

HỌC VIỆN CÔNG NGHỆ BUUTURE CHÍNH VIỄN THÔNG
KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO BÀI TẬP LỚN

Lập trình Python

Nhóm 10: Phân tích điểm thi THPT quốc gia từ năm 2020-2024

Sinh viên thực hiện:

Lưu Minh Hiển-B22DCCN290: Phân tích tương quan số điểm giữa các môn thi và khối thi.

Ngô Đức Tuấn Cường-B22DCCN094: Phân tích theo từng môn.

Nguyễn Tiến Đạt-B22DCCN196: Phân tích theo khối thi.

Nguyễn Duy Hải-B22DCCN266:.Load data, làm sạch dữ liệu và phân tích số lượng tăng trưởng thí sinh.

Tô Quang Huy-B22DCCN396: Phân tích theo khu vực

Hà Nội, ngày 19/11/2024

A. Tải và làm sạch dữ liệu

```
[49]    import numpy as np
        import pandas as pd
    ✓  0.0s

    Loading data

[50]    data1 = pd.read_csv('F:/python nhom4/data/archive/thpt2020.csv')
        data2 = pd.read_csv('F:/python nhom4/data/archive/thpt2021.csv')
        data3 = pd.read_csv('F:/python nhom4/data/archive/thpt2022.csv')
        data4 = pd.read_csv('F:/python nhom4/data/archive/thpt2023.csv')
        data5 = pd.read_csv('F:/python nhom4/data/archive/thpt2024.csv')
    ✓  3.3s
```

Gọi thư viện và tải dữ liệu vào chương trình

```
▶ v  [51]  data1.info()
    ✓  0.0s
...
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 870486 entries, 0 to 870485
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   SBD              870486 non-null   int64  
 1   Toan             866581 non-null   float64 
 2   NguVan          856565 non-null   float64 
 3   VatLy            293287 non-null   float64 
 4   HoaHoc           295536 non-null   float64 
 5   SinhHoc          290377 non-null   float64 
 6   LichSu            568581 non-null   float64 
 7   DiaLy             555072 non-null   float64 
 8   GDCH             482980 non-null   float64 
 9   NgoaiNgu          772098 non-null   float64 
 10  MaMonNgoaiNgu     772098 non-null   object  
dtypes: float64(9), int64(1), object(1)
memory usage: 73.1+ MB
```

Hàm info(): cho biết thông tin các thuộc tính trong data

```

  ↴
    ✓ data1.sort_values(by = "SBD", ascending= True).head[5]
    ✓ 0.0s

```

	SBD	Toan	NguVan	VatLy	HoaHoc	SinhHoc	LichSu	DiaLy	GDCD	NgoaiNgu	MaMon	NgoaiNgu
0	2001000001	7.0	6.0	NaN	NaN	NaN	8.00	NaN	NaN	NaN		NaN
1	2001000002	9.0	6.0	7.25	8.75	7.25	NaN	NaN	NaN	9.4		N1
2	2001000003	8.8	6.5	8.25	8.75	4.75	NaN	NaN	NaN	6.0		N1
3	2001000004	7.0	NaN	NaN	6.00	NaN	NaN	NaN	NaN	8.2		N1
4	2001000005	9.0	7.5	NaN	NaN	NaN	5.75	7.25	9.25	9.2		N1

data1.sort_values(by='SBD', ascending=True).head(5)

In ra 5 bản ghi đầu tiên của data1 được sắp xếp theo SBD

```

  ↴
    ✓ data1[['SBD', 'Toan', 'NguVan', 'VatLy', 'HoaHoc', 'SinhHoc', 'LichSu', 'DiaLy', 'GDCD', 'NgoaiNgu']]
    ✓ data1.columns = ['sbd', 'toan', 'ngu_van', 'vat_ly', 'hoa_hoc', 'sinh_hoc', 'lich_su', 'dia_ly', 'gdcd', 'ngoai_ngu']
    ✓ data1.sort_values(by = "sbd", ascending= True).head[5]
    ✓ 0.0s

```

	sbd	toan	ngu_van	vat_ly	hoa_hoc	sinh_hoc	lich_su	dia_ly	gdcd	ngoai_ngu
0	2001000001	7.0	6.0	NaN	NaN	NaN	8.00	NaN	NaN	NaN
1	2001000002	9.0	6.0	7.25	8.75	7.25	NaN	NaN	NaN	9.4
2	2001000003	8.8	6.5	8.25	8.75	4.75	NaN	NaN	NaN	6.0
3	2001000004	7.0	NaN	NaN	6.00	NaN	NaN	NaN	NaN	8.2
4	2001000005	9.0	7.5	NaN	NaN	NaN	5.75	7.25	9.25	9.2

Lấy các thuộc tính cần và thay đổi các tên thuộc tính về dạng chuẩn cùng với các bảng dữ liệu khác

```
data1['year'] = '2020'
data2['year'] = '2021'
data3['year'] = '2022'
data4['year'] = '2023'
data5['year'] = '2024'
] ✓ 0.0s
```



```
data_all = (
    pd.concat([data1, data2, data3, data4, data5], ignore_index=True)
)
data_all.info()
] ✓ 0.3s
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4864150 entries, 0 to 4864149
Data columns (total 11 columns):
 #   Column      Dtype  
 --- 
 0   sbd        int64  
 1   toan       float64 
 2   ngu_van    float64 
 3   vat_ly     float64 
 4   hoa_hoc    float64 
 5   sinh_hoc   float64 
 6   lich_su    float64 
 7   dia_ly     float64 
 8   gdcd       float64 
 9   ngoai_ngu  float64 
 10  year       object  
dtypes: float64(9), int64(1), object(1)
memory usage: 408.2+ MB
```

Gắn cột năm cho từng bảng và gộp chung tất cả các bảng dữ liệu lại với nhau

```

69] data_all['sbd'] = data_all['sbd'].astype(str).str.zfill(8)
       data_all['sbd'] = data_all['sbd'].str[-8:]
       data_all.head(5)
✓ 2.6s

..      sbd  toan  ngu_van  vat_ly  hoa_hoc  sinh_hoc  lich_su  dia_ly  gdcd  ngoai_ngu  year
0  01000001    7.0      6.0    NaN     NaN     NaN    8.00    NaN    NaN     NaN  2020
1  01000002    9.0      6.0    7.25    8.75    7.25    NaN    NaN    NaN     9.4  2020
2  01000003    8.8      6.5    8.25    8.75    4.75    NaN    NaN    NaN     6.0  2020
3  01000004    7.0      NaN    NaN     6.00    NaN    NaN    NaN     NaN     8.2  2020
4  01000005    9.0      7.5    NaN     NaN     NaN    5.75    7.25   9.25     9.2  2020

70] data_all['dia_phuong'] = data_all['sbd'].str[:2]
       data_all.head(5)
✓ 0.7s

..      sbd  toan  ngu_van  vat_ly  hoa_hoc  sinh_hoc  lich_su  dia_ly  gdcd  ngoai_ngu  year  dia_phuong
0  01000001    7.0      6.0    NaN     NaN     NaN    8.00    NaN    NaN     NaN  2020        01
1  01000002    9.0      6.0    7.25    8.75    7.25    NaN    NaN    NaN     9.4  2020        01
2  01000003    8.8      6.5    8.25    8.75    4.75    NaN    NaN    NaN     6.0  2020        01
3  01000004    7.0      NaN    NaN     6.00    NaN    NaN    NaN     NaN     8.2  2020        01
4  01000005    9.0      7.5    NaN     NaN     NaN    5.75    7.25   9.25     9.2  2020        01

```

Chỉnh sửa các giá trị sbd sao cho về cùng 1 định dạng

Trong sbd thì 2 chữ số đầu chính là mã địa phương tạo thêm cột địa phuong

```

File Edit Selection View Go Run Terminal Help ⏪ BTL
EXPLORER ... phan_tich_theo_mon_hoc.ipynb tuong_quan_giu_cac_mon_thi_va_khoi.thipyrb loading_and_transform.ipynb tuong_quan_giu_cac_mon_thi_va_khoi.thi.ipynb Python 3.11.9
data_sach + Markdown ▶ Run All ⌂ Restart Clear All Outputs Variables Outline ...
data_all.csv ... '58': 'Trà Vinh',
... '59': 'Sóc Trăng',
... '60': 'Bạc Liêu',
... '61': 'Cà Mau',
... '62': 'Điện Biên',
... '63': 'Đăk Nông',
... '64': 'Hậu Giang'
}
data_all['dia_phuong'] = data_all['dia_phuong'].replace(locality_mapping)
data_all.head(5)
✓ 16.5s

..      sbd  toan  ngu_van  vat_ly  hoa_hoc  sinh_hoc  lich_su  dia_ly  gdcd  ngoai_ngu  year  dia_phuong
0  01000001    7.0      6.0    NaN     NaN     NaN    8.00    NaN    NaN     NaN  2020        Hà Nội
1  01000002    9.0      6.0    7.25    8.75    7.25    NaN    NaN    NaN     9.4  2020        Hà Nội
2  01000003    8.8      6.5    8.25    8.75    4.75    NaN    NaN    NaN     6.0  2020        Hà Nội
3  01000004    7.0      NaN    NaN     6.00    NaN    NaN    NaN     NaN     8.2  2020        Hà Nội
4  01000005    9.0      7.5    NaN     NaN     NaN    5.75    7.25   9.25     9.2  2020        Hà Nội

Xuất data sạch ra ngoài

df1 = pd.DataFrame(data1)
df1.to_csv('data2020.csv', index = False)
df2 = pd.DataFrame(data2)
df2.to_csv('data2021.csv', index = False)
df3 = pd.DataFrame(data3)
df3.to_csv('data2022.csv', index = False)
df4 = pd.DataFrame(data4)
df4.to_csv('data2023.csv', index = False)
df5 = pd.DataFrame(data5)
df5.to_csv('data2024.csv', index = False)
df6 = pd.DataFrame(data_all)
df6.to_csv('data_all.csv', index = False)

0.0s
Ln 66, Col 17 (1575 selected) Spaces: 4 CRLF Cell 7 of 26 ENG INTL 3:30 PM 11/20/2024
28°C Có nắng

```

Từ mã địa phương gắn tên địa phương của thí sinh dùng hàm .replace()

Sau khi chỉnh sửa xong thì xuất dữ liệu data ra ngoài bằng hàm .to_csv với file xuất ra là csv.

B.Phân tích tương quan

I. Tương quan điểm số giữa các môn thi

1. Mối tương quan nổi bật giữa các môn học

TƯƠNG QUAN GIỮA CÁC MÔN THI

```
# Tạo khung hình lớn cho các heatmap
plt.figure(figsize=(20, 15))

# Danh sách các năm và dữ liệu tương ứng
years = [2020, 2021, 2022, 2023, 2024]
data_list = [data1, data2, data3, data4, data5] # Đây là các DataFrame tương ứng cho từng năm

# Danh sách tiêu đề cho từng năm
titles = ['Năm 2020', 'Năm 2021', 'Năm 2022', 'Năm 2023', 'Năm 2024']

# Vòng lặp qua các năm để tạo heatmap cho từng năm
for i, (year, data) in enumerate(zip(years, data_list)):
    # Chọn các môn cần phân tích
    subjects_scores = data[['toan', 'ngu_van', 'ngoai_ngu', 'vat_ly', 'hoa_hoc', 'sinh_hoc', 'lich_su', 'dia_ly', 'gdcd']]

    # Tính ma trận tương quan
    correlation_matrix = subjects_scores.corr()

    # Tạo subplot (sắp xếp theo 2 hàng, 3 cột)
    plt.subplot(2, 3, i + 1)

    # Vẽ heatmap
    sns.heatmap(correlation_matrix, annot=True, cmap='YlGn', fmt='.2f', vmin=-1, vmax=1, linewidths=0.5)

    # Thiết lập tiêu đề và nhãn
    plt.title(f'Heatmap của ma trận tương quan điểm {titles[i]}')
    plt.xlabel('Môn học')
    plt.ylabel('Môn học')

    # Điều chỉnh layout để các biểu đồ không bị chồng lấn
    plt.tight_layout()
plt.show()
```

plt.figure(figsize=(20,15)): tạo khung hình cho heatmap

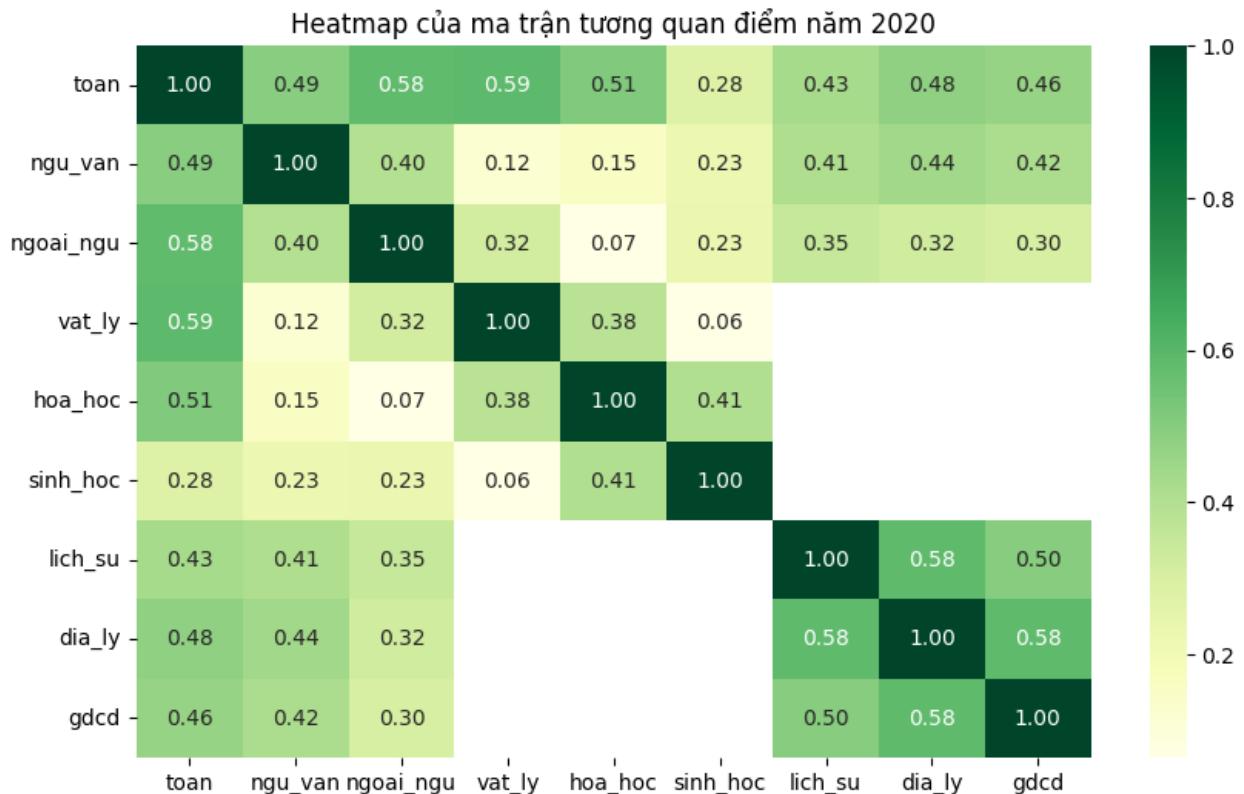
subjects_scores: lấy dữ liệu DataFrame chỉ chứa điểm số của các môn trên

correlation_matrix = subjects_score.corr(): Phương thức này tính toán ma trận tương quan giữa các cột trong DataFrame subjects_score

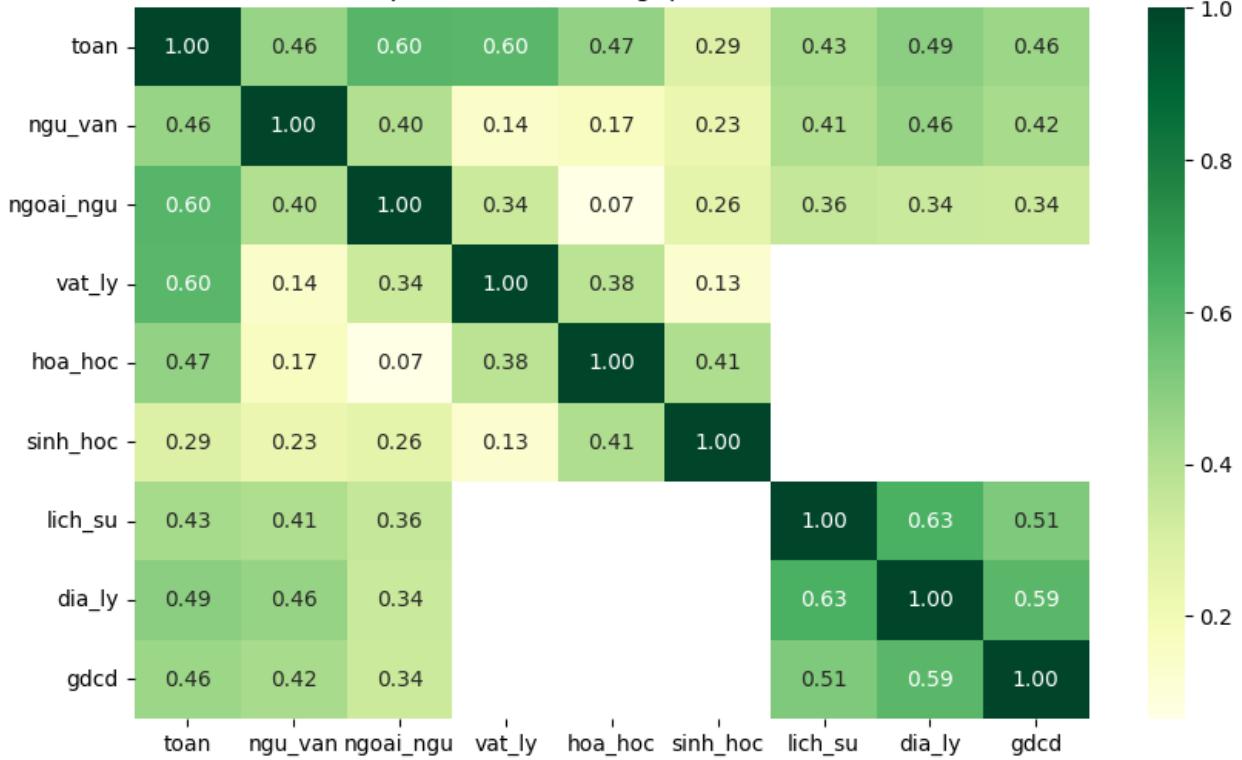
sns.heatmap(correlation_matrix, annot=True, cmap='YlGn', fmt='%.2f', vmin=-1, linewidths=0.5):

- **sns.heatmap:** Hàm vẽ biểu đồ nhiệt dựa trên ma trận tương quan.
- **correlation_matrix:** Dữ liệu để vẽ (ma trận đã tính ở trên).
- **annot=True:** Hiển thị giá trị tương quan trực tiếp trên ô trong biểu đồ.

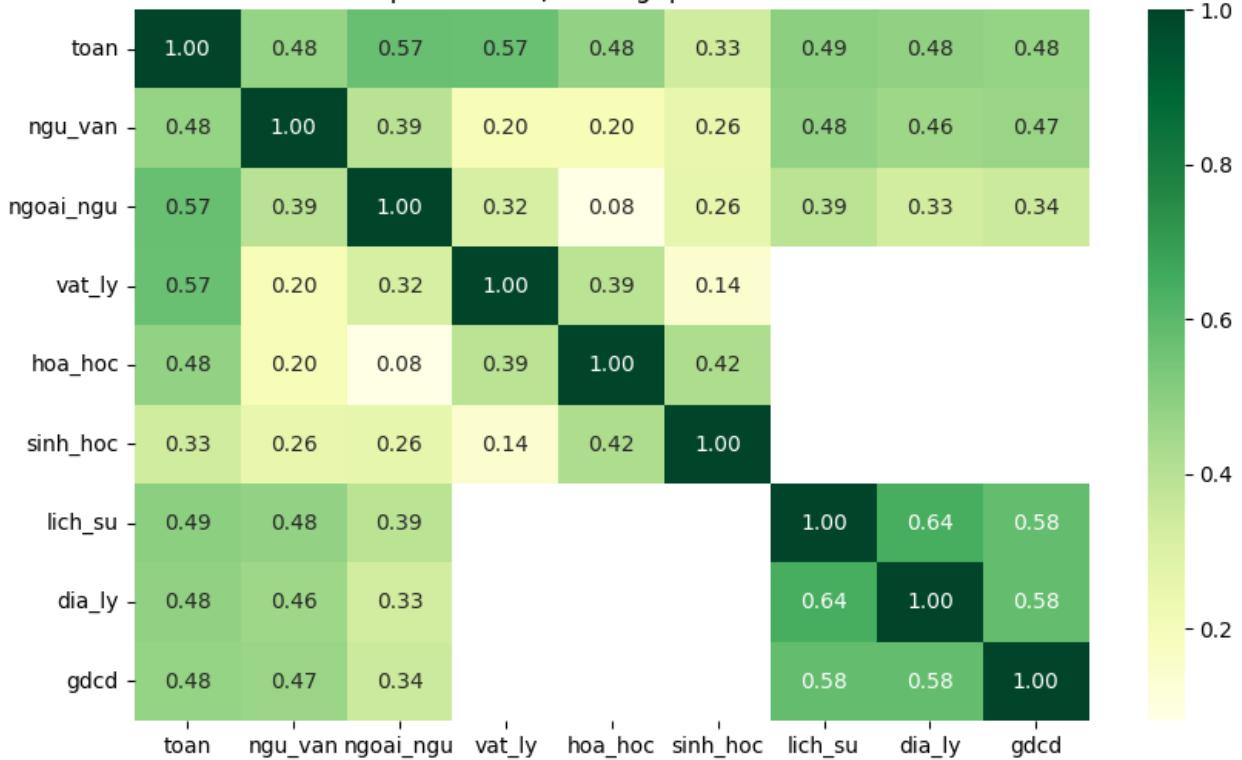
- **cmap='YlGn'**: Chỉ định bảng màu được sử dụng cho biểu đồ (màu **vàng-xanh lá**).
- **fmt='%.2f'**: Định dạng các giá trị hiển thị, ở đây hiển thị dưới dạng số thực với 2 chữ số thập phân.
- **vmin = -1**: Giá trị thấp nhất tương ứng với màu đầu tiên trong bảng màu.
- **linewidths**: Độ rộng của đường kẻ (lưới) ngăn cách giữa các ô trong heatmap.



Heatmap của ma trận tương quan điểm năm 2021



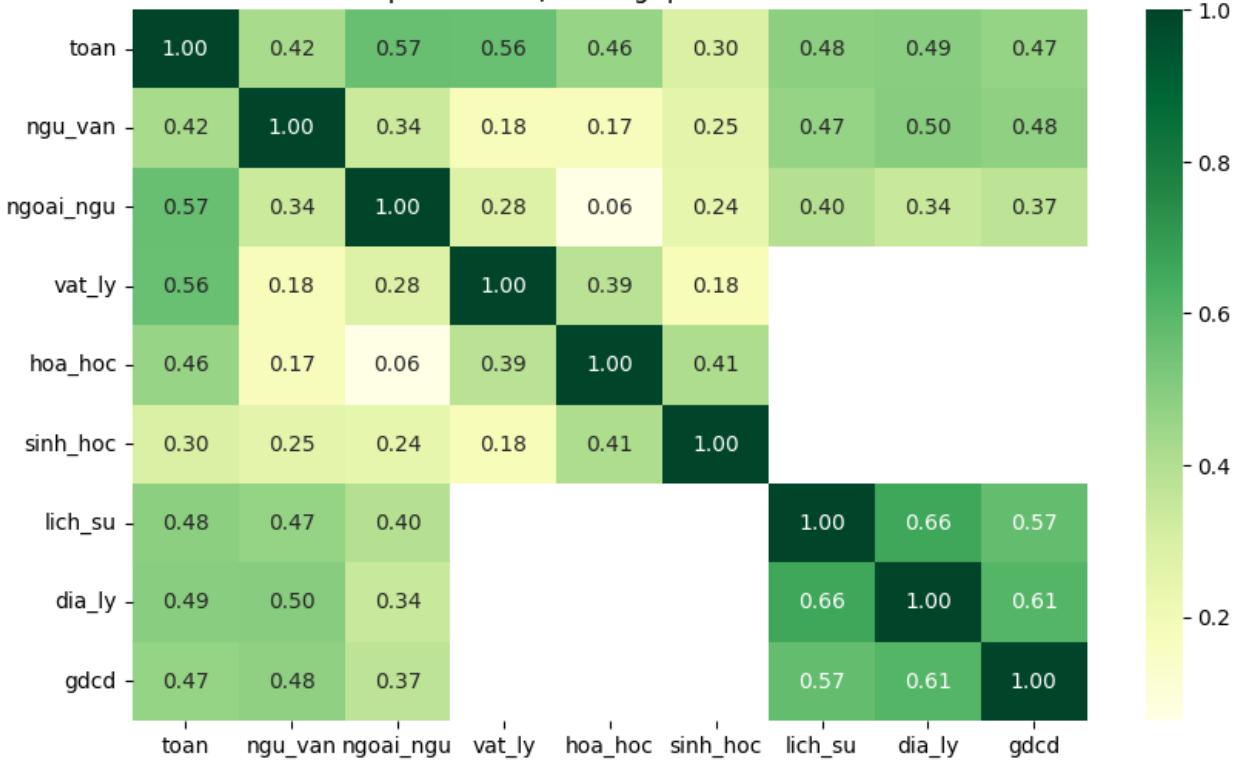
Heatmap của ma trận tương quan điểm năm 2022



Heatmap của ma trận tương quan điểm năm 2023



Heatmap của ma trận tương quan điểm năm 2024



- **Tương quan cao:**

- **Ngữ Văn - Lịch Sử - Địa Lý:** Đây là nhóm môn khối C, và mối tương quan cao giữa các môn này là điều dễ hiểu, do các môn này có liên quan về kỹ năng phân tích văn bản và tư duy logic trong xã hội.
 - **Vật Lý - Toán - Hóa Học:** Các môn thuộc khối A, A1 thường có mối tương quan cao, đặc biệt Toán với Vật Lý. Điều này hợp lý vì khả năng tư duy logic và tính toán được yêu cầu cao trong các môn này.
 - **GDCD - Địa Lý - Lịch Sử:** Mối liên kết chặt chẽ giữa các môn xã hội được duy trì khá ổn định qua các năm, đặc biệt ở các năm 2023 và 2024.
- **Tương quan thấp:**
 - **Ngoại Ngữ với các môn tự nhiên (Toán, Lý, Hóa):** Mối tương quan ở mức thấp hoặc trung bình. Điều này phản ánh rằng năng lực ngoại ngữ thường không quá liên quan đến các kỹ năng tính toán và tư duy logic.
 - **Sinh Học với các môn khác ngoài khối B (Toán, Hóa):** Sinh Học có mối tương quan trung bình hoặc thấp với các môn khối khác, trừ Hóa và Toán, do đặc thù môn học thiên về sinh lý và khoa học đời sống.

2. Xu hướng thay đổi tương quan qua các năm

- **Ôn định qua các năm:**
 - Các nhóm môn cùng khối (như Toán - Lý - Hóa trong khối A) duy trì mối tương quan ổn định qua các năm, cho thấy không có nhiều thay đổi lớn trong cách học và dạy.
 - Mối tương quan giữa Văn - Sử - Địa trong khối C cũng không thay đổi nhiều, thể hiện tính đồng nhất trong năng lực của học sinh ở các môn xã hội.
- **Biến động nhẹ:**
 - Môn Ngoại Ngữ có sự thay đổi về mức độ tương quan với các môn còn lại, đặc biệt với Toán và Văn. Có thể điều này phản ánh sự cải tiến trong cách giảng dạy Ngoại Ngữ qua các năm hoặc sự phân hóa năng lực giữa các nhóm học sinh.

3. Điểm khác biệt nổi bật trong từng năm

- **Năm 2020:** Tương quan giữa Văn và các môn tự nhiên (Toán, Lý) tương đối thấp, điều này có thể là do sự phân hóa mạnh trong giai đoạn học sinh vừa trải qua kỳ thi đặc biệt (COVID-19).
- **Năm 2024:** Tương quan giữa Lý - Ngoại Ngữ giảm nhẹ, trong khi Lịch Sử - Địa Lý tăng cao. Điều này có thể liên quan đến sự thay đổi chương trình hoặc cách học của học sinh trong các môn học xã hội.

4. Kết luận

- Các nhóm môn trong cùng khối thi vẫn duy trì mối tương quan mạnh, phản ánh sự ổn định trong phân chia khối và định hướng học tập.
- Môn Ngoại Ngữ có mối tương quan thấp với hầu hết các môn tự nhiên, nhưng vai trò của nó ngày càng được cải thiện, đặc biệt trong các năm gần đây.
- Mối liên kết giữa các môn xã hội (Văn - Sử - Địa - GD&CD) vẫn rất chặt chẽ, phù hợp với đặc thù của nhóm môn này.

II. Tương quan điểm số giữa các khối thi

1. Tương quan giữa các khối thi theo từng năm

```
# Thêm khối D07: Toán, Hóa Học, Ngoại Ngữ và tính tổng điểm cho các khối
for data in data_list:
    data['A'] = data['toan'] + data['vat_ly'] + data['hoa_hoc'] # Khối A: Toán, Vật Lý, Hóa Học
    data['A1'] = data['toan'] + data['vat_ly'] + data['ngoai_ngu'] # Khối A1: Toán, Vật Lý, Ngoại Ngữ
    data['B'] = data['toan'] + data['hoa_hoc'] + data['sinh_hoc'] # Khối B: Toán, Hóa Học, Sinh Học
    data['C'] = data['ngu_van'] + data['lich_su'] + data['dia_ly'] # Khối C: Ngữ Văn, Lịch Sử, Địa Lý
    data['D'] = data['toan'] + data['ngu_van'] + data['ngoai_ngu'] # Khối D: Toán, Ngữ Văn, Ngoại Ngữ
    data['D07'] = data['toan'] + data['hoa_hoc'] + data['ngoai_ngu'] # Khối D07: Toán, Hóa Học, Ngoại Ngữ

# Tạo khung hình lớn cho các heatmap
plt.figure(figsize=(20, 15))

# Danh sách tiêu đề cho từng năm
titles = ['Năm 2020', 'Năm 2021', 'Năm 2022', 'Năm 2023', 'Năm 2024']

# Vòng lặp qua các năm để tạo heatmap cho từng năm
for i, (data, title) in enumerate(zip(data_list, titles)):
    # Chọn các cột điểm số của các khối thi
    group_scores = data[['A', 'A1', 'B', 'C', 'D', 'D07']]

    # Tính toán ma trận tương quan
    correlation_matrix = group_scores.corr()

    # Tạo subplot (sắp xếp theo 2 hàng, 3 cột)
    plt.subplot(2, 3, i + 1)

    # Vẽ heatmap
    sns.heatmap(correlation_matrix, annot=True, cmap='YlGn', fmt='%.2f', vmin=-1, vmax=1, linewidths=0.5)

    # Thiết lập tiêu đề và nhãn
    plt.title(f'Heatmap của ma trận tương quan {title}')
    plt.xlabel('Khối thi')
    plt.ylabel('Khối thi')

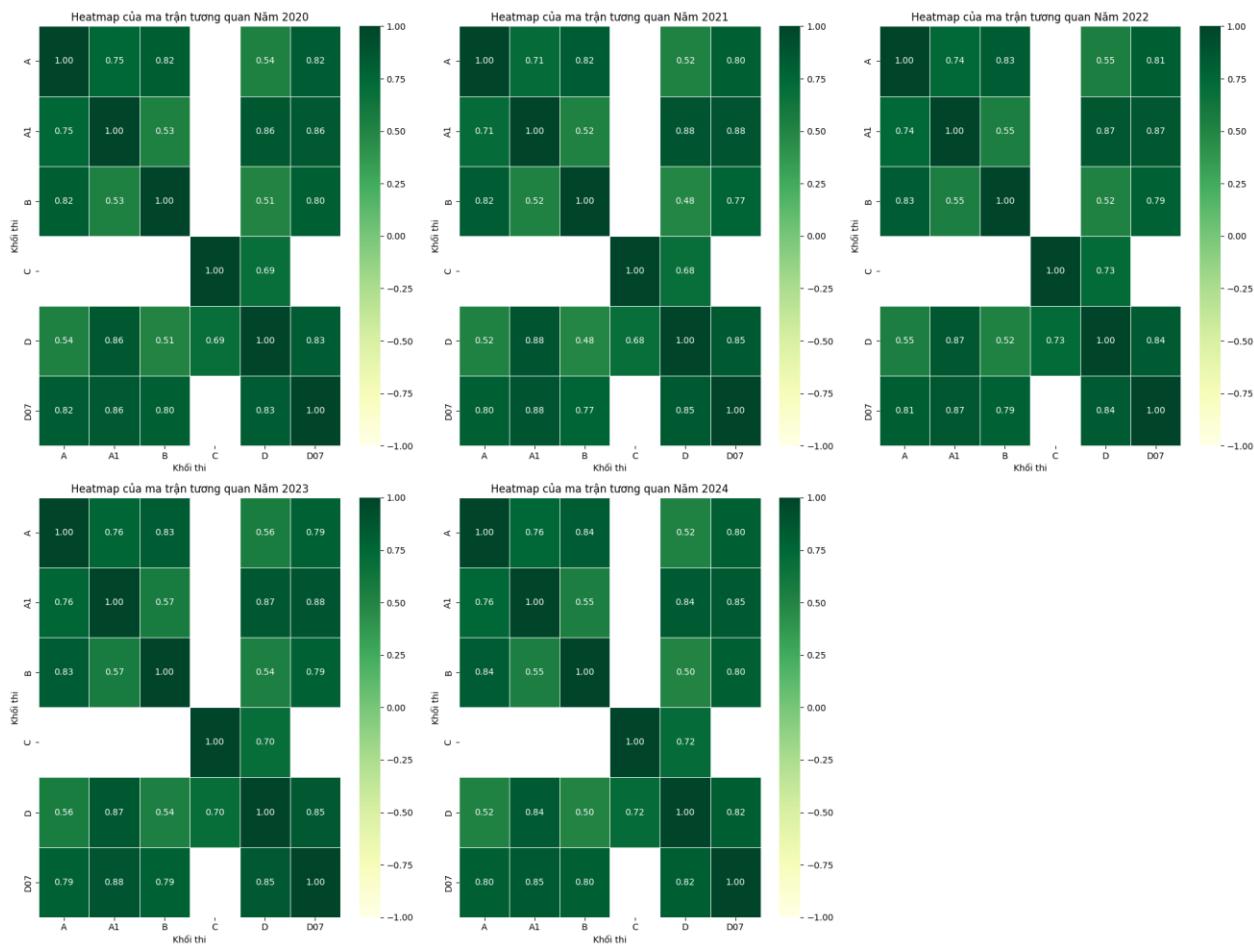
    # Điều chỉnh layout để các biểu đồ không bị chồng lấn
    plt.tight_layout()
plt.show()
```

group_scores: lấy dữ liệu DataFrame chỉ chứa điểm số của các môn theo từng khối.

correlation_matrix = group_score.corr(): Phương thức này tính toán ma trận tương quan giữa các cột trong DataFrame group_score.

`sns.heatmap(correlation_matrix, annot=True, cmap='YlGn', fmt='.2f', vmin=-1, linewidths=0.5):`

- **sns.heatmap:** Hàm vẽ biểu đồ nhiệt dựa trên ma trận tương quan.
- **correlation_matrix:** Dữ liệu để vẽ (ma trận đã tính ở trên).
- **annot=True:** Hiển thị giá trị tương quan trực tiếp trên từng ô trong biểu đồ.
- **cmap='YlGn':** Chỉ định bảng màu được sử dụng cho biểu đồ (màu **vàng-xanh lá**).
- **fmt='.2f':** Định dạng các giá trị hiển thị, ở đây hiển thị dưới dạng số thực với 2 chữ số thập phân.
- **vmin = -1:** Giá trị thấp nhất tương ứng với màu đầu tiên trong bảng màu.
- **linewidhts:** Độ rộng của đường kẻ (lưới) ngăn cách giữa các ô trong heatmap.



- **Năm 2020:**

- Khối A và A1 có tương quan cao nhất (**0.75**) trong các khối, cho thấy cấu trúc điểm của hai khối này tương đồng đáng kể.
- Khối C có mức tương quan thấp nhất với khối B (**0.69**), điều này có thể xuất phát từ sự khác biệt về môn thi giữa hai khối.

- Khối D và D07 có mức tương quan cao (**0.86**), cho thấy xu hướng đồng nhất về điểm giữa các môn thuộc hai khối này.
- Năm 2021:
 - Tương quan giữa **khối A và khối D07** đạt **0.88**, cao hơn năm 2020, phản ánh mối liên hệ tăng dần giữa hai khối.
 - Tuy nhiên, tương quan giữa **khối B và khối C** giảm từ **0.69** xuống **0.68**, thể hiện sự tách biệt về điểm số giữa các khối này.
 - **Khối A1 và B** có tương quan giảm nhẹ từ **0.53 (2020)** xuống **0.52**.
- Năm 2022:
 - Xu hướng tương tự các năm trước khi tương quan giữa khối D và D07 vẫn duy trì ở mức rất cao (**0.84**).
 - Tương quan giữa khối A và khối B tăng nhẹ từ **0.77 (2021)** lên **0.79**, cho thấy sự hội tụ điểm số giữa hai khối này.
- Năm 2023:
 - Tương quan giữa A1 và D07 vẫn cao (**0.88**), phản ánh sự ổn định trong cấu trúc điểm của các khối này.
 - Khối C duy trì mức tương quan trung bình với các khối khác, dao động trong khoảng từ **0.68 - 0.70**, cho thấy tính độc lập tương đối của khối này so với các khối còn lại.
- Năm 2024:
 - Khối D và D07 tiếp tục duy trì tương quan cao nhất (**0.85**).
 - Tương quan giữa A1 và B có xu hướng tăng nhẹ (**0.55**), cho thấy sự đồng nhất dần dần giữa hai khối này.

2. Phân tích theo khối thi

- **Khối A và A1:**
 - Luôn có mức tương quan cao trong suốt 5 năm, dao động từ **0.74 (2022)** đến **0.76 (2024)**.
 - Điều này phản ánh cấu trúc điểm số tương đồng do phần lớn môn thi tương tự nhau (Toán, Lý).
- **Khối B và C:**
 - Tương quan thấp nhất trong các cặp khối thi, dao động từ **0.68 - 0.73**, thể hiện sự khác biệt rõ rệt về cấu trúc điểm thi. Điều này có thể là do khối B tập trung vào các môn Khoa học Tự nhiên, trong khi khối C thiên về các môn Khoa học Xã hội.
- **Khối D và D07:**
 - Đây là cặp khối có tương quan cao nhất qua các năm (**0.83 - 0.87**), cho thấy xu hướng điểm số đồng đều. Sự tương đồng này có thể đến từ môn thi tiếng Anh và các môn Khoa học Tự nhiên như Toán, Hóa.

III. Nhận xét và đánh giá

1. Tính ổn định:

- Hầu hết các khối thi có mức độ tương quan ổn định qua các năm, phản ánh cấu trúc điểm số của các khối không có sự biến động lớn. Điều này có thể là do cơ chế chấm điểm và xu hướng học tập của học sinh không thay đổi quá nhiều.

2. Biến động:

- Một số cặp khối thi có sự thay đổi nhẹ về tương quan qua các năm, ví dụ:
 - **Khối A1 và B:** giảm nhẹ từ **0.53 (2020)** xuống **0.52 (2021)**, nhưng sau đó tăng lên **0.55 (2024)**.
 - **Khối A và B:** có sự gia tăng ổn định từ **0.77 (2021)** lên **0.80 (2024)**.

3. Đặc điểm từng khối thi:

- **Khối A và A1:** Tương quan ổn định và cao, thể hiện cấu trúc điểm thi đồng nhất.
- **Khối C:** Luôn có mức tương quan thấp với các khối khác, phản ánh đặc thù môn thi và xu hướng điểm số của khối này.
- **Khối D và D07:** Tương quan cao và ổn định nhất trong các cặp khối, minh chứng cho sự tương đồng về cách học và thi cử.

C.Phân tích theo địa phương

```

# Tính số lượng thí sinh mỗi năm theo khu vực
studen_region = data_all.groupby(['year', 'khu_vuc']).size().unstack()

# Tính tốc độ tăng trưởng
growth_region = studen_region.pct_change() * 100

growth_region = growth_region.apply(pd.to_numeric, errors='coerce')

# Xử lý các giá trị vô cùng và thay thế bằng NaN thay NaN bằng 0
growth_region.replace([np.inf, -np.inf], np.nan, inplace=True)
growth_region.fillna(0, inplace=True)

plt.figure(figsize=(12, 4))

# Duyệt qua từng khu vực để vẽ đường cho mỗi khu vực
for region in growth_region.columns:
    x = growth_region.index
    y = growth_region[region]

    x_num = pd.to_numeric(x, errors='coerce')

    # làm mịn đường
    x_new = np.linspace(x_num.min(), x_num.max(), 300)
    spl = make_interp_spline(x_num, y, k=3)
    y_smooth = spl(x_new)

    # Vẽ đường
    plt.plot(x_new, y_smooth, label=region, linewidth=2, alpha=0.7)

plt.title('Tốc độ tăng trưởng số lượng thí sinh theo khu vực')
plt.xlabel('Năm')
plt.ylabel('Tốc độ tăng trưởng (%)')
plt.legend(loc='upper left', bbox_to_anchor=(1.05, 1))
plt.grid(False)
plt.xticks(x_num)
plt.tight_layout()
plt.show()
]

```

studen_region = data_all.groupby(['year', 'khu_vuc']).size().unstack()

- **data_all:** DataFrame chứa dữ liệu gốc về số lượng thí sinh, bao gồm các cột như year (năm) và khu_vuc (khu vực).
- **groupby(['year', 'khu_vuc']):** Nhóm dữ liệu theo year và khu_vuc để tính toán cho từng nhóm (kết hợp năm và khu vực).

- **.size()**: Đếm số lượng thí sinh trong mỗi nhóm (số hàng thuộc về mỗi year và khu_vuc).
- **.unstack()**: Chuyển đổi từ dạng nhóm (multi-index) thành bảng 2D, với các hàng là year và các cột là khu_vuc.

```
growth_region = studen_region.pct_change() * 100
```

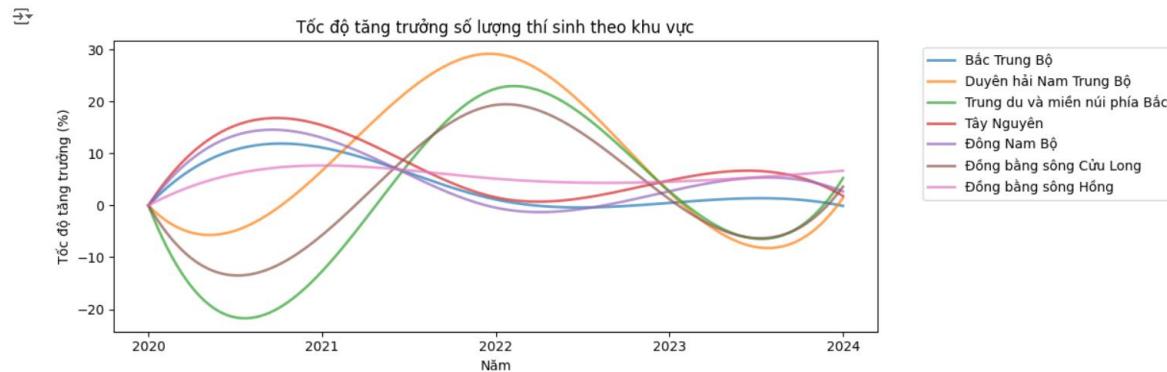
- **.pct_change()**: Tính phần trăm thay đổi giữa các hàng liên tiếp trong DataFrame (so sánh số liệu giữa các năm).
 - Công thức: $(\text{Năm hiện tại} - \text{Năm trước})/\text{Năm trước} * 100$
- **100**: Chuyển đổi từ dạng tỷ lệ (0.1 = 10%) thành dạng phần trăm.

```
growth_region = growth_region.apply(pd.to_numeric, errors='coerce')
growth_region.replace([np.inf, -np.inf], np.nan, inplace=True)
```

```
growth_region.fillna(0, inplace=True)
```

- **apply(pd.to_numeric, errors='coerce')**: Chuyển đổi tất cả các giá trị về dạng số, thay các giá trị không thể chuyển đổi (chuỗi, ký hiệu lạ) bằng NaN.
- **replace([np.inf, -np.inf], np.nan, inplace=True)**: Thay thế các giá trị vô cùng (inf và -inf) bằng NaN.
- **fillna(0, inplace=True)**: Thay tất cả các giá trị NaN bằng 0.

*Tốc độ tăng trưởng thí sinh theo từng khu vực



I. Phân tích chi tiết theo từng khu vực

1. Bắc Trung Bộ (màu xanh lam nhạt):

- **Xu hướng:** Tốc độ tăng trưởng khá ổn định từ năm 2020 đến 2024, dao động quanh mức 0-10%.
- **Điểm nổi bật:**
 - Giai đoạn 2020-2022, khu vực này có sự tăng trưởng nhẹ, đạt đỉnh vào khoảng năm 2021 (~10%).
 - Sau đó, tốc độ tăng trưởng giảm dần và quay về gần mức 0 vào năm 2024.

- **Đánh giá:** Đây là khu vực có tốc độ tăng trưởng ổn định, không có sự biến động mạnh.

2. Duyên hải Nam Trung Bộ (màu cam):

- **Xu hướng:** Khu vực này có mức biến động lớn nhất trong biểu đồ.
- **Điểm nổi bật:**
 - Giai đoạn 2020-2021, tốc độ tăng trưởng đạt đỉnh cao nhất (~30%), chứng tỏ sự gia tăng mạnh số lượng thí sinh trong năm này.
 - Giai đoạn 2022-2023, tốc độ tăng trưởng giảm mạnh, rơi xuống âm (~-20%).
 - Năm 2024, phục hồi nhẹ, quay lại mức xấp xỉ 0%.
- **Đánh giá:** Đây là khu vực có sự biến động mạnh mẽ nhất trong biểu đồ, có thể do những thay đổi trong chính sách giáo dục hoặc điều kiện kinh tế - xã hội ở khu vực này.

3. Trung du và miền núi phía Bắc (màu xanh lá):

- **Xu hướng:** Có mức tăng trưởng thấp và biến động mạnh nhất về phía âm.
- **Điểm nổi bật:**
 - Năm 2020-2021: Tốc độ tăng trưởng giảm sâu xuống ~-20%.
 - Sau đó, tốc độ phục hồi dần từ năm 2022 và quay lại mức 0% vào năm 2024.
- **Đánh giá:** Khu vực này có mức giảm mạnh nhất trong giai đoạn đầu, cho thấy sự giảm sút số lượng thí sinh, có thể do điều kiện địa lý khó khăn hoặc xu hướng chuyển đổi lao động.

4. Tây Nguyên (màu đỏ):

- **Xu hướng:** Tốc độ tăng trưởng tăng mạnh và đạt đỉnh giữa giai đoạn 2021 (~20%).
- **Điểm nổi bật:**
 - Sau năm 2021, tốc độ giảm dần và quay về gần mức 0% vào năm 2024.
- **Đánh giá:** Khu vực Tây Nguyên có sự tăng trưởng tích cực trong giai đoạn đầu, nhưng giảm dần về cuối.

5. Đông Nam Bộ (màu tím):

- **Xu hướng:** Khu vực này có mức tăng trưởng ổn định nhất trong tất cả các khu vực.
- **Điểm nổi bật:**
 - Duy trì tốc độ tăng trưởng trong khoảng 5-10% suốt từ 2020 đến 2024.
- **Đánh giá:** Đông Nam Bộ là khu vực có sự ổn định đáng kể trong tăng trưởng số lượng thí sinh.

6. Đồng bằng sông Cửu Long (màu nâu):

- **Xu hướng:** Có sự dao động nhẹ quanh mức 0%.
- **Điểm nổi bật:**
 - Giai đoạn 2020-2022: Tốc độ tăng trưởng chậm, giảm nhẹ trong một vài năm.
 - Giai đoạn 2023-2024: Phục hồi nhẹ nhưng vẫn ở mức thấp.
- **Đánh giá:** Tăng trưởng chậm và không có biến động lớn, thể hiện sự bão hòa trong số lượng thí sinh tại khu vực này.

7. Đồng bằng sông Hồng (màu hồng):

- **Xu hướng:** Tốc độ tăng trưởng thấp, ổn định và có xu hướng tăng dần.
- **Điểm nổi bật:**
 - Từ 2020-2021: Gần như không có tăng trưởng.
 - Giai đoạn 2022-2024: Tốc độ tăng trưởng cải thiện nhẹ, đạt mức khoảng 5% vào năm 2024.
- **Đánh giá:** Đây là khu vực có sự tăng trưởng đều và ít biến động, cho thấy sự ổn định về số lượng thí sinh.

II. Nhận xét và đánh giá tổng quan

1. Khu vực có sự tăng trưởng mạnh nhất:

- **Duyên hải Nam Trung Bộ** đạt đỉnh tốc độ tăng trưởng (~30%) vào năm 2021, tuy nhiên sau đó lại giảm mạnh.
- **Tây Nguyên** cũng là khu vực có tăng trưởng tích cực trong giai đoạn đầu.

2. Khu vực ổn định:

- **Đông Nam Bộ** là khu vực có tốc độ tăng trưởng ổn định nhất (5-10% mỗi năm).
- **Đồng bằng sông Hồng** cũng có xu hướng tăng trưởng đều đặn trong các năm.

3. Khu vực có biến động tiêu cực:

- **Trung du và miền núi phía Bắc** có mức giảm mạnh nhất (~-20%) vào giai đoạn 2020-2021, nhưng đã phục hồi dần vào các năm sau.

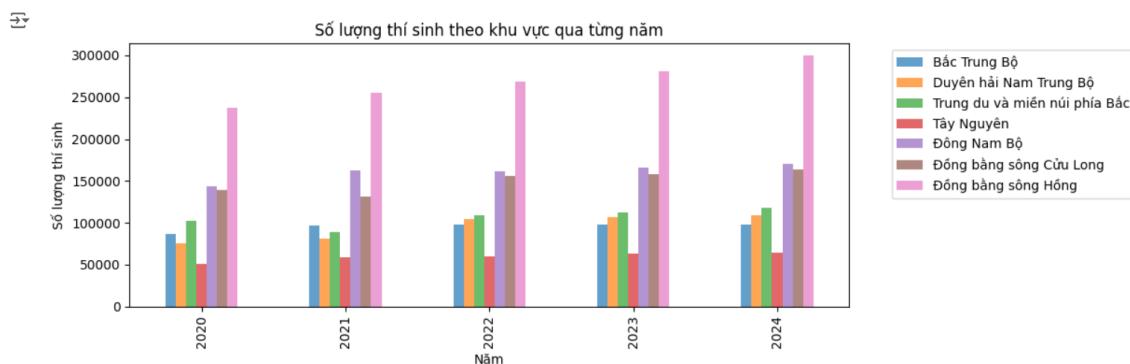
4. Những khu vực cần chú ý:

- Sự biến động mạnh tại **Duyên hải Nam Trung Bộ** có thể là do các yếu tố khách quan như chính sách giáo dục hoặc sự biến đổi kinh tế - xã hội.
- Khu vực **Trung du và miền núi phía Bắc** cho thấy thách thức lớn trong việc duy trì số lượng thí sinh, cần có các biện pháp hỗ trợ để khuyến khích thí sinh ở khu vực này.

III. Kết luận

Biểu đồ phản ánh sự khác biệt đáng kể trong xu hướng tăng trưởng số lượng thí sinh giữa các khu vực. Một số khu vực như Đông Nam Bộ và Đồng bằng sông Hồng duy trì ổn định, trong khi các khu vực như Duyên hải Nam Trung Bộ và Trung du miền núi phía Bắc có sự biến động mạnh. Điều này nhấn mạnh tầm quan trọng của các chiến lược giáo dục phù hợp với từng khu vực nhằm duy trì và tăng cường số lượng thí sinh dự thi.

*Số lượng thí sinh theo khu vực qua từng năm



1. Khu vực Đồng bằng sông Hồng (màu hồng)

- Đây là khu vực có số lượng thí sinh cao nhất qua mọi năm.
- Năm 2020, số lượng thí sinh đạt **xấp xỉ 250.000**.
- Từ năm 2021 đến 2024, số lượng thí sinh tăng nhẹ và duy trì ở mức **trên 270.000**.
- Điều này phản ánh đặc trưng về dân số đông và sự phát triển của hệ thống giáo dục tại khu vực này.

2. Khu vực Đồng bằng sông Cửu Long (màu nâu)

- Đồng bằng sông Cửu Long cũng duy trì số lượng thí sinh cao, đứng thứ hai trong các khu vực.
- Số lượng thí sinh dao động trong khoảng **180.000 - 200.000** qua các năm.
- Tuy nhiên, từ năm 2023 đến 2024, số liệu cho thấy một mức giảm nhẹ, từ **190.000** xuống khoảng **185.000**.

3. Khu vực Đông Nam Bộ (màu tím)

- Số lượng thí sinh của khu vực này khá ổn định và dao động trong khoảng **160.000 - 180.000**.
- Có sự gia tăng nhẹ từ năm 2020 (khoảng **165.000**) lên năm 2024 (khoảng **175.000**).

4. Khu vực Tây Nguyên (màu đỏ)

- Tây Nguyên là khu vực có số lượng thí sinh thấp nhất trong tất cả các năm, dao động từ **40.000** đến **60.000**.

- Mặc dù dân số ít hơn so với các khu vực khác, nhưng sự chênh lệch này cũng phản ánh điều kiện giáo dục chưa đồng đều.

5. Khu vực Trung du và miền núi phía Bắc (màu xanh lá cây)

- Khu vực này duy trì số lượng thí sinh trong khoảng **70.000 - 90.000**.
- Số lượng thí sinh có xu hướng giảm nhẹ qua các năm, từ **85.000** năm 2020 xuống khoảng **75.000** năm 2024.

6. Các khu vực Bắc Trung Bộ (màu xanh dương) và Duyên hải Nam Trung Bộ (màu cam)

- Cả hai khu vực đều có số lượng thí sinh ở mức trung bình so với cả nước:
 - Bắc Trung Bộ dao động từ **90.000** (năm 2020) lên khoảng **100.000** (năm 2024).
 - Duyên hải Nam Trung Bộ duy trì trong khoảng **100.000 - 110.000** qua các năm.
- Điều này cho thấy sự ổn định về quy mô dân số và sự tham gia giáo dục tại các khu vực này.

Nhận xét tổng quan

- Khu vực có số lượng thí sinh vượt trội:** Đồng bằng sông Hồng dẫn đầu qua mọi năm, với số lượng thí sinh cao gấp đôi so với một số khu vực như Tây Nguyên hoặc Trung du và miền núi phía Bắc.
- Khu vực có số lượng thí sinh thấp:** Tây Nguyên liên tục ở mức thấp nhất, phản ánh sự hạn chế trong tiếp cận giáo dục và điều kiện dân cư.
- Xu hướng qua các năm:**
 - Tổng thể, số lượng thí sinh ở tất cả các khu vực không biến động quá lớn.
 - Một số khu vực như Đồng bằng sông Cửu Long và Trung du miền núi phía Bắc có xu hướng giảm nhẹ.

Kết luận và đề xuất

- Đồng bằng sông Hồng và các khu vực như Đồng bằng sông Cửu Long cần tiếp tục được đầu tư để duy trì chất lượng giáo dục, đáp ứng nhu cầu số lượng lớn thí sinh.
- Các khu vực có số lượng thí sinh thấp như Tây Nguyên nên được chú trọng phát triển về cơ sở hạ tầng giáo dục và chính sách hỗ trợ để khuyến khích sự tham gia của học sinh.
- Việc phân bổ tài nguyên giáo dục cần được điều chỉnh hợp lý để thu hẹp khoảng cách giữa các khu vực.

*Chi tiết từng khu vực

1.Khu vực trung du miền núi phía Bắc

year	dia_phuong	2020	2021	2022	2023	2024	Ti_lệ_21	Ti_lệ_22	Ti_lệ_23	Ti_lệ_24
7	Phú Thọ	13810.000000	16151.000000	15808.000000	15874.000000	16382.000000	17%	-2%	0%	3%
9	Thái Nguyên	15190.000000	14840.000000	15083.000000	16153.000000	16713.000000	-2%	2%	7%	3%
6	Lạng Sơn	8510.000000	9530.000000	9608.000000	9353.000000	9469.000000	12%	1%	-3%	1%
3	Hòa Bình	9151.000000	9478.000000	9575.000000	9690.000000	9823.000000	4%	1%	1%	1%
10	Tuyên Quang	7995.000000	8680.000000	8425.000000	8392.000000	8984.000000	9%	-3%	-0%	7%
5	Lào Cai	6355.000000	7331.000000	7204.000000	7866.000000	8356.000000	15%	-2%	9%	6%
12	Điện Biên	5608.000000	6268.000000	6413.000000	6685.000000	7216.000000	12%	2%	4%	8%
2	Hà Giang	5627.000000	5603.000000	5863.000000	6286.000000	6920.000000	-0%	5%	7%	10%
1	Cao Bằng	4598.000000	4738.000000	5043.000000	5046.000000	5506.000000	3%	6%	0%	9%
4	Lai Châu	3298.000000	3563.000000	3662.000000	3838.000000	4188.000000	8%	3%	5%	9%
0	Bắc Kạn	2851.000000	2893.000000	3053.000000	2917.000000	3174.000000	1%	6%	-4%	9%
8	Sơn La	11626.000000	nan	11325.000000	11773.000000	12584.000000	nan%	nan%	4%	7%
11	Yên Bai	7404.000000	nan	8048.000000	8235.000000	8687.000000	nan%	nan%	2%	5%

Khu vực này bao gồm các địa phương như Phú Thọ, Thái Nguyên, Sơn La, Yên Bai, và các tỉnh miền núi khác. Số liệu thể hiện xu hướng tăng trưởng khác nhau qua các năm từ 2020 đến 2024.

Xu hướng số lượng thí sinh qua các năm:

- Phú Thọ:** Tăng trưởng ổn định, từ **13.810 thí sinh (2020)** lên **16.382 thí sinh (2024)**, tương ứng tăng **19%** trong 4 năm.
- Thái Nguyên:** Biến động nhẹ với mức giảm năm 2021 (-2%) nhưng tăng trưởng trở lại, đạt **16.713 thí sinh năm 2024**, tăng tổng cộng **10%** so với năm 2020.
- Sơn La:** Số liệu bị thiếu năm 2021-2022 nhưng đạt **12.584 thí sinh** năm 2024, tăng **7%** so với 2023.
- Các tỉnh như **Hòa Bình, Tuyên Quang, và Lạng Sơn** có xu hướng tăng trưởng nhẹ, dao động trong khoảng từ **1-7%** qua các năm.

Địa phương có sự tăng trưởng nổi bật:

- Điện Biên và Hà Giang** có mức tăng trưởng đáng kể, đặc biệt năm 2024 tăng lần lượt **8%** và **10%**, phản ánh nỗ lực cải thiện giáo dục tại các tỉnh vùng sâu, vùng xa.

Địa phương có số lượng thấp:

- Bắc Kạn và Lai Châu** duy trì số lượng thấp nhất trong khu vực, lần lượt đạt **3.174** và **4.188 thí sinh** năm 2024. Tuy nhiên, cả hai địa phương này vẫn cho thấy sự tăng trưởng tích cực qua từng năm.

2. Khu vực Đồng bằng sông Hồng

year	dia_phuong	2020	2021	2022	2023	2024	Ti_lệ_21	Ti_lệ_22	Ti_lệ_23	Ti_lệ_24
3	Hà Nội	78860	100691	96942	102095	107867	28%	-4%	5%	6%
6	Hải Phòng	18533	23241	22399	22723	25530	25%	-4%	1%	12%
9	Thái Bình	19592	22858	20145	20845	22580	17%	-12%	3%	8%
5	Hải Dương	19583	22263	20422	21934	23366	14%	-8%	7%	7%
7	Nam Định	18667	20917	19769	20405	21760	12%	-5%	3%	7%
8	Quảng Ninh	14547	16367	15624	16025	17848	13%	-5%	3%	11%
1	Bắc Ninh	14621	16336	15850	16724	17614	12%	-3%	6%	5%
10	Vĩnh Phúc	12171	14002	13873	14042	15461	15%	-1%	1%	10%
2	Hà Nam	8575	9670	9183	9635	9564	13%	-5%	5%	-1%
4	Hưng Yên	12780	8217	13885	15386	16329	-36%	69%	11%	6%
0	Bắc Giang	19298	932	20553	21080	21755	-95%	2105%	3%	3%

Đây là khu vực tập trung đông dân cư và phát triển giáo dục, bao gồm các địa phương như Hà Nội, Hải Phòng, Bắc Ninh, và Thái Bình.

Xu hướng số lượng thí sinh qua các năm:

- Hà Nội:** Dẫn đầu toàn khu vực với sự gia tăng từ **78.860 thí sinh (2020)** lên **107.867 thí sinh (2024)**, tương ứng tăng **37%** trong 4 năm.
- Hải Phòng:** Số lượng thí sinh tăng đều qua các năm, từ **18.533 (2020)** lên **25.530 (2024)**, tương ứng tăng **38%**, phản ánh tốc độ phát triển giáo dục mạnh mẽ.
- Nam Định, Hải Dương, Thái Bình:** Các tỉnh này duy trì mức tăng trưởng ổn định khoảng **7-8%**, riêng Nam Định đạt **21.760 thí sinh** năm 2024.

Địa phương có sự bất thường:

- Hưng Yên:** Số liệu cho thấy sự giảm mạnh vào năm 2021 (-36%) nhưng phục hồi nhanh chóng với mức tăng **11%** vào năm 2023 và tiếp tục duy trì tăng trưởng năm 2024.
- Bắc Giang:** Ghi nhận một bất thường lớn vào năm 2021 (giảm tới -95%), nhưng sau đó tăng mạnh trở lại vào năm 2022 và ổn định vào năm 2023-2024.

Địa phương nổi bật:

- Hà Nội và Hải Phòng:** Là hai địa phương có mức tăng trưởng cả về số lượng và tỷ lệ đáng chú ý, đóng vai trò trọng tâm trong phát triển giáo dục toàn vùng.

3.Khu vực Bắc Trung Bộ

year	dia_phuong	2020	2021	2022	2023	2024	Ti_lệ_21	Ti_lệ_22	Ti_lệ_23	Ti_lệ_24
1	Nghệ An	31072	34327	36564	36886	36729	10%	7%	1%	-0%
0	Hà Tĩnh	15308	17268	17304	17249	17004	13%	0%	-0%	-1%
5	Thừa Thiên -Huế	12514	13291	13296	13029	12774	6%	0%	-2%	-2%
2	Quảng Bình	10658	11902	11115	11168	11189	12%	-7%	0%	0%
4	Thanh Hóa	9219	11050	10501	11149	11749	20%	-5%	6%	5%
3	Quảng Trị	7897	8485	8643	8413	8359	7%	2%	-3%	-1%

Nghệ An:

Tăng trưởng ổn định với mức tăng đáng kể vào 2021 (+10%) và 2022 (+7%).

Tuy nhiên, số lượng thí sinh chững lại vào năm 2023 (+1%) và giảm nhẹ (-0%) vào 2024, phản ánh xu hướng bão hòa sau giai đoạn phục hồi mạnh.

Hà Tĩnh:

Năm 2021 ghi nhận mức tăng trưởng ấn tượng (+13%), nhưng từ 2022 đến 2024 tốc độ tăng trưởng giảm rõ rệt (-0% đến -1%).

Điều này có thể cho thấy tác động từ dân số hoặc số lượng thí sinh đến tuổi thi giảm.

Thừa Thiên-Huế:

Số lượng thí sinh ổn định trong các năm đầu (2021: +6%), nhưng dần giảm (-2% vào 2023 và 2024).

Đây có thể là dấu hiệu của sự cạnh tranh hoặc giảm nhu cầu thi tuyển tại khu vực này.

Quảng Bình:

Mức tăng trưởng tốt vào 2021 (+12%), nhưng sau đó suy giảm (-7% vào 2022) và gần như không đổi vào 2024 (0%).

Điều này thể hiện một giai đoạn giảm sức hút đối với các thí sinh.

Thanh Hóa:

Tăng trưởng đột phá năm 2021 (+20%), nhưng lại giảm nhẹ vào 2022 (-5%). Từ năm 2023, số lượng thí sinh tăng trở lại ở mức ổn định (5%-6%).

Đây là tỉnh duy trì sự phục hồi tốt nhất trong nhóm.

Quảng Trị:

Tăng trưởng nhẹ qua các năm, dao động từ +7% (2021) đến +2% (2023), nhưng giảm nhẹ vào 2024 (-1%).

Điều này thể hiện sự ổn định, mặc dù có dấu hiệu giảm trong dài hạn.

4. Duyên hải Nam Trung Bộ

year	dia_phuong	2020	2021	2022	2023	2024	Ti_lê_21	Ti_lê_22	Ti_lê_23	Ti_lê_24
5	Quảng Nam	7280.000000	16701.000000	16739.000000	17197.000000	17292.000000	129%	0%	3%	1%
1	Bình Định	16876.000000	14551.000000	18489.000000	18897.000000	19342.000000	-14%	27%	2%	2%
2	Khánh Hòa	13094.000000	13847.000000	14025.000000	14487.000000	14388.000000	6%	1%	3%	-1%
7	Đà Nẵng	nan	12631.000000	12575.000000	13133.000000	13414.000000	nan%	-0%	4%	2%
0	Bình Thuận	10893.000000	12036.000000	12727.000000	12900.000000	13133.000000	10%	6%	1%	2%
3	Ninh Thuận	5717.000000	5972.000000	5875.000000	6121.000000	6288.000000	4%	-2%	4%	3%
4	Phú Yên	10078.000000	3549.000000	10874.000000	10666.000000	10622.000000	-65%	206%	-2%	-0%
6	Quảng Ngãi	11966.000000	1696.000000	13295.000000	13870.000000	14354.000000	-86%	684%	4%	3%

- Quảng Nam:

- Tăng trưởng đột biến vào năm 2021 (+129%) do một cú hích lớn (có thể là chính sách giáo dục hoặc sự kiện thúc đẩy).

- Từ 2022 đến 2024, ổn định hơn với mức tăng nhẹ (+1% đến +3%).

- Điều này cho thấy sự phục hồi và duy trì tốt ở khu vực.

- Bình Định:

- Sụt giảm vào năm 2021 (-14%) nhưng phục hồi mạnh vào 2022 (+27%).

- Từ 2023 đến 2024, tăng trưởng chậm lại (+2%), cho thấy tỉnh đã ổn định sau biến động ban đầu.

- Khánh Hòa:

- Tăng trưởng ổn định qua các năm, dao động từ -1% đến +6%.

- Đây là một trong những tỉnh có xu hướng bền vững, không biến động lớn.

- Đà Nẵng:

- Thiếu dữ liệu năm 2020, nhưng từ 2021 trở đi, số lượng thí sinh tăng trưởng nhẹ qua các năm (-0% đến +4%).
- Là khu vực có sự ổn định, nhưng không có cú hích mạnh.

- Bình Thuận:

- Duy trì mức tăng ổn định, dao động từ +1% đến +10%.
- Mức tăng giảm dần vào năm 2024 (+2%), phản ánh dấu hiệu bão hòa.

- Ninh Thuận:

- Tăng trưởng khiêm tốn, dao động từ -2% đến +4%.
- Đây là tỉnh có mức tăng trưởng ổn định nhưng chậm.

- Phú Yên:

- Biến động mạnh: sụt giảm nghiêm trọng năm 2021 (-65%) nhưng tăng đột phá vào 2022 (+206%).
- Từ năm 2023 trở đi, tăng trưởng gần như dừng lại (-2% đến -0%), cho thấy tỉnh đã đạt ngưỡng giới hạn.

- Quảng Ngãi:

- Biến động cực lớn: giảm mạnh vào 2021 (-86%) nhưng phục hồi mạnh năm 2022 (+684%).
- Sau đó, tăng trưởng chậm lại ở mức +3% đến +4%, biểu hiện của sự ổn định trở lại.

5.Tây Nguyên

year	dia_phuong	2020	2021	2022	2023	2024	Ti_lệ_21	Ti_lệ_22	Ti_lệ_23	Ti_lệ_24
4	Đăk Lăk	13945	19767	20348	21042	20865	42%	3%	3%	-1%
2	Lâm Đồng	13497	14169	14102	14687	15372	5%	-0%	4%	5%
0	Gia Lai	13315	13937	14075	14861	15201	5%	1%	6%	2%
3	Đăk Nông	6212	6679	6852	7384	7681	8%	3%	8%	4%
1	Kon Tum	4294	4631	4730	5028	5038	8%	2%	6%	0%

- **Đăk Lăk:**

- Năm 2021 tăng mạnh nhất với 42% so với năm 2020. Các năm tiếp theo tăng trưởng chậm hơn, dao động từ 3%-3%, và giảm nhẹ -1% vào năm 2024.

- **Lâm Đồng:**

- Năm 2022 có mức tăng trưởng gần như bằng 0% (-0%), nhưng đã phục hồi mạnh mẽ với 4% (2023) và 5% (2024).

- **Gia Lai:**

- Có xu hướng tăng trưởng đều đặn qua các năm, dao động từ 1%-6%, cho thấy sự ổn định trong phát triển giáo dục.

- **Đăk Nông:**

- Có mức tăng trưởng tương đối đồng đều từ 3%-8%, đặc biệt năm 2021 và 2024 đạt mức tăng cao (8%).

- **Kon Tum:**

- Mặc dù có số lượng thí sinh thấp nhất trong khu vực, Kon Tum vẫn duy trì tăng trưởng ổn định ở mức 8% (2021), sau đó giảm nhẹ về 0% vào năm 2024.

Nhận xét và kết luận:

- **Xu hướng chung:** Khu vực Tây Nguyên có xu hướng tăng trưởng số lượng thí sinh qua các năm, nhưng mức độ tăng không đồng đều giữa các tỉnh. Các tỉnh lớn như Đăk Lăk, Gia Lai và Lâm Đồng dẫn đầu về số lượng và có vai trò quan trọng trong sự phát triển của khu vực.
- **Khó khăn và thách thức:** Một số tỉnh (như Đăk Lăk, Lâm Đồng) đã có dấu hiệu chững lại hoặc giảm nhẹ, điều này có thể phản ánh các yếu tố bên ngoài như dân số học, hệ thống giáo dục, hoặc cơ hội tiếp cận giáo dục.
- **Hướng phát triển:** Các tỉnh như Đăk Nông và Kon Tum cần tận dụng đà tăng trưởng hiện tại, đồng thời cần đầu tư nhiều hơn vào cơ sở vật chất và chính sách giáo dục để thu hút thêm học sinh tham gia.

6.Đồng bằng sông Cửu Long

year	dia_phuong	2020	2021	2022	2023	2024	Ti_lệ_21	Ti_lệ_22	Ti_lệ_23	Ti_lệ_24
7	Long An	13850.000000	16353.000000	15321.000000	15653.000000	15771.000000	+18%	-6%	+2%	+1%
9	Tiền Giang	14283.000000	16202.000000	15028.000000	15455.000000	16273.000000	+13%	-7%	+3%	+5%
6	Kiên Giang	12434.000000	13700.000000	14325.000000	14423.000000	15035.000000	+10%	+5%	+1%	+4%
2	Bến Tre	11313.000000	13178.000000	12301.000000	12117.000000	12145.000000	+16%	-7%	-1%	0%
4	Cần Thơ	10648.000000	12113.000000	12195.000000	12061.000000	12841.000000	+14%	+1%	-1%	+6%
11	Vĩnh Long	10215.000000	11327.000000	10327.000000	10435.000000	10913.000000	+11%	-9%	+1%	+5%
3	Cà Mau	9637.000000	10914.000000	10692.000000	9776.000000	10151.000000	+13%	-2%	-9%	+4%
8	Sóc Trăng	8514.000000	9625.000000	9985.000000	10045.000000	10642.000000	+13%	+4%	+1%	+6%
10	Trà Vinh	7890.000000	8806.000000	9461.000000	9196.000000	9686.000000	+12%	+7%	-3%	+5%
5	Hậu Giang	5813.000000	6815.000000	6549.000000	6937.000000	7261.000000	+17%	-4%	+6%	+5%
1	Bạc Liêu	5589.000000	6332.000000	6413.000000	6382.000000	6384.000000	+13%	+1%	-0%	0%
0	An Giang	15239.000000	5505.000000	18574.000000	19942.000000	20251.000000	-64%	+237%	+7%	+2%
12	Đồng Tháp	13386.000000	nan	15038.000000	15678.000000	16464.000000	nan%	nan%	+4%	+5%

a. Tỉnh có xu hướng tăng trưởng ổn định

- Long An: Tăng trưởng khá tốt năm 2021 (+18%), sau đó giảm nhẹ năm 2022 (-6%), và phục hồi dần vào các năm tiếp theo (+2% năm 2023, +1% năm 2024). Sự ổn định cho thấy tỉnh có khả năng duy trì nguồn lực thí sinh đều đặn.
- Tiền Giang: Mức tăng trưởng khá mạnh vào 2021 (+13%), giảm vào 2022 (-7%), nhưng dần phục hồi vào 2023 (+3%) và 2024 (+5%). Phản ánh sự phát triển bền vững trong khu vực.
- Cần Thơ: Tăng trưởng mạnh mẽ vào 2021 (+14%), tuy giảm nhẹ vào 2022 (-7%) và 2023 (-1%), nhưng đạt mức tăng trưởng tốt trở lại vào năm 2024 (+6%). Là một trung tâm lớn của vùng, Cần Thơ vẫn duy trì vị thế thu hút thí sinh.
- Vĩnh Long: Duy trì sự ổn định từ 2021 đến 2024 với mức tăng trưởng lần lượt là +11%, -9%, -1%, và +5%. Phục hồi đáng kể sau giai đoạn suy giảm, thể hiện nỗ lực cải thiện số lượng thí sinh.

b. Tỉnh có sự biến động lớn

- Hậu Giang: Ghi nhận mức tăng trưởng mạnh vào 2021 (+17%), nhưng giảm nhẹ vào 2022 (-4%) và phục hồi ổn định từ 2023 (+6%) đến 2024 (+5%). Phản ánh khả năng điều chỉnh tốt từ các chính sách địa phương.
- An Giang: Biến động mạnh nhất trong vùng. Năm 2021, số lượng thí sinh giảm sâu (-64%), nhưng tăng đột biến vào 2022 (+237%). Từ 2023 đến 2024, mức tăng trưởng dần ổn định (+7% và +2%). Các biến động này có thể liên quan đến sự thay đổi trong quản lý giáo dục hoặc các yếu tố dân số.

- Bạc Liêu: Mặc dù tăng trưởng tốt năm 2021 (+13%), nhưng từ 2022 đến 2024, số lượng thí sinh hầu như không thay đổi (0% đến -0%). Đây có thể là dấu hiệu của sự bão hòa trong nguồn thí sinh.

c. Tỉnh có xu hướng giảm dần hoặc bão hòa**

- Kiên Giang: Tăng trưởng đều đặn năm 2021 (+10%) và 2022 (-5%), nhưng chững lại vào 2023 (+1%) và 2024 (+4%). Tỉnh đang bước vào giai đoạn tăng trưởng chậm hoặc bão hòa.

- Sóc Trăng: Xu hướng tương tự với mức tăng trưởng +13% (2021) và giảm dần ở các năm tiếp theo (+4%, +1%, +5%).

- Bến Tre: Tăng trưởng tốt năm 2021 (+16%), sau đó giảm nhẹ (-7%) và gần như bão hòa vào 2023 (-1%) và 2024 (0%).

- Trà Vinh: Biến động nhẹ qua các năm, từ +12% (2021) đến -3% (2023), nhưng phục hồi vào 2024 (+5%).

- Đồng Tháp: Thiếu dữ liệu 2021, nhưng từ 2022 đến 2024 ghi nhận sự tăng trưởng ổn định (+4% đến +5%).

c. Kết luận

- Đồng bằng sông Cửu Long cho thấy sự phục hồi mạnh mẽ vào năm 2021, nhưng từ 2022 trở đi, tốc độ tăng trưởng số lượng thí sinh bắt đầu giảm hoặc ổn định, đặc biệt ở các tỉnh lớn như Cần Thơ, Long An, và Tiền Giang. Một số tỉnh nhỏ hơn như An Giang và Hậu Giang có mức biến động lớn, cho thấy ảnh hưởng từ các yếu tố đặc thù địa phương hoặc chính sách giáo dục chưa đồng bộ.

7. Đồng Nam Bộ

year	dia_phuong	2020	2021	2022	2023	2024	Ti_lệ_21	Ti_lệ_22	Ti_lệ_23	Ti_lệ_24
3	Hồ Chí Minh	74451	86259	84596	84851	87322	16%	-2%	0%	3%
5	Đồng Nai	28254	30366	31363	33158	33800	7%	3%	6%	2%
1	Bình Dương	11386	12999	12797	14218	15239	14%	-2%	11%	7%
0	Bà Rịa-Vũng Tàu	11441	12959	12661	12925	12635	13%	-2%	2%	-2%
2	Bình Phước	9774	10362	10702	10930	11304	6%	3%	2%	3%
4	Tây Ninh	8544	9644	9763	10229	10486	13%	1%	5%	3%

- **Hồ Chí Minh:**
 - Năm 2021 tăng trưởng mạnh nhất (16%), nhưng đã giảm -2% trong năm 2022 và duy trì mức tăng chậm từ 0%-3% trong hai năm tiếp theo. Điều này có thể do tác động từ dân số ổn định và nguồn lực giáo dục sẵn có.
- **Đồng Nai:**
 - Có xu hướng tăng trưởng ổn định qua các năm, dao động từ 2%-7%. Mức tăng cao nhất là 7% vào năm 2021 và thấp nhất 2% vào 2024.
- **Bình Dương:**
 - Ghi nhận mức tăng trưởng mạnh mẽ (14% năm 2021) và hồi phục đáng kể vào năm 2023 (11%), cho thấy tiềm năng lớn trong việc thu hút học sinh mới.
- **Bà Rịa - Vũng Tàu:**
 - Tăng trưởng khá ổn định trong giai đoạn 2021 (13%), tuy nhiên năm 2022 và 2024 giảm -2%, cho thấy dấu hiệu suy giảm cần được chú ý.
- **Bình Phước:**
 - Tăng trưởng tương đối thấp so với các tỉnh khác (dao động từ 3%-6%), nhưng vẫn duy trì sự ổn định trong suốt giai đoạn.
- **Tây Ninh:**
 - Tăng trưởng đáng kể ở mức 13% (2021), nhưng giảm mạnh về mức 1%-5% trong các năm tiếp theo.

. Nhận xét và kết luận:

- **Xu hướng chung:** Đông Nam Bộ là khu vực có mức tăng trưởng số lượng thí sinh khá ổn định, đặc biệt là các tỉnh có nền kinh tế phát triển mạnh như Hồ Chí Minh, Đồng Nai, và Bình Dương. Tuy nhiên, một số tỉnh như Bà Rịa - Vũng Tàu và Tây Ninh có dấu hiệu giảm tăng trưởng ở một số năm, cần đánh giá lại các yếu tố ảnh hưởng.
- **Thách thức:** Tỷ lệ giảm nhẹ trong năm 2022 ở hầu hết các tỉnh (đặc biệt là Hồ Chí Minh, Bà Rịa - Vũng Tàu) có thể liên quan đến các yếu tố xã hội hoặc giáo dục.

***Số lượng thí sinh có tổng điểm cao (>=54) theo khu vực từ 2020 đến 2024**

Số lượng thí sinh đạt điểm cao(>=54) tổng tất cả các môn theo khu vực

```
# tính và lọc số lượng thí sinh>=54 điểm
data_all['tong_diem'] = data_all[['toan', 'ngu_van', 'vat_ly', 'hoa_hoc', 'sinh_hoc', 'lich_su', 'dia_ly', 'gdcd', 'ngoai_ngu']].sum(axis=1)

data_all['year'] = pd.to_numeric(data_all['year'], errors='coerce')

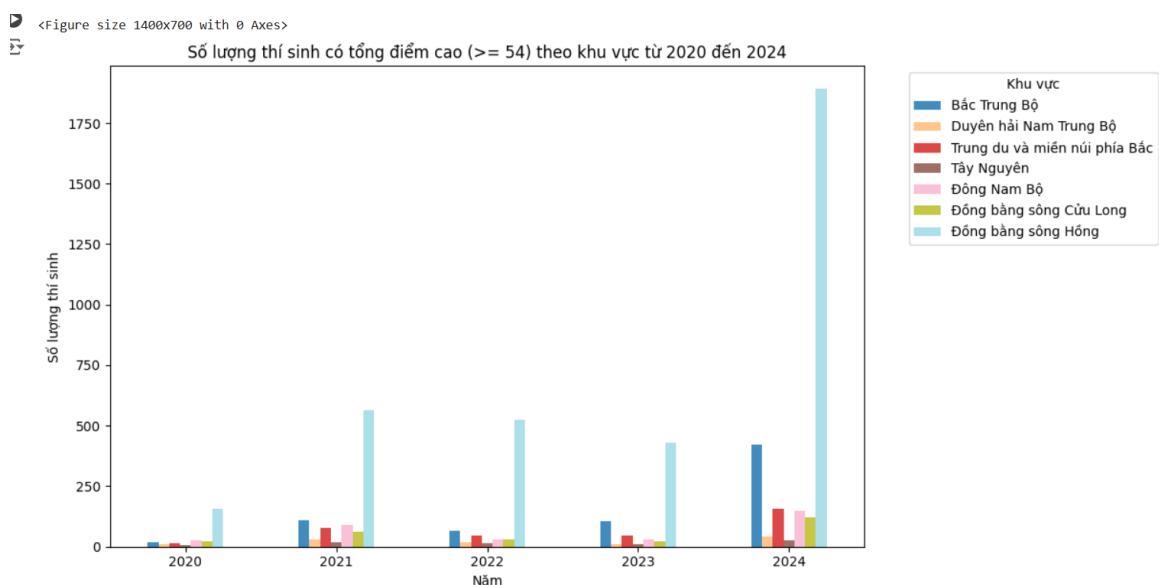
data_tong_diem_54 = data_all[(data_all['tong_diem'] >= 54) & (data_all['year'].between(2020, 2024))]

#Đếm số lượng học sinh có tổng điểm >= 54 cho mỗi khu vực và mỗi năm.
students_region_tong_54 = data_tong_diem_54.groupby(['year', 'khu_vuc']).size().unstack(fill_value=0)

plt.vẽ màu
plt.figure(figsize=(14, 7))
students_region_tong_54.plot(kind='bar', stacked=False, colormap='tab20', alpha=0.85, figsize=(12, 6))

plt.title('Số lượng thí sinh có tổng điểm cao (>= 54) theo khu vực từ 2020 đến 2024')
plt.xlabel('Năm')
plt.ylabel('Số lượng thí sinh')

# cẩn chỉnh
plt.xticks(rotation=0, ha='center')
plt.legend(title="Khu vực", bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.show()
```



1. Tổng quan xu hướng:

- Đồng bằng sông Hồng vượt trội so với các khu vực khác về số lượng thí sinh đạt điểm cao trong suốt giai đoạn 2020-2024.

- Các khu vực còn lại (như Bắc Trung Bộ, Tây Nguyên, và Đông Nam Bộ) có số lượng thí sinh đạt điểm cao khá khiêm tốn và tương đối ổn định qua các năm.

2. Phân tích chi tiết theo từng khu vực:

- Đồng bằng sông Hồng:

- Chiếm ưu thế rõ rệt, đặc biệt vào năm 2024 với mức tăng trưởng vượt bậc so với các năm trước đó (gần 1800 thí sinh). Điều này cho thấy chất lượng giáo dục và sự cạnh tranh cao tại khu vực này.

- Bắc Trung Bộ:

- Có sự tăng trưởng ổn định qua các năm, đặc biệt nổi bật vào năm 2024 với số lượng thí sinh tăng đáng kể, phản ánh sự cải thiện trong hệ thống giáo dục hoặc nguồn lực đào tạo.

- Duyên hải Nam Trung Bộ và Trung du và miền núi phía Bắc:

- Hai khu vực này có mức tăng trưởng khá khiêm tốn, duy trì ở mức thấp và không có đột phá lớn trong giai đoạn.

- Tây Nguyên:

- Gần như không có sự thay đổi lớn, số lượng thí sinh đạt điểm cao vẫn rất thấp, cho thấy hạn chế trong việc tạo điều kiện để học sinh đạt thành tích xuất sắc.

- Đông Nam Bộ:

- Có mức tăng trưởng nhẹ qua các năm, nhưng vẫn chưa đạt được mức độ cạnh tranh cao so với các khu vực như Đồng bằng sông Hồng.

- Đồng bằng sông Cửu Long:

- Số lượng thí sinh đạt điểm cao tương đối ít và không có sự biến động đáng kể qua các năm, có thể do những hạn chế trong chất lượng đào tạo.

3. Nhận xét và kết luận:

- Sự chênh lệch khu vực: Đồng bằng sông Hồng dẫn đầu cả nước về số lượng thí sinh đạt điểm cao, trong khi các khu vực như Tây Nguyên, Duyên hải Nam Trung Bộ và Đồng bằng sông Cửu Long có thành tích kém hơn đáng kể.

- Nguyên nhân tiềm năng:

- Chất lượng giáo dục, cơ sở hạ tầng, và các chính sách đào tạo khác biệt giữa các khu vực.

- Điều kiện kinh tế - xã hội ảnh hưởng lớn đến khả năng đầu tư cho giáo dục và mức độ cạnh tranh của học sinh.

*Số lượng thí sinh có tổng điểm thấp (<=30)theo khu vực từ năm 2020 đến năm 2024

Số lượng thí sinh đạt điểm thấp(<=30) tổng tất cả các môn theo khu vực

```
# tính và lọc số lượng thí sinh>=54 điểm
data_all['tong_diem'] = data_all[['toan', 'ngu_van', 'vat_ly', 'hoa_hoc', 'sinh_hoc', 'lich_su', 'dia_ly', 'gdcd', 'ngoai_ngu']].sum(axis=1)
data_tong_diem_30 = data_all[(data_all['tong_diem'] <= 30) & (data_all['year'].between(2020, 2024))]

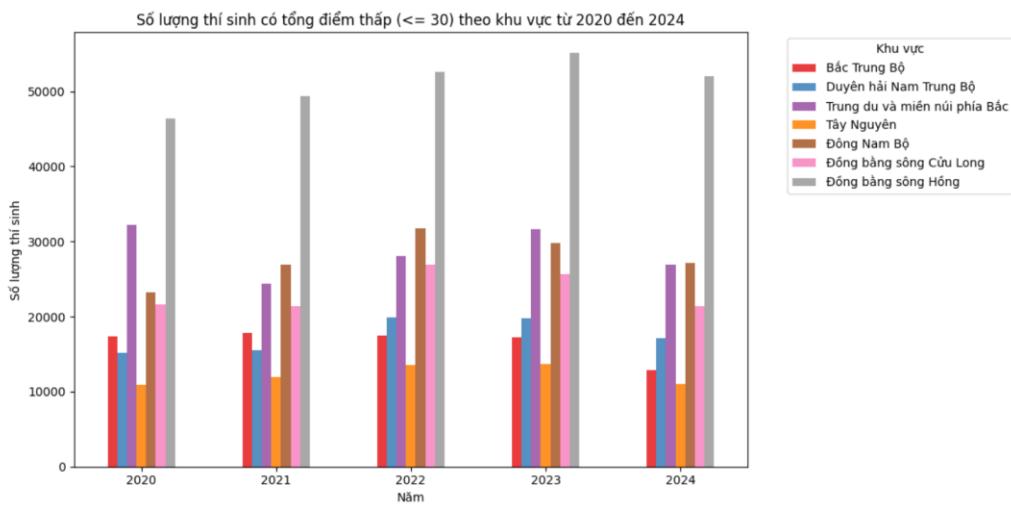
#Đếm số lượng học sinh có tổng điểm >= 54 cho mỗi khu vực và mỗi năm.
students_region_tong_30 = data_tong_diem_30.groupby(['year', 'khu_vuc']).size().unstack(fill_value=0)

#Vẽ màu
plt.figure(figsize=(14, 7))
students_region_tong_30.plot(kind='bar', stacked=False, colormap='Set1', alpha=0.85, figsize=(12, 6))

plt.title('Số lượng thí sinh có tổng điểm thấp (<= 30) theo khu vực từ 2020 đến 2024')
plt.xlabel('Năm')
plt.ylabel('Số lượng thí sinh')

# cẩn chỉnh
plt.xticks(rotation=0, ha='center')
plt.legend(title="Khu vực", bbox_to_anchor=(1.05, 1), loc='upper left')

plt.tight_layout()
plt.show()
```



1. Khu vực Bắc Trung Bộ:

- Có số lượng thí sinh cao nhất trong suốt giai đoạn 2020-2024.
- Số lượng tăng đều từ 2020 đến 2022, đạt đỉnh khoảng 50.000 thí sinh vào năm 2022.
- Sau đó giảm nhẹ vào năm 2023 và 2024 nhưng vẫn duy trì ở mức cao.

2. Khu vực Duyên hải Nam Trung Bộ:

- Xếp thứ hai về số lượng thí sinh.
- Có xu hướng tăng liên tục từ 2020 đến 2024, từ khoảng 20.000 lên trên 30.000 thí sinh.
- Tốc độ tăng nhanh nhất trong giai đoạn này.

3. Khu vực Trung du và miền núi phía Bắc:

- Số lượng thí sinh khá ổn định ở mức khoảng 20.000-25.000 từ 2020 đến 2024.
- Tăng nhẹ vào năm 2022 rồi giảm dần về mức ban đầu vào năm 2024.

4. Các khu vực còn lại:

- Tây Nguyên, Đồng Nam Bộ, Đồng bằng sông Cửu Long và Đồng bằng sông Hồng đều có số lượng thí sinh thấp hơn so với 3 khu vực trên.

- Có xu hướng tăng nhẹ hoặc tương đối ổn định trong giai đoạn này.

5. Tổng thể:

- Tổng số lượng thí sinh có xu hướng tăng dần từ 2020 đến 2023, đạt đỉnh vào năm 2022-2023.

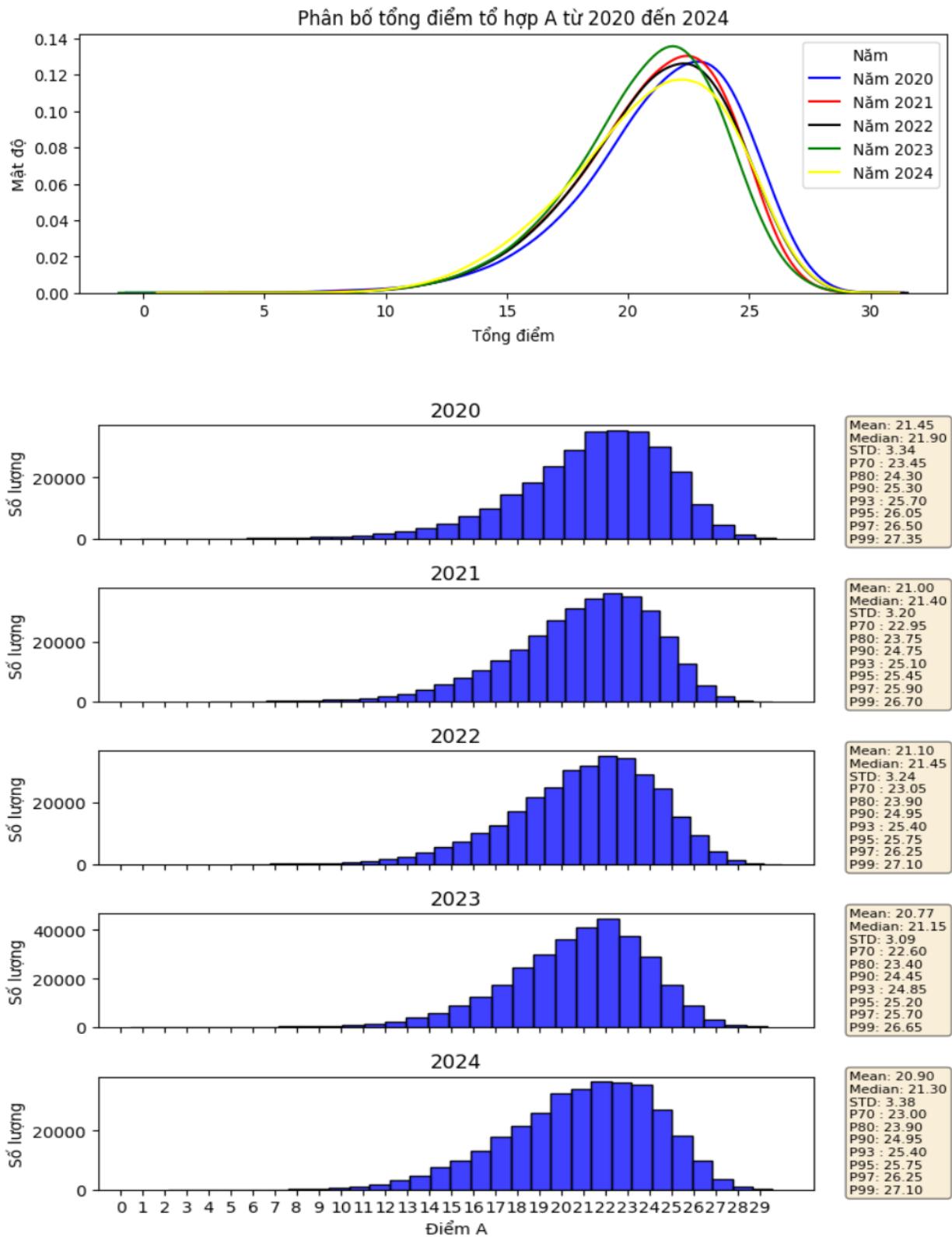
- Sau đó có xu hướng giảm nhẹ vào năm 2024.

D.Phân tích theo khối thi

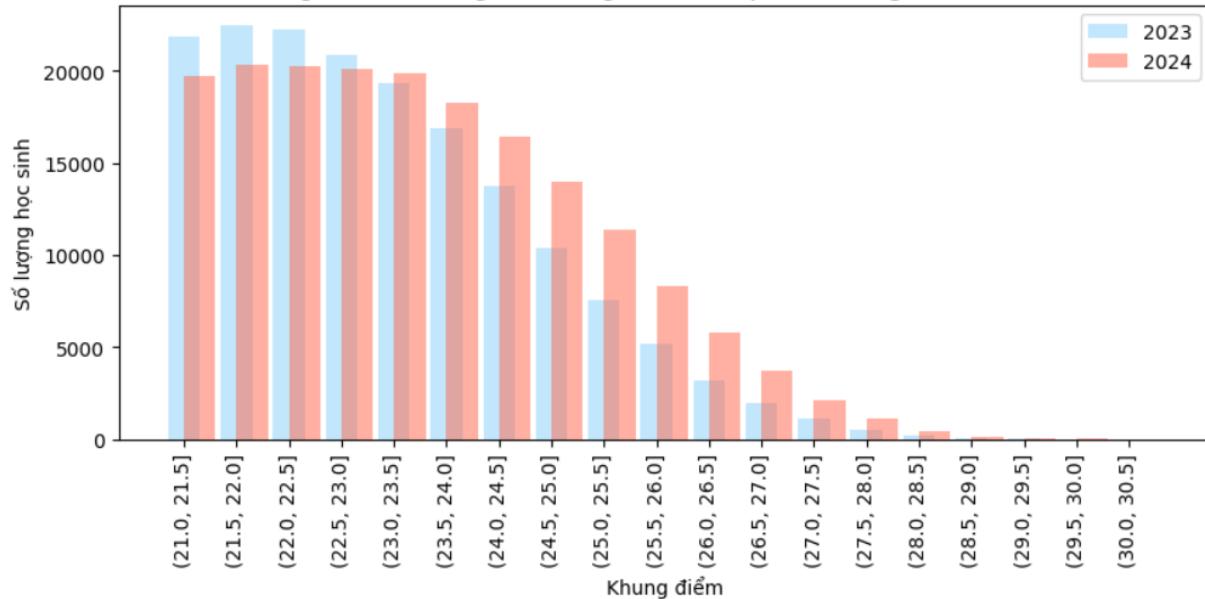
```
#tao cot tong diem cho tung to hop mon
data_all['A'] = data_all['toan'] + data_all['vat_ly'] + data_all['hoa_hoc']
data_all['A1'] = data_all['toan'] + data_all['vat_ly'] + data_all['ngoai_ngu']
data_all['B'] = data_all['toan'] + data_all['hoa_hoc'] + data_all['sinh_hoc']
data_all['C'] = data_all['ngu_van'] + data_all['lich_su'] + data_all['dia_ly']
data_all['D'] = data_all['toan'] + data_all['ngoai_ngu'] + data_all['ngu_van']
data_all['D07'] = data_all['toan'] + data_all['ngoai_ngu'] + data_all['hoa_hoc']
#cac to hop mon can phan tich
combinations = ['A', 'A1', 'B', 'C', 'D', 'D07']
#thiet lap cai dat cho bieu do
colors = ['blue', 'red', 'black', 'green', 'yellow']
#ve
for combo in combinations:
    plt.figure(figsize=(10, 3))
    for year, color in zip([2020, 2021, 2022, 2023, 2024], colors):
        data = data_all[data_all['year'] == str(year)][combo].dropna()
        sns.kdeplot(data, label=f'Năm {year}', color=color, bw_adjust=2)
    plt.title(f'Phân bố tổng điểm tổ hợp {combo} từ 2020 đến 2024')
    plt.xlabel('Tổng điểm')
    plt.ylabel('Mật độ')
    plt.legend(title='Năm')
    plt.grid(False)
    plt.show()
```

Dùng biểu đồ kdeplot để biết được mật độ điểm theo từng năm

*Khối A :

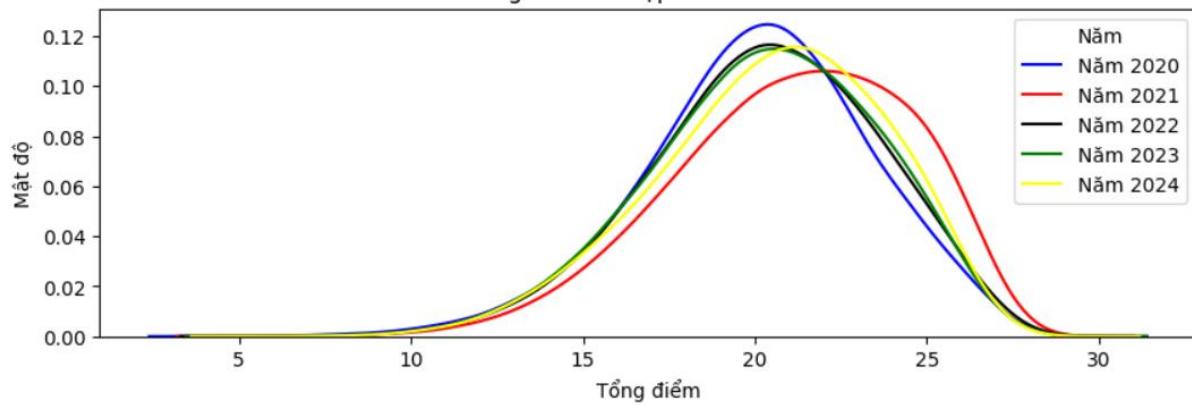


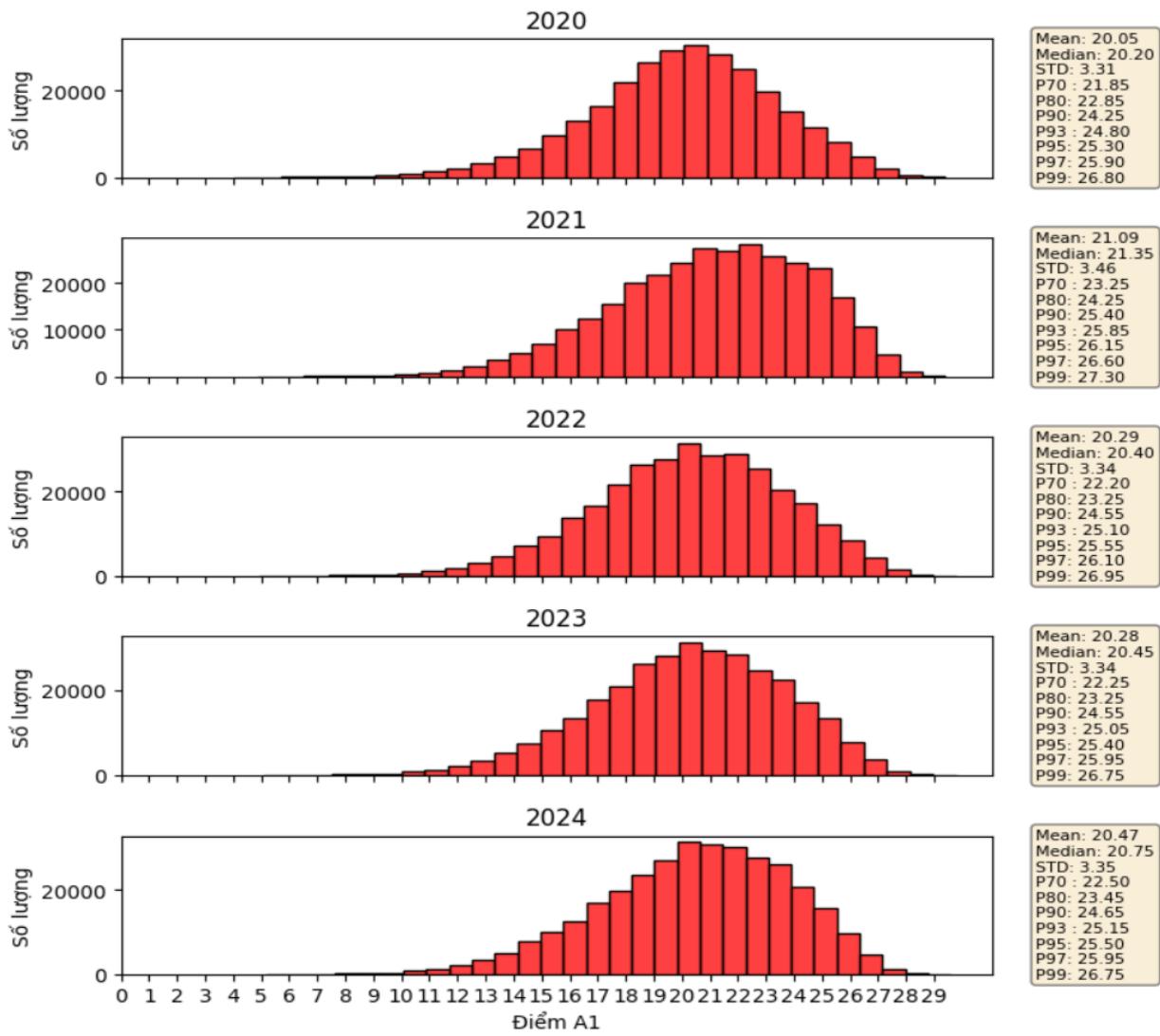
So sánh số lượng học sinh trong các khung điểm tổ hợp {combo} giữa các năm 2023 và 2024

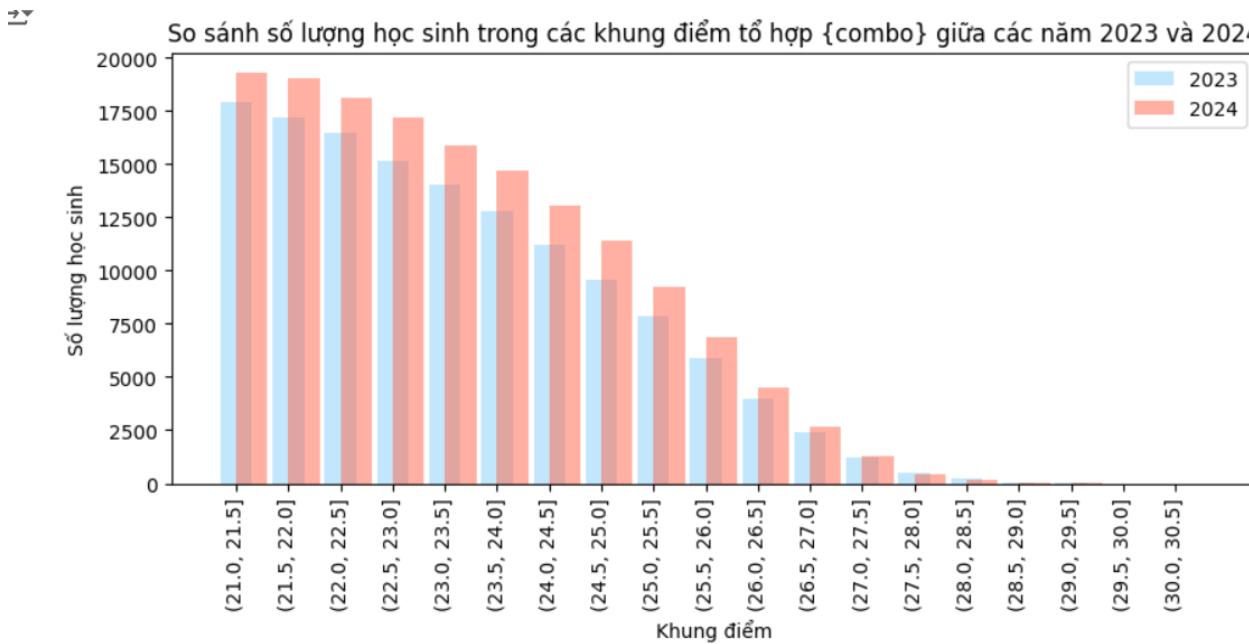


*Khối A1

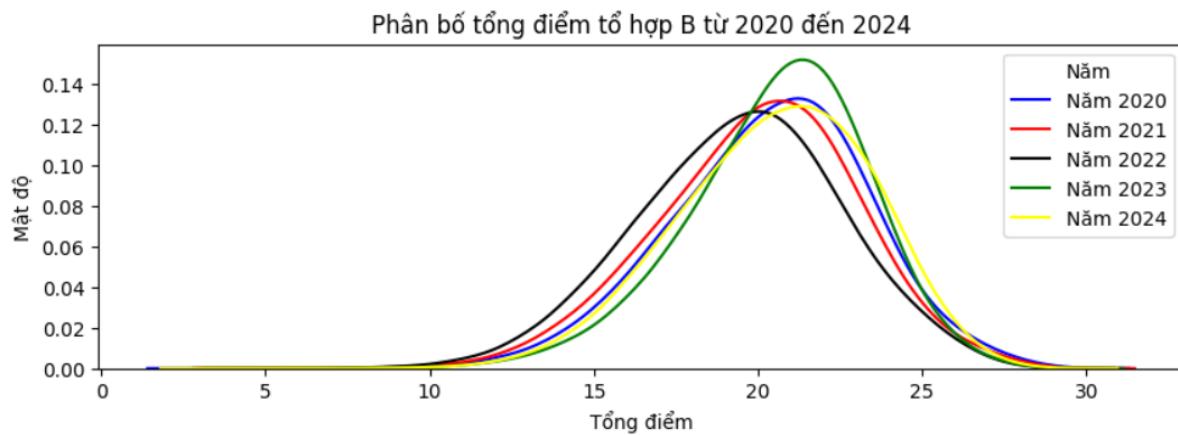
Phân bố tổng điểm tổ hợp A1 từ 2020 đến 2024

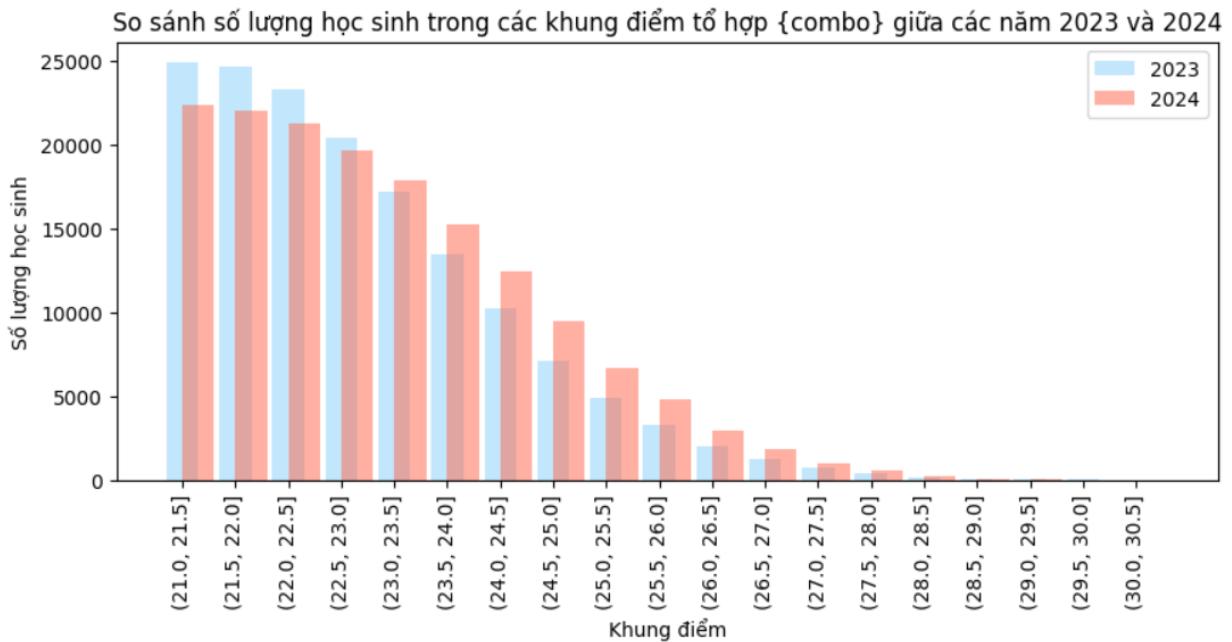
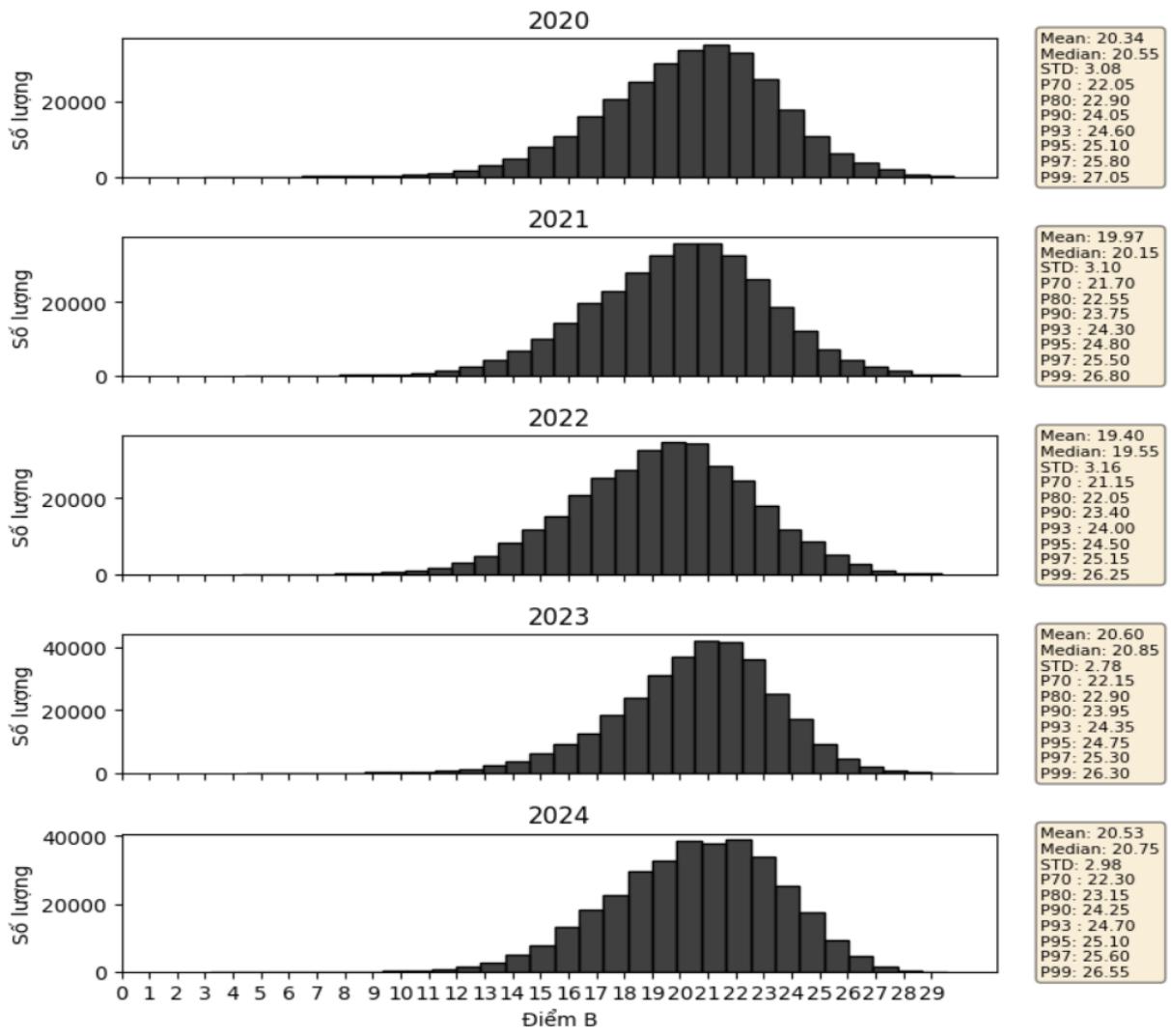




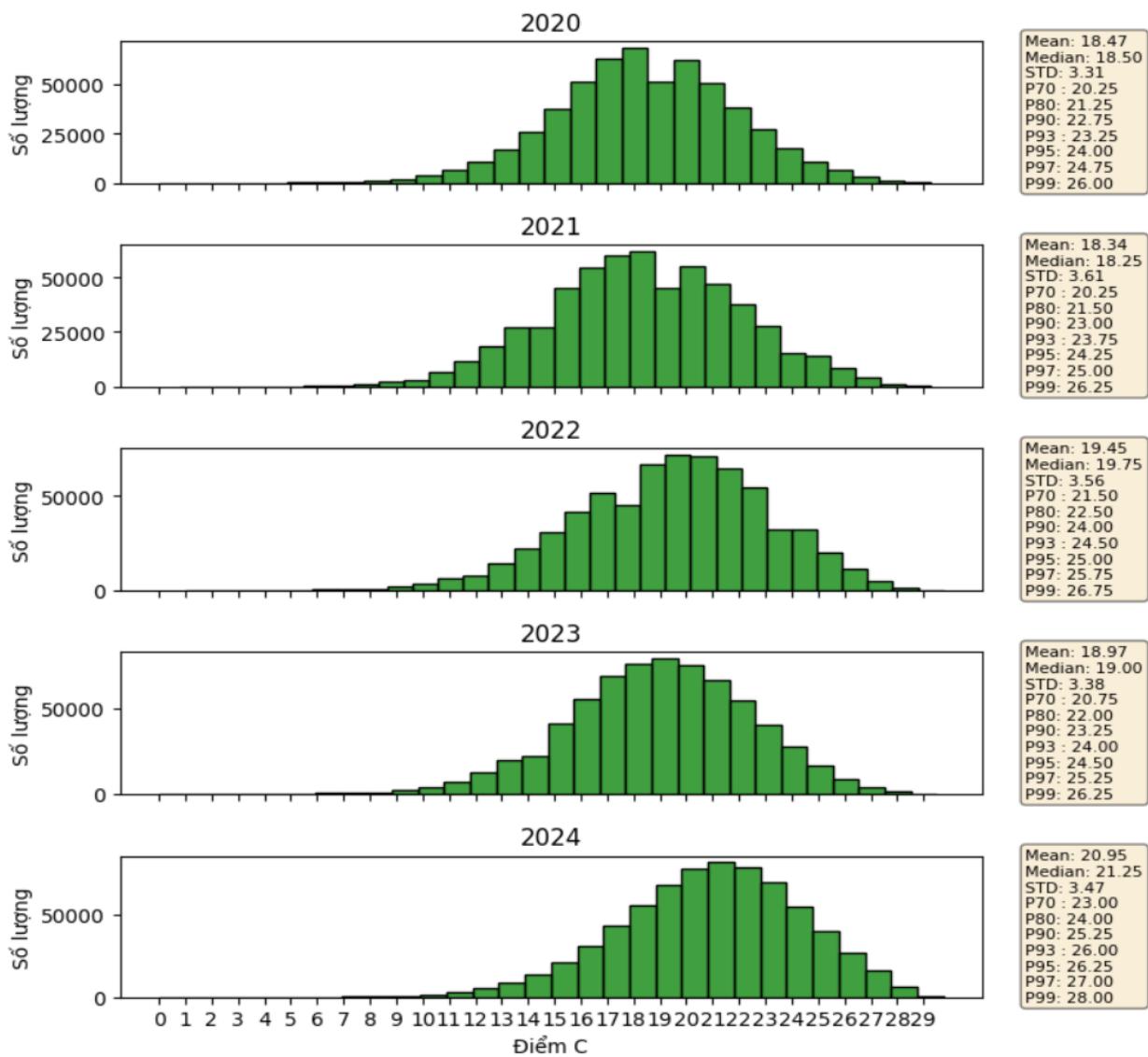
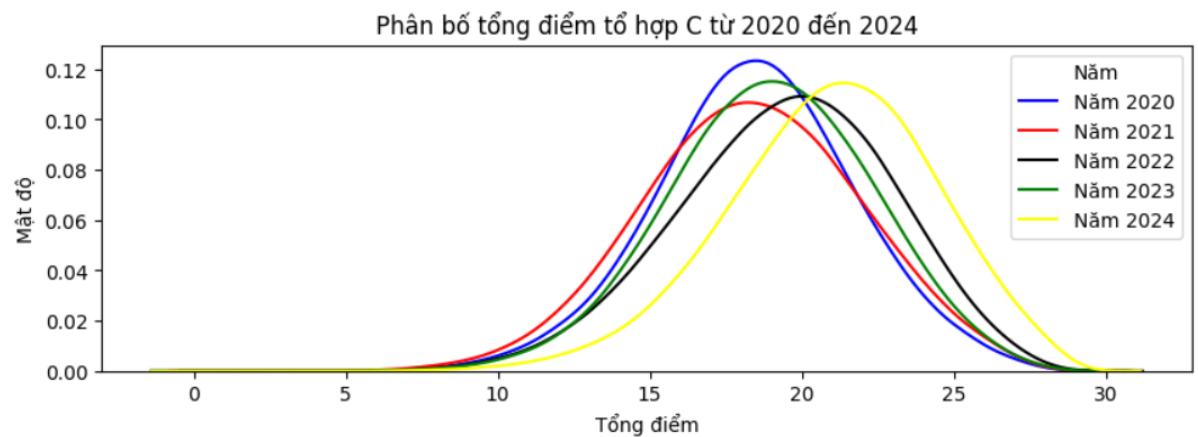


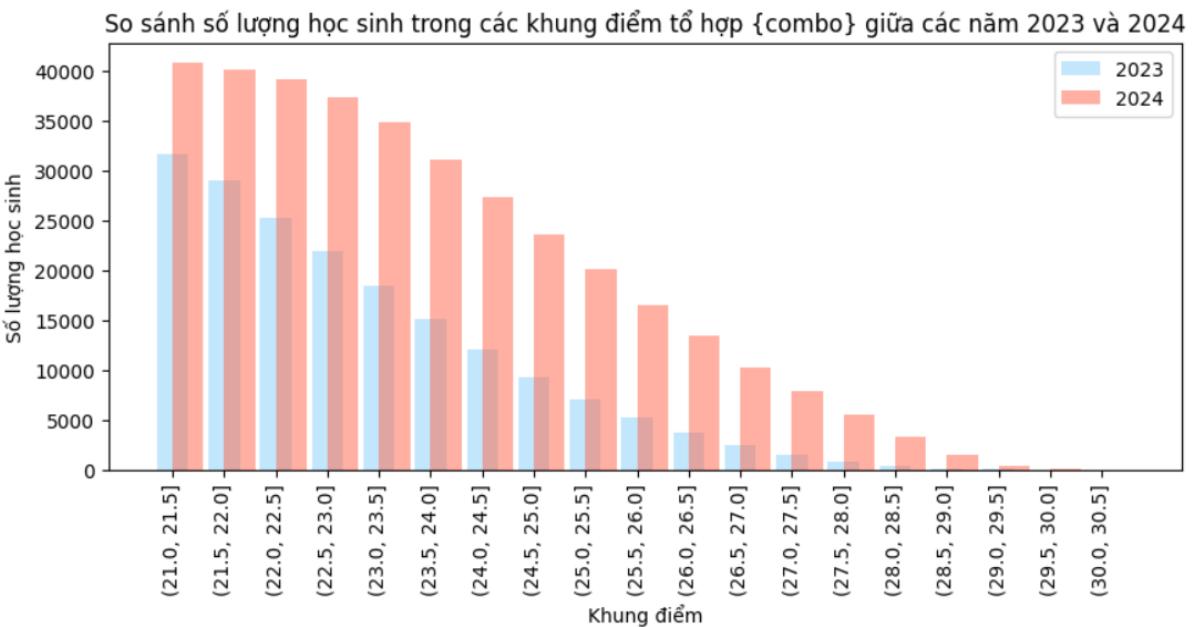
*Khối B



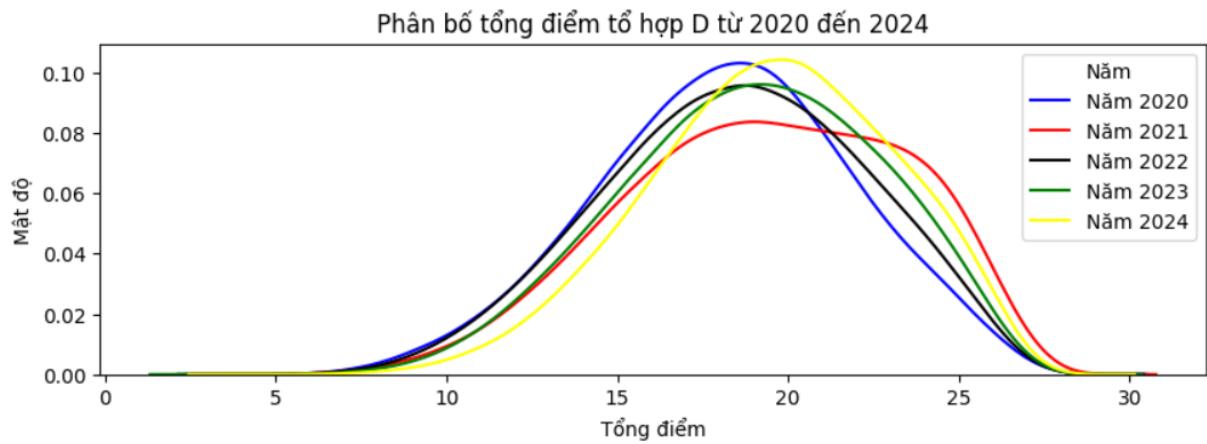


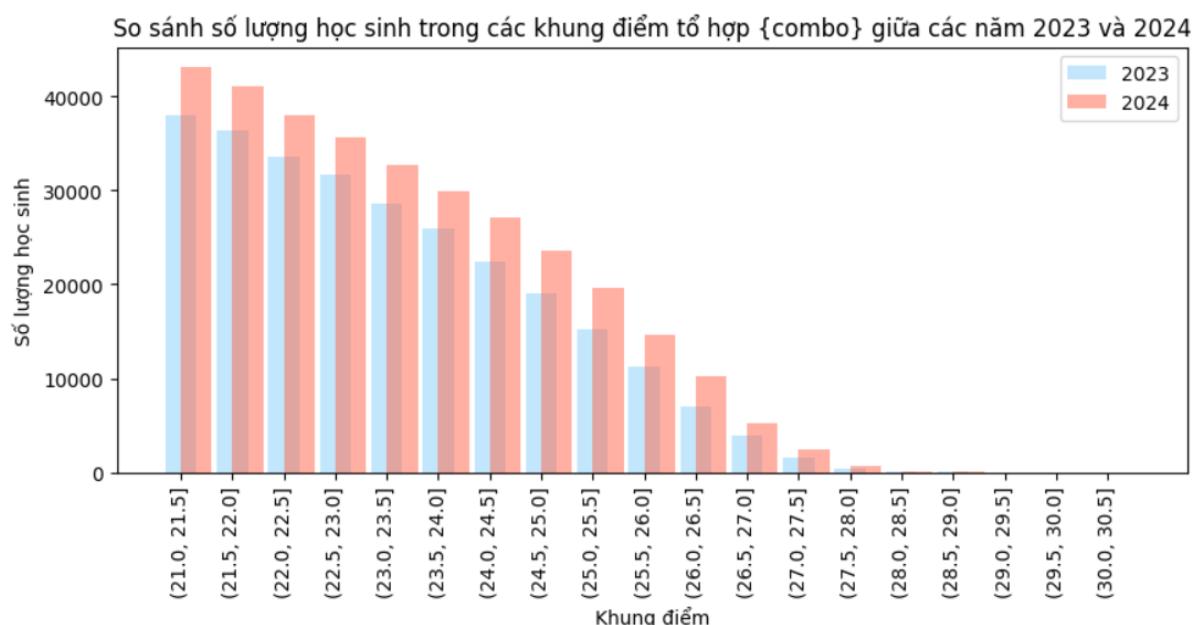
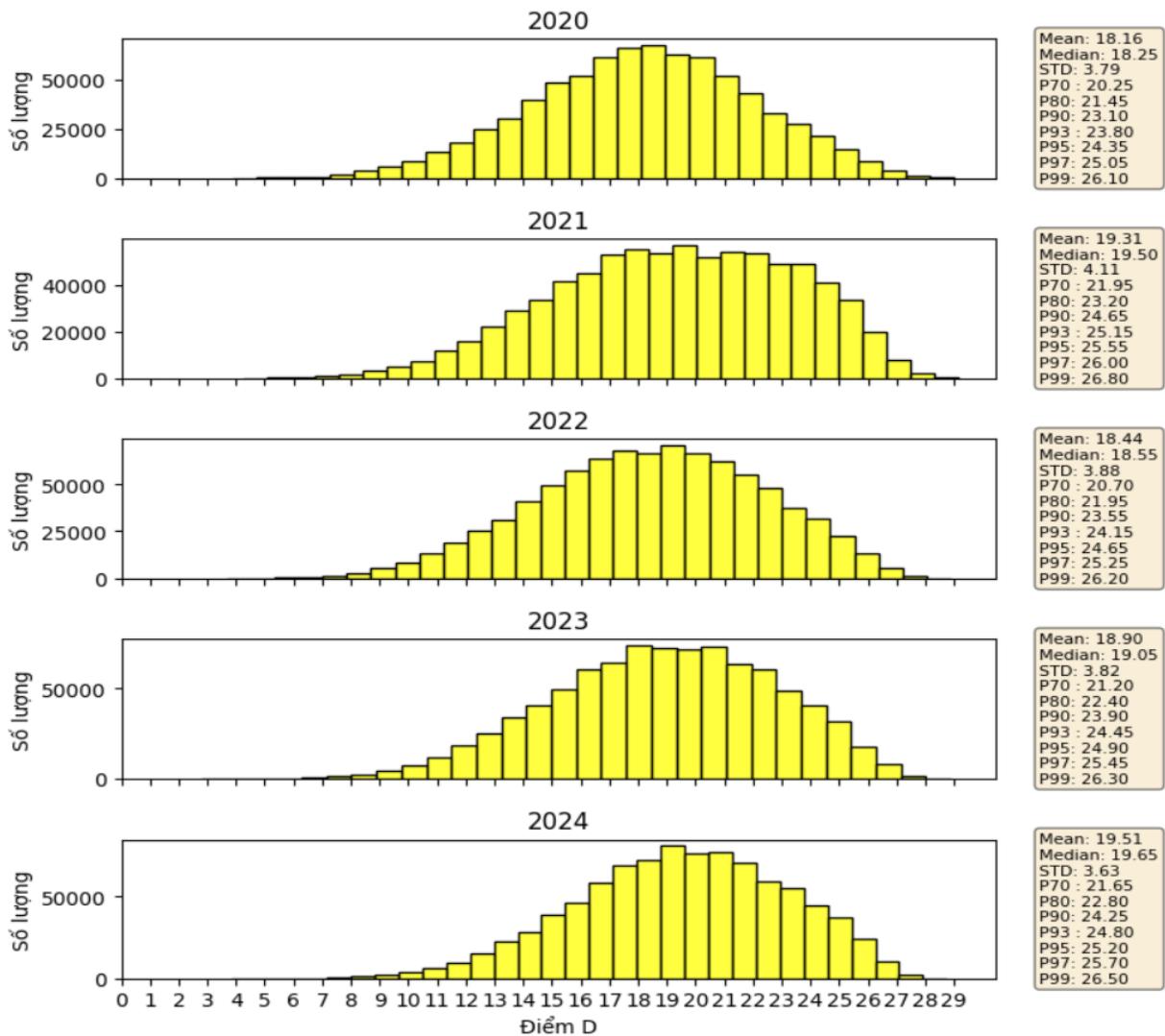
*Khối C





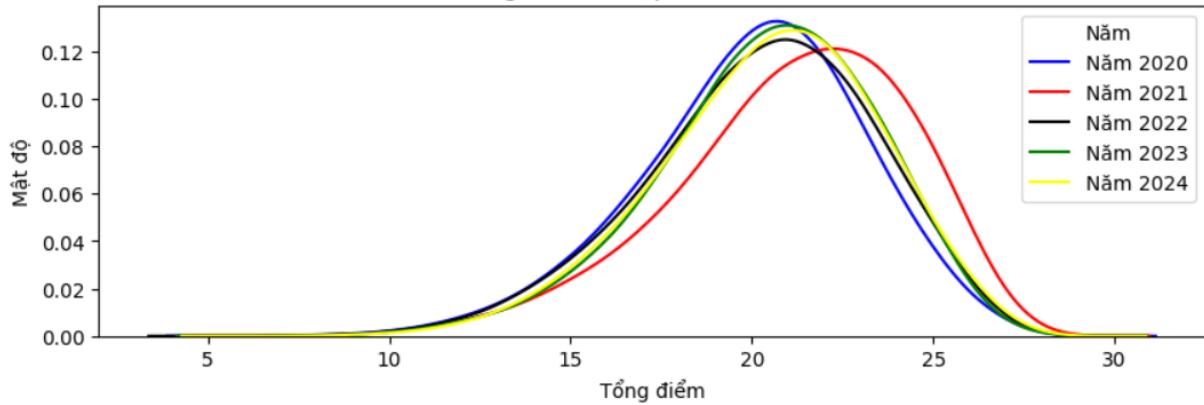
*Khối D



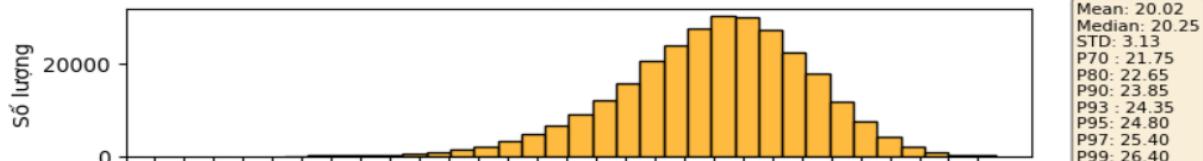


*Khối D07

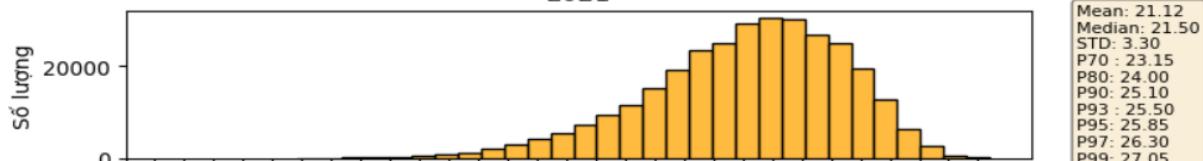
Phân bố tổng điểm tổ hợp D07 từ 2020 đến 2024



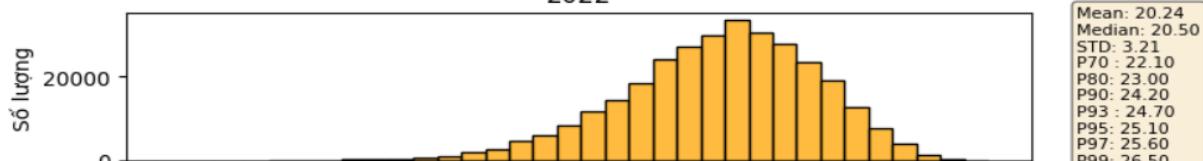
2020



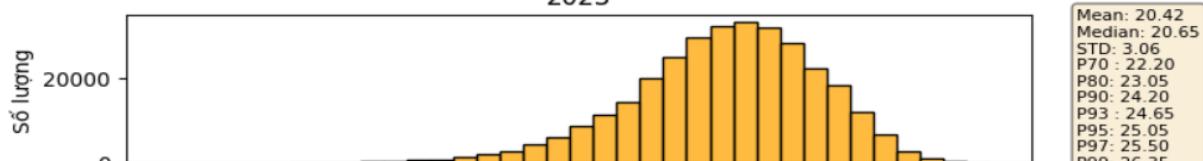
2021



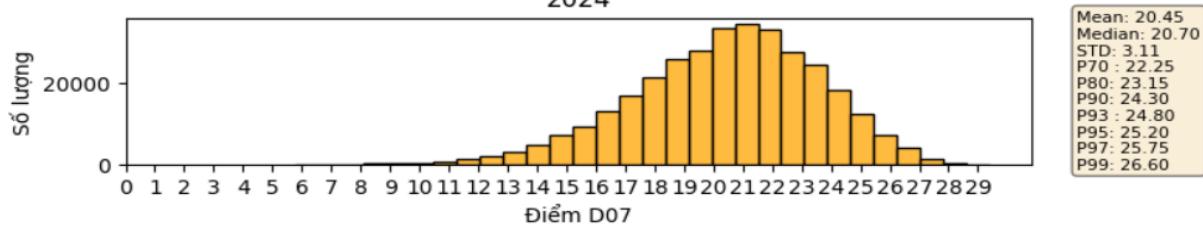
2022

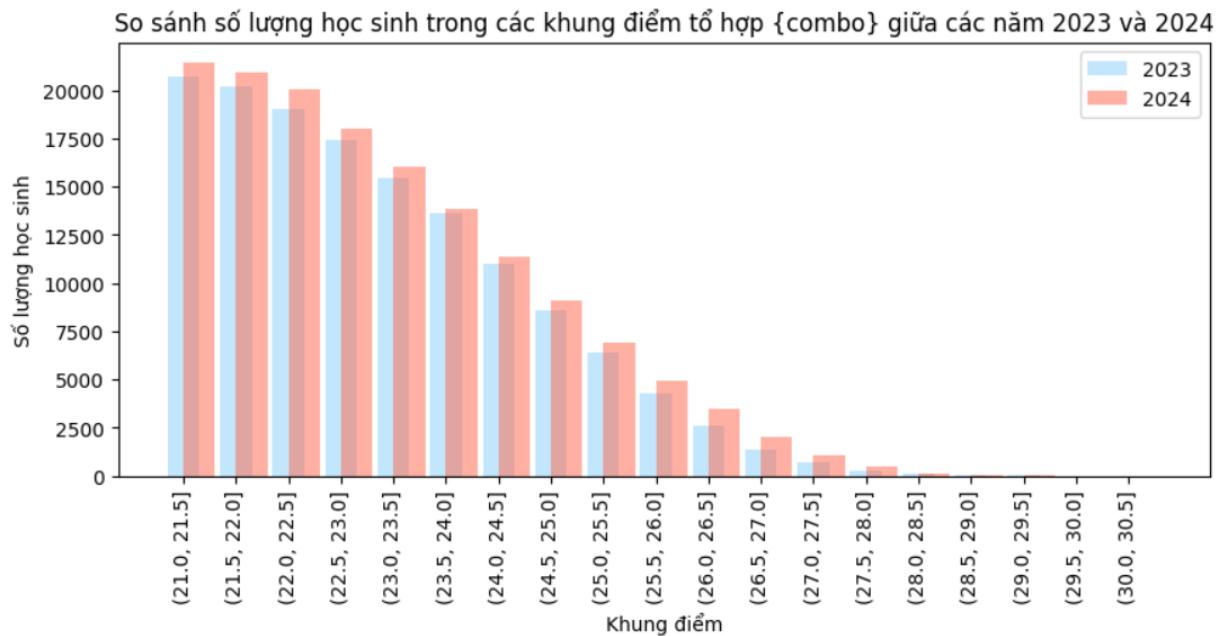


2023



2024





- **Khối A (Toán, Lý, Hóa):**

- Từ năm 2020 đến 2024, biểu đồ phân phối mật độ cho thấy sự ổn định với đỉnh phân phối nằm trong khoảng 18–22 điểm.
- Sự thay đổi giữa các năm là không quá lớn, nhưng năm 2024 có phân phối trải rộng hơn ở phần điểm thấp.

- **Khối A1 (Toán, Lý, Anh):**

- Xu hướng điểm khối A1 khá tương đồng với khối A nhưng có sự biến thiên mạnh hơn giữa các năm, đặc biệt năm 2021 có đỉnh phân phối cao hơn.
- Năm 2024 có xu hướng giảm nhẹ ở mức điểm cao (>25 điểm), cho thấy mức độ cạnh tranh không cao như các năm trước.

- **Khối B (Toán, Hóa, Sinh):**

- Đỉnh phân phối của khối B dịch chuyển từ mức 19 điểm (2020) lên mức 20–21 điểm (2023).
- Năm 2024 có sự phân tán rõ rệt hơn với một số thí sinh đạt mức điểm thấp hơn 10.

- **Khối C (Văn, Sử, Địa):**

- Biểu đồ khối C cho thấy sự ổn định qua các năm, với đỉnh phân phối duy trì trong khoảng 15–20 điểm.
- Năm 2024 có sự cải thiện nhẹ ở phần đuôi điểm cao (>25 điểm), chứng tỏ khả năng làm bài của thí sinh tốt hơn.

- **Khối D (Toán, Văn, Anh):**

- Điểm khối D khá đồng đều, với đỉnh phân phối nằm trong khoảng 18–22 điểm.
- Năm 2021 có sự chênh lệch nhẹ khi đỉnh phân phối thấp hơn so với các năm khác.

- **Khối D07 (Toán, Hóa, Anh):**

- Tương tự khối A1, khối D07 có sự biến động mạnh giữa các năm, đặc biệt năm 2021 và năm 2024.
- Năm 2024, sự phân tán ở mức điểm thấp (<10 điểm) nhiều hơn, nhưng đỉnh phân phối vẫn nằm ở mức 20–22 điểm.

3. Kết luận

- Các khối tổ hợp có sự khác biệt nhẹ về mức độ ổn định qua các năm.
- Năm 2024 cho thấy sự phân tán điểm rộng hơn ở một số khối (như A, B, và D07), điều này có thể liên quan đến thay đổi về đề thi hoặc trình độ của thí sinh.
- Điểm trung bình của các khối vẫn tập trung chủ yếu trong khoảng 18–22, phản ánh mức độ chuẩn hóa điểm khá tốt qua các năm.

D.Phân tích theo môn thi

*Toán

```
Toán

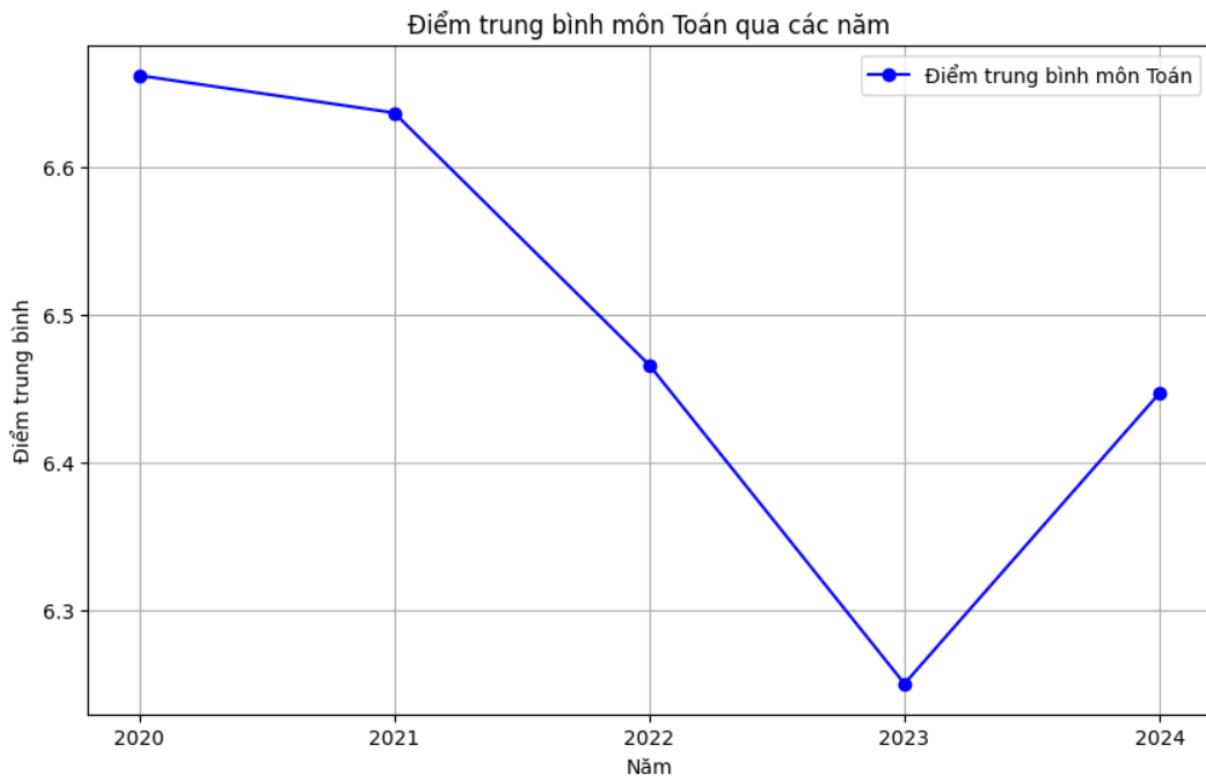
data_all['toan'] = pd.to_numeric(data_all['toan'], errors='coerce')

df_mean = data_all.groupby('year')['toan'].mean()
print(df_mean)

plt.figure(figsize=(10, 6))
plt.plot(df_mean.index, df_mean.values, marker='o', color='b', label='Điểm trung bình môn Toán')
plt.title('Điểm trung bình môn Toán qua các năm')
plt.xlabel('Năm')
plt.ylabel('Điểm trung bình')
plt.legend()
plt.grid(True)
plt.show()

year
2020    6.662271
2021    6.636990
2022    6.466254
2023    6.250557
2024    6.447309
Name: toan, dtype: float64
```

Dùng hàm .mean() để tính toán lấy điểm trung bình môn



```

data_all['year'] = pd.to_numeric(data_all['year'])

score = list(range(0, 11, 1))

fig, axes = plt.subplots(5, 1, figsize=(8, 8), sharex=True)

for i, year in enumerate([2020, 2021, 2022, 2023, 2024]):
    sns.histplot(data_all[data_all['year']==year]['toan'].dropna(), bins=20, kde=False, ax=axes[i])
    axes[i].set_title(f'{year}')
    axes[i].set_xlabel('diem toan')
    axes[i].set_ylabel('so luong')

    year_data = data_all[data_all['year']==year]['toan'].dropna()
    mean = year_data.mean()
    median = year_data.median()
    std_dev = year_data.std()
    p70 = year_data.quantile(0.70)
    p80 = year_data.quantile(0.80)
    p90 = year_data.quantile(0.90)
    p93 = year_data.quantile(0.93)
    p95 = year_data.quantile(0.95)
    p97 = year_data.quantile(0.97)
    p99 = year_data.quantile(0.99)
    textstr = (f'Mean: {mean:.2f}\nMedian: {median:.2f}\nSTD: {std_dev:.2f}\n'
               f'P70 : {p70:.2f}\nP80: {p80:.2f}\nP90: {p90:.2f}\n'
               f'P93 : {p93:.2f}\nP95: {p95:.2f}\nP97: {p97:.2f}\nP99: {p99:.2f}')
    props = dict(boxstyle='round', facecolor='wheat', alpha=0.5)
    axes[i].text(1.05, 0.5, textstr, transform=axes[i].transAxes, fontsize=8,
                verticalalignment='center', horizontalalignment='left', bbox=props)
plt.xticks(score)
plt.tight_layout()
plt.show()

```

Dùng hàm histplot() để vẽ biểu đồ histplot điểm của môn học

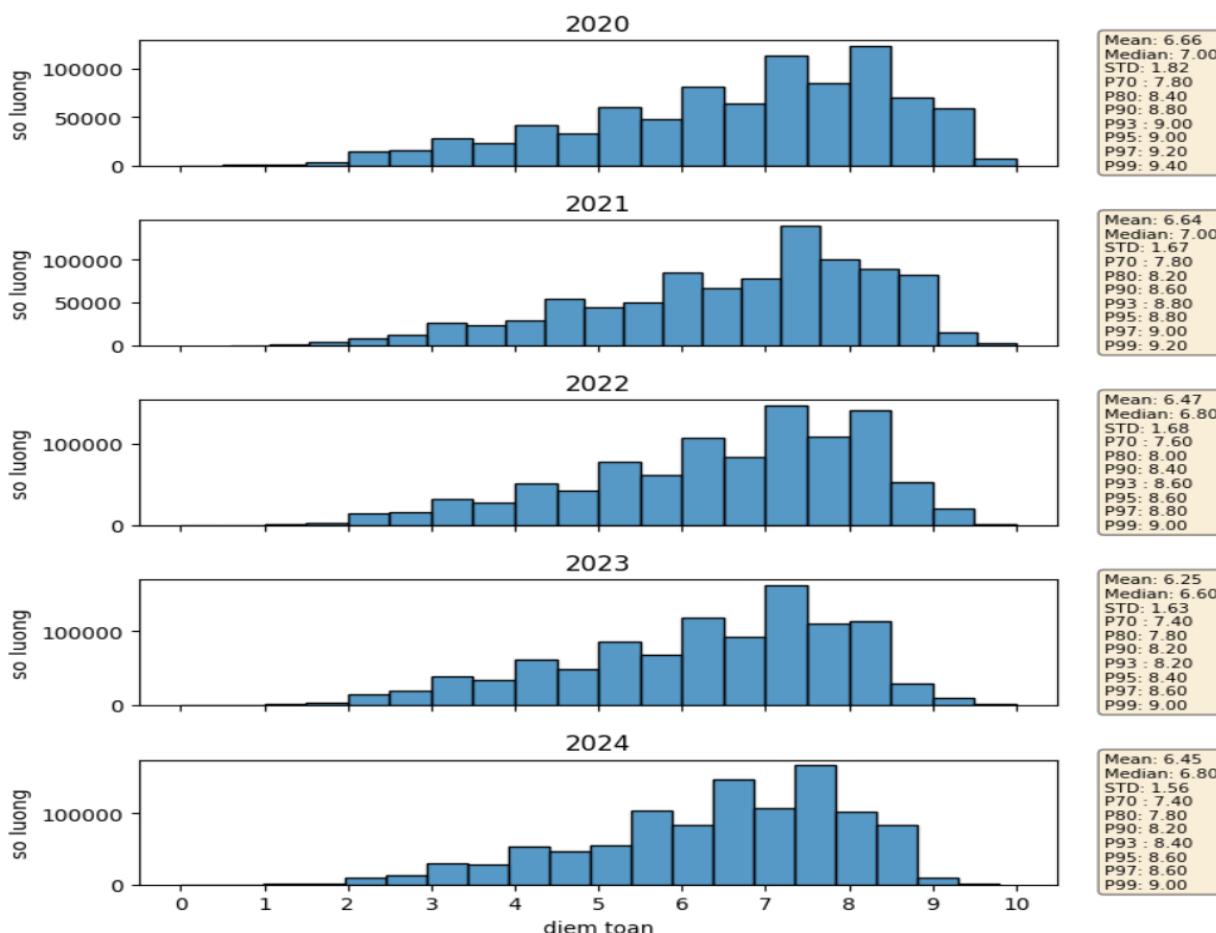
- `data_all[data_all['year'] == year]`: Lọc dữ liệu cho từng năm.
- `['toan']`: Chọn cột 'toan', giả sử đây là điểm thi Toán.
- `.dropna()`: Loại bỏ các giá trị NaN trong dữ liệu.
- `bins=20`: Chia dữ liệu thành 20 nhóm (bin) trong biểu đồ histogram.
- `kde=False`: Tắt biểu đồ mật độ (Kernel Density Estimate).
- `ax=axes[i]`: Vẽ biểu đồ vào biểu đồ con thứ i (theo chỉ số).

`year_data = data_all[data_all['year']==year]['toan'].dropna()`:

- Lọc dữ liệu cho năm year và chỉ lấy cột 'toan', loại bỏ các giá trị NaN.

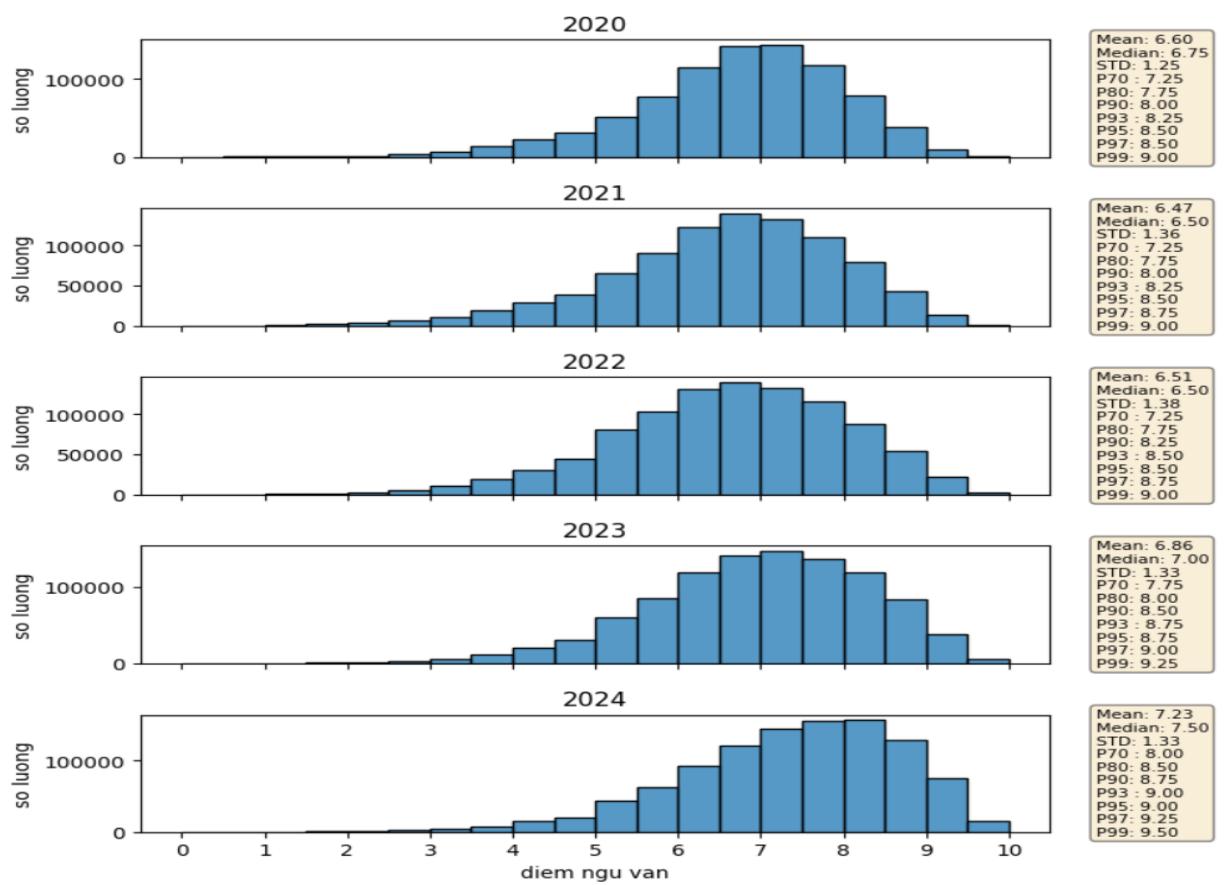
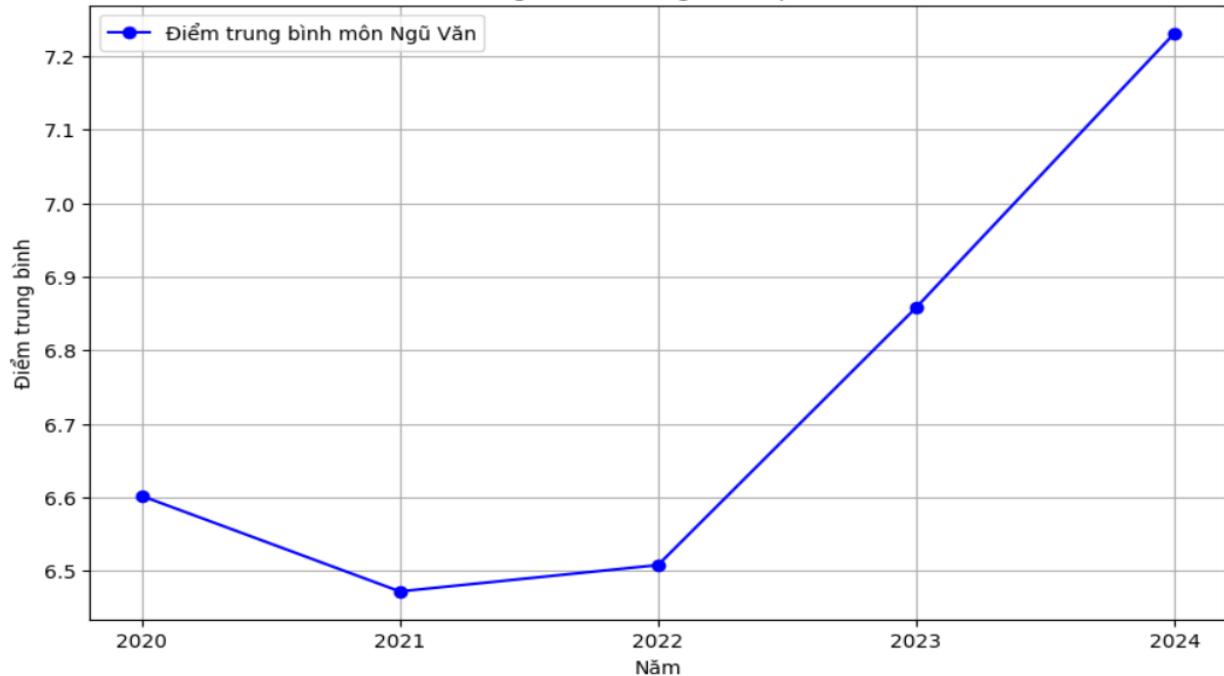
Các dòng tiếp theo tính toán các giá trị thống kê:

- **mean**: Trung bình.
- **median**: Trung vị.
- **std_dev**: Độ lệch chuẩn.
- **p70, p80, p90, p93, p95, p97, p99**: Các phân vị (quantiles) của dữ liệu tại các mức 70%, 80%, 90%, 93%, 95%, 97%, và 99%.



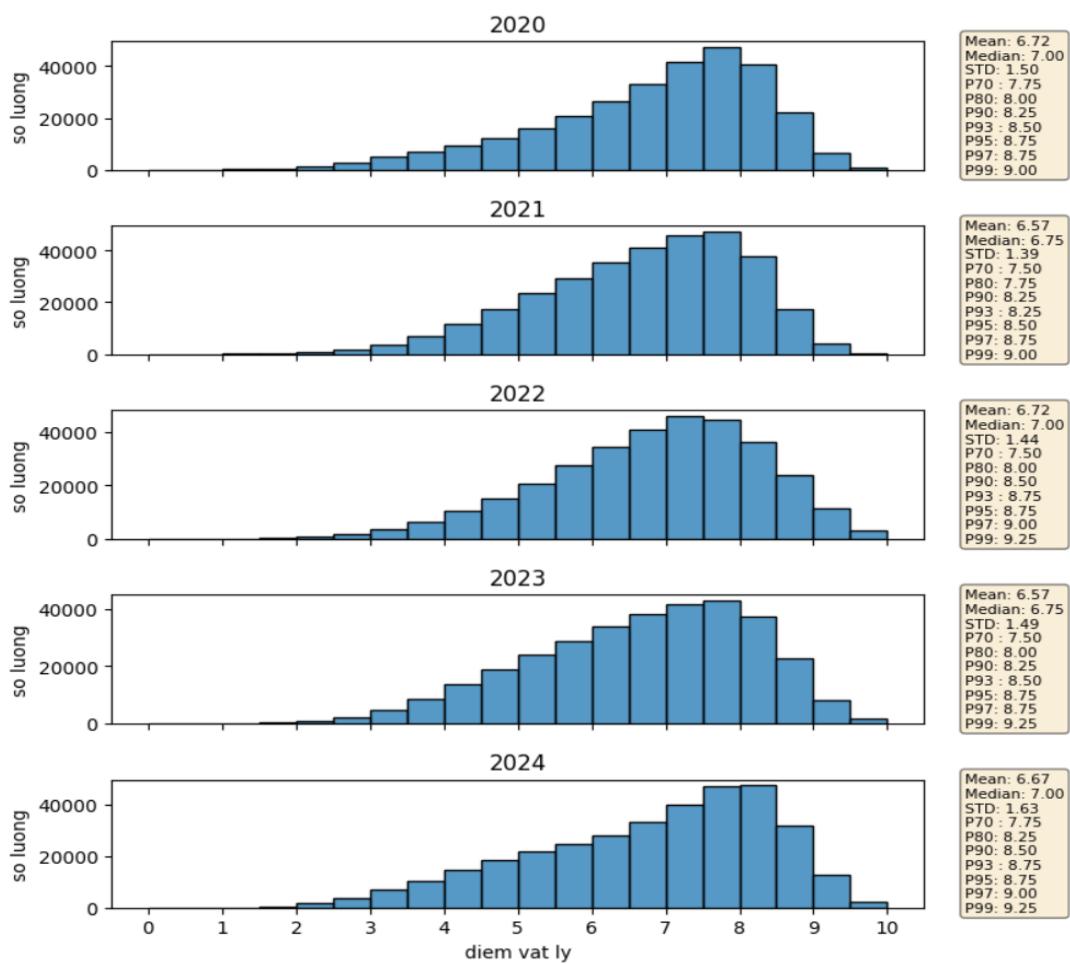
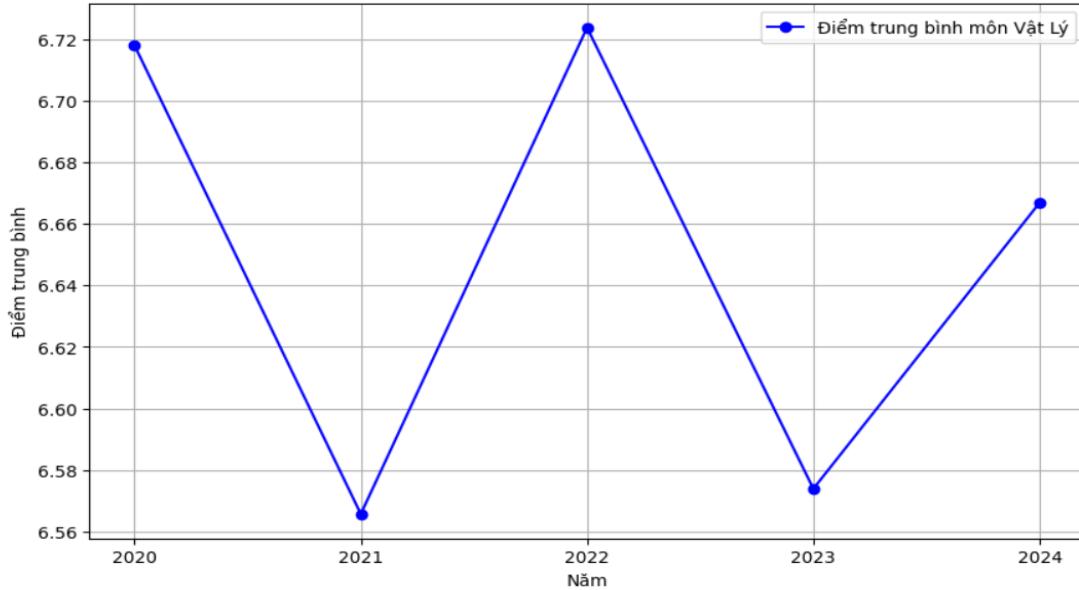
*Điểm Ngữ Văn

Điểm trung bình môn Ngữ Văn qua các năm

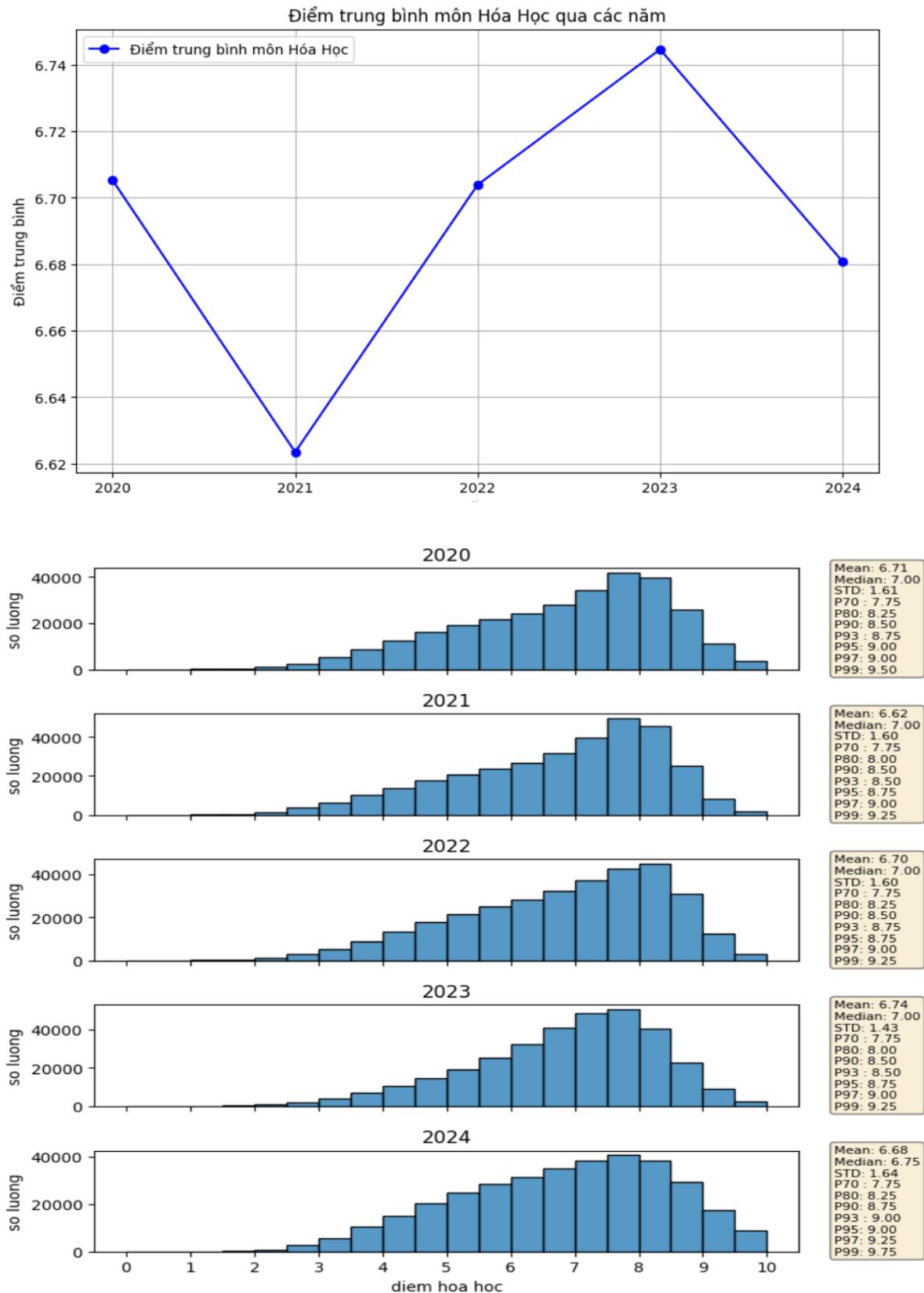


*Điểm Vật Lý

Điểm trung bình môn Vật Lý qua các năm

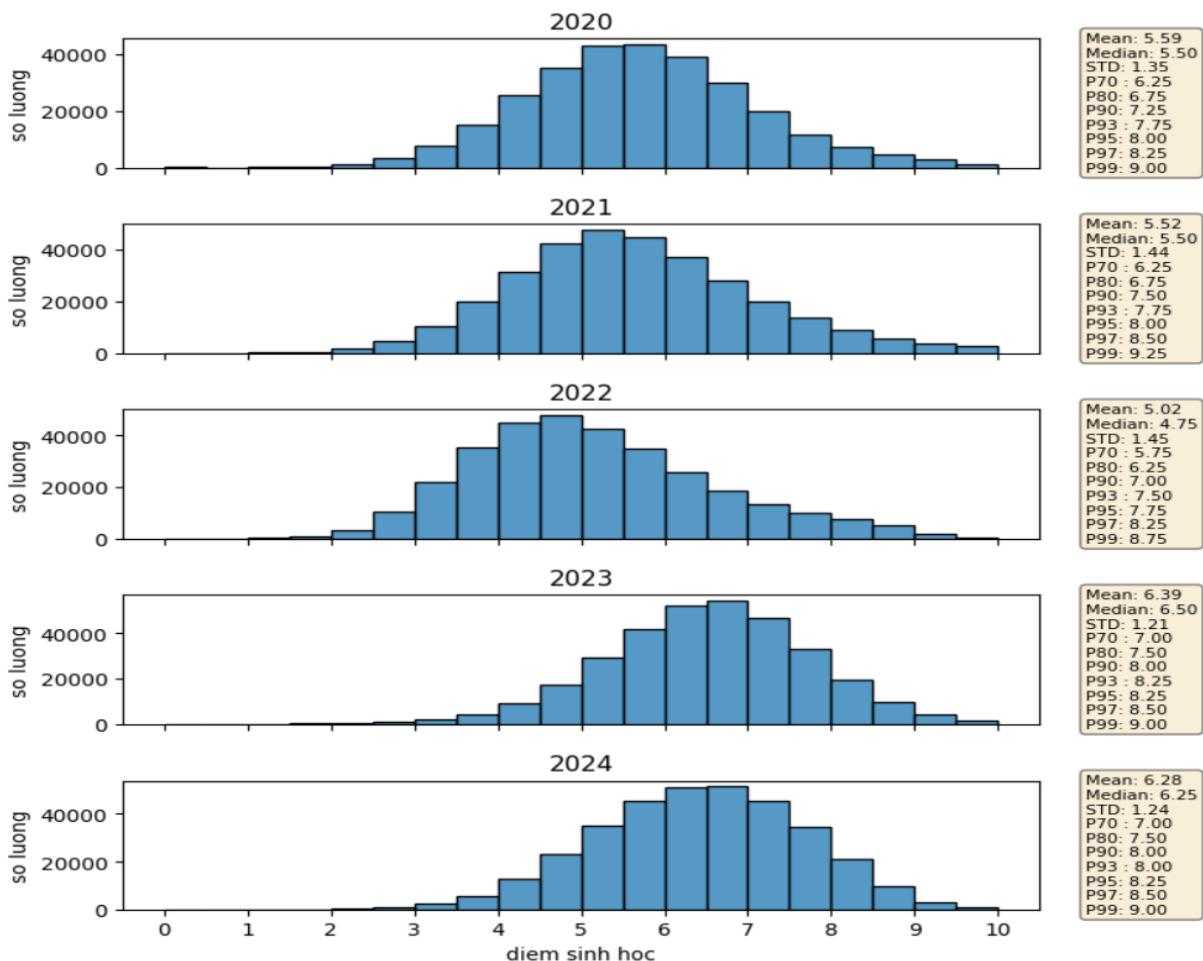
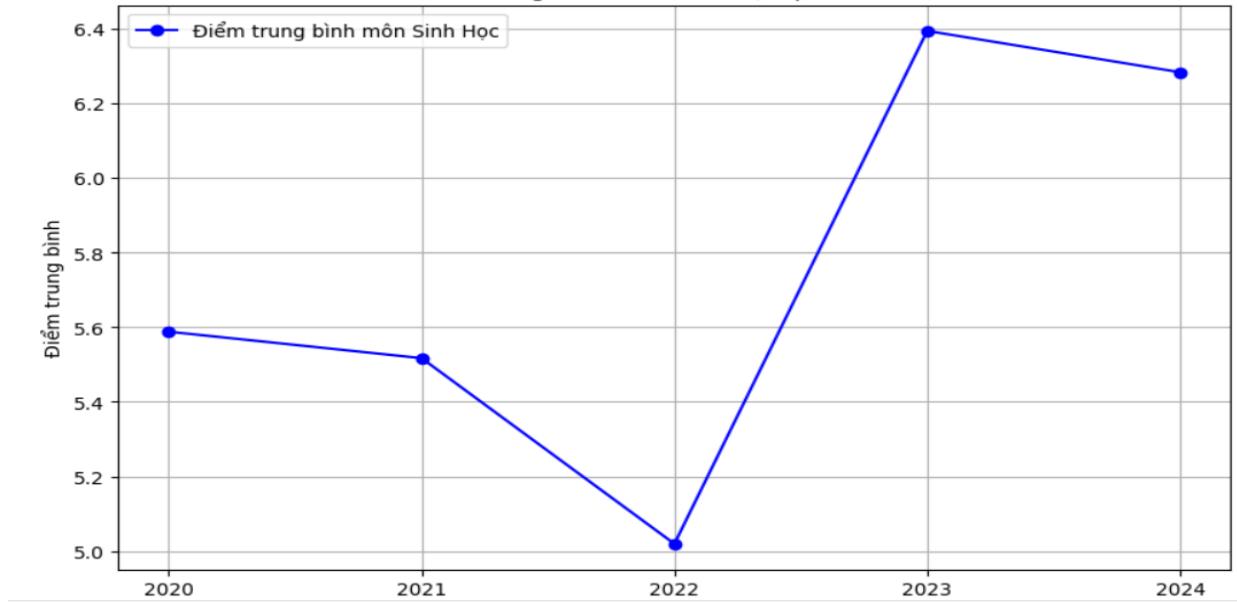


*Điểm Hóa Học

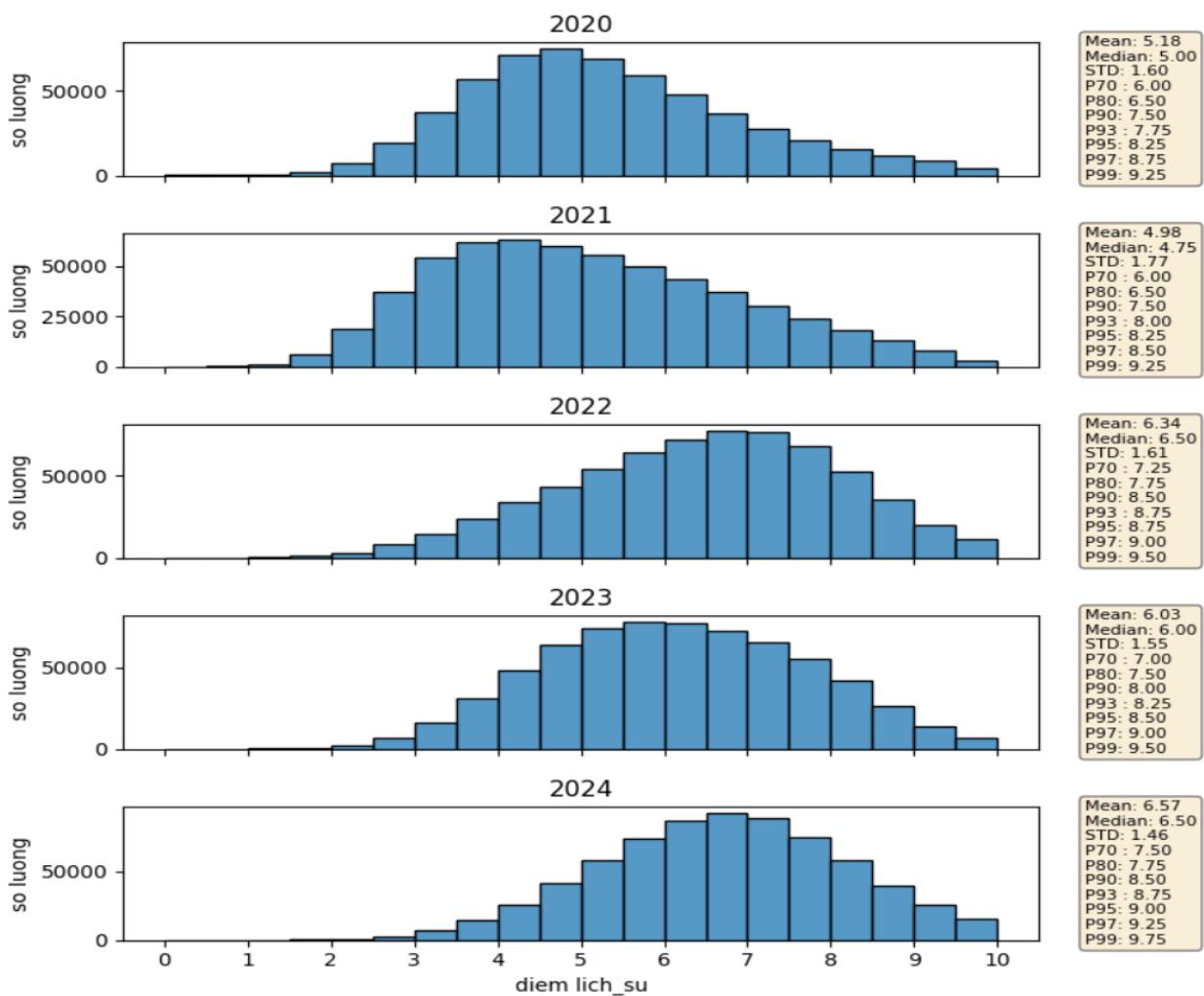
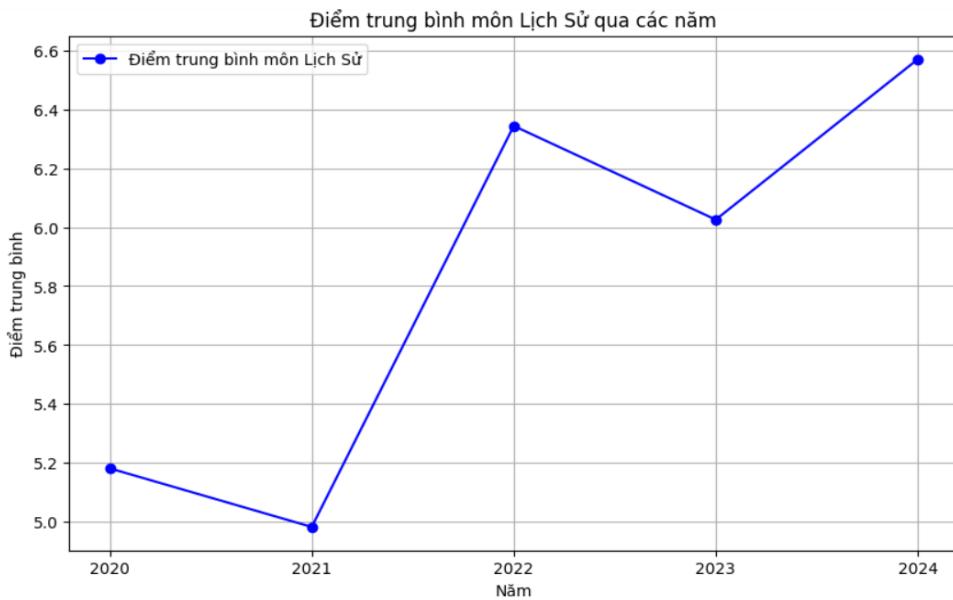


*Điểm Sinh Học

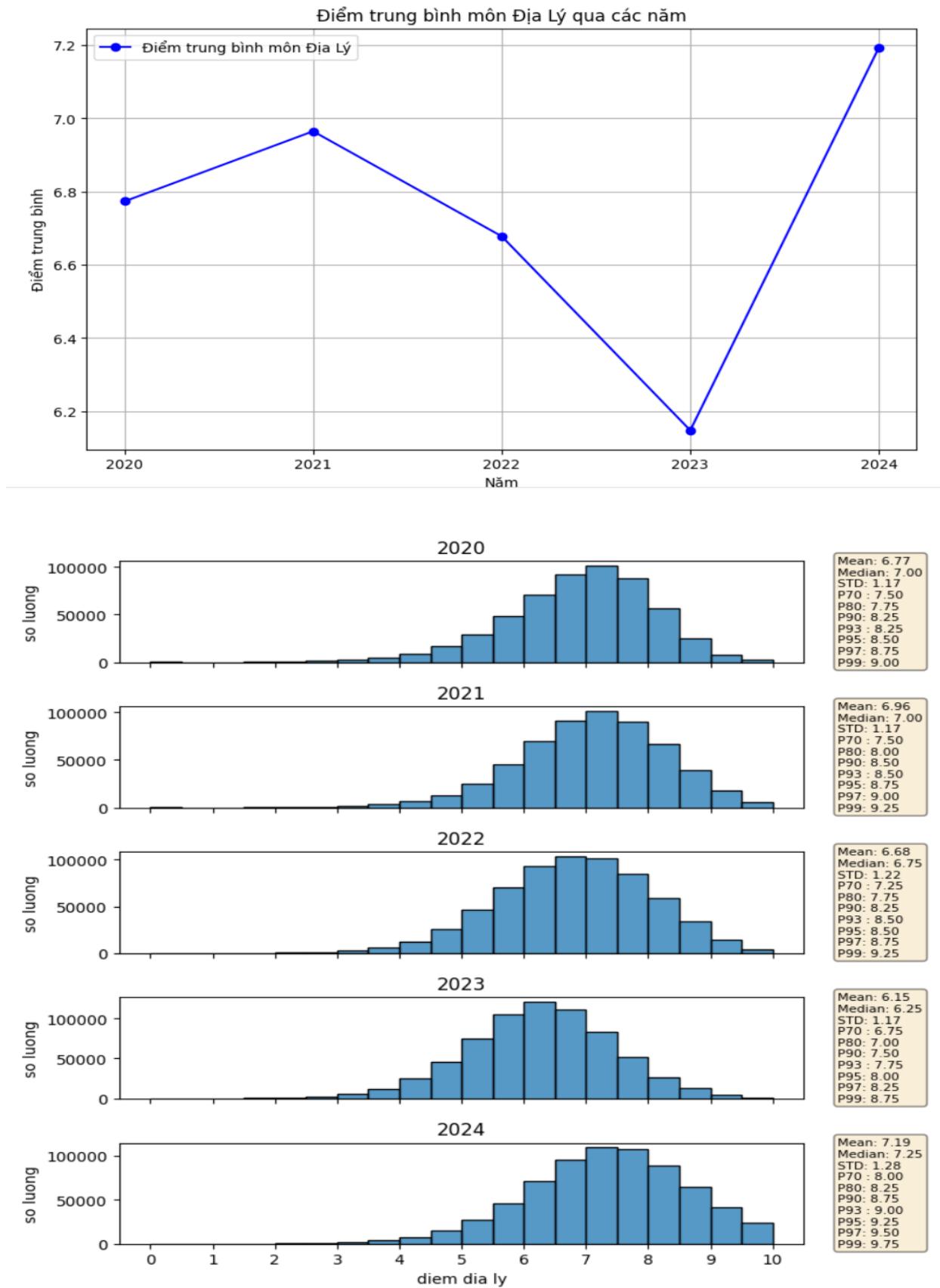
Điểm trung bình môn Sinh Học qua các năm



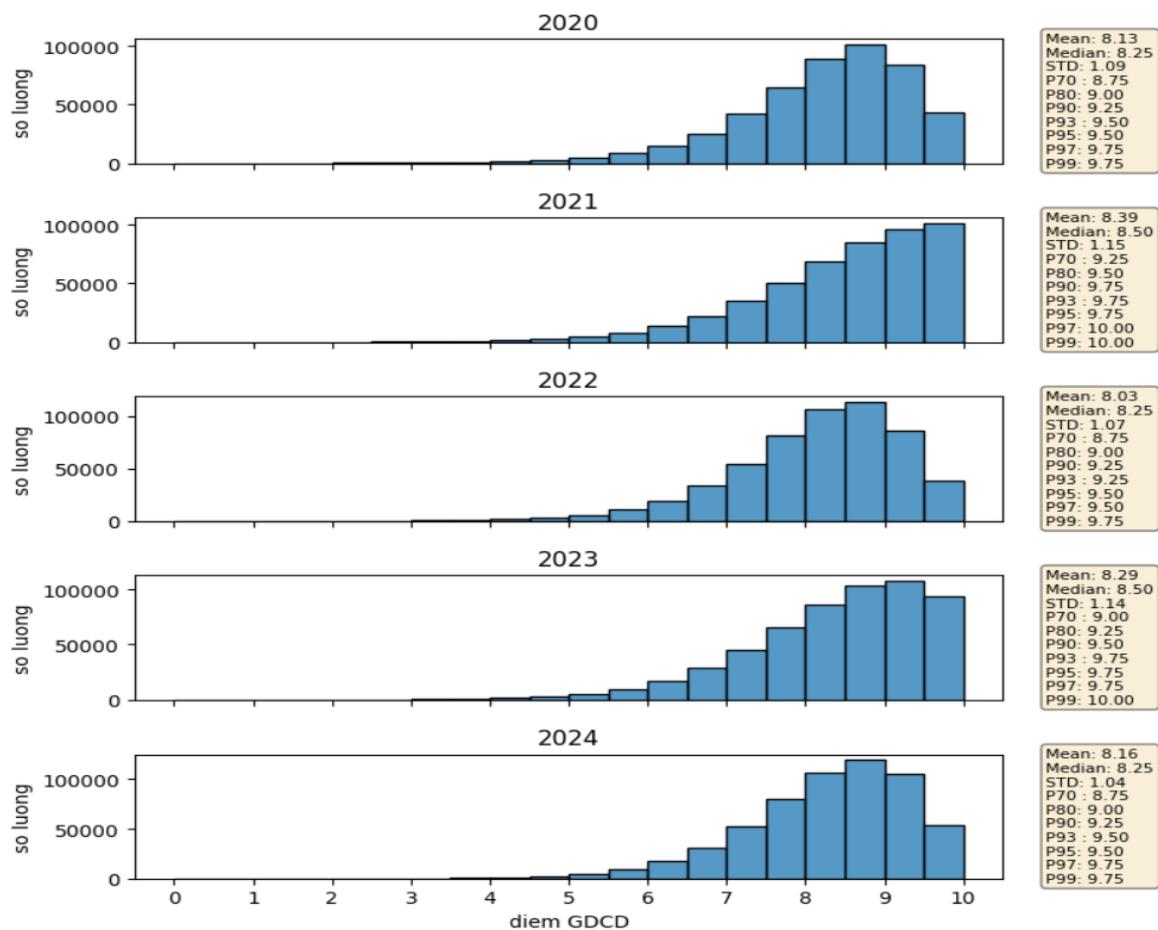
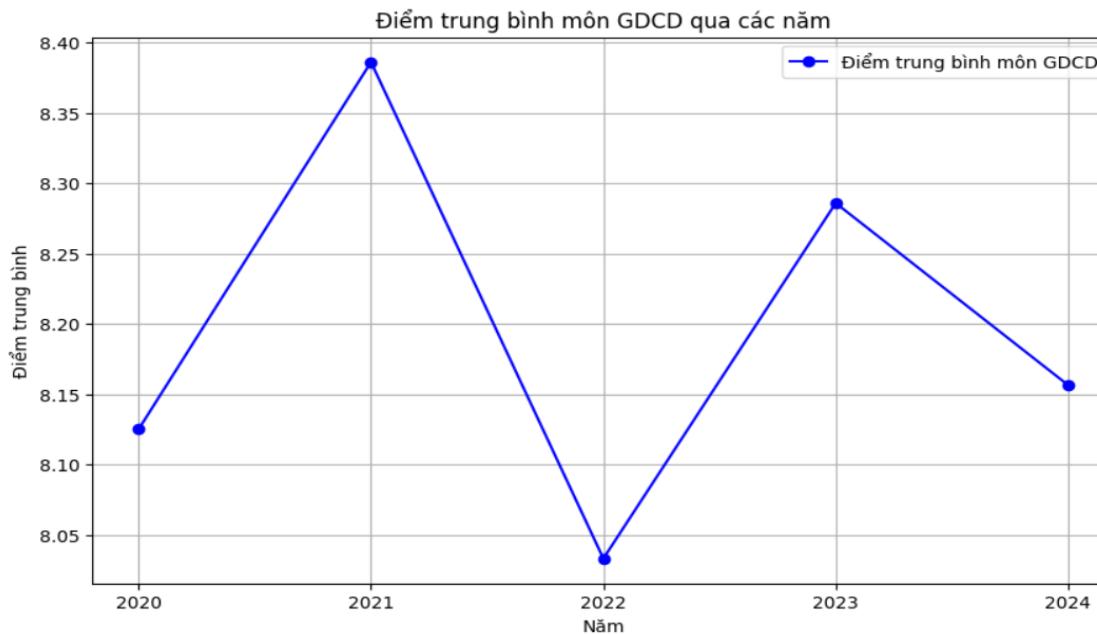
*Điểm Lịch Sử



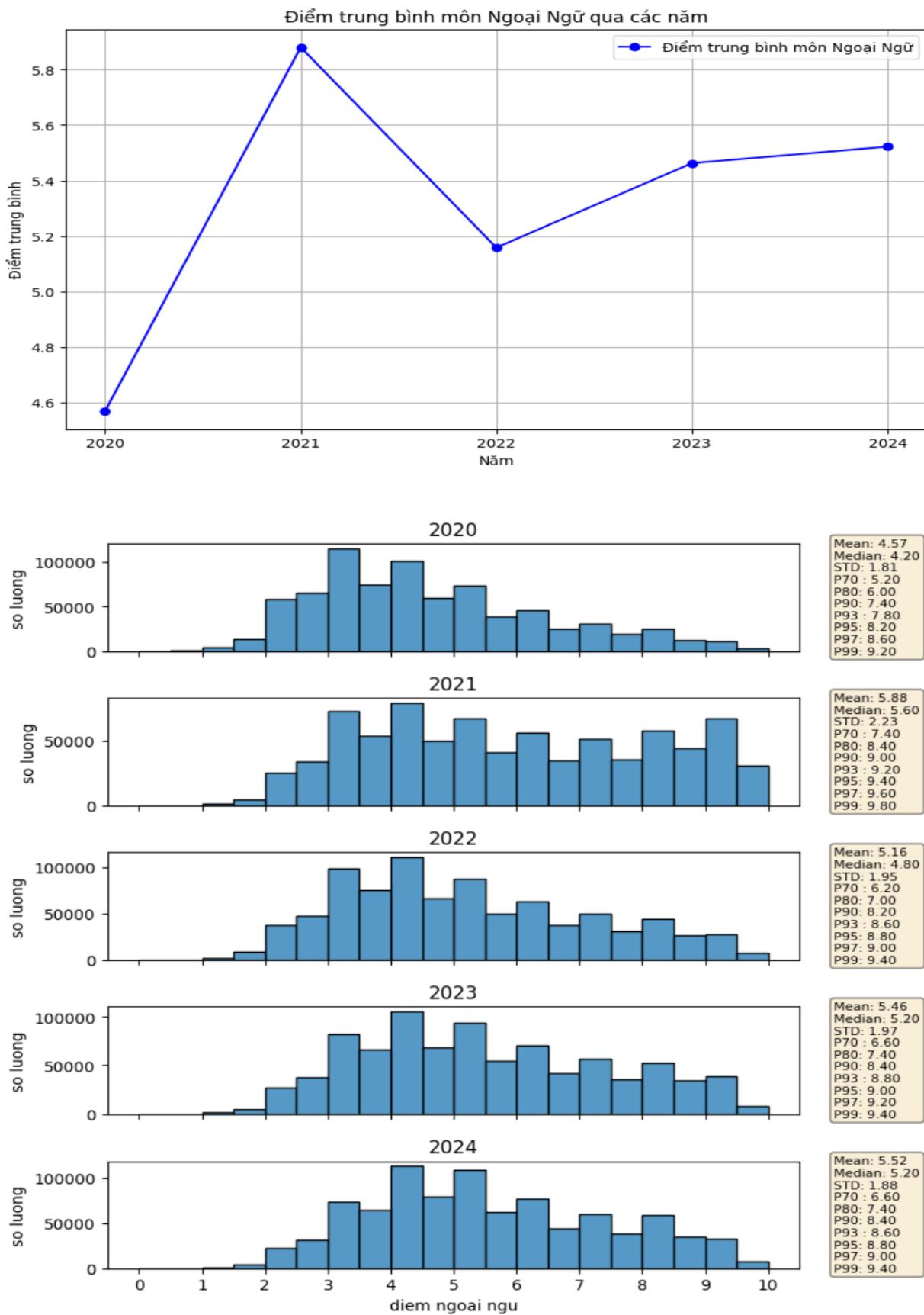
*Địa Lý



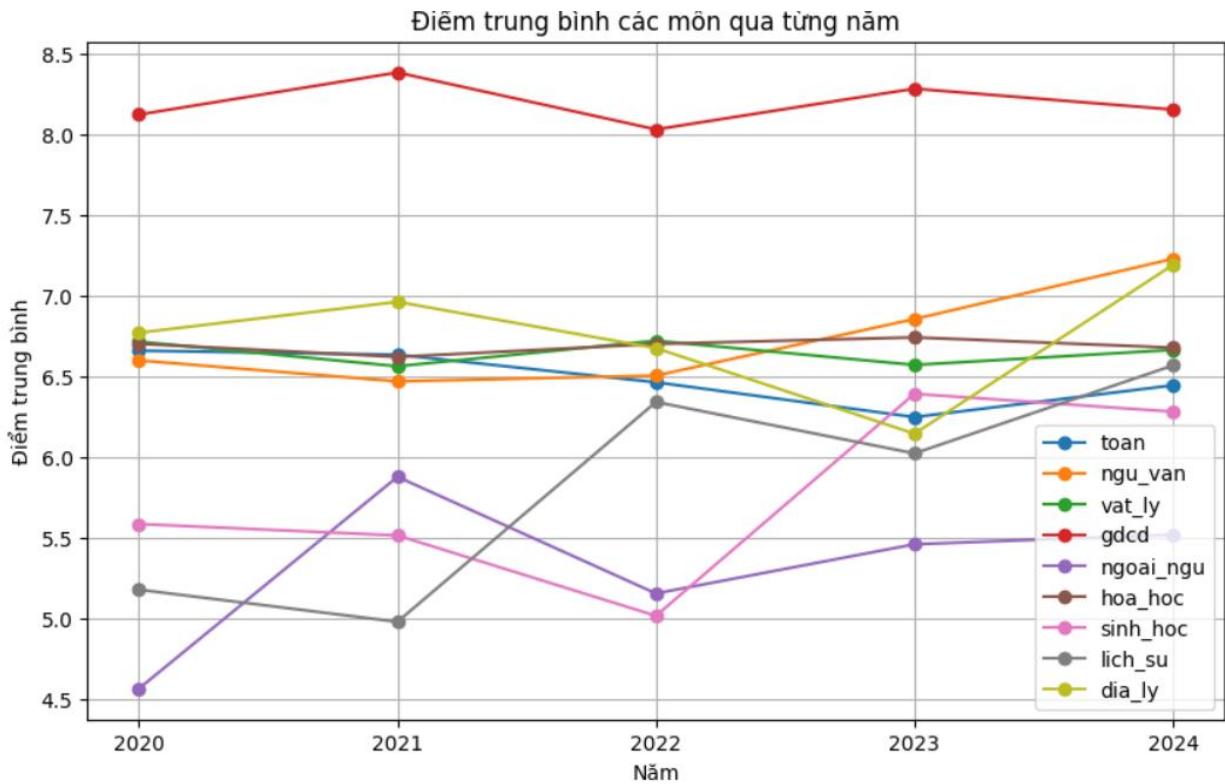
*GDCD



*Điểm Ngoại Ngữ



*Điểm trung bình các môn qua từng năm



```
# Thiết lập dữ liệu cho từng năm
data_2020 = data_all[data_all['year'] == "2020"]
data_2021 = data_all[data_all['year'] == "2021"]
data_2022 = data_all[data_all['year'] == "2022"]
data_2023 = data_all[data_all['year'] == "2023"]
data_2024 = data_all[data_all['year'] == "2024"]

# Danh sách môn học và màu sắc
subjects = ['toan', 'ngu_van', 'vat_ly', 'hoa_hoc', 'sinh_hoc', 'lich_su', 'gdcd', 'ngoai_ngu']
colors = ['#FF6347', '#4682B4', '#32CD32', '#FFD700', '#8A2BE2', '#DC143C', '#FF4500', '#2E8B57', '#D2691E']
years = [2020, 2021, 2022, 2023, 2024]

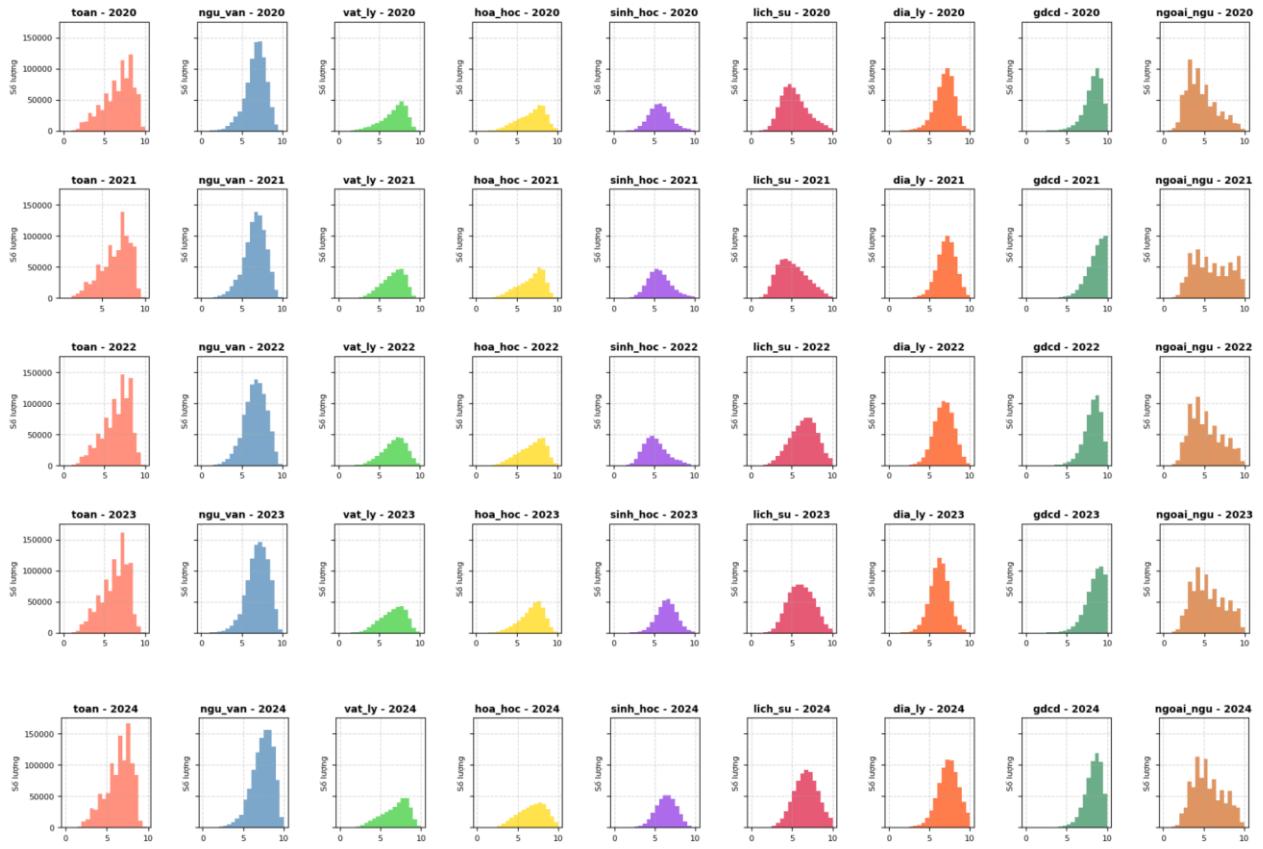
# Tạo lưới 5 hàng x 9 cột
fig, axes = plt.subplots(5, 9, figsize=(18, 12), sharey=True)

# Lặp qua từng năm và từng môn để tạo biểu đồ
for j, year in enumerate(years):
    data_year = data_all[data_all['year'] == str(year)]
    for i, subject in enumerate(subjects):
        ax = axes[j, i] # Truy cập vào trực con tại vị trí [j, i]
        color = colors[i]

        # Vẽ biểu đồ histogram cho từng môn học của từng năm
        ax.hist(data_year[subject].dropna(), bins=20, alpha=0.7, color=color)
        ax.set_title(f'{subject} - {year}', fontsize=10, fontweight='bold')
        ax.set_ylabel('Số lượng', fontsize=8)
        ax.grid(True, linestyle='--', alpha=0.5)
        ax.tick_params(axis='both', labelsize=8)

# Điều chỉnh khoảng cách giữa các biểu đồ
plt.tight_layout(pad=3.0)
plt.show()
```

Dùng biểu đồ histogram để vẽ phổ điểm từng môn theo từng năm để so sánh



*Phân tích

a. Phân tích xu hướng từng môn học

1. Môn Toán:

- Điểm trung bình của môn Toán tương đối ổn định trong cả 5 năm, duy trì trong khoảng từ 6.0 đến 6.5.
- Đây là môn có sự dao động nhỏ và không ghi nhận những bước tăng trưởng hoặc suy giảm đáng kể. Điều này có thể phản ánh sự ổn định trong chất lượng giảng dạy và khả năng học tập của học sinh.
- Tuy nhiên, điểm trung bình môn Toán chưa nằm trong nhóm dẫn đầu, cho thấy cần tiếp tục cải thiện để đẩy mạnh kết quả học tập môn học quan trọng này.

2. Môn Ngữ Văn:

- Ngữ Văn là môn có sự cải thiện ấn tượng nhất về điểm trung bình qua các năm. Điểm tăng từ mức khoảng 7.5 vào năm 2020 lên gần 8.5 vào năm 2024, biến môn này trở thành môn có điểm trung bình cao nhất trong biểu đồ.
- Xu hướng này cho thấy sự hiệu quả trong các phương pháp giảng dạy hoặc việc học sinh ngày càng chú trọng hơn vào môn học này. Ngữ Văn có thể được xem là thế mạnh cần phát huy thêm.

3. Môn Vật Lý:

- Điểm trung bình của môn Vật Lý dao động trong khoảng 6.5 đến 7.0 qua các năm, cho thấy sự ổn định nhưng không có nhiều đột phá.
- Duy trì mức điểm này giúp Vật Lý nằm trong nhóm các môn có thành tích khá tốt, tuy nhiên, sự thiếu biến động cũng đặt ra câu hỏi về việc cần đổi mới để tạo bước phát triển vượt trội hơn.

4. Môn Giáo dục công dân (GDCD):

- GDCD luôn giữ vững điểm trung bình ở mức cao từ 7.0 đến gần 8.0, là một trong những môn có thành tích xuất sắc nhất suốt cả giai đoạn.
- Mức điểm này thể hiện sự quan tâm của cả giáo viên và học sinh đối với môn học, cũng như sự liên kết tốt giữa lý thuyết và thực tiễn trong giảng dạy.

5. Môn Ngoại Ngữ:

- Ngoại Ngữ bắt đầu với mức điểm trung bình thấp hơn so với nhiều môn khác (dưới 6.0 vào năm 2020), nhưng đã ghi nhận sự tăng trưởng mạnh mẽ, đặc biệt trong hai năm 2023 và 2024, vượt qua nhiều môn để đạt mức gần 7.5.
- Đây là tín hiệu tích cực, cho thấy học sinh dần cải thiện khả năng ngoại ngữ - một kỹ năng quan trọng trong bối cảnh hội nhập quốc tế.

6. Môn Lịch Sử:

- Lịch Sử có điểm khởi đầu thấp nhất (4.5 vào năm 2020), phản ánh sự yếu kém trong việc học tập môn này ở giai đoạn đầu.
- Tuy nhiên, từ năm 2022 đến 2024, điểm số đã tăng mạnh, đạt gần 6.0 vào năm 2024, chứng tỏ sự cải thiện đáng kể, có thể nhờ vào các chính sách nâng cao nhận thức về môn học.

7. Môn Sinh Học:

- Sinh Học có mức điểm trung bình dao động nhẹ trong giai đoạn 2020-2022, nhưng ghi nhận sự tăng trưởng mạnh trong hai năm cuối, từ khoảng 5.0 lên hơn 6.0.
- Xu hướng này cho thấy những nỗ lực cải thiện chất lượng giảng dạy đã đạt được kết quả tích cực.

8. Môn Địa Lý và Hóa Học:

- Cả hai môn đều duy trì sự ổn định, với mức điểm trung bình dao động từ 5.5 đến 6.5 qua các năm.
- Mặc dù không có sự đột phá rõ rệt, hai môn này vẫn nằm trong vùng an toàn, không rơi vào nhóm thấp.

b. So sánh tổng quan các môn học

- **Ngữ Văn** là môn có sự cải thiện tốt nhất và dẫn đầu về điểm trung bình vào năm 2024.
- **Lịch Sử và Sinh Học** là hai môn khởi đầu yếu kém nhưng đã có những bước tiến đáng kể, đặc biệt từ năm 2022 trở đi.
- **Toán, Vật Lý, và Hóa Học** tuy ổn định nhưng thiếu sự đột phá, cần có những giải pháp đổi mới để nâng cao chất lượng.

- Ngoại Ngữ và GD&CD đều cho thấy xu hướng cải thiện tích cực, với Ngoại Ngữ ghi nhận sự tăng trưởng mạnh về điểm trung bình trong những năm cuối.

E. Phân tích theo số lượng thí sinh

```

data_year = data_all
#Tính toán số lượng thí sinh mỗi năm
students_per_year = data_year['year'].value_counts().sort_index()

#Tính tỉ lệ tăng trưởng
growth_rate = students_per_year.pct_change() * 100

#Tạo Dataframe để hiển thị kết quả
students_growth = pd.DataFrame({
    'Số lượng thí sinh': students_per_year,
    'Tỉ lệ tăng trưởng năm': growth_rate #·Corrected·column·name·here
}).reset_index().rename(columns={'index': 'year'})

#Vẽ biểu đồ
fig, ax = plt.subplots(2, 1, figsize=(6, 6))

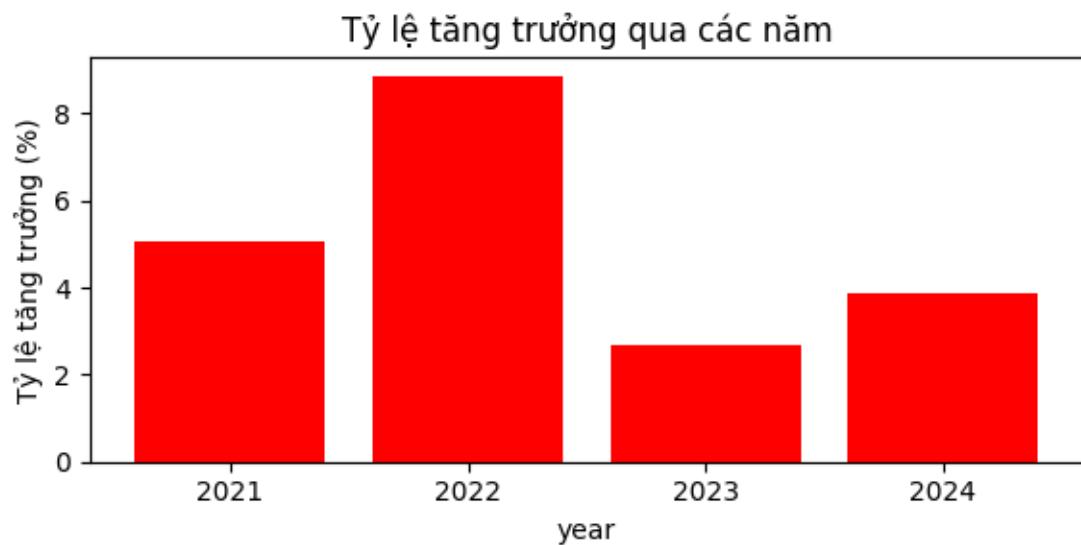
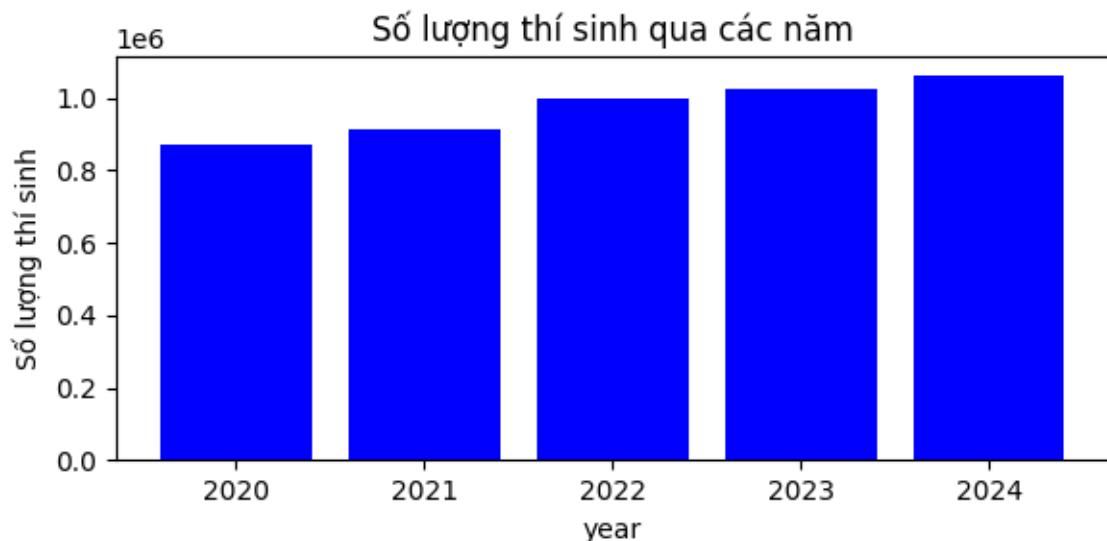
#Vẽ số lượng thí sinh
ax[0].bar(students_growth['year'], students_growth['Số lượng thí sinh'], color='blue')
ax[0].set_xlabel('year')
ax[0].set_ylabel('Số lượng thí sinh')
ax[0].set_title('Số lượng thí sinh qua các năm')
ax[0].grid(False)

#Vẽ tỉ lệ tăng trưởng
ax[1].bar(students_growth['year'], students_growth['Tỉ lệ tăng trưởng năm'], color='red')
ax[1].set_xlabel('year')
ax[1].set_ylabel('Tỷ lệ tăng trưởng (%)')
ax[1].set_title('Tỷ lệ tăng trưởng qua các năm')
ax[1].grid(False)

plt.tight_layout()
plt.show()

```

Đếm số lượng thí sinh và tính toán tỉ lệ tăng trưởng thí sinh qua từng năm



Số lượng thí sinh tăng dần qua các năm chúng ta có thể thấy điểm chuẩn xét đỗ đại học từng năm luôn thay đổi tăng nhẹ 1 ít so với các năm trước.

Nhìn trong bảng tăng trưởng thí sinh vào năm 2022 có sự tăng trưởng mạnh lên đến hơn 8% thì như chúng ta đã biết năm 2021 thì sau khi báo điểm THPT quốc gia. Tất cả các thí sinh sinh năm 2003 lúc đó là lớp 12 thi đều được điểm cao do đó có rất nhiều trường hợp thi đc 27-28đ xét tuyển đại học bị trượt. Vì vậy rất nhiều thí sinh sinh năm 2003 quyết định thi lại năm sau cùng khóa sinh năm 2004 làm cho số lượng thí sinh tăng mạnh lên như vậy.

