

Cafe Sales Data Cleaning Report

1. Introduction

This report summarizes the data cleaning process applied to the Cafe Sales Dataset. The dataset originally contained missing values, duplicate entries, inconsistent formatting, and incorrect data types. The cleaning process was performed using Pandas and NumPy to ensure data accuracy and consistency.

2. Data Issues Identified

- **Missing Values:** Detected in columns such as *Items, price per unit, quantity etc.*
- **Duplicate Records:** No duplicate rows were found
- **Incorrect Data Types:** Issues found in price per unit and items **Inconsistent Formatting:** Irregular capitalization and extra spaces in categorical columns

3. Cleaning Steps Performed

3.1 Handling Missing Values

- Filled missing values in **item** with “**Unknown**”
- Dropped rows where critical data was missing

3.2 Checked for Duplicates

- No duplicate rows were found, hence no need to remove duplicates

3.3 Standardizing Data Formats

- Converted column names to lowercase and removed spaces
- Converted **Price per unit** to a numeric format
- Standardized date format in **Transaction Date**

3.4 Fixing Inconsistent Data

- Trimmed extra spaces and corrected text capitalization

- Mapped similar categories to a standard format

4. Final Data Quality Check

- **Total Rows After Cleaning:** *10000*
- **Total Columns:** *9*
- **No missing values or duplicates detected**
- **Data types correctly formatted**

5. Conclusion

The data cleaning process improved the dataset's reliability for analysis by addressing missing values, duplicates, incorrect formatting, and invalid entries. This cleaned dataset can now be used for sales trend analysis, customer insights, and business forecasting.