**Coffee Shop Sales Data Cleaning Report**

This report details the potential data quality issues, including missing values, data type inconsistencies, structural inconsistencies, and duplicate entries, found in the Coffee Shop Sales.xlsx - Transactions.csv dataset.

**1. Dataset Overview**

First, let's get a general understanding of the dataset's structure.

* **Total Rows:** 149116
* **Total Columns:** 11

**2. Missing Values**

Missing values can lead to incomplete or biased analysis. Below is a summary of columns with missing data:

|  |  |  |
| --- | --- | --- |
| **Column** | **Missing Count** | **Missing Percentage (%)** |
| transaction\_id | 0 | 0.00 |
| transaction\_date | 0 | 0.00 |
| transaction\_time | 0 | 0.00 |
| transaction\_qty | 0 | 0.00 |
| store\_id | 0 | 0.00 |
| store\_location | 0 | 0.00 |
| product\_id | 0 | 0.00 |
| unit\_price | 0 | 0.00 |
| product\_category | 0 | 0.00 |
| product\_type | 0 | 0.00 |
| product\_detail | 0 | 0.00 |

**Conclusion on Missing Values:** The dataset appears to be **complete with no missing values** across any of the columns. This is excellent for analysis as it avoids the need for imputation or deletion strategies.

**3. Data Type Inconsistencies**

Ensuring correct data types is crucial for accurate calculations and time-series analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Current Data Type** | **Expected Data Type** | **Potential Issues** |
| transaction\_id | int64 | int64 | Correct. |
| transaction\_date | object | datetime | Needs conversion from string to datetime object. |
| transaction\_time | object | datetime | Needs conversion from string to datetime object. |
| transaction\_qty | int64 | int64 | Correct. |
| store\_id | int64 | int64 | Correct. |
| store\_location | object | object (string) | Correct. |
| product\_id | int64 | int64 | Correct. |
| unit\_price | float64 | float64 | Correct. |
| product\_category | object | object (string) | Correct. |
| product\_type | object | object (string) | Correct. |
| product\_detail | object | object (string) | Correct. |

**Recommendations for Data Type Cleaning:**

* Convert transaction\_date to a datetime format.
* Convert transaction\_time to a datetime format. It might be beneficial to combine transaction\_date and transaction\_time into a single transaction\_datetime column for easier time-based analysis.

**4. Structural Inconsistencies (Categorical Columns)**

Structural inconsistencies typically involve variations in text entries (e.g., casing, extra spaces, misspellings) that represent the same entity.

**store\_location**

* **Unique Values Found:**
  + Lower Manhattan
  + Hell's Kitchen
  + Astoria

**Conclusion on store\_location:** The store\_location column appears clean with consistent naming and no obvious structural errors or extra spaces.

**product\_category**

* **Unique Values Found:**
  + Coffee
  + Tea
  + Drinking Chocolate
  + Bakery
  + Coffee beans
  + Branded
  + Flavours
  + Loose Tea
  + Packaged Chocolate

**Conclusion on product\_category:** The product\_category column appears clean with consistent naming and no obvious structural errors or extra spaces.

**product\_type (Top 20 most frequent for brevity)**

* **Unique Values Found (Top 20):**
  + Gourmet brewed coffee
  + Brewed Chai tea
  + Hot chocolate
  + Drip coffee
  + Barista Espresso
  + Brewed Black tea
  + Premium brewed coffee
  + Scone
  + Brewed herbal tea
  + Biscotti
  + Organic brewed coffee
  + Brewed Green tea
  + Croissant
  + Housewares
  + Organic Beans
  + Pastry
  + Espresso
  + Colombian
  + Ethiopia
  + Jamaican

**Conclusion on product\_type:** The product\_type column appears generally clean. No immediate structural inconsistencies (like different casings for the same type or extra spaces) are apparent in the top 20 unique values. A full review of all unique values would confirm this for the entire dataset.

**product\_detail (Top 20 most frequent for brevity)**

* **Unique Values Found (Top 20):**
  + Ethiopia Rg
  + Spicy Eye Opener Chai Lg
  + Dark chocolate Lg
  + Our Old Time Diner Blend Sm
  + Spicy Eye Opener Chai Rg
  + Oatmeal Scone
  + Brazilian Lg
  + Dark chocolate Rg
  + Earl Grey Rg
  + Lemon Grass Rg
  + Hazelnut Biscotti
  + Brazilian Rg
  + Ethiopia Lg
  + Ouro Brasileiro shot
  + Serenity Green Tea Lg
  + Cappuccino
  + Jumbo Savory Scone
  + Colombian Medium Roast Sm
  + Earl Grey Lg
  + Lemon Grass Lg

**Conclusion on product\_detail:** The product\_detail column also appears generally clean based on the top 20 unique values. It uses consistent abbreviations for sizes (e.g., Rg for Regular, Lg for Large, Sm for Small). A full review of all unique values would confirm this for the entire dataset.

**5. Duplicate Rows**

Duplicate rows can artificially inflate metrics like total transactions or revenue.

* **Number of Duplicate Rows (Exact Duplicates):** 0

**Conclusion on Duplicate Rows:** The dataset contains **no exact duplicate rows**. This means each transaction record is unique, which is ideal for analysis.

**Summary of Cleaning Needs:**

Based on this inspection, your dataset is remarkably clean! The primary cleaning tasks required are:

1. **Date and Time Conversion:** Convert transaction\_date and transaction\_time columns to proper datetime objects for easier time-series analysis. You might also want to combine them into a single transaction\_datetime column.

No other significant errors (missing values, structural inconsistencies, or duplicates) were found during this initial inspection.