Hello,

At first, I want to thank Sprocket Central Pty Ltd for giving us an opportunity to analyse Your data.

This is Himanshu Pandey from KPMG Data Analytics (Virtual Internship) team. Me and my team have thoroughly gone through the data sets provided by you, and we have come with few conclusions, based on the discrepancies and irregularities in data.

And due to these errors in the data set, we suggest the following changes in order to improve the data quality, which will help us in understanding better customer acquisition and changing market needs.

The following are the details of error encountered in the data set.

Transactions Sheet: 20000 Records

FIELD NAME	ERRORS
Online_order	360 records are Blanks
brand	197 records are Blanks
product_line	197 records are Blanks
product_class	197 records are Blanks
product_size	197 records are Blanks
standard_cost	197 records are Blanks
product_first_sold_date	197 records are Blanks

New Customer List: 1000 Records

FIELD NAME	ERRORS
Last_name	29 records are Blanks
gender	Apart from Female and Male, gender U is given, which might mean Unspecified or Unknown,.
DOB	17 records are Blanks
Job_title	106 records are Blanks
Job_indusrty_category	165 records are Blanks

Customer demographic: 4000 Records

FIELD NAME	ERRORS
DOB	87 records are Blanks One of the record has DOB in year 1843, The age of this person seems an error
last_name	125 records are Blanks
Gender	88 records gender 'U' Values are not consistence M, Male, F, Female, Femal, U
job_title	506 records are Blanks
job_industry	656 records mention 'N/A'
Default	There are records with 'special characters' , and also 302 blanks
Tenure	87 records are Blanks

Customer Address: 3999 Records

FIELD NAME	ERRORS
State	Some data is in abbreviation form and some in complete format. New south wales , NSW are given separately, Similar is the case with Victoria and VIC This may create a problem in understanding the data set for some.

Customer Address dataset does not have any blank values.

Suggestions to improve dataset:

1. Replace gender 'U' with reference to the customer name and make a consistency.
2. For tenure values, we can take a mean of rest of the values and assign the mean value to the missing fields in order to maintain the consistency of data.
3. Remove all blanks, as they give a wrong estimation about number of people.
4. Eliminate the blank orders considering fake orders.
5. Change the product_first_sold_date into date format
6. Wherever the money is being mentioned, we can correctly mention the currency associated with it.
7. Avoid Full name while also using the abbreviations of some data, for e.g. New south wales and NSW etc.
8. All N/A values should be removed, as it is similar to blanks.

If we keep in mind these suggestions, and bring some changes to the dataset, I think the market would be a lot easier to understand and so will be New customer acquisition.In the end, again I want to thank Sprocket Central Pty L for giving me and my team this opportunity to analyze and understand the data.

Thanks and Regards,
Himanshu Pandey
KPMG (Data Analytics Team)