Dear ----,

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

**Summary Table:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** |
| **Customer Demographic** | * DOB inaccurate * Age: missing | * Job Title:   Blanks   * Customer id: incomplete | * Gender: inconsistency | * Deceased customer: filler out | * Default column: delete |  |
| **Customer Address** |  | * Customer id: incomplete | * State: inconsistency |  |  |  |
| **Transactions** | * Profit: missing | * Customer id: incomplete * Online orders: blanks * Brands : blanks |  |  | * Canceled status order : filler out | * List Price: format * Product sold date: format |

Below are more in depth description of data quality:

**Accuracy Issue:**

* **DOB was in accurate for “Customer Demographic” and missing Age\_column; missing a profit column for “Transaction”.**

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 “Transaction” to check accuracy for sales.

Creating additional columns for age and profit will allow for easier identification of error. The profit\_column will assits in future monetary analysis.

**Completeness**

* **Additional Customer\_ids were inconsistent among “Customer Demographic”, “Customer Address”, and “Transactions”.**

Mitigation: filter all customer\_ids

Recommendation: ensure that table is up-to-date (from the same time periods)

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 occurrence it is encouraged it is to cross check.

* **Blanks in job\_title for “customer Demographic” in online\_order and brand\_column for “Transactions”.**

Mitigation: filter out all blanks for job\_title, online\_order and brand\_column.

Recommendation: simplify job\_title to another category such as industry or provide drop down options for job\_title. Provide drop down for online\_order and brand\_column.

Blanks are treated as incomplete data and can be skew future analysis the result . In addition of drop down option will allow to have complete data and will results in more accurate analysis.

**Consistency**

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

Mitigation: filter out all :

**Gender: 1. M -> Male**

1. **F -> Female**
2. **Femal -> Female**

**States: 1. New South Wales -> NSW**

1. **Victoria -> VIC**

Recommendation: Provide drop down for all state and gender.

Drop down options , minimize manual entry and human error. Allows for increase of consistency of terminology. Gender identity can be a sensitive topic , proceed with caution when creating options.

**Currency**

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

Mitigation: filter out all customer check ‘Y’

Recommendation: can be difficult to check for deceased\_indicator, but once this information received one should update data accordingly.

Deceased customer are not the current customers, removing them from data will increase currency of data and will result in more accurate estimate in future analysis.

**Relevancy**

* **Lack of relevancy or comprehensibility in default\_column for “Customer Demographic” and order\_status for “Transactions”**

Mitigation: filter out all ‘canceled’ order\_status and deleted default\_column

Recommendation: check for incomprehensible metadata and delete or format to made comprehensible.

‘canceled’ order\_status is the irrelevant data for future analysis.

**Validity**

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

Mitigation: Format product\_sale\_date to short date and list\_price to currency.

Recommendation: setup 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 not allowing for either 2 or 3 decimals placed consistently will increase the readability.

That summaries the 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 improved the analysis tat 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 question regarding mitigation or any data quality issues identified.

Kind regards,

Poulami Bakshi