Respected Manager,

I want to start by saying thank you for allowing me the chance to clean the data sets.

Following is a detailed examination of the four datasets provided, a list of data-quality problems found during cleaning, and suggestions for future improvements.

The given 4 datasets are:

1. Transactions(Data 1)
2. New Customer List (Data 2)
3. Customer Demographic (Data 3)
4. Customer Address (Data 4)

The Data Quality Assessment has been performed on the basis of the Data Quality Dimensions Framework mentioned as follows:

1. Accuracy
2. Completeness
3. Consistency
4. Timeliness
5. Relevancy
6. Uniqueness
7. Validity

Accuracy

- Dataset 2: The 'DOB' column contains false information.
- The faulty data were filtered as a form of mitigation.
- It is advised to use conditional formatting so that incorrect values are automatically prevented.

Completeness

- Dataset 1: Some records for the first sales date, brand, product line, product class, and online order are missing.
- Dataset 2: There are blank values in the columns "last_name," "DOB," "job_title," and "job_industry_category."
- Dataset 3: There are blank entries in the columns "last_name," "DOB," "job_title," "job_industry_category," and "tenure."
- Mitigation: In all three datasets, the missing values were filled using the fillna() function and the reverse and forward filling methods.
- It is advised that conditional formatting or a drop-down function be used to automatically prevent null values.

Consistency

- Dataset 1: The formats of the columns 'list price' and 'standard_costs' were inconsistent.
- Two decimal places were added, and consistency was achieved.
- Dataset 3: The values in the 'gender' column are inconsistent. Its values were highly variable.
- Dataset 4: The values in the column 'states' were inconsistent.

- Mitigation: The incorrect format was used to replace the inconsistent values in datasets 1, 3, and 4.
- It is advised to use a categorical data type to prevent problems with inconsistent data. Dropdown menus reduce discrepancies and human error in manual data entry by various staff members and enhance the reading and interpretation of the data.

Timeliness

- Two clients were identified as deceased in Timeliness Dataset 3 (TD3).
- Mitigation: I removed any deceased customers from the list.
- The data should be updated frequently to prevent inaccuracies of this nature

Relevancy

- 'order_status', a field in Dataset 1, showed cancelled orders.Thus, they were eliminated.
- Dataset 2: Some 'Unnamed' hidden columns were present in between the pertinent columns and may have been deceptive. These columns were so removed.
- Dataset 3: The column 'default' contains corrupt data.
- Mitigation: I removed the rows containing cancelled orders as well as the "default" column.
- It is advised to either delete or reformat corrupted material.

Uniqueness

- In the uniqueness domain, I discovered no problems.

Validity

- First Dataset: Dates were presented in integer format in the "product_first_sold_date" column. I changed the type of data and made the necessary corrections.
- Reduction: I changed the integer format into a conventional one.
- It is advised that the values' data types be standardised to prevent inconsistencies.

Therefore, these were the data quality problems that we found in the provided data during the data cleaning and analysis procedure.

Regarding the aforementioned subjects, I would be delighted to talk about any feedback or inquiries. I would be open to any dialogue to make sure that the assumptions made are understood by Sprocket Central Ltd.

Thanking You

Regards,
Ujjwal Srivastava
KPMG