Dear Sir/Ma’am,

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

**Summary Table:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Datasets** | **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** |
| **Customer Demographic** | DOB:  inaccurate  Age:  missing | Job Title: blanks  Customer id: incomplete | Gender: inconsistency | Deceased Customers:  filter out | Default Column:  delete |  |
| **Customer**  **Address** |  | Customer id:  incomplete | States:  inconsistency |  |  |  |
| **Transactions** | 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 description of data quality issues discovered and methods of migration used. Recommendation 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.

**Issues:**

* **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 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 Issues**:

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

* **Currency Issues:**

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 estimate in future analysis.

* **Relevancy Issues:**

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 overestimate.

* **Validity Issues:**

Format of **list\_price, product\_sale\_date** for **“Transactions”.**

*Mitigation:* Format **product\_sale\_date** to short data 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 male data to 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 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 questions regarding mitigation or any data quality issues identified.

Have a nice day.

Kind regards,

Mandar Vaidya