Dear Sprocket Central Pvt Ltd,

Thank you for providing the dataset, we have reviewed the dataset completely and we have summarized the following Data Quality issue with the Dataset. We have further given our comments about how we are going to tackle the quality issue.

|  |  |
| --- | --- |
| Worksheet Name | Data Quality Issues |
| Transactions | Completeness and Relevance |
| New Customer List | Completeness and Consistency |
| Customer Demographics | Completeness, Consistency and relevancy |
| Customer Address | Consistency |

We’ve addressed the above data quality issues and took the 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 Online order, order status, brand, product lines, product class, product size, standard cost columns with **null values** and Product\_first\_sold column converted to Date format, Removed all the Null value column.

a. We identified various blank value in the columns mentioned above, it is important to remove the blank values from the dataset as it will solve the completeness issue, or it may lead to inaccurate results while modelling.

b. The column Product\_first\_sold\_date had an integer value integer value instead of date value, which is not relevant in this case, however it is changed to Date format so that it is easy to interpret.

2. **Worksheet name: New Customer List** where we identified **null values** for last name, dob, job\_title, job\_industry columns and there were also inconsistent values for Gender.

The null value in the last\_name can be ignored in this case but there were significant number of null values in the dob, job\_title, job\_industry columns where in this case we have removed for better completeness. However, there was the field “U” in the Gender column which was irrelevant value, so it has been discarded from the column. However, if more clarity is provided it will be great or for now the gender “U” has been removed.

Replaced NSW, QLD, VIC to New South Wales, Queensland, Victoria in state column, which had consistency issue.

Removed all unnamed columns, null columns, rank column, value column where we find it to be irrelevant for the analysis purpose.

3. **Worksheet : Customer Demographic**

a. The null value in the last\_name can be ignored in this case but there were significant number of blank values in the dob, job\_title, job\_industry, tenure columns where in this case we have removed for better completeness.

b. was the field “U” in the Gender column which was irrelevant value, so it has been discarded from the column. However, if more clarity is provided it will be great or for now the gender “U” has been removed. The field was changed to Male/Female.

c. The column DOB had an integer value integer value instead of date value, which is not relevant in this case, however it is changed to Date format so that it is easy to interpret, removed irrelevant DOB value which has an year 1843 which is removed, Removed all null values in Dob.

d. Replaced null by N/A in Job title column.

e. Removed deceased customers.

f. Irrelevant column default, columns with null values has been removed since it had no relationship with the data.

4. **Worksheet name : Customer Address**

a. Replaced NSW, QLD, VIC to New South Wales, Queensland, Victoria in state column, which had consistency issue.

b. columns with null values have been removed since it had no relationship with the data.

Moving Forward our team will continue Data Cleaning, 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 our data SME, to ensure all assumptions are in line with the Sprocket Central Pvt Ltd understanding.  
  
Regards,  
Abhishek Naik

Junior Data consultant.