Experiment No. 6
Apply hierarchical clustering on the Wholesale Customers Dataset
Date of Performance:
Date of Submission:

Aim: Apply appropriate Unsupervised Learning Technique on the Wholesale Customers Dataset.

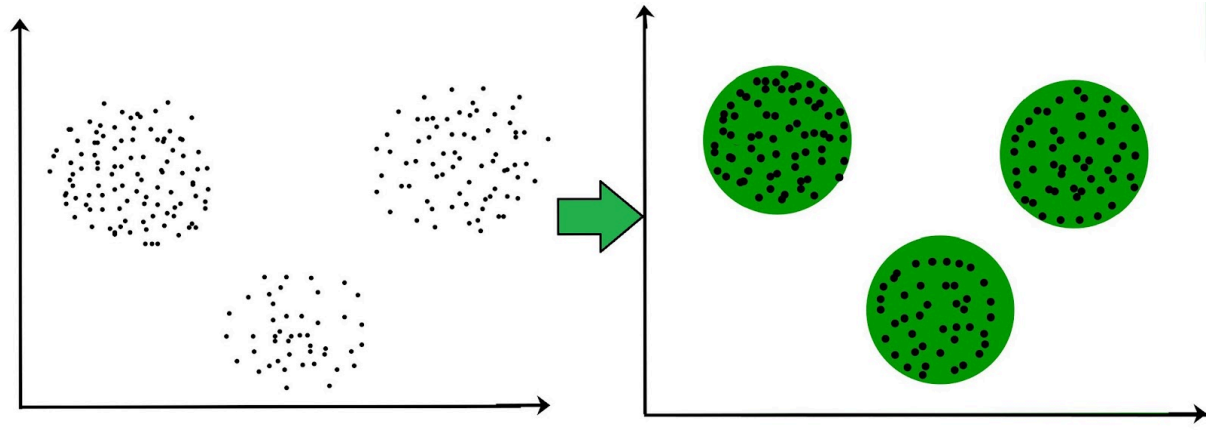
Objective: Able to perform various feature engineering tasks, apply Clustering Algorithm on the given dataset.

Theory:

It is basically a type of unsupervised learning method. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

For ex– The data points in the graph below clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the below picture.



Dataset:

This data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories. The wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The dataset consist of 440 large retailers annual spending on 6 different varieties of product in 3 different regions (lisbon , oporto, other) and across different sales channel (Hotel, channel)

Detailed overview of dataset

Records in the dataset = 440 ROWS

Columns in the dataset = 8 COLUMNS

FRESH: annual spending (m.u.) on fresh products (Continuous)

MILK:- annual spending (m.u.) on milk products (Continuous)

GROCERY:- annual spending (m.u.) on grocery products (Continuous)

FROZEN:- annual spending (m.u.) on frozen products (Continuous)

DETERGENTS_PAPER :- annual spending (m.u.) on detergents and paper products (Continuous)

DELICATESSEN:- annual spending (m.u.)on and delicatessen products (Continuous);

CHANNEL: - sales channel Hotel and Retailer

REGION:- three regions (Lisbon, Oporto, Other)

Code & Result:

```
#Hierarchical Clustering, ,

import os

import pandas as pd

import matplotlib.pyplot as plt
```

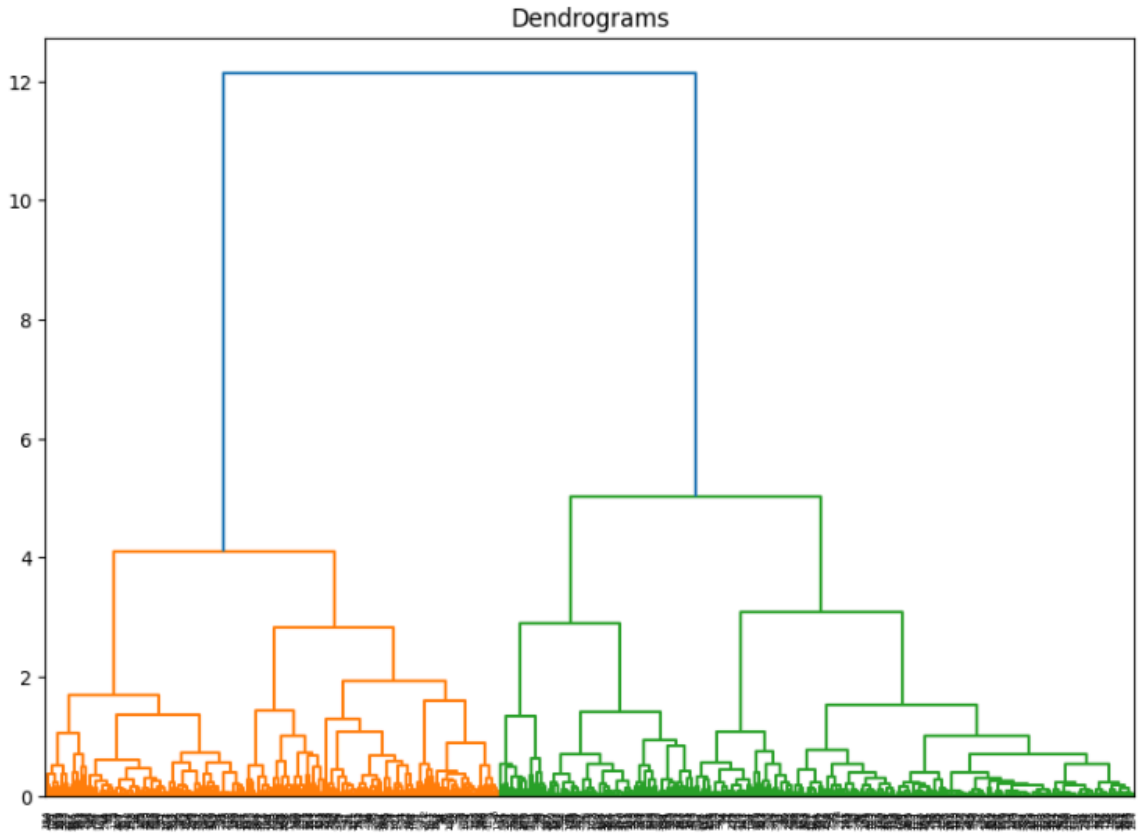
```
from sklearn.preprocessing import normalize

import scipy.cluster.hierarchy as shc

from sklearn.cluster import AgglomerativeClustering

data = pd.read_csv('Wholesale customers data.csv')

print(data.head())
```

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	2	3	12669	9656	7561	214	2674	1338
1	2	3	7057	9810	9568	1762	3293	1776
2	2	3	6353	8808	7684	2405	3516	7844
3	1	3	13265	1196	4221	6404	507	1788
4	2	3	22615	5410	7198	3915	1777	5185

```
data_scaled = normalize(data)

data_scaled = pd.DataFrame(data_scaled, columns=data.columns)

print(data_scaled.head())
```

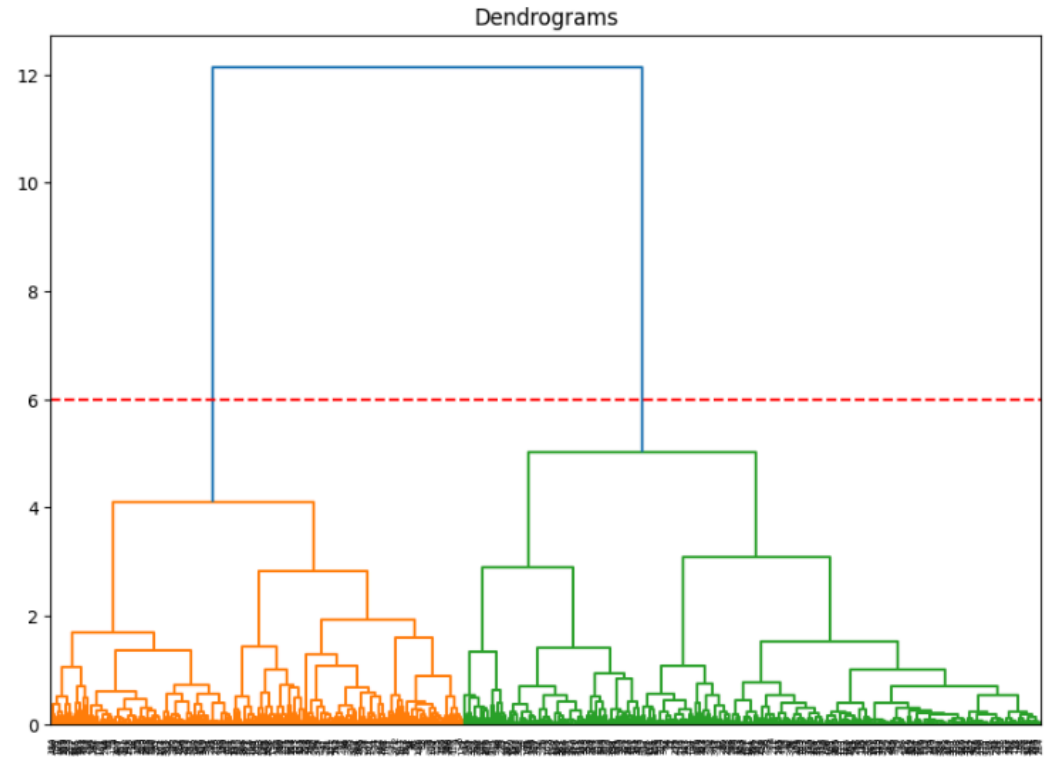
	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	0.000112	0.000168	0.708333	0.539874	0.422741	0.011965	0.149505	0.074809
1	0.000125	0.000188	0.442198	0.614704	0.599540	0.110409	0.206342	0.111286
2	0.000125	0.000187	0.396552	0.549792	0.479632	0.150119	0.219467	0.489619
3	0.000065	0.000194	0.856837	0.077254	0.272650	0.413659	0.032749	0.115494
4	0.000079	0.000119	0.895416	0.214203	0.284997	0.155010	0.070358	0.205294

```
plt.figure(figsize=(10, 7))

plt.title("Dendrograms")

d = shc.dendrogram(shc.linkage(data_scaled, method='ward'))
```




```
plt.figure(figsize=(10, 7))

plt.title("Dendrograms")

d = shc.dendrogram(shc.linkage(data_scaled, method='ward'))

plt.axhline(y=6, color='r', linestyle='--')
```

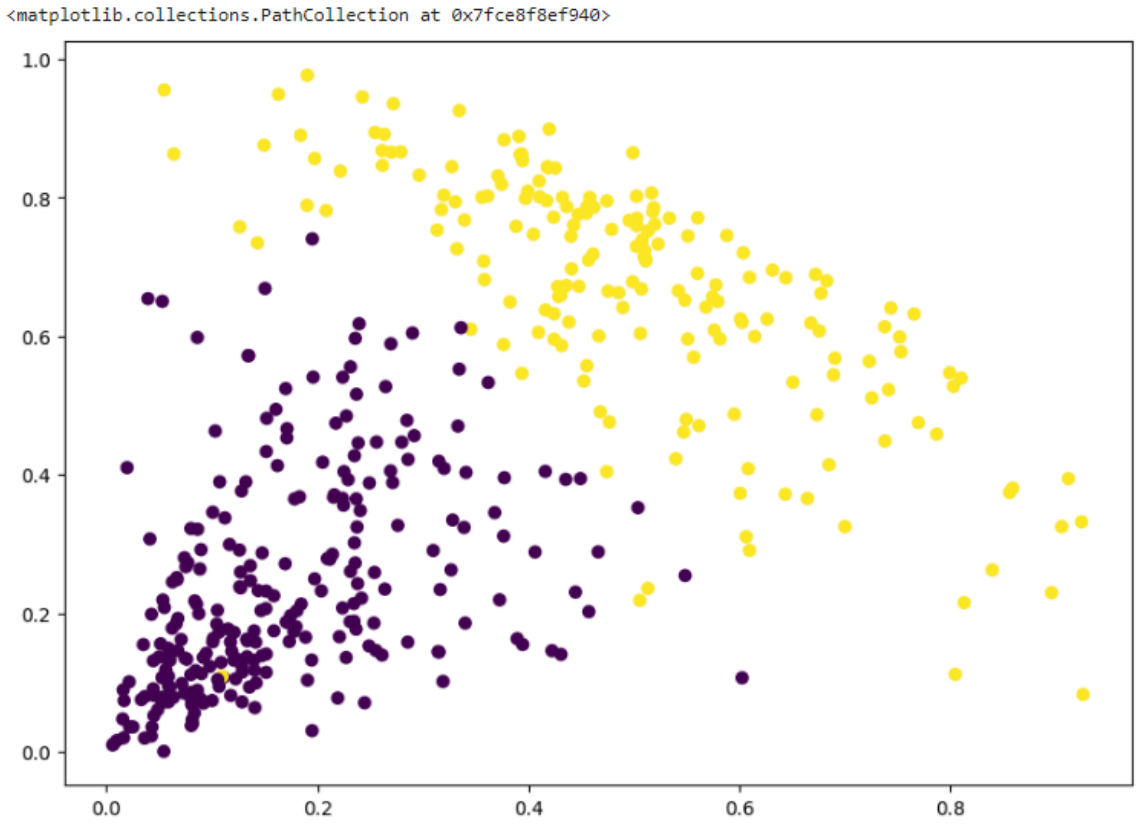
```
<matplotlib.lines.Line2D at 0x7fce8fcc7f70>
```




```
cluster = AgglomerativeClustering(n_clusters=2, linkage='ward')

print(cluster.fit_predict(data_scaled))
```

```
[1 1 1 0 0 1 0 1 1 1 1 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 0 1 1 0
 1 1 0 0 0 1 1 1 1 1 1 1 1 0 1 0 1 0 1 1 1 0 1 0 1 1 1 0 1 1 0 1 0 0 0 0 0
 1 0 0 1 0 1 0 1 1 0 0 1 1 0 0 0 0 0 0 1 0 1 1 1 0 0 0 1 1 1 0 0 0 1 1 1 1 0
 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 0 0 1 0 0
 0 0 0 1 0 1 0 1 1 0 1 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 1 1 0 0 0 1 0 0 1 1 1
 0 0 1 1 1 0 0 0 1 0 0 0 1 0 0 1 1 0 1 1 1 0 1 1 1 0 1 0 1 1 1 1 0 1 0 0 1
 0 0 0 0 0 0 1 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 1 0 0 0 0
 0 0 0 0 1 1 1 1 0 1 0 0 1 1 0 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0
 0 0 1 1 0 1 1 1 1 1 1 0 0 1 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 0 0 1 0 0 0 1 0
 1 0 0 0 0 0 0 1 1 1 1 0 1 1 0 1 1 0 1 1 1 1 0 1 0 1 1 1 0 0 1 0 0 1 0 0 0 0
 0 0 1 0 0 0 1 0 1 1 0 0 0 0 1 0 1 0 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
 1 0 0 0 1 1 0 0 1 1 1 1 1 0 1 0 0 0 0 0 1 0 1 0 1 0 0 1 0 0 0 1 0 1]
```

```
plt.figure(figsize=(10, 7))

plt.scatter(data_scaled['Milk'],                    data_scaled['Grocery'],
c=cluster.labels_)
```



Conclusion:

The above code performs hierarchical clustering on the "Wholesale customers data" dataset, normalizing the data for effective clustering. A dendrogram visualizes the hierarchical relationships, and a red dashed line indicates a suitable cutoff for forming two clusters. The AgglomerativeClustering model assigns each data point to one of these clusters, which are then visualized in a scatter plot based on 'Milk' and 'Grocery' purchases. The results reveal distinct customer segments, highlighting their differing purchasing behaviors. Overall, the combination of dendrograms and scatter plots effectively illustrates the clustering results and the data's underlying structure.