Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Experiment No. 5
Apply appropriate Unsupervised Learning Technique on the Wholesale Customers Dataset
Date of Performance:
Date of Submission:

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

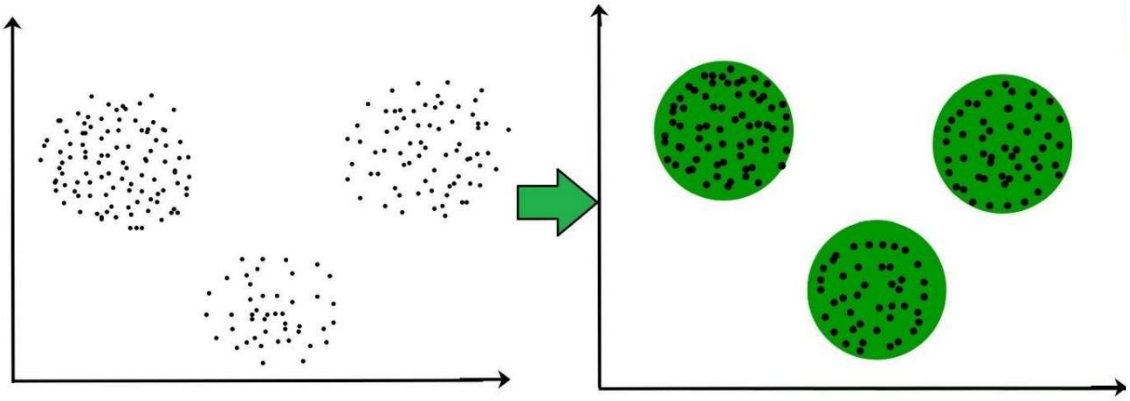
Objective: Able to perform various feature engineering tasks, apply Clustering Algorithm on the given dataset.

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 example: 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)

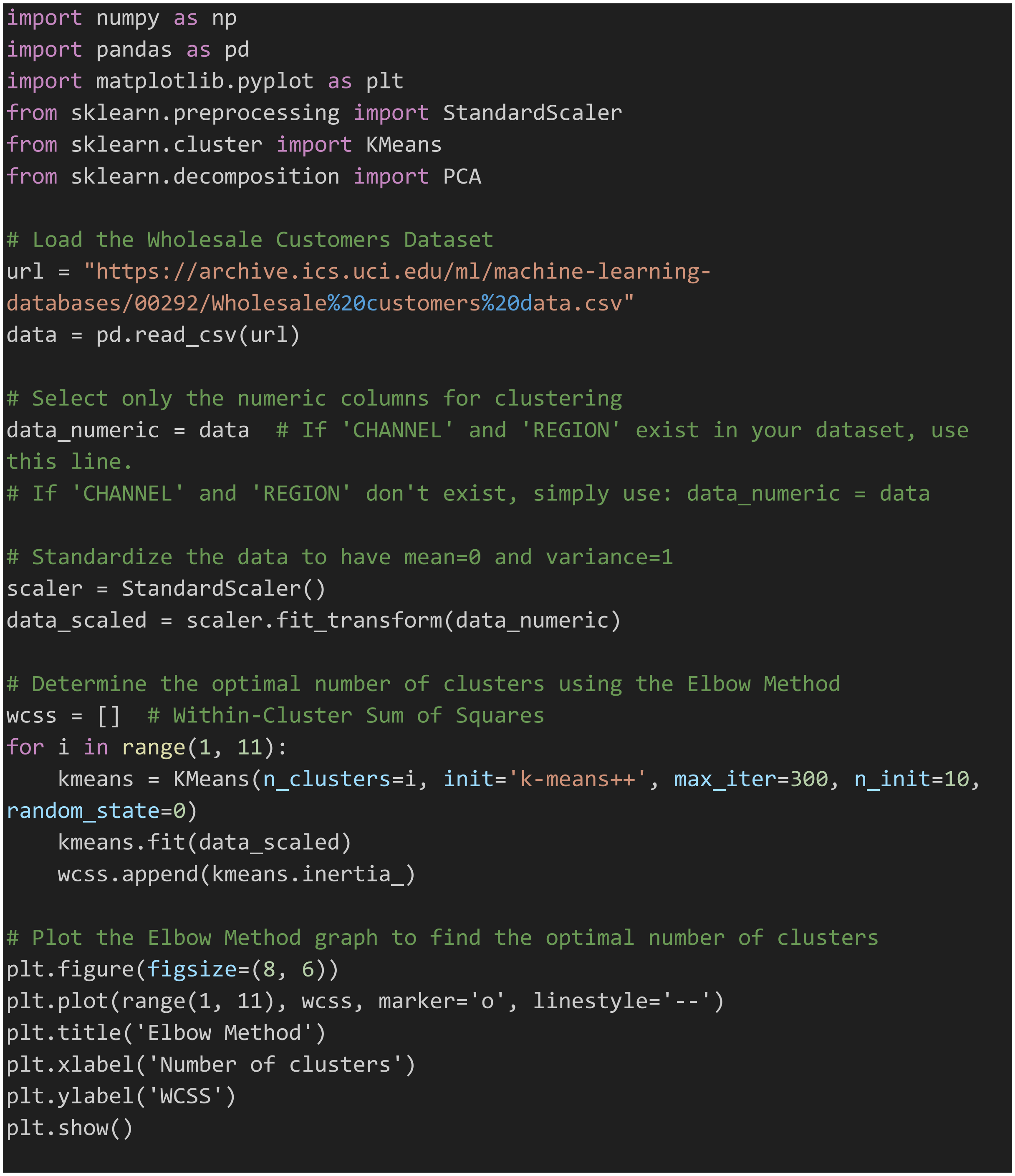
DETERGENTS_PAPER :- annual spending (m.u.) on detergents and paper products (Continuous)

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

CHANNEL: - sales channel Hotel and Retailer

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

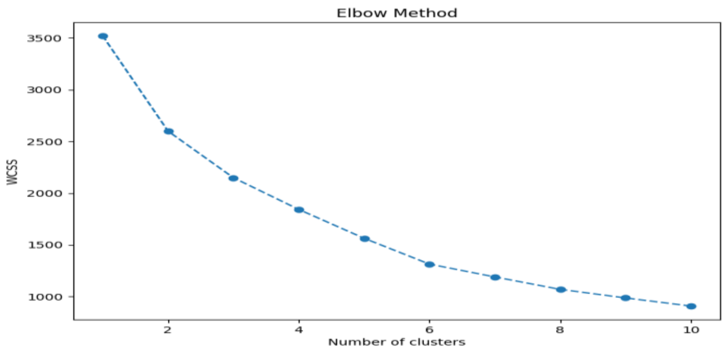
Code:

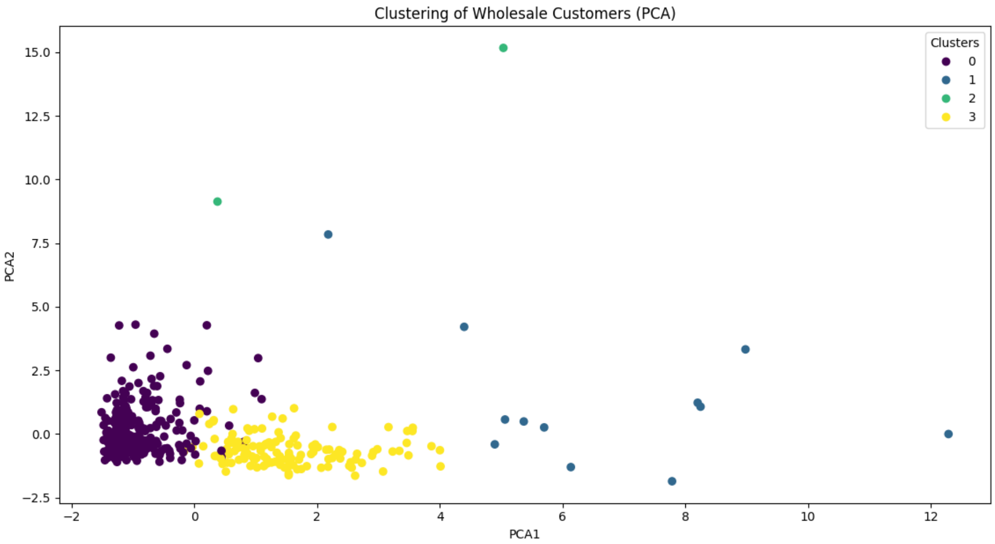


```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA

# Load the Wholesale Customers Dataset
url = "https://archive.ics.uci.edu/ml/machine-learning-
databases/00292/Wholesale%20customers%20data.csv"
data = pd.read_csv(url)

# Select only the numeric columns for clustering
data_numeric = data  # If 'CHANNEL' and 'REGION' exist in your dataset, use
this line.
# If 'CHANNEL' and 'REGION' don't exist, simply use: data_numeric = data

# Standardize the data to have mean=0 and variance=1
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data_numeric)

# Determine the optimal number of clusters using the Elbow Method
wcss = []  # Within-Cluster Sum of Squares
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10,
random_state=0)
    kmeans.fit(data_scaled)
    wcss.append(kmeans.inertia_)

# Plot the Elbow Method graph to find the optimal number of clusters
plt.figure(figsize=(8, 6))
plt.plot(range(1, 11), wcss, marker='o', linestyle='--')
plt.title('Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```

```
# Based on the Elbow Method, choose the optimal number of clusters (e.g., 4)
optimal_clusters = 4

# Apply K-Means clustering with the optimal number of clusters
kmeans = KMeans(n_clusters=optimal_clusters, init='k-means++', max_iter=300,
n_init=10, random_state=0)
cluster_labels = kmeans.fit_predict(data_scaled)

# Add the cluster labels to the original dataset
data['Cluster'] = cluster_labels

# Reduce dimensionality for visualization (you can skip this step if you
prefer)
pca = PCA(n_components=2)
data_pca = pca.fit_transform(data_scaled)
data['PCA1'] = data_pca[:, 0]
data['PCA2'] = data_pca[:, 1]

# Visualize the clustered data using PCA
plt.figure(figsize=(10, 8))
scatter = plt.scatter(data['PCA1'], data['PCA2'], c=data['Cluster'],
cmap='viridis')
plt.title('Clustering of Wholesale Customers (PCA)')
plt.xlabel('PCA1')
plt.ylabel('PCA2')
plt.legend(*scatter.legend_elements(), title='Clusters')
plt.show()

# You can now analyze the clusters and draw conclusions based on the results.
```

Output:





Conclusion:

In conclusion, applying unsupervised learning techniques, such as clustering, to the Wholesale Customers Dataset provides valuable insights for the wholesale distributor's business. It enables customer segmentation, demand forecasting, service customization, targeted marketing, and informed expansion planning, all of which can lead to more effective and efficient business operations.