Hello Client,  
Hope you are doing well.  
  
Thank you for providing the four datasets to our team. The team went through the dataset and has compiled the table below which highlights the summary statistics. Please inform us if our figures don’t match with your understanding about the dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| Table Name | Date Data Received | No of Records | Distinct Customer ID’s |
| Customer Demographic | 10/10/2020 | 4000 | 4000 |
| Customer Address | 10/10/2020 | 4003 | 4003 |
| Transaction Data | 10/10/2020 | 20,000 | 3494 |
| New Customer List | 10/10/2020 | 1000 | 1000 |

The following data quality issues were encountered in the dataset and the following are the ways in which we can provide strategies for mitigating these issues.

1. **Few incorrect DOB entries & Inconsistent Data Types [Accuracy & Validity Issue]** -

**Mitigation** : Setting a lower limit for the birth year to be entered so it doesn’t give us impossible ages like 176 years. Convert selected records in characters to numeric and removing nonnumeric characters from strings. There must be constraints on data types.

1. **Additional Customer ID’s in the Transactions Table and customer address table. [Uniqueness Issue]** –

This indicates that the analysis results may be skewed if there are missing data records which may be due to unsynchronous behavior of the data.

**Mitigation** : Time based verification that means checking if all tables are from the same period. Training set for our model should only constitute customers from Customer Demographic table.

1. **Various Columns like brand purchase, tenure and DOB are empty [Completeness Issue**] –

Less than 1% of transactions have missing fields so we remove those records from our training dataset.

**Mitigation :** Make sure all data is being entered only once all are parameters. If only a small amount of rows are empty then we can delete those entries if the values are not predictor variables for our model.

1. **Inconsistent Values under same Attributes [ Consistency Issues] :**

Column with more than 2 values for the same terms for instance Victoria being shown as “V”, “Vic”, “Victoria. the data has been cleaned to avoid multiple representations of the same value.

**Mitigation** : Reducing this by removing redundant values with the third option getting rid of the third record [alternative]. Using regular expression to replace these extended values into abbreviations. This will help ensuring consistency.

Moving ahead, the team will continue with cleansing and transforming the data. All assumptions along the way will be documented.

As per the issues discussed and the strategies we would like to discuss the dataset with you before going ahead with the analysis.  
Do let us know your availability.  
  
Look forward to hearing from you,  
Thanks with regards,  
Ajinkya Bhavik