OLIST E-COMMERCE ANALYTICS PROJECT

End-to-End Business Analytics Report

Data Engineering | SQL Analytics | Python EDA | Power BI Dashboards

Primary Business Question:

"Why is customer retention critically low, and what does it imply for long-term business growth?"

Dataset: Olist Brazilian E-Commerce Public Dataset
Stack: Python | PostgreSQL | Power BI | DAX

1. Project Overview

This document presents the complete analytical journey of the Olist E-Commerce Analytics Project — a full-stack, industry-grade data analytics initiative built to simulate a real business intelligence environment. The project spans every stage of modern analytics: raw data ingestion, warehouse design, SQL-based KPI engineering, Python exploratory data analysis, and executive-level Power BI dashboards.

1.1 Business Context

Olist is a Brazilian e-commerce marketplace that connects small sellers to major retail platforms. The dataset represents actual transactional data from 2016 to 2018 and provides a rich base for understanding customer behavior, logistics performance, revenue distribution, and marketplace health.

1.2 Primary Business Question

Central Analytical Problem

"Why is customer retention on the Olist marketplace critically low — sitting at approximately 3% — and what does this structural weakness imply for long-term business sustainability and revenue growth?"

1.3 Project Objectives

This project was designed to go beyond surface-level dashboards and deliver deep business diagnostics. Specific objectives included:

- Build a reproducible, industry-standard ETL pipeline to ingest raw data into a structured data warehouse.
- Design a clean Star Schema in PostgreSQL to serve as a reliable analytics foundation.
- Engineer meaningful business KPIs at the SQL layer before any visualization.
- Conduct systematic hypothesis testing to identify the true drivers of low retention.
- Perform advanced Python EDA to uncover revenue distribution, skewness, and growth patterns.
- Design an executive Power BI dashboard with three purpose-built analytical pages.
- Translate findings into actionable strategic recommendations for the business.

1.4 Technologies & Architecture

Layer	Tools Used
Data Ingestion	Python, Pandas, SQLAlchemy
Data Warehouse	PostgreSQL, pgAdmin
Analytics Layer	SQL (Advanced), Window Functions, Aggregations
EDA & Python Analytics	Python, Pandas, Matplotlib, NumPy
Visualization	Power BI, DAX (Advanced Measures)
Version Control	Git, structured folder architecture (raw → analytics → reporting)

2. Dataset Overview

The Olist Brazilian E-Commerce Public Dataset is a multi-table relational dataset covering orders placed between 2016 and 2018. It includes information across eight distinct data files, each capturing a different dimension of the marketplace.

2.1 Source Data Files

Eight CSV files were loaded into the raw schema of the PostgreSQL warehouse, each representing a separate domain entity:

Table	Content Description
customers	Customer profiles with IDs and location data
orders	Order header data including timestamps and status
order_items	Line-level detail: product, seller, price, freight per item
order_payments	Payment methods and amounts per order
order_reviews	Customer review scores and comments post-delivery
products	Product attributes including category and dimensions
sellers	Seller profiles with location information
product_category_name_translation	Portuguese-to-English category name mapping

2.2 Critical Dataset Discovery: Customer Identifier

A significant data modeling issue was identified early in the project. The raw data contains two customer ID fields — customer_id and customer_unique_id — and understanding the distinction between them was essential to accurate analysis.

Key Discovery

customer_id changes with every new order placed by the same person — it is a session-level identifier, not a person-level identifier.

customer_unique_id is the true business key that identifies a unique individual customer across all their orders.

Using customer_id incorrectly showed ~0% repeat customers. Correcting to customer_unique_id revealed the true repeat rate of approximately 3.05% — a critical modeling correction with significant business implications.

3. Data Engineering Phase

The data engineering phase established the foundational infrastructure of the entire analytics system. The goal was to build a reproducible, industry-standard pipeline that would convert raw CSV files into a clean, structured warehouse ready for analysis.

3.1 ETL Pipeline — Raw Layer Ingestion

A reusable ETL script (load_raw.py) was developed using Python, Pandas, and SQLAlchemy to automate the loading of all eight source CSV files into a dedicated raw schema in PostgreSQL. The script was designed to be modular and reproducible, allowing any data refresh to be completed without manual intervention.

The ETL pipeline was built around four core principles:

- Automated CSV detection and loading — no hardcoded file references, making the pipeline extensible.
- Reproducibility — the script can be re-run cleanly at any point without duplicating data.
- Industry ETL standards — structured separation between raw ingestion and analytical transformation layers.
- Elimination of manual imports — all data flows programmatically from source files to the database.

3.2 Data Cleaning in SQL

After raw ingestion, a series of SQL-based cleaning operations were applied to ensure data quality and analytical integrity. The following issues were identified and resolved:

Issue Identified	Resolution Applied
Null freight values	Handled with appropriate null treatment logic in SQL
Date format inconsistency	Normalized all date fields to consistent timestamp format
Duplicate records	Validated and removed via deduplication logic
Data type mismatches	Corrected column types to match expected analytical use
Portuguese category names	Joined translation table to standardize all categories to English
Integrity consistency	Cross-validated referential integrity across linked tables

3.3 Folder Architecture

The project maintained a structured folder hierarchy to enforce clean separation between data stages. This mirrors professional data engineering practice and ensures clarity throughout the pipeline:

- Raw layer — original CSV files and initial database ingestion
- Analytics layer — cleaned, transformed, and modeled tables (Star Schema)
- Reporting layer — final views and aggregations ready for visualization

4. Data Warehouse Design — Star Schema

Rather than connecting Power BI directly to raw transactional tables, a clean Star Schema was designed within the PostgreSQL analytics schema. This architectural decision reduced redundancy, improved query performance, and enabled clean dimensional analysis.

4.1 Fact Table — analytics.fact_orders

The fact table was designed at the order-item grain, meaning each row represents one product item within an order. This grain provides the highest level of analytical flexibility, supporting revenue analysis at both the order and item levels.

The fact table contains the following fields:

Field	Type	Description
order_id	Key	Unique identifier for each order
customer_unique_id	FK → dim_customers	Corrected customer identifier (not customer_id)
product_id	FK → dim_products	Product purchased in this line item
seller_id	FK → dim_sellers	Seller fulfilling this order item
order_date	FK → dim_date	Date of the order for time-based analysis
price	Measure	Product price for this line item
freight_value	Measure	Shipping cost associated with this item
total_payment	Measure	Standardized revenue metric used across all KPIs
avg_review_score	Measure	Customer satisfaction score linked to the order
delivery_days	Derived Metric	Engineered field: days between order and delivery

4.2 Dimension Tables

Four dimension tables were built to support clean slicing and filtering of the fact data:

- dim_customers — Customer profile data using customer_unique_id as the primary key, along with geographic attributes (city, state).
- dim_products — Product catalog with English-translated category names (resolved from the Portuguese source data).
- dim_sellers — Seller profiles including location, enabling seller-level performance analysis.

- dim_date — A dedicated date dimension table configured as the official date table in Power BI, with fields for year, month, quarter, and a corrected YearMonth sort to ensure proper chronological ordering in visuals.

4.3 Relationships

One-to-Many relationships were defined from each dimension table to the fact table, using single cross-filter direction to maintain performance and predictable filter propagation. Bidirectional filtering was deliberately avoided except where analytically required.

5. SQL Analytics & KPI Engineering

Before moving to any visualization tool, all core business logic and KPIs were validated at the SQL layer. This approach ensures that the numbers driving decisions are grounded in accurate, tested queries — not calculated in an uncontrolled visual environment.

5.1 Core Business KPIs Computed

Metric	Value	Interpretation
Total Revenue	₹20.3 Million	Measured using total_payment as the standardized revenue metric
Total Orders	~99,000	Count of distinct orders processed on the platform
Total Customers	~95,420	Unique individuals identified by customer_unique_id
Average Order Value (AOV)	₹205.83	Stable across time — no upward pricing trend observed
One-Time Customers	92,507	96.95% of total customer base — single-purchase buyers
Repeat Customers	2,913	3.05% of total customer base — bought more than once
Repeat Rate	3.05%	Critically low for a marketplace; industry benchmark is typically 20–30%
Churn Rate	96.95%	Effectively 97 in every 100 customers never return
Repeat Revenue Contribution	~5.62%	Despite 2x higher spend per person, small volume limits impact
Avg Spend — One-Time	₹138.67	Baseline spend for single-purchase customers
Avg Spend — Repeat	₹262.03	Repeat customers spend approximately 2x more per transaction

5.2 Advanced SQL Transformations

Beyond basic aggregations, the following SQL techniques were applied to build analytically rich outputs:

- Delivery days calculation — computed as the difference between order placement and actual delivery timestamps, creating the delivery_days derived field.
- Customer order ranking — window functions (ROW_NUMBER) were used to rank each order per customer, enabling first-order vs subsequent-order analysis.
- Frequency grouping — customers were grouped by total order count (1 order, 2 orders, 3+ orders) to analyze behavioral patterns across frequency segments.
- Repeat behavior identification — flagged customers with more than one distinct order for segmentation and retention analysis.
- Delivery bucket classification — orders were categorized into fast, standard, slow, and very slow delivery buckets based on delivery_days thresholds.
- Revenue aggregation with percent contribution — window functions (SUM() OVER()) were used to calculate each category and state as a percentage of total revenue.
- Month-over-Month growth — LAG() window function used to compute percentage change in revenue and order volume across consecutive months.
- Performance optimization — heavy aggregation logic was deliberately moved to SQL to prevent Power BI calculated column overload, avoiding resource exhaustion errors.

6. Retention Hypothesis Testing

With the 3% repeat rate established as the central business problem, a systematic hypothesis elimination process was conducted to identify what was — and critically, what was not — driving low retention. Each potential cause was tested empirically using data, not assumption.

6.1 Hypothesis 1 — Is Delivery Speed Causing Churn?

Segment	Avg Delivery Days	Conclusion
One-Time Customers	~12.5 days	No meaningful difference
Repeat Customers	~12.5 days	No meaningful difference

Verdict

Delivery speed is NOT a driver of churn. Both customer groups experienced identical average delivery times, eliminating logistics as the root cause of one-time purchasing behavior.

6.2 Hypothesis 2 — Is First-Order Experience Causing Churn?

Segment	Avg First-Order Review Score	Conclusion
One-Time Customers	4.10 / 5	Virtually identical satisfaction
Repeat Customers	4.15 / 5	Virtually identical satisfaction

Verdict

First-purchase satisfaction is NOT a driver of churn. Both groups rated their initial experience similarly, ruling out poor product or service quality as the explanation for non-return behavior.

6.3 Hypothesis 3 — Is Category Type Causing Churn?

Repeat rates were calculated per product category to assess whether certain purchase categories inherently discourage return visits.

Category	Repeat Rate	Observation

furniture_decor	~7.2%	Above average, but still low in absolute terms
bed_bath_table	~6.5%	Slightly higher stickiness
sports_leisure	~5.3%	Above platform average
Major categories (avg)	1.7% – 3%	Around or below overall platform rate

Verdict

Category is NOT a primary driver of churn. Even product types that logically encourage repeat purchases (beauty, fashion, home goods) show low retention, confirming that the problem is not product-category-specific but is structural to the platform itself.

6.4 Hypothesis 4 — Is Geography Causing Churn?

Repeat rates were compared across Brazilian states to determine if region-specific factors (logistics, culture, market maturity) explained the retention problem.

Finding & Verdict

Repeat rate is uniformly ~3% across all states — including high-revenue states like São Paulo and lower-penetration states.

Geography is NOT driving churn. The retention problem is systemic and platform-wide, not concentrated in any specific region.

6.5 Root Cause Conclusion

After systematically eliminating delivery speed, first-order experience, product category, and geography as explanations, the evidence points to a structural marketplace problem rather than any operational or product-level deficiency.

Structural Root Causes Identified

1. Transaction-based platform design — the marketplace functions as a search-and-buy engine with no mechanisms to encourage return visits.
2. No loyalty or retention infrastructure — no points system, membership tier, subscription model, or post-purchase engagement.
3. Price-comparison behavior — customers likely use Olist for a specific purchase, compare prices, buy once, and do not develop platform affinity.
4. First-order discount dependency — promotional pricing may attract first-time buyers who are not genuinely loyal to the platform.

7. Advanced SQL & Behavioral Analytics

Beyond hypothesis testing, a deeper set of behavioral and cohort analyses were conducted to fully characterize how customers interact with the platform over time.

7.1 Cohort Retention Analysis

A cohort retention matrix was constructed using DATE_TRUNC to group customers by their first purchase month and then tracking whether they returned in subsequent months. This revealed the exact shape of the churn curve.

Finding

Month 1 cohort retention (the share of first-month buyers who return in the immediately following month) was less than 1%.

This confirms that churn is not gradual — it is immediate and sharp. Customers are not slowly drifting away over several months; they are not returning after their very first purchase.

7.2 Purchase Gap Analysis

For the small segment of repeat customers that does exist, the time between their first and second order was analyzed to determine whether any habitual purchasing cycle is emerging.

Metric	Value	Interpretation
Median Days to Second Order	28 days	The typical repeat customer returns within about a month
Average Days to Second Order	80 days	The mean is inflated by a long tail of very late returnees
Distribution Range	0 to 600+ days	Extremely wide spread — no consistent purchasing rhythm visible

The wide gap between median (28 days) and average (80 days) indicates a heavy-tailed distribution — most repeat buyers return relatively quickly, but a significant minority takes months or never returns within a meaningful window. The overall picture does not support the existence of a habitual purchase cycle on the platform.

7.3 RFM Segmentation

A full RFM (Recency, Frequency, Monetary) segmentation was applied to classify the entire customer base based on their behavioral characteristics. Each customer was scored across three dimensions:

- Recency — how recently did they place their last order (measured from the dataset maximum date)?
- Frequency — how many distinct orders have they placed in total?
- Monetary — what is their total cumulative spend?

This produced the following customer segments:

RFM Segment	Business Meaning
Champions	High frequency, high spend, purchased recently — most valuable customers
Loyal Customers	Consistent purchasing behavior with solid monetary value
Potential Loyalists	Recent buyers with growth potential if engaged correctly
Big Spenders	High monetary value but lower frequency — premium one-time buyers
At Risk	~22,000 customers — previously engaged but showing signs of disengagement
Lost Customers	No recent activity, low frequency — effectively churned

Critical Insight

The "At Risk" segment of approximately 22,000 customers represents the most immediately actionable opportunity — these customers have shown prior willingness to purchase and could be re-engaged through targeted remarketing before they move to the "Lost" category.

8. Python EDA — Revenue Distribution & Growth Analysis

Complementing the SQL-layer analysis, an advanced Python EDA phase was conducted using Pandas, Matplotlib, and NumPy. This phase provided statistical depth to the business findings, particularly around revenue distribution, growth dynamics, and customer behavior.

8.1 Revenue Distribution Analysis

Metric	Value	Interpretation
Total Unique Customers	95,420	Full addressable base in the dataset
Mean Revenue per Customer	₹212.82	Pulled up by high spenders at the right tail
Median Revenue per Customer	₹113.15	The "typical" customer spends significantly less than the mean
Revenue Skewness	69.1	Extremely right-skewed — a small group of customers drives disproportionate revenue
Pareto Finding	42% of customers → 80% revenue	The distribution does not perfectly follow 80/20 but shows meaningful concentration

A skewness of 69.1 is extremely high, confirming that revenue is not evenly distributed across the customer base. However, importantly, the Pareto analysis (42% of customers generating 80% of revenue) indicates that this is broad-base concentration — not elite-customer dependency. The business is not precariously reliant on a tiny group of mega-spenders, but still carries meaningful concentration risk.

8.2 Revenue & Growth Trend Analysis

Monthly revenue and order volume were analyzed across the full dataset timeline to characterize the platform growth lifecycle:

- 2017 — Hyper-growth phase. Strong month-over-month revenue increases. Customer acquisition was the primary driver of this growth surge.
- 2018 — Stabilization phase. Revenue plateaued as new customer acquisition slowed. Monthly Active Customers (MAC) mirrored this pattern exactly.
- November spike — A clear seasonal revenue peak was observed every November, consistent with Black Friday and year-end promotional activity.
- AOV stability — Average Order Value remained flat in the ₹190–₹220 range throughout both growth and plateau phases. Revenue growth was entirely volume-driven, not monetization-

driven. The platform was not extracting more value from existing customers — it was simply acquiring new ones.

8.3 Correlation Analysis

A correlation analysis between delivery_days and avg_review_score was conducted to quantify the relationship between logistics performance and customer satisfaction.

Correlation Finding

Correlation coefficient: -0.3 (negative, moderate strength)

Slower delivery does reduce customer satisfaction meaningfully. However, this correlation does not extend to improved retention — even customers who received fast deliveries and gave high review scores overwhelmingly did not return. This confirms that operational efficiency is necessary but not sufficient to solve the retention problem.

9. Power BI Dashboard Design

The final phase of the project converted all SQL-validated analytical findings into an executive-ready, three-page Power BI dashboard. The objective was professional visual storytelling — not generic chart building — with every visual tied to a specific business question.

9.1 Data Connection & Model Setup

Connecting PostgreSQL to Power BI

Power BI was connected directly to the PostgreSQL analytics schema (olist_analytics) using the built-in PostgreSQL connector. Only the five clean analytics-layer tables were loaded — no raw schema tables were imported into the visual layer.

Relationship Modeling

Relationships were configured as One-to-Many connections from each dimension table to the central fact table, using single cross-filter direction. Bidirectional filtering was avoided to prevent ambiguous filter propagation and performance issues. This mirrors the Star Schema design established at the SQL layer.

Date Table Configuration

The dim_date table was explicitly marked as the Date Table in Power BI. A critical fix was applied to resolve month sorting — the default alphabetical sort caused months to display out of chronological order. A YearMonth integer field was created and used as the sort key to ensure all time-series visuals display in correct temporal order.

9.2 DAX Measures Layer

All KPIs were implemented as explicit DAX measures rather than calculated columns. This architectural choice was made to prevent the query resource exceeded error encountered during development, where heavy calculated columns caused performance crashes. Measures evaluate only when needed (filter context), while calculated columns evaluate for every row — making measures essential for large datasets.

The following measures were created across the model:

DAX Measure	Business Purpose
Total Revenue	Sum of total_payment — primary revenue metric across all visuals
Total Orders	Distinct count of order_id for volume tracking
Total Customers	Distinct count of customer_unique_id

Average Order Value (AOV)	Total Revenue / Total Orders — basket size indicator
One-Time Customers	Customers with exactly 1 distinct order
Repeat Customers	Customers with more than 1 distinct order
Repeat Rate %	Repeat Customers / Total Customers — core retention KPI
Churn Rate %	1 - Repeat Rate % — mirrors acquisition dependency
Repeat Revenue	Revenue generated exclusively by repeat customers
Monthly Repeat Rate	Month-by-month repeat rate trend using time intelligence
AOV by Order Group	Average spend segmented by 1st, 2nd, 3rd, and 4+ order frequency
Revenue Rank	RANKX with SKIP mode to correctly rank customers by revenue without percentile distortion
Cumulative Revenue %	Running total of revenue percentage for Pareto analysis
Revenue Bucket Segmentation	Classifies customers into Top 10%, 20%, 30%, 40%, and Remaining 60%

Note: RANKX was implemented using SKIP mode (rather than DENSE) to prevent percentile distortion in the Pareto segmentation. DENSE ranking assigns the same rank to ties and then skips no numbers, which artificially inflates lower-tier bucket membership when used for percentage groupings.

10. Dashboard Architecture — Three-Page Design

The dashboard was architected as three purpose-built pages, each answering a distinct business question. Visuals were selected to serve analytical storytelling rather than aesthetic variety.

10.1 Page 1 — Executive Overview

Page Objective

Provide leadership with an immediate, high-level view of platform health — revenue performance, order volumes, and geographic and category concentration — without requiring any drill-down.

KPI Cards

Four headline KPI cards were placed at the top of the page for instant executive readability:

Total Revenue	Total Orders	Total Customers	Avg Order Value
₹20.3 Million	~99,000	~95,000	₹205.83

Visuals on This Page

- Monthly Revenue Trend — Line chart showing the full revenue growth arc from 2016 through 2018, revealing the hyper-growth phase and subsequent plateau.
- Monthly Orders Trend — Companion chart tracking order volume, confirming that revenue growth mirrors volume growth rather than price increases.
- Top 5 States by Revenue — Bar chart highlighting geographic concentration, with São Paulo contributing approximately 37% of total revenue.
- Top 10 Products by Revenue — Horizontal bar chart identifying category-level revenue leaders.
- Year Slicer — Interactive filter enabling executives to segment all visuals by year.

Key Findings from Page 1

- Revenue grew strongly through 2017 on the back of customer acquisition volume, then plateaued through 2018 as acquisition growth moderated.
- São Paulo and the top three states together account for approximately 62% of platform revenue — a significant geographic concentration risk.
- No single product category represents more than 8–9% of total revenue, indicating healthy category-level diversification.

- Average Order Value remained stable at ₹190–₹220 throughout the entire period, confirming that growth was volume-driven, not monetization-driven.

10.2 Page 2 — Customer Retention & Revenue Intelligence

Page Objective

Diagnose the sustainability of the revenue model by quantifying retention weakness, understanding the value gap between one-time and repeat buyers, and identifying where revenue is concentrated.

Step 1: Customer Classification & Retention KPIs

The page opens with the most critical business metrics: the classification of customers into one-time and repeat segments, with associated retention and churn rates prominently displayed.

Segment	Customer Count	% of Total	Avg Spend per Person
One-Time Customers	92,507	96.95%	₹138.67
Repeat Customers	2,913	3.05%	₹262.03

Step 2: Pareto & Revenue Concentration Analysis

Customers were ranked by total lifetime spend using RANKX (SKIP) and then grouped into revenue buckets to visualize how revenue concentrates across the customer distribution.

- Revenue buckets defined: Top 10%, Top 20%, Top 30%, Top 40%, and Remaining 60%.
- A Combo Chart was used — column bars representing revenue by bucket, with a line overlay showing Cumulative Revenue % — providing a Pareto visualization that clearly communicates concentration.
- Sorting of buckets required a custom sort fix to prevent alphabetical ordering (which would show "Remaining 60%" before "Top 10%").
- The analysis validated the Python EDA finding: approximately 42% of customers generate 80% of platform revenue.

Step 3: RFM Segmentation Visual

The RFM segments (Champions, Loyal Customers, Potential Loyalists, Big Spenders, At Risk, Lost Customers) were visualized to give the business a clear picture of where their customer value is concentrated and where risk is highest. The prominent "At Risk" group of ~22,000 customers stands out as the immediate priority for retention marketing.

Step 4: AOV by Order Group

Average Order Value was calculated and visualized separately for customers who placed their 1st, 2nd, 3rd, and 4+ orders. This analysis confirms a clear pattern: the more orders a customer places, the higher their average spend per order. This rising-AOV-with-engagement insight strengthens the business case for investing in retention programs — the financial reward of converting a one-time buyer to a repeat buyer goes beyond just additional orders.

Step 5: Monthly Retention Trend & State-Level Retention

- Monthly Repeat Rate % — a trend line confirming that retention remains consistently low (~3%) and has not improved organically over time. There is no evidence of the platform naturally developing stronger repeat behavior.
- Repeat Rate % by State — confirms that the 3% rate is uniform across all major states. No region shows significantly stronger loyalty behavior, reinforcing the conclusion that the problem is structural rather than geographic.

10.3 Page 3 — Logistics & Operations Intelligence

Page Objective

Assess operational delivery performance, quantify the true revenue at risk from logistics failures, and identify which states and categories experience the greatest delivery challenges.

Step 1: Core Operational KPIs

Metric	Value	Interpretation
Average Delivery Days	12.41 days	Overall average across all orders and states
Delivery Variability (Std Dev)	9.46 days	High variability indicates inconsistent logistics performance
Fast Delivery %	33.5%	Only 1 in 3 orders qualifies as fast delivery
Late Delivery %	27.9%	Approximately 1 in 4 orders is delivered late
On-Time Reliability %	82.35%	Platform delivers on time 4 out of 5 orders

Step 2: Operational Risk Revenue Modeling

A refined risk model was developed to move beyond simply labeling all late orders as "at risk." True operational revenue risk was defined as orders that were both slow AND received a low review score. This combination — delivery failure plus customer dissatisfaction — represents the genuine financial exposure.

Risk Model Component	Value / Finding
Revenue from Very Slow Deliveries	₹2.68 Million
Operational Risk Revenue %	12.88% of total revenue
Key Insight	27.9% of orders are late, but only 12.88% of revenue is truly at risk — not all delays translate to dissatisfaction

Step 3: Logistics Breakdown & Regional Analysis

- Revenue Share vs Order Share by Delivery Bucket — a visual comparison confirming whether slower deliveries are disproportionately concentrated in high-value or low-value orders.
- Review impact analysis — confirming the -0.3 correlation between delivery speed and customer satisfaction quantified in the EDA phase.
- Category delivery scatter analysis — plotting delivery time against revenue by category to identify whether any specific product categories systematically experience worse logistics.
- Regional logistics comparison table — delivery time by state reveals São Paulo at approximately 8.6 days, major states at 11–15 days, and remote states at 20–24 days.

Logistics Conclusion

The platform is moderately reliable (82% on-time) but operationally inconsistent — the high standard deviation (9.46 days) reveals significant tail delays.

Underperforming states outside SP are experiencing longer delivery times driven by demand and geographic reach, not operational failure per se.

However, delivery variability does not meaningfully explain churn — the retention problem exists even in São Paulo, which has the fastest average delivery time on the platform.

11. Integrated Business Diagnosis & Strategic Findings

Combining all layers of the analysis — data engineering, SQL analytics, hypothesis testing, Python EDA, cohort analysis, RFM segmentation, and Power BI visualization — the following integrated diagnosis emerges.

11.1 Growth Model Identified

Platform Business Model (As Diagnosed)

Acquire new customer → One-time purchase → Immediate churn → Acquire again

The platform is structurally operating as an acquisition engine, not a retention machine. Revenue is generated by continuously attracting new buyers rather than cultivating ongoing relationships with existing ones.

11.2 Key Business Problems Identified

#	Problem	Evidence
1	**Critically low repeat rate**	3.05% repeat rate vs typical e-commerce benchmark of 20–30%
2	**Acquisition dependency**	Revenue plateau in 2018 directly mirrors slowing customer acquisition, not pricing or product issues
3	**No habitual purchase cycle**	Purchase gap analysis: wide distribution from 0 to 600+ days, no consistent behavioral rhythm
4	**Geographic revenue concentration**	São Paulo contributes 37% of revenue; top 3 states drive 62% combined — significant regional risk
5	**Logistics inconsistency**	Std deviation of 9.46 delivery days; remote states at 20–24 day averages vs São Paulo at 8.6
6	**Revenue skewness risk**	Skewness of 69.1; ~42% of customers drive 80% of revenue — moderate but real concentration risk

11.3 High-Value Opportunity

Retention ROI Calculation

Current state: Repeat rate = 3.05%, Repeat revenue = ~5.62% of total.

Repeat customers spend 2x more per transaction (₹262 vs ₹139) but are too few to meaningfully contribute to total revenue.

Doubling the repeat rate from 3% to 6% — by engaging just 2,900 additional customers to return once — could significantly increase revenue without acquiring a single new customer.

This is the highest-ROI opportunity available to the business, and it is currently being left entirely untapped.

12. Strategic Recommendations

Based on the complete analytical diagnosis, four strategic priority areas are recommended. Each recommendation is grounded in specific data findings from this project.

12.1 Build a Retention & Loyalty Infrastructure

Priority: Critical — This is the single highest-impact intervention available.

- Introduce a loyalty points program tied to repeat purchases to create tangible incentives for return visits.
- Implement personalized post-purchase remarketing campaigns triggered 14–30 days after a first order (aligned with the median 28-day return window identified in the gap analysis).
- Develop targeted re-engagement campaigns specifically for the ~22,000 At Risk RFM customers before they move to the Lost segment.
- Consider a subscription or membership model for high-frequency categories (bed/bath, sports, beauty) that already show above-average retention potential.

12.2 Stabilize Logistics & Reduce Delivery Variability

Priority: High — While delivery does not directly cause churn, inconsistency erodes trust.

- Focus on reducing the delivery standard deviation (currently 9.46 days) rather than just the average — customers respond negatively to unpredictability.
- Establish SLA commitments by state and product category, with fast-track routing for high-value orders (protecting the 12.88% revenue exposure segment).
- Expand logistics partnerships in remote states currently averaging 20–24 day delivery windows.

12.3 Protect & Grow the High-Value Customer Segment

Priority: High — Revenue concentration in a small group creates vulnerability.

- Identify Champions and Loyal Customers from the RFM model and provide them with priority service, early access to new products, or dedicated account support.
- Monitor the operational risk segment (slow delivery + low review score) in real-time and proactively intervene with compensation or re-delivery for high-value orders.
- Implement customer lifetime value (CLV) modeling to differentiate acquisition spend — investing more to acquire customers resembling the Repeat/Champion profile.

12.4 Diversify Geographic Revenue Exposure

Priority: Medium — Current concentration in São Paulo creates strategic fragility.

- Design state-specific marketing campaigns targeting underpenetrated regions with demonstrated purchase intent but lower current order volumes.
- Establish local logistics partnerships in growth-opportunity states to reduce delivery times and make the platform competitive outside the São Paulo corridor.
- Analyze category demand patterns by state to tailor seller recruitment and product availability to regional preferences.

13. Project Summary & Technical Architecture

This project represents a complete, full-stack analytics implementation — from raw data ingestion through to executive visual storytelling — built to professional industry standards.

13.1 Project Maturity & Scope

This is not a beginner dashboard or a single-tool analysis. The project spans multiple technical layers and analytical disciplines:

Layer	What Was Built
Data Engineering	Automated ETL pipeline, raw schema ingestion, reproducible data refresh process
Data Modeling	Star schema with one fact table and four dimension tables; correct business key identification
KPI Engineering	SQL-validated revenue, retention, cohort, and operational metrics before any visualization
Hypothesis Testing	Systematic elimination of four potential churn drivers through data evidence
Retention Analytics	Cohort matrix, RFM segmentation, purchase gap analysis, repeat rate trend analysis
Revenue Intelligence	Pareto analysis, skewness calculation, growth phase identification, AOV by frequency
Operational Analytics	Logistics risk modeling, regional delivery comparison, satisfaction correlation
Executive Visualization	Three-page Power BI dashboard with advanced DAX, correct sorting, and performance-optimized measures

13.2 Technical Challenges Solved

Several non-trivial technical problems were identified and resolved throughout the project:

- Wrong customer identifier — correcting from session-level customer_id to true business key customer_unique_id, which fundamentally changed the retention picture from "0% repeat" to "3.05% repeat."
- Portuguese category names — resolved by loading and joining the translation table into the dimension build, creating an English-language product catalog for international readability.
- Power BI resource exhaustion — caused by overuse of calculated columns for heavy aggregation. Resolved by migrating all intensive logic to DAX measures (evaluated at filter time, not row time).

- Month sort misalignment — default alphabetical sorting of month names caused time-series visuals to display out of order. Fixed by creating a YearMonth integer sort key in dim_date.
- Pareto percentile distortion — RANKX with DENSE mode assigned overlapping rank positions that skewed bucket membership. Corrected to SKIP mode for accurate percentile groupings.
- Bidirectional filter risk — deliberately avoided in relationship modeling to prevent ambiguous filter propagation across the Star Schema.

13.3 Next Analytical Phase

The following analytical extensions are the logical next steps for this project:

- Customer Lifetime Value (CLV) Modeling — predict the long-term revenue value of different customer acquisition channels and behavioral profiles.
- Cohort Heatmap — a full month-by-month retention heatmap across all acquisition cohorts to visualize how quickly and how consistently customers churn.
- Profitability View — incorporating seller-level cost data to move from revenue analysis to margin analysis.
- Seller Performance Intelligence — identifying top and underperforming sellers by delivery reliability, review score, and revenue contribution.
- Discount Impact Analysis — quantifying whether promotional pricing is genuinely driving acquisition of long-term customers or simply attracting one-time discount seekers.
- Price Sensitivity Analysis — understanding customer elasticity to inform pricing strategy and AOV growth potential.

End of Report

Olist E-Commerce Analytics Project | Full-Stack Business Analytics Documentation