Hi There,

I am Disha Gupta. Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. As per the first module, data quality assessment of the three datasets was supposed to be done. Below are the observations and solutions to them.

For the dataset Customer Demographics:

1. In the gender column, there are rows with value ‘’U” which is not defined what U stands for. So instead of U, unknown can be used
2. Again, in the gender column, there are values given as F, Female and Femal which specifies as ‘’Female” and M and Male for ‘’Male’’. Only one type of value should be used. Male for men and Female for women
3. The default column has got garbage values, so the column can be removed
4. Columns DOB and Tenure have got some Null values. As they are less compared to the data entirely, they can be removed or else can be filled by some default values

For the dataset Customer Demographics:

1. There are dates in DOB which need to have special details like ones in the years 1938 and 1941
2. The null values need to be checked with the customer itself if possible, in columns like job\_title and industry

For the dataset Transactions:

1. There is large amount of null values in the columns like product line, product class, product size which can be filled using the product id
2. The online order column that consists of null values, can be filled depending on other related record of data

For the dataset Customer address:

1. There are some duplicate records in the address as the same address cannot have different post code. Thus, depending on the property value may be the duplicate value must be removed
2. There are values like VIC and Victoria , NSW and New South Wales which represent the same place yet different notations are given

Common for all datasets , datatypes need to be proper for all the columns.

Regards,

Disha Gupta