Data Cleaning and Sales Analysis Report: Cafe Operations 2023

Prepared by: Kaung Khant Lin, Pyae Sone Aung

Date: December 14, 2025

Abstraction

This report outlines the end-to-end data cleaning and exploratory analysis of a synthetic "Dirty Cafe Sales" dataset containing 10,000 raw transaction records. The primary objective was to remediate severe data quality issues—including missing values, text errors, and inconsistent data types—to enable accurate business reporting.

Key Findings:

- **Data Recovery:** Successfully restored 9,974 valid transactions (99.7% recovery rate) using logical imputation techniques.
- **Revenue Overview:** The cafe generated a total revenue of **$89,042** in 2023.
- **Operational Insight:** "Takeaway" orders dominate the business model, accounting for **70%** of all transactions.
- **Customer Behavior:** "Digital Wallets" are the preferred payment method (55%), indicating a tech-savvy, convenience-focused customer base.

Tables of Contents

2. Introduction

2.1 Background

Reliable sales data is essential for inventory planning and financial forecasting. The dataset provided, *dirty_cafe_sales.csv*, contained 10,000 rows of synthetic transaction data intentionally corrupted with errors such as "UNKNOWN" values, "ERROR" strings, and missing fields, mimicking real-world system failures.

The dataset was obtained from Kaggle and is publicly available at this link.

2.2 Objectives

The analysis covers sales transactions from January 1, 2023, to December 31, 2023. The project scope included:

1. **Data Cleaning:** Identifying and rectifying inconsistencies using Python.
2. **Data Restoration:** Imputing missing financial and categorical data rather than deleting it.
3. **Visualization:** Creating an interactive Tableau dashboard to present key performance indicators (KPIs).

2.3 Data Source and Licensing

This dataset was created by **Ahmed Mohamed**, and full credit is respectfully attributed to the original author. The *Dirty Cafe Sales* dataset is publicly available on Kaggle and is released under the **Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) License**.

3. Methodology

3.1 Tools & Technologies

- **Python (Pandas, NumPy):** Used for data manipulation, type conversion, and logical imputation.
- **Tableau Public:** Used for dashboard creation and visual storytelling.

3.2 Cleaning Procedure

A systematic 8-step approach was applied to the dirty dataset:

1. **Type Conversion:** Core numeric columns (Quantity, Price per unit, Total spent) were forced to numeric types, coercing errors to *NaN*.
2. **Error Handling:** Placeholder strings like "UNKNOWN" and "ERROR" were standardized to null values (*np.nan*).
3. **Logical Imputation (Financials):** Missing values were calculated using the relationship:

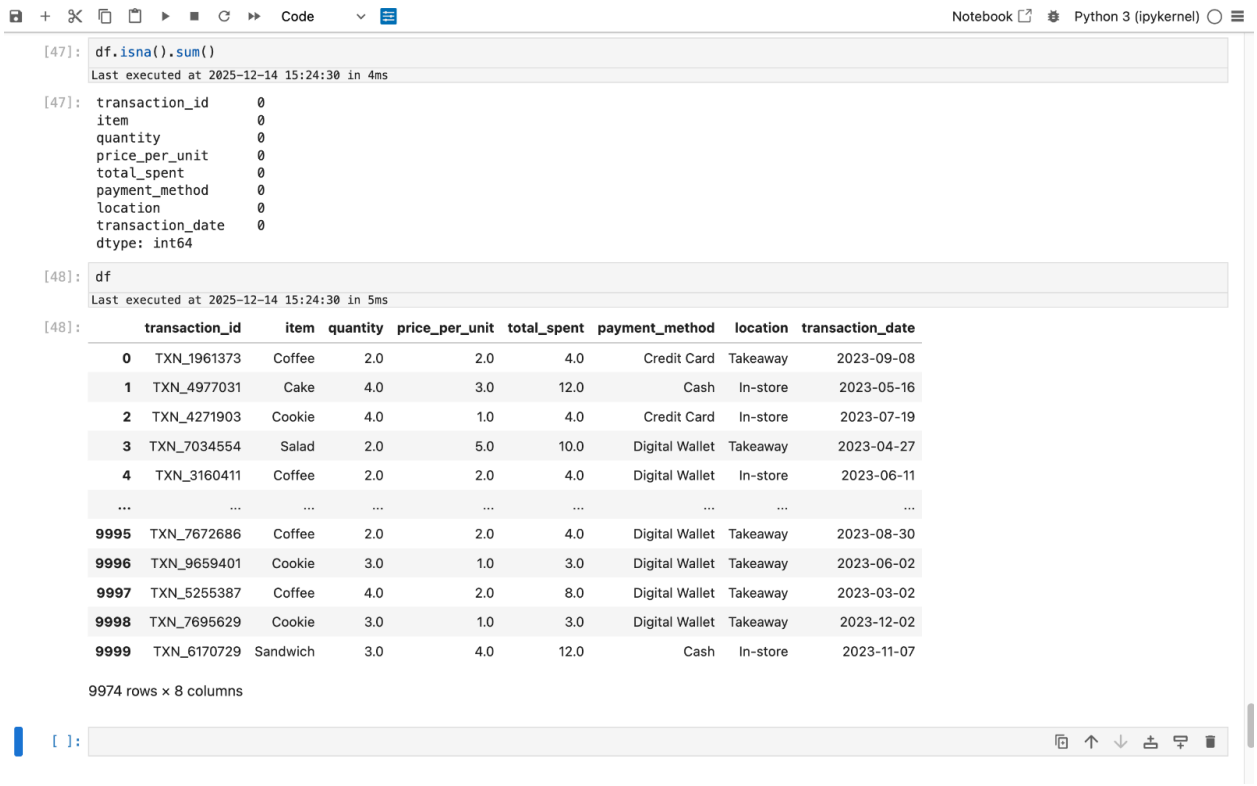
$$Total\ spent\ =\ Quantity\ *\ Price\ per\ unit$$

4. **Bidirectional Mapping:** Two reference dictionaries were created to cross-fill missing values based on available context:
    a. *price_to_item*: Mapped known prices to missing item names (e.g., $2.00 → "Coffee").
    b. *menu_prices*: Mapped known item names to missing *price_per_unit* values (e.g., "Salad" → $5.00).
5. **Irrecoverable Data Removal:** Rows with simultaneously missing Quantity and Total Spent were dropped as they could not be mathematically recovered.
6. **Date formatting:** The Transaction Date column was converted to datetime objects to allow for time-series analysis.
7. **Temporal Imputation:** Missing dates were filled using forward-fill (*ffill( )*) and backward-fill (*bfill( )*) methods to maintain timeline continuity.
8. **Categorical Imputation:** Missing Payment method and Location entries were filled using the statistical *mode* (most frequent value).

4. Data & Results

4.1 Cleaned Data Results

After completing the data cleaning process, the dataset (Fig. 1.1) was fully resolved with respect to missing values as well as invalid entries such as “UNKNOWN” and “ERROR”. A total of 26 unrecoverable records, in which both Quantity and Total Spent values were missing, were removed from the dataset. As a result, the final cleaned dataset consists of 9,974 valid transaction records, representing a complete and analysis-ready sample.



```
[47]: df.isna().sum()
Last executed at 2025-12-14 15:24:30 in 4ms

[47]: transaction_id      0
      item                0
      quantity            0
      price_per_unit      0
      total_spent         0
      payment_method      0
      location            0
      transaction_date    0
      dtype: int64

[48]: df
Last executed at 2025-12-14 15:24:30 in 5ms
```

[48]:

	transaction_id	item	quantity	price_per_unit	total_spent	payment_method	location	transaction_date
0	TXN_1961373	Coffee	2.0	2.0	4.0	Credit Card	Takeaway	2023-09-08
1	TXN_4977031	Cake	4.0	3.0	12.0	Cash	In-store	2023-05-16
2	TXN_4271903	Cookie	4.0	1.0	4.0	Credit Card	In-store	2023-07-19
3	TXN_7034554	Salad	2.0	5.0	10.0	Digital Wallet	Takeaway	2023-04-27
4	TXN_3160411	Coffee	2.0	2.0	4.0	Digital Wallet	In-store	2023-06-11
...	...	...	...	...	...	...	...	...
9995	TXN_7672686	Coffee	2.0	2.0	4.0	Digital Wallet	Takeaway	2023-08-30
9996	TXN_9659401	Cookie	3.0	1.0	3.0	Digital Wallet	Takeaway	2023-06-02
9997	TXN_5255387	Coffee	4.0	2.0	8.0	Digital Wallet	Takeaway	2023-03-02
9998	TXN_7695629	Cookie	3.0	1.0	3.0	Digital Wallet	Takeaway	2023-12-02
9999	TXN_6170729	Sandwich	3.0	4.0	12.0	Cash	In-store	2023-11-07

9974 rows × 8 columns

Fig. 1.1 Cleaned Data Results

4.2 Visual Findings

Following the completion of the data cleaning process, the cleaned dataset was used to create data visualizations. These visualizations were developed using Tableau, which enabled effective exploration and presentation of sales trends, total revenue, total transactions, average order value, best selling items, location performance and payment method preference.

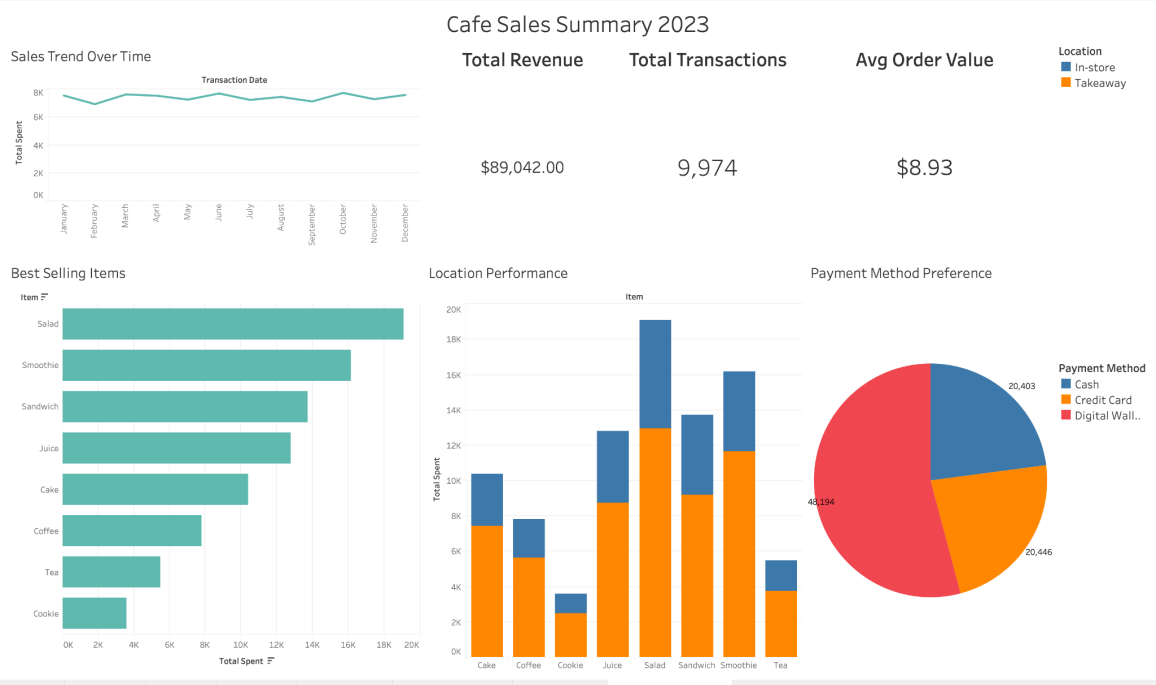


Fig. 1.2 Cafe Sales Summary 2023 dashboard

4.3 Data Visualization Explanation

The dashboard(Fig. 1.2 Cafe Sales Summary 2023 dashboard) provides a high-level overview of café performance using the cleaned dataset. Key performance indicators show a Total Revenue of $89,042.00, 9,974 total transactions, and an Average Order Value (AOV) of $8.93, reflecting stable overall sales activity after data cleaning.

Sales Trends:

Sales remained relatively consistent throughout the year, with only minor monthly fluctuations. This indicates steady customer demand and the absence of strong seasonal effects on revenue.

Category Performance:

- Salad, Smoothies, and Sandwiches are the top-performing items in terms of total revenue.
- Lower-priced items such as Cookies and Tea generated lower total revenue despite frequent purchases, suggesting high volume but limited revenue contribution.

Channel Distribution:

- Takeaway sales slightly outperform In-store sales across most product categories, indicating a clear customer preference for takeaway orders.
- This pattern highlights the importance of optimizing takeaway services, including packaging and order efficiency.

Payment Preference:

- Digital Wallets are the most commonly used payment method, followed by Credit Cards and Cash.
- This trend demonstrates a strong shift toward cashless payment methods, emphasizing the need for reliable digital payment infrastructure.

Overall Insight:

The dashboard effectively translates the cleaned dataset into actionable business insights, supporting data-driven decision-making related to product strategy, sales channels, and payment systems.

5. Conclusion

5.1 Conclusion

The practice successfully demonstrated that dirty data can be salvaged through logical cleaning pipelines. The cafe is performing well as a high-volume takeaway establishment, with healthy revenue flow driven by beverages (**Juice/Smoothies**) and high-value food items (**Salads**).

5.2 Recommendations

1. **POS System Audit:** The high frequency of "UNKNOWN" and "ERROR" entries indicates a flaw in the Point of Sale system. An IT audit is recommended to ensure data is captured correctly at the source.
2. **Inventory Management:** Given that **Smoothies** is top seller by volume, inventory levels for fresh fruits should be prioritized to prevent stockouts.
3. **Marketing:** Capitalize on the **Takeaway** trend by offering **Digital Walle**t exclusive promotions to further speed up service and reward loyal, tech-savvy customers.