LinkedIn:-https://www.linkedin.com/in/shudhanshu-awasthi

CASE STUDY: BRAZILIAN OLIST E-COMMERCE DATA ANALYSIS

BY

SHUDHANSHU AWASTHI

DUAL SPECIALIZATION

(MBA BUSINESS ANALYTICS & MARKETING)

TABLE OF CONTENTS

Case Study: Brazilian Olist E-Commerce Data Analysis

Company Overview

Olist is a leading Brazilian e-commerce platform that connects small and medium-sized businesses with customers across major online marketplaces like Mercado Livre, Amazon, and Magalu. Instead of building their own online stores, sellers use Olist to list, manage, and promote their products under the trusted Olist brand. This allows local merchants to reach customers nationwide while focusing on sales rather than operations.

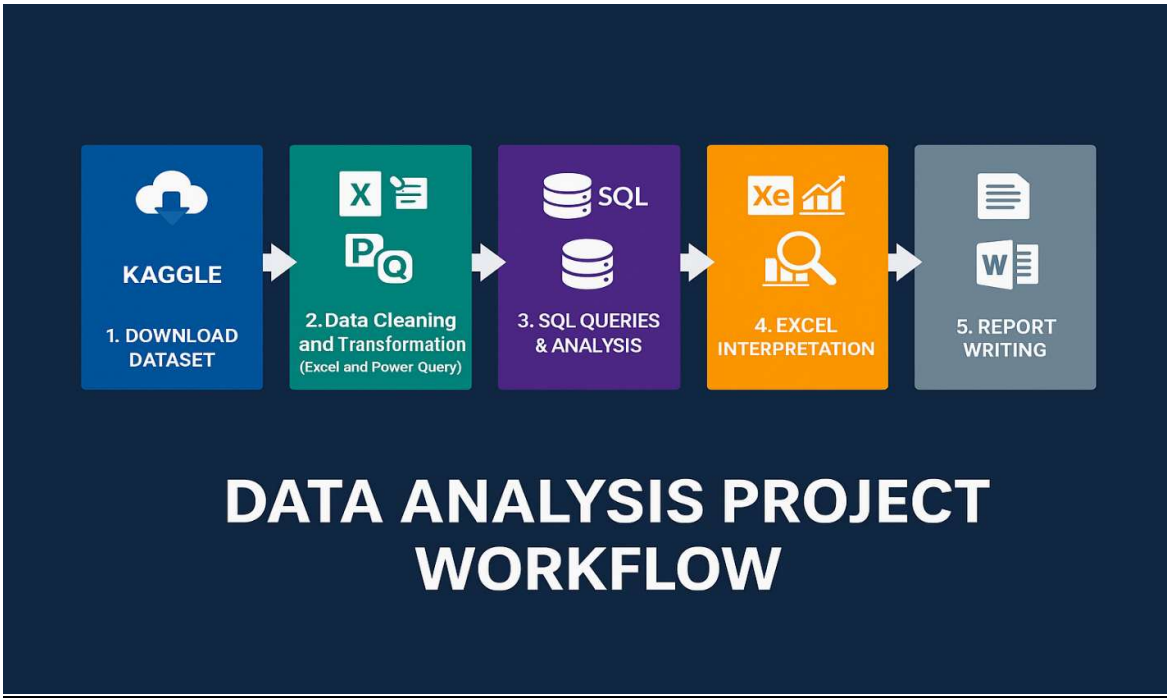
Business Model

Olist follows a **commission-based marketplace model**, earning revenue through transaction and service fees on each completed sale. The company provides end-to-end support — from product listing and order management to payment processing and logistics coordination. By partnering with multiple delivery and fulfillment providers, Olist ensures quick and reliable shipping.

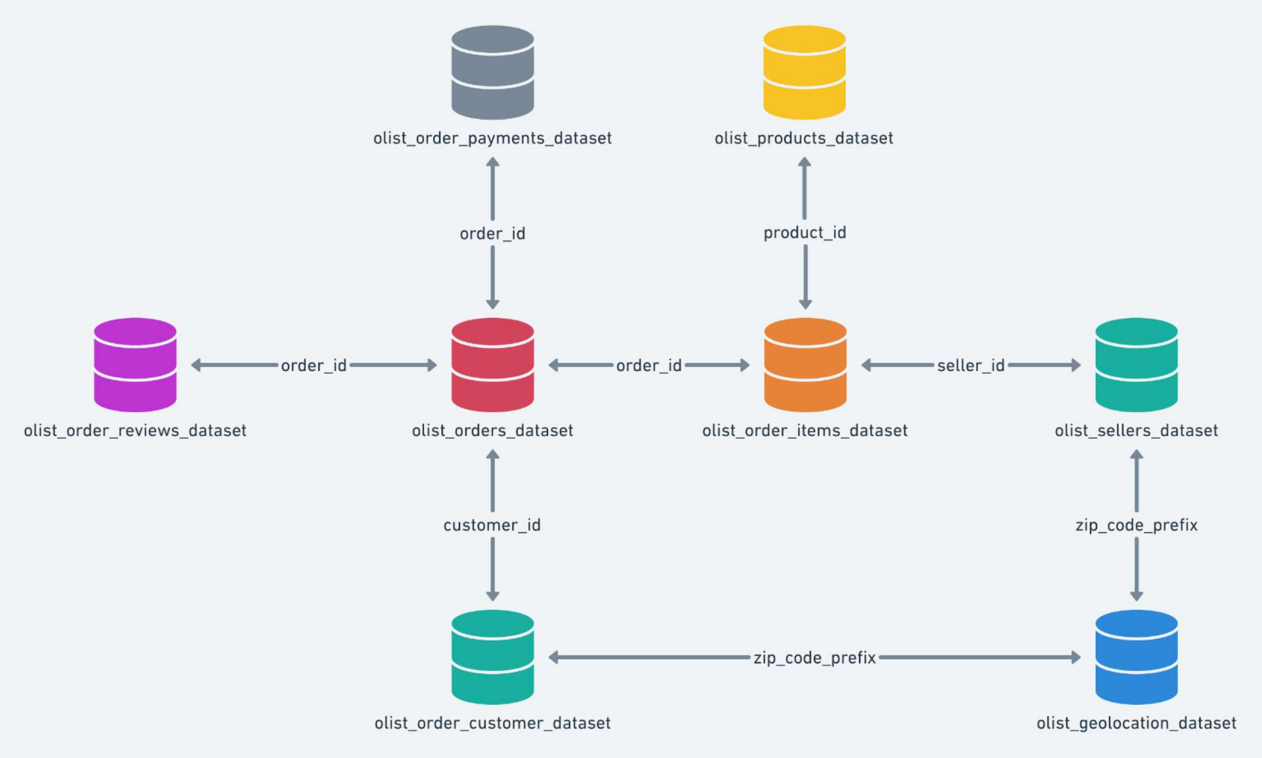
Problem statement

1. **How big and healthy is the business?** Show total revenue and orders, and how both change month to month so we can spot growth, declines or seasonal peaks.
2. **Where should we invest locally?** Show which states and cities bring the most revenue and orders so marketing and logistics can be focused where they'll have the most impact.
3. **How do customers pay?** Show which payment methods customers use for successful (delivered) orders and whether customers prefer installments or certain payment types.
4. **Which products earn money, and which only sell a lot?** Rank categories and SKUs by revenue and units sold so we can prioritise high-value items and spot popular but low-margin categories.
5. **Where are the quality or logistics problems?** Identify categories (and later sellers) with the most cancellations, refunds, or late deliveries and calculate the revenue lost — so we can fix the biggest problems first.
6. **Do we have enough sellers where customers are?** Map sellers and customers by state to find places with few or no sellers, which can hurt delivery speed and assortment.
7. **Is the marketplace concentrated?** Measure how much revenue the top 10% of sellers contribute to see if we're too dependent on a few players.
8. **Are customers coming back and who are the best ones?** Measure repeat buyers and how much revenue they bring, see how review scores relate to sales, and identify the top 1% of customers by lifetime spend for targeted retention.

DATA ANALYSIS PROJECT WORK FLOW



DATA SCHEMA DIAGRAM (SQL)



MySQL Questions for Analysis

Business Questions

Revenue & Orders (Business Health)

1. What is the total revenue & orders generated by Olist, and how has it changed month by month to identify growth or decline trends?

2. Which customer states or cities contribute the most revenue, indicating regions for marketing focus?

3. Which payment methods are most commonly used for delivered orders, and what proportion of total payments does each method represent? Additionally, what insights can be drawn about customer payment preferences on Olist?

Products & Categories:-

4. Which product categories generate the highest revenue, showing where the business earns most?

5. Which product categories sell the most units, indicating popularity vs revenue efficiency?

6. Which product categories have the highest average revenue per order , suggesting opportunities for upselling?

7. What are Top 3 product categories by number of order cancellations or refunds, indicating potential quality or logistics issues?

8. What are Top 5 product categories by revenue lost due to order cancellations ?

Sellers & Marketplace

9. How Sellers & Customers are distributed across different states and which states have low or no Seller/Customers, explaining how it can effect the business

10. How has the number of active sellers changed month by month from 2016 to 2018, highlighting marketplace growth and peak engagement periods ?

11. What percentage of total revenue is contributed by the top 10% of sellers, revealing dependency on key sellers?

Customers & Retention

12. What percentage of customers are repeat buyers, and what share of revenue do they contribute, showing customer loyalty impact?

13. How do customer reviews and ratings affect sales and revenue performance on Olist ?

14. Who are the top 1% of customers by lifetime spend, and what percentage of total revenue do they represent, highlighting high-value customers?

Olist E Commerce Data Set

STEP 1:- Data Cleaning:-

1. **Olist Customer Data Set:-**

   Cleaned this data set by checking for blanks, inconsistencies and capitalized each words of city column with the help of power query transform option, also found that customer_id column had unique values while customer_unique_id has duplicate values, since for every order it has unique customer id also one customer can place multiple orders so chances are there that customer_unique_id is repeated in this data set so I choose not to de duplicate the values as we can always handle it in sql if needed.

2. **Olist Payment Data Set:-**

   Cleaned this data set by checking for blanks, inconsistencies and capitalized each words of payment type column with the help of power query transform option

3. **Olist Order Review Data Set:-**

   Cleaned this data set by checking for blanks, inconsistencies and found that review comment title and review comment message column had null values since it can't significantly impact analysis as per our objectives so I choose to replace null values with N/A with the help of power query replace value option

4. **Olist Products Data Set:-**

   Cleaned this data set by checking for blanks, inconsistencies and found that product category name column had null values so I choose it to replace with "Uncategorized" with the help of power Query, also product name length , description length had null values so I replaced it with 0 , corrected spelling of columns

5. **Olist Seller Data Set**

   Cleaned this data set by checking for blanks, inconsistencies and found that seller city had 1 numeric value zip code prefix instead of city name so I have checked it against geo location data set found the city name & replaced it with name of city , also capitalized each word of seller city column with the help of power query

6. **Product Category Name Translation Data Set:-**

   Cleaned this data set by checking for blanks, inconsistencies and found that column header was not in proper format , promoted header using power query transform:- use first row as a header option also capitalized each word of product category names .

7. Olist Geo Location Data Set:-

Cleaned this data set by checking for blanks, inconsistencies and capitalizing each words of geo location city names column.

8. Olist Order Data Set:-

Cleaned this data set by checking for blanks, inconsistencies and found that despite of order status "delivered" columns such as order approved date ,"order delivered carrier date, order delivered customer date had null values in 14 records so I decided to impute the values for the order status "delivered" by following steps:-

a). Column:- **Order Approved at** (Imputation)

1. Transforming the columns into correct date format:- dd-mm-yyy

2. Find the difference between order approved time and order purchase time to find out difference in time from placing order by the customer to approving the purchase of item from seller

   Formulae Used:-
   =IF(ISBLANK([@[order_delivered_carrier_date]]), " ", F2-E2)
   =MEDIAN([Order Approved-Order Purchare Time])

3. Finding out the median of difference between dates using excel statistics formulae and found that median difference was 0.014~ 0 day means order is approved on the same day

4. Adding the median difference to the order purchase time to find out missing values in order approved at column

a) Column:- **Order Delivered Carrier Date** (Imputation)

5. Transforming the columns into correct date format:- dd-mm-yyy

6. Find the difference between order delivered carrier date and order approved at to find out difference in time from ordering to handing over order to the carrier

   Formulae Used:-
   =IF(ISBLANK([@[order_delivered_carrier_date]]), " ", F2-E2)
   =MEDIAN([Carrier Date-Approved Date])

7. Finding out the median of difference between dates using excel statistics formulae and found that median difference was 1.81~2 days

8. Adding the median difference to the order approved at date to find out missing values in order delivered carrier date

b). Column:- **Order Delivered Customer Date** (Imputation)

1. Transforming the columns into correct date format:- dd-mm-yyy

2. Find the difference between order estimated delivery date and order delivered carrier date to find out difference in time from handing over order to the carrier to final delivery to the customer

Formulae Used:-
=IF(ISBLANK([@[order_delivered_customer_date]]), " ", H2-G2)
=MEDIAN([Estimated Date-Customer Delivered Date])

3. Finding out the median of difference between dates using excel statistics formulae and found that median difference was 11.94~ 12 days

4. Adding the median difference to the order delivered carrier date to find out missing values in order delivered customer date column

After Cleaning 9 files additional columns used for imputation of missing values were deleted to ensure files remain light for exporting it to mysql for further analysis

STEP 2:-*Importing Cleaned CSV Files into MySQL*

After Cleaning CSV Files, I have Imported CSV files into MySQL using load data infile command the imported data sets were validated against csv files for it's data types and to ensure each records were correctly imported without any data loss following are the code given below

```
-- ============================================================================
-- Olist E-commerce Database Schema
-- ============================================================================
-- 1. Create and use the database

CREATE DATABASE IF NOT EXISTS ecommerce;
USE ecommerce;

-- ============================================================================
-- Table: customers
-- ============================================================================

DROP TABLE IF EXISTS customers;
CREATE TABLE customers
(
   customer_id VARCHAR(50),
   customer_unique_id VARCHAR(50),
   customer_zip_code_prefix INT,
   customer_city VARCHAR(255),
   customer_state CHAR(2),
   PRIMARY KEY (customer_id)
) CHARACTER SET utf8mb4 COLLATE utf8mb4_unicode_ci;

LOAD DATA INFILE 'D:/Business Analyst Training/Analytics
Projects/CSV_Cleaned_Data_Sets/Customers_Data.csv'
INTO TABLE customers
```

```
CHARACTER SET utf8mb4
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\r\n'
IGNORE 1 ROWS
(customer_id, customer_unique_id, customer_zip_code_prefix, customer_city, customer_state);



-- ========================================================================
-- Table: geolocation
-- ========================================================================


DROP TABLE IF EXISTS geolocation;
CREATE TABLE geolocation
 (
   geolocation_zip_code_prefix INT,
   geolocation_lat DECIMAL(11, 8),
   geolocation_lng DECIMAL(11, 8),
   geolocation_city VARCHAR(255),
   geolocation_state CHAR(2)
) CHARACTER SET utf8mb4 COLLATE utf8mb4_unicode_ci;


LOAD DATA INFILE 'D:/Business Analyst Training/Analytics
Projects/CSV_Cleaned_Data_Sets/Geo_Location_Data.csv'
INTO TABLE geolocation
CHARACTER SET utf8mb4
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\r\n'
IGNORE 1 ROWS
```

```
(geolocation_zip_code_prefix, geolocation_lat, geolocation_lng, geolocation_city, geolocation_state);



-- ========================================================================
-- Table: order_items
-- ========================================================================

DROP TABLE IF EXISTS order_items;
CREATE TABLE order_items (
    order_id VARCHAR(50),
    order_item_id INT,
    product_id VARCHAR(50),
    seller_id VARCHAR(50),
    shipping_limit_date DATE,
    price DECIMAL(10, 2),
    freight_value DECIMAL(10, 2),
    PRIMARY KEY (order_id, order_item_id)
) CHARACTER SET utf8mb4 COLLATE utf8mb4_unicode_ci;

LOAD DATA INFILE 'D:/Business Analyst Training/Analytics
Projects/CSV_Cleaned_Data_Sets/Order_Items_Data.csv'
INTO TABLE order_items
CHARACTER SET utf8mb4
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\r\n'
IGNORE 1 ROWS
(order_id, order_item_id, product_id, seller_id, shipping_limit_date, price, freight_value);
```

```
-- ============================================================================
-- Table: order_payments
-- ============================================================================

DROP TABLE IF EXISTS order_payments;
CREATE TABLE order_payments (

 order_id VARCHAR(50),
    payment_sequential INT,
    payment_type VARCHAR(50),
    payment_installments INT,
    payment_value DECIMAL(10, 2),
    PRIMARY KEY (order_id, payment_sequential)
) CHARACTER SET utf8mb4 COLLATE utf8mb4_unicode_ci;

LOAD DATA INFILE 'D:/Business Analyst Training/Analytics
Projects/CSV_Cleaned_Data_Sets/Order_Payments_Data.csv'
INTO TABLE order_payments
CHARACTER SET utf8mb4
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\r\n'
IGNORE 1 ROWS
(order_id, payment_sequential, payment_type, payment_installments, payment_value);

-- ============================================================================
-- Table: order_reviews
-- ============================================================================
DROP TABLE IF EXISTS order_reviews;
CREATE TABLE order_reviews
(
```

```
  review_id VARCHAR(50),
    order_id VARCHAR(50),
    review_score INT,
    review_comment_title TEXT,


    review_comment_message TEXT,
    review_creation_date DATE,
    review_answer_timestamp DATE,
    PRIMARY KEY (review_id)

) CHARACTER SET utf8mb4 COLLATE utf8mb4_unicode_ci;


LOAD DATA INFILE 'D:/Business Analyst Training/Analytics
Projects/CSV_Cleaned_Data_Sets/Order_Reviews_Data.csv'
INTO TABLE order_reviews
CHARACTER SET utf8mb4
FIELDS TERMINATED BY ','
ENCLOSED BY ""
LINES TERMINATED BY '\r\n'
IGNORE 1 ROWS
(review_id,
order_id,review_score,review_comment_title,review_comment_message,review_creation_date,review_a
nswer_timestamp);



-- ========================================================================
-- Table: orders
-- ========================================================================


DROP TABLE IF EXISTS orders;
CREATE TABLE orders
```

```
(
    order_id VARCHAR(50),
    customer_id VARCHAR(50),
    order_status VARCHAR(50),
    order_purchase_timestamp DATE,
    order_approved_at DATE NULL,

    order_delivered_carrier_date DATE NULL,
    order_delivered_customer_date DATE NULL,
    order_estimated_delivery_date DATE,
    PRIMARY KEY (order_id)
) CHARACTER SET utf8mb4 COLLATE utf8mb4_unicode_ci;

LOAD DATA INFILE 'D:/Business Analyst Training/Analytics
Projects/CSV_Cleaned_Data_Sets/Orders_Data.csv'
INTO TABLE orders
CHARACTER SET utf8mb4
FIELDS TERMINATED BY ',' ENCLOSED BY '"'
LINES TERMINATED BY '\r\n'
IGNORE 1 ROWS
(
  order_id,
  customer_id,
  order_status,
  order_purchase_timestamp,
  @order_approved_at,
  @order_delivered_carrier_date,
  @order_delivered_customer_date,
  order_estimated_delivery_date
```

```
)
SET
  order_approved_at = NULLIF(@order_approved_at, ''),
  order_delivered_carrier_date = NULLIF(@order_delivered_carrier_date, ''),
  order_delivered_customer_date = NULLIF(@order_delivered_customer_date, '');



-- ========================================================================
-- Table: products
-- ========================================================================

DROP TABLE IF EXISTS products;
CREATE TABLE products (
    product_id VARCHAR(50),
    product_category_name VARCHAR(255) NULL,
    product_name_length INT NULL,

    product_description_length INT NULL,
    product_photos_qty INT NULL,
    product_weight_g INT NULL,
    product_length_cm INT NULL,
    product_height_cm INT NULL,
    product_width_cm INT NULL,
    PRIMARY KEY (product_id)
) CHARACTER SET utf8mb4 COLLATE utf8mb4_unicode_ci;

LOAD DATA INFILE 'D:/Business Analyst Training/Analytics
Projects/CSV_Cleaned_Data_Sets/Products_Data.csv'
INTO TABLE products
CHARACTER SET utf8mb4
FIELDS TERMINATED BY ','
```

```
ENCLOSED BY '"'
LINES TERMINATED BY '\r\n'
IGNORE 1 ROWS
(product_id, product_category_name, product_name_length, product_description_length,
product_photos_qty, product_weight_g, product_length_cm, product_height_cm, product_width_cm);


-- ========================================================================
-- Table: sellers
-- ========================================================================


DROP TABLE IF EXISTS sellers;
CREATE TABLE sellers
(
   seller_id VARCHAR(50),
   seller_zip_code_prefix INT,
   seller_city VARCHAR(255),
   seller_state CHAR(2),

PRIMARY KEY (seller_id)
) CHARACTER SET utf8mb4 COLLATE utf8mb4_unicode_ci;


LOAD DATA INFILE 'D:/Business Analyst Training/Analytics
Projects/CSV_Cleaned_Data_Sets/Sellers_Data.csv'
INTO TABLE sellers
CHARACTER SET utf8mb4
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\r\n'
IGNORE 1 ROWS
(seller_id, seller_zip_code_prefix, seller_city, seller_state);
```

```
-- ============================================================================
-- Table: product_category_name_translation
-- ============================================================================
DROP TABLE IF EXISTS product_category_name_translation;
CREATE TABLE product_category_name_translation
(
    product_category_name VARCHAR(255),
    product_category_name_english VARCHAR(255),
    PRIMARY KEY (product_category_name)
) CHARACTER SET utf8mb4 COLLATE utf8mb4_unicode_ci;



LOAD DATA INFILE 'D:/Business Analyst Training/Analytics
Projects/CSV_Cleaned_Data_Sets/Product_Category_Name_Translation_Data.csv'
INTO TABLE product_category_name_translation
CHARACTER SET utf8mb4
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\r\n'
IGNORE 1 ROWS
(product_category_name, product_category_name_english);
```

STEP 3:- Writing Queries to Fetch Data and Solve Questions in MYSQL

After Importing Data Sets in MySQL, I have written queries to fetch relevant data from the tables to solve data analysis questions following are the codes give below:-

SHOW DATABASES;

USE ECOMMERCE;

-- Q1 : What is the total revenue & orders generated by Olist, and how has it changed month by month to identify growth or decline trends?

-- Explanation: Calculates total revenue earned by Olist and number of orders month by month to track business growth or decline.

```
CREATE OR REPLACE VIEW OLIST_REVENUE_PERFORMANCE AS
WITH Order_Revenue AS (
   SELECT
      order_id,
      SUM(payment_value) AS total_payment
   FROM order_payments
   GROUP BY order_id
)
SELECT
   YEAR(orders.order_purchase_timestamp) AS Year,
   DATE_FORMAT(orders.order_purchase_timestamp, '%m-%Y') AS sales_month,
   ROUND(SUM(Order_Revenue.total_payment), 2) AS revenue,
   COUNT(DISTINCT orders.order_id) AS orders
FROM orders
JOIN Order_Revenue
   ON orders.order_id = Order_Revenue.order_id
WHERE orders.order_status = 'delivered'
GROUP BY Year, sales_month
```

```
ORDER BY Year, sales_month;


-- Total revenue
SELECT ROUND(SUM(revenue),2) FROM olist_revenue_performance;


-- Monthly revenue & orders
SELECT * FROM olist_revenue_performance;


-- Total orders
SELECT SUM(orders) FROM olist_revenue_performance;


-------------------------------------------------------
-- Q2  Which customer states contribute the most revenue, indicating regions for marketing focus?

SELECT
   customers.customer_state AS state,
   ROUND(SUM(order_payments.payment_value), 2) AS total_revenue
FROM orders
JOIN customers
   ON orders.customer_id = customers.customer_id
JOIN order_payments
   ON orders.order_id = order_payments.order_id
WHERE orders.order_status = 'delivered'  -- consider only completed orders
GROUP BY customers.customer_state
ORDER BY total_revenue DESC;
-------------------------------------------------------
-- Q3 Which payment methods are most commonly used for delivered orders,
-- and what proportion of total payments does each method represent?
-- Additionally, what insights can be drawn about customer payment preferences on Olist?
```

-- Shows how often each payment type was used and its share of total payments

---

```
WITH Delivered_Orders AS (
    -- 1) Select only delivered orders
    SELECT DISTINCT
        orders.order_id
    FROM orders
    WHERE orders.order_status = 'delivered'
),

Delivered_Payments AS (
    -- 2) Join to payments to get all payment entries for delivered orders
    SELECT
        order_payments.order_id,
        order_payments.payment_type
    FROM order_payments
    LEFT JOIN Delivered_Orders
      ON order_payments.order_id = Delivered_Orders.order_id
    WHERE order_payments.payment_type IS NOT NULL
)

-- 3) Count total times each payment type was used and its percentage of all payments
SELECT
    Delivered_Payments.payment_type AS payment_type,
    COUNT(*) AS times_used,
    ROUND(
        COUNT(*) * 100.0 / (SELECT COUNT(*) FROM Delivered_Payments),
        3
    ) AS percentage_of_total_payments
```

```
FROM Delivered_Payments
GROUP BY Delivered_Payments.payment_type
ORDER BY times_used DESC;
```

-- Q4. Which product categories generate the highest revenue, showing where the business earns most?

-- Q5. Which product categories sell the most units, indicating popularity vs revenue efficiency?

-- Q6. Which product categories have the highest average revenue per order , suggesting opportunities for upselling?

-- Explanation: total orders, revenue, and avg revenue per order for each product category.

```
WITH Delivered_Orders AS (
    SELECT order_id
    FROM orders
    WHERE order_status = 'delivered'
),
Order_Revenue AS (
    SELECT order_id, SUM(payment_value) AS total_payment
    FROM order_payments
    GROUP BY order_id
),
Item_Counts AS (
    SELECT order_id, COUNT(*) AS total_items
    FROM order_items
    GROUP BY order_id
),
Order_Items_With_Category AS (
    SELECT
```

```
  order_items.order_id,
      order_items.product_id,
      COALESCE(product_category_name_translation.product_category_name_english,'Unknown') AS product_category,
      1 AS item_count
   FROM order_items
   LEFT JOIN products ON order_items.product_id = products.product_id
   LEFT JOIN product_category_name_translation
      ON products.product_category_name = product_category_name_translation.product_category_name
)
SELECT
   product_category,
   COUNT(DISTINCT Order_Items_With_Category.order_id) AS total_orders,
   ROUND(SUM(Order_Revenue.total_payment * Order_Items_With_Category.item_count / Item_Counts.total_items),2) AS total_revenue,
   ROUND(SUM(Order_Revenue.total_payment * Order_Items_With_Category.item_count / Item_Counts.total_items) / COUNT(DISTINCT Order_Items_With_Category.order_id),2) AS avg_revenue_per_order
FROM Order_Items_With_Category
JOIN Delivered_Orders ON Order_Items_With_Category.order_id = Delivered_Orders.order_id
LEFT JOIN Order_Revenue ON Order_Items_With_Category.order_id = Order_Revenue.order_id
LEFT JOIN Item_Counts ON Order_Items_With_Category.order_id = Item_Counts.order_id
GROUP BY product_category
ORDER BY total_revenue DESC;
```

-- Q7 What are Top 3 product categories by number of order cancellations or refunds, indicating potential quality or logistics issues?

-- Q8 What are Top 5 product categories by revenue lost due to order cancellations ?

```
CREATE OR REPLACE VIEW CANCELLED_ORDERS AS
SELECT
   COALESCE(Product_Category_Name_Translation.product_category_name_english, 'Unknown') AS
product_category_name,
   COUNT(DISTINCT Orders.order_id) AS cancelled_orders,
   ROUND(COALESCE(SUM(Order_Items.price), 0), 2) AS revenue_lost
FROM
   Orders
   LEFT JOIN Order_Items ON Orders.order_id = Order_Items.order_id
   LEFT JOIN Products ON Order_Items.product_id = Products.product_id
   LEFT JOIN Product_Category_Name_Translation
      ON Products.product_category_name =
Product_Category_Name_Translation.product_category_name
WHERE
   Orders.order_status = 'canceled'
GROUP BY
   COALESCE(Product_Category_Name_Translation.product_category_name_english, 'Unknown')
ORDER BY
   revenue_lost DESC;

SELECT * FROM CANCELLED_ORDERS;
SELECT SUM(revenue_lost) FROM CANCELLED_ORDERS;
SELECT SUM(cancelled_orders) FROM CANCELLED_ORDERS;

-----------------------------------------------------------------
-- Q9: Distribution of Customers Across States

CREATE OR REPLACE VIEW Customers_Location_Info AS
SELECT
   customer_state AS State,
```

```
 COUNT(DISTINCT customers.customer_unique_id) AS No_Of_Customers,
   ROUND(
      COUNT(DISTINCT customers.customer_unique_id) * 100.0
      / SUM(COUNT(DISTINCT customers.customer_unique_id)) OVER (),
      2
   ) AS Percentage_Of_Total_Customers
FROM customers
GROUP BY customer_state;

SELECT * FROM Customers_Location_Info ORDER BY Percentage_Of_Total_Customers DESC;

-----------------------------------------------------
```

-- Q9: Distribution of sellers across states

```
CREATE OR REPLACE VIEW SELLERS_LOCATION_INFO AS
SELECT
   seller_state AS State,
   COUNT(DISTINCT seller_id) AS No_Of_Sellers,
   ROUND(
      COUNT(DISTINCT seller_id) * 100.0
      / SUM(COUNT(DISTINCT seller_id)) OVER (),
      2
   ) AS Percentage_Of_Total_Sellers
FROM sellers
GROUP BY seller_state;

SELECT * FROM SELLERS_LOCATION_INFO ORDER BY Percentage_Of_Total_Sellers DESC;
```

-- Q9.How Sellers & Customers are distributed across different states and which states have low or no Seller/Customers,

-- explaining how it can effect the business

-- Combined customer & seller info by state

-- Explanation: Finds states with customers but no sellers, and vice versa.

```
SELECT
   COALESCE(CUSTOMERS_LOCATION_INFO.State, SELLERS_LOCATION_INFO.State) AS State,
   IFNULL(CUSTOMERS_LOCATION_INFO.No_Of_Customers,0) AS No_Of_Customers,
   IFNULL(SELLERS_LOCATION_INFO.No_Of_Sellers,0) AS No_Of_Sellers
FROM CUSTOMERS_LOCATION_INFO
LEFT JOIN SELLERS_LOCATION_INFO
   ON CUSTOMERS_LOCATION_INFO.State = SELLERS_LOCATION_INFO.State
UNION
SELECT
   COALESCE(CUSTOMERS_LOCATION_INFO.State, SELLERS_LOCATION_INFO.State) AS State,
   IFNULL(CUSTOMERS_LOCATION_INFO.No_Of_Customers,0) AS No_Of_Customers,
   IFNULL(SELLERS_LOCATION_INFO.No_Of_Sellers,0) AS No_Of_Sellers
FROM CUSTOMERS_LOCATION_INFO
RIGHT JOIN SELLERS_LOCATION_INFO
   ON CUSTOMERS_LOCATION_INFO.State = SELLERS_LOCATION_INFO.State
ORDER BY No_Of_Customers DESC;
```

-- Q10.How has the number of active sellers changed month by month from 2016 to 2018,

-- highlighting marketplace growth and peak engagement periods ?

-- Explanation: Counts active sellers per quarter to understand marketplace engagement.

```
SELECT
    YEAR(order_purchase_timestamp) AS Active_Year,
    Monthname (order_purchase_timestamp) AS Month_Name,
    QUARTER(order_purchase_timestamp) AS Active_Quarter,
    COUNT(DISTINCT order_items.seller_id) AS Total_Active_Sellers
FROM orders
JOIN order_items ON order_items.order_id=orders.order_id
WHERE orders.order_status NOT IN ('canceled','unavailable')
GROUP BY Active_Year,Month_Name, Active_Quarter;


-----------------------------------------------------
-- Q11: What percentage of total revenue is contributed by the top 10% of sellers, revealing dependency on key sellers?


-- Explanation: Identifies top sellers and calculates their share of total revenue.


WITH Seller_Revenue AS (
    SELECT
        sellers.seller_id,
        SUM(order_payments.total_payment * seller_items.seller_item_count / total_items.total_items) AS total_revenue
    FROM sellers
    LEFT JOIN (
        SELECT order_id, seller_id, COUNT(*) AS seller_item_count
        FROM order_items
        GROUP BY order_id, seller_id
    ) seller_items ON sellers.seller_id = seller_items.seller_id
    LEFT JOIN (
        SELECT order_id, COUNT(*) AS total_items
        FROM order_items
```

```
GROUP BY order_id
    ) total_items ON seller_items.order_id = total_items.order_id
    LEFT JOIN (
        SELECT order_id, SUM(payment_value) AS total_payment
        FROM order_payments
        GROUP BY order_id
    ) order_payments ON seller_items.order_id = order_payments.order_id
    LEFT JOIN orders ON seller_items.order_id = orders.order_id
        AND orders.order_status = 'delivered'
    GROUP BY sellers.seller_id
),
Seller_Rank AS (
    SELECT
        seller_id,
        total_revenue,
        PERCENT_RANK() OVER (ORDER BY total_revenue DESC) AS revenue_rank
    FROM Seller_Revenue
)
SELECT
    ROUND(
        SUM(CASE WHEN revenue_rank <= 0.1 THEN total_revenue ELSE 0 END)
        * 100.0 / SUM(total_revenue), 2
    ) AS percentage_revenue_from_top_10_percent
FROM Seller_Rank;
```

-- Q12: What percentage of customers are repeat buyers, and what share of revenue do they contribute, showing customer loyalty impact?

-- Explanation: Shows customer loyalty impact by identifying repeat buyers and their revenue contribution.

```
WITH Order_Revenue AS (
   SELECT order_id, SUM(payment_value) AS total_payment
   FROM Order_Payments
   GROUP BY order_id
),
Customers_Summary AS (
   SELECT
      customers.customer_unique_id,
      COUNT(DISTINCT orders.order_id) AS number_of_orders,
      SUM(Order_Revenue.total_payment) AS total_revenue
   FROM Customers
   JOIN Orders ON customers.customer_id = orders.customer_id
   JOIN Order_Revenue ON orders.order_id = Order_Revenue.order_id
   WHERE orders.order_status = 'delivered'
   GROUP BY customers.customer_unique_id
)
SELECT
   ROUND(COUNT(CASE WHEN number_of_orders > 1 THEN customer_unique_id END) * 100.0 / COUNT(customer_unique_id), 2) AS percentage_of_repeat_customers,
   ROUND(SUM(CASE WHEN number_of_orders > 1 THEN total_revenue ELSE 0 END) * 100.0 / SUM(total_revenue), 2) AS percentage_of_revenue_from_repeaters
FROM Customers_Summary;
```

-- Q13: How do customer reviews and ratings affect sales and revenue performance on Olist ?

-- -- This query shows how many orders and how much revenue each review score generated

-- Only delivered orders are considered to ensure accurate sales performance

```
SELECT
    order_reviews.review_score,              -- Review rating (1 to 5 stars)
    COUNT(DISTINCT order_reviews.order_id) AS total_orders,  -- Number of orders for that score
    ROUND(SUM(order_payments.payment_value), 2) AS total_revenue -- Total revenue from those
orders
FROM order_reviews
JOIN orders
    ON order_reviews.order_id = orders.order_id
JOIN order_payments
    ON order_reviews.order_id = order_payments.order_id
WHERE orders.order_status = 'delivered'  -- Only include successfully delivered orders
GROUP BY order_reviews.review_score
ORDER BY order_reviews.review_score ASC;
```

-- Q14: Who are the top 1% of customers by lifetime spend,

-- and what percentage of total revenue do they represent, highlighting high-value customers?

-- Explanation: Identifies high-value customers and their share of total revenue.

```
WITH CustomerLifetimeValue AS (
    SELECT
        Customers.customer_unique_id,
        SUM(Order_Payments.payment_value) AS lifetime_spend
    FROM Customers
    JOIN Orders ON Customers.customer_id = Orders.customer_id
    JOIN Order_Payments ON Orders.order_id = Order_Payments.order_id
    WHERE Orders.order_status = 'delivered'
    GROUP BY Customers.customer_unique_id
```

```
),
CustomerPercentiles AS (
    SELECT
        CustomerLifetimeValue.customer_unique_id,
        CustomerLifetimeValue.lifetime_spend,
        NTILE(100) OVER (ORDER BY CustomerLifetimeValue.lifetime_spend DESC) AS
percentile_rank
    FROM CustomerLifetimeValue
)
SELECT
    ROUND(SUM(CASE WHEN CustomerPercentiles.percentile_rank = 1 THEN
CustomerPercentiles.lifetime_spend ELSE 0 END) * 100.0 /
SUM(CustomerPercentiles.lifetime_spend),2) AS percentage_revenue_from_top_1_percent
FROM CustomerPercentiles;


------------------------------------------------------------
-- ************************************************************
```

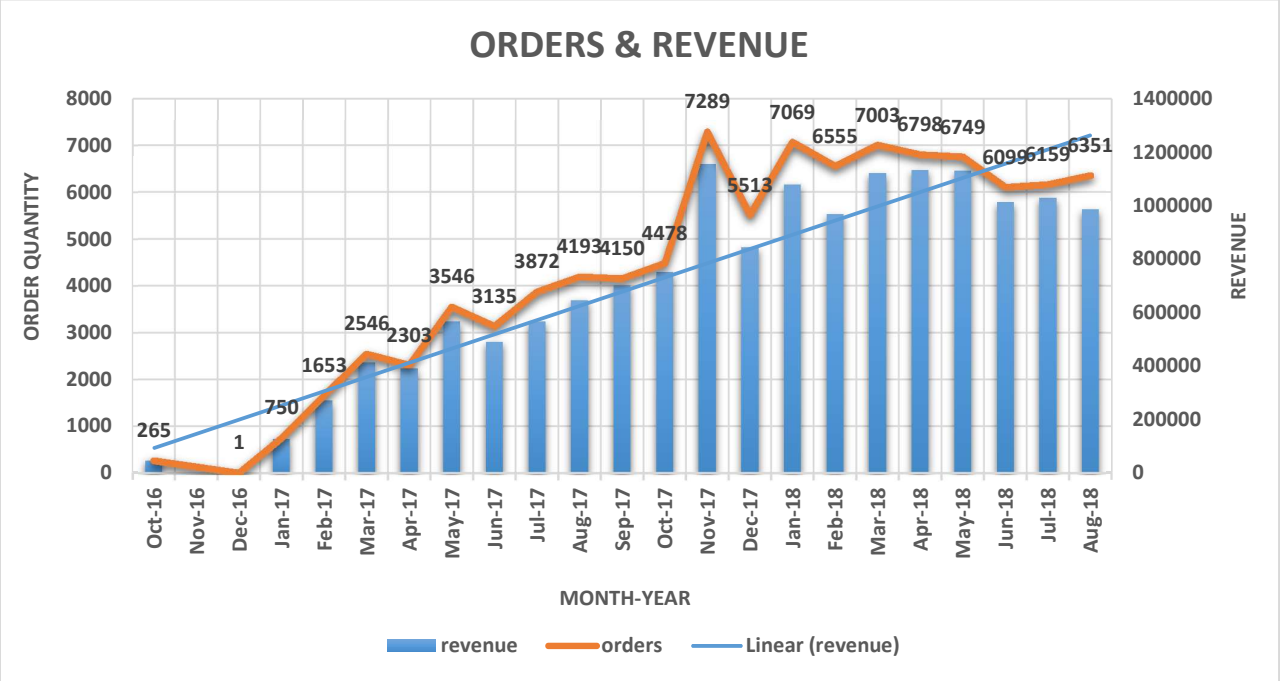
STEP 4:-Data Analysis & Interpretation in Excel (Business Questions)

After writing SQL Queries to fetch relevant data to solve questions, I have exported the Results of Queries in form of tables into Excel for further interpretation , it allows me to filter ,visualize and write interpretation for the data without writing complex queries.

ANSWERS:- Business Questions

Revenue & Orders (Business Health)

1. What is the total revenue generated by Olist, and how has it changed month by month to identify growth or decline trends?



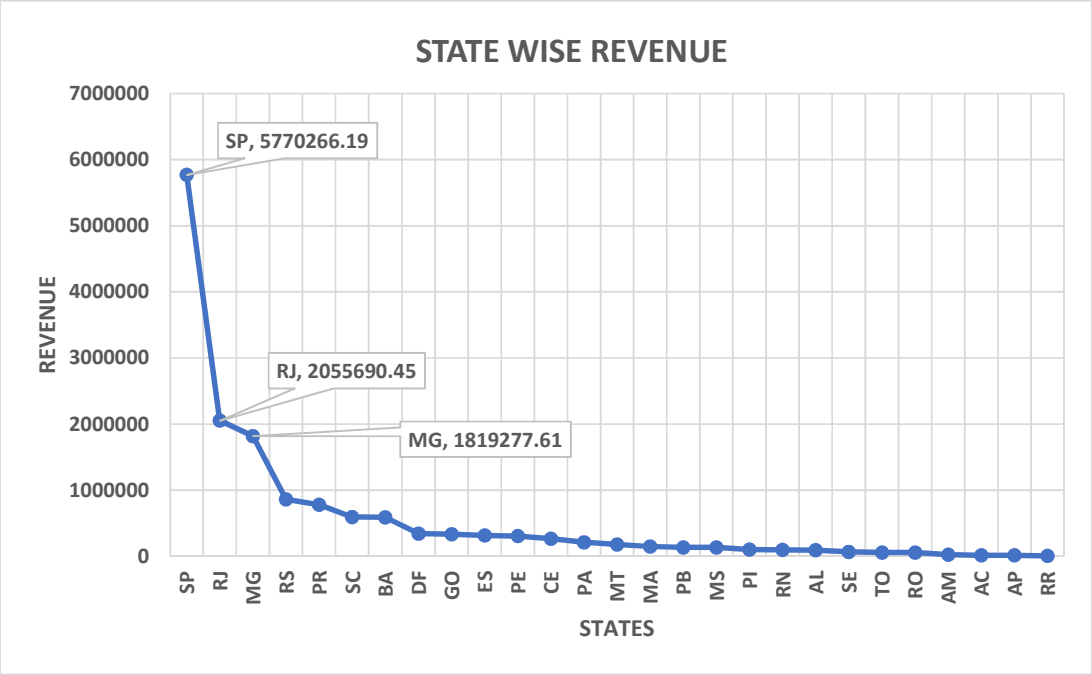
Interpretation:

Olist generated a **total revenue of R$ 15.32 million** from delivered orders between October 2016 and August 2018. Revenue grew from R$ 1,27,546 in January 2017 to a (**peak of R$ 11,53,528, 7289 orders) in November 2017**, then stabilized between R$ 9,85,414–R$ 11,32,934 through 2018. Orders followed a similar trend, rising from 750 to 7,289 and then staying around 6,099–7,069, indicating that growth was driven by increasing sales volume. The trend reflects strong growth in 2017 followed by a stable performance phase in 2018.

Actionable Insight:

To sustain growth, Olist should focus on increasing monthly orders through promotions or targeted campaigns and explore ways to boost sales during slower months.

2. Which customer states contribute the most revenue, indicating regions for marketing focus?



Interpretation:

The state of **São Paulo (SP)** overwhelmingly leads with **total revenue of R$ 5,770,266.19**, contributing the **largest share** of Olist's sales. It's followed by **Rio de Janeiro (RJ)** with **R$ 2,055,690.45** and **Minas Gerais (MG)** with **R$ 1,819,277.61**. Together, these top three states generate more than 60% of total revenue, highlighting them as core markets.

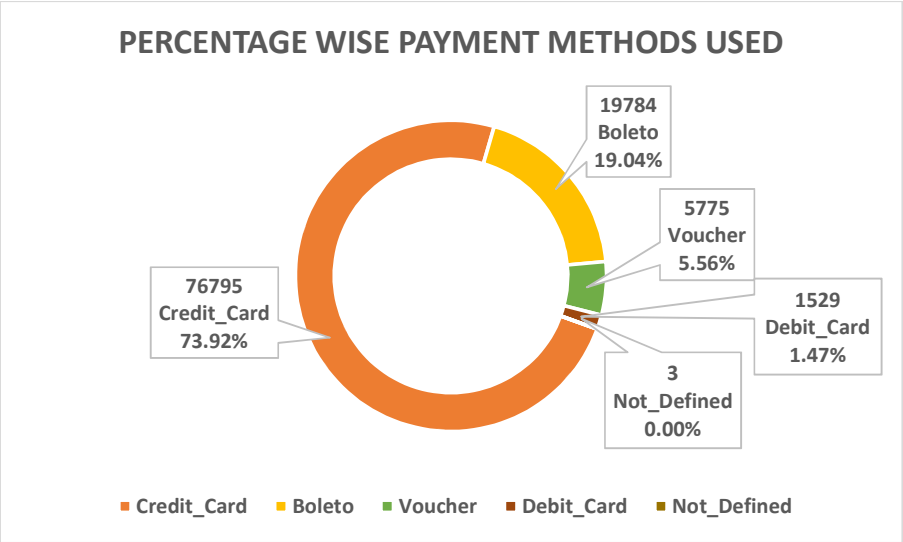
In contrast, northern and north-eastern states like Roraima (RR, R$ 9,039.52), Amapá (AP, R$ 16,141.81), and Acre (AC, R$ 19,586.25) contribute minimally, signalling low market penetration in these areas.

Actionable Insights:-

Focus marketing and logistics on high-performing states (SP, RJ, MG) to strengthen brand dominance.

Expand outreach and awareness in underrepresented northern states to tap untapped customer potential.

3. Which payment methods are most commonly used for delivered orders, and what proportion of total payments does each method represent? Additionally, what insights can be drawn about customer payment preferences on Olist?



Interpretation:-

The **majority of Olist customers prefer credit cards**, **accounting for 73.92 % of payments (76,795 transactions)**. **Boleto** — a popular Brazilian payment slip — **follows with 19.04 %** (19,784 transactions), appealing to customers without credit access. Vouchers (5.56%) and debit cards (1.47%) are used less frequently, while "Not Defined" payments are negligible.

This indicates that Olist's customers strongly favor credit-based purchases, likely due to installment flexibility, while cash-equivalent methods like boleto remain relevant for budget-conscious or unbanked users.

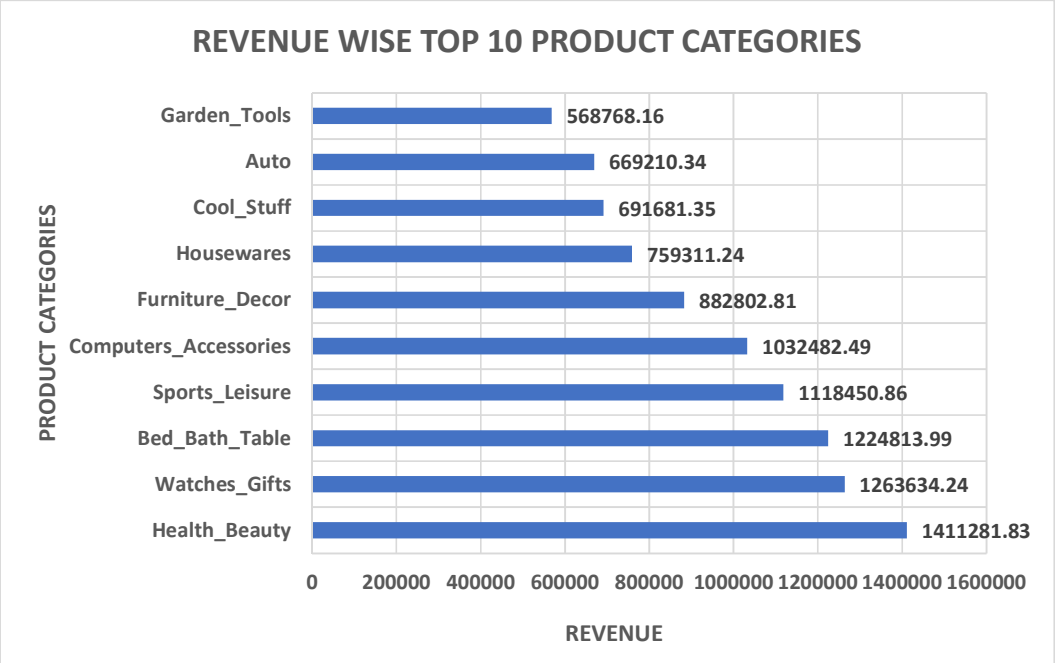
Actionable Insights:-

Optimize checkout and promotions around credit card and boleto options to support customer convenience.

Encourage debit and voucher adoption through discounts or loyalty rewards to diversify payment reliance.

Products & Categories:-

4. Which top 3 product categories generate the highest revenue, showing where the business earns most?



Interpretation:

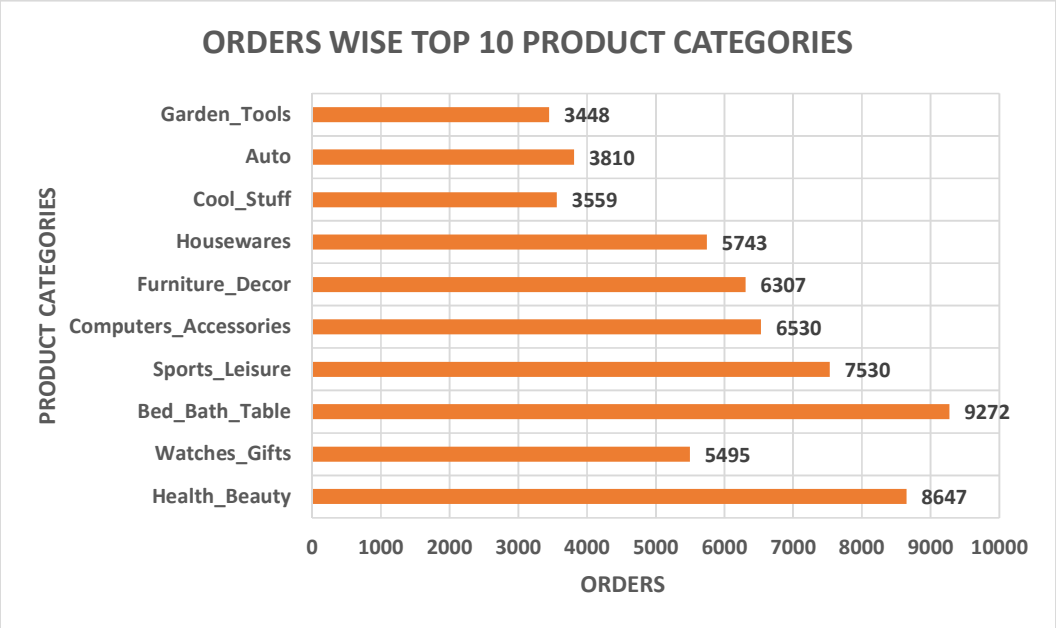
The **top revenue-generating categories are Health_Beauty (R$ 1,411,282), Watches_Gifts (R$ 1,263,634), and Bed_Bath_Table (R$ 1,224,814)**, driven by high order volumes and consistent customer demand. Meanwhile, categories such as Computers (R$ 227,459) and Home_Appliances_2 (R$ 118,130) show high revenue per order but contribute less overall due to lower sales volume.

Overall, Olist's revenue strength lies in high-volume lifestyle categories, indicating that broad appeal and repeat purchases drive growth more than premium pricing alone.

Actionable Insight:

Focus marketing and inventory expansion on high-volume categories like Health_Beauty and Bed_Bath_Table, while strategically promoting high-value, low-volume segments (e.g., Computers, Small_Appliances) to balance volume and profitability.

5. Which product categories sell the most units, indicating popularity vs revenue efficiency?



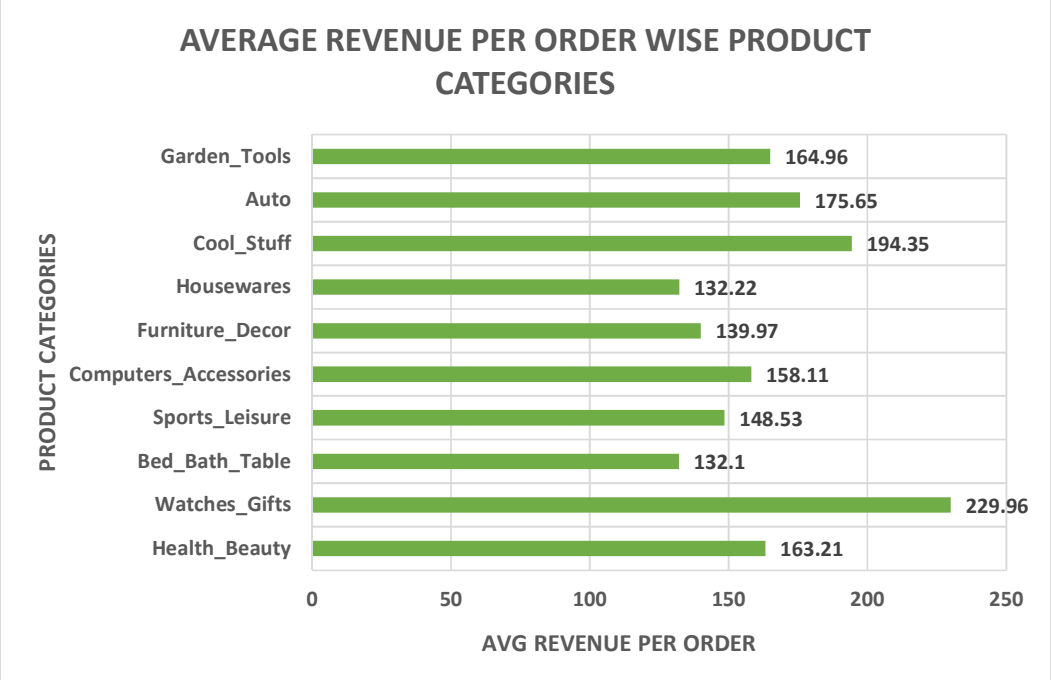
Interpretation:

The categories with the most units sold are **Bed_Bath_Table (9,272 orders), Health_Beauty (8,647 orders), and Sports_Leisure (7,530 orders)**, highlighting strong customer demand in home and personal care products. While these categories drive volume, Watches_Gifts (R$ 1.26 M) and Health_Beauty (R$ 1.41 M) deliver higher revenue efficiency, balancing both popularity and profitability. In contrast, categories like Computers (177 orders, R$ 227 K) show low volume but high revenue per sale.

Actionable Insight:

Maintain strong inventory and promotional focus on high-volume categories to sustain growth, while improving visibility of high-value, low-volume segments to maximize overall sales efficiency.

6. Which top 5 product categories have the highest average revenue per order , suggesting opportunities for upselling?



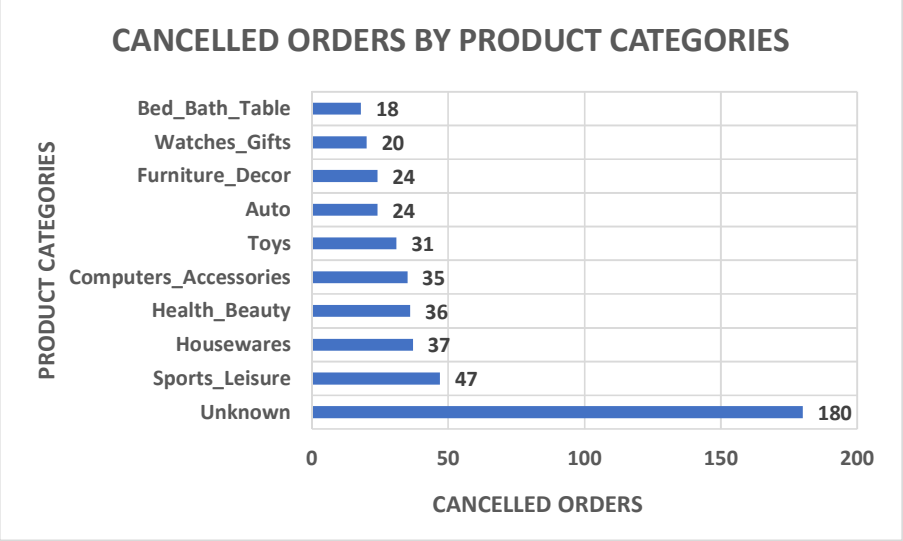
Interpretation:

The **top 5 categories** by average revenue per order are **Computers (R$ 1,285.08), Small_Appliances_Home_Oven_And_Coffee (R$ 684.29), Home_Appliances_2 (R$ 520.40), Agro_Industry_And_Commerce (R$ 430.56), and Musical_Instruments (R$ 330.91)**. These categories have high per-order value but lower total orders, indicating niche products with high spending per purchase.

Actionable Insight:

Focus marketing and upselling strategies on these high-value categories to increase revenue efficiency, while ensuring stock availability and targeted promotions to convert niche demand into higher sales.

7. What are Top 3 product categories by number of order cancellations or refunds, indicating potential quality or logistics issues?



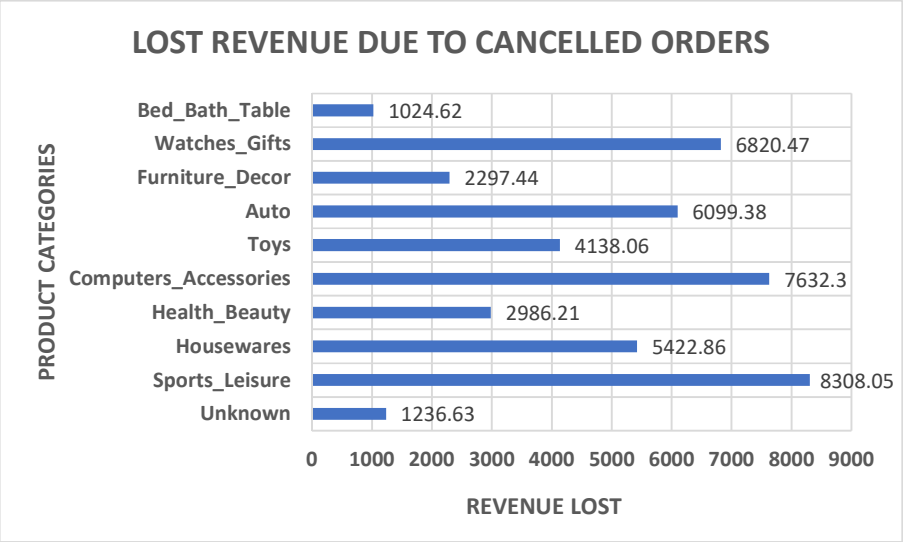
Interpretation:

The categories with the **most cancelled orders** are **Sports_Leisure (47 cancellations, R$8,308 lost), Housewares (37 cancellations, R$5,423 lost), and Health_Beauty (36 cancellations, R$2,986 lost).** Although Cool_Stuff had fewer cancellations (15), the revenue lost was the highest at R$14,455, indicating high-value items. The data suggests potential quality, logistics, or customer satisfaction issues in these categories.

Actionable Insight:

The operations and quality teams should investigate frequent cancellation reasons in high-impact categories like Cool_Stuff, Sports_Leisure, and Housewares, implement process improvements, and enhance customer communication to reduce revenue loss.

8. What are Top 5 product categories by revenue lost due to order cancellations ?



Interpretation:

Using the product price to calculate revenue lost due to cancellations, the highest impact categories are **Cool_Stuff (R$ 14,455.46), Sports_Leisure (R$ 8,308.05), Computers_Accessories (R$ 7,632.30), Watches_Gifts (R$ 6,820.47), and Auto (R$ 6,099.38).**

This method shows the potential revenue lost if these orders had been fulfilled, independent of any actual payments received. The data highlights high-value categories where cancellations cause significant missed revenue opportunities.

Actionable Insight:

The operations and product teams should prioritize reducing cancellations in high-price categories by improving stock accuracy, logistics reliability, and clear customer communication to protect potential revenue

---

Sellers & Marketplace

9. How Sellers & Customers are distributed across different states and which states have low or no Seller/Customers, explaining how it can effect the business

Interpretation:

The data shows that Olist's marketplace is **heavily concentrated in** a few southern and **south-eastern states. São Paulo (40,302 customers, 1,849 sellers), Minas Gerais (11,259 customers, 244 sellers), and Paraná (4,882 customers, 349 sellers)** lead in both customers and sellers.

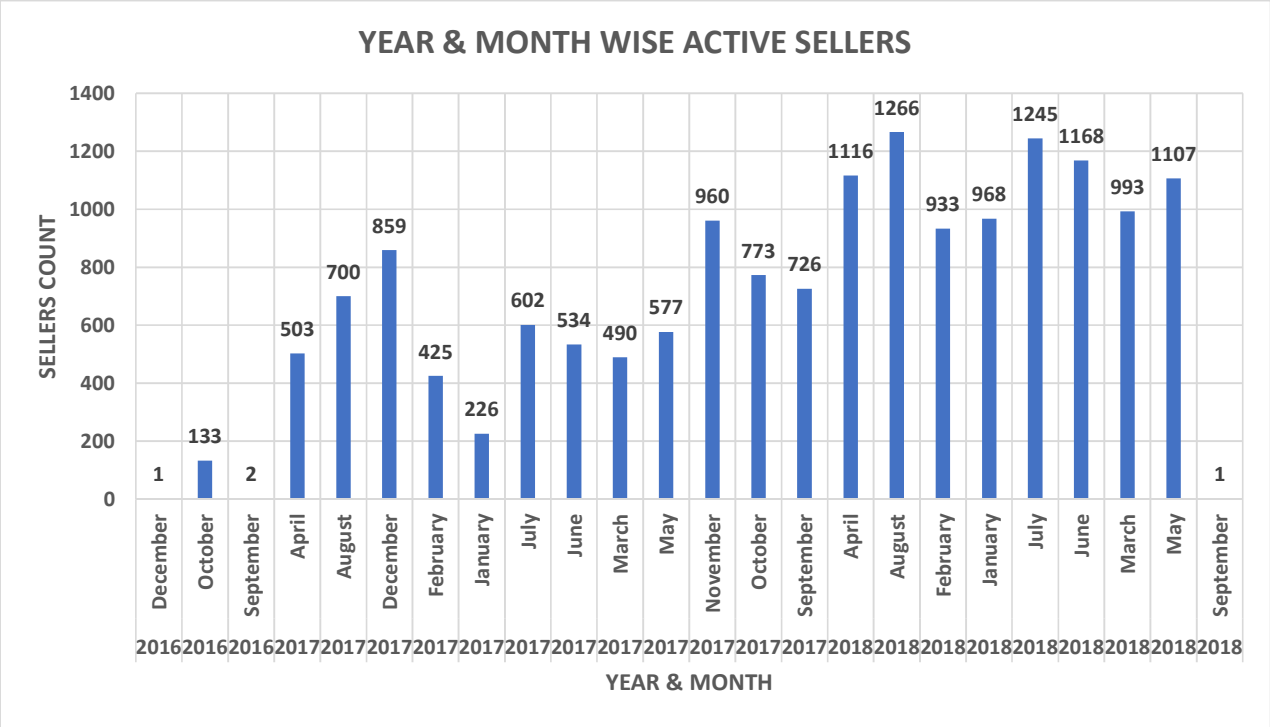
However, states like Bahia (3,277 customers, 19 sellers) and Rio de Janeiro (12,384 customers, 171 sellers) show strong demand but fewer sellers, causing regional imbalance.

Several states — Amapá, Roraima, Tocantins, and Alagoas — have no active sellers, while others like Pará and Maranhão have only one, showing limited reach in northern regions.

Actionable Insight:

The regional growth and seller acquisition teams should expand seller recruitment in high-demand but low-supply states such as Bahia and Rio de Janeiro, and build partnerships in no-seller states to strengthen coverage and capture untapped demand.

10. How has the number of active sellers changed month by month from 2016 to 2018, highlighting marketplace growth and peak engagement periods ?



Interpretation:

The number of active sellers grew steadily from 133 in October 2016 to over 1,200 in mid-2018, reflecting strong marketplace expansion and increasing engagement. There are minor anomalies in the early months (like 1 seller in September 2016 and September 2018), likely due to incomplete reporting. The overall trend shows consistent growth across quarters, with the **highest active sellers in August 2018 (1,266).**

Actionable Insight:

The marketplace operations team should continue onboarding initiatives and support programs to sustain seller growth, while the marketing team can engage top-performing sellers to encourage higher sales and retention.

11. What percentage of total revenue is contributed by the top 10% of sellers, revealing dependency on key sellers?

Interpretation:

The top 10% of sellers contribute 66.73% of total revenue, indicating that a relatively small group of sellers drives the majority of marketplace earnings. This reveals a significant dependency on key sellers and highlights their critical role in overall business performance.

Actionable Insight:

The marketplace management team should focus on retaining and supporting these top sellers through loyalty programs, priority support, and targeted incentives to reduce risk and ensure sustained revenue growth.

Customers & Retention

12. What percentage of customers are repeat buyers, and what share of revenue do they contribute, showing customer loyalty impact?

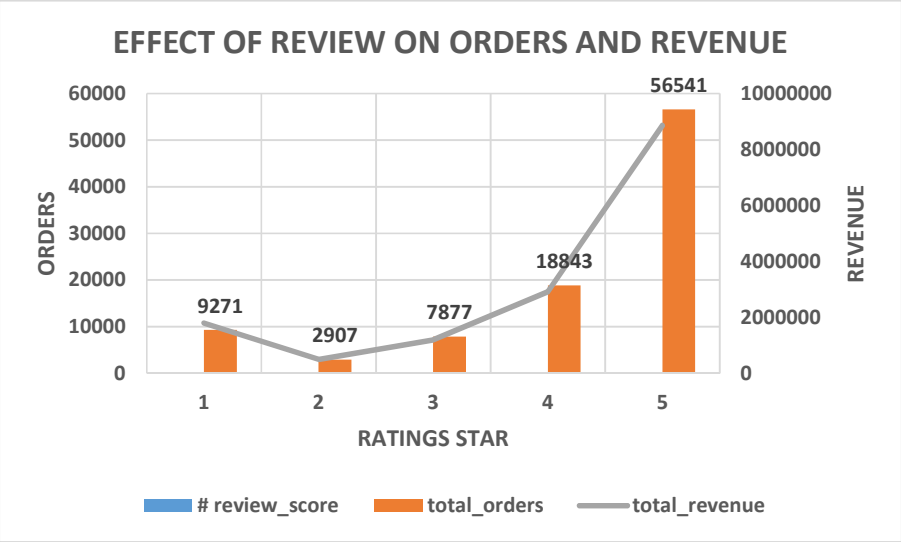
Interpretation:

Only 3% of customers made repeat purchases, contributing 5.6% of total revenue. This indicates that while most customers buy only once, the small base of loyal buyers still generates a slightly higher revenue share, showing moderate value concentration among repeaters. The low repeat rate highlights limited customer retention and suggests that most growth currently depends on acquiring new buyers rather than maintaining existing ones.

Actionable Insight:

The marketing and CRM teams should develop loyalty and re-engagement programs — such as personalized offers, post-purchase follow-ups, or subscription models — to increase repeat purchase rates and strengthen long-term customer value.

13. How do customer reviews and ratings affect sales and revenue performance on Olist ?



Interpretation:-

Analysis of Olist customer reviews shows a strong positive relationship between review scores and both total orders and revenue. One-star and two-star products account for only 9,271 and 2,907 orders, generating R$ 17,93,847.7 and R$ 4,94,708.73, respectively, while four- and five-star products contribute 18,843 and 56,541 orders, generating R$ 29,18,549.73 and R$ 88,62,157.38. The **correlation between review scores and orders is ~0.95, and with revenue ~0.97, confirming higher-rated products drive performance.**

Actionable insights:

Prioritize quality improvements for low-rated products and encourage reviews for high-rated products to boost trust, sales, and revenue.

14. Who are the top 1% of customers by lifetime spend, and what percentage of total revenue do they represent, highlighting high-value customers?

Interpretation:

The **top 1% of customers contribute 10.35% of total revenue**, showing that a small but powerful group of high-value buyers significantly impacts overall sales. This indicates moderate revenue concentration, meaning Olist benefits from loyal, high-spending customers but isn't overly dependent on them.

Actionable Insight:

Olist's CRM and marketing teams should create exclusive retention and loyalty programs for these high-value buyers—such as early access to deals or personalized recommendations—to increase repeat purchases and lifetime value.

CONCLUSION

Olist — Executive snapshot (delivered orders, Oct 2016–Aug 2018)

Summary:-

Total delivered-order revenue = **R$ 15,422,461.77**. Growth peaked in **November 2017** and then settled into a stable range through 2018.

Key facts (quick numbers)

- **Total revenue:** R$ **15,422,461.77**

- **Top states (combined):**

  São Paulo (R$ 5,770,266.19), Rio de Janeiro (R$ 2,055,690.45), Minas Gerais (R$ 1,819,277.61)→ R$ 9,645,234.25 (~62.5%) of revenue.

- **Payment mix (by number of transactions including payment in installments):**

  Credit card 76,795 (73.92%), Boleto 19,784 (19.04%), Voucher 5,775 (5.56%), Debit card 1,529 (1.47%).

- **Top 3 categories by revenue:**

  Health_Beauty R$ **1,411,281.83**, Watches_Gifts R$ **1,263,634.24**, Bed_Bath_Table R$ **1,224,813.99**.

- **High average revenue per order:**

  Computers **R$ 1,285.08/order**, Small_Appliances_Home_Oven_And_Coffee **R$ 684.29/order**, Home_Appliances_2 **R$ 520.40/order**.

- **Major cancellation losses (price-based):**

  Cool_Stuff R$ **14,455.46**, Sports_Leisure R$ **8,308.05**, Computers_Accessories R$ **7,632.30**.

- **Marketplace concentration:**

  Top 10% sellers → **66.73%** of revenue.

- **Customer loyalty:**

  Repeat buyers = **3%** of customers, contributing **5.6%** of revenue. Top 1% customers = **10.35%** of revenue.

- **Reviews effect:**

  Ratings strongly correlate with performance (orders r ≈ **0.95**, revenue r ≈ **0.97**).

Short conclusions (what this means for Olist)

- Revenue is concentrated in a few states and a few product categories — **scale wins** (volume in popular categories) plus **select premium products** (high ARPO) drive results.
- Payments: most customers prefer **credit cards** (installments matter) — checkout & financing are strategic levers.
- Cancellations are small in count but costly in select categories — operational fixes can recover measurable revenue.
- Marketplace risk: heavy dependence on top sellers — retention of these sellers is business critical.
- Reviews matter: higher ratings → substantially more orders and revenue.

Top 3 immediate actions (who should act & what to do)

1. **Commercial / Growth:** Prioritize São Paulo, Rio de Janeiro and Minas Gerais for targeted acquisition, inventory buffers and faster fulfilment. Run 4–6 week promo pushes for **Health_Beauty** and **Bed_Bath_Table** to lift monthly orders.
2. **Marketing / CRM:** Launch a loyalty & re-engagement program (email + incentives + easy financing offers) to increase repeat purchase rate; target high-ARPO buyers with tailored bundles and financing.
3. **Operations / Product:** Root-cause cancellations in **Cool_Stuff** and **Sports_Leisure** (stock accuracy, seller SLAs, returns). Implement short-term seller performance audits and clearer product pages to reduce refund/cancel rates.

Quick KPIs to track (minimum weekly)

- Delivered revenue (R$) — weekly & M/M trend
- Cancellation rate and **R$ lost by category**
- Share of revenue from top 10% sellers (%)
- Repeat-customer rate (%) and revenue from repeaters (%)
- Payment-method mix (%) and installment use