Dear Client,

Thank you for sharing the datasets from Sprocket Central Pty Ltd with us. The following table summarizes the major quality concerns identified in the three datasets.

Data Quality Dimension	Accuracy	Completeness	Consistency	Currency	Relevancy	Validity
Customer Demographic	DOB - inaccurate Age column is missing	customer_id - incomplete DOB - null values job title - null values tenure – null values	gender - inconsistency past_3_years_bike_related_ purchases - format change tenure - format change	deceased_indicator - filter out (N)	default - delete column	DOB col has 1843 as year record which is not valid and must be removed.
Customer Address		customer_id - incomplete	state - inconsistency postcode - format change property_valuation - format change			
New Customer List	Age column is missing	Last name – null values Job title – null values DOB – null values				Past_3_years-bike_related_purchase – format change Post code – format change Property_valuation – format change DOB – format change
Transaction	profit column is missing	customer_id - incomplete order_online - null values brand - null values	customer_id - format change product_id - format change transaction_id - format change list_price - format change online_order - format change	Customer_id 5034 is not up to date.	Remove “cancelled” order from order status.	transaction_date - format change product_first_sold_date - format change

Below are more detailed descriptions of the data quality issues observed and the approaches taken to mitigate them. To avoid future data quality problems, recommendations and explanations have been included. Following recommendations will improve accuracy of data used to influence business decisions of Sprocket Central Pty Ltd in the future.

Accuracy Issues: -

1. The ‘DOB’ column in Customer Demographic has one redundant record and blank records (remove 1843 year record and also the nulls).
2. The ‘age’ column was not included in the Customer Demographic dataset.
3. The ‘profit’ column was not included in the Transactions dataset.

Mitigation: Remove the null records and filter out the incorrect records in DOB.

Recommendation: Create an 'age' column in the *Customer Demographic* dataset to provide a more thorough view and make error detection easier. Create a 'profit' column in the *Transactions* dataset to ensure that sales are accurate and to assist in future monetary analysis.

Completeness Issues: -

1. The customer id 5034 in transaction data must be removed as it is not present in customer demographic data and change to text format.
2. The 'last_name', 'DOB' and 'job_title' columns in the Customer Demographics and New Customer List dataset contains null records.
3. The 'order_online' and 'brand' columns in the Transaction dataset contains null records.

Mitigation: Consider data of customer_id from 1 to 3500 as many of them have all the information present in all the three datasets and filter out all the null records from the above-mentioned columns in their respective datasets.

Recommendation: Ensure that all three datasets are up to date and don't have null records or incomplete data, as this will lead to skewed results in our study. The addition of dropdown options will allow for more reliable data, resulting in more detailed analysis.

Consistency Issues: -

1. The 'gender' column in the Customer Demographic dataset contains inconsistent values.
2. The 'state' column in the Customer Address dataset contains inconsistent values

Mitigation: Filter all (M) under Male, filter all (F) and (Female) under Female for 'gender' column in the *Customer Demographic* dataset. Filter all (New South Wales) under NSW and (Victoria) under VIC for 'state' column in the *Customer Address* dataset.

Recommendation: To avoid human error and improve terminology consistency, provide a dropdown menu for the gender and state columns when inputting values in the dataset.

3. The 'past_3_years_bike_related_purchases' and 'tenure' columns in the Customer Demographic dataset have inconsistent format.
4. The 'customer_id', 'postcode' and 'property_valuation' columns in the Customer Address dataset have inconsistent format.
5. The 'customer_id', 'product_id' and 'transaction_id' columns in the Transaction dataset have inconsistent format.

Mitigation: The values of 'past_3_years_bike_related_purchases' and 'tenure' columns in the *Customer Demographic* dataset are converted to numeric format. The values of 'customer_id' and 'postcode' columns in the *Customer Address* dataset are converted to text format and 'property_valuation' column are converted to numeric format. The values of 'customer_id', 'product_id' and 'transaction_id' columns in the *Transaction* dataset are converted to text format.

Recommendation: Set the column formats beforehand, such as text, numeric, and the number of decimal places necessary, based on the values that will be entered in those columns.

6. The format of the 'list_price' column is inconsistent.

Mitigation: The values of 'list_price' column in the *Transaction* dataset is converted to currency format with proper number of decimals to make it consistent with the standard_cost column present in *Transaction* dataset.

Recommendation: Set the column formats beforehand, such as text, numeric, currency and the number of decimal places necessary, based on the values that will be entered in those columns.

Currency Issues: -

1. The ‘deceased_indicator’ column in the Customer Demographic dataset contains records of customers who are no longer active.

Mitigation: Filter out customers whose ‘deceased_indicator’ column in the *Customer Demographic* dataset is set to ‘Y'.

Recommendation: Always make sure that datasets are up to date and only contains active customer records. By removing the deceased customer records, the data will be more relevant, resulting in more accurate estimations in future analyses.

Relevancy Issues: -

1. The ‘default’ column in the Customer Demographic dataset is irrelevant and unnecessary for further analysis.
2. The ‘order_status’ column in the Transaction dataset is irrelevant and unnecessary for further analysis.

Mitigation: Delete the ‘default’ column having metadata in the *Customer Demographic* dataset. Filter out the records with the ‘order_status’ column as cancelled in the *Transaction* dataset.

Recommendation: Examine for any metadata that isn't apparent and either eliminate that data or format it to make more accurate. Ensure that the customer_id in the dataset is correct and up to date. The cancelled order status is irrelevant information that will distort our future analysis results.

Validity Issues: -

1. The ‘transaction_date’ and ‘product_first_sold_date’ columns in the Transaction dataset have wrong format.
2. The ‘DOB’ col in Transaction dataset has 1843 as year record which is not valid and must be filtered out for further analysis.

Mitigation: The values of ‘transaction_date’ and ‘product_first_sold_date’ columns in the *Transaction* dataset are converted to date format (YYYY-MM-DD). Filter out the record having ‘DOB’ of the year 1843.

Recommendation: Set the column formats beforehand, such as date and the number of decimal places necessary, based on the values that will be entered in those columns.

This covers the detailed descriptions of all the data quality issues discovered through first stage of data quality analysis. The suggested mitigation strategies are simple and effective techniques for increasing data quality in preparation for future analysis. Please feel free to contact us if you have any questions about the issues discussed.

Thanks & Regards,

Neha Patil