Subject: Data Quality Assessment and Enhancement Plan for Sprocket Central Pty Ltd

Dear [Client Point-of-Contact],

I trust this message finds you in good health. We are grateful for the opportunity to partner with Sprocket Central Pty Ltd and are fully committed to elevating your data quality standards, thereby facilitating more precise and insightful analyses.

Our comprehensive data quality assessment has uncovered specific issues within your datasets that warrant immediate attention. Here is a detailed breakdown of these challenges, along with our proposed solutions and recommendations aimed at preventing similar issues in the future:

Problem 1: Completeness

- Issue: Our analysis has revealed missing values across several columns in the datasets, including but not limited to `online_order`, `brand`, `product_line`, `product_class`, `product_size`, `standard_cost`, `product_first_sold_date`, `last_name`, `DOB`, `job_title`, `job_industry_category`, `default`, and `tenure`.
- Mitigation: We recommend the removal of records containing missing values, which accounts for approximately **2.28%** in CustomerDemographic (CustomerDemographic + CustomerAddress), **0.98%** in Transactions, and **1.7%** in NewCustomerList. In cases where missing values are numeric or relate to date of birth (DOB), we suggest employing imputation techniques while marking missing categorical and text data as 'no data.' We maintain data type integrity during this process.
- Recommendation: Establish robust data validation checks at the point of data entry to ensure completeness in future submissions. Clear data entry guidelines should be implemented to minimize instances of missing values.

Problem 2: Validity

- Issue: Inconsistencies in data types have been identified, necessitating conversions such as transforming `tenure` to an integer, `product_first_sold_date` to a datetime format, and several columns (`online_order`, `order_status`, `brand`, `product_line`, `product_class`, `product_size`) to the category data type for memory optimization and consistency. Addressing 'n/a' values in the `job_industry_category` column is also crucial.
- Mitigation: We propose executing data type conversions for the designated columns, ensuring data consistency. Additionally, we recommend removing the `default` column due to its heterogeneous data types.
- Recommendation: Establish uniform data type standards for various columns and enforce data constraints to prevent future data type-related issues. Maintain a standardized data dictionary as a reference point and encourage the use of drop-down lists, rather than free-text fields, for categorical data entry.

Problem 3: Accuracy

- Issue: Discrepancies have been observed in customer IDs between tables, with customer_id 5034 in the 'Transactions' table and customer_ids 4001, 4002, 4003 in the 'Customer Address' table not present in the 'Customer Demographic' table.
- Mitigation: To preserve data accuracy, we recommend aligning all tables to the same reference period and restricting analyses to customers listed in the 'Customer Demographic' table.
- Recommendation: The disparities indicate potential data synchronization issues, which could skew analysis results due to missing data records.

Problem 4: Consistency

- Issue: Inconsistencies have been noted in the representation of attributes, such as the `gender` and `state` columns, which sometimes contain full names and abbreviations across the tables.
- Mitigation: To ensure data consistency, we propose standardizing values using regular expressions.
- Recommendation: Provide clear data entry guidelines that include standardized naming conventions for attributes to maintain consistency in future data submissions. Enforce the use of drop-down lists for data entry to further enhance uniformity.

Problem 5: Relationship Between Tables

- Issue: The 'NewCustomerAddress' table lacks any meaningful relationship with the 'Transactions' table and lacks a unique identifier like a customer ID.
- Mitigation: We recommend excluding this table from our analysis as it does not contribute to our fact table or yield valuable insights.
- Recommendation: Assign a unique customer ID to every new customer for precise identification and future analyses.

In summary, we have outlined these data quality issues, along with corresponding mitigation strategies and recommendations, to ensure their effective resolution. Our dedicated team is committed to data cleaning, standardization, and transformation to prepare the data for insightful analysis. Should any questions or clarifications arise during this process, we will meticulously document them and seek your valuable input.

Our efforts have also yielded significant memory savings, reducing data storage requirements by **34.24%** in CustomerDemographic (CustomerDemographic + CustomerAddress), **35%** in Transactions, and **40.71%** in NewCustomerList tables.

As we move forward, we would greatly appreciate the opportunity to collaborate with your data Subject Matter Expert (SME) to align with Sprocket Central's data quality standards and validate the improvements made.

Thank you for entrusting our team with this endeavor. We eagerly anticipate the opportunity to enhance your data quality and analysis capabilities.

Best regards,
Rameez Khan
Data Analyst
KPMG Analytics Team