Dear Manager,

Thank you for providing us with the three Sprocket Central Pty Ltd datasets. The summary table below highlights the key quality issues identified in the three data sets.

Summary Table:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** |
| **Transactions** | **Profit:** | **Customer id:** |  |  | **Cancelled Status order:** Filter out | **List Price:** Format |
| Missing | Incomplete | **Product Sold Date:** Format |
|  | **Online Order:** |  |
|  | Blanks |  |
|  | **Brand:** Blanks |  |
| **Customer Demographic** | **DOB:** Inaccurate | **Job title:** Blanks | **Gender:** Inconsistency | **Deceased Customers:** Filter out | **Default Column:** Delete |  |
| **Age:** | **Customer id:** |
| Missing | Incomplete |
| **Customer Address** |  | **Customer id:** | **States:** Inconsistency |  |  |  |
| Incomplete |

More detailed descriptions of data quality issues discovered and mitigation methods used are presented below. Recommendations and explanations have also been included in order to minimize potential data quality issues. The following recommendations will improve the accuracy of data used to influence Sprocket Central Pty Ltd's company decisions in the future.

* **Accuracy Issues:**

DOB was incorrect for "Customer Demographic" and was missing an age\_column; "Transactions" was missing a profit column.

* Mitigation: Remove outliers from DOB.
* Recommendation: Add an age\_column to make data more understandable and easier to check for errors. Create a profit\_column in "Transactions" to check the accuracy of sales.

Creating additional columns for age and profit will allow for easier error detection. Profit\_column will be useful in future financial analysis.

* **Completeness:**

Additional customer\_id varied between "Customer Demographic," "Customer Address," and "Transactions."

* Mitigation: Filter all customer\_id between 1 and 3500.
* Recommendation: Manage tables up to date (from the same time period). Only customer\_id ranging from 1 to 3500 will be used in our model because they contain complete data.

The data received may not be in sync across all spreadsheets, and analysis results may be skewed due to incomplete data. This is a completeness issue; to avoid future occurrences, cross-check spreadsheets and sync data.

Blanks in job\_title for "Customer Demographic," as well as in online\_order and brand\_column for "Transactions."

* Mitigation: Remove 'blanks' from job\_title, online\_order, and brand\_column.
* Recommendation: Move job\_title to another category, such as industry\_industry, or provide job\_title dropdown options. Give online\_order and brand\_column dropdown options.

Blanks are treated as incomplete data, which can skew the results of subsequent analyses. The addition of dropdown options will result in more complete data and more accurate analysis.

* **Consistency:**

Gender inconsistency in "Customer Demographic" and "Customer Address"

* Mitigation: For gender, filter all 'M' under 'Male', all 'Femal' and 'F' under 'Female'. For states, change all 'New South Wales' to 'NSW' and all 'Victoria' to 'VIC'.
* Recommendation: Create dropdown options for 'Male,' 'Female,' and 'U' in gender. Make a dropdown menu for all state abbreviations.

Dropdown menu options reduce manual entry and human error. Allows for increased terminology consistency. Gender identity is a delicate subject; use caution when developing options.

* **Currency:**

People with letter 'Y' in the deceased indicator are not current "Customer Demographic" customers.

* Mitigation: Filter out customers who checked 'Y' in deceased indicator.
* Recommendation: Checking for deceased customers can be difficult, but once this information is received, data should be updated accordingly.

Deceased customers are no longer active customers; removing them from data will improve data currency and result in more accurate estimates in future analysis.

* **Relevancy:**

In default column for "Customer Demographic" and order status for "Transactions," there is a lack of relevance or comprehensibility.

* Mitigation: Removed Metadata from default column. Remove order status that is 'Cancelled'.
* Recommendation: Check for incoherent Metadata and delete for format to make it comprehensible.

Canceled order\_status is irrelevant for future analysis because it can skew data—for example, the total number of customers per year will be overestimated.

* **Validity:**

List\_price and product\_sale\_date formats for "Transactions"

* Mitigation: product\_sale\_date should be converted to short date format, and list\_price should be converted to currency.
* Recommendation: When entering new data, setup columns so that formats such as price and decimals are already present.

Tolerable values will make data easier to interpret. Formatting the price and allowing for 2 or 3 decimals to be placed consistently will improve readability. This will have a positive impact on the speed and accuracy of business decisions.

This summarizes all of the data quality issues discovered during the first stage of data quality analysis. The mitigation strategies proposed are simple and effective methods of improving data quality for future analysis. They will not only improve the quality of analysis that can be performed within the company, but will also raise the level of analysis that KPMG and other hired analysis teams can perform.

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

Best regards,

Shabrina