Subject: Data Quality Assessment and Strategies for Improvement

Dear Client,

I hope this message finds you well. We have conducted a comprehensive data quality assessment of the datasets provided, and I wanted to share our findings and recommend strategies to address the identified issues. Our evaluation is based on accuracy, completeness, consistency, currency, relevancy, validity, and uniqueness.

The below table highlights the summary statistics from the three datasets received. Please let us know if the figures are not aligned with your understanding.

Table Name	No. of records	Distinct Customer IDs
Customer Demographic	4,000	3,493
Customer Address	3,999	3,999
Transaction Data	20,000	3,494

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.

Transactions Data Sheet:

- **Null Values:** We found null values in several columns, including **online_order**, **brand**, **product_line**, **product_size**, **standard_cost**, and **product_first_sold_date**.
- **Invalid Values:** The **product_first_sold_date** column contains invalid values of numbers.

Recommendation: To improve data quality in the Transaction Data sheet, we suggest addressing these null values and correcting the data type issue in the **product_first_sold_date** column. You may consider imputing missing data and converting **product_first_sold_date** to a valid date format.

CustomerDemographic Data Sheet:

- **Null Values:** Null values were found in the **last_name**, **DOB**, **job_title**, **job_industry_category**, **default**, and **tenure** columns.
- **Incorrect Values:** The **default** column contains incorrect values as strings.
- **Invalid Year of Birth:** The **DOB** column includes a year of birth as early as 1843.
- **Gender Values:** The **gender** column contains inconsistent values (F, M, Female, Female, Male).

Recommendation: To enhance data quality in the CustomerDemographic Data sheet, we suggest addressing the null values, correcting the **default** column values, and validating the **DOB** column for realistic birth years. Normalize the **gender** values for consistency.

CustomerAddress Data Sheet:

- **Inconsistent Values:** Inconsistent values were identified for locations, such as "Victoria" and "VIC," as well as "NSW" and "New South Wales."

Recommendation: To improve data quality in the CustomerAddress Data sheet, we recommend deduplicating location values to ensure uniqueness.

In summary, addressing these data quality issues will contribute to more accurate, complete, consistent, current, relevant, valid, and unique datasets. We recommend implementing data cleaning, validation, and standardization processes to mitigate these issues and ensure high-quality data for your analysis.

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,
Fola Oluwatosin.