

ĐỒ ÁN MÔN HỌC

THU THẬP VÀ PHÂN TÍCH DỮ LIỆU ĐIỂM THI THPT QUỐC GIA 2025 BẰNG CÔNG CỤ SELENIUM

Môn học: Mã nguồn mở trong khoa học dữ liệu

GVHD: ThS Lê Nhật Tùng

Học kỳ: 1B

Năm học: 2025 - 2026



THÀNH VIÊN NHÓM



Họ và tên	MSSV
Nguyễn Đức Vinh (Leader)	2386400052
Nguyễn Đăng Khoa	2386400026
Phan Xuân Dương	2386400966

Giới thiệu đề tài



Lý do chọn đề tài

Kỳ thi THPT Quốc Gia có vai trò quan trọng trong việc xét tốt nghiệp và tuyển sinh đại học. Việc thu thập và phân tích dữ liệu điểm thi giúp đánh giá khách quan về hiệu quả dạy học, cũng như hỗ trợ học sinh, giáo viên và nhà quản lý giáo dục trong việc đưa ra các quyết định phù hợp.



Mục tiêu nghiên cứu

Đề tài nhằm thu được bức tranh tổng quan về kết quả thi, nhận diện xu hướng và mức độ phân hóa giữa các môn học, từ đó rút ra nhận xét và đề xuất góp phần nâng cao chất lượng dạy học trong những năm tiếp theo.



Công cụ sử dụng



Ngôn ngữ lập trình bậc cao với cú pháp đơn giản, dễ học. Có hệ sinh thái thư viện phong phú hỗ trợ thu thập và xử lý dữ liệu



Thư viện Python dùng để vẽ biểu đồ cơ bản, hỗ trợ nhiều loại biểu đồ khác nhau giúp trực quan hóa dữ liệu



Công cụ tự động hóa trình duyệt web cho phép mô phỏng thao tác người dùng.



seaborn

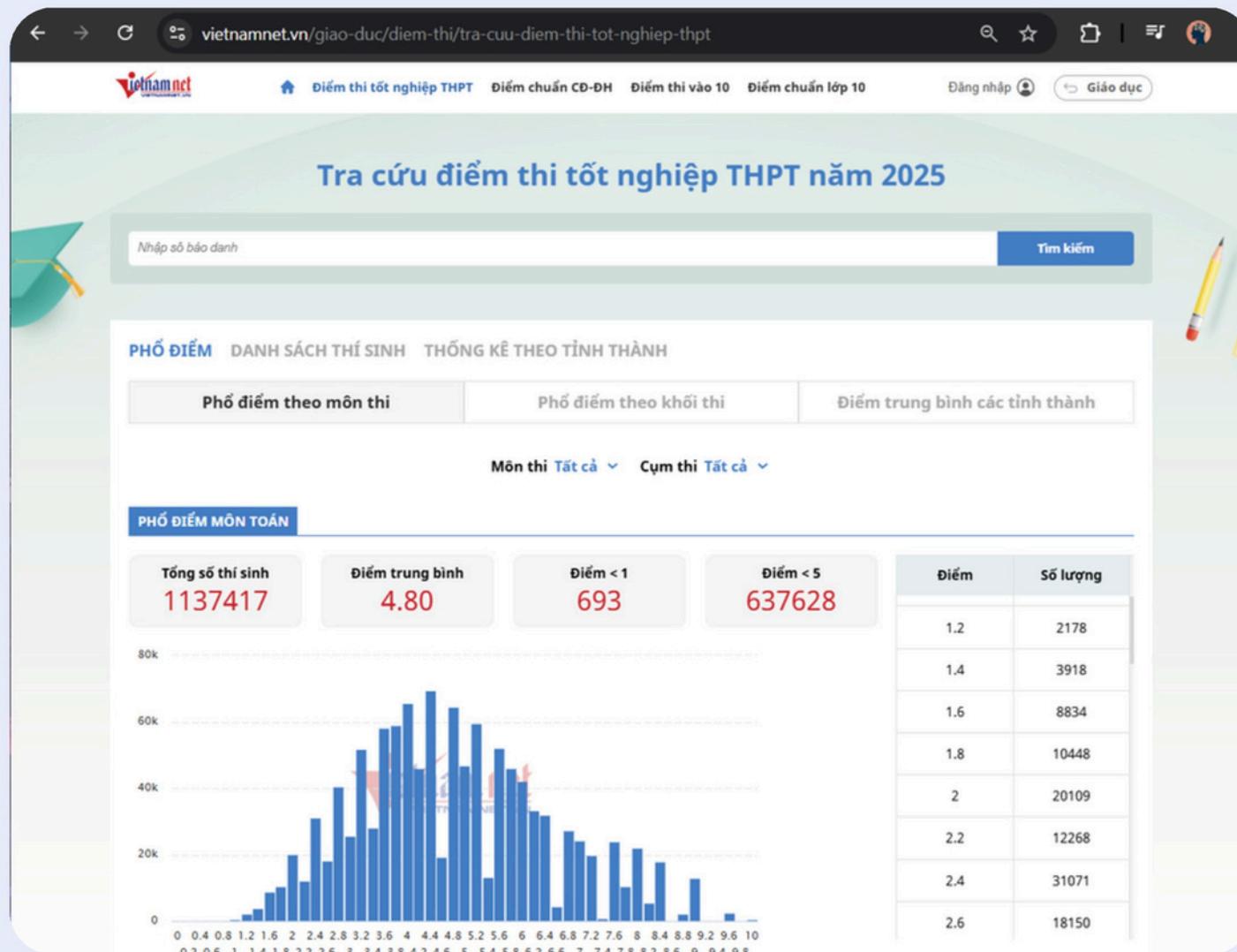
Seaborn mở rộng thư viện Matplotlib bằng việc cung cấp nhiều biểu đồ đẹp, hiện đại hơn giúp trực quan hóa nâng cao



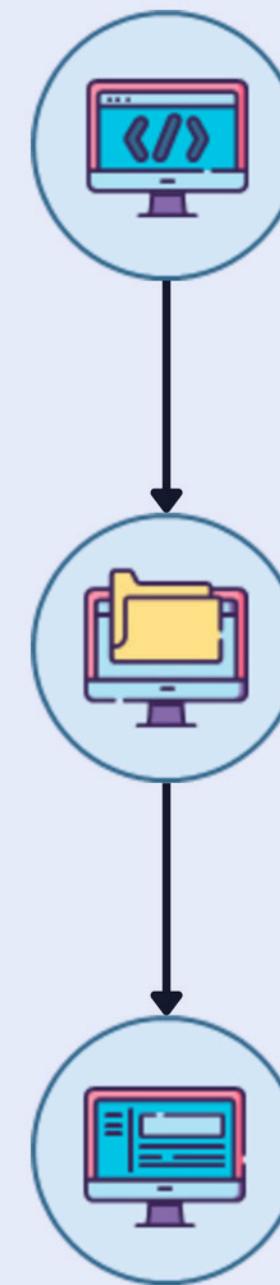
Hệ quản trị cơ sở dữ liệu nhẹ, giúp lưu trữ dữ liệu điểm thi để phục vụ phân tích và thống kê



Nguồn và quy trình hệ thống



Nguồn: <https://vietnamnet.vn/giao-duc/diem-thi/tra-cuu-diem-thi-tot-nghiep-thpt>



Thu thập & làm sạch dữ liệu

Thu thập dữ liệu từ trang đích và làm sạch dữ liệu vừa thu thập được

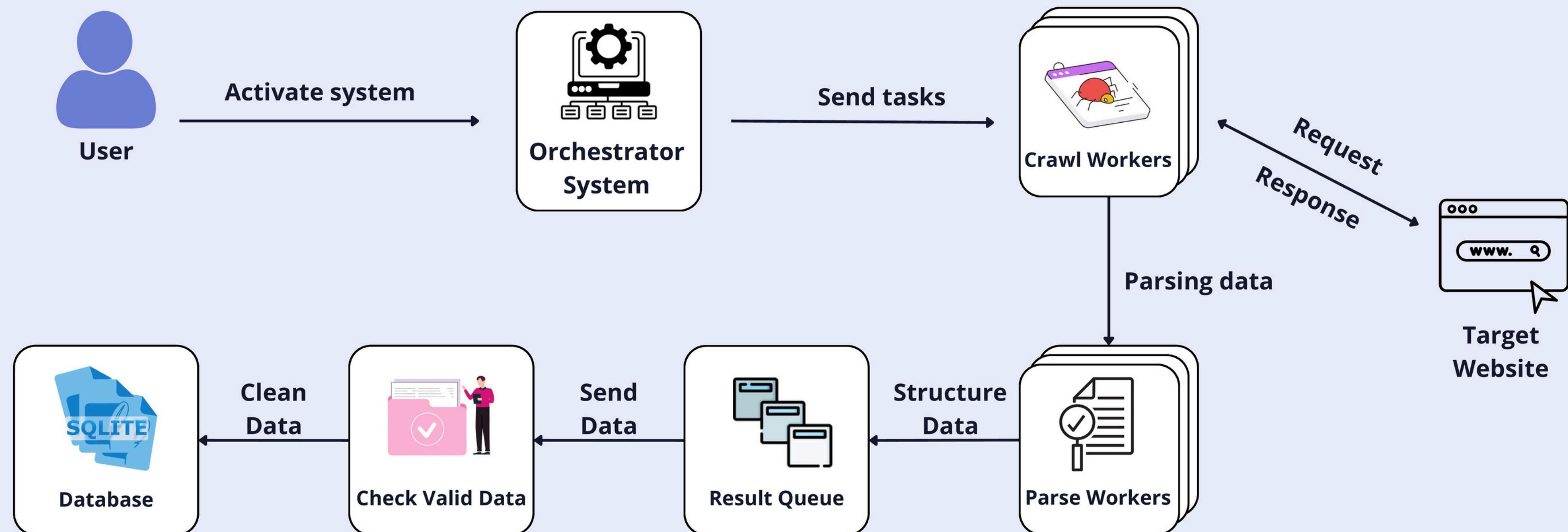
Lưu trữ dữ liệu

Lưu trữ dữ liệu có cấu trúc sau khi thu thập được vào CSDL SQLite

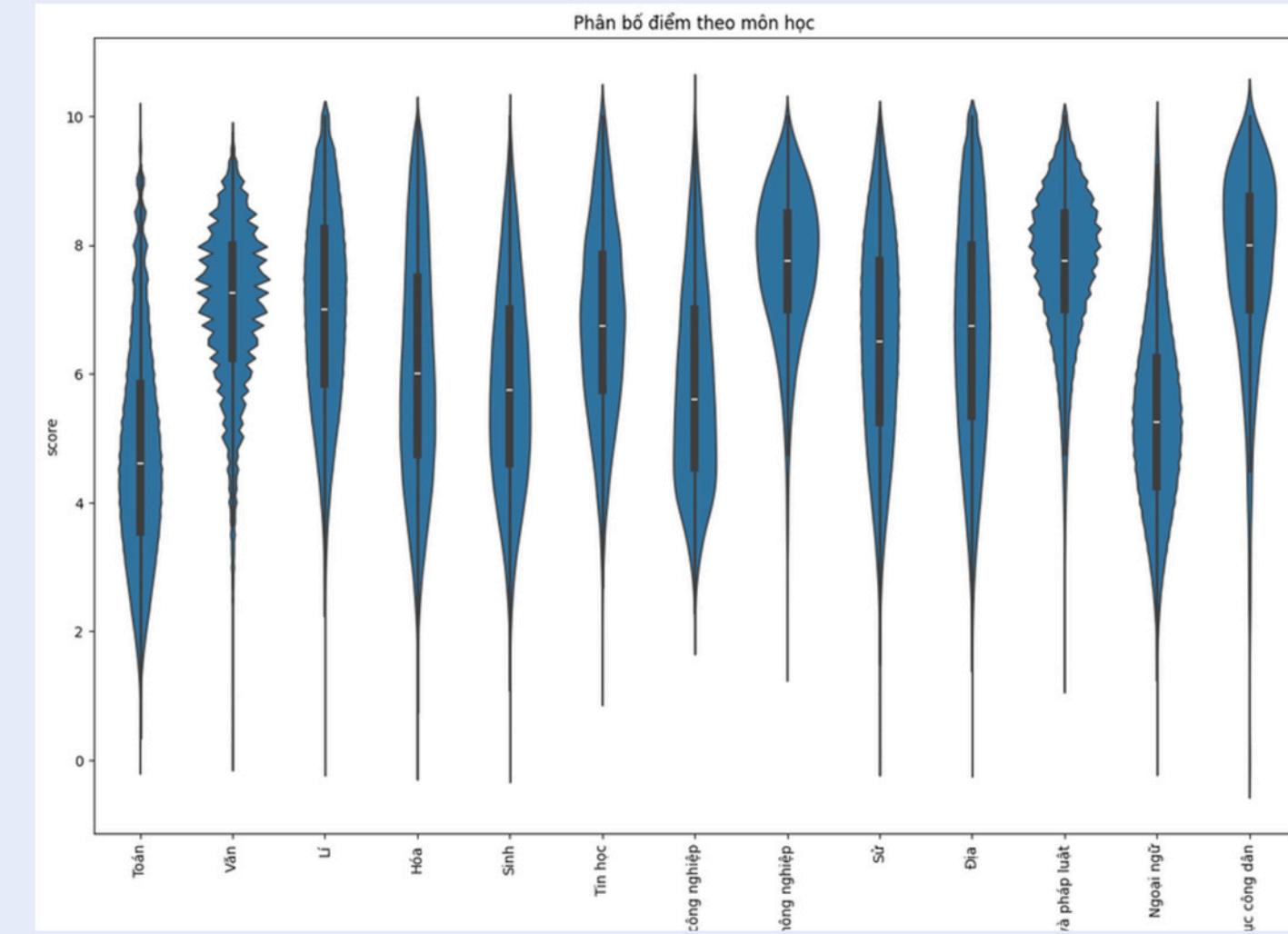
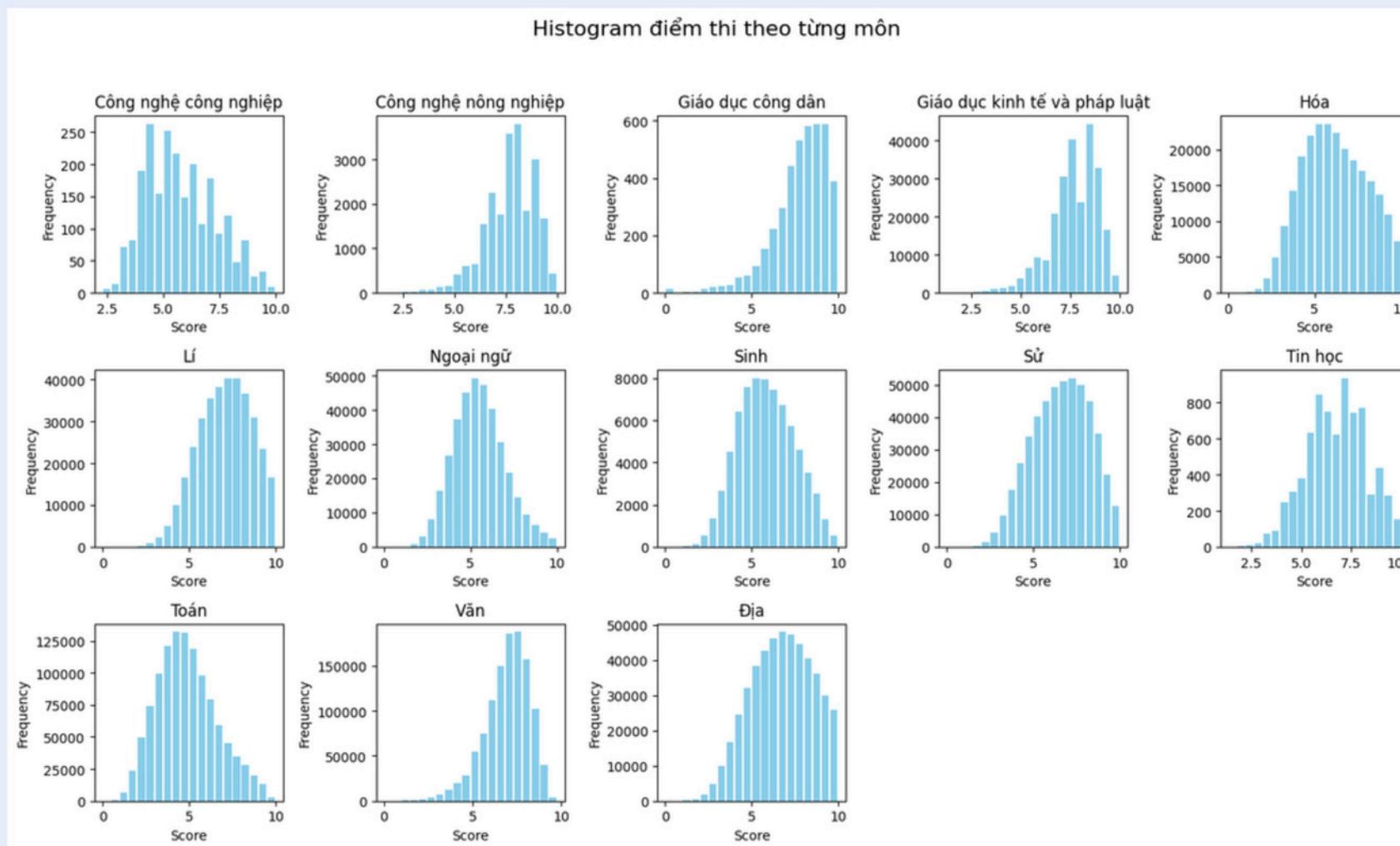
Phân tích dữ liệu

Tiến hành truy vấn và phân tích bộ dữ liệu thô nhằm tìm ra các thông tin hữu ích

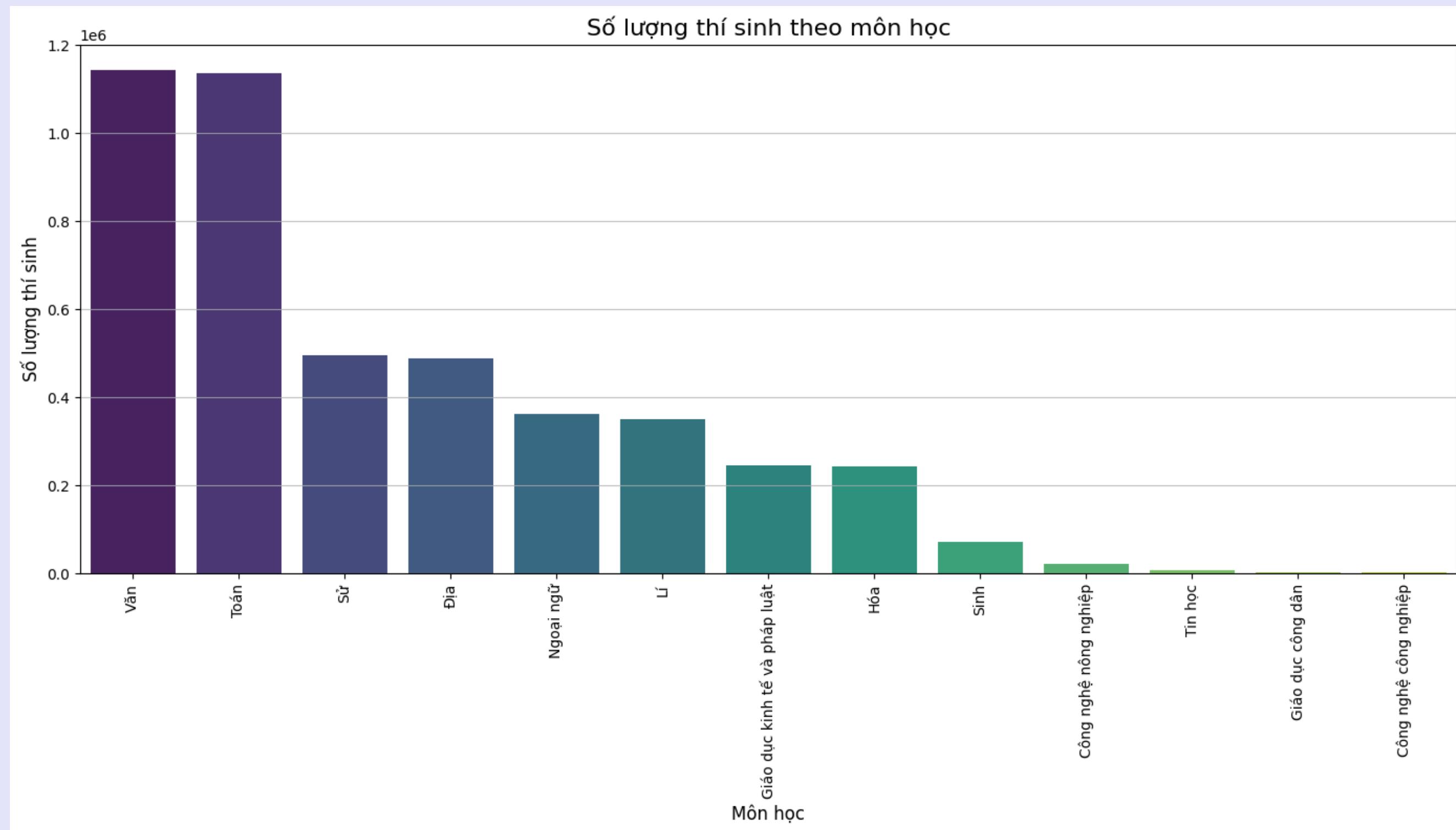
Xây dựng hệ thống thu thập dữ liệu



Phân tích và trực quan hóa dữ liệu

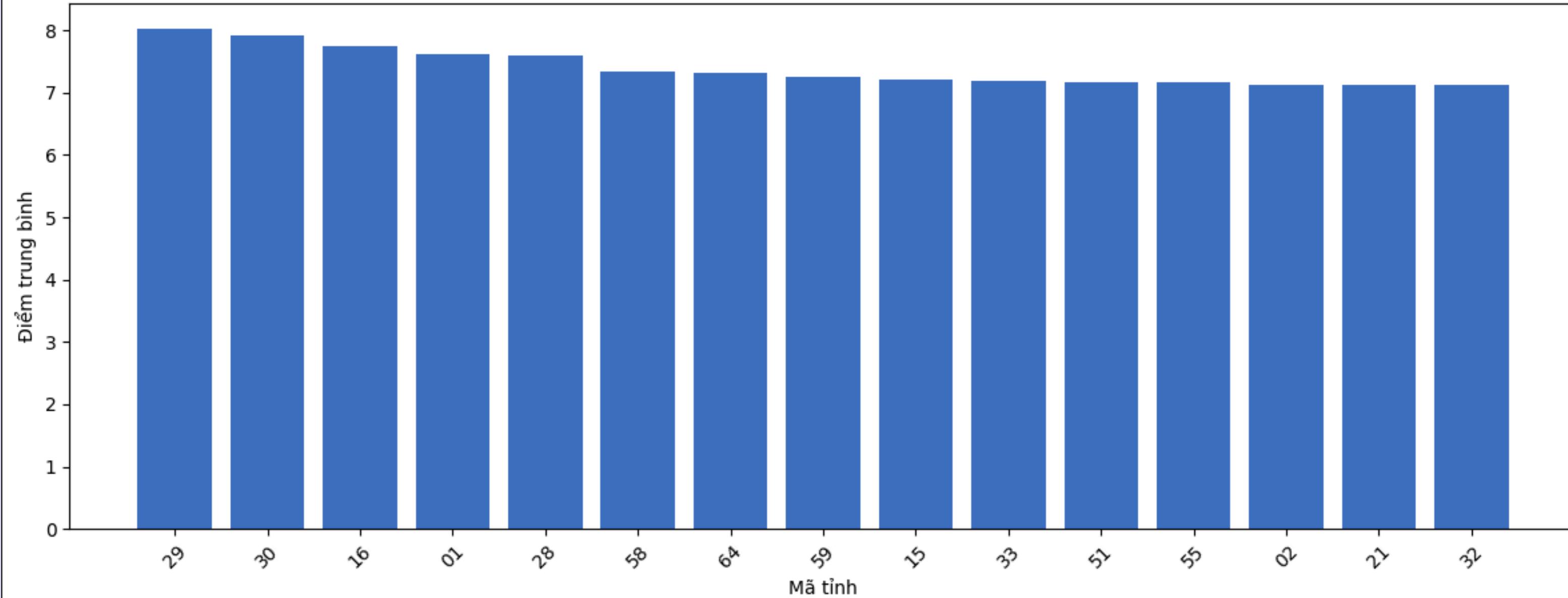


Nhận xét: Nhìn chung, bức tranh toàn cảnh về phổ điểm cho thấy một cấu trúc phân hóa năng lực thí sinh khá ổn định và phân tách rõ ràng theo hai mục tiêu của kỳ thi (xét tốt nghiệp và xét tuyển đại học). Đại đa số các môn học thuật cốt lõi đều tuân theo phân phối chuẩn (hình chuông) với đỉnh tập trung ở trung bình (5-6 điểm), cho thấy đề thi có độ tin cậy cao để sàng lọc thí sinh.



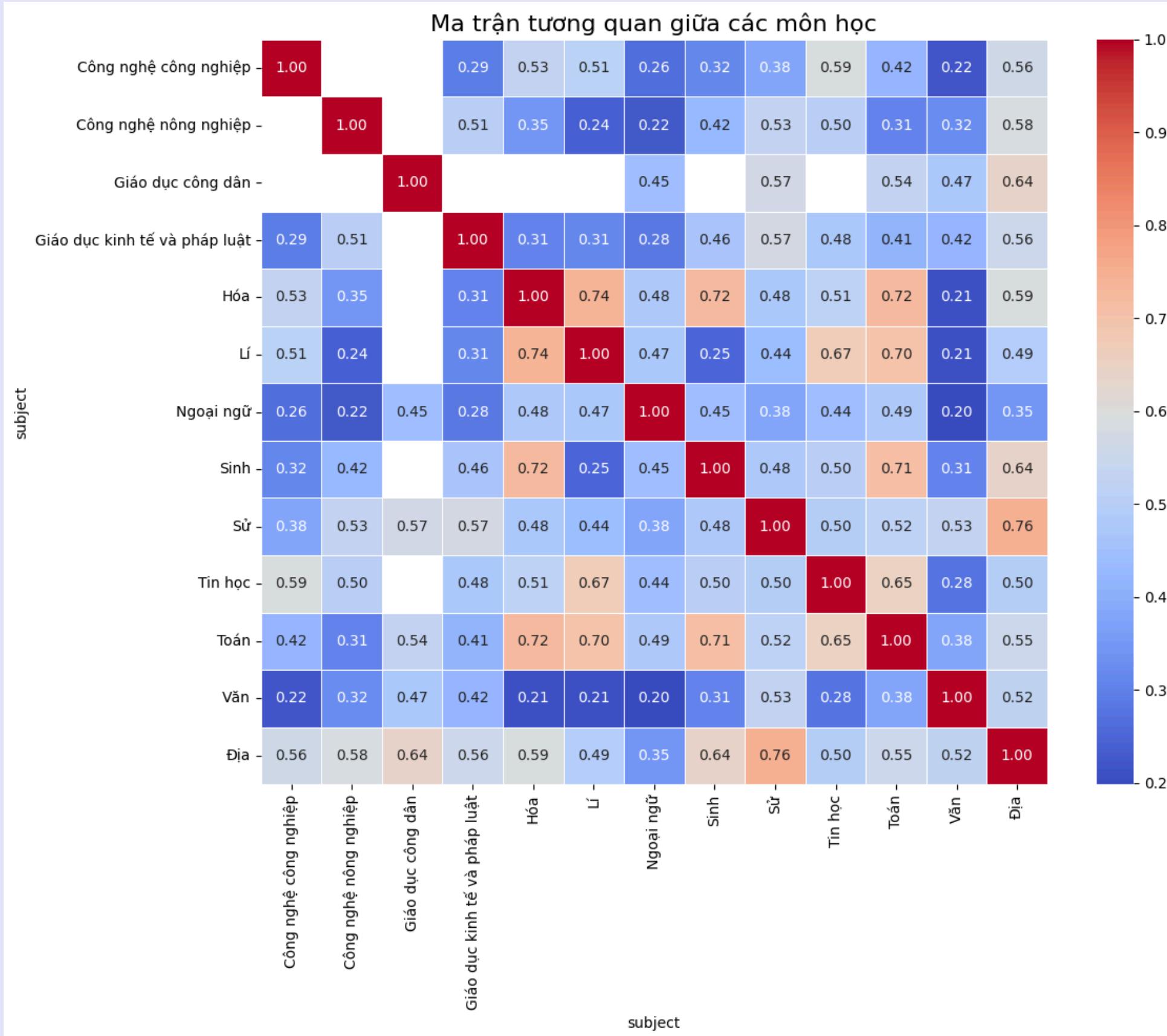
Biểu đồ số lượng thí sinh đạt điểm tối đa theo từng môn .đa số thí sinh đạt điểm tối đa môn văn và môn toán một số môn còn tương đối ít đối được chọn nên điểm còn ít .

Top 15 tỉnh có điểm trung bình Môn TOÁN cao nhất



Biểu đồ top tỉnh có điểm trung bình môn toán cao nhất

Qua biểu phân cụm thành các tỉnh dựa trên số báo danh (2 số đầu) ta thấy 3 tỉnh có điểm trung bình toán lớn nhất là 29 Hải Phòng, 30 Hà Nội và 16 Phú thọ.



Biểu đồ tương quan giữa điểm số của các môn học

- Các môn Toán, Lý, Hóa thể hiện sự tương quan rất cao (hệ số từ 0.70 đến 0.74)
- Toán và Tin học: Môn Toán có sự tương quan chặt chẽ với Tin học (0.65) và Sinh học (0.71)
- Môn Văn và Ngoại ngữ có hệ số tương quan thấp nhất với các môn còn lại (phần lớn dưới 0.40)

Hướng phát triển

1. Mở rộng phân tích dữ liệu

- So sánh phổ điểm các năm trước (ví dụ 2023-2025) để đánh giá xu hướng thay đổi mức độ khó của đề thi.
- Phân tích chi tiết theo khu vực, trường học, khối thi để phát hiện sự chênh lệch hoặc ưu thế vùng miền.

2. Ứng dụng công nghệ nâng cao

- Áp dụng các kỹ thuật học máy (Machine Learning) để dự đoán điểm hoặc nhóm thí sinh theo năng lực.
- Tích hợp dashboard trực quan (ví dụ bằng Plotly hoặc Power BI) để báo cáo kết quả phân tích theo thời gian thực.

3. Cải thiện công cụ thu thập và xử lý dữ liệu

- Tự động cập nhật dữ liệu từ các website công bố điểm mới mỗi năm.
- Kết hợp với cơ sở dữ liệu lớn hơn (ví dụ MySQL, PostgreSQL) để quản lý dữ liệu toàn quốc, phục vụ nghiên cứu sâu hơn.



CẢM ƠN THẦY VÀ CÁC BẠN ĐÃ LẮNG NGHE



QUESTION & ANSWER

