Respected Sir/Mam,

I had gone through the dataset of Sprocet central . I had suggested some changes needed to be done in the dataset . And also required some imformation related to dataset . All the required information related data cleaning as below :

**CUSTOMER DEMOGRAPHICS TABLE :**

* **Customer\_id :** Should be unique.
* **Last Name :** There are blank values in last name column . W e can replace null or blank values with 'NaN' Or '-' or can keep same ''.
* **Gender :** ( Femal,Female and F ) Three type of values for Female category. (Male,M) two type values for Male category.(U) value in Gender column meaning is not described. We can replace 'Femal' and 'F' with 'Female' . We can replace 'M' with 'MALE'. IF 'U' is for unknown we can keep as it is in column . [**Constriants :** ('Male' or 'Female' or 'U')].
* **DOB** : There are dates which are from 1931 till 2002 .There is date of year 1843 customer which makes customer 180 year old which is not possible. And there are also null values in this column . We can either delete outlier date record i.e. 1843 . And we can replace null values with '-'.
* **Deceased Indicator :** There are 2 customers who are dead that cannot be used to do marketing. So we can remove this all column for analysis after deleting above two records as all our customers will be alive.
* **Default** : This column contain all values that are gibberish which doesnot have any meaning. We can delete entire column as it is not of any use for analysis.

**Columns to keep : '**customer\_id', 'first\_name', 'last\_name', 'gender', 'past\_3\_years\_bike\_related\_purchases', 'DOB', 'job\_title', 'job\_industry\_category' ,'wealth\_segment'.

**CUSTOMER ADDRESS TABLE :**

* **Customer\_id :** Should be unique.Extra 3 customers detail as compared to customer demographic table.
* **property\_valuation :** Please mention meaning of property valuation in this table and unit to measure it.

**Columns to keep :** ALL

**NEW CUSTOMERS LIST TABLE :**

* **Last Name :** There are blank values in last name column . W e can replace null or blank values with 'NaN' Or '-' or can keep same ''.
* **Gender : [Constriants :** ('Male' or 'Female' or 'U')].
* **DOB** : There are null values in DOB column. We can replace null values with '-'.
* **job\_title :** There are null values in job\_title column.We can replace null values with 'n/a'.
* **tenure** : Please specify the meaning of tenure column in this table and unit to measure .
* **property\_valuation :** Please mention meaning of property valuation in this table and unit to measure it. Some values are in integer format and some values are in float format we neet to keep values in same datatype.
* **Rank :** Please specify on which basis rank is given in this table.
* **Value :** Please specify the meaning of value in this table.

**Columns to keep :** ALL

**TRANSACTION TABLE :**

* **transaction\_id :** This column should have unque constraint.
* **product\_id :** Please specify meaning of value 0 in product\_id is ti relevent or not. Not null constraint.
* **customer\_id :** Not null constraint.
* **online\_order :** online\_order column should not be null either it should be 'True' or 'False'.
* **brand :** There are null values in brand column we can replace them with 'Unknown'.
* **product\_line :** Please specify meaning of this column and it also has blank values that we need to replace it.
* **product\_class** : There are some blanks in product class .There should be only 3 categories according to data i.e. high, medium, low.
* **product\_first\_sold\_date :** This column is in number datatype , it should be in date datatype.

**Columns to keep :** ALL

Thank you.

Akhtar Shabbir Attar

KPMG