Hello,

This is Swapnil Wagh 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 mitigates 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 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 |

**DATA QUALITY ASSESMENT**

**INCOMPLETNESS**

\* Many missing values are present in dataset

**INACCURACY**

\* column like default have unwanted script instead of proper data .it conation mixture of both the number ,text ,etc

**INCONSISTENCY**

\* many features have inconsistency

**REUNDANCY**

\* Gender column have redundancy like Male is written as Male and M & Feamle as Femal,Female,F

**NON-STANDARD**

\* many places data is not standard

\* Date also available in datetime format instead of date

**CONCLUSION:**

From above point we can say that the given Data have Poor Quality

Regards,

Swapnil Wagh

KPMG (Data Analytics Team)