

In [1]:

```
from sklearn.cluster import KMeans
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
from matplotlib import pyplot as plt
%matplotlib inline
```

In [3]:

```
df=pd.read_csv("https://raw.githubusercontent.com/codebasics/py/master/ML/13_kmeans/income.csv")
df.head()
```

Out[3]:

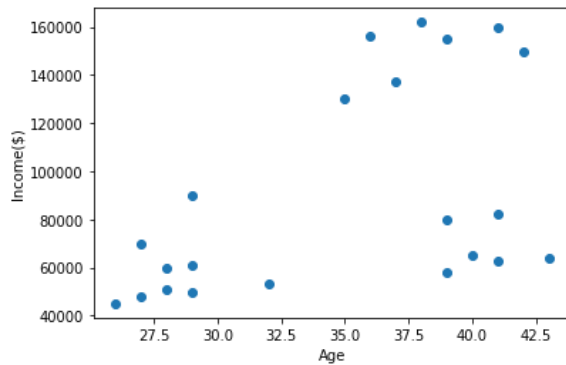
	Name	Age	Income(\$)
0	Rob	27	70000
1	Michael	29	90000
2	Mohan	29	61000
3	Ismail	28	60000
4	Kory	42	150000

In [4]:

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

Out[4]:

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



In [5]:

```
km = KMeans(n_clusters=3)
y_predicted = km.fit_predict(df[['Age','Income($)']])
y_predicted
```

Out[5]:

array([0, 0, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 0, 0, 2])

In [6]:

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

Out[6]:

	Name	Age	Income(\$)	cluster
0	Rob	27	70000	0
1	Michael	29	90000	0
2	Mohan	29	61000	2
3	Ismail	28	60000	2
4	Kory	42	150000	1

In [7]:

km.cluster_centers_

Out[7]:

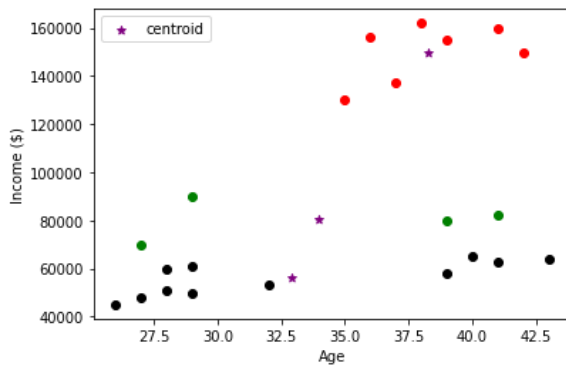
```
array([[3.40000000e+01, 8.05000000e+04],
       [3.82857143e+01, 1.50000000e+05],
       [3.29090909e+01, 5.61363636e+04]])
```

In [8]:

```
df1 = df[df.cluster==0]
df2 = df[df.cluster==1]
df3 = df[df.cluster==2]
plt.scatter(df1.Age,df1['Income($)',color='green')
plt.scatter(df2.Age,df2['Income($)',color='red')
plt.scatter(df3.Age,df3['Income($)',color='black')
plt.scatter(km.cluster_centers_[0],km.cluster_centers_[1],color='purple',marker='*',label='centroid')
plt.xlabel('Age')
plt.ylabel('Income ($)')
plt.legend()
```

Out[8]:

<matplotlib.legend.Legend at 0x21c9406fc10>



In [9]:

```
scaler = MinMaxScaler()

scaler.fit(df[['Income($)']])
df['Income($)'] = scaler.transform(df[['Income($)']])

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

In [10]:

df.head()

Out[10]:

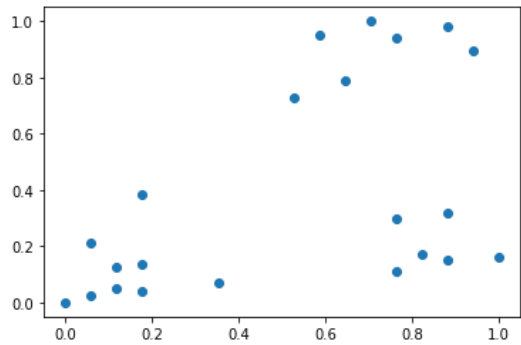
	Name	Age	Income(\$)	cluster
0	Rob	0.058824	0.213675	0
1	Michael	0.176471	0.384615	0
2	Mohan	0.176471	0.136752	2
3	Ismail	0.117647	0.128205	2
4	Kory	0.941176	0.897436	1

In [11]:

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

Out[11]:

<matplotlib.collections.PathCollection at 0x21c940f3850>



In [12]:

```
km = KMeans(n_clusters=3)
y_predicted = km.fit_predict(df[['Age','Income($)']])
y_predicted
```

Out[12]:

array([1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2])

In [13]:

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

Out[13]:

	Name	Age	Income(\$)	cluster
0	Rob	0.058824	0.213675	1
1	Michael	0.176471	0.384615	1
2	Mohan	0.176471	0.136752	1
3	Ismail	0.117647	0.128205	1
4	Kory	0.941176	0.897436	0

In [14]:

```
km.cluster_centers_
```

Out[14]:

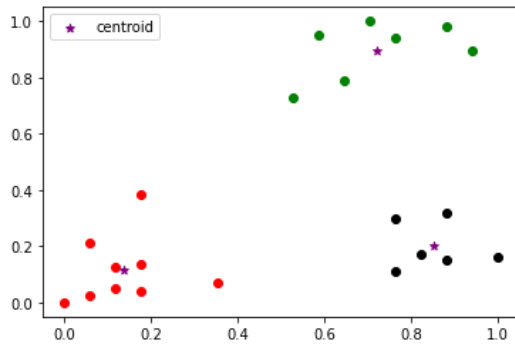
array([[0.72268908, 0.8974359],
 [0.1372549 , 0.11633428],
 [0.85294118, 0.2022792]])

In [15]:

```
df1 = df[df.cluster==0]
df2 = df[df.cluster==1]
df3 = df[df.cluster==2]
plt.scatter(df1.Age,df1['Income($)'],color='green')
plt.scatter(df2.Age,df2['Income($)'],color='red')
plt.scatter(df3.Age,df3['Income($)'],color='black')
plt.scatter(km.cluster_centers_[0],km.cluster_centers_[1],color='purple',marker='*',label='centroid')
plt.legend()
```

Out[15]:

<matplotlib.legend.Legend at 0x21c9416c100>



In []: