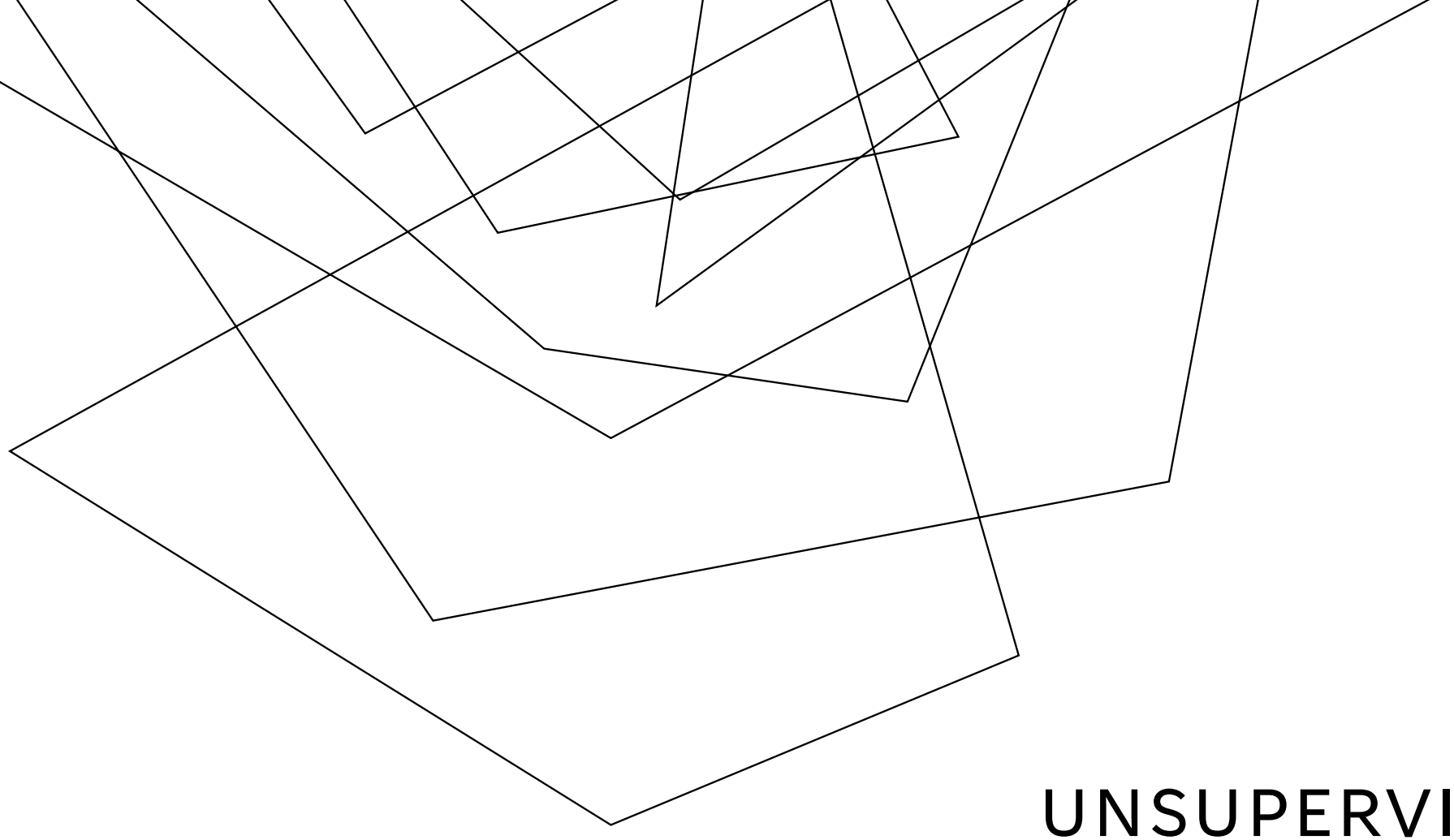


UNSUPERVISED LEARNING

Ethan Lam

AGENDA

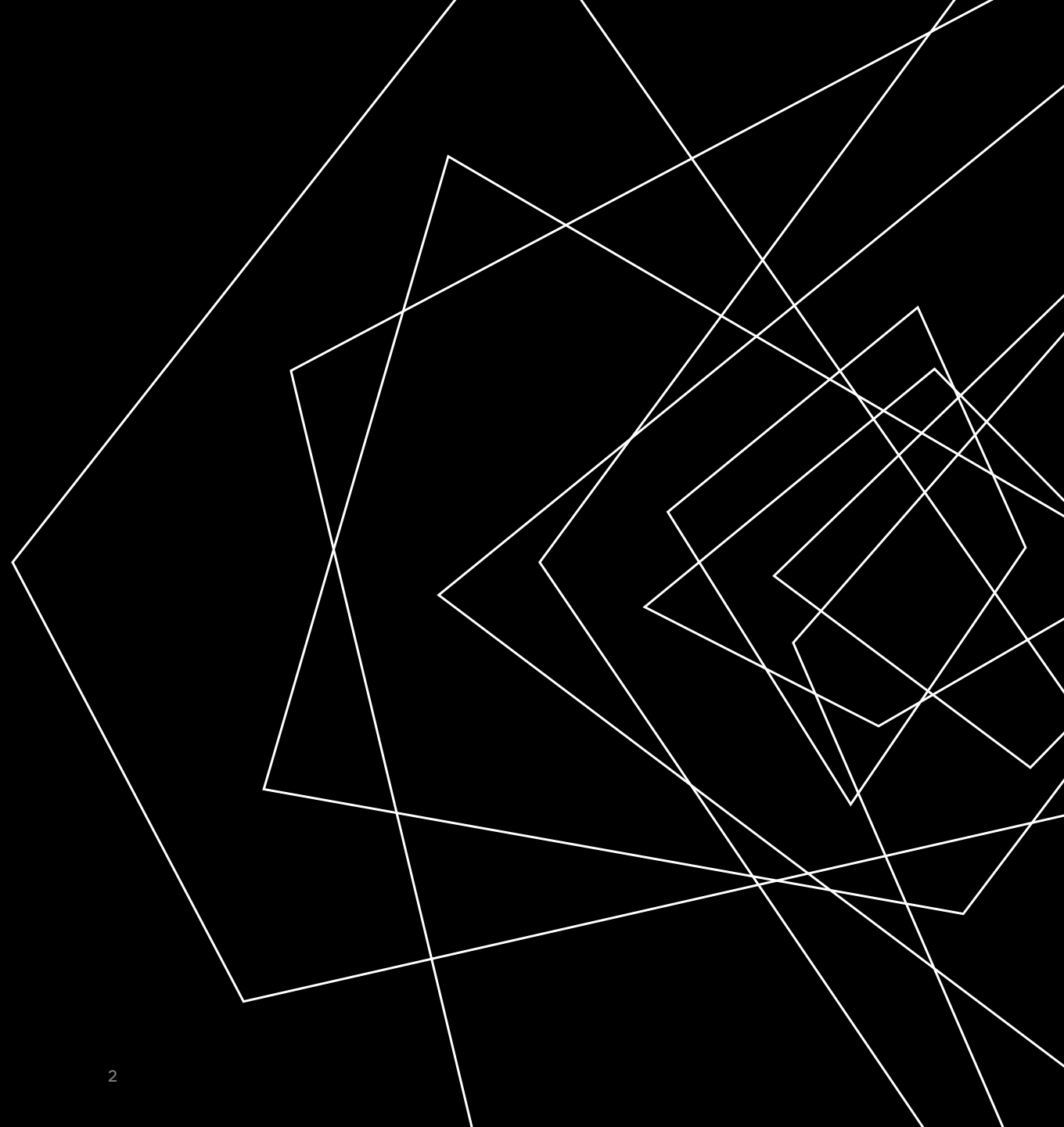
Introduction

EDA

KMeans

Hierrachy

PCA



INTRODUCTION

The dataset is “Wholesale Data” from Kaggle. It refers to a client of a wholesale distributor.

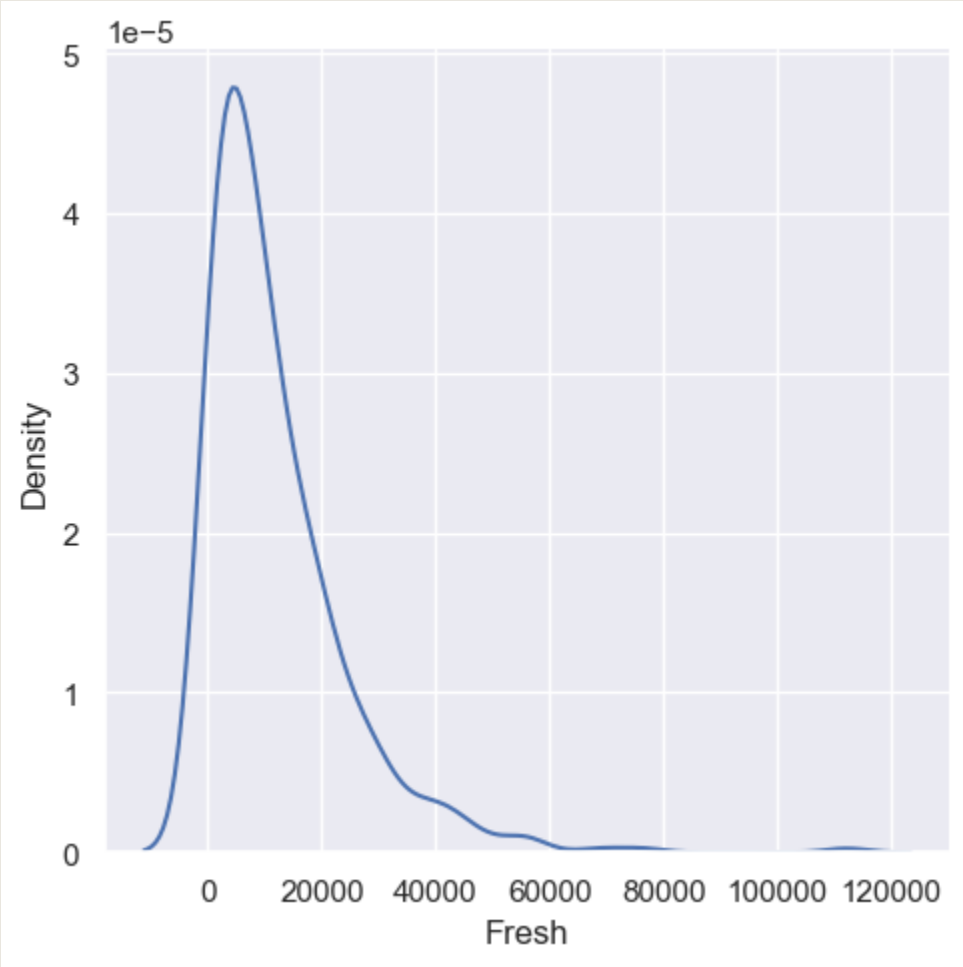
In this project, we will do the EDA and use K-means, hierarchy to identify the number of clusters.

We will also use PCA for feature reductions

EDA

During the EDA, there are some steps have been done

- Check missing values
- Check outliers
- Use data visualization to show the distribution of data
- Show the pairplot

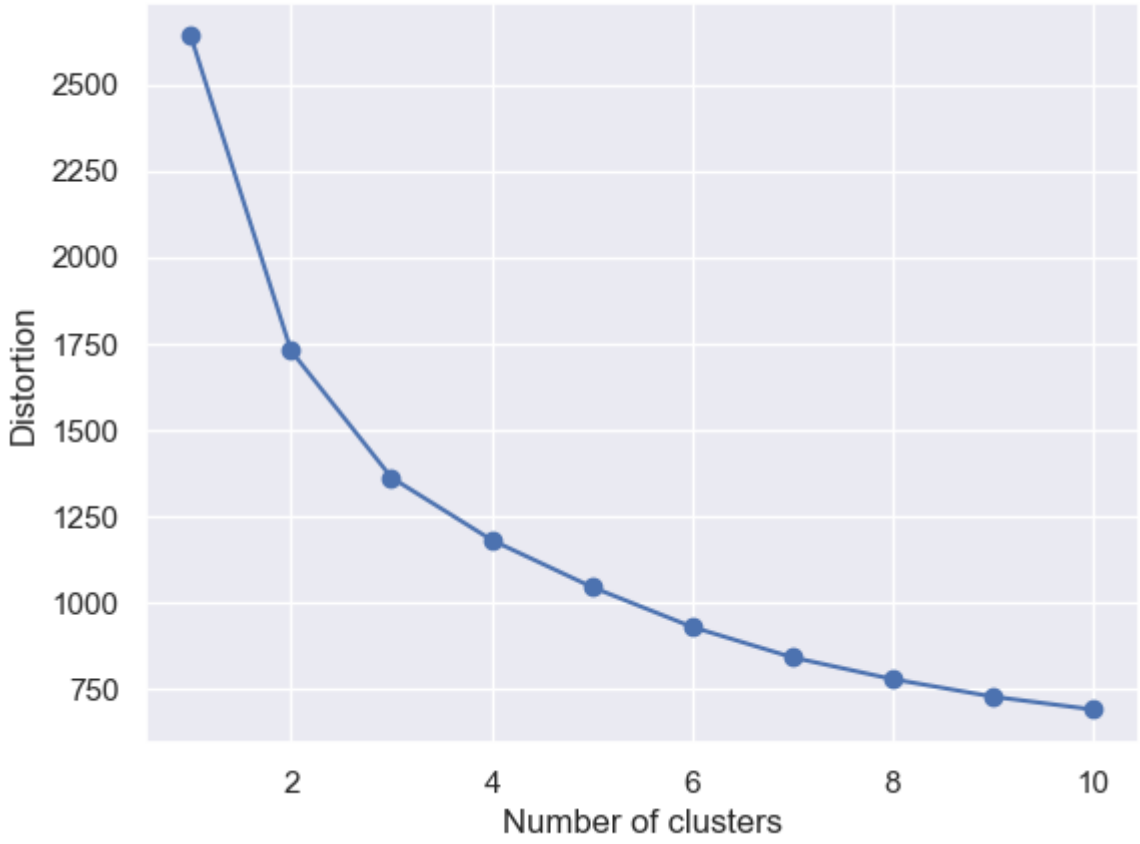


Summary

- There are 440 datapoints (observations)
- There are 8 features, including 2 categorical features (`Channel`, `Region`) and 6 numerical features
- There are no missing values.
- There are outliers in numerical features, which should be handled.

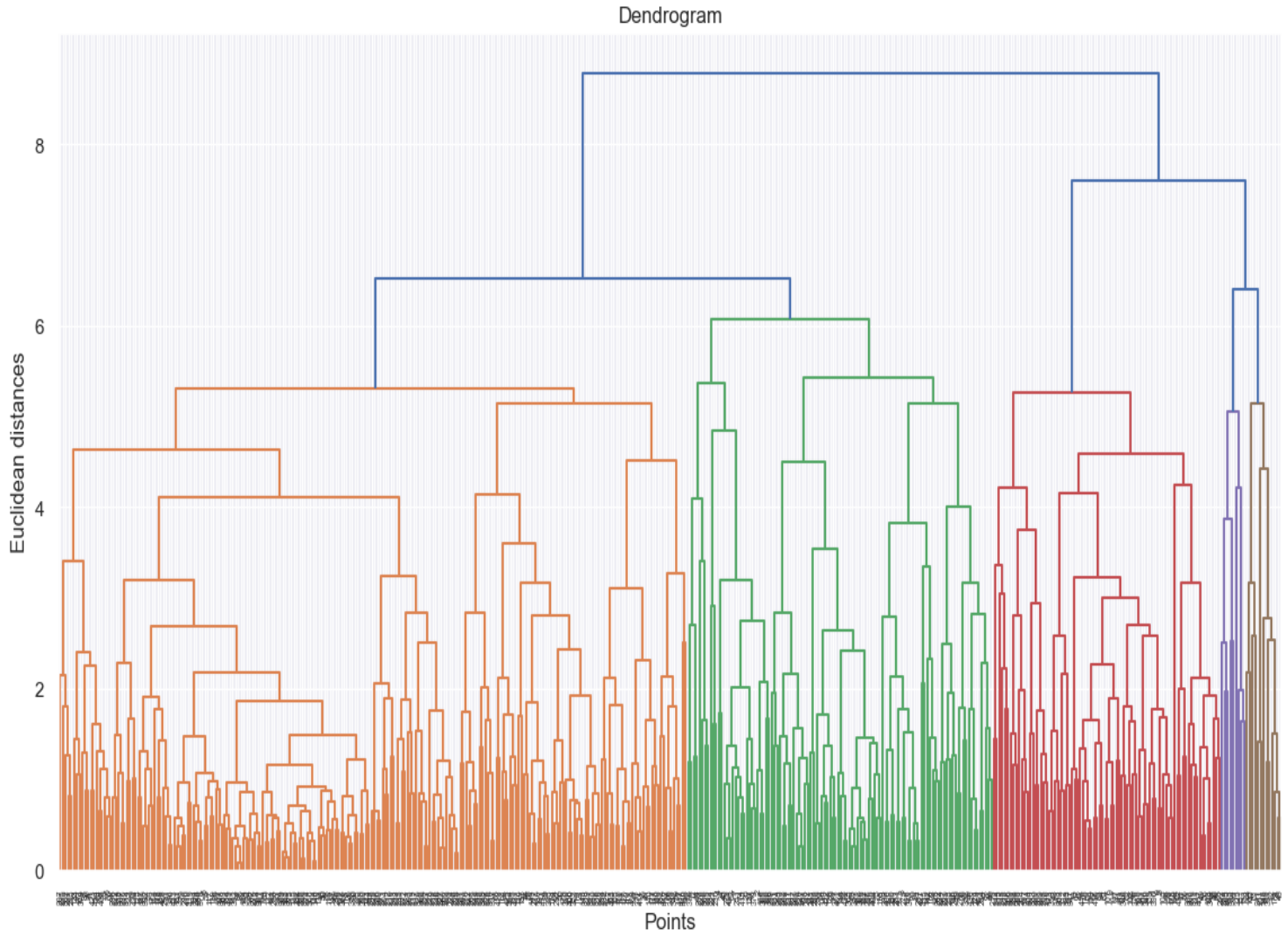
K-MEANS

- Using the Elbow method to identify the optimum number of clusters in K-Mean
- Summary: The optimum number can be from 5-6



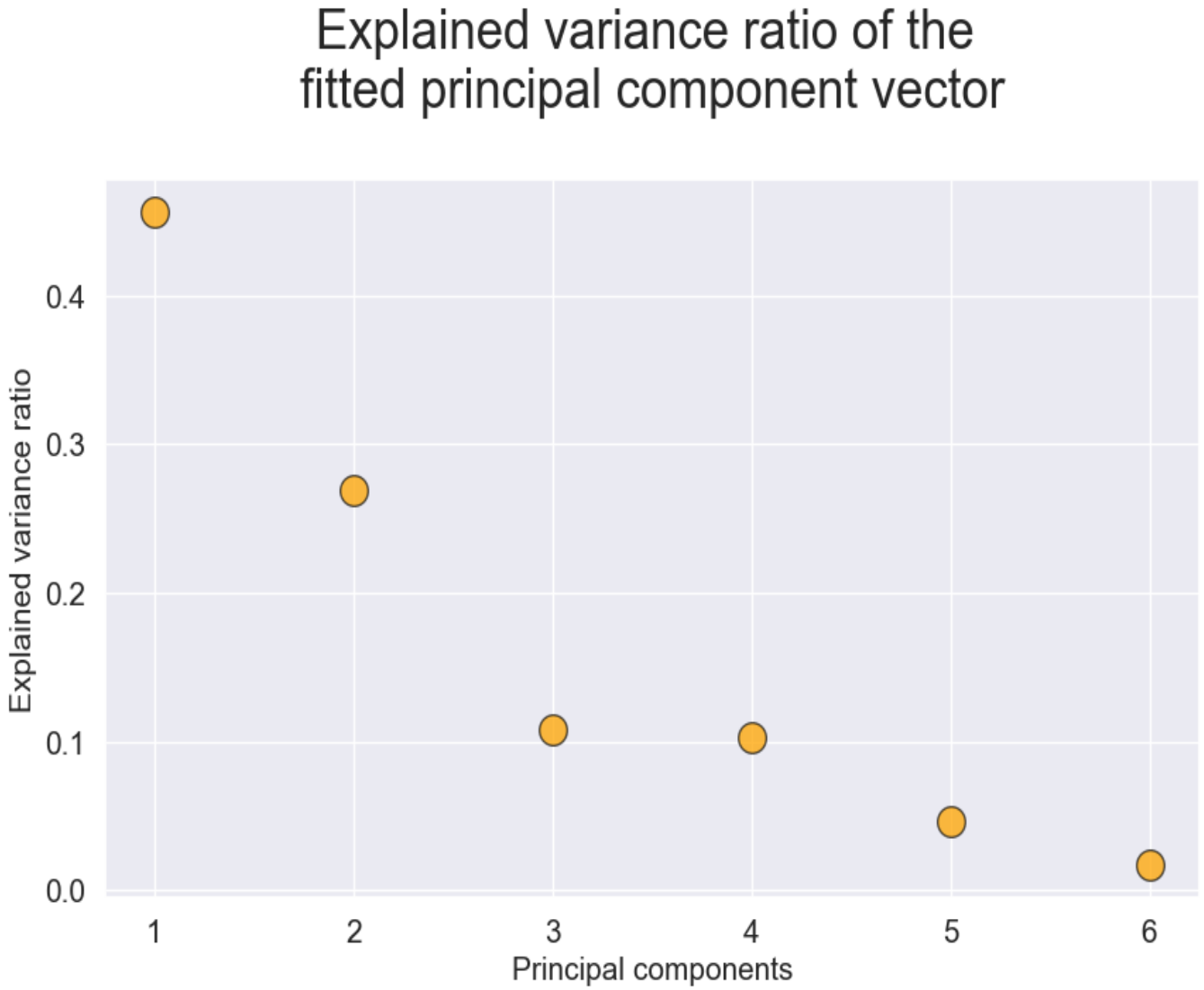
HIERRACHY

- Using the Dendogram to check the number of clustering
- Summary: The optimum number is 5



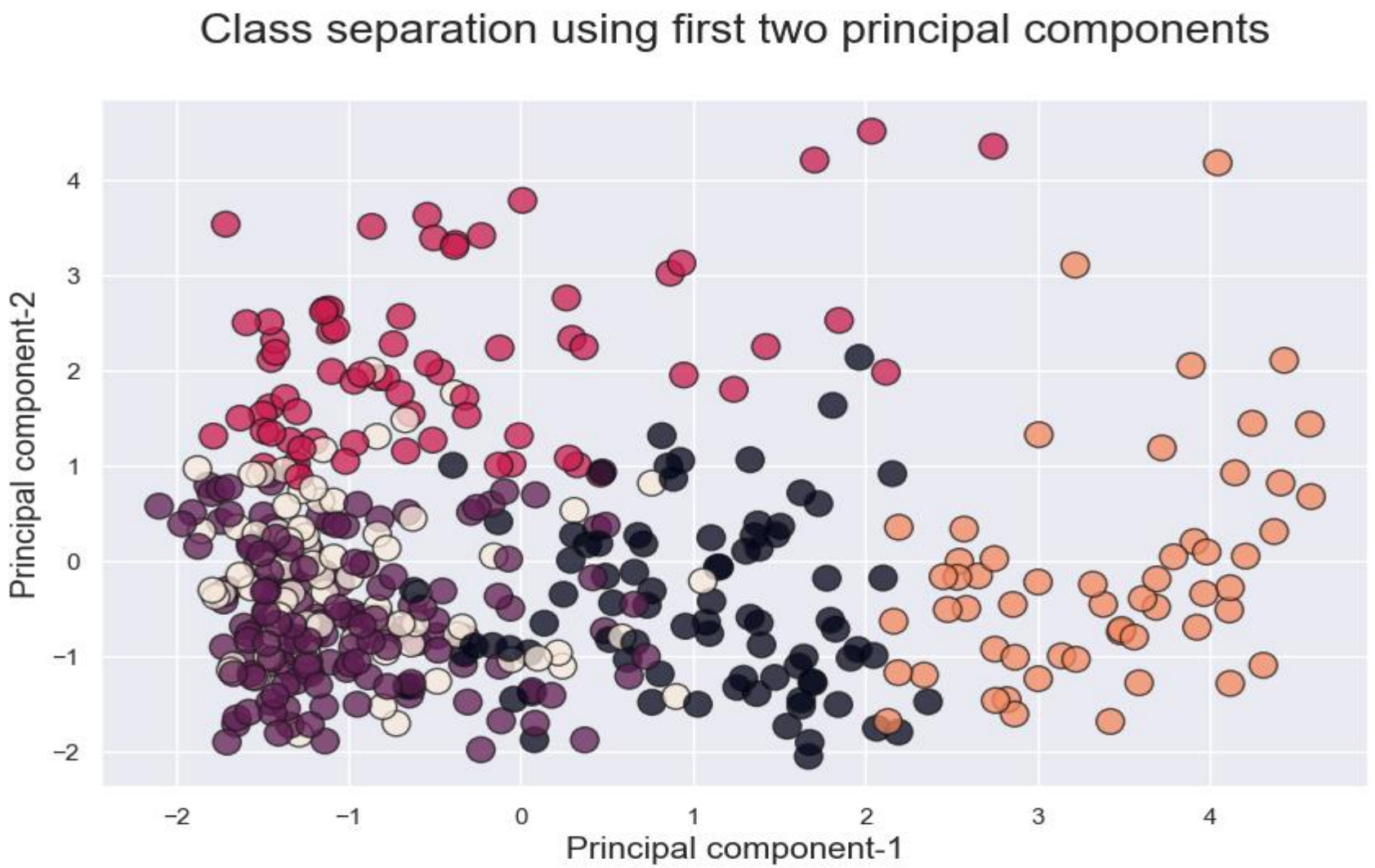
PCA

- Check the variance ratio
- Summary: The first 2 components has most of the information



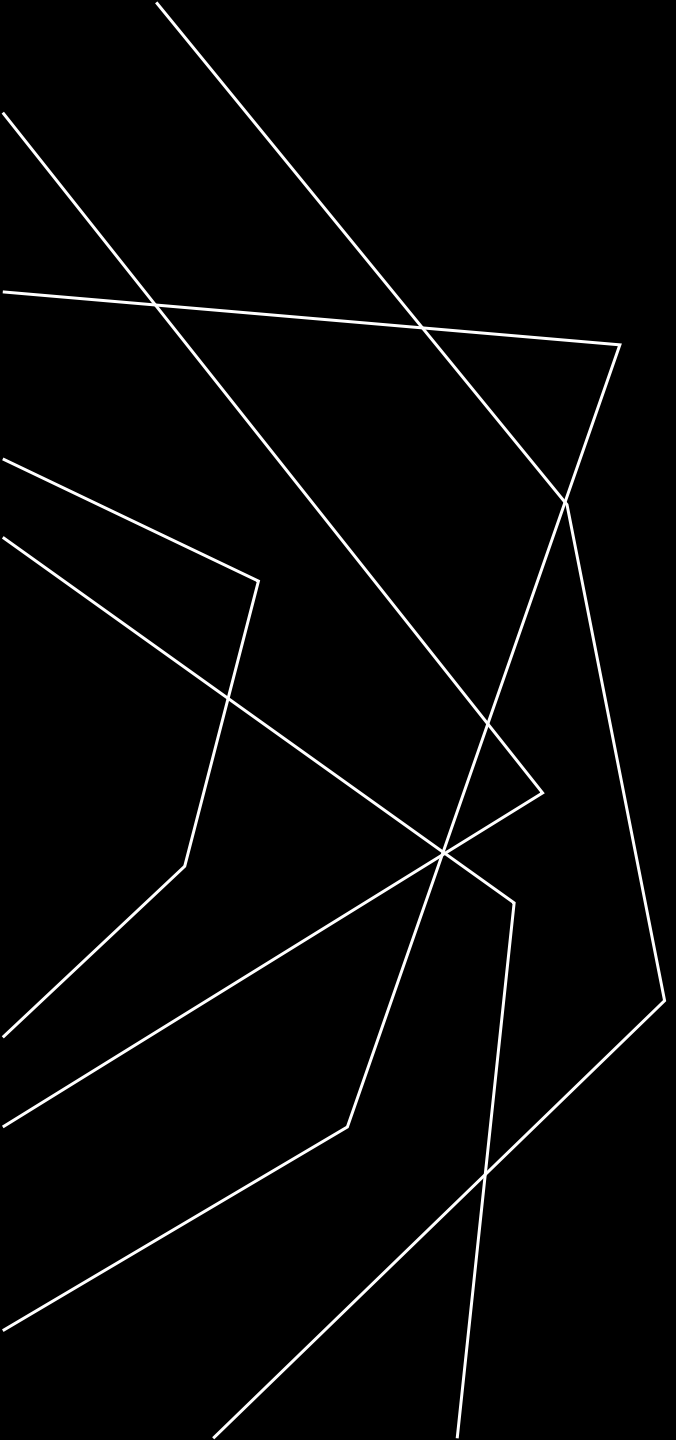
PCA

- The class separation using the first 2 components
- We can see there are 5 clusters



CONCLUSION

- There are outliers in the dataset that must be handled
- The data should be scaled for learning
- The K-means clustering and hierarchy clustering can be used together to cross-check the optimum cluster number
- PCA has 2 components, the first holds 45% and the second holds 27% of the information
- By using Clustering, we can segment the customers into 5 groups
- Each groups has different needs about product and spending which is good for company to have the insight so that they can plan or take action properly



THANK YOU