K-Means Clustering

In [3]:

```
import pandas as pd
from matplotlib import pyplot as plt
%matplotlib inline
```

In [5]:

```
df=pd.read_csv(r"C:\Users\thara\Downloads\Income.csv")
df
```

Out[5]:

	Gender	Age	Income($)
0	Male	19	15
1	Male	21	15
2	Female	20	16
3	Female	23	16
4	Female	31	17
...	...	...	...
195	Female	35	120
196	Female	45	126
197	Male	32	126
198	Male	32	137
199	Male	30	137

200 rows × 3 columns

In [6]:

```
df.head()
```

Out[6]:

	Gender	Age	Income($)
0	Male	19	15
1	Male	21	15
2	Female	20	16
3	Female	23	16
4	Female	31	17

In [7]:

```
df.tail()
```

Out[7]:

	Gender	Age	Income($)
195	Female	35	120
196	Female	45	126
197	Male	32	126
198	Male	32	137
199	Male	30	137

In [8]:

```
df.shape
```

Out[8]:

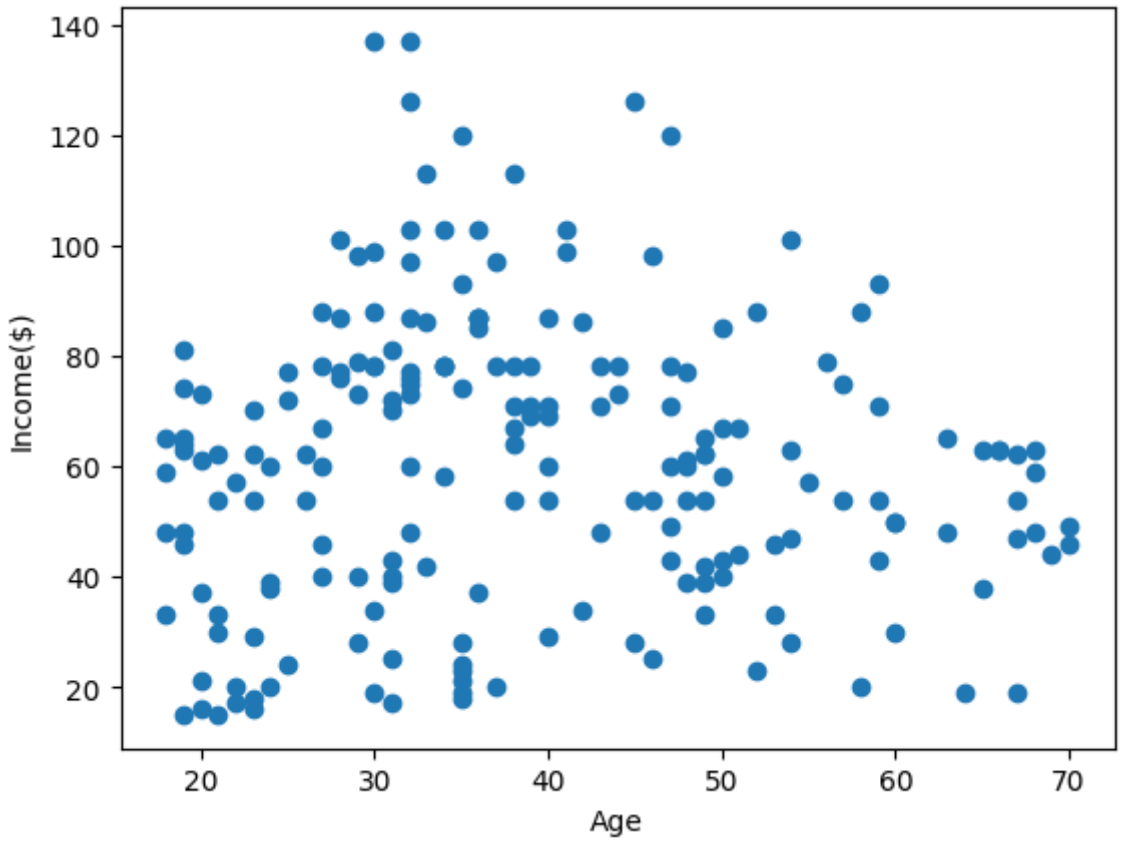
```
(200, 3)
```

In [9]:

```
plt.scatter(df["Age"],df["Income($)"])
plt.xlabel("Age")
plt.ylabel("Income($)")
```

Out[9]:

```
Text(0, 0.5, 'Income($)')
```



In [10]:

```
from sklearn.cluster import KMeans
```

In [11]:

```
kM=KMeans()
kM
```

Out[11]:

KMeans()

In [12]:

```
y_predicted = kM.fit_predict(df[["Age","Income($)"]])
y_predicted
```

```
C:\Users\thara\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
```

Out[12]:

```
array([5, 5, 5, 5, 5, 5, 5, 5, 1, 5, 1, 5, 1, 5, 5, 5, 5, 5, 1, 5, 5, 5,
       1, 5, 1, 5, 1, 5, 1, 5, 1, 5, 1, 6, 1, 6, 1, 6, 6, 6, 4, 6, 1, 6,
       1, 6, 1, 6, 6, 6, 1, 6, 6, 4, 1, 1, 1, 4, 6, 4, 4, 6, 4, 4, 4, 6,
       2, 4, 6, 6, 4, 4, 4, 4, 4, 6, 2, 2, 6, 2, 4, 2, 4, 2, 6, 2, 4, 6,
       2, 2, 4, 3, 2, 2, 3, 3, 2, 3, 2, 3, 3, 2, 4, 3, 2, 3, 4, 2, 4, 4,
       4, 3, 2, 3, 3, 3, 4, 2, 2, 2, 3, 2, 2, 2, 3, 3, 2, 2, 2, 2, 2, 2,
       3, 3, 3, 3, 2, 3, 3, 3, 2, 3, 3, 3, 3, 3, 2, 3, 3, 3, 2, 3, 2, 3,
       2, 3, 3, 3, 3, 3, 2, 3, 3, 3, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7,
       7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 0, 0, 0, 0, 0, 0,
       0, 0])
```

In [13]:

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

Out[13]:

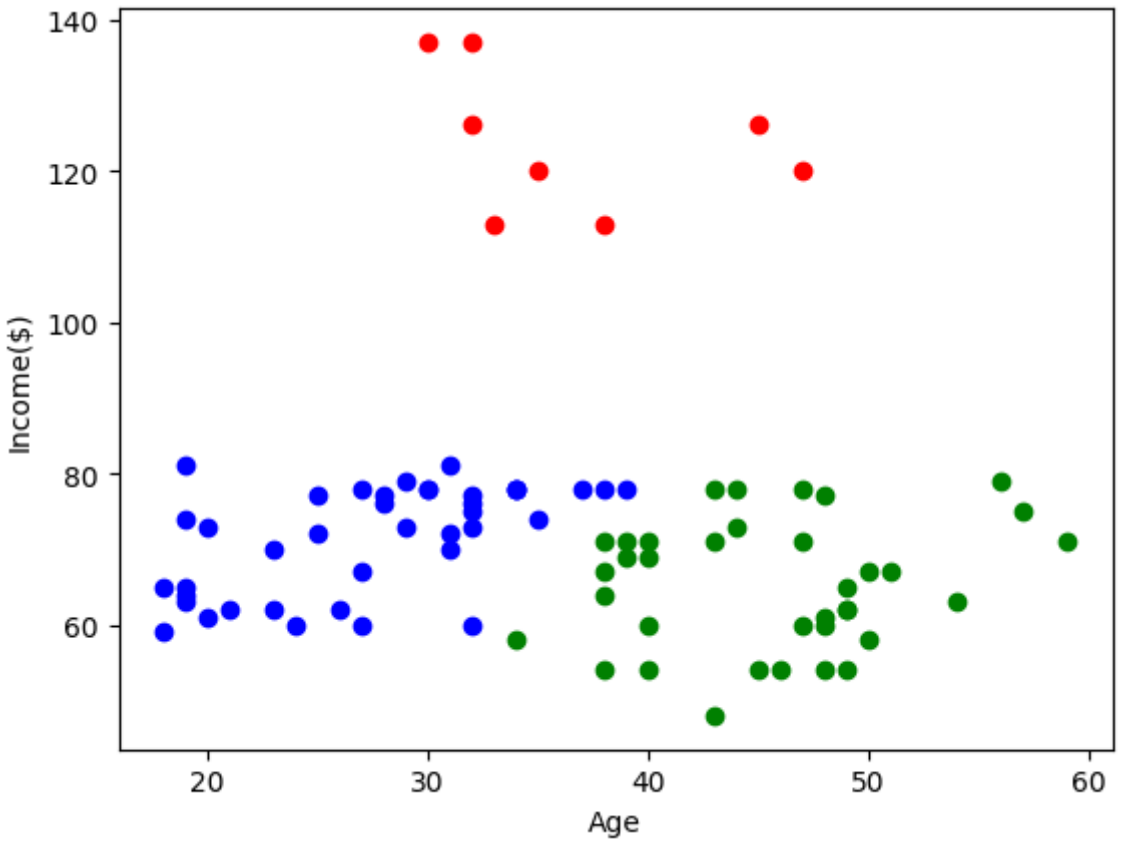
	Gender	Age	Income($)	cluster
0	Male	19	15	5
1	Male	21	15	5
2	Female	20	16	5
3	Female	23	16	5
4	Female	31	17	5

In [14]:

```
df1 = df[df.cluster == 0]
df2 = df[df.cluster == 2]
df3 = df[df.cluster == 3]
plt.scatter(df1["Age"],df1["Income($)"],color="red")
plt.scatter(df2["Age"],df2["Income($)"],color="green")
plt.scatter(df3["Age"],df3["Income($)"],color="blue")
plt.xlabel("Age")
plt.ylabel("Income($)")
```

Out[14]:

```
Text(0, 0.5, 'Income($)')
```



In [15]:

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

Out[15]:

	Gender	Age	Income($)	cluster
0	Male	19	0.000000	5
1	Male	21	0.000000	5
2	Female	20	0.008197	5
3	Female	23	0.008197	5
4	Female	31	0.016393	5

In [16]:

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

Out[16]:

	Gender	Age	Income($)	cluster
0	Male	0.019231	0.000000	5
1	Male	0.057692	0.000000	5
2	Female	0.038462	0.008197	5
3	Female	0.096154	0.008197	5
4	Female	0.250000	0.016393	5

In [17]:

```
km=KMeans()
```

In [18]:

```
y_predicted=km.fit_predict(df[["Age","Income($)"]])
y_predicted
```

```
C:\Users\thara\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
```

Out[18]:

```
array([6, 6, 6, 6, 0, 6, 0, 6, 4, 0, 4, 0, 7, 6, 0, 6, 0, 6, 7, 0, 0, 6,
       7, 0, 7, 0, 7, 0, 0, 6, 4, 6, 7, 6, 7, 6, 7, 0, 0, 6, 4, 6, 7, 0,
       7, 6, 7, 0, 0, 0, 7, 0, 0, 4, 7, 7, 7, 4, 0, 7, 4, 3, 4, 7, 4, 3,
       7, 4, 3, 0, 4, 7, 4, 4, 4, 3, 7, 7, 3, 7, 4, 2, 4, 7, 3, 7, 1, 3,
       2, 1, 4, 3, 1, 2, 2, 3, 1, 3, 1, 3, 3, 1, 4, 3, 1, 3, 4, 1, 4, 4,
       4, 3, 2, 3, 3, 3, 4, 1, 1, 1, 3, 2, 2, 2, 3, 2, 1, 2, 1, 2, 1, 2,
       3, 2, 3, 2, 1, 2, 3, 2, 1, 2, 2, 2, 3, 2, 1, 2, 2, 2, 1, 2, 1, 2,
       1, 2, 2, 2, 2, 2, 1, 2, 3, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2,
       1, 2, 1, 2, 5, 5, 5, 5, 5, 5, 1, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5,
       5, 5])
```

In [19]:

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

Out[19]:

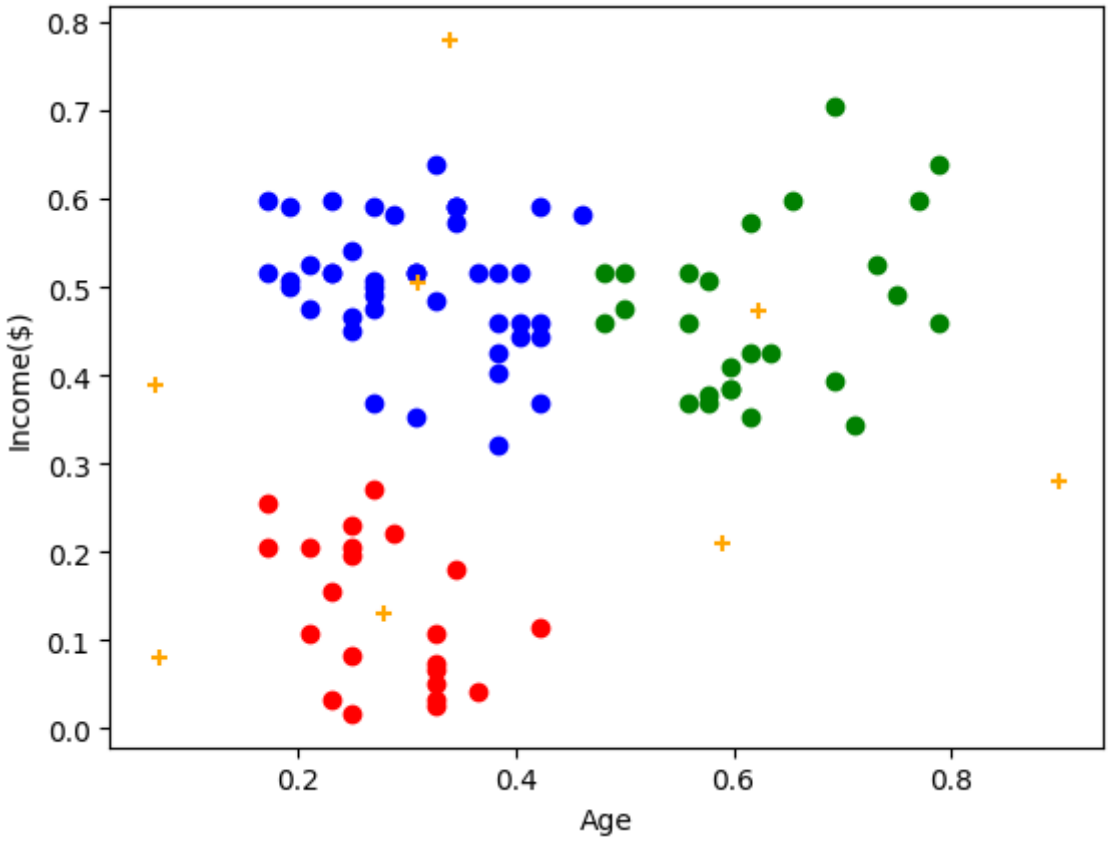
	Gender	Age	Income($)	cluster	New Cluster
0	Male	0.019231	0.000000	5	6
1	Male	0.057692	0.000000	5	6
2	Female	0.038462	0.008197	5	6
3	Female	0.096154	0.008197	5	6
4	Female	0.250000	0.016393	5	0

In [20]:

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

Out[20]:

```
Text(0, 0.5, 'Income($)')
```



In [21]:

```
km.cluster_centers_
```

Out[21]:

```
array([[0.27884615, 0.13040238],
       [0.62352071, 0.47225725],
       [0.30944056, 0.50428465],
       [0.06923077, 0.38786885],
       [0.89799331, 0.28011404],
       [0.34008097, 0.77998274],
       [0.07239819, 0.08003857],
       [0.58974359, 0.20969945]])
```

In [23]:

```
k_rng=range(1,10)
sse=[]
for k in k_rng:
    km=KMeans(n_clusters=k)
    km.fit(df[["Age","Income($)"]])
    sse.append(km.inertia_)
    print(sse)
```

```
C:\Users\thara\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\thara\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\thara\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\thara\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\thara\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\thara\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\thara\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
```

```
[23.583906150363607]
[23.583906150363607, 13.028938428018286]
[23.583906150363607, 13.028938428018286, 7.49302484330499]
[23.583906150363607, 13.028938428018286, 7.49302484330499, 6.0558586448125
47]
[23.583906150363607, 13.028938428018286, 7.49302484330499, 6.0558586448125
47, 4.713811834695168]
[23.583906150363607, 13.028938428018286, 7.49302484330499, 6.0558586448125
47, 4.713811834695168, 3.8651257592912622]
[23.583906150363607, 13.028938428018286, 7.49302484330499, 6.0558586448125
47, 4.713811834695168, 3.8651257592912622, 3.0547174363693586]
[23.583906150363607, 13.028938428018286, 7.49302484330499, 6.0558586448125
47, 4.713811834695168, 3.8651257592912622, 3.0547174363693586, 2.646060977
430515]
[23.583906150363607, 13.028938428018286, 7.49302484330499, 6.0558586448125
47, 4.713811834695168, 3.8651257592912622, 3.0547174363693586, 2.646060977
430515, 2.3135720353543285]
```

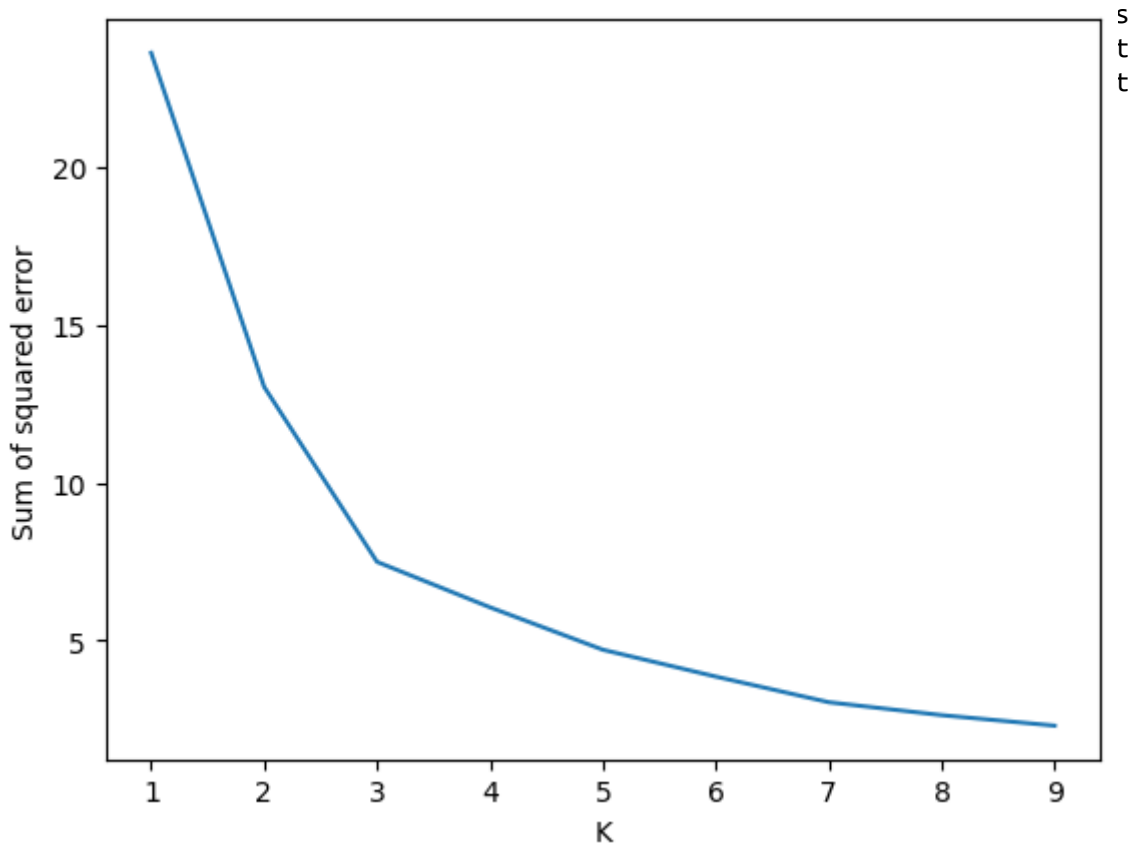
In [24]:

```
plt.plot(k_rng,sse)
plt.xlabel("K")
plt.ylabel("Sum of squared error")
```

```
C:\Users\thara\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
```

Out[24]:

```
Text(0, 0.5, 'Sum of squared error')
```



In [ ]: