BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN – ĐHQG TPHCM

Báo cáo

Chủ đề: Phân tích và dự đoán điểm thi THPT qua các năm từ năm 2020 - 2022

Học phần: Nhập môn Khoa Học Dữ Liệu
Giảng viên: Hà Văn Thảo
Lớp: 20KDL1
Nhóm 21

Họ và tên	MSSV
Trần Hoàng Anh	20280004
Nguyễn Quốc Bảo	20280006
Nguyễn Minh Hoàng Đạt	20280014
Hà Thành Long	20280061

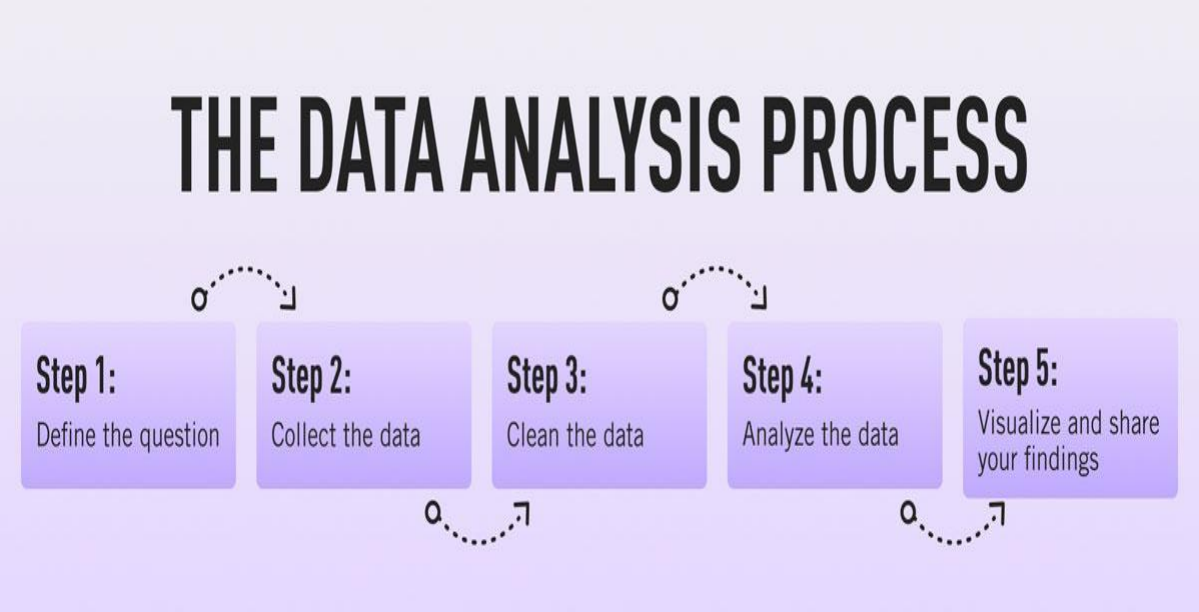
Mục lục

Source: QuocBao02/Python-for-Data-Science---Final-Project (github.com)

Đặt vấn đề

Cuộc thi trung học phổ thông quốc gia là một trong những cuộc thi quan trọng nhất đối với mỗi học sinh. Mỗi năm sẽ có khoảng 1 triệu học sinh đến từ các trường trải dài từ bắc vào nam tham gia cuộc thi để có thể xét tốt nghiệp trung học phổ thông cũng như là điểm thi để xét tuyển các trường đại học. Như vậy, dữ liệu về điểm thi các môn học của học sinh sẽ ngày càng nhiều, phân bố điểm sẽ khác nhau giữa các môn học cũng như giữa các tỉnh, thành phố.

Do đó, bài báo cáo sẽ đi vào việc phân tích dữ liệu về điểm thi trung học phổ thông ở TP.HCM qua các năm từ 2020 đến 2022.Từ các thao tác ban đầu là lấy dữ liệu từ trang web uy tín,sau đó xử lý dữ liệu ta thu được các file chứa dữ liệu: mã số học sinh và điểm của các môn thi tương ứng.Tiếp theo,ta trực quan hóa dữ liệu bằng cách vẽ các biểu đồ cột thể hiện phổ điểm các môn thi, số điểm liệt và số điểm tối đa, trung bình các khối thi,... Bước cuối cùng ta dựa vào dữ liệu thu được từ năm 2020, 2021 dùng mô hình Linear Regression trong thư viện Sklearn của python để dự đoán phổ điểm của một môn học bất kì trong năm 2022 sau đó ta đối chứng với dữ liệu thực của năm 2022 để kiểm định độ chính xác mô hình.



1. Các thư viện được sử dụng:

```
import numpy as np
```

*Thư viện numpy: là thư viện mã nguồn mở dùng xử lý mảng trong python được sử dụng rộng rãi. Numpy dễ học, dễ tiếp cận là thư viện cơ bản. Có khoảng hơn 450 thư viện trong python phụ thuộc vào numpy ví dụ như: Pandas, Scipy, Keras….

Cách cài đặt thư viện numpy: !pip install numpy

```
import pandas as pd
```

* Thư viện pandas: là thư viện mã nguồn mở cung cấp các công cụ phân tích dữ liệu và cấu trúc dữ liệu hiệu suất cao trong python. Có thể sử dụng để phân tích, thao tác với dữ liệu, xử lý, lọc, tổng hợp dữ liệu. Hai thành phần chính của Pandas là Series và DataFrame. Một Series về cơ bản là một cột và một Data Frame là một mảng đa chiều. Trong bài báo cáo lần này chúng ta sử dụng chính là DataFrame, từ việc khởi tạo, lưu trữ và trích xuất dữ liệu, phục vụ cho việc phân tích, trực quan hóa và dự đoán điểm thi. Với các hàm được sử dụng là pandas.read_csv(), khởi tạo Data Frame là pandas.DataFrame, Pandas.DataFrame.columns(),để mô tả dữ liệu ta dùng DataFrame.describe() ,và để lưu Data Frame là DataFame.to_csv(),...

Cách cài đặt thư viện pandas: !pip install pandas

```
from bs4 import BeautifulSoup
```

*Thư viện BeautifulSoup: là thư viện hỗ trợ lấy thông tin từ các trang web là các tệp HTML và XML. Nó hoạt động cùng với các parser (trình phân tích cú pháp) cung cấp cho bạn các cách để điều hướng, tìm kiếm và chỉnh sửa trong parse tree (cây phân tích được tạo từ parser). Nhờ các parser này nó đã giúp các lập trình viên tiết kiệm được nhiều giờ làm việc. BeautifulSoup cũng là một thư viện nổi tiếng thường được dùng trong Web crawling của python. Với các hàm hỗ trợ có sẵn ta dễ dàng tìm được một thẻ (tab), một class_id, ... trong một file HTML hoặc XML.

Cách cài đặt thư viện BeautifulSoup: !pip install beautifulsoup4

```
from urllib.request import urlopen
```

*urllib: là một module của Python có thể dùng để mở các URL. Nó định nghĩa các hàm và lớp giúp thao tác với URL.Với nó chúng ta có thể truy cập và trích xuất dữ liệu từ internet như XML, HTML, JSON, ...

*urllib.request: là mô-đun định nghĩa các hàm và lớp giúp mở URL (chủ yếu là HTTP) trong một thế giới phức tạp - cơ bản và thông báo xác thực, chuyển hướng, cookie,...

Cách cài đặt thư viện: !pip install urllib3

```
import matplotlib.pyplot as plt
```

*matplotlib: là thư viện vẽ đồ thị mạnh trong python với numpy. Được sử dụng để trực quan hóa dữ liệu. Module được sử dụng nhiều nhất của Matplotlib là Pyplot. Cách cài đặt thư viện matplotlib:

*pyplot: là một module của Matplotlib cung cấp cách vẽ đồ thị gồm các hàm đơn giản để thêm các thành phần plot như lines, images, text, v.v. vào các axes trong figure.

```
from sklearn.linear_model import LinearRegression
```

*Scikit-learn (Sklearn) là thư viện mạnh mẽ nhất dành cho các thuật toán học máy được viết trên ngôn ngữ Python. Thư viện cung cấp một tập các công cụ xử lý các bài toán machine learning và statistical modeling gồm: classification, regression, clustering, và dimensionality reduction.

*Sklearn.linear_model: thực hiện các mô hình hồi quy tuyến tính. Trong trường hợp này, chúng ta sử dụng LinearRegression. Là một mô hình hồi quy tuyến tính cổ điển.

```
from sklearn.preprocessing import PolynomialFeatures
```

*Sklearn.preprocessing : model tiền xử lý và chuẩn hóa dữ liệu. Có các phương pháp như: chia tỷ lệ, định tâm, chuẩn hóa, nhị phân hóa. Trong trường hợp này chúng ta sử dụng PolynomialFeatures. Có tác dụng tạo ma trận đối tượng mới bao gồm tất cả các tổ hợp đa thức của các đối tượng có bậc nhỏ hơn hoặc bằng bậc đã chỉ định.

```
from sklearn.model_selection import train_test_split
```

*Sklearn.model_selection: dùng để lựa chọn mô hình.Sử dụng train_test_split để tách các mảng hoặc ma trận thành các tập con kiểm tra và huấn luyện ngẫu nhiên.

```
from sklearn.metrics import r2_score
```

*Sklearn.metrics: chứa các hàm về điểm số, số liệu hiệu suất và số liệu theo cặp và tính toán khoảng cách. Sử dụng r2_score là hàm điểm hồi quy.

*Để cài đặt thư viện sklearn: !pip install scikit-learn

2. Thu thập và lọc dữ liệu:

Dữ liệu sẽ được lấy từ trang web (web scraping)[2] Tra cứu điểm thi tốt nghiệp THPT 2022 (vietnamnet.vn) nhằm thu thập dữ liệu về điểm thi của các thí sinh ở thành phố Hồ Chí Minh bắt đầu từ mã số 02000001 (mã tỉnh là 2 số đầu tiên: 02).

Dùng hàm urlopen() để mở object dưới dạng html theo địa chỉ thì rất khó thu thập nên viết hàm Parse_web(url) để thu thập dữ liệu dưới dạng list

Parse_web(url): là hàm nhận vào địa chỉ url của trang web đích và trả về là một list chứa điểm và môn học tương ứng nếu không truy cập được url đó thì kết quả tự động trả về None. Với việc sử dụng thư viện BeautifulSoup để chọn các thẻ mong muốn trong tập tin HTML được trả về từ urllib.request.

Add_value_into_DataFrame(root_url, start_id, end_id):để tạo một data frame từ học sinh đầu tiên đến học sinh cuối cùng và thêm chúng vào một dataframe duy nhất. Trong data frame chứa các thông tin cột tương ứng gồm: ID và tên các môn học như:Toán, Lí,Hóa,Sinh,Ngoại ngữ,Sử, Địa,GDCD,Văn.

Province_id dùng để lọc lại mã tỉnh, vì id đưa vào hàm với kiểu dữ liệu string nếu sử dụng range cho for tiếp theo thì phải đưa về int mà mã tỉnh có str là “01”, “02”, … sẽ thành số 1, 2, … Nên province_id dùng để gặp trường hợp trên và lưu str “0” ở đầu lại để truyền đúng địa chỉ url vào parse_web.

Vòng lặp for i để thu thập dữ liệu của thí sinh có id từ start_id đến end_id + 1. Tạo dict có keys ứng với các cột dataframe và values của keys môn học sẽ mặc định là -1 nếu không có (vì điểm >=0 và <= 10). Nếu parse_web …

Sau đó ta lưu dataframe lưu thành file csv bằng dataframe.to_csv().Lặp lại các thao tác trên ta thu được 3 file dữ liệu chứa điểm thi THPTQG ở TP.HCM qua các năm 2020 đến 2022.

3. Phân tích và trực quan hóa dữ liệu:

3.1. Phân tích dữ liệu:

Sử dụng hàm có sẵn dataframe.describe() trong thư viện pandas để có được cái nhìn tổng quan về dataframe với các thông tin có được như là: số lượng

dòng, giá trị nhỏ nhất, lớn nhất, trung bình, độ lệch chuẩn và tứ phân vị trên từng cột. Ví dụ về dữ liệu năm 2022:

Data Statistics

```
dataframe2022.describe()
```

	Toán	Lí	Hóa	Sinh	Văn	Ngoại ngữ	GDCD	Địa	Sử
count	84094.000000	46752.000000	47182.000000	46804.000000	83072.000000	73249.000000	29545.000000	36217.000000	36476.000000
mean	7.057077	6.467969	6.591868	4.987090	6.340630	6.396445	8.201075	6.741903	6.450351
std	1.246951	1.353473	1.497060	1.433235	1.042226	1.891149	0.896812	1.062059	1.484106
min	1.000000	0.000000	0.000000	0.000000	0.250000	0.800000	0.000000	0.000000	1.000000
25%	6.400000	5.500000	5.500000	4.000000	5.750000	5.000000	7.750000	6.000000	5.500000
50%	7.200000	6.500000	6.750000	4.750000	6.500000	6.600000	8.250000	6.750000	6.500000
75%	8.000000	7.500000	7.750000	5.750000	7.000000	8.000000	8.750000	7.500000	7.500000
max	10.000000	10.000000	10.000000	10.000000	9.250000	10.000000	10.000000	10.000000	10.000000

Từ việc phân tích dữ liệu của năm 2022 ở TP.HCM.Ta thấy số lượng học sinh tham gia thi tổ hợp khoa học tự nhiên cao hơn nhiều so với tổ hợp môn khoa học xã hội, số lượng thí sinh tham gia thi các môn có sự chênh lệch do có môn Toán và môn Văn là môn thi bắt buộc.

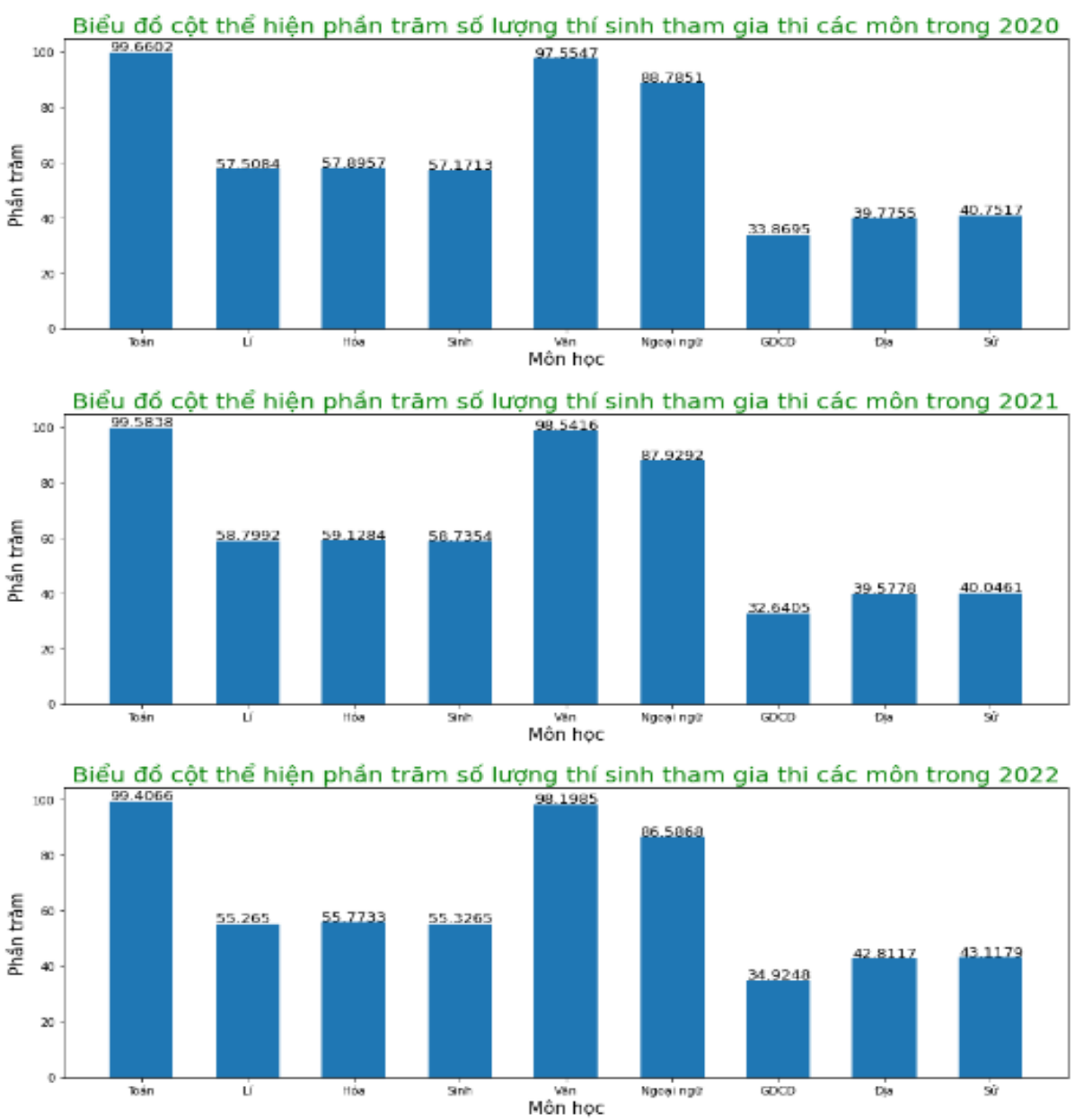
- Điểm trung bình của môn GDCD là cao nhất đạt 8.2 chứng tỏ tỉ lệ học sinh đạt điểm giỏi chiếm đa số so với tổng số thí sinh tham gia thi môn GDCD.
- Điểm trung bình của các môn còn lại không có sự chênh lệch quá nhiều đều nằm trong khoảng điểm 6.5 cho thấy sự cân bằng về mặt bằng điểm thi các môn.
- Điểm trung bình của môn Sinh học là thấp nhất chưa đến 5 điểm, cho thấy sự chênh lệch lớn với mặt bằng điểm thi chung. Từ các thông số khác ta thấy số điểm cao chiếm tỷ lệ nhỏ khi mà các điểm tứ phân vị chỉ nằm ở mức điểm trung bình và mức điểm cao nhất là 10. Các môn đều có điểm liệt khi giá trị min<=1 và điểm cao nhất của các môn là 10 trừ môn Văn là 9.25

3.2. Trực quan hóa dữ liệu:

Biểu đồ cột thể hiện số lượng thí sinh tham gia:

Hàm CountStudentOfSubject: cho ta biết số lượng học sinh tham gia thi của môn học.

Hàm PercentChart: đầu tiên ta tạo một mảng rỗng y nhằm lưu trữ phần trăm số lượng học sinh tham gia thi một môn trên tổng số học sinh tham gia thi. Cuối cùng biểu diễn nó trên đồ thị cột bằng hàm plt.bar().



Biểu đồ cột thể hiện phổ điểm của môn học

Barplot: Vì mỗi môn có phổ điểm khác nhau nên từ điển default_score có keys là phổ điểm và values là số lượng thí sinh đạt được số điểm đó sẽ chia ra 3 trường hợp:

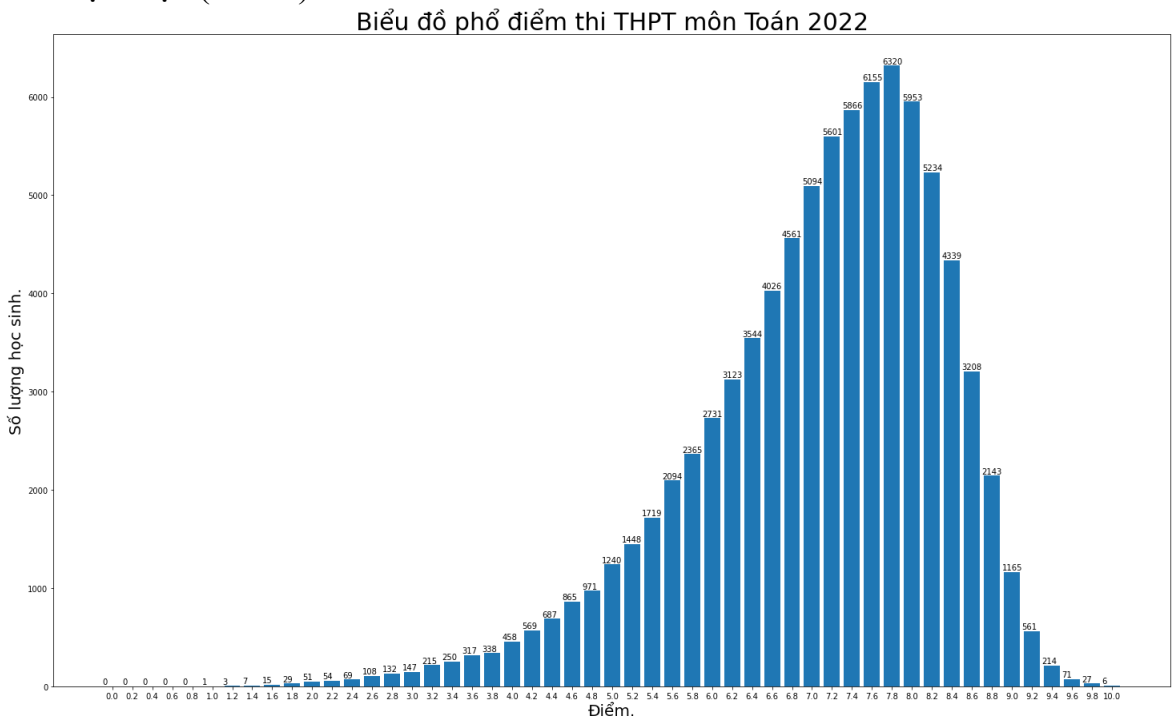
Môn Toán và Ngoại Ngữ có phổ điểm cách nhau 0.2đ

Môn Văn số điểm nó không có định bước nhảy, có thể là 0.23 hay 0.25 … Nên biến Van_score để set lại các giá trị khác nhau trong dữ liệu điểm thi môn văn

Các môn còn lại cách nhau 0.25đ

Tạo mảng 1 chiều df để lưu số điểm của môn học. Sau đó dùng vòng lặp để đếm số điểm của thí sinh đạt được (ứng với keys)

Từ keys và values của từ điển default_score ta dùng hàm bar của thư viện matplotlib để vẽ biểu đồ cột với trung hoành là phổ điểm (keys) trục tung là số thí sinh đạt được (values)

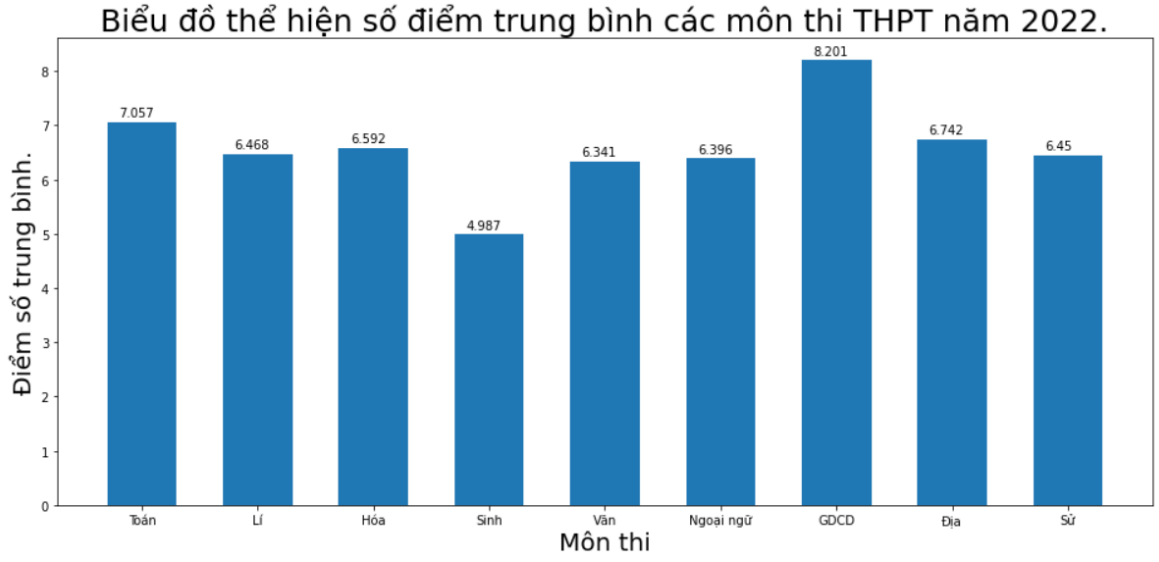


Biểu đồ phổ điểm môn Toán năm 2022 cho thấy số điểm môn Toán phổ biến chủ yếu từ 7 - 8.8 khá là cao. Đồng thời có 19 thí sinh đạt điểm 10 và có 1 thí sinh nào rơi vào điểm liệt dưới 1. Điều này cho thấy trung bình điểm môn Toán khá cao nên sẽ góp phần nâng tỉ lệ tốt nghiệp của các thí sinh

Biểu đồ cột thể hiện điểm trung bình các môn thi

Meanplot: Tạo danh sách y_mean lưu điểm trung bình tất cả thí sinh của từng môn học

Sau đó dùng hàm bar của thư viện matplotlib để vẽ biểu đồ cột với dữ liệu trục hoành là tên các môn học, trục tung là y_mean (thang điểm 10) các môn tương ứng

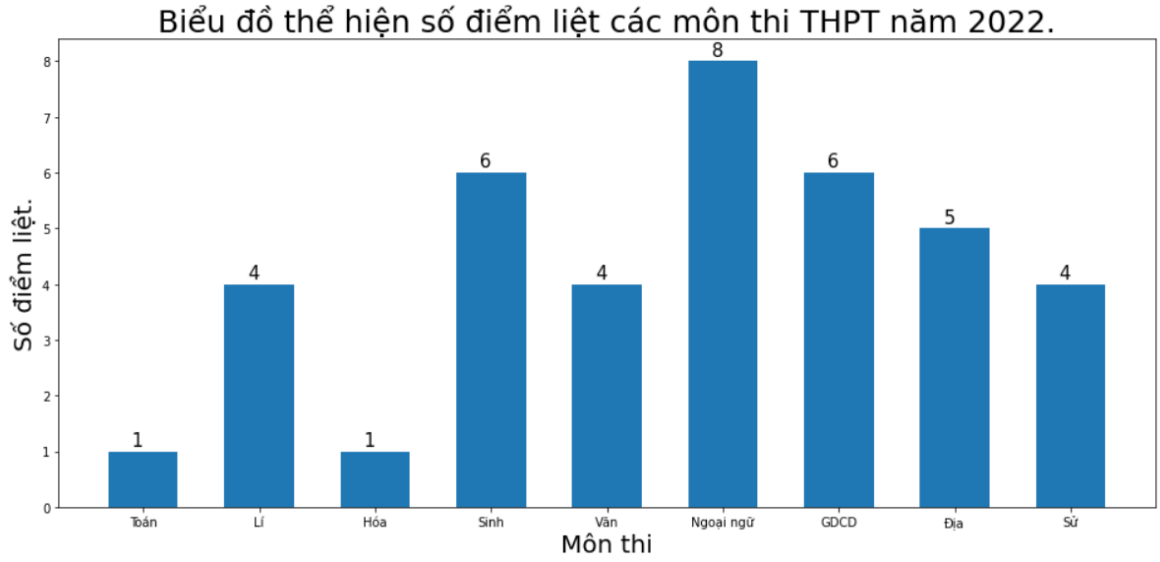


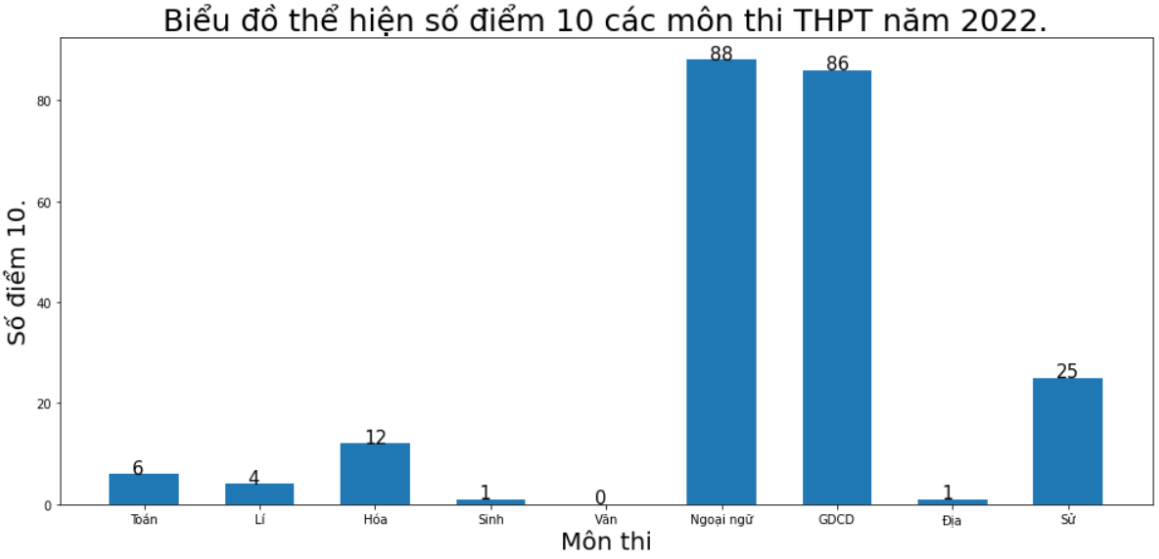
Điểm trung bình các môn thi không có sự chênh lệch quá nhiều tuy nhiên giữa môn GDCD và Sinh lại có sự chênh lệch khá nhiều cho thấy mặt bằng chung về điểm GDCD khá cao kể cả là xu hướng của học sinh lại thiên về các môn tự nhiên hơn.

Biểu đồ cột thể hiện số điểm liệt và điểm 10 các môn thi

SpecialScoreChart: tương tự như hàm Meanplot, hàm sẽ có thêm mode: maximum – điểm 10, paralysis – điểm liệt (<1).

Tạo từ điển count_score với keys là tên các môn học, values là số thí sinh đạt điểm 10 (hoặc bị điểm liệt) sau đó dùng hàm bar của thư viện matplotlib để vẽ biểu đồ cột với dữ liệu trục hoành là keys tên các môn học, trục tung là values số thi sinh các môn tương ứng





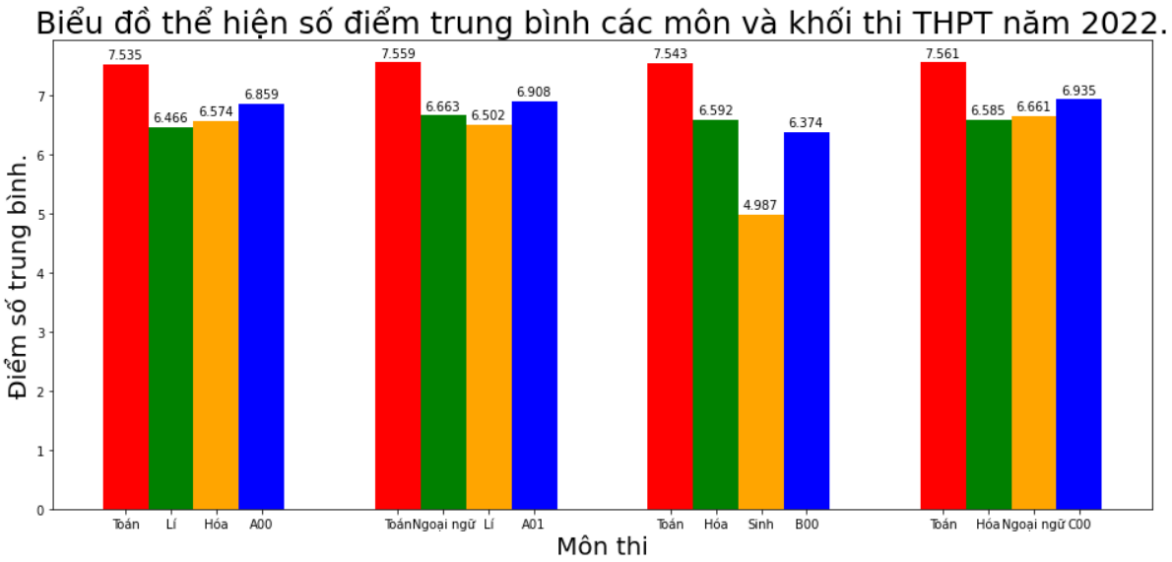
Từ 2 biểu đồ cột trên dễ thấy rằng mặc dù các môn học vẫn xuất hiện thí sinh bị điểm liệt tuy nhiên không có môn nào có quá 10 thí sinh điểm liệt ở các môn so với tổng số lượng thí sinh từng môn lên tới hàng chục ngàn.

Bên cạnh đó, số điểm 10 của các môn năm 2022 lại có sự chênh lệch khá lớn. Ở môn Ngoại Ngữ và GDCD có rất nhiều thí sinh đạt điểm 10, trái lại thì môn Văn, Sinh hay Địa gần như bất khả thi.

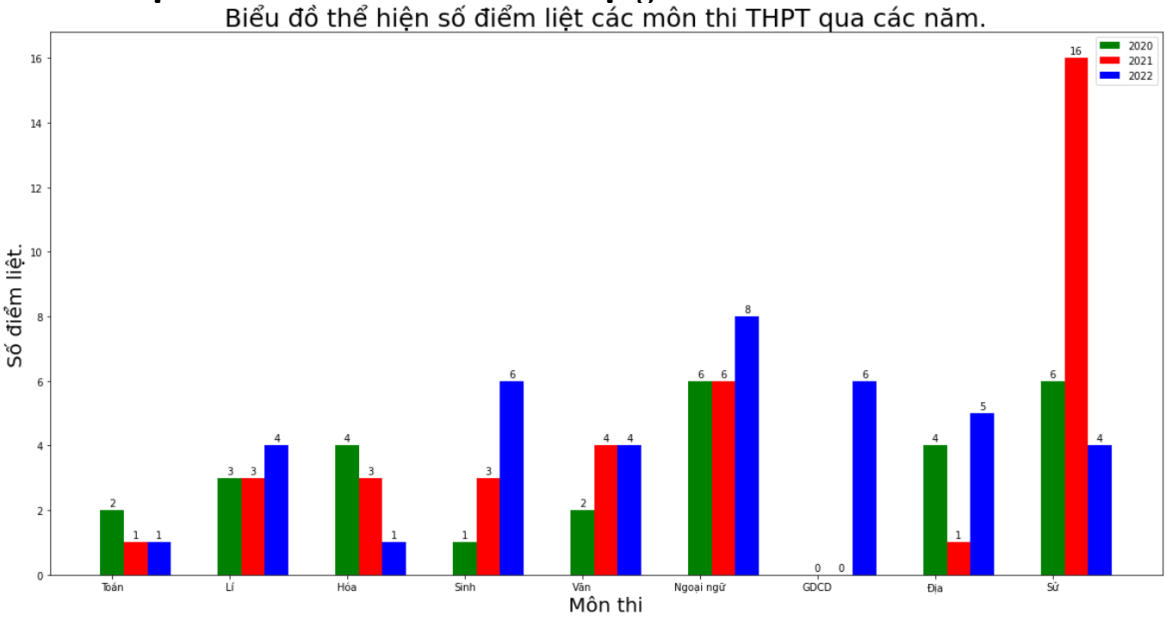
Biểu đồ cột thể hiện số điểm trung bình giữa các tổ hợp môn

MeanBlockChart: Ta có đầu vào là dataframe và blocksubject – từ điển có keys là tên tổ hợp, values là các môn học tổ hợp đó

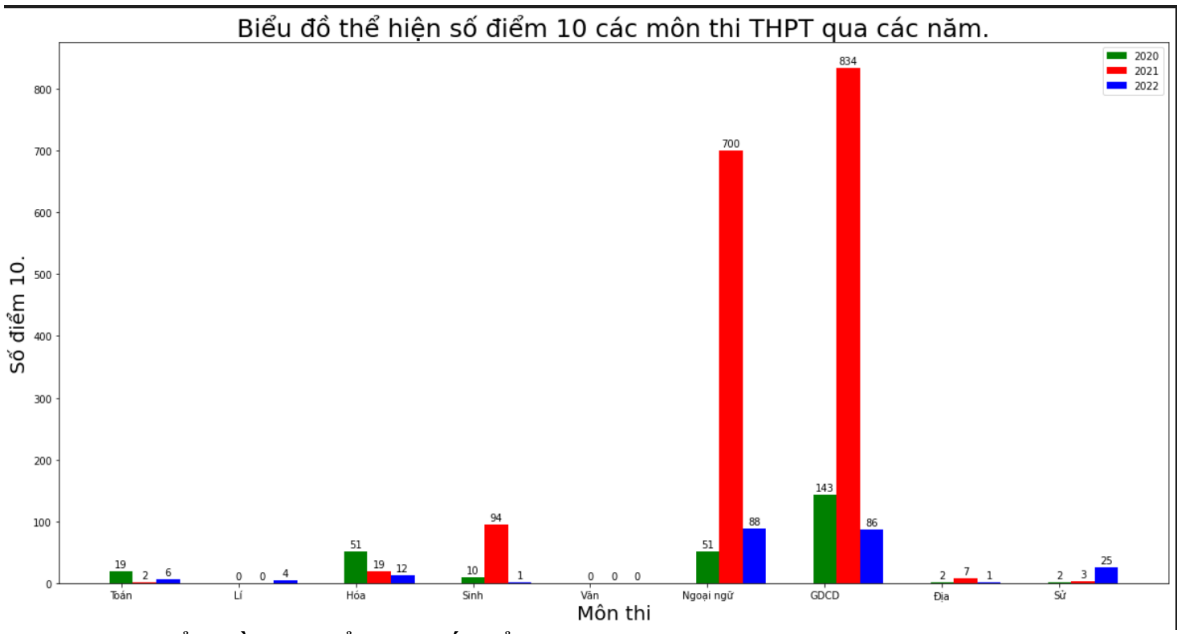
Tạo từ điển MeanBlock có keys tương tự như blocksubject nhưng values thì sẽ có 4 giá trị - 3 giá trị đầu là trung bình điểm thi của 3 môn tổ hợp, giá trị thứ 4 là trung bình điểm thi của cả 3 môn tổ hợp.



Biểu đồ cột so sánh số điểm 10 và điểm liệt giữa các năm



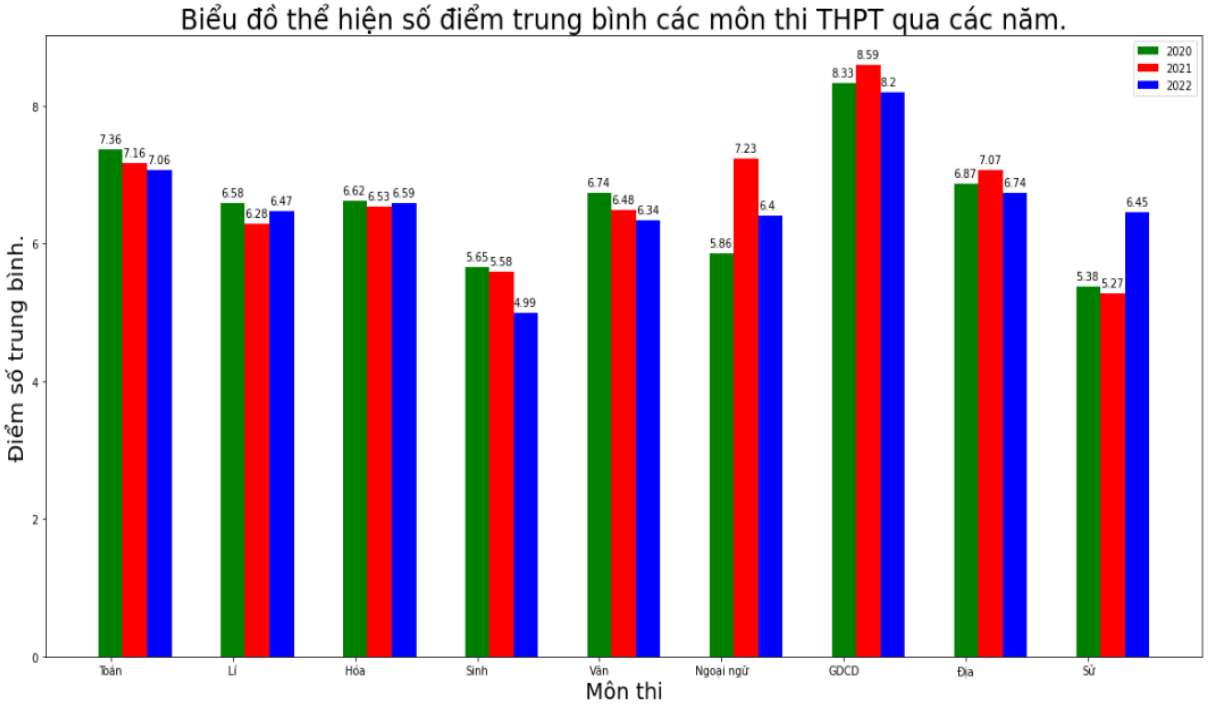
Nhìn chung số lượng điểm liệt giữa các môn không có nhiều sự khác biệt. Tuy nhiên môn Sử năm 2021 lại có số điểm liệt nhiều so với năm 2020 và 2022. Còn môn GDCD năm 2022 có 6 thí sinh bị điểm liệt trong khi 2 năm còn lại thì không xuất hiện trường hợp nào



Cũng như biểu đồ cột thể hiện số điểm 10 năm 2022 thì trong 3 năm 2020-2022 cũng không có sự khác biệt nhiều về tỉ lệ số điểm 10 giữa các môn trong 1 năm. Còn giữa các năm với nhau thì có sự khác biết khá lớn của năm 2021 so với 2 năm còn lại đặc biệt là ở môn Ngoại Ngữ, GDCD hay môn Sinh có số điểm 10 rất nhiều (khoảng 10 lần).

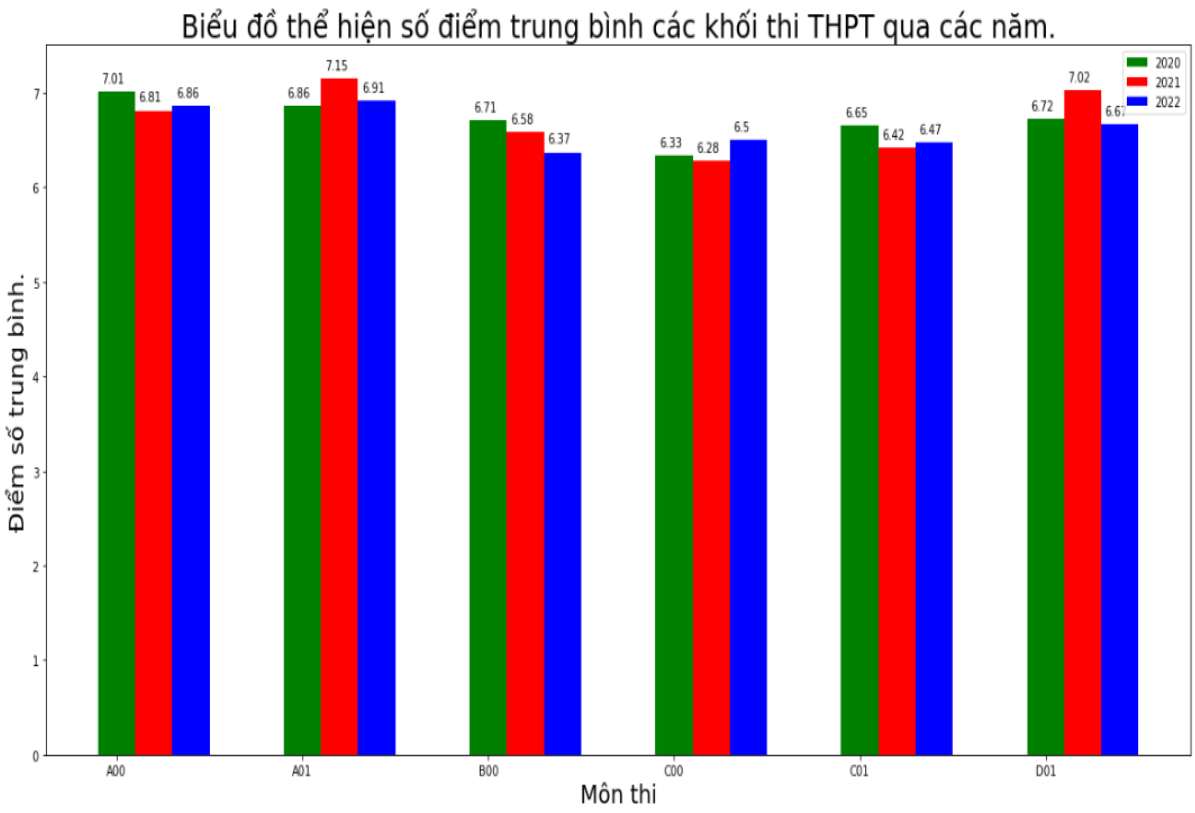
Môn Lí năm 2022 xuất hiện 4 điểm 10 trong khi năm 2020 và 2021 không có điểm 10 nào. Môn Toán và Hóa năm 2020 thì có số điểm 10 nhiều hơn. Dù vậy môn Văn thì số điểm này gần như là bất khả thi.

Biểu đồ cột so sánh điểm trung bình của các môn qua các năm:

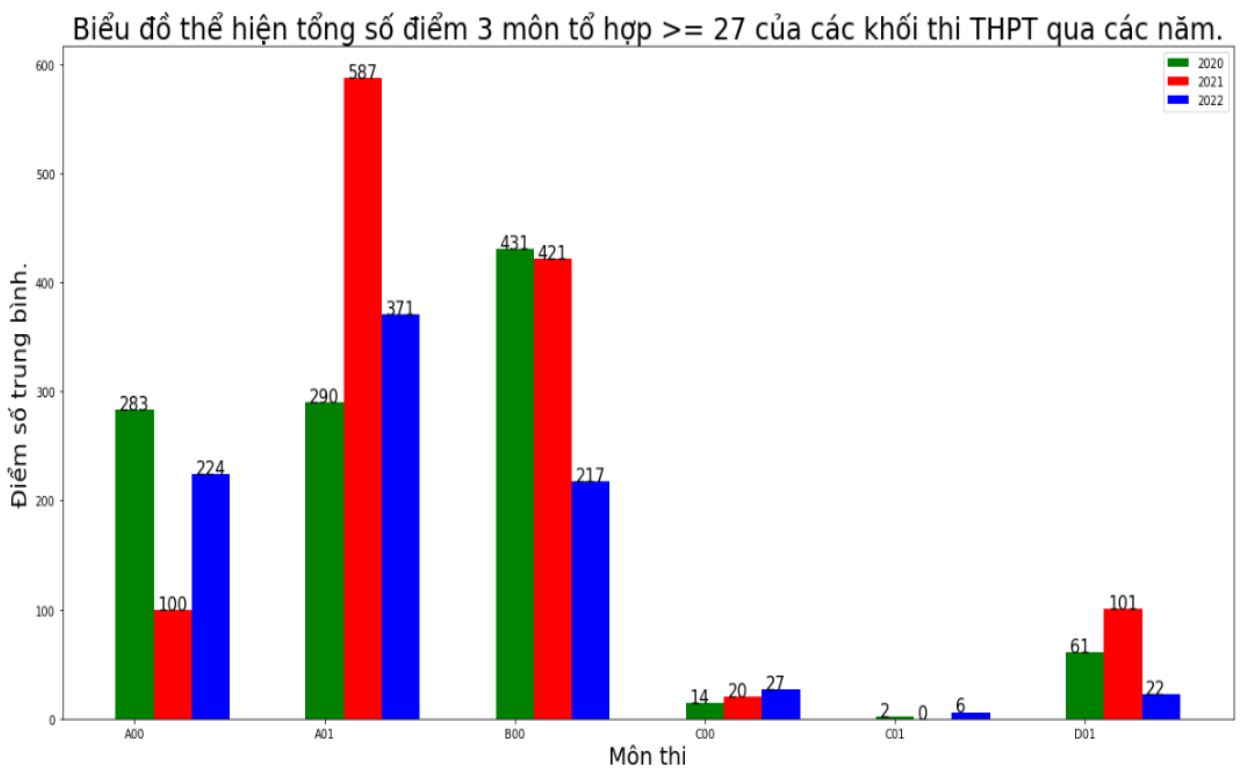


- Từ biểu đồ ta có thể đưa ra các nhận xét tổng quan về phổ điểm trung bình của các môn qua các năm. Phần nào có thể đánh giá một cách chủ quan về độ khó của đề thi.

Biểu đồ cột so sánh điểm trung bình của các khối thi qua các năm



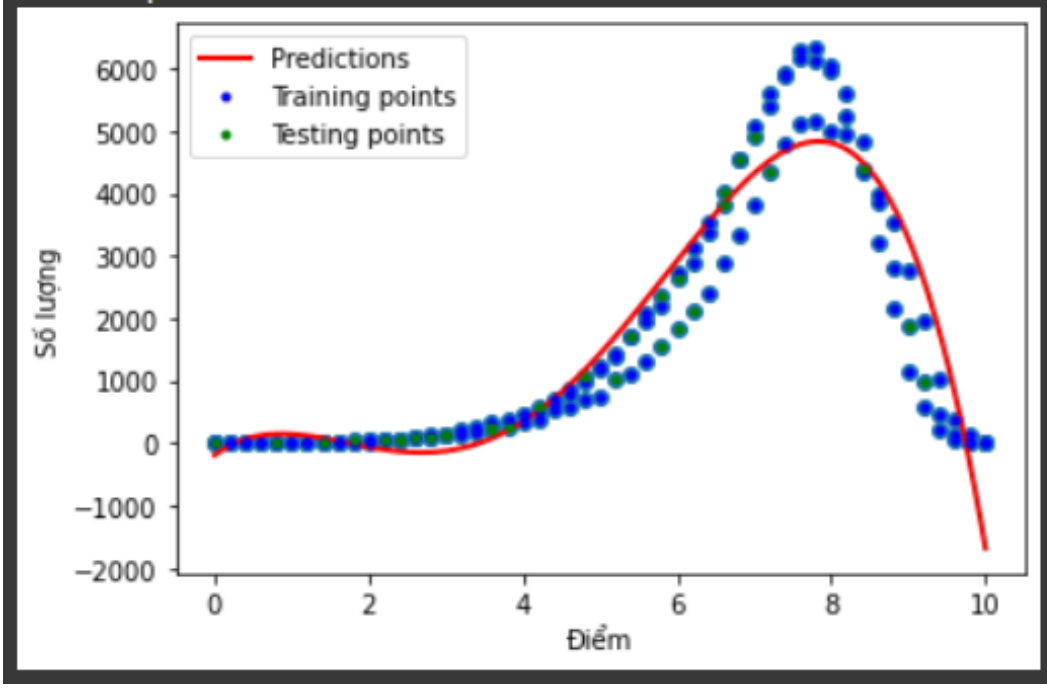
Biểu đồ cột so sánh tổng điểm 3 môn tổ hợp lớn hơn hoặc bằng 27 điểm của các khối thi qua các năm:



4. Dự đoán điểm thi trong năm 2022:

Từ dữ liệu của năm 2020,2021 ta sử dụng các hàm PolynomiaFratures, LinearRegression trong thư viện sklearn để huấn luyện và kiểm tra mô hình dự đoán điểm thi. Dựa vào quan sát ta thấy dữ liệu theo dạng phân phối chuẩn và tập dữ liệu chứa hai giá trị x, y nên khi ta sử dụng LinearRegression thì mô hình sẽ dự đoán ra hàm số hồi quy là một đường thẳng, thì nó mâu thuẫn với phân phối chuẩn cho nên lí giải cho việc sử dụng PolynomiaFratures để tạo ra một cái feature theo phân phối chuẩn. Rồi sau đó ta huấn luyện mô hình LinearRegression với giá trị x là điểm thi và y là số lượng học sinh đạt được điểm thi đó. Ta train mô hình với hàm số hồi quy là hàm bậc 4. Từ hàm số hồi quy đó ta đưa ra đồ thị sau:

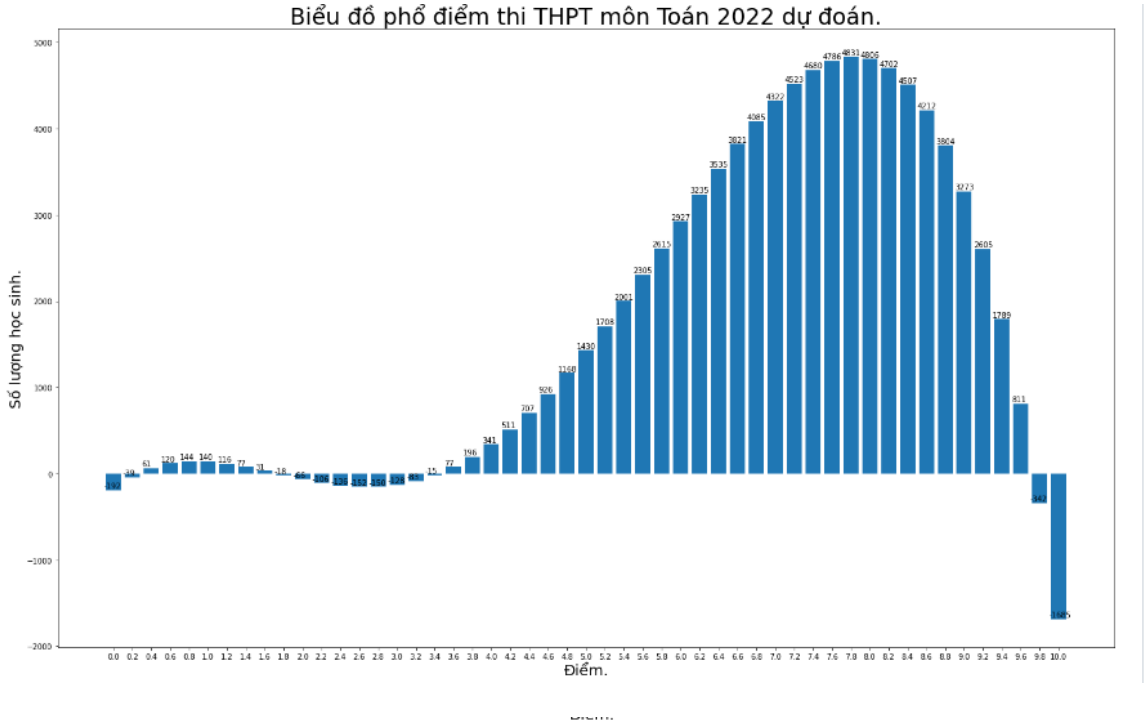
```
Intercept of model: -192.93908030908256
Hệ số hàm bậc 4
 [ 907.56621344 -750.23931177  189.00242391  -12.45471967]
R2_score = 0.8749334190763194
Mean Absolute Error: 387.4955223709709
Mean Squared Error: 326723.02579100966
```

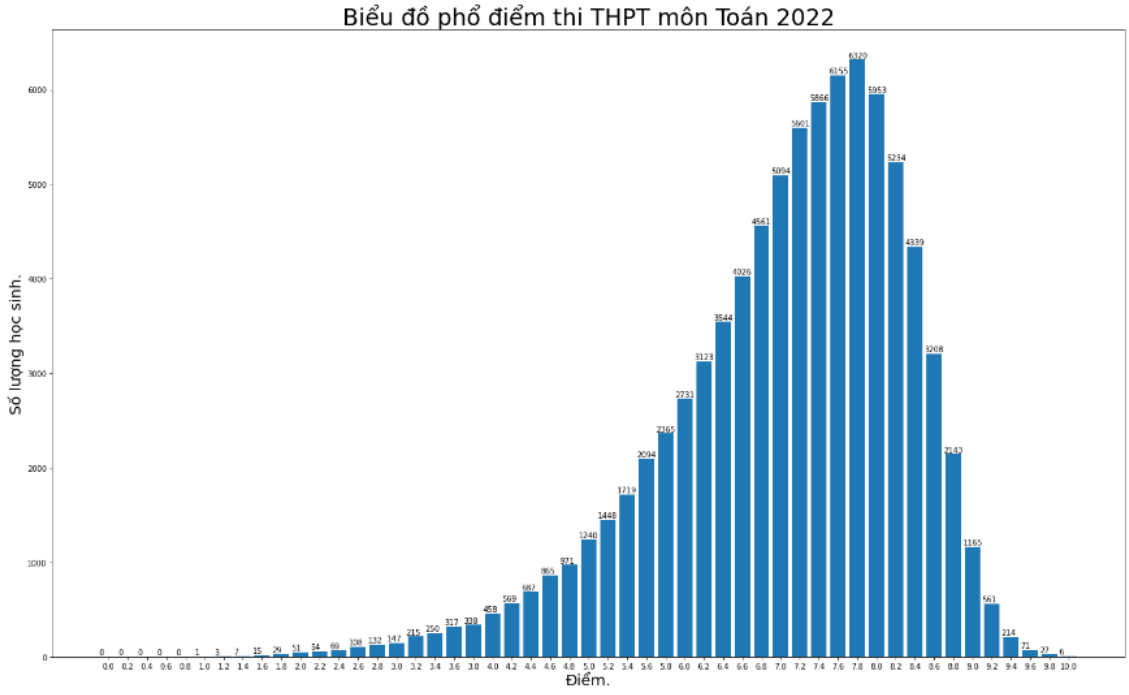


Sau khi training mô hình LinearRegression chúng ta có các hệ số (coefficient của hàm bậc 4) của mô hình

Intercept – hệ số chặn của mô hình là 1 số âm: -192.939 tuy nhiên về mặt ý nghĩ của dữ liệu khi training thì số lượng thí sinh đạt được tại điểm x thì phải luôn không âm mà hệ số chặn được hiểu như giá trị trung bình biến phản hồi khi tất cả các biến dự đoán mô hình bằng 0. Tức là ngay cả khi không có thí sinh nào đạt được số điểm đó thì Intercept = 0.

Sau đó ta đưa ra dự đoán kiểm định cho kết quả của năm 2022.





Và so sánh với kết quả thực tế:

5. Kết luận:

Từ các metric R2_score, Mean Squared Error và Mean Absolute Error: [4]

- Độ chính xác của mô hình (R2_score) dự đoán là xấp xỉ 0.87 tuy vậy có những giá trị âm thể hiện mô hình này không hoàn toàn chính xác.
- MSE – bình phương trung bình sai số = 326723.026 so với miền giá trị từ 0 – 8000 cực kì lớn do tại giá trị cận điểm 10 là âm khá nhiều trong khi thực tế dữ liệu không thể xảy ra và tại điểm 8 mô hình dự đoán khá xa dữ liệu thực dẫn đến sự chênh lệch rất lớn
- MAE – độ lớn trung bình sai số = 387.496 sự chênh lệch lớn do tương tự như MSE tuy nhiên MAE tổng hòa hơn cho tất cả giá trị dữ đoán bởi MSE khi tính toán bị ảnh hưởng bởi các outlier và sự phân bố giá trị dữ liệu trong khi MAE không quan tâm xu hướng lỗi như thế nào, metric này xử lí các lỗi như nhau.

Nguyên nhân mô hình LinearRegression không chính xác có thể là:

- Do tập dữ liệu trên chỉ để phân tích, khó có thể sử dụng để dự đoán cho năm tiếp theo.
- Điểm thi phụ thuộc vào: độ khó của đề thi, tình hình xã hội qua các năm, lượng kiến thức ôn thi, …

Bài báo cáo chỉ là quy trình cơ bản về phân tích và trực quan hóa dữ liệu. Từ đó ta có thể có cái nhìn tổng quát về bộ dữ liệu, phân tích các thành phần bên trong của dữ liệu từ đó ta đưa ra được cái insight và storytelling về điểm thi qua các năm 2020 đến 2022.

Tài liệu tham khảo

[1] A Step-by-Step Guide to the Data Analysis Process [2023] (careerfoundry.com)

[2] Tìm hiểu chung về Web Scraping và các vấn đề cần quan tâm (viblo.asia)

[3] Beautiful Soup: Build a Web Scraper With Python – Real Python

[4] 3 Best metrics to evaluate Regression Model? | by Songhao Wu | Towards Data Science