Dear Client,

Thank you for trusting us with the job and providing with the datasets from Sprocket Central Pty Ltd. The summary table below highlights the major quality concerns that we discovered in the data shared with us.

Do let us know if you have any further query.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Quality Dimensions** | **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** |
| **Customer Demographics** | * DOB: inaccurate * Age: missing | * Job Title: blanks * DOB: blank * Customer Id: incomplete | * Gender: inconsistency * past\_3\_years\_bike   \_related\_purchases: format   * tenure: format | * Deceased customers: filter out | * Default column: delete |  |
| **Customer Address** |  | * Customer id: Incomplete | * States: inconsistency |  |  |  |
| **Transactions** | * Profit: missing | * Customer Id: incomplete * Online orders: blank * Brand: blank | * Customer\_id: format * Transaction\_id: format * Product\_id: format |  | * Cancelled status order: filter out * Customer\_id: filter out | * List price: format * Product sold date: format * Transaction date: format |

Below is more detailed description of data quality issues observed and the approach taken to mitigate them. To avoid further confusions, recommendations and explanations have been listed. Following description will help in improving accuracy of data used to influence data driven decisions in the organization.

# Accuracy Issues: -

1. The ‘DOB’ column in the Customer Demographic dataset was incorrect and had null values.
2. Missing ‘age’ column in the Customer Demographic dataset.
3. Missing ‘profit’ column in the Transactions dataset.

**Mitigation**: Remove the null values and filter out outliers in DOB column.

**Recommendation**: Create an ‘age’ column in the *Customer Demographic* dataset allowing for comprehensive data and easy error checking. Create a ‘profit’ column in the *Transactions* dataset to ensure accuracy.

# Completeness Issues: -

1. The ‘customer\_id’ column has incomplete values nacross all the three datasets.
2. The ‘DOB’ and ‘job\_title’ columns in the Customer Demographics dataset contains null values.
3. The ‘order\_online’ and ‘brand’ columns in the Transaction dataset contains null values.

**Mitigation**: Filter all customer\_id from 1 to 3500 as many of them have all the information present in all the three datasets and filter out all the null records from the above-mentioned columns in their respective datasets.

**Recommendation**: The data received may not be in sync across all the spreadsheets. This is a ‘completeness’ issue. To prevent future occurrences, make sure to cross check all the datasets.

# Consistency Issues: -

1. The ‘gender’ column in the Customer Demographic dataset contains inconsistent values.
2. The ‘state’ column in the Customer Address dataset contains inconsistent values

**Mitigation**: Filter all (M) under Male, filter all (F) and (Femal) under Female for ‘gender’ column in the *Customer Demographic* dataset. Filter all (New South Wales) under NSW and (Victoria) under VIC for ‘state’ column in the *Customer Address* dataset.

**Recommendation**: To avoid human error and improve terminology consistency, provide a dropdown menu for the gender and state columns when inputting values in the dataset.

1. The ‘past\_3\_years\_bike\_related\_purchases’ and ‘tenure’ columns in the Customer Demographic dataset have inconsistent format.
2. The ‘customer\_id’, ‘postcode’ and ‘property\_valuation’ columns in the Customer Address dataset have inconsistent format.
3. The ‘customer\_id’, ‘product\_id’ and ‘transaction\_id’ columns in the Transaction dataset have inconsistent format.

**Mitigation**: The values of ‘past\_3\_years\_bike\_related\_purchases’ and ‘tenure’ columns in the *Customer Demographic* dataset are converted to numeric format. The values of ‘customer\_id’ and ‘postcode’ columns in the *Customer Address* dataset are converted to text format and ‘property\_valuation’ column are converted to numeric format. The values of ‘customer\_id’, ‘product\_id’ and ‘transaction\_id’ columns in the *Transaction* dataset are converted to text format.

**Recommendation**: Set the column formats beforehand, such as text, numeric, and the number of decimal places necessary, based on the values that will be entered in those columns.

# Currency Issues: -

1. The ‘deceased\_indicator’ column in the Customer Demographic dataset contains records of customers who are no longer active.

**Mitigation**: Filter out customers in ‘deceased\_indicator’ column in the *Customer Demographic* dataset where value is set to ‘Y'.

**Recommendation**: Always make sure that datasets are up to date and only contains active customer records. By removing the deceased customer records, the data will be more relevant, resulting in more accurate estimations in future analyses.

# Relevancy Issues: -

1. The ‘default’ column in the Customer Demographic dataset is irrelevant and unnecessary for further analysis.
2. The ‘customer\_id’ and ‘order\_status’ column in the Transaction dataset is irrelevant and unnecessary for further analysis.

**Mitigation**: Delete the ‘default’ column having metadata in the *Customer Demographic* dataset. Filter out the records with

customer\_id ‘5034’ and the ‘order\_status’ column as cancelled in the *Transaction* dataset

**Recommendation**: Examine for any metadata that isn't apparent and either eliminate that data or format it to make more accurate. Ensure that the customer\_id in the dataset is correct and up to date. The cancelled order status is irrelevant information that will distort our future analysis results.

# Validity Issues: -

1. The ‘transaction\_date’, ‘list\_price’ and ‘product\_first\_sold\_date’ columns in the Transaction dataset have wrong format

**Mitigation**: The values of ‘transaction\_date’ and ‘product\_first\_sold\_date’ columns in the *Transaction* dataset are converted to date format. The values of ‘list\_price’ column in the *Transaction* dataset is converted to currency format with proper number of decimals.

**Recommendation**: Set the column formats beforehand, such as date, currency and the number of decimal places necessary, based on the values that will be entered in those columns.

This covers the detailed descriptions of all the data quality issues discovered through first stage of data quality analysis. The suggested mitigation strategies are simple and effective techniques for increasing data quality in preparation for future analysis. Please feel free to contact us if you have any questions about the issues discussed.

Thanks & Regards,

Diana Satapathy