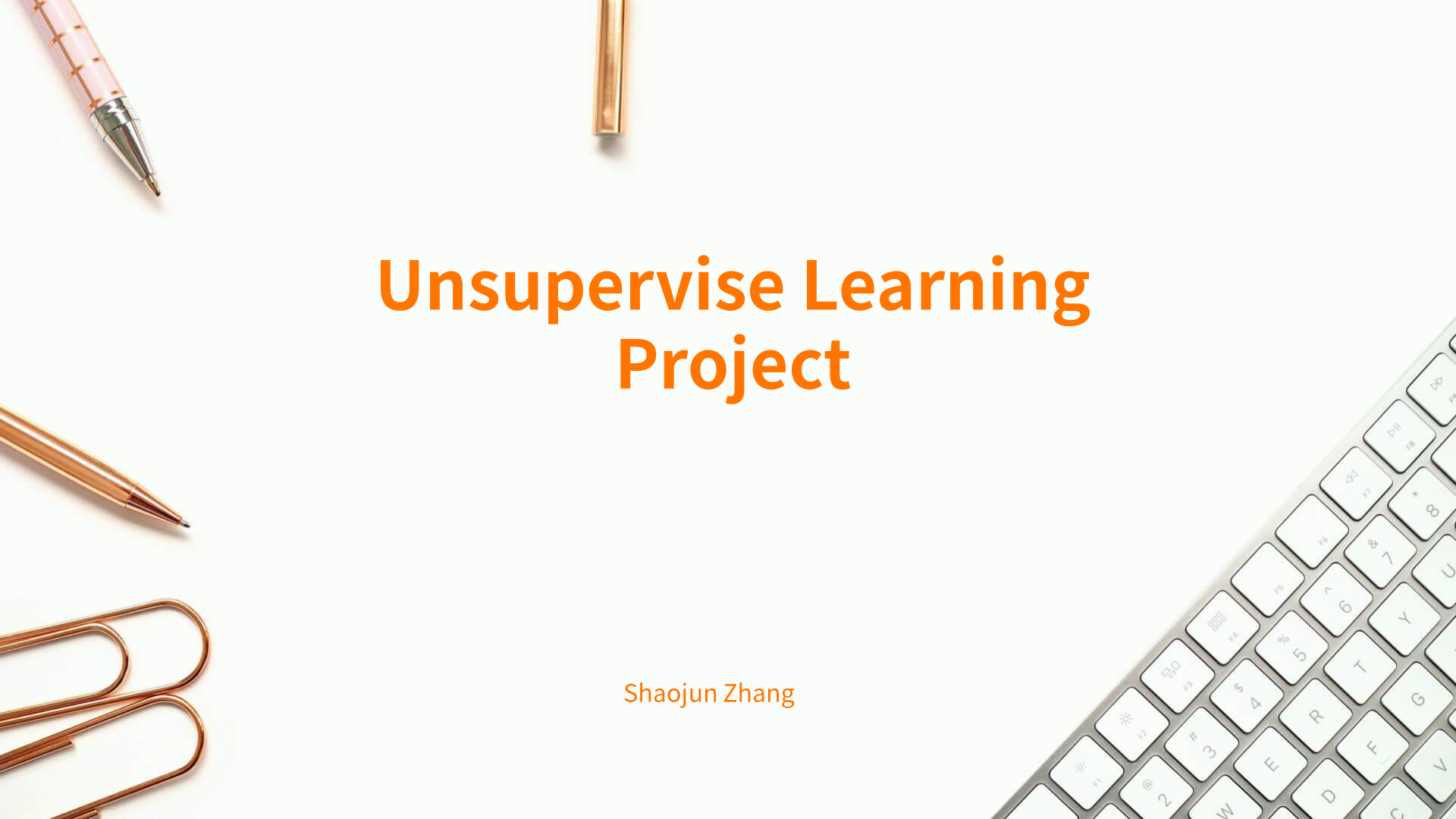
Unsupervise Learning Project

Shaojun Zhang

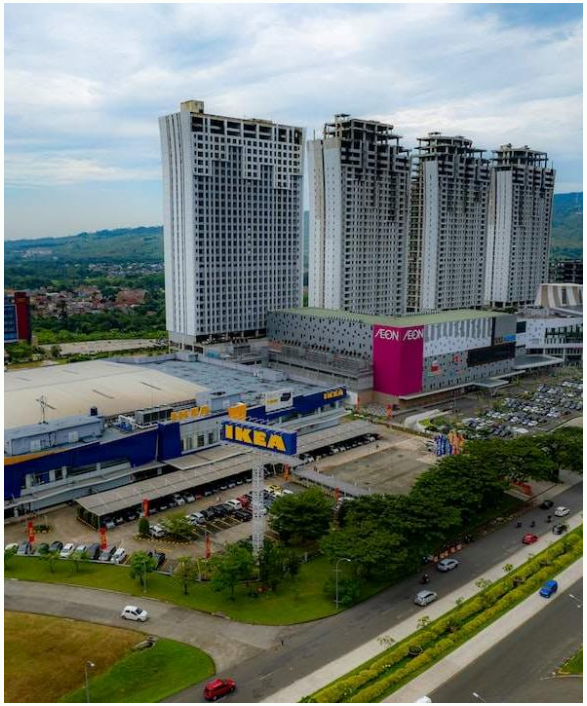


1. Section One: Data Introduction
2. Section Two: Project Objective
3. Section Three: Methodology Overview
4. Section Four: Results Analysis
5. Section Five: Challenges Encountered
6. Section Six: Conclusion

01

Section One: Data Introduction

Section One: Data Introduction



This dataset pertains to wholesalers, capturing detailed purchase behaviors across different channels and regions. The dataset primarily includes the following fields:

This dataset pertains to wholesalers, capturing detailed purchase behaviors across different channels and regions. The dataset primarily includes the following fields:

Channel:
The sales avenue. This dictates how consumers are accessing and purchasing the product, be it through traditional retail stores or online platforms.

Region:
The location of the consumers. Regional factors might influence purchasing preferences.

Fresh:
The amount purchased of fresh produce. This helps us gauge the demand for fresh food items by consumers.

Milk:
Amount of milk and dairy products purchased. Dairy products are typically everyday needs for most households.

Grocery:
Amount of daily goods purchased. This encapsulates all non-perishable goods.

Frozen:
Amount of frozen food products purchased. Frozen foods typically have a longer shelf life.

Detergents_Paper:
Amount of detergents and paper products purchased. These are items that consumers might purchase on a recurring basis.

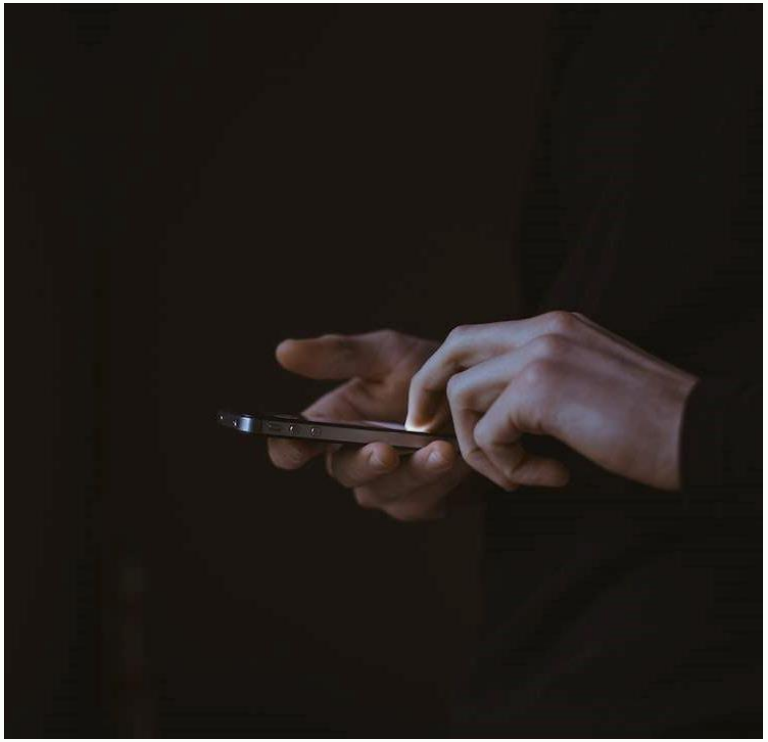
Delicassen:
Amount of delicacies or ready-to-eat foods purchased. These are foods prepared for the consumer's convenience.

02

Section Two: Project Objective

Section Two: Project Objective

Beyond classifying consumers based on their purchase records, we aim to offer wholesalers a more precise market segmentation strategy, facilitating more effective marketing and inventory management.



03

Section Three: Methodology Overview

Section Three: Methodology Overview

1. **Data Cleaning**:
Firstly, outliers and missing values within the dataset were addressed to ensure the integrity of our analysis.

2. **Clustering Analysis**:
Two different methodologies were employed for clustering, determining the optimal grouping strategy.

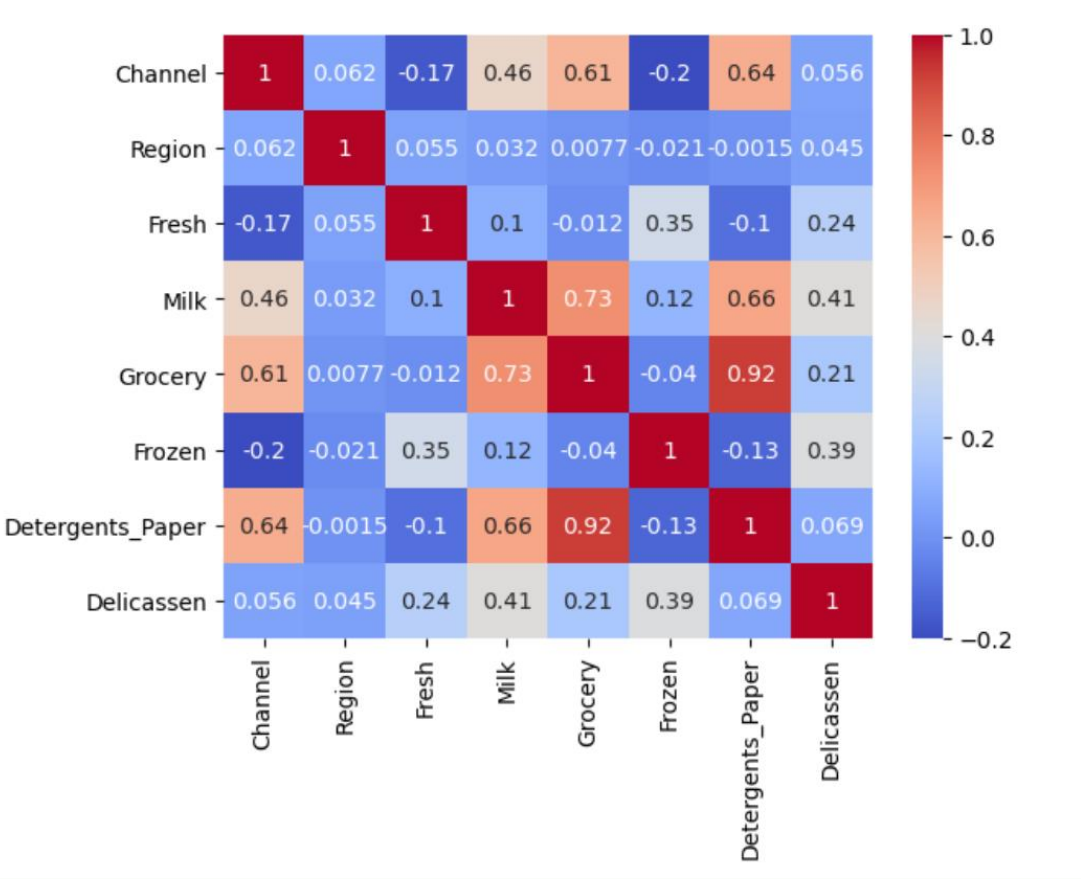
3. **Dimensionality Reduction**:
PCA enabled us to extract crucial information from multidimensional data without significant loss of feature details.

4. **Model Validation**:
Predictions from the model were compared against actual data to ascertain predictive performance.

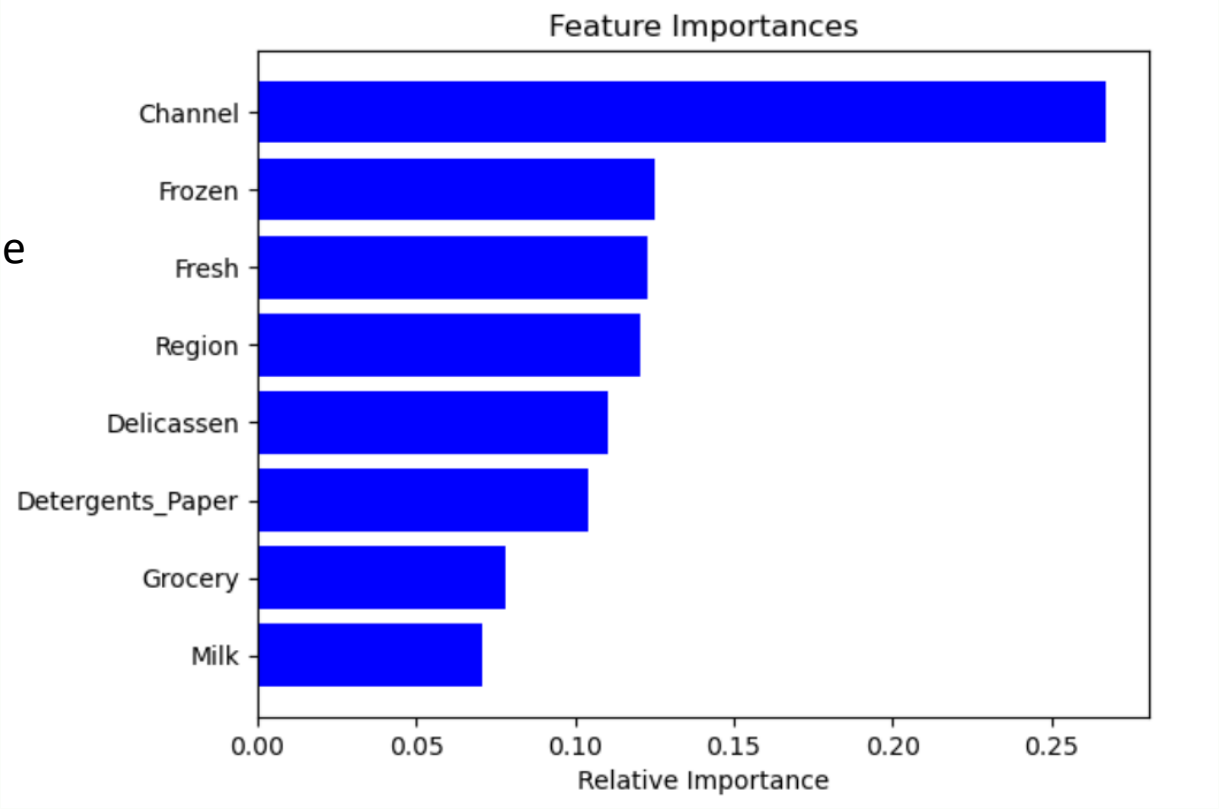
04

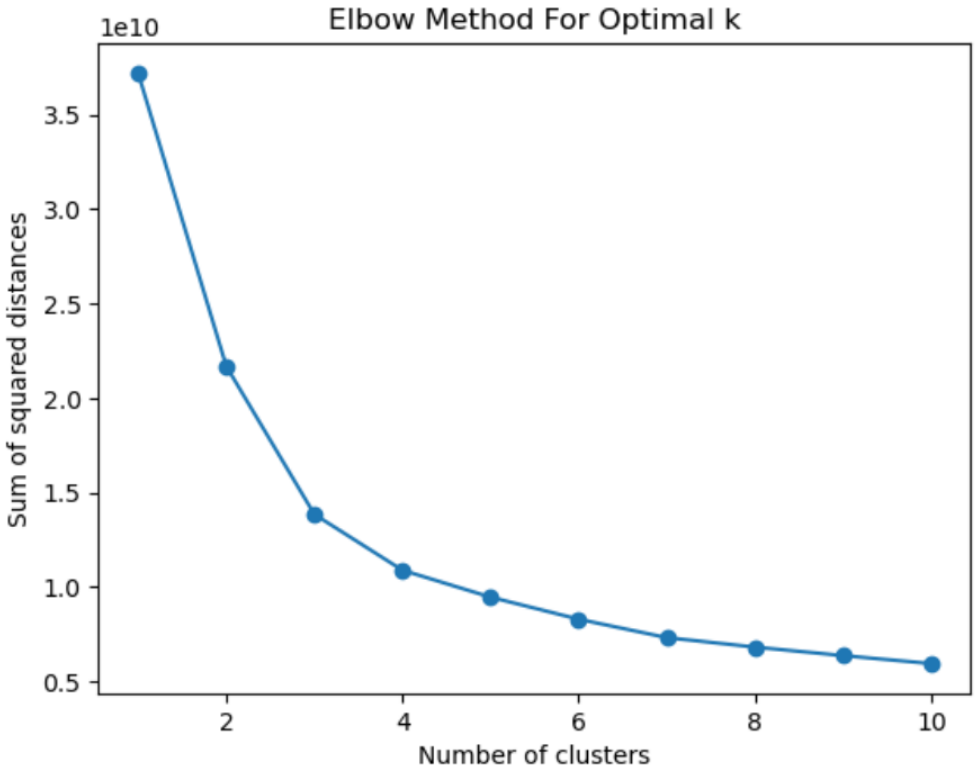
Section Four: Results Analysis

From the heatmap, we can identify some features that have a high correlation, such as Milk and Grocery, Paper and Channel, Grocery and Channel, Paper and Grocery, and Paper and Milk.



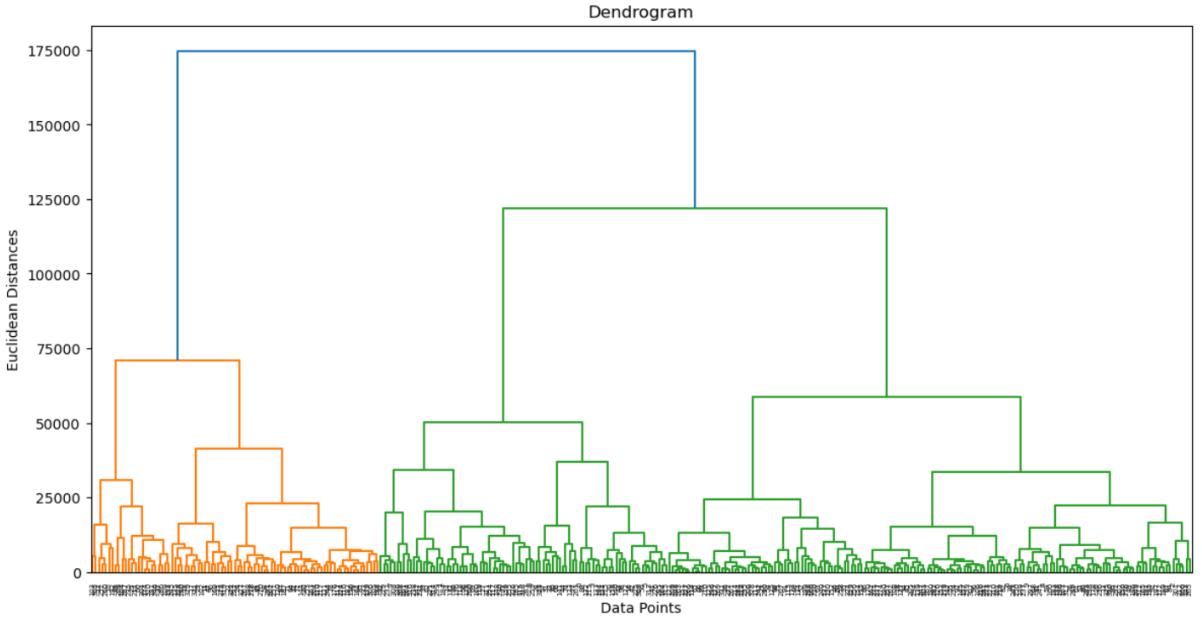
Using Random Forest, the ranking of the most important features has been listed. Apart from Channel, the two most important features are Frozen and Fresh.



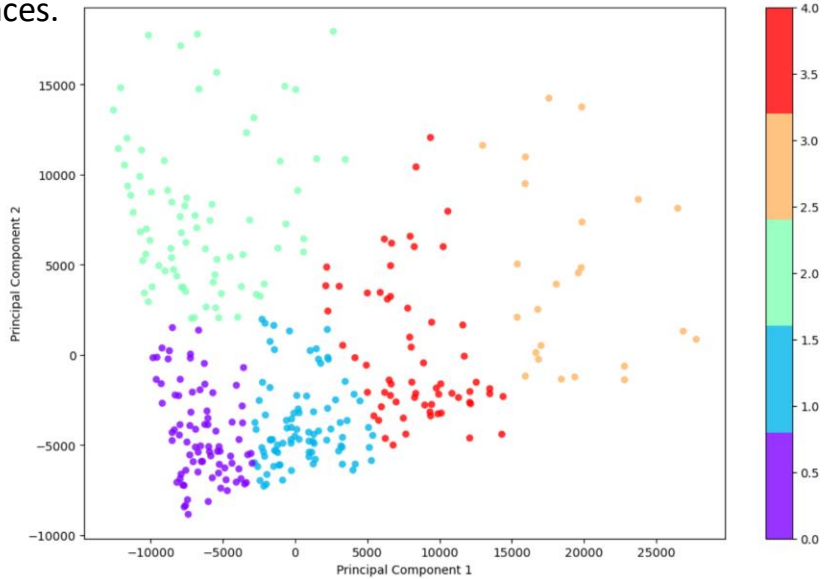


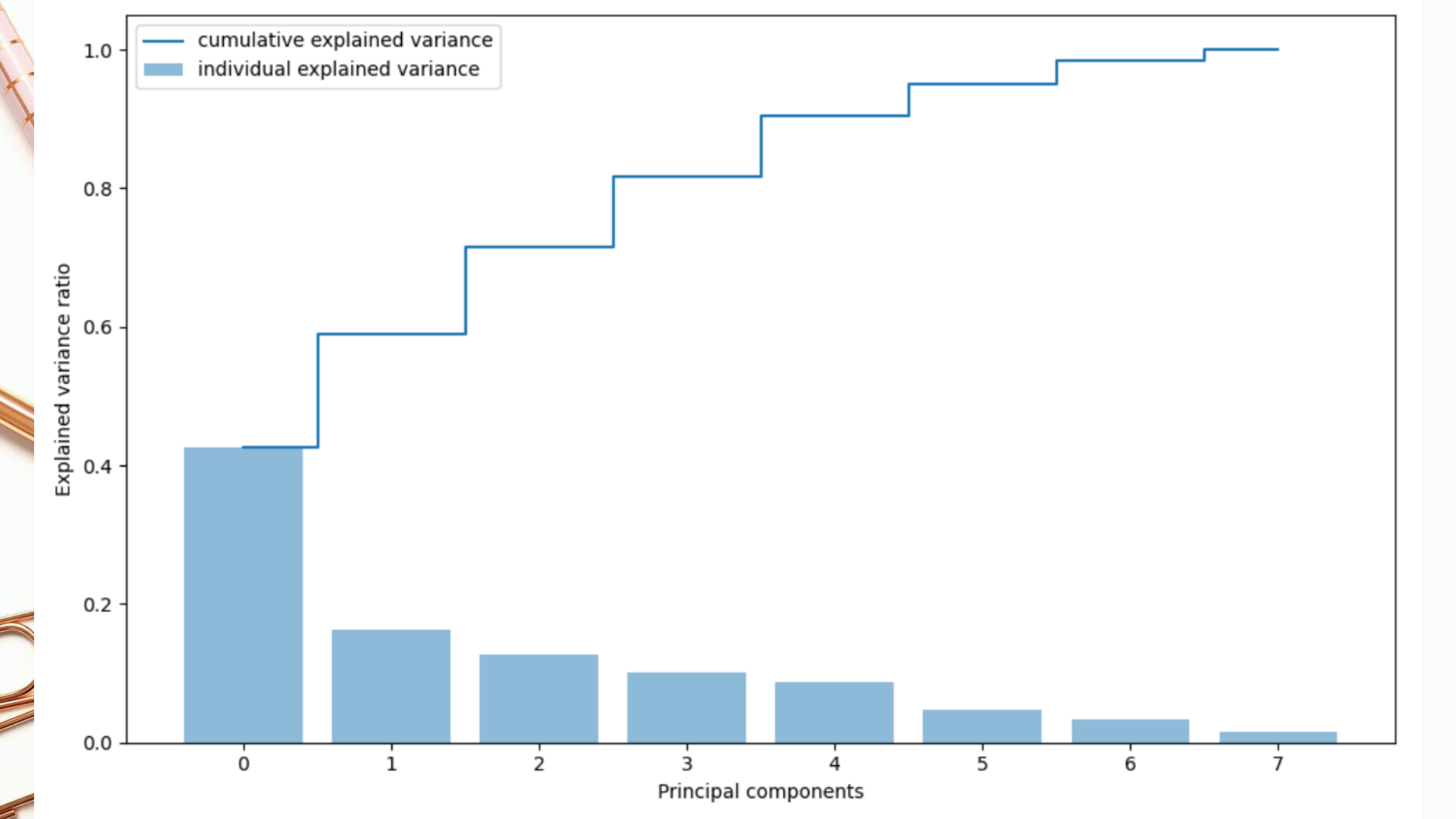
We used the Elbow method to determine the optimal number of clusters. There is minimal change in the slope after five clusters, so we believe that five clusters are optimal.

Each branch represents a cluster, and its length signifies the distance or difference between two clusters. From this graph, we can see that 5-6 is a good number of clusters.



Using five clusters, we created a PCA scatter plot. We can clearly see distinct differences.




cumulative explained variance
individual explained variance
Explained variance ratio
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
Principal components

05

Section Five: Challenges Encountered

Section Five: Challenges Encountered

01

1. **Data Scale**:
Choosing the right computational tools and techniques was challenging when dealing with large datasets.

02

2. **Feature Selection & Transformation**:
Deciding on which features to use for clustering and PCA, and how to transform these features was pivotal.

03

3. **Interpreting Clustering Outcomes**:
Different clustering algorithms can yield varied results; deciding which is more interpretable was a task.

06

Section Six: Conclusion

Section Six: Conclusion

This project illuminated insights into consumer behaviors through unsupervised learning. We hope that these findings bring tangible business value to wholesalers. Thank you for your participation, and we look forward to a deeper discussion with all of you.

THE END

THANKS

版权声明：

PPT中所用字体下载："Noto Sans SC", "Noto Sans KR"

https://www.aliyundrive.com/s/Q1VAkGau9PE