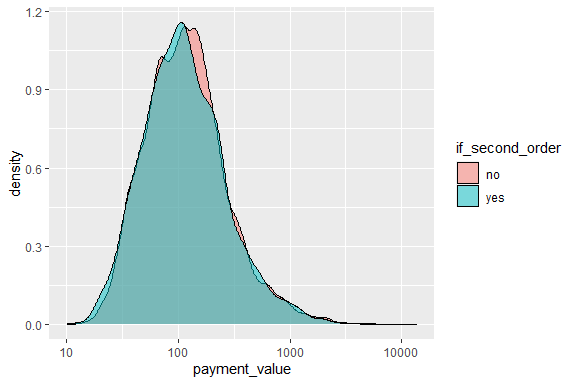
Analysis of model results

Libraries, loading the dataset etc

#### Payment value

On the plot … the values of payment for each order are presented. I have used the Kernel Density Estimation technique to smoothen the plot. As the distribution is highly right-skewed, I have logarithmed the values. The density plot is grouped by the fact whether the particular customer also created a second order later. It can be seen that the 2 densities almost overlap. This means that payment value would not be a good predictor in an univariate approach - although maybe it can be interacted with other features and start having predictive power.

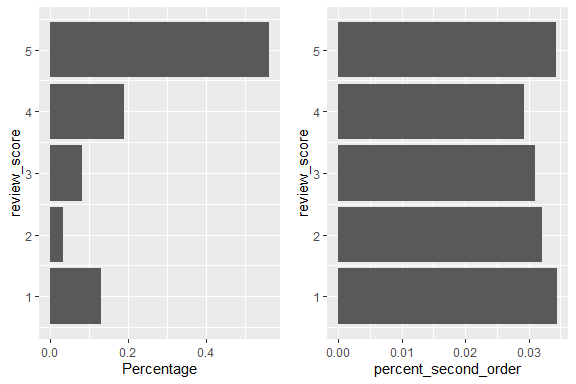


#### Review score

On the plot .. percentages of orders that were given x stars in the review are shown. On the right subplot percentages of the customers that made a second order are presented. Most of the reviews are positive - the scores 4 and 5 make up for 75% of the whole dataset. Another thing worth noticing is the tendency to the negative score polarization - if the customer is unsatisfied with the order, it is more likely for her to give the lowest review.

The Relationship between making a second order and review score for the first one is somehow surprising. One would expect that if the client is unsatisfied for the first time, she will never buy in this store again. In the case of this dataset it is the opposite - the customers that gave one-star review are also the most likely to make the second order. It is worth noting is that the differences between the groups are very small - between 2.9% for review 4 (smallest one), and 3.45% for review 1. One can wonder if this can come simply from random reasons, and that the review score does not influence the probability to come back at all. In particular, the difference between the percentages for the scores 1 and 5 (0.003%) is that small that it most likely for random reasons.

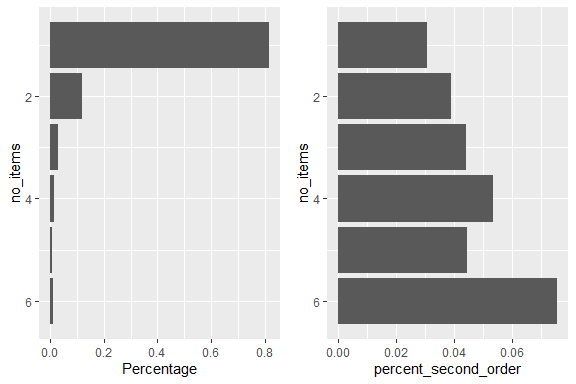
One should bear in mind that the observations avaliable in the dataset are not the complete customers data that the Olist company has. Rather, they were somehow sampled. The dataset authors claim that it represents the customers base in a complete manner. However, there was some sampling bias introduced while creating the dataset that is based on the value of the review score. Namely, in real life case customers do not have to provide star review of every order. The authors of the dataset sampled the orders database in such a way that they excluded the orders, for which the review was not given. One should bear in mind that the analysis of review score is incomplete because of that - one would wonder if there are factors that influence the customer to provide the review, and the very fact of providing the review changes the probability to buy for the second time for that particular customer.



#### No items - numbers and percentage

On the plot .. analysis of number of items in the order is presetned. There were also orders with number of items above 6, however they make up for 0.2% of the dataset only, that is why I excluded them for clarity of the plot. On the left subplot is shown the percentage share in the full dataset, while on the right one - percentage of the customers that put second order after ordering x items for the first time.

A trend is clearly visible - the more items the customer has bought in the first order, the more likely she is to also put the second order. This difference is pretty strong - between 1 and 6 items the percentage increase in the response is almost 150%. However, one should bear in mind that big orders are extremely rare - 93% of the customers buy one or two items only.

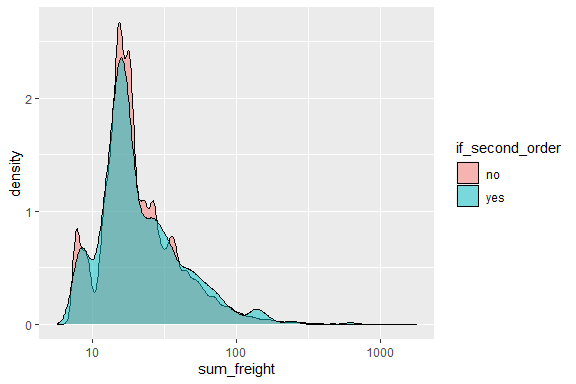


sum\_freigt

Sum of the final transportation costs that the customer had to make. The distribution is highly right-skewed, so for clarity I have log-transformed the values.

No clear distinction between two groups of observations are apparent.

An interesting thing to check is the relationship between the value of the ordered products and the transportation cost. Pearson correlation between these two is 0.5, meaning that the value of the items ordered somehow influences the rest of the costs.



Map

Load libraries, process the dataset

Draw map

