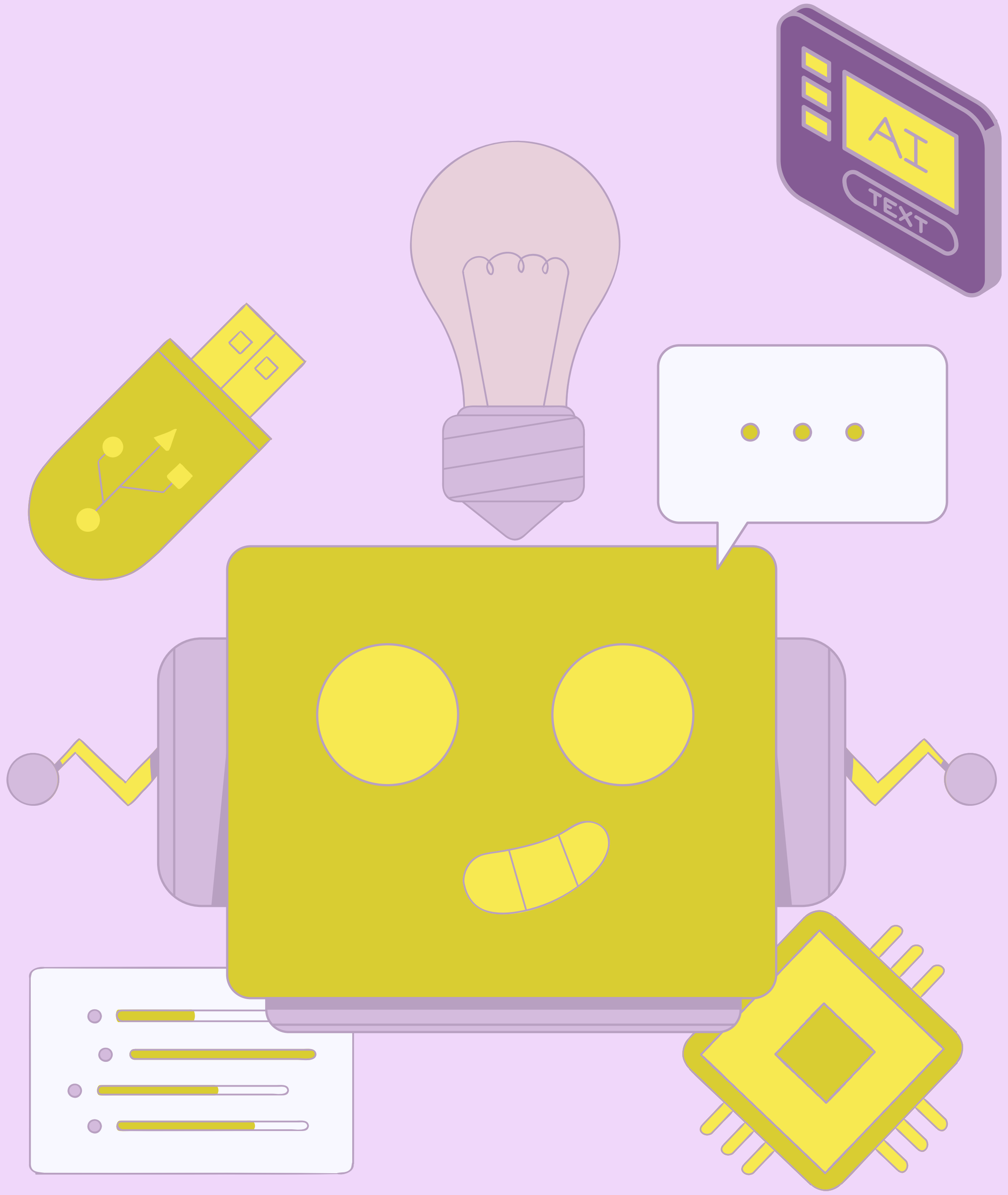
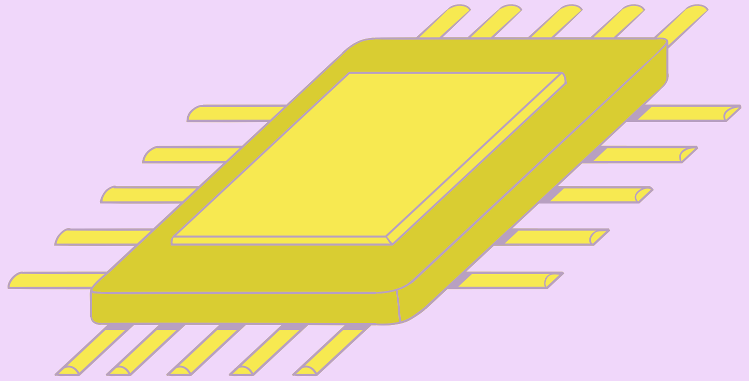
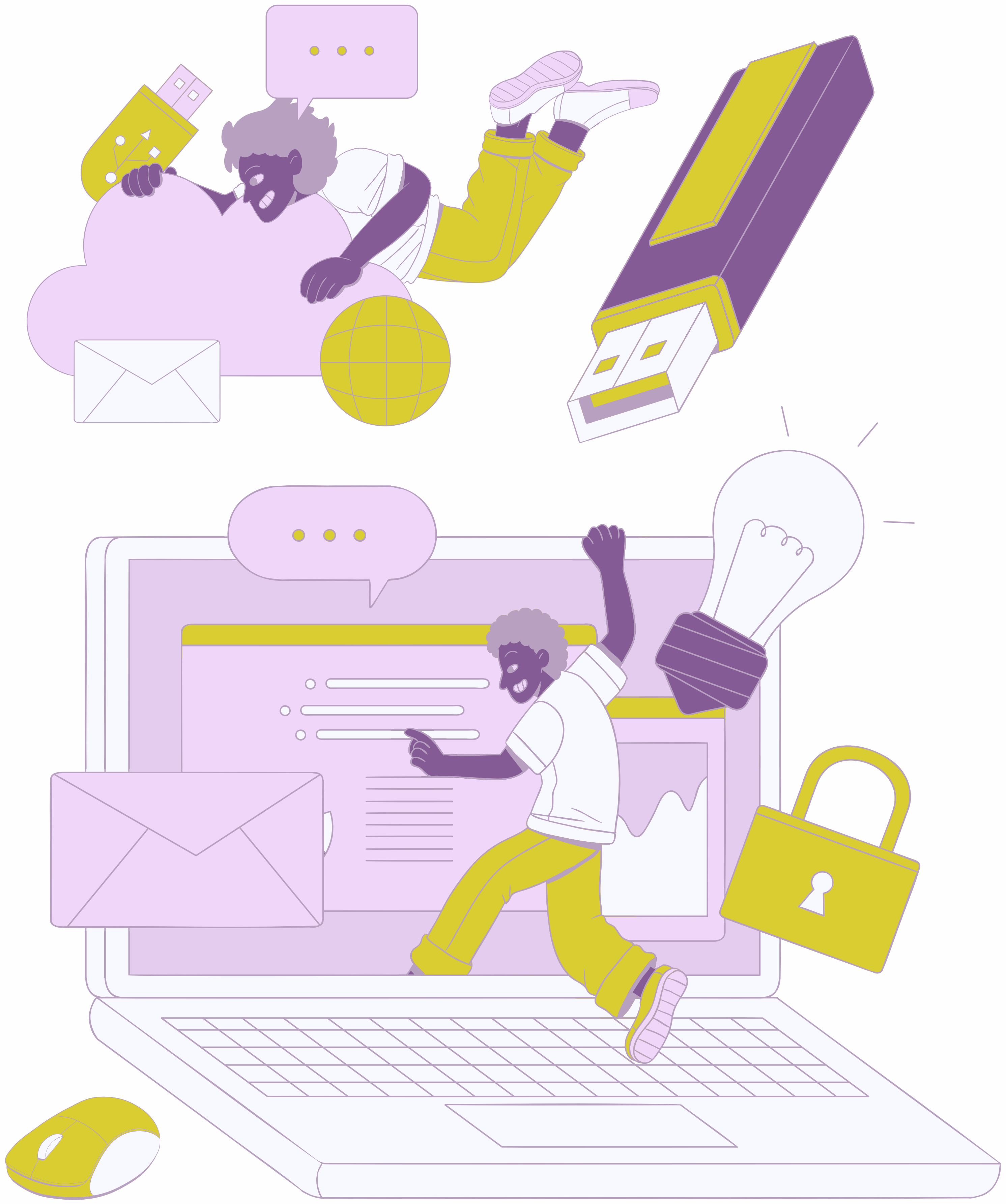


CUSTOMER SEGMENTATION WITH MACHINE LEARNING

Patricia Díez

April 2025





INDEX

- Introduction
- Dataset Description
- Exploratory Data Analysis (EDA)
- Data Preprocessing
- Applied Clustering Techniques
- Comparison of Results
- Final Model Selection
- Conclusions
- Results
- Potential Project Improvements

INTRODUCTION

OBJECTIVE

Segmenting wholesale customers based on their purchasing behavior, with the help of Machine Learning.

WHAT CAN WE USE IT FOR?

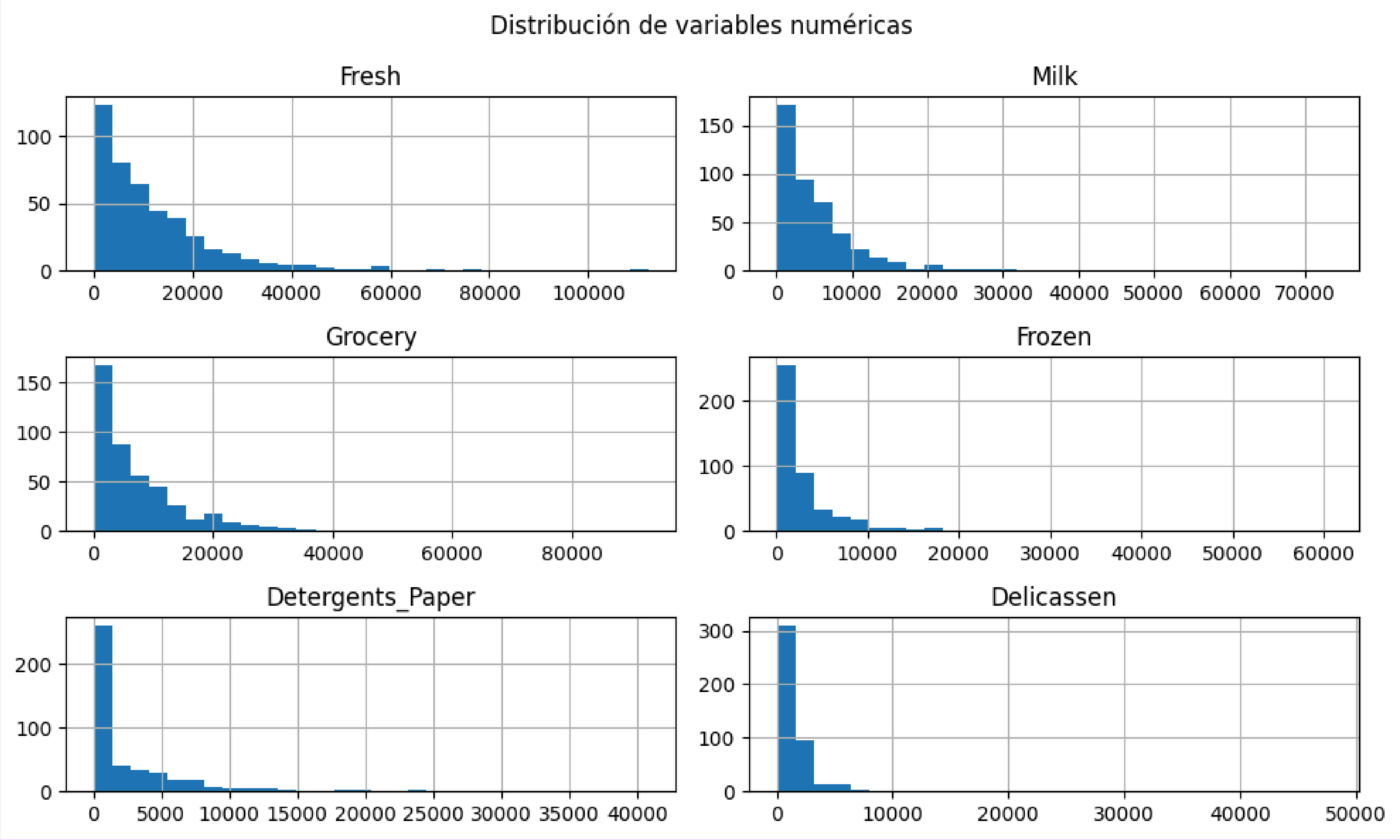
This segmentation enables the design of specific marketing strategies for each customer group.

DATASET DESCRIPTION

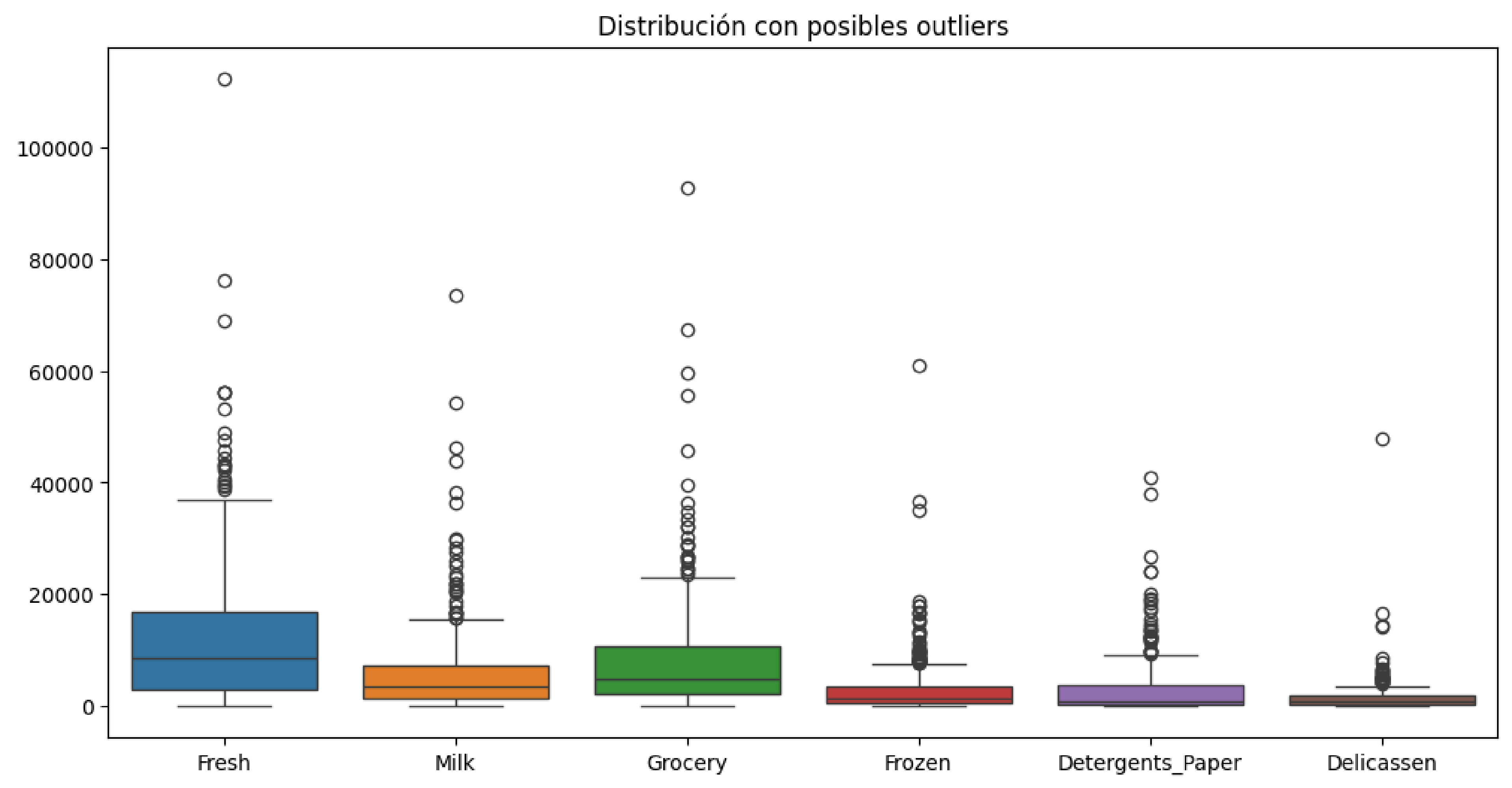
WHOLESALE CUSTOMERS

- Public dataset from Kaggle
- Contains spending information across 8 categories for 440 wholesale customers.
- Collected variables:
  - Numerical variables: Fresh, Milk, Grocery, Frozen, Detergents_Paper, Delicassen
  - Categorical variables: Region and Channel

PREPROCESSING AND EDA

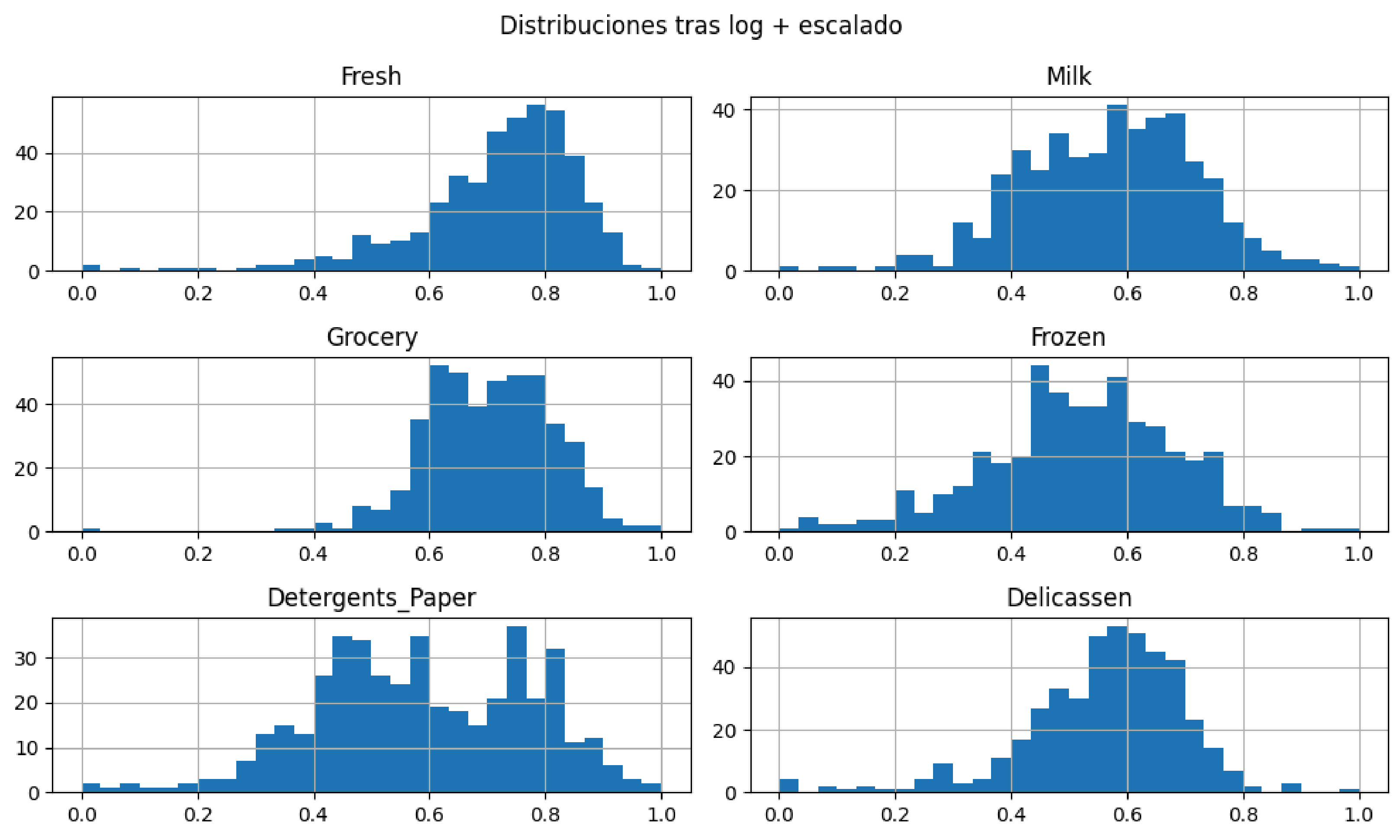


RIGHT SKEWNESS



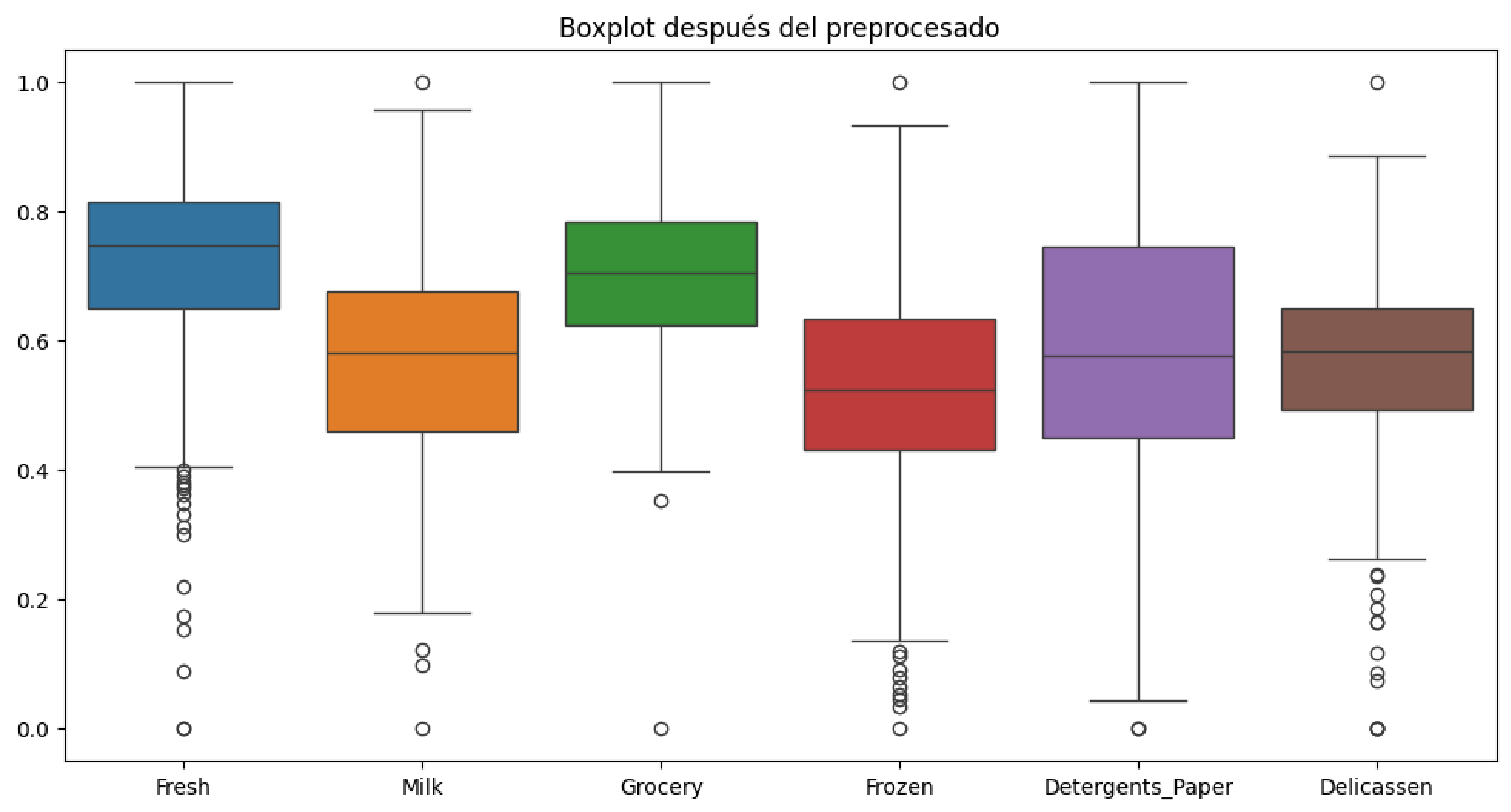
HIGH NUMBER OF OUTLIERS

PREPROCESSING AND EDA

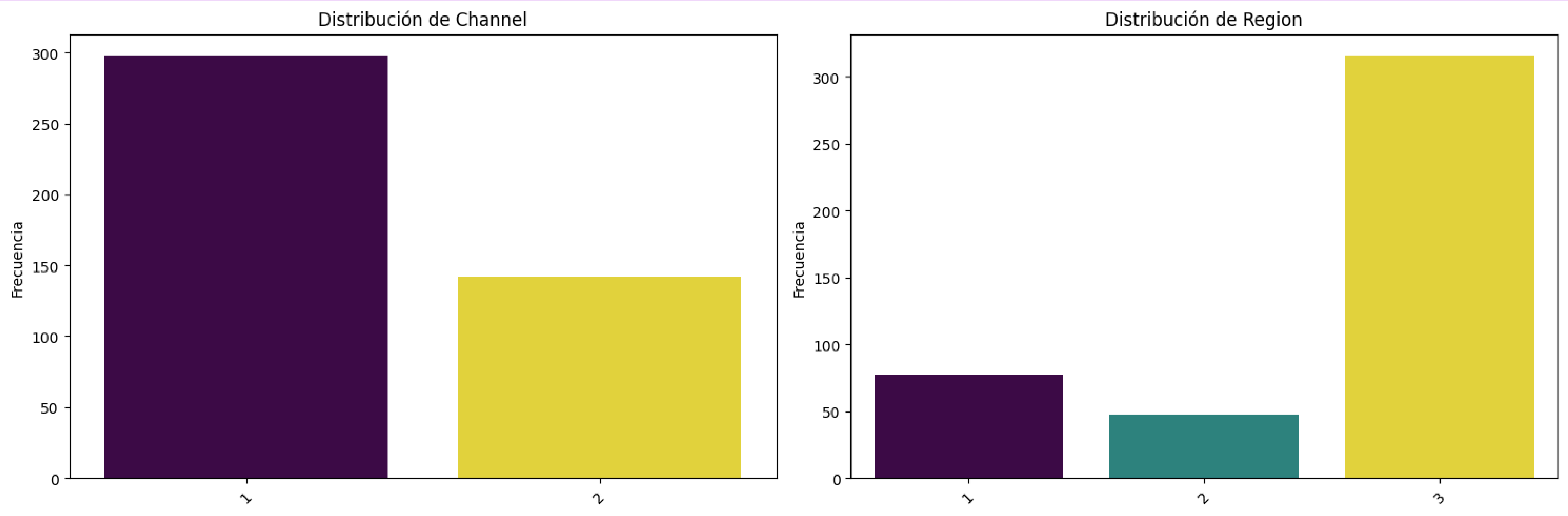


LOGARITHM

FEATURE SCALING



PREPROCESSING AND EDA



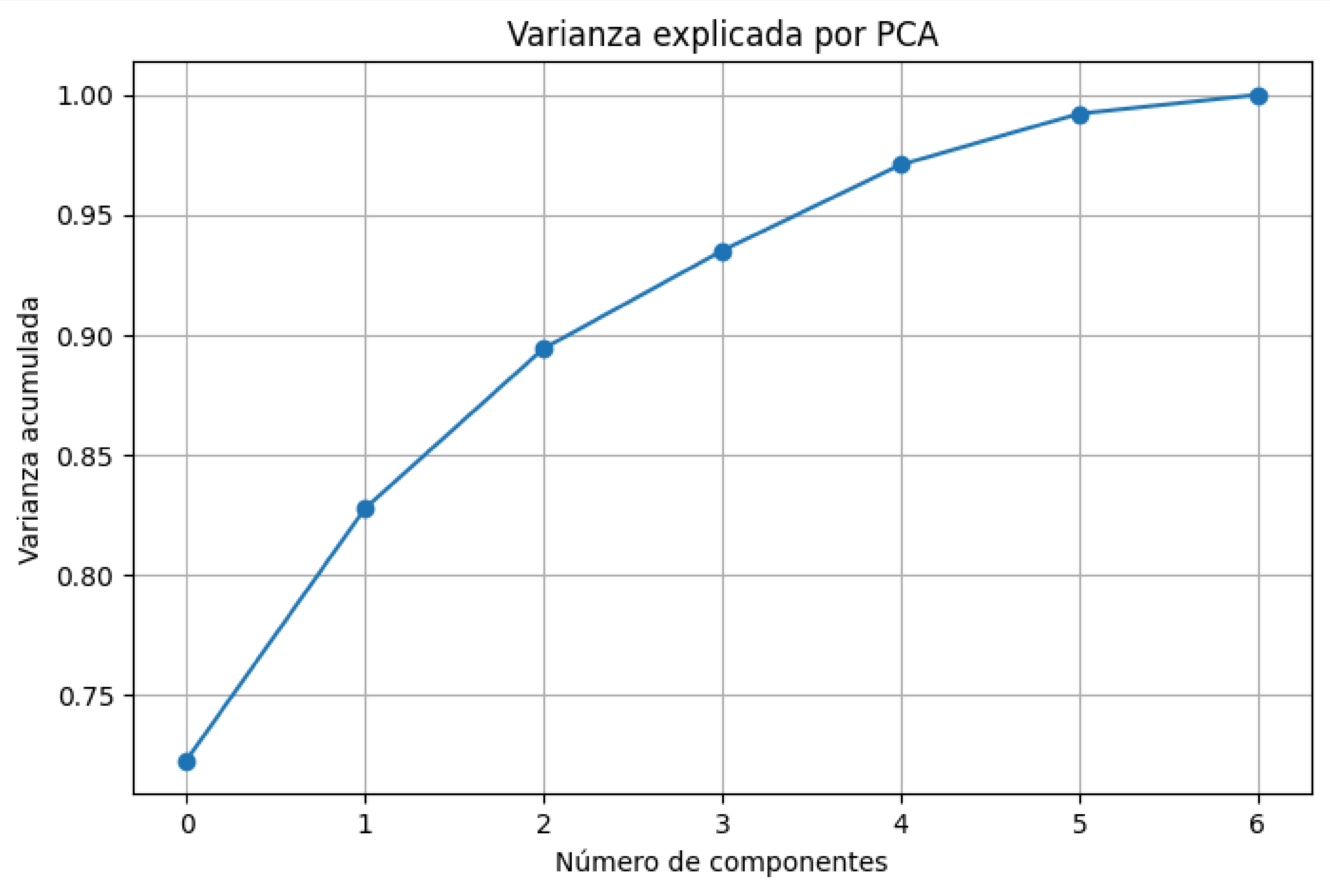
Distribución de Channel

Distribución de Region

ONE - HOT ENCODING OF CHANNEL

DROP REGION

PCA – PRINCIPAL COMPONENT ANALYSIS

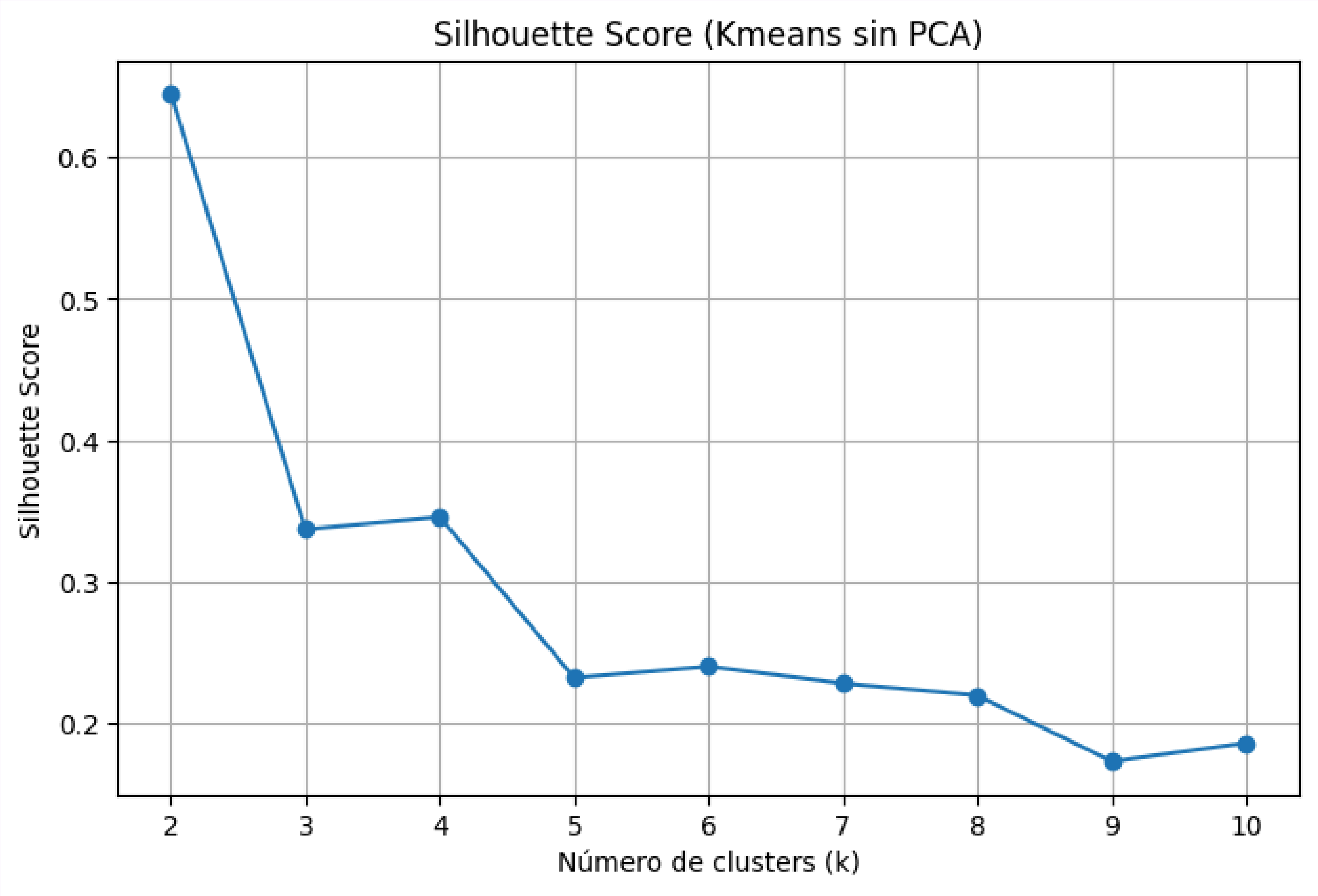


- With 2 components, 89% of the total variance is explained

- With 3 components, 94%

- With 4 components, 97%

K = 2 IN KMEANS

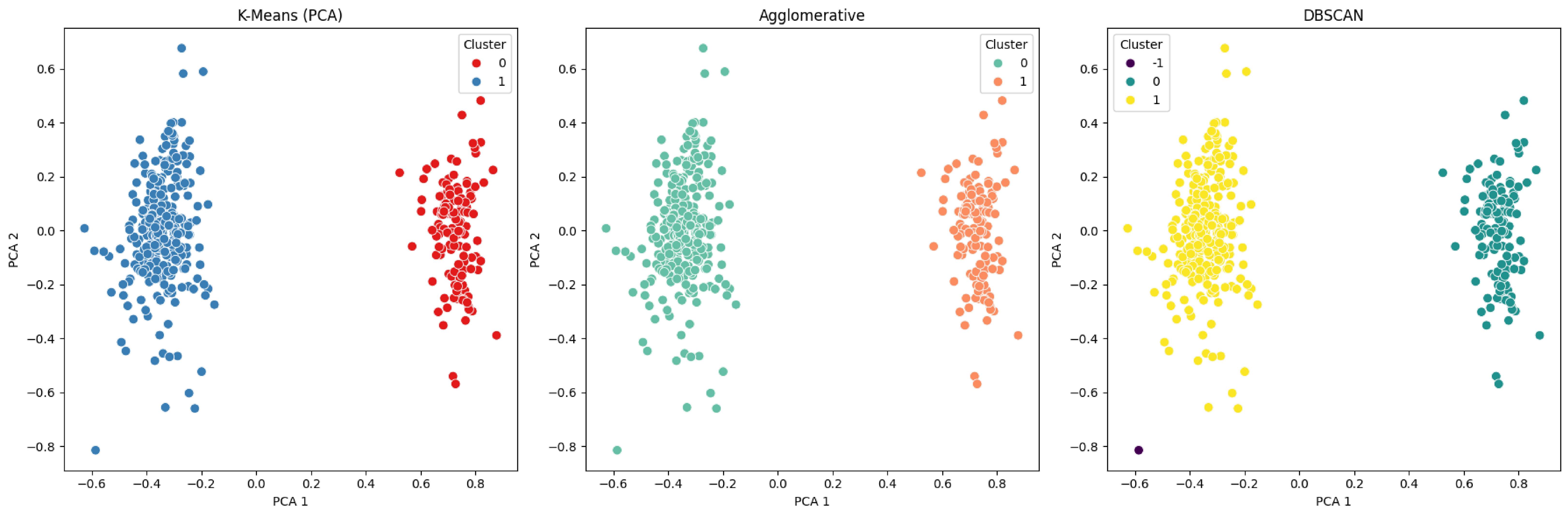


- We tested values of k between 2 and 10
- Elbow method: not conclusive
- Silhouette Score: highest value at k = 2

RESULTS COMPARISON

Algorithm	Silhouette Score
KMeans (with PCA)	0.64
KMeans (without PCA)	0.69
DBSCAN	0.79
Agglomerative	0.78

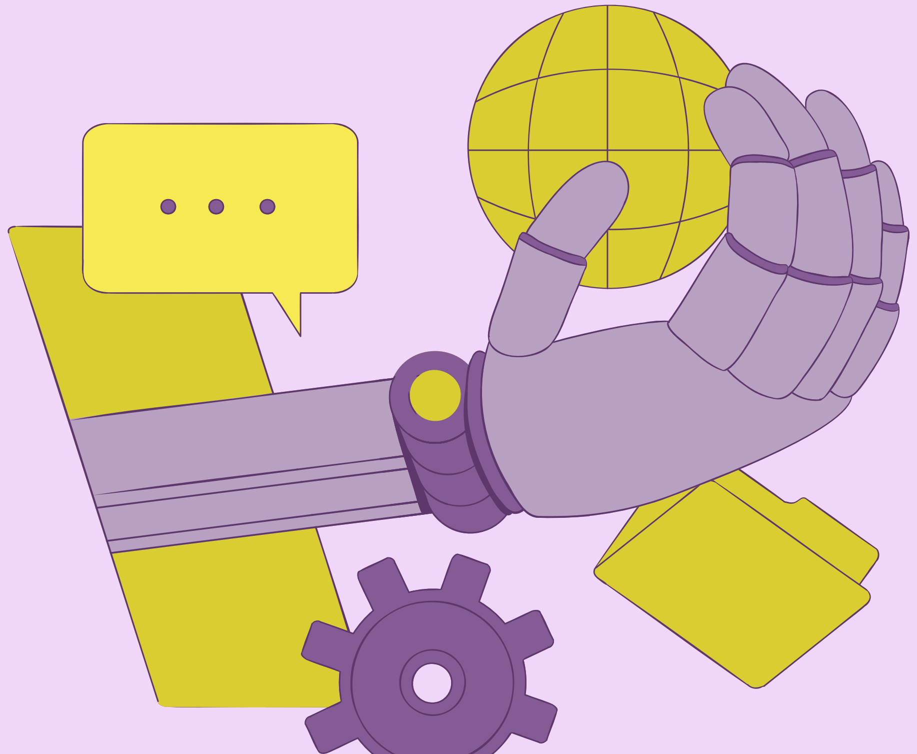
MODEL SELECTION



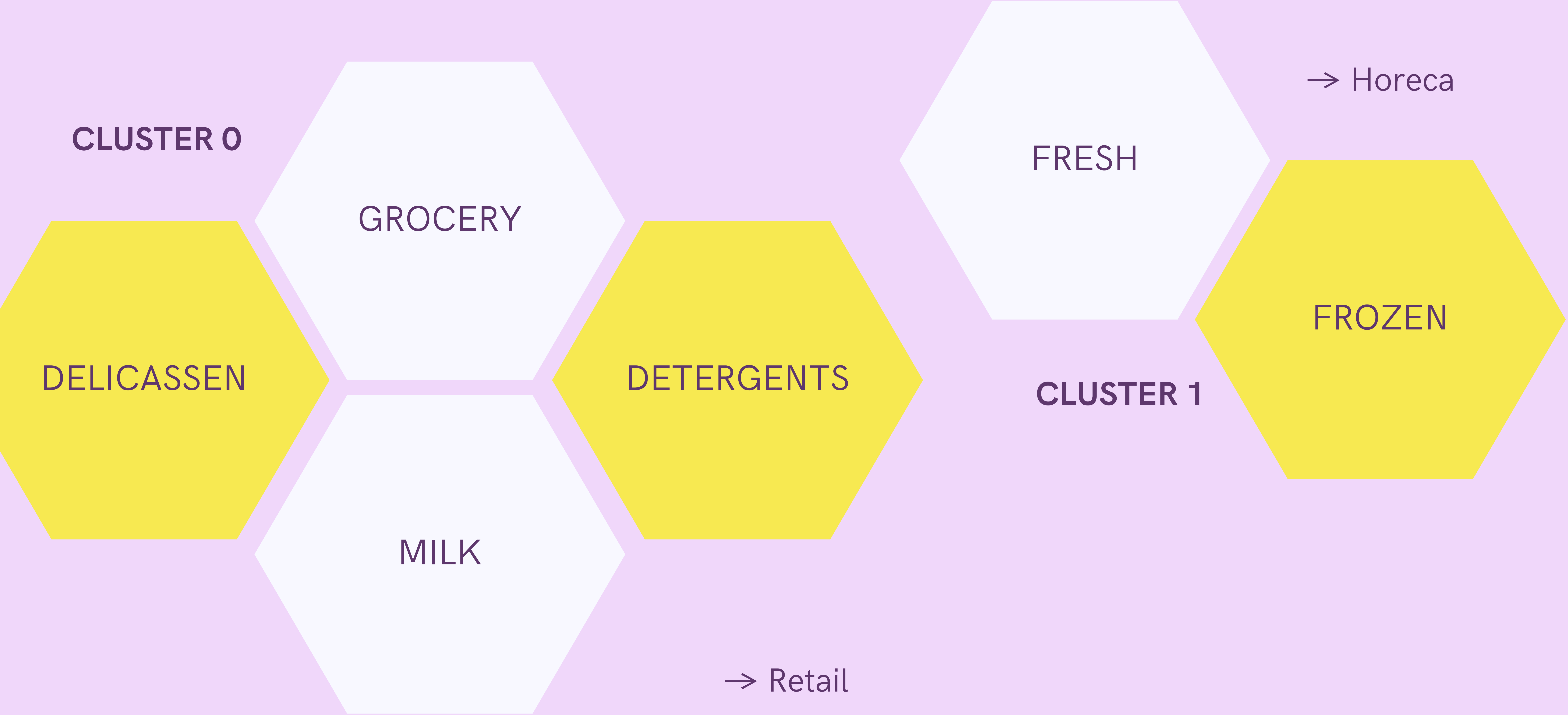
FINAL MODEL SELECTED = KMEANS WITH PCA

CONCLUSIONS

- We tested three models: KMeans, DBSCAN, and Agglomerative Clustering.
- Although DBSCAN and Agglomerative achieved higher Silhouette Scores, we selected KMeans with PCA as the final model due to:
  - Simplicity and speed: Fast to train and easy to interpret.
  - Stability: Less sensitive to parameters compared to DBSCAN.
  - Generalization: Easily applicable to new, unseen data.
  - Interpretability: Allows clear analysis of the key features in each cluster



RESULTS



CLUSTER 0

GROCERY

DELICASSEN

DETERGENTS

MILK

→ Retail

CLUSTER 1

FRESH

FROZEN

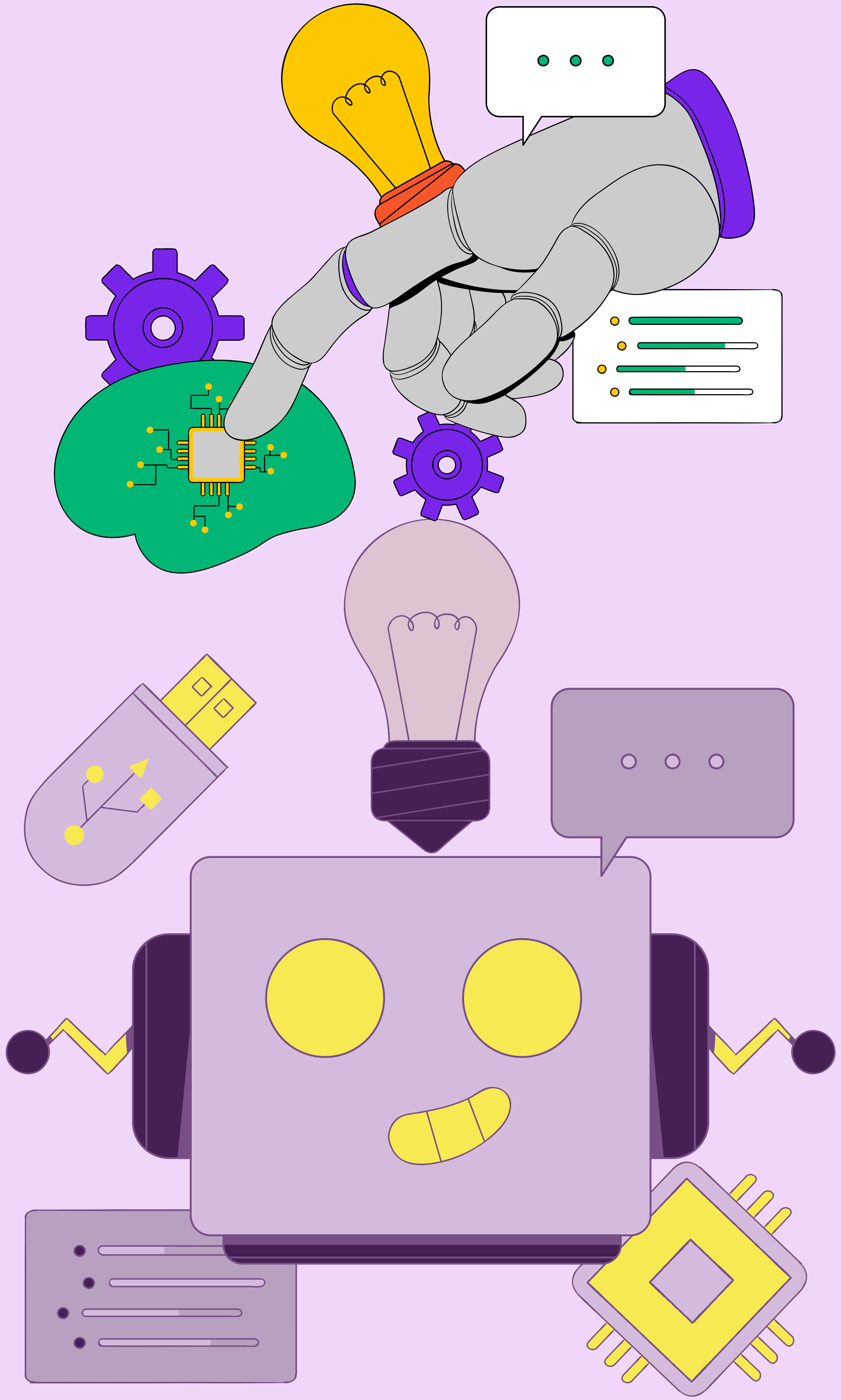
→ Horeca

POTENTIAL IMPROVEMENTS

This system is useful for customer segmentation and marketing strategy — future improvements could include:

PARAMETER TUNING

OULIER ANALYSIS



THANK YOU!