

Hello Manager,

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The summary table below highlights key quality issues that we discovered within the three datasets. Please let us know if you have any queries regarding the issues presented.

### Summary Table

	Accuracy	Completeness	Consistency	Currency	Relevancy	Validity
Customer Demographic	<ul style="list-style-type: none"><li>• DOB: Inaccurate</li><li>• Age: missing</li></ul>	<ul style="list-style-type: none"><li>• Job title: Blanks</li><li>• Customer id: incomplete</li></ul>	<ul style="list-style-type: none"><li>• Gender: inconsistency</li></ul>	<ul style="list-style-type: none"><li>• Deceased Customers: filter out</li></ul>	<ul style="list-style-type: none"><li>• Default column: delete</li></ul>	
Customer Address		<ul style="list-style-type: none"><li>• Customer id: incomplete</li></ul>	<ul style="list-style-type: none"><li>• State inconsistency</li></ul>			
Transactions	<ul style="list-style-type: none"><li>• Profit: missing</li></ul>	<ul style="list-style-type: none"><li>• Customer id: incomplete</li><li>• Online order: blanks</li><li>• Brand: blanks</li></ul>			<ul style="list-style-type: none"><li>• Cancelled status order: Filter out</li></ul>	<ul style="list-style-type: none"><li>• List price: format</li><li>• Product sold date format</li></ul>

Below are more 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 businesses decisions of Sprocket Central Pty Ltd in the future.

### Accuracy Issues

- **DOB Was inaccurate for “Customer Demographics” and missing an age \_column; missing 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. Create a **profit\_column** in “**Transaction**” 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 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 one 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 Demographics”, in online\_order and brand\_column for “Transaction”**

*Mitigation: Filter out ‘blanks’ for **job\_title**, **online\_order**, and **brand\_column**.*

*Recommendation: Simply **job\_title** to another category such as **industry\_industry** or provide dropdown options for **job\_title**. Provide dropdown options for **online\_order** and **brand\_column**.*

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 ‘female’ 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.*

Drop down options, minimize manual entry and human error. Allow 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 Demographics.”**

*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 the 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 – for example total number of customers per annum will be an overestimate.

## Validity

- **Format of list\_price, product\_sale\_date for "Transaction"**

*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 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.

That summarizes all data quality issues discovered through the first stage of the data quality analysis. The mitigation strategies suggested are simple and effective ways of improving data quality for future analysis. This will not only improve the analysis output that can be performed within the company what 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 mitigations or any data quality issue identified.

Kind regards,  
Taranbeer Singh