

HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.



ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Phân tích điểm thi THPT Quốc gia 2020 sử dụng Apache Spark

Sinh viên: Lê Huỳnh Đức

MSSV: 20205067

Giảng viên hướng dẫn: TS. Trần Việt Trung

ONE LOVE. ONE FUTURE.

Nội dung

1. Giới thiệu đề tài
2. Kiến trúc hệ thống
3. Đánh giá hệ thống
4. Phân tích dữ liệu
5. Kết luận và hướng phát triển

Nội dung

1. Giới thiệu đề tài
2. Kiến trúc hệ thống
3. Đánh giá hệ thống
4. Phân tích dữ liệu
5. Kết luận và hướng phát triển

1. Giới thiệu đề tài

Đặt vấn đề

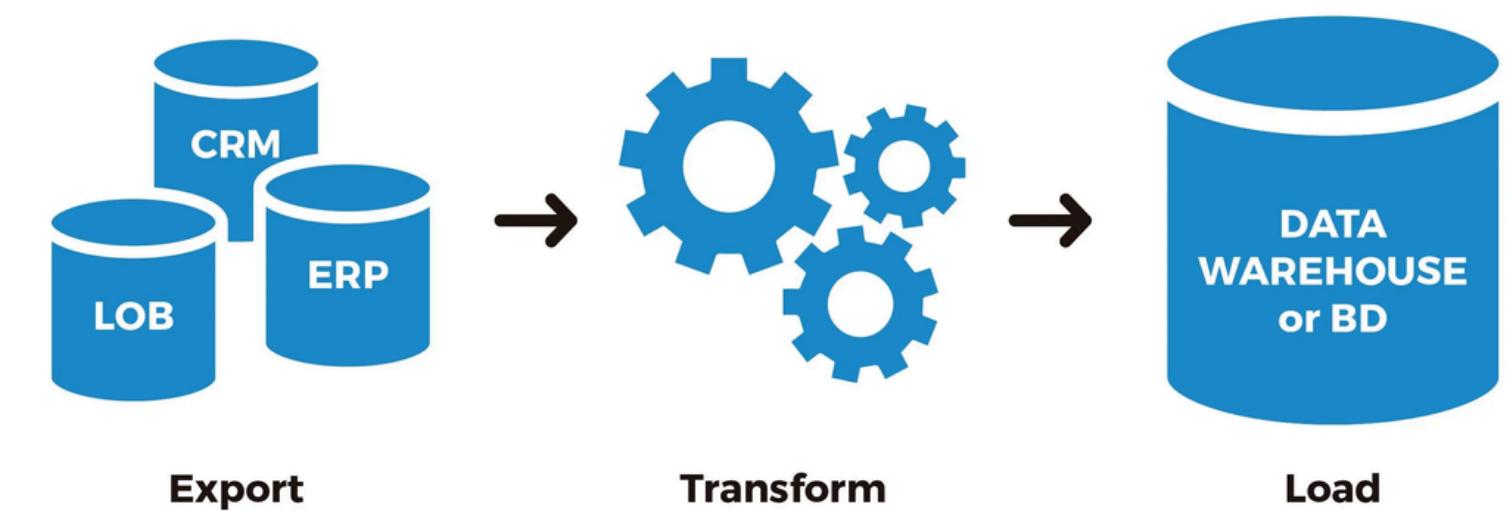
- Giáo dục là yếu tố quan trọng trong quá trình phát triển của mỗi quốc gia.
- Kỳ thi THPT Quốc gia là kỳ thi quan trọng, đánh dấu bước ngoặt mới trong cuộc đời của các bạn học sinh
=> Cần có những đánh giá để nâng cao chất lượng dạy - học để các bạn học sinh có chuẩn bị tốt nhất cho kỳ thi THPT Quốc gia.



1. Giới thiệu đề tài

Mục tiêu và phạm vi đề tài

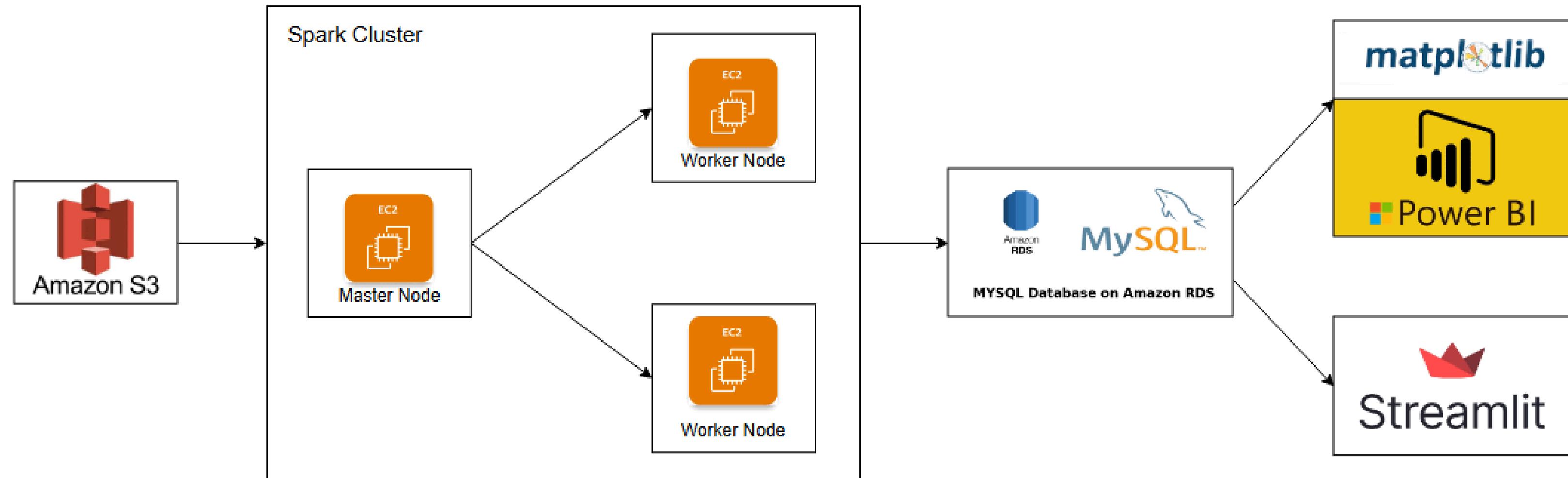
Thực hiện quy trình ETL dữ liệu điểm thi THPT Quốc gia năm 2020 để rút ra những phát hiện về kỳ thi, cũng như các yếu tố ảnh hưởng đến kết quả học tập của học sinh.



Nội dung

1. Giới thiệu đề tài
2. Kiến trúc hệ thống
3. Đánh giá hệ thống
4. Phân tích dữ liệu
5. Kết luận và hướng phát triển

2. Kiến trúc hệ thống



Kiến trúc hệ thống

2. Kiến trúc hệ thống

Thành phần Amazon S3



File dữ liệu đầu vào là file text, các bản ghi trong file có dạng HTML

File có 900,000 bản ghi, kích thước 2.3GB

2. Kiến trúc hệ thống

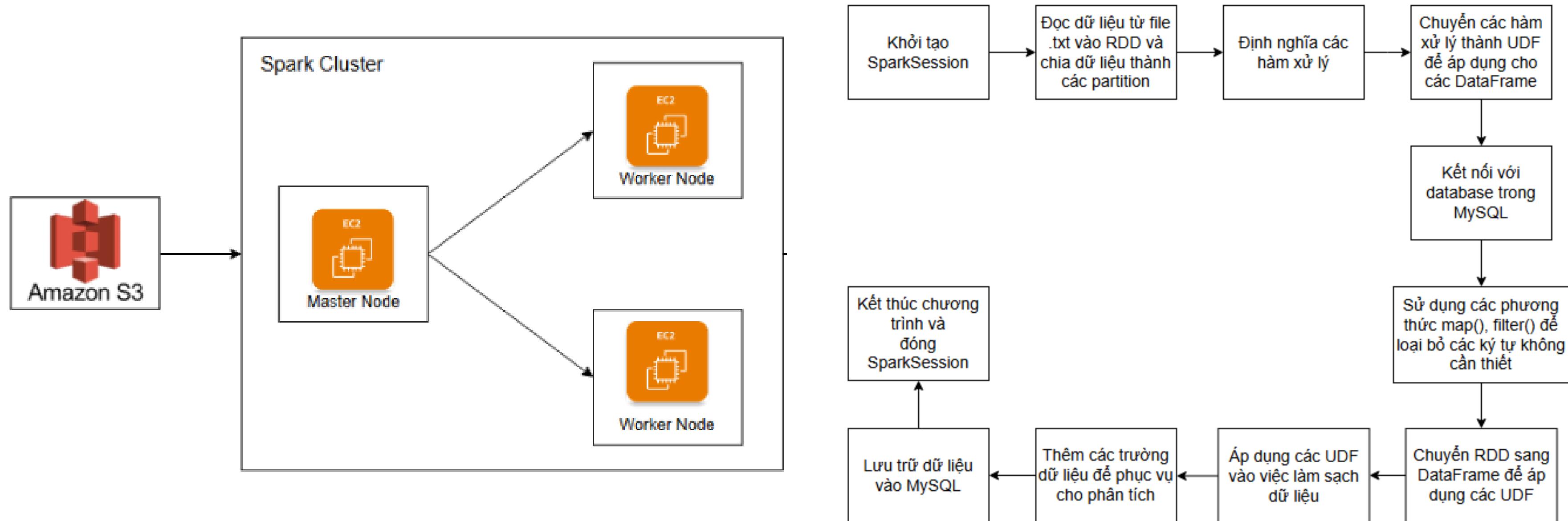
```
<table style="margin-top:50px; border: 1px solid; width:100%">\r
<tr>\r
    <td style="border: 1px solid ; font-weight: bold ">\r
        H\xe1\xbb\x8d v\xc3\xa0 T\xc3\xaan\r
    </td>\r
    <td style="border: 1px solid ; font-weight: bold">\r
        Ng\xc3\xa0y sinh\r
    </td>\r
    <td style="border: 1px solid ; font-weight: bold">\r
        \xc4\x90\xe1\xbb\x83m thi\r
    </td>\r
\r
</tr>\r
\r
<tr>\r
    <td style="border: 1px solid">\r
        PH\xe1\xba\x0M HO&#192;NG H\xc6\xaf\xc6\xa0NG &#193;|\r
    </td>\r
    <td style="border: 1px solid">\r
        04/11/2002\r
    </td>\r
    <td style="border: 1px solid">\r
        To&#225;n: 6.60 Ng\xe1\xbb\xaf v\xc4\x83n: 6.25 L\xe1\xbb\x8bch s\xe1\xbb\xad: 5.75 \xc4\x90\xe1\xbb\x8ba l&#237;: 7.00
GDCD: 7.25 KHXH: 6.67 T\xe1\xba\xbfng Anh: 5.20 \r
    </td>\r
\r
</tr>\r
\r
</table>\r
\r
<br />\r
\r
<br />
</div>

</body>
</html>
```



2. Kiến trúc hệ thống

Thành phần Amazon EC2



2. Kiến trúc hệ thống

Thành phần MySQL



Tên trường	Kiểu dữ liệu	Mô tả
sbd	INT	Số báo danh của thí sinh
name	VARCHAR(255)	Họ và tên thí sinh
dob	DATE	Ngày tháng năm sinh của thí sinh
toan	FLOAT	Điểm thi môn Toán của thí sinh
ngu_van	FLOAT	Điểm thi môn Ngữ văn của thí sinh
tieng_anh	FLOAT	Điểm thi môn Tiếng Anh của thí sinh
lich_su	FLOAT	Điểm thi môn Lịch sử của thí sinh
dia_ly	FLOAT	Điểm thi môn Địa lý của thí sinh
gdcd	FLOAT	Điểm thi môn GDCD của thí sinh
khxh	FLOAT	Điểm trung bình tổ hợp KHXH của thí sinh
sinh_hoc	FLOAT	Điểm thi môn Sinh học của thí sinh
vat_li	FLOAT	Điểm thi môn Vật lý của thí sinh
hoa_hoc	FLOAT	Điểm thi môn Hóa học của thí sinh
khtn	FLOAT	Điểm trung bình tổ hợp KHTN của thí sinh
tinh	VARCHAR(100)	Tỉnh thành nơi thí sinh dự thi
khu_vuc	VARCHAR(255)	Khu vực nơi thí sinh dự thi
so_bai_thi	INT	Số bài thi mà thí sinh đã làm
nam_thi	INT	Năm thí sinh dự thi
ban_khtn	INT	Thí sinh chọn ban KHTN
ban_khxh	INT	Thí sinh chọn ban KHXH

Cơ sở dữ liệu MySQL là nơi lưu trữ dữ liệu đã làm sạch



2. Kiến trúc hệ thống

Thành phần Trực quan hóa dữ liệu



Power BI

matplotlib

Các công cụ trực quan hóa dữ liệu phổ biến, cũng là công cụ em chọn để hỗ trợ cho việc phân tích dữ liệu trong đồ án này

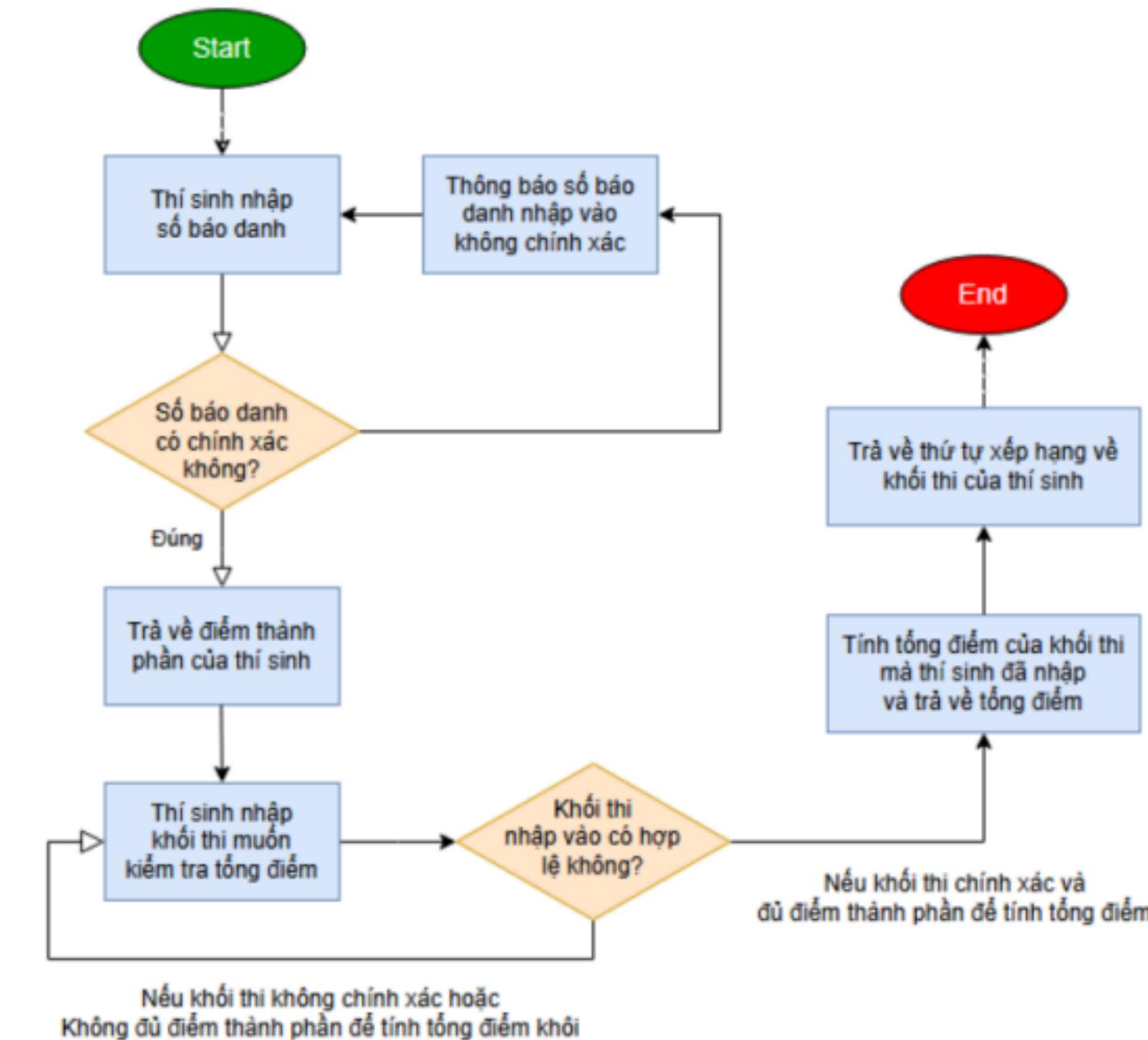


2. Kiến trúc hệ thống

Thành phần Streamlit



Streamlit



Streamlit được sử dụng để xây dựng một trang web
giúp thí sinh tra cứu điểm và thứ tự xếp hạng của mình

Nội dung

1. Giới thiệu đề tài
2. Kiến trúc hệ thống
3. Đánh giá hệ thống
4. Phân tích dữ liệu
5. Kết luận và hướng phát triển

3. Đánh giá hệ thống

Kết quả đánh giá hệ thống

Lần thử	Xử lý và lưu trữ dữ liệu với cụm Spark2	Xử lý và lưu trữ dữ liệu khi không sử dụng Spark
Lần thử thứ nhất	8 phút 12 giây	30 phút 03 giây
Lần thử thứ 2	8 phút 6 giây	30 phút 08 giây
Lần thử thứ 3	8 phút 12 giây	29 phút 57 giây
Lần thử thứ 4	8 phút 18 giây	30 phút 20 giây
Lần thử thứ 5	8 phút 12 giây	29 phút 56 giây
Lần thử thứ 6	8 phút 6 giây	30 phút 8 giây
Lần thử thứ 7	8 phút 12 giây	30 phút
Lần thử thứ 8	8 phút 12 giây	30 phút 05 giây
Lần thử thứ 9	8 phút 6 giây	30 phút 06 giây
Lần thử thứ 10	8 phút	30 phút 10 giây
Trung bình thời gian xử lý của 10 lần thử	8 phút 10 giây	30 phút 05 giây

- Tốc độ xử lý dữ liệu trung bình khi không sử dụng Spark là **30 phút 05 giây**
- Tốc độ xử lý dữ liệu trung bình với cụm Spark **nhanh hơn 268%**, với chỉ **8 phút 10 giây** xử lý.

Bảng kết quả thực nghiệm khi so sánh việc xử lý dữ liệu khi sử dụng Spark và khi không sử dụng Spark



Nội dung

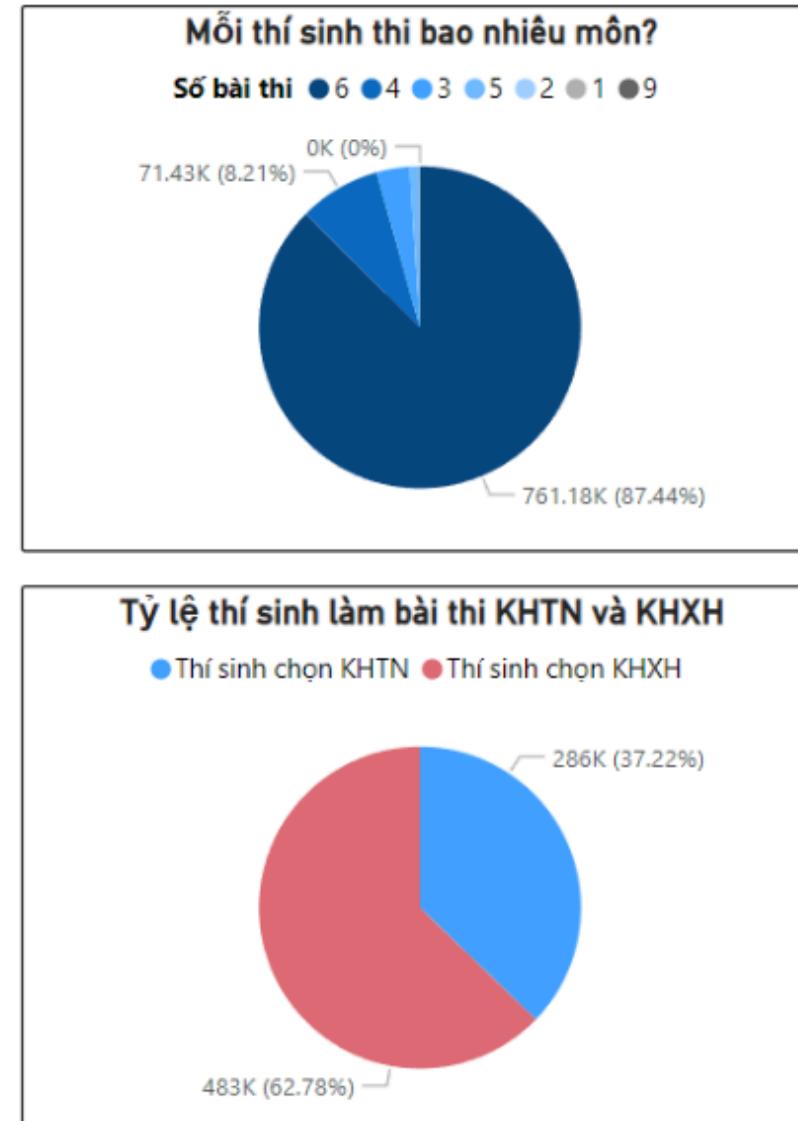
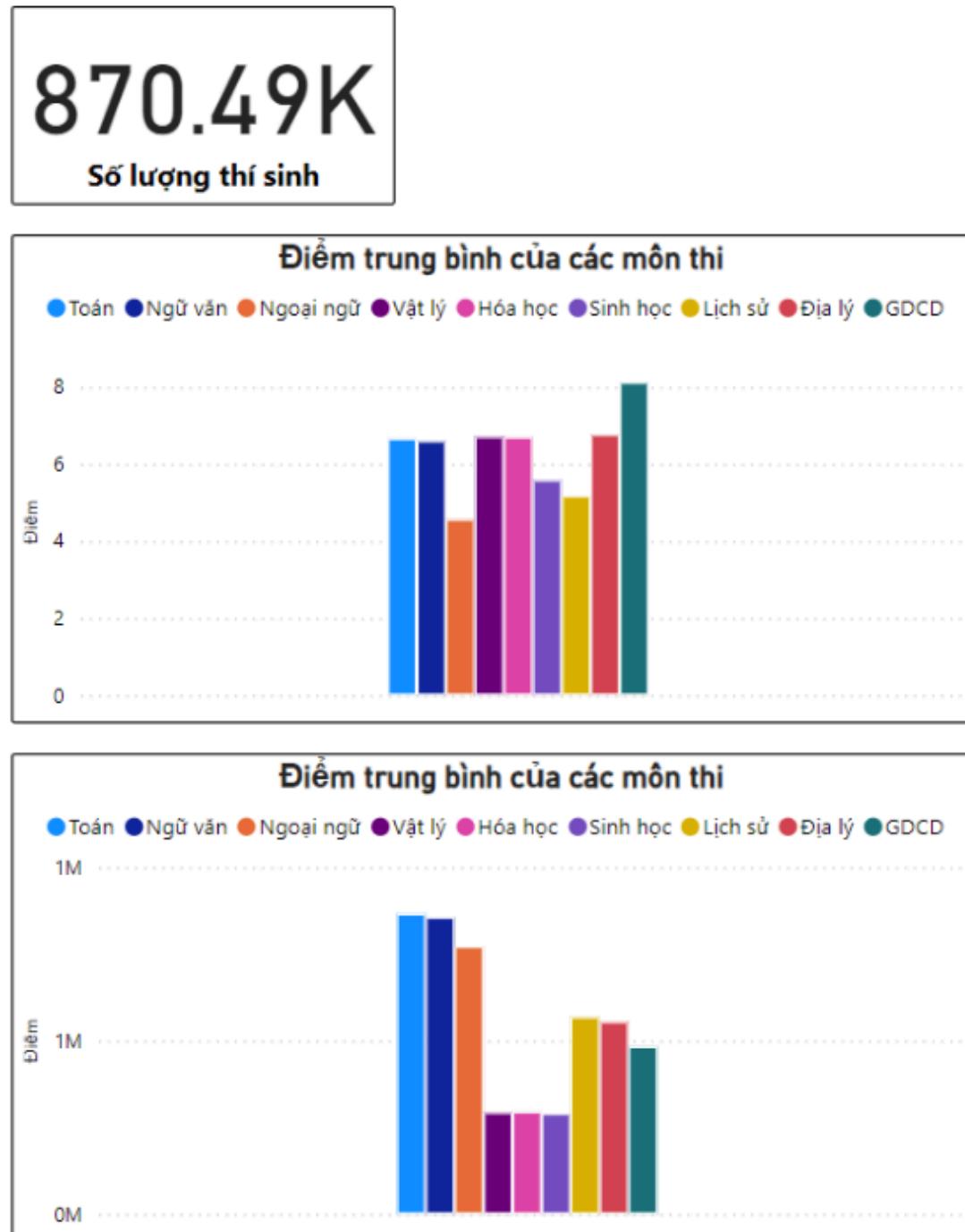
1. Giới thiệu đề tài
2. Kiến trúc hệ thống
3. Đánh giá hệ thống
4. Phân tích dữ liệu
5. Kết luận và hướng phát triển

4. Phân tích dữ liệu điểm thi

4.1. Phân tích điểm thi THPT Quốc gia 2020



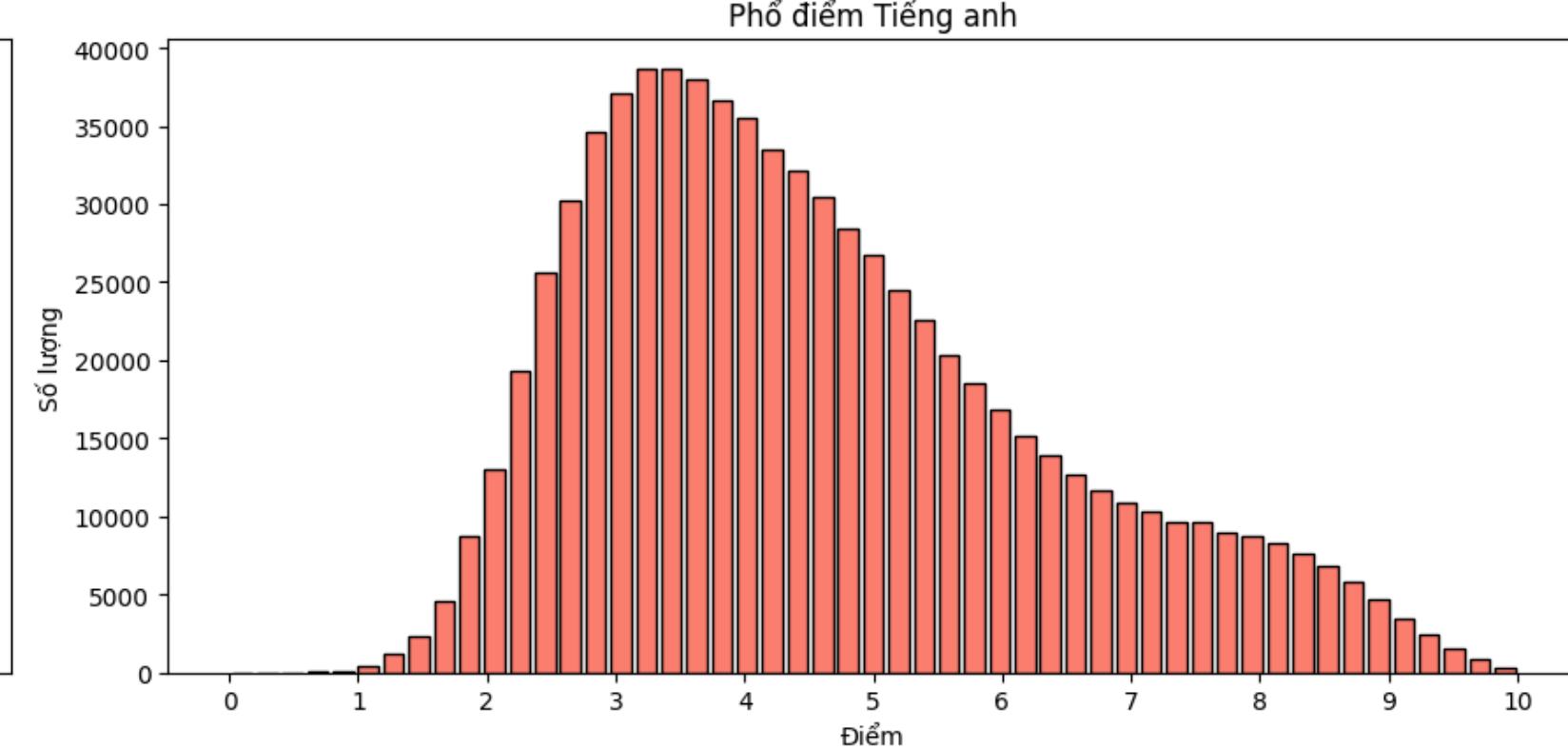
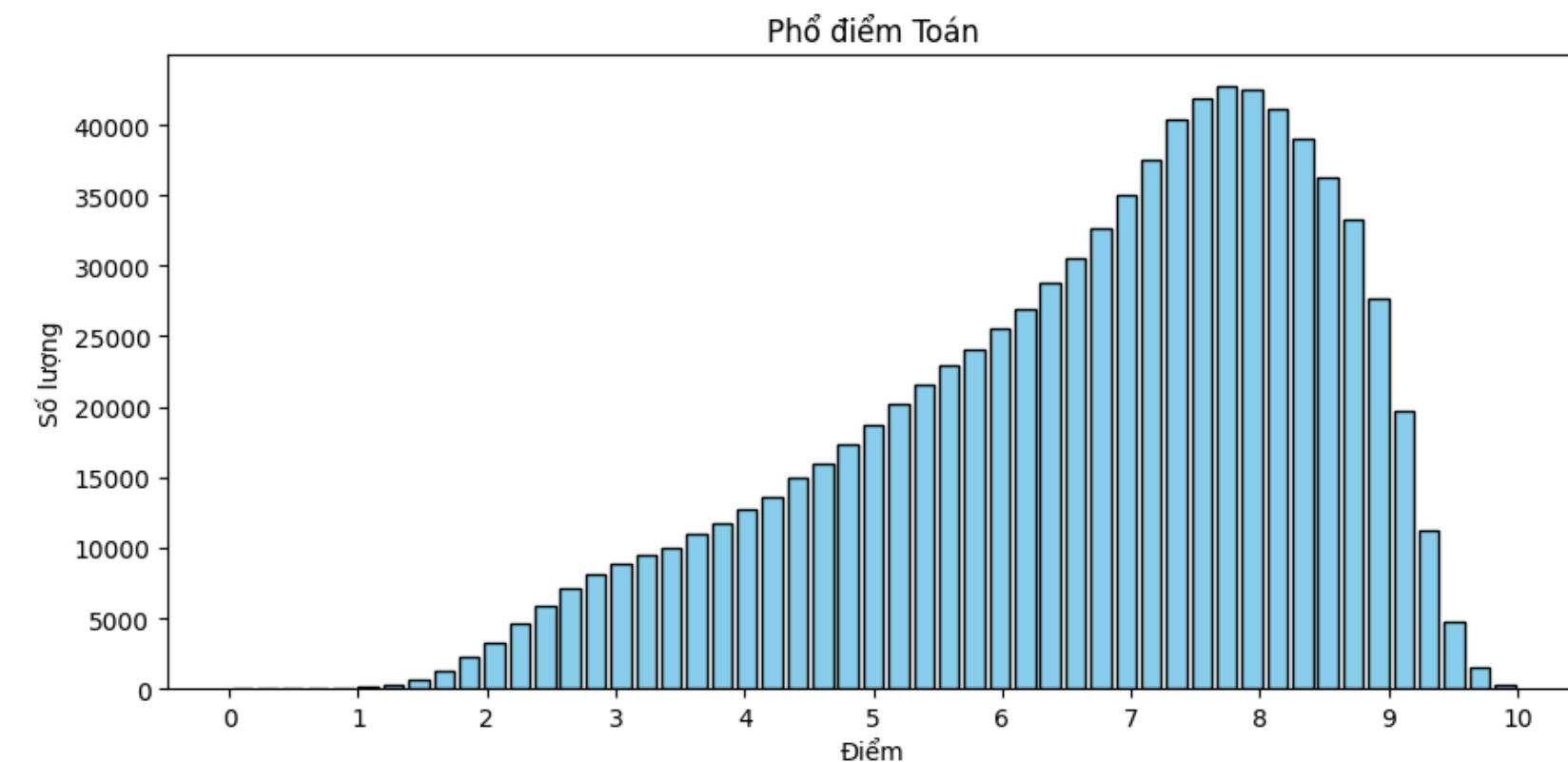
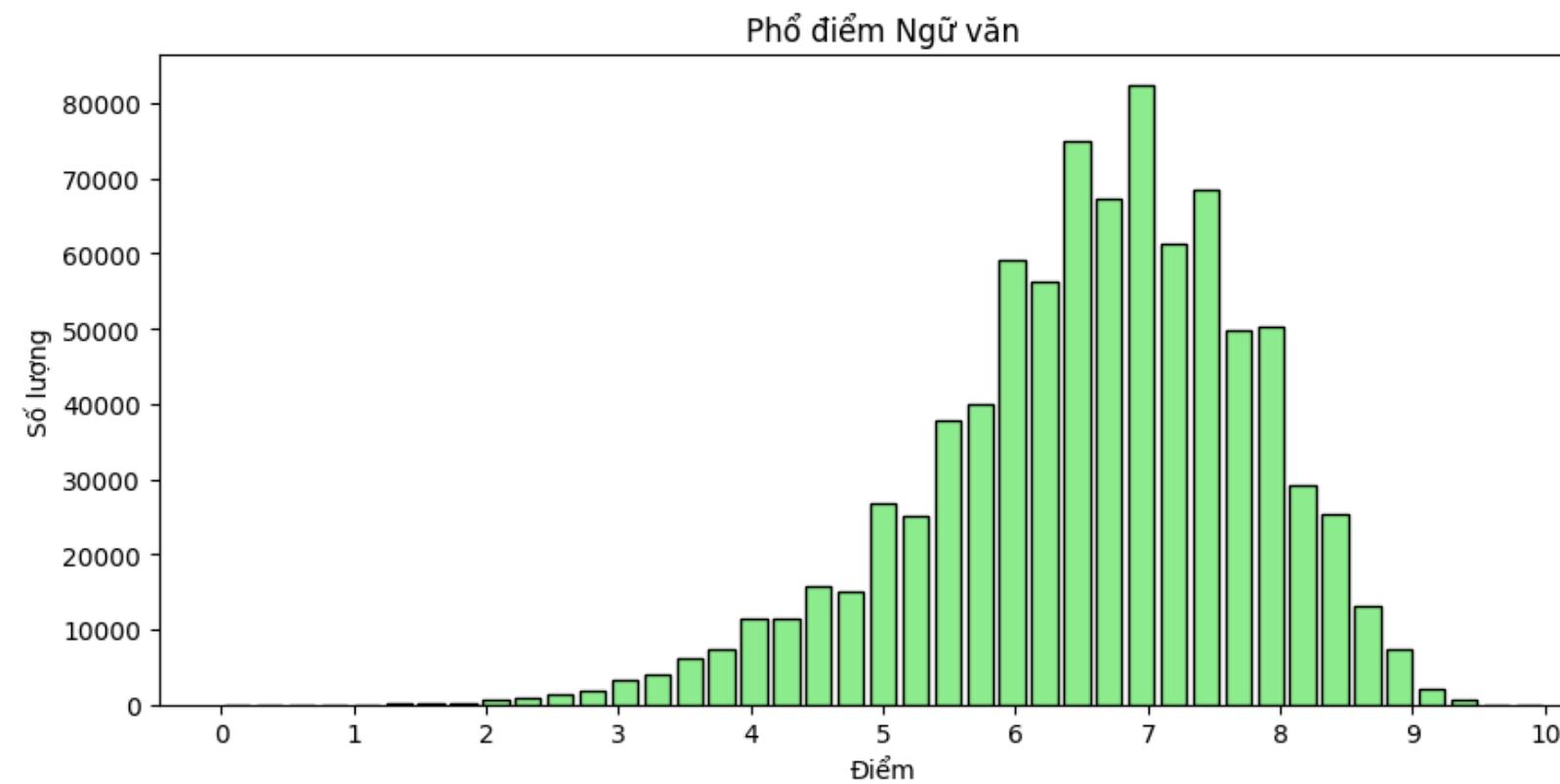
4. Phân tích dữ liệu điểm thi



- Có hơn 870 nghìn thí sinh dự thi (thiếu 30 nghìn thí sinh tại Đà Nẵng)
- Số thí sinh chọn tổ hợp KHXH nhiều gấp gần 2 lần số thí sinh chọn tổ hợp KHTN
- Phần lớn (87.44%) thí sinh tham dự bài thi với 6 môn thi. Đây chủ yếu là những thí sinh lần đầu thi THPT Quốc gia

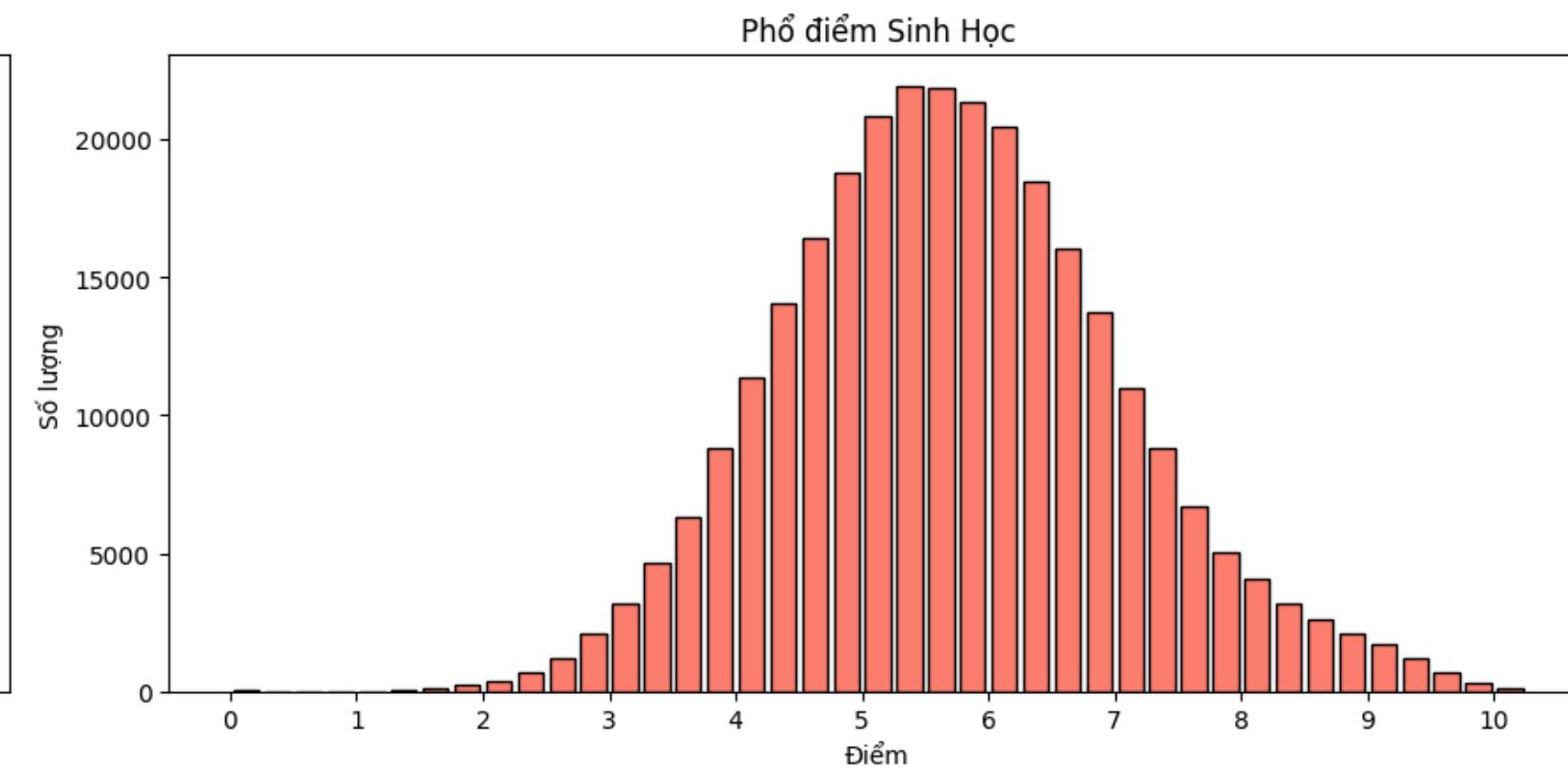
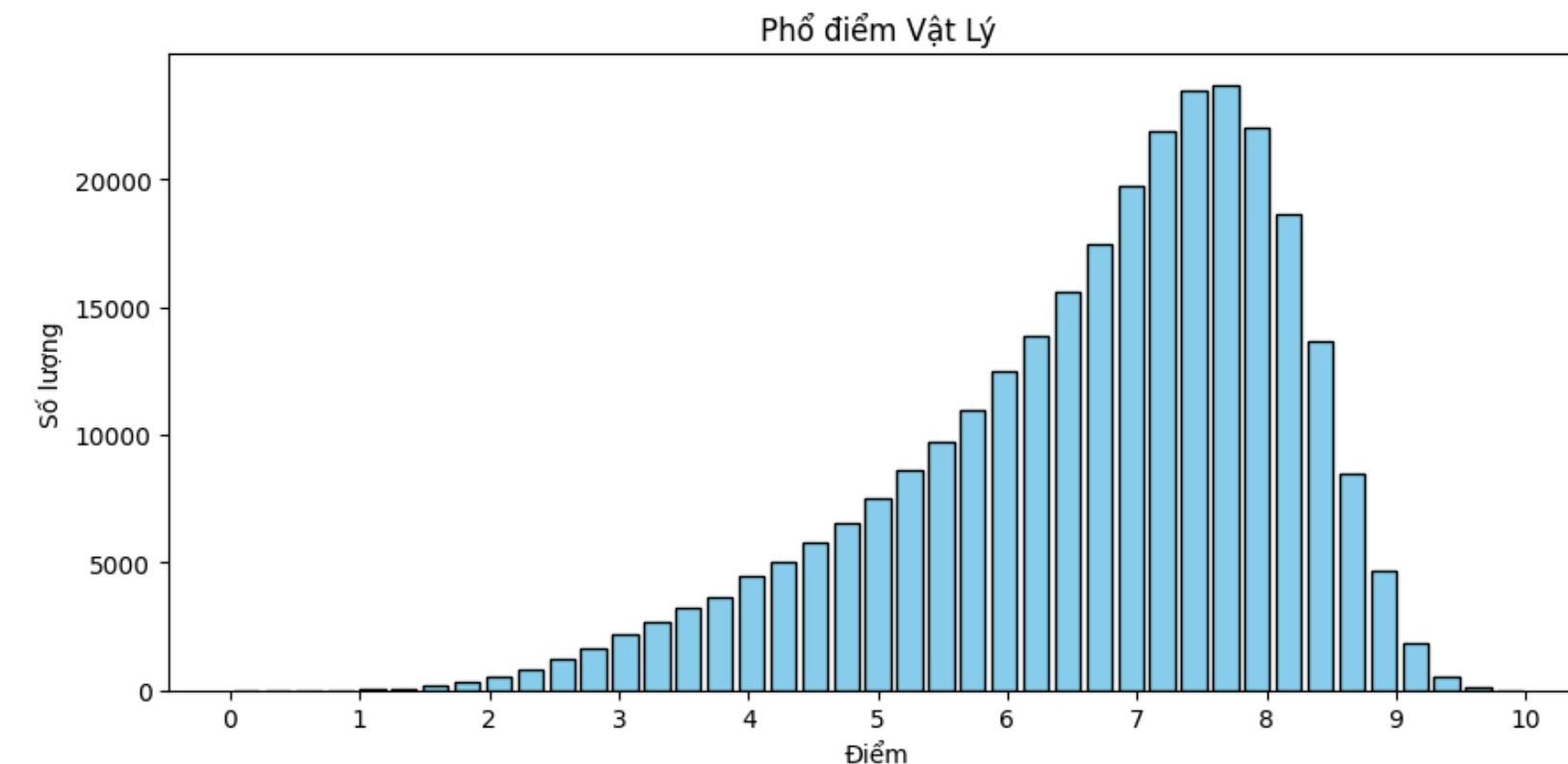
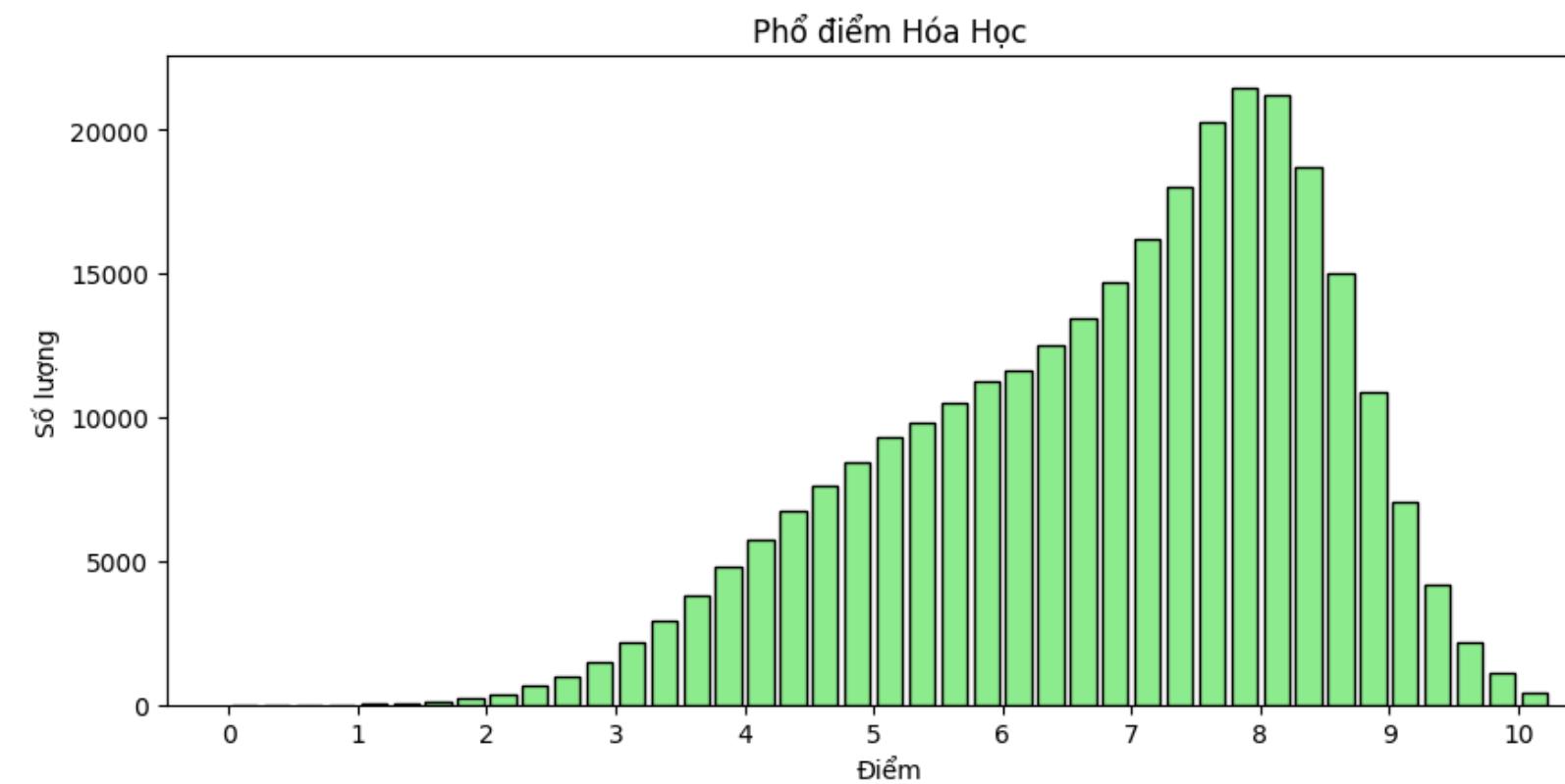
4. Phân tích dữ liệu điểm thi

Phổ điểm các môn thi bắt buộc



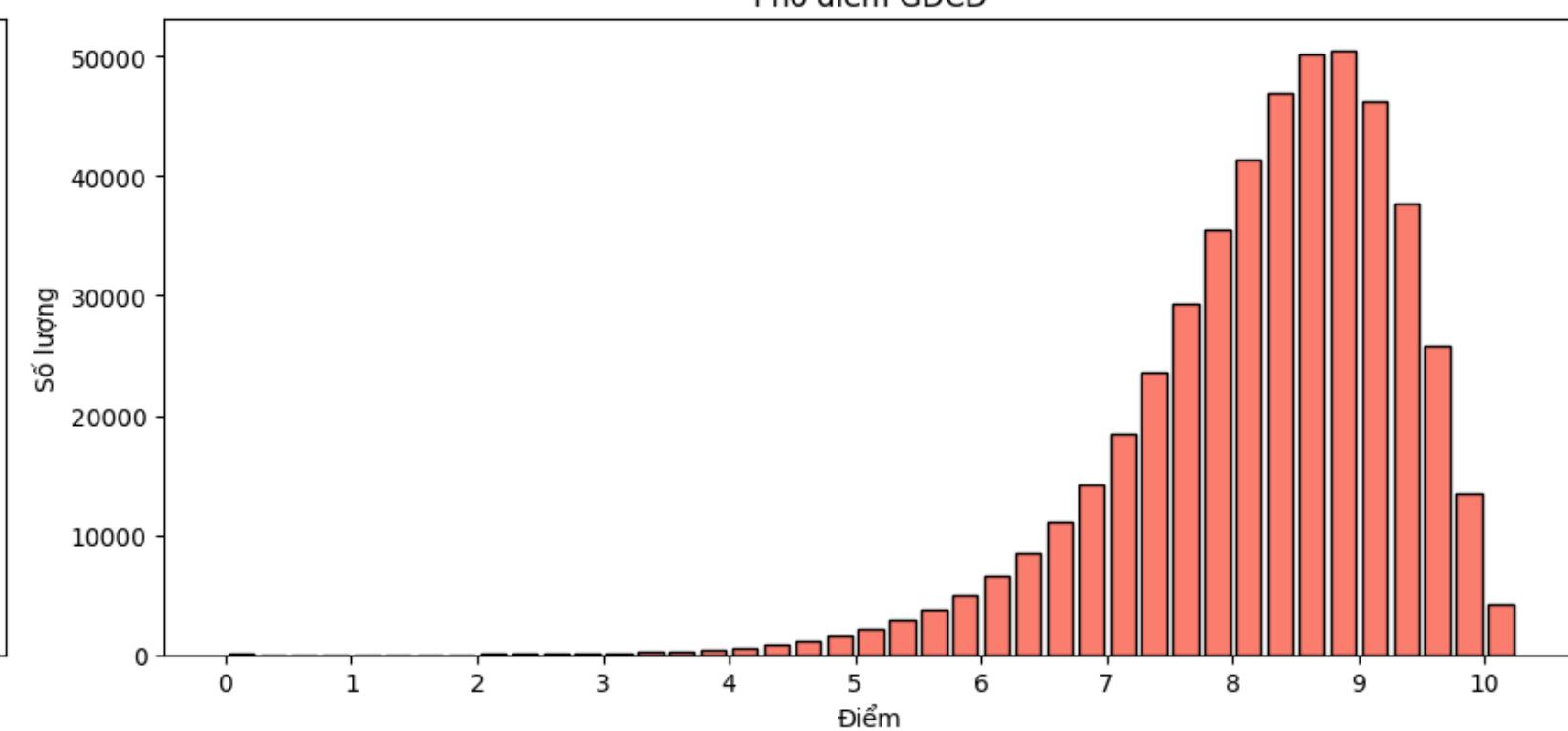
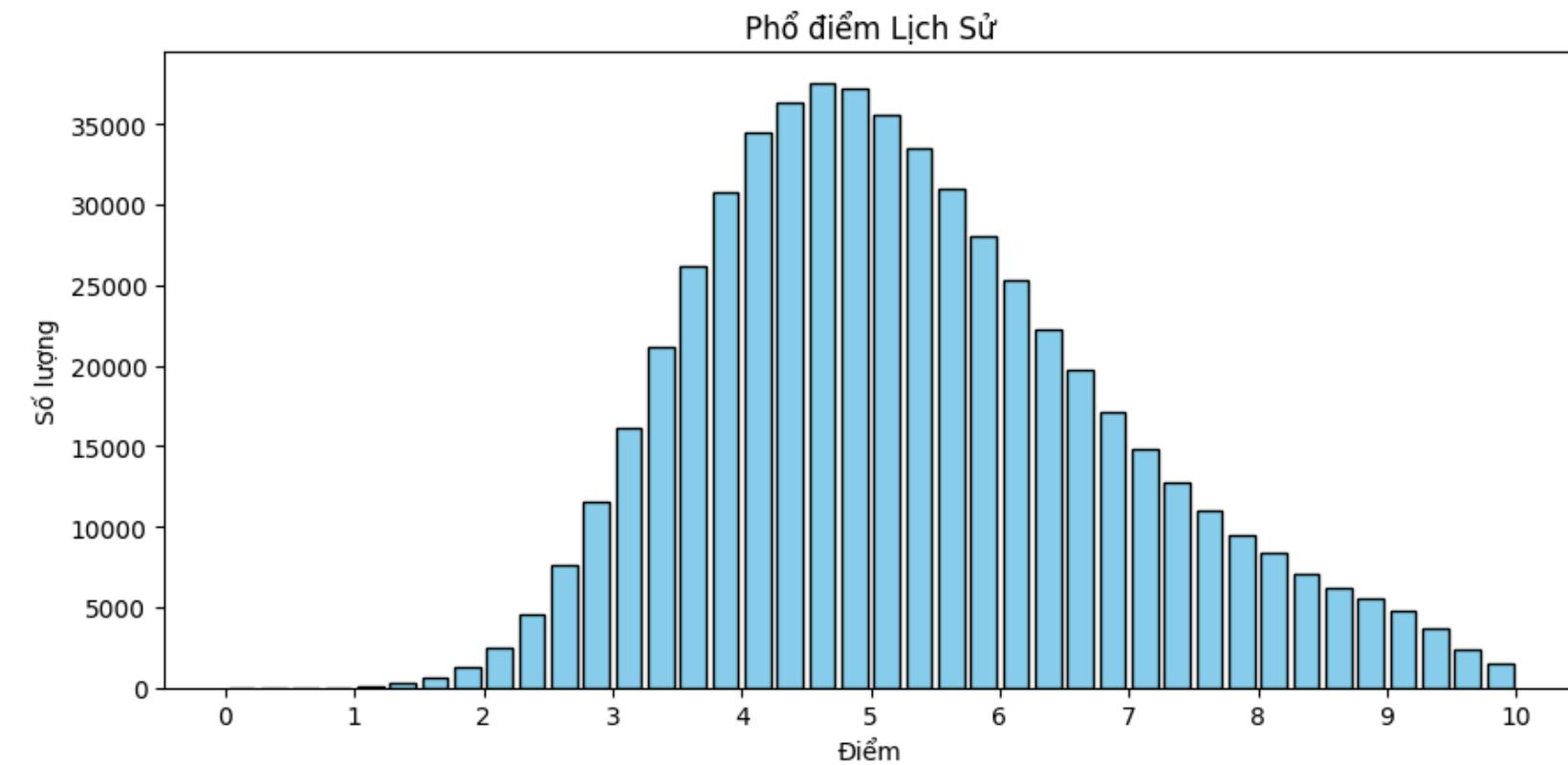
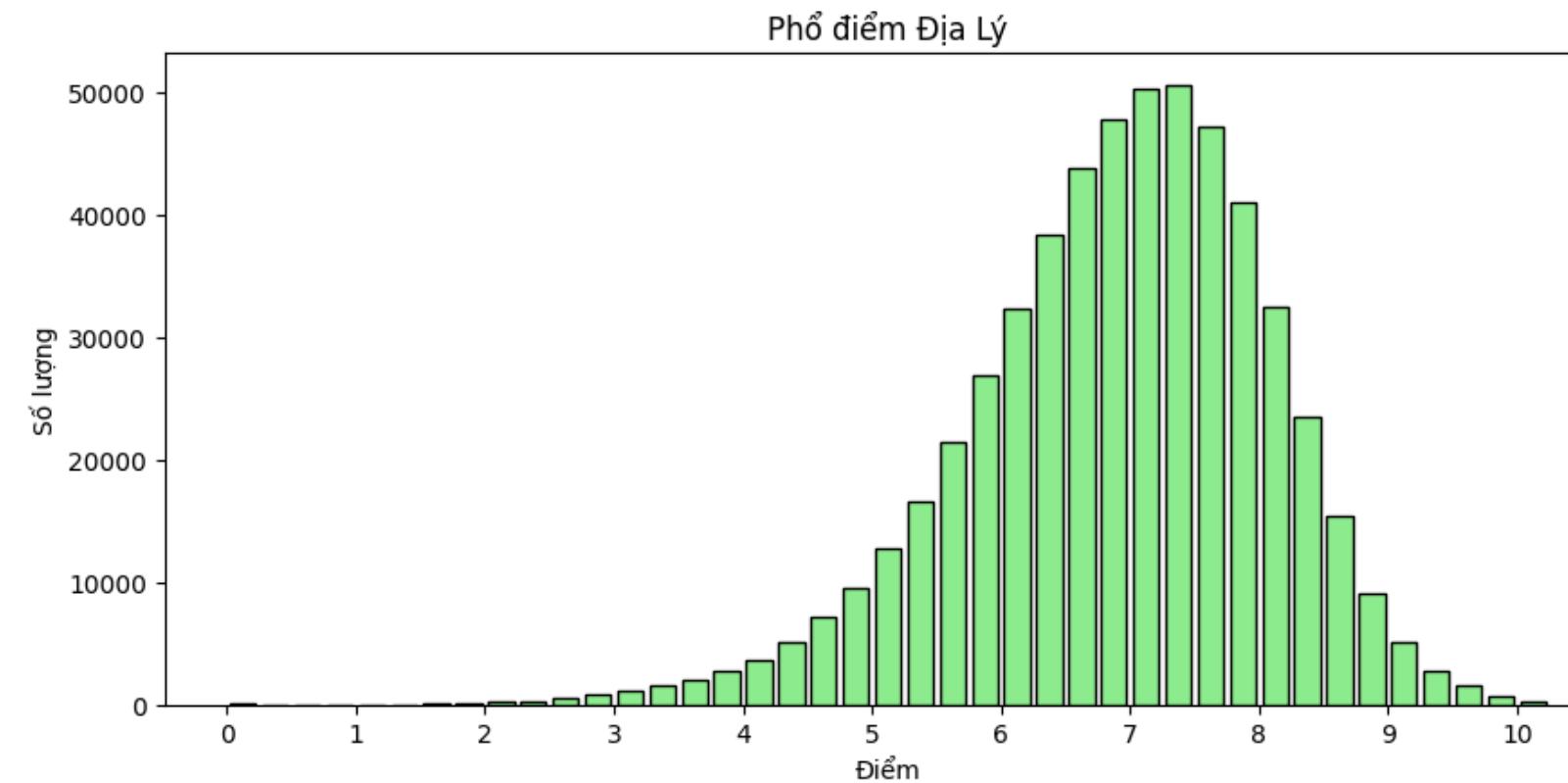
4. Phân tích dữ liệu điểm thi

Phổ điểm các môn thi KHTN



4. Phân tích dữ liệu điểm thi

Phổ điểm các môn thi KHXH



4. Phân tích dữ liệu điểm thi

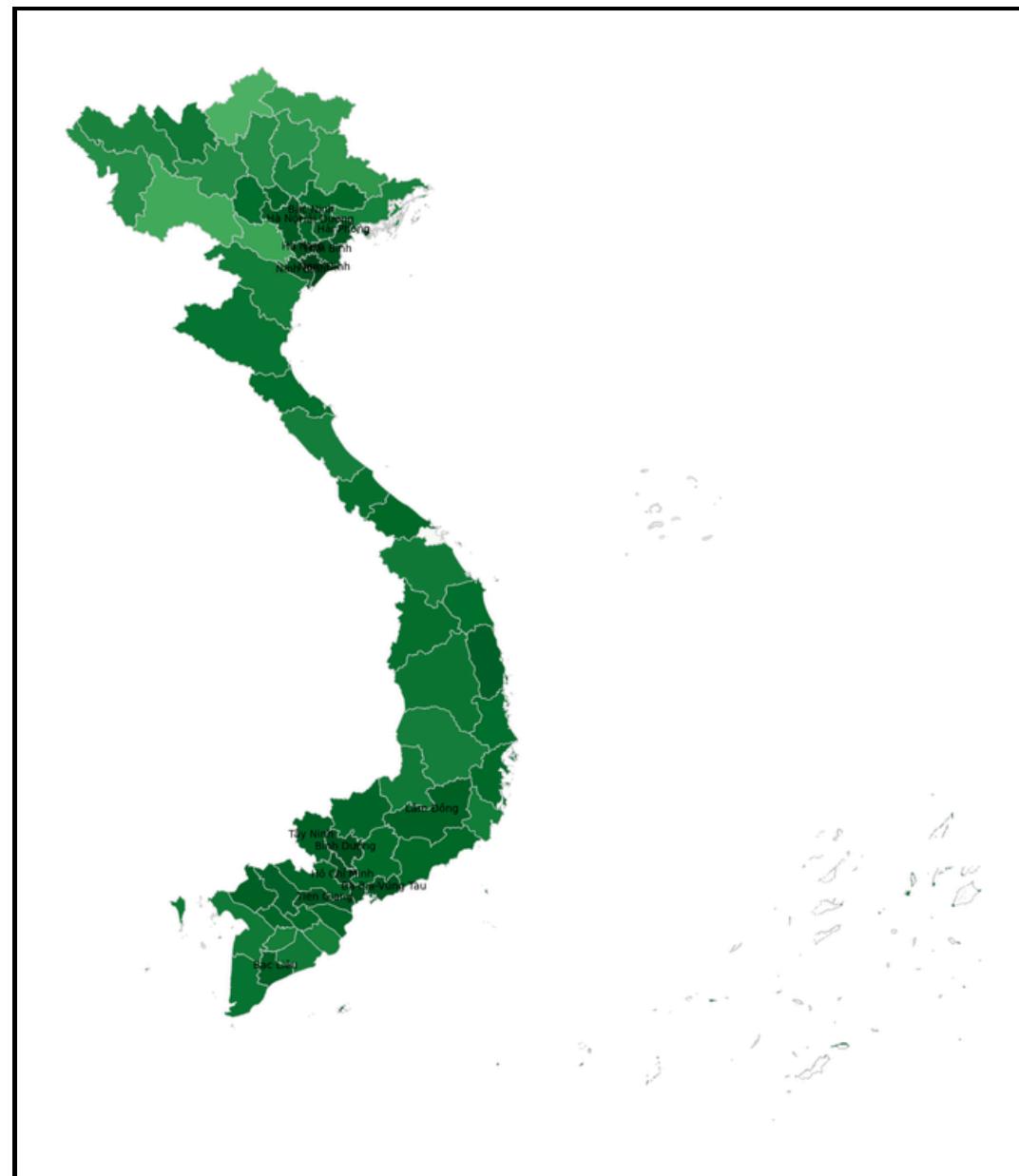
Nhận xét:

Từ biểu đồ trên, có thể thấy:

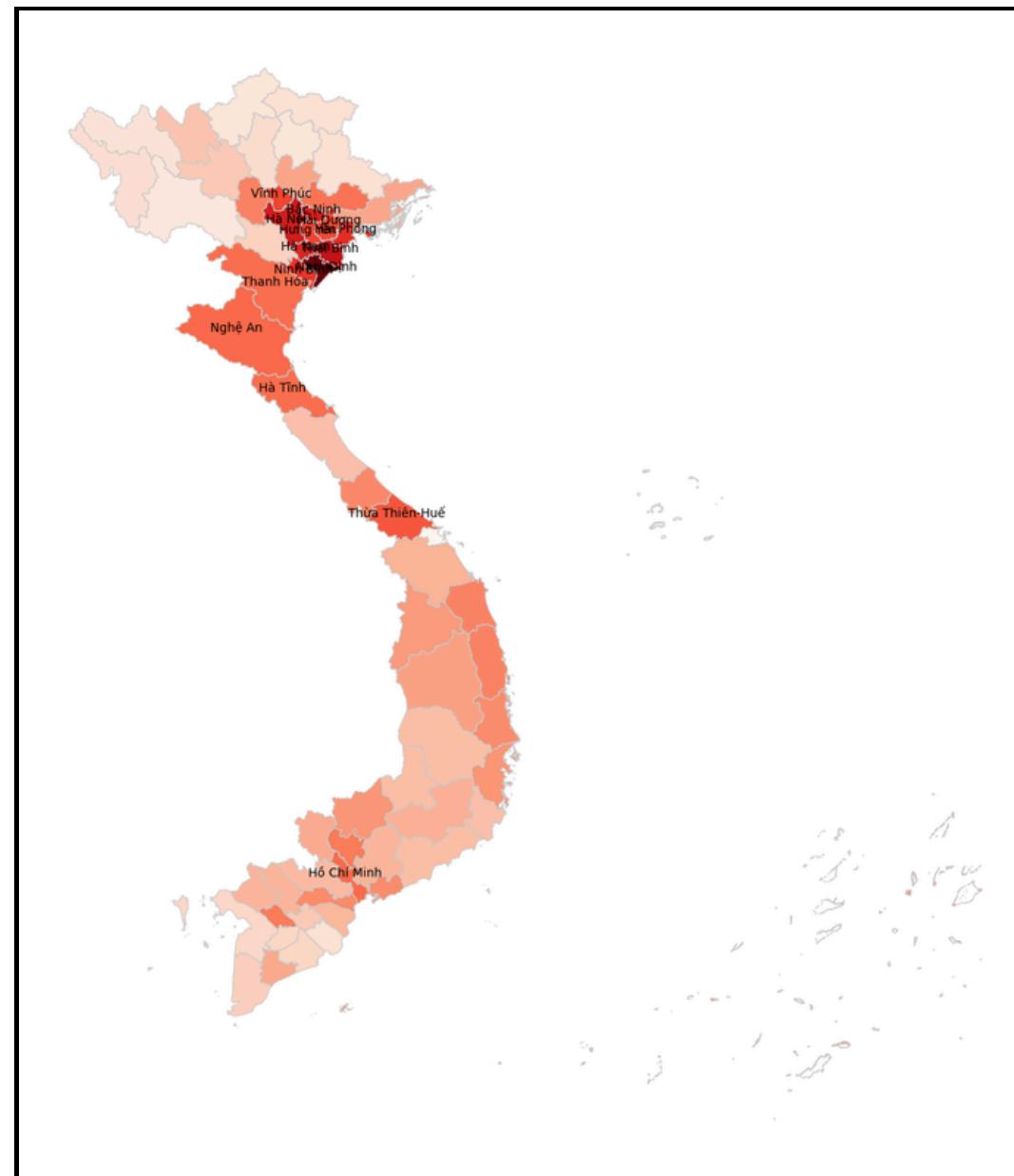
- Các môn Tiếng anh, Sinh học, Lịch sử đang có phổ điểm thấp hơn các môn học khác. Vì vậy, cần nâng cao, cải thiện chất lượng các môn học này.
- Điểm thi của 6 môn học còn lại có điểm trung bình ở mức 6-8 điểm, điều này cho thấy thế mạnh của học sinh ở những môn học này. Các trường THPT cần phát huy điểm mạnh này của mình.



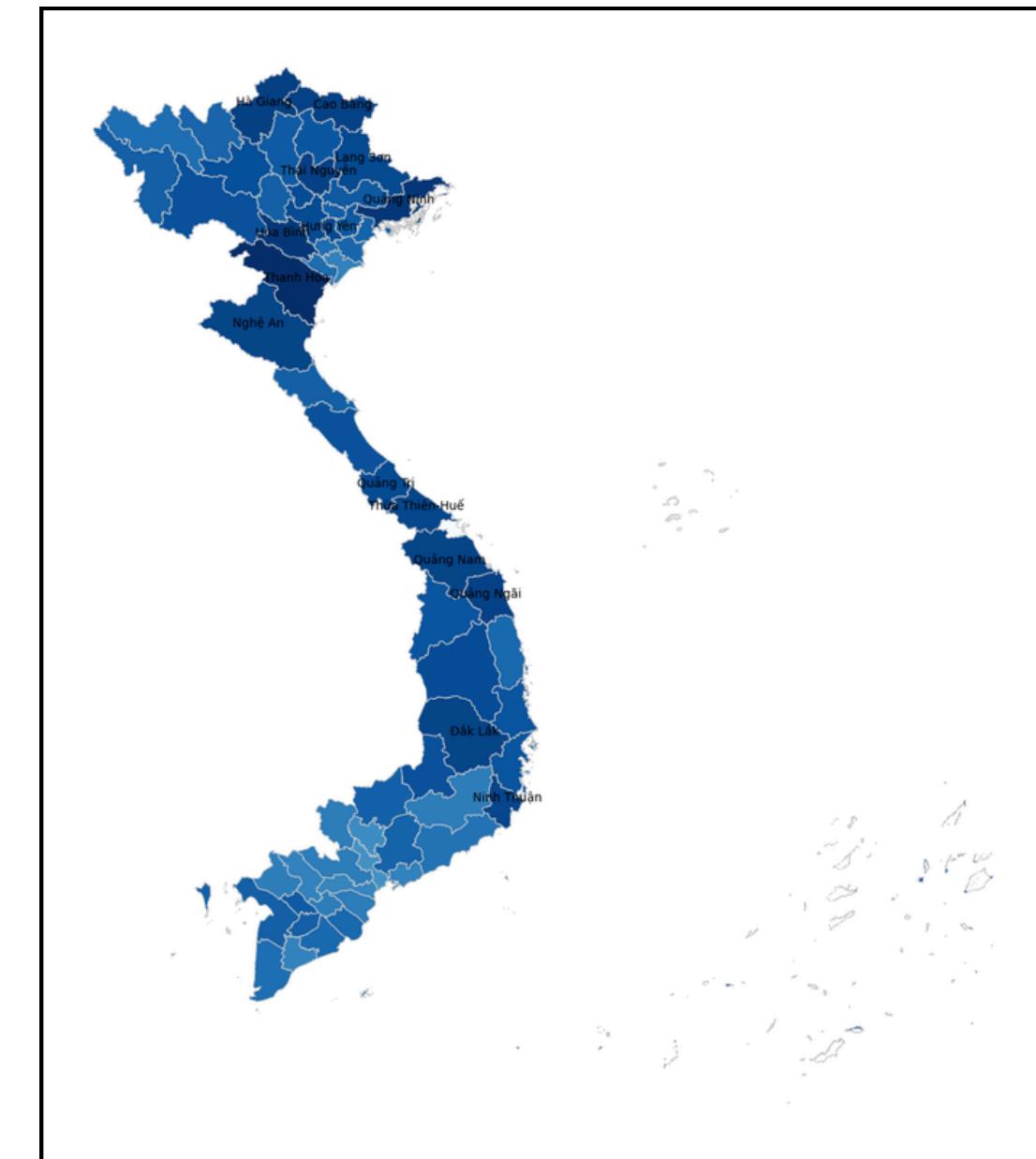
4. Phân tích dữ liệu điểm thi



Điểm trung bình môn Toán
của các tỉnh

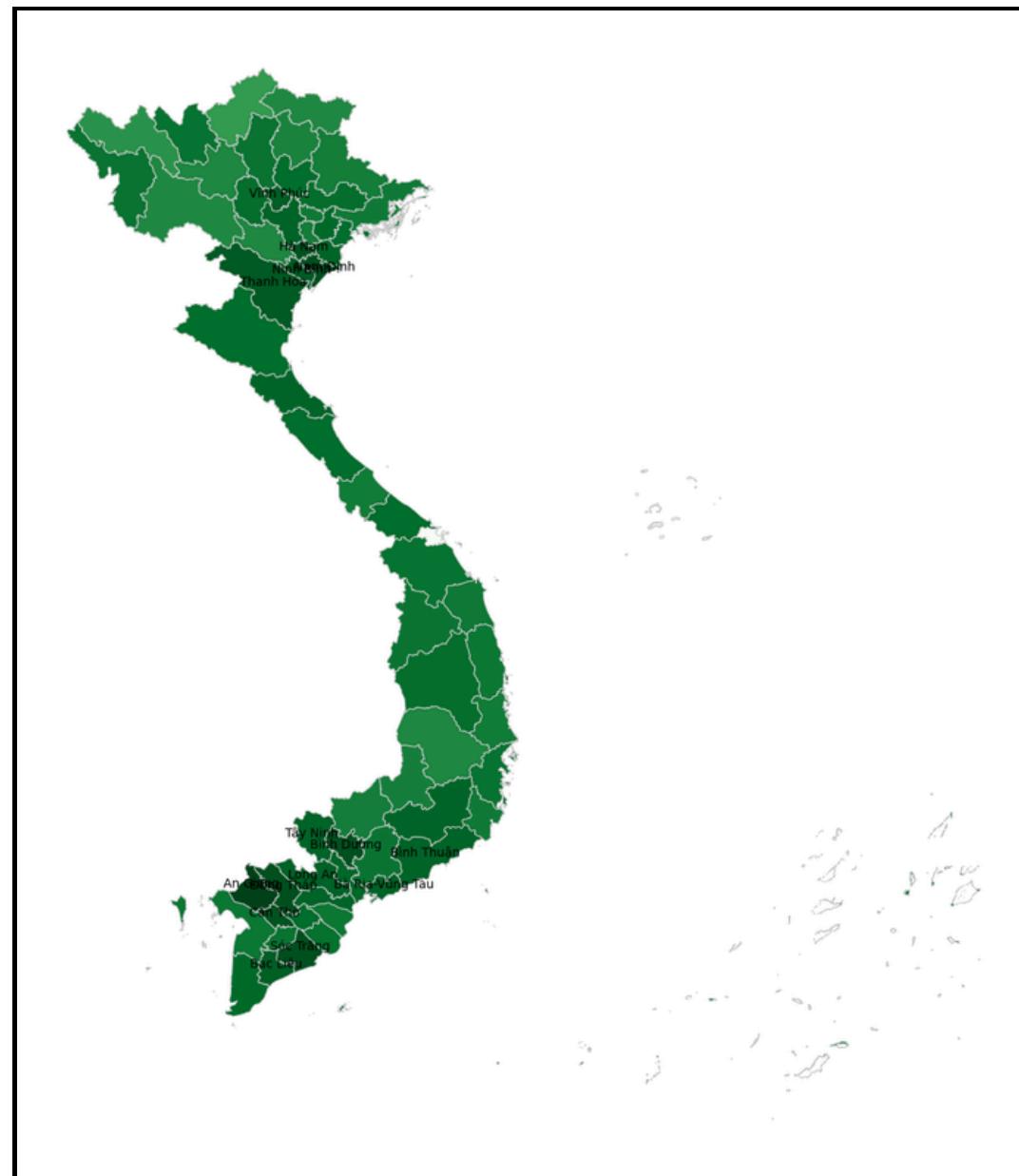


Tỷ lệ học sinh có điểm Toán
 ≥ 9 của các tỉnh

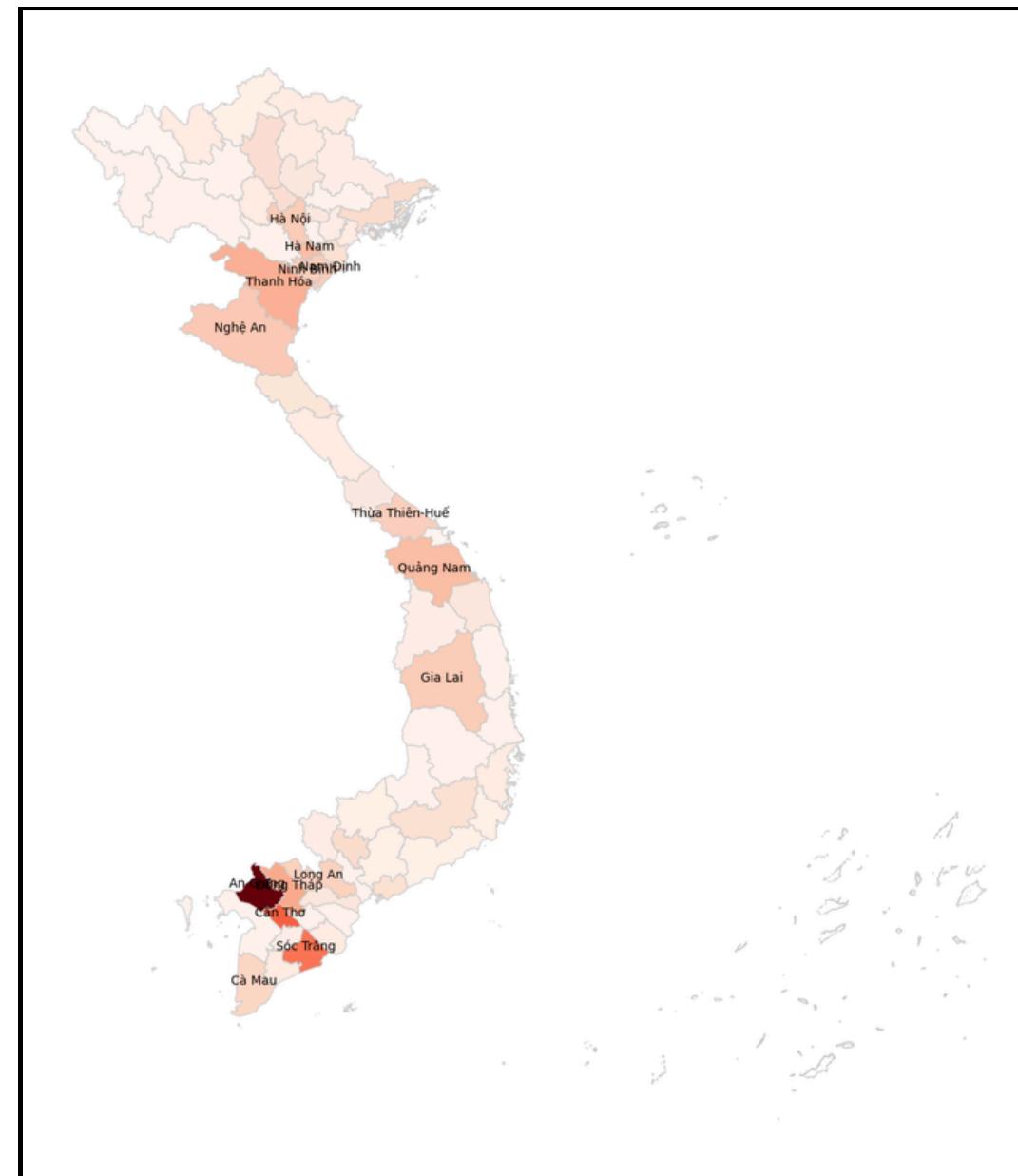


Phương sai điểm Toán
của các tỉnh

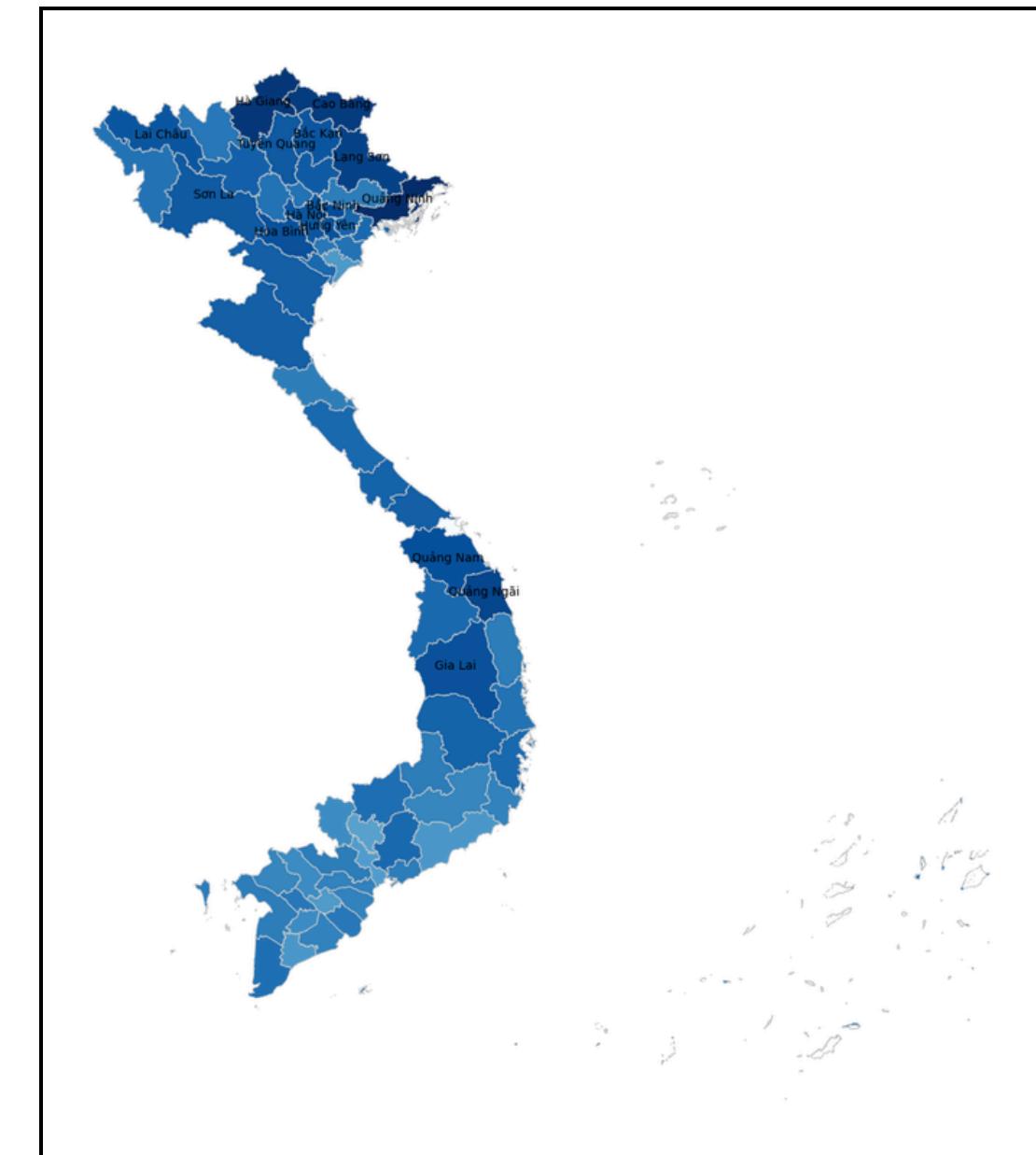
4. Phân tích dữ liệu điểm thi



Điểm trung bình môn Ngữ văn của các tỉnh

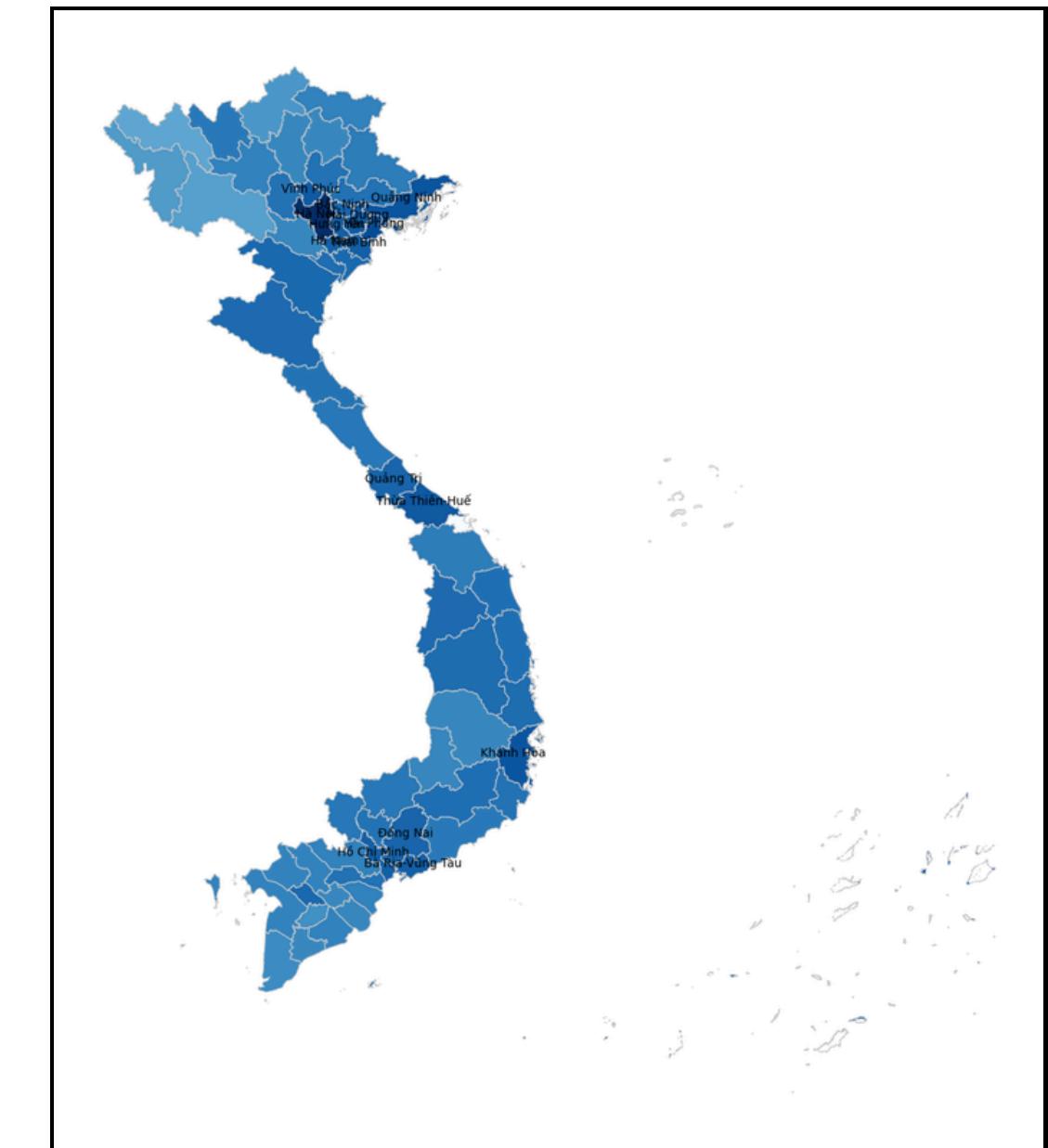
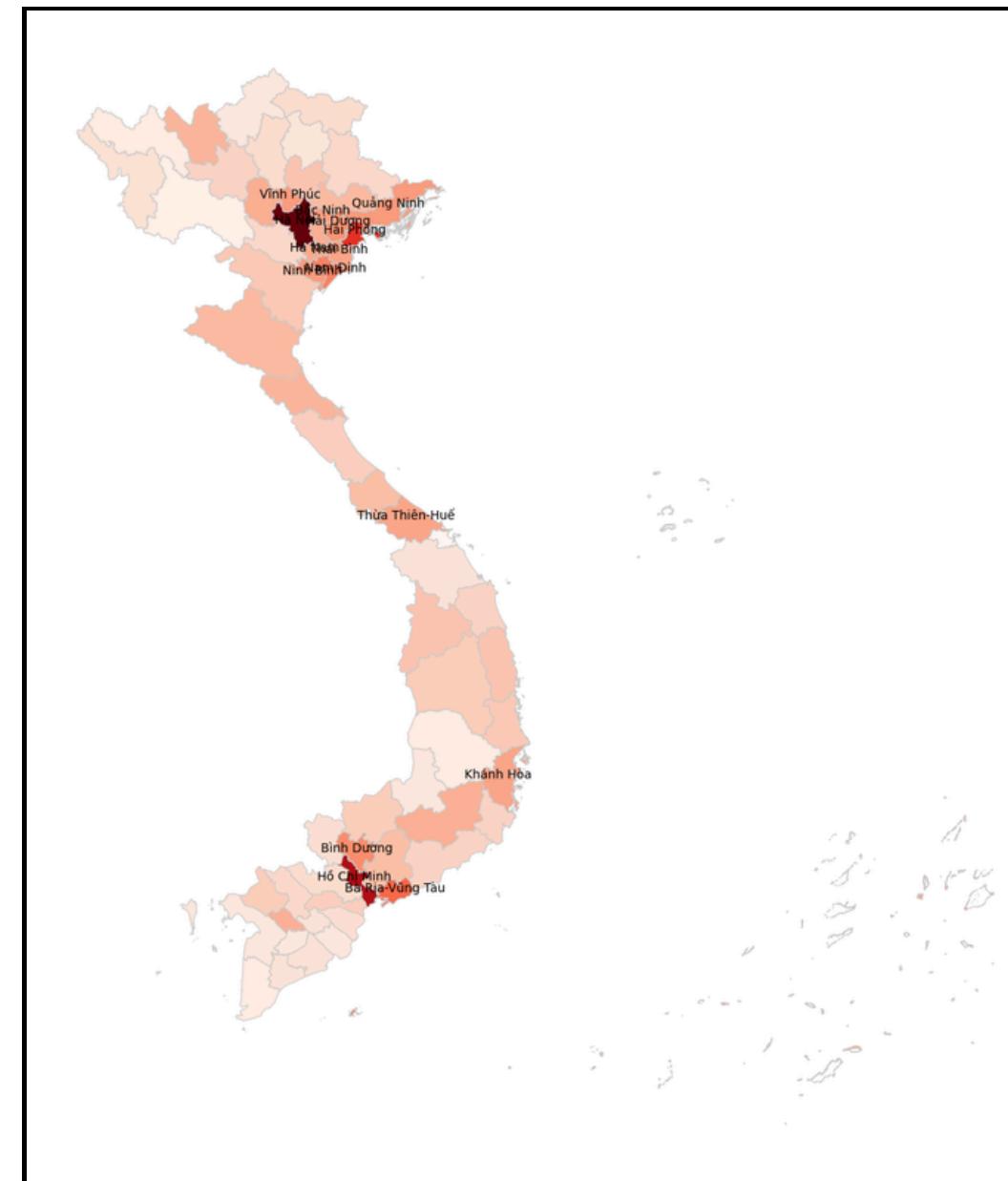
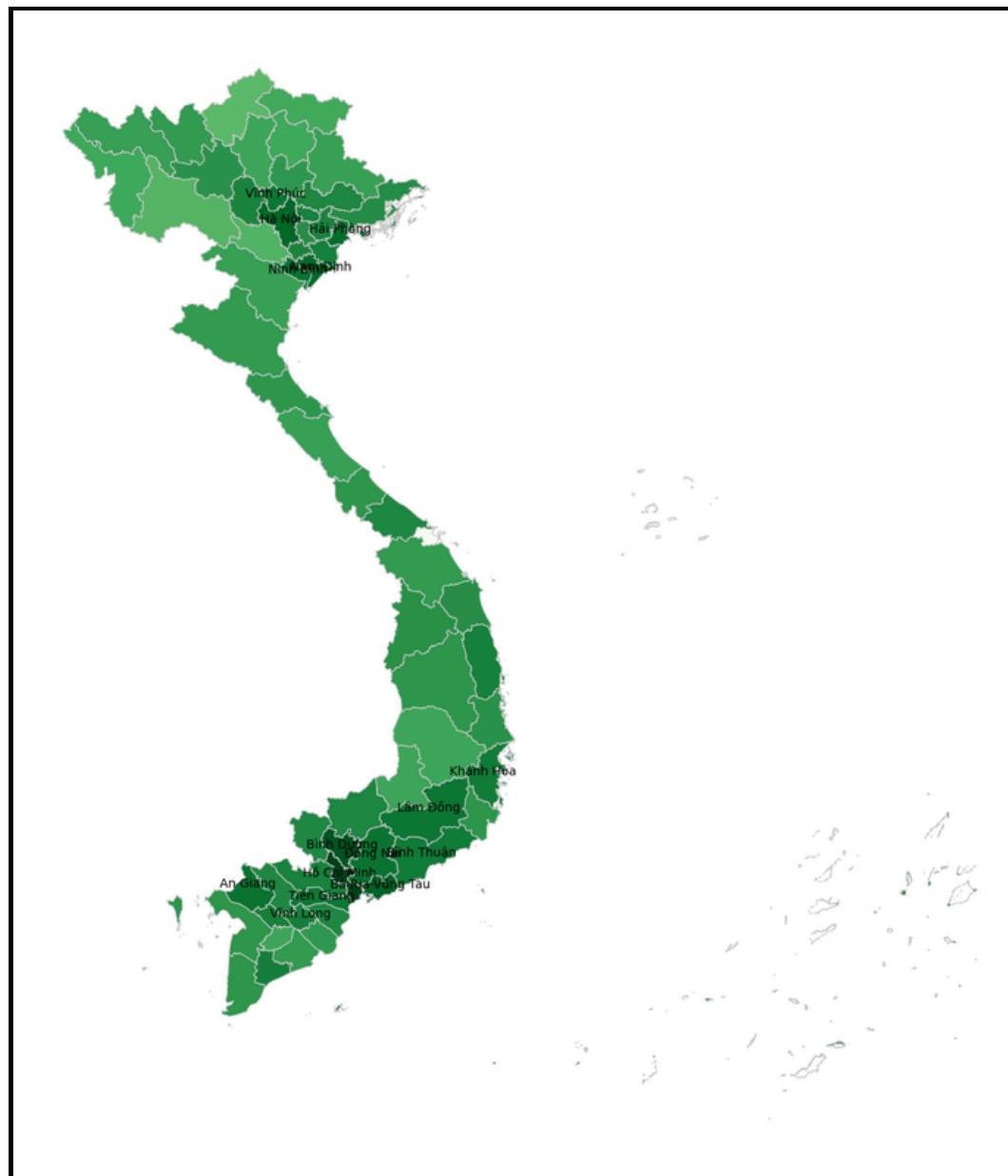


Tỷ lệ học sinh có điểm môn Ngữ văn ≥ 9 của các tỉnh



Phương sai điểm mòn Ngữ văn của các tỉnh

4. Phân tích dữ liệu điểm thi



4. Phân tích dữ liệu điểm thi

Nhận xét:

Từ biểu đồ trên, có thể thấy:

- Vị trí địa lý là một yếu tố có ảnh hưởng lớn đến điểm số.
- Các môn KHTN là điểm mạnh của học sinh Việt Nam.
- Một số tỉnh ở khu vực miền núi phía Bắc, khu vực Duyên hải Nam Trung Bộ và khu vực Tây Nam Bộ có điểm trung bình cũng như tỷ lệ học sinh giỏi chưa cao.
- Ở các môn thi đều có những khu vực có sự không đồng đều về điểm số, có thể là do sự phân loại về “lớp chọn” và “lớp thường”.

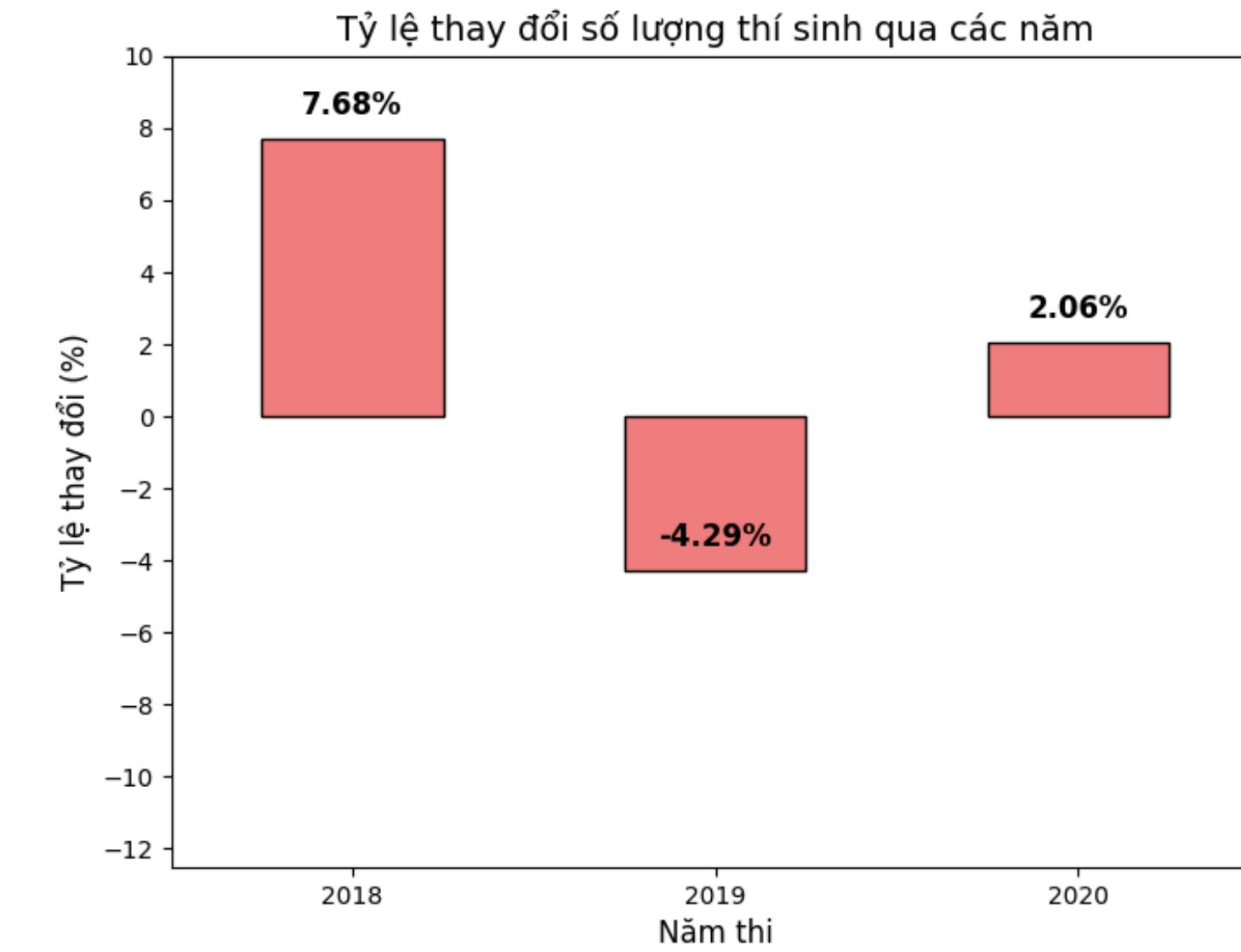
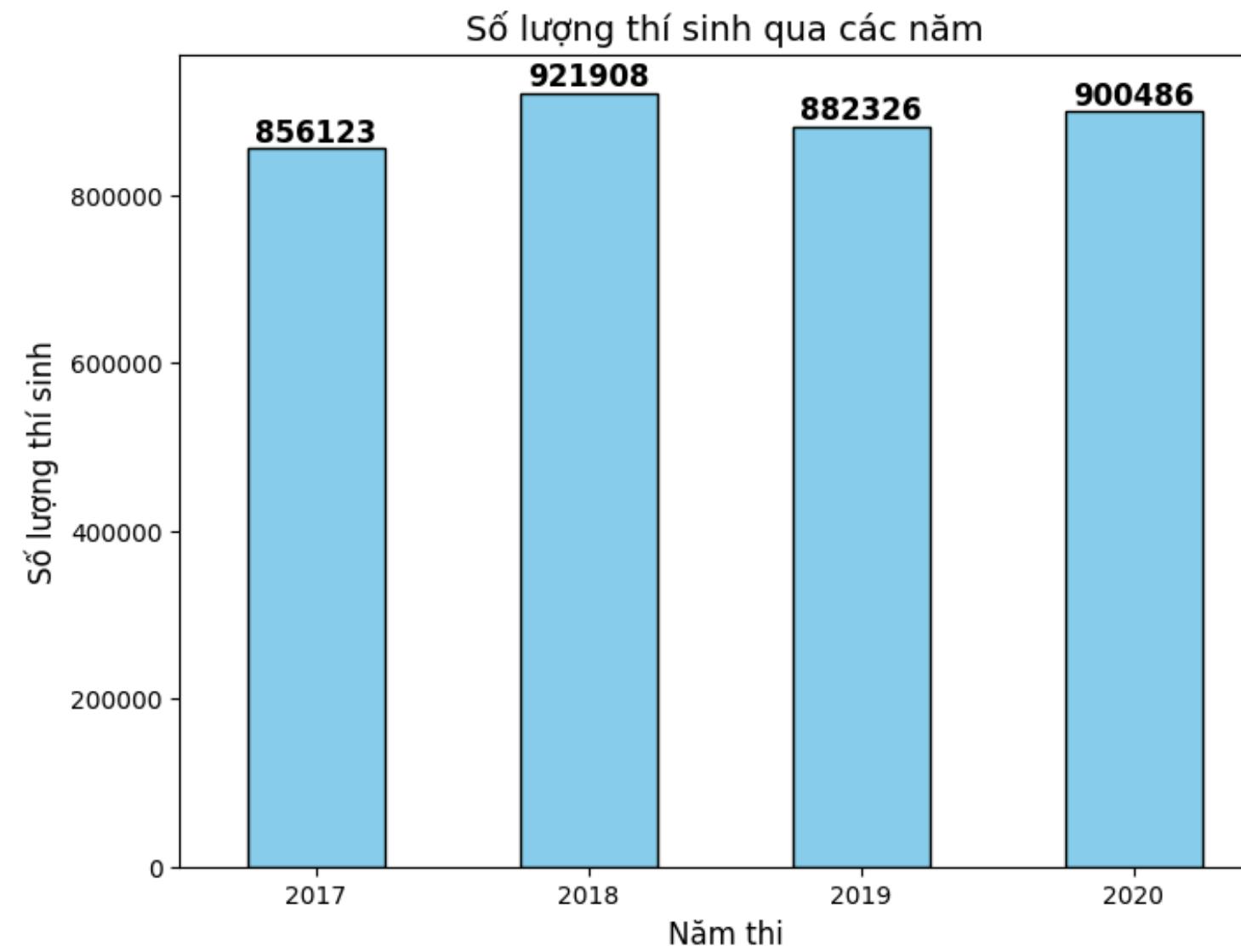


4. Phân tích dữ liệu điểm thi

4.2. So sánh điểm thi THPT Quốc gia 2020 với các năm 2017, 2018, 2019

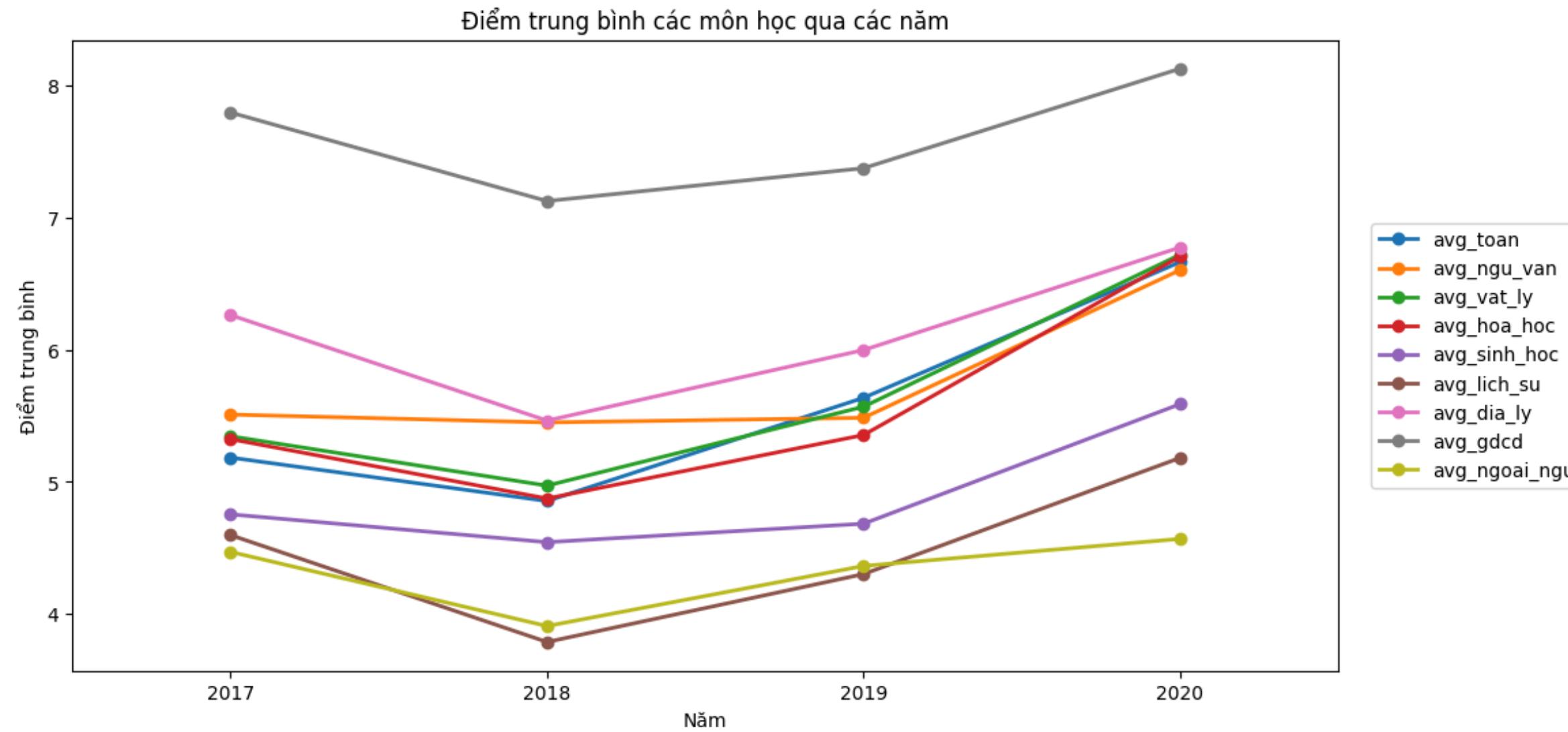


4. Phân tích dữ liệu điểm thi



Trong giai đoạn 2017 - 2020, số lượng thí sinh chỉ có 1 chút biến động.
Số lượng thí sinh năm 2020 chỉ tăng nhẹ so với năm 2019.
-> Mức độ cạnh tranh năm 2020 có thể tăng lên một chút so với 2019.

4. Phân tích dữ liệu điểm thi

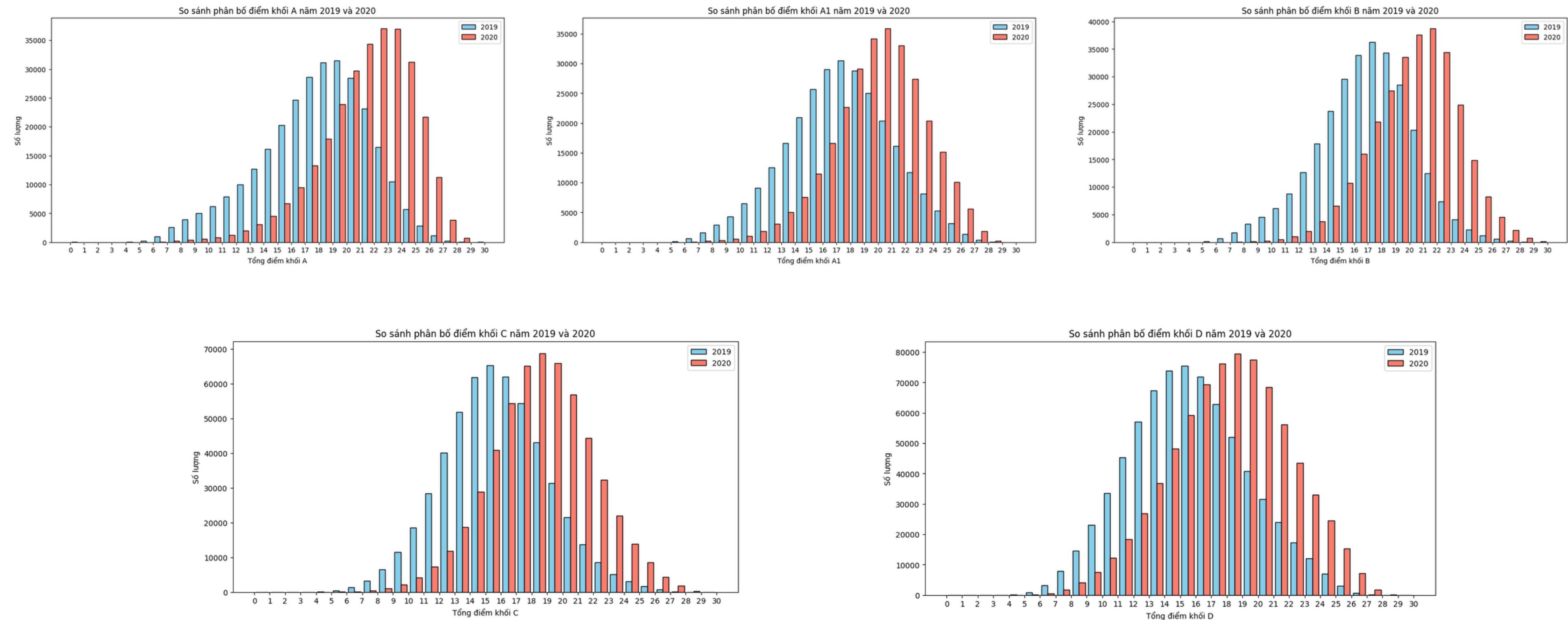


Nhận xét:

Năm 2020, điểm trung bình của tất cả các môn học đều tăng so với năm 2019. Điều này có thể là do đề thi 2020 dễ hơn.

-> Có thể thấy điểm chuẩn năm 2020 sẽ tăng so với 2019

4. Phân tích dữ liệu điểm thi



4. Phân tích dữ liệu điểm thi



**Sau khi biết điểm thi
năm 2020 sẽ tăng so
với năm 2019, thí sinh
nên đăng ký nguyện
vọng xét tuyển Đại
học như thế nào?**

4. Phân tích dữ liệu điểm thi

WEB TRA CỨU ĐIỂM THI CỦA THÍ SINH

Nhập số báo danh của bạn

Điểm thi thành phần của bạn là:

Toán	Ngữ văn	Ng.ngữ	Vật lý	Hóa học	Sinh học	Lịch sử	Địa lý	GDCD
9.0	6.0	3.0	8.5	9.0	4.25	nan	nan	nan

Nhập khối thi của bạn. Cần nhập đúng dạng (A00, A01,..., B00, B01,...)

Điểm khối A00 của bạn là: 26.5

Xếp hạng điểm khối A00 của bạn tại TỈNH THANH HÓA là: 364 trên tổng số 9066 thí sinh, bạn thuộc top 4.02 %

Xếp hạng điểm khối A00 của bạn trên toàn quốc là: 8468 trên tổng số 291252 thí sinh, bạn thuộc top 2.91 %

Điểm số của bạn tương đương với mức điểm 24.2 năm 2019 (dựa trên top điểm thi trên toàn quốc là 2.91 %)

Trang web giúp thí sinh tra cứu vị trí xếp hạng theo khối thi



Nội dung

1. Giới thiệu đề tài
2. Kiến trúc hệ thống
3. Đánh giá hệ thống
4. Phân tích dữ liệu
5. Kết luận và hướng phát triển

5. Kết luận và hướng phát triển

- Việc cài đặt các Node đang thủ công, sẽ gặp khó khăn khi xử lý dữ liệu quy mô lớn hơn
 - > Có thể sử dụng Kubernetes làm quản lý cụm.
- Xây dựng mô hình Học máy để đánh giá độ khó, độ hay của đề thi vì đây là các yếu tố quan trọng giúp phân loại học sinh.
- Có nhiều yếu tố ảnh hưởng đến điểm thi như độ khó của đề thi hoặc điều kiện kinh tế - xã hội ví dụ như COVID-19
 - > Có thể sử dụng thêm các thông tin này để đưa ra các dự đoán chính xác hơn về điểm thi THPT Quốc gia.



A large, faint watermark of the HUST logo is visible across the background of the slide. The logo consists of the letters "HUST" in a white, bold, sans-serif font, with a red, dotted, ribbon-like graphic extending from the right side of the "U" towards the bottom left.

HUST

THANK YOU !