Hello,

This is Soni Kumari 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.

The data quality analysis is the core phase and due to errors in the data set, we suggest the following improvements in order to improve the data quality, which will eventually help us to driven the better analytics, results for your company.

* We can take a mode year value for the missing records of customers DOB.
* We can assign a uniform last name of customers, which values are missing.
* Replace gender ‘U’ with reference to the customer name and make a consistency.
* 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.
* The 'transaction\_date' and 'product\_first\_sold\_date' columns in the transaction sheet are stored as integers. Both colums should be changed to date.
* The integers in the date columns are not correct as it shows everything happening the same day at different times. We would need to refer back to the author of the data to seek more classification on how the data was collected.
* Missing values should be dropped on basis of our analysis.
* A profit section should be added in the transaction sheet.
* The default section in the Customer Demographics sheet can be deleted as it does not make any sense.
* In the customer address sheet the state column the ‘NSW’ and ‘Vic’ should be replaced as ‘New South Wales’ and ‘Victoria’ respectively.

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

# Customer Demographic (Total records 4000)

|  |  |
| --- | --- |
| **FIELD NAME** | **ERRORS** |
| DOB | 01 record 1843  87 records Blanks |
| last\_name | 125 records Blanks |
| Gender | 88 records gender ‘U’  Values are not consistence M, Male, F, Female, Femal, U |
| 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 20000 -past 3months)**

|  |  |
| --- | --- |
| **FIELD NAME** | **ERRORS** |
| Online\_order | 94 records Blanks |
| brand | 48 records Blanks |
| product\_line | 48 records Blanks |
| product\_class | 48 records Blanks |
| product\_size | 48 records Blanks |
| standard\_cost | 48 records Blanks |
| product\_first\_sold\_date | 48 records Blanks |

**Customer Address (Total records 4000)**

|  |  |
| --- | --- |
| **FIELD NAME** | **ERRORS** |
| States | Values are not consistent.  NSW, Vic, QLD, New South Wales, Victoria. |

Regards,

KPMG (Data Analytics Team)