k-means-clustering-10-11-02-pdf

February 13, 2025

[ ]:

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

[ ]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

[ ]:

```
file_path = ('/content/income.csv')
```

[ ]:

```
df = pd.read_csv(file_path)
```

[ ]:

```
df
```

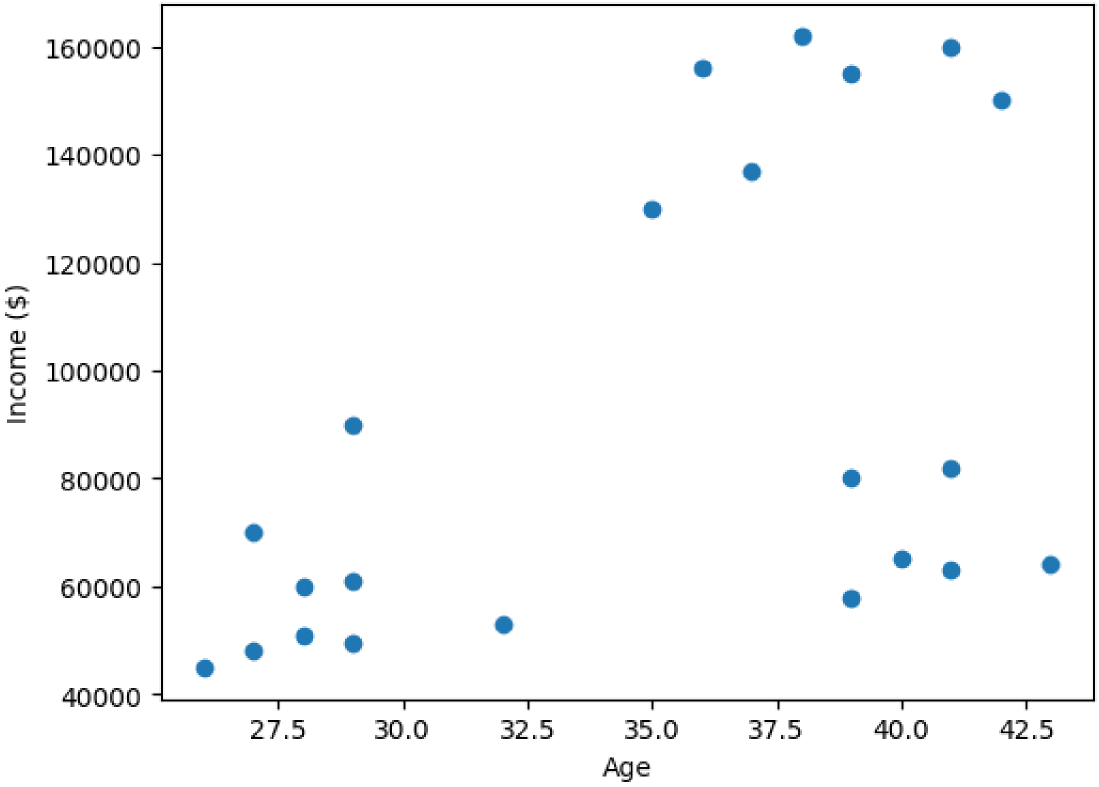
[ ]:

```
        Name  Age  Income($)
0        Rob   27      70000
1    Michael   29      90000
2      Mohan   29      61000
3     Ismail   28      60000
4       Kory   42     150000
5     Gautam   39     155000
6      David   41     160000
7     Andrea   38     162000
8       Brad   36     156000
9   Angelina   35     130000
10    Donald   37     137000
11       Tom   26      45000
12    Arnold   27      48000
13     Jared   28      51000
14     Stark   29      49500
15    Ranbir   32      53000
16    Dipika   40      65000
17  Priyanka   41      63000
18      Nick   43      64000
19      Alia   39      80000
```

```
20        Sid    41       82000
21      Abdul    39       58000
```

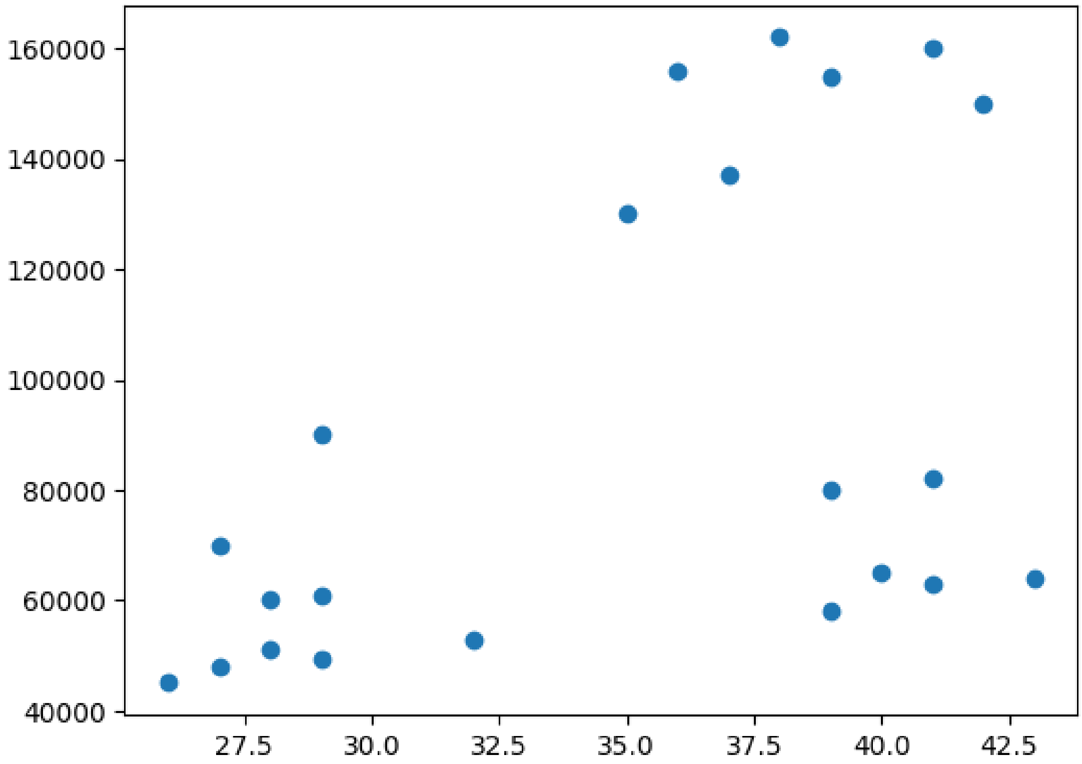
```
[ ]: plt.scatter(df.Age, df['Income($)'])
     plt.xlabel('Age')
     plt.ylabel('Income ($)')
```

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



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

```
[ ]: <matplotlib.collections.PathCollection at 0x7e83137fc150>
```



[ ]:

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

[ ]:

```
array([1, 1, 1, 1, 2, 2, 2, 2, 2, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1],
      dtype=int32)
```

[ ]:

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

[ ]:

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

[ ]:

```
km.cluster_centers_
```

[ ]:

```
array([[3.60000000e+01, 1.33500000e+05],
       [3.32000000e+01, 6.26333333e+04],
       [3.92000000e+01, 1.56600000e+05]])
```

[ ]:

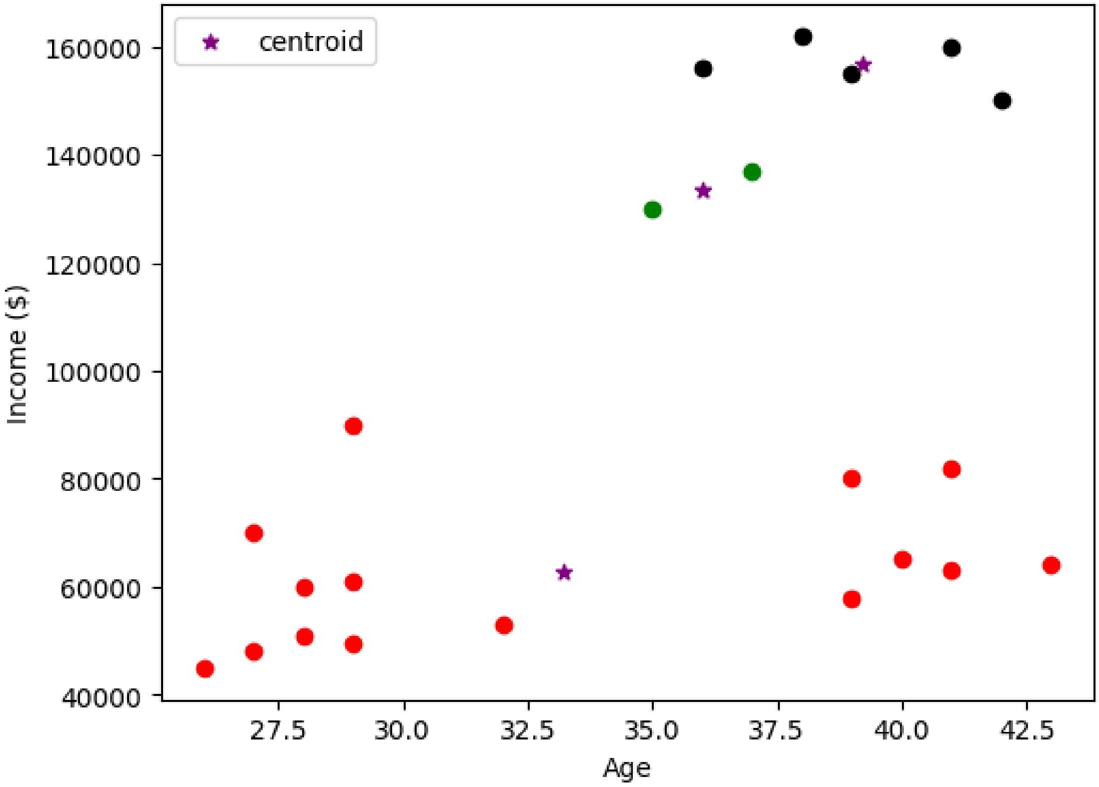
```
df1 = df[df.cluster==0]
df2 = df[df.cluster==1]
df3 = df[df.cluster==2]
```

[ ]:

```
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()
```

[ ]:

```
<matplotlib.legend.Legend at 0x7e831380e390>
```



[ ]:

```
scaler = MinMaxScaler()

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

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

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

[ ]:

```
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call
drive.mount("/content/drive", force_remount=True).
```

[ ]:

```
df.head()
```

[ ]:

```
      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        2
```

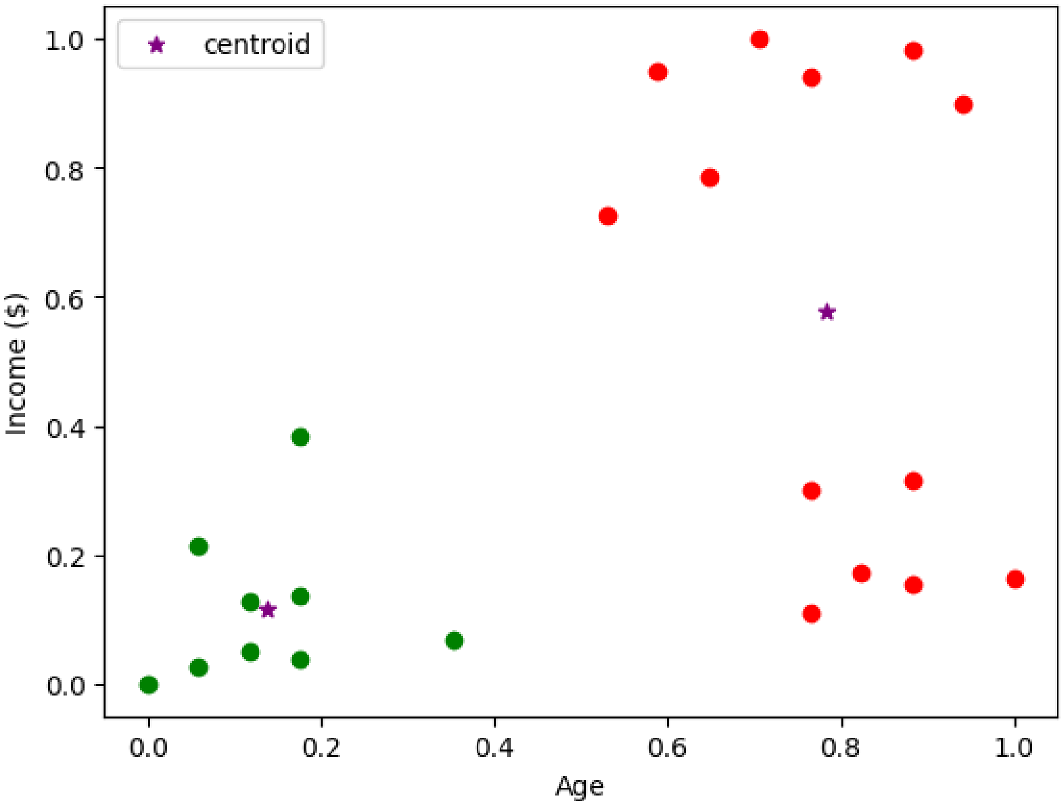
[ ]:

```
#kmeans cluster with k=2

import matplotlib.pyplot as plt
km = KMeans(n_clusters=2)
y_predicted = km.fit_predict(df[['Age', 'Income($)']])
y_predicted
df['cluster']=y_predicted
df.head()
km.cluster_centers_
df1 = df[df.cluster==0]
df2 = df[df.cluster==1]
plt.scatter(df1.Age, df1['Income($)'],color='green')
plt.scatter(df2.Age, df2['Income($)'],color='red')
plt.scatter(km.cluster_centers_[:,0],km.cluster_centers_[:
  ↪,1],color='purple',marker='*',label='centroid')
plt.xlabel('Age')
plt.ylabel('Income ($)')
plt.legend()
```

[ ]:

```
<matplotlib.legend.Legend at 0x7e830db21610>
```

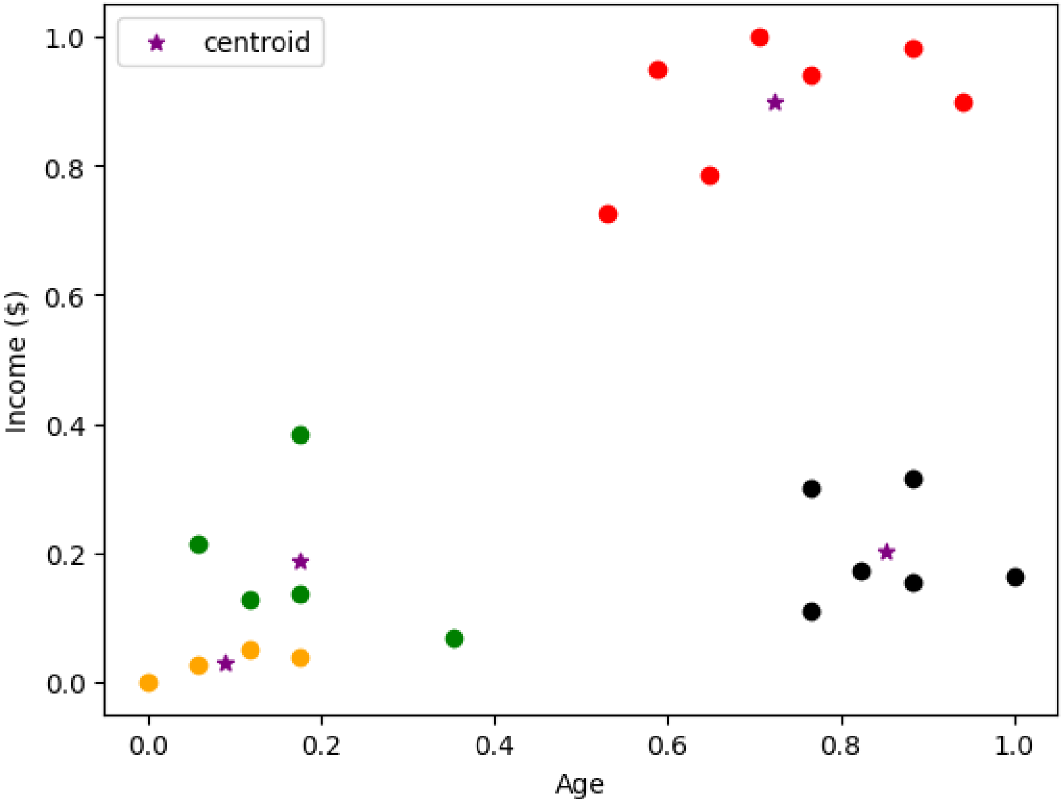


[ ]:

```
#Kmeans cluster with k =4

import matplotlib.pyplot as plt
km = KMeans(n_clusters=4)
y_predicted = km.fit_predict(df[['Age', 'Income($)']])
y_predicted
df['cluster']=y_predicted
df.head()
km.cluster_centers_
df1 = df[df.cluster==0]
df2 = df[df.cluster==1]
df3 = df[df.cluster==2]
df4 = df[df.cluster==3]
plt.scatter(df1.Age, df1['Income($)'],color='green')
plt.scatter(df2.Age, df2['Income($)'],color='red')
plt.scatter(df3.Age, df3['Income($)'],color='black')
plt.scatter(df4.Age, df4['Income($)'],color='orange')
plt.scatter(km.cluster_centers_[:,0],km.cluster_centers_[:
  ↪,1],color='purple',marker='*',label='centroid')
plt.xlabel('Age')
plt.ylabel('Income ($)')
plt.legend()
```

```
[ ]: <matplotlib.legend.Legend at 0x7e830dab0810>
```



```
[ ]:
```

```
[ ]:
```

```
[ ]:
```