

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



Xác suất thống kê MT2013

Báo cáo bài tập lớn

Phân tích dữ liệu GPU

Advisor(s): Advisor h

Student(s): Student 1 ID 1

Student 2 ID 2

Student 3 ID 3

HO CHI MINH CITY, NOVEMBER 2025

Mục lục

1	Tổng quan dữ liệu	1
2	Kiến thức nền	3
2.1	Thống kê mô tả và thống kê suy diễn	3
2.2	Các đặc trưng của tổng thể và mẫu	3
2.2.1	Khái niệm	3
2.2.2	Tỷ lệ	3
2.2.3	Trung bình	3
2.2.4	Phương sai, độ lệch chuẩn	4
2.2.5	Các đặc trưng khác	5
2.3	Ước lượng	6
2.4	Kiểm định giả thuyết thống kê	6
2.4.1	Khái niệm chung về kiểm định	6
2.4.2	Quy tắc kiểm định:	7
2.4.3	Các sai lầm trong bài toán kiểm định	7
2.4.4	Các bước thực hiện kiểm định	8
2.5	Các mô hình kiểm định được sử dụng trong báo cáo	8
2.5.1	Bài toán kiểm định trung bình 1 mẫu	8
2.5.2	Bài toán kiểm định 2 mẫu	9
2.5.3	Phân tích phương sai (anova)	9
2.5.4	Phân tích phương sai 1 yếu tố	10
2.5.5	Hồi quy tuyến tính bội	11
3	Tiền xử lý dữ liệu	12
3.1	Đọc dữ liệu vào R	12
3.2	Làm sạch dữ liệu	13

Danh sách hình vẽ

3.1	Hiển thị 5 dòng đầu tiên của dữ liệu sau khi đọc vào R	12
3.2	Tỷ lệ dữ liệu khuyết thiếu trong các đặc trưng	14
3.3	Các cột còn lại sau khi loại bỏ các cột có tỷ lệ NA > 10%	15
3.4	Dữ liệu sau khi làm sạch	15

Danh sách bảng

1.1	Bảng mô tả một vài thông số quan trọng của tập dữ liệu GPU	2
2.1	Công thức của bài toán kiểm định tỷ lệ & trung bình 1 mẫu	8
2.2	Các dạng toán kiểm định 2 mẫu	9
2.3	Bảng tóm tắt ANOVA 1 yếu tố	11

Danh sách đoạn mã

1	Đọc dữ liệu từ file trong R	12
2	Thay thế giá trị rác thành NA	13
3	Loại bỏ các cột có tỷ lệ NA > 10%	13
4	Loại bỏ các cột có tỷ lệ NA > 10%	14
5	Chuẩn hoá các đơn vị đo trong dữ liệu	15



1 Tổng quan dữ liệu

Trong thời đại công nghệ phát triển nhanh chóng, bộ xử lý đồ họa (GPU – Graphics Processing Unit) đã trở thành một trong những thành phần quan trọng nhất của máy tính hiện đại. Ban đầu, GPU được thiết kế với mục đích chính là xử lý hình ảnh và đồ họa trong các trò chơi điện tử, phần mềm thiết kế và dựng phim. Tuy nhiên, cùng với sự tiến bộ của công nghệ, vai trò của GPU đã vượt xa khỏi phạm vi đồ họa thuần túy. Ngày nay, GPU đóng vai trò cốt lõi trong các lĩnh vực như trí tuệ nhân tạo (AI), học sâu (Deep Learning), mô phỏng khoa học, và xử lý dữ liệu lớn. Nhờ vào cấu trúc song song mạnh mẽ với hàng nghìn lõi xử lý, GPU có thể thực hiện hàng loạt phép tính phức tạp cùng lúc, giúp rút ngắn đáng kể thời gian xử lý so với CPU truyền thống. Dưới đây là tập dữ liệu khảo sát các yếu tố về GPU cung cấp những thông tin chi tiết như tốc độ xung nhịp, nhiệt độ tối đa, mức tiêu thụ điện năng, kích thước khuôn chip, ngày phát hành, giá bán, và nhiều đặc trưng kỹ thuật khác. Việc phân tích các dữ liệu này giúp ta hiểu rõ hơn về sự phát triển của công nghệ GPU qua các giai đoạn, từ đó đánh giá xu hướng tiến hóa về hiệu năng, giá thành, và mức độ tối ưu năng lượng. Bên cạnh đó, người nghiên cứu có thể khám phá mối quan hệ giữa giá thành và hiệu suất hoạt động, tìm hiểu xem liệu có nhà sản xuất nào nổi bật trong một phân khúc nhất định hay không. Thông qua việc khai thác và phân tích dữ liệu GPU, chúng ta có thể dự đoán xu hướng của các thế hệ GPU tương lai, phục vụ cho các ứng dụng như học máy, đồ họa máy tính, và tính toán hiệu năng cao.

Tập dữ liệu All_GPUs, bao gồm 34 thông số của 3406 bộ xử lý đồ họa (GPU) khác nhau đến từ 3 nhà sản xuất chính là NVIDIA, AMD và Intel. Dữ liệu được quan sát và thu thập từ trang web [Kaggle](#). Tập dữ liệu này có tương đối nhiều thông số, trong đó có thể kể đến một số thông số được nêu ra trong bảng 1.1 dưới đây:



STT	Tên biến	Đơn vị	Mô tả
1	Manufacturer		Hãng của sản xuất GPU
2	Release_Date		Năm sản xuất của GPU
3	Memory_Bandwidth	Gigabyte/giây (GB/s)	Lượng dữ liệu tối đa mà bộ nhớ GPU có thể truyền tải trong mỗi giây
4	Memory_Speed	MHz	Dộ rộng của bus bộ nhớ, ảnh hưởng đến tốc độ truy cập và hiệu suất của bộ nhớ GPU
5	L2_Cache	KB	Bộ nhớ đệm cấp 2 giúp GPU truy cập nhanh hơn vào dữ liệu được sử dụng thường xuyên, tối ưu hóa hiệu suất
6	Memory_Bus	Bit	Dộ rộng của kênh truyền dữ liệu trong RAM (Memory).
7	Memory		Dung lượng bộ nhớ đồ họa (VRAM) của GPU, quyết định khả năng xử lý và lưu trữ dữ liệu hình

Table 1.1: Bảng mô tả một vài thông số quan trọng của tập dữ liệu GPU

Như đã đề cập ở trên, tập dữ liệu có tổng cộng 34 thông số khác nhau, tuy nhiên nếu liệt kê hết ở bảng 1.1 thì sẽ rất dài. Ngoài ra, trong các thông số đó, có nhiều thông số mang tính kỹ thuật cao và không phổ biến, những dữ liệu này sẽ được xử lý sau để dễ dàng hơn trong việc phân tích và trực quan hóa dữ liệu.



2 Kiến thức nền

2.1 Thống kê mô tả và thống kê suy diễn

Thống kê mô tả (descriptive statistics): là quá trình thu thập, biểu diễn, tổng hợp và xử lý dữ liệu để biến đổi dữ liệu thành thông tin.

Thống kê suy diễn (Inferential statistics): xử lý các thông tin có được từ thống kê mô tả, từ đó đưa ra các cơ sở cho những dự đoán (predictions), dự báo (forecasts) và các ước lượng (estimations).

2.2 Các đặc trưng của tổng thể và mẫu

2.2.1 Khái niệm

Tổng thể thống kê (population): là tập hợp các phần tử thuộc đối tượng nghiên cứu, cần được quan sát, thu thập và phân tích theo một hoặc một số đặc trưng nào đó. Các phần tử tạo thành tổng thể thống kê được gọi là đơn vị tổng thể.

Mẫu (sample): là một số đơn vị được chọn ra từ tổng thể theo một phương pháp lấy mẫu nào đó. Các đặc trưng mẫu được sử dụng để suy rộng ra các đặc trưng của tổng thể nói chung.

Đặc điểm thống kê (dấu hiệu nghiên cứu): là các tính chất quan trọng liên quan trực tiếp đến nội dung nghiên cứu và khảo sát cần thu thập dữ liệu trên các đơn vị tổng thể. Người ta chia làm 2 loại: đặc điểm thuộc tính và đặc điểm số lượng.

2.2.2 Tỷ lệ

Với một tổng thể có N phần tử và M phần tử mang tính chất A nào đó. Tỷ lệ tổng thể (kí hiệu: p) được tính bởi công thức:

$$p = \frac{M}{N}$$

Với một mẫu có n phần tử và có m phần tử mang tính chất A nào đó. Tỷ lệ mẫu (kí hiệu: f hay \bar{p}) được tính bởi công thức:

$$p = f = \frac{m}{n}$$

2.2.3 Trung bình

Trung bình (mean): là đại lượng thường được sử dụng nhất để đo giá trị trung tâm của dữ liệu (Trung bình bị ảnh hưởng bởi các giá trị ngoại lai). Với một tổng thể có N phần tử,



trung bình tổng thể (kí hiệu: μ hay \bar{X}) tính bởi công thức:

$$\mu = \frac{1}{N} \sum_{i=1}^N X_i = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

Với một mẫu có n phần tử, trung bình mẫu (kí hiệu: \bar{x}) tính bởi công thức:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Trong trường hợp X có bảng phân phối tần số như sau:

X	x_1	x_2	x_3	\dots	x_k
Tần số	n_1	n_2	n_3	\dots	n_k

Ta lại có trung bình mẫu tính bởi công thức:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{x_1 n_1 + x_2 n_2 + \cdots + x_k n_k}{n}$$

2.2.4 Phương sai, độ lệch chuẩn

Phương sai (Variance): là trung bình của bình phương độ lệch các giá trị so với trung bình. Phương sai phản ánh độ phân tán hay sự biến thiên của dữ liệu.

Độ lệch chuẩn (Standard deviation): Độ lệch chuẩn (Standard deviation) là căn bậc hai của phương sai. Độ lệch chuẩn dùng để đo sự biến thiên, biểu diễn sự biến thiên xung quanh trung bình và cũng có cùng đơn vị đo với dữ liệu gốc.

Với một tổng thể có N phần tử, phương sai tổng thể (kí hiệu: σ^2) tính bởi công thức:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

Khi đó: σ được gọi là độ lệch chuẩn của tổng thể.

Với một mẫu có n phần tử, phương sai mẫu (kí hiệu: s^2) tính bởi công thức:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Trong trường hợp X có bảng phân phối tần số như sau:

X	x_1	x_2	x_3	\dots	x_k
Tần số	n_1	n_2	n_3	\dots	n_k



Ta lại có phương sai mẫu tính bởi công thức:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^k n_i(x_i - \bar{x})^2$$

Khi đó: s được gọi là độ lệch chuẩn mẫu.

2.2.5 Các đặc trưng khác

Yếu vị (Mode): là giá trị của phần tử có số lần xuất hiện lớn nhất trong mẫu. Yếu vị không bị ảnh hưởng bởi các điểm ngoại lai.

Hệ số biến thiên (Coefficient of variation): đo lường mức độ biến động tương đối của mẫu dữ liệu, được dùng khi người ta muốn so sánh mức độ biến động của các mẫu không cùng đơn vị đo. Đơn vị tính bằng %.

$$CV(\text{tongthe}) = \frac{\sigma}{\mu} \times 100\%$$

$$CV(\text{mau}) = \frac{s}{\bar{x}} \times 100\%$$

Sai số chuẩn (Standard Error): là giá trị đại diện cho độ lệch chuẩn của giá trị trung bình trong tập dữ liệu. Nó phục vụ như một thước đo biến động cho các biến ngẫu nhiên hay độ lượn độ phân tán. Độ phân tán càng nhỏ, dữ liệu càng chính xác.

Trung vị (Median): Giả sử X có N quan sát, sắp các quan sát này theo thứ tự tăng dần. Trung vị là giá trị nằm chính giữa dây số này và chia nó thành 2 phần bằng nhau. Cụ thể:

Giả sử mẫu có kích thước n được sắp xếp tăng dần theo giá trị được khảo sát:

$$x_1 \leq x_2 \leq \dots \leq x_{n-1} \leq x_n$$

Nếu $n = 2k + 1$ (n lẻ) thì trung vị mẫu là giá trị x_{k+1}

Nếu $n = 2k$ (n chẵn) thì trung vị mẫu là giá trị $\frac{x_k + x_{k+1}}{2}$

Trung vị không bị ảnh hưởng bởi các điểm ngoại lai (outliers).

Tứ phân vị (Quartiles): Giá trị trung vị chia mẫu dữ liệu đã sắp thứ tự thành 2 tập có số phần tử bằng nhau. Trung vị của tập dữ liệu nhỏ hơn là Q_1 (gọi là tứ phân vị dưới) và trung vị của tập dữ liệu lớn hơn là Q_3 (gọi là tứ phân vị trên). Q_2 được lấy bằng giá trị trung vị. Độ trải giữa, hay là khoảng tứ phân vị $IQR = Q_3 - Q_1$.

Điểm Outlier: còn gọi là điểm dị biệt, điểm ngoại lệ, điểm ngoại lai.... Đó là các phần tử của mẫu có giá trị nằm ngoài khoảng

$$(Q_1 - 1.5 \times IQR; Q_3 + 1.5 \times IQR)$$



2.3 Ước lượng

Lý thuyết ước lượng là một nội dung trọng tâm trong thống kê và nghiên cứu khoa học, tập trung vào việc xác định giá trị của các tham số (parameters) của quần thể dựa trên những mẫu (samples) được chọn ra từ quần thể đó. Mục tiêu chính của ước lượng là tìm ra các giá trị gần đúng cho những đại lượng đặc trưng của quần thể như trung bình tổng thể (μ), phương sai tổng thể (σ^2), và tỷ lệ phần tử có đặc điểm nhất định trong quần thể (p).

- **Khoảng Tin Cậy (Confidence Interval - CI):** Là một loại ước lượng khoảng được sử dụng để chỉ ra phạm vi mà ta tin rằng tham số của tổng thể nằm trong đó. Khoảng tin cậy thường được xác định bởi hai giới hạn: giới hạn dưới và giới hạn trên. Ví dụ, một khoảng tin cậy 95% cho trung bình tổng thể có thể là (20, 30), nghĩa là ta tin rằng với độ tin cậy 95%, trung bình thực sự của tổng thể nằm trong khoảng từ 20 đến 30.
- **Mức Ý Nghĩa (Significance Level - α):** Là ngưỡng mà ta chọn để quyết định ý nghĩa. Ví dụ mức ý nghĩa thường được chọn ở mức 0.05 - nghĩa là khả năng kết quả quan sát sự khác biệt được nhìn thấy trên số liệu là ngẫu nhiên chỉ là 5%.
- **Độ Tin Cậy (Confidence Level):** Được biểu thị dưới dạng một tỷ lệ phần trăm chỉ mức độ tin tưởng hoặc sự chắc chắn mà khoảng tin cậy ước lượng của chúng ta bao gồm tham số tổng thể thực sự. Ví dụ: nếu ta xây dựng khoảng tin cậy với mức tin cậy 95%, ta tin chắc rằng 95 trên 100 lần ước tính sẽ nằm giữa giá trị trên và giá trị dưới được chỉ định bởi khoảng tin cậy.

$$\gamma = 1 - \alpha$$

Có hai phương pháp ước lượng thường được sử dụng là ước lượng điểm (point estimation) và ước lượng khoảng (interval estimation), tuy nhiên trong phạm vi bài này, nhóm chỉ nhắc đến ước lượng bằng khoảng tin cậy.

2.4 Kiểm định giả thuyết thống kê

2.4.1 Khái niệm chung về kiểm định

Trong thống kê, **kiểm định (hypothesis testing)** là quá trình đánh giá một giả thuyết về dữ liệu để xác định xem liệu có đủ bằng chứng để chấp nhận hay bác bỏ giả thuyết đó. Mục tiêu của kiểm định là đưa ra quyết định dựa trên dữ liệu mẫu có sẵn để rút ra những kết luận về tổng thể.

Quá trình kiểm định thường bắt đầu bằng việc xây dựng hai giả thuyết:



- **Giả thuyết không (null hypothesis, ký hiệu H_0):** là giả thiết về yếu tố cần kiểm định của tổng thể ở trạng thái bình thường, không chịu tác động của các hiện tượng liên quan. Yếu tố trong H_0 phải được xác định cụ thể.
- **Giả thuyết thay thế - giả thuyết đối (alternative hypothesis, ký hiệu H_1):** là một mệnh đề mâu thuẫn với H_0 , H_1 thể hiện xu hướng cần kiểm định.

Miền Bác Bỏ (Rejection Region): là miền số thực thỏa $P(G \in RR | H_0 \text{ đúng}) = \alpha$. α là một số khá bé, thường không quá 10% và được gọi là mức ý nghĩa của kiểm định. Một ký hiệu khác của miền bác bỏ được dùng trong bài: W_α

Miền Chấp Nhận (Acceptance Region): phần bù của miền bác bỏ trong R

Tiêu chuẩn kiểm định: là hàm thống kê $G = G(X_1, X_2, \dots, X_n, \theta_0)$, xây dựng trên mẫu ngẫu nhiên $W = (X_1, X_2, \dots, X_n)$ và tham số θ_0 liên quan đến H_0 ; Điều kiện đặt ra với thống kê G là nếu H_0 đúng thì quy luật phân phối xác suất của G phải hoàn toàn xác định.

2.4.2 Quy tắc kiểm định:

Từ mẫu thực nghiệm, ta tính được một giá trị cụ thể của tiêu chuẩn kiểm định, gọi là **giá trị kiểm định thống kê**:

$$g_{qs} = G(x_1, x_2, \dots, x_n, \theta_0)$$

Theo nguyên lý xác suất bé, biến cỗ $G \in RR$ có xác suất nhỏ nên với 1 mẫu thực nghiệm ngẫu nhiên, nó không thể xảy ra.

Do đó:

- + Nếu $g_{qs} \in RR$ thì bác bỏ H_0 , thừa nhận giả thiết H_1 .
- + Nếu $g_{qs} \notin RR$: ta chưa đủ dữ liệu khẳng định H_0 sai. Vì vậy ta chưa thể chứng minh được H_1 đúng.

2.4.3 Các sai lầm trong bài toán kiểm định

Kết luận của một bài toán kiểm định có thể mắc các sai lầm sau:

- **Sai lầm loại I:** Bác bỏ giả thiết H_0 trong khi H_0 đúng. Xác suất mắc sai lầm này nếu H_0 đúng chính bằng mức ý nghĩa α . Nguyên nhân mắc phải sai lầm loại I thường có thể do kích thước mẫu quá nhỏ, có thể do phương pháp lấy mẫu ...
- **Sai lầm loại II:** Thừa nhận H_0 trong khi H_0 sai, tức là mặc dù thực tế H_1 đúng nhưng giá trị thực nghiệm g_{qs} không thuộc RR.



		Tình huống
Quyết định	H_0 đúng	H_0 sai
Bắc bở H_0	Sai lầm loại I. Xác suất = α	Quyết định đúng
Không bác bỏ H_0	Quyết định đúng	Sai lầm loại II. Xác suất = β

2.4.4 Các bước thực hiện kiểm định

- Phát biểu giả thuyết và đối thuyết của bài toán.
- Tính giá trị thống kê kiểm định (tiêu chuẩn kiểm định) cho bài toán.
- Xác định miền bác bỏ tốt nhất cho bài toán.
- Dựa ra kết luận.

2.5 Các mô hình kiểm định được sử dụng trong báo cáo

2.5.1 Bài toán kiểm định trung bình 1 mẫu

Dạng bài	Phân bố của tổng thể	Giả thiết H_0	Giả thiết đối H_1	Miền bác bỏ RR (miền bác bỏ H_0 với mức ý nghĩa α)	Hàm thống kê kiểm định (Tiêu chuẩn kiểm định)
Kiểm định tỷ lệ 1 mẫu	* X có phân phối Không -một. * $n \geq 30$. (1)	$p = p_0$	$p \neq p_0$	$(-\infty; -z_{\alpha/2}) \cup (z_{\alpha/2}; +\infty)$	$Z_{qs} = \frac{f - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx N(0, 1)$
			$p < p_0$	$(-\infty; -z_\alpha)$	
			$p > p_0$	$(z_\alpha; +\infty)$	
Kiểm định trung bình 1 mẫu	* X có phân phối chuẩn. * Đã biết σ^2 . (2a)	$\mu = \mu_0$	$\mu \neq \mu_0$	$(-\infty; -z_{\alpha/2}) \cup (z_{\alpha/2}; +\infty)$	$Z_{qs} = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$
			$\mu < \mu_0$	$(-\infty; -z_\alpha)$	
			$\mu > \mu_0$	$(z_\alpha; +\infty)$	
Kiểm định trung bình 1 mẫu	* X có phân phối chuẩn. * Chưa biết σ^2 . (2b)	$\mu = \mu_0$	$\mu \neq \mu_0$	$(-\infty; -t_{\alpha/2;(n-1)}) \cup (t_{\alpha/2;(n-1)}; +\infty)$	$T_{qs} = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \sim T_{(n-1)}$
			$\mu < \mu_0$	$(-\infty; -t_{\alpha;(n-1)})$	
			$\mu > \mu_0$	$(t_{\alpha;(n-1)}; +\infty)$	
	* X có phân phối tùy ý. * $n \geq 30$. (2c) X không có giả thiết PPC	$\mu = \mu_0$	$\mu \neq \mu_0$	$(-\infty; -z_{\alpha/2}) \cup (z_{\alpha/2}; +\infty)$	$Z_{qs} = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \approx N(0, 1)$ Nếu chưa biết σ thì thay bởi s
			$\mu < \mu_0$	$(-\infty; -z_\alpha)$	
			$\mu > \mu_0$	$(z_\alpha; +\infty)$	

Table 2.1: Công thức của bài toán kiểm định tỷ lệ & trung bình 1 mẫu



2.5.2 Bài toán kiểm định 2 mẫu

Phân bố của tổng thê	GT H_0	GT H_1	Miền bác bỏ RR	T/chuẩn kiểm định
* 2 mẫu độc lập * X_1, X_2 có pp chuẩn. * Đã biết σ_1^2 và σ_2^2 (4a)	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$(-\infty; -z_{\alpha/2}) \cup (z_{\alpha/2}; +\infty)$	$Z_{qs} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
		$\mu_1 < \mu_2$	$(-\infty; -z_\alpha)$	
		$\mu_1 > \mu_2$	$(z_\alpha; +\infty)$	
* 2 mẫu độc lập * X_1, X_2 có pp chuẩn * Chưa biết σ_1^2, σ_2^2 ; $\sigma_1^2 = \sigma_2^2$ (4b)	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$(-\infty; -t_{\alpha/2;(n_1+n_2-2)}) \cup (t_{\alpha/2;(n_1+n_2-2)}; +\infty)$	$T_{qs} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$
		$\mu_1 < \mu_2$	$(-\infty; -t_{\alpha;(n_1+n_2-2)})$	
		$\mu_1 > \mu_2$	$(t_{\alpha;(n_1+n_2-2)}; +\infty)$	
* 2 mẫu độc lập * X_1, X_2 có pp chuẩn * Chưa biết σ_1^2, σ_2^2 ; $\sigma_1^2 \neq \sigma_2^2$ (4c)	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$(-\infty; -t_{\alpha/2;(\nu)}) \cup (t_{\alpha/2;(\nu)}; +\infty)$	$T_{qs} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
		$\mu_1 < \mu_2$	$(-\infty; -t_{\alpha;(\nu)})$	
		$\mu_1 > \mu_2$	$(t_{\alpha;(\nu)}; +\infty)$	
* 2 mẫu độc lập * X_1, X_2 có pp tùy ý * Mẫu lớn: $n_1, n_2 \geq 30$ * Đã biết hoặc chưa biết σ_1^2, σ_2^2 (4d)	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$(-\infty; -z_{\alpha/2}) \cup (z_{\alpha/2}; +\infty)$	$Z_{qs} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
		$\mu_1 < \mu_2$	$(-\infty; -z_\alpha)$	
		$\mu_1 > \mu_2$	$(z_\alpha; +\infty)$	
* 2 mẫu phụ thuộc tương ứng theo cặp * X_1, X_2 có pp chuẩn * Chưa biết σ_D^2 (4e)	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$(-\infty; -t_{\alpha/2;(n-1)}) \cup (t_{\alpha/2;(n-1)}; +\infty)$	$T_{qs} = \frac{\bar{X}_D - \mu_0}{s_D} \sqrt{n}$
		$\mu_1 < \mu_2$	$(-\infty; -t_{\alpha;(n-1)})$	
		$\mu_1 > \mu_2$	$(t_{\alpha;(n-1)}; +\infty)$	
* 2 mẫu phụ thuộc tương ứng theo cặp * 2 mẫu lớn: $n \geq 30$ * Đã biết hoặc chưa biết σ_D^2 (4f)	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$(-\infty; -z_{\alpha/2}) \cup (z_{\alpha/2}; +\infty)$	$Z_{qs} = \frac{\bar{X}_D - \mu_0}{\sigma_D} \sqrt{n}$
		$\mu_1 < \mu_2$	$(-\infty; -z_\alpha)$	
		$\mu_1 > \mu_2$	$(z_\alpha; +\infty)$	

Table 2.2: Các dạng toán kiểm định 2 mẫu

2.5.3 Phân tích phương sai (anova)

Phân tích phương sai là một mô hình dùng để xem xét sự biến động của một biến ngẫu nhiên định lượng X chịu tác động trực tiếp của một hay nhiều yếu tố nguyên nhân (định tính).



Được làm hai loại là phân tích phương sai 1 yếu tố và phân tích phương sai 2 yếu tố.

2.5.4 Phân tích phương sai 1 yếu tố

Giả thiết

- Các tổng thể có phân phối chuẩn $N(\mu_i; \sigma_i^2)$, $i = 1, 2, \dots, k$ với k là tổng thể (thông thường $k \geq 3$).
- Phương sai các tổng thể chưa biết nhưng được giả định là bằng nhau ($\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$).
- Các mẫu quan sát (từ k tổng thể) được lấy độc lập.

Các bước thực hiện

Bước 1: Đặt giả thiết kiểm định

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k,$$

$$H_1 : \exists \mu_i \neq \mu_j \quad (i \neq j)$$

Bước 2: Tính trung bình mẫu của các nhóm $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ theo công thức:

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}, \quad (i = 1, 2, \dots, k)$$

Bước 3: Tính tổng các bình phương lệch (tổng bình phương):

$$\begin{aligned} SSW &= \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \\ SSB &= \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2 \\ SST &= SSW + SSB = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 \end{aligned}$$

Bước 4: Tính các phương sai:

$$MSW = \frac{SSW}{N - k}, \quad MSB = \frac{SSB}{k - 1}$$

Bước 5: Tính thống kê kiểm định (tiêu chuẩn kiểm định, giá trị quan sát):

$$F = \frac{MSB}{MSW}$$

Bước 6: Xác định miền bác bỏ của bài toán:

$$RR = (F_{\alpha; k-1; N-k}; +\infty)$$

Tìm giá trị $F_{\alpha; k-1; N-k}$ tra bảng Fisher mức ý nghĩa α và cột $k - 1$ và dòng $N - k$.

Bước 7: Đưa ra kết luận:



- Nếu $F > F_{\alpha;k-1;N-k} \iff F \in RR \Rightarrow$ Bác bỏ H_0 , chấp nhận H_1
- Nếu $F < F_{\alpha;k-1;N-k} \iff F \notin RR \Rightarrow$ không bác bỏ H_0 (chưa bác bỏ được H_0 , chấp nhận H_0)

Table 2.3: Bảng tóm tắt ANOVA 1 yếu tố

Nguồn của sự biến thiên	SS	df	MS	F
Giữa các nhóm	SSB	k-1	MSB	$F = \frac{MSB}{MSW}$
Trong từng nhóm	SSW	N-k	MSW	
Toàn bộ	SST	N-1		

2.5.5 Hồi quy tuyến tính bội

Khái niệm: Hồi quy tuyến tính là một kỹ thuật phân tích dữ liệu dự đoán giá trị của dữ liệu không xác định bằng cách sử dụng một giá trị dữ liệu liên quan và đã biết khác.

Bài toán phân tích hồi quy là bài toán nghiên cứu mối liên hệ phụ thuộc của một biến (gọi là biến phụ thuộc) vào một hay nhiều biến khác (gọi là các biến độc lập), với ý tưởng ước lượng được giá trị trung bình (tổng thể) của biến phụ thuộc theo giá trị của các biến độc lập, dựa trên mẫu được biết trước.

3 Tiết xử lý dữ liệu

3.1 Đọc dữ liệu vào R

Ta đọc dữ liệu từ file dữ liệu đã cho dưới định dạng CSV vào R bằng hàm `read.csv()` như sau:

Listing 1: Đọc dữ liệu từ file trong R

```
1 df  <- read.csv("./data_sets/All_GPUs.csv")  
2 head(df, 5)
```

Sau khi đọc dữ liệu xong, ta có thể sử dụng hàm `head()` để hiển thị 5 dòng đầu tiên của dữ liệu nhằm kiểm tra xem dữ liệu đã được đọc đúng chưa (Hình 3.1).

```

> df <- read.csv("./data_sets/All_GPUs.csv")
> head(df, 5)
#> #>   Architecture Best_Resolution Boost_Clock Core_Speed DVI_Connection Dedicated Direct_X
#> 1   Tesla G92b                      738 MHz        2 Yes    DX 10.0
#> 2   R600 XT      1366 x 768          \n-           2 Yes    DX 10
#> 3   R600 PRO     1366 x 768          \n-           2 Yes    DX 10
#> 4   RV630        1024 x 768          \n-           2 Yes    DX 10
#> 5   RV630        1024 x 768          \n-           2 Yes    DX 10
#> #>   DisplayPort_Connection HDMI_Connection Integrated L2_Cache Manufacturer Max_Power Memory Memory_Bandwidth
#> 1   NA             0 No            0KB Nvidia 141 Watts 1024 MB 64GB/sec
#> 2   NA             0 No            0KB AMD 215 Watts 512 MB 106GB/sec
#> 3   NA             0 No            0KB AMD 200 Watts 512 MB 51.2GB/sec
#> 4   NA             0 No            0KB AMD 256 MB 36.8GB/sec
#> 5   NA             0 No            0KB AMD 45 Watts 256 MB 22.4GB/sec
#> #>   Memory_Bus Memory_Speed Memory_Type Name Notebook_GPU Open_GL
#> 1   256 Bit      1000 MHz   GDDR3 GeForce GTS 150 No 3.3
#> 2   512 Bit      828 MHz   GDDR3 Radeon HD 2900 XT 512MB No 3.1
#> 3   256 Bit      800 MHz   GDDR3 Radeon HD 2900 Pro No 3.1
#> 4   128 Bit      1150 MHz  GDDR4 Radeon HD 2600 XT Diamond Edition No 3.3
#> 5   128 Bit      700 MHz   GDDR3 Radeon HD 2600 XT No 3.1
#> #>   PSU Pixel_Rate Power_Connector Process RDPs Release_Date Release_Price Resolution_WxH
#> 1   450 Watt & 38 Amps 12 GPixel/s None 55nm 16 \n01-Mar-2009 2560x1600
#> 2   550 Watt & 35 Amps 12 GPixel/s None 80nm 16 \n14-May-2007 2560x1600
#> 3   550 Watt & 35 Amps 10 GPixel/s None 80nm 16 \n07-Dec-2007 2560x1600
#> 4   3 GPixel/s      3 GPixel/s   None 65nm 4 \n01-Jul-2007 2560x1600
#> 5   400 Watt & 25 Amps 3 GPixel/s None 65nm 4 \n28-Jun-2007 2560x1600
#> #>   SLI_Crossfire Shader TMUs Texture_Rate VGA_Connection
#> 1   Yes 4 64 47 GTexel/s 0
#> 2   Yes 4 16 12 GTexel/s 0
#> 3   Yes 4 16 10 GTexel/s 0
#> 4   Yes 4 8 7 GTexel/s 0
#> 5   Yes 4 8 6 GTexel/s 0

```

Figure 3.1: Hiển thị 5 dòng đầu tiên của dữ liệu sau khi đọc vào R

Như vậy, ta đã hoàn thành việc đọc dữ liệu từ file CSV vào R và có thể tiến hành các bước tiền xử lý dữ liệu tiếp theo.



3.2 Làm sạch dữ liệu

Vì trong file dữ liệu ban đầu, có thể tồn tại các giá trị bị thiếu (NA) hoặc các giá trị không hợp lệ, ta cần thực hiện các bước làm sạch dữ liệu để đảm bảo tính chính xác và độ tin cậy của phân tích sau này.

Trước tiên, ta sẽ phải thay thế tất cả các giá trị rác, không hợp lệ thành giá trị *NA* trong R. Ví dụ, nếu một ô bất kì có giá trị là chuỗi rỗng "" hoặc ký tự đặc biệt như "N/A", ta sẽ thay thế chúng bằng *NA* như sau:

```

1 df <- df %>%
2   mutate(across(where(is.character), trimws))
3 df[df == ""] <- NA
4 df[df == "N/A"] <- NA
5 df[df == "NA "] <- NA
6 df[df == "-"] <- NA
7 df[df == "Unknown Release Date"] <- NA
8 # Chi lay nam san xuat, khong lay ngay cu the
9 df$Release_Date <- as.Date(df$Release_Date, format = "%d-%b-%Y")
10 df$Release_Date <- format(df$Release_Date, "%Y")

```

Listing 2: Thay thế giá trị rác thành NA

Sau khi thay thế các giá trị rác, nhóm nhận thấy rằng có rất nhiều yếu tố có số lượng *NA* lớn, điều này buộc nhóm phải lựa chọn giữa loại bỏ và chuẩn hóa. Trong nội dung bài báo cáo này, nhóm sẽ loại bỏ những đặc điểm (cột) có số lượng giá trị *NA* vượt quá 10% tổng số dòng dữ liệu. Để thực hiện việc này, ta có thể sử dụng đoạn mã sau:

```

1 # Dem so luong gia tri NA trong moi cot
2 missing_counts = freq.na(df)
3 # Ve do thi ty le du lieu khuyet
4 ggplot(missing_counts, aes(x = rownames(missing_counts), y =
  missing_counts[,2], )) +
  geom_bar(stat = "identity", fill = "cyan") +
  geom_text(aes(label = paste0(missing_counts[,2], "%")), vjust =
    -0.5, size = 2) +
  labs(title = "Missing rate", x = "Feature", y = "Rate (%)") +
  theme_minimal() +
  theme(axis.text.x = element_text(
    size = 10,

```

```

11    angle = 90,
12    hjust = 1
13  ))

```

Listing 3: Loại bỏ các cột có tỷ lệ NA > 10%

Kết quả trực quan hóa tỷ lệ dữ liệu khuyết thiêU được thể hiện trong Hình 3.2.

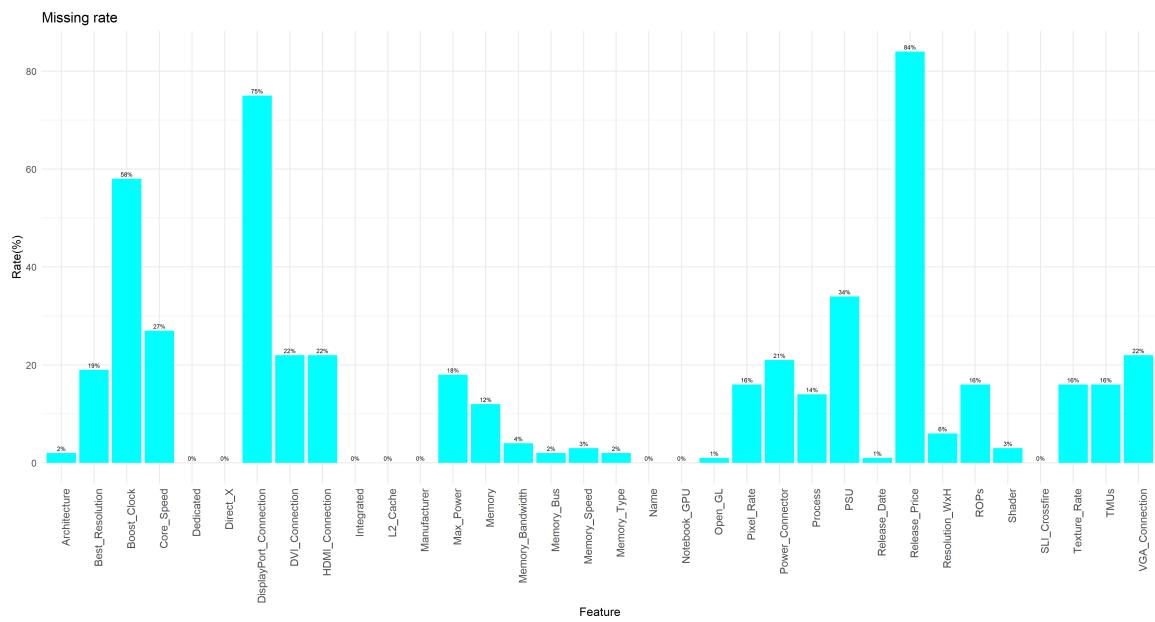


Figure 3.2: Tỷ lệ dữ liệu khuyết thiêU trong các đặc trưng

Tiếp theo sau đó, ta sẽ loại bỏ các cột có tỷ lệ giá trị NA vượt quá 10% tổng số dòng dữ liệu như sau:

```

1 missing_counts_df <- data.frame(
2   feature = rownames(missing_counts),
3   percent = missing_counts[,2]
4 )
5 cols_to_keep <- missing_counts_df$feature[missing_counts_df$
6   percent <= 10 & missing_counts_df$feature != "Architecture" &
7   missing_counts_df$feature != "Name"]
8 df_filtered <- df[, cols_to_keep, drop = FALSE]
9 head(df_filtered, 5)
10 df_filtered <- na.omit(df_filtered)

```

Listing 4: Loại bỏ các cột có tỷ lệ NA > 10%



Bằng câu lệnh `print(names(df_filtered))`, ta có thể kiểm tra lại các cột còn lại sau khi đã loại bỏ các cột có tỷ lệ giá trị *NA* vượt quá 10% (Hình 3.3).

```
> print(names(df_filtered))
[1] "Resolution_WxH"   "Memory_Bandwidth" "Shader"          "Memory_Speed"    "Memory_Bus"
[6] "Memory_Type"     "Open_GL"        "Release_Date"   "Dedicated"      "Integrated"
[11] "Direct_X"        "L2_Cache"       "Manufacturer"  "Notebook_GPU"  "SLI_Crossfire"
```

Figure 3.3: Các cột còn lại sau khi loại bỏ các cột có tỷ lệ NA > 10%

Mặc dù đã làm sạch dữ liệu bằng cách loại bỏ các cột có tỷ lệ khuyết hoặc số lượng giá trị *NA* cao, vẫn còn một yếu tố khiến việc phân tích dữ liệu trở nên khó khăn, đó là các đơn vị đo, do đó ta cần chuẩn hoá các đơn vị đo trong dữ liệu bằng cách loại bỏ chúng.

```
1 # Chuan hoa don vi do trong cac cot
2 remove_unit_cols <- c("Memory_Bandwidth", "Memory_Speed", "Memory_Bus",
3                         "Direct_X")
4 main_df <- df_filtered
5 main_df[remove_unit_cols] <- lapply(df_filtered[remove_unit_cols],
6                                     function(x) {
7                                         as.numeric(gsub("[^0-9.]", "", x))
8                                     })
9 clean_cache <- function(x) {
10     main <- as.numeric(sub("KB.*", "", x))           # 2304
11     mult      <- as.numeric(sub(".*\\"\\(x([0-9]+)\\)\\)", "\\\\$1", x))  #
12     2
13     if (is.na(mult)) mult <- 1
14     return(main * mult)
15 }
16 main_df$L2_Cache <- sapply(main_df$L2_Cache, clean_cache)
```

Listing 5: Chuẩn hoá các đơn vị đo trong dữ liệu

Dữ liệu đã được làm sạch được lưu vào biến `main_df` và được hiện trong Hình 3.4.

```
> head(main_df, 5)
  Manufacturer Release_Date Memory_Bandwidth Memory_Speed L2_Cache Open_GL Memory_Type Resolution_WxH
1      Nvidia    2009            64.0        1000         0     3.3      GDDR3  2560x1600
2       AMD     2007           106.0        828         0     3.1      GDDR3  2560x1600
3       AMD     2007            51.2        800         0     3.1      GDDR3  2560x1600
4       AMD     2007            36.8       1150         0     3.3      GDDR4  2560x1600
5       AMD     2007            22.4        700         0     3.1      GDDR3  2560x1600
  Direct_X Shader Memory_Bus Dedicated Integrated Notebook_GPU SLI_Crossfire
1       10      4      256     Yes     No      No      Yes
2       10      4      512     Yes     No      No      Yes
3       10      4      256     Yes     No      No      Yes
4       10      4      128     Yes     No      No      Yes
5       10      4      128     Yes     No      No      Yes
```

Figure 3.4: Dữ liệu sau khi làm sạch