Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM

Khoa Công nghệ Thông tin

---o0o---

ĐỒ ÁN CUỐI KÌ

Môn: Trực Quan Hóa Dữ Liệu

Giáo Viên Hỗ Trợ:
- **Bùi Tiến Lên**
- **Lê Ngọc Thành**

Sinh Viên:
- **Vương Thành An**
- **Ngô Quốc Phát**
- **Trần Minh Thiện**
- **Lê Tâm Anh**

Mục lục

1. Tổng quan:

1.1 Thông tin nhóm:

STT	MSSV	Họ và Tên
1	19127330	Lê Tâm Anh
2	19127326	Vương Thành An
3	19127503	Ngô Quốc Phát
4	19127281	Trần Minh Thiện

1.2 Đánh giá mức độ hoàn thành của mỗi yêu cầu:

STT	Yêu cầu	Hoàn thành
1	Lựa chọn dữ liệu	100%
2	Khám phá dữ liệu	100%
3	Thống kê mô tả dữ liệu	100%
4	Trực quan hóa dữ liệu	100%
5	Phát triển mô hình học máy	100%

1.3 Mức độ hoàn thành của từng thành viên:

MSSV	Họ và Tên	Nhiệm vụ	Hoàn thành
19127326	Vương Thành An	- Lấy và tiền xử lí dữ liệu - Trực quan hóa dữ liệu - Khám phá dữ liệu	100%
19127330	Lê Tâm Anh	- Trực quan hóa dữ liệu - Tìm hiểu thông tin về dữ liệu	100%
19127503	Ngô Quốc Phát	- Trực quan hóa dữ liệu - Viết báo cáo - Tìm hiểu về Tableau	100%
19127281	Trần Minh Thiện	- Lấy dữ liệu - Trực quan hóa dữ liệu	100%

2. Thu Thập Dữ Liệu

2.1. Dữ liệu của làm về chủ đề gì ? Bạn lấy được nguồn của dữ liệu từ đâu?

- Đồ án nói về dữ liệu : dữ liệu về điểm thi của sinh viên năm 2021 gồm các môn trong kỳ thi cuối cấp trung học phổ thông để xét tốt nghiệp và xét tuyển vào các trường đại học trong đó có các môn : Toán, Văn, Lý,Hóa, Sinh, Địa Lý, Lịch Sử, Giáo Dục Công Dân, và các môn ngoại ngữ như : Anh,Pháp, Đức, Trung Quốc, Nga, Nhật Bản.

- Dữ liệu này thuộc về TÙNG DƯƠNG BÙI (link kaggle : https://www.kaggle.com/datasets/tdbui1209/vietnam-national-hs-

graduation-examination-2021?select=update.csv&fbclid=IwAR39wx-vMSJLZF-sAxXtP4ZLNX_rw4XYEH0Xe0SgR1obY8d50pGGLbdXyso), nguồn mà tác giả ghi trong phần SOURCE: https://moet.gov.vn/Pages/home.aspx)

2.2. Tác giả có cho phép bạn được sử dụng dữ liệu của họ không ?

- Theo như điều khoản của chủ sở hữu thì data được chia sẻ rộng rãi và mục đích học tập, nghiên cứu không vi phạm điều khoản của chủ sỡ hữu đưa ra

3. Load dữ liệu vào:

- Dữ liệu có 2 file csv:
  - data.csv chứa file điểm và thông tin thí sinh
  - province.csv chứa id và tên của 63 tỉnh thành ở Việt Nam nơi mà các thí sinh tham gia thi

3.1. Load file data.csv

```
data_df = pd.read_csv('data.csv',error_bad_lines=False)
data_df.head()
```

	id_examinee	math	physics	chemistry	biology	history	geography	literature	civic_education
0	1000002	9.2	NaN	NaN	NaN	5.75	9.75	8.25	9.25
1	1000003	4.4	NaN	NaN	NaN	4.25	4.00	6.25	NaN
2	1000004	8.4	4.00	3.00	3.50	NaN	NaN	6.75	NaN
3	1000005	8.8	8.25	5.75	5.25	NaN	NaN	8.25	NaN
4	1000006	8.0	NaN	NaN	NaN	5.00	6.50	8.75	9.25

3.2. Load file province.csv

```
province_df = pd.read_csv('province.csv',error_bad_lines=False)
province_df.head()
```

	id_province	name_province
0	1	ha noi
1	2	ho chi minh
2	3	hai phong
3	4	da nang
4	5	ha giang

3.3. Kết hợp dữ liệu từ province_df vào data_df

- Bởi vì 1 hoặc 2 số đầu tiên của examinee_id(số báo danh của thí sinh) thể hiện lên id của tỉnh thành nơi mà thí sinh dự thi. Do đó ta sẽ từ examinee_id để suy ra được id_province của thí sinh bằng cách chia examinee cho 1000000 và đưa về kiểu int
- Sau đó ta sẽ join 2 dataframe lại với nhau thông qua cột id_province của 2 dataframe từ đây data_df sẽ có thêm 2 cột là id_province và name_province ta sẽ biết được tên tỉnh mà thí sinh dự thi.

```
data_df['id_province']=0
data_df['id_province']=(data_df.id_examinee/1000000).astype(int)
```

```
data_df=pd.merge(data_df, province_df, how='inner', left_on = 'id_province', right_on = 'id_province')
```

3.4. Tìm vùng miền mà thí sinh dự thi

- Từ các tỉnh mà thí sinh tham gia dự thi ta có thể suy ra được thí sinh dự thi ở miền nào của Việt Nam và đất nước Việt Nam được chia làm 3 miền:
  - Miền Trung hiện có 19 tỉnh thành phố gồm: Thanh Hoá, Nghệ An, Hà Tĩnh, Quảng Bình, Quảng Trị và Thừa Thiên-Huế, Kon Tum, Gia Lai, Đắc Lắc, Đắc Nông và Lâm Đồng, Đà Nẵng, Quảng Nam, Quảng Ngãi, Bình Định, Phú Yên, Khánh Hoà, Ninh Thuận và Bình Thuận
  - Miền Bắc gồm 25 tỉnh: Lào Cai, Yên Bái, Điện Biên, Hoà Bình, Lai Châu, Sơn La; Tỉnh Hà Giang, Cao Bằng, Bắc Kạn, Lạng Sơn, Tuyên Quang, Thái Nguyên, Phú Thọ, Bắc Giang, Quảng Ninh; Tỉnh Bắc Ninh, Hà Nam, Hà Nội, Hải Dương, Hải Phòng, Hưng Yên, Nam Định, Ninh Bình, Thái Bình, Vĩnh Phúc
  - Miền Nam gồm 19 tỉnh: Bình Phước, Bình Dương, Đồng Nai, Tây Ninh, Bà Rịa-Vũng Tàu và Thành phố Hồ Chí Minh, Long An, Đồng Tháp, Tiền Giang, An Giang, Bến Tre, Vĩnh Long,

Trà Vinh, Hậu Giang, Kiên Giang, Sóc Trăng, Bạc Liêu, Cà Mau và Thành phố Cần Thơ.

- Ở đây em liệt kê ra danh sách tỉnh thuộc từng miền dựa theo trang web https://en.wikipedia.org/wiki/Vietnam
- Từ đó ta sử dụng phương thức apply hàm find_region để kiểm tra từng name_province của mỗi thí sinh để tìm ra vùng miền mà thí sinh dự thi rồi gán vào cột name_region

```
central_list=['thanh hoa','nghe an', 'ha tinh','quang binh','quang tri','thua thien - hue','kon tum',
              'gia lai','dak lak','dak nong','lam dong','da nang','quang nam','quang ngai','binh dinh',
              'phu yen','khanh hoa','ninh thuan','binh thuan']
northern_list=['lao cai','yen bai','dien bien','hoa binh','lai chau','son la','ha giang','cao bang','bac kan','lang son',
               'tuyen quang','thai nguyen','phu tho','bac giang','quang ninh','ha nam','ha noi','hai duong','hai phong',
               'hung yen','nam dinh','ninh binh',' thai binh','vinh phuc']
southern_list=set(province_df.name_province)-set(central_list)
southern_list=list(southern_list-set(northern_list))
```

```
def find_region(province_name):
    if province_name in central_list:
        return 'central'
    elif province_name in northern_list:
        return 'northern'
    else:
        return 'southern'
```

```
data_df['name_region']=data_df['name_province'].apply(find_region)
```

```
display(data_df[['id_examinee','name_province','name_region']].head())
```

	id_examinee	name_province	name_region
0	1000002	ha noi	northern
1	1000003	ha noi	northern
2	1000004	ha noi	northern
3	1000005	ha noi	northern
4	1000006	ha noi	northern

3.5. Tìm ra tổ hợp môn dự thi của mỗi thí sinh

- Vì đối tượng tham gia thi kì thi Trung Học Phổ Thông quốc gia đa số là những học sinh lớp 12, các bạn này có thể chọn tổ hợp môn để dự thi là tổ hợp tự nhiên hoặc tổ hợp xã hội hoặc cả hai. Trường hợp số ít thí sinh còn lại thì sẽ là thí sinh tự do những thí sinh này được tự do lựa chọn môn để thi.
- Do đó ở đây bọn em sẽ tìm ra tổ hợp môn dự thi của các thí sinh để có thêm thông tin để phân tích dựa trên việc các môn mà điểm số của các bạn không bị missing.
- Cụ thể ở đây em sẽ chia thành 3 nhóm tổ hợp mà thí sinh sẽ thi cùng 3 môn thi bắt buộc là buộc Toán, Văn, Anh Văn:
  - science: gồm 3 môn thuộc tổ hợp khoa học tự nhiên: Vật Lý, Hóa Học, Sinh Học
  - social: gồm 3 thuộc tổ hợp khoa học xã hội: Lịch Sử, Địa Lý, Giáo Dục Công Dân
  - other: trường hợp thí sinh thi cả 2 tổ hợp môn hoặc thí sinh tự do(được tự do lựa chọn những môn dự thi)
- Bước 1: Tạo một temp dataframe từ data_df sau đó drop bỏ các cột không xét bao gồm id_examinee, id và name của province, tên vùng

miền và các môn không thuộc 2 tổ hợp phía trên. Bây giờ trong temp_df sẽ chứa điểm của 6 môn trong đó 3 môn thuộc tổ hợp tự nhiên và 3 môn thuộc tổ hợp xã hội.

- Bước 2: Điền giá trị 0 vào các môn không dự thi để sort và lấy ra tên của 3 môn mà thí sinh có điểm thi cao nhất.
- Bước 3: Kiểm tra tổ hợp sử dụng hàm find_combination_sub(data):
  - Nếu 3 môn đó đều thuộc tổ hợp tự nhiên là 'physics','chemistry','biology' thì chứng tỏ tổ hợp thí sinh đó chọn thi là tổ hợp tự nhiên ('science').
  - Nếu 3 môn đó đều thuộc tổ hợp xã hội là 'history','geography','civic_education' thì chứng tỏ tổ hợp thí sinh đó chọn thi là tổ hợp xã hội ('social').
  - Nếu 3 môn đều khác với 2 tổ hợp trên thì ta sẽ cho vào nhóm 'other': có hai trường hợp cho nhóm này.
    - Thí sinh này tham gia thi cả 2 tổ hợp môn dẫn đến các môn thi cao nhất của thí sinh có sự trộn lẫn giữa tổ hợp tự nhiên và xã hội nên sẽ kiểm tra ra không đúng. Ví dụ ['physics','geography','chemistry'] ==> không thuộc tổ hợp nào.
    - Thí sinh là thí sinh thi tự do dẫn đến số môn thi không đủ để xếp vào bất kì tổ hợp nào. Ví dụ thí sinh đó chỉ thi Toán, Vật Lý, Hóa Học để xét tuyển khối A00 thì số môn kiểm tra được sẽ là ['physics','chemistry'] ==> không thuộc tổ hợp nào.

```
temp_df=data_df
list_drop=['id_examinee', 'math','english','literature', 'russian',
       'french', 'chinese', 'german', 'japanese', 'id_province',
       'name_province','name_region']
temp_df=temp_df.drop(list_drop,axis=1)
temp_df.fillna(0)
data_df['com_sub']=0
data_df['com_sub']=list(temp_df.columns.values[np.argsort(-temp_df.values, axis=1)[:, :3]])
data_df['com_sub']
```

```
0         [geography, civic_education, history]
1                 [history, geography, physics]
2                 [physics, biology, chemistry]
3                 [physics, chemistry, biology]
4         [civic_education, geography, history]
                          ...
934612    [civic_education, history, geography]
934613    [civic_education, geography, history]
934614    [civic_education, history, geography]
934615    [civic_education, geography, history]
934616    [civic_education, geography, history]
Name: com_sub, Length: 934617, dtype: object
```

- Ta có thể thấy được 3 môn mà thí sinh thi cao nhất tương ứng với tổ hợp môn mà thí sinh đăng ký dự thi

```
scicent_list=['physics','chemistry','biology']
social_list=['history','geography','civic_education']
```

```
def find_combination_sub(data):
    if sorted(data) == sorted(scicent_list):
        return 'science'
    elif sorted(data) == sorted(social_list):
        return 'social'
    else:
        return 'others'
```

```
data_df['com_sub']=data_df['com_sub'].apply(find_combination_sub)
```

```
display(data_df[['id_examinee','com_sub']].head())
```

	id_examinee	com_sub
0	1000002	social
1	1000003	others
2	1000004	science
3	1000005	science
4	1000006	social

- Kết quả tổ hợp môn tìm được của 5 thí sinh đầu tiên

4. Khám phá dữ liệu

4.1. Dữ liệu gồm có bao nhiêu dòng và bao nhiêu cột?

```
num_rows,num_cols=data_df.shape
print('Dữ liệu bao gồm: {} dòng và {} cột'.format(num_rows,num_cols))
```

```
Dữ liệu bao gồm: 934617 dòng và 19 cột
```

4.2. Mỗi dòng có ý nghĩa gì? Có vấn đề các dòng có ý nghĩa khác nhau không?

- Mỗi dòng thể hiện các thông tin về thông tin và điểm số của một thí sinh tham gia dự thi kì thì tốt nghiệp Trung Học Phổ Thông quốc gia. Có vẻ không có dòng nào khác loài.

4.3. Dữ liệu có các dòng bị lặp không?

```
have_duplicated_rows=data_df.duplicated().any()
print(have_duplicated_rows)
```

```
False
```

4.4. Mỗi cột có ý nghĩa gì?

- id_examinee: số báo banh của thí sinh
- math: Điểm môn Toán của thí sinh (0-10)
- physics: Điểm môn Vật Lý của thí sinh (0-10)
- chemistry: Điểm môn Hóa của thí sinh (0-10)
- biology: Điểm môn Sinh của thí sinh (0-10)
- history: Điểm môn Sử của thí sinh (0-10)
- geography: Điểm môn Địa Lý của thí sinh (0-10)
- literature: Điểm môn Văn của thí sinh (0-10)
- civic_education: Điểm môn Giáo Dục Công Dân của thí sinh (0-10)
- english: Điểm môn ngôn ngữ Anh của thí sinh (0-10)
- russian: Điểm môn ngôn ngữ Nga của thí sinh (0-10)
- french: Điểm môn ngôn ngữ Pháp của thí sinh (0-10)
- chinese: Điểm môn ngôn ngữ Trung Quốc của thí sinh (0-10)
- german: Điểm môn ngôn ngữ Đức của thí sinh (0-10)
- japanese: Điểm môn ngôn ngữ Nhật Bản của thí sinh (0-10)
- id_province: id của tỉnh mà thí sinh dự thi
- name_province: tên của tỉnh mà thí sinh dự thi
- com_sub: tổ hợp môn mà thí sinh lựa chọn thi cùng với 3 môn thi bắt buộc Toán, Văn, Anh Văn
    - science: gồm 3 môn thuộc tổ hợp khoa học tự nhiên: Vật Lý, Hóa Học, Sinh Học
    - social: gồm 3 thuộc tổ hợp khoa học xã hội: Lịch Sử, Địa Lý, Giáo Dục Công Dân
    - other: trường hợp thí sinh thi cả 2 tổ hợp môn hoặc thí sinh tự do(được tự do lựa chọn những môn dự thi)
- name_region: được suy ra từ tỉnh thành mà thí sinh tham gia thi theo 3 miền của nước Việt Nam
    - northern: Miền Bắc
    - central: Miền Trung
    - southern: Miền Nam

4.5. Mỗi cột hiện đang có kiểu dữ liệu gì? Có cột nào có kiểu dữ liệu chưa phù hợp để có thể xử lý tiếp không?

```
col_dtypes=data_df.dtypes
print(col_dtypes)
```

```
id_examinee          int64
math               float64
physics            float64
chemistry          float64
biology            float64
history            float64
geography          float64
literature         float64
civic_education    float64
english            float64
russian            float64
french             float64
chinese            float64
german             float64
japanese           float64
id_province          int32
name_province       object
name_region         object
com_sub             object
dtype: object
```

- Ở đây ta tìm được 3 cột có kiểu dữ liệu dạng object là name_province, com_sub, name_region. Cùng khám phá xem bên trong từng giá trị của nó có kiểu dữ liệu là gì bằng hàm open_object_dtype

```
def open_object_dtype(s):
    df=pd.Series(s)
    a=df.apply(lambda x : type(x)).unique()
    dtypes = set(a)
    return dtypes
```

```
objects_key=(data_df.loc[:, data_df.dtypes == object]).keys()
print("numer of object columns:",len(objects_key))
for key in objects_key:
    print("types of ",key," is ",open_object_dtype(data_df[key]))
```

```
numer of object columns: 3
types of  name_province  is  {<class 'str'>}
types of  name_region  is  {<class 'str'>}
types of  com_sub  is  {<class 'str'>}
```

- Như vậy giá trị ở cả 3 cột đều có kiểu dữ liệu là string. Tuy nhiên để chắc chắn có phải là string hay không thì ta in ra xem thử

```
print(data_df[['name_province','com_sub','name_region']])
```

```
       name_province  com_sub name_region
0             ha noi   social    northern
1             ha noi   others    northern
2             ha noi  science    northern
3             ha noi  science    northern
4             ha noi   social    northern
...              ...      ...         ...
934612   tuyen quang   social    northern
934613   tuyen quang   social    northern
934614   tuyen quang   social    northern
934615   tuyen quang   social    northern
934616   tuyen quang   social    northern

[934617 rows x 3 columns]
```

- Ta có thể nhìn thấy rõ ràng 3 kiểu dữ liệu này dạng string không bị trộn lẫn giữa số do đó ta thực hiện bước tiếp theo

4.6. Chuyển kiểu dữ liệu từ object sang string

- Ta sẽ kiểu dữ liệu của cột từ object thành string bằng phương thức asptype

```
data_df['name_province']=data_df['name_province'].astype('string')
data_df['com_sub']=data_df['com_sub'].astype('string')
data_df['name_region']=data_df['name_region'].astype('string')
print(data_df[['name_province','com_sub','name_region']].dtypes)
```

```
name_province    string
com_sub          string
name_region      string
dtype: object
```

4.7. Với mỗi cột có kiểu dữ liệu dạng numeric, các giá trị được phân bố như thế nào?

```
cate_cols=['id_examinee','id_province','name_province','com_sub', 'name_region']
numeric_cols=list(set(data_df.keys())-set(cate_cols))
```

```
nume_col_profiles_df=pd.DataFrame(index=['missing_ratio','min','max'],
                                  columns=numeric_cols)

nume_col_profiles_df.loc['missing_ratio',numeric_cols]=np.float64(data_df[numeric_cols].isna().mean()*100)

nume_col_profiles_df.loc['min',numeric_cols]=np.float64(data_df[numeric_cols].min())

nume_col_profiles_df.loc['max',numeric_cols]=np.float64(data_df[numeric_cols].max())

nume_col_profiles_df=nume_col_profiles_df.astype(np.float64)


print(nume_col_profiles_df)
```

	russian	french	physics	history	chemistry	\
missing_ratio	99.988979	99.917292	65.110307	35.693765	64.940719	
min	3.600000	2.000000	0.000000	0.000000	0.000000	
max	10.000000	10.000000	10.000000	10.000000	10.000000	

	geography	math	biology	literature	german	\
missing_ratio	36.289304	1.066319	65.496883	1.356277	99.988337	
min	0.000000	0.600000	0.000000	0.000000	2.200000	
max	10.000000	10.000000	10.000000	10.000000	10.000000	

	japanese	chinese	civic_education	english
missing_ratio	99.896214	99.778733	46.059509	12.556801
min	1.200000	1.200000	0.000000	0.000000
max	10.000000	10.000000	10.000000	10.000000

- Ta thấy 3 môn thi bắt buộc là Toán, Ngữ Văn, Anh Văn có tỉ lệ được thí sinh tham gia rất cao do đây là những môn bắt buộc
- Tỉ lệ thí sinh không chọn các môn thuộc tổ hợp xã hội là khoảng 35-45% thấp hơn nhiều so với các nhóm tự nhiên là ở khoảng 65%
- Tất cả các môn đều có thí sinh đạt điểm 10
- Điểm thấp nhất ở các môn chủ yếu dao động từ 0-2đ

4.8. Với mỗi cột có kiểu dữ liệu dạng categorical, các giá trị được phân bố như thế nào ?

```
def missing_ratio(s):
    return s.isna().mean() * 100
def num_diff_vals(s):
    return s.dropna().nunique()

def diff_vals(s):
    return s.dropna().unique()
```

```
index=['missing_ratio', 'num_diff_vals', 'diff_vals']
cate_col_profiles_df=pd.DataFrame(
     index=index,
     columns=cate_cols)
data_cate_col_df = data_df[cate_cols]
cate_col_profiles_df = data_cate_col_df.agg([missing_ratio,num_diff_vals,diff_vals])

print(cate_col_profiles_df)
```

```
                                                    id_examinee  \
missing_ratio                                               0.0
num_diff_vals                                            934617
diff_vals      [1000002, 1000003, 1000004, 1000005, 1000006, ...

                                                    id_province  \
missing_ratio                                               0.0
num_diff_vals                                                62
diff_vals      [1, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 2,...

                                                  name_province  \
missing_ratio                                               0.0
num_diff_vals                                                62
diff_vals      [ha noi, lang son, bac kan, thai nguyen, yen b...

                                com_sub                    name_region
missing_ratio                       0.0                            0.0
num_diff_vals                         3                              3
diff_vals      [social, others, science]  [northern, southern, central]
```

- Không có kiểu dữ liệu nào bị thiếu và không có sự bất thường ở cột id_examinee, com_sub và name_region
- Tuy nhiên số tỉnh thành ở 2 cột id_provice và name_province chỉ ra là dữ liệu chỉ có thông tin của 62 tỉnh thành trong khi Việt Nam có tận 63 tỉnh thành.

4.9. Tìm ra dữ liệu điểm thi của tỉnh nào bị thiếu

```
miss_provice=set(province_df.id_province)-set(data_df.id_province.unique())
```

```
miss_provice
```

```
{50}
```

```
miss_provice=set(province_df.name_province)-set(data_df.name_province.unique())
miss_provice
```

```
{'dong thap'}
```

- Như vậy ta không có dữ liệu của các thí sinh dự thi ở tỉnh Đồng Tháp. Tuy nhiên ở tỉnh cuối cùng của province_id được đánh số là 64 như vậy là có 1 mã số nào đó từ 1 đến 64 không được chọn làm mã tỉnh dự thi. Cùng tìm thử xem đó là con số nào

```
province_df.id_province
```

```
0      1
1      2
2      3
3      4
4      5
      ..
58    60
59    61
60    62
61    63
62    64
Name: id_province, Length: 63, dtype: int64
```

```
miss_provice_id=set(np.arange(1,65))-set(province_df.id_province.unique())
```

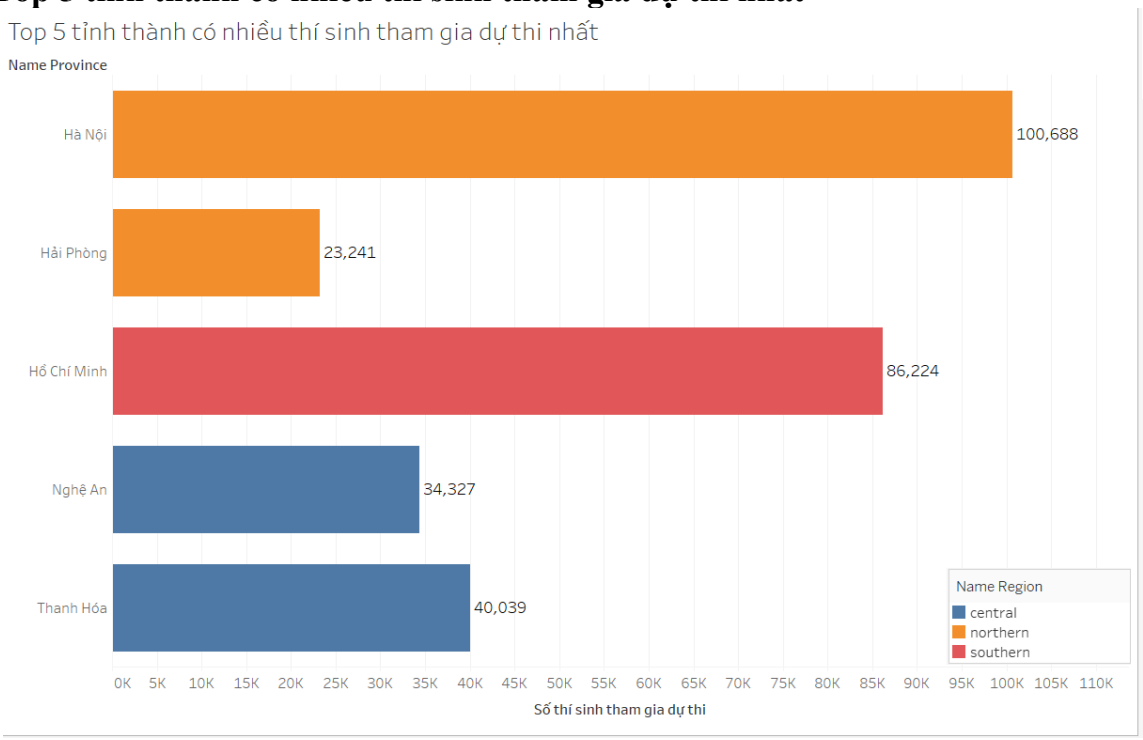
```
miss_provice_id
```

```
{20}
```

- Con số không được chọn là 20. Em đã thử google tìm danh sách mã tỉnh dự thi Trung Học Phổ Thông quốc gia ở Việt Nam thì đúng là không có tỉnh số 20. Nhưng người ta cũng không có đề cập tại sao lại bỏ đi con số này.

5. Thống kê mô tả

5.1. Top 5 tỉnh thành có nhiều thí sinh tham gia dự thi nhất

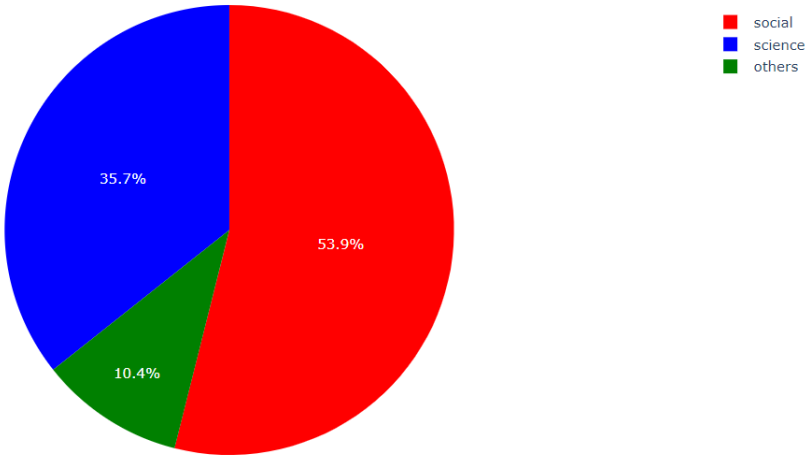


- Qua biểu đồ cho ta thấy hai thành phố trung tâm là nơi có nhiều thí sinh tham gia dự thi nhất với TP Hà Nội là hơn 100000 thí sinh và TP Hồ Chí Minh là hơn 86000 thí sinh
- Miền trung cũng có 2 tỉnh góp mặt là Thanh Hóa và Nghệ An với lần lượt là khoảng 40000 và 34000
- Miền Bắc thì còn có thêm Hải Phòng với hơn 23000. Trong khi, miền Nam chỉ có mỗi TP Hồ Chí Minh

5.2. So sánh số lượng thí sinh tham gia ở 3 miền đất nước

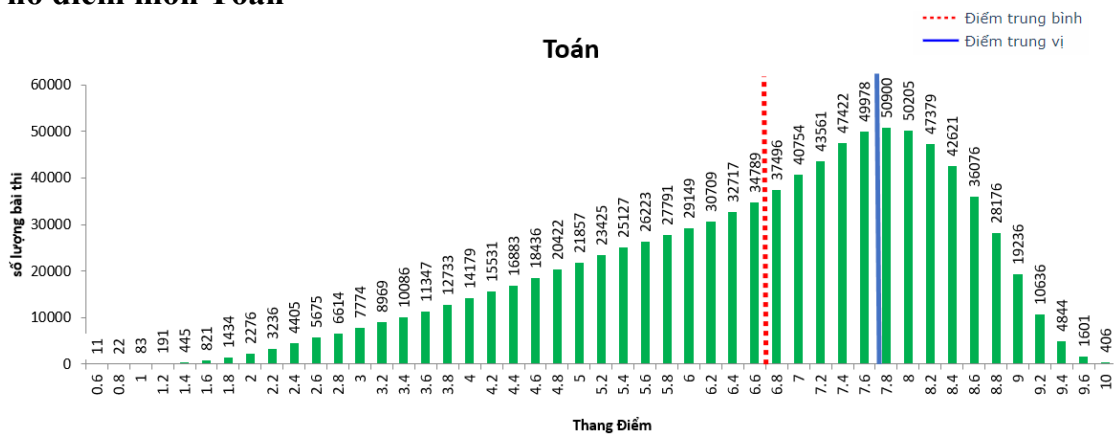
- Miền Bắc là nơi có nhiều thí sinh tham gia nhất với 36.2%
- Miền Nam xếp thứ hai với 34.6%
- Miền Trung thì ít hơn một chút là 29.2% do dân số chủ yếu tập trung ở 2 miền còn lại.

5.3. So sánh số lượng thí sinh đăng ký theo 2 loại tổ hợp và nhóm thí sinh tự do



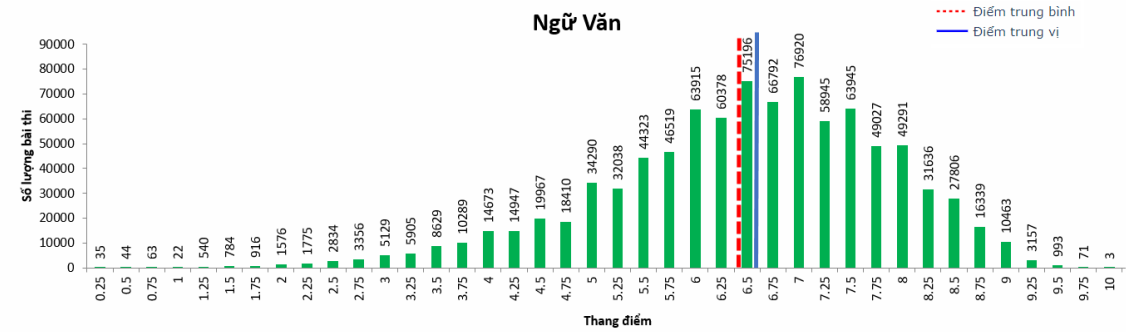
- Ta có thể số lượng học sinh dự thi tổ hợp Xã Hội có phần áp đảo với 53.9%
- Phần trăm những thí sinh tham gia tổ hợp môn Tự nhiên thì có phần ít hơn chỉ với 35.7%
- Những thí sinh tự do chiếm 10% trên tổng số

5.4. Phổ điểm môn Toán



- Ta có thể thấy phân bố điểm Toán có sự không đối xứng khi phổ điểm hơi lệch về bên phải. Dẫn đến số lượng thí sinh đạt điểm trên trung bình cao chủ yếu tập trung ở mức 7.4-8.2đ, nhưng nó giảm mạnh khi số điểm càng tiến về 10 chứng tỏ đề thi có sự phân loại rõ rệt giữa điểm 8 và điểm 9 10(giữa học sinh khá và học sinh giỏi).
- Với 924651 thí sinh tham gia thi môn Toán
- Điểm trung bình là 6.61 điểm
- Điểm trung vị là 7.0 điểm
- Điểm số có nhiều thí sinh đạt nhất là 7.8 điểm
- Phương sai là 2.83
- Độ lệch chuẩn là 1.68
- Số thí sinh có điểm <= 1 là 116 (chiếm tỷ lệ 0.01%)
- Số thí sinh có điểm dưới trung bình (<=5) là 161573 (với 17.47%)
- Số thí sinh có điểm >=9 là 36723 (với 3.97%)

5.5. Phổ điểm môn Ngữ Văn

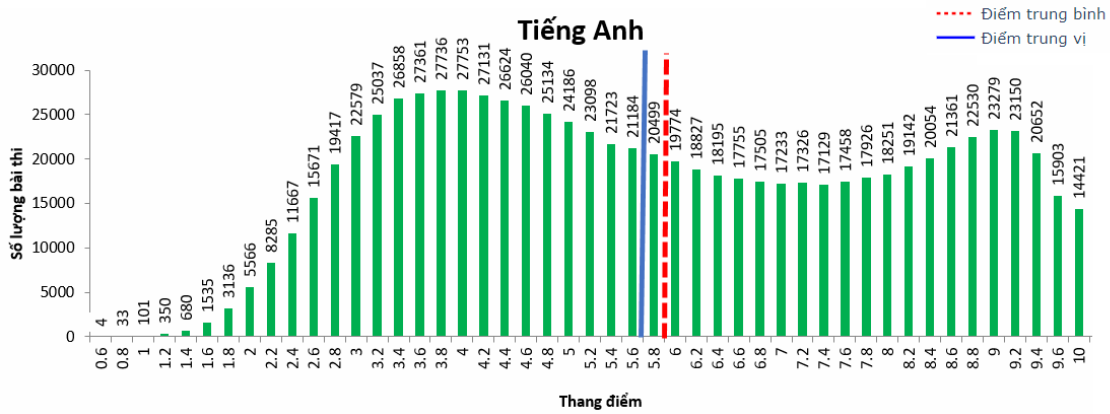


- Phổ điểm Ngữ Văn cũng có sự lệch nhẹ về bên phải mặc dù không nhiều bằng môn Toán nhưng điều đặc biệt ở phổ điểm Ngữ Văn mà tất cả các phổ điểm môn khác không có đó là phân bố theo hiệu ứng hình "răng cưa".

- Cụ thể hơn, đó là số lượng thí sinh đạt số điểm n.0 hoặc n.5 với n là số tự nhiên từ (4-8) sẽ cao hơn so với số thí sinh đạt n.25 hoặc n.75. Ví dụ, nhìn lên biểu đồ ta thấy số thí sinh đạt 6.0 nhiều hơn 6.25 và số thí sinh đạt 6.5 điểm nhiều hơn 6.75 điểm.
- Nguyên nhân là do môn Ngữ Văn là môn thi tự luận, trong đề thi sẽ có khoảng 4 câu với mức điểm tối đa là n.0 hoặc n.5 điểm và nếu thí sinh làm đúng yêu cầu thì sẽ dễ nhận được số điểm kiểu này hơn so với n.25 và n.75 điểm.

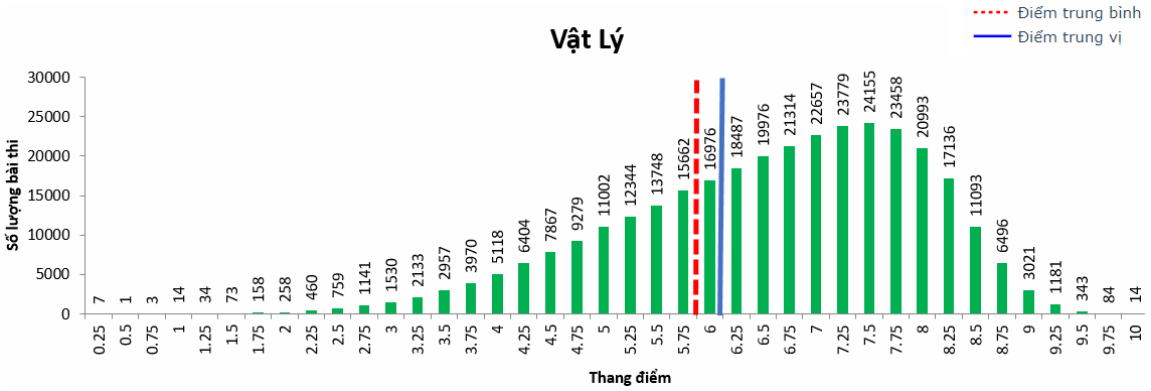
- Với 921941 thí sinh tham gia thi môn Ngữ Văn
- Điểm trung bình là 6.48 điểm
- Điểm trung vị là 6.5 điểm
- Điểm số có nhiều thí sinh đạt nhất là 7.0 điểm
- Phương sai là 1.86
- Độ lệch chuẩn là 1.36
- Số thí sinh có điểm <= 1 là 164 (chiếm tỷ lệ 0.01%)
- Số thí sinh có điểm dưới trung bình (<=5) là 110089 (với 11.94%)
- Số thí sinh có điểm >=9 là 14684 (với 1.59%)

5.6. Phổ điểm môn Tiếng Anh(Anh Văn)



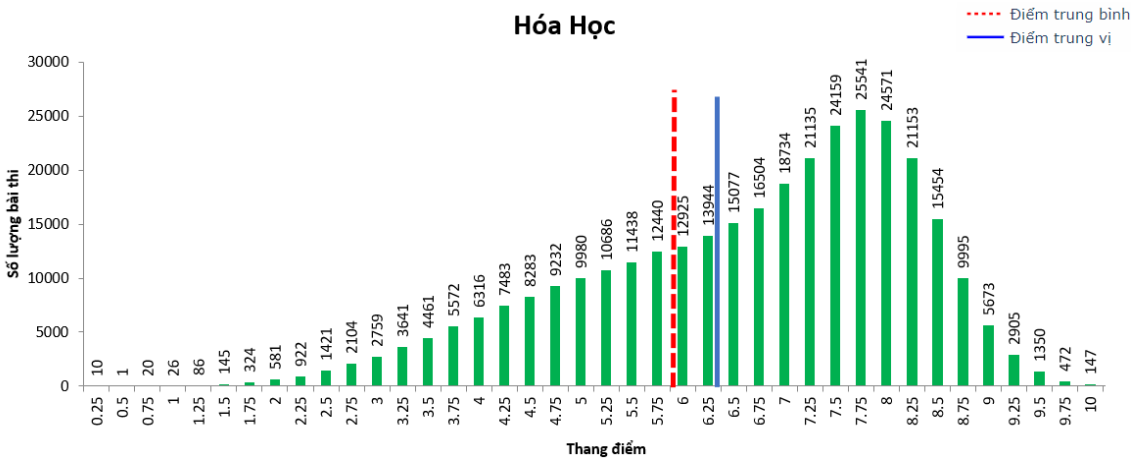
- Phổ điểm môn Anh Văn cũng có sự bất thường quá rõ ràng khi phổ điểm các môn khác khi nó thuộc dạng bimodal(có 2 đỉnh) và càng bất thường hơn khi 2 đỉnh này nằm ở hai thái cực là điểm 4 (điểm dưới trung bình) và điểm 9 (điểm giỏi) chứng tỏ có sự mất cân bằng về trình độ Tiếng Anh giữa các bạn học sinh, một nhóm thì học quá giỏi, một nhóm thì quá tệ.
- Với 817259 thí sinh tham gia thi môn Anh Văn
- Điểm trung bình là 5.85 điểm
- Điểm trung vị là 5.6 điểm
- Điểm số có nhiều thí sinh đạt nhất là 4.0 điểm
- Phương sai là 4.92
- Độ lệch chuẩn là 2.22
- Số thí sinh có điểm <= 1 là 138 (chiếm tỷ lệ 0.02%)
- Số thí sinh có điểm dưới trung bình (<=5) là 161573 (với 40.22%)
- Số thí sinh có điểm >=9 là 97405 (với 11.92%)

5.7. Phổ điểm môn Vật Lý



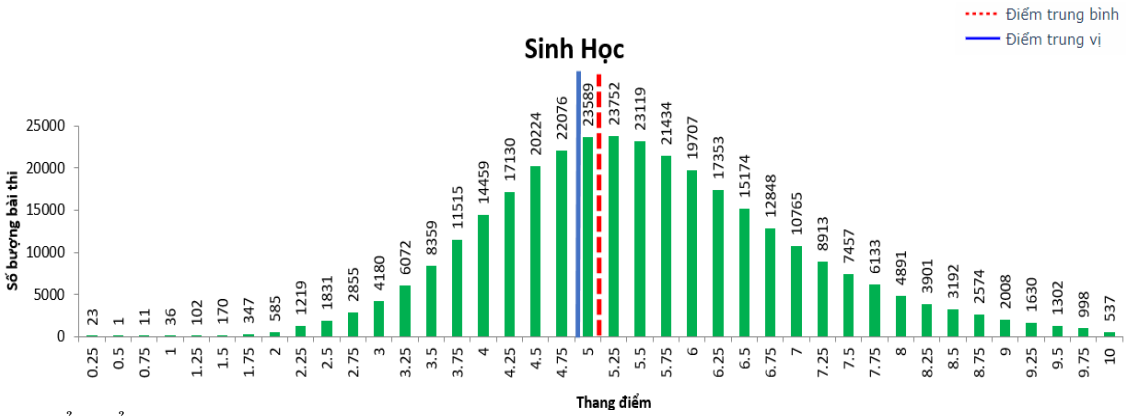
- Phân bố điểm của môn thi đầu tiền trong tổ hợp Khoa Học Tự Nhiên có sự tương đồng với môn Toán khi hơi lệch về bên phải và có sự phân loại giữa điểm khá và giỏi.
- Với 326085 thí sinh tham gia thi môn Vật Lý
- Điểm trung bình là 6.58 điểm
- Điểm trung vị là 6.75 điểm
- Điểm số có nhiều thí sinh đạt nhất là 7.5 điểm
- Phương sai là 1.94
- Độ lệch chuẩn là 1.39
- Số thí sinh có điểm <= 1 là 25 (chiếm tỷ lệ 0.01%)
- Số thí sinh có điểm dưới trung bình(<=5) 42166 (với 12.93%)
- Số thí sinh có điểm >=9 là 4643 (với 1.42%)

5.8. Phổ điểm môn Hóa Học



- Gần như không có sự khác biệt về phân bố giữa điểm Vật Lý và Hóa học. Điều này có thể lý giải vì môn Toán và các môn tự nhiên còn lại là các môn mang tính tư duy, tính toán dẫn đến việc các bạn học tốt các môn Tự Nhiên thì sẽ học tốt Môn Toán và ngược lại.
- Với 327670 thí sinh tham gia thi môn Hóa Học
- Điểm trung bình là 6.63 điểm
- Điểm trung vị là 7.0 điểm
- Điểm số có nhiều thí sinh đạt nhất là 7.75 điểm
- Phương sai là 2.55
- Độ lệch chuẩn là 1.6
- Số thí sinh có điểm (<=1) là 57 (chiếm tỷ lệ 0.02%)
- Số thí sinh có điểm dưới trung bình (<=5) 53387 (chiếm tỷ lệ 16.29%)
- Số thí sinh có điểm giỏi (>=9) 10547 (chiếm tỷ lệ 3.22%)

5.9. Phổ điểm môn Sinh Học

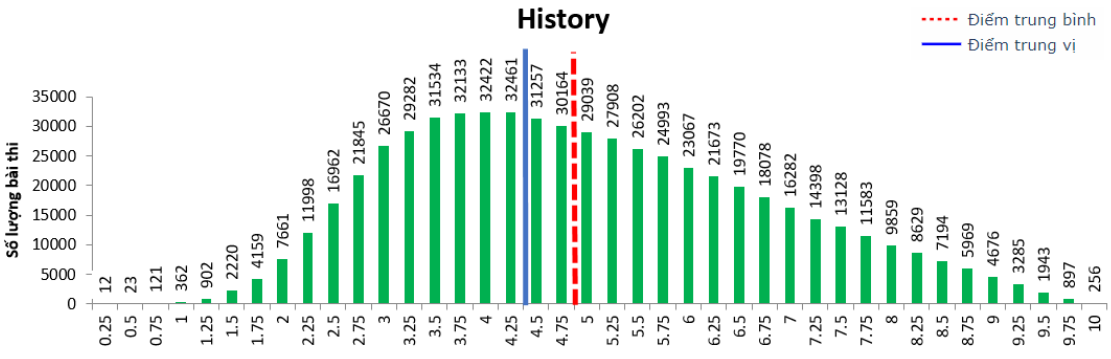


- Phổ điểm của môn Sinh Học thì hơi khác biệt so với hai môn còn lại trong cùng tổ hợp với hình cái chuông úp ngược ngay đúng giữa biểu đồ đúng với mong muốn của nhiều người khi vẽ ra biểu đồ histogram tuy nhiên với tính chất của một kì thi vừa để xét tuyển đại học vừa để tốt nghiệp cấp 3 thì chưa tốt bằng điểm của Vật Lý và Hóa Học do điểm thi này chiếm 70% điểm tốt nghiệp của các bạn học sinh nêu

điểm không quá cao dễ dẫn đến các bạn bị trượt tốt nghiệp. Thay vào đó đề dễ từ mức 7 điểm và có sự phân loại từ 8-10 điểm sẽ phù hợp hơn

- Với 322472 thí sinh tham gia thi môn Sinh Học
- Điểm trung bình là 5.52 điểm
- Điểm trung vị là 5.5 điểm
- Điểm số có nhiều thí sinh đạt nhất là 5.25 điểm
- Phương sai là 2.08
- Độ lệch chuẩn là 1.44
- Số thí sinh có điểm (<=1) là 71 (chiếm tỷ lệ 0.02%)
- Số thí sinh có điểm dưới trung bình (<=5) 111195 (chiếm tỷ lệ 34.48%)
- Số thí sinh có điểm giỏi (>=9) 6475   (chiếm tỷ lệ 2.01%)

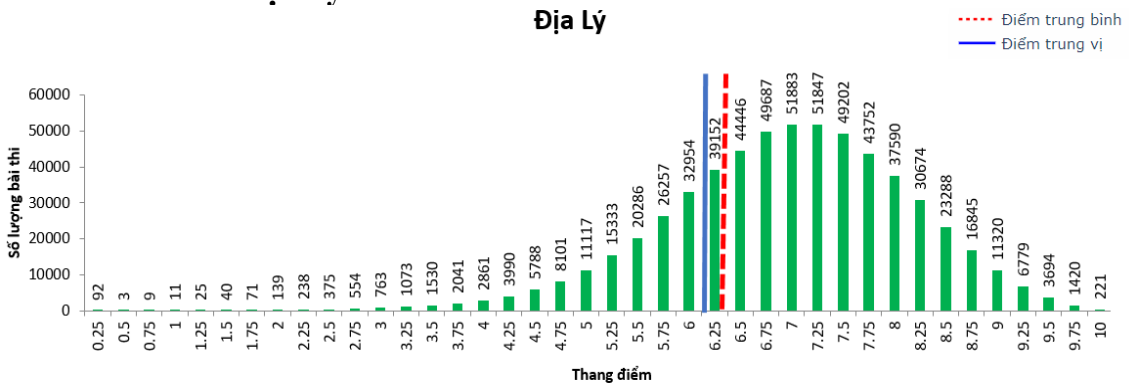
5.10. Phổ điểm môn Lịch Sử



- Lịch sử là môn duy nhất có phổ điểm lệch về bên trái với đỉnh của biểu đồ nằm ở mức 4-4.25 điểm. Là môn duy nhất có điểm trung bình dưới 5 điểm. Ở đây nếu xét cả hai tiêu chí vừa để tốt nghiệp và xét tuyển đại học thì đều không phải là một mô hình chuẩn(một đề thi tốt). Tuy nhiên việc điểm môn Sử thấp đã là vấn đề nan giải của những năm trước đây, chỉ mới năm trước điểm trung bình của môn vượt trên 5 nhưng năm nay đã tuột xuống cùng với sự xuất hiện bất ngờ vì số điểm của môn Tiếng Anh đây sẽ là vấn đề nan giải cho bộ giáo dục cũng như những thầy cô trực tiếp giảng dạy các bạn học sinh trong năm tới.
- Với 601017 thí sinh tham gia thi môn Lịch Sử
- Điểm trung bình là 4.98 điểm
- Điểm trung vị là 4.25 điểm

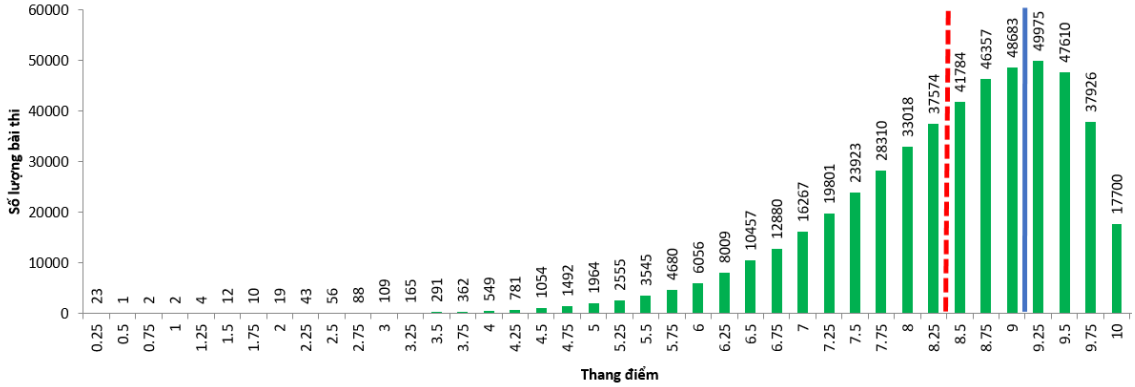
- Điểm số có nhiều thí sinh đạt nhất là 5.25 điểm
- Phương sai là 3.14
- Độ lệch chuẩn là 1.77
- Số thí sinh có điểm (<=1) là 518 (chiếm tỷ lệ 0.09%)
- Số thí sinh có điểm dưới trung bình (<=5) 312188 (chiếm tỷ lệ 51.94%)
- Số thí sinh có điểm giỏi (>=9) 11057 (chiếm tỷ lệ 1.84%)

5.11. Phổ điểm môn Địa Lý



- Trong 3 môn thuộc tổ hợp Xã Hội thì Địa Lý là môn có phổ điểm tốt nhất khi đạt đỉnh ở mức 7-8 điểm và giảm mạnh khi số điểm tăng cao. Chứng tỏ đây là một bài thi tốt cho việc xét tốt nghiệp cũng như chọn lọc ra những sinh viên tương lai ưu tú cho các trường đại học.
- Với 595451 thí sinh tham gia thi môn Địa Lý
- Điểm trung bình là 6.96 điểm
- Điểm trung vị là 7.0 điểm
- Điểm số có nhiều thí sinh đạt nhất là 7.0 điểm
- Phương sai là 1.39
- Độ lệch chuẩn là 1.18
- Số thí sinh có điểm (<=1) là 115 (chiếm tỷ lệ 0.02%)
- Số thí sinh có điểm dưới trung bình (<=5) 27704 (chiếm tỷ lệ 4.65%)
- Số thí sinh có điểm giỏi (>=9) 23434 (chiếm tỷ lệ 3.94%)

5.12. Phổ điểm môn Giáo Dục Công Dân

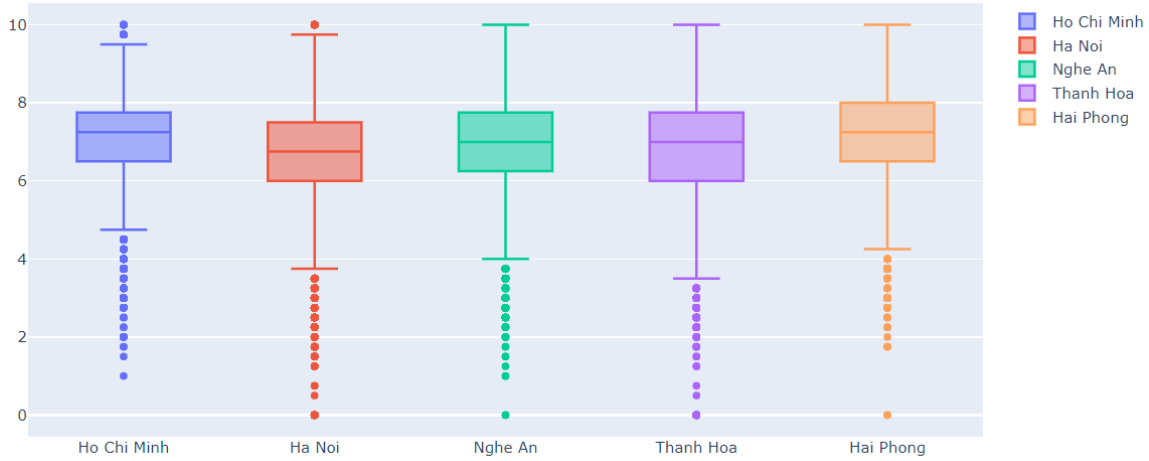


- Từ trước đến giờ, Giáo Dục Công Dân được biết đến là một môn dễ lấy điểm cao, tuy nhiên ở kì thi năm 2021 số điểm này quá dễ để đạt điểm

cực cao khi phổ điểm lệch quá kinh khủng về phía bên phải cụ thể là ở điểm 9-10 tăng nhiều hơn rất nhiều so với những năm trước đây.
    - Chứng tỏ nếu đề xét tốt nghiệp thì đây được xem như là vị cứu tinh cho các bạn lỡ làm bài điểm thấp các môn còn lại tuy nhiên để xét tuyển đại học thì phổ điểm này thật sự khó mà chấp nhận được.
- Với 504137 thí sinh tham gia thi môn Giáo Dục Công Dân
- Điểm trung bình là 8.38 điểm
- Điểm trung vị là 8.5 điểm
- Điểm số có nhiều thí sinh đạt nhất là 9.25 điểm
- Phương sai là 1.34
- Độ lệch chuẩn là 1.16
- Số thí sinh có điểm (<=1) là 28 (chiếm tỷ lệ 0.01%)
- Số thí sinh có điểm dưới trung bình (<=5) 5063 (chiếm tỷ lệ 1.0%)
- Số thí sinh có điểm giỏi (>=9) 201894 (chiếm tỷ lệ 40.05%)

6. Trực quan và phân tích dữ liệu

6.1 Thống kê về điểm môn địa của 5 tỉnh lớn nhất



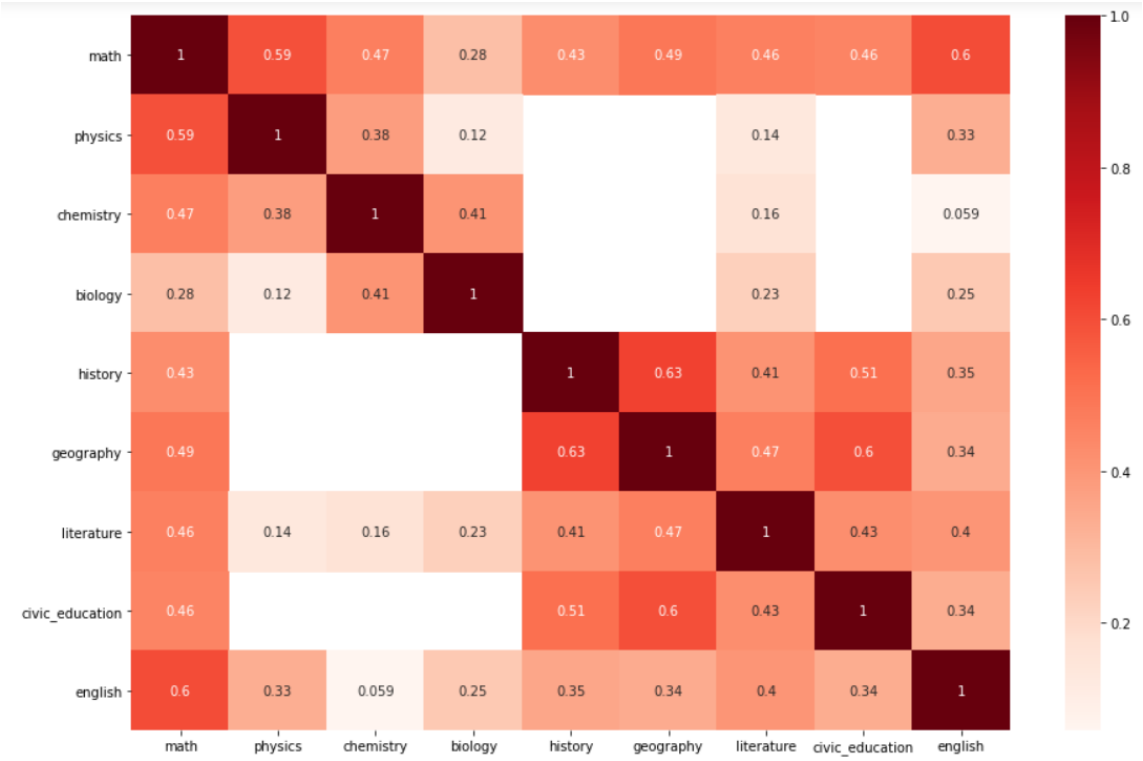
- Dựa theo biểu đồ boxplot, ta thấy được mức chênh lệch về tỉ lệ điểm môn địa của 5 thành phố lớn là Hồ Chí Minh.

- Nhìn chung các thành phố có số điểm điểm phân bố khá giống nhau, điểm chủ yếu phân phối ở mức từ 6-8 điểm trong đó thành phố Hồ Chí Minh là khu vực duy nhất không có thí sinh đạt 0 điểm môn địa, ngoài ra Hồ Chí Minh cũng là thành phố có mức điểm trung bình của học sinh cao nhất trong 5 thành phố cùng với thành phố Hải Phòng với 7.25 điểm, lower fence của thành phố Hồ Chí Minh cũng ở mức cao hơn sao với các thành phố còn lại.

- Về thành phố Hà Nội có điểm địa phân bố nhìn chung thấp hơn so với các thành phố còn lại, có thể nói điểm mạnh của học sinh Hà Nội không phải là môn địa.

- Có một điểm đáng chú ý ở đây là mặc dù ở hai thành phố lớn nhất Việt Nam là Hà Nội và thành phố Hồ Chí Minh nhưng ta có thể thấy số học sinh đạt điểm 9,10 khá thấp so với mức điểm phổ biến khi hai mức điểm 9,10 là những mức điểm outlier ở hai thành phố này, tuy nhiên mức điểm này so với 3 thành phố còn lại thì không thuộc các điểm outlier, có nghĩa là các học sinh đạt điểm 9,10 ở 3 thành phố Nghệ An, Thanh Hóa, Hải Phòng không hề ít nếu so với mức điểm phổ biến ở 3 thành phố này, điều này chứng minh có khá nhiều học sinh học khá giỏi môn địa ở 3 khu vực này.

6.2 Mối tương quan giữa các môn học



- Mối tương quan giữa các môn thi, càng gần 1 thì chỉ số tương quan giữa các môn càng mạnh và ngược lại, nếu môn này có mối tương quan mạnh với môn kia thì có thể học sinh khi giỏi môn này có tỉ lệ cao cũng sẽ giỏi môn kia .

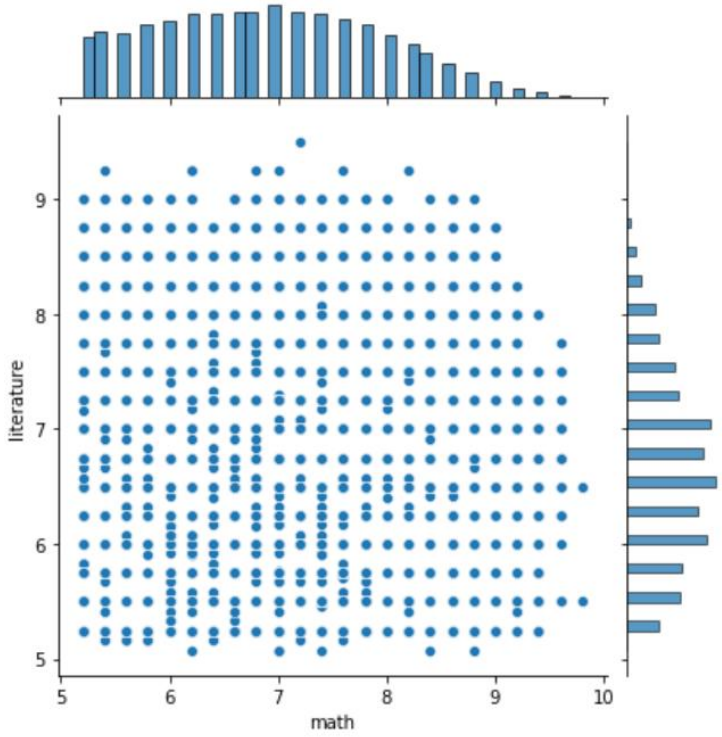
- Ta thấy môn toán có mối tương quan cao nhất là với lý,tiếng anh và môn địa, các môn còn lại cũng có mối tương quan khá cao là lớn hơn 0.4 , trừ môn sinh là có mối liên hệ thấp nhất với môn toán khi mối tương quan giữa hai môn này chỉ có 0.28, có nghĩa là học sinh đạt điểm cao môn toán thường sẽ không đạt điểm cao trong môn sinh. Đáng chú ý là tiếng anh và môn địa là những môn thuộc xã hội nhưng có hệ số

tương quan cao với môn toán thuộc khoa tự nhiên, điều này có thể là do dù ở bất kì khoa nào nền giáo dục của người Á Đông nói chung và Việt Nam nói riêng cũng đề cao môn toán và học sinh cần phải nắm vững được môn này.

- Ngoài ra không có sự tương quan nào giữa các môn học lớn hơn 0.7, lớn nhất là lịch sử với địa 0.63, theo sau là toán và anh 0.6.

6.3 Các thí sinh có điểm dưới trung bình môn Tiếng Anh trong khi 2 môn Toán và Văn đều trên trung bình là vì sao? Liệu rằng việc học ngoại ngữ đang được xem nhẹ dẫn đến tình trạng "học lệch"?

- Mục tiêu của câu hỏi này là trực quan và đưa ra nhận xét cũng như giải pháp.
- Đối tượng sử dụng: mọi người muốn tìm hiểu về điểm thi.
- Ở câu hỏi này, ta sử dụng biểu đồ jointplot kết hợp giữa scatter và histogram để xem phân bố điểm Toán và Văn của các thí sinh có điểm trên 5 nhưng lại dưới trung bình môn Tiếng Anh.
- Ta sử dụng biểu đồ này là vì có thể dễ dàng phân tích được rằng một bộ phận thí sinh có đang học lệch hay là họ đang xem nhẹ việc học ngoại ngữ.
- Dữ liệu được trực quan bao gồm 2 cột math và literature được lọc ra ở 2 thành phố lớn là Hà Nội, Hồ chí Minh. Với biến math được ánh xạ vào trục x và biến literature được ánh xạ vào trục y.
- Các bước thực hiện:
    - Bước 1: Lọc ra dữ liệu cần cho việc trực quan với điều kiện: math > 5, literature > 5, english < 5 và các thí sinh phải thuộc 2 thành phố đã nêu.
    - Bước 2: vẽ biểu đồ trực quan cho 2 cột dữ liệu.
    - Bước 3: Nhận xét về biểu đồ và rút ra kết luận.



Nhận xét:

- Tuy là ở 2 thành phố lớn là Hồ Chí Minh và Hà Nội, nơi có điều kiện học Tiếng Anh rất thuận lợi nhưng lại có đến hơn 26700 thí sinh dưới trung bình môn Tiếng Anh.
- Dựa vào phổ điểm phân hóa ở 2 môn Toán và Văn thì có thể thấy điểm Toán và Văn của các thí sinh này chủ yếu trong mức 6 đến 7. Thậm chí còn có khá nhiều thí sinh còn có điểm Toán và văn ở mức giỏi từ 8 đến 10.
- Việc các thí sinh có điểm ở mức 6 và 7 thì có thể lý giải được là do các thí sinh có học lực ở mức trung bình và khá nên có thể xem nhẹ việc học Tiếng Anh và một phần 2 môn Toán và Văn cũng có thể gọi là 2 môn học nền tảng nên các thí sinh có điểm trên trung bình.
- Nhưng ở chiều hướng ngược lại với các thí sinh có điểm Toán hoặc Văn hay cả 2 môn đều trên 8 nhưng lại xuất hiện trong danh sách các thí sinh dưới trung bình môn Tiếng Anh có thể là do tình trạng xem nhẹ việc học ngoại ngữ, chỉ chú tâm học 2 môn chính là Toán với Văn dẫn đến hậu quả là khi bước chân vào môi trường đại học thì sẽ cảm thấy không tự tin với các môn học cần kiến thức nền tảng về Tiếng Anh, cũng như trong tương lai các bạn sẽ cần Tiếng Anh để giao tiếp trong môi trường làm việc.

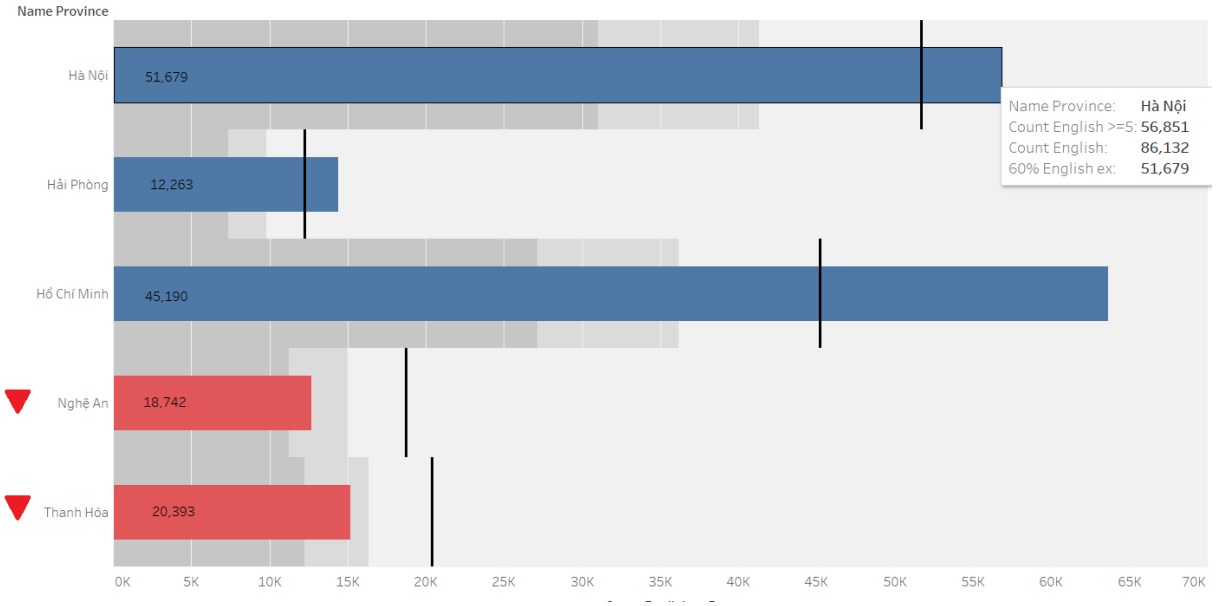
Đề xuất giải pháp:

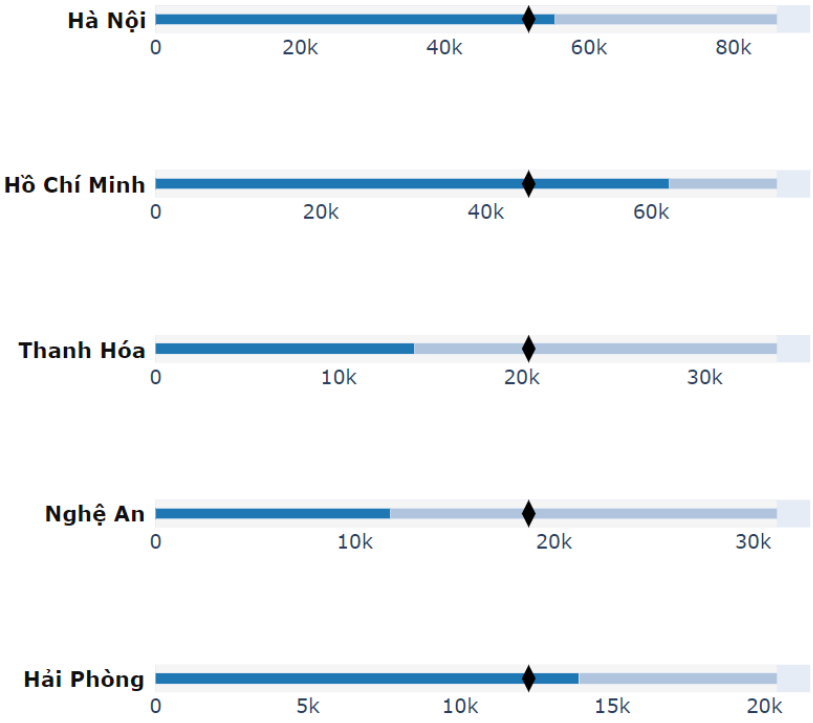
- Các giáo viên cần có trao đổi với nhau để có thể sớm nhận biết các học sinh đang học lệch nhằm chấn chỉnh sớm.
- Cần có các buổi tuyên truyền về lợi ích của việc học ngoại ngữ với học sinh.

- Trường học cần có các câu lạc bộ, hội nhóm trao đổi ngoại ngữ để tạo hứng thú cũng như tạo ra sân chơi cho các học sinh có cơ hội học hỏi thêm về ngoại ngữ.

6.4 Số lượng thí sinh có điểm Tiếng Anh trên trung bình ở 5 tỉnh thành có số lượng thí sinh thi thpt quốc gia nhiều nhất là bao nhiêu? Liệu có sự chênh lệch lớn giữa các tỉnh thành này hay không?

- Mục tiêu của câu hỏi này là trực quan, phân tích điểm thi Tiếng Anh giữa các tỉnh thành.
- Đối tượng sử dụng: mọi người muốn tìm hiểu về điểm thi.
- Câu hỏi này sẽ giúp chúng ta so sánh được tổng thể số lượng thí sinh có điểm Tiếng Anh trên trung bình ở 5 thành phố lớn. Đồng thời ta cũng đặt ra một tiêu chí tối thiểu về số lượng thí sinh cần đạt điểm trên trung bình để xem các tỉnh thành này đã vượt qua con số đó chưa.
- Ta sẽ sử dụng biểu đồ bullet chart để thể hiện số lượng thí sinh trên trung bình so tới tổng thí sinh của tỉnh thành đó, biểu đồ này cũng có thể trực quan được một reference line để thể hiện mục tiêu cần đạt của các tỉnh thành. Qua đó, người dùng có thể dễ dàng quan sát các giá trị mà ta muốn nói đến.
- Dữ liệu mà chúng ta dùng sẽ gồm 2 cột là english và name_province. Các điểm dữ liệu sẽ được lọc ra theo tiêu chí english > 5 và theo các tỉnh thành. Sau đó mỗi bullet chart sẽ tương ứng với 1 tỉnh thành với số lượng thí sinh trên trung bình và tổng số thí sinh của tỉnh thành đó.
- Các bước thực hiện:
  + Bước 1: Lọc ra các cột cần thiết và đưa chúng về dạng chung trong một list để vẽ biểu đồ.
  + Bước 2: Vẽ biểu đồ bullet chart.
  + Bước 3: Nhận xét và phân tích biểu đồ.





Nhận xét:

- Tuy rằng đây là 5 thành phố có số lượng thi nhiều nhất nhưng chỉ có thành phố Hồ Chí Minh, Hà Nội, Hải Phòng là vượt qua mốc 60% còn 2 thành phố còn lại thì cách mốc 60% khá xa.
- Điều đáng nói ở đây là với một thành phố lớn như Hà Nội mà số thí sinh trên trung bình môn Tiếng Anh chỉ có thể vượt qua mốc 60% tầm 4000 thí

sinh. Đây là thành phố lớn và có điều kiện học ngoại ngữ rất tốt nhưng kết quả thì lại không được như mong đợi.
- Thành phố Hải Phòng cũng chỉ vượt qua mốc 60% được khoảng 1000 thí sinh tuy nhiên với chỉ tổng số 20000 thí sinh thì đây cũng là một con số chấp nhận được.
- Qua đó, có thể dễ dàng nhận thấy được chất lượng dạy và học ngoại ngữ vẫn còn ở mức kém và chưa hiệu quả đặc biệt là ở Thanh Hóa và Nghệ An. Việc học ngoại ngữ trong những năm gần đây đã được đề cao tầm quan trọng tuy nhiên thì có vẻ như học sinh vẫn chưa "mặn mà" lắm với việc biết thêm một ngoại ngữ khác.

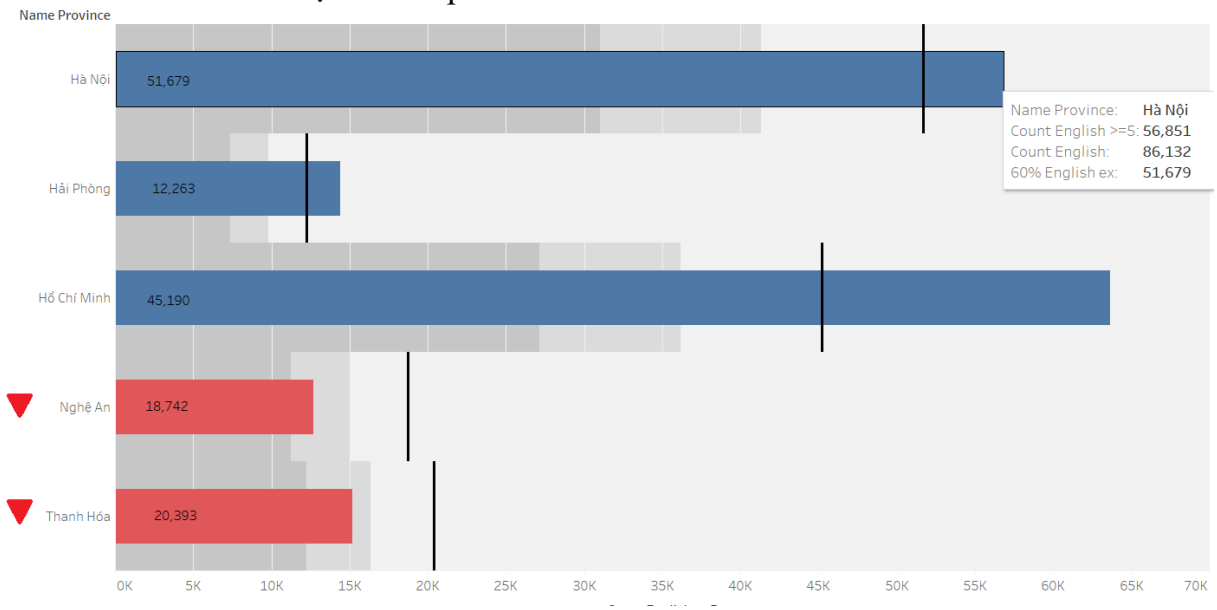
Đề xuất giải pháp:

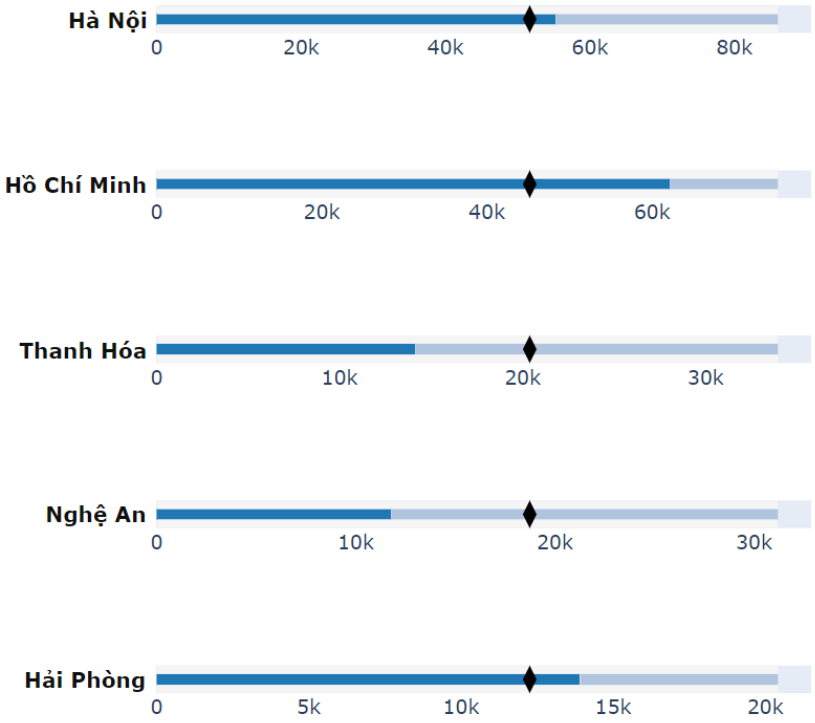
- Các tỉnh thành cần có phương án nâng cao chất lượng dạy và học môn Tiếng Anh như: bồi dưỡng lực lượng giáo viên, đổi mới phương pháp dạy và học, tạo ra các môi trường trao đổi để học sinh tiếp cận Tiếng Anh dễ dàng hơn, khuyến khích việc dạy và học Tiếng Anh,...
- Đặt ra tiêu chí cụ thể cho thành phố để từ đó nâng cao tỉ lệ thí sinh có điểm thi Tiếng Anh trên trung bình qua các năm.
- Thường xuyên lấy ý kiến học sinh về tình hình học Tiếng Anh cũng như khó khăn trong quá trình học để kịp thời hỗ trợ.

6.4 Số lượng thí sinh có điểm Tiếng Anh trên trung bình ở 5 tỉnh thành có số lượng thí sinh thi thpt quốc gia nhiều nhất là bao nhiêu? Liệu có sự chênh lệch lớn giữa các tỉnh thành này hay không?

- Mục tiêu của câu hỏi này là trực quan, phân tích điểm thi Tiếng Anh giữa các tỉnh thành.
- Đối tượng sử dụng: mọi người muốn tìm hiểu về điểm thi.
- Câu hỏi này sẽ giúp chúng ta so sánh được tổng thể số lượng thí sinh có điểm Tiếng Anh trên trung bình ở 5 thành phố lớn. Đồng thời ta cũng đặt ra một tiêu chí tối thiểu về số lượng thí sinh cần đạt điểm trên trung bình để xem các tỉnh thành này đã vượt qua con số đó chưa.
- Ta sẽ sử dụng biểu đồ bullet chart để thể hiện số lượng thí sinh trên trung bình so tới tổng thí sinh của tỉnh thành đó, biểu đồ này cũng có thể trực quan được một reference line để thể hiện mục tiêu cần đạt của các tỉnh thành. Qua đó, người dùng có thể dễ dàng quan sát các giá trị mà ta muốn nói đến.
- Dữ liệu mà chúng ta dùng sẽ gồm 2 cột là english và name_province. Các điểm dữ liệu sẽ được lọc ra theo tiêu chí english > 5 và theo các tỉnh thành. Sau đó mỗi bullet chart sẽ tương ứng với 1 tỉnh thành với số lượng thí sinh trên trung bình và tổng số thí sinh của tỉnh thành đó.
- Các bước thực hiện:
  + Bước 1: Lọc ra các cột cần thiết và đưa chúng về dạng chung trong một list để vẽ biểu đồ.

+ Bước 2: Vẽ biểu đồ bullet chart.
+ Bước 3: Nhận xét và phân tích biểu đồ.





Nhận xét:

- Tuy rằng đây là 5 thành phố có số lượng thi nhiều nhất nhưng chỉ có thành phố Hồ Chí Minh, Hà Nội, Hải Phòng là vượt qua mốc 60% còn 2 thành phố còn lại thì cách mốc 60% khá xa.

- Điều đáng nói ở đây là với một thành phố lớn như Hà Nội mà số thí sinh trên trung bình môn Tiếng Anh chỉ có thể vượt qua mốc 60% tầm 4000 thí sinh. Đây là thành phố lớn và có điều kiện học ngoại ngữ rất tốt nhưng kết quả thì lại không được như mong đợi.
- Thành phố Hải Phòng cũng chỉ vượt qua mốc 60% được khoảng 1000 thí sinh tuy nhiên với chỉ tổng số 20000 thí sinh thì đây cũng là một con số chấp nhận được.
- Qua đó, có thể dễ dàng nhận thấy được chất lượng dạy và học ngoại ngữ vẫn còn ở mức kém và chưa hiệu quả đặc biệt là ở Thanh Hóa và Nghệ An. Việc học ngoại ngữ trong những năm gần đây đã được đề cao tầm quan trọng tuy nhiên thì có vẻ như học sinh vẫn chưa "mặn mà" lắm với việc biết thêm một ngoại ngữ khác.

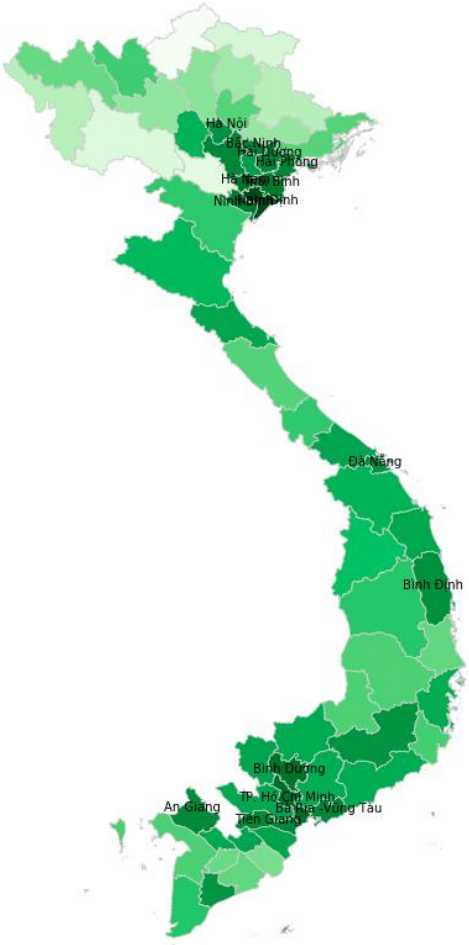
Đề xuất giải pháp:

- Các tỉnh thành cần có phương án nâng cao chất lượng dạy và học môn Tiếng Anh như: bồi dưỡng lực lượng giáo viên, đổi mới phương pháp dạy và học, tạo ra các môi trường trao đổi để học sinh tiếp cận Tiếng Anh dễ dàng hơn, khuyến khích việc dạy và học Tiếng Anh,...
- Đặt ra tiêu chí cụ thể cho thành phố để từ đó nâng cao tỉ lệ thí sinh có điểm thi Tiếng Anh trên trung bình qua các năm.
- Thường xuyên lấy ý kiến học sinh về tình hình học Tiếng Anh cũng như khó khăn trong quá trình học để kịp thời hỗ trợ.

6.5 Có sự bất thường nào về lực học của 3 vùng miền với nhau hay không?

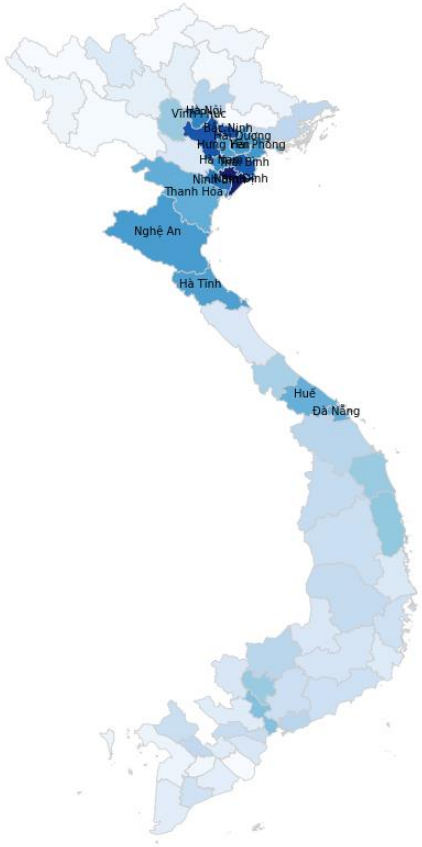
a. Toán:

Biểu đồ thể hiện điểm trung bình môn Toán của từng tỉnh



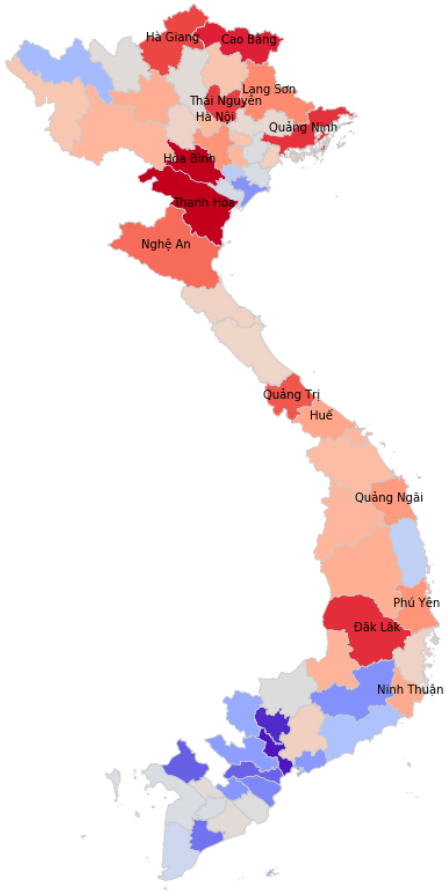
- Ta có thể thấy một xu hướng rất rõ rệt, những tỉnh có điểm trung bình môn toán cao nhất đều tập trung xung quanh 2 trung tâm của đất nước là Hà Nội và TP. Hồ Chí Minh. Miền trung chỉ góp vào 2 đại diện là Đà Nẵng và Bình Định.
  - Từ đây, ta thấy được sự khác biệt về vị địa lí, sự phát triển của kinh tế cũng ảnh hưởng rất lớn đến trình độ học vấn của từng vùng miền.

Biểu đồ thể hiện số điểm môn Toán >9 của từng tỉnh



- Tuy nhiên, miền Nam không phải là nơi có nhiều điểm toán cao nhất trên cả nước, thậm chí là ko thể so bì với miền Trung. Riêng miền Bắc với truyền thống hiếu học được truyền từ đời này sang đời khác, luôn có số lượng điểm toán lớn hơn 9 là rất nhiều và có thể nói đây là khu vực sinh ra nhiều nhân tài nhất.
- Miền trung tuy điểm trung bình toán không cao nhưng luôn có rất nhiều cá nhân có thành tích rất xuất sắc.

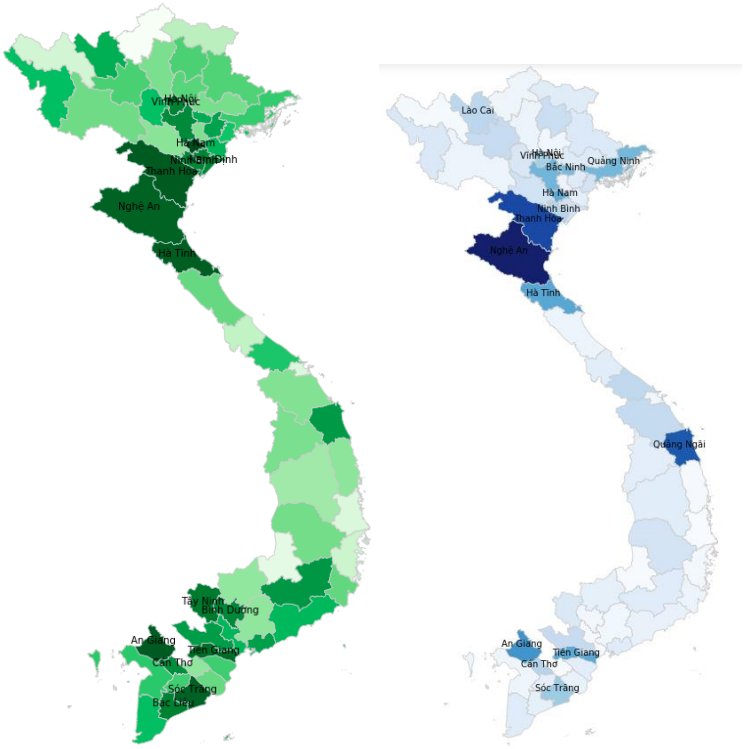
Biểu đồ thể hiện phương sai điểm môn Toán của từng tỉnh



- Tiếp theo, ta có thể thấy được sự chênh lệch điểm của từng tỉnh qua biểu đồ trên.
- Miền Nam tuy không có nhiều điểm cao nhưng về mặt bằng chung thì điểm khá đồng đều, không xuất hiện sự bất bình đẳng.
- Miền Bắc học toán rất tốt nhưng vẫn còn 1 số nơi xuất hiện việc bất bình đẳng giáo dục như Hòa Bình, Thái Nguyên.
- Nam Định thể hiện được sự vượt trội của mình khi vừa có có điểm trung bình cao, số lượng học sinh điểm cao nhiều nhất mà lại không xuất hiện sự bất bình đẳng giáo dục.
- Miền trung xuất hiện sự bất bình đẳng giáo dục khá lớn, đặc biệt là khu vực Thanh Hóa, Nghệ An, Đăk Lăk. Có thể vì sự khắc nghiệt của khí hậu miền Trung, cũng như là nơi có nền kinh tế chưa được phát triển tốt như 2 miền còn lại nên việc phổ cập về giáo dục chưa được hiệu quả.

b. Môn Văn:

Biểu đồ thể hiện điểm trung bình và số điểm lớn hơn 9 của môn Văn trên

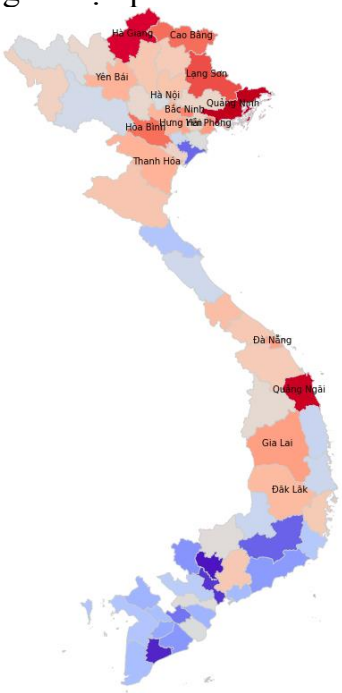


từng tỉnh

Qua 2 biểu đồ trên, ta có thể thấy rằng:

- Các tỉnh miền Nam luôn có điểm trung bình môn văn khá cao, đa số tập trung ở khu vực An Giang, Tiền Giang.
- 3 tỉnh Thanh Hóa, Nghệ An, Hà Tĩnh có sự trỗi dậy rất mạnh mẽ về cả điểm trung bình lẫn số điểm tốt.
- Hà Nam là tỉnh có điểm trung bình văn dẫn đầu cả nước.

- Miền Nam vẫn là nơi có nền giáo dục phát triển rất tốt khi không xuất hiện hiện

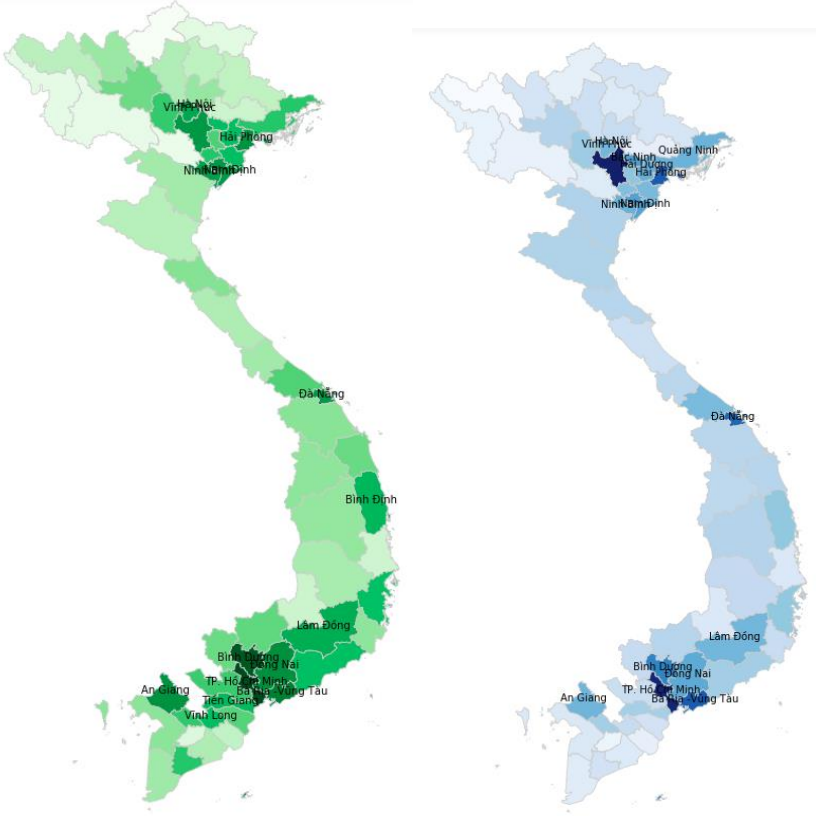


tượng bất bình đẳng giáo dục.

- Sự bất bình đẳng giáo dục thường xuất hiện ở khu vực duyên hải miền Trung và khu vực miền Bắc, đặc biệt là khu vực miền núi Đông Bắc. Ở nơi đây, việc phổ cập giáo dục lên các khu vực miền núi là vô cùng khó khăn vì sự khắc nghiệt do thời tiết giá lạnh, cũng như việc di chuyển bị hạn chế nên có rất nhiều nơi vẫn đang gặp khó khăn rất nhiều.

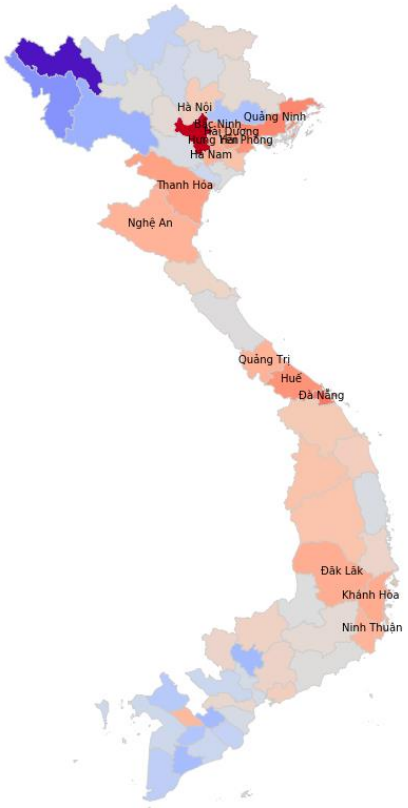
c. Môn Tiếng Anh:

Biểu đồ thể hiện điểm trung bình và điểm lớn hơn 9 của môn Anh trên từng tỉnh



- Một lần nữa, ta thấy được các tỉnh giỏi môn tiếng anh chủ yếu tập trung ở quanh khu vực Hà Nội và TP. Hồ Chí Minh.Đây là những nơi vừa phát triển về kinh tế, vừa phát triển về du lịch và dịch vụ, nên sự quan tâm của phụ huynh cũng như học sinh về việc học môn tiếng anh là rất cao. Số lượng các trung tâm dạy môn tiếng anh cũng như các giáo viên dạy tiếng anh chất lượng làm cho học sinh ở 2 khu vực này có trình độ tiếng anh rất cao.
- Với việc phát triển du lịch và dịch vụ làm cho tiếng anh len lỏi vào cuộc sống hằng ngày của mỗi người dân, trình độ tiếng anh của người dân ở khu vực phía Nam là rất cao.
- Miền Trung luôn là nơi có sự tiếp cận đến tiếng Anh rất kém, một phần do các bậc phụ huynh và học sinh không được tiếp xúc và không hiểu được tầm quan trọng của Tiếng Anh nên điểm môn tiếng Anh thua xa 2 miền còn lại.

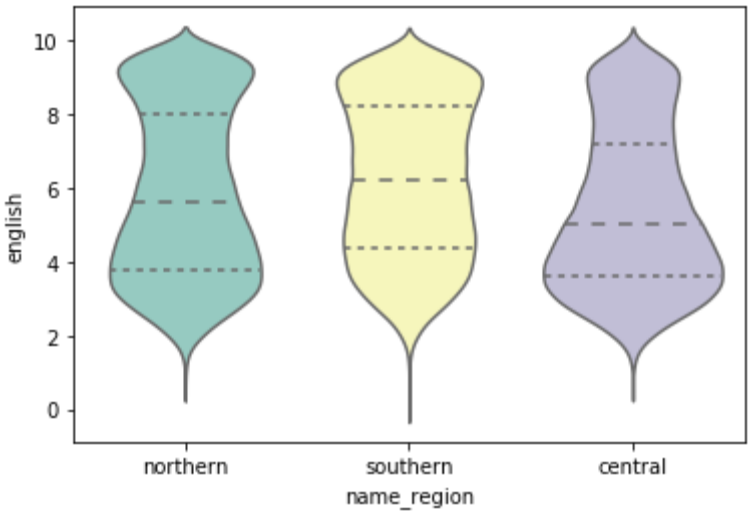
Biểu đồ thể hiện phương sai điểm môn Anh trên từng tỉnh.



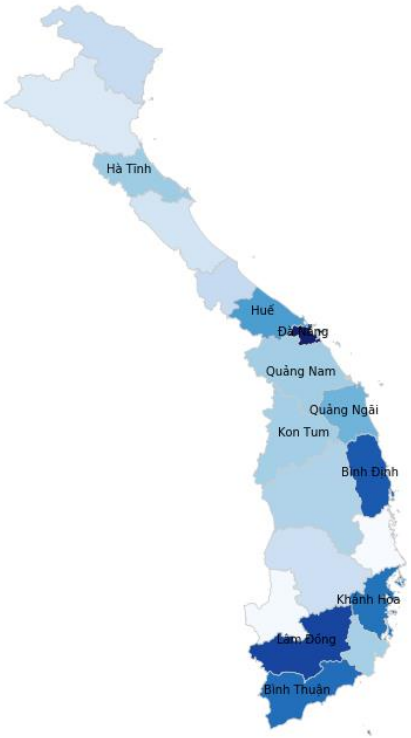
- Miền Nam xứng đáng là một nơi không hề xuất hiện sự bất bình đẳng giáo dục, chưa có một môn học nào có sự chênh lệch điểm số cao. Đây đúng là thành phố phát triển nhất cả nước, mũi nhọn về cả kinh tế và giáo dục.

6.6. Chúng ta đã thấy được sự bất thường về lực học môn Tiếng Anh ở miền Trung, liệu có thể tìm ra được nguyên nhân và đề xuất được giải pháp khắc phục cho vấn đề này không?

Biểu đồ thể hiện sự phân bố điểm môn tiếng Anh của 3 vùng miền

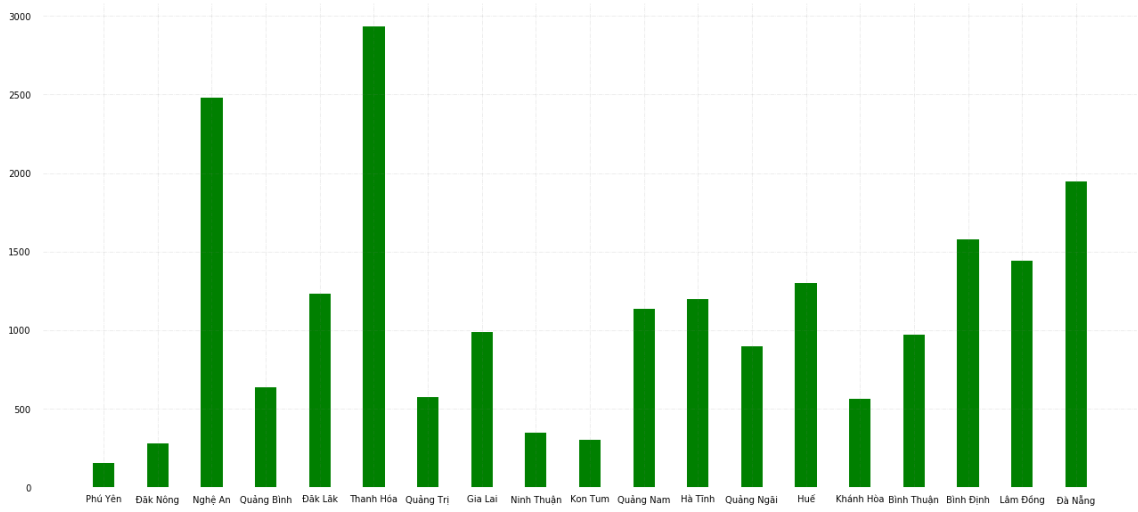


Qua biểu đồ trên, ta thấy được sự khác biệt về phổ điểm tiếng anh của cả 3 vùng miền.
- Miền Nam có 1 phổ điểm rộng phần trên và hẹp dần, phần rộng nhất là phần có điểm 9. Từ đây cho thấy lực học môn tiếng anh ở miền nam rất tốt, số lượng điểm 9 rất nhiều và không có nhiều sự chênh lệch
- Miền Bắc thì lại có hình dạng 2 đầu rộng ra, nhưng phần dưới rộng hơn phần trên. Cho thấy được tuy điểm cao ở miền bắc nhiều nhưng điểm thấp cũng rất nhiều, có lẽ đây là sự khác nhau giữa các trường chuyên và trường không chuyên. Các trường chuyên thường bắt buộc học sinh phải học tốt môn tiếng anh, còn các trường không chuyên thường chỉ đào tạo đủ chứ không nhất thiết học sinh phải tốt môn tiếng anh.
- Miền Trung là nơi có kết quả tệ nhất, hình dạng cây violin với phần dưới rất rộng, hẹp dần và điểm trung bình thấp hơn hẳn so với 2 miền còn lại. Số lượng điểm dưới 5 ở miền trung rất nhiều, thể hiện rõ sự chênh mảng trong việc học tiếng anh của học sinh.

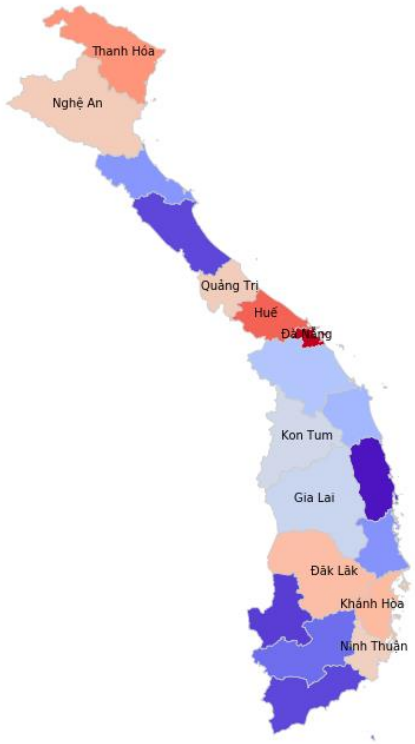


Biểu đồ thể hiện điểm trung bình môn Anh của các tỉnh miền Trung

Biểu đồ thể hiện số lượng điểm cao môn Anh ở miền Trung



Biểu đồ thể hiện phương sai điểm môn Anh các tỉnh miền Trung



- Những tỉnh có điểm trung bình đứng đầu là Đà Nẵng, Bình Định và Lâm Đồng, Bình Thuận. Các tỉnh khác có vẻ có điểm khá là thấp.
- Tuy vậy nhưng Thanh Hóa, Nghệ An lại là 2 tỉnh có số học sinh đạt điểm trên 9 môn tiếng Anh cao nhất trong các tỉnh miền Trung.
- Sự bất bình đẳng giáo dục được thể hiện rõ ở Đà Nẵng và vùng Thanh Hóa Nghệ Tĩnh.

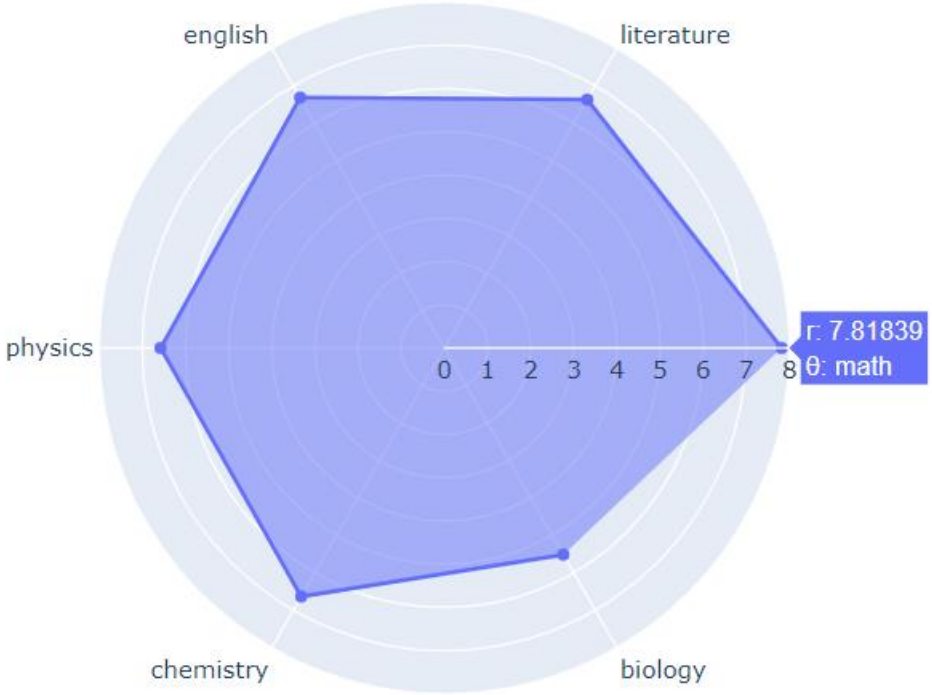
Vậy **nguyên nhân** khiến cho điểm môn tiếng Anh của các học sinh miền Trung thấp là:
- Một phần lớn các thí sinh thi là người miền núi, nông thôn hoặc gia đình không đủ điều kiện để cho học sinh có thể học ngoại ngữ một cách tốt nhất. Mặc dù đa số học sinh đều được học Tiếng Anh từ bậc THCS nhưng đa số đều bị mất gốc dẫn đến lên tới THPT không còn thời gian để đầu tư vào môn tiếng Anh.
- Bởi vì khí hậu khắc nghiệt và kinh tế không phát triển, dẫn đến các dịch vụ du lịch cũng không phát triển theo, vì vậy nên người dân ở miền Trung thường ít được tiếp xúc với tiếng Anh. Những tỉnh như Đà Nẵng, Bình Định và Khánh Hòa là các thành phố du lịch nên thường các tỉnh này có điểm trung bình cao hơn.
- Chất lượng đào tạo ở các vùng nông thôn chưa thật sự được đảm bảo chất lượng, Các giáo viên đa số được cấp bằng một cách dễ dàng và có rất ít kĩ năng thực hành, chưa đảm bảo được chất lượng giảng dạy.
- Nhà trường cũng như phụ huynh học sinh chưa thực sự quan tâm đến môn tiếng anh, đến bây giờ họ vẫn giữ quan niệm là phải tập trung toán, lý, hóa, sinh, văn. Từ đấy làm cho tâm lý của học sinh vẫn coi tiếng anh như một môn phụ và không thực sự nghiêm túc với môn này.
- Ở các thành phố lớn, đa só học sinh đều được học tiếng anh ở ngoài trường học cụ thể là các trung tâm, các khóa học. Còn ở vùng nông thôn thì thường không đủ điều kiện để theo học một khóa như vậy, nhưng thời gian học 1 tuần 2 tiết tiếng Anh trên trường là chưa thực sự đủ đối với học sinh. Từ đây xuất hiện ra sự khác biệt về trình độ tiếng anh của nông thôn và thành thị

Giải pháp:
- Tăng chất lượng dạy và học Tiếng Anh ở các lớp học bằng cách tăng trình độ của các giáo viên.Các giáo viên Tiếng Anh cần được tập huấn, giao lưu với người nước ngoài và phải tăng cường các kĩ năng, kinh nghiệm trong giảng dạy.
- Thúc đẩy các ngành dịch vụ du lịch để thu hút khách nước ngoài, từ đây tạo cơ hội giao tiếp cho người dân, đưa tiếng anh len lỏi vào cuộc sống hàng ngày.
- Đề xuất các trường trung học tổ chức các kì thi, các câu lạc bộ tiếng anh để các học sinh được giao tiếp tiếng anh một cách tự nhiên, từ đây có thể tạo động lực cho học sinh học tiếng anh một các tự nhiên hơn.
- Đề xuất các trường đại học cần có chuẩn tiếng Anh đầu vào để học sinh THPT nhất thiết phải học tiếng Anh để có thể được đăng ký vào trường, từ đây có thể thúc đẩy học sinh học tiếng anh nhiều hơn.

6.7. Giữa tổ hợp Tự Nhiên và Xã Hội, làm sao để các em học sinh đưa ra sự lựa chọn khôn ngoan?

6.7.1 Điểm mạnh và điểm yếu của các học sinh ban Tự Nhiên

- Lọc ra những học sinh thuộc ban Tự Nhiên
- Tính điểm trung bình từng môn học trên tất cả các thí sinh được chọn
- Sử dụng biểu đồ Radar chart

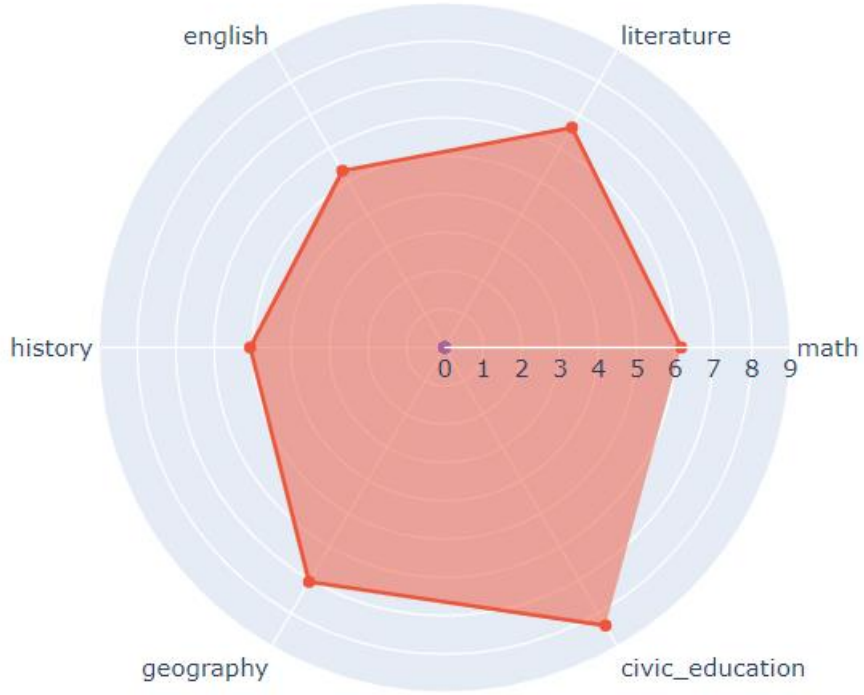


- Không có gì quá bất ngờ khi điểm toán là điểm áp đảo nhất của các thí sinh thuộc ban Tự Nhiên với gần 8.0
- Bên cạnh đó điểm trung bình Tiếng Anh và Ngữ văn của các thí sinh trong ban tương đối ổn với hơn 6.5 điểm
- Điểm trung bình môn Vật lý và Học cũng tương đối cao lần lượt là 6.57 và 6.63
- Điểm đáng chú ý ở đây là môn Sinh Học khi điểm của nó thấp hơn nhiều so với mặt bằng chung các môn học khác mà thí sinh ban Tự Nhiên phải dự thi, với điểm trung bình chỉ vỏn vẹn 5.5 điểm. Như đã nói ở trên thì môn Sinh Học đang là vấn đề trong nền giáo dục của nước ta. Các người đứng đầu trong ngành giáo dục nên đúng ra chịu trách nhiệm với vấn đề của môn học này, cùng như những thầy cô nên thay đổi về phương pháp dạy học hay một hướng khác đó là truyền cảm hứng giúp các bạn có thêm niềm hứng thú với môn Sinh Học. Do thời đại tiếp, công nghệ tế bào sẽ trở thành xu hướng

chúng ta cần có nhiều nhân tài trong lĩnh vực sinh học để nước ta không bị hụt hơi trong cuộc đua với các nước khác trên thế giới.

- Như vậy điểm mạnh của các học sinh ban tự nhiên chính là môn Toán và điểm yếu chính là môn Sinh Học

6.7.2. Điểm mạnh và điểm yếu của các học sinh ban Xã hội

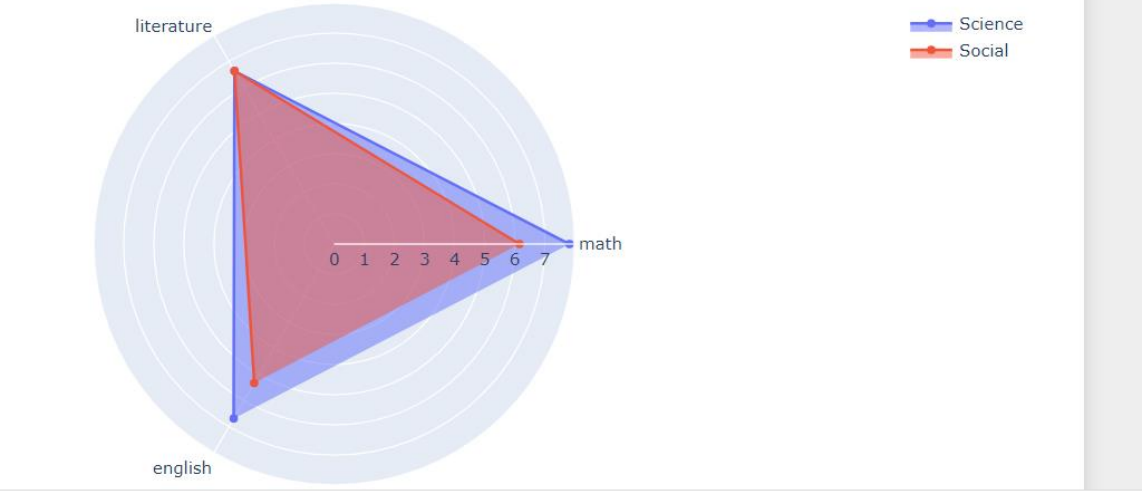


- Khác với tổ hợp Tự Nhiên môn mà các học sinh ban Xã Hội có một môn thi mà luôn luôn các thí sinh sẽ có điểm thi tốt mà không cần đầu tư quá nhiều thời gian đó là Giáo Dục Công Dân.
- Bên cạnh đó Địa Lý cũng là môn có điểm cao thứ 2 trong 6 môn mà các bạn thuộc tổ hợp xã hội phải thi với điểm trung bình là 7.0
- Lịch sử là môn có điểm trung bình quá thấp khi chỉ nó ở mức 5.06 điểm. Đây là vấn đề đã có từ những năm gần đây khi bài thi Lịch Sử thay đổi cách đặt câu hỏi qua đó đòi hỏi thí sinh phải có sự tư duy để lựa chọn đáp án đúng. Nên cần thêm thời gian để các bạn học sinh có thích nghi cũng như qua nhiều đợt thi các thầy cô sẽ tích lũy được kinh nghiệm và giúp các học trò của mình có thể làm được bài tốt hơn.
- Môn Toán là một môn các bạn ban Xã Hội học không quá tốt chỉ với 6.15
- Điểm đáng chú ý đó là điểm thi Ngữ Văn và Tiếng Anh cũng hơi tệ với lần lượt là 6.6 và 5.3 đáng lý ra phải là điểm mạnh của các bạn thuộc ban Xã Hội vì rất nhiều bạn thí sinh dự định xét tuyển theo

khối D01(Toán, Ngữ Văn, Tiếng Anh) thì được định hướng sang học Xã Hội.

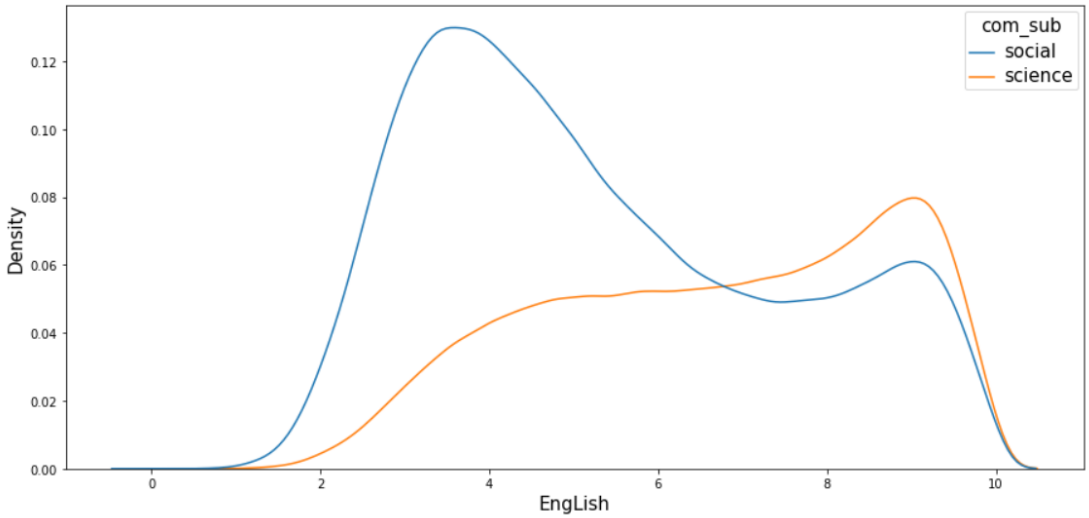
- Tạm hời điểm yếu của các bạn ban Xã Hội có thể nhìn thấy đó là môn Lịch Sử. Tuy nhiên khi so sánh điểm trung bình 3 môn thi bắt buộc là Toán, Ngữ Văn, Anh Văn Sẽ cho ta cái nhìn cụ thể hơn

6.7.3.So sánh thực lực giữa 2 ban tự nhiên và xã hội ở 3 môn học bắt buộc



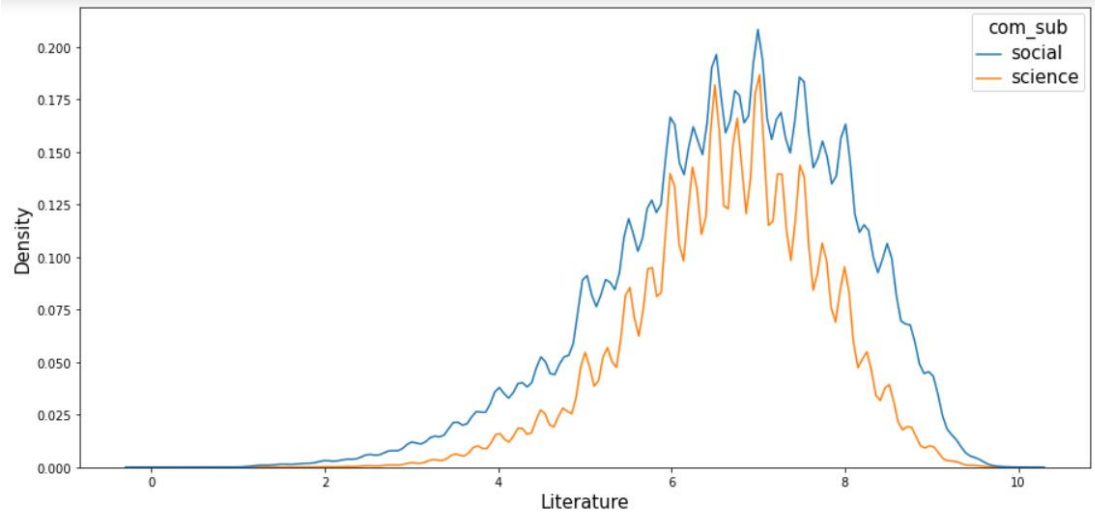
- Nhìn sơ qua ta có thể thấy sự mất cân bằng rất lớn giữa điểm thi của các thí sinh ở hai ban khác nhau
- Với môn Ngữ Văn đây được xem là thế mạnh của các bạn bên Xã Hội tuy nhiên bây giờ nó đã bị bắt kịp bởi các bạn thuộc ban Tự Nhiên
- Ngược lại thì điểm Toán của các thí sinh thuộc ban Tự Nhiên vẫn giữ sự áp đảo tự trước đến giờ
- Cuối cùng là môn Tiếng Anh được dự đoán là có sự cân bằng về trình độ giữa hai ban tuy nhiên điểm của các thí sinh bên Nhiên là 6.68 lớn hơn rất nhiều so với các bạn bên Xã Hội là 5.32

6.7.4 Mật độ phân bố điểm thi Tiếng Anh và Ngữ Văn giữa 2 ban Tự Nhiên và Xã Hội



- Sự mất cân bằng quá rõ ràng về trình độ tiếng Anh của hai tổ hợp môn Tự Nhiên và Xã Hội

- Khi điểm của các thí sinh bên Tự Nhiên thì có mật độ tăng dần theo số điểm và đặt đỉnh ở mức 9 điểm
- Ngược lại số điểm mà nhiều thí sinh thuộc ban xã hội đạt được nhiều nhất là dưới 4 điểm, nhiều gấp 3 lần so với các thí sinh thuộc ban Tự Nhiên, mật độ thí sinh đạt được điểm cũng giảm dần khi số điểm tăng lên chỉ có tăng nhẹ khi từ 8.6 lên 9 điểm tuy nhiên vẫn thấp hơn so với ban Tự Nhiên.



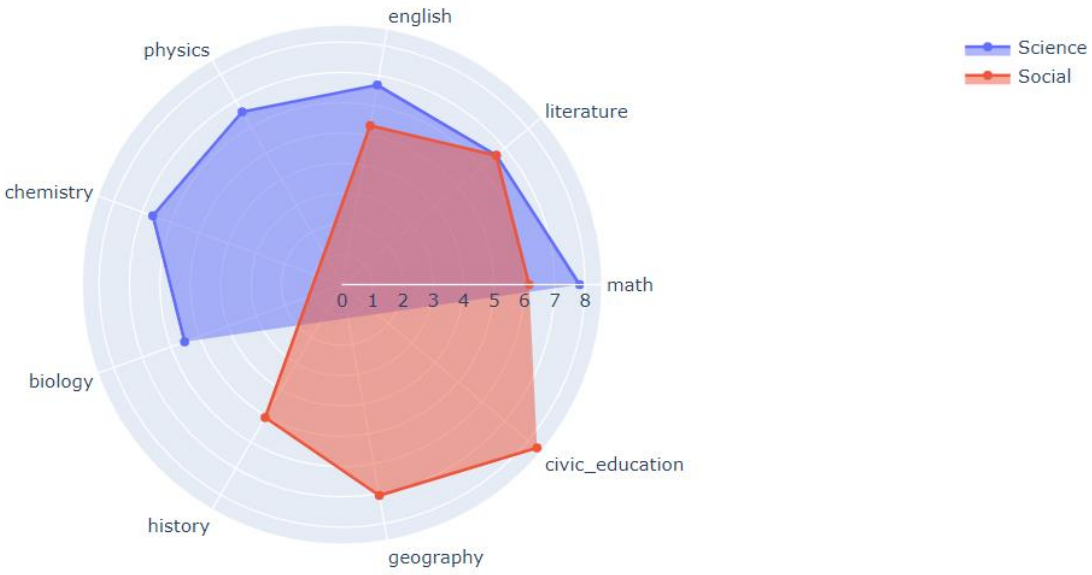
- Một tín hiệu đáng mừng là điểm Ngữ Văn có sự cân bằng giữa các thí sinh và số lượng thí sinh đạt điểm tốt của môn Văn của ban Xã Hội vẫn nhiều hơn kha khá so với ban Tự Nhiên. Như vậy, Ngữ Văn vẫn là điểm mạnh của những thí

sinh chuyên về Xã Hội, việc điểm trung bình Ngữ Văn của 2 ban gần bằng nhau có thể giải thích bằng hai lý do:

- Thứ nhất là các thí sinh thuộc Ban Tự Nhiên đã học tốt Môn Ngữ Văn.
- Thứ hai là một lượng lớn thí sinh tham gia thi tổ hợp Xã Hội và họ đạt điểm kém môn Ngữ Văn đã kéo điểm trung bình của môn xuống.

- Câu hỏi đặt ra liệu việc chọn tổ hợp Xã Hội làm cho học sinh ngày càng mất động lực học tập dẫn đến kết quả sa sút so với tổ hợp Tự Nhiên.
  - Các môn thuộc tổ hợp xã hội dễ học hơn và dễ được điểm cao hơn thuận lợi cho các thí sinh có như cầu xét tốt nghiệp
  - Điểm trung bình 3 môn chung Toán, Ngữ Văn, Anh Văn nhìn chung đều thấp hơn so với Tự Nhiên
  - Phải học chung với lượng lớn thí sinh xem nhẹ việc học do họ chỉ chú trọng vào chuyện đậu tốt nghiệp
  - Có thể được nhập học vào 1 trường đại học quá dễ dàng bằng phương thức xét học bạ dẫn đến thí sinh không còn muốn học nữa.
- Em có một số đề xuất cho vấn đề này:
  - Hãy loại bỏ và thay thế một ngành ở đại học bằng những khóa học nghề do có một số ngành trong các trường đại học là không cần thiết ví dụ từ đó việc vào đại học sẽ thử thách hơn đòi hỏi các em học sinh phải chăm chỉ mới có thể có tương lai tốt

6.7.5 Cách đưa ra lựa chọn tổ hợp môn hợp lý cho các bạn học sinh



- Mỗi tổ hợp đều có một môn học khó tiếp thu đó là Sinh Học ở ban Tự Nhiên và Lịch Sử ở ban xã hội
- Tuy nhiên nếu ta "sớm" biết được mục đích của chúng ta đối với kì thi là tốt nghiệp hay để xét tuyển đại học thì vấn đề trên có thể giải quyết được(có thể là từ năm lớp 10):
  - Trường hợp xét tốt nghiệp: rõ ràng rất chi là phù hợp khi ta chọn tổ hợp Xã Hội vì nó có nhiều môn dễ lấy điểm cao và điểm thi THPT chiếm 70% trong điểm tốt nghiệp.
  - Trường hợp xét tuyển đại học: chúng ta nên lựa chọn tổ hợp Tự nhiên vì
    - Các môn thuộc tổ hợp Xã Hội là Lịch Sử, Địa Lý, Công Dân gần như không xét được các ngành về công nghệ, kĩ thuật những ngành sẽ giúp đất nước phát triển mạnh mẽ trong tương lai
    - Chất lượng đào tạo tốt hơn, khi rõ ràng kết quả của các thí sinh thi Tự nhiên ở 3 môn Toán, Ngữ Văn, Tiếng Anh đã có sự tiến bộ rất nhiều trong khi các thí sinh bên Xã Hội ngày càng tệ đi. Đặc biệt các bạn đang có dự định thi khối D01 thì cũng nên học tổ hợp Tự Nhiên thay vì xã hội do các bạn sẽ được học cùng với các bạn giỏi Toán và Tiếng Anh thay vì chỉ có những bạn giỏi Ngữ Văn ở Xã Hội.

7. Mô hình dự đoán điểm thi Tiếng Anh(Anh Văn)

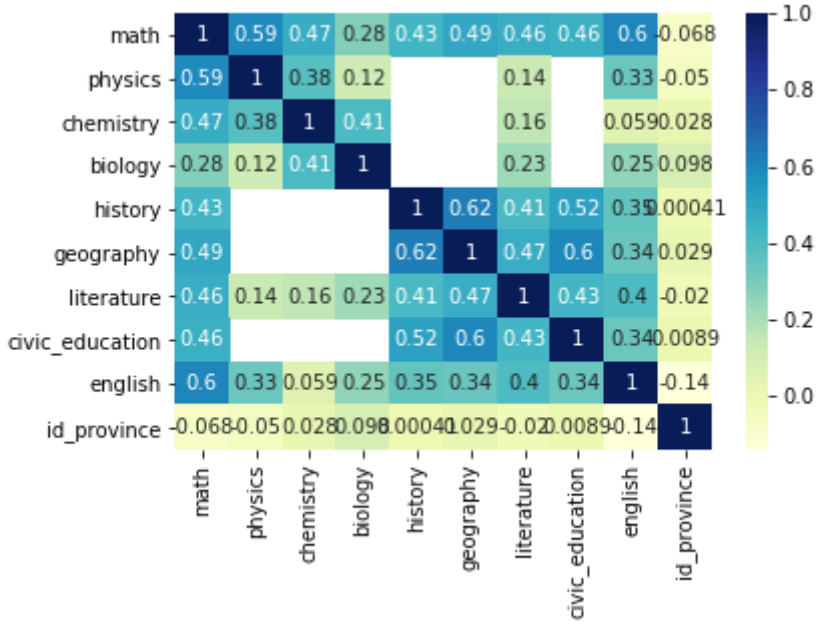
7.1. Loại bỏ những cột không cần thiết:

Ta có thể thấy được những cột như 'id_examinee', 'russian', 'french', 'chinese',' german', 'japanese' đều là những cột không có nhiều ảnh hưởng đến môn toán lắm. Nên ta loại bỏ những cột này ngay từ đầu.

```
data=data.drop(['id_examinee','russian','french','chinese','german','japanese'],axis=1)
```

7.2 Xét sự tương quan của các cột còn lại đối với cột Toán:

- Lý do chọn Toán toán có độ tương quan cao với các môn khác hay nó có thể được xem là biến phụ thuộc vào những môn khác. Chúng ta có thể quan sát trên heat map. Bên cạnh đó tính chất của model linear regression là dự đoán biến phụ thuộc(Toán) bằng những biến độc lập với nhau (điểm những môn học còn lại).



7.3. Mã hóa các biến categorical:

Có 3 cột có type là project nên ta sẽ mã hóa cả 3 cột bằng phương pháp LabelEncoder.

```
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
data['name_region']= label_encoder.fit_transform(data['name_region'])
data['com_sub']=label_encoder.fit_transform(data['com_sub'])
data['name_province']=label_encoder.fit_transform(data['name_province'])
```

Sau đó ta thực hiện điền các giá trị thiếu bằng 0.

7.4 Sử dụng mô hình hồi quy tuyến tính để dự đoán điểm môn Toán:

Kết quả sau khi sử dụng hồi quy tuyến tính để train trên tập dữ liệu.

```
model_lin = sm.OLS.from_formula("math ~ com_sub + english + physics + chemistry +  biology + history + geography
result_lin = model_lin.fit()
result_lin.summary()
```

Dựa vào bảng trên, ta có được các chỉ số khi train mô hình:

:

OLS Regression Results

Dep. Variable:	math	**R-squared:**	0.578
Model:	OLS	**Adj. R-squared:**	0.578
Method:	Least Squares	**F-statistic:**	1.150e+05
Date:	Thu, 28 Apr 2022	**Prob (F-statistic):**	0.00
Time:	17:15:06	**Log-Likelihood:**	-1.3951e+06
No. Observations:	924651	**AIC:**	2.790e+06
Df Residuals:	924639	**BIC:**	2.790e+06
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.4917	0.009	396.427	0.000	3.474	3.509
com_sub	-0.7160	0.007	-109.398	0.000	-0.729	-0.703
english	0.2042	0.001	347.164	0.000	0.203	0.205
physics	0.1895	0.001	155.979	0.000	0.187	0.192
chemistry	0.2749	0.001	210.838	0.000	0.272	0.277
biology	-0.0176	0.001	-13.737	0.000	-0.020	-0.015
history	0.1321	0.001	131.169	0.000	0.130	0.134
geography	0.0223	0.001	18.073	0.000	0.020	0.025
literature	0.1392	0.001	155.452	0.000	0.137	0.141
civic_education	0.1689	0.002	107.701	0.000	0.166	0.172
name_province	-0.0028	7.17e-05	-39.518	0.000	-0.003	-0.003
name_region	-0.0324	0.001	-22.011	0.000	-0.035	-0.030

Omnibus:	7810.363	**Durbin-Watson:**	1.764
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	11985.789
Skew:	-0.068	**Prob(JB):**	0.00
Kurtosis:	3.541	**Cond. No.**	324.

- R-squared khá tốt : R-squared=0.578. Đây có ý nghĩa là các thuộc tính khác biểu diễn được 57,8% thuộc tính Math, còn lại là các biến nhiễu và các biến ngoại lệ.
- P>|t| của tất cả các thuộc tính đều bằng 0. điều này có nghĩa là các thuộc tính đều độc lập với thuộc tính Math.

Từ đây ta có thể thấy có thể sử dụng hồi quy tuyến tính để dự đoán điểm toán và ta không cần phải loại bỏ bất kì thuộc tính nào.

- Chúng ta cũng có thể chia tập dữ liệu ra làm tập train và tập test để có thể xem xét mô hình có thực sự tốt hay không.

Ta thấy được rằng sau khia chia tập dữ liệu, train và test thì ta nhận được các chỉ số tương tự như lần train trước.

```
X = data[predictors].values
y = data[target_column].values

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=0)
```

```
lr = LinearRegression()
full_pipeline=Pipeline(steps=[("scal", StandardScaler()),("lr",lr)])
full_pipeline.fit(X_train, y_train)
```

```
]: Pipeline(memory=None,
          steps=[('scal',
                  StandardScaler(copy=True, with_mean=True, with_std=True)),
                 ('lr',
                  LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
                                   normalize=False))],
          verbose=False)
```

Type *Markdown* and LaTeX: α^2

```
pred_train_lr= full_pipeline.predict(X_train)
print(np.sqrt(mean_squared_error(y_train,pred_train_lr)))
print(r2_score(y_train, pred_train_lr))

pred_test_lr= full_pipeline.predict(X_test)
print(np.sqrt(mean_squared_error(y_test,pred_test_lr)))
print(r2_score(y_test, pred_test_lr))
```

```
1.090260395496559
0.5802929872433558
1.090766440714573
0.5805464907168283
```

Chỉ số ở cả tập test và tập train đều bằng nhau.

- Sai số trung bình khoảng 1 điểm
- chỉ số R square cũng khá tốt.

Tóm lại ta có thể sử dụng mô hình để có thể dự báo điểm môn Toán từ các điểm môn khác.