Trường Đại học Khoa học Tự nhiên – Đại học Quốc gia TP.HCM
KHOA CÔNG NGHỆ THÔNG TIN

Môn: TRỰC QUAN HÓA DỮ LIỆU
ĐỒ ÁN MÔN HỌC HỌC KÌ II – NĂM HỌC 2021-2022

Thành viên

- 19120330 - Nguyễn Đoan Phúc
- 19120606 - Nguyễn Đình Hoàng Nguyên
- 19120643 - Đào Thị Thiện Tâm
- 19120683 - Thái Trung Tín
- 19120715 - Nguyễn Kha Vĩ

Nội dung chính

1. Giới thiệu đề tài.
2. Dữ liệu của đề tài.
3. Tiền xử lý dữ liệu.
4. Trực quan hóa dữ liệu.
5. Phân tích dữ liệu.

1.

Giới thiệu đề tài

1. Giới thiệu đề tài

Kỳ thi tốt nghiệp THPT là một kỳ thi quan trọng trong hệ thống giáo dục Việt Nam và dành cho học sinh lớp 12. Mục đích của kỳ thi này là công nhận việc hoàn tất chương trình học phổ thông của học sinh và là điều kiện cần để tham gia xét tuyển các bậc đại học và cao đẳng. Đề tài lần này đó chính là dùng dữ liệu điểm thi năm 2020 để phân tích và trực quan.

2.

Dữ liệu của đề tài

2. Dữ liệu của đề tài

- Dữ liệu được sử dụng trong đề tài này là điểm thi THPT năm 2020.
- Dữ liệu được lấy từ trang https://github.com/beecost/bee-university

```
# 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

3.

Tiền xử lý dữ liệu

3. Tiền xử lý dữ liệu

a. Khám phá dữ liệu.

- Bộ dữ liệu điểm thi THPT Quốc gia 2020 có 870486 dòng, mỗi dòng tương ứng với một học sinh và 12 biến, chứa thông tin về điểm thi các môn học, mã môn ngoại ngữ, số báo danh.
- Dữ liệu từng cột:

```
# 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
```

3. Tiền xử lý dữ liệu

b. Làm sạch dữ liệu.

+ 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!
```

3. Tiền xử lý dữ liệu

b. Làm sạch dữ liệu.

- Xử lý các giá trị “NaN”.
- 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” => Các thí sinh này được miễn thi môn ngoại ngữ và không có thí sinh nào bị mất điểm môn ngoại ngữ.

```
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
```

3. Tiền xử lý dữ liệu

b. Làm sạch dữ liệu.

- Thêm biến “Cum_thi”.
- Ta sẽ thêm tên cụm dự thi 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)])
```

3. Tiền xử lý dữ liệu

c. 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)
```

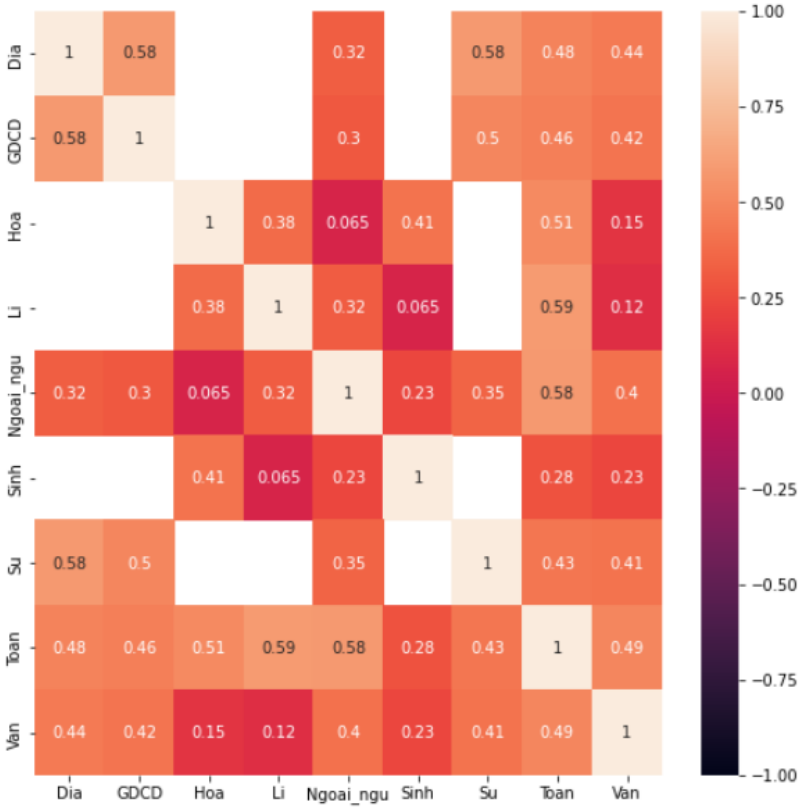
	Unnamed: 0	Dia	GDCD	Hoa	Li	Ma_mon_ngoai_ngu	Ngoai_ngu	Sinh	Su	Toan	Van	sbd	Cum_thi
0	0	7.00	6.50	NaN	NaN	N1	4.2	NaN	4.75	6.4	6.75	18014547	Bắc Giang
1	1	7.75	7.75	NaN	NaN	N1	2.8	NaN	3.75	7.6	6.00	18014530	Bắc Giang
2	2	6.50	NaN	NaN	NaN	NaN	NaN	NaN	4.00	4.8	4.75	18014521	Bắc Giang
3	3	8.00	9.50	NaN	NaN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

4.

Trực quan hóa dữ liệu

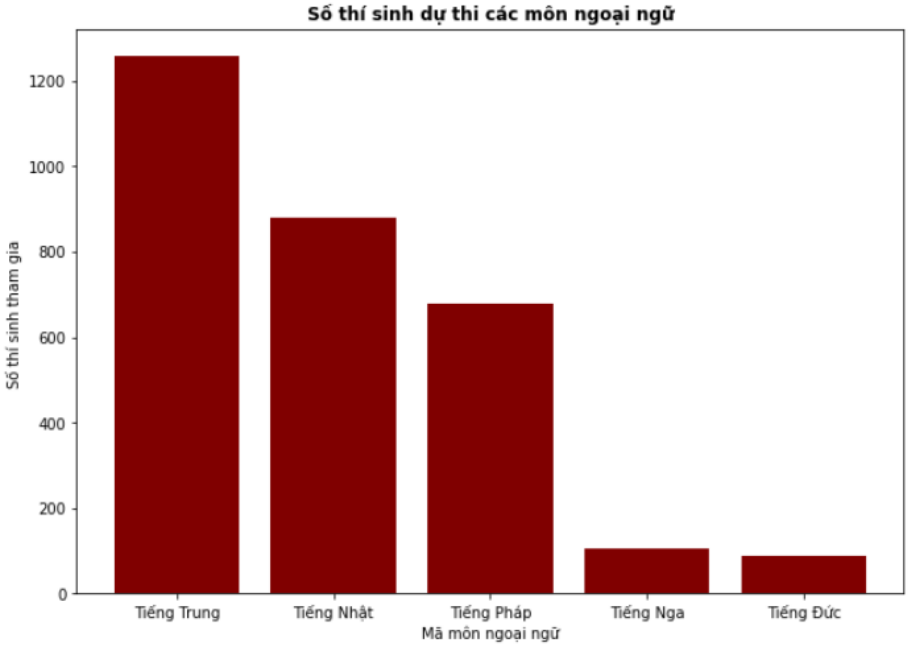
4. Trực quan hóa dữ liệu

- Mối quan hệ tương quan giữa điểm thi các môn học.
- Sử dụng biểu đồ heatmap để trực quan câu hỏi này.



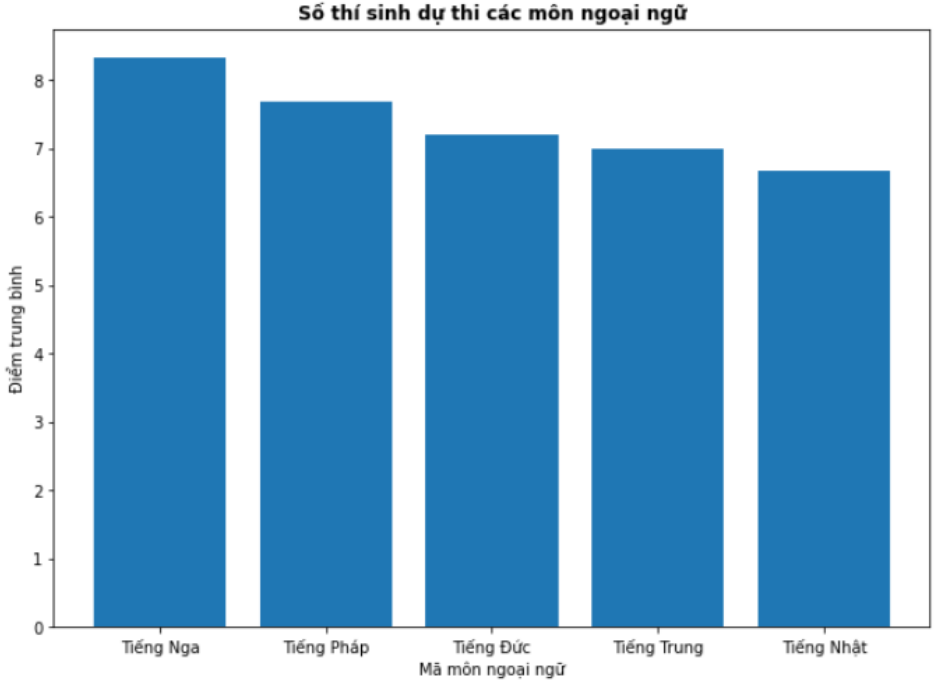
4. Trực quan hóa dữ liệu

- Môn ngoại ngữ nào (ngoại trừ Tiếng Anh) có nhiều thí sinh tham gia nhất?
- Sử dụng biểu đồ cột (bar chart) để trực quan câu hỏi này.



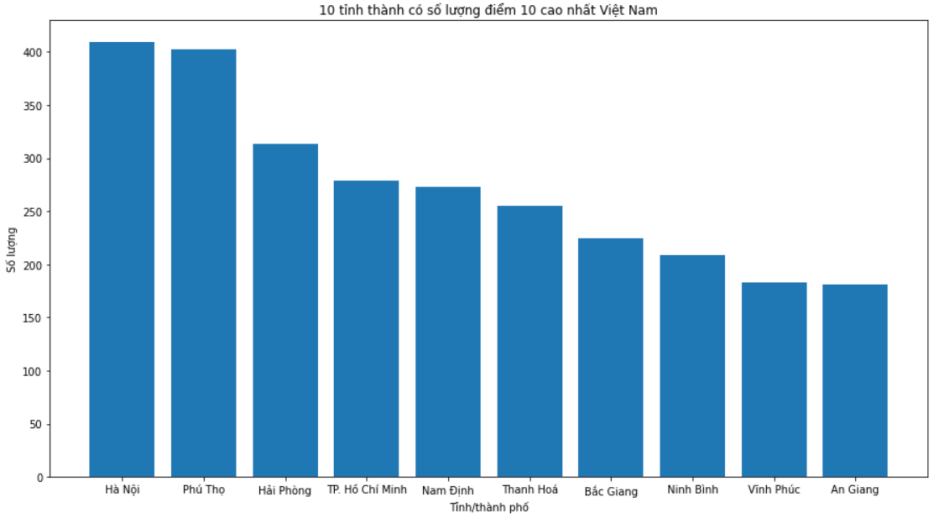
4. Trực quan hóa dữ liệu

- Môn ngoại ngữ nào có điểm thi trung bình cao nhất?
- Ta cũng sử dụng biểu đồ cột (bar chart) để trực quan câu hỏi này.



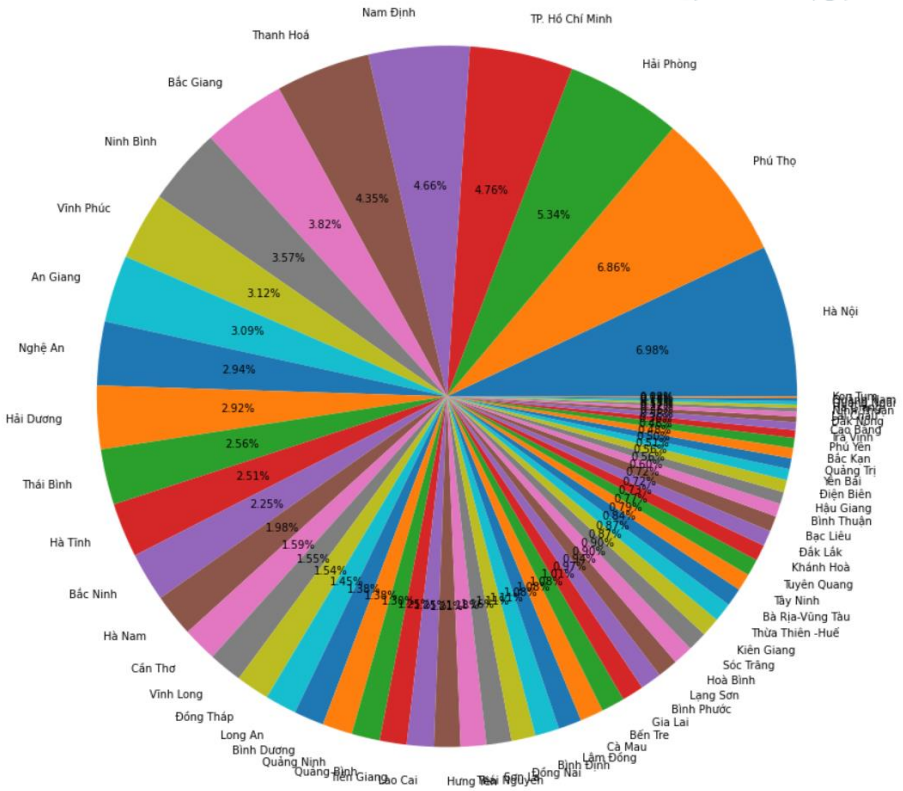
4. Trực quan hóa dữ liệu

- 10 tỉnh thành có số lượng điểm 10 cao nhất.
- Ta cũng sử dụng biểu đồ cột (bar chart) để trực quan câu hỏi này.



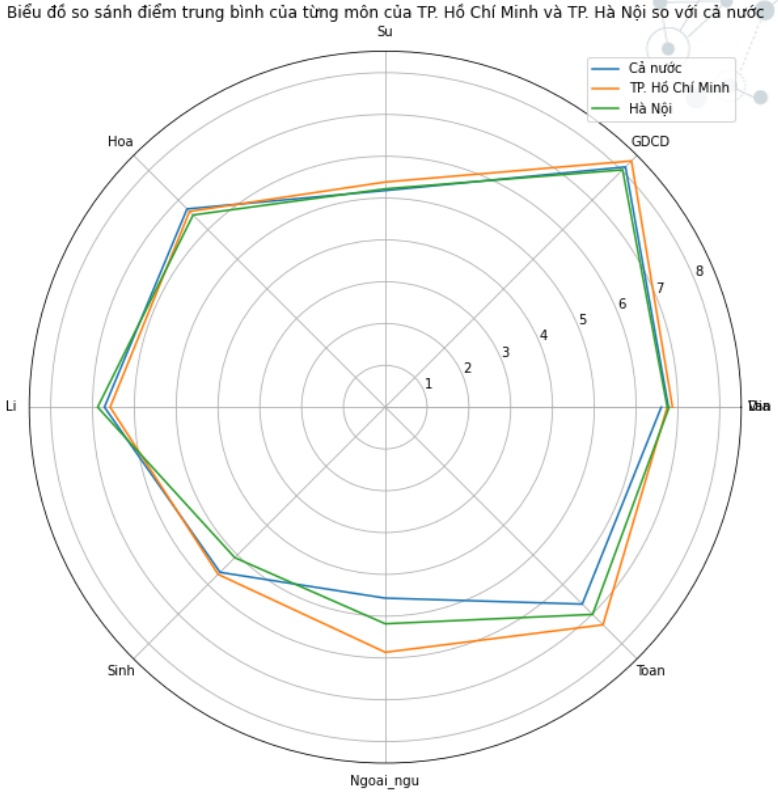
4. Trực quan hóa dữ liệu

- Tỉ lệ số lượng điểm 10 ở tất cả các tỉnh thành với cả nước.
- Sử dụng biểu đồ tròn (pie chart) để trực quan câu hỏi này.



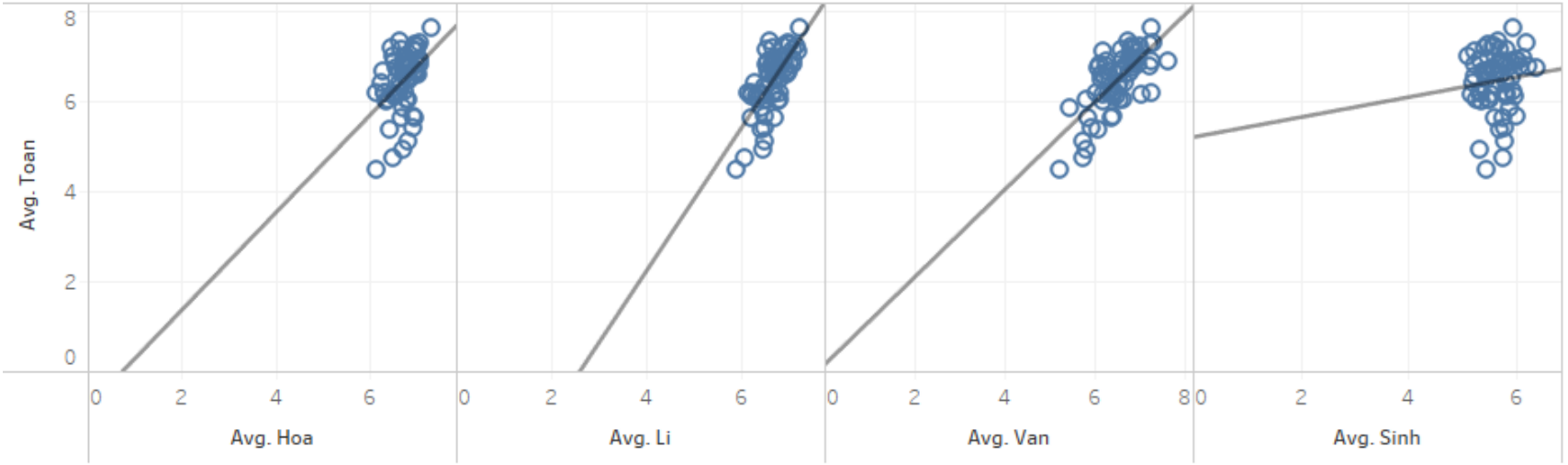
4. Trực quan hóa dữ liệu

- So sánh điểm trung bình tất cả các môn giữa TP. Hồ Chí Minh, TP. Hà Nội và cả nước.
- Sử dụng biểu đồ radar (radar chart) để trực quan câu hỏi này.



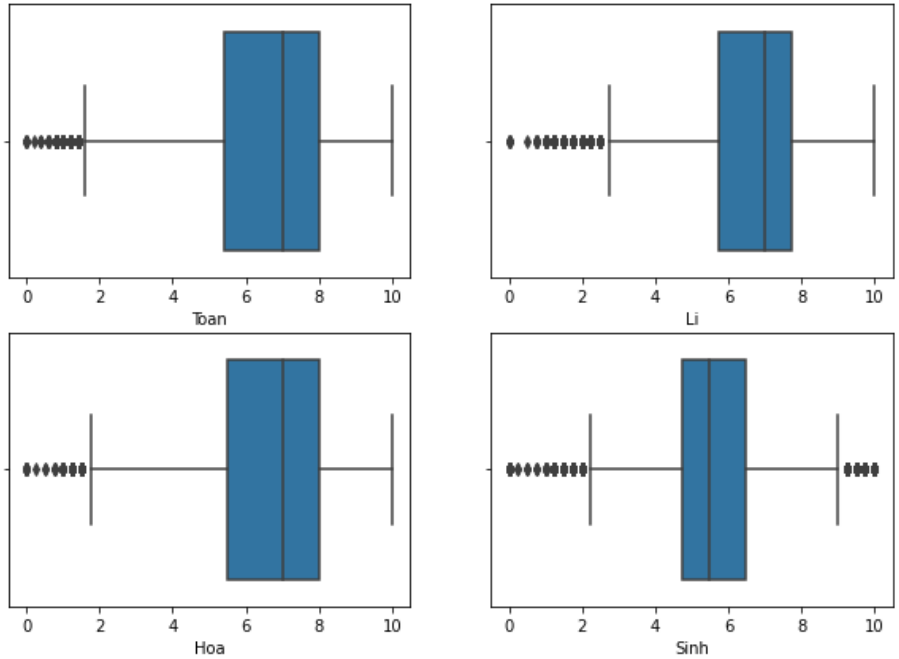
4. Trực quan hóa dữ liệu

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



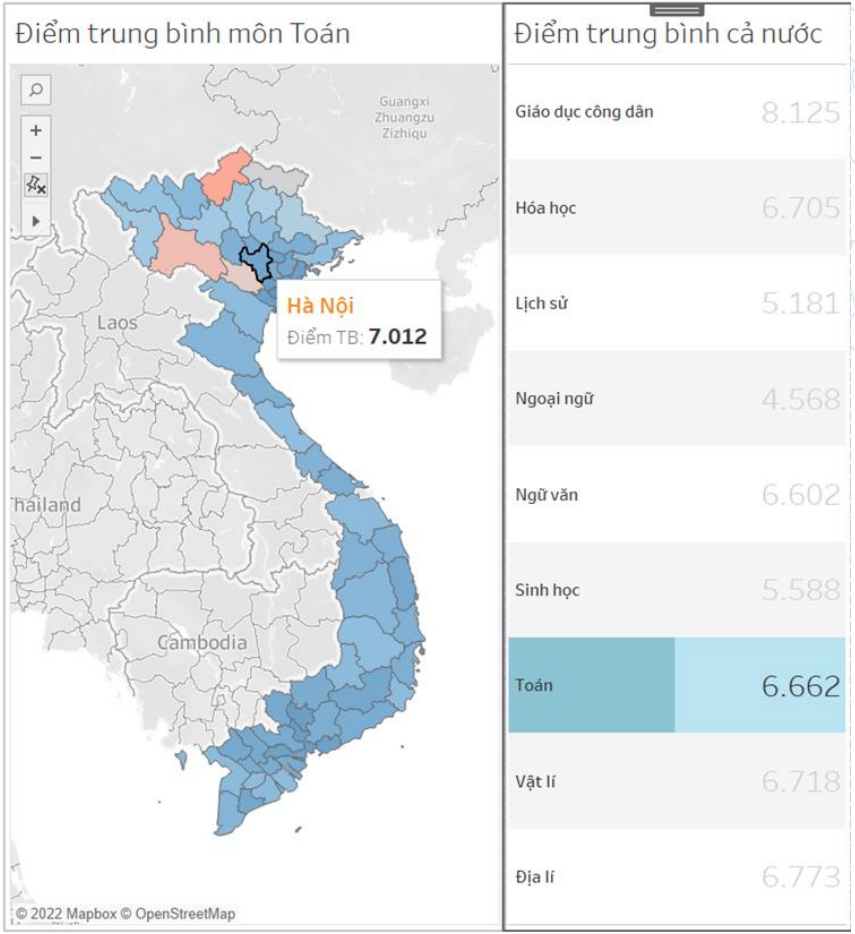
4. Trực quan hóa dữ liệu

- Sự phân bố và phân tán dữ liệu của các môn tự nhiên.
- Sử dụng biểu đồ box (box plot) để trực quan câu hỏi này.



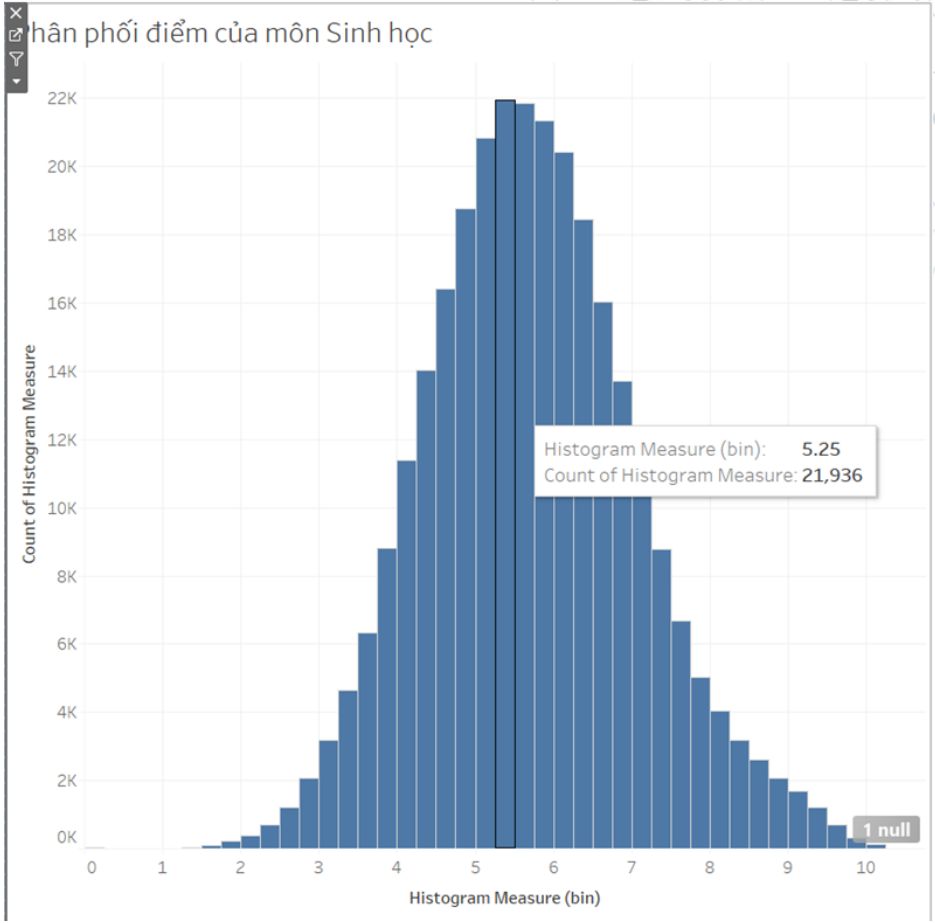
4. Trực quan hóa dữ liệu

- Điểm trung bình các môn trên phạm vi cả nước và từng địa phương.
- Sử dụng biểu đồ bản đồ để trực quan câu hỏi này.

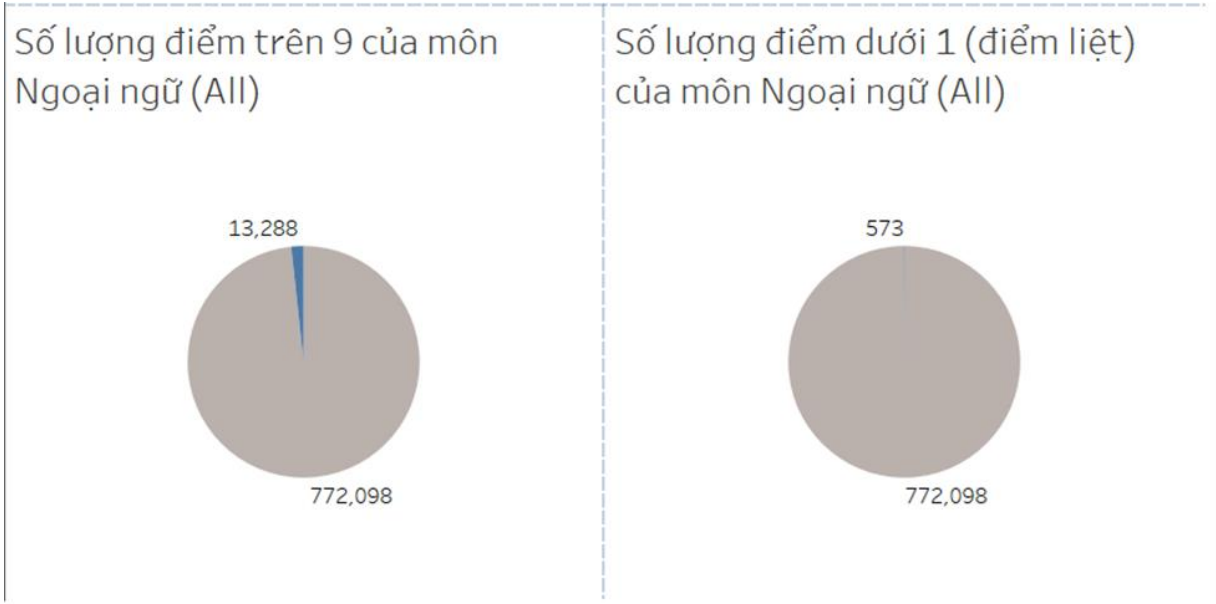


4. Trực quan hóa dữ liệu

- Phân phối điểm của các môn trên phạm vi cả nước và từng địa phương.
- Sử dụng biểu đồ histogram để trực quan câu hỏi này.



4. Trực quan hóa dữ liệu



Số lượng điểm giỏi và điểm liệt của các môn trên phạm vi cả nước và từng địa phương. Sử dụng biểu đồ tròn để trực quan câu hỏi này.

5.

Phân tích dữ liệu

5. Phân tích dữ liệu

a. 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

5. Phân tích dữ liệu

b. 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 cơ bản như sau:

- 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.

5. Phân tích dữ liệu

c. 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)).

5. Phân tích dữ liệu

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

- 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

5. Phân tích dữ liệu

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

- Kết quả mô hình:

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

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

5. Phân tích dữ liệu

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

- Trực quan mô hình:

5. Phân tích dữ liệu

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

➢ 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.

Thank you for listening!


Questions?


Ask
us!