Subject: Assessment of Data Quality and Recommendations for Improvement

Dear Sprocket Central,

I hope this message finds you well. Following our recent discussions, our team has conducted an initial assessment of the data provided by Sprocket Central Pty Ltd. We have identified several data quality issues that might impact our analysis and would like to share our findings and recommendations with you.

Customer Demographic Data:

Missing Data: There are missing values in several columns including last_name, DOB, job_title, job_industry_category, and tenure. Depending on the analysis to be performed, these missing values could cause issues.

Invalid Entries: The default column contains nonsensical and inconsistent entries. We might need your assistance to understand what this column represents.

Customer Addresses Data:

Inconsistent Categories: The state names are not consistently formatted, with some states abbreviated and others not.

Transactions Data:

Missing Data: We observed missing values in the online_order, brand, product_line, product_class, product_size, and standard_cost columns.
Join Keys:

We noticed that the customer_id column, which is the key to join these datasets, appears to be consistent and without missing values. However, we will need to further investigate to ensure there are matching records across all datasets.

Recommendations:

Imputation: For missing numerical data, we can impute missing values with a central tendency measure (mean, median) or use a predictive modeling method. For categorical data, the mode or a predictive modeling method can be used.

Removal: If missing data is non-random or represents a large proportion of the data, it might be best to remove these data entries.

Standardization: We can standardize state names in the Customer Addresses dataset to ensure consistency.

Validation Rules: For the default column in the Customer Demographic dataset, we can work together to understand its significance and establish validation rules to clean it.

Advanced Techniques: For complex missing data issues, we can use techniques such as data mining or machine learning to predict and fill missing values.

Please note that our approach will be guided by the nature of the data, the extent of the data quality issues, and the specific requirements of our analysis. We look forward to discussing these findings and our next steps with you.

Best Regards,
Utuapo Kakuva
Data Analyst