DATA QUALITY ASSESSMENT REPORT

Dear Client,

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

Dataset Name	No. of Records	Distinct Customer IDs	Data Recorded Date
Customer Address	• 3999 rows • 26 columns: 6 with data, 20 empty columns.	3999	13/02/2023
Customer Demographic	• 4000 rows • 26 columns: 13 with data, 13 empty columns.	4000	13/02/2023
Transaction Data	• 20000 rows • 26 columns: 13 with data, 13 empty columns.	3494	13/02/2023

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 recurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

- Less Number of Customer IDs in the Customer address and Transaction Data.
  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.
- Various columns, such as brand, DOB, and Job title, have empty values in certain records.
  Mitigation: If only a small number of rows are empty, filter out the record entirely from the training set. Else, it is a core field, impute based on the distribution in the training dataset.
  For key datasets such as transactions, less than 1% of transactions (totaling less than 0.1% of revenue) have missing fields. These records have been removed from the training dataset.
- Inconsistent values for the same attribute (e.g., Victoria being represented as “V”, “Vic”, and “Victoria “)
  Mitigation: Use regular expressions to replace extended values with 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.

- Inconsistent data type for the same attribute
  (e.g., numeric values for some fields and strings for others)
  Mitigation: Convert selected records in characters to numeric. Remove non-numeric characters from a string.
  Recommendation: Ensure that fact tables in the given field make it difficult to interpret results at a later stage. Therefore, appropriate data transformations are made to ensure consistent data types for a given field.

  Moving forward, the team will continue with data cleaning, standardization, and transformation 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,
  Alberta Adu