

Dear Mr. or Mrs. Client,

Thank you for providing us with the datasets from Sprocket Central Pty Ltd. The below table highlights the summary statistics from the three datasets received. Please let us know if the figures are not aligned with your understanding.

	Accuracy	Completeness	Consistency	Timeliness	Validity
Customer Demographic	Age: cannot be found	Customer id: incomplete Job title: blanks	Gender: inconsistency	Filter out deceased customers	Default: cannot be interpreted
Customer Address		Customer id: incomplete	States: inconsistency		
Transactions	Profit: missing	Customer id: incomplete Online order: blanks Brands: blanks			Cancelled status: filter out

Below are more in depth descriptions of data quality issues uncovered and solutions and recommendations on curbing such data quality issues and improve the accuracy of the underlying data to drive business decisions in the future.

Accuracy Issues

- **Age and Profit columns is missing in Customer demographic and Transactions respectively**

Mitigation: create an age column and profit column for more convenient future analyses.

Age will allow easier verification of errors and can assist in customer analysis. Profit will be helpful in analyzing finance.

Completeness Issues

- **Incomplete customer_id values across data sheets.**

Mitigation: Use only customer ids from Customer Demographic data
When data received are not in sync, the analysis results might be skewed.

- **Blanks are found in Job Title, Online order and Brands column**

Mitigation: Removing a small number of rows would not affect the overall data. For a large number of blanks, or it is a necessary field, the missing values should be implemented based on the distribution.

The purpose is to analyze the full (or nearly full) set of data without losing information or destabilize the current distribution.

Consistency Issues

- **Inconsistent values in Gender (F, Female,..) and States (V, Vic, Victoria,..) columns**

Mitigation: Change all different values with similar meaning into just one value that could represent the same meaning.

Selectable options should be implemented rather than a free text field.

Timeliness Issues

- **Customers with deceased_indicator value of 'Y' are not current customers**
Mitigation: Filter out such customers will result in a more accurate analysis of current customer.

Validity Issues

- **Default column cannot be interpreted in a usable way**
Mitigation: Remove the default column
An unusable column would only cause confusion with the analysis process.
- **Order Status of 'Cancelled' has no relevancy.**
Mitigation: Remove all cancelled orders
Cancelled orders are not relevant and might skew the data and result in inaccurate analyses.

That summarizes all data quality issues discovered through the first stage of data quality analysis. The mitigation techniques are simple and effective ways of improving data quality for future analyses and produce more precise results for discovering insights and business decisions. If any further questions arise, please let us know and they will be addressed moving on.

Kind regards,
Michael Nguyen