**Dear Sprocket Central Pty Ltd,**

Thank you for providing us with Sprocket Central dataset, we have reviewed the dataset and summarised the data quality issue. We have carefully addressed these issues. We have given our views about how we have tackled these data quality issue and laid out a plan to move forward the data cleaning process.

|  |  |
| --- | --- |
| **Worksheet name** | **Data quality issue** |
| **Transactions** | relevancy and completeness |
| **New Customer List** | consistency and completeness |
| **Customer Demographic** | consistency, completeness & relevancy |

In the table, we measure some of the data quality problems in the Sprocket Central dataset, we identify and implement appropriate measures to address these problems, and provide data to prevent data quality problems from recurring and improve the accuracy of the underlying data used to drive business decisions.

1. Additional **‘Customer ids’** is not present in the **‘new customer list’.**

This means that the received data may not match up, which can affect the outcome of the analysis if there are any missing data records.

1. Worksheet name ‘**Transactions’** where we find the null values for columns name – “online\_order” & “brand”. The “product\_first\_sold\_date” column was converted into date & time format.
   1. We recognize various blank values in the columns mentioned above, it is important to remove blank values from the dataset and raise the data quality issue for completeness and may lead to inaccurate results while modelling.
   2. The column ‘product\_first\_sold\_date’ converted into date & time format, which is easy to explain, the data value is in **“int”** form this problem may arise while exporting the data from any database, however we changing it to date & time format to make easier to explain.
2. Worksheet name ‘**New Customer List’** where we find null values, there was also inconsistent value for **‘gender’** & We make a new **‘Rank’** with the column being **“U”** because it connects to the **‘value’** column.
   1. We recognize blank values in this worksheet also for column name – **“job\_title”** & **“job\_industry\_category”**. There’re some null values in the **“last\_name”** column there’s no issue with that because we only need first name.
   2. The column **“gender”** has inconsistency, there’re spelling error for female, and some rows had abbreviations. We had changed to the gender column for being male for “M” and female for “F” & the column also contain the irrelevant value “U” we discarded from the column. However, if more clarity provided on this it would be grateful else for now it is irrelevant for the column.
   3. There’re are some hidden columns after the **P** column, the rank column is not assigned to any particular column, because the selected column **"U"** had a rank formula where the reference **"W"** represents the **"Value"** column, we changed the rank column. to the reference column **“W”.**
3. Worksheet name **“CustomerDemographic”** which has inconsistency for gender & DOB, there were missing values and irrelevant field called “default”.
   1. The null values present in **‘job title’** and **‘job industry’,** we removed null values.
   2. Irrelevant field called “default” was removed as it had no relationship with the dataset.
   3. DOB had one inconsistent value – 1843, please provide more accurate data with this dob section
   4. This worksheet also has similar **‘gender’** inconsistency, we changed the field of gender to M/F, there’s also an irrelevant value **“U”** which was removed successfully.

Moving forward, as part of the model analysis activities, the team will be continuing with data cleansing, standardisation and transformation. During the journey, questions will arise and assumptions will be documented. When we've finished, it would be great to have a little time with your Data SME and make sure that all assumptions are in line with what Sprocket Central understands.

Kind Regards,

Swapnil Srivastava