Hello Sprocket Central Pty Ltd,

I am writing this after analyzing the datasets you provided us. The following are data quality issues that I found in your datasets as well as recommendations:

* Transaction Data Dataset:
  + In this dataset I encountered a lot of missing values as some rows had no inputs. The columns affected and the number of null inputs, in brackets, include: online\_order(360), brand(197), product\_line(197), product\_class(197), product\_size(197), standard\_cost(197).
  + Also, the column ‘product\_first\_sold\_date’ contains data that are incorrect that is they are in number forms instead of date form as the name of the column indicates.
* CustomerDemographic Dataset:
  + There were also missing values in this dataset in the following columns: last\_name(3725), DOB(87), job\_title(506), job\_industry\_category(656), default(302), tenure(87)
  + In the ‘gender’ column, the value ‘U’ is seen in some rows and it seems contradictory on the conventional notations of gender. It’s usage or meaning should therefore be explained.
  + Additionally, in the ‘gender’ column, values such as ‘Femal’, ‘F’, ‘M’ are observed which gives a wide range of values. An improvement to this would be to use either ‘Female’ or ‘F’ to indicated feminine and not both. This also applies for the male gender use either ‘Male’ or ‘M’ and not both for consistency purposes. Double checking during data entry should also help avoid incorrect entries.
  + In the ‘DOB’ column, I observed a value ‘1843-12-21’ which is completely inaccurate as there are values for that customer in the ‘past\_3\_years\_bike\_related\_purchases’ column.
  + The ‘default’ column also contains metadata in varying formats and thus is not useful as is.
  + In the ‘job\_industry\_category’ column, the value ‘n/a’ is not clear. It could be understood in various ways such as the particular customer is unemployed, they are in the informal job sector as most categories listed are in the formal sector or their job category is not listed in your system. An improvement to this would be replacing ‘n/a’ with values such as ‘unemployed’, ‘self-employed’.
* CustomerAddress Dataset:
  + In the ‘property\_valuation’ column the metrics for measurement of the valuation have not been defined. Indication where it is thousands, millions and in which currency would greatly improve the data quality.
  + In the ‘customer\_id’ column, some customer ids such ids 3, 10 and 23 are missing thus their address information is also missing.

Some strategies that your organization could utilize to mitigate the above issues include:

* First would be to ensure that for any new customer all the required information is collected and entered into the system to avoid having missing data. In cases where there are missing fields for existing customers, these can be filled on customers’ next interaction with your organization. Requests can also be sent out to the customers to fill in the missing information.
* The various field should also be formatted in ways the restrict inputs to only the required formats. For instance, where columns with numbers should only contain numbers and no characters. Same case applies for date columns. The gender column can be restricted to contain specific values for either gender to ensure consistency. Columns can also be restricted to contain values in a certain range such as in the ‘DOB’ column can be restricted to only 150 years back from current year or any reasonable number of years guided by factors such as life expectancy in the country.
* Metadata entered into the datasets should be cross checked to ascertain that the data is relevant to the column, dataset or even to the organization. This ensures that data collected is correct, relevant and useful to your organization.
* Finally naming of datasets and their columns should be simple and as clear as possible. For instance in the CustomerDemographic dataset, the column ‘tenure’ is not clear on it’s meaning and may be ambiguous or sound useless to someone who is not involved in the day to day running of your organization’s activities. Better naming of the column would help the data analysis team to understand the data better.

With the above issues addressed and strategies implemented the quality of your organization’s data is bound to get better and better.

Kind Regards,

KPMG.