Brazilian E-Commerce Public OLIST

From Raw Data to Pricing Strategy
Comprehensive Exploratory Data Analysis (EDA) and Machine Learning Model

Dr. Mohamed Badawy

Eng. Hosny

Meet the Team

Mohamed Younis

Jovanie Hanie

Abdelrahman Omar

Project Goals & Objectives

The Problem

The main objective is to understand the entire customer journey analysis within the Olist platform and the subtle factors that affect the quality of the trial purchase, and then use the analysis and machine learning to better improve forecasting.

Key Objectives

- Analyze Olist data to understand the entire journey.
- Clean and prepare data for analysis.
- Study customer reviews and understand the reasons for their decline.
- Analyze seller and product performance.
- Extract insights to improve the customer experience.
- Build a machine learning model

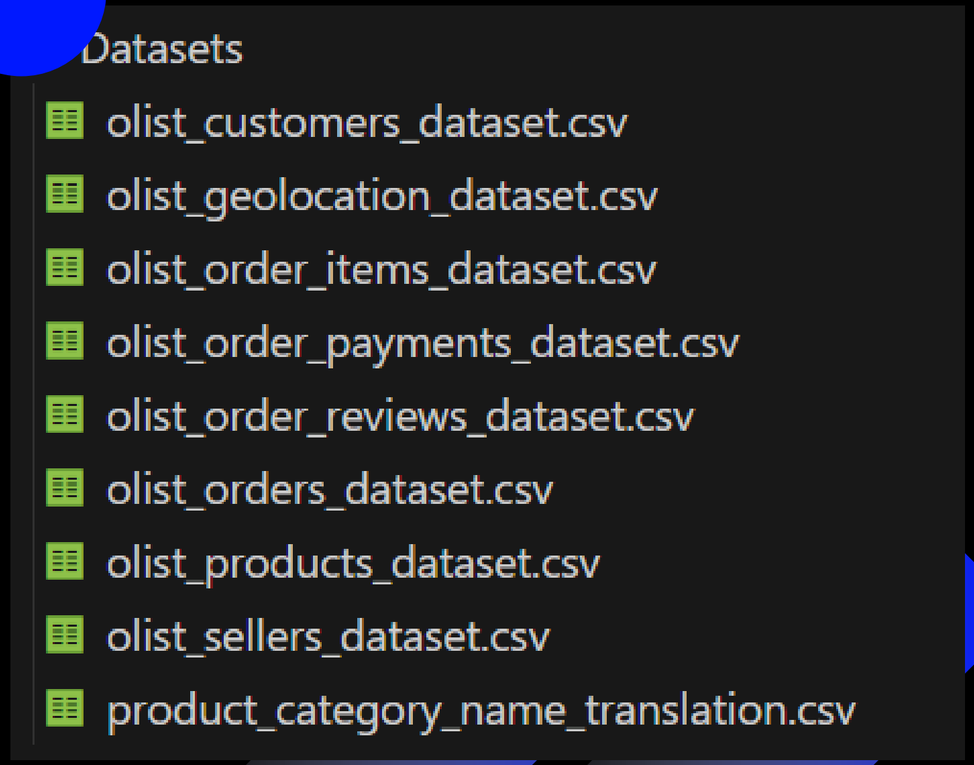
Data Overview & Initial Challenges

Source:
The project contains nine main tables, each representing a different part of the order journey (Customers, Orders, Order Items, Payments, Reviews, Products, Sellers, Locations, etc.).

Initial Challenges:

- Columns not formatted as datetime made it difficult to calculate delivery times and features.
- Missing values and duplicates appeared in some tables.
- Some columns exhibited inconsistencies (categories, text values).
- The importance of processing this data became clear because it is fundamental before delving into deep EDA and machine learning.

Datasets Link



Data Cleaning:

- The focus was on correcting the data type, especially the datetime columns, to accurately calculate shipping time and delays.
- Some duplicates in tables, such as geolocation, were removed.
- A small number of missing values were addressed, depending on the nature of each table.
- Some inconsistent values in certain columns, such as category names, were standardized.
- The goal was to adequately prepare the data for EDA and Machine Learning without complex cleaning operations.

Part of Data Cleaning
Order Payment DATASET

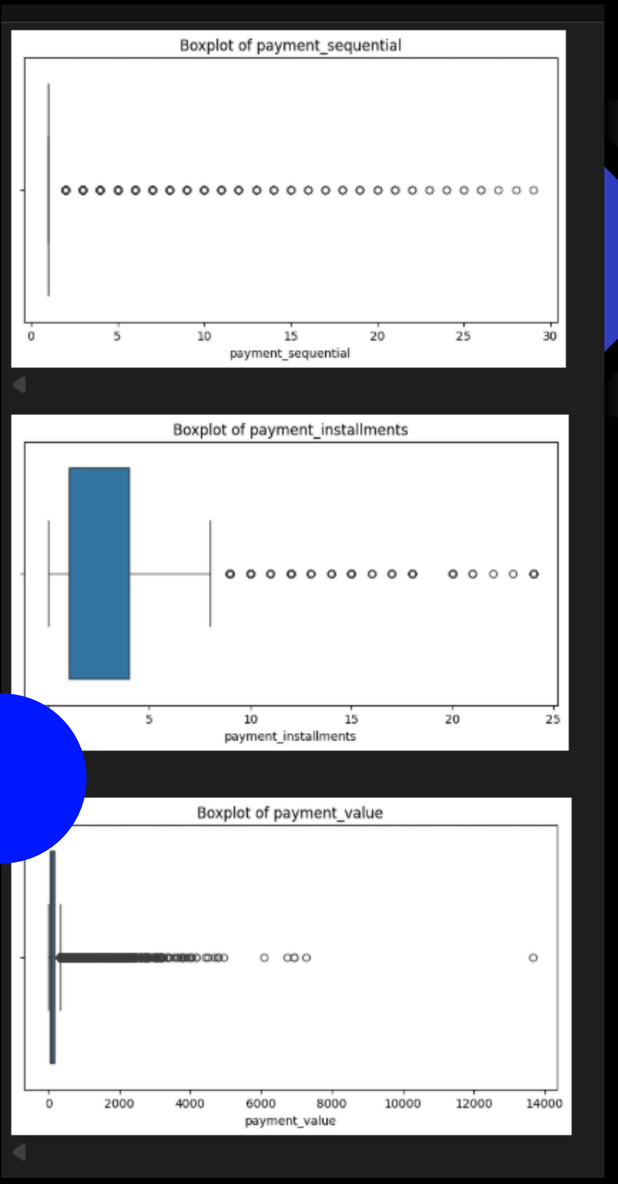
In the Order Payment Dataset, we noticed significant variations in payment values between orders.

We used the Interquartile Range (IQR) to identify payment outliers.

The IQR helped us pinpoint values that deviated from the normal payment range because they could:

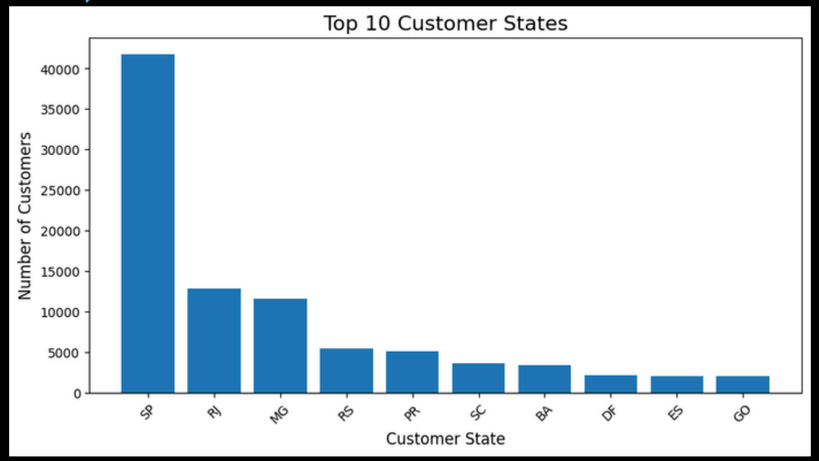
Influence statistical analysis

Inflate the average in a misleading way



Insights Before Merge

Key Insights & EDA



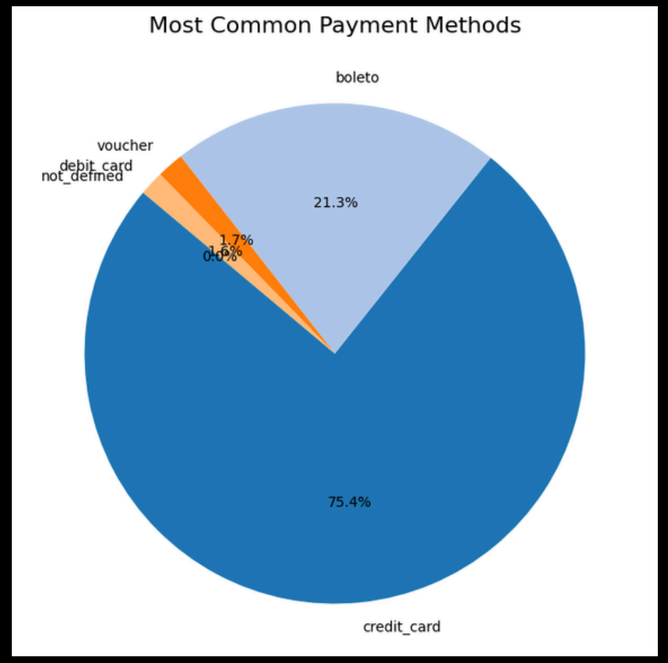
Customer Geographical Distribution Analysis
São Paulo Dominance: The state of São Paulo (SP) holds the top spot by a significant margin in the number of customers, making it the primary center of demand.

The majority of purchasing power is concentrated in the Southeast region, while density is lower in the northern regions.

This distribution means that most shipping operations take place within a specific geographic area, which presents an opportunity to reduce costs if properly leveraged.

Insights Before Merge

Key Insights & EDA



Credit card usage (credit card dominance):
Over 75% of transactions are made via credit cards.
This indicates a significant customer base using this payment method.

Installment payments as an attraction (installment as a key driver):
A large percentage of transactions are divided into 2-4 installments.
Options are constantly available to increase sales volume.

Other: Pay slip (Boleto Bancário):
The second most common payment method, and is considered essential for people who do not have a credit card.

Insights Before Merge
Key Insights & EDA

Product Portfolio Analysis: Applying the 80/20 Rule

Top Revenue Categories:
Domination: Categories such as home goods, health and beauty, and sports are the primary revenue drivers.

These categories should be prioritized in inventory management.

The Pareto Principle (80/20):
Analysis shows that 20% of product categories generate approximately 80% of total revenue.

This specific percentage should be monitored, and sufficient stock should be maintained to avoid out-of-stock issues.

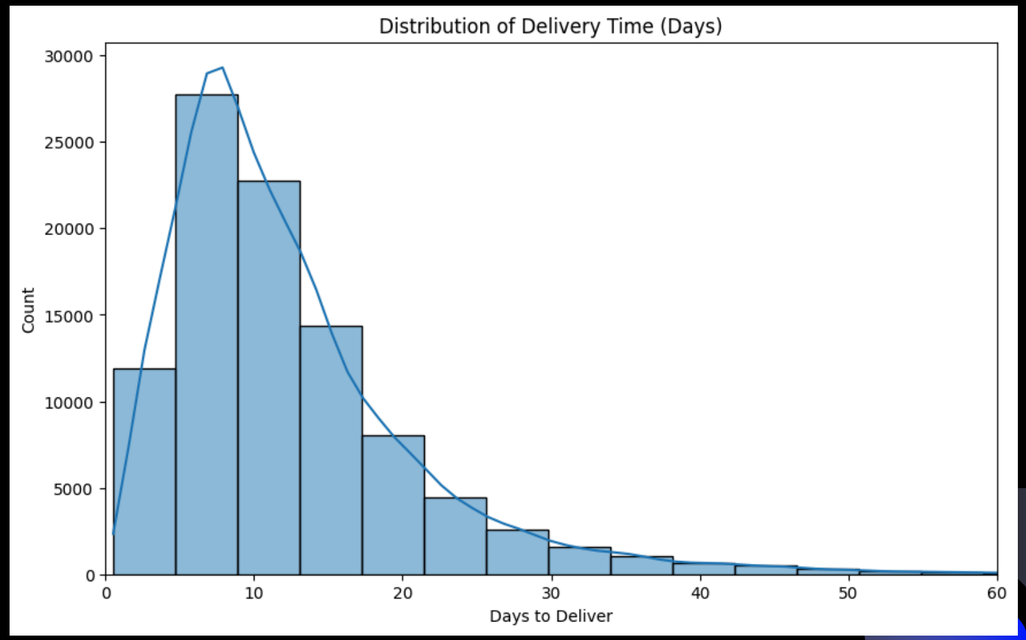
Product Diversity: A large number of long-tail categories with low revenues require continuous evaluation of their economic viability.

Insights Before Merge
Key Insights & EDA

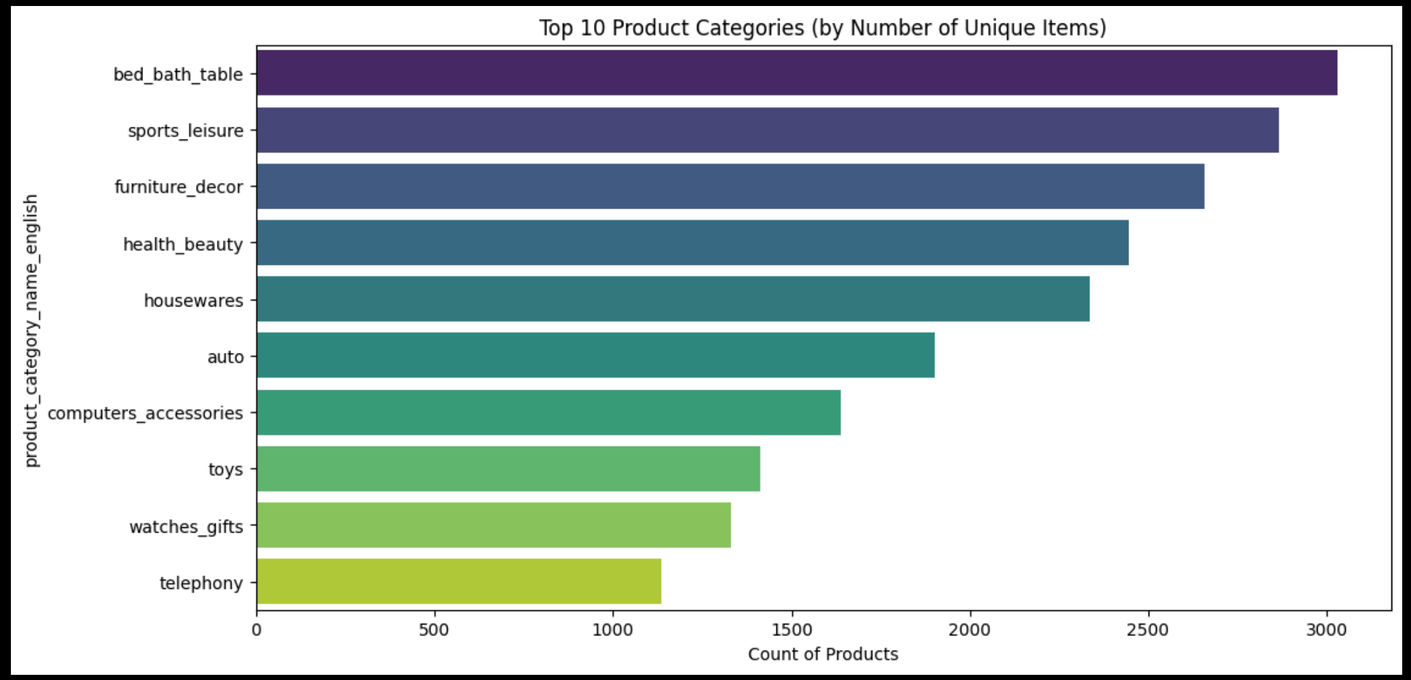
Delivery Time Distribution: Service Speed Analysis
Average Delivery Time (Service Average):
The average time required to deliver an order to the customer is 12.65 days.

Distribution Skewness:
The graph shows a rightward skewed distribution, meaning that the majority of orders arrive faster than average.
However, a small number of orders (late orders) take exceptionally long (over 30-40 days).

Outliers:
Late orders (outliers) are responsible for raising the overall average (12.65 days). The reasons for these individual delays must be identified.



Insights Before Merge
Key Insights & EDA



Key product categories (inventory distribution)
Data shows that Olist's main strength lies in home and leisure products. The "bed_bath_table" and "sports_leisure" categories are the most diverse, each exceeding 3,000 unique items, demonstrating the depth of its inventory of essential everyday goods. Categories like "furniture_decor" and "health_beauty" complement this strong lifestyle base.

In contrast, the technology and related subcategories represent a real growth opportunity, with "telephony" and watches_gifts" lagging behind with fewer" than 1,500 items each. The platform should focus its efforts on attracting new sellers to increase diversity in these sectors, fill inventory gaps, and expand its customer base in the .electronics and gifts market

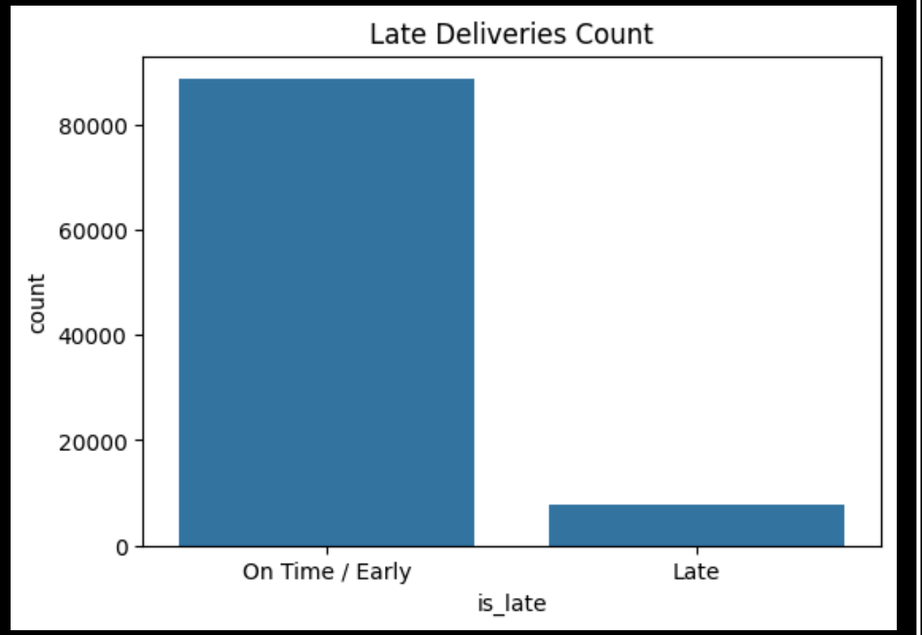
Feature Engineering

Creating Metrics for Delivery Performance

1. Actual Delivery Time Measurement:
Equation: Delivery Days = Customer Receive Date - Purchase Date
Value: This variable is the target column for the machine learning model.

2. Lateness Quantification:
Equation: Diff_Estimated = Estimated Delivery Date - Actual Delivery Date
Value: Helps evaluate the logistics team's performance.
Positive value (+ve): The order arrived later than expected.
Negative value (-ve): The order arrived earlier than expected.

3. Binary Lateness Flag:
Variable: is_late (Yes/No)
Value: Converts the lateness into a simple classification to understand the prevalence of the lateness problem.



	order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date	orde
0	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15	2017-10-04 19:55:00	2017-10-10 21:25:13	
1	53cdb2fc8bc7dce0b6741e2150273451	b0830fb4747a6c6d20dea0b8c802d7ef	delivered	2018-07-24 20:41:37	2018-07-26 03:24:27	2018-07-26 14:31:00	2018-08-07 15:27:45	
2	47770eb9100c2d0c44946d9cf07ec65d	41ce2a54c0b03bf3443c3d931a367089	delivered	2018-08-08 08:38:49	2018-08-08 08:55:23	2018-08-08 13:50:00	2018-08-17 18:06:29	
3	949d5b44dbf5de918fe9c16f97b45f8a	f88197465ea7920adcdbec7375364d82	delivered	2017-11-18 19:28:06	2017-11-18 19:45:59	2017-11-22 13:39:59	2017-12-02 00:28:42	
4	ad21c59c0840e6cb83a9ceb5573f8159	8ab97904e6daea8866dbdbc4fb7aad2c	delivered	2018-02-13 21:18:39	2018-02-13 22:20:29	2018-02-14 19:46:34	2018-02-16 18:17:02	

Data After Merge

Shape of Final Dataframe: (110823, 49)

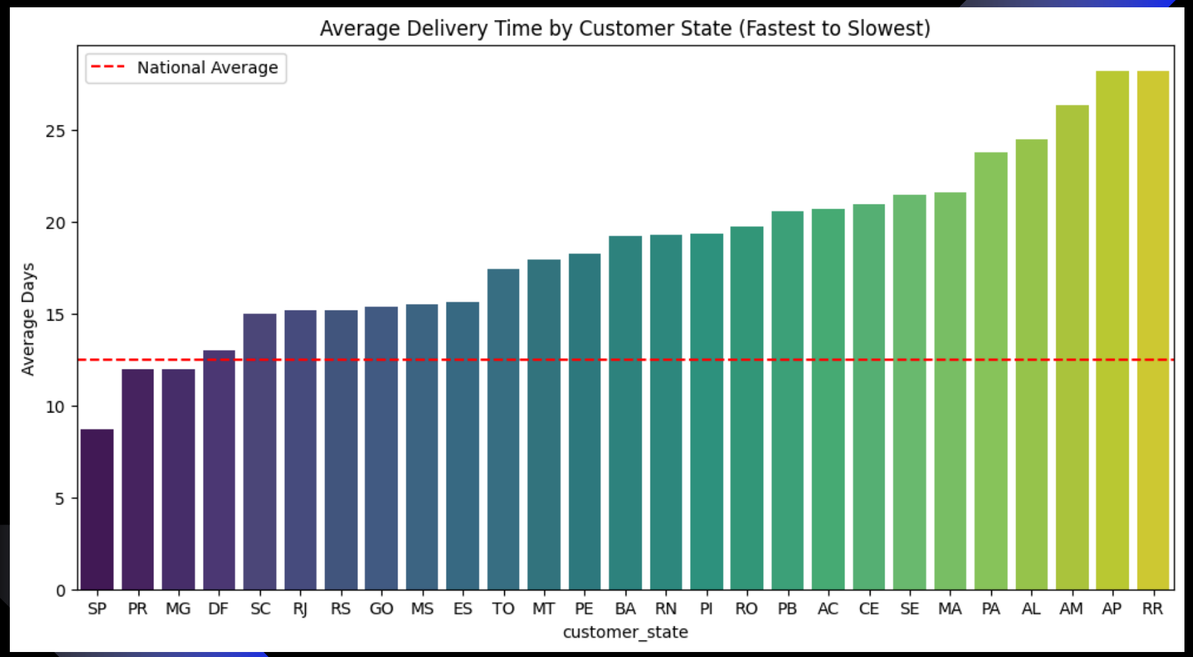
Important Columns: Customer State - Seller State (Location)- Product Weight (g)
Product Dimensions (cm^3) (Product Specifications) - Price
Shipping Cost (Cost) - Product Category Name (English) (Type)
Month of Purchase - Day of Purchase (Time) - Expected Delivery Date

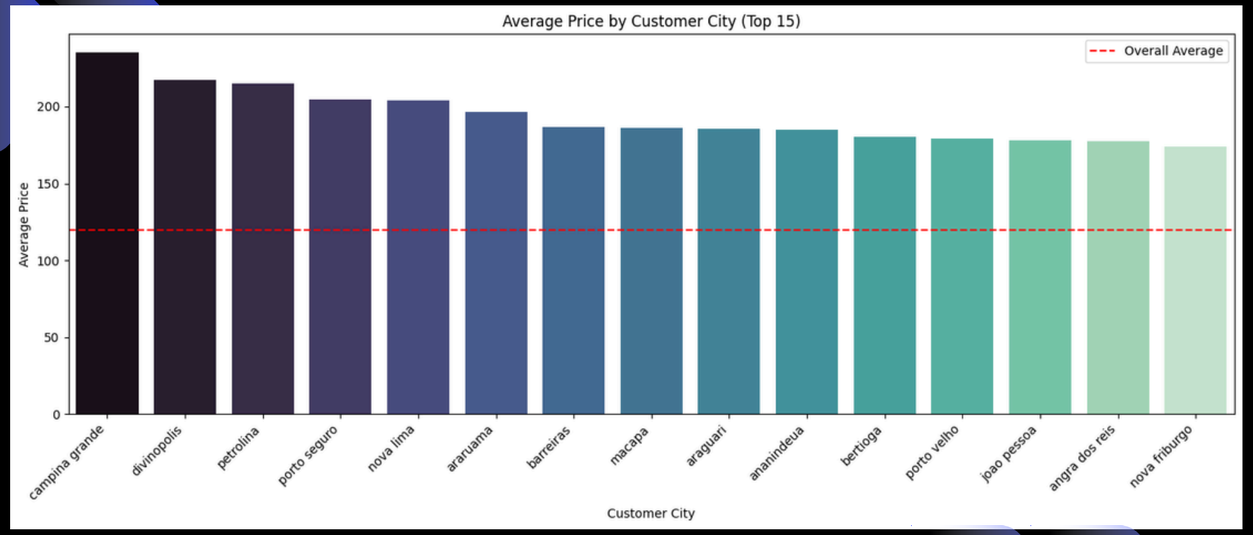
Key Insights & EDA

Average Delivery Time Analysis by Customer State
Fastest Performance (Best Performers):
States in SP, PR, MG, DF, and SC all perform faster than the national average.This indicates high efficiency in the supply and distribution chains within these regions.

Slowest Performance (Areas Needing Improvement):
States on the far right, such as AL, AM, AP, and RR, have very high average delivery times, ranging from 25 to over 28 days, indicating significant logistical challenges, possibly due to distance or infrastructure.

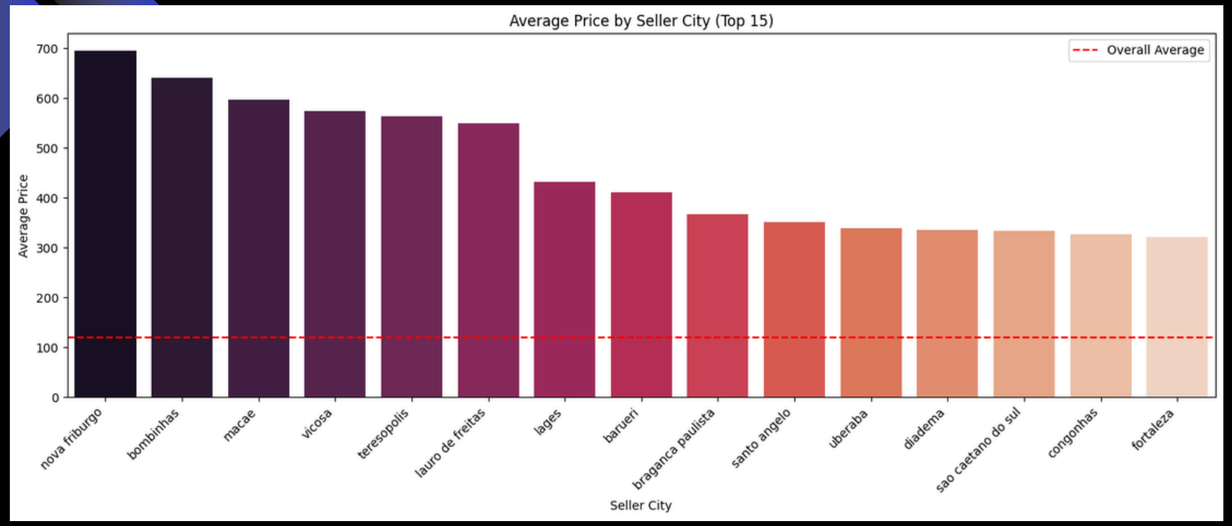
A large group of states in the middle, such as MT, PE, BA, RN, and PI, have average delivery times of 17 to 20 days, which are significantly slower than the national average and require review.





Key Insights & EDA

The problem of uniform pricing



1. Price Variation for the Same Product Across Different Cities: The immediate problem: The first graph shows that the price of the same product varies significantly between customer cities.
Impact: This price variation for the same product indicates pricing inefficiency and can drive customers to seek better prices outside the platform, damaging trust and loyalty.

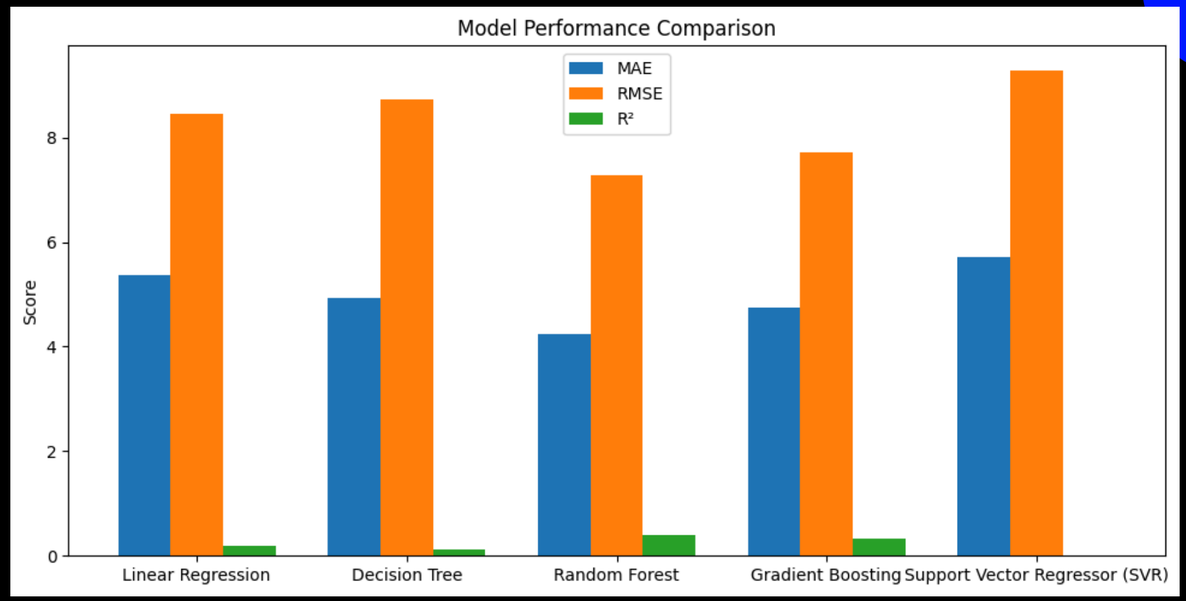
2. Average Price by Seller City: Huge Variation: There is a huge gap between the highest and lowest average prices for sellers.
Possible Explanation: This may indicate that sellers in cities with high prices sell niche/luxury products, or that they add very high operating or shipping costs to the product price.

Machine Learning

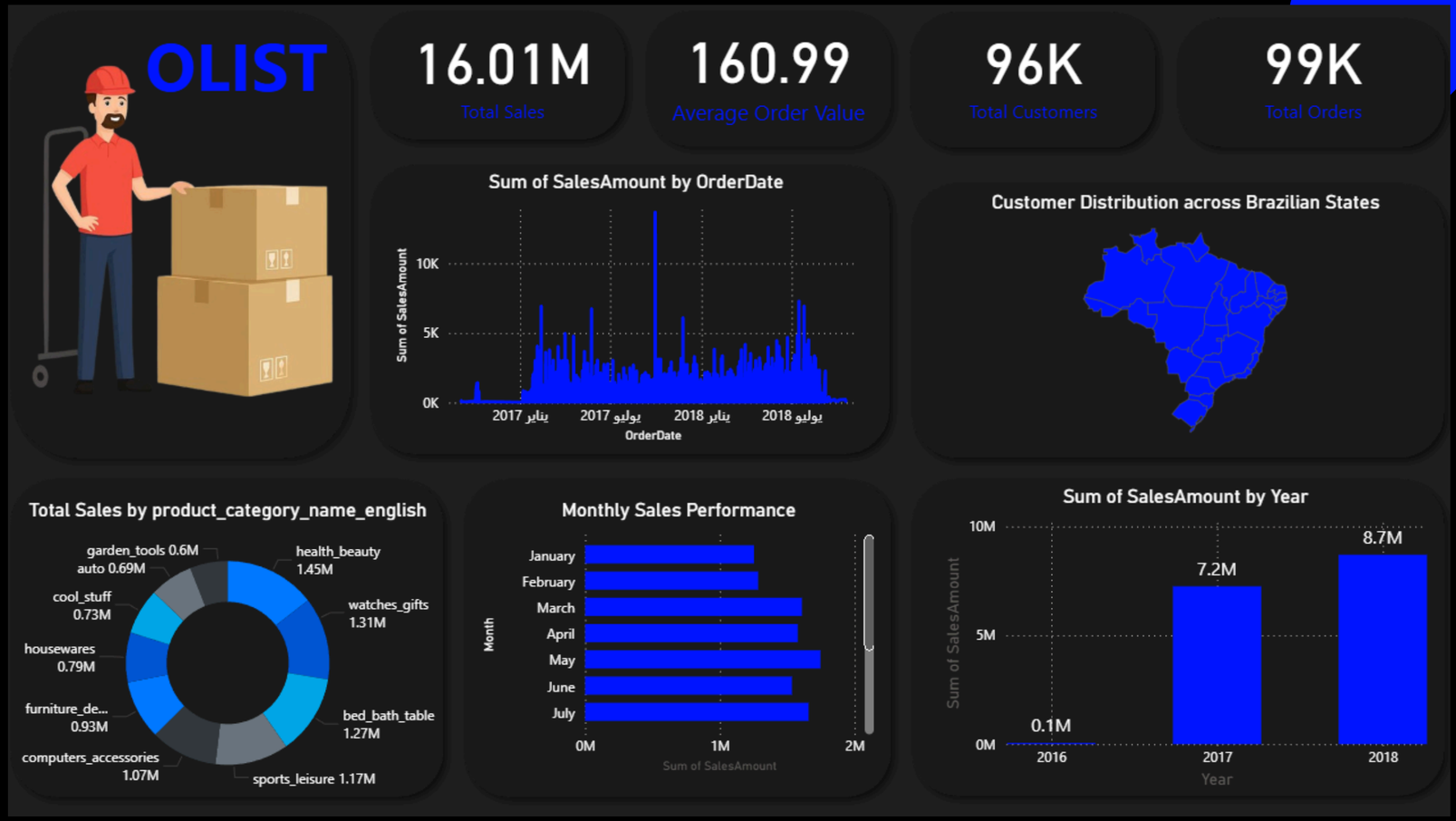
Models

1. **Objective**:
   - Prediction: To use various regression models to accurately forecast key variables.

2. Key Models Employed:
   - Five distinct regression models were trained and compared for performance:
     - Linear Regression (Baseline)
     - Decision Tree
     - Random Forest (Ensemble Method)
     - Gradient Boosting (Ensemble Method)
     - SVR (Support Vector Regressor)



Interactive Dashboard Overview



Thank You