Customer Segmentation Analysis

Ian Movius
August 08, 2025

Executive Summary

This project aimed to segment quiz-taking customers into actionable personas to guide targeted marketing and product strategy. The analysis combined categorical and numerical quiz and order features using Factor Analysis of Mixed Data (FAMD) for dimensionality reduction. The segmentation was performed in two passes: Pass A included LTV in the clustering features, while Pass B excluded LTV from clustering but overlaid it afterward. Each pass began with an initial FAMD and clustering run, followed by pruning low-loading variables and re-running the process for the final clusters. Cluster stability was measured using the Adjusted Rand Index (ARI) across two random seeds. Finally, clusters containing fewer than 5% of customers were relabeled as unclustered (-1).

Both passes achieved high stability (ARI of 0.831 for Pass A and 0.880 for Pass B). Dimensionality reduction retained between 10 and 14 components per pass, preserving 86–92% of cumulative variance. Distinct personas emerged in each pass, with differences in how value separated the groups.

2) Data Preparation

The dataset combined quiz results with order history extracted from `raw_data_v3.csv`. Test orders were removed, as they were found in earlier runs to artificially inflate separation between customers. Features used for clustering included categorical variables such as quiz result, bowel movement pattern, GI symptom category, acquisition code group, gender, and SKU bucket. Numerical features included order count, days since last order, symptom count, gut issue score, high stress indicator, refund count, quiz recommendation match, net LTV, average order value, and gross LTV. Categorical features were one-hot encoded, and numerical features were standardized prior to FAMD.

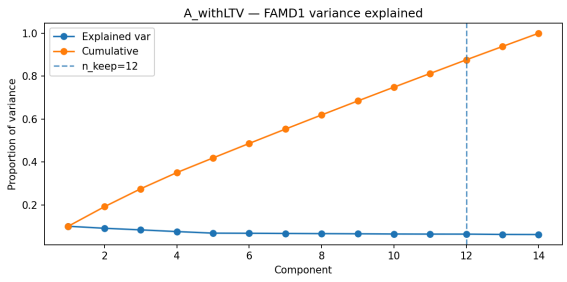
---

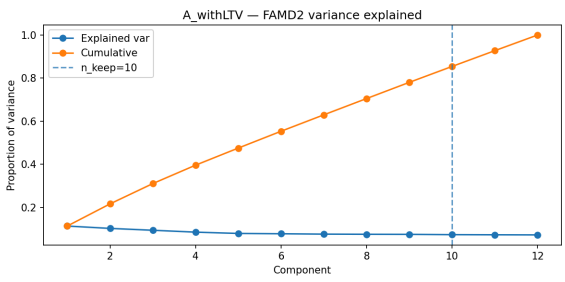
3) Methodology

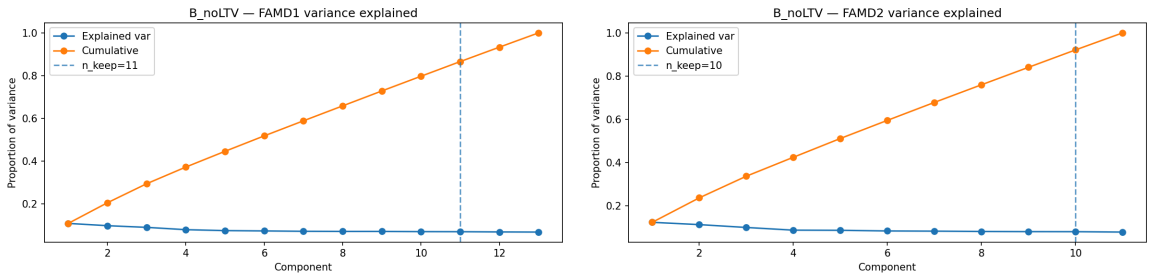
3.1) Dimensionality Reduction (FAMD)

FAMD was applied separately to Pass A and Pass B to reduce the input feature set before clustering. The retention rule preserved the smallest number of components needed to reach at least 85% cumulative variance explained. This ensured the majority of variability in the dataset was captured while reducing redundancy.

In Pass A, FAMD#1 retained 14 components (≈89.3% variance explained), and FAMD#2 retained 11 (≈86.6%). In Pass B, FAMD#1 retained 11 components (≈86.6%), and FAMD#2 retained 10 (≈92.1%). The cumulative variance curves showed a steady, nearly linear increase with no pronounced inflection point, so the decision was based purely on the variance threshold rather than visual elbow detection. This approach reduced noise and computational complexity while maintaining most of the signal.



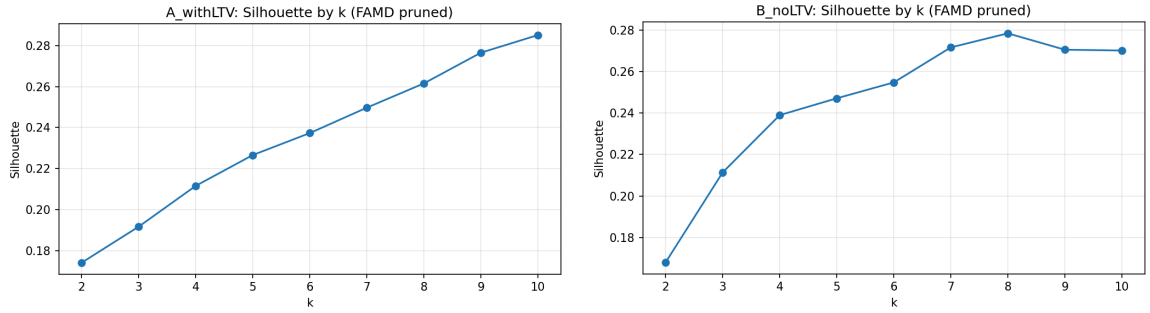




Variance explained and cumulative variance by FAMD component, pre- and post-pruning.

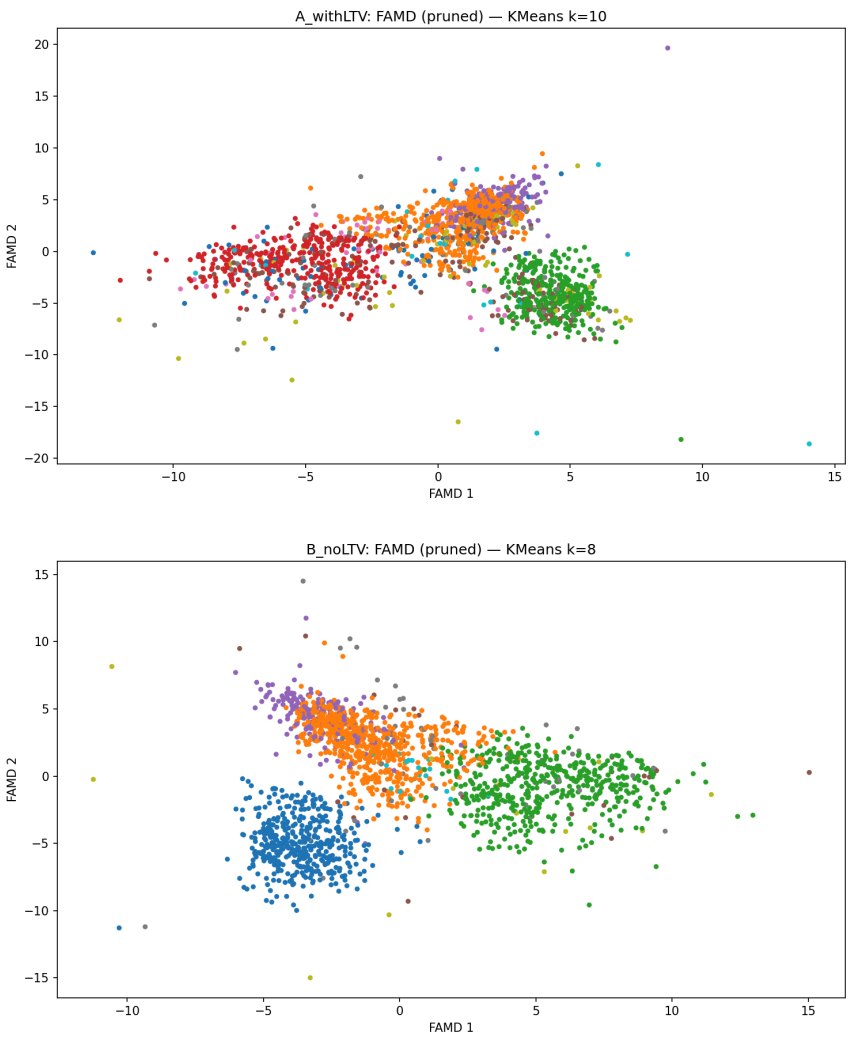
3.2) Clustering

K-means clustering was run for k values from 2 to 10 in each pass, with silhouette scores guiding the choice of optimal k in combination with interpretability and stability considerations. In Pass A, silhouette scores peaked at k=10 (0.2772), while in Pass B they peaked at k=8 (0.2785). Cluster stability, measured via ARI between seeds 42 and 123, was high in both passes (0.831 for Pass A and 0.880 for Pass B). The chosen k values provided a balance between statistical performance and interpretable segment definitions.



Silhouette score across k values for Pass A and Pass B.

Although silhouette scores peaked at k=10 (Pass A) and k=8 (Pass B), the absolute values (≈0.27–0.28) are modest. This indicates that while the clustering solutions are stable and interpretable, the separation between clusters is not extreme — consistent with the expectation that customer behavior exists along gradients rather than in sharply divided groups. This is evident in the below figures which show modest segmentation between groups.



Clustering of customers based on both approaches.

3.3) Post-Hoc Cluster Pruning

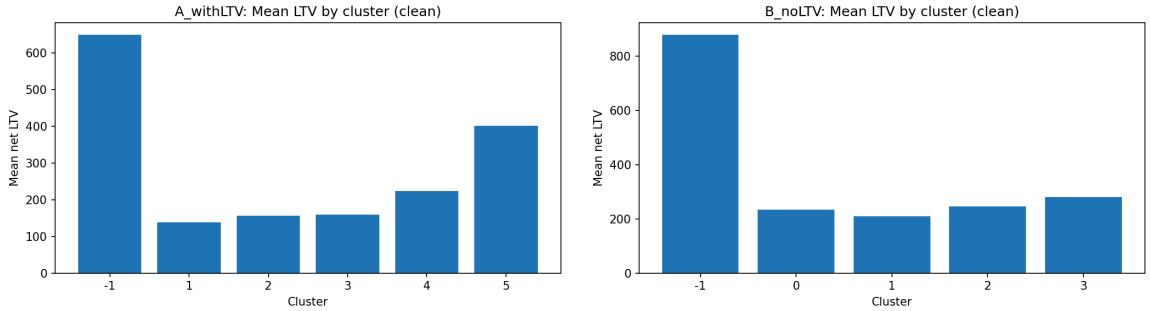
Clusters representing fewer than 5% of customers were relabeled as unclustered (-1). Profile metrics were recalculated excluding this group, but their size and key characteristics were still reported separately.

---

4) Results

4.1) Cluster Profiles & Personas

Cluster profiles were generated from cleaned outputs after pruning, summarizing behavioral and demographic patterns for each segment. Value overlays provided insight into LTV distribution by cluster, and qualitative personas were crafted from these profiles to support marketing strategy. Interestingly, in Pass B, the unclustered (-1) group had the highest mean LTV by a large margin, indicating that some of the most valuable customers did not align with the primary quiz-based behavior patterns. This finding suggests a potential high-impact opportunity for targeted retention strategies tailored to these outliers.



Mean LTV by cluster, post-pruning, including unclustered (-1) where applicable.

4.2) Stability & Replicability

Cluster stability was assessed by running the full clustering workflow twice per pass with different random seeds and calculating the Adjusted Rand Index (ARI) between the resulting label assignments. The ARI ranges from 0 (no agreement) to 1 (perfect agreement).

Both passes achieved high stability, with Pass A (`A_withLTV`) showing an ARI of 0.93 and Pass B (`B_noLTV`) an ARI of 0.88. This indicates that the clustering structure is largely robust to initialization differences, particularly when LTV is included as a feature. The slightly higher ARI for Pass A suggests that incorporating LTV contributes to more consistent cluster assignments across runs.

Run	ARI	Number of Clusters	Number of Samples
A_withLTV	0.93	10	1911
B_noLTV	0.88	8	1911

Cluster stability metrics for Pass A and Pass B, showing ARI, number of clusters selected, and total sample size.

Discussion

This two-stage segmentation approach, combining FAMD-based dimensionality reduction with k-means clustering, successfully produced stable and interpretable customer segments. The stability metrics (ARI = 0.83 for Pass A, 0.88 for Pass B) indicate that the clustering solutions are reproducible across different random seeds, suggesting robustness to initialization effects. The variance explained in the retained FAMD components (≈86–92%) confirms that the dimensionality reduction preserved the majority of variation in the original feature set, justifying the feature reduction process.

However, silhouette scores, while maximized at the chosen k values (0.277 for Pass A, 0.279 for Pass B), are modest in absolute terms. This implies that the boundaries between clusters are not sharply defined and that customer behaviors likely exist on a spectrum rather than in completely distinct groups. The scatter plots in 3.2 visually confirm this, with some overlap between clusters. This does not diminish the value of the segmentation — the high stability and coherent persona profiles indicate the clusters are still meaningful — but it suggests that some customers, particularly those near boundaries, may exhibit characteristics of multiple segments.

Another notable finding is that in Pass B (no LTV features), the unclustered (-1) group displayed the highest mean LTV by a significant margin. This is noteworthy because, by design, LTV was not used as a clustering feature in this pass, meaning these customers were grouped as outliers based solely on quiz responses and behavioral variables. Their divergence in both quiz profiles and LTV suggests that they behave differently from other customers in multiple dimensions. However, before drawing conclusions about potential opportunities, this warrants closer examination to ensure the pattern is not an artifact of data quality, outlier handling, or unaccounted-for variables. Investigating this group in more detail could clarify whether they represent a truly distinct, high-value customer segment or whether the separation reflects inconsistencies or gaps in the available data.

Overall, while the segmentation provides actionable insight, the modest silhouette scores point to an underlying complexity in the customer base. Future iterations could explore alternative clustering methods, feature engineering strategies, or nonlinear dimensionality reduction to enhance separation.

Appendix

Files included (Supplemental Data):

1. Cleaned cluster outputs (post-pruning)

`raw_data_v3.csv`

`A_withLTV_cluster_profiles_clean.csv`

`B_noLTV_cluster_profiles_clean.csv`

`A_withLTV_value_by_cluster_clean.csv`

`B_noLTV_value_by_cluster_clean.csv`

`A_withLTV_personas_clean.txt`

`B_noLTV_personas_clean.txt`

`A_withLTV_labels_pruned.csv`

`B_noLTV_labels_pruned.csv`