k means clustering

In [1]:
```python
import pandas as pd
from matplotlib import pyplot as plt
%matplotlib inline
df=pd.read_csv(r"C:\Users\ubinl\Downloads\Income.csv")
df
```

Out[1]:

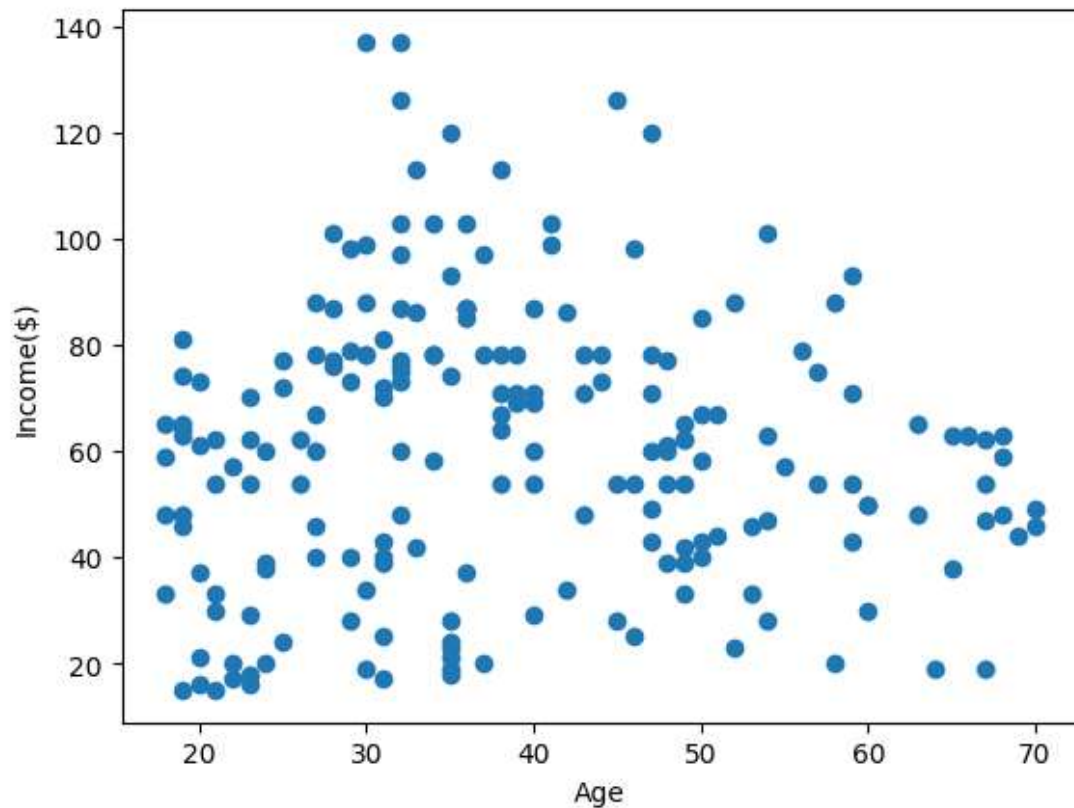|	Gender	Age	Income($)
0	Male	19	15
1	Male	21	15
2	Female	20	16
3	Female	23	16
4	Female	31	17
...	...	...	...
195	Female	35	120
196	Female	45	126
197	Male	32	126
198	Male	32	137
199	Male	30	137

200 rows × 3 columns

In [2]:
```
plt.scatter(df["Age"],df["Income($)"])
plt.xlabel("Age")
plt.ylabel("Income($)")
```

Out[2]: Text(0, 0.5, 'Income($)')

In [3]:
```
from sklearn.cluster import KMeans
km=KMeans()
km
```

Out[3]:
```
▼ KMeans
KMeans()
```

In [4]:
```python
y_predicted=km.fit_predict(df[["Age","Income($)"]])
y_predicted
```

C:\Users\ubinl\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\cl
uster_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[4]:
```
array([4, 4, 4, 4, 4, 4, 4, 4, 1, 4, 1, 4, 1, 4, 4, 4, 4, 4, 1, 4, 4, 4,
       1, 4, 1, 4, 1, 4, 1, 4, 1, 4, 1, 4, 1, 4, 4, 4, 5, 4, 1, 4,
       1, 4, 1, 4, 4, 4, 1, 4, 6, 5, 1, 1, 1, 5, 6, 5, 5, 6, 5, 5, 5, 6,
       0, 5, 6, 6, 5, 0, 5, 5, 5, 6, 0, 0, 6, 0, 5, 0, 5, 0, 6, 0, 5, 6,
       6, 0, 5, 6, 0, 0, 6, 6, 0, 6, 0, 6, 6, 0, 5, 6, 0, 6, 5, 0, 5, 5,
       5, 6, 0, 6, 6, 6, 5, 0, 0, 0, 6, 0, 0, 0, 2, 2, 0, 0, 0, 2, 0, 2,
       2, 2, 2, 2, 0, 2, 2, 2, 0, 2, 2, 2, 2, 2, 0, 2, 2, 2, 2, 2, 2, 2,
       0, 2, 2, 2, 2, 2, 0, 2, 2, 2, 7, 2, 7, 2, 7, 2, 7, 2, 7, 7, 7, 2,
       7, 2, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 3, 3, 3, 3, 3, 3,
       3, 3])
```

In [5]:
```python
df["cluster"]=y_predicted
df.head()
```

Out[5]:

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

In [6]:
```python
df1=df[df.cluster==0]
df2=df[df.cluster==1]
df3=df[df.cluster==2]
plt.scatter(df1["Age"],df1["Income($)"],color="purple")
plt.scatter(df2["Age"],df2["Income($)"],color="skyblue")
plt.scatter(df3["Age"],df3["Income($)"],color="violet")
plt.xlabel("Age")
plt.ylabel("Income($)")
```

Out[6]: Text(0, 0.5, 'Income($)')
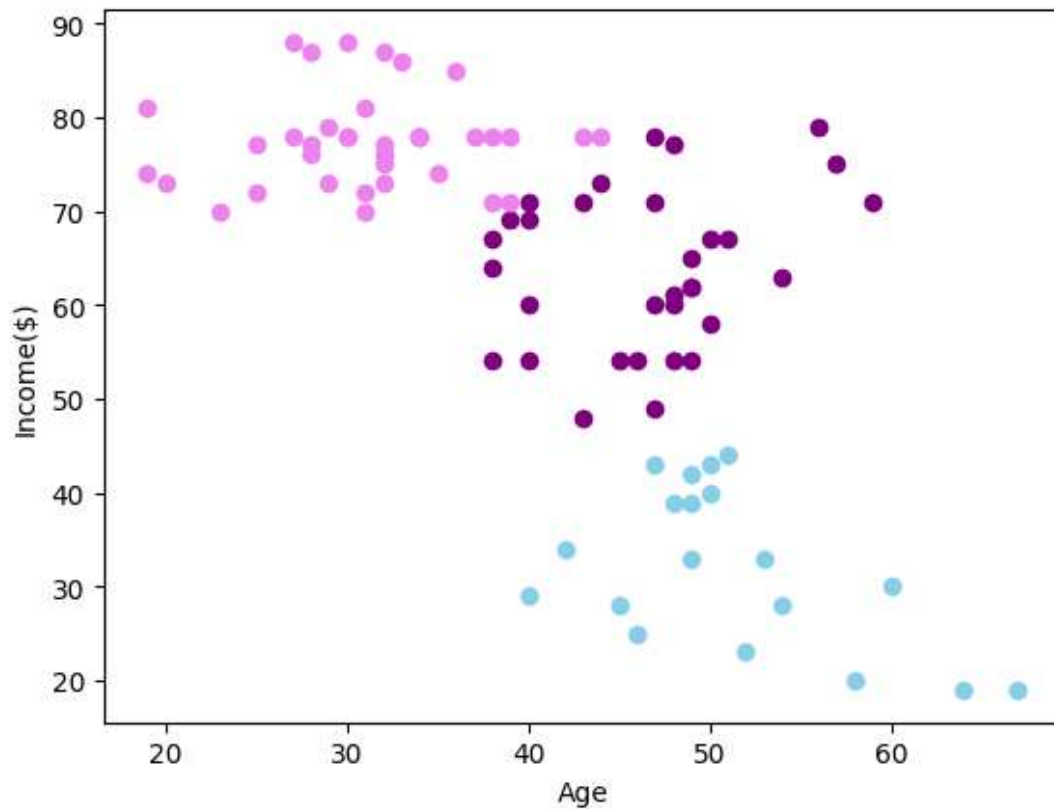
In [7]:
```python
from sklearn.preprocessing import MinMaxScaler
```

In [8]:
```python
Scaler=MinMaxScaler()
```

In [9]:
```python
Scaler.fit(df[["Income($)"]])
df["Income($)"]=Scaler.transform(df[["Income($)"]])
df.head()
```

Out[9]:

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

In [10]:
```python
Scaler.fit(df[["Age"]])
df["Age"]=Scaler.transform(df[["Age"]])
df.head()
```

Out[10]:

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

In [11]:
```python
km=KMeans()
km
```

Out[11:

```
▼ KMeans
KMeans()
```

In [12]:
```python
y_predicted=km.fit_predict(df[["Age","Income($)"]])
y_predicted
```

C:\Users\ubinl\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\cl
uster_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[12]:
```
array([7, 7, 7, 7, 0, 7, 0, 7, 6, 0, 6, 0, 2, 7, 0, 7, 0, 7, 2, 0, 0, 7,
       2, 0, 2, 0, 2, 0, 0, 7, 6, 7, 2, 7, 2, 7, 2, 0, 0, 7, 6, 7, 2, 0,
       2, 7, 2, 0, 0, 0, 2, 0, 0, 6, 2, 2, 2, 6, 3, 2, 6, 3, 6, 2, 6, 3,
       2, 6, 3, 0, 6, 2, 6, 6, 6, 3, 2, 2, 3, 2, 6, 1, 6, 2, 3, 2, 5, 3,
       1, 5, 6, 3, 5, 1, 1, 3, 5, 3, 5, 3, 3, 5, 6, 3, 5, 3, 6, 5, 6, 6,
       6, 3, 1, 3, 3, 3, 6, 5, 5, 5, 3, 1, 1, 1, 3, 1, 5, 1, 5, 1, 5, 1,
       3, 1, 3, 1, 5, 1, 3, 1, 5, 1, 1, 1, 3, 1, 5, 1, 1, 1, 5, 1, 5, 1,
       5, 1, 1, 1, 1, 1, 5, 1, 3, 1, 5, 1, 1, 1, 1, 1, 1, 1, 1, 1, 5, 1,
       5, 1, 5, 1, 4, 4, 5, 4, 4, 4, 5, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4])
```

In [13]:
```python
df1=df[df.cluster==0]
df2=df[df.cluster==1]
df3=df[df.cluster==2]
plt.scatter(df1["Age"],df1["Income($)"],color="hotpink")
plt.scatter(df2["Age"],df2["Income($)"],color="SpringGreen")
plt.scatter(df3["Age"],df3["Income($)"],color="plum")
plt.xlabel("Age")
plt.ylabel("Income($)")
```

Out[13]: Text(0, 0.5, 'Income($)')
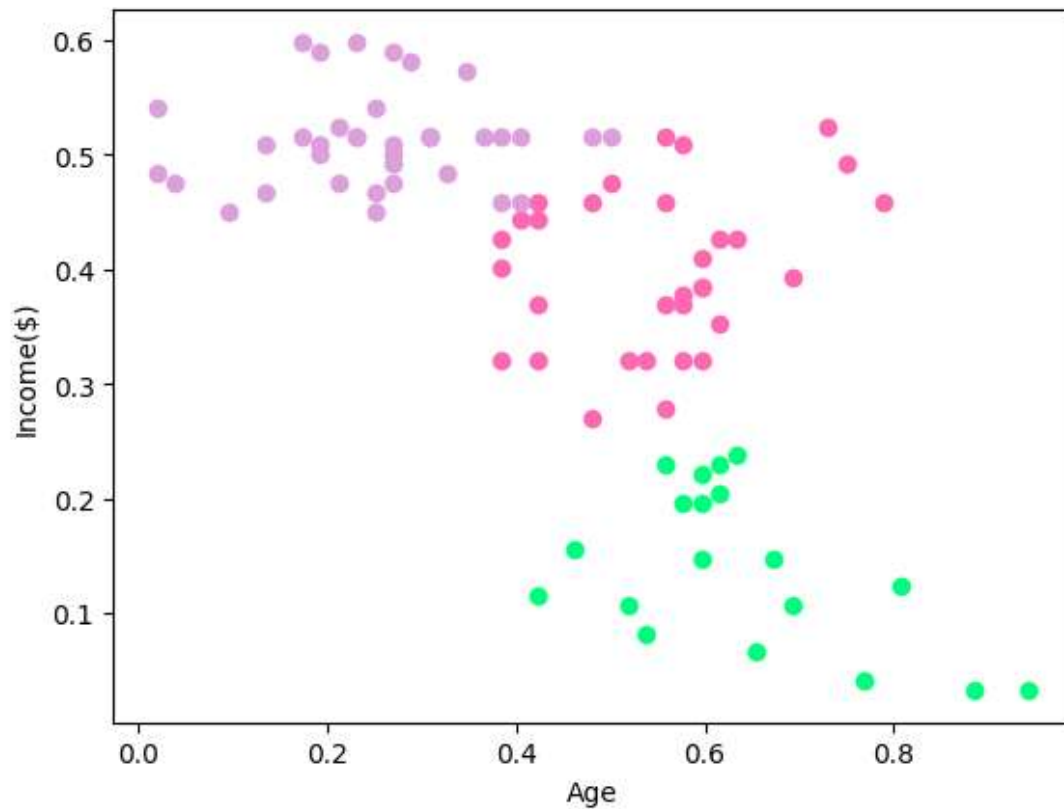
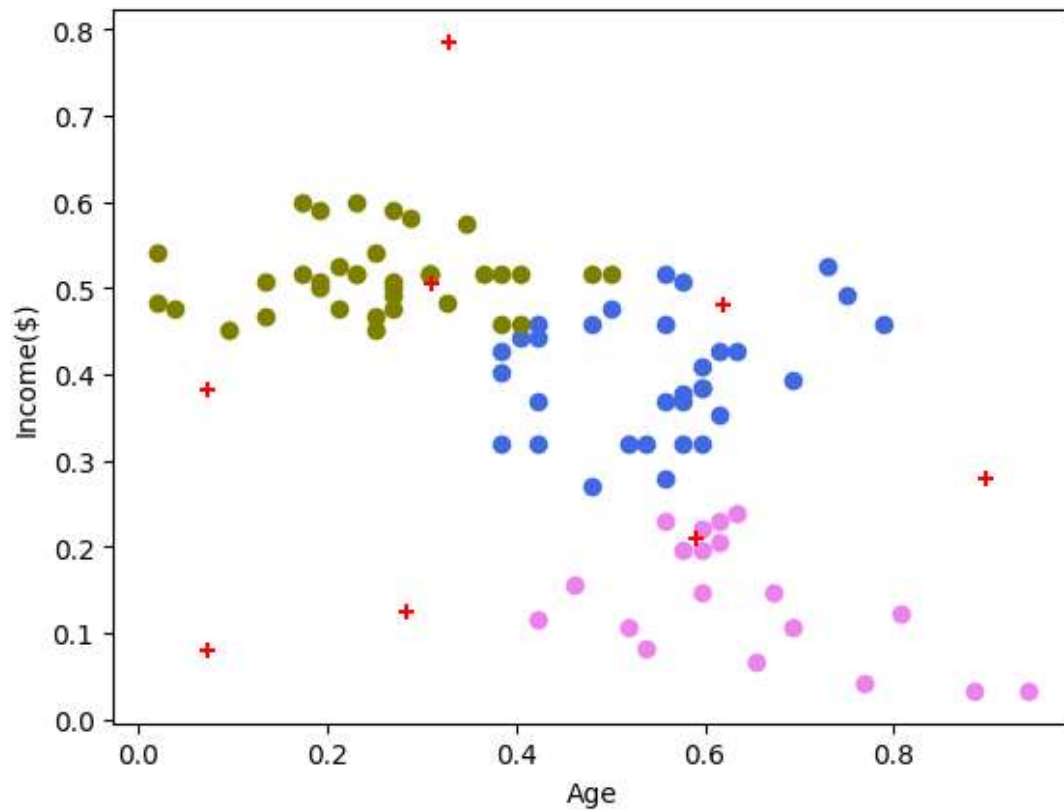In [14]:
```python
km.cluster_centers_
```

Out[14]:
```
array([[0.28388278, 0.1245121 ],
       [0.30944056, 0.50428465],
       [0.58974359, 0.20969945],
       [0.07322485, 0.38272383],
       [0.32905983, 0.78551913],
       [0.62037037, 0.47996357],
       [0.89799331, 0.28011404],
       [0.07239819, 0.08003857]])
```

```
In [16]: df1=df[df.cluster==0]
         df2=df[df.cluster==1]
         df3=df[df.cluster==2]
         plt.scatter(df1["Age"],df1["Income($)"],color="royalblue")
         plt.scatter(df2["Age"],df2["Income($)"],color="violet")
         plt.scatter(df3["Age"],df3["Income($)"],color="olive")
         plt.scatter(km.cluster_centers_[:,0],km.cluster_centers_[:,1],color="red",marker="+"
         plt.xlabel("Age")
         plt.ylabel("Income($)")
```

Out[16]: Text(0, 0.5, 'Income($)')

In [17]:
```python
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
```

```
C:\Users\ubinl\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\cl
uster\_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\ubinl\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\cl
uster\_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\ubinl\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\cl
uster\_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\ubinl\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\cl
uster\_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\ubinl\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\cl
uster\_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\ubinl\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\cl
uster\_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\ubinl\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\cl
uster\_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\ubinl\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\cl
uster\_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\ubinl\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\cl
uster\_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[17]:
```
[23.583906150363607,
 13.028938428018286,
 7.492113413237458,
 6.055858644812547,
 4.713025598595382,
 3.8596812260865807,
 3.0547174363693586,
 2.652565114951915,
 2.3135720353543285]
```

In [18]:
```python
plt.plot(k_rng,sse)
plt.xlabel("k")
plt.ylabel("sum of squared Error")
```

Out[18]: Text(0, 0.5, 'sum of squared Error')