Dear Client,

I am writing to you in regards to the three datasets we have received from Sprocket Central Pty Ltd. to share our conclusions of data quality assessment. The below table highlights the summary statistics from the datasets provided. Please let us know if the figures are not aligned with your understanding.

|  |  |  |
| --- | --- | --- |
| Table Name | No. of record | Distinct Customer IDs |
| Customer Demographic | 4000 | 4000 |
| Customer Address | 3999 | 3999 |
| Transaction Data | 20000 | 3494 |

Significant data quality issues and methods used to mitigate them are listed below. Additionally, recommendations have been provided to avoid the reoccurrence of data quality issues and to ensure data quality for business decisions.

* **Product\_ids refer to products of multiple brands, product\_lines etc.**

*Mitigation: Omit the product\_id records in the training set.*

*Recommendation: Create and link the product database to the product related fields.* Drop-down lists must be enforced on the fields like product\_id, brand, product\_line, product\_class, and product\_size, limiting the field choices bounded by the product database.

* **Customer\_ids present in the ‘Transactions table’ and ‘Customer Address table’ but missing in ‘Customer Demographic’**

*Mitigation: Please ensure that all the tables are synchronized. Only records from ‘Customer Demographic’ will be used for the model training.*

The data provided might not be synchronized, which consequently might skew the results. Please refer to the file ‘data\_outliers.xlsx’ for the list of outliers between tables.

* **Various columns, such as brand, product\_size, or job\_title, have missing values**

*Mitigation: For columns with small number of rows with missing data, omit them in the training set. Otherwise, if it is a core field, impute the missing values based on the distribution.*

Less than 1% of transactions (totaling less than 0.1% of revenue) in the ‘Transactions’ dataset has missing values – those records have been removed from the training set.

* **Inconsistent values for the same attribute (e.g., Victoria being represented as “V,” “VIC,” and “Victoria”)**

*Mitigation: Replaced the values to an abbreviation to keep the records consistent*

*Recommendation: Enforce a drop-down list for the user entering the data.*

Moving forward, the team will continue with data cleaning, standardization, and transformation processes for model analysis. Questions, concerns, and assumptions will be documented along the way. After it has been completed, it would be great to spend some time with your data SME to ensure our assumptions are aligned with Sprocket Central’s understanding.

Best regards,

Yegor Ganin