6.1 Sourcing Open Data

By. Amani Abdelwahab

Data Source:

The dataset is sourced from Kaggle. it is a Brazilian ecommerce public dataset of orders made at Olist Store. The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allow viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers.

The primary purpose of this dataset is to analyse factors that influence sales performance, customer satisfaction, and delivery efficiency. It helps identify patterns in order value, customer reviews, and delivery times across different regions, supporting data-driven decisions to improve logistics, customer service, and profitability.

The data can be access here:

Brazilian E-Commerce Public Dataset by Olist

Data Collection:

The dataset was collected from Olist, a Brazilian e-commerce platform that connects small businesses to major online marketplaces. The data represents real transactions made between 2016 and 2018 and includes various aspects of the e-commerce process, from order placement to customer reviews.

The data was gathered from multiple sources within Olist's platform, including:

- Order Information: Details of orders, including product categories, payment methods, and order status.
- Customer Data: Anonymized customer locations and purchase behaviour.
- Seller Data: Information about sellers, their locations, and performance.
- Geographical Information: Customer and seller locations by city and state.
- Time Variables: Order purchase dates, shipping estimates, and delivery dates.

The dataset is publicly available on Kaggle and adheres to open-source data standards, making it accessible for educational and analytical purposes.

Data Limitations:

- Anonymized Data: No personal details of customers or sellers.
- Geographic Scope: Limited to Brazil.

- Time Frame: Data from 2016-2018 may not reflect current trends.
- Missing Values: Some delivery dates, reviews, and product categories are incomplete.
- Product Details: Limited descriptions and broad categories.
- No Marketing Data: Lacks information on campaigns or promotions.
- External Factors: Delivery performance is influenced by factors not captured in the data.

Why this data

It is well-suited for my project because:

- Covers a Commercial Domain: It focuses on e-commerce, a highly relevant industry for business analysis.
- Rich Dataset: Includes sales, customer reviews, payments, delivery, and seller performance data.
- Geographical Component: Contains location-based data (Brazilian states and cities) for geographic analysis.
- Meets Project Requirements: Includes continuous variables (prices, payment values, delivery time) and categorical variables (product categories, payment types, customer locations).
- Sufficient Data Volume: Over 100,000 rows, exceeding the 2,000-row requirement.
- Real-World Insights: Enables analysis of sales trends, customer behavior, and logistics performance.
- Open-Source: Freely available on Kaggle for transparent, unbiased exploration.

This dataset provides ample opportunities for data cleaning, visualization, and hypothesis testing, making it a great choice for your project.

Ethical Considerations

- Privacy Protection: Customer and seller identities are anonymized to safeguard privacy.
- Data Usage: Ensure the data is used for educational and analytical purposes only.
- Bias Awareness: Be mindful of potential regional or demographic biases since data is limited to Brazil.
- Transparency: Clearly communicate any assumptions or data transformations made during analysis.
- Fair Representation: Avoid misrepresenting businesses or customers based on partial or incomplete data.
- Respect for Intellectual Property: Acknowledge Kaggle and Olist as data sources.

These considerations ensure ethical and responsible use of the dataset.

Data profiling summery:

1. Dataset Overview

- Source: Kaggle (Open source)
- Domain: E-commerce (Commercial field)
- Geographical Scope: Brazil (Nationwide, regional data available)
- Time Frame: 2016 - 2018
- Rows: +100,000 (meets 2,000-row minimum)
- Key Tables: Orders, Products, Customers, Reviews, Payments, and Geolocation

2. Key Variables

- Categorical Variables (3+):
  - `product_category_name` (Product category)
  - `payment_type` (Credit card, boleto, voucher, etc.)
  - `customer_state` (Geographic region)
- Continuous Variables (3+):
  - `payment_value` (Payment amount)
  - `freight_value` (Shipping cost)
  - `product_weight_g` (Product weight in grams)
- Time-dependent Variables:
  - `order_purchase_timestamp` (Order date)
  - `order_delivered_customer_date` (Delivery date)

3. Main Tables and Columns:

1. **Orders:**
   - `order_id` (Unique identifier)
   - `customer_id` (Customer identifier)
   - `order_status` (Delivered, shipped, canceled, etc.)
   - `order_purchase_timestamp` (Purchase date)
   - `order_delivered_customer_date` (Delivery date)
2. **Products:**
   - `product_id` (Unique product identifier)
   - `product_category_name` (Product category)
   - `product_weight_g`, `product_length_cm`, `product_height_cm`, `product_width_cm` (Product dimensions)
3. **Customers:**
   - `customer_id` (Unique customer identifier)
   - `customer_city`, `customer_state` (Geographic location)
4. **Reviews:**
   - `review_id` (Unique review identifier)

  - `review_score` (Rating from 1 to 5)
  - `review_comment_message` (Customer feedback)

5. **Payments:**
   - `payment_sequential` (Sequential payment number)
   - `payment_type` (Credit card, boleto, voucher, etc.)
   - `payment_installments` (Number of installments)
   - `payment_value` (Payment amount)
6. **Geolocation:**
   - `geolocation_zip_code_prefix` (Postal code prefix)
   - `geolocation_lat`, `geolocation_lng` (Latitude and longitude)
7. **Sellers:**
   - `seller_id` (Unique seller identifier)
   - `seller_city`, `seller_state` (Seller location)

Data Cleaning

1. **Handling Missing Values:**
2. **Dropping Irrelevant Columns:**
3. **Optimizing Data Types:**
4. **Data Integrity:**
5. **Saving the Cleaned Dataset**
6. **Descriptive Analysis:**

A descriptive analysis and data cleaning performed in a Jupyter notebook.

Questions to Explore:

1. **Sales Performance:**
   - What are the monthly sales trends, and do they vary across regions?
   - Which product categories generate the highest revenue?
   - Why do certain months show higher or lower sales?
   - Why do specific product categories outperform others in revenue?
2. **Customer Behaviour:**
   - How do customer reviews affect repeat purchases?
   - Is there a correlation between delivery time and customer satisfaction?
   - Why do some customers leave negative reviews despite timely delivery?
   - Why do repeat customers prefer specific product categories?
3. **Logistics Efficiency:**
   - What is the average delivery time across different regions?
   - How do delivery delays impact customer reviews and overall sales?
   - Why do delivery delays occur more frequently in certain regions?

    - Why do orders from specific states take longer to deliver?

4. **Payment Insights:**
    - What is the most common payment method, and how does it impact order value?
    - Is there a relationship between instalment payments and higher purchase amounts?
    - Why do customers choose instalment payments over single payments?
    - Why do certain payment methods lead to higher average order values?

1. **Geographical Analysis:**
    - Which regions have the highest number of orders?
    - How does location influence delivery speed and customer satisfaction?
    - Why do customers from certain regions leave higher reviews?
    - Why do urban areas have faster delivery times compared to rural areas?