Dear [Client point-of-contact],

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	4000	4000	
Customer Address	3999	3999	
Transactions Data	20000	3494	2017-01-01 ~ 2017-12-30

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 Master (Customer Demographic)`**

  Mitigation: Please ensure that all tables are from the same period. Only customers in customer master list will be used as a training set for our model

  Recommendation: This indicated that the data received may not be in sync with each other which may skew the analysis results of there are missing records. Please refer to excel file “data_outliers.csv” for the list of outliers between tables

- **Various columns, such as online order, the brand of a purchase, product line, product class, product size, standard cost, product first sold date, last name, DOB, job title, job industry category, tenure have empty values in certain records**

  Mitigate: 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.

  Recommendation: For key datasets, such as transactions, less than 1% of transactions (total 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 'F', 'Femal' and 'Female')**

  Mitigate: 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 as missing value.

- **Inconsistent data type for the same attribute (e.g. numeric values for some fields and strings for others)**

  Mitigate: 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 fields 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, standardizing 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,
Murong (Sophie) Cui