MySQL – Project 3

Northeastern University – ALY 6030

**edX Database**

Full Name \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Overview**

In this project, you will query your database from project 2 to learn about the dataset. Recall, this database contains information about users who took courses taught on the edX platform.

**Recall the attribute definitions:**

|  |  |
| --- | --- |
| Column Name | Description |
| course\_id | three-part identifier for a course number |
| Course Short Title | Short title for the course |
| Course Long Title | Long title for the course |
| userid\_DI | Individual user ID |
| registered | Whether the user is registered (1/0) |
| viewed | Whether the user has viewed the contents (1/0) |
| explored | Whether the user has explored the course (1/0) |
| certified | Whether the user is certified (1/0) |
| Country | User’s country of origin |
| LoE\_DI | User’s level of education |
| YoB | User’s year of birth |
| Age | User’s age |
| gender | User’s gender |
| grade | User’s grade in the course |
| nevents | Number of events the user has done on the site |
| ndays\_act | Number of actions taken by the user |
| nplay\_video | Number of video plays done by the user |
| nchapters | Number of chapters read by the user |
| nforum\_posts | Number of forum posts made by the user |
| roles | Any roles the user has |
| incomplete\_flag | Whether the user has an incomplete for the course |

Now that our database is set up, let's start to ask some questions about our data! From now on we can use the three tables in our data model rather than going back to the original "flat table." If we wanted to, we could even drop that table from our database.

**For each question, enter your respective MySQL queries needed to obtain the required result. Furthermore if the output is a single numerical value (e.g., average, sum, …), include it with your answer.**

**Part 1: An exploration of 6.00x**

1. How many enrollees are in the class with course\_id MITx/6.00x/2012\_Fall?

Assumed enrolled is the same as "registered."

1. That same course was also taught again the following term with course\_id "MITx/6.00x/2013\_Spring". Did enrollment go up or down?
2. What was the average grade in each term of 6.00x?
3. Uh, those seem too low. Remove all of the zeroes for people who haven't taken any tests, and calculate the average from the remaining data for both terms of 6.00x?

1. What was the total number of enrollments for 6.00x over both terms?
2. How many people were enrolled (registered) in both terms of 6.00x?

**Part 2: A look at all courses**

1. Let's take this one step further, let's list the number of enrollees in every course.
2. Make the returned result nicer by changing the header from COUNT(\*) to enrollees.
3. Order the previous results to show the number of enrollees in descending order. Which class had the most enrollees?

**Part 3: A deeper look at all of the courses**

***FIRST JOIN***

1. Create the same list of enrollees in descending order, but include only the course long title and the counts of enrollees in the returned result. Note this will group together courses with the same name, which were taught in different terms.
2. Now we need to find out which are the courses that had at least 4,000 Enrollees? You may do this by course id.

**Part 4: Course engagement**

1. Now let's put together some statistics for the engagement for each course. How many people are registered, have viewed, have explored, and have become certified for each course by course\_id? Rename the columns.

1. Challenge (Bonus) : Redo the above question using course long title instead of course\_id. Note this will group together courses with the same name, which were taught in different terms.
2. What fraction of the users view, explore, or certify in the content in each course once they have registered? Rename the columns.
3. For a second challenge (Bonus), do this again with the long course title, noting that this will group courses with the same name across terms.

**Part 5: More challenging questions**

1. Find the list of courses that are hosted at HarvardX and have more than 4000 enrollees?

***Hint/Challenge – try two different approaches:***

***1. JOIN: Use the institution field in Courses, and JOIN Courses with Course\_Users***

***2. NO JOIN: Try using LIKE with some wildcards***

1. How many users who have registered more than 3 courses?
2. How many users are there by country? Order these alphabetically by country.
3. What is the average grade by country for users who have become certified in any course? Order by average grade, descending.

***Hint: JOIN the Course\_Users and Users table together, and use the Country field***

1. Which country has the highest average grade for people certified in a course? Which country has the lowest average grade for people certified in a course?

**Part 6: Harvard and MIT, course grades**

1. What is the average grade by country for users who have become certified in any ***HARVARD*** course? Order by average grade, descending.

***Hint: TRIPLE JOIN!***

1. Which country has the highest average grade for people certified in a HARVARD course? Which country has the lowest average grade for people certified in a HARVARD course?
2. Repeat the above query for MIT.