Dear [Name of the client],

Hope you are doing well!

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. To confirm that the data we have received matches with yours, we compiled a short summary of the datasets as given below. Please let us know if the statistics are not aligned with your understanding.

Table Name	Total No. of Records	No. of distinct Customer ID's
Customer Demographic	50955	4000
Customer Address	23994	3999
Transactions	258458	20000

We have conducted a data quality analysis of the same as per the guidelines of the 'Data Quality Framework Table', and have found a few issues as listed below. Moreover, recommendations have been provided to help mitigate these issues and ensure a consistent data quality to drive future business decisions.

1. Missing values (Completeness)

Problem: There are missing values in columns 'Online Order,' 'Brand Name', 'Product Line', 'Product Class', 'Product Size', 'Standard Cost', and 'product_first_sold_date' for the Transactions spreadsheet. Also, there are records missing in the 'job_title,' 'job_industry_category' and 'tenure' columns of the Customer Demographic spreadsheet.

Recommendation: The number of missing values is very small compared to the total number of values in a column, so we can simply delete these missing records from the respective spreadsheets. If these datapoints are crucial, then impute based on the distribution of the dataset.

2. Inconsistency of the number of values (Consistency)

Problem: The number of 'customer_ids' entries are not consistent across the Transactions, Customer Demographic and Customer Address spreadsheets. The additional number of entries in the Transactions table implies that the data is sourced from different time periods.

Recommendation: The data must be consistent and correctly synced for the analysis to take place. Only the customers in the Customer Demographic table are to be used as a training set.

3. Duplicated Values (Uniqueness)

Problem: There are instances of duplicated values. For example, in the Customer Demographic sheet, there are 3 separate ways in which the female gender is specified viz. 'Female', 'Femal', and 'F' and 2 separate ways the male gender is specified viz. 'Male' and 'M'. Moreover, in the Customer Address sheet, there are duplicated values in the 'state' column such as 'New South Wales' & 'NSW', and 'Victoria' & 'VIC'.

Recommendation: The genders can be specified using the abbreviated norms ‘M’ and ‘F’ for the male and female genders respectively; while the states can also be represented using the abbreviated forms ‘NSW’, ‘QLD’ and ‘VIC’. To ensure uniformity, have a drop-down list rather than a text field.

4. Invalid Values (Validity)

Problem: There are instances of an unspecified gender ‘U’ present in both the New Customer List sheet as well as the Customer Demographic sheet – these are invalid values.

Recommendation: The gender values called ‘U’ can be randomly given either the ‘M’ or ‘F’ values keeping the distribution of those values in the dataset in mind.

5. Incorrect values (Accuracy)

Problem: There an obviously incorrect values such as the DOB value of ‘1843-12-21’ in the Customer Demographic sheet. This would make the customer’s age 180 years in 2023 which is impossible

Recommendation: Since such incorrect values are rarely found in the data, they can be safely deleted without it affecting the overall analysis of the dataset.

6. Inconsistent Data Types for the same attribute

Problem: For same (and similar) attributes across the tables, different data types such as numeric and text data have been used. This can make it difficult to analyze and interpret results at a later stage.

Recommendation: Convert all such data to numeric and remove all the non-numeric characters from the string.

Kindly investigate the above data-quality issues along with the recommendations to counter them to ensure a consistent quality of the dataset across all the spreadsheets. Moving forward, our team will continue the ETL process and document our assumptions and observations. It would be great to have a meeting scheduled with your data team to ensure smooth operability in aligned with Sprocket Central’s vision and understanding.

Regards,
Rohan Deo