Dear Client,

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. Below is a summary table which highlights key quality issues that we discovered within the three datasets. Please do let us know if you have any queries surrounding the issues presented.

Summary Table:

	Accuracy	Completeness	Consistency	Currency	Relevancy	Validity
Customer Demographics	DOB: inaccurate Age: missing	Job title: blanks Customer id: incomplete	Gender: inconsistent	Deceased customers: filter out	Default column: delete	
Customer Address		Customer id: incomplete	States: inconsistent			
Transactions	Profit: missing	Customer id: incomplete Online order: blanks Brands: blank			Cancelled status order: filter out	List price: format Product sold date: format

Please find below an in-depth descriptions of data quality issues discovered, and methods of mitigation used. Recommendations and explanations have also been included to avoid further data quality issues in the future. The following recommendations will improve accuracy of data used to influence business decisions of Sprocket Central Pty Ltd in the future.

Accuracy Issues

- DOB was inaccurate for "Customer Demographic" and was missing an age_column, a profit column for "Transactions".

  *Mitigation: Filter out outlier in **DOB**.*
  *Recommendation: Create an **age_column**, allowing for more comprehensible data and easier to check for errors. Also create a **profit_column** in **"Transactions"** to check for accuracy of sales.*

  Creating additional columns for age and profit will allow for easier identification of errors. The **profit_column** will assist in future monetary analysis.

Completeness

- Additional customer_ids were inconsistent among "Customer Demographic", "Customer Address", and "Transactions"

  *Mitigation: Filter all **customer_ids** from 1 to 3500.*

*Recommendation: Ensure tables are up to date (from the same time period). For our model, only **customer_ids from 1 to 3500** will be used as they have complete data.*

The data received may not be in sync across all spreadsheets, with incomplete data the analysis results may be skewed. This is a 'completeness' issue, to prevent future occurrences, it is encouraged to cross check spreadsheets and sync data.

- Blanks in job_title for "Customer Demographic", in online_order and brand column for "Transactions"

  *Mitigation: Filter out 'blanks' for **job_title, online_order, and brand_column.***
  *Recommendation: Simplify job_title to another category such as **industry** or provide dropdown options for **job_title**. Also provide dropdown options for **online_order** and **brand_columns.***

  Blanks are treated as incomplete data and can skew further analysis results. The addition of dropdown options will allow to have more complete data and will result in more accurate analysis.

Consistency

- Inconsistency in gender for "Customer Demographic" and "Customer Address" respectively

  *Mitigation: Filter all 'M' under category of 'Male', Filter all 'Femal' and 'F' under 'Female' for **gender.** Filter all 'New South Wales' to 'NSW' and 'Victoria' to 'VIC' for **states.***
  *Recommendation: Create dropdown options for 'Male', 'Female', and 'U' in **gender.** Create dropdown options for all **state** abbreviations.*

  Dropdown options minimise manual entry and human error. This allows for increase of consistency of terminology. Gender identity can be a sensitive topic, proceed with caution when creating options.

Currency

- People that are 'Y' in deceased_indicator are not current customers for "Customer Demographic"

  *Mitigation: Filter out customers checked **'Y'** in **deceased_indicator.***
  Recommendation: Can be difficult to check for deceased customers, but once this information is received one should update data accordingly

  Deceased customers are not current customers, removing them from data will increase currency of data and will result in more accurate estimates in future analysis.

Relevancy

- Lack of relevancy or comprehensibility in default column for “Customer Demographic” and order_status for “Transactions”

  *Mitigation: Deleted Metadata in default_column. Filter out **“Cancelled” order_status.***
  Recommendation: Check for incomprehensible Metadata and delete or format to make comprehensible.

  ***‘Cancelled’ order_status** is irrelevant information for future analysis, as it can skew data- e.g. total number of customers per annum will be an overestimate.*

Validity

- Format of list_price, product_sale_data for “Transactions”

  *Mitigation: Format **product_sale_data** to short date format, format **list_price** to currency.*
  Recommendation: Set up columns so that formats such as price and decimals are already in place when entering new data.

  These formatting will make data to be interpreted more easily. Formatting into price and allowing for either 2 or 3 decimals placed consistency will increase readability. This will reflect positively on speed and accuracy of analysis for business decisions.

That summarises all data quality issues discovered through the first stage of data quality analysis. The mitigation strategies suggested are simple and effective ways of improving data quality for future analysis. They will not only improve the analysis output that one can perform within the company but will increase the level of analysis that can be performed by KPMG and other hired analysis teams.

Please let us know if you have any questions regarding mitigation or any data quality issues identified.

Kind regards,
Tooba Khullat