

Dear John Smith,

Thank you for providing us the four datasets from the Sprocket Central Pvt Ltd. The datasets have lots of quality issue that we encountered while working with the dataset. The summary table below highlights key quality issue with the dataset.

Please let us know if you have any queries regarding the issue presented.

Summary Table

	Accuracy	Completeness	Consistency	Currency	Relevancy	Validity
Customer Demographic	DOB: inaccurate Age: Missing	Job title: Blanks Last Name: Blanks DOB: Blanks Job Industry: N/A	Gender: Inconsistency	Deceased Customers: Filter out	Default column: Irrelevant	Customer id: format DOB: Format Tenure: Format Past_3_year: Format
Customer Address		Customer id: Incomplete	State: Inconsistency			Post code: Format Property Valuation: Format
Transactions	Profit: Missing	Customer id: Incomplete Product line: Blanks Product Class: Blanks Product sold date: Blanks. Product size: Blanks Online Order: Blanks Brand: Blanks	Standard cost: inconsistent			Customer id: Format Transaction date: Format List Price: Format Product sold date: format
New Customer list	DOB: inaccurate Age: Missing	Job title: Blanks Last Name: Blanks DOB: Blanks			4 unknown columns: Irrelevant	Post code: Format Property Valuation: Format Past_3_year: Format

Below are the more details descriptions of data quality issues discovered and the methods used to tackle those issue.

ACCURACY ISSUES

- DOB was inaccurate for “*Customer Demographic*” and “*New Customer list*” and missing an age_column to filter out the outliers easily.

Mitigation: Filter out outlier in DOB.

- “*Transactions*” is missing a profit_column to check the profitability of sales.

Recommendation: Create an age_column, allowing for more comprehensible data and easier to check for errors. Create a profit_column in “*Transactions*” to check accuracy of sales.

COMPLETENESS

- Job-title, DOB, Last Name is “*Customer Demographic*” and “*New Customer list*” have blanks.
- Customer id, Product line, Product Class, Product sold date, Product size, Online Order, Brand in the “*Transactions*” have blanks.

Mitigation: Filter out 'blanks' for **job_title, online_order, and brand_column**.

- “Customer Demographic” have 4000 customer id whereas “Transactions” has only 3500 which also inconsistent.

Mitigation: Filter all customer_ids from 1 to 3500

- Additional Customer ids in “Customer Address” is inconsistent.

CONSISTENCY

- Inconsistency in gender for “Customer Demographic”.

Mitigation: Filter all 'M' under category of 'Male,' filter all 'Femal' and 'F' under 'Female' for gender.

- Inconsistency in states for “Customer Address”.

Mitigation: Filter all 'New South Wales' to 'NSW' and 'Victoria' to 'VIC' for states.

- Inconsistency in standard cost for “Transactions”.

Mitigation: Manually Adding “\$” to the row missing this.

CURRENCY

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

Mitigation: Filter out customers checked 'Y' in deceased_indicator.

RELEVANCY

- Irrelevant column default in "Customer Demographic "

Mitigation: Deleted the default column.

- Found 4 hidden columns in without any name in “New Customer List”.

Mitigation: All 4 are calculated columns. 3 of them are irrelevant so deleted them. One of the columns is calculating RANK so renamed is as rank and check for its consistency.

VALIDITY

- Format of Transaction date, list price, product_sale_date for "Transactions".

Mitigation: Convert the date columns into Date format and price columns into Currency format.

- Format of Customer id, DOB, Tenure, Past_3_year for "Customer Demographic " and “New Customer List”.

Mitigation: Convert the DOB into Date format and rest in number format.

- Format of Post code, Property Valuation for “Customer Address”.

Mitigation: Convert the of Post code Property Valuation in number format.

That summarises all data quality issues discovered analysis. The mitigation strategies suggested quality for future analysis. They will not only perform within the company but will increase KPMG and other hired analysis teams. Please let us know if you have questions identified.

Kind regards,

Sudhanshu Gosain