Dear Manager,

I hope you are keeping well.

Thank you for sharing the datasets. The below table shows the number of records and number of unique customers in each table. Request you to please revert us back if the numbers are not matched as per your understanding.

|  |  |  |
| --- | --- | --- |
| **Dataset Name** | **Number of Records** | **Number of Unique Customer ID** |
| CustomerDemographic (Master Table) | 4000 | 4000 |
| CustomerAddress | 3999 | 3999 |
| Transactions | 20000 | 3496 |

The issues related to Data Quality the action taken for creating a training dataset is mentioned below:  
**A. Drop Irrelevant columns:**Please find the below table representing a columns names which are not included in the training dataset. Please let us know if you want us to keep them for analysis.

|  |  |  |
| --- | --- | --- |
| **Customer Demographic** | **Customer Address** | **Transactions** |
| first\_name | address | transaction\_id |
| last\_name | country | list\_price |
| deceased\_indicator | property\_valuation | product\_first\_sold\_date |
| default |  |  |

**B. Missing Records:**1. Customer Demographic information is missing for 4 Customers.   
Their Id’s: {4001, 4002, 4003, 5034}  
2. Address information is missing for 5 Customers.   
Their Id’s {3, 10, 22, 23, 5034}  
3. Transaction information is missing for 507 Customers.   
Their Id’s are listed in the attached file name “TransactionDataMissing.xlsx”  
**Mitigation:** The data received is not in sync as there are missing data records. The data analysis will be done on the customers present in the Master table. Hence, missing information of these customers in the Customer Address table and Transactions table needs to be added.   
*Training Data: Only those customers are included in the training dataset whose data is present in Address and Transaction table.*

**C. Missing Values:**  
Below table shows the column names containing the number of missing values:

|  |  |  |  |
| --- | --- | --- | --- |
| **Customer Demographic** | Number of Missing Values | **Transactions** | Number of Missing Values |
| DOB | 87 | Brand | 197 |
| job\_title | 504 | product\_line | 197 |
| job\_industry\_category | 655 | standard\_cost | 197 |
| Tenure | 87 | product\_class | 197 |
|  |  | product\_size | 197 |

**Mitigation:** Please ensure that the correct data should be filled for above mentioned columns. This would help us understand the customers interest and their buying behaviour.

**D. Inconsistent Values:**1. In Customer Demographic table column name ‘gender’ have inconsistent values as below:  
['F', 'Male', 'Female', 'U', 'Femal', 'M'] which has been replaced by ['Female', 'Male', 'U']  
Please ensure the value for category ‘U’

2. In Customer Demographic table column name ‘DOB’ of customer id 34 is incorrect as for this value the age of the customer shows 177.

3. In Customer Address table column name ‘state’ have inconsistent values as below:  
['New South Wales', nan, 'QLD', 'VIC', 'NSW', 'Victoria'] which has been replaced by ['New South Wales', nan, 'QLD', 'Victoria']

Could you please direct me to the person who can guide with the information needed for creating a better dataset for training the model.

Thank you.

Kind regards,  
Sunita Pandharmise