

# **Olist Brazilian E-commerce Analysis**

**Analyzing Sales, Seller Performance, and  
Customer Experience.**

**Prepared by: OLUWATOMIWA ADELEKE**  
**E-mail: [oluwatomiwaadeleke@gmail.com](mailto:oluwatomiwaadeleke@gmail.com)**  
**[LinkedIn](#), [Github](#)**

# 1. PROJECT INTRODUCTION

## 1.1 Project Overview

This project provides a deep-dive analysis of the Brazilian E-commerce landscape using data from Olist, the largest department store in Brazilian marketplaces. By integrating multiple relational data points from order placement to final customer review, this study identifies the drivers of revenue and the bottlenecks in logistics that influence customer satisfaction.

## 1.2 Objectives

- **Analyze Sales:** Identify seasonal trends, high-value categories, and revenue concentration.
- **Customer Behavior:** Quantify acquisition, retention, and the profile of repeat vs. one-time buyers.
- **Operational Efficiency:** Measure delivery lead times, seller processing speed, and their impact on customer satisfaction.
- **Customer Experience:** Connect review sentiment (score and text) to key operational metrics.

## 1.3 Data Source & Tools

- **Data Source:** The dataset was sourced from [Kaggle](#).
- **Tools:**
  - Python (Pandas/NumPy): For data wrangling and ETL
  - Power BI: For data modelling and interactive visualization.
  - DAX: To create complex measures like "Average Delivery Delay" and "Repeat Purchase Rate".
  - Microsoft Excel: Supporting data inspection
  - Jupyter Notebook: Exploratory analysis and data preparation

# 2. Dataset Engineering & Descriptions

## 2.1 Dataset Overview

The Olist ecosystem is composed of several datasets that can be linked by unique keys (Order IDs, Customer IDs, etc.). This relational structure allows for a holistic view of the "Order-to-Review" lifecycle.

## 2.2 Datasets Description

- **Orders:** Central table tracking the status and timing of every transaction.  
Key Columns: order\_id, customer\_id, order\_status, order\_purchase\_timestamp.

- Order items: Contains line-item details, including shipping costs and seller links.  
Key Columns: order\_id, product\_id, seller\_id, price, freight\_value
- Customers: Differentiates between a single order ID and a unique customer identity.  
Key Columns: customer\_id, customer\_unique\_id, customer\_city, customer\_state.
- Sellers: Data on the origin of the products.  
Key Columns: seller\_id, seller\_city, seller\_state
- Products: Catalog details for categorization.  
Key Columns: product\_id, product\_category\_name
- Payments: Details on how the customer paid and the total transaction value.  
Key Columns: order\_id, payment\_type, payment\_installments, payment\_value
- Reviews: Post-purchase feedback used for sentiment analysis.  
Key Columns: review\_id, order\_id, review\_score, review\_comment\_message
- Product category name translation: Maps original Portuguese product category names to their description in English equivalent  
Key Columns: product\_category\_name, product\_category\_name\_english
- Geolocation: Provides latitude and longitude coordinates mapped to Brazilian zip code prefixes.  
Key Columns: geolocation\_zip\_code\_prefix, geolocation\_lat, geolocation\_lng, geolocation\_city, geolocation\_state.

## **2.3 Data Cleaning and Transformation (ETL - Transformation)**

The raw data, consisting of 9 separate CSV files, required extensive cleaning and feature engineering in Python (Pandas/NumPy) and Power Query to prepare for modelling.

- Date/Time Conversion: Converted all timestamp columns (e.g., order\_purchase\_timestamp, order\_delivered\_customer\_date) from object/string type to datetime objects. This enables time-series analysis and date-based calculations (like lead time).
- Category Translation: Merged the product category name translation table with the products table to convert Portuguese product categories to English, ensuring the dashboard is universally understandable.
- Missing Value Handling Reviews: For missing review\_score values, these records were typically for canceled or undelivered orders,

they were kept but excluded from average score calculations. Product Metrics: Missing physical dimensions/weight were often imputed with the median for relevant calculation columns or ignored, as they are not critical for revenue analysis.

- Outlier Management Order Status: Filtered out orders with order\_status other than 'delivered' to focus on completed sales and reliable delivery metrics.
- Feature Engineering: Created new analytical columns such as: Delivery Delay, Average order value, Unit Sold, Total Customers, New Customers, Returning customers, Repeat Purchase Rate, customer type, Average Review Score, Seller Revenue, Average seller Delivery time, Average selling rating.
- Geolocation Aggregation: Consolidated geolocation data by ZIP code prefix to optimize mapping performance and regional analysis.

## **2.4 Data Merging and Modelling (ETL - Loading & Modelling)**

The 9 relational datasets were merged to create a single analytical table structure suitable for a star schema in Power BI. The merging process was conducted primarily using Left Joins on common key columns to ensure all orders remained, while enriching them with related details.

## **2.5 Exploratory Data Analysis**

EDA was conducted to uncover underlying patterns and validate hypotheses before dashboard construction. The primary focus areas included:

- Revenue Distribution: Mapping how revenue flows across various product categories.
- Seasonality: Identifying monthly and daily sales peaks (e.g., Black Friday impact).
- Purchase Frequency: Analyzing the behavior of unique customers over time.
- Logistics Patterns: Investigating bottlenecks in the delivery chain and frequency of delays.
- Correlation Analysis: Quantifying the relationship between delivery performance and customer review scores.
- Financial Behavior: Studying payment method preferences and the use of installments.

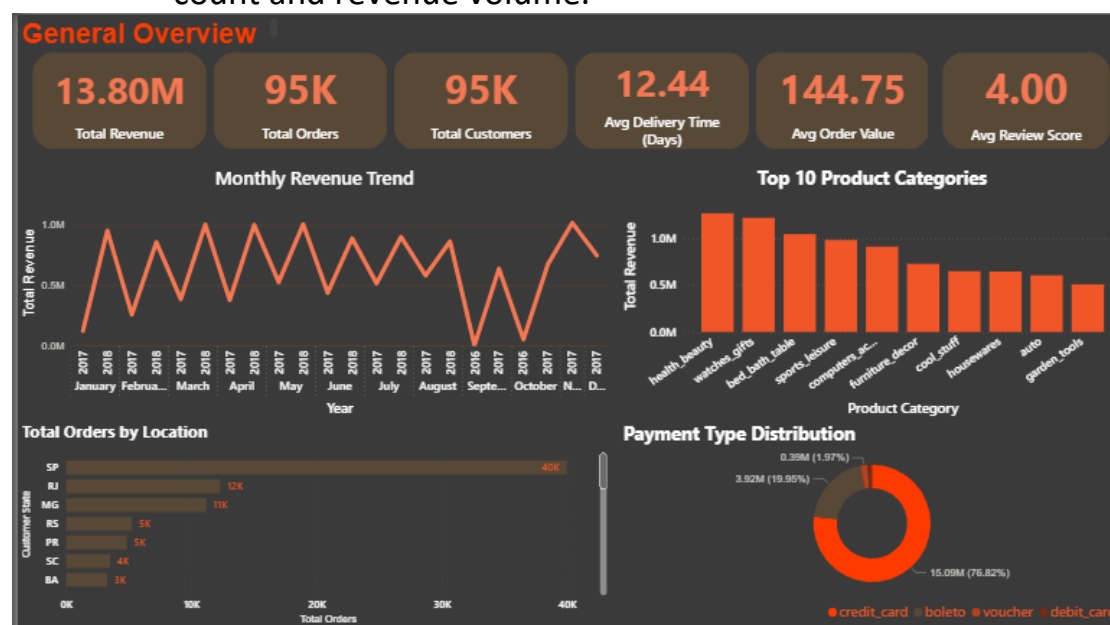
Insights from the EDA directly guided the Power BI dashboard design, ensuring that the most critical KPIs, such as the correlation between

shipping delays and low ratings, were placed prominently for stakeholders.

## 3. KEY FINDINGS

### 3.1 Sales & Revenue: High Concentration and Seasonality

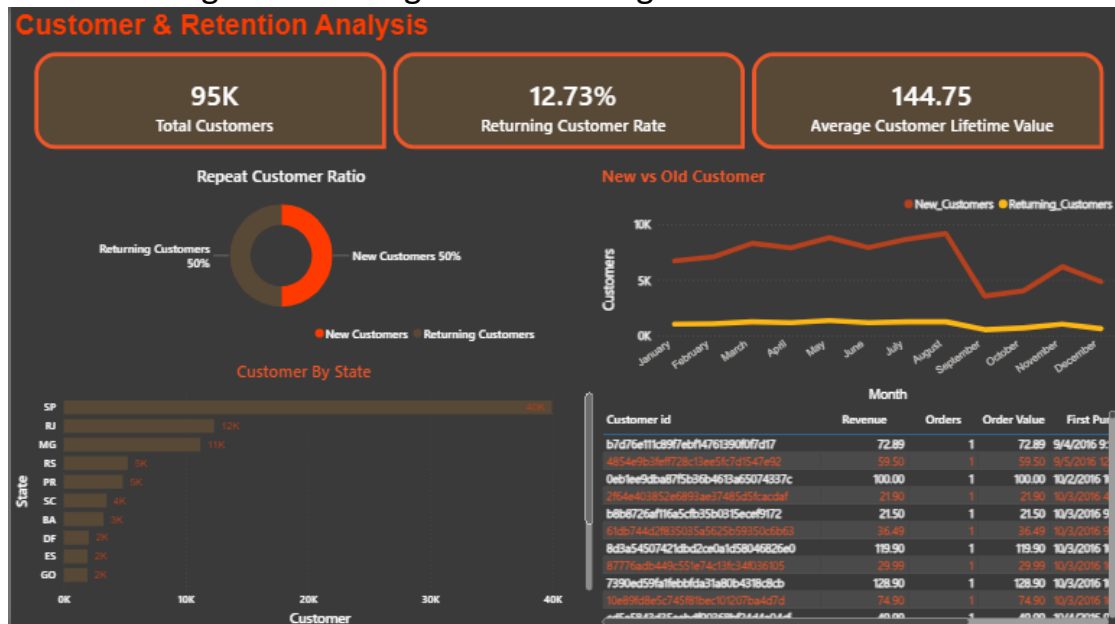
- Revenue Growth shows seasonal peaks: Peak sales consistently occur in November (Black Friday), with a notable dip in the subsequent months. The marketplace exhibits year-over-year growth, but is prone to strong seasonal volatility.
- Small subset of categories drives revenue: The Top 5 product categories (e.g., Health & Beauty, Watches, Sports & Leisure) often account for over 40% of total revenue. This concentration poses a business risk if one category faces disruption.
- Geographic Revenue Skew: Sales are heavily dominated by the Southeast region, especially the state of São Paulo (SP), which represents the largest market share by both customer count and revenue volume.



### 3.2 Customer Behavior & Retention: Acquisition vs. Loyalty

- Majority of customers are first-time buyers: Over 90% of unique customer IDs (customer\_unique\_id) have only one purchase record. This is a critical indicator of a high churn rate and weak customer lifetime value.
- Repeat purchase rate is relatively low: The calculated repeat purchase rate is low. This highlights the platform's strength in acquisition but severe weakness in retention.

- Repeat buyers are high-value: Although few in number, the repeat customers have a significantly higher Average Order Value (AOV) than first-time buyers, making them a crucial segment for targeted marketing.



### 3.3 Seller Performance & Customer Experience

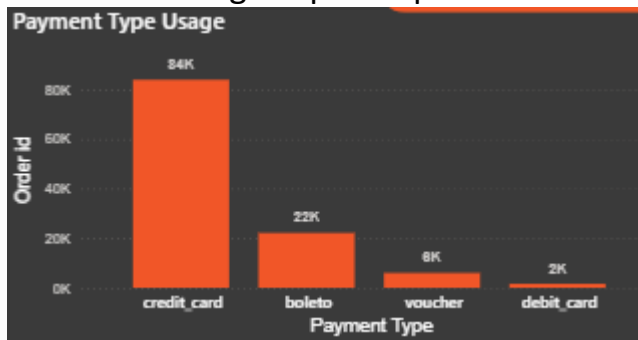
- Faster delivery times correlate with higher customer ratings: Orders delivered faster than the estimated date achieve an average review score of approximate 4.8, while delayed orders typically fall to approximate 2.5. Delivery speed is the single most controllable factor impacting satisfaction.
- High seller revenue does not guarantee high review scores: Some top revenue-generating sellers may have average or even poor review scores due to high volume masking operational issues. A balanced metric is needed.

### 3.4 Logistics & Payments

- Delivery delays vary by region: Remote Northern and Northeastern states experience average delivery times that are 50% longer than the Southeast, leading to a higher concentration of 1-star reviews in those regions.



- Credit cards dominate payment methods: Credit cards account for the highest percentage of transactions and total value. The use of installment payments is strongly associated with higher-priced products.



## 4. Recommendations for Business Improvement

### 4.1 Operational Improvements

- Seller Tiering: Implement a "Gold/Silver/Bronze" seller ranking system based on the calculated Seller Score (Revenue + Speed + Review Score).
- Investigate logistics partners or establish regional hubs in the high-delay Northern/Northeastern states to close the gap between actual and estimated delivery times

### 4.2 Marketing & Retention

- Loyalty Incentives: Launch a "Repeat Buyer" discount program to address the low retention rate found in the data. Target the approximately 90 of first-time buyers with a compelling "Welcome Back" offer 30-45 days after their first purchase to convert them into repeat customers.
- Payment Optimization: Leverage the insight that installments drive high-value orders by actively promoting installment options for new product categories to help them break into the top revenue-generating tiers.

### 4.3 Customer Experience

- Automated Feedback Loop: Use Python-based sentiment analysis on "Review Comments" to flag negative reviews for immediate

customer service follow-up before they impact the seller's long-term rating.

## **5. CONCLUSION**

The marketplace demonstrates strong sales performance and generally positive customer sentiment. However, improvements in customer retention, delivery efficiency, and seller quality management could significantly enhance long-term performance. Evaluating sellers through a combination of revenue, delivery reliability, and customer feedback provides a more balanced view of performance.