Dear Client,

Thank you for providing us with the datasets. The below table highlights the summary statistics of the data. Please let us know if the figures are not aligned with your understanding.

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Set** | **No. of records** | **Distinct customer ids** | **Date - Data Received** |
| Customer demographic | 4000 | 4000 | 13-06-2021 |
| Customer address | 3999 | 3999 | 13-06-2021 |
| Transaction data | 20000 | 3494 | 13-06-2021 |

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

*Note: Furthermore, recommendations have been provided to avoid the reoccurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.*

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

Recommendation: Please ensure that all tables are from the same period. Only customers in the Customer Demographics list will be considered 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.

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

Recommendation: If only a small number of rows are empty, filter out the record entirely. Else, if it is a core field, impute values based on distribution in the considered dataset. For key datasets, such as transactions, less than 1% of transactions 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”).

Recommendation: Use regular expression to replaced extended values into abbreviations to ensure consistency across addresses. 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 type for the same attribute (e.g., numeric values for some fields and strings for others).

Recommendation: Convert selected records in characters to numeric. Remove non-numeric characters from string. 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, standardisation, and transformation process for the purpose of model analysis. After we have completed this, it would be great to spend some time with your data to ensure that all assumptions are aligned with Sprocket Central’s understanding.

Kind regards,

Somya Bhushan