Dear Sir,

This is to inform you that I’ve gone through the dataset that I’ve got, and would like to point the major issues that I’ve found in the dataset.

1. **Transactions**:

* online\_order, brand, product\_line, product\_class, product\_size, standard\_cost, product\_first\_sold\_date- has null values (approximately 1%), we may try imputation techniques to replace the null values.

1. **NewCustomerList**:

* Column no. 16, 17, 18, 19, 20 are missing. Please update me on the same.
* 'last\_name', 'DOB', 'job\_title'(10.6%), 'job\_industry\_category'(16.5%)- has null values which can be taken care of applying imputation techniques.
* ‘deceased\_indicator’ is a constant feature. Is it an important feature to consider?
* property\_valuation - Unit not mentioned.

Mitigation: You may add unit while defining the attribute so that units are consistent.

* Value- Unit not mentioned.

Mitigation: You may add unit while defining the attribute so that units are consistent.

1. **CustomerDemographic**:

* Gender- has typing mistakes.

Mitigation: You may try to use dropdowns so that entries are consistent.

* Default- has many inconsistent/garbage values. Please update me on the same.

1. **CustomerAddress**:

* property\_valuation - Unit not mentioned.

Mitigation: You may add unit while defining the attribute so that units are consistent.

1. **Additional customer\_ids in the ‘Transactions table’ and ‘Customer Address table’ but not in ‘Customer Master (Customer Demographic)’**

* Mitigation: Please ensure that all tables are from the same period. Only customers in the Customer Master list will be used as a training set for our model.
* This indicates that the data received may not be in sync with each other which may skew the analysis results if there are missing data records. Please refer to excel file ‘data\_outliers.xlsx’ for the list of outliers between tables.

1. **Inconsistent values for the same attribute (e.g. Victoria being represented as “V”, “Vic” and “Victoria”)**

* Mitigation: Use regular expression to replaced extended values into 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’ has been replaced based on the distribution from the training dataset.

I’ll need a confirmation on the above mentioned inconsistencies. Moving forward, the team will continue with the data cleaning, standardization and transformation process for the purpose of model analysis. Questions will be raised along the way and assumptions documented. After we have completed this, it would be great to spend some time with your data SME to ensure that all assumptions are aligned with Sprocket Central’s understanding.

Warm regards,

Dipankar Dey

Junior Data Analyst, KPMG