DAI ASSIGNMENT 1: DATA ANALYSIS OF COFFEE SHOP SALES DATASET

1. Introduction

The goal of this analysis is to clean, preprocess, and explore the coffee sales dataset. This includes handling missing values, correcting inconsistencies, detecting outliers, and performing exploratory data analysis (EDA) to gain insights into sales patterns.

---

2. Data Cleaning

2.1 Handling Missing Values

- Categorical columns with missing values were filled using the mode of each respective column.
- Numerical columns with missing values were imputed using the median to maintain the data distribution.

2.2 Standardization of Text Data

- Converted categorical values to a consistent format (e.g., title-casing names and locations, trimming whitespace).
- Standardized item names (e.g., replacing 'ERROR' and 'UNKNOWN' with mode values).

2.3 Handling Outliers (Winsorization)

- Instead of removing outliers outright, Winsorization was applied by capping extreme values at the 5th and 95th percentiles for numerical variables.

---

3. Exploratory Data Analysis (EDA)

3.1 Univariate Analysis

- **Distribution of Sales:** Histogram and box plots revealed skewness in total sales values.
- **Quantity Sold:** Most frequent quantity was between 1-3 units per transaction.
- **Price per Unit:** Most items were priced within a reasonable range, with a few premium-priced products.
- **Most Popular Items:** Bar charts indicated the top-selling coffee varieties.

3.2 Bivariate Analysis

- **Correlation Analysis:**
  - Positive correlation between Quantity Sold and Total Sales.
  - Weak correlation between Price Per Unit and Total Sales, indicating pricing strategy does not directly affect bulk purchasing.

3.3 Multivariate Analysis

- **Pairwise Relationship (Pairplot Analysis)**:
    - Confirmed trends in pricing, quantity, and total sales.
- **Impact of Payment Methods on Sales:**
    - Digital payments were more commonly used than cash, as seen in the grouped bar charts.
- **Time-Based Trends:**
    - Sales fluctuated during different times of the day and across days of the week.
    - Sales dropped in februrary and there were other time based trades as well which can be used to improve the total sales of the coffee shop

---

4. Data Visualization Summary

- **Histograms & Boxplots:** Illustrated data distributions and outlier effects.
- **Correlation Heatmaps:** Provided insight into interdependencies among numerical variables.
- **Bar Charts:** Displayed categorical trends such as the most frequently purchased items.

  Also helped in analysis of sales by months, days and year.
- **Line Charts:** Showed time-based sales fluctuations.

---

5. Key Insights & Recommendations

1. **Pricing Strategy:** Since Price per Unit has largeimpact on total sales, consider offering discounts on larger purchases to increase quantity sold.
2. **Popular Items:** The most popular items is juice with more frequency than any other item. The shop categories itself as a coffee shop and should consider rebranding to put more focus on juices since it seems to be a bestseller.
3. **Store Locations:** The coffe shop franchise should also add a location column to their data so the stores locations with most sales can be analyzed and expanded.
4. **Time-Based Sales Promotions:** It should be noted that sales drop a bit in February and there are some other time based variations analysed. The sales can be boosted using these
5. **Digital Payment Optimization:** With a high percentage of digital transactions, ensuring a seamless payment experience may enhance customer satisfaction.

---

6. Conclusion

This analysis provided a structured approach to cleaning, processing, and exploring the coffee sales dataset. Winsorization was used for handling outliers instead of outright removal, missing values were carefully imputed, and multiple statistical and visual techniques were applied to derive meaningful insights.