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 | No. of records | Distinct Customer IDs | Date Data Received |
| Customer Demographic | *-insert value-* | *-insert value-* | *-insert value-* |
| Customer Address | *-insert value-* | *-insert value-* | *-insert value-* |
| Transaction Data | *-insert value-* | *-insert value-* | *-insert value-* |

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 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’ and ‘Customer Address table’ but not in ‘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.*
* **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, assign values based on distribution in the dataset.*

* **Inconsistent values for the same attribute (e.g. Victoria being represented as “V”, “Vic” and “Victoria”)**  
  *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.
* **Inconsistent data type for the same attribute (e.g. numeric values for some fields and strings for others)**   
  *Mitigation: Having different data types for a given field make it difficult to interpret results at the later stage.* *Convert selected records in characters to numeric. Remove non-numeric characters from string.*

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,

Chetan Oli

Data Analyst