**E-commerce Analytics Capstone Project Report**

**1. Dataset Overview**

The dataset used for this project was sourced from Kaggle and published by Olist, one of Brazil's leading marketplace platforms. It comprises approximately 100,000 orders made between 2016 and 2018 across various product categories and marketplaces. The dataset provides a multi-faceted view of e-commerce transactions, with anonymized customer information and a rich set of features including:

* Order status and timestamps
* Pricing and freight details
* Customer geolocation and demographics
* Product descriptions and categories
* Customer feedback and review scores

**2. Dataset Schema and Entity Relationships**

The original dataset consists of several related tables. For this project, we selected a reduced subset of five key tables to ensure performance and maintain analytical consistency between SQL and Excel:

* orders (primary table linking all transactions)
* customers
* order\_items
* products (renamed as products\_new after translation)
* order\_reviews\_clean (cleaned version of order\_reviews)

The analytical dataset is built from five core tables: orders, customers, order\_items, products, and order\_reviews, all interconnected through key relationships that trace the complete customer journey. The orders table serves as the central hub, uniquely identified by order\_id, and links directly to customers via customer\_id to provide customer location and demographic details. Each order can contain multiple entries in the order\_items table, joined via order\_id, which captures product-level details such as price and freight value. The order\_items table connects to the products table through product\_id, enabling access to product attributes including category and description metrics. Customer feedback is captured in the order\_reviews table, also linked by order\_id, which records review scores and timestamps. Together, these tables form a relational schema that supports comprehensive analysis across sales performance, product trends, delivery efficiency, and customer satisfaction.

3. Research Questions & Methodology

|  |  |  |
| --- | --- | --- |
| Section | Research Question | Methodology |
| Customer & Delivery | What is the average delivery time per state? | DATEDIFF between purchase and delivery, grouped by state |
| Customer & Delivery | How many orders fall under each status? | COUNT grouped by order\_status |
| Sales & Revenue | Which categories generate the highest revenue? | SUM of price grouped by product category |
| Sales & Revenue | Total revenue and best month for orders | SUM and COUNT grouped by month |
| Reviews | What is the average review score? | AVG of review\_score |
| Reviews | Which state has fastest delivery? | DATEDIFF + AVG grouped by customer\_state |

**4. Key Findings**

**Customer Metrics**:

* Unique customers: 95,978
* State with most customers: SP (40,315)
* Order Status Breakdown:
* Delivered: 107,577 (97% of total)
* Cancelled: 509
* Others: invoiced, shipped, processing, unavailable

**Sales Metrics:**

Top revenue-generating category: Health & Beauty (BRL 1.25M)\

* Lowest: Security and Services (BRL 283)
* Top order month: November 2017 (8,443 orders)
* Total sales: BRL 13 million+
* Review Insights:
* Average review score: 4.1
* SP state has lowest delivery delay (8 days average), despite highest order volume

**5. Challenges and Solutions**

|  |  |
| --- | --- |
| Challenge | Solution Applied |
| Missing quantity column in order\_items | Interpreted rows as units based on order\_item\_id and repeated product\_id |
| Duplicate review\_id with different order\_ids | Used ROW\_NUMBER in CTE to retain only the first occurrence |
| Large dataset causing Excel performance issues | Reduced to five essential tables and removed NULLs in key columns |
| Column translations missing headers | Manually excluded faulty header row and joined based on values |

**6. Recommendations**

Customer Retention: Focus marketing efforts on SP, the state with most orders and fastest deliveries.

Product Expansion: Consider increasing stock or visibility for top-performing categories like Health & Beauty.

Delivery Optimization: Replicate SP’s logistics model in other high-volume states.

**7. Appendix: SQL Queries**

This query combines key information from orders, customers, order\_items, products\_new, and order\_reviews\_clean to form a consolidated dataset for analysis.

-- Create a unified table combining order, customer, product, and review details

-- Output table: joined\_orders\_dataset

SELECT

-- Order details

o.order\_id,

o.customer\_id,

o.order\_status,

o.order\_purchase\_timestamp,

o.order\_approved\_at,

o.order\_delivered\_customer\_date,

o.order\_estimated\_delivery\_date,

-- Customer details

c.customer\_unique\_id,

c.customer\_zip\_code\_prefix,

c.customer\_city,

c.customer\_state,

-- Order item and pricing details

oi.order\_item\_id,

oi.product\_id,

oi.seller\_id,

oi.shipping\_limit\_date,

oi.price,

oi.freight\_value,

-- Product metadata

p.product\_category\_name\_english,

p.product\_name\_lenght,

p.product\_description\_lenght,

p.product\_photos\_qty,

-- Customer review score

r.review\_score

-- Store result into a new table for persistent analysis

INTO joined\_orders\_dataset

FROM orders o

JOIN customers c

ON o.customer\_id = c.customer\_id

JOIN order\_items oi

ON o.order\_id = oi.order\_id

LEFT JOIN products\_new p

ON oi.product\_id = p.product\_id

LEFT JOIN order\_reviews\_clean r

ON o.order\_id = r.order\_id;

**Q1: What is the average delivery time by state?**

sql

CopyEdit

SELECT

customer\_state,

ROUND(AVG(delivery\_days), 2) AS avg\_delivery\_days,

COUNT(DISTINCT order\_id) AS total\_orders

FROM joined\_orders\_dataset

GROUP BY customer\_state

ORDER BY avg\_delivery\_days DESC;

💡 Insight: Helps identify logistics efficiency and delivery performance across states.

***Q2: What is the count of each order status?***

sql

CopyEdit

SELECT

order\_status,

COUNT(\*) AS total\_orders

FROM joined\_orders\_dataset

GROUP BY order\_status

ORDER BY total\_orders DESC;

💡 Insight: Reveals overall distribution and handling outcomes of orders.

***--Q3: Which product categories generate the highest revenue?***

SELECT

product\_category\_name\_english,

ROUND(SUM(price), 2) AS total\_revenue,

COUNT(product\_id) AS items\_sold

FROM joined\_orders\_dataset

GROUP BY product\_category\_name\_english

ORDER BY total\_revenue DESC;

💡 Insight: Reveals best performing category

***--Q4: What is the average freight value by category?***

SELECT

product\_category\_name\_english,

ROUND(AVG(freight\_value), 2) AS avg\_freight

FROM joined\_orders\_dataset

GROUP BY product\_category\_name\_english

ORDER BY avg\_freight DESC;

***--Q5: What is the average review score per product category?***

SELECT

product\_category\_name\_english,

ROUND(AVG(review\_score), 0) AS avg\_review\_score,

COUNT(review\_score) AS review\_count

FROM joined\_orders\_dataset

WHERE review\_score IS NOT NULL

GROUP BY product\_category\_name\_english

ORDER BY avg\_review\_score ASC;

***Q6:--Do delivery delays affect review scores?***

SELECT

review\_score,

ROUND(AVG(delivery\_days), 2) AS avg\_delivery\_days,

COUNT(\*) AS review\_count

FROM joined\_orders\_dataset

WHERE review\_score IS NOT NULL AND delivery\_days IS NOT NULL

GROUP BY review\_score

ORDER BY review\_score;