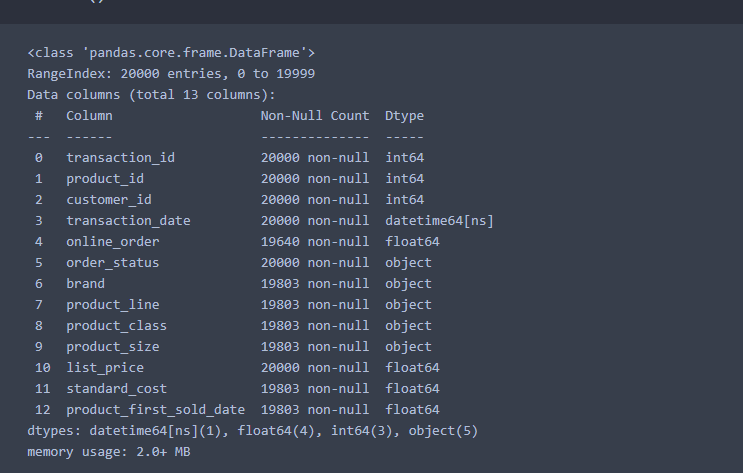
Dear Sir

I have reviewed the dataset provided by you and I found certain loopholes in it. If we are able to correct them, then we will certainly achieve good results as accurate data analysis is the outcome of analyzing accurate data.

I’ll mention below the dataset glitches according to the datasets itself:

1. **TRANSACTIONS**

* The last column i.e. ‘product\_first\_sold\_date’ does not hold a valid date, instead contains a number.
* The field ‘list\_price’ does not contain any currency value and without currency value it is just another number.
* There are many extra columns in the dataset which should ideally be removed as they contain Null values.
* In the given fields there are many NaN or Null values which are not suitable to make an ideal dataset. Hence, those NaN /Null values should be substituted with some other value. We can also consider taking Mode of the dataset and fill up those values.
* Here is the statistics of the incomplete data. In total we have 20,000 entries.



* We have one field called ‘transaction\_date’. Along with this we can also store the corresponding day of the week. Using the day, we can draw wonderful insights from the data and figure out on which days we have maximum and minimum sales.

1. **NEW CUSTOMER LIST**

* In the field, ‘product\_first\_sold\_date’, the format of date is incorrect.
* There are some fields in this dataset which are ambiguous or they can be clearly explained before introducing the dataset. Those fields are:

1. wealth\_segment
2. deceased\_indicator
3. property\_valuation
4. rank
5. value
6. **CUSTOMER DEMOGRAPHIC**

* Some people have middle name too, so we can create one extra column for writing middle name instead of just putting the first name and middle name in the same field.
* There is one column called ‘default’ and it is ambiguous and of no use. It can be removed from the dataset.
* In the gender field, some genders are represented by ‘F’ or ‘M’ while some are represented by ‘Female’ or ‘Male’. We should maintain consistency in data by using only one kind of representation.
* The format of ‘DOB’ field can be changed to ‘DD-MM-YYYY’ because it the simple to read and understand.
* There are some fields in this dataset which are ambiguous or they can be clearly explained before introducing the dataset. Those fields are:

1. wealth\_segment
2. deceased\_indicator
3. **CUSTOMER ADDRESS**

* In the ‘state’ field, some entries are represented in abbreviated form such as ‘VIC’ or ‘QLD’ while the name of some states are given in full-form such as ‘New South Wales’. In order to make the dataset idea, we should stick to only one kind of data.
* There are some fields in this dataset which are ambiguous or they can be clearly explained before introducing the dataset. Those fields are:

1. property\_valuation