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 comments 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 & Relevancy |
| New Customer List | Completeness & Consistency |
| Customer Demographic | Completeness, Consistency & Relevancy |
| Customer Address | Completeness & Relevancy |

The table above, outlines a few data quality issues and we have identified them and given recommendations below to avoid these issues from arising again.

**1. Worksheet Name – Transaction we have identified blank values for columns namely “online\_order” and “brand”. We have converted the column “product\_first\_sold\_date” into a date format.**

1. We have identified the blanks in the mentioned columns above and removed them. It is important to remove blank values from dataset as the it arises the issue on data Completeness and may lead to inaccurate results while modelling.
2. The column “product\_first\_sold\_date” is converted to date format as this problem may arise when exporting the data from the third party which may convert date value to integer.

2. **Worksheet Name – New Customer List, we have identified blank values, and inconsistent values for gender.**

1. As mentioned above, blanks values were discovered in the column “last\_name” however it is not an important issue as we may only use the first\_name, but still they were blank and hence been removed. Also, there were blanks in columns “job\_title”, “job\_industry”.
2. The column “gender” which is a categorical variable has inconsistency, there were spelling errors for female, some rows even had abbreviations. This was changed to the columns being M for Male and F for Female. The column also consisted an irrelevant variable “U” which was discarded from the column. However, if more clarity would be provided on this it would be great else for now it’s irrelevant.

**3. Worksheet Name – Customer Demographic, we have identified inconsistent values for gender, and also there were missing and irrelevant field called “default”.**

1. This gender issue was similar to the New customer List worksheet and we removed “U” as it was irrelevant.
2. Null values were removed from columns “job\_title”, job\_industry”.
3. Irrelevant field name default was removed as it had no relationship to the dataset.

**4. Worksheet Name – Customer Address, we have identified abbreviations and names for State column.**

1. The column “state” had names as well as their abbreviations. So, we made “New South Wales” as “NSW” and “Victoria” as “VIC” for state column.

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 time with your data SME, to ensure all our assumptions are inline with the Sprocket central Pty Ltd understanding.

Kind Regards,

Shreya J Gandhi

Junior Data Consultant.