Thank you for providing us with the datasheets from Sprocket Central Pty Ltd. The following table highlights the quality issues found during the data quality check within the datasheets provided.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Accuracy | Completeness | Consistency | Currency | Relevance | Validity |
| Customer Demographic | * DOB: inaccurate * Age: missing | * Job Title: Blanks * Customer id: incomplete | * Gender: inconsistency | * Deceased customers: Filtered out | * Default column: deleted |  |
| Customer Address |  | * Customer id: incomplete | * States: inconsistency |  |  |  |
| Transactions | * Profit: missing | * Customer id: incomplete * Online Order: blanks * Brand: blanks |  |  | * Cancelled statis order: filtered out | * List price: format * Product sold date: format |

Below are some of the methods of mitigation used. In depth descriptions are provided as well to improve the accuracy of data which will be used to influence business decisions of Sprocket Central Pty Ltd.

**Accuracy Issues:**

DOB was inaccurate for “Customer Demographic” and missing an age\_Column; missing a profit column for “transactions”

*Mitigation*: Filter out outliner DOB

*Recommendation*: Create age\_column, allowing for more comprehensible data and easier to check for errors. Create a profit\_column in “transactions” to check inaccuracy of sales.

**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.

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” and 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 in “Transactions’

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

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

Blanks are considered as incomplete data and can skew the results of future analysis.

**Consistency Issues:**

Inconsistency in gender for “customer demographic” and states in “customer address” respectively

*Mitigation*: Filter all ’M’ under the category of ‘Male’ filter all ‘F’ under the category of ‘Female’ for gender. Filter all ‘New South Wales’ under the category ‘NSW’ and ‘Victoria’ in ‘VIC’ for states.

*Recommendation*: Create dropdown options for ‘Male’, ‘Female’ and ‘U’ in gender. Create dropdown options for all states abbreviations.

Dropdown options minimize the manual entry and human error. Allows for increase of consistency of terminology.

**Currency Issues:**

People 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 would increase data accuracy and result in more accurate results in future analysis.

**Relevancy Issues:**

Lack of relevancy or comprehensibility I default\_column for “Customer Demographic” and order\_status for “Transactions”

*Mitigation*: Deleted metadata in default\_column. Filter out “Cancelled” in order\_status

*Recommendation*: Check for incomprehensible Metadata, delete or filter out to make comprehensible.

“Cancelled” order\_status is irrelevant data as it can skew data and can be responsible for overestimation for the total number of customers annually.

**Validity Issues:**

Format of list\_price, product\_sale\_date for “transactions”

*Mitigation*: Format product\_sale\_date to short date. Format list\_price to currency.

*Recommendation*: Set up columns such that the format for currency and dates is already in place while entering the data.

Allowable values will make data to be interpreted more easily. Formatting into price and allowing for wither 2 or 3 decimal placed will increase readability.

This summarizes the data quality issues discovered through the first stage of data quality analysis. The mitigations strategies and suitable recommendations have been provided as to increase the effectiveness of the analysis in the future.

If there are any queries regarding mitigations or any data quality issues feel free to contact us.

Regards,

Akshat Bhardwaj