**About Dataset**

**Dirty Cafe Sales Dataset**

**Overview**

The Dirty Cafe Sales dataset contains 10,000 rows of synthetic data representing sales transactions in a cafe. This dataset is intentionally "dirty," with missing values, inconsistent data, and errors introduced to provide a realistic scenario for data cleaning and exploratory data analysis (EDA). It can be used to practice cleaning techniques, data wrangling, and feature engineering.

**File Information**

* **File Name:** dirty\_cafe\_sales.csv
* **Number of Rows:** 10,000
* **Number of Columns:** 8

**Columns Description**

| **Column Name** | **Description** | **Example Values** |
| --- | --- | --- |
| Transaction ID | A unique identifier for each transaction. Always present and unique. | TXN\_1234567 |
| Item | The name of the item purchased. May contain missing or invalid values (e.g., "ERROR"). | Coffee, Sandwich |
| Quantity | The quantity of the item purchased. May contain missing or invalid values. | 1, 3, UNKNOWN |
| Price Per Unit | The price of a single unit of the item. May contain missing or invalid values. | 2.00, 4.00 |
| Total Spent | The total amount spent on the transaction. Calculated as Quantity \* Price Per Unit. | 8.00, 12.00 |
| Payment Method | The method of payment used. May contain missing or invalid values (e.g., None, "UNKNOWN"). | Cash, Credit Card |
| Location | The location where the transaction occurred. May contain missing or invalid values. | In-store, Takeaway |
| Transaction Date | The date of the transaction. May contain missing or incorrect values. | 2023-01-01 |

**Data Characteristics**

1. **Missing Values:**
   * Some columns (e.g., Item, Payment Method, Location) may contain missing values represented as None or empty cells.
2. **Invalid Values:**
   * Some rows contain invalid entries like "ERROR" or "UNKNOWN" to simulate real-world data issues.
3. **Price Consistency:**
   * Prices for menu items are consistent but may have missing or incorrect values introduced.

**Menu Items**

The dataset includes the following menu items with their respective price ranges:

|  |  |
| --- | --- |
| **Item** | **Price($)** |
| Coffee | 2 |
| Tea | 1.5 |
| Sandwich | 4 |
| Salad | 5 |
| Cake | 3 |
| Cookie | 1 |
| Smoothie | 4 |
| Juice | 3 |

**Use Cases**

This dataset is suitable for:

* Practicing **data cleaning** techniques such as handling missing values, removing duplicates, and correcting invalid entries.
* Exploring **EDA techniques** like visualizations and summary statistics.
* Performing **feature engineering** for machine learning workflows.

**Cleaning Steps Suggestions**

To clean this dataset, consider the following steps:

1. **Handle Missing Values:**
   * Fill missing numeric values with the median or mean.
   * Replace missing categorical values with the mode or "Unknown."
2. **Handle Invalid Values:**
   * Replace invalid entries like "ERROR" and "UNKNOWN" with NaN or appropriate values.
3. **Date Consistency:**
   * Ensure all dates are in a consistent format.
   * Fill missing dates with plausible values based on nearby records.
4. **Feature Engineering:**
   * Create new columns, such as Day of the Week or Transaction Month, for further analysis.

**License**

This dataset is released under the [CC BY-SA 4.0 License](https://creativecommons.org/licenses/by-sa/4.0/). You are free to use, share, and adapt it, provided you give appropriate credit.

**Feedback**

If you have any questions or feedback, feel free to reach out through the dataset's discussion board on Kaggle.