Dear Sprocket Central Pty Ltd,

Thanks a lot for providing us with the three datasets from Sprocket. We have analyzed the datasets and summarized the data quality issues with the datasets below. We have further given our comments about how we have tackled these data quality issues and laid out a plan to move forward with the data cleaning.

Worksheet Name	**Data Quality Issues**
Transaction	Completeness & Relevancy
NewCustomerList	Completeness, Consistency & Relevancy
CustomerDemographic	Completeness, Consistency & Relevancy
CustomerAddress	Consistency

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

1. **Worksheet name "Transactions" where we identified blank values for columns 'online_order' and 'brand'. The column 'product_first_sold_date' was converted into the date/time format and column 'list_price' was converted into the currency format.**
    a. We identified various blank values in the columns 'online_order' and 'brand'. It is important to remove blank values from the datasets so as to raise the data quality issues for the completeness and may lead to inaccurate results while modelling.
    b. The columns 'product_first_sold_date' and 'list_price' was converted to date/time format and currency format respectively which is easy to interpret. These problems may arise when exporting data from third party, however they are not easy to interpret therefore, changing the format makes it easier to interpret data.

2. **Worksheet name "NewCustomerList" where we identified blank values, inconsistent values and relevancy issues.**
    a. In this sheet, there were some blank values observed in column 'last_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 columns 'job_title' and 'job_industry_category'.
    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 being Male for male and Female for female. The column also consisted 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.

   c. The columns ‘postcode’, ‘property_valuation’ and ‘past_3_years_bike_related_purchases’ were converted from text format into the integers. There problems may come up when exporting data from third party, however they are not easy to interpret therefore, changing the format makes it easier to interpret data.

3. **Worksheet name “CustomerDemographic” has inconsistency for ‘gender’, there were missing values and irrelevant field called ‘default’.**
   a. This worksheet was dealt with in a similar way to the worksheet for new customer list, the field for gender was changed to Male/Female. “U” which was an irrelevant value in the field was removed from the field.
   b. Null values were removed from ‘last_name’, ‘job_title’ and ‘job_industry_category’.
   c. Irrelevant field called ‘default’ was removed as it had no relationship to the data.

4. **Worksheet name “CustomerAddress” has inconsistency in column ‘state’.**
   a. The column ‘state’ has some inconsistency in it. There were some rows that were not abbreviated. The data was changed to NSW for New South Wales and VIC for Victoria. This made the data more understandable by whom so ever would have a look at it.

Moving forward, the team will continue the data cleaning and data transformation process for modelling. Questions will be raised along the was 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

Naman Arora

Junior Data Consultant