E-Commerce Data Analysis

# 1. Data Dictionary

**1. \*\*Customers Dataset\*\*:**  
 - `customer\_id`: Unique identifier for each customer transaction.  
 - `customer\_unique\_id`: Unique identifier for each customer.  
 - `customer\_zip\_code\_prefix`: Zip code prefix of customer location.  
 - `customer\_city`: City of the customer.  
 - `customer\_state`: State of the customer.  
  
**2. \*\*Order Items Dataset\*\*:**  
 - `order\_id`: Unique identifier for each order.  
 - `order\_item\_id`: Sequence item number if the order has more than one item.  
 - `product\_id`: Unique identifier for the product.  
 - `seller\_id`: Unique identifier for the seller.  
 - `shipping\_limit\_date`: Deadline for shipping the item.  
 - `price`: Item price.  
 - `freight\_value`: Freight price.  
  
**3. \*\*Order Payments Dataset\*\*:**  
 - `order\_id`: Unique identifier for each order.  
 - `payment\_sequential`: A customer may pay an order with more than one payment method. If so, a sequence will be created to accommodate all payments.  
 - `payment\_type`: Method of payment.  
 - `payment\_installments`: Number of installments chosen for the payment.  
 - `payment\_value`: Transaction value.  
  
**4. \*\*Order Reviews Dataset\*\*:**  
 - `review\_id`: Unique identifier for each review.  
 - `order\_id`: Unique identifier for each order.  
 - `review\_score`: Score given by the customer on a scale of 1 to 5.  
 - `review\_comment\_title`: Title of the review.  
 - `review\_comment\_message`: Comment of the review.  
 - `review\_creation\_date`: Date of the review.  
 - `review\_answer\_timestamp`: Timestamp when the review was answered.  
  
**5. \*\*Orders Dataset\*\*:**  
 - `order\_id`: Unique identifier for each order.  
 - `customer\_id`: Unique identifier for each customer transaction.  
 - `order\_status`: Current status of the order.  
 - `order\_purchase\_timestamp`: Timestamp of the purchase.  
 - `order\_approved\_at`: Approval timestamp.  
 - `order\_delivered\_carrier\_date`: Carrier shipping timestamp.  
 - `order\_delivered\_customer\_date`: Customer receiving timestamp.  
 - `order\_estimated\_delivery\_date`: Estimated delivery timestamp.  
  
**6. \*\*Products Dataset\*\*:**  
 - `product\_id`: Unique identifier for each product.  
 - `product\_category\_name`: Category name in Portuguese.  
 - `product\_name\_lenght`: Length of the product name.  
 - `product\_description\_lenght`: Length of the product description.  
 - `product\_photos\_qty`: Number of photos posted for the product.  
 - `product\_weight\_g`: Weight of the product in grams.  
 - `product\_length\_cm`: Length of the product in cm.  
 - `product\_height\_cm`: Height of the product in cm.  
 - `product\_width\_cm`: Width of the product in cm.  
  
**7. \*\*Sellers Dataset\*\*:**  
 - `seller\_id`: Unique identifier for each seller.  
 - `seller\_zip\_code\_prefix`: Zip code prefix of seller location.  
 - `seller\_city`: City of the seller.  
 - `seller\_state`: State of the seller.  
  
**8. \*\*Category Translation Dataset\*\*:**  
 - `product\_category\_name`: Category name in Portuguese.  
 - `product\_category\_name\_english`: Category name translated to English.

# 2. Business Problems

1. Understanding customer preferences and buying behavior to improve sales.  
2. Enhancing product recommendations based on customer reviews and purchasing history.  
3. Optimizing the supply chain and logistics for better delivery times and lower costs.  
4. Analyzing the performance of sellers to maintain a high-quality marketplace.  
5. Identifying potential areas of expansion based on geographical sales data.

# 3. Tool Selection

**- \*\*Python\*\*:** For data cleaning, preprocessing, analysis, and machine learning tasks.  
**- \*\*SQL\*\*:** For querying and managing the data efficiently.  
**- \*\*Tableau\*\*:** For visualizing the data and gaining insights.

# 4. Questions

1. What are the top selling products and categories?  
2. How does the price and freight value affect customer satisfaction (review score)?  
3. Which sellers are performing well, and which ones are not? How does their performance impact customer satisfaction?  
4. How does the delivery time (actual vs estimated) impact customer satisfaction?  
5. What are the trends in sales over time? Are there any seasonal patterns?