Cafe Data Cleaning Report

Cleaned By: Vicki Yang

Transaction ID Adjustments

I standardized the Transaction ID column by removing unnecessary prefixes (such as "TXN_"). This ensured IDs were uniform and allowed for accurate duplicate detection.

Duplicate Records

To verify data integrity, I grouped records by Transaction ID and flagged any that appeared more than once. This step helped confirm that each transaction was represented only once in the dataset.

Item Column

The Item column contained invalid entries such as blanks, "UNKNOWN", and "ERROR". These rows were removed from the dataset to maintain reliable item-level information.

Quantity, Price Per Unit, and Total Spent

The Quantity, Price Per Unit, and Total Spent columns were checked for consistency. Some Total Spent values were missing, invalid, or did not match the calculation (Quantity * Price Per Unit). These cases were corrected by recalculating and updating Total Spent to ensure accuracy.

Payment Method

The Payment Method column included irregular entries (blank, "ERROR"). These were standardized by replacing them with "UNKNOWN". This preserved records while ensuring the column contained only valid categories.

Location Column

The Location field had some missing or invalid entries. These were standardized to "UNKNOWN" for consistency, ensuring all rows had a valid location reference.

Transaction Date Column

The Transaction Date column included invalid formats and placeholder values. I identified entries not matching the standard YYYY-MM-DD format and replaced blanks and "ERROR" with "UNKNOWN". This step ensured that the column could be reliably used for time-based reporting.