ĐẠI 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

Danh sách hình vẽ

Danh sách bảng

Danh sách đoạn mã

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 hoạ (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	Độ 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	Độ 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 sử lý sau để dễ dàng hơn trong việc phân tích và trực quan hoá 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:

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

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	...	x_k
Tần số	n_1	n_2	n_3	...	n_k

Ta lại có trung bình mẫu tính bởi công thức:

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

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:

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

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:

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

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

X	x_1	x_2	x_3	...	x_k
Tần số	n_1	n_2	n_3	...	n_k

Ta lại có phương sai mẫu tính bởi công thức:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2 = \frac{1}{n-1}[n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \cdots + n_k(x_k - \bar{x})^2]$$

Khi đó: s được gọi là độ lệch 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.

$$SE(\text{tongthe}) = \frac{\sigma}{\sqrt{N}}$$

$$SE(\text{mau}) = \frac{s}{\sqrt{n}}$$

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 dãy 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:

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

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).

- **Ước lượng điểm (Point Estimation):** là dùng một tham số thống kê mẫu đơn lẻ để ước lượng giá trị tham số của tổng thể.

$$\mu \approx \bar{x}$$
$$\sigma^2 \approx s^2$$
$$p \approx f$$

- **Ước lượng bằng khoảng tin cậy (Interval Estimation):** -Ước lượng bằng khoảng tin cậy chính là tìm ra khoảng ước lượng $(G_1; G_2)$ cho tham số θ trong tổng thể sao cho ứng với độ tin cậy (confidence) bằng γ cho trước, $P(G_1 < \theta < G_2) = \gamma$.

2.3.1 Ước lượng điểm (Point Estimation)

Một **ước lượng (estimator)** của một tham số (của tổng thể): là một biến ngẫu nhiên có giá trị phụ thuộc vào thông tin của mẫu, giá trị của nó là một xấp xỉ cho tham số chưa biết của tổng thể. Một giá trị cụ thể của biến ngẫu nhiên này gọi là một **giá trị ước lượng điểm**.

Xét đại lượng ngẫu nhiên X có phân phối $F(x; \theta)$ với tham số θ chưa biết.

Chọn một mẫu ngẫu nhiên cỡ n từ $X_1, X_2, \ldots, X_n$.

Thống kê $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$ gọi là một ước lượng điểm cho θ.

Với một mẫu cụ thể $(x_1, x_2, \ldots, x_n)$, ta gọi $\hat{\theta} = h(x_1, x_2, \ldots, x_n)$ là một giá trị ước lượng điểm cụ thể cho θ.

2.4 Ước lượng bằng khoảng tin cậy (Interval Estimation)

Cho tham số θ của tổng thể và $X_1, X_2, \ldots, X_n$ là các quan sát ngẫu nhiên. Ta gọi khoảng (c, d) là khoảng ước lượng (hay khoảng tin cậy) của tham số θ với độ tin cậy γ nếu: $P(c < \theta < d) = \gamma$. Có thể nói, độ tin cậy γ cho khoảng ước lượng của tham số θ chính là xác suất để ta đúng khi ước lượng tham số θ bằng khoảng (c, d). Ngược lại, xác suất mà ta cho phép sai khi ước lượng θ được gọi là mức ý nghĩa. Kí hiệu là α. Ta có $\alpha + \gamma = 1$.

Xác định khoảng ước lượng đối xứng của trung bình tổng thể dựa vào một mẫu đã cho, với kích thước là n, trung bình mẫu là $\bar{x}$ và phương sai mẫu là s^2 hoặc phương sai tổng thể là σ^2. Mục tiêu là xác định một khoảng ước lượng đối xứng xung quanh μ của mẫu với một mức độ tin cậy cụ thể.

Để giải quyết vấn đề này, ta cần đi đến việc xác định epsilon (ϵ) - sai số ước lượng, dựa trên các thông tin đã biết về mẫu. Khoảng tin cậy sẽ được biểu diễn bằng khoảng $\bar{x} \pm \epsilon$. Tùy thuộc vào giả định về phân phối của dữ liệu và các thông tin đã biết về phương sai, cách tính ϵ sẽ thay đổi như sau:

Dạng	Giả định	Loại	Ngưỡng sai số	Khoảng tin cậy
Tỷ lệ	$n \geq 30$	Đối xứng	$\epsilon = z_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}$	$f - \epsilon < p < f + \epsilon$
		Bên phải	$\epsilon = z_{\alpha}\sqrt{\frac{f(1-f)}{n}}$	$0 < p < f + z_{\alpha}\sqrt{\frac{f(1-f)}{n}}$
		Bên trái	$\epsilon = z_{\alpha}\sqrt{\frac{f(1-f)}{n}}$	$f - z_{\alpha}\sqrt{\frac{f(1-f)}{n}} < p < 1$
Trung bình	$X \sim N(\mu, \sigma^2)$ Đã biết σ	Đối xứng	$\epsilon = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\bar{x} - \epsilon < \mu < \bar{x} + \epsilon$
		Bên phải	$\epsilon = z_{\alpha}\frac{\sigma}{\sqrt{n}}$	$-\infty < \mu < \bar{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}}$
		Bên trái	$\epsilon = z_{\alpha}\frac{\sigma}{\sqrt{n}}$	$\bar{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}} < \mu < \infty$
	$X \sim N(\mu, \sigma^2)$ Chưa biết σ	Đối xứng	$\epsilon = t_{\alpha/2;n-1}\frac{s}{\sqrt{n}}$	$\bar{x} - \epsilon < \mu < \bar{x} + \epsilon$
		Bên phải	$\epsilon = t_{\alpha;n-1}\frac{s}{\sqrt{n}}$	$-\infty < \mu < \bar{x} + t_{\alpha;n-1}\frac{s}{\sqrt{n}}$
		Bên trái	$\epsilon = t_{\alpha;n-1}\frac{s}{\sqrt{n}}$	$\bar{x} - t_{\alpha;n-1}\frac{s}{\sqrt{n}} < \mu < \infty$
Phân phối tùy ý, mẫu lớn ($n \geq 30$). Nếu chưa biết σ thì thay bằng s		Đối xứng	$\epsilon = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\bar{x} - \epsilon < \mu < \bar{x} + \epsilon$
		Bên phải	$\epsilon = z_{\alpha}\frac{\sigma}{\sqrt{n}}$	$-\infty < \mu < \bar{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}}$
		Bên trái	$\epsilon = z_{\alpha}\frac{\sigma}{\sqrt{n}}$	$\bar{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}} < \mu < \infty$

Table 2.1: Phân phối XS của trung bình mẫu và tỷ lệ mẫu

2.5 Kiểm định giả thuyết thống kê

2.5.1 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 H0 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 $G = G(X_1, X_2, \ldots, X_n, \theta_0)$, xây dựng trên mẫu ngẫu nhiên $W = (X_1, X_2, \ldots, X_n)$ 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.

3 Tiền 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

```
df <- read.csv("./data_sets/All_GPUs.csv")
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 ROPs   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           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:

```
df <- df %>%
  mutate(across(where(is.character), trimws))
df[df == ""] <- NA
df[df == "N/A"] <- NA
df[df == "NA"] <- NA
df[df == "-"] <- NA
df[df == "Unknown Release Date"] <- NA
# Chi lay nam san xuat, khong lay ngay cu the
df$Release_Date <- as.Date(df$Release_Date, format = "%d-%b-%Y")
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 hoá. 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:

```
# Dem so luong gia tri NA trong moi cot
missing_counts = freq.na(df)
# Ve do thi ty le du lieu khuyet
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,
```

```
    angle = 90,
    hjust = 1
  ))
```

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

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

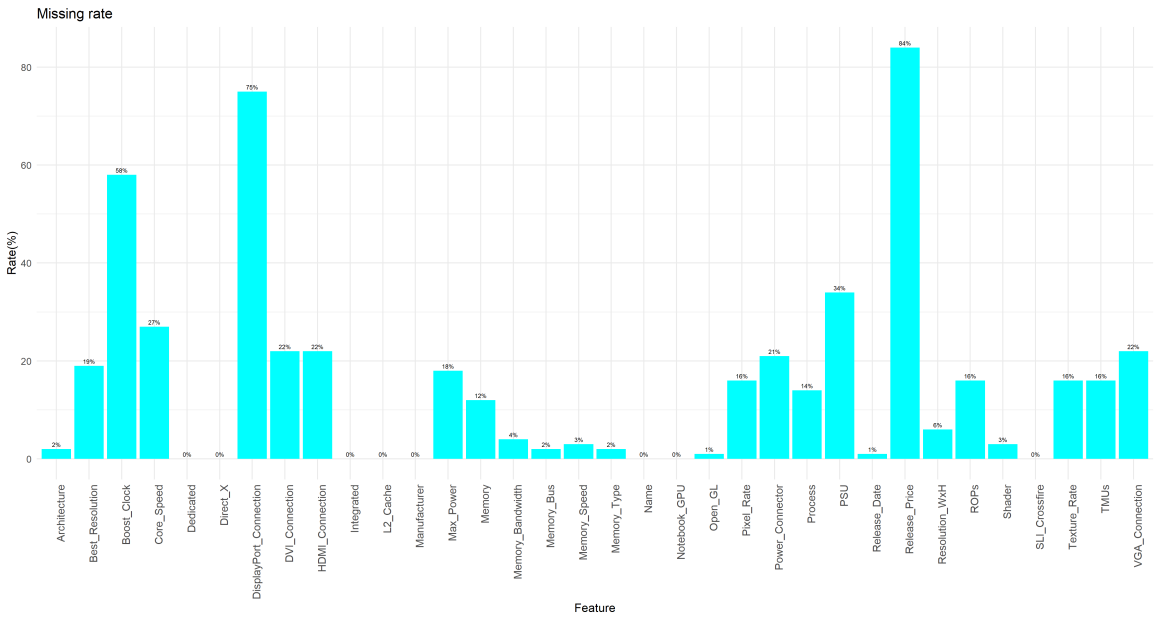

Figure 3.2: Tỷ lệ dữ liệu khuyết thiếu trong các đặc trưng

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:

```
missing_counts_df <- data.frame(
  feature = rownames(missing_counts),
  percent = missing_counts[,2]
)
cols_to_keep <- missing_counts_df$feature[missing_counts_df$
   percent <= 10 & missing_counts_df$feature != "Architecture" &
   missing_counts_df$feature != "Name"]
df_filtered <- df[, cols_to_keep, drop = FALSE]
head(df_filtered, 5)
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.

```r
# Chuan hoa don vi do trong cac cot
remove_unit_cols <- c("Memory_Bandwidth", "Memory_Speed", "Memory_Bus", "Direct_X")
main_df <- df_filtered
main_df[remove_unit_cols] <- lapply(df_filtered[remove_unit_cols], function(x) {
  as.numeric(gsub("[^0-9.]", "", x))
})
clean_cache <- function(x) {
  main <- as.numeric(sub("KB.*", "", x))       # 2304
  mult       <- as.numeric(sub(".*\\(x([0-9]+)\\)", "\\1", x))   # 2
  if (is.na(mult)) mult <- 1
  return(main * mult)
}
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