

Dear client,

First of all, thank you for providing us with the three datasets from Sprocket Central Pty Ltd. We did some data quality assessment on these datasets and found some issues in the data. The details on the same are provided below.

Transaction data:

- Contains 20000 records and 13 fields.
- Contains no duplicated data.
- The following 7 fields contain null/missing values.
 - online_order, brand, product_line, product_class, product_size, standard_cost, product_first_sold_date
- missing profit column.
- product_first_sold_date not in correct format/datatype.
- There are more customer ids present in transaction data compared to customer demographic data.

Customer demographic data:

- Contains 4000 records and 13 fields.
- Contains no duplicated data.
- The following 6 fields contain null/missing values.
 - last_name, DOB, job_title, job_industry_category, default, tenure
- doesn't contain age column.
- values in gender column is not consistent also contains unknown gender 'U'
- No relevancy in the values of default column.

Customer address:

- Contains 3999 records and 6 fields.
- Contains no duplicated data.
- Contains no null/missing values.
- Values in state column is not consistent
- There are more customer ids present in customer address data compared to customer demographic data.

Recommendation: If the rows with null values are much less compared to the entire rows, then we can just remove the rows with null values or else if the data is important then we can use imputation methods to add data to those missing cells like using mode or mean.

After identifying the issues in the datasets, we will further move on to the data cleaning stage. Questions will be asked along the way.

Regards,

Mansi Patel