Identifying High-Value Customers & Product Opportunities on Olist

Leveraging data analytics to determine which customers to prioritise and what to sell

Presenter
Teo Hwee Sze | Data Analytics Capstone | General Assembly

What problem are we solving?

The Challenge of Scale for Olist

- **Many customers, many products**
- **No clear view of who / what drives revenue**
- **One-size-fits all approach = missed opportunity**

Core Strategic Questions

1. Who are our **most valuable & active** customers?
2. Which **product categories** drive the most revenue & are often bought together (affinity)?
3. How should we **segment** our customers for high-impact marketing and product bundling recommendations?

🎯 Project Goal:

Turn transactional data into a clear focus on customers to prioritise, determine what they buy, and how to use that to build actionable segments and product bundling opportunities.

Who is this project for?

Nathan,
Lead Product Manager

Role: Owns product strategy - assortment, pricing, bundling.

Needs & Goals: Identify categories and bundles that truly drive revenue and margin.

Frustration: No statistical evidence on which product relationships and cross-sell opportunities are actually strong.

How this project helps: market basket analysis framework to identify high-impact categories & bundles + dashboard

Hazel,
Marketing Manager

Role: Plans campaigns, CRM programmes, and acquisition activity.

Needs & Goals: Clear customer segments and product insights to target the right people with the right messaging.

Frustrations: No data-driven segmentation framework for marketing campaigns.

How this project helps: segmentation framework for personalised marketing campaigns + dashboard

Jim,
Head of Strategy (leadership)

Role: sets business strategy, revenue targets, and investment priorities.

Needs & Goals: consolidated view of customers & categories that drive growth to guide strategy

Frustrations: Siloed, inconsistent data, making it hard to decide where to invest.

How this project helps: consolidated, high-level insights on key segments and categories via dashboard

Approach & Data Source

The Data

Olist Brazilian E-Commerce Public Dataset (100+K records): Kaggle

Open-source dataset released by Olist:

The dataset contains real, anonymised transactional data from the Olist platform, covering:

I. Time period: 2016 - 2020

II. Geography: customers across multiple states in Brazil

III. Units of analysis:
a) Item-level - individual products within each order
b) Order-level - order status, total value, delivery dates
c) Customer-level - unique customer identifiers & their purchase history

Analytical Pipeline

1. Understand Features & Engineer Features

Explore basic patterns in customer spend, frequency, and recency | Clean and prepare RFM variables accordingly for clustering

2. Segment Customers

Test K-Means, Gaussian Mixture Models (GMM), Hierarchical Density-based Spatial Clustering (HDBSCAN)) | Compare cluster quality, then select K-Means with 3 clusters as the final, business-friendly segmentation

3. Analyse Products & Affinities

Rank product categories by revenue & orders to find the core assortment | Run Market Basket Analysis (Apiori & FP-Growth) at Category level to uncover top category-pairs overall & by segment

4. Dashboard Development

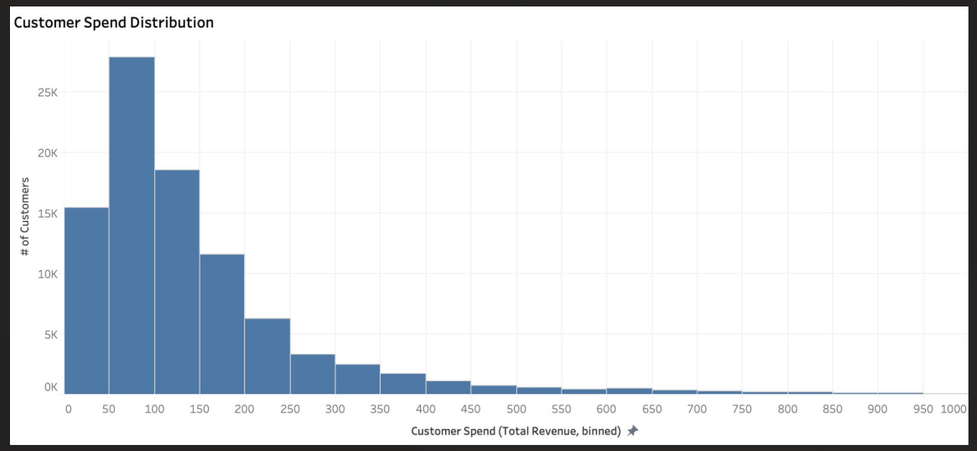
Create interactive Tableau dashboard that surfaces key customer segments & top product category insights.

Findings

Customer behaviour at a glance

Few power customers dominate revenue contribution

Customer Spend Distribution



Olist customers' spend distribution is heavily skewed, highlighting that most customers sit in the lower spend band, with only small number contributing to high spend.

Customer Spend, Frequency & Recency

Total Revenue (Customer Spend)

$15.4M

Average Orders Per Customer

1.03

Average Days Recency (Dormancy)

333

$15.4M in revenue, but from mostly one-off, dormant customers.

Customer Segmentation

3 behavioural segments as the recommended approach

K-Means clustering revealed three distinct, behaviour-based segments:

1. High-Value Active Buyers

~27% customers driving **~52% of revenue.**

- High avg instalments (~6.3 per order)
- Credit cards dominate

Description: **Big-ticket, credit-heavy customers;** small group with outsized financial impact.

2. Low-Value, Dormant Buyers

~32k customers driving **~23% of revenue.**

- Long average recency (~465 days).
- Boleto reliant

Description: **Low spend customers who are mostly inactive** - essentially a churn pool.

3. Active Value Buyers

~35.5k customers driving **~24% of revenue.**

- Shorter recency (~212 days),
- Card led but meaningful Boleto share.

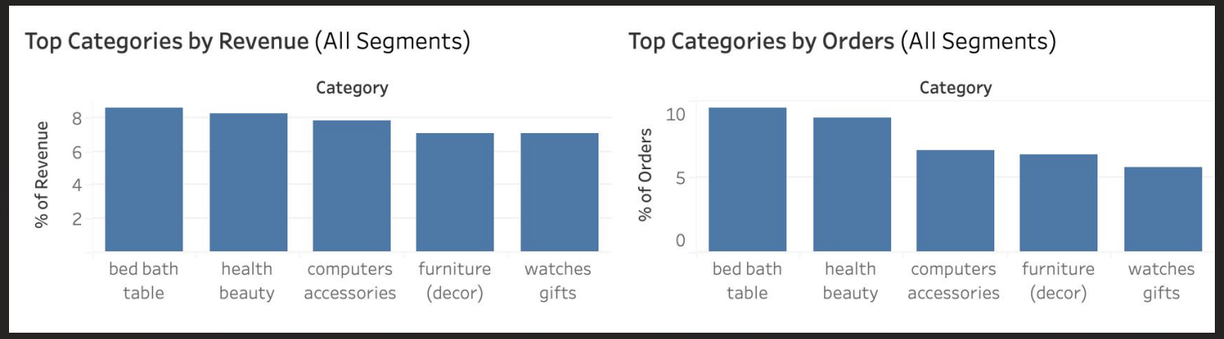
Description: **Frequent, value-oriented customers** with potential for basket-building and upselling.

Chosen clustering method: K-Means with 3 clusters was chosen as the final model: best trade-off between separation and compactness, and simple enough to apply in marketing.

Product Category Insights

A small set of categories drives most of the business

Olist's Core Assortment



Top categories

Home: bed bath table, furniture (decor)

Tech/Gifts: computers accessories, watches gifts

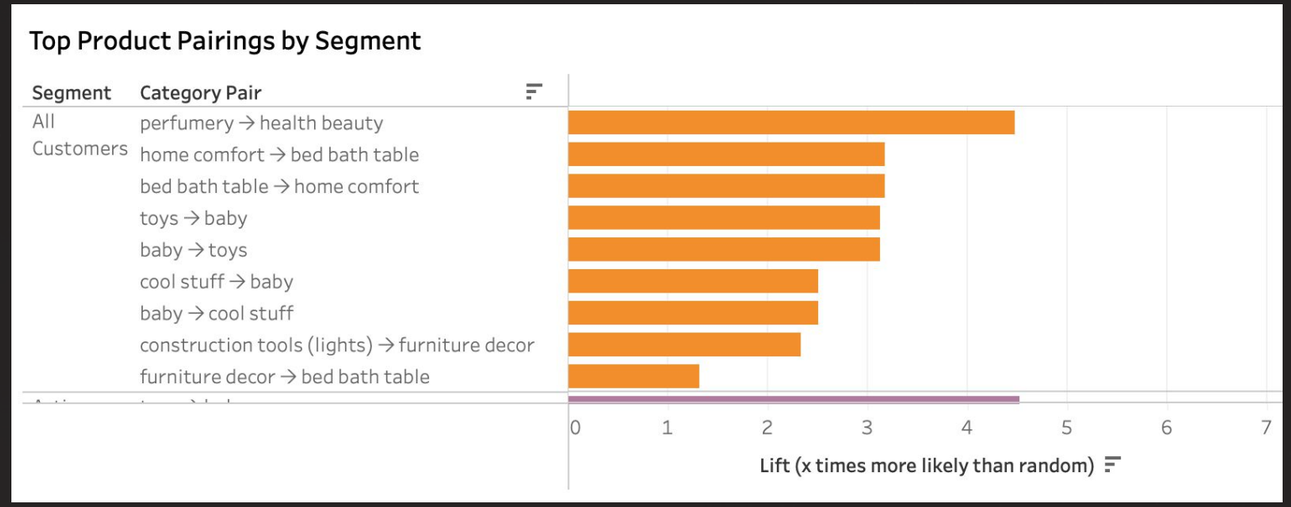
Personal Care: health beauty

Segment	Category Skew by Revenue	Implication
High-value Active Buyers	Premium home/tech gifting	Focus on **high AOV** cross-sells.
Low-value, Dormant Buyers	Hobby/leisure (sports leisure), but low spend volume.	Focus on **reactivation offers** tied to past interest.
Active Value Buyers:	Everyday self-care & home essentials (health beauty, bed bath table).	Focus on **frequency** and replenishment cycles.

Most revenue flows through a handful of home, beauty, and tech categories, but each segment leans into this core assortment in differently.

Product Affinities

What customers put together in the same basket



Market Basket Analysis (MBA) revealed that multi-category orders fall into 3 distinct customer "objectives," providing clear bundling opportunities.

1. Home Upgrade / Improvement

home comfort → bed bath table

High Confidence (>80%): If a customer buys home comfort, they almost always add a bed & bath item.

2. Family / Kids

baby ↔ toys / cool_stuff

Strong (lifts ~3.2×), symmetric relationships. Kids baskets frequently mix essentials with fun extras (e.g., a baby item plus a toy).

3. Beauty & Self-care

perfumery → health_beauty

High Lift (~4.5×): Perfume acts as a strong anchor into the wider beauty category.

How do product affinities differ by segment?

Same objective, different segment tactics

While all segments buy around *home, family, and beauty*, each one leans into these missions in a different way.

Segment	Strongest Affinities	Implication
High-value Active Buyers	Premium beauty cross-sell via perfumery and health beauty. There is also a pronounced "renovate then decorate" & home-comfort upgrading objectives.	**With high AOV** potential, focus on premium cross-sells (e.g., designer furniture, luxury beauty).
Low-value, Dormant Buyers	Sports leisure & health beauty in both directions, along with watches gifts and telephony.	Pairings indicate that segment tends to centre around hobby, lifestyle, and gifting. Focus on tying in MBA in **reactivation offers** tied to past purchases.
Active Value Buyers	Everyday home patterns with toys and baby. Home-upgrade behaviour is also evident through rules such as construction tools (lights) → furniture decor, home comfort → bed bath table, and art → furniture decor	Focus on building basket size with "home essentials" type of bundles

Key Takeaways

The big picture: strategic summary

The analysis provides four clear, actionable takeaways for Olist's Product, Marketing, and Leadership teams.

Revenue is Highly Concentrated

A small, but engagement-heavy group (High-Value Active) drives **>50%the total revenue.** This segment must be protected and given VIP service.

Growth Sits in the Mid-Value Segment

Active Value Buyers as a group is large, active & value-oriented. Ideal for focused marketing to increase **frequency and average basket value**.

Low-value Dormant Buyers is a Churn Pool

Treat this low-spend, dormant group as a **reactivation target** for targeted, highly-discounted offers, but do not rely on them for core revenue.

Products & Baskets Give Clear Plays

We have concrete **bundle ideas** derived from the home, family, and beauty purchase objectives, that can be directly implemented in checkout and marketing flows.

Recommendations, Data Limitations & Next Steps

Recommendations for working teams

Marketing & Product Moves (What to do now)

1. Protect the revenue engine (High-Value Active Buyers)

- Treat them as a VIP tier: priority service, instalment-led offers, and premium home + beauty bundles.

2. Grow mid-value segment (Active Value Buyers)

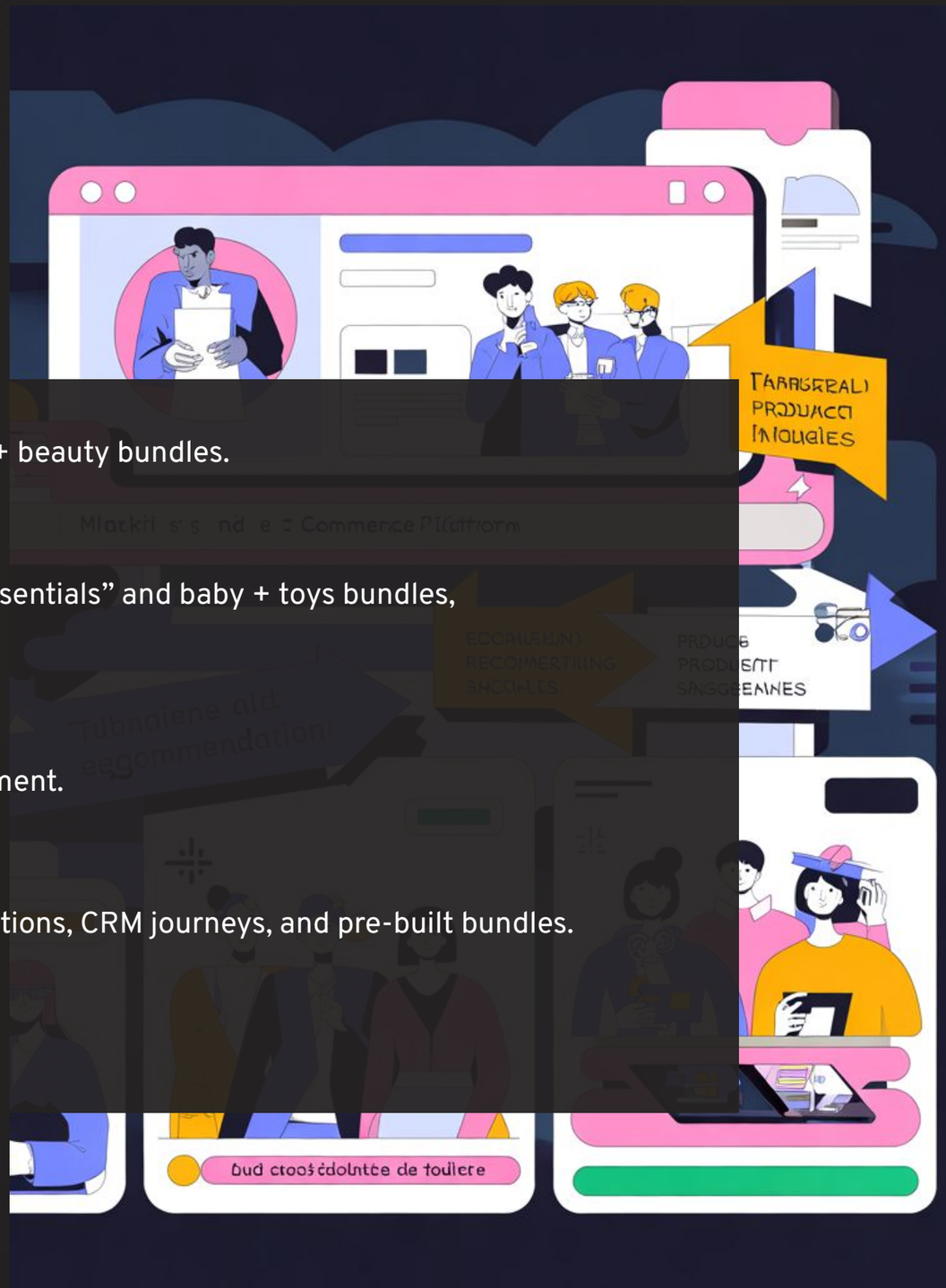
- Use them as the volume & growth engine: frequency-building campaigns, “home essentials” and baby + toys bundles, loyalty/replenishment offers.

3. Treat Low-value Dormant Buyers as reactivation pool

- However, as they will not contribute to revenue, avoid high investments on this segment.

4. Operationalise bundles & affinities

Embed key category rules (home upgrade, family/kids, beauty) into on-site recommendations, CRM journeys, and pre-built bundles.



Data Limitations & Next Steps

Key Data Limitations *(What this analysis can't yet do)*

1. Limited customer demographics

The dataset contains location data (city, state, ZIP prefix) but no information on age, gender, income, or household profile. As a result, segmentation is based on behavioural patterns, not personas

2. Limited product details

The dataset only contains product id and category name, not the name of the actual product itself. As a result, it limits learnings on market basket analysis on the product level.

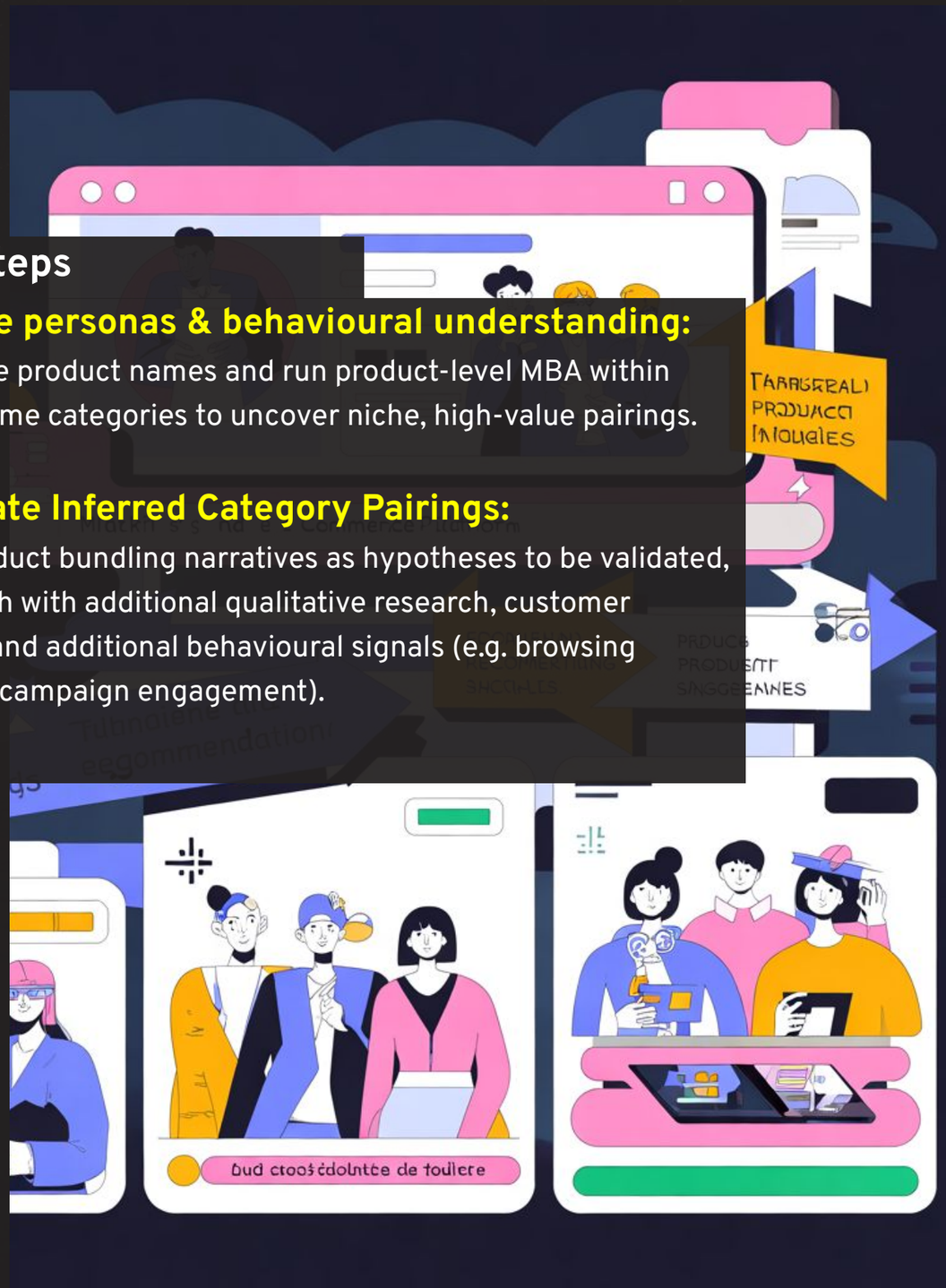
Next steps

1. Refine personas & behavioural understanding:

determine product names and run product-level MBA within high-volume categories to uncover niche, high-value pairings.

1. Validate Inferred Category Pairings:

Treat product bundling narratives as hypotheses to be validated, and enrich with additional qualitative research, customer surveys, and additional behavioural signals (e.g. browsing patterns, campaign engagement).



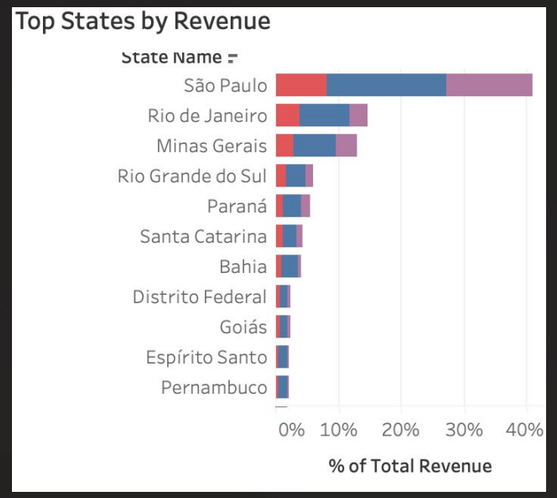
Olist Dashboard Demo

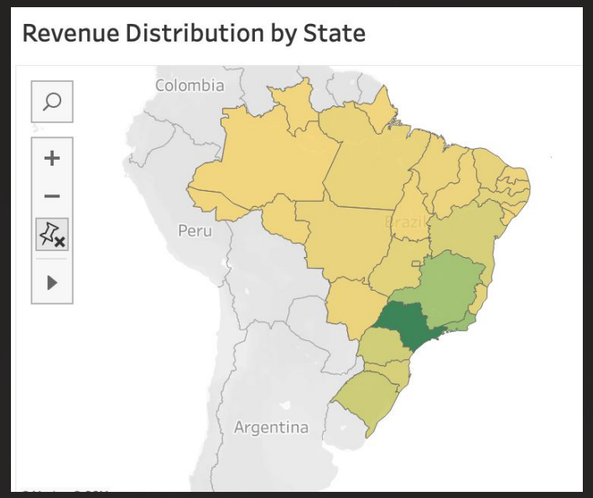
Q&A

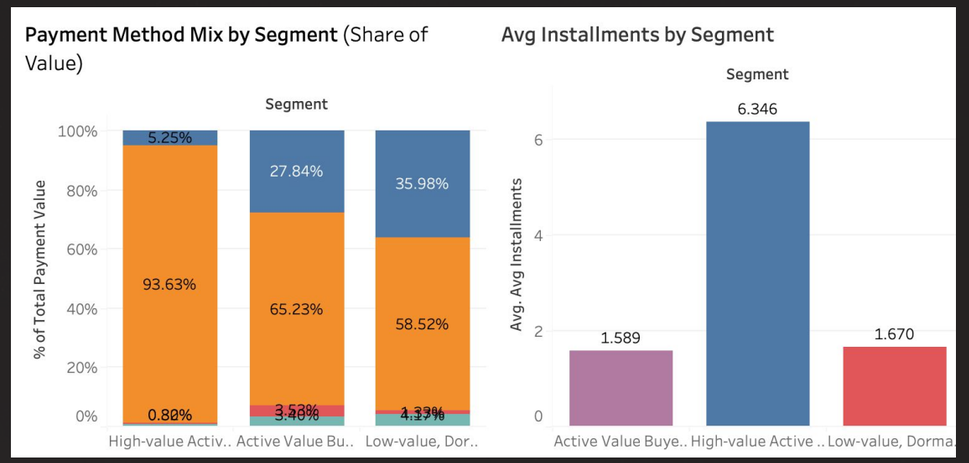
Appendix

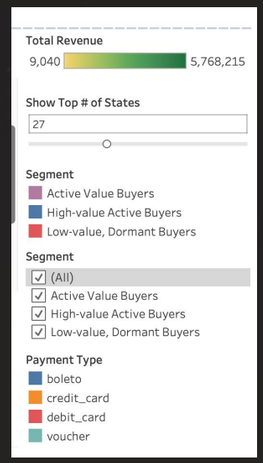
Segment Insights

Same regions, different payment behaviours









Geographic Consistency

The same 4 states - São Paulo, Rio de Janeiro, Minas Gerais and Rio Grande do Sul - account for ~⅔ of sales across **all three segments.**

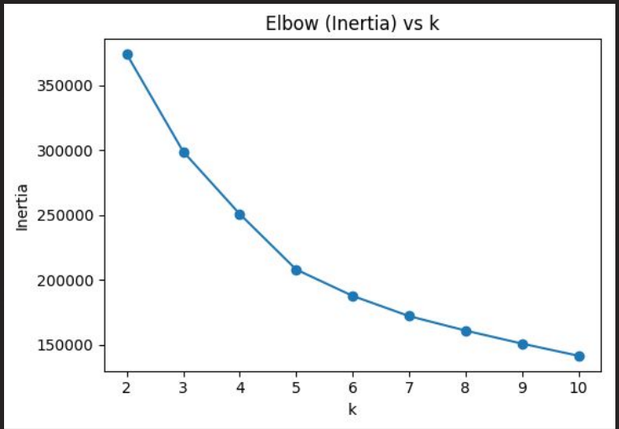
Pattern indicates that segmentation is primarily behavioural rather than regional; the same core states appear across all segments, but customers within those states behave differently.

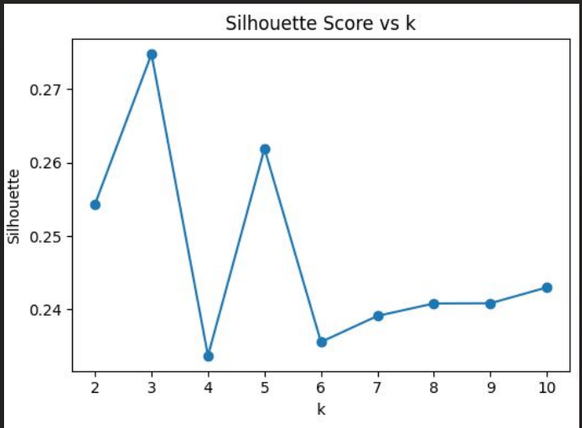
Payment Channel Split

Credit card is the dominant payment method, but its importance varies by segment:

- **High-value Active Buyers:** credit cards dominate, aligns closely with their profile as big-ticket, instalment-heavy profile.
- **The Low-value, Dormant Buyers:** Boleto accounts for a sizable proportion, making this the most boleto-reliant segment.
- **The Active Value Buyers:** sit in the middle - card led but meaningful Boleto share.

Clustering model





Selected Model and Cluster Quality

To select a segmentation model, **Elbow** (Inertia) **Method and Silhouette curves** for K-Means across different values of *k* (Figure 2) were employed.

I assess clustering performance of each model with quality scores:

A. Silhouette Score – Measures how well-separated the clusters are.
Benchmark: An acceptable silhouette score in real world application is a range between 0.2-0.5.

B. Average Intra-Cluster Variance – Captures how compact each cluster is.
Benchmark: This will be relative to the performance of all the models that I test in this project, but the model with the lowest average intra-cluster variance will be regarded as the most compact model.

Clustering model

Model	K-value	Silhouette Score	Avg Intra-Cluster Variance (Mean)	Max Cluster Variance
K-Means	3	0.209	3.559	4.351
K-Means	5	0.263	3.573	8.923
GMM	3	0.204	5.123	5.123
GMM	5	0.091	4.057	4.057
HDBSCAN	5	0.061	2.520	3.124

Table 1: Clustering Results

Selected K-Means with 3 clusters as the final segmentation model.

A 3-cluster solution also keeps the segmentation simple enough for marketing teams to understand and target, while still capturing meaningful differences in customer behaviour.

K-Means (k = 3) offers a good balance of cluster separation and compactness, with a competitive silhouette score and relatively low average intra-cluster variance.

K-Means (k = 5) achieves a slightly higher silhouette score but produces less balanced clusters with higher maximum variance, making the segmentation harder to use operationally.

GMM (k = 3 and k = 5) provides more flexible, overlapping clusters, but with lower silhouette scores and higher variance, reducing clarity between segments.

HDBSCAN finds compact clusters with a low average variance, but its very low silhouette score makes it less suitable as the primary segmentation for business use.

Market Basket Analysis

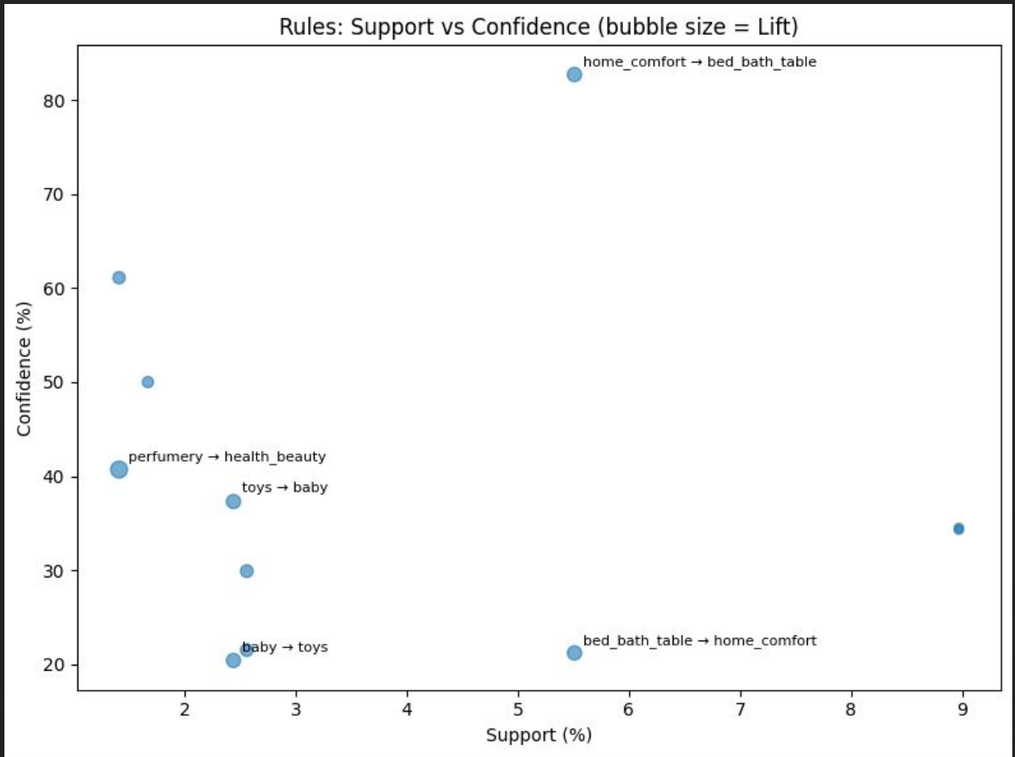
1. Modeling Approach & Key Association Rules
   After selecting category-level granularity, I generated association rules for category pairs using both the Apriori and FP-Growth algorithms:

   1. Apriori is treated as a conceptual baseline
   2. FP-Growth: offers a more scalable alternative on larger subsets of data.

   Because both algorithms operate on the same category-level baskets, the comparison focuses on runtime and the number of rules produced, rather than on accuracy.

   Rules were filtered using minimum thresholds on support, confidence, and lift (support > 2%, confidence > 50%, lift > 1.2) to obtain a compact, interpretable set suitable for cross-sell and bundling strategies.

Market Basket Analysis



I evaluate market basket analysis association rules for each product category pair using Support, Confidence, and Lift metrics:

A. Support – Measures how frequent a product combination appears. If support is too low, the pattern may be noise or overfitted.
Benchmark: >2%. Retail datasets often use 1–5% as a lower bound.

B. Confidence – measures how often an antecedent product “A” is bought with a consequent product “B”.
Benchmark: >50%. A score of more than 50% indicates a reasonably strong conditional probability that “If A is bought, B is bought at least 50% of the time.”

C. Lift – compares how likely products are brought together, compared to how often we’d expect it to be bought by chance. This is determined by dividing confidence score by support.
Benchmark: > 1.2. It is generally considered meaningfully above random chance, implying actionable co-purchase potential.