

**Brazilian E-Commerce Public Dataset**

**ROBERT DAVIES**

1. **DATA SUMMARY**

**Data Sourcing**

This is an internal dataset of orders made at Olist Store from 2016 to 2018 made publicly available by Olist, the operator of an e-commerce site in Brazil. It is real commercial data that has been anonymized for privacy purposes. It is trustworthy data as it has been released by the data owner. The data was directly sourced from Kaggle at the link below:

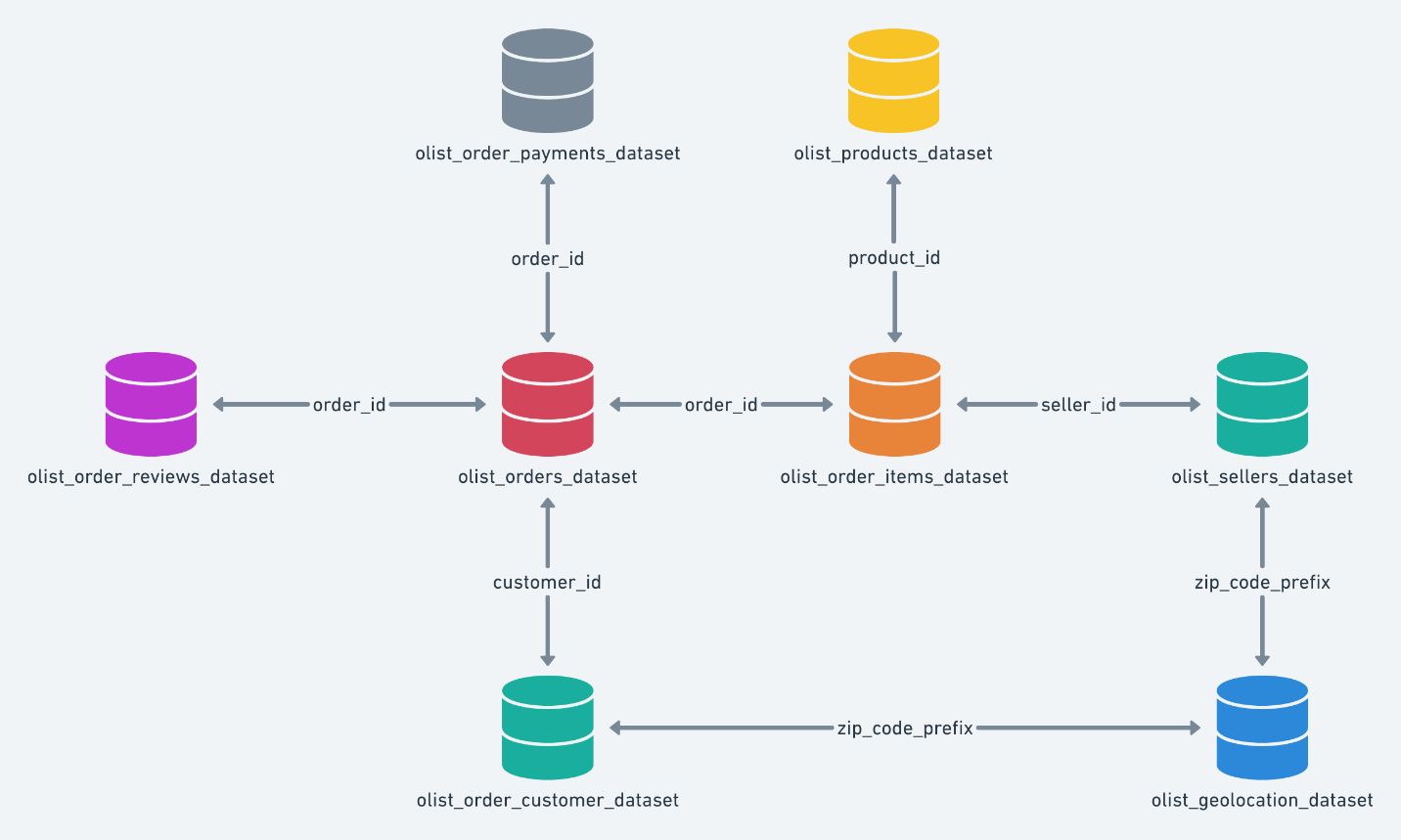
Brazilian E-Commerce Public Dataset

**Data Collection**

The data is administrative, collected by Olist from its internal data systems. Olist connects small businesses from all over Brazil to sales channels. Those merchants sell their products through the Olist Store and ship them directly to customers using Olist logistics partners. After a customer purchases a product from Olist Store, a seller is notified to fulfill that order. After a sale, a satisfaction survey is e-mailed to the customer where they can give a rating for the purchase experience and write down some comments. Data relating to each of these steps in the sales process has been made publicly available by Olist for non-commercial purposes.

**Data Contents**

The dataset is divided into multiple tables for cleaner organization. It contains information relating to about 100k customer orders from 2016 to 2018 made at the Olist store website across Brazil. Its features allow viewing of details of orders from multiple dimensions including product, price, freight costs, payment total, seller that fulfilled each order, customer location, and even text of customer reviews. The data schema is shown below (excluding a product category translation table):



**Data Relevance**

The dataset meets the requirements of the project brief including coming from an authoritative source, containing multiple data types, and including a geographical component. I chose this dataset because it is consistent with my personal objective of applying data analysis tools and techniques to a couple of specific use cases:

1. Forecasting future sales performance.
2. Generating data insights and recommendations to improve business performance.
3. **DATA PROFILE OF RAW DATA**
4. **Data Tables**

**Table 1: orders**

The core data about each order.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **time-variant/**  **-invariant** | **structured/ unstructured** | **qualitative/**  **quantitative** | **nominal/ordinal/binary**  **discrete/continuous** |
| order\_id | time-invariant | semi-structured | qualitatative | nominal |
| customer\_id | time-invariant | semi-structured | qualitatative | nominal |
| order\_status | time-variant | structured | qualitatative | ordinal |
| order\_purchase\_  timestamp | time\_variant | structured | quantitative | continuous |
| order\_approved\_at | time\_variant | structured | quantitative | continuous |
| order\_delivered\_  carrier\_date | time\_variant | structured | quantitative | continuous |
| order\_delivered\_  customer\_date | time\_variant | structured | quantitative | continuous |
| order\_estimated\_  delivery\_date | time\_variant | structured | quantitative | discrete |

**Table 2: order\_reviews**

Data about the reviews made by customers in relation to their orders.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **time-variant/**  **-invariant** | **structured/ unstructured** | **qualitative/**  **quantitative** | **nominal/ordinal/binary**  **discrete/continuous** |
| review\_id | time-invariant | semi-structured | qualitatative | nominal |
| order\_id | time-invariant | semi-structured | qualitatative | nominal |
| review\_score | time-invariant | structured | quantitative | discrete |
| review\_comment\_title | time-invariant | unstructured | qualitatative | nominal |
| review\_comment\_  message | time-invariant | unstructured | qualitatative | nominal |
| review\_creation\_date | time\_variant | structured | quantitative | discrete |
| review\_answer\_  timestamp | time\_variant | structured | quantitative | continuous |

**Table 3: order\_payments**

Data about the payment value and options for orders.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **time-variant/**  **-invariant** | **structured/ unstructured** | **qualitative/**  **quantitative** | **nominal/ordinal/binary**  **discrete/continuous** |
| order\_id | time-invariant | semi-structured | qualitatative | nominal |
| payment\_sequential | time-invariant | structured | quantitative | discrete |
| payment\_type | time-invariant | structured | qualitatative | nominal |
| payment\_installments | time-invariant | structured | quantitative | discrete |
| payment\_value | time\_variant | structured | quantitative | continuous |

**Table 4: customers**

Data about Olist customers and their location.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **time-variant/**  **-invariant** | **structured/ unstructured** | **qualitative/**  **quantitative** | **nominal/ordinal/binary**  **discrete/continuous** |
| customer\_id | time-invariant | semi-structured | qualitatative | nominal |
| customer\_unique\_id | time-invariant | semi-structured | qualitatative | nominal |
| customer\_zip\_code\_  prefix | time-invariant | structured | qualitatative | nominal |
| customer\_city | time-invariant | structured | qualitatative | nominal |
| customer\_state | time-invariant | structured | qualitatative | nominal |

**Table 5: products**

Data about the products sold by Olist.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **time-variant/**  **-invariant** | **structured/ unstructured** | **qualitative/**  **quantitative** | **nominal/ordinal/binary**  **discrete/continuous** |
| product\_id | time-invariant | semi-structured | qualitatative | nominal |
| product\_category\_name | time-invariant | structured | qualitatative | nominal |
| product\_name\_length | time-invariant | structured | quantitative | discrete |
| product\_description\_  length | time-invariant | structured | quantitative | discrete |
| product\_photos\_qty | time-variant | structured | quantitative | discrete |
| product\_weight\_g | time\_invariant | structured | quantitative | discrete |
| product\_length\_cm | time\_invariant | structured | quantitative | discrete |
| product\_height\_cm | time\_invariant | structured | quantitative | discrete |
| product\_width\_cm | time\_invariant | structured | quantitative | discrete |

**Table 6: order\_items**

Data about the items purchased within each order.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **time-variant/**  **-invariant** | **structured/ unstructured** | **qualitative/**  **quantitative** | **nominal/ordinal/binary**  **discrete/continuous** |
| order\_id | time-invariant | semi-structured | qualitatative | nominal |
| order\_item\_id | time-invariant | structured | quantitative | discrete |
| product\_id | time-invariant | semi-structured | qualitatative | nominal |
| seller\_id | time-invariant | semi-structured | qualitatative | nominal |
| shipping\_limit\_date | time\_variant | structured | quantitative | continuous |
| price | time\_variant | structured | quantitative | continuous |
| freight\_value | time\_variant | structured | quantitative | continuous |

**Table 7: sellers**

Data about the sellers that fulfilled orders made at Olist.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **time-variant/**  **-invariant** | **structured/ unstructured** | **qualitative/**  **quantitative** | **nominal/ordinal/binary**  **discrete/continuous** |
| seller\_id | time-invariant | semi-structured | qualitatative | nominal |
| seller\_zip\_code\_  prefix | time-invariant | structured | qualitatative | nominal |
| seller\_city | time-invariant | structured | qualitatative | nominal |
| seller\_state | time-invariant | structured | qualitatative | nominal |

**Table 8: geolocation**

Data about Brazilian zip codes and their latitude/longitude coordinates.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **time-variant/**  **-invariant** | **structured/ unstructured** | **qualitative/**  **quantitative** | **nominal/ordinal/binary**  **discrete/continuous** |
| geolocation\_ zip\_code\_  prefix | time-invariant | structured | qualitatative | nominal |
| geolocation\_ lat | time-invariant | structured | quantitative | continuous |
| geolocation\_ lng | time-invariant | structured | quantitative | continuous |
| geolocation\_city | time-invariant | structured | qualitatative | nominal |
| geolocation\_state | time-invariant | structured | qualitatative | nominal |

**Table 9: product\_category\_name\_translation**

Portuguese to English translation of Olist product name categories.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **time-variant/**  **-invariant** | **structured/ unstructured** | **qualitative/**  **quantitative** | **nominal/ordinal/binary**  **discrete/continuous** |
| product\_category\_  name | time-invariant | structured | qualitatative | nominal |
| product\_category\_  name\_english | time-invariant | structured | qualitatative | nominal |

1. **Data Checks**

Basic data cleaning and consistency checks, and descriptive statistical analysis performed in Jupyter in the following notebook: 6.1 Olist exploratory analysis.ipynb

1. **Data Limitations and Ethics**

Since the dataset was open source with a license, I have no ethical concerns relating to how it was sourced. A review of the dataset confirmed that all personally identifiable customer and seller data has been anonymized, therefore I have no data privacy concerns relating to its usage. The review of the data didn’t raise any equality concerns or potential bias due to the nature of the dataset as the specifics of actual sales transactions via a well-known Brazilian e-commerce company.

The only potential limitation is that the dataset is internal, with no way for an external party to validate accuracy of the order details or to be certain that it is a complete dataset and not just a partial set of orders that Olist made publicly available. Organizations are usually motivated to share data that makes them look good. However, there is no specific reason to distrust the accuracy of the data released by Olist.

1. **EXPLORATORY QUESTIONS**

There are many questions I have relating to different aspects of the dataset. Below are a few and I will choose to focus on one area for the purposes of this project once I have conducted some further preliminary analysis.

**Sales Prediction**

* What will monthly total sales be for the Olist platform in the remainder of 2018?
* What are the trends in regional sales on the Olist platform?

**Order Fulfilment Performance**

* Is there a correlation between delivery time and customer satisfaction?
* Are sellers or logistics partners primarily responsible for late deliveries?
* Which sellers are low performers in terms of late deliveries?
* Are there specific regions or product categories with much higher frequency of late deliveries?
* What are the trends in delivery performance over time?

**Product Performance**

* Which product categories are experiencing the fastest growth?
* Which product categories have higher customer dissatisfaction?
* Which product categories have higher order cancellation rates?

**Customer Base**

* What is the concentration of orders among top customers?
* What is the level of customer satisfaction among top customers?
* Is there a relationship between freight costs and customer satisfaction?

**Payment Options**

* Does payment in instalments help with higher customer satisfaction?
* Do orders with instalment payments have higher values than other orders?
* What are the trends in payment types used by customers over time?