QMB 6358 Final Project

End-to-End Data Analysis Using the Olist E-Commerce Dataset

Akio Azevedo Maebayashi

November 30, 2025

1 Dataset Selection

The dataset used in this analysis originates from the Brazilian e-commerce retailer Olist, a large marketplace platform that connects sellers and buyers nationwide. The full dataset is publicly available through the Olist Kaggle repository, containing over 100,000 orders, 3,000 product categories, and detailed transactional, logistical, and customer-level information.

This dataset was selected because it offers:

- A rich combination of numerical and categorical variables.
- A realistic business context involving payments, shipping, customer behavior, and product-level information.
- Sufficient size and complexity to perform descriptive statistics, visualizations, correlation analysis, and regression modeling.

2 Introduction

This report presents a complete end-to-end data analysis of the Brazilian Olist e-commerce dataset. The objective is to demonstrate the full analytics workflow, including data cleaning, descriptive statistics, visualization, correlation analysis, regression modeling, and interpretation.

3 Business Questions and Objectives

We address the following business questions:

1. What operational factors influence customer review scores?
2. How strongly does delivery time affect customer satisfaction?
3. Do customers from different states behave differently?
4. Which regions generate the highest revenue?

4 Methods Overview

The analysis was conducted using R and tidyverse. The workflow included:

1. **Data Cleaning:** Missing values were inspected, timestamps were standardized into proper datetime formats, and outliers were evaluated. Additional variables such as delivery time and total order value were engineered to support later analysis.
2. **Descriptive Statistics:** Summary statistics were produced for all major numerical and categorical variables to understand overall patterns, distributions, missingness, and group frequencies.
3. **Visualizations:** Key business questions were explored through histograms, bar charts, and trend plots built with `ggplot2`. These visualizations highlight patterns in spending behavior, logistics performance, geographic demand, and satisfaction.
4. **Correlation Analysis:** Pairwise correlations were calculated to quantify linear relationships between operational metrics (delivery time, installments, spending) and review score. This provided an initial sense of which factors might meaningfully predict customer satisfaction.
5. **Regression Modeling:** A multiple linear regression model was estimated to measure the combined influence of delivery time, installments, order value, and geographic location on review scores. The model allowed for formal statistical testing and interpretation of effect sizes and significance.

5 Descriptive Analysis

5.1 Missing Values

	variable	missing_count
1	order_id	0
2	customer_id	0
3	order_status	0
4	order_purchase_timestamp	0
5	order_approved_at	14
6	order_delivered_carrier_date	1
7	order_delivered_customer_date	0
8	order_estimated_delivery_date	0
9	customer_unique_id	0
10	customer_zip_code_prefix	0

5.2 Numeric Summary

total_price	total_freight	product_count	payment_value
Min. : 0.85	Min. : 0.00	Min. : 1.000	Min. : 9.59
1st Qu.: 45.90	1st Qu.: 13.84	1st Qu.: 1.000	1st Qu.: 61.80
Median : 86.00	Median : 17.16	Median : 1.000	Median : 105.13
Mean : 136.65	Mean : 22.76	Mean : 1.142	Mean : 159.44
3rd Qu.: 149.90	3rd Qu.: 23.99	3rd Qu.: 1.000	3rd Qu.: 176.09
Max. :13440.00	Max. :1794.96	Max. :21.000	Max. :13664.08

max_installments	review_score	delivery_time	order_value
Min. : 1.000	Min. :1.000	Min. : 0.53	Min. : 9.59
1st Qu.: 1.000	1st Qu.:4.000	1st Qu.: 6.76	1st Qu.: 61.79
Median : 2.000	Median :5.000	Median : 10.21	Median : 105.08
Mean : 2.929	Mean :4.156	Mean : 12.52	Mean : 159.41
3rd Qu.: 4.000	3rd Qu.:5.000	3rd Qu.: 15.69	3rd Qu.: 176.02
Max. :24.000	Max. :5.000	Max. :208.35	Max. :13664.08

delivery_delay	is_SP

```
 Min.   :-59.00         Min.   :0.0000
 1st Qu.: -5.00         1st Qu.:0.0000
 Median :  0.00         Median :0.0000
 Mean   :  4.80         Mean   :0.4201
 3rd Qu.:  9.00         3rd Qu.:1.0000
 Max.   :133.00         Max.   :1.0000
```

5.3 Categorical Summary

5.3.1 Review Score Distribution

```
  review_score     n percent
1            5 57060   59.2
2            4 18987   19.7
3            1  9409    9.76
4            3  7961    8.26
5            2  2941    3.05
```

5.3.2 Top 10 States

```
   customer_state     n
 1 SP             40478
 2 RJ             12284
 3 MG             11355
 4 RS              5363
 5 PR              4918
 6 SC              3534
 7 BA              3246
 8 DF              2089
 9 ES              1978
10 GO              1963
```

5.3.3 Order Status Distribution

```
1 delivered    96352
2 canceled         6
```

5.3.4 Top 10 Cities

	customer_city	n
1	sao paulo	15041
2	rio de janeiro	6571
3	belo horizonte	2702
4	brasilia	2080
5	curitiba	1485
6	campinas	1401
7	porto alegre	1345
8	salvador	1180
9	guarulhos	1137
10	sao bernardo do campo	916

5.3.5 Installment Usage Summary

max_installments	n	percent
1	46746	48.5
2	11994	12.4
3	10121	10.5
4	6867	7.13
5	5134	5.33
6	5071	5.26
7	4128	4.28
8	3797	3.94
9	1555	1.61
10	614	0.64

5.3.6 Delivery Time (Days)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.5334	6.7637	10.2120	12.5234	15.6875	208.3518

6 Data Cleaning and Preparation

The raw dataset required several preprocessing steps to ensure consistency, accuracy, and suitability for statistical analysis. The main cleaning and preparation steps included:

- **Timestamp standardization:** All date and time variables (purchase, approval, carrier delivery, customer delivery, and estimated delivery) were converted into proper datetime formats to enable accurate time-based calculations.
- **Feature engineering:** Several analytical variables were created, including:
  - *delivery_time*: number of days between purchase and customer delivery.
  - *order_value*: total spending per order, computed as price + freight.
  - *is_SP*: a binary indicator flagging whether a customer is located in the state of São Paulo.
- **Handling missing and inconsistent data:** Rows missing essential fields such as delivery timestamps, review scores, or payment values were removed. These represent a small portion of the dataset and their removal prevents distortions in downstream analysis.
- **Data integration:** The final analytic dataset was obtained by merging multiple Olist tables—orders, order items, payments, customers, and reviews—into a single, unified structure where each row represents one completed order.
- **Filtering extreme or invalid values:** Outliers (e.g., excessively long delivery times or extremely high order values) were trimmed only for visualization purposes, ensuring clarity of interpretation while preserving all data for statistical modeling.

This preprocessing pipeline ensured a clean, structured dataset suitable for descriptive analysis, visualization, and regression modeling.

7 Visualizations

7.1 Order Value Distribution

This figure provides an overview of how much customers typically spend on each purchase. Values above 300 were removed to prevent extreme outliers from distorting the distribution.

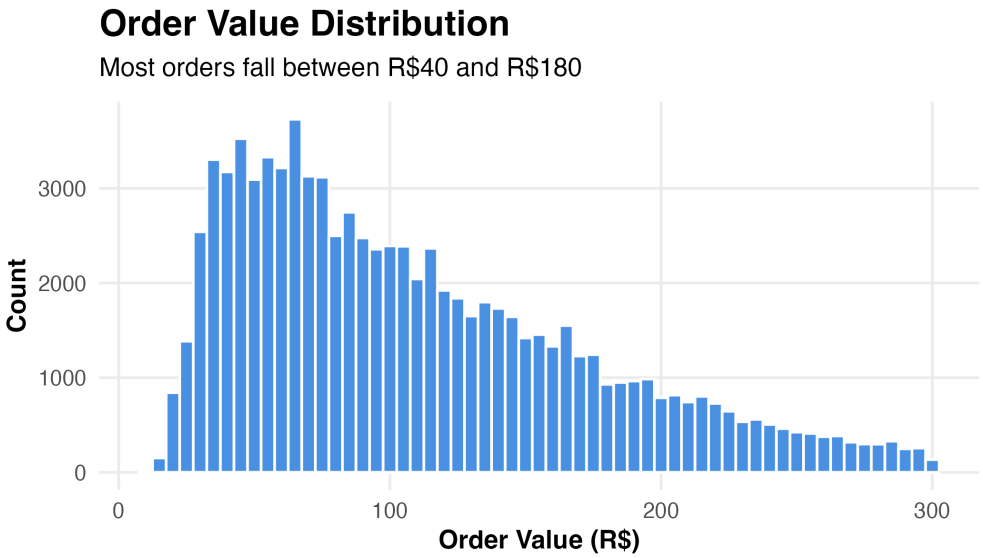

Figure 1: Distribution of order values (orders below R$300).

7.2 Delivery Time Distribution

This figure shows the distribution of delivery times for orders that arrived within 60 days. It helps evaluate the consistency and typical speed of Olist's logistics network.

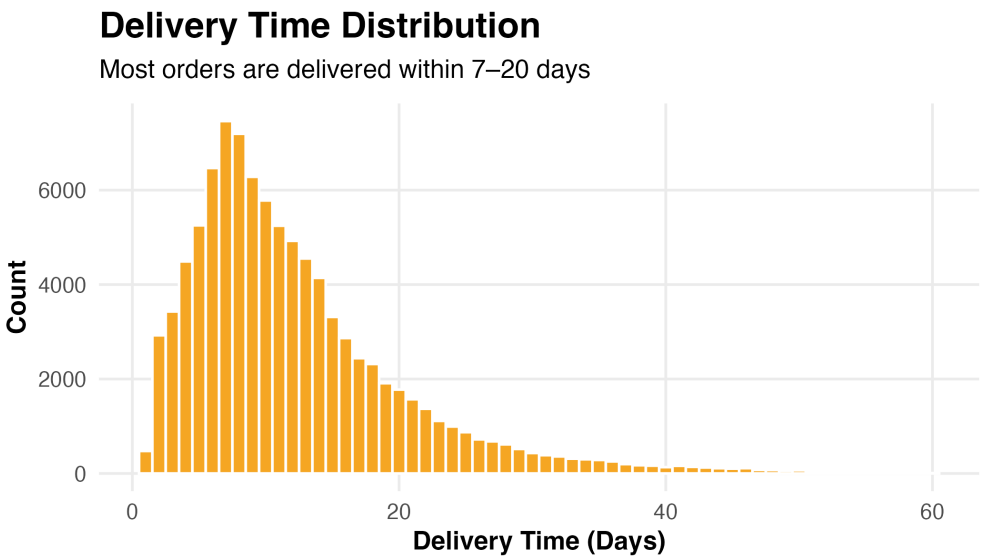

Figure 2: Distribution of delivery times (0–60 days).

7.3 Review Score vs Delivery Time

This figure illustrates how average customer ratings vary with delivery duration. It helps visualize whether slower delivery is associated with lower satisfaction.

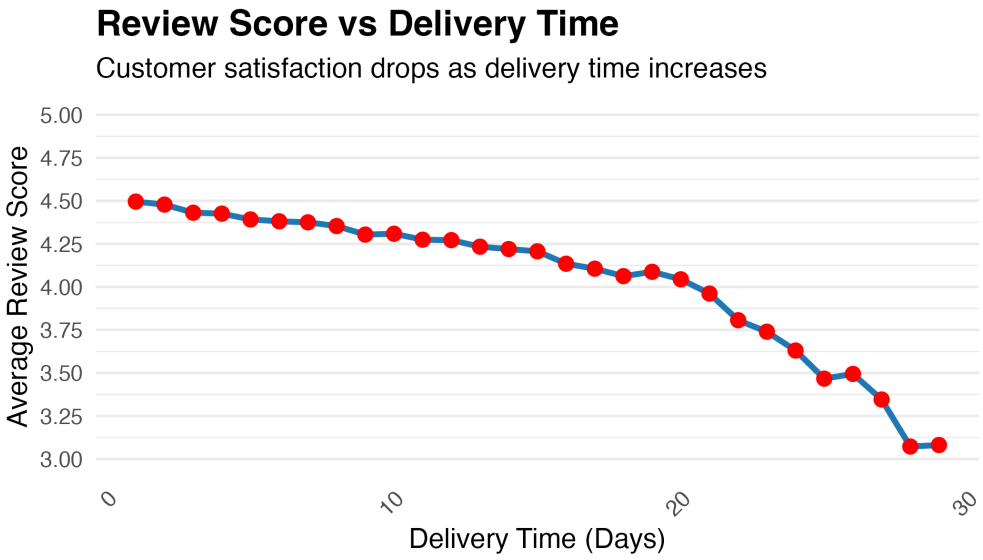

Figure 3: Average review score by delivery time.

7.4 Total Revenue by State

This figure compares total revenue generated across Brazilian states. It highlights regional differences in purchasing volume and identifies the platform's strongest markets.

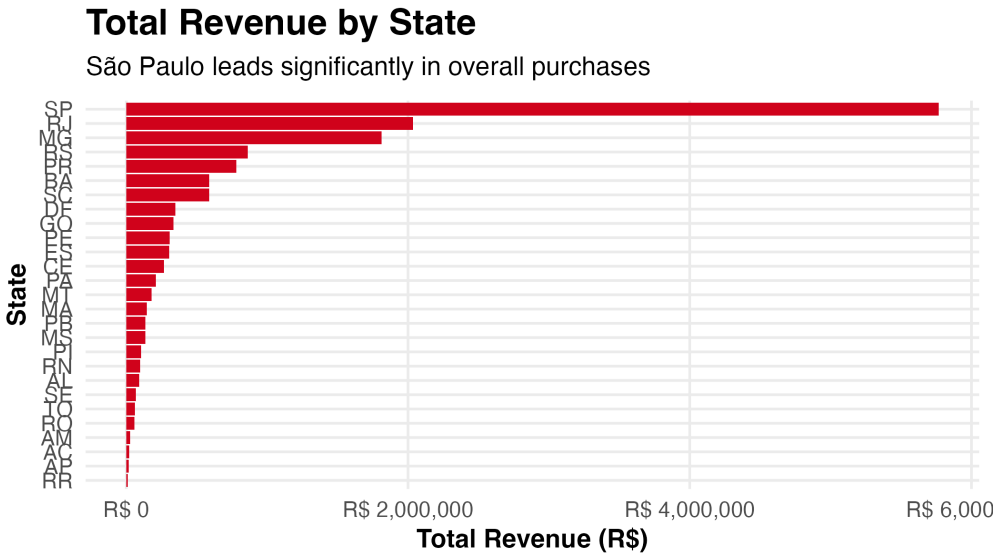

Figure 4: Revenue per Brazilian state.

8 Statistical Analysis

8.1 Correlation Analysis

```
1 delivery_time vs review_score     -0.334
2 order_value vs review_score       -0.0421
3 installments vs review_score      -0.0306
4 SP vs review_score                 0.0600
```

9 Regression Analysis

This section presents both the simple and the multiple regression models used to evaluate how operational features such as delivery time, installment usage, order value, and customer location influence customer review scores. The simple regressions isolate the effect of each variable on its own, while the multiple regression examines their combined impact. Across all

models, the effects are statistically significant but not large in magnitude, which suggests that these variables influence customer satisfaction but do not fully determine review outcomes.

9.1 Simple Regression Models

9.1.1 Model 1: Review Score ∼ Delivery Time

The first model shows a strong and highly significant negative relationship between delivery time and review score. Each additional day of delivery is associated with a lower rating. This pattern is meaningful and consistent, although delivery time alone explains only a portion of the total variation in satisfaction.

9.1.2 Model 2: Review Score ∼ Max Installments

The second model indicates a small but statistically significant negative effect. Orders paid using a higher number of installments tend to receive slightly lower ratings. The magnitude of the effect is modest, indicating that installment usage alone is not a dominant driver of review outcomes.

9.1.3 Model 3: Review Score ∼ Order Value

The third model reveals a very small yet statistically significant negative effect for order value. Higher spending is associated with marginally lower ratings, possibly reflecting higher expectations among customers who make larger purchases. The practical impact of this relationship remains limited.

9.1.4 Model 4: Review Score ∼ São Paulo Indicator

The fourth model examines whether customers from the state of São Paulo behave differently. The coefficient is negative and statistically significant, indicating that customers from São Paulo tend to give lower ratings on average. Although this effect is measurable, it remains modest.

9.2 Multiple Regression Model

To analyze the combined effect of all predictors, a multiple linear regression model was estimated. This model accounts for potential relationships among variables such as delivery time, order value, and customer location.

The fitted multiple regression equation is presented below:

$$\hat{y}_i = \beta_0 + \beta_1 \cdot \text{delivery\_time}_i + \beta_2 \cdot \text{max\_installments}_i + \beta_3 \cdot \text{order\_value}_i + \beta_4 \cdot \text{is\_SP}_i$$

where $\hat{y}_i$ represents the predicted review score for order i.

This equation shows that delivery time has the largest effect on customer satisfaction. Each additional day is associated with a decrease of approximately 0.048 points in the review score. The effects of installment usage, order value, and living in São Paulo are smaller in magnitude but remain statistically significant.

The full coefficient table from R is shown below:

```
Coefficients:
                   Estimate Std. Error  t value Pr(>|t|)
(Intercept)       4.856e+00  9.552e-03  508.395  < 2e-16 ***
delivery_time    -4.796e-02  4.384e-04 -109.389  < 2e-16 ***
max_installments -5.204e-03  1.517e-03   -3.430 0.000605 ***
order_value      -1.080e-04  1.897e-05   -5.696 1.23e-08 ***
is_SP            -1.610e-01  8.407e-03  -19.150  < 2e-16 ***
```

9.2.1 Interpretation

The results indicate that delivery time is the most influential operational factor affecting customer reviews. The negative and highly significant coefficient confirms that customers place considerable importance on fast delivery when evaluating their shopping experience.

Order value and the number of installments also show statistically significant relationships with review scores, although the effects are small. Customers who spend more or use more installments may have higher expectations, which can result in slightly lower ratings.

The indicator for São Paulo is also negative and significant, suggesting that customers from Brazil's largest e commerce region tend to be more critical. Yet the effect is modest

compared with the influence of delivery time.

The model has an R squared of approximately 0.11, meaning that about 11 percent of the variation in review scores is explained by these operational variables. This level of explanatory power is typical for customer review data, which is strongly affected by subjective elements such as product expectations, personal preferences, seller communication, and other factors not included in the dataset. While the model highlights meaningful relationships, most of the variation in review behavior arises from influences outside the available operational metrics.

10 Conclusion

This project carried out a complete end to end analytical workflow using the Olist e commerce dataset. The process included data cleaning, descriptive statistics, visual exploration, and regression modeling, which together provided a clear understanding of customer behavior and operational performance.

The results show that delivery time is the most influential operational factor affecting customer satisfaction. Longer delivery times are consistently linked to lower review scores. Other factors such as order value, installment usage, and customer location also matter, although their effects are statistically significant but small.

The low R squared values indicate that many subjective and unobserved elements, including product expectations, seller behavior, and personal preferences, play an important role in shaping customer reviews. The geographic analysis further shows strong regional differences, with São Paulo responsible for a large share of orders and revenue.

Overall, the findings suggest that logistical improvements can help increase satisfaction, but a deeper understanding of review behavior requires richer data and more advanced modeling. I also want to note that I truly enjoyed this project. The dataset contains a large amount of information, which made it engaging to explore. This was a very useful and enjoyable experience, and I hope to apply the skills I learned here to continue improving my analytical abilities.