Dear Sprocket Central Pty Ltd,

Thank you for making available the dataset from Sprocket Central Pty Ltd. We have reviewed the attached dataset and summarised the following data quality issues.The table below highlights key data quality issues with the Sprocket central Pty Ltd dataset.

Summary Table

Worksheet Name	Accuracy	Completeness	Consistenc y	Currency	Relevancy	Validity
Transaction	a. Profit: missing	a. Customer id: incomplete b. Online order: blanks c. Brand: blanks			a. Cancelled status order: filter out	a. List price: format b. Product sold date: format
Customer Address		Customer id: incomplete	States: inconsistent			
Customer Demographic	a. DOB: Inaccurate b. Age: Missing	a. Job title: blanks b. Customer id: incomplete	a. Gender: inconsisten cy	a. Deceased customers: filtered out	a. Default column: dropped	

The data quality issues discovered have been further described below and methods of mitigation used. We have given recommendations below to avoid subsequent data quality issues.

Accuracy Issues

- **DOB was inaccurate for “Customer Demographic” and missing an age_column; missing a profit column for “Transactions”**

  Mitigation: Filter out the outlier in DOB.
  Recommendation: Create an “age”, allowing for more comprehensible data and ease of checking errors. Create a “profit” in “Transactions” to check accuracy of sales.

  Creating additional columns for age and profit will allow for easier identification of errors. The profit column will assist in future financial analysis.

Completeness Issues

- **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 columns.
  Recommendation: Simplify job_title to another category such as industry_industry or provide dropdown options for job_title. 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 for more complete data and will result in more accurate analysis.

Consistency Issues

- **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,” ‘Femal,’ and ‘U’ in gender. Create dropdown options for all state abbreviations.

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

Currency Issues

- **Customers 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 Issues

- **Lack of relevancy or comprehensibility in default_column for “Customer Demographic” and order_status for “Transactions”**

  Mitigation: Delete 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. For example, the total number of customers per annum will be an overestimate.

Validity Issues

- **Format of list_price, product_sale_date for “Transactions”**

  Mitigation: Format product_sale_date 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.

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

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 can be performed within the company but will increase the level of analysis that can be performed by KPMG’s analytics team.
Kindly let us know if you have any questions as regards the identified data quality issues, the mitigations and recommendations.

Kind Regards

Tochukwu Collins
Junior Data Consultant