Hello,

This is Bhumika K R from KPMG Data Analytics (Virtual Internship) team. We have reviewed the data sets which were provided by your company and during the data quality analysis, we have found the some errors in the data sets.

Table Name	No of Records	Distinct Customer's ID
Customer Demographic	4000	4000
Customer Address	4003	4003
Transaction Data	20000	3494
New Customer List	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.

2) **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 asynchronous behaviour 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.

3) **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.

4) **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.

These assumptions will be considered for moving ahead with the cleaning process.

Thank you