Dear Sprocket Central Pty Ltd,

Thank you for providing us with the datasets. The tables below, highlights the summary statistics from the three data sets received. Table 2 shows the key quality issues that we discovered within the dataset, from your company Sprocket Central Pty Ltd.

Table 1: Data Summary

|  |  |  |  |
| --- | --- | --- | --- |
| Table name | No. of records | Customer IDs | Date Data Received |
| Customer Demographic | 4000 Entries | 4000 IDs | 12/4/2021 |
| Customer Address | 3999 Entries | 3999 IDs | 12/4/2021 |
| Transaction Data | 20000 Entries | 20000 IDs | 12/4/2021 |

Table 2: Data Quality Overview

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table name | Accuracy | Completeness | Consistency | Currency | Relevancy | Validity |
| Customer Demographic | \*DOB:  Inaccurate  \*Age; Missing | \*Job Title:  Blanks | \*Gender  inconsistency | \*Deceased customers:  Filter out | \*Default column: delete |  |
| Customer Address |  | \*Customer id: Incomplete | \*States:  inconsistency |  |  |  |
| Transactions | \*Profit: Missing | \*Online Order: blanks  \*Brand: blanks |  |  | \*Cancelled status order: filter out | \*List price: format  \*Product sold date: format |

Below are notable data quality issues that were discovered and the methods used to mitigate the identified data inconsistencies. Furthermore, recommendations have been provided to avoid the reoccurrence of data quality issues and improve the accuracy of the underlying data used to make business decisions.

**ACCURACY**

* **\***DOB is inaccurate for dataset “Customer Demographic” and missing an “AGE’ Column.
* \*For “Transactions” dataset, “Profit column” is missing.

*Mitigation: Filter out outliers in DOB.*

*Recommendation: Add an “AGE” column, allowing for more comprehensive data, also create a “PROFIT” column for “Transactions” to check accuracy of sales.*

Adding columns “AGE” and “PROFIT” will allow for easier identification of errors and assist in monetary analysis.

**COMPLETENESS**

* **\***Customer IDs are inconsistent for “Customer Address” dataset.

*Mitigation: Filter all Customers IDs*

*Recommendation: Ensure tables are up to date (from the same time period).*

The data received may not be in sync across all spreadsheets, with incomplete data the analysis results may not be accurate. It is recommended cross check spreadsheets and sync data.

* \*Blanks in Job titles for “Customer Demography”, in online order and brand column for “Transactions”.

*Mitigation: Remove “blanks” for job title, online order and brand column.*

*Recommendation: Provide dropdown options for job title, online order, and brand columns.*

Blanks are treated as incomplete data and can affect analysis results. The addition of dropdown options will result to more accurate analysis.

**Consistency**

* **\***Inconsistency in gender for **“**Customer Demographic” and “Customer Address”.

*Mitigation: Replace all “M” with ‘Male’ under category Male and “female”, ‘F’, with “Female” for gender. Replace all “New South Wales” with “NSW’, and “Victoria” with “VIC” for states.*

*Recommendation: Provide dropdown options for all state abbreviations.*

Dropdown options, minimizes human error, increases consistency of terminology.

**Currency**

* \*People that are “Y” in deceased indicator are not current customers for “Customer Demographic”

*Mitigation: Remove 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.*

**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.*

**Validity**

* \*Format of list price, product sale, date for “Transactions”

*Mitigation: Format product sale date to short format, and list price to currency.*

*Recommendation: Set up columns so that formats such as price and decimal 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.

This summarizes the data quality issues discovered, moving forward, the team will continue with the data cleaning, standardization and transformation process for the purpose of model analysis.

Please let us know, if you have any questions or reservations regarding this data quality issues discovered.

Best Regards,

RACHAEL OJOCHIDE IDAKWO