K-MEANS CLUSTERING

In [28]:

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

import warnings
warnings.filterwarnings("ignore")
```

In [29]:

```
df=pd.read_csv(r"C:\Users\91949\Downloads\Income.csv")
df
```

Out[29]:

	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

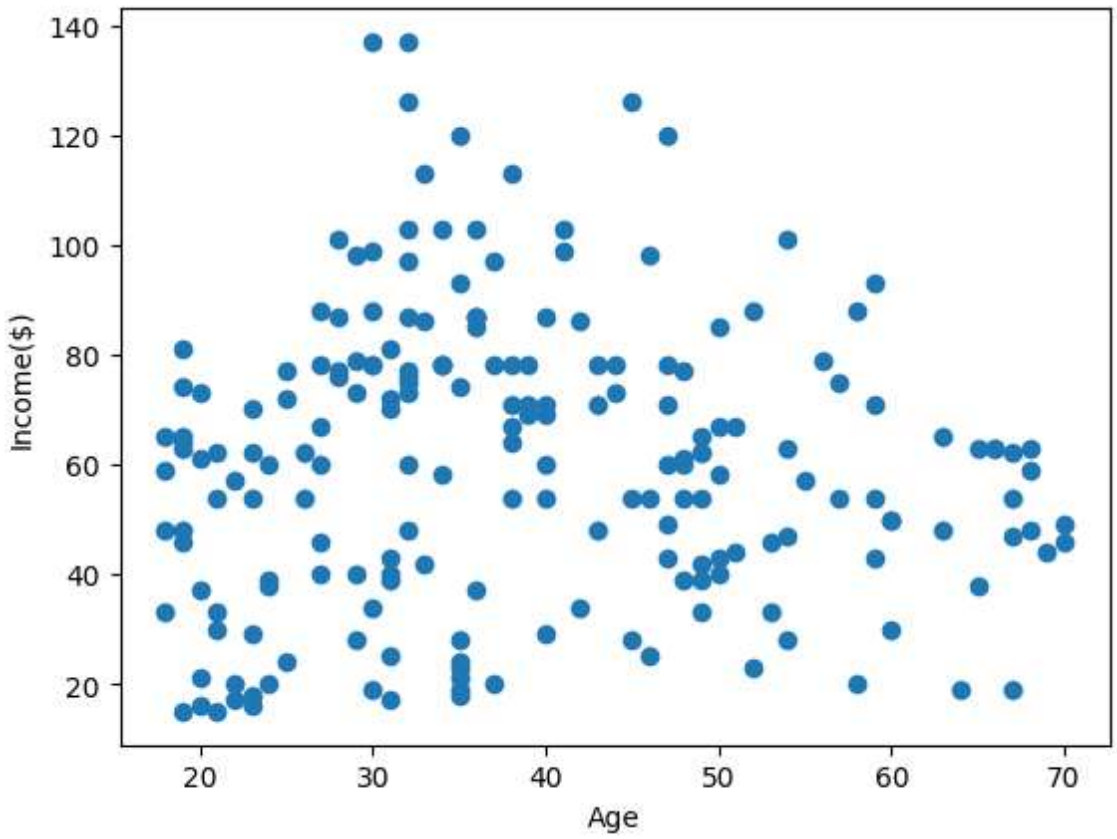
200 rows × 3 columns

In [30]:

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

Out[30]:

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



In [31]:

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

In [32]:

```
km = KMeans()
km
```

Out[32]:

```
KMeans()
```

In [33]:

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

Out[33]:

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

In [34]:

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

Out[34]:

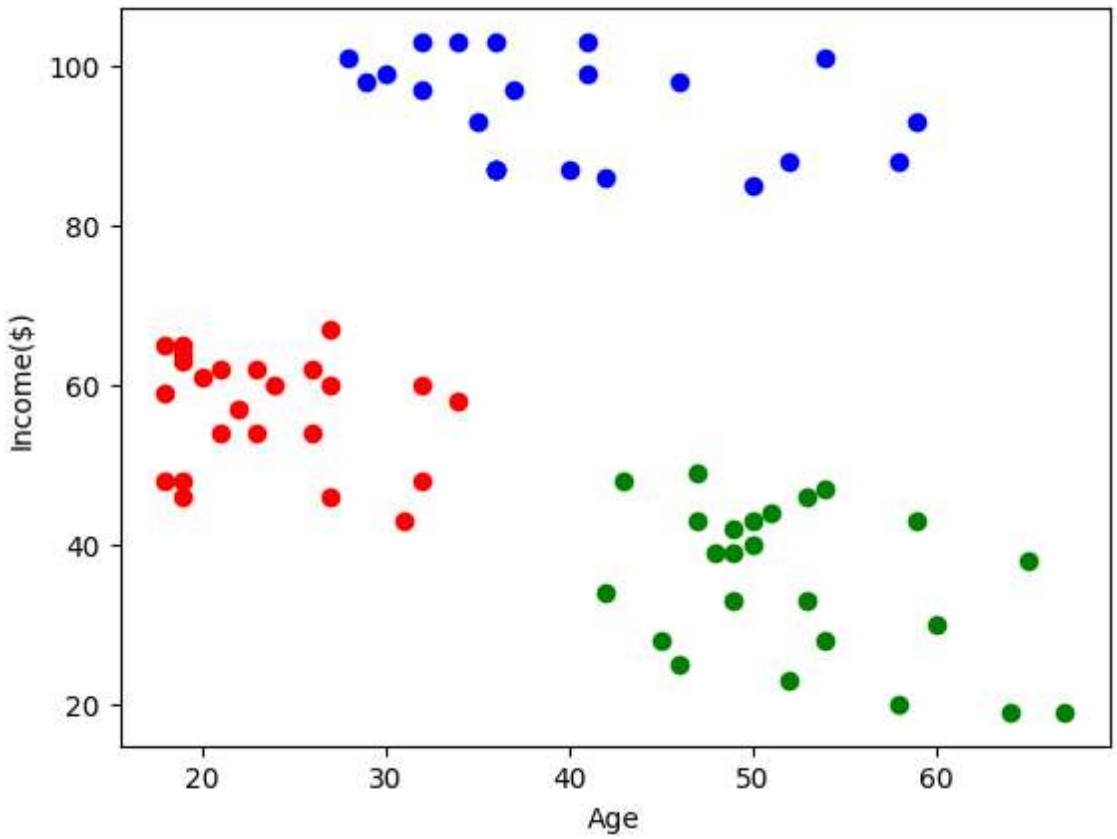
	Gender	Age	Income($)	Cluster
0	Male	19	15	3
1	Male	21	15	3
2	Female	20	16	3
3	Female	23	16	3
4	Female	31	17	3

In [35]:

```
df1=df[df.Cluster==0]
df2=df[df.Cluster==1]
df3=df[df.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.xlabel("Age")
plt.ylabel("Income($)")

```

Out[35]:

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



In [36]:

```
from sklearn.preprocessing import MinMaxScaler
```

In [37]:

```
scaler=MinMaxScaler()
```

In [39]:

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

Out[39]:

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

In [40]:

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

Out[40]:

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

In [41]:

```
km=KMeans()
```

In [43]:

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

Out[43]:

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

In [44]:

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

Out[44]:

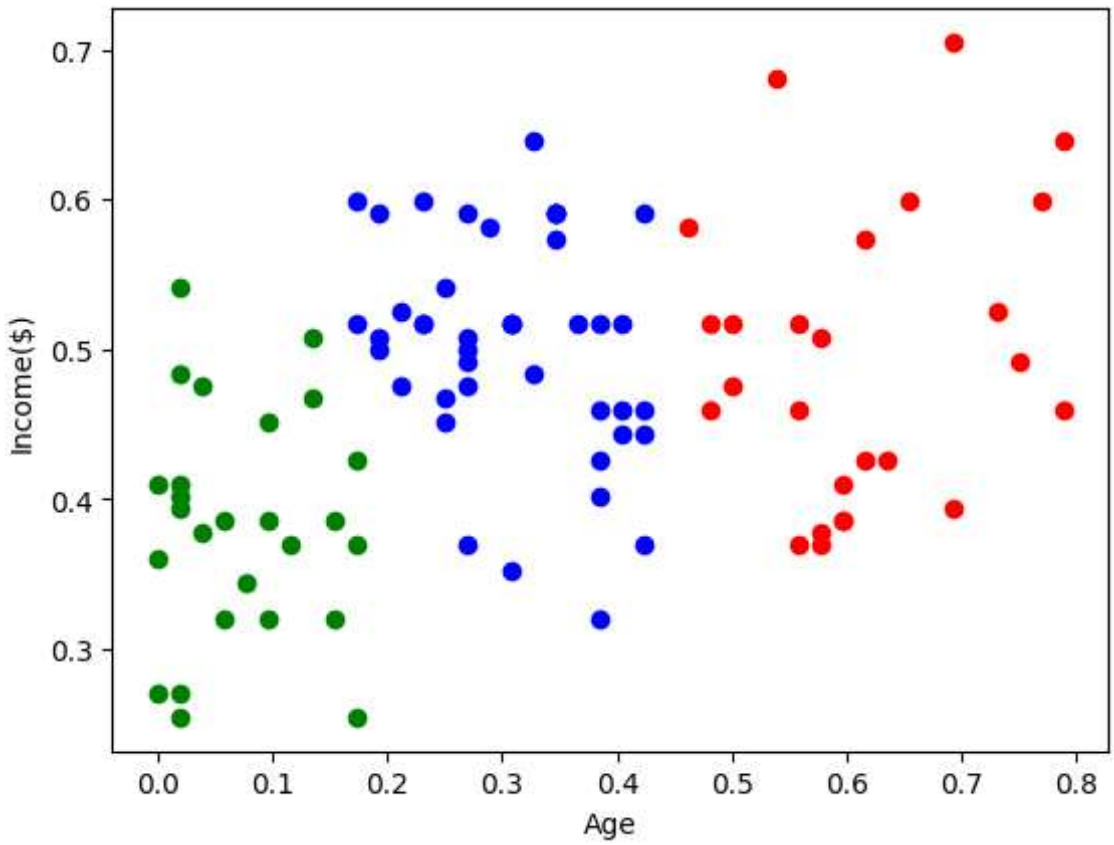
	Gender	Age	Income($)	Cluster	New Cluster
0	Male	0.019231	0.000000	3	6
1	Male	0.057692	0.000000	3	6
2	Female	0.038462	0.008197	3	6
3	Female	0.096154	0.008197	3	6
4	Female	0.250000	0.016393	3	3

In [45]:

```
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.xlabel("Age")
plt.ylabel("Income($)")

```

Out[45]:

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



In [46]:

```
km.cluster_centers_
```

Out[46]:

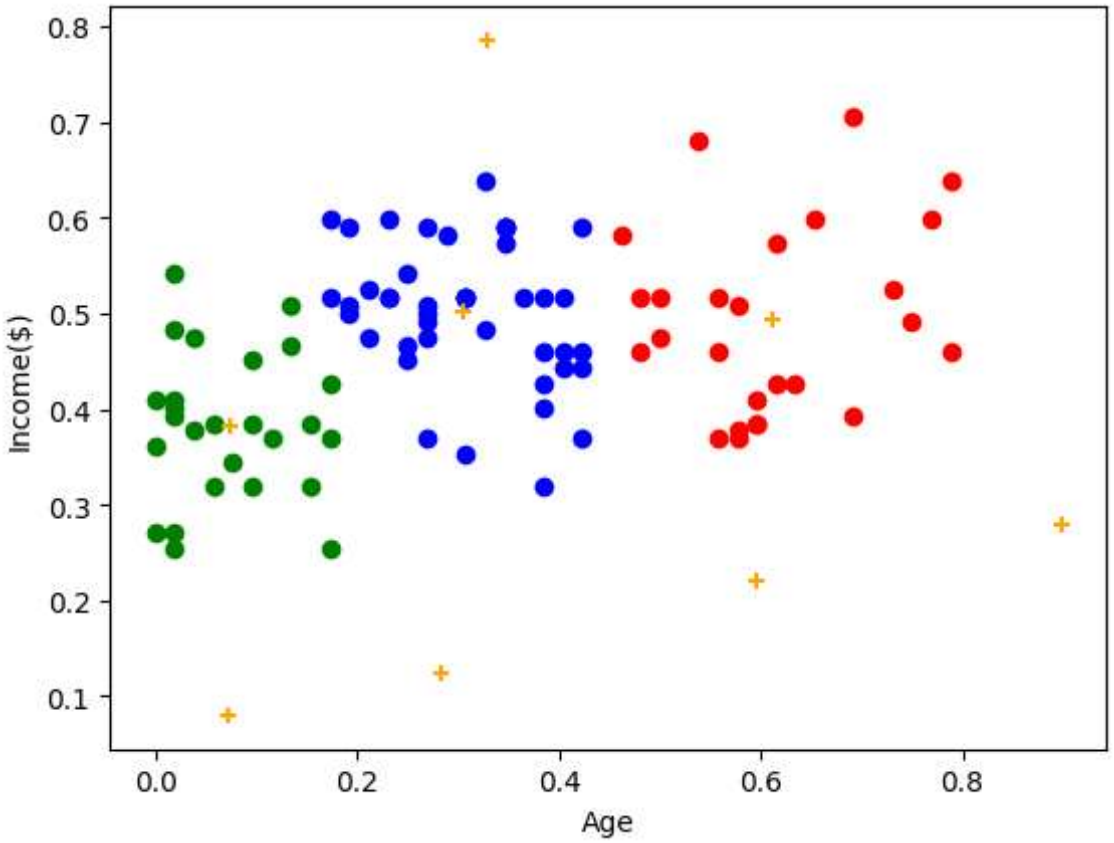
```
array([[0.61094675, 0.49401009],
       [0.07322485, 0.38272383],
       [0.3059034 , 0.50247808],
       [0.28388278, 0.1245121 ],
       [0.89799331, 0.28011404],
       [0.32905983, 0.78551913],
       [0.07239819, 0.08003857],
       [0.5954142 , 0.2203657 ]])
```

In [47]:

```
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[47]:

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



In [51]:

```
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_)
sse
```

Out[51]:

```
[23.583906150363603,
 13.028938428018286,
 7.492107868586012,
 6.055858644812547,
 4.713416604872824,
 3.8616447037115407,
 3.054717436369358,
 2.643789119086916,
 2.3135720353543285]
```

In [52]:

```
plt.plot(k_rng,sse)
plt.xlabel("k")
plt.ylabel("Sum of Squared Error")
```

Out[52]:

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