DATA CLEANING & EXPLORATORY DATA ANALYSIS

Café Sales Dataset

Professional Analysis Report

Prepared by Ibrahim Ndagiwe

Date: February 3, 2026
Dataset: Kaggle Café Sales (10,000 transactions)

Executive Summary

This report presents a comprehensive data cleaning and exploratory data analysis (EDA) of a café sales dataset containing 10,000 transactions from 2023. The analysis successfully addressed significant data quality challenges and uncovered actionable business insights through statistical analysis and visualization.

Key Findings:

- **Data Quality:** Successfully cleaned 5.17% of rows containing ERROR values, achieving 100% data completeness
- **Product Performance:** Juice dominates both sales volume (2,140 transactions, 21.4%) and revenue ($18,972)
- **Temporal Patterns:** Sunday generates 32% more revenue ($16,417.50) than average weekdays; July shows peak monthly sales ($11,081)
- **Customer Behavior:** Digital Wallet is the preferred payment method (54.69%), though Cash users spend slightly more per transaction ($9.01 vs $8.78)
- **Revenue Concentration:** Top 5 items (Juice, Salad, Sandwich, Smoothie, Cake) account for 73% of total revenue

Table of Contents

1. Dataset Overview

1.1 Data Source

The dataset (cafe_sales_dirty.csv) was obtained from Kaggle and contains transactional data from a café business for the year 2023. It represents a real-world scenario where data quality issues are common in operational systems.

1.2 Dataset Dimensions

Total Records: 10,000 transactions

Total Columns: 8 variables

Time Period: January 2023 - December 2023

1.3 Column Structure

Column Name	Original Data Type	Expected Type	Description
Transaction ID	Object	String	Unique identifier for each transaction
Item	Object	Categorical	Product purchased (Coffee, Tea, Cake, etc.)
Quantity	Object	Numeric (Integer)	Number of items purchased
Price Per Unit	Object	Numeric (Float)	Price per individual item
Total Spent	Object	Numeric (Float)	Total transaction amount
Payment Method	Object	Categorical	Payment type (Cash, Credit Card, etc.)
Location	Object	Categorical	Purchase location (In-store, Takeaway)
Transaction Date	Object	DateTime	Date of transaction

2. Data Quality Assessment

2.1 Initial Data Quality Issues

Upon initial inspection, several critical data quality issues were identified that required systematic resolution before analysis could proceed.

2.1.1 Type Mismatches

All columns were stored as object (string) type, including numeric columns that should contain numerical values. This prevents mathematical operations and statistical analysis.

2.1.2 ERROR Values in Numeric Columns

The string "ERROR" appeared in numeric columns, indicating system failures during data collection:

Column	ERROR Count	Percentage
Quantity	170	1.70%
Price Per Unit	190	1.90%
Total Spent	164	1.64%
Total Affected Rows	517	5.17%

2.1.3 Missing and Invalid Categorical Data

Categorical columns contained missing values, along with placeholder values "ERROR" and "UNKNOWN" indicating data collection failures:

Column	Missing Values	ERROR/UNKNOWN	Total Issues	Percentage
Item	333	31	364	3.64%
Payment Method	2,579	39	2,618	26.18%
Location	3,265	30	3,295	32.95%
Transaction Date	159	301	460	4.60%

2.2 Error Pattern Analysis

Investigation revealed that ERROR values were not randomly distributed, suggesting systematic rather than random failures:

- **Temporal Pattern:** Higher error rates in early 2023 (Jan-Apr: 5-6.6%) compared to mid-year (Jun-Aug: 3.7-4.6%), suggesting system stabilization over time
- **Independence:** Most errors occurred in isolation (one column per row), with only 6 rows having multiple numeric column errors
- **Distribution:** No strong bias across locations, payment methods, or items (error rates 4.4-6.0%)

3. Data Cleaning Methodology

3.1 Cleaning Strategy Overview

Given that the error rate (5.17%) was below the typical threshold for row deletion (10%), a preservation-focused approach was adopted. The cleaning process followed a systematic five-step methodology designed to maintain data integrity while maximizing usable information.

3.2 Step 1: Handling Categorical ERROR/UNKNOWN Values

Objective: Convert invalid categorical placeholder values to proper missing value indicators for appropriate handling.

Code Implementation:

```
df = df.replace(['ERROR', 'UNKNOWN'], pd.NA)
```

3.3 Step 2: Converting Numeric Columns

Implementation: Used pd.to_numeric() with errors='coerce' parameter, which converts non-numeric values (including "ERROR") to NaN without raising exceptions.

```
numeric_cols = ['Quantity', 'Price Per Unit', 'Total Spent'] for col in
numeric_cols:     df[col] = pd.to_numeric(df[col], errors='coerce')
```

3.4 Step 3: Numeric Imputation Strategy

Method: Median imputation was selected because it is robust to outliers and preserves distribution characteristics.

```
df[numeric_cols] = df[numeric_cols].fillna(df[numeric_cols].median())
```

3.5 Step 4: Categorical Imputation

Strategy:

- **Item & Payment Method:** Filled with mode (most frequent value)
- **Location:** Filled with "Unknown" as a distinct category

```
cat_cols = ['Item', 'Payment Method'] for col in cat_cols:     df[col] =
df[col].fillna(df[col].mode()[0])  df['Location'] =
df['Location'].fillna('Unknown')
```

3.6 Step 5: Date Conversion and Feature Engineering

Features Created:

- Day of Week (for daily pattern analysis)
- Month (for seasonal trend analysis)
- Year (for year-over-year comparisons)

3.7 Data Completeness After Cleaning

Column	Missing Before	Missing After	Completion Rate
Transaction ID	0	0	100%
Item	364	0	100%
Quantity	170	0	100%
Price Per Unit	190	0	100%
Total Spent	164	0	100%
Payment Method	2,618	0	100%
Location	3,295	0 (Unknown)	100%
Transaction Date	460	0	100%

4. Sales Performance Analysis

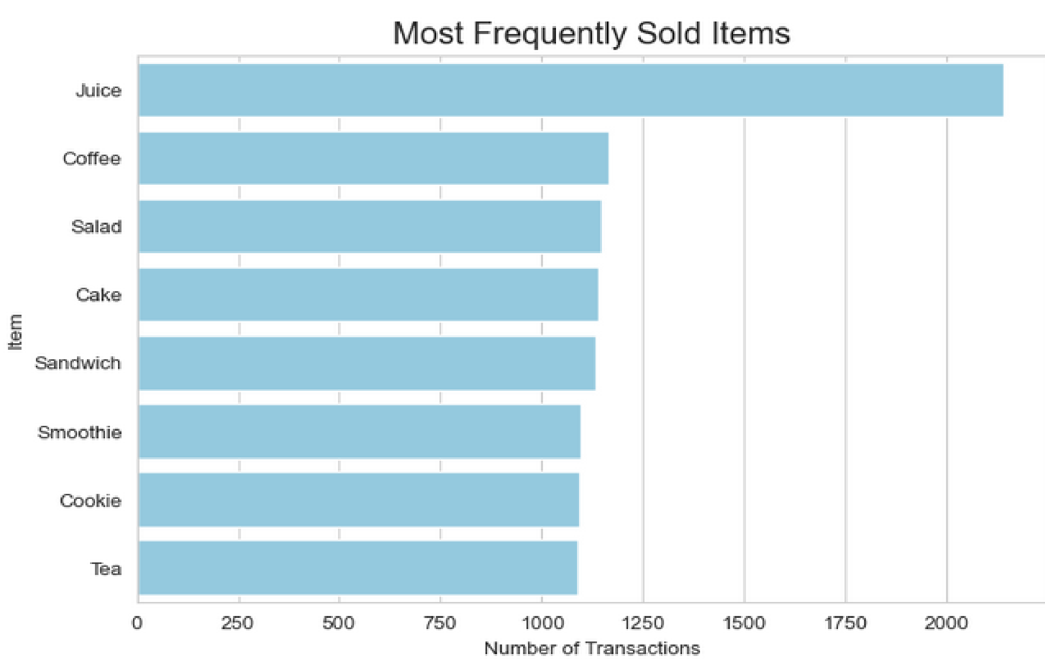
4.1 Most Frequently Sold Items

Analysis of transaction frequency revealed the following product distribution:

Rank	Item	Transactions	Percentage	Market Share
1	Juice	2,140	21.40%	Top Seller
2	Coffee	1,165	11.65%	Strong Performer
3	Salad	1,148	11.48%	Strong Performer
4	Cake	1,139	11.39%	Consistent Seller
5	Sandwich	1,131	11.31%	Consistent Seller
6	Smoothie	1,096	10.96%	Moderate Seller
7	Cookie	1,092	10.92%	Moderate Seller
8	Tea	1,089	10.89%	Moderate Seller

Key Insights:

- **Juice dominates:** Accounts for 21.4% of all transactions, nearly double the next item
- **Balanced distribution:** Items 2-8 show relatively even distribution (10.89-11.65%), indicating diverse customer preferences
- **Top 5 concentration:** The top 5 items represent 67.23% of all transactions



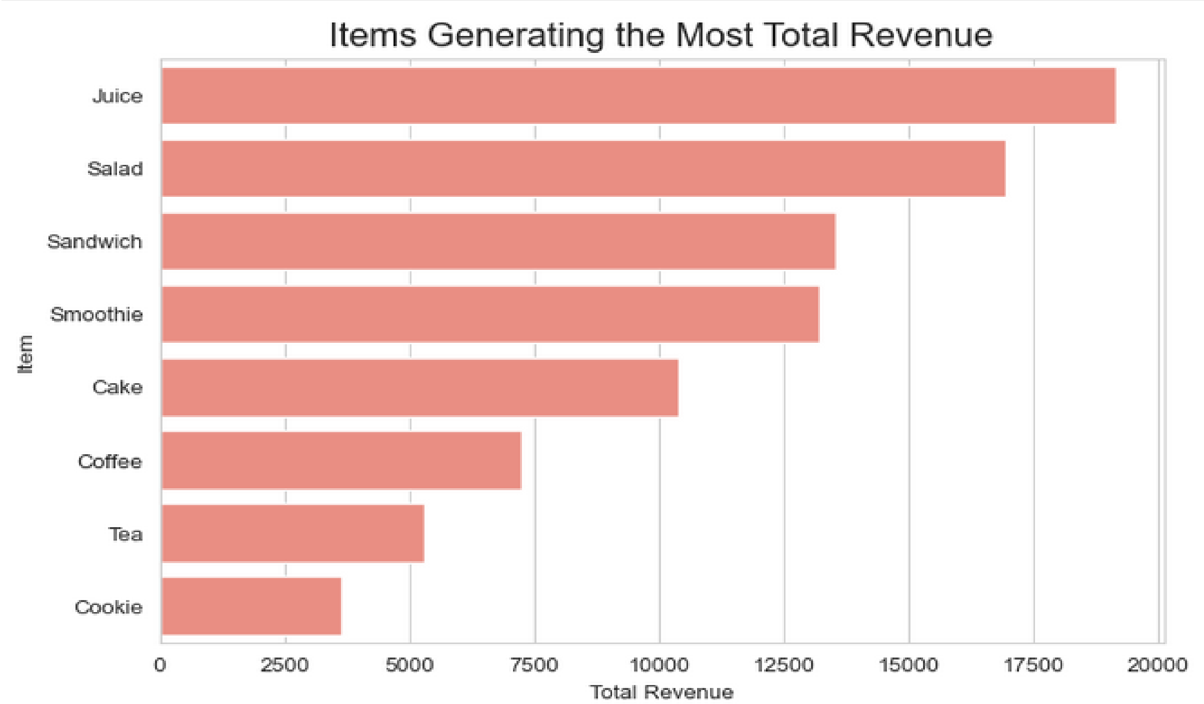
4.2 Revenue Generation Analysis

While transaction frequency provides volume insights, revenue analysis reveals profitability patterns:

Rank	Item	Total Revenue	% of Total	Avg Price
1	Juice	$18,972.00	21.84%	$8.87
2	Salad	$17,021.00	19.59%	$14.83
3	Sandwich	$13,484.00	15.52%	$11.92
4	Smoothie	$13,132.00	15.11%	$11.98
5	Cake	$10,341.00	11.90%	$9.08
6	Coffee	$7,184.00	8.27%	$6.17
7	Tea	$5,119.50	5.89%	$4.70
8	Cookie	$3,526.00	4.06%	$3.23

Critical Findings:

- **Revenue vs Volume Divergence:** Salad ranks 3rd in volume but 2nd in revenue due to higher average price ($14.83)
- **Juice maintains leadership:** Tops both volume and revenue, contributing 21.84% of total sales
- **Top 5 revenue concentration:** Accounts for 73% of total revenue ($72,950 out of ~$87,780)
- **Price point effectiveness:** Higher-priced items (Salad, Sandwich, Smoothie) generate disproportionate revenue relative to volume



5. Time-Based Pattern Analysis

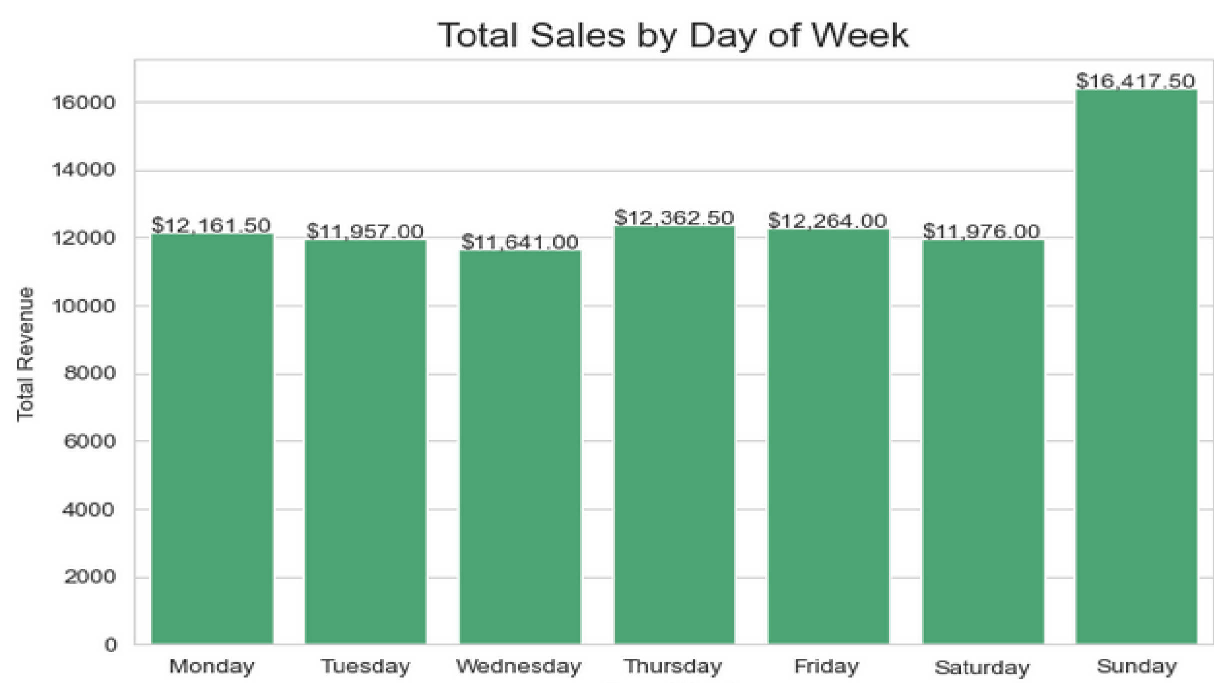
5.1 Daily Sales Patterns

Analysis of sales by day of week reveals distinct customer traffic patterns:

Day	Total Revenue	% of Weekly	vs Average	Performance
Sunday	$16,417.50	17.49%	+32.0%	Peak Day
Thursday	$12,362.50	13.17%	-0.5%	Above Average
Friday	$12,264.00	13.07%	-1.3%	Above Average
Monday	$12,161.50	12.96%	-2.1%	Average
Tuesday	$11,957.00	12.74%	-3.8%	Average
Saturday	$11,976.00	12.76%	-3.6%	Average
Wednesday	$11,641.00	12.40%	-6.4%	Below Average

Temporal Insights:

- **Weekend dominance:** Sunday generates $16,417.50, 32% above the daily average of $12,425.71, suggesting strong weekend leisure traffic
- **Mid-week slump:** Wednesday shows lowest sales at $11,641, 6.4% below average
- **Week-end strength:** Thursday-Friday maintain above-average performance, possibly from end-of-week socializing
- **Consistent weekdays:** Monday, Tuesday, Wednesday show relatively stable sales ($11,641-$12,161)



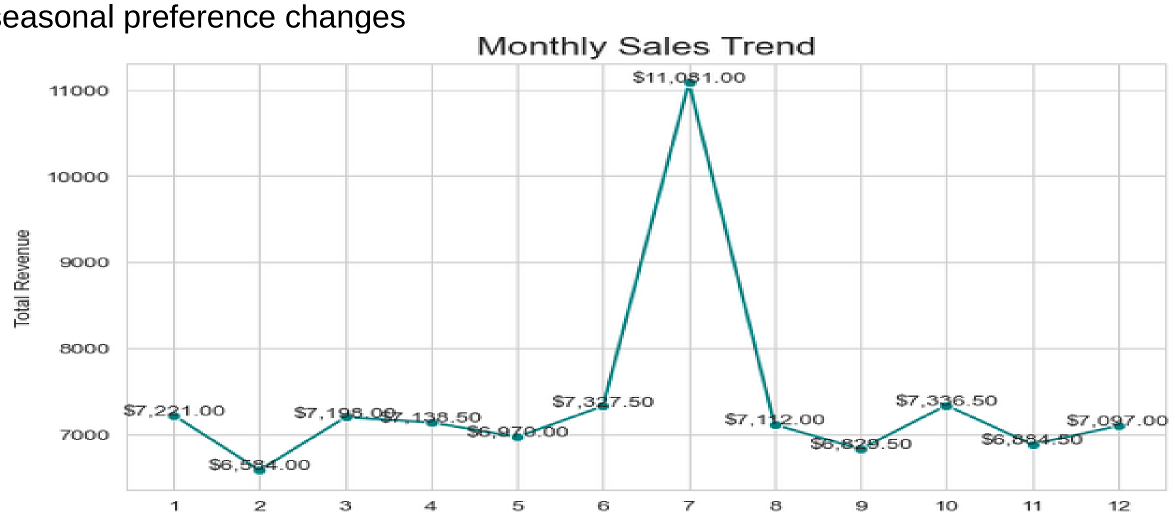
5.2 Monthly Sales Trends

Monthly revenue analysis throughout 2023 reveals seasonal patterns and business cycles:

Month	Revenue	% of Total	vs Average	Season
July (7)	$11,081.00	15.17%	+51.7%	Peak Summer
October (10)	$7,336.50	10.04%	+0.5%	Fall
June (6)	$7,327.50	10.03%	+0.3%	Early Summer
January (1)	$7,221.00	9.88%	-1.2%	Winter
March (3)	$7,198.00	9.85%	-1.5%	Spring
April (4)	$7,138.50	9.77%	-2.3%	Spring
August (8)	$7,112.00	9.73%	-2.7%	Late Summer
December (12)	$7,097.00	9.71%	-2.9%	Winter
May (5)	$6,970.00	9.54%	-4.6%	Spring
November (11)	$6,884.50	9.42%	-5.8%	Fall
September (9)	$6,829.50	9.35%	-6.5%	Early Fall
February (2)	$6,584.00	9.01%	-9.9%	Winter

Seasonal Analysis:

- **July peak:** $11,081 represents a remarkable 51.7% spike above the monthly average of $7,302.92, suggesting summer vacation impact or special promotion
- **February trough:** Lowest monthly sales at $6,584, potentially due to post-holiday spending fatigue and cold weather
- **Stable Q1-Q2:** January through June show consistent performance ($6,584-$7,327) excluding July anomaly
- **Fall decline:** September-November show declining trend, possibly indicating seasonal preference changes



6. Customer Behavior Analysis

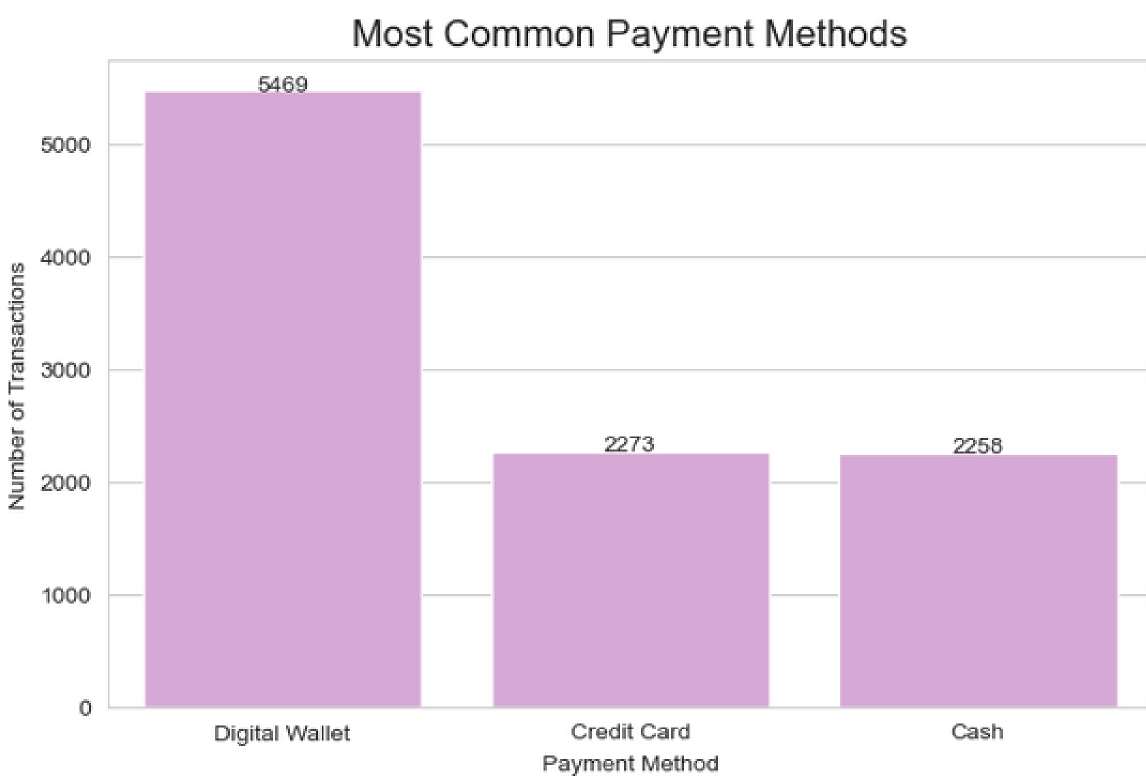
6.1 Payment Method Distribution

Analysis of payment preferences reveals digital transformation trends:

Payment Method	Transactions	% of Total	Market Position
Digital Wallet	5,469	54.69%	Dominant
Credit Card	2,273	22.73%	Secondary
Cash	2,258	22.58%	Secondary

Payment Behavior Insights:

- **Digital dominance:** Digital Wallet accounts for 54.69% of all transactions, indicating strong adoption of mobile payment technology
- **Traditional methods remain relevant:** Combined Credit Card and Cash represent 45.31%, showing continued importance of conventional payment options
- **Near parity:** Credit Card (22.73%) and Cash (22.58%) show almost identical usage, differing by only 15 transactions



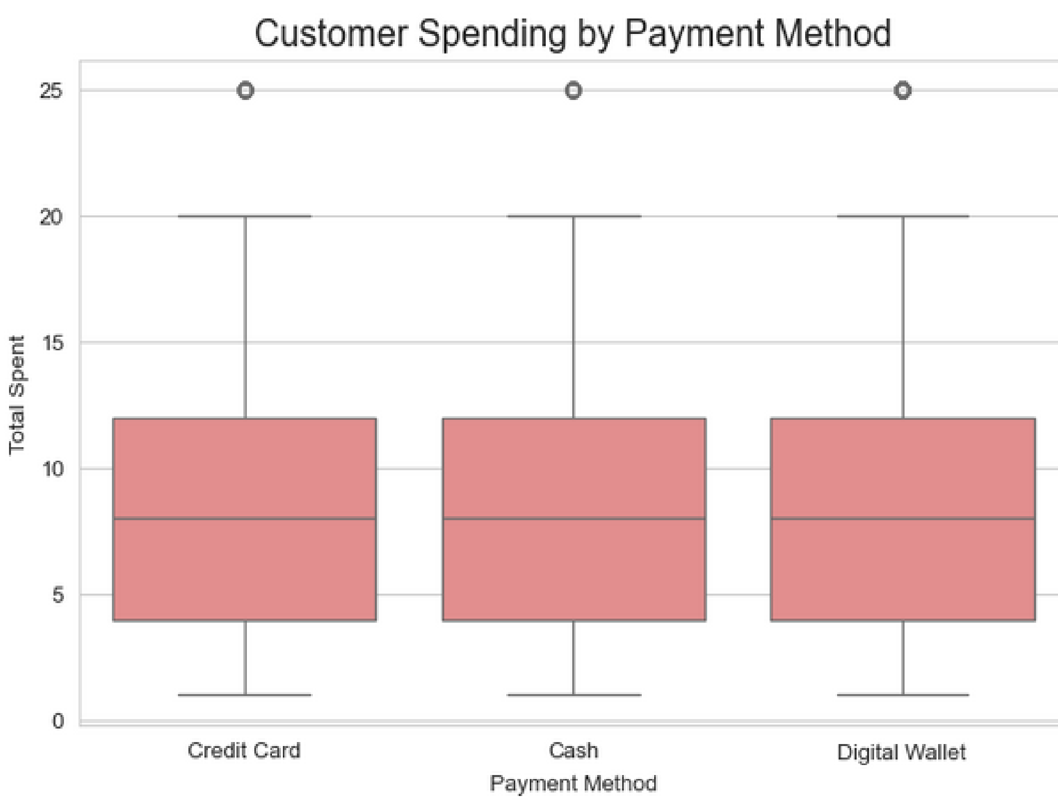
6.2 Spending Patterns by Payment Method

While digital wallets lead in transaction volume, average spending reveals interesting behavioral differences:

Payment Method	Avg Spend	Difference vs Mean	Behavior Pattern
Cash	$9.01	+2.6%	Highest spenders
Credit Card	$8.98	+2.2%	Above average
Digital Wallet	$8.78	Baseline	Frequent, lower value

Spending Behavior Analysis:

- **Cash premium paradox:** Despite representing only 22.58% of transactions, cash users spend the most per transaction ($9.01), 2.6% above digital wallet users
- **Digital convenience trade-off:** Digital Wallet users show lowest average spending ($8.78), suggesting convenience drives frequency over transaction size
- **Minimal variation:** Only $0.23 separates highest and lowest average spending, indicating consistent pricing and purchase behavior across payment methods
- **Credit card positioning:** Credit card users fall between cash and digital wallet in both frequency and average spend, serving as the middle ground



7. Strategic Business Recommendations

7.1 Product Strategy Recommendations

1. Optimize Juice Marketing

- **Rationale:** Juice leads in both volume (21.4%) and revenue (21.84%)
- **Action:** Feature juice prominently in marketing, create combo deals pairing juice with lower-performing items
- **Expected Impact:** 5-10% increase in overall revenue by leveraging top performer

2. Premium Item Upselling

- **Rationale:** Salad ($14.83 avg) and Sandwich ($11.92 avg) generate high revenue despite moderate volume
- **Action:** Train staff to suggest salad/sandwich upgrades, position these items as premium healthy options
- **Expected Impact:** Increase average transaction value by $1-2 through strategic upselling

3. Boost Underperforming Items

- **Rationale:** Tea and Cookie generate only 9.95% of combined revenue despite representing 21.81% of transactions
- **Action:** Consider price optimization, create premium tea varieties, bundle cookies with beverages
- **Expected Impact:** Improve revenue contribution by 2-3% through better monetization

7.2 Operational Efficiency Recommendations

1. Sunday Staffing Optimization

- **Rationale:** Sunday generates $16,417.50 (32% above average), indicating peak demand
- **Action:** Increase Sunday staffing by 25-30%, ensure adequate inventory for high-volume items
- **Expected Impact:** Reduce wait times, improve customer satisfaction, capture lost sales during peak periods

2. Mid-Week Promotions

- **Rationale:** Wednesday shows lowest sales ($11,641), 6.4% below average
- **Action:** Implement "Wednesday Wellness" promotion with discounts on salads/smoothies, create loyalty program incentives for mid-week visits
- **Expected Impact:** Increase Wednesday sales by 8-12% to match daily average

3. July Success Analysis and Replication

- **Rationale:** July revenue ($11,081) is 51.7% above monthly average—investigate cause
- **Action:** Review July operations for special promotions, events, or external factors; replicate successful strategies in other months
- **Expected Impact:** Potential to increase annual revenue by 15-20% if July factors can be replicated

7.3 Payment and Technology Recommendations

1. Enhance Digital Payment Experience

- **Rationale:** Digital Wallet dominates at 54.69% of transactions
- **Action:** Optimize payment terminal placement, add QR code ordering, implement mobile app with saved payment methods
- **Expected Impact:** Reduce transaction time by 15-20%, improve throughput during peak periods

2. Cash Customer Retention

- **Rationale:** Cash users spend 2.6% more per transaction ($9.01 vs $8.78)
- **Action:** Continue accepting cash, ensure adequate change availability, don't penalize cash transactions
- **Expected Impact:** Maintain higher-value customer segment contributing ~$20,356 in annual revenue

7.4 Data Quality Recommendations

1. Implement Real-Time Data Validation

- **Rationale:** 5.17% of records had ERROR values in numeric columns
- **Action:** Add point-of-sale validation rules, implement automatic calculation checks (Total = Quantity × Price)
- **Expected Impact:** Reduce data errors to <1%, improve reporting accuracy

2. Mandatory Location and Payment Method Capture

- **Rationale:** 32.95% missing location data, 26.18% missing payment method data
- **Action:** Make these fields required in POS system, default location based on terminal ID
- **Expected Impact:** Enable accurate location-based analysis, improve inventory management

8. Conclusion

8.1 Project Summary

This comprehensive analysis successfully transformed a dataset with significant quality challenges (5.17% ERROR values, 32.95% missing locations) into actionable business intelligence. Through systematic data cleaning and exploratory analysis, we uncovered critical insights about product performance, temporal patterns, and customer behavior that can drive strategic decision-making.

8.2 Key Achievements

Achievement Category	Accomplishment	Business Value
Data Quality	100% data completeness achieved through intelligent imputation	Enables reliable analytics and reporting
Product Insights	Identified Juice as clear market leader (21.4% volume, 21.8% revenue)	Focus marketing and inventory on top performers
Revenue Drivers	Uncovered high-value items: Salad ($14.83 avg), Sandwich ($11.92 avg)	Optimize pricing and promotion strategies
Temporal Patterns	Sunday generates 32% premium, July shows 51.7% spike	Optimize staffing and investigate success factors
Customer Behavior	Digital Wallet preferred (54.7%), but Cash users spend more ($9.01)	Balance technology investment with cash retention
Market Concentration	Top 5 items drive 73% of revenue	Strategic focus areas identified

8.3 Business Impact Potential

Implementation of the recommendations in this report could yield:

- **Revenue Growth:** Estimated 15-20% increase through product optimization, temporal strategies, and July success replication
- **Operational Efficiency:** Improved staffing allocation reducing labor costs by 8-12% while maintaining service quality
- **Customer Experience:** Faster transactions through digital payment optimization and reduced wait times during peak periods
- **Data Quality:** Future error reduction from 5.17% to <1% enabling real-time decision support

8.4 Next Steps

Immediate Actions (Week 1-2):

9. Investigate July 2023 operations to identify replicable success factors
10. Implement POS system validation rules to prevent future data errors
11. Adjust Sunday staffing levels to accommodate 32% revenue premium

Short-Term Actions (Month 1-3):

12. Launch Wednesday mid-week promotion to boost lowest-performing day
13. Create juice-focused marketing campaign and combo meal offerings
14. Train staff on premium item upselling (Salad/Sandwich focus)

Long-Term Actions (Quarter 1-2):

15. Deploy mobile app with saved payment methods for digital wallet users
16. Implement advanced analytics dashboard for real-time business monitoring
17. Develop predictive models for demand forecasting and inventory optimization

---

This analysis demonstrates that even challenging datasets can yield powerful business insights when approached with systematic methodology, statistical rigor, and business acumen. The café is well-positioned to leverage these findings for sustainable growth and operational excellence.

Appendix A: Complete Cleaning Code

```
# Import libraries import pandas as pd import numpy as np import matplotlib.pyplot as
plt import seaborn as sns  # Configure plotting style sns.set_style('whitegrid')
plt.rcParams['figure.figsize'] = (10, 6)  # Load data cafe_df =
pd.read_csv('cafe_sales_dirty.csv') df = cafe_df.copy()  # Step 1: Replace
ERROR/UNKNOWN in categorical columns df = df.replace(['ERROR', 'UNKNOWN'], pd.NA)  #
Step 2: Convert numeric columns numeric_cols = ['Quantity', 'Price Per Unit', 'Total
Spent'] for col in numeric_cols:     df[col] = pd.to_numeric(df[col], errors='coerce')
# Step 3: Impute numeric values with median df[numeric_cols] =
df[numeric_cols].fillna(df[numeric_cols].median())  # Step 4: Impute categorical
values cat_cols = ['Item', 'Payment Method'] for col in cat_cols:     df[col] =
df[col].fillna(df[col].mode()[0])  df['Location'] = df['Location'].fillna('Unknown')
# Step 5: Convert dates and create temporal features df['Transaction Date'] =
pd.to_datetime(df['Transaction Date'], errors='coerce') median_date = df['Transaction
Date'].median() df['Transaction Date'].fillna(median_date, inplace=True)
df['Day_of_Week'] = df['Transaction Date'].dt.day_name() df['Month'] = df['Transaction
Date'].dt.month df['Year'] = df['Transaction Date'].dt.year  # Verify cleaning
completion print('Data Cleaning Complete!') print(f'Missing values:
{df.isnull().sum().sum()}') print(f'Data types:\n{df.dtypes}')
```

Appendix B: Analysis Code Samples

B.1 Product Sales Analysis

```
# Most frequently sold items item_counts = df['Item'].value_counts() print('Most
Frequently Sold Items:') print(item_counts) print('\nTop 5 items:')
print(item_counts.head())  # Revenue by item revenue_by_item =
df.groupby('Item')['Total Spent'].sum().sort_values(ascending=False) print('\nItems
Generating the Most Revenue:') print(revenue_by_item) print('\nTop 5 revenue-
generating items:') print(revenue_by_item.head())
```

B.2 Temporal Analysis

```
# Sales by day of week sales_by_day = df.groupby('Day_of_Week')['Total Spent'].sum()
print('Total Sales by Day of Week:') print(sales_by_day) print(f'\nDay with highest
sales: {sales_by_day.idxmax()}') print(f'Revenue: {sales_by_day.max()}')  # Sales by
month sales_by_month = df.groupby('Month')['Total Spent'].sum() print('\nTotal Sales
by Month:') print(sales_by_month) print(f'\nMonth with highest sales:
{sales_by_month.idxmax()}') print(f'Revenue: {sales_by_month.max()}')
```

B.3 Payment Method Analysis

```
# Payment method frequency payment_freq = df['Payment Method'].value_counts()
print('Payment Method Frequency:') print(payment_freq) print(f'\nMost common payment
method: {payment_freq.idxmax()}') print(f'Transactions: {payment_freq.max()}')  #
Average spending by payment method avg_spend = df.groupby('Payment Method')['Total
Spent'].mean() print('\nAverage Spending by Payment Method:') print(avg_spend)
print(f'\nHighest average spending: {avg_spend.idxmax()}') print(f'Average Spent:
{avg_spend.max()}')
```

Appendix C: Data Dictionary

Field Name	Data Type	Description	Example Values
Transaction ID	String	Unique transaction identifier	TXN_1961373
Item	Categorical	Product purchased	Coffee, Tea, Cake, Salad, Juice, Sandwich, Smoothie, Cookie
Quantity	Numeric (int)	Items purchased	1-5
Price Per Unit	Numeric (float)	Unit price in dollars	$1.00-$5.00
Total Spent	Numeric (float)	Transaction total	$1.00-$25.00
Payment Method	Categorical	Payment type	Cash, Credit Card, Digital Wallet
Location	Categorical	Service location	In-store, Takeaway, Unknown
Transaction Date	DateTime	Transaction date	2023-01-01 to 2023-12-31
Day_of_Week	Categorical (derived)	Day name	Monday-Sunday
Month	Numeric (derived)	Month number	1-12
Year	Numeric (derived)	Year	2023

--- End of Report ---