EDA of CAFE SALES data

The dataset is loaded and initial inspections are performed.

The dataset contains transaction records from a cafe business with the following key features:

- Transaction details (Transaction ID, Transaction Date)
- Product information (Item, Quantity, Price Per Unit, Total Spent)
- Payment information (Payment Method)
- Store information (Location)

Handling missing values

1] Replaced "ERROR", "UNKNOWN", and empty strings with NaN values
2] Converted numerical columns ('Quantity', 'Price Per Unit', 'Total Spent') to numeric data types, with errors coerced to NaN
3] Converted 'Transaction Date' column to datetime format
4] Categorical columns are converted to 'object' type

Actions performed on missing values:

- Numerical columns (Quantity, Price Per Unit, Total Spent) had missing values replaced with column means
- Categorical columns (Item, Payment Method, Location) had missing values replaced with "Missing"
- Rows with missing Transaction Dates were removed.

Duplicate detection

There were no duplicate Transaction IDs in the dataset

Outlier Detection and Removal

Outliers are detected and removed using the Interquartile Range (IQR) method

For each numerical column (Quantity, Price Per Unit, Total Spent):

- Used IQR method and calculated Q1 and Q3
- Identified and removed values outside these bounds

- The percentage of outliers for each column was reported

Univariate Analysis

- Summary statistics using describe() provided mean, standard deviation, min, max, and quartiles
- Created histograms and box plots for numerical columns ('Quantity', 'Price Per Unit', 'Total Spent')

Bivariate Analysis

Numerical Relationships:

- Correlation matrix between numerical variables (Quantity, Price Per Unit, Total Spent)
- Scatter plot of Quantity vs. Total Spent to examine their relationship

Categorical vs. Numerical Relationships:

- Bar plots showing average Total Spent by Item categories
- Box plots showing distribution of Total Spent across Item categories
- Violin plots showing distribution of Total Spent by Payment Method
- Box plots showing distribution of Total Spent by Location

Multivariate Analysis

- Pair plots of numerical variables colored by Item categories
- Correlation heatmap including dummy-encoded categorical variables
- Grouped comparisons using:
    - Box plots of Total Spent by Item, grouped by Payment Method
    - Box plots of Total Spent by Item, grouped by Location
    - Violin plots of Total Spent by Payment Method, grouped by Location
    - Scatter plots of Quantity vs. Total Spent, colored by Item

Conclusion

The visualizations and analyses reveal:
The data shows distinct patterns in customer spending behaviour across different products, payment methods, and locations.

1. The dataset contains information on various items sold, payment methods used, and transaction locations.
2. There were missing values and outliers in the data, which were addressed through imputation and removal

3. The correlation matrix shows relationships between numerical variables
4. The distribution of total spent varies across different items, payment methods, and locations, as shown in the box plots and bar charts.
5. The pair plot reveals potential patterns in the relationships between numerical variables, segmented by item type