Good morning Sprocket Central Pty Ltd,

thank you for your three data set, I checked them and summaried here.

|  |  |  |
| --- | --- | --- |
| Table name | unique Customer\_ID | number of reocrd |
| CustomerDemographic | 4000 | 4000 |
| CustomerAddress | 4000 | 4000 |
| Transaction | 3494 | 20000 |

In the CustomerDemographic table

1. There are 125/4000 rows missing last name, since the name has no impact on the analysis of user sales, it is directly ignored here or filled with any value.
2. In the gender column, there are multiple expressions for male, female and unknown, so we change some of them to make a uniform expression.
3. In the birthday column, there is an outlier, based on experience, it should be 1943 rather than 1843. In addition, there are 87/4000 missing birth date values.
4. 506/4000 rows are missing job title,
5. 656/4000 rows are missing job industry,
6. 87/4000 rows are missing tenure.

In the CustomerAddress table

1. the state expressed in different ways, I just unified up the forms of expression
2. In the country column, all users are from Australia, so this row can be deleted

In the Transaction table

1. online order, brand, product line, product class, product size, have blank cells
2. list price were not in currency format, then I changed that to currency
3. there are several value missed in standard cost and some of them accurated to many decimal places, but we only need two decimal, so I changed that
4. the column product\_first\_sold\_date has meaningless value, according to the feature, I changed that to date, which makes the data in that column much clearer

In the NewCustomerList table:

1. Some property\_valuation are accurate to two decimal places, most retain integers, so only integers are retained uniformly
2. Value accuracy is not uniform, so all are retained to two decimal places only
3. job\_title and DOB miss some value
4. There is a hidden column R.S.T.U.V. through checking its function, seems that this is to add a coefficient based on customers’ conditions and a random number. for now it will not influent on the analyze, but maybe in the future it works, so I just keep them while hide them

I added an age column in CustomerDemographic and NewCustomerList, calculated with the formula “=int((Now()-DOB)/365)”, also added a profit column in Transactions, calculated with the formula “=list\_price-standard\_cost”. In order to make the analyze easier.

For the missing value, the analysis will be more accurate if you can provide more complete data, otherwise, I will decide to delete the data or complete the data based on experience and other relevant databases according to the analysis process.

If you have additional explanations for this, please feel free to contact us, but for now, I'll remove the row change for the rest of the analysis.