Dear Sprocket Central,

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The below table highlights the summary statistics from the three datasets received.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **No. of Records** | **Distinct Customer IDs** | **Date Data Received** |
| **Transaction Data** | 20000 | 3494 | 13-10-2022 |
| **Customer Demographic** | 4000 | 4000 | 13-10-2022 |
| **Customer Address** | 3999 | 3999 | 13-10-2022 |

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

**Customer Demographic (Total records 4000)**:

|  |  |
| --- | --- |
| **FIELD NAME** | **ERRORS** |
| DOB | 1 record of Year: 1843.  87 Blank records. |
| last\_name | 125 Blank records. |
| Gender | 88 records of gender ‘U’  Values are inconsistent. E.g.: M, Male, F, Female, Femal, U. |
| job\_title | 506 Blank records. |
| job\_industry\_category | 656 Blank records. |
| Default | 3317 records had value ‘special characters’ and Blank records. |
| Tenure | 87 Blank records. |

**Transactions (Total records 20000 - 3 months)**:

|  |  |
| --- | --- |
| **FIELD NAME** | **ERRORS** |
| online\_order | 360 Blank records. |
| brand | 197 Blank records. |
| product\_line | 197 Blank records. |
| product\_class | 197 Blank records. |
| product\_size | 197 Blank records. |
| standard\_cost | 197 Blank records. |
| product\_first\_sold\_date | 197 Blank records. |

We suggest the following mitigates in order to improve the data quality, which will eventually help us to drive better analytical results for your company:

* We can take a mean year value for the missing records of customers DOB.
* We can assign a uniform last name for customers whose values are missing.
* Replace gender ‘U’ with reference to the customer’s name and make consistency.
* For tenure values, we can take mean of rest of the values and assign the mean value to the missing fields in order to maintain the consistency of data.
* Eliminate the blank orders considering fake orders.
* 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.
* Please ensure that all the tables are from the same period. Only customers in the Customer Demographic list will be used as a training set for our model.
* Convert selected records in characters to numeric. Remove non-numeric characters from string.

Please let us know if the figures are not aligned with your understanding.

Regards,

Bhavin Sheth.