Dear Manager,

Thank you for providing us with three datasets from Sprocket Central Pty Ltd. The table below shows the summary statistics from the three datasets. Please let us know if you have any queries surrounding the issues presented.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Accuracy | Completeness | Consistency | Currency | Relevancy | Validity |
| Customer Demographic | DOB: inaccurate  Age: Missing | Job Title: Blanks  Customer\_id: incomplete | Gender: Inconsistency | Decreased 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 solid date: format |

|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **No of Records** | **Distinct Records** | **Date data received** |
| **Transactions** | 200001 | 3496 | 07/20/2022 |
| **Customer Address** | 3999 | 3999 | 07/20/2022 |
| **Customer Demographic** | 4000 | 4000 | 07/20/2022 |

* NULL values: various columns such as brand columns in Transactions have many empty cells. These records have been removed from the datasets. If there are only few rows are empty, filter out the records entirely from the datasets. Else, If it is core field, impute based on distribution in the training dataset. For key datasets, such as transactions, less than 1% of transactions (total less than 0.1% of revenue) have missing fields.
* Inconsistent data type for the same attributes: Convert them into correct format, for example, convert from numeric to date format, convert a string to numeric. Recommendation: Make sure that columns given in the table are consistent in data types. It is difficult to interpret results from having different data types. Additionally, gender records where ‘U’ have been replaced based on the distribution from the training dataset.
* Key field (customer\_id) is contained in Transactions, Customer Address and Customer Demographic tables but not exists in New Customer List. 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.
* Non-sense data: The contents in default column of Customer Demographic have been removed.

After we have completed this, it would be great to spend some time with your data SME to ensure that all assumptions are aligned with Sprocket Central’s understanding.

Kind regards,

Minh Thu Nguyen