Hello,

This is Umika from KPMG Data Analytics Team.

We have reviewed the data sets provided by Sprocket Central Pty Ltd. And during data quality analysis, we have come across few errors in the data sets.

We can give your organisation a holistic view of your data, enabling you to learn from and use it to make better business decisions, grow revenue, enhance operational capabilities, and manage enterprise risks and compliance mandates.

As we know that data quality analysis is the core of the process, and it is essential to clean the underlying data and mitigate these issues in order to improve data analysis and provide better solution to your company.

* Empty DOB cells can be either filled with the median value or the rows can be deleted as they account for just 87 out of 4000 rows.
* Gender column has no standard value, it is specified as F, M, Female, Male, U.
* Default column is of no use as it just contains random values.
* Assign a uniform last name to the customers such as ‘\_’ as there are about 125 null last names.
* Delete empty online orders considering them as fake orders as they account for just 360 out of 20000 orders.
* The product\_first\_sold\_date needs to be renamed as when we converted the column to datetime, it showed wrong values, so it does not relate to the data set.

The errors in the data sets are described as follows:

**Customer Demographic data set**

|  |  |
| --- | --- |
| **Column Name** | **Errors** |
| last\_name | 125 NULL VALUES |
| DOB | 87 NULL VALUES, A DOB WITH YEAR 1843 WHICH IS NOT POSSIBLE. |
| gender | No uniformity in gender specification as M, F, Male, Female, U. |
| Job\_title | 506 null values |
| Job\_industry\_category | 656 null values |
| default | 302 null values but it is a random column and has no importance. |
| tenure | 87 null values |

**Transaction Data Set**

|  |  |
| --- | --- |
| **Column\_name** | **Errors** |
| online\_order | 360 null values |
| brand | 197 null values |
| product\_line | 197 null values |
| product\_class | 197 null values |
| product\_size | 197 null values |
| standard\_cost | 197 null values |
| product\_first\_sold\_date | 197 null values, The column name says date but the data type of it is not datetime and the column name needs to be renamed. |

Regards,

KPMG DATA ANALYTICS TEAM