TO: Sprocket Central Pty Ltd

Subject: Data Quality Framework Table

Dear Sir/Madam,

Below are identified data quality issues and strategies to mitigate these issues:

Sheet name	Unique columns	Number of rows before cleaning	Number of rows after cleaning
Transactions	*transaction_id*	20000	19946
NewCutomerList	N/A (since it is the target sheet)	1000	852
CustomerDemographic	*customer_id*	4000	3308
CustomerAddress	*customer_id*	4000	4000

1. Accuracy (Correct values)

- *Transactions* → *product_first_sold_date* ⇒ Change data type from number to date

Mitigate step:

The Data Handling team should check the columns' data type to mitigate this issue.

2. Completeness (Data fields with values)

- *Transactions* → *online_order*, *brand*, *product_line*, *product_class*, *product_line*, *standard_cost*, and *product_first_sold_date* contain null values.

- *NewCustomerList* → *last_name*, *DOB*, *job_title* contain null values.

- *CustomerDemographic* → *last_name*, *DOB*, *job_title*, *default*, *tenure* contain null values.

Mitigate step:

The Data Handling team can tally the old data or relevant records to fill in some null values.

3. Consistence (Values free from contradiction)

- *CustomerDemographic* → *gender* → Change “F” and “Femal” to “Female” and “M” to “Male”.

- *CustomerAddress* → *state* → Change “New South Wales” to “NSW” and “Victoria” to “VIC”.

Mitigate step:

The Data Handling team check the value format to have all the value in a consistent form.

4. Currency (Values up to Date)

- *CustomerDemographic* → *DOB* contains the meaningless value “1843-12-21.”

Mitigate step:

The Data Handling team should check the latest value to keep the value up to date.

5. Relevancy (Data Items with value Meta-data)

- NewCustomerList → *gender* → Check value “U” → Delete rows with this value due to the inconsistency of whether it represents “Male” or “Female.”
- CustomerDemographics → *gender* → Check value “U” → Delete rows with this value due to the inconsistency of whether it represents “Male” or “Female.”
- CustomerDemographics → *deceased_indicator* → Deleted value "Y" because the analysis wants only alive customers
- CustomerDemographics → *default* → Deleted this column since it is not decodable.

Mitigate step:

The Data Handling team should tally the record of "U" in gender to find the correct values.

6. Validity (Data containing allowable values)

No issue from the database relevant to this aspects

7. Uniqueness (Records that are duplicates)

No issue from the database relevant to this aspects

Moving forward, the team will continue with the data cleaning, standardization and transformation process for the purpose of model analysis. Questions will be raised along the way, and assumptions will be documented. After we have completed this, it would be great to spend some time with your data SME to ensure that all assumptions are aligned with Sprocket Central’s understanding.

Kind regards,

Theo Nguyen.