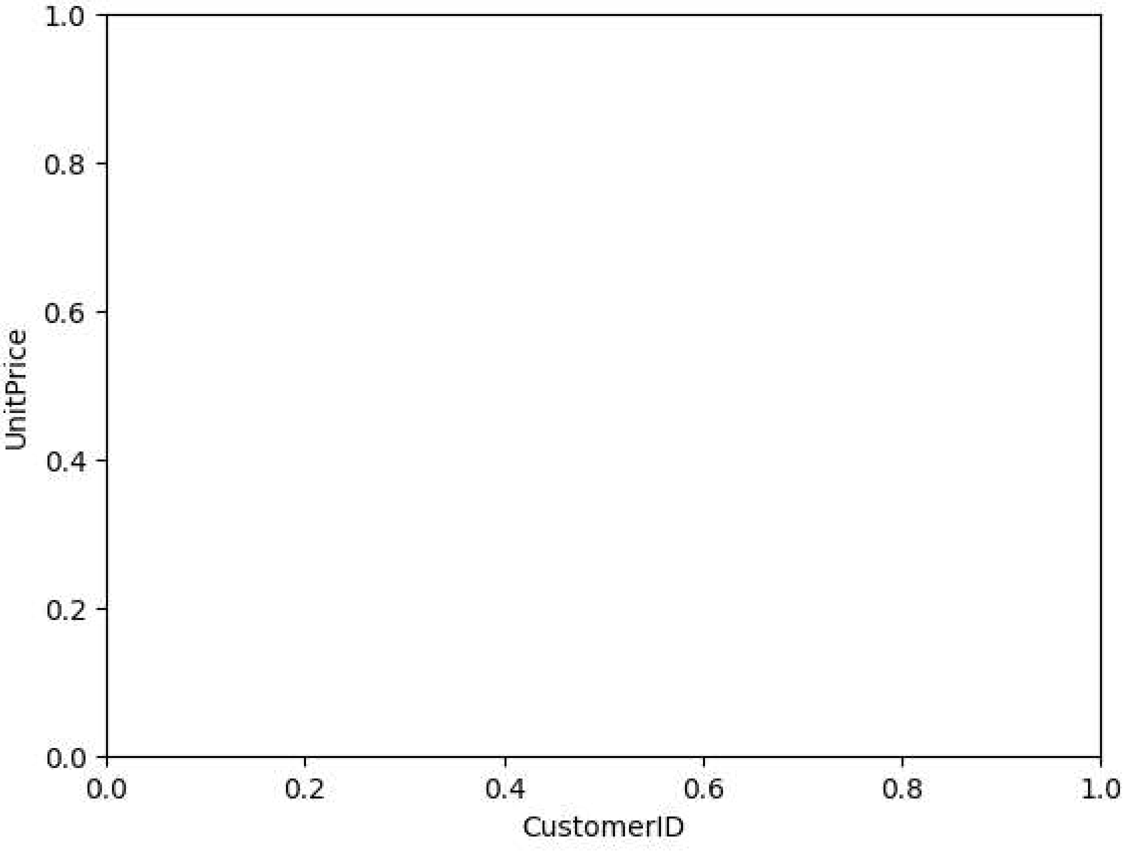
| | InvoiceNo | StockCode | Description | Quantity | InvoiceDate | UnitPrice | CustomerID | Country | Cl |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 536365 | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | 6 | 1/12/2010 8:26 | 2.55 | 17850.0 | United Kingdom | |
| **1** | 536365 | 71053 | WHITE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

In [14]:
```
df1=df[df.Cluster==0]
df2=df[df.Cluster==1]
df3=df[df.Cluster==2]

plt.scatter(df1["CustomerID"],df1["UnitPrice"],color="red")
plt.scatter(df2["CustomerID"],df2["UnitPrice"],color="green")
plt.scatter(df3["CustomerID"],df3["UnitPrice"],color="blue")
```

Out[14]:
```
<matplotlib.collections.PathCollection at 0x249d5862d50>
```



```
In [15]: plt.xlabel("CustomerID")
         plt.ylabel("UnitPrice")

Out[15]: Text(0, 0.5, 'UnitPrice')
```



```
In [16]: from sklearn.preprocessing import MinMaxScaler
         scaler=MinMaxScaler()
```

In [17]:

```
scaler.fit(df[["UnitPrice"]])
df["UnitPrice"]=scaler.transform(df[["UnitPrice"]])
df.head()
```

Out[17]:

| | InvoiceNo | StockCode | Description | Quantity | InvoiceDate | UnitPrice | CustomerID | Country | Cl |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 536365 | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | 6 | 1/12/2010 8:26 | 0.221150 | 17850.0 | United Kingdom | |
| **1** | 536365 | 71053 | WHITE METAL LANTERN | 6 | 1/12/2010 8:26 | 0.221167 | 17850.0 | United Kingdom | |
| **2** | 536365 | 84406B | CREAM CUPID HEARTS COAT HANGER | 8 | 1/12/2010 8:26 | 0.221154 | 17850.0 | United Kingdom | |
| **3** | 536365 | 84029G | KNITTED UNION FLAG HOT WATER BOTTLE | 6 | 1/12/2010 8:26 | 0.221167 | 17850.0 | United Kingdom | |
| **4** | 536365 | 84029E | RED WOOLLY HOTTIE WHITE HEART. | 6 | 1/12/2010 8:26 | 0.221167 | 17850.0 | United Kingdom | |

In [18]:

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

Out[18]:

| | InvoiceNo | StockCode | Description | Quantity | InvoiceDate | UnitPrice | CustomerID | Country | Cl |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 536365 | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | 6 | 1/12/2010 8:26 | 0.221150 | 0.926443 | United Kingdom | |
| **1** | 536365 | 71053 | WHITE METAL LANTERN | 6 | 1/12/2010 8:26 | 0.221167 | 0.926443 | United Kingdom | |
| **2** | 536365 | 84406B | CREAM CUPID HEARTS COAT HANGER | 8 | 1/12/2010 8:26 | 0.221154 | 0.926443 | United Kingdom | |
| **3** | 536365 | 84029G | KNITTED UNION FLAG HOT WATER BOTTLE | 6 | 1/12/2010 8:26 | 0.221167 | 0.926443 | United Kingdom | |
| **4** | 536365 | 84029E | RED WOOLLY HOTTIE WHITE HEART. | 6 | 1/12/2010 8:26 | 0.221167 | 0.926443 | United Kingdom | |

In [19]:

```
km=KMeans()
```

In [20]:

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

```
C:\Users\Lenovo\AppData\Local\Programs\Python\Python311\Lib\site-packages\skl
earn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_init` wil
l change from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly to su
ppress the warning
  warnings.warn(
```

Out[20]:

```
array([4, 4, 4, ..., 1, 1, 1])
```

In [21]:

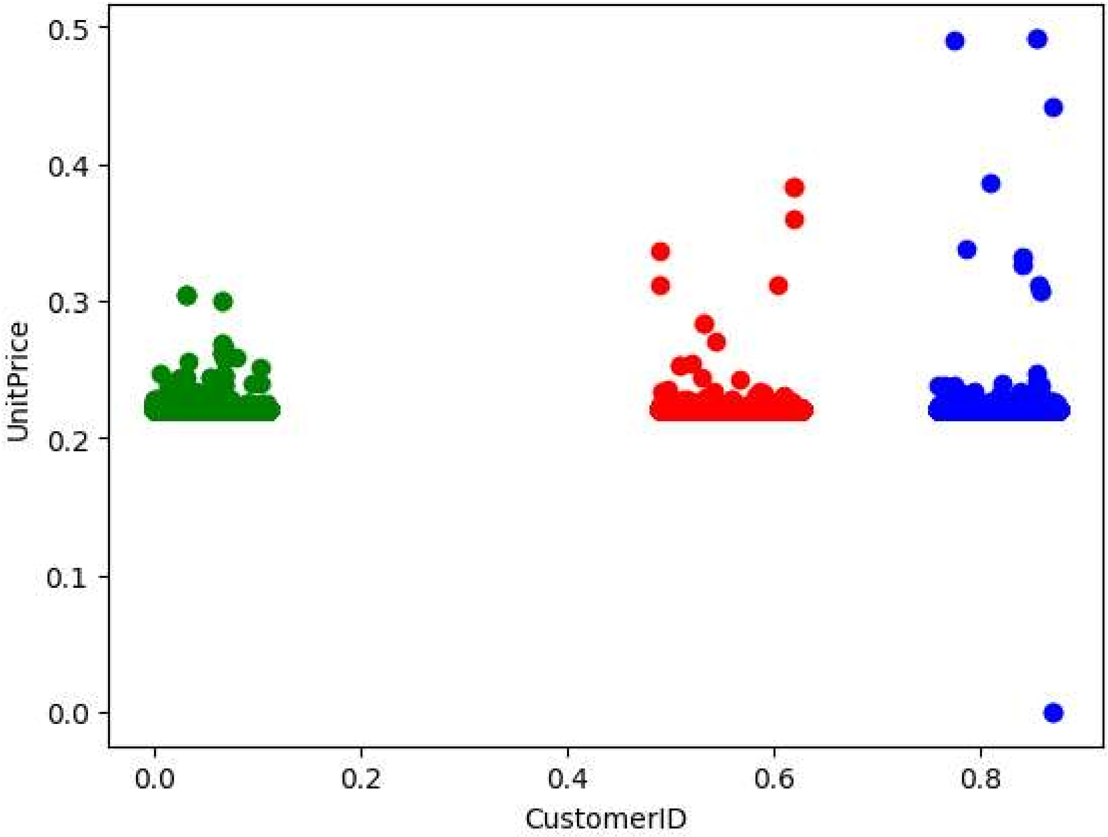
```
df["New Cluster"]=y_predicted
df.head()
```

Out[21]:

| | InvoiceNo | StockCode | Description | Quantity | InvoiceDate | UnitPrice | CustomerID | Country | Cl |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 536365 | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | 6 | 1/12/2010 8:26 | 0.221150 | 0.926443 | United Kingdom | |
| **1** | 536365 | 71053 | WHITE METAL LANTERN | 6 | 1/12/2010 8:26 | 0.221167 | 0.926443 | United Kingdom | |
| **2** | 536365 | 84406B | CREAM CUPID HEARTS COAT HANGER | 8 | 1/12/2010 8:26 | 0.221154 | 0.926443 | United Kingdom | |
| **3** | 536365 | 84029G | KNITTED UNION FLAG HOT WATER BOTTLE | 6 | 1/12/2010 8:26 | 0.221167 | 0.926443 | United Kingdom | |
| **4** | 536365 | 84029E | RED WOOLLY HOTTIE WHITE HEART. | 6 | 1/12/2010 8:26 | 0.221167 | 0.926443 | United Kingdom | |

In [23]:
```
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="red")
plt.scatter(df2["CustomerID"],df2["UnitPrice"],color="green")
plt.scatter(df3["CustomerID"],df3["UnitPrice"],color="blue")

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

Out[23]:
```
Text(0, 0.5, 'UnitPrice')
```



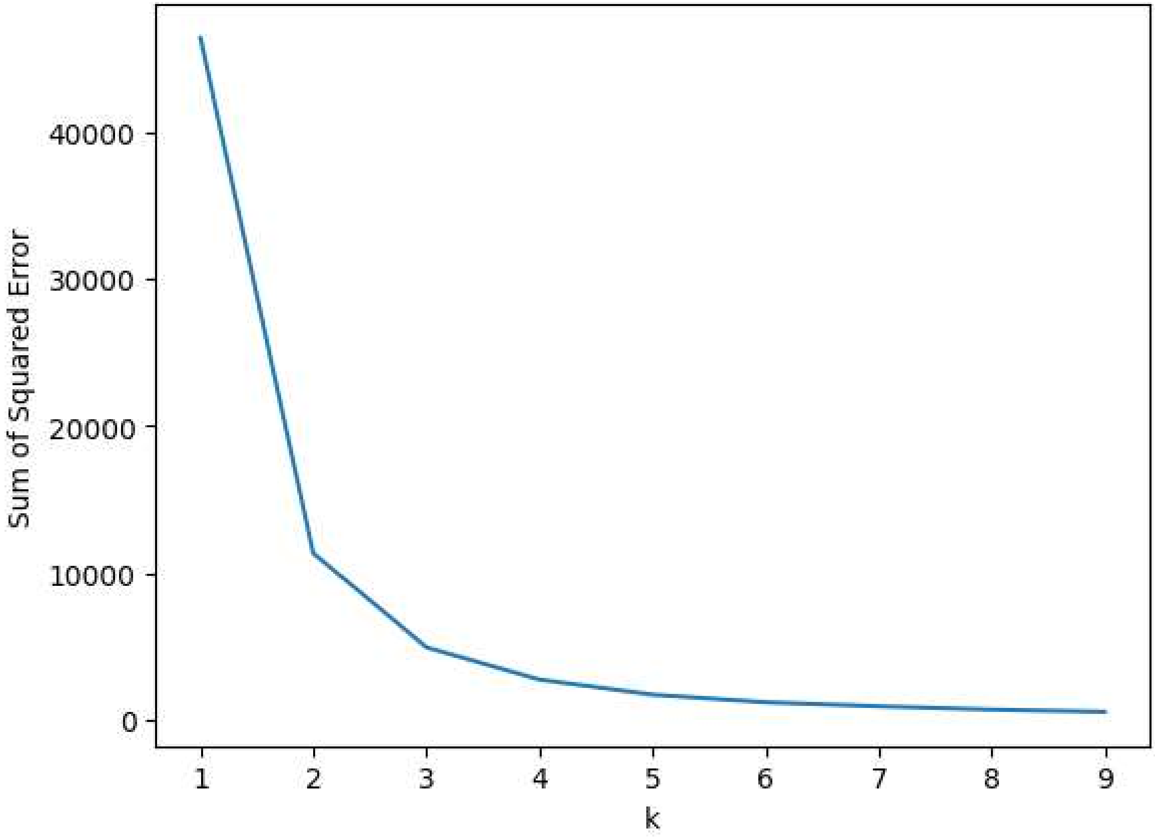
In [24]:
```
k_rng=range(1,10)
sse=[]
```

In [25]:
```
for k in k_rng:
    km=KMeans(n_clusters=k)
    km.fit(df[["CustomerID","UnitPrice"]])
    sse.append(km.inertia_)
sse
```

```
C:\Users\Lenovo\AppData\Local\Programs\Python\Python311\Lib\site-packages\skl
earn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_init` wil
l change from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly to su
ppress the warning
  warnings.warn(
C:\Users\Lenovo\AppData\Local\Programs\Python\Python311\Lib\site-packages\skl
earn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_init` wil
l change from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly to su
ppress the warning
  warnings.warn(
C:\Users\Lenovo\AppData\Local\Programs\Python\Python311\Lib\site-packages\skl
earn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_init` wil
l change from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly to su
ppress the warning
  warnings.warn(
C:\Users\Lenovo\AppData\Local\Programs\Python\Python311\Lib\site-packages\skl
earn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_init` wil
l change from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly to su
ppress the warning
  warnings.warn(
C:\Users\Lenovo\AppData\Local\Programs\Python\Python311\Lib\site-packages\skl
earn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_init` wil
l change from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly to su
ppress the warning
  warnings.warn(
C:\Users\Lenovo\AppData\Local\Programs\Python\Python311\Lib\site-packages\skl
earn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_init` wil
l change from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly to su
ppress the warning
  warnings.warn(
C:\Users\Lenovo\AppData\Local\Programs\Python\Python311\Lib\site-packages\skl
earn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_init` wil
l change from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly to su
ppress the warning
  warnings.warn(
C:\Users\Lenovo\AppData\Local\Programs\Python\Python311\Lib\site-packages\skl
earn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_init` wil
l change from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly to su
ppress the warning
  warnings.warn(
C:\Users\Lenovo\AppData\Local\Programs\Python\Python311\Lib\site-packages\skl
earn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_init` wil
l change from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly to su
ppress the warning
  warnings.warn(
Out[25]: [46375.89020547866,
          11337.110496294004,
          4919.4931454647085,
          2724.56079103382,
          1696.1075931587568,
          1179.4690017913558,
          905.5886528414202,
          678.2463155005453,
          529.3923176277611]
```

```
In [26]: plt.plot(k_rng,sse)
         plt.xlabel("k")
         plt.ylabel("Sum of Squared Error")
```

```
Out[26]: Text(0, 0.5, 'Sum of Squared Error')
```



**CONCLUSION : For the given dataset,we used clustering algorithm named KMeans and we got an accuarte graph(Elbow curve).**