Dear Sir,

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The below table highlights the summary statistics and quality issue from the three datasets received. Please let us know if the figures are not aligned with your understanding.

**Summary Table:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Accuracy of Data** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** |
| **Customer Demographic** | 1. DOB inaccurate 2. Age missing |  | Inconsistency for gender column | Deceased customers was 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 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Table name** | **No. of records** | **Distinct customer ids** | **Date data received** |
| **Customer demographic** | 4000 | 4000 | 07-06-2020 |
| **Customer address** | 3999 | 3999 | 07-06-2020 |
| **Transaction data** | 20000 | 3496 | 07-06-2020 |

Notable data quality issues that were encountered and the methods used to mitigate the identified data inconsistencies are as follows. Furthermore, 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:**

DOB was inaccurate for “customer demographic” and age column was missing from dataset, 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 for checking 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 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. 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

*Mitigation:* Use regular expression to replaced extended values into abbreviations to ensure consistency across addresses

*Recommendation:* Dropdown option should be created for the ‘Male’, ‘Femal’ and ‘U’ in the gender column. Enforce a drop-down list for the user entering the data rather than a free text field. In order to construct meaningful variables for the model, the data has been cleaned to avoid multiple representations of the same value. Additionally, gender records where ‘U’ have been replaced based on the distribution from the training dataset.

**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. Therefore, appropriate data transformations are made to ensure consistent data types for a given field.

**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.

‘Canceled’ 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 in 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 any question on this, or any other data quality issue identified.

**With thanks and sincere regards,**

**Shinjini Das**