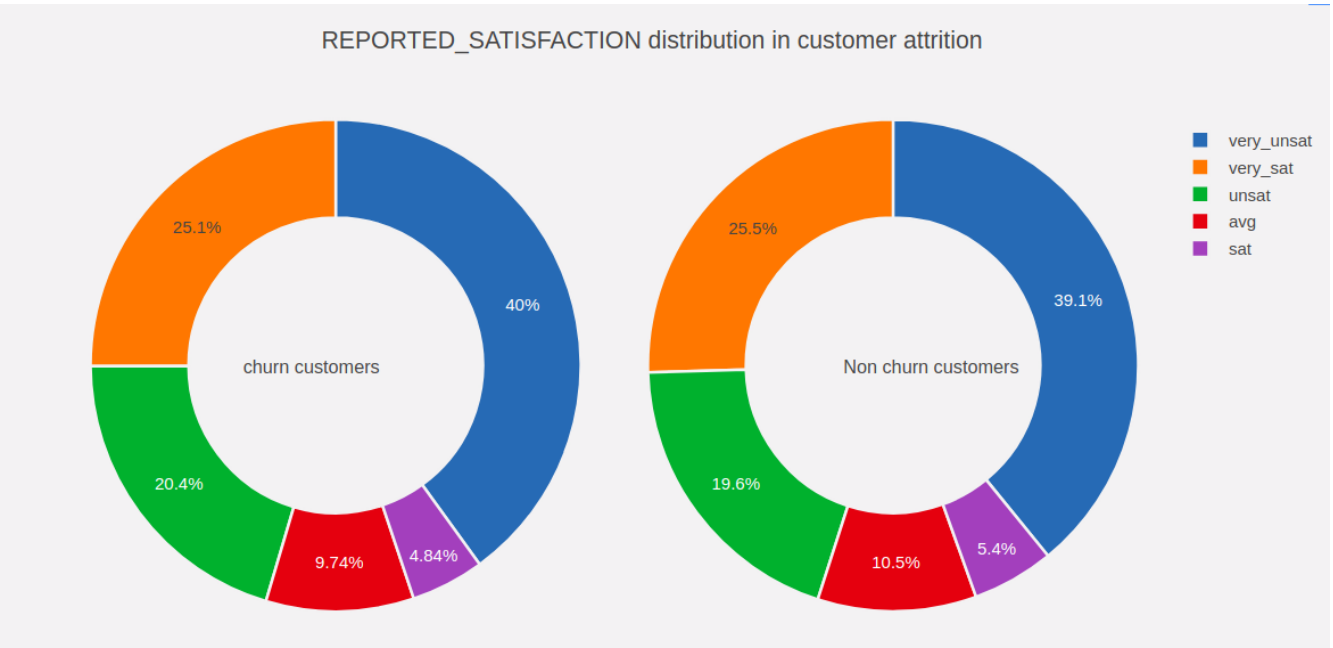
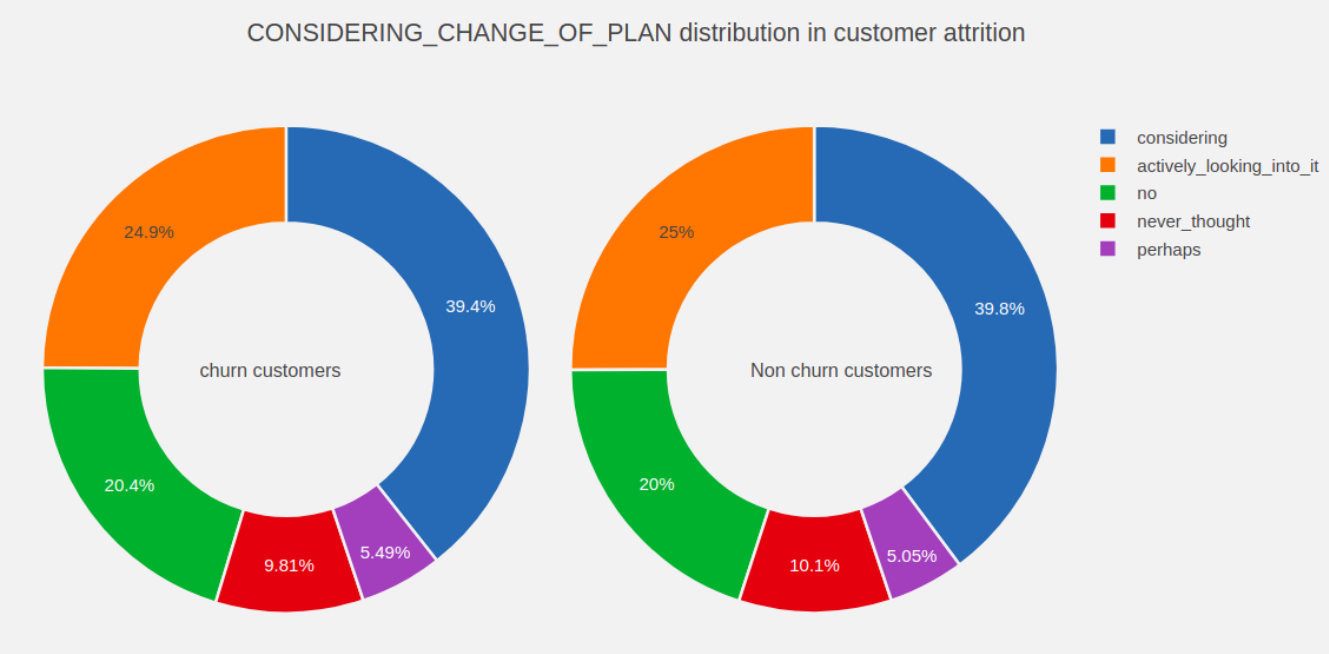
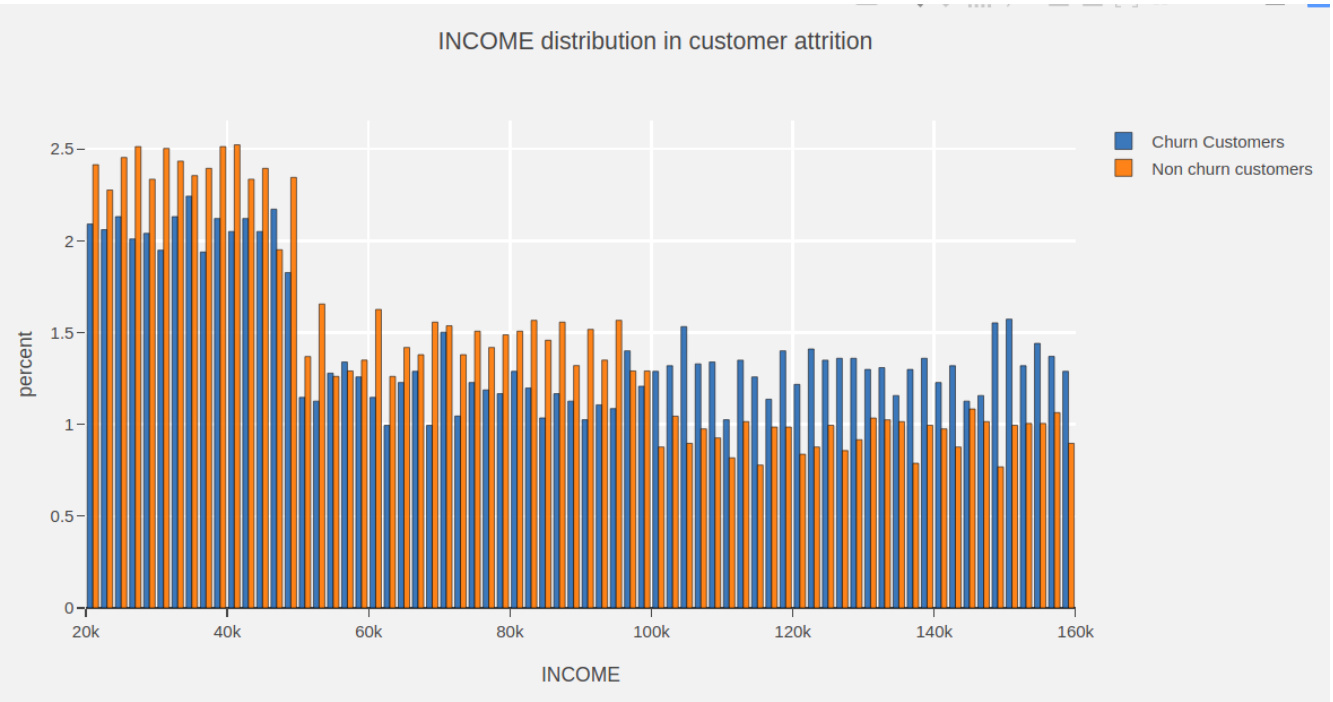
Observations: Customer Churn Dataset



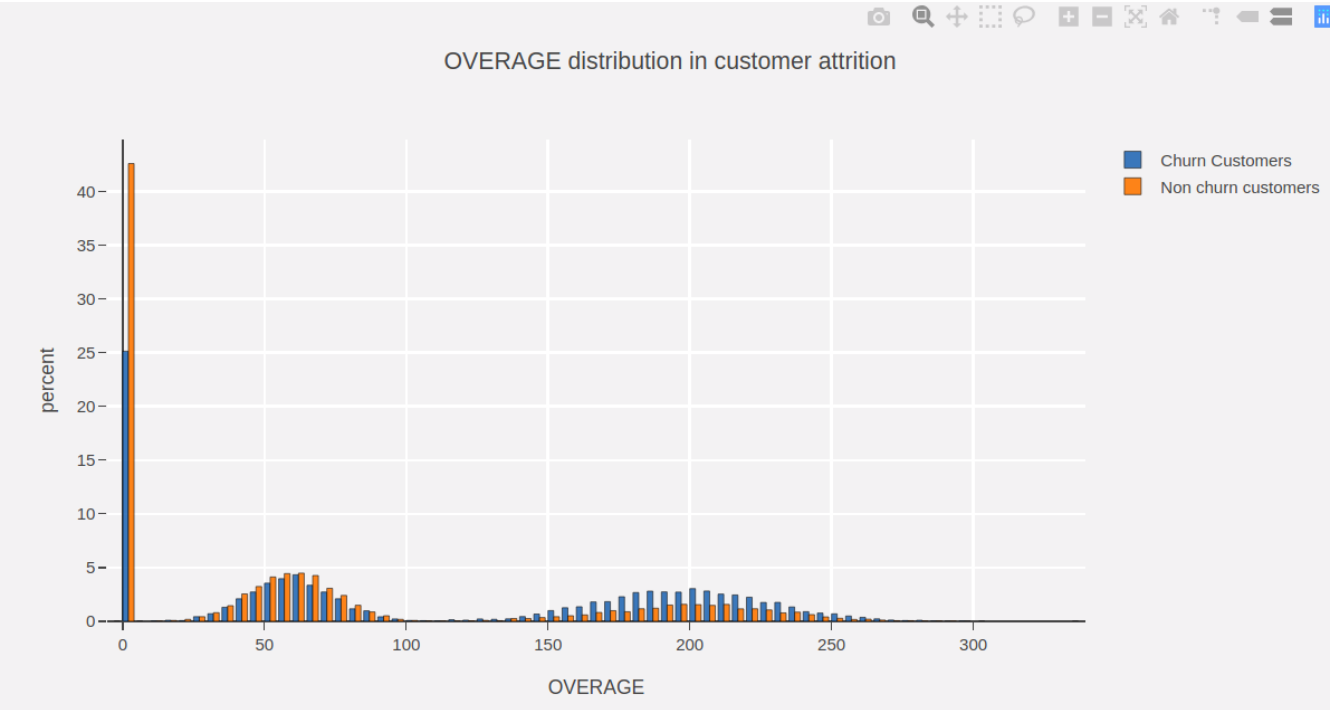


In all the donut graphs, we can see that the distribution is very symmetric. Considering the data generation process which consist of selection of random customers and letting them enjoy the service and then letting them decide whether they want to leave or not, this kind of symmetry is very ususual. My point is, this data is artificial or made up.

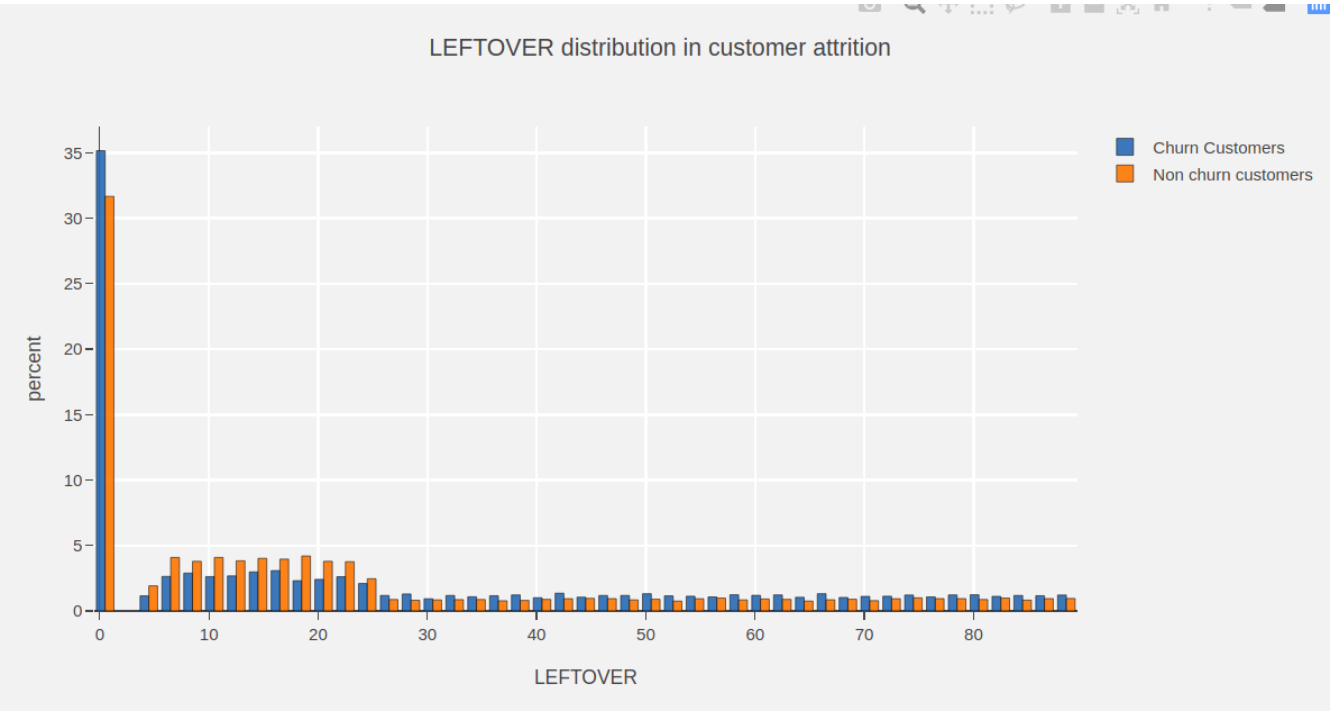
We will try to plot further continuos variable graphs to understand whether we can support this hypothesis or not.



This graph shows that below 90K income, there are more non-churners than churners, and after 90K, the population consists more of churning customers. It's surprising we have such division in customers. There should be something which is influencing this kind of trend.



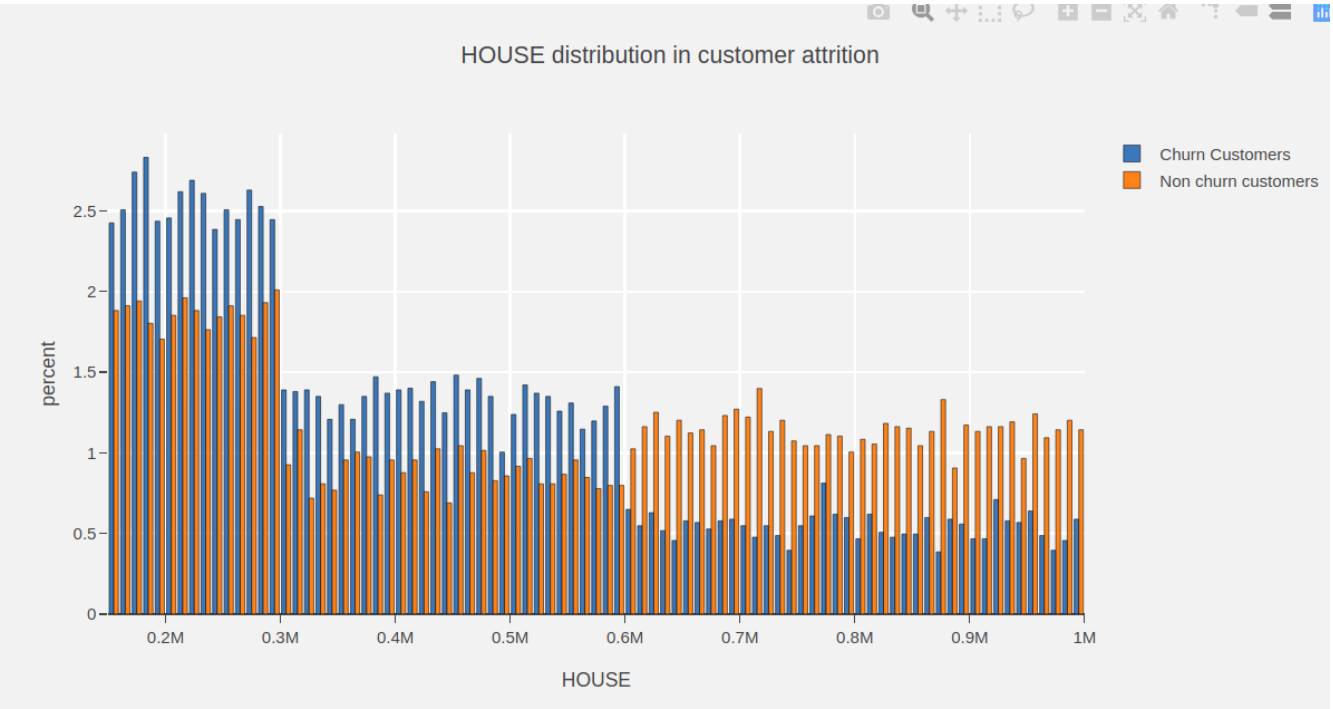
The overage distribution is quite surprising, as the graphs are very symmetric. There must be something wrong in the data generation process as such kind of symmetry is very unusual. Either all the customers are same in groups or the company is regulating the customers which they are already aware of.



Same story.The leftover is also symmetric as the number increases in the x axis. I can say with confidence now, that this part of the dataset is artificial.

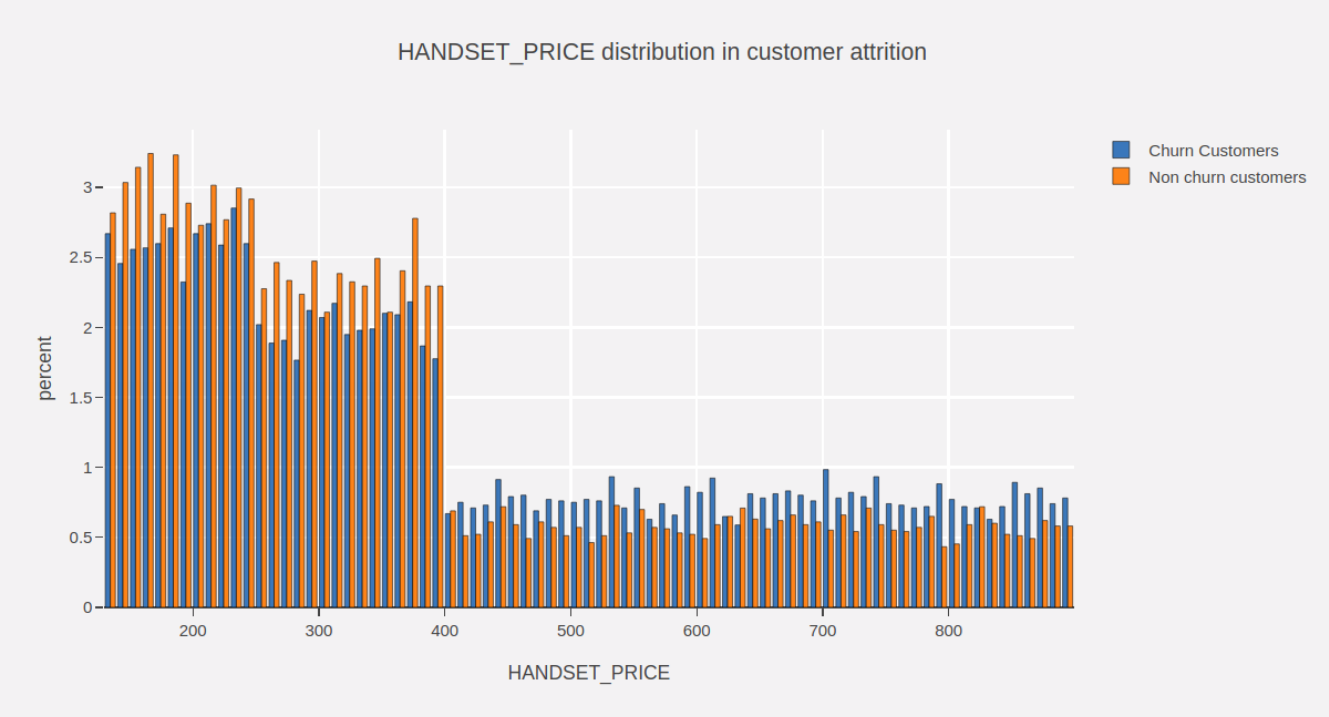
If the data comes from the randomized trail, and there is no influencing effect on the outcome of customer's action, this kind of symmetric distribution is highly unlikely.

Moreover, this dataset is highly skewed and missing up a lot of datapoints which can make this in accord with the data generation process.

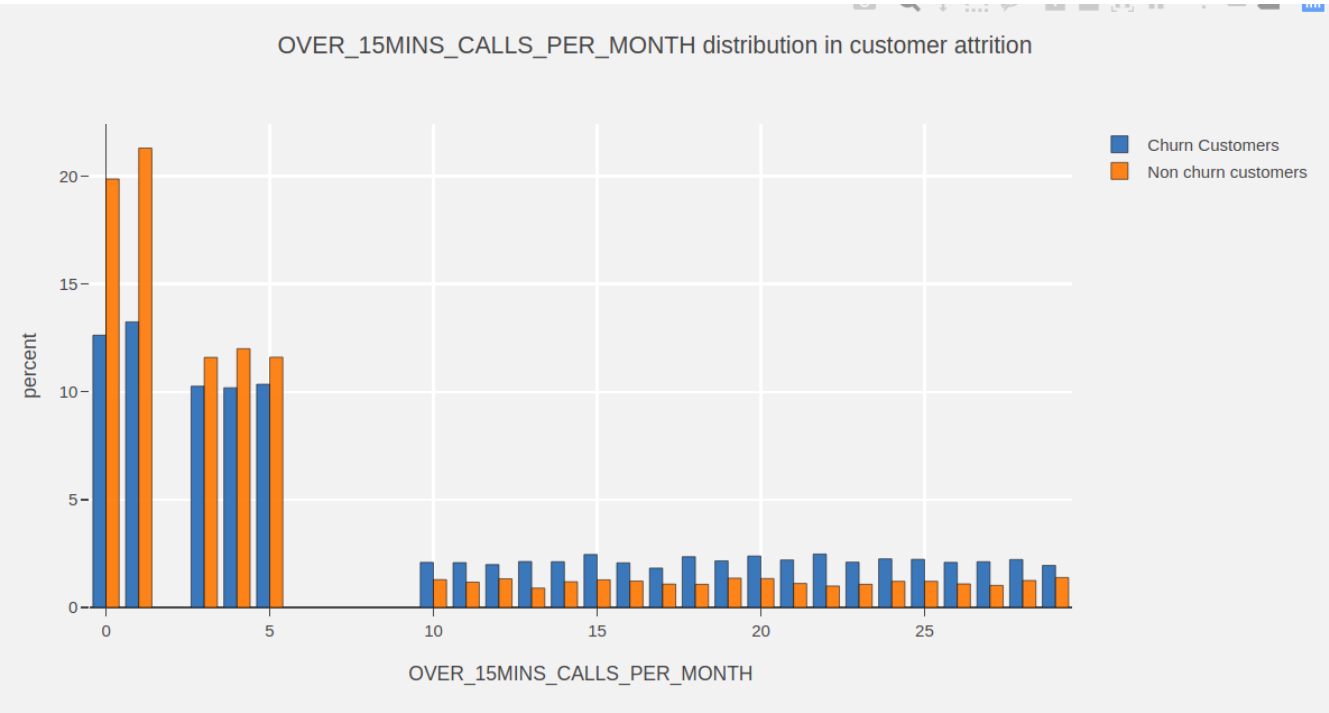


The distribution shows that before the 0.6M mark, the churning customers have higher house evaluation and after the 0.6M mark, the non-churning customers have the higher house evaluation. This is not in sync with the income graphs, as the income graph mentioned that the churning customers has higher income.

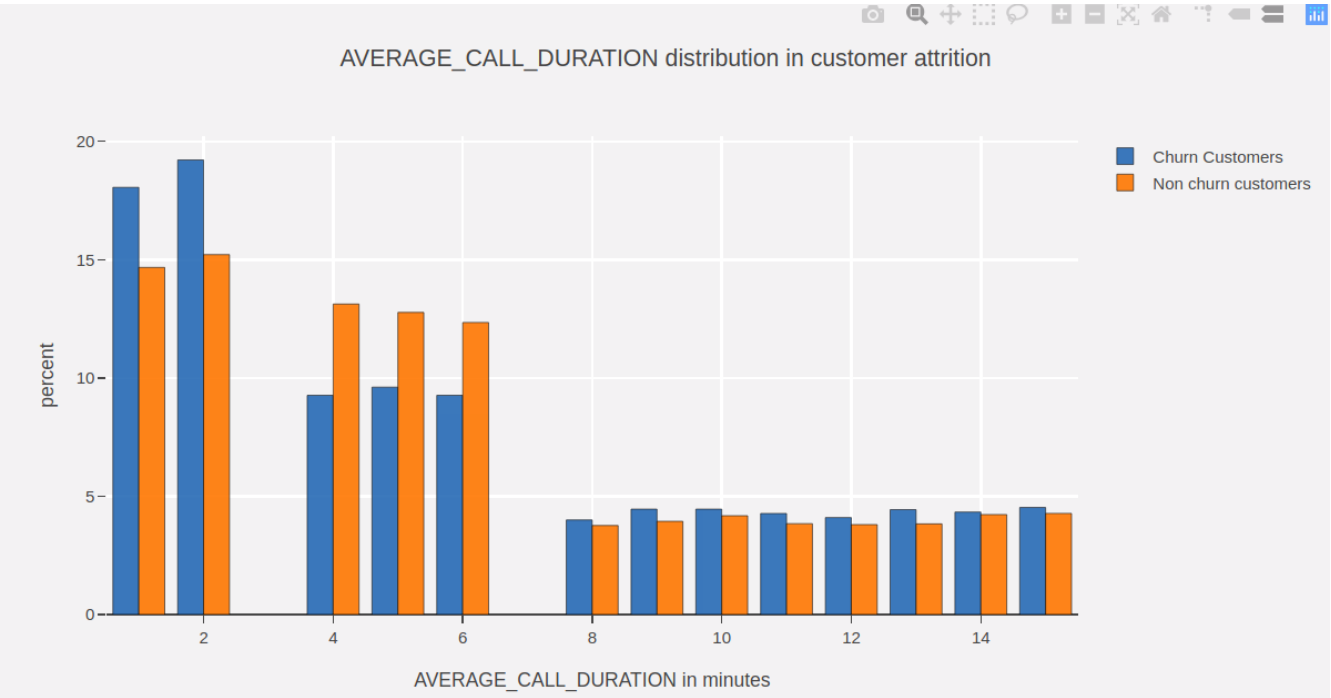
The house distribution here consist the value of churning and non-churning customers in the same ratio. This also seems to be artificial. The sudden drop after 0.3M and then the shift in distribution after the 0.6M mark, seems to be not alright.



The handset price distribution shows that after the 400 dollar mark, there is a sharp fall and there is very clear boundary between the churners and non-churners. This data is not reliable as majority of the people in USA use iPhone which costs over $600 (according to statista.com)



In this distribution for over 15 minute calls per month, we can see that all the data points are very much symmetric. Considering the data generation process which I discussed above, this is highly unlikely and it suggests that the customers which are making long calls tend to leave. This might be a result of planning of the call rates by the telecom company.



In this distribution for average call duration, we can see that all the data are very much symmetric. The people who have a high average call duration tend to leave more than the people who have not churned.

The above graphs present us a fair insight into the data. The main take away is, the data is artificial, missing major data point and does not takes the data generation process
into account.