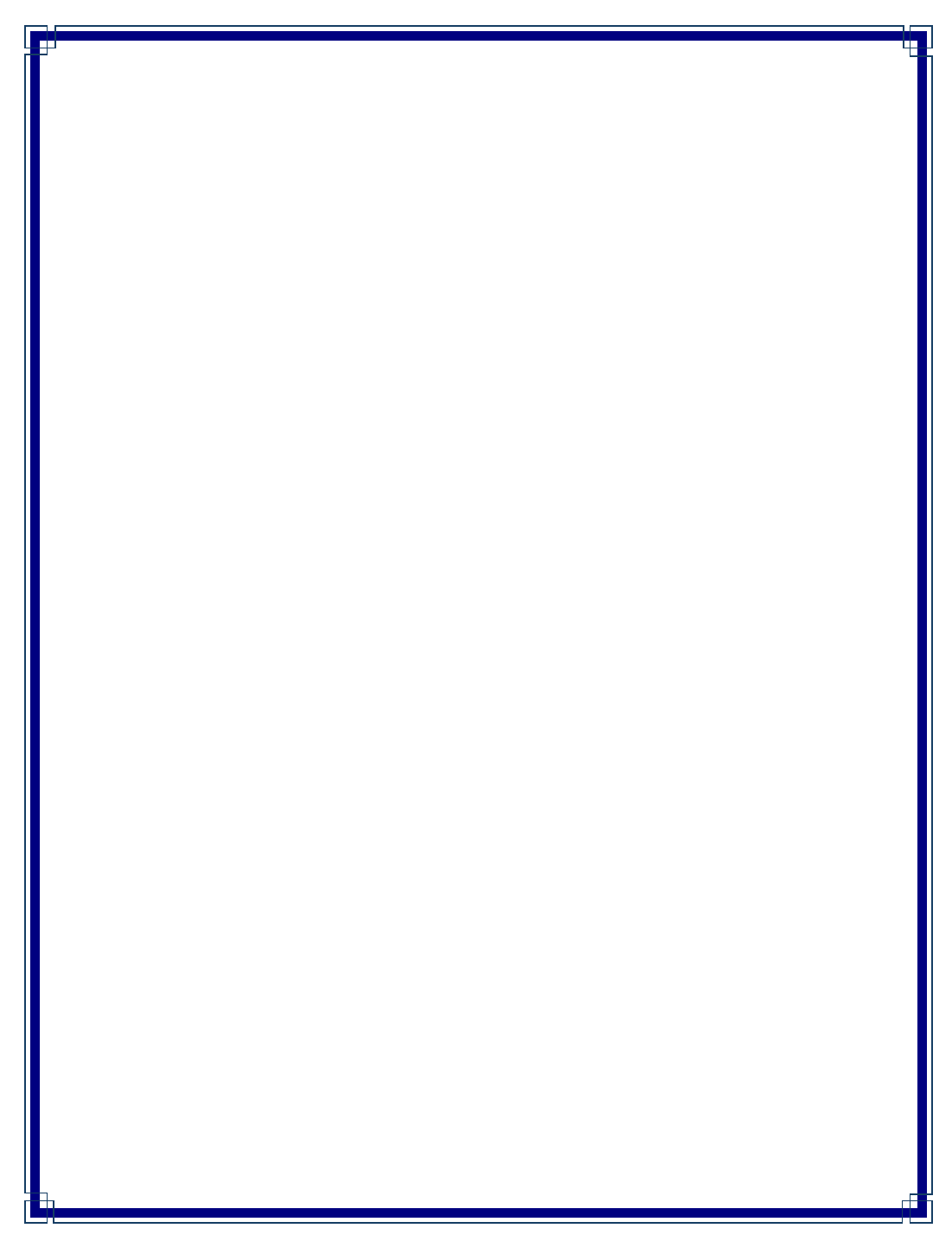
**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**



**KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG**

**CHIẾN LƯỢC GÁN NHÃN**

**GVHD:** ThS. Nguyễn Thị Anh Thư

**Lớp:** CS313.P22

**Nhóm 6**

**Họ và tên MSSV**

Nguyễn Hồng Phát 22521072

Đặng Thanh Ngân 22520929

Phạm Thanh Thảo 22521373

Đinh Hữu Phước 22521150

Lê Dương Minh Thiên 22521386

🙡🙢 Tp. Hồ Chí Minh, 4/2025 🙠

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

....................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................

**Mục lục**

[**I. Trực quan và giới hạn input 5**](#_3m87z7cyf4kp)

[1. Loại bỏ những khóa học chỉ có ít học viên đăng kí 5](#_gmjh8v9xmxob)

[Các khóa học có nhiều học viên đăng ký thường: 7](#_4x1vzt31dydy)

[2. Loại bỏ những khóa học có ít tài liêu hoặc nhiều tài liệu 8](#_8p965xwrlzu2)

[3. Loại bỏ những khóa học không có bài tập và không có video 11](#_g0om2olsxwy0)

[4. Loại bỏ những khóa học không giới hạn thời gian 12](#_v3t1ph5unuc7)

[5. Loại bỏ những học viên chỉ đăng kí những không làm bài tập 13](#_w9ypykedjrle)

[6. Thông tin input cuối cùng 14](#_b2bphdifhchh)

[**II. Thêm thông tin về khóa học 18**](#_iogzc7zd04g)

[**III. Chiến lược gán nhãn 19**](#_2ftqjott5eug)

[1. Cách tính điểm trong các khóa của Xuetang 19](#_yrwp87cu1yq)

[2. Xếp loại kết quả của học viên trong XuetangX 20](#_s6g9qf7z2rdl)

[3. Xếp loại kết quả của học viên của nhóm 21](#_kdips8xxjq1d)

[Thành phần điểm: 21](#_gxeiiwytha58)

[Công thức tính điểm tổng kết: 21](#_mqwhzfmurefi)

[Nguyên tắc gán nhãn: 22](#_wjpy2jl7x400)

[Bảng phân loại chi tiết: 22](#_gl3fllehy4qe)

# 

# Trực quan và giới hạn input

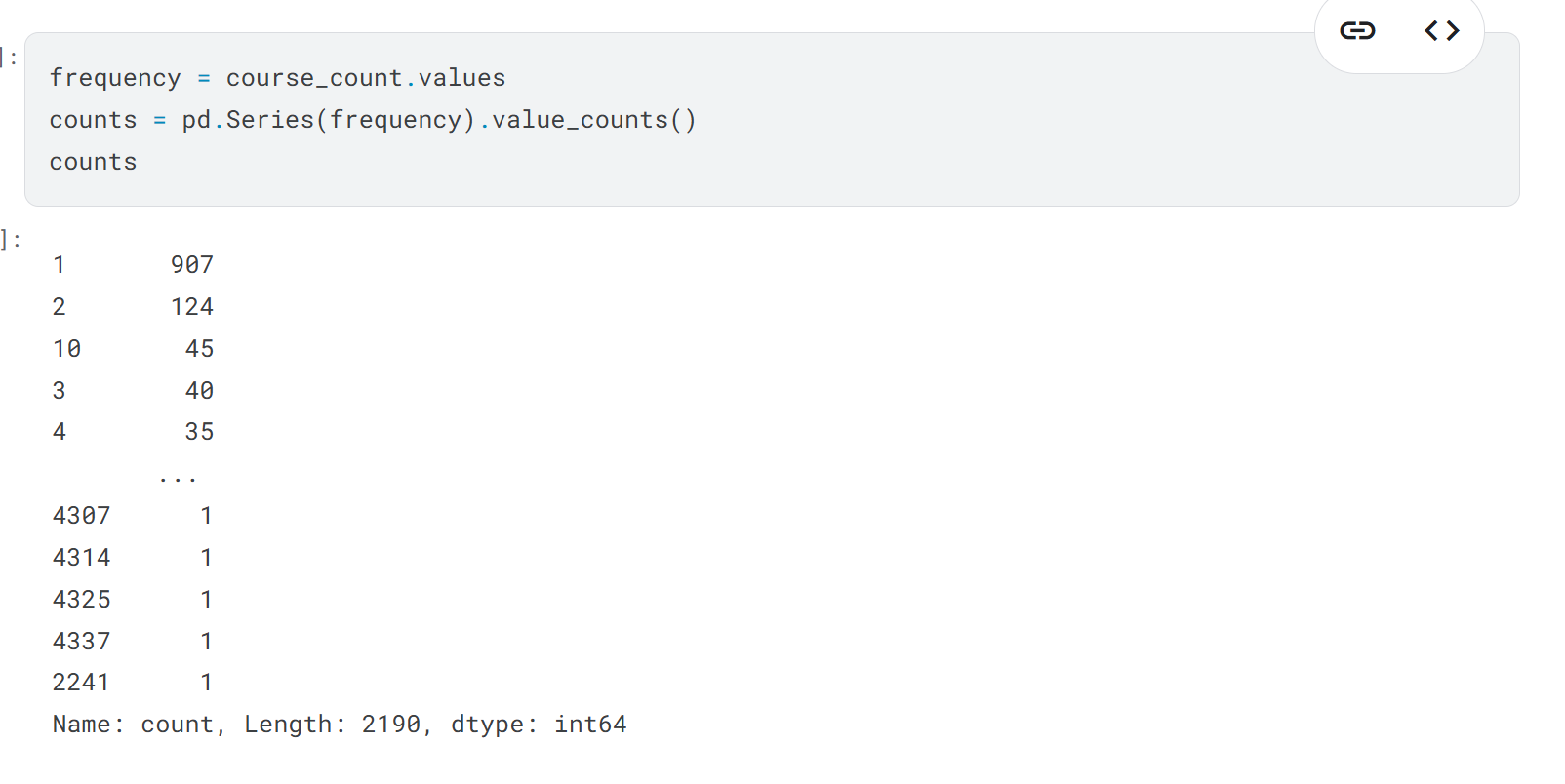
## Loại bỏ những khóa học có ít học viên đăng kí

****

Tổng số khóa học được đăng ký: 4,701

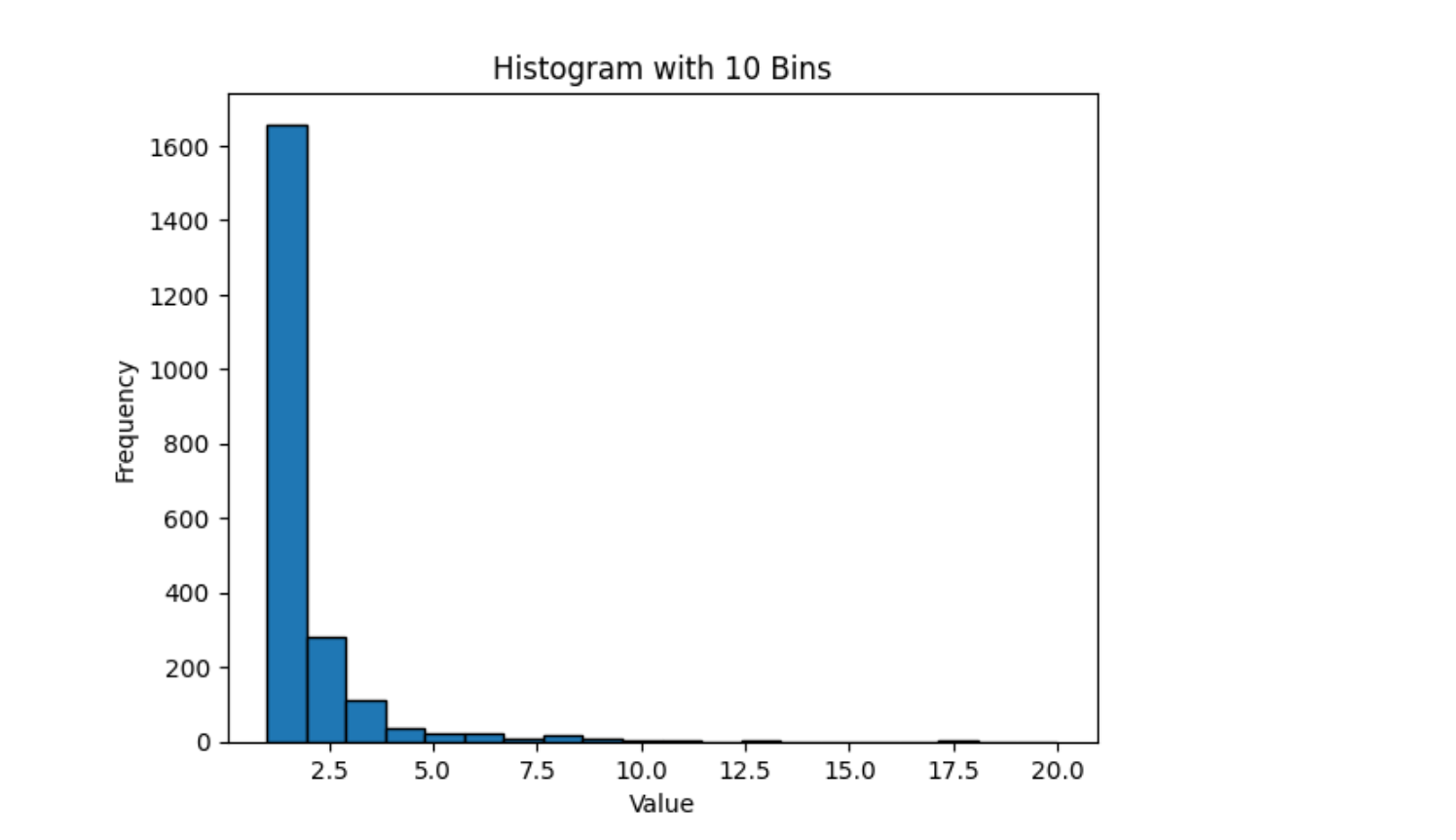
Số lượng khóa học thực tế có dữ liệu: ~3,900 ⇒ 800+ khóa học không có thông tin chi tiết hoặc đã bị loại bỏ trong dữ liệu gốc.

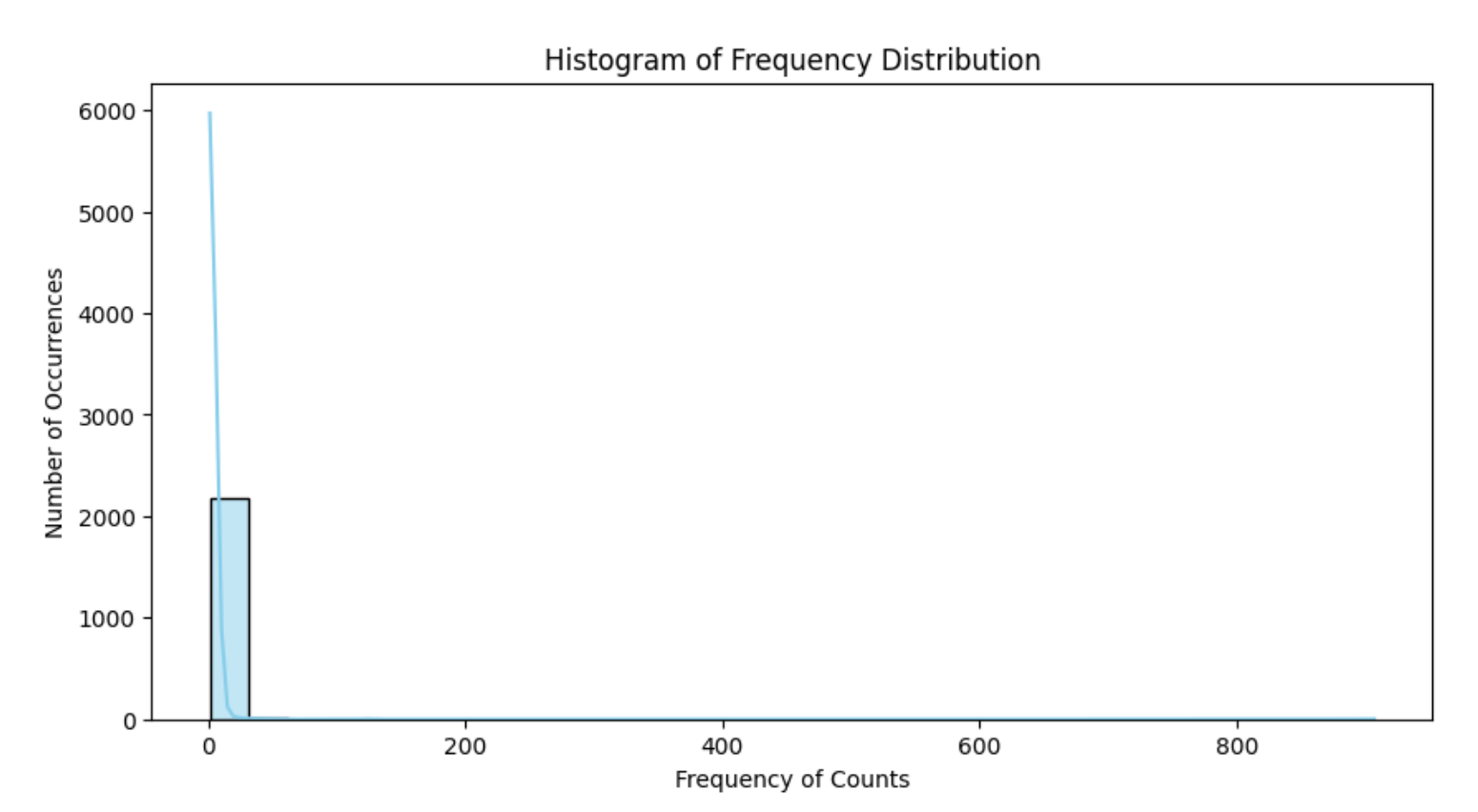
Số lượng khóa học chỉ có 1 học viên đăng ký: 907 khóa học ⇒ Đây là nhóm chiếm tỷ lệ cao nhất trong tất cả các khóa học.

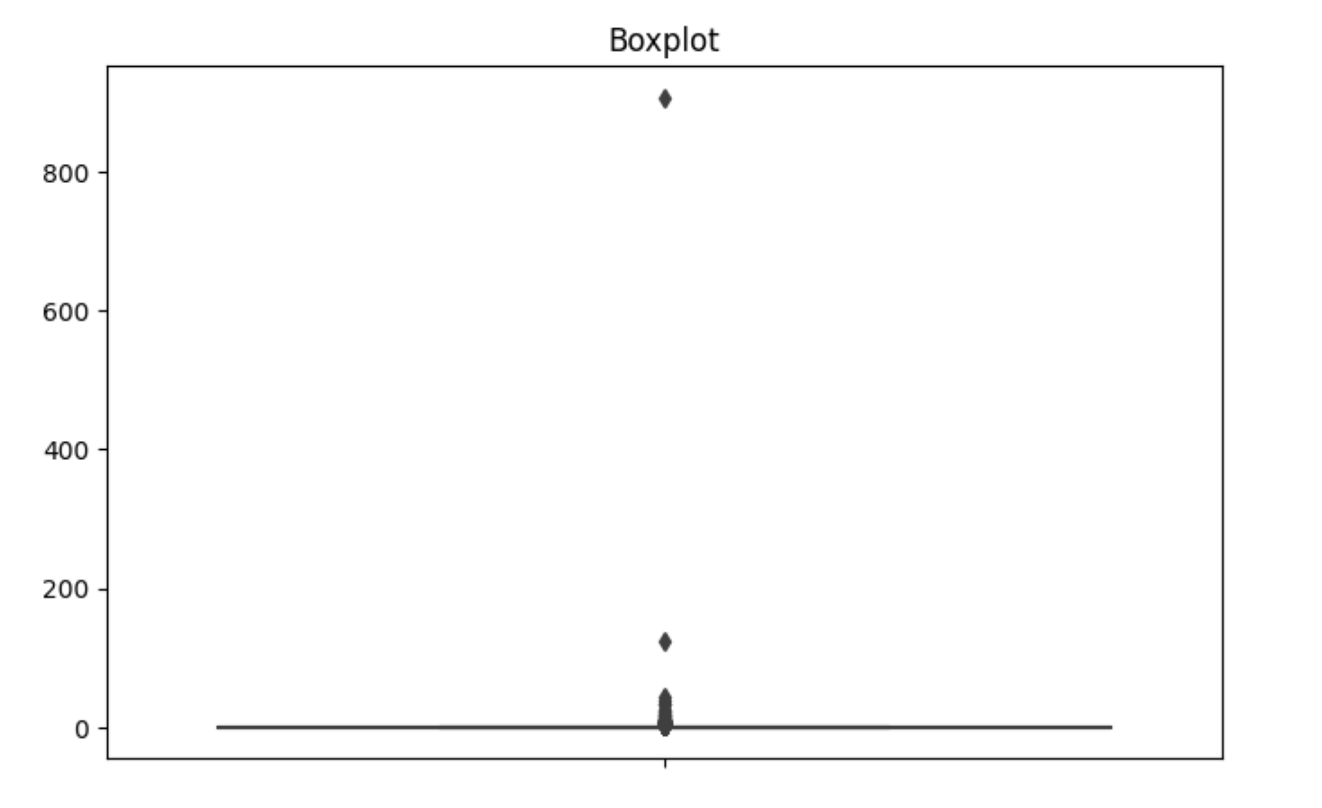
****

Phần lớn các khóa học có rất ít học viên đăng ký (đa phần từ 1–3 người).

Chỉ một phần nhỏ khóa học có trên 10 học viên đăng ký.





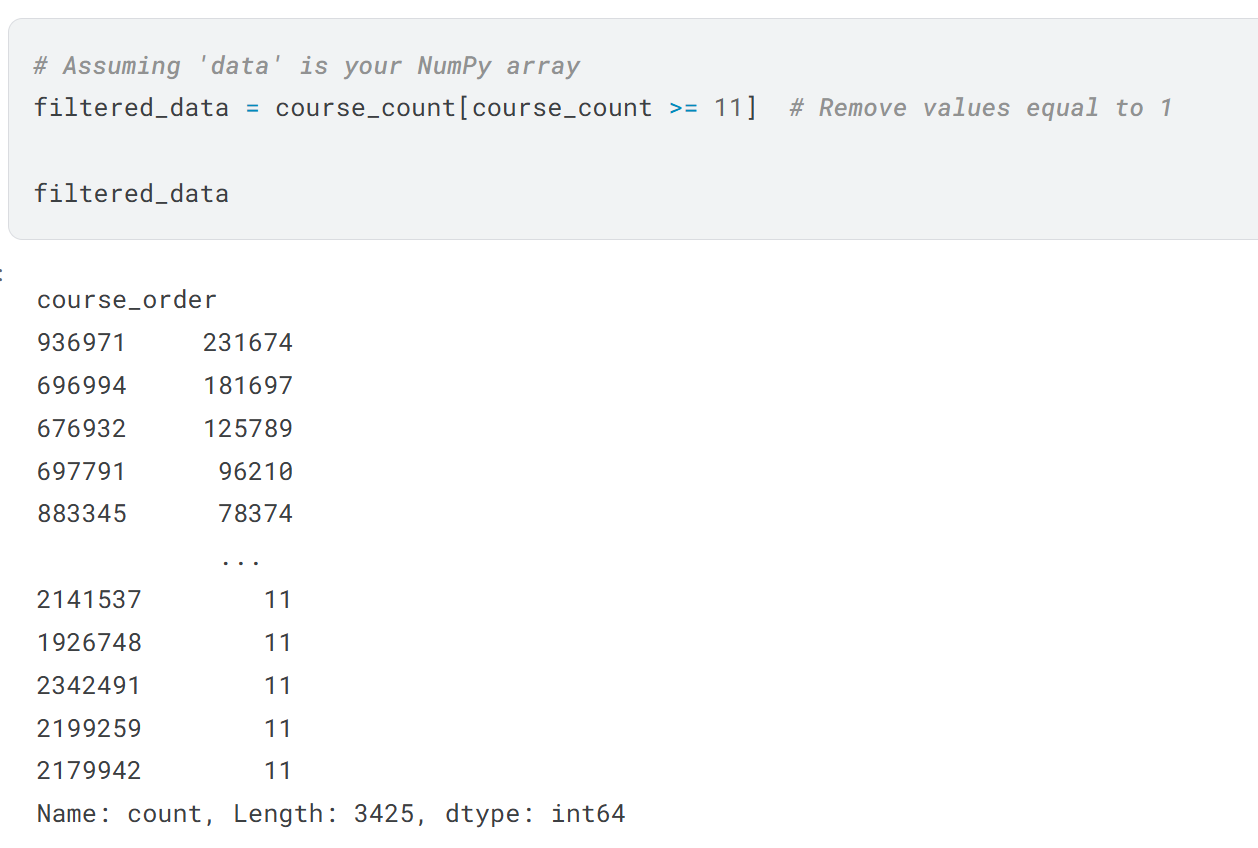


Lọc những khóa học chỉ có hơn 10 người đăng kí học việc này giúp tập trung vào các khóa học "phổ biến" và tránh trường hợp bị null.

Các khóa học có nhiều học viên đăng ký thường:

* + Có nội dung hấp dẫn.
  + Được giảng viên nổi bật giảng dạy.
  + Có chất lượng hoặc đánh giá tốt.

Việc lọc giúp phân tích xu hướng học tập thực sự phổ biến, thay vì bị nhiễu bởi hàng trăm khóa học chỉ có 1-2 người học.



## Loại bỏ những khóa học có ít tài liêu hoặc nhiều tài liệu

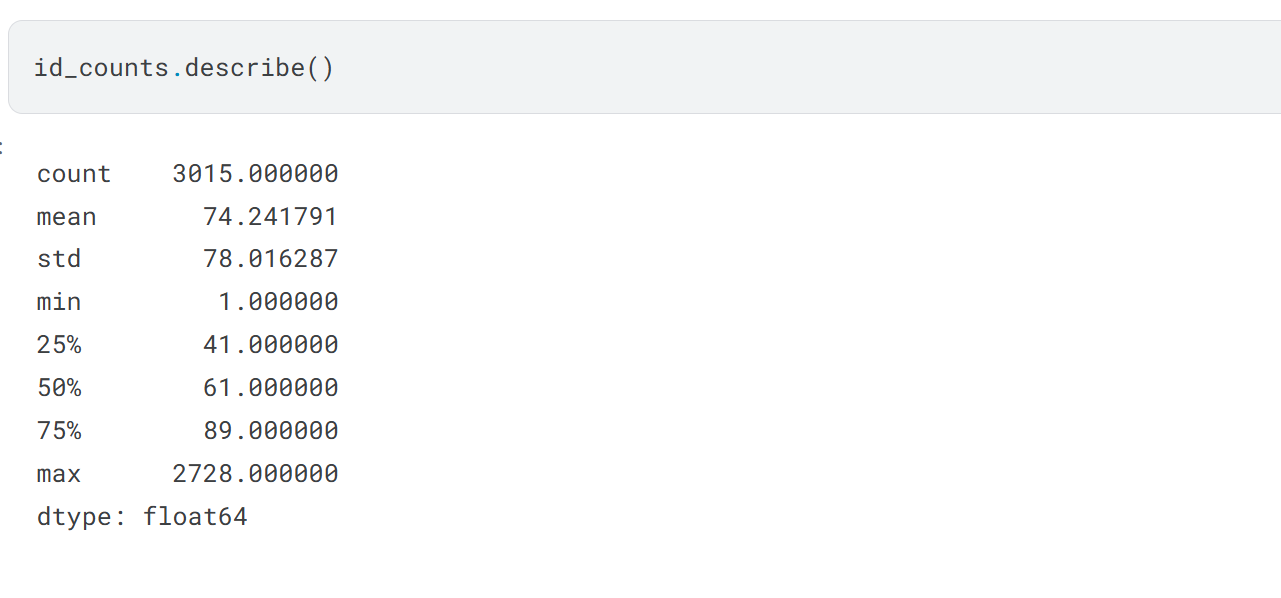
Sử dụng hàm explode() để tách từng resource trong danh sách ra thành một dòng riêng, giúp tính toán chính xác số lượng tài nguyên trên mỗi khóa học.



Số lượng resource trong mỗi khóa học

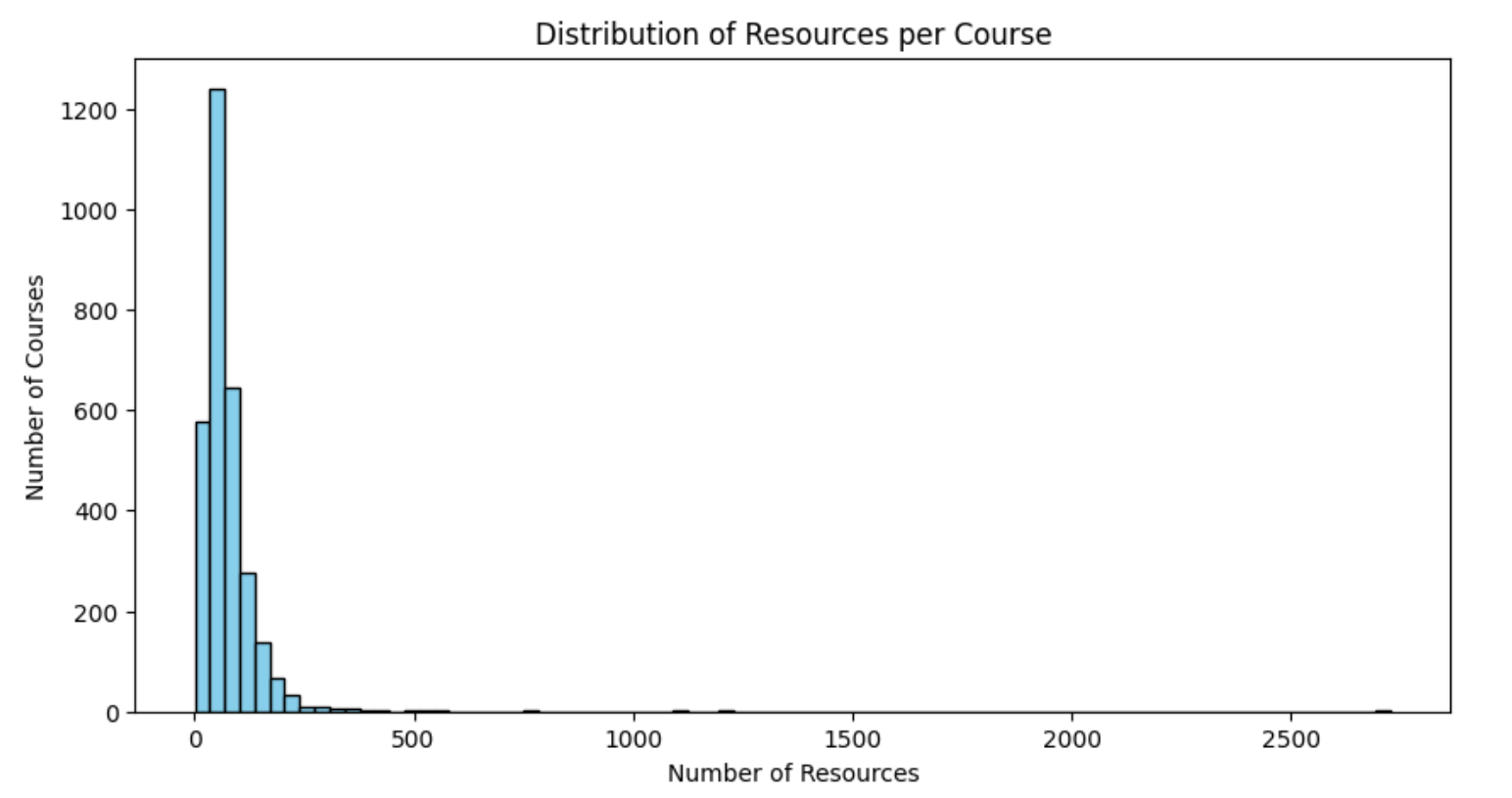


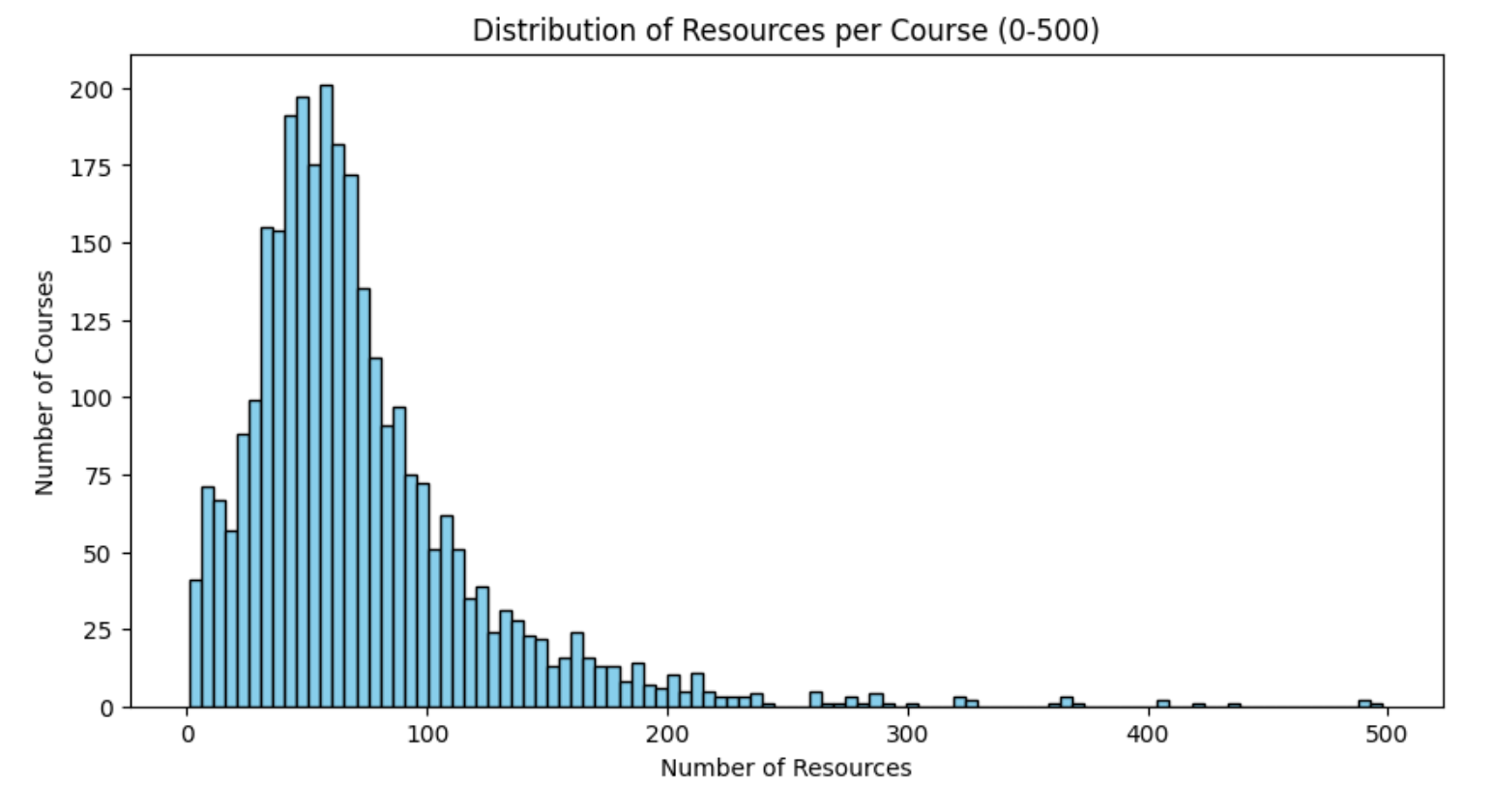
Áp dụng hàm describe() để tổng hợp:

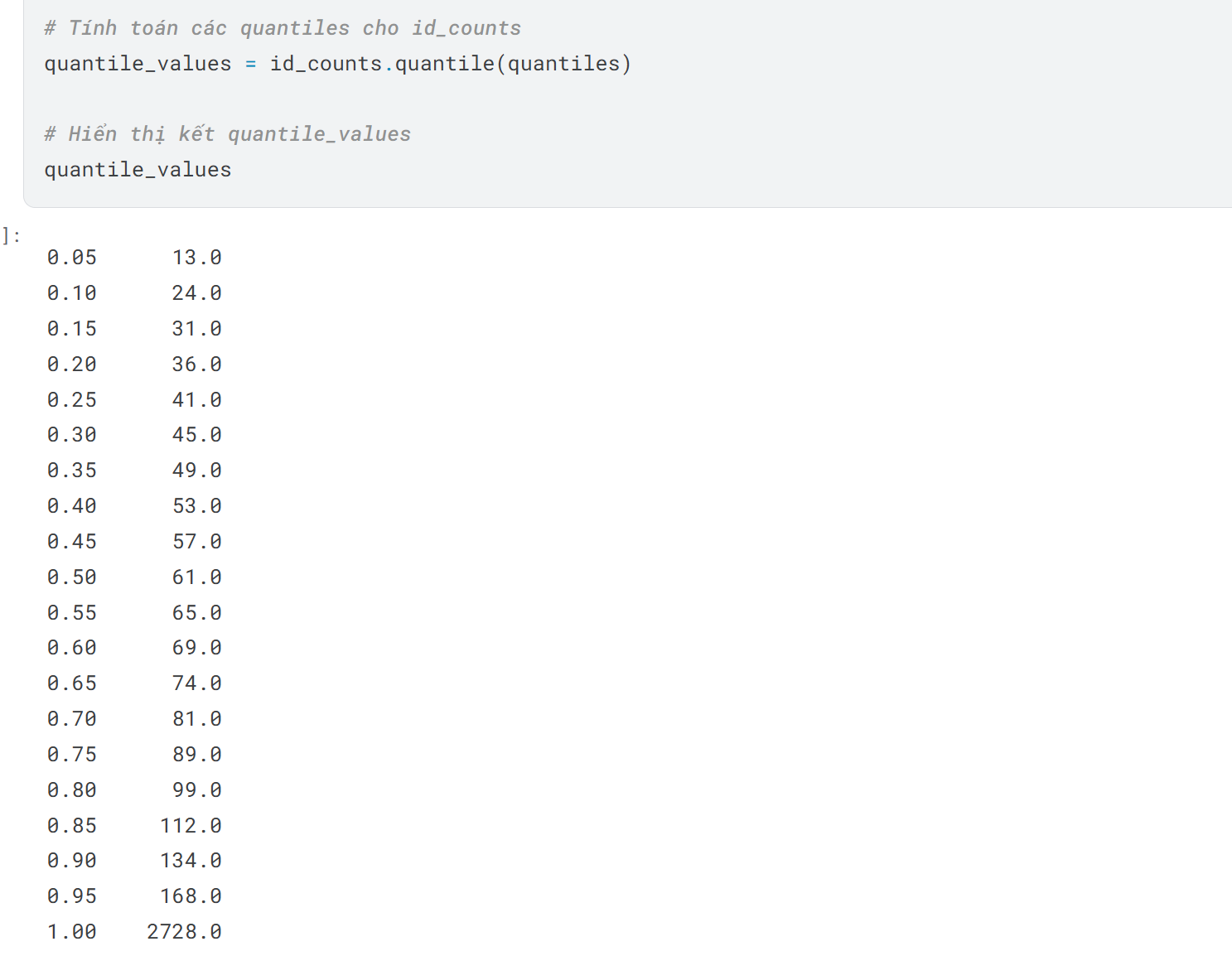


Phân tích số liệu:

* Min = 1 → Có những khóa học rất ít tài nguyên, có thể do lỗi dữ liệu hoặc không hữu ích.
* Max = 2728 → Có một số khóa học có tài nguyên cực lớn, có thể là ngoại lệ.
* Mean = 74.24 → Giá trị trung bình khá hợp lý.
* 50% (Median) = 61 → Một nửa khóa học có tài nguyên ≤ 61.
* 75% (Q3) = 89 → 75% khóa học có tài nguyên ≤ 89.
* Std (Độ lệch chuẩn) = 78.01 → Khoảng cách giữa các giá trị khá lớn







Giới hạn những khóa học có từ 13 đến 168 resource quatile là 0.05 đến 0.95 tức giới hạn từ 13 đến 168 resource

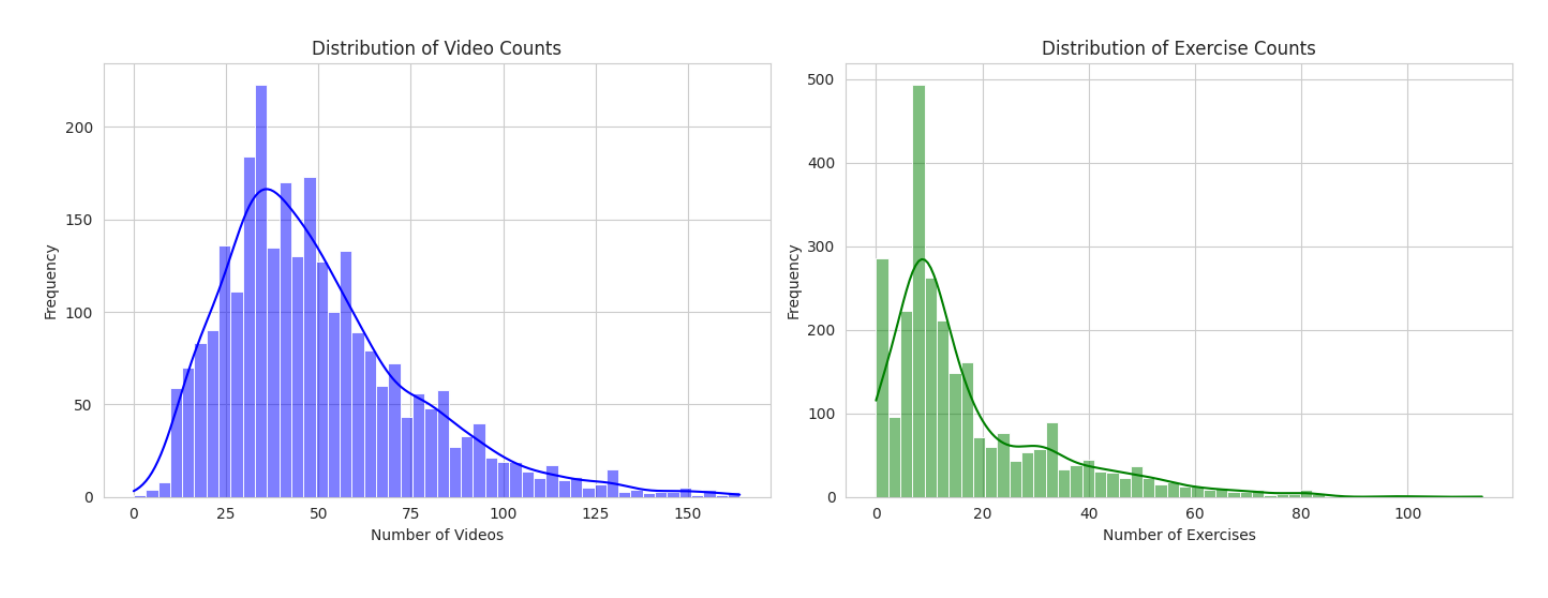
Giới hạn dưới (5% quantile): 13 tài nguyên.

Giới hạn trên (95% quantile): 168 tài nguyên.

⇒ Chỉ giữ lại các khóa học có từ 13 đến 168 tài nguyên để phân tích sâu hơn.

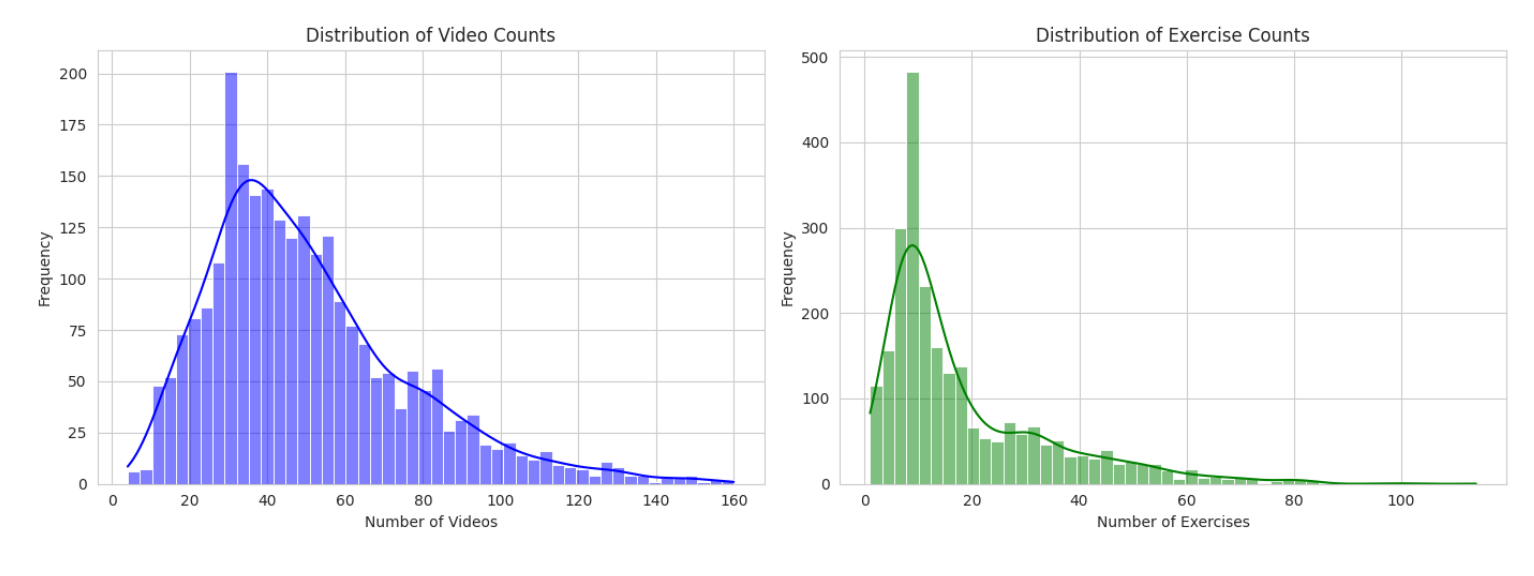
Nhằm loại bỏ các trường hợp quá ít (có thể là lỗi) hoặc quá nhiều (có thể là ngoại lệ), giúp dữ liệu tập trung và đáng tin cậy hơn. Tránh trường hợp khóa học quá ngắn thời gian học ngắn khó dự đoán sớm

## Loại bỏ những khóa học không có bài tập và không có video

****

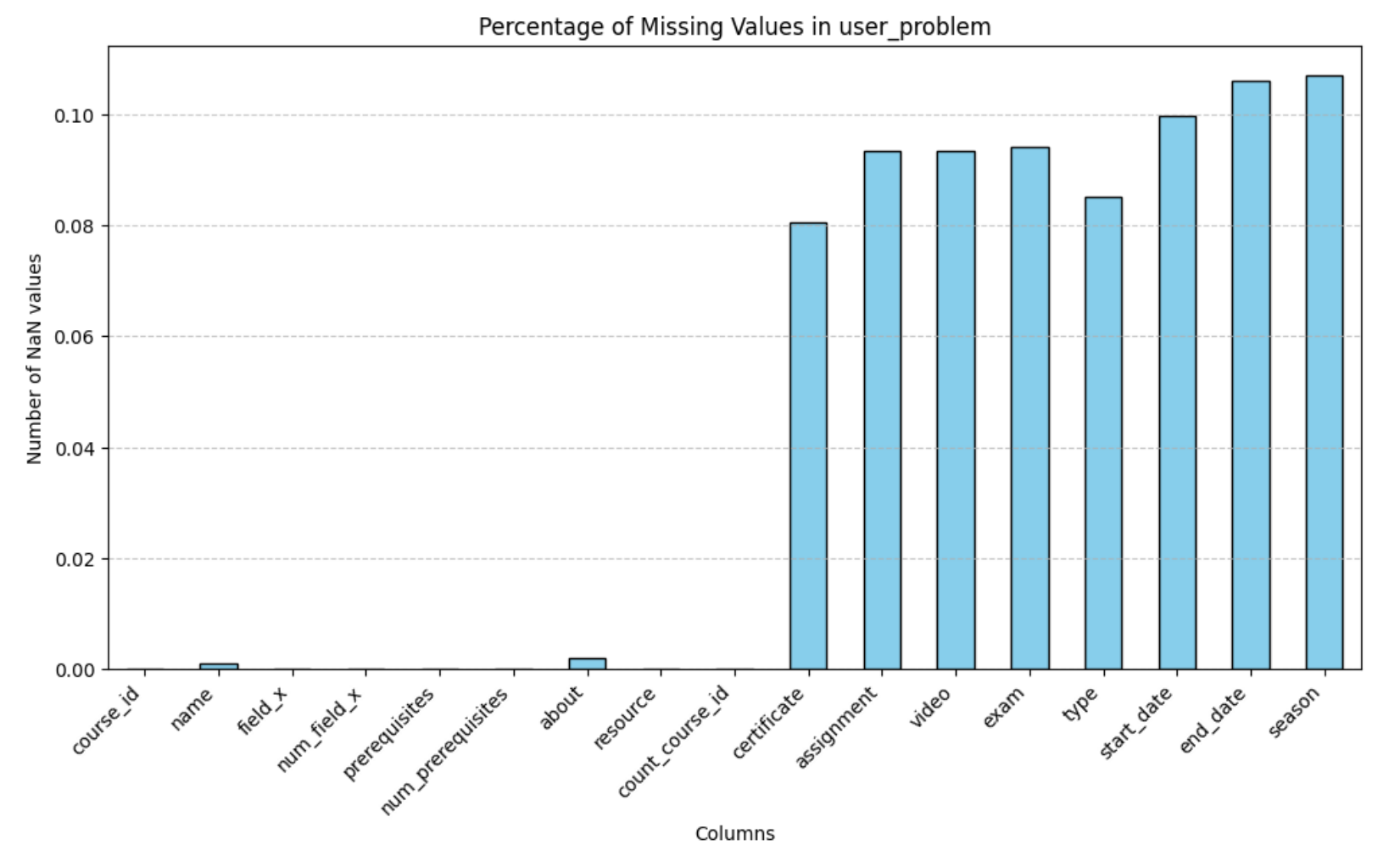
*Bảng phân phối số lượng video và exercise trong từng khóa học*

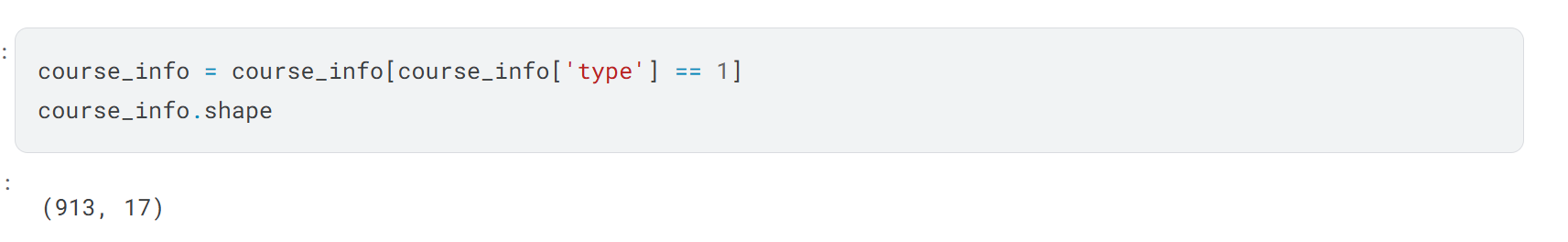




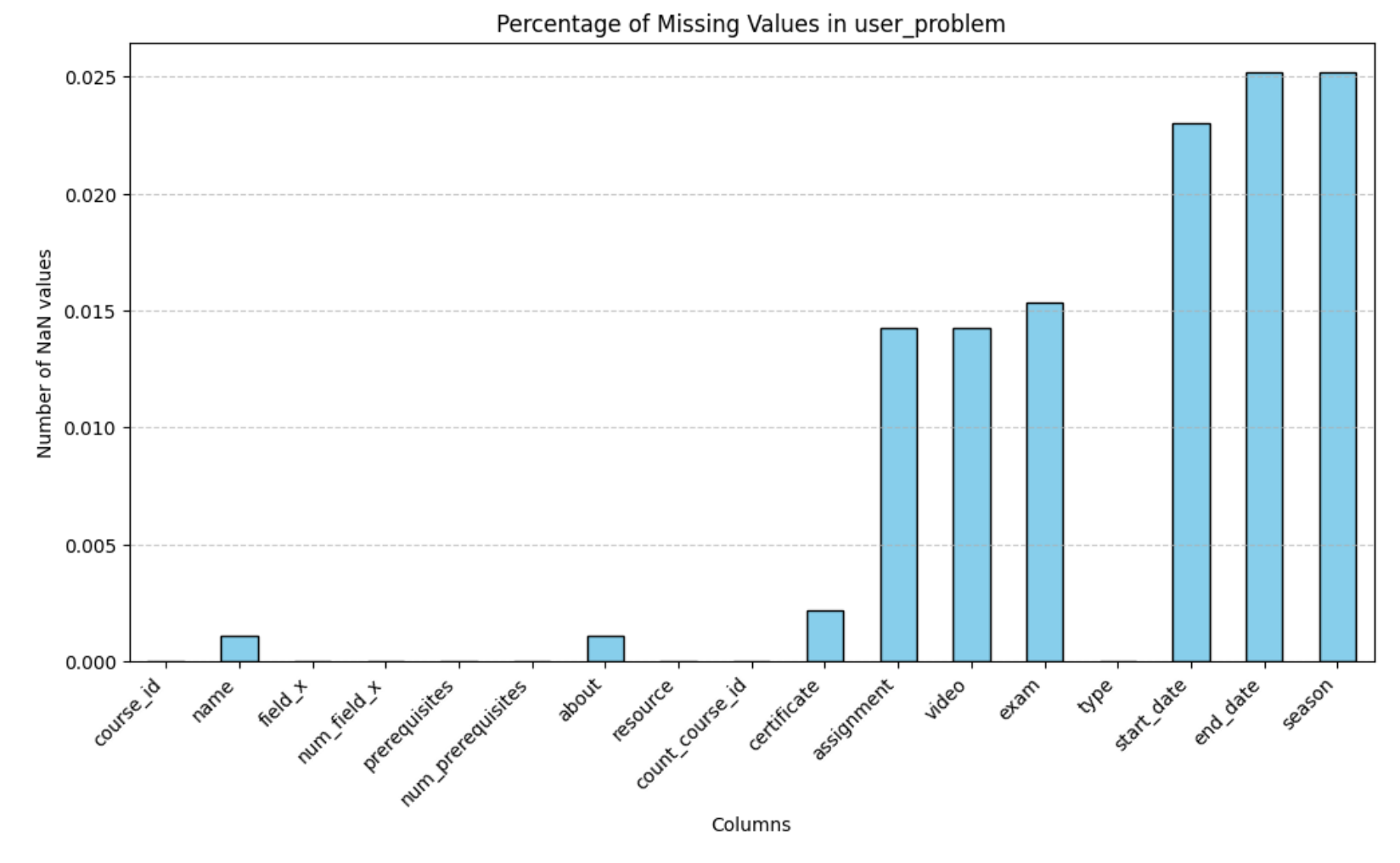
Bảng đồ sau khi phân phối dữ liệu

## Loại bỏ những khóa học không giới hạn thời gian

****

****

Chỉ giữ lại những khóa học loại một - có thời gian bắt đầu và kết thúc của khóa học.

****

## Loại bỏ những học viên chỉ đăng kí những không làm bài tập

****

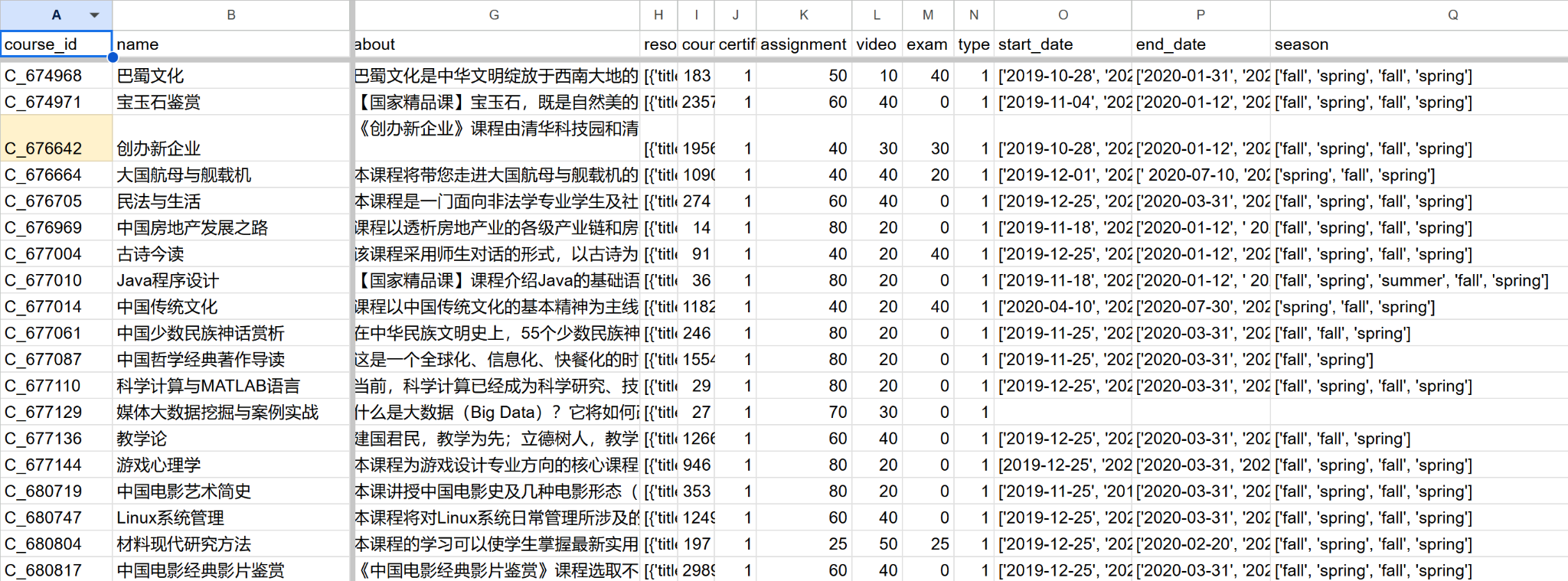
## Thông tin input cuối cùng

## 

| **Nhóm** | **Tên cột** | **Ý nghĩa** |
| --- | --- | --- |
| **Thông tin khóa học (course\_\*)** | course\_id | Mã khóa học |
|  | num\_prerequisites | Số lượng môn học tiên quyết |
|  | field\_x | Lĩnh vực khóa học (optional) |
|  | num\_field\_x | Số lượng lĩnh vực liên quan |
|  | start\_date | Ngày bắt đầu khóa học |
|  | end\_date | Ngày kết thúc khóa học |
|  | duration\_days | Độ dài của khóa học (ngày) |
| **Tài liệu khóa học (resource\_\*)** | video\_count | Tổng số lượng video |
|  | exercise\_count | Tổng số lượng bài tập |
|  | chapter\_count | Tổng số lượng chương |
| **Thành phần điểm (score\_\*)** | assignment | Điểm bài tập |
|  | exam | Điểm bài thi cuối kỳ |
|  | video | Điểm từ xem video |
|  | certificate | Có nhận chứng chỉ hay không (0, 1) |
| **Thông tin người dùng (user\_\*)** | user\_id | ID người dùng |
|  | school | Trường học |
|  | user\_enroll\_time | Thời gian người dùng đăng ký khóa học |
|  | user\_past\_course\_count | Số lượng khóa học đã đăng ký trước đó |
|  | user\_time\_since\_last\_course | Khoảng cách (giờ) từ lần đăng ký hiện tại đến lần gần nhất |
| **Hành vi học tập - Bình luận (comment\_\*)**  **- Theo từng đợt** | comment\_count\_phase{i} | Số lượng bình luận trong giai đoạn i |
|  | total\_words\_phase{i}\_x | Tổng số từ trong các bình luận trong giai đoạn i |
|  | entropy\_time\_comment\_phase{i} | Mức độ phân tán thời gian của bình luận (entropy) trong giai đoạn i |
|  | total\_positive{i} | Tổng bình luận và phản hồi tích cực của học viên trong giai đoạn i |
|  | total\_negative1 | Tổng bình luận và phản hồi tiêu cực của học viên trong giai đoạn i |
|  | total\_neutral1 | Tổng bình luận và phản hồi trung lập của học viên trong giai đoạn i |
| **Hành vi học tập - Phản hồi (reply\_\*)**  **- Theo từng đợt** | reply\_count\_phase{i} | Số lượng phản hồi trong giai đoạn i |
|  | total\_words\_phase{i}\_y | Tổng số từ trong các phản hồi trong giai đoạn i |
|  | entropy\_time\_reply\_phase{i} | Mức độ phân tán thời gian của phản hồi (entropy) trong giai đoạn i |
|  | total\_positive{i} | Tổng bình luận và phản hồi tích cực của học viên trong giai đoạn i |
|  | total\_negative1 | Tổng bình luận và phản hồi tiêu cực của học viên trong giai đoạn i |
|  | total\_neutral1 | Tổng bình luận và phản hồi trung lập của học viên trong giai đoạn i |
| **Hành vi học tập - Bài tập (exercise\_\*)**  **- Theo từng đợt** | exercise\_id\_count\_{i} | Số lượng bài tập (exercise) được làm trong giai đoạn i |
|  | exercise\_correct\_sum\_{i} | Tổng số câu trả lời đúng trong giai đoạn i |
|  | exercise\_correct\_mean\_{i} | Tỷ lệ trả lời đúng trung bình của mỗi bài tập trong giai đoạn i |
|  | exercise\_num\_problem\_sum\_{i} | Tổng số câu hỏi trong tất cả các bài tập trong giai đoạn i |
|  | exercise\_num\_problem\_mean\_{i} | Số câu hỏi trung bình mỗi bài trong giai đoạn i |
|  | exercise\_attempts\_sum\_sum\_{i} | Tổng số lần thử làm các bài tập trong giai đoạn i |
|  | exercise\_attempts\_sum\_mean\_{i} | Số lần thử trung bình mỗi bài trong giai đoạn i |
|  | exercise\_attempts\_mean\_mean\_{i} | Trung bình số lần thử mỗi câu hỏi (problem) trong giai đoạn i |
|  | exercise\_date\_from\_enroll\_min\_{i} | Ngày đầu tiên làm bài tập (so với ngày đăng ký) trong giai đoạn i |
|  | exercise\_date\_from\_enroll\_mean\_{i} | Ngày trung bình làm bài (so với ngày đăng ký) trong giai đoạn i |
|  | exercise\_date\_from\_enroll\_max\_{i} | Ngày cuối cùng làm bài (so với ngày đăng ký) |
|  | exercise\_context\_sum\_{i} | Tổng số ngữ cảnh làm bài tập trong giai đoạn i |
|  | exercise\_context\_mean\_{i} | Trung bình số ngữ cảnh làm bài tập trong giai đoạn i |
|  | exercise\_langugage\_binary\_mean\_{i} | Trung bình giá trị ngôn ngữ bài tập (1 = Tiếng Anh, 0 = Tiếng Trung) trong giai đoạn i |
|  | exercise\_diff\_sum\_{i} | Tổng thời gian làm bài tập trong giai đoạn i |
|  | exercise\_diff\_mean\_{i} | Trung binh thời gian làm mỗi bài tập trong giai đoạn i |
|  | exercise\_diff\_min\_{i} | Thời gian làm bài ít nhất trong giai đoạn i |
|  | exercise\_diff\_max\_{i} | Thời gian làm bài nhiều nhất trong giai đoạn i |
|  | exercise\_perc\_goal\_correct\_sum\_{i} | Tổng của tỷ lệ trả lời đúng giai đoạn i |
|  | exercise\_perc\_goal\_correct\_mean\_{i} | Trung bình của tỷ lệ trả lời đúng trong giai đoạn i |
|  | exercise\_perc\_goal\_score\_sum\_{i} | Tổng điểm đạt được trong giai đoạn i |
|  | exercise\_perc\_goal\_score\_mean\_{i} | Trung bình điểm đạt được của mỗi bài tập trong giai đoạn i |
|  | exercise\_perc\_real\_completed\_sum\_{i} | Tổng tỷ lệ hoàn thành bài tập thực tế trong giai đoạn i |
|  | exercise\_perc\_real\_completed\_mean\_{i} | Trung bình tỷ lệ hoàn thành thực tế |
|  | exercise\_perc\_real\_completed\_std\_{i} | Độ lệch chuẩn của tỷ lệ hoàn thành thực tế |
|  | exercise\_perc\_real\_correct\_sum\_{i} | Tổng tỷ lệ đúng thực tế |
|  | exercise\_perc\_real\_correct\_mean\_{i} | Tỷ lệ đúng thực tế trung bình |
|  | exercise\_perc\_real\_correct\_std\_{i} | Độ lệch chuẩn của tỷ lệ đúng thực tế |
|  | exercise\_perc\_real\_score\_sum\_{i} | Tổng điểm thực tế đạt được |
|  | exercise\_perc\_real\_score\_mean\_{i} | Trung bình điểm thực tế |
|  | exercise\_perc\_real\_score\_std\_{i} | Độ lệch chuẩn điểm thực tế |
|  | exercise\_hour\_entropy\_{i} | Entropy của thời gian làm bài tập theo giờ (phân tán trong ngày) |
| **Hành vi học tập - Video (video\_\*) - Theo từng đợt** | video\_watched\_count\_{i} | Số lượng video được xem trong đợt i (user coi ở giây cuối cùng được coi là hoàn thành) |
|  | video\_watched\_percentage\_{i} | Tỷ lệ video đã xem trên tổng số video của khóa học trong từng đợt |
|  | video\_watch\_time\_{i} | Tổng thời gian user xem video trong đợt i |
|  | video\_time\_{i} | Tổng thời gian video được chọn để coi đợt i |
|  | video\_percentage\_watch\_{i} | Phần trăm thời lượng user coi trên tổng thời lượng của video |
|  | video\_pause\_count\_{i} | Số lượng ngắt quãng trong tổng số video đợt i |
|  | video\_pause\_avg\_{i} | Trung bình số lần ngắt quãng đợt i |
|  | video\_pause\_std\_{i} | Độ lệch chuẩn số lần ngắt quãng đợt i |
|  | video\_rewatch\_count\_{i} | Số lần xem lại đonạ video (segment) đợt i |
|  | video\_rewatch\_avg\_{i} | Trung bình số lần xem lại trên mỗi video đợt i |
|  | video\_rewatch\_std\_{i} | Độ lệch chuẩn số lần xem lại trên mỗi video đợt i |
|  | video\_time\_between\_views\_avg\_{i} | Trung bình thời gian giữa các lần xem video đợt i |
|  | video\_speed\_avg\_{i} | Trung bình tốc độ xem video đợt i |
|  | video\_entropy\_time\_{i} | Entropy thời điểm user coi video đợt i |
|  | video\_final\_score\_percentage\_{i} | Phần trăm điểm đặt được vào điểm cuối khóa |
| **Nhãn** | label | A, B, C, D, E thể hiện kết quả học tập cuối cùng của học viên |

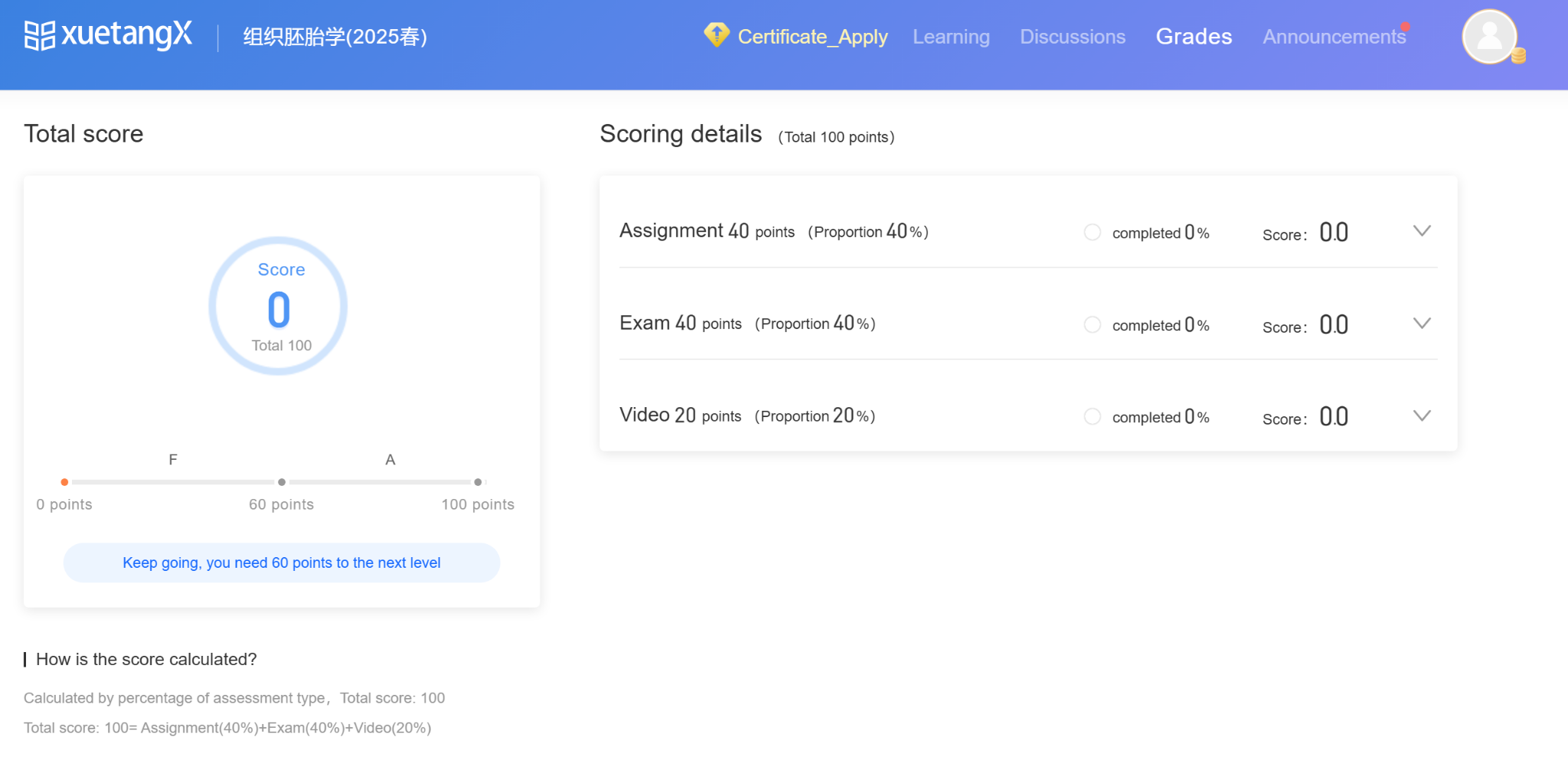
# Thêm thông tin về khóa học

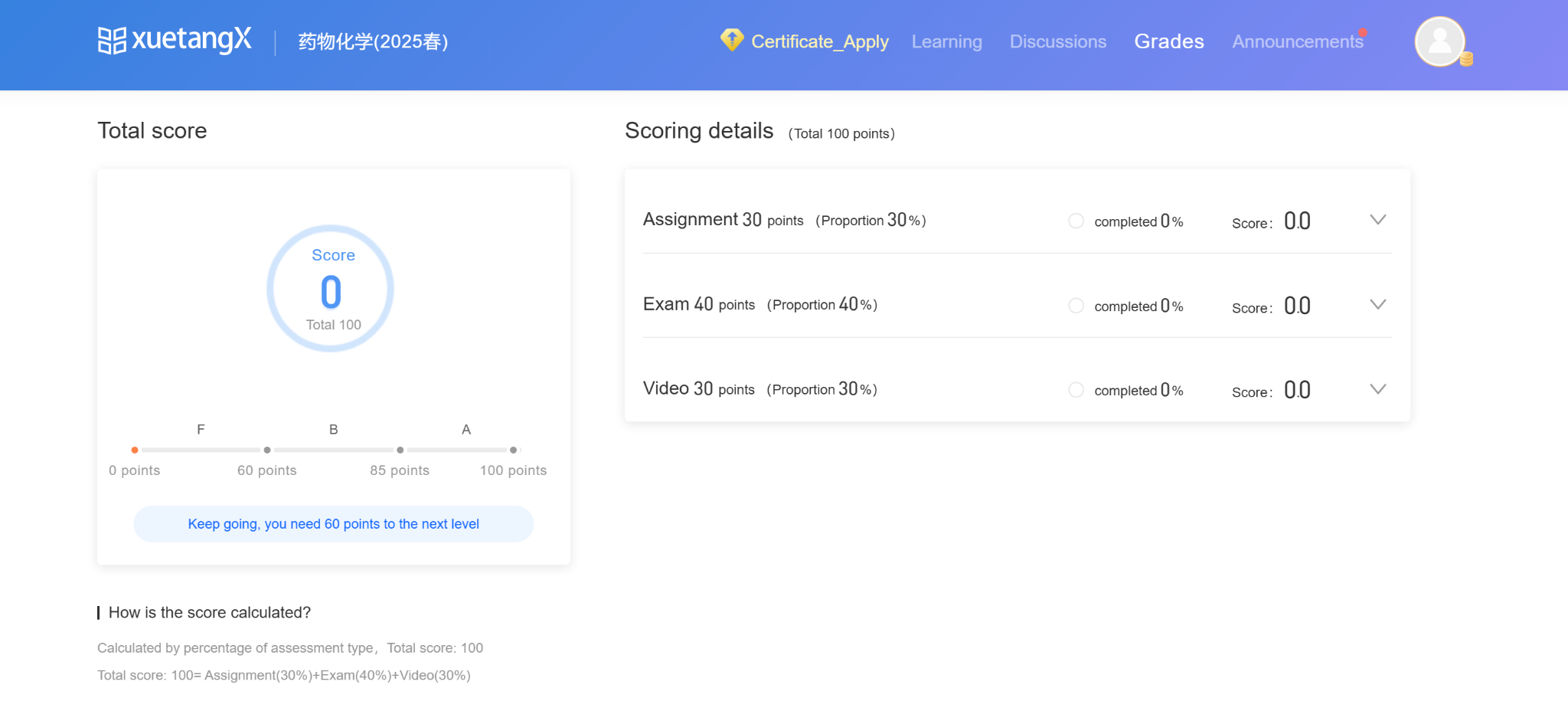
Nhóm vào từng khóa học trong XuetangX để tìm thêm thông tin về thành phần điểm assignment (điểm bài tập), video (điểm dành cho video), exam (điểm dành cho các bài kiểm tra), và type (loại khóa học), start\_date (ngày bắt đầu khóa học), end\_date (ngày kết thúc của khóa học), season (mùa khai giảng).

****

# Chiến lược gán nhãn

## Cách tính điểm trong các khóa của Xuetang





Trên nền tảng học trực tuyến **XuetangX** (một trong những MOOC nổi tiếng tại Trung Quốc, tương tự Coursera hay edX), cách tính điểm và phân loại học sinh có thể hơi khác nhau tùy vào từng khóa học, nhưng thường sẽ tuân theo một cấu trúc chung như sau:

Các khóa học thường có các thành phần điểm sau:

| **Thành phần** | **Mô tả** | **Tỉ lệ điểm (tham khảo)** |
| --- | --- | --- |
| **Bài tập (Assignment)** | Bao gồm câu hỏi trắc nghiệm, tự luận, v.v… sau mỗi video hoặc cuối chương | 0–40% |
| **Xem tài liệu (Video, Reading)** | Một bài kiểm tra lớn giữa khóa học. | 20–30% |
| **Kiểm tra Giữa kì, Cuối kì** | Bài kiểm tra quan trọng nhất, tổng kết toàn bộ khóa học. | 30–50% |
| **Tương tác (Discussion/Participation)** | Có thể được chấm điểm nếu giáo viên yêu cầu tham gia diễn đàn. | 0–10% |

Điểm tổng kết sẽ được tính theo công thức:

**Điểm cuối cùng = (Điểm bài tập × 0.x) + (Điểm kiểm tra× 0.y) + (Điểm xem tài liệu × 0.z) + (Điểm tương tác × 0.k)**

Các hệ số (x, y, z, k) phụ thuộc vào cấu trúc của từng khóa học.

## Xếp loại kết quả của học viên trong XuetangX

Trên nền tảng học trực tuyến **XuetangX**, quá trình đánh giá kết quả học tập của học viên được tổ chức theo mô hình **hai mức xếp loại chính thức**, tương ứng với trạng thái **“Đạt” (A)** và **“Không đạt” (F)**. Cách phân loại này không chỉ giúp đơn giản hóa quy trình đánh giá mà còn tạo điều kiện thuận lợi cho học viên tự theo dõi tiến độ học tập của mình trong suốt khóa học.

| **Mức xếp loại** | **Nhãn (Label)** | **Mô tả chi tiết** |
| --- | --- | --- |
| **A (Đạt)** | A | Học viên được xếp loại A khi hoàn thành khóa học với **tổng điểm đánh giá cuối cùng đạt từ 60% trở lên**. Điểm này là kết quả tổng hợp từ các thành phần như bài tập, kiểm tra, tương tác lớp học... tùy theo cấu trúc khóa học. Loại A đồng nghĩa với việc học viên **đủ điều kiện nhận chứng chỉ hoàn thành khóa học**. |
| **F (Không đạt)** | F | Học viên được xếp loại F khi **tổng điểm cuối khóa dưới 60%** hoặc không hoàn thành các nhiệm vụ bắt buộc (như bỏ lỡ bài kiểm tra chính, không nộp bài tập, không tham gia phần nội dung cốt lõi). Kết quả F đồng nghĩa với việc **không được cấp chứng chỉ**, bất kể tỷ lệ học tập đã hoàn thành là bao nhiêu. |

Một số khóa học có cách xếp loại học sinh khác.

| **Mức xếp loại** | **Nhãn (Label)** | **Mô tả chi tiết** |
| --- | --- | --- |
| **A (Giỏi)** | A | Học viên được xếp loại A khi hoàn thành khóa học với **tổng điểm đánh giá cuối cùng đạt từ 85% trở lên**. |
| **B (Đạt)** | B | Từ **60% đến dưới 85%** tổng điểm đánh giá cuối khóa. |
| **F (Không đạt)** | F | Học viên được xếp loại F khi **tổng điểm cuối khóa dưới 60%** hoặc không hoàn thành các nhiệm vụ bắt buộc (như bỏ lỡ bài kiểm tra chính, không nộp bài tập, không tham gia phần nội dung cốt lõi). Kết quả F đồng nghĩa với việc **không được cấp chứng chỉ**, bất kể tỷ lệ học tập đã hoàn thành là bao nhiêu. |

Chứng chỉ trên XuetangX là minh chứng chính thức cho việc hoàn thành khóa học, và chỉ được cấp trong những trường hợp sau:

* Tổng điểm đánh giá cuối cùng đạt từ 60% trở lên (tương ứng xếp loại B, A).
* Học viên hoàn thành đầy đủ các yêu cầu của khóa học, bao gồm nhưng không giới hạn:
  + Nộp đầy đủ các bài tập bắt buộc.
  + Tham gia và hoàn thành bài kiểm tra giữa kỳ/cuối kỳ (nếu có).
  + Đáp ứng điều kiện về thời gian học tập (nộp bài đúng hạn).
* Đối với một số khóa học nâng cao, chứng chỉ chỉ được cấp khi:
  + Học viên **xác minh danh tính** trên hệ thống (Verified Certificate).
  + Điểm đánh giá được giám sát và phê duyệt bởi đội ngũ giảng viên hoặc trợ giảng.

1. **Xếp loại học viên một số khóa học trực tiếp khác**

### Coursera

Thang điểm 5 mức độ của Coursera, Coursera là một trong những nền tảng MOOC hàng đầu, và họ sử dụng một thang điểm phổ biến để đánh giá kết quả học tập của người dùng. Thang điểm này có thể được tóm tắt như sau:

* **Distinction (Xuất sắc)**: Điểm số 85-100% (A) Người học đạt được thành tích vượt trội trong việc nắm vững kiến thức và kỹ năng của khóa học, thường trên 85%. Thành tích này cho thấy sự hiểu biết sâu sắc và khả năng ứng dụng kiến thức vào thực tiễn.
* **Merit (Giỏi)**: Điểm số 70-84% (B) Người học hoàn thành tốt các yêu cầu của khóa học, đạt từ 70% đến 84%. Đây là một thành tích tốt, thể hiện sự hiểu biết vững chắc và khả năng vận dụng kiến thức.
* **Pass (Đạt yêu cầu)**: Điểm số 50-69% (C) Người học đạt đủ yêu cầu tối thiểu của khóa học với điểm số từ 50% đến 69%. Điều này cho thấy người học đã nắm được những kiến thức cơ bản nhưng vẫn cần cải thiện thêm.
* **Fail (Không đạt)**: Dưới 50% (D) Người học không đạt được yêu cầu tối thiểu của khóa học, với điểm số dưới 50%. Kết quả này đòi hỏi phải nỗ lực học tập thêm để đạt được kiến thức cần thiết.
* **Incomplete (Chưa hoàn thành)**: (E) Người học không hoàn thành khóa học hoặc không đạt đủ điểm để được xếp loại, thường là do thiếu bài kiểm tra hoặc bài tập quan trọng.

### edX

Trên nền tảng **edX**, việc phân loại và tính điểm cho học viên được thực hiện thông qua hệ thống đánh giá của khóa học. Quá trình này có thể khác nhau tùy vào từng khóa học và cách thiết kế của giảng viên hoặc tổ chức cung cấp khóa học. Tuy nhiên, một số phương pháp phổ biến mà giảng viên sử dụng để phân loại và tính điểm học viên trên edX bao gồm:

1. Hệ thống điểm (Grading System)

* **Bài kiểm tra (Quizzes)**: Bài kiểm tra thường được chia thành nhiều câu hỏi nhỏ (multiple choice, đúng/sai, điền từ...) và được chấm điểm tự động ngay sau khi học viên hoàn thành.
* **Bài tập (Assignments)**: Những bài tập này có thể là bài viết, bài thuyết trình, hoặc các công việc lớn hơn. Giảng viên sẽ chấm điểm thủ công và phản hồi.
* **Thảo luận (Discussion Boards)**: Đôi khi, học viên sẽ tham gia thảo luận trực tuyến. Điểm có thể được cấp dựa trên chất lượng và số lượng đóng góp của học viên trong các cuộc thảo luận này.

2. Tính điểm tổng kết (Final Score Calculation)

* Mỗi hoạt động (bài kiểm tra, bài tập, thảo luận, v.v.) có thể có một trọng số điểm khác nhau. Ví dụ, bài kiểm tra có thể chiếm 40% tổng điểm, bài tập chiếm 30%, và thảo luận chiếm 30%.
* Học viên sẽ nhận điểm cho từng phần của khóa học, và tổng điểm cuối cùng sẽ được tính toán theo tỷ lệ đã quy định.

3. Các mức xếp loại (Grade Levels)

* **Pass/Fail**: Một số khóa học sẽ phân loại học viên theo hình thức đậu/loại, có thể dựa trên điểm số cuối cùng hoặc việc hoàn thành tất cả các nhiệm vụ trong khóa học.
* **Letter Grades**: Một số khóa học sẽ cấp điểm theo hệ thống chữ cái (A, B, C, D, F), hoặc có thể sử dụng hệ thống điểm số (0-100) để phân loại học viên.
* **Certificates**: Khi học viên hoàn thành khóa học và đạt điểm yêu cầu (thường là 50% hoặc cao hơn), họ có thể nhận chứng chỉ hoàn thành khóa học từ edX.

4. Chính sách về điểm tham gia

* **Điểm tham gia**: Học viên có thể được tính điểm tham gia nếu hoàn thành bài tập và các hoạt động đúng hạn, tham gia vào các cuộc thảo luận và tương tác với bạn học.
* **Điểm thưởng**: Trong một số khóa học, học viên có thể nhận điểm thưởng nếu hoàn thành các nhiệm vụ ngoài chương trình học hoặc tham gia vào các hoạt động học tập bổ sung.

5. Chấm điểm tự động và thủ công

* **Chấm điểm tự động**: Các bài kiểm tra với câu hỏi khách quan như trắc nghiệm, đúng/sai thường được chấm tự động và ngay lập tức.
* **Chấm điểm thủ công**: Bài tập viết, dự án, thuyết trình và các dạng bài tập không thể chấm điểm tự động thường sẽ được giảng viên hoặc trợ giảng chấm điểm thủ công.

6. Hệ thống điểm cụ thể của edX

* Trên edX, điểm số có thể được theo dõi thông qua **Dashboard** của học viên, nơi bạn có thể xem các điểm số của mình trong từng bài kiểm tra, bài tập, và các hoạt động khác trong khóa học.

Thang đo phân loại học viên trên edX

Thang đo phân loại học viên trong edX chủ yếu được thiết lập dựa trên điểm số cuối cùng của học viên sau khi hoàn thành tất cả các bài kiểm tra, bài tập và các nhiệm vụ trong khóa học. Một số thang đo phổ biến bao gồm:

a. Hệ thống phân loại theo điểm số (Numeric Grade Scale)

* **Điểm số** thường được chấm từ **0 đến 100** (hoặc theo một tỷ lệ tương đương). Mỗi hoạt động học tập sẽ có trọng số điểm khác nhau và tổng điểm cuối cùng sẽ được tính từ tất cả các bài kiểm tra, bài tập, và các hoạt động tham gia.

| **Thang đo học viên** | **Tỉ lệ điểm** |
| --- | --- |
| **A** (Xuất sắc) | 90 - 100 |
| **B** (Khá) | 80 - 89 |
| **C** (Trung bình) | 70 - 79 |
| **D** (Yếu) | 60 - 69 |
| **F** (Không đạt) | Dưới 60 |

* **Lý do lựa chọn kết hợp các thang điểm của các khóa học theo 5 mức độ:**
* **Phân loại mức độ thành công của người học**

Trong các khóa học trực tuyến, hiệu suất học tập của học viên có thể được phân thành nhiều mức độ khác nhau dựa trên các chỉ số như điểm số, thời gian học, mức độ hoàn thành bài tập, và hoạt động tham gia. Việc sử dụng thang điểm 5 mức độ giúp chia người học thành các nhóm khác nhau theo cách sau:

* + - * A (xuất sắc): Học viên có thành tích vượt trội.
      * B (giỏi): Học viên có kết quả tốt, hoàn thành hầu hết các yêu cầu.
      * C (trung bình): Học viên hoàn thành yêu cầu nhưng có thể còn nhiều điểm cần cải thiện.
      * D (kém): Học viên có thành tích dưới mức trung bình, chưa hoàn thành đầy đủ các yêu cầu.
      * E (rất kém): Học viên hầu như không tham gia hoặc có kết quả rất thấp.
* **Độ chính xác và đa dạng trong phân loại phân loại**

Việc phân chia thành 5 mức độ giúp tạo ra độ chi tiết cao hơn trong phân loại so với các hệ thống chỉ có 2 hoặc 3 mức. Điều này giúp cho các mô hình học máy có thể dự đoán với độ phân giải tốt hơn và dễ dàng phát hiện các mẫu học tập khác nhau:

* Hệ thống 2 mức độ (pass/fail) quá đơn giản và không phản ánh được sự đa dạng trong hành vi và kết quả học tập.
* Hệ thống 3 mức độ (tốt/trung bình/kém) thiếu độ chi tiết để phân biệt các học viên khá (giỏi và xuất sắc) hoặc những người có thành tích thấp (kém và rất kém).
* **Dễ hiểu và phổ biến**

Thang điểm 5 mức độ là thang đo phổ biến trong các hệ thống giáo dục, bao gồm cả các hệ thống giáo dục truyền thống và trực tuyến. Học viên và các giảng viên đều quen thuộc với cách thức đánh giá này, giúp việc giải thích và hiểu kết quả dễ dàng hơn.

* **Cân bằng giữa độ phân giải và tính khả thi**
  + - * **Độ phân giải cao hơn**: Một số hệ thống có thể chọn thang điểm chi tiết hơn như 7 mức hoặc 10 mức. Tuy nhiên, điều này có thể làm giảm tính khả thi khi mô hình học máy cần xử lý nhiều hơn các lớp (classes), đồng thời làm tăng sự phức tạp của dữ liệu.
      * **Tính khả thi**: 5 mức là đủ chi tiết để phân loại người học mà vẫn duy trì tính khả thi trong việc triển khai các mô hình học máy và thống kê trên các nền tảng giáo dục trực tuyến.
* **Thích hợp với các chỉ số hiệu suất**

Các khóa học trực tuyến thường theo dõi nhiều chỉ số như tổng điểm, thời gian học, số lượng bài tập hoàn thành, số lần tương tác, và điểm kiểm tra cuối kỳ. Khi sử dụng các chỉ số này, 5 mức độ cung cấp độ phân loại rõ ràng hơn giữa các học viên có hiệu suất khác nhau.

* **Chiến lược gán nhãn thuận tiện cho mô hình học máy**

Việc chọn 5 lớp nhãn cũng thuận tiện hơn cho các mô hình học máy, nhất là trong việc huấn luyện và kiểm định các mô hình phân loại (classification models). Các mô hình phân loại đa lớp (multi-class classification) thường hoạt động hiệu quả với số lượng lớp vừa phải như 5 lớp, đồng thời cho phép mô hình học cách phân biệt giữa các mức độ khác nhau của học viên dựa trên các đặc trưng như thời gian học, điểm số, và các hoạt động tương tác.

1. Xếp loại kết quả của học viên của nhóm
2. Cơ sở tính điểm

Nhóm vẫn duy trì phương pháp tính điểm như trên nền tảng **XuetangX**, vốn sử dụng mô hình đánh giá dựa trên nhiều thành phần để phản ánh toàn diện quá trình học tập của học viên. Tuy nhiên, để phục vụ mục đích gán nhãn và phân tích dữ liệu học tập, nhóm xác định rõ **3 thành phần điểm chính** như sau:

Thành phần điểm:

| **Thành phần** | **Bao gồm** | **Ghi chú thêm** |
| --- | --- | --- |
| **Assignment** | Bài tập trắc nghiệm, tự luận, câu hỏi chương | Gộp cả điểm Discussion (nếu có) |
| **Exam** | Chỉ tính điểm Final Exam | Bỏ qua điểm giữa kỳ nếu có |
| **Video** | Tỉ lệ xem video, mức độ hoàn thành tài liệu học | Gộp thêm phần Reading (nếu có) |

**Công thức tính điểm tổng kết:**

**Score = (w₁ × Assignment) + (w₂ × Exam) + (w₃ × Video)**

* Trong đó: w₁ + w₂ + w₃ = 1
* Trọng số w₁, w₂, w₃ được lấy theo từng khóa học cụ thể trên XuetangX.
* Các khóa học có thể thay đổi trọng số, nhưng nhìn chung:
  + Assignment: 20% – 40%
  + Exam (final): 30% – 50%
  + Video/Reading: 20% – 30%

1. Chiến lược xếp loại kết quả của học viên.

Nền tảng XuetangX mặc định chia thành 2 mức:

* **A**: Đạt ≥ 60%
* **F**: Không đạt < 60%

Tuy nhiên, nhóm mở rộng thành **5 mức** để:

* Phân tích tốt hơn chất lượng học tập ở các cấp độ khác nhau
* Phục vụ gán nhãn mô hình học máy
* Vẫn đảm bảo **tính tương thích với hệ thống gốc** bằng cách **giữ nguyên ngưỡng 60% là điều kiện đạt**

Nguyên tắc gán nhãn:

* **Mốc 60%** là ranh giới rõ ràng để phân biệt học viên **đạt** và **không đạt**.
* Các học viên có điểm **< 60%** đều bị gán nhãn là **Fail**, không đủ điều kiện nhận chứng chỉ.
* Các học viên có điểm **≥ 60%** sẽ được chia tiếp thành **4 mức độ thành công**, phản ánh mức độ hoàn thành từ cơ bản đến xuất sắc.

Bảng phân loại chi tiết:

| **Nhãn** | **Điểm số (Score)** | **Mức độ hoàn thành** |
| --- | --- | --- |
| **A (Excellent)** | 85 ≤ Score ≤ 100 | Hoàn thành xuất sắc; thành thạo kiến thức và kỹ năng. |
| **B (Good)** | 70 ≤ Score < 85 | Nắm vững nội dung; vượt yêu cầu cơ bản. |
| **C (Pass)** | 60 ≤ Score < 70 | Đủ điều kiện để nhận tín chỉ của khóa học. Đáp ứng phần lớn mục tiêu học tập. |
| **D (Fail)** | 30 ≤ Score < 60 | Không đạt để nhận tín chỉ. |
| **E (Inactive)** | 0 ≤ Score < 30 | Học viên hầu như không hoạt động. |

Dữ liệu sau khi gán nhãn

