To,
Sprocket Central Pty Ltd

Subject - Data Quality issues and strategies to mitigate these issues.

Dear Sir/Madam,
Thank you for providing us with four datasets.Below is a table that highlights the summary statistics from the four datasets received.

Table name	Unique column	No. of Rows	Date data recieved
Transactions	transaction_id	2000	25-08-2023
NewCustomerList	N/A	1000	25-08-2023
CustomerDemographic	customer_id	4000	25-08-2023
CustomerAddress	customer_id	4003	25-08-2023

Notable data quality issues were encountered and the methods used to mitigate the identified data inconsistencies are as follows. Furthermore, recommendations have been provided to avoid the recurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions

1. **Accuracy (Values not in the correct format)**

Missing a profit column in “Transactions”; DOB is inaccurate in “Customer Demographic” and missing Age column.
Mitigation: Filter out outlier in DOB,
Recommendations: Create profit column in “Transactions” to check accuracy of sales. Create Age column to check errors more accurately in “Customer Demographic”.

2. Completeness (Fields with missing values):

Transactions sheet → *online_order, brand, product_line, product_class, product_size, standard cost, product first sold date* contain null values.
NewCustomerList → *DOB, job title, job industry* contain null values.
CustomerDemographic → *DOB, job title, job industry* contain null values.
Mitigation: Filter out blanks from columns
Recommendations: Provide dropdown options for columns missing values

3. Consistency Issues
Inconsistency in gender for “Customer Address” and “Customer Demographic” respectively.
Mitigation:: Use regular expression to replace extended values into abbreviations to ensure consistency across by filtering all ‘M’ under ‘Male’ and all ‘Femal’ and ‘F’ under ‘Female’ for gender. Filter all ‘New South Whale’ to ‘NSW’ and ‘Victoria’ to ‘VIC’ for states.
Recommendation: Enforce a drop-down list for the user entering the data rather than a free text field.

4. Currency
People that are ‘Y’ in deceased_indicator for “Customer Demographic” are not current customers
Mitigation: filter out customers which are checked as ‘Y’ in “Customer Demographic”
Recommendations: When information is received, one must update data.

5.Relevancy
Lack of relevancy in default column for “Customer Demographic” and NewCustomerList
Mitigation: filter out customers which are checked as ‘Y’ in “Customer Demographic” because want only alive customers
Filter out customers with “U” because don’t know whether it is male or female,
Recommendations: Check for incomprehensible Metadata and delete or format to make comprehensible.

6. Validity
Format of list_price; product_sold_date for “Transaction”
Mitigations: Format product_sold_date to short data format; format list_price to currency
Recommendations: Set up columns so that formats such as price and decimals are already in place when entering new data

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 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,

Lavender Echessa.