**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**A black screen with a black border

Description automatically generated

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC MÁY TÍNH**

A blue logo with a black background

Description automatically generated with low confidence

**BÀI TẬP PHÂN TÍCH DỮ LIỆU**

**BÁO CÁO PHÂN TÍCH BỘ DỮ LIỆU**

**MÔN HỌC: KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG (CS313)**

**Nhóm 2**

**GVHD**

ThS. Nguyễn Anh Thư

**TP. HO CHI MINH, 3/2025**

**DANH SÁCH THÀNH VIÊN**

| **STT** | **Họ và tên** | **MSSV** |
| --- | --- | --- |
| 1 | **Võ Đình Trung** | **22521571** |
| 2 | Trương Phúc Trường | 22521587 |
| 3 | Mai Dương | 22520302 |
| 4 | Trần Qui Linh | 22520779 |
| 5 | Triệu Tấn Huy | 22520581 |
| 6 | Trần Nguyễn Anh Phong | 22521092 |
| 7 | Nguyễn Tấn Lợi | 22520801 |

**LỜI CẢM ƠN**

Chúng tôi xin chân thành cảm ơn cô giáo, ThS. Nguyễn Thị Anh Thư, người đã định hướng, giúp đỡ, trực tiếp hướng dẫn và tận tình chỉ bảo chúng tôi trong suốt quá trình nghiên cứu, xây dựng và hoàn đồ án này.

Chúng tôi cũng xin được cảm ơn tới gia đình, những người thân và bạn bè thường xuyên quan tâm, động viên, chia sẻ kinh nghiệm, cung cấp các tài liệu hữu ích trong thời gian học tập, nghiên cứu cũng như trong suốt quá trình thực hiện đồ án.

TP. HCM, ngày 18 tháng 3 năm 2025

**ĐÁNH GIÁ CỦA GIẢNG VIÊN HƯỚNG DẪN**

………………………………………………………………………………………...

………………………………………………………………………………………...………………………………………………………………………………………...………………………………………………………………………………………...………………………………………………………………………………………...………………………………………………………………………………………...………………………………………………………………………………………...………………………………………………………………………………………...………………………………………………………………………………………...………………………………………………………………………………………...………………………………………………………………………………………...………………………………………………………………………………………...………………………………………………………………………………………...………………………………………………………………………………………...………………………………………………………………………………………...………………………………………………………………………………………...………………………………………………………………………………………...………………………………………………………………………………………...………………………………………………………………………………………...………………………………………………………………………………………...………………………………………………………………………………………...…………………………………………………………………………………………

TP. HCM, ngày 3 tháng 3 năm 2025

**GVHD**

**Nguyễn Anh Thư**

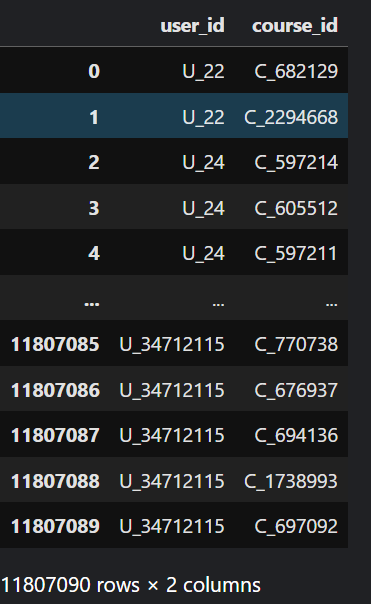
# **Tìm hiểu và tiến hành hợp nhất dữ liệu:**

1. **Nguồn dữ liệu ban đầu (~11.8 triệu dòng):**

- Đọc file user.json.

- Trích xuất tất cả các cặp (user\_id, course\_id) từ trường course\_order của mỗi người dùng. Đây sẽ là 2 id chính cho dataframe hợp nhất

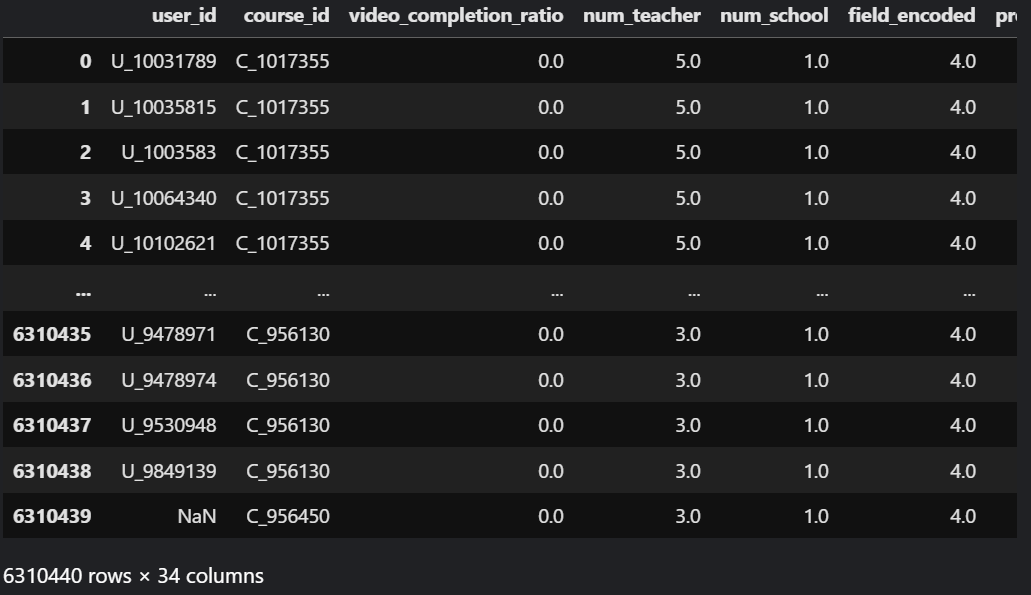
- DataFrame được tạo ra chứa 11,807,090 dòng

****

1. **Tiêu chí loại bỏ chính dẫn đến giảm số dòng:**

* Loại bỏ các dòng liên quan đến người dùng tham gia cực nhiều khóa học: Tất cả các dòng dữ liệu (user\_id, course\_id) thuộc về những người dùng đã đăng ký nhiều hơn 20 khóa học đã bị loại bỏ. Ảnh hưởng đến khoảng 5,288,059 dòng (chiếm 44.79% dữ liệu gốc). Đây là nguyên nhân chính gây ra sự giảm sút lớn.
* Loại bỏ các dòng liên quan đến khóa học rất ít người đăng ký: Tất cả các dòng dữ liệu thuộc về những khóa học có ít hơn 2 người dùng đăng ký đã bị loại bỏ. Điều này ảnh hưởng đến khoảng 907 dòng (chiếm 0.01% dữ liệu gốc).
* Sau khi áp dụng các tiêu chí lọc cải tiến trên, số dòng dữ liệu giảm từ 11,807,090 xuống còn 6,310,440 dòng (giữ lại 55.21%).

DataFrame đã lọc này đã lọc ra các user\_id, course\_id đủ quan trọng để đánh giá, tiếp tục tiến hành các bước hợp nhất và xử lí dữ liệu tiếp theo.



1. **Cách hợp nhất và tạo, xử lí các đặc trưng dữ liệu để tạo 1 dataframe input duy nhất ở trên:**
2. **Cách tạo output:**

- Tính Tỷ lệ Hoàn thành Video (video\_completion\_ratio):

* Đo lường mức độ người dùng đã xem video trong khóa học.
* Công thức cơ bản (hoặc ý tưởng): (Tổng số video đã xem của user) / (Tổng số video mặc định của khóa học)
* Tính toán điều này bằng cách dùng dữ liệu từ user\_video\_interaction.csv
* Giá trị này thường được giới hạn ở mức tối đa là 1 (hoặc 100%).

- Tính Tỷ lệ Hoàn thành Bài tập/Vấn đề (problem\_ratio):

* Đo lường mức độ người dùng đã tương tác hoặc hoàn thành các bài tập/vấn đề.
* Tính dựa trên: min(1, Tổng số bài tập đã làm của user / Tổng số bài tập trong khóa học).
* Giá trị này cũng thường được giới hạn ở mức tối đa là 1.
* Xác định Trọng số (Alpha - α):
* Đây là yếu tố quan trọng để cân bằng giữa video và bài tập. Trọng số này phụ thuộc vào cấu trúc của khóa học.

- Ý tưởng: Một khóa học có nhiều bài tập và ít video nên đặt trọng số cao hơn cho việc hoàn thành bài tập, và ngược lại.

* Tính tỷ lệ: ratio = num\_problems / (num\_videos + epsilon) (epsilon để tránh chia cho 0).

Chuẩn hóa tỷ lệ về khoảng [0, 1): alpha = ratio / (1 + ratio).

Xử lý trường hợp đặc biệt:

Nếu chỉ có bài tập (num\_videos == 0), alpha = 1.

Nếu chỉ có video (num\_problems == 0), alpha = 0.

Nếu không có cả hai, alpha = 0.5 (trung lập).

* Tính Điểm Hoàn thành Tổng hợp (composite\_completion):

Đây là trung bình có trọng số của hai tỷ lệ hoàn thành đã tính ở trên.

Công thức chung:

composite\_completion = (alpha \* problem\_completion\_ratio) + ((1 - alpha) \* video\_completion\_ratio)

- Trong đó:

* alpha là trọng số cho việc hoàn thành bài tập.
* (1 - alpha) là trọng số cho việc hoàn thành video.

1. **Khám phá dữ liệu sau khi hợp nhất**
   1. **Thống Kê Mô Tả**

* Ý nghĩa các cột dữ liệu:
* user\_id: Mã định danh duy nhất của người dùng/học viên
* course\_id: Mã định danh duy nhất của khóa học
* video\_completion\_ratio: Tỷ lệ hoàn thành video của học viên trong khóa học
* num\_teacher: Số lượng giáo viên giảng dạy khóa học
* num\_school: Số lượng trường học liên kết với khóa học
* field\_encoded: Mã hóa lĩnh vực/chủ đề của khóa học
* prerequisites\_encoded: Mã hóa các điều kiện tiên quyết của khóa học
* num\_videos: Tổng số video trong khóa học
* num\_exercises: Tổng số bài tập trong khóa học
* num\_problems: Tổng số bài toán/câu hỏi trong khóa học
* num\_students: Tổng số học viên đăng ký khóa học
* total\_default\_video\_time: Tổng thời lượng mặc định của tất cả video trong 1 khóa học
* total\_comments: Tổng số bình luận trong khóa học
* total\_replies: Tổng số phản hồi trong khóa học
* avg\_comments\_per\_student: Số bình luận trung bình mỗi học viên của mỗi khóa học
* avg\_replies\_per\_student: Số phản hồi trung bình mỗi học viên của mỗi khóa học

total\_problem\_attempts: Tổng số lần thử giải bài tập của mỗi khóa học

* avg\_problem\_attempts\_per\_student: Số lần thử giải bài tập trung bình mỗi học viên
* total\_completion\_rate: Tổng tỷ lệ hoàn thành của khóa học của mỗi khóa học
* avg\_completion\_rate: Tỷ lệ hoàn thành trung bình của khóa học của mỗi khóa học
* total\_video\_watch\_time: Tổng thời gian xem video của tất cả học viên của mỗi khóa học
* avg\_video\_watch\_time\_per\_student: Thời gian xem video trung bình mỗi học viên
* iscorrect\_ratio: Tỷ lệ trả lời đúng của học viên
* attempts\_ratio: Tỷ lệ số lần thử làm bài
* score\_ratio: Tỷ lệ điểm số đạt được
* lang\_ratio: Tỷ lệ liên quan đến ngôn ngữ (có thể là tỷ lệ sử dụng ngôn ngữ cụ thể)
* option\_ratio: Tỷ lệ các tùy chọn (có thể liên quan đến câu hỏi trắc nghiệm)
* type\_ratio: Tỷ lệ theo loại (có thể liên quan đến loại bài tập/câu hỏi)
* problem\_ratio: Tỷ lệ hoàn thành bài tập/bài toán
* total\_video\_watch\_time\_1: Phiên bản thay thế của tổng thời gian xem video
* avg\_video\_watch\_time\_1: Phiên bản thay thế của thời gian xem video trung bình
* video\_watched: Số lượng video đã được xem bởi học viên
* alpha: Hệ số alpha dùng trong công thức tính mức độ hoàn thành tổng hợp
* **composite\_completion**: Chỉ số hoàn thành tổng hợp (kết hợp video và bài tập). Đây cũng là output của bài toán

Các cột này chứa thông tin về người dùng, khóa học, và mức độ tương tác của người dùng với nội dung khóa học, bao gồm xem video và làm bài tập.

* 1. **Phân Tích Thống Kê**

**2.2.1. Kiểm định T-test**

Tiến hành kiểm định t-test trên 2 trường là video\_completion\_ratio và composite\_completion. Ta giả định so sánh những record có video\_completion\_ratio = 0 và những record có video\_completion\_ratio khác 0 để kiểm định giả thuyết: “Trung bình của composite\_completion giữa hai nhóm video\_completion\_ratio == 0 và video\_completion\_ratio != 0 là bằng nhau”



* Ý nghĩa thống kê: Với p-value cực kỳ nhỏ (< 0.05). Điều này có nghĩa là có sự khác biệt đáng kể về trung bình composite\_completion giữa hai nhóm video\_completion\_ratio == 0 và video\_completion\_ratio != 0.
* Hướng của sự khác biệt: Vì statistic là âm (15.167403154980034), điều này cho thấy trung bình composite\_completion của nhóm video\_completion\_ratio == 0 (group1) thấp hơn so với nhóm video\_completion\_ratio != 0 (group2). Trong ttest\_ind, nếu statistic âm, nhóm đầu tiên (group1) có trung bình nhỏ hơn nhóm thứ hai (group2).
* Ý nghĩa thực tiễn: Học sinh ở các trường có video\_completion\_ratio == 0 có số lần thử giải bài tập trung bình (composite\_completion) ít hơn đáng kể so với học sinh ở các trường có video\_completion\_ratio != 0.

**2.2.2. Kiểm định Anova**

Tiến hành kiểm định Anova về vấn đề tỉ lệ composite\_completion giữa các course\_id với nhau.



Với kết quả thu được:

* Ý nghĩa thống kê:

Với p-value = 0.0 (nhỏ hơn ngưỡng 0.05). Điều này có nghĩa là có sự khác biệt đáng kể về trung bình composite\_completion giữa các nhóm trường học (dựa trên course\_id).

* Statistic (F-value):

Giá trị F = 129.60853678975826 là rất lớn, cho thấy sự khác biệt giữa các nhóm rất rõ rệt. Trong ANOVA, F-value lớn thường đi kèm với p-value nhỏ, củng cố kết luận rằng các trung bình không giống nhau.

* Ý nghĩa thực tiễn:

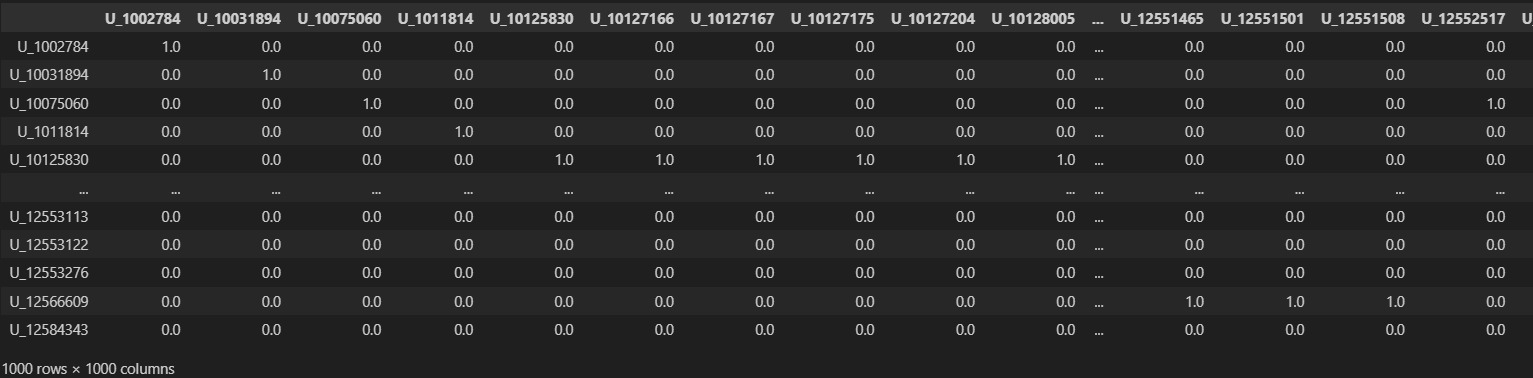
Kết quả cho thấy tỷ lệ hoàn thành trung bình (composite\_completion) của học sinh khác nhau đáng kể giữa các trường học (được mã hóa bởi course\_id).

Một số trường có thể có tỷ lệ hoàn thành cao hơn hoặc thấp hơn đáng kể so với các trường khác, có thể do các yếu tố như chất lượng giảng dạy, tài liệu học tập, hoặc sự tham gia của học sinh.

**2.2.3. Kiểm định Jaccard**

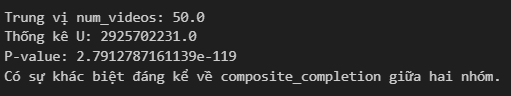
Tiến hành so sánh độ tương đồng về course mà các user đã tham gia (với điều kiện composite\_completion >= 0.5).

Những user nào có độ tương đồng với nhau về các course đã đăng ký sẽ có giá trị là 1 và ngược lại.



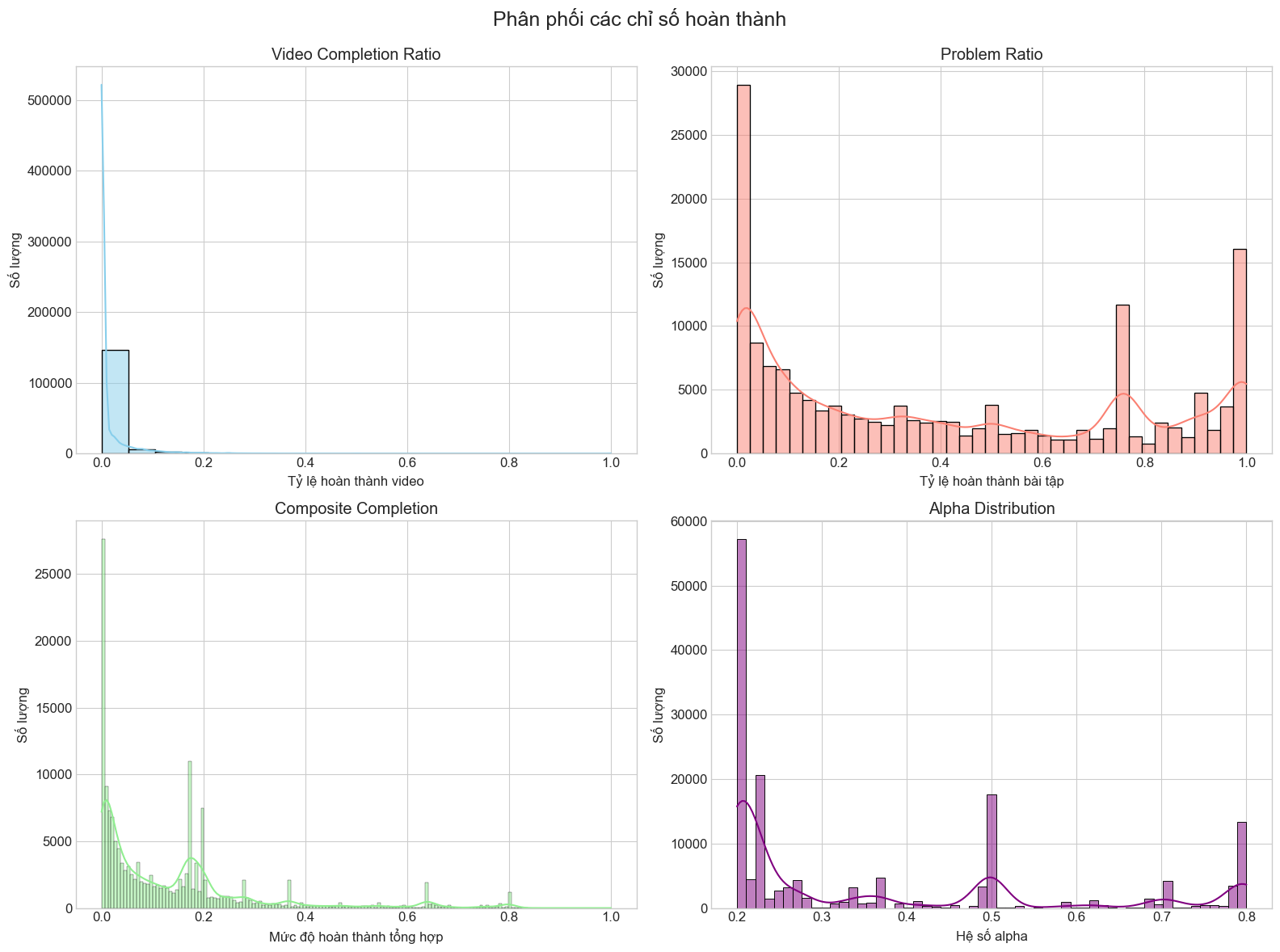
**2.2.4. Kiểm định Mann-WhitneyU**

So sánh composite\_completion giữa hai nhóm độc lập, ví dụ: người dùng tham gia khóa học có num\_videos trên và dưới trung vị.

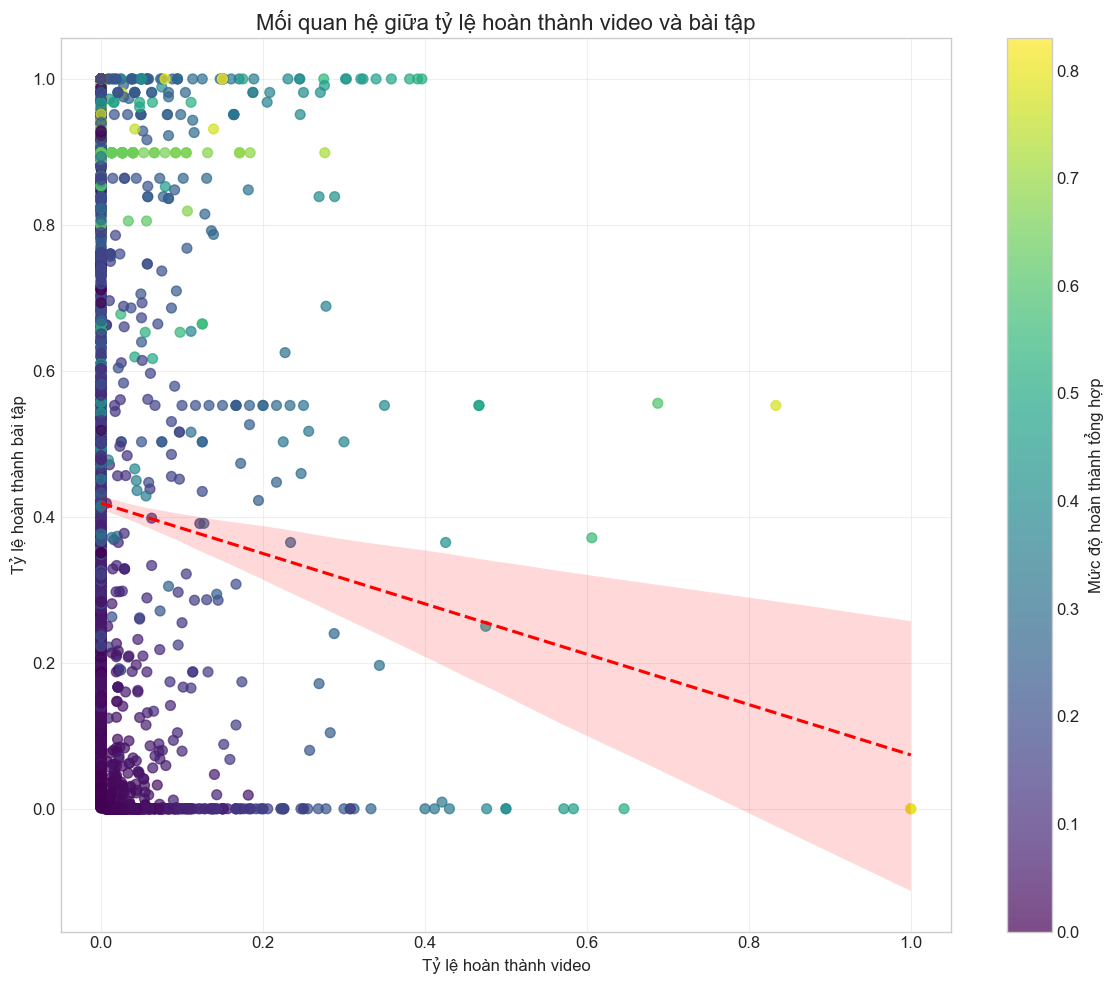


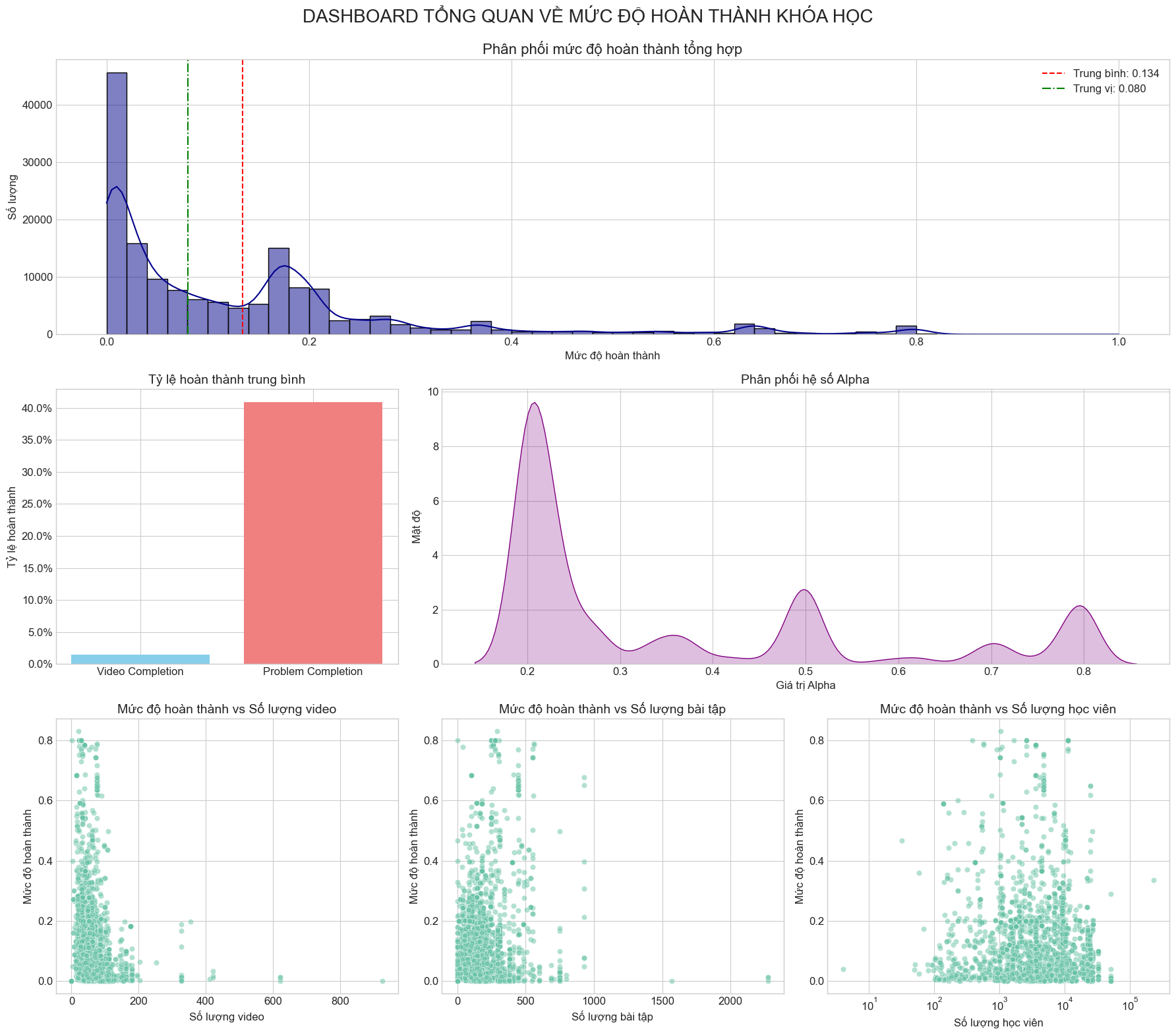
Từ output trên ta có thể thấy, những course có số lượng video nhỏ hơn trung vị sẽ thường có composite\_completion thấp so với những course có số lượng video lơn hơn trung vị. Từ đó, ta có thể nhận ra rằng những course có nhiều tài nguyên học tập, được đầu tư và trao chuốc kĩ lưỡng hơn với những video hướng dẫn cho user sẽ giúp học sinh dễ dàng học tập cũng như gắn bó hơn cho đến khi hoàn thành khóa học.

* 1. **Trực Quan Hóa**



* Video Completion Ratio (Tỷ lệ hoàn thành video):





* 1. **Khai Phá Tri Thức**

1. Phân tích tổng quan

* Dữ liệu người học được phân thành 3 cụm (cluster) dựa trên các đặc trưng hành vi như: thời lượng xem video, số lượt làm bài, tỉ lệ đúng, và mức độ tương tác (bình luận, trả lời).
* Biểu đồ phân tán bên dưới minh họa sự phân cụm theo:
* Trục X: điểm hoàn thành video (video\_completion\_score)
* Trục Y: lượt làm bài trung bình trên mỗi bài tập (attempts\_per\_problem)

2. Tóm tắt đặc trưng theo từng cụm

| Cụm | Mô tả tóm tắt |
| --- | --- |
| Cluster | Thời lượng xem video trung bình ~20,000s, số lần làm bài ~10,000, tỉ lệ đúng rất thấp (0.02), nhiều bình luận (~142,000) |
| Cluster 1 | Không xem video và không làm bài, tương tác thấp (gần như bỏ cuộc) |
| Cluster 2 | Thời lượng xem video rất cao (~256,800s), làm bài rất nhiều (trung bình 3.6 triệu lượt), tỉ lệ đúng cao nhất (0.62), tương tác rất nhiều (bình luận ~358,000) |

1. Diễn giải kết quả

| Cụm | Nhận định hành vi |
| --- | --- |
| Cụm 0 | Người học có mức tương tác khá cao, làm bài nhiều, nhưng tỉ lệ đúng rất thấp → có thể là người học chăm chỉ nhưng chưa hiệu quả |
| Cụm 1 | Người học không hoạt động, không xem video, không làm bài → có thể đã rời bỏ khóa học hoặc mất động lực |
| Cụm 2 | Người học xuất sắc: xem video nhiều, làm bài nhiều, đúng nhiều, tương tác cao → có thể là người học điển hình để làm gương hoặc hỗ trợ người khác |

1. Rút trích tri thức – Gợi ý hành động

| Cụm | Gợi ý |
| --- | --- |
| Cụm 0 | Cung cấp hỗ trợ học tập (ví dụ: tài liệu dễ hiểu hơn, tutor kèm cặp) để tăng hiệu quả học tập. |
| Cụm 1 | Gửi thông báo nhắc nhở, cá nhân hóa động viên, hoặc khảo sát lý do mất kết nối để can thiệp sớm. |
| Cụm 2 | Giao thêm thử thách nâng cao, mời làm mentor/tutor hỗ trợ học viên khác hoặc tham gia đánh giá khóa học. |

1. Kiến nghị

* Tích hợp kết quả phân cụm vào hệ thống quản lý học tập để tự động gợi ý hỗ trợ học viên theo cụm.
* Kết hợp thêm yếu tố thời gian (ví dụ: sự thay đổi hành vi theo tuần) để phát hiện sớm nguy cơ bỏ cuộc.
* Áp dụng mô hình học máy để dự đoán người học sẽ rơi vào cụm nào trong tương lai từ hành vi ban đầu.