Dear sir or madam,

I am Qing from KPMG Data Analytics (Virtual Internship) team. We have reviewed the datasets provided by your company and we have some questions about them.

The data quality is the core in our analysis. However, we would like to discuss the following problems found in your datasets, which will help to lead to the better analysis or insights for your company’s success.

* Missing values in customers DOB and product\_first\_sold\_date. If they are relevant to the analysis, we might predict them by regression or use random imputation to get them;
* Missing values in last name of customers, either in new customer list or customer demographic. If last name is not relevant to the analysis, we might use the same last name to replace those missing values to keep those records;
* ‘U’ as a value in gender of either new customer list or customer demographic. We need to check with the author of the data if gender is relevant to the analysis;
* For tenure values, we can use mean or median imputation, regression imputation or other imputation methods to get the values to maintain the consistency of data, which depends on the objective of analysis;
* We need more explanation of the data values and collections of a few columns in customer related columns, such as property\_valuation, decreased indicator, tenure, default, Rank and Value. Especially, all the values in country column are Australia. We might could drop this column since all customer are from the same country;
* For customer born in 1843, we need to figure out if it is a typo and find the exact time of the year of birth. Besides, we could drop it as it is a fake records;
* Those missing values in product\_line, product\_class, product\_size columns in transactions and job\_title and job\_industry\_category columns in customer related datasets are hard to predict, we might use empty string to keep those records to maintain the variety of data for a better analysis.

The following are the details of problems encountered in all datasets:

# Customer NewCustomerList (Total records 1000)

|  |  |
| --- | --- |
| **column** | **problems** |
| last\_name | 29 missing values |
| DOB | 17 missing values |
| gender | 17 records with value ‘U’ |
| job\_title | 106 missing values |
| job\_industry\_category | 165 missing values |
| Rank, Value | Explanation of the meaning of data |
| country | Values are the same |
| postcode | datatype is int, due to the pandas importing |

# Customer CustomerDemographic (Total records 4000)

|  |  |
| --- | --- |
| **column** | **problems** |
| last\_name | 125 missing values,  DOB in 1843 |
| gender | 88 records with value ‘U’  Femal, M, F are not consistent |
| DOB | 87 missing values |
| job\_title | 506 missing values |
| job\_industry\_category | 656 missing values |
| decreased\_indicator | Explanation of the meaning of data |
| default | unreadable values |
| tenure | 87 missing values |
| country | Values are the same |

# Customer CustomerAddress (Total records 3999)

|  |  |
| --- | --- |
| **column** | **problems** |
| postcode | datatype is int, due to the pandas importing |
| country | Values are the same, Australia |
| state | New South Wales, QLD, VIC, NSW, Victoria are not consistent |
| property\_valuation | Explanation of the meaning of data |
| This table could merge with the CustomerDemographic, since they share the same customer\_id | |

**Transactions (Total records 20000 )**

|  |  |
| --- | --- |
| **column** | **problems** |
| Online\_order | 360 missing values |
| brand | 197 missing values |
| product\_line | 197 missing values |
| product\_class | 197 missing values |
| product\_size | 197 missing values |
| standard\_cost | 48 records Blanks |
| product\_first\_sold\_date | 197 missing values, datatype is float due to excel serial number as date |

Please feel free to contact me if you have any questions regarding the this quality analysis.

Best regards,

Qing