Hello,

This is Tuhin Halder 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.

Data quality analysis is require to process of scientifically and statistically evaluating data in order to determine and meet the quality required for business processes . We suggest the following mitigates in order to improve the data quality, which will eventually help us to driven the better analytics, results for your company.

* We can change Float to datetime format for the product sold date.
* We can take a mode year value for the missing records of customers DOB.
* Replace gender ‘U’ with reference to the customer name and make a consistency.
* Exploring the missing values in each dataset.
* For tenure values, we can take a mean of rest of the values and assign the mean value to the missing fields in order to maintain the consistency of data.
* Eliminate the blank orders considering fake orders.
* Check duplicate Values in each data set

The following are the details of error encountered in the data set.

**Customer Demographic (Total records 4000)**

|  |  |
| --- | --- |
| **FIELD NAME** | **ERRORS** |
| last\_name | 125 records Blanks |
| Gender | 88 records gender ‘U’, 1 record ‘Femal’, 1 record ‘F’, 1 record ‘M’  Values are not consistence M, Male, F, Female, Femal, U |
| DOB | 87 records Blanks |
| job\_title | 506 records Blanks |
| job\_industry | 656 records mention ‘N/A’ |
| Default | 3317 records value ‘special characters’ includes null and Blanks |
| Tenure | 87 records Blanks |

**Transactions (Total records 20,000)**

|  |  |
| --- | --- |
| **FIELD NAME** | **ERRORS** |
| Online\_order | 360 records Blanks |
| brand | 197 records Blanks |
| product\_line | 197 records Blanks |
| product\_class | 197 records Blanks |
| product\_size | 197 records Blanks |
| standard\_cost | 197 records Blanks |
| product\_first\_sold\_date | 197 records Blanks |

Regards,

KPMG (Data Analytics Team)