

DẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH



BÀI TẬP LỚN KHAI PHÁ DỮ LIỆU

NHÓM 19 - LỚP L02

Phân tích Điểm thi THPT Quốc gia năm 2025 ở Thành phố Hồ Chí Minh

Giảng viên: Đỗ Thanh Thái, CSE-HCMUT

Sinh viên: Võ Hữu Khang - 2211479

Tống Duy Khang - 2211467

Đỗ Hoàng Linh - 2211844

Email liên hệ nhóm: khang.vohuu@hcmut.edu.vn

TP.HỒ CHÍ MINH, THÁNG 11/2025



Mục lục

1 Giới Thiệu	4
1.1 Bối Cảnh Nghiên Cứu	4
1.2 Câu Hỏi Nghiên Cứu	4
2 Khung Lý Thuyết: Các Khái Niệm Thống Kê và Học Máy	4
2.1 ANOVA (Analysis of Variance - Phân tích phương sai)	4
2.2 Cohen's d - Effect Size (Kích Thước Hiệu Ứng)	5
2.3 Lợi Thế Đô Thị (Urban Advantage)	6
2.4 Thống Kê Mô Tả (Descriptive Statistics)	6
2.4.1 Trung Bình (Mean)	6
2.4.2 Trung Vị (Median)	6
2.4.3 Mode (Mốt)	7
2.4.4 Độ Lệch Chuẩn (Standard Deviation - SD)	7
2.4.5 Độ Lệch Tuyệt Đối Trung Vị (Median Absolute Deviation - MAD)	7
2.5 Trực Quan Hóa Dữ Liệu	7
2.5.1 Biểu Đồ Histogram và Kernel Density Estimation (KDE)	7
2.5.2 Biểu Đồ Boxplot (Box Plot)	8
2.5.3 Biểu Đồ Heatmap (Bản Đồ Nhiệt)	8
2.6 Phân Tích Tương Quan	8
2.6.1 Ma Trận Tương Quan (Pearson và Spearman)	8
2.7 Các Kiểm Định Thống Kê	9
2.7.1 Kiểm Định Post-Hoc (Ví Dụ: Tukey's HSD)	9
2.7.2 Kolmogorov-Smirnov Test (KS Test)	9
2.7.3 t-test (Kiểm Định t)	9
2.8 Mô Hình Hóa Dự Báo (Predictive Modeling)	10
2.8.1 Hồi Quy Tuyến Tính (Linear Regression)	10
2.8.2 Mô Hình Ensemble Nâng Cao (Random Forest & Gradient Boosting)	10
2.9 Dánh Giá và Diễn Giải Mô Hình	10
2.9.1 K-Fold Cross Validation (Kiểm Định Chéo K-Fold)	10
2.9.2 R ² (R-squared - Hệ số xác định)	11
2.9.3 RMSE (Root Mean Squared Error - Sai số toàn phương trung bình gốc)	11
2.9.4 MAE (Mean Absolute Error - Sai số tuyệt đối trung bình)	11
2.9.5 Diễn Giải Mô Hình (XAI - Explainable AI)	11
2.9.6 Kỹ Thuật SHAP (SHapley Additive exPlanations)	11
3 Phương Pháp Nghiên Cứu	12
3.1 Dữ Liệu và Chuẩn Bị	12
3.1.1 Tập Dữ Liệu Gốc	12
3.1.2 Làm sạch dữ liệu	12
3.1.3 Phát hiện dữ liệu	12
3.1.4 Làm Giàu Dữ Liệu (Feature Engineering)	12
3.2 Các Hướng Phân Tích	13
4 Thực hiện	13
4.1 Hướng 1: Phân tích Phổ điểm, Ảnh hưởng Nhóm tuổi và Lợi thế Đô thị	13
4.1.1 Phương pháp luận	14



4.1.2	Kết quả Phân tích	14
4.1.2.a	Thống kê mô tả điểm thi TPHCM so với Toàn quốc	14
4.1.2.b	Kiểm định thống kê	15
4.1.2.c	Phân tích Lợi thế Đô thị (Urban Advantage)	16
4.1.2.d	Phân tích Tổ hợp 3 Môn	16
4.1.3	Nhận xét	17
4.1.4	Kết luận và Quyết định	22
4.2	Hướng 2: Phân tích Xu hướng Học tập (Cohen's d) và Mô hình hóa Dự báo	22
4.2.1	Phương pháp luận	23
4.2.2	Kết quả Phân tích	23
4.2.2.a	Phân tích Xu hướng học (Cohen's d)	23
4.2.2.b	So sánh chi tiết nhóm 'Thiên Văn' và 'Thiên Toán'	25
4.2.2.c	Phân tích Tương quan Môn học	27
4.2.2.d	Mô hình hóa Dự báo Tổng điểm	28
4.2.2.e	Điều giải Mô hình (XAI) với SHAP	29
4.2.3	Nhận xét và Quyết định (Decision Making)	30
4.3	Hướng 3: Phân tích Hiệu ứng Tuổi Tương đối (Relative Age Effect)	31
4.3.1	Phương pháp luận	32
4.3.2	Kết quả Phân tích	32
4.3.2.a	Thống kê mô tả theo Quý sinh	32
4.3.2.b	Kết quả Kiểm định Giả thuyết (ANOVA)	33
4.3.2.c	Trực quan hóa Phân bố điểm	33
4.3.3	Kết luận và Nhận xét	35
4.4	Hướng 4: Phân tích Khám phá: Cung Hoàng Đạo và Kết quả thi	36
4.4.1	Phương pháp luận	36
4.4.2	Kết quả Phân tích	36
4.4.2.a	Thống kê mô tả theo Cung Hoàng Đạo	36
4.4.2.b	Kết quả Kiểm định Giả thuyết (ANOVA)	37
4.4.2.c	Trực quan hóa Phân bố điểm	37
4.4.3	Kết luận và Nhận xét	39
5	Kết Luận Chung	39



Danh sách thành viên & Phân chia công việc

STT	Họ và tên	MSSV	Lớp	Phân công
1	Võ Hữu Khang	2211479	L02	Hướng 1 và 2
2	Tống Duy Khang	2211467	L02	Data Cleaning và Hướng 3
3	Đỗ Hoàng Linh	2211844	L02	Featuring Engineering và Hướng 4



1 Giới Thiệu

1.1 Bối Cảnh Nghiên Cứu

Kỳ thi Tốt nghiệp Trung học Phổ thông (THPT) quốc gia năm 2025 đánh dấu một bước ngoặt quan trọng trong lịch sử giáo dục Việt Nam. Lần đầu tiên, kỳ thi áp dụng chương trình giáo dục phổ thông mới (Chương trình giáo dục phổ thông 2018) với những thay đổi cơ bản:

- Giảm số môn thi:** Chỉ còn 4 môn (Toán, Ngữ văn là 2 môn bắt buộc và 2 môn tổ hợp tự chọn) thay vì 6 môn trước đây, nhằm giảm áp lực cho thí sinh
- Thay đổi cấu trúc đề thi:** Tất cả môn đều có sự thay đổi ít nhiều. Đặc biệt môn Ngữ văn chuyển từ đánh giá kiến thức ghi nhớ sang đánh giá năng lực tư duy, sáng tạo
- Tăng tỷ trọng xét tuyển:** Điểm học bạ chiếm 50% trong xét công nhận tốt nghiệp
- Đề thi khoa học hơn:** Tập trung vào năng lực thực tiễn và giải quyết vấn đề

Những thay đổi này tạo ra một bối cảnh đặc biệt để nghiên cứu sự thích ứng của học sinh TP. Hồ Chí Minh (TPHCM), với vị thế là trung tâm kinh tế-giáo dục hàng đầu, là một trường hợp điển hình để phân tích “lợi thế đô thị” (urban advantage).

1.2 Câu Hỏi Nghiên Cứu

Nghiên cứu này trả lời bốn câu hỏi chính:

- RQ1 (Hướng 1):** Cải cách thi năm 2025 đã tác động đến phân bố điểm của thí sinh TPHCM như thế nào so với mặt bằng cả nước? Có sự tồn tại của “lợi thế đô thị” không?
- RQ2 (Hướng 2):** Mỗi tương quan giữa các môn thi đã thay đổi ra sao? Học sinh có xu hướng học thiên về Văn hay Toán, và nhóm nào có kết quả cao hơn?
- RQ3 (Hướng 3):** Các yếu tố nhân khẩu học như độ tuổi và tháng sinh (Hiệu ứng Tuổi Tương đối - Relative Age Effect), Cung Hoàng Đạo có ảnh hưởng đến kết quả học tập hay không?
- RQ4 (Hướng 2 & 4):** Yếu tố nào (môn thi, đặc điểm cá nhân, xu hướng học tập) là quan trọng nhất trong việc dự đoán tổng điểm?

2 Khung Lý Thuyết: Các Khái Niệm Thống Kê và Học Máy

2.1 ANOVA (Analysis of Variance - Phân tích phương sai)

ANOVA là một kiểm định thống kê dùng để so sánh giá trị trung bình của **ba nhóm trở lên**, nhằm xác định xem có ít nhất một nhóm khác biệt đáng kể so với các nhóm còn lại hay không.

Giả thuyết kiểm định:

- Giả thuyết null (H_0):** Trung bình của tất cả các nhóm là bằng nhau ($\mu_1 = \mu_2 = \dots = \mu_k$).
- Giả thuyết đối (H_a):** Có ít nhất một cặp trung bình nhóm khác nhau.

Công thức F-statistic: Chỉ số F được tính bằng tỷ lệ giữa phương sai *giữa các nhóm* (sự thay đổi giải thích được) và phương sai *trong nội bộ từng nhóm* (sự thay đổi không giải thích được).

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{\sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2 / (k-1)}{\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 / (N-k)} \quad (1)$$

trong đó:

- k = số nhóm so sánh



- n_i = kích thước (số quan sát) của nhóm i
- \bar{X}_i = trung bình của nhóm i
- \bar{X} = trung bình tổng thể của tất cả các quan sát
- N = tổng số quan sát ($N = \sum n_i$)
- MS_{between} : Mean Square Between (Trung bình bình phương giữa các nhóm), tử số.
- MS_{within} : Mean Square Within (Trung bình bình phương nội bộ nhóm), mẫu số.

Điễn giải:

- Nếu giá trị **F lớn và p-value < 0.05** (hoặc mức ý nghĩa α đã chọn), chúng ta bác bỏ giả thuyết H_0 . Điều này có nghĩa là có bằng chứng thống kê cho thấy **tồn tại sự khác biệt đáng kể** về giá trị trung bình của ít nhất một nhóm so với các nhóm khác.
- **Lưu ý:** ANOVA không cho biết *cặp nhóm cụ thể nào* khác nhau. Để xác định điều này, cần thực hiện kiểm định Post-Hoc.

Các giả định quan trọng của ANOVA:

1. **Độc lập:** Các quan sát trong mỗi nhóm và giữa các nhóm phải độc lập với nhau.
2. **Phân phối chuẩn:** Dữ liệu (hoặc phần dư) trong mỗi nhóm phải tuân theo phân phối chuẩn.
3. **Đồng nhất phương sai (Homoscedasticity):** Phương sai của các nhóm phải tương đương nhau (có thể kiểm tra bằng Levene's Test).

Ứng dụng trong nghiên cứu:

- So sánh điểm trung bình giữa các nhóm tuổi (18, >18, thi lại) để xem tuổi tác có ảnh hưởng đến kết quả không.
- So sánh điểm trung bình giữa các tổ hợp môn (A, A1, B, D) để xác định tổ hợp nào có điểm trung bình cao/thấp hơn.
- So sánh điểm trung bình giữa các xu hướng học tập (lệch tự nhiên, lệch xã hội, cân bằng).

2.2 Cohen's d - Effect Size (Kích Thước Hiệu Ứng)

Trong khi p-value cho biết liệu sự khác biệt có ý nghĩa thống kê hay không, Cohen's d đo lường *mức độ lớn* (magnitude) của sự khác biệt đó. Đây là một thước đo tiêu chuẩn hóa, độc lập với kích thước mẫu.

Công thức (cho hai nhóm độc lập):

$$d = \frac{M_1 - M_2}{SD_{\text{pooled}}} \quad (2)$$

trong đó SD_{pooled} (Độ lệch chuẩn gộp) được tính bằng:

$$SD_{\text{pooled}} = \sqrt{\frac{(n_1 - 1) \cdot SD_1^2 + (n_2 - 1) \cdot SD_2^2}{n_1 + n_2 - 2}} \quad (3)$$

- M_1, M_2 : Trung bình của nhóm 1 và nhóm 2.
- SD_1, SD_2 : Độ lệch chuẩn của nhóm 1 và nhóm 2.
- n_1, n_2 : Kích thước mẫu của nhóm 1 và nhóm 2.



Điễn giải (theo Cohen (1988)): Giá trị d cho biết sự khác biệt giữa hai trung bình là bao nhiêu đơn vị *độ lệch chuẩn*.

- $|d| < 0.2$: Kích thước hiệu ứng rất nhỏ (không đáng kể).
- $0.2 \leq |d| < 0.5$: Kích thước hiệu ứng nhỏ.
- $0.5 \leq |d| < 0.8$: Kích thước hiệu ứng trung bình.
- $|d| \geq 0.8$: Kích thước hiệu ứng lớn.

Ứng dụng:

- Trong nghiên cứu này, Cohen's d được dùng để **định lượng và phân loại** xu hướng học tập. Bằng cách tính d cho chênh lệch điểm Văn - Toán (đã chuẩn hóa), chúng ta có thể phân loại mức độ chênh lệch từ "cân bằng hoàn toàn" ($|d| < 0.05$) đến "lệch mạnh" ($|d| \geq 0.8$) một cách khách quan.

2.3 Lợi Thế Đô Thị (Urban Advantage)

Lợi thế đô thị là một khái niệm đề cập đến sự chênh lệch có hệ thống về cơ hội, tài nguyên, và kết quả phát triển (như giáo dục, thu nhập, sức khỏe) giữa khu vực đô thị (thành thị) và khu vực nông thôn.

Mô hình hóa trong thống kê: Trong phân tích định lượng, "lợi thế đô thị" thường được đưa vào mô hình (ví dụ: hồi quy) dưới dạng một **biến giả (dummy variable)**.

Điễn giải: Một hệ số dương và có ý nghĩa thống kê cho biến KhuVuc sẽ khẳng định sự tồn tại của "lợi thế đô thị" trong bối cảnh dữ liệu đang xét, cho thấy học sinh thành thị có kết quả trung bình cao hơn.

Ứng dụng thực tế:

- Phân tích chênh lệch thu nhập hoặc cơ hội việc làm giữa cư dân đô thị và nông thôn.
- Dánh giá tác động của chính sách đô thị hóa đến chất lượng cuộc sống.
- Trong giáo dục: Xác định và đo lường khoảng cách về điểm số hoặc tỷ lệ đỗ đại học giữa học sinh thành phố và nông thôn.

2.4 Thống Kê Mô Tả (Descriptive Statistics)

2.4.1 Trung Bình (Mean)

Trung bình cộng là thước đo **xu hướng trung tâm** phổ biến nhất, đại diện cho giá trị "điển hình" của tập dữ liệu.

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad (4)$$

- **Điễn giải:** Đây là điểm cân bằng của dữ liệu. Nó rất hữu ích khi dữ liệu có phân phối đối xứng (ví dụ: phân phối chuẩn).
- **Hạn chế:** Trung bình **rất nhạy cảm** với các giá trị ngoại lệ (outliers). Một giá trị quá lớn hoặc quá nhỏ có thể kéo trung bình lệch về phía nó.
- **Ứng dụng:** Tính điểm trung bình môn học, thu nhập trung bình.

2.4.2 Trung Vị (Median)

Trung vị là giá trị nằm chính giữa của tập dữ liệu đã được sắp xếp theo thứ tự. Nó là **phân vị thứ 50 (50th percentile)**.

- **Điễn giải:** 50% dữ liệu có giá trị nhỏ hơn hoặc bằng trung vị, và 50% dữ liệu có giá trị lớn hơn hoặc bằng trung vị.

- **Ưu điểm:** Trung vị là thước đo xu hướng trung tâm **bền vững (robust)**, nghĩa là nó **không bị ảnh hưởng** (hoặc ít bị ảnh hưởng) bởi các giá trị ngoại lệ.
- **Ứng dụng:** Phù hợp cho dữ liệu bị lệch (skewed data) như giá nhà, thu nhập (nơi có một vài giá trị rất cao).

2.4.3 Mode (Mốt)

Mode là giá trị xuất hiện với **tần suất cao nhất** trong tập dữ liệu. Một tập dữ liệu có thể không có mode, có một mode (unimodal), hoặc nhiều mode (bimodal, multimodal).

- **Điễn giải:** Đại diện cho giá trị phổ biến nhất.
- **Ưu điểm:** Đây là thước đo xu hướng trung tâm duy nhất có thể dùng cho **dữ liệu định tính (categorical data)**.
- **Ứng dụng:** Xác định sản phẩm bán chạy nhất, màu sắc áo được ưa chuộng nhất.

2.4.4 Độ Lệch Chuẩn (Standard Deviation - SD)

Độ lệch chuẩn là thước đo **mức độ phân tán (dispersion)** của dữ liệu xung quanh giá trị trung bình.

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{N}} \quad (\text{cho quần thể}) \quad (5)$$

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} \quad (\text{cho mẫu}) \quad (6)$$

- **Điễn giải:** Giá trị SD càng lớn, dữ liệu càng phân tán rộng so với trung bình. Nếu SD nhỏ, các giá trị có xu hướng cụm lại gần trung bình.
- Trong phân phối chuẩn, quy tắc 68-95-99.7 cho biết: khoảng 68% dữ liệu nằm trong ± 1 SD, 95% trong ± 2 SD, và 99.7% trong ± 3 SD so với trung bình.
- **Hạn chế:** Giống như trung bình, SD nhạy cảm với giá trị ngoại lệ.

2.4.5 Độ Lệch Tuyệt Đối Trung Vị (Median Absolute Deviation - MAD)

MAD là một thước đo mức độ phân tán **bền vững (robust)**, dựa trên trung vị.

$$\text{MAD} = \text{median}(|X_i - \text{median}(X)|) \quad (7)$$

- **Điễn giải:** Đây là "trung vị" của khoảng cách từ mỗi điểm dữ liệu đến "trung vị" của toàn bộ dữ liệu. Nó là giải pháp thay thế cho độ lệch chuẩn khi dữ liệu có nhiều ngoại lệ hoặc không phân phối chuẩn.
- **Ứng dụng:** Thường dùng trong phát hiện ngoại lệ (outlier detection) vì nó không bị ảnh hưởng bởi chính các ngoại lệ đó.

2.5 Trực Quan Hóa Dữ Liệu

2.5.1 Biểu Đồ Histogram và Kernel Density Estimation (KDE)

Cả hai đều dùng để **trực quan hóa hình dạng phân phối** của một biến liên tục.

- **Histogram:** Chia dữ liệu thành các khoảng (bins) và đếm số lượng quan sát trong mỗi khoảng, thể hiện bằng các cột. Hình dạng của histogram phụ thuộc vào cách chọn số lượng bins.
- **KDE:** Là phiên bản "mượt mà" hơn của histogram. Nó ước tính hàm mật độ xác suất liên tục của dữ liệu, cho cái nhìn rõ hơn về hình dạng, tính đối xứng (skewness) và số lượng đỉnh (modality) của phân phối.



2.5.2 Biểu Đồ Boxplot (Box Plot)

Boxplot (biểu đồ hộp) tóm tắt phân phối dữ liệu qua 5 thông số (Ngũ vị):

1. Giá trị tối thiểu (Min)
2. Tứ phân vị dưới Q1 (25%)
3. Trung vị Q2 (50%)
4. Tứ phân vị trên Q3 (75%)
5. Giá trị tối đa (Max)

- **IQR (Interquartile Range):** Khoảng giữa Q3 và Q1 ($IQR = Q3 - Q1$), chứa 50% dữ liệu ở giữa.
- **Điễn giải:** Boxplot rất hiệu quả trong việc **so sánh nhanh phân phối** giữa nhiều nhóm và **phát hiện các giá trị ngoại lệ** (outliers) - thường là các điểm nằm ngoài khoảng $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$.

2.5.3 Biểu Đồ Heatmap (Bản Đồ Nhiệt)

Heatmap là một biểu đồ sử dụng màu sắc để biểu thị các giá trị trong một ma trận 2D.

- **Điễn giải:** Cường độ màu (ví dụ: từ nhạt đến đậm) tương ứng với giá trị (từ thấp đến cao).
- **Ứng dụng:** Thường dùng để trực quan hóa **ma trận tương quan**, giúp nhanh chóng phát hiện các cặp biến có tương quan mạnh (âm hoặc dương).

2.6 Phân Tích Tương Quan

2.6.1 Ma Trận Tương Quan (Pearson và Spearman)

Ma trận tương quan đo lường mức độ và chiều hướng của mối quan hệ giữa các cặp biến.

- **Hệ số tương quan Pearson (r):**

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \quad (8)$$

- Đo lường mối quan hệ **tuyến tính** (đường thẳng).
- Giá trị từ -1 (tương quan âm hoàn hảo) đến +1 (tương quan dương hoàn hảo). 0 nghĩa là không có tương quan **tuyến tính**.
- Yêu cầu dữ liệu phải có phân phối chuẩn (hoặc gần chuẩn).

- **Hệ số tương quan Spearman (ρ):**

- Đo lường mối quan hệ **đơn điệu** (không nhất thiết phải là đường thẳng) dựa trên **thứ hạng** (rank) của dữ liệu.
- Đây là phương pháp phi tham số, không yêu cầu phân phối chuẩn và bền vững hơn với ngoại lệ.

CẢNH BÁO QUAN TRỌNG: **Tương quan không đồng nghĩa với quan hệ nhân quả (Correlation does not imply causation).** Chỉ vì hai biến di chuyển cùng nhau không có nghĩa là biến này gây ra biến kia.



2.7 Các Kiểm Định Thống Kê

2.7.1 Kiểm Định Post-Hoc (Ví Dụ: Tukey's HSD)

- **Mục đích:** Khi ANOVA trả về kết quả có ý nghĩa ($p < 0.05$), nó chỉ cho biết "có ít nhất một nhóm khác biệt". Kiểm định Post-Hoc (kiểm định "sau đó") được dùng để thực hiện **so sánh cặp đôi** tất cả các nhóm để xác định *cụ thể* nhóm nào khác biệt với nhóm nào.
- **Tại sao cần thiết:** Nếu chạy nhiều kiểm định t-test cho từng cặp, nguy cơ mắc sai lầm Loại I (bắc bỏ H_0 trong khi H_0 đúng) sẽ tăng lên. Các kiểm định Post-Hoc như Tukey's HSD (Honestly Significant Difference) điều chỉnh p-value để kiểm soát tỷ lệ lỗi này.
- **Ứng dụng:** Sau khi chạy ANOVA cho các tổ hợp môn, nếu $p < 0.05$, dùng Tukey HSD để xem điểm tổ hợp A có khác D không, A có khác B không, v.v.

2.7.2 Kolmogorov-Smirnov Test (KS Test)

KS Test là một kiểm định phi tham số dùng để so sánh các phân phối.

- **Loại 1 (One-Sample KS Test):** Kiểm tra xem dữ liệu của một mẫu có tuân theo một phân phối lý thuyết cụ thể hay không (ví dụ: phân phối chuẩn).
- **Loại 2 (Two-Sample KS Test):** Kiểm tra xem hai mẫu dữ liệu có đến từ cùng một phân phối hay không.

$$D = \sup_x |F_1(x) - F_2(x)| \quad (9)$$

(Thống kê D là khoảng cách tối đa giữa hai hàm phân phối tích lũy $F(x)$ của hai mẫu).

- **Điễn giải:** Nếu p-value < 0.05 , chúng ta bác bỏ giả thuyết null, kết luận rằng hai phân phối là khác nhau.

2.7.3 t-test (Kiểm Định t)

t-test được dùng để so sánh trung bình của **hai nhóm**.

- **Giả thuyết kiểm định:**

- H_0 : Trung bình hai nhóm bằng nhau ($\mu_1 = \mu_2$).
- H_a : Trung bình hai nhóm khác nhau ($\mu_1 \neq \mu_2$).

- **Các loại t-test:**

- **Independent Samples t-test (Kiểm định 2 mẫu độc lập):** So sánh trung bình của hai nhóm hoàn toàn riêng biệt (ví dụ: nam vs. nữ, đô thị vs. nông thôn).

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (\text{Công thức Welch's t-test}) \quad (10)$$

- **Paired Samples t-test (Kiểm định mẫu cặp):** So sánh trung bình của cùng một nhóm tại hai thời điểm khác nhau (ví dụ: điểm trước và sau khóa học).

- **Giả định:** Dữ liệu (hoặc chênh lệch) phải phân phối chuẩn.
- **Điễn giải:** p-value < 0.05 chỉ ra rằng sự khác biệt trung bình giữa hai nhóm có ý nghĩa thống kê.

2.8 Mô Hình Hóa Dự Báo (Predictive Modeling)

2.8.1 Hồi Quy Tuyến Tính (Linear Regression)

Hồi quy tuyến tính là mô hình cơ bản dùng để **mô hình hóa mối quan hệ tuyến tính** và **dự đoán** một biến phụ thuộc (liên tục) Y dựa trên một hoặc nhiều biến độc lập X .

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \quad (11)$$

- **Điễn giải:**

- β_0 (Hệ số chặn): Giá trị dự đoán của Y khi tất cả X bằng 0.
- β_i (Hệ số): Mức thay đổi trung bình trong Y khi X_i **tăng 1 đơn vị**, với điều kiện giữ nguyên các biến X khác.

- **Giả định chính (L.I.N.E):**

1. **Linearity:** Mối quan hệ giữa X và Y là tuyến tính.
2. **Independence:** Các phần dư (ϵ) độc lập với nhau.
3. **Normality:** Phần dư (ϵ) phân phối chuẩn.
4. **Equal Variance (Homoscedasticity):** Phương sai của phần dư là không đổi.

2.8.2 Mô Hình Ensemble Nâng Cao (Random Forest & Gradient Boosting)

Đây là các mô hình học máy mạnh mẽ, phi tuyến tính, thường cho độ chính xác cao.

- **Random Forest (Rừng Ngẫu Nhiên):**

- Là một thuật toán tập hợp (ensemble) dựa trên **Bagging**.
- Nó xây dựng **nhiều cây quyết định (decision trees)** một cách độc lập.
- Dự đoán cuối cùng được lấy bằng cách **lấy trung bình** (cho hồi quy).
- **Ưu điểm:** Rất hiệu quả, xử lý tốt dữ liệu phi tuyến, chống overfitting (quá khớp) cao, có thể tính toán độ quan trọng của biến (feature importance).

- **Gradient Boosting (Tăng Cường Gradient):**

- Là một thuật toán ensemble dựa trên **Boosting**.
- Nó xây dựng các cây quyết định một cách **tuần tự**.
- Mỗi cây mới được huấn luyện để **sửa lỗi** mà cây trước đó đã mắc phải.
- **Ưu điểm:** Thường cho độ chính xác rất cao.
- **Nhược điểm:** Nhạy cảm hơn với nhiễu và dễ bị overfitting nếu không được tinh chỉnh.

Ứng dụng: Sử dụng Random Forest/Gradient Boosting để dự báo tổng điểm (biến Y) dựa trên các biến đầu vào như điểm 9 môn học và nhóm tuổi (biến X).

2.9 Đánh Giá và Diễn Giải Mô Hình

2.9.1 K-Fold Cross Validation (Kiểm Định Chéo K-Fold)

Đây là kỹ thuật để đánh giá hiệu suất của mô hình một cách **bền vững** và **tránh overfitting**.

- **Cách thức:** Dữ liệu được chia thành K "phân" (folds) bằng nhau.
- Mô hình được huấn luyện K lần. Mỗi lần, K-1 phần được dùng làm dữ liệu huấn luyện (train) và 1 phần còn lại được dùng làm dữ liệu kiểm tra (validation).
- Hiệu suất cuối cùng của mô hình (ví dụ: RMSE) là **trung bình** của K lần chạy đó.
- **K=10:** Là lựa chọn phổ biến, cân bằng giữa độ tin cậy của ước lượng và chi phí tính toán.



2.9.2 R² (R-squared - Hệ số xác định)

R² đo lường **tỷ lệ phần trăm phương sai** của biến phụ thuộc (Y) mà mô hình có thể giải thích được.

$$R^2 = 1 - \frac{SS_{\text{residual}}}{SS_{\text{total}}} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} \quad (12)$$

- **Điễn giải:** R² có giá trị từ 0 đến 1. R² = 0.75 có nghĩa là mô hình giải thích được 75% sự biến động của Y.
- **Hạn chế:** Nên sử dụng **Adjusted R²** (**R² hiệu chỉnh**) vì nó phạt mô hình nếu thêm các biến không cần thiết.

2.9.3 RMSE (Root Mean Squared Error - Sai số toàn phương trung bình gốc)

RMSE là độ lệch chuẩn của các sai số dự đoán (phần dư).

$$\text{RMSE} = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{n}} \quad (13)$$

- **Điễn giải:** RMSE có **cùng đơn vị** với biến phụ thuộc Y. Giá trị RMSE càng nhỏ, mô hình dự đoán càng chính xác.
- **Đặc điểm:** Do bình phương sai số, RMSE **phạt rất nặng** các lỗi dự đoán lớn (outliers).

2.9.4 MAE (Mean Absolute Error - Sai số tuyệt đối trung bình)

MAE là trung bình của giá trị tuyệt đối của các sai số.

$$\text{MAE} = \frac{\sum|Y_i - \hat{Y}_i|}{n} \quad (14)$$

- **Điễn giải:** Cũng có cùng đơn vị với Y. MAE = 1.5 có nghĩa là, trung bình, dự đoán của mô hình lệch 1.5 điểm so với giá trị thực tế.
- **So với RMSE:** MAE **bền vững hơn** với các ngoại lệ vì nó không bình phương sai số.

2.9.5 Diễn Giải Mô Hình (XAI - Explainable AI)

XAI là các kỹ thuật giúp con người hiểu được tại sao một mô hình học máy (thường là "hộp đen" như Random Forest) lại đưa ra một dự đoán cụ thể.

2.9.6 Kỹ Thuật SHAP (SHapley Additive exPlanations)

SHAP là một phương pháp XAI tiên tiến dựa trên Lý thuyết trò chơi (Shapley values) để giải thích đầu ra của bất kỳ mô hình học máy nào.

$$f(x) = \phi_0 + \sum_{i=1}^M \phi_i \quad (15)$$

- **Điễn giải:** SHAP gán cho mỗi biến (feature) một giá trị ϕ_i (gọi là SHAP value) cho *từng dự đoán cụ thể*.
- ϕ_0 là giá trị dự đoán cơ sở (baseline), thường là trung bình của tất cả các dự đoán.
- ϕ_i đại diện cho **sự đóng góp** của biến i vào việc đưa dự đoán từ giá trị baseline ϕ_0 đến giá trị dự đoán cuối cùng $f(x)$.
- **Ưu điểm:** SHAP cung cấp cả **giải thích toàn cục (global)** (biến nào quan trọng nhất) và **giải thích cục bộ (local)** (tại sao mô hình dự đoán cho *thí sinh A* như vậy).
- **Ứng dụng:** Giúp trả lời RQ4 một cách minh bạch, xác định yếu tố nào (Toán, Văn, nhóm tuổi) ảnh hưởng nhiều nhất đến dự báo tổng điểm.



3 Phương Pháp Nghiên Cứu

3.1 Dữ Liệu và Chuẩn Bị

3.1.1 Tập Dữ Liệu Gốc

- Kích thước:** 129,148 thí sinh thi kỳ thi THPT 2025 tại TP.HCM
- Cột dữ liệu:** 22 cột bao gồm số báo danh, họ tên, ngày sinh, năm thi, điểm của từng môn trong 9 môn thi, tổng điểm
- Thời gian:** Năm 2025

3.1.2 Làm sạch dữ liệu

Lỗi phát hiện	Số lượng	Hành động
Giá trị ngoại lai (điểm thi không trong [0,10])	0	Kiểm tra, loại nếu lỗi
Ngày sinh không phù hợp (không đúng định dạng DD/MM/YYYY hoặc không hợp lệ)	0	Kiểm tra, nếu lỗi có cách khắc phục
Giá trị trùng lặp dựa trên Số báo danh	0	Loại bỏ nếu trùng
Hàng có >6 cột điểm thi có dữ liệu	0	Kiểm tra, loại nếu có
Kết quả	129,148	Hợp lệ

Bảng 1: Quy trình làm sạch dữ liệu

3.1.3 Phát hiện dữ liệu

- Viết đoạn code để kiểm tra hàng có $>=5$ cột điểm thi có dữ liệu:
⇒ Phát hiện dữ liệu không chỉ có những thí sinh thi đúng năm (tốt nghiệp THPT vào năm 2025), mà còn những thí sinh thi lại/thí sinh tự do thi chương trình cũ
⇒ Có thể dùng yếu tố "Nhóm tuổi" để so sánh hoặc là một yếu tố để thêm vào các mô hình ở dưới.
- Viết đoạn code để kiểm tra hàng có thí sinh thi cả nhóm 1 (Vật lí hoặc Hóa học) và nhóm 2 (Lịch sử hoặc Địa lí):
⇒ Phát hiện dữ liệu các thí sinh năm 2025 không còn phân chia "Khoa học tự nhiên"(KHTN) và "Khoa học xã hội"(KHXH) như các năm trước
⇒ Không thể dùng tổ hợp KHTN/KHXH để so sánh hoặc là một yếu tố để thêm vào các mô hình ở dưới.

3.1.4 Làm Giàu Dữ Liệu (Feature Engineering)

Các đặc trưng mới được tạo ra:

- Nhóm tuổi:** Phân loại thí sinh dựa vào năm sinh và số môn có điểm:
 - "18 tuổi": Sinh năm 2007.
 - ">18 tuổi": Sinh trước 2007, có đúng 4/9 môn thi có điểm.
 - "Thi lại": Sinh trước 2007, các trường hợp còn lại (không đủ điều kiện trên).
- Quý sinh:** Chia ngày sinh thành các quý để kiểm tra hiệu ứng tuổi tương đối:
 - Q1: Tháng 1–3
 - Q2: Tháng 4–6
 - Q3: Tháng 7–9
 - Q4: Tháng 10–12



- **Tổ hợp tự chọn:** Xác định tổ hợp môn thi dựa trên số môn có điểm và nhóm tuổi (3, 4 hoặc 6 môn).
- **Điểm trung bình bắt buộc:** $\frac{\text{Toán} + \text{Ngữ văn}}{2}$
- **Điểm trung bình tổ hợp:** Trung bình cộng các môn tự chọn (không gồm Toán, Ngữ văn).
- **Chênh lệch Văn-Toán:** Ngữ văn – Toán
- **Cung hoàng đạo:** Xác định từ ngày sinh:
 - **Bạch Dương** (Aries): 21/3 – 19/4
 - **Kim Ngưu** (Taurus): 20/4 – 20/5
 - **Song Tử** (Gemini): 21/5 – 20/6
 - **Cự Giải** (Cancer): 21/6 – 22/7
 - **Sư Tử** (Leo): 23/7 – 22/8
 - **Xử Nữ** (Virgo): 23/8 – 22/9
 - **Thiên Bình** (Libra): 23/9 – 22/10
 - **Bọ Cạp** (Scorpio): 23/10 – 21/11
 - **Nhân Mã** (Sagittarius): 22/11 – 21/12
 - **Ma Kết** (Capricorn): 22/12 – 19/1
 - **Bảo Bình** (Aquarius): 20/1 – 18/2
 - **Song Ngư** (Pisces): 19/2 – 20/3

3.2 Các Hướng Phân Tích

Bốn hướng phân tích chính được thực hiện trong báo cáo, bao gồm từ thống kê mô tả cơ bản đến mô hình hóa học máy nâng cao và phân tích khám phá.

Hướng	Các kỹ thuật chính được sử dụng	Các chỉ Số & Công cụ đánh giá
1	Thống kê Mô tả (<i>Mean, Median, Mode, Std, MAD</i>), ANOVA (Nhóm tuổi, Tổ hợp 3 môn), K-S Test , Phân tích Lợi thế Đô thị.	\bar{X} , Median, σ , MAD, F-statistic, p-value (cho ANOVA và K-S Test).
2	Phân loại xu hướng học bằng Cohen's d (Effect Size), Ma trận Tương quan, Hồi quy Random Forest/Gradient Boosting , K-Fold CV, Diễn giải mô hình bằng SHAP .	Cohen's d , r (Pearson), R^2 , RMSE, MAE, SHAP values.
3	Thống kê mô tả chi tiết theo Quý sinh , ANOVA so sánh <i>Tổng điểm</i> và <i>điểm từng môn</i> giữa các Quý sinh.	$\bar{X}_{\text{Quý sinh}}$, σ , F-statistic, p-value.
4	Thống kê mô tả chi tiết theo Cung Hoàng Đạo , ANOVA so sánh <i>Tổng điểm</i> và <i>điểm từng môn</i> giữa 12 cung hoàng đạo. (Phân tích Khám phá)	\bar{X}_{Cung} , σ , F-statistic, p-value.

Bảng 2: Tóm tắt 4 hướng phân tích chính trong báo cáo

4 Thực hiện

4.1 Hướng 1: Phân tích Phổ điểm, Ảnh hưởng Nhóm tuổi và Lợi thế Đô thị

Hướng phân tích này đóng vai trò nền tảng, giải quyết câu hỏi nghiên cứu cốt lõi: "Cải cách kỳ thi 2025 đã tác động đến phân bố điểm như thế nào và liệu có sự tồn tại của 'lợi thế đô thị' (urban advantage) hay không?"



4.1.1 Phương pháp luận

Để kiểm tra giả thuyết này, chúng tôi đã áp dụng một chuỗi các phân tích thống kê từ tệp h1.py.

- Thống kê Mô tả:** Các chỉ số xu hướng trung tâm là **Điểm trung bình (Mean)**, **Trung vị (Median)**, **Điểm nhiều thí sinh đạt được nhất (Mode)** và độ phân tán như **Độ lệch chuẩn (Standard Deviation - Std)**, **Độ lệch tuyệt đối trung vị (Median Absolute Deviation - MAD)**. Cùng với các chỉ số mà Bộ GD&ĐT công bố như **Số thí sinh điểm <5**, **Số thí sinh điểm <5 theo %**, **Số thí sinh điểm ≥7**, **Số thí sinh điểm ≥7 theo %**, **Số thí sinh đạt điểm ≤1**, **Số thí sinh đạt điểm ≤1 theo %**, **Số thí sinh đạt điểm 10**, **Số thí sinh đạt điểm 0**, **Tỉ lệ điểm 10/1000 thí sinh** đã được tính toán cho tất cả 9 môn thi. Các chỉ số này được so sánh trực tiếp với dữ liệu phổ điểm toàn quốc do Bộ GD&ĐT công bố (biến national_stats trong h1.py).
- Kiểm định Phân phối chuẩn (Kolmogorov-Smirnov):** Chúng tôi đã thực hiện kiểm định K-S (Kolmogorov-Smirnov) cho tất cả các biến và Tổng điểm. Mục đích là để xác định tính chuẩn của dữ liệu, một giả định quan trọng cho nhiều kiểm định tham số.
- Phân tích Phương sai (ANOVA):** Kiểm định F (ANOVA) một chiều được sử dụng để so sánh giá trị trung bình giữa nhiều hơn hai nhóm. Cụ thể:
 - So sánh Tổng điểm (gốc) giữa ba nhóm tuổi (18, >18, Thi lại).
 - So sánh Điểm tổ hợp 3 môn giữa các tổ hợp 3 môn khác nhau (A00, A01, B00, C00, v.v.).
- Tính toán Lợi thế Đô thị:** Lợi thế đô thị được định lượng bằng cách lấy điểm trung bình của mẫu TP.HCM trừ đi điểm trung bình toàn quốc cho từng môn học.

4.1.2 Kết quả Phân tích

4.1.2.a Thống kê mô tả điểm thi TPHCM so với Toàn quốc

Bảng dưới đây tóm tắt các chỉ số thống kê quan trọng cho 9 môn thi, so sánh giữa dữ liệu của TPHCM và Toàn quốc.

Chỉ số	Toán	Văn	Anh	Lí	Hóa	Sinh	Sử	Địa	GDCD/KTPL
Sĩ số (TPHCM)	128218	127901	62809	57611	36252	9689	32987	34962	20288
<i>Điểm Trung bình (Mean)</i>									
TPHCM	5.26	7.06	5.67	6.98	5.95	6.29	6.67	6.82	7.97
Toàn quốc	4.78	7.00	5.38	6.99	6.06	5.78	6.52	6.63	7.69
Chênh lệch	0.48	0.06	0.29	-0.01	-0.11	0.51	0.15	0.19	0.28
<i>Điểm Trung vị (Median)</i>									
TPHCM	5.25	7.25	5.50	7.00	5.75	6.25	6.75	6.85	8.00
Toàn quốc	4.60	7.25	5.25	7.00	6.00	5.75	6.60	6.75	7.75
Chênh lệch	0.65	0.00	0.25	0.00	-0.25	0.50	0.15	0.10	0.25
<i>Độ lệch chuẩn (Std)</i>									
TPHCM	1.52	1.01	1.45	1.45	1.81	1.52	1.50	1.56	1.03
Toàn quốc	1.68	1.28	1.45	1.52	1.81	1.58	1.63	1.75	1.18
Chênh lệch	-0.16	-0.27	0.00	-0.07	0.00	-0.06	-0.13	-0.19	-0.15
<i>Độ lệch tuyệt đối trung vị (MAD)</i>									
TPHCM	1.00	0.50	1.00	1.00	1.27	1.15	1.00	1.15	0.75
Toàn quốc	1.35	1.00	1.16	1.25	1.51	1.30	1.36	1.45	0.92
Chênh lệch	-0.35	-0.50	-0.16	-0.25	-0.24	-0.15	-0.36	-0.30	-0.17
<i>Điểm Mode (Mode)</i>									
TPHCM	5.25	7.50	5.25	7.50	5.75	7.25	7.00	7.75	8.25
Toàn quốc	4.75	7.50	5.25	7.50	6.25	6.50	7.25	7.75	8.25
Chênh lệch	0.50	0.00	0.00	0.00	-0.50	0.75	-0.25	0.00	0.00
<i>Số thí sinh điểm < 5</i>									



Chỉ số	Toán	Văn	Anh	Lí	Hóa	Sinh	Sử	Địa	GDCD/KTPL
TPHCM	55232	3645	19069	4780	11503	1965	4598	4497	203
Toàn quốc	635102	70308	134478	34029	70910	22674	89665	89054	6324
<i>Tỉ lệ điểm < 5 (%)</i>									
TPHCM (%)	43.077	2.850	30.360	8.297	31.731	20.281	13.939	12.863	1.001
Toàn quốc (%)	56.395	6.240	38.220	9.790	29.529	32.440	18.630	18.690	2.567
Chênh lệch (%)	-13.32	-3.39	-7.86	-1.49	2.20	-12.16	-4.69	-5.83	-1.57
<i>Số thí sinh điểm >= 7</i>									
TPHCM	19269	80402	12583	30498	11318	3468	15715	17305	17516
Toàn quốc	137741	671209	53251	186531	80847	17579	210702	215695	192613
<i>Tỉ lệ điểm >= 7 (%)</i>									
TPHCM (%)	15.028	62.863	20.034	52.938	31.220	35.793	47.640	49.497	86.337
Toàn quốc (%)	12.231	59.572	15.135	53.663	33.667	25.151	43.778	45.269	78.171
Chênh lệch (%)	2.80	3.29	4.90	-0.72	-2.45	10.64	3.86	4.23	8.17
<i>Số thí sinh đạt điểm 10</i>									
TPHCM	43	0	30	634	59	20	54	314	166
Toàn quốc	513	0	141	3929	625	82	1518	6907	1451
Tỉ lệ TPHCM/TQ (%)	8.38	N/A	21.28	16.14	9.44	24.39	3.56	4.55	11.44
<i>Tỉ lệ điểm 10/1000 thí sinh</i>									
TPHCM	0.3354	0.0000	0.4776	11.0048	1.6275	2.0642	1.6370	8.9812	8.1822
Toàn quốc	0.4555	0.0000	0.4007	11.3033	2.6027	1.1732	3.1540	14.4961	5.8888
Chênh lệch	-0.12	0.00	0.08	-0.30	-0.98	0.89	-1.52	-5.51	2.29
<i>Số thí sinh đạt điểm 0</i>									
TPHCM	2	4	2	1	2	2	2	1	1
Toàn quốc	6	7	2	1	0	0	2	3	0
Tỉ lệ TPHCM/TQ (%)	33.33	57.14	100.0	100.0	N/A	N/A	100.0	33.33	N/A
<i>Số thí sinh đạt điểm <= 1</i>									
TPHCM	28	15	5	1	5	3	3	4	1
Toàn quốc	777	87	28	3	8	1	13	19	0
<i>Tỉ lệ điểm <= 1 (%)</i>									
TPHCM (%)	0.022	0.012	0.008	0.002	0.014	0.031	0.009	0.011	0.005
Toàn quốc (%)	0.069	0.008	0.008	0.001	0.003	0.001	0.003	0.004	0.000
Chênh lệch (%)	-0.05	0.00	0.00	0.00	0.01	0.03	0.01	0.01	0.00

Bảng 3: Thống kê mô tả điểm thi TPHCM so với Toàn quốc

4.1.2.b Kiểm định thống kê

Các kiểm định ANOVA và Kolmogorov-Smirnov (K-S) đã được thực hiện để xác định các khác biệt có ý nghĩa thống kê và kiểm tra tính chuẩn của dữ liệu.



Bảng 4: Kết quả các kiểm định thống kê

Kiểm định	Chi tiết	Kết luận (tại $\alpha = 0.05$)
ANOVA	So sánh TỔNG ĐIỂM (gốc) giữa các nhóm tuổi	Có sự khác biệt có ý nghĩa thống kê ($F=4366.48, p=0$)
ANOVA	So sánh ĐIỂM TỔ HỢP 3 MÔN giữa các TỔ HỢP	Có sự khác biệt có ý nghĩa thống kê ($F=1170.95, p=0$)
K-S Test	Kiểm tra phân phối chuẩn cho môn Toán	Dữ liệu KHÔNG tuân theo phân phối chuẩn ($p=0$)
K-S Test	Kiểm tra phân phối chuẩn cho môn Ngữ văn	Dữ liệu KHÔNG tuân theo phân phối chuẩn ($p=0$)
K-S Test	Kiểm tra phân phối chuẩn cho môn Tiếng Anh	Dữ liệu KHÔNG tuân theo phân phối chuẩn ($p=0$)
K-S Test	Kiểm tra phân phối chuẩn cho môn Vật lí	Dữ liệu KHÔNG tuân theo phân phối chuẩn ($p=0$)
K-S Test	Kiểm tra phân phối chuẩn cho môn Hóa học	Dữ liệu KHÔNG tuân theo phân phối chuẩn ($p=0$)
K-S Test	Kiểm tra phân phối chuẩn cho môn Sinh học	Dữ liệu KHÔNG tuân theo phân phối chuẩn ($p=0$)
K-S Test	Kiểm tra phân phối chuẩn cho môn Lịch sử	Dữ liệu KHÔNG tuân theo phân phối chuẩn ($p=0$)
K-S Test	Kiểm tra phân phối chuẩn cho môn Địa lí	Dữ liệu KHÔNG tuân theo phân phối chuẩn ($p=0$)
K-S Test	Kiểm tra phân phối chuẩn cho môn GDСD/KTPL	Dữ liệu KHÔNG tuân theo phân phối chuẩn ($p=0$)
K-S Test	Kiểm tra phân phối chuẩn cho Tổng điểm	Dữ liệu KHÔNG tuân theo phân phối chuẩn ($p=0$)

4.1.2.c Phân tích Lợi thế Đô thị (Urban Advantage)

Phân tích này so sánh điểm trung bình của TPHCM với điểm trung bình toàn quốc để định lượng "lợi thế đô thị".

Môn	TP.HCM	Toàn Quốc	Chênh Lệch	%
Toán	5.26	4.78	+0.48	+10.04%
Sinh học	6.29	5.78	+0.51	+8.82%
Tiếng Anh	5.67	5.38	+0.29	+5.39%
Lịch sử	6.67	6.52	+0.15	+2.30%
Địa lí	6.82	6.63	+0.19	+2.87%
GDСD/KTPL	7.97	7.69	+0.28	+3.64%
Ngữ văn	7.06	7.00	+0.06	+0.86%
Vật lý	6.98	6.99	-0.01	-0.14%
Hóa học	5.95	6.06	-0.11	-1.82%

Bảng 5: So sánh lợi thế đô thị dựa trên điểm trung bình

4.1.2.d Phân tích Tổ hợp 3 Môn

Thống kê các tổ hợp 3 môn được sắp xếp theo điểm trung bình (từ cao xuống thấp) và sắp xếp theo số lượng (từ cao xuống thấp) để xác định các tổ hợp có hiệu suất cao nhất.

Bảng 6: Thống kê tổ hợp 3 môn (sắp xếp theo điểm trung bình 3 môn)

Tổ hợp 3 môn	Điểm TB 3 môn	Độ lệch chuẩn	Số lượng
D66	20.95	2.35	8709
C20	20.77	2.83	6048
C19	20.41	3.00	3890
D11	20.19	2.71	31164
D14	20.19	2.77	7528
D15	20.13	2.68	7642
C08	20.12	3.50	7310
C01	19.96	3.00	56975
C00	19.95	3.61	19531
C05	19.81	3.36	21418
...
<i>Các tổ hợp điểm trung bình thấp nhất:</i>			
C12	16.21	2.57	251
A02	15.62	3.37	512
B02	14.81	3.14	151
A05	14.47	3.20	224
A06	13.56	3.24	432

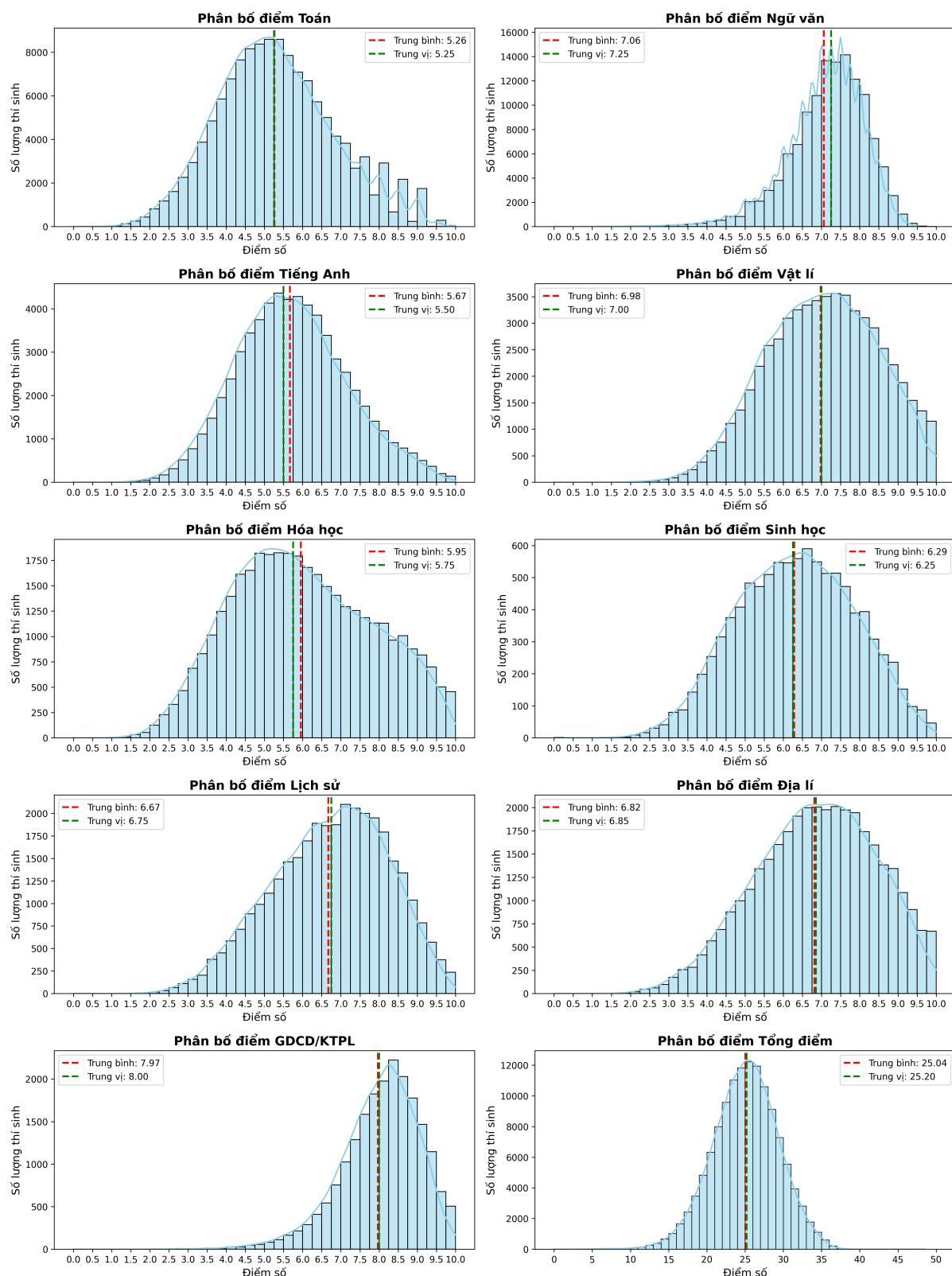
Bảng 7: Thống kê tổ hợp 3 môn (sắp xếp theo số thí sinh)

Tổ hợp 3 môn	Điểm TB 3 môn	Độ lệch chuẩn	Số lượng
D01	18.59	2.74	62318
C01	19.96	3.00	56975
C02	18.83	3.40	35364
C04	17.92	3.08	34140
C03	17.76	3.05	32136
A01	19.02	3.52	31473
D11	20.19	2.71	31164
A00	18.62	4.23	21777
C05	19.81	3.36	21418
C14	19.41	2.58	20199
...
<i>Các tổ hợp có số thí sinh thấp nhất:</i>			
A02	15.62	3.37	512
A06	13.56	3.24	432
C12	16.21	2.57	251
A05	14.47	3.20	224
A11	17.29	2.41	166

4.1.3 Nhận xét

4.1.3.1 Đặc điểm Phân bố Phổ điểm: Kết quả kiểm định Kolmogorov-Smirnov (K-S) cho thấy một phát hiện quan trọng: dữ liệu của tất cả 9 môn học và cả **Tổng điểm** đều **không tuân theo phân phối chuẩn** (tất cả $p).$

Điều này là hợp lý về mặt lý thuyết do đặc tính của dữ liệu điểm thi: (1) dữ liệu bị chặn ở hai đầu (bounded) trong khoảng $[0, 10]$, và (2) sự phân bố bị lệch (skewed) do mức độ khó/dễ của đề thi và năng lực của thí sinh.



Hình 1: Phân bố điểm (Histogram và KDE) của 9 môn thi và Tổng điểm. Các đường nét dứt màu đỏ và xanh lá cây lần lượt biểu thị giá trị Trung bình (Mean) và Trung vị (Median).

Quan sát từ Hình 1 cho thấy:

- Môn **Toán** có điểm trung bình (5.26) và trung vị (5.25) gần như trùng nhau, nhưng phổ điểm rộng, cho thấy tính phân hóa cao.

- Môn **Ngữ văn** (Mean = 7.06) và **GDCD/KTPL** (Mean = 7.97) cho thấy sự lệch trái rõ rệt (Mean < Median), cho thấy đề thi các môn này có xu hướng "dễ thở" hơn, giúp đa số thí sinh đạt điểm khá. "Toán" và "Tiếng Anh" có vẻ gần đối xứng hơn nhưng vẫn có đỉnh nhọn.
- Tổng điểm** có phân bố gần giống chuẩn nhưng cũng không vượt qua được kiểm định K-S (KS-statistic = 0.9999, $p = 0$).
- Tỉ lệ thí sinh TPHCM có điểm dưới 5 thấp hơn đáng kể so với toàn quốc ở hầu hết các môn, đặc biệt là môn Toán (TPHCM: 43.077% vs TQ: 56.395%).

4.1.3.2 Phân tích Ánh hưởng của Nhóm tuổi Kiểm định ANOVA một chiều được thực hiện để so sánh Tổng điểm giữa ba nhóm tuổi. Kết quả cho thấy sự khác biệt có ý nghĩa thống kê rất cao: $F(2, N) = 4366.49, p < 0.001$.

Giá trị F cực lớn này khẳng định rằng phuơng sai *giữa các nhóm* lớn hơn rất nhiều so với phuơng sai *trong nội bộ từng nhóm*. Nói cách khác, nhóm tuổi là một yếu tố ảnh hưởng thực sự đến kết quả thi, chứ không phải do ngẫu nhiên.

Bảng 8: Thống kê mô tả theo Nhóm tuổi

Nhóm tuổi	Tổng điểm			Điểm trung bình (Mean) theo môn học								
	Mean	Std	Count	Toán	Văn	Anh	Hóa	Lí	Sinh	Sử	Địa	GDCD
18	25.33	4.09	118,436	5.33	7.11	5.68	6.00	7.02	6.33	6.76	6.89	8.01
>18	22.62	4.17	7,842	4.07	6.41	5.07	4.67	5.89	5.14	6.16	6.29	7.61
Thi lại	19.32	5.97	2,869	5.79	6.42	5.66	6.36	6.35	6.63	6.45	6.95	7.56

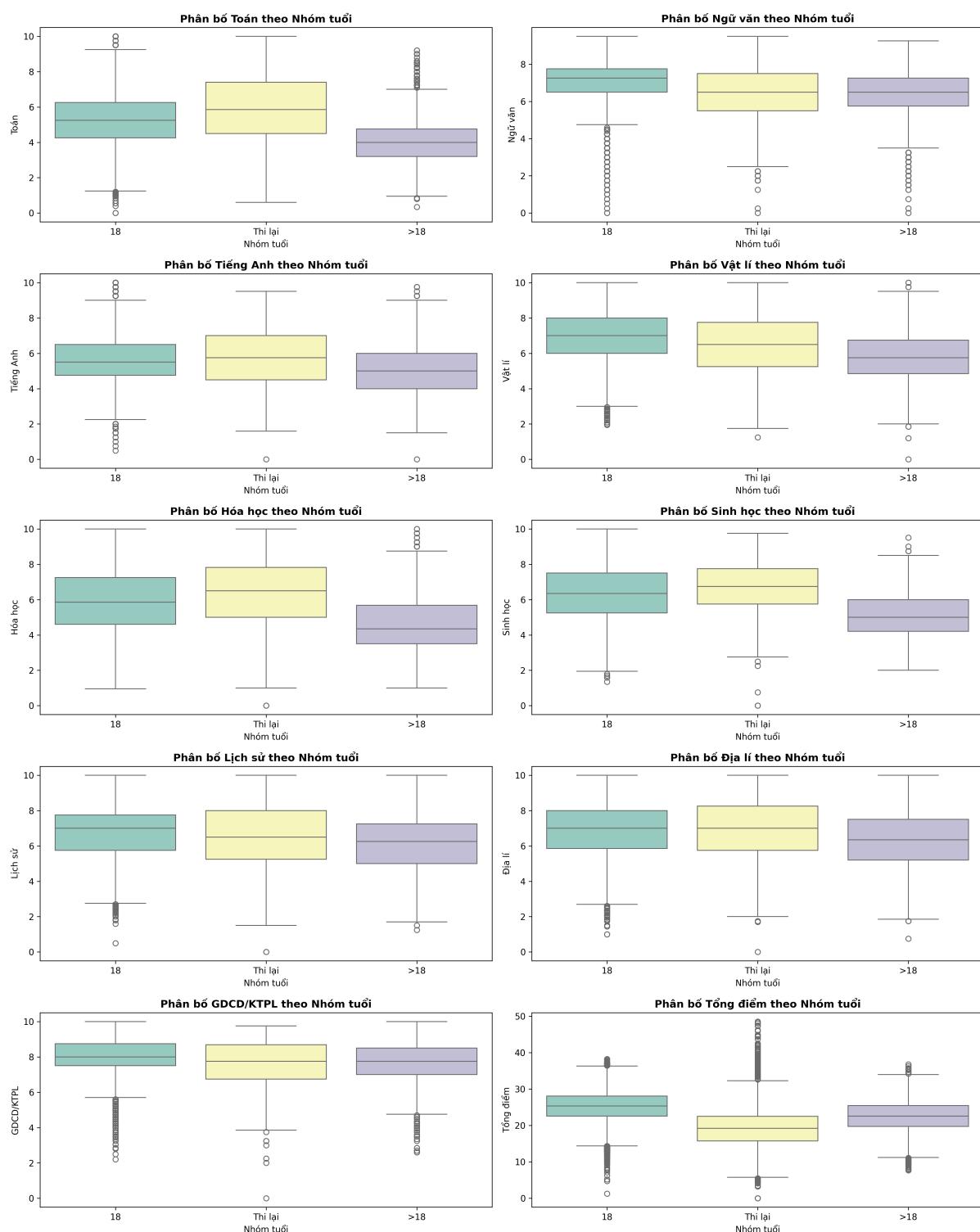
Từ Bảng 8 ta có thể thấy rõ 2 điều:

1. **Chiến lược “học lẹch” của nhóm “Thi lại”:**

Các môn gánh team của nhóm “Thi lại” chênh khá nhiều so với 2 nhóm còn lại: **Toán: 5.79** (so với **5.33**), **Hóa: 6.36** (so với **6.00**), **Sinh: 6.63** (so với **6.33**), **Địa: 6.95** (so với **6.89**). Điều này cho thấy nhóm “Thi lại” có chiến lược ôn thi rất tập trung — hay còn gọi là “học tủ” / ôn theo tổ hợp: họ dồn sức vào các môn cần thiết để xét tuyển (ví dụ: các môn khối B nhiều nhất) và có thể “buông” các môn khác. Kết quả là **điểm các môn họ tập trung** thì **rất cao**, nhưng **tổng điểm trung bình** (do bao gồm cả các môn bị điểm thấp) lại bị **kéo xuống**.

2. **Sự phân hóa lớn nhất ở nhóm “Thi lại”:**

Nhìn vào cột Std (độ lệch chuẩn) của Tổng điểm: Nhóm 18: **4.09**, Nhóm >18: **4.17**, Nhóm “Thi lại”: **5.97**. Độ lệch chuẩn của nhóm “Thi lại” cao vượt trội. Điều này có nghĩa là **điểm số trong nhóm này rất phân hóa và không đồng đều**. Có thể có **những người thi lại đạt điểm rất cao** (do ôn tập tập trung để thi lại vào ngành/trường khác), nhưng cũng có **nhiều người điểm rất thấp**, từ đó **kéo trung bình (Mean) xuống** (có thể thi lại do năm ngoái không đậu tốt nghiệp). Nhóm 18 tuổi có **điểm số đồng đều hơn nhiều**.



Hình 2: Biểu đồ Boxplot so sánh phân bố điểm của 9 môn và Tổng điểm giữa ba nhóm tuổi (18, Thi lại, >18).

Từ Hình 2, có thể thấy rõ:

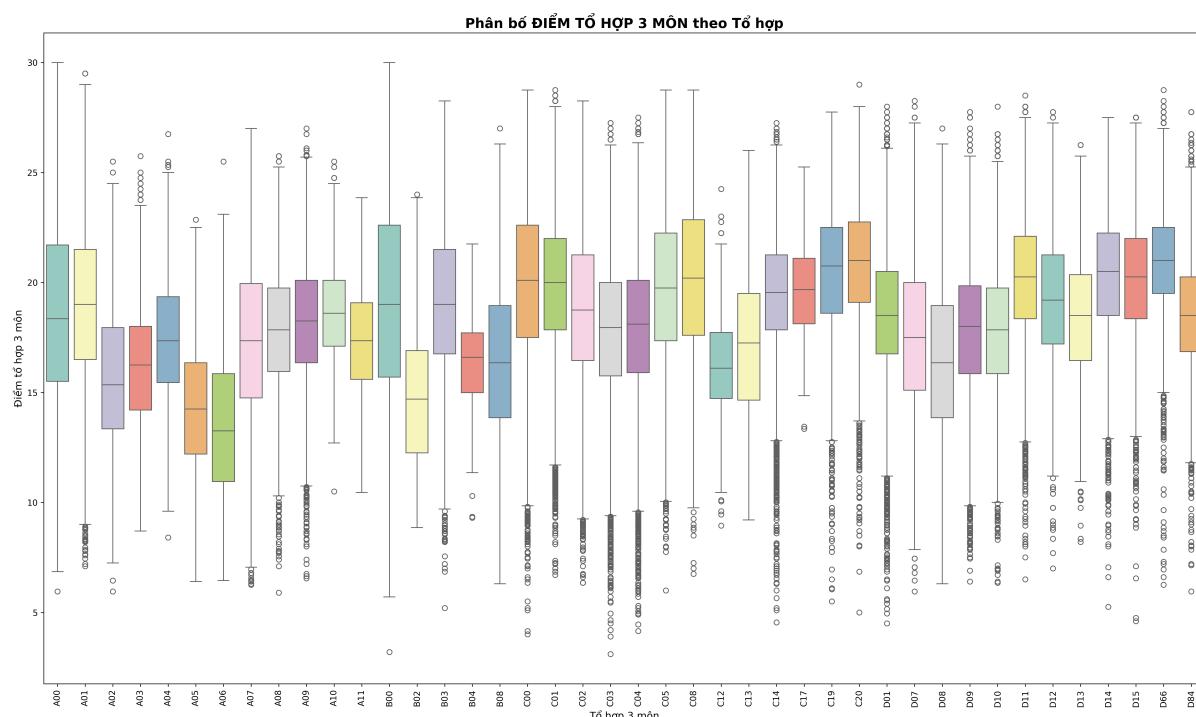
- Nhóm **18 tuổi** có mức điểm trung vị cao nhất và độ phân tán (chiều dài hộp) nhỏ nhất ở hầu hết các môn, đặc biệt là **Tổng điểm**.
- Nhóm "**Thi lại**" có kết quả thấp nhất một cách rõ rệt. Trung vị **Tổng điểm** của nhóm này thấp hơn đáng kể so với hai nhóm còn lại. Vì có vài thí sinh thi lại chỉ thi 3 môn đúng tổ hợp này nên có sự chênh lệch đó (4 môn ở nhóm 18 tuổi).

- Nhóm ">18" có kết quả nằm giữa nhóm 18 tuổi và nhóm Thi lại, nhưng có độ phân tán lớn (nhiều outliers).

4.1.3.3 Phân tích theo Tổ hợp 3 môn Để đánh giá công bằng hơn, điểm của các tổ hợp 3 môn đã được tính lại bằng cách chỉ cộng điểm của 3 môn đó và loại bỏ các thí sinh thiếu điểm bất kỳ môn nào trong tổ hợp.

Kết quả ANOVA so sánh Điểm tổ hợp 3 môn giữa các tổ hợp khác nhau cũng cho thấy sự khác biệt có ý nghĩa thống kê rất cao: $F(k, N) = 1170.95, p < 0.001$.

Hình 3 trực quan hóa sự khác biệt này.



Hình 3: Phân bố điểm tổ hợp 3 môn theo từng tổ hợp.

Các phát hiện chính từ Hình 3 và tệp h1.txt:

- Tổ hợp phổ biến nhất:** D01 (Toán, Văn, Anh) có số lượng thí sinh đăng ký nhiều nhất ($N = 62,318$).
- Tổ hợp điểm cao nhất:** D66 (Văn, GD&CD, Anh) đạt điểm trung bình 3 môn cao nhất (Mean = 20.95).
- Tổ hợp điểm thấp nhất:** A06 (Toán, Hóa, Địa) có điểm trung bình 3 môn thấp nhất (Mean = 13.56).
- Độ phân tán:** Rất khác biệt, cho thấy một số tổ hợp (như A00) có sự phân hóa thí sinh rất lớn, trong khi các tổ hợp khác (như D66) có điểm số tập trung hơn.

⇒ Điều này cho thấy các môn KHXH (đặc biệt là GD&CD) đang kéo điểm trung bình của các tổ hợp lên cao hơn so với các môn KHTN (đặc biệt là Toán).

4.1.3.4 Phân tích Lợi thế Đô thị (Urban Advantage) Bảng 9 tóm tắt kết quả so sánh điểm trung bình của mẫu TP.HCM so với toàn quốc.

Kết quả cho thấy một "lợi thế đô thị" rõ rệt ở hầu hết các môn, đặc biệt là **Toán** (+0.48 điểm) và **Sinh học** (+0.51 điểm). Môn **Tiếng Anh** cũng cho thấy lợi thế đáng kể (+0.29 điểm), có thể lý giải bằng việc học sinh ở đô thị có nhiều cơ hội tiếp cận và học thêm ngoại ngữ hơn.



Bảng 9: So sánh điểm trung bình TP.HCM và Toàn quốc

Môn	TP.HCM (Mean)	Toàn quốc (Mean)	Chênh lệch (Điểm)	Chênh lệch (%)
Toán	5.26	4.78	+0.48	+10.04%
Sinh học	6.29	5.78	+0.51	+8.82%
Tiếng Anh	5.67	5.38	+0.29	+5.39%
GDCD/KTPL	7.97	7.69	+0.28	+3.64%
Địa lí	6.82	6.63	+0.19	+2.87%
Lịch sử	6.67	6.52	+0.15	+2.30%
Ngữ văn	7.06	7.00	+0.06	+0.86%
Vật lí	6.98	6.99	-0.01	-0.14%
Hóa học	5.95	6.06	-0.11	-1.82%

Tuy nhiên, một phát hiện đáng ngạc nhiên là sự xuất hiện của "bất lợi đô thị" ở hai môn KHTN là **Vật lí** (-0.01 điểm) và **Hóa học** (-0.11 điểm).

4.1.4 Kết luận và Quyết định

Các kết quả từ Hướng 1 cung cấp một bức tranh đa chiều về kỳ thi.

Thứ nhất, việc tất cả các môn đều **không tuân theo phân phối chuẩn** ($p < 0.001$) là một khẳng định về mặt thống kê. Điều này cho thấy việc sử dụng các kiểm định phi tham số (non-parametric) hoặc các mô hình robust (như Random Forest trong Hướng 2) là cần thiết và phù hợp hơn so với các phương pháp dựa trên giả định chuẩn.

Thứ hai, kết quả ANOVA về **nhóm tuổi** ($p < 0.001$) khẳng định nhóm "Thi lại" là một nhóm đối tượng riêng biệt, có kết quả học tập thấp hơn đáng kể (như Hình 2 đã chỉ ra). Điều này cho thấy việc thi lại không đảm bảo cải thiện điểm số và nhóm này cần các chiến lược hỗ trợ đặc biệt.

Thứ ba, sự khác biệt lớn về điểm số giữa các **tổ hợp 3 môn** (Hình 3) cho thấy việc lựa chọn tổ hợp có ảnh hưởng lớn đến tổng điểm xét tuyển. Sự chênh lệch giữa tổ hợp cao nhất (D66, Mean = 20.95) và thấp nhất (A06, Mean = 13.56) là rất lớn, phản ánh sự khác biệt về độ khó tương đối giữa các môn và/hoặc năng lực của các nhóm thí sinh lựa chọn các tổ hợp này. Đối với học sinh/phụ huynh: Cần có tư vấn hướng nghiệp rõ ràng hơn, rằng việc chọn tổ hợp có môn "GDCD/KTPL" có thể mang lại tổng điểm cao hơn đáng kể. Đối với các trường Đại học: Các trường cần nhận thức rằng điểm đầu vào của các tổ hợp rất khác nhau. Việc đặt điểm sàn tương đương cho các tổ hợp D01 (điểm TB 18.59) và D66 (điểm TB 20.95) có thể không công bằng.

Thứ tư, cảnh báo về "Điểm liệt" và khả năng phân loại. Môn Toán có tỉ lệ điểm < 5 rất cao (43.077%), cho thấy đây là môn học gây khó khăn lớn nhất. Ngược lại, các môn như Địa lí và GDCD/KTPL có rất nhiều điểm 10. Điều này đặt ra câu hỏi về khả năng phân loại thí sinh giỏi của đề thi các môn này.

Cuối cùng, phân tích **lợi thế đô thị** mang lại kết quả thú vị nhất. Lợi thế ở môn Toán (+10.04%) và Tiếng Anh (+5.39%) xác nhận các giả định phổ biến về chất lượng giảng dạy và nguồn lực vượt trội ở thành thị. Tuy nhiên, kết quả âm ở môn Hóa học (-1.82%) và Vật lí là một nghịch lý. Điều này có thể xuất phát từ hai khả năng: (1) Đề thi hai môn này có độ khó cao, gây ảnh hưởng đến cả nhóm học sinh vốn được coi là "học tốt" ở TP.HCM; hoặc (2) Có sự dịch chuyển trong lựa chọn tổ hợp của học sinh TP.HCM, khiến số lượng (và có thể là chất lượng đầu vào) của nhóm thi hai môn này giảm đi. Đây là một câu hỏi mở cần được nghiên cứu sâu hơn về dữ liệu đăng ký dự thi (self-selection bias).

4.2 Hướng 2: Phân tích Xu hướng Học tập (Cohen's d) và Mô hình hóa Dự báo

Hướng phân tích này giải quyết hai câu hỏi nghiên cứu cốt lõi: (RQ3) Mỗi tương quan giữa các môn thi đã thay đổi ra sao? và (RQ4) Yếu tố nào là quan trọng nhất trong việc dự đoán tổng điểm của thí sinh?

Chúng tôi tiếp cận vấn đề này bằng cách:



1. Lượng hóa xu hướng học "Thiên Văn" hoặc "Thiên Toán" của từng thí sinh bằng chỉ số Cohen's d.
2. Phân tích tương quan giữa 9 môn thi.
3. Xây dựng và so sánh các mô hình học máy để dự đoán **Tổng điểm**.
4. Sử dụng kỹ thuật Diện giải Mô hình (XAI) SHAP để xác định các yếu tố "then chốt" ảnh hưởng đến kết quả.

Tất cả phân tích được thực hiện bằng tệp `h2.py`.

4.2.1 Phương pháp luận

1. **Phân tích Xu hướng học (Cohen's d):** Để đo lường mức độ một thí sinh "thiên" về Ngữ văn hay Toán, chúng tôi sử dụng Cohen's d. Đây là một phép đo "effect size" (kích thước ảnh hưởng) chuẩn trong nghiên cứu giáo dục. Nó chuẩn hóa sự khác biệt giữa điểm Ngữ văn và điểm Toán theo độ lệch chuẩn chung ($d = (\text{Điểm Văn} - \text{Điểm Toán})/\text{SD}_{\text{chung}}$).

- $d > 0.2$: Được phân loại là "Thiên Văn".
- $d < -0.2$: Được phân loại là "Thiên Toán".
- $-0.2 \leq d \leq 0.2$: Được phân loại là "Cân bằng".

Vì sao dùng Cohen's d?: Vì độc lập với kích thước mẫu (không bị ảnh hưởng bởi N lớn).Thêm vào đó chuẩn hóa chênh lệch (cho phép so sánh giữa các môn khác nhau) và đó là tiêu chuẩn quốc tế trong giáo dục

2. **Phân tích Tương quan (Correlation):** Chúng tôi sử dụng ma trận tương quan Pearson (Pearson Correlation Matrix) để đo lường mức độ quan hệ tuyến tính giữa 9 môn thi. Hệ số tương quan r từ -1 đến +1, trong đó $r > 0.7$ được coi là tương quan mạnh.
3. **Mô hình hóa Dự báo (Predictive Modeling):** Chúng tôi huấn luyện ba mô hình để dự đoán **Tổng điểm** dựa trên điểm 9 môn học và nhóm tuổi:
 - **Linear Regression (Hồi quy Tuyến tính):** Mô hình cơ sở (baseline).
 - **Random Forest Regressor:** Một mô hình ensemble mạnh, có khả năng bắt các mối quan hệ phi tuyến tính.
 - **Gradient Boosting Regressor:** Một mô hình ensemble mạnh khác, thường cho độ chính xác cao.
4. **Đánh giá Mô hình (Model Validation):** Hiệu suất của các mô hình được đánh giá bằng **R-squared** (R^2) (mức độ giải thích phương sai) và **RMSE** (Root Mean Squared Error - Sai số Trung bình Căn bậc hai, đo lường độ lệch trung bình của dự đoán so với thực tế, tính bằng điểm).
Chúng tôi cũng sử dụng **Kiểm định chéo 10-lần (10-Fold Cross Validation)** để đảm bảo các mô hình có khả năng tổng quát hóa tốt, không bị overfitting.
5. **Diễn giải Mô hình (Explainable AI - XAI):** Chúng tôi sử dụng **SHAP (SHapley Additive exPlanations)** trên mô hình Gradient Boosting (mô hình hoạt động tốt) để hiểu rõ: (1) Môn học nào quan trọng nhất? và (2) Môn học đó ảnh hưởng đến **Tổng điểm** theo chiều hướng (âm/dương) và cường độ (mạnh/yếu) như thế nào.

4.2.2 Kết quả Phân tích

4.2.2.a Phân tích Xu hướng học (Cohen's d)

Phân tích Cohen's d cho thấy một sự mất cân bằng lớn trong xu hướng học của thí sinh.

- 4.2.2.1 **Phân bố thí sinh theo xu hướng học:** Dữ liệu từ `h2.txt` cho thấy sự áp đảo của nhóm "Thiên Văn": Theo xu hướng học:

- **Thiên Văn:** 109,272 thí sinh
- **Thiên Toán:** 14,678 thí sinh

- **Cân bằng hoàn toàn:** 3,039 thí sinh
- **Không xác định:** 2,159 thí sinh

Theo xu hướng phân loại: $-|d| < 0.2$: Cân bằng (không đáng kể) $-0.2 \leq |d| < 0.5$: Cân bằng (lệch nhỏ) $-0.5 \leq |d| < 0.8$: Lệch rõ (lệch trung bình) $-|d| \geq 0.8$: Lệch mạnh (lệch lớn)

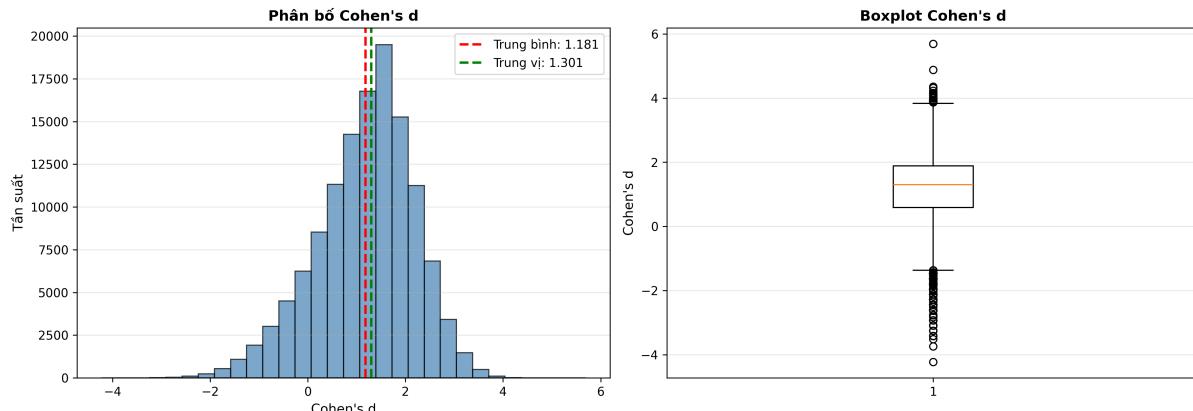
- **Lệch mạnh:** 93,434 thí sinh
- **Cân bằng (lệch nhỏ):** 14,427 thí sinh
- **Lệch rõ:** 9,182 thí sinh

- **Cân bằng:** 6,882 thí sinh
- **Cân bằng hoàn toàn:** 3,064 thí sinh
- **Không xác định:** 2,159 thí sinh

4.2.2.2 Thống kê chỉ số Cohen's d:

- **Trung bình (Mean):** 1.1805
- **Trung vị (Median):** 1.3013

- **Độ lệch chuẩn (Std):** 1.0000
- **Min:** -4.2292
- **Max:** 5.6932



Hình 4: Phân bố (Histogram) và Boxplot của chỉ số Cohen's d (Văn - Toán). Giá trị dương có nghĩa là "Thiên Văn", giá trị âm có nghĩa là "Thiên Toán".

Từ Hình 4 và các số liệu thống kê, chúng ta thấy:

- Phân bố lệch hẳn về bên phải (lệch dương), với giá trị trung bình (1.18) và trung vị (1.30) đều lớn hơn 0.
- Điều này khẳng định đa số thí sinh có điểm Ngữ văn cao hơn điểm Toán một cách đáng kể.

Bảng 10 tóm tắt sự phân bố và kết quả chi tiết của các nhóm này.

Bảng 10: Thống kê mô tả chi tiết theo Xu hướng học (Cohen's d)

Xu hướng	Tổng điểm			Diểm trung bình (Mean) các chỉ số					
	Mean	Std	Count	Toán	Văn	Cohen's d	TB	Bắt buộc	TB
Thiên Toán	28.78	4.61	14,678	7.49	6.46		-0.67	6.98	7.51
Cân bằng	27.69	4.16	3,039	6.84	6.84		0.00	6.84	7.08
Thiên Văn	24.57	3.85	109,272	4.91	7.16		1.46	6.03	6.35
N/A	19.54	5.54	2,159	6.43	6.34		NaN	NaN	6.62

4.2.2.3 Kiểm định ANOVA (Tổng điểm): Kết quả kiểm định ANOVA cho Tổng điểm giữa 3 nhóm (Thiên Văn, Thiên Toán, Cân bằng) là:

- **F-statistic:** 7027.9234



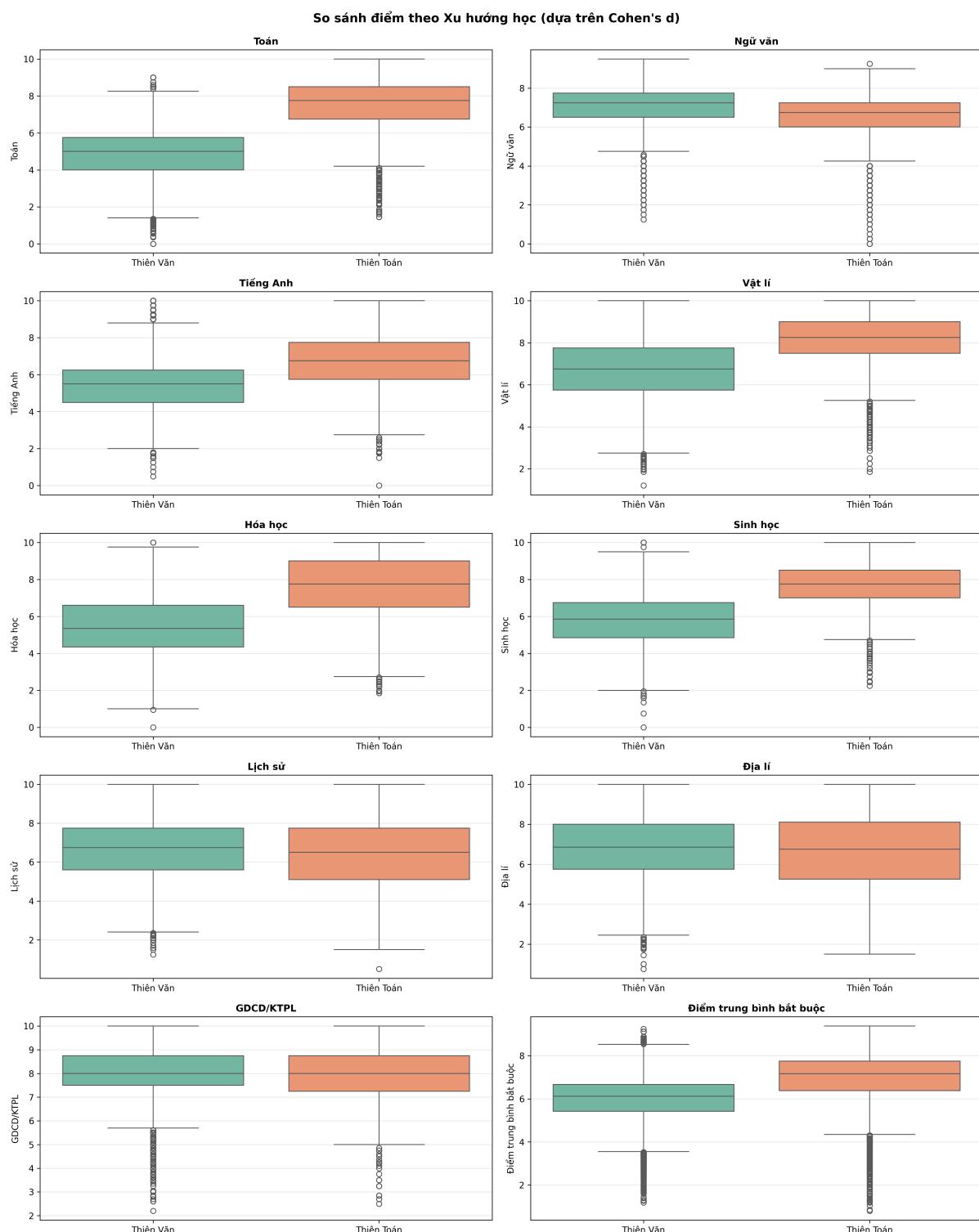
- **p-value:** 0.000000
- **Kết luận:** Có sự khác biệt có ý nghĩa thống kê RẤT LỚN giữa các nhóm ($p < 0.05$).

Phát hiện quan trọng từ Bảng 10 là:

1. **Nghịch lý "Thiên Văn":** Mặc dù nhóm "Thiên Văn" chiếm đa số (109,272 thí sinh), nhóm này lại có **Tổng điểm** trung bình (24.57) thấp hơn đáng kể so với cả nhóm "Cân bằng"(27.69) và nhóm "Thiên Toán"(28.78).
2. **Chênh lệch Lớn:** Nhóm "Thiên Toán" có **tổng điểm** cao hơn nhóm "Thiên Văn"tới **4.21 điểm** (khoảng 17%).

4.2.2.b So sánh chi tiết nhóm 'Thiên Văn' và 'Thiên Toán'

Hình 5 đi sâu vào so sánh thành tích của hai nhóm đối lập này trên tất cả các môn.



Hình 5: So sánh phân bố điểm 9 môn và điểm bắt buộc giữa nhóm Thiên Văn (Xanh) và Thiên Toán (Cam).

Quan sát từ Hình 5 cho thấy một phát hiện đáng ngạc nhiên:

- Nhóm "Thiên Toán"(màu cam) không chỉ vượt trội ở môn Toán, mà còn có **điểm trung bình cao hơn ở TẤT CẢ các môn học khác**, bao gồm cả Ngữ văn, Lịch sử, Địa lí, và GDCD/KTPL.
- Điều này được xác nhận bằng loạt kiểm định ANOVA cho từng môn (Bảng 11), tất cả đều cho thấy sự khác biệt có ý nghĩa thống kê.

Bảng 11: Phân tích ANOVA cho từng môn học theo Xu hướng học (Thiên Văn vs Thiên Toán)

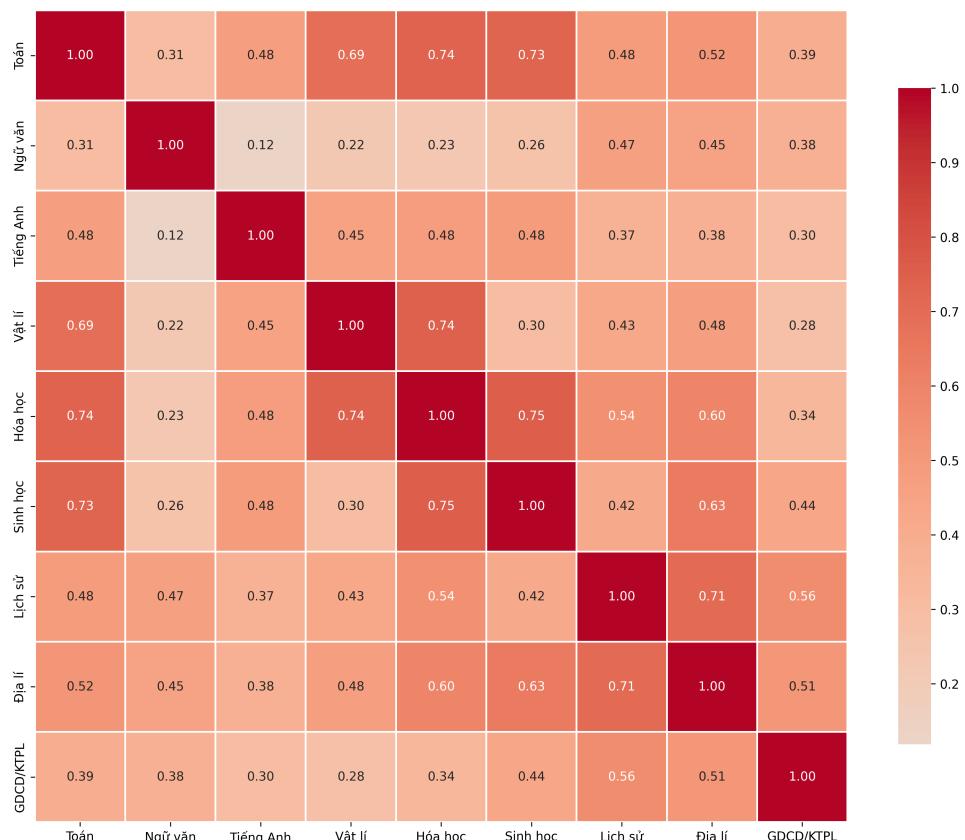
Môn học	F-statistic	p-value	Kết luận ($p < 0.05$)
Toán	55534.4185	0.000000	CÓ sự khác biệt
Ngữ văn	6495.0093	0.000000	CÓ sự khác biệt
Tiếng Anh	5412.9911	0.000000	CÓ sự khác biệt
Vật lí	10921.7042	0.000000	CÓ sự khác biệt
Hóa học	8407.3516	0.000000	CÓ sự khác biệt
Sinh học	2646.9073	0.000000	CÓ sự khác biệt
Lịch sử	32.7424	0.000000	CÓ sự khác biệt
Địa lí	10.2442	0.001372	CÓ sự khác biệt
GDCD/KTPL	10.7869	0.001024	CÓ sự khác biệt

Kết quả này cho thấy "Thiên Toán" dường như là một chỉ báo của một nhóm học sinh **có năng lực học tập tổng thể cao hơn**. Ngược lại, nhóm "Thiên Văn" có điểm số thấp hơn trên diện rộng, với điểm yếu chí mạng là môn Toán (trung bình chỉ 4.91).

4.2.2.c Phân tích Tương quan Môn học

Ma trận tương quan (Hình 6) cho thấy các môn học liên kết với nhau như thế nào.

Ma trận tương quan Pearson giữa các môn thi (Năm 2025)



Hình 6: Ma trận tương quan Pearson giữa 9 môn thi. Màu càng đỏ, tương quan càng mạnh (gần 1.0).

Từ Hình 6 và tệp h2.txt, chúng ta rút ra các cụm liên kết rõ rệt:

- **Cụm Khoa học Tự nhiên (KHTN):** Có sự liên kết rất chặt chẽ.

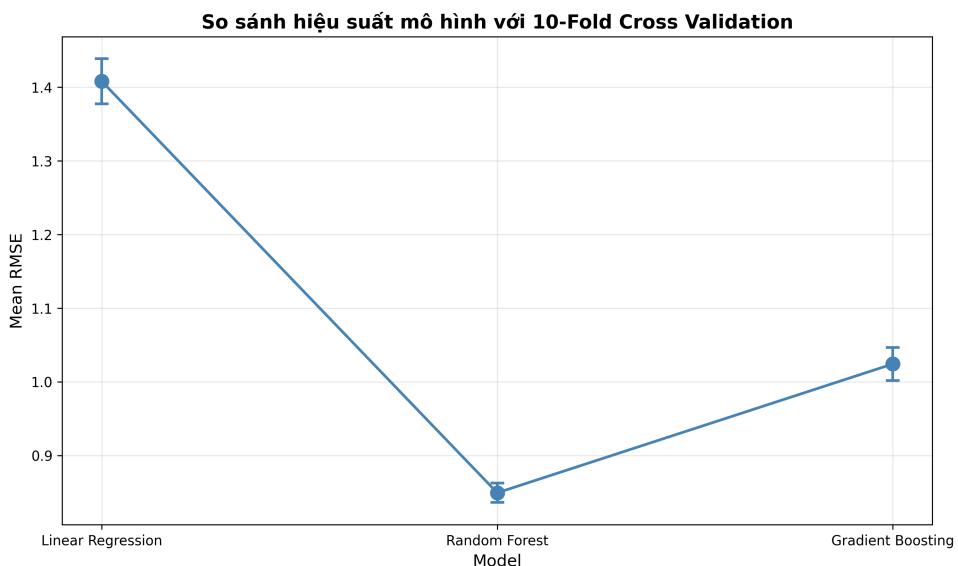
- Hóa học - Sinh học: $r = 0.75$
 - Vật lí - Hóa học: $r = 0.74$
 - Toán - Hóa học: $r = 0.74$
 - Toán - Sinh học: $r = 0.73$
 - Toán - Vật lí: $r = 0.69$
- **Cụm Khoa học Xã hội (KHXH):** Củng liên kết chặt chẽ với nhau.
 - Lịch sử - Địa lí: $r = 0.71$
 - Lịch sử - GDCH/KTPL: $r = 0.56$
 - Địa lí - GDCH/KTPL: $r = 0.51$
 - **Môn liên kết yếu:** Ngữ văn có tương quan rất yếu với các môn KHTN (ví dụ: Văn-Lý $r = 0.22$, Văn-Hóa $r = 0.23$). Tiếng Anh cũng tương đối độc lập.

4.2.2.d Mô hình hóa Dự báo Tổng điểm

Chúng tôi so sánh 3 mô hình để tìm ra mô hình dự báo Tổng điểm tốt nhất.

Bảng 12: So sánh hiệu suất các mô hình dự báo Tổng điểm

Mô hình	R ² Score	RMSE (Test set)	10-Fold CV RMSE
Linear Regression	0.8946	1.3902	1.4081
Random Forest	0.9582	0.8750	0.8492
Gradient Boosting	0.9453	1.0017	1.0242



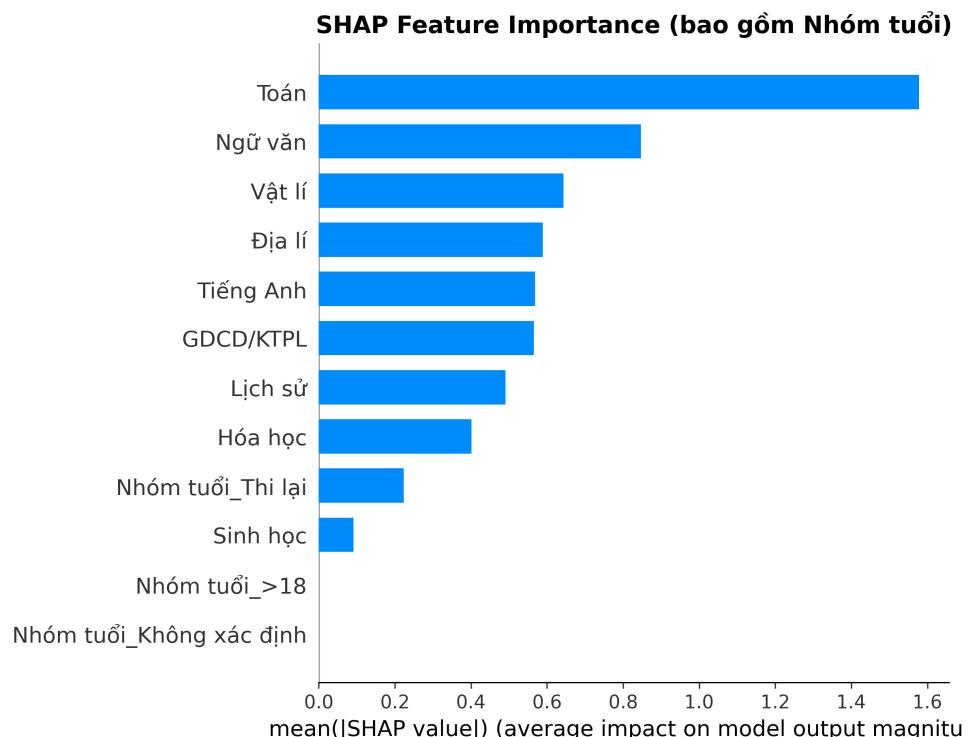
Hình 7: So sánh Sai số Trung bình (Mean RMSE) của 3 mô hình qua 10-Fold Cross Validation.

Kết quả từ Bảng 12 và Hình 7 rất rõ ràng:

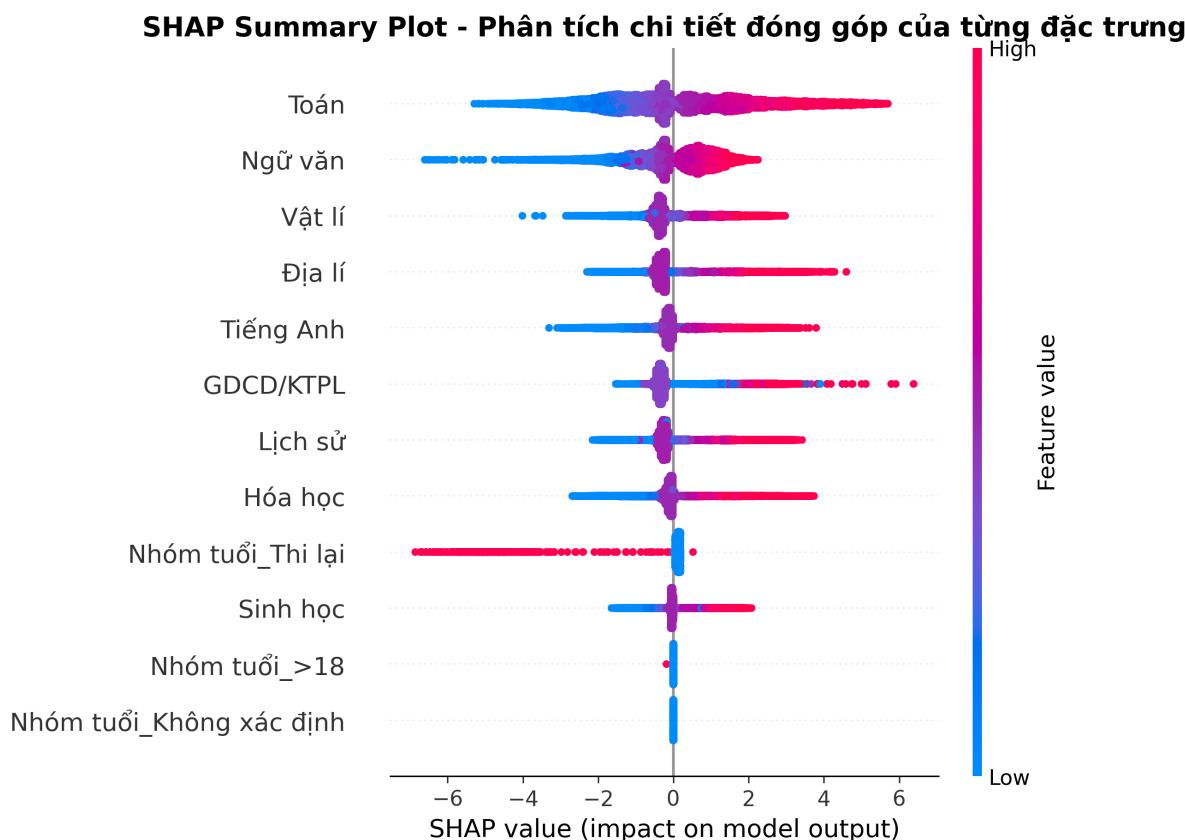
- **Random Forest** là mô hình vượt trội nhất, với $R^2 = 0.9582$ (giải thích được 95.8% sự biến thiên của tổng điểm).
- Quan trọng hơn, 10-Fold CV RMSE của Random Forest (0.8492) là thấp nhất và rất gần với RMSE của test set (0.8750), cho thấy mô hình **rất ổn định và không bị overfitting**.
- Hiệu suất cao của Random Forest so với Linear Regression cho thấy mối quan hệ giữa điểm các môn và tổng điểm là **phi tuyến tính (non-linear)** và phức tạp.

4.2.2.e Diễn giải Mô hình (XAI) với SHAP

Sử dụng SHAP trên mô hình Gradient Boosting (vì Random Forest tính SHAP rất lâu) để hiểu "tại sao" một thí sinh có điểm cao/thấp.



Hình 8: Mức độ quan trọng của các đặc trưng (SHAP Feature Importance). Trục hoành là tác động trung bình tuyệt đối lên dự đoán tổng điểm.



Hình 9: Phân tích chi tiết SHAP (Beeswarm plot). Trục hoành là "SHAP value"(tác động lên tổng điểm). Màu đỏ = giá trị đặc trưng cao (điểm cao). Màu xanh = giá trị đặc trưng thấp (điểm thấp).

Hình 8 và 9 cung cấp những insight quan trọng nhất của toàn bộ phân tích:

- Toán là Vua (Math is King):** Môn Toán (SHAP value = 1.578) là yếu tố quan trọng **gấp đôi** môn quan trọng thứ hai là Ngữ văn (SHAP value = 0.847).
- Tác động Đối xứng của Toán:** Nhìn vào Hình 9, môn Toán có độ "dàn trải" lớn nhất.
 - Điểm Toán cao (chấm đỏ) có tác động tích cực CỰC LỚN, đẩy tổng điểm dự đoán tăng lên tối +7 điểm.
 - Điểm Toán thấp (chấm xanh) có tác động tiêu cực CỰC LỚN, kéo tổng điểm dự đoán giảm xuống tối -5 điểm.
- Tác động của các môn khác:** Các môn như Ngữ văn, Vật lí, Địa lí, Tiếng Anh cũng quan trọng, nhưng tác động của chúng (cả tích cực và tiêu cực) đều **nhỏ hơn đáng kể** so với môn Toán.
- Yếu tố Nhân khẩu học:** Nhóm “Thi lại” là một yếu tố dự báo tiêu cực. Các chấm đỏ (nghĩa là Nhóm tuổi_Thi lại = True) tập trung gần như hoàn toàn ở phía âm, nghĩa là việc thi lại kéo tổng điểm dự đoán xuống.

4.2.3 Nhận xét và Quyết định (Decision Making)

Dựa trên các kết quả phân tích, chúng tôi đưa ra các nhận xét và khuyến nghị mang tính quyết định sau:

4.2.3.1 Quyết định 1: Ưu tiên chiến lược "Chinh phục môn Toán".

- Phát hiện:** Da số học sinh (75.8%) thuộc nhóm "Thiên Văn"(điểm Văn > Toán), nhưng nhóm này lại có tổng điểm thấp hơn 4.21 điểm so với nhóm "Thiên Toán".



- **Lý do:** SHAP (Hình 9) chỉ ra Toán là môn có "đòn bẩy" lớn nhất. Điểm Toán thấp (như mức trung bình 4.91 của nhóm "Thiên Văn") là yếu tố "trừng phạt" (penalty) mạnh nhất, kéo sập tổng điểm. Ngược lại, điểm Toán cao là yếu tố "thưởng" (bonus) lớn nhất.
- **Khuyến nghị (Decision):**
- *Dối với Học sinh/Phụ huynh:* Không thể "bù đắp" điểm Toán yếu bằng các môn khác. Để đạt tổng điểm cao, việc đầu tư nâng cao điểm Toán là **chiến lược hiệu quả nhất**.
- *Dối với Nhà trường:* Cần có các chương trình phụ đạo và tăng cường môn Toán đặc biệt cho nhóm học sinh có thiên hướng KHXH, vì đây là "gót chân Achilles" của các em.

4.2.3.2 Quyết định 2: Tái định nghĩa "Thiên Văn" và "Thiên Toán".

- **Phát hiện:** Nhóm "Thiên Toán" học giỏi hơn nhóm "Thiên Văn" ở **tất cả các môn**, kể cả Ngữ văn (Bảng 11).
- **Lý do:** Chỉ số Cohen's d (Văn - Toán) dường như không đo lường "sở thích" (KHXH vs KHTN), mà đang vô tình phân tách **hai nhóm năng lực học tập** (một nhóm học toàn diện và một nhóm học yếu hơn, đặc biệt yếu Toán).
- **Khuyến nghị (Decision):**
- *Dối với Giáo viên:* Khi thấy một học sinh "Thiên Văn" (theo định nghĩa này), không nên chỉ hiểu là em đó "giỏi Văn", mà phải hiểu là em đó **có nguy cơ cao bị điểm thấp** do yếu Toán.

4.2.3.3 Quyết định 3: Tư vấn chọn tổ hợp dựa trên các "Cụm Năng lực".

- **Phát hiện:** Ma trận tương quan (Hình 6) cho thấy các môn học đi theo "cụm" rất rõ ràng (KHTN: Toán-Lý-Hóa-Sinh; KHXH: Văn-Sử-Địa-GDCD).
- **Lý do:** Năng lực tư duy logic cần cho Toán cũng cần cho Lý, Hóa. Năng lực tư duy xã hội cần cho Sử cũng cần cho Địa.
- **Khuyến nghị (Decision):**
- *Dối với Tư vấn Hướng nghiệp:* Thay vì chọn môn rời rạc, học sinh nên được tư vấn xác định mình thuộc "cụm năng lực" nào (KHTN hay KHXH) và **chọn tổ hợp trọn vẹn trong cụm đó** để tối đa hóa điểm số, vì các môn này hỗ trợ lẫn nhau.

4.2.3.4 Quyết định 4: Sử dụng Mô hình AI để Cảnh báo Sớm.

- **Phát hiện:** Mô hình Random Forest có thể dự đoán **Tổng điểm** với độ chính xác cao (RMSE ≈ 0.85 điểm).
- **Lý do:** Mô hình đã "học" được các mối quan hệ phức tạp và phi tuyến tính (ví dụ: điểm Toán từ 1-5 tác động tiêu cực ra sao, từ 8-10 tác động tích cực ra sao).
- **Khuyến nghị (Decision):**
- *Dối với Nhà trường:* Có thể áp dụng các mô hình tương tự (sử dụng điểm thi thử) để **dự báo sớm và chính xác** kết quả của học sinh.
- *Hành động:* Khi mô hình dự báo một học sinh có tổng điểm thấp, nhà trường có thể dùng SHAP để **xác định chính xác nguyên nhân** (ví dụ: "điểm Toán của em đang kéo em xuống 3.5 điểm") và đưa ra can thiệp tùy chỉnh, thay vì một lời khuyên chung chung.

4.3 Hướng 3: Phân tích Hiệu ứng Tuổi Tương đối (Relative Age Effect)

Hướng phân tích này tập trung giải quyết câu hỏi nghiên cứu RQ2: "Các yếu tố nhân khẩu học như tháng sinh (Hiệu ứng Tuổi Tương đối) có ảnh hưởng đến kết quả học tập trong kỳ thi mới hay không?" Mục tiêu là kiểm chứng giả thuyết khoa học từ OECD, cho rằng những học sinh sinh vào đầu năm học (ví dụ, Quý 1 hoặc Quý 4) thường có lợi thế về mặt nhận thức và thể chất, dẫn đến kết quả học tập cao hơn so với những em sinh vào cuối năm (ví dụ, Quý 4 hoặc Quý 2).



4.3.1 Phương pháp luận

Để kiểm tra giả thuyết này, chúng tôi đã thực hiện các bước phân tích dựa trên tệp h3.py:

- Tạo đặc trưng "Quý sinh" (Feature Engineering):** Dựa trên cột ngày sinh của thí sinh, chúng tôi tạo ra một biến phân loại mới là Quý sinh (Birth Quarter). Biến này được định nghĩa như sau:
 - Q1:** Sinh tháng 1, 2, 3 (nhóm "lớn tuổi" nhất, giả định có lợi thế).
 - Q2:** Sinh tháng 4, 5, 6.
 - Q3:** Sinh tháng 7, 8, 9.
 - Q4:** Sinh tháng 10, 11, 12 (nhóm "nhỏ tuổi" nhất, giả định bất lợi).
- Thống kê Mô tả:** Chúng tôi tính toán các chỉ số thống kê mô tả (Mean, Std Dev, Count) cho Tổng điểm và điểm trung bình (Mean) cho tất cả 9 môn thi, được phân nhóm theo 4 Quý sinh.
- Kiểm định Giả thuyết (ANOVA):** Sử dụng Phân tích Phutong sai một chiều (One-Way ANOVA) để so sánh giá trị trung bình của Tổng điểm giữa bốn nhóm Quý sinh. Giả thuyết không (H_0) là không có sự khác biệt về điểm trung bình giữa các nhóm. Chúng tôi sử dụng mức ý nghĩa $\alpha = 0.05$.
- Trực quan hóa Dữ liệu:** Sử dụng biểu đồ hộp (Boxplot) để so sánh trực quan sự phân bố điểm của Tổng điểm và 9 môn thi giữa 4 nhóm quý sinh.

4.3.2 Kết quả Phân tích

4.3.2.a Thống kê mô tả theo Quý sinh

Bảng 13 tóm tắt các chỉ số thống kê chính được trích xuất từ tệp h3.txt. Nhìn chung, sự khác biệt về điểm trung bình giữa các quý là rất nhỏ. Q1 có Tổng điểm trung bình cao nhất (25.07), trong khi Q2 có điểm trung bình thấp nhất (24.99). Sự chênh lệch lớn nhất chỉ là 0.08 điểm, một con số không đáng kể.

Bảng 13: Thống kê mô tả điểm thi theo Quý sinh (dữ liệu từ h3.txt)

Quý sinh	Tổng điểm			Điểm trung bình (Mean) các môn								
	Mean	Std	Count	Toán	Văn	Anh	Lí	Hóa	Sinh	Sử	Địa	GD&CD
Q1	25.07	4.30	28590	5.27	7.08	5.69	6.98	5.95	6.33	6.69	6.84	7.99
Q2	24.99	4.27	28880	5.24	7.08	5.66	6.96	5.94	6.26	6.64	6.80	7.97
Q3	25.03	4.30	31146	5.26	7.05	5.67	6.98	5.96	6.31	6.70	6.83	7.97
Q4	25.04	4.27	40532	5.28	7.05	5.67	6.98	5.96	6.26	6.67	6.80	7.94

4.3.2.1 Phân tích điểm trung bình cao nhất theo môn: Dựa trên dữ liệu từ h3.txt, chúng tôi ghi nhận quý sinh có điểm trung bình cao nhất cho từng môn học:

- Tổng điểm:** Q1
- Ngữ văn:** Q1, Q2
- Toán:** Q4
- Tiếng Anh:** Q1
- Vật lí:** Q1, Q3, Q4
- Hóa học:** Q3, Q4
- Sinh học:** Q1
- Lịch sử:** Q3
- Địa lí:** Q1
- GD&CD/KTPL:** Q1

4.3.2.2 Phân tích tổ hợp điểm cao nhất theo Quý sinh: Phân tích sâu hơn về điểm trung bình của các tổ hợp 3 môn phổ biến cho thấy một số khác biệt nhỏ (được trích xuất từ h3.txt). Mặc dù Q1 chiếm ưu thế ở hầu hết các tổ hợp, đặc biệt là các tổ hợp chứa môn KHXH, nhưng Q4 lại nhỉnh hơn ở một số tổ hợp KHTN:



- **Tổ hợp A00 (Toán, Vật lí, Hóa học):** Q4 có tổng điểm trung bình 3 môn cao nhất (18.23).
- **Tổ hợp D07 (Toán, Hóa học, Tiếng Anh):** Q4 có tổng điểm trung bình 3 môn cao nhất (16.91).
- **Tổ hợp A01 (Toán, Vật lí, Tiếng Anh):** Q1 có tổng điểm trung bình 3 môn cao nhất (17.94).
- **Tổ hợp B00 (Toán, Hóa học, Sinh học):** Q1 có tổng điểm trung bình 3 môn cao nhất (17.55).
- **Tổ hợp C00 (Ngữ văn, Lịch sử, Địa lí):** Q1 có tổng điểm trung bình 3 môn cao nhất (20.61).
- **Tổ hợp C14 (Ngữ văn, Toán, GDCD/KTPL):** Q1 có tổng điểm trung bình 3 môn cao nhất (20.34).
- **Tổ hợp D01 (Ngữ văn, Toán, Tiếng Anh):** Q1 có tổng điểm trung bình 3 môn cao nhất (18.04).
- ...

Những quan sát này, cùng với dữ liệu trong Bảng 13, cung cấp cho kết luận rằng mặc dù Q1 có xu hướng nhỉnh hơn ở nhiều môn xã hội và tổng điểm, sự khác biệt là rất nhỏ và không có ý nghĩa thống kê, như sẽ được trình bày ở phần kiểm định ANOVA.

4.3.2.b Kết quả Kiểm định Giả thuyết (ANOVA)

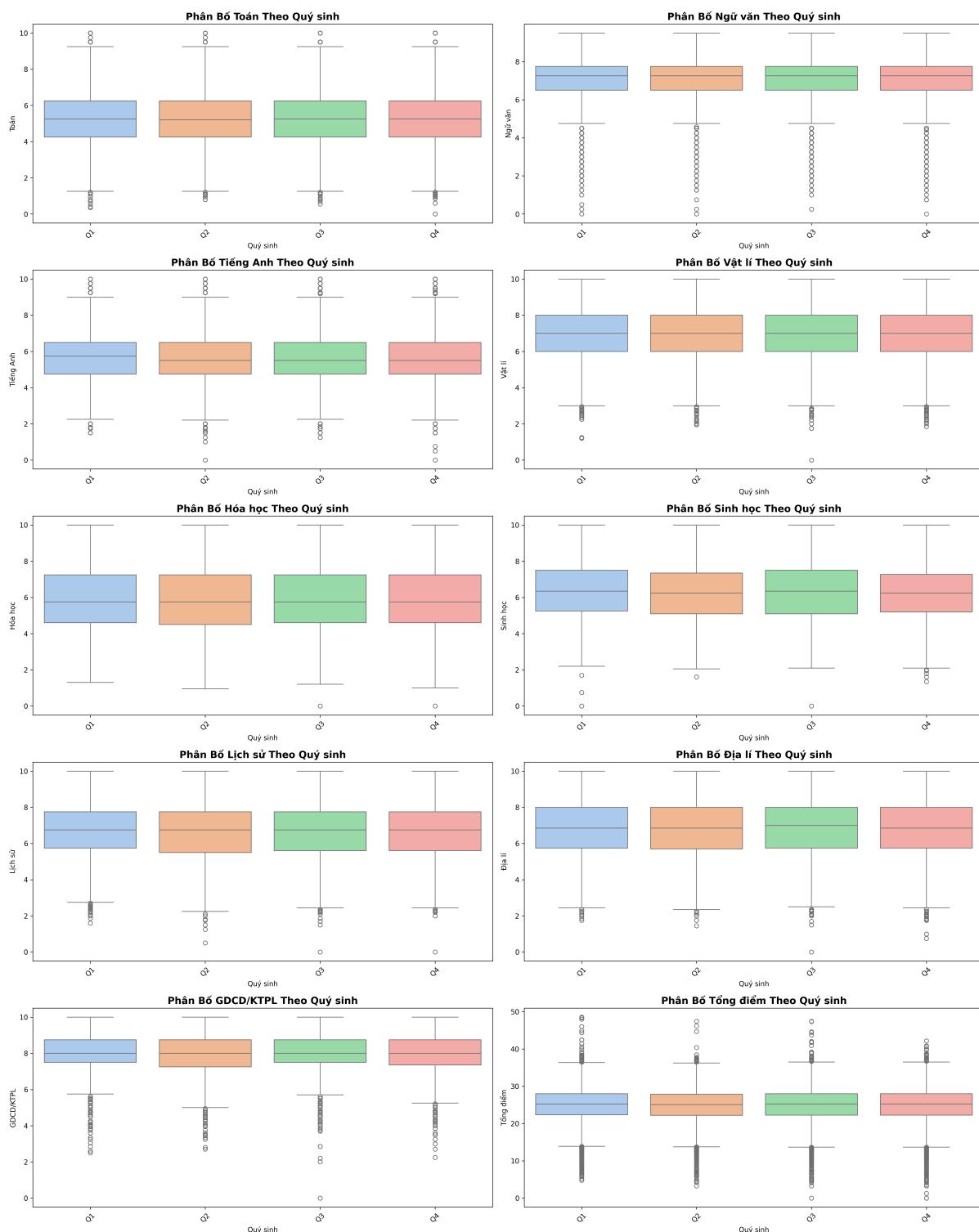
Kết quả kiểm định ANOVA (Bảng 14) cung cấp bằng chứng thống kê rõ ràng.

Bảng 14: Kết quả kiểm định ANOVA so sánh Tổng điểm giữa 4 Quý sinh

Chỉ số	Giá trị
F-statistic	1.8480
p-value	0.136041
Mức ý nghĩa α	0.05
Kết luận	KHÔNG CÓ bằng chứng đủ mạnh để bác bỏ giả thuyết H_0 . ($p = 0.136 > 0.05$)

4.3.2.c Trực quan hóa Phân bố điểm

Biểu đồ hộp trong Hình 10 trực quan hóa kết quả của kiểm định ANOVA.



Hình 10: Phân bố điểm (Boxplot) của 9 môn thi và **Tổng điểm** theo 4 Quý sinh (Q1, Q2, Q3, Q4). Hình ảnh được tạo ra bởi h3.py.

Quan sát Hình 10, đặc biệt là biểu đồ "Phân Bố Tổng điểm" (góc dưới bên phải), có thể thấy rõ:

- Bốn "hộp" (đại diện cho 50% thí sinh ở giữa) của 4 quý gần như **giống nhau** về vị trí (giá trị trung vị - vạch cam) và kích thước (độ phân tán - chiều cao hộp).
- Sự phân bố của các môn học riêng lẻ (Toán, Văn, Anh, v.v.) cũng cho thấy sự tương đồng rất cao giữa 4 nhóm quý sinh.



Trực quan này hoàn toàn ủng hộ kết luận từ kiểm định ANOVA rằng tháng sinh không tạo ra sự khác biệt đáng kể.

4.3.3 Kết luận và Nhận xét

4.3.3.1 Bác bỏ Giả thuyết Hiệu ứng Tuổi Tương đối (RAE) Phát hiện quan trọng nhất từ hướng phân tích này là **Hiệu ứng Tuổi Tương đối (RAE) không tồn tại** trong bối cảnh kỳ thi THPT 2025 tại TP.HCM. Giả thuyết của OECD, vốn được ghi nhận ở nhiều nước phương Tây, đã bị bác bỏ trong môi trường giáo dục Việt Nam dựa trên bộ dữ liệu này. Chênh lệch 0.08 điểm giữa nhóm "già nhất" (Q1) và "trẻ nhất" (Q2, trong trường hợp này) là không có ý nghĩa thống kê ($p = 0.136$).

4.3.3.2 Lý giải cho sự vắng mặt của RAE tại Việt Nam Sự khác biệt này có thể được lý giải bởi các đặc thù của hệ thống giáo dục Việt Nam, như đã được đề cập trong tài liệu phân tích:

- Không phân loại học sinh sớm (No early tracking):** Không giống như một số hệ thống giáo dục phương Tây, nơi học sinh có thể được phân vào các lớp "năng khiếu" hoặc "hỗ trợ" từ rất sớm (ví dụ 6 tuổi), hệ thống của Việt Nam có xu hướng giữ học sinh trong các lớp học đồng nhất. Điều này ngăn chặn lợi thế nhỏ ban đầu về sự trưởng thành (của nhóm sinh sớm) tích tụ thành một lợi thế học thuật lớn theo thời gian.
- Sự linh hoạt của tổ hợp môn:** Kỳ thi cho phép học sinh chọn các tổ hợp môn (KHTN hoặc KHXH) phù hợp với thế mạnh của mình. Nếu một học sinh sinh muộn (Q4) cảm thấy yếu thế ở các môn tự duy trì tương ứng (như Toán, Lý), em đó có thể chọn tổ hợp KHXH (Văn, Sử, Địa) và vẫn có cơ hội đạt tổng điểm rất cao, qua đó làm mất đi bất lợi RAE.
- Học thêm phổ biến:** Văn hóa học thêm rộng rãi tại Việt Nam cho phép những học sinh yếu hơn (bất kể tháng sinh) có cơ hội nhận hỗ trợ bổ sung để bù đắp kiến thức, làm giảm khoảng cách thành tích.
- Bão hòa về sự trưởng thành (Maturity saturation):** Ở độ tuổi 17-18, sự chênh lệch về nhận thức và thể chất do chênh lệch vài tháng tuổi có thể không còn đáng kể như ở cấp tiểu học.

4.3.3.3 Phân tích sâu về Tổ hợp môn (Phát hiện nhỏ) Mặc dù RAE tổng thể không tồn tại, tệp h3.txt và tài liệu phân tích cho thấy một vài xu hướng nhỏ (nhưng không có ý nghĩa thống kê) khi chia theo tổ hợp (Bảng 15):

Bảng 15: So sánh điểm trung bình tổ hợp 3 môn (KHTN vs KHXH) theo Quý sinh

Tổ hợp	Quý 1 (Mean)	Quý 4 (Mean)
A00 (Toán-Lý-Hóa)	18.12	18.23
C00 (Văn-Sử-Địa)	20.61	20.42
C14 (Văn-Toán-GDCD)	20.34	20.01

Có một xu hướng rất nhẹ cho thấy nhóm sinh muộn (Q4) nhỉnh hơn một chút ở tổ hợp KHTN (A00), trong khi nhóm sinh sớm (Q1) lại nhỉnh hơn ở các tổ hợp KHXH (C00, C14). Tuy nhiên, như kết luận chung đã khẳng định, những khác biệt này là quá nhỏ và không đủ cơ sở để đưa ra bất kỳ khuyến nghị nào.

4.3.3.4 Kết luận cho Hướng 3: Nghiên cứu bác bỏ giả thuyết về Hiệu ứng Tuổi Tương đối trong kỳ thi THPT 2025 tại TP.HCM. Khuyến nghị thực tiễn cho học sinh, phụ huynh và nhà trường là **không cần quan tâm đến tháng sinh** như một yếu tố ảnh hưởng đến kết quả học tập hay lựa chọn tổ hợp môn. Sự thành công trong kỳ thi phụ thuộc vào các yếu tố khác như năng lực, phương pháp học và sự lựa chọn tổ hợp môn phù hợp (như đã phân tích ở Hướng 2), chứ không phải ngẫu nhiên về ngày sinh.



4.4 Hướng 4: Phân tích Khám phá: Cung Hoàng Đạo và Kết quả thi

Hướng phân tích này mang tính khám phá, không dựa trên các giả thuyết khoa học đã được thiết lập, nhưng có thể mang lại những góc nhìn thú vị và đổi chiều các niềm tin phổ biến trong văn hóa đại chúng với dữ liệu thực tế.

4.4.1 Phương pháp luận

Để khám phá xem liệu có tồn tại sự khác biệt có ý nghĩa thống kê về kết quả thi giữa 12 nhóm cung hoàng đạo hay không, chúng tôi đã thực hiện các bước phân tích dựa trên tệp h4.py:

- Tạo Đặc trưng (Feature Engineering):** Dựa vào cột ngày sinh của mỗi thí sinh, tiến hành tạo một cột dữ liệu mới là Cung Hoàng Đạo. Quá trình này ánh xạ mỗi khoảng ngày sinh trong năm vào một trong 12 cung hoàng đạo tương ứng.
- Thống kê Mô tả:** Tính toán các chỉ số thống kê (Mean, Std Dev, Count) cho Tổng điểm và điểm trung bình (Mean) cho tất cả 9 môn thi, được phân nhóm theo 12 cung hoàng đạo.
- Kiểm định Giả thuyết (ANOVA):** Sử dụng Phân tích Phương sai một chiều (One-Way ANOVA) để so sánh giá trị trung bình của Tổng điểm giữa 12 nhóm. Giả thuyết không (H_0) là không có sự khác biệt về điểm trung bình giữa các nhóm, với mức ý nghĩa $\alpha = 0.05$.
- Trực quan hóa Dữ liệu:** Sử dụng biểu đồ hộp (Boxplot) để so sánh trực quan sự phân bố điểm của Tổng điểm và 9 môn thi giữa 12 nhóm cung hoàng đạo.

4.4.2 Kết quả Phân tích

4.4.2.a Thống kê mô tả theo Cung Hoàng Đạo

Bảng 16 tóm tắt các chỉ số thống kê chính được trích xuất từ tệp h4.txt. Nhìn lướt qua, sự khác biệt về điểm trung bình giữa các cung là cực kỳ nhỏ. Cung có Tổng điểm trung bình cao nhất là Ma Kết (25.10) và thấp nhất là Cự Giải (24.97). Sự chênh lệch lớn nhất chỉ là 0.13 điểm.

Bảng 16: Thống kê mô tả điểm thi theo Cung Hoàng Đạo (dữ liệu từ h4.txt)

Cung Hoàng Đạo	Tổng điểm			Điểm trung bình (Mean) các môn								
	Mean	Std	Count	Toán	Văn	Anh	Lí	Hóa	Sinh	Sử	Địa	GDCD
Ma Kết	25.10	4.27	10553	5.29	7.07	5.70	6.98	5.98	6.29	6.63	6.83	7.98
Song Ngư	25.07	4.32	9276	5.26	7.08	5.67	6.98	5.94	6.37	6.73	6.86	7.97
Bảo Bình	25.05	4.28	9135	5.28	7.08	5.68	7.00	5.94	6.30	6.70	6.81	7.98
Xử Nữ	25.05	4.29	11017	5.26	7.06	5.67	6.98	5.94	6.39	6.73	6.83	7.94
Bạch Dương	25.04	4.25	9633	5.25	7.08	5.66	6.95	5.91	6.27	6.69	6.82	8.00
Sư Tử	25.04	4.32	10049	5.25	7.05	5.65	6.95	5.95	6.27	6.67	6.84	8.02
Thiên Bình	25.03	4.24	12166	5.25	7.05	5.66	6.99	5.97	6.28	6.68	6.81	7.96
Nhân Mã	25.03	4.29	13625	5.27	7.04	5.69	6.97	5.96	6.32	6.67	6.80	7.94
Song Tử	25.02	4.29	9701	5.25	7.08	5.67	6.99	5.97	6.20	6.62	6.82	7.99
Bọ Cạp	25.02	4.31	13960	5.28	7.06	5.64	6.99	5.95	6.18	6.68	6.78	7.95
Kim Ngưu	25.01	4.27	9873	5.25	7.07	5.66	6.97	5.96	6.32	6.63	6.76	7.97
Cự Giải	24.97	4.29	10160	5.26	7.06	5.68	6.98	5.96	6.29	6.66	6.82	7.92

4.4.2.1 Phân tích điểm trung bình cao nhất theo môn: Dữ liệu từ h4.txt cho thấy sự phân bố "cao nhất" dường như là ngẫu nhiên và bị chia sẻ giữa nhiều cung:

- Tổng điểm:** Ma Kết
- Ngữ văn:** Bạch Dương, Bảo Bình, Song Ngư, Song Tử
- Toán:** Ma Kết
- Tiếng Anh:** Ma Kết



- **Vật lí:** Bảo Bình
- **Hóa học:** Ma Kết
- **Sinh học:** Xử Nữ
- **Lịch sử:** Song Ngư, Xử Nữ
- **Địa lí:** Song Ngư
- **GDCD/KTPL:** Sư Tử

4.4.2.2 Phân tích tổ hợp điểm cao nhất theo Cung Hoàng Đạo: Tương tự, khi xét các tổ hợp 3 môn, không có một cung nào chiếm ưu thế tuyệt đối. Thay vào đó, vị trí dẫn đầu được chia sẻ:

- **Tổ hợp A00 (Toán, Vật lí, Hóa học):** Ma Kết (tổng mean: 18.25)
- **Tổ hợp A01 (Toán, Vật lí, Tiếng Anh):** Ma Kết (tổng mean: 17.97)
- **Tổ hợp B00 (Toán, Hóa học, Sinh học):** Xử Nữ (tổng mean: 17.60)
- **Tổ hợp C00 (Ngữ văn, Lịch sử, Địa lí):** Song Ngư (tổng mean: 20.67)
- **Tổ hợp C14 (Ngữ văn, Toán, GDCD/KTPL):** Ma Kết (tổng mean: 20.34)
- **Tổ hợp C20 (Ngữ văn, Địa lí, GDCD/KTPL):** Sư Tử (tổng mean: 21.92)
- **Tổ hợp D01 (Ngữ văn, Toán, Tiếng Anh):** Ma Kết (tổng mean: 18.06)
- ...

Những chênh lệch nhỏ này, mặc dù thú vị, nhưng không đủ để kết luận về mặt thống kê.

4.4.2.b Kết quả Kiểm định Giả thuyết (ANOVA)

Để xác nhận xem các khác biệt nhỏ quan sát được có ý nghĩa thống kê hay chỉ là ngẫu nhiên, chúng tôi thực hiện kiểm định ANOVA. Kết quả (Bảng 17) là rõ ràng.

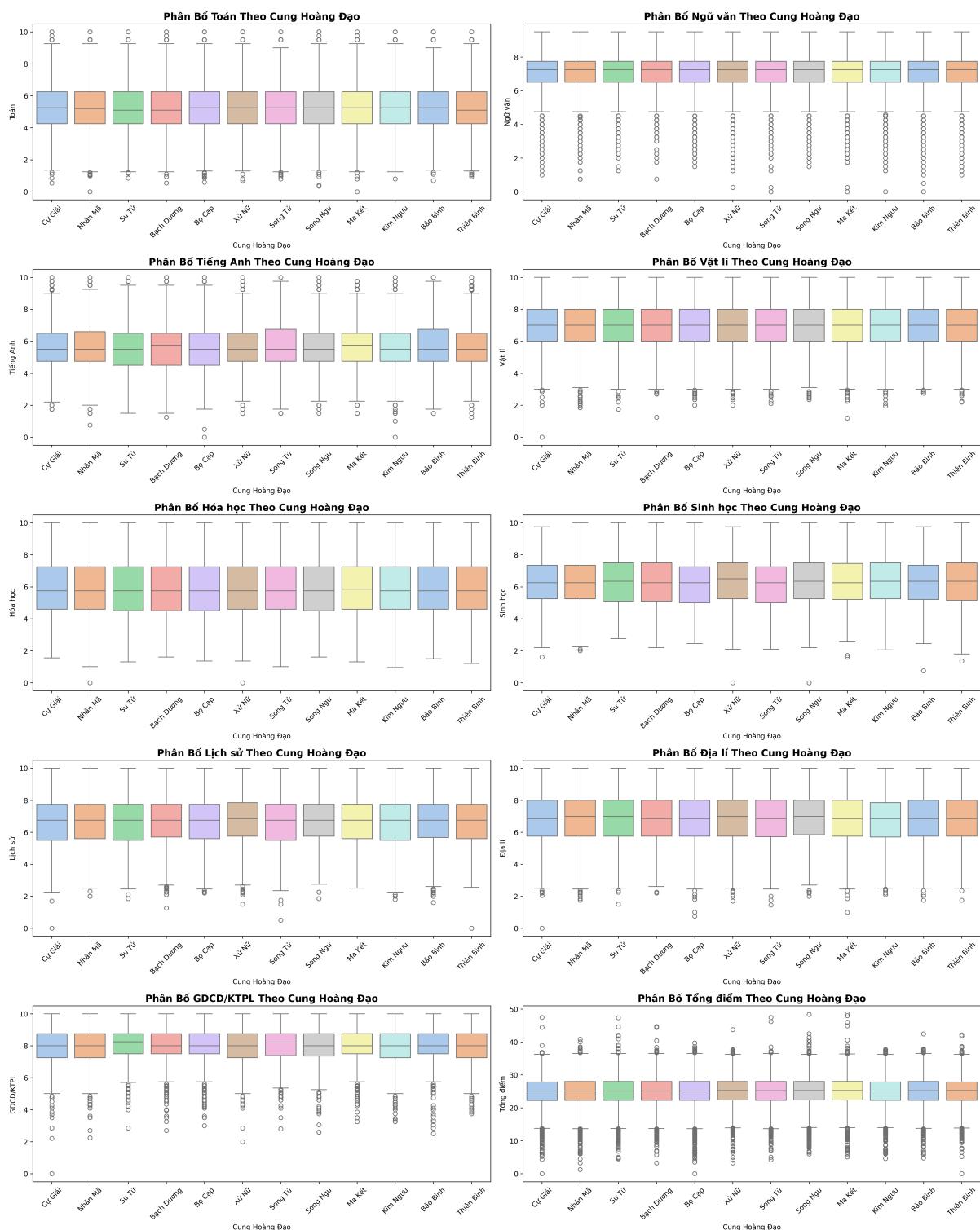
Bảng 17: Kết quả kiểm định ANOVA so sánh Tổng điểm giữa 12 Cung Hoàng Đạo

Chỉ số	Giá trị
F-statistic	0.5592
p-value	0.863106
Mức ý nghĩa α	0.05
Kết luận	KHÔNG CÓ bằng chứng đủ mạnh để bác bỏ giả thuyết H_0 . ($p = 0.863 > 0.05$)

Với giá trị p -value = 0.8631, cao hơn rất nhiều so với mức ý nghĩa 0.05, chúng ta kết luận rằng **không có sự khác biệt có ý nghĩa thống kê** về tổng điểm trung bình giữa các thí sinh thuộc 12 cung hoàng đạo.

4.4.2.c Trực quan hóa Phân bố điểm

Biểu đồ hộp trong Hình 11 cung cấp bằng chứng trực quan mạnh mẽ, cung cấp cho kết luận của ANOVA.



Hình 11: Phân bố điểm (Boxplot) của 9 môn thi và Tổng điểm theo 12 Cung Hoàng Đạo. Hình ảnh được tạo ra bởi h4.py.

Quan sát Hình 11 (đặc biệt là biểu đồ "Phân Bố Tổng điểm" ở góc dưới bên phải), có thể thấy rõ:

- Mười hai "hộp" (đại diện cho 50% thí sinh ở giữa) của 12 cung hoàng đạo gần như **giống hệt nhau**.
- Vị trí của vạch trung vị (màu cam) và chiều cao của các hộp (biểu thị độ phân tán) là tương đồng một cách đáng kinh ngạc trên tất cả các nhóm.



Trực quan này bác bỏ một cách thuyết phục bất kỳ giả định nào về sự khác biệt trong phân bố điểm thi giữa các cung hoàng đạo.

4.4.3 Kết luận và Nhận xét

4.4.3.1 Bác bỏ Giả thuyết về Cung Hoàng Đạo Phát hiện quan trọng nhất từ hướng phân tích này là **không có bất kỳ mối liên hệ nào** có ý nghĩa thống kê giữa cung hoàng đạo của thí sinh và kết quả thi của họ. Kiểm định ANOVA với p -value = 0.8631 đã khẳng định điều này một cách rõ ràng.

4.4.3.2 Diễn giải sự khác biệt nhỏ Mặc dù thống kê mô tả cho thấy Ma Kết (25.10) có điểm trung bình cao hơn một chút so với Cự Giải (24.97), chênh lệch 0.13 điểm này là **hoàn toàn ngẫu nhiên** và không có ý nghĩa thống kê. Các "chiến thắng" nhỏ ở từng môn học hay tổ hợp môn (ví dụ Ma Kết ở tổ hợp A00, Song Ngư ở tổ hợp C00) chỉ đơn thuần là "nhiều" (random noise) trong một bộ dữ liệu lớn.

4.4.3.3 Cảnh báo về Diễn giải (Quan trọng) Cần phải nhấn mạnh rằng phân tích này mang tính **khám phá thuần túy** và **không có cơ sở khoa học**. Cung hoàng đạo là một khái niệm chiêm tinh, không phải là một yếu tố nhân khẩu học hay sinh học có ảnh hưởng đến năng lực nhận thức.

- Khuyến nghị:** Kết quả này **không nên được sử dụng** dưới bất kỳ hình thức nào để tư vấn hướng nghiệp, lựa chọn môn học, hay đánh giá năng lực của học sinh.
- Kết luận cuối cùng:** Nếu lặp lại khảo sát này vào năm 2026, thứ hạng "cao nhất" của các cung hoàng đạo gần như chắc chắn sẽ thay đổi, một lần nữa khẳng định đây chỉ là kết quả của sự ngẫu nhiên.

5 Kết Luận Chung

Kỳ thi THPT 2025 tại TP. Hồ Chí Minh là một sự kiện có ý nghĩa lịch sử, đánh dấu sự chuyển đổi của hệ thống giáo dục Việt Nam từ mô hình tập trung vào kiến thức sang mô hình phát triển năng lực. Phân tích dữ liệu 129,148 thí sinh với các phương pháp thống kê tiên tiến và học máy hiện đại cho phép chúng ta rút ra các kết luận quan trọng:

- Lợi thế rõ nét** ở Toán, Sinh, Tiếng Anh, phản ánh chất lượng giáo dục cao hơn ở TPHCM. Tuy nhiên, một phát hiện đáng ngạc nhiên là sự xuất hiện của "bất lợi rõ thị" ở hai môn KHTN là Vật lí (-0.01 điểm) và Hóa học (-0.11 điểm). Cần tìm hiểu nguyên nhân một cách rõ ràng và có biện pháp cải thiện chất lượng giảng dạy hai môn này tại TPHCM.
- Xu hướng học tập thiên Văn chiếm ưu thế** (75.6%), có thể do đề Văn dễ hơn hoặc năng lực của học sinh, nhưng không hiệu quả - cần khuyến khích cân bằng hoặc phát triển Toán.
- Tổ hợp các môn có Văn (có thể xem là tổ hợp các môn Khoa học xã hội) chiếm đa số:** Tổ hợp đồng thí sinh như D01 (62318 thí sinh), C01 (56975 thí sinh).
- Thí sinh chọn tổ hợp có môn **GDCD/KTPL** lại có điểm số **cao nhất**.
- Hiệu ứng tuổi tương đối không đáng kể** - khác biệt so với các nước phương Tây
- Toán là yếu tố quyết định** - SHAP value 1.578, cao gấp 1.9 lần Văn
- Mô hình dự báo cực hiệu quả** - Random Forest $R^2 = 0.9582$

Kết quả này cung cấp bằng chứng khoa học để các nhà hoạch định chính sách, các trường học, giáo viên, và học sinh ra quyết định hợp lý trong bối cảnh kỳ thi mới.