In [3]:
```
import pandas as pd
from matplotlib import pyplot as plt
%matplotlib  inline
```

In [4]:
```
df=pd.read_csv(r"C:\Users\Jayadeep\Downloads\Income.csv")
df
```

Out[4]:

	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

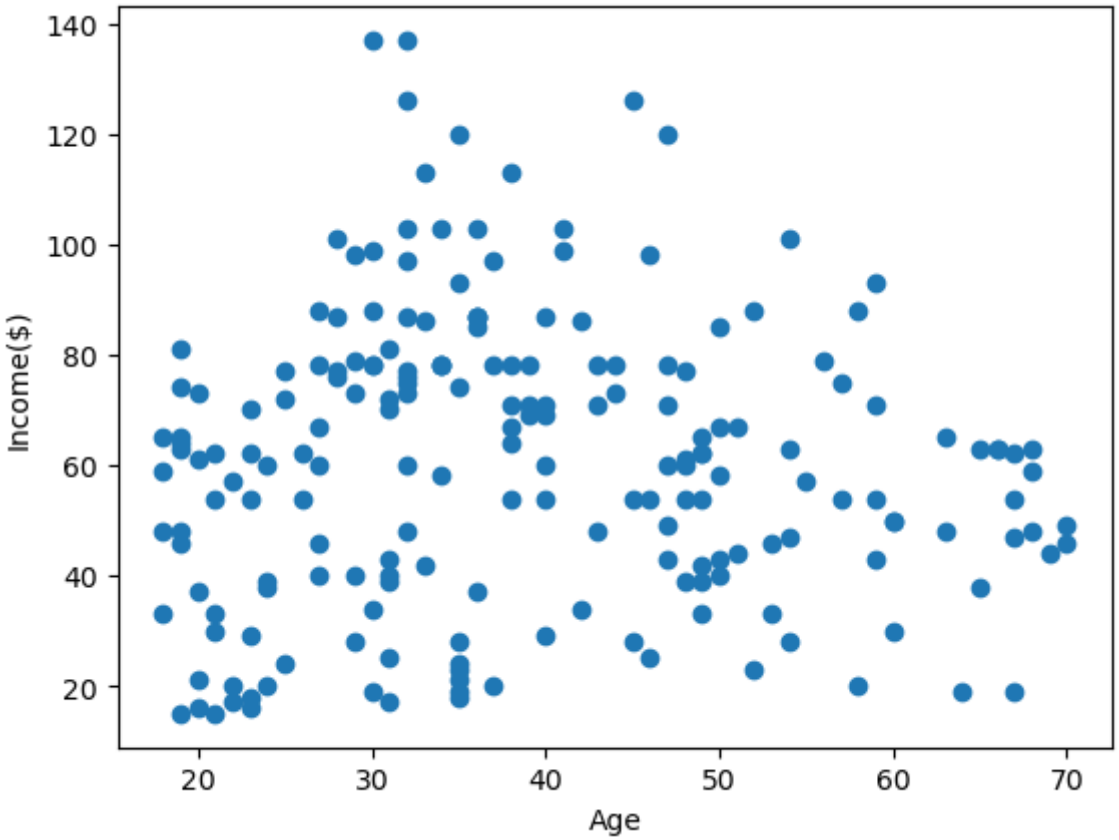
In [5]:
```
df.head()
```

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

In [6]:
```
plt.scatter(df["Age"],df["Income($)"])
plt.xlabel("Age")
plt.ylabel("Income($)")
```

Out[6]:
```
Text(0, 0.5, 'Income($)')
```



In [7]:
```
from sklearn.cluster import KMeans
```

In [8]:
```
km=KMeans()
km
```

Out[8]:
```
KMeans()
```

In [10]:
```
y_predicted=km.fit_predict(df[["Age","Income($)"]])
y_predicted
```

Out[10]:
```
array([2, 2, 2, 2, 2, 2, 2, 2, 6, 2, 6, 2, 6, 2, 2, 2, 2, 2, 6, 2, 2, 2,
       6, 2, 6, 2, 6, 2, 6, 2, 6, 2, 6, 5, 6, 5, 6, 5, 5, 5, 6, 5, 6, 5,
       6, 5, 6, 5, 5, 5, 6, 5, 5, 6, 6, 6, 6, 0, 5, 6, 0, 5, 0, 6, 0, 5,
       6, 0, 5, 5, 0, 6, 0, 0, 0, 5, 1, 1, 5, 1, 0, 1, 0, 1, 5, 1, 0, 5,
       1, 1, 0, 7, 1, 1, 7, 7, 1, 7, 1, 7, 7, 1, 0, 7, 1, 7, 0, 1, 0, 0,
       0, 7, 1, 7, 7, 7, 0, 1, 1, 1, 7, 1, 1, 1, 7, 7, 1, 1, 1, 1, 1, 1,
       7, 7, 7, 7, 1, 7, 7, 7, 1, 7, 7, 7, 7, 7, 1, 7, 7, 7, 1, 7, 1, 7,
       1, 7, 7, 7, 7, 7, 1, 7, 7, 7, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
       3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4,
       4, 4])
```

In [11]:
```
df["cluster"]=y_predicted
```

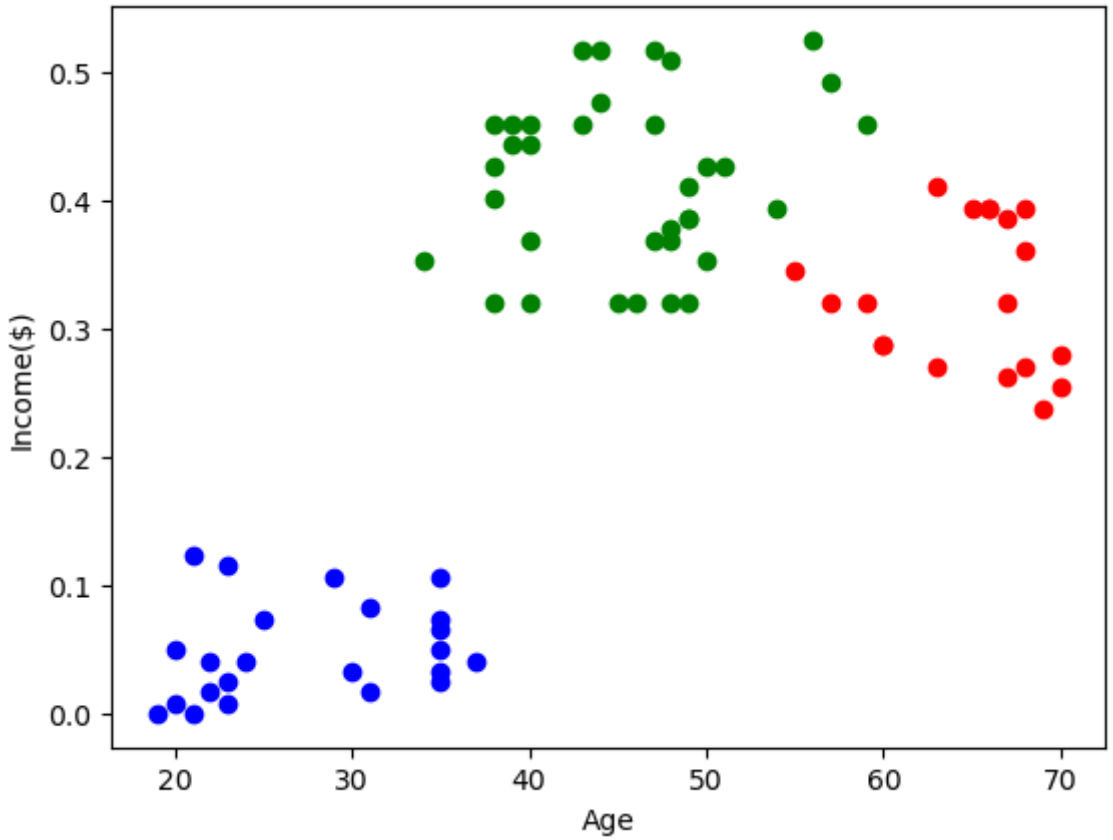
In [12]:
```
df.head()
```

Out[12]:

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

In [22]:
```
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[22]:
```
Text(0, 0.5, 'Income($)')
```



In [23]:
```
from sklearn.preprocessing import MinMaxScaler
```

In [24]:
```
Scaler=MinMaxScaler()
```

In [25]:
```
Scaler.fit(df[["Income($)"]])
df["Income($)"]=Scaler.transform(df[["Income($)"]])
df.head()
```

Out[25]:

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

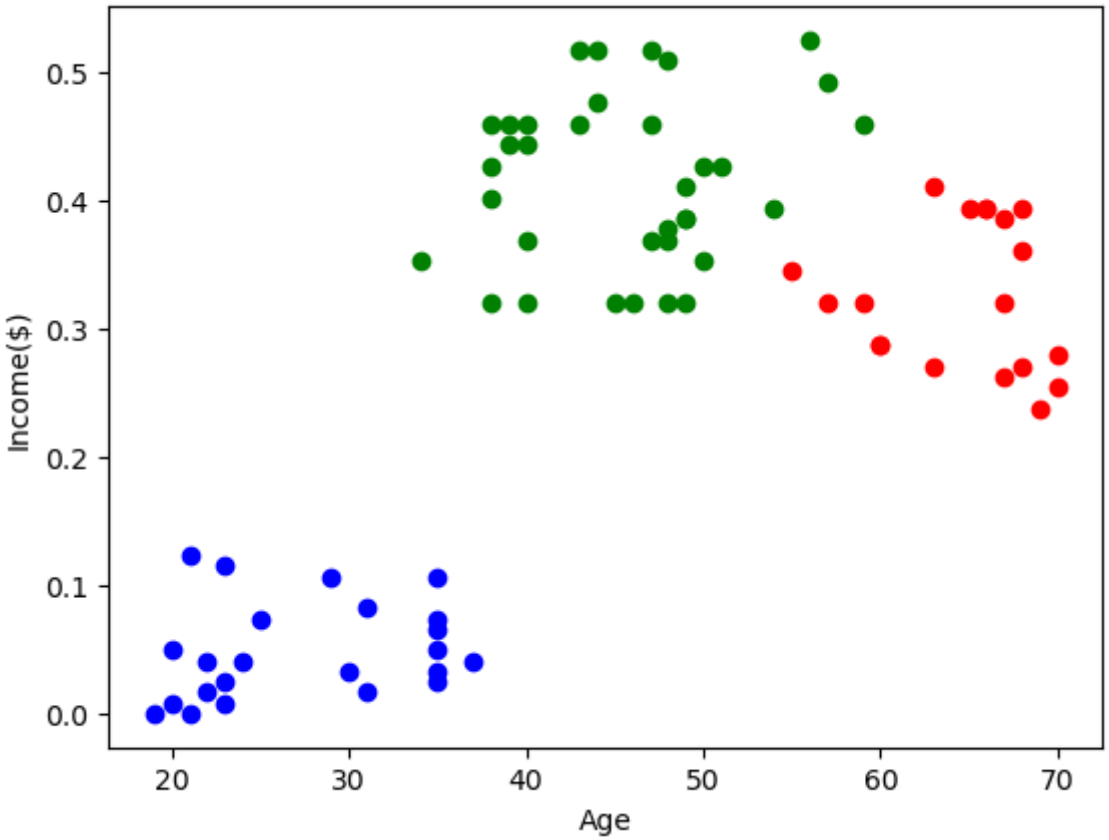
In [26]:
```
km=KMeans()
```

In [27]:
```
y_predicted=km.fit_predict(df[["Age","Income($)"]])
y_predicted
```

Out[27]:
```
array([5, 5, 5, 2, 7, 5, 0, 2, 3, 7, 3, 0, 1, 2, 0, 5, 0, 5, 4, 0, 0, 2,
       4, 7, 1, 7, 6, 0, 6, 2, 1, 5, 1, 5, 4, 5, 6, 7, 0, 5, 3, 2, 4, 7,
       4, 2, 4, 2, 7, 7, 4, 7, 7, 1, 4, 4, 4, 3, 2, 1, 3, 5, 3, 1, 3, 5,
       6, 3, 5, 7, 3, 4, 1, 1, 1, 2, 6, 6, 2, 4, 1, 0, 3, 4, 5, 4, 1, 5,
       0, 4, 3, 5, 4, 6, 7, 2, 4, 2, 4, 5, 2, 4, 3, 2, 4, 5, 3, 1, 3, 3,
       3, 5, 0, 5, 5, 5, 3, 4, 4, 4, 2, 0, 6, 0, 2, 7, 6, 6, 1, 0, 4, 0,
       2, 7, 5, 7, 6, 7, 5, 0, 1, 7, 2, 7, 2, 2, 4, 7, 0, 0, 6, 0, 6, 0,
       4, 2, 0, 7, 0, 7, 1, 7, 5, 7, 4, 0, 6, 7, 0, 7, 6, 2, 0, 0, 4, 7,
       1, 2, 1, 0, 0, 7, 4, 7, 6, 7, 1, 2, 6, 0, 0, 7, 7, 0, 4, 0, 6, 7,
       7, 7])
```

In [28]:
```
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[28]:
```
Text(0, 0.5, 'Income($)')
```

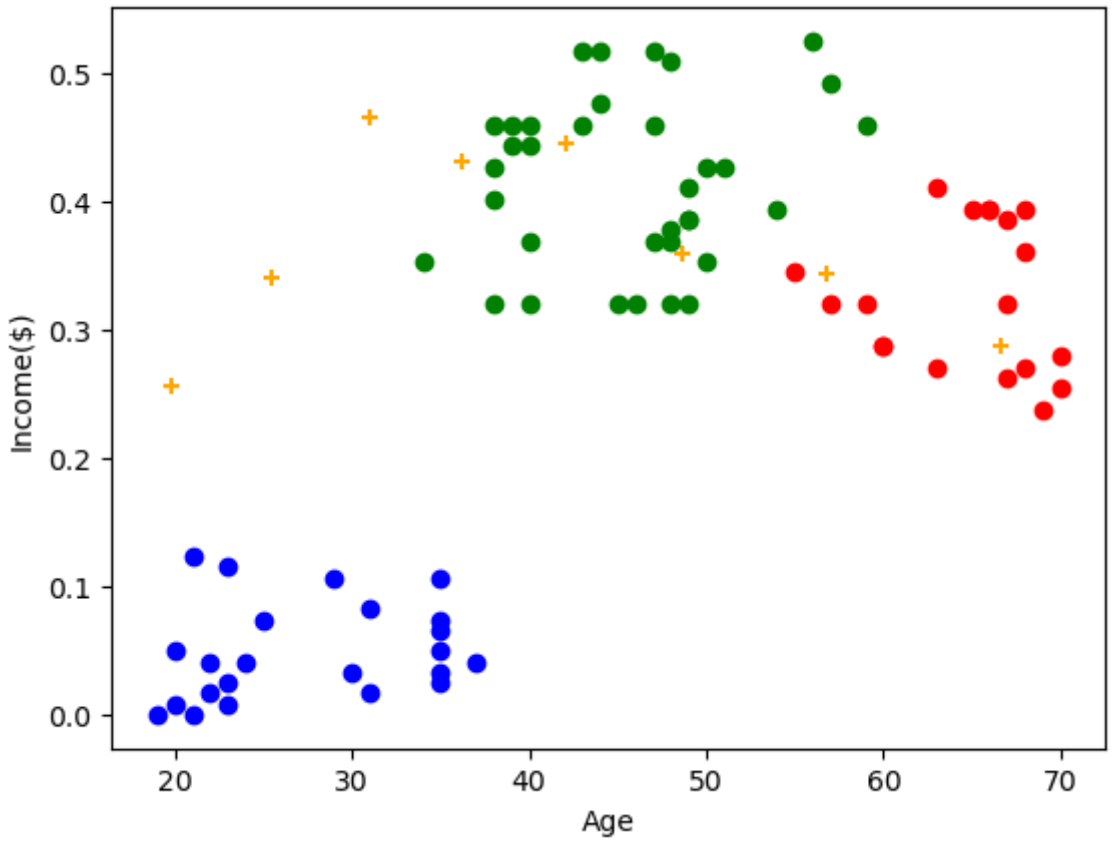


In [29]:
```
km.cluster_centers_
```

Out[29]:
```
array([[36.15625   ,  0.43058402],
       [56.78947368,  0.34383089],
       [25.4       ,  0.34098361],
       [66.64705882,  0.28688525],
       [48.63333333,  0.35819672],
       [19.8       ,  0.25639344],
       [42.11111111,  0.44535519],
       [31.        ,  0.46480231]])
```

In [30]:
```
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($)"],color="blue")
plt.scatter(km.cluster_centers_[:,0],km.cluster_centers_[:,1],color="orange",marker="+")
plt.xlabel("Age")
plt.ylabel("Income($)")
```

Out[30]:
```
Text(0, 0.5, 'Income($)')
```



In [32]:
```
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\Jayadeep\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:1036: UserWarning: KMeans is known to have
a memory leak on Windows with MKL, when there are less chunks than available threads. You can avoid it by setting th
e environment variable OMP_NUM_THREADS=1.
  warnings.warn(
```

Out[32]:
```
[38840.72314431605,
 10558.82532963463,
 5678.436799754746,
 2521.859262993656,
 1628.957209547009,
 1031.7292698285246,
 786.4533149505711,
 604.885525249197,
 474.9095178891467]
```

In [33]:
```
plt.plot(k_rng,sse)
plt.xlabel("k")
plt.ylabel("sum of squared error")
```

Out[33]:
```
Text(0, 0.5, 'sum of squared error')
```