Dear Sprocket Central Pty Ltd,

Thank you for providing the three datasets from Sprocket Central Pty Ltd. The summary table below highlights key quality issues that we discovered within the three datasets. Please let us know if you have any queries regarding the issues presented.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Accuracy** | **Completeness** | **Consistency** | **currency** | **Relevancy** | **Validity** |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Customer** **demographic** | **DOB:** Inaccurate  **Age**: missing | **Job title:** banks  **Customer id:** incomplete | **Gender:**  inconsistency | **Deceased,**  **Customers:** filter out |  |  |
| **Customer Address** |  | **Customer id:** incomplete | **States:** inconsistence |  |  |  |
| **Transactions** | **Profit:** missing | **Customer id:**  incomplete  **Online order:**  Blanks  **Brand:** blanks |  |  | **Cancelled status order:** filter out | **List price:** format  **Product sold date**: format |

Below is more depth description of data description of data quality issues discovered and method of mitigation used. Recommendations and explanations have also been included to avoid further data quality issues in the future. The following recommendations will improve the accuracy of data used to influence the 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 for, allowing for more comprehensible data and easier to check for errors. Create a **profit column** in “**Transactions”** to check accuracy of sales.

**Completeness**

* **Additional customer id was inconsistent among “customer Demographic”, “Customer address” and “Transactions.”**

Mitigation: Filter all **customer ids from 1 to 3500**

Recommendations: Ensure tables are up to date (from the same time). For our model, only **customer id from 1 to 3500** will be used as they have complete data.

The data received may not be synced 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 the job title for “Customer Demographic” in online order and brand column for “Transactions”**

Mitigation: Filter out “blanks” for **job title and brand column.**

Recommendation: Simply job title to another category such as **industry** or dropdown options for **job title.** Provide dropdown option for **online order and brand column.**

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

**Consistency**

* **Inconsistency in gender for “Customer Demographic” and “Customer Address” respectively**

Mitigation: Filter out all ‘M’ under category of “Male” filter all ‘F’ under “Female” for gender. Filter all “New South Wales” to “NSW” and “Victoria” to “VIC” for **states.** Recommendation: Create dropdown option for “Male”, “Female” and ‘U’ for all **state** abbreviations.

Dropdown options, minimizes manual entry and human errors. Allow for increase of consistency of terminology. Gender identifies can be a sensitive topic, so proceed with caution when creating options.

**Currency**

* **People that are ‘Y’ in deceased indicator are not current customers for “Customer Demographic” column.**

Mitigation: Filter out customer check ‘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 the 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 format, **list price** to currency.

Recommendation: set up column so that formats such as price and decimals are already in place when entering new data.

Allowable values will make data interpreted more easily. Format

Into price and allowing for either 2 or 3 decimals placed consistently will increase readability. This will reflect the positivity regarding speed and accuracy of analysis for business decisions.

That summaries 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 any questions regarding mitigation or any data quality issues identified.

Kind regards,

Prakash