**1. Introduction**

The Brazilian E-commerce Public Dataset by Olist comprises a collection of data regarding orders made at Olist's e-commerce platform in Brazil. The dataset covers orders placed between 2016 and 2018 and includes information on order status, prices, payment, delivery performance, customer location, product attributes, and reviews. This documentation provides an overview of the dataset structure, its components, and guidance on how to use the data for analysis.

**Dataset URL:** [Kaggle: Brazilian E-commerce](https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce)

**2. Dataset Overview**

The dataset contains multiple CSV files capturing different aspects of the e-commerce transactions. The files are interlinked through various keys, allowing for a comprehensive analysis of different facets of e-commerce activities.

**2.1. Files Included**

* **olist\_customers\_dataset.csv**: Contains customer information, including customer IDs and location.
* **olist\_orders\_dataset.csv**: Details of each order placed, including timestamps and status.
* **olist\_order\_items\_dataset.csv**: Each product ordered is represented as an item in this file, including product IDs, quantities, and pricing.
* **olist\_order\_payments\_dataset.csv**: Information on payment methods and installments for each order.
* **olist\_order\_reviews\_dataset.csv**: Customer reviews and ratings for products in each order.
* **olist\_products\_dataset.csv**: Details of products, including names, categories, and dimensions.
* **olist\_sellers\_dataset.csv**: Information about the sellers, including their locations.
* **olist\_geolocation\_dataset.csv**: Geographic coordinates for postal codes.
* **product\_category\_name\_translation.csv**: Mapping of product category names from Portuguese to English.

**3. Data Schema and Structure**

Each dataset has specific columns that relate to others using key fields. The following is a detailed schema for each dataset:

**3.1. olist\_customers\_dataset.csv**

* customer\_id (string): Unique ID for each customer.
* customer\_unique\_id (string): Unique identifier for each customer, appearing multiple times for customers who place multiple orders.
* customer\_zip\_code\_prefix (int): First five digits of the customer zip code.
* customer\_city (string): Customer's city.
* customer\_state (string): Customer's state.

**3.2. olist\_orders\_dataset.csv**

* order\_id (string): Unique ID for each order.
* customer\_id (string): Unique ID for each customer (links to olist\_customers\_dataset).
* order\_status (string): Status of the order (e.g., delivered, shipped).
* order\_purchase\_timestamp (timestamp): Time when the order was placed.
* order\_approved\_at (timestamp): Time when the order was approved.
* order\_delivered\_carrier\_date (timestamp): Time when the order was shipped.
* order\_delivered\_customer\_date (timestamp): Time when the order was delivered to the customer.
* order\_estimated\_delivery\_date (timestamp): Estimated delivery date.

**3.3. olist\_order\_items\_dataset.csv**

* order\_id (string): Unique ID for each order (links to olist\_orders\_dataset).
* order\_item\_id (int): ID for each item in the order.
* product\_id (string): Unique ID for each product.
* seller\_id (string): Unique ID for each seller.
* shipping\_limit\_date (timestamp): Deadline for the seller to ship the product.
* price (float): Price of the item.
* freight\_value (float): Shipping cost of the item.

**3.4. olist\_order\_payments\_dataset.csv**

* order\_id (string): Unique ID for each order (links to olist\_orders\_dataset).
* payment\_sequential (int): Sequential number of payments.
* payment\_type (string): Payment method used (e.g., credit card, boleto).
* payment\_installments (int): Number of payment installments.
* payment\_value (float): Total amount paid.

**3.5. olist\_order\_reviews\_dataset.csv**

* review\_id (string): Unique ID for each review.
* order\_id (string): Unique ID for each order (links to olist\_orders\_dataset).
* review\_score (int): Rating given by the customer (1-5).
* review\_comment\_title (string): Title of the review comment.
* review\_comment\_message (string): Detailed review comment.
* review\_creation\_date (timestamp): Date when the review was created.
* review\_answer\_timestamp (timestamp): Date when the review was answered.

**3.6. olist\_products\_dataset.csv**

* product\_id (string): Unique ID for each product.
* product\_category\_name (string): Category of the product in Portuguese.
* product\_name\_length (int): Length of the product name.
* product\_description\_length (int): Length of the product description.
* product\_photos\_qty (int): Number of product photos.
* product\_weight\_g (float): Weight of the product in grams.
* product\_length\_cm (float): Length of the product in cm.
* product\_height\_cm (float): Height of the product in cm.
* product\_width\_cm (float): Width of the product in cm.

**3.7. olist\_sellers\_dataset.csv**

* seller\_id (string): Unique ID for each seller.
* seller\_zip\_code\_prefix (int): First five digits of the seller zip code.
* seller\_city (string): Seller's city.
* seller\_state (string): Seller's state.

**3.8. olist\_geolocation\_dataset.csv**

* geolocation\_zip\_code\_prefix (int): First five digits of the zip code.
* geolocation\_lat (float): Latitude.
* geolocation\_lng (float): Longitude.
* geolocation\_city (string): City.
* geolocation\_state (string): State.

**3.9. product\_category\_name\_translation.csv**

* product\_category\_name (string): Product category name in Portuguese.
* product\_category\_name\_english (string): Product category name in English.

**4. EXTRACT TRANSFORM LOAD**

**4.1 Extract**

The extraction process involved loading data from the nine CSV files into dataframes using a data manipulation library like pandas.

Data Extraction Process

Using pandas, we read each CSV file into a DataFrame:

A screenshot of a computer code

Description automatically generated

**4.1.1** Identify data types for each table, check for null values, size and shape of the dataset

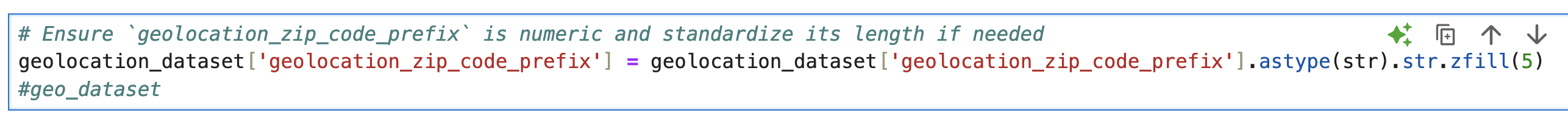
A white rectangular object with blue and white lines

Description automatically generated with medium confidence

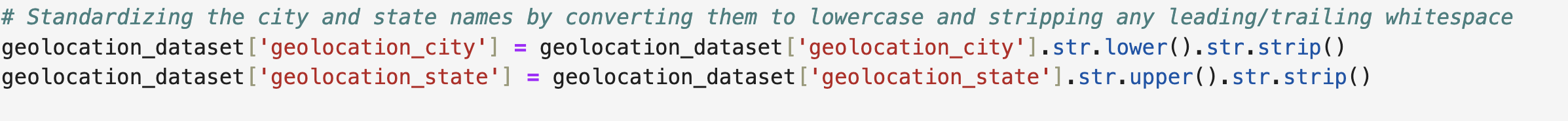
**5. Transform**

The transformation phase includes data cleaning, type conversion, normalization and data enrichment to prepare the data for loading.

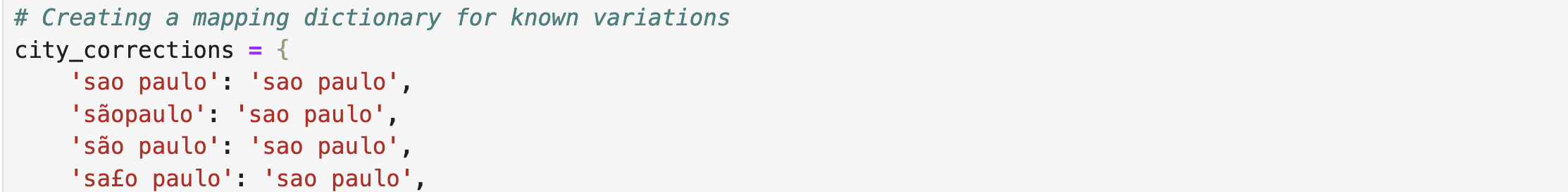
5.1. Standardizing geolocation\_zip\_code to 5 characters



5.2. Standardizing the city, state names to lowercase & removing whitespaces



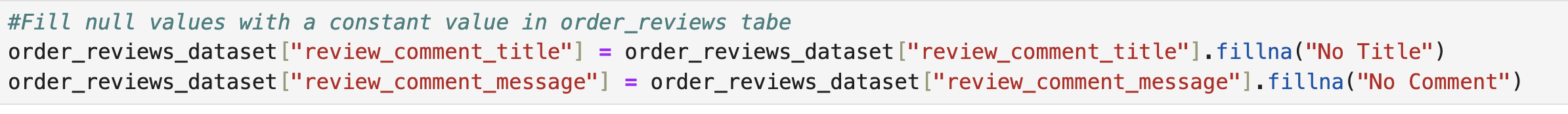
5.3. Creating a mapping dictionary for known city name variations



5.4. Convert 'shipping\_limit\_date' to datetime format



5.5. Fill null values with a constant value in order\_reviews\_dataset



5.6. Handling date time for all columns

A screen shot of a computer

Description automatically generated

5.7. Renaming columns in the product dataset



5.8. Replacing product\_category\_name with translated version & corrected spelling in values

**A screenshot of a computer

Description automatically generated**

5.9. Fill empty values with appropriate values to preserve the row

A screen shot of a computer code

Description automatically generated

**6. Load**

6.1 Table Creation involves creating orders, order\_reviews, order\_payments, order\_reviews, customers tables

A screen shot of a computer code

Description automatically generated

6.2. Adding in constraints & foreign keys to respective tables IN ORDER order\_reviews, order\_payments, customers, order\_items

A white screen with blue text

Description automatically generated

6.3. Inserting records to tables in reverse order of creation geolocation, sellers, products, customers, order\_items, order\_payments, order\_reviews, orders

A screenshot of a computer

Description automatically generated

**7. Conclusion**

This documentation provides an overview and detailed schema of the Brazilian e-commerce public dataset. By understanding the data's structure, users can perform various analyses, from customer behaviour to product performance, and uncover valuable insights in the Brazilian e-commerce sector.