

Dear Sprocket Central Pty Ltd,

Thank you for providing us with the dataset from Sprocket, we have reviewed the dataset and 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.

	Accuracy	Completeness	Consistency	Currency	Relevancy	Validity
Customer Demographic	DOB: Inaccurate. Age: Missing.	Job Title: Blanks	Gender: Inconsistent	Deceased Customer: Filtered out	Default Column: Delete	
Customer Address			States: Inconsistent			
Transactions	Profit: Missing	Online Orders: Blanks. Brands: Blanks.			Cancelled Status Order: Filtered Out	List Price: Format Product Sold Date: Format

The table outlines a few data quality issues with the central Pty Ltd dataset. We have taken relevant steps to identify these issues and given recommendations to avoid these data quality issues from arising again.

1. Worksheet name transaction where we identified blank values for columns "online_order", "brand", "product_line", "product_class" and "product_size". The column for "product_first_sold_date" was converted into a date/time format. "list_price", and "standard_cost" was converted into a Currency format.
 - a. We identified various blank values in the columns mentioned above, it is essential to remove blank values from them as they raise the data quality issue for completeness and may lead to inaccurate results while modelling.
 - b. The column for "product_first_sold_date" is converted into a date/time format which is easy to interpret, this problem may arise when exporting data from a third —party which may convert the date value to an integer, however, they are not easy to interpret therefore changing it to date/time format makes it easier to interpret data. Also the same case in "list_price" and "standard_cost" was converted into a Currency format.
2. Worksheet Name New Customer List where we identified blank values, there were also inconsistent values for gender.
 - a. As mentioned above, blank values were discovered in the sheet for the column "last_name". However, it is not an issue as we may only use the first name instead, therefore it is as important, however, there were still blank values. These were followed by more blank and null values in columns "job_title" and "job_industry_category".
 - b. The column for "gender" which is a categorical variable has inconsistency, there are spelling errors for females, and some rows had abbreviations. This was changed to the columns being M for male and F for female. The column also consisted of an irrelevant variable "U" which was discarded from the column. However, if more clarity could be provided on this it would be great, or else for now it is irrelevant for the column.

3. Worksheet name customer demographic which has inconsistency for gender, there were missing values and an irrelevant field called "default"
 - a. This worksheet was dealt with in a similar way to the worksheet for the new customer list, the field for gender was changed to M / F. "U" which was an irrelevant value in the field was removed from the field.
 - b. Null Values were removed from "last_name", "job _ title" and "job _ industry_category"
 - c. 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. It would be great to spend some time with your data SME, to ensure all our assumptions are in line with the Sprocket Central Pty Ltd understanding.

Kind Regards

Rejoice Raju

Junior Data Analyst