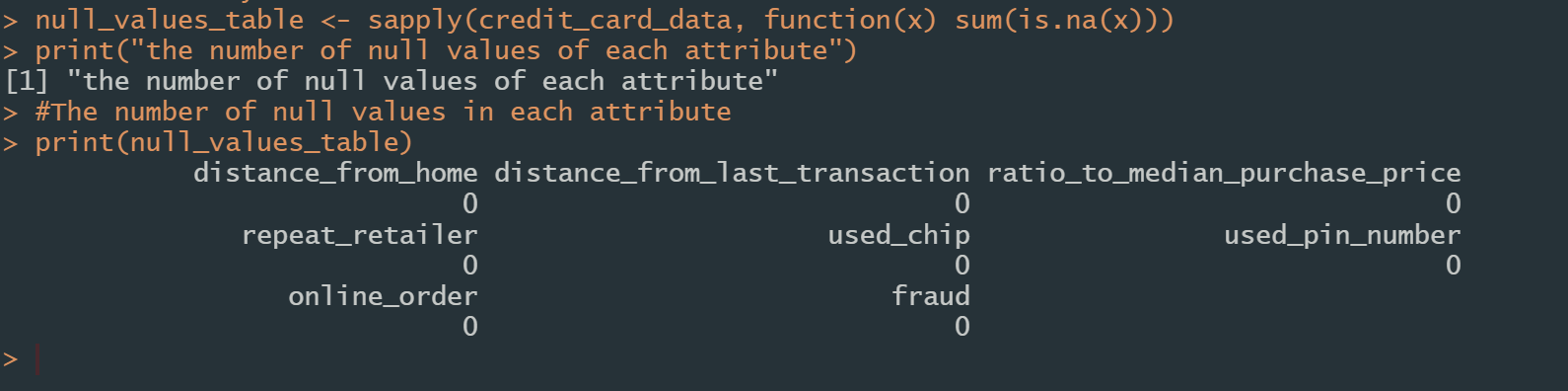


As shown in the above table, there are 8 attributes and 1,000,000 observations in our dataset.

**Outlier Analysis**

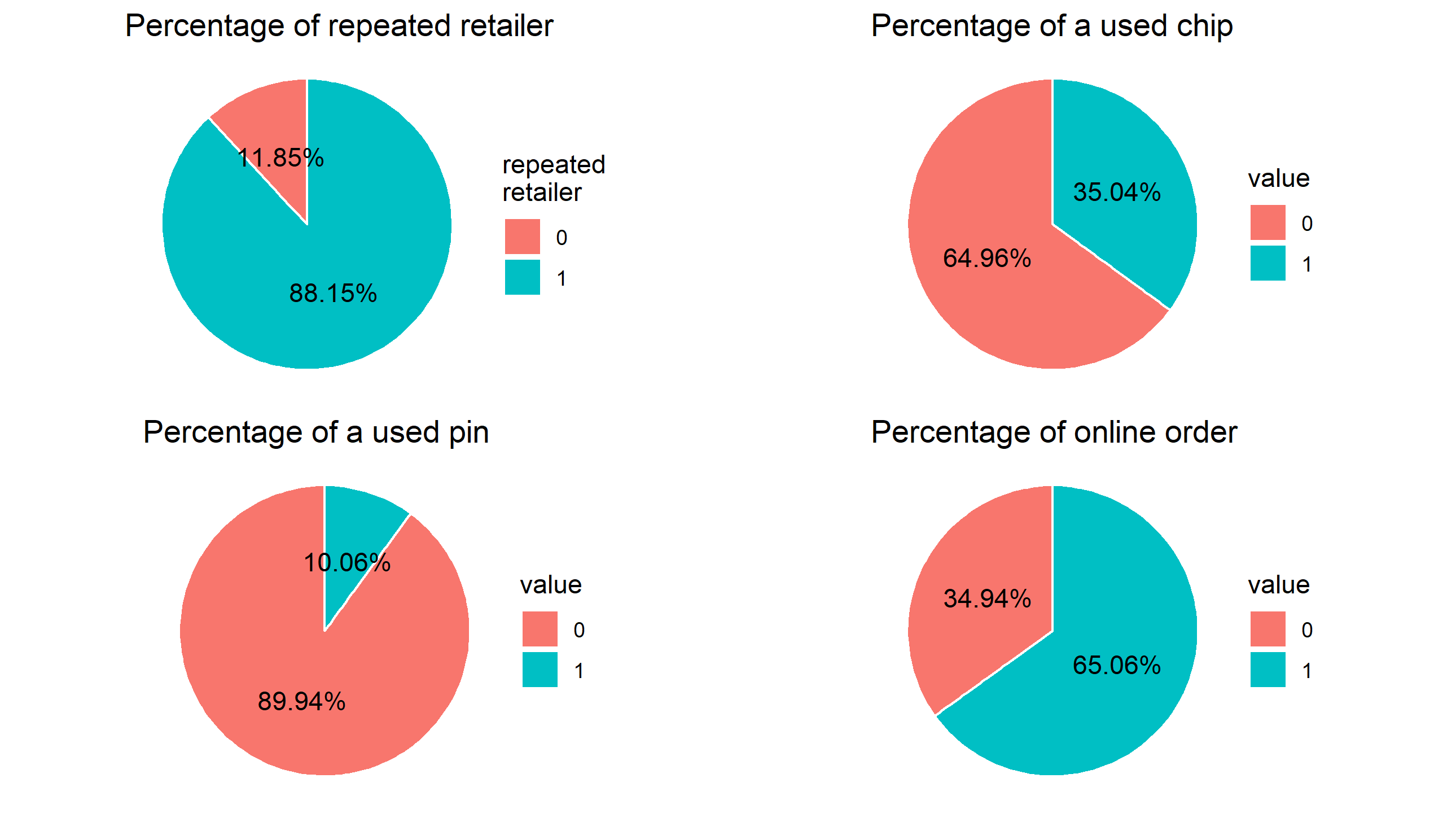


|  |  |  |
| --- | --- | --- |
|  |  |  |

There are no missing values in our dataset, and no negative value in every attribute. Although there are some extreme cases, we can’t verify that they are mistakes or just some special cases, so we can’t remove them from the dataset.

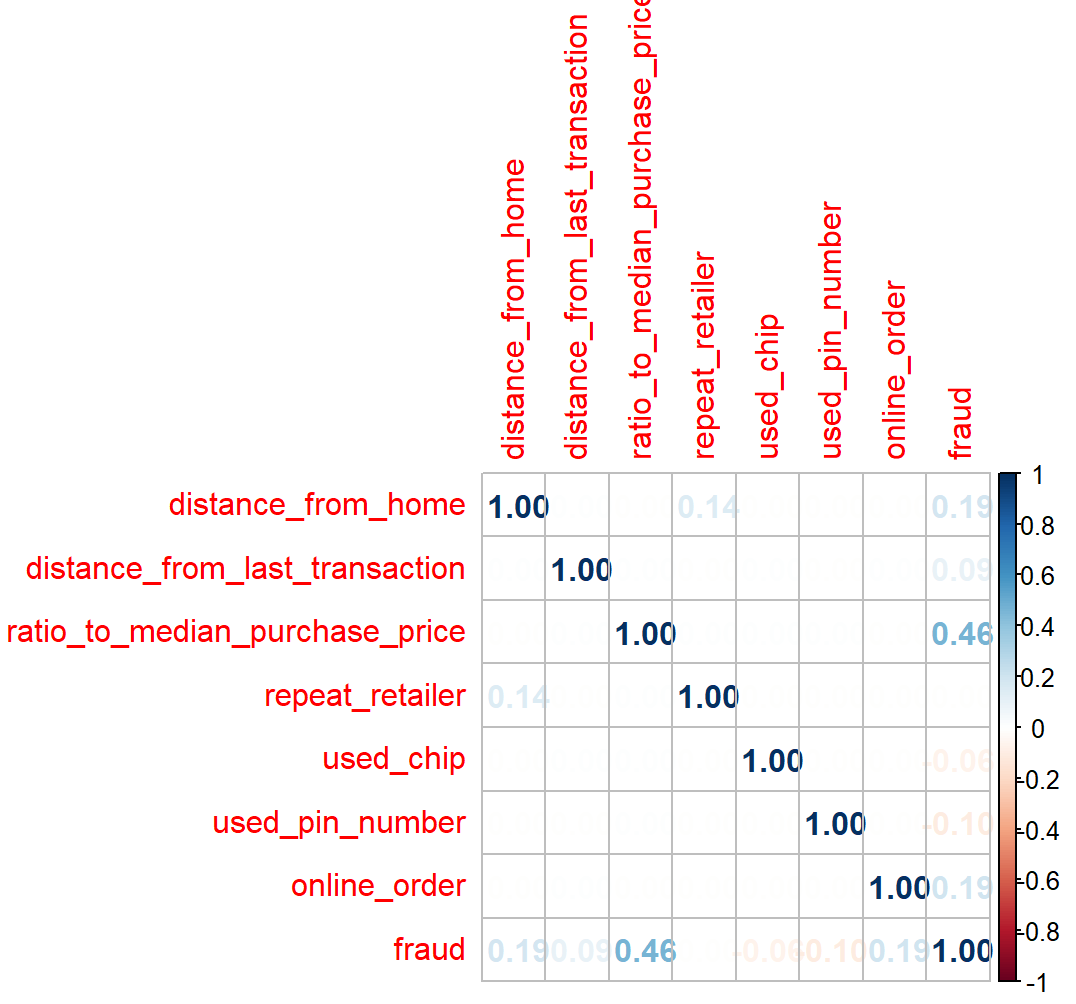
**Univariate Analysis**

|  |  |
| --- | --- |
|  | This is an unbalanced dataset as normal cases takes up to more than 90 percent of all cases. |

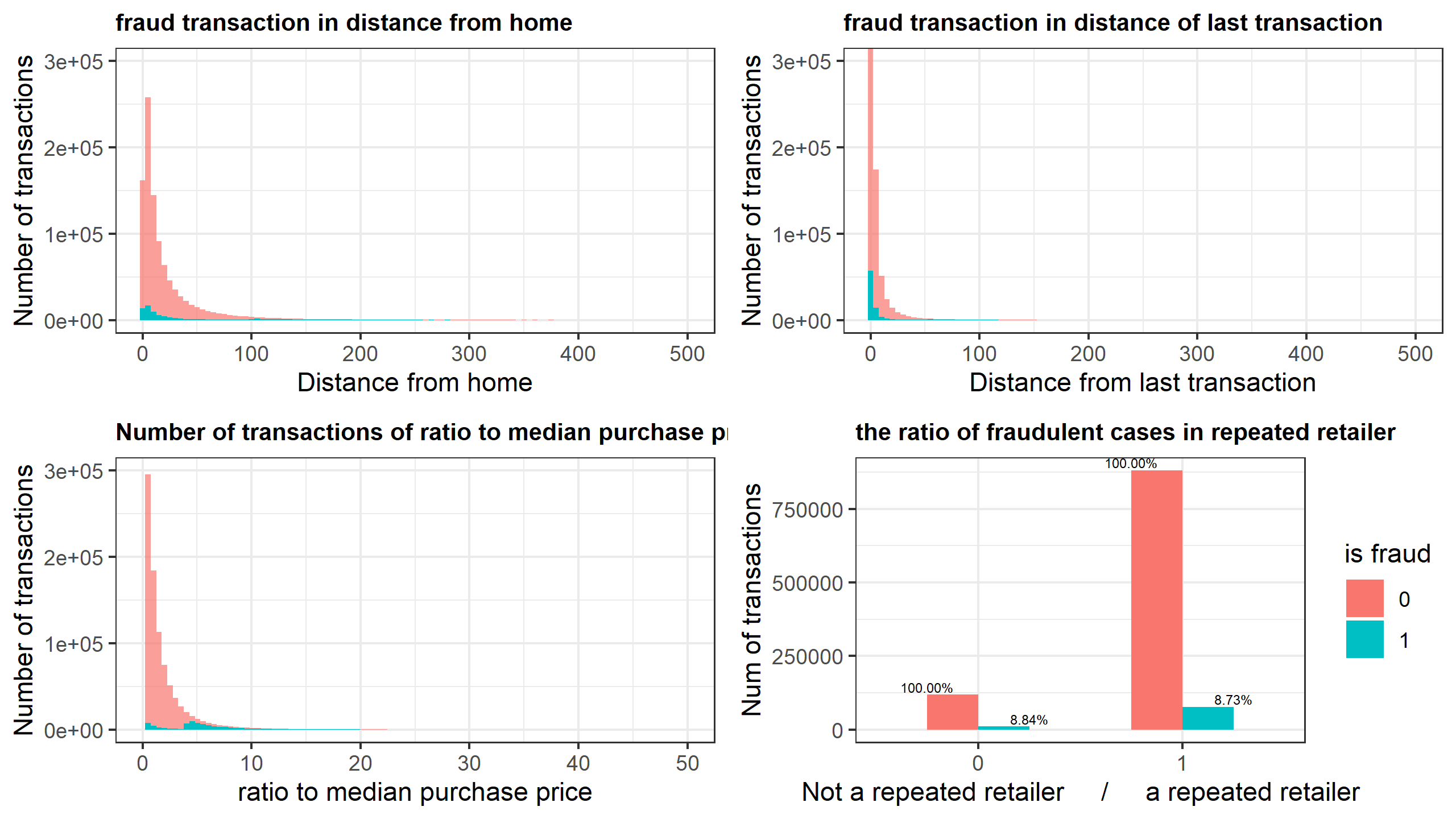


In the above graph, we can see that the majority transactions are online orders and literally people do not use pin number when they do a payment. Moreover, it indicates that repeated retailer cases are the majority in the dataset.

**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 goes up. The fraud percentage of repeated retailer cases are almost distributed evenly, which means this attribute is not relevant to fraud attribute while the others more less has some relation with fraud.