Data Organization

* We deleted data whose value of *inconsistent\_flag* is 1. The value of “1” means there’s something wrong with the Tracking Logs for a class or a student.
* We deleted the variables with named *LoE\_DI*, *YoB* and *gender* which are N/A.
* We deleted dataset of course code CB22x, 8.MReV, CS50x, ER22x, PH207x, PH278x which has no video events for the course.
* We calculated the interval between the date for their first time event (*start\_time\_DI*) and the date for the last time event (*last\_event\_DI*),  labeled as *days\_involed.*
* We calculated student’s average events everyday by using the number of students’ interactions with the course (*nevents*) and the *days\_involed*, labeled as *events\_per\_day*.
* We calculated student’s active rate in the course using number of unique days student interacted with the course ( *nday\_act* )and *days\_involed,* labeled as *rate\_active\_days.*
* We calculated the ratio between  the number of video played by each student (*nplay\_video*) and the number of video uploaded by instructors, labeled as *rate\_play\_video.*
* We calculate the ratio between the number of chapters viewed by each student (*nchapters*) and the number of chapters designed by instructors, labeled as *rate\_chapter.*

Data Description

Based on the accessible data set, we changed our purpose to discover how students’ engagement to an online course predicts their grade. The engagement is defined and modeled by released data from the Edx.

Dependent variable: grade

Independent variables (features)-the degree of engagement:

* *viewed*: administrative, 0/1; anyone who accessed the ‘Courseware’ tab (the home of the videos, problem sets, and exams) within the edX platform for the course. Note that there exist course materials outside of the ‘Courseware’ tab, such as the Syllabus or the Discussion forums.
* *explored*: administrative, 0/1; anyone who accessed at least half of the chapters in the courseware (chapters are the highest level on the “courseware” menu housing course content).
* *certified*: administrative, 0/1; anyone who earned a certificate. Certificates are based on course grades, and depending on the course, the cutoff for a certificate varies from 50% = 80%.
* *final\_cc\_cname\_DI*: mix of administrative (computed from IP address) and user= provided (filled in from student address if available when IP was indeterminate); during de=identification, some country names were replaced with the corresponding continent/region name. Examples: “Other South Asia” or “Russian Federation”.
* *nforum\_posts*: administrative, number of posts to the Discussion Forum. Example: “8”.
* *days\_involved*: the days that students involved in this course. It is presented by the interval between the date for their first time event (*start\_time\_DI)* and the date for the last time event (*last\_event\_DI).*
* *events\_per\_day*: the number of events that emerged by each student per day. It is calculated by dividing the number of students’ interactions with the course (*nevents)*  by the *days\_involed*
* *rate\_active\_days*: the number of active days (*ndays\_act*) v.s. *days\_involved*.
* *rate\_play\_videos*: the number of video played by each student v.s. the number of video uploaded by instructors.
* *rate\_chapters*: the number of chapters viewed by each student v.s. the number of chapters designed by instructors.

Report for Discussion

We (Shan & Tiantian) both brought the following questions to assigned partners for this week and discussed with them.

1. We don’t have access to all courses, since some of them have closed enrollment or are not open anymore. How should we deal this that?
2. Our original plan is to compare features at the course level. However, we found that if we do that, we only have 3 available courses which means we only have 3 sample to compare. That is obviously not sufficient.
3. If we change the plan and to conduct analysis at the student level, what meaningful thing we could do ?

Bringing with suggestions from each assigned partner, we two get together and discussed the problems and possible solutions. We made the final decision about how to solve previous questions.

1. Since the data on the course is very limited we decided to give up the original plan and conducted analysis at the student level--discovering the relationship between students’ engagement and grades.
2. In order to narrow down our topic, we decided to only keep courses that provide videos and not consider those without videos.
3. To better define students’ engagement, we generate new variables, such as events\_per\_day, rate\_play\_videos, etc.
4. During discussion, we also found some useful literatures that provide valuable information.

Another question we concurred is the current stage we generated the *rate\_active\_days*  column, but there are students who only involved in course for 1 day and have 1 event. For those cases, students’ active rate is 1, which makes them seem to be very “active”. To deal with this problem, we are thinking to set certain cut-off point to better grouping student who are actually taking the course or at least have intention to get involved in the course . For example, if the value of *days\_involved* is under 5, then no matter what active rate this student has he/she will be considered as not actually taking this course. But we are not sure whether this method is appropriate and how to set the cut-off point. This is the new question we need to solve in future.