



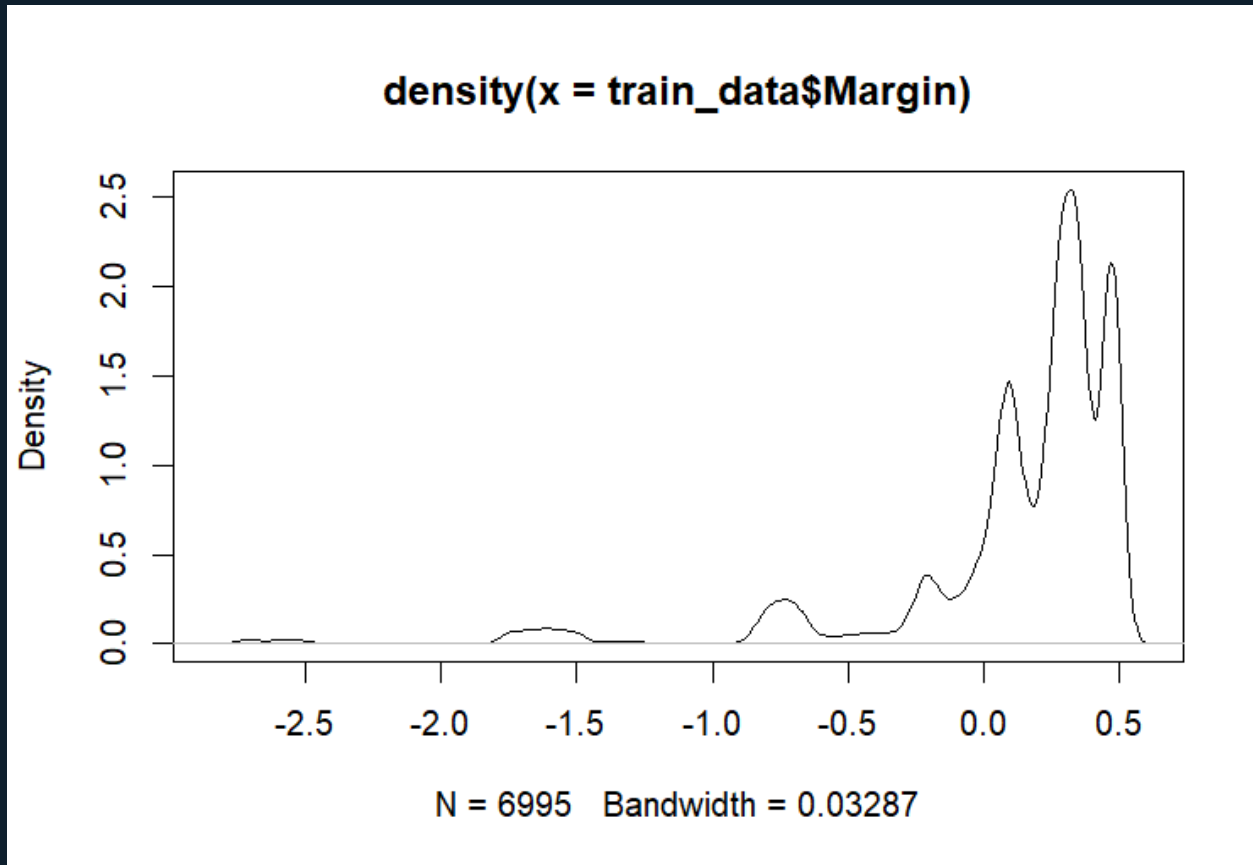
Margin Analysis with Mixture Models and Bootstrap

Prediction and Inference on a
Superstore Transactions Dataset
using R

Matheus Khatib – 12/20/25

Exploratory Data Analysis

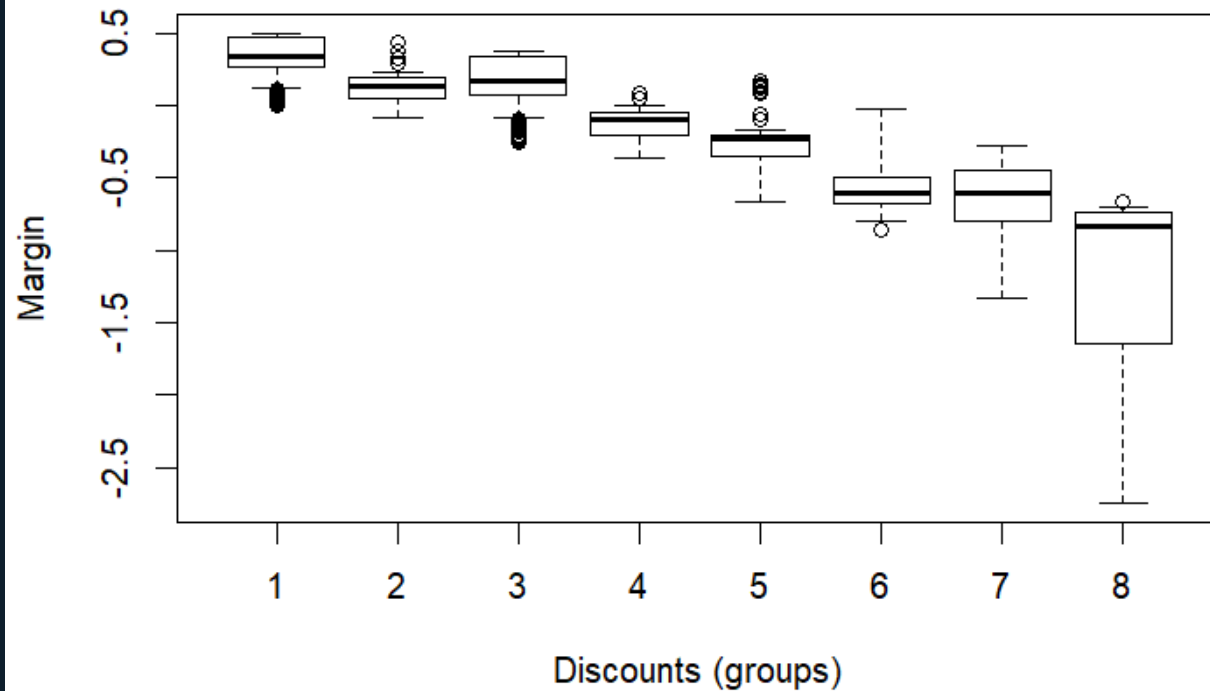
The profit margin shows clear signs of non-normality



- Heavy tail in negative values
- Clear presence of subgroups
- Small number of negative margins

Preparation of Covariates for Modeling

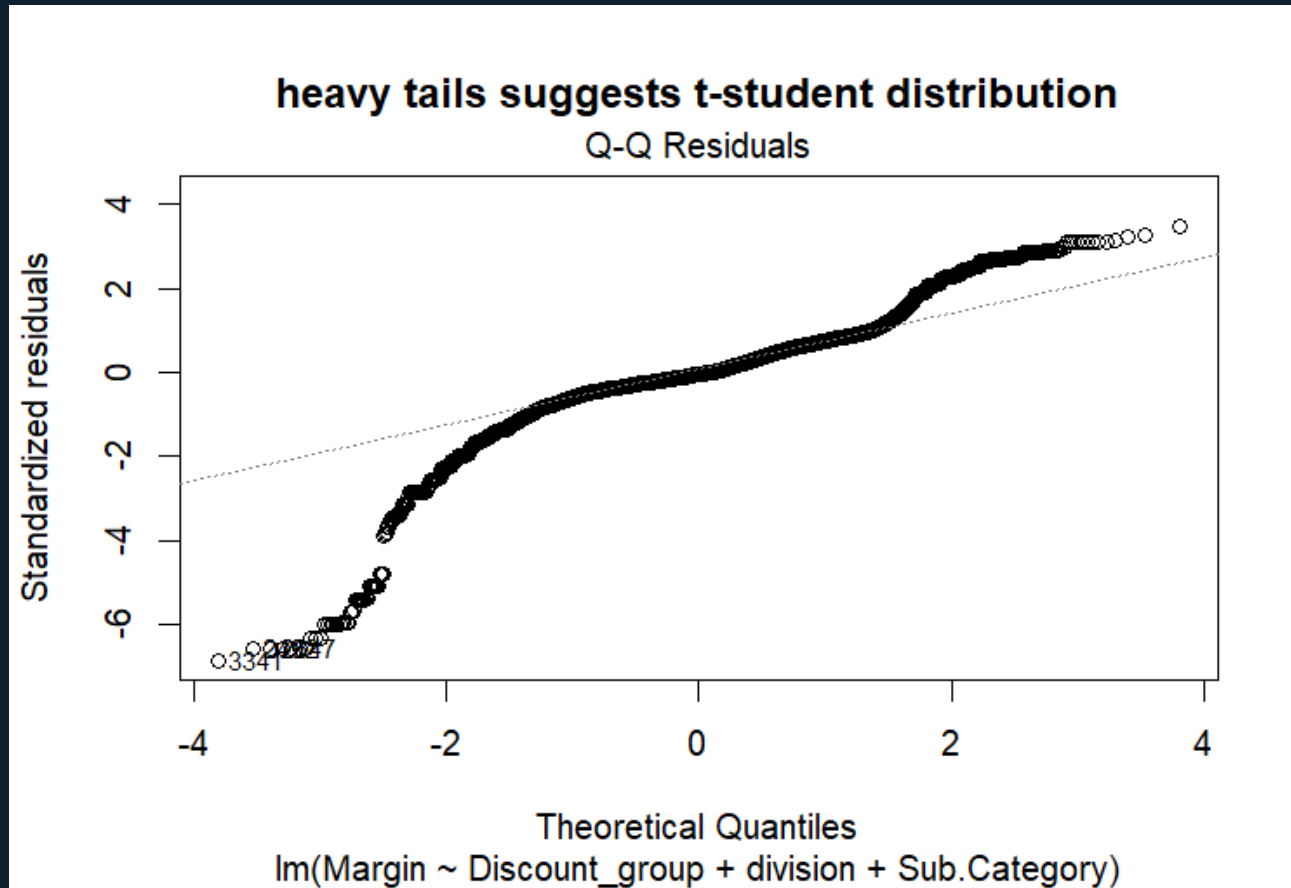
Example of grouped variable: Discount



- Rare categories were grouped
- No missing values (NAs)
- Variables with many factor levels were initially excluded
- Data split into training, validation, and test sets

How linear models predict the
response?

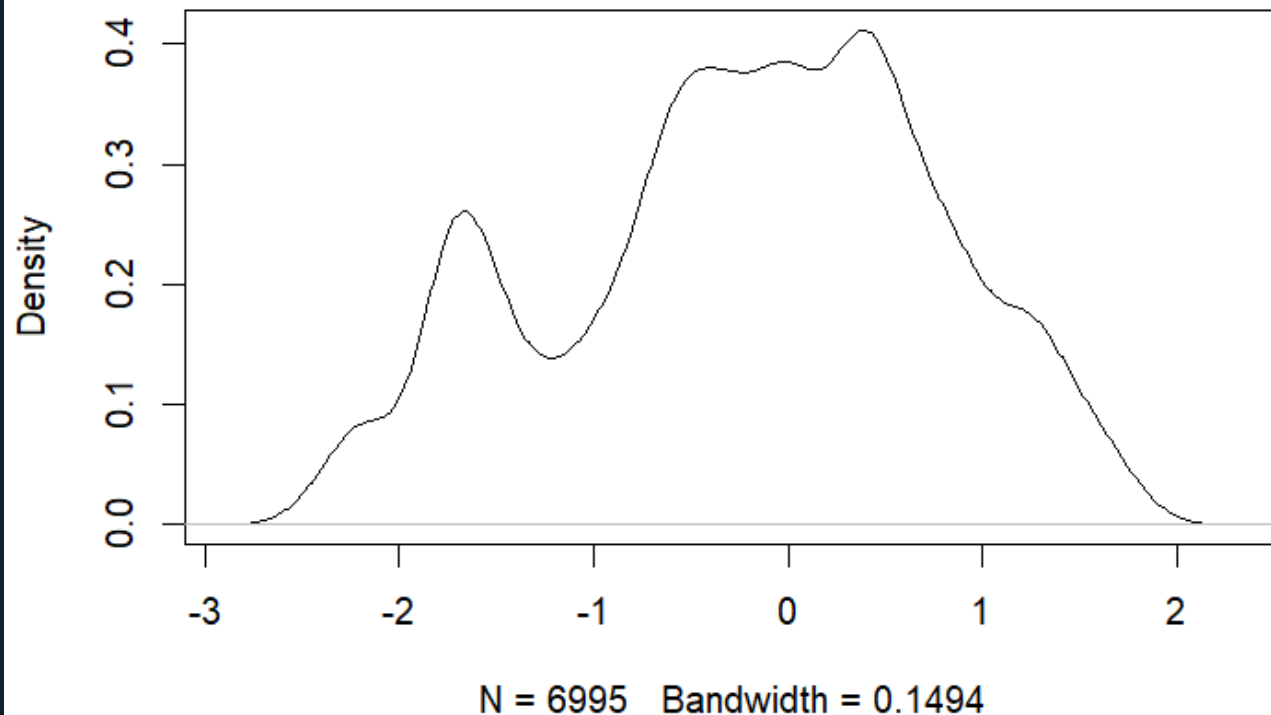
Multiple linear regression model



- MAE = 0.114
- RMSE = 0.169
- Predicted $R^2 = 0.877$
- It showed strong performance metrics given its simplicity

Generalized Additive Model with a Student's t response (GAMLSS)

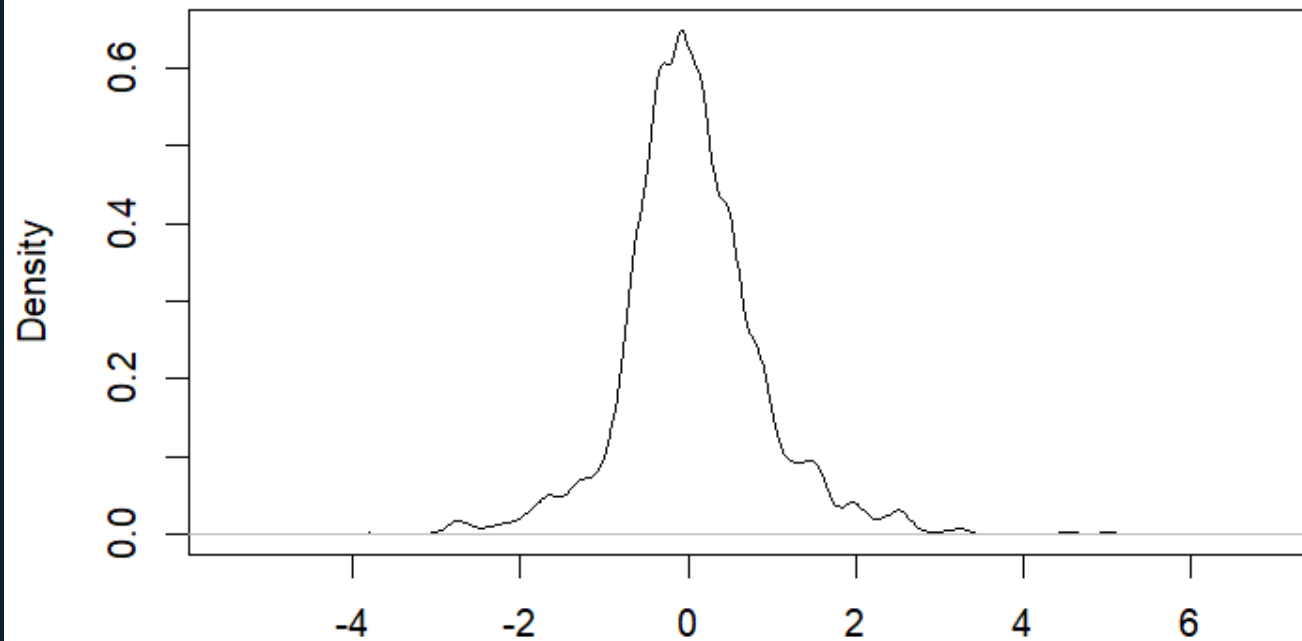
Presence of bimodality in the residuals' density



- MAE = 0.101
- RMSE = 0.228
- Predictive $R^2 = 0.777$
- It showed inferior predictive power compared to the simpler model

Gaussian Mixture Generalized Additive Model(Flexmix)

Residual density approximates a normal distribution (good fit)



N = 8494 Bandwidth = 0.09561

- MAE = 0.043
- RMSE = 0.075
- Predictive $R^2 = 0.978$
- Model with the best predictive performance metrics

How do other models perform
on this dataset?

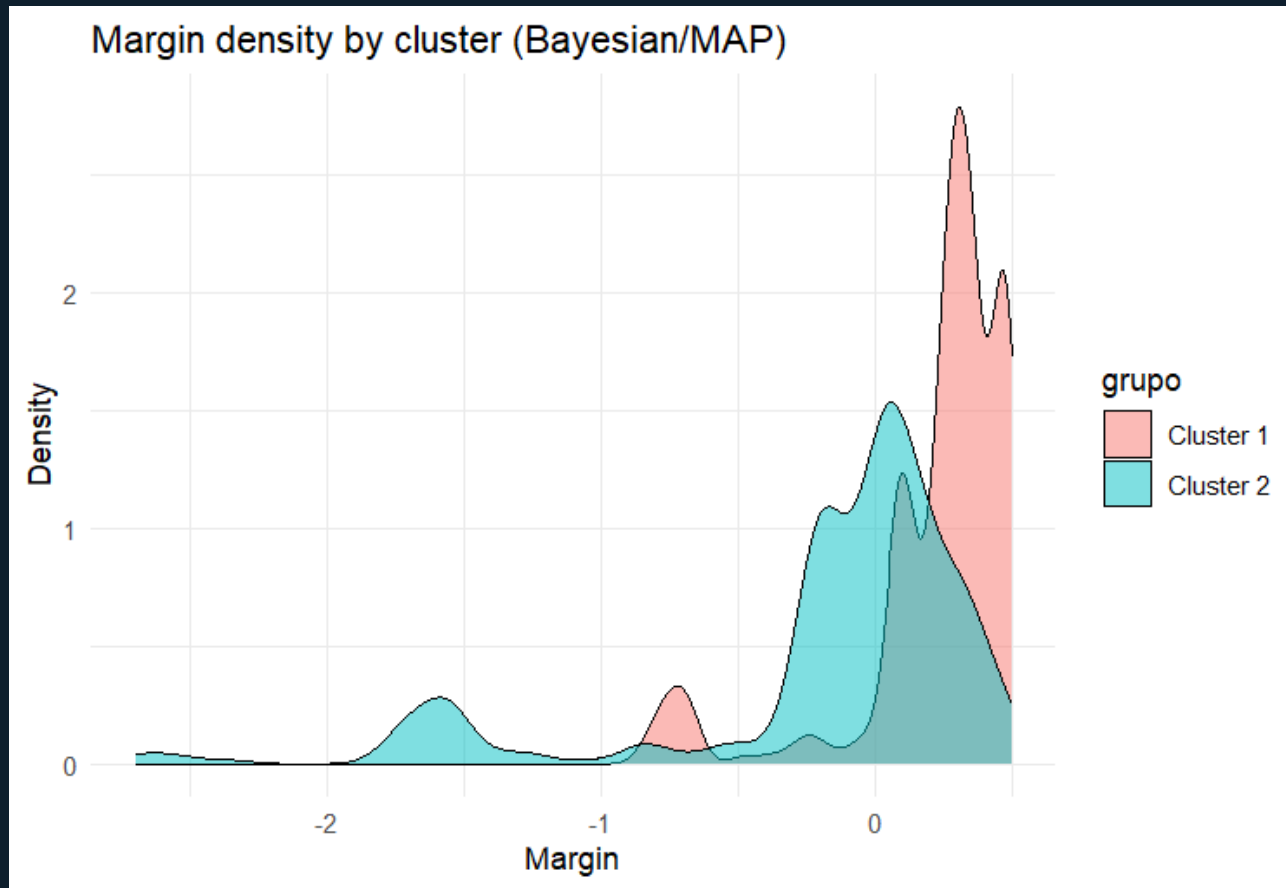
Predictive metrics of all fitted models

	modelo	rmse	mae	r2_adj
1	Linear Gaussian model without interactions	0.168867060750838	0.11397572686095	0.877432744106153
2	Linear Gaussian model with interactions	0.155074617837345	0.0924882623026571	0.896636783927642
3	GAMLSS Student-t model without interactions	0.22765350501957	0.100815018447282	0.777242027219483
4	GAMLSS Student-t model with interactions	0.193176201546834	0.0918093501361032	0.8396046332497
5	Additive Gaussian mixture model	0.0747695699878902	0.0430496057868475	0.975971075082614
6	Heteroskedastic Gaussian mixture model	0.191308142458227	0.103269535694373	0.842691755614484
7	Two-part hurdle model	0.188174827323718	0.0920458823302395	0.848006745866307
8	Regression tree	0.153263024092494	0.113054854078366	0.89903767210855
9	Pruned regression tree	0.159642059502327	0.0986989010517414	0.890458365157249
10	Bagged regression tree	0.132275057526771	0.0847922218089116	0.924796077596159
11	Random forest	0.144080193862036	0.093022259908126	0.910773646872462
12	Boosted regression tree	0.119724511493462	0.0758424007051833	0.938390069078774
13	XGBoost	0.121513597539088	0.0754919026706865	0.936534993029231
14	LightGBM	0.123344872714506	0.0776112265259605	0.934607675308276

The Gaussian mixture model achieves the best predictive performance while still allowing inference

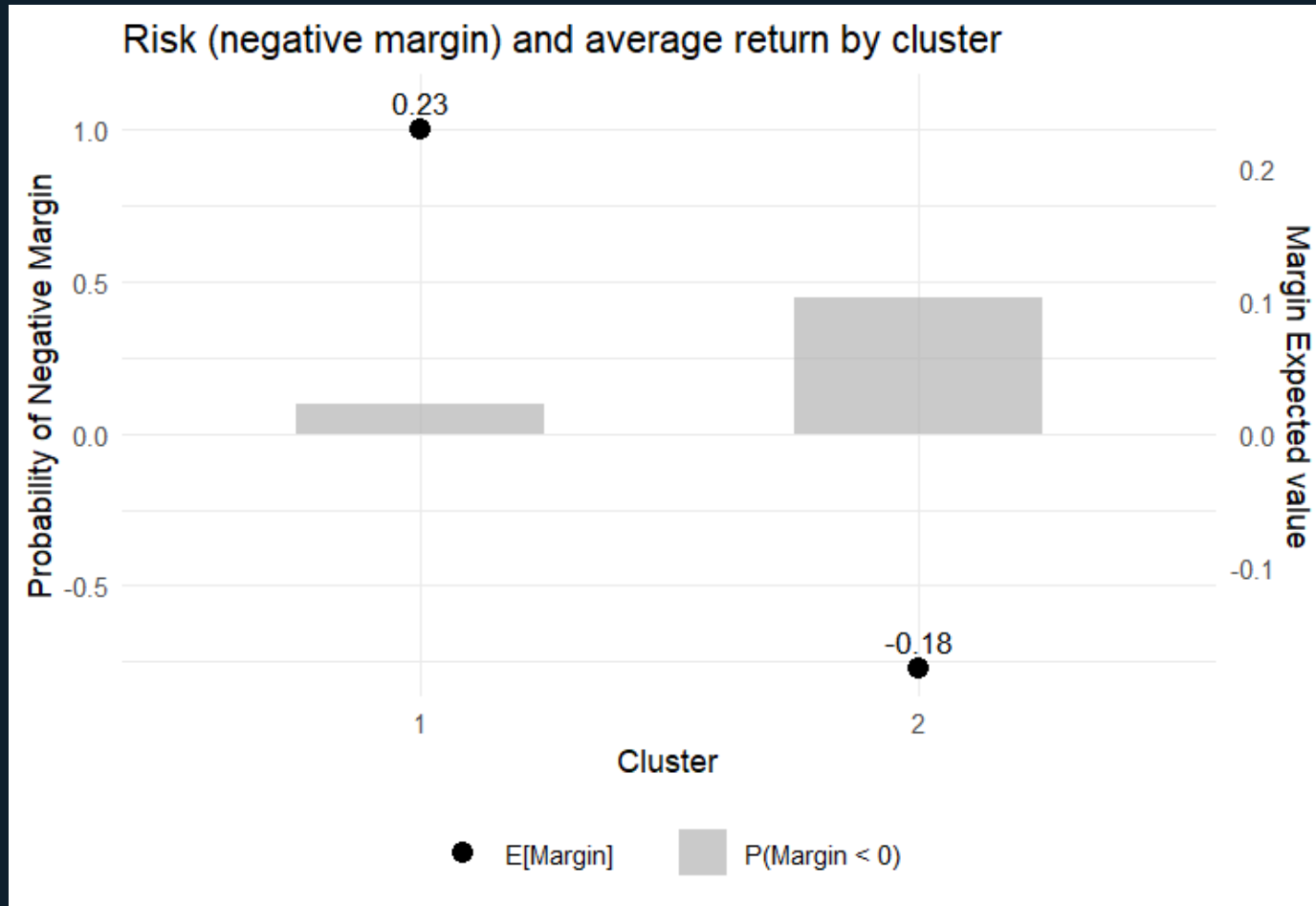
Which variables distinguish the profit
margin between clusters?

Cluster 2: higher variability and risk



- The second group exhibits more extreme negative margins
- The first group has a slightly higher median margin
- Margin variability is substantially higher in the second cluster

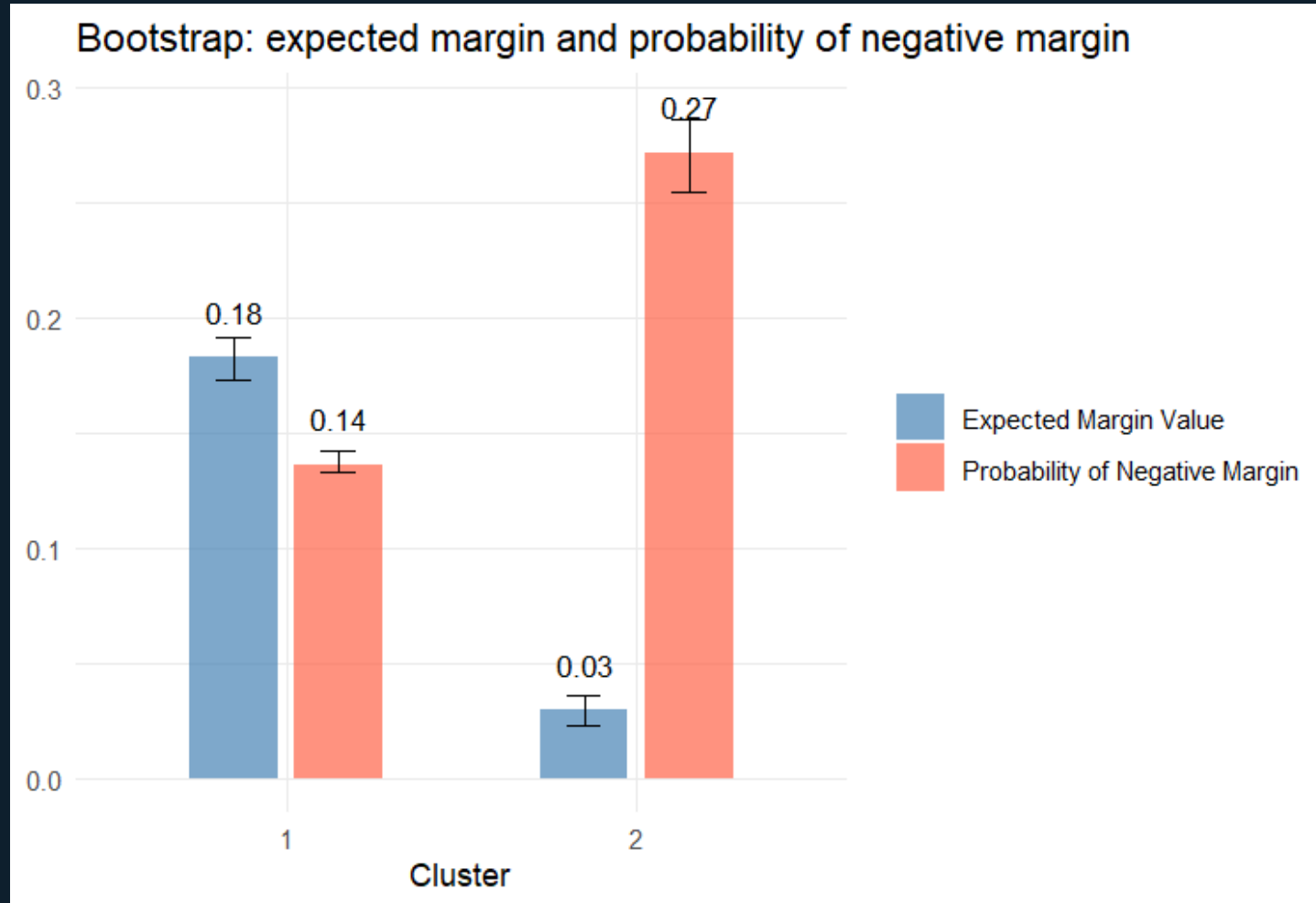
Cluster 2: greater probability of negative margin



- Group 2 is more likely to generate negative margins
- Cluster 2 has a negative expected profit margin
- Cluster 1 shows a positive expected profit margin

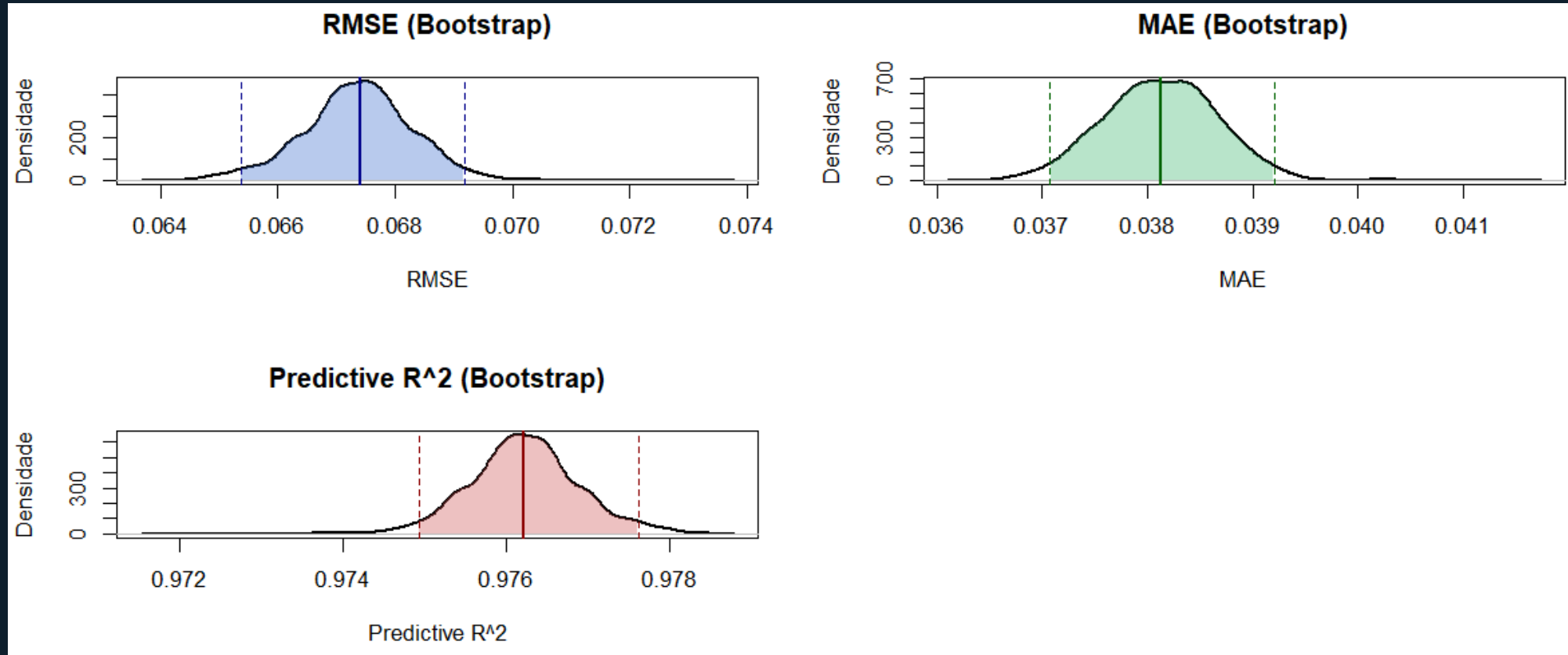
But could these results be
generalized for future transactions?

Bootstrap: Cluster 2 shows a higher risk of negative margin ($B = 1000$)



- Cluster 2's expected margin may be underestimated due to limited bootstrap iterations (computational cost)
- Cluster 2 is more likely to produce negative margins
- Cluster 1 shows a positive expected margin

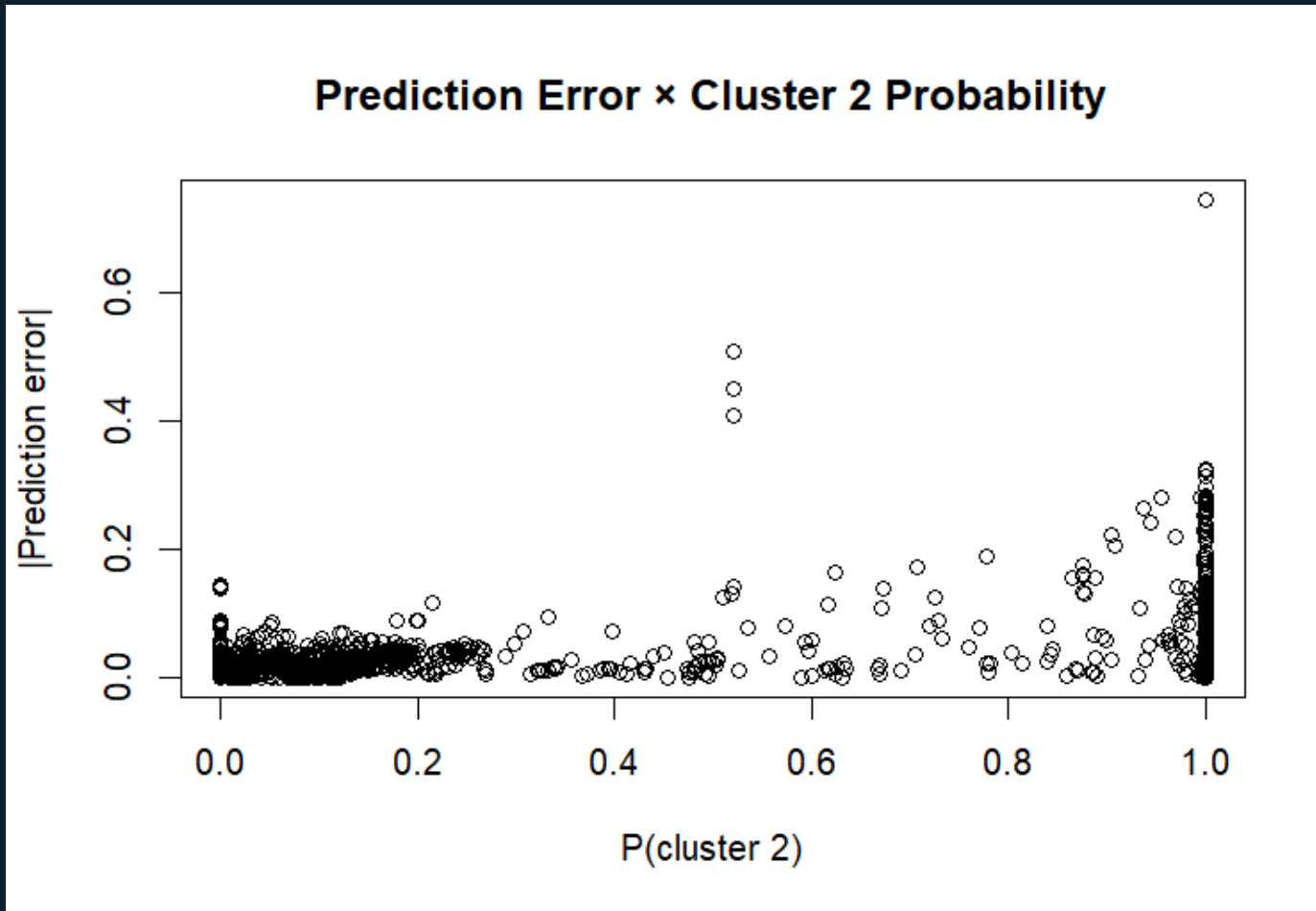
Bootstrap: Predictions show low variability ($B = 1000$)



Both metrics show intervals concentrated around small error values

Does the model sufficiently
distinguish between the clusters?

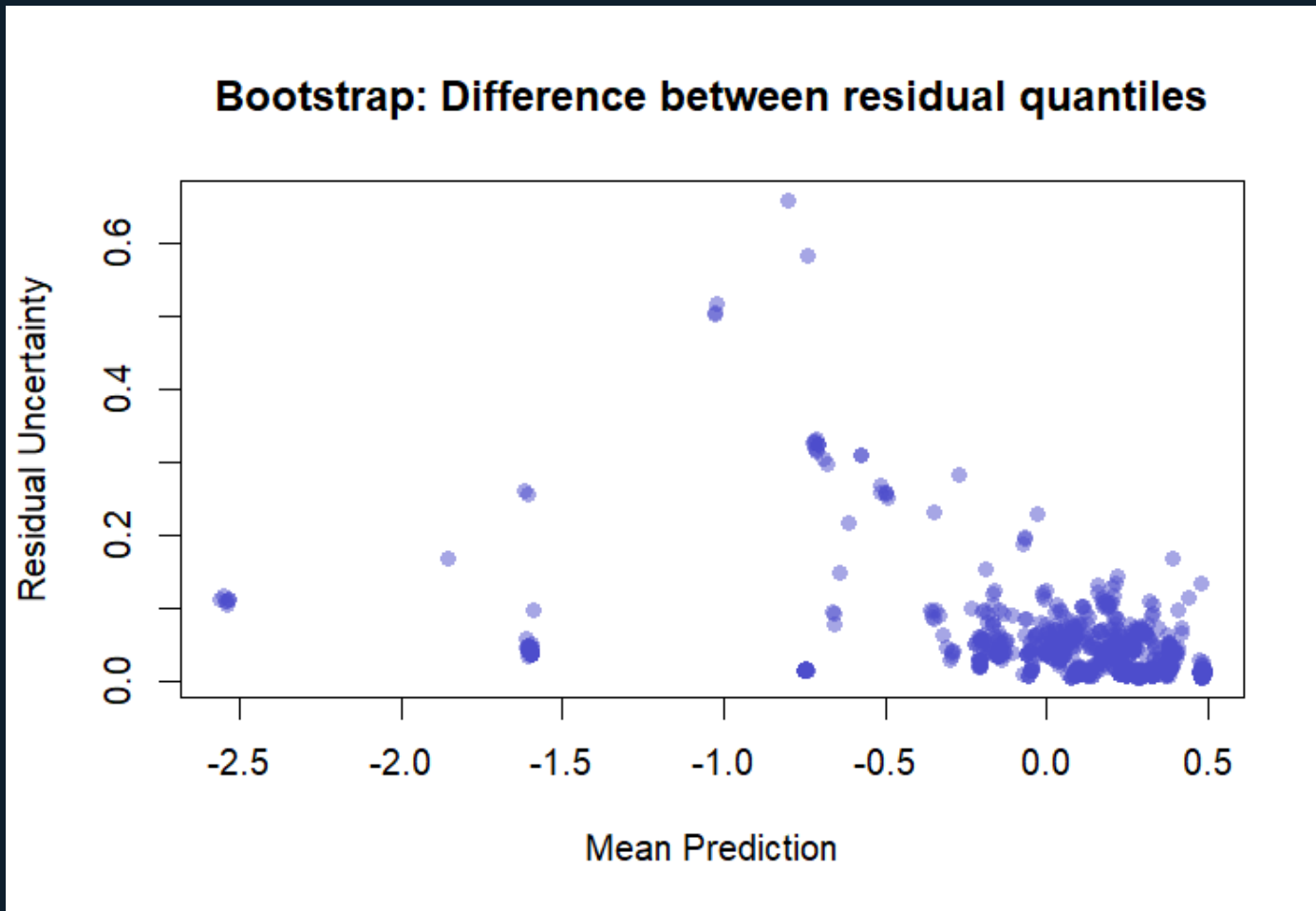
The margin of Cluster 1 is also sensitive to discounts



- Cluster 1 is classified with higher accuracy
- Classification errors are higher for Cluster 2
- Low entropy (0.281) indicates high confidence in cluster assignment

Is it possible to improve the model?

Mixture model with heavy tails and heteroscedasticity

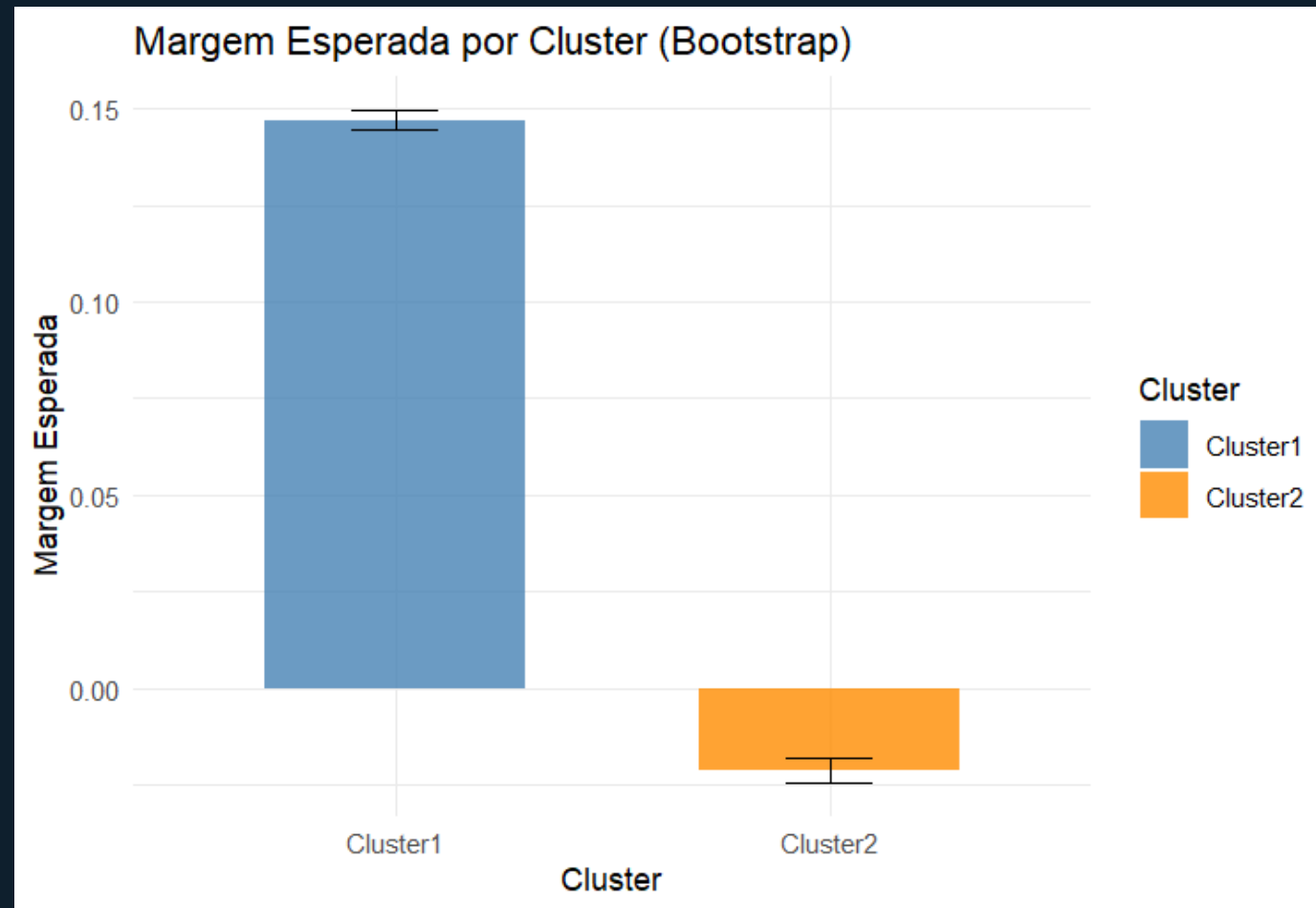


- Heavy tails were identified early in the analysis
- The left-hand plot suggests heteroscedasticity
- A heteroscedastic Student's t mixture model is recommended

How do the clusters differ in terms
of margin?

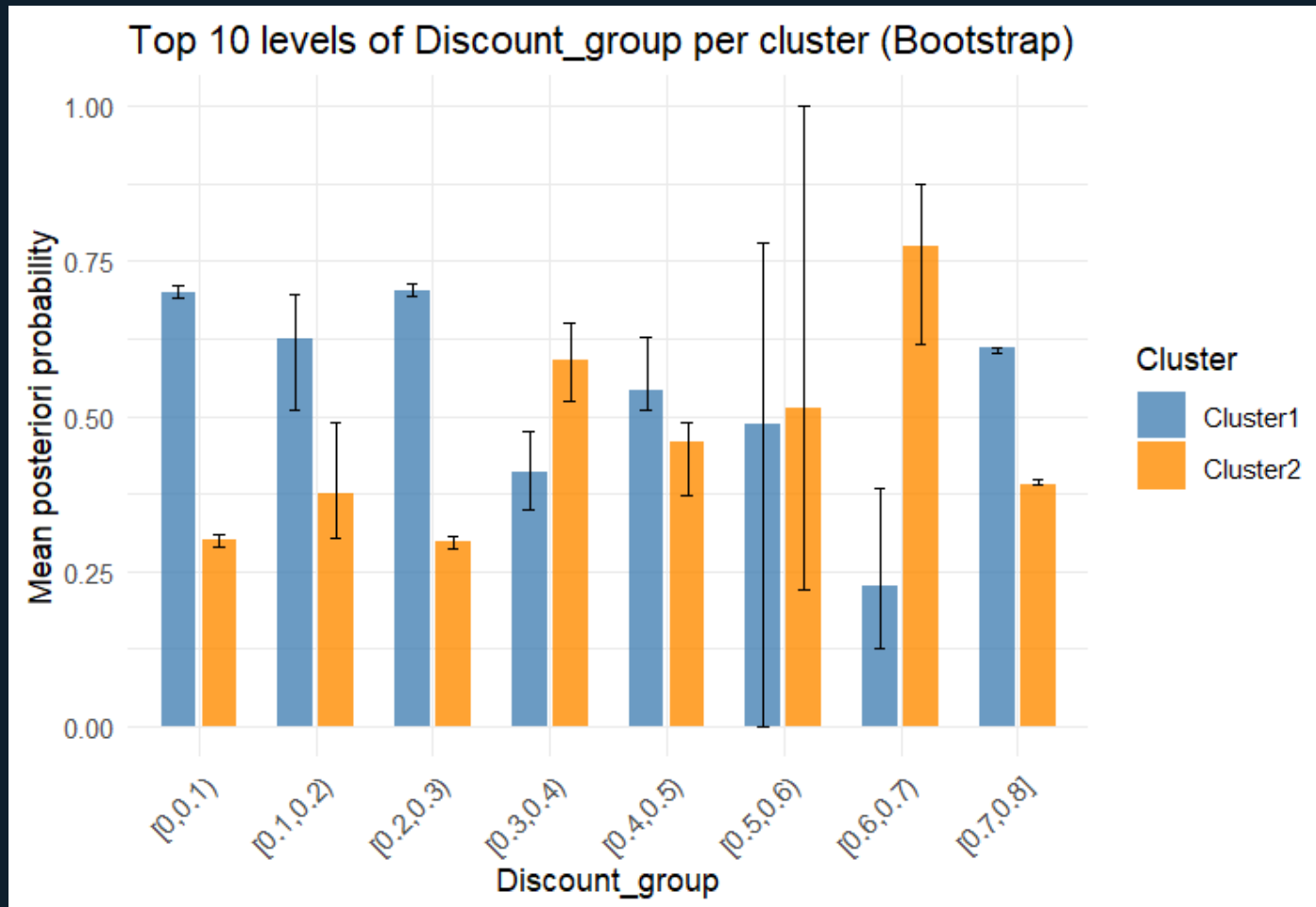
Cluster 1: positive expected profit

Cluster 2: negative expected profit



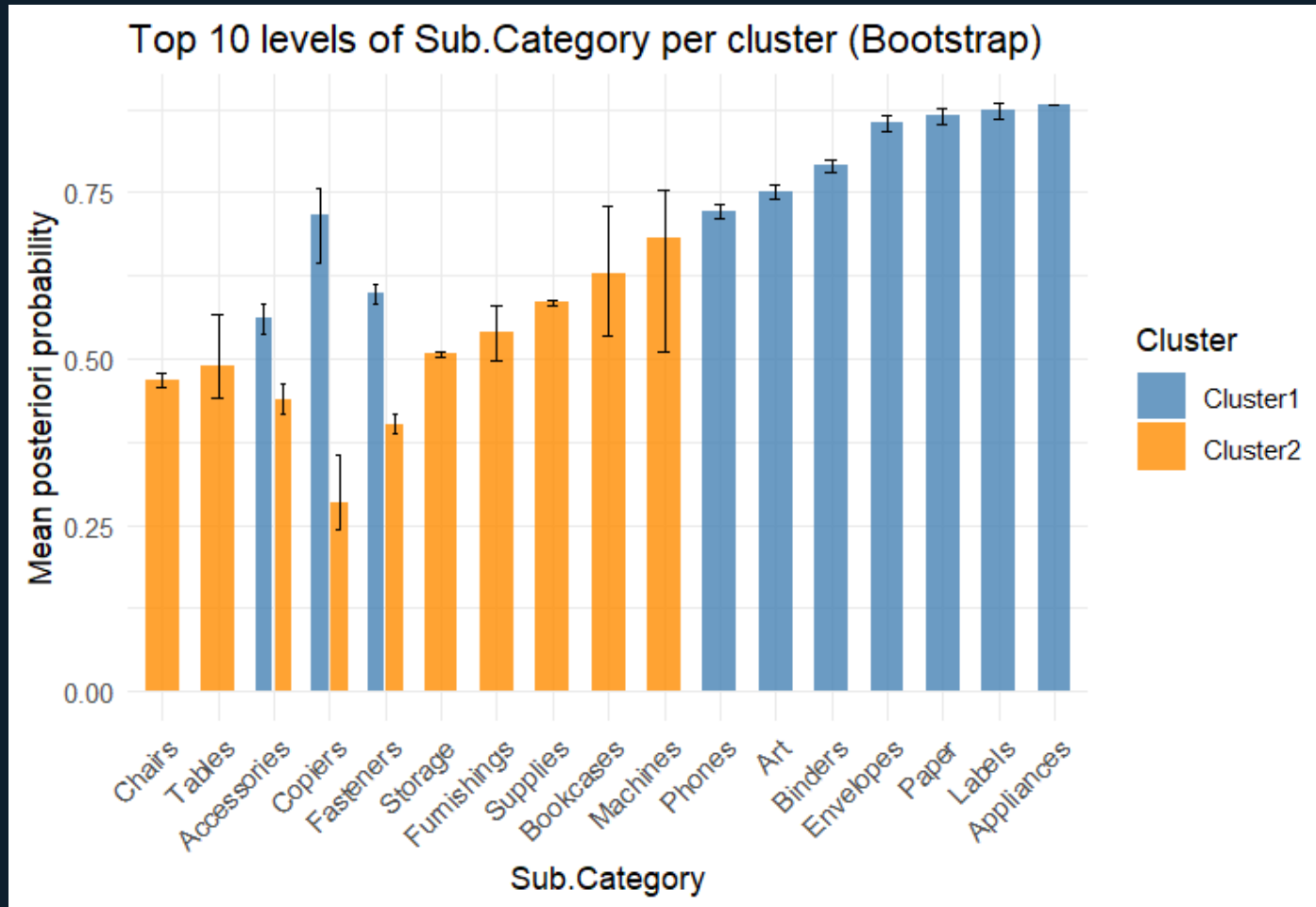
Which covariates primarily
characterize each cluster?

Cluster 2 shows higher probability across three discount ranges



- Cluster 2 (low-margin) is more prevalent at medium and high discount levels
- Uncertainty peaks in the 50%–70% discount range
- Other discount levels exhibit low variability

Subcategory is the most decisive factor in margin profitability



- Low-margin products (Cluster 2) are concentrated in machines, bookcases, supplies and furnishings
- High-margin products (Cluster 1) are more prevalent in office supplies, art, and phones
- Uncertainty across categories is limited

What are the final recommendations to
increase profit margin?

Review discount policies and closely monitor product sales.

- Review discount policies for the 30%–40% and 50%–70% ranges
- Closely monitor sales of low-margin products such as furniture, bookcases, machines, and supplies
- Increase cross-selling of high-margin products, particularly office supplies, art, and phones