**Data Quality Report  
Project Name:** InfoSysTrade Retail Data Analysis  
**Author:** Eze Ifeanyi

**1. Introduction**This report outlines the data quality assessment and preprocessing steps undertaken to ensure the integrity and reliability of the dataset used for the InfoSysTrade Retail Data Analysis. The methodology includes handling missing values, removing duplicates, data transformation, and feature engineering.

**2. Data Cleaning Methodology  
2.1 Duplicate Handling**During initial data inspection, duplicate records were identified in the dataset. To prevent data redundancy and ensure unique transaction records, duplicate rows were checked using the Pandas duplicated() function. A total of **one duplicate transaction** was detected and removed to maintain data accuracy.

**2.2 Missing Value Treatment**The dataset contained missing values across multiple columns. The following strategies were applied:

* **Low Occurrence Missing Values:** Since missing values were randomly distributed and constituted a minimal percentage of the dataset, they were removed to maintain statistical integrity.
* **Impact Analysis:** The removal of these values was assessed to ensure no significant impact on key trends and patterns.

**2.3 Feature Engineering & Transformation**Several transformations were performed to enhance the dataset’s analytical potential:

* **Date Conversion:** The "Date" column was converted into datetime format for time-based analysis.
* **Month Extraction:** A new feature, **“Month”**, was created by extracting month names from the Date column, enabling insights into seasonal trends and customer purchasing behaviors.

**2.4 Data Consistency & Formatting**

* Categorical variables were standardized to maintain uniform representation.
* Numeric fields were normalized where necessary to align with industry best practices.
* Inconsistent data formats were rectified to ensure seamless integration with modeling techniques.

**3. Data Integrity Assessment**Post-cleaning, a thorough validation process was performed:

* **Statistical Summary Checks:** Ensured logical consistency in numerical data distributions.
* **Data Type Verification:** Confirmed correct data types for all features.
* **Outlier Detection:** Examined extreme values to confirm relevance.  
  These steps ensured a structured, clean dataset, ready for **Exploratory Data Analysis (EDA)** and **Predictive Modeling**.

**4. Conclusion**The applied cleaning methodology significantly improved data quality, reducing errors and inconsistencies while preserving critical business insights. The refined dataset is now well-structured for advanced analysis and model development.