Chapter 12 - Exercise 2: Xclara

Cho dữ liệu như xclare.csv

1. Đọc dữ liệu, chuẩn hóa dữ liệu nếu cần
2. Dùng dendrogram để xác định số nhóm/cụm
3. Áp dụng thuật toán
4. Trực quan hóa kết quả, nhận xét
5. Cho X_test = np.array([[20, 20], [40,60], [70,5]]), cho biết những phần tử này thuộc cụm nào?

In [1]:
```
# from google.colab import drive
# drive.mount("/content/gdrive", force_remount=True)
```

In [2]:
```
# %cd '/content/gdrive/My Drive/LDS6_MachineLearning/practice/Chapter12_Hierarch
```

In [3]:
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

In [4]:
```
data = pd.read_csv("xclara.csv", sep=",")
print(data.shape)
data.head()
```

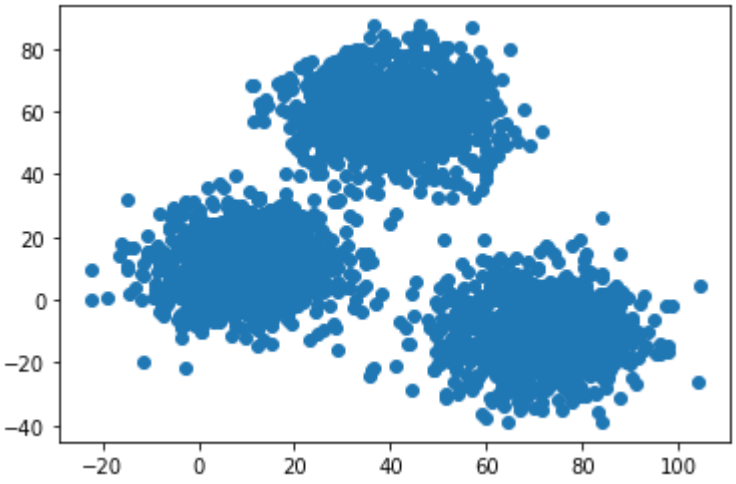
```
(3000, 2)
```

Out[4]:

	V1	V2
0	2.072345	-3.241693
1	17.936710	15.784810
2	1.083576	7.319176
3	11.120670	14.406780
4	23.711550	2.557729

In [5]:
```
plt.scatter(data.V1, data.V2)
```

Out[5]:
```
<matplotlib.collections.PathCollection at 0x1d3563faf60>
```



In [6]:
```
import scipy.cluster.hierarchy as shc
plt.figure(figsize=(10,6))
plt.title("Customer Dendograms")
dend = shc.dendrogram(shc.linkage(data, method='ward'))
plt.show()
```

In [7]:
```
# cluster = 3
from sklearn.cluster import AgglomerativeClustering
cluster = AgglomerativeClustering(n_clusters=3,
                                  affinity='euclidean',
                                  linkage='ward')
cluster.fit(data)
```

Out[7]:
```
AgglomerativeClustering(affinity='euclidean', compute_full_tree='auto',
                        connectivity=None, distance_threshold=None,
                        linkage='ward', memory=None, n_clusters=3,
                        pooling_func='deprecated')
```

In [8]:
```
cluster.labels_
```

Out[8]:
```
array([2, 2, 2, ..., 1, 1, 1], dtype=int64)
```

In [9]:
```
X_test = np.array([[20, 20], [40,60], [70,5]])
```

In [10]:
```
plt.figure(figsize=(8,8))
plt.scatter(data.V1, data.V2, c=cluster.labels_, cmap='rainbow')
plt.scatter(X_test[:,0],X_test[:,1], color="black", marker='s')
plt.xlabel("V1")
plt.ylabel("v2")
plt.show()
```