Customer Segmentation

The aim of the project is to build customers segments, where each segment has its own characteristics based on the features we use for this segmentation. We are using what is called in machine learning by « unsupervised learning » to detect patterns in data. In this case study of phone users, we are using Kmeans clustering algorithm, which calculates similarities between all the data rows and groups them to clusters. Clusters are « similar » within them and « dissimilar » between each other.

The features we use here are :

- Average call duration
- Average number of calls per day
- Average # SMS per day
- % daytime calls (9am -3pm)
- % evening time calls (6pm-10pm)
- % of weekday calls (Sunday - Thursday)
- % of Friday calls
- % of Saturday calls

Features extraction methodology

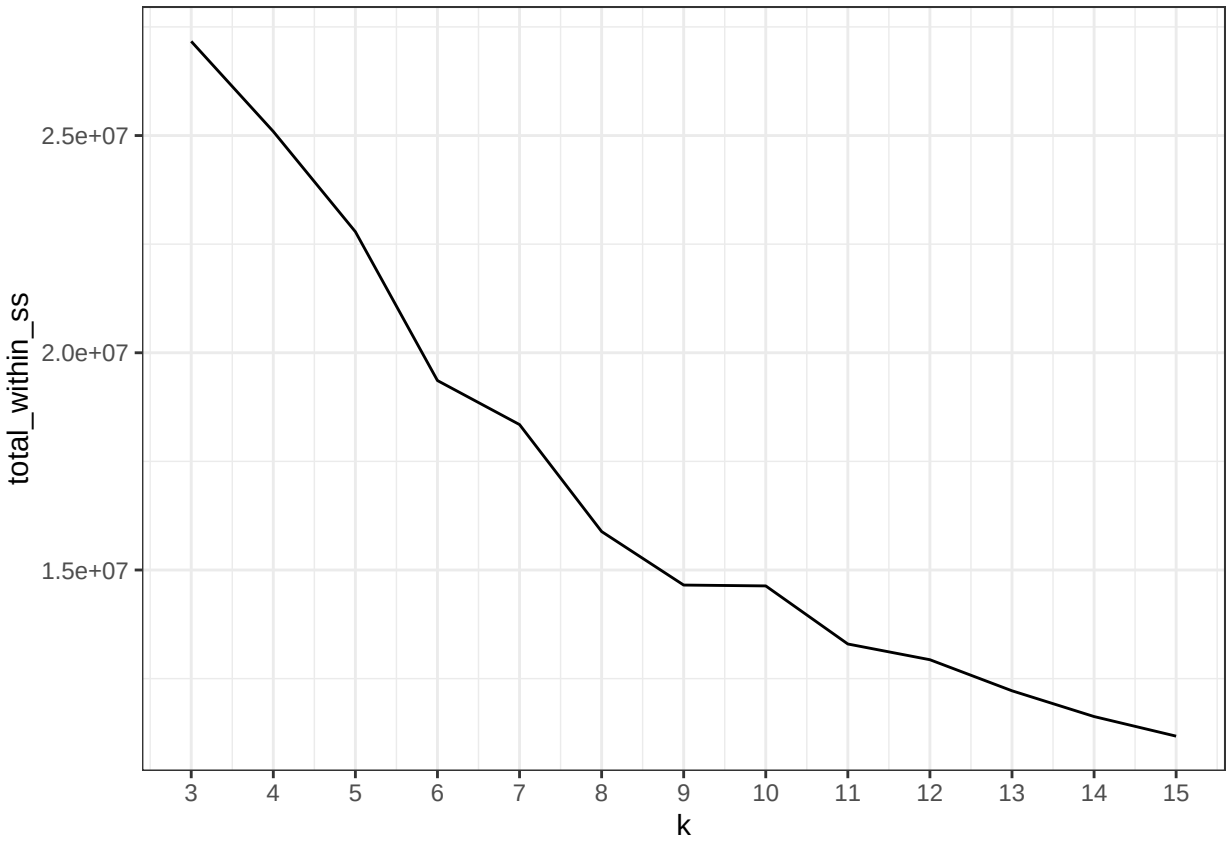
For features extracting or (features engineering), we use the data already processed (not the raw data) with 7 variables. We have used the phone number as a user id for aggregation. The date and the hour of calls/sms help to calculate all the % attributes.

The script for features extraction must run over the entire data to aggregate by user. Since the data are large and processing it once will consume memory, we have split it into nine splits with distinct users, that is we cannot find any user appeared in more than one split of data. The final features data will have one row for each user with the 8 attributes we have mentioned.

Note that the features don't vary in the same scale, % attributes related to time do vary in the same range of [0,1]. Average number of calls and sms are both integer variables, but may have widely different ranges. Average call duration has its own scale. This difference of scales will be considered in results interpretation.

In order to calculate the similarities between users based on mentioned attributes, the Kmeans algorithm uses scaled data which has united scale for all features. For this purpose, we have rescaled the features data before running the algorithm.

The Kmeans algorithm require the number of clusters as input. The choice of an optimal number of clusters will be done using the elbow method after running Kmeans algorithm many times with different values of number of clusters. The elbow method consists of plotting what is called "total within sum of squares" (which is a measure of similarity within clusters) as a function of number of clusters. The function is intuitively decreasing, since we tend to have more homogenous clusters as long we are considering more clusters. The idea is to choose the number of clusters which our criteria ("total within sum of squares") does not improve comparing to considering a smaller number of clusters.



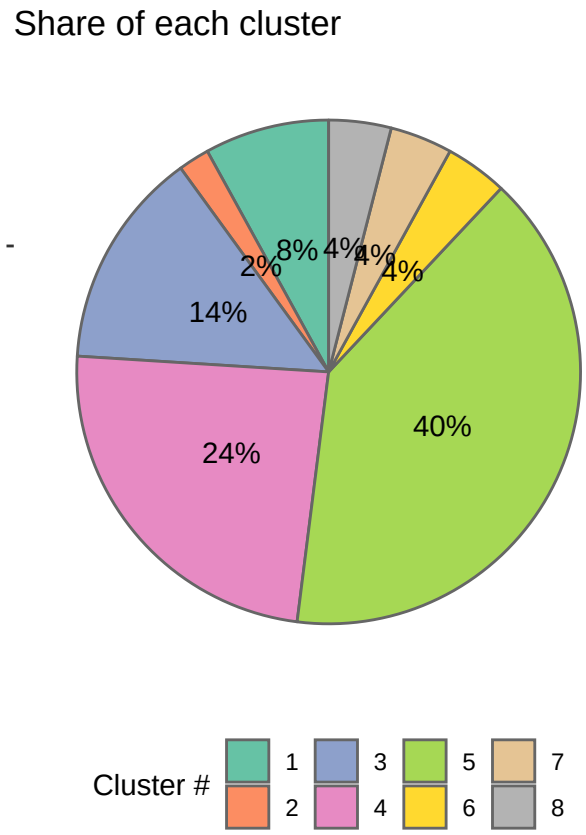
Our elbow plot suggests to use 8 clusters for our segmentation.

Results

The Kmeans algorithm ends by assigning each user to a cluster. The interpretation of clustering results remains as an analysis task of the distributions of features by cluster, which will give an overall idea of what characterizes each cluster from the other clusters.

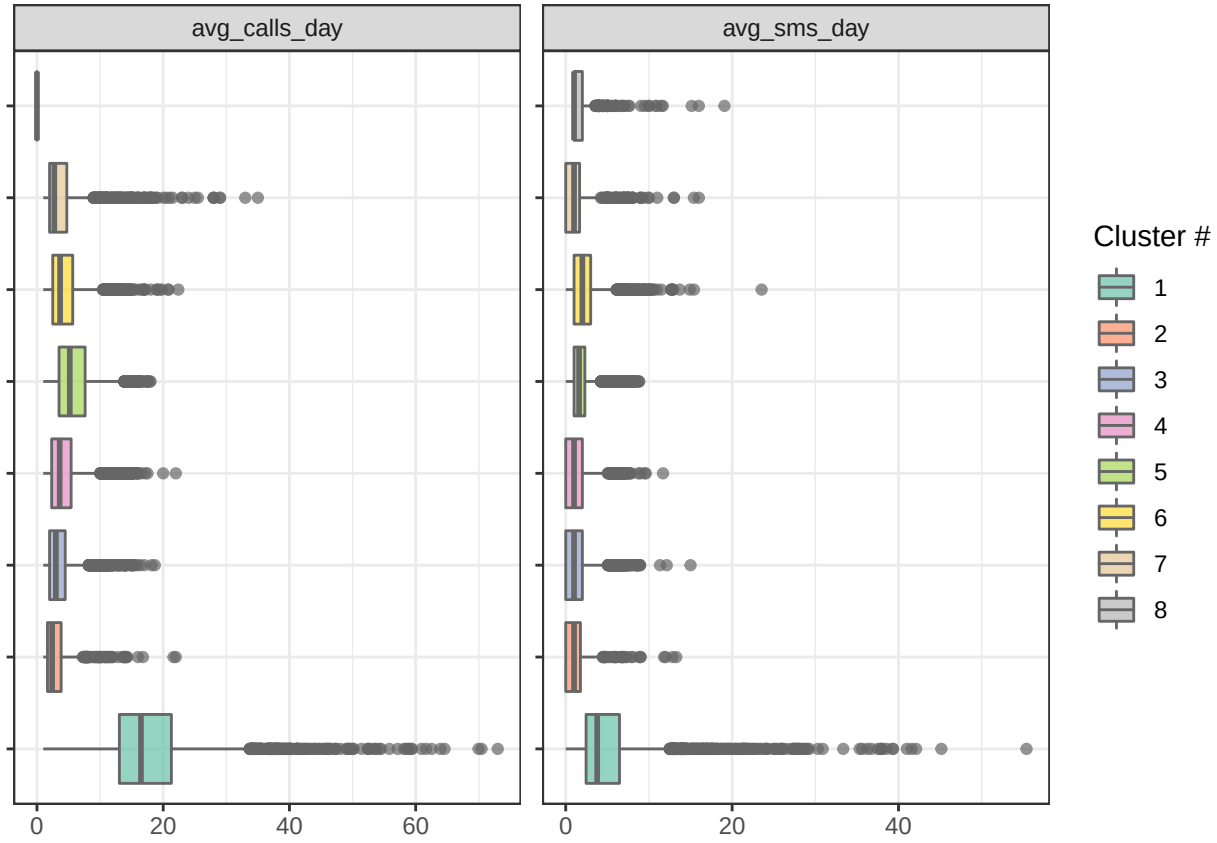
First, we start by analyzing the weight or the share of each cluster.

Share of each cluster



Next, we will go through the average number of calls/sms

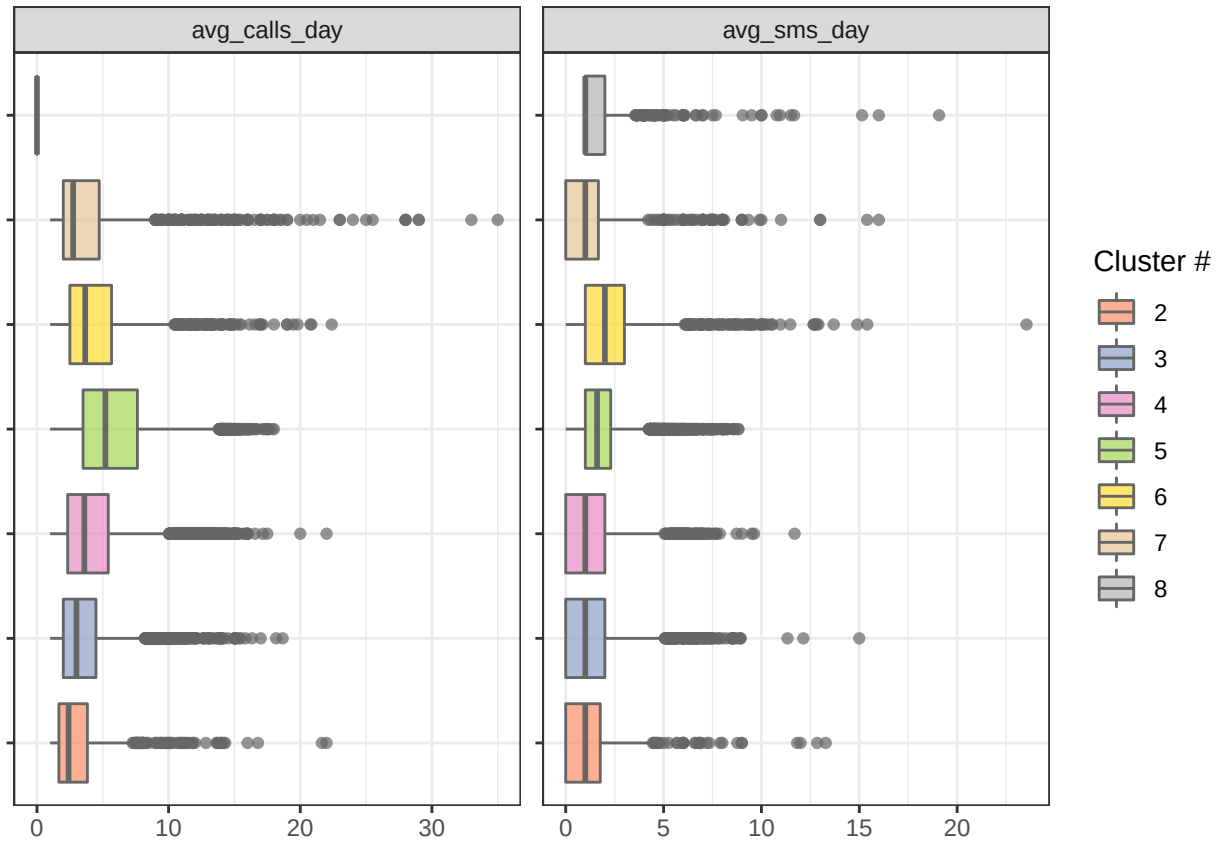
Average number of calls/sms



It seems clearly that the cluster num. 1 (which represents 8% of the total users in the data) regroups users that do higher number of calls per day (higher average number of sms as well) comparing to other users. The difference is pretty clear in the graph.

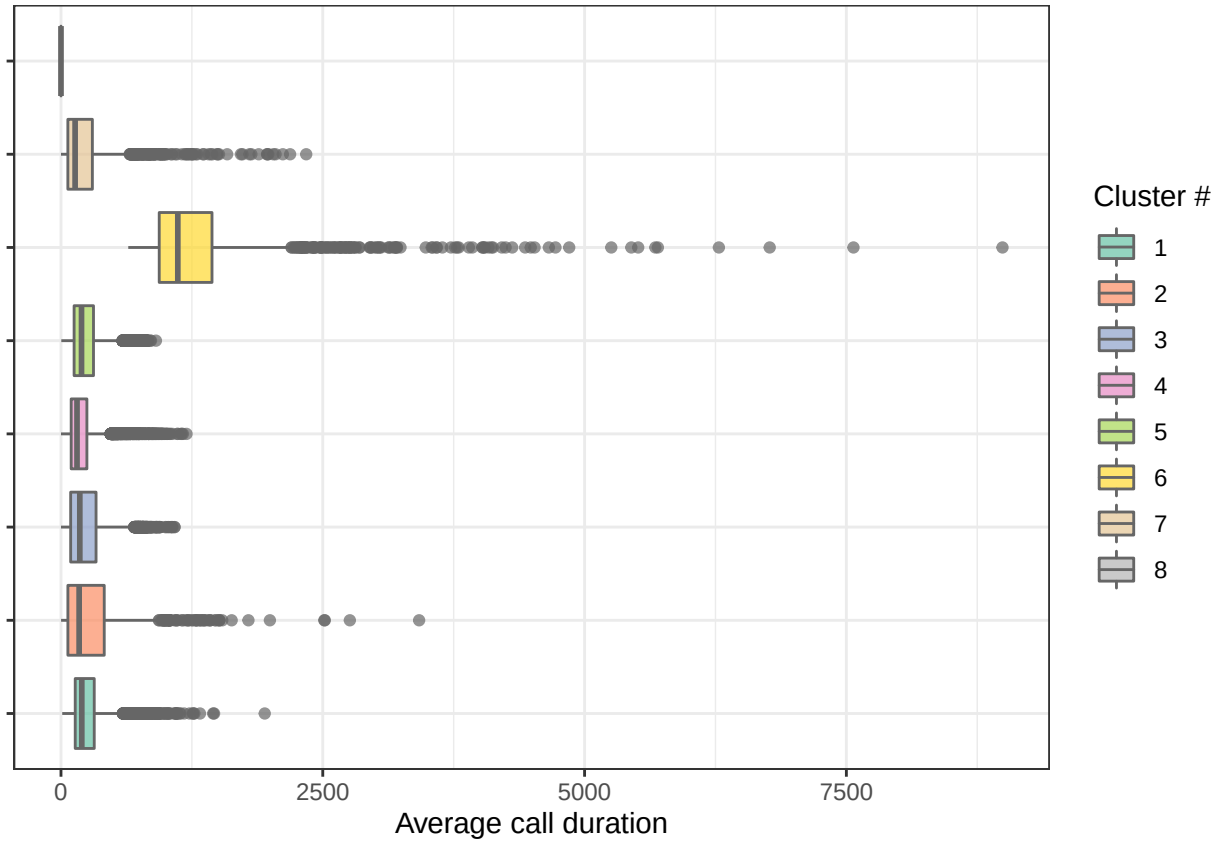
The cluster num. 8 regroups the users that do not use phone for calls at all. This segment of users represents 4% of all users.

All other six clusters seem to have almost similar distributions of the average number of calls per day. Let's remove the cluster num. 1 to improve the axis scale:

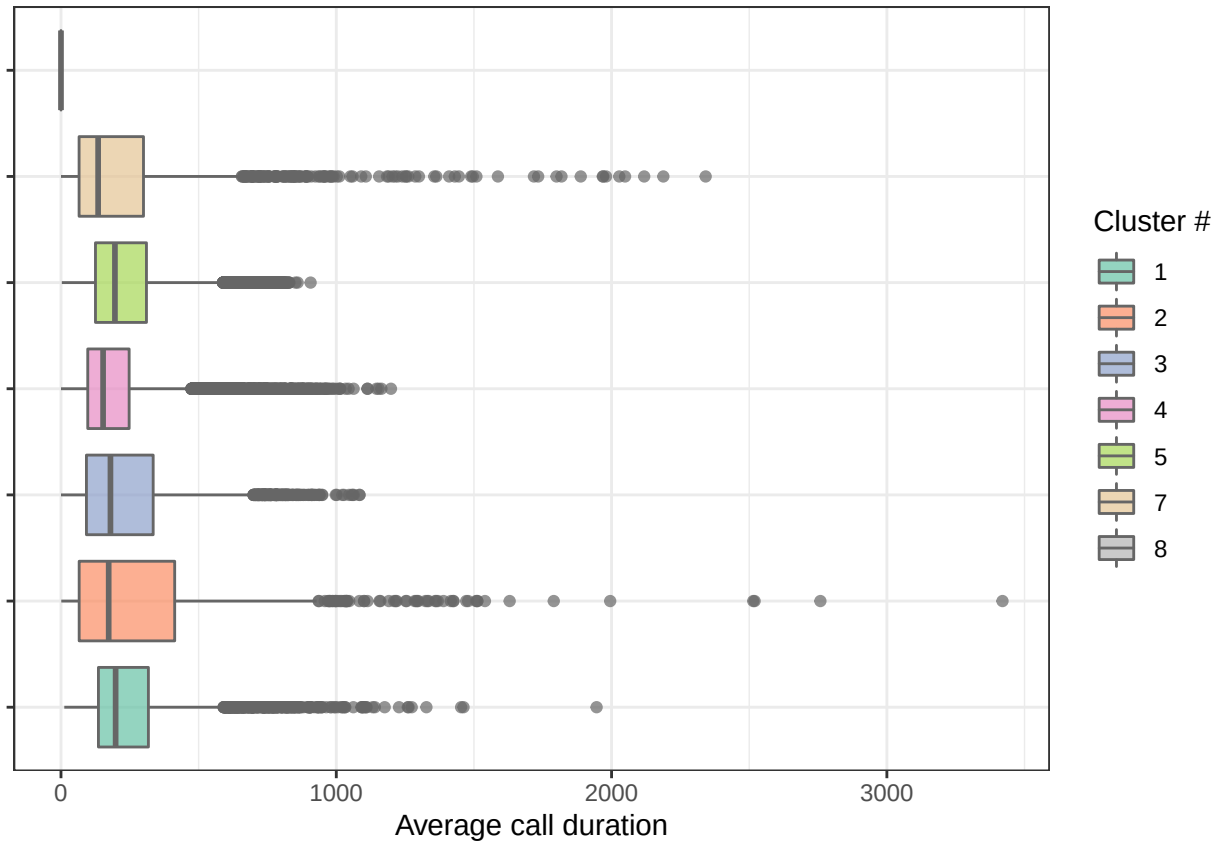


No significant difference is noted between clusters.

Average call duration



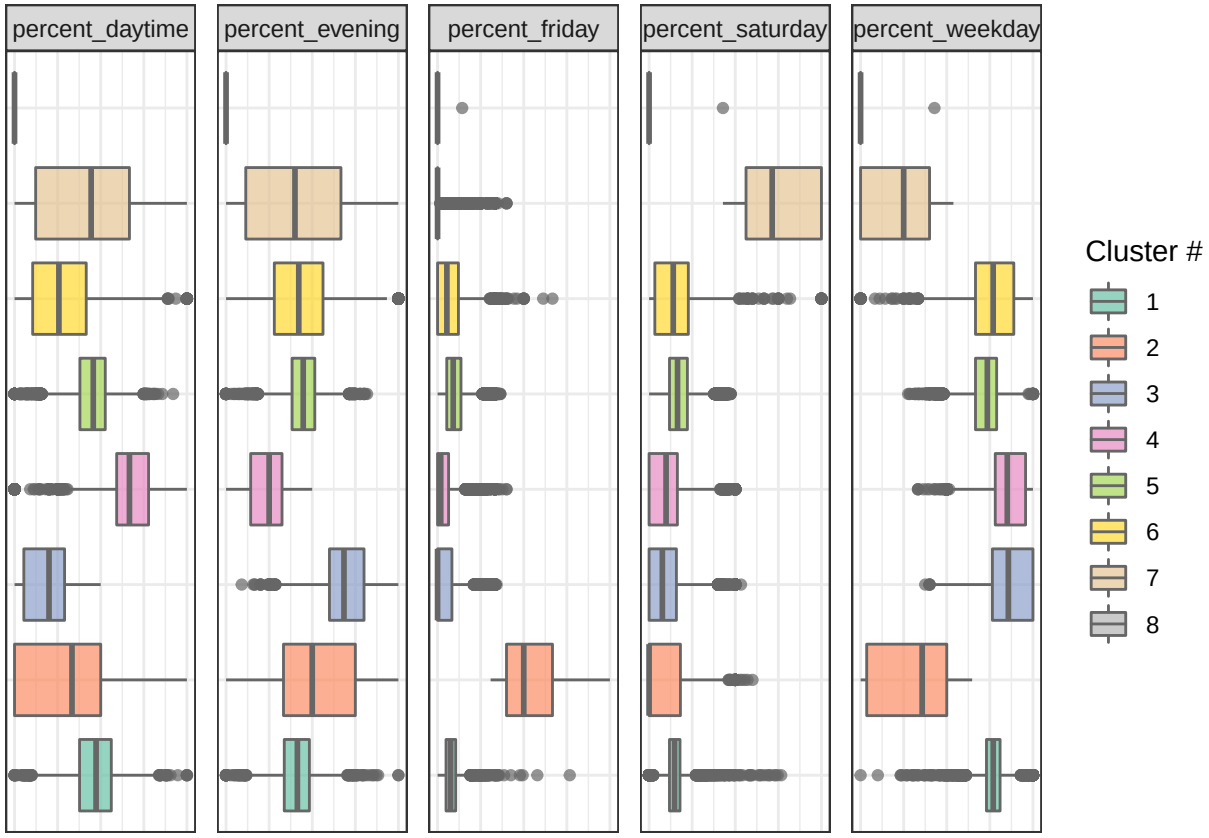
For the average call duration, the cluster num. 6 is characterized by higher call duration. 4% of users belongs to each cluster.



No significant difference is observed between other clusters.

With the three features examined before, only three clusters have specific distributions. Let's explore the attributes related to calling time (evening, morning, weekday, weekend) to see which distinguish other clusters.

Time-related features



Cluster num. 1 (8% of total users who make higher average number of calls per day) : This segment tends to make calls during the week (much more than on weekends) with no difference between day and evening.

Cluster num. 2 (only 2% of total users) : make less calls on Saturdays, pretty much calls in evenings than daytime.

Cluster num. 3 (14% of total users) : This segment makes calls very often during the week and rarely on weekends. More often in the evening than during the day.

Cluster num. 4 (24% of total users) : similar to cluster num. 3 in terms of weekend/weekday, but more calls in day than the evening.

Cluster num. 5 (40% of total users) : The segment with the highest share is characterized by the fact that it makes more calls during the week than on weekends, either during the day or in the evening.

Cluster num. 6 (4% of total users who make the longest calls) : They make most of their calls on weekdays, less often on weekend.

Cluster num. 7 (4% of total users) : Those users make rare calls on Fridays, but more often on Saturdays than weekdays.

Cluster num. 8 (4% of total users) : This special segment contains users who don't use phone for calls.

To sum up, the following table presents the mean of each feature for each cluster :

Cluster	share	avg call duration	avg calls day	avg sms day	percent daytime	percent evening	percent friday	percent saturday	percent weekday
Overall	100%	264.3	5.8	1.9	0.44	0.41	0.08	0.16	0.72
Segment 1	8.5%	259.5	17.9	5.4	0.47	0.41	0.08	0.15	0.77
Segment 2	2.2%	302.3	3.2	1.1	0.35	0.50	0.59	0.10	0.31
Segment 3	14%	238.8	3.5	1.3	0.19	0.72	0.05	0.10	0.86
Segment 4	23.6%	194.5	4.3	1.3	0.68	0.23	0.04	0.11	0.86
Segment 5	40.3%	239.0	5.8	1.8	0.45	0.45	0.10	0.17	0.72
Segment 6	3.8%	1330.2	4.6	2.3	0.28	0.43	0.08	0.15	0.76
Segment 7	3.9%	239.5	3.9	1.1	0.44	0.43	0.02	0.75	0.22
Segment 8	3.7%	0.0	0.0	1.6	0.00	0.00	0.00	0.00	0.00