Dear Sprocket Central Pty Ltd,

Thank you for making available the dataset from Sprocket. We have reviewed the attached dataset and summarised the following data quality issues. Further to this, we have provided some comments on how we have handled these issues and stipulated plans for subsequent data cleaning.

The table below shows a list of few data quality issues with the Sprocket central Pty Ltd dataset. We have taken relevant steps to identify these issues in the following worksheets and have given recommendations below to avoid subsequent data quality issues.

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

1. The “Transaction” worksheet
    a. The “online_order” and “Brands” columns have blank values. It is important to remove blank values from the dataset as this may lead to inaccurate results while modelling.
    b. The column for “product_first_sold_date” was converted into a date / time format which is easy to interpret. This problem may arise when exporting data from the third party which may convert date value to integer. However, they are not easy to interpret therefore converting it to date / time format makes it easier to interpret.

2. The “New Customer List” worksheet.
    a. Blank values were discovered under the column named “last_name” however this may not be significant since the first name is available in the dataset. Furthermore, there were still blank values in “job _title” and “industry” columns.
    b. The column named “gender” which is a categorical variable has some inconsistencies. There are spelling errors for “Female” while many other rows used the abbreviation. This has been changed to “F” for female and “M” for male. The column also has an irrelevant variable “U” which has been removed from the column. For now it is irrelevant for the column, however, we would appreciate it if more clarity could be provided on this.

3. The “Customer demographic worksheet
    a. The “gender” column has been treated in the same manner as that in the previous worksheet.
    b. b. The columns named “job_title” and “job_industry” had Null Values were removed from the dataset.
    c. The column named “default” was removed as it had no relationship to the data and was considered as irrelevant.

Majority of the data quality issues identified above if ignored would result in an inaccurate and faulty result while carrying out data modelling. Carrying out a data quality check ensures that we harmonise our assumptions with that of Sprocket Central Pty Ltd. In the absence of further details, we would proceed with data cleaning and data transformation process for modelling. We would document further questions and assumptions while working on during this course of this work.

Kind Regards

Tochukwu Collins
Junior Data Consultant