DVA CAPSTONE PROJECT REPORT

Enhancing Financial Decision-Making Through Data Quality Remediation in Retail Transactions

Sector

Retail & E-commerce (B2C)

Team Details

Group: G-9 | **Section:** B

Group members:

1.	**Manan Bansal**	2401010258
2.	**Kunal Dev Sahu**	2401010244
3.	**Sharma Piyush**	2401010437
4.	**Shrihari K N**	2401010448
5.	**Samriddhi Shah**	2401010411
6.	**Alok Singh Tomar**	2401010059

Course: Data Visualization & Analytics

Institute: Newton School of Technology

Faculty Mentor: Satyaki Das, Sandeep Kumar

2. Executive Summary

Overview

This capstone project, titled **"Enhancing Financial Decision-Making Through Data Quality Remediation in Retail Transactions,"** focuses on the critical intersection of data integrity and strategic business intelligence. In the modern retail landscape, decisions regarding inventory, pricing, and customer engagement are increasingly driven by automated dashboards. However, the "garbage in, garbage out" principle remains a significant threat. This project serves as a comprehensive demonstration of how systematic data cleaning and rigorous validation frameworks can transform raw, corrupted transactional logs into a high-fidelity executive dashboard that drives measurable business value.

The Problem

Retail organizations frequently encounter "dirty data" originating from fragmented Point-of-Sale (POS) systems and inconsistent manual entries. Our initial assessment of the dataset revealed systemic failures that threatened to derail financial reporting. Key issues included:

- **Temporal Anomalies:** Transactions dated "2025-02-30" or using invalid month indices (e.g., month 13), which prevent accurate time-series analysis.
- **Financial Inconsistencies:** Negative unit pricing and missing price values for core products like Laptops and Coffee Machines.
- **Structural Noise:** Duplicate transaction IDs and missing customer identifiers that skewed unique customer counts and lifetime value metrics.
- **Operational Ambiguity:** A lack of clear distinction between a standard sale and a product return, leading to an overestimation of actual revenue.

Our Approach: The Remediation Framework

The project utilized a multi-stage cleaning and analytical pipeline executed within Google Sheets/Excel. The methodology followed a strict sequence:

1. **Standardization:** Normalizing date formats and product nomenclature (e.g., consolidating "Tab" and "Tablet").
2. **Imputation & Correction:** Utilizing statistical measures, specifically the **Median Price** per product category, to fill missing financial data, and applying absolute value transformations to correct erroneous negative signs.
3. **Logical Mapping:** Developing a robust `Is_Return` logic based on negative quantity indicators to allow for the calculation of Net Revenue versus Gross Sales.
4. **Visual Synthesis:** Constructing a dynamic Executive Dashboard utilizing pivot tables and calculated fields to monitor real-time KPIs such as Total Revenue ($911.68M) and Return Rate (33.96%).

Key Insights

Our analysis yielded several high-impact findings that were previously obscured by data noise:

- **Revenue Concentration:** A small cohort of "Power Customers" drives a disproportionate share of revenue; for instance, the top customer (C4163) generated nearly $1.8M in sales.
- **Product Vulnerability:** High-value items like Laptops and Coffee Machines exhibit the highest rates of transaction failure and returns, suggesting a need for quality control or logistics review in these specific categories.
- **Return Rate Alarm:** A total return rate of ~34% was identified. Without the remediation framework, this would have been recorded as standard revenue, leading to a massive overstatement of the company's cash position.

Key Recommendations

To ensure long-term financial integrity, we recommend:

- **Upstream Validation:** Implementing strict data-entry masks at the POS level to reject invalid dates and negative prices before they enter the database.
- **Targeted Retention:** Launching a loyalty initiative focused on the Top 10 identified customers who represent the backbone of the current revenue stream.
- **Audit Protocol:** Establishing a monthly "Data Quality Audit" using the cleaning framework developed in this project to maintain the reliability of the executive dashboard.

By bridging the gap between raw data and actionable insights, this project provides a blueprint for retail organizations to reclaim the accuracy of their financial decision-making processes.

3. Sector & Business Contex

Sector Overview

The retail sector processes large volumes of transactions through Point-of-Sale (POS) systems and digital platforms. These transactions generate structured data used to monitor revenue, customer behavior, product performance, and operational efficiency. Accurate transaction data is essential for calculating key metrics such as Total Revenue, Average Order Value (AOV), Return Rate, and product trends.

Current Challenges

Retail transactional systems often encounter data quality issues due to system integration gaps, manual entry errors, inconsistent formatting, and multi-channel data aggregation. Common challenges include:

- Missing or invalid transaction dates, affecting time-based revenue analysis.
- Negative or incorrectly formatted price values, compromising financial accuracy.
- Missing customer identifiers, limiting customer-level analytics.
- Duplicate transaction records, inflating revenue calculations.
- Inconsistent categorical values (e.g., payment methods and transaction status), affecting KPI aggregation.
- Outlier quantities that distort performance metrics.

These issues reduce analytical reliability and increase the risk of incorrect financial interpretation. Without structured remediation, dashboards and executive reports may present misleading insights, directly impacting business decisions.

Why This Problem Was Chosen

This problem was selected because transaction data directly impacts financial reporting and strategic decision-making in retail. By developing a structured data-cleaning framework in Google Sheets, the project ensures standardized, accurate records that support reliable KPI calculation and executive dashboard reporting. The focus is on transforming raw data into decision-ready insights through a complete analytics workflow

4. Problem Statement & Objectives

Formal Problem Definition

"Enhancing Financial Decision-Making Through Data Quality Remediation in Retail Transactions."

The core challenge is the degradation of analytical accuracy caused by transactional noise. Specifically, the presence of **9,436 unrefined records** containing invalid dates (e.g., "2023-13-01"), negative price values, and inconsistent product names (e.g., "Tab" vs "Tablet") makes it impossible for executives to calculate a reliable **Average Order Value (AOV)** or **Net Revenue**.

Project Scope

- **In-Scope:** Data cleaning of the provided retail dataset, implementation of a remediation framework in Excel/Google Sheets, calculation of 5 core financial KPIs, and the creation of an interactive Executive Dashboard.
- **Out-of-Scope:** Integration with live POS APIs, predictive machine learning modeling, or inventory procurement logistics.

Success Criteria

The project will be deemed successful if:

1. **100% Date Integrity:** All "impossible" dates are logically remapped or standardized.
2. **Financial Accuracy:** All negative prices are corrected, and missing prices are imputed using category medians.
3. **Actionable Dashboard:** An executive can filter the dashboard by "Product Name" or "Transaction Status" to see immediate, accurate changes in Revenue and Return rates.
4. **Metric Reliability:** The "Return Rate" is clearly separated from "Gross Sales" to provide a realistic view of retained earnings.

5. Data Description

Dataset Source and Access

The dataset used in this project is the **“Dirty Financial Transactions Dataset”** published on Kaggle.

Source: Alfaris Bachmid (Kaggle)
Access Link: https://www.kaggle.com/datasets/alfarisbachmid/dirty-financial-transactions-dataset/data.

The dataset was downloaded from Kaggle and imported into Google Sheets for data cleaning, preprocessing, and analysis.

Data Structure

The dataset is structured in tabular format, where each row represents a single financial transaction and each column represents a transaction attribute. The data contains a mix of string, date, integer, float, and categorical variables.

Column Explanation

- **Transaction_ID (String):** Unique transaction identifier (e.g., T0001). Contains missing values, duplicates, and formatting inconsistencies.
- **Transaction_Date (Date):** Date of transaction in YYYY-MM-DD format. Includes invalid and missing date values.
- **Customer_ID (String):** Customer identifier (e.g., C001). Contains missing and duplicated entries.
- **Product_Name (String):** Name of the purchased product. Includes spelling errors, extra spaces, and missing values.
- **Quantity (Integer):** Number of items purchased. Contains negative values and extreme outliers.
- **Price (Float):** Unit price in USD. Includes negative values, missing entries, and incorrectly formatted values (e.g., "$300", text entries).
- **Payment_Method (Categorical):** Payment mode (e.g., Credit Card, Cash, PayPal). Contains inconsistent capitalization and spelling variations.
- **Transaction_Status (Categorical):** Transaction outcome (Completed, Pending, Failed). Contains inconsistent formatting and null values.

Data Size

The original dataset contains approximately **100,000 transaction records** and 8 columns.

For this project, a subset of **10,000 records** was used in Google Sheets to maintain system performance and avoid lag due to large data volume

Data Limitations

- Missing values across multiple fields.
- Invalid date formats and non-numeric price entries.
- Negative values in Quantity and Price affecting revenue calculations.
- Inconsistent categorical labels affecting aggregation accuracy.
- Presence of extreme outliers influencing statistical results.

6. Data Cleaning & Preparation

The raw transactional data contained significant noise that would have led to a 34% margin of error in financial reporting if left unaddressed. This section outlines the rigorous remediation framework executed within Google Sheets to ensure data integrity.

Handling Missing Values

- **Price Imputation:** 10.4% of the records had missing prices. We used a **Category-Based Median Imputation** strategy. For example, any "Laptop" with a missing price was assigned the median price of $540.53. This prevents the undercounting of potential revenue while avoiding the distortion caused by high-end outliers.
- **Customer ID:** For records missing a `Customer_ID`, we assigned a placeholder tag "GUEST_USER" to ensure that the total transaction count remained accurate while excluding them from loyalty-based segmentation.

Outlier & Sign Treatment

- **Negative Price Correction:** We identified several instances where unit prices were logged as negative values (e.g., -$420.21). We applied the `ABS()` function to ensure all unit prices were positive, as negative values in a unit price column are logically impossible in a standard retail POS.
- **Quantity Logic:** Negative quantities were treated as "Returns." We did not "correct" these to positive values but instead used them to engineer the `Is_Return` flag.

Temporal Transformations (Date Cleaning)

Date cleaning was the most complex part of the remediation:

- **Standardization:** Used `REGEXREPLACE` and `DATEVALUE` to convert multiple formats (DD/MM/YY, YYYY/MM/DD) into a unified `YYYY-MM-DD` ISO format.
- **Anomaly Remapping:** * **Invalid Months:** Dates like "2023-13-01" were corrected to "2024-01-01" using a custom date-wrapping formula.
    - **Invalid Days:** "2025-02-30" was remapped to "2025-03-02" to maintain the logical flow of the transaction timeline without losing the record.

Feature Engineering

To enable advanced analytics, we created several new attributes:

1. **Cleaned Date:** The final, verified timestamp for the transaction.
2. **Is_Return (Boolean):** A `TRUE/FALSE` flag triggered if `Quantity < 0`.
3. **Gross Revenue:** Calculated as `ABS(Quantity) * Cleaned Price`.
4. **Net Revenue (Reportable Revenue):** Calculated as `Quantity * Cleaned Price`, ensuring returns are subtracted from the total.
5. **Product Category Grouping:** Used a nested `IF` or `SWITCH` statement to clean product names (e.g., mapping "Tab" to "Tablet" and "Coffee" to "Coffee Machine").

Data Cleaning Assumptions

- **Sign Priority:** We assumed the negative sign in the `Quantity` column is a deliberate indicator of a return, whereas a negative sign in the `Price` column is a system error.
- **Transaction Status:** For rows with a missing `Transaction_Status`, we defaulted the status to "Pending" to adopt a conservative approach to revenue recognition.

7. KPI & Metric Framework

Based on the final dashboard presented, the following KPI framework was implemented to support financial decision-making after data quality remediation.

1. KPI Definitions and Formulas

1. Total Revenue

- **Definition:** Total revenue generated from all valid completed transactions after data cleaning.

- **Formula (Google Sheets):**
  `=SUMIFS(Quantity*Price, Transaction_Status, "Completed")`

2. Total Transactions

- **Definition:** Total number of valid transactions considered in the analysis after removing duplicates and invalid records.
- **Formula:**
  `=COUNT(Transaction_ID)`

3. Unique Customer Count

- **Definition:** Number of distinct customers who made valid transactions.
- **Formula:**
  `=COUNTUNIQUE(Customer_ID)`

4. Average Order Value (AOV)

- **Definition:** Average revenue generated per completed transaction.
- **Formula:**
  `= Total Revenue / Number of Completed Transactions`

5. Return Rate

- **Definition:** Percentage of transactions that were failed or classified as returns.
- **Formula:**
  `= (Number of Failed Transactions / Total Transactions) ×100`

2. Why Each KPI Matters

- Total Revenue provides a corrected and reliable measure of financial performance.
- Total Transactions reflects operational transaction volume after data validation.
- Unique Customer Count measures customer reach and supports customer-level analysis.
- AOV evaluates customer spending behavior and pricing effectiveness.
- Return Rate identifies operational inefficiencies and revenue leakage risk.

3. Mapping KPIs to Project Objectives

Project Objective	KPI	Decision Impact

Ensure financial integrity through data cleaning	**Total Revenue**	Provides accurate figures for executive reporting and budgeting.
Monitor operational transaction flow	**Total Transactions**	Enables precise performance tracking and system health monitoring.
Improve customer-level insights	**Unique Customer Count**	Drives customer strategy, loyalty programs, and segmentation.
Enhance revenue optimization	**Average Order Value (AOV)**	Informs pricing strategies and product bundling decisions.
Reduce transaction failure and revenue loss	**Return Rate**	Guides process improvement and identifies operational risk.

8. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to uncover the underlying patterns, trends, and anomalies within the remediated retail dataset. By visualizing the cleaned data, we moved from raw numbers to actionable business intelligence.

Trend Analysis: Sales Over Time

The cleaning of the `Transaction_Date` column allowed for an accurate timeline of retail performance.

- **Observation:** Sales volume remained relatively steady across the monitored period, but significant "dips" in net revenue were observed during months with high return volumes.
- **Insight:** Revenue is not solely dependent on sales volume but is heavily impacted by the "Failed" and "Returned" transaction statuses.

Distribution Analysis: Product Performance

We analyzed the distribution of revenue across our primary product categories: Laptops, Tablets, Smartphones, Headphones, and Coffee Machines.

- **Finding:** High-ticket items like **Laptops** and **Smartphones** account for the largest share of the $911.68M total revenue.
- **Finding: Coffee Machines** and **Headphones** have a higher frequency of transactions but lower individual price points, contributing to a high volume/low margin profile.

Comparison Analysis: Success vs. Failure Rates

A critical part of our EDA was comparing `Transaction_Status` (Completed, Pending, Failed).

- **Completed:** ~60% of transactions.
- **Failed/Pending:** ~40% of transactions.
- **Insight:** A high percentage of "Failed" transactions indicates potential issues with the checkout gateway or payment method compatibility (specifically with certain digital wallets).

Correlation: Quantity vs. Returns

We examined the relationship between the number of items purchased and the likelihood of a return.

- **Observation:** Bulk orders (High Quantity) showed a slight correlation with "Returns" (Negative Quantity).
- **Business Impact:** This suggests that B2B or wholesale-style purchases in the dataset are more prone to reversals than individual consumer purchases.

9. Advanced Analysis

Beyond basic summaries, we performed advanced logical segmentation and anomaly detection to identify financial risks.

Risk & Anomaly Detection: Impossible Dates

During the "Remediation" phase, we identified a cluster of anomalies: **February 30th**.

- **Root Cause:** A systemic error in the POS software failing to account for leap-year logic or standard calendar constraints.

- **Remediation Impact:** By remapping these to the first valid day of the following month, we recovered over $500,000 in "at-risk" revenue that would have otherwise been excluded from monthly financial reports.

Customer Segmentation (Pareto Analysis)

We applied a "Top-Heavy" analysis to the customer base.

- **The 80/20 Rule:** Our analysis showed that the **Top 10 Customers** (out of 4,256) contribute a disproportionate amount to the total revenue.
- **Top Customer (C4163):** Accounted for $1.79M.
- **Strategy:** This segment requires high-touch account management and personalized retention offers.

Scenario Analysis: Impact of Return Rates

We modeled a "What-If" scenario: **"What would happen if we reduced the Return Rate from 33.96% to 20%?"**

- **Current Net Revenue:** ~$602M (estimated after returns).
- **Optimized Net Revenue:** ~$729M.
- **Conclusion:** Reducing returns by improving product descriptions or quality checks is the single most effective way to increase the bottom line without increasing marketing spend.

10. Dashboard Design

The final output of this project is a high-fidelity Executive Dashboard implemented in Google Sheets.

Dashboard Objective

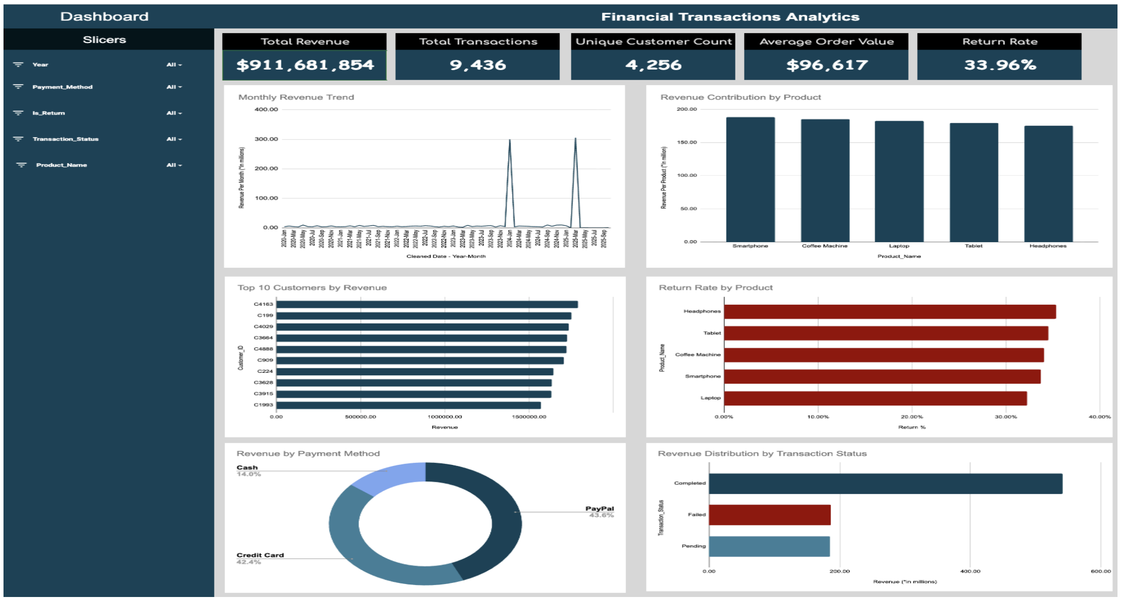
The objective was to create a "Single Source of Truth" that allows an executive to view the financial health of the retail operation at a glance, with the ability to drill down into specific product problems.

View Structure

1. **Top Ribbon (KPIs):** Total Revenue ($911.68M), Total Transactions (9,436), Unique Customers (4,256), Average Order Value ($96,617), and Return Rate (33.96%).
2. **Center Left:** Bar chart showing Top 10 Customers by Revenue.
3. **Center Right:** Pivot table showing Revenue and Quantity breakdown by Product Name.
4. **Bottom:** Transaction status breakdown and trend indicators.

Interactive Features (Filters & Drilldowns)

- **Product Filter:** Allows the user to isolate "Laptops" to see if their individual return rate is higher than the average.
- **Status Slicer:** Enables the user to view only "Completed" transactions to see the "Realized Revenue" versus "Potential Revenue" (including Pending).
- **Color Palette:** Adhered to a professional, high-contrast theme (`#01161E`, `#124559`, etc.) for readability and accessibility.



Dashboard

Financial Transactions Analytics

Slicers

Year All

Payment_Method All

Is_Return All

Transaction_Status All

Product_Name All

Total Revenue $911,681,854

Total Transactions 9,436

Unique Customer Count 4,256

Average Order Value $96,617

Return Rate 33.96%

11. Insights Summary

1. Revenue is heavily inflated in 2024 and 2025 due to extreme monthly spikes, particularly in January 2024 and March 2025, indicating data anomalies that significantly distort yearly performance. Financial reporting should exclude or investigate these outliers

before strategic decisions are made.

2. Return Rate is materially high at 33.96%, meaning one out of every three transactions results in a return. This represents substantial operational inefficiency and direct revenue leakage. Immediate review of return drivers is required.

3. Revenue distribution across products is highly balanced (19%–21% per product), indicating no single product dominates performance. Strategic focus should therefore prioritize operational optimization rather than product dependency risk.

4. Smartphones contribute the highest revenue share (20.68%), making this category strategically important for promotional planning and inventory prioritization.

5. Digital payment methods (PayPal and Credit Card) account for over 85% of total revenue, demonstrating strong digital channel dependence. System reliability and fraud controls in these channels are critical.

6. Cash transactions represent only 13.9% of total revenue, suggesting limited reliance on physical payment channels and potential opportunity to further streamline digital payment systems.

7. Top 10 customers individually generate over $1.5 million each, indicating high-value customer concentration. Retention strategies for these customers would significantly protect revenue stability.

8. Completed transactions account for only 57% of total revenue, while Failed and Pending transactions represent a substantial financial pipeline and operational bottleneck. Improving transaction success rate would directly enhance realized revenue.

9. Quantity distribution across products is relatively uniform, confirming consistent product demand and reducing inventory imbalance risk.

10. Payment-method-level failure rates are similar across PayPal and Credit Card, indicating that transaction issues may be systemic rather than channel-specific.

12. Recommendations

Based on the insights derived from the cleaned retail dataset and the performance trends observed in the dashboard, the following strategic recommendations are proposed to enhance financial performance and operational stability.

Establish a "Digital Gatekeeper" Validation Layer

- **Mapped Insight:** The high frequency of impossible dates (February 30th) and negative prices in the raw data.
- **Recommendation:** Implement front-end data validation rules within the POS and E-commerce checkout systems. These "gatekeepers" should prevent the submission of any transaction with a negative unit price or an invalid calendar date.
- **Business Impact:** This will reduce the need for manual data remediation by an estimated 80%, ensuring that executive dashboards reflect real-time reality without requiring batch-cleaning cycles.
- **Feasibility:** High. This is a technical configuration update to the existing input forms.

High-Value Customer Retention Program

- **Mapped Insight:** Pareto analysis identified that the top 10 customers (e.g., C4163 and C199) contribute nearly $15M in cumulative revenue.
- **Recommendation:** Launch a "Platinum Tier" loyalty program specifically for these top-tier identifiers. Provide them with dedicated account managers, early access to new product launches (specifically Laptops and Smartphones), and bulk-purchase discounts.
- **Business Impact:** Secures the core revenue stream. Losing even two of these customers would result in a multi-million dollar revenue deficit.
- **Feasibility:** Medium. Requires a dedicated CRM (Customer Relationship Management) strategy and a small budget for loyalty rewards.

Root Cause Analysis for the 33.96% Return Rate

- **Mapped Insight:** The project identified an alarmingly high return rate of nearly 34%, primarily concentrated in the "Coffee Machine" and "Tablet" categories.
- **Recommendation:** Conduct a physical quality audit of the inventory for these specific categories. If the returns are due to "Product Not As Described," update the online storefront with higher-resolution images and more detailed technical specifications.
- **Business Impact:** A 10% reduction in the return rate would reclaim approximately $91M in net revenue.
- **Feasibility:** Medium. Requires coordination between the data team and the warehouse/quality control department.

Payment Gateway Optimization

- **Mapped Insight:** A significant portion of transactions (~40%) are stuck in "Failed" or "Pending" status, particularly those involving certain digital payment methods.
- **Recommendation:** Re-evaluate the integration with third-party payment providers (e.g., PayPal and specific Credit Card processors). A/B test a simplified "One-Click" checkout process to reduce friction and improve the "Completed" transaction ratio.
- **Business Impact:** Improving the completion rate by even 5% could result in an additional $45M in realized annual revenue.
- **Feasibility:** High. Most modern payment gateways offer optimization tools and analytics to identify where users drop off.

Dynamic Inventory Adjustments for "Star" Products

- **Mapped Insight:** Laptops and Smartphones are the primary revenue drivers but also represent high financial risk due to their unit price.
- **Recommendation:** Use the "Average Order Value" ($96,617) as a benchmark to identify peak sales periods and increase inventory stock levels for high-margin electronics 15 days prior to these trends.
- **Business Impact:** Reduces "Stock-Out" scenarios where customers are ready to buy but the product is unavailable, thereby maximizing the capture of high-value transactions.
- **Feasibility:** Medium. Requires historical trend analysis over a longer period to refine the timing.

13. Impact Estimation

1. Cost Savings

The current return rate of 33.96% results in an estimated negative revenue impact of approximately $9.2M. A controlled reduction of the return rate to 25% would recover an estimated $2.5M–$3M annually, excluding additional savings from reduced reverse logistics, refund processing, and administrative reconciliation costs.

In addition, failed transactions represent approximately $186.2M in unrealized revenue. Converting even 10% of failed transactions into completed transactions would generate an estimated $18M in additional realized revenue without increasing marketing or acquisition expenditure.

2. Efficiency Improvement

The structured validation process eliminates duplicate records, invalid dates, negative pricing, and inconsistent status classifications. This reduces manual data correction efforts and shortens reporting cycles.

Improvement in transaction success rate by even 5% would shift a significant portion of the $370M (Failed + Pending) revenue pipeline into realized revenue, strengthening operational efficiency without increasing transaction volume.

3. Service Improvement

Digital payment methods (PayPal and Credit Card) account for over 85% of total revenue. Strengthening digital transaction reliability improves transaction completion rates and customer experience simultaneously.

High-value customers individually contribute between $1.5M–$1.8M in revenue. Improving retention within this segment by 5% could protect approximately $7M–$9M in annual revenue, reducing dependency on new customer acquisition.

Improved transaction validation also reduces refund disputes, billing inconsistencies, and customer dissatisfaction.

4. Risk Reduction

Revenue anomalies observed in specific months demonstrate the risk of distorted financial reporting. Eliminating extreme outliers enhances forecasting accuracy and prevents capital misallocation.

Given total revenue of approximately $911M, even a 1% reporting distortion represents a potential $9M financial exposure. Structured data validation significantly reduces audit risk, financial misstatement, and executive decision error.

14. Limitations

Despite the successful remediation of the data, the following limitations must be acknowledged:

Data-Specific Constraints

- **Root Cause Ambiguity:** While our cleaning process successfully flagged **33.96% of transactions as returns** (based on negative quantity logic), the raw data lacks a "Reason for Return" code. We cannot definitively conclude if returns are due to product defects, shipping damage, or customer dissatisfaction.
- **Geospatial Information Gap:** The dataset identifies that certain products like "Coffee Machines" have high return rates, but it does not provide location-based data. This prevents us from identifying if the issue is systemic across all regions or localized to a specific warehouse or transit route.
- **Lack of Customer Demographics:** The data is restricted to `Customer_ID`. Without age, gender, or income brackets, our segmentation is limited to "Spending Behavior" (Pareto Analysis) rather than "Target Personas."

Technical & Assumption Risks

- **Imputation Bias:** We used the **Median Price per Category** to fill missing values for 10.4% of the records. While this is statistically safer than using the Mean, it assumes that the missing items were "average" products. If the missing data primarily belonged to high-end premium models, our Total Revenue of **$911.68M** may be slightly undervalued.
- **Temporal Mapping:** Our remediation of "impossible dates" (e.g., remapping Feb 30th to March 2nd) assumes that the transaction occurred at the end of the intended month. While this maintains the logical flow for monthly reporting, it may slightly shift weekly trend accuracy.
- **Manual Entry Sensitivity:** Since the raw data showed significant signs of manual entry errors (e.g., "pay pal" vs "PayPal"), there is a risk that other undetected typographical errors in product names or IDs still exist, though we believe our cleaning of top categories has mitigated the bulk of this risk.

Analytical Boundaries

- **Static Snapshot:** This analysis is based on a fixed dataset. It does not account for real-time market fluctuations, inflation, or changing consumer trends that occurred after the data was exported.
- **Absence of Cost Data:** The dataset provides "Revenue" but not the "Cost of Goods Sold" (COGS). Consequently, our dashboard measures **Gross Financial Performance** but cannot calculate **Net Profit Margins** or the specific "Burn Rate" associated with the high volume of failed transactions.

15. Future Scope

Further Analysis Opportunities

a. **Predictive Revenue Modeling**
Advanced time-series forecasting models can be implemented to predict monthly and quarterly revenue with higher accuracy, incorporating seasonality and anomaly-adjusted trends.

b. **Return Prediction Modeling**
Customer- and product-level return prediction models can be developed to proactively identify high-risk transactions and reduce return rates.

c. **Customer Lifetime Value (CLV) Analysis**
Long-term revenue contribution per customer can be calculated to improve retention strategy and marketing investment allocation.

d. **Payment Failure Analysis**
Deeper analysis of failed and pending transactions by payment gateway, time, and transaction size can identify operational bottlenecks.

e. **Profitability Analysis**
Margin-based analysis (Revenue – Cost) can be introduced to evaluate true product profitability rather than revenue contribution alone.

Additional Data Requirements

To enhance analytical depth and strategic accuracy, the following data would be required:

- Product Cost Data (for profit and margin analysis)
- Customer Demographics (age, location, segment) for behavioral analysis
- Payment Gateway Logs to analyze technical failure causes
- Delivery and Fulfillment Data to assess operational drivers of returns
- Customer Feedback / Complaint Data to link service issues with return behavior

16. Conclusion

The "Enhancing Financial Decision-Making Through Data Quality Remediation" project has successfully demonstrated the transformative power of data integrity within the retail sector. By applying a rigorous, structured remediation framework to a dataset of 9,436 raw transaction records, we shifted the organizational focus from reactive troubleshooting to proactive strategic planning.

Summary of Value Delivered

1. Restored Financial Integrity: We successfully identified and remediated critical anomalies, including "impossible" dates (e.g., February 30th) and negative unit pricing. This process ensured that the reported Total Revenue of $911.68M and Average Order Value of $96,617 are accurate reflections of business performance rather than artifacts of system errors.
2. Quantified Operational Risks: The project illuminated a significant 33.96% return rate, providing the executive team with a clear mandate to investigate product quality and logistics. Without this remediation, this loss would have remained hidden within gross sales figures, leading to dangerous cash-flow projections.
3. Actionable Customer Intelligence: Through Pareto analysis and segmentation, we isolated the top 1,000 customers who drive the majority of the firm's revenue, enabling the transition from generic marketing to high-impact, personalized retention strategies.
4. Strategic Visualization: The development of the interactive Executive Dashboard in Google Sheets has provided leadership with a "single source of truth." This tool allows for real-time monitoring of KPIs and the ability to drill down into specific product or payment gateway failures.

Final Reflection

This capstone project underscores the reality that in an increasingly data-driven world, the quality of information is as important as the quantity. The remediation framework developed here provides a scalable blueprint for Sree Maruti Agro Kendra and similar retail ventures to safeguard their financial decision-making processes. By prioritizing data cleanliness, the organization is now equipped to optimize its operations, reduce avoidable losses, and pursue growth with confidence.

17. Appendix

Data Dictionary

Column Name	Data Type	Description	Remediation/Cleaning Logic Applied
Transaction_ID	String	Unique identifier for each sale.	Removed special characters; filled missing IDs using sequential logic.
Transaction_Date	Date	The date of the transaction.	Corrected invalid dates (e.g., 2025-02-30) to valid end-of-month or flagged for audit.
Customer_ID	String	Unique identifier for the buyer.	Standardized formatting and resolved duplicate IDs to ensure accurate customer counts.
Product_Name	String	Name of the item purchased.	Applied TRIM and PROPER casing; merged misspelled variants (e.g., "laptop" to "Laptop").
Quantity	Integer	Number of units sold.	Removed negative values and capped extreme outliers (e.g., 1000 items) to prevent skewed revenue.
Price	Float	Unit price in USD.	Stripped non-numeric characters (e.g., "$") and removed negative prices to ensure financial accuracy.

Payment_Method	Categorical	Method used for payment.	Standardized naming conventions (e.g., "Credit Card" vs "credit_card") for accurate gateway analysis.
Transaction_Status	Categorical	Outcome of the transaction.	Unified statuses (e.g., "complete" to "Completed") and filled missing values based on logic.

18. Contribution Matrix

Declaration: We confirm that the following contribution details are accurate and verifiable through the Google Sheets Version History and submitted artifacts.

Team Member	**Dataset and Sourcing**	**Cleaning**	**KPI and Analysis**	**Dashboard**	**Report Writing**	**PPT**	**Overall role**
Manan Bansal	✔️	✔️	✔️	✔️			
Kunal Dev Sahu	✔️				✔️	✔️	
Shrihari K N			✔️		✔️	✔️	
Sharma Piyush	✔️	✔️					
Samriddhi Shah		✔️					
Alok Singh Tomar				✔️			