**Dear Client,**

Thank you for providing the datasets of Sprocket Pty Ltd. The table below highlights the summary of statistics from all the datasets we received. Please let us know if the information below is not aligned with your understanding.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Accuracy | Completeness | Consistency | Currency | Relevancy | Validation |
| Customer Demographic | DOB: Inappropriate  Age: Null | Job title: Blanks  Customer ID: Incomplete | Gender: Inconsistent | Deceased pointer: removed dead customers. | Default:  Irrelevant column |  |
| Transaction | Profit column added | Customer ID: Incomplete  Online Order: Blank  Brand: Blanks |  | Status: Cancelled orders are filtered |  | List price : Format issues  Product sold date: Format issue |
| Customer Address |  | Customer ID:  Incomplete | State: Inconsistency |  |  |  |

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

**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*** *check accuracy for sales and easier to check.*

Creating additional columns for age and profit will allow for error.The ‘profit’ column will assist in future monetary analysis.

**Completeness:**

* *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.

* *Blanks in job\_title for "Customer Demographic," in online order and brand\_column for "Transactions"*

***Mitigation:*** *Filter out 'blanks' for job\_title, online order, and brand\_column.* ***Recommendation:*** *Simplify job\_title to another category such as industry\_industry or provide dropdown options for job\_title. Provide dropdown options for online order and brand\_column.*

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

**Consistency:**

* *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, Femal,' and 'U' 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.

**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 data accordingly.*

Deceased customers are not current customers, removing them from data will increase currency of data and will result in more 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' 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 an overestimate.

**Validity:**

* Format of list price, product\_sale\_date for "Transactions"

***Mitigation:*** *Format product\_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 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.

With regards,

Mahilan Sathasivam.