Hello,

This is Akaanksha Mishra from KPMG Data Analytics (Virtual Internship) team. Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. We have examined the data sets and during the data quality analysis, we have found some issues in the data sets.

The data quality analysis is the core stage and due to faults in the data set, we suggest the following mitigates in order to improve the data quality, which will ultimately help us to driven the better analytics and outcomes 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; for the values that 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 to keep the consistency of data.
* Eliminate the blank orders considering fake orders.

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

**Customer Demographic (Total records 4000)**

|  |  |
| --- | --- |
| **FIELD NAME** | **FAULTS** |
| DOB | 01 record for 1843-12-21  87 records for Blanks |
| last name | 125 records for Blanks |
| Gender | 88 records for gender ‘U’  Values are inconsistent for M, Male, F, Female, Femal, U |
| job title | 506 records for Blanks |
| job industry | 656 records mentioned as ‘N/A’ |
| Tenure | 1. records for Blanks |

For Gender records, ‘M’ has been replaced with ‘Male’ and ‘F’ and ‘Femal’ with ‘Female’.

**Transactions (Total records 20000 -past 3months)**

|  |  |
| --- | --- |
| **FIELD NAME** | **FAULTS** |
| Online order | 360 records for Blanks |
| brand | 197 records for Blanks |
| product line | 197 records for Blanks |
| product class | 197 records for Blanks |
| product size | 197 records for Blanks |
| standard cost | 197 records for Blanks |
| product\_first\_sold\_date | 197 records for Blanks |

**Customer Address (Total records 3999)**

|  |  |
| --- | --- |
| **FIELD NAME** | **FAULTS** |
| state | Values are inconsistent for NSW and VIC |

For state records, ‘NSW’ has been replaced with ‘New South Wales’ and ‘VIC’ with ‘Victoria’.

Therefore, appropriate data transformations have been made to ensure consistent data types for a given field. For additional progress, the team will continue with the data cleaning, standardization, and transformation process for the objective of model analysis.

Kind regards,

Akaanksha Mishra

KPMG (Data Analytics Team)