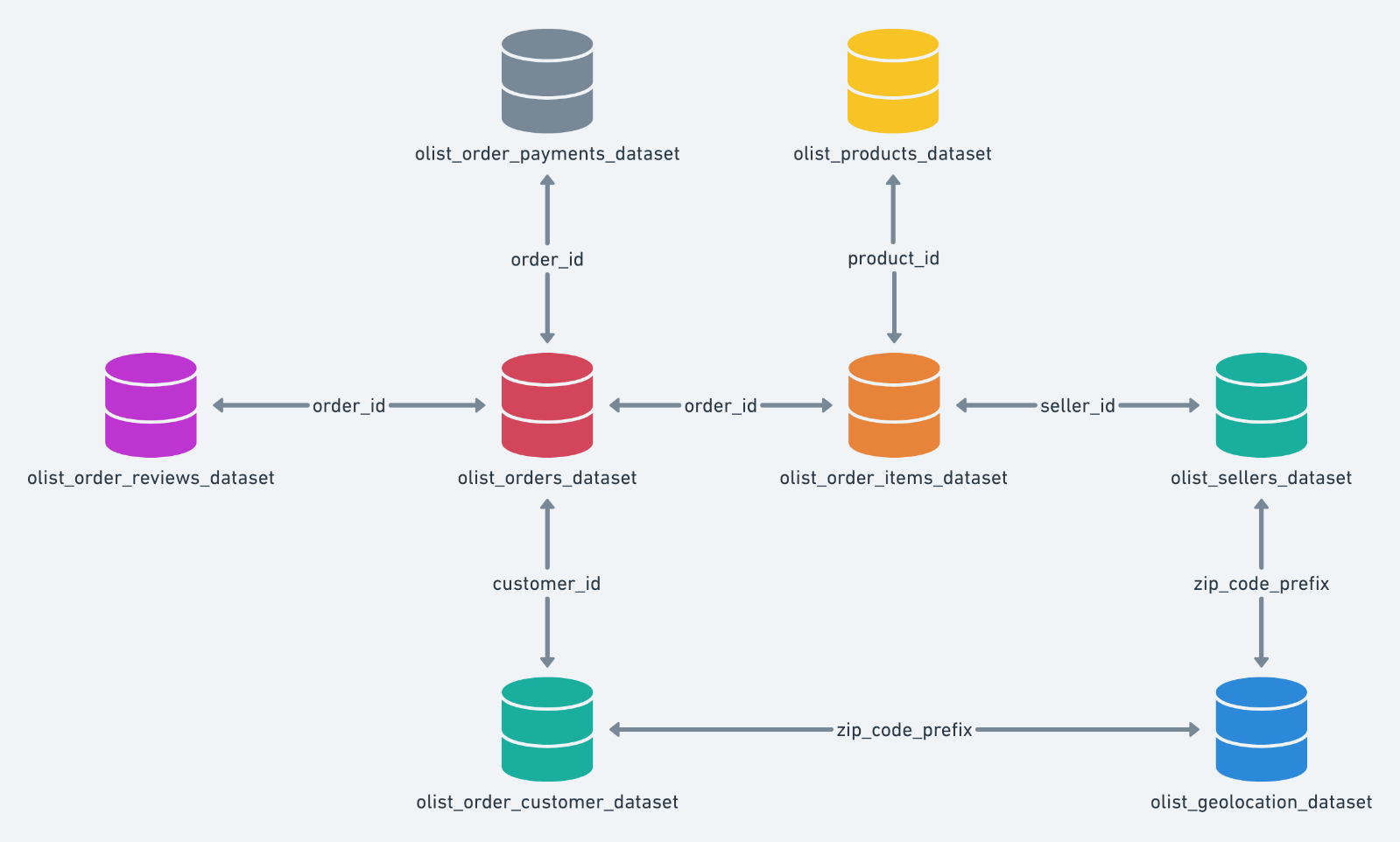
|  |
| --- |
| Close-up image showing the leaf-sides of two oversized books side-by-side on a bookshelf, with additional books in soft focus background |
| **Brazilian E-Commerce Public Dataset by Olist** |
| |  |  | | --- | --- | |  |  | |
|  |
|  |
|  |
| Ankit Narang |

**INTRODUCTION**

This 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 allows 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.

For the purpose of our analysis, we will start with merging the necessary files as there are separate tables that contain information on different aspects of the transactions. Most of these files would be essential to our analysis apart from a few which tend to cater more towards the logistics and supply chain departments.

The structure of the data schema as provided by Olist can be seen below.



**Data Quality Assessment**

1. **olist\_customers\_dataset.csv –** This dataset consists of the customers’ demographics data like customers’ zip code, city and state. Also, it includes the customer id and customer unique id. The data quality issues are as follows:
   1. The customer\_id and customer\_unique\_id are very long which might take more computational resources. Hence, we need to decide how it should be treated.
   2. The zip\_code\_prefix provided are not the real zip codes of the area in which the customers are residing.

**Findings:**

* **Around 42 % of the customers lives in the Sao Paulo state.**
* **There are 99441 customers in total in the dataset.**

1. **olist\_sellers\_dataset.csv –** This dataset includes the demographics data of the sellers similar to the customer’s dataset. It has seller\_zip\_code\_prefix, seller\_city, seller\_state and seller\_id. Please find the quality issues and commentsvelow:
   1. The seller\_id is very long which might take more computational resources. Hence, we need to decide how it should be treated.
   2. The zip\_code\_prefix provided are not the real zip codes of the area in which the sellers are located.

**Findings:**

* **Around 60% of the sellers lives in the Sao Paulo state.**
* **There are 3095 sellers in total in the dataset.**

1. **olist\_products\_dataset.csv –** This datasets includes features related to the products like product\_category\_name, product\_name\_lenght, product\_description\_lenght, product\_photos\_qty, product\_weight\_g, product\_length\_cm, product\_height\_cm, product\_width\_cm.
   1. product\_ids are too long and need to be processed accordingly.
   2. The product category names are in brazillian language and hence need to be translated.
   3. Variable names like product\_name\_lenght is wrong and should be corrected.
   4. There are some missing values in the dataset. We can either remove them or impute them.
   5. Scales of the measurement of the products are very different. So, we can apply standardization for the sake of scaling the data.
2. **olist\_orders\_dataset.csv –** This dataset mainly contains the dates when the product was purchased, order approved data, estimated delivery date, final delivery date etc. It also contains the status of the delivery of the products.
   1. order\_ids and customer\_ids are too long and need to be processed accordingly.
   2. Around 97% of the products were successfully delivered.
   3. More than 99% of the purchases were made in the years 2017 and 2018. There is very less data for the year 2016.
   4. There are missing values in the date columns.
3. **olist\_order\_reviews\_dataset.csv –** The datasets contains the data related to the reviews and the dates on which reviews were written. It also has review\_counts.
   1. review\_ids and order\_ids are too long and need to be processed accordingly.
   2. 57% of the products received a review count of 5.
   3. 88% of the data for the reviews title is missing.
   4. A large number of customers have not provided a review and hence, there would be less data if we choose to work with NLP.
   5. There are missing values in the date columns.
4. **olist\_order\_payments\_dataset.csv – It provides us the information about the mode and number of instalments of payments etc.**
   1. order\_ids are too long and need to be processed accordingly.
   2. 74% of the customers have used credit card as the mode of payments.7
   3. Half of the customers have made the payment at once but other likes to pay in instalments.
5. **olist\_order\_items\_dataset.csv –** This includes the number of items ordered, price and freight\_value etc.
   1. 99.5% of the shipping limit date are from the years 2017 and 2018.
   2. Product\_ids, seller\_ids and order\_ids are too long and need to be processed accordingly.
6. **olist\_geolocation\_dataset.csv –** The geospatial data is also provided in this dataset and it can be easily used to do geospatial analysis.

**Findings:**

* There are more than 1 million data points in this dataset.
* 40% of the observations are from Sao Paulo state.

1. **product\_category\_name\_translation.csv –** This is the final dataset in the data corpus. It includes the translation of the product categories which can be mapped to the Brazilian names in other datasets.