

Data Quality Report

This report outlines the data cleaning methodology applied to the dataset. The primary focus was on ensuring data quality by addressing missing values, removing duplicate value, incorrect data types, and outliers.

Data Cleaning Methodology:

1. Missing Values:

The dataset was initially checked for missing values. It was found that there were no missing entries, which is a positive indicator of data quality.

2. Duplicate values:

There is no duplicate value found in the data, which means we have a clean data ready for analysis.

3. Data Types:

The data types of the columns were examined, revealing that several numeric columns were incorrectly classified as objects.

These columns included 'Unit price', 'Quantity', 'Total', etc. To address this, the relevant columns were converted to numeric types using the 'pd.to_numeric' function with error coercion to handle any non-numeric values.

4. Outlier Detection:

After ensuring the correct data types, outlier detection was performed using the Interquartile Range (IQR) method.

This involved calculating the first (Q1) and third (Q3) quartiles and identifying any values that fell outside the range defined by $Q1 - 1.5 * IQR$ and $Q3 + 1.5 * IQR$.

Conclusion:

The cleaning methodology applied has ensured that the dataset is now in a suitable format for analysis. The next steps will involve further exploratory data analysis and modeling.