

Identifying High-value Ecommerce Segments & Product Opportunities in Olist e-commerce platform

**(General Assembly Capstone Project
Technical Report)**

Submitted to:

Waseem Sheriff
DSI Instructor

Benjamin Tan
Instructor

Report prepared by:

Teo Hwee Sze
General Assembly student

6 November, 2025

Table of Contents

1.0	Introduction & Problem Statement	4
2.0	Goals for this Project	5
2.1	Customer Profiling & Insights	5
2.2	Product Category Insights	6
2.3	Business Recommendations	6
3.0	Project Approach	7
3.1	Customer Segmentation	7
3.1.I	Cluster Quality Diagnostics	8
3.1.II	Apply Clustering	7
3.2	Segment Profiling & High-Value Analysis	8
3.3	Determining Product Affinities via MBA	8
4.0	Criteria for Success	9
4.1	Analytical Performance Metrics	9
4.1.I	Customer Segmentation Quality	9
4.1.II	Product Affinity Strength	9
4.2	Business Relevance	10
5.0	Intended Audience	11
5.1	Product Managers	11
5.2	Marketing Teams	11
5.3	Leadership	11
5.4	Data Analysts & Data Scientists	12
6.0	Data Source	13
6.1	Origin and Scope	13
6.2	Data Quality and Limitations	13
6.3	Key Tables Used	14
7.0	Feature Engineering	16
7.1	Customer-Level Behavioural Features	16
7.2	Category-Level Performance Features	18
7.3	Rule-Level Features for Market Basket Analysis	19
8.0	Patterns, Trends, and Insights	22
8.1	Customer Behaviour Overview	22
8.1.I	Distribution of Spend, Frequency, and Recency	22

8.2 Customer Segmentation Results	23
8.2.I Selected Model and Cluster Quality	23
8.2.II Segment Profiles: Behavioural Summary	25
8.2.III Geographical Distribution by Segment	28
8.2.IV Payment Behaviour by Segment	29
8.3 Product Category Performance	27
8.3.I Top Categories by Revenue and Orders	30
8.3.II Category Mix by Segment	31
8.4 Product Affinities (Market Basket Analysis)	32
8.4.I Determining Basket Structure and Granularity	32
8.4.II Modeling Approach & Key Association Rules	33
8.4.III Segment-Specific Product Affinities	35
9.0 Conclusion: What These Patterns Mean for the Business	39
9.1 Recommendations for Overall Marketing & Product Strategy	39
9.2 Next Steps and Further Analysis	41
10.0 Appendix	41
10.1 Data Dictionary	41

1.0 Introduction & Problem Statement

Olist is a Brazilian e-commerce platform that helps small and medium-sized businesses reach customers across the country. It operates as an online marketplace where merchants list their products and services, and customers can browse, compare, and purchase them directly through the platform.

For a business like Olist, a core challenge is building a clear, data-driven view of customer behaviour. Specifically, identifying high-value segments and uncovering their needs and preferences across both the Olist brand and its network of sellers.

A second challenge lies in defining the right product mix for each customer segment to maximise marketing return on investment (ROI) across engagement, remarketing, and acquisition campaigns. In practice, this means surfacing high-affinity products alongside both owned and paid media in ways that are most likely to resonate with each customer.

Amid rising advertising competition within the Brazilian market¹, the brand must leverage data analytics and machine learning to recommend the right products to the right audiences in the right campaigns. Doing so can improve targeting precision, strengthen customer engagement, and enhance overall marketing ROI.

This project explores these challenges, with a focus on customer segmentation and product category associations that can directly inform marketing and product strategies.

¹ Shubham. (2025, October). *Brazil online advertising and digital media market*. Ken Research. <https://www.kenresearch.com/brazil-online-advertising-and-digital-media-market> Ken Research

2.0 Goals for this Project

This project is designed to answer priority business and marketing questions for the Olist marketplace across three themes: customer profiling, product category insights, and actionable recommendations.

2.1 Customer profiling & insights

- I. What is the distribution of customers by:
 - A. total spend,
 - B. order frequency, and
 - C. recency (days since last purchase)?
- II. Who are our top customers contributing the most to total revenue, and what percentage of total revenue do they account for?
- III. How should we segment our customer base for marketing and business purposes?
- IV. Based the recommended segmentation, what is the:
 - A. overall revenue share,
 - B. average order frequency, and
 - C. average recency per segment?
- V. Which states contribute the highest sales and number of customers? How are these states distributed across segments?
- VI. What are the most common payment methods and how do payment preferences vary by segment?

2.2 Product category insights

- VII. What are the top 10 product categories by revenue and orders?
- VIII. What share of total revenue is driven by the top 10 categories?
- IX. What share of total orders is driven by the top 10 categories?
- X. How do category revenue share and order share differ by segment?
- XI. Applying market basket analysis:
 - A. Which categories are most frequently purchased together?
 - B. Are there strong associations between specific categories (e.g., electronics → accessories)?

2.3 Business Recommendations

- XII. What should be the overall marketing and product strategy for Olist based on the findings?
- XIII. What segment-specific strategies should be recommended for:
 - A. Targeting and Acquisition,
 - B. Cross-sell and Upsell, and
 - C. Product/Category prioritisation?

3.0 Project Approach

This project applies clustering, statistical analysis, and association rule mining to uncover customer and product insights that directly address the business questions outlined above.

3.1 Customer Segmentation

First, customer profiling and clustering is to identify distinct, interpretable segments.

I. Cluster quality diagnostics

- A. Compute inertia for a range of k values to assess how well data points fit within clusters.
- B. Compute silhouette scores for each k to compare how well-separated clusters are.

II. Apply Clustering

Clustering is applied to identify distinct customer groups based on purchase frequency, spending behaviour, and product preferences.

Here, three clustering models are tested to determine the best number of clusters and model for our segmentation for marketing & business insights:

- A. K-means clustering: Groups customers into k clusters by minimising the distance between each data point and its cluster centroid.
Model assumption: clusters are roughly spherical and similar in size.
- B. Gaussian Mixture Models (GMM): Models the data as a mixture of several Gaussian (bell-shaped) distributions, each with its own mean and covariance. Performs soft clustering, assigning each customer a probability of belonging to each cluster.
Assumption: clusters can be elliptical and overlapping, reflecting the reality that customers may share traits across segments.
- C. Hierarchical Density-based Spatial Clustering (HDBSCAN): A density-based method that identifies clusters based on how closely data points are packed together and automatically detects clusters of varying density and labels points that don't belong clearly to any group as noise.

The model and numbers of clusters with the best trade-off between separation, compactness, and business interpretability will be selected for segmentation analysis.

Note: Feature engineering and rationale behind derived variables used for clustering (e.g. total spend, frequency, recency, delivery time, payment behaviour) are detailed in a later section of this report.

3.2 Segment Profiling & High-Value Analysis

After selecting the final clustering model, statistical analyses were conducted to profile each customer segment:

- III. **Identify high-value segments and products:** Use statistical analyses to deepen the profiling of our proposed customer segments. Here, insights were determined to answer the key business and marketing questions related to our customer segments. In this section, top product categories that contribute most significantly to overall revenue and order frequency, per segment were determined.

3.3 Determining Product Affinities via Market Basket Analysis (MBA)

Steps to determining product category relationships using MBA:

- IV. **Determine basket granularity:** Run basket level distribution to assess the level of granularity to use.
- V. **Association Rule Mining:** Conduct MBA with the recommended level of granularity to identify product categories that are frequently purchased together within each segment.
- VI. **Share business and marketing recommendations:** Use statistical analyses to deepen our business strategy across marketing and product initiatives. Here, I also uncover insights to help answer the key business and marketing questions related to product categories too.

4.0 Criteria for Success

The success of this project is evaluated using a mix of analytical performance metrics and business relevance criteria.

4.1 Analytical Performance Metrics

I. Customer Segmentation Quality

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.

II. Product Affinity Strength

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.

Business Relevance

Beyond model metrics, the project is considered successful if it delivers:

- I. **Actionable Segmentation:** Customer clusters should reveal clear behavioural differences that can inform marketing and product strategies.
- II. **Cross-Sell Opportunities Identified:** Association rules should highlight meaningful product pairings or bundles that can increase average order value.
- III. **Marketing ROI Improvement:** Insights that can be operationalised via interactive dashboards to help marketers:
 - A. target the right segments,
 - B. feature the right product mix, and
 - C. design more efficient campaigns that have potential to improve conversion and engagement.

5.0 Intended Audience

This project is designed with multiple stakeholders in a given e-commerce organisation in mind:

5.1 Product managers

- I. Background: Responsible for defining and evolving the product and category strategy, including assortment, bundling, and prioritisation of product lines.
- II. Needs and Goals: Understand which product categories and combinations drive revenue, and identify bundling and cross-sell opportunities that can enhance basket size and margin.
- III. Frustrations: Limited visibility into which product relationships are actually backed by data. Difficulty prioritising categories without a clear, data-driven framework.
- IV. How this project helps: Provides an interactive dashboard and statistically grounded Market Basket Analysis to identify high-impact categories and bundles, helping to inform the product roadmap and prioritise high-performing categories.

5.2 Marketing teams

- I. Background: Plan and execute campaigns, CRM programmes, and personalised communications to acquire, retain, and grow customers.
- II. Needs and Goals: Clear customer segments with distinct behaviours and value profiles, as well as insights into which products to promote to which segments.
- III. Frustrations: Broad, undifferentiated campaigns that yield low ROI. Limited ability to personalise messaging due to lack of structured segmentation.
- IV. How this project helps: Delivers an interactive dashboard and segmentation framework to design targeted campaigns, personalised recommendations, and more efficient marketing spend.

5.3 Leadership

- I. Background: Set overall business strategy, revenue targets, and investment priorities.

- II. Needs and Goals: Understand which customer segments and product categories drive the most revenue and growth, so that they can align strategic focus and resources to the most attractive opportunities.
- III. Frustrations: Lack of consolidated, data-driven view of customers and products, leading to fragmented insights across teams.
- IV. How this project helps: Provides consolidated, high-level insights via dashboards and summaries on key segments and categories, enabling better strategic decisions on focus areas and resource allocation.

5.4 Data Analysts & Data Scientists:

- I. Background: Build and maintain analytical models, dashboards, and data pipelines to support decision-making.
- II. Needs and Goals: A reproducible, well-documented workflow for customer segmentation and Market Basket Analysis. Clarity on feature engineering choices, model selection, and evaluation metrics.
- III. Frustrations: Lack of shared standards for segmentation and product affinity analysis.
- IV. How this project helps: Offers a documented, reusable workflow (including feature engineering, clustering, and MBA) that can be applied to ongoing customer behaviour monitoring and future segmentation projects.

6.0 Data Source

6.1 Origin and Scope

This project uses the Brazilian E-Commerce Public Dataset by Olist, an open-source dataset released by Olist, a Brazilian online marketplace, and hosted on Kaggle².

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

- I. Time period: 2016 to 2020
- II. Geography: customers and sellers across multiple states in Brazil
- III. Units of analysis:
 - A. Item-level (e.g. product lines within each order)
 - B. Order-level (e.g. status, total value)
 - C. Customer-level (e.g. unique customer identifiers and location)

The dataset was chosen as it has the appropriate units of analysis suitable for both customer segmentation and product-level Market Basket Analysis and a large base of >100+K records.

6.2 Data Quality and Limitations

While the dataset is rich and suitable for segmentation and Market Basket Analysis, there are several important limitations:

I. 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

II. 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.

III. Product categories and language

Original product categories are in Portuguese, and can be ambiguous or overlapping. The translation file was used to map category names back to English, but there may still be category noise.

IV. Selection bias

² https://www.kaggle.com/datasets/olistbr/brazilian-e-commerce?select=olist_customers_dataset.csv

The data represents customers who shop via the Olist marketplace, and does not include offline purchases or other channels. Insights are therefore limited to this online context.

V. Cancelled and non-delivered orders

Orders in cancelled or undelivered status are present in the raw data, which were excluded when analysing realised revenue and building the models for clustering and market basket analysis.

VI. Multiple payments per order

Some orders have more than one payment record (e.g. part credit card, part voucher). Payment data was aggregated at the order level to avoid double-counting and to correctly capture total order value and instalment behaviour.*

Basic preprocessing (such as handling duplicates, filtering cancelled orders, and aligning date/time features) is applied before analysis.

6.3 Key Tables Used

For this project, I focus on six core tables from the Olist dataset:

I. **olist_orders_dataset**

Defines each order and its key timestamps, including purchase date, approval date, shipping date, delivery date, and estimated delivery date. This table is the core dataset for order-level analysis.

II. **olist_customers_dataset**

Contains customer identifiers and location fields such as state, and ZIP code prefix. This table enables customer-level aggregation and geographical analysis.

III. **olist_order_payments_dataset**

Stores payment information for each order, including payment method, number of instalments, and payment value. Some orders have multiple payment records (e.g. split between credit card and voucher), which is important for understanding payment behaviour.

IV. **olist_order_items_dataset**

Provides item-level detail for each order, including product ID, seller ID, price, freight value, and shipping limit date. This table is the basis for basket-level

analysis.

V. **olist_products_dataset**

Contains product-level attributes such as product category, dimensions, weight, and text length features (e.g. product name length). It is used to map products to product categories and to enrich item-level analysis.

VI. **product_category_name_translation.csv**

Provides translations from Portuguese product category names to English. This is used to create an English-friendly category field for analysis and visualisation.

7.0 Feature Engineering

To support customer segmentation and Market Basket Analysis, I engineered a set of features at three main levels:

- I. **Customer-level behavioural features** for clustering and profiling
- II. **Category-level performance features** for revenue and volume analysis
- III. **Rule-level metrics** for Market Basket Analysis

7.1 Customer level behavioural features

Customer-level features are aggregated at the `customer_unique_id` level to build a behavioural profile for each customer. These features are inspired by an **RFM-style framework** (Recency, Frequency, Monetary), extended with **delivery experience** and **payment behaviour**.

Feature name	Level	Definition	Rationale
<code>order_total</code>	Order	Total amount paid per <code>order_id</code> , summed from all payment records.	Provides monetary value of each order, even when payments are split.
<code>total_revenue</code>	Customer / Segment	Sum of <code>order_total</code> across all completed orders for a customer or group.	Used to identify high-value customers and revenue contributions by segment.
<code>n_orders / order_frequency</code>	Customer	Number of unique completed orders per customer (<code>nunique(order_id)</code>).	Captures purchase frequency, a core dimension for segmentation.

recency_days	Customer	Number of days between the customer's most recently completed order and the analysis reference date.	Measures how recently a customer has purchased (active vs lapsed).
delivery_time_days	Order	Days between purchase and delivery (<code>order_delivered_customer_date - order_purchase_timestamp</code>).	Captures the delivery experience at the order level.
avg_delivery_days	Customer	Average <code>delivery_time_days</code> across all completed orders for a customer.	Allows analysis of whether delivery speed is linked to behaviour or churn.
avg_order_total	Customer	Average <code>order_total</code> across completed orders.	Distinguishes customers who place few large orders vs many small orders.
order_total_log	Order	Log-transformed order value (e.g. <code>log(order_total + 1)</code>).	Reduces skew from very large orders, making features more suitable for clustering.
avg_order_total_log	Customer	Average of <code>order_total_log</code> per customer.	Stabilises variance while preserving differences between low- and high-spend customers.
order_installments	Order	Maximum number of payment installments used for each <code>order_id</code> .	Reflects credit usage for a given order when multiple payment records exist.

avg_installments	Customer	Average <code>order_installments</code> across a customer's orders.	Captures payment behaviour and reliance on instalments over time.
avg_installments_log	Customer	Log-transformed version of <code>avg_installments</code> .	Handles skew where a small subset of customers use very high installments.

7.2 Category-level performance features

At the **product category** level, I derive features from the `order_products` dataframe to understand which categories drive revenue and volume, and how they are priced and shipped.

Feature name	Level	Definition	Rationale
<code>category_final</code>	Item / Category	English product category name, derived from the translation table.	Makes categories interpretable for stakeholders and dashboards.
<code>category_canonical</code>	Category	Cleaned / consolidated version of the category field (e.g. merged rare or noisy categories).	Ensures categories have sufficient volume and are stable enough for analysis and MBA.
<code>revenue_share_pct</code>	Category	Percentage of total revenue contributed by each category.	Identifies top revenue-driving categories and their relative importance.

n_orders (category)	Category	Number of unique orders that contained at least one item from the category.	Shows how widely the category appears across baskets .
n_items (category)	Category	Total number of items sold from the category.	Highlights high-volume categories, even if unit price is low.
avg_price	Category	Average item price per category.	Positions categories as premium vs mass and informs pricing and promotion strategy.
avg_freight	Category	Average freight (shipping) cost per item per category.	Provides insight into logistics cost and potential margin considerations.

7.3 Rule-Level Features for Market Basket Analysis

For Market Basket Analysis, I treat each **order** as a basket of product categories and generate association rules of the form:

IF basket contains $A \rightarrow \text{THEN}$ it is likely to also contain B

For each rule, I derive standard association metrics:

Feature name	Level	Definition	Rationale
Antecedent	Rule	The “ IF ” side of the rule (category or category set A).	Identifies the trigger category or combination.

Consequent	Rule	The “THEN” side of the rule (category or category set B).	Identifies the recommended / associated category.
support_count	Rule	Number of orders that contain both A and B together.	Ensures rules are based on a meaningful absolute volume of transactions.
support_pct	Rule	support_count divided by total number of orders.	Measures how frequent the combination is across all transactions.
conf_pct (confidence)	Rule	Proportion of orders containing A that also contain B ($P(A \text{ and } B) / P(A)$)	Used to measure how reliable a rule is for cross-sell.
lift	Rule	Ratio of confidence to the baseline probability of B ($P(A \text{ and } B) / P(B)$)	Verifies if the association between A and B is stronger than random chance.
ante_support_pct	Rule	Percentage of orders that contain the antecedent A.	Puts confidence into context by showing how common A is on its own.
cons_support_pct	Rule	Percentage of orders that contain the consequent B.	Shows how popular B is independently of A.

These metrics are used to **prioritise rules** that are both:

- I. **Statistically robust** (sufficient support, lift above random), and
- II. **Commercially meaningful** (involving important categories with material revenue or volume).

These engineered features form the basis for the clustering and Market Basket Analysis techniques, and are later surfaced through dashboards for business stakeholders. Full data dictionary in appendix section.

8.0 Patterns, Trends and Insights

This section presents the key patterns, trends, and insights uncovered from my analyses. I first explore overall customer behaviour, then examine segment-level differences and product category performance, and finally highlight product affinities identified through Market Basket Analysis. These insights directly address the business questions outlined in Section 2.0.

8.1 Customer Behaviour Overview

Distribution of Spend, Frequency & Recency

To understand overall customer behaviour, I analysed how total spend, order frequency, and recency (days since last purchase) are distributed across the customer base, and how they differ by segment (Figure 1/Dashboard).

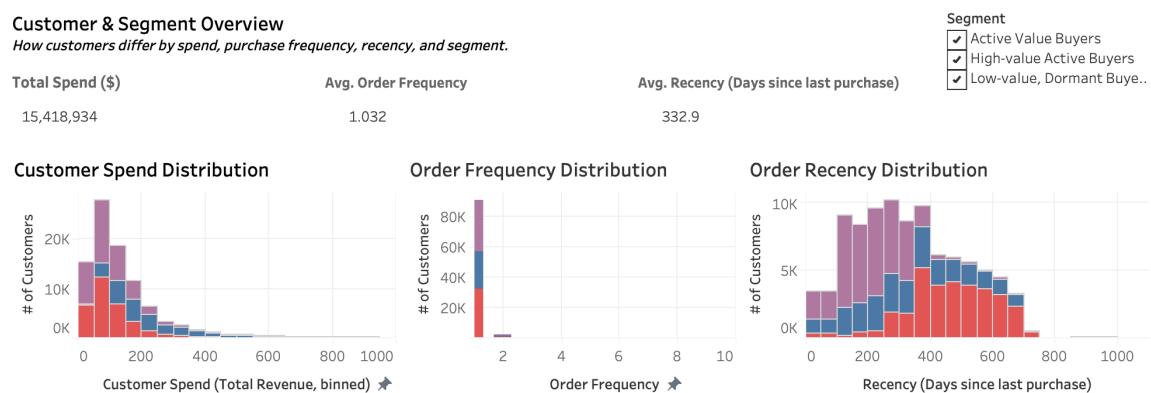


Figure 1/Dashboard: Customer Spend, Frequency, and Recency

Dashboard summarizing total customer spend, average order frequency, and days since last purchase, broken down by customer segment. Refer to Tableau file to view interactive dashboard.

Key Takeaways from Figure 1/Dashboard:

A. Customer spend is heavily skewed.

Across all customers, total online spend on Olist is about **\$15.4M**. The “Customer Spend Distribution” chart shows a classic **right-skewed** pattern: most customers sit in the lower spend bands, with a small number contributing much higher spend. This confirms that a relatively small group of customers drives a large share of revenue.

B. Most customers purchase only once.

From the “Order Frequency Distribution” chart, the **average order frequency is 1.03 orders per customer**, and almost all customers appear in the 1-order bar. Only a small fraction place two or more orders.

This indicates that Olist currently has a large base of one-time buyers, with significant potential to grow repeat purchases.

C.

Recency indicates a large dormant base.

The “Order Recency Distribution” shows that the **average recency is around 333 days** (roughly one year since last purchase). Overall, these patterns suggest that while Olist has a small but important group of high-spend, recently active customers, the majority of the customer base is infrequent and dormant.

Together, these distributions show that while Olist has a small but important group of high-spend, active customers, the majority of the customer base is **low-frequency and relatively inactive**. This reinforces the need for targeted retention and reactivation strategies, which are explored further in the segmentation and recommendation sections that follow.

8.2 Customer Segmentation Results

I. 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.

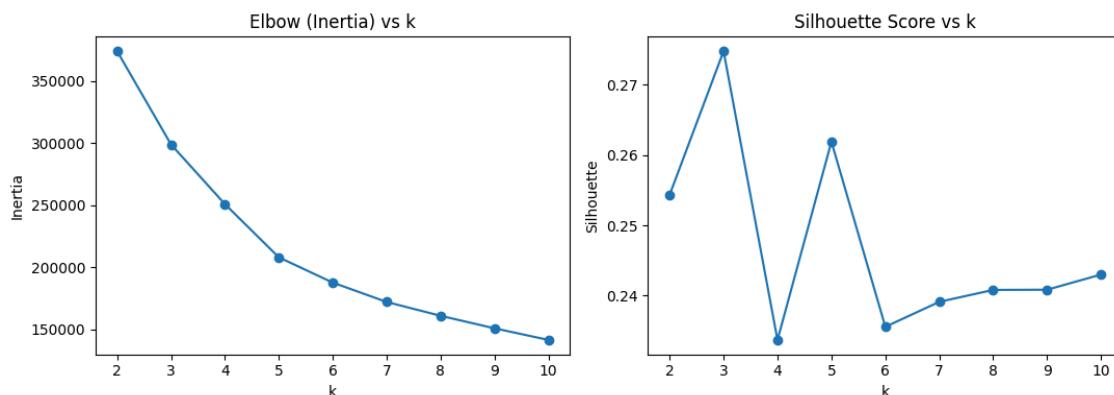


Figure 2: Elbow (Inertia) and Silhouette Score Across k for K-Means

Diagnostic plots used to narrow the choice of k ; inertia flattens after $k \approx 4-5$, while silhouette peaks at $k = 3$, guiding the selection of a 3-cluster solution for further evaluation.

The **Elbow curve** showed a steep drop in inertia from **$k = 2$ to 5** , after which the curve started to flatten, indicating diminishing returns in compactness beyond **$k \approx 4-5$** . The **Silhouette scores** peaked around **$k = 3$** , with **$k = 5$** also performing reasonably well.

Based on this, k-value was narrowed down to **k = 3 and k = 5** and compared three clustering methods: **K-Means Clustering, GMM, and HDBSCAN** (Table 1).

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

Key Findings

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

Taking both the statistical diagnostics into account, I 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.

II. Segment Profiles: Behavioural Summary

To understand how the three clusters differ in practice, I summarised **customer count, revenue contribution, average spend, order frequency, recency, and instalment behaviour** for each segment (Figures 3 and 4).



Figure 3: Customer Spend, Frequency, and Recency

Dashboard summarizing total customer spend, average order frequency, and days since last purchase, broken down by customer segment. Refer to Tableau file to view interactive dashboard.

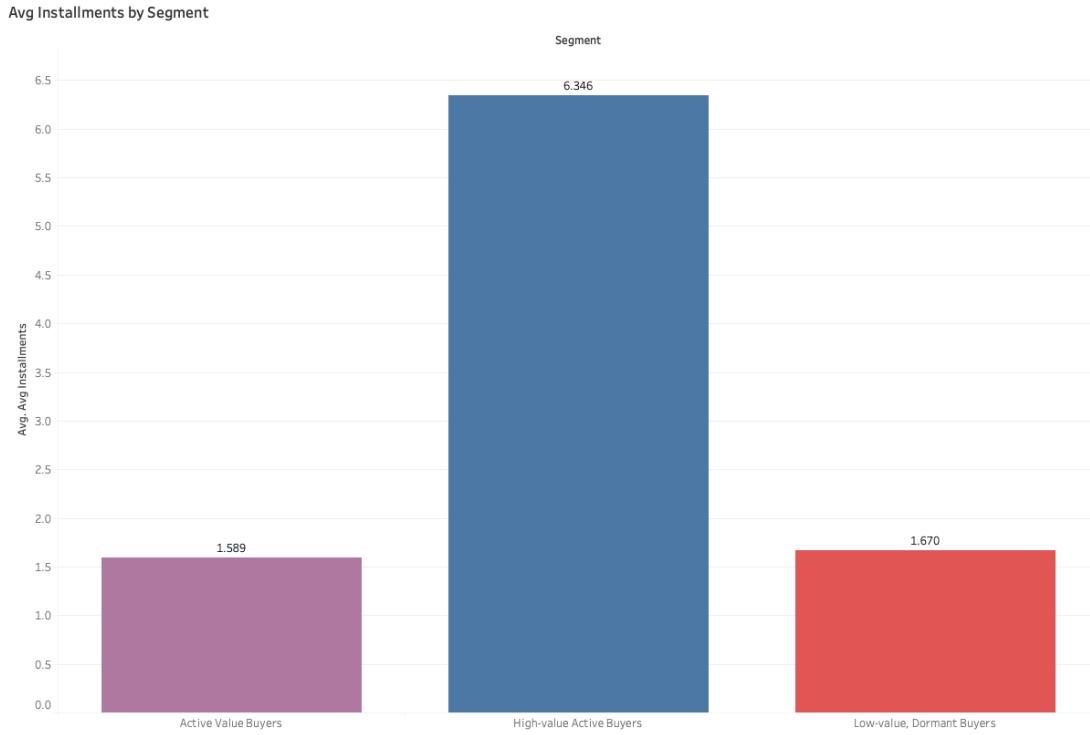


Figure 4: Average Installments Per Cluster

Bar Chart summarizing average number of installments per customer, broken down by customer segment.

Key Findings

- A. **The first cluster** (represented by blue bars) is the smallest in size but the most important commercially. It contains roughly 25.5k customers, around 27% of the base, yet generates approximately 52% of total customer revenue.

Average revenue per customer is about \$315, nearly three times that of the other segments.

Customers in this group use an average of around 6.3 instalments per order, accounting for almost 60% of all instalments in the dataset.

This segment is therefore characterised and defined as: High-value Active Buyers - a small group of big-ticket, credit-dependent buyers that underpins platform revenue.

- B. **The second cluster** (represented by red bars) is larger in size but much weaker in monetisation. It includes around 32.4k customers and contributes approximately 23% of total revenue.

Average revenue per customer is around \$110, and installment usage is modest, at roughly 1.7 instalments per order.

This group also has the highest average recency (around 465 days since last purchase), indicating that many of these customers have not transacted for a long time.

Overall, **this segment is therefore characterised and defined as:**

Low-value Dormant Buyers - consisting of customers with low spend and low recent activity.

- C. **The third cluster** (represented by purple bars) is the largest group with approximately 35.5k customers. It generates about **24.7% of the total revenue**, with average revenue per customer of roughly **\$107**, similar to Cluster 1.

However, behaviourally it is quite different: this segment records the lowest average recency (~212 days), indicating more recent purchases and higher ongoing engagement.

Installment usage is moderate at around 1.6 instalments per order. Inferring the results, these customers tend to make smaller, more frequent, value-oriented purchases, with light to moderate reliance on credit.

As such, this segment can be characterized & defined as: Active Value Buyers - representing a broad base of engaged but mid-value customers with clear potential for growth.

- D. Taken together, the three segments reveal a clear structure: a small, high-value, instalment-heavy core; a large but low-value dormant base; and a large, mid-value but active group.

III. Geographical Distribution by Segment

To assess whether the segments were driven by geography or by behaviour, I analysed the distribution of revenue by state and then examined how each segment is represented across those states.

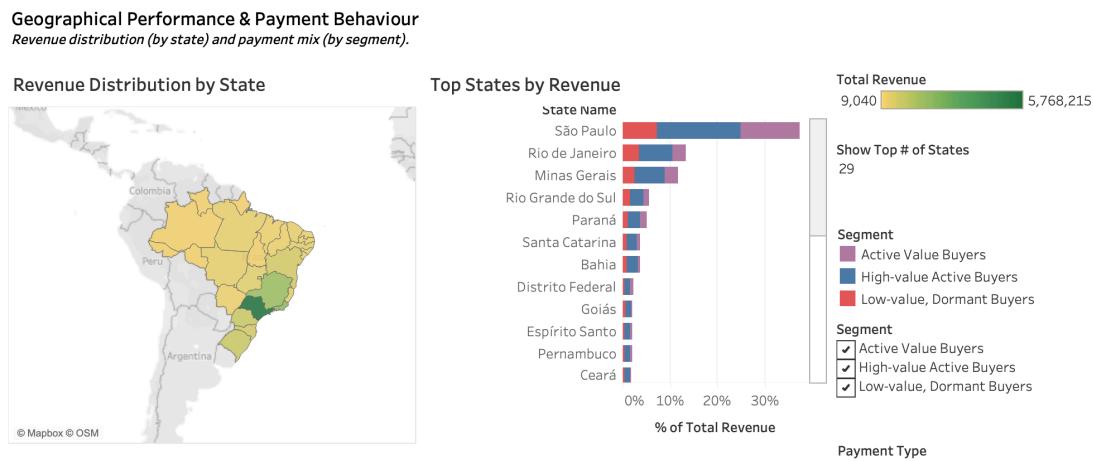


Figure 5: Geographical Performance by Segment

Dashboard showing revenue distribution by state and segment, highlighting the top revenue-contributing states. Refer to Tableau file to view interactive dashboard.

Key Findings

- Across all segments**, the same four states dominate revenue: São Paulo (SP), Rio de Janeiro (RJ), Minas Gerais (MG), and Rio Grande do Sul (RS). Together, these states account for roughly two-thirds or more of revenue in each cluster, with all other states contributing relatively small shares. This pattern indicates that the segmentation is primarily **behavioural rather than regional**; the same core states appear across all segments, but customers within those states behave differently.
- Within this shared footprint, there are some notable differences particularly for **Active Value Buyers** segment. In this cluster, SP accounts for a substantially larger share of revenue than in the other segments, while the contribution from smaller states is reduced. This concentration is consistent with the behavioural profile of this cluster: São Paulo is a dense, mature e-commerce market with better logistics and higher digital adoption, which aligns with the **lower average delivery times and lower recency** observed for these customers.

IV. Payment Behaviour by Segment

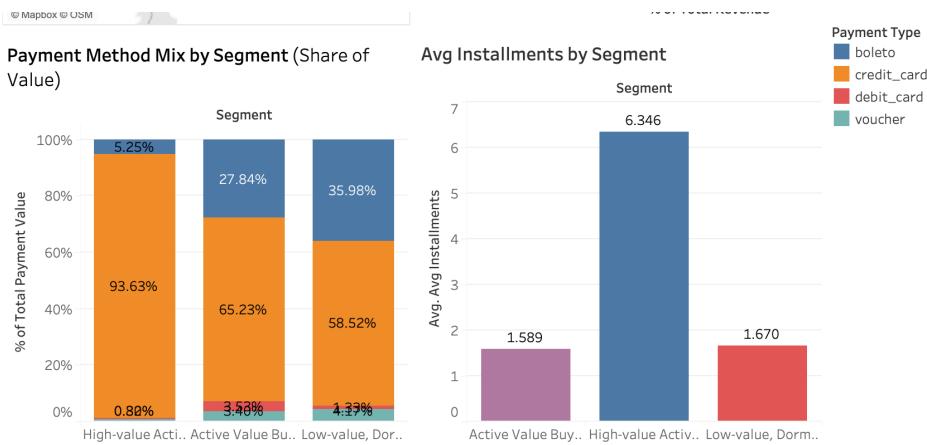


Figure 6: Payment Method Mix & Average Installments by Segment

Dashboard showing share of payment value by method (Credit card, Boleto, Debit card, Voucher) and average installments per order, broken down by customer segment. Refer to Tableau file to view interactive dashboard.

Across the customer base, credit cards are the dominant payment method, but its importance varies by segment. Key Findings:

- A. Among High-value Active Buyers, credit cards account for the vast majority of payment value, with only a small share paid via Boleto and negligible use of debit or voucher.** This aligns closely with their profile as big-ticket, instalment-heavy customers, a pattern reinforced by their high average number of instalments per order (over six on average).
- B. The Low-value, Dormant Buyers segment presents a more mixed payment profile.** Credit cards still represent the largest share of payment value, but Boleto accounts for a sizable proportion, making this the most boleto-reliant segment. Average installment usage is much lower than in the high-value cluster. This combination of lower spend, higher recency, and greater use of Boleto suggests more cautious or constrained purchasing behaviour, and a lower reliance on credit.
- C. The Active Value Buyers segment sits between the two extremes.** Credit cards remain the primary payment method, but Boleto still represents a meaningful share of payment value. Average installments per order are modest and similar to those of the dormant segment. This is consistent with the segment's broader behavioural pattern: customers are active and engaged,

making smaller, value-oriented purchases with light to moderate use of credit, rather than relying heavily on financing.

8.3 Product Category Performance

This section summarizes the key findings of which product categories drive performance on the platform and how their importance differs by customer segment. I focus on revenue and order volume within each segment.

I. Top Categories by Revenue and Orders

To understand which categories matter most overall, I aggregated both **revenue** and **order counts** by product category (using the cleaned `category_canonical` field). Figure 7 shows the category & performance & segment mix.



Figure 7: Category Performance and Segment Mix

Dashboard showing top categories by revenue and orders, and category revenue/order share by segment. Refer to Tableau file to view interactive dashboard.

Key Findings

A. Across all customers, performance is concentrated in a small set of **home, lifestyle, and tech-related categories** that dominate the top 5. In particular, the following categories form the core of both revenue and volume:

- a. bed bath table
- b. health beauty
- c. computers accessories
- d. furniture decor
- e. watches gifts

These categories appear at the top of both the **revenue** and **order** rankings, indicating that they are not only widely purchased but also generate a substantial share of total sales. Together, they define Olist's effective "core assortment" in this dataset.

II. Category Mix by Segment

To understand how category preferences differ by customer type, I computed category revenue share and order share within each segment. The bottom panels of Figure 7 show how the same set of categories contributes to each of the three clusters.

Key Findings

A. For **High-Value Active Buyers**, the mix is skewed towards **gifting and premium home/tech purchases**. Categories are as such:

- a. watches gifts,
- b. bed bath table,
- c. furniture decor, and
- d. computers accessories

This combination suggests that these customers are making **larger, "upgrade" style purchases** in the home and technology space, often financed via instalments, rather than focusing on day-to-day essentials.

B. The **Low-value, Dormant Buyers** segment shows a different emphasis. They still purchase from core home categories such as:

- a. bed_bath_table
- b. furniture_decor

However, **their category mix tilts more towards hobby and leisure items** (e.g. sports leisure) and lower overall spend.

- C. While having similar trends as Low-value, Dormantbuyers, **the Active Value Buyers segment** appears to be anchored in everyday self-care and home essentials, having purchased more **health beauty**.

This pattern aligns with their behavioural profile: they make more recent, smaller-value, “routine” purchases, often for personal care or household basics, with light to moderate use of credit.

Overall, these differences in category mix provide a clear basis for **segment-specific merchandising and promotion strategies**, which will be developed further in the Recommendations section.

8.4 Product Affinities (Market Basket Analysis)

This section investigates which product categories tend to be purchased together, and how those affinities differ by customer segment, using Market Basket Analysis (MBA) using association rules.

I. Determining Basket Structure and Granularity

A key methodological decision for the Market Basket Analysis was the level of granularity at which to model affinities: individual products or broader product categories.

Product-level rules can be very specific, but they require a sufficient number of multi-item baskets to produce stable patterns. Category-level rules are less granular but typically more robust and easier to interpret for business stakeholders

To inform this choice, I first examined the basket size distribution by treating each order ID as a basket and counting the number of distinct products per order (Figure 8).

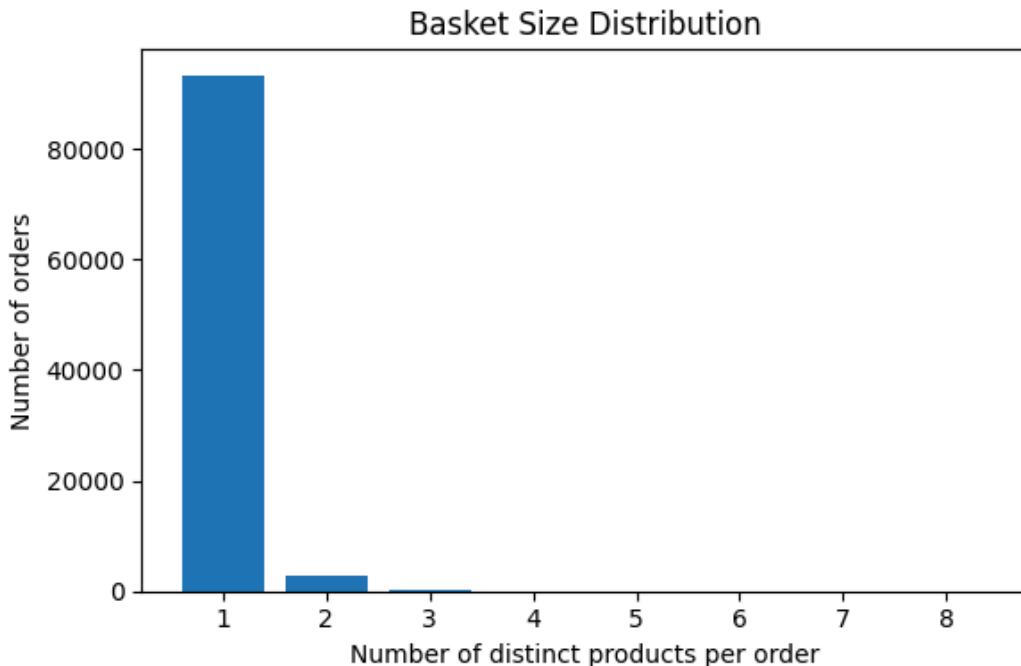


Figure 8: Basket Size Distribution
Number of distinct products per order across all completed orders.

Key Findings

Across 96,454 orders, the vast majority contain only **one product**, with a long tail of small multi-item baskets. This indicates that product-level co-occurrence is inherently sparse in this dataset: most baskets simply do not contain enough different SKUs to support reliable product–product rules.

Given this structure, I conducted the Market Basket Analysis at the **category level**. Aggregating products into categories increases support for meaningful patterns and yields rules that are **more stable, more generalisable, and easier to act on** (e.g. “bed & bath with home comfort”) than product-specific combinations.

V. 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:

- A. Apriori is treated as a conceptual baseline
- B. 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** ($\text{support} > 2\%$, $\text{confidence} > 50\%$, $\text{lift} > 1.2$) to obtain a compact, interpretable set suitable for cross-sell and bundling strategies.

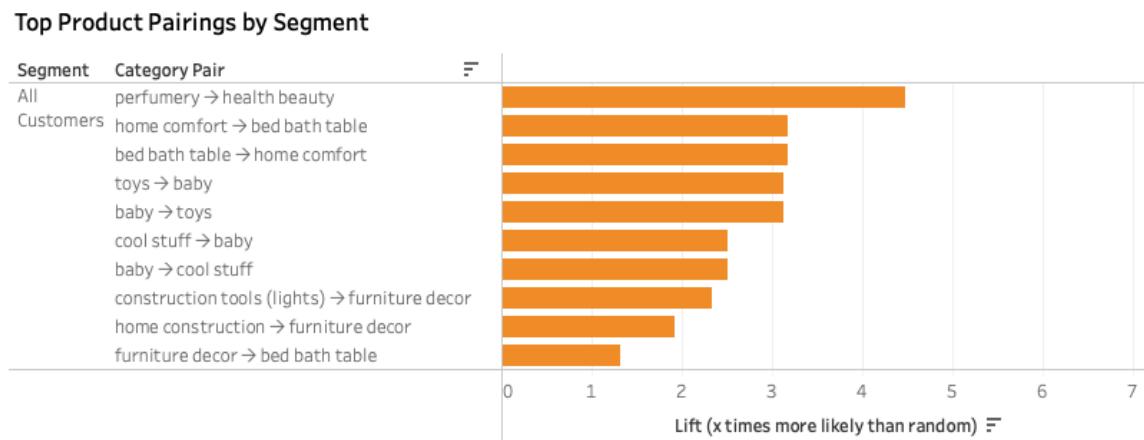


Figure 9: Top Product Pairings & Basket Insights
Top category pairs by lift and their support: confidence profiles across all customers. Refer to Tableau file to view interactive dashboard.

Figure 9 summarises the strongest category pairings across all customers, with lift (how much more often two categories occur together than expected by chance) as the key metric.

Key Findings

Across the full customer base, three broad purchase behaviours emerge in multi-category baskets:

A. Beauty and self-care

The rule **perfumery → health beauty** stands out, with confidence around 40% and lift of roughly 4.5×. Customers buying perfumery are far more likely than random to include broader health and beauty products in the same order.

This makes perfumery a strong anchor category for premium beauty bundles and “complete your self-care routine” recommendations.

B. Home upgrade / home comfort

One of the strongest rules is **home comfort → bed bath table**, with very high confidence (over 80%) and lift of around 3x. When customers buy home-comfort items, they almost always add bed and bath products to the same basket.

The reverse rule, **bed bath table → home comfort**, has lower confidence but higher reach, indicating that home comfort frequently acts as an add-on to a bed & bath mission.

More generally, combinations of **bed bath table**, **furniture decor**, **home comfort**, and occasionally **construction tools (lights)** or **home construction** form a clear home-upgrade pattern in which customers refresh décor, and sometimes tools or lighting within the same order.

C. Family / kids

There are tight two-way links between **baby** and **toys**, with confidence around 20–40% and lifts above 3x, indicating that baskets containing one category are much more likely than random to contain the other.

Additional rules such as **cool stuff → baby** and **baby → cool stuff** (with lifts above 2x) reinforce a family and gifting mission.

Taken together, the all-customer rules show that the most robust cross-category affinities are concentrated in **home**, **family/kids**, and **beauty**. These rules are both statistically strong and operationally meaningful, providing clear candidates for cross-sell prompts, bundled offers, and recommendation strategies.

VI. Segment-Specific Product Affinities

To understand how product affinities vary across customer types, I repeated the Market Basket Analysis within each segment. Figures 10-13 highlights the top rules by lift for each cluster.

Product Pairings & Basket Insights

Which category pairs are most often bought together, and how strong those relationships are by segment.

Top Product Pairings by Segment

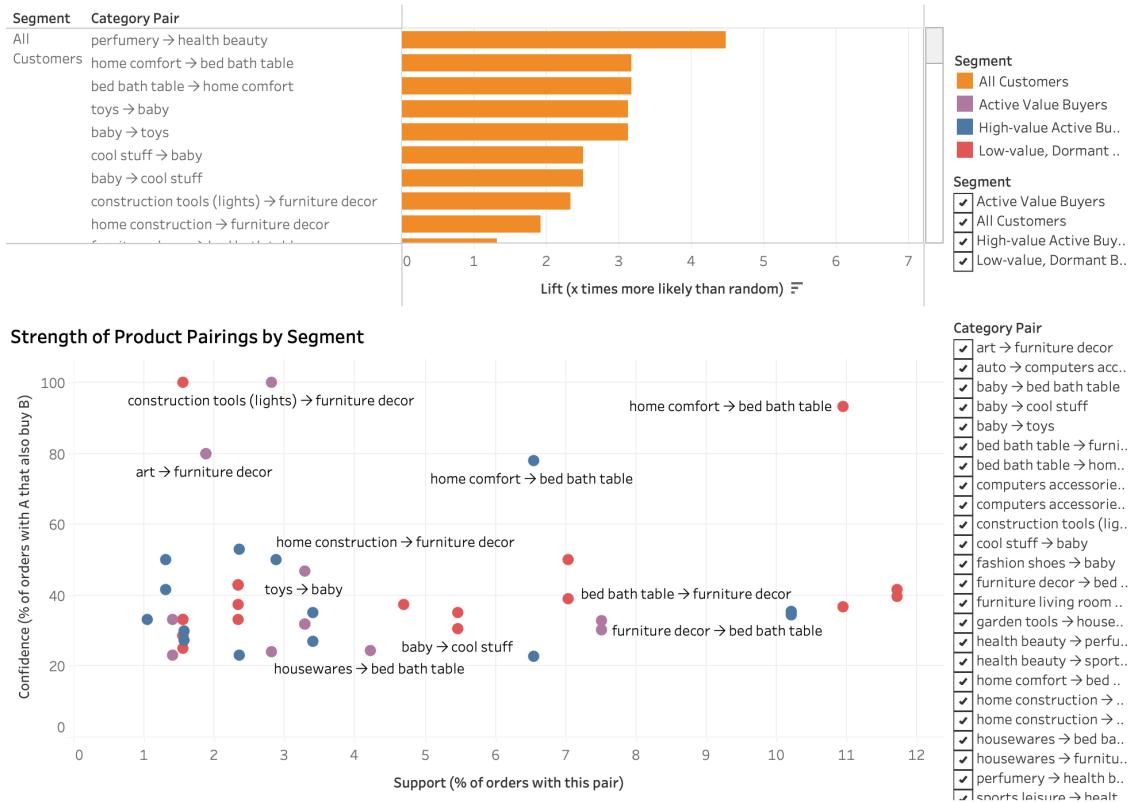


Figure 10: Top Product Pairings & Basket Insights

Top category pairs by lift and their support, and strength of product pairings by segment. Refer to Tableau file to view interactive dashboard.

Top Product Pairings by Segment

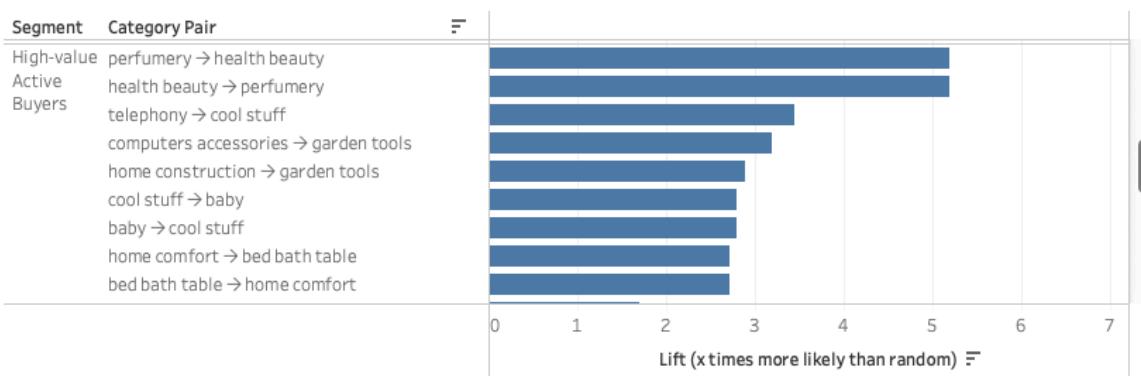


Figure 11: Top Product Pairings & Basket Insights for High-value Active Buyers

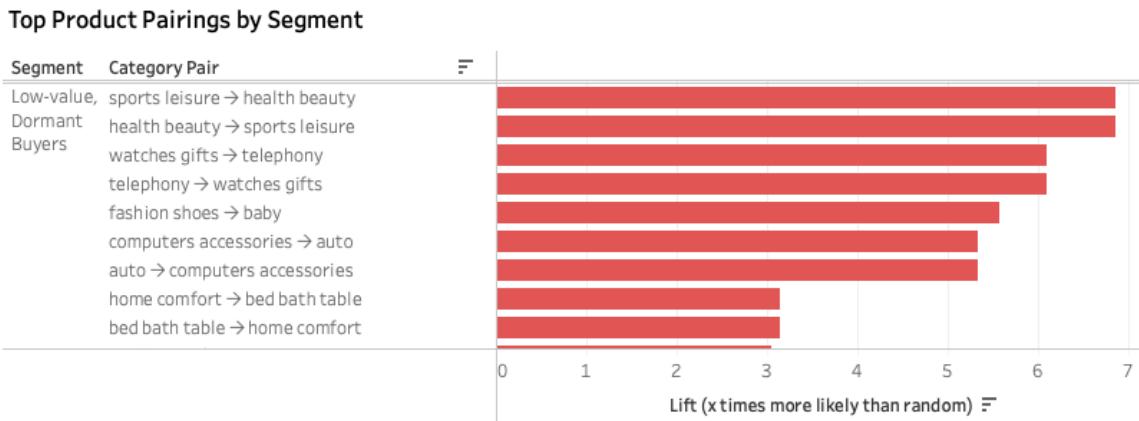


Figure 12: Top Product Pairings & Basket Insights for Low-value Dormant Buyers

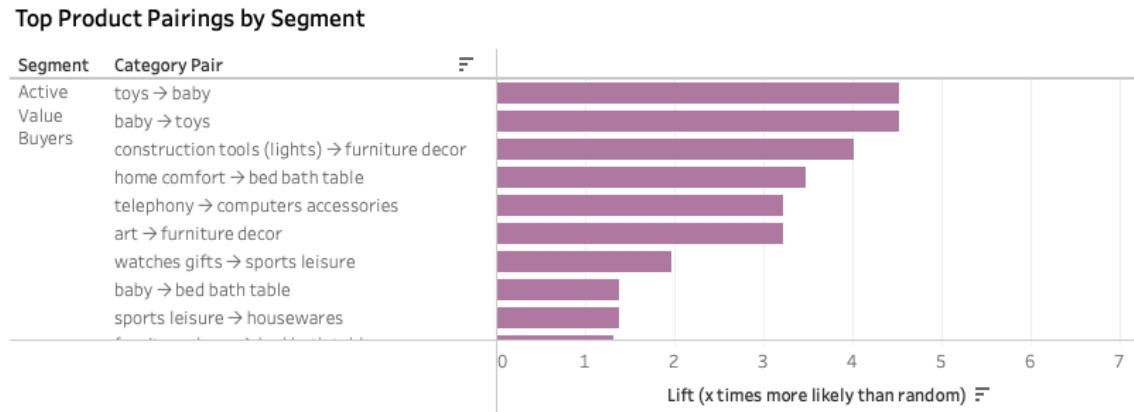


Figure 13: Top Product Pairings & Basket Insights for Active Value Buyers

Key Findings

A. High-value Active Buyers

Figure 11 shows that High-value Active Buyers are characterised by a mix of premium beauty and home-upgrade affinities. The most powerful rules link **perfumery** and **health beauty** in both directions, indicating that perfume purchasers are highly likely to buy wider beauty products in the same basket and vice versa. This supports the view of perfumery as a premium anchor for self-care bundles.

Alongside beauty, there are meaningful links from **home construction** and **computers accessories** to **garden tools**, and from **home comfort** to **bed bath table** (with the reverse rule also present).

These patterns suggest a pronounced “renovate then decorate” and home-comfort upgrading objective: when these customers invest in tools,

tech, or comfort items, they often complement the purchase with garden or bed & bath products.

Occasional pairings between **cool stuff** and **baby** further point to secondary family/gifting objectives within this high-value group.

B. Low-value, Dormant Buyers

Figure 12 highlights a slightly different emphasis for this segment. The strongest affinities connect **sports leisure** and **health beauty** in both directions, along with rules linking **watches gifts** and **telephony**, and **fashion shoes** to **baby**.

These pairings indicate that when this segment does place multi-category orders, they tend to centre around hobby, lifestyle, and gifting rather than strictly routine purchases.

Home-related rules are still present: most notably **home comfort** and **bed bath table** (and the reverse), but they sit alongside pairings between **computer accessories** and **auto**, suggesting that this segment's multi-category baskets are infrequent but highly goal-driven (e.g. a specific home refresh, a tech/auto hobby purchase, or a targeted gifting occasion).

C. Active Value Buyers (Cluster 2)

As shown in Figure 13, this segment displays the strongest family and everyday-home patterns. The dominant rules link **toys** and **baby** in both directions, with high lift and solid support, indicating tightly structured kids/family baskets.

This confirms that a significant share of this segment's multi-category orders revolve around children's needs and family gifting.

Home-upgrade behaviour is also evident through rules such as **construction tools (lights) → furniture decor**, **home comfort → bed bath table**, and **art → furniture decor**, suggesting a "renovate and decorate" and home styling mission within this group.

Additional pairings, including **telephony → computers accessories** and **watches gifts → sports leisure**, point to smaller hobby or gadget-oriented baskets that complement the core family and home activity.

D. Overall, the segment-level analysis shows that while all three segments participate in home, family, and beauty missions, they do so with **different intensity and focus**: High-value Active Buyers skew towards premium beauty and upgrade-style home baskets, Low-value Dormant Buyers exhibit occasional but concentrated lifestyle and gifting baskets, and Active Value Buyers are driven by frequent family and everyday home combinations. These differences provide a clear basis for **segment-specific bundles and cross-sell strategies** in the recommendations that follow.

9.0 Conclusion: What These Patterns Mean for the Business

9.1 Recommendations for Overall Marketing & Product Strategy

I. Prioritise high-value and high-potential segments. In order of priority:

A. **High-value Active Buyers as the primary revenue engine. Focus on retention, satisfaction, and high-margin cross-sell/upsell:**

1. For targeting, prioritise this segment for high-value prospecting and creation of lookalike audiences.
2. Maintain depth and premium positioning in home-upgrade and beauty categories.
3. Tie these categories to installment-led propositions, reflecting the segment's strong financing usage.
4. For cross-selling & upselling:
 - a) *Push home-upgrade bundles or finish the “renovate then decorate” journey narratives: “Bed & Bath + Décor + Home Comfort + Construction + Furniture Decor” when any of those categories appear in basket.*
 - b) *Use perfumery as a trigger: if perfume is in the cart, surface high-margin **health beauty** add-ons.*
 - c) *When **baby** is present, introduce **gifting / family extras** (**cool_stuff, toys**).*

B. Use Active Value Buyers as the volume and engagement engine, growing their value via frequency-building tactics and smart bundling:

1. For targeting, Use this group as the model for mass acquisition and CRM, given its size and engagement.
2. Use this group to test new value propositions (smaller bundles, replenishment offers, loyalty mechanics) before scaling.
3. Keep strong, competitively priced ranges in everyday home and kids categories.
4. For cross-selling & upselling:
 - a) *Run always-on home living bundles, framed as everyday improvement: suggest “finish the room” and/or “make your home more comfortable” type of narrative.*
 - b) *Make baby ↔ toys the core cross-sell engine with “Baby + Toys essentials”, “New parent starter packs” kind of prompts.*

C. Manage Low-value, Dormant Buyers as a churn and reactivation pool but avoid high investments. Keep segment but piggyback on existing hero categories rather than designing assortment specifically for this segment.

II. Double down on core categories

- A. Protect and grow the key **home** and **beauty** categories that already drive most revenue.
- B. Use segment-level category performance to align assortment, merchandising, and promos with the clusters that over-index in each category.

III. Operationalise affinities in the customer journey

- A. Embed the strongest category rules into:
 1. on-site recommendation modules, such as homepage, cart add-on recommendations, post-purchase,

2. lifecycle communications, and
3. pre-built bundles or “complete the set” offers.

IV. Align credit and payment propositions with segments

- A. For installment-heavy segments, attach premium bundles and upgrades to attractive installment options.
- B. For Boleto-heavy segments, keep boleto-friendly, low-friction offers and simple bundles.

9.2 Next Steps and Further Analysis

Overall, the segmentation, category performance, and product affinity analyses provide a **coherent foundation** for designing segment-specific targeting, cross-sell strategies, and product and category prioritisation. Here are the next steps that I would take to finetune the insights and strategy:

I. Refine personas and behavioural understanding

- A. Drill down to product-level analysis: use product-level MBA within high-volume categories to uncover niche, high-value pairings, while keeping category-level rules as the main operational layer for recommendations and bundles.
- B. Validate inferred category pairings: Treat home-upgrade, family/kids, and beauty narratives as hypotheses to be validated, and enrich with additional qualitative research, customer surveys, and additional behavioural signals (e.g. browsing patterns, campaign engagement).

II. Test and iterate on marketing activation

- A. Segment-based campaign experiments: A/B test segment-specific messages, incentives, and bundles and measure incremental lift in AOV, frequency, and reactivation.
- B. Operationalise MBA in product and UX: Test & implement the strongest rules as recommendation logic on homepage, cart, and post-purchase pages, and track attach rates.

10.0 Appendix

Data Dictionary

1. Original fields from Olist public dataset

Dataset original source	Field	Description
customers_dataset.csv	customer_id	Internal order-level customer key used in orders_*. Each order has a unique customer_id.
	customer_unique_id	Unique identifier of a customer.
	customer_zip_code_prefix	First five digits of the customer's zip code
	customer_city	Customer's city name
	customer_state	Customer state code
order_items_dataset.csv	order_id	Unique order identifier
	order_item_id	Sequential number identifying number of items included in the same order.
	product_id	Unique product identifier
	seller_id	Unique seller identifier
	price	Item price charged for a given line.
	freight_value	Item freight value item (if an order has more than one item the freight value is split between items)
order_payments_dataset.csv	order_id	Unique order identifier
	payment_sequential	Sequence number of each payment leg. A customer may pay an order with more than one payment method. If he does so, a sequence will be created to accommodate all payments.

	payment_type	Method of payment chosen by the customer.
	payment_installments	Number of installments chosen by the customer.
	payment_value	Transaction value paid in the given leg.
orders_dataset_core.csv	order_id	order unique identifier
	customer_id	Internal order-level customer key used in orders_*. Each order has a unique customer_id.
	order_status	Reference to the order status (delivered, shipped, etc).
	order_purchase_timestamp	Shows the purchase timestamp.
	order_approved_at	Shows the payment approval timestamp.
	order_delivered_carrier_date	Shows the order posting timestamp when it was handed to the logistic partner.
	order_delivered_customer_date	Shows the actual order delivery date to the customer.
	order_estimated_delivery_date	Shows the estimated delivery date that was informed to the customer at the purchase moment.
product_category_name_transla tion.csv	product_category_name	Category name in Portuguese
	product_category_name_english	Category name in English

2. Engineered Features

2.1 Customer level behavioural features

Feature name	Level	Definition	Rationale
order_total	Order	Total amount paid per <code>order_id</code> , summed from all payment records.	Provides monetary value of each order, even when payments are split.
total_revenue	Customer / Segment	Sum of <code>order_total</code> across all completed orders for a customer or group.	Used to identify high-value customers and revenue contributions by segment.
n_orders / order_frequency	Customer	Number of unique completed orders per customer (<code>nunique(order_id)</code>).	Captures purchase frequency, a core dimension for segmentation.
recency_days	Customer	Number of days between the customer's most recently completed order and the analysis reference date.	Measures how recently a customer has purchased (active vs lapsed).
delivery_time_days	Order	Days between purchase and delivery (<code>order_delivered_customer_date - order_purchase_timestamp</code>).	Captures the delivery experience at the order level.
avg_delivery_days	Customer	Average <code>delivery_time_days</code> across all completed orders for a customer.	Allows analysis of whether delivery speed is linked to behaviour or churn.

avg_order_total	Customer	Average <code>order_total</code> across completed orders.	Distinguishes customers who place few large orders vs many small orders.
order_total_log	Order	Log-transformed order value (e.g. <code>log(order_total + 1)</code>).	Reduces skew from very large orders, making features more suitable for clustering.
avg_order_total_log	Customer	Average of <code>order_total_log</code> per customer.	Stabilises variance while preserving differences between low- and high-spend customers.
order_installments	Order	Maximum number of payment installments used for each <code>order_id</code> .	Reflects credit usage for a given order when multiple payment records exist.
avg_installments	Customer	Average <code>order_installments</code> across a customer's orders.	Captures payment behaviour and reliance on instalments over time.
avg_installments_log	Customer	Log-transformed version of <code>avg_installments</code> .	Handles skew where a small subset of customers use very high installments.

2.2 Category-level performance features

Feature name	Level	Definition	Rationale
category_final	Item / Category	English product category name, derived from the translation table.	Makes categories interpretable for stakeholders and dashboards.
category_canonical	Category	Cleaned / consolidated version of the category field (e.g. merged rare or noisy categories).	Ensures categories have sufficient volume and are stable enough for analysis and MBA.
revenue_share_pct	Category	Percentage of total revenue contributed by each category.	Identifies top revenue-driving categories and their relative importance.
n_orders (category)	Category	Number of unique orders that contained at least one item from the category.	Shows how widely the category appears across baskets .
n_items (category)	Category	Total number of items sold from the category.	Highlights high-volume categories, even if unit price is low.
avg_price	Category	Average item price per category.	Positions categories as premium vs mass and informs pricing and promotion strategy.
avg_freight	Category	Average freight (shipping) cost per item per category.	Provides insight into logistics cost and

			potential margin considerations.
--	--	--	----------------------------------

2.3 Rule-Level Features for Market Basket Analysis

Feature name	Level	Definition	Rationale
Antecedent	Rule	The “IF” side of the rule (category or category set A).	Identifies the trigger category or combination.
Consequent	Rule	The “THEN” side of the rule (category or category set B).	Identifies the recommended / associated category.
support_count	Rule	Number of orders that contain both A and B together.	Ensures rules are based on a meaningful absolute volume of transactions.
support_pct	Rule	support_count divided by total number of orders.	Measures how frequent the combination is across all transactions.
conf_pct (confidence)	Rule	Proportion of orders containing A that also contain B ($P(A \text{ and } B)$)	Used to measure how reliable a rule is for cross-sell.
lift	Rule	Ratio of confidence to the baseline probability of B ($\frac{P(A \text{ and } B)}{P(B)}$)	Verifies if the association between A and B is stronger than random chance.

ante_support_pct	Rule	Percentage of orders that contain the antecedent A.	Puts confidence into context by showing how common A is on its own.
cons_support_pct	Rule	Percentage of orders that contain the consequent B.	Shows how popular B is independently of A.