Olist Dataset Documentation

Gathered by : Saeed Shirini

Olist Dataset Documentation

About Dataset

Brazilian E-Commerce Public Dataset by Olist

Welcome! This is a Brazilian e-commerce public dataset of orders made at Olist Store. The dataset has information on 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allow viewing orders from multiple dimensions: from order status, price, payment, and freight performance to customer location, product attributes, and finally reviews written by customers. We also released a geolocation dataset that relates Brazilian zip codes to lat/lng coordinates.

This is real commercial data, it has been anonymized, and references to the companies and partners in the review text have been replaced with the names of Game of Thrones great houses.

Join it With the Marketing Funnel by Olist

We have also released a Marketing Funnel Dataset. You may join both datasets and see an order from a Marketing perspective now!

Instructions on joining are available on this Kernel.

Context

This dataset was generously provided by Olist, the largest department store in Brazilian marketplaces. Olist connects small businesses from all over Brazil to channels without hassle and with a single contract. Those merchants are able to sell their products through the Olist Store and ship them directly to the customers using Olist logistics partners. See more on our website: www.olist.com

After a customer purchases the product from Olist Store a seller gets notified to fulfill that order. Once the customer receives the product, or the estimated delivery date is due, the customer gets a satisfaction survey by email where he can give a note for the purchase experience and write down some comments.

Attention

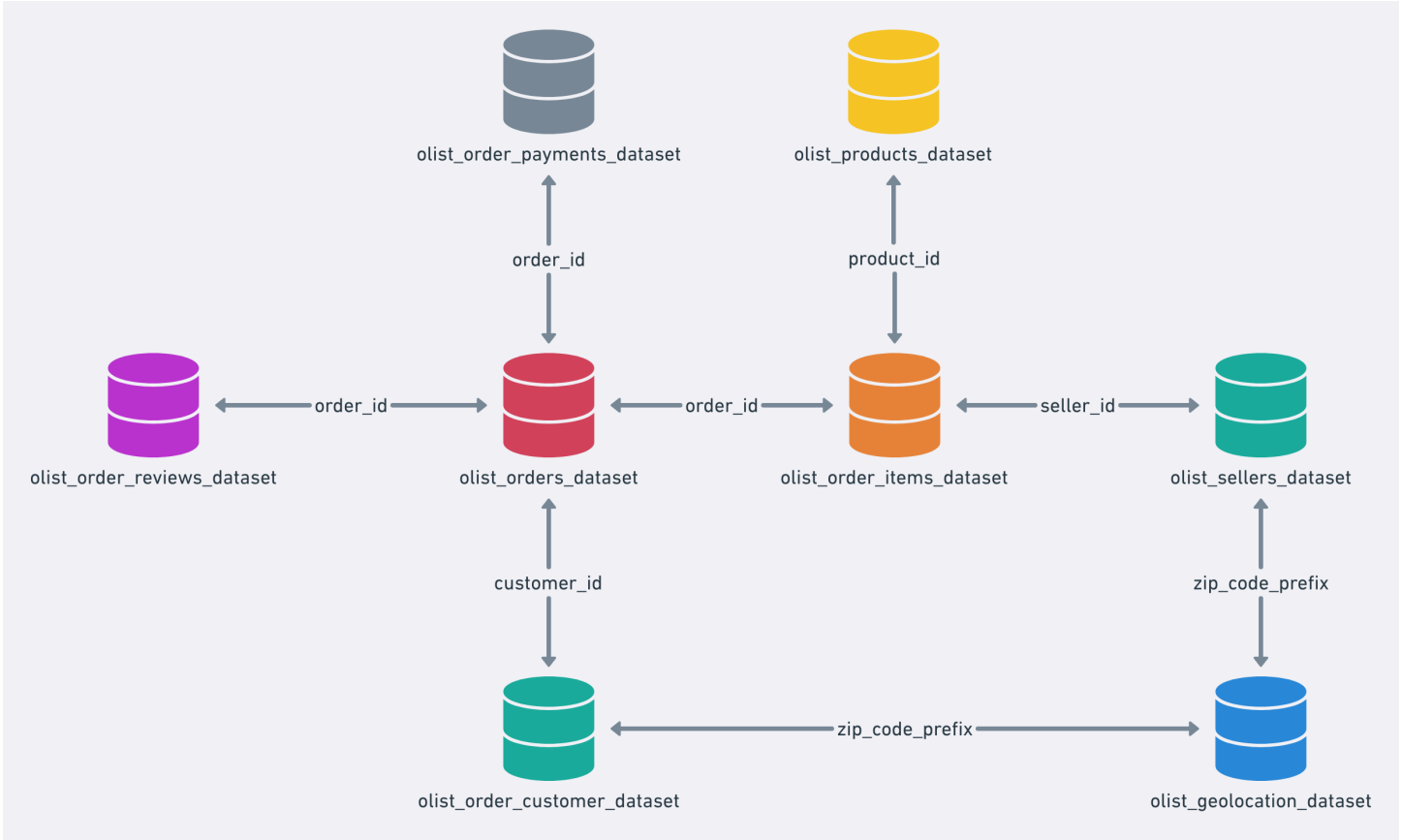
1. An order might have multiple items.

2. Each item might be fulfilled by a distinct seller.
3. All text identifying stores and partners where replaced by the names of Game of Thrones great houses.

Example of a product listing on a marketplace

Data Schema

The data is divided into multiple datasets for better understanding and organization. Please refer to the following data schema when working with it:



Classified Dataset

We had previously released a classified dataset, but we removed it in *Version 6*. We intend to release it again as a new dataset with a new data schema. While we haven't finished it, you may use the classified dataset available at *Version 5* or previous.

Inspiration

Here are some inspirations for possible outcomes from this dataset.

NLP:

This dataset offers a supreme environment to parse out the review text through its multiple dimensions.

Clustering:

Some customers didn't write a review. But why are they happy or mad?

Sales Prediction:

With purchase date information you'll be able to predict future sales.

Delivery Performance:

You will also be able to work through delivery performance and find ways to optimize delivery times.

Product Quality:

Enjoy yourself discovering the product categories that are more prone to customer dissatisfaction.

Feature Engineering:

Create features from this rich dataset or attach some external public information to it.

Acknowledgments

Thanks to Olist for releasing this dataset.

Customer Dataset

About This file

This dataset has information about the customer and its location. Use it to identify unique customers in the orders dataset and to find the order's delivery location.

In our system, each order is assigned to a unique customer_id. This means that the same customer will get different IDs for different orders. The purpose of having a customer_unique_id on the dataset is to allow you to identify customers that made repurchases at the store. Otherwise, you would find that each order had a different customer associated with it.

Geolocation Dataset

This dataset has information on Brazilian zip codes and their lat/lng coordinates. Use it to plot maps and find distances between sellers and customers.

Order Items Dataset

This dataset includes data about the items purchased within each order.

Example:

The order_id = 00143d0f86d6fbd9f9b38ab440ac16f5 has 3 items (same product). Each item has the freight calculated according to its measures and weight. To get the total freight value for each order you just have to sum.

The total order_item value is: 21.33 * 3 = 63.99

The total freight value is: 15.10 * 3 = 45.30

The total order value (product + freight) is: 45.30 + 63.99 = 109.29

Payments Dataset

This dataset includes data about the orders payment options.

Order Reviews Dataset

This dataset includes data about the reviews made by the customers.

After a customer purchases the product from Olist Store a seller gets notified to fulfill that order. Once the customer receives the product, or the estimated delivery date is due, the customer gets a satisfaction survey by email where he can give a note about the purchase experience and write down some comments.

Order Dataset

This is the core dataset. From each order, you might find all other information.

Products Dataset

This dataset includes data about the products sold by Olist.

Sellers Dataset

This dataset includes data about the sellers who fulfilled orders made at Olist. Use it to find the seller's location and to identify which seller fulfilled each product.

Category Name Translation

Translates the product_category_name to English.