Analysis Report: Cafe Sales Data

1. Data Overview

The dataset contains 10,000 transactions from a cafe, including details such as items purchased, quantity, price, total spent, payment method, location, and transaction date. Key steps in data preparation included handling missing values, converting data types, imputing missing numeric values with medians, replacing placeholder strings, and treating outliers.

2. Key Relationships & Insights

A. Numeric Variables

Total Spent vs. Quantity & Price
Expected Relationship: Total Spent should ideally equal Quantity × Price Per Unit.
Observation: Discrepancies were found (e.g., 'ERROR' in the original data). After imputation, the median was used, which may not reflect true values. This introduces potential inaccuracies in analyzing exact sales figures.
Recommendation: Recalculate Total Spent using Quantity × Price Per Unit where possible to improve data accuracy.

Correlation Analysis
Quantity vs. Price Per Unit: Weak correlation, suggesting items vary in pricing (e.g., coffee vs. cake).
Total Spent vs. Quantity/Price: Strong positive correlation, as expected. Outliers were capped during cleaning, which may reduce variance but stabilizes trends.

B. Categorical Variables

Popular Items
Top Items: Coffee, Juice, Salad, and Cake were the most frequently sold.
Least Popular: Items marked as 'ERROR' or 'UNKNOWN' (data quality flags).

Payment Methods
Dominant Method: Digital Wallet (most common), followed by Credit Card and Cash.
Data Issues: 2,579 missing values in Payment Method, imputed with the most frequent value (Digital Wallet).

Location

Sales Distribution:
In-store: 3,017 transactions.
Takeaway: 3,022 transactions.
Data Issues: 3,265 missing values, imputed with 'Takeaway' (most frequent).

C. Temporal Trends
Transaction Dates: Extracted into Year, Month, Day, Weekday, and Hour.
Monthly Sales: Potential spikes in April and September (highest transaction counts).
Hourly Trends: Peaks during morning (8–10 AM) and afternoon (2–4 PM), aligning with typical cafe rush hours.

3. Data Quality Issues
Total Spent Calculation
Original data included 'ERROR' values. Imputation with median may not reflect true sales.
Impact: Affects accuracy of revenue analysis.
Missing Values
Significant missing data in Location (3,265), Payment Method (2,579), and Item (333).
Impact: May skew categorical distributions.

Outliers
Outliers in Quantity, Price, and Total Spent were capped, potentially masking extreme but valid transactions.

4. Recommendations
Recalculate Total Spent using Quantity × Price Per Unit to ensure accuracy.
Investigate Missing Data: Determine why Location and Payment Method have high missing rates.
Seasonal Promotions: Leverage peak hours/months for targeted marketing (e.g., morning coffee discounts).
Item Popularity: Stock more high-demand items (Coffee, Juice) and review low-sales items (e.g., 'ERROR' entries).

5. Visualizations
Correlation Matrix: Show relationships between Quantity, Price, and Total Spent.
Item Sales Distribution: Bar chart of top-selling items.
Hourly Sales Trend: Line graph highlighting peak hours.
Payment Method Preference: Pie chart showing method prevalence.

Conclusion
The dataset reveals strong sales for coffee and digital wallet usage, with consistent in-store and takeaway demand. Data quality improvements (e.g., recalculating Total Spent) would enhance reliability. Temporal trends suggest opportunities for optimizing staffing and promotions during peak hours.