Dear Client,

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. 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.

|  |  |  |
| --- | --- | --- |
| **Table name** | **Data found to be cleaned** | **Received date** |
| Transactions | Duplicate Customer Id and and blank rows. | 03 June 2023 |
| New customer list | Many blank rows found | 03 June 2023 |
| Customer demographic | Many blank rows found | 03 June 2023 |
| Customer Address | Many blank rows found |  |

* **Additional customer\_ids** in the ‘Transactions table’ but not in ‘Customer Master (Customer Demographic)’ 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 if there are missing data records. Please refer to excel file ‘data\_outliers.xlsx’ for the list of outliers between tables.
* **Various columns**, such as the **brand of a purchase, 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 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.
* **Inconsistent values for the same attribute** (e.g. female being represented as “female” and “f”) Mitigation: Use regular expression to replaced extended values into 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

Moving forward, the team will continue with the data cleaning, standardisation 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.

Kind regards,

Tushar Pal