Data Quality Assessment

Assessment of data quality and completeness in preparation for analysis.

Here are some data quality issues and our suggestions based on these three tables which included Customer Demographic, Customer Address and Transaction. If you have any further questions about the following part, please let me know.

- **More than one columns contain many empty values. (e.g. job title from customer demographic.)**
  - *Suggestion: Remove those empty values in the training process if there are small number of null values.*
  - Solution: Only less than 1% data of transaction are empty. These records have been removed from training dataset.
- **"Customer id" can't match properly between these three tables. (e.g. Only transaction contains customer id equal 5043.)**
  - *Suggestion: Please check if the three customer tables from the same periods.*
  - Solution: Only transaction customer id will be used due to all three tables contain customer id from 0 to 3500.
- **Different data type in the same column(attribute). (e.g. Parts of DOB from customer demographic are text while the rest are date format.)**
  - *Suggestion: Make sure that fact tables in a given database have restrictions on data types.*
  - Solution: Change the format of the entire columns to the same format by using Excel or Python.
- **Inconsistent values for the same attribute. (e.g. Some of gender from customer demographic are F or Female.)**
  - *Suggestion: Enforce a drop-down list for the user entering the data instead of a free text field.*
  - Solution: Replace extended values into abbreviations by python to ensure consistency.