Dear Manager,

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

Summary	Customer Address Table	Customer Demographics Table	New Customer List Table	Transactions Table
Number of Records	This table has 3999 records and 6 features or columns ranging from [customer id] to [property valuation]	The table has 4000 records and 13 columns	1000 records and 23 columns	This table has 20000 records and 13 columns
Distinct Customer Ids	3999 distinct customer ids	There are 4000 distinct customer ids	None	20000 distinct transaction ids
Date Data Received	05/05/2024	05/05/2024	05/05/2024	05/05/2024

The team has noted several data quality issues. The section highlights the issues encountered, the methods used to mitigate the identified data inconsistencies as well as recommendations to avoid the reoccurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions as follows.

1. Inconsistent values for the same attribute

(e.g., New South Wales represented as “New South Wales”, and “NSW”, Victoria being represented as “VIC”, and “Victoria” in Customer Address data set

Female represented as “Female”, “Femal”, and “F”, Male being represented as “Male”, and “M” in Customer Demographic data set)

Mitigation: Use regular expression to replace 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.

2. Inconsistent data type for the same attribute

(e.g., numeric values for some fields and strings for others, some dates stored as numeric)

Mitigation: Convert selected records in characters to numeric. Remove non-numeric characters from string. Convert dates in characters and numeric to date type.

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.

3. **Missing values: Various columns have empty values in certain record**

(e.g., last name is 3% empty, DOB is 2% empty, job title 12% empty, tenue 2% empty, online order is 2% empty)

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.

For key datasets, such as transactions, less than 1% of transactions have missing fields. These records have been removed from the training dataset.

4. **Relevancy issues: Various columns, default, and 5 columns in Customer List table**

(e.g., in demographic data set, default column is irrelevant to our model, in New Customer List table, there are 5 columns without descriptions or column headers)

Mitigation: The default column provides unreadable data, so it would be irrelevant for our model. It's been removed.

Also, the columns mentioned above provide numeric record for unspecified features. It's been removed.

However, they may be referring to a very important metric within your organization. Consider providing additional description for the column headers.

5. **Inconsistent records, additional customer ids in the Transactions table and Customer Address table but not in Customer Demographic. New Customer List table has no Customer ids)**

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, indicating that the data received may not be in sync with each other which may skew the analysis results if there are missing data records.

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,

David Owino

Data Analyst