Sir/Madam,

Thank You for providing the dataset for Sprocket Central Pt. Ltd. The following statistics reflect data as received by our team, please reach out if there is a misunderstanding:

|  |  |  |  |
| --- | --- | --- | --- |
| **DATASET** | **NO. OF RECORDS** | **CUSTOMER IDs** | **DATA RECIEVED** |
| **TRANSACTION** | 20,000 | 3,494 | 22 July 2021 |
| **CUSTOMER DEMOGRAPHIC** | 4,000 | 4,000 | 23 July 2021 |
| **CUSTOMER ADDRESS** | 3,999 | 3,999 | 23 July 2021 |

After receiving the dataset, we went ahead and evaluated them as per standard data quality dimensions to identify any data quality issues. It is also identified certain strategies that shall help un mitigating these issues in the future that shall help in future analysis and growth of the company.

The following table summarizes the various fields having issues that came to the notice:

|  |  |  |  |
| --- | --- | --- | --- |
| **PERFORMANCE** | **TRANSACTIONS** | **CUSTOMER DEMOGRAPHIC** | **CUSTOMER ADDRESS** |
| **ACCURACY** |  | * DOB: Inaccurate * Job Industry Category: Misspelled |  |
| **COMPLETENESS** | * Standard Cost: Blank * Brand: Blank * Product Line,Size,Class,First Sold Date: Blank * Customer Id: Not Synchronized | * DOB: Blank * Job Title: Blank * Job Industry Category: Blank * Tenure: Blank * Customer Id: Not Synchronized | Customer Id: Not Synchronized |
| **CONSISTENCY** |  | Gender: Inconsistent | States: Inconsistent |
| **CURRENCY** |  | Deceased Customer: Filtered |  |
| **RELEVANCY** | Order Status: Exclude Cancelled | Default: Exclude Field |  |
| **VALIDITY** | Product First Sold Date: Format |  |  |

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

● **Additional customer ids** in the ‘Transactions table’ and ‘Customer Address table’ but not in ‘Customer Master (Customer Demographic)’ **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.

● Various columns, such as the brand of a purchase, or job title, have empty values in certain records **Mitigation:** If only a small number of rows are empty, filter out the record entirely from the training set for prediction. Else, if it is a core field, impute based on distribution in the training dataset. For key datasets, such as transactions, less than 1% of transactions (totalling less than 0.1% of revenue) have missing fields. These records have been removed from the training dataset.

● Inconsistent values for the same attribute (e.g. Victoria being represented as “V”, “Vic” and “Victoria”) Mitigation: Use regular expression to replaced extended values into abbreviations to ensure consistency across addresses.

● 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.

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.

Regards

Surmeet Mohanty