TRỰC QUAN HÓA DỮ LIỆU

Nhóm thực hiện:

19120330 – Nguyễn Đoan Phúc

19120606 – Nguyễn Đình Hoàng Nguyên

19120643 – Đào Thị Thiện Tâm

19120683 – Thái Trung Tín

19120715 – Nguyễn Kha Vĩ

ĐỒ ÁN MÔN HỌC

HỌC KỲ II – NĂM HỌC 2021-2022

THÔNG TIN CÁ NHÂN

Mã nhóm ***14***

MSSV	Họ tên	Email	Điện thoại
19120606	Nguyễn Đình Hoàng Nguyên	19120606@student.hcmus.edu.vn	0918195530
19120683	Thái Trung Tín	19120683@student.hcmus.edu.vn	0337399029
19120643	Đào Thị Thiện Tâm	19120643@student.hcmus.edu.vn	0377753521
19120715	Nguyễn Kha Vĩ	19120715@student.hcmus.edu.vn	0374246292
19120330	Nguyễn Đoan Phúc	19120330@student.hcmus.edu.vn	0945706609

MỤC LỤC

A. PHÂN CÔNG CÔNG VIỆC VÀ ĐÁNH GIÁ MỨC ĐỘ HOÀN THÀNH CỦA CÁC THÀNH VIÊN

Người phụ trách	MSSV	Công việc	Mức độ hoàn thành (%)
Nguyễn Kha Vĩ	19120715	- Tiền xử lý dữ liệu và viết báo cáo liên quan. - Trực quan hóa và viết báo cáo một câu hỏi về dữ liệu : Trực quan số lượng điểm giỏi (>=9) và điểm liệt (<=1) của các môn trên phạm vi cả nước và từng địa phương - Kiểm tra và nộp đồ án.	100%
Nguyễn Đình Hoàng Nguyên	19120606	- Trực quan hóa và viết báo cáo hai câu hỏi về dữ liệu : + Tỉ lệ số lượng điểm 10 ở tất cả các tỉnh thành với cả nước và 10 tỉnh thành có số lượng điểm 10 cao nhất. + So sánh điểm trung bình tất cả các môn giữa TP. Hồ Chí Minh, TP. Hà Nội và cả nước. - Soạn slide thuyết trình - Phân tích dữ liệu	100%

Thái Trung Tín	19120683	- Trực quan hóa và viết báo cáo hai câu hỏi về dữ liệu : + Trực quan sự phân bố và phân tán dữ liệu của các môn tự nhiên. + Điểm trung bình môn Toán với một số môn khác của các tỉnh/thành phố. - Soạn slide thuyết trình	100%
Đào Thị Thiện Tâm	19120643	- Trực quan hóa và viết báo cáo hai câu hỏi về dữ liệu : + Trực quan điểm trung bình các môn trên phạm vi cả nước và từng địa phương + Trực quan phân phối điểm của các môn trên phạm vi cả nước và từng địa phương - Chỉnh sửa báo cáo, slide	100%
Nguyễn Đoan Phúc	19120330	- Tìm các thông tin dữ liệu về đề tài thực hiện - Trực quan hóa và viết báo cáo hai câu hỏi về dữ liệu : + Mối quan hệ tương quan giữa điểm thi các môn học + Môn ngoại ngữ nào (ngoại trừ Tiếng Anh) có nhiều thí sinh tham gia nhất? Môn ngoại ngữ nào có điểm thi trung bình cao nhất? - Quay video trình bày	100%

B. BÁO CÁO ĐỒ ÁN

I. MÔ TẢ THÔNG TIN VỀ DỮ LIỆU VÀ CÁC BƯỚC TIỀN XỬ LÝ

1. Khám phá dữ liệu

1.1 Xem kích thước bộ dữ liệu

```
# Xem kích thước dữ liệu
shape = df.shape
print(f'Dữ liệu có {shape[0]} dòng và {shape[1]} biến')
```

```
Dữ liệu có 870486 dòng và 12 biến
```

- Bộ dữ liệu điểm thi THPT Quốc gia 2020 có 870486 dòng, mỗi dòng tương ứng với một học sinh và 12 biến, mỗi biến ứng với một môn học trong kỳ thi.

1.2 Xem các dòng dữ liệu và kiểu dữ liệu từng cột

```
# Xem dữ liệu ở các dòng đầu tiên
df.head()
```

	Unnamed: 0	Dia	GDCD	Hoa	Li	Ma_mon_ngoai_ngu	Ngoai_ngu	Sinh	Su	Toan	Van	sbd
0	0	7.00	6.50	NaN	NaN	N1	4.2	NaN	4.75	6.4	6.75	18014547
1	1	7.75	7.75	NaN	NaN	N1	2.8	NaN	3.75	7.6	6.00	18014530
2	2	6.50	NaN	NaN	NaN	NaN	NaN	NaN	4.00	4.8	4.75	18014521
3	3	8.00	9.50	NaN	NaN	N1	5.8	NaN	8.25	8.0	7.00	18014517
4	4	NaN	NaN	8.5	8.0	N1	4.0	5.0	NaN	8.2	6.50	18014523

```
# Xem kiểu dữ liệu từng cột
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 870486 entries, 0 to 870485
Data columns (total 12 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Unnamed: 0        870486 non-null  int64
 1   Dia               555072 non-null  float64
 2   GDCD              482980 non-null  float64
 3   Hoa               295536 non-null  float64
 4   Li                293287 non-null  float64
 5   Ma_mon_ngoai_ngu  772098 non-null  object
 6   Ngoai_ngu         772098 non-null  float64
 7   Sinh              290377 non-null  float64
 8   Su                568581 non-null  float64
 9   Toan              866581 non-null  float64
 10  Van               856565 non-null  float64
 11  sbd               870486 non-null  int64
dtypes: float64(9), int64(2), object(1)
memory usage: 79.7+ MB
```

- Ta thấy dữ liệu trên có khá nhiều giá trị "NaN". Tuy nhiên, đối với kì thi THPT Quốc gia, thí sinh có quyền chỉ chọn 1 trong các tổ hợp KHTN và KHXH, đồng thời cũng có các thí sinh được miễn thi môn ngoại ngữ. Nên dẫn đến dữ liệu có nhiều giá trị "NaN" cũng là hợp lý.

- Kiểu dữ liệu cũng đã phù hợp trong quá trình trực quan, phân tích. Tuy nhiên, ta nên thêm biến "Cum_thi" để dễ dàng phân loại thí sinh theo cụm.

2. Làm sạch dữ liệu

2.1 Đếm và xoá các giá trị trùng nhau

```
# Đếm các cặp giá trị trùng nhau
duplicated = df.duplicated().sum()
if duplicated > 0:
  df = df.drop_duplicates(ignore_index = True)
else:
  print("Không có giá trị nào trùng nhau!")
```

```
Không có giá trị nào trùng nhau!
```

- Ở bộ dữ liệu này, không có các cặp dữ liệu nào trùng nhau.

2.2 Xử lý các giá trị "NaN"

- Như đã trình bày ở trên, dữ liệu có nhiều giá trị "NaN", và chúng là hợp lý.

- Ở đây ta sẽ xét một trường hợp cụ thể là môn Ngoại ngữ. Số lượng giá trị "NaN" ở biến "Ngoai_ngu" bằng số lượng giá trị "NaN" ở biến "Ma_mon_ngoai_ngu", từ đó ta có thể suy đoán rằng:

+ Các thí sinh này được miễn thi môn ngoại ngữ.

+ Đồng thời cũng không có thí sinh nào bị mất điểm môn ngoại ngữ.

Điều này càng khẳng định các giá trị "NaN" là hợp lý.

```
NaNNN = pd.isnull(df['Ngoai_ngu'])
NaNMMNN = pd.isnull(df['Ma_mon_ngoai_ngu'])

if sum(NaNNN ^ NaNMMNN) == 0:
  print("Không có thí sinh nào bị mất điểm môn ngoại ngữ")
```

```
Không có thí sinh nào bị mất điểm môn ngoại ngữ
```

2.3 Thêm biến "Cum_thi"

- Ta sẽ thêm tên cụm thi dự vào 2 số đầu số báo danh của mỗi thí sinh để dễ dàng hơn trong quá trình trực quan dữ liệu giữa các tỉnh với nhau.

```
# Các cụm thi và số thứ tự
map = {
    1: "Hà Nội",
    2: "TP. Hồ Chí Minh",
    3: "Hải Phòng",
    4: "Đà Nẵng",
    5: "Hà Giang",
    6: "Cao Bằng",
    7: "Lai Châu",
    8: "Lào Cai",
    9: "Tuyên Quang",
    10: "Lạng Sơn",
    11: "Bắc Kạn",
    12: "Thái Nguyên",13: "Yên Bái",14: "Sơn La",15: "Phú Thọ",16: "Vĩnh P
}
```

```
# Mapping số báo danh với thứ tự cụm thi dự vào 2 số đầu của số báo danh
df["Cum_thi"] = df["sbd"].apply(lambda x: map[int(x/1_000_000)])
```

- Kết quả sau khi gom nhóm theo từng cụm như ảnh bên dưới

```
# Xem số lượng thí sinh ở từng cụm
df.groupby(["Cum_thi"])["sbd"].agg('count')
```

```
Cum_thi
An Giang            15239
Bà Rịa-Vũng Tàu     11441
Bình Dương          11386
Bình Phước           9774
Bình Thuận          10893
                    ...
Điện Biên            5608
Đăk Nông             6212
Đắk Lắk             13945
Đồng Nai            28254
Đồng Tháp           13386
Name: sbd, Length: 62, dtype: int64
```

2.4 Kết quả sau khi xử lý dữ liệu

```
shape = df.shape
print(f'Dữ liệu có {shape[0]} dòng và {shape[1]} biến')
```

```
Dữ liệu có 870486 dòng và 13 biến
```

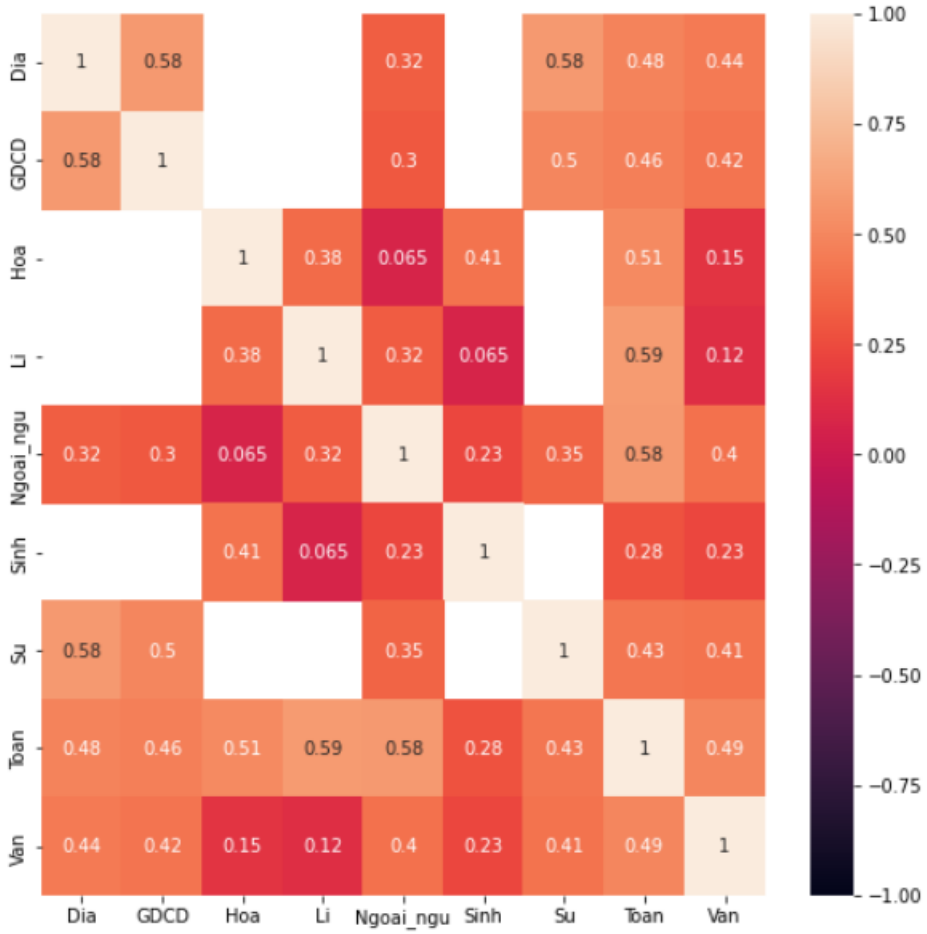
```
df.head(5)
```

	Unnamed: 0	Dia	GDCD	Hoa	Li	Ma_mon_ngoai_ngu	Ngoai_ngu	Sinh	Su	Toan	Van	sbd	Cum_thi
0	0	7.00	6.50	NaN	NaN	N1	4.2	NaN	4.75	6.4	6.75	18014547	Bắc Giang
1	1	7.75	7.75	NaN	NaN	N1	2.8	NaN	3.75	7.6	6.00	18014530	Bắc Giang
2	2	6.50	NaN	NaN	NaN	NaN	NaN	NaN	4.00	4.8	4.75	18014521	Bắc Giang
3	3	8.00	9.50	NaN	NaN	N1	5.8	NaN	8.25	8.0	7.00	18014517	Bắc Giang
4	4	NaN	NaN	8.5	8.0	N1	4.0	5.0	NaN	8.2	6.50	18014523	Bắc Giang

II. TRỰC QUAN HÓA DỮ LIỆU

1. Mối quan hệ tương quan giữa điểm thi các môn học

- Sử dụng biểu đồ heatmap để trực quan câu hỏi này. Hệ số tương quan dao động trong khoảng [-1,1], càng gần 1 thì mối liên hệ giữa 2 biến càng mạnh, càng gần -1 thì mối liên hệ giữa hai biến càng ít, hoặc gần như không có.

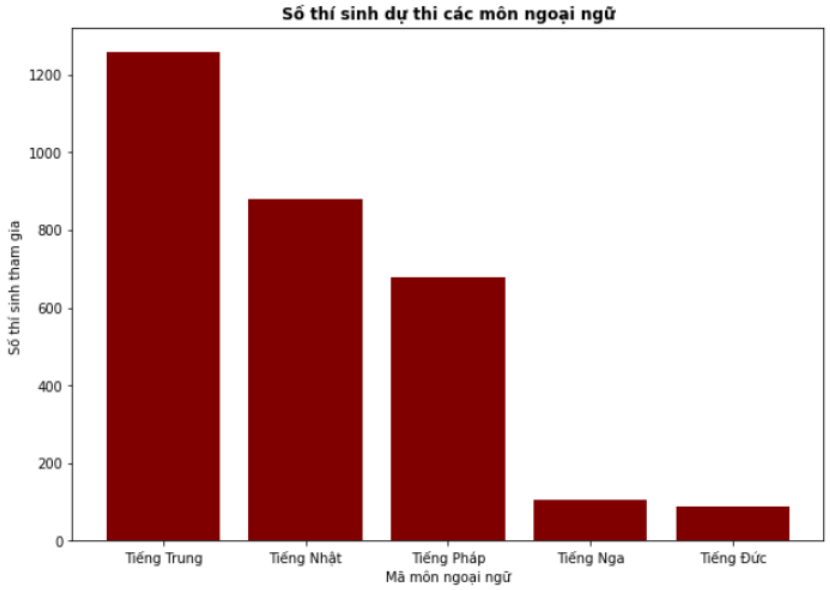


- Nhìn vào biểu đồ, ta thấy các môn xã hội như Địa, GDCD, Sử không có sự tương quan với các môn tự nhiên như Lý, Hóa, Sinh.

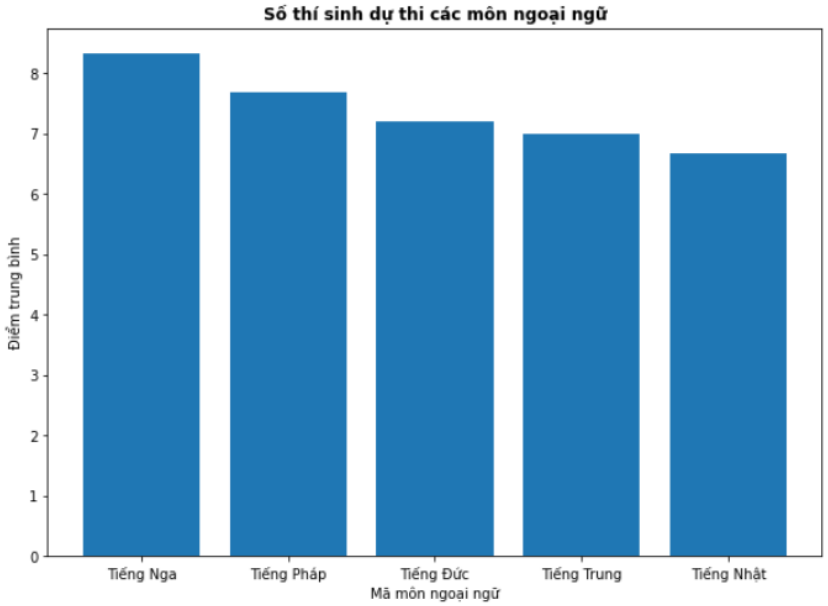
- Môn toán có sự tương quan lớn nhất với môn Lí, Ngoại Ngữ, Văn.

- Hệ số tương quan mạnh nhất thuộc về hai môn Toán - Lí (0.59)

2. **Môn ngoại ngữ nào (ngoại trừ Tiếng Anh) có nhiều thí sinh tham gia nhất? Môn ngoại ngữ nào có điểm thi trung bình cao nhất?**

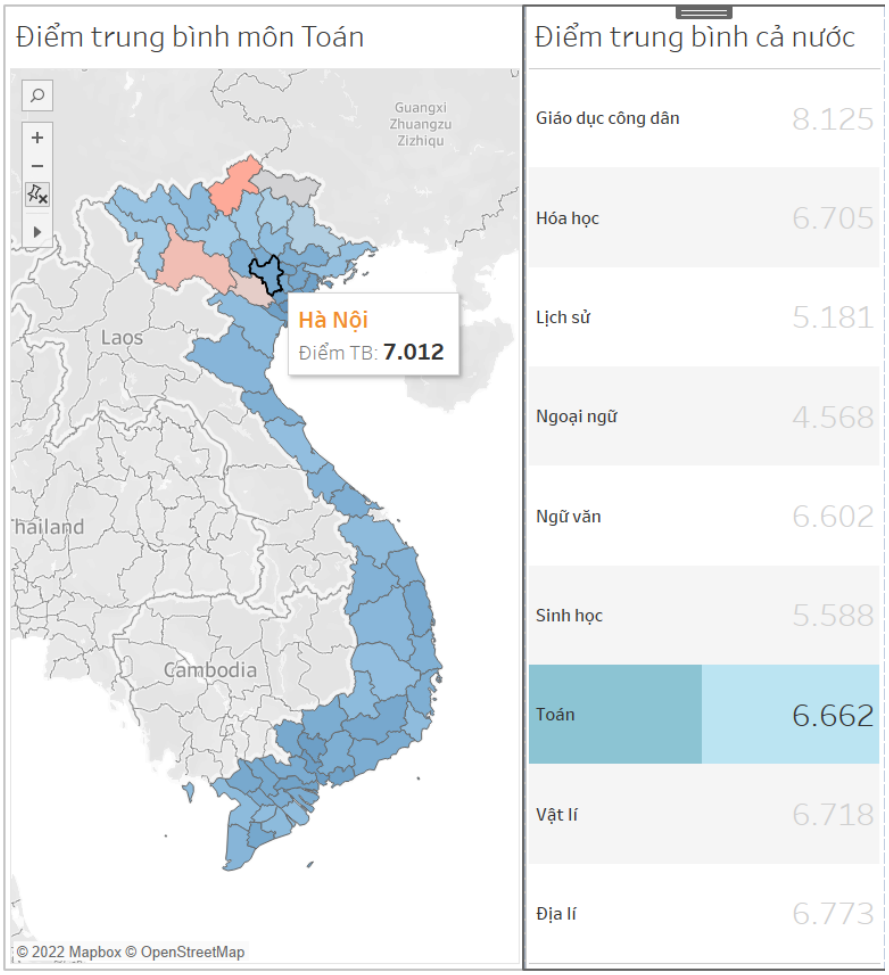


- Vậy môn ngoại ngữ khác (ngoài tiếng Anh) được nhiều thí sinh lựa chọn nhất là tiếng Trung.



- Điểm trung bình cao nhất thuộc về môn Tiếng Nga. Ta nhận thấy một điều thú vị, các môn ngoại ngữ có ít thí sinh tham gia lại có điểm trung bình cao hơn các môn ngoại ngữ phổ biến.

3. Trực quan điểm trung bình các môn trên phạm vi cả nước và từng địa phương



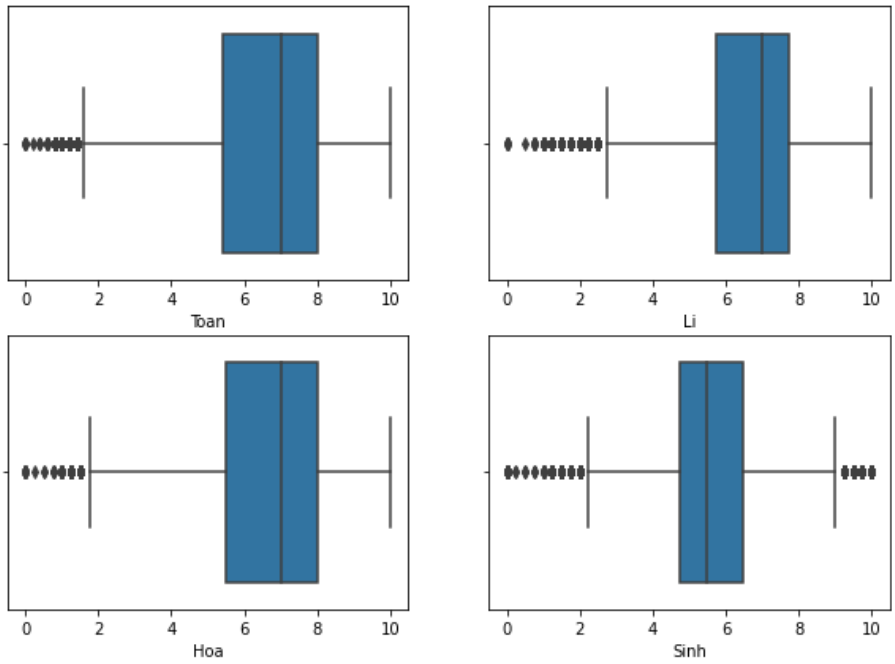
- Biểu đồ cho thấy được điểm trung bình môn được chọn trên phạm vi cả nước và từng khu vực. Ở phạm vi khu vực, sử dụng Maps để trực quan. Sử dụng màu sắc: đỏ đối với điểm dưới 5, xanh đối với điểm trên 5.

- Dễ dàng so sánh chênh lệch học lực giữa các địa phương hoặc so với trung bình cả nước. Nhìn vào biểu đồ trên, dễ dàng nhận thấy đối với môn Toán, điểm trung bình thấp nhất là Hà Giang (4,504), cao nhất là Nam Định (7,633). Điểm trung bình ở Hà Nội (7,012) cao hơn cả nước (6,662).

- Tương tác: Chọn môn ở biểu đồ điểm trung bình cả nước. Khi đó, biểu đồ điểm trung bình môn được chọn trên từng địa phương sẽ được cập nhật tương ứng.

- Link Dashboard: tai đây

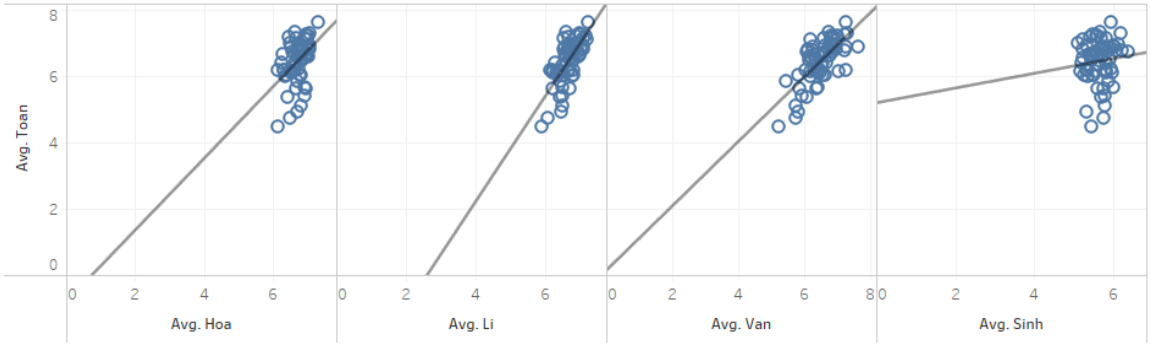
4. Trực quan sự phân bố và phân tán dữ liệu của các môn tự nhiên



- Nhìn vào 4 biểu đồ trên ta thấy các outlier của Toán Lí Hóa ở dưới mức minimum trong khi đó outlier của Sinh lại có đủ các outlier của maximum và minimum. Có thể suy luận ra số lượng điểm 10 năm nay nhiều hơn.

- Các giá trị Q1, Q3, median của Toán, Lí, Hóa có giá trị tương đối giống nhau trong khi đó đối với môn Sinh thì các giá trị trên bị giảm đi. Tuy điểm 10 nhiều hơn nhưng về mặt đa số thì các môn khác vẫn có điểm cao hơn so với môn Sinh.

5. Điểm trung bình môn Toán của các tỉnh/thành với một số môn khác

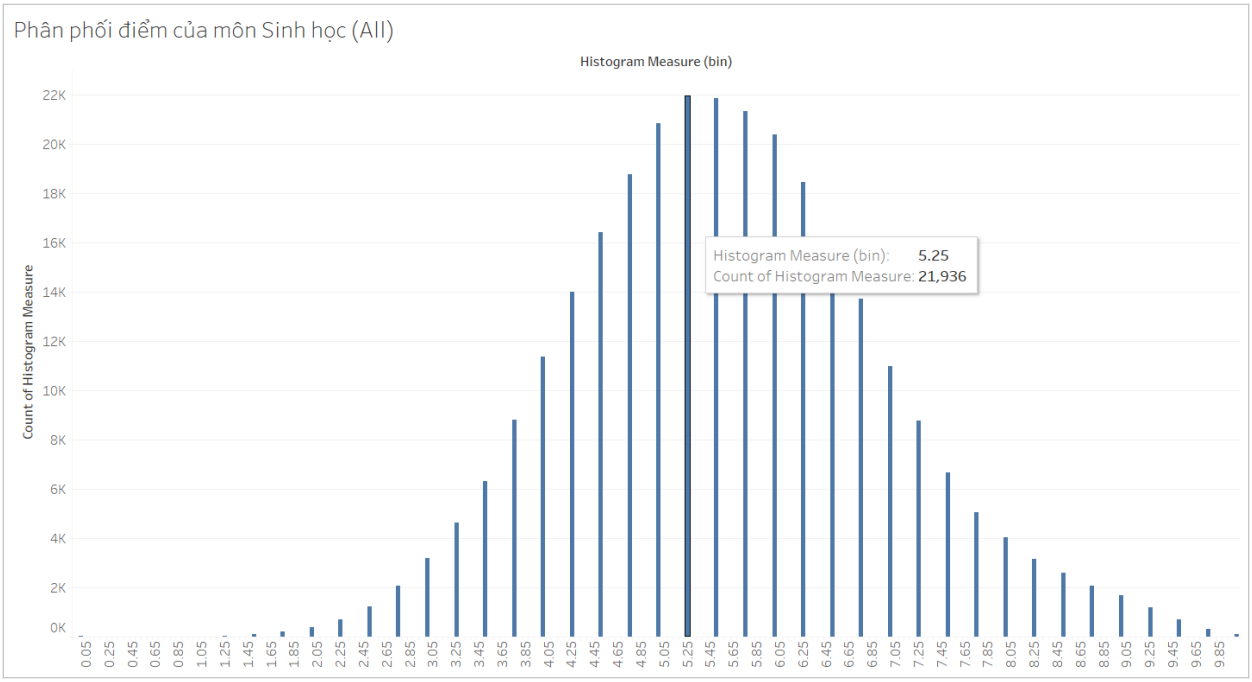


- Các biểu đồ trên thể hiện mối quan hệ giữa điểm toán trung bình của các tỉnh/thành phố so với điểm trung bình của các môn Hóa, Lí, Văn, Sinh.

- Nhìn vào biểu đồ, dựa vào độ dốc của đường xu hướng, ta thấy học sinh giỏi toán thường sẽ giỏi cả Hóa, Lí, Văn. Đáng ngạc nhiên là đường xu hướng của đồ thị 4 có độ dốc thấp hơn hẳn so với các đồ thị còn lại kể cả đồ thị Toán - Văn, nghĩa là học sinh giỏi Toán có khả năng giỏi Văn cao hơn giỏi Sinh.

- Link Dashboard: tại đây

6. Trực quan phân phối điểm của các môn trên phạm vi cả nước và từng địa phương



- Sử dụng biểu đồ histogram cho thấy được phân phối điểm môn được chọn trên phạm vi cả nước và từng khu vực.

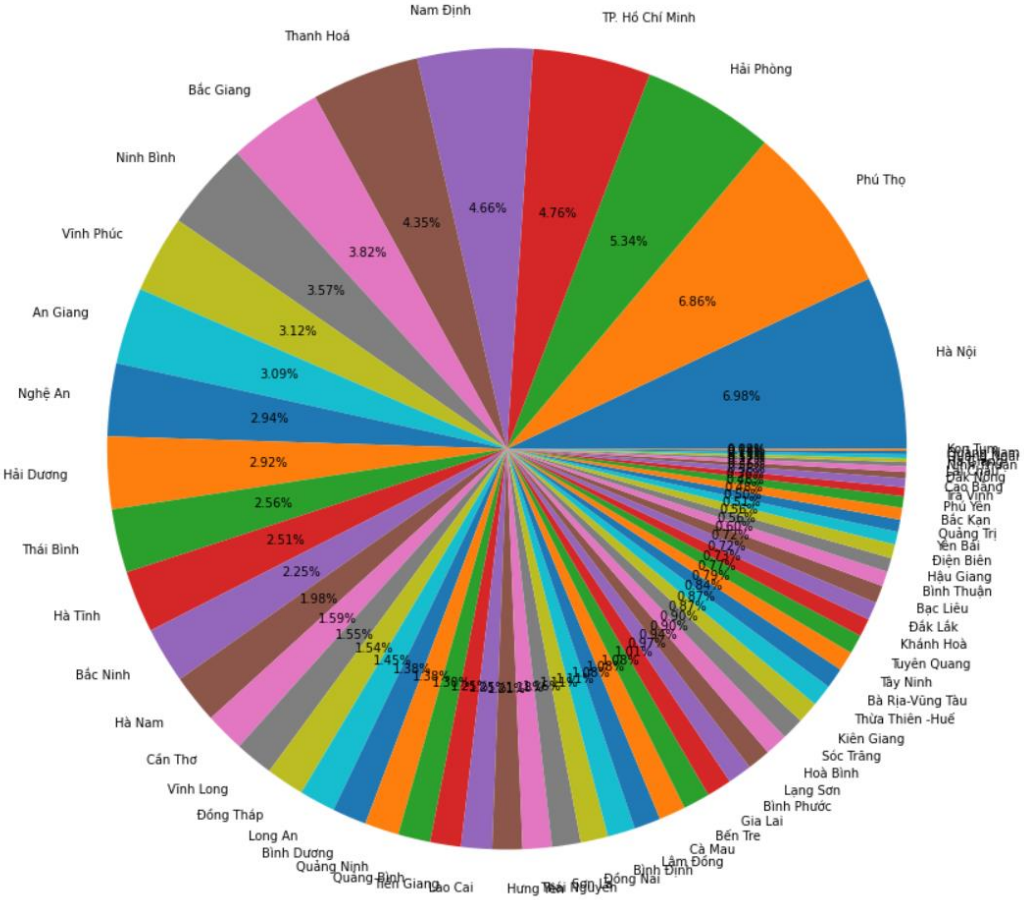
- Có thể xem xét được số điểm thí sinh đạt được hầu hết là bao nhiêu.

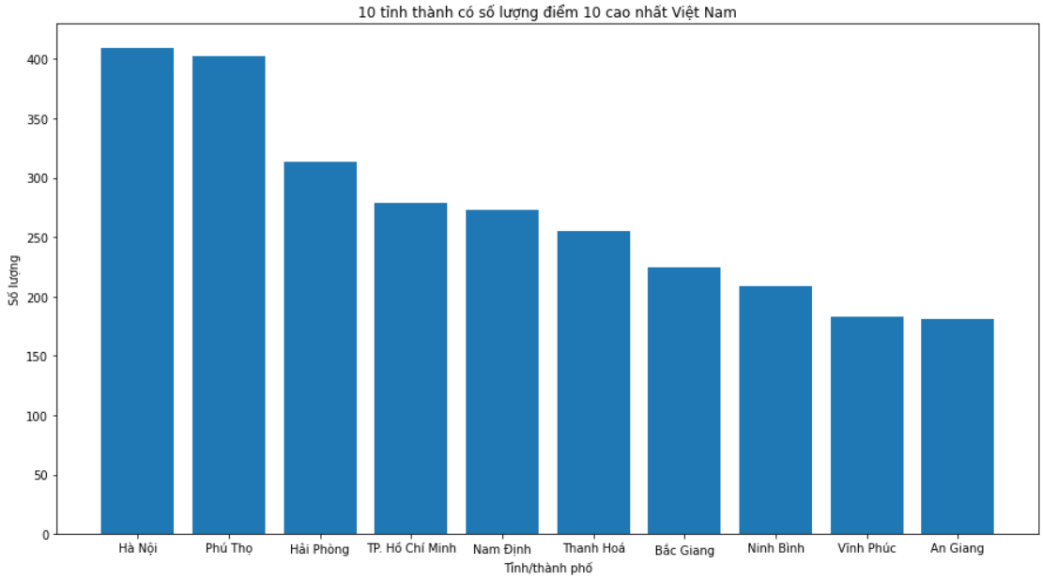
- Nhìn vào biểu đồ trên, đối với môn Sinh học trên phạm vi toàn quốc, số điểm nhiều thí sinh đạt được nhất là 5,25.

- Có thể xem điểm có tuân theo phân phối chuẩn hay không, từ đó suy ra được một số kết luận.

- Tương tác: Chọn môn ở biểu đồ điểm trung bình cả nước, chọn khu vực ở Maps (có thể chọn một/nhiều tỉnh hoặc cả nước). Khi đó, biểu đồ phân phối điểm của môn trên khu vực được chọn sẽ được cập nhật tương ứng.

7. Tỉ lệ số lượng điểm 10 ở tất cả các tỉnh thành với cả nước và 10 tỉnh thành có số lượng điểm 10 cao nhất.



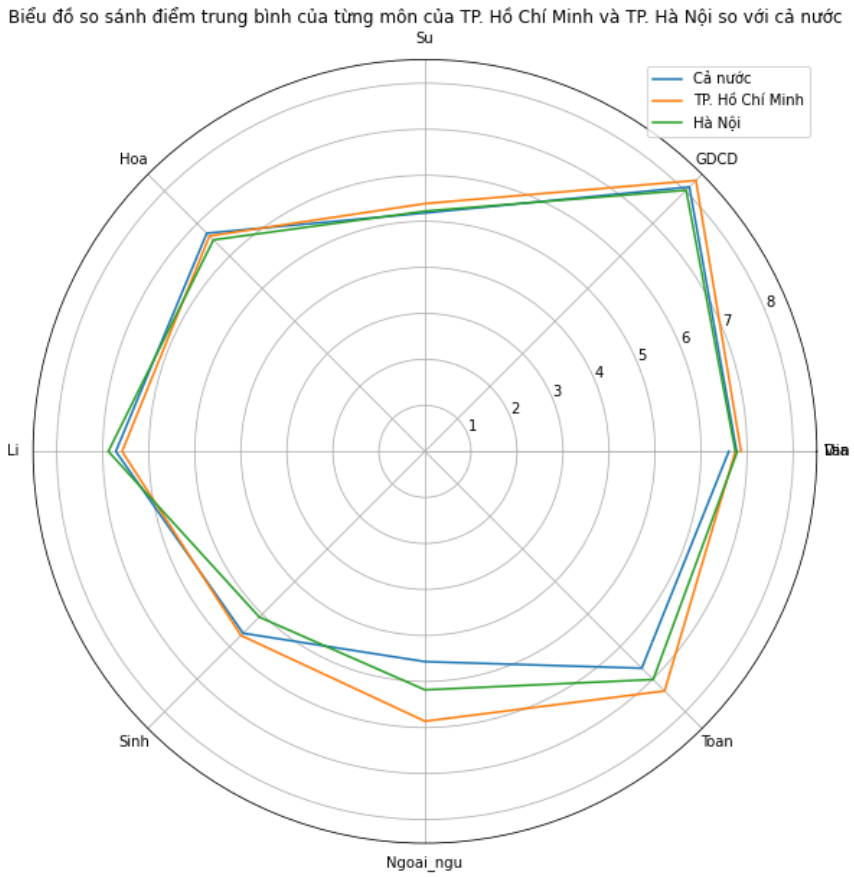


- Biểu đồ cho ta thấy được số lượng điểm 10 của thành phố Hà Nội là cao nhất cả nước chiếm tỉ lệ là 6.98%, số lượng điểm 10 của tỉnh Kon Tum là thấp nhất và chiếm tỉ lệ là 0.09%.

- Trong top 10 tỉnh/thành phố có số lượng điểm 10 cao nhất lần lượt là: Hà Nội, Phú Thọ, Hải Phòng, TP. Hồ Chí Minh, Nam Định, Thanh Hóa, Bắc Giang, Ninh Bình, Vĩnh Phúc và An Giang có 2728 điểm 10 chiếm gần 50% số lượng điểm 10 cụ thể là 46,57% của cả nước.

- Hầu hết các tỉnh/thành phố (8/10 tỉnh/thành phố) có số lượng điểm 10 cao nhất là đến từ miền Bắc Việt Nam.

8. So sánh điểm trung bình tất cả các môn giữa TP. Hồ Chí Minh, TP. Hà Nội và cả nước

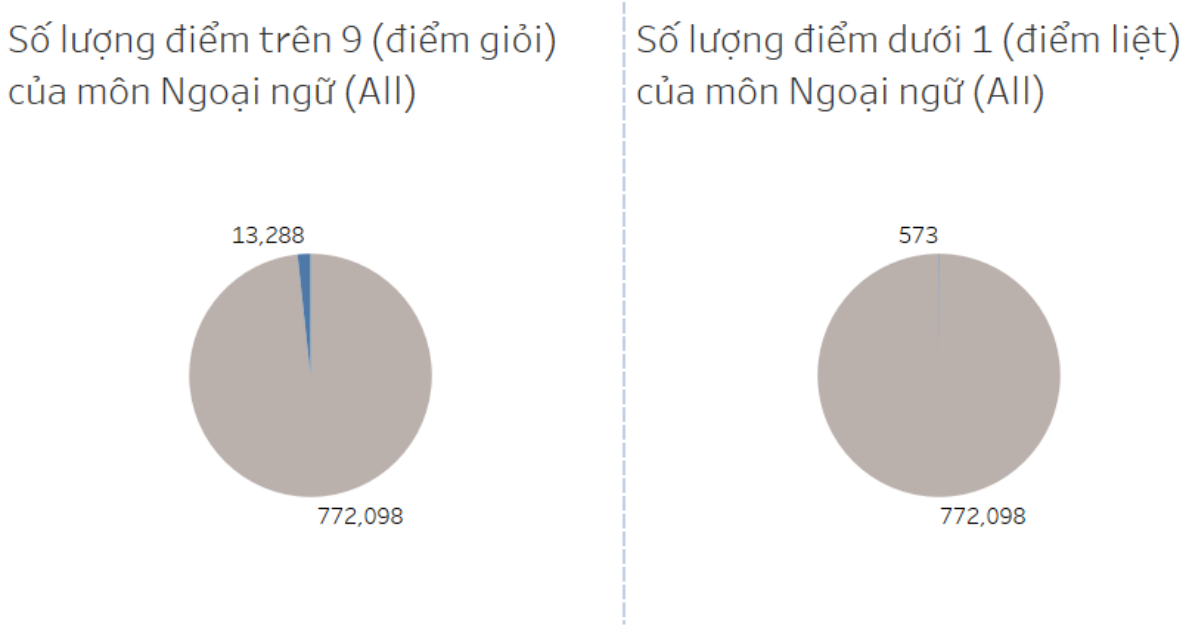


- Ở các môn GDCD, Địa, Hóa, Lí thì sự chênh lệch giữa TP. Hồ Chí Minh, TP. Hà Nội so với cả nước là không đáng kể.

- Môn Sinh của TP. Hà Nội có điểm trung bình thấp hơn so với TP. Hồ Chí Minh và cả nước.

- Sự chênh lệch điểm trung bình lớn nhất nằm ở môn Ngoại ngữ và môn Toán: Ở đây TP. Hồ Chí Minh có điểm trung bình cao nhất ở cả 2 môn, theo sau là TP. Hà Nội và điểm trung bình của cả nước ở 2 môn này thì thấp hơn đáng kể.

9. **Trực quan số lượng điểm giỏi (>=9) và điểm liệt (<=1) của các môn trên phạm vi cả nước và từng địa phương**



- Sử dụng biểu đồ tròn dễ dàng so sánh được số lượng điểm >=9 hoặc <=1 so với tổng thể, hoặc so sánh với môn khác.

- Nhìn vào biểu đồ trên, ta thấy được số điểm liệt của môn Ngoại ngữ (573) là rất cao so với tổng thể, cũng rất cao so với 2 môn bắt buộc khác là Toán (208) và Văn (137). Số lượng học sinh đạt điểm giỏi cũng rất thấp so với Toán (65150), cho thấy chất lượng học môn Ngoại ngữ ở nước ta là chưa tốt.

- Tương tác: Chọn môn ở biểu đồ điểm trung bình cả nước, chọn khu vực ở Maps (có thể chọn một/nhiều tỉnh hoặc cả nước). Khi đó, 2 biểu đồ số lượng điểm giỏi và điểm liệt của môn trên khu vực được chọn sẽ được cập nhật tương ứng.

III. PHÂN TÍCH DỮ LIỆU

1. Thống kê mô tả

	Dia	GDCD	Hoa	Li	Ngoai_ngu	Sinh	Su	Toan	Van	sbd
count	555072.000000	482980.000000	295536.000000	293287.000000	772098.000000	290377.000000	568581.000000	866581.000000	856565.000000	8.704860e+05
mean	6.773215	8.125164	6.705225	6.718001	4.568337	5.588041	5.181097	6.662271	6.601825	2.800845e+07
std	1.171702	1.090310	1.607924	1.497900	1.808446	1.350572	1.598987	1.819042	1.248463	1.898843e+07
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000001e+06
25%	6.000000	7.500000	5.500000	5.750000	3.200000	4.750000	4.000000	5.400000	6.000000	1.200657e+07
50%	7.000000	8.250000	7.000000	7.000000	4.200000	5.500000	5.000000	7.000000	6.750000	2.801803e+07
75%	7.500000	9.000000	8.000000	7.750000	5.600000	6.500000	6.250000	8.000000	7.500000	4.400927e+07
max	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	6.400582e+07

2. Phân tích dữ liệu đơn giản

Từ các số liệu từ bảng thống kê mô tả, ta có thể rút ra một số nhận xét:

- Không có môn học nào có số lượng thí sinh là tối đa.

- Số lượng thí sinh thi môn Toán là cao nhất và môn Sinh là thấp nhất.

- Số lượng thí sinh thi khối Khoa học Xã hội gồm Sử, Địa và Giáo dục công dân cao gấp đôi so với số lượng thí sinh thi khối Khoa học Tự nhiên gồm Lí, Hóa và Sinh.

- Đa số điểm trung bình của các môn học nằm trong khoảng [5; 7] điểm.

- Điểm trung bình của môn Giáo dục công dân là cao nhất (8.125 đ) và của môn Ngoại ngữ là thấp nhất (4.568 đ).

- Chỉ có duy nhất điểm trung bình của môn Ngoại ngữ là dưới trung bình

- Tất cả các môn đều có điểm thấp nhất là 0 điểm và điểm cao nhất là 10 điểm.

3. Phân tích hồi quy, giải thích và dự báo

- Ta sử dụng mô hình logistic regression để xem xét điểm Toán sẽ ảnh hưởng đến tỉ lệ vào được Nhóm ngành Máy tính và công nghệ thông tin của Trường Đại học Khoa học Tự nhiên – ĐHQG TP.HCM. (Điểm chuẩn năm 2020 của 4 khối A00, A01, D07, B08 là 27.2).

- Ta xét đến khối A00 (Toán, Lí, Hóa) và khối A01 (Toán, Lí, Ngoại ngữ N1 (Tiếng Anh)).

- Xây dựng mô hình:

```
1 x_Khoi_A = df_Khoi_A["Toan"]
2 y_Khoi_A =df_Khoi_A["Ket_qua"]
3 MLE_Khoi_A = Logit(y_Khoi_A, sm.add_constant(x_Khoi_A)).fit()
4 MLE_Khoi_A.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.040406
         Iterations 12
```

Logit Regression Results

Dep. Variable:	Ket_qua	**No. Observations:**	291252
Model:	Logit	**Df Residuals:**	291250
Method:	MLE	**Df Model:**	1
Date:	Wed, 22 Jun 2022	**Pseudo R-squ.:**	0.3910
Time:	11:37:47	**Log-Likelihood:**	-11768.
converged:	True	**LL-Null:**	-19323.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-48.1823	0.550	-87.671	0.000	-49.259	-47.105
Toan	4.8775	0.059	83.270	0.000	4.763	4.992

```
1 x_Khoi_A1 = df_Khoi_A1["Toan"]
2 y_Khoi_A1 = df_Khoi_A1["Ket_qua"]
3 MLE_Khoi_A1 = Logit(y_Khoi_A1, sm.add_constant(x_Khoi_A1)).fit()
4 MLE_Khoi_A1.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.021769
         Iterations 12
```

Logit Regression Results

Dep. Variable:	Ket_qua	**No. Observations:**	283114
Model:	Logit	**Df Residuals:**	283112
Method:	MLE	**Df Model:**	1
Date:	Wed, 22 Jun 2022	**Pseudo R-squ.:**	0.3123
Time:	11:37:58	**Log-Likelihood:**	-6163.0
converged:	True	**LL-Null:**	-8962.0
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-44.0927	0.746	-59.096	0.000	-45.555	-42.630
Toan	4.3339	0.079	54.588	0.000	4.178	4.489

- Kết quả mô hình:

$$g = \frac{1}{1 + e^{-(-48.1823 + 4.8775x)}}$$

$$g = \frac{1}{1 + e^{-(-44.0927 + 4.3339x)}}$$

- Trực quan mô hình:

- Nhận xét:

+ Nếu chỉ đạt điểm 9 trở xuống thì tỉ lệ đậu vào Trường gần như là bằng 0.

+ Dù đạt được điểm tuyệt đối thì tỉ lệ đậu vào Trường của cả 2 khối lần lượt là hơn 60% và hơn 30%. Đây là một tỉ lệ không cao. Chứng tỏ chỉ một mình điểm Toán vẫn chưa giúp thí sinh có cơ hội đậu vào Trường.

+ Tỉ lệ đậu vào Trường của khối A00 cao hơn khối A01 gần như là gấp đôi. Chứng tỏ các tỉ thí sinh sẽ chọn khối A00 để vào Trường sẽ cao hơn khối A01.

C. TÀI LIỆU THAM KHẢO

GitHub Repositories 'bee-university', Tran Minh Tuan, github.com/beecost/bee-university