

## Data Analytics and Insights Take-home Exercise

**Due date:** 3 business days after receiving this exercise. We recommend spending about 2 - 3 hours on this exercise.

Course Hero is an online learning platform where members can access over millions of course-specific study resources contributed by a community of students and educators. The platform includes practice problems, study guides, videos, class notes, and step-by-step explanations to help with every college class - from economics to literature, biology to history, accounting to psychology, and everything in between.

As part of the corporate goal, content growth, especially quality content in the form of documents is of great importance. Our 2020 goal for document acquisition is to grow the cumulative count as well as proportion of quality documents in our content library. The current content growth rate is high enough which allows us to invest in long-term strategic improvements including document quality scoring and content tagging. Ultimately the goal is to be able to accept and deny documents based on content quality and value.

Data Analytics and Insights team is interested in supporting Product Management in forming the document quality improvement strategy analytically. To help answer the question, there are two sample datasets, one of which contains document attributes such as school and subject; the other one includes document performance in the form of paid unlocks. We currently define a high quality document if it has more than 3 paid unlocks (>=3) within 24 months of upload to the website where a paid unlock is recorded when customers pay to view the document. We would like you to use the sample datasets to help recommend an analytics plan to help increase growth of quality documents, and offer suggestions to operationalize those insights to help Course Hero.

**Instructions:**

This exercise is intended to assess your analytical thinking and storytelling skills through data analysis and data science solutions. The focus should be on the structure of your findings and analytics plan.  
Please analyze the data linked [**here**](https://drive.google.com/drive/u/0/folders/1IWyyRtjtN-jgC803kdjzKy4B0NBMgak1) and provide a report with your findings along with any code you used (language of your choosing: SQL, R, Python, etc). We recommend spending around **2 - 3 hours** on this exercise. See below for a detailed description of the datasets.

1. *Define the problem* - What is the document quality status today based on the datasets? For example, what fraction of the documents are of high quality per current definition?
2. *Understand the data* - Perform any EDA on the data set to derive insights, identify opportunities and etc.
3. *Discuss your analytics plan/steps* - How would you build a prediction model for high quality documents? Actual model development is not required.
4. *Bonus question* - Discuss how Course Hero might leverage the insights gained from the analytics deliverable to improve?

Dataset ‘DS\_interview\_doc’:

|  |  |
| --- | --- |
| **Column** | **Description** |
| New\_ID | Unique identifier for each document |
| date | Date of document upload |
| year\_month | Year and month of document upload |
| byte\_size | Byte size of the document |
| filetype | File type, i.e. pdf, docx, and etc. |
| school\_country | Country of the school associated with the document |
| school\_name | Name of the school associated with the document |
| school\_type | The type of school the document is associated with, i.e. ‘High’ for high school and ‘Higher’ for universities, college and such. |
| enrollment | Enrollment of the school associated with the document. (Number of students). |
| first\_subject\_name | The subject linked to a document,i.e, history |
| tag | Tag associated to a document, i.e, paper, notes, and etc. |
| page\_count | Count of pages in the document |
| hqd\_score | A score of the document readability via a machine learning model, the higher the more readable. |
| is\_mcq | 1 if the document contains multiple choice questions. Null otherwise. |
| top\_keywords | Top keywords extracted from the documents |
| language | Document language |
| course\_id | Course id associated if applicable |
| user\_id | Document uploaders’ user id |

Dataset ‘DS\_interview\_paid\_unlock’:

|  |  |
| --- | --- |
| **Column** | **Description** |
| New\_ID | Unique identifier for each document |
| unlock\_date | Date of document getting paid unlocks |
| unlock\_count | Count of paid unlocks on that day |