Dear Sprocket Central Pty Ltd,

Thank you for providing us with the four datasets from Sprocket Central Pty Ltd. The below table highlights the summary statistics from the four datasets received. Please let us know if the figures are not aligned with you understanding.

Table_Name	No_Of_Records	Distinct_Customers	Date_Data_Received
Customer Demographic	4_000	4_000	13-07-2023
New Customer List	1_000	1_000	13-07-2023
Customer Address	3_999	3_999	13-07-2023
Transactions	20_000	20_000	13-07-2023

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

1) Empty values

- In “Customer Demographic” dataset having empty values in various columns like “last_name, DOB, job_title, tenure”.
- In ”New Customer List” dataset having empty records in columns like “last_name, DOB, job_title”.
- In "Transactions" dataset having empty records in various columns like "Online_order, brand, product_line, product_class, product_size, standard_cost, product_first_sold_date".

- *Mitigation*: If only a small number of rows are empty, filter out the record entirely from the training set for prediction. Else, if it is a core field, impute based on distribution in the training dataset.

2) Invalid count of records in data sets

- Additional cutomer_id's or Improper/missing Customer_id's are in "Customer Address".
- "Cutomer Demographic" is of 4000 records and "New Customer List" is of 1000, totally we are having 5000 active customers.But not accurate/matching count of records found in "Customer Address" i.e., of 3999 and "Transactions" with of 20000 records.

Mitigation: Please ensure that all tables are from the same period. Only customers in the Customer Master list will be used as a training set for our model. This indicates that the data received may not be in sync with each other which may skew the analysis results if there are missing data records.

Please refer to excel file ‘sprocket_data_outliers.xlsx’ for the list of outliers between tables.

3) In consistent values for the same attributes

e.g. “Female” attribute represented as "F", "Femal", and “Male” represented as "M"

- In "Cutomer Demographic" and "New Customer List" the column "job_industry_category" having value "n/a" which is not a valid data for analyzation.

Mitigation: Use regular expression to replaced extended values into abbreviations to ensure consistency across addresses.

Recommendation: Enforce a drop-down list for the user entering the data rather than a free text field. In order to construct meaningful variables for the model, the data has been cleaned to avoid multiple representations of the same value. Additionally, gender records where 'U' have been replaced based on the distribution from the training dataset.

4) Inconsistent data type for the same attribute

(e.g : combination of integers and string values in same column)

- In "New Customer List" columns of "postcode, property_valuation, past_3_years_bike_related_purchases" having such combinations.
- In "Transactions" the column "product_first_sold_date" is of invalid data type of date type.
- In "Customer Demographic" column "default" is of undefined data type/garbage values.

Mitigation: Convert selected records in characters to numeric. Remove non-numeric characters from string. *Recommendation:* Ensure that fact tables in the given database have constraints on data types. Having different data types for a given field make it difficult to interpret results at the later stage. Therefore, appropriate data transformations are made to ensure consistent data types for a given field.

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,

Ganesh Reddy.