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.

|  |  |  |  |
| --- | --- | --- | --- |
| Table Name | No. of Records | Distinct Customer IDs | Date of Data Received |
| Transaction | 20000 | 3494 | NA |
| Customer Demographic | 4000 | 4000 |
| Customer Address | 3999 | 3999 |

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.

* **More customer\_ids in the “transaction” table (32 additional ids) and “customer address” table (3 additional ids) than in the “Customer Master (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. Please refer to excel file ‘data\_outliers.xlsx’ for the list of outliers between tables.

* **Varies columns have empty values in certain records.**
  + In the transaction sheet, columns are: online\_order, brand, product\_line, product\_class, product\_size, standard\_cost and product\_first\_sold\_date
  + In the customer\_demographic sheet: DOB, job\_title\_default, tenure

*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.**
  + In the customer\_demographic sheet: gender for F, Femal, Male, U, Female, M
  + In the customer\_address sheet: State has VIC and Victoris, NSW and New South Wales, QLD

*Mitigation: Use regular expression to replace 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 same attribute.**
  + transaction sheet has numbers for product\_first\_sold\_date and customer\_demographic has string for DOB, which are both dates

*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.

* **Illogical data values.**
  + In the customer\_demographic sheet, a customer with id=34 was born in 1843

*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.*

Recommendation: Enforce a select range for the user entering the data rather than a free selection. Since only one row was spotted with this problem, this row will be dropped from analyzing later.

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,

[Junior Consultant Name]