# ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
# TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
# KHOA CÔNG NGHỆ THÔNG TIN


# Đồ án

# Trực quan hóa dữ liệu

## Lab 02: Làm việc với dữ liệu dạng Time-Series

**LỚP: CQ2022/21**

Giảng viên:

Bùi Tiến Lên

Lê Nhựt Nam

Lê Ngọc Thành

**Sinh viên thực hiện:**

| MSSV | Tên thành viên |
|---|---|
| 21120119 | Hứa Minh Quân |
| 21120406 | Lê Viết Đạt Trọng |
| 21120441 | Dương Huỳnh Anh Duy |
| 22120380 | Hồ Nhất Trí |

Học kỳ I – Năm học 2024-2025

# Mục lục

# I. Đóng góp của các thành viên

| Thành viên | Công việc | Mức độ hoàn thành | Phần trăm đóng góp |
|---|---|---|---|
| Hứa Minh Quân | | 100% | 25% |
| Lê Viết Đạt Trọng | Ăn ngủ cùng AI | 100% | 25% |
| Dương Huỳnh Anh Duy (Nhóm trưởng) | | 100% | 25% |
| Hồ Nhất Trí | | 100% | 25% |

# II. Thu thập dữ liệu

- Đề tài được chọn: Điểm thi trung học phổ thông quốc gia của Việt Nam năm 2023
- Bộ dữ liệu: Vietnamese National HS Graduation Exam 2023

## 1. Nội dung bộ dữ liệu

Tập dữ liệu cung cấp thông tin về điểm thi bao gồm các môn bắt buộc và các môn thuộc hai tổ hợp tự nhiên và xã hội được quy định trong kỳ thi Trung học phổ thông quốc gia năm 2023

## 2. Cách bộ dữ liệu được xây dựng

Crawl dữ liệu từ điểm thi 2023 công bố bởi bộ giáo dục

## 3. Cách sử dụng và tính pháp lý

- Dataset tải về trên Kaggle dưới dạng file csv.
- Dataset được đăng tải với giấy phép CC0: Public domain, tức là người đăng không yêu cầu về bản quyền và các quyền liên quan, người sử dụng hoàn toàn có thể điều chỉnh, nghiên cứu, phân phối, thậm chí cho mục đích thương mại.

# III. Khám phá dữ liệu

## 1. Các thư viện đã sử dụng

- **NumPy**: là thư viện cốt lõi cho tính toán khoa học trong Python. Nó cung cấp các cấu trúc dữ liệu hiệu quả cao cho các mảng lớn và đa chiều (ndarray), cùng với một bộ sưu tập lớn các hàm toán học để thực hiện các phép toán số trên các mảng này.
- **Pandas**: được xây dựng trên NumPy để cung cấp các cấu trúc dữ liệu linh hoạt và dễ sử dụng hơn, đặc biệt là cho việc làm việc với dữ liệu có cấu trúc như bảng dữ liệu (DataFrame) và các chuỗi thời gian (Series).
- **Matplotlib**: Matplotlib là một thư viện đồ họa 2D phổ biến trong Python, cung cấp các công cụ để tạo ra các loại đồ thị khác nhau như đường, cột, hình tròn, ...

- **Seaborn**: được xây dựng trên Matplotlib để tạo ra các đồ thị thống kê hấp dẫn và có tính thẩm mỹ cao. Nó cung cấp một giao diện đơn giản hơn và các hàm cấp cao để trực quan hóa các mối quan hệ giữa các biến trong dữ liệu.

## 2. Đọc dữ liệu từ file

Sử dụng *pd.read_csv*() để dọc dữ liệu từ file '*score.csv*', được chuyển thành dạng DataFrame, do dữ liệu thí sinh năm 2023 rất lớn lên đến 1,5 triệu dòng nên nhóm lấy xử lý dữ liệu tượng trưng là 7000 dòng

## 3. Ý nghĩa mỗi cột

Các cột:

- StudentID: Mã định danh duy nhất cho mỗi học sinh.
- Mathematics: Điểm số của học sinh trong môn Toán.
- Literature: Điểm số của học sinh trong môn Ngữ văn.
- Foreign Language: Điểm số của học sinh trong môn Ngoại ngữ.
- Physics: Điểm số của học sinh trong môn Vật lý.
- Chemistry: Điểm số của học sinh trong môn Hóa học.
- Biology: Điểm số của học sinh trong môn Sinh học.
- History: Điểm số của học sinh trong môn Lịch sử.
- Geography: Điểm số của học sinh trong môn Địa lý.
- Civic Education: Điểm số của học sinh trong môn Giáo dục công dân.
- Foreign Language Code: Mã đại diện cho ngoại ngữ mà học sinh đã chọn học..

## 4. Hợp nhất các cột điểm của tổ hợp

Khi phân tích tập dữ liệu, các cột điểm tổ hợp tự nhiên sẽ bị trống do thí sinh chọn tổ hợp xã hội và ngược lại nên nhóm thực hiện xử lý:

- Tạo một cột mới tên là Subject Combination gồm 2 giá trị là Society và Nature để phân loại học sinh theo tổ hợp
- Hợp nhất điểm các môn, các môn có mối quan hệ tương đương được hợp nhất thành cột mới để tránh mất dữ liệu
  - Physics/History: Gộp điểm của môn Vật lý (hoặc Lịch sử nếu thiếu điểm Vật lý).
  - Chemistry/Geography: Gộp điểm của môn Hóa học (hoặc Địa lý nếu thiếu điểm Hóa học).
  - Biology/Civic education: Gộp điểm của môn Sinh học (hoặc Giáo dục công dân nếu thiếu điểm Sinh học).

## 5. Kiểm tra và xóa các dòng trùng lặp

Trong quá trình nhập dữ liệu có thể xảy ra sai sót, sử dụng *duplicated*().*sum*() để xác định cũng như đưa ra thông báo về các dòng đó. Kết hợp với *drop_duplicates*() để xóa các dòng đó.

## 6. Kiểm tra xử lý dữ liệu missing

Sau khi hợp nhất các môn nếu vẫn còn dữ liệu missing thì đến từ các thí sinh bỏ môn thi bắt buộc, nhóm sẽ xử lí là xóa các dòng missing chỉ tập trung những thí sinh thực hiện thi đầy đủ.

## 7. Thêm cột tổng điểm

Để trực quan hơn dữ liệu, nhóm thực hiện thêm một cột tính tổng điểm tốt nghiệp tên là Total

## 8. Ý nghĩa các cột sau khi xử lý

- Student ID: Mã định danh duy nhất cho mỗi học sinh.
- Mathematics: Điểm thi của học sinh trong môn Toán.
- Literature: Điểm thi của học sinh trong môn Ngữ văn.
- Foreign language: Điểm thi của học sinh trong môn Ngoại ngữ.
- Foreign language code: Mã đại diện cho ngoại ngữ mà học sinh đã chọn học.
- Subject combination: Tổ hợp môn của học sinh:
  - Nature: Học sinh thuộc tổ hợp tự nhiên (Vật lý, Hóa học, Sinh học).
  - Society: Học sinh thuộc tổ hợp xã hội (Lịch sử, Địa lý, Giáo dục công dân).
- Physics/History: Điểm thi học sinh  môn Vật lý hoặc Lịch sử
- Chemistry/Geography: Điểm thi học sinh  môn Hóa học hoặc Địa lý
- Biology/Civic education: Điểm thi học sinh môn Sinh học hoặc Giáo dục công dân
- Total: Tổng điểm tốt nghiệp học sinh từ tất cả các môn được tính.

## 9. Các thuộc tính có kiểu số

Xác định các giá trị đặc biệt của các thuộc tính có kiểu dữ liệu ‘int64’:

- Giá trị nhỏ nhất.
- Giá trị lớn nhất.
- Tỉ lệ giá trị bị thiếu.
- Các khoảng tứ phân vị.

## 10. Xóa các giá trị gây nhiễu

- Sử dụng phương pháp Z-score để lọc giá trị nhiễu theo các bước:
  - Tính giá trị trung bình và độ lệch chuẩn.
  - Tính Z-score cho từng điểm dữ liệu:

$$Z = \frac{x - \mu}{\sigma}$$

  - Thiết lập ngưỡng Z-score để xác định outliers (ở đây là 3).
  - Loại bỏ các điểm dữ liệu có giá trị Z-score vượt quá ngưỡng đã thiết lập.
- Lý do sử Z-score thay vì dùng IQR (khoảng tứ phân vị):
  - Z-score chuyển đổi các giá trị về cùng một thang đo, cho phép so sánh trực tiếp giữa các biến khác nhau, ngay cả khi chúng có đơn vị đo khác nhau.
  - Z-score không phụ thuộc vào hình dạng của phân phối dữ liệu.

- Dễ dàng điều chỉnh ngưỡng Z-score để phù hợp với mức độ nghiêm ngặt trong việc xác định ngoại lệ.

- Sau khi xem xét thì nhóm quyết định chỉ phân tích với trường hợp xe đã qua sử dụng (**Status** = Used), vì xe đã qua sử dụng chiếm 95.77% số xe và sự cách biệt về giá, số km đã đi giữa xe cũ và các xe khác là rất lớn có thể gây nhiễu dữ liệu và khó phân tích cũng như xóa đi cột thuộc tính này.

### 11. Lưu bộ dữ liệu sau khi xử lý

Dùng *to_excel*() để lưu lại bộ dữ liệu đã xử lý thành 1 file xlsx mới vào thư mục Datasets 'preprocessed_score.xlsx'.

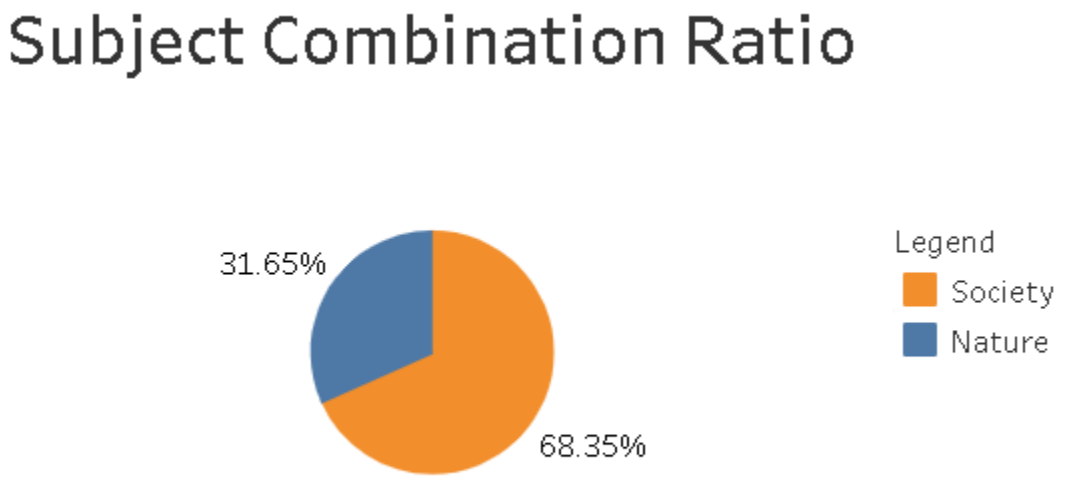
# IV. Trực quan hóa dữ liệu

## 1. Phân tích các thuộc tính

### *a) Sự phân bố theo phần trăm các bài thi theo tổ hợp*

Biểu đồ tròn:

- Dễ dàng so sánh các phần cấu thành nên một tổng thể. Mỗi phần của hình tròn đại diện cho một tỷ lệ nhất định
- Truyền đạt thông tin một cách nhanh chóng và rõ ràng, đặc biệt khi muốn so sánh các phần của một tổng thể.



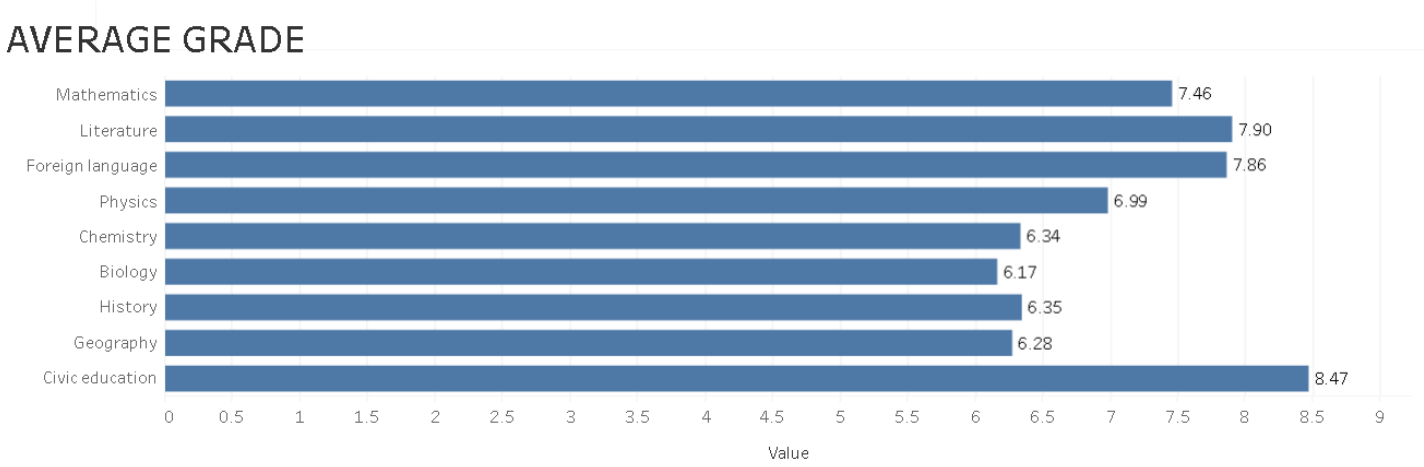

Nhận xét: Biểu đồ pie chart trên thể hiện tỷ lệ phần trăm phân bố số bài thi theo tổ hợp môn tự nhiên hoặc xã hội

- **Society**: số bài thi tổ hợp xã hội chiếm 68.35% số bài thi trong tập dữ liệu
- **Nature**: số bài thi tổ hợp tự nhiên chiếm 31.65% số bài thi trong tập dữ liệu
- Tỷ lệ số học sinh chọn tổ hợp xã hội cao hơn gấp đôi so với tổ hợp tự nhiên, phản ánh học sinh có xu hướng quan tâm hơn đến các môn học xã hội và có sự chênh lệch trong sự phân bổ học sinh vào các tổ hợp.

### *b) Tổng quan điểm trung bình theo từng môn học*

Biểu đồ thanh ngang:

- Dễ dàng so sánh các nhóm: So sánh nhóm với các thành phần có số lượng khác nhau.
- Phân loại rõ ràng: Giúp phân loại và trực quan hóa các thành phần trong nhóm một cách rõ ràng.



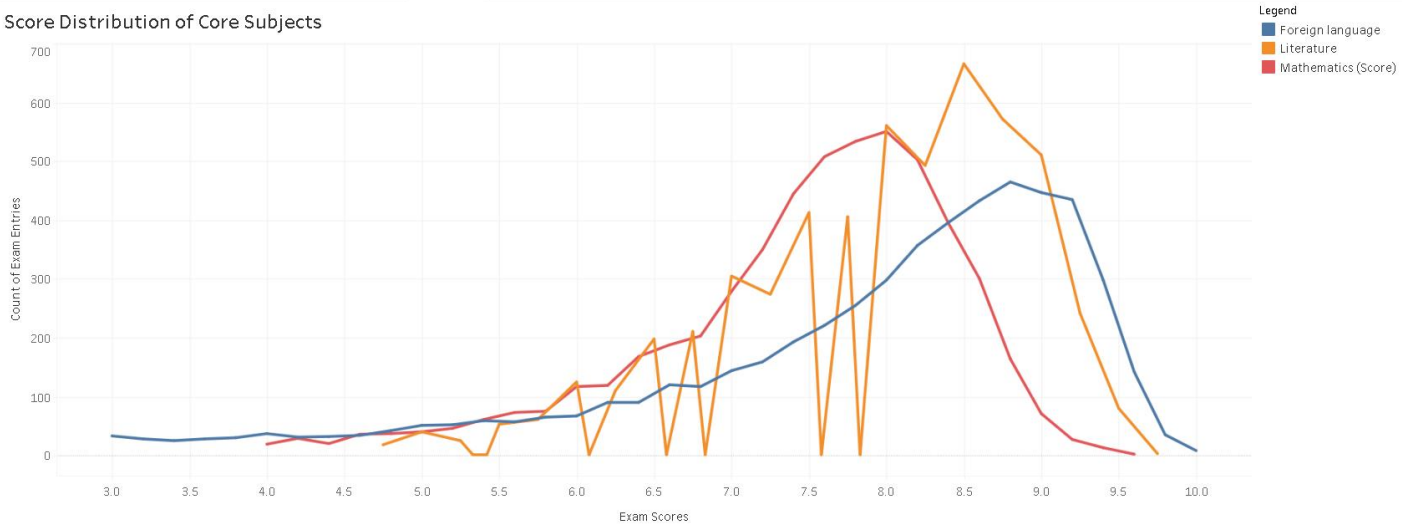

Nhận xét: Biểu đồ thanh ngang trên cung cấp cái nhìn tổng quan về phân bổ điểm trung bình theo từng môn. Các thông tin quan sát được từ biểu đồ như sau:

- Giáo dục công dân (Civic Education) là môn có điểm trung bình cao nhất đạt 8.47. Dựa vào thực tế và số liệu có thể phản ánh điểm thi môn giáo dục công dân được đánh giá với tiêu chí nhẹ nhàng hơn các môn thi khác
- Các môn Toán, Văn, Anh có điểm trung bình tương đối cao lần lượt là 7.46, 7.9, 7.86. Đây là ba môn chính trong kỳ thi và được đầu tư nhiều trong quá trình học tập nên số liệu hoàn toàn hợp lý
- Các môn tổ hợp còn lại với phổ điểm trung bình trong khoảng 6 đến 7, điều này phản ánh các môn tổ hợp ít được đầu tư hơn 3 môn chính hoặc các điểm không được đầu tư đồng đều do có nhiều khối thi, trong đó Sinh học là môn có điểm trung bình thấp nhất chỉ đạt 6.17

### *c) Biểu đồ phân phối điểm theo 3 môn thi chính (Toán, Văn, Anh)*

Biểu đồ Đường:

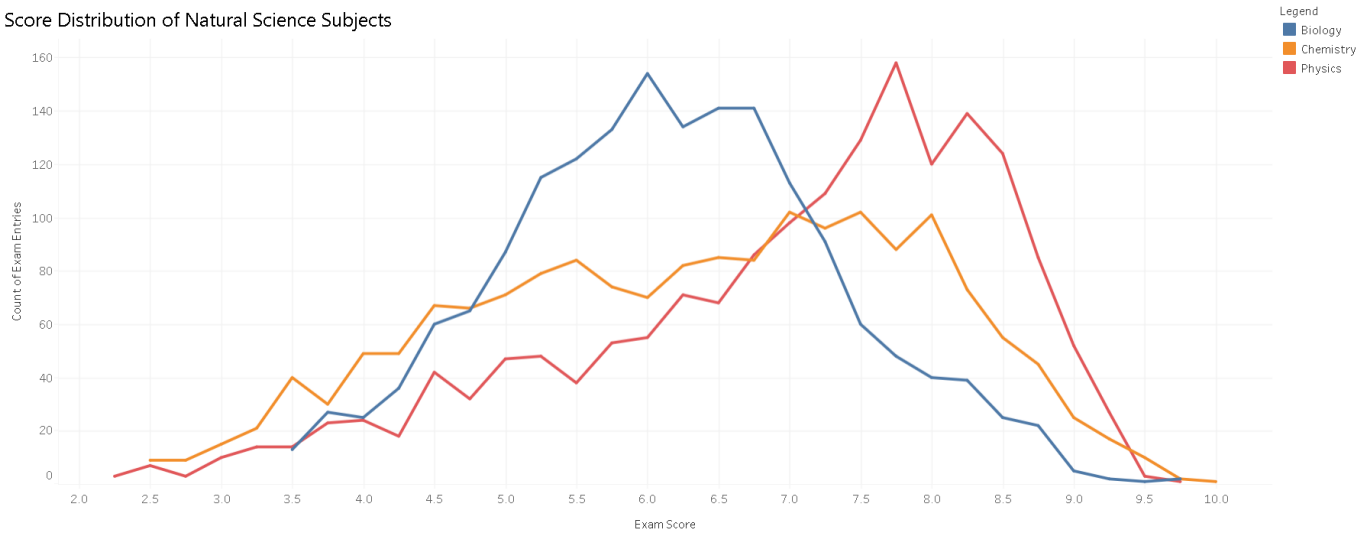
- Thể hiện sự thay đổi liên tục và xu hướng của dữ liệu, dễ nhận diện các mức điểm phổ biến hoặc sự phân tán điểm giữa các môn học.
- Trực quan hóa dữ liệu liên tục: Thể hiện rõ ràng phân phối của dữ liệu có nhiều giá trị.
- Dễ hiểu và đơn giản: Dễ dàng giải thích cho người dùng và giúp họ hiểu nhanh chóng về phân phối dữ liệu.
- Phân tán dữ liệu: Giúp nhận biết mức độ phân tán trong dữ liệu.
- Xác định mô hình phân phối: Giúp nhận diện các phân phối phổ biến trong dữ liệu.




Nhận xét::

- Phân phối điểm của môn ngoại ngữ (Foreign language) - đường màu xanh
  - Đỉnh biểu đồ nằm ở mức 8.5 – 9.0 cho thấy phần lớn thí sinh đạt diểm rất cao
  - Phân phối điểm cao hơn mức trung bình (6.5 - 9.0) chiếm ưu thế, phản ánh các nội dung kiểm tra tập trung vào những kiến thức cơ bản hoặc kỹ năng mà học sinh đã được chuẩn bị tốt
  - So với các môn khác, đường biểu đồ của Ngoại ngữ ít dao động và không có hiện tượng đột ngột tăng/giảm, cho thấy phân phối điểm đồng đều hơn.
- Phân phối điểm của môn Toán (Mathematics) – đường màu đỏ
  - Đỉnh biểu đồ nằm ở mức điểm 8.0 cho thấy phần lớn các thí sinh đạt điểm khá tốt cho môn toán
  - Dải điểm trải dài từ 4.0 đến 10.0, nhưng số lượng bài thi đạt điểm thấp (dưới 5.0) và điểm rất cao (trên 9.0) không nhiều
  - So với Ngoại ngữ, môn Toán có phân phối điểm khá đồng đều nhưng hẹp hơn và tập trung vào khoảng khá – giỏi
- Phân phối điểm môn Ngữ Văn (Literature) – đường màu vàng
  - Đỉnh biểu đồ nằm ở mức điểm 8.5, cho thấy phần lớn thí sinh đạt điểm khá cao
  - Đường biểu đồ có nhiều dao động, lên xuống đột ngột ở nhiều mức điểm cho thấy không có sự đồng đều trong phân phối điểm vì môn Ngữ Văn có thể cảm nhận và chấm điểm theo nhiều góc độ không có đáp án cụ thể nên việc có nhiều mức điểm lạ gây dao động là hoàn toàn dễ hiểu
  - Tuy nhiên điểm số tập trung ở mức 8.0-9.0 cao hơn nhiều so với Toán và Ngoại ngữ.

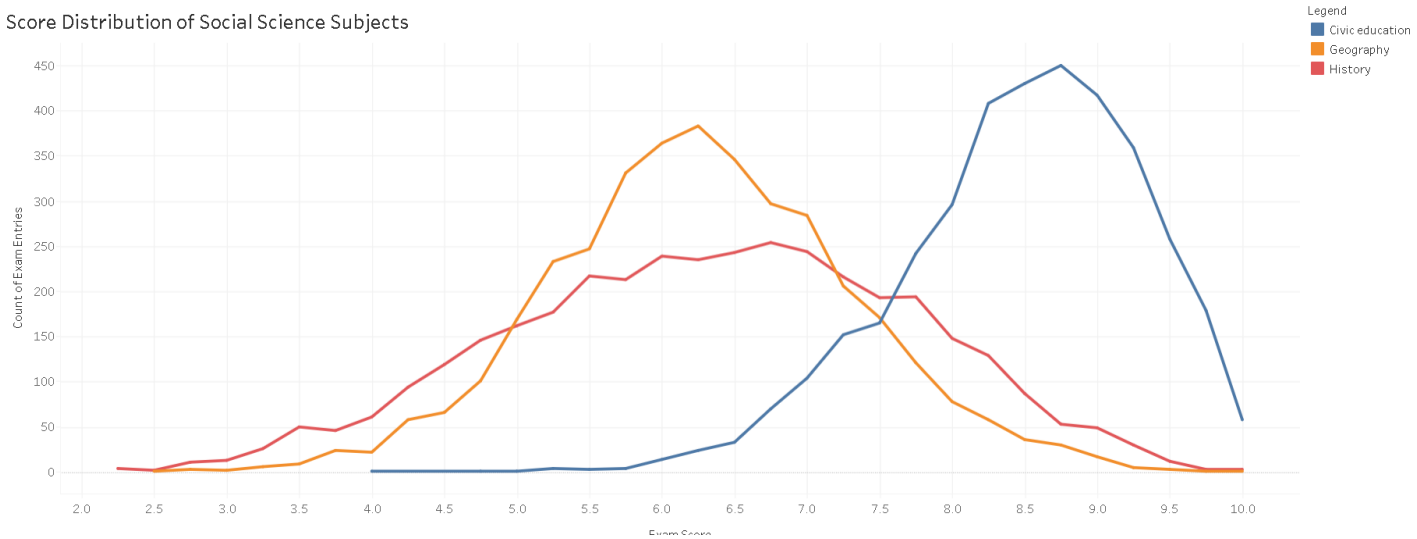
### *d) Biểu đồ phân phối điểm theo 3 môn thi tổ hợp tự nhiên (Lý, Hóa, Sinh)*




Nhận xét::

- Phân phối điểm của môn Sinh học (Biology) - đường màu xanh
  - Điểm số trải dài từ 2.0 đến 9.5, nhưng tập trung chủ yếu trong khoảng 4.0 - 6.5
  - Đỉnh phân phối ở mức 5.5, là mức điểm phổ biến nhất, cho thấy đa số học sinh chỉ đạt mức trung bình trong môn Sinh học.
  - Điểm ở mức giỏi từ 8.0 thấp hơn rất nhiều so với 2 môn tự nhiên còn lại, có thể lí giải một phần do nhiều khối thi không lấy điểm môn sinh trong tổ hợp tự nhiên mà chủ yếu từ 2 môn còn lại
- Phân phối điểm của môn Vật lý (Physics) – đường màu đỏ
  - Điểm số trải dài từ 2.0 đến 9.5, với đỉnh phân phối ở mức 7.5-8.0
  - Học sinh đạt điểm trong khoảng 6.5 đến 8.5 chiếm đa số, phản ánh kết quả thi môn vật lý khá tốt so với 2 môn tự nhiên còn lại
- Phân phối điểm môn Hóa học (Chemistry) – đường màu vàng
  - Điểm số trải dài từ 2.0 đến 10.0, nhưng phân phối không tập trung mạnh vào một mức điểm cụ thể.
  - Điểm phổ biến nhất nằm ở mức 7.0, tuy nhiên không nổi bật nhiều so với các mức lân cận (6.5 - 7.5).
  - Phần lớn học sinh đạt điểm trong khoảng 5.5 đến 8.0, thể hiện năng lực học tập ổn định và đồng đều ở môn Hóa học..

## *e) Biểu đồ phân phối điểm theo 3 môn thi tổ hợp xã hội (Sử, Địa, GDCD)*



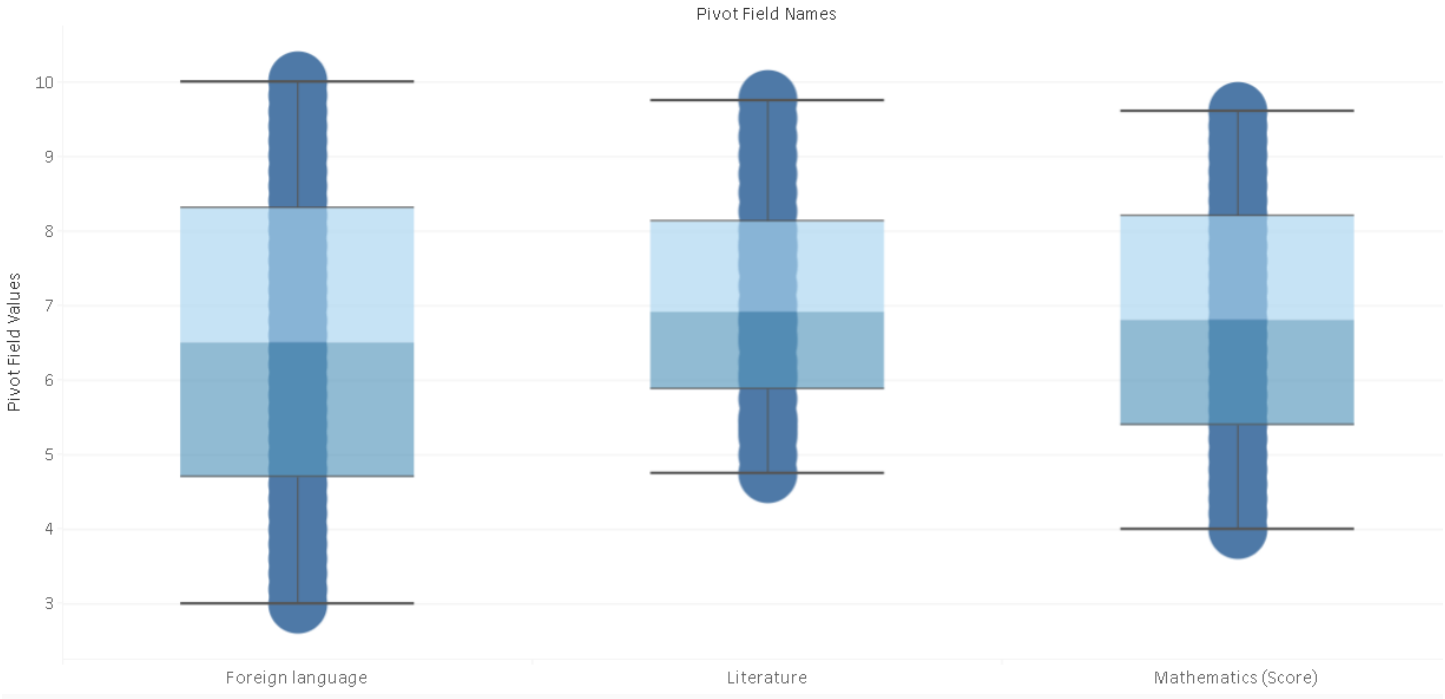

Nhận xét:

- Phân phối điểm của môn Giáo dục công dân (Civic Education) - đường màu xanh
  - Đỉnh biểu đồ nằm ở mức 8.5 – 9.0 cho thấy phần lớn thí sinh đạt điểm rất cao
  - Điểm số dao động từ 4.0- 10, với số lượng bài thi phân bổ chủ yếu ở mức khá giỏi, đường không có dao động lớn, phản ánh sự ổn định trong kết quả.
  - .Là môn có kết quả cao nhất trong ba môn xã hội, số lượng học sinh đạt điểm rất cao (trên 9.0) đáng kể
- Phân phối điểm của môn Lịch sử (History) – đường màu đỏ
  - Điểm số môn Lịch sử phân bổ tương đối đều trong khoảng từ 3.0 đến 8.0, với đỉnh điểm ở mức 6.0.
  - Phân phối điểm không tập trung mạnh vào một mức điểm cụ thể, mà tương đối đều trong khoảng từ 4.0 đến 7.0, cho thấy sự đa dạng trong kết quả thi
- Phân phối điểm môn Địa lý(Geography) – đường màu vàng
  - Điểm tập trung ở mức 5.5 - 7.5, với đỉnh ở khoảng 6.5,
  - Điểm số dao động từ 2.0 đến 9.0, với số lượng bài thi phân bổ tương đối đồng đều trong khoảng trung bình đến khá.
  - Điểm cao (trên 8.0) không phổ biến, cho thấy đề thi Địa lý năm 2023 tương đối thách thức.

## *f) Biểu đồ phân tích kết quả điểm thi các môn chính (Toán, Văn, Anh)*

Biểu đồ hộp:

- Hiển thị trực quan toàn diện: nhanh chóng nắm bắt được trung vị (median), khoảng giá trị chính (IQR), điểm thấp nhất và cao nhất, cũng như các điểm ngoại lai (outliers) trong dữ liệu
- Dễ dàng phát hiện sự biến động: Các điểm ngoại lai của boxplot dễ dàng được nhận diện, giúp xác định dữ liệu bất thường hoặc giá trị xa trung tâm
- So sánh nhiều tập dữ liệu: giúp phát hiện sự khác biệt hoặc tương đồng giữa các đối tượng trên biểu đồ

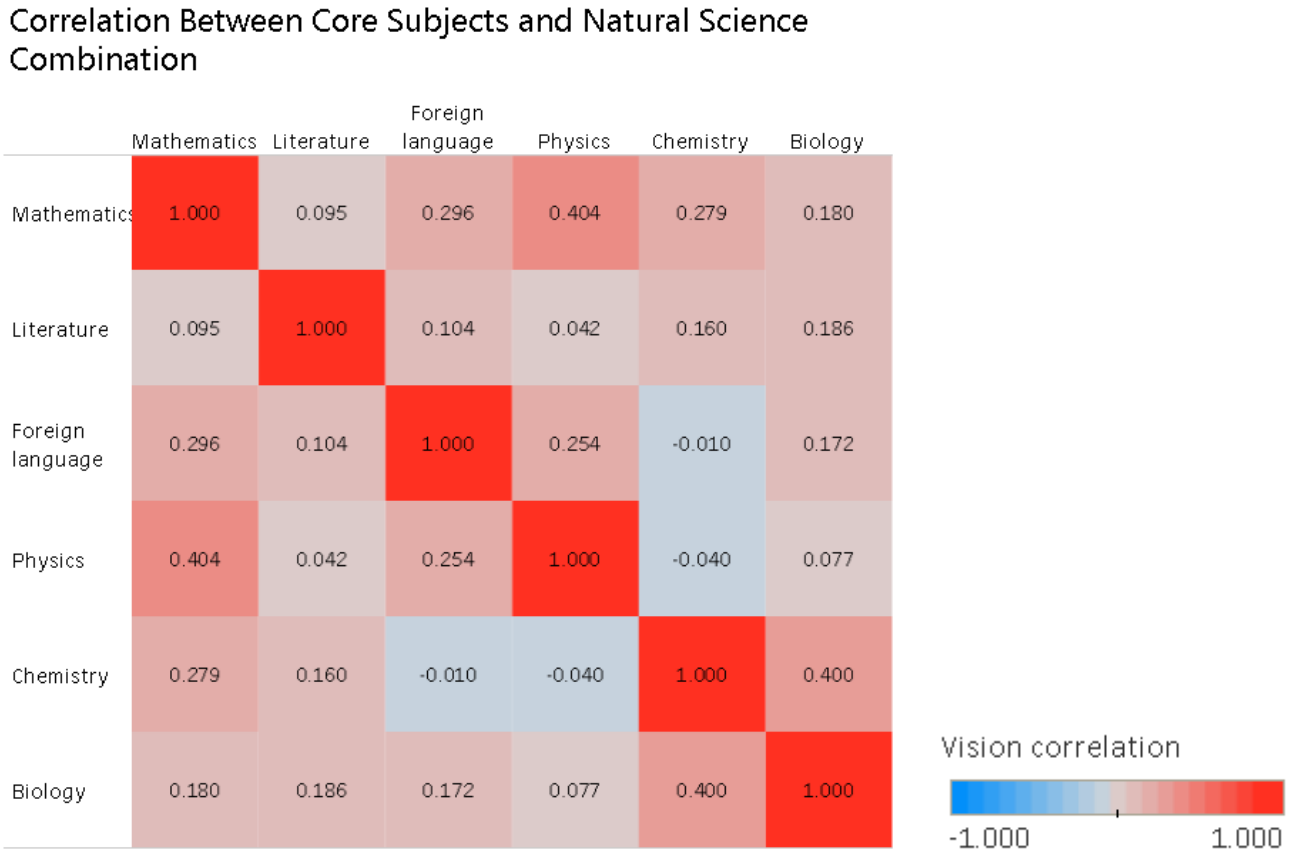



Nhận xét:

- Trung vị và phân phối:
  - Trung vị của cả ba môn đều nằm trong khoảng từ 6.5 - 7.0, cho thấy điểm số tập trung ở mức khá trở lên.
  - Môn Ngoại ngữ có trung vị thấp hơn một chút so với Toán và Ngữ văn.
- Độ rộng của phân phối:
  - Cả ba môn đều có điểm thấp nhất và cao nhất trải dài từ 3.0 đến 10.0, nhưng Ngoại ngữ có sự phân tán lớn hơn so với Toán và Ngữ văn.
- Kết quả đồng đều:
  - Kết quả học tập ở ba môn khá ổn định, với phần lớn học sinh đạt từ mức trung bình đến khá giỏi, và rất ít điểm ngoại lai.

## 2. Phân tích mối quan hệ giữa các thuộc tính

### *a) Mối quan hệ giữa điểm các môn chính (Toán, Văn, Anh) với các môn tổ hợp tự nhiên (Lý, Hóa, Sinh)*




Biểu đồ nhiệt:

- Dễ dàng nhận biết xu hướng và mối tương quan giữa các biến.
- Dễ so sánh với các kiểu dữ liệu dạng số

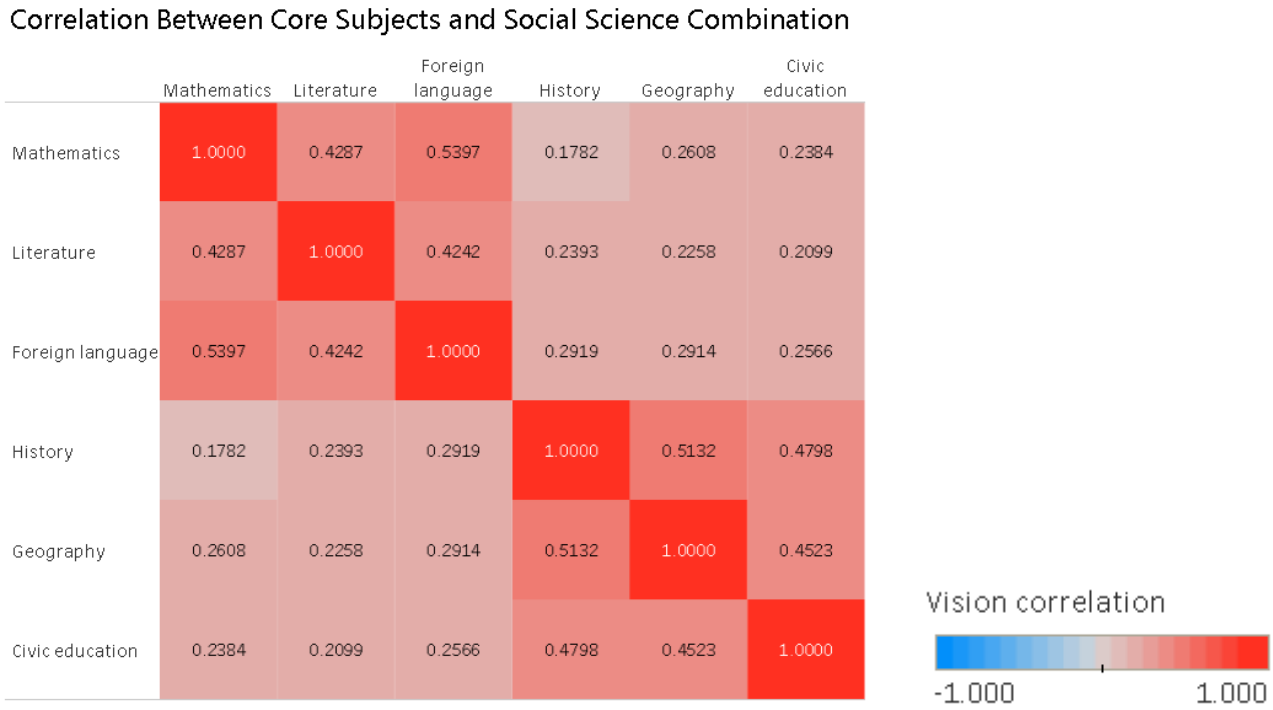
Nhận xét:

Qua biểu đồ nhóm có thể thấy được tương quan chính giữa các môn

| Môn thi | Hệ số tương quan | Nhận xét |
|---|---|---|
| Toán và Lý | 0.404 | Tương quan mạnh, toán là nền tảng quan trọng cho các môn tự nhiên đặc biệt là Lý |
| Hóa và Sinh | 0.4 | Tương quan mạnh do sự liên hệ về kiến thức giữa 2 môn |
| Toán và Hóa | 0.279 | Tương quan trung bình, Toán có tác động đến Hóa nhưng không mạnh bằng Lý |
| Ngoại ngữ và Lý | 0.254 | Tương quan trung bình, Học sinh điểm tốt Ngoại ngữ có xu hướng điểm tốt môn Lý hơn các môn tự nhiên khác |

- Mức tương quan thấp giữa Toán và Ngữ văn có thể được lý giải bởi đây là dữ liệu của các thí sinh thi tổ hợp tự nhiên. Những thí sinh chọn thi tổ hợp tự nhiên thường không chú trọng vào các môn xã hội như Ngữ văn, hoặc họ sẽ chọn tổ hợp xã hội trong trường hợp thi khối thi Toán – Văn - Anh để dễ đạt điểm hơn.
- Các khối thi phổ biến nhất trong tổ hợp tự nhiên là Toán - Lý - Hóa, Toán - Hóa - Sinh và Toán - Lý - Anh. Điều này lý giải tại sao Toán có tương quan mạnh với Lý và Hóa, trong khi Ngoại ngữ cũng có mối liên hệ nhất định với Lý, bởi khối Toán - Lý - Anh thường được lựa chọn bởi những học sinh có năng lực đồng đều giữa Toán, Lý và Ngoại ngữ.
- Nhìn chung, sự tương quan giữa điểm thi các môn học này phản ánh hoàn toàn hợp lý cách học và định hướng chọn khối thi của học sinh trong tổ hợp tự nhiên.

### *b) Mối quan hệ giữa điểm các môn chính (Toán, Văn, Anh) với các môn tổ hợp xã hội (Sử, Địa, GDCD)*




Nhận xét:

Qua biểu đồ nhóm có thể thấy được tương quan chính giữa các môn

| Môn thi | Hệ số tương quan | Nhận xét |
|---|---|---|
| Toán và Văn | 0.4287 | Các môn bắt buộc có tương quan mạnh với nhau, thí sinh thi tổ hợp xã hội có xu hướng học đều hết 3 môn bắt buộc thay vì bỏ qua văn và ngoại ngữ như tổ hợp tự nhiên |
| Toán và Ngoại ngữ | 0.5397 | |
| Văn và Ngoại ngữ | 0.4242 | |
| Các môn tổ hợp xã hội với nhau | > 0.4 | Tương quan mạnh, một môn điểm tốt thì điểm 2 môn còn lại cũng có xu hướng tốt theo |

- Sự tương quan rất cao giữa Toán, Văn, và Ngoại ngữ trong tập dữ liệu này có thể được lý giải bởi đây là dữ liệu được trích xuất từ một phần nhỏ của tập gốc. Tập dữ liệu này chủ yếu đến từ các thí sinh có định hướng khối Toán - Văn - Anh, dẫn đến mối tương quan mạnh giữa ba môn bắt buộc này.
- Các khối thi xã hội, như Văn - Sử - Địa (khối C), chỉ chiếm tỷ lệ không đáng kể trong tập dữ liệu, nên mối tương quan giữa Ngữ văn với các môn xã hội (Lịch sử, Địa lý, Giáo dục công dân) chỉ đạt mức trung bình.
- Tuy nhiên, trong nhóm các môn xã hội, mối tương quan giữa Lịch sử, Địa lý và GDCD vẫn cao vì các thí sinh có xu hướng ôn luyện đồng đều ba môn này khi học tổ hợp xã hội  để đảm bảo kết quả tốt

# V. Tài liệu tham khảo

https://www.tableau.com/chart

https://youtu.be/X0g7GQsPSIA?si=5pc4uq7HCo3RT7Us

https://www.tableau.com/blog/beginners-guide-tableau-public

https://chartio.com/resources/tutorials/what-is-a-box-plot/

https://www.geeksforgeeks.org/z-score-for-outlier-detection-python/