Hi,

Thank you for providing the datasets. I have conducted a data quality assessment on the datasets and following issues were identified. I have provided some strategies for data cleaning and to improve the data quality. I have also attached a clean dataset for your reference.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset/Data Quality Dimension** | **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** | **Uniqueness** |
| Customer Demographic | 01) DOB: some of the customer data is inaccurate | 01) last\_Name: 125 missing data  02) DOB: 87 missing data  03) job\_title: 506 missing data  04) tenure: 87 missing data  05)Customer\_id: incomplete | 01) gender: data is inconsistent | 01)deceased\_indicator: unnecessary data about deceased customer | 01) Default: data is not relevant to the data set. | No issue found | No issue found |
| CustomerAddress | No issue found | 01)Customer\_id: incomplete | 01) state: data is inconsistent | No issue found | No issue found | No issue found | No issue found |
| Transaction | No issue found | 01) online\_order = 360 data missing  02) brand = 179 data missing  03) product\_line= 179 data missing  04) product\_class= 179 data missing  05) product\_size= 179 data missing  06)standard\_cost= 179 data missing  07)product\_first\_sold\_date= 179 data missing  08) 01)Customer\_id: incomplete | No issue found | No issue found | 01) order\_status: cancelled order data is not required or irrelevant | 01)list\_price: data is not in valid format (not shown as currency)  02) product\_first\_sold\_date: data is not in valid format. | No issue found |

**Mitigation Strategy:**

**Dataset: CustomerDemographic**

**Issue/ Strategy:**

**Accuracy:** some of the data is inaccurate like DOB of customer\_id 34 is 21/12/1843. Similar data can be filtered out.

**Completeness:** Missing data (blank cells) can be deleted. Customer\_id is not complete across spreadsheets. Only customer\_id from 1 to 3500 will be used in the testing dataset.

**Consistency:** Data about gender need to be changed to make them consistent (e.g. change all male to ‘Male’ or ‘M’)

**Currency:** Deceased data can be deleted from the dataset as these data is not required.

**Relevancy:** Default column data is not relevant to the dataset; hence this column can be deleted.

**Dataset: CustomerAddress**

**Issue/ Strategy:**

**Consistency:** Data about state need to be changed to make them consistent (e.g., change all Victoria to ‘VIC’).

**Dataset: Transaction**

**Issue/ Strategy:**

**Completeness:** Missing data (blank cells) can be deleted.

**Relevancy:** Data with ‘cancelled’, order\_status can be deleted as this is not required for analysis.

**Validity:** Data in list\_price and product\_first\_order\_date columns are not in valid format. Format of these data can be changed to valid one.

**Recommendation:**

* Adding an age column in CustomerDemographic data set and profit column in Transaction data set will be helpful for analysis and error detection purpose.
* Data type is not consistent across spreadsheets. In some places data type is char where same data is saved in numeric value. Convert recodes to numeric in ‘NewCustomerList’ dataset to make it consistent with other dataset.
* Constraint on data type will ensure consistency.
* Drop down menu for selecting state, gender etc will help reduce error in dataset.

Thank you again for the opportunity to work with you.

Best Regards,

Neelakshi Bhattacharya