Dear Sprocket Pvt Ltd,

Thank you for providing us with the dataset from Sprocket, we have reviewed the dataset extensively and summarized the following data quality issues with the dataset. We have further given our comments about how we have tackled those data quality issues and laid out a plan for the purpose of data cleaning.

Worksheet Name	**Data Quality Issue**
Transaction	Completeness and Relevancy
New Customer List	Completeness and Consistency
Customer Demographic	Completeness, Consistency and Relevancy

The above table outlines a few data quality issues with sprocket Central Pvt Ltd dataset. We have taken relevant steps to identify these issues and have given essential recommendations below to avoid these data quality issues from arising again.

1) **Worksheet Name – Transaction, here we identified blank values for columns “online_order” and “brand”. Also, the column for “product_first_sold_date” was converted correctly into date/time format. Few columns were not named and contained null values; such columns were deleted.**
    - We identified various blank values in the columns mentioned above, it is important to remove blank values from the dataset and raise the data quality issue for completeness and may lead to inaccurate results while modelling.
    - The column for “product_first_sold_date” was converted into date/time format which is easy to interpret, this problem may arise while exporting data from third party which may lead to conversion of date/time format to integer format, however they are not easy to interpret therefore, changing it into date/time format makes it easier to interpret data.
    - Some redundant columns were also present which were not names and serving only as a redundant column, dropping such columns make the analysis further easy.

2) Worksheet Name- New Customer List, here we identified blank values and inconsistent values for ‘Gender” column.

- As mentioned above, blank values were discovered in the sheet for the column “second_name” however it may be not a real issue as we may only use first name instead therefore it is not as important, however there were null values present. These were followed by more blank and null values in columns “job_title” and “Industry”.
- The column for "gender" which is a categorical variable has inconsistency, there are spelling errors for female, and some rows had abbreviations. This was changed to the columns being M for male and F for female. The column also consisted of an irrelevant variable "U" which

was discarded from the column. However, if more clarity could be provided on this it would be great or else for now it is irrelevant for the column.

3) Worksheet name- customer, demographic which has inconsistency for gender, there were missing values and irrelevant field called "default"

- This worksheet was dealt with in a similar way to the worksheet for new customer list, the field for gender was changed to M/F "U" which was an irrelevant value in the field was removed from the field.
- Null Values were removed from "job_title" and "job_industry"
- Irrelevant field called default was removed as it had no relationship to the data.

Moving forward, the team will continue the data cleaning and data transformation process for modelling. Questions will be raised along the way and assumptions will be documented separately. It would be great to spend some time with your data Team, to ensure all our assumptions are in line with the Sprocket Central Pty Ltd understanding.

With Regards,

Abhishek Tripathi

Data Consultant