**E-MAIL DRAFT**

**Subject:** Data Quality Assessment and Recommendations for Sprocket Central Pty Ltd

Dear Client,

I hope this email finds you well. I wanted to provide you with an initial assessment of the data you've shared with us and offer recommendations to improve its quality to support our analysis for Sprocket Central Pty Ltd.

Upon reviewing the datasets, we have identified several data quality issues that need attention. These issues may impact the accuracy and reliability of our analysis. Below, I've outlined the main concerns of each data table along with strategies to mitigate them:

**1) NewCustomerList Data:**

* Hidden Columns: There are five hidden columns with unknown headers containing numerical values. We recommend providing metadata or descriptions for these columns to facilitate meaningful analysis. We recommend removing this column for data cleanliness.
* Gender Classification: The "Gender" column includes values such as 'U,' which is unclear in its meaning. Additionally, 17 rows with 'U' gender lack a date of birth (DOB). We suggest clarifying the gender categories and addressing the missing DOB issue or dropping these for cleanliness.
* Zero Bike Purchases: Nine customers have a '0' value for "past\_3\_years\_bike\_related\_purchases," which may be irrelevant to our analysis. These records can be reviewed and potentially corrected or removed.
* Missing DOB and Last Name: 17 rows are missing DOB, and 29 people lack a last name. Data validation checks should be implemented to ensure completeness.
* Job Title and Job Industry Category: The "Job\_title" column has multiple titles with unclear classifications (I, II, III, IV), and about 106 rows are missing. Consider simplifying job titles and providing more descriptive classifications. Additionally, "job\_industry\_category" has 165 empty rows, with 22 rows not having job titles. Providing metadata for these categories would enhance data understanding.
* Deceased Indicator: All records in the "deceased\_indicator" column have 'N' values, which is expected but can be maintained for data consistency.
* Country and Property Valuation: "Country" has consistent values of 'Australia,' while the "property\_valuation" column contains float values. Consider converting property valuations to integers for uniformity.
* Value Column Precision: The "value" column should be made consistent in terms of decimal units, either to 4 or 2 decimal places, to ensure data consistency.

**2) CustomerDemographic:**

* Missing Records: Records {4001, 4002, 4003} are missing from this dataset, which are present in the CustomerAddress table. Ensure data completeness by addressing this issue.
* Missing Last Names: There are 126 records with missing last names. Implement data validation checks to capture this information.
* Gender Classification: The "Gender" column includes multiple values referring to the same category (e.g., Male and M, Female, Femal and F), leading to inconsistency. Simplify gender categories to 'Male' and 'Female' and address the 'U' category, which has 88 rows without data.
* Zero Bike Purchases: 38 customers have '0' purchases in the "past\_3\_years\_bike\_related\_purchases" column, which may be irrelevant to our analysis. Review and potentially correct or remove these records.
* DOB Inaccuracy: One customer's DOB is recorded as 1843, which appears incorrect and should be reviewed and potentially corrected. Additionally, there are 87 missing DOB values, all associated with the 'U' gender category.
* Job Title and Job Industry Category: Similar to the NewCustomerList dataset, the "Job\_title" column contains multiple titles with unclear classifications (I, II, III, IV), and about 504 rows are missing. Simplify job titles and provide more descriptive classifications. The "job\_industry\_category" has 655 empty rows, with 104 rows not having job titles. Providing metadata for these categories would enhance data understanding.
* Deceased Indicator: Two customers are marked in the "deceased\_indicator" column. Ensure the accuracy of this data, and consider removing these records if necessary.
* Tenure: The "Tenure" column has 87 missing values, all associated with the 'U' gender category. As this represents a significant data gap, consider dropping columns with the 'U' gender category.
* Default Column: The "default" column contains unrelevant data with unidentified symbols and characters. We recommend removing this column for data cleanliness.

**3) CustomerAddress:**

* Missing Records: Records {3, 10, 22, 23} are missing from this dataset, which are present in the CustomerDemographic table. To maintain data consistency, address this issue.
* Country: The "country" column has consistent values of 'Australia.'

**4) Transactions:**

* Missing "online\_order" Data: There are 360 missing rows in the "online\_order" column. Implement validation checks to capture this information accurately.
* Missing Data for Multiple Columns: Columns 'brand,' 'product\_line,' 'product\_class,' 'product\_size,' 'standard\_cost,' and 'product\_first\_sold\_date' have 197 rows of missing data. Review and seek to retrieve or eliminate these records for robust analysis.
* Standard Cost Precision: Ensure that "standard\_cost" values are made consistent in terms of decimal units, specifically to 2 decimal places, to ensure data consistency.

Additionally, we recommend providing metadata or descriptions for critical columns, especially "wealth\_segment" and "job\_industry\_category," to facilitate data analysis for new customers. The "product\_first\_sold\_date" column from the Transactions dataset is also essential for a complete understanding of the data.

Please let us know if you have any questions or would like to discuss these findings further. We are committed to working collaboratively to improve the quality of your data and support your business objectives.

Best Regards,

Karan Agrawal

Data Analyst

KPMG's Lighthouse & Innovation Team