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.

**Error 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 | * States: inconsistency |  |  |  |
| **Transaction** | * Profit: missing | * Customer id: incomplete * Online Order:   Blanks   * Brand: blanks |  |  | * Cancelled status order: filter out | * List price: format * Product sold date: format |

Below are more in depth descriptions of data quality issues discovered and methods of mitigation used. Recommendations and 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 Issues**

* **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 data that are comprehensible and easier to check for errors. Create a **profit\_column** in **“Transactions”** to check accuracy of sales.

Creating additional columns for age and profit will allow for easier identification of errors. The **Profit\_column** will assist in future monetary analysis.

**Completeness**

* **Addition 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’ issues, to prevent future occurrences it is encouraged to cross check spreadsheets and sync data.

* **Blanks in job\_title for “Coustomer Demographic,” in online\_order and brand\_column for “Transactions”**

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

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

Blanks are treated as incomplete data and can skew further analysis results. The addition of dropdown options will allow to have more complete data and will result in more accurate analysis.

**Consistency**

* **Inconsistency is 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 **‘VLC’** for states.

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

Dropdown options minimize manual entry and human error. Allows for increase of consistency of terminology. Gender identity can be 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 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\_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 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 **product\_sale\_date** to short date format, format **list\_pice** 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 interpret more easily. Formatting into price and allowing for either 2 or 3 decimal placed consistently will increase readability. This will reflect positively on speed and accuracy of analysis for business decisions.

That summarise 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 me know if you have any question regarding mitigation or any data quality issues identified.

Kind regards,

A.Yash Kumar Rao