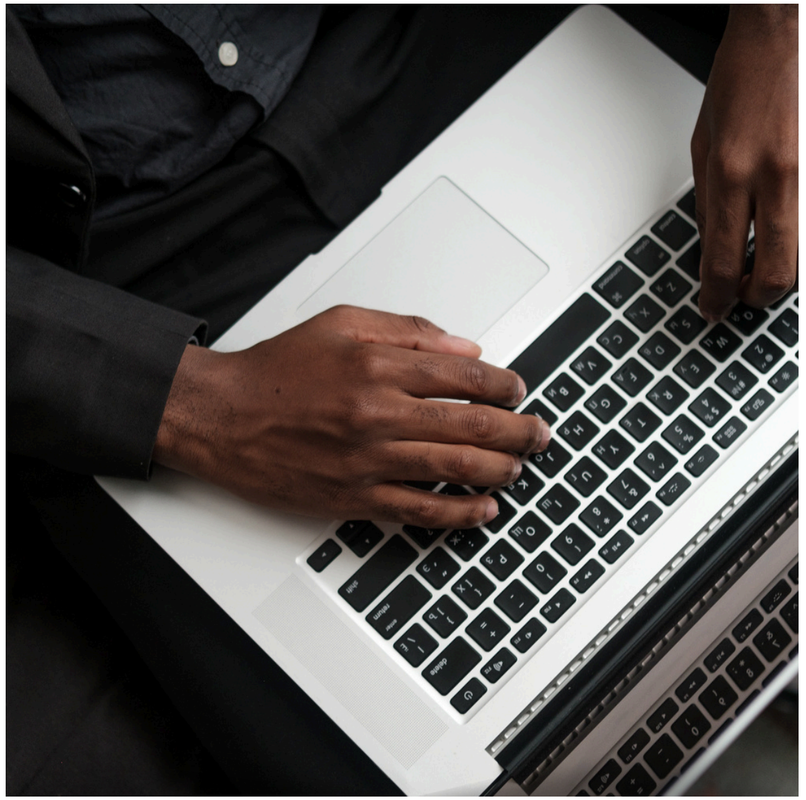
EXPLORING MULTIVARIATE METHODS



Sofia Knutas (A01831481) & Kristina Kristiansen Stapnes (A01830792)
Aplicación de métodos multivariados en ciencia de datos (Gpo 601)
Group 6

OBJECTIVE AND ROADMAP OF THE PORTFOLIO

The objective of this portfolio is to demonstrate the implementation, evaluation, and interpretation of multivariate analytical techniques - specifically factor extraction, supervised discriminant classification, and unsupervised cluster modeling - using standardized preprocessing, statistical validation, and model performance assessment.

Project 1: Factor Analysis

Identify latent dimensions in customer satisfaction survey data

Project 2: Discriminant Analysis

Classify loan applicants as defaulters or non-defaulters

Project 3: Customer Segmentation

Segment retail customers into meaningful groups using clustering

FACTOR ANALYSIS

Objective: Identify latent dimensions in customer satisfaction survey data

Outcome

1 5 key satisfaction dimensions

Method

2 Exploatory Factor Analysis (EFA)

Dataset

3 3,400 survey responses, 23 satisfaction variables

01

Data Preparation

- n = 3400 observations
- 23 survey variables retained
- Missing values < 5% removed
- Variables standardized

02

Suitability Tests

- KMO = 0.959 (excellent)
- Bartlett's test = significant
- Average correlation = 0.337

02

Factor Extraction

- Scree plot & Kaiser criterion
- 5 factors retained
- Varimax & Promax rotations applied

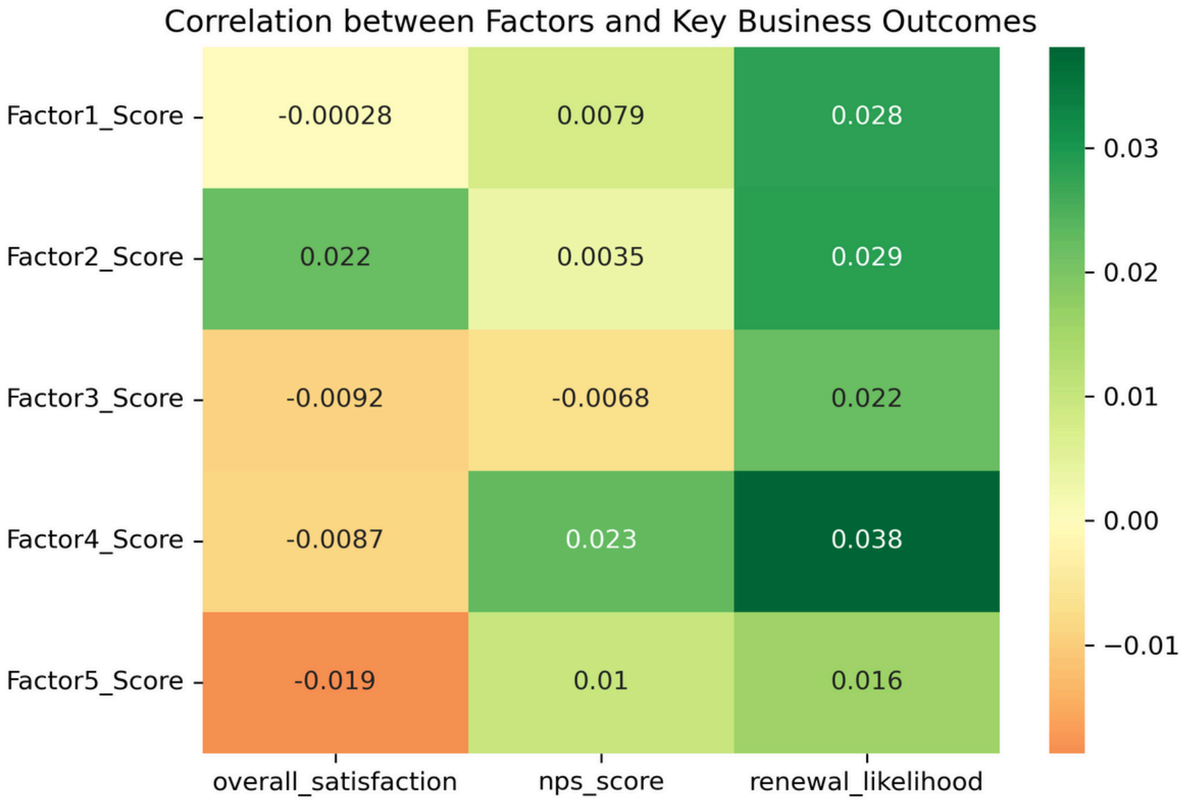
FACTOR ANALYSIS

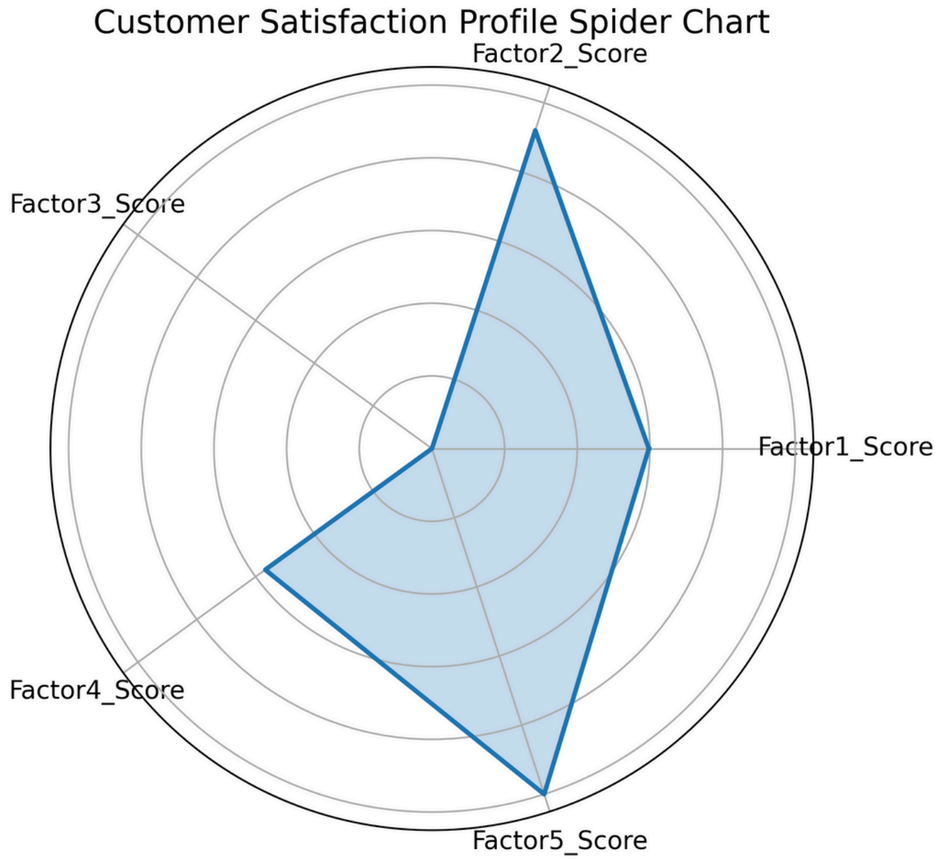
1. **Technical excellence** - Innovation, problem-solving, expertise
2. **Relationship Management & Trust** - Communication, reliability, responsiveness
3. **Project Delivery & Performance** - On-time delivery, project management, quality
4. **Financial Value & Transparency** - Cost fairness, ROI, billing accuracy
5. **Support & Service Quality** - Training, documentation, after-project help

FACTOR ANALYSIS

The spider chart shows that Factors 2 and 5 have the strongest scores, indicating they contribute most to the customer satisfaction structure, while Factor 3 has the weakest influence.

The correlation heatmap shows that all factors have low correlations with overall satisfaction, NPS, and renewal likelihood, meaning the extracted factors do not strongly predict business outcomes. Factor 4 shows the highest association with renewal, but still at a weak level.





DISCRIMINANT ANALYSIS

LendSmart

CREDIT RISK ANALYSIS

Objective: Classify loan applicants as defaulters or non-defaulters

Dataset Overview

• n = 2500 applicants
• Binary target variable
• Mix of numerical & categorical predictors

Preprocessing

• Dummy encoding of categorical variables
• Train/test split (80/20)
• Standardization applied

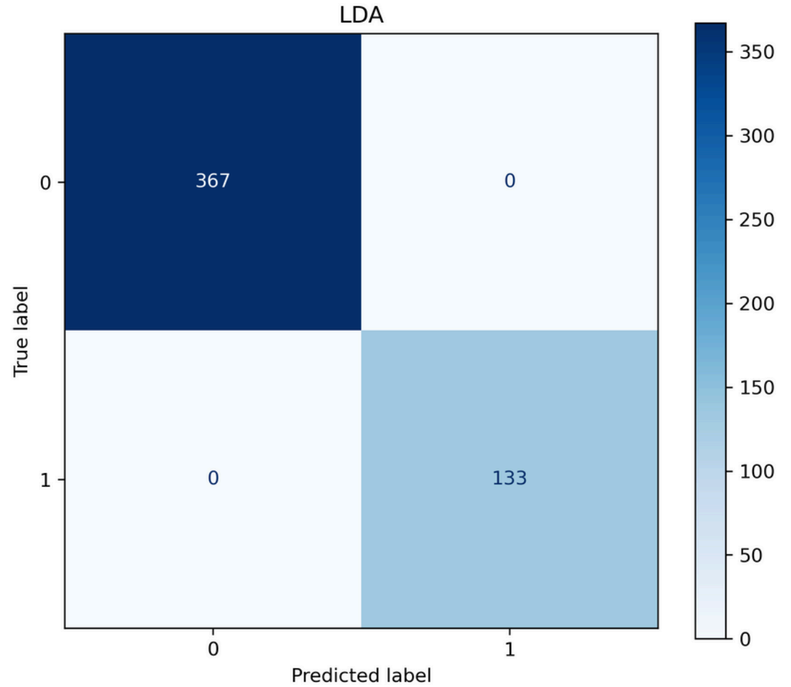
LDA and QDA model

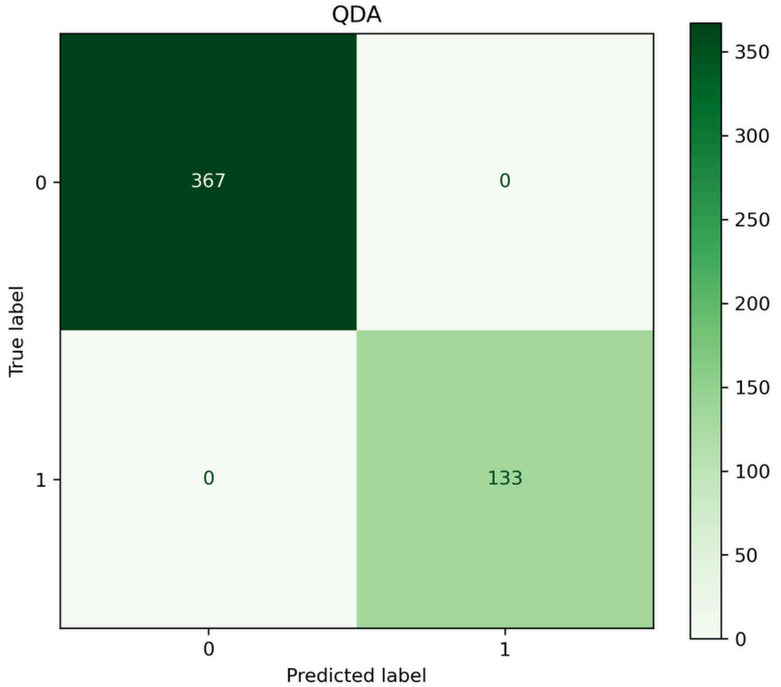
• Confusion matrices
• Classification metrics
• ROC curves & AUC

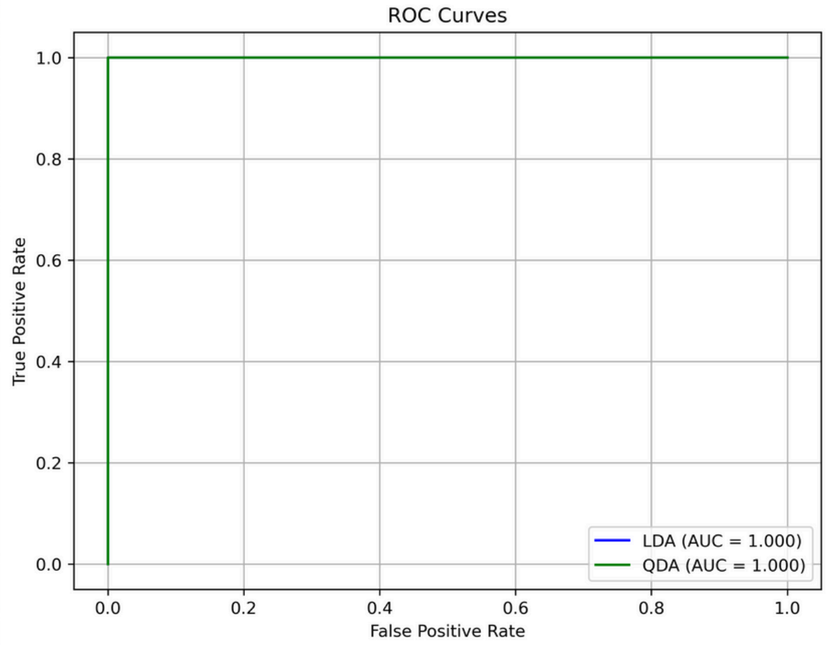
DISCRIMINANT ANALYSIS

Discriminant Conclusions

- Both models perform strongly
- LDA preferred (simpler + similar accuracy)







CLUSTER ANALYSIS

Objective: Segment retail customers into meaningful groups using clustering

01

Data & Preparation

- n = 3000 customers
- 9 behavioral variables
- Standardized for distance-based algorithms

02

Hierarchical Clustering

Uses data analysis to generate insights, inform strategies, support decision making across multiple functions.

CLUSTER ANALYSIS

03 Hierarchical Clustering

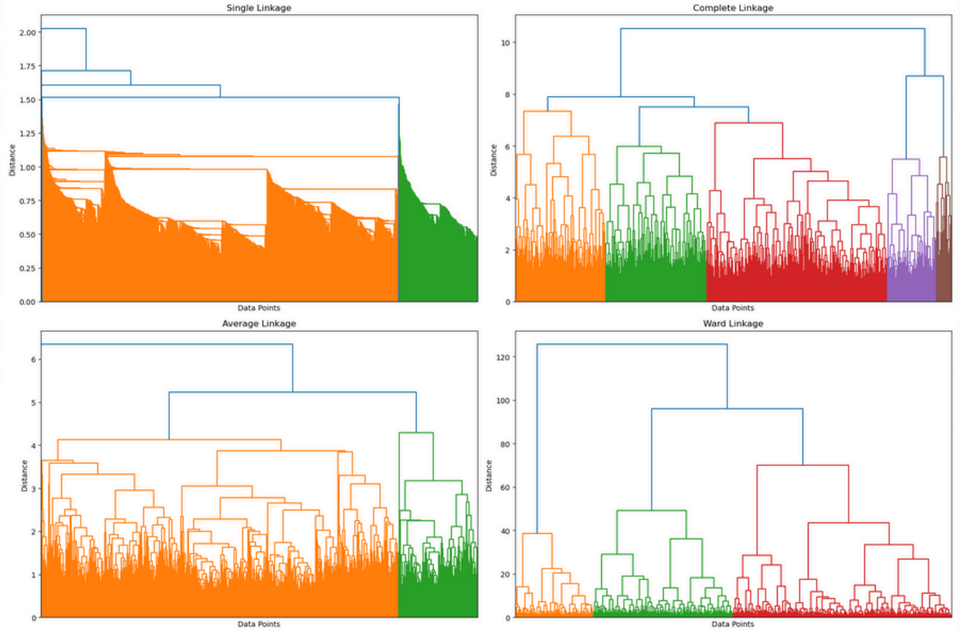
- Single linkage → chaining
- Complete & average → weak structure
- Ward's method → best separation

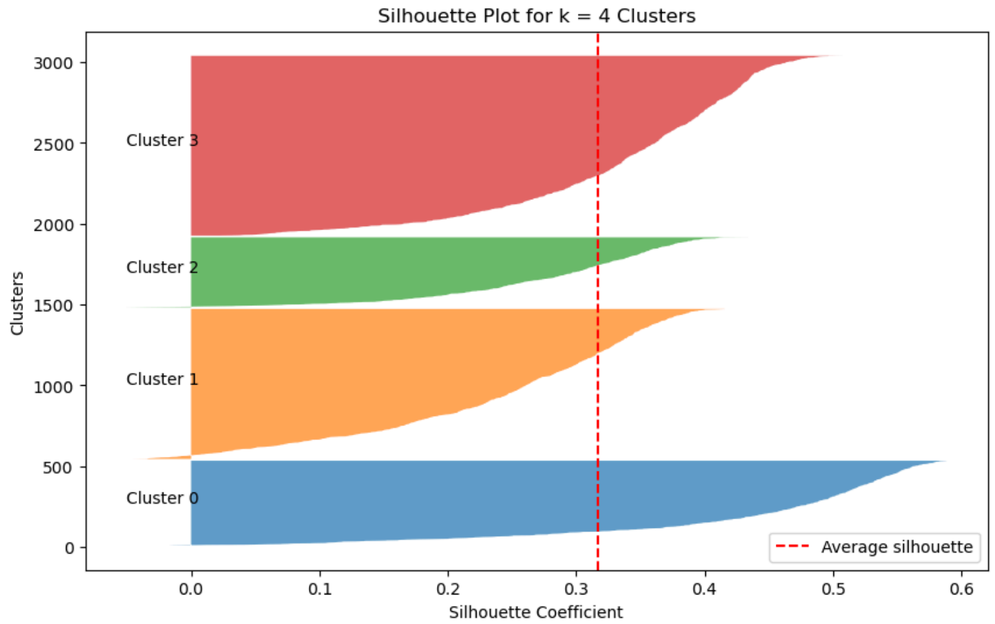
04 Determining k

- Silhouette scores tested
- Elbow method applied
- Optimal number of clusters: 4

05 Cluster Validation

- Silhouette score strong overall
- Clear boundaries for clusters 0 & 3
- Cluster 2 weakest separation





CLUSTER ANALYSIS

Segment 1: High-Value Frequent Shoppers

Alex

- Shops often
- High spending
- Engaged
- Low returns

Segment 2: Low-Activity Window-Shoppers

‘Amanda

- Browses a lot
- Rarely buys
- High returns
- Low engagement

Segment 3: “Bob”

Luxury Occasional Shoppers

Bob

- Buys rarely
- High-value purchases
- Seasonal activity
- Interested in premium products

Segment 4: ‘Alice

Budget-Conscious Regular Buyers

- Shops consistently
- Moderate spending
- Values deals
- Largest segment

Alice the Practical Planner

TOOLS USED

- Python
- pandas
- numpy
- seaborn / matplotlib
- scikit-learn
- factor_analyzer
- scipy

Methods Used

- Standardization
- KMO / Bartlett
- Scree plot
- Varimax / Promax
- LDA / QDA
- Confusion matrices
- ROC / AUC
- K-Means
- Silhouette analysis
- PCA projection

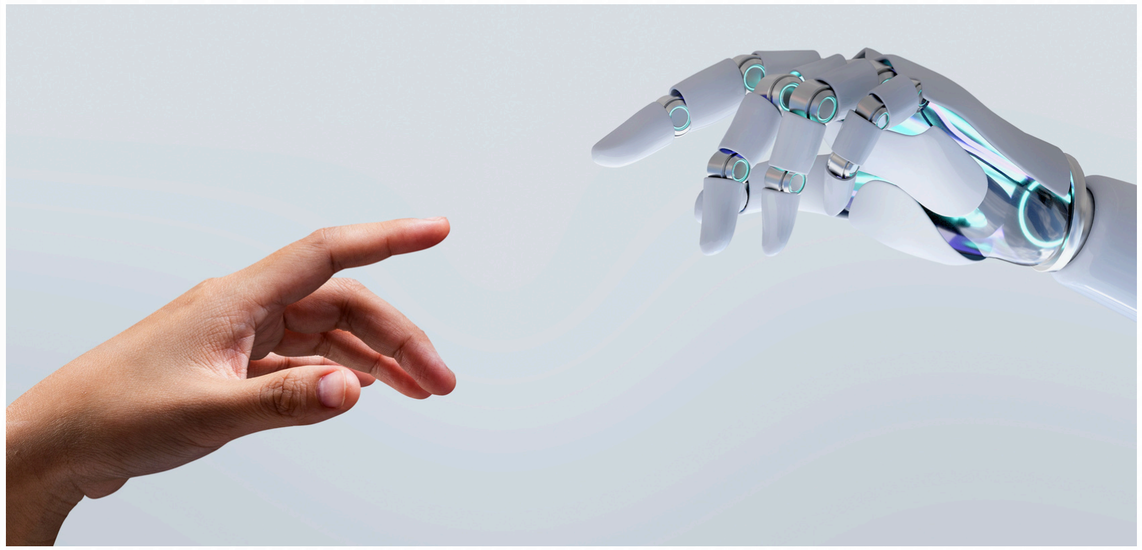
REFLECTION

01 **Factor analysis confirmed 5 latent satisfaction dimensions**

02 **Discriminant analysis classified loan default risk effectively**

03 **Clustering identified 4 meaningful customer groups**

REFLECTION



Technical skills gained:

- Data preprocessing & scaling
- Factor analysis and PCA
- LDA & QDA classification
- Hierarchical and K-Means clustering
- Model evaluation (AUC, accuracy, silhouette)
- Data visualization and interpretation

Analytical skills gained:

- Identifying patterns and structures in data
- Comparing models and selecting best fit
- Drawing conclusions from statistical results

THANKS FOR LISTENING!

Go to the GitHub repository

Sofia Knutas (A01831481) & Kristina Kristiansen Stapnes (A01830792)
Aplicación de métodos multivariados en ciencia de datos (Gpo 601)
Group 6