**Dear Sprocket Pty Ltd,**

Thank you for providing us with the dataset from Sprocket, we have reviewed the dataset and summarized the following data quality issue with the dataset. 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 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 recommendations below to avoid these data quality issues from arising again.

1. **Worksheet name “Transaction” where we identified blank values for columns “online\_order” and “brand”. The columns for “product\_first\_sold\_date” was converted into a date/time format.**
2. We identified various blank values in the columns mentioned above, it is important to remove blank values from the dataset as it is the reason for the data quality issue for completeness and may lead to inaccurate results while modelling.
3. 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 a data from third-party which may convert date value to integer, however they are not easy to interpret. Therefore, changing it to date/time format makes it easier to interpret data.
4. **Worksheet name “New Customer List “ where we identified blank values, there was also inconsistent values for gender.**
5. 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 names instead therefore it is not as important, however there were still blank values. These were followed by more blank and null values in columns “job\_title” and “industry”.
6. 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 column being M for male and F for female. The column also consisted and irrelevant variable “U” which was discarded from the column. However, if more clarity could be provided on it, it would be great or it would be irrelevant for the column as of now.
7. **Worksheet name “Customer Demographic” which has inconsistency for gender, there were missing values and irrelevant field called “default”**
8. This worksheet was dealt with in a similar way as the worksheet for “New Customer List”, the field of gender was changed to M/F. “U” which was irrelevant value in the field was removed from the field.
9. Null Values were removed from “job\_title” and “job\_industry”.
10. 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.

Thanking You,

Your’s faithfully,

Mohammad Nehal

Jr. Data Consultant