Dear Client,

We have reviewed the data sets and we discovered some data quality issues during the analysis.The summary table below highlights key quality issues :

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Accuracy | Completeness | Consistency | Currency | Relevancy | Validity | Uniqueness |
| Customer Demographic | “Default”: values are not correct | Missing data:   * last\_name * DOB * job\_title * job\_industry\_category * tenure | Inconsistence Data:  Gender | Archive or filter out *deceased* Customers | incomprehensible information in the « default »  column | / | No duplicated values |
| Customer Address | / | No missing data | Inconsistence Data:  States | / | / | / | No duplicated values |
| Transactions | / | Missing Data:   * online\_order * brand * product\_line * product\_class * product\_size * standard\_cost * product\_first\_sold\_date   Incomplete Data:   * Customer id(no transaction for customers between 3501-5033) |  | / | “order\_status” is irrelevant for analysis. | Data Type:   * Product Sold Date:   Float>Date | No duplicated values |
| New Customer List | 5 unamed Columns | Mssing/incomplete Data:  Customer id is missing |  | / | / | / | No duplicated values |

You could find below the detailed descriptions of data quality issues with some recommendations and explanations to avoid data quality issues:

**Transactions DataSet :**

* 20000 records, 13 columns
* The 'product\_first\_sold\_date' is stored as floats,we should change it to date
* The missing values : (since total percentage of all missing values is less than 0.05,you could simply remove these missing data)

| **Column** | **numberOfMissingData** | **Percentage** |
| --- | --- | --- |
| **online\_order** | 360 | 0.01800 |
| **brand** | 197 | 0.00985 |
| **product\_line** | 197 | 0.00985 |
| **product\_class** | 197 | 0.00985 |
| **product\_size** | 197 | 0.00985 |
| **standard\_cost** | 197 | 0.00985 |
| **product\_first\_sold\_date** | 197 | 0.00985 |

* The records that have a standard\_cos t>=1665.04 are outliers values,
* nb of these outlier :189 (0.009%)
* No duplicated Values

**New Customer List Dataset :**

* 1000 records and 23 attributes
* There are 5 unamed Columns
* The missing values :

|  |  |  |
| --- | --- | --- |
| **Column** | **numberOfMissingData** | Note |
| last\_name | 29 | Replace missing with uniform last name or “undefined” |
| DOB | 17 | Impute missing values with Backfill&frontfill /Machine learning methods |
| job\_title | 106 | Impute missing values with Machine learning methods |
| job\_industry\_category | 165 | Impute missing values with Machine learning methods(knn,…) |

* No duplicated Values
* For the gender columns we shold change “U” to “Unspecified”

**Customer Demographic Dataset**

* 4000 rows × 13 columns
* Values in the "default" column are not correct (special characters, null and Blanks)
* The Gender column has inconsistent values, we need to replace inconsistent values with appropriate values

(M, Male, F, Female, Femal, U 🡪 should replace by🡪 Male ,Female ,Unspecified )

* The missing values:

|  |  |
| --- | --- |
| **Column** | **numberOfMissingData** |
| last\_name | 125(Replace missing with uniform last name or “undefined”) |
| DOB | 87 |
| job\_title | 506 |
| job\_industry\_category | 656 |
| tenure | 87(replace missing with the mean) |

**Customer Adress Datasets**

* 3999 rows × 6 columns
* No missing data
* there are 3999 records ,in the customer demographic dataset there are 4000 for customers (id:4003,4002,4001 we have their adress but we don't have their demographic information ) for customer (id :3 we don't have his adress ) since we have these problems only with these 4 customers, an inner merge of these 2 dataset will be the right chose

Now we need to megre all these datasets:

* The new cutomers list have 1000 cutomers (without cutomer\_id)
* The customer demographic dataset has 4000 cutomers(these are the old cutomers)
* The demographic information for all new customers in the newcustomerlist dataset are missing
* we can't do the demographic study with the new customers(and also we can't merge it since the customer-id is missing in the new customer\_list)

solution:

* we can numerate(id) new customers from 4001 to 5001 and then merge the three dataframes

Note:

* since there is no information in Transaction dataset about customer between 3500 and 5034 , we don't have to use this solution and at the moment we should merge only Customer demographic dataset(Adress/demographic) with Transcation dataset

Regards,

KPMG (Data Analytics Team)