Assignment on K-mean clustering.

Apply K-mean clustering on Income data set to form 3 Clusters and display there clasters using scatter graph.

In [41]:
```
from sklearn.cluster import KMeans
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
from matplotlib import pyplot as plt
%matplotlib inline
```

Loading Dataset

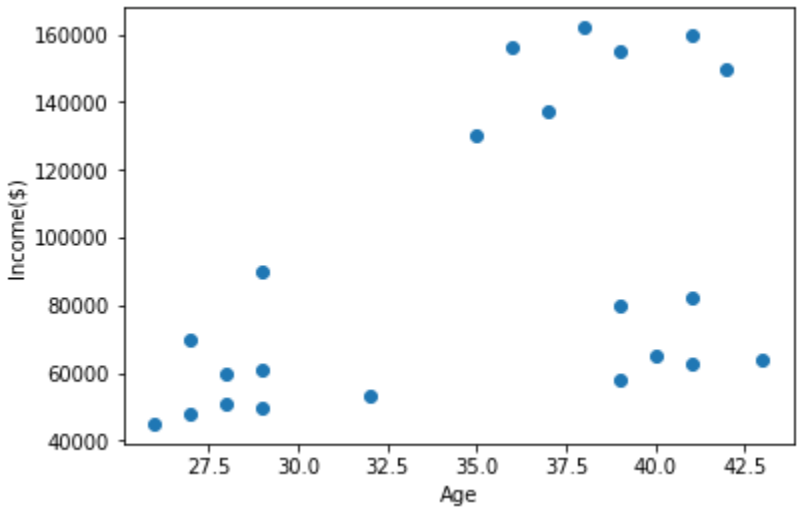
In [42]:
```
df = pd.read_csv("income.csv")
df.head()
```

Out[42]:

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

In [43]:
```
plt.scatter(df.Age,df['Income($)'])
plt.xlabel('Age')
plt.ylabel('Income($)')
```

Out[43]: Text(0,0.5,'Income($)')



Create a KMeans instance with 3 clusters

In [44]:
```
km = KMeans(n_clusters=3)
y_predicted = km.fit_predict(df[['Age','Income($)']])
y_predicted
```

Out[44]: array([2, 2, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0])

In [45]:
```
df['cluster']=y_predicted
df.head()
```

Out[45]:

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

In [46]:
```
km.cluster_centers_
```

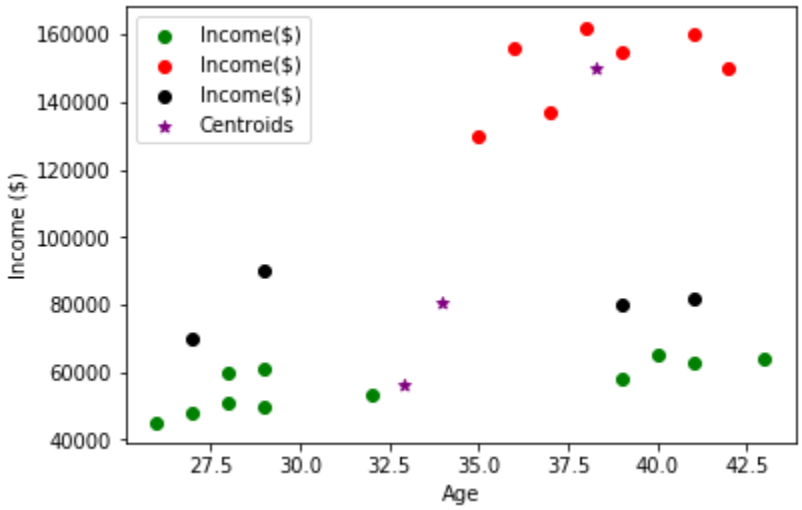
Out[46]:
```
array([[3.29090909e+01, 5.61363636e+04],
       [3.82857143e+01, 1.50000000e+05],
       [3.40000000e+01, 8.05000000e+04]])
```

Visualize the results

[47]:
```
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($)'],color='black')
plt.scatter(km.cluster_centers_[:, 0], km.cluster_centers_[:, 1], color='purple',
plt.xlabel('Age')
plt.ylabel('Income ($)')
plt.legend()
```

Out[47]: <matplotlib.legend.Legend at 0x2a35ad29dd8>

In



Preprocessing using min max scaler

In [48]:
```
scaler = MinMaxScaler()
scaler.fit(df[['Income($)']])
df['Income($)'] = scaler.transform(df[['Income($)']])
scaler.fit(df[['Age']])
df['Age'] = scaler.transform(df[['Age']])
```

In [49]:
```
df.head()
```

Out[49]:

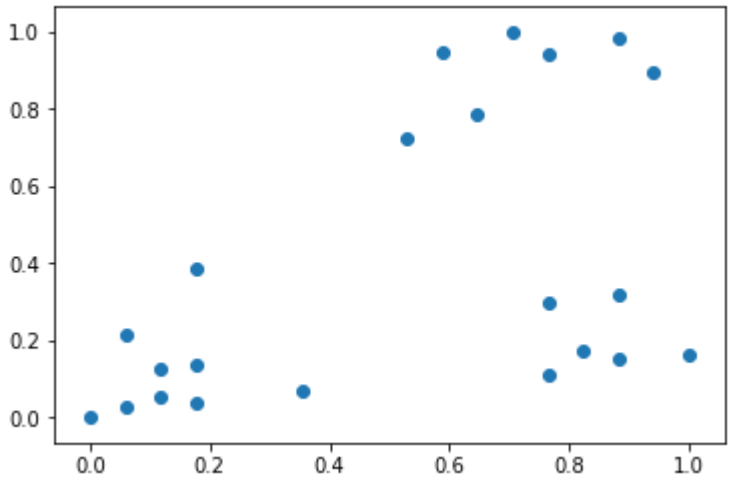
	Name	Age	Income($)	cluster
0	Rob0.058824	0.213675		2
1	Michael0.176471	0.384615		2
2	Mohan0.176471	0.136752		0
3	Ismail0.117647	0.128205		0
4	Kory0.941176	0.897436		1

Plot the clustered data

[50]:
```
plt.scatter(df.Age,df['Income($)'])
```

Out[50]: `<matplotlib.collections.PathCollection at 0x2a35ad96940>`

In



In [51]:
```
km = KMeans(n_clusters=3)
y_predicted = km.fit_predict(df[['Age','Income($)']])
y_predicted
```

Out[51]:
```
array([1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0])
```

In [52]:
```
df['cluster']=y_predicted
df.head()
```

Out[52]:

	Name	Age	Income($)	cluster
0	Rob0.058824	0.213675		1
1	Michael0.176471	0.384615		1
2	Mohan0.176471	0.136752		1
3	Ismail0.117647	0.128205		1
4	Kory0.941176	0.897436		2

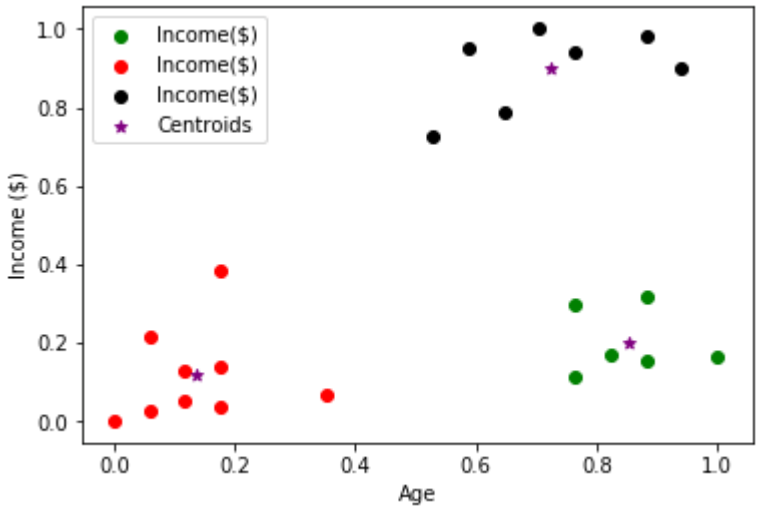
In [53]:
```
km.cluster_centers_
```

Out[53]:
```
array([[0.85294118, 0.2022792 ],
       [0.1372549 , 0.11633428],
       [0.72268908, 0.8974359 ]])
```

[54]:
```
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($)'],color='black')
plt.scatter(km.cluster_centers_[:, 0], km.cluster_centers_[:, 1], color='purple',
plt.xlabel('Age') plt.ylabel('Income ($)') plt.legend()
```

Out[54]:
```
<matplotlib.legend.Legend at 0x2a35adf7f98>
```

In



In [55]:
```
sse = []
k_rng = range(1,10)
for k in k_rng:
 km = KMeans(n_clusters=k)
 km.fit(df[['Age','Income($)']])
 sse.append(km.inertia_)
sse
```

Out[55]:
```
[5.434011511988179,
 2.091136388699078,
 0.4750783498553095,
 0.3491047094419565,
 0.2664030124668416,
 0.2173883310613267,
 0.18275153026579993,
 0.13265419827245162,
 0.10188787724979426]
```

[56]:
```
plt.xlabel('K')
plt.ylabel('Sum of squared error')
plt.plot(k_rng,sse)
```

Out[56]:
```
[<matplotlib.lines.Line2D at 0x2a35ae5d5f8>]
```

In