Olist E-Commerce Growth & Retention Analysis

Executive Summary

This project analyzes growth and retention patterns for Olist, a Brazilian e-commerce marketplace, using PostgreSQL and Tableau Public. The analysis reveals that revenue growth is primarily acquisition-led, driven by increasing order volume rather than higher spend per order.

Repeat customers contribute less than 6% of total revenue, indicating weak retention and low customer lifetime value. Repeat behavior is delayed rather than immediate, with most returning customers purchasing again only after long gaps.

A product-level deep dive shows that while revenue is highly concentrated in a small number of categories, these high-revenue categories exhibit low repeat behavior. In contrast, a small subset of consumable and lifestyle categories shows strong loyalty potential.

The findings suggest that future growth should rebalance from acquisition-only strategies toward targeted retention, category-specific loyalty initiatives, and repeat-purchase acceleration.

Business Context & Key Questions

Olist operates a large multi-category e-commerce marketplace where sustainable growth depends not only on acquiring new customers but also on retaining existing ones.

This analysis was guided by the following business questions:

- Is revenue growth driven by higher order volume or higher spend per order?
- How dependent is the business on one-time customers?
- How quickly do customers return after their first purchase?
- Which product categories show strong repeat behavior?
- Is revenue diversified across categories or highly concentrated?

Dataset & Tools

Dataset: Public Olist e-commerce dataset (orders, order items, products, customers)

https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Tools Used:

- PostgreSQL: Data ingestion, data modeling, CTEs, window functions, and metric creation.
- Tableau Public: Executive dashboards and product-level storytelling.

SQL was used to create analytically correct, Tableau-ready output tables, while Tableau was used exclusively for visualization and communication.

Data Modelling & Definitions

Data Model Overview

The raw Olist dataset consists of multiple normalized tables representing different business entities, including orders, order items, customers, products, and sellers. To enable scalable and accurate analysis, the data was organized into a star-schema-inspired analytical model.

At the center of the model is an order-item–level fact table, supported by descriptive dimension tables.

Fact Table: Order Items

The primary fact table is built at the order item grain, where each row represents a single product purchased within an order.

This grain was intentionally chosen because:

- A single order can contain multiple products.
- Revenue, pricing, freight cost, and seller information exist at the item level.
- Aggregating directly at the order level would undercount revenue and distort KPIs.

This design ensures that all revenue and volume metrics are calculated correctly.

Dimension Tables

Several dimension tables were used to enrich the fact table:

- Customers – customer identifiers and customer uniqueness
- Products – product attributes and product category
- Sellers – seller-level attributes
- Orders – order timestamps and order lifecycle metadata

Dimension tables provide descriptive context while keeping the fact table lean and performant.

Customer Definition (Critical Design Choice)

A key modeling decision was to use customer_unique_id rather than customer_id to define customers.

This is critical because:

- A single real-world customer can have multiple customer_id values.
- Using customer_id would incorrectly inflate the customer count.
- Repeat customer analysis would be fundamentally wrong.

All retention and repeat-purchase metrics are therefore based on unique customers, ensuring analytical correctness.

Separation of Raw and Analytics Layers

Two logical schemas were used:

- Raw schema – preserves source data without modification.
- Analytics schema – contains transformed, analysis-ready tables.

This separation:

- Protects source data integrity.
- Makes analytical logic explicit and auditable.

- Reflects real-world analytics engineering best practices.

Metric Definitions

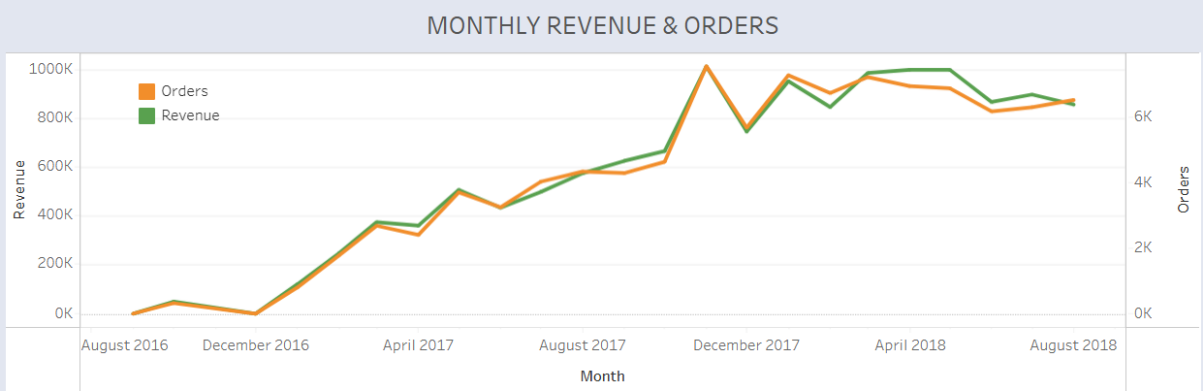
Key business metrics were defined as follows:

- Revenue: Sum of item-level price values.
- Orders: Count of distinct order IDs.
- Average Order Value (AOV): Revenue divided by total orders.
- Repeat Customer: A customer with more than one distinct order.
- Repeat Revenue Share: Percentage of revenue generated by repeat customers.

All metrics were computed in SQL and validated before being used in Tableau.

Analysis & Key Findings

(A) Revenue & Orders Over Time (Monthly)



Key Patterns:

1. **Growth is order-led:** Revenue increases because more customers are placing orders not because customers are spending more per order. This aligns with: Flat AOV & strong order growth.
2. **No pricing or basket expansion effect:** If the business had successful upselling, bundling, or pricing optimization we would expect AOV to rise meaningfully. That is not happening here.
3. **Growth quality question:** This table alone raises an important question, "If growth slows, do we have any lever besides acquiring more customers?" This question connects directly to repeat customer analysis.

Revenue growth has been strong and consistent, but it is primarily driven by increasing order volume rather than higher spend per order(both orange and green lines are synchronous over time), indicating acquisition-led growth rather than value-led or loyalty-led growth.

Actionable Recommendations:

1. **Shift focus from acquisition-only growth**: Continue acquisition, but introduce post-purchase upsell flows & promote complementary products.

2. **Experiment with AOV levers:** such as bundles, threshold-based free shipping, or Category-specific cross-sell.
3. **Retention initiatives:** Increasing repeat frequency would increase orders and increase revenue without proportional acquisition cost.

(B) Revenue by Customer Type (One-time customer or Repeat customer)

customer_type	total_revenue
One-time	12812821.73
Repeat	778821.97

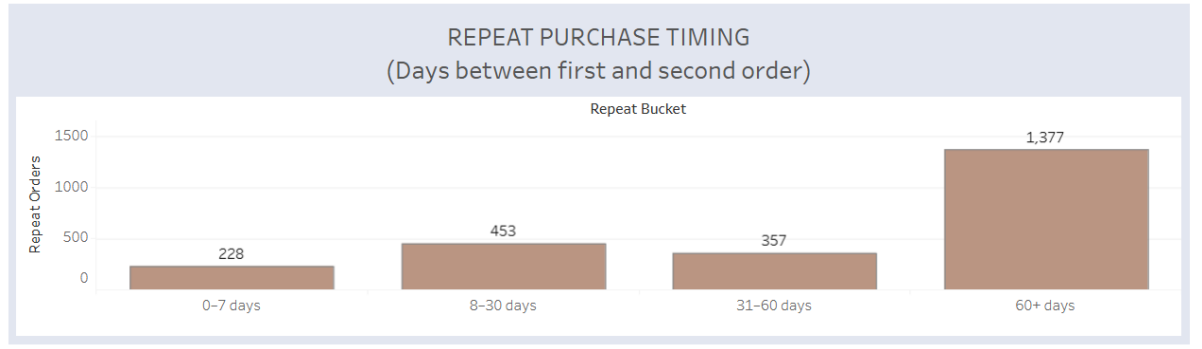
Over 94% of total revenue comes from one-time customers & repeat customers contribute less than 6% of the revenue. The revenue gap between one-time and repeat customers is massive, not marginal which is not a subtle pattern.

The business relies almost entirely on first-time purchases for revenue, with repeat customers contributing a disproportionately small share, indicating weak retention and low customer lifetime value.

Actionable Recommendations (Tied Directly to the Data)

1. **Make retention a top-3 growth KPI:** Even a small lift in repeat revenue would materially impact total revenue
2. **Invest in early lifecycle engagement:** Focus on first 30 days post-purchase such as follow-up communication, personalized recommendations, etc.
3. **Rebalance growth spend:** Invest towards retention experiments rather than just acquisition.

(C) Repeat Order Timing



This analysis groups repeat customers based on the number of days between their **first and second purchase**, segmented into four time buckets: 0-7 days, 8-30 days, 31-60 days, and 60+ days.

Key Patters:

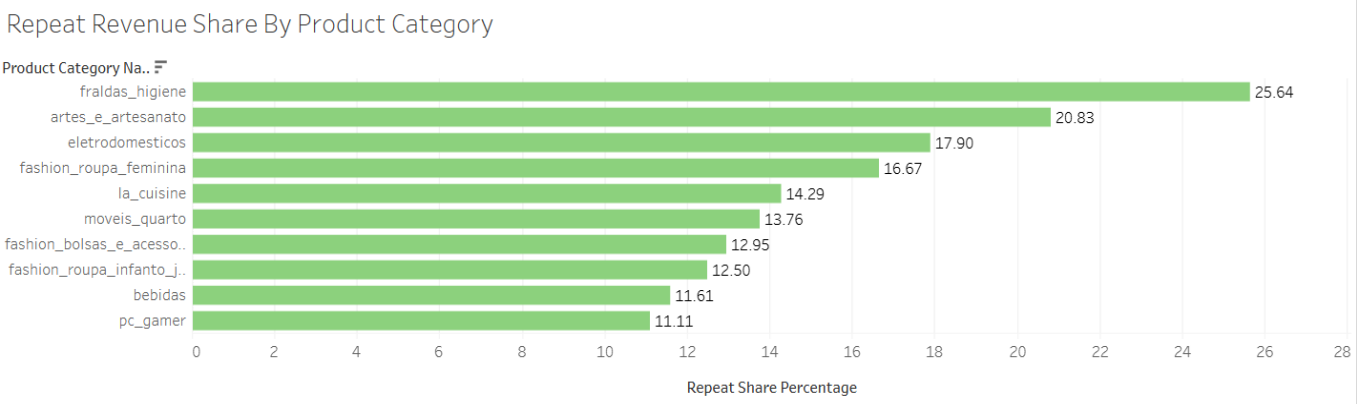
1. **Delayed repeat behavior dominates:** This suggests customers are not building immediate loyalty and that the purchases are episodic, not habitual.

Customers who do return tend to do so after long gaps, indicating that repeat behavior is driven by delayed need rather than strong post-purchase engagement or brand loyalty. This insight explains: Why repeat revenue is so low? Why AOV is flat? Why growth is acquisition-led?

Actionable Recommendations :

1. Focus retention efforts in first 30 days: Post-purchase emails, product recommendations, incentives for second purchases.
2. Design category-specific re-engagement: For categories with long replacement cycle promote replenishment nudges and early cross sell.

(D) Repeat revenue Share – Category-wise



Top 10(out of 78) repeat share categories with percentage:

category	English	total_order_items	repeat_items	repeat_share_perc
fraldas_higiene	Diapers & Hygiene	39	10	25.641
artes_e_artesanato	Arts & Crafts	24	5	20.833
eletrodomesticos	Home Appliances	771	138	17.899
fashion_roupa_feminina	Women's Clothing	48	8	16.667
la_cuisine	Kitchen & Culinary	14	2	14.286
moveis_quarto	Bedroom Furniture	109	15	13.761
fashion_bolsas_e_acessorios	Fashion Bags & acc	2031	263	12.949
fashion_roupa_infanto_juvenil	Children's Clothing	8	1	12.5
bebidas	Beverages	379	44	11.609

Key patterns:

1. **Consumables & lifestyle categories repeat more:** High-repeat categories tend to be consumables (e.g., hygiene-related) or lifestyle / personal interest categories(fashion). These naturally support replenishment and habit formation.
2. **Many categories are inherently one-time:** Low-repeat categories are Occasion-based (flowers), Media / one-off purchases, Service-like products. Trying to "fix" retention here would be inefficient.
3. **Volume is not equal to Loyalty:** Some categories with High order volume Still have low repeat share. This reinforces why repeat share (not repeat count) was the right metric.

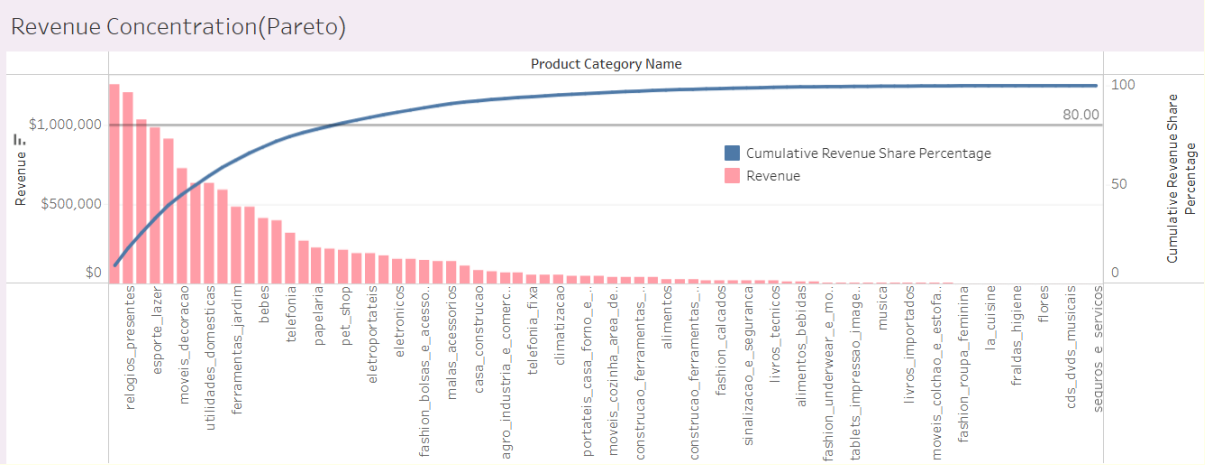
Actionable Recommendations (Tied Directly to the Data):

1. Explore subscriptions where repeat share is highest.
2. Prioritize retention spend by category: Focus CRM, lifecycle, and loyalty efforts on top repeat-share categories (e.g., hygiene, lifestyle) and measure success by increase in repeat share and earlier repeat timing.

Treating all categories with the same retention strategy is inefficient. Categories with higher repeat share represent natural candidates for loyalty programs, replenishment reminders, and long-term customer engagement. In contrast, low-repeat categories are better positioned as acquisition or cross-sell entry points rather than retention drivers.

(E) Revenue concentration by product categories:

Top to bottom categories based on revenue and their cumulative share.



This analysis examines how total revenue is distributed across product categories using a Pareto (80/20) framework. The bar chart represents revenue by category, while the cumulative line shows the percentage of total revenue contributed as categories are added from highest to lowest revenue.

Key patterns:

1. **Revenue is highly concentrated:** This means that the business is not diversified across categories. A small subset of categories dominates financial performance.

2. **The business's highest-revenue categories are not the categories that build customer loyalty**, creating a tension between short-term revenue generation and long-term customer value.

Actionable Recommendations:

1. **Protect high-revenue categories:** For top revenue categories, we need to optimize pricing, supply reliability and conversion. Although these categories are acquisition heavy but loyalty gains will be incremental.
2. **Grow high-repeat categories intentionally:** For high repeat-share categories: Increase visibility and cross-sell placement. We can also explore Subscriptions and replenishment reminders (promo emails/texts).

The business is financially dependent on a narrow set of high-revenue categories. This creates risk: any demand slowdown, supply disruption, or competitive pressure in these categories would have a disproportionate impact on overall revenue. At the same time, many of these high-revenue categories exhibit low repeat behavior, limiting long-term revenue stability.

Business Recommendations

Based on the combined findings across growth trends, retention behavior, and product category dynamics, the following recommendations focus on improving revenue durability, not just short-term growth.

Rather than addressing isolated metrics, these actions are designed to rebalance the growth model toward higher customer lifetime value while managing revenue concentration risk.

1. Rebalance Growth Strategy from Acquisition-Led to Retention-Enabled

Decision:
Shift growth measurement from primarily "new orders acquired" to repeat purchase rate and time-to-second-purchase as core performance indicators.

Rationale:
Current revenue growth is driven almost entirely by one-time customers, creating long-term dependency on continuous acquisition. Improving early retention would allow revenue to compound without proportional increases in acquisition spend.

Action:

- Make "time to second purchase" a tracked KPI alongside revenue and orders.
- Hold marketing and product teams accountable for post-purchase engagement performance.

2. High Repeat Potential, Low Volume Categories

Example: Diapers & Hygiene, Arts & Crafts, Kitchen & Culinary

Decision:
Abandon one-size-fits-all retention tactics and implement category-specific lifecycle strategies.

Rationale:
Repeat behavior varies significantly by category. Some categories naturally lend themselves to

repeat purchases, while others function primarily as entry points. Treating them equally dilutes ROI.

Action:

- Introduce replenishment reminders (e.g., time-based nudges for hygiene and consumables)
- Pilot subscription or auto-reorder options where applicable
- Increase visibility of these categories in post-purchase emails and home page recommendations

3. Use High-Revenue Categories as Cross-Sell Gateways, Not Loyalty Anchors

Decision:
Avoid over-investing in retention mechanics for high-revenue but low-repeat categories.

Rationale:
Pareto analysis shows revenue concentration in categories that do not exhibit strong repeat behavior. Attempting to force loyalty in these categories is likely inefficient.

Action:

- Optimize high-revenue categories for conversion and margin, not repeat.
- Design post-purchase pathways that guide customers into higher-repeat categories.
- Measure success by downstream category migration, not repeat within the same category.

4. Compress the Repeat Purchase Timeline

Decision:
Explicitly target earlier repeat behavior as a strategic objective.

Rationale:
Most repeat purchases currently occur after long gaps, indicating weak habit formation. Shortening the repeat window improves retention probability and lifetime value.

Action:

- Concentrate engagement efforts in the first 30 days post-purchase
- Test incentives, reminders, and personalized recommendations aimed at accelerating second purchases.
- Evaluate success based on shift in repeat timing distribution, not just repeat count.

5. Manage Revenue Concentration Risk Proactively

Decision:
Treat revenue concentration as an ongoing risk metric, not just a descriptive insight.

Rationale:
Dependence on a small number of categories exposes the business to demand and supply volatility. Growth in smaller but loyal categories can act as a stabilizing counterbalance.

Action:

- Monitor category-level revenue concentration quarterly.

- Invest selectively in growing repeat-heavy categories even if short-term revenue impact is modest.
- Use diversification as a long-term resilience strategy rather than a near-term revenue lever.

Project Limitations

While this analysis provides meaningful insights into growth, retention, and product category dynamics, several limitations should be acknowledged:

- No customer lifetime value (CLV) modeling:
  The analysis focuses on repeat behavior and revenue contribution but does not estimate long-term customer value or profitability.
- Limited behavioral depth:
  Repeat customers are defined at the platform level; repeat purchases within the same product or category were not separately analyzed.
- No marketing or acquisition channel data:
  Without channel-level attribution, it is not possible to assess how different acquisition sources impact retention and repeat behavior.
- Static time-based analysis:
  The analysis does not account for cohort effects or changes in customer behavior over time based on acquisition period.

Links

- Project Documents (Github): https://github.com/Janarthan-Anuraag/OList-Ecommerce-Growth-Retention-Analysis/tree/main

- Tableau: https://public.tableau.com/app/profile/janarthan.anuraag/viz/OlistE-CommerceGrowthRetentionAnalysis/ExecutiveGrowthRetentionOverview

- Database: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce