Dear Client,

Thank you for providing us with the four datasets from Sprocket Central Pty Ltd. The below table highlights the summary statistics from the four datasets received.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sheet Name** | **No. of Records** | **Distinct Customer ID** | **No. of Attributes** |
| Transactions | 20000 | 3503 | 13 |
| New Customer List | 1000 | \_ | 18 |
| Customer Demographic | 4000 | 4000 | 13 |
| Customer Address | 4000 | 4000 | 6 |

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 re-occurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

1. **Additional customer\_ids in the ‘Transactions table’ and ‘Customer Address table’ but not in ‘Customer Demographic’**

Please ensure that all tables are from the same period. Only customers in Customer Masterlist 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 if there are missing data records.

1. **Various columns, such as the brand of purchase, or job title, have empty values in certain records**

If only a small number of rows are empty, filter out the record entirely from the training set for prediction. (e.g. some null values in job\_industry\_category and blank values in job\_title.) Else, if it is a core field, impute based on the distribution in the training dataset.

1. **Inconsistent values for the same attribute**

(e.g. Victoria being represented as “V”, “Vic” and “Victoria” and gender attribute shows male/female but contains F/M/U) Use regular expression to replaced extended values into abbreviations to ensure consistency across addresses.

1. **The inconsistent data type for the same attribute**

(e.g. numeric values for some fields and strings for others and ‘default’ attribute has some random now readable values) Convert selected records in characters to numeric. Remove non-numeric characters from a string. Therefore, appropriate data transformations are made to ensure consistent data types for a given field.