**TRƯỜNG ĐẠI HỌC SƯ PHẠM – ĐẠI HỌC ĐÀ NẴNG**

**KHOA VẬT LÍ**

****

**ĐỒ ÁN CHUYÊN NGÀNH**

**ĐỀ TÀI: PHÂN TÍCH DỮ LIỆU VỚI PYTHON**

**Sinh viên thực hiện:**

Đinh Văn Lộc 21CNTT3

**Người hướng dẫn:** GV. Vũ Thị Trà

**Đà Nẵng, 12/2023**

**LỜI CẢM ƠN**

Đầu tiên, em xin chân thành cảm ơn các thầy các cô trong khoa Tin học đã trang bị những kiến thức cho em trong suốt quá trình học tập tại Trường Đại học Sư phạm – Đại học Đà Nẵng vừa qua. Chính nhờ công lao giảng dạy, chỉ bảo tận tình của các thầy các cô mà em mới có được những kiến thức chuyên ngành công nghệ thông tin để có thể thực hiện tiếp chặng đường học tập, vận dụng và sáng tạo ra những sản phẩm hữu ích góp phần phục vụ các lĩnh vực khác nhau.

Thứ hai, chúng em xin chân thành cảm ơn TS. Vũ Thị Trà người đã tận tình hướng dẫn, chỉ bảo, góp ý, hỗ trợ cho chúng em trong suốt quá trình thực hiện đề tài này*.* Đó là những góp ý hết sức quý báu không chỉ trong quá trình thực hiện đồ án này mà còn là hành trang tiếp bước cho em trong quá trình học tập và lập nghiệp sau này.

Mặc dù đã có nhiều cố gắng để hoàn thành đồ án nhưng trong phạm vi và khả năng cho phép, chắc chắn đồ án không tránh khỏi những thiếu sót. Em rất mong nhận được sự thông cảm, góp ý và tận tình chỉ bảo của quý thầy cô.

Chúng em xin chân thành cảm ơn!

*Đà Nẵng, tháng 12 năm 2023* **Tác giả**

**MỤC LỤC**

[MỞ ĐẦU 5](#_Toc154371179)

[1. Lý do chọn đề tài 5](#_Toc154371180)

[2. Mục tiêu nghiên cứu 5](#_Toc154371181)

[3. Nhiệm vụ nghiên cứu 5](#_Toc154371182)

[4. Đối tượng và phạm vi nghiên cứu 6](#_Toc154371183)

[4.1 Đối tượng nghiên cứu 6](#_Toc154371184)

[4.2 Phạm vi nghiên cứu 6](#_Toc154371185)

[CHƯƠNG 1: CƠ SỞ LÍ THUYẾT 7](#_Toc154371186)

[1. cURL 7](#_Toc154371187)

[1.1 Khái niệm 7](#_Toc154371188)

[1.2 Ưu điểm 7](#_Toc154371189)

[1.3 Nhược điểm 7](#_Toc154371190)

[2. String Manipulation trong Python 8](#_Toc154371191)

[2.1 Khái niệm 8](#_Toc154371192)

[2.2 Một số cách thức làm sạch dữ liệu khác 8](#_Toc154371193)

[3. Python 8](#_Toc154371194)

[4. Thư viện Matplotlib 9](#_Toc154371195)

[4.1 Khái niệm 9](#_Toc154371196)

[4.2 Đặc điểm chính 9](#_Toc154371197)

[CHƯƠNG 2: KHAI PHÁ DỮ LIỆU ĐIỂM THI THPT CỤM ĐÀ NẴNG 2023 10](#_Toc154371198)

[1. Quy trình khai phá dữ liệu 10](#_Toc154371199)

[1.1 Xác định bài toán 10](#_Toc154371200)

[1.2 Thu thập dữ liệu 10](#_Toc154371201)

[1.3 Xử lí và làm sạch dữ liệu 10](#_Toc154371202)

[1.4 Phân tích dữ liệu 10](#_Toc154371203)

[1.5 Đánh giá và lựa chọn mô hình 10](#_Toc154371204)

[2. Xác định bài toán 11](#_Toc154371205)

[3. Thu thập dữ liệu 11](#_Toc154371206)

[3.1 Nguồn dữ liệu 11](#_Toc154371207)

[3.2 Phương pháp thu thập dữ liệu 11](#_Toc154371208)

[3.3 Phạm vi và loại dữ liệu thu thập 12](#_Toc154371209)

[3.3.1 Phạm vi 12](#_Toc154371210)

[3.3.2 Loại dữ liệu 12](#_Toc154371211)

[4. Xử lí và làm sạch dữ liệu 13](#_Toc154371212)

[4.1 Dữ liệu tiền xử lí 13](#_Toc154371213)

[4.2 Làm sạch dữ liệu 14](#_Toc154371214)

[4.2.1 Loại bỏ ký tự đặc biệt 14](#_Toc154371215)

[4.2.2 Loại bỏ thẻ HTML 14](#_Toc154371216)

[4.2.3 Chọn thông tin liên quan và ý nghĩa 15](#_Toc154371217)

[4.2.4 Mã hóa chuỗi 17](#_Toc154371218)

[4.2.5 Xử lí dữ liệu 18](#_Toc154371219)

[5. Xây dựng mô hình 18](#_Toc154371220)

[5.1 Biểu đồ tròn – số môn học sinh thi 19](#_Toc154371221)

[5.2 Biểu đồ cột 20](#_Toc154371222)

[CHƯƠNG 3: ĐÁNH GIÁ MÔ HÌNH 26](#_Toc154371223)

[1. Biểu đồ điểm trung bình theo số lượng môn thi 26](#_Toc154371224)

[1.1 Biểu đồ 26](#_Toc154371225)

[1.2 Đánh giá 26](#_Toc154371226)

[2. Biểu đồ số học sinh bỏ thi hoặc không đăng kí mỗi môn 26](#_Toc154371227)

[2.1 Biểu đồ 26](#_Toc154371228)

[2.2 Đánh giá 27](#_Toc154371229)

[3. Biểu đồ số môn học sinh thi 27](#_Toc154371230)

[3.1 Biểu đồ 27](#_Toc154371231)

[3.2 Đánh giá 28](#_Toc154371232)

[4. Phổ điểm 28](#_Toc154371233)

[4.1 Biểu đồ 28](#_Toc154371234)

[4.2 Đánh giá 32](#_Toc154371235)

[KẾT LUẬN 34](#_Toc154371236)

[TÀI LIỆU THAM KHẢO 35](#_Toc154371237)

# MỞ ĐẦU

## 1. Lý do chọn đề tài

Trong thời đại số hóa ngày càng phát triển, dữ liệu từ Internet đã trở thành nguồn tài nguyên vô cùng quý báu cho việc nghiên cứu và phân tích. Trong khu vực giáo dục, dữ liệu về điểm thi tốt nghiệp THPT không chỉ là thông tin quan trọng về kết quả học tập mà còn là cơ sở để hiểu rõ hơn về chất lượng giáo dục cũng như tiềm năng phát triển của các cụm thi trên địa bàn.

Đồ án này tập trung vào việc khai thác và phân tích dữ liệu điểm thi tốt nghiệp THPT tại cụm Đà Nẵng năm 2023 từ nguồn Internet. Mục tiêu là tìm hiểu các phương pháp làm sạch dữ liệu, khai phá tri thức và áp dụng các kỹ thuật phân tích dữ liệu sẽ giúp chúng ta có cái nhìn sâu hơn về hình ảnh chung về chất lượng giáo dục.

## 2. Mục tiêu nghiên cứu

Nắm vững các phương pháp làm sạch dữ liệu (data cleaning) để xử lý dữ liệu điểm thi tốt nghiệp THPT từ nguồn Internet.

Áp dụng các kỹ thuật khai phá dữ liệu (data mining) để tìm ra thông tin quan trọng từ dữ liệu điểm thi.

Phân tích và trình bày các mô hình, biểu đồ, số liệu minh họa giúp hiểu rõ hơn về đặc điểm, xu hướng và mối liên hệ giữa các yếu tố với điểm thi.

## 3. Nhiệm vụ nghiên cứu

Thu thập dữ liệu điểm thi tốt nghiệp THPT cụm Đà Nẵng năm 2023 từ các nguồn trên Internet.

Tiến hành làm sạch dữ liệu, xử lý các giá trị thiếu, dữ liệu không đồng nhất, và loại bỏ nhiễu.

Áp dụng các kỹ thuật khai phá dữ liệu để phân tích đặc điểm, xu hướng và quy luật từ dữ liệu điểm thi.

Tạo các biểu đồ, mô hình dữ liệu minh họa để trình bày và diễn giải kết quả phân tích.

## 4. Đối tượng và phạm vi nghiên cứu

### 4.1 Đối tượng nghiên cứu

Đối tượng nghiên cứu của đồ án này bao gồm dữ liệu điểm thi tốt nghiệp THPT của các thí sinh tại cụm Đà Nẵng năm 2023.

### 4.2 Phạm vi nghiên cứu

Phạm vi nghiên cứu của đồ án này tập trung vào việc phân tích dữ liệu điểm thi tốt nghiệp THPT tại cụm Đà Nẵng năm 2023 từ nguồn dữ liệu trên Internet. Phạm vi cũng bao gồm các phương pháp xử lý dữ liệu, khai phá tri thức, và các phương pháp phân tích dữ liệu để hiểu rõ hơn về chất lượng giáo dục và các yếu tố ảnh hưởng đến kết quả học tập.

# CHƯƠNG 1: CƠ SỞ LÍ THUYẾT

## 1. cURL

### 1.1 Khái niệm

cURL là một command line tool và thư viện được sử dụng để truyền dữ liệu thông qua nhiều giao thức khác nhau (như HTTP, HTTPS, FTP, FTPS, SCP, SFTP, FILE, IMAP, SMTP, POP3, RTSP và RTMP…).

Với giao thức HTTP, cURL hỗ trợ việc gửi đi một request với tất cả các phương thức hiện có và việc truyền dữ liệu qua giao thức HTTPS. Ta có thể dùng cURL trong Php để GET và POST dữ liệu.

### 1.2 Ưu điểm

Dễ Sử Dụng: Curl rất dễ dàng sử dụng, đặc biệt trong việc thực hiện các yêu cầu HTTP cơ bản như tải trang web, tải tập tin, hoặc gửi dữ liệu form.

Hỗ Trợ Nhiều Giao Thức: Curl hỗ trợ nhiều giao thức, bao gồm HTTP, HTTPS, FTP, SFTP, SCP, và nhiều giao thức khác, giúp bạn tải về dữ liệu từ nhiều nguồn khác nhau.

Đa Dạng Các Tùy Chọn và Cấu Hình: Curl cung cấp nhiều tùy chọn và cấu hình linh hoạt, cho phép bạn điều chỉnh các tham số như User-Agent, header, cookies, và proxy.

### 1.3 Nhược điểm

Không Thân Thiện với Python: Curl không phải là một thư viện được tích hợp sẵn trong Python, điều này có thể làm cho việc tích hợp và tương tác với các chương trình Python khó khăn hơn.

Khó Điều Chỉnh Trong Quá Trình Thực Thi: Curl không linh hoạt trong việc xử lý dữ liệu trả về trực tiếp. Bạn có thể gặp khó khăn khi muốn xử lý dữ liệu sau khi tải về từ trang web.

Yêu Cầu Kiến Thức Về Cú Pháp Command-Line: Để sử dụng curl hiệu quả, bạn cần có kiến thức về cú pháp command-line, điều này có thể là rào cản đối với những người không quen thuộc với môi trường dòng lệnh.

## 2. String Manipulation trong Python

### 2.1 Khái niệm

String manipulation là quá trình thay đổi, xử lý và tác động lên chuỗi ký tự trong ngôn ngữ lập trình Python.

Chuỗi (string) là một dạng dữ liệu quan trọng trong Python, chứa một chuỗi các ký tự, số hoặc ký tự đặc biệt được bao quanh bởi dấu ngoặc đơn ('') hoặc dấu nháy kép ("").

### 2.2 Một số cách thức làm sạch dữ liệu khác

BeautifulSoup: là một thư viện mạnh mẽ trong Python giúp phân tích và trích xuất thông tin từ các trang web HTML/XML

Regex: là một công cụ mạnh mẽ cho việc tìm kiếm và xử lý chuỗi, có thể sử dụng Regex trong Python để tìm kiếm và thay thế các mẫu chuỗi không mong muốn, loại bỏ các đoạn văn bản không cần thiết, hoặc trích xuất thông tin cần thiết từ dữ liệu lấy về từ Internet.

Pandas: Nếu dữ liệu bạn thu thập có cấu trúc như CSV hoặc bảng, bạn có thể sử dụng thư viện Pandas để đọc dữ liệu và thực hiện các thao tác làm sạch dữ liệu như loại bỏ cột không cần thiết, xử lý dữ liệu trùng lặp, hoặc xử lý giá trị thiếu.

## 3. Python

#### 3.1 Định nghĩa

Python là một ngôn ngữ lập trình được sử dụng rộng rãi trong các ứng dụng web, phát triển phần mềm, khoa học dữ liệu và máy học (ML). Các nhà phát triển sử dụng Python vì nó hiệu quả, dễ học và có thể chạy trên nhiều nền tảng khác nhau***.***

#### 3.2 Ưu điểm khi sử dụng Python trong khai phá dữ liệu

Thư viện phong phú cho phép thực hiện các công việc khai phá dữ liệu từ việc xử lý, phân tích đến xây dựng mô hình dễ dàng và hiệu quả.

Python được biết đến với cú pháp đơn giản và dễ đọc, giúp người mới học và người có kinh nghiệm dễ dàng tiếp cận. Điều này giúp giảm thời gian học và triển khai các dự án khai phá dữ liệu.

#### 

## 4. Thư viện Matplotlib

### 4.1 Khái niệm

Matplotlib là một thư viện Python mạnh mẽ dùng để tạo đồ thị, biểu đồ và hình vẽ một cách đa dạng và linh hoạt. Được phát triển bởi John D. Hunter, Matplotlib được thiết kế để vẽ đồ thị và biểu đồ một cách dễ dàng và linh hoạt, phục vụ cho nhu cầu trực quan hóa dữ liệu trong phân tích và khai phá dữ liệu

### 4.2 Đặc điểm chính

Tính linh hoạt và đa dạng: Matplotlib cung cấp các hàm và công cụ để tạo ra nhiều loại đồ thị và biểu đồ khác nhau như biểu đồ đường, cột, hình tròn, histogram, scatter plot, heatmap, và nhiều loại biểu đồ phức tạp khác.

Dễ sử dụng: Với cú pháp đơn giản và dễ hiểu, Matplotlib là một công cụ linh hoạt cho việc tạo các hình vẽ và biểu đồ theo ý muốn của người dùng.

Tích hợp tốt với các công cụ khác: Matplotlib có thể tích hợp dễ dàng với các thư viện và công cụ khác trong hệ sinh thái khoa học dữ liệu của Python như NumPy, Pandas, Seaborn và Jupyter Notebook.

Hỗ trợ đa nền tảng: Matplotlib có khả năng hoạt động trên nhiều nền tảng khác nhau bao gồm Windows, Linux, MacOS và có thể tạo ra các hình ảnh với định dạng khác nhau như PNG, JPEG, SVG, PDF, và nhiều định dạng khác.

# CHƯƠNG 2: KHAI PHÁ DỮ LIỆU ĐIỂM THI THPT CỤM ĐÀ NẴNG 2023

## 1. Quy trình khai phá dữ liệu

## *1.1 Xác định bài toán*

Xác định mục tiêu cụ thể muốn đạt được thông qua việc khai phá dữ liệu. Điều này bao gồm việc hiểu rõ vấn đề cần giải quyết, câu hỏi cần trả lời hoặc mục tiêu kinh doanh cụ thể.

## *1.2 Thu thập dữ liệu*

Thu thập dữ liệu từ các nguồn khác nhau như cơ sở dữ liệu, tệp tin, web, sensor, và các nguồn dữ liệu khác phù hợp với mục tiêu.

## *1.3 Xử lí và làm sạch dữ liệu*

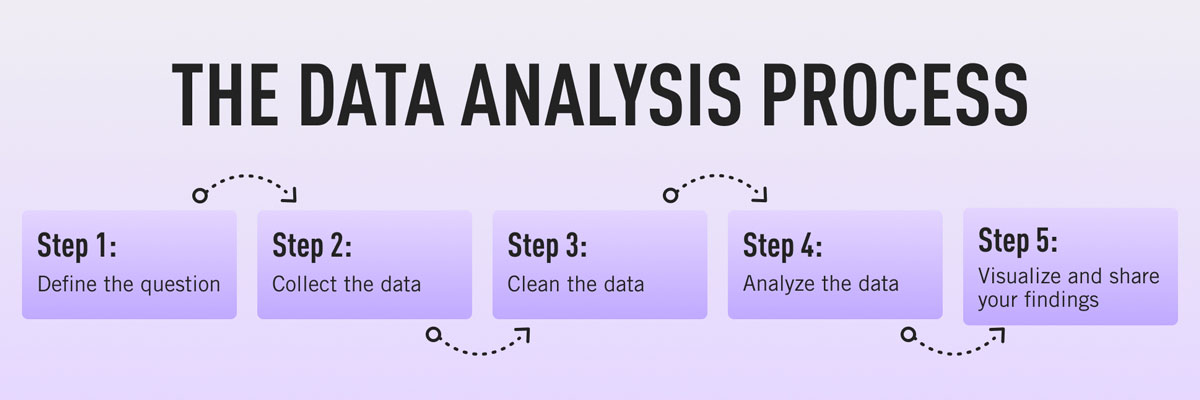
Bước này bao gồm xử lý và làm sạch dữ liệu để chuẩn hóa, loại bỏ giá trị thiếu, loại bỏ nhiễu, điều chỉnh định dạng và chuẩn hóa dữ liệu để chuẩn bị cho quá trình khai phá.

## *1.4 Phân tích dữ liệu*

Áp dụng các phương pháp và thuật toán khai phá dữ liệu để tìm ra mẫu, quy luật, thông tin tiềm ẩn và mối quan hệ trong dữ liệu. Các phương pháp có thể bao gồm phân tích thống kê, học máy, kỹ thuật khai phá dữ liệu, và các phương pháp khác.

## *1.5 Đánh giá và lựa chọn mô hình*

Đánh giá kết quả từ việc phân tích dữ liệu, lựa chọn mô hình tốt nhất dựa trên mục tiêu của bạn



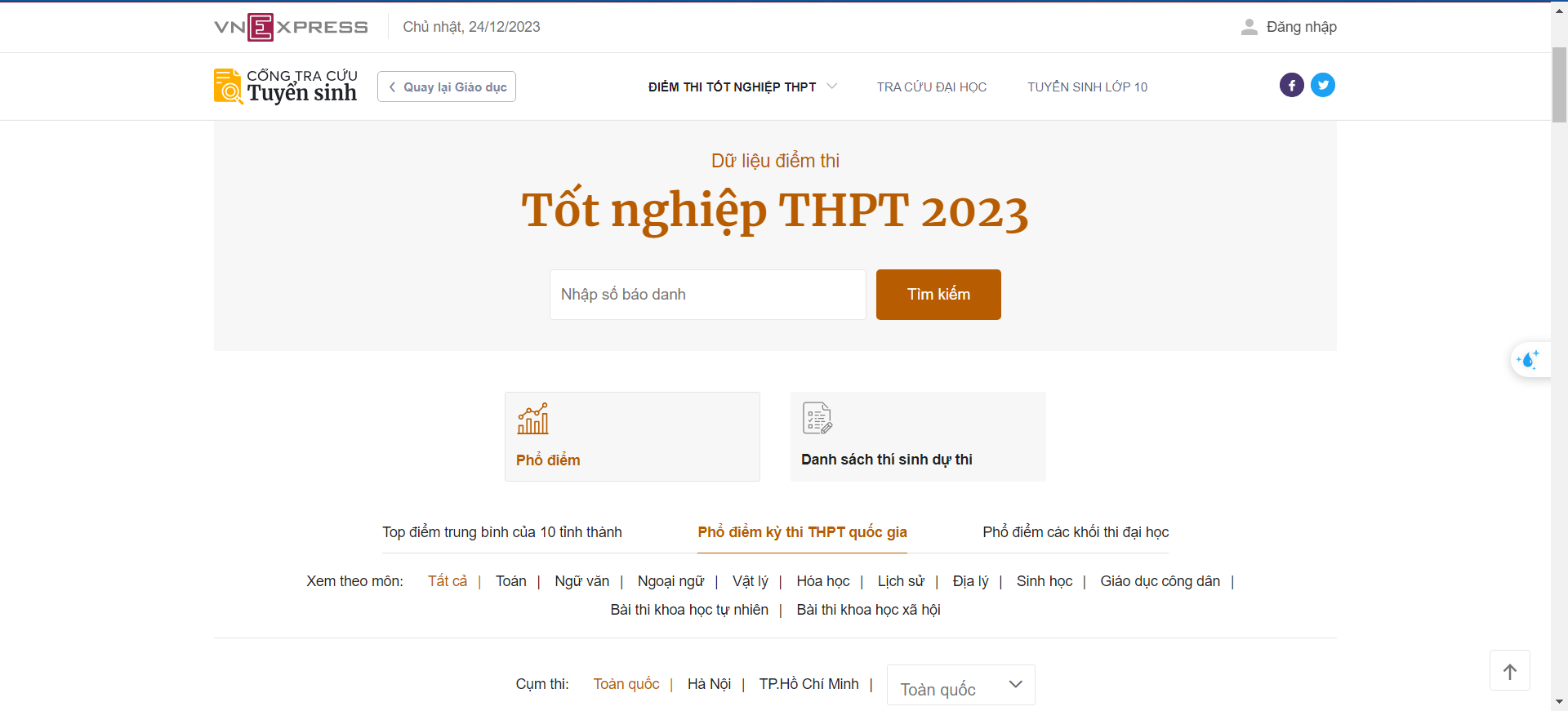
## 2. Xác định bài toán

Mục tiêu là thực hiện khai phá dữ liệu điểm thi THPT quốc gia 2023 tại cụm Đà Nẵng để hiểu rõ hơn về phổ điểm, mối liên hệ giữa các môn học, và ảnh hưởng của mỗi môn học đối với tổng điểm thi.

## 3. Thu thập dữ liệu

## *3.1 Nguồn dữ liệu*

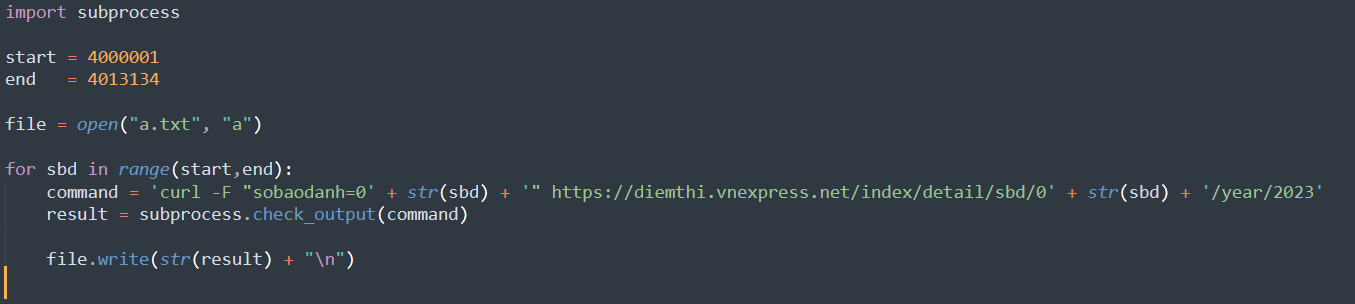
Dữ liệu được thu thập từ trang web https://diemthi.vnexpress.net/, nơi cung cấp thông tin về điểm thi THPT quốc gia.



Hình 1. Website VNExpress

## *3.2 Phương pháp thu thập dữ liệu*

Sử dụng cURL để lấy dữ liệu từ trang web. Trang web cung cấp thông tin điểm thí sinh dựa trên số báo danh và năm thi.



Hình 2. Mã nguồn thu thập dữ liệu

subporocess: Là một thư viện Python cho phép tạo, chạy và quản lý các tiến trình hệ thống từ chương trình Python của bạn.

curl: Là một công cụ dùng để truy cập và tải xuống dữ liệu từ URL thông qua giao thức HTTP, cũng được sử dụng để gửi dữ liệu lên server.

-F "sobaodanh=0' + str(sbd) + '": Tùy chọn -F trong cURL là để gửi dữ liệu dưới dạng form data. Trong trường hợp này, "sobaodanh=0' + str(sbd) + '" là thông tin số báo danh của thí sinh.

https://diemthi.vnexpress.net/index/detail/sbd/0' + str(sbd) + '/year/2023: Đây là URL của trang web mà chúng ta đang truy cập để lấy thông tin về điểm thi dựa trên số báo danh của thí sinh. Phần 0' + str(sbd) + '/year/2023 trong URL thay thế số báo danh để truy cập vào trang thông tin điểm thi cụ thể của từng thí sinh.

## *3.3 Phạm vi và loại dữ liệu thu thập*

### 3.3.1 Phạm vi

Dữ liệu được thu thập từ kết quả điểm thi THPT quốc gia 2023 của thí sinh tại cụm 04 – Đà Nẵng. Số báo danh được trải dài từ số báo danh 04000001 cho đến 04013133, giúp thu thập thông tin điểm thi của các thí sinh trong phạm vi này.

### 3.3.2 Loại dữ liệu

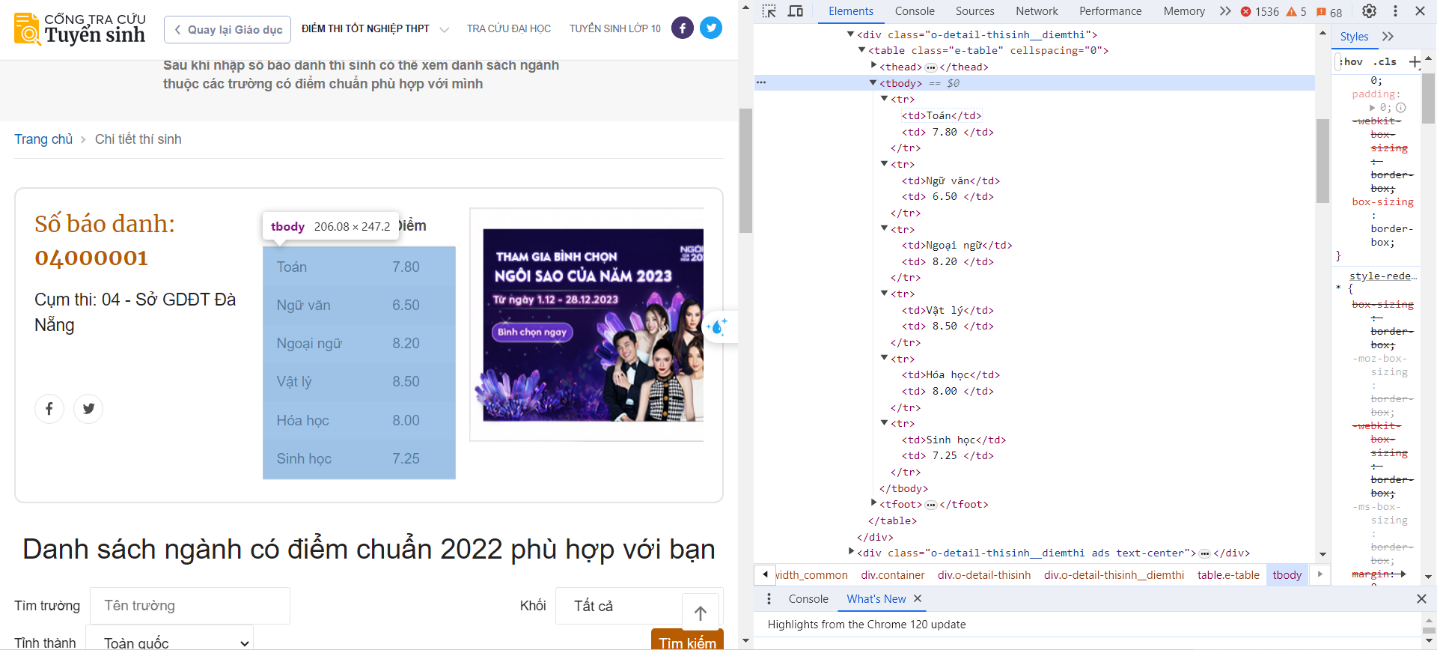
Số báo danh của thí sinh: dùng để xác định danh tính từng thí sinh.

Điểm thi của từng thí sinh trong các môn học, bao gồm các môn thuộc các khối thi như Toán, Ngữ Văn, Lịch Sử, Địa Lý, Giáo Dục Công Dân, Sinh Học, Vật Lý, Hóa Học và Ngoại Ngữ.

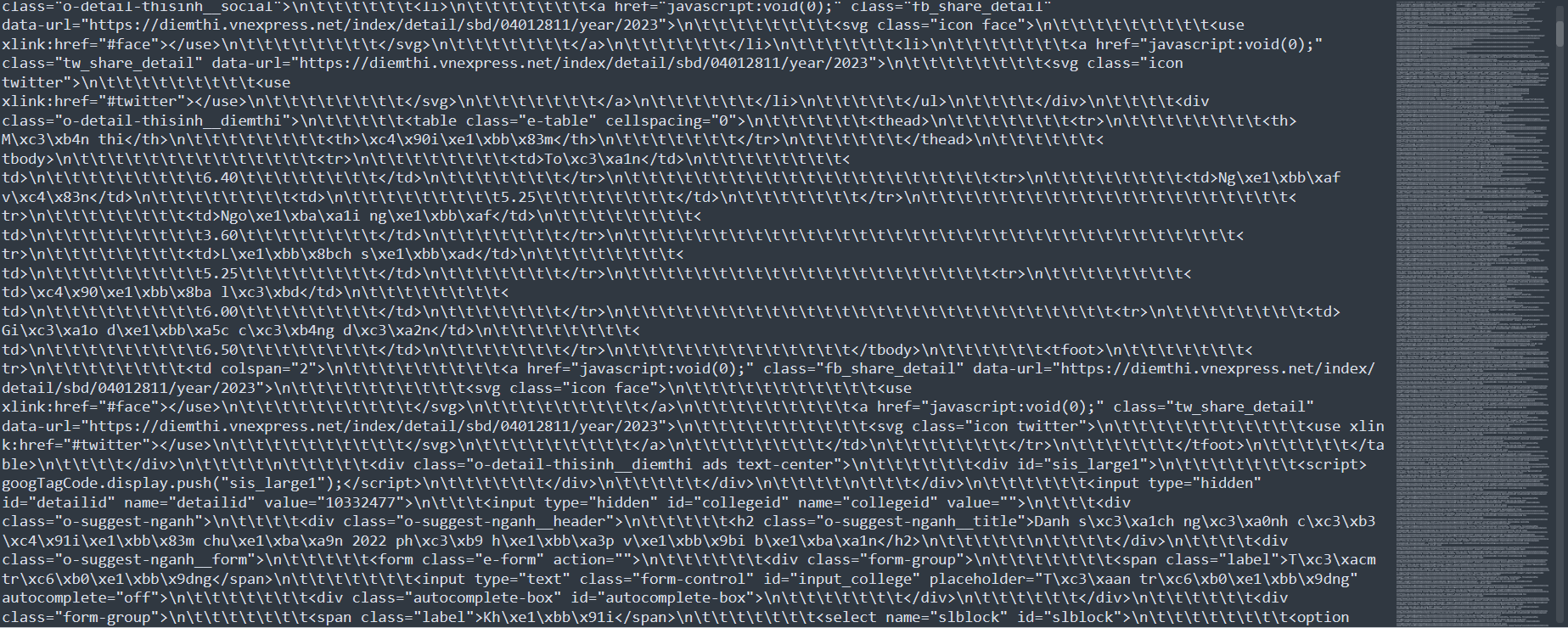
## 4. Xử lí và làm sạch dữ liệu

### 4.1 Dữ liệu tiền xử lí

Trước khi rút trích nội dung trang web, đây là nội dung ta cần đặc biệt lưu ý trước khi làm sạch dữ liệu. Gồm nội dung thẻ môn thi và điểm, phòng trường hợp thí sinh không thi môn nào nhưng vẫn có đăng ký thi.

Hình 3. Kiểm tra elements của website (dữ liệu trước khi trích xuất)

Sau khi rút trích dữ liệu website về file, các định dạng chữ bị biến đổi thành chuỗi Unicode (UTF-8). Khi mã hóa một chuỗi để lưu trữ hoặc truyền qua mạng, nó có thể được mã hóa theo các bảng mã như UTF-8.

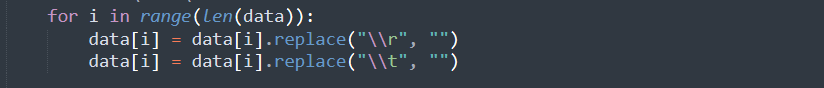


Hình 4. Dữ liệu được trích xuất về file

### 4.2 Làm sạch dữ liệu

### 4.2.1 Loại bỏ ký tự đặc biệt

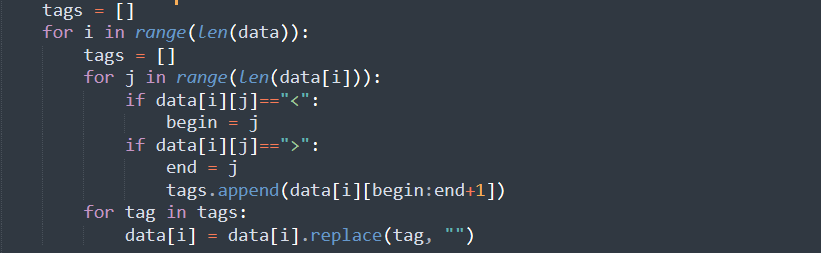
Loại bỏ ký tự đặc biệt \r (Carriage Return) và \t (Tab) khỏi dữ liệu. Đây là các ký tự thường dùng được sử dụng để thực hiện công việc định dạng, tạo khoảng trống, hoặc điều chỉnh bố cục trong mã nguồn HTML.



Hình 5. Sử dụng hàm replace để loại bỏ ký tự đặc biệt

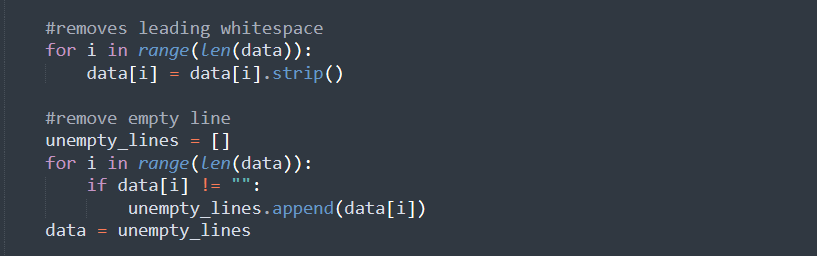
### 4.2.2 Loại bỏ thẻ HTML

Loại bỏ các thẻ trong HTML, bằng cách tạo một danh sách các thẻ mà vòng lặp tìm được thông qua quy luật so sánh chuỗi khi duyệt hết dữ liệu. Rồi sử dụng hàm replace để loại bỏ các chuỗi giống với thẻ trong danh sách.



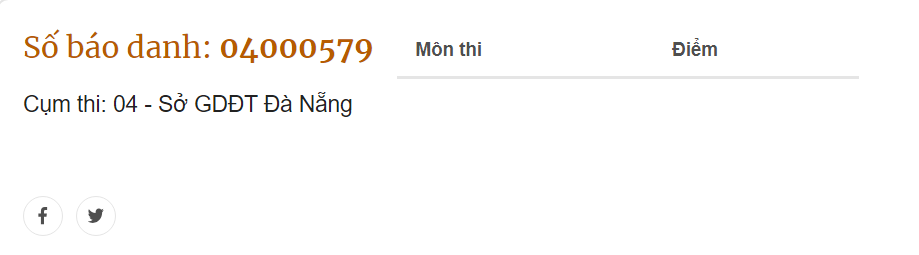
Hình 6. Loại bỏ thẻ HTML trong dữ liệu

Loại bỏ các dấu cách đầu dòng thông qua hàm strip và các dòng trống thông qua duyệt các chuỗi và bỏ qua dòng trống.

Hình 7. Loại bỏ dấu cách đầu dòng và các dòng trống

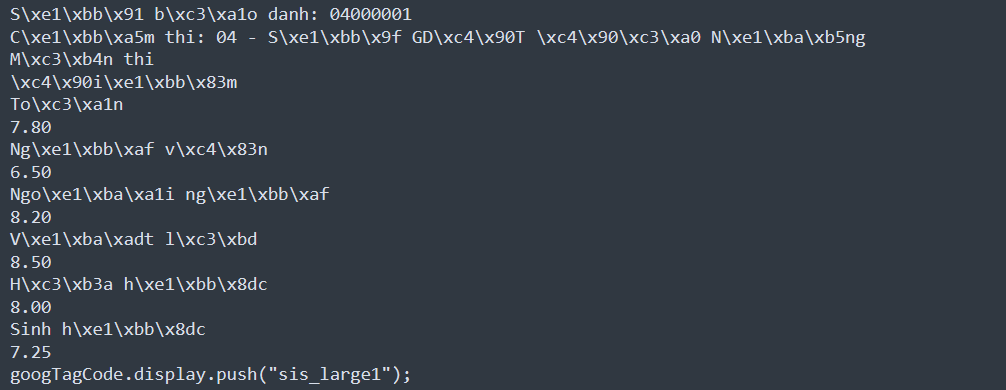
### 4.2.3 Chọn thông tin liên quan và ý nghĩa

Theo thông thường thì ta có thể duyệt tìm các điểm thi của một thí sinh rồi trích ra vào file dữ liệu sạch. Nhưng cũng có một số trường hợp các thí sinh chỉ đăng ký thi mà không tới thi.



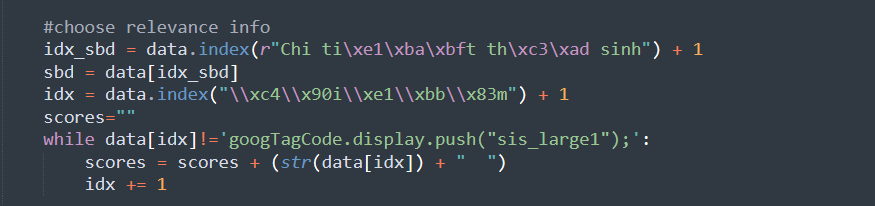
Hình 8. Thông tin thí sinh 04000579

Vì thế ta sẽ sử dụng hàm index để tìm ra chuỗi đứng trước số báo danh và cả sau điểm thi làm cột mốc để duyệt lấy nội dung. Đây là hình ảnh dữ liệu được làm sạch với các bước trên, giờ số báo danh và điểm số đã hiện ra một cách rõ ràng.



Hình 9. Sau khi làm sạch với các bước trên

Chọn hai chuỗi chưa được mã hóa “Chi tiết thí sinh” và “Điểm” là mốc đầu và “googTagCode.display.push("sis\_large1");” làm mốc sau để duyệt lấy điểm.



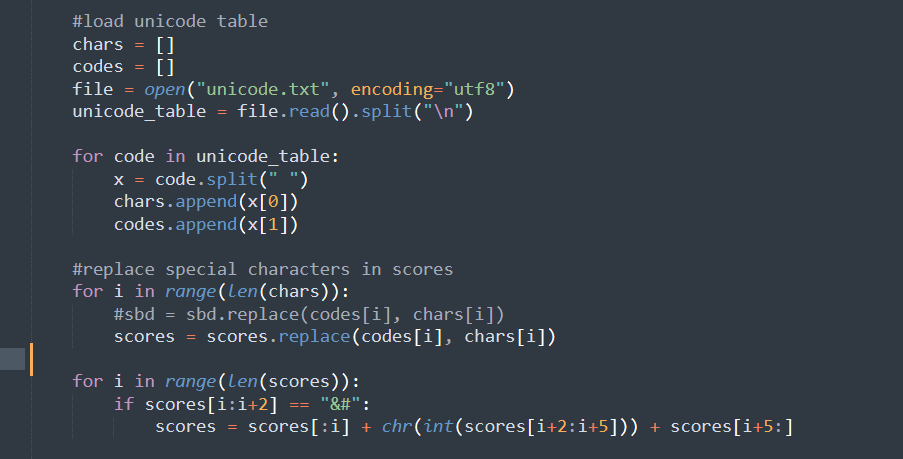
Hình 10. Mã nguồn duyệt chọn thông tin liên quan

4.2.4 Mã hóa chuỗiCác chuỗi nội dung trong file dữ liệu bị mã hóa thành dưới định dạng UTF-8. Ta sẽ tạo một file tên là unicode chứa các chuỗi mã hóa và các chữ tiếng việt thay thế tương ứng.



Hình 11. Nội dung file unicode.txt

Ta cho các ký tự tiếng việt vào danh sách chars và các chuỗi chưa được mã hóa sang danh sách codes. Sau đó ta duyệt dự liệu và thay thế các chuỗi chưa được mã hóa bằng hàm replace. Ở một số python cũ, có khả năng sẽ có ký tự “&#” này xuất hiện thì ta có thể cắt bằng indexing của python.



Hình 12. Mã nguồn mã hóa chuỗi

### 4.2.5 Xử lí dữ liệu

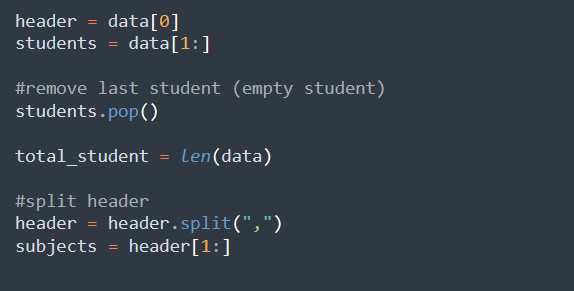
Ta thêm dữ liệu trong danh sách scores\_list theo thứ tự header được quy định sẵn ở khi tạo file dữ liệu sạch. Ở những môn không có điểm thì ta điền giá trị -1.

******

Hình 13. Đưa danh sách số báo danh và điểm vào dữ liệu sạch

## 5. Xây dựng mô hình

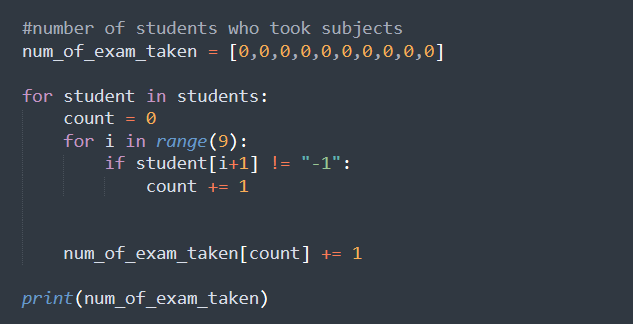
Quy tắc chung, sau khi mở file dữ liệu sạch. Ta gán các nhãn vào biến header và dữ liệu tất cả thí sinh vào biến subjects.

****

Hình 14. Nhãn và dữ liệu thí sinh

### 5.1 Biểu đồ tròn – số môn học sinh thi

Biểu đồ tròn (Pie chart) thường được sử dụng trong việc trình bày phần trăm hoặc tỷ lệ của các phần thành của một tập dữ liệu so với tổng thể.



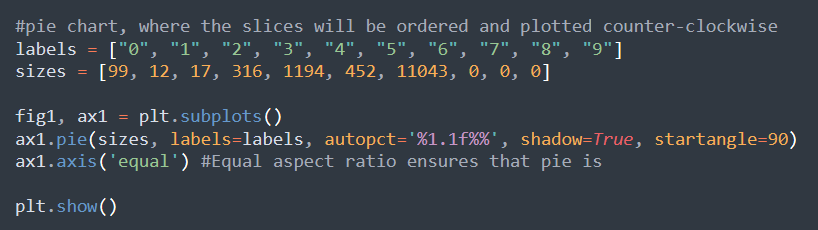
Hình 15. Mã nguồn đếm số môn học sinh thi

num\_of\_exam\_taken là một list gồm 10 phần tử ban đầu đều là 0, dùng để lưu trữ số lượng học sinh theo số môn học họ tham gia. Index từ 0 đến 9 biểu thị số lượng môn học tham gia từ 0 đến 9.

Trong vòng lặp bên trong for i in range(9):, code kiểm tra điểm của học sinh ở các môn học từ 1 đến 9. Nếu điểm khác -1 (giả sử -1 là điểm của học sinh không tham gia môn học đó), biến count sẽ được tăng lên 1.

Tiếp theo, num\_of\_exam\_taken[count] sẽ được tăng lên 1, tức là tại index tương ứng với số môn học mà học sinh tham gia, giá trị tương ứng trong list num\_of\_exam\_taken sẽ được tăng lên để đếm số lượng học sinh tham gia số môn học đó.

Cuối cùng, kết quả sẽ được in ra bởi lệnh print(num\_of\_exam\_taken), hiển thị số lượng học sinh tham gia từng số môn học khác nhau (từ 0 đến 9 môn học)



Hình 16. Mã nguồn xây dựng biểu đồ

fig1, ax1 = plt.subplots(): Tạo ra một subplot để chứa biểu đồ tròn. ax1 sẽ là trục để vẽ biểu đồ.

ax1.pie(sizes,labels=labels,autopct='%1.1f%%',shadow=True,startangle=90): Tạo biểu đồ tròn với sizes là dữ liệu, labels là nhãn, autopct để hiển thị phần trăm trên mỗi phần, shadow=True để thêm hiệu ứng bóng đổ, và startangle=90 để bắt đầu vẽ biểu đồ từ góc 90 độ.

ax1.axis('equal'): Dòng này giúp đảm bảo biểu đồ tròn sẽ có tỉ lệ chiều cao và chiều rộng bằng nhau, tạo ra một hình tròn hoàn chỉnh thay vì hình bầu dục.

### 5.2 Biểu đồ cột

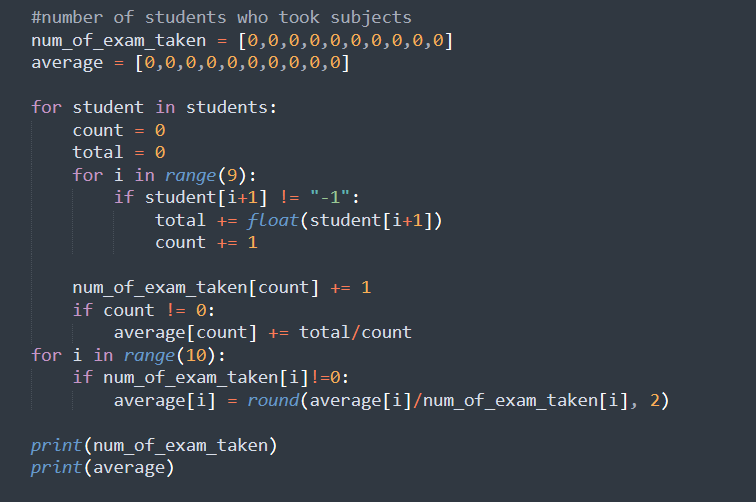
Biểu đồ cột (Bar Chart) là một dạng biểu đồ thường được sử dụng để trực quan hóa và so sánh dữ liệu giữa các danh mục hoặc biến phụ thuộc. Nó biểu diễn dữ liệu bằng các cột dọc hoặc ngang, với chiều dài của mỗi cột thể hiện giá trị của một biến.

#### 5.2.1 Biểu đồ điểm trung bình theo số lượng môn thi

num\_of\_exam\_taken và average là hai list ban đầu được khởi tạo với 10 phần tử có giá trị ban đầu là 0 để đếm số học sinh theo số lượng môn học họ tham gia, average để tính điểm trung bình theo số lượng môn học đó.

Trong vòng lặp for student in students:, vòng lặp lần lượt duyệt qua danh sách học sinh (students) để tính tổng điểm của học sinh và số môn học họ tham gia. Sau đó, cập nhật num\_of\_exam\_taken và average tương ứng.

Sau khi tính toán xong, đoạn mã sử dụng vòng lặp for i in range(10) để chia điểm trung bình cho số lượng học sinh tương ứng (num\_of\_exam\_taken). Điều này giúp tính toán điểm trung bình thực sự theo số lượng môn thi, và được làm tròn đến 2 chữ số thập phân.



Hình 17. Mã nguồn phần tính điểm trung bình theo số môn

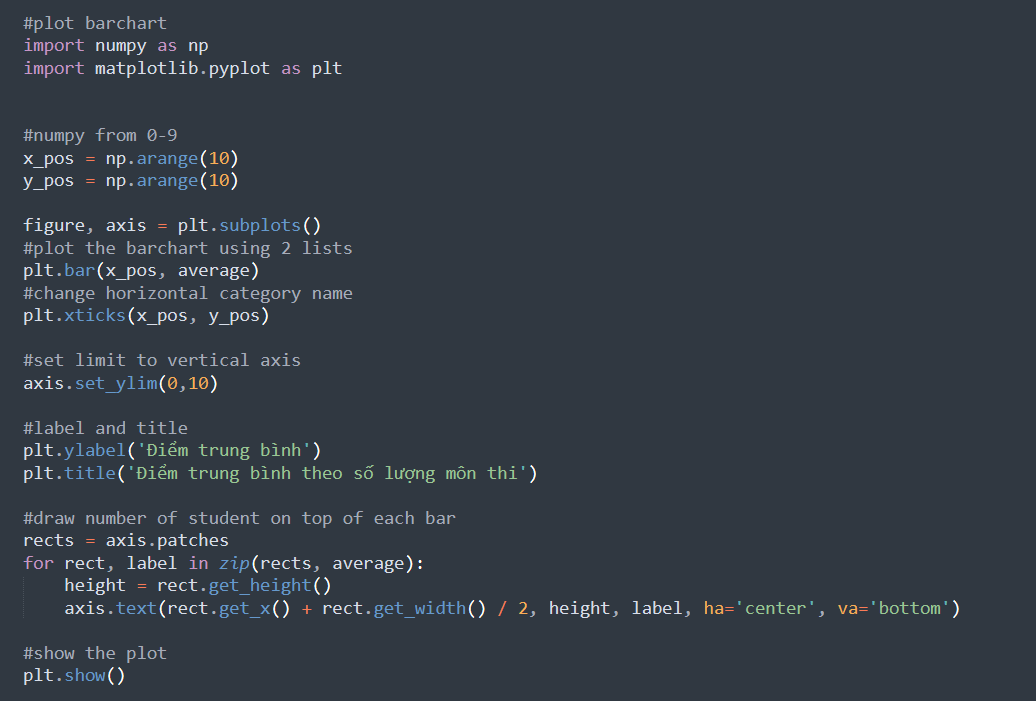
Tiếp theo, dữ liệu được sử dụng để vẽ biểu đồ cột (bar chart). Thư viện matplotlib và numpy được sử dụng để tạo biểu đồ cột.

x\_pos và y\_pos là các mảng numpy từ 0 đến 9, đại diện cho số lượng môn học tham gia.

Với plt.bar(x\_pos, average), biểu đồ cột được vẽ, trong đó x\_pos là trục x, average là trục y (điểm trung bình).

Các cài đặt còn lại như plt.xticks(), axis.set\_ylim(), plt.ylabel(), plt.title() là để cấu hình trục x, trục y, nhãn trục và tiêu đề cho biểu đồ.

Cuối cùng, axis.text() được sử dụng để hiển thị số lượng học sinh tương ứng trên đỉnh của mỗi cột trên biểu đồ.



Hình 18. Phần vẽ biểu đồ điểm trung bình theo số lượng môn thi

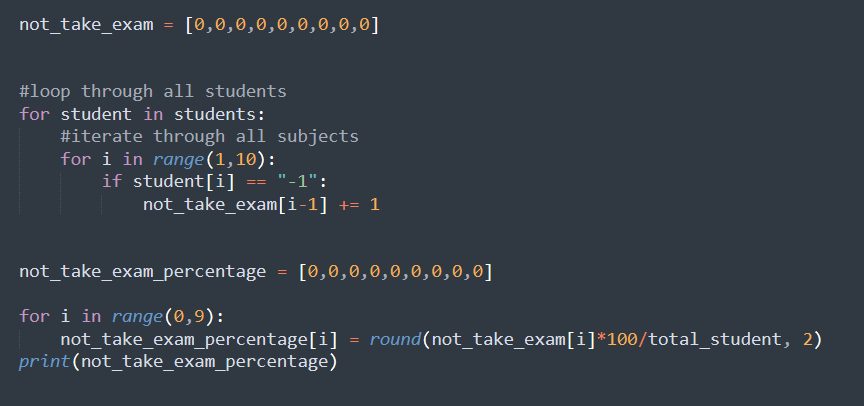
#### 5.2.2 Biểu đồ số học sinh bỏ thi hoặc không đăng ký của mỗi môn

Dùng list not\_take\_exam có độ dài 9 (tương ứng với số môn học) để lưu số lượng học sinh không tham gia mỗi môn học.

Vòng lặp duyệt qua danh sách học sinh. Trong mỗi học sinh, vòng lặp tiếp theo duyệt qua từng môn học.

Nếu điểm thi của học sinh tương ứng với môn học là "-1" (giả sử -1 là điểm của học sinh không tham gia môn học đó), thì số lượng học sinh không tham gia môn học đó sẽ được tăng lên 1.

Tạo list not\_take\_exam\_percentage để lưu trữ tỷ lệ phần trăm học sinh không tham gia từng môn học.



Hình 19. Mã nguồn phần tính toán số học sinh bỏ thi hoặc không đăng ký của mỗi môn

y\_pos = np.arange(len(subjects)): Tạo một mảng y\_pos bằng cách sử dụng np.arange() để tạo các giá trị từ 0 đến độ dài của danh sách subjects. Đây là trục y trên biểu đồ.

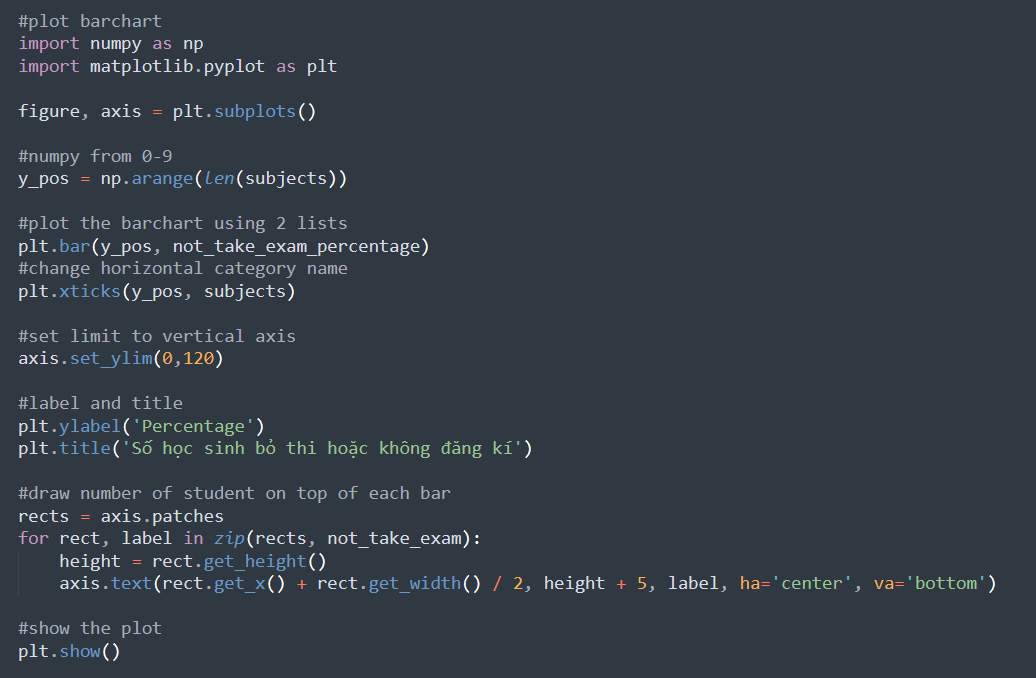
plt.bar(y\_pos, not\_take\_exam\_percentage): Vẽ biểu đồ cột, sử dụng y\_pos làm trục x và not\_take\_exam\_percentage làm trục y. Đây là biểu đồ cột thể hiện tỷ lệ phần trăm học sinh bỏ thi hoặc không đăng kí từng môn học.

plt.xticks(y\_pos, subjects): Đặt nhãn cho trục x, sử dụng subjects (danh sách các môn học) ứng với từng vị trí trên trục x.

axis.set\_ylim(0,120): Đặt giới hạn trên trục y từ 0 đến 120. Điều này có thể là để điều chỉnh biểu đồ sao cho dữ liệu vẫn có thể hiển thị một cách rõ ràng.

plt.ylabel('Percentage') và plt.title('Số học sinh bỏ thi hoặc không đăng kí'): Đặt nhãn trục y và tiêu đề biểu đồ.

for rect, label in zip(rects, not\_take\_exam):...: Với mỗi cột trong biểu đồ, hiển thị số lượng học sinh không tham gia hoặc bỏ thi tương ứng trên đỉnh của cột



Hình 20. Phần xây dựng biểu đồ số học sinh bỏ thi hoặc không đăng ký của mỗi môn

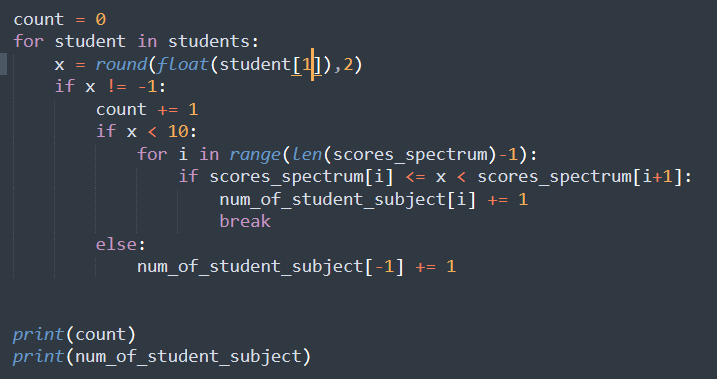
#### 5.2.3 Biểu đồ phổ điểm

num\_of\_student\_subject: Mảng lưu số lượng học sinh theo phổ điểm.

scores\_spectrum: Mảng lưu các khoảng phổ điểm từ 0 đến 10 với bước nhảy là 0.2 (từ 0.0 đến 10.0).

Vòng lặp duyệt qua danh sách học sinh. Với mỗi học sinh, xác định điểm thi môn Toán (x), sau đó tăng số lượng học sinh ở khoảng phổ điểm tương ứng trong num\_of\_student\_subject.

Tính tổng điểm trung bình (diemtb) dựa trên phổ điểm và số lượng học sinh trong từng khoảng phổ điểm.

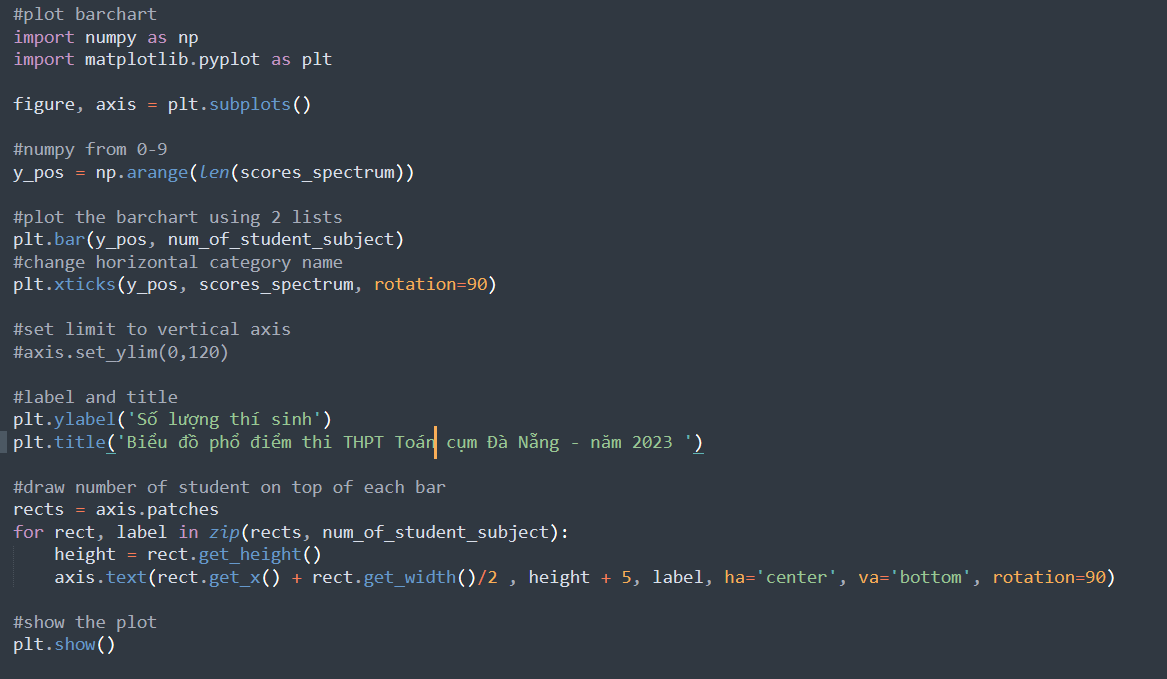


Hình 21. Mã nguồn phần tính toán phổ điểm

Tạo biểu đồ cột với trục x là scores\_spectrum và trục y là num\_of \_student\_ subject.

plt.xticks(y\_pos, scores\_spectrum, rotation=90): Đặt nhãn trên trục x là các phổ điểm, xoay chúng 90 độ để dễ đọc.

axis.text(): Hiển thị số lượng học sinh trên đỉnh mỗi cột của biểu đồ.

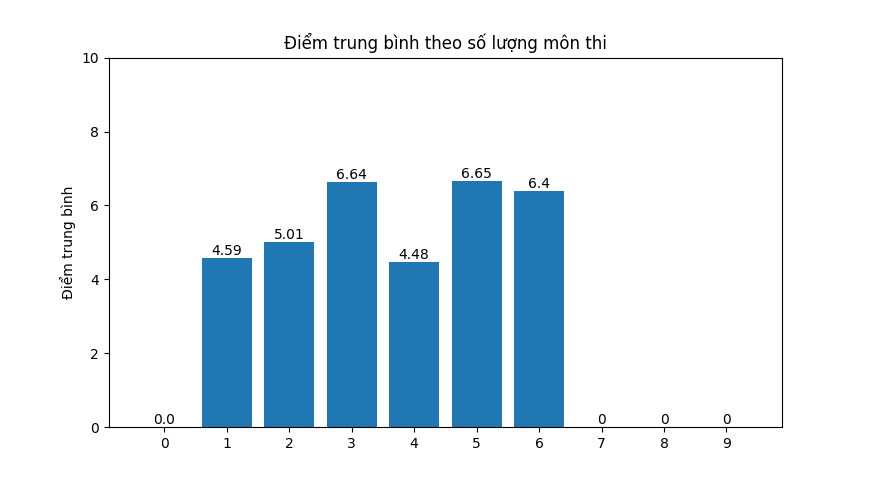


Hình 22. Phần tính xây dựng biểu đồ phổ điểm

# CHƯƠNG 3: BIỂU ĐỒ

## 1. Biểu đồ điểm trung bình theo số lượng môn thi

### 1.1 Biểu đồ



Hình 23. Biểu đồ điểm trung bình theo số lượng môn thi

### 1.2 Đánh giá

ta có thể nhận thấy rằng điểm trung bình môn thi tăng dần khi số lượng môn tăng từ 0 đến 3. Từ 3 môn trở đi, điểm trung bình duy trì ở mức cao, xấp xỉ 6.64. Một số nhóm thí sinh thi (0, 4, 7, 8, 9 môn) có điểm trung bình là 0, trong khi các nhóm thi nhiều môn (3, 5, 6 môn) có điểm trung bình cao hơn (6.64 và 6.65).

## 2. Biểu đồ số học sinh bỏ thi hoặc không đăng kí mỗi môn

### 2.1 Biểu đồ

#### 

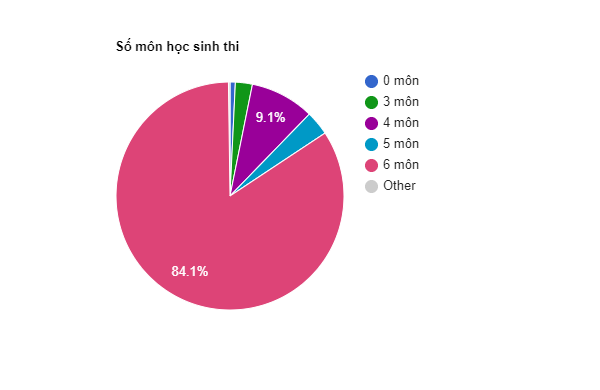
#### Hình 24. Biểu đồ số học sinh bỏ thi hoặc không đăng kí

### 2.2 Đánh giá

Tổng quan, dữ liệu cho thấy có một số lượng đáng kể học sinh không tham gia kỳ thi ở các môn học khác nhau trong cụm Đà Nẵng năm 2023. Điều này có thể do nhiều nguyên nhân như sự chuẩn bị không đầy đủ, học sinh không quan tâm đến môn học đó, hoặc các vấn đề cá nhân khác. Nhưng những môn bắt buộc thì có rất ít thí sinh không thi. Thí sinh thi các môn khối xã hội có xu hướng ít bỏ thi hơn các môn khối tự nhiên.

## 3. Biểu đồ số môn học sinh thi

### 3.1 Biểu đồ

****

Hình 25. Biểu đồ số môn học sinh thi

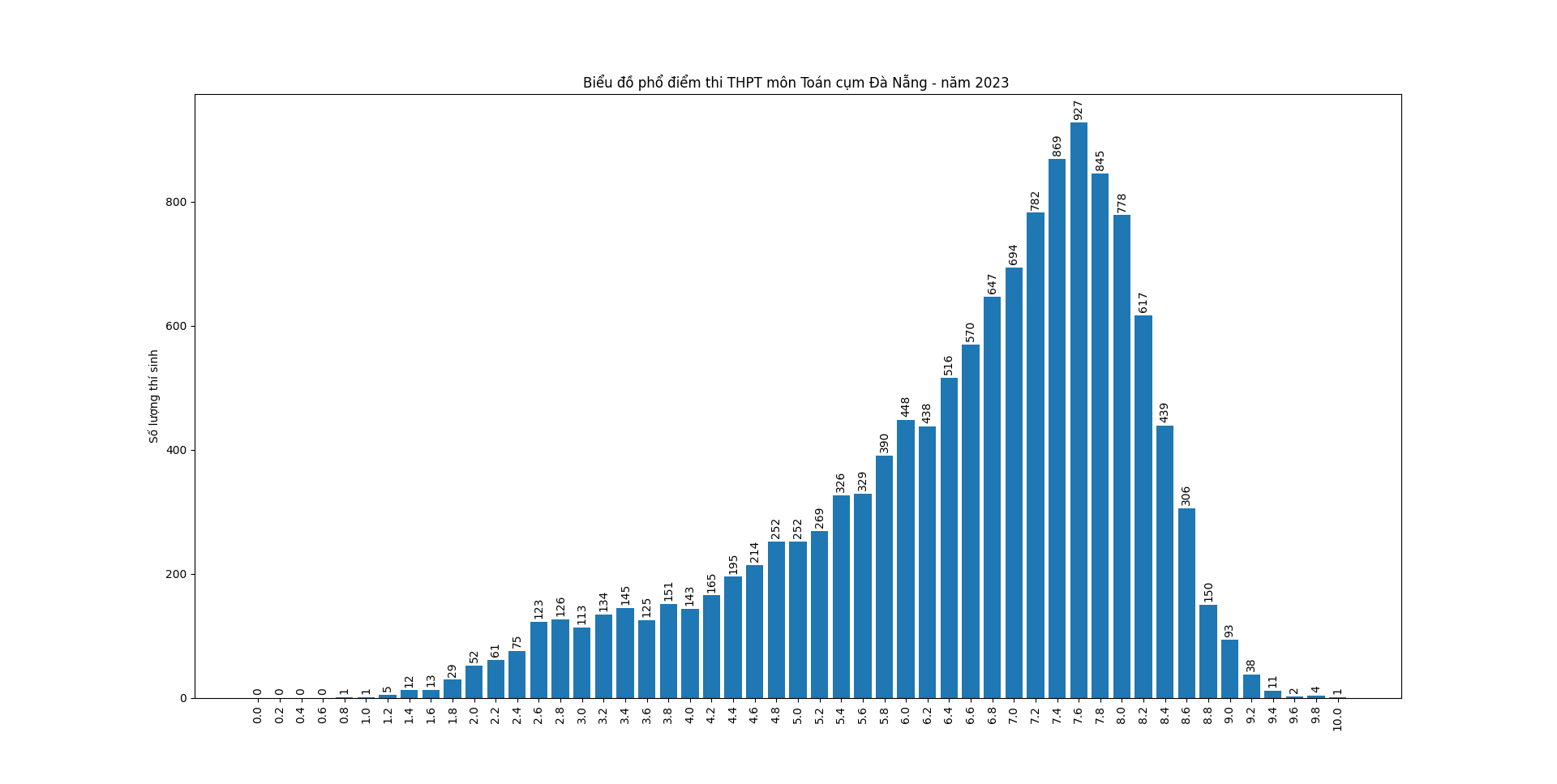
### 3.2 Đánh giá

Hầu hết học sinh thi 6 môn để đủ các môn thuộc các khối tự nhiên hoặc xã hội để lấy điểm thi đại học. Có một số ít thi dưới 6 môn có thể là chỉ thi để xét tốt nghiệp cấp ba.

## 4. Phổ điểm

### 4.1 Biểu đồ

#### 4.1.1 Phổ điểm Toán

****

Hình 26. Phổ điểm toán

#### 4.1.2 Phổ điểm Ngữ Văn

#### 

#### Hình 27. Phổ điểm Ngữ Văn

#### 4.1.3 Phổ điểm Lịch Sử

#### 

#### Hình 28. Phổ điểm Lịch Sử

#### 4.1.4 Phổ điểm Địa Lý

#### 

#### Hình 29. Phổ điểm Địa Lý

#### 4.1.5 Phổ điểm Giáo Dục Công Dân

#### 

#### Hình 30. Phổ điểm Giáo Dục Công Dân

#### 4.1.6 Phổ điểm Sinh Học

#### 

#### Hình 31. Phổ điểm Sinh Học

#### 4.1.7 Phổ điểm Vật Lý

#### 

#### Hình 32. Phổ điểm Vật Lý

#### 4.1.8 Phổ điểm Hóa Học

#### 

#### Hình 33. Phổ điểm Hóa Học

#### 4.1.9 Phổ điểm Ngoại Ngữ

#### 

#### Hình 34. Phổ điểm Ngoại Ngữ

### 4.2 Đánh giá

#### Kết quả phân tích phổ điểm thi tốt nghiệp THPT môn Toán của cụm Đà Nẵng năm 2023 cho thấy có 12876 thí sinh tham gia bài thi Toán, trong đó điểm trung bình là 6.53; điểm số có nhiều thi sinh đạt nhất là 7.6 điểm.

#### Kết quả phân tích phổ điểm thi tốt nghiệp THPT môn Ngữ Văn của cụm Đà Nẵng năm 2023 cho thấy có 12871 thí sinh tham gia bài thi Ngữ Văn, trong đó điểm trung bình là 6.02; điểm số có nhiều thi sinh đạt nhất là 6.0 điểm.

#### Kết quả phân tích phổ điểm thi tốt nghiệp THPT môn Vật Lý của cụm Đà Nẵng năm 2023 cho thấy có 5391 thí sinh tham gia bài thi Vật Lý, trong đó điểm trung bình là 6.51; điểm số có nhiều thi sinh đạt nhất là 7.5 điểm.

#### Kết quả phân tích phổ điểm thi tốt nghiệp THPT môn Giáo Dục Công Dân của cụm Đà Nẵng năm 2023 cho thấy có 6173 thí sinh tham gia bài thi Giáo Dục Công Dân, trong đó điểm trung bình là 8.07; điểm số có nhiều thi sinh đạt nhất là 8.25 điểm.

#### Kết quả phân tích phổ điểm thi tốt nghiệp THPT môn Lịch Sử của cụm Đà Nẵng năm 2023 cho thấy có 7509 thí sinh tham gia bài thi Lịch Sử, trong đó điểm trung bình là 5.56; điểm số có nhiều thi sinh đạt nhất là 5.0 và 5.25 điểm.

#### Kết quả phân tích phổ điểm thi tốt nghiệp THPT môn Địa Lí của cụm Đà Nẵng năm 2023 cho thấy có 7480 thí sinh tham gia bài thi Địa Lí, trong đó điểm trung bình là 5.77; điểm số có nhiều thi sinh đạt nhất là 5.5 điểm.

#### Kết quả phân tích phổ điểm thi tốt nghiệp THPT môn Hóa Học của cụm Đà Nẵng năm 2023 cho thấy có 5408 thí sinh tham gia bài thi Hóa Học, trong đó điểm trung bình là 6.47; điểm số có nhiều thi sinh đạt nhất là 7.5 điểm.

#### Kết quả phân tích phổ điểm thi tốt nghiệp THPT môn Ngoại Ngữ của cụm Đà Nẵng năm 2023 cho thấy có 11221 thí sinh tham gia bài thi Ngoại Ngữ, trong đó điểm trung bình là 6.11; điểm số có nhiều thi sinh đạt nhất là 7.8 điểm.

# KẾT LUẬN

Dựa trên kết quả phân tích phổ điểm thi tốt nghiệp THPT của cụm Đà Nẵng năm 2023, chúng ta có thể rút ra kết luận rằng hiệu suất thi và phân bố điểm thi trong các môn học khác nhau có sự đa dạng.

Trước hết, chúng ta có thể thấy rằng điểm trung bình của các môn thi dao động từ 5.02 (môn Lịch Sử) đến 8.07 (môn Giáo Dục Công Dân). Các môn thi bắt buộc như Toán, Ngữ Văn và Ngoại Ngữ có sự tham gia bài thi cao và điểm trung bình khá ổn định, trong khi các môn không bắt buộc như Địa Lý và Lịch Sử có điểm trung bình thấp hơn. Điều này có thể cho thấy sự quan tâm và sự chuẩn bị khác nhau của học sinh đối với các môn học khác nhau.

Cũng đáng chú ý là điểm số phổ biến nhất trong mỗi môn thi cũng có sự biến đổi. Điểm số phổ biến nhất trong các môn thi từ 5.0 (môn Lịch Sử) đến 8.25 (môn Giáo Dục Công Dân), và điểm số này có thể thay đổi theo từng môn. Điều này cho thấy rằng có sự đa dạng trong kết quả thi của các thí sinh và không có một điểm số nhất định mà đạt được bởi tất cả thí sinh.

Bên cạnh đó, có một sự tăng dần của điểm trung bình khi số lượng môn tăng từ 0 đến 3, và điểm trung bình duy trì ở mức cao, xấp xỉ 6.64 từ 3 môn trở đi. Điều này có thể cho thấy rằng việc tham gia và học tập đồng thời các môn học có thể góp phần nâng cao hiệu suất thi của học sinh.

Từ các kết quả trên, chúng ta có thể nhận thấy rằng việc đánh giá và phân tích phổ điểm thi tốt nghiệp THPT có thể cung cấp thông tin hữu ích về hiệu suất học tập và phân bố điểm thi của học sinh trong các môn học khác nhau. Thông qua việc hiểu rõ điểm mạnh và điểm yếu của học sinh trong từng môn thi, nhà trường và giáo viên có thể tùy chỉnh quá trình giảng dạy và học tập để đạt được kết quả tốt hơn trong tương lai. Đồng thời, học sinh cũng có thể sử dụng thông tin này để định hướng học tập và nâng cao kỹ năng trong các môn học mà họ có điểm thấp để đạt được thành tích tốt hơn.

# TÀI LIỆU THAM KHẢO

1.https://careerfoundry.com/en/blog/data-analytics/the-data-analysis-process-step-by-step/

2. Wes McKinney, Python for Data Analysis 2nd edition.