Dear Client,

I am Tijani Mubarak Adewale from KPMG data analytics (virtual internship) team. My team and I have discovered some key quality issues with the three data sets provided based on the following data quality dimensions: Accuracy, Completeness, Consistency, Currency, Relevancy, Validity, and uniqueness. The summary table below shows a concise report on the dimensions mentioned above. Thanks for providing us with the data sets from Sprocket Central Pty Ltd. Please do not hesitate to let us know if you have any queries whatsoever, concerning the issues presented.

**Data Quality Assessment Table**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** |
| **Transactions** | * Profit Column:   Missing. | * Customer id:   Incomplete   * Online order:   Blank cells   * Brand:   Blank cells. |  |  | * Cancelled order status:   Filter out. | * List price: Format. * Product sold date:   Format. |
| **Customer Address** |  | * Customer id:   Incomplete. | * States:   Inconsistency. |  |  |  |
| **Customer**  **Demographic** | * DOB:   Inaccurate   * Age:   Missing | * Job title:   Blank cells. | * Gender:   Inconsistency. | * Deceased customers:   Filter out. | * Default column:   Drop/remove. |  |

For a more detailed Assessment report, we dim it fit to give a more in-depth description of data quality issues discovered and methods of mitigation used. To avoid cases of data quality issues in future we have made detailed explanations and recommendations on how data should be represented and handled to improve accuracy of data used to influence business decisions of Sprocket Central Pty Ltd in the future.

**Accuracy Issues:**

* **Inaccurate DOB in “Customer Demographic” and missing age\_column;**

**Missing profit column in “Transactions”.**

*Mitigation: Filter out outlier in* ***DOB.***

*Recommendation: Take a mode year value for the missing records of customer’s* ***DOB****. Create an* ***age\_column*** *to store ages as it is more comprehensive and easier to examine for errors. Create a* ***profit\_column*** *in “****Transactions****” to store profits for easy examination of accuracy of sales.*

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

**Completeness:**

* **Inconsistency of customer\_ids among “Customer Demographic”, “Customer Address”, and “Transactions”.**

*Mitigations: Filter thoroughly all* ***customer\_ids from 1 to 3500***

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

Skewed analysis results is caused by incomplete data, now data received may not be sync across all spreadsheets. This is a ‘completeness issue, to prevent future occurrences, it is encouraged to cross check spreadsheet and sync data.

* **Blank cells in job\_title in “Customer Demographic”, and blank cells in online\_order and brand\_column for “Transactions”.**

*Mitigation: filter out blank cells for* ***job\_title****,* ***online\_order****, and* ***brand\_column****.*

*Recommendation: Endeavour to provide a dropdown option for* ***job\_title****,* ***online\_order*** *and* ***brand\_column.***

The use of dropdown options will allow to have more complete data to work with and will result in more accurate analysis, as blank cells are treated as incomplete data and can skew further analysis results.

Consistency

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

*Mitigation: Filter all ‘M’ under the category of “Male” and replace with “Male”, filter out all ‘femal’ and ‘F’ under female and replace with “Female” in* ***gender****. Filter all “New South Wales” and replace with ‘NSW’, and filter all ‘Victoria’ and replace with ‘VIC’ for* ***states****.*

*Recommendation: Create dropdown for ‘Male’, ‘Femal’, and ‘U’ in gender. Create a 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 sensitive topic, proceed with caution when creating options.*

Currency:

* **‘Y’ in deceased\_indicator in “Customer Demographic” indicates that those people are dead and are not current customers.**

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

*Recommendation: Information should be updated as soon as it is received accordingly.*

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

Relevancy

* **irrelevancy or incomprehensibility of default\_column in “Customer Demographic” and cancelled in order\_status in “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 in future analysis, as it can skew data, for example, total number of customers per annum will be overestimated.*

Validity:

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

*Mitigation:* Change Format **of product\_sales\_date** to short date format, and format **list\_price** to currency.

*Recommendation: Set up columns such that format such as price in decimal are already in place when entering new data.*

*Allowable values will make data to be interpreted more easily. Formatting into price and allowing for 2 or 3 decimal places consistently, will increase readability. This will reflect positively on speed and accuracy of analysis for business decisions.*

Here you have it, a summary of 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 also 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.

Kind regards,

Tijani Mubarak.