Hello,

Hope you are doing well. Thank you for providing us with the datasets from Sprocket Central Pty Ltd. The summary table highlights the quality issue that I discovered in the data. Please go through the below table, and please let us know if you have any queries in this:

**Summary table:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Accuracy of Data** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** |
| **Customer Demographic** | 1. DOB inaccurate 2. Age missing |  | Inconsistency for gender column | Deceased customers were removed | Default column deleted |  |
| **Customer Address** |  | Customer id was incomplete |  |  |  |  |
| **Transactions** | Profit Column was missing | 1. Online order status was blank 2. Brand column was blank |  |  | Filtered out cancelled order status | Product sold date format |

Please find the detailed description of the data quality issues discovered and methods of mitigation used. Additionally, recommendations and explanations have been provided to avoid the reoccurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

**Accuracy issue of dataset:**

Date of birth (DOB) was inaccurate for “customer demographic” and age column was missing from dataset, also profit column was missing from the dataset.

All the outliers were removed from DOB column.

*Recommendation:* Age column was created, for checking the errors

Profit Column was created to cross check the accuracy of the sales.

In addition, few additional columns were inserted in the “Transactions” sheet, which further will help in the identification of the errors in the dataset.

**Empty values in dataset:**

Blanks are present in job\_title in “Customer Demographic” sheet, and many blanks are present in online\_order column and brand\_column of “Transactions” sheet. Various columns, such as the brand of a purchase, or job title, have empty values in certain

*Mitigation:* If the number of empty rows in the dataset is less than 2% of the entire dataset, then that rows can be easily removed, form the training dataset. In addition, if the number of empty rows is greater than 2%, then it can be imputed with mean or median of the row based on the type of dataset.

In this dataset, many rows contain the blank row, which is contributed to less than 1% of the total dataset. These records have been removed from the training dataset.

**Consistency in the data:**

Inconsistency in gender column for “customer demographic” sheet.

*Mitigation:* Replaced all ‘M’ with ‘Male’, ‘f’ and ‘Femal’ with ‘Female’ for gender column. Moreover, replaced ‘NSW’ with ‘New South Wales’, and ‘VIC’ with ‘Victoria’ for states column.

*Recommendation:* Dropdown option should be created for the ‘Male’, ‘Femal’ and ‘U’ in the gender column.

**Inconsistent data type:**

Inconsistent data type for the same attribute (e.g. numeric values for some fields and strings for others).

*Mitigation*: Convert selected records in characters to numeric. Remove non-numeric characters from string.

*Recommendation*: Ensure that fact tables in the given database have constraints on data types. Having different data types for a given field make it difficult to interpret results at the later stage.

**Relevancy:**

Lack of relevancy of data in default column for “Customer Demographic” and “order- status” column in “Transaction” sheet

*Mitigation:* Filtered out ‘Canceled’ order\_status.

*Recommendations:* Check for incomprehensible metadata and delete or format to make the data comprehensible.

Cancelled oreder\_status is an irrelevant data for the future analysis, as it is a skewed data.

**Validity:**

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

*Mitigation:* Format product\_sale\_date to short date format, format list\_price to currency.

Recommendations: Set up columns so that format such as price and decimals are already in place when entering the data.

Allowable values will make the data to be interpreted more easily. Formatting in the price and allowing for 2 or 3 decimal places, will majorly increase the reliability of the data.

**Unsynchronized data**

Ensured that the provided data re firm the same time period. Additional customer\_ids are present in “Transactions table” and “Customer Address table” but not in “Customer Demographic”

Provided data is skewed data, which will impact the analysis of the dataset, which will reduce the efficiency of the dataset.

*Mitigation:* Please ensure that all tables are from the same period. Only customers in the Customer Master list will be used as a training set for our model. This indicates that the data received may not be in sync with each other which may skew the analysis results if there are missing data records.

That summarizes all the data quality issues discovered through the first stage of the data quality analysis. The mitigation strategy suggested are simple and effective ways of improving the data quality for future analysis. This will majorly reduce the time required for the analysis and improves the analysis output.

Please let us know if you have any queries with the data quality issues or the mitigation strategy suggested. Feel free to connect with us.

Thanks and regards,

Onkar Swami.