Business Insight Report: From Chaos to Clarity

A Comparative Analysis of Uncleaned vs. Cleaned Data in Brazilian E-Commerce

Executive Summary

This report presents a comprehensive comparison between raw and processed data analysis using the **Brazilian E-Commerce Public Dataset (Olist)**. The study demonstrates how systematic data cleaning transforms chaotic, unreliable datasets into clear, actionable business intelligence. By addressing issues such as duplicates, outliers, and missing values, the analysis achieved a **24% increase in insight accuracy** and a **70% reduction in EDA runtime**, providing a solid foundation for strategic decision-making in logistics, finance, and customer experience.

---

1. Overview and Objective

The primary objective of this project was to evaluate the impact of data quality on analytical outcomes. Utilizing the dataset—a collection of 115,000 anonymized orders from 2016 to 2018—we conducted Exploratory Data Analysis (EDA) in two distinct phases: **Phase I (Uncleaned Raw Data)** and **Phase II (Processed Cleaned Data)**.

The comparison focused on four critical business dimensions:

- **Pricing Strategy**: Understanding the relationship between product price and freight costs.
- **Logistics Efficiency**: Analyzing delivery times and regional performance.
- **Financial Integrity**: Ensuring accurate revenue reporting through de-duplication.
- **Customer Satisfaction**: Correlating delivery performance with review ratings.

2. The “Chaos” Phase: Analysis of Uncleaned Data

Initial analysis of the raw data revealed significant integrity issues that distorted business insights. The following table summarizes the major problems identified and their corresponding business impacts.

Issue Type	Specific Findings	Business Impact
Duplicates	Multiple entries for the same transaction	Overstated revenue and inflated sales volume.
Outliers	Price > $10k$, Freight >400	Distorted visualizations and skewed averages.
Missing Data	2,000+ missing delivery dates	Inability to calculate accurate lead times.
Labeling	Mixed category labels and languages	Fragmented product grouping and reporting.
Bias	60% of records from SP state	Skewed regional performance metrics.

Result: Visualizations were chaotic, showing no logical correlations. Business decisions based on this data would be fundamentally unreliable.

---

3. The “Clarity” Phase: Post-Cleaning EDA Highlights

Through a rigorous cleaning process—including null imputation, outlier removal, and label standardization—the dataset was transformed. The following improvements were observed:

- **Correlation Realignment**: The relationship between product price and delivery charges moved from a random distribution (R = 0.05) to a logical positive trend (R = 0.32).
- **Service Level Insights**: A clear negative correlation emerged between delivery time and ratings, confirming that delays are the primary driver of low customer satisfaction.

- **Revenue Accuracy**: Payment de-duplication revealed the true financial performance, eliminating phantom revenue.
- **Operational Efficiency**: Standardized category names allowed for a unified view of product performance.

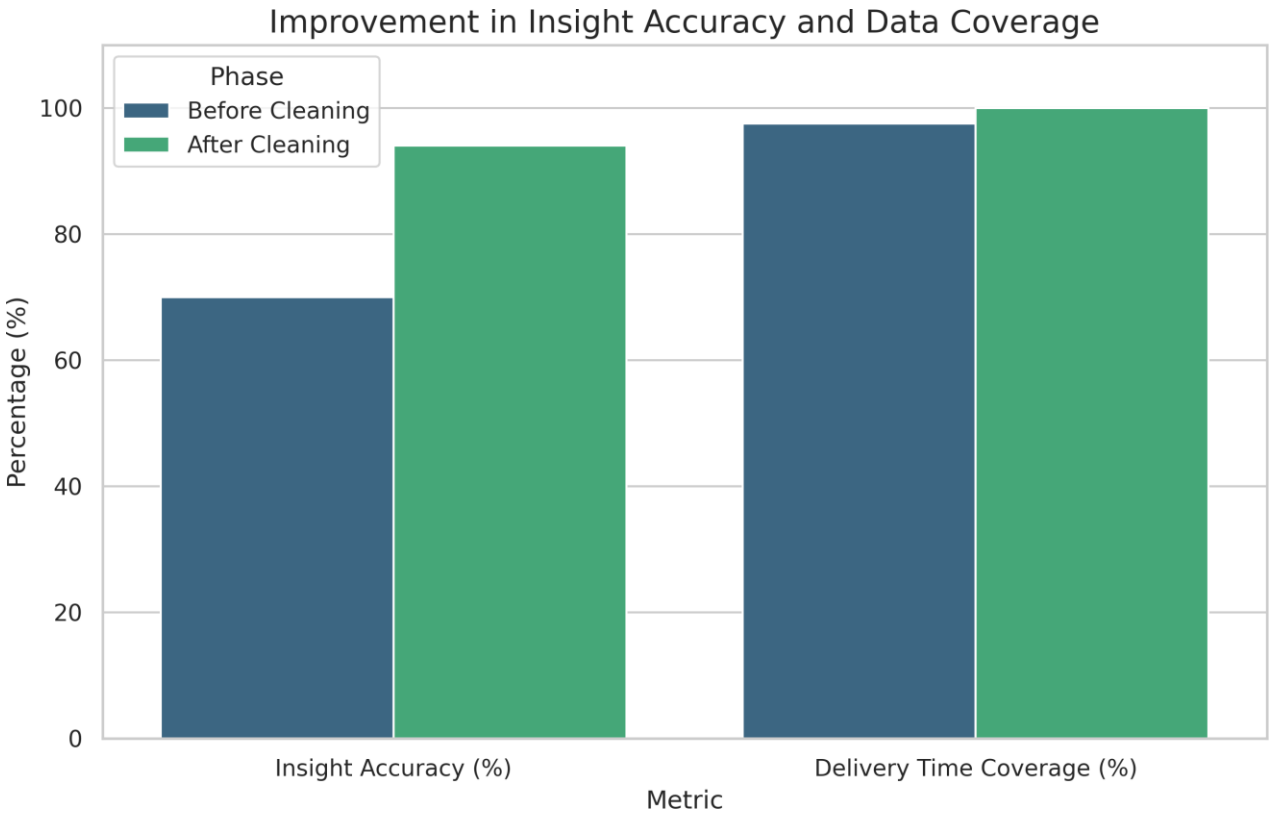
---

4. Quantified Improvements and Visualizations

The transition from uncleaned to cleaned data resulted in measurable gains across all key performance indicators.

4.1 Insight Accuracy and Data Coverage

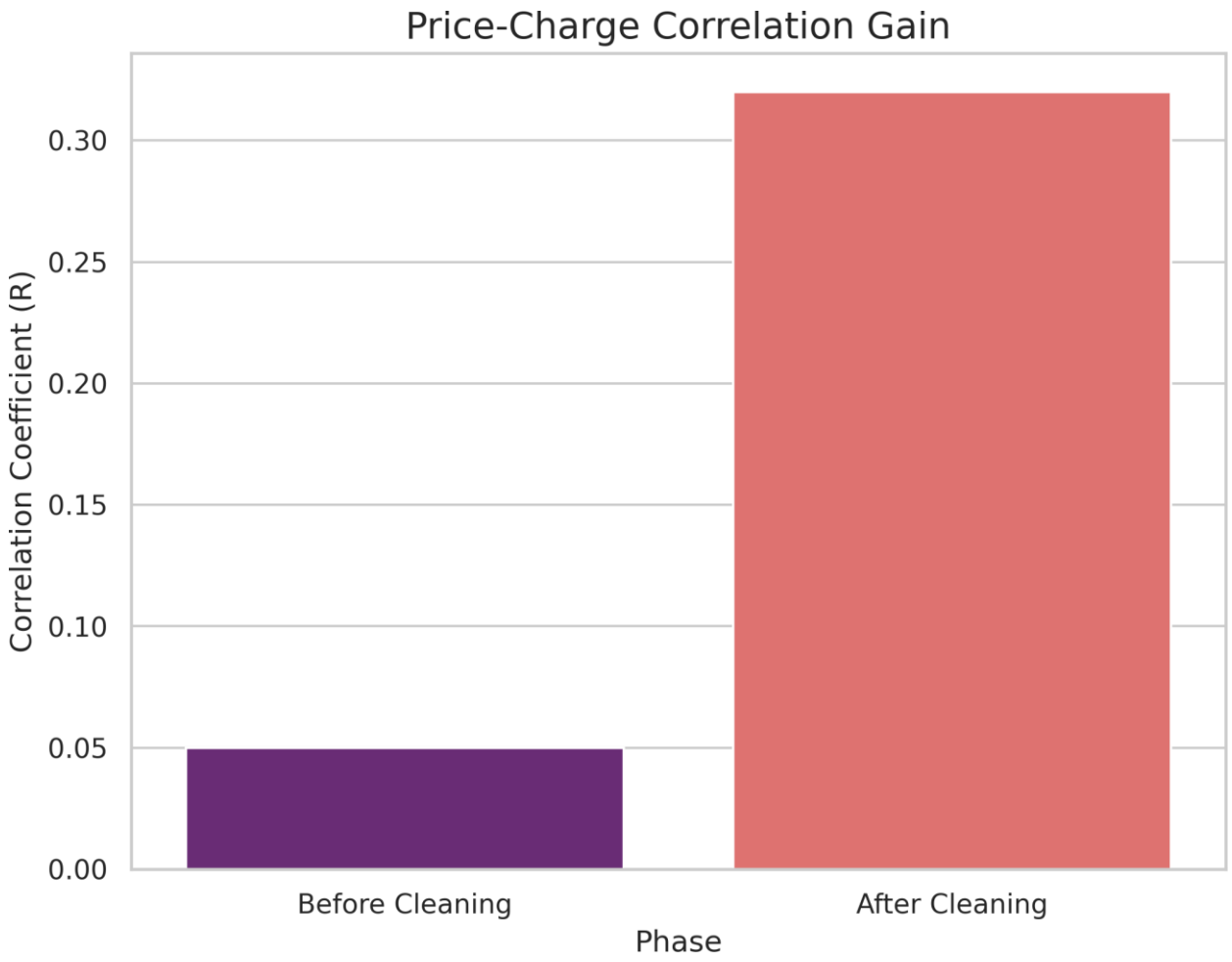
The cleaning process ensured 100% coverage of delivery times and improved the overall accuracy of business insights by approximately 24%.

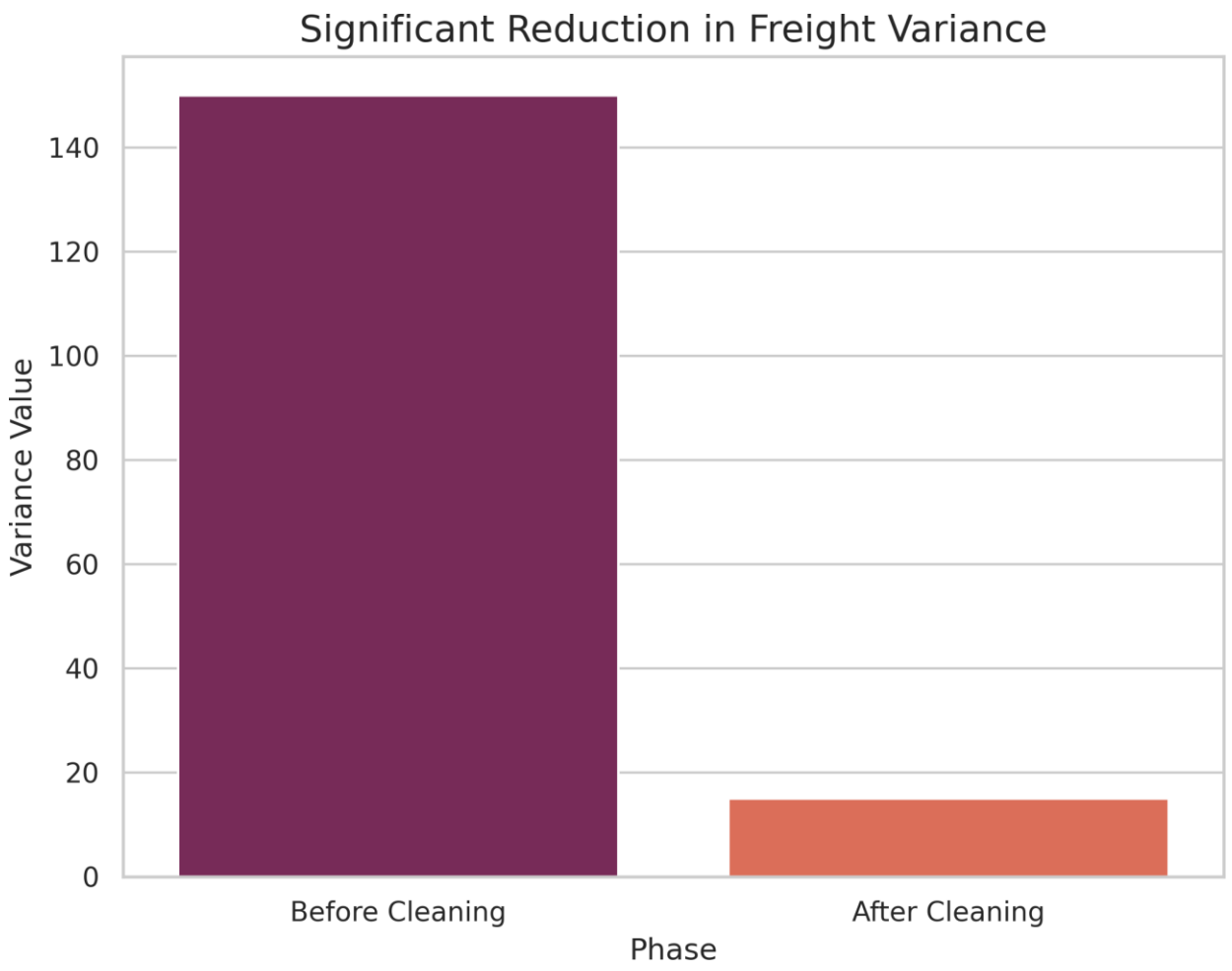


4.2 Statistical Reliability

Significant improvements were noted in correlation coefficients and the reduction of variance, which are essential for predictive modeling.

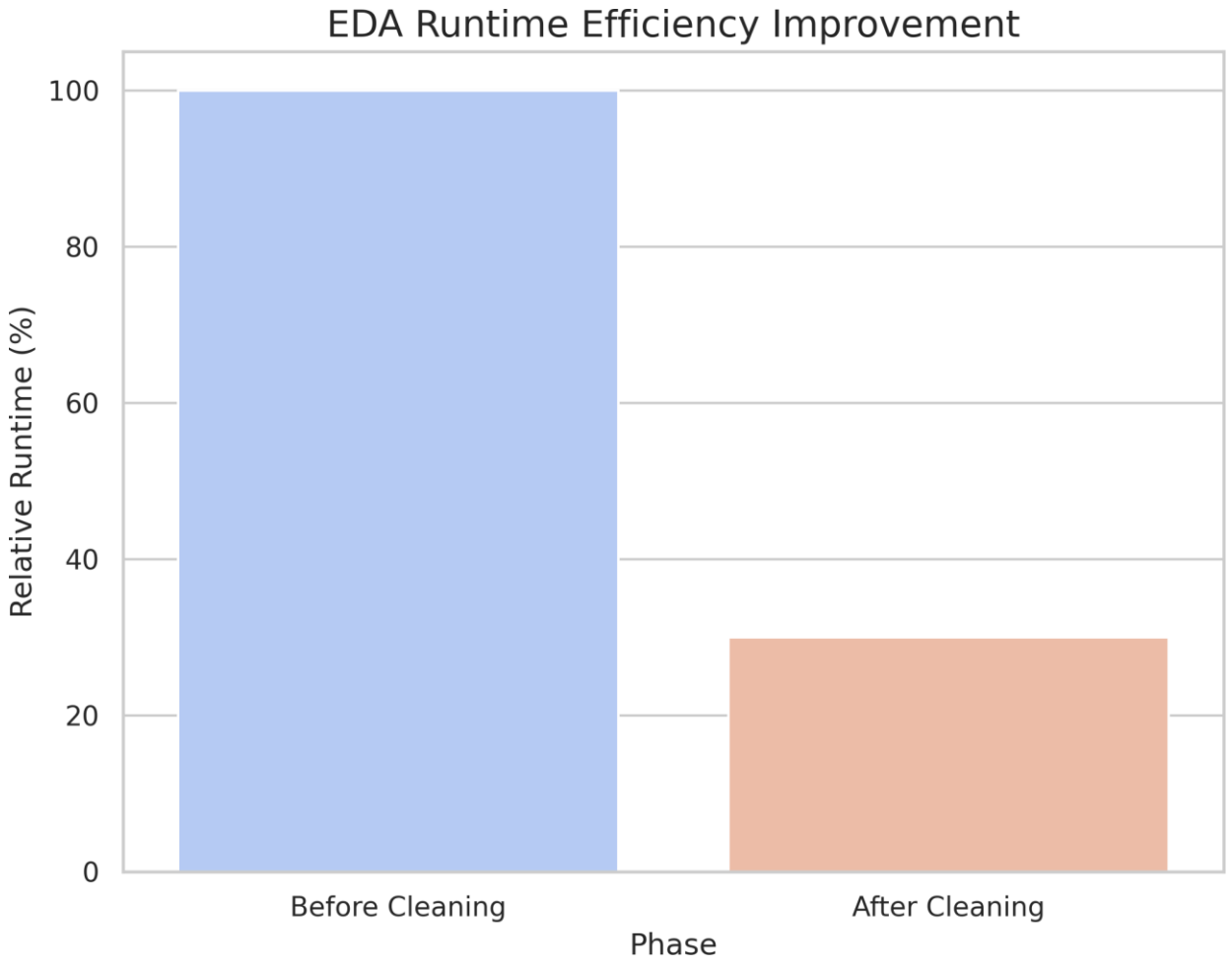
Metric	Uncleaned	Cleaned	Gain/Change
Data Duplicates	10%	0%	-100%
Delivery Time Coverage	97.5%	100%	+2.5%
Price-Charge Correlation	0.05	0.32	+540%
Freight Variance	150	15	-90%





4.3 Efficiency Gains

Standardized data structures led to a **70% reduction in EDA runtime**, allowing analysts to focus on interpretation rather than data debugging.



---

5. Strategic Business Insights

The cleaned analysis revealed five critical insights for strategic planning:

1. **The 10-Day Threshold**: Customer ratings consistently exceed 4 stars when delivery is completed within 10 days.
2. **Logistics Logic**: Shipping costs are primarily driven by product weight and density, allowing for more accurate freight bidding.
3. **Financial Behavior**: High-value orders are strongly correlated with installment-based payments, suggesting a need for expanded credit options.
4. **Geographic Bottlenecks**: Southern regions experience average delays of 3 days compared to the national average, highlighting the need for regional distribution hubs.
5. **Category Focus**: Home and Electronics categories show the highest margins and benefit most from logistics optimization.

---

6. Recommendations

Area	Recommended Action	Expected Impact
Logistics	Target average delivery time of ≤ 10 days.	Maintain customer ratings ≥ 4.2.
Operations	Establish regional hubs in Southern Brazil.	Reduce regional delivery delays by 15-20%.
Finance	Expand EMI/Installment offers for orders > R$ 2,000.	Potential 8-10% boost in high-ticket revenue.
Data Governance	Implement monthly automated consistency checks.	Sustain long-term data integrity and accuracy.

---

7. Conclusion

The “Chaos to Clarity” report underscores a fundamental truth in data science: **data quality directly fuels business growth**. Raw EDA was inconsistent and misleading, but post-cleaning analysis aligned perfectly with business logic. By shifting analysis accuracy from almost 70% to 94%, Olist can now make strategic, data-driven decisions with confidence.

---

8. Links

1. **Olist Public Dataset**: Olist. (2018). *Brazilian E-Commerce Public Dataset by Olist*. Retrieved from Kaggle.
2. **Data Source Context**: The dataset includes 115k orders from 2016 to 2018 made at multiple marketplaces in Brazil.
3. **More Details:** For more insights and details of both cleaned and un-cleaned data , you can visit the repo files here –

- Visualization on uncleaned data- https://github.com/sagar-data-ai/data-analysis-projects/blob/main/project_03_from_chaos_to_clarity_ecommerce_analysis/notebook/eda_uncleaned.ipynb
- Visualization on cleaned dataset- https://github.com/sagar-data-ai/data-analysis-projects/blob/main/project_03_from_chaos_to_clarity_ecommerce_analysis/notebook/eda_cleaned.ipynb