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.

**Data Quality Framework Table**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Datasets** | **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** |
| **Customer**  **Demographic** | Inaccurate D.O.B  Age  Missing | Job title: blanks | Gender: inconsistency | Filter out deceased  customers | Delete: default  column |  |
| **Customer**  **Address** |  |  | States: inconsistency |  |  |  |
| **Transactions** | Profit: missing | Online order: blanks  Brand: blanks |  |  | Cancelled order status:  filter out | List price: format  Product first 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"
* Profit column and age column are missing in the “Transactions” dataset.
* **Mitigate:** Filter out DOB.
* **Recommendations:** Create an age\_column, allowing for more comprehensible data 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**

* Additional customer\_ids were inconsistent among "Customer Demographic," "Customer Address," and "Transactions"
* Blanks in job title for the “Customer Demographic” dataset.
* Blanks in online order and brand column in the “Transactions” dataset.
* **Mitigation:**
  + - * Filter out job title; brand; and online order columns.
      * Filter all customer\_ids from 1 to 3500
* **Recommendations:** 
  + - * Blank values in columns affect data analysis in a negative way, so avoid null values.
      * 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 our 'blanks' for job\_title, online order, and 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 order 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 in gender column for “Customer Demographic” dataset.
* Inconsistency in states column for “Customer Address” dataset
* **Mitigate:** Filter out ‘M’ as ‘Male’ and ‘F’ as ‘Female’ in gender column. Filter out ‘New South Wales’ as ‘NSW’ and ‘Victoria’ as ‘VIC’ in the states column.

**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 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 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 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,

**Joshua Akintayo Oladeji**