Hello Kathleen,

Thank you for providing us with Sprocket Central Pty Ltd.’s datasets. I’ve highlight key data quality issues in the following summary table:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Datasets** | **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** |
| Customer Demographic | DOB:  Error  Age:  missing | Job title:  Blank  Customer ID:  incomplete | Gender:  inconsistency | Deceased Customers:  Filter out | Default  Columns: delete |  |
| Customer Address |  | Customer ID:  incomplete | States:  inconsistency |  |  |  |
| Transactions | Profit:  missing | Customer ID: incomplete  Online order:  Blank  Brand: blank |  |  | Cancelled Status order: filter out | List price: format  Product sold date: format |

I’ve recommended the below improvement of data accuracy that will impact accuracy of data in the future. These mitigation are very effective ways of improving data quality of three datasets.

**Accuracy Issues**

Issue: Inaccurate DOB in customer demographic dataset, missing age column and profit column in transaction dataset

Mitigation: Filter out outlier in DOB.

Recommendation: Create an age column that will allow for easier identification of errors. A profit column will assist future analysis.

**Completeness**

Issue: Incomplete additional customer among customer demographic, customer address, and transactions datasets.

Mitigation: Filter all customer\_ids

Recommendation: use only customer\_ids that have complete data. Cross check datasets and sync data to prevent future completeness issues.

**Consistency**

Issues: Inconsistent gender in Customer demographic and customer address datasets

Mitigation: Filter all “M” under category of “Male” and “F” under “female” category; filter all “new South Wales” to “NSW” and “Victoria” to “VIC” for states. Create dropdown options for all state abbreviation to minimize human error and manual entry.

**Currency**

Issues: deceased individuals are in the customer demographic dataset

Mitigation: Filter out “Y” in deceased\_indicator.

Recommendation: update deceased customers’ data; remove them from dataset to get more accurate estimates.

**Relevancy**

Issues: Lack of relevancy in default\_column in customer demographic and order status in transactions datasets

Mitigation: Deleted Metadata in default\_column. Filter out “cancelled” order status.

Recommendation: Check for incomprehensible metadata.

**Validity**

Issues: list\_price and product\_sale\_date format in transactions dataset

Mitigation: Fromat product\_sale\_date to short date format; format list\_price to currency.

Recommendation: Fix columns to increase reliability and validity.

Please let me know if you have any concerns regarding data quality.

Best,

Farhana Akter