ĐẠI HỌC QUỐC GIA TP HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
Khoa Khoa Học Và Kỹ Thuật Máy Tính

BÀI TẬP LỚN
MÔN XÁC SUẤT THỐNG KÊ

Đánh giá hiệu năng của CPU sử dụng mô hình hồi quy tuyến tính

Giảng viên hướng dẫn: Phan Thị Hường
Lớp L13 – Nhóm 25

Danh sách thành viên:
Đỗ Huy Minh Dũng - 2210568
Trần Thị Lài - 2220035

Thành phố Hồ Chí Minh – 2023Đ

Đóng góp của nhóm:

STT	Họ và tên	MSSV	Tỉ lệ đóng góp	Điểm cộng
1.	Đỗ Huy Minh Dũng	2210568	50%	+0
2.	Trần Thị Lài	2220035	50%	-0

Đánh giá của giảng viên hướng dẫn:

STT	Họ và tên	MSSV	Đánh giá
1.	Đỗ Huy Minh Dũng	2210568	
2.	Trần Thị Lài	2220035	
3.	Tổng cộng		

Mục lục

1.Giới thiệu dữ liệu

1.1 Mô tả tóm tắt dữ liệu

Bộ dữ liệu trong bài tập lớn này chứa những thông tin về các thông số kỹ thuật, ngày phát hành và giá cả của các bộ phận máy vi tính, gồm file Intel_CPUs.csv thống kê những con chip xử lí trung tâm (Central Processing Unit).

1.2 Central Processing Unit (CPU)

1.2.1. Khái niệm về CPU

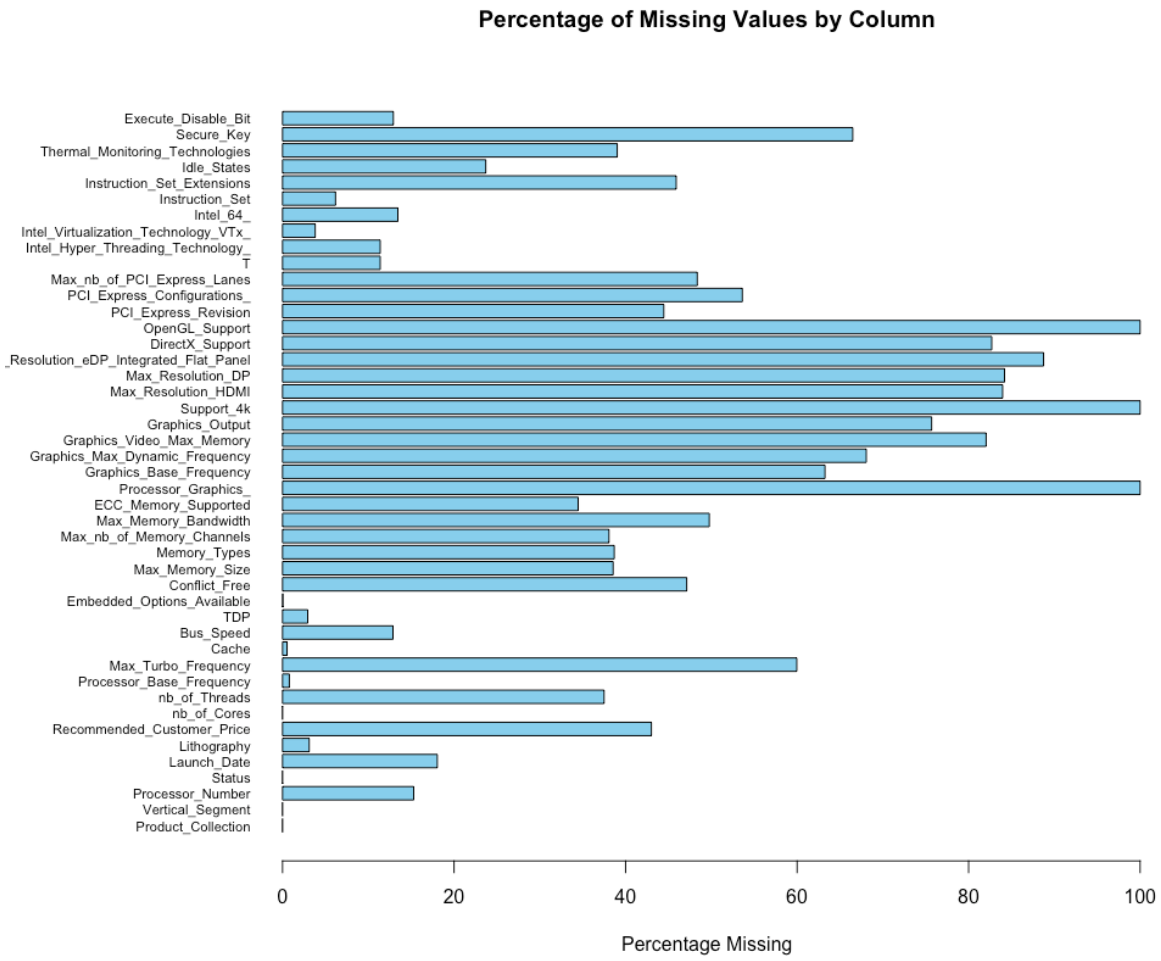
CPU là một con chip, cụ thể là một vi mạch tích hợp (IC - Integrated Circuit). Nó là một linh kiện điện tử phức tạp được thiết kế để thực hiện các phép tính toán và xử lý dữ liệu trong máy tính và các thiết bị điện tử khác.

1.2.2 Các biến và giá trị quan trắc

Bộ dữ liệu Intel_CPUs.csv có tổng cộng 2283 giá trị quan trắc và 45 biến. Với việc chọn biến TDP (Công suất thiết kế) làm biến phụ thuộc, nhóm thực hiện hai việc để lọc ra những biến độc lập phù hợp nhất:
- Khảo sát về độ khuyết của dữ liệu:

Product_Collection	Vertical_Segment	Processor_Number
0.00000000	0.00000000	15.28690320
Status	Launch_Date	Lithography
0.00000000	18.04643014	3.10994306
Recommended_Customer_Price	nb_of_Cores	nb_of_Threads
43.01357862	0.00000000	37.49452475
Processor_Base_Frequency	Max_Turbo_Frequency	Cache
0.78843627	59.96495839	0.52562418
Bus_Speed	TDP	Embedded_Options_Available
12.87779238	2.93473500	0.04380201
Conflict_Free	Max_Memory_Size	Memory_Types
47.13096802	38.54577311	38.67717915
Max_nb_of_Memory_Channels	Max_Memory_Bandwidth	ECC_Memory_Supported
38.06395094	49.75908892	34.47218572
Processor_Graphics_	Graphics_Base_Frequency	Graphics_Max_Dynamic_Frequency
100.00000000	63.25010951	68.06833114
Graphics_Video_Max_Memory	Graphics_Output	Support_4k
82.04117389	75.68988173	100.00000000
Max_Resolution_HDMI	Max_Resolution_DP	Max_Resolution_eDP_Integrated_Flat_Panel
83.96846255	84.18747262	88.74288217
DirectX_Support	OpenGL_Support	PCI_Express_Revision
82.69820412	100.00000000	44.45904512
PCI_Express_Configurations_	Max_nb_of_PCI_Express_Lanes	T
53.61366623	48.35742444	11.38852387
Intel_Hyper_Threading_Technology_	Intel_Virtualization_Technology_VTx_	Intel_64_
11.38852387	3.81077530	13.44721857
Instruction_Set	Instruction_Set_Extensions	Idle_States
6.17608410	45.90451161	23.69689006
Thermal_Monitoring_Technologies	Secure_Key	Execute_Disable_Bit
39.02759527	66.49145861	12.92159439

Hình 1: Phần trăm giá trị bị mất tính toán được[1]



Hình 2: Biểu đồ cột phần trăm dữ liệu bị mất của từng thuộc tính.

- Tìm hiểu các công bố về những thuộc tính có ảnh hưởng lớn tới công suất hoạt động của CPU:

1. Lithography[1]: Công nghệ bán dẫn có thể cải thiện hiệu suất và hiệu quả năng lượng của CPU. Các quy trình sản xuất mới dẫn đến sự thay đổi về kích thước của CPU và cung cấp hiệu quả năng lượng tốt hơn.

2. Cores and Threads [2]: Số lõi và luồng có thể ảnh hưởng đến khả năng đa nhiệm và đa luồng của CPU. Trong một số trường hợp, việc sử dụng nhiều lõi có thể giúp làm giảm thời gian hoàn thành công việc, nhưng cũng có thể tăng tiêu thụ năng lượng.

[1] Impact of CPU lithography on performance and idle power

[2] Rearch about Power and Energy of Number of Thread: Rearch document

3. Cache[3]: Kích thước và cấu trúc của cache có thể ảnh hưởng đến khả năng lưu trữ và truy xuất nhanh chậm. Điều đó dẫn đến giảm thời gian chờ đợi và khả năng tiêu thụ năng lượng.

4. Bandwidth [4]: Băng thông hệ thống ảnh hưởng trực tiếp đến khả năng truyền dữ liệu giữa CPU và bộ nhớ. Nếu băng thông của hệ thống cao, CPU có thể nhanh chóng truy cập và xử lý dữ liệu, giảm thời gian chờ đợi và giảm tiêu thụ năng lượng.

5. Instruction set(tập lệnh) [5] : Các lệnh mới và cải tiến có thể cung cấp các chức năng hiệu quả năng lượng và giúp tối ưu hóa hiệu suất.

6. Frequency[6] : Tần số xung nhịp cao có thể cung cấp hiệu suất tốt hơn trong một số tác vụ, nhưng cũng đi kèm với tiêu thụ năng lượng và sản sinh nhiệt độ cao hơn.

Đây là những thuộc tính bị khuyết dữ liệu ít nhất đồng thời có ảnh hưởng tới công suất hoạt động của CPU. Chi tiết về khái niệm và đơn vị đo của từng thuộc tính sẽ được đề cập ở mục 1.3.3

1.2.3 Các biến phân tích

Với mục đích phân tích hiệu suất của máy tính phát triển theo thời gian nhóm tập trung vào mối quan hệ của các biến sau:

1. Lithography: Công nghệ bán dẫn được sử dụng để sản xuất mạch tích hợp, được thống kê trong file với đơn vị nanômét (nhóm đã sửa tên thành ‘lithography’).

2. nb_of_Cores: Số lượng các đơn vị xử lý trung tâm độc lập (nhóm đã sửa tên thành ‘number_cores’).

3. nb_of_Threads: Được gọi là luồng hoặc luồng thực thi, mô tả số lượng chuỗi lệnh cơ bản được thực hiền theo trìn[2]h tự có thể truyền qua (nhóm đã sửa tên thành ‘number_threads’)

4. Processor_Base_Frequency: Mô tả tốc độ mà các transistor của bộ xử lý thực hiện mở và đóng, nói cách khác đây là tần số cơ bản của bộ xử lý (nhóm đã sửa tên thành ‘Pfrequency’).

5. Cache: Là một khu vực bộ nhớ nhanh được đặt trên bộ xử lý, có đơn vị là KB, MB,... (nhóm đã sửa tên thành ‘cache’).

6.TDP: Công suất trung bình, tính bằng watt, đặc trưng cho mức năng lượng mà bộ xử lý tiêu thụ khi hoạt động ở tần số cơ bản với tất cả các lõi (nhóm đã sửa tên thành ‘power’).

---

[3] Architectural Techniques for Improving the power Consumtion of NoCBase CMPs.
[4] The Impact of Bandwidth Constraints on the Energy Consumption of Wireless Sensor Networks
[5] A Study on the Impact of Instruction Set Architectures on Processor's Performance
[6] On the Impact of Sampling Frequency on Software Energy Measurements: Research document

7. Max_Memory_Bandwidth: Tốc độ tối đa mà dữ liệu có thể được đọc từ hoặc lưu trữ vào bộ nhớ bán dẫn bởi bộ xử lý, có đơn vị là GB/s (nhóm đã sửa tên thành 'bandwidth').

8. Instruction_Set: Một bộ các lệnh hoặc mã hướng dẫn mà một bộ xử lý (processor) hoặc một máy tính có khả năng thực hiện, có đơn vị là bit (nhóm đã sửa tên thành 'instruction_set').

1.3 Nguồn của dữ liệu

Bộ dữ liệu được cung cấp ở đây phần lớn thuộc về Intel, Game-Debate và những công ty khác có sự tham gia vào sản xuất các sản phẩm này.

2. Cơ sở lý thuyết

2.1 Một số loại biểu đồ được sử dụng trong bài tập lớn.

Scatterplot (biểu đồ phân tán): là một loại biểu đồ dạng chấm được sử dụng để biểu diễn dữ liệu hai chiều hoặc nhiều chiều trong một hệ trục tọa độ. Nhóm sẽ sử dụng biểu đồ này để thể hiện mối quan hệ giữa biến power (TDP) so với những biến phân tích khác.

Histogram (biểu đồ tần số): là một biểu đồ dạng cột sử dụng để biểu diễn phân phối tần số hoặc tần suất của một tập hợp dữ liệu.

Boxplot (biểu đồ hộp) là một loại biểu đồ thống kê được sử dụng để biểu diễn phân phối của một tập dữ liệu và tóm tắt các thông tin thống kê quan trọng về dữ liệu, như giá trị trung bình, phạm vi, median, và các giá trị ngoại lệ, thường được vẽ dưới dạng hộp chữ nhật với đường thẳng ngang qua bên trong.

2.2 Hồi quy tuyến tính

Giả sử Y phụ thuộc vào k biến độc lập X_1, … X_k. Mô hình hồi quy tuyến tính bội có dạng:

$Y = \alpha+\beta_1X_1 +\beta_2X_2 +...+\beta_kX_k +U$

Trong đó:

• β_j: được gọi là các hệ số hồi quy riêng, thể hiện mức độ biến thiên Y khi X_j thay đổi một đơn vị, các biến còn lại không đổi.

• U: là sai số

Có 3 giả định cần được đáp ứng khi thực hiện xây dựng mô hình hồi quy tuyến tính:

- Tính chuẩn tắc: phần dư (chênh lệch giữa giá trị quan sát được và giá trị dự đoán) tuân theo phân phối chuẩn.
- Tính tuyến tính: Mối quan hệ giữa các biến độc lập và biến phụ thuộc phải tuyến tính, nghĩa là giá trị kỳ vọng của biến phụ thuộc thay đổi theo đường thẳng khi các biến độc lập thay đổi, giữ cho các biến khác không đổi.
- Không có hiện tượng đa cộng tuyến: Các biến độc lập trong mô hình hồi quy không có mối tương quan cao với nhau. Đa cộng tuyến xảy ra sẽ gây khó khăn cho việc xác định tác động riêng lẻ của chúng lên biến phụ thuộc.

Phương trình hồi quy bội của mẫu:
Gọi các hệ số $a, b_1, ..b_k$ là ước lượng cho $\alpha, \beta_1, ...\beta_k$ được xác định bởi phương pháp bình phương bé nhất:

$$f = \sum_{i=1}^{n} (y_i - a - b_i x_{1i} - ... - b_k x_{ki})^2 \longrightarrow min$$

Từ điều kiện trên, ta có hệ:

H0 : $\beta = \beta_0$

H1 : $\beta > \beta_0$

Giải hệ phương trình, ta sẽ tìm được nghiệm $(a, b_1, ...b_k)$.
Phương trình $y = a + b_1x_1 + ... + b_kx_k$ được gọi là phương trình hồi quy bội của mẫu.
Chúng ta cũng có thể tìm được nghiệm $(a, b_1, ...b_k)$ bằng phương pháp ma trận, tuy nghiên dù phương pháp nào đi nữa thì việc tìm nghiệm bằng phương pháp thủ công là rất phức tạp. Với công nghệ máy tính phát triển, các phần mềm thống kê được phát triển thì việc tìm nghiệm trở nên dễ dàng hơn. Chính vì vậy, chúng ta không nên quá quan tâm đến việc tìm nghiệm bằng phương pháp thử công như thế nào. Phương pháp bình phương bé nhất phải thoả mãn điều kiện tương tự như đối với hồi quy tuyến tính.

Chỉ số hiệu suất mô hình:

R bình phương (R^2): là một độ đo thống kê được sử dụng trong mô hình hồi quy để đo lường mức độ biến động của biến phụ thuộc mà mô hình có thể giải thích. Giá trị R^2 nằm trong khoảng từ 0 đến 1. Giá trị càng gần 1 cho thấy mô hình giải thích một phần lớn sự biến động của biến phụ thuộc.

Mean Squared Error (MSE) đo lường sự chênh lệch giữa giá trị dự đoán từ mô hình và giá trị thực tế, sau đó lấy giá trị trung bình bình phương của sự chênh lệch đó. MSE càng nhỏ thì mô hình càng chính xác.

3. Thống kê tả

3.1. Tiền xử lý dữ liệu

- Import data:

Sau khi import file Intel_CPUs.csv ta có dữ liệu như hình:

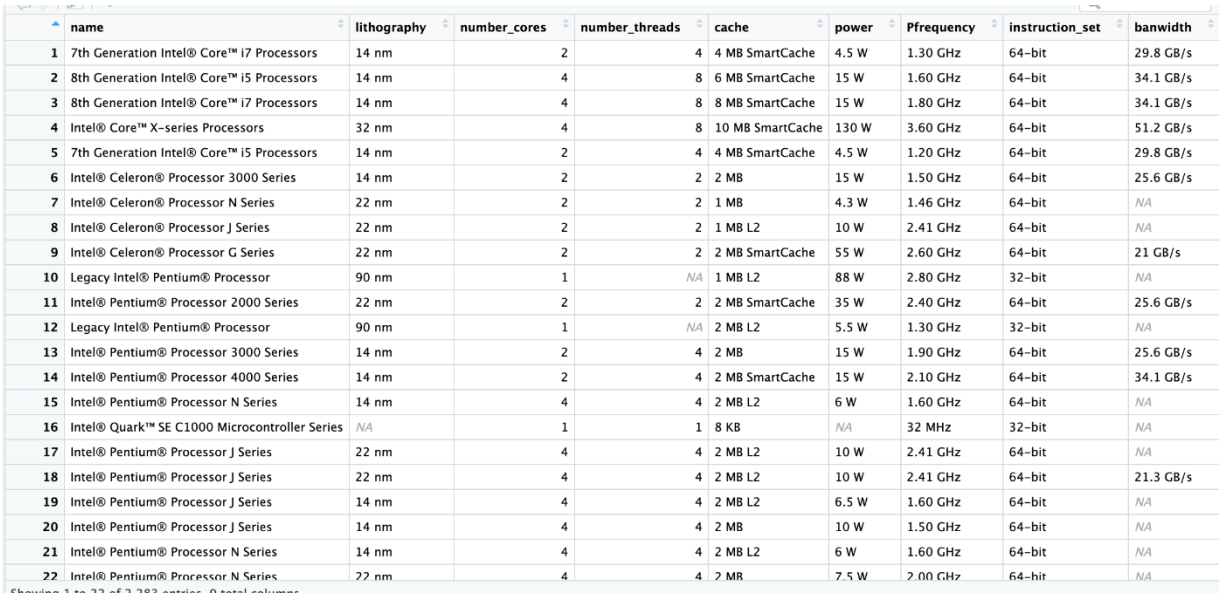
	Product_Collection	Vertical_Segment	Processor_Number	Status	Launch_Date	Lithography	Recommended_Customer_Price
1	7th Generation Intel® Core™ i7 Processors	Mobile	i7-7Y75	Launched	Q3'16	14 nm	$393.00
2	8th Generation Intel® Core™ i5 Processors	Mobile	i5-8250U	Launched	Q3'17	14 nm	$297.00
3	8th Generation Intel® Core™ i7 Processors	Mobile	i7-8550U	Launched	Q3'17	14 nm	$409.00
4	Intel® Core™ X-series Processors	Desktop	i7-3820	End of Life	Q1'12	32 nm	$305.00
5	7th Generation Intel® Core™ i5 Processors	Mobile	i5-7Y57	Launched	Q1'17	14 nm	$281.00
6	Intel® Celeron® Processor 3000 Series	Mobile	3205U	Launched	Q1'15	14 nm	$107.00
7	Intel® Celeron® Processor N Series	Mobile	N2805	Launched	Q3'13	22 nm	N/A
8	Intel® Celeron® Processor J Series	Desktop	J1750	Launched	Q3'13	22 nm	N/A
9	Intel® Celeron® Processor G Series	Desktop	G1610	Launched	Q1'13	22 nm	$42.00
10	Legacy Intel® Pentium® Processor	Mobile	518	End of Interactive Support		90 nm	N/A
11	Intel® Pentium® Processor 2000 Series	Mobile	2020M	Launched	Q3'12	22 nm	$134.00
12	Legacy Intel® Pentium® Processor	Mobile	773	End of Interactive Support		90 nm	N/A
13	Intel® Pentium® Processor 3000 Series	Mobile	3825U	Launched	Q1'15	14 nm	$161.00
14	Intel® Pentium® Processor 4000 Series	Mobile	4405U	Launched	Q3'15	14 nm	$161.00
15	Intel® Pentium® Processor N Series	Mobile	N3710	Launched	Q1'16	14 nm	$161.00
16	Intel® Quark™ SE C1000 Microcontroller Series	Embedded	C1000	Launched	Q4'15		$7.75
17	Intel® Pentium® Processor J Series	Desktop	J2850	Launched	Q3'13	22 nm	$94.00
18	Intel® Pentium® Processor J Series	Desktop	J2900	Launched	Q4'13	22 nm	$94.00
19	Intel® Pentium® Processor J Series	Desktop	J3710	Launched	Q1'16	14 nm	N/A
20	Intel® Pentium® Processor J Series	Desktop	J4205	Launched		14 nm	$161.00
21	Intel® Pentium® Processor N Series	Mobile	N3700	Launched	Q1'15	14 nm	$161.00
22	Intel® Pentium® Processor N Series	Mobile	N3510	Launched	Q3'13	22 nm	N/A

howing 1 to 22 of 2,283 entries, 45 total columns

Hình 3: Dữ liệu gốc

- Liệt kê dữ liệu khuyết của từng thuộc tính để tiến hành chọn ra những biến khả quan cho mô hình:

- Lọc biến cần sử dụng trong ta có bảng (data) như hình:

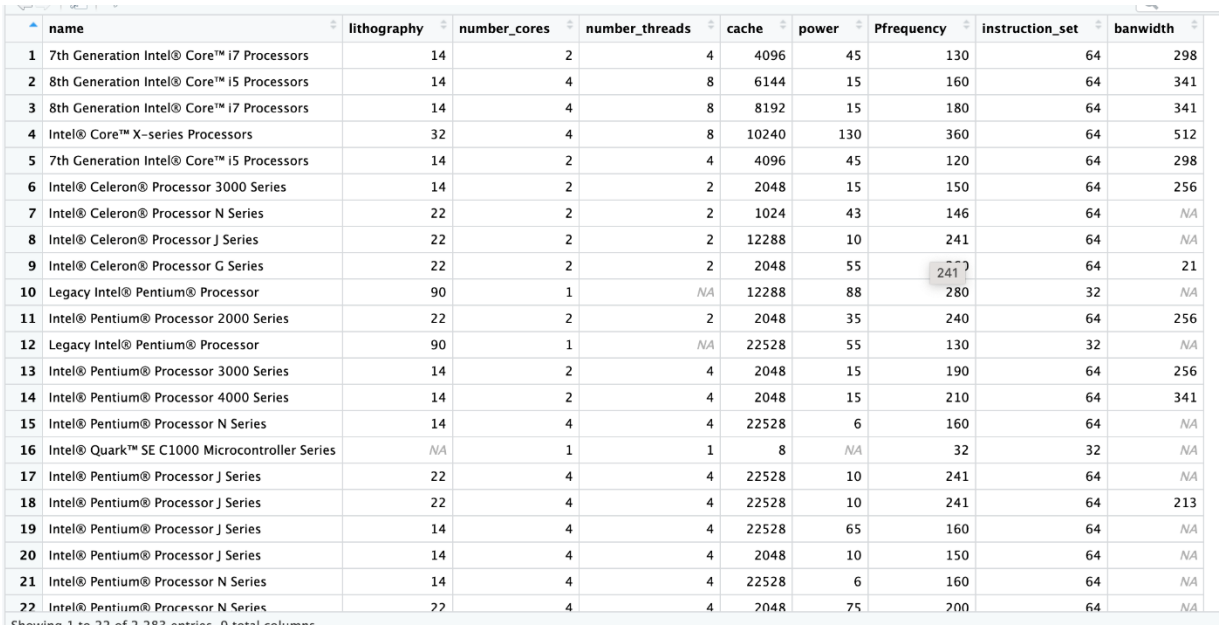


	name	lithography	number_cores	number_threads	cache	power	Pfrequency	instruction_set	banwidth
1	7th Generation Intel® Core™ i7 Processors	14 nm	2	4	4 MB SmartCache	4.5 W	1.30 GHz	64-bit	29.8 GB/s
2	8th Generation Intel® Core™ i5 Processors	14 nm	4	8	6 MB SmartCache	15 W	1.60 GHz	64-bit	34.1 GB/s
3	8th Generation Intel® Core™ i7 Processors	14 nm	4	8	8 MB SmartCache	15 W	1.80 GHz	64-bit	34.1 GB/s
4	Intel® Core™ X-series Processors	32 nm	4	8	10 MB SmartCache	130 W	3.60 GHz	64-bit	51.2 GB/s
5	7th Generation Intel® Core™ i5 Processors	14 nm	2	4	4 MB SmartCache	4.5 W	1.20 GHz	64-bit	29.8 GB/s
6	Intel® Celeron® Processor 3000 Series	14 nm	2	2	2 MB	15 W	1.50 GHz	64-bit	25.6 GB/s
7	Intel® Celeron® Processor N Series	22 nm	2	2	1 MB	4.3 W	1.46 GHz	64-bit	NA
8	Intel® Celeron® Processor J Series	22 nm	2	2	1 MB L2	10 W	2.41 GHz	64-bit	NA
9	Intel® Celeron® Processor G Series	22 nm	2	2	2 MB SmartCache	55 W	2.60 GHz	64-bit	21 GB/s
10	Legacy Intel® Pentium® Processor	90 nm	1	NA	1 MB L2	88 W	2.80 GHz	32-bit	NA
11	Intel® Pentium® Processor 2000 Series	22 nm	2	2	2 MB SmartCache	35 W	2.40 GHz	64-bit	25.6 GB/s
12	Legacy Intel® Pentium® Processor	90 nm	1	NA	2 MB L2	5.5 W	1.30 GHz	32-bit	NA
13	Intel® Pentium® Processor 3000 Series	14 nm	2	4	2 MB	15 W	1.90 GHz	64-bit	25.6 GB/s
14	Intel® Pentium® Processor 4000 Series	14 nm	2	4	2 MB SmartCache	15 W	2.10 GHz	64-bit	34.1 GB/s
15	Intel® Pentium® Processor N Series	14 nm	4	4	2 MB L2	6 W	1.60 GHz	64-bit	NA
16	Intel® Quark™ SE C1000 Microcontroller Series	NA	1	1	8 KB	NA	32 MHz	32-bit	NA
17	Intel® Pentium® Processor J Series	22 nm	4	4	2 MB L2	10 W	2.41 GHz	64-bit	NA
18	Intel® Pentium® Processor J Series	22 nm	4	4	2 MB L2	10 W	2.41 GHz	64-bit	21.3 GB/s
19	Intel® Pentium® Processor J Series	14 nm	4	4	2 MB L2	6.5 W	1.60 GHz	64-bit	NA
20	Intel® Pentium® Processor J Series	14 nm	4	4	2 MB	10 W	1.50 GHz	64-bit	NA
21	Intel® Pentium® Processor N Series	14 nm	4	4	2 MB L2	6 W	1.60 GHz	64-bit	NA
22	Intel® Pentium® Processor N Series	22 nm	4	4	2 MB	7.5 W	2.00 GHz	64-bit	NA

Hình 4: Dữ liệu sau khi lọc các biến chính

- Làm sạch dữ liệu gồm chuyển dữ liệu về đúng định dạng và lọc ra dữ liệu bị khuyết:

+ chuyển về đúng định dạng :



	name	lithography	number_cores	number_threads	cache	power	Pfrequency	instruction_set	banwidth
1	7th Generation Intel® Core™ i7 Processors	14	2	4	4096	45	130	64	298
2	8th Generation Intel® Core™ i5 Processors	14	4	8	6144	15	160	64	341
3	8th Generation Intel® Core™ i7 Processors	14	4	8	8192	15	180	64	341
4	Intel® Core™ X-series Processors	32	4	8	10240	130	360	64	512
5	7th Generation Intel® Core™ i5 Processors	14	2	4	4096	45	120	64	298
6	Intel® Celeron® Processor 3000 Series	14	2	2	2048	15	150	64	256
7	Intel® Celeron® Processor N Series	22	2	2	1024	43	146	64	NA
8	Intel® Celeron® Processor J Series	22	2	2	12288	10	241	64	NA
9	Intel® Celeron® Processor G Series	22	2	2	2048	55	[illegible]	64	21
10	Legacy Intel® Pentium® Processor	90	1	NA	12288	88	280	32	NA
11	Intel® Pentium® Processor 2000 Series	22	2	2	2048	35	240	64	256
12	Legacy Intel® Pentium® Processor	90	1	NA	22528	55	130	32	NA
13	Intel® Pentium® Processor 3000 Series	14	2	4	2048	15	190	64	256
14	Intel® Pentium® Processor 4000 Series	14	2	4	2048	15	210	64	341
15	Intel® Pentium® Processor N Series	14	4	4	22528	6	160	64	NA
16	Intel® Quark™ SE C1000 Microcontroller Series	NA	1	1	8	NA	32	32	NA
17	Intel® Pentium® Processor J Series	22	4	4	22528	10	241	64	NA
18	Intel® Pentium® Processor J Series	22	4	4	22528	10	241	64	213
19	Intel® Pentium® Processor J Series	14	4	4	22528	65	160	64	NA
20	Intel® Pentium® Processor J Series	14	4	4	2048	10	150	64	NA
21	Intel® Pentium® Processor N Series	14	4	4	22528	6	160	64	NA
22	Intel® Pentium® Processor N Series	22	4	4	2048	75	200	64	NA

Hình 5: Dữ liệu sau định dạng

+ Thống kê dữ liệu missing

	missing	%
banwidth	1136	50
number_threads	856	37
instruction_set	141	6
lithography	71	3
power	67	3
Pfrequency	18	1

Hình 6: Thống kê dữ liệu mising

Nhận xét: do số lượng dữ liệu missing quá lớn nên nhóm đề xuất sử dụng phương pháp Mean/Model/ Median Imputation để điền vào giá trị còn thiếu trong bảng (cleaned_data)

Bảng dữ liệu hoàn chỉnh dùng phân tích (cleaned_data)

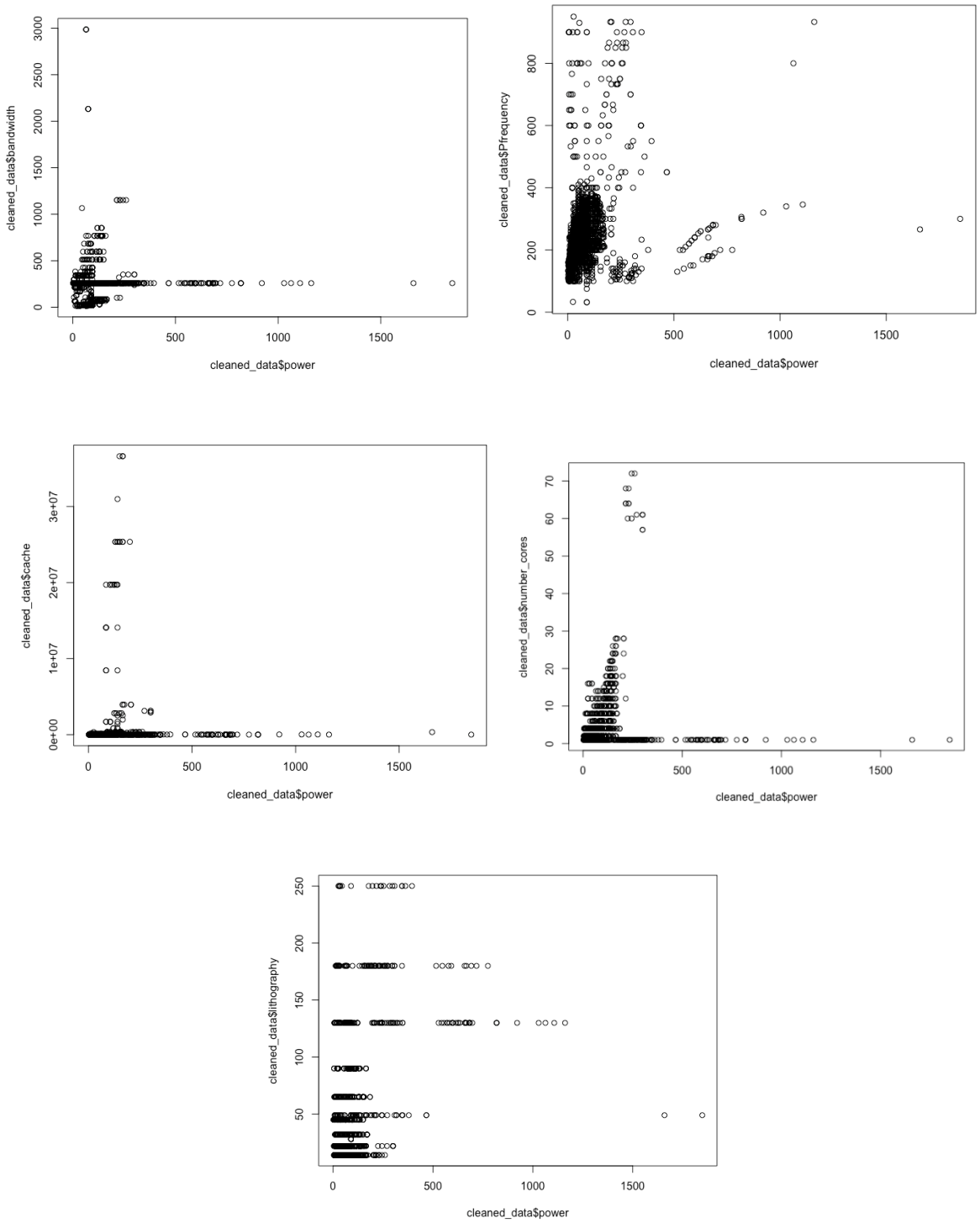
	name	lithography	number_cores	number_threads	cache	power	Pfrequency	instruction_set	banwidth
1	7th Generation Intel® Core™ i7 Processors	14	2	4	4096	45	130	64	298
2	8th Generation Intel® Core™ i5 Processors	14	4	8	6144	15	160	64	341
3	8th Generation Intel® Core™ i7 Processors	14	4	8	8192	15	180	64	341
4	Intel® Core™ X-series Processors	32	4				360	64	512
5	7th Generation Intel® Core™ i5 Processors	14	2	4	4096	45	120	64	298
6	Intel® Celeron® Processor 3000 Series	14	2	2	2048	15	150	64	256
7	Intel® Celeron® Processor N Series	22	2	2	1024	43	146	64	259
8	Intel® Celeron® Processor J Series	22	2	2	12288	10	241	64	259
9	Intel® Celeron® Processor G Series	22	2	2	2048	55	260	64	21
10	Legacy Intel® Pentium® Processor	90	1	9	12288	88	280	32	259
11	Intel® Pentium® Processor 2000 Series	22	2	2	2048	35	240	64	256
12	Legacy Intel® Pentium® Processor	90	1	9	22528	55	130	32	259
13	Intel® Pentium® Processor 3000 Series	14	2	4	2048	15	190	64	256
14	Intel® Pentium® Processor 4000 Series	14	2	4	2048	15	210	64	341
15	Intel® Pentium® Processor N Series	14	4	4	22528	6	160	64	259
16	Intel® Quark™ SE C1000 Microcontroller Series	49	1	1	8	90	32	32	259
17	Intel® Pentium® Processor J Series	22	4	4	22528	10	241	64	259
18	Intel® Pentium® Processor J Series	22	4	4	22528	10	241	64	213
19	Intel® Pentium® Processor J Series	14	4	4	22528	65	160	64	259
20	Intel® Pentium® Processor J Series	14	4	4	2048	10	150	64	259
21	Intel® Pentium® Processor N Series	14	4	4	22528	6	160	64	259
22	Intel® Pentium® Processor N Series	22	4	4	2048	75	200	64	259

column 4: numeric with range 0 - 60

Hình 7: Dữ liệu sau khi làm sạch

3.2 Thống kê tả

3.2.1 So sánh tỉ số tương quan



Hình 8: Scatter plots của các chỉ số theo power

Nhận xét:

- Bandwidth và power: Mức độ phân bố của tập dữ liệu không đều và có xu hướng trải đều sang hai hướng và có thể đây là mô hình tuyến tính bậc 2, cho thấy với bộ dữ liệu hiện tại bandwidth có vài outlier bên ngoài làm tăng mức độ tiêu thụ năng lượng.
- Pfrequency và power: Dữ liệu cho có xu hướng đi lên từ trái sang phải , cho thấy khi cpu có tầng số càng lớn thì mức độ sử dụng năng lượng càng cao có thể hình thành mối quan hệ tuyến tính giữa Pfrequency và power.
- Cache và power: Dữ liệu tập trung chủ yếu dưới 100 cho thấy cache của CPU càng bé thì mức độ sử dụng năng lượng càng lớn. Các outlier bên ngoài do có cache lớn nên độ tiêu thụ năng lượng có dấu hiệu giảm dần.
- Number_cores và power: Dữ liệu có xu hướng phân tán sang phải, tuy nhiên có 1 vài outlier bên ngoài dẫn đến mất dữ liệu có thể giữa cores và power có quan hệ tuyến tính theo chiều tăng.
- Lithography và power: Với mỗi kích thước lithography lại có sự tăng các power khác nhau, có thể năng lượng tăng do sự tăng nhiệt độ trên lithography, quan hệ tuyến tính giữa lithography và power chưa xác định được.

3.2.2 Tóm tắt dữ liệu

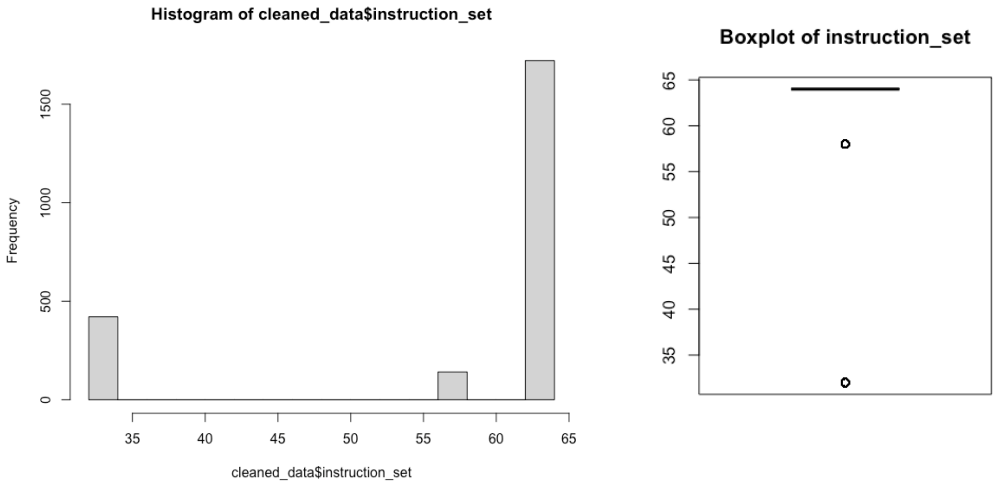
Về tên của các cpu

```
> cat( "Số loại cpu: ", name_count, "\n")
Số loại cpu:  75
> cat("Tên có tần suất cao nhất là:", most_frequent_name, "\n")
Tên có tần suất cao nhất là: Legacy Intel® Core™ Processors
```

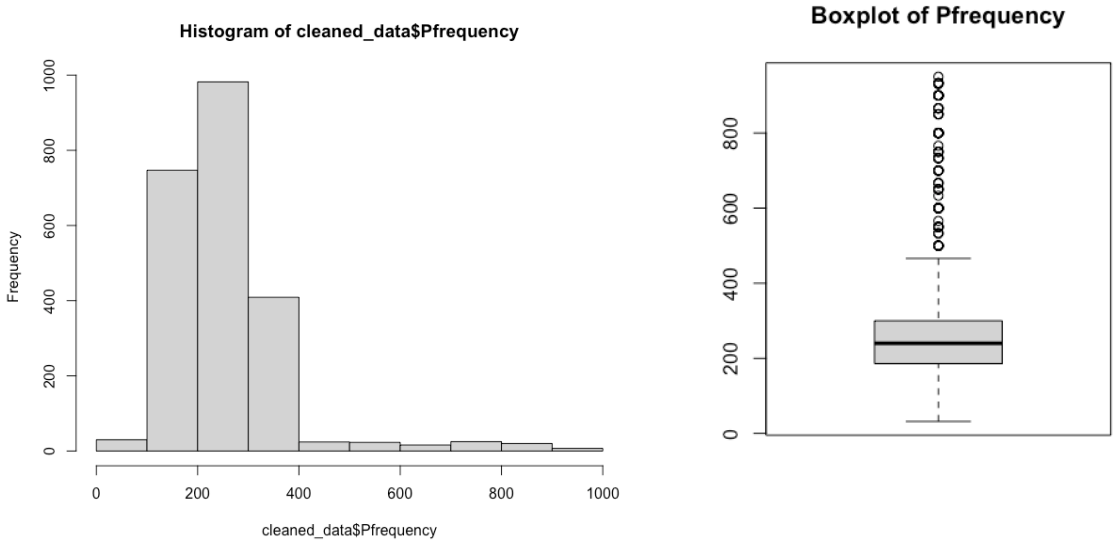
Nhận xét: Có 75 tên CPU được tìm thấy, theo thống kê tên có tần suất cao nhất là Legacy IntelR Core^{TM} Processors (375 lần)

	Statistic	Lithography	Cores	Threads	Cache	Power	Instruction_set	Pfrequency	Bandwidth
1	Median	32.00000	2.000000	9.000000	8192.0	65.0000	64.00000	240.0000	259.0000
2	Mean	48.98642	4.066579	8.830048	334651.4	89.2247	57.72843	258.8909	258.9645
3	Minimum	14.00000	1.000000	1.000000	8.0	2.0000	32.00000	32.0000	16.0000
4	Maximum	250.00000	72.000000	56.000000	36611072.0	1848.0000	64.00000	950.0000	2986.0000

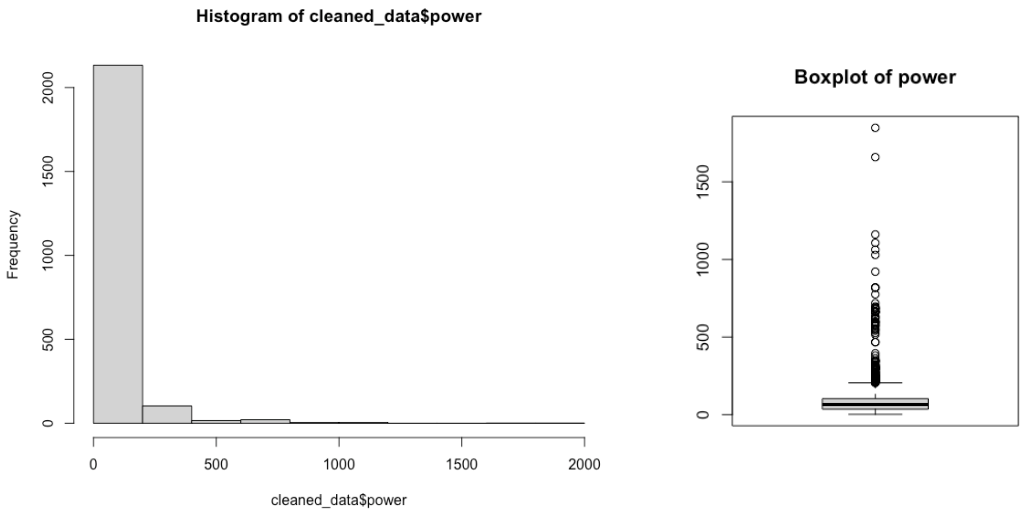
Hình 9: Thống kê median, mean, min, max của các biến



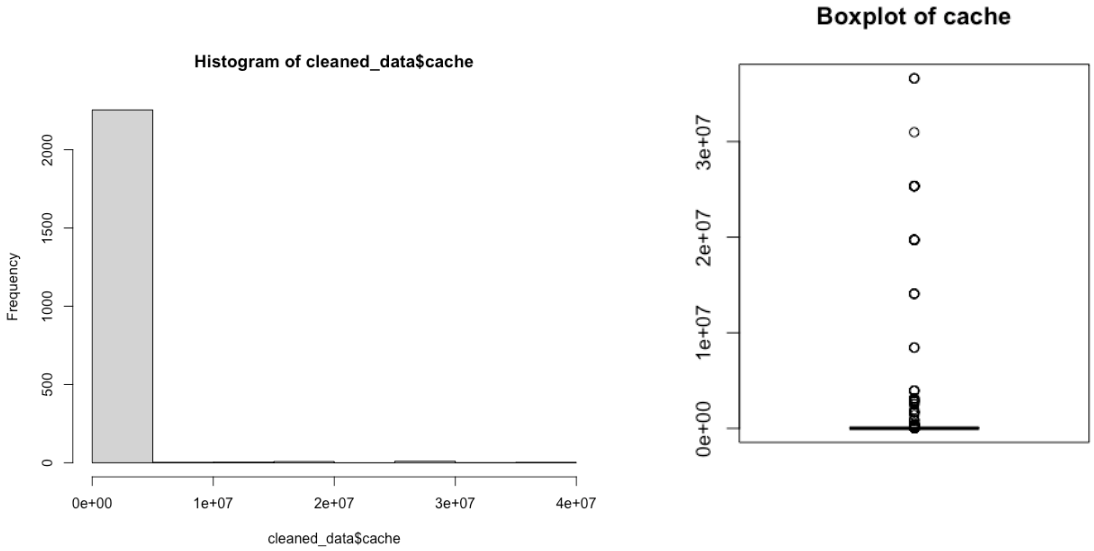
Hình 10: Histogram và Boxplot của instruction_set



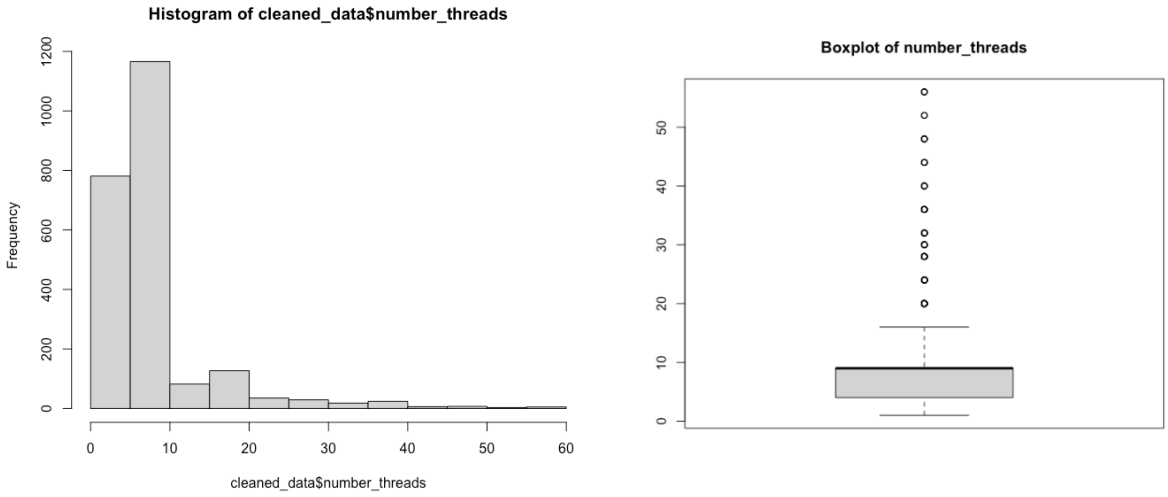
Hình 11: Histogram và Boxplot của Pfrequency



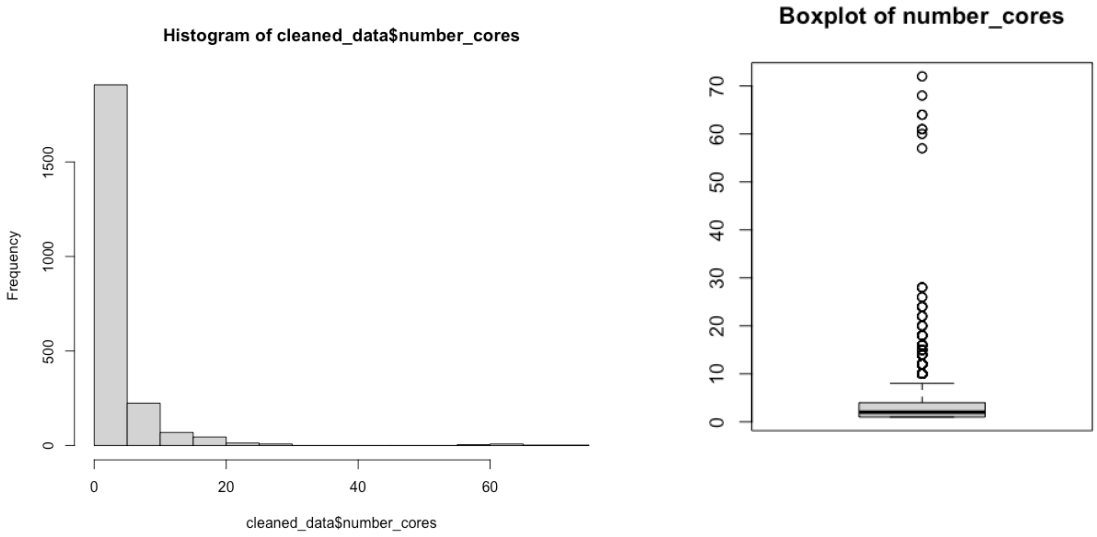
Hình 12: Histogram và Boxplot của power



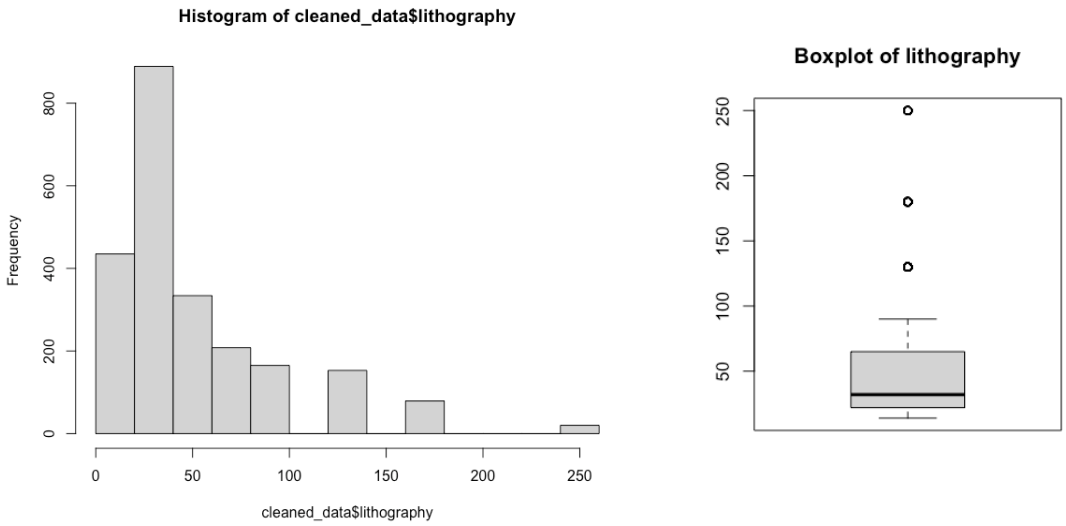
Hình 13: Histogram và Boxplot của cache



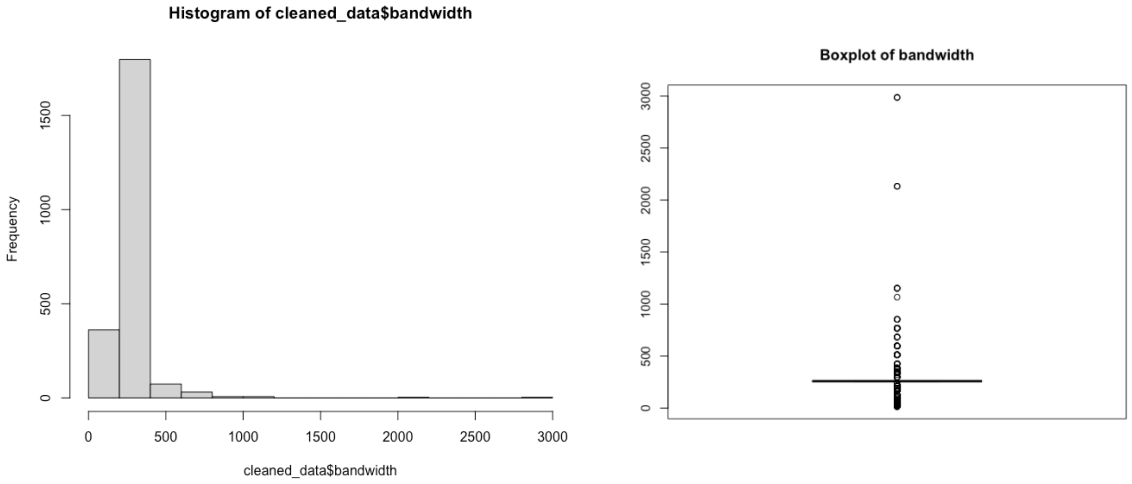
Hình 14: Histogram và Boxplot của number_threads



Hình 15: Histogram và Boxplot của number_cores



Hình 16: Histogram và Boxplot của lithography



Hình 17: Histogram và Boxplot của bandwidth

Tóm tắt về các biểu đồ:

- Lithography: kích thước của các lithography giao động từ 14 -250 nm với trung bình 48.98 nm. Vì giá trị median (32) < giá trị trung bình nên phân bố bị lệch trái cho thấy cpu có xu hướng giảm kích thước lithography.
- Cores : số cores của cpu giao động từ 1 – 72 với trung bình 4. Vì median (2) < trung bình nên phân bố lệch trái cho thấy số cores của cpu chủ yếu ở mức thấp.
- Threads : thread giao động từ 1 – 56 với trung bình 8.83 ~ 9 . Vì median ~ trung bình nên phân bố của cores gần như phân bố đều.
- Cache : số cache giao động từ 8 đến 36611072 kB với trung bình 334651. Vì median (8192) < trung bình nên phân bố lệch trái.

- Power: giá trị giao động từ 2 – 1848 W với trung bình 89.22 W. Vì median (65) < trung bình nên phân bố lệch trái cho thấy đa phần cpu sử dụng năng lượng tương đối ít.
- Instruction_set : giá trị giao động từ 32 – 64 (bit) với trung bình ~58 . Vì median (64) > trung bình và bằng giá trị lớn nhất nên phân bố lệch phải.Cho thấy đa số cpu được thiết kế với instruction_set bằng 64 bit.
- Pfrequency: giá trị giao động từ 32 -950 với trung bình ~ 259. Vì median(240) < trung bình nên phân bố lệch trái, cho thấy đa số cpu hoạt động với tầng số tương đối cao
- Bandwidth : giá trị giao động từ 16 – 2986 với trung bình ~259. Vì median (259) ~ trung bình nên phân bố gần đều.

4. Phân tích mối tương quan giữa công suất của CPU với các thuộc tính sử dụng mô hình hồi quy tuyến tính.

Nhóm tiến hành xây dựng mô hình hồi quy tuyến tính với power là biến phụ thuộc vào các biến lithography, number_cores, number_threads, cache, Pfrequency, instruction_set, bandwidth. Với mục tiêu xây dựng mô hình dự đoán giá trị power. Hàm **lm** trong R sẽ được sử dụng để ước tính mối quan hệ giữa các biến độc lập với biến phụ thuộc.

4.1 Giả định

4.1.1 Phân phối đều

Bước 1: Tạo mô hình hồi quy tuyến tính
Bước 2: Lấy phần dư từ mô hình
Bước 3 Thực hiện kiểm định bằng Shapiro-Wilk cho phần dư có phân phối chuẩn
Kết quả đạt được:

shapiro_test_result	list [4] (S3: htest)	List of length 4
statistic	double [1]	0.5299201
W	double [1]	0.5299201
p.value	double [1]	3.005988e-61
method	character [1]	'Shapiro-Wilk normality test'
data.name	character [1]	'residuals'

Hình 18: Kết quả kiểm định Shapiro-Wilk

Giá trị của p trong phần dư xấp xỉ 0 .Vì vậy có thể kết luận phần dư sau khi sử dụng mô hình hồi quy tuyến tính có phân phối chuẩn.

4.1.2 Độ tuyến tính

```
Call:
lm(formula = power ~ lithography + number_cores + number_threads +
    cache + Pfrequency + instruction_set + bandwidth, data = cleaned_data)

Residuals:
    Min      1Q  Median      3Q     Max
-260.37  -27.53  -10.59   15.27 1737.92

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.287e+01  1.882e+01   3.340  0.00085 ***
lithography      8.026e-01  7.655e-02  10.485  < 2e-16 ***
number_cores     3.469e+00  4.335e-01   8.003 1.92e-15 ***
number_threads   1.546e+00  3.749e-01   4.125 3.84e-05 ***
cache            3.011e-07  8.947e-07   0.337  0.73649
Pfrequency       7.984e-02  1.841e-02   4.336 1.51e-05 ***
instruction_set -1.090e+00  2.619e-01  -4.161 3.29e-05 ***
bandwidth        5.449e-03  1.259e-02   0.433  0.66515
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 103.1 on 2275 degrees of freedom
Multiple R-squared:  0.2112,    Adjusted R-squared:  0.2087
F-statistic: 86.99 on 7 and 2275 DF,  p-value: < 2.2e-16
```

Hình 19: Kiểm tra độ tuyến tính

Trong hồi quy tuyến tính , hệ số xác định thường được gọi là R bình phương (R^2) là thước đo thống kê biểu thị tỉ lệ phương sai của biến phụ thuộc có thể giải thích bằng các biến độc lập trong mô hình hồi quy tuyến tính. Mức giao động của R bình phương hiệu chỉnh từ 0 đến 1.

Với R-squared = 0.2112 nghĩa là sự biến thiên của biến phụ thuộc theo các biến độc lập trong mô hình thấp. Kết luận mối quan hệ giữa power với các biến khác là mối quan hệ tuyến tính yếu ($R^2 < 0.5$).

4.1.3 Kiểm tra tính đa cộng tuyến

```
lithography    number_cores  number_threads          cache
   2.499865        1.617170       1.573442       1.166743
 Pfrequency instruction_set      bandwidth
   1.145808        2.235821       1.038852
```

Hình 20: coefficient của các biến

Đa công tuyến được dùng để kiểm tra các biến trong mô hình bằng Variance Inflation Factor (VIF). VIF là thước đo lường mức độ tăng lên của phương sai của ước lượng hệ số hồi quy do hiện tượng đa cộng tuyến.
VIF của các biến tương ứng: lithography = 2.499, number_cores = 1.61717, number_threads = 1.57344, cache = 1.1667, Pfrequency = 1.1458, instruction_set = 2.2358, bandwidth = 1.0388.
VIF nếu lớn hơn 5 hoặc 10 cho thấy mức độ đa cộng tuyến của mô hình có vấn đề. Trong trường hợp trên VIF của các biến đều bé hơn 5 nên VIF có thể sử dụng trong mô hình.

4.2 Phân tách dữ liệu

Phân tách dữ liệu giúp ngăn chặn việc mô hình bị điều chỉnh quá mức, điều nay thường xảy ra khi mô hình ghi nhớ dữ liệu huấn luyện quá tốt và không thể khái quát hóa lên thành dữ liệu mới. Vì vậy, nhóm chia dữ liệu thành các tập huyến luyện và kiểm thử theo lượng dữ liệu 80% và 20%.

4.3 Hồi quy tuyến tính đa biến

Trước tiên, nhóm sẽ hiện thực mô hình hồi quy tuyến tính đa biến bậc nhất.
Phương trình tổng quát của mô hình hồi quy tuyến tính đa biến có dạng:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$$

4.3.1 Điều chỉnh mô hình

```
Residuals:
    Min      1Q  Median      3Q     Max
-260.37  -27.53  -10.59   15.27 1737.92

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.287e+01  1.882e+01   3.340  0.00085 ***
lithography      8.026e-01  7.655e-02  10.485  < 2e-16 ***
number_cores     3.469e+00  4.335e-01   8.003 1.92e-15 ***
number_threads   1.546e+00  3.749e-01   4.125 3.84e-05 ***
cache            3.011e-07  8.947e-07   0.337  0.73649
Pfrequency       7.984e-02  1.841e-02   4.336 1.51e-05 ***
instruction_set -1.090e+00  2.619e-01  -4.161 3.29e-05 ***
bandwidth        5.449e-03  1.259e-02   0.433  0.66515
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 103.1 on 2275 degrees of freedom
Multiple R-squared:  0.2112,    Adjusted R-squared:  0.2087
F-statistic: 86.99 on 7 and 2275 DF,  p-value: < 2.2e-16
```

Hình 21: Mô hình hồi quy tuyến tính đa biến - Kết quả

Từ kết quả nhóm có những nhận định:

Mô hình hồi quy tuyến tính bội đã được điều chỉnh và có giá trị phần dư (Residuals) nằm trong khoảng từ -260.37 đến 1737.92 với tứ phân vị thứ nhất là -27.53, trung vị là -10.59 và tứ phân vị thứ ba là 15.57. Điều này cho thấy sự chênh lệch trong các sai số dự đoán là rất lớn. Bên cạnh đó, khoảng cách giá trị giữa ba tứ phân vị không quá cao trong khoảng cách giữa tứ phân vị thứ nhất với giá trị phần dư nhỏ nhất và khoảng cách giữa tứ phân vị thứ ba và giá trị phần dư lớn nhất lại vô cùng lớn chứng tỏ rằng sai số dự đoán có phân bố khá đều ở trung tâm nhưng lại rất dốc ở vùng rìa.

Các tham số ước lượng (coefficients) của mô hình thể hiện hướng và độ lớn của mối quan hệ giữa các biến độc lập đối với biến phụ thuộc. Trong trường hợp này, tất cả các biến độc lập đều có mối quan hệ dương với biến phụ thuộc ngoại trừ biến instruction_set có mối quan hệ âm. Ngoài ra, nhóm nhận thấy 2 biến độc lập cache và bandwidth có chỉ số p-value khá cao (lần lượt là 0.73649 và 0.66515), lớn hơn đáng kể giá trị thông thường là 0.05 hay nói cách khác là 2 biến dự đoán này không có ý nghĩa thống kê, không có đóng góp đáng kể vào mô hình.

Các biến độc lập còn lại có chỉ số p-value khá thấp cho thấy chúng là các yếu tố dự đoán có ý nghĩa thống kê. Lithography là biến có chỉ số p-value nho nhất được quan sát cho thấy đây là biến dự đoán có ý nghĩa cao.

Sai số chuẩn dư (Residual standard error) của mô hình là khoảng 103.1 khá cao, giá trị này là một thước đo về độ lệch thông thường của giá trị thực tế so với giá trị dự đoán.

Hệ số R bình phương đa biến (0.2112) và giá trị R bình phương được điều chỉnh (0.2087) có độ lớn khá thấp cho thấy mô hình này không thực sự phù hợp để biểu diễn mối quan hệ giữa biến phụ thuộc power đối với các biến độc lập khác. Giá trị R bình phương được điều chỉnh, tức là hệ số R bình phương đa biến được tính lại sau khi loại bỏ những biến không có ý nghĩa thống kê cao có độ lớn chỉ nhỏ hơn một chút so với hệ số R bình phương cũ, cho thấy việc loại bỏ 2 biến trên không ảnh hưởng nhiều đến mô hình hồi quy.

Chỉ số F-statistic (86.99) và p-value tương ứng (< 2.2e-16) để kiểm tra giả thuyết rằng có ít nhất một trong các biến dự đoán có ý nghĩa trong việc giải thích biến phản ứng. Với chỉ số p-value rất nhỏ, nhóm quyết định bác bỏ giả thuyết H0, nghĩa là giả thuyết H1, ít nhất một biến dự đoán đóng góp đáng kể vào việc dự đoán biến phản ứng được nhận định là đúng.

4.3.2 Hồi quy từng bước (Stepwise Regression)

Nhóm quyết định sẽ áp dụng phương pháp hồi quy từng bước để chọn ra mô hình phù hợp nhất:

```
Start:  AIC=21173.91
power ~ lithography + number_cores + number_threads + cache +
    Pfrequency + instruction_set + bandwidth

                  Df Sum of Sq      RSS   AIC
- cache            1      1204 24176405 21172
- bandwidth        1      1991 24177192 21172
<none>                         24175201 21174
- number_threads   1    180820 24356021 21189
- instruction_set  1    183973 24359175 21189
- Pfrequency       1    199793 24374995 21191
- number_cores     1    680520 24855721 21235
- lithography      1   1168279 25343480 21280

Step:  AIC=21172.03
power ~ lithography + number_cores + number_threads + Pfrequency +
    instruction_set + bandwidth

                  Df Sum of Sq      RSS   AIC
- bandwidth        1      1876 24178281 21170
<none>                         24176405 21172
+ cache            1      1204 24175201 21174
- instruction_set  1    187693 24364098 21188
- Pfrequency       1    201082 24377487 21189
- number_threads   1    205114 24381518 21189
- number_cores     1    684715 24861120 21234
- lithography      1   1175183 25351588 21278

Step:  AIC=21170.2
power ~ lithography + number_cores + number_threads + Pfrequency +
    instruction_set

                  Df Sum of Sq      RSS   AIC
<none>                         24178281 21170
+ bandwidth        1      1876 24176405 21172
+ cache            1      1088 24177192 21172
- instruction_set  1    188367 24366647 21186
- Pfrequency       1    202937 24381217 21187
- number_threads   1    204410 24382691 21187
- number_cores     1    714830 24893111 21235
- lithography      1   1174024 25352304 21277
```

Hình 22: Kết quả của hồi quy từng bước

Kết quả trên cho thấy:

Quy trình điều chỉnh mô hình hồi quy đã được áp dụng vào mô hình hồi quy tuyến tính ban đầu, trong đó bao gồm tất cả các biến (lithography, number_cores, number_threads, cache, Pfrequency, instruction_set, bandwidth). Chỉ số AIC (Akaike Information Criterion) được sử dụng để so sánh các mô hình với nhau. Giá trị AIC thấp hơn cho thấy mô hình có độ phù hợp tốt hơn.

Chỉ số AIC của mô hình ban đầu là 21173.91, quy trình hồi quy từng bước bắt đầu với mô hình bằng cách xem xét tác động của việc loại bỏ từng biến trên AIC. Quy trình quyết định loại bỏ biến cache đầu tiên, việc loại bỏ này làm cho giá trị tổng bình phương dư (RSS) tăng lên ít nhất đồng thời làm giảm giá trị AIC từ 21173.91 xuống 21172.03 được xem là hợp lí.

Quy trình điều chỉnh được tiếp tục bằng cách xem xét việc thêm biến cache lại vào mô hình hoặc loại bỏ bất kỳ biến còn lại nào khác. Cuối cùng quy trình hồi quy từng bước quyết định loại bỏ biến bandwidth ra khỏi mô hình. Việc loại bỏ bandwidth làm chỉ số RSS tăng lên ít nhất và cũng làm giảm giá trị AIC từ 21172.03 xuống 21170.2

Quy trình tiếp tục xem xét việc thêm lại các biến đã loại bỏ (cache, bandwidth) vào mô hình hoặc loại bỏ bất kỳ biến còn lại nào (lithography, number_cores, number_threads, Pfrequency, instruction_set). Tuy nhiên việc thêm lại hoặc loại bỏ dù theo cách nào cũng sẽ làm tăng chỉ số AIC nên quy trình kết thúc.

Sau khi thực hiện điều chỉnh, mô hình cuối cùng bao gồm lithography, number_cores, number_threads, Pfrequency và instruction_set đã có số AIC thấp hơn ban đầu, cho thấy mô hình được chọn bằng phương pháp hồi quy từng bước sẽ phù hợp hơn theo tiêu chí này.

4.3.3 Phương trình tuyến tính

Sau khi áp dụng hồi quy từng bước và kiểm tra giả định, nhóm đã có mô hình hồi quy tuyến tính bội cuối cùng. Các hệ số hồi quy cho mỗi biến độc lập trong mô hình này như sau:

```
power = 64.51055 + 0.7991562 * lithography + 3.505458 * number_cores + 1.578585
* number_threads + 0.08035478 * Pfrequency + -1.098782 * instruction_set
```

Hình 23: Các tham số của hồi quy tuyến tính

Mặc dù đã áp dụng các quy trình điều chỉnh, tuy nhiên mô hình không thực sự có tính chính xác cao, cho thấy các biến lithography, number_cores, number_threads, Pfrequency, instruction_set không có ảnh hưởng lớn tới power trong hồi quy tuyến tính, được chỉ ra bởi giá trị R^2 khá thấp là 0.2111

```
Call:
lm(formula = power ~ lithography + number_cores + number_threads +
    Pfrequency + instruction_set, data = cleaned_data)

Residuals:
    Min      1Q  Median      3Q     Max
-260.03  -27.54  -10.38   14.66 1737.67

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     64.51055   18.52546   3.482 0.000507 ***
lithography      0.79916    0.07600  10.515  < 2e-16 ***
number_cores     3.50546    0.42724   8.205 3.82e-16 ***
number_threads   1.57859    0.35979   4.388 1.20e-05 ***
Pfrequency       0.08035    0.01838   4.372 1.29e-05 ***
instruction_set -1.09878    0.26088  -4.212 2.63e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 103 on 2277 degrees of freedom
Multiple R-squared:  0.2111,    Adjusted R-squared:  0.2093
F-statistic: 121.8 on 5 and 2277 DF,  p-value: < 2.2e-16
```

Hình 24: Kết quả của mô hình hồi quy tuyến tính cuối cùng

4.3.4 Kiểm tra khả năng dự đoán của mô hình

Kiểm thử phương trình trong hình 4.4 để dự đoán các giá trị power trong tập dữ liệu kiểm thử mà nhóm đã phân chia ở mục 4.2.1.

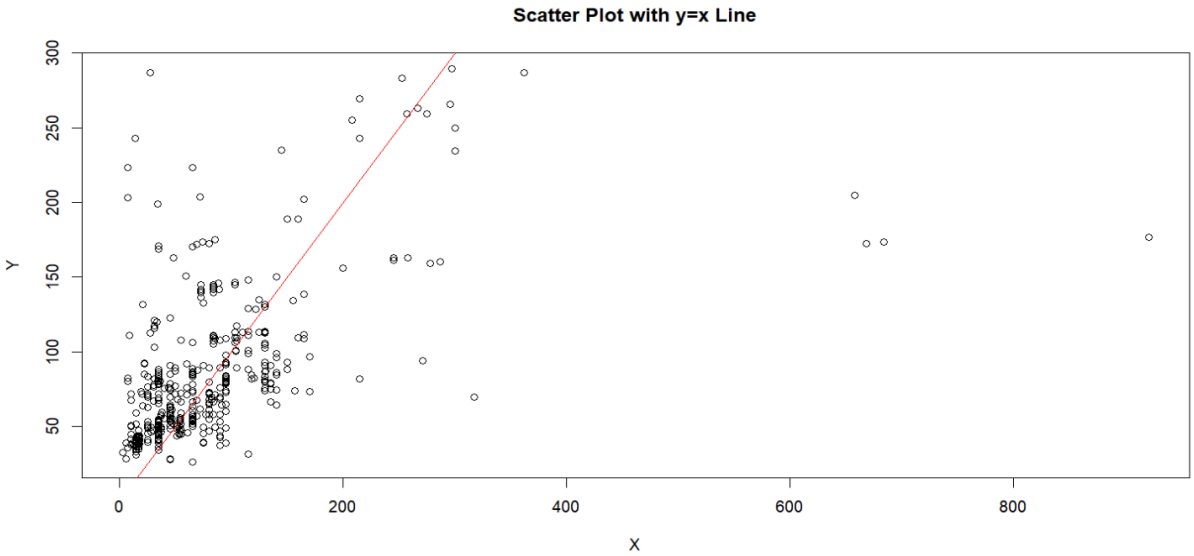
	test_set	prediction
4	130	75.33971
5	45	28.34450
8	10	41.30350
11	35	41.22315
16	90	76.16356
20	10	37.76606
21	6	38.56960
24	75	49.46272
31	15	30.81216
32	115	31.58057
50	59	150.50472
59	115	147.91402
65	287	160.03263
67	84	111.14589
68	684	173.45188

Hình 25: So sánh giá trị kiểm thử và dự đoán

Một số quan sát:

1. Đối với quan sát đầu tiên (ở vị trí 4), power thực tế là 130, nhưng mô hình đã đánh giá thấp, dự đoán xấp xỉ 75.34.

2. Đối với quan sát thứ hai (ở vị trí 5) , power thực tế là 45 và mô hình cũng đánh giá thấp với dự đoán xấp xỉ 28.34
3. Đối với quan sát ở vị trí 21 , power thực tế là 6 nhưng mô hình đánh giá cao với dự đoán xấp xỉ 38.57



Hình 26: Biểu đồ phân tán giữa giá trị dự đoán và giá trị thực tế

Mức độ tập trung của các điểm dữ liệu càng gần với đường thẳng y = x càng tốt. Có thể thấy đa số các điểm tập trung khá chặt chẽ xung quanh đường thẳng nhưng không quá gần, một số điểm còn nằm rất xa cho sự khác biệt giữa dự đoán của mô hình với giá trị power thực tế.
Nhóm tính được chỉ số MSE của mô hình là 5170.5232, một giá trị MSE lớn như thế này thường được coi là không tốt, chỉ ra rằng mô hình không dự đoán tốt trên dữ liệu kiểm thử. Do đó, cần phải có các điều chỉnh thích hợp để mô hình cải thiện các dự đoán này.

4.4 Mở rộng: Hồi quy đa thức

Chúng ta có thể mô hình hóa giá trị kì vọng của y dưới dạng đa thức bậc n, mang lại mô hình hồi quy đa thức tổng quát:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon \ .$$

Nhóm tiến hành xây dựng mô hình hóa giá trị kỳ vọng của y dưới dạng đa thức bậc n với mô hình hồi quy đa thức.

4.4.1 Hồi quy đa thức bậc 2

```
Call:
lm(formula = power ~ poly(lithography, 2) + poly(number_cores,
    2) + poly(number_threads, 2) + poly(cache, 2) + poly(Pfrequency,
    2) + poly(instruction_set, 2) + poly(bandwidth, 2), data = cleaned_data)

Residuals:
    Min      1Q  Median      3Q     Max
-236.97  -25.90   -5.59   10.22 1738.47

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                  89.225      2.124  42.002  < 2e-16 ***
poly(lithography, 2)1      2296.790    206.367  11.130  < 2e-16 ***
poly(lithography, 2)2      -545.133    128.763  -4.234 2.39e-05 ***
poly(number_cores, 2)1     2010.123    225.308   8.922  < 2e-16 ***
poly(number_cores, 2)2    -1049.796    253.487  -4.141 3.58e-05 ***
poly(number_threads, 2)1   -776.228    289.178  -2.684  0.00732 **
poly(number_threads, 2)2   -102.442    116.507  -0.879  0.37935
poly(cache, 2)1            -134.133    118.244  -1.134  0.25676
poly(cache, 2)2              50.811    107.214   0.474  0.63560
poly(Pfrequency, 2)1        523.491    112.509   4.653 3.46e-06 ***
poly(Pfrequency, 2)2       -599.652    107.936  -5.556 3.09e-08 ***
poly(instruction_set, 2)1  -791.044    156.637  -5.050 4.77e-07 ***
poly(instruction_set, 2)2  -598.490    121.009  -4.946 8.14e-07 ***
poly(bandwidth, 2)1         113.909    106.300   1.072  0.28403
poly(bandwidth, 2)2         -23.852    104.812  -0.228  0.82000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.5 on 2268 degrees of freedom
Multiple R-squared:  0.2376,    Adjusted R-squared:  0.2329
F-statistic: 50.48 on 14 and 2268 DF,  p-value: < 2.2e-16
```

Hình 27: Kết quả của mô hình hồi quy tuyến tính bậc 2

Output hiển thị kết quả của hồi quy tuyến tính đa biến trong đó biến phụ thuộc power được mô hình hóa dưới dạng hàm đa thức bậc hai của các biến lithography, number_cores, number_threads, cache, bandwidth, Pfrequency và instruction_set.
R bình phương của mô hình là 0.2376 nghĩa là chỉ có 23.76 biến power được giải thích bằng biến đã chọn. Thống kê F-statistic là 50.48 và p gần bằng 0 nên cho thấy rằng có ít nhất 1 biến là yếu tố dự đoán có liên quan đến power.

4.4.2 Hồi quy đa thức bậc 3

Do biến instruction_set không có nhiều hơn 3 giá trị khác nhau nên nhóm sẽ loại bỏ biến này ra khỏi mô hình, nhóm cũng sẽ thực hiện tương tự với mô hình đa thức bậc 4 ở mục 4.4.3

```
Call:
lm(formula = power ~ poly(lithography, 3) + poly(number_cores,
    3) + poly(number_threads, 3) + poly(cache, 3) + poly(Pfrequency,
    3) + poly(bandwidth, 3), data = cleaned_data)

Residuals:
    Min      1Q  Median      3Q     Max
-257.34  -28.75   -9.96   12.65 1769.31

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 89.225      2.114  42.214  < 2e-16 ***
poly(lithography, 3)1     2592.825    191.245  13.558  < 2e-16 ***
poly(lithography, 3)2     -213.127    138.660  -1.537   0.1244
poly(lithography, 3)3     -859.602    110.692  -7.766 1.22e-14 ***
poly(number_cores, 3)1    1258.624    232.916   5.404 7.21e-08 ***
poly(number_cores, 3)2    -268.137    258.100  -1.039   0.2990
poly(number_cores, 3)3    -163.544    147.024  -1.112   0.2661
poly(number_threads, 3)1    23.403    286.102   0.082   0.9348
poly(number_threads, 3)2  -292.886    149.346  -1.961   0.0500 *
poly(number_threads, 3)3    52.545    124.417   0.422   0.6728
poly(cache, 3)1            105.304    112.339   0.937   0.3487
poly(cache, 3)2            -73.323    104.788  -0.700   0.4842
poly(cache, 3)3            109.728    123.422   0.889   0.3741
poly(Pfrequency, 3)1       271.859    116.128   2.341   0.0193 *
poly(Pfrequency, 3)2      -714.364    110.060  -6.491 1.05e-10 ***
poly(Pfrequency, 3)3       472.417    112.395   4.203 2.73e-05 ***
poly(bandwidth, 3)1        -25.077    108.582  -0.231   0.8174
poly(bandwidth, 3)2         90.032    105.961   0.850   0.3956
poly(bandwidth, 3)3       -208.996    113.687  -1.838   0.0661 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101 on 2264 degrees of freedom
Multiple R-squared:  0.2466,    Adjusted R-squared:  0.2406
F-statistic: 41.16 on 18 and 2264 DF,  p-value: < 2.2e-16
```

Hình 28: Kết quả của mô hình hồi quy tuyến tính bậc 3

Output của mô hình hồi quy tuyến tính đa biến trong đó biến phụ thuộc power được mô hình hóa dưới dạng hàm đa thức bậc ba của các biến lithography, number_cores, number_threads, cache, bandwidth và Pfrequency.

R bình phương của có độ lớn chỉ bằng 0.2466, nghĩa là chỉ có 24.66% biến power được giải thích bằng biến đã chọn. Thống kê F-statistic là 41.16 và p gần bằng 0 cho thấy có ít nhất 1 biến là yếu tố dự đoán có liên quan đến power.

4.4.3 Mô hình đa thức bậc 4

```
Residuals:
    Min      1Q  Median      3Q     Max
-342.14  -26.11   -9.32   10.66 1768.51

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 89.225      2.107  42.349  < 2e-16 ***
poly(lithography, 4)1     2657.319    194.834  13.639  < 2e-16 ***
poly(lithography, 4)2     -216.628    140.272  -1.544  0.12264
poly(lithography, 4)3     -909.137    111.275  -8.170 5.07e-16 ***
poly(lithography, 4)4      -82.822    115.149  -0.719  0.47205
poly(number_cores, 4)1    1074.750    366.132   2.935  0.00336 **
poly(number_cores, 4)2    -157.657    419.372  -0.376  0.70700
poly(number_cores, 4)3     -45.884    253.651  -0.181  0.85647
poly(number_cores, 4)4    -101.498    282.892  -0.359  0.71979
poly(number_threads, 4)1   257.669    576.944   0.447  0.65520
poly(number_threads, 4)2  -208.538    264.203  -0.789  0.43001
poly(number_threads, 4)3     2.900    156.361   0.019  0.98520
poly(number_threads, 4)4    49.729    109.544   0.454  0.64990
poly(cache, 4)1            121.005    114.829   1.054  0.29210
poly(cache, 4)2            -66.583    106.295  -0.626  0.53111
poly(cache, 4)3            137.386    126.749   1.084  0.27852
poly(cache, 4)4           -161.487    110.187  -1.466  0.14290
poly(Pfrequency, 4)1       235.858    120.582   1.956  0.05059 .
poly(Pfrequency, 4)2      -729.254    114.134  -6.389 2.02e-10 ***
poly(Pfrequency, 4)3       480.020    114.343   4.198 2.80e-05 ***
poly(Pfrequency, 4)4       417.388    105.108   3.971 7.38e-05 ***
poly(bandwidth, 4)1        -22.180    110.043  -0.202  0.84028
poly(bandwidth, 4)2         86.500    106.560   0.812  0.41702
poly(bandwidth, 4)3       -204.280    116.087  -1.760  0.07859 .
poly(bandwidth, 4)4         90.230    110.781   0.814  0.41545
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 100.7 on 2258 degrees of freedom
Multiple R-squared:  0.2533,    Adjusted R-squared:  0.2454
F-statistic: 31.92 on 24 and 2258 DF,  p-value: < 2.2e-16
```

Hình 29: Kết quả của mô hình hồi quy tuyến tính bậc 4

Đầu ra trình bày kết quả của hồi quy tuyến tính đa biến bậc 4 trong đó biến phụ thuộc power được mô hình hóa bởi các biến độc lập lithography, number_cores, number_threads, cache, bandwidth và Pfrequency

Độ lớn R bình phương của mô hình là 0,2533, cho thấy có khoảng 25,33% biến thể trong power có thể được giải thích bằng mô hình này.

Thống kê F là 31.92 với giá trị p gần bằng 0, cho thấy ít nhất một trong các biến độc lập có quan hệ khá chặt chẽ, đáng kể đến biến phụ thuộc.

4.4.4 Mô hình hồi quy đa thức phù hợp nhất và hệ số hồi quy đa thức.

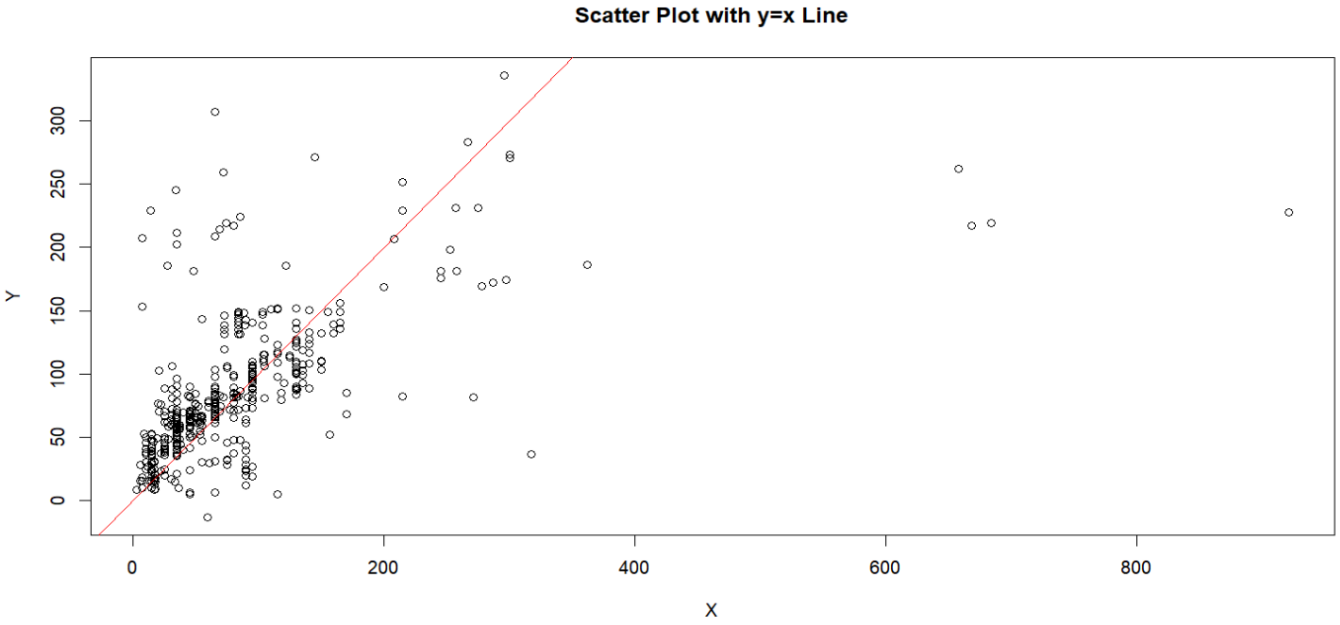
Với việc có chỉ số R-Square cao nhất nên nhóm quyết định chọn phương trình đa thức bậc 4:

```
power = 89.2247 + 2657.319 *lithography + -216.6285 *lithography^2 + -909.1375 *lithography^3
+ -82.82193 *lithography^4+ 1074.75 *number_cores + -157.6568 *number_cores^2
+ -45.88354 *number_cores^3 + -101.4981 *number_cores^4 + 257.6687 *number_threads
+ -208.5378 *number_threads^2 + 2.900209 *number_threads^3 + 49.72889 *number_threads^4
+ 121.0052 *cache + -66.58336 *cache^2 + 137.3859 *cache^3 + -161.487 *cache^4
+ 235.8576 *Pfrequency + -729.2536 *Pfrequency^2 + 480.0203 *Pfrequency^3
+ 417.3879 *Pfrequency^4 + -22.18042 *bandwidth + 86.50002 *bandwidth^2
+  -204.28 *bandwidth^3 +  90.23032 *bandwidth^4
```

Hình 30: Hệ số của phương trình hồi quy bậc 4

4.5 So sánh mô hình

Nhóm sẽ phân tích biểu đồ phân tán và chỉ số MSE của mô hình hồi quy đa thức bậc 4 để so sánh với mô hình hồi quy đa biến bậc nhất.



Hình 31: So sánh giá trị kiểm thử và dự đoán

Có thể thấy mật độ tập trung của các điểm dữ liệu gần đường thẳng y = x hơn so với mô hình hồi quy bậc nhất. Các điểm dữ liệu được tập trung khá chặt chẽ cho thấy mô hình này có khả năng dự đoán chính xác hơn. Tuy nhiên, vẫn có một số điểm dữ liệu nằm ngoài và rất xa đường thẳng, nghĩa là khả năng dự đoán của mô hình vẫn chưa cao. Nhóm tính được chỉ số MSE của mô hình là 4739.1242 giảm rõ rệt so với chỉ số MSE của mô hình quy bậc nhất, đây là một tín hiệu tích cực về khả năng dự đoán của mô hình mới.

Thông qua những phân tích trên, nhóm kết luận mô hình hồi quy đa thức hiệu quả hơn vì mật độ tập trung của các điểm dữ liệu xung quanh đường thẳng y = x trong biểu đồ phân tán chặt chẽ hơn và có chỉ số MSE thấp hơn hồi quy đa biến bậc nhất.

5. Thảo luận

5.1 Ưu điểm

Mô hình hồi quy tuyến tính rất dễ hiểu và giải thích, đặc biệt là khi có một số biến độc lập, có thể mở rộng mô hình để xử lý tình huống phức tạp hơn bằng cách thêm các biến độc lập hoặc biến đa thức, có thể sử dụng các phương pháp để lựa chọn biến quan trọng và có thể mô hình hóa mối quan hệ giữa biến phụ thuộc và nhiều biến độc lập.

5.2 Nhược điểm

Mô hình giả định rằng mối quan hệ giữa biến phụ thuộc và biến độc lập là tuyến tính. Điều này có thể không phù hợp với mọi dữ liệu, mô hình có thể bị ảnh hưởng nhiều bởi dữ liệu nhiễu hoặc giá trị ngoại lai và nếu có tương quan mạnh giữa các biến độc lập, mô hình có thể không hoạt động hiệu quả.

6. Mã và nguồn dữ liệu

6.1 Mã nguồn

Link: Mã nguồn

6.2 Nguồn dữ liệu

Link: Dữ liệu

7. Tài liệu tham khảo

[1] Sách Douglas C. Montgomery, Applied Statistics and Probability for Engineers
[2] John Verzani, simpleR - Using R for Introductory Statistics
[3] Bộ Giáo dục đào tạo, Giáo trình xác suất và thống kê, NXB: Đại học Quốc gia Thành phố Hồ Chí Minh.