Data Cleaning Process Report

Overview

The data cleaning process ensured the accuracy, consistency, and usability of the dataset for reliable analysis. Cleaning focused on four interconnected tables—**customers**, **location**, **orders**, and **products**—addressing issues such as duplicates, null values, formatting inconsistencies, and referential integrity. These steps provided a robust foundation for generating actionable insights.

Cleaning Steps by Table

1. Customers Table

- **Trimming and Standardization**:
  - Removed extra spaces from `customer_id` and `customer_name`.
  - Standardized `customer_name` to uppercase for uniformity.
- **Handling Null Values**:
  - Replaced null or blank `customer_name` values with **'UNKNOWN'** to maintain completeness.
- **Duplicate Management**:
  - Identified and resolved duplicate `customer_id` entries using SQL. Retained the first occurrence for consistency.
- **Validation**:
  - Verified the removal of duplicate IDs.
  - Ensured `customer_id` values were linked to corresponding `orders`.

2. Location Table

- **Trimming and Case Consistency**:
  - Standardized fields:
    - `city` and `region`: lowercase.
    - `state` and `country`: uppercase.
- **Handling Null Values**:
  - Flagged null values in `city` and `state` for future resolution without impacting analysis.
- **Validation**:
  - Confirmed the absence of blanks in critical fields like `postal_code`.

3. Orders Table

- **Trimming and Standardization**:
  - Standardized `ship_mode` and `segment` fields to lowercase for consistency.
- **Date Formatting**:
  - Converted `order_date` and `ship_date` to MySQL DATE format (YYYY-MM-DD).
- **Duplicate Management**:
  - Validated that duplicate `customer_id` and `product_id` entries represented legitimate transactions.
- **Rounding**:
  - Rounded `sales` and `profit` values to two decimal places for precision.
- **Validation**:
  - Ensured all `customer_id` and `product_id` values had corresponding records in the `customers` and `products` tables, respectively.

4. Products Table

- **Trimming and Standardization**:
  - Standardized fields:
    - `category` and `product_name`: uppercase.
    - `sub_category`: lowercase.
- **Duplicate Management**:
  - Retained rows with the most complete data, such as the longest `product_description`.
- **Validation**:
  - Verified the integrity of `product_id` values and ensured linkage to the `orders` table.

Referential Integrity

- **Cross-Table Validation**:
  - Confirmed that all `customer_id` and `product_id` values in the `orders` table matched records in the `customers` and `products` tables.
- **Orphaned Records**:
  - Removed `orders` with missing `customer_id` or `product_id` to maintain integrity.

Outcomes and Impact

1. **Duplicates Eliminated**: Ensured unique records across all tables.
2. **Data Standardized**: Text and date fields were cleaned and made uniform.
3. **Null Values Resolved**: Key fields were completed for better accuracy.
4. **Integrity Ensured**: Cross-table relationships were validated, enabling seamless analysis.