Report milestone 1

Hou Yumeng, Alexis Jacq

November 2, 2015

1 Data description

This dataset describes the MOOC user behaviors during the first 3 weeks on 2 courses: C++ and JAVA. It has 26 columns/dimensions and 40028 rows, presenting the engagement information, grades, level of achievement, country and course info of 40028 students. Regarding to the engagement/activity parts, contents of this dataset could be described in the following way using a set: $S(\text{Activity}) = \{\text{AssignmentSubmission, AssignmentReSubmission, LectureReView, LectureView, ForumView, ThreadLaunch, PostonThread}\}$

2 Engagement index

We could define an engagement index as a sum of all these variables. But we could loose lot of information : if one of the variables has hight variance and another a small one, the variable with small variance will be neglected in the sum.

We can use the R function 'scale' to rescale all the variables: that way, all variable are centred, and variance=1. Now we can create variables by summing.

Our first index is the "*overall engagement*" : this is the sum of all the variables. Then, we define the "*passive engagement*" that is the sum (LectureReView + LectureView + ForumView) and the "*active engagement*" based on the sum (AssignmentSubmission, AssignmentReSubmission,ThreadLaunch, PostonThread).

We have to make sure that it makes sense to separate those two values. If they are correlated, we can just use the overall engagement as an unique index. The pearson test gives a small coefficient of correlation (0.39) with hight p-value to be between 0.36 and 0.41 : we are sure the variables have no linear connections. We can plot the active vs passive engagement data to visualize their coexistence.

We can also plot the distribution of these different variables using histograms (figure 2).

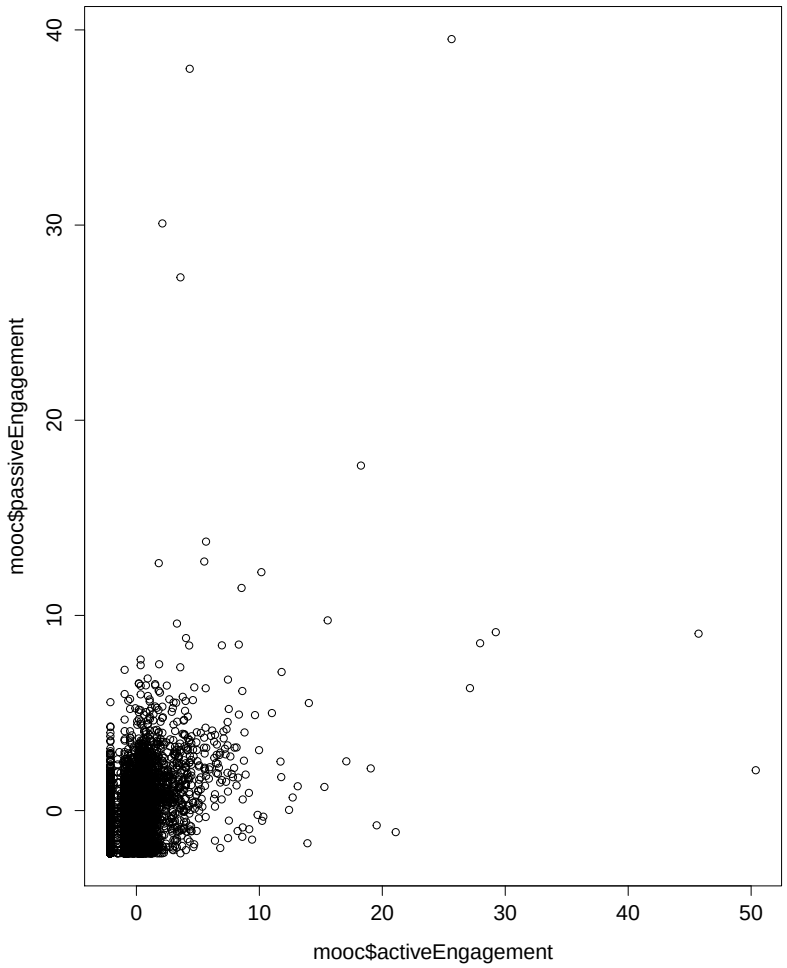

Figure 1: active engagement VS passive engagement: no correlation.

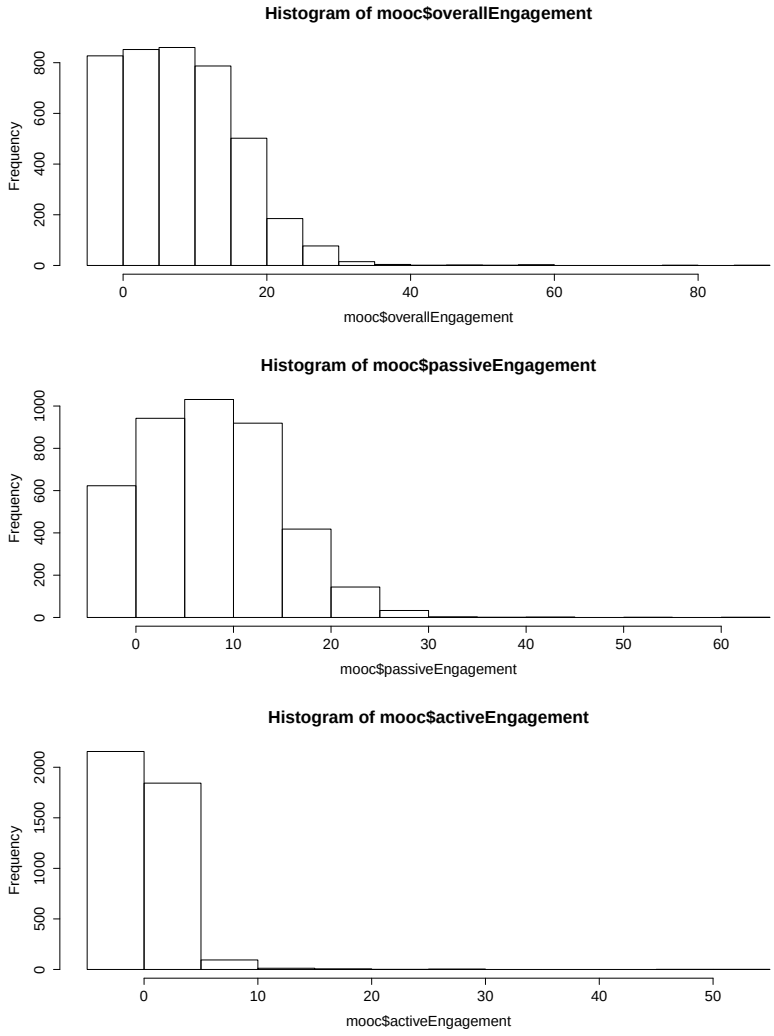

Figure 2: distribution of different indexes of engagement.

3 Comparing engagement in c++ and java courses

3.1 Method

(1) We use boxplot to show the intuitive distribution of engagement, by presenting the distribution of outliers (since most of the data is 0 in the first week). (2) We use line plot to show the overview of engagement statistically and compare engagement drop. (3) We can perform t-test to compare the mean of engagement in cpp and java.

3.2 Results

Based on the results shown in the following figures (especially explicit in the line chart), students registered the course of c++ and showed active engagement during the first week. But starting from the 2nd week, engagement of java becomes better than c++. During the first 3 weeks, it's similar within the two courses that the overall engagement as well as passive engagement went down in the 2nd week, then surge in the 3rd week. But interestingly, active engagement kept growing during that time (Figure 3). Regarding to engagement drop, overall, C++ suffered a higher engagement drop starting from the 2nd week (Figure 4).

We then perform t-test to compare the means of both weekly and global engagement in c++ and java. Unfortunately, the only significant difference is obtained when we compare the mean of active engagement in java and c++ during the week 3 (p-value = 0.031).

4 The most beneficial forum activity

4.1 Method

(1)We compare the coef of correlation between the value of final grades and the three variables that correspond to a forum activity to analyse their effect separately. (2)If no activity is highly correlated with the final grade, we assume they have a joint effect and built a linear model to explain final grade with all the possible activities. We use stepAIC from the package MASS to choose the variables of the model. Then we choose the forum activity with the highest effect.

4.2 Results

The coefficients of correlation with final grade given by forum activities was too small to conclude: forumView: 0.079 (p-value<0.05), threadLaunch: 0.052 (p-value<0.05, postOn: 0.037 (p-value>0.05).

After AIC steps, the linear model returned just explains variable with the forum views (p-value<0.023) and the other activities (p-value¡2e-16). We can check that this model is better than the initial full model by comparing RMSE between real grades and predictions : the RMSE ratio is 0.95. We conclude that the most effective forum activity is the forum view.

An interesting point: if we do not rescale the values before the linear regression, the most effective forum activity becomes the thread launch but no forum activities effects is longer significant. The stepAIC remove all the forum activities.

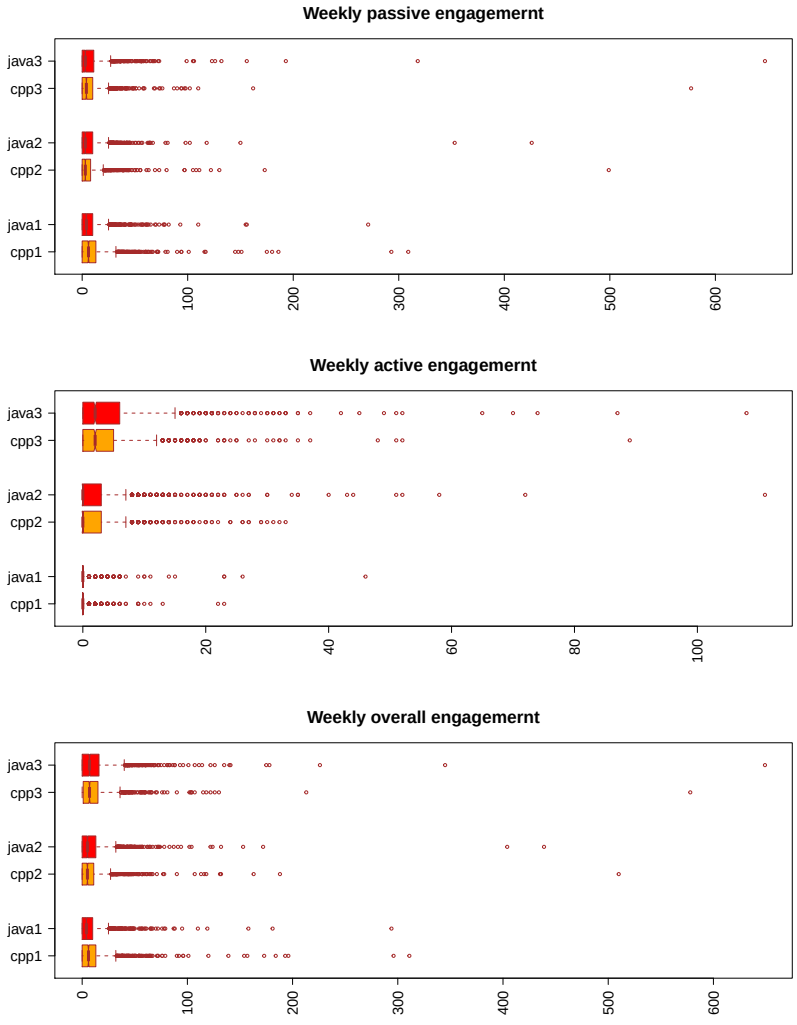

Figure 3: boxplot for weekly engagement.

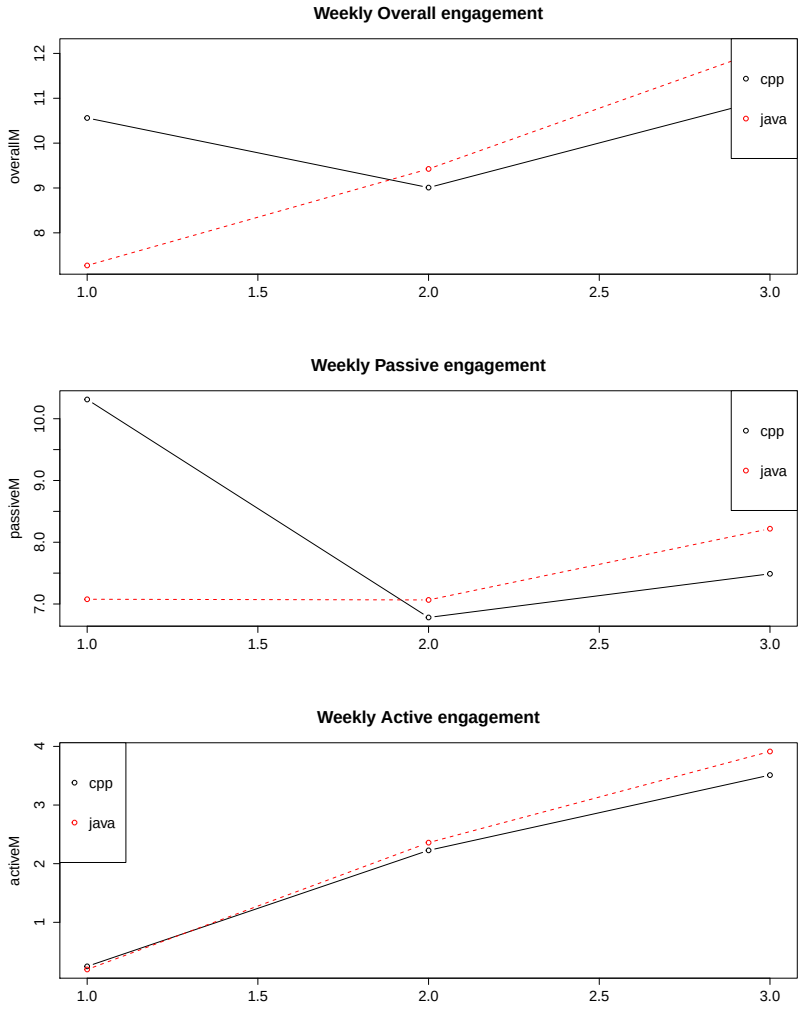

Figure 4: Line chart for weekly engagement.

5 Clustering student's behaviours

5.1 Methods

The first idea is to use our two indexes of engagement (passive and active) to reduce the dimension of data: we look at a 2D description of behaviours. We can also try to reduce the dimension via PCA (but we have seen that in our case PCA is not really helpful). An third idea is to look at three different variables: forum activities, lectures views and submissions. Then we have 3 variables. But we know that we want to study the behaviour looking their effect on the final grade. We can study which couple of two of these 3 variables is the best to separate different patterns that have visible consequences on the final grade. Then, we cluster in the 2D space given by this couple of variable.

5.2 Results

Using passive and active engagement, the result of Kmean with 4 clusters seems to be "forced" (figure 5). The different clusters cannot be interpreted. We can observe that use 4 cluster does not make sense with these variables by looking the "within cluster sum of square" as a function of the number of cluster. If passing from 3 to 4 is visible improvement, it make sense, otherwise it does not. As we can see on figure 6, 4 clusters is not better than 2,3 or 5.

Now we take a look to normal grade and the three kind of activities (submissions, forum and lectures). We separate the data in two groups : students with a final grade > 90 and student with final grade < 90. Then, we plot the clouds given by the 3 possible couples of variables and we assign a red color to the point corresponding to hight grades and a blue color to the others (figure 7). We can see that with the couple (forum vs submissions), two directions seems to separate the hight grades and others (low forum, hight submissions grade<90, and hight forum, low submissions grade>90). We chose those variables to find clusters. The figure 8 shows that the "within cluster sum of square" is improved by 4 cluster much more than with 5 and 4 clusters seems to be optimal. We finally plot those clusters (figure 9). The cluster in black contain no hight grades. It can be interpreted as the completely disengaged students. The blue one is the students that did not visit the forum at all. The frequency of hight grades in this cluster is small. The green cluster represent can be interpreted as the "engaged students" that both used forums and submitted assignments. It contains a little bit more hight grades than lower grades. And finally, the red one represent the student that used extensively the forum. Almost all of them received hight final grades.

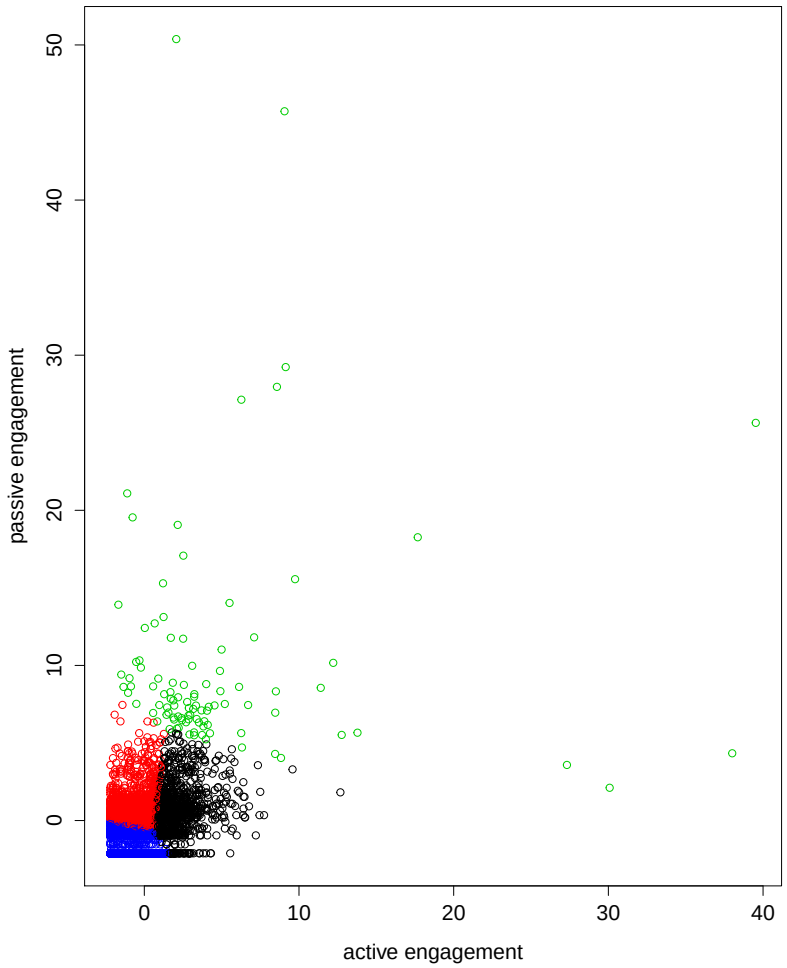

Figure 5: Result of kmeans in passive/active engagement space, k=4.

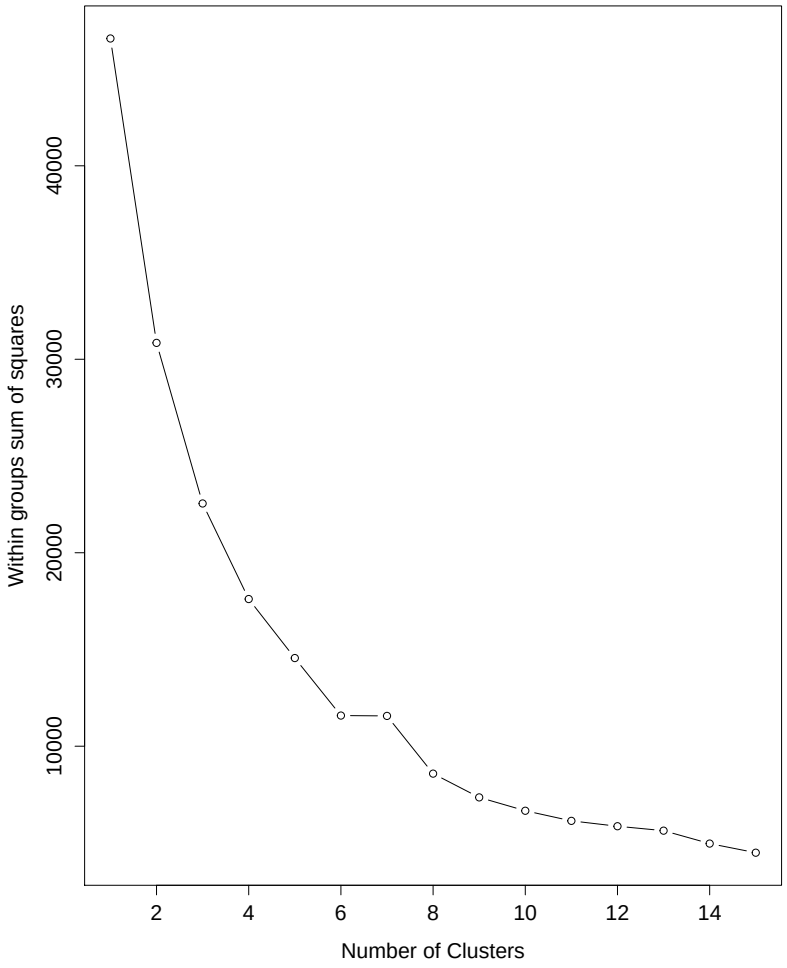

Figure 6: within-cluster sum of square as a function of the number of cluster in passive/active engagement space.

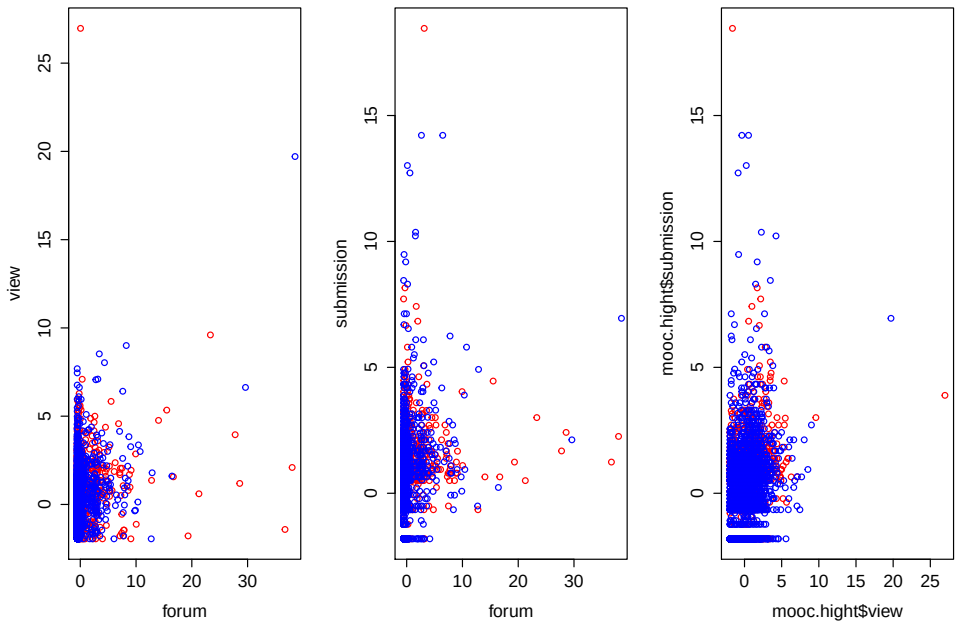

Figure 7: repartition of hight grades (red) and other grades (blue) in the 3 possible 2D space obtained with the variables (submissions, forum and lectures).

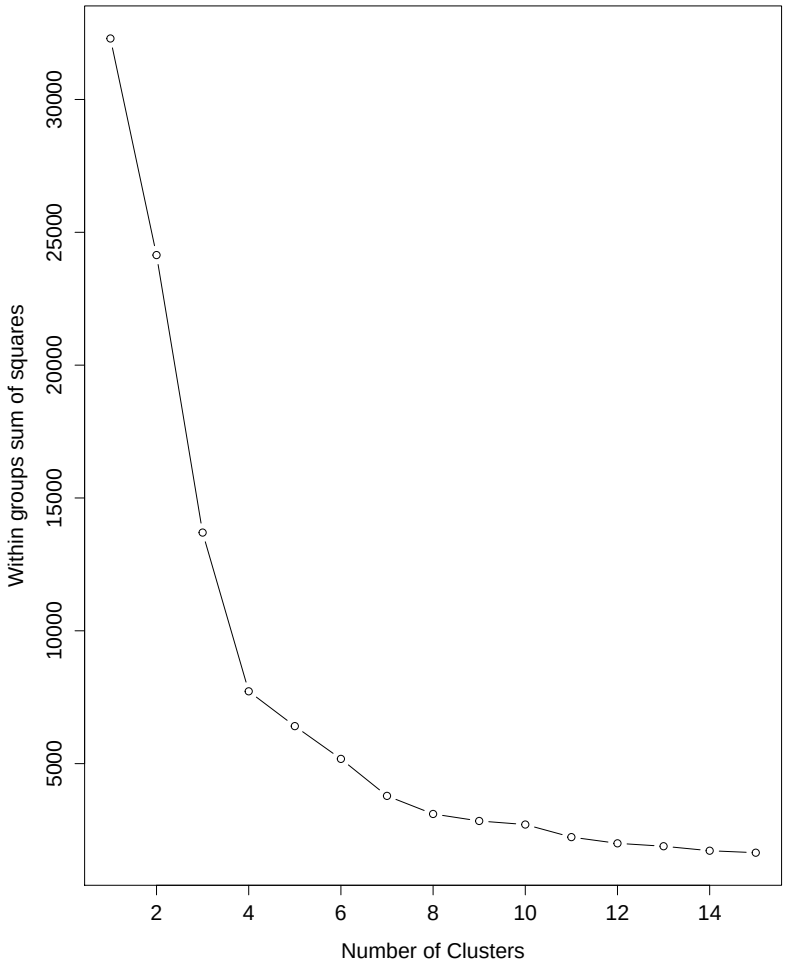

Figure 8: within-cluster sum of square as a function of the number of cluster in passive/active engagement space.

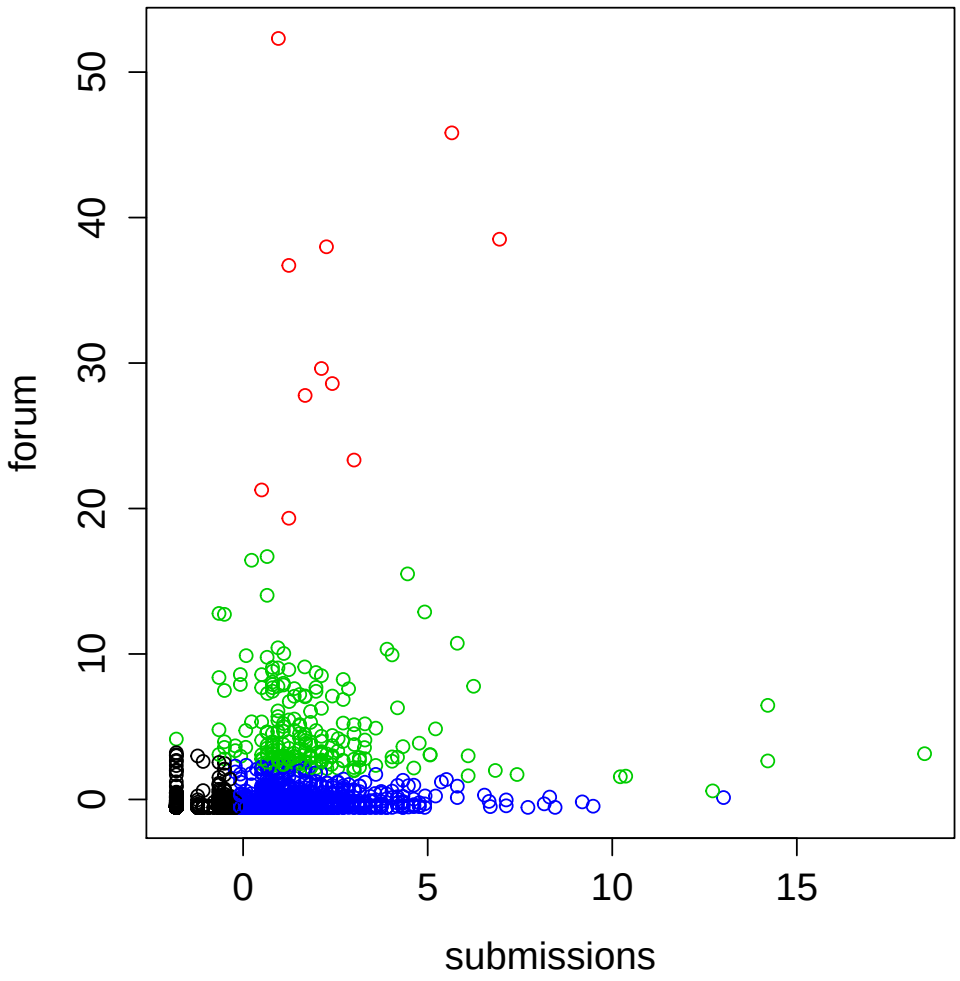

Figure 9: Result of kmeans in passive/active engagement space, k=4.