Hi,

As per the instructions, I examined the data and have carried out a data quality check for Sprocket Central data. I have observed the following issues with the data quality with regards to the parameters of Accuracy, Completeness, Consistency, currency, validity, uniqueness and Relevancy (also mentioning how wecan deal with the same). I also removed some data and merged all the relevant columns to the transaction datasheet so I can carry out my model building using it .:

- **The records in variables like 'last name', 'DOB', 'job_title', 'job_industry_category' and 'Tenure' are missing**
  - ➢ Please note the full name, including the last name, when collecting the data as 'last name' was missing in almost 3.125% of data in the Customer Demographics table.
    Solution: The variable 'last name' is not majorly creating an issue in data analysis; therefore, we are not dropping the data set with missing 'last name' records.
  - ➢ Variable 'DOB' in the Customer Demographic table is also missing with around 2.175%, which is less than 5%.
    Solution: Since 'DOB' is an important variable, therefore it would be better if you could provide all the missing DOB. If not, then for the analysis purpose, we will drop the data point with the missing DOB. In some cases we can use it to find the Age while conducting the analysis .
  - ➢ Columns "job_title" and "job_industry_category" in Customer Demographics table have sizable numbers of null values (12.65% & 16.5% respectively).
    Solution: Since these are categorical columns, therefore our approach will be to define another category to replace these.
  - ➢ Column 'Tenure' data in the Customer Demographics table is also missing with around 2.175%, which is the same as the 'DOB' missing values percentage. And it is found that the same data points are missing for 'DOB' and 'Tenure'.
    Solution: We can drop the records, since missing records are < 5%.
- **Inconsistency in the values and data type**
  - ➢ Column 'gender' in the Customer Demographics table has been defined in different ways. For example, Female is defined in 3 different ways - Female, F and Femal. Similarly, male is defined in two ways- Male and M.

Solution: For analysis, data need to be consistent, and for the same reason, I have defined the gender as Female and Male.

Recommendation: I would recommend you to please create a drop-down list for such variable to avoid inconsistency.

- All the variables should have consistency in terms of type of data (integer, string, date time, etc.) present in the variable.

  Solution: For the analysis purpose I have modified the data type of the variables as required.

  Recommendation: I would suggest that take the input in the required data type for example if the column is showing the variable related to date then column data type should be in date time format and not in numeric form.

- **Accuracy needs to be maintained while filling the data.**

  Valid 'DOB' should be given: for customer id 34, 'DOB' mentioned is 1843-12-21, which is practically impossible (Customer Demographics table). For now, I have removed the data point customer id 34.

I hope this would help in improving the data quality in future. Looking forward to hearing from you soon regarding the problem with the data quality in the provided dataset. I will also attach an excel with the data cleaned and filtered by me .

Warm Regards,

Daniella Brito