BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC THĂNG LONG

BÀI TẬP LỚN

THU THẬP VÀ PHÂN TÍCH DỮ LIỆU GIÁ BÁN CỦA HAI CỬA HÀNG BÁN ĐIỆN THOẠI CELLPHONES VÀ THẾ GIỚI DI ĐỘNG

SINH VIÊN THỰC HIỆN: **HOÀNG ĐỨC LINH – A41406**
NGUYỄN HỮU LÂM – A41856
NGÀNH: **TRÍ TUỆ NHÂN TẠO**

HÀ NỘI-2023

MỤC LỤC

CHƯƠNG 1. GIỚI THIỆU CHUNG **1**

1.1. Đối tượng **1**

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT **2**

2.1. Khoa học dữ liệu là gì **2**

2.2. Các gói thư viện hỗ trợ **2**

2.2.1. Pandas *2*

2.2.2. Numpy *2*

2.2.3. Matplotlib *2*

2.3. Một số thuật ngữ về điện thoại và laptop **2**

2.3.1. RAM *2*

2.3.2. ROM *3*

2.3.3. Chipset *3*

2.3.4. Hệ điều hành (Operating System) *4*

CHƯƠNG 3. THU THẬP DỮ LIỆU **5**

3.1. Nguồn dữ liệu **5**

3.2. Thu thập dữ liệu **5**

3.2.1. Điện thoại *6*

3.2.2. Laptop *8*

CHƯƠNG 4. XỬ LÍ DỮ LIỆU **10**

4.1. Kiểm tra và đánh giá chất lượng dữ liệu **10**

4.1.1. Thế giới di động *10*

4.1.1.1. Laptop 10

4.1.2. CellphoneS *11*

4.1.2.1. Laptop 11

4.2. Phương pháp xử lí **12**

4.2.1. Làm sạch dữ liệu *12*

4.2.2. Kết quả *12*

4.2.2.1. Thế giới di dộng (Laptop) 12
4.2.2.2. CellphoneS (Laptop) 12
4.2.2.3. Thê giới di động (Điện thoại) 13
4.2.2.4. CellphoneS (Điện thoại) 13
CHƯƠNG 5. EXPLORE DATA ANLYSIS 14
5.1. Laptop 14
5.1.1. Giá cả sản phẩm 14
5.1.2. Loại Ram được sử dụng 16
5.1.3. Loại Rom được sử dụng 17
5.1.4. Chip và card đồ họa 18
5.2. Điện thoại 19
5.2.1. Phân khúc giá điện thoại 19
5.2.2. RAM 21
5.2.3. ROM 23
5.2.4. Hệ điều hành (OS) 24
5.2.5. Kích thước màn hình 26
CHƯƠNG 6. TỔNG KẾT 27
6.1. Phân chia công việc 27
6.2. Khó khăn trong quá trình thực hiện 27
6.2.1. Quá trình thu thập dữ liệu 27
6.2.2. Xử lý dữ liệu 27

DANH MỤC BẢNG, HÌNH, ẢNH, ĐỒ THỊ

Ảnh 5.1. Biểu đồ thể hiện số lượng sản phẩm theo mức giá ở CellphoneS 15
Ảnh 5.2. Biểu đồ thể hiện số lượng sản phẩm theo mức giá ở Thế giới di động 15
Ảnh 5.3. Biểu đồ sản phẩm được phân theo Ram ở thế giới di động..... 16
Ảnh 5.4. Biểu đồ phân loại sản phẩm theo Ram ở CellphoneS 17
Ảnh 5.5. Biểu đồ thể hiện sản phẩm theo Rom ở Thế giới di động 17
Ảnh 5.6. Biểu đồ thể hiện sản phẩm theo cpu ở thế giới di động 18
Ảnh 5.7. Biểu đồ thể hiện sản phẩm theo Card đồ họa tích hợp ở CellphoneS 19
Ảnh 5.8 Biểu đồ thể hiện giá sản phẩm điện thoại ở CellphoneS 20
Ảnh 5.9 Biểu đồ thể hiện giá sản phẩm điện thoại ở TGDD..... 21
Ảnh 5.10 Biểu đồ thể hiện số lượng RAM có trong sản phẩm ở TGDD 22
Ảnh 5.11 Biểu đồ thể hiện số lượng RAM có trong sản phẩm ở CellphoneS 22
Ảnh 5.12 Biểu đồ thể hiện số dung lượng của sản phẩm tại CellphoneS 23
Ảnh 5.13 Biểu đồ thể hiện số dung lượng có trong sản phẩm ở TGDD 23
Ảnh 5.14 Biểu đồ số lượng hệ điều hành Android và iOS ở CellphoneS 24
Ảnh 5.15 Biểu đồ số lượng hệ điều hành Android và iOS ở TGDD..... 25
Ảnh 5.16 Biểu đồ số lượng kích thước màn hình ở CellphoneS..... 26

CHƯƠNG 1. GIỚI THIỆU CHUNG

1.1. Đối tượng

Đề tài giới hạn đối tượng nghiên cứu trong phạm vi các loại điện thoại di động và laptop thông dụng có trên thị trường ở hai cửa hàng bán thiết bị di động là CellphoneS và Thế Giới Di Động. Sử dụng các phương pháp phân tích số liệu thống kê và biểu đồ để so sánh giá cả của các thiết bị di động được bày bán ở cả hai cửa hàng.

Dữ liệu về sản phẩm điện thoại và laptop được lấy trong khoảng thời gian mới nhất 6/2023.

Tài liệu giới hạn phạm vi đối tượng người đọc là những người quan tâm đến lĩnh vực công nghệ, thị trường điện thoại di động, laptop và các doanh nghiệp, các hãng sản xuất điện thoại, laptop lớn muốn nghiên cứu thị trường để đưa ra chiến lược kinh doanh.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1. Khoa học dữ liệu là gì

Khoa học dữ liệu là sự kết hợp giữa toán học, thống kê, lập trình chuyên biệt, phân tích nâng cao, trí tuệ nhân tạo (AI) và máy học với kiến thức chuyên môn về chủ đề cụ thể để khám phá những thông tin chi tiết hữu ích ẩn chứa trong các tập dữ liệu. Những hiểu biết sâu này có thể được sử dụng để định hướng việc ra quyết định và lập kế hoạch chiến lược.

2.2. Các gói thư viện hỗ trợ

2.2.1. Pandas

Pandas là một mô-đun mạnh mẽ được tối ưu hóa trên Numpy và cung cấp một tập hợp các cấu trúc dữ liệu đặc biệt phù hợp với chuỗi thời gian và phân tích dữ liệu kiểu bảng tính (giống bảng tổng hợp trong Excel).

- Pandas là một thư viện của Python được sử dụng để làm việc với các tập dữ liệu. Nó có các chức năng phân tích, làm sạch, khám phá và khai thác dữ liệu.
- Cho phép phân tích dữ liệu lớn và đưa ra kết luận dựa trên lý thuyết về thống kê. Pandas có thể dọn dẹp các tập dữ liệu lộn xộn, làm cho chúng dễ đọc và trở nên phù hợp. Dữ liệu chuẩn rất quan trọng trong khoa học dữ liệu.

2.2.2. Numpy

Để sử dụng một thư viện khoa học đã biên dịch, bộ nhớ được cấp phát trong trình thông dịch Python bằng cách nào đó phải biến được thư viện này làm đầu vào. Hơn nữa, đầu ra từ các thư viện này cũng phải trả về trình thông dịch Python. Trao đổi bộ nhớ hai chiều này về cơ bản là chức năng cốt lõi của mô-đun Numpy(mảng số trong Python). Numpy là tiêu chuẩn thực tế cho mảng số trong Python. Nó xuất hiện như một nỗ lực của Travis Oliphant và những người khác nhằm thống nhất các mảng số đã có trong Python.

2.2.3. Matplotlib

Matplotlib là một thư viện vẽ sơ đồ có sẵn cho ngôn ngữ lập trình Python như một thành phần của NumPy, một tài nguyên xử lý số dữ liệu lớn. Matplotlib sử dụng một API hướng đối tượng để nhúng các sơ đồ trong các ứng dụng Python.

2.3. Một số thuật ngữ về điện thoại và laptop

2.3.1. RAM

RAM (Random access memory) là bộ nhớ ngắn hạn của, nơi lưu trữ dữ liệu mà bộ xử lý hiện đang sử dụng. Thiết bị của bạn có thể truy cập bộ nhớ RAM nhanh hơn nhiều

so với dữ liệu trên ổ cứng, SSD hoặc thiết bị lưu trữ dài hạn khác, đó là lý do dung lượng RAM rất quan trọng đối với hiệu suất hệ thống.

Các thiết bị cũ hoặc cấp thấp hơn thường có RAM 4GB, trong khi các máy cao cấp hơn(và đắt tiền hơn) có RAM 8GB hoặc 16GB. Và bạn có thể tìm thấy các máy tính chuyên nghiệp, đồ họa, cao cấp có nhiều RAM hơn.

4 GB RAM: Nếu bạn sử dụng thiết bị để duyệt web, làm việc với các ứng dụng Office tiêu chuẩn và chỉnh sửa ảnh nhẹ, sẽ ổn với 4GB RAM.

RAM 8 GB: Người dùng đa nhiệm nặng hoặc game thủ nhẹ nhàng nên chọn thiết bị có RAM 8GB.

Hơn 16 GB RAM: Một số tác vụ sử dụng nhiều điện toán, chẳng hạn như chơi game nghiêm túc, chỉnh sửa video, lập trình hoặc phải chạy đồng thời nhiều tác vụ chuyên sâu. Những người yêu cầu hiệu suất ở mức độ chuyên nghiệp mà không bị chậm sẽ cần RAM 16+GB.

2.3.2. ROM

Bộ nhớ chỉ đọc (ROM) là một loại bộ nhớ không thay đổi được sử dụng trong máy tính và các thiết bị điện tử khác. Dữ liệu được lưu trữ trong ROM không thể được sửa đổi bằng điện tử sau khi sản xuất. Bộ nhớ chỉ đọc hữu ích cho việc lưu trữ phần mềm hiếm khi thay đổi trong suốt vòng đời của hệ thống, đôi khi được gọi là phần sụn (firmware). Các ứng dụng phần mềm cho các thiết bị lập trình có thể được phân phối dưới dạng các hộp mực bổ trợ có chứa bộ nhớ chỉ đọc (băng trò chơi video game).

"ROM (hoặc flash) được sử dụng để lưu trữ phần sụn khởi động (firmware) cơ bản cho bộ xử lý chính, cũng như các phần firmware khác nhau cần thiết để điều khiển các thiết bị độc lập như card đồ họa, đĩa cứng, ổ đĩa DVD, màn hình, v.v.. trong hệ thống.

Ngày nay, nhiều bộ nhớ "chỉ đọc" này - đặc biệt là BIOS - thường được thay thế bằng bộ nhớ Flash, để cho phép lập trình lại khi cần nâng cấp firmware. Tuy nhiên, các hệ thống phụ đơn giản (chẳng hạn như bàn phím hoặc một số bộ điều khiển giao tiếp trong các mạch tích hợp trên bo mạch chính chẳng hạn) có thể sử dụng ROM (lập trình một lần).

2.3.3. Chipset

Chipset là một tập hợp chip.Với máy tính thì khi nhắc đến chipset dùng để đề cập đến chip đặc biệt trên mainboard hoặc trên các card mở rộng. Đối với PC, chip để nhắc đến **chip cầu bắc và chip cầu nam**, là chip trên bo mạch chính. Chip bắc cầu là chip nằm trên cùng phía bắc của bo mạch chủ. Nó vận hành và quản lý tác vụ nặng hơn so với chip nam cầu. Phần về 2 loại chip này mình sẽ nói rõ hơn ở phần sau.

Chipset được xem là nơi kết nối giữa phần mềm và phần cứng giữa mainboard. Và giúp tìm ra thiết bị ngoại vi phù hợp cho máy tính. Những thiết bị ngoại vi này bao gồm CPU, RAM, ổ cứng…

- Bộ xử lý trung tâm (CPU): Là thành phần quan trọng nhất của chip điện thoại, chịu trách nhiệm xử lý và điều khiển các tác vụ chính trên điện thoại.
- Bộ xử lý đồ họa (GPU): Là thành phần quản lý các tác vụ liên quan đến đồ họa, bao gồm chơi game, xem video và hiển thị ảnh.
- Bộ điều khiển bộ nhớ (Memory Controller): Là thành phần quản lý các tác vụ liên quan đến bộ nhớ của điện thoại, bao gồm quản lý bộ nhớ RAM và ROM.

Một số nhà cung cấp Chip phổ biến như: Qualcomm, Mediatek, Intel, AMD, Samsung, Apple...

2.3.4. Hệ điều hành (Operating System)

Hệ điều hành là một phần mềm nền tảng cho phép điều hành, quản lý toàn bộ các thành phần khác trên một thiết bị điện tử. Ví dụ như Linux và Windows hay IOS,…

Một số hệ điều hành phổ biến như:

- Hệ điều hành Android: Hệ điều hành Android là hệ điều hành phổ biến nhất hiện nay. Nó là một hệ điều hành di động dựa trên Linux Kernel và phần mềm mã nguồn mở. Hệ điều hành Android được phát triển bởi Google. Thiết bị Android đầu tiên được ra mắt vào năm 2008.
- Hệ điều hành iPhone/ iOS: iOS được phát triển bởi Apple inc để sử dụng trên thiết bị của họ. Hệ điều hành iOS là hệ điều hành phổ biến nhất hiện nay. Nó là một hệ điều hành rất an toàn. Hệ điều hành iOS không khả dụng cho bất kỳ điện thoại di động nào khác.
- Hệ điều hành Windows: Hệ điều hành window là hệ điều hành được phát triển bởi Microsoft. Nó được thiết kế cho máy tính không phải macbook và hỗ trợ đầy đủ các ứng dụng bạn cần và nhiều hơn rất nhiều, vượt trội hơn so với các hệ điều hành khác.

CHƯƠNG 3. THU THẬP DỮ LIỆU

3.1. Nguồn dữ liệu

Đường dẫn tới hai trang web CellphoneS và Thế Giới Di Động:

- https://cellphones.com.vn/ và https://www.thegioididong.com/

Khi thu thập dữ liệu, thông tin cần quan tâm là các giá trị có thể ảnh hưởng đến giá điện thoại như: thương hiệu, bộ xử lý (chip), công nghệ màn hình.

Do lượng dữ liệu tương đối lớn và rộng nên để dữ liệu thu thập có ích cần đặt ra một số tiêu chí:

- Phụ thuộc vào việc các sản phẩm có chứa đầy đủ thông tin hay không, lựa chọn các thông tin mà đa phần các sản phẩm trên web đều có.
- Dữ liệu về thông số cần viết có quy tắc để dễ dàng lấy được giá trị cần thiết.

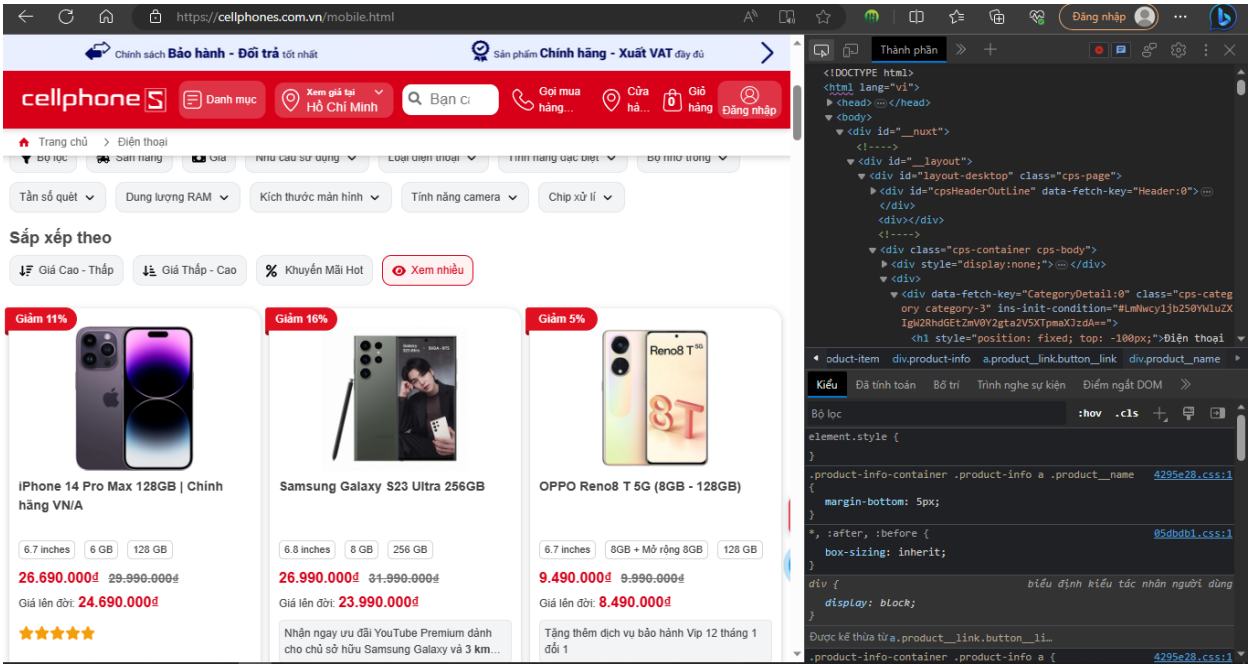
Khi thu thập dữ liệu thô thì thường có nhiều chữ xuất hiện kèm trong dữ liệu cần quan tâm nên kiểu dữ liệu đa phần ở dạng chuỗi. Dữ liệu sau khi làm sạch sẽ được thay đổi kiểu dữ liệu.

Trang web chứa nhiều thông tin về điện thoại uy tín ở Việt Nam: thegioididong.com, cellphones.com.vn... Tuy nhiên dữ liệu đa phần chỉ gồm các sản phẩm khoảng 2-3 năm trở lại đây và trên thị trường Việt Nam. Sau khi tìm hiểu thì đa phần chỉ khoảng dưới 300 sản phẩm về điện thoại và 600 sản phẩm về laptop và điện thoại. Vì vậy, có thể sẽ có nhiều thông tin bị thiếu hoặc bị lẫn với nhau.

3.2. Thu thập dữ liệu

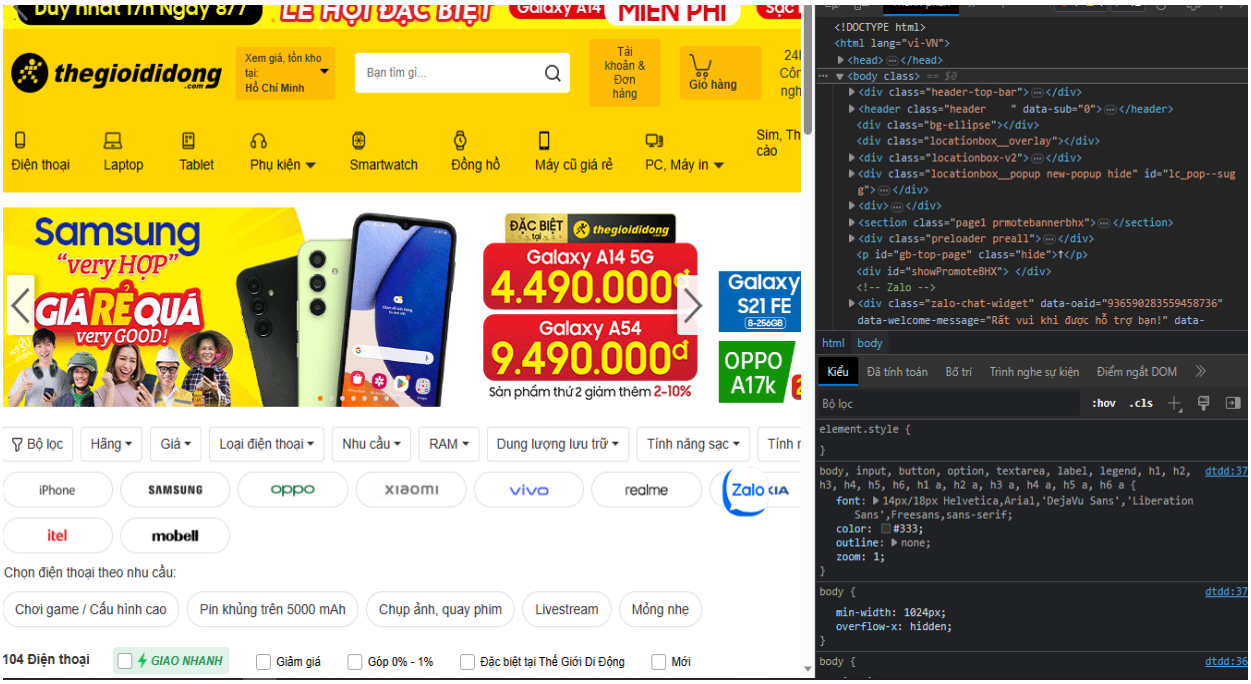
Phương pháp thu thập thông qua HTML, sử dụng thư viện selenium để truy cập vào trang web và tương tác với các yếu tố cần thu thập.

3.2.1. *Điện thoại*



Ảnh 3.1. Ảnh lấy dữ liệu từ trang web CellphoneS thông qua html của trang web

Đối với trang web CellphoneS, tiến hành thu thập dữ liệu về điện thoại: Tên sản phẩm, giá cả, kích thước màn hình, ram, dung lượng, hệ điều hành thông qua đường Link của các sản phẩm đã được thu thập trước đó.



Ảnh 3.2 Ảnh lấy dữ liệu điện thoại thông qua html của trang web TGDD

Đối với trang web TGDD, tiến hành thu thập dữ liệu về điện thoại: Tên sản phẩm, giá cả, ram, dung lượng, hệ điều hành thông qua đường Link của các sản phẩm đã được thu thập trước đó.

Kết quả thu được:

– CellphoneS:

+ Số lượng mẫu: 205 mẫu

	Name	Inches	RAM	ROM	OS	Price	Links
0	OPPO Reno7 4G (8GB - 128GB)	6.43 inches	8 GB	128 GB	Android 11, ColorOS 12	79900000.0	https://cellphones.com.vn/oppo-reno7-128gb.html
1	OPPO A16K	6.52 inches	3 GB	32 GB	Android 11	36900000.0	https://cellphones.com.vn/oppo-a16k.html
2	OPPO A95	6.43 inches	8 GB	128 GB	Android 11	5090000.0	https://cellphones.com.vn/oppo-a95.html
3	OPPO Reno7 (5G)	6.43 inches	8 GB	256 GB	Android 11, ColorOS 12	129900000.0	https://cellphones.com.vn/oppo-reno-7.html
4	OPPO Reno7 Z (5G)	6.43 inches	8 GB	128 GB	Android 11 - ColorOS 12	99900000.0	https://cellphones.com.vn/oppo-reno7-z.html
...	...	...	...	...	...	...	...
200	Vsmart Bee 5	6.0 inches	Không Có	16 GB	9.0 (Pie)	1590000.0	https://cellphones.com.vn/vsmart-bee-5.html
201	Huawei P30 Lite	6.15 inches	6 GB	128 GB	Android v9.0 (Pie)	7490000.0	https://cellphones.com.vn/huawei-p30-lite-1.html
202	Huawei P30 Pro	6.4 inches	8 GB	256 GB	Android v9.0 (Pie)	23990000.0	https://cellphones.com.vn/huawei-p30-pro-1.html
203	Huawei P30	6.1 inches	8 GB	128 GB	Android v9.0 (Pie)	17990000.0	https://cellphones.com.vn/huawei-p30-1.html
204	Samsung Galaxy A51 6GB	6.5 inches	Không Có	64 GB	10	7990000.0	https://cellphones.com.vn/samsung-galaxy-a51-1...

205 rows x 7 columns

Ảnh 3.3 Thông tin dữ liệu Điện thoại của CellphoneS

– Thế giới di động

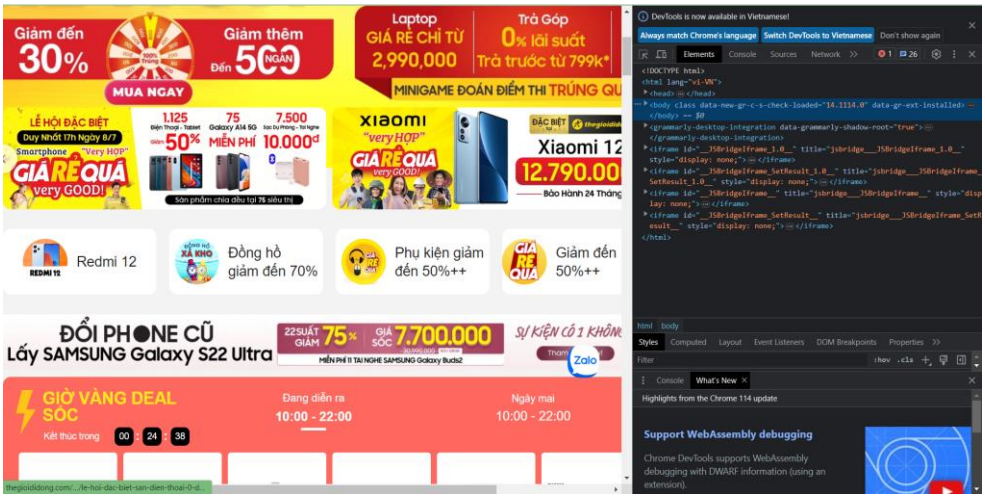
+ Số lượng mẫu: 97 mẫu

	Name	Price	OS	RAM	ROM	Links
0	Điện thoại Samsung Galaxy A14 6GB	4990000.0	Android 13	6 GB	128 GB	https://www.thegioididong.com/dtdd/samsung-gal...
1	Điện thoại OPPO Find N2 Flip 5G	19990000.0	Android 13	8 GB	256 GB	https://www.thegioididong.com/dtdd/oppo-find-n...
2	Điện thoại iPhone 14 Pro Max 128GB	26680000.0	iOS 16	6 GB	128 GB	https://www.thegioididong.com/dtdd/iphone-14-p...
3	Điện thoại iPhone 14 Pro 128GB	24790000.0	iOS 16	6 GB	128 GB	https://www.thegioididong.com/dtdd/iphone-14-pro
4	Điện thoại Vivo Y36	6990000.0	Android 13	8 GB	256 GB	https://www.thegioididong.com/dtdd/vivo-y36
...	...	...	...	...	...	...
92	Điện thoại Mobell F209 4G	620000.0	Không Có	Không Có	Không Có	https://www.thegioididong.com/dtdd/mobell-f209
93	Điện thoại Masstel IZI 26 4G	600000.0	Không Có	Không Có	Không Có	https://www.thegioididong.com/dtdd/masstel-izi...
94	Điện thoại Itel it9010	580000.0	Không Có	Không Có	Không Có	https://www.thegioididong.com/dtdd/itel-it9010
95	Điện thoại Masstel Lux 10 4G	570000.0	Không Có	Không Có	Không Có	https://www.thegioididong.com/dtdd/masstel-lux-10
96	Điện thoại Masstel IZI 10 4G	390000.0	Không Có	Không Có	Không Có	https://www.thegioididong.com/dtdd/masstel-izi...

97 rows x 6 columns

Ảnh 3.4 Thông tin dữ liệu Điện thoại của TGDD

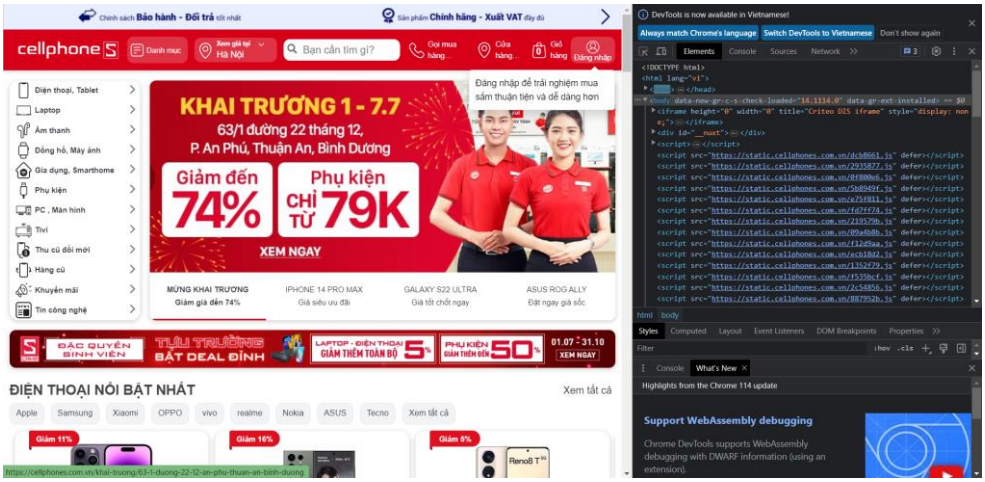
3.2.2. Laptop



Ảnh 3.5. Lấy dữ liệu thông qua html của trang web TGDĐ

Các thông tin lấy cần là những thông tin mà đa phần mỗi sản phẩm trên web đều có.

Tiến hành thu thập dữ liệu về laptop đối với trang Thế giới di động: Tên sản phẩm, giá cả, ram, rom, cpu thông qua các đường Link của sản phẩm đã được thu thập trước đó.



Ảnh 3.6. Lấy dữ liệu thông qua html của trang web CellphoneS

Đối với trang web CellphoneS, tiến hành thu thập dữ liệu về laptop: Tên sản phẩm, giá cả, ram, card đồ họa thông qua đường Link của các sản phẩm đã được thu thập trước đó.

Kết quả thu thập được:

– Thế giới di động:

+ Số lượng mẫu: 267 mẫu

	Name	Price	Cpu	Ram	Rom
0	Laptop Dell Inspiron 16 5620 i7 1255U/8GB/512G...	249900000.0	Corei7	8GB	512GBSSD
1	Laptop HP Pavilion X360 14 ek0134TU i5 1235U/8...	184900000.0	Corei5	8GB	512GBSSD
2	Laptop HP Envy X360 13 bf0090TU i7 1250U/16GB/...	279900000.0	Corei7	16GB	512GBSSD
3	Laptop HP EliteBook 630 G9 i7 1255U/16GB/512GB...	241900000.0	Corei7	16GB	512GBSSD
4	Laptop Asus Zenbook 14 OLED UX3402ZA i5 1240P/...	209900000.0	Corei5	8GB	512GBSSD
...	...	...	...	...	...
261	Laptop Apple MacBook Air 13 inch M2 2022 8-cor...	273900000.0	AppleM2	8GB	256GBSSD
262	Laptop Apple MacBook Pro 13 inch M2 2022 8-cor...	295900000.0	AppleM2	8GB	256GBSSD
263	Laptop Apple MacBook Air 13 inch M2 2022 8-cor...	379900000.0	AppleM2	8GB	256GBSSD
264	Laptop Apple MacBook Pro 13 inch M2 2022 8-cor...	383900000.0	AppleM2	8GB	256GBSSD
265	Laptop Apple MacBook Pro 13 inch M1 2020 8-cor...	349900000.0	AppleM1	8GB	256GBSSD

266 rows × 5 columns

Ảnh 3.7. Thông tin cột của laptop ở Thế giới di động

– CellphoneS:

+ Số lượng mẫu: 476 mẫu

	Name	Card	Price	Ram	Link
0	Laptop Acer Aspire 3 A315-58-53S6 NX.AM0SV.005	INTEL	15990000.0	8GB	https://cellphones.com.vn/laptop-acer-aspire-3...
1	Laptop MSI Crosshair 15 B12UEZ-460VN	NVIDIA	34990000.0	16GB	https://cellphones.com.vn/laptop-msi-gaming-cr...
2	Laptop Lenovo Ideapad Gaming 3 15IAH7 82S9006YVN	NVIDIA	22300000.0	8GB	https://cellphones.com.vn/laptop-lenovo-ideapa...
3	Laptop Gigabyte G5 GD-51VN123SO	NVIDIA	19990000.0	16GB	https://cellphones.com.vn/laptop-gigabyte-g5-g...
4	Laptop Lenovo Ideapad 3 15IAU7 82RK001GVN	INTEL	13600000.0	8GB	https://cellphones.com.vn/laptop-lenovo-ideapa...
...	...	...	...	...	...
470	Laptop ASUS Gaming ROG Strix G17 G713RW-LL157W	NVIDIA	59990000.0	16GB	https://cellphones.com.vn/laptop-asus-gaming-r...
471	Laptop LG Gram 2022 16ZD90Q-G.AX51A5	INTEL	36990000.0	8GB	https://cellphones.com.vn/laptop-lg-gram-2022-...
472	Laptop LG Gram 2022 14ZD90Q-G.AX31A5	INTEL	23290000.0	8GB	https://cellphones.com.vn/laptop-lg-gram-2022-...
473	Laptop HP 240 G8 3D3H7PA	INTEL	19190000.0	8GB	https://cellphones.com.vn/laptop-hp-240-g8-3d3...
474	Laptop Dell Vostro 5502 NT0X01	NVIDIA	22590000.0	8GB	https://cellphones.com.vn/laptop-dell-vostro-5...

475 rows × 5 columns

Ảnh 3.8. Thông tin cột của laptop ở CellphoneS

Một số lưu ý khi thu thập dữ liệu

Chỉ cần thu thập dữ liệu thô, không cần xử lý trong quá trình thu thập luôn vì có thể có các trường hợp đặc biệt mà phương pháp xử lý đó không phù hợp. Chẳng hạn như một số mẫu dữ liệu ROM, CPU, Card đồ họa sau khi thu thập đã bị sai.

CHƯƠNG 4. XỬ LÍ DỮ LIỆU

4.1. Kiểm tra và đánh giá chất lượng dữ liệu

4.1.1. Thế giới di động

4.1.1.1. Laptop

Số cột: 5

Số hàng: 267

Thông tin về các cột dữ liệu:

+ Name: tên sản phẩm
+ Price: Giá của sản phẩm
+ Cpu: Bộ xử lí trung tâm
+ Ram: Bộ nhớ tạm thời (GB)
+ Rom: Dung lượng lưu trữ (GB/TB)

Vấn đề của tập dữ liệu:

- Kiểu dữ liệu đều là object trong khi một số cột cần là kiểu dữ liệu số nguyên hoặc số thực.
- Name: không có vấn đề về thiếu, trùng lặp hay bất thường
- Rom, Ram không bị thiếu không quá nhiều và cũng có các giá trị nan. Nguyên nhân do là sản phẩm cũ, hoặc mới về hang nên trang web chưa kịp cập nhật
- Cpu bị thiếu nhiều và cũng bị lẫn lộn dữ liệu của màn hình hay pin laptop. Nguyên nhân có thể do bên trang web chưa sắp xếp đồng đều cho bảng thông số hoặc có các máy đời cũ đã ngưng bán hang hoặc máy mới nhưng chưa cập nhật thông tin lên trang web.
- Đa phần dữ liệu bị thiếu thông tin. Nguyên nhân do các dòng máy đời cũ, các máy không có thông tin rõ ràng.
- Các vấn đề mà nguyên nhân do web chứa quá nhiều thông tin chi tiết:
  + Ram chứa them thông tin về chủng loại, thông tin thay thế, nâng cấp.
  + Rom cũng tương tự như vậy, có nhiều thông tin về loại Rom và khả năng nâng cấp của Rom.

+ Cpu có nhiều thông tin về card đồ họa được tích hợp vào cpu và có nhiều chủng cùng loại nhưng khác nhau về hiệu năng

4.1.2. CellphoneS

4.1.2.1. Laptop

Số cột: 5

Số hàng: 476

Thông tin về các cột dữ liệu:

+ Name: tên sản phẩm
+ Price: Giá của sản phẩm
+ Card: Bộ xử lí đồ họa
+ Ram: Bộ nhớ tạm thời (GB)

Vấn đề của tập dữ liệu:

- Kiểu dữ liệu đều là object trong khi một số cột cần là kiểu dữ liệu số nguyên hoặc số thực.
- Name: không có vấn đề về thiếu, trùng lặp hay bất thường.
- Ram không bị thiếu không quá nhiều và cũng có các giá trị nan. Nguyên nhân do là sản phẩm cũ, hoặc mới về hang nên trang web chưa kịp cập nhật
- Rom bị thiếu nhiều, có nhiều giá trị bị nan và bị lần lộn thông tin nên chưa thể lấy được dữ liệu trọn vẹn từ trang web
- Card đồ họa không bị thiếu thông tin, không trùng lặp hay có sự bất thường
- Đa phần dữ liệu bị thiếu thông tin. Nguyên nhân do các dòng máy đời cũ, các máy không có thông tin rõ ràng.
- Các vấn đề mà nguyên nhân do web chứa quá nhiều thông tin chi tiết:
  + Ram chứa them thông tin về chủng loại, thông tin thay thế, nâng cấp.
  + Rom cũng tương tự như vậy, bị lẫn lộn thông tin sản phẩm, chưa thể thu thập hết thông tin về laptop.
  + Card đồ họa có nhiều thông tin về chủng loại, thế hệ và nhiều thông tin khác như khả năng ép xung, khả năng tiêu thụ điện năng….

4.2. Phương pháp xử lí

4.2.1. Làm sạch dữ liệu

Price tiến hành thu thập và xử lí tách chuỗi

Thay thế giá trị bị thiếu ở cột Price và chuyển cột Price sang float

Rom, Ram tiến hành thu thập và xử lí tách chuỗi

Chuyển cột thuộc tính Ram về 4 loại Ram: 8GB, 16GB, 32GB, 64GB

Chuyển cột thuộc tính Rom về 4 loại Ram: 256GB, 512GB, 1TB, 2TB

Xóa bỏ thông tin không cần thiết về card đồ họa, cpu, rom, ram giữ lại loại card, cpu được tích hợp ở trong laptop

4.2.2. Kết quả

4.2.2.1. Thế giới di dộng (Laptop)

Sau khi làm sạch, tập dữ liệu gồm 5 cột và 267 mẫu

Các thuộc tính:

+ Name: Tên sản phẩm
+ Rom: Dung lượng lưu trữ
+ Ram: Bộ nhớ đệm
+ Price: Giá của sản phẩm
+ Cpu: Bộ xử lí trung tâm

Dữ liệu sau khi làm sạch sử dụng để phân tích, khám phá dữ liệu

4.2.2.2. CellphoneS (Laptop)

Sau khi làm sạch, tập dữ liệu gồm 5 cột và 476 mẫu

Các thuộc tính:

+ Name: Tên sản phẩm
+ Ram: Bộ nhớ đệm
+ Price: Giá của sản phẩm
+ Card: Bộ xử lí hình ảnh
+ Link: đường link dẫn tới từng sản phẩm

Dữ liệu sau khi làm sạch sử dụng để phân tích, khám phá dữ liệu

4.2.2.3. Thê giới di động (Điện thoại)

Sau khi làm sạch, tập dữ liệu gồm 6 cột và 98 dữ liệu

Các thuộc tính:

+ Name: Tên sản phẩm
+ Ram: Bộ nhớ đệm
+ Rom: Dung lượng lưu trữ
+ OS: Hệ điều hành
+ Price: Giá của sản phẩm
+ Link: Đường dẫn của từng sản phẩm

4.2.2.4. CellphoneS (Điện thoại)

Sau khi làm sạch, tập dữ liệu gồm 7 cột và 205 dữ liệu

Các thuộc tính:

+ Name: Tên sản phẩm
+ Inches: Kích thước màn hình
+ Ram: Bộ nhớ đệm
+ Rom: Dung lượng lưu trữ
+ OS: Hệ điều hành
+ Price: Giá của sản phẩm
+ Link: Đường dẫn của từng sản phẩm

CHƯƠNG 5. EXPLORE DATA ANLYSIS

5.1. Laptop

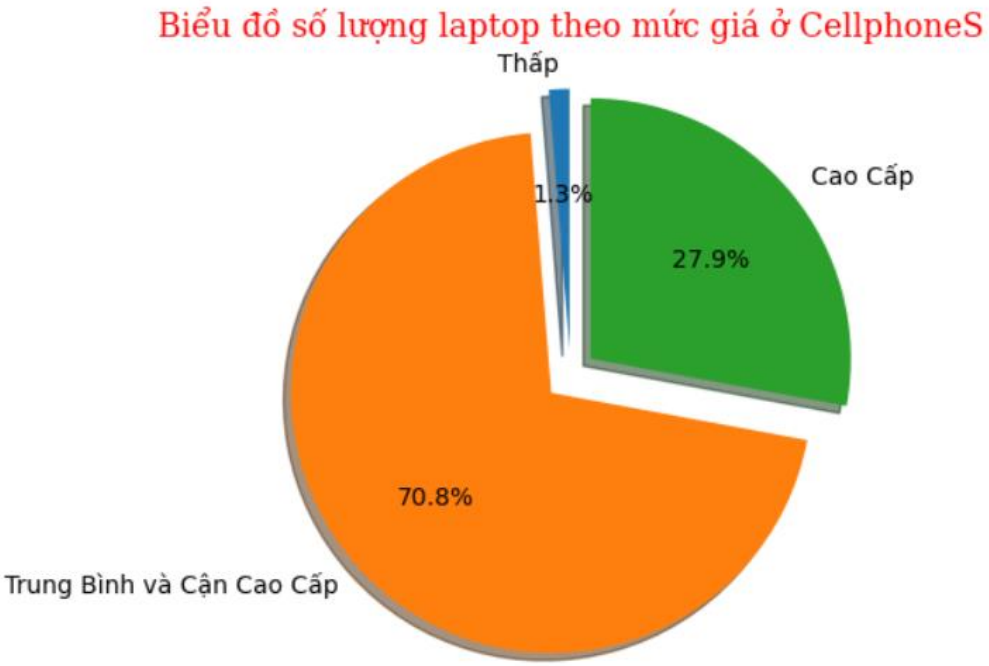
5.1.1. Giá cả sản phẩm

Phân chia dữ liệu dựa vào mức giá của Laptop làm 3 loại:

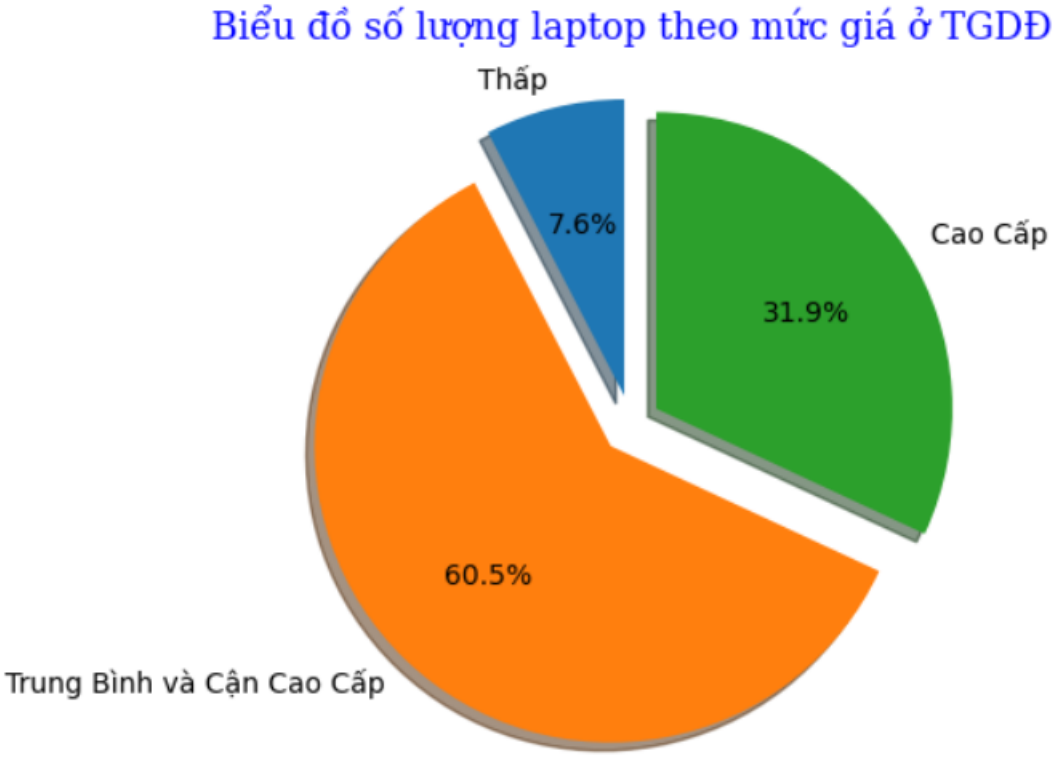
Laptop có mức giá thấp: dưới 10 triệu VNĐ. Trong phân khúc này, các mẫu laptop thường có cấu hình và tính năng tương đối cơ bản, chủ yếu chạy hệ điều hành Windows phiên bản cũ và hiệu suất xử lý thấp hơn so với các phân khúc khác vì chủ yếu laptop ở phân khúc này thường được sử dụng nhiều bởi các em học sinh, sinh viên chưa với mức chi tiêu hạn hẹp.

Laptop có mức giá trung bình và cận cao cấp: từ 10 đến 30 triệu VNĐ. Trong phân khúc này, các Laptop có tính năng và cấu hình tốt hơn, bao gồm màn hình đẹp, bộ vi xử lí cao, vượt trội về hiệu năng, hiệu suất xử lý mạnh mẽ hơn, và hỗ trợ nhiều tính năng tiên tiến hơn, đặc biệt trang bị card đồ họa rời. Điều này có thể hỗ trợ người sử dụng với nhiều mục đích với tác vụ nặng hay nhu cầu như thiết kế, render video, chạy máy ảo, lập trình nhúng,…

Laptop có mức giá cao cấp: từ 30 triệu VNĐ trở lên. Đây được coi là phân khúc cao cấp là dành cho những người dùng có ngân sách dồi dào và yêu cầu cao về tính năng, hiệu suất và thiết kế. Trong phân khúc này, các điện thoại thường có cấu hình và tính năng tốt nhất, bao gồm màn hình độ phủ màu cao có thể giúp các bạn học thiết kế có thể làm việc với hình ảnh mà không cần căn chỉnh màu, sự nhỏ gọn tiện lợi mà nó hướng đến như các dòng laptop cao cấp dành cho doanh nhân hay laptop chuyên về chơi hỗ trợ đầy đủ tính năng cho một game thủ hoặc streamer,….



Ảnh 5.1. Biểu đồ thể hiện số lượng sản phẩm theo mức giá ở CellphoneS



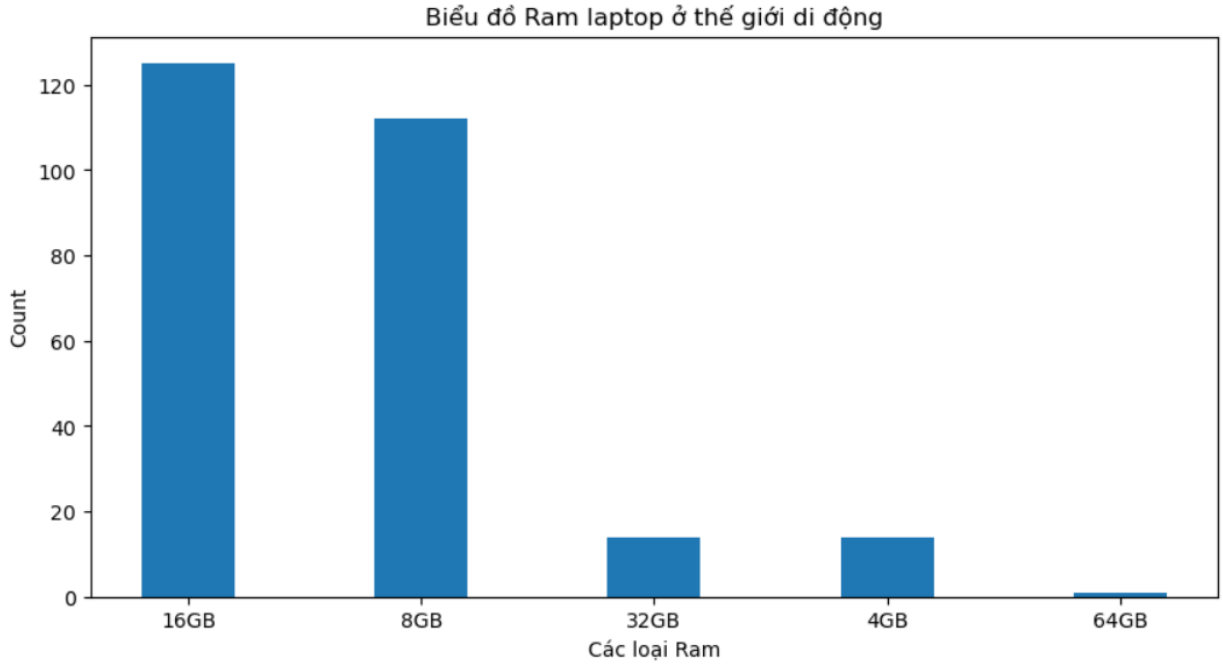
Ảnh 5.2. Biểu đồ thể hiện số lượng sản phẩm theo mức giá ở Thế giới di động

Ta có thể thấy qua biểu đồ thì hầu như phân khúc giá từ trung bình cho đến cao cấp chiếm phần lớn so với các sản phẩm khác. Phân khúc giá từ 10 Tr trở xuống có vẻ không được quan tâm nhiều trong thời gian gần đây. Có thể vì chất lượng chưa tương xứng và

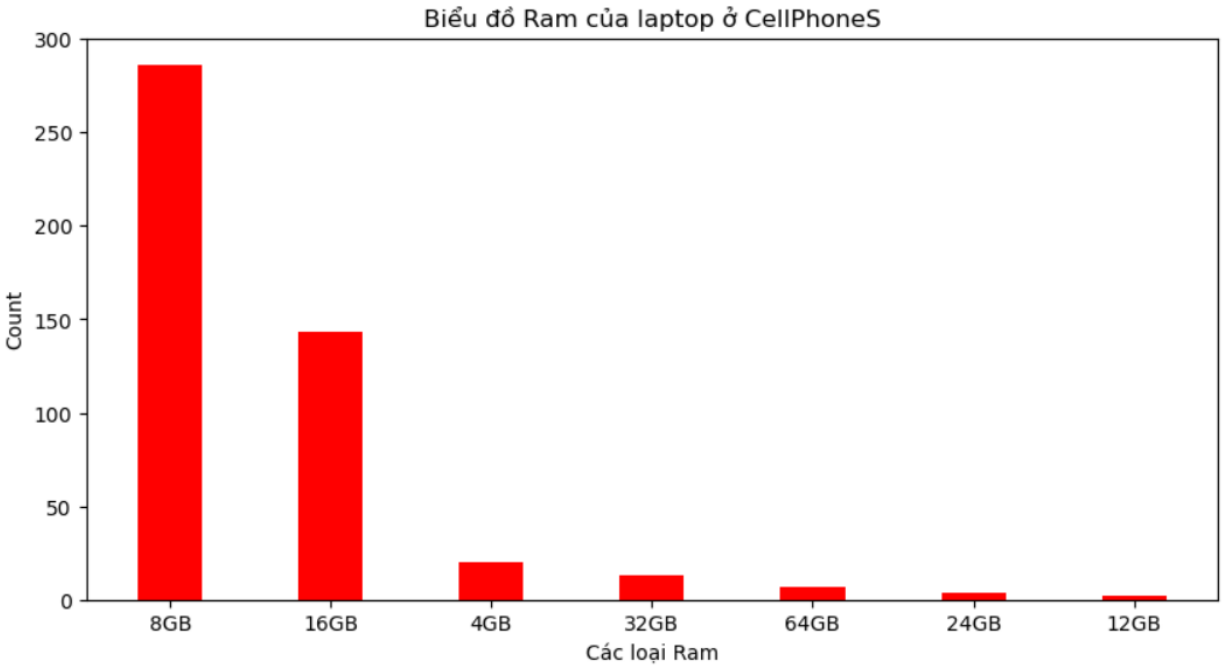
chưa đạt yêu cầu của người dung. Phân khúc trung bình và cao cấp được nhiều bán nhiều nhất và được người tiêu dung chọn mua nhiều vì hợp túi tiền mà hiệu quả đem lại còn vượt ngoài mong đợi.

Phân khúc giá cao cấp từ 30 Tr trở lên thì sẽ nhắm tới khách hang có điều kiện tài chính dư dả và đi lại nhiều nên cần một chiếc laptop mong nhẹ có thể dễ đem đi đem lại mà hiệu quả cho ra cũng vô cùng cao như Macbook, hay các dòng doanh nhân của Dell, …

5.1.2. Loại Ram được sử dụng



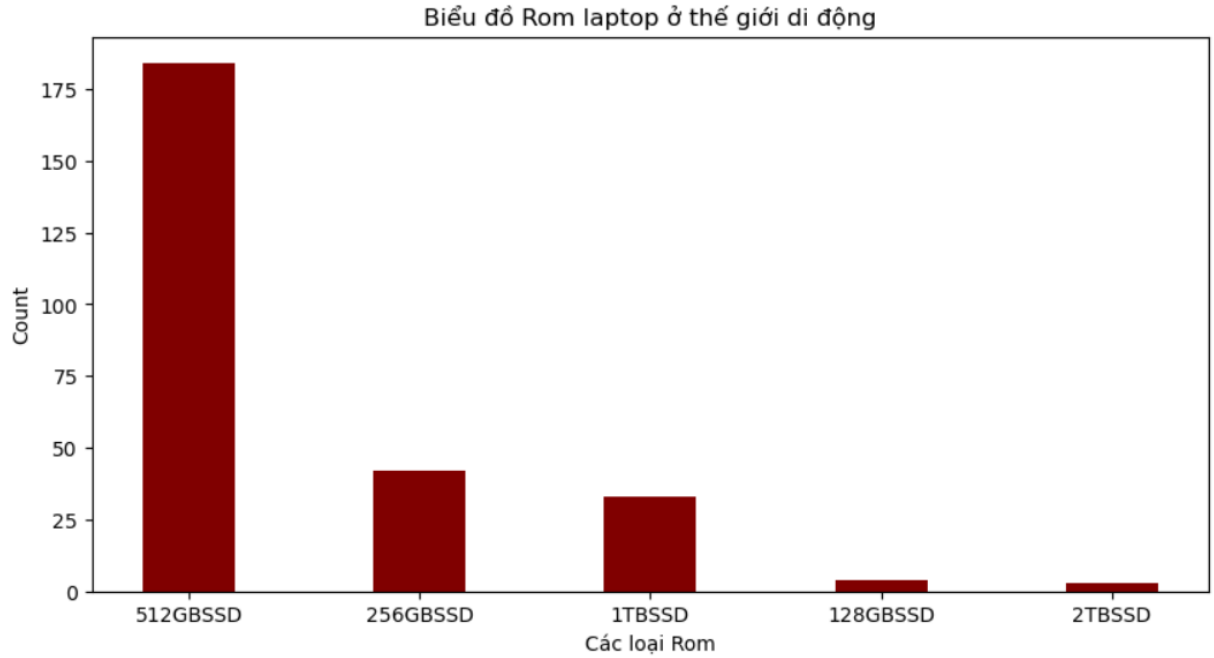
Ảnh 5.3. Biểu đồ sản phẩm được phân theo Ram ở thế giới di động



Ảnh 5.4. Biểu đồ phân loại sản phẩm theo Ram ở CellphoneS

Ta có thể thấy được Ram 8 GB và 16 GB được các hang sản xuất sử dụng tích hợp vào nhiều mẫu máy khác nhau. Hơn 50% sản phẩm đều được trang bị thanh Ram 8GB hoặc 16GB để có thể đáp ứng được với nhu cầu sử dụng cao của khách hang.

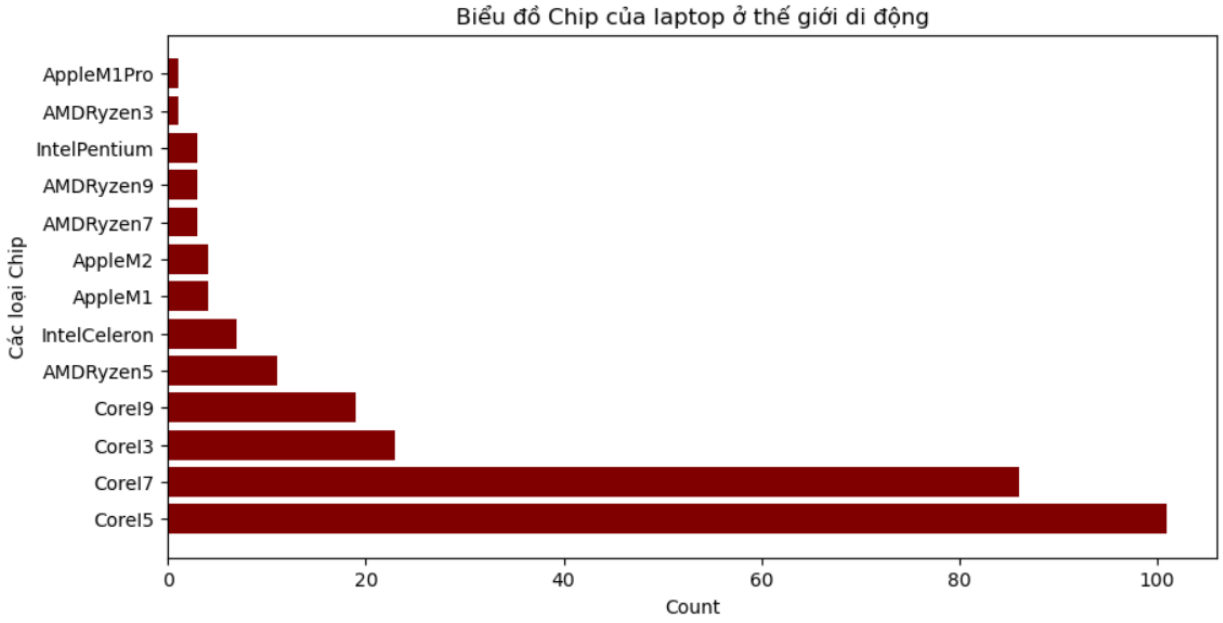
5.1.3. Loại Rom được sử dụng



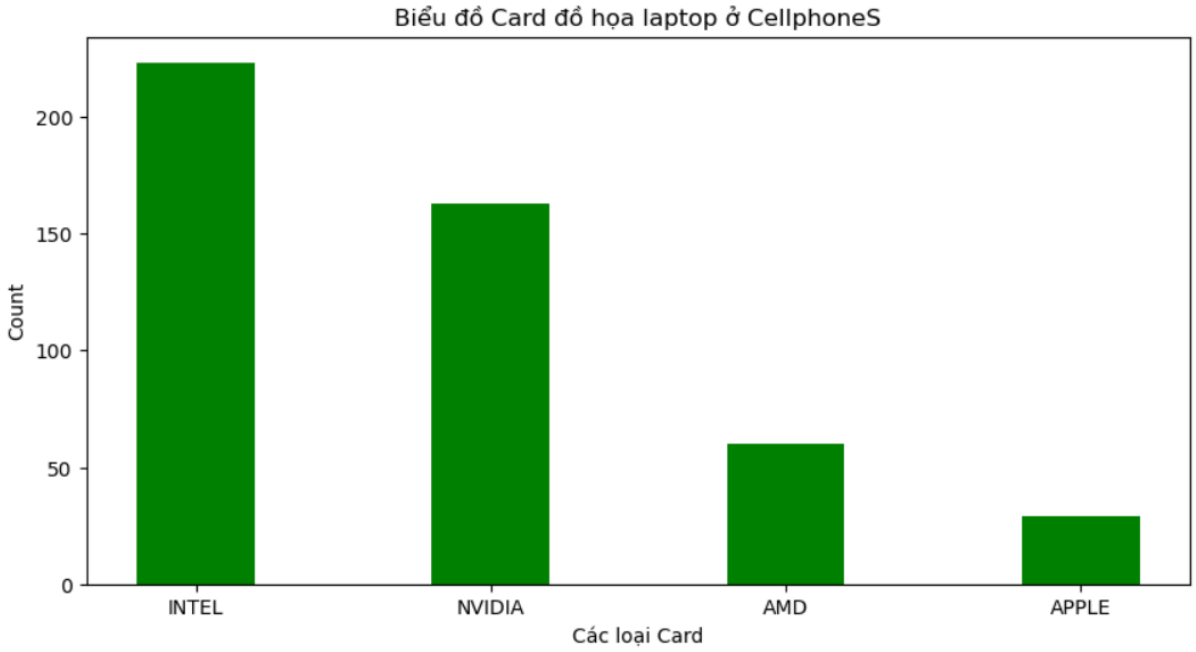
Ảnh 5.5. Biểu đồ thể hiện sản phẩm theo Rom ở Thế giới di động

Có thể thấy bộ nhớ của laptop thường được các nhà sản xuất sử dụng là 512GB và dều là loại SSD có thể đọc, ghi dữ liệu ở tốc độ cao. Điều này có thể đáp ứng được các như cầu từ cơ bản cho đến nâng cao của khách hang. Bên cạnh đó còn có các loại dung lượng bộ nhớ khác như 256GB chủ yếu dung trong laptop văn phòng hoặc Macbook, dung lượng 1TB cx được tích hợp ở trong laptop gaming, đồ họa đáp ứng nhu cầu của các creator, editor từ nghiệp dư tới chuyên nghiệp.

5.1.4. Chip và card đồ họa



Ảnh 5.6. Biểu đồ thể hiện sản phẩm theo cpu ở thế giới di động



Ảnh 5.7. Biểu đồ thể hiện sản phẩm theo Card đồ họa tích hợp ở CellphoneS

Thông qua biểu đồ về chipset thì ta có thể thấy Intel là nhà sản xuất chiếm lợi thế trong cạnh tranh sản laptop và loại cpu được dung nhiều nhất là Core I5 với nhiều biến thể khác nhau tiếp đó là Core I7 cũng có phần không kém cạnh so với các loại cpu khác. Cpu AMD tuy có tiếng trong lĩnh vực sản chip và card đồ họa nhưng có vẻ chưa thực sự cạnh tranh được so với ông lớn Intel.

Card đồ họa thì Intel chiếm phần lớn chủ yêu là card tích hợp dung cho các sản phẩm laptop cao cấp. Tiếp sau đó là Nvidia cũng không kém cạnh, được nhiều ông lớn trong ngành sản xuất laptop tin dung nhờ sự hiệu quả và tính ổn định của nó trong việc xử lí nhiều tác vụ nặng.

5.2. Điện thoại

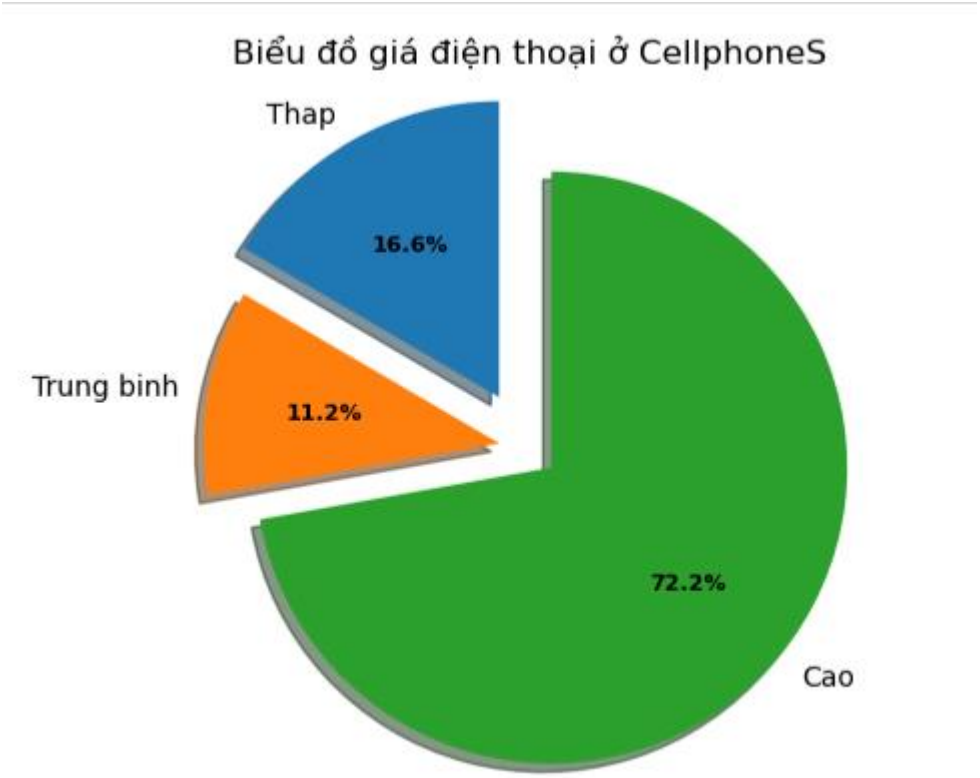
5.2.1. Phân khúc giá điện thoại

- Phân loại dữ liệu dựa vào mức giá của sản phẩm làm 3 loại:

- Sản phẩm ở mức giá thấp (dưới 10 triệu VND): Ở phân khúc này, các sản phẩm thường có cấu hình thấp và tính năng tương đối cơ bản, chạy hệ điều hành Android phiên bản cũ hoặc không chạy hệ điều hành nào, màn hình ở phân khúc này rất kém, RAM và ROM thấp nên hiệu suất xử lý kém hơn các sản phẩm ở phân khúc khác.

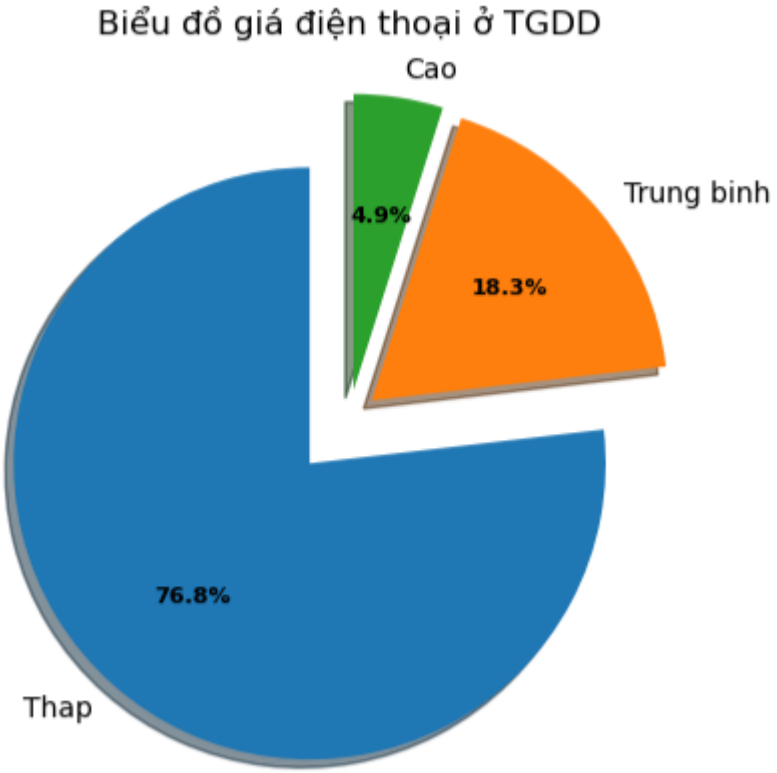
- Sản phẩm ở mức giá trung bình (từ 10 triệu VND – 25 triệu VND): Ở phân khúc này, sản phẩm có tính năng và cấu hình tốt hơn, bao gồm màn hình đẹp hơn, camera chất lượng cao hơn, hiệu suất xử lý mạnh mẽ hơn và hỗ trợ nhiều tính năng tiến tiến hơn.

- Sản phẩm ở mức giá cao cấp (trên 25 triệu VND): Ở phân khúc này, các sản phẩm sẽ thường được những người có tài chính cao nhắm tới vì yêu cầu về tính năng, hiệu suất và thiết kế có phần khắt khe hơn. Sản phẩm ở tầm giá này, sẽ có nhiều tính năng, cấu hình rất chất lượng, bao gồm màn hình ở mức tốt nhất, camera chất lượng cao, khả năng chống nước và bụi, hỗ trợ sạc nhanh, sạc không dây và nhiều tính năng tiên tiến khác.



Ảnh 5.8 Biểu đồ thể hiện giá sản phẩm điện thoại ở CellphoneS

Có thể thấy, ở CellphoneS, thị phần thấp và trung bình được nhắm tới ít hơn chủ yếu là tập trung vào thị phần cao cấp. Tỷ lệ sản phẩm ở mức giá trung bình rất thấp chỉ chiếm 11,1%, sản phẩm ở mức giá thấp chiếm 16,6% và thị phần cao cấp chiếm tỷ lệ nhiều nhất 72,2%.

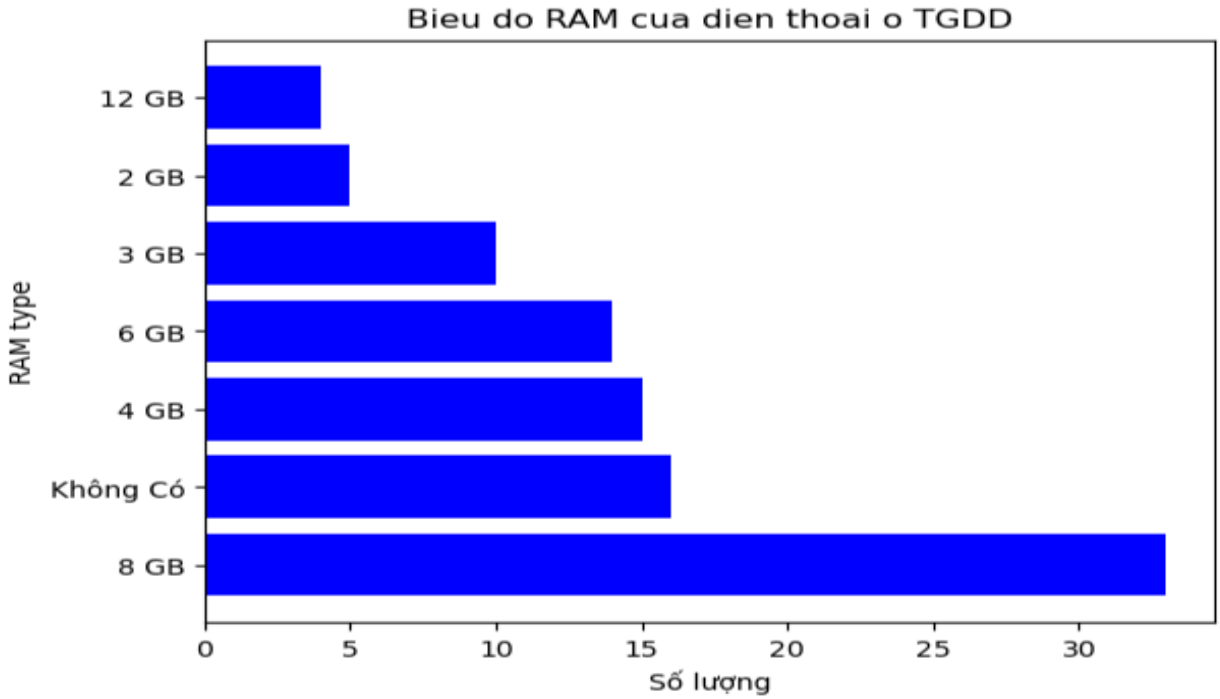


Ảnh 5.9 Biểu đồ thể hiện giá sản phẩm điện thoại ở TGDD

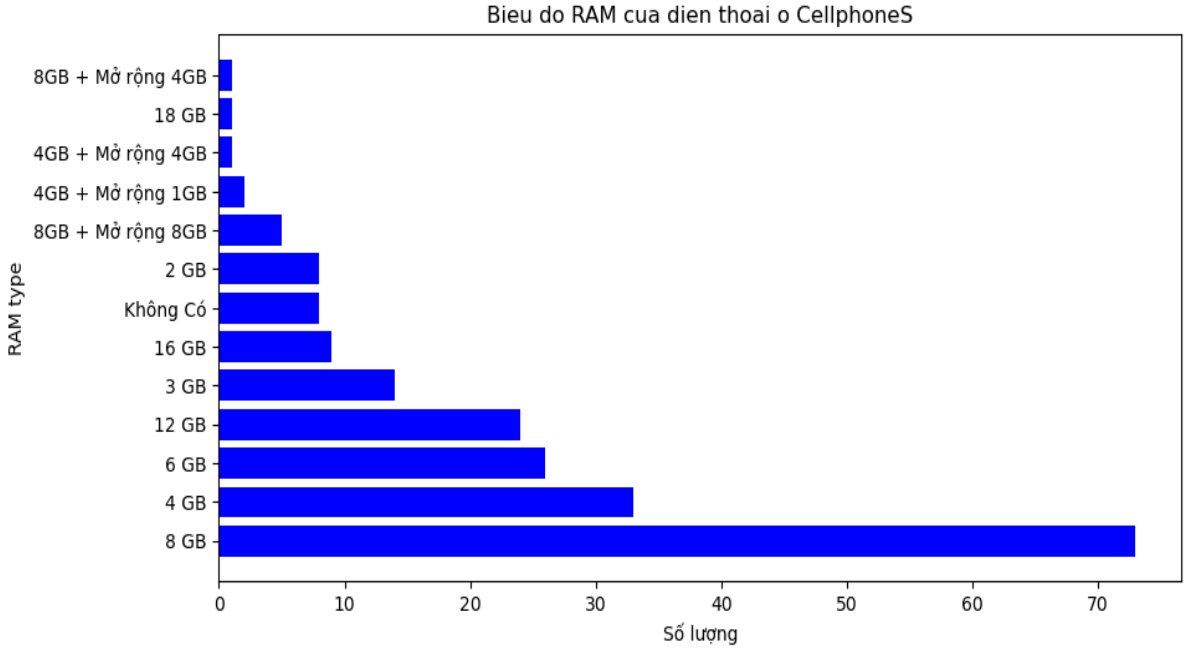
Ở trong web TGDD, tỷ lệ sản phẩm ở mức giá thấp chiếm đại đa số sản phẩm (76,8%) và số lượng sản phẩm ở mức giá trung bình chiếm 18,3% và số lượng sản phẩm cao cấp chiếm phần nhỏ sản phẩm có trong cửa hàng chỉ khoảng 4,9%

5.2.2. RAM

Đơn vị của RAM là GB



Ảnh 5.10 Biểu đồ thể hiện số lượng RAM có trong sản phẩm ở TGDD

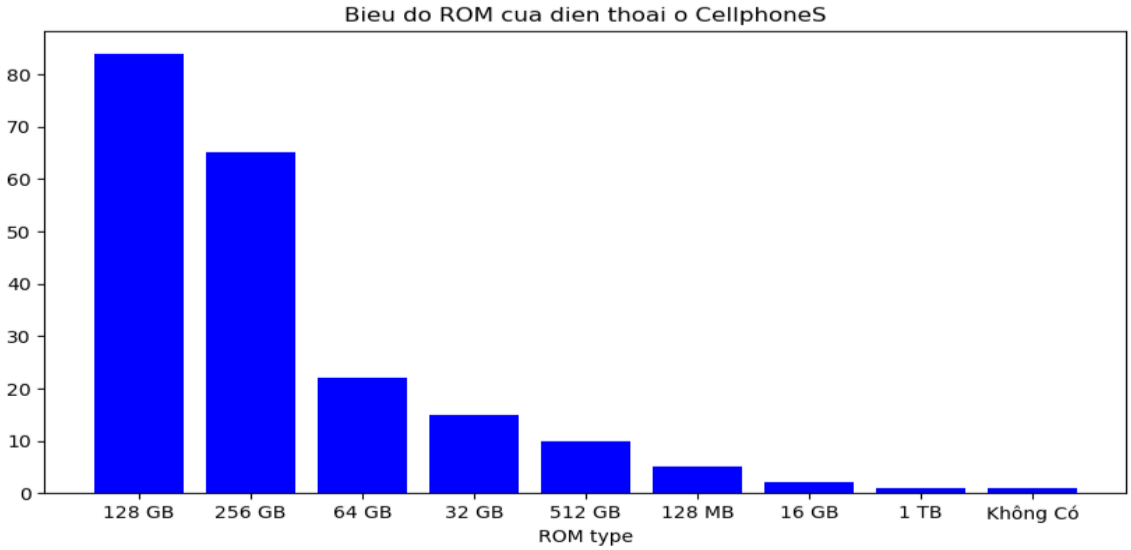


Ảnh 5.11 Biểu đồ thể hiện số lượng RAM có trong sản phẩm ở CellphoneS

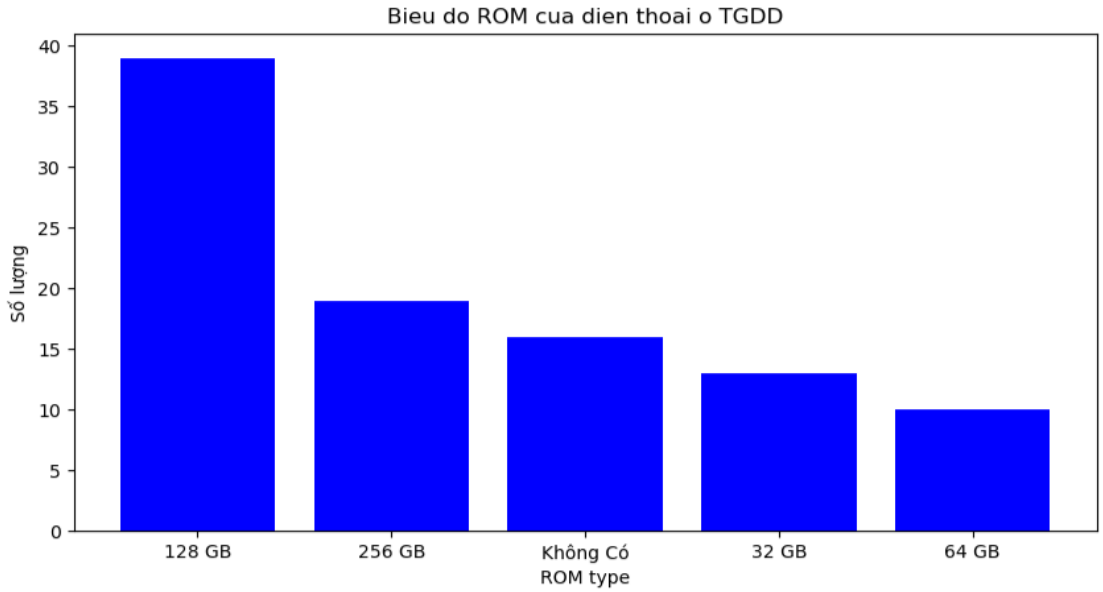
Có thể thấy, các sản phẩm gần đây đều chú trọng vào các việc thực hiện nhiều tác vụ cũng như chơi game giải trí nên số lượng RAM cũng tăng lên từ thấp nhất là (những chiếc điện thoại Nokia được sản xuất từ rất lâu) nên không có RAM cho đến 18GB RAM . Nhưng cơ bản và có số lượng nhiều nhất vẫn là 8GB RAM. Một vài sản phẩm có thể mở rộng thêm số lượng RAM. Tuy nhiên, số lượng RAM cao cũng khiến giá thành cao hơn

5.2.3. ROM

Đơn vị của ROM là GB



Ảnh 5.12 Biểu đồ thể hiện số dung lượng của sản phẩm tại CellphoneS

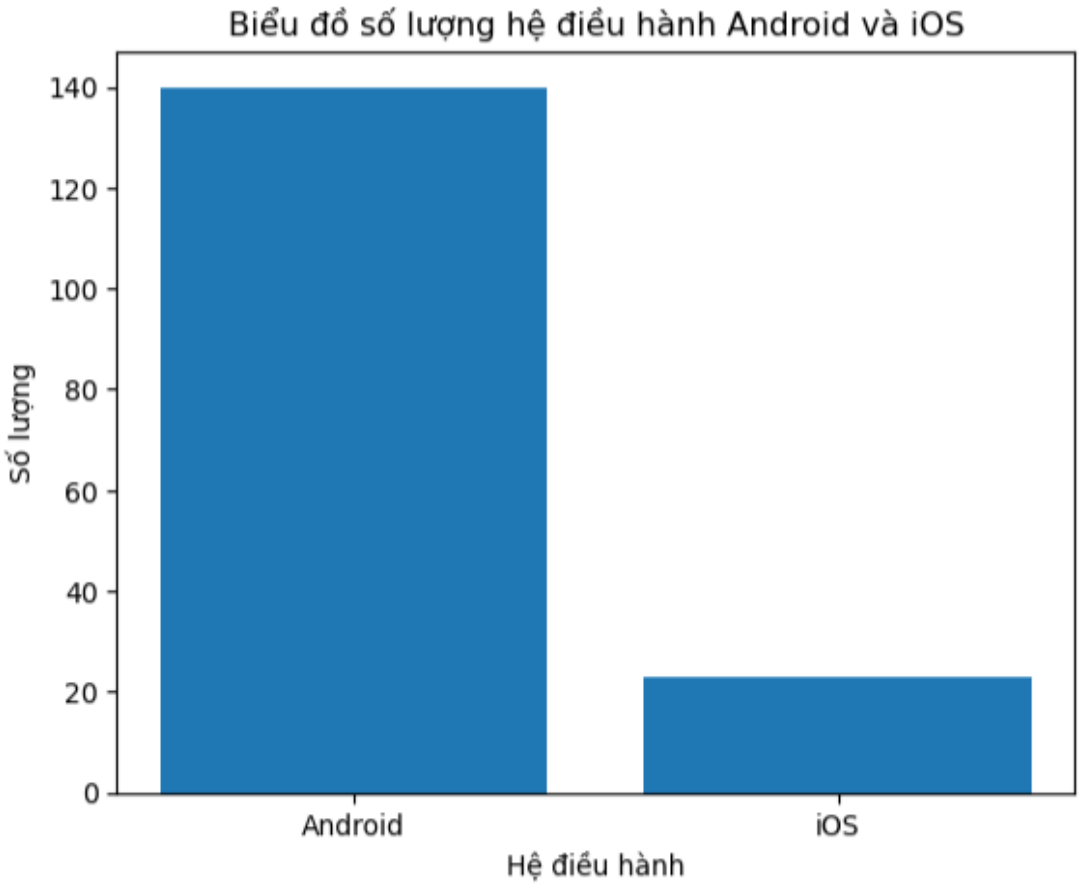


Ảnh 5.13 Biểu đồ thể hiện số dung lượng có trong sản phẩm ở TGDD

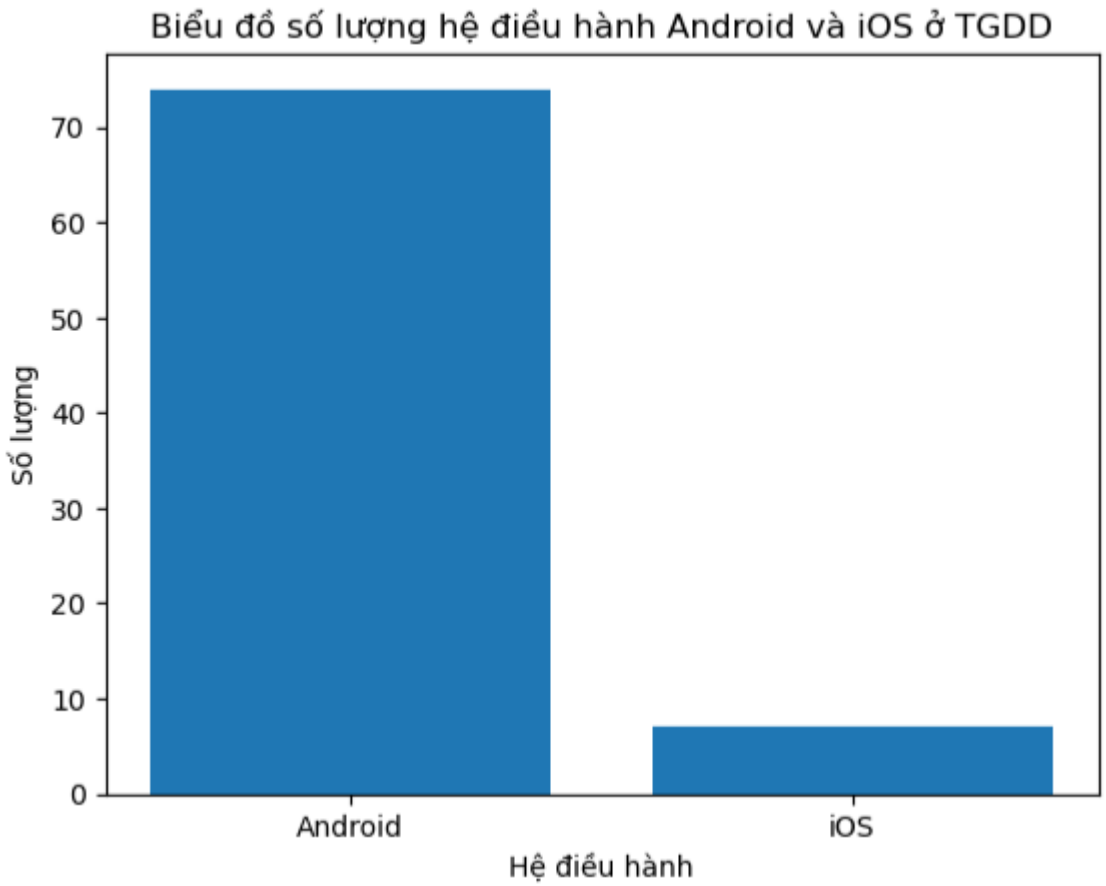
ROM hay còn được gọi là bộ nhớ trong, nơi lưu trữ dữ liệu của thiết bị điện thoại. ROM đa có dung lượng lưu trữ trong khoảng từ 16GB cho đến 256GB, có một vài thiết bị do quá lâu đời nên chỉ có 128MB dung lượng lưu trữ. Bên cạnh đó, cũng có nhưng sản phẩm mới ra mắt được trang bị khả năng lưu trữ lên tới 512GB và 1TB. Việc trang bị số lượng ROM lớn sẽ làm tăng giá thành, chi phí sản xuất và chip được tích hợp.

5.2.4. Hệ điều hành (OS)

Ở đây ta có 2 hệ điều hành chính, đó là Android và IOS.



Ảnh 5.14 Biểu đồ số lượng hệ điều hành Android và iOS ở CellphoneS

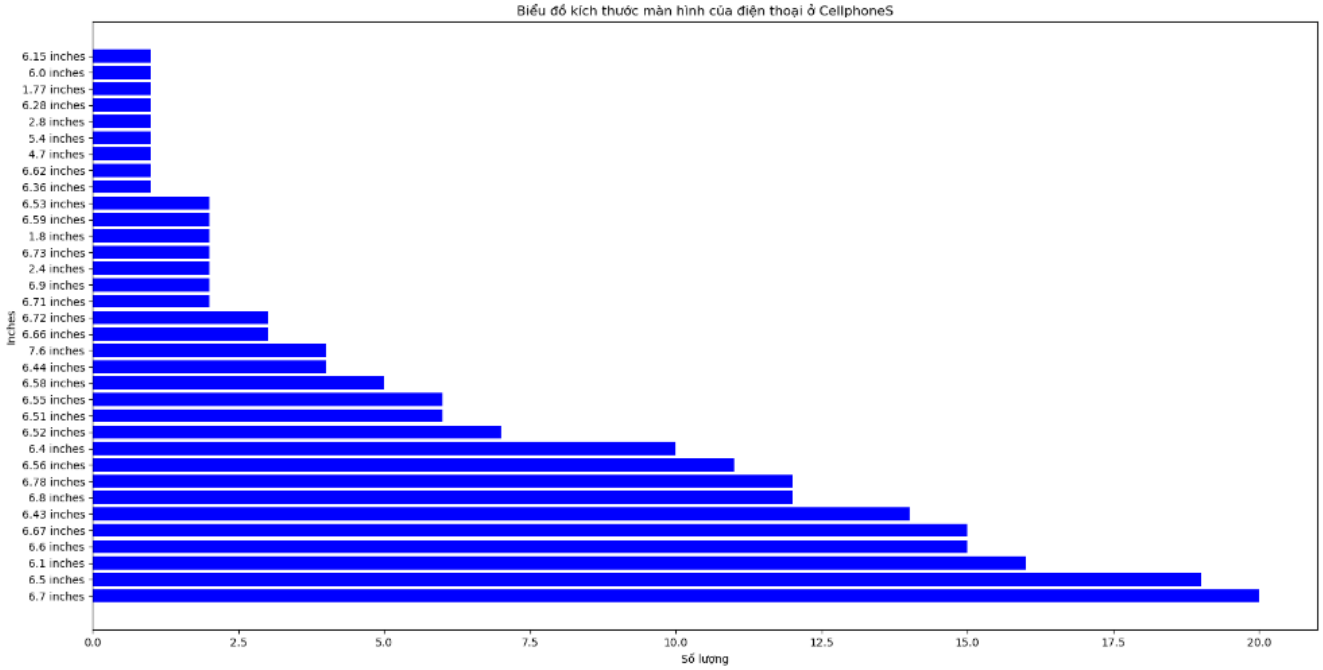


Ảnh 5.15 Biểu đồ số lượng hệ điều hành Android và iOS ở TGDD

Số lượng sản phẩm sử dụng Android nhiều gấp nhiều lần sản phẩm iOS, là vì iOS là hệ điều hành độc quyền của Apple nên rất khó có sản phẩm nào khác ngoài Apple được cấp quyền sử dụng hệ điều hành này.

Trái lại với anh bạn Apple "ích kỷ", hệ điều hành Android lại rất được các hãng sản xuất điện thoại sử dụng vì 1 phần Android có mã nguồn mở được chia sẻ bởi Google nên có rất nhiều hãng lấy về và biến tấu thêm 1 chút khác biệt để đưa vào sản xuất và sử dụng. Sự đa dạng thiết bị, Android có thể được sử dụng trên nhiều thiết bị khác nhau như: tablet, smartwatch và các thiết bị thông minh khác.

5.2.5. Kích thước màn hình



Ảnh 5.16 Biểu đồ số lượng kích thước màn hình ở CellphoneS

CHƯƠNG 6. TỔNG KẾT

6.1. Phân chia công việc

- Code thu thập dữ liệu – Lâm, Linh
- Cào dữ liệu:
  + Laptop (CellphoneS và TGDD) – Lâm, Linh
  + Điện thoại (CellphoneS và TGDD) – Lâm, Linh
- Tìm các vấn đề và làm sạch dữ liệu –Lâm, Linh
- Phân tích dữ liệu:
  + Điện thoại – Lâm
  + Laptop – Linh

6.2. Khó khăn trong quá trình thực hiện

6.2.1. Quá trình thu thập dữ liệu

- Khó khăn là 2 trang web khác nhau nên cách lấy dữ liệu từ 2 trang web cũng khác nhau. Có những dữ liệu mà trang web này có nhưng trang web kia không nên rất khó để lấy ra dữ liệu đồng bộ cho cả 2 bên.
- Thu thập dữ liệu mất nhiều thời gian khi phải cào 2 trang web để lấy ra cùng 1 loại dữ liệu, việc nhầm lẫn 2 dữ liệu với nhau rất hay xảy ra. Cũng như lúc thu thập dữ liệu về thông tin sản phẩm (cấu hình) hay bị tình trạng dữ liệu bị lẫn. Ví dụ khi em lấy thông tin về RAM thì kết quả trả về lại là OS.
- Khi thu thập link sản phẩm của TGDD, dữ liệu thu về có nhiều link lạ, link lỗi chứa sản phẩm, mất rất nhiều thời gian để loại bỏ nhưng link lỗi đấy.

6.2.2. Xử lý dữ liệu

- Mặc dù 2 trang web có rất nhiều sản phẩm để thu thập, tuy nhiên dữ liệu lại chứa khá nhiều dữ liệu rỗng, để fill thì lại rất nhiều thời gian tìm kiếm nên đành phải xóa bỏ.