

Dear Client,

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

Name of table	No. of records	Distinct Customer IDs	Date Data Received
Transaction Data	200002	3495	18.10.2022
Customers Demographic	4002	4002	18.10.2022
Customer Address	4001	4001	18.10.2022

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

1. The number of customer IDs in the Transaction Data and Customer Address may be not synced with the Customer Demographic table. This may skew the analysis results if there are missing data records.

Mitigation: Please ensure that all tables are from the same period. Only customers in the Customer Demographic list will be used as an analyzing set for our model.

2. Various columns such as last_name, or 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 set for prediction. Else, if it is a core field, impute based on the distribution in the dataset.

3. Inconsistent values for the same attribute. For example, Victoria is represented as "V", "Vic" and "Victoria"; Female is represented as "F" and "Female".

Mitigation: Use regular expressions to replace extended values with abbreviations to ensure consistency across addresses. Enforce a drop-down list for the user entering the data rather than a free text field. To construct meaningful variables for the model, the data has been cleaned to avoid multiple representations of the same value.

4. Inconsistent data type for the same attribute. For example, strings for numeric values columns.

Mitigation: Convert selected records in characters to numeric. Ensure that fact tables in the given database have constraints on data types.

5. Correctness of values. For example, the date of birth of some customers is incorrect when they are converted into ages.

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.

Have a nice day!

Best regards,

Nhien Chau.