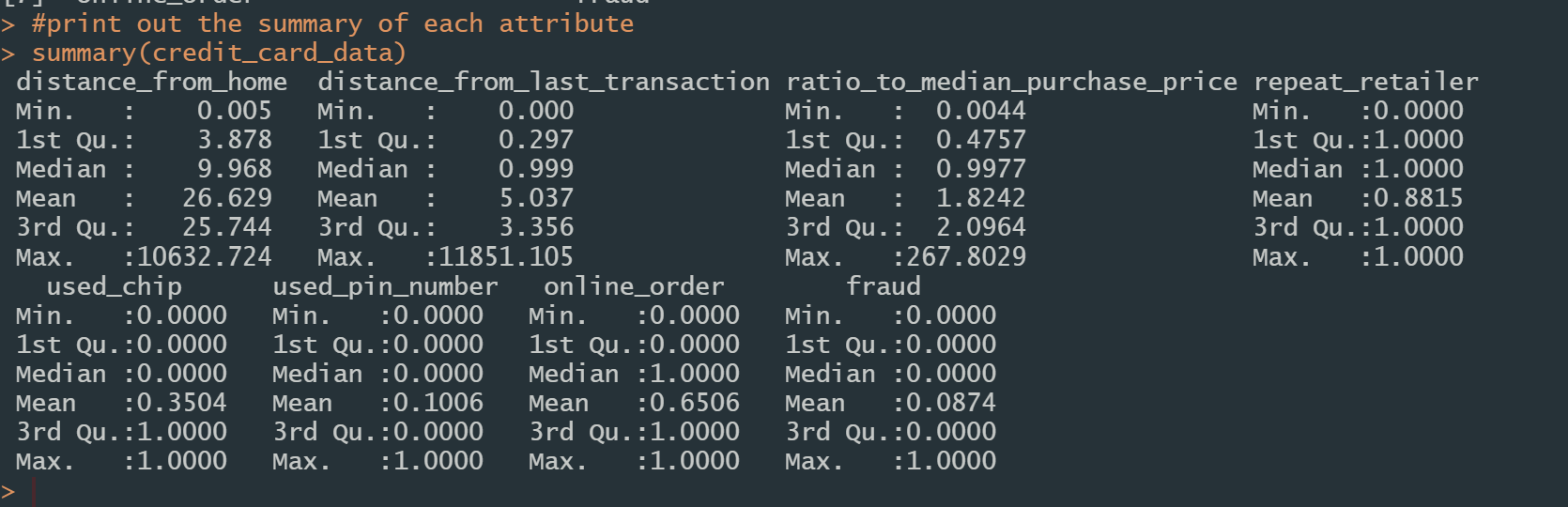
**Univariate Analysis**



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Distance from home | Distance from last transaction | Ratio to median purchase price | Repeat retailer |
| Min | 0.005 | 0.000 | 0.0044 | 0.0000 |
| Max | 10,632.724 | 11,851.105 | 267.8029 | 1.0000 |
| Median | 9.968 | 0.999 | 0.9977 | 1.0000 |
| Mean | 26.629 | 5.037 | 1.8242 | 0.8815 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Used chip | Used pin number | Online order | fraud |
| Min | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Max | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| Median | 0.0000 | 0.0000 | 1.0000 | 0.0000 |
| Mean | 0.3504 | 0.1006 | 0.6506 | 0.0874 |

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

The attribute distance from home represents the distance from home where the transaction happened.

The attribute distance from last transaction represents the distance from last transaction happened.

The attribute ratio to median purchase price means that the ratio of purchase price transaction to median purchase price.

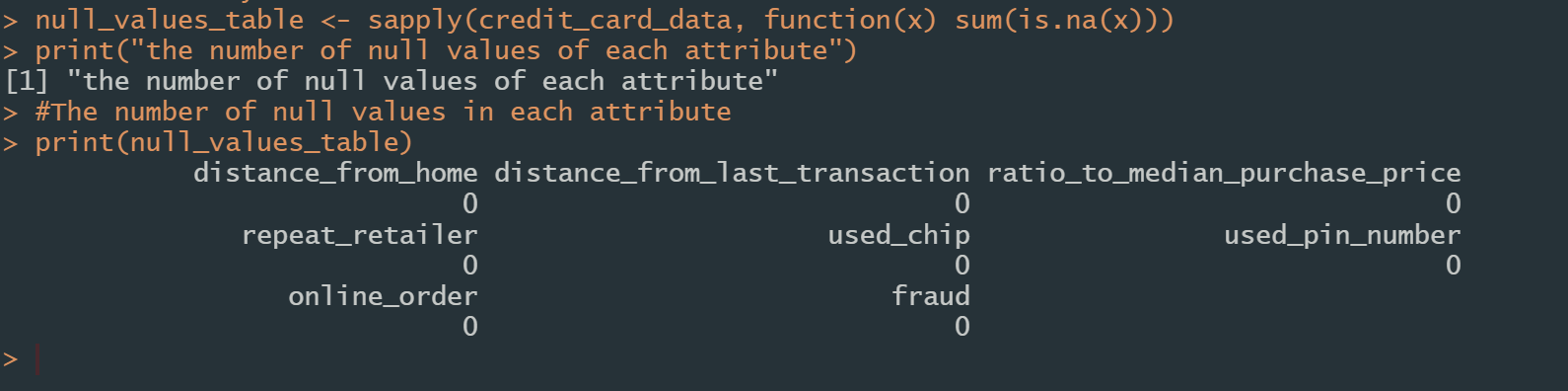
Attribute repeat retailer represents that is the transaction happened from the same retailer before.

The attribute used chip represents that is the transaction through chip (credit card).

The attribute used pin number represents that is the transaction happened by using PIN number.

Attribute online order means that is the transaction an online order.

Attribute fraud means that the transaction is fraudulent.

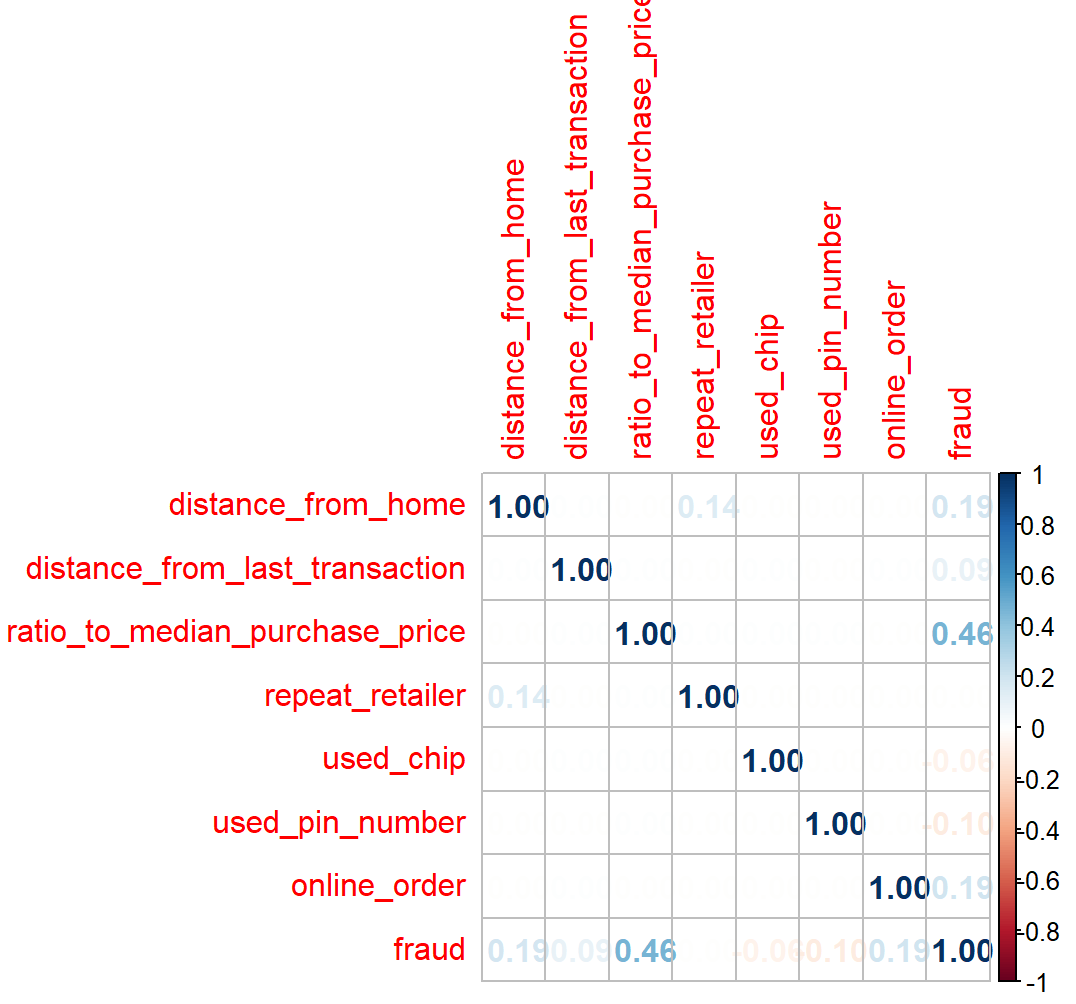


There are no missing values in our dataset.

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

In the above graph, it indicates that non-fraud cases and repeated retailer cases are the majority in the dataset. They take up 88.15 and 91.26% of all transactions respectively.

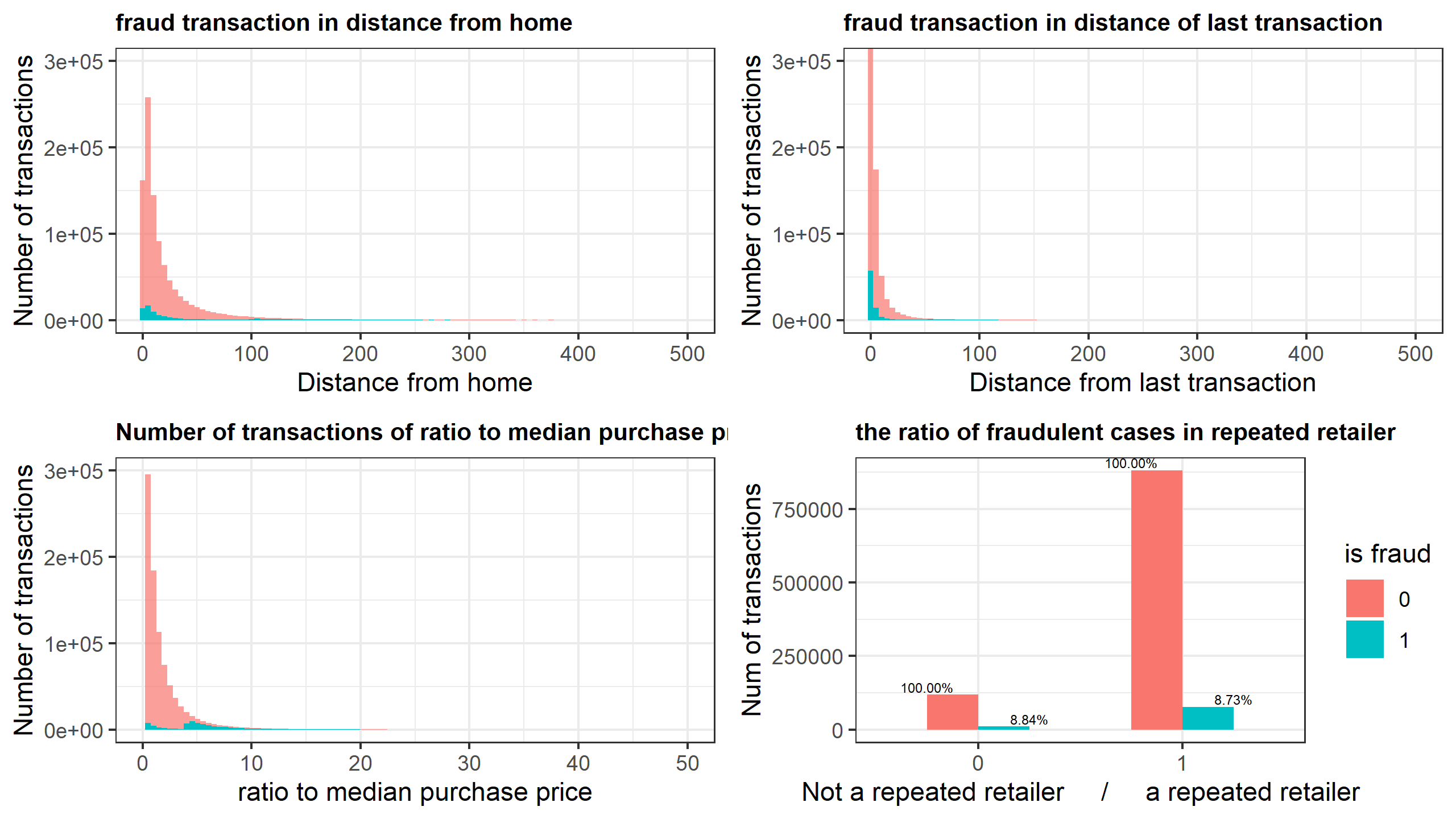
**Bivariate analysis**



In the above heat map, we can see that the relationship between the attribute of median purchase price and fraud cases is much higher than others.

There is a weak relation between distance from home and repeat retailer. The transactions that are not a repeat retailer all have a small value (less than 2). Compared with the other cases, the value of distance from home is distributed in a wide range.

There are also some weak relations between the attribute of fraud and the other attribute. We will illustrate it in the following content.



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.

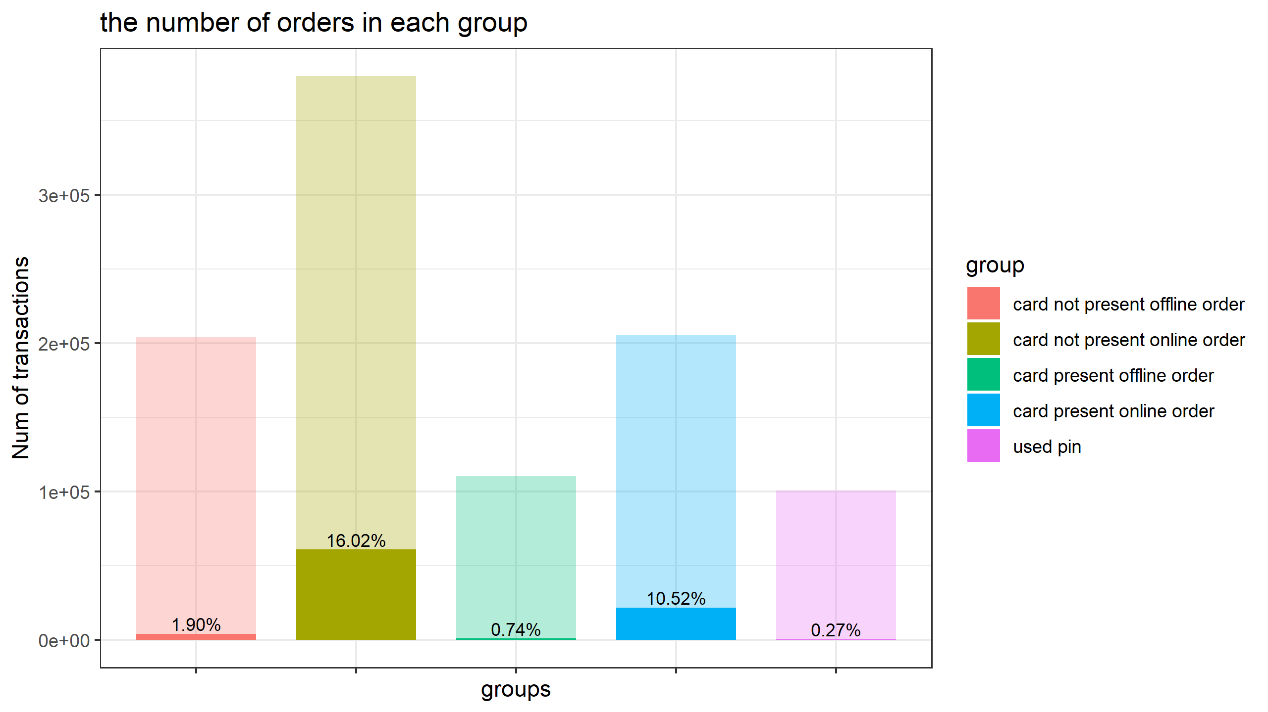
In the top left graph, we can see that the majority of distance from home have a small value. The larger the distance from home, the number of corresponding transaction decreases. However, the distribution of fraud cases does not follow the pattern of non-fraud cases. We can divide fraud cases into two major parts. The first part is composed of small values. The second part is distributed widely in large values. The number of fraud cases in the first part is higher than in the second part.

In the top right graph, the distribution pattern is similar to the distance from home attribute except the value is smaller and the gap between the two parts is narrower. The values of the attribute distance of the last transactions in all transactions are smaller than the distance from home. We can see the majority of cases are concentrated in the first bin. We can also divide the fraud cases into two parts; the part with smaller values is mainly grouped in the first and second bins. The part with larger values is distributed in a wide range. The first part has many more cases than the second part.

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

In the bottom right graph, it is the distribution of fraud cases in repeated retailer. As we can see in the graph, the percentage of fraud cases in both partitions is similar, which means that there is no relationship between repeater retailer and fraud cases. We can also notice this information in the heat map.

**Multivariate analysis**



We divide all the transactions into 5 groups based 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 use pin number, will be divided into other 4 group. In other words, 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) do not contain transactions that use pin numbers.

In group card not present offline order (red group), it is composed of the transactions that do not used a chip and it is an offline order.

In group card not present online order (yellow group), it is composed of the transactions that do not used a chip and it is an online order.

In group card present offline order (green group), it is composed of the transactions that used a chip and it is an offline order.

In group card present online order (blue group), it is composed of the transactions that used a chip and it is an online order.

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.

Moreover, used chip transactions have a slightly lower rate than fraud case.

As the distribution of each group in attribute distance from home, distance from the last transaction, and ratio to the median purchase price are similar to ungrouped statistics, which is plotted in bivariate analysis. We will not provide a detailed graph here. As our group is based on attributes used chip, used pin number and online order, we can see there is no relationship between these attributes with attribute distance from home, distance from last transaction and ratio to median purchase price in the heat map.

The distribution in attribute repeat retailer of each group is similar to all the number of transactions in each group, which is plotted in multivariate analysis. As we can see in the heat map, there is no relation between repeat retailer and attribute used chip, used pin number and online order.