To: Sprocket Central Pty Ltd.

From: KPMG Analytics, Information & Modeling team.

Subject: Data Quality of datasets provided.

Dear Sir/Ma.

Thank you for providing us with the datasets from Sprocket Central Pty Ltd. This mail is to inform you that our team at KPMG Analytics has carried out a data quality assessment on the datasets you provided.

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

Name of dataset	No. of Columns	No. of Rows	Date Received
Transactions	13	20000	Feb 28th 2023
Customer Demographic	13	4000	Feb 28th 2023
Customer Address	6	3999	Feb 28th 2023
New Customer List	23	1000	Feb 28th 2023

Below are the notable data quality issues we encountered, methods used to mitigate them, as well as recommendations to avoid the reoccurrence of these issues to improve the accuracy of the underlying data used to drive business decisions.

• Transactions dataset

Major data quality issues found in the transactions dataset:

1. **Missing values**: The dataset had 360 (1.8%) missing values in the 'online_order' column. The columns 'brand', 'product_line', 'product_class', 'product_size', 'standard_cost', 'product_first_sold_date' all had 197 (0.985%) missing values spanning across the same rows.
2. **Wrong data type of columns**: Some columns had values presented in the wrong data types. e.g 'product_first_sold_date'.
3. **Column accuracy**: In the product_id column, some product ids belong to only one brand while others belong to multiple brands. This may lead us to assume that no brand of product has a unique product id. We would like you to confirm if that is the case.

Mitigation: A total of 2.8% missing values were removed from the dataset and appropriate data transformations were made to ensure consistent data types for a given field.

Recommendation: Ensure that fact tables in the given database have constraints on data types.

• Customer Demographic dataset

Major data quality issues in the Customer Demographic dataset:

1. **Missing Values**: The dataset had 506 (12.65%) missing values in job_title and 656 (16.4%) missing values in job_industry_category columns. It also had 125 (3.125%) customers whose last names were not documented and another 87 (2.175%) had both missing dates of birth and tenure values.

2. **Accuracy of columns**: The columns gender and default had accuracy issues. In the gender column, other categories outside of 'female' and 'male' were documented. These categories were: 'U', 'F', 'Femal', and 'M'. We assume the last three categories to be input errors, (please clarify if this is inaccurate). We would however appreciate more clarity on what gender 'U' is. In the default column, the values are completely incomprehensible.

Mitigation: Replaced the missing values in job_title and job_industry_category with 'N/A' (Not Available) to reduce the number of missing values from 1370 to 415. Replaced misspelt categories to ensure consistency.

Recommendation: Enforce a drop-down list for the user entering the data rather than a free text field to increase column accuracy.

- ## Customer Address dataset

We have no notable quality issues with this dataset.

- ## New Customer List dataset

Major data quality issues in the New Customer List dataset:

1. **Missing Values**: The dataset had 106 (10.6%) missing values in job_title and 165 (16.5%) missing values in job_industry_category columns. It also had 29 customers (2.9%) whose last names were not documented and another 17 (1.7%) had missing dates of birth.
2. **Column Accuracy**: The gender column had accuracy issues. It had another category 'U' other than 'female' and 'male'. We would appreciate clarity on what the category implies.
3. **Missing column titles**: Five columns, positioned between property_valuation and rank had no column titles.

Mitigation: Replaced the missing values in job_title and job_industry_category with 'N/A' (Not Available) to reduce the number of missing values from 285 to 46.

Recommendation: Enforce a drop-down list for the user entering the data rather than a free text field to increase column accuracy. Review all data carefully before separating them into customer demographic and customer address datasets.

- ## Customer ID Primary Keys across datasets.

There are 190 customers who have performed transactions but either have no customer demographic records or customer address records attached to their customer id (186 customers had no customer demographic records, 5 customers had no customer address records and only one customer had no customer demographic and customer address record).

Please refer to the excel file ‘customer_id.xlsx’ for the list of outliers between tables as well as transaction details.

The table below highlights the summary statistics from the four datasets after data cleaning (please compare with the initial table).

Name of dataset	No. of Columns	No. of Rows
Transactions	13	19,445
Customer Demographic	12	3792
Customer Address	6	3999
New Customer List	23	954

Please let us know if any part of this mail was confusing, or if there is any additional information you would like us to know about the datasets.

Thank you for your time.

Oladapo Adepeju Kairat

Data Analyst Intern, KPMG