

2. DATASET

2.1 Overview of the Olist Brazilian Ecommerce Dataset

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

To illustrate this project, we are employing a sample comprehensive dataset from Kaggle, posted by Olist: a Brazilian e-commerce platform (real data, anonymized) that connects sellers with marketplace and their own e-commerce channels. The following is an overview of the dataset:

2.1.1 Dataset Structure & Size

The dataset consists of **9 CSV files** containing **1.5+ million total records**:

- **olist_orders_dataset.csv**: 99,442 orders
- **olist_customers_dataset.csv**: 99,442 customers
- **olist_order_items_dataset.csv**: 112,651 order items (some orders have multiple items)
- **olist_order_payments_dataset.csv**: 103,887 payment records
- **olist_order_reviews_dataset.csv**: 104,720 customer reviews
- **olist_products_dataset.csv**: 32,952 products
- **olist_sellers_dataset.csv**: 3,096 sellers
- **olist_geolocation_dataset.csv**: 1,000,164 location records (geospatial data)
- **product_category_name_translation.csv**: 71 category translations

2.1.2 Temporal Coverage

- **Date Range**: September 2016 - October 2018 (approximately 25 months)
- Most active period appears to be 2017-2018

2.1.3 Key Data Tables & Relationships

1. Orders Table (olist_orders_dataset.csv)

- Order lifecycle tracking: purchase → approval → carrier delivery → customer delivery
- Order statuses: 96,478 delivered, 1,107 shipped, 625 canceled, etc.
- Key timestamps for delivery analysis

2. Customer Information (olist_customers_dataset.csv)

- Customer locations (city/state/ZIP)
- Unique customer IDs for privacy/anonymization
- Geographic distribution across Brazil

3. Order Items (`olist_order_items_dataset.csv`)

- Product prices and shipping costs
- Links products to sellers and orders
- Supports multi-item orders

4. Products (`olist_products_dataset.csv`)

- Product categories (71 categories, Portuguese names with English translations)
- Product dimensions and weight
- Product descriptions and photo counts

5. Sellers (`olist_sellers_dataset.csv`)

- Seller locations and distribution
- Marketplace dynamics analysis

6. Payments (`olist_order_payments_dataset.csv`)

- Payment methods: credit card, boleto, etc.
- Installment information
- Payment sequencing for multi-payment orders

7. Reviews (`olist_order_reviews_dataset.csv`)

- Customer satisfaction scores (1-5 scale)
- Review comments and titles
- Response timestamps from sellers

8. Geolocation (`olist_geolocation_dataset.csv`)

- Latitude/longitude coordinates
- ZIP code to location mapping
- Enables geospatial analysis

2.1.4 Opportunities Analysis

This dataset is excellent for:

1. **Customer Behavior Analysis:** Purchase patterns, repeat customers, geographic preferences
2. **Product Performance:** Best-selling categories, pricing strategies, seasonal trends
3. **Seller Performance:** Sales volume, delivery times, customer satisfaction
4. **Logistics & Delivery:** Delivery performance, shipping costs, geographic challenges
5. **Payment Analysis:** Preferred payment methods, installment trends
6. **Market Basket Analysis:** Product associations and recommendations
7. **Geospatial Analysis:** Regional performance, delivery optimization
8. **Time Series Analysis:** Seasonal patterns, growth trends

2.1.5 Data Quality Notes

- Well-structured relational database design
- Consistent data types and foreign key relationships
- Geographic coverage across Brazil's diverse regions
- Rich temporal data for time-based analysis
- Both quantitative metrics and qualitative review data

This is a robust enough dataset for our purposes, offering a comprehensive foundation for e-commerce analytics. It covers the entire customer journey from browsing to post-purchase reviews. The Brazilian market context adds interesting cultural and regional dimensions to the analysis which might enrich the findings of our analysis.