Dear client,

First of all, thank you for the dataset, I would like to give you an overview of the data quality assignment. We went through your data and the following is the assessment.

|  | **Transaction sheet** | **Customer Demographic sheet** | **Customer Address sheet** |
| --- | --- | --- | --- |
| No of records | 2000 | 4000 | 3999 |
| No of unique records | 3494 | 4000 | 3999 |
| No of blank records | \* | \* | none |
| Receiving date | 24 June 2023 | 24 June 2023 | 24 June 2023 |

\* - contains blank records, look below for details

Problem with the dataset

1. **Blank records - Problem of data loss-** *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 the 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.
2. **Inconsistent datatype -** *Mitigation:* Convert selected records in characters to numeric. Remove non-numeric characters from a string. Recommendation: Ensure that fact tables in the given database have constraints on data types. Having different data types for a given field makes it difficult to interpret results at a later stage. Therefore, appropriate data transformations are made to ensure consistent data types for a given field
3. **Different/inconsistent values for same attributes** - *Mitigation:* Use regular expressions to replace extended values with abbreviations to ensure consistency across addresses. Recommendation: Enforce a drop-down list for the user entering the data rather than a free text field. In order to construct meaningful variables for the model, the data has been cleaned to avoid multiple representations of the same value. Additionally, gender records where ‘U’ have been replaced based on the distribution from the training dataset.
4. **Possible Data corruption and garbage values-** *Mitigation:* Make sure that Define criteria or rules that data must meet to be considered valid. Use data validation techniques such as data type checks, range checks, pattern matching, and consistency checks. Filter out or flag any data that does not meet the validation criteria.

#### Problems with - Transaction sheet

1. 360 values are blanks in online order True / False (E)- 18% of record
2. 197 blank cells under brand (G)- 9.8% of record
3. 197 blank cells under product\_line (H) - 9.8% of record
4. 197 blank cells under product class (I) - 9.8% of record
5. 197 blank cells under product\_size (J) - 9.8% of record
6. 197 bank cells under standard\_cost (L)- 9.8% of record
7. 197 bank cells under product\_first\_sold (M)- 9.8% of record
8. The data type is inconsistent in product\_first\_sold (M) and is not in date format

### Problems with - Customer Demographic sheet

1. 125 blank cells under last name (C)- 3.1% of records
2. inconsistent data under gender (D)
3. 1- out of range data - under - past\_3\_years\_bike\_related\_purchases
4. 87 - blank records - under - past\_3\_years\_bike\_related\_purchases-2.2% of records
5. 506 - blank records under - job title-12.65% of records
6. 87 - blank records - under-tenure- 2.2% of records
7. default column has garbage value

We are moving forward with the process of data cleaning, standardization and transformation for the purpose of data modelling. Questions and assumptions will be documented. It will be great to spend some time with your data.

Regards

Maitrayi Rajawat