**OBJECTIVE**

As a data analyst for Olist Store Analysis, provide insights on key aspects of e-commerce performance and customer behavior to support decision-making.

**Weekday vs Weekend Performance:** Total number of orders comparison and Average payment value for orders.

**High Review Scores and Payment Method:** Total number of orders with a review score of 5 and payment type as credit card.

**Delivery Performance:** Average number of days taken for order delivery (customer's perspective) for pet\_shop.

**City-Based Insights:** Average price and payment values from customers in São Paulo.

**Shipping Days vs Review Scores:**Analyze the relationship between shipping days (calculated as the difference between order\_delivered\_customer\_date and order\_purchase\_timestamp) and review scores.

**DOMAIN KNOWLEDGE**

In the e-commerce domain, revenue generation typically follows several key mechanisms. Here’s a breakdown of how money flows in an e-commerce business:

**Sales Revenue**: Olist earns a commission on products sold by third-party sellers.

**Transaction Fees**: A fee for processing payments on each sale.

**Advertising**: Sellers pay for better visibility (sponsored listings) and third-party ads.

**Affiliate Marketing**: Influencers earn commissions for driving sales.

**Shipping Fees**: Olist charges for logistics and shipping.

**Subscription Models**: Premium seller accounts or buyer memberships.

**Returns & Refunds**: Olist may charge for processing returns or restocking.

**DATA KNOWLEDGE**

**Customers\_Dataset Table(99441 rows)**

* customer\_id: Unique identifier for each customer.
* customer\_unique\_id: A unique, anonymized ID for the customer.
* customer\_zip\_code\_prefix: Postal code prefix of the customer's address.
* customer\_city: City where the customer resides.
* customer\_state: State where the customer resides.

**Geolocation\_Dataset Table(738332 rows)**

* geolocation\_zip\_code\_prefix: Zip code prefix linked to geolocation data.
* geolocation\_lat: Latitude of the location.
* geolocation\_lng: Longitude of the location.
* geolocation\_city: City associated with the location.
* geolocation\_state: State associated with the location.

**Order\_Items\_Dataset Table(112650 rows)**

* order\_id: Unique identifier for each order.
* order\_item\_id: Unique identifier for each order item.
* product\_id: Unique identifier for each product in the order.
* seller\_id: ID of the seller fulfilling the order.
* shipping\_limit\_date: Last date for shipping the order.
* price: Price of the product.
* freight\_value: Shipping cost for the order.

**Order\_Payments\_Dataset Table(103886 rows)**

* order\_id: Unique identifier for the order.
* payment\_sequential: Sequential order of payments for the order.
* payment\_type: Type of payment (e.g., credit card, boleto).
* payment\_installments: Number of installments in the payment.
* payment\_value: Total payment amount.

**Order\_Review\_Dataset Table(99224 rows)**

* review\_id: Unique identifier for each review.
* order\_id: Unique identifier for the related order.
* review\_score: Rating given by the customer (1-5).
* review\_comment\_title: Title of the review comment.
* review\_comment\_message: Detailed review message.
* review\_creation\_date: Date when the review was created.
* review\_answer\_timestamp: Time when the review was answered by the seller.

**Orders\_Dataset Table(99441 rows)**

* order\_id: Unique identifier for the order.
* customer\_id: Unique identifier for the customer placing the order.
* order\_status: Current status of the order (e.g., delivered, cancelled).
* order\_purchase\_timestamp: Timestamp when the order was placed.
* order\_approved\_at: Timestamp when the order was approved.
* order\_delivered\_carrier\_date: Date when the carrier delivered the order.
* order\_delivered\_customer\_date: Date when the customer received the order.
* order\_estimated\_delivery\_date: Estimated delivery date.

**Product\_Category\_Name\_Translation Table(71)**

* product\_category\_name: Category name in the original language.
* product\_category\_name\_english: English translation of the category name.

**Products\_Dataset Table(32951 rows)**

* product\_id: Unique identifier for each product.
* product\_category\_name: Category of the product.
* product\_name\_length: Length of the product's name.
* product\_description\_length: Length of the product description.
* product\_photos\_qty: Number of photos associated with the product.
* product\_weight\_g: Weight of the product in grams.
* product\_length\_cm: Length of the product in centimeters.
* product\_height\_cm: Height of the product in centimeters.
* product\_width\_cm: Width of the product in centimeters.

**Seller\_Dataset Table(3095 rows)**

* seller\_id: Unique identifier for each seller.
* seller\_zip\_code\_prefix: Postal code prefix of the seller's address.
* seller\_city: City where the seller is located.
* seller\_state: State where the seller is located.

**DATA CLEANING**

**Customers\_Dataset**

* Checked for null values.
* Capitalized each word in the customer\_city column.
* Removed duplicate rows.
* Checked and corrected data types.

**Geolocation\_Dataset**

* Checked for null values.
* Removed duplicate rows.
* Checked and corrected data types.
* Standardized the geolocation\_city column by changing all variations of "Sao Paulo" to a single consistent format (e.g., "São Paulo").

**Order\_Items\_Dataset**

* Checked and corrected data types.
* Removed duplicate rows.
* Checked for null values.

**Order\_Payments\_Dataset**

* Checked and corrected data types.
* Removed duplicate rows.
* Checked for null values.
* Cleaned the payment\_type column by:
  + Replacing underscores ("\_") with spaces (" ").
  + Capitalizing each word.

**Order\_Review\_Dataset**

* Checked and corrected data types.
* Removed duplicate rows.
* Checked for null values in review\_comment\_title and review\_comment\_message columns.
* Replaced null values with "No Title" and "No Message", respectively.

**Orders\_Dataset**

* Capitalized each word in the order\_status column.
* Removed duplicate rows.
* Checked and corrected data types.
* Checked for null values in order\_delivered\_carrier\_date and order\_delivered\_customer\_date.
* Retained null values for potential future analysis.

**Product\_Category\_Name\_Translation**

* Used the first row as headers.
* Removed duplicate rows.
* Checked and corrected data types.
* Checked for null values.
* Cleaned both columns by replacing underscores ("\_") with spaces (" ") and capitalizing each word.

**Products\_Dataset**

* Removed duplicate rows.
* Checked and corrected data types.
* Corrected the header names for name\_length and description\_length (corrected spelling errors).
* Replaced null values in product\_name\_length and product\_description\_length with 0.
* Cleaned the product\_category\_name column by:
  + Replacing underscores ("\_") with spaces (" ").
  + Replacing null values with "Unknown".
  + Capitalizing each word.

**Seller\_Dataset**

* Removed duplicate rows.
* Checked for null values.
* Checked and corrected data types.
* Standardized the seller\_city column by replacing the value "04482258" with "Rio De Janeiro" based on zip\_code and seller\_state information.
* Capitalized each word in the seller\_city column.
* Standardized the seller\_city column by changing all variations of "Sao Paulo" to a single consistent format (e.g., "São Paulo").

**RELATIONSHIPS(JOINS)**

**Step 1 :**

Merge Olist\_products\_dataset (column : product\_category\_name) & product\_category\_name\_Translation(column:product\_category\_name)

**Step 2:**

Remove duplicates from geolocation table with column “geolocation\_zip\_code\_prefix” for establishing a relationship.

**Step 3:**

Create Date Table and join it with order\_approved\_date column in orders\_dataset.



**DO VISIT THE INTERACTIVE DASHBOARD**

**THANKS 😊**