Dear [Client point-of-contact],

Thank you for providing us with the necessary datasets from Sprocket Central Pty Ltd. The below table shows the summary statistics from the three datasets received. Please, inform us if the data not correct,

|  |  |  |  |
| --- | --- | --- | --- |
| Table name | No. of records | Distinct. Customer IDs | Date Data Received |
| Customer  Demographic | 4000 | 4000 | 03/08/2021 |
| Customer Address | 3999 | 3999 | 03/08/2021 |
| Transaction Data | 20000 | 3494 | 03/08/2021 |

The most important data features, which reflects data inconsistency will be represented below. Our analytics team provide you with the additional information how to eliminate similar problems in the data in the future. This will help you to significantly improve data quality and accuracy.

* [customer\_ids] in the ‘Transactions table’ and ‘Customer Address table’ are not equal with ‘Customer Demographic’ table

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. We made an additional table with the list of missing [customer\_id] for each table. Total unique customer id records are 4004. From this customers in ‘Transaction table’ are 3494, in ‘Demographic’ – 4000 and in ‘Address’ – 3999.

* 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. 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: Convert selected records in characters to numeric. Remove non-numeric characters from string. Recommendation: 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, 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.

Kind regards,

[Junior Consultant Name]