Supplementary information **Sergey Platonov**
BCG Case Study March 2018

Tools description

For the solution of the problem, I used python libraries: pandas (to load and manipulate the data), seaborn and matplotlib (to plot it).

Loading the data

The data is given in ten csv files and the first step is to merge the data into a single file. I did it with the help of the terminal in Mac Os (or command line in Windows). Another possible solution would be loading file by file into many dataframes and then merging them together (this way it helps to remove headers from every file in the beginning, here I do it at the step of removing NAN values).

The data had two important problems for the loading procedure. First, some of the lines contained more elements then it was stated in the description. I skipped these rows (altogether it was $\propto$ 18 rows) and later checked that they didn't influence the result. Second, there were two symbol sets used as a separator: | and $\&SDF*$. Pandas readcsv has a specified option 'separator' to solve this problem.

Structuring

After loading I removed the quotes signs from the column names and the dollar sign from the column 'BillingRate'. Next step was to convert billing month and rate to numeric data type. For simplicity, I also changed names of the columns to their simplified 2-3 letters version. The final step was to remove all the NAN values from the dataframe with pandas dropna.

Analysis

In order to find the total number of customers, I grouped customers by the billing month and then found the number of unique customer ID for every month. To calculate how many customers leave every month I took differences between total number of customers for next neighboured months. I show the result in Table(1).

Unfortunately provided dataset didn't contain transactions for the previous year and I couldn't calculate how many customers left in January. I found that the highest rate of attrition happened in July. The averaged attrition rate also increased from 46 customers per month to 99 customers per month after July.

I also performed analysis for different customer regions and found that attrition happened in every region: South – 23.7 ; North – 21.6 ; West – 43.3 ; East – 17 customers per year. The highest attrition happened in the biggest by the number of customers 'West' region. Averaged attrition rate also increased for all regions after July from 15.2 to 21 for South, from 8.4 to 20.6 for North, from 17.4 to 41 for West and from 5.4 to 15.8 customers per month for East.

Month	Attrition rate
Feb	171
Mar	-57
Apr	-55
May	149
Jun	24
Jul	440
Aug	96
Sep	66
Oct	98
Nov	198
Dec	35

Table 1: Customer Attrition rate for every month in the observation year.

Averaged customer billing rate

In order to find the reason for the highest attrition happened in July, I analyzed the customer billing rate. First, I looked in the customer distribution by billing rate. I found that the rate was distributed uniformly from the minimum value of around \$162 to maximum $\propto$ \$275. Then I made analysis of the averaged customer billing rate for different customer regions over the year. It revealed that the billing rate changed by \$5.4 uniformly for all the customers in the 'West' region. Averaged billing rates for all other regions stayed almost constant (with changes less than 5 cents). This can be one possible reason for highest attrition rate in July.

Additional data

The attrition rate as a function of month strongly suggests seasonality feature in the concerned data. For example, this can happen because students left the area for summer break or more generally this can be related to the fact that people pay less attention to television/internet in summer and spend more time outdoors. Therefore additional data can include more observation periods (at least several years) to prove seasonality. I can also simplify the task by joining months into the quartiles (as it is usually done in TV/internet provider).

I also found a strong evidence of the billing rate influence on customer attrition for July. At the end of the analysis, I also checked the distribution of left customers over the billing rates. The distribution was approximately uniform (the lack of samples didn't allow us to make a solid conclusion), however, more info about customers and their subscription will also be helpful to analyze the impact of the rate change on the customer attrition.