Dear Team,

Thank you for providing the three datasets and 1 dataset containing new customer information from Sprocket Central Pty Ltd. Below is the summary statistics of the 3 dataset that I have received. Kindly let me know if there is any discrepancy in my understanding.

|  |  |
| --- | --- |
|  | **TotalNoOfRecords** |
| **Transaction** | 20000 |
| **CustomerDemographics** | 4000 |
| **CustomerAddress** | 3999 |

Below is the description of the dataset individually

**Transaction**

* online\_order, brand, product\_line, product\_class, product\_size, standard\_cost, product\_first\_sold\_date contains null values.
* Customer\_id present in Transaction table are not present in Customer\_Demographics table.
* Inconsistent data types for some columns such as transaction\_date

**New Customer List**

* 5 columns doesn’t have header
* Last\_nem, job\_title have null values.

**CustomerDemographic**

* last\_name, job\_title, tenure have null values
* gender has spelling mistakes in Male and Female
* DOB has “1843-12-21”, 177 years which is impossible
* Job\_title, job\_industry\_category contains n/a values
* default column contains some absurd values and ASCII characters.
* Inconsistent data types for some columns such as DOB

**CustomerAddress**

* Inconsistent values for some columns such as “Victoria: is written as “Vic” in state.

**Strategies**:

* I would replace the null and NA values by using mode of the observations.
* Replace the year 1843 by 1943, probably there was typing error.
* Remove the unnamed columns because no meta-data description about these columns are given.
* Drop the default column because it is irrelevant.