Dear John Doe,

Thank you for providing us with 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 surrounding the issues presented.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevance** | **Validity** |
| **Customer Demographic** | DOB: Inaccurate  Age: Missing | Job title: Blanks  Customer id: incomplete | Gender: Inconsistency | Deceased customer: 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 descriptions of data quality issues discovered and methods and mitigation used. Recommendations and explanations have also been included to avoid further data quality issues in the future. The following recommendations will improve accuracy of data used to influence 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**, allowing for more comprehensible data and easier to check for errors. Create a **profit\_column** in “Transactions” to check for accuracy in sales.

Creating additional columns for age and profit will allow for easier identification of errors. The **profit\_column** will assist in future analysis.

**Completeness**

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

***Mitigation***: Filter out all **customer\_ids** from 1 to 3500.

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

The data obtained may not be synced across all spreadsheets, with incomplete data the analysis results may be skewed. This is a completeness issue and to prevent such cases, it is encouraged to cross check spreadsheets to sync data.

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

***Mitgation***: Filter out “blanks” for **job\_title, online\_order** and **brand\_column**.

***Recommendation***: Simplify **job\_title** or provide a drop down options for **job\_title**, **online\_order**, and **brand\_column**

Blanks are treated as incomplete data and can skew results. Dropdown options will allow for more complete data and will result in accurate analysis.

**Consistency**

* Inconsistency in **gender** for “Customer Demographics” and “Customer Address”

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

***Recommendation***: Create a dropdown options for ‘Male’, ‘Femal’, and ‘U’ in **gender**. Create dropdown options for all **state** abbreviations.

Dropdown options will minimize manual entry and human error. Allows for increased consistency of terminology.

**Currency**

* 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 the date accordingly.

Deceased customers are not current customers so removing them will greatly increase currency of the data and result in 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’ in **order\_status.**

***Recommendation****:* Check for incomprehensible metadata and delete or format to make it comprehensible.

‘Cancelled’for **order\_status** is irrelevant information and can skew data.

**Validity**

* Format of **list\_price**, **product\_sale\_date** for ‘Transactions’

***Mitigation****:* Format **product\_sale\_date** to short date format and 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 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 only improve the analysis output that one can perform within the company but will increase the level of analysis that be performed by KPMG and other hired analysis teams.

Please let us know if you have any questions regarding mitigation or data quality issues identified.

Kind regards,

Mohammad.