

Dear Sprocket Central Pty Ltd Team,

I trust this message finds you in good health and high spirits. Following a meticulous review of the datasets provided, I have identified several data quality issues that warrant your attention. Below, I outline these findings and propose specific recommendations to bolster the data's integrity and utility.

Customer Demographic Dataset:

- **Missing Data:** Fields such as last name, date of birth, job title, job industry, default status, and tenure duration exhibit missing data. To rectify this:
 - Replace missing 'last name,' 'job title,' 'job industry,' and 'default' values with 'unknown'.
 - Impute missing 'date of birth' and 'tenure duration' with median values.
- **Inconsistent Gender Labels:** The gender field contains multiple labels (Female, Male, F, Femal, M, U). To ensure consistency:
 - Standardize 'Female,' 'Femal,' and 'F' to 'Female'.
 - Standardize 'Male' and 'M' to 'Male'.
 - Address 'U' as 'Unknown'.
- **Outdated DOB Record:** One entry has an unrealistic DOB of 1843. This will be updated to the median DOB value across the dataset.
- **Default Field Inconsistencies:** This field includes non-meaningful numeric and non-numeric characters. All nonsensical entries will be replaced with 'unknown'.
- **Recommendations:**
 - Implement mandatory fields in data entry forms to prevent null values.
 - Utilize dropdown menus with predefined categories in data collection.
 - Set validation rules to verify that dates and other inputs are within a reasonable and expected range.

Customer Address Dataset:

- **Customer ID Discrepancies:** The dataset is missing entries for customer IDs 3, 10, 22, and 23, which are present in the Customer Demographic dataset.
- **State Field Inconsistency:** Variations such as "NSW" and "New South Wales," "VIC," and "Victoria" are present. These will be mapped to their respective abbreviations for uniformity.
- **Recommendations:**
 - Introduce referential integrity checks within your databases to ensure that all customer records are complete and accurately linked.
 - Use dropdown options for fields like state and country.

Transactions Dataset:

- **Missing Online Order Data:** The 'online_order' column is missing 360 values. These will be removed from the dataset.
- **Product Information Gaps:** Columns 'brand,' 'product_line,' 'product_class,' 'product_size,' 'standard_cost,' and 'product_first_sold_date' have 197 missing entries, all linked to a product ID of 0.
- **Product First Sold Date Format:** This column currently contains numerical values rather than dates. The format will be corrected to reflect accurate date information.

- **Recommendations:**

- Review the product information management system to ensure a fail-safe process for capturing complete product details.
- Apply a non-nullable constraint to critical fields in your database.

The above issues have been systematically addressed through strategic imputation, standardization of labels, and mapping of state abbreviations, enhancing the dataset's reliability for subsequent analyses.

Please feel free to reach out for any clarifications or further assistance with the implementation of these data quality improvements.

Warm regards,
Deepak Bisht