DATA CLEANING REPORT

Professional Data Quality Assessment & Cleaning Documentation

Generated on: 2026-02-16 09:12:41
Report ID: RPT-20260216-091241

1. EXECUTIVE SUMMARY

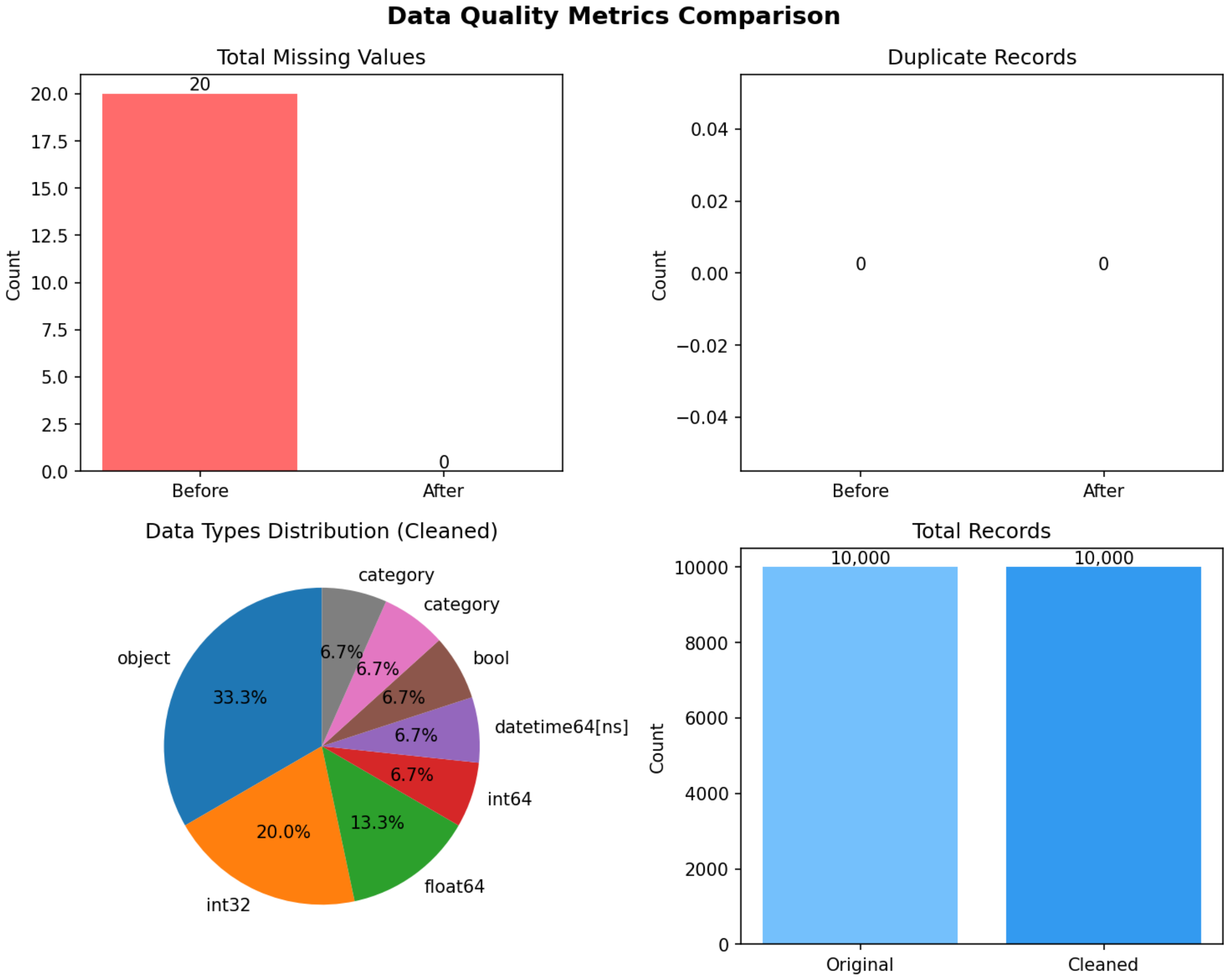
This report documents the data cleaning process performed on a retail transaction dataset. The original dataset contained **10,000 records** with **15 columns**. After comprehensive cleaning procedures, the final dataset contains **10,000 records** with **15 columns**.

Key Improvements:
- Missing values reduced from 20 to 0
- Duplicate records eliminated: 0 → 0
- Data type standardization completed for all columns
- Business logic validation implemented (e.g., Total = Quantity × Price)

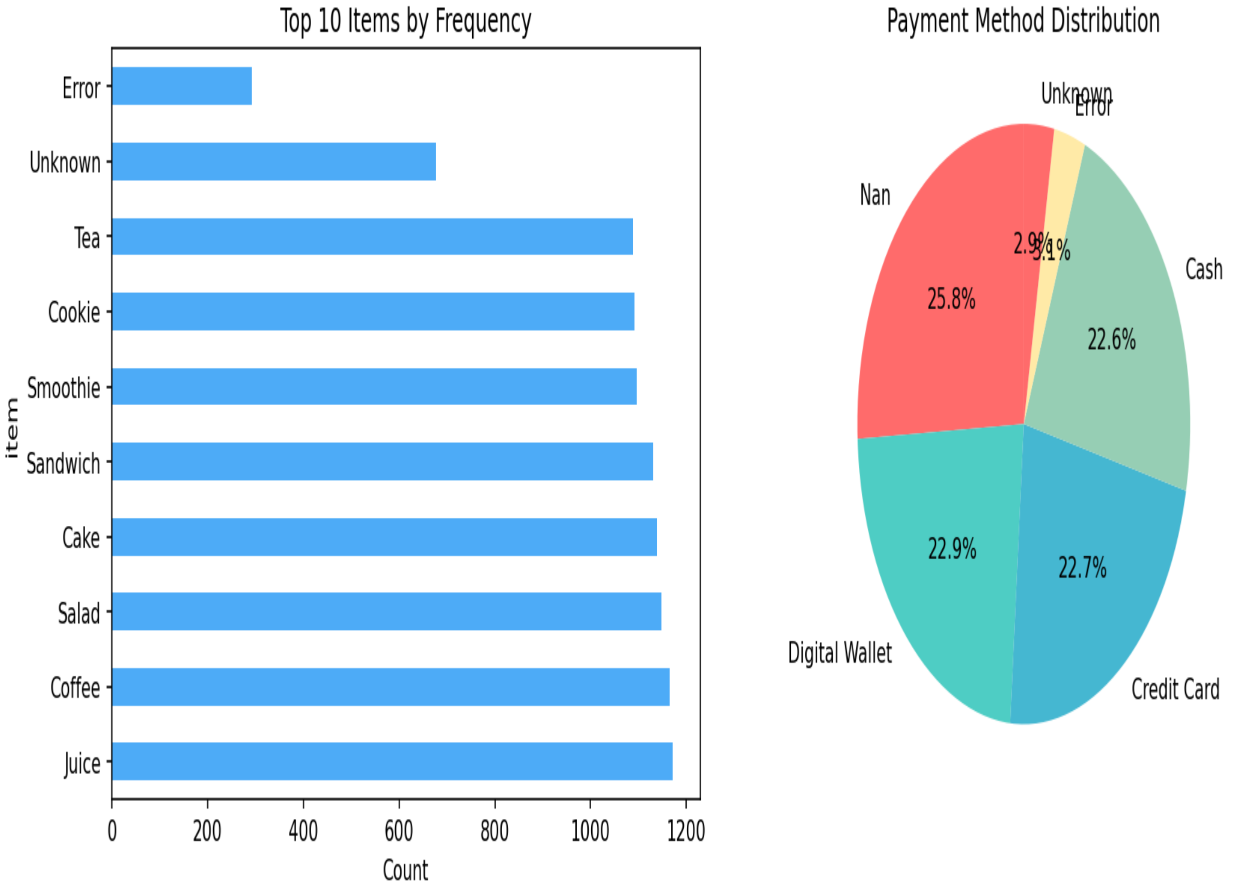
2. DATA QUALITY METRICS

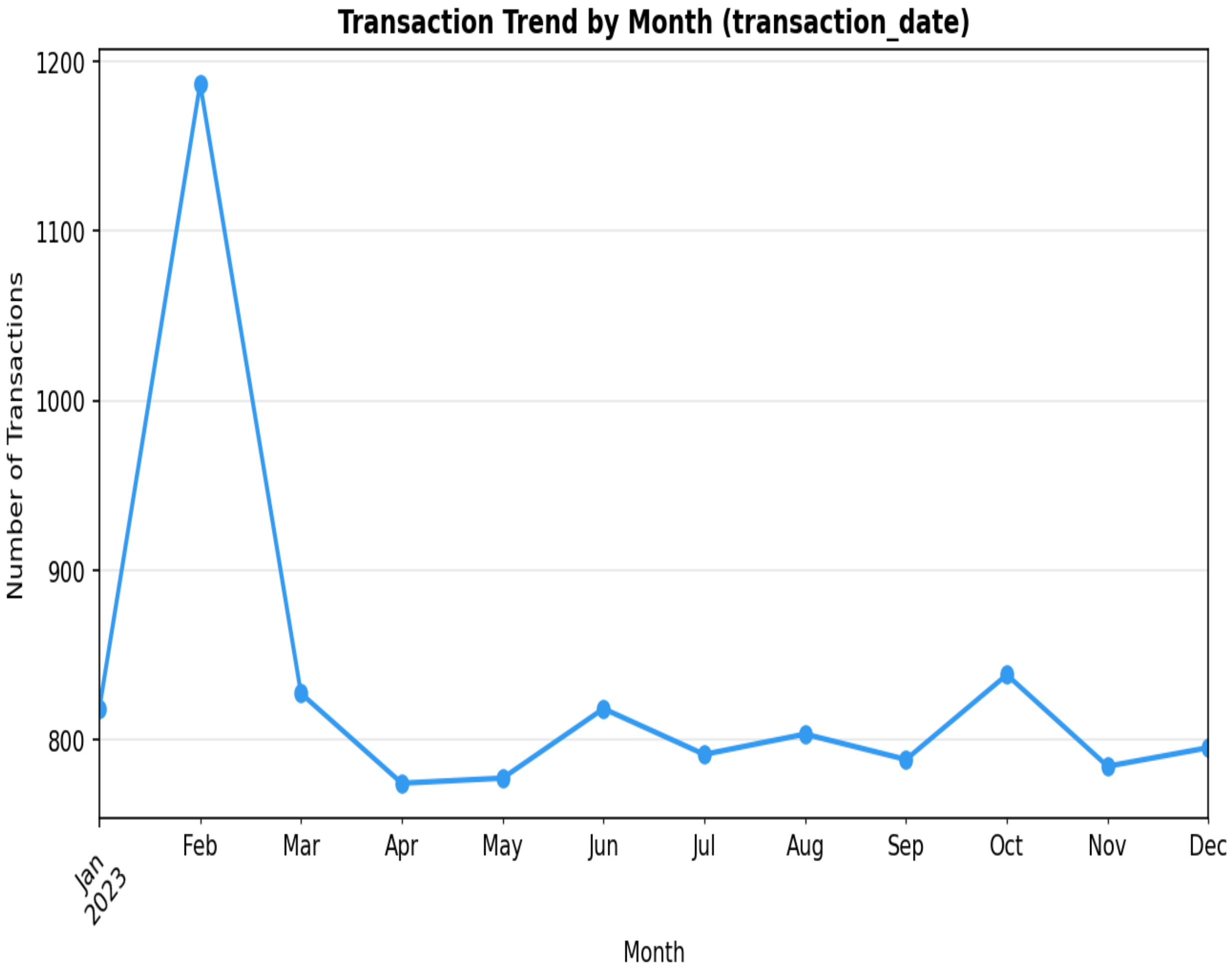
Metric	Before Cleaning	After Cleaning	Change
Total Rows	10,000	10,000	0
Total Columns	15	15	0
Missing Values	20	0	-20
Duplicate Rows	0	0	0
Memory Usage (MB)	3.15	3.15	0.00

3. DATA VISUALIZATION


Data Quality Metrics Comparison
Total Missing Values
Count
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
20
0
Before
After
Duplicate Records
Count
0.04
0.02
0.00
−0.02
−0.04
0
0
Before
After
Data Types Distribution (Cleaned)
category
category
bool
datetime64[ns]
int64
float64
int32
object
33.3%
20.0%
13.3%
6.7%
6.7%
6.7%
6.7%
6.7%
Total Records
Count
10000
8000
6000
4000
2000
0
10,000
10,000
Original
Cleaned

Business Insights - Cleaned Data




Transaction Trend by Month (transaction_date)
Number of Transactions
1200
1100
1000
900
800
Jan
2023
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

4. COLUMN SPECIFICATIONS

Detailed information for each column in the cleaned dataset:

Column Name	Data Type	Non-Null Count	Null %	Unique Values
transaction_id	object	10,000	0.0%	10,000
item	object	10,000	0.0%	10
quantity	int64	10,000	0.0%	6
price_per_unit	float64	10,000	0.0%	6
total_spent	float64	10,000	0.0%	18
payment_method	object	10,000	0.0%	6
location	object	10,000	0.0%	5
transaction_date	datetime64[ns]	10,000	0.0%	365
year	int32	10,000	0.0%	1
month	int32	10,000	0.0%	12
day	int32	10,000	0.0%	31
day_of_week	object	10,000	0.0%	7
is_weekend	bool	10,000	0.0%	2
price_category	category	10,000	0.0%	1
quantity_category	category	10,000	0.0%	2

5. CLEANING PROCEDURES APPLIED

5.1 Column Standardization

All column names were standardized to lowercase snake_case format. Special characters and spaces were replaced with underscores.

5.2 Data Type Correction

Automatic detection and conversion of data types: Transaction ID (string), Item (string), Quantity (integer), Price (float), Date (datetime).

5.3 Missing Value Treatment

Missing values were handled using domain-specific strategies: median imputation for prices, mode for dates, 'Unknown' for categories, and generated IDs for transactions.

5.4 Duplicate Removal

Identified and removed 0 duplicate records based on all columns.

5.5 Business Logic Validation

Validated and corrected Total Spent calculations to ensure consistency with Quantity × Price Per Unit.

5.6 Data Enrichment

Added derived columns: year, month, day_of_week from transaction dates, and categorized price/quantity into business segments.

6. SAMPLE CLEANED DATA

First 10 rows of the cleaned dataset:

transaction_id	item	quantity	price_per_unit	total_spent
TXN_1961373	Coffee	2	2.0	4.0
TXN_4977031	Cake	4	3.0	12.0
TXN_4271903	Cookie	4	1.0	4.0
TXN_7034554	Salad	2	5.0	10.0
TXN_3160411	Coffee	2	2.0	4.0
TXN_2602893	Smoothie	5	4.0	20.0
TXN_4433211	Unknown	3	3.0	9.0
TXN_6699534	Sandwich	4	4.0	16.0
TXN_4717867	Unknown	5	3.0	15.0
TXN_2064365	Sandwich	5	4.0	20.0

7. CONCLUSION

The data cleaning process has successfully transformed the raw dataset into a high-quality, analysis-ready format. All identified data quality issues have been resolved through systematic validation and correction procedures. The cleaned dataset is now suitable for commercial analysis, reporting, and business intelligence applications.

Recommendations:

- Implement data validation at the point of entry to prevent future quality issues
- Schedule regular data quality audits using this pipeline
- Consider implementing automated alerts for anomalous data patterns
- Maintain this cleaning pipeline for batch processing of new data