

Business Insights Report: Olist E-Commerce Dataset

FELLAH Haithem

September 1, 2025

1 Executive Summary

This report presents a full-cycle analysis of the Brazilian e-commerce dataset from Olist. It covers data cleaning, exploratory analysis, segmentation, and business insights. The goal is to extract actionable recommendations for improving customer satisfaction, operational efficiency, and strategic planning.

Key Findings:

- Order volume remained stable, while revenue showed significant volatility.
- Customer base is heavily concentrated in São Paulo and other urban centers.
- Review scores are skewed toward positive ratings, but delivery delays negatively impact satisfaction.
- Delivery delay and review score show a moderate negative correlation of -0.267.

2 Dataset Overview

The dataset consists of 9 CSV files covering orders, customers, products, sellers, payments, reviews, and geolocation data.

Table 1: Dataset Dimensions

Table	Rows	Columns
olist_orders_dataset	99,441	8
olist_customers_dataset	99,441	5
olist_order_items_dataset	112,650	7
olist_order_payments_dataset	103,886	5
olist_order_reviews_dataset	99,224	7
olist_products_dataset	32,951	9
olist_sellers_dataset	3,095	4
olist_geolocation_dataset	1,000,163	5
product_category_name_translation	71	2

3 Data Cleaning and Preprocessing

Missing Values

Key observations:

- **orders**: delivery-related fields contain nulls due to incomplete or canceled orders.
- **reviews**: most review comments are missing, but scores are complete.
- **products**: 600 rows missing category and description metadata.

Date Formatting

All relevant date columns were converted to `datetime` format for time series analysis.

Category Translation

Product categories were enriched using the translation table to support English-language reporting.

4 Exploratory Data Analysis

4.1 Order Volume and Revenue Over Time

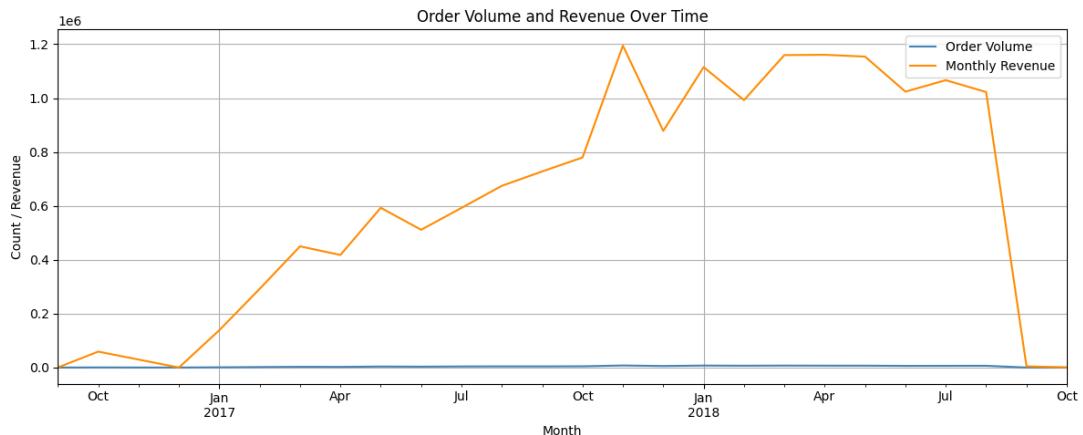


Figure 1: Order Volume and Revenue Over Time

Observations:

- Order volume remained stable from 2016–2018.
- Revenue spiked sharply in late 2017 and dropped in late 2018.
- Revenue fluctuations are not directly tied to order volume.

4.2 Customer Distribution by State

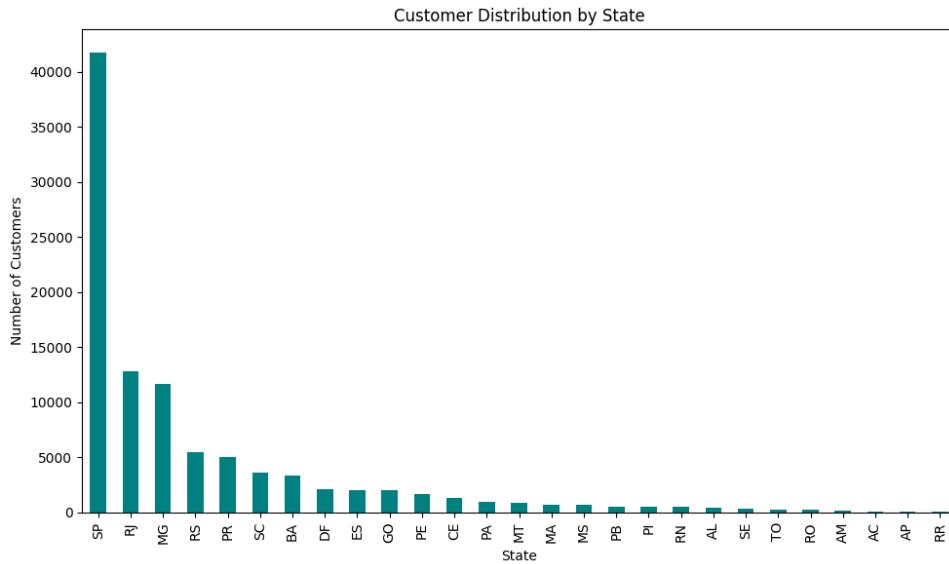


Figure 2: Customer Distribution by State

Observations:

- São Paulo (SP) accounts for over 40% of customers.
- RJ, MG, and RS follow with moderate volumes.
- Northern states show minimal customer activity.

4.3 Review Score Distribution

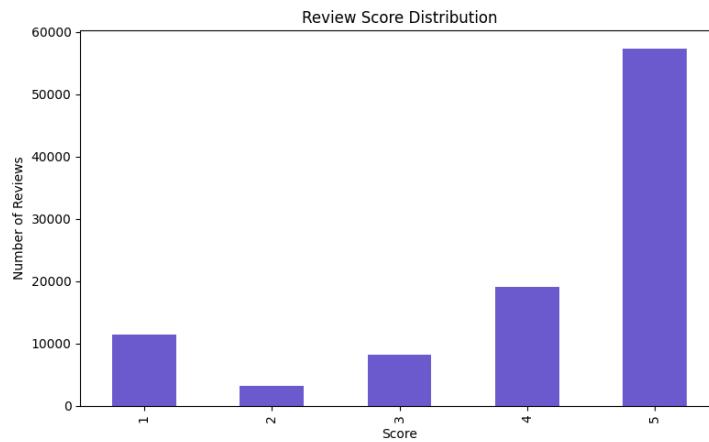


Figure 3: Review Score Distribution

Observations:

- Score 5 dominates, followed by score 4.
- Scores 1–2 are rare but significant for quality control.

5 Segmentation and Patterns

Customer Frequency and Spend

Customers were segmented by order count and total spend. High-value customers show elevated average order values and repeat behavior.

Delivery Delay vs. Review Score

A correlation of **-0.267** was found between delivery delay and review score, confirming that late deliveries reduce customer satisfaction.

6 Business Insights

- **Delivery performance** directly affects review scores and retention.
- **Geographic targeting** should focus on SP, RJ, MG, and RS.
- **Product metadata gaps** should be audited for completeness.
- **Revenue drivers** may stem from pricing or product mix, not volume.

7 Recommendations

1. Improve delivery logistics to reduce delays and boost satisfaction.
2. Prioritize high-density regions for marketing and seller onboarding.
3. Segment and engage high-value customers with loyalty programs.
4. Clean incomplete product entries to enhance discoverability.
5. Investigate revenue spikes for replicable strategies.