Hello,

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

The data quality analysis is one of the core phases and due to some issues in the data set, we suggest the following mitigations in order to improve the data quality.

- We can take a mode year value for the missing records of customersDOB.
- We can assign a uniform last name of customers, which values aremissing.
- Replace gender 'U' with reference to the customer name and make aconsistency.
- 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 ofdata.
- Eliminate the blank orders considering fakeorders.

The following are the details of quality issues encountered in the dataset.

Customer Demographic (Total records 4000)

Field Name	**Data Quality Issue**
DOB	DOB inaccurate and Age Missing
deceased_indicator	Not updated
Gender	Inconsistent (M, Male, F, Female, Femal, U)
job_title	Incomplete (several blanks)
job_industry	Incomplete (several blanks)
Default	Irrelavant column
Tenure	Incomplete (several blanks)

Transactions (Total records 20000)

Field Name	**Data Quality Issue**
Online_order	Incomplete (several blanks)
brand	Incomplete (several blanks)
order_status	Not updated(several cancelled entries present)
profit	Missing
standard_cost	Invalid format
List_price	Invalid format
product_first_sold_date	Invalid format

Customer Address (Total records 3999)

Field Name	**Data Quality Issue**
States	Inconsistent entries

New Customer List (Total records 1000)

Field Name	Data Quality Issue
Rank	Duplicate column
job_title	Incomplete (several blanks)

Below are more in depth descriptions of the data quality issues discovered and also the methods of mitigation are noted along with some recommendations. Following recommendations will help improve the overall data quality.

Customer Demographic

- Creating an *age column* will help with a better analysis.
- *Deceased list* needs to be updated to filter out deceased customers.
- *Gender* entries need to be consistent for easier processing.
- The *Default column* is irrelevant as it contains data not suitable for analysis.
- Several incomplete columns such as, *job_title, job_industry* and *Tenure* were present which have blank entries.

Transactions

- Columns having costs and prices were not formatted properly.
- *product_first_sold_date* is also improperly formatted.
- Adding a *profit column* will help with a better analysis.
- Several columns have blank entries.
- Some cancelled entries need to be filtered out from *the order_status* column.

Customer Address

- *States* entries need to be consistent for easier processing.

New Customer List

- The *Rank column* has a duplicate column which is linked with the other columns using a formula. So it can be renamed to Rank and the current rank column needs to deleted to avoid duplicate values.
- The *Job_Title* column has several blank entries.

That sums up all the data quality issues discovered through data quality analysis. The recommendations provided above are sure to improve the analysis output.

Please feel free to let us know if you have any queries regarding the issues presented.

Regards,
KPMG (Data Analytics Team)