

Data Validation & Quality Check Plan

E-commerce Operations & Customer Experience Analytics – Olist

Purpose

This document defines the data validation and quality assurance process to ensure all insights are derived from accurate, reliable, and business-trustworthy data.

Data Sources in Scope

Orders, order items, sellers, customer reviews, and payments tables representing transactional e-commerce operations.

Critical Data Elements

order_id, purchase timestamp, delivered customer date, estimated delivery date, review score, seller_id, price, and quantity.

Completeness Checks

Missing delivery dates are excluded from delay analysis. Orders without reviews are retained but excluded from satisfaction metrics. Records without seller identifiers are removed from seller-level analysis.

Uniqueness Checks

Duplicate order identifiers and order-item combinations are investigated and deduplicated to prevent double counting.

Validity Checks

Review scores must fall between 1 and 5. Delivery days must be non-negative. Prices must be greater than zero. Extreme outliers are flagged for review.

Consistency Checks

Orders delivered before purchase timestamps are removed. Cancelled orders with reviews are excluded from satisfaction analysis. Delivery status is validated against timestamps.

Timeliness Checks

Date ranges are validated to ensure full period coverage. Partial or incomplete periods are excluded from trend analysis when necessary.

Derived Metric Validation

Delivery delay is calculated as actual minus estimated delivery date. Revenue is computed as price multiplied by quantity. On-time delivery rate is validated across aggregation levels.

Handling Data Issues

Missing reviews are excluded from rating metrics. Extreme delays are analyzed separately. Cancelled orders are excluded from revenue calculations.

Output of This Step

The output includes a clean validated dataset, a data quality summary, and documented assumptions ensuring confidence in downstream analysis.