

As shown in the above table, there are 8 attributes in our dataset.

Distance\_from\_home: the distance from home where the transaction happened.

Distance\_from\_last\_transaction: the distance from last transaction happened.

Ratio\_to\_median\_purchase\_price: ratio of purchase price transaction to median.

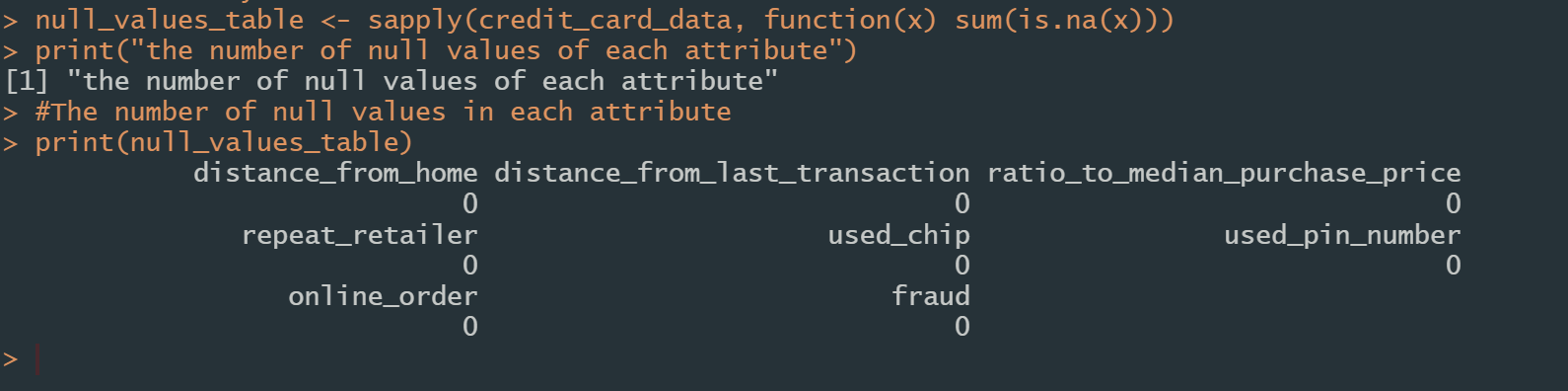
Repeat\_retailer: the transaction happened from the same retailer before.

Used\_chip: the transaction used a chip to transact (credit card).

Used\_pin\_number: the transaction happened by using PIN number.

Online\_order: the transaction is an online order.

Fraud: the transaction is fraudulent.



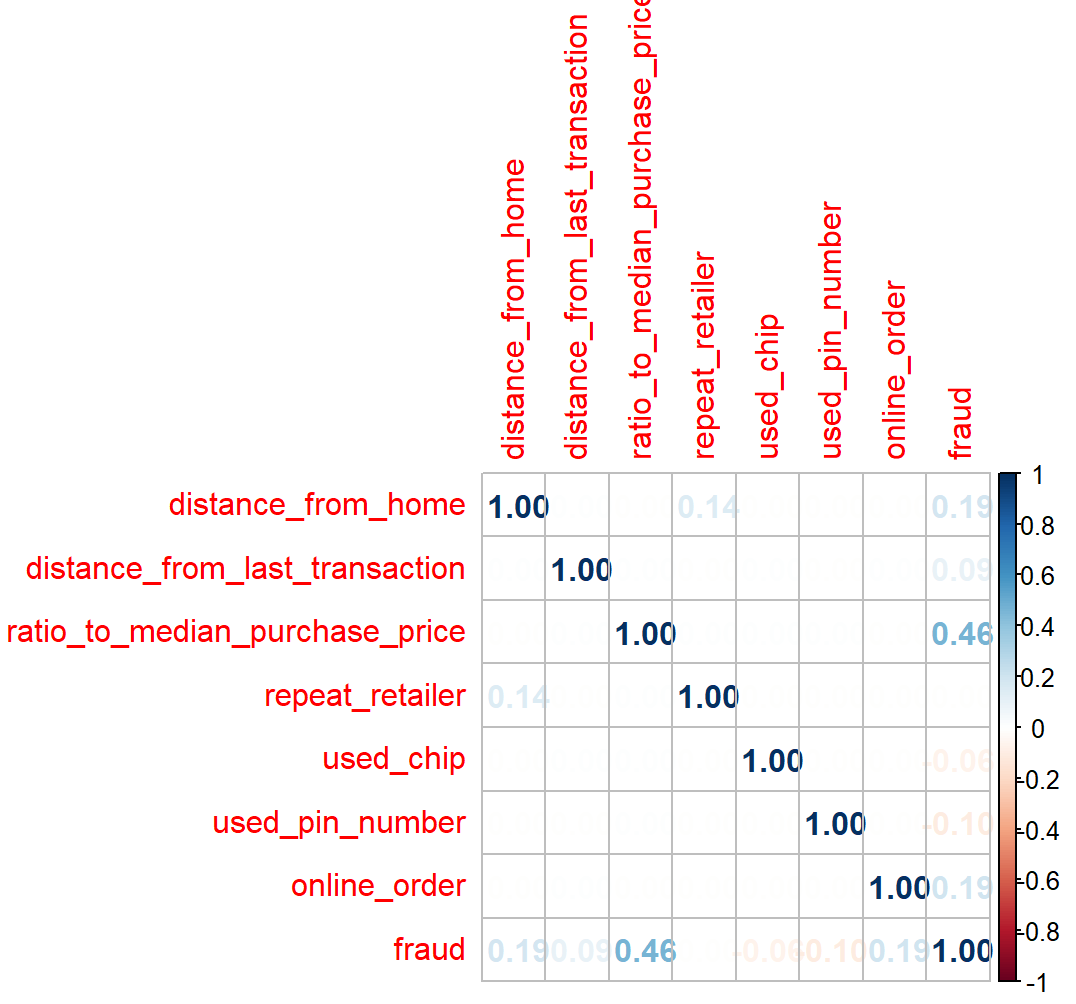
There are no missing values in our dataset.

**Univariate Analysis**

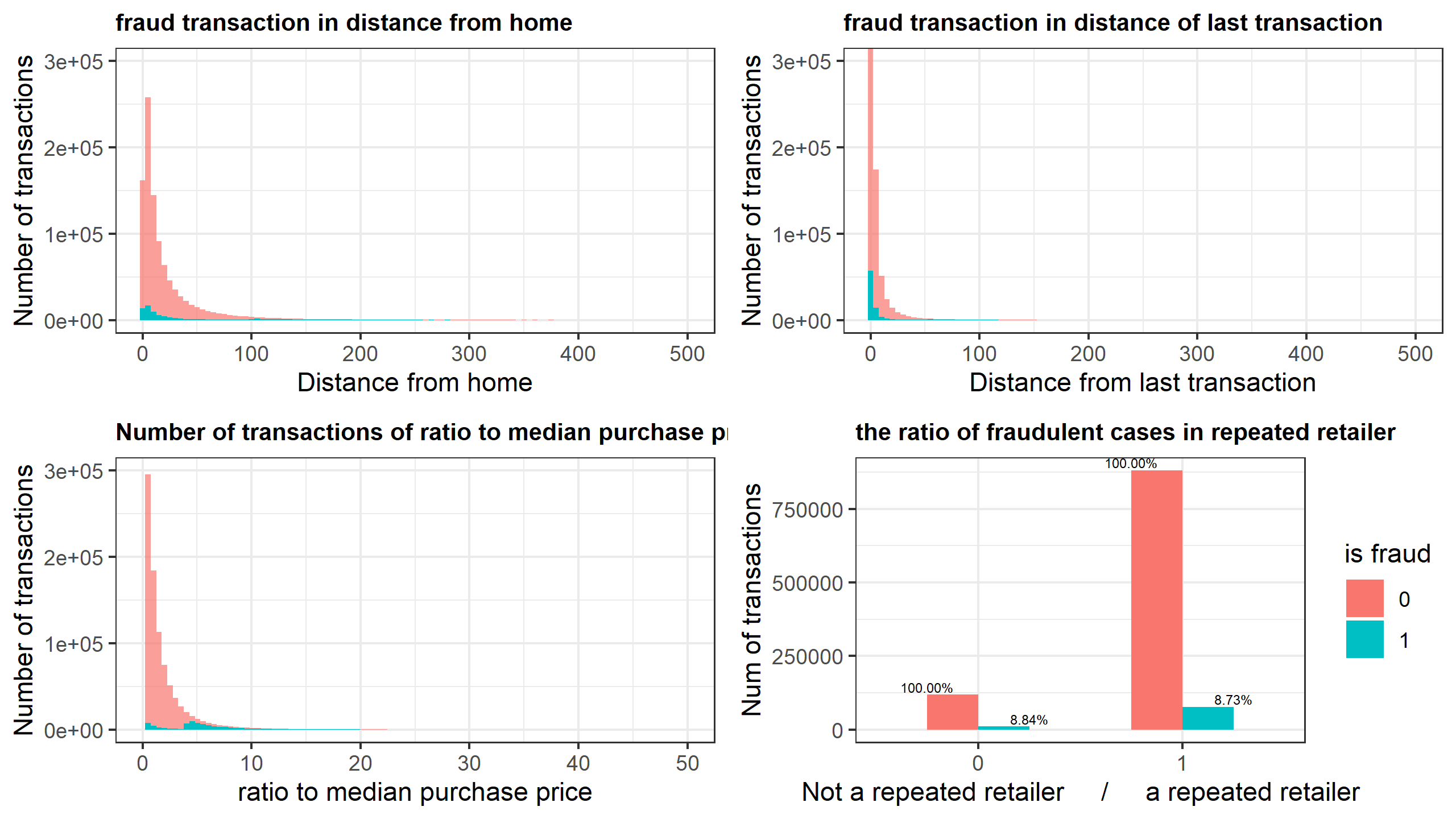
|  |  |
| --- | --- |
|  |  |

In the above graph, it indicates that non-fraud cases and repeated retailer cases are the majority in the dataset.

**Bivariate analysis**



We can see that the relationship between attribute ratio to median purchase price and fraud cases are much higher than others. There are also some weak relations between the fraud attribute and attribute distance\_from\_home and online\_order.

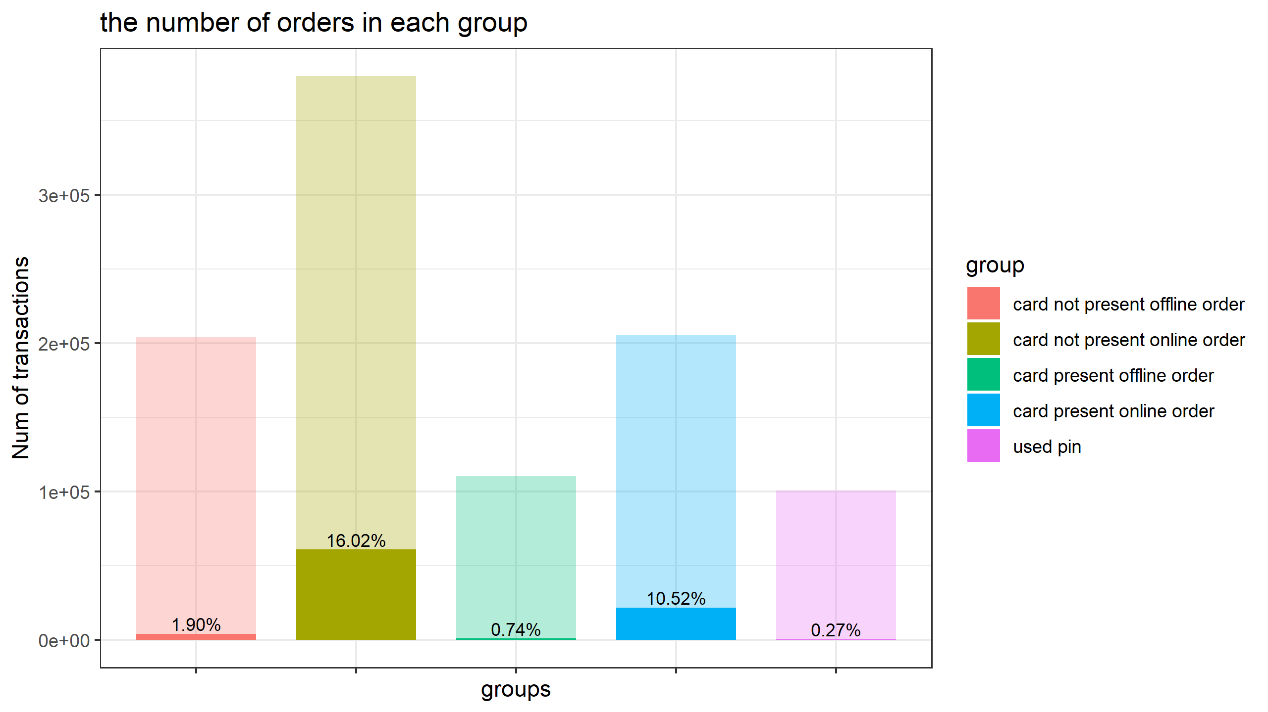


The above graph plotted the number of transactions of fraudulent cases and all cases in dataset. The red area indicates the number of all transactions in its corresponding attribute. The blue area shows the fraud cases in that attribute.

The majority of distance\_from\_home and distance\_of\_last\_transaction have a small value. However, there also some fraud cases distribute widely in large values.

The distribution of ratio to median purchase in shown in bottom left graph. We can also divide the fraud cases into two parts. The part with small ratio has fewer cases than the higher ratio part. The fraud percentage is increases when the ratio to median purchase price is high.

**Multivariate analysis**



We divide all the transactions into 5 groups bases on attribute used chip, used pin number and online order.

When pin number is used, the transaction will be divided into used pin group (purple group) The remaining cases, which do not used pin number, will be divided into other 4 groups: card not present offline order group (red group), card not present online order (yellow group), card present offline order (green group) and card present online order (blue group).

In the above graph, we can see the number of transactions in each group and its fraud percentage. The fraud percentage of offline order (green group, red group) and used pin group (purple group) is much lower than other two. Although the number of transactions in online order (yellow group, blue group) is large, they have a high fraud rate. The online transactions are more likely to be a fraud case.

As the distribution of each group in attribute distance from home, distance from last transaction and ratio to median purchase price are similar to ungrouped statistics, which is plotted in bivariate analysis. We will not provide the detailed graph here.