Hi,

Thank you for providing the dataset of Sprocket Central Pvt Lmt company. The first task before getting into analysis is to check the quality of the data that has been provided. The table below points out the key quality issues within the dataset.

|  | Transactions | CustomerDemographic | CustomerAddress | NewCustomerList |
| --- | --- | --- | --- | --- |
| Completeness | Columns like online\_order,brand,product\_line,product\_class, product\_size,standard\_cost,product,first\_sold\_date have empty values | Columns like DOB,job\_title have empty values  Job\_industry column has “n/a” values | Values in customer\_id is missing | Columns like DOB, job\_title, job\_industry\_category have empty cells. |
| Accuracy | Profit column is missing | Inaccurate values in column DOB.  Column age is missing |  | No primary key to join with other tables.  age column is missing |
| Validity | product\_first\_sold\_date:  Values not in date format.  list\_price: not in currency format. |  |  |  |
| Consistency |  | Gender column with inconsistent values | Column state has inconsistent values |  |
| Relevancy |  | Column default is not relevant for analysis |  |  |

These are quality issues that we came across and have some suggestions to mitigate these issues.

**Completeness**

Updating the data in each table is one of the ways to fill the empty cells. Also there are some customer\_id missing in the table “CustomerAddress”. This needs to be fixed by adding the missing customer and their details to the table.

**Accuracy**

There is an entry in the CustomerDemographic table with the DOB "1843-12-21", which is not possible. It needs to be checked and updated.

Profit and age columns are missing in “Transaction” and “CustomerDemographic” tables. Inorder to mitigate this issue, it is possible to add columns “profit” (list\_price-standard\_cost) and the current age of the customer to the respective tables.

**Validity**

The format of the columns “product\_first\_sold\_date” and “list\_price” in the transaction table are not appropriate.

**Consistency**

Values in column gender in CustomerDemographic table and “state” in CustomerAddress table are not consistent to each other.This can be avoided in future if a drop down option is available while entering the customer details.

**Relevancy**

The "default" column in the CustomerDemographic table is not necessary for analysis; if it's for knowing the customer review or suggestion, then predefined templates can be provided, which will be time-consuming for customers.

These are the suggestions and recommendations to rectify the quality issue that was found during the first stage of analysis, which will lead to a positive impact on the output of the analysis.

If any query or suggestion regarding the quality issues pointed out above feel free to contact us.

Regards

Sunooja