Dear Sprocket Central Team,

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

**Summary Table**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** |
| **Customer Demographic** | DOB: inaccurate  Age: Missing | Job title: blanks  Customer id: incomplete | Gender: inconsistency | Deceased customer: filter out | Default column: delete |  |
| **Customer Address** |  | Customer id: incomplete | States: inconsistency |  |  |  |
| **Transactions** | Profit: missing | Customer id: incomplete  Online orders: blanks  Brands: blanks |  |  | Cancelled order status: filter out | List price: format  Product sold date: format |

Below are more in depth description of data quality issues discovered & methods of mitigation used. Recommendation & explanations have also been included to avoid further data quality issues in the future.

Following recommendations will improve accuracy of data used to influence business decisions of Sprocket Central Pty Ltd in the future

**Accuracy Issue**

* **DOB was inaccurate for “Customer Demographic” 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 & easier to check for errors. Create a***profit\_column***in***“Transactions”***to check accuracy of sales.

**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 1 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 Demographic” in online\_order and brand\_column for “Transactions”**

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

*Recommendation*: Simplified job title to another category such as industry underscore industry or provide drop down options for job title provide drop down options for online order and brand collar

Blanks are treated as incomplete data and can skew further analysis results. The addition of dropdown options will allow having 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 'Femal' 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.

Dropdown options minimize manual entry and human error. Allows 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 Demographic"**

*Mitigation*: Filter out customers checked '***Y***' in ***deceased\_indicator*** .

*Recommendation*: Can be difficult to check for disease customers but was this information is received one should update data accordingly.

Deceased customers are not current customers, removing them from data will increase accuracy 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 relevant 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 transactions**

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

*Recommendation*: Setup columns so that formats such as price and decimals are already in place when entering new data.

Allowable values will make data to be 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.

This 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. They will not only improve the analysis output that one can perform within the company but will increase the level of analysis that can be performed by KPMG and other hired analysis teams.

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

**Kind regards,**

**Pranay Kumar**