Subject: Addressing Data Quality Issues and Proposed Strategies for Sprocket Central Pty Ltd

Dear sir/mam,

I hope this email finds you well. I wanted to bring to your attention some data quality issues we have identified in three different datasets: Customer Demographic, Customer Address, and Transaction. Additionally, I would like to propose strategies to mitigate these issues and ensure the accuracy and reliability of your data.

Customer Demographic Dataset:

1. "Gender" Column: There are several inconsistencies in the gender categories, including "U - 88," "M - 1," "Femal - 1," and "F - 1." To address this, we recommend standardizing the gender categories to "Male," "Female," and "Unknown." This will ensure consistent and meaningful analysis of customer demographics.

2. "Last\_name" Column: We have identified 125 missing values in the "Last\_name" column. It is essential to capture complete customer information for effective personalization and identification. We suggest implementing a data validation process during data entry to reduce missing values in this column.

3. "DOB" Column: There are 87 missing values in the "DOB" column, along with an entry for "1843-12-21," which indicates an individual with an age of 180 years. We recommend conducting data cleansing to fill in missing values using techniques such as imputation based on existing customer data. Additionally, the erroneous entry for the date of birth should be investigated and corrected.

4. "Job\_title" Column: We have identified 507 missing values in the "Job\_title" column. It is crucial to have this information for better customer segmentation and targeting. We suggest implementing mandatory field requirements during data collection to minimize missing values.

5. "Job\_industry\_category" Column: There are 657 missing values in the "Job\_industry\_category" column. This information can provide valuable insights into customer preferences and behavior. We propose using data imputation techniques based on existing customer data to fill in the missing values.

6. "Default" Column: The "Default" column is unclear and does not provide any actionable insights. Moreover, there are 302 missing values in this column. We recommend conducting a thorough analysis of the column's relevance and, if necessary, consider removing it from the dataset.

7. "Tenure" Column: We have identified 87 missing values in the "Tenure" column. Tenure information is essential for understanding customer loyalty and retention. We suggest capturing this information consistently during customer onboarding and implementing data validation processes to minimize missing values.

Customer Address Dataset:

1. "STATE" Column: Inconsistencies exist in the representation of states. "New South Wales" is represented as "NSW" in several rows, and "Victoria" is represented as "VIC" in multiple rows. We recommend standardizing the state abbreviations to ensure consistent and accurate reporting.

2. "QLD" Representation: There is a question regarding the representation of "QLD" in the dataset. It is unclear whether it stands for "Queensland." We suggest verifying the intended meaning and updating the column accordingly.

Transaction Dataset:

1. "Product\_first\_sold\_date" Column: The date format in the "Product\_first\_sold\_date" column is incorrect. We recommend transforming the date format to ensure consistency and compatibility with the intended analysis. Additionally, there are 197 missing values in this column that need to be addressed through data imputation or investigation.

2. Missing Values: Several columns in the Transaction dataset have missing values, including "online\_order," "brand," "product\_line," "product\_class," "product\_size," "standard\_cost," and "product\_first\_sold\_date." We suggest conducting data imputation using appropriate techniques to fill in these missing values and ensure a complete dataset for analysis.

To mitigate these data quality issues and ensure accurate insights, we propose the following strategies:

1. Implement standardized data collection processes with mandatory fields to minimize missing values.

2. Conduct data cleansing and validation exercises regularly to identify and correct inconsistencies and errors in the data.

3. Utilize data imputation techniques to fill in missing values based on existing customer data or relevant statistical methods.

4. Establish data quality monitoring mechanisms to identify and resolve issues promptly.

5. Provide training and guidelines to the data entry team to ensure consistent data entry practices.

By addressing these data quality issues and implementing the proposed strategies, Sprocket Central Pty Ltd will have cleaner and more reliable data, enabling better decision-making and improved customer understanding.

Please let us know if you have any questions or require further assistance in implementing these measures. We are committed to helping you optimize your data quality and achieve your business objectives.

Thank you for your attention to this matter.

Best regards,

Vinayak Shivanagutti

Data Analyst

KPMG