Dear Manager,

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table name | Accuracy | Completeness | Consistency | Currency | Relevancy | Validity |
| Customer demographic | DOB:  Inaccurate  Age:  Missing | Job title:  Blank | Gender:  Inconsistency | Deceased  Customer:  Filter out | Default  Column:  Delete |  |
| Customer address |  | Customer id:  Incomplete | State:  Inconsistency |  |  |  |
| Transaction data | Profit:  Missing | Customer id:  Incomplete  Online orders:  Blanks  Brand: Blanks |  |  | Cancelled  Status  Order:  Filter out | List price: Format  Product sold date: Format |

Notable data quality issues that were encountered and the methods used to mitigate the identified data inconsistencies are as follows. Furthermore, recommendations have been provided to avoid the reoccurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

* **DOB was inaccurate for “Customer Demographic” and missing an 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 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.

* **Additional customer\_ids were inconsistent among “Customer Demographic”, “Customer Address”, and “Transaction”**

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 occurrence

It is encouraged to cross check spreadsheets and sync data.

* **Blanks in job\_title for “Customer Demographic”, in online\_order and brand\_column for “Transaction”**

Mitigation: Filter out ‘blanks’ for **job\_title, online\_order**, and **brand\_column**.

Recommendation: Simplify **job\_title** to another category suck as industry or provide dropdown option for **job\_title**. Provide dropdown options for **online\_order** and **brand\_column**.

Blanks are treated as incomplete data and can skew further analysis results. The addition of dropdown option will allow to have more complete data and will result in more accurate analysis.

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

Mitigation: Filter all ‘M’ under category of ‘Male’, filter all ‘Femal’ and ‘F’ under ‘Female’ for **gender**. Filter all ‘New South Wales’ to ‘NSW’ and ‘Victoria’ to ‘VIC’ for **states**.

Recommendation: Create dropdown options for ‘Male’, ‘Female’, and ‘Unspecified’ in **gender**. Create dropdown options for all **state** abbreviations.

Dropdown options minimizes 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.

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

Mitigation: Filter out customers checked ‘**Y**’ in **deceased\_indicator.**

Recommendation: Can be difficult to check for deceased customer, but once this information is received one should update data accordingly.

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

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

Mitigation: Deleted metadata in default\_columns. 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 customer per annum will be an overestimate.

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

Mitigation: Format **product\_first\_sale\_date** to short date 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 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.

That summarises 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 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 or any data quality issues identified.

Kind regards,

Amandeep Singh