Respected Manager,

First of all, I would like to thank you for giving me this opportunity to clean the data sets.

Listed below are an in-depth analysis of the given 4 datasets, the data-quality issues discovered while cleaning along with recommendations on how to improve the quality of data in the future.

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 has inaccurate data.

**Mitigation** : Those inaccurate data were filtered.

**Recommendation** : Some conditional formatting can be done so that inaccurate values can be automatically prevented.

* Completeness

Dataset 1: Some records of the online order, brand, product line, product class, product size, standard cost and first sold date are missing.

Dataset 2 : The columns ‘last\_name’, ‘DOB’,’job\_title’, ‘job\_industry\_category’ has missing values.

Dataset 3 : The columns ‘last\_name’, ‘DOB’,’job\_title’, ‘job\_industry\_category’ and ‘tenure’ has missing values.

**Mitigation** : The missing values were filled using fillna() function and backward and forward filling method in all the 3 datasets.

**Recommendation** : Some conditional formatting or a drop down function can be implemented so that null values can be automatically prevented.

* Consistency

Dataset 1: Columns ‘list price’ and ‘standard\_costs’ had inconsistent formats.

The formats were changed to 2 decimal places and made consistent.

Dataset 3 : The column ‘gender’ has inconsistent values.It had many variations in the values.

Dataset 4 : The column ‘states’ had inconsistent values.

**Mitigation** : The inconsistent values in dataset 1, 3 and 4 were replaced with the correct format.

**Recommendation** : A categorical data type can be implemented to avoid the issues of inconsistent data. Dropdown options minimise inconsistencies and human error in manual entries by different personnel and improves the data interpretability and readability.

* Timeliness

Dataset 3 : 2 customers were listed as deceased.

**Mitigation** : I filtered the customers listed as deceased.

**Recommendation** : The data needs to be regularly updated to avoid such kind of errors.

* Relevancy

Dataset 1 : The column ‘ order\_status’ displayed cancelled orders.So, they were filtered out.

Dataset 2 : Some hidden columns named as ‘Unnamed’ were present in between the relevant columns which could be misleading. So, those columns were dropped.

Dataset 3 : There is a column named ‘default’ which consists of corrupt data.

**Mitigation** : I dropped the ‘default’ column and the rows with cancelled orders.

**Recommendation** : Corrupted data can be either removed or reformatted.

* Uniqueness

I found no issues in the Uniqueness domain.

* Validity

Dataset 1 : The column ‘product\_first\_sold\_date’ had dates in integer format.So, I converted the type of data and corrected it.

**Mitigation** : I converted the integer format into a standard format.

**Recommendation** : The data type of the values needs to be standardised in order to avoid any descrepancies.

So, these were the data quality issues present in the provided data which we discovered in during the data cleaning and analysis process.

I would be grateful to discuss any comments or questions regarding the above issues. I would be happy for any discussion to ensure that all assupmtions applied align with Sprocket Central Ltd’s understanding.

Thanking You

Regards,

Anannya Manojawas

KPMG