

BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC KINH TẾ TP. HỒ CHÍ MINH



ĐỒ ÁN MÔN HỌC
BIỂU DIỄN TRỰC QUAN DỮ LIỆU

TRỰC QUAN HÓA DỮ LIỆU ĐIỂM THI THPTQG
GIAI ĐOẠN 2020-2023

Họ và tên	MSSV	Lớp
Vũ Nguyễn Thảo Vy	31211027686	DS001
Nguyễn Quốc Việt	31211027687	DS001
Nguyễn Thanh Vy	31211027689	DS002
Nguyễn Nhật Thảo Vy	31211025542	DS001

GVHD: TS. Nguyễn An Tế
TP. Hồ Chí Minh, Ngày 25 tháng 11 năm 2023

Mục lục

1 CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI	7
1.1 Giới thiệu đề tài	7
1.2 Tổng quan về bộ dữ liệu	7
2 CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU	8
2.1 Xóa và đồng nhất tên cột	13
2.2 Thêm thuộc tính	15
2.3 Xử lý dữ liệu bị thiếu	16
2.4 Xử lý dữ liệu không nhất quán	18
2.5 Xử lý dữ liệu không hợp lệ	19
3 CHƯƠNG 3: THÔNG KÊ MÔ TẢ	20
3.1 Kiểm tra các đại lượng về xu thế trung tâm	20
3.1.1 Tính toán các đại lượng về xu thế trung tâm	20
3.1.2 Biểu diễn trực quan các đại lượng về xu thế trung tâm	22
3.2 Kiểm tra các đại lượng về độ phân tán	28
3.2.1 Mức độ phân tán điểm thi các môn vào năm 2020	30
3.2.2 Mức độ phân tán điểm thi các môn vào năm 2021	31
3.2.3 Mức độ phân tán điểm thi các môn vào năm 2022	32
3.2.4 Mức độ phân tán điểm thi các môn vào năm 2023	33
3.3 Kiểm tra các đại lượng về hình dáng phân phối	35
3.3.1 Phân phối điểm thi các môn bắt buộc	36
3.3.2 Phân phối điểm thi các môn thuộc tổ hợp KHTN	40
3.3.3 Phân phối điểm thi các môn thuộc tổ hợp KHXH	42
3.4 Kiểm tra đại lượng về sự tương quan	45
4 CHƯƠNG 4: BIỂU DIỄN TRỰC QUAN DỮ LIỆU	47
4.1 Biểu đồ thể hiện số lượng thí sinh theo tổ hợp	47
4.2 Biểu đồ thể hiện tỷ lệ thí sinh theo tổ hợp	49
4.3 Số lượng thí sinh dự thi trong giai đoạn 2020 – 2023	51
4.4 Biểu đồ thống kê top 10 tỉnh/ thành phố có số lượng thí sinh dự thi cao nhất từ năm 2020 đến năm 2023	52
4.5 Pie chart theo khoảng điểm	56
4.6 Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc (Toán, Ngữ văn, Ngoại ngữ) thấp nhất	59
4.7 Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc (Toán, Ngữ văn, Ngoại ngữ) cao nhất	62

4.8	Tổng số lượng điểm 10 trong giai đoạn 2020 – 2023	66
4.8.1	Số lượng điểm 10 ở các môn học trong năm 2020	68
4.8.2	Số lượng điểm 10 ở các môn học trong năm 2021	68
4.8.3	Số lượng điểm 10 ở các môn học trong năm 2022	69
4.8.4	Số lượng điểm 10 ở các môn học trong năm 2023	69
4.9	Số lượng điểm 10 theo từng môn học	70
4.10	Biểu đồ Parallel Set thể hiện số thí sinh đạt điểm 10 môn Ngữ Văn theo Năm, Tổ hợp và Miền	74
4.11	Top 10 tỉnh có số lượng thí sinh đạt điểm 10 theo từng môn nhiều nhất .	77
4.12	Top 10 tỉnh có số lượng thí sinh đạt điểm liệt (<=1) theo từng môn nhiều nhất	79
4.13	Biểu đồ thể hiện số lượng bài thi có điểm liệt (<=1) theo từng môn . .	81
4.14	Biểu đồ thể hiện tỷ lệ % bài thi bị điểm liệt của 2 tổ hợp theo từng năm .	84
4.15	Biểu đồ điểm trung bình bài thi tổ hợp	85
4.16	Biểu đồ phân phối tổng điểm thi 2 khối KHTN và KHXH qua các năm .	87
4.17	Phân phối các khối xét tuyển đại học (A00, A01, B, C, D) theo từng năm	89
4.18	Biểu đồ thể hiện sự tương quan giữa điểm Toán và điểm trung bình 3 môn trong tổ hợp KHTN	93
4.19	Biểu đồ cột thể hiện số lượng thí sinh thi các môn ngoại ngữ khác . .	94
4.20	Biểu đồ cột thể hiện điểm trung bình các môn ngoại ngữ khác	97
4.21	Biểu đồ cột ngang thống kê số lượng thí sinh thi các môn ngoại ngữ khác theo tỉnh thành	97
4.22	Bản đồ Choropleth số lượng thí sinh thi tiếng Trung năm 2023	98
4.23	Bản đồ thể hiện điểm trung bình môn Tiếng Anh theo tỉnh/thành	99
5	CHƯƠNG 5: Giảm chiều dữ liệu bằng phương pháp phân tích thành phần chính PCA và phân cụm dữ liệu bằng KMeans	102
5.1	Giảm chiều dữ liệu bằng phương pháp phân tích thành phần chính PCA	102
5.2	Phân cụm dữ liệu bằng thuật toán KMeans	105
6	CHƯƠNG 6: Kết luận	108

Danh sách hình vẽ

1	Biểu đồ Histogram điểm thi môn Toán THPTQG giai đoạn 2020-2023	23
2	Biểu đồ Histogram điểm thi môn Ngữ Văn THPTQG giai đoạn 2020-2023	24
3	Biểu đồ Histogram điểm thi môn Ngoại Ngữ THPTQG giai đoạn 2020-2023	24
4	Biểu đồ Histogram điểm thi môn Vật Lý và Hóa học THPTQG giai đoạn 2020-2023	25
5	Biểu đồ Histogram điểm thi môn Sinh học THPTQG giai đoạn 2020-2023	25
6	Biểu đồ Histogram điểm thi môn Lịch Sử THPTQG giai đoạn 2020-2023	26
7	Biểu đồ Histogram điểm thi môn Địa Lý THPTQG giai đoạn 2020-2023	27
8	Biểu đồ Histogram điểm thi môn GDCD THPTQG giai đoạn 2020-2023	27
9	Biểu đồ Box Plot các môn thi THPTQG năm 2020	30
10	Biểu đồ Box Plot các môn thi THPTQG năm 2021	31
11	Biểu đồ Box Plot các môn thi THPTQG năm 2022	32
12	Biểu đồ Box Plot các môn thi THPTQG năm 2023	33
13	Biểu đồ phân phối xác suất môn Toán giai đoạn 2020-2023	37
14	Biểu đồ phân phối xác suất môn Ngữ Văn giai đoạn 2020-2023	38
15	Biểu đồ phân phối xác suất môn Tiếng Anh giai đoạn 2020-2023	39
16	Biểu đồ phân phối xác suất môn Vật Lý giai đoạn 2020-2023	40
17	Biểu đồ phân phối xác suất môn Hóa học giai đoạn 2020-2023	41
18	Biểu đồ phân phối xác suất môn Sinh học giai đoạn 2020-2023	42
19	Biểu đồ phân phối xác suất môn Lịch sử giai đoạn 2020-2023	43
20	Biểu đồ phân phối xác suất môn Địa lý giai đoạn 2020-2023	44
21	Biểu đồ phân phối xác suất môn GDCD giai đoạn 2020-2023	45
22	Heatmap thể hiện sự tương quan giữa các điểm thi	46
23	Biểu đồ thể hiện thí sinh theo tổ hợp các năm	48
24	Biểu đồ thể hiện tỷ lệ số lượng thí sinh theo tổ hợp qua các năm	50
25	Số lượng thí sinh dự thi trong giai đoạn 2020 – 2023	52
26	Số lượng thí sinh theo tỉnh/thành phố năm 2020	54
27	Số lượng thí sinh theo tỉnh/thành phố năm 2021	54
28	Số lượng thí sinh theo tỉnh/thành phố năm 2022	55
29	Số lượng thí sinh theo tỉnh/thành phố năm 2023	55
30	Pie chart theo khoảng điểm môn Toán	57
31	Pie chart theo khoảng điểm môn Ngữ Văn	57
32	Pie chart theo khoảng điểm môn Ngoại Ngữ	58
33	Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc thấp nhất năm 2020	60

34	Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc thấp nhất năm 2021	60
35	Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc thấp nhất năm 2022	61
36	Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc thấp nhất năm 2023	61
37	Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc cao nhất năm 2020	63
38	Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc cao nhất năm 2021	64
39	Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc cao nhất năm 2022	65
40	Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc cao nhất năm 2023	65
41	Tổng số lượng điểm 10 trong giai đoạn 2020 – 2023	67
42	Biểu đồ thể hiện số lượng điểm 10 theo từng môn năm 2020	68
43	Biểu đồ thể hiện số lượng điểm 10 theo từng môn năm 2021	68
44	Biểu đồ thể hiện số lượng điểm 10 theo từng môn năm 2022	69
45	Biểu đồ thể hiện số lượng điểm 10 theo từng môn năm 2023	69
46	Tổng số lượng điểm 10 môn Toán qua các năm	71
47	Tổng số lượng điểm 10 môn Ngữ Văn qua các năm	72
48	Tổng số lượng điểm 10 môn Ngoại Ngữ qua các năm	73
49	Tổng số lượng điểm 10 các môn theo từng năm	74
50	Số lượng thí sinh đạt điểm 10 môn Ngữ Văn theo năm, tổ hợp thi, và miền	76
51	Top 10 tỉnh có số lượng thí sinh đạt điểm 10 theo từng môn nhiều nhất	78
52	Top 10 tỉnh có số lượng thí sinh đạt điểm liệt (<=1) theo từng môn nhiều nhất	80
53	Biểu đồ thể hiện số lượng thí sinh đạt điểm liệt (<=1) theo từng môn năm 2020	82
54	Biểu đồ thể hiện số lượng thí sinh đạt điểm liệt (<=1) theo từng môn năm 2021	82
55	Biểu đồ thể hiện số lượng thí sinh đạt điểm liệt (<=1) theo từng môn năm 2022	83
56	Biểu đồ thể hiện số lượng thí sinh đạt điểm liệt (<=1) theo từng môn năm 2023	83
57	Biểu đồ thể hiện số lượng thí sinh đạt điểm liệt (<=1) theo từng môn năm 2023	85

58	Biểu đồ điểm trung bình bài thi tổ hợp	86
59	Top 10 tỉnh có số lượng thí sinh đạt điểm liệt (<=1) theo từng môn nhiều nhất	88
60	Phân phối các khối A0	90
61	Phân phối các khối A1	90
62	Phân phối các khối B	91
63	Phân phối các khối C	92
64	Phân phối các khối D	92
65	Biểu đồ thể hiện sự tương quan giữa điểm Toán và điểm trung bình 3 môn trong tổ hợp KHTN	94
66	Biểu đồ cột thể hiện số lượng thí sinh thi các môn ngoại ngữ khác	96
67	Biểu đồ cột thể hiện điểm trung bình các môn ngoại ngữ khác	97
68	Biểu đồ cột thể hiện số lượng thí sinh thi các môn ngoại ngữ khác	98
69	Bản đồ thể hiện điểm trung bình môn Tiếng Anh theo tỉnh/thành	101
70	Quy trình thực hiện PCA. <i>Nguồn: Machine Learning Cơ Bản</i>	102
71	Đồ thị biểu diễn % phương sai tích lũy theo số features	104
72	Dữ liệu trước và sau khi thực hiện giảm chiều bằng phương pháp PCA. .	105
73	Dữ liệu đã được phân cụm sử dụng phương pháp Kmeans	107

Danh sách bảng

1	Bảng mô tả các thuộc tính trong bộ dữ liệu	8
---	--	---

LỜI MỞ ĐẦU

Trong kỷ nguyên kỹ thuật số, dữ liệu đã, đang, và sẽ tiếp tục đóng một vai trò quan trọng hơn bao giờ hết trong đa dạng các lĩnh vực. Dữ liệu đồng thời đang trải qua sự bùng nổ chưa từng có tiền lệ. Theo báo cáo của chuyên trang Exploding Topics, khoảng 28.77 triệu terabytes dữ liệu đang được tạo ra mỗi ngày. Việc trực quan hóa dữ liệu vì vậy đang đóng một vai trò quan trọng hơn bao giờ hết trong các lĩnh vực khác nhau. Trực quan hóa dữ liệu giúp chuyển đổi dữ liệu phức tạp thành các hình ảnh, biểu đồ, từ đó tăng cường sự hiểu biết, hỗ trợ ra quyết định và hỗ trợ việc giao tiếp và truyền tải các thông điệp.

Dữ liệu có thể được xem là một nguồn tài nguyên quý giá giúp hỗ trợ và tối ưu hóa quá trình ra quyết định trong mọi mặt của cuộc sống, trong đó bao gồm cả lĩnh vực giáo dục. Theo báo cáo của Tập đoàn Tư vấn đa quốc gia McKinsey & Company năm 2011, Giáo dục là một trong những lĩnh vực phải đổi mới với những thách thức khi không có nguồn tài nguyên dữ liệu cần thiết. Kỳ thi Trung học Phổ Thông Quốc Gia (THPTQG), diễn ra trong khoảng thời gian từ tháng 6 - tháng 7 hàng năm là một trong những sự kiện giáo dục nhận được sự quan tâm nhất của toàn thể phụ huynh và học sinh trên phạm vi cả nước. Kỳ thi nhằm mục đích xét tốt nghiệp bậc học THPT đồng thời điểm thi sẽ được dùng để các thí sinh xét tuyển vào Đại học. Đồ án **Trực quan hóa dữ liệu điểm thi THPTQG giai đoạn 2020-2023** vì vậy góp phần giúp phân tích xu hướng điểm số, đánh giá hiệu suất học tập, và tìm ra các thông tin quan trọng về chất lượng giáo dục. Bằng cách trực quan hóa dữ liệu, đề tài này giúp chúng ta hiểu rõ hơn về chất lượng hệ thống giáo dục trên phạm vi cả nước và cung cấp cơ sở để cải thiện chất lượng học tập.

Chúng em xin chân thành cảm ơn TS Nguyễn An Tế, người phụ trách hướng dẫn trực tiếp học phần Biểu Diễn Trực Quan Dữ Liệu. Bằng tất cả những kinh nghiệm, sự nhiệt huyết và tận tâm, thầy luôn đưa đến chúng em những bài học đầy bổ ích cho quá trình sự nghiệp trở thành những người làm việc trong lĩnh vực Khoa học Dữ Liệu sau này. Thầy luôn truyền tải những khái niệm, nguyên tắc để chúng em hiểu rõ từng bản chất một cách dễ hiểu, gần gũi và chân thật nhất. Dù đôi lúc khắt khe, chúng em nhận thức được rằng đó là điều cần thiết để có thể phát triển bản thân trong học tập và sự nghiệp. Một lần nữa, chúng em xin chân thành cảm ơn thầy.

1 CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

1.1 Giới thiệu đề tài

Kỳ thi Trung Học Phổ Thông Quốc Gia (THPTQG) bắt đầu từ năm 2015, là sự kết hợp giữa Kỳ thi tốt nghiệp trung học phổ thông và Kỳ thi tuyển sinh đại học và cao đẳng theo quyết định số 358 của Bộ Giáo Dục và Đào Tạo. Kể từ năm 2017, hình thức thi được chuyển sang trắc nghiệm 100% trừ môn Ngữ Văn. Trong đó, thí sinh phải thi ba môn bắt buộc bao gồm Toán, Văn, Anh và một trong hai tổ hợp Khoa Học Tự Nhiên (KHTN) hoặc Khoa học Xã Hội (KHXH). Các môn thuộc tổ hợp KHTN bao gồm các môn Vật Lý, Hoá học, Sinh học và tổ hợp KHXH bao gồm các môn Lịch Sử, Địa Lý, GDCCD.

Kỳ thi THPTQG luôn là mối quan tâm hàng đầu của nhiều người, từ học sinh, phụ huynh, các đơn vị giáo dục và đơn vị quản lý ngành giáo dục. Kết quả thi phản ánh khách quan kết quả học tập của các thí sinh và chất lượng dạy học ở các địa phương và còn là một phương tiện để các trường đại học đánh giá xét tuyển. Vì vậy, việc phân tích, xem xét kỹ kết quả của kỳ thi qua các năm cho chúng ta cái nhìn cơ bản về hiệu quả của việc dạy và học cũng như giúp các thí sinh có những định hướng phù hợp với họ.

Đồ án này được thực hiện dựa trên kết quả điểm thi THPTQG giai đoạn 2020-2023. Mục tiêu của đồ án là khảo sát, so sánh, đánh giá và biểu diễn trực quan kết quả thi của các thí sinh trên cả nước. Từ đó đưa ra những đánh giá, nhận xét về kỳ thi này.

1.2 Tổng quan về bộ dữ liệu

Bộ dữ liệu “Điểm thi THPT Quốc gia” chứa dữ liệu về thông tin cơ bản của thí sinh, và điểm thi các môn trong kỳ thi tốt nghiệp THPT từ năm 2020 đến năm 2023. Trong đó, điểm thi của từng năm được khai thác từ các website tra cứu điểm thi và được đăng tải tại các trang Github lần lượt cho các năm 2020, 2021¹, 2022², và 2023³.

Bộ dữ liệu sử dụng trong bài được tổng hợp và điều chỉnh từ 4 bộ dữ liệu của từng năm (2020, 2021, 2022, 2023). Bộ dữ liệu tổng hợp bao gồm 12 thuộc tính và 3.875.378 quan sát. Các thuộc tính trong bộ dữ liệu hoàn chỉnh sau khi đã được tiền xử lí có thể được mô tả trong [Bảng 1](#).

¹<https://github.com/khoingo123/diem-thi-dai-hoc-2021>

²<https://github.com/khoingo123/diem-thi-dai-hoc-2022>

³https://github.com/anhdung98/diem_thi_2023

STT	Tên Thuộc Tính	Mô tả	Ghi chú
1	ID	Số báo danh của thí sinh	Hai chữ số đầu trong SBD của thí sinh đại diện cho mã tỉnh, thông tin này sẽ được dùng để thêm thuộc tính Tỉnh ở Phân 2.2
2	Math	Điểm thi môn Toán	
3	Literature	Điểm thi môn Ngữ văn	
4	Physics	Điểm thi môn Vật lý	
5	Chemistry	Điểm thi môn Hóa học	
6	Biology	Điểm thi môn Sinh học	
7	Geography	Điểm thi môn Địa lý	
8	Ethics	Điểm thi môn Giáo dục công dân	
9	ForeignLanguage	Điểm thi môn Ngoại ngữ	
10	Year	Năm thi của thí sinh	Được bổ sung trong bước tiền xử lý
11	Province	Tỉnh thành nơi thí sinh dự thi	Được bổ sung trong bước tiền xử lý
12	Complex	Lựa chọn bài thi tổ hợp của thí sinh	Được bổ sung trong bước tiền xử lý

Bảng 1: Bảng mô tả các thuộc tính trong bộ dữ liệu

2 CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU

Vì chi tiết của các 4 bộ dữ liệu của từng năm không đồng bộ, cần tiến hành một số bước xử lý để đồng bộ và tổng hợp thành 1 bộ dữ liệu nhất quán.

Thông tin của bộ dữ liệu năm 2020:

¹ df_2020.info()

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 870486 entries, 0 to 870485
Data columns (total 14 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   Unnamed: 0    870486 non-null   int64  
 1   SBD         870486 non-null   int64  
 2   Tên        870486 non-null   float64 
 3   Ngày Sinh  870486 non-null   float64 
 4   Giới tính   870486 non-null   float64 
 5   Toán        866581 non-null   float64 
 6   Văn          856565 non-null   float64 
 7   Lý           293287 non-null   float64 
 8   Hoá          295536 non-null   float64 
 9   Sinh         290377 non-null   float64 
 10  Lịch Sử     568581 non-null   float64 
 11  Địa Lý     555072 non-null   float64 
 12  GDCD        482980 non-null   float64 
 13  Ngoại Ngữ  772098 non-null   float64 
dtypes: float64(12), int64(2)
memory usage: 93.0 MB

```

Quan sát dữ liệu

```

1 print(df_2020.head(), '\n-----')
2 print(df_2020.tail())

```

	Unnamed: 0	SBD	Tên	Ngày Sinh	Giới tính	Toán	Văn	Lý	Hoá	\
0	0	18014547	NaN	NaN	NaN	6.4	6.75	NaN	NaN	
1	1	18014530	NaN	NaN	NaN	7.6	6.00	NaN	NaN	
2	2	18014521	NaN	NaN	NaN	4.8	4.75	NaN	NaN	
3	3	18014517	NaN	NaN	NaN	8.0	7.00	NaN	NaN	
4	4	18014523	NaN	NaN	NaN	8.2	6.50	8.0	8.5	
		Sinh	Lịch Sử	Địa Lý	GDCD	Ngoại Ngữ				
0	NaN	4.75	7.00	6.50		4.2				
1	NaN	3.75	7.75	7.75		2.8				
2	NaN	4.00	6.50	NaN		NaN				
3	NaN	8.25	8.00	9.50		5.8				
4	5.0	NaN	NaN	NaN		4.0				

	Unnamed: 0	SBD	Tên	Ngày Sinh	Giới tính	Toán	Văn	Lý	Hoá	\
870481	870481	17014593	NaN	NaN	NaN	4.6	4.00	NaN	NaN	
870482	870482	17014597	NaN	NaN	NaN	7.4	8.00	NaN	NaN	
870483	870483	17014595	NaN	NaN	NaN	8.2	8.75	NaN	NaN	
870484	870484	17014598	NaN	NaN	NaN	4.6	6.50	NaN	NaN	
870485	870485	60005601	NaN	NaN	NaN	8.0	7.75	NaN	NaN	
		Sinh	Lịch Sử	Địa Lý	GDCD	Ngoại Ngữ				
870481	NaN	3.75	6.75	6.25		2.0				
870482	NaN	5.50	7.50	9.50		2.6				
870483	NaN	4.50	6.50	8.50		6.2				
870484	NaN	4.50	7.50	9.00		3.6				
870485	NaN	4.25	6.50	8.25		3.4				

Thông tin của bộ dữ liệu năm 2021:

```
1 df_2021.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 987391 entries, 0 to 987390
Data columns (total 14 columns):
 #   Column      Non-Null Count  Dtype  
 --- 
  0   Unnamed: 0    987391 non-null   int64  
  1   SBD         987391 non-null   int64  
  2   Tên          86367 non-null    object  
  3   Ngày Sinh   86359 non-null    object  
  4   Giới tính    86367 non-null    object  
  5   Toán         977151 non-null   float64 
  6   Văn          974146 non-null   float64 
  7   Lý           345024 non-null   float64 
  8   Hóa          346650 non-null   float64 
  9   Sinh         341241 non-null   float64 
  10  Lịch Sử       634465 non-null   float64 
  11  Địa Lý        628649 non-null   float64 
  12  GDCD         532021 non-null   float64 
  13  Ngoại Ngữ    868006 non-null   float64 
dtypes: float64(9), int64(2), object(3)
memory usage: 105.5+ MB
```

Quan sát dữ liệu

```
1 print(df_2021.head(), '\n-----')
2 print(df_2021.tail())
```

```
Unnamed: 0      SBD  Tên Ngày Sinh Giới tính  Toán  Văn   Lý   Hoá  Sinh \
0            0  1000043  NaN     NaN     NaN  8.0  8.50  NaN  NaN  NaN
1            1  1000163  NaN     NaN     NaN  3.8  7.50  NaN  NaN  NaN
2            2  1000040  NaN     NaN     NaN  3.8  2.00  NaN  NaN  NaN
3            3  1000007  NaN     NaN     NaN  9.0  5.25  7.25  4.75  3.5
4            4  1000180  NaN     NaN     NaN  8.8  8.50  NaN  NaN  NaN

Lịch Sử  Địa Lý  GDCD  Ngoại Ngữ
0      5.75    6.50   8.75     8.8
1      6.75    7.00    NaN     NaN
2      2.50    4.50    NaN     NaN
3      NaN     NaN     NaN     9.0
4      5.25    7.25   8.00     9.6

Unnamed: 0      SBD  Tên Ngày Sinh Giới tính  Toán  Văn   Lý   Hoá \
987386  987386  32008526  NaN     NaN     NaN  5.8  6.00  NaN  NaN
987387  987387  32008527  NaN     NaN     NaN  5.6  6.00  NaN  NaN
987388  987388  32008528  NaN     NaN     NaN  6.6  8.00  NaN  NaN
987389  987389  32008529  NaN     NaN     NaN  6.0  7.25  NaN  NaN
987390  987390  32008530  NaN     NaN     NaN  3.8  5.75  NaN  NaN

Sinh  Lịch Sử  Địa Lý  GDCD  Ngoại Ngữ
987386  NaN     3.75    5.50    7.5     3.2
987387  NaN     3.00    5.00    6.5     4.8
987388  NaN     4.75    6.50    8.0     4.0
987389  NaN     4.50    7.00    8.5     4.6
987390  NaN     2.25    5.25    6.5     2.8
```

Thông tin của bộ dữ liệu năm 2022:

```
1 df_2022.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 995441 entries, 0 to 995440
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   sbd        995441 non-null   int64  
 1   toan       982726 non-null   float64 
 2   ngu_van    981407 non-null   float64 
 3   ngoai_ngu  870609 non-null   float64 
 4   vat_li     325523 non-null   float64 
 5   hoa_hoc    327367 non-null   float64 
 6   sinh_hoc   322198 non-null   float64 
 7   lich_su    659662 non-null   float64 
 8   dia_li     657421 non-null   float64 
 9   gcdcd     554343 non-null   float64 
dtypes: float64(9), int64(1)
memory usage: 75.9 MB
```

Quan sát dữ liệu

```
1 print(df_2022.head(), '\n-----')
2 print(df_2022.tail())
```

```
      sbd  toan  ngu_van  ngoai_ngu  vat_li  hoa_hoc  sinh_hoc  lich_su \
0  1000001    3.6    5.00      4.0    NaN     NaN     NaN     2.75
1  1000002    8.4    6.75      7.6    NaN     NaN     NaN     8.50
2  1000003    5.8    7.50      5.0    NaN     NaN     NaN     7.25
3  1000004    7.4    7.50      8.6    NaN     NaN     NaN     7.50
4  1000005    7.2    8.50      9.0    NaN     NaN     NaN     8.00

      dia_li  gcdcd
0      6.0    8.75
1      7.5    8.25
2      5.5    8.75
3      6.5    7.50
4      8.5    8.25
-----
      sbd  toan  ngu_van  ngoai_ngu  vat_li  hoa_hoc  sinh_hoc  lich_su \
995436  64006584    8.4    6.75      4.6    NaN     NaN     NaN
995437  64006585    5.6    6.50      2.8    NaN     NaN     NaN
995438  64006586    5.8    6.00      6.6    NaN     NaN     NaN
995439  64006587    7.6    6.75      7.0    NaN     NaN     NaN
995440  64006588    6.6    4.50      3.2    NaN     NaN     NaN

      lich_su  dia_li  gcdcd
995436    6.50    6.75    9.00
995437    6.25    6.75    8.50
995438    7.25    8.00    8.00
995439    8.75    7.25    9.75
995440    3.00    6.00    7.50
```

Thông tin của bộ dữ liệu năm 2023:

```
1 df_2023.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1022060 entries, 0 to 1022059
Data columns (total 11 columns):
 #   Column      Non-Null Count  Dtype  
 ---  --          --          --      
 0   sbd         1022060 non-null  int64  
 1   toan        1003373 non-null  float64 
 2   ngu_van    1008239 non-null  float64 
 3   ngoai_ngu  880997 non-null  float64 
 4   vat_li     327189 non-null  float64 
 5   hoa_hoc    328118 non-null  float64 
 6   sinh_hoc   324625 non-null  float64 
 7   lich_su    683447 non-null  float64 
 8   dia_li     682134 non-null  float64 
 9   gdcd       565452 non-null  float64 
 10  ma Ngoai_ngu 880997 non-null  object  
dtypes: float64(9), int64(1), object(1)
memory usage: 85.8+ MB
```

Quan sát dữ liệu

```
1 print(df_2023.head(), '\n-----')
2 print(df_2023.tail())
```

```
      sbd  toan  ngu_van  ngoai_ngu  vat_li  hoa_hoc  sinh_hoc  lich_su \
0  1000001   8.4    8.50      9.2     NaN     NaN     NaN     6.75
1  1000002   7.2    8.50      9.2     NaN     NaN     NaN     8.75
2  1000003   NaN    6.50      NaN     NaN     NaN     NaN     9.25
3  1000004   7.8    8.25      7.8     NaN     NaN     NaN     4.50
4  1000005   7.2    8.00      7.8     NaN     NaN     NaN     4.75

      dia_li  gdcd ma Ngoai_ngu
0      6.00  9.00      N1
1      6.50  8.50      N1
2      7.50  NaN       NaN
3      6.25  8.25      N1
4      6.75  8.25      N1
-----
```

```
      sbd  toan  ngu_van  ngoai_ngu  vat_li  hoa_hoc  sinh_hoc  lich_su \
1022055  64006933   7.8    6.75      5.4     NaN     NaN     NaN
1022056  64006934   7.4    7.50      6.0     6.0    5.75     6.25
1022057  64006935   6.4    7.00      3.0     NaN     NaN     NaN
1022058  64006936   6.6    7.00      5.8     NaN     NaN     NaN
1022059  64006937   NaN    5.00      NaN     NaN     NaN     NaN

      lich_su  dia_li  gdcd ma Ngoai_ngu
1022055    8.50    7.75  9.75      N1
1022056    NaN     NaN  NaN       N1
1022057    5.50    5.75  7.25      N1
1022058    6.50    6.50  9.25      N1
1022059    4.25    4.75  NaN       NaN
```

2.1 Xóa và đồng nhất tên cột

Bộ dữ liệu của năm 2020 và 2021 có chứa các cột: ‘Tên’, ‘Ngày sinh’, ‘Giới tính’ đều không có dữ liệu. Đây là thông tin cá nhân của thí sinh nên không thể thu thập được dữ liệu cho 3 cột này. Vì vậy cần tiến hành bỏ đi các cột trên.

Bộ dữ liệu của năm 2023 có cột ‘ma Ngoai ngu’ chứa dữ liệu về mã ngoại ngữ tương ứng với từng ngôn ngữ của môn thi Ngoại ngữ. Tuy nhiên 3 bộ dữ liệu còn lại không có thuộc tính này. Để đồng bộ khi tiến hành tổng hợp 4 bộ dữ liệu, cần bỏ đi cột này. Các phân tích liên quan đến thuộc tính ‘ma Ngoai ngu’ sẽ được tiến hành trên bộ dữ liệu riêng của năm 2023.

```
1 #Xóa cột
2 df_2020 = df_2020.drop(columns= ['Unnamed: 0', 'Tên', 'Ngày Sinh', 'Giới tính'])
3 df_2021 = df_2021.drop(columns= ['Unnamed: 0', 'Tên', 'Ngày Sinh', 'Giới tính'])
4 df_2023 = df_2023.drop(columns=[ 'ma Ngoai ngu'])
```

Tên của cùng 1 thuộc tính trong 4 bộ dữ liệu không giống nhau, vì vậy cần tiến hành đổi tên các cột cho nhất quán.

```
1 #Đồng bộ tên cột
2 df_2020.rename(columns={
3     'SBD' : 'ID',
4     'Toán' : 'Math',
5     'Văn' : 'Literature',
6     'Lý' : 'Physics',
7     'Hoá' : 'Chemistry',
8     'Sinh' : 'Biology',
9     'Lịch Sử' : 'History',
10    'Địa Lý' : 'Geography',
11    'GDCH' : 'Ethics',
12    'Ngoại Ngữ' : 'ForeignLanguage',
13    'Year' : 'Year',
14    'code' : 'Code',
15    'province' : 'Province'
16 }, inplace=True)
17
18
19 df_2021.rename(columns={
20     'SBD' : 'ID',
21     'Toán' : 'Math',
22     'Văn' : 'Literature',
```

```

23     'Lý'          : 'Physics',
24     'Hoá'         : 'Chemistry',
25     'Sinh'        : 'Biology',
26     'Lịch Sử'    : 'History',
27     'Địa Lý'     : 'Geography',
28     'GDCD'       : 'Ethics',
29     'Ngoại Ngữ'  : 'ForeignLanguage',
30     'Year'        : 'Year',
31     'code'        : 'Code',
32     'province'   : 'Province'
33 }, inplace=True)

34

35 df_2022.rename(columns={
36     'sbd'          : 'ID',
37     'toan'         : 'Math',
38     'ngu_van'      : 'Literature',
39     'ngoai_ngu'    : 'ForeignLanguage',
40     'vat_li'       : 'Physics',
41     'hoa_hoc'      : 'Chemistry',
42     'sinh_hoc'     : 'Biology',
43     'lich_su'      : 'History',
44     'dia_li'       : 'Geography',
45     'gdcd'         : 'Ethics'
46 }, inplace=True)

47

48

49 df_2023.rename(columns={
50     'sbd'          : 'ID',
51     'toan'         : 'Math',
52     'ngu_van'      : 'Literature',
53     'ngoai_ngu'    : 'ForeignLanguage',
54     'vat_li'       : 'Physics',
55     'hoa_hoc'      : 'Chemistry',
56     'sinh_hoc'     : 'Biology',
57     'lich_su'      : 'History',
58     'dia_li'       : 'Geography',
59     'gdcd'         : 'Ethics'
60 }, inplace=True)

```

2.2 Thêm thuộc tính

Trước khi tổng hợp 4 bộ dữ liệu thành 1 bộ dữ liệu cuối cùng, cần tạo thêm cột ‘Year’ và gán giá trị năm tương ứng cho mỗi bộ dữ liệu.

```
1 #Gán nhãn năm
2 df_2020['Year'] = '2020'
3 df_2021['Year'] = '2021'
4 df_2022['Year'] = '2022'
5 df_2023['Year'] = '2023'
```

Tổng hợp các bộ dữ liệu của các năm 2020, 2021, 2022, 2023 vào 1 DataFrame.

```
1 #Hợp nhất data các năm
2 df = pd.concat([df_2020, df_2021, df_2022, df_2023], ignore_index=True)
```

Sau khi quan sát dữ liệu, tiến hành thêm cột ‘Province’ thể hiện thông tin về tỉnh thành nơi mà thí sinh dự thi. Bằng cách dựa vào số báo danh (‘ID’) của thí sinh, có thể lấy được mã của từng tỉnh tương ứng (hàm `create_province_code(x)`). Đọc file ‘diaphantinh.geojson’ và đưa vào DataFrame `map_df`, trong đó có dữ liệu mã tỉnh ứng với tên tỉnh. Tạo dictionary mapping có mã tỉnh là khóa (key) và tên tỉnh là giá trị (value) tương ứng. Tạo cột ‘Province’, trong đó giá trị của mỗi hàng được xác định bằng cách áp dụng hàm `create_province_code(x)` cho ‘ID’ của hàng đó, sau đó lấy giá trị tương ứng từ từ điển mapping.

```
1 #Xác định province_code
2 def create_province_code(x):
3     if len(str(x)) == 7:
4         return '0' + str(x)[0]
5     return str(x)[:2]
6
7 import geopandas as gpd
8 !gdown 14EBf0pzL4-UQxGwW4xsIzDiIERid3LDP
9
10 map_df = gpd.read_file('diaphantinh.geojson')
11 map_df.head()
12
13 code_provinces = ['51', '52', '18', '11', '60', '19', '56',
14                 '37', '44', '43', '47', '61', '55', '06',
15                 '04', '40', '63', '62', '48', '50', '38',
```

```

16             '05', '24', '01', '30', '21', '03', '64',
17             '23', '22', '41', '54', '36', '07', '42',
18             '10', '08', '49', '25', '29', '27', '45',
19             '15', '39', '31', '34', '35', '17', '32',
20             '59', '14', '46', '26', '12', '28', '33',
21             '53', '02', '58', '09', '57', '16', '13']
22
23 mapping = {code_provinces[i]: map_df['ten_tinh'].unique().tolist()[i] for i in
24     range(len(code_provinces))}
```

2.3 Xử lý dữ liệu bị thiếu

Tạo hàm check_missing_value để kiểm tra số lượng và tỷ lệ dữ liệu bị thiếu của từng cột.

```

1 def check_missing_values(df):
2     missing_values = df.isnull().sum()
3     missing_percentage = (missing_values / df.shape[0]) * 100
4     return pd.DataFrame({'Missing values': missing_values, 'Percentage (%)':
5         round(missing_percentage,2)})
6 check_missing_values(df)
```

	Missing values	Percentage (%)
ID	0	0.00
Math	45547	1.18
Literature	55021	1.42
Physics	2584355	66.69
Chemistry	2577707	66.51
Biology	2596937	67.01
History	1329223	34.30
Geography	1352102	34.89
Ethics	1740582	44.91
ForeignLanguage	483668	12.48
Year	0	0.00
Province	0	0.00

Các trường hợp thiếu dữ liệu:

- Thiếu điểm các môn thuộc bài thi tổ hợp (Lý, Hóa, Sinh, Lịch sử, Địa lý, Giáo dục công dân)
- Thiếu điểm các môn chính (Toán, Ngữ Văn, Ngoại ngữ)

Các môn thuộc bài thi của 2 tổ hợp, Khoa học Tự nhiên (viết tắt là KHTN) gồm các môn thi thành phần Vật lý, Hóa học, Sinh học; tổ hợp Khoa học Xã hội (viết tắt là KHXH) gồm các môn thi thành phần Lịch sử, Địa lý, Giáo dục công dân. Theo quy chế thi, mỗi thí sinh chỉ có thể đăng ký dự thi 1 trong 2 tổ hợp này. Vì vậy, trong 6 môn thí sinh chỉ có thể chọn 3 môn, 3 môn không chọn sẽ bị thiếu dữ liệu.

Tuy nhiên vẫn có những thí sinh không có đủ điểm của bài thi tổ hợp. Để lọc ra những trường hợp này cần tạo thêm 1 cột 'Complex' để phân loại thí sinh theo bài thi tổ hợp. Các thí sinh có đủ điểm 3 môn Lý, Hóa, Sinh ('Physics', 'Chemistry', 'Biology') được gán nhãn 'KHTN'. Thí sinh có đủ điểm 3 môn Sử, Địa, Giáo dục công dân ('History', 'Geography', 'Ethics') được gán nhãn 'KHXH'. Trường hợp không có dữ liệu của cả 6 môn sẽ gán giá trị 'None'. Nếu có điểm của cả 6 môn được gán giá trị 'Unknown', đây có thể xem là trường hợp dữ liệu bất thường.

```
1 df['Complex'] = np.where(  
2     df[['Physics', 'Chemistry', 'Biology',  
3           'History', 'Geography', 'Ethics']].notna().all(axis=1), 'Unknown',  
4     np.where(df[['Physics', 'Chemistry', 'Biology']].notna().all(axis=1), 'KHTN',  
5     np.where(df[['History', 'Geography', 'Ethics']].notna().all(axis=1),  
6           'KHXH', 'None'  
7     )))  
8 df[df.Complex=='Unknown']
```

Để xử lý dữ liệu bị thiếu của các trường hợp trên, cần xem xét đến nguyên nhân dẫn đến việc thiếu dữ liệu của từng thuộc tính để đưa ra hướng giải quyết cụ thể. Nguyên nhân của việc thiếu dữ liệu (thiếu điểm Toán hoặc Ngữ Văn hoặc điểm bài thi tổ hợp):

- Dữ liệu bị mất trong quá trình thu thập dữ liệu
- Thí sinh không tham dự đủ các bài thi

Đối với các trường hợp không có đủ điểm bài thi tổ hợp hoặc điểm bài thi môn Toán hoặc Ngữ văn, xử lý bằng cách loại bỏ các dòng này. Vì những thí sinh không đủ điều kiện để xét công nhận tốt nghiệp. Đồng thời không thể đưa ra những phân tích chính xác cho các dòng thiếu dữ liệu này.

```
1 #Xóa các dòng không có điểm tổ hợp
2 df = df.drop(df.loc[df.Complex=='None'].index)
3 #Xóa các dòng không có điểm môn Toán hoặc Ngữ văn
4 df_nan_math_literature = df[df[['Math', 'Literature']].isna().any(axis=1)]
5 df = df.drop(df_nan_math_literature.index)
```

Riêng đối với trường hợp thiếu dữ liệu môn Ngoại ngữ, không thể trực tiếp loại bỏ các dòng này. Vì trong số những dữ liệu bị thiếu này có thể bao gồm cả trường hợp các thí sinh được miễn thi môn Ngoại ngữ. Theo quy chế thi, có 2 đối tượng được miễn thi môn Ngoại ngữ là thành viên đội tuyển quốc gia dự thi Olympic quốc tế môn Ngoại ngữ hoặc có một trong các chứng chỉ ngoại ngữ được quy định bởi Bộ Giáo Dục. Nếu loại bỏ các dòng thiếu dữ liệu này có thể ảnh hưởng đến độ chính xác của các phân tích sau này. Hơn nữa, một số thư viện của Python khi xử lý dữ liệu NaN có thể tự động bỏ qua chúng để đảm bảo kết quả của phân tích không bị ảnh hưởng.

2.4 Xử lý dữ liệu không nhất quán

Xem lại dữ liệu của cột ‘Province’ để kiểm tra có lỗi phát sinh hay không.

```
1 #Xem lại dữ liệu cột Province
2 code_province = df[['Province']].drop_duplicates().sort_values('Province')
3
4 code_province['Province'].unique()
```

```

array(['An Giang', 'Bà Rịa - Vũng Tàu', 'Bình Dương', 'Bình Phước',
       'Bình Thuận', 'Bình Định', 'Bạc Liêu', 'Bắc Giang', 'Bắc Kạn',
       'Bắc Ninh', 'Bến Tre', 'Cao Bằng', 'Cà Mau', 'Cần Thơ', 'Gia Lai',
       'Hà Giang', 'Hà Nam', 'Hà Nội', 'Hà Tĩnh', 'Hòa Bình', 'Hưng Yên',
       'Hải Dương', 'Hải Phòng', 'Hậu Giang', 'Khánh Hòa', 'Kiên Giang',
       'Kon Tum', 'Lai Châu', 'Long An', 'Lào Cai', 'Lâm Đồng',
       'Lạng Sơn', 'Nam Định', 'Nghệ An', 'Ninh Bình', 'Ninh Thuận',
       'Phú Thọ', 'Phú Yên', 'Quảng Bình', 'Quảng Nam', 'Quảng Ngãi',
       'Quảng Ninh', 'Quảng Trị', 'Sóc Trăng', 'Sơn La',
       'TP. Hồ Chí Minh', 'Thanh Hóa', 'Thái Bình', 'Thái Nguyên',
       'Thừa Thiên Huế', 'Tiền Giang', 'Trà Vinh', 'Tuyên Quang',
       'Tây Ninh', 'Vĩnh Long', 'Vĩnh Phúc', 'Yên Bái', 'Điện Biên',
       'Đà Nẵng', 'Đăk Lăk', 'Đăk Nông', 'Đồng Nai', 'Đồng Tháp'],
      dtype=object)

```

Nhận xét: Xuất hiện lỗi chính tả, lỗi đánh máy trong tên của 3 tỉnh: Quảng Bình, Cần Thơ, Kiên Giang, Bà Rịa - Vũng Tàu. Tiến hành sửa các lỗi này để đảm bảo tính chính xác cho dữ liệu.

```

1 #Xử lý typo error
2 df['Province'] = df['Province'].replace({
3     'Quản Bình': 'Quảng Bình',
4     'Cần Thơ': 'Cần Thơ',
5     'Kien Giang': 'Kiên Giang',
6     'Bà Rịa -Vũng Tàu': 'Bà Rịa - Vũng Tàu'
7 })

```

2.5 Xử lý dữ liệu không hợp lệ

Tiếp theo ta tiến hành kiểm tra dữ liệu không hợp lệ đối với các cột điểm. Cụ thể, các điểm có giá trị bé hơn 0 hoặc lớn hơn 10 là không hợp lệ. Có điểm cả 2 tổ hợp hoặc không có điểm tất cả các môn cũng là dữ liệu không hợp lệ nên ta sẽ tiến hành bỏ các dòng dữ liệu này.

```

1 #Thí sinh có điểm thi <0 hoặc >10
2 subjects = ['Math', 'Literature', 'Physics', 'Chemistry', 'Biology', 'History',
   ↳ 'Geography', 'Ethics', 'ForeignLanguage']
3 abnormal = ((df[subjects] < 0) | (df[subjects] > 10)).any(axis=1)
4 df = df.drop(df[abnormal].index)

5
6 #Thí sinh có điểm bài thi cả 2 tổ hợp
7 df = df.drop(df.loc[df.Complex=='Unknown'].index)

```

Lưu dữ liệu

Tiến hành lưu bộ dữ liệu vừa tổng hợp và đã được tiền xử lý để thuận tiện cho việc sử dụng ở công đoạn tiếp theo.

```
1 df.to_csv('diem_thi_thptqg_2020_2023.csv')
```

3 CHƯƠNG 3: THÔNG KÊ MÔ TẢ

3.1 Kiểm tra các đại lượng về xu thế trung tâm

3.1.1 Tính toán các đại lượng về xu thế trung tâm

Nói đến các đại lượng về trung tâm, 3 thước đo được sử dụng rộng rãi nhằm biểu diễn một giá trị thể hiện vị trí/xu thế “trung tâm” của tập dữ liệu bao gồm: trung bình (mean – trung tâm về mặt giá trị), trung vị (median – trung tâm về mặt vị trí) và yếu vị (mode – trung tâm về mức độ tập trung dữ liệu).

Đối với bộ dữ liệu “Điểm thi THPTQG từ năm 2020-2023”, ta sẽ quan sát mean, median và mode điểm của từng môn học theo từng năm.

Trung bình Pythagore (Pythagorean Means)

Để tính điểm thi trung bình cho từng môn học theo từng năm, ta sử dụng trung bình cộng (Arithmetic mean) và làm tròn đến 2 chữ số thập phân.

```
1 # Điểm thi trung bình của 9 môn học theo từng năm
2 subjects = ['Math', 'Literature', 'Physics', 'Chemistry', 'Biology', 'History',
3             'Geography', 'Ethics', 'ForeignLanguage']
4 print('{:^100}'.format('ĐIỂM THI TRUNG BÌNH CỦA 9 MÔN HỌC THEO TỪNG NĂM'))
5 print()
6 df.groupby('Year')[subjects].mean().round(2)
```

ĐIỂM THI TRUNG BÌNH CỦA 9 MÔN HỌC THEO TỪNG NĂM

Year	Math	Literature	Physics	Chemistry	Biology	History	Geography	Ethics	ForeignLanguage
2020	6.87	6.73	6.74	6.70	5.57	5.25	6.86	8.13	4.56
2021	6.80	6.61	6.57	6.62	5.51	5.06	7.06	8.38	5.84
2022	6.64	6.65	6.72	6.68	5.01	6.46	6.77	8.03	5.14
2023	6.45	7.00	6.58	6.74	6.39	6.14	6.23	8.29	5.45

Tuy nhiên, giá trị của trung bình cộng dễ bị ảnh hưởng bởi các giá trị ngoại lệ và các phân phối bất đối xứng. Ví dụ, nếu trong năm đó một số thí sinh xuất sắc thi được môn Toán với điểm rất cao, đặc biệt là so với các thí sinh khác, thì điểm Toán trung bình của năm đó sẽ bị sai lệch theo hướng điểm cao hơn. Điều này xảy ra vì điểm Toán trung bình được tính bằng cách lấy tổng điểm Toán của tất cả các thí sinh và chia cho số thí sinh. Những điểm thi rất cao như vậy sẽ có ảnh hưởng lớn đến điểm Toán trung bình, làm cho nó bị lệch.

Do đó, nếu phân bố dữ liệu không đối xứng và có các giá trị ngoại lệ, thì việc sử dụng giá trị trung bình một cách độc lập có thể không phản ánh chính xác trung tâm của dữ liệu. Trong trường hợp này, các phép đo trung tâm khác như trung vị (median) có thể là sự lựa chọn tốt hơn.

Trung vị (Median)

Trung vị là thước đo trung tâm tốt hơn đối với các tập dữ liệu bất đối xứng hay tập dữ liệu bị tác động bởi giá trị ngoại lệ. Trung vị là một số tách giữa nửa lớn hơn và nửa bé hơn của một mẫu, một quần thể, hay một phân bố xác suất. Trung vị là giá trị giữa, có nghĩa $\frac{1}{2}$ quan sát sẽ có các giá trị nhỏ hơn hay bằng số trung vị, và $\frac{1}{2}$ quan sát sẽ có giá trị bằng hoặc lớn hơn số trung vị.

Điểm thi trung vị cho từng môn học theo từng năm, làm tròn đến 2 chữ số thập phân:

```

1 # Điểm thi trung vị của 9 môn học theo từng năm
2 print('{:^100}'.format('ĐIỂM THI TRUNG VỊ CỦA 9 MÔN HỌC THEO TỪNG NĂM'))
3 print()
4 df.groupby('Year')[subjects].median()

```

ĐIỂM THI TRUNG VỊ CỦA 9 MÔN HỌC THEO TỪNG NĂM

	Math	Literature	Physics	Chemistry	Biology	History	Geography	Ethics	Foreign Language
Year									
2020	7.2	6.75	7.00	7.0	5.50	5.0	7.00	8.25	4.2
2021	7.2	6.75	6.75	7.0	5.50	5.0	7.00	8.50	5.6
2022	7.0	6.75	7.00	7.0	4.75	6.5	6.75	8.25	4.8
2023	6.8	7.00	6.75	7.0	6.50	6.0	6.25	8.50	5.2

Mặc dù giá trị trung vị không chịu ảnh hưởng của các giá trị ngoại lệ và rất dễ tính toán. Tuy nhiên trung vị không thể dùng để dự đoán vì không chính xác bằng trung bình, trung vị thường được dùng để thay thế hoặc bổ sung nhằm điều chỉnh 1 số hạn chế khi sử dụng giá trị trung bình.

Yếu vị (Mode)

Yếu vị là giá trị xuất hiện nhiều lần nhất trong tập dữ liệu. Có tập dữ liệu có 1 mode, có tập dữ liệu có đến 2 hoặc 3 mode và cũng có thể có tập dữ liệu không có mode nào.

```
1 # Mode điểm thi của 9 môn học theo từng năm
2 print('{:^100}'.format('MODE ĐIỂM THI CỦA 9 MÔN HỌC THEO TỪNG NĂM'))
3 print()
4 df.groupby('Year')[subjects].apply(lambda x: x.mode().iloc[0]) # Nếu có hơn 1
→ mode thì lấy mode đầu tiên
```

MODE ĐIỂM THI CỦA 9 MÔN HỌC THEO TỪNG NĂM

	Math	Literature	Physics	Chemistry	Biology	History	Geography	Ethics	ForeignLanguage
Year									
2020	7.8	7.0	7.75	7.75	5.25	4.75	7.25	8.75	3.4
2021	7.8	7.0	7.50	7.75	5.25	4.25	7.25	9.25	4.0
2022	7.8	7.0	7.25	8.00	4.50	7.00	7.00	8.50	3.8
2023	7.6	7.0	7.50	7.50	6.50	6.00	6.25	9.00	4.2

Yếu vị cũng không bị ảnh hưởng bởi các giá trị ngoại lệ. Tuy nhiên, yếu vị chỉ ổn định khi lượng giá trị nhiều và sẽ khó xác định rõ nếu dữ liệu chỉ có một số ít giá trị. Khi dữ liệu ít, một giá trị nào đó có thể xuất hiện nhiều nhất do ngẫu nhiên chứ không thực sự đại diện cho xu hướng chung. Khi có rất nhiều dữ liệu, xác suất giá trị xuất hiện nhiều nhất (mode) là giá trị đặc trưng của quần thể là rất cao.

3.1.2 Biểu diễn trực quan các đại lượng về xu thế trung tâm

Cả 3 đại lượng trên đều không thể hiện quá chính xác về xu thế trung tâm khi đứng một mình và khi không thấy được phân phối của dữ liệu, vì vậy ta cần xem đến biểu đồ tần số khi thể hiện cả 3 đại lượng cùng lúc của 9 môn qua từng năm.

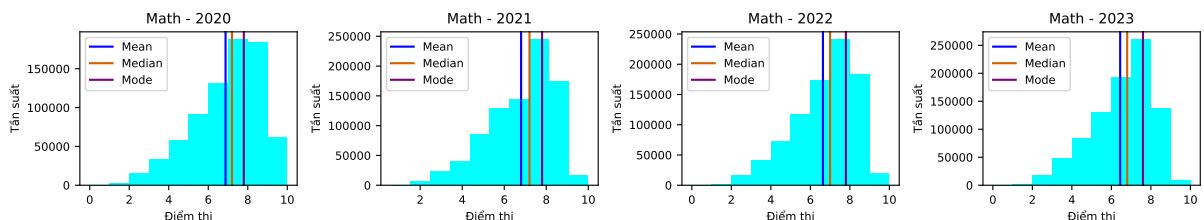
```
1 subjects = ['Math', 'Literature', 'Physics', 'Chemistry', 'Biology',
2             'History', 'Geography', 'Ethics', 'ForeignLanguage']
3 fig, axes = plt.subplots(9,4, figsize=(12, 20))
4
5 for i, subject in enumerate(subjects):
6     mean = df.groupby('Year')[subject].mean()
7     median = df.groupby('Year')[subject].median()
```

```

8     mode = df.groupby('Year')[subject].apply(lambda x: x.mode())
9
10    for j, year in enumerate(df['Year'].unique()):
11        axes[i, j].hist(df[df['Year'] == year][subject], color = 'cyan')
12        axes[i, j].axvline(mean[year], color='b', linewidth=1.5, label='Mean')
13        axes[i, j].axvline(median[year], color='#D55E00', linewidth=1.5,
14                           label='Median')
15        axes[i, j].axvline(mode[year].values[0], color='purple', linewidth=1.5,
16                           label='Mode')
17        axes[i, j].set_title(f'{subject} - {year}', fontsize = 10)
18        axes[i, j].set_xlabel('Điểm thi', fontsize=8)
19        axes[i, j].set_ylabel('Tần suất', fontsize=8)
20        axes[i, j].legend(fontsize=8)
21
22    plt.tight_layout()
23    plt.savefig("figs/Biểu đồ Mean Mode Median.pdf")
24    plt.show()

```

Điểm thi môn Toán các năm 2020 - 2023:



Hình 1: Biểu đồ Histogram điểm thi môn Toán THPTQG giai đoạn 2020-2023

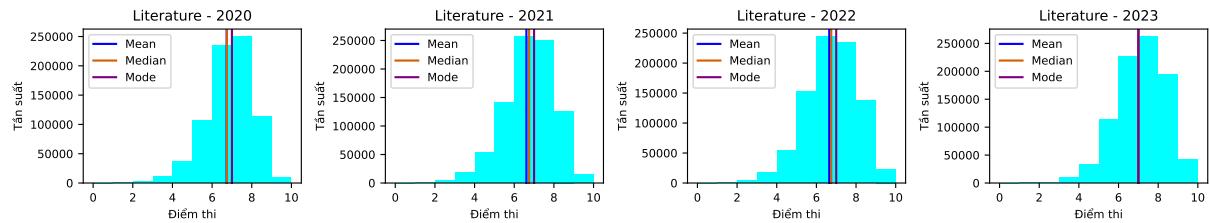
Điểm thi môn Toán của 4 năm đều có $\text{mean} < \text{median} < \text{mode}$ vì thế phân phối không đối xứng và có xu hướng lệch trái.

Kết hợp với số liệu đã tính toán ở trên, ta có thể thấy từ năm 2021-2023, đa số các thí sinh đều có mức điểm dưới 8 và chỉ có phần ít các thí sinh đạt điểm từ 8 trở lên. Điều này ngụ ý rằng đề thi Toán cả 3 năm này đều có sự phân loại cao cho mức điểm dưới 8 và trên 8. Lý do có sự phân loại rõ rệt này là vì trong 50 câu trắc nghiệm, 40 câu đầu là các câu hỏi thuộc cấp độ nhận biết - thông hiểu, chỉ cần nắm vững nền tảng kiến thức là có thể dễ dàng giải quyết, còn 10 câu cuối đòi hỏi mức độ vận dụng cao.

Giá trị mean, mode, median của điểm thi Toán năm 2020 không chênh lệch nhiều so với 3 năm còn lại, tuy nhiên nhìn vào biểu đồ Histogram năm 2020, tỷ lệ thí sinh đạt điểm trên 8 so với tỷ lệ thí sinh đạt điểm dưới 8 không chênh lệch quá nhiều như những năm

2021-2023. Thật vậy, đây là năm mà cả nước phải chịu ảnh hưởng của dịch Covid bùng phát, toàn bộ thí sinh phải học trực tuyến tại nhà, nên nội dung thi được tinh giản dẫn đến đề thi Toán năm 2020 có phần dễ hơn và ít phân hóa hơn.

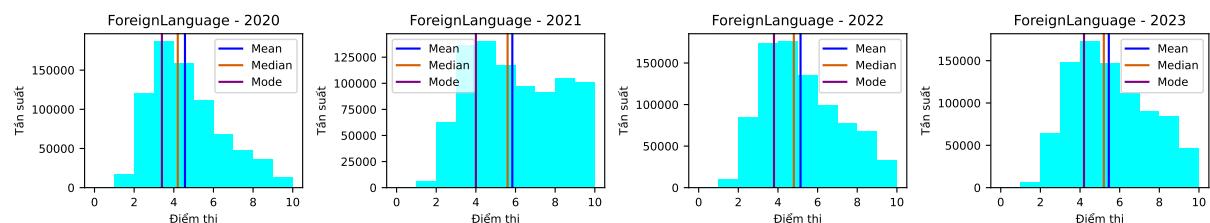
Điểm thi môn Văn các năm 2020 - 2023:



Hình 2: Biểu đồ Histogram điểm thi môn Ngữ Văn THPTQG giai đoạn 2020-2023

Biểu đồ Histogram thể hiện điểm thi môn Văn của 4 năm đều có đường mean, median và mode gần nhau, thậm chí trùng nhau ở năm 2023. Điều này cho thấy phân phối điểm khá đồng đều xung quanh giá trị trung bình là khoảng 7 điểm và ít xuất hiện các trường hợp ngoại lệ có điểm quá cao hay quá thấp. Đề thi Văn của cả 4 năm đều vừa sức, quen thuộc, các kiểu dạng câu hỏi không bất ngờ với thí sinh.

Điểm thi môn Ngoại ngữ các năm 2020 - 2023:



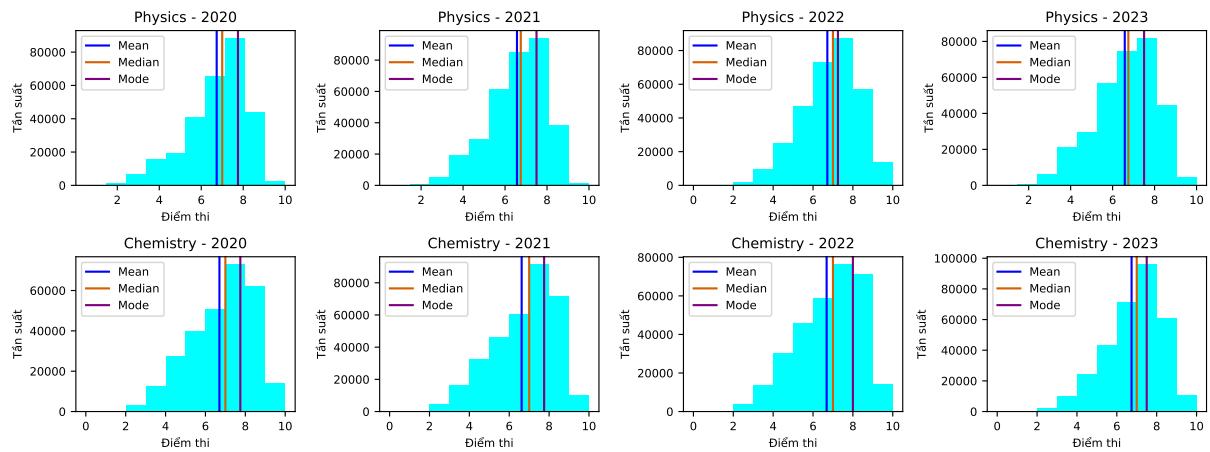
Hình 3: Biểu đồ Histogram điểm thi môn Ngoại Ngữ THPTQG giai đoạn 2020-2023

Điểm thi môn Ngoại ngữ của 4 năm đều có mode < median < mean vì thế phân phối không đối xứng và có xu hướng lệch phải. Đây cũng là môn học có điểm mean, median, mode thấp nhất trong 9 môn (đều dưới 6 điểm). Nguyên nhân thứ nhất là do, nếu không chọn tổ hợp xét tuyển đại học có tiếng Anh như D01 (Toán, Văn, Anh), A01 (Toán, Lý, Anh), A07 (Toán, Hóa, Anh), thí sinh không quá quan tâm đến môn học, mục tiêu chỉ cần không bị điểm liệt. Thứ hai, thí sinh ở các xã nông thôn không có điều kiện học ngoại ngữ. Thậm chí một số nơi thiếu giáo viên, các thí sinh còn không được học bộ môn này.

Tuy nhiên, biểu đồ Histogram điểm Ngoại ngữ năm 2021 có đường mode cách xa đường mean và median, điểm thi tập trung nhiều ở mức 4-5 điểm nhưng điểm trung bình và trung vị lại xấp xỉ 6. Điều này thể hiện rằng đề thi Ngoại ngữ 2021 tương đối dễ, ít phân

hóa ở mức khá (6-8 điểm) và giỏi (8-10 điểm) nhưng sẽ vẫn có nhiều thí sinh không làm được bài và đạt điểm thấp bởi những lý do để kể trên.

Điểm thi môn Vật lý, Hóa học các năm 2020 - 2023:



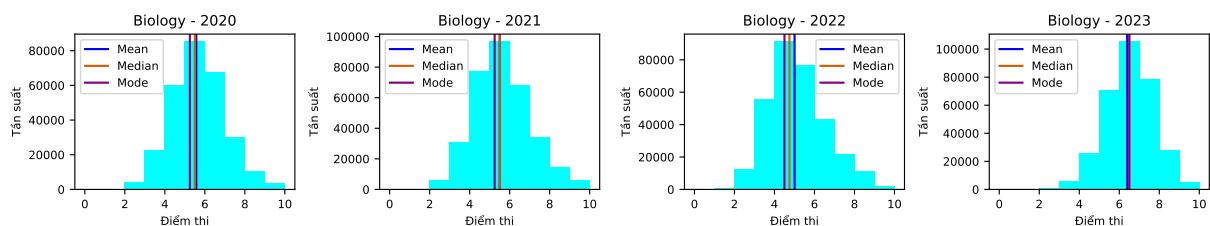
Hình 4: Biểu đồ Histogram điểm thi môn Vật Lý và Hóa học THPTQG giai đoạn 2020-2023

Tương tự với điểm Toán, điểm thi môn Vật lý và Hóa học của 4 năm đều có $\text{mean} < \text{median} < \text{mode}$ vì thế phân phối không đối xứng và có xu hướng lệch trái.

Đối với môn Vật lý, đa số các thí sinh đều có mức điểm dưới 8 và chỉ có phần ít các thí sinh đạt điểm từ 8 trở lên. Có thể thấy rằng đề thi Vật lý cả 4 năm đều có sự phân loại cao cho mức điểm dưới 8 và trên 8.

Đối với môn Hóa học, số lượng thí sinh đạt điểm trên 8 cao hơn so với môn Vật lý, tuy nhiên đa số các thí sinh vẫn đạt điểm dưới 8 và chỉ có phần ít các thí sinh đạt điểm từ 8 trở lên. Vì vậy, đề thi Hóa học cả 4 năm vẫn có sự phân loại tốt cho mức điểm dưới 8 và trên 8.

Điểm thi môn Sinh học các năm 2020 - 2023:

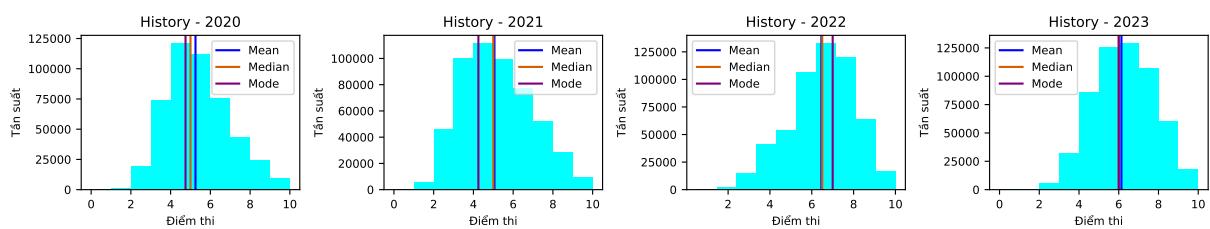


Hình 5: Biểu đồ Histogram điểm thi môn Sinh học THPTQG giai đoạn 2020-2023

Tương tự với điểm Văn, điểm thi môn Sinh học của 4 năm đều có đường mean, median và mode gần nhau, thậm chí trùng nhau ở năm 2023. Điều này cho thấy phân phối điểm

khá đồng đều xung quanh giá trị trung bình và ít xuất hiện các trường hợp ngoại lệ có điểm quá cao hay quá thấp. Tuy nhiên, điểm trung bình, trung vị và yếu vị năm 2023 có phần cao hơn những năm còn lại (trên 6 điểm). Theo thầy Đinh Đức Hiền, giáo viên môn Sinh Hệ thống giáo dục FPT, đề thi Sinh năm 2023 đã có điều chỉnh, loại bỏ gần như hoàn toàn các dạng Toán không mang bản chất Sinh học như trước kia, không còn các dạng Toán phải sử dụng mẹo mực, công thức. Vì thế, điểm thi Sinh năm 2023 có tăng cao hơn những năm còn lại.

Điểm thi môn Lịch sử các năm 2020 - 2023:



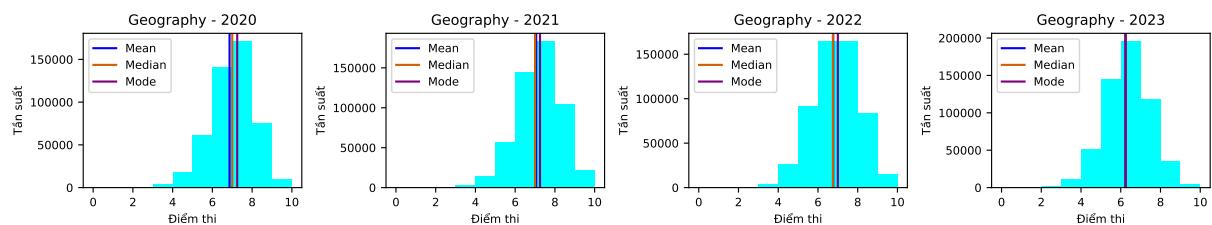
Hình 6: Biểu đồ Histogram điểm thi môn Lịch Sử THPTQG giai đoạn 2020-2023

Điểm thi môn Lịch sử năm 2020 và 2021 có mode < median < mean vì thế phân phối không đối xứng và có xu hướng lệch phải. Tuy nhiên, sự chênh lệch giữa các giá trị mean, median và mode không quá lớn cho thấy rằng điểm phân bố đồng đều xung quanh giá trị trung bình là xấp xỉ 5 điểm. Môn Lịch sử là môn thứ hai có điểm trung bình thấp sau môn Ngoại ngữ. Nguyên nhân thứ nhất là do môn Lịch sử cần sự ghi nhớ và tập trung cao độ nên thí sinh thường có tâm lý sợ, chán môn học này. Nguyên nhân thứ hai là do định hướng của cha mẹ, bởi cha mẹ thường hướng con theo những tổ hợp dễ chọn nghề, chọn trường và dễ tìm việc làm. Trong số những ngành nghề này, ít xuất hiện bồng dáng của môn Lịch sử. Và nguyên nhân cuối cùng phải kể đến là chương trình Lịch sử cho cấp THPT còn dài cùng với phương pháp dạy môn Lịch sử chưa thực sự lôi cuốn.

Với sự điều chỉnh trong chương trình môn Lịch sử, từ năm 2022, điểm thi môn Lịch sử cải thiện rõ rệt, điểm trung bình, trung vị và yếu vị đều tăng, phân phối điểm có xu hướng lệch trái. Theo Phó Giáo sư Nguyễn Mạnh Hưởng, đề thi môn Lịch sử năm 2022 được điều chỉnh nhẹ nhàng hơn so với các năm trước và chỉ có 2 câu thực sự phân hóa, đây cũng là một lý do góp phần cải thiện phổ điểm so với 2 năm trước.

Điểm trung bình, trung vị và yếu vị năm 2023 có phần thấp hơn so với năm 2022 nhưng vẫn cao hơn năm 2020 và 2021. Các đường mean, median, mode trong biểu đồ Histogram môn Lịch sử 2023 trùng nhau cho thấy phân phối điểm khá đồng đều xung quanh giá trị trung bình và ít xuất hiện các trường hợp ngoại lệ có điểm quá cao hay quá thấp.

Điểm thi môn Địa lý các năm 2020 - 2023:

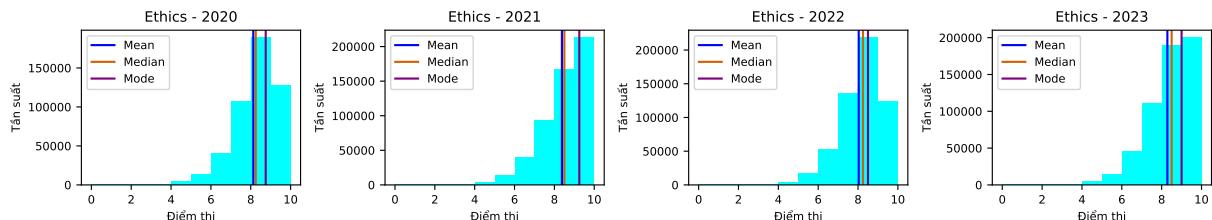


Hình 7: Biểu đồ Histogram điểm thi môn Địa Lý THPTQG giai đoạn 2020-2023

Biểu đồ Histogram điểm thi môn Địa lý các năm 2020-2022 đều có đường mean, median, mode gần nhau. Điều này cho thấy phân phối điểm khá đồng đều xung quanh giá trị trung bình là khoảng 7 điểm và ít xuất hiện các trường hợp ngoại lệ có điểm quá cao hay quá thấp.

Riêng năm 2023, điểm thi trung bình, trung vị và yếu vị môn Địa lý có phần thấp hơn, rơi vào khoảng 6.25 điểm. Đề thi Địa lý năm 2023 được cho là “khó và lạ” hơn vì những năm trước đây, cụ thể là từ 2017-2022, các câu hỏi phần Atlat luôn ghi rõ số trang. Còn năm 2023, câu hỏi chỉ ghi tiêu đề. Thí sinh muốn biết trang Atlat đó nằm ở đâu phải tra mục lục, mất thêm thời gian nhưng những nội dung khó nhất bài thi đều nằm ở câu biểu đồ.

Điểm thi môn GD&CD các năm 2020 - 2023:



Hình 8: Biểu đồ Histogram điểm thi môn GD&CD THPTQG giai đoạn 2020-2023

Điểm thi trung bình, trung vị và yếu vị của môn GD&CD đều cao nhất khi so với 8 môn còn lại, rơi vào khoảng 8-9.25 điểm. Theo Tạp chí giáo dục Việt Nam, nguyên nhân là do chương trình môn GD&CD có số lượng bài học ít, khối lượng kiến thức ít, bên cạnh đó, môn học có nhiều bài tập tình huống yêu cầu học sinh xử lý. Trong khi, nhiều những tình huống xảy ra xung quanh cuộc sống hàng ngày nên học sinh dễ dàng nắm bắt và lựa chọn được những đáp án chính xác.

Nhìn vào biểu đồ Histogram năm 2020 và 2021, ta có thể thấy tỷ lệ thí sinh đạt điểm trên 9 năm 2021 cao hơn nhiều so với tỷ lệ thí sinh đạt điểm trên 9 năm 2020. Năm 2021

không có sự thay đổi về chương trình môn GD&CD, vì vậy lí do có thể đến từ việc đề thi GD&CD năm 2021 tương đối dễ, không có tính phân loại thí sinh.

Giống với cặp biểu đồ Histogram năm 2020 và 2021, khi nhìn vào biểu đồ Histogram năm 2022 và 2023, ta có thể thấy tỷ lệ thí sinh đạt điểm trên 9 năm 2023 cao hơn nhiều so với tỷ lệ thí sinh đạt điểm trên 9 năm 2021.

3.2 Kiểm tra các đại lượng về độ phân tán

Tổng quát về các đại lượng dùng để đo mức độ phân tán của tập dữ liệu:

Tứ phân vị (Quartile)

Tứ phân vị có 3 giá trị, đó là tứ phân vị thứ nhất, thứ nhì, và thứ ba. Ba giá trị này chia một tập hợp dữ liệu đã sắp xếp theo thứ tự thành 4 phần có số lượng quan sát đều nhau.

- Giá trị tứ phân vị thứ nhất Q1 bằng trung vị phần dưới, tương đương với bách phân vị thứ 25.
- Giá trị tứ phân vị thứ hai Q2 chính bằng giá trị trung vị, tương đương với bách phân vị thứ 50.
- Giá trị tứ phân vị thứ ba Q3 bằng trung vị phần trên, tương đương với bách phân vị thứ 75.

Khoảng trai giữa (InterQuartile Range – IQR)

Khoảng trai giữa hay còn gọi là khoảng tứ phân vị của tập dữ liệu. Khoảng trai giữa là một con số cho biết mức độ lan truyền của nửa giữa của tập dữ liệu. IQR thường được sử dụng thay cho khoảng biến thiên (Range) vì nó loại trừ hầu hết giá trị bất thường hay giá trị ngoại lệ (Outliers) của dữ liệu. Công thức tính IQR có dạng: $IQR = Q3 - Q1$

IQR có thể giúp xác định các giá trị ngoại lệ. Một giá trị bị nghi ngờ là một giá trị ngoại lệ nếu nó nhỏ hơn $1,5 * IQR$ dưới phần tư đầu tiên ($Q1 - 1,5 * IQR$) hoặc lớn hơn ($1,5 * IQR$) trên phần tư thứ ba ($Q3 + 1,5 * IQR$).

Để kiểm tra các đại lượng về độ phân tán, nhóm sử dụng biểu đồ hộp (Box Plot) để có cái nhìn tổng quan về phân phối điểm số của từng môn học trong giai đoạn 2020-2023.

- Hộp (box): Đại diện cho phạm vi giữa tứ phân vị thứ nhất (Q1) và tứ phân vị thứ ba (Q3). Kích thước của hộp cho biết sự phân tán của dữ liệu trong khoảng này.

- Các giá trị IQR (Interquartile Range) thể hiện phạm vi giữa Q1 và Q3, và cho biết sự khác biệt giữa dữ liệu trong khoảng này. Khoảng IQR càng lớn, dữ liệu càng phân tán mạnh.
- Đường kẻ ngang trong hộp: Đại diện cho giá trị trung vị của tập dữ liệu.
- Dây nối (whiskers): Đại diện cho phạm vi của dữ liệu, ngoại trừ các điểm ngoại lệ (outliers). Dây nối thường bắt đầu từ đầu hộp và kết thúc tại dây nằm xa hộp, biểu thị phạm vi phân bố dữ liệu.
- Điểm ngoại lệ (outliers): Là các giá trị cách xa phạm vi của dữ liệu. Các điểm ngoại lệ thường được đánh dấu trên biểu đồ.

```

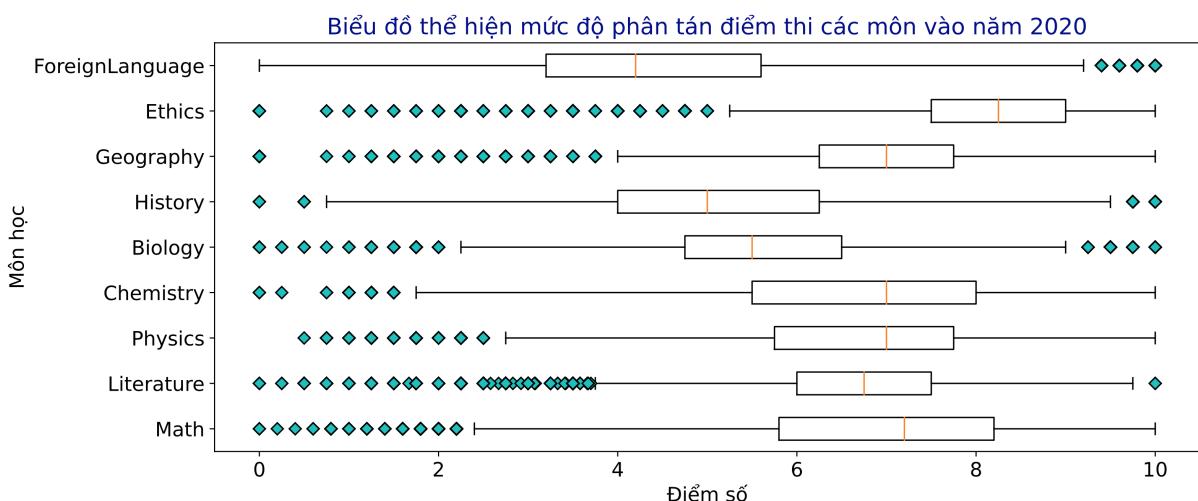
1 subjects = ["Math", "Literature", "Physics", "Chemistry", "Biology", "History",
2   ↵ "Geography", "Ethics", "ForeignLanguage"]
3 fig, axes = plt.subplots(4, 1, figsize=(12, 20))
4 years = [2020, 2021, 2022, 2023]
5 outlier_props = dict(marker='D', markerfacecolor='c', markersize=6)
6
7 for i, year in enumerate(years):
8     data_for_year = df[df['Year'] == year]
9     data_by_subject = [data_for_year[subject].dropna() for subject in subjects]
10    axes[i].boxplot(data_by_subject, labels=subjects, vert=False,
11      ↵ flierprops=outlier_props)
12    axes[i].set_title(f"Biểu đồ thể hiện mức độ phân tán điểm thi các môn vào
13      ↵ năm {year}", fontsize=16, color='darkblue')
14    axes[i].set_xlabel("Điểm số", fontsize=14)
15    axes[i].set_ylabel("Môn học", fontsize=14)
16    axes[i].tick_params(axis='x', labelsize=14)
17    axes[i].tick_params(axis='y', labelsize=14)
18
19 plt.tight_layout()
20 plt.savefig(f'figs/Mức độ phân tán (gộp) năm {year}.pdf')
21 plt.show()

```

3.2.1 Mức độ phân tán điểm thi các môn vào năm 2020

Các giá trị Min, Max được tính sau khi đã xác định được và loại bỏ các giá trị ngoại lai (outliers):

```
1 result = pd.DataFrame(columns=["Subject", "Min", "1st Quartile", "Median", "3rd
   ↵ Quartile", "IQR", "Max"])
2
3 for subject in subjects:
4     data_for_subject = df[df['Year'] == 2020][subject].dropna()
5     q1 = data_for_subject.quantile(0.25)
6     median = data_for_subject.median()
7     q3 = data_for_subject.quantile(0.75)
8     iqr = q3 - q1
9     min_value = data_for_subject[data_for_subject >= (q1 - 1.5 * iqr)].min()
10    max_value = data_for_subject[data_for_subject <= (q3 + 1.5 * iqr)].max()
11
12    result = result.append({"Subject": subject, "Min": min_value, "1st
      ↵ Quartile": q1, "Median": median, "3rd Quartile": q3, "IQR": iqr, "Max":
      ↵ max_value}, ignore_index=True)
13 print('{:^80}'.format('Mức độ phân tán của điểm thi các môn năm 2020'))
14 print()
15 result
```



Hình 9: Biểu đồ Box Plot các môn thi THPTQG năm 2020

Các chỉ số thống kê:

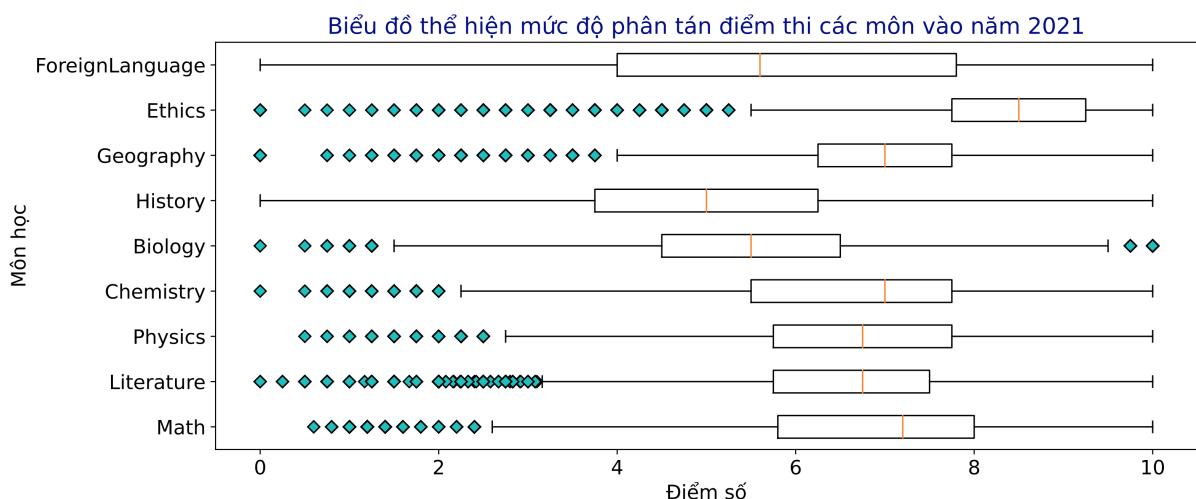
Mức độ phân tán của điểm thi các môn năm 2020							
	Subject	Min	1st Quartile	Median	3rd Quartile	IQR	Max
0	Math	2.40	5.80	7.20	8.20	2.40	10.00
1	Literature	3.75	6.00	6.75	7.50	1.50	9.75
2	Physics	2.75	5.75	7.00	7.75	2.00	10.00
3	Chemistry	1.75	5.50	7.00	8.00	2.50	10.00
4	Biology	2.25	4.75	5.50	6.50	1.75	9.00
5	History	0.75	4.00	5.00	6.25	2.25	9.50
6	Geography	4.00	6.25	7.00	7.75	1.50	10.00
7	Ethics	5.25	7.50	8.25	9.00	1.50	10.00
8	ForeignLanguage	0.00	3.20	4.20	5.60	2.40	9.20

Trong năm 2020, môn học có sự biến động lớn nhất là môn Hoá (Chemistry) với IQR khoảng 2.50 điểm, cho thấy rằng trong các môn học này có sự chênh lệch rất lớn giữa điểm số của các thí sinh.

Các môn học khác như Toán, Vật lý, Lịch sử, và Ngoại ngữ (Math, Physics, History, Foreign Language) cũng có độ phân tán tương đối lớn với IQR khoảng từ 2.00 - 2.40 điểm. Điều này cho thấy sự chênh lệch trong điểm số của học sinh trong các môn này cũng tương đối lớn.

Các môn còn lại như Ngữ văn, Sinh học, Địa lý, Giáo dục công dân (Literature, Biology, Geography, Ethics) có phạm vi điểm hẹp hơn với IQR khoảng từ 1.50 – 1.75 điểm, cho thấy sự đồng đều hơn trong điểm số của các thí sinh.

3.2.2 Mức độ phân tán điểm thi các môn vào năm 2021



Hình 10: Biểu đồ Box Plot các môn thi THPTQG năm 2021

Các chỉ số thống kê:

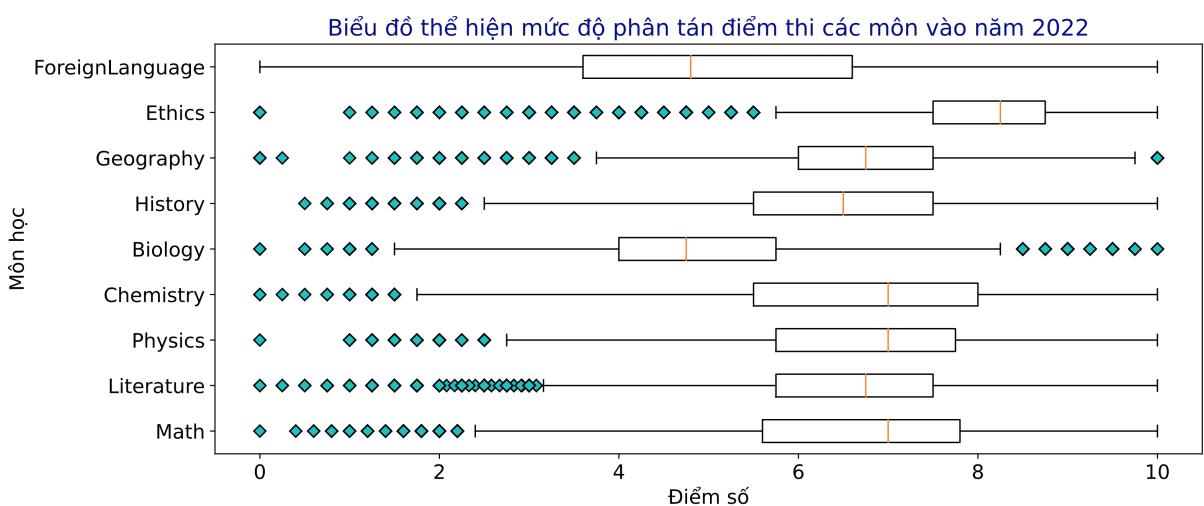
Mức độ phân tán của điểm thi các môn năm 2021								
	Subject	Min	1st Quartile	Median	3rd Quartile	IQR	Max	
0	Math	2.60	5.80	7.20	8.00	2.20	10.0	
1	Literature	3.16	5.75	6.75	7.50	1.75	10.0	
2	Physics	2.75	5.75	6.75	7.75	2.00	10.0	
3	Chemistry	2.25	5.50	7.00	7.75	2.25	10.0	
4	Biology	1.50	4.50	5.50	6.50	2.00	9.5	
5	History	0.00	3.75	5.00	6.25	2.50	10.0	
6	Geography	4.00	6.25	7.00	7.75	1.50	10.0	
7	Ethics	5.50	7.75	8.50	9.25	1.50	10.0	
8	ForeignLanguage	0.00	4.00	5.60	7.80	3.80	10.0	

Khác với năm 2020, trong năm 2021, môn học có sự biến động điểm số lớn nhất là môn Ngoại ngữ (Foreign Language) với IQR là 3.80 điểm, tiếp đến là môn Lịch sử (History) với khoảng IQR là 2.50 điểm, cho thấy mức độ phân tán của dữ liệu ở 2 môn học này là rất lớn.

Các môn có sự phân tán hay biến động dữ liệu tương đối lớn là Toán, Vật lý, Hoá học, Sinh học (Math, Physics, Chemistry, Biology) với khoảng IQR từ 2.00 – 2.25.

Các môn học còn lại có phạm vi phân tán hẹp với khoảng IQR từ 1.50 – 1.75, cho thấy sự đồng đều trong điểm số của các thí sinh.

3.2.3 Mức độ phân tán điểm thi các môn vào năm 2022



Hình 11: Biểu đồ Box Plot các môn thi THPTQG năm 2022

Các chỉ số thống kê:

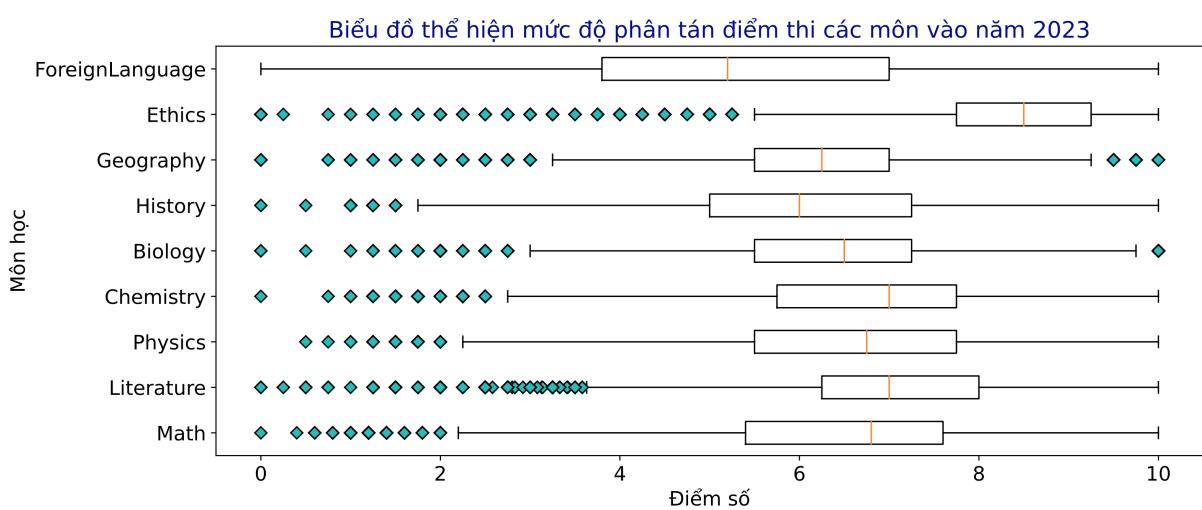
Mức độ phân tán của điểm thi các môn năm 2022								
	Subject	Min	1st Quartile	Median	3rd Quartile	IQR	Max	
0	Math	2.40	5.60	7.00	7.80	2.20	10.00	
1	Literature	3.16	5.75	6.75	7.50	1.75	10.00	
2	Physics	2.75	5.75	7.00	7.75	2.00	10.00	
3	Chemistry	1.75	5.50	7.00	8.00	2.50	10.00	
4	Biology	1.50	4.00	4.75	5.75	1.75	8.25	
5	History	2.50	5.50	6.50	7.50	2.00	10.00	
6	Geography	3.75	6.00	6.75	7.50	1.50	9.75	
7	Ethics	5.75	7.50	8.25	8.75	1.25	10.00	
8	ForeignLanguage	0.00	3.60	4.80	6.60	3.00	10.00	

Tương tự năm 2021, môn Ngoại ngữ (Foreign Language) là môn học có điểm số phân tán nhiều nhất với khoảng IQR là 3.00 điểm.

Các môn như Toán, Vật lý, Hóa học, Lịch sử có độ biến động dữ liệu tương đối cao với khoảng IQR từ 2.00 – 2.50 điểm.

Các môn còn lại có độ biến động dữ liệu nhỏ, phân tán hẹp, cho thấy sự đồng đều hơn trong điểm số của các thí sinh.

3.2.4 Mức độ phân tán điểm thi các môn vào năm 2023



Hình 12: Biểu đồ Box Plot các môn thi THPTQG năm 2023

Các chỉ số thống kê:

Mức độ phân tán của điểm thi các môn năm 2023							
	Subject	Min	1st Quartile	Median	3rd Quartile	IQR	Max
0	Math	2.20	5.40	6.80	7.60	2.20	10.00
1	Literature	3.63	6.25	7.00	8.00	1.75	10.00
2	Physics	2.25	5.50	6.75	7.75	2.25	10.00
3	Chemistry	2.75	5.75	7.00	7.75	2.00	10.00
4	Biology	3.00	5.50	6.50	7.25	1.75	9.75
5	History	1.75	5.00	6.00	7.25	2.25	10.00
6	Geography	3.25	5.50	6.25	7.00	1.50	9.25
7	Ethics	5.50	7.75	8.50	9.25	1.50	10.00
8	ForeignLanguage	0.00	3.80	5.20	7.00	3.20	10.00

Trong năm 2023, điểm số ở môn Ngoại ngữ vẫn giữ mức độ phân tán mạnh nhất, với khoảng IQR cao hơn năm 2022 là 0.20 điểm, cho thấy sự đa dạng và không đồng đều trong phân phối.

Các môn học có mức độ biến động tương đối cao là: Toán, Vật lý, Hoá học, Lịch sử. Điểm số ở các môn còn lại: Ngữ văn, Sinh học, Địa lý, Giáo dục công dân có mức độ phân tán hẹp.

Nhận xét chung:

- Trong giai đoạn 2021 – 2023, điểm số ở môn Ngoại ngữ bắt đầu có sự phân tán rộng, biến động mạnh mẽ, với khoảng IQR từ 3.00 – 3.80 điểm, thể hiện sự chênh lệch lớn giữa các giá trị dữ liệu.
- Trong giai đoạn 4 năm từ 2020 đến 2023, điểm số ở các môn: Toán, Vật lý, Hoá học, Lịch sử luôn có sự biến động lớn, cho thấy sự đa dạng trong dữ liệu và không có sự đồng đều trong phân phối.
- Ngược lại, điểm số các môn như Ngữ Văn, Địa lý, Sinh học, Giáo dục công dân lại có mức độ phân tán dữ liệu hẹp, ít biến động, phân bố tương đối đều trong khoảng giữa phần tư thứ 1 và phần tư thứ 3.

Nguyên nhân:

- Độ khó của từng môn học: Các môn học có độ khó khác nhau. Một số môn có nội dung khó hơn, yêu cầu kiến thức sâu rộng hơn, và do đó có thể dẫn đến sự biến động lớn hơn trong điểm số. Các môn học khó hơn thường có IQR cao hơn.
- Sự đa dạng trong chương trình học: Sự đa dạng trong chương trình học và giảng dạy

có thể ảnh hưởng đến sự biến động trong điểm số. Một số trường học hoặc khu vực có chương trình học đa dạng hơn.

- **Sự khác biệt cá nhân:** Mỗi học sinh có sự khác biệt về năng lực, hiệu suất học tập, sự chuẩn bị, và nỗ lực, dẫn đến sự biến động trong điểm số.

3.3 Kiểm tra các đại lượng về hình dáng phân phối

Tổng quát về các phương pháp đo lường và kiểm tra độ đối xứng trong hình dáng phân phối

Để kiểm tra độ nghiêng (*skewness*) hay độ đối xứng trong hình dáng phân phối, ta có thể sử dụng các chỉ số như skewness hay kurtosis. Trong đó, độ đo skewness được định nghĩa như sau:

$$S_k = \frac{1}{n} \sum_{i=1}^n \frac{(x_i - \bar{x})^3}{s^3}$$

với n là số điểm dữ liệu, s là độ lệch chuẩn, và \bar{x} là trung bình các điểm dữ liệu.

Độ nhọn *Kurtosis* thể hiện độ cao phần “trung tâm” hay độ “dày” phần đuôi của phân phối và có công thức như sau:

$$\text{kurtosis} = \frac{1}{|X|} \sum_{x_i \in X} \frac{(x_i - \bar{x})^4}{s^4}$$

Trong đó, độ đo kurtosis của các phân phối thường được so sánh với phân phối chuẩn (có chỉ số kurtosis = 3). Ngoài ra, ta có thể so sánh các giá trị mean và median để kiểm tra độ nghiêng của hình dáng phân phối, trong đó nếu:

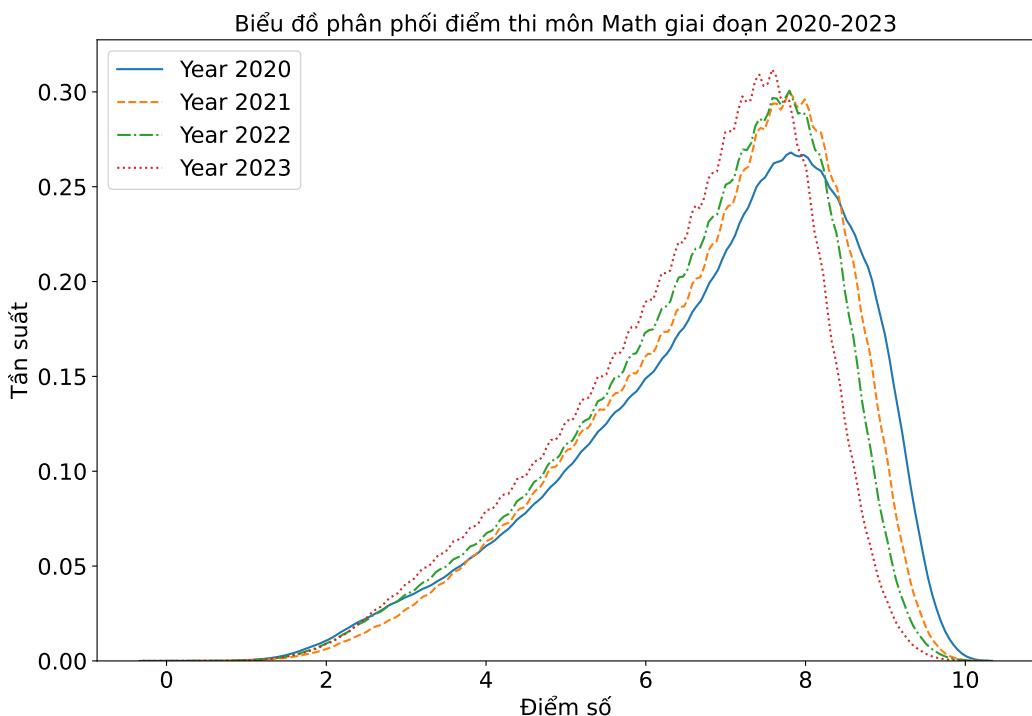
- Mean > median (positive skewness): hình dáng phân phối sẽ có đuôi bên phải dài hơn đuôi bên trái; khi đó các giá trị lớn (có thể là các outliers) đẩy giá trị mean về phía cuối.
- Mean < median (negative skewness): hình dáng phân phối sẽ có đuôi bên trái dài hơn đuôi bên phải; khi này các giá trị nhỏ (outliers) sẽ đẩy mean về phía đầu.

3.3.1 Phân phối điểm thi các môn bắt buộc

```
1  linestyles = [':', '--', '-.', ':']
2
3  grouped_data = df.groupby('Year')
4
5  subjects = [
6      'Math', 'Literature', 'Physics', 'Chemistry',
7      'Biology', 'History', 'Geography',
8      'Ethics', 'ForeignLanguage'
9  ]
10
11 for subject in subjects:
12     plt.figure(figsize=(12, 8))
13
14     for i, (year, group) in enumerate(grouped_data):
15         sns.kdeplot(group[subject], fill=False, label=f'Year {year}',
16                     linestyle=linestyles[i])
17
18     plt.title(f"Biểu đồ phân phối điểm thi môn {subject} giai đoạn 2020-2023",
19               fontsize=16)
20     plt.xlabel("Điểm số", fontsize=16)
21     plt.ylabel("Tần suất", fontsize=16)
22     plt.legend(fontsize=16)
23     plt.xticks(fontsize=16)
24     plt.yticks(fontsize=16)
25     plt.savefig(f'figs/Phân phối môn {subject}.pdf')
```

Môn Toán

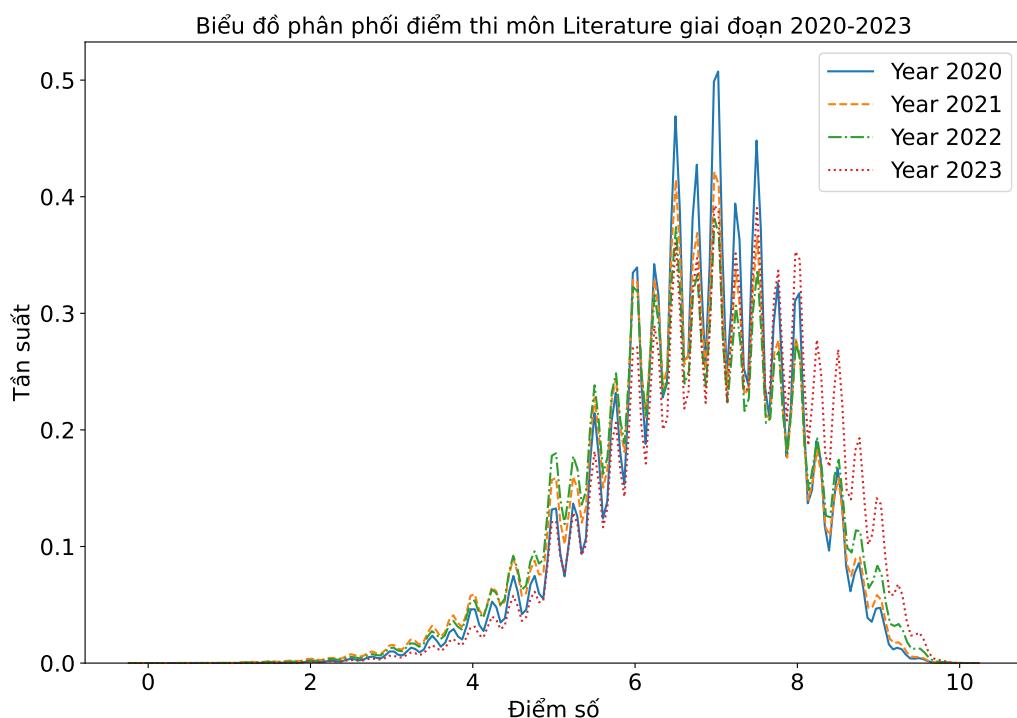
Có thể thấy phân phối của môn Toán qua các năm là khá gần nhau, với mode dao động trong khoảng điểm từ 7 đến 8. Năm 2023, mode của môn Toán là thấp nhất, mode điểm thi Toán của các năm 2020, 2021, và 2022 khá gần nhau. Điểm trung bình của môn Toán dao động trong khoảng 6.45 đến 6.87, trong đó cao nhất là 6.87 vào kì thi năm 2020, và thấp nhất là năm 2023 với điểm trung bình 6.45. Phổ điểm lệch trái do trong cả bốn năm vì điểm thi trung bình (6.45-6.87) nhỏ hơn mode (7.6 - 7.8).



Hình 13: Biểu đồ phân phối xác suất môn Toán giai đoạn 2020-2023

Môn Ngữ Văn

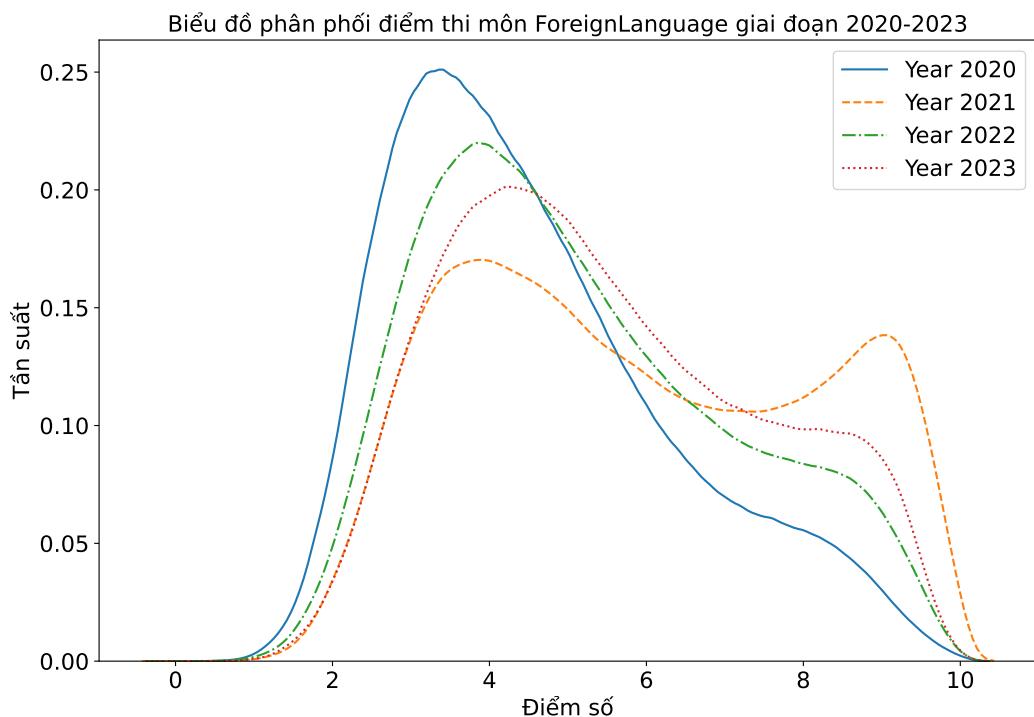
Phổ điểm thi môn Ngữ Văn dao động khá mạnh, tuy nhiên phân phối điểm qua từng năm có xu hướng giống nhau và không chênh lệch quá nhiều. Thí sinh chủ yếu đạt điểm từ 6-7 đối với môn Ngữ Văn. Cụ thể, điểm trung bình đối với môn Ngữ Văn qua các năm lần lượt là 6.73; 6.61; 6.65; và 7.00. Trong khi đó, điểm có tần suất xuất hiện nhiều nhất là 7.00 qua bốn năm.



Hình 14: Biểu đồ phân phối xác suất môn Ngữ Văn giai đoạn 2020-2023

Môn Tiếng Anh

Phổ điểm thi môn Tiếng Anh của ba năm 2020, 2022, và 2023 có hình dáng khá giống nhau, trong đó phân phối điểm thi năm 2020 có xu hướng lệch trái so với 2 năm 2022 và 2023. Phổ điểm môn tiếng Anh năm 2021 được nhận định là “kì lạ” bởi sự xuất hiện của hai đỉnh.

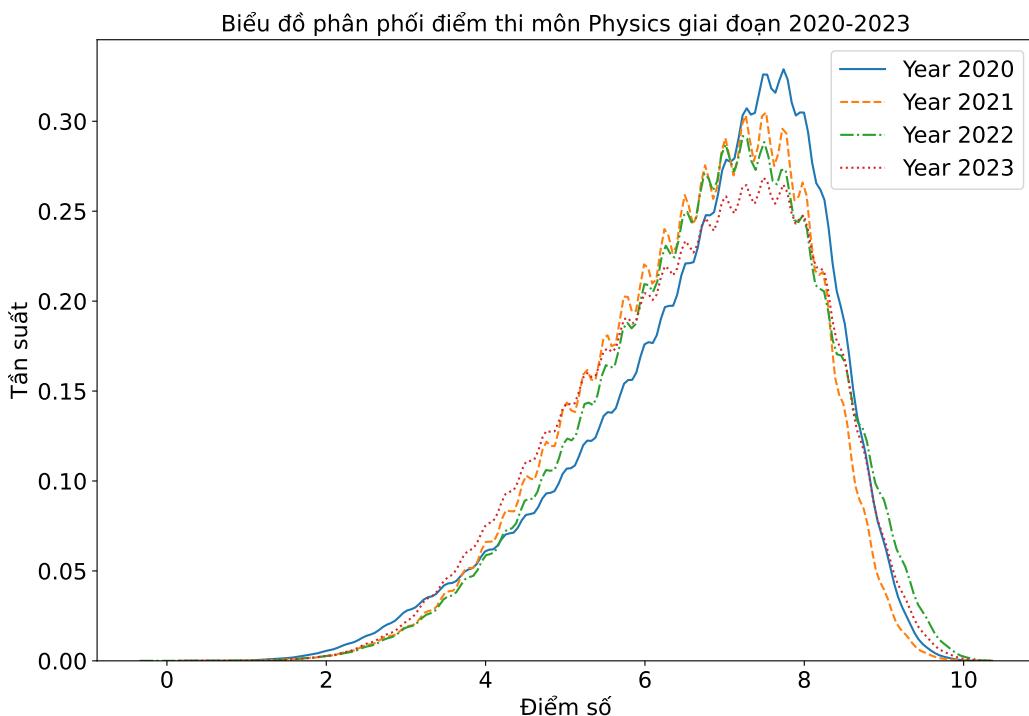


Hình 15: Biểu đồ phân phối xác suất môn Tiếng Anh giai đoạn 2020-2023

Theo phân tích của nhiều chuyên gia, nguyên nhân khiến phổ điểm môn tiếng Anh có 2 đỉnh là do phân hóa về điều kiện dạy học (như cơ sở vật chất, chất lượng giáo viên, sự quan tâm đầu tư của phụ huynh và học sinh đối với môn học này). So với kết quả những năm trước, GS Nguyễn Đình Đức (ĐH Quốc gia Hà Nội) cho biết phổ điểm tiếng Anh năm 2021 đã có sự điều chỉnh theo hướng tích cực hơn. Đỉnh bên trái của phổ (khoảng 4 - 5 điểm) cao hơn đỉnh của năm 2020 (3 - 3.8 điểm). Trong khi đó, phổ điểm môn tiếng Anh vào năm 2022 và 2023 là khá giống nhau, đỉnh bên trái dao động vào khoảng 4-4.2 điểm.

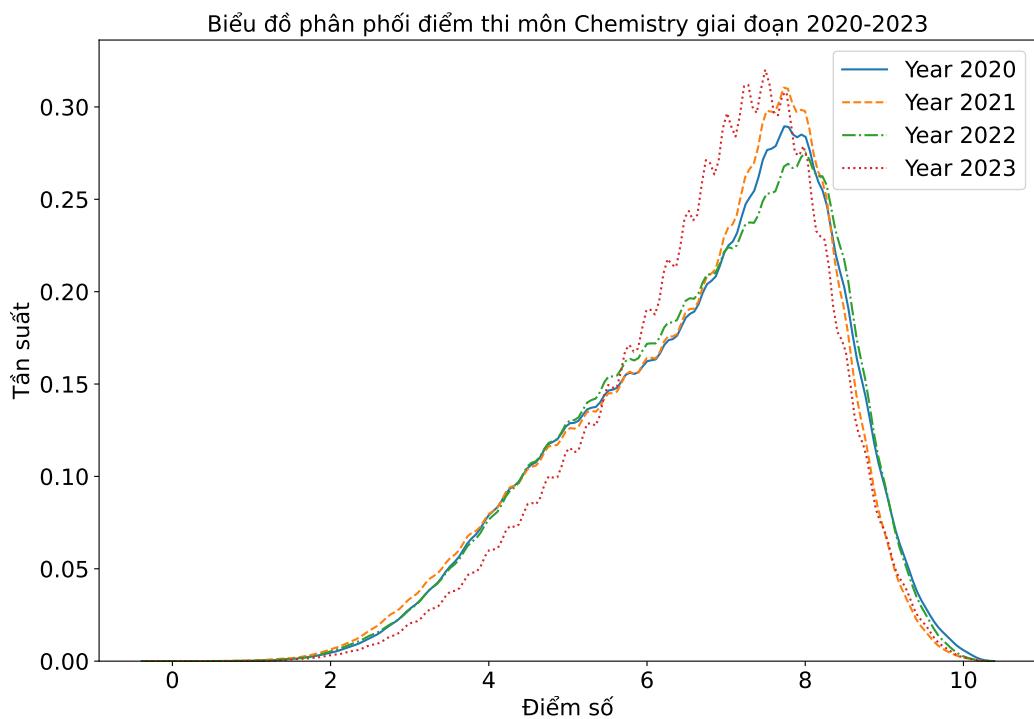
3.3.2 Phân phối điểm thi các môn thuộc tổ hợp KHTN

Đối với môn Vật Lý, không có sự chênh lệch nhiều giữa các năm trong giai đoạn từ 2020 đến 2023. Nhìn chung, năm 2020 là năm đỉnh đạt cao nhất, với điểm mode rơi vào khoảng 7.75 và điểm trung bình là 6.74 . Theo phân tích và nhận định, đề thi Vật Lý năm 2020 nhẹ nhàng, dễ đạt điểm trên trung bình nhưng phân hóa khá rõ đối với học sinh khá, giỏi. 2021 là năm điểm thi trung bình môn Vật Lý thấp nhất với 6.57 điểm và điểm mode là 7.25.



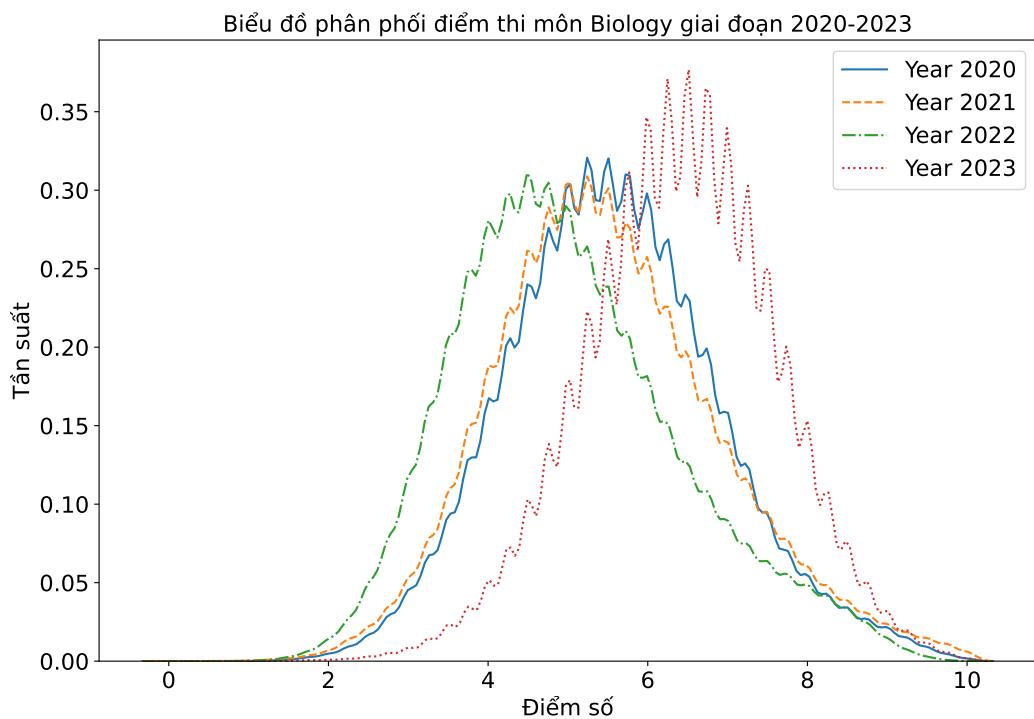
Hình 16: Biểu đồ phân phối xác suất môn Vật Lý giai đoạn 2020-2023

Tương tự với các môn học khác, phổ điểm thi môn Hoá từ năm 2020 đến năm 2023 là khá giống nhau và có xu hướng lệch phải, điểm trung bình dao động trong khoảng từ 7.8 đến 8. Cụ thể, điểm trung bình của từng năm lần lượt là 6.70; 6.62; 6.68; và 6.74 Điểm thi mode trong năm 2023 là 7.50, thấp nhất trong các năm. Năm 2023, đỉnh của phân phối có sự dịch chuyển sang trái và thấp hơn các năm trước. Theo [Báo Tuổi Trẻ](#), điều này là do đề thi có câu từ chưa chặt chẽ, đưa thêm thông tin bổ sung chưa chính xác hoặc thừa, và cách ‘thiết kế’ bài toán các bài toán hóa thiếu logic.



Hình 17: Biểu đồ phân phối xác suất môn Hoá học giai đoạn 2020-2023

Trái lại với hai môn khác trong tổ hợp KHTN là Vật Lý và Hoá học , phổ điểm thi môn Sinh học giai đoạn 2020-2023 có khá nhiều biến động. Dù phân phối điểm thi vào năm 2020 và 2021 là khá giống nhau, đỉnh dao động trong khoảng 4.5-6.50. Đáng chú ý là vào năm 2022, khi điểm mode rơi vào khoảng 4.50, thấp nhất trong tất cả các năm. Trong khi đó, vào năm 2023, phổ điểm thi môn Sinh tăng mạnh với dao động đỉnh nằm trong khoảng 7 điểm.

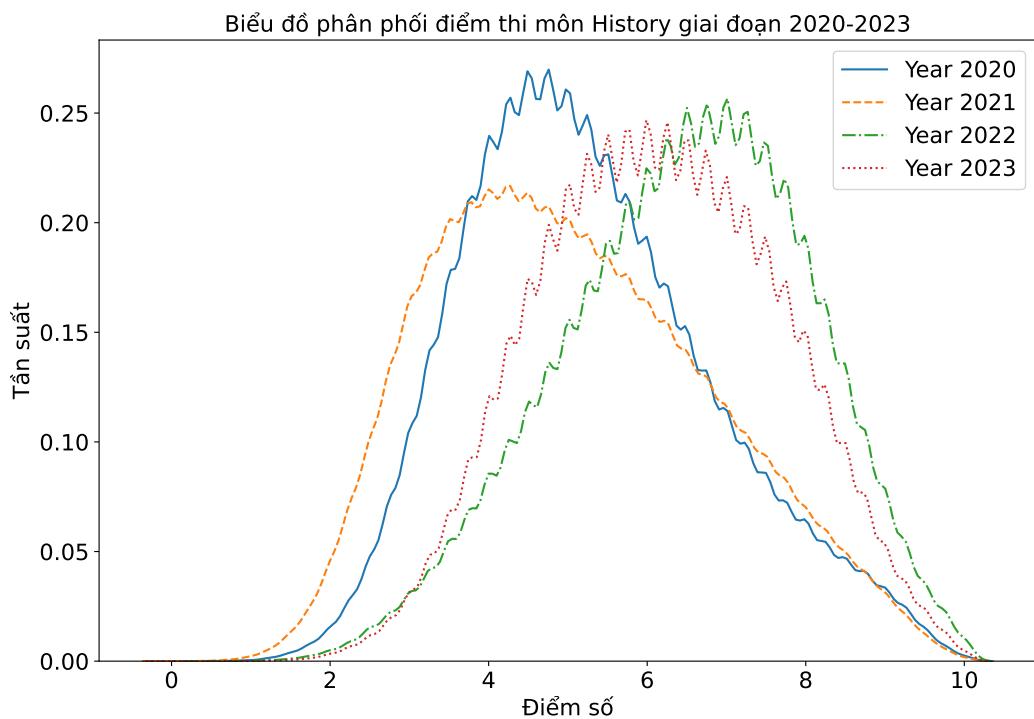


Hình 18: Biểu đồ phân phối xác suất môn Sinh học giai đoạn 2020-2023

3.3.3 Phân phối điểm thi các môn thuộc tổ hợp KHXH

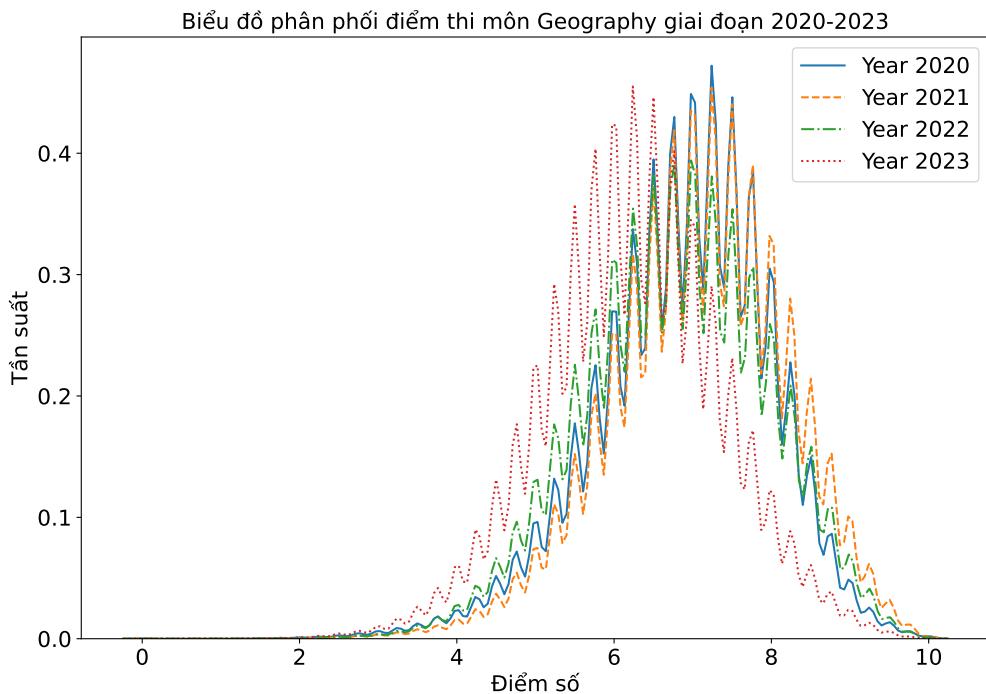
Môn Lịch Sử

Phổ điểm thi môn Lịch sử qua bốn năm có sự khác nhau khá rõ rệt. Trong đó, phân phối có đỉnh thấp nhất là vào năm 2021, rơi vào mức 4.25 điểm. Năm 2022, phổ điểm lệch hẳn sang phải so với năm 2020 và 2021. Trong khi đó, vào năm 2023, đỉnh dao động ở mức 6 điểm.



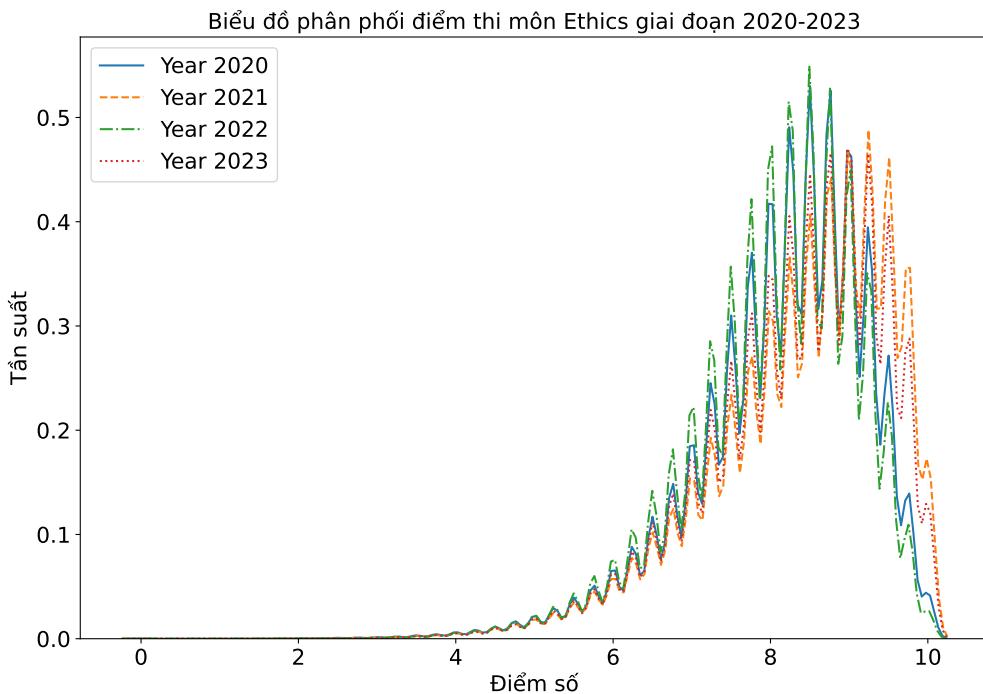
Hình 19: Biểu đồ phân phối xác suất môn Lịch sử giai đoạn 2020-2023

Phổ điểm môn Địa lý các năm 2020, 2021, và 2022 tương đối giống nhau về hình dáng phân phối lân đinh, dao động ở mức từ 6-7. Trong khi đó, đỉnh của phổ điểm thi Địa Lý vào năm 2023 lệch sang trái so với các năm trước, với mức đỉnh nằm ở khoảng 6. Theo báo [Quân đội nhân dân](#), một số thí sinh cho rằng đề Địa lý thuộc bài thi tổ hợp Khoa học Xã hội kỳ thi tốt nghiệp THPT 2023 là “khó và lạ”, trong đó những nội dung khó nhất bài thi nằm ở câu biểu đồ.



Hình 20: Biểu đồ phân phối xác suất môn Địa lý giai đoạn 2020-2023

Điểm thi môn GD&CD ở cả 4 năm đều có xu hướng lệch phải, với đỉnh dao động trong khoảng 8.6-8.8. Điểm trung bình môn GD&CD qua tất cả các năm trong giai đoạn 2020-2023 đều trên mức 8. Cụ thể, điểm trung bình lần lượt cho các năm là 8.13; 8.38; 8.03; và 8.29. Điểm mode ở năm 2021 là cao nhất với 9.25 điểm, thấp nhất là vào năm 2022 với 8.50 điểm.



Hình 21: Biểu đồ phân phối xác suất môn GDCD giai đoạn 2020-2023

3.4 Kiểm tra đại lượng về sự tương quan

Hai đại lượng phổ biến được sử dụng để đo lường mức độ tương quan là Hiệp phương sai (Covariance) và Hệ số tương quan Pearson (Pearson Correlation).

Hiệp phương sai (Covariance) của một tập hợp là giá trị trung bình của các tích số sai lệch của mỗi lần quan sát. Tuy nhiên, giá trị số của hiệp phương sai không được chuẩn hóa, nên rất khó giải thích cường độ của mối quan hệ, vì nó phụ thuộc vào các đơn vị đo lường của các biến.

Hệ số tương quan Pearson của 2 biến ngẫu nhiên X, Y được tính bằng cách chia hiệp phương sai cho tích của độ lệch chuẩn. Độ lệch chuẩn đo lường độ biến thiên tuyệt đối của tập dữ liệu, do đó khi chia các giá trị hiệp phương sai cho độ lệch chuẩn, nó sẽ chia tỷ lệ giá trị xuống một phạm vi giới hạn từ -1 đến $+1$. Khi hệ số tương quan Pearson của 2 biến X, Y lớn hơn 0 , thì 2 biến X và Y có quan hệ tuyến tính thuận với nhau. Khi hệ số tương quan Pearson của 2 biến X, Y nhỏ hơn 0 , thì 2 biến X và Y có quan hệ tuyến tính nghịch với nhau. Khi hệ số tương quan Pearson của 2 biến X, Y bằng 0 , thì 2 biến X và Y không có mối quan hệ với nhau hoặc có mối quan hệ rõ ràng giữa các biến nhưng mối quan hệ này không tuyến tính và cũng không có tương quan.

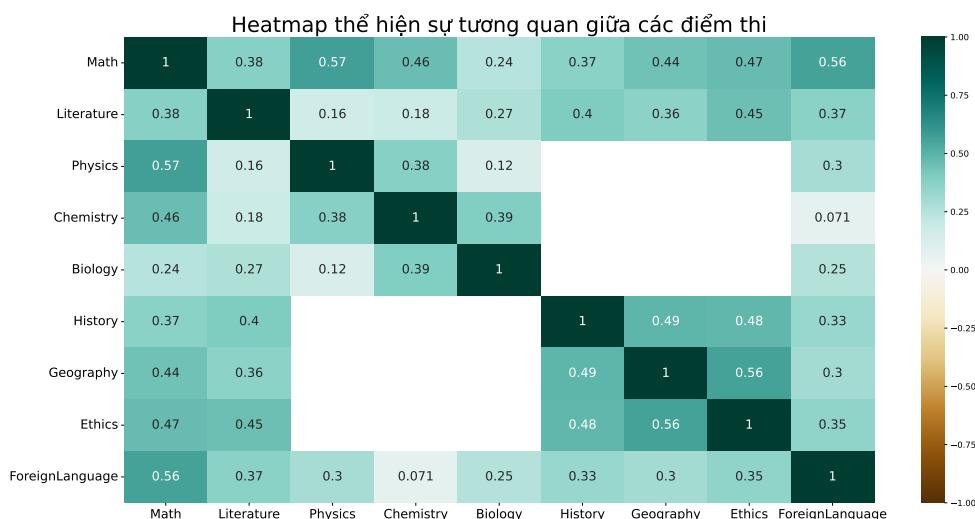
Để xem mối tương quan giữa từng cặp điểm thi với nhau, ta dùng ma trận hiệp phương

sai, mức độ tương quan của mỗi cặp được đo lường bằng hệ số tương quan Pearson để giá trị được chuẩn hóa về đoạn [-1, 1]. Và biểu diễn ma trận này dưới dạng heatmap.

```

1 subjects = ['Math', 'Literature', 'Physics', 'Chemistry', 'Biology', 'History',
2   ↪ 'Geography', 'Ethics', 'ForeignLanguage']
3 annot_kws = {"size": 15} #chỉnh fontsize của các hệ số
4 sbn.heatmap(df[subjects].corr(method='pearson'), vmin=-1, vmax=1, annot=True,
5   ↪ cmap='BrBG', annot_kws=annot_kws)
6 plt.xticks(fontsize=15)
7 plt.yticks(rotation=0, fontsize=15)
8 plt.title('Heatmap thể hiện sự tương quan giữa các điểm thi', fontsize=25)
9 plt.savefig("figs/correlation_heatmap.pdf")
10 plt.show()

```



Hình 22: Heatmap thể hiện sự tương quan giữa các điểm thi

Heatmap bị trống các hệ số tương quan giữa điểm của các môn thuộc tổ hợp KHTN với điểm của các môn thuộc tổ hợp KHXH vì các thí sinh chỉ chọn 1 trong 2 tổ hợp để thi, vậy nên mỗi thí sinh chỉ có điểm thi của 1 trong 2 tổ hợp, điểm thi của tổ hợp còn lại sẽ là các giá trị NaN.

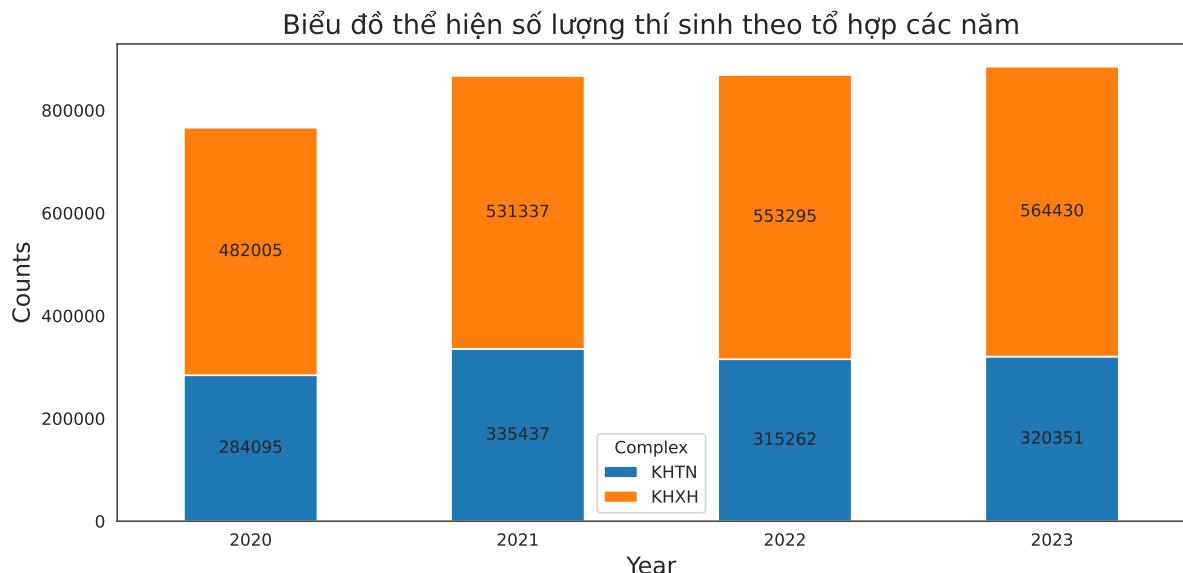
Ta có thể thấy, tất cả các cặp điểm đều tuyến tính thuận, tuy nhiên sự phụ thuộc lại không cao, cao nhất là điểm Toán và điểm Lý (0.57); điểm Hóa và điểm Ngoại ngữ gần như không có mối quan hệ tuyến tính với nhau (0.071).

4 CHƯƠNG 4: BIỂU ĐỒ DIỄN TRỰC QUAN ĐỨC LIỆU

4.1 Biểu đồ thể hiện số lượng thí sinh theo tổ hợp

```
1 df_gb_complex_year = df.groupby(['Complex', 'Year']).size().reset_index(name =
2     ↪ 'Counts')
3 df_pivot = df_gb_complex_year.pivot(index='Year', columns='Complex',
4     ↪ values='Counts')
5
6 ax = df_pivot.plot(kind='bar', stacked=True, figsize=(10,5))
7
8 plt.xlabel('Year', fontsize=14)
9 plt.ylabel('Counts', fontsize=14)
10 plt.title('Biểu đồ thể hiện số lượng thí sinh theo tổ hợp các năm', fontsize=16)
11 plt.ticklabel_format(style='plain', axis = 'y')
12 plt.xticks(rotation=0)
13
14 for p in ax.patches:
15     width, height = p.get_width(), p.get_height()
16     x, y = p.get_xy()
17     ax.text(x+width/2,
18             y+height/2,
19             '{:.0f}'.format(height),
20             horizontalalignment='center',
21             verticalalignment='center')
22
23 plt.tight_layout()
24 plt.savefig('figs/Biểu đồ thể hiện số lượng thí sinh theo tổ hợp các năm.pdf')
25 plt.show()
```

Số lượng thí sinh đăng ký dự thi bài thi tổ hợp KHTN luôn thấp hơn số lượng thí của tổ hợp KHXH. Việc này cho thấy xu hướng chọn giải pháp an toàn của thí sinh trong việc xét công nhận tốt nghiệp THPT.



Hình 23: Biểu đồ thể hiện thí sinh theo tổ hợp các năm

Hiện tượng này có thể được lý giải bằng tính chất của đề thi các môn tổ hợp KHTN và KHXH. Trong khi môn Địa lý được sử dụng Atlat địa lý trong phòng thi, và 2 môn còn lại của tổ hợp KHXH có thể suy luận tình huống để suy đoán đáp án; thì để vượt qua bài thi các môn KHTN thí sinh cần tư duy logic và hệ thống kiến thức dàn trải. Bên cạnh đó, đối với các thí sinh có nguyện vọng xét tuyển đại học bằng tổ hợp D01, việc lựa chọn bài thi tổ hợp KHXH giúp họ có thể tập trung vào 3 môn chính để đạt được điểm xét tuyển cao, trong khi điểm bài thi tổ hợp chỉ cần đủ tiêu chuẩn xét tốt nghiệp.

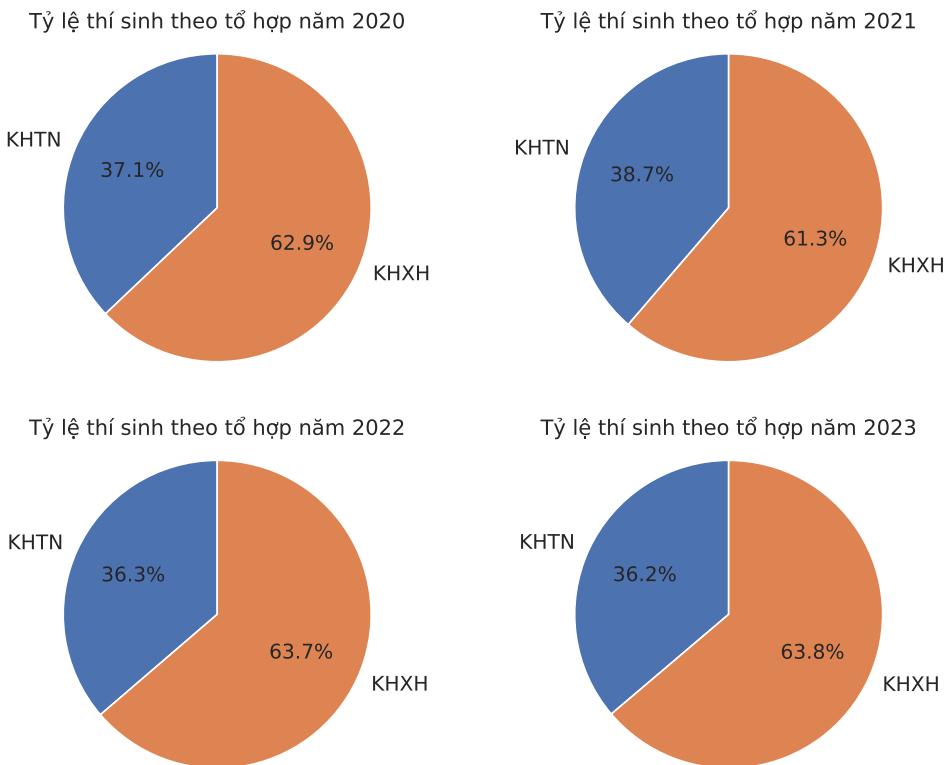
Tuy nhiên, việc bài thi KHXH được lựa chọn nhiều hơn cũng dẫn đến một bất lợi đối với các thí sinh đăng ký xét tuyển bằng tổ hợp có các môn Sử, Địa, GD&CD. Điểm chuẩn của các ngành xét tuyển tổ hợp KHXH nhiều khả năng sẽ tăng dần và tỷ lệ cạnh tranh cũng gay gắt hơn. Để đạt được điểm cao trong bài thi tổ hợp KHXH cũng cần nhiều nỗ lực và kiến thức tổng quát. Thực tế độ khó của bài thi qua các năm vẫn được đánh giá là có tính phân hóa cao, hoàn toàn không dễ để đạt được điểm 9, 10 ở môn Lịch sử, Địa lý.

Vậy đối với các thí sinh chỉ ưu tiên mục tiêu xét tốt nghiệp hoặc định hướng xét tuyển bằng khối D01, tổ hợp KHXH có thể là một lựa chọn thực tế. Đối với mục tiêu xét tuyển, thí sinh nên lựa chọn theo năng lực học tập và định hướng nghề nghiệp của bản thân. Bất kỳ lựa chọn nào cũng đòi hỏi thí sinh phải có chiến lược học tập hiệu quả để đạt được điểm cao trong bài thi tổ hợp.

4.2 Biểu đồ thể hiện tỷ lệ thí sinh theo tổ hợp

Qua biểu đồ thể hiện số lượng thí sinh theo tổ hợp, ta có thể thấy, cả 4 năm đều có số lượng thí sinh thi tổ hợp KHXH nhiều hơn số lượng thí sinh thi tổ hợp KHTN. Cụ thể hơn, để biết được tỷ lệ giữa số lượng thí sinh thi tổ hợp KHXH với số lượng thí sinh thi tổ hợp KHTN, ta sẽ biểu diễn tỷ lệ này dưới dạng biểu đồ tròn.

```
1 grouped = df.groupby(['Year', 'Complex']).size().unstack()
2
3 years = df['Year'].unique()
4 plt.figure(figsize=(10, 8))
5 for i, year in enumerate(years, 1):
6     plt.subplot(2, 2, i)
7     count_by_complex = grouped.loc[year].dropna()
8     labels = count_by_complex.index
9     sizes = count_by_complex.values
10    plt.subplot(2, 2, i)
11    plt.pie(sizes, labels=labels, autopct='%.1f%%', startangle=90,
12             textprops={'fontsize': 12})
13    plt.axis('equal')
14    plt.title(f'Tỷ lệ thí sinh theo tổ hợp năm {year}', fontsize=12.5)
15
16 plt.savefig("figs/Biểu đồ thể hiện tỷ lệ thí sinh theo tổ hợp qua từng
→ năm.pdf")
plt.show()
```



Hình 24: Biểu đồ thể hiện tỷ lệ số lượng thí sinh theo tổ hợp qua các năm

Qua 4 năm, tỷ lệ thí sinh thi KHXH (hơn 60%) áp đảo so với tỷ lệ thí sinh thi KHTN (chỉ hơn 30%), thậm chí sự chênh lệch này còn có xu hướng tăng dần khi tỷ lệ thí sinh thi KHTN năm 2021 là 38.7% giảm còn 36.3% (năm 2022) và 36.2% (năm 2023).

Mặc dù các thí sinh chọn bài thi tổ hợp KHXH để tránh điểm thấp, giảm áp lực ôn tập để dành thời gian cho các môn xét tuyển đại học là lựa chọn sáng suốt nhưng một số chuyên gia cũng khuyên cáo điều này có thể dẫn đến một số hệ lụy trong việc đào tạo nguồn nhân lực.

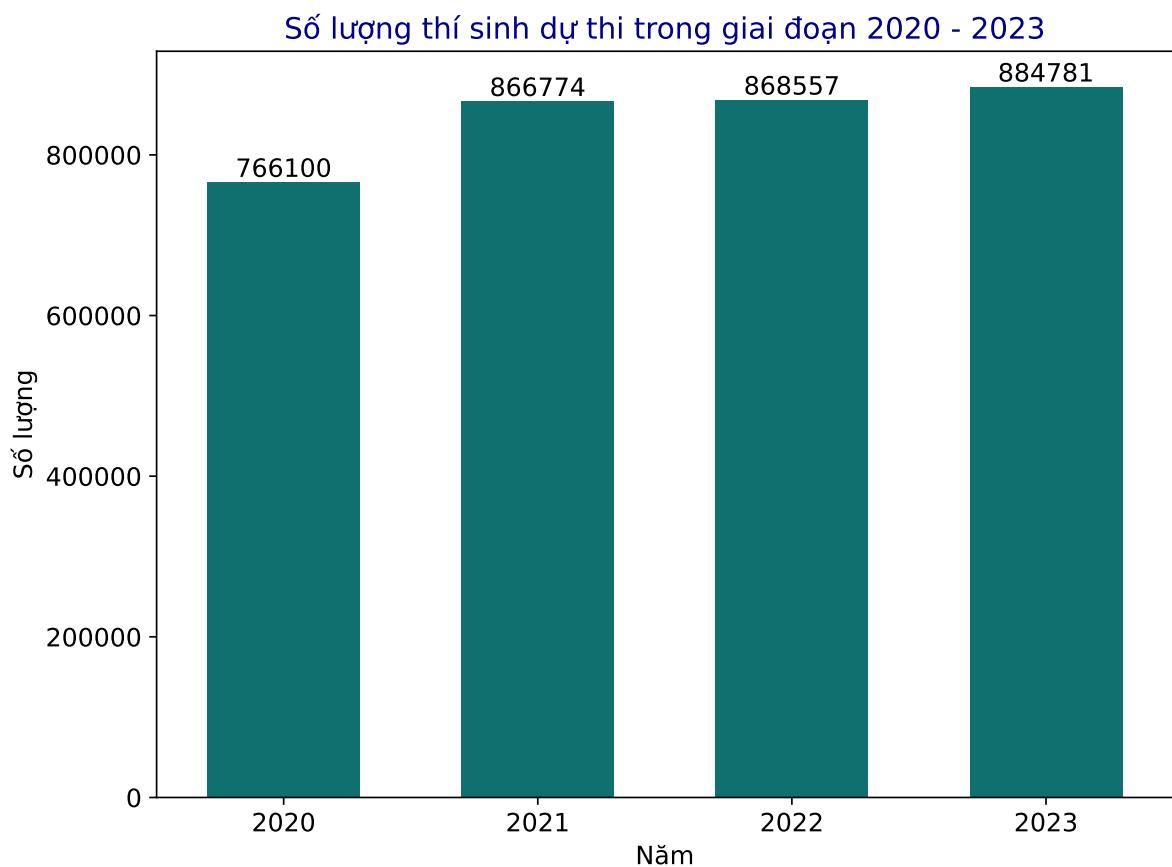
Với nền kinh tế ngày càng đòi hỏi nhiều lao động liên quan đến khoa học-công nghệ, đặc biệt là các ngành nghề liên quan đến STEM (khoa học, công nghệ, kỹ thuật và toán). Nếu thí sinh chọn thi tổ hợp KHTN ngày càng giảm, không chỉ ảnh hưởng đến nguồn nhân lực của đất nước, mà còn ảnh hưởng đến đời sống và thu nhập của các thí sinh sau này. Xu hướng thế giới cho thấy các ngành nghề STEM ngày càng dễ có việc làm, có thu nhập cao, trong khi các ngành xã hội, cơ hội việc làm hạn chế hơn.

Nhóm đề xuất giải pháp để cải thiện tình trạng này là cần cân bằng độ khó đề thi giữa tổ

hợp KHXH và KHTN, nếu không, có thể dẫn tới nguy cơ thí sinh thiên về chọn tổ hợp KHXH để an toàn trong việc xét tốt nghiệp THPT mà không phải chọn tổ hợp thi theo xu hướng nghề nghiệp và đam mê của bản thân.

4.3 Số lượng thí sinh dự thi trong giai đoạn 2020 – 2023

```
1 plt.figure(figsize=(10, 6))
2 counts = df['Year'].value_counts().sort_index()
3 ax = sns.barplot(x=counts.index, y=counts.values, palette='teal', width=0.5)
4
5 for p in ax.patches:
6     ax.annotate(f"{int(p.get_height())}", (p.get_x() + p.get_width() / 2,
7         p.get_height()), ha='center', va='bottom')
8
9 plt.title('Số lượng thí sinh dự thi trong giai đoạn 2020 – 2023', fontsize=16,
10           color='darkblue')
11 plt.xlabel('Năm', fontsize=14)
12 plt.ylabel('Số lượng thí sinh', fontsize=14)
13 plt.xticks(fontsize=14)
14 plt.yticks(fontsize=14)
15 plt.savefig(f'figs/Số lượng thí sinh dự thi trong giai đoạn 2020 – 2023.pdf')
16 plt.show()
```



Hình 25: Số lượng thí sinh dự thi trong giai đoạn 2020 – 2023

Số lượng thí sinh dự thi có xu hướng tăng dần từ năm 2020 đến năm 2023. Theo số liệu thống kê, có tổng cộng 766,100 thí sinh dự thi vào năm 2020.

Có 866,774 thí sinh dự thi tốt nghiệp THPT năm 2021, tăng hơn 100,674 thí sinh so với năm trước. Số lượng thí sinh tăng mạnh ở năm này có lý do quan trọng là tình hình dịch Covid-19 diễn biến phức tạp, nhiều thí sinh không thể du học và nhiều em học phổ thông, học đại học ở nước ngoài cũng đã trở về Việt Nam để học tập.

Năm 2022 và năm 2023 có số lượng thí sinh dự thi lần lượt là 868,557 và 884,781 thí sinh. Nhận thấy rằng số lượng thí sinh dự thi tốt nghiệp THPT quốc gia năm 2023 tăng đáng kể so với năm 2022, khoảng 16,000 thí sinh.

4.4 Biểu đồ thống kê top 10 tỉnh/ thành phố có số lượng thí sinh dự thi cao nhất từ năm 2020 đến năm 2023

```

1 years = [2020, 2021, 2022, 2023]
2 for year in years:
3     df_year = df[df['Year'] == year]
4     top_10_provinces = df_year['Province'].value_counts().nlargest(10)

```

```

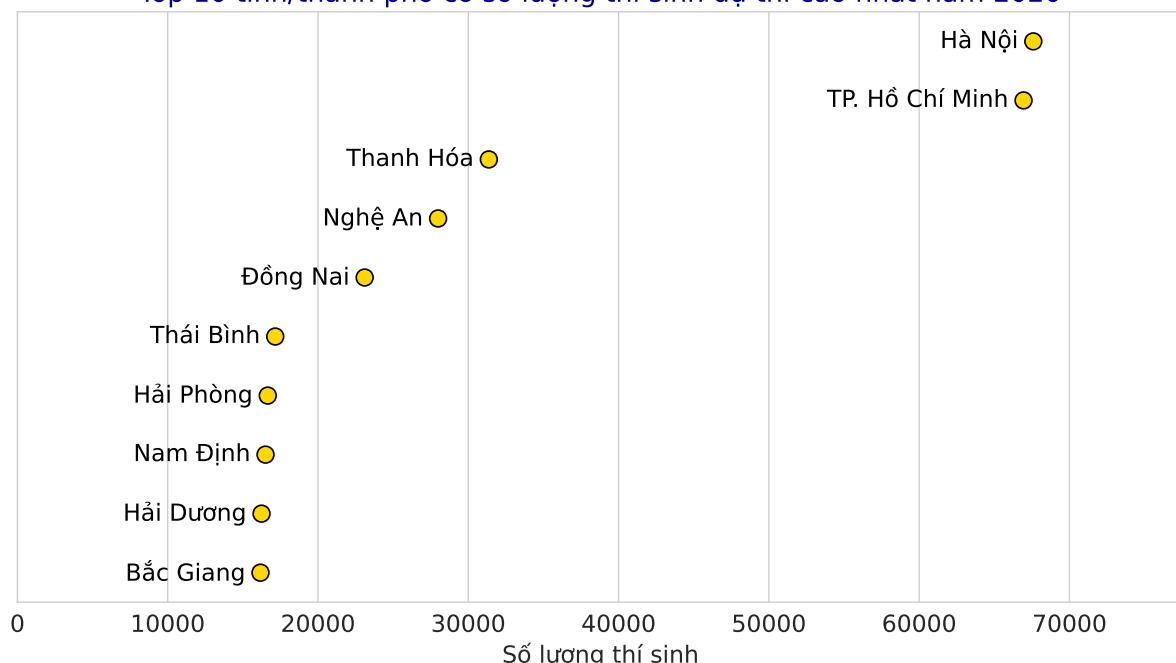
5
6     plt.figure(figsize=(10, 6))
7     ax = sns.stripplot(x=top_10_provinces.values, y=top_10_provinces.index,
8                          color='gold', size=10, edgecolor='black', linewidth=1)
9
10    plt.xlabel('Số lượng thí sinh', fontsize=14)
11    plt.title(f'Top 10 tỉnh/thành phố có số lượng thí sinh dự thi cao nhất năm
12          {year}', fontsize=16, color='darkblue')
13    plt.xticks(fontsize=14)
14
15    for i, (value, province) in enumerate(zip(top_10_provinces.values,
16                                              top_10_provinces.index)):
17        ax.text(value - 1000, i, f'{province}', color='black', va='center',
18                fontsize=14, ha='right')
19
20    ax.set_yticks([])
21    plt.xlim(0, max(top_10_provinces.values) + 10000)
22    plt.tight_layout()
23    plt.savefig(f'figs/Dot_Plot_So_luong_thi_sinh_theo_thanh_pho_nam_{year}.pdf')
24    plt.show()

```

Trong đó, Hà Nội là địa phương dẫn đầu cả nước về số lượng thí sinh dự thi tốt nghiệp THPT quốc gia từ năm 2020 đến 2023 với số liệu thống kê nổi bật. Kể từ năm 2021, số lượng thí sinh dự thi của Hà Nội luôn vượt mức hơn 80.000 thí sinh.

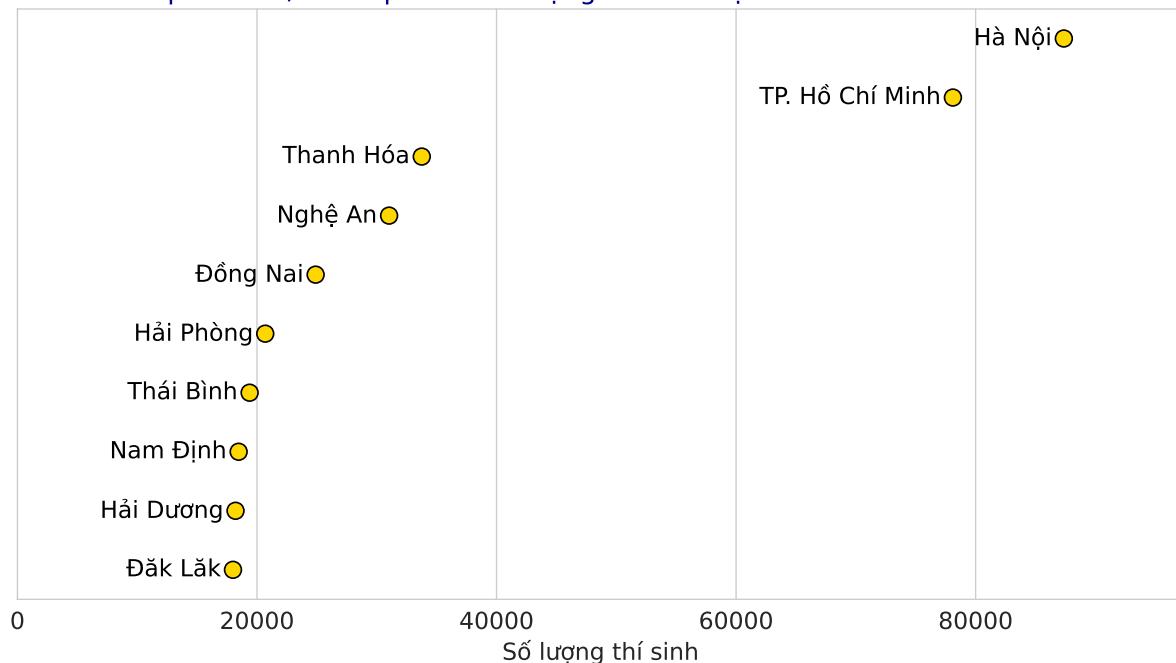
Xếp thứ 2 là Thành phố Hồ Chí Minh với số lượng thí sinh nhiều đáng kể, nhưng vẫn ở mức thấp hơn so với Hà Nội.

Top 10 tỉnh/thành phố có số lượng thí sinh dự thi cao nhất năm 2020



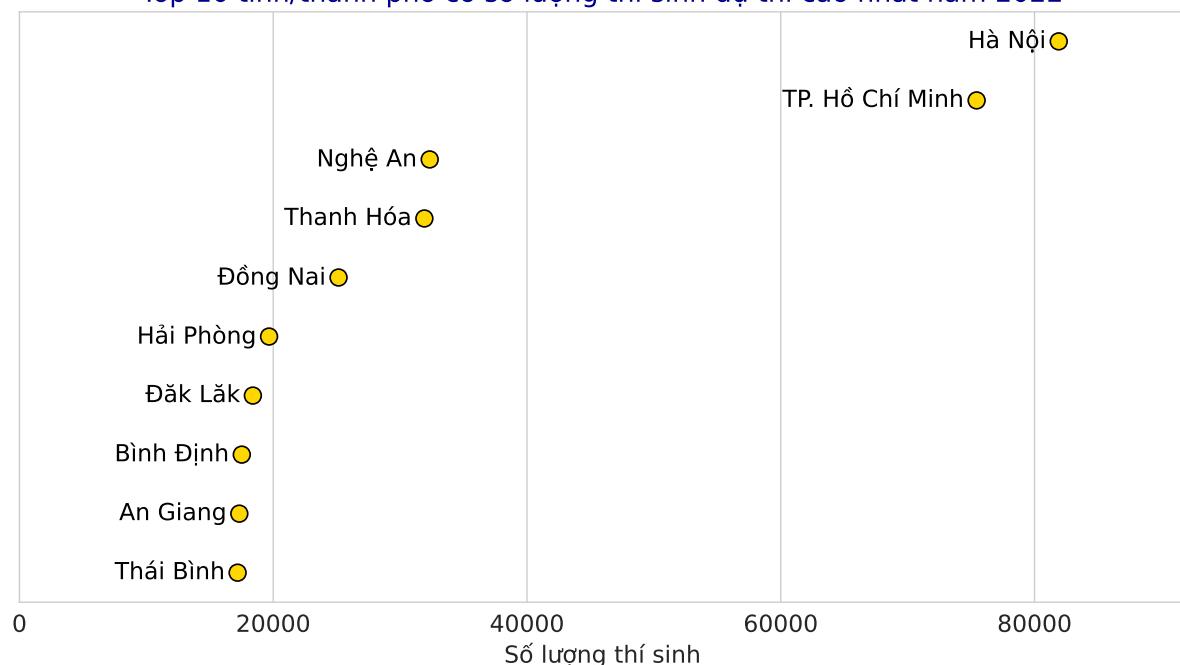
Hình 26: Số lượng thí sinh theo tỉnh/thành phố năm 2020

Top 10 tỉnh/thành phố có số lượng thí sinh dự thi cao nhất năm 2021



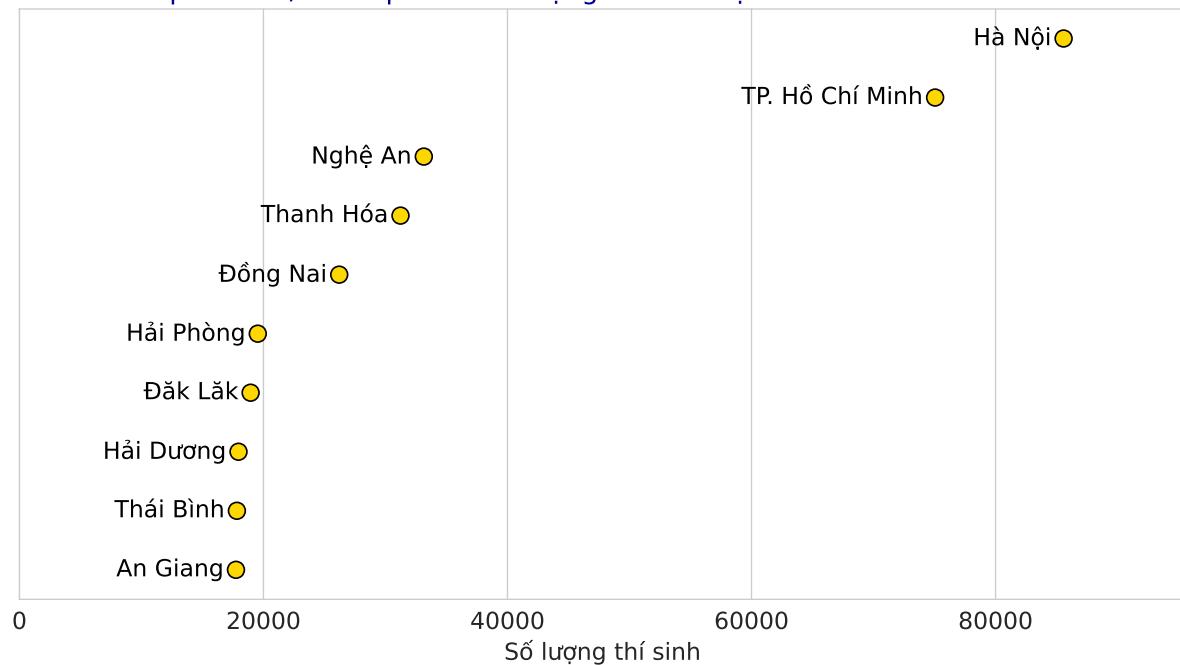
Hình 27: Số lượng thí sinh theo tỉnh/thành phố năm 2021

Top 10 tỉnh/thành phố có số lượng thí sinh dự thi cao nhất năm 2022



Hình 28: Số lượng thí sinh theo tỉnh/thành phố năm 2022

Top 10 tỉnh/thành phố có số lượng thí sinh dự thi cao nhất năm 2023

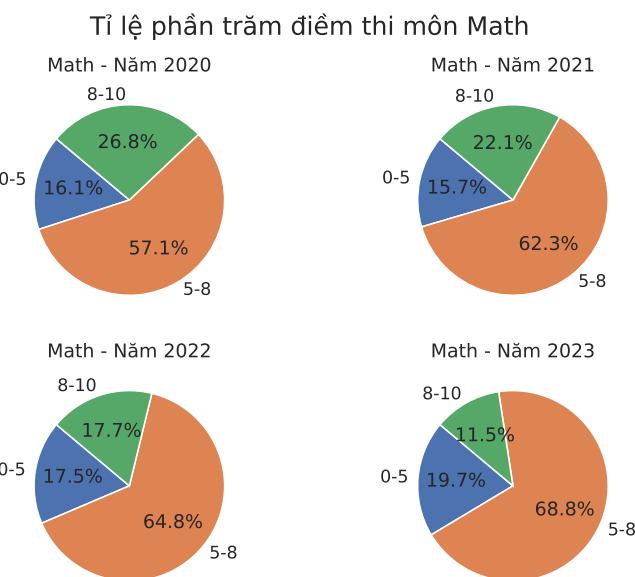


Hình 29: Số lượng thí sinh theo tỉnh/thành phố năm 2023

4.5 Pie chart theo khoảng điểm

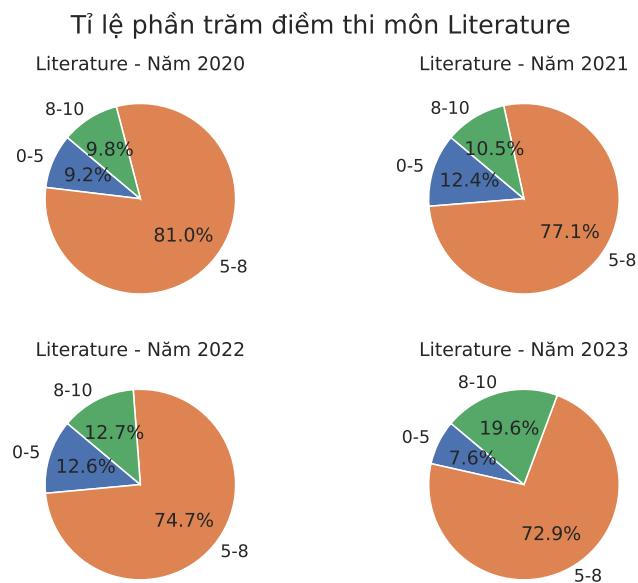
```
1 score_ranges = ['0-5', '5-8', '8-10']
2
3 for subject in subjects:
4     plt.figure(figsize=(8, 6))
5
6     for year in range(2020, 2024):
7         scores = df[(df['Year'] == year)][subject]
8
9         range_counts = [len(scores[(scores >= 0) & (scores <= 5)]),
10                         len(scores[(scores > 5) & (scores <= 8)]),
11                         len(scores[(scores > 8) & (scores <= 10)])]
12
13     plt.subplot(2, 2, year - 2020 + 1)
14     plt.pie(range_counts, labels=score_ranges, autopct='%.1f%%',
15             startangle=140)
16     plt.title(f'{subject} - Năm {year}')
17
18 plt.suptitle(f'Tỉ lệ phần trăm điểm thi môn {subject}', fontsize=16)
plt.savefig(f'figs/Biểu đồ tròn môn {subject.lower()}.pdf', format='pdf')
```

Tỉ lệ phần trăm số thí sinh dưới trung bình môn Toán trong giai đoạn 2020-2023 dao động ở mức từ 16% đến gần 20%. Trong đó, tỉ lệ phần trăm số thí sinh có điểm thi Toán dưới trung bình năm 2023 là cao nhất với 19.7%, tỉ lệ thí sinh đạt ngưỡng điểm từ 8 đến 10 trong năm này cũng là thấp nhất so với 3 năm còn lại với 11.5%. Năm có tỉ lệ phần trăm thí sinh đạt điểm từ 8 đến 10 môn Toán cao nhất là năm 2020 với 26.8%.



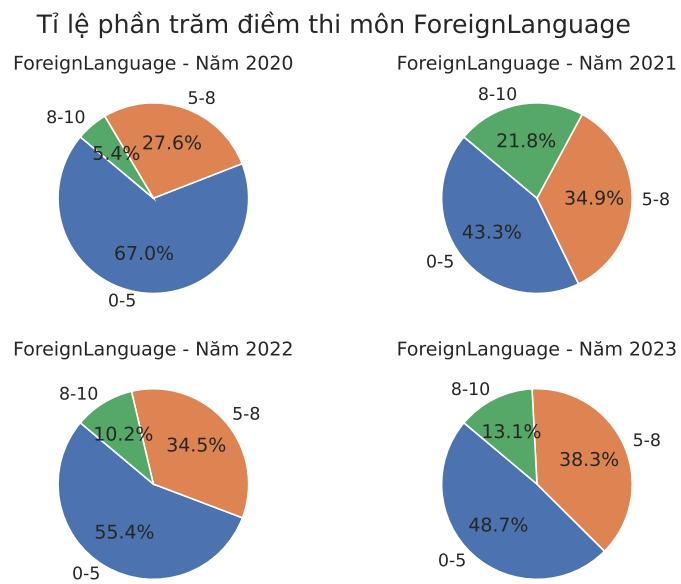
Hình 30: Pie chart theo khoảng điểm môn Toán

Đa phần các thí sinh dự thi đạt điểm 5-8 đối với môn Ngữ Văn, dao động trong khoảng từ 72 đến 81%. Đáng chú ý, năm 2023, có đến gần 20% số thí sinh đạt điểm thi môn Ngữ Văn trên 8. Số tỉ lệ thí sinh có điểm thi Ngữ Văn dưới trung bình nằm ở khoảng 9-12% trong giai đoạn 2020-2023.



Hình 31: Pie chart theo khoảng điểm môn Ngữ Văn

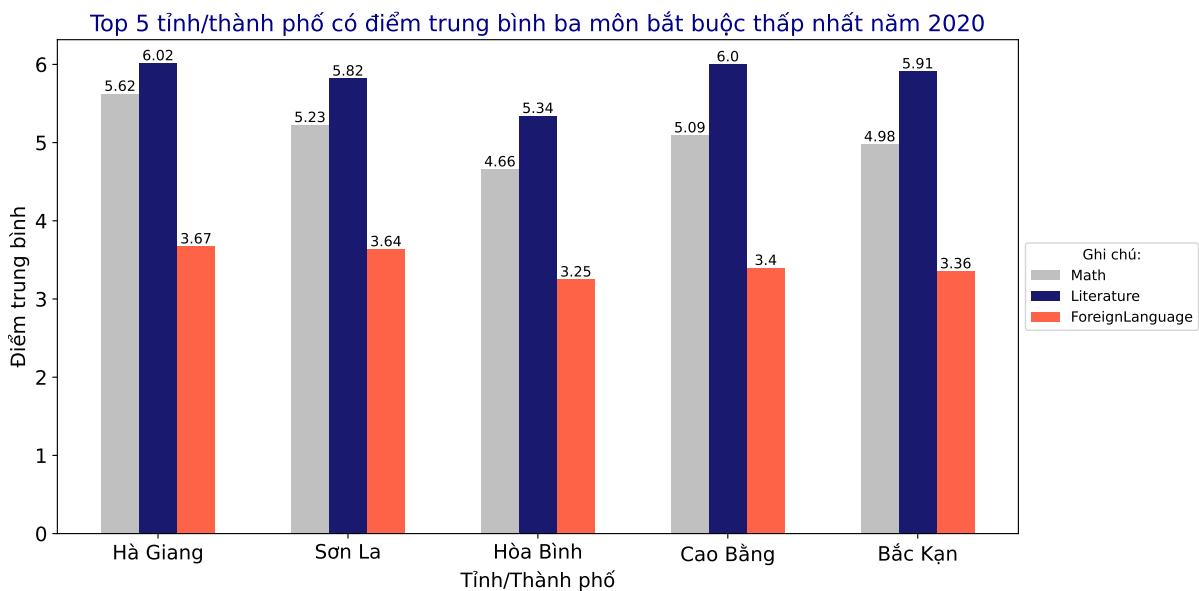
Đối với môn Tiếng Anh, vào năm 2020 và 2022, tỉ lệ phần trăm số thí sinh có điểm dưới trung bình là hơn nửa, với lần lượt 67% và 55.4%. Vào năm 2021 và 2023, số thí sinh có điểm thi dưới trung bình là 43.3% và 48.7%. Năm 2020, tỉ lệ thí sinh thi tiếng Anh đạt điểm trên 8 là chỉ khoảng 5.4%. Năm 2021, tỉ lệ thí sinh đạt điểm trên 8 là cao nhất trong 4 năm với 21.8%.



Hình 32: Pie chart theo khoảng điểm môn Ngoại Ngữ

4.6 Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc (Toán, Ngữ văn, Ngoại ngữ) thấp nhất

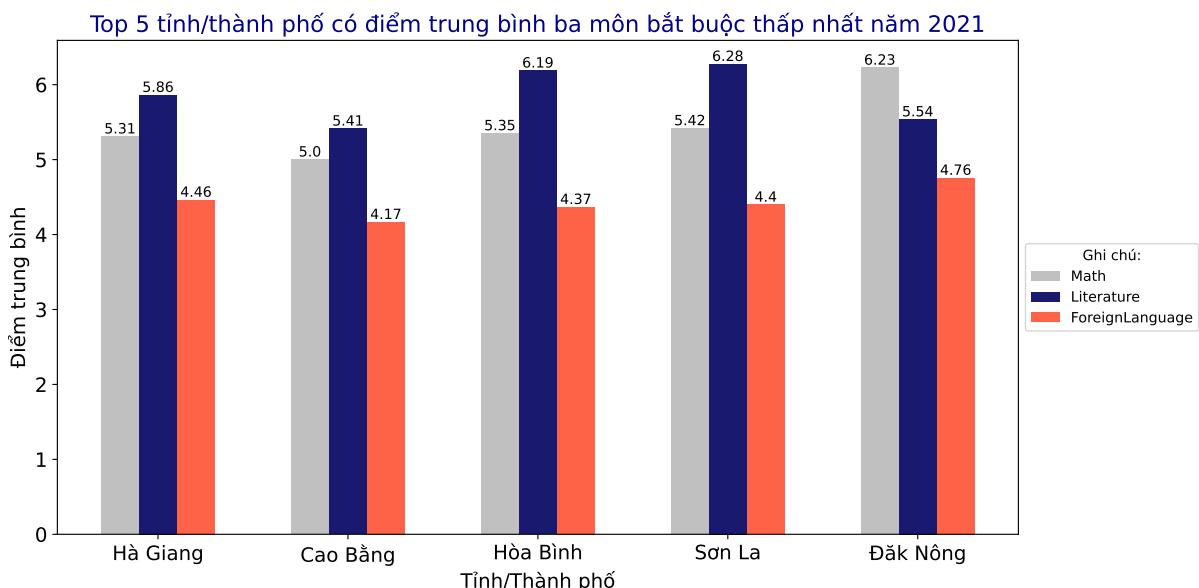
```
1 years = [2020, 2021, 2022, 2023]
2 for year in years:
3     df_year = df[df['Year'] == year]
4     bottom_5_provinces = df_year.groupby('Province')[['Math', 'Literature',
5         ↪ 'ForeignLanguage']].mean().mean(axis=1).nsmallest(5).index
6     bottom_5_df = df_year[df_year['Province'].isin(bottom_5_provinces)]
7     subjects = ["Math", "Literature", "ForeignLanguage"]
8     bar_width = 0.2
9     index = range(len(bottom_5_provinces))
10    colors = sbn.color_palette(['silver', 'midnightblue', 'tomato'])
11
12    plt.figure(figsize=(12, 6))
13    for i, subject in enumerate(subjects):
14        values = bottom_5_df.groupby('Province')[subject].mean()
15        bars = plt.bar([x + i * bar_width for x in index], values, bar_width,
16            ↪ label=subject, color=colors[i])
17
18        for bar in bars:
19            height = bar.get_height()
20            plt.text(bar.get_x() + bar.get_width() / 2, height,
21                ↪ str(round(height, 2)), ha='center', va='bottom')
22
23    plt.xlabel('Tỉnh/Thành phố', fontsize=14)
24    plt.ylabel('Điểm trung bình', fontsize=14)
25    plt.title(f'Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc thấp
26        ↪ nhất năm {year}', fontsize=16, color='darkblue')
27    plt.xticks([x + bar_width for x in index], bottom_5_provinces)
28    plt.xticks(fontsize=14)
29    plt.yticks(fontsize=14)
30    plt.legend(title='Ghi chú:', loc='center left', bbox_to_anchor=(1, 0.5))
31    plt.tight_layout()
32    plt.savefig(f'figs/Top 5 tỉnh thành có điểm trung bình ba môn bắt buộc thấp
33        ↪ nhất năm {year}.pdf', bbox_inches='tight')
34    plt.show()
```



Hình 33: Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc thấp nhất năm 2020

Trong năm 2020, Hà Giang, Sơn La, Hòa Bình, Cao Bằng, Bắc Kạn là 5 tỉnh/ thành phố có điểm trung bình 3 môn bắt buộc Toán, Ngữ văn, Ngoại ngữ thấp nhất.

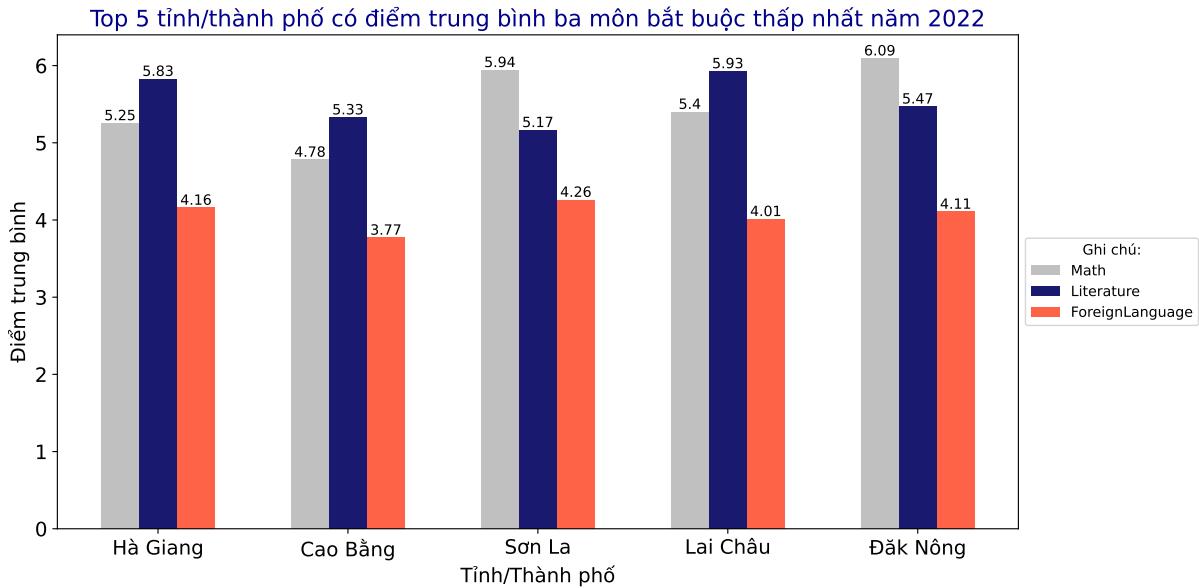
Trong đó, Hòa Bình là tỉnh thành có số điểm trung bình ở cả 3 môn bắt buộc thấp nhất với lần lượt 4.66 điểm, 5.34 điểm và 3.25 điểm. Môn Tiếng Anh vốn không phải là thế mạnh của các tỉnh khu vực miền núi, vùng sâu vùng xa khó khăn.



Hình 34: Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc thấp nhất năm 2021

So với năm 2020, năm 2021, 5 địa phương có điểm trung bình 3 môn bắt buộc thấp nhất không nhiều thay đổi, bao gồm Hà Giang, Cao Bằng, Hòa Bình, Sơn La, Đăk Nông.

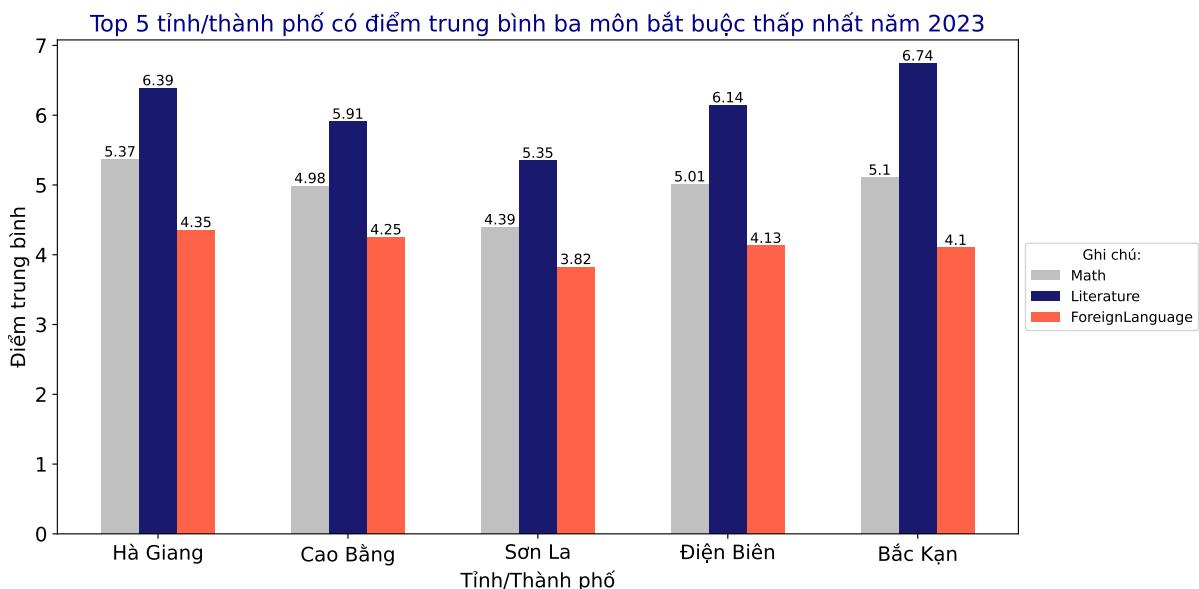
Trong đó Cao Bằng là nơi có mức điểm trung bình ở cả 3 môn Toán, Ngữ văn và Ngoại ngữ thấp nhất trong tất cả các tỉnh với lần lượt số điểm là 5.0, 5.41 và 4.17 điểm.



Hình 35: Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc thấp nhất năm 2022

Năm 2022, top 5 tỉnh/ thành phố có điểm trung bình 3 môn bắt buộc thấp nhất là Hà Giang, Cao Bằng, Sơn La, Lai Châu, Đăk Nông.

Cao Bằng tiếp tục là tỉnh/ thành phố có mức điểm trung bình ở môn Toán và Ngoại ngữ thấp nhất, tương ứng với 4.78 điểm và 3.77 điểm. Đối với môn ngữ Văn, Sơn La có điểm trung bình là 5.17 điểm, đây là mức điểm trung bình thấp nhất ở môn này kể từ năm 2020.



Hình 36: Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc thấp nhất năm 2023

Trong năm 2023, các tỉnh thành như Hà Giang, Cao Bằng, Sơn La tiếp tục nằm trong top các tỉnh/ thành phố có điểm trung bình ở 3 môn Toán, Ngữ văn, Ngoại ngữ thấp nhất.

Trong đó, Sơn La là tỉnh thành có điểm trung bình ở cả 3 môn thấp nhất, thấp hơn khá nhiều so với mức trung bình chung của cả nước với lần lượt là 4.39 điểm, 5.35 điểm và 3.82 điểm.

Nhận xét chung:

Từ năm 2020 đến 2023, các tỉnh thành như Hà Giang, Cao Bằng, Sơn La luôn nằm trong top 5 các tỉnh/ thành phố có mức điểm trung bình 3 môn Toán, Ngữ văn, và Ngoại ngữ thấp nhất cả nước.

Theo Ths. Phạm Thái Sơn, giám đốc Trung tâm Tuyển sinh và Truyền thông, Trường Đại học Công nghiệp Thực phẩm TP.HCM cho hay: "Hà Giang là tỉnh có nhiều học sinh dân tộc, có xuất phát điểm đã khó khăn về kinh tế cùng với việc học tập cũng khó khăn không kém. Đây là một trong những lý do khiến điểm thi tốt nghiệp THPT những năm qua ở Hà Giang luôn thấp là điều dễ hiểu".

Thí sinh ở các tỉnh/thành phố trong giai đoạn 2020-2023 đều gặp khó khăn ở 3 môn bắt buộc, đặc biệt là môn Toán và Ngoại ngữ. Cần phải đưa ra các chương trình học cụ thể và các biện pháp hỗ trợ để cải thiện hiệu suất và nâng cao chất lượng giáo dục ở mỗi địa phương.

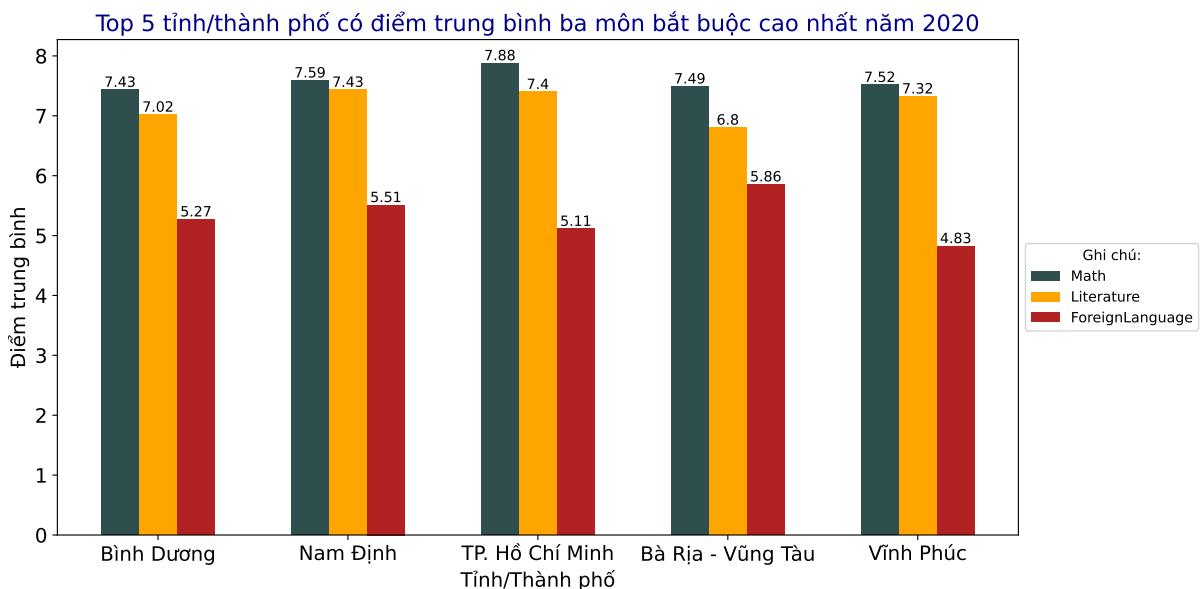
4.7 Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc (Toán, Ngữ văn, Ngoại ngữ) cao nhất

```
1 years = [2020, 2021, 2022, 2023]
2 for year in years:
3     df_year = df[df['Year'] == year]
4     bottom_5_provinces = df_year.groupby('Province')[['Math', 'Literature',
5         ↪ 'ForeignLanguage']].mean().mean(axis=1).nlargest(5).index
6     bottom_5_df = df_year[df_year['Province'].isin(bottom_5_provinces)]
7     subjects = ["Math", "Literature", "ForeignLanguage"]
8     bar_width = 0.2
9     index = range(len(bottom_5_provinces))
10    colors = sns.color_palette(['darkslategrey', 'orange', 'firebrick'])
11
12    plt.figure(figsize=(12, 6))
13    for i, subject in enumerate(subjects):
14        values = bottom_5_df.groupby('Province')[subject].mean()
15        bars = plt.bar([x + i * bar_width for x in index], values, bar_width,
16            ↪ label=subject, color=colors[i])
```

```

15
16     for bar in bars:
17         height = bar.get_height()
18         plt.text(bar.get_x() + bar.get_width() / 2, height,
19                   str(round(height, 2)), ha='center', va='bottom')
20
21     plt.xlabel('Tỉnh/Thành phố', fontsize=14)
22     plt.ylabel('Điểm trung bình', fontsize=14)
23     plt.title(f'Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc cao
24             ↑ nhất năm {year}', fontsize=16, color='darkblue')
25     plt.xticks([x + bar_width for x in index], bottom_5_provinces)
26     plt.xticks(fontsize=14)
27     plt.yticks(fontsize=14)
28     plt.legend(title='Ghi chú:', loc='center left', bbox_to_anchor=(1, 0.5))
29     plt.tight_layout()
30     plt.savefig(f'figs/Top 5 thành phố có điểm trung bình ba môn bắt buộc cao
31             ↑ nhất năm {year}.pdf', bbox_inches='tight')
32     plt.show()

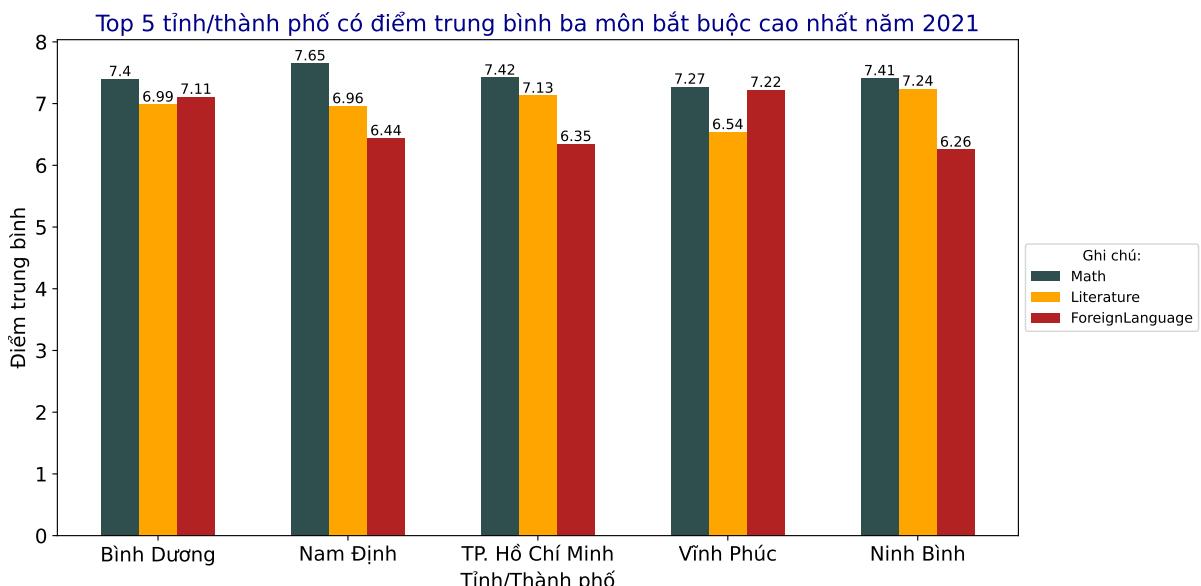
```



Hình 37: Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc cao nhất năm 2020

Top 5 các tỉnh/ thành phố có điểm trung bình 3 môn bắt buộc cao nhất năm 2020 là Bình Dương, Nam Định, TP. Hồ Chí Minh, Bà Rịa – Vũng Tàu, Vĩnh Phúc.

Trong đó, TP. Hồ Chí Minh dẫn đầu cả nước ở điểm thi trung bình môn Toán với 7.88 điểm, tiếp đến là Nam Định, Vĩnh Phúc, Bà Rịa – Vũng Tàu và Bình Dương. Ở môn Ngữ văn, Nam Định đứng đầu với 7.43 điểm. Ngoại ngữ là Bà Rịa – Vũng Tàu với 5.86 điểm.

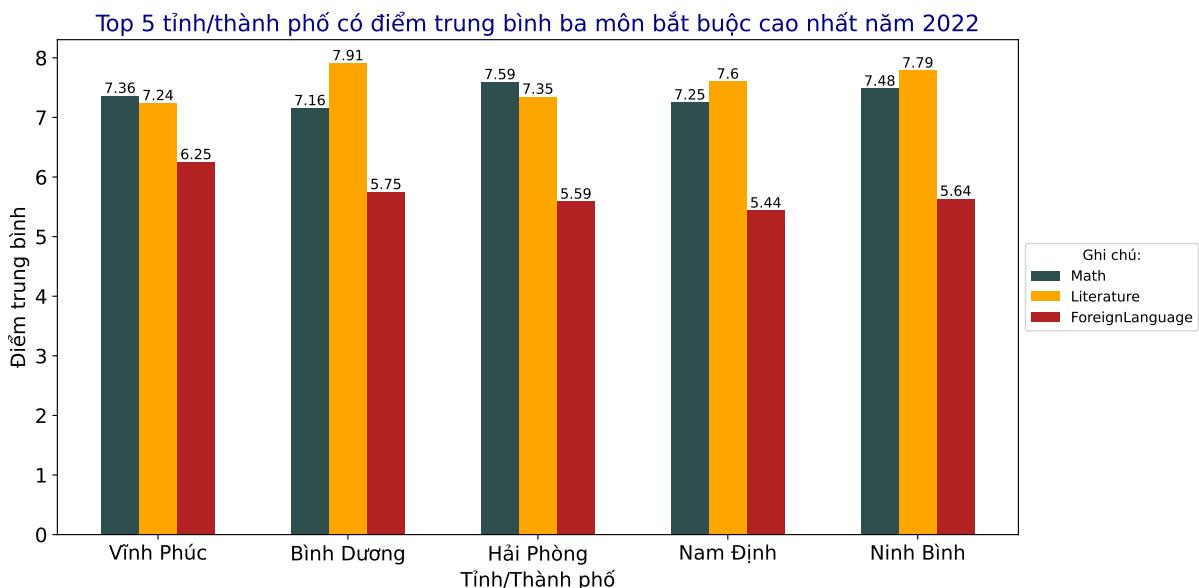


Hình 38: Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc cao nhất năm 2021

Các tỉnh/ thành phố có điểm trung bình 3 môn bắt buộc cao nhất năm 2021 không có quá nhiều thay đổi so với năm 2020 vẫn là Bình Dương, Nam Định, TP. Hồ Chí, Vĩnh Phúc. Tuy nhiên, năm nay, Ninh Bình đã vươn lên trong bảng xếp hạng top 5 với số điểm thống kê nổi bật ở 3 môn Toán, Ngữ văn, Ngoại ngữ.

Nam Định xếp hạng nhất ở điểm trung bình môn Toán với 7.65 điểm. Ở môn Ngữ văn, Ninh Bình là địa phương có điểm số cao nhất là 7.24 điểm.

Đối với môn Ngoại ngữ, Vĩnh Phúc vươn lên dẫn đầu cả nước với điểm số trung bình là 7.22 điểm, tăng 2.39 điểm so mức điểm trung bình của địa phương này vào năm 2020.

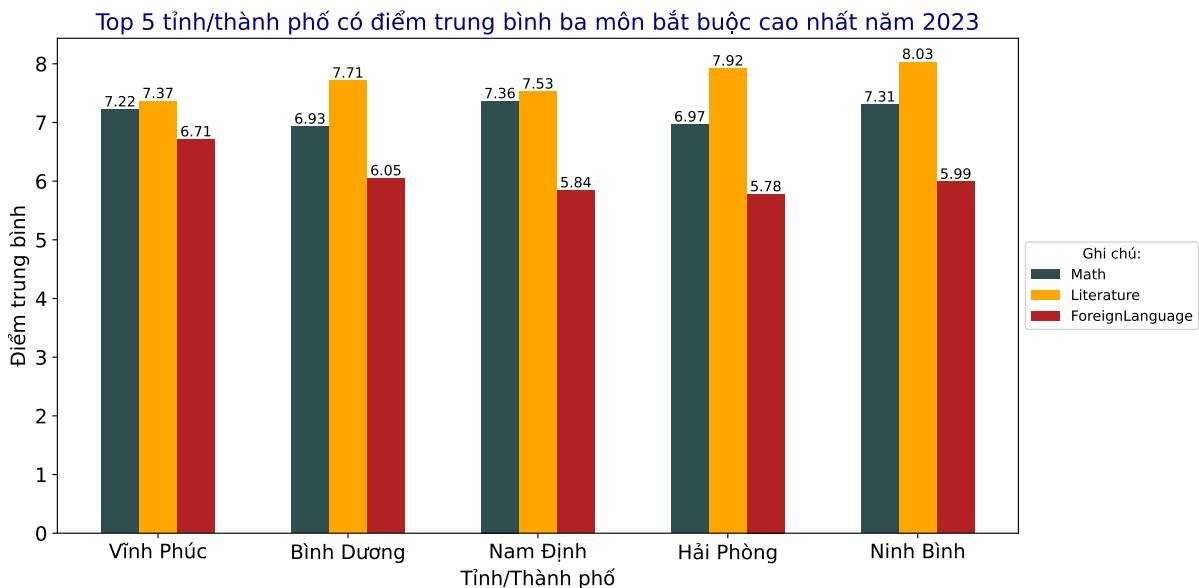


Hình 39: Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc cao nhất năm 2022

Trong năm 2022, Hải Phòng lọt vào top 5 tỉnh/ thành phố có điểm trung bình 3 môn bắt buộc cao nhất, đồng thời dẫn đầu điểm trung bình môn Toán với 7.59 điểm.

Ở môn Ngữ văn, tỉnh có điểm trung bình cao nhất là Bình Dương với 7.91 điểm.

Vĩnh Phúc tiếp tục là địa phương có điểm trung bình môn Ngoại ngữ cao nhất, tuy nhiên có sự giảm xuống đáng kể so với năm 2021, với 6.25 điểm.



Hình 40: Top 5 tỉnh/thành phố có điểm trung bình ba môn bắt buộc cao nhất năm 2023

Trong kỳ thi THPT quốc gia năm 2023, top 5 tỉnh/ thành phố có điểm trung bình 3 môn bắt buộc Toán, Ngữ Văn, Ngoại ngữ cao nhất không thay đổi so với năm 2022.

Với 7.36 điểm, Nam Định trở thành địa phương có điểm trung bình môn Toán cao nhất cả nước vào năm 2023, tuy nhiên vẫn có sự chênh lệch nhẹ so với mức điểm vào năm 2021 là 0.29 điểm.

Tỉnh thành có điểm trung bình môn Ngữ văn cao nhất là Hải Phòng ở ngưỡng 7.92 điểm, và đây cũng là mức điểm cao nhất ở môn học này trong giai đoạn 2020-2023.

Vĩnh Phúc giữ vững vị trí đầu bảng trong 3 năm liên tiếp từ năm 2021 đến năm 2023 với điểm trung bình môn Ngoại ngữ cao nhất, 6.71 điểm, tăng 0.46 điểm so với năm ngoái.

Nhận xét chung:

Trong giai đoạn 2020 – 2023, Bình Dương, Nam Định, Vĩnh Phúc là 3 khu vực luôn nằm trong top 5 các tỉnh/ thành phố có mức điểm trung bình 3 môn bắt buộc cao nhất cả nước, tuy nhiên vẫn có sự chênh lệch giữa các năm. Qua đó, cho ta thấy được rằng chất lượng giáo dục phổ thông ở các tỉnh/ thành phố này rất tốt và ổn định qua các năm.

Trong đó, có thể thấy sự tăng trưởng ngoạn mục của Bình Dương tại kỳ thi tốt nghiệp THPT quốc gia tương đồng với các thành tựu kinh tế vượt trội của tỉnh này trong những năm qua. Theo số liệu của Tổng cục thống kê, kết quả công bố sơ bộ "Khảo sát mức sống dân cư năm 2022", Bình Dương là địa phương có thu nhập bình quân đầu người cao nhất cả nước với 8,076 triệu đồng/người/tháng. Con số này vượt lên trên Hà Nội (6,423 triệu đồng/người/tháng) và TPHCM (6,392 triệu đồng/người/tháng).

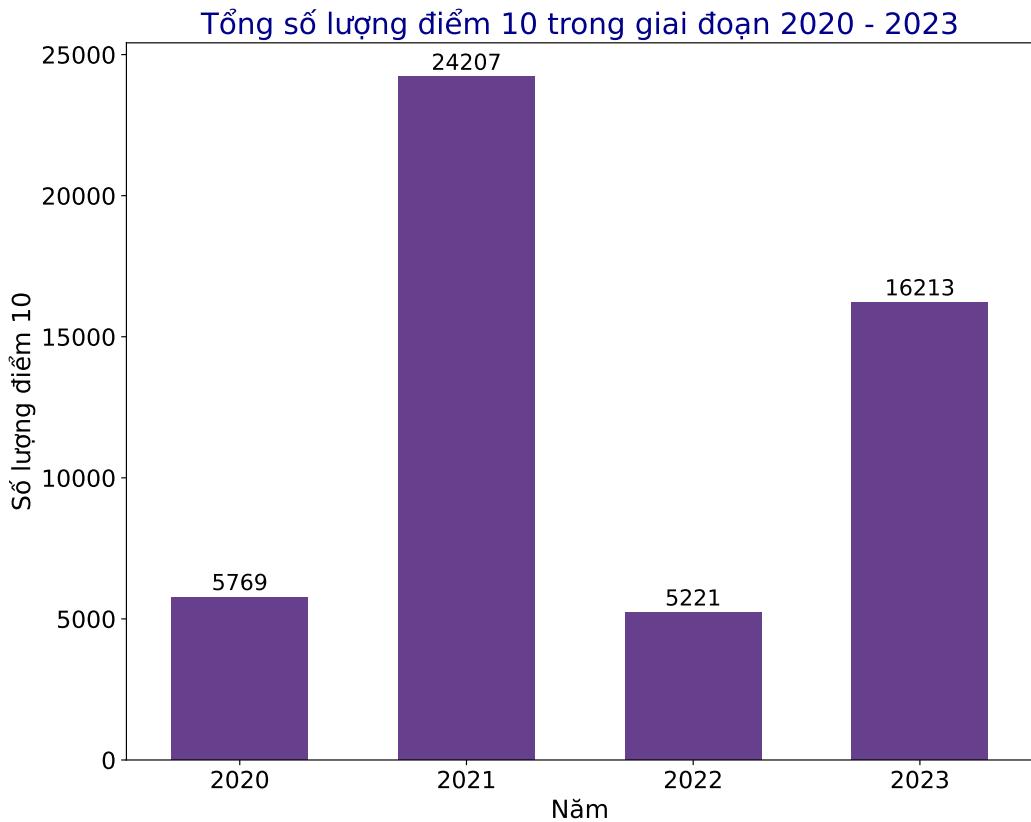
4.8 Tổng số lượng điểm 10 trong giai đoạn 2020 – 2023

```
1 tong_diem_10_theo_nam = df[df.iloc[:, 1:] ==
2     ↳ 10].count(axis=1).groupby(df['Year']).sum()
3 plt.figure(figsize=(10, 8))
4 barplot = sns.barplot(x=tong_diem_10_theo_nam.index,
5     ↳ y=tong_diem_10_theo_nam.values, color='rebeccapurple', width=0.55)
6
7 for p in barplot.patches:
8     barplot.annotate(int(p.get_height()),
9         ↳ (p.get_x() + p.get_width() / 2., p.get_height()),
10        ha='center', va='center',
11        xytext=(0, 9),
12        textcoords='offset points',
13        fontsize=15)
14
15 plt.xlabel('Năm', fontsize=16)
16 plt.ylabel('Số lượng điểm 10', fontsize=16)
17 plt.title('Tổng số lượng điểm 10 trong giai đoạn 2020 – 2023',
18     ↳ color='darkblue', fontsize=20)
```

```

16 plt.xticks(fontsize=16)
17 plt.yticks(fontsize=16)
18 plt.tight_layout()
19 plt.savefig(f'figs/Tổng số lượng điểm 10 theo từng năm.pdf')
20 plt.show()

```



Hình 41: Tổng số lượng điểm 10 trong giai đoạn 2020 – 2023

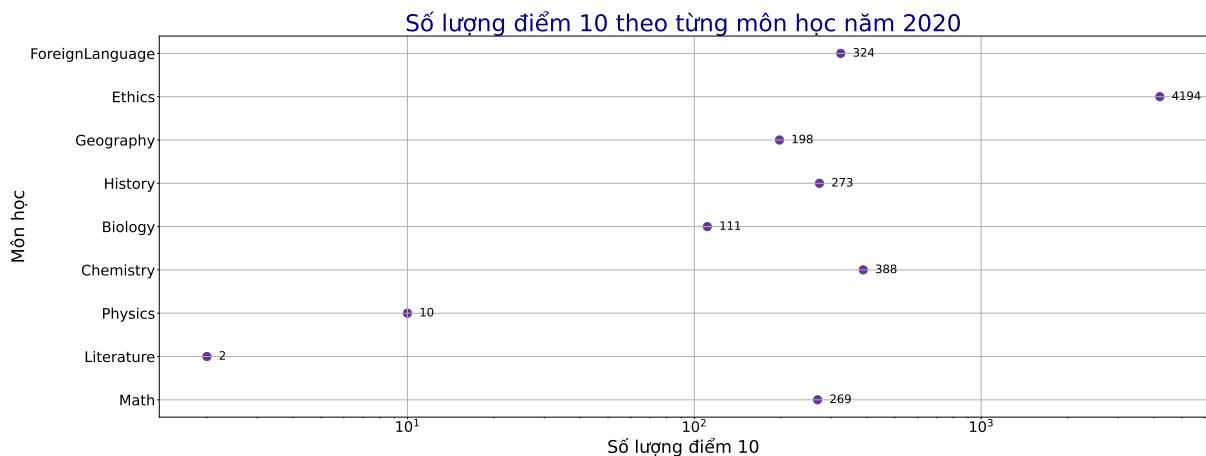
Năm 2020, cả nước có khoảng 5,769 điểm 10 thi tốt nghiệp THPT. Năm 2021, số lượng điểm tuyệt đối là 24,207 điểm, cao hơn gấp 4.1 lần so với năm trước.

Đến kỳ thi tốt nghiệp THPT 2022 có tất cả 5,221 điểm 10, giảm sút so với năm 2021. So với năm 2021, mức này giảm tới 4.6 lần, số lượng này tương đương so với năm 2020.

Có thể thấy trong năm 2023, số lượng điểm 10 tăng đáng kể, lên tới 16,427 điểm 10, cao gấp 3.1 lần so với tổng số điểm tuyệt đối trong năm 2022.

Nhìn chung, 2021 là năm có số lượng điểm 10 cao nhất trong giai đoạn 2020 – 2023.

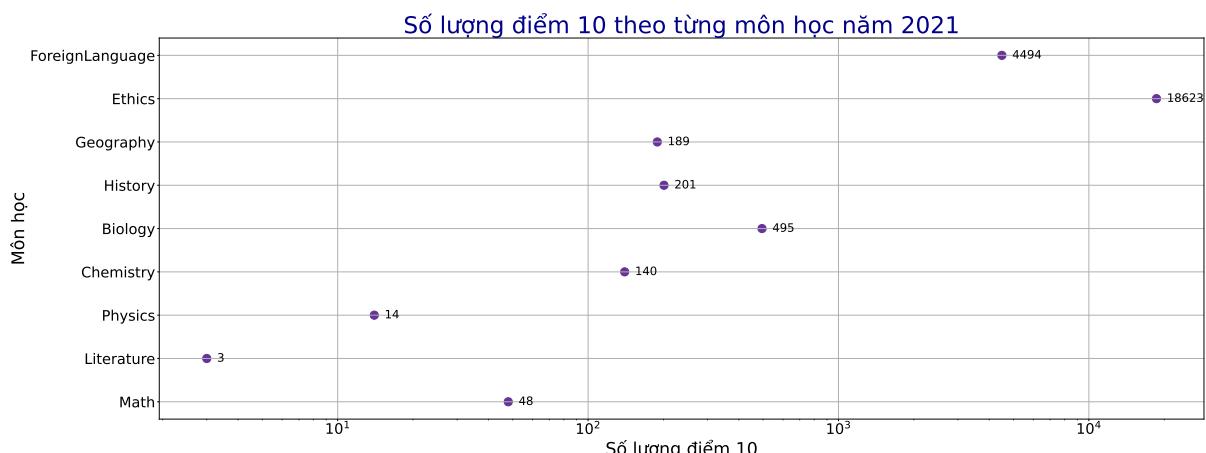
4.8.1 Số lượng điểm 10 ở các môn học trong năm 2020



Hình 42: Biểu đồ thể hiện số lượng điểm 10 theo từng môn năm 2020

Giáo dục công dân có số bài thi đạt điểm 10 cao nhất với 4,194. Các môn như Toán, Hóa học, Lịch sử, Ngoại ngữ có khoảng 273 - 388 điểm 10. Số lượng bài thi đạt điểm tuyệt đối ở 2 môn Ngữ văn và Vật lý rất ít, lần lượt là 2 và 10 bài thi.

4.8.2 Số lượng điểm 10 ở các môn học trong năm 2021



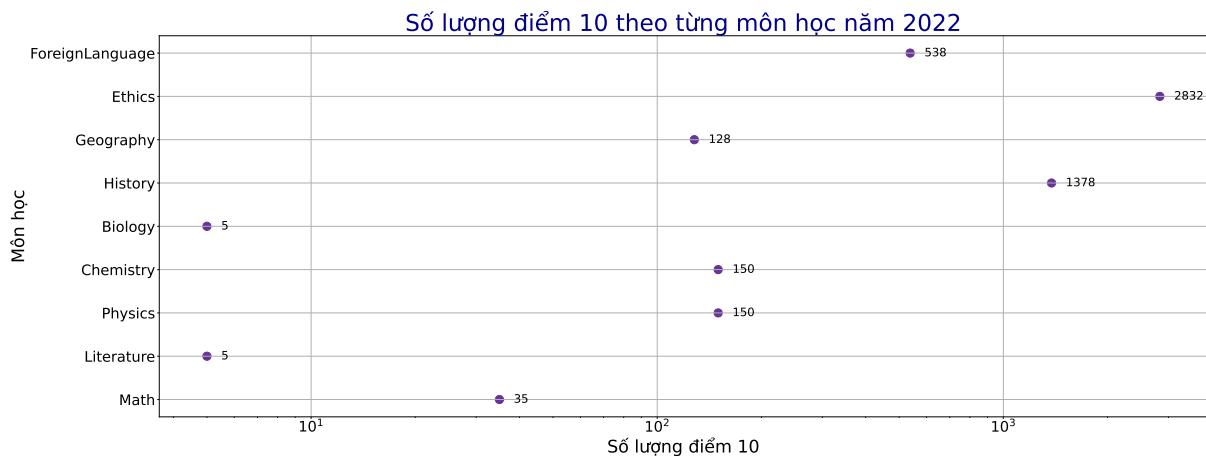
Hình 43: Biểu đồ thể hiện số lượng điểm 10 theo từng môn năm 2021

Cũng như năm 2020, Giáo dục công dân tiếp tục là môn thi có số lượng thí sinh được điểm 10 nhiều nhất trong kỳ thi tốt nghiệp THPT năm 2021, với hơn 18,680 bài thi đạt điểm 10 - gấp 4.45 lần so với năm 2020.

Môn Ngoại ngữ năm nay có số điểm 10 cao vọt so với các năm trước, có hơn 4,345 bài thi điểm 10 - gấp 19.3 lần so với năm ngoái.

Tiếp đến là môn Sinh học, với 582 bài thi điểm 10. Riêng môn Lịch sử được đánh giá là có đề thi khó, nhưng cũng có 201 thí sinh được điểm 10.

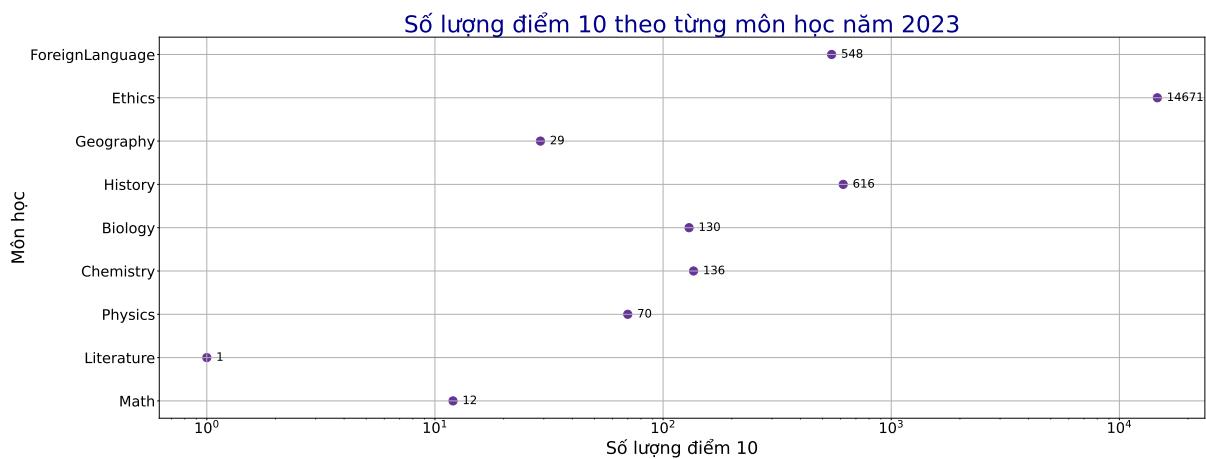
4.8.3 Số lượng điểm 10 ở các môn học trong năm 2022



Hình 44: Biểu đồ thể hiện số lượng điểm 10 theo từng môn năm 2022

Môn Giáo dục công dân có nhiều điểm 10 nhất với 2,832 bài. Đặc biệt, chúng ta có thể chứng kiến sự thay đổi rõ rệt ở môn Lịch sử năm nay khi số lượng thí sinh đạt điểm 10 tăng vọt từ 201 năm 2021 lên 1,378. Trong khi đó, môn Ngữ văn và Sinh học có ít điểm 10 nhất với 5 bài thi mỗi môn.

4.8.4 Số lượng điểm 10 ở các môn học trong năm 2023



Hình 45: Biểu đồ thể hiện số lượng điểm 10 theo từng môn năm 2023

Giáo dục công dân vẫn là môn học ghi nhận số điểm tuyệt đối đạt kỷ lục cao nhất qua các năm. Theo kết quả thống kê, năm nay cả nước ta có 14,671 thí sinh đạt điểm 10 môn học này, gấp hơn 5 lần so với năm 2022, tuy nhiên vẫn thấp hơn so với năm 2021.

Lịch sử và Ngoại ngữ lần lượt xếp vị trí thứ 2, 3 về tổng số điểm 10 (cụ thể lần lượt là 616 và 548).

Tổng số điểm 10 ở các môn đa số đều có xu hướng giảm mạnh: Toán, Lý, Hóa, Sử, Địa,

Ngữ văn và Ngoại ngữ. Duy nhất Giáo dục công dân và Sinh học có số bài thi đạt điểm 10 tăng hơn so với năm ngoái. Năm 2022, môn Sinh học chỉ có 5 bài thi đạt tuyệt đối thì con số năm nay là 130.

Nhận xét chung:

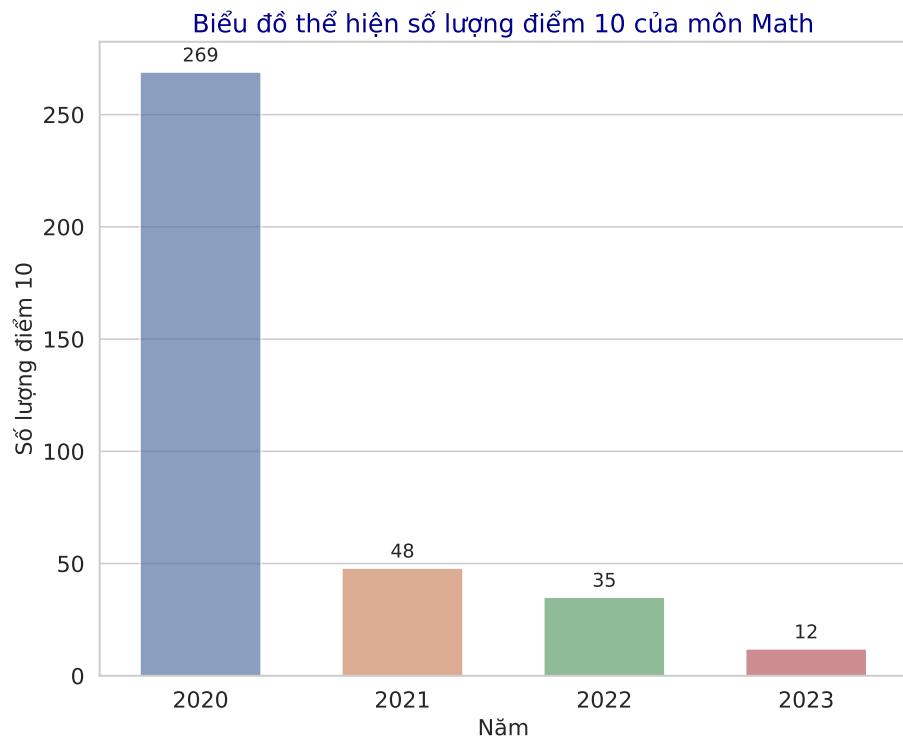
Số lượng điểm tuyệt đối cao nhất luôn nằm ở môn Giáo dục công dân trong giai đoạn 2020 – 2023, kỉ lục vào năm 2021 với 18,623 điểm 10, có thể đưa ra một số lí do như sau: Nội dung kiến thức môn học hiện nay khá ngắn gọn, súc tích. Kiến thức gắn liền với đời sống, nên thí sinh dễ dàng trả lời được các câu hỏi gắn với thực tiễn.

Đánh giá chung, điểm thi các môn trong kỳ thi năm 2021 đều nhỉnh hơn so với năm 2020. Nguyên nhân một phần là do đề thi tốt nghiệp THPT năm 2021 được ra theo hướng nhẹ nhàng, phù hợp với mục tiêu xét tốt nghiệp và điều kiện học sinh phải nghỉ học dài ngày do ảnh hưởng của dịch bệnh.

4.9 Số lượng điểm 10 theo từng môn học

```
1 subjects = ['Math', 'Literature', 'ForeignLanguage', 'Physics', 'Chemistry',
2             → 'Biology', 'History', 'Geography', 'Ethics']
3 plt.figure(figsize=(25, 20))
4
5 for i, subject in enumerate(subjects, 1):
6     ax = plt.subplot(3, 3, i)
7     sns.countplot(x='Year', data=df[df[subject] == 10], alpha=0.7, width=0.6)
8     plt.title(f'Biểu đồ thể hiện số lượng điểm 10 của môn {subject}', 
9               → color='darkblue', fontsize=16)
10    plt.xlabel('Năm', fontsize=14)
11    plt.ylabel('Số lượng điểm 10', fontsize=14)
12    plt.xticks(fontsize=14)
13    plt.yticks(fontsize=14)
14
15    for p in ax.patches:
16        ax.annotate(f'{int(p.get_height())}', (p.get_x() + p.get_width() / 2.,
17                                              → p.get_height(),
18                                              ha='center', va='center', xytext=(0, 10),
19                                              → textcoords='offset points')
20
21    plt.tight_layout()
22    plt.savefig(f'figs/Tổng số lượng điểm 10 của môn {subject}.pdf')
23    plt.show()
```

Môn Toán:



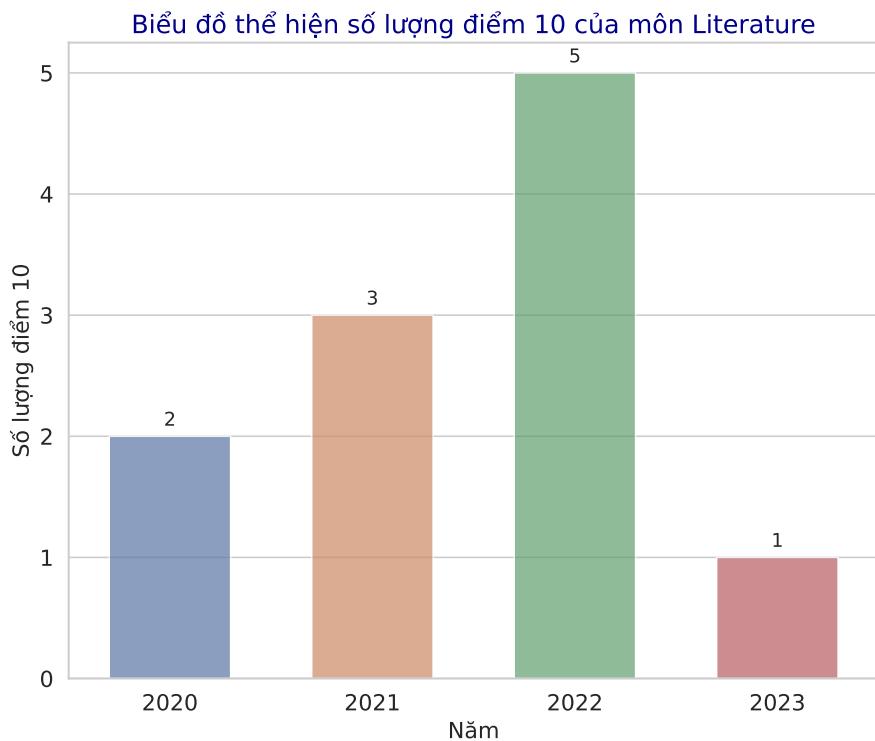
Hình 46: Tổng số lượng điểm 10 môn Toán qua các năm

Từ năm 2020 đến năm 2023, cả nước ghi nhận 364 điểm 10 môn toán trong kỳ thi THPT quốc gia. Trong đó năm 2020, số lượng điểm 10 chạm mức 269 điểm 10, cao nhất ở môn này trong giai đoạn 2020 - 2023.

Mặc dù số lượng thí sinh dự thi tăng đột ngột vào năm 2021, nhưng số lượng bài thi đạt điểm tuyệt đối ở môn Toán lại giảm mạnh, thấp hơn 5.6 lần so với năm 2020.

Với 12 điểm 10 môn toán năm 2023, số lượng điểm 10 kỳ thi tốt nghiệp năm nay ít nhất trong 4 năm trở lại đây.

Môn Ngữ văn:

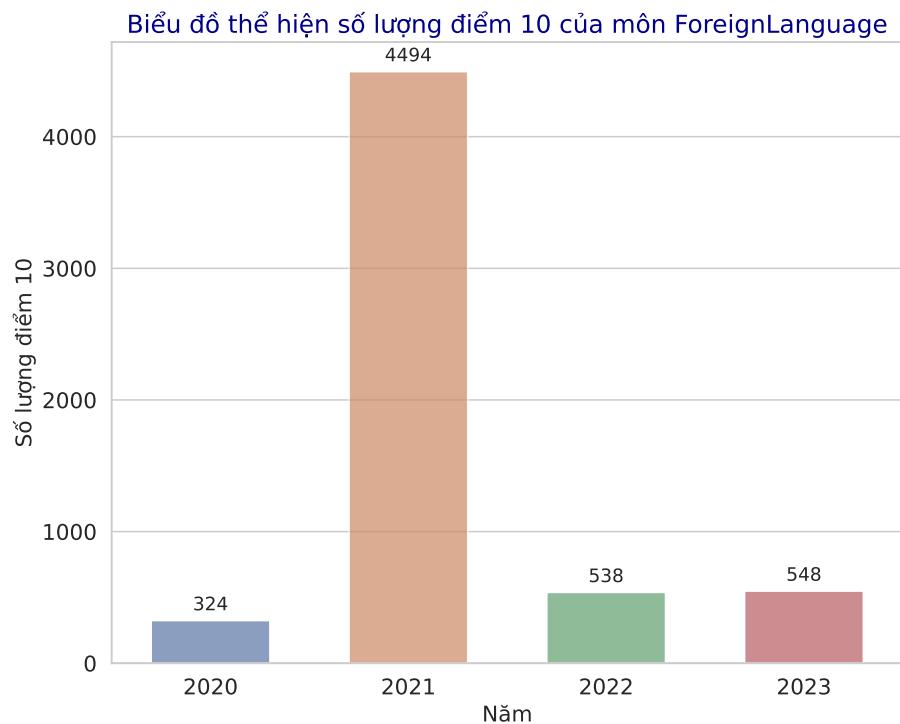


Hình 47: Tổng số lượng điểm 10 môn Ngữ Văn qua các năm

Ở môn Ngữ văn, số điểm 10 luôn ở mức "khan hiếm". Năm 2023, cả nước chỉ có 1 học sinh đạt điểm 10. Con số 1 điểm 10 ở môn văn năm 2023 được xem là tương đối khiêm tốn so với 5 điểm 10 năm 2022.

Trong khi đó, ở kỳ thi tốt nghiệp THPT năm 2021, cả nước ghi nhận 3 điểm 10 môn văn. Ở kỳ thi năm 2020, có 2 thí sinh đạt điểm tuyệt đối.

Môn Ngoại ngữ:



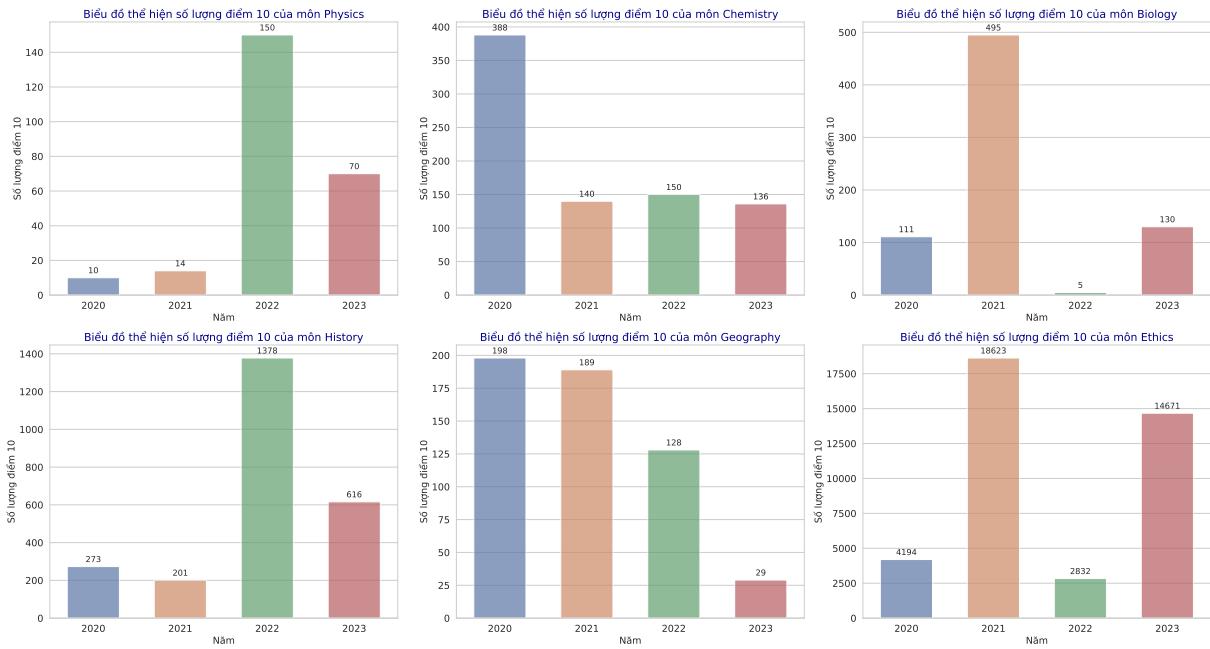
Hình 48: Tổng số lượng điểm 10 môn Ngoại Ngữ qua các năm

Trong giai đoạn 4 năm (2020 - 2023), số thí sinh đạt điểm tối đa môn Ngoại ngữ nhiều nhất là vào năm 2021, với con số "khủng" 4,494 điểm 10. Số lượng điểm tuyệt đối vào năm 2021 gấp hơn 3 lần tổng số điểm 10 ở cả 3 năm 2020, 2022 và 2023 cộng lại. Năm nay cả nước ghi nhận 548 điểm 10, tăng đôi chút so với năm 2022 (538 bài thi đạt điểm 10).

Dưới đây là biểu đồ biến động số điểm 10 các môn còn lại trong giai đoạn 2020 - 2023:

Từ biểu đồ, ta có thể nhận thấy mặc dù môn Giáo dục công dân là môn luôn có số lượng điểm 10 cao nhất, nhưng vẫn có sự biến động mạnh theo từng năm, ghi nhận số lượng nhiều nhất vào năm 2021.

Đáng chú ý có môn Địa lý, ghi nhận xu hướng giảm đều số lượng điểm tuyệt đối trong 4 năm liền. Cụ thể, môn học này ghi nhận 198 điểm 10 (năm 2020), 189 điểm 10 (năm 2021), 128 điểm 10 (năm 2022) và 29 điểm 10 (năm 2023).



Hình 49: Tổng số lượng điểm 10 các môn theo từng năm

4.10 Biểu đồ Parallel Set thể hiện số thí sinh đạt điểm 10 môn Ngữ Văn theo Năm, Tỉnh và Miền

Đầu tiên ta tiến hành thêm cột Miền (gồm miền Bắc, miền Nam, miền Trung vào dữ liệu). Định nghĩa danh sách các tỉnh thuộc ba miền này và một Dictionary với key là tên miền và values là danh sách các tỉnh tương ứng với ba miền. Sau đó dùng hàm apply() để map các tỉnh trong dữ liệu với miền tương ứng:

```

1 # thêm cột miền vào dữ liệu
2
3 # miền Bắc
4 north = [
5     "Lào Cai", "Yên Bái", "Điện Biên", "Hoà Bình", "Lai Châu", "Sơn La", "Hà
6         ↳ Giang", "Cao Bằng",
7     "Bắc Kạn", "Lạng Sơn", "Tuyên Quang", "Thái Nguyên", "Phú Thọ", "Bắc
8         ↳ Giang", "Quảng Ninh",
9     "Bắc Ninh", "Hà Nam", "Hà Nội", "Hải Dương", "Hải Phòng", "Hưng Yên", "Nam
10        ↳ Định", "Ninh Bình",
11     "Thái Bình", "Vĩnh Phúc"
12 ]
13
14 # miền Trung
15 mid = [
16     "Thanh Hoá", "Nghệ An", "Hà Tĩnh", "Quảng Bình", "Quảng Trị", "Thừa Thiên
17         ↳ Huế",
18     "Đà Nẵng", "Quảng Nam", "Quảng Ngãi", "Bình Định", "Phú Yên", "Khánh Hòa",
19 ]

```

```

15     "Ninh Thuận", "Bình Thuận", "Kon Tum", "Gia Lai", "Đăk Lăk", "Đăk Nông",
16     ↪ "Lâm Đồng"
17 ]
18 # miền Nam
19 south = [
20     'TP. Hồ Chí Minh', 'Bình Phước', 'Bình Dương', 'Đồng Nai', 'Tây Ninh', 'Bà
21     ↪ Rịa - Vũng Tàu',
22     'Long An', 'Đồng Tháp', 'Tiền Giang', 'An Giang', 'Bến Tre', 'Vĩnh Long',
23     ↪ 'Trà Vinh',
24     'Hậu Giang', 'Kiên Giang', 'Sóc Trăng', 'Bạc Liêu', 'Cà Mau', 'Cần Thơ'
25 ]
26
27 regions = {
28     'Miền Bắc': north,
29     'Miền Trung': mid,
30     'Miền Nam': south
31 }
32 df['Miền'] = df['Province'].apply(
33     lambda province: next((region for region, provinces in regions.items() if
34         ↪ province in provinces), ''))
35 )

```

Việc vẽ Parallel set sẽ được thực hiện dựa vào thư viện Plotly như sau:

```

1 import plotly.io as pio
2
3 fig = px.parallel_categories(df[df['Literature']==10], color="Year",
4                             dimensions=['Year', 'Complex', 'Miền'],
5                             labels={'Year': 'Year', 'Complex': 'Complex',
6                                     ↪ 'Miền': 'Region'},
7                             title = 'Biểu đồ Parallel Set thể hiện số thí sinh
8                                     đạt điểm 10 môn Ngữ Văn theo Năm, Tổ hợp và
9                                     Miền')
10
11 fig.update_coloraxes(showscale=False)
12
13 fig.update_layout(
14     font=dict(
15         size=24
16     )
17 )

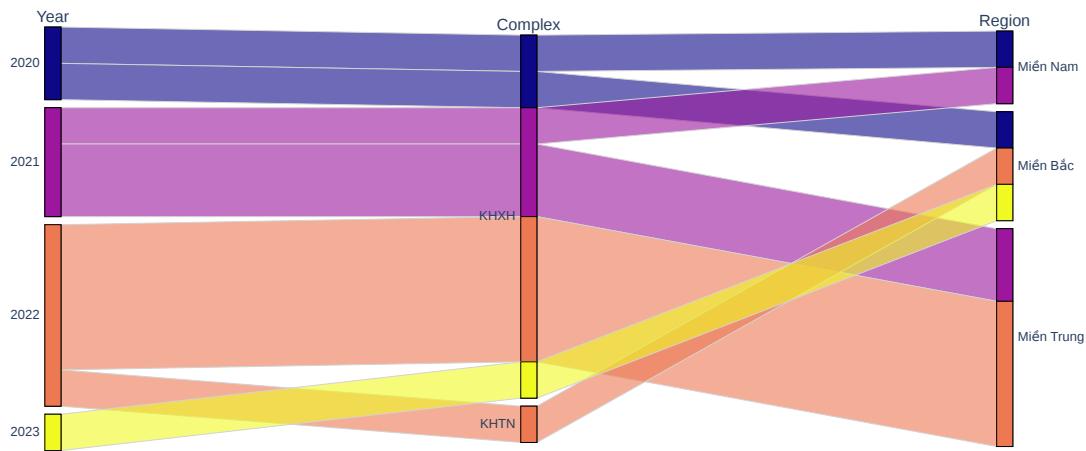
```

```

14 )
15
16 fig.update_xaxes(tickfont=dict(size=24))
17 fig.update_yaxes(tickfont=dict(size=24))
18
19 fig.write_html('figs/literature10.html')
20 pio.write_image(fig, 'figs/literature10.pdf')
21 fig.show()

```

Biểu đồ Parallel Set thể hiện số thí sinh đạt điểm 10 môn Ngữ Văn theo Năm, Tổ hợp và Miền



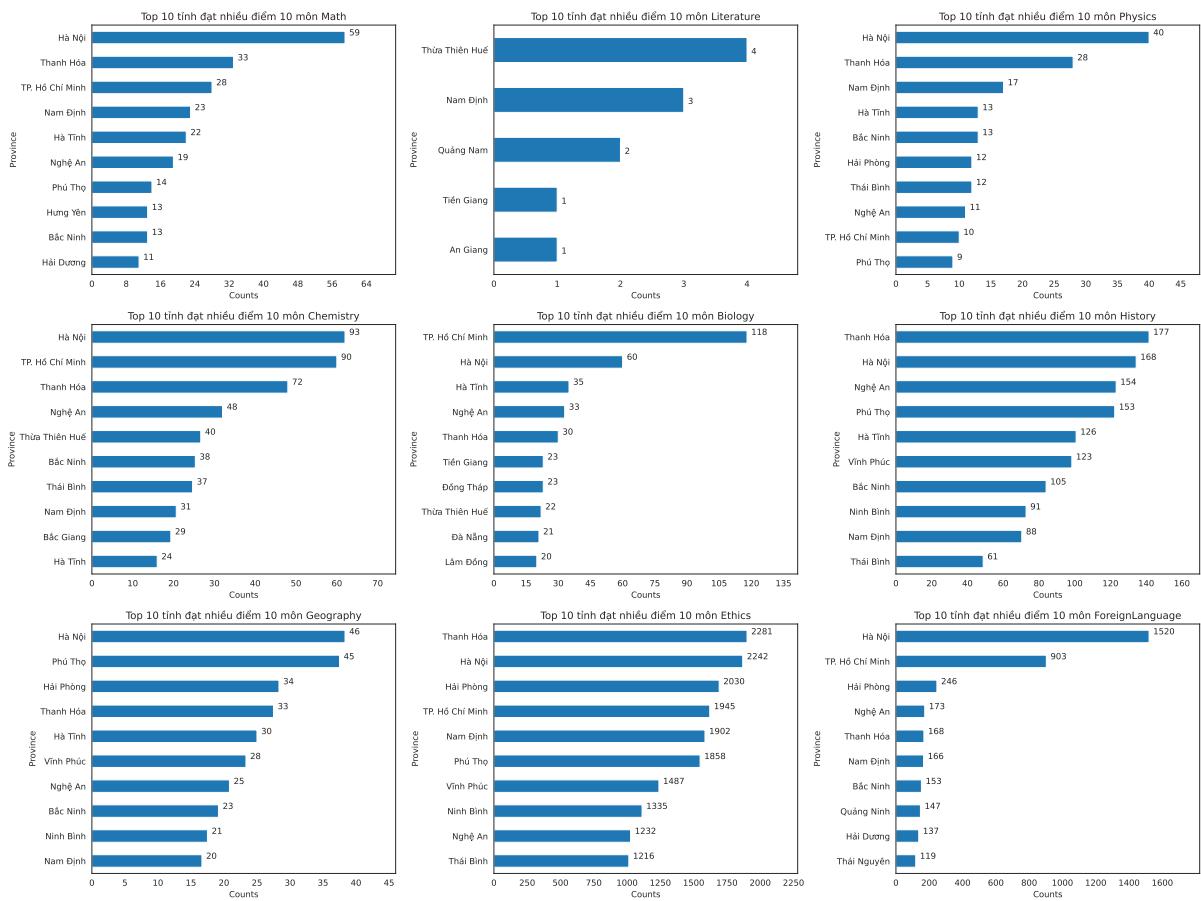
Hình 50: Số lượng thí sinh đạt điểm 10 môn Ngữ Văn theo năm, tổ hợp thi, và miền

Đa phần các thí sinh đạt điểm 10 môn Ngữ Văn ở miền Trung và chọn thi khối KHXH. Năm 2022, có 1 thí sinh thuộc khu vực miền Bắc đạt điểm 10 môn Ngữ Văn và chọn khối thi KHTN, đây cũng là điều chưa từng có tiền lệ trong năm 2020 và 2021.

Cũng trong năm 2022, số thí sinh đạt điểm 10 môn Ngữ Văn là cao nhất với tổng cộng 5 thí sinh, trong đó có 4 thí sinh thuộc khu vực miền Trung, và có 3 thí sinh thuộc tỉnh Thừa Thiên - Huế, theo Báo Lao Động. Đối với khu vực miền Nam, có một thí sinh duy nhất đạt điểm 10 môn Văn vào năm 2020 và một thí sinh vào năm 2021

4.11 Top 10 tỉnh có số lượng thí sinh đạt điểm 10 theo từng môn nhiều nhất

```
1 import matplotlib.pyplot as plt
2 from matplotlib.ticker import MaxNLocator
3
4 subjects = ['Math', 'Literature', 'Physics', 'Chemistry', 'Biology', 'History',
5             → 'Geography', 'Ethics', 'ForeignLanguage']
6 fig, axs = plt.subplots(3, 3, figsize=(20,15)) # Tao 3 hàng, 3 cột subplot
7
8 for i, subject in enumerate(subjects):
9     df_subject = df[df[subject] == 10] # Lọc ra các thí sinh đạt điểm 10
10    top_provinces =
11        → df_subject['Province'].value_counts().head(10).sort_values(ascending=True)
12        → # Lấy top 10 tỉnh có số lượng thí sinh đạt điểm 10 nhiều nhất
13    row = i // 3
14    col = i % 3
15    ax = top_provinces.plot(kind='barh', ax=axs[row, col]) # Vẽ biểu đồ trên
16        → subplot tương ứng
17    axs[row, col].xaxis.set_major_locator(MaxNLocator(integer=True)) # Đặt trục
18        → hoành là số nguyên
19    axs[row, col].set_title(f'Top 10 tỉnh đạt nhiều điểm 10 môn {subject}')
20    axs[row, col].set_ylabel('Province')
21    axs[row, col].set_xlabel('Counts')
22    axs[row, col].set_xticklabels(axs[row, col].get_xticklabels(), rotation=0)
23
24    for p in ax.patches:
25        ax.annotate(str(p.get_width()), (p.get_x() + p.get_width(), p.get_y()),
26                    → xytext=(5, 10), textcoords='offset points')
27    ax.set_xlim(0, top_provinces.max() * 1.2)
28
29 plt.tight_layout()
30 plt.savefig(f'figs/Top 10 tỉnh đạt nhiều điểm 10 theo từng môn.pdf')
31 plt.show()
```



Hình 51: Top 10 tỉnh có số lượng thí sinh đạt điểm 10 theo từng môn nhiều nhất

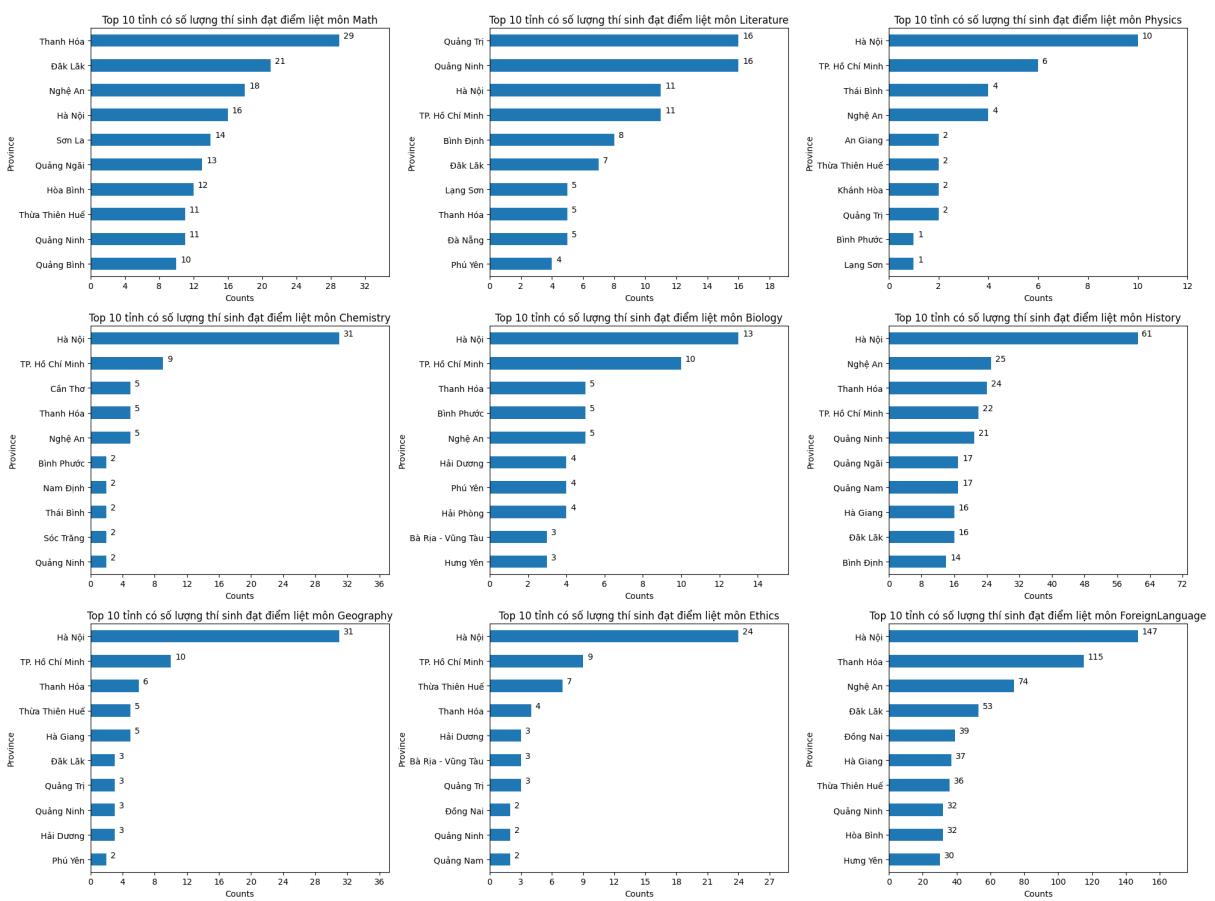
Các tỉnh dẫn đầu về số lượng điểm 10 nổi bậc có: Hà Nội, TP HCM, Nam Định, Thanh Hóa, Nghệ An, Hà Tĩnh. Trong đó Hà Nội đứng đầu với 5 trên 9 môn học có số lượng điểm 10 nhiều nhất cả nước.

Đặc biệt đối với một số môn như Toán, Vật lý, Sinh học hay Ngoại ngữ, cách biệt giữa tỉnh đứng đầu và tỉnh kế tiếp là khá đáng kể, 2 thành phố trung tâm của cả nước là Hà Nội và TP HCM có số lượng điểm 10 vượt xa các tỉnh còn lại. Có thể thấy vị trí địa lý là nhân tố quan trọng quyết định chất lượng của thí sinh. Đối với môn Ngữ văn, vì đặc thù rất ít thí sinh được chấm điểm tuyệt đối, nên cả 4 năm qua số lượng thí sinh đạt điểm 10 môn này rất ít. Đó là lý do biểu đồ chỉ thể hiện được 5 tỉnh thành xếp hạng cao nhất.

Một vấn đề khác có thể thấy là trong danh sách các tỉnh dẫn đầu về số lượng thí sinh đạt điểm tuyệt đối, khu vực phía Bắc chiếm đa số. Điều này chỉ ra rằng chất lượng giáo dục của các khu vực chưa thật sự cân bằng. Các khu vực còn lại cần nhìn nhận và nỗ lực hơn trong việc cải thiện hiệu quả giảng dạy.

4.12 Top 10 tỉnh có số lượng thí sinh đạt điểm liệt (<=1) theo từng môn nhiều nhất

```
1 import matplotlib.pyplot as plt
2 from matplotlib.ticker import MaxNLocator
3
4 df = pd.read_csv('diem_thi_thptqg_2020_2023.csv')
5 subjects = ['Math', 'Literature', 'Physics', 'Chemistry', 'Biology', 'History',
6   ↵ 'Geography', 'Ethics', 'ForeignLanguage']
7 fig, axs = plt.subplots(3, 3, figsize=(20,15)) # Tao 3 hàng, 3 cột subplot
8
9 for i, subject in enumerate(subjects):
10    df_subject = df[df[subject] <= 1] # Lọc ra các thí sinh đạt điểm 1
11    top_provinces =
12      ↵ df_subject['Province'].value_counts().head(10).sort_values(ascending=True)
13      ↵ # Lấy top 10 tỉnh có số lượng thí sinh đạt điểm 10 nhiều nhất
14    row = i // 3
15    col = i % 3
16    ax = top_provinces.plot(kind='barh', ax=axs[row, col]) # Vẽ biểu đồ trên
17      ↵ subplot tương ứng
18    axs[row, col].xaxis.set_major_locator(MaxNLocator(integer=True)) # Đặt trục
19      ↵ hoành là số nguyên
20    axs[row, col].set_title(f'Top 10 tỉnh có số lượng thí sinh đạt điểm liệt
21      ↵ môn {subject}')
22    axs[row, col].set_ylabel('Province')
23    axs[row, col].set_xlabel('Counts')
24    axs[row, col].set_xticklabels(axs[row, col].get_xticklabels(), rotation=0)
25
26    for p in ax.patches:
27      ax.annotate(str(p.get_width()), (p.get_x() + p.get_width(), p.get_y()),
28        ↵ xytext=(5, 10), textcoords='offset points')
29    ax.set_xlim(0, top_provinces.max() * 1.2)
30
31 plt.tight_layout()
32 plt.savefig(f'figs/Top 10 tỉnh có số lượng thí sinh đạt điểm liệt theo từng
33      ↵ môn.pdf')
34 plt.show()
```



Hình 52: Top 10 tỉnh có số lượng thí sinh đạt điểm liệt (≤ 1) theo từng môn nhiều nhất

Có thể thấy tên các tỉnh đứng đầu cả nước về số lượng điểm 10 ở trên vẫn xuất hiện trong danh sách các tỉnh có số lượng thí sinh bị điểm liệt cao nhất, tiêu biểu là Hà Nội, TP HCM, Thanh Hóa, Nghệ An. Điều này chỉ ra sự không đồng đều ở chất lượng thí sinh và chất lượng giáo dục giữa các đơn vị của từng địa phương. Tuy nhiên, việc này là không thể tránh khỏi đối với những nơi có số lượng thí sinh đông đặc biệt là 2 thành phố trung tâm là Hà Nội và TP HCM, vì là nơi tập trung dân số đông đúc. Số lượng thí sinh lớn dẫn đến việc số lượng điểm liệt cũng cao hơn.

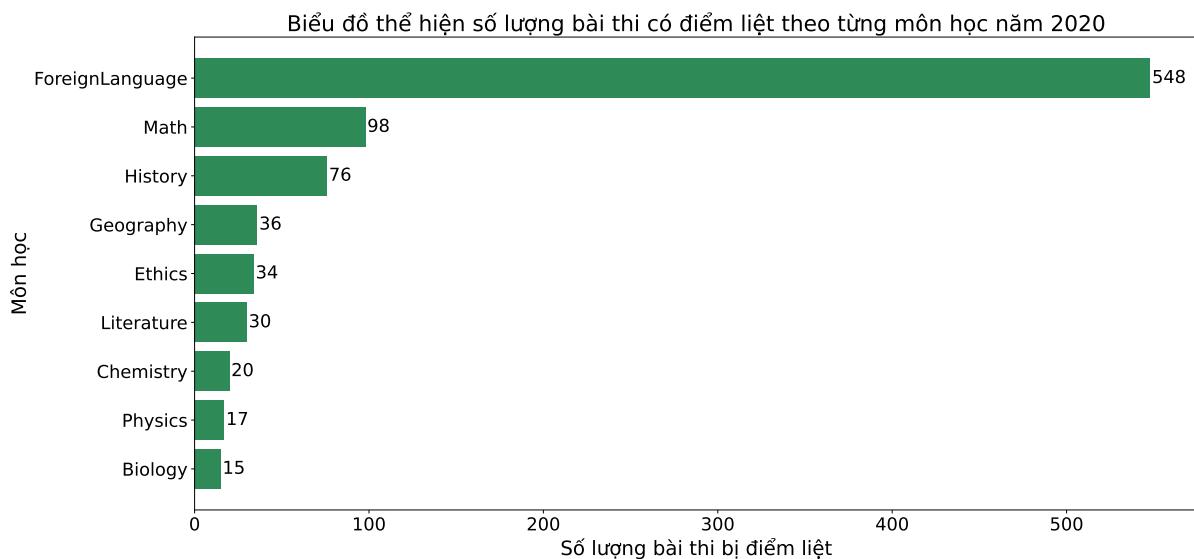
Tuy nhiên, các địa phương cũng cần có sự nhìn nhận về mặt bằng chung của chất lượng giáo dục cả nước để chủ động tích cực nâng cao năng lực đào tạo của địa phương mình, nhằm giảm thiểu sự chênh lệch. Nhà nước cũng nên quan tâm đến việc hỗ trợ, tạo điều kiện cho các khu vực miền núi, vùng nông thôn có cơ hội tiếp cận với tiêu chuẩn giáo dục tốt hơn.

Một hiện tượng khác là môn Ngoại ngữ có số lượng điểm liệt cao hơn nhiều so với các môn còn lại, mặc dù số lượng thí sinh đạt điểm 10 môn này lại tương đối cao. Hiện tượng này cho thấy sự phân cực rõ rệt trong trình độ ngoại ngữ giữa các thí sinh. Vẫn đề khác biệt hệ chương trình đào tạo ngoại ngữ ở bậc phổ thông là một nguyên nhân dẫn tới chất

lượng học sinh không đồng đều, kết quả thi không đồng đều. Hiệu quả của hệ chương trình 7 năm (từ lớp 6) thường không tốt bằng hệ 10 năm (từ lớp 3). Tuy vậy hệ 10 năm chủ yếu được triển khai ở những tỉnh, thành phố có điều kiện kinh tế - xã hội phát triển. Có thể thấy việc hiệu quả của việc dạy và học môn này là không đồng đều. Vấn đề này đặt ra yêu cầu cần cải thiện cách dạy, cách học, cách kiểm tra, đánh giá cũng như tăng cường đầu tư về ngoại ngữ từ các đơn vị giáo dục.

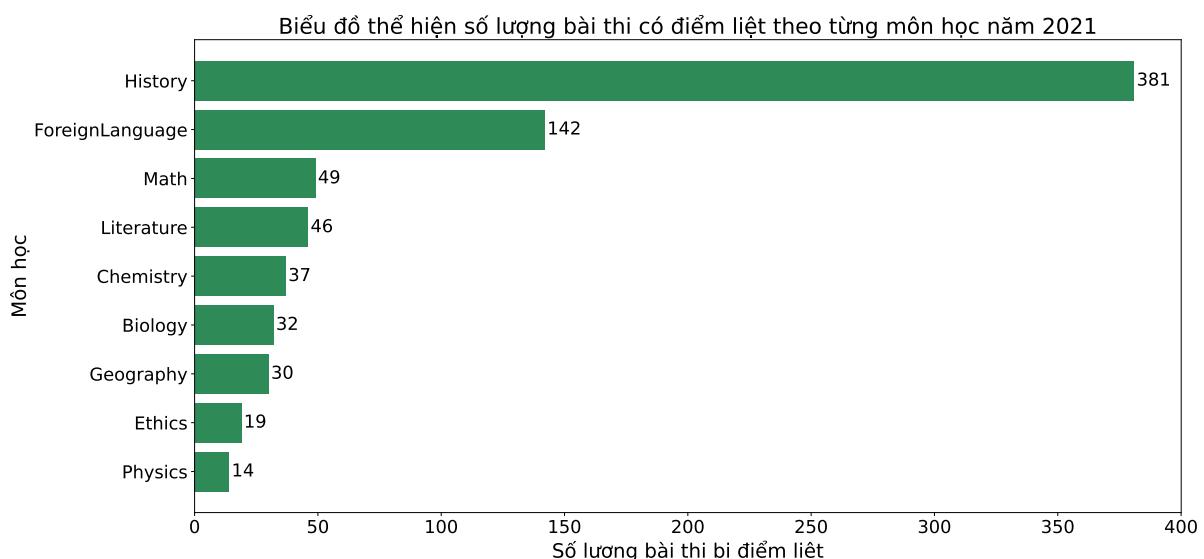
4.13 Biểu đồ thể hiện số lượng bài thi có điểm liệt (<=1) theo từng môn

```
1 subjects = ["Math", "Literature", "Physics", "Chemistry", "Biology", "History",
2   ↪ "Geography", "Ethics", "ForeignLanguage"]
3 for i,year in enumerate(df.Year.unique(),1):
4     df_year = df[df['Year'] == year]
5     plt.figure(figsize=(17, 8))
6     u1 = df_year[df_year[subjects] <= 1][subjects].count()
7     sorted_u1 = u1.sort_values(ascending=True)
8
9     for subject in sorted_u1.index:
10         value = sorted_u1[subject]
11         plt.barh(subject, value, label=f'{subject}: {value}', color='seagreen')
12         plt.text(value+1, subject, str(value), ha='left', va='center', fontsize=18)
13
14     plt.ylabel('Môn học', fontsize=20)
15     plt.xlabel('Số lượng bài thi bị điểm liệt', fontsize=20)
16     plt.title(f'Biểu đồ thể hiện số lượng bài thi có điểm liệt theo từng môn học
17   ↪ năm {year}', fontsize=22)
18     plt.xticks(fontsize=18)
19     plt.yticks(fontsize=18)
20     plt.tight_layout()
21     plt.savefig(f'figs/Barchart Số lượng bài thi có điểm liệt theo từng môn học
22   ↪ năm {year}.pdf')
23     plt.show()
```



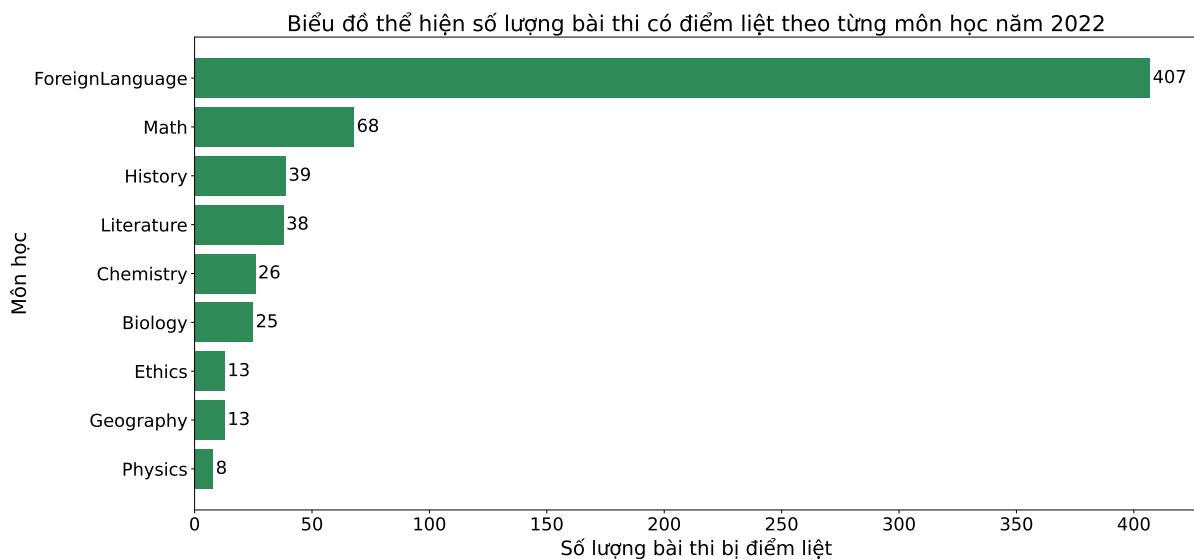
Hình 53: Biểu đồ thể hiện số lượng thí sinh đạt điểm liệt (≤ 1) theo từng môn năm 2020

Năm 2020, 3 môn học có nhiều bài thi bị điểm liệt nhất lần lượt là môn Ngoại ngữ (549 bài thi, cao vượt trội so với các môn còn lại), kế đến là môn Toán (98 bài thi) và môn Lịch sử (76 bài thi). Cũng trong năm 2020, các môn học có số lượng bài thi bị điểm liệt ít nhất đều là 3 môn trong tổ hợp KHTN, môn Sinh (15 bài thi), môn Lý (17 bài thi), môn Hóa (20 bài thi).



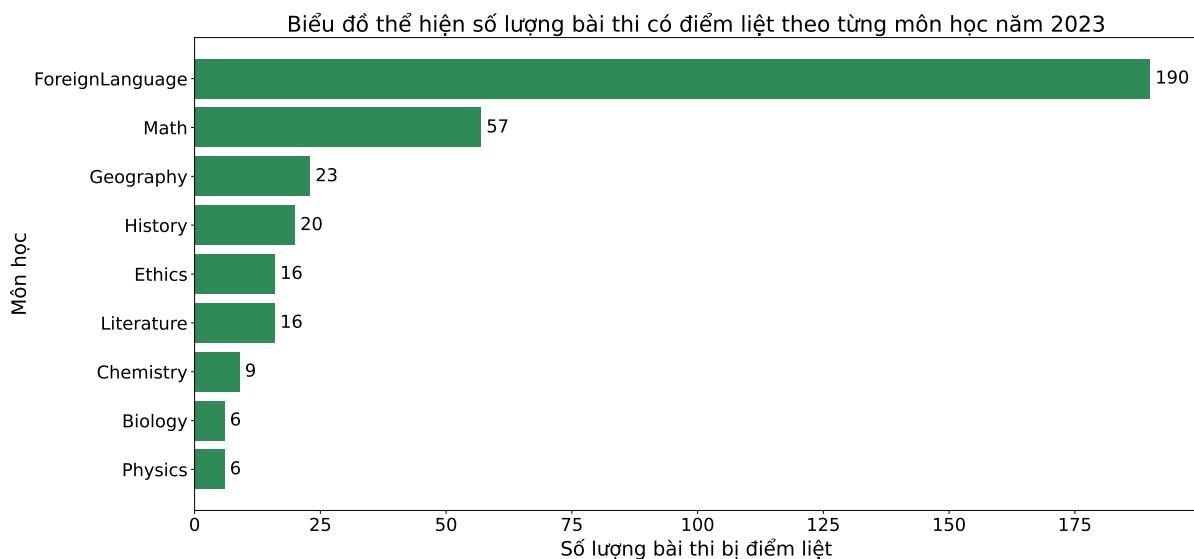
Hình 54: Biểu đồ thể hiện số lượng thí sinh đạt điểm liệt (≤ 1) theo từng môn năm 2021

Năm 2021, 3 môn học có nhiều bài thi bị điểm liệt nhất vẫn là Ngoại ngữ, Lịch sử và Toán nhưng có sự thay đổi về thứ tự: đứng đầu là môn Lịch sử (381 bài thi), kế đến là môn Ngoại ngữ (142 bài thi) và môn Toán (49 bài thi). 3 môn học có số lượng bài thi bị điểm liệt ít nhất lần lượt là môn Lý (14 bài thi), môn GD&CD (19 bài thi) và môn Địa (30 bài thi).



Hình 55: Biểu đồ thể hiện số lượng thí sinh đạt điểm liệt (≤ 1) theo từng môn năm 2022

Năm 2022, 3 môn học có nhiều bài thi bị điểm liệt nhất vẫn là Ngoại ngữ, Lịch sử và Toán: đứng đầu là Ngoại ngữ (407 bài thi), kế đến là môn Toán (68 bài thi) và môn Lịch sử (39 bài thi, nhiều hơn môn Văn 1 bài thi). 3 môn học có số lượng bài thi bị điểm liệt ít nhất lần lượt là môn Lý (8 bài thi), môn GDCD và môn Địa (13 bài thi).



Hình 56: Biểu đồ thể hiện số lượng thí sinh đạt điểm liệt (≤ 1) theo từng môn năm 2023

Năm 2023 là năm duy nhất trong 4 năm mà môn Lịch sử không nằm trong 3 môn có số lượng bài thi bị điểm liệt nhiều nhất, thay vào đó là môn Địa (23 bài thi). Dẫn đầu vẫn là môn Ngoại ngữ (190 bài thi) và môn Toán (57 bài thi). Các môn học có số lượng bài thi bị điểm liệt ít nhất đều là 3 môn trong tổ hợp KHTN, môn Lý (6 bài thi), môn Sinh (6 bài thi), môn Hóa (9 bài thi).

Nhận xét chung:

Qua 4 năm, tuy môn Ngoại ngữ là một trong số 3 môn có số lượng điểm 10 nhiều nhất nhưng cũng là một trong số 3 môn có số bài thi bị điểm liệt nhiều nhất, chứng tỏ có sự chênh lệch lớn về trình độ ở môn Ngoại ngữ, các bài thi đa phần sẽ thuộc vào 1 trong 2 loại: có điểm thi xuất sắc và có điểm thi rất thấp.

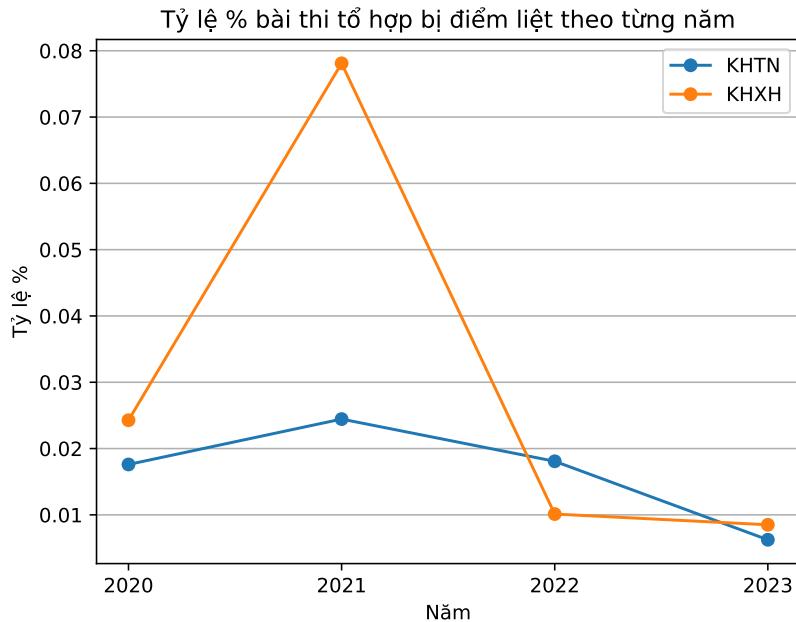
Nhìn chung, các môn trong tổ hợp KHTN đều có số lượng bài thi bị điểm liệt thấp hơn khi so với tổ hợp KHXH. Tuy nhiên, vì số lượng thí sinh thi tổ hợp KHTN ít hơn nhiều so với thí sinh thi tổ hợp KHXH nên không thể vội kết luận rằng thi tổ hợp KHXH sẽ dễ bị điểm liệt hơn tổ hợp KHTN.

4.14 Biểu đồ thể hiện tỷ lệ % bài thi bị điểm liệt của 2 tổ hợp theo từng năm

Để đánh giá tổ hợp nào có tỷ lệ bài thi bị điểm liệt nhiều hơn, ta sẽ so sánh tỷ lệ % bài thi bị điểm liệt trong tổ hợp KHTN (bằng số bài thi tổ hợp KHTN có ít nhất 1 trong 3 môn bị điểm liệt chia tổng số bài thi KHTN và nhân với 100) với tỷ lệ % bài thi bị điểm liệt trong tổ hợp KHXH qua 4 năm, sau đó biểu diễn dưới dạng line plot.

```
1 years = df['Year'].unique()
2 khtn = df[df['Complex']=='KHTN'][['Physics', 'Chemistry', 'Biology']]
3 khxh = df[df['Complex']=='KHXH'][['History', 'Geography', 'Ethics']]
4 khtn_rate = []
5 for y in years:
6     khtn_yr = khtn[df['Year'] == y]
7     rate = (khtn_yr <= 1).any(axis=1).sum() / len(khtn_yr) * 100
8     khtn_rate.append(rate)
9
10 khxh_rate = []
11 for y in years:
12     khxh_yr = khxh[df['Year'] == y]
13     rate = (khxh_yr <= 1).any(axis=1).sum() / len(khxh_yr) * 100
14     khxh_rate.append(rate)
15
16 plt.plot(years, khtn_rate, label='KHTN', marker = 'o')
17 plt.plot(years, khxh_rate, label='KHXH', marker = 'o')
18 plt.title('Tỷ lệ % bài thi tổ hợp bị điểm liệt theo từng năm')
19 plt.xlabel('Năm')
20 plt.ylabel('Tỷ lệ %')
21 plt.grid(axis = 'y')
22 plt.legend()
23 plt.savefig(f'figs/Biểu đồ thể hiện tỷ lệ bài thi bị điểm liệt của 2 tổ hợp
→ theo từng năm.pdf')
```

```
24 plt.show()
```



Hình 57: Biểu đồ thể hiện số lượng thí sinh đạt điểm liệt (<=1) theo từng môn năm 2023

Năm 2020 và 2021 có tỷ lệ % bài thi bị điểm liệt trong tổ hợp KHXH đều cao hơn so với tổ hợp KHTN, thậm chí cao hơn rất nhiều tại năm 2021. Năm 2022 là năm duy nhất có tỷ lệ % bài thi bị điểm liệt trong tổ hợp KHXH thấp hơn so với tổ hợp KHTN. Và tới năm 2023, không còn có sự chênh lệch nhiều giữa 2 tổ hợp. Có thể thấy rằng từ năm 2022, tỷ lệ % bài thi bị điểm liệt giảm ở cả 2 tổ hợp là một tín hiệu đáng mừng. Nhưng nhìn chung, tỷ lệ bài thi bị điểm liệt trong tổ hợp KHXH vẫn cao hơn so với tổ hợp KHTN.

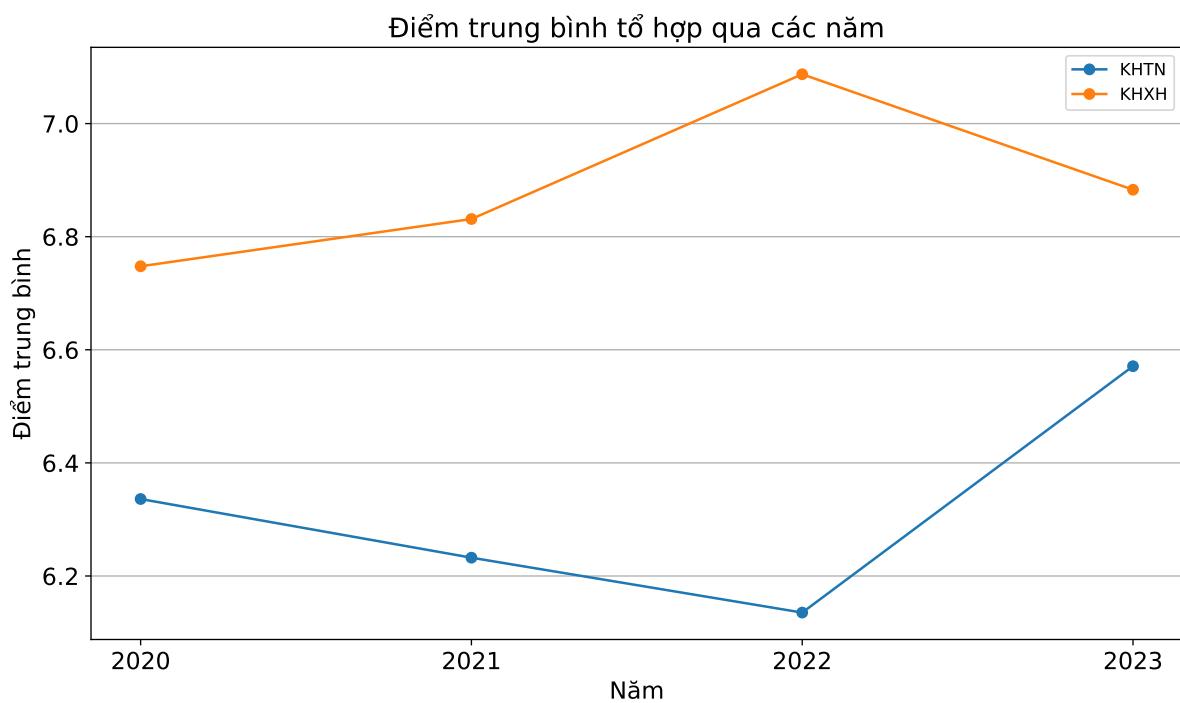
4.15 Biểu đồ điểm trung bình bài thi tổ hợp

```
1 # Tính điểm trung bình bài thi tổ hợp
2 df.loc[df['Complex'] == 'KHTN', 'Complex_avg'] = df[df['Complex'] ==
   ↵ 'KHTN'][['Physics', 'Chemistry', 'Biology']].mean(axis=1)
3 df.loc[df['Complex'] == 'KHXH', 'Complex_avg'] = df[df['Complex'] ==
   ↵ 'KHXH'][['History', 'Geography', 'Ethics']].mean(axis=1)
4
5 df_complex_avg = df.groupby(['Year',
   ↵ 'Complex'])['Complex_avg'].mean().reset_index()
6 df_complex_avg['Year'] = df_complex_avg['Year'].astype('category')
7
8 #Biểu đồ
```

```

9 plt.figure(figsize=(10, 6))
10
11 for complex in df_complex_avg['Complex'].unique():
12     df_complex = df_complex_avg[df_complex_avg['Complex'] == complex]
13     plt.plot(df_complex_avg['Year'].unique(), df_complex['Complex_avg'],
14             marker='o', label=complex)
15
16 plt.xlabel('Năm', fontsize=14)
17 plt.xticks(df_complex_avg['Year'].unique())
18 plt.ylabel('Điểm trung bình', fontsize=14)
19 plt.title('Điểm trung bình tổ hợp qua các năm', fontsize=16)
20 plt.grid(axis = 'y')
21 plt.xticks(fontsize=14)
22 plt.yticks(fontsize=14)
23 plt.tight_layout()
24 plt.legend()
25 plt.savefig(f'figs/Điểm trung bình bài thi tổ hợp.pdf')
26 plt.show()

```



Hình 58: Biểu đồ điểm trung bình bài thi tổ hợp

Điểm trung bình của tổ hợp KHTN và KHXH chênh lệch rõ rệt, đáng chú ý nhất là năm 2022 khi cách biệt của 2 tổ hợp là gần 1 điểm. Lý giải cho hiện tượng này có thể dựa vào độ khó không cân bằng của đề thi giữa 2 tổ hợp cũng như sự khác biệt về khối lượng

kiến thức của các môn thi thành phần. Thực tế qua các kỳ thi những năm gần đây, bài thi tổ hợp KHXH thường được đánh giá là “dễ thở hơn” so với tổ hợp KHTN. Đặc biệt, môn GD&K kế từ lần đầu tiên được đưa vào bài thi tổ hợp đã có không ít ý kiến trái chiều về sự phù hợp của nó khi đưa vào kỳ thi quan trọng này.

Có thể hiểu việc đưa GD&K vào kỳ thi THPTQG là một giải pháp chấn chỉnh của Bộ GD đối với hiện tượng lơ là trong việc dạy và học môn học thường được xem là “môn phụ” cũng như là một cách giúp giảm bớt áp lực thi cử. Tuy nhiên, cần nhìn nhận rằng hạ thấp yêu cầu đầu ra cũng dẫn đến việc giảm chất lượng học sinh sau THPT. Việc điểm thi trung bình của môn GD&K luôn cao hơn hẳn so với mặt bằng chung các môn còn lại trong cả 2 tổ hợp cho thấy vẫn còn một số bất cập. Điều này một phần khiến cho điểm trung bình của bài thi tổ hợp KHXH luôn cao hơn KHTN, trong khi tỉ lệ điểm liệt của tổ hợp KHTN thường thấp hơn tổ hợp KHXH. Nếu chỉ so sánh điểm trung bình bài thi tổ hợp sẽ không phản ánh đúng chất lượng của 2 nhóm thí sinh.

Năm học 2022-2023 là năm đầu tiên của lộ trình thay thế môn GD&K bằng môn Giáo dục Kinh tế và Pháp luật ở cấp THPT, theo chương trình giáo dục phổ thông 2018. Có thể thấy những nhà hoạch định giáo dục đã và đang có nhiều nỗ lực nâng cao chất lượng của chương trình đào tạo và giải quyết những bất cập còn tồn tại trong kỳ thi. Vì vậy phương án thi THPT năm 2025 sẽ có nhiều thay đổi mạnh mẽ điều chỉnh theo chương trình giáo dục đổi mới và được kỳ vọng sẽ được xây dựng phù hợp hơn, kế thừa được những ưu điểm và giải quyết được những bất cập của kỳ thi hiện hành.

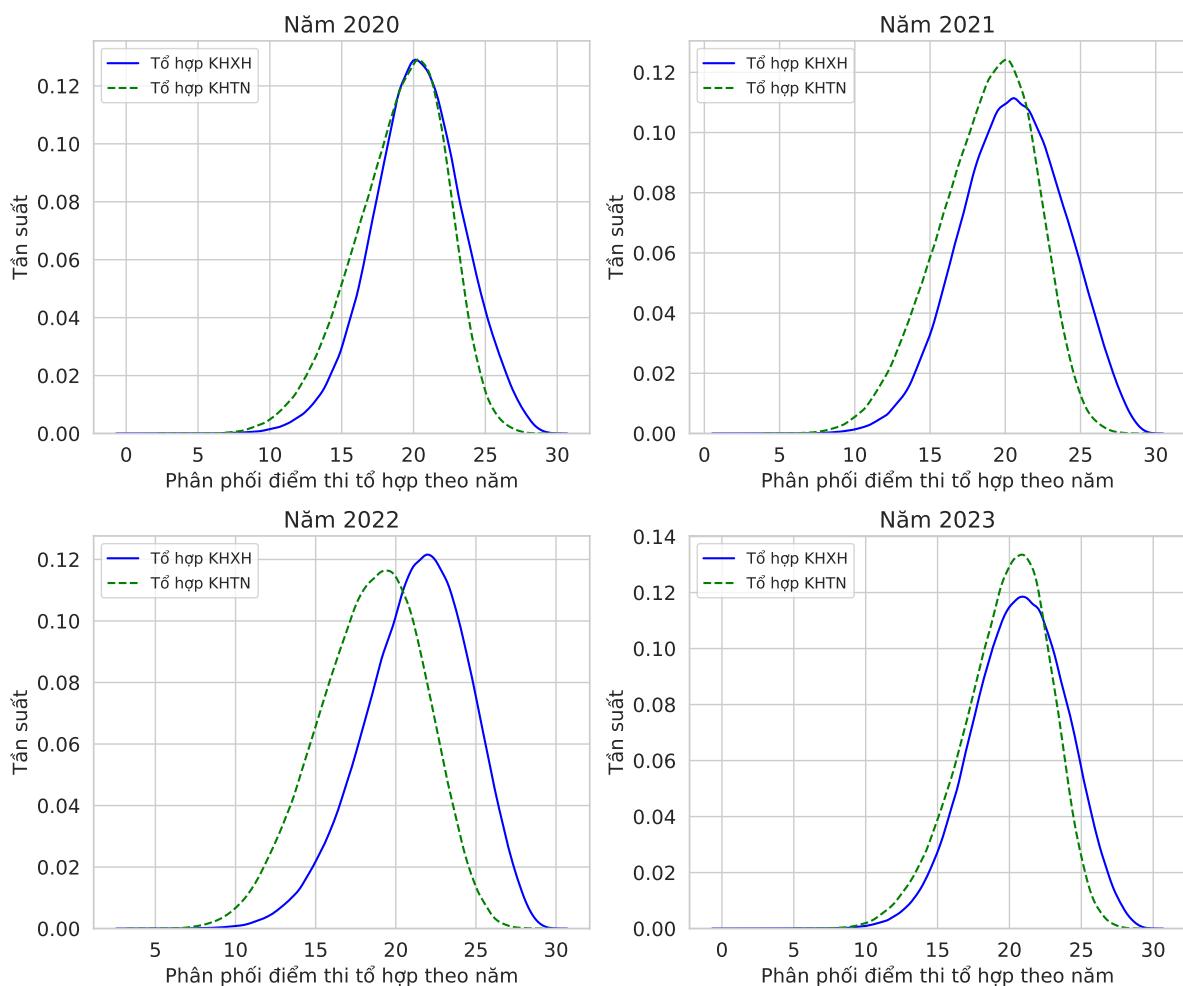
4.16 Biểu đồ phân phối tổng điểm thi 2 khối KHTN và KHXH qua các năm

```
1 years = [2020, 2021, 2022, 2023]
2
3 linestyles = ['-', '--', ':', '-.']
4 colors = ['blue', 'green', 'red', 'purple']
5
6 fig, axes = plt.subplots(2, 2, figsize=(12, 10))
7
8 for i, year in enumerate(years):
9     ax = axes[i // 2, i % 2]
10
11     data_year = df[df['Year'] == year]
12
13     for j, complex_value in enumerate(data_year['Complex'].unique()):
14         data_complex = data_year[data_year['Complex'] == complex_value]
15         sbn.kdeplot(data=data_complex['Điểm thi tổ hợp'], ax=ax, label=f'Tổ hợp
16             {complex_value}',
```

```

16                         linestyle=linestyles[j], color=colors[j])
17
18     ax.set_title(f'Năm {year}', fontsize=16)
19     ax.set_xlabel('Phân phối điểm thi tổ hợp theo năm', fontsize=14)
20     ax.set_ylabel('Tần suất', fontsize=14)
21
22     ax.tick_params(axis='both', labelsize=14)
23     ax.legend(fontsize=14)
24     ax.legend()
25
26 plt.tight_layout()
27 plt.savefig('figs/complex.pdf')
28 plt.show()

```



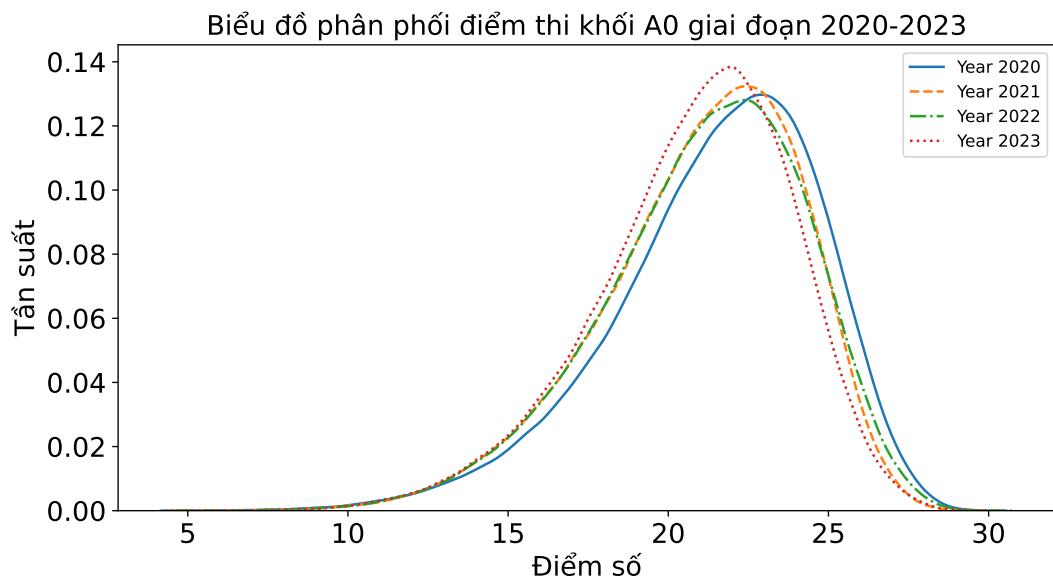
Hình 59: Top 10 tỉnh có số lượng thí sinh đạt điểm liệt (≤ 1) theo từng môn nhiều nhất

Nhìn chung, tổng điểm thi các môn giữa hai tổ hợp khá tương đồng nhau, với đỉnh nằm ở khoảng 20 điểm. Mức chênh lệch về điểm trung bình giữa hai tổ hợp KHTN và KHXH

cao nhất là vào năm 2022, với điểm trung bình của tổ hợp KHTN là 18.40 và tổ hợp KHXH lên đến mức 21.26. Trong đó, theo báo Lao Động , vào năm 2022, trái ngược với tổ KHTN, nhiều thí sinh thi tổ hợp môn KHXH cho rằng đề thi khá dễ. Song, điểm mode của tổ hợp KHTN vào năm 2023 là cao nhất khi so sánh với điểm mode của tổ hợp KHXH và KHTN vào các năm trước, với mức điểm 21.00.

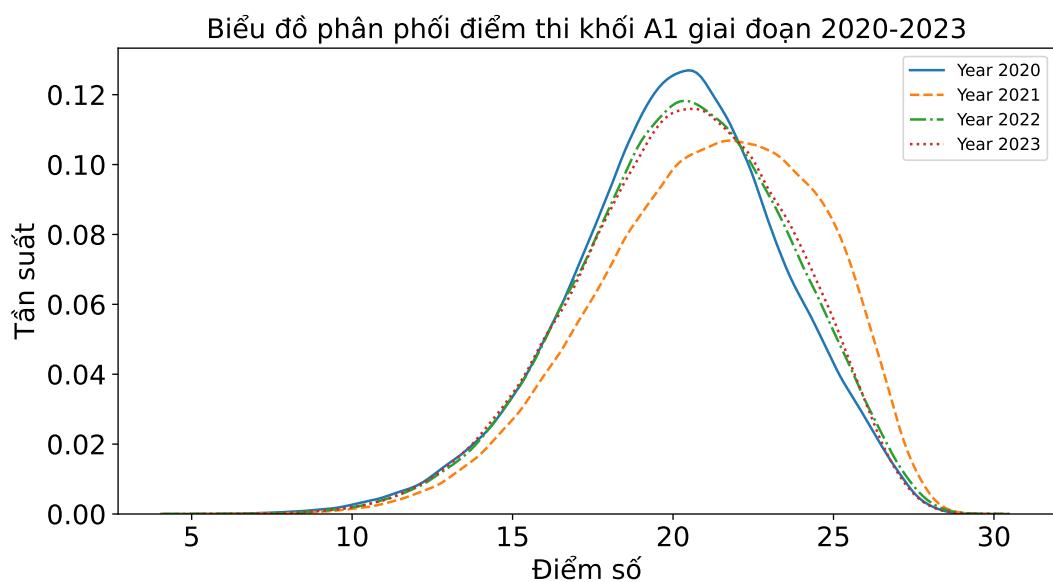
4.17 Phân phối các khối xét tuyển đại học (A00, A01, B, C, D) theo từng năm

```
1   '''
2   Khối A0: Toán - Lý - Hóa
3   Khối A1: Toán - Lý - Anh
4   Khối B: Toán - Hóa - Sinh
5   Khối C: Văn - Sử - Địa
6   Khối D: Toán - Anh - Văn
7   '''
8
9   # thêm các cột khối thi theo môn tương ứng
10
11 df['A0'] = df['Math']+df['Physics']+df['Chemistry']
12 df['A1'] = df['Math']+df['Physics']+df['ForeignLanguage']
13 df['B'] = df['Math']+df['Biology']+df['Chemistry']
14 df['C'] = df['Literature']+df['History']+df['Geography']
15 df['D'] = df['Math']+df['ForeignLanguage']+df['Literature']
16
17 khoi_thi = ['A0', 'A1', 'B', 'C', 'D']
18
19 grouped_data = df.groupby('Year')
20 for khoi_thi in khoi_thi:
21     plt.figure(figsize=(10, 5))
22
23     for i, (year, group) in enumerate(grouped_data):
24         sns.kdeplot(group[khoi_thi], fill=False, label=f'Year {year}', 
25                     linestyle=linestyles[i])
26
27     plt.title(f"Biểu đồ phân phối điểm thi khối {khoi_thi} giai đoạn
28             → 2020-2023", fontsize=16)
29     plt.xlabel("Điểm số", fontsize=16)
30     plt.ylabel("Tần suất", fontsize=16)
31     plt.legend()
32     plt.xticks(fontsize=16)
33     plt.yticks(fontsize=16)
34     plt.savefig(f'figs/Phân phối khối {khoi_thi}.pdf')
```



Hình 60: Phân phối các khối A0

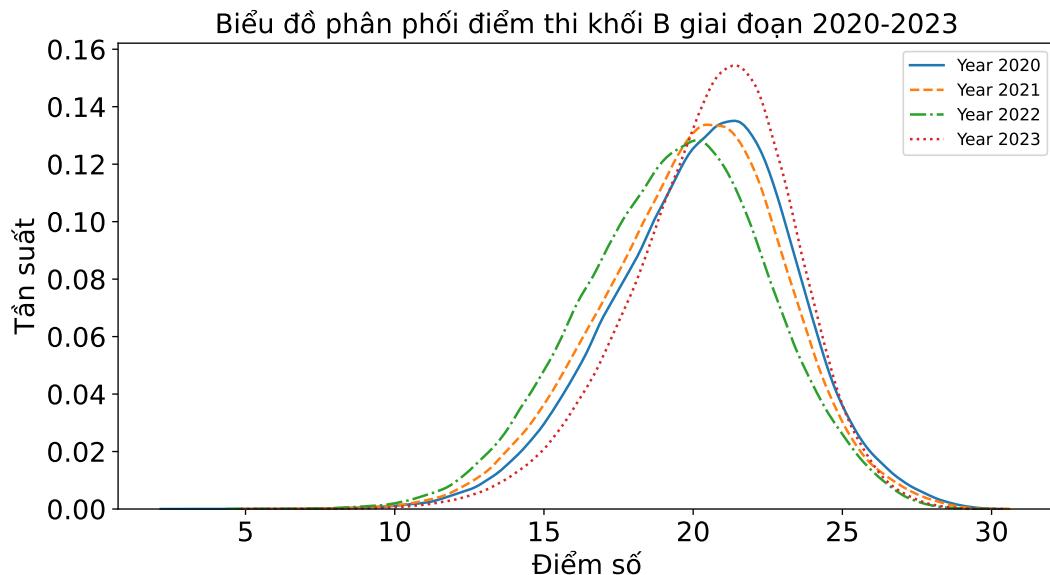
Phổ điểm của khối A0 qua các năm là khá tương đồng nhau, với điểm mode dao động trong khoảng 22.10 đến 23.20 điểm . Trong đó, điểm mode của khối A0 cao nhất là vào năm 2020, với 23.20 điểm, tiếp đến là năm 2021 và 2022, với cùng mức điểm mode là 22.75. Điểm mode của năm 2023 là thấp nhất với 22.10 điểm. Ngoài ra, điểm trung bình năm 2020 của khối A1 là cao nhất với 21.48 điểm, năm 2023 là thấp nhất với 20.78 điểm.



Hình 61: Phân phối các khối A1

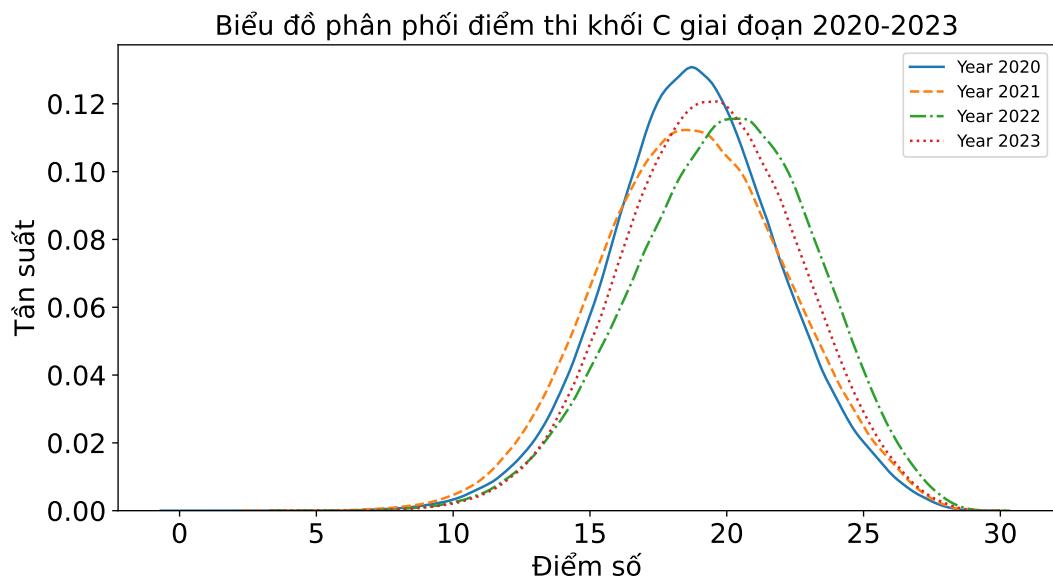
Phổ điểm trong ba năm 2020, 2021, và 2023 của tổ hợp thi khối A1 (bao gồm các môn Toán học, Vật lý, Tiếng Anh) là khá tương đồng nhau, với đỉnh dao động ở mức 20.43

điểm. Nổi bật là năm 2021, khi mức điểm mode rơi vào 22.25, cao hơn khá nhiều so với các năm trước . Mức điểm đỉnh ở các năm 2020, 2022, 2023 lần lượt là 19.50; 20.00; và 20.50. Điểm trung bình của khối thi A1 vào năm 2021 cũng là cao nhất với 21.05 điểm, thấp nhất là năm 2020 với 20.05 điểm.



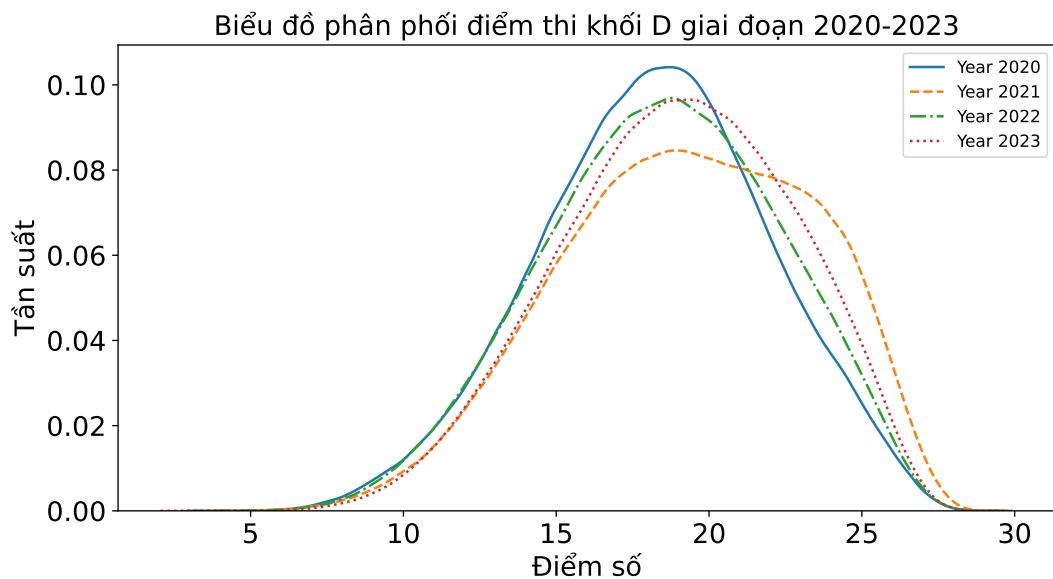
Hình 62: Phân phối các khối B

Đối với khối B (bao gồm các môn Toán học, Hóa học, Sinh học), điểm thi trung bình dao động ở mức 19.36 đến 20.59 trong giai đoạn 2020-2023 và các đỉnh dao động ở mức 20.45 đến 21.60. Điểm đỉnh và điểm trung bình vào năm 2023 là cao nhất với lần lượt rơi vào mức 21.60 và 20.59; năm 2022 là năm điểm đỉnh và điểm trung bình của khối B thấp nhất, rơi vào mức 20.45 và 19.36 điểm.



Hình 63: Phân phối các khối C

Đối với khối C, điểm trung bình của các năm dao động trong mức 18.73 đến 19.87. Trong đó, điểm trung bình lần lượt của các năm là 18.81; 18.73; 19.87; và 19.38. Điểm đỉnh của năm 2021 là thấp nhất với 18.50 và cao nhất là vào năm 2022 với 20.75 điểm. Điểm đỉnh của 2 năm 2020 và 2023 lần lượt là 18.75 và 19.75.



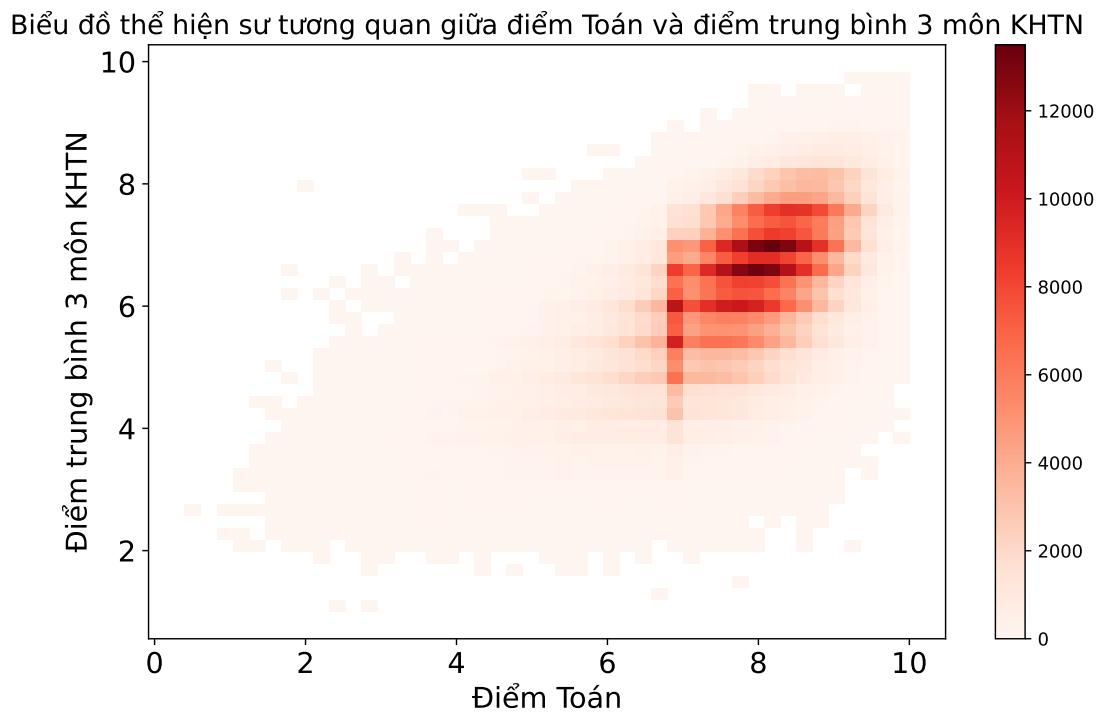
Hình 64: Phân phối các khối D

Đối với khối D, điểm trung bình dao động ở mức từ 18.16 đến 19.25. Trong đó thấp nhất là vào năm 2020 với mức điểm trung bình 18.16, cao nhất là vào năm 2021 với điểm trung bình 19.25. Điểm đỉnh ở các năm lần lượt là 18.4; 19.4; 19.0; và 19.6.

4.18 Biểu đồ thể hiện sự tương quan giữa điểm Toán và điểm trung bình 3 môn trong tổ hợp KHTN

Môn Toán được xem là một môn thuộc khối ngành khoa học tự nhiên nhưng kỳ thi THPTQG không tính môn Toán vào tổ hợp KHTN mà thay vào đó môn Toán là môn bắt buộc. Từ đó, có thể điểm môn Toán và điểm trung bình 3 môn KHTN có mối tương quan thuận, tức là, những thí sinh có điểm trung bình 3 môn tổ hợp KHTN cao cũng sẽ có điểm Toán cao hoặc những thí sinh có điểm trung bình 3 môn tổ hợp KHTN thấp cũng sẽ có điểm Toán thấp. Để kiểm chứng điều này, ta sẽ cần biểu diễn biểu đồ thể hiện sự tương quan giữa điểm Toán và điểm trung bình 3 môn tổ hợp KHTN. Với tập dữ liệu lớn, lên đến hơn 1 triệu thí sinh thi tổ hợp KHTN, việc biểu diễn sự tương quan này bằng scatter plot thông thường sẽ dẫn đến hiện tượng chồng lấp (overplotting), vì vậy ta sẽ quan sát mối tương quan này dưới dạng 2D Histogram.

```
1 df_khtn = df[df['Complex'] == 'KHTN'][['Math', 'Physics', 'Chemistry', 'Biology']]
2 df_khtn['avg_KHTN'] = df_khtn[['Physics', 'Chemistry', 'Biology']].mean(axis=1)
3
4 plt.figure(figsize=(10, 6))
5 sbn.histplot(data=df_khtn, x='Math', y='avg_KHTN', cmap='Reds', bins=45,
6   ↪ cbar=True, alpha=1)
7
8 plt.xlabel('Điểm Toán', fontsize=16)
9 plt.ylabel('Điểm trung bình 3 môn KHTN', fontsize=16)
10 plt.title('Biểu đồ thể hiện sự tương quan giữa điểm Toán và điểm trung bình 3
11   ↪ môn KHTN', fontsize=15)
12 plt.xticks(fontsize=16)
13 plt.yticks(fontsize=16)
14 plt.savefig('figs/Biểu đồ thể hiện sự tương quan giữa điểm Toán và điểm trung
15   ↪ bình 3 môn KHTN.pdf')
16 plt.show()
```



Hình 65: Biểu đồ thể hiện sự tương quan giữa điểm Toán và điểm trung bình 3 môn trong tổ hợp KHTN

Màu sắc tại mỗi ô trên biểu đồ phản ánh mật độ điểm thí sinh tại tọa độ đó. Các ô màu đậm tập trung chủ yếu ở khoảng điểm KHTN từ 4 đến 6 điểm, tương ứng với đó là điểm Toán nằm trong khoảng 6 đến 9, thể hiện phần lớn thí sinh tập trung vào nhóm điểm này.

Khi phân tích riêng phần dữ liệu chính này, có thể thấy rõ mối tương quan thuận giữa điểm Toán và điểm trung bình các môn KHTN: khi điểm KHTN tăng lên thì điểm Toán có xu hướng tăng theo.

Tuy nhiên, khi xem xét toàn bộ dữ liệu, mối tương quan thuận trên không còn rõ ràng.

4.19 Biểu đồ cột thể hiện số lượng thí sinh thi các môn ngoại ngữ khác

Kì thi THPTQG cho phép các thí sinh chọn lựa thi các ngoại ngữ khác ngoài Tiếng Anh, trong đó mỗi ngoại ngữ được kí hiệu như sau:

- N1 - Tiếng Anh
- N2 - Tiếng Nga
- N3 - Tiếng Pháp
- N4 - Tiếng Trung Quốc

- N5 - Tiếng Đức
- N6 - Tiếng Nhật
- N7 - Tiếng Hàn

Chỉ có bộ dữ liệu điểm thi năm 2023 có chưa thông tin về các mã ngoại ngữ này. Nên ta sẽ tiến hành trực quan hóa và phân tích sự lựa chọn các ngoại ngữ của các thí sinh trong năm 2023.

Đầu tiên ta tiến hành đọc lại bộ dữ liệu điểm thi của năm 2023 như sau:

```

1 df_2023 = pd.read_csv("/content/diem_thi_thptqg_2023.csv")
2 df_2023.rename(columns={'sbd': 'ID'}, inplace=True)

```

Để trực quan hóa số lượng thí sinh theo từng môn ngoại ngữ, ta sẽ dùng hàm count và tạo ra một dataframe mới tên là language như sau:

```

1 language = df_2023['ma ngoai ngu'].dropna().value_counts()
2 language = language.reset_index(name='count')

```

Sau đó ta sẽ tiến hành tính điểm trung bình của từng môn ngoại ngữ đồng thời thêm cột dữ liệu chỉ rõ môn ngoại ngữ dựa theo mã ngoại ngữ tương ứng như sau:

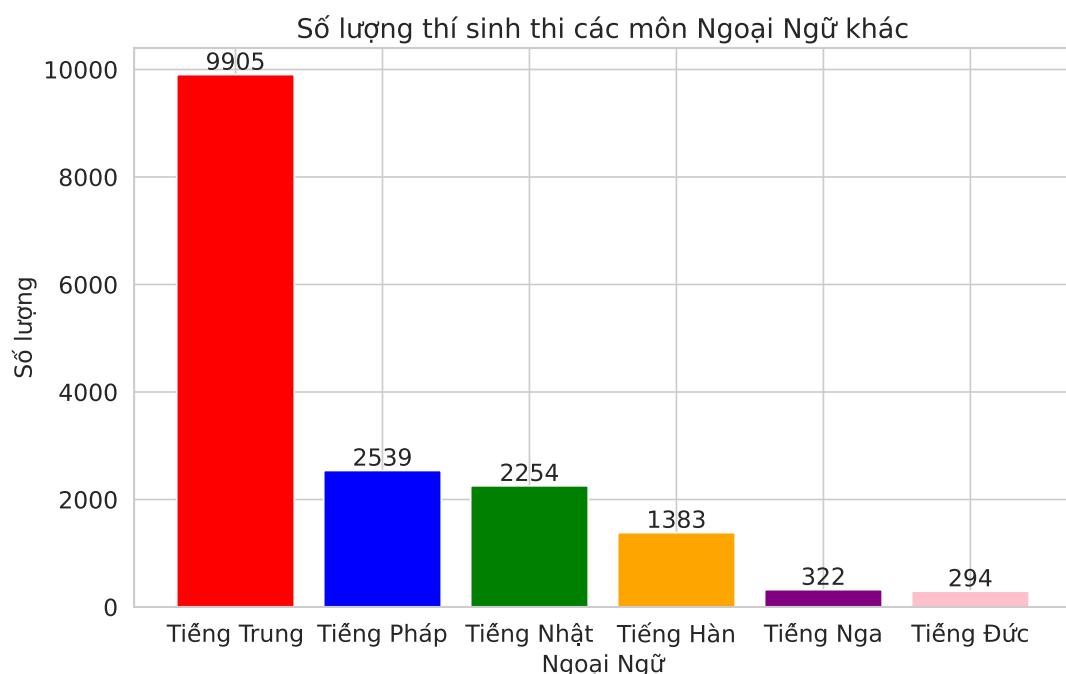
```

1 language['avg_score'] = np.nan
2 for i in range(len(language)):
3     language.loc[i, 'avg_score'] = np.mean(df_2023[df_2023['ma ngoai ngu'] ==
4         → language.loc[i, 'index']]['ngoai ngu'].dropna())
5
6 language['subject'] = np.nan
7 subjects = ['Tiếng Anh', 'Tiếng Trung', 'Tiếng Pháp', 'Tiếng Nhật', 'Tiếng Hàn',
8     → 'Tiếng Nga', 'Tiếng Đức', ]
9 for i in range(len(subjects)):
10    language.loc[i, 'subject'] = subjects[i]

```

Sau đó ta sẽ vẽ biểu đồ cột thể hiện số lượng thí sinh thi các môn ngoại ngữ khác (trừ tiếng Anh) như sau:

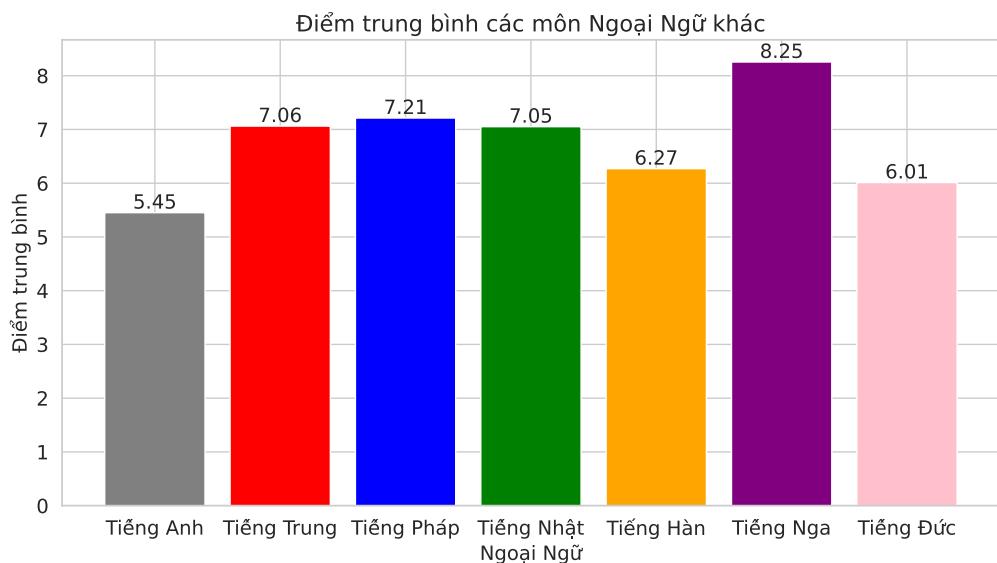
```
1 fl = language.tail(6)
2
3 plt.figure(figsize=(10,6))
4 plt.bar(fl['subject'], fl['count'], color=['red', 'blue', 'green', 'orange',
5     → 'purple', 'pink', 'gray'])
6 plt.xlabel('Ngoại Ngữ', fontsize=14)
7 plt.ylabel('Số lượng', fontsize=14)
8 plt.xticks(fontsize=14)
9 plt.yticks(fontsize=14)
10
11 for i, count in enumerate(fl['count']):
12     plt.text(i, count, str(count), ha='center', va='bottom', fontsize=14)
13
14 plt.title('Số lượng thí sinh thi các môn Ngoại Ngữ khác', fontsize=16)
15 plt.savefig('figs/f11.pdf')
16 plt.show()
```



Hình 66: Biểu đồ cột thể hiện số lượng thí sinh thi các môn ngoại ngữ khác

4.20 Biểu đồ cột thể hiện điểm trung bình các môn ngoại ngữ khác

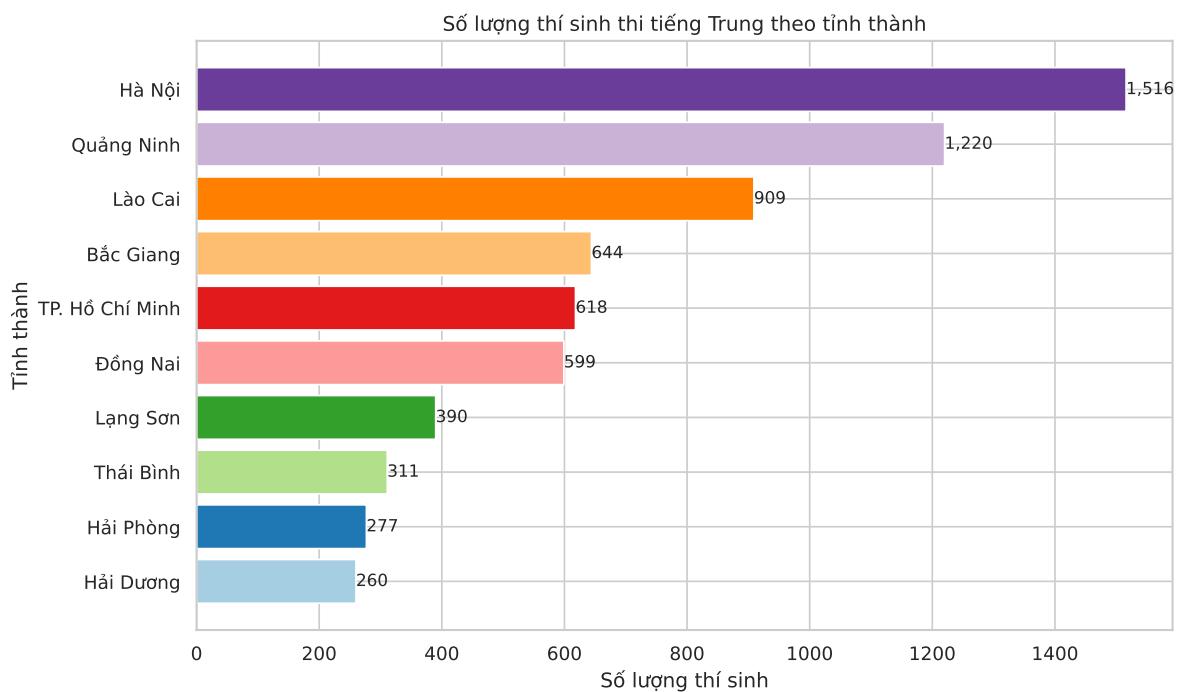
Tiếng Anh là Ngoại Ngữ có điểm thi trung bình thấp nhất vào năm 2023 với 5.45. Các môn Ngoại Ngữ khác có điểm trung bình khá cao, nằm trong khoảng từ 7-8. Trong đó cao nhất là tiếng Nga với điểm trung bình 8.25. Tiếng Đức và tiếng Hàn là hai môn Ngoại Ngữ có điểm trung bình nằm ở mức tương đối thấp, với lần lượt 6.07 và 6.27.



Hình 67: Biểu đồ cột thể hiện điểm trung bình các môn ngoại ngữ khác

4.21 Biểu đồ cột ngang thông kê số lượng thí sinh thi các môn ngoại ngữ khác theo tỉnh thành

Hà Nội là tỉnh có đông học sinh thi tiếng Trung nhất. Tiếp theo là Quảng Ninh, Lào Cai, và Bắc Giang, đây là các tỉnh ở Việt Nam có cửa khẩu Trung Quốc. Các tỉnh khác có cửa khẩu Trung Quốc nhưng không nằm trong danh sách top 10 bao gồm Cao Bằng, Hà Giang, Lai Châu, và Điện Biên. Ngoài ra TP HCM và Đồng Nai là 2 thành phố lớn và là nơi có nhiều người Hoa sinh sống.



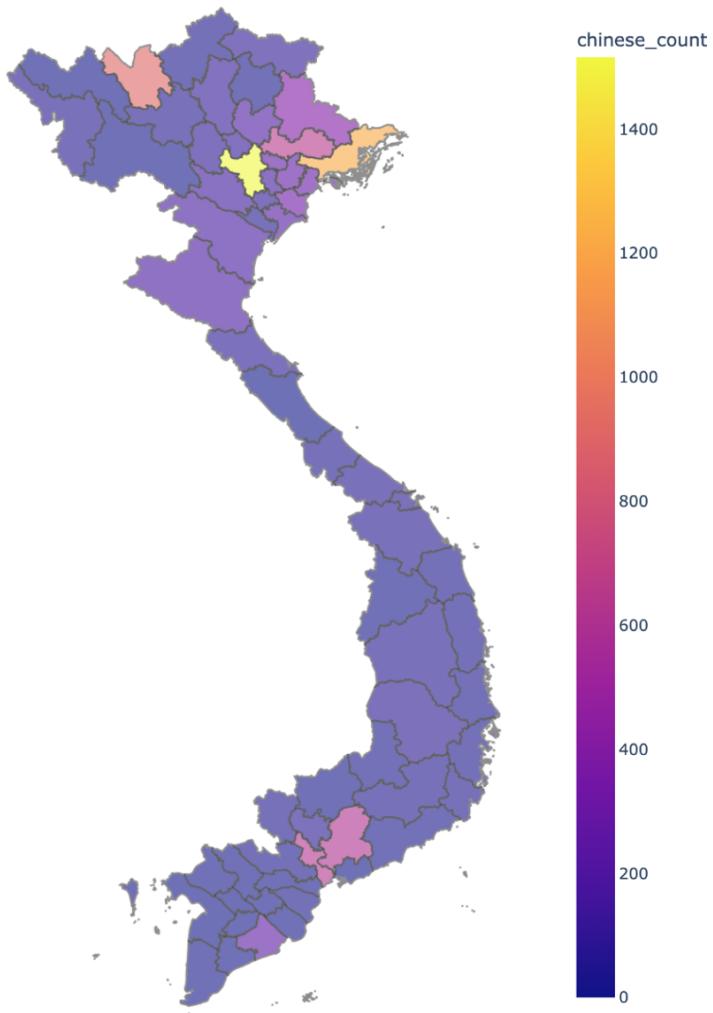
Hình 68: Biểu đồ cột thể hiện số lượng thí sinh thi các môn ngoại ngữ khác

4.22 Bản đồ Choropleth số lượng thí sinh thi tiếng Trung năm 2023

```

1 import plotly.express as px
2
3 fig = px.choropleth(merged,
4                     geojson=merged.geometry,
5                     locations=merged.index,
6                     color='chinese_count',
7                     hover_name='ten_tinh',
8                     title='Số lượng thí sinh thi tiếng Trung theo tỉnh thành
9                     → năm 2023')
10
11 fig.update_geos(fitbounds="locations", visible=False)
12 fig.update_traces(marker_opacity=0.6)
13
14 fig.show()
15 fig.write_html('figs/chinese_count_map.html')

```



4.23 Bản đồ thể hiện điểm trung bình môn Tiếng Anh theo tỉnh/thành

Qua các biểu đồ Histogram, biểu đồ thể số lượng điểm 10 và biểu đồ thể hiện số lượng điểm liệt môn Ngoại ngữ, nhóm nhận thấy có sự chênh lệch về trình độ rõ rệt giữa các thí sinh, lý do có thể đến từ việc điều kiện học ngoại ngữ tại các khu vực là khác nhau. Thực tế là học sinh nông thôn không hoặc rất ít có điều kiện học thêm với các giáo viên nước ngoài, thậm chí nhiều nơi còn không có giáo viên người Việt dạy ngoại ngữ. Vì đa số thí sinh chọn môn Tiếng Anh cho bài thi Ngoại ngữ, nhóm sẽ dùng bản đồ thể hiện điểm trung bình môn Tiếng Anh theo tỉnh/thành để kiểm tra sự khác biệt.

```

1 merged = merged.rename(columns={'ForeignLanguage': 'Điểm trung bình
2 môn Tiếng Anh'})
3
4 fig = px.choropleth(merged,
5                     geojson=merged.geometry,
6                     locations=merged.index,
7                     color='Điểm trung bình môn Tiếng Anh',

```

```

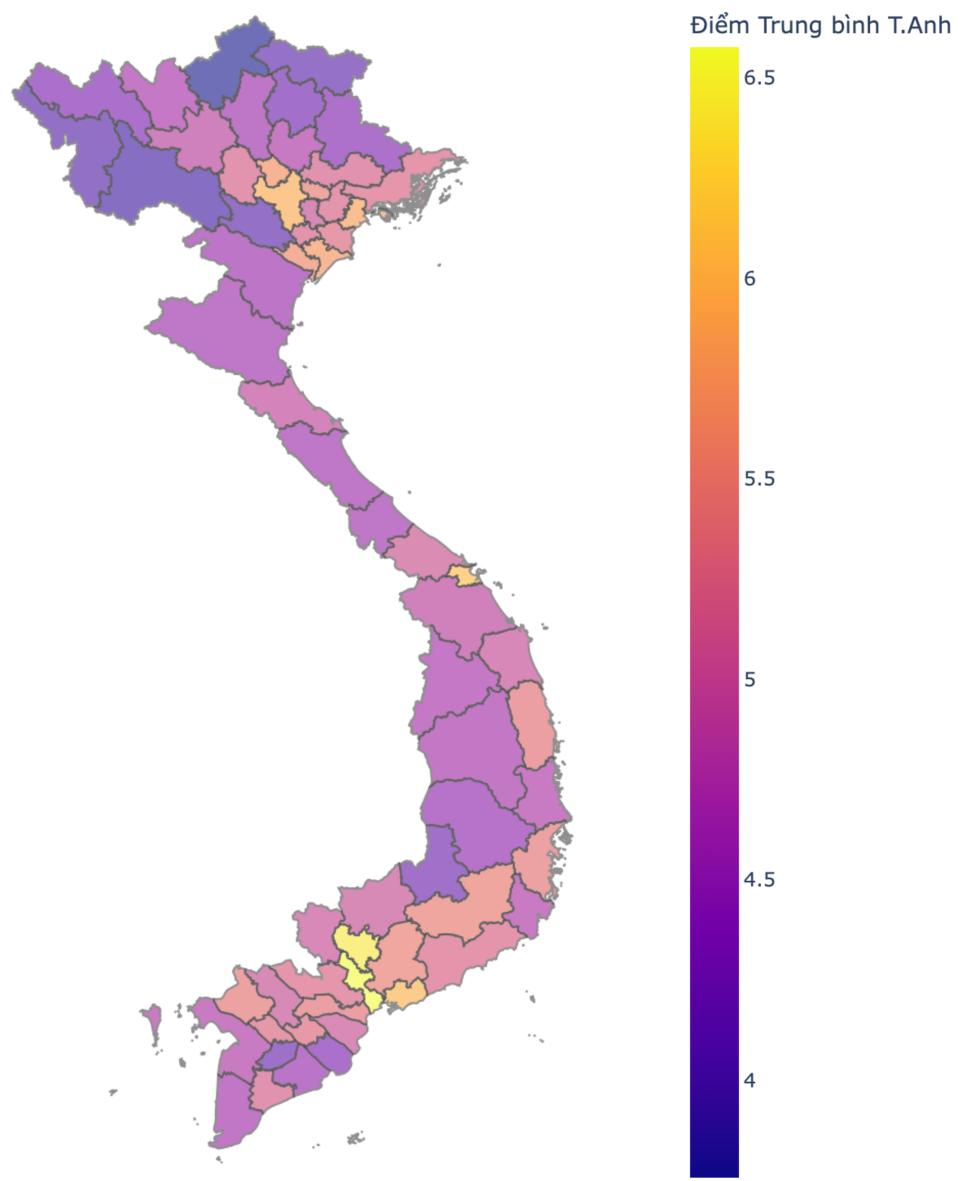
8         hover_name='ten_tinh',
9         title='Điểm trung bình môn Tiếng Anh theo tỉnh thành giai
10        đoạn 2020-2023')
11 fig.update_geos(fitbounds="locations", visible=False)
12 fig.update_traces(marker_opacity=0.6)
13
14 fig.show()
15 fig.write_html('figs/mean_eng_map.html')

```

Bản đồ cho thấy các khu vực có điểm trung bình môn Tiếng Anh cao nhất là TPHCM, Đà Nẵng, Hà Nội và các khu vực lân cận. Các khu vực có điểm Tiếng Anh thấp nhất là các tỉnh khu vực miền núi phía Bắc giáp ranh giới Trung Quốc, Lào.

Nhìn chung, ngoại trừ các thành thị và khu vực lân cận, những khu vực còn lại đều có điểm trung bình Tiếng Anh rất thấp, điều này phản ánh chân thực sự chênh lệch chất lượng dạy và học ngoại ngữ nói chung và Tiếng Anh nói riêng giữa vùng miền.

Một giải pháp có thể kể đến là luân chuyển giáo viên từ các thành phố lớn về khu vực khó khăn hơn công tác trong một thời gian nhất định. Tuy nhiên, lương giáo viên cả nước đều thấp chung, nhưng cơ hội để có thêm thu nhập từ những việc làm khác liên quan đến ngoại ngữ ở thành thị cho giáo viên cao hơn hẳn nông thôn. Vì vậy, cần thêm một giải pháp đánh vào trọng tâm, trọng điểm các vùng để nâng cao chất lượng dạy và học Ngoại ngữ.



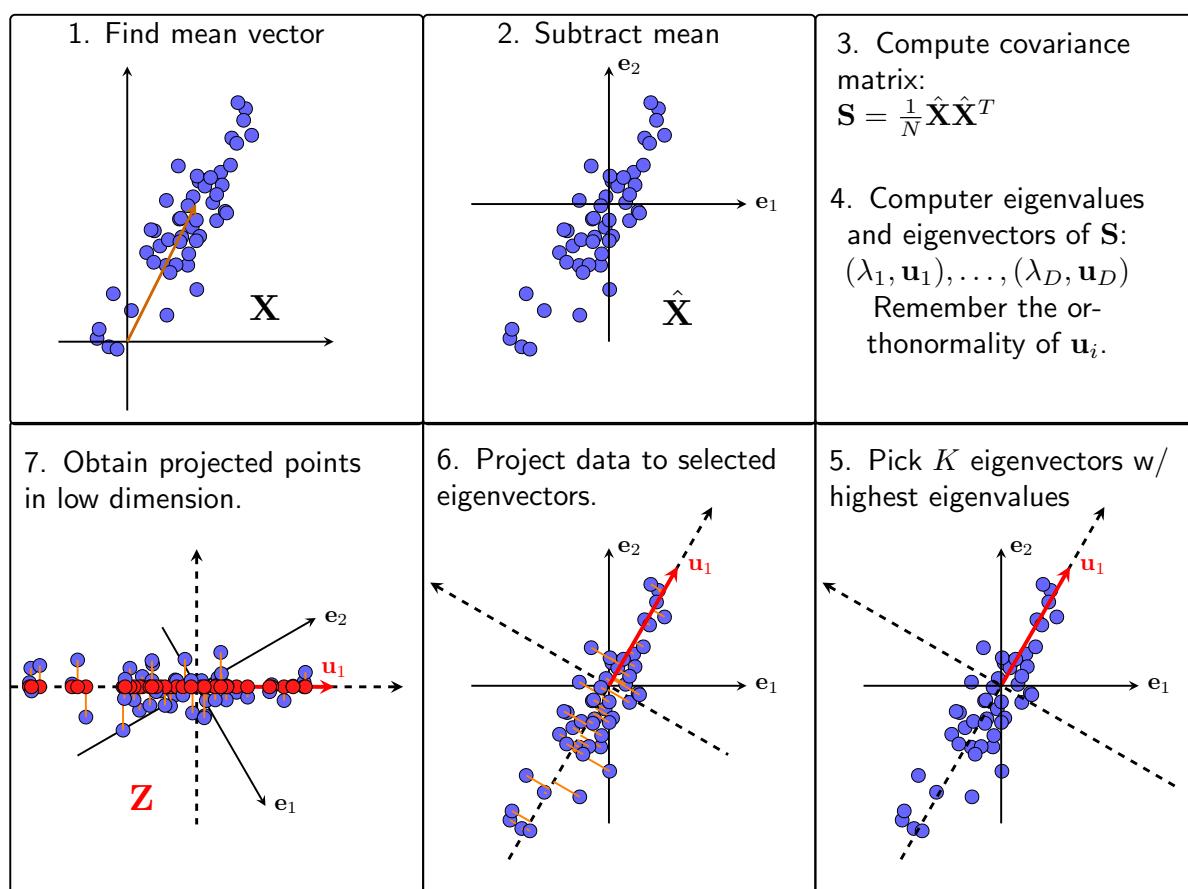
Hình 69: Bản đồ thể hiện điểm trung bình môn Tiếng Anh theo tỉnh/thành

5 CHƯƠNG 5: Giảm chiều dữ liệu bằng phương pháp phân tích thành phần chính PCA và phân cụm dữ liệu bằng KMeans

5.1 Giảm chiều dữ liệu bằng phương pháp phân tích thành phần chính PCA

Phân tích thành phần chính (Principal Component Analysis (PCA)) là một thuật toán học máy đơn giản dựa trên các kiến thức của đại số tuyến tính. PCA đóng vai trò là một kỹ thuật cơ bản trong phân tích dữ liệu và giảm chiều dữ liệu. Việc thực hiện PCA giúp chuyển đổi dữ liệu nhiều chiều sang chiều thấp hơn trong khi vẫn giữ được thông tin cần thiết.

PCA procedure



Hình 70: Quy trình thực hiện PCA. Nguồn: Machine Learning Cơ Bản

Các bước thực hiện PCA như sau:

1. Tính vector trung bình $\bar{\mathbf{x}}$
2. Tính vector $\hat{\mathbf{x}} = \mathbf{x} - \bar{\mathbf{x}}$
3. Tính $v = \text{covariance}(\hat{\mathbf{x}})$
4. Tính các trị riêng (eigenvalues) $\lambda_1, \dots, \lambda_n$ và vector riêng (eigenvectors) $\mathbf{u}_1, \dots, \mathbf{u}_n$

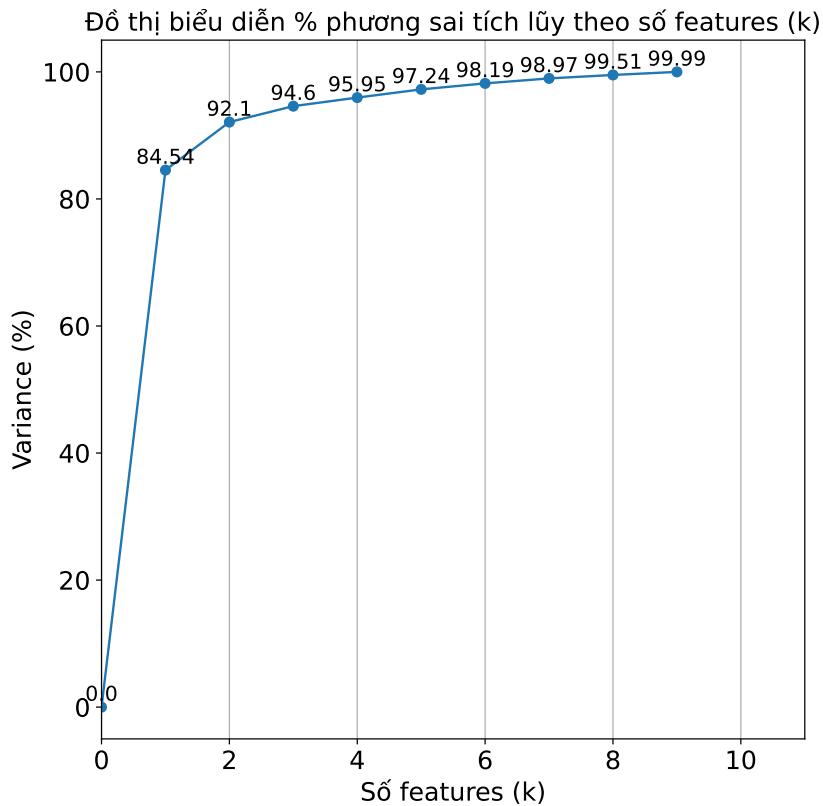
của v

5. Chọn k eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_k$ có eigenvalues lớn nhất

6. Thực hiện phép chiếu $\mathbf{x}_1, \dots, \mathbf{x}_n$ và xuông cở sở $\mathbf{u}_1, \dots, \mathbf{u}_n$

Để có thể chọn được số k features phù hợp, ta sẽ tiến hành vẽ đồ thị biểu diễn phần trăm phương sai tích lũy k như sau:

```
1 ## Vẽ đồ thị biểu diễn % phương sai tích lũy theo số features --> chọn k theo
2 → điểm "gãy"
3
4 numeric_data = df.drop(['Unnamed: 0', 'ID', 'Year', 'Province', 'Complex'],
5 → axis=1)
6 numeric_data.fillna(0, inplace=True)
7
8 pca = PCA().fit(numeric_data)
9 points = np.cumsum(pca.explained_variance_ratio_) * 100 # Các điểm dữ liệu
10 points = np.insert(points, 0, 0) # Thêm điểm k = 0, variance = 0
11 x_i = np.arange(0, numeric_data.shape[1] + 1)
12 y_i = (points[-13:])//0.01/100
13
14 plt.figure(figsize = (8, 8))
15 plt.plot(points, marker = 'o')
16 plt.xlabel('Số features (k)', fontsize = 16)
17 plt.ylabel('Variance (%)', fontsize = 16)
18 plt.title('Đồ thị biểu diễn % phương sai tích lũy theo số features (k)',
19 → fontsize = 16)
20 plt.xlim([0, numeric_data.shape[1] + 1 + 1])
21 plt.xticks(fontsize = 16)
22 plt.yticks(fontsize = 16)
23 plt.grid(axis = 'x')
24
25 for i in x_i:
26     plt.text(i, y_i[i] + 1, y_i[i], ha = 'center', va = 'baseline', fontsize =
→ 13) # tung độ của text cao hơn point 1 đơn vị
27
28 plt.savefig('explained_variance_ratio.pdf')
29 plt.show()
```



Hình 71: Đồ thị biểu diễn % phương sai tích lũy theo số features

Kể từ $k = 2$, ta nhận xét thấy số phần trăm phương sai tích lũy tăng không nhiều khi k tăng, vì vậy ra sẽ chọn $k = 2$. Tiếp theo ta sẽ trực quan hóa dữ liệu trước và sau khi thực hiện PCA để thấy được sự hiệu quả của thuật toán này:

```

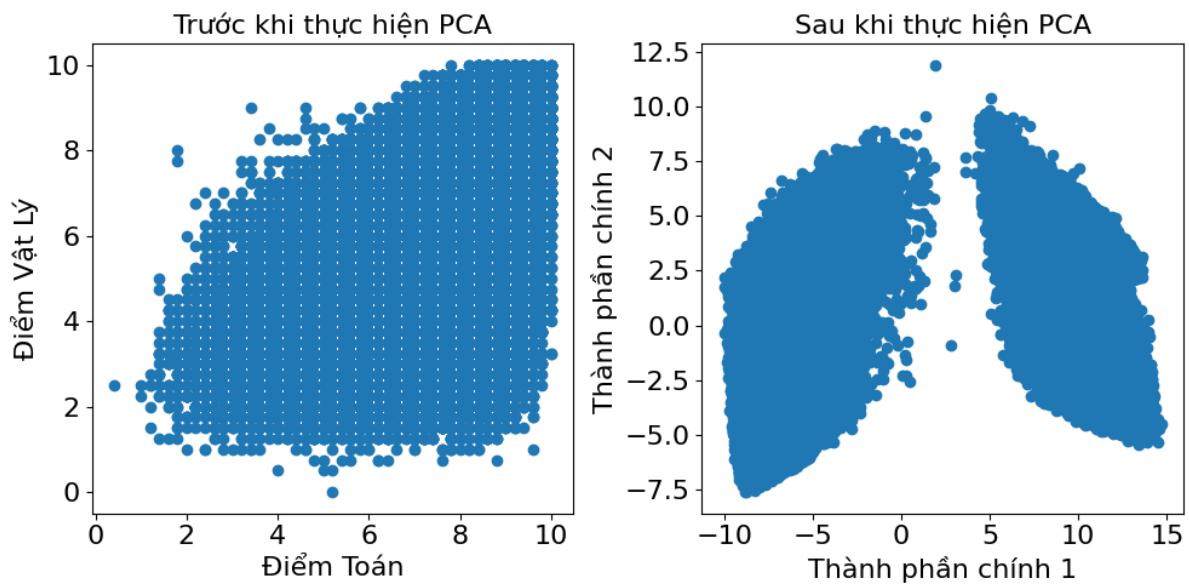
1 reduced_data = pca.fit_transform(numeric_data)
2 plt.figure(figsize=(10, 5))
3
4 plt.subplot(1, 2, 1)
5 plt.scatter(df['Math'], df['Physics'])
6 plt.title('Trước khi thực hiện PCA', fontsize = 16)
7 plt.xlabel('Điểm Toán', fontsize = 16)
8 plt.ylabel('Điểm Vật Lý', fontsize = 16)
9 plt.xticks(fontsize = 16)
10 plt.yticks(fontsize = 16)
11
12 # After PCA
13 plt.subplot(1, 2, 2)
14 plt.scatter(reduced_data[:, 0], reduced_data[:, 1])
15 plt.title('Sau khi thực hiện PCA', fontsize = 16)
16 plt.xlabel('Thành phần chính 1', fontsize = 16)

```

```

17 plt.ylabel('Thành phần chính 2', fontsize = 16)
18 plt.xticks(fontsize = 16)
19 plt.yticks(fontsize = 16)
20
21 plt.tight_layout()
22 plt.savefig('pca.png')
23 plt.show()

```



Hình 72: Dữ liệu trước và sau khi thực hiện giảm chiều bằng phương pháp PCA.

Có thể thấy sau khi thực hiện PCA dữ liệu đã phân thành hai cụm khá tách bạch nhau.

5.2 Phân cụm dữ liệu bằng thuật toán KMeans

Thuật toán Kmeans là một phương pháp học không giám sát nền tảng trong học máy được sử dụng để phân cụm dữ liệu thành các nhóm k dựa trên sự tương đồng của dữ liệu. Quá trình này diễn ra trong một số bước chính.

Ban đầu, thuật toán Kmeans sẽ tính toán các trọng tâm của dữ liệu. Trong đó, xét trong không gian n chiều, gồm n cột. Trọng tâm (về mặt biểu diễn) là một điểm ảo, là một vector n chiều. Thông thường, các trọng tâm tùy ý được chọn để bắt đầu quá trình.

Tiếp theo, thuật toán tính toán khoảng cách giữa mỗi điểm dữ liệu và hai trọng tâm. Các tính toán này giúp xác định trọng tâm nào mà mỗi điểm dữ liệu gần nhất, gán mỗi điểm cho màu cụm tương ứng dựa trên khoảng cách với trọng tâm.

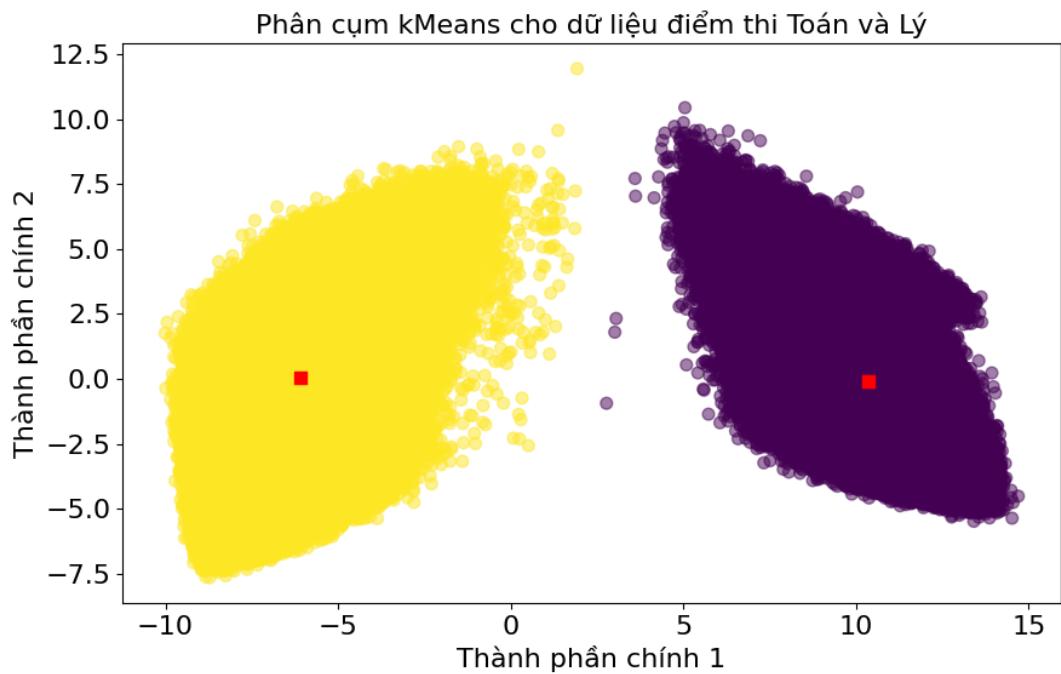
Việc gán này dẫn đến màu sắc của các điểm dữ liệu, thể hiện trực quan mối liên hệ cụm

của chúng. Ở đây bắt đầu quá trình lặp, trong đó thuật toán liên tục tính toán lại tâm và gán lại các điểm dữ liệu cho tâm gần nhất. Vòng lặp này tiếp tục cho đến khi tâm ổn định và thuật toán hội tụ, có nghĩa là không còn sự thay đổi nào trong tâm giữa các lần lặp.

Đối với bộ dữ liệu điểm thi THPTQG, sau khi đã giảm chiều dữ liệu bằng phương pháp PCA, ta sẽ thực hiện phân cụm dữ liệu bằng thuật toán Kmeans như sau:

```
1 from sklearn.cluster import KMeans
2
3 k = 2
4
5 kmeans = KMeans(n_clusters=k, random_state=42)
6 clusters = kmeans.fit_predict(reduced_data[:])
7 centroids = kmeans.cluster_centers_
8
9 df['Cluster'] = clusters
10 df['cluster'] = kmeans.labels_
11
12 print(df['Cluster'].value_counts())
13
14 plt.scatter(reduced_data[:, 0], reduced_data[:, 1], c=clusters, cmap='viridis',
15             s=50, alpha=0.5)
15 plt.scatter(centroids[:, 0], centroids[:, 1], marker="s", s=50, color='r') # 
16             ← Plot centroids for all clusters
16 plt.title('Phân cụm kMeans cho dữ liệu điểm thi Toán và Lý', fontsize=16)
17 plt.xlabel('Thành phần chính 1', fontsize = 16)
18 plt.ylabel('Thành phần chính 2', fontsize = 16)
19 plt.xticks(fontsize=16)
20 plt.yticks(fontsize=16)
21 plt.savefig('kmeans.png')
22 plt.show()
```

Kết quả phân cụm được trực quan hóa như ở [Hình 73](#). Dù có vài điểm dữ liệu nhiễu nhưng một cách tổng quan, thuật toán KMeans đang phân cụm khá tốt.



Hình 73: Dữ liệu đã được phân cụm sử dụng phương pháp Kmeans

Tuy nhiên, thuật toán Kmeans không phải là không có giới hạn. Một nhược điểm chính là sự cần thiết phải xác định trước số lượng cụm k . Yêu cầu này có thể là thách thức vì giá trị k lý tưởng thường không được biết trước. Các kỹ thuật như biểu diễn trực quan hoặc Grid Search trên một loạt các giá trị k (ví dụ từ 2 đến 100) có thể hỗ trợ xác định số lượng cụm tối ưu, coi k là một siêu tham số.

Hơn nữa, thuật toán Kmeans chỉ hiệu quả đối với các tập dữ liệu lồi (convex). Ngoài ra, nhược điểm khác của thuật toán Kmeans là mỗi một điểm dữ liệu chỉ được phép thuộc vào một cụm. Trong thực tế, các điểm dữ liệu có thể thuộc về nhiều cụm, khiến khía cạnh này bị hạn chế.

Một vài thuật toán được sử dụng để cải tiến thuật toán Kmeans bao gồm Fuzzy C-means và Hierarchical Agglomerative Clustering (HAC).

Mặc dù có những hạn chế nêu trên, Kmeans vẫn là một thuật toán được sử dụng rộng rãi do tính đơn giản và hiệu quả của nó. Vì là một thuật toán thuộc lớp học không giám sát, việc diễn giải ý nghĩa của các cụm cần kiến thức của các chuyên gia, như các nhà hoạch định chính sách giáo dục và các nhà giáo dục học.

6 CHƯƠNG 6: Kết luận

Với đề tài “Biểu diễn trực quan dữ liệu”, nhóm đã sử dụng bộ Điểm thi tốt nghiệp THPT quốc gia giai đoạn 2020 - 2023 để tiến hành biểu diễn trực quan các dữ liệu và phát triển đề tài theo hướng trực quan hoá hơn, dễ tiếp nhận thông tin hơn. Nhóm đã sử dụng đa dạng các loại biểu đồ cũng như ứng dụng các nguyên tắc về thiết kế màu sắc để biểu diễn, phù hợp với từng mục tiêu khác nhau trong việc phân tích ý nghĩa của dữ liệu. Bên cạnh những chỉ số cơ bản về điểm thi, đồ án còn phân tích sự chênh lệch giữa các môn học, các bài thi tổ hợp, các tỉnh thành và khu vực.

Thông qua việc trực quan hoá dữ liệu, đồ án đã cung cấp cái nhìn tổng quan và chi tiết hơn về những xu hướng, kết quả thi tốt nghiệp THPT quốc gia của các thí sinh trong giai đoạn 2020 - 2023. Từ đó tiến hành phân tích, so sánh và đưa ra những nhận định sâu sắc về tình hình giáo dục tại Việt Nam. Điều này có thể đóng vai trò quan trọng trong quá trình đưa ra các quyết định và đề xuất cải thiện chất lượng giáo dục ở cấp độ quốc gia.

Tuy nhiên, đồ án vẫn tồn tại một số hạn chế. Trong quá trình tiền xử lý, nhóm phải chấp nhận một rủi ro nhất định để bỏ đi các dòng dữ liệu bị thiếu. Vì không có đủ dữ kiện để phân biệt giữa nhóm thí sinh tự do, thí sinh thuộc chương trình GDTX với nhóm thí sinh không tham gia đủ bài thi đã đăng ký để giữ lại những thí sinh hợp lệ. Việc loại bỏ này dẫn đến một đánh đổi rằng những đánh giá của nhóm đã bỏ qua thành phần các đối tượng đủ điều kiện xét tốt nghiệp (thí sinh hệ GDTX) và các đối thủ cạnh tranh tiềm năng trong việc xét tuyển đại học (thí sinh tự do).

Bên cạnh đó bộ dữ liệu cũng không thỏa một số điều kiện về tính độc lập của mẫu nên nhóm không tiến hành kiểm định thống kê. Do đó một số giả thuyết của nhóm đưa ra vẫn chưa thể kiểm chứng.

BẢNG PHÂN CÔNG

Thành viên	Phân công	Đánh giá
Vũ Nguyễn Thảo Vi	Kiểm tra các đại lượng về xu thế trung tâm Kiểm tra đại lượng về sự tương quan Biểu diễn trực quan dữ liệu và phân tích	100%
Nguyễn Quốc Việt	Kiểm tra các đại lượng về hình dáng phân phối Biểu diễn trực quan dữ liệu và phân tích PCA và phân cụm	100%
Nguyễn Thanh Vy	Kiểm tra các đại lượng về độ phân tán Biểu diễn trực quan dữ liệu và phân tích Kết luận	100%
Nguyễn Nhật Thảo Vy	Tổng quan Tiền xử lý dữ liệu Biểu diễn trực quan dữ liệu và phân tích Kết luận	100%

Toàn bộ source code và bộ dữ liệu được đăng tải tại <https://github.com/quocviethere/UEH-Data-Visualization-Final-Project>.

TÀI LIỆU THAM KHẢO

1. Quyên, Đ. (n.d.). Nguyên nhân khiến điểm thi môn Lịch sử liên tiếp “đội sổ”. Retrieved from <https://kinhtedothi.vn/5-nguyen-nhan-khien-diem-thi-mon-lich-su-lien-tiep-doi-so.html>
2. Sang, D. (n.d.). Hệ thống thông tin thống kê khoa học và công nghệ. Retrieved from <http://thongke.cesti.gov.vn/dich-vu-thong-ke/tai-lieu-phan-tich-thong-ke/861-thong-ke-mo-ta-trong-nghien-cuu-dai-luong-tuong-quan>
3. Vĩnh Hà, H. H. (n.d.). Đề thi tốt nghiệp THPT. Retrieved from <https://tuoitre.vn/de-thi-tot-nghiep-thpt-nhe-nhang-diem-chuan-se-tang-20200811075331749.htm>