**Dear Sprocket Central Pvt. Ltd.**

Thankyou for providing us 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 |
| Customer Demographic | Completeness, Consistency and Relevancy |
| Customer Address | Accuracy, Completeness and Consistency |

The table above addresses a few data quality issues with Sprocket central Pvt. Ltd. Dataset. We have taken some necessary steps to identify these issues and have given our recommendations below to avoid these data quality issues from coming again.

1. First worksheet named transaction where we identified blank values for variables “online\_order” and “brand”. The column “product\_first\_sold\_date” was initially integer value which was converted into date/time format.
   1. We identified various blank values in the columns mentioned above, it is important to remove blank values from the data because it may give misleading results while modelling and generating insights.
   2. The columns “product\_first\_sold\_date” was converted into date/time format which is more interpretable, and can be used in a better way while modelling and data analysis.
2. Second worksheet named customer demographic where we identified some inconsistency for gender, also there were missing values and there was an irrelevant field called “default”.
   1. This worksheet was dealt in a similar way as of transaction. The gender column was converted with values as M and F and u was dropped because it was unknown.
   2. Null values were removed from “job\_title” and “job\_industry”.
   3. Irrelevant field called default was removed as it contained an unreadable language and was irrelevant.
3. Third, worksheet named customer address where we identified some accuracy problem with data field “state” and field named “country” contained only one value so that is irrelevant for the further analysis.
   1. The “state” contained New South Wales and Victoria states with both short form and entire name so they were replaced with one common name.
   2. The field “country” was removed from analysis because it is of no use.

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 more time with the data SME, to ensure all our assumptions are in line with the Sprocket Central Pvt. Ltd. Understanding.

Regards

Prabhat Mishra