

Dear Sophia Williams,

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The significant problems related to data quality that were faced and the strategies utilised to address the detected inconsistencies in the data are listed below. Additionally, suggestions have been given to prevent the recurrence of data quality issues and enhance the precision of the fundamental data employed to support business decisions. Please let us know if the figures are not aligned with your understanding.

## Accuracy

- DOB was inaccurate for the customer demographic dataset.
- Recommended : Created an **Age** column in customer demographic, a **Profit** and **Transaction Month** columns in transactions.
- Creating the age column will help in easier identification of errors in the DOB column.
- And adding a Profit Column will help in monetary analysis in the future.

## Completeness

- **Customer id**, **job title** and **last name** columns in customer demographic had missing values.
- **Last name**, **DOB** and **job title** columns in Newcustomerlist had missing values.
- **Order status** and **Brand** columns in Transactions had missing values.
- Blanks are treated as incomplete data and can skew further analysis results. The addition of dropdown options will allow us to have more complete data and will result in accurate analysis. Recommended : Filter out empty rows in customer id, last name, dob, order status and brand columns. And replace empty rows in job title column across the datasets with "**N/A**"

## Consistency

- Data is inconsistent in the **Gender** column in customer demographic and **State** column in customer address.
- When the gender of individuals in a dataset are incorrectly labelled or missing, it leads to inaccuracies in any analysis or conclusions drawn from the data. Therefore replace 'M' with 'Male', 'Femal' and 'F' with 'Female' in the gender column.
- Recommended : In the customer address dataset, replace 'New South Wales' with 'NSW' and 'Victoria' with 'VIC' in the State column.

## Relevancy

- **Default** column in the customer demographic lacks relevance so it is best to delete the column.
- The **order status** column in transactions contains 'Cancelled' orders too which will mess up the analysis and moreover it is irrelevant. So filtered out Cancelled from order status.

## Validity

- Formatted the **product\_first\_sold\_date** in transactions to short date format so that the data is interpreted correctly for analysis.
- Formatted the **list\_price** in transactions to custom currency format so that it will increase consistency and readability.

In conclusion, this overview encompasses all of the data quality concerns uncovered during the initial stage of the data quality analysis. The recommended mitigation strategies are uncomplicated and efficient methods of enhancing data quality for future analysis. A link to the data after quality analysis :

[KPMG\\_Task\\_1](#)

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

Kind regards,  
Sneha Chavakula