Khai báo các thư viện:

```
import numpy as np
import pandas as pd
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
```

Đọc file dữ liệu

```
df = pd.read_csv("bank-data.csv")
df.shape  #Số lượng mẫu và Số thuộc tính quan sát
```

```
(600, 12)
```

```
df.head(10) # Xem 10 dòng dữ liệu đầu tiên
```

Xem kiểu dữ liệu của các thuộc tính

```
df.dtypes
```

```
ID              int64
age             int64
sex            object
region         object
income        float64
married        object
children        int64
car            object
save_act       object
current_act    object
mortgage       object
pep            object
dtype: object
```

Xóa cột ID

```
df.drop('ID',axis=1)
```

	age	sex	region	income	married	children	car	save_act	current_act	mortgage	pep
0	48	0	0	17546.00	0	1	0	0	0	0	1
1	40	1	3	30085.10	1	3	1	0	1	1	0
2	51	0	0	16575.40	1	0	1	1	1	0	0
3	23	0	3	20375.40	1	3	0	0	1	0	0

Đổi dữ liệu từ dạng định danh (categorical) về dạng số

```
from sklearn.preprocessing import LabelEncoder
lb_make = LabelEncoder()
df["sex"] = lb_make.fit_transform(df["sex"])
df["region"] = lb_make.fit_transform(df["region"])
df["married"] = lb_make.fit_transform(df["married"])
df["car"] = lb_make.fit_transform(df["car"])
df["save_act"] = lb_make.fit_transform(df["save_act"])
df["current_act"] = lb_make.fit_transform(df["current_act"])
df["mortgage"] = lb_make.fit_transform(df["mortgage"])
df["pep"] = lb_make.fit_transform(df["pep"])
df.head(10)
```

	ID	age	sex	region	income	married	children	car	save_act	current_act	mortgage	pep
0	1	48	0	0	17546.00	0	1	0	0	0	0	1

```
X = df.values[:, 0:df.shape[1]]
```

2	3	51	0	0	16575.40	1	0	1	1	1	0	0

```
clus = 3
kmeans = KMeans(n_clusters=clus).fit(X)
print('Centers found by scikit-learn:')
print(kmeans.cluster_centers_)
pred_label = kmeans.predict(X)
print(pred_label)
```

```
Centers found by scikit-learn:
[[3.03000000e+02 3.23699634e+01 5.12820513e-01 1.21611722e+00
  1.64878906e+04 6.81318681e-01 9.96336996e-01 4.32234432e-01
  6.08058608e-01 7.50915751e-01 3.44322344e-01 3.77289377e-01]
 [3.05009434e+02 5.90566038e+01 5.00000000e-01 1.13207547e+00
  4.92589066e+04 6.79245283e-01 1.07547170e+00 5.09433962e-01
  1.00000000e+00 8.01886792e-01 3.20754717e-01 5.94339623e-01]
 [2.95248869e+02 4.67873303e+01 4.84162896e-01 1.29864253e+00
  3.07320385e+04 6.24434389e-01 1.00000000e+00 5.61085973e-01
  6.42533937e-01 7.46606335e-01 3.66515837e-01 4.88687783e-01]]
[0 2 0 0 1 2 0 2 2 2 1 2 0 1 0 0 0 1 2 0 1 0 2 0 0 1 0 0 2 2 0 0 0 2 2 0 0
 0 0 2 2 2 1 2 0 1 0 0 2 0 0 2 0 2 1 1 2 0 0 2 2 2 2 0 0 0 0 2 0 0 0 0 2 0
 1 2 2 1 0 2 0 2 0 0 2 2 0 0 0 2 2 2 2 1 1 0 0 2 0 0 2 0 0 2 2 2 1 0 0 2 2
 2 2 2 1 1 0 0 0 1 2 1 0 2 1 2 2 1 0 0 0 1 0 2 1 2 0 2 1 1 2 0 0 1 0 1 0 0
 2 2 2 2 1 2 0 0 2 0 0 0 0 0 0 0 2 2 0 0 2 2 1 0 2 2 1 1 2 2 0 2 0 1 2 0 2
 2 0 2 2 0 1 0 1 0 2 0 0 1 0 0 2 1 0 0 0 0 1 0 2 0 2 2 0 0 0 0 0 2 2 2 2 1
 2 1 1 2 1 2 2 0 2 2 0 2 1 0 1 0 0 0 0 1 2 0 2 2 0 0 0 0 1 2 0 0 0 2 2 0 0
 0 0 2 0 1 0 0 2 2 2 0 1 1 0 2 2 0 2 0 2 2 0 0 1 2 0 0 2 2 0 2 0 2 1 0 0 2
 2 0 0 0 0 0 2 1 1 0 1 2 2 0 2 2 2 0 0 1 0 2 0 0 2 0 0 2 2 0 2 0 1 0 1 0 2
 0 0 2 2 2 2 0 2 1 1 0 2 1 2 1 2 2 2 2 2 1 2 1 0 0 2 0 1 0 2 0 1 0 2 1 0 2
 1 2 0 2 1 0 2 2 0 0 2 2 0 2 2 2 0 2 0 0 1 0 2 0 2 0 2 2 1 2 0 0 2 0 0 2 1
 0 0 0 0 0 1 0 1 0 0 0 2 0 1 1 0 1 0 2 2 2 0 2 2 1 0 2 0 0 0 2 0 0 2 1 2 2
 0 2 1 2 1 2 2 2 2 1 0 0 0 1 0 1 0 2 0 0 0 2 0 0 0 2 1 2 0 2 0 0 2 2 0 1 0
 1 2 0 0 1 0 0 0 0 2 0 2 1 1 2 1 1 1 0 2 2 2 0 1 2 0 0 1 0 0 0 1 0 2 0 0 2
 0 2 0 0 0 0 0 0 0 0 2 1 1 2 2 2 0 0 2 1 2 0 0 2 0 1 0 2 0 0 0 1 0 2 1 2 2
 0 0 2 2 2 0 0 1 0 0 2 0 0 2 0 2 2 2 0 2 2 0 1 0 2 1 0 0 2 0 0 0 2 1 0 2 2
 0 1 2 1 0 0 0 2]
```

Vẽ đồ thị minh họa kết quả phân cụm

```
plt.xlabel('X')
plt.ylabel('y')
plt.title('Kết quả phân cụm')
plt_colors = ['b', 'g', 'r', 'c', 'm', 'y', 'k', 'w'] # danh sách các màu hỗ trợ

for i in range(clus):
    # lấy dữ liệu của cụm i
    data = X[pred_label == i]
    # Vẽ cụm i lên đồ thị
    plt.plot(data[:, 1], data[:, 4], plt_colors[i] + '^', markersize = 4, label = 'cluster' + str(i))
    # Vẽ tâm cụm i lên đồ thị
    plt.plot(kmeans.cluster_centers_[i][0], kmeans.cluster_centers_[i][1],plt_colors[i+4] +
                                                    'o', markersize = 10, label = 'center' + str(i))
plt.legend() # Hiện bảng chú thích
plt.show()
```