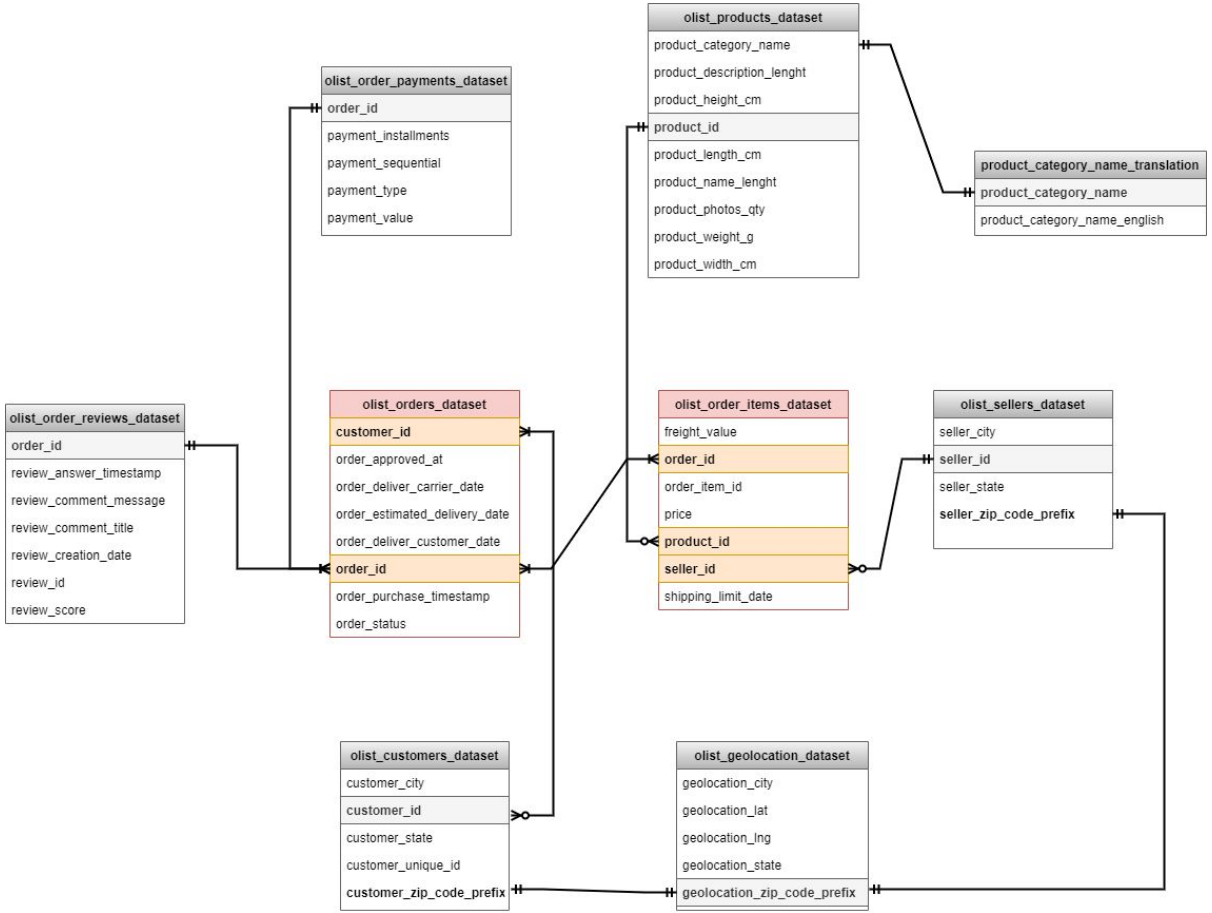
DATA EXPLORATION AND VISUALIZATION

1. Data Exploration: EDA

Draw an ERD to explore the relationship between the dataset



1. Data Exploration: EDA

Why I do EDA?

- Identify the faulty points in data -> easily remove them and clean your data
- Help to understand the relationship between variables which gives us wider perspective on data

Steps to do EDA:

1. Understand dataset
   - Kind of data I working at, number of columns and rows, how it actually look likes
2. Clean your data from redundancies
3. Analysis of Relationship between variables:
   - Estimate the Central
   - Estimate the Variability
   - Explore the Distribution
   - Correlation

2. Business Acumen

1. Identify and measure at least three key suitable metrics to analyse the company performance and deliver insights for improvement
2. Explain your rationale behind those metrics
3. Provide supporting analysis and visualization to communicate your ideas to the CEO
4. Which additional data/dataset do you think the company should collect and why

Empathize

1. Understanding goals

- Who is the report for? - CEO
- What is the purpose of the report? -
  - Maximize GMV
  - Optimize spending
- What is the desire of the viewer? - Make a decision/ suggestion

2. Understanding dataset

8 tables:

- Customers
- Geolocation
- Items
- Payments
- Reviews
- Orders
- Products
- Sellers

Empathize

1. Customers
   - Entity: customer
   - Dimension: id, state, city, unique_id
   - Metrics: most common state, city, retention rate

2. Geolocation
   - Entity: geolocation
   - Dimension: city, state
   - Metrics: distribution of customers in Brazil

3. Order Items
   - Entity: order_id, order_item_id, product_id, seller_id
   - Dimension: price, freight_value
   - Metrics: popular products, revenue per order, shipping spending, revenue per seller, revenue per product
   - Time: shipping_limit_date

4. Payments
   - Entity: order_id, payment_type
   - Dimension: payment_installments, payment_sequential, payment value
   - Metrics: most popular payment types, total payment_value by each type, tendency of installment

Empathize

5. Reviews
   - Entity: order_id, review_id
   - Dimension: review_score
   - Metrics: total score per order
   - Time: review_creation_date, review_answer_timestamp

6. Orders
   - Entity: customer_id, order_id, order_status
   - Dimension: time purchase, time approved, time delivery carrier, estimated delivery time, time to delivery customer on hand
   - Metrics: On-time rate
   - Time: all except entity

7. Products
   - Entity: product_id, product_category_name,
   - Dimension: length, height, photo_qty, weight, width
   - Metrics: relationship between product size and time shipping?

8. Sellers
   - Entity: seller_id
   - Dimension: seller_city, seller_state
   - Metrics: location of seller and relationship with location of customer, time shipping

Define

- Define important attributes
- List down key metrics
- Brainstorming question list
- Ranking -> Pick

1. Scale up GMV and optimize spending are hard to do at the same time. Depend on the situation at each different stage of the company, they have to decide to increase GMV or cut cost.

   Let come up with Maximize GMV first:

- Most important metric: GMV (related to revenue)
- Sub-metric: Thinking in different dimension:
  Revenue = number of orders* AOV
  Revenue = activation+reactivation
  Revenue = revenue by channel
  Revenue = revenue by category

Revenue = number of orders * AOV

- ✓ Increase number of orders
- ✓ Increase AOV
- ✓ Increase both

Number of orders

- What is your targeted customers?
- They live focusly on what region of country?
- How many orders of each region?
- How many orders of each group customers?

AOV

1. Want customers buy products with bigger prices
   - What are main strategy of sales?/ How products are selling?
     - Up-selling?
     - Gross-selling?
   - Can we find other sellers who manufacture products with higher prices?
   - Can we sell more expensive products?
2. Make customers buy more products -> make more selections
   - How many products are selling
   - What category has the most sold products

Revenue = activation+reactivation

1. Get new users
   - Marketing channels
2. Make old users by again and again
   - Strategy for them

customer life cycle
retention rate
churn rate
=> not good to have more new users but churn rate are high

Cost to promotion to get new users vs
Cost to increase user retention, which ones is bigger? -> suggest strategy

Revenue = revenue by category

- What category are the most favorite?
- What category has biggest revenue?
- What category has the most selection?

Optimize spending

1. Shipping cost
   - What are distances from seller to customers?
   - Decrease the distant or suggest sellers who have the same products in customer’s region to join/ sellers to move to region which have lots of customers purchasing their products
   - Or say ‘no’ for shipping, let sellers do shipping at their own
2. Warehouse cost

How to have the shortest time for warehouse?

- Classify orders in the most convenient way
- Eliminate abandoned orders -> sell with lower price

Recommendation

Maximize GMV	Assumption	Recommendation
Increase number of orders	- target customers: Female, (Age), who interested in health and beauty - number of orders come from East size which the population density is high - defined unique_customer_id - defined which customers buy most	- designed target promotion for each target customer group - send discount code for top frequent customers via their own account: unique_customer_id - increase brand awareness in West size
Increase AOV 1. Make people buy more products 2. Make people buy products with higher price	- need data about sales method, number of products selling with combo, hot deal - Bed_bath_table are having the biggest quantity of products sold	1.1 Create more selection for customers -> In top 5 categories: bed_bath_table, furniture_decor, health_beauty, sports_leisure and computer accessories 2.1 Up-selling: sell with combo, instead of single 2.2 Gross-selling: sell with hot deal 2.3 Find sellers in the same category but higher product price -> suggest they sell in the company website 2.4 Suggest more products in higher range prices

Maximize GMV	Assumption	Recommendation
Activation and Reactivation	- need customer life cycle value - retention rate are low: 3.12% - need churn rate - need number of old customers return	Sometimes new customers are increasing but churn rate are high => Should focus to get promotion for old users - increase retention rate Are cost of products reasonable? -> Customers do not like high freight cost
Category	- Health and beauty bring the most valuable GMV - Sold most products in bed_bath_table - Bed_bath_table and sports leisure have the most selections	Focus on these categories
Channel	- Need data of marketing channels	

Optimize Spending	Assumption	Recommendation
Shipping expenses	- Some are high, because of shipping oversea - Most sellers are in Sao Paulo but most customers are in coastal cities	- Optimize distances - Increase orders in the same region, area - Distribute sellers better ● Suggest sellers in SP that most of customers of their products are in coastal cities -> they should move to these cities ● Find sellers who sell the same products but live in coastal cities -> suggest them to join ecommerce - Cut cost: let sellers do their shipping
Warehouse expenses	Need data in warehouse	- Classify orders in right way - Eliminate the inventory
Marketing expenses	Need data about marketing expenses to get new users and to increase retention rate	- Compare the expenses -> to action that we should acquire more customers or focus on old customers, loyalty customers