Dear Sprocket Central Pty, Ltd,

Thank you for providing us with the dataset from Sprocket, we have reviewed the dataset and summarised the following data quality issue with the dataset. We have further given our commands about how we have tackled these data quality issue and laid out a plan to move forward with the data cleaning.

Worksheet name	Data Quality issue
Transaction	Completeness and Relevancy
New Customer List	Completeness and Consistency
Customer Demographic	Completeness , Consistency and Relevancy

The table above, outlines a few data quality issues with Sprocket central Pty ltd dataset. We have taken relevant steps to identify these issues and given recommendation below to avoid these data quality issue from arising again.

1. Worksheet name - Transaction : Where we identified blank values from columns "Online order" , "Brand" and "Product line". The column for "Product_first_Sold_date" was converted into a date/time format.

a. We identified various blank values in the columns mentioned above, it is important to remove blank value from the data set as the raise the data quality issue for completeness and may lead to inaccurate results while modelling.

b. The column for "Product_first_sold_data" was converted into a date/time format which is easy to interpret, this problem may arise when exporting data from third-party which may convert date value to integer, however they are not easy to interpret therefore changing it to data/time format makes it easier to interpret data.

2. Worksheet name – New Customer List : Where we identified blank values from columns "Last name" , "Job title" and "Jon industry",there was also inconsistent values for gender.

a. As mentioned above blank values were discovered in the sheet for the column "Second_name" however it is not an issue as we may only use first name instead therefore it is not as important, however there were still blank values. There were followed by more blank and null values in columes "Jon_title" and "Job_industry".

b. 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 Male and Female. The column also consist 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"

a. This worksheet was default 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.

b. The Null values were removed from "Job_title","Job_industry", "Last_name" and "Customer_ID".

c. 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 SME, to ensure all our assumptions are in line with the Sprocket Central pty Ltd understanding.

Kind Regards

Lavanya

Junior Data Consultant.