

A B2B2C Marketplace Growth and Retention Analytics Framework: Data Science Documentation

Data Science Project Documentation
Marketplace Analytics Case Study
Olist Brazilian E-Commerce Dataset, 2017–2018
Email: setyonugrohodwibudi@gmail.com

Abstract—This paper documents a comprehensive end-to-end data science project for diagnosing growth efficiency and retention health in a two-sided B2B2C marketplace. Using the Olist Brazilian E-Commerce public dataset, the project connects business-to-business merchant acquisition performance with business-to-consumer customer retention outcomes. The central problem is the “leaky bucket” marketplace paradox: a platform may acquire many sellers and customers, yet fail to compound growth if acquisition channels produce low-value merchants or customers do not return after the first purchase. The proposed framework follows a structured analytics lifecycle covering business understanding, analytic approach, data requirements, data collection, data understanding, data preparation, modeling, evaluation, and deployment. The final analytical products include a B2B acquisition channel yield model, a B2C monthly cohort retention matrix, SQL-based data marts, and executive dashboard specifications. The documentation is designed to be understandable to business stakeholders while remaining detailed enough for data science, analytics engineering, and portfolio review.

Index Terms—B2B2C marketplace, cohort analysis, customer retention, marketing analytics, SQL analytics, data science lifecycle, Olist, funnel analytics, revenue yield, Power BI.

I. INTRODUCTION

Modern digital marketplaces operate as multi-sided ecosystems. They must attract sellers who provide inventory and attract customers who generate demand. In practice, these two growth engines are often managed separately. Marketing and sales teams focus on acquiring merchants, while product and customer success teams focus on consumer loyalty. This separation creates an analytical blind spot: the business may spend heavily on lead generation without understanding whether those leads become valuable sellers, and it may acquire many customers without measuring whether those customers return.

This project addresses that blind spot through a unified B2B2C marketplace analytics framework. The analysis uses the Olist Brazilian E-Commerce Dataset, which contains transactional marketplace records and marketing funnel data from 2017 to 2018. The goal is not merely to report sales volume, but to trace the complete growth path from merchant lead generation to customer repeat purchasing behavior.

The resulting documentation provides a professional data science project narrative in IEEE conference format. It translates raw project notes into a structured paper that is suitable for academic-style presentation, business communication, and portfolio demonstration.

II. BUSINESS UNDERSTANDING

A. Executive Context

The marketplace under study faces two strategic risks. First, merchant acquisition budget may be allocated based on marketing qualified lead volume rather than downstream seller value. Second, customer acquisition may generate one-time buyers rather than a loyal customer base. If both risks occur simultaneously, the business becomes dependent on continuous paid acquisition to sustain revenue.

The project frames this problem as a “leaky bucket” paradox. A marketplace can appear healthy because new leads, sellers, orders, and customers continue entering the system. However, if sellers acquired from certain channels produce low gross merchandise value (GMV), and customers rarely return after their first order, the marketplace is not building durable growth.

B. Problem Statement

The project investigates two linked business problems:

- 1) **B2B capital allocation risk:** Marketing budget is potentially being optimized for lead volume rather than conversion quality, sales velocity, or seller-generated GMV.
- 2) **B2C retention crisis:** Customer acquisition may be masking weak repeat-purchase behavior, producing a dependency on constant new-customer acquisition.

C. Objectives

The project has two primary objectives:

- **Objective 1: B2B efficiency.** Identify acquisition channels that generate sellers with the highest revenue yield per lead and the fastest days-to-close.
- **Objective 2: B2C health.** Quantify monthly customer cohort retention to determine whether customers return after their first purchase.

D. Stakeholders

The analysis supports several decision makers. The Chief Marketing Officer can reallocate spending toward high-yield channels. Product and customer success leaders can justify investments in loyalty programs, re-engagement campaigns, and post-purchase experiences. Sales operations teams can prioritize channels with faster sales cycles and higher conversion quality.

III. ANALYTIC APPROACH

The project uses a descriptive and diagnostic analytics approach. Instead of building a predictive model as the first step, the project creates reliable performance measures that explain where growth is efficient and where value leaks from the marketplace.

A. Analytical Questions

The analysis is guided by the following questions:

- Which marketing channels produce the highest number of leads?
- Which channels convert leads into sellers most efficiently?
- Which channels produce sellers that generate the most GMV?
- How long does it take for leads from each channel to close?
- How many customers return in the months after their first purchase?
- Does the marketplace show evidence of sustainable retention or one-time purchasing behavior?

B. Methodological Design

The solution combines funnel analytics and cohort analysis. Funnel analytics evaluates the B2B lead-to-seller process, while cohort analysis evaluates B2C repeat behavior over time. This dual-method design is appropriate because marketplace growth depends on both supply-side quality and demand-side loyalty.

IV. DATA REQUIREMENTS

The project requires data that can connect three analytical layers: marketing leads, converted sellers, and customer orders. Table I summarizes the main data requirements.

TABLE I
CORE DATA REQUIREMENTS

Data Area	Required Fields	Purpose
Marketing funnel	Lead identifier, origin channel, first contact date, seller identifier, won date	Measure lead volume, conversion, and days-to-close
Seller orders	Seller identifier, order identifier, item price, freight value	Attribute GMV to seller acquisition channels
Customer orders	Customer identifier, unique customer identifier, order purchase timestamp	Build customer cohorts and repeat-purchase timelines
Order status	Order status, approval and delivery dates	Exclude invalid or incomplete transactions when needed
Product and geography	Product category, city, state, geolocation	Enable future segmentation and operational analysis

A critical requirement is the availability of stable identifiers. The customer-level analysis must use a unique customer identity rather than an order-specific customer key, because retention requires tracking the same individual across purchases.

V. DATA COLLECTION

The data source is the Olist Brazilian E-Commerce public dataset. The collection layer includes transactional order files, order item records, customer records, seller records, product metadata, and marketing funnel tables. The dataset represents a Brazilian marketplace context and covers the main period from January 2017 through August 2018.

The project assumes a local analytical database implementation, such as SQL Server, where raw files are ingested into staging tables before being transformed into curated analytics views. This approach improves reproducibility because raw data, cleaned data, and final modeled outputs are separated.

VI. DATA UNDERSTANDING

Data understanding focuses on identifying the analytical grain, data quality limitations, and relationships among entities. The most important grains are lead-level records for B2B funnel analysis, order-item-level records for GMV attribution, and customer-month-level records for retention analysis.

A. Entity Relationships

The primary relationships are:

- Marketing qualified leads connect to closed deals through seller identifiers.
- Sellers connect to marketplace transactions through order item records.
- Customers connect to orders through customer identifiers and unique customer identifiers.
- Orders connect to order items, payments, reviews, and timestamps.

B. Data Quality Considerations

Common quality considerations include missing marketing origin values, leads that never convert, order status filtering, duplicated business keys, and incomplete timestamp fields. Missing channel values are not automatically discarded because they may represent a meaningful operational category. In this project, unknown or missing channels are retained as an explicit group when evaluating channel yield.

VII. DATA PREPARATION

The preparation process transforms raw marketplace tables into analysis-ready datasets. The design follows analytics engineering principles: preserve raw data, clean and standardize fields, create reusable intermediate models, and publish final business-facing views.

A. Cleaning and Standardization

Preparation tasks include converting timestamp strings into date or datetime types, standardizing marketing channel labels, validating seller and customer identifiers, and filtering transaction records to valid orders. Derived date fields such as purchase month, cohort month, and lead close duration are created to support time-based analysis.

B. Feature Engineering

The project engineers several analytical features:

- **Days to close:** Difference between first contact date and won date.
- **Lead conversion flag:** Indicator of whether a marketing lead became a seller.
- **Seller GMV:** Sum of order item value associated with a seller.
- **First purchase month:** Earliest purchase month for each unique customer.
- **Month index:** Number of months between a customer's cohort month and later purchase months.

VIII. MODELING

The modeling layer consists of two analytical models: a B2B channel yield model and a B2C cohort retention model. These are not machine learning models; they are business intelligence models designed for decision support.

A. B2B Channel Yield Model

The B2B model aggregates lead and seller outcomes by marketing channel. Its key performance indicators are shown in Table II.

TABLE II
MARKETPLACE ANALYTICS KPIS

KPI	Formula	Interpretation
Conversion Rate	Converted sellers / total MQLs	Channel efficiency
Average Days to Close	Won date minus first contact date	Sales velocity
Revenue Yield per MQL	Total GMV / total MQLs	Financial return per lead
Cohort Size	Customers with first purchase in month	Retention baseline
Month-N Retention	Returning customers in month N / cohort size	Loyalty health
Churn Rate	1 - retention rate	Customer loss rate

Revenue yield per MQL is treated as the B2B north-star metric because it combines lead generation and downstream economic value. A channel with many leads but low GMV can be less valuable than a smaller channel that converts high-quality sellers.

B. B2C Cohort Retention Model

The B2C model assigns each unique customer to the month of their first purchase. It then measures whether that customer purchased again in later months. This creates a retention matrix where each row is a cohort month and each column is a month index after acquisition.

The generic retention formula is:

$$Retention_{c,n} = \frac{Customers_{c,n}}{Customers_{c,0}} \times 100 \quad (1)$$

where c is the cohort month, n is the number of months after first purchase, $Customers_{c,0}$ is the cohort size, and

$Customers_{c,n}$ is the number of cohort members active in month n .

IX. EVALUATION

Evaluation determines whether the analytical outputs answer the original business questions. Because the models are descriptive and diagnostic, evaluation focuses on interpretability, metric validity, and decision usefulness rather than predictive accuracy.

A. B2B Evaluation

The B2B report should identify channels that differ meaningfully across lead volume, conversion rate, sales velocity, and GMV yield. A valuable result is a ranking of channels by revenue yield per MQL, because it exposes whether high-volume acquisition channels are actually generating marketplace value.

B. B2C Evaluation

The B2C cohort matrix should reveal whether customers continue purchasing after their acquisition month. If Month-1 retention is extremely low, the marketplace demonstrates leaky-bucket behavior. In that case, growth is driven more by new acquisition than by retained customer value.

C. Business Interpretation

The combined interpretation is more powerful than either analysis alone. A marketplace can have strong merchant acquisition but weak customer retention, or strong customer demand but inefficient seller acquisition. By showing both sides together, the framework helps leaders decide whether to invest in marketing reallocation, seller quality improvement, product retention programs, or post-purchase lifecycle communication.

X. DEPLOYMENT

The deployment plan contains two complementary outputs: an executive dashboard and a portfolio visualization layer.

A. Executive Dashboard

The recommended Power BI dashboard is a single-page executive command center. The top row contains KPI cards for total MQLs, converted sellers, average days-to-close, and average Month-1 retention. The left side presents B2B channel performance using a combination of bar charts, line charts, and detailed tables. The right side presents a B2C retention heatmap or matrix with conditional formatting.

This design helps executives compare acquisition efficiency and retention health on one screen. It prevents the organization from optimizing merchant acquisition while ignoring customer loyalty, or improving retention while starving the platform of quality seller inventory.

B. Portfolio Visualization

A Python-based visualization layer can complement the dashboard. Recommended charts include a revenue-yield-per-MQL bar chart by channel and a cohort retention heatmap. These charts demonstrate technical capability with SQL, pandas, seaborn, and matplotlib while communicating the business story clearly.

XI. REPRODUCIBILITY AND GOVERNANCE

A professional data science project must be maintainable. The recommended repository structure separates documentation, SQL scripts, notebooks, reports, and raw or processed data. SQL views should contain clear names such as `analytics.b2b_growth_yield` and `analytics.b2c_cohort_retention`. Documentation should define assumptions, metric formulas, and known limitations.

Governance considerations include protecting customer and seller identifiers, avoiding unnecessary exposure of personally identifiable information, and documenting whether analyses are based on raw, cleaned, or filtered transaction records.

XII. LIMITATIONS

The project is subject to several limitations. First, the dataset is historical and represents a specific Brazilian e-commerce marketplace context from 2017 to 2018. Second, marketing channel values may contain missing or ambiguous categories. Third, descriptive cohort retention does not explain causality by itself; additional experimentation or causal inference would be required to prove why customers do or do not return. Fourth, GMV-based seller value may not capture profitability if costs, discounts, refunds, and marketing spend are unavailable.

XIII. RECOMMENDATIONS

Based on the framework, the project recommends the following actions:

- 1) Reallocate marketing investment toward channels with high revenue yield per MQL, not merely high lead volume.
- 2) Investigate channels with long days-to-close to identify sales process bottlenecks.
- 3) Launch retention initiatives if Month-1 retention is weak, including lifecycle email, loyalty incentives, personalized recommendations, and post-purchase support.
- 4) Segment retention by product category, seller quality, geography, and delivery experience to identify drivers of repeat purchase.
- 5) Convert the descriptive framework into a recurring dashboard monitored by marketing, product, and sales leadership.

XIV. CONCLUSION

This paper presents a detailed professional documentation artifact for a B2B2C marketplace data science project. The framework connects merchant acquisition and customer retention, two areas that are often analyzed separately but jointly determine marketplace health. By combining B2B funnel analytics with B2C cohort retention analysis, the project identifies where acquisition capital creates value and where customer value leaks from the system. The final deliverables—SQL-modeled analytics views, executive dashboard specifications, and portfolio visualizations—provide a practical foundation for data-driven marketplace strategy.