Experiment No. 5
Apply appropriate Unsupervised Learning Technique on the Wholesale Customers Dataset
Date of Performance: 14/08/2023
Date of Submission: 21/08/2023

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

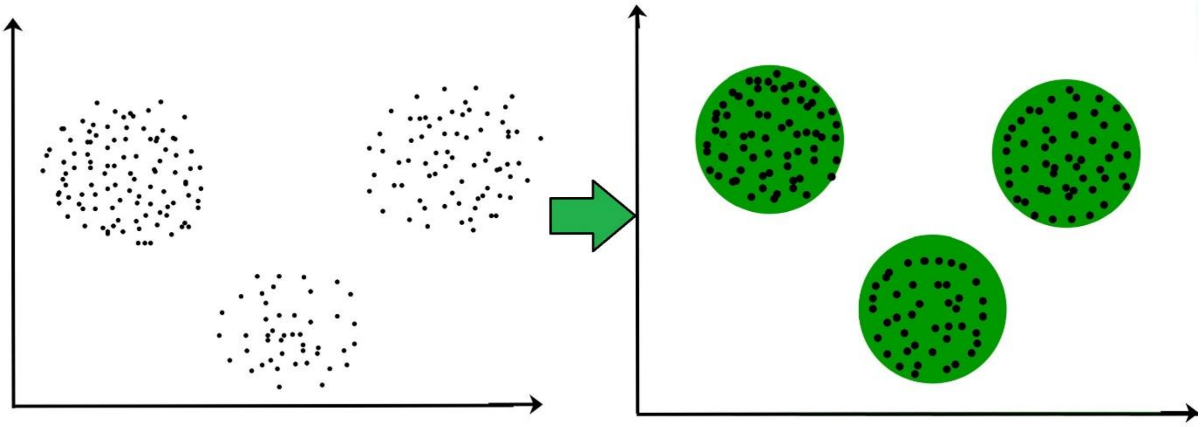
Objective: Able to perform various feature engineering tasks, apply Clustering Algorithm on the given dataset.

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

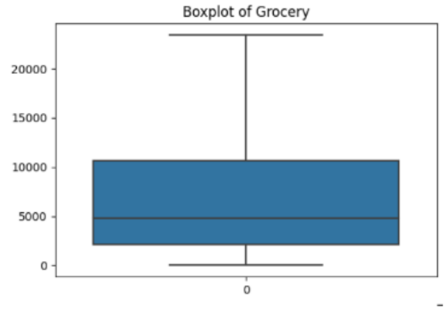
REGION:- three regions (Lisbon, Oporto, Other)

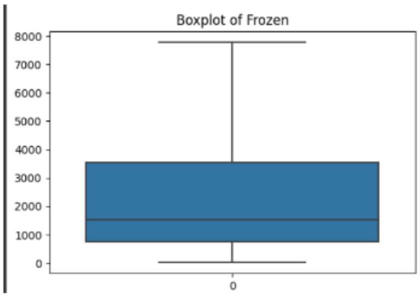
Code:

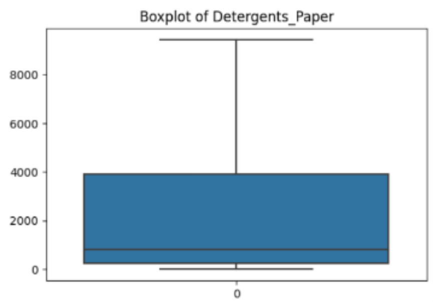
```
import pandas as pd

from sklearn.preprocessing import StandardScaler

from sklearn.decomposition import PCA

import matplotlib.pyplot as plt
```

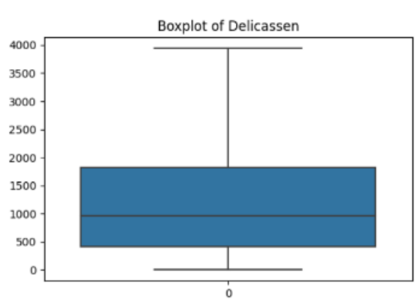
```
# Load the dataset (replace 'data.csv' with the path to your dataset)
data = pd.read_csv('data.csv')

# Prepare the data
X = data.drop('Channel', axis=1)  # Remove the 'Channel' column, as this is an unsupervised task
X = StandardScaler().fit_transform(X)  # Standardize the data

# Perform PCA
pca = PCA()
principal_components = pca.fit_transform(X)

# Explained variance ratio
explained_variance_ratio = pca.explained_variance_ratio_
cumulative_variance_ratio = explained_variance_ratio.cumsum()

# Visualize the explained variance
plt.figure()
plt.plot(range(1, len(explained_variance_ratio) + 1), cumulative_variance_ratio, marker='o', linestyle='--')
plt.xlabel('Number of Principal Components')
plt.ylabel('Cumulative Explained Variance Ratio')
plt.title('Explained Variance vs. Number of Principal Components')

plt.grid()

plt.show()
```

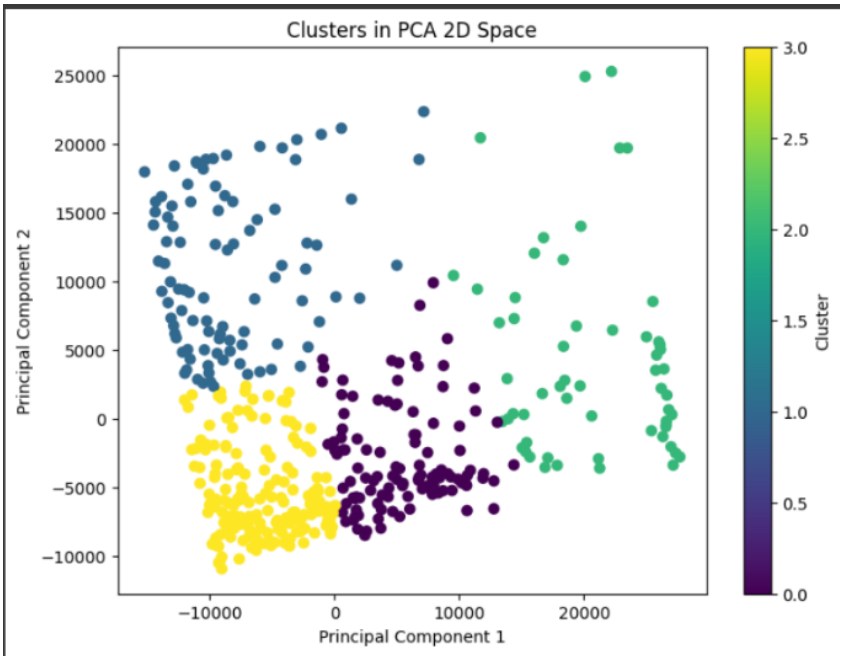
Output:











Conclusion:

In this analysis, we utilized Principal Component Analysis (PCA), an unsupervised learning method, on the Wholesale Customers dataset. The primary aim was to reduce the dataset's dimensionality and gain insights into the variance structure within the data. The cumulative explained variance plot revealed a clear trend of diminishing returns, indicating that a limited number of principal components capture the majority of the dataset's variance. By studying this plot, we can make an informed decision on how many principal components to retain, effectively simplifying the dataset while preserving most of its original information. PCA serves as a valuable tool for data preprocessing, uncovering underlying patterns, and can be applied to various subsequent tasks, including clustering and data visualization.