Dear Correspondent @ Sprocket Central Pty Ltd,

I’ve carefully analyzed your datasets and have identified several data quality issues as described below for each dataset table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Transaction | New Customer List | Customer Demographic | Customer Address |
| Accuracy |  |  | i. Inaccurate value in ‘deceased\_indicator’ |  |
| Completeness | i. Missing values in ‘online\_order’, ‘product\_class, ‘product\_size’, ‘brand’ and ‘product\_line’ | i. Missing values in ‘last\_name’, ‘DOB’ and ‘job\_title’  ii. n/a values in ‘job\_industry\_category’ | i. Missing values in ‘job\_title’ and ‘DOB’  ii. n/a values in ‘job\_industry\_category’ | i. Missing values in ‘customer\_id’ |
| Consistency | i. No currency in ‘list\_price’  ii. Some values without currency in ‘standard\_cost’ |  |  |  |
| Validity | i. Invalid ‘customer\_id’ value |  |  | i. Invalid ‘customer\_id ’ values |
| Uniqueness |  |  | i. ‘gender’ has redundant values |  |
| Relevant |  | i. Four unnamed columns |  |  |
| Currency | i. ‘product\_first\_sold\_date’ values in non-standard format | i. ‘DOB’ values in non-standard format |  |  |

1. **Transaction Table**
   1. ‘customer\_id’ column has a value ‘5034’ which is invalid as it’s nonexistent in the customer tables.
      1. *This means that the value of this variable is mistaken and can be hinted that there is other replacable missing value in the data.*
   2. ‘online\_order’, ‘product\_class’, ‘brand’, ‘product\_size’, and ‘product\_line’ columns are incomplete due to missing values.
      1. *This denotes the total elimination of missing values as it just forms approx. 5% of the dataset and hence approved for cleaning.*
   3. ‘standard\_cost’ column have some values without any currency and ‘list\_price’ column have values without any currency and hence these variables are inconsistent.
      1. *This can introduce some mistakes in all sorts of aggregation and hence it’s suggested to homogenize it.*
   4. ‘product\_first\_sold\_date’ have date values not in standard format.
      1. *All the data analysis software have a standard format which avoids them to miss-represent a date-time variable.*
2. **New Customer List Table**
   1. ‘last\_name’, ‘DOB’ and ‘job\_title’ columns are incomplete due to missing values. ‘job\_industry\_category’ column has n/a values and hence it’s incomplete as well.
      1. *Here the missing values constitute less than 5% of the dataset and hence it’s approved for cleaning.*
   2. ‘DOB’ column has dates in non-standard format.
   3. There are four unnamed columns between ‘propery\_valuation’ and ‘rank’ columns having decimal values, the relevancy is questionable.
      1. *These columns were shrinked between the named columns and these variables had some influence over the target ‘rank’ and with ‘property\_valuation’ as a variable as well. This was determined by the pearson correlation coefficient being quite significantly positive.*
3. **Customer Demographic Table**
   1. ‘gender’ column has three values for Female, i.e. ‘F’, ‘Femal’ and ‘Female’ and two values for Male, i.e. ‘M’ and ‘Male’ which adds redundancy.
      1. *This suggests homogenizing this variable.*
   2. ‘job\_title’ and ‘DOB’ columns have missing values, while ‘job\_industry\_category’ column has n/a values and hence these columns are incomplete.
   3. There is a customer with DOB year as ‘1843’ and with deceased\_indicator as ‘N’ which signifies an error and hence it’s inaccurate.
      1. *This indicates that there has been a mistake entering the date value rather than in describing as deceased.*
4. **Customer Address Table**
   1. Some values in the ‘customer\_id’ column are missing i.e. 4, 10, 22 and 23 whereas there are some new values in the column which are nonexistent in other customer tables i.e. 4001, 4002 and 4003 which describes incompleteness and invalidity.
      1. *Suggests that the customer actually having the values 4, 10, 22 and 23 may by miss-inputted as 4001, 4002, and 4003 and one possibility of duplication as well.*
   2. ‘state’ column has two values indicating New South Wales and two values indicating Victoria and hence there’s redundancy.

Regards,

Yash Sinojia

Virtual Intern @ KPMG