To: Sprocket Central Pty LTD

Subject: Regarding the quality of the dataset.

Hello Team,

Thanks for providing the three datasets:

Customer Demographic, Customer Address and Transactions.

Listing the common problems:

1. There are more empty columns which are not a part of the dataset. It is with due consideration that these columns are to be deleted before processing as it might lead to more memory consumption and can accidentally be included in visualizations.
2. Since customer demographic is the primary dataset used, there is a complete disparity in the customer\_id values for all the three datasets. Inconsistent amount of customer ids and different ids are present. Please refer to the attached excel sheet. Correct customer id details should be provided by going through the details.
3. Lot of missing values are present in all the three datasets. For some fields, the missing percentage is about 2 while for others it is more. These small portion of missing values can be treated by deleting or if they are important and if the missing proportion is used, these values need to be imputed accordingly. Please refer to the following columns: customers’ surname, job title, tenure column, online\_order, brand, product\_line, product\_class, product\_size, standard\_cost and product\_first\_sold\_date.
4. Inconsistent data types are specified and the values that are used are erroneous not holding any vital information.

* The DOB field is not created in a correct format. The data type has to be Date however it is created as “Any” data type, also unacceptable range of DOBs are given and this might lead to the prediction of wrong target audience. This can be easily solved by changing the data-type of that column to date and providing correct data.
* The default column is erroneous as it does not hold any meaningful information. It is good to drop this off.
* The product\_first\_sold\_date as the field name suggests must have a date data type. However the values are wrong and no dates were present but whole number data. This should be treated.
* Since the entire dataset is specific to the country Australia, the column country in this dataset can be removed.
* In the customer Address, the column state has many different values: New South Wales is expressed as it is and as NSW, Victoria as itself and VIC. These values should be expressed in a standard form as each different form creates a category of its own and leading to mismatch/ wrong prediction.