08/06/2021

Dear Kathleen,

Thank you for providing us the information(dataset) from Sprocket Central Pty Ltd. Analysing the dataset from the three datasets received, we can observe some important points below.

Table name	Num of records	Num distinct customer id	Date Data Received
Customer Demographic	4000	4000	07/06/2021
Customer Address	3999	3999	07/06/2021
Transaction Data	3999	3494	07/06/2021

Some data quality issues that were found in the dataset are described following standard data quality, observing accuracy, completeness, consistency, validity and uniqueness:

- **Some customer_ids** in the Transaction Data table and Customer Address table **are not present in Customer Demographic**, where it has the identity of the customer as full name and date of birthday. To analyze the information about the customer it will be used only customer that is present(synchronized) in the Customer Demographic table (master table). Here some examples of non-synchronized between the tables.

  Customer_id 5034 is present only in Transaction Data table

  Customer_ids 4001,4002,4003 are only present in Customer Address

- **Some records are incomplete**, as job_industry_category(16.5%) , job_title(12.7%), last_name(3.2%) and DOB(2.2%). For records that totaling less than 1%, they have been removed from the training dataset.
- **Inconsistent values for the same attribute**. For example: Gender: Male/M, Female/F/Femal, U and State: New South Wales/NSW, Victoria/VIC. It recommends that create a list of valid values to avoid inconsistency. Before use the dataset it will be necessary to fix the inconsistent values.
- **Column/Field default** on the Customer Demographic table **have special characters**, with no clear information. It will be ignored in the training dataset.
- **Column/Field product_first_sold_date** on the Transaction Data **table does not have Date format.** It will be casting in Date format to be use in the training dataset.

Moving forward, the team will continue with the data cleaning, standardization and transformation process for the purpose of model analysis. Questions will be raised along the way and assumptions documented. After we have completed this, it would be great to spend some time with your data SME to ensure that all assumptions are aligned with Sprocket Central's understanding.

Kind regards,

Cynthyá Belloni