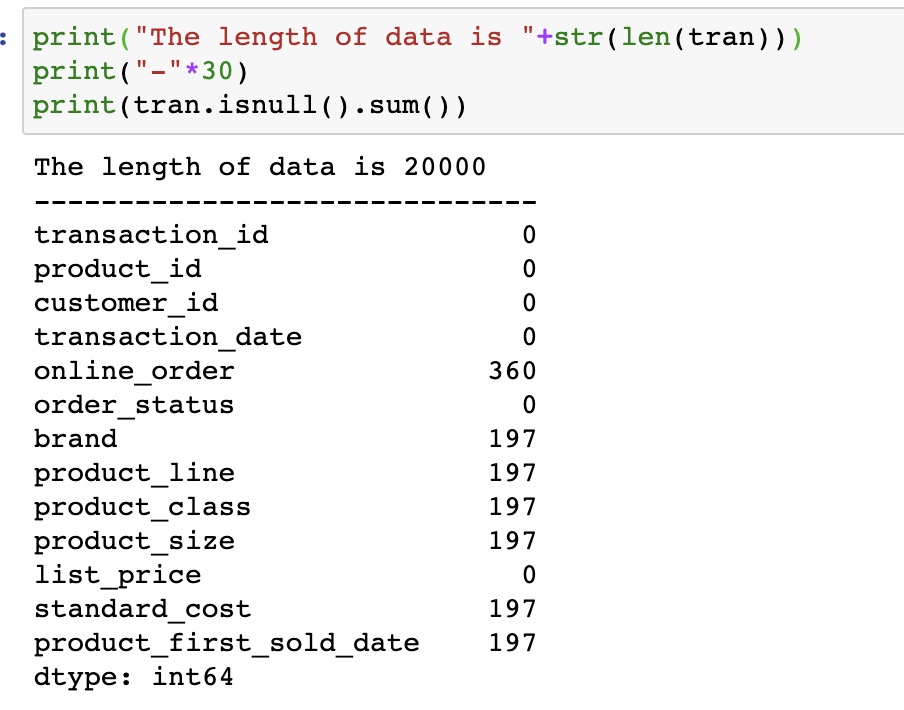
# Data Quality Assessment

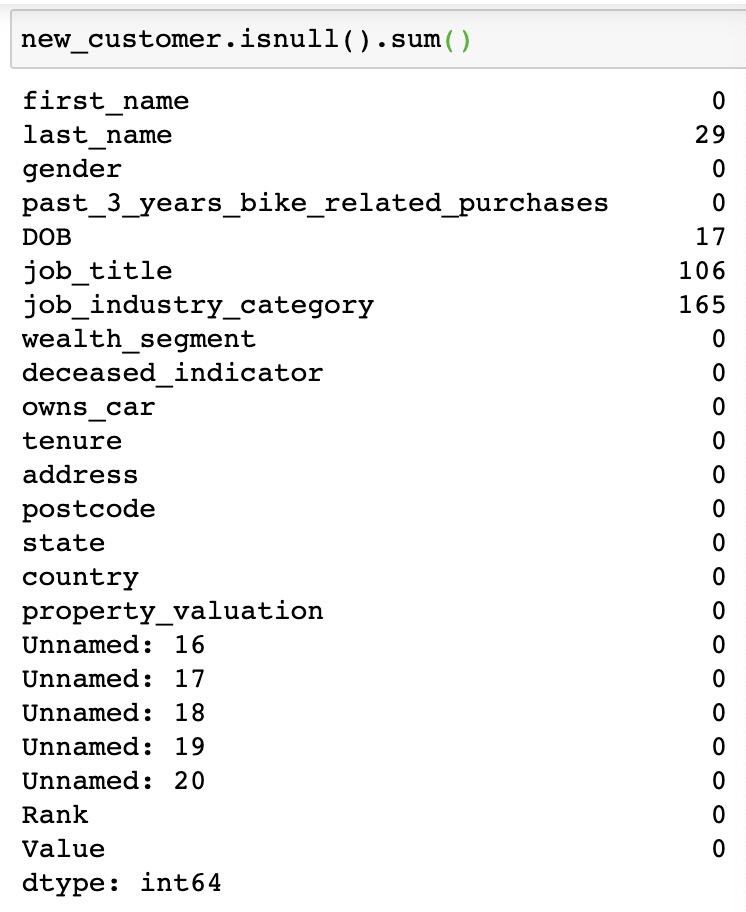
The data has four tables: transactions, newcustomerlist, customerdemographic and customeraddress. The quality of the data set is evaluated by the following dimensions: 1. Integrity; 2. Accuracy; 3. Consistency

## Integrity

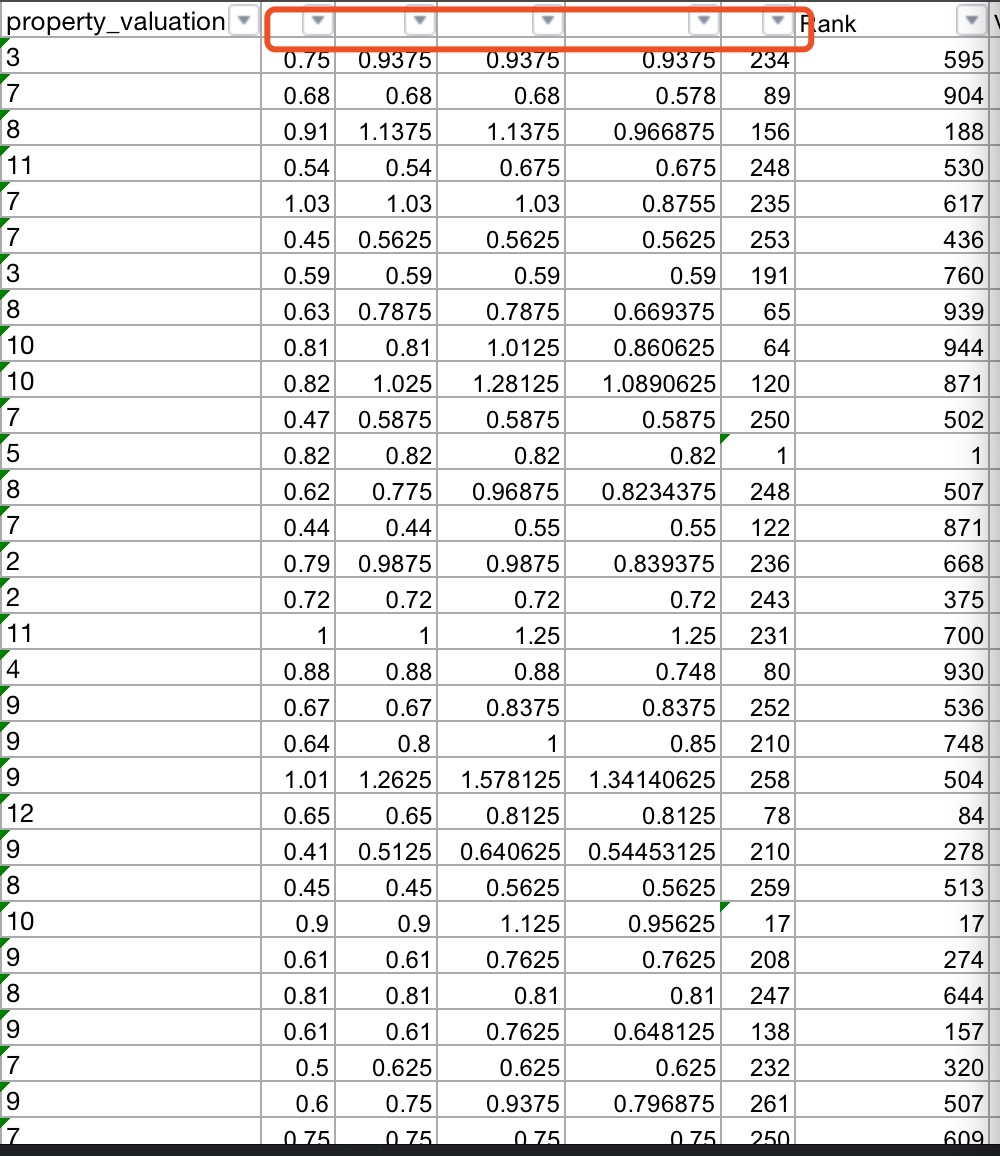
For the transactions table, the table has 20000 rows, online\_ There are 360 orders missing and 170 other trading commodity information missing.



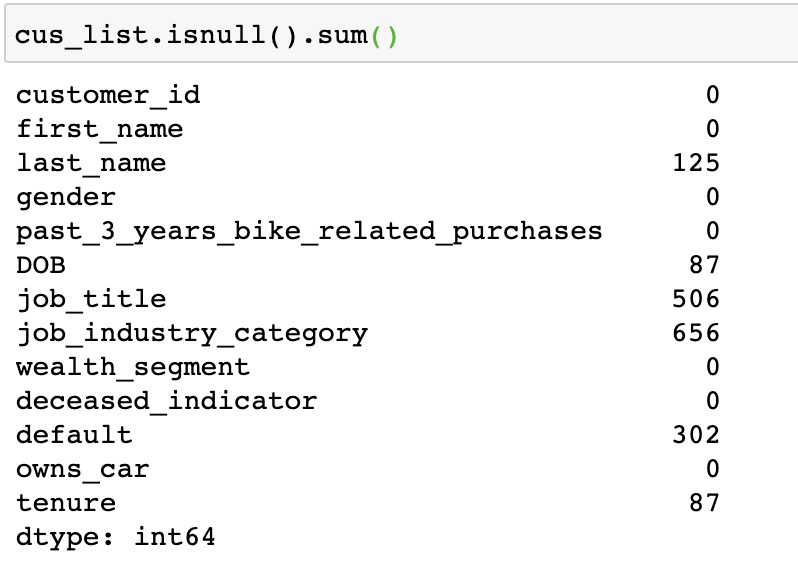
For the NewCustomerList table, there are 1000 rows in the table, and the number of missing information in each field is as follows:



In addition, there are four missing field names in the NewCustomerList table:



For the CustomerDemographic table, the table has a total of 4000 rows, and the specific missing number of each field is as follows:



## Consistency

1. By comparing the customer address table with the customer demographic table, we can find that CD table has 4000 customer identification IDS, while CA table has 3999 customer identification IDs;

In the customeraddress table, there are 4001, 4002 and 4003 customers, and their information is missing in the customerdemographic table; in the customerdemographic table, there are customers 3, 10, 22, 23, and their information is missing in the customer address table.

2. In the state field of the customeraddress table, new south walls and NSW express the same meaning, while Victoria and Vic express the same meaning;

3. In the gender field in the customerdemographic table, F, female and female express the same meaning, while m and male express the same meaning;

## Accuracy

1. There is an abnormal value of "1843-12-21" in DOB field of CustomerDemographic table;

2. There is a representation of "U" in the gender field in the CustomerDemographic table, but this gender expression is not clear