Dear John Doe,

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The summary table below highlights key quality issues that we discovered within the three data sets. 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: Blank * Customer id: incomplete | * Gender: Inconsistency | * Deceased customers: Filter out | * Default column:   Delete |  |
| **Customer Address** |  | * Customer id: Incomplete | * Status: Inconsistency |  |  |  |
| **Transactions** | * Profit is missing * Transaction month is missing | * Customer id: Incomplete * Brand: Blank * Online Order: Blank |  |  | * **Cancelled Status order: filter out** | * List Price: format * Product Sold Date: format |

Below are more in depth description of data quality issues discovered and methods of mitigation used. Recommendations and explanation have also been included to avoid further data quality issues in the future. Following recommendations will improve the accuracy of the data which is used to influence to business decisions of Sprocket Central Pty Ltd in the future.

**Accuracy Issue**

* **DOB was inaccurate for “Customer Demographic” and missing an Age columns;**

**Missing a Transaction month and Profit columns for “Transactions”**

*Mitigation: Filter out outlier in* ***DOB****.*

*Recommendation: Create an* ***Age column****, allowing for more comprehensible data and easier to check for errors. Create a* ***Profit column*** *in* ***Transactions*** *to check accuracy for sales. Also created* ***Transaction month column*** *in* ***“Transactions”*** *for better visibility in which month most of the transactions are happening.*

Creating additional columns for Age, Profit and Transaction month will allow for easier identification of errors. The **profit column** will assist in future monetary analysis.

**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 upto 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 result may be skewed. This is ‘completeness’ issue to prevent future occurrences, it is encouraged to cross check the 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: Simplify* ***job\_title*** *to another category or provide dropdown option for* ***job\_title****. Provide dropdown options for* ***online\_order*** *and* ***brand column.***

Blanks are treated as incomplete data and can skew further analysis results. The addition of dropdown option will allow to 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 category of ‘Female’ for* ***gender****. Filter all ‘NSW’ to ‘New South Wales’ and ‘VIC’ to ‘Victoria’ for* ***states****.*

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

Dropdown option minimizes manual entry and human error. Allows for increase of consistency of terminology. Gender identity can be a sensitive topic, proceed with cautions when creating options.

**Currency**

* **People that are ‘Y’ in deceased\_indicator are not current customers in “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 from data will increase currency of data and will result in more accurate estimates in future analysis.

**Relevancy**

* **Lack of Relevancy or Comprehensibility in default\_columns for “Customer Demographics” and order\_status for “Transactions”**

Mitigation: Deleted Metadata in default\_columns. Filter out **“Cancelled” order\_status**.

Recommendation: Check for incomprehensible Metadata and delete or format to make comprehensible.

**‘Cancelled’ order\_status** is irrelevant 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 of **product\_sale\_date** to short date\_format, format **list\_price** to currency.

Recommendation: Set up 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.

That summarizes all data quality issues discovered through the first stage of data quality analysis. The mitigation strategies suggested are simple and effective ways of improving data quality for future analysis. This will not only improve the analysis output that one can perform within the company but will also increase the level of analysis that can be performed by KPMG and other hired analysis teams.

Please let me know, if you have questions regarding mitigation or any data quality issues identified.

Kind Regards,

Rajat Singh