Gmail **Islam Omar <islamomar27@gmail.com>**

Data Quality Assessment

Islam Omar <islamomar27@gmail.com> Tue, Aug 15, 2023 at 12:58 AM
To: Islam Omar <islamomar27@gmail.com>

Dear [Client],

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd.
After the Analysis for the data, We encountered some data quality issues. The methods used to mitigate this identified data inconsistencies have been provided.Furthermore, recommendations have been provided to avoid the recurrence of these data quality issues and improve the quality of the data used to drive business decisions.

- **Number of customer_ids in the 'Transactions table' and 'Customer Address table' is less than that in 'Customer Master (Customer Demographic)'**
  Mitigation: Only customers in the Transaction table and in Customer Master and in the 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.

- **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. Female being represented as "F", "Femal" and "Female")**
  Mitigation: Replace extended values into standardized values to ensure consistency across gender and addresses, during the model development process.
  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 in the columns default and product_first_sold_date (e.g. values in the product first sold date are float numbers not in the date format)**
  Mitigation: These columns have been removed due to the unavailability to make data transformation to ensure consistent data types for a given field
  Recommendation: Ensure that fact tables in the given database have constraints on data types.
  Having different data types for a given field makes it difficult to interpret results at the later stage. Therefore, appropriate data transformations are made to ensure consistent data types for a given field.

We appreciate the opportunity to collaborate on this data analysis project. Please let us know if you require any clarification or have additional questions as we progress with the analysis. We look forward to continuing our productive working relationship as we leverage the data to uncover actionable insights for your organization.

Regards,
Islam Omar