Subject: Identification of data quality issues and mitigation strategies – Client Sprocket Central Pty Ltd

Dear Sprocket Central Pty Ltd.

First of all, I would like to thank you for contacting and trusting our Analysis, Information and Modeling team at KPMG to carry out the analysis of your customers' dataset and transactions with the aim of identifying your data quality problems. I hope this email can help you understand how to mitigate these issues so you can grow your business.

After evaluating the supplied KPMG\_VI\_New\_raw\_data\_update\_final spreadsheet and following the data quality dimensions (accuracy, integrity, consistency, timeliness, relevance and uniqueness), using Microsoft Excel, below is a summary of the identified problems and corresponding strategies:

Regarding the CUSTOMER DEMOGRAPHY worksheet:

1. CUSTOMER\_ID column without ID information: We notice that the CUSTOMER\_ID column is missing the ID information for customers 4001, 4002 and 4003 as they are registered in the CUSTOMER\_ADDRESS worksheet. To mitigate this issue, we recommend synchronizing the data between the two spreadsheets to ensure consistent customer identification across all relevant records.

2. LAST\_NAME column with missing values: The LAST\_NAME column contains 125 empty lines, indicating missing values. These missing values ​​can lead to a loss of data integrity. To resolve this, we suggest reviewing and updating missing surnames to ensure complete and accurate client records.

3. GENDER column with inconsistent values: The GENDER column has different values ​​representing the same gender. For example, "F", " FEMALE" and "FEMAL" represent the female gender. Likewise, "M" and "MALE" represent the male gender. To improve data consistency, we recommend standardizing gender values ​​to a single representation (for example, "Female" and "Male"). In addition, there are 88 customers registered with just the letter "U", which can lead to misinterpretations. It is advisable to review and update these records accordingly to help improve accuracy.

4. PAST\_3\_YEARS\_BIKE\_RELATED\_PURCHASES column in the wrong worksheet: The PAST\_3\_YEARS\_BIKE\_RELATED\_PURCHASES column represents the number of purchases made by each customer in the last 3 years. Since this information relates to purchasing data and not demographic information, we suggest moving this column to a separate worksheet dedicated to transactional data.

5. DOB column with missing values ​​and incorrect entry: The DOB (Date of Birth) column contains 87 empty rows, which indicate missing values. To maintain data integrity, we recommend reviewing and updating these missing values. Also, an incorrect value was entered for the customer with ID = 34, giving a birth year of 1843, which would make the customer 175 years old. Correcting this entry is essential to ensure accurate data analysis. We also suggested that they change the column name to DATE\_OF\_BIRTH, as the DOB name is not intuitive and can generate doubts when analyzed.

6. JOB\_TITLE column with missing values ​​and unrelated positions: The JOB\_TITLE column has 506 empty rows, indicating missing values. In addition, there are numerous positions that do not match the company's activities, such as TEACHER, TECHNICAL WRITER, TEACHER, ASSISTANT TEACHER, DENTAL HYGIENIST, among others. To resolve this, we recommend reviewing and updating the missing values ​​and ensuring that the job titles listed align with the company's activities.

7. JOB\_INDUSTRY\_CATEGORY column with missing values ​​and unrelated categories: The JOB\_INDUSTRY\_CATEGORY column has 656 rows with the value "N/A", indicating missing values. Also, some of the listed categories do not align with the company's activities. We suggest reviewing and updating the missing values ​​and aligning the categories with the actual activities of the company's sector.

8. DECEASED\_INDICATOR Column: The DECEASED\_INDICATOR column contains two clients who have passed away. We recommend assessing the need to keep these records in the database and considering appropriate treatment, such as archiving or removing the records.

9. DEFAULT column with incorrect values: The DEFAULT column presented numerous incorrect, imprecise and meaningless values, therefore, we recommend evaluating the need for this column to remain in the column or carrying out a total review of these values.

10. TENURE column with missing values: The TENURE column has 87 empty rows, indicating missing values. Of these 87, 50 customers reported that they own a car, therefore, for these customers, the value 1 must be recorded at least.

Regarding the CUSTOMER ADDRESS worksheet:

1. CUSTOMER\_ID column missing values: The CUSTOMER\_ID column in the CUSTOMER ADDRESS worksheet is missing information for IDS 3, 10, 22, and 23, which are recorded in the CUSTOMER\_DEMOGRAPHIC worksheet. To maintain consistency, we advise synchronizing the data between the two worksheets.

2. STATE column with inconsistent values: The STATE column has different values ​​that represent the same state. For example, "VIC" and "Victoria" refer to the state of Victoria, while "NSW" and "New South Wales" represent the state of New South Wales. Standardizing state names to a single representation will promote consistency and accuracy.

Regarding the TRANSACTION worksheet:

1. PRODUCT\_ID column with missing values: The PRODUCT\_ID column contains 1378 orders with ID = 0, which do not refer to the same product. Consequently, these 1378 orders were registered without the sold product ID. To fix this issue, we recommend reviewing and updating these orders with the correct Product IDs to ensure data integrity.

2. CUSTOMER\_ID column with wrong data: The CUSTOMER\_ID column includes 3 orders with ID = 5034. However, the customer with ID = 5034 does not exist in the CUSTOMER\_DEMOGRAPHIC and CUSTOMER\_ADDRESS worksheets, indicating erroneous data entry. It is advisable to identify the correct ID of the actual customer and update these records accordingly.

3. TRANSACTION\_DATE Column: No issues were identified regarding the data quality of this column. However, notice that the worksheet contains data for the year 2017, not the last 3 months of 2018 as expected. Reviewing and updating the data to reflect the correct time period will ensure accuracy in analysis and reporting.

4. ONLINE\_ORDER column with missing values: The ONLINE\_ORDER column has 360 empty lines, indicating missing values. These missing values ​​can result in a loss of data integrity. To resolve this issue, we recommend revising and updating the missing values ​​to accurately reflect whether orders were placed online or through other channels.

5. BRAND column with missing values: The BRAND column contains 197 empty lines, indicating missing values. These missing values ​​can lead to a loss of data integrity. We recommend reviewing and updating missing brand information to maintain accurate records.

6. PRODUCT\_LINE, PRODUCT\_CLASS, PRODUCT\_SIZE, STANDARD\_COST and PRODUCT\_FIRST\_SOLD\_DATE columns with missing values: These columns also have 197 empty rows, indicating missing values. Resolving this issue involves reviewing and updating the missing values ​​for each respective column to ensure complete and accurate information.

7. Data Formatting: The PRODUCT\_FIRST\_SOLD\_DATE column is not formatted as a date, while the LIST\_PRICE and STANDARD\_COST columns are not formatted as currencies (float). We recommend applying appropriate formatting to these columns to improve readability and data consistency.

In none of the 3 worksheets were duplicated lines identified.

By implementing these strategies, you can improve the quality of your transactional data, enabling better decision making and analysis. If you have any questions or need further assistance, please don't hesitate to get in touch. We are here to support you.

Yours sincerely,

Fernanda G Clemente

Data analyst, KPMG