# Business problem

Business need to focus on the right segmentation so that they can focus on that segmentation

# Data definition: -

This eCommerce dataset is provided by companies. This is the Company, which provides the connection of merchants and their products to the major marketplaces of various places. This e-Commerce public dataset contains nine CSV files. Each CSV holds different information about the eCommerce public dataset. All nine CSV and their details are given below.

**CSV FILE 1 (customer’s dataset): -** This CSV contains customer dataset which contains all the information of the customer. Features of the customer dataset are customer id, customer zip code, and customer state.

customer\_id: key to the orders dataset. Each order has a unique customer\_id.

customer\_unique\_id: unique identifier of a customer.

customer\_zip\_code\_prefix: first five digits of customer zip code

customer\_city: customer city name

customer\_state: customer state

**CSV FILE 2 (geolocation dataset): -** This dataset is geolocation dataset that contains all the information of location in the dataset. Features of geolocation dataset are geolocation state, geolocation city, and geolocation zip code, and geo-location latitude and longitude.

geolocation\_zip\_code\_prefix: first 5 digits of zip code

geolocation\_lat: latitude

geolocation\_long: longitutude

geolocation\_city: city name

geolocation\_state: state

**CSV FILE 3 (order items dataset): -** The third dataset contains all the information about order item. The features of item dataset are order id, order item id, product id, seller id, shipping limit time, price, and freight value.

order\_id: order unique identifier

order\_item\_id: sequential number identifying number of items included in the same order.

product\_id: product unique identifier

seller\_id: seller unique identifier

shipping\_limit\_date: Shows the seller shipping limit date for handling the order over to the logistic partner.

Price: item price

freight\_value: item freight value item (if an order has more than one item the freight value is splitted between items)

**CSV FILE 4 (order payments detail dataset): -** This fourth dataset contains all the information on the payment of the orders. the features of the order payment details are order id, payment type, payment value, payment sequential, and payment installments.

order\_id: unique identifier of an order.

payment\_sequential: a customer may pay an order with more than one payment method. If he does so, a sequence will be created to

payment\_type: method of payment chosen by the customer.

payment\_installments: number of installments chosen by the customer.

payment\_value: transaction value.

**CSV FILE 5 (order reviews dataset): -** The fifth dataset contains all the information of order review. the features of the order reviews dataset are review id, order id, review comment title, review comment message, review creation date and review answer timestamp.

review\_id: unique review identifier

order\_id: unique order identifier

review\_score: Note ranging from 1 to 5 given by the customer on a satisfaction survey.

review\_comment\_title: Comment title from the review left by the customer, in Portuguese.

review\_comment\_message: Comment message from the review left by the customer, in Portuguese.

review\_creation\_date: Shows the date in which the satisfaction survey was sent to the customer.

review\_answer\_timestamp: Shows satisfaction survey answer timestamp.

**CSV FILE 6 (orders dataset): -** This sixth dataset contains all the information of orders. the features of the order dataset are order id, customer id, order status, order purchase timestamp, order approved at, order delivered carrier date, order delivered customer date, and order estimated delivery date.

order\_id: unique identifier of the order.

customer\_id: key to the customer dataset. Each order has a unique customer\_id.

order\_status: Reference to the order status (delivered, shipped, etc).

order\_purchase\_timestamp: Shows the purchase timestamp.

order\_approved\_at: Shows the payment approval timestamp.

order\_delivered\_carrier\_date: Shows the order posting timestamp. When it was handled to the logistic partner.

order\_delivered\_customer\_date: Shows the actual order delivery date to the customer.

order\_estimated\_delivery\_date: Shows the estimated delivery date that was informed to customer at the purchase moment.

**CSV FILE 7 (products dataset): -** This seventh dataset contains all the information about the product. the features of the product dataset are product id, product category name, product name length, product description length, product photos qty, product weight g, product length cm, product height cm, and product width cm.

product\_id: unique product identifier

product\_category\_name: root category of product, in Portuguese.

product\_name\_lenght: number of characters extracted from the product name.

product\_description\_lenght: number of characters extracted from the product description.

product\_photos\_qty: number of product published photos

product\_weight\_g: product weight measured in grams.

product\_length\_cm: product length measured in centimeters.

product\_height\_cm: product height measured in centimeters.

product\_width\_cm: product width measured in centimeters.

**CSV FILE 8 (sellers’ dataset): -** The dataset contains all the information of sellers. the features of the seller’s dataset are seller id, seller zip code, seller city, and seller state.

seller\_id: seller unique identifier

seller\_zip\_code\_prefix: first 5 digits of seller zip code

seller\_city: seller city name

seller\_state: seller state

**CSV FILE 9 (product category name translation dataset): -** This dataset contains all the information of product category names. the features of the product category name dataset are product category name and product category name English.

product\_category\_name: category name in Portuguese

product\_category\_name\_english: category name in English

# Data problem: -

1. There are 9 csv file so first problem is data merging problem. All data are relational data here so the first step will merge all CSV. So we need to convert all CSV into a single CSV. For merge, CSV will apply merge on left. Here will merge two CSV at a time.

Here when will add all these csv and will convert into 1 csv.

2. Will found a lot of null values and outliers in dataset so second task will remove all the null values and treat the outlier with the help of different technique in machine learning like mean median and mode etc.

3. Problem because of categorical variable if there will be categorical variable in the dataset then categorical variable will convert that categorical variable into integer variables with the help of various technique like one hot encoding etc. Then will apply cluster algorithm for market segmentation of the customer.

4. Apply clustering techniques for finding the right set of customer segmentation