Deliverables & Documentation

Subtask:

Summarize the cleaning and transformation steps performed, provide an overview of the final dataset's characteristics, and discuss any key observations or insights gained during the preparation process.

Summary of Data Processing and Final Dataset Characteristics

This section summarizes the data cleaning and transformation steps performed on the dirty_cafe_sales.csv dataset, provides an overview of the resulting df_sales DataFrame, and highlights key observations.

1. Initial Data Quality Issues Identified

During the 'Data Quality Assessment' phase, several issues were identified:

- **Missing Values**: Significant missing values were present across multiple columns, including Item, Quantity, Price Per Unit, Total Spent, Payment Method, Location, and Transaction Date.
- **Inconsistent Data Entries**: Numerical columns (Quantity, Price Per Unit, Total Spent) and categorical columns (Item, Payment Method, Location, Transaction Date) contained non-standard string values such as 'ERROR' and 'UNKNOWN'. These entries effectively represented missing or invalid data but were not initially recognized as NaN.
- **Incorrect Data Types**: All columns were initially imported as object type. Specifically, Quantity, Price Per Unit, and Total Spent should have been numeric (float/int), and Transaction Date should have been a datetime object.
- **No Duplicate Rows**: Fortunately, no duplicate rows were found in the dataset.

2. Cleaning Steps Undertaken

The following cleaning steps were executed to address the identified issues:

- **Standardizing Missing Values**: 'ERROR' and 'UNKNOWN' strings in Quantity, Price Per Unit, Total Spent, Item, Payment Method, Location, and Transaction Date columns were replaced with np.nan (or NaT for datetime).
- **Correcting Data Types**: Quantity, Price Per Unit, and Total Spent were converted to float64 after standardizing missing values, coercing any remaining errors to NaN. Transaction Date was converted to datetime64[ns] after standardizing missing values, coercing errors to NaT.
- **Conditional Imputation for Numerical Columns**: For Quantity, Price Per Unit, and Total Spent, missing values were conditionally imputed where two of the three related fields were present. For example, if Total Spent was missing but Quantity and Price Per Unit were available, Total Spent was calculated as Quantity * Price Per Unit.
- **Median Imputation for Remaining Numerical Missing Values**: Any remaining NaN values in Quantity, Price Per Unit, and Total Spent after conditional imputation were filled using the median of their respective columns.

- **Mode Imputation for Categorical Missing Values**: Missing values in Item, Payment Method, and Location were imputed using the mode (most frequent value) of each column.
- **Removal of Rows with Missing Transaction Dates**: Rows with NaT (Not a Time) in the Transaction Date column were dropped, as these critical entries could not be reliably imputed without losing analytical integrity.

3. Data Transformation Steps

New features were created to enrich the dataset for further analysis:

- **Date Component Extraction**: From the Transaction Date column, the following new features were extracted:
    - Transaction_Year
    - Transaction_Month
    - Transaction_Day
    - Transaction_DayOfWeek (numeric, 0 for Monday to 6 for Sunday)
    - Transaction_DayName (e.g., 'Monday', 'Tuesday')
- **Weekday/Weekend Indicator**: A new categorical feature Transaction_WeekdayWeekend was created, categorizing transactions as 'Weekday' or 'Weekend' based on Transaction_DayOfWeek.
- **Consistency of 'Total Spent'**: The Total Spent column was re-calculated as the product of Quantity and Price Per Unit across all rows to ensure absolute consistency and correct any potential discrepancies introduced by earlier imputation or original dirty data.

4. Overview of the Final df_sales Data Frame

After all cleaning and transformation steps, the df_sales DataFrame is now prepared for analysis:

- **Dimensions**: The DataFrame has been reduced from 10,000 rows to 9,540 rows and now contains 14 columns.
- **Missing Values**: There are no missing values (isnull().sum() reports 0 for all columns), confirming a clean dataset.
- **Data Types**: All columns now have appropriate data types, including float64 for numerical values, datetime64[ns] for the transaction date, int32 for extracted date components, and object for categorical strings, ready for analysis.
- **Descriptive Statistics**: The df_sales.describe() output shows reasonable distributions for numerical columns, with Quantity ranging from 1 to 5, Price Per Unit from 1.0 to 5.0, and Total Spent from 1.0 to 25.0. The date-related descriptive statistics show that all transactions occurred in 2023.

5. Key Observations and Insights

Based on the cleaned and transformed dataset:

- **Transaction Volume**: Approximately 95.4% of the original records were retained, indicating a robust dataset after cleaning.

- **Item Distribution**: While not explicitly quantified here, the Item column now contains 9 distinct product categories (excluding NaN). Further analysis could focus on the popularity of different items.
- **Payment Methods**: Payment Method now has 3 distinct categories. Digital Wallet, Cash, and Credit Card. Further analysis could involve transaction trends per payment type.
- **Location Analysis**: Location is categorized into Takeaway and In-store. This allows for comparison of sales performance across different service models.
- **Seasonal/Daily Trends**: The newly created date features (Year, Month, Day, DayOfWeek, DayName, Weekday/Weekend) will enable detailed time-series analysis to identify peak sales periods, popular days of the week, or monthly trends.
- **Consistency**: The re-calculation of Total Spent ensures that all financial calculations in the dataset are internally consistent, which is crucial for accurate sales reporting and financial analysis.