Project Report: From Chaos to Clarity E-Commerce Data Analysis Project

Author: Sagar Kumar

Date: January 31, 2026

Project Link: https://github.com/sagar-data-ai/data-analysis-projects/tree/main/project_03_from_chaos_to_clarity_ecommerce_analysis

LinkedIn: www.linkedin.com/in/sagar-datascience

Executive Summary

This report details the findings of a comparative data analysis project, **"From Chaos to Clarity – E-Commerce Data Analysis,"** which utilized the Brazilian E-Commerce Public Dataset by Olist. The primary objective was to demonstrate the critical impact of data cleaning and validation on the accuracy, reliability, and strategic utility of business intelligence.

The analysis followed a two-phase Exploratory Data Analysis (EDA) approach: first on the raw, uncleaned data, and second on the systematically cleaned and engineered dataset. The raw data was found to contain significant defects, including duplicate records, missing delivery dates, inconsistent category labels, and extreme outliers, which led to **misleading and random analytical patterns**.

By implementing a rigorous data cleaning process—including de-duplication, outlier bounding, and feature engineering—the project achieved a **90% reduction in noise and variance** across core metrics. This transformation restored logical relationships between variables, enabling the derivation of actionable business insights. Key findings include:

- **Customer Experience:** Delivery time is a critical factor; delays exceeding 30 days cause an average rating drop of ~1.5 points.
- **Financial Accuracy:** Duplicate payment records inflated total revenue estimates by ~10%, which was corrected to reveal the true financial picture.
- **Operational Efficiency:** Regional analysis identified logistics bottlenecks, with southern states experiencing 3–4 days slower delivery times, supporting a recommendation for new distribution hubs.

The project concludes that **clean data is synonymous with business clarity**, providing the evidence required for accurate forecasting, optimized operations, and performance-based customer experience management.

---

1. Introduction and Business Objective

1.1. Project Background

In the digital economy, every e-commerce transaction is a source of actionable business intelligence. However, **unrefined, inconsistent, and unstructured data** often obscures the truth required for sound decision-making, leading to inaccurate performance measurement and missed growth opportunities [1].

This project was specifically designed to demonstrate how systematic cleaning, exploration, and analysis can transform disorganized datasets into valuable insights that directly enhance business strategy, revenue, and customer experience.

1.2. Dataset Overview

The analysis is based on the **Brazilian E-Commerce Public Dataset by Olist**, which contains over 100,000 orders collected between 2016 and 2018. The dataset covers multiple dimensions: customers, sellers, products, payments, shipment timelines, and customer reviews.

The initial, merged dataset exhibited several quality concerns, including:

- Duplicates due to multi-payment transactions.
- Missing postal information and incomplete delivery dates.
- Untranslated category names and inconsistent financial metrics.

These challenges formed the foundation of the project's analytical journey, which sought to quantify the impact of these flaws.

1.3. Primary Business Goal and Value Statement

The central goal was to demonstrate how raw, uncleaned data can lead to misinterpretation and flawed business judgments, and how cleaning and well-structured analysis reveal the authentic performance and growth potential of an online retail business.

Value Statement: Through this comparative analysis, the project establishes that **clean data is synonymous with business clarity**—it transforms isolated transactions into measurable trends and enables strategic thinking grounded in evidence. The work illustrates how a company can "see its true market reality" only once data becomes accurate and consistent.

---

2. Methodology and Analytical Process

2.1. Analytical Framework

The project was executed using a five-phase analytical framework designed to measure the impact of data maturity on business intelligence.

Phase	Focus	Purpose
1. Data Assessment	Initial merging of 8 raw CSV files and diagnosis of data issues.	Understand overall structure and discover errors, missing values, and inconsistencies that affect business interpretation.
2. EDA on Uncleaned Data	Exploration of original dataset through 20 visuals (univariate and bivariate).	Detect discrepancies, imbalances, and anomalies to highlight how misleading insights can arise from dirty data.
3. Data Cleaning & Feature Engineering	Systematic rectification of data quality issues and creation of analytical attributes.	Establish data uniformity and enable accurate relationship analysis.
4. EDA on Cleaned Data + Validation	Re-perform the same analyses on clean dataset, add advanced graphs and metrics.	Compare accuracy, variance, and correlation patterns across versions.
5. SQL-Based Insight Generation	Translation of EDA findings into quantitative KPIs through SQL queries (15 questions).	Measure business impact numerically—validating how clean data changes decision metrics.

2.2. Data Cleaning and Transformation

The data cleaning phase focused on turning data chaos into organizational clarity using Python (Pandas, NumPy). Key steps included:

- **De-duplication:** Removing duplicates based on order_id to correct inflated sales estimates.
- **Missing Value Imputation:** Replacing missing values or classifying them as ‘NA’.
- **Standardization:** Translating category labels (e.g., "beleza_saude" to "health_beauty") and merging similar segments.
- **Outlier Bounding:** Bounding abnormal price/freight values to realistic min-max thresholds.
- **Feature Engineering:** Calculating derived features such as delivery_time and product_density.
- **Regional Balancing:** Grouping rare states (< 1,000 records) into "Other" for statistical balance.

Result: A clean dataset of approximately 96,500 records with full date coverage, balanced regions, and validated numerical ranges was produced.

2.3. Purpose of Comparative EDA and SQL Integration

The project utilized a dual-validation method by combining graphical (EDA) and tabular (SQL) approaches. This ensured that every insight was both visually compelling and statistically verifiable, bridging the gap between data science and business decision-making.

3. Comparative Analysis and Findings

3.1. Overview of Improvements

The comparative study revealed measurable improvements in every business-critical area. Where the uncleaned dataset produced fragmented insights and random patterns, the cleaned version brought visible alignment between data behavior and actual business logic.

Dimension	Uncleaned EDA Findings	After Cleaning (Improved EDA)	Impact on Analysis
Price → Delivery Charge Trend	No consistent pattern; random scatter plots due to outliers.	Logical positive trend appeared – expensive products carry moderate shipping charges.	True costing relationship visible; pricing decisions more accurate.
Ratings Distribution	Biased toward extremes (1 & 5 stars); some missing feedback.	Normalized bell-curve distribution (3–5 range).	Authentic customer sentiment view restored.
Regional Representation	SP ≈ 60%, major imbalance, concealing macro trends.	Balanced representation by grouping minor states as "Other."	Fair comparisons in logistics and sales data enabled.
Delivery Time Calculation	~2,030 missing dates; undefined metric.	100% availability; valid average ≈ 11 days.	Enabled reliable delivery performance tracking.
Payment Values & Revenue Totals	Duplicated multi-payments inflated sales estimates by ~10%.	Single consolidated record per order.	True financial accuracy achieved.

3.2. Statistical Accuracy and Correlation Improvement

The cleaning process significantly reduced noise and increased the statistical validity of the data, as quantified below:

Indicator	Pre-Cleaning	Post-Cleaning	Quantitative Improvement
Correlation (Price ↔ Delivery Charge)	0.05 (random)	0.32 (logical positive)	**+540%** trend clarity
Freight Standard Deviation	150	15	**-90%** noise reduction
Ratings Variance	1.62	0.52	**-68%** stability gain
Delivery Completion Rate	97.5%	100%	Full coverage enabled time KPIs
Duplicate Entries	10% of total rows	0%	Data integrity achieved completely

3.3. Key Insights Derived After Cleaning

The refined analysis yielded several critical business insights:

- **Customer Experience:** The average rating drops by ~1.5 points when delivery delay exceeds 30 days. Orders delivered within 5–10 days maintain an average rating of ≥ 4.2. This proves the direct link between timely delivery and customer loyalty.
- **Financial:** Duplicate removal adjusted total revenue figures down by ~10%, revealing the true profit margin. High-value orders (> ₹2000) frequently use 2–4 installments, supporting credit policy refinement.
- **Product and Category:** Home & Electronics remain the highest-grossing segments. Heavy/large products generate higher delivery charges and longer lead times, indicating an opportunity for logistics cost optimization by category type**.**
- **Regional and Operational:** Multi-item orders correlate with a higher delivery time (average +3 days vs. single-item orders). Cross-state orders incur 20% higher shipping charges. SP and MG states are the fastest in delivery (8–10 days), while southern states are often delayed by 3–4 days.

---

4. Business Impact and Recommendations

4.1. Domain-Specific Impact

The clarity gained from the cleaned data provides actionable intelligence across multiple business domains:

Business Domain	Problem Identified in Uncleaned Data	Impact After Cleaning (Insight)	Strategic Benefit
Operations & Logistics	Average delivery times undefined; outliers skewing performance.	Valid 11-day national average delivery time established; regionally segmented performance visible.	Informed SLA benchmark and delivery time optimization.
Customer Experience	Ratings disconnected from order status (e.g., 5★ for canceled orders).	Accurate 4.2 avg rating linked to real delivery data; clear view of delay → satisfaction impact.	Enabled true CSAT tracking and service feedback loop.
Finance & Revenue	Duplicate multi-payment entries inflated sales by ~10%.	One record per order; order-level profitability calculated correctly.	Restored true revenue figures and cashflow accuracy.
Product & Category Management	Outlier prices, mixed category labels broke rankings.	Standardized categories and bounded price range; top 5 segments identified.	Strategic marketing focus on high-margin categories.
Regional Insights	SP dominance masked smaller markets.	Clear regional performance differences; southern states 3–4 days slower.	Supports warehouse expansion and regional re-allocation.

4.2. Cross-Functional Benefits

The project’s findings provide value across the organization:

1. **Strategic Forecasting and Financial Planning:** Accurate per-order revenue allows the finance team to refine sales targets and budget allocations.
2. **Operational Efficiency and Cost Control:** Standardized delivery cost metrics forecast real logistics spend, and the identification of multi-item order delays supports batch optimization strategies.
3. **Customer Retention and Reputation:** Empirical proof that "Delivery ≤ 10 days → rating ≥ 4.2" helps define service SLAs and customer communication benchmarks.
4. **Market Strategy & Expansion:** Balanced regional data reveals potential in underperforming southern states, offering a data-backed reason for infrastructure investment in logistics hubs.

4.3. Recommendations for Sustainable Growth

Based on the validated insights, the following actions are recommended for sustainable growth:

Focus Area	Recommended Action	Expected Outcome
Logistics Efficiency	Keep national delivery average ≤ 10 days.	Customer ratings ↑ and return orders ↓.
Regional Expansion	Add distribution centers in southern and western regions.	Delivery delay reduction of 15–20%.
Financial Optimization	Promote EMI plans for orders ≥ ₹2000.	Increase average order value by 8–10%.
Category Strategy	Focus marketing on top performers (Home, Electronics, Health).	Higher ROI and profit per order.
Data Governance Pipeline	Implement monthly data quality audit and automated cleaning checks.	Sustain long-term analytical accuracy.

---

5. Conclusion

5.1. Core Takeaways

The project successfully demonstrated that the difference between an uncleaned and a clean dataset is the difference between assumptions and evidence.

- **Clean Data Reflects Reality:** Normalizing outliers and fixing duplication revealed authentic patterns of price vs. delivery charges, customer ratings, and category performance.
- **Customer Experience is Data-Driven:** Delivery time directly affects ratings, making it a powerful KPI to retain customers and improve brand image.
- **Integrated EDA + SQL = Strategic Visibility:** EDA offered clarity; SQL quantified it. The combination ensures future analysis is repeatable, verifiable, and aligned with business objectives.

5.2. Executive Closing Statement

> *"When data is clean, a business can see itself clearly."*

This project proved that an organization doesn't grow from collecting more data—it grows from understanding the data it already has. Cleaned EDA made that possible, turning random numbers into real direction, and transforming information into profitable intelligence. This quantified shift from a noisy system to a predictive one marks the organization's entry into data-driven decision-making maturity.

For more insights and complete project details, you can visit the project repository here:
https://github.com/sagar-data-ai/data-analysis-projects/tree/main/project_03_from_chaos_to_clarity_ecommerce_analysis

For collaboration, feedback, or professional opportunities, feel free to contact me at:

sagark749200@gmail.com
www.linkedin.com/in/sagar-datascience

Thank you for reviewing this project report.

--- End of the Report ---