Dear Client,

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

Table Name	**No. of records**	**Distinct CustomerIDs**	**Date Data Recieved**
Customer Demographic	4000	4000	1 – 1 – 2021
Customer Address	3999	3999	1 – 1 – 2021
Transaction Data	20000	20000	1 – 1 – 2021

Some data quality issues that were encountered and the methods used to mitigate the identified data inconsistencies are as follows:

- **Missing Values in various columns of "Customer Demographics" and "Transactions" table.**
  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 (totalling less than 0.1% of revenue) have missing fields. These records have been removed from the training dataset.

- **Additional customer ids in "Transactions" (5004) and "Customer Address" (4001, 4002, 4003) but not in "Customer Demographic".**
  Mitigation: Please ensure that all tables are from the same period. Only customers in the Customer Demographic will be used as a training set.
  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.

- **Inconsistent values for the same attribute (For Example: New South Wales and Victoria is also represented as "NSW" and "VIC" respectively).**
  Mitigation: Use regular expression to replaced extended values into abbreviations to ensure consistency across State names
  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.

- **Inconsistent Data types for the same attribute.**
  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, data pre-processing 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 (Subject Matter Expert) to ensure that all assumptions are aligned with Sprocket Central's understanding.

Kind regards,

Shivam Negi