ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

NHẬP MÔN KHOA HỌC DỮ LIỆU - CQ2022/21

ĐỀ TÀI

Phân Tích Thị Trường Điện Thoại Và Dự Đoán Giá Bán Tại Mobile City

Giảng viên: Lê Ngọc Thành
Lê Nhựt Nam

Sinh viên:		
	Trương Tiến Anh	22120017
	Lê Nguyễn Gia Bảo	22120023
	Nguyễn Hữu Bền	22120029
	Đoàn Minh Cường	22120043
	Bùi Đình Bảo	21120201

Chuyên ngành Khoa học máy tính/Khoa học dữ liệu

Ho Chi Minh City, 12/2024

Mục lục

I Giới thiệu nhóm và tiến độ công việc **4**

II Báo cáo bài tập **4**

1 Giới thiệu đề tài đồ án **4**

2 Data Collecting (Thu thập dữ liệu) **5**

3 Data Pre-processing (Tiền xử lý dữ liệu) **6**

4 Data Exploring (Khám phá dữ liệu) **8**
4.1 Khái quát . . . 8
4.2 Đặt ra các câu hỏi . . . 9
4.2.1 Có bao nhiêu dòng và cột trong bộ dữ liệu? . . . 9
4.2.2 Mỗi dòng có ý nghĩa gì? Có quan trọng không khi mỗi dòng có ý nghĩa khác nhau? 9
4.2.3 Có dòng dữ liệu nào bị trùng lặp không? . . . 9
4.2.4 Các cột (thuộc tính) có ý nghĩa như thế nào? . . . 9
4.2.5 Mỗi cột hiện tại có kiểu dữ liệu gì? Có cột nào có kiểu dữ liệu không phù hợp cho việc xử lý tiếp theo không? . . . 11
4.2.6 Tỷ lệ giá trị bị thiếu là bao nhiêu? . . . 11
4.2.7 Đối với kiểu dữ liệu dạng numeric, dữ liệu được phân bố như thế nào? . . . 12
4.2.8 Đối với kiểu dữ liệu dạng **category**, dữ liệu được phân bố như thế nào? . . . 13
4.2.9 Thời gian bảo hành? . . . 13
4.2.10 Màu sắc? . . . 14
4.2.11 Hệ điều hành? . . . 14
4.2.12 Có bao nhiêu điện thoại cũ trong bộ dữ liệu? . . . 15
4.2.13 CPU? . . . 15
4.2.14 RAM? . . . 16
4.2.15 Dung lượng Pin? . . . 16
4.2.16 Công suất sạc? . . . 17
4.2.17 Hãng điện thoại? . . . 17
4.2.18 Loại màn hình? . . . 18
4.2.19 Độ phân giải cam trước? . . . 19
4.2.20 Độ phân giải cam sau? . . . 19
4.2.21 Phân tích tương quan giữa: Giá, Dung lượng pin, RAM, Bộ nhớ trong, Kích thước màn hình? . . . 20
4.2.22 Mối quan hệ giữa Giá và các thuộc tính: RAM, Bộ nhớ trong . . . 21

5 Defining the Problem (Xác định các vấn đề) **21**
5.1 Mức giá giảm có ảnh hưởng đến lượt đánh giá và mức độ hài lòng của khách hàng không? 21
5.1.1 Mục đích . . . 21
5.1.2 Tiền xử lý . . . 22
5.1.3 Trực quan hóa và nhận xét . . . 22
5.2 Thời gian bảo hành có ảnh hưởng đến sự hài lòng của khách hàng không? . . . 24
5.2.1 Mục đích . . . 24
5.2.2 Tiền xử lý . . . 24
5.2.3 Trực quan hóa và nhận xét . . . 24
5.3 Dòng điện thoại nào được quan tâm nhiều nhất từ trước đến nay dựa trên số lượt đánh giá và hỏi đáp? (Top 10) . . . 25
5.3.1 Mục đích . . . 25
5.3.2 Tiền xử lý . . . 25
5.3.3 Trực quan hóa và nhận xét . . . 26

5.4 Dòng điện thoại nào được quan tâm nhiều nhất từ trước đến nay dựa trên số lượt đánh giá và hỏi đáp? (Top 10) . . . 26
5.4.1 Mục đích . . . 26
5.4.2 Tiền xử lý . . . 26
5.4.3 Trực quan hóa và nhận xét . . . 27
5.5 Trong các thông số kỹ thuật của một chiếc điện thoại, những thông số nào thường có ảnh hưởng lớn nhất đến giá bán? . . . 27
5.5.1 Mục đích . . . 27
5.5.2 Tiền xử lý . . . 28
5.5.3 Trực quan hóa và nhận xét . . . 28
5.6 Kiểu thiết kế điện thoại nào phổ biến nhất hiện nay, dựa trên các kiểu thiết kế của các mẫu điện thoại hiện có trong cửa hàng? . . . 29
5.6.1 Mục đích . . . 29
5.6.2 Tiền xử lý . . . 30
5.6.3 Trực quan hóa và nhận xét . . . 30
5.7 Hãng điện thoại có ảnh hưởng đến giá cả không? . . . 31
5.7.1 Mục đích . . . 31
5.7.2 Tiền xử lý . . . 31
5.7.3 Trực quan hóa và nhận xét . . . 31

6 Data Modeling (Xây dựng mô hình) 32
6.1 Tiền xử lý dữ liệu cho các mô hình . . . 32
6.2 Tạo các mô hình . . . 33
6.2.1 Model: XGBRegressor, DecisionTreeRegressor, RandomForestRegressor . . . 33
6.2.2 Đánh giá và so sánh các mô hình: GBRegressor, DecisionTreeRegressor, RandomForestRegressor . . . 37

III Tổng kết 40

Lời mở đầu

Nhóm chúng em xin gửi lời cảm ơn chân thành đến Trường Đại học Khoa học Tự nhiên TP.HCM vì đã đưa môn **Nhập Môn Khoa Học Dữ Liệu** vào chương trình học. Đặc biệt, chúng em trân trọng cảm ơn **Thầy Lê Ngọc Thành** và **Thầy Lê Nhựt Nam**, những giảng viên tận tâm đã hướng dẫn và đồng hành cùng chúng em trong suốt học phần này. Dù chỉ tiếp cận những kiến thức cơ bản của Khoa học Dữ liệu, môn học đã mang đến cho chúng em nhiều kỹ năng hữu ích, đặc biệt về khả năng phân tích và suy luận.

Với sự cố gắng và tinh thần trách nhiệm, nhóm chúng em đã hoàn thiện đồ án kết thúc học phần. Tuy nhiên, trong quá trình thực hiện đồ án khó tránh khỏi thiếu sót. Chúng em mong nhận được sự góp ý quý báu từ thầy để cải thiện trong tương lai.

Trong bối cảnh chuyển đổi số đang trở thành xu hướng tất yếu trong các lĩnh vực kinh tế, văn hóa, và xã hội, **Khoa học Dữ liệu** đóng vai trò quan trọng trong việc thu thập, phân tích và dự báo thông tin, hỗ trợ các nhà quản trị đưa ra quyết định chính xác. Đồ án của nhóm tập trung phân tích bộ dữ liệu **Mobile** để cung cấp góc nhìn tổng quan hơn về thị trường điện tử hiện nay.

Nội dung báo cáo bao gồm: giới thiệu đề tài, mục tiêu nghiên cứu, các quy trình xử lý dữ liệu, ứng dụng thực tế vào bài toán, đánh giá kết quả mô hình, và phần đưa ra những quan điểm của nhóm khi hoàn thành đồ án.

Phần I
Giới thiệu nhóm và tiến độ công việc

Bảng 1: *Thông tin các thành viên trong nhóm*

STT	Họ Tên	MSSV	Email
1	Trương Tiến Anh	22120017	22120017@student.hcmus.edu.vn
2	Lê Nguyễn Gia Bảo	22120023	22120023@student.hcmus.edu.vn
3	Nguyễn Hữu Bền	22120029	22120029@student.hcmus.edu.vn
4	Đoàn Minh Cường	22120043	22120043@student.hcmus.edu.vn
5	Bùi Đình Bảo	21120201	21120201@student.hcmus.edu.vn

Bảng 2: *Công việc nhóm*

STT	Họ Tên	Công việc được giao	Tiến độ	Vấn đề
1	Trương Tiến Anh	Phân tích 2 vấn đề, xây dựng mô hình, viết báo cáo.	Hoàn thành	Không có
2	Lê Nguyễn Gia Bảo	Phân tích 2 vấn đề, xây dựng và đánh giá mô hình.	Hoàn thành	Không có
3	Nguyễn Hữu Bền	Quản lí nhóm, khám phá, xử lí dữ liệu, viết báo cáo.	Hoàn thành	Không có
4	Đoàn Minh Cường	Phân tích 2 vấn đề, xây dựng mô hình, làm thuyết trình.	Hoàn thành	Không có
5	Bùi Đình Bảo	Tạo template, thu thập, tiền xử lí dữ liệu, hỗ trợ nhóm.	Hoàn thành	Không có

Phần II
Báo cáo bài tập

1 Giới thiệu đề tài đồ án

Lý do chọn đề tài:

Trong bối cảnh công nghệ ngày càng phát triển, nhu cầu tìm hiểu thông tin về các sản phẩm điện thoại di động trở nên rất phổ biến. Đặc biệt, với sự xuất hiện của các trang web thương mại điện tử như MobileCity.vn, người dùng có thể dễ dàng tìm kiếm và so sánh thông tin về các dòng điện thoại, giá cả, và các đặc điểm kỹ thuật.

MobileCity.vn là một trong những trang web nổi bật tại Việt Nam chuyên cung cấp thông tin về các sản phẩm điện thoại di động, bao gồm các thương hiệu nổi tiếng như Apple, Samsung, Oppo, Xiaomi, và nhiều hãng khác. Trang web này cung cấp một lượng lớn thông tin về các mẫu điện thoại, từ thông số kỹ thuật, giá cả, khuyến mãi cho đến đánh giá của người tiêu dùng.

Tuy nhiên, việc thu thập và xử lý thông tin từ trang web này một cách tự động có thể gặp khó khăn do số lượng sản phẩm rất lớn và thay đổi liên tục. Vì vậy, việc thu thập thông tin điện thoại từ trang web MobileCity.vn không chỉ mang lại giá trị thực tiễn trong việc hỗ trợ người tiêu dùng trong việc lựa chọn sản phẩm phù hợp mà còn giúp nghiên cứu các xu hướng thị trường điện thoại di động hiện nay.

Đề tài này nhằm mục đích xây dựng một hệ thống thu thập thông tin từ trang web MobileCity.vn, giúp người dùng dễ dàng tiếp cận dữ liệu cập nhật về các sản phẩm điện thoại. Hệ thống này có thể được áp dụng trong các nghiên cứu thị trường, phân tích xu hướng tiêu dùng, hoặc phục vụ cho các mục đích cá nhân như so sánh giá và tìm kiếm các ưu đãi tốt nhất.

Mục tiêu nghiên cứu:

Nghiên cứu này dựa trên việc thu thập và phân tích các thông tin điện thoại từ trang web MobileCity.vn, một trong những trang thương mại điện tử hàng đầu tại Việt Nam chuyên cung cấp các sản phẩm điện thoại di động. Mục tiêu của nghiên cứu là xây dựng một hệ thống thu thập thông tin tự động về các dòng điện thoại, bao gồm các đặc điểm như tên sản phẩm, giá cả, hãng sản xuất, cấu hình kỹ thuật và các khuyến mãi hiện có. Thông qua đó, nghiên cứu này hy vọng sẽ cung cấp một giải pháp hữu ích cho người tiêu dùng trong việc lựa chọn và so sánh các sản phẩm điện thoại phù hợp.

Bộ dữ liệu thu thập được từ MobileCity.vn sẽ giúp phân tích các xu hướng tiêu dùng và thị trường điện thoại, đồng thời cung cấp thông tin giá trị cho các nghiên cứu thị trường và đánh giá các chiến lược marketing của các thương hiệu điện thoại. Các biến độc lập trong nghiên cứu bao gồm các yếu tố như giá, thương hiệu, cấu hình và đánh giá của người dùng. Mô hình thu thập và phân tích này không chỉ giúp người tiêu dùng dễ dàng tiếp cận thông tin về các sản phẩm điện thoại mà còn có thể hỗ trợ các cửa hàng và nhà sản xuất điện thoại trong việc hiểu rõ hơn về nhu cầu của người dùng và tối ưu hóa chiến lược kinh doanh của mình.

Thông qua việc xây dựng và triển khai hệ thống thu thập dữ liệu này, chúng em hy vọng có thể cung cấp một công cụ hữu ích cho những người quan tâm đến việc lựa chọn điện thoại, từ đó cải thiện quá trình ra quyết định mua sắm và phát triển thị trường điện thoại di động.

2 Data Collecting (Thu thập dữ liệu)

Nguồn dữ liệu
Bộ dữ liệu các dòng điện thoại được chúng em cào từ trang web MobileCity.vn.
Dữ liệu gồm 8855 dòng (đối tượng) và 26 cột thuộc tính như sau:

#	Thuộc tính	Kiểu dữ liệu	Mô tả thuộc tính
0	Tên	object	Tên của điện thoại
1	Loại điện thoại	object	Loại của điện thoại (smartphone, feature phone, v.v.)
2	Thời gian bảo hành	float64	Thời gian bảo hành của điện thoại (tính bằng năm)
3	Đánh giá	float64	Điểm đánh giá trung bình của điện thoại từ người dùng
4	Số lượng bình luận	int64	Số lượng bình luận người dùng đã để lại cho điện thoại
5	Đường dẫn	object	Đường dẫn đến trang chi tiết của điện thoại trên website
6	Màu sắc	object	Màu sắc của điện thoại
7	Giá mới	float64	Giá của điện thoại khi mới phát hành
8	Giá cũ	float64	Giá của điện thoại khi đã qua sử dụng
9	Hệ điều hành	object	Hệ điều hành được cài đặt trên điện thoại (Android, iOS, v.v.)
10	CPU	object	Loại vi xử lý (CPU) của điện thoại
11	RAM	float64	Dung lượng RAM của điện thoại (tính bằng GB)
12	Bộ nhớ trong	float64	Dung lượng bộ nhớ trong của điện thoại (tính bằng GB)
13	Dung lượng pin	float64	Dung lượng pin của điện thoại (tính bằng mAh)
14	Thiết kế	object	Thiết kế của điện thoại (thân kim loại, nhựa, v.v.)
15	Hãng điện thoại	object	Hãng sản xuất điện thoại (Samsung, Apple, v.v.)
16	Là điện thoại cũ	bool	Chỉ ra nếu điện thoại đã qua sử dụng hay không
17	Kích thước màn hình	float64	Kích thước màn hình của điện thoại (tính bằng inch)
18	Tần số quét	float64	Tần số quét màn hình của điện thoại (tính bằng Hz)
19	Độ sáng màn hình	float64	Độ sáng tối đa của màn hình điện thoại (tính bằng nits)
20	Loại màn hình	object	Loại màn hình (AMOLED, LCD, v.v.)
21	Số thẻ SIM	int64	Số lượng thẻ SIM mà điện thoại hỗ trợ
22	Loại pin	object	Loại pin của điện thoại (Li-ion, Li-Po, v.v.)
23	Công suất sạc	float64	Công suất sạc của điện thoại (tính bằng watt)
24	Độ phân giải cam sau	object	Độ phân giải của camera sau (tính bằng megapixel)

#	Thuộc tính	Kiểu dữ liệu	Mô tả thuộc tính
25	Độ phân giải cam trước	object	Độ phân giải của camera trước (tính bằng megapixel)

3 Data Pre-processing (Tiền xử lý dữ liệu)

Tiền xử lý dữ liệu là một giai đoạn quan trọng để làm sạch, chuẩn hóa và chuẩn bị dữ liệu cho quá trình phân tích và xây dựng mô hình. Hãy xem qua những dòng đầu tiên của tập dữ liệu để nắm được cấu trúc và kiểu dữ liệu.

	ten	loai_dien_thoai	thoi_gian_bao_hanh	danh_gia	so_luong_binh_luan	duong_dan	mau_sac	gia_moi	gia_cu	he_dieu_hanh	cpu	ram	bo_nho_trong	dung_luong_pin	thiet_ke	hang_dien_thoai
0	Điện thoại Xiaomi Redmi 12C (Helio G85)	Redmi	12.0	5.0	7	https://mobilecity.vn/dien-thoai/xiaomi-redmi-...	Xanh Đậm	1650000.0	2950000.0	Android 12	Mediatek MT6769Z Helio G85 (12nm)\n8 nhân (2x2...	4.0	64.0	5000.0	Thanh + Cảm ứng	Xiaomi
1	Điện thoại Xiaomi Redmi 12C (Helio G85)	Redmi	12.0	5.0	7	https://mobilecity.vn/dien-thoai/xiaomi-redmi-...	Xanh Đậm	1950000.0	2950000.0	Android 12	Mediatek MT6769Z Helio G85 (12nm)\n8 nhân (2x2...	4.0	128.0	5000.0	Thanh + Cảm ứng	Xiaomi
2	Điện thoại Xiaomi Redmi 12C (Helio G85)	Redmi	12.0	5.0	7	https://mobilecity.vn/dien-thoai/xiaomi-redmi-...	Xanh bạc hà	1650000.0	2950000.0	Android 12	Mediatek MT6769Z Helio G85 (12nm)\n8 nhân (2x2...	4.0	64.0	5000.0	Thanh + Cảm ứng	Xiaomi
3	Điện thoại Xiaomi Redmi 12C (Helio G85)	Redmi	12.0	5.0	7	https://mobilecity.vn/dien-thoai/xiaomi-redmi-...	Xanh bạc hà	1950000.0	2950000.0	Android 12	Mediatek MT6769Z Helio G85 (12nm)\n8 nhân (2x2...	4.0	128.0	5000.0	Thanh + Cảm ứng	Xiaomi
4	Điện thoại Xiaomi Redmi 12C (Helio G85)	Redmi	12.0	5.0	7	https://mobilecity.vn/dien-thoai/xiaomi-redmi-...	Tím	1650000.0	2950000.0	Android 12	Mediatek MT6769Z Helio G85 (12nm)\n8 nhân (2x2...	4.0	64.0	5000.0	Thanh + Cảm ứng	Xiaomi

(a)

la_dien_thoai_cu	kich_thuoc_man_hinh	tan_so_quet	do_sang_man_hinh	loai_man_hinh	so_the_sim	loai_pin	cong_suat_sac	do_phan_giai_cam_sau	do_phan_giai_cam_truoc
False	6.71	NaN	500.0	LCD	2	Li-Po	10.0	[50.0, 0.08]	[5.0]
False	6.71	NaN	500.0	LCD	2	Li-Po	10.0	[50.0, 0.08]	[5.0]
False	6.71	NaN	500.0	LCD	2	Li-Po	10.0	[50.0, 0.08]	[5.0]
False	6.71	NaN	500.0	LCD	2	Li-Po	10.0	[50.0, 0.08]	[5.0]
False	6.71	NaN	500.0	LCD	2	Li-Po	10.0	[50.0, 0.08]	[5.0]

(b)

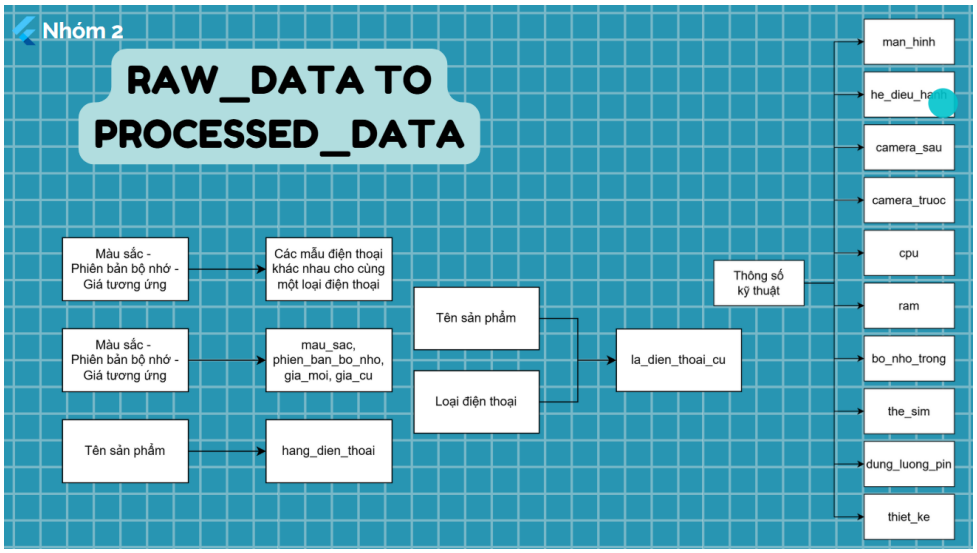
Hình 1: *5 dòng đầu của dữ liệu*

Các bước tiền xử lý được thực hiện cụ thể như sau:

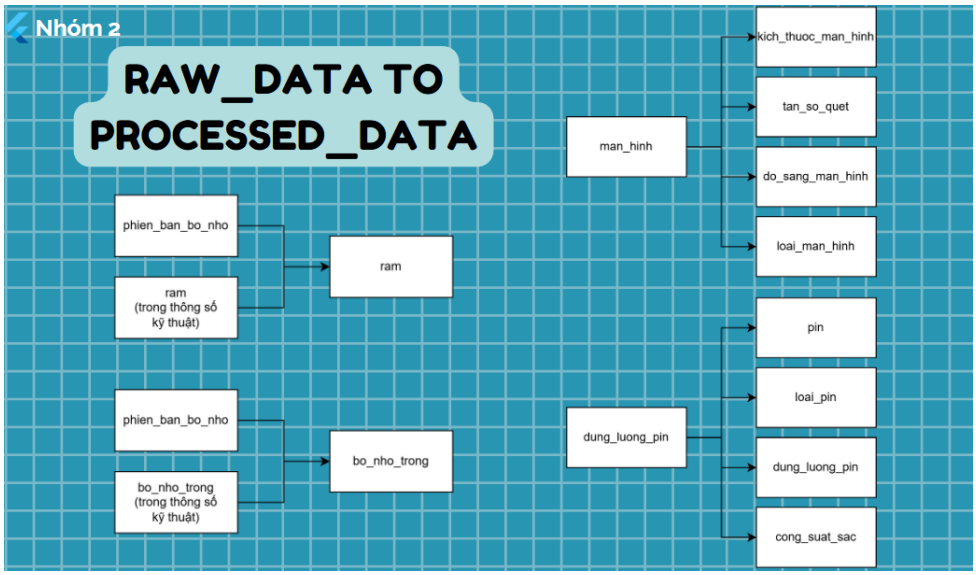
- Đọc dữ liệu từ file CSV: Dữ liệu được đọc từ file `raw_data.csv` vào DataFrame `data`.
- Xóa các bản ghi trùng lặp: Sử dụng `drop_duplicates()` để loại bỏ các dòng trùng lặp.

- Đặt lại tên cột: Tên các cột được đặt lại để dễ sử dụng hơn với danh sách tên mới. Ví dụ: "Tên điện thoại" được đổi thành `ten`.
- Loại bỏ các dòng không liên quan:
  - Xóa dòng có chứa `"Điện thoại Xiaomi SU7 (Ô tô điện chạy 800km/1 lần sạc)"`.
  - Loại bỏ các dòng chứa `Apple Watch` và `Máy Chơi Game PC` dựa trên cột `loai_dien_thoai`.
- Trích xuất thông tin thuộc tính từ chuỗi JSON:
  - Tách `mau_sac`, `phien_ban_bo_nho`, `gia_moi`, và `gia_cu` từ cột `mau_sac_phien_ban_bo_nho_gia_moi_gia_cu`.
  - Tách thông tin về màn hình, hệ điều hành, camera, CPU, RAM, bộ nhớ, SIM, pin, và thiết kế từ cột `thong_so_ky_thuat`.
- Xác định thuộc tính:
  - Hãng điện thoại: Trích xuất tên hãng từ cột `ten`.
  - Có phải điện thoại cũ hay không: Dựa vào từ khóa "cũ" trong cột `ten` và `loai_dien_thoai`.
  - Phiên bản bộ nhớ: Tách thông tin RAM và bộ nhớ trong từ cột `phien_ban_bo_nho`.
- Xử lý dữ liệu màn hình:
  - Trích xuất kích thước màn hình (`kich_thuoc_man_hinh`), tần số quét (`tan_so_quet`), độ sáng (`do_sang_man_hinh`), và loại màn hình (`loai_man_hinh`).
  - Loại bỏ cột `man_hinh`.
- Trích xuất thông tin về pin:
  - Loại pin (`loai_pin`), dung lượng pin (`dung_luong_pin`), và công suất sạc (`cong_suat_sac`) từ cột `dung_luong_pin`.
- Xử lý camera:
  - Trích xuất độ phân giải camera trước và sau (`do_phan_giai_cam_truoc`, `do_phan_giai_cam_sau`).
- Chuẩn hóa thông tin hệ điều hành: Dựa vào thông tin mô tả để xác định `Android`, `iOS` hoặc "Khác".
- Xử lý dữ liệu thiếu:
  - Điền giá trị mặc định cho cột `thoi_gian_bao_hanh` (0 nếu thiếu).
  - Thay chuỗi rỗng trong cột `mau_sac` bằng `None`.
- Lưu dữ liệu đã xử lý vào file `processed_data.csv`.

Bảng tham chiếu thuộc tính từ `raw_data.csv` sang `processed_data.csv`



(a)



(b)

4 Data Exploring (Khám phá dữ liệu)

4.1 Khái quát

Bước Data Exploring là một phần quan trọng trong quy trình xử lý dữ liệu. Đây là quá trình phân tích và khám phá dữ liệu để hiểu rõ hơn về cấu trúc, đặc điểm, và mối quan hệ giữa các biến. Trong bước này, thường bao gồm:

- Tóm tắt thống kê: Sử dụng các hàm như describe() hoặc info() để xem thông tin cơ bản như số lượng dữ liệu, kiểu dữ liệu, giá trị thiếu, giá trị trung bình, độ lệch chuẩn, v.v.

- Phân tích dữ liệu thiếu: Xác định các cột hoặc hàng chứa giá trị bị thiếu để quyết định loại bỏ hoặc xử lý dữ liệu.
- Phân phối dữ liệu: Quan sát phân phối của từng cột để phát hiện outliers, dữ liệu phân bố lệch, hoặc các bất thường khác.
- Mối quan hệ giữa các biến: Sử dụng biểu đồ hoặc ma trận tương quan để hiểu các mối quan hệ giữa các biến độc lập và biến mục tiêu.
- Trực quan hóa dữ liệu: Biểu đồ histogram, boxplot, scatter plot thường được sử dụng để làm rõ đặc tính dữ liệu.

4.2 Đặt ra các câu hỏi

Ở giai đoạn này ta sẽ đặt ra các câu hỏi để có thể hiểu rõ hơn về dữ liệu.

4.2.1 Có bao nhiêu dòng và cột trong bộ dữ liệu?

Theo như phân tích và tính toán nhóm em đã thống kê được kết quả: bộ dữ liệu có 8855 dong và 26 cột.

4.2.2 Mỗi dòng có ý nghĩa gì? Có quan trọng không khi mỗi dòng có ý nghĩa khác nhau?

Mỗi dòng đại diện cho một mẫu điện thoại di động với các thông số kỹ thuật và thông tin khác nhau, các thuộc tính giúp hiểu rõ hơn về đặc điểm, cấu hình và giá trị của từng mẫu điện thoại.
Rất quan trọng! Vì bộ dữ liệu này chứa thông tin về nhiều mẫu điện thoại khác nhau, mỗi dòng cần phải có các giá trị đúng và phản ánh chính xác từng mẫu điện thoại. Nếu các dòng có ý nghĩa khác nhau hoặc không đồng nhất, dữ liệu sẽ trở nên khó hiểu và không chính xác, ảnh hưởng đến các phân tích và kết luận bạn có thể rút ra từ dữ liệu.

4.2.3 Có dòng dữ liệu nào bị trùng lặp không?

Sau khi kiểm tra trùng lặp bằng cách **data.duplicated().sum()**, ta phát hiện được 8 dùng trùng lặp vì vậy ta sẽ loại bỏ để thuật tiện cho việc tính toán: **data = data.drop_duplicates()**

4.2.4 Các cột (thuộc tính) có ý nghĩa như thế nào?

Bảng 4: *Ý nghĩa các thuộc tính*

Thuộc tính	**Ý nghĩa**
ten	Tên của điện thoại (mẫu điện thoại)
loai_dien_thoai	Loại điện thoại (ví dụ: Smartphone, Feature Phone, v.v.)
thoi_gian_bao_hanh	Thời gian bảo hành của điện thoại (thường tính bằng tháng hoặc năm)
danh_gia	Đánh giá trung bình của người dùng cho sản phẩm (thường là điểm từ 1 đến 5)
so_luong_binh_luan	Số lượng bình luận hoặc đánh giá mà sản phẩm nhận được từ người dùng
duong_dan	Đường dẫn URL đến hình ảnh hoặc trang sản phẩm trên website
mau_sac	Màu sắc của điện thoại (ví dụ: Đen, Trắng, Vàng, v.v.)
gia_moi	Giá mới của điện thoại (giá bán lẻ hiện tại)
gia_cu	Giá cũ của điện thoại (nếu có, thường là giá bán ban đầu hoặc giá khuyến mãi trước đây)
he_dieu_hanh	Hệ điều hành mà điện thoại sử dụng (ví dụ: Android, iOS, v.v.)
cpu	Bộ vi xử lý (CPU) của điện thoại (ví dụ: Snapdragon, Apple A-series, v.v.)

Thuộc tính	Ý nghĩa
ram	Bộ nhớ RAM của điện thoại (đơn vị tính là GB)
bo_nho_trong	Bộ nhớ trong của điện thoại (đơn vị tính là GB)
dung_luong_pin	Dung lượng pin của điện thoại (thường tính bằng mAh - miliampere giờ)
thiet_ke	Thiết kế của điện thoại (ví dụ: kiểu dáng, chất liệu, thiết kế mỏng nhẹ, v.v.)
hang_dien_thoai	Thương hiệu hoặc nhà sản xuất điện thoại (ví dụ: Samsung, Apple, Xiaomi, v.v.)
la_dien_thoai_cu	Liệu điện thoại là hàng cũ hay không (1: có, 0: không)
kich_thuoc_man_hinh	Kích thước màn hình của điện thoại (đơn vị tính là inch)
tan_so_quet	Tần số quét màn hình (hertz, ví dụ: 60Hz, 120Hz, v.v.)
do_sang_man_hinh	Độ sáng tối đa của màn hình (thường tính bằng nits)
loai_man_hinh	Loại màn hình của điện thoại (ví dụ: AMOLED, LCD, v.v.)
so_the_sim	Số lượng khe SIM hỗ trợ (ví dụ: 1 SIM, 2 SIM)
loai_pin	Loại pin của điện thoại (ví dụ: Li-ion, Li-polymer, v.v.)
cong_suat_sac	Công suất sạc tối đa của điện thoại (thường tính bằng W - watt)
do_phan_giai_cam_sau	Một list các độ phân giải của camera sau (đơn vị tính là megapixels - MP)
do_phan_giai_cam_truoc	Một list các độ phân giải của camera trước (đơn vị tính là megapixels - MP)

Bảng 5: *Phân loại dữ liệu của các thuộc tính*

Thuộc tính	Phân loại dữ liệu
ten	Categorical, Discrete, Nominal
loai_dien_thoai	Categorical, Discrete, Nominal
thoi_gian_bao_hanh	Numerical, Discrete, Ordinal
danh_gia	Numerical, Continuous, Ordinal
so_luong_binh_luan	Numerical, Discrete, Ratio
duong_dan	Categorical, Discrete, Nominal
mau_sac	Categorical, Discrete, Nominal
gia_moi	Numerical, Continuous, Ratio
gia_cu	Numerical, Continuous, Ratio
he_dieu_hanh	Categorical, Discrete, Nominal
cpu	Categorical, Discrete, Nominal
ram	Numerical, Continuous, Ratio
bo_nho_trong	Numerical, Continuous, Ratio
dung_luong_pin	Numerical, Continuous, Ratio
thiet_ke	Categorical, Discrete, Nominal
hang_dien_thoai	Categorical, Discrete, Nominal
la_dien_thoai_cu	Categorical, Discrete, Nominal
kich_thuoc_man_hinh	Numerical, Continuous, Ratio
tan_so_quet	Numerical, Continuous, Ordinal
do_sang_man_hinh	Numerical, Continuous, Ratio
loai_man_hinh	Categorical, Discrete, Nominal
so_the_sim	Categorical, Discrete, Nominal
loai_pin	Categorical, Discrete, Nominal
cong_suat_sac	Numerical, Continuous, Ratio
do_phan_giai_cam_sau	Categorical, Discrete, Nominal
do_phan_giai_cam_truoc	Categorical, Discrete, Nominal

4.2.5 Mỗi cột hiện tại có kiểu dữ liệu gì? Có cột nào có kiểu dữ liệu không phù hợp cho việc xử lý tiếp theo không?

Sau khi kiểm tra kiểu dữ liệu của các thuộc tính bằng **dtypes** ta sẽ tiến hành chuyển đổi một số thuộc tính phân loại đang có kiểu dữ liệu là 'object' sang 'category'.

Bảng 6: *Loại dữ liệu cập nhật của các thuộc tính*

Thuộc tính	Loại dữ liệu
ten	object
loai_dien_thoai	category
thoi_gian_bao_hanh	float64
danh_gia	float64
so_luong_binh_luan	int64
duong_dan	object
mau_sac	category
gia_moi	float64
gia_cu	float64
he_dieu_hanh	category
cpu	category
ram	float64
bo_nho_trong	float64
dung_luong_pin	float64
thiet_ke	category
hang_dien_thoai	category
la_dien_thoai_cu	category
kich_thuoc_man_hinh	float64
tan_so_quet	float64
do_sang_man_hinh	float64
loai_man_hinh	category
so_the_sim	int64
loai_pin	category
cong_suat_sac	float64
do_phan_giai_cam_sau	object
do_phan_giai_cam_truoc	object

4.2.6 Tỷ lệ giá trị bị thiếu là bao nhiêu?

Sau khi phân tích và tín toàn bằng các hàm hỗ trợ ta thống kê được số liệu sau.

Bảng 7: *Số lượng và tỉ lệ phần trăm giá trị thiếu của các thuộc tính*

Thuộc tính	Số lượng giá trị thiếu	Tỉ lệ giá trị thiếu (%)
gia_cu	3689.0	41.697751
do_sang_man_hinh	3448.0	38.973663
loai_pin	2166.0	24.482876
tan_so_quet	2101.0	23.748163
gia_moi	1504.0	17.000113
cong_suat_sac	487.0	5.504691
kich_thuoc_man_hinh	335.0	3.786594
loai_man_hinh	330.0	3.730078
bo_nho_trong	319.0	3.605742
mau_sac	279.0	3.153611
dung_luong_pin	200.0	2.260653
thiet_ke	114.0	1.288572
ram	17.0	0.192156

Thuộc tính	Số lượng giá trị thiếu	Tỉ lệ giá trị thiếu (%)
cpu	15.0	0.169549
do_phan_giai_cam_sau	0.0	0.000000
so_the_sim	0.0	0.000000
ten	0.0	0.000000
la_dien_thoai_cu	0.0	0.000000
hang_dien_thoai	0.0	0.000000
loai_dien_thoai	0.0	0.000000
he_dieu_hanh	0.0	0.000000
duong_dan	0.0	0.000000
so_luong_binh_luan	0.0	0.000000
danh_gia	0.0	0.000000
thoi_gian_bao_hanh	0.0	0.000000
do_phan_giai_cam_truoc	0.0	0.000000

Nhận xét:

- Các tỷ lệ dữ liệu bị thiếu trong bộ dữ liệu này dao động khá rộng, với "gia_cu"và "do_sang_man_hinh"có tỷ lệ thiếu cao nhất, khoảng 41% và 40% tương ứng. Các thuộc tính khác như "tan_so_quet"và "loai_pin"cũng có tỷ lệ thiếu đáng kể, khoảng 24%. Trong khi đó, một số thuộc tính khác có tỷ lệ dữ liệu bị thiếu thấp, dưới 5%, còn một số thuộc tính còn lại thì không bị thiếu dữ liệu.
- Ta sẽ tiến hành loại bỏ các cột có tỷ lệ thiếu $> 50\%$

4.2.7 Đối với kiểu dữ liệu dạng numeric, dữ liệu được phân bố như thế nào?

Để hiểu rõ hơn về dữ liệu ta sẽ tính các giá trị 'missing_ratio', 'min', 'lower_quartile', 'median', 'upper_quartile', 'max'

Bảng 8: *Thống kê chi tiết các thuộc tính dạng* ***Numeric***

Thuộc tính	missing ratio (%)	min	lower quartile	median	upper quartile	max
thoi_gian_bao_hanh	0.0	0.0	6.0	12.0	12.0	30.0
danh_gia	0.0	4.0	5.0	5.0	5.0	5.0
so_luong_binh_luan	0.0	0.0	0.0	7.0	67.0	5490.0
gia_moi	17.0	100000.0	4750000.0	7650000.0	13250000.0	51250000.0
gia_cu	41.7	825000.0	5450000.0	8450000.0	13950000.0	68000000.0
ram	0.19	0.75	6.0	8.0	12.0	24.0
bo_nho_trong	3.61	16.0	128.0	256.0	256.0	1024.0
dung_luong_pin	2.26	4.5	4500.0	5000.0	5000.0	15500.0
kich_thuoc_man_hinh	3.79	3.5	6.5	6.67	6.74	8.7
tan_so_quet	23.75	60.0	120.0	120.0	120.0	240.0
do_sang_man_hinh	38.97	400.0	800.0	1200.0	1600.0	6000.0
so_the_sim	0.0	1.0	2.0	2.0	2.0	2.0
cong_suat_sac	5.5	1.0	25.0	45.0	80.0	240.0

4.2.8 Đối với kiểu dữ liệu dạng category, dữ liệu được phân bố như thế nào?

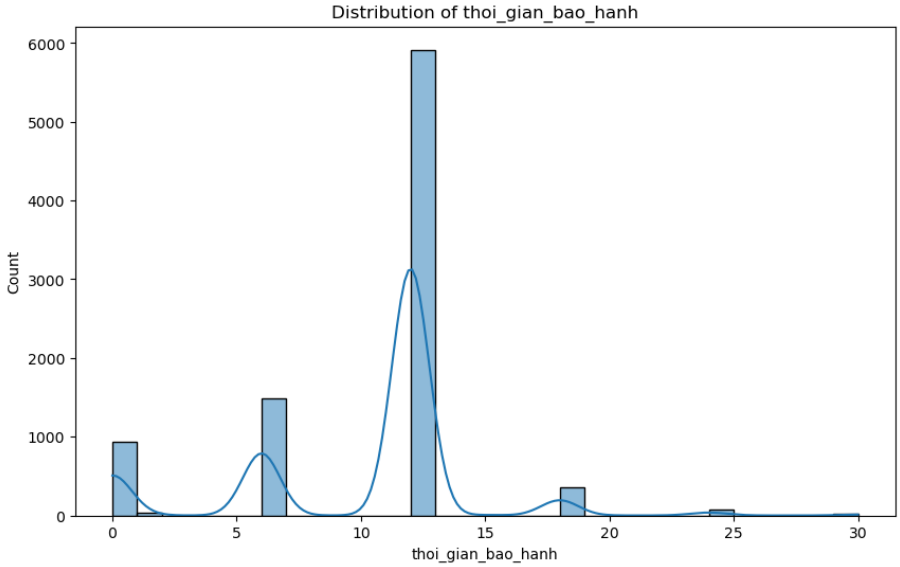
Để hiểu rõ hơn về dữ liệu ta sẽ tính các giá trị 'missing_ratio', 'num_values', 'value_ratios'.

Bảng 9: *Thống kê chi tiết các thuộc tính dạng* ***Category***

Thuộc tính	**missing ratio (%)**	**num values**	**value ratios (%)**
loai_dien_thoai	0.0	43	'Vivo': 13.2, 'Redmi': 12.5, 'Samsung Chính hãng': 12.0
mau_sac	3.15	477	'Đen': 23.0, 'Xanh': 13.3, 'Trắng': 13.0, 'Tím': 10.0
he_dieu_hanh	0.0	23	'Android 13': 19.5, 'Android 14': 18.8, 'Android 12': 15.5
cpu	0.17	701	'Qualcomm SM8475 Snapdragon 8+ Gen 1 (4 nm)': 15.0, 'MediaTek Dimensity 9200 (4 nm)': 12.0
thiet_ke	1.29	677	'Thanh + Cảm ứng': 20.8, 'Thanh + cảm ứng': 20.2, 'Nguyên khối': 15.5
hang_dien_thoai	0.0	56	'Xiaomi': 27.9, 'Samsung': 14.0, 'Vivo': 13.6
la_dien_thoai_cu	0.0	2	False: 79.4, True: 20.6
loai_man_hinh	3.73	3	'AMOLED': 57.3, 'OLED': 21.6, 'LCD': 21.0
loai_pin	24.48	3	'Li-Po': 70.9, 'Li-Ion': 23.8, 'Si/C': 5.3

4.2.9 Thời gian bảo hành?

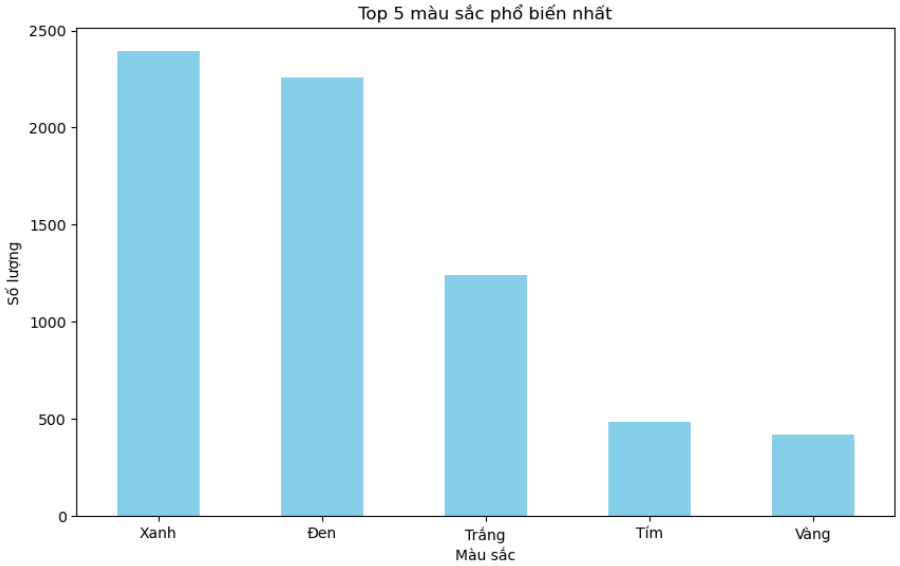
Thời gian bảo hành là khoảng thời gian mà nhà sản xuất hoặc nhà bán lẻ cam kết sửa chữa hoặc thay thế miễn phí các lỗi kỹ thuật của điện thoại phát sinh trong quá trình sử dụng, miễn là lỗi đó nằm trong điều kiện bảo hành. Ta sẽ trực quan thời gian bảo hành theo hình dưới đây:



Nhận xét: Biểu đồ phân bố thời gian bảo hành cho thấy có ba điểm nhấn chính ở các khoảng thời gian bảo hành 0, 12, và 24 tháng. Điều này cho thấy hầu hết các sản phẩm có thời gian bảo hành tập trung chủ yếu ở 12 tháng, với một số lượng lớn sản phẩm được bảo hành trong 24 tháng và một số ít không có bảo hành (0 tháng).

4.2.10 Màu sắc?

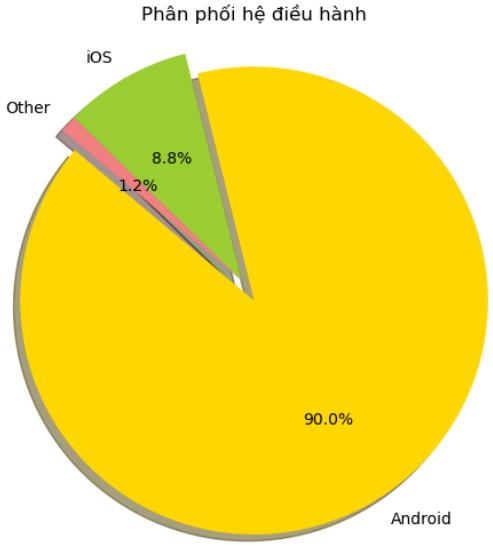
Màu sắc của điện thoại là tùy chọn về màu ngoại hình của thiết bị, thường bao gồm mặt lưng, khung viền và các chi tiết thiết kế bên ngoài. Màu sắc giúp điện thoại trở nên thẩm mỹ, cá nhân hóa và phù hợp với gu thẩm mỹ của người dùng. Các màu phổ biến bao gồm: đen, trắng, bạc, vàng, xanh, đỏ, và các phiên bản màu đặc biệt khác.



4.2.11 Hệ điều hành?

Hiện tại, điện thoại thường sử dụng một trong hai hệ điều hành chính:

- Hệ điều hành Android: được phát triển bởi Google và là hệ điều hành mở, được sử dụng rộng rãi trên nhiều thương hiệu điện thoại như Samsung, Xiaomi, Oppo, Vivo, Huawei và các hãng khác. Hệ điều hành này có nhiều phiên bản khác nhau, thường xuyên được cập nhật để cải thiện tính năng và hiệu năng cho người dùng.
- Hệ điều hành iOS: được phát triển độc quyền bởi Apple và chỉ có trên các thiết bị iPhone. Đây là hệ điều hành đóng, được tối ưu hóa cao, với nhiều phiên bản được cập nhật định kỳ để nâng cấp trải nghiệm người dùng và bảo mật hệ thống.

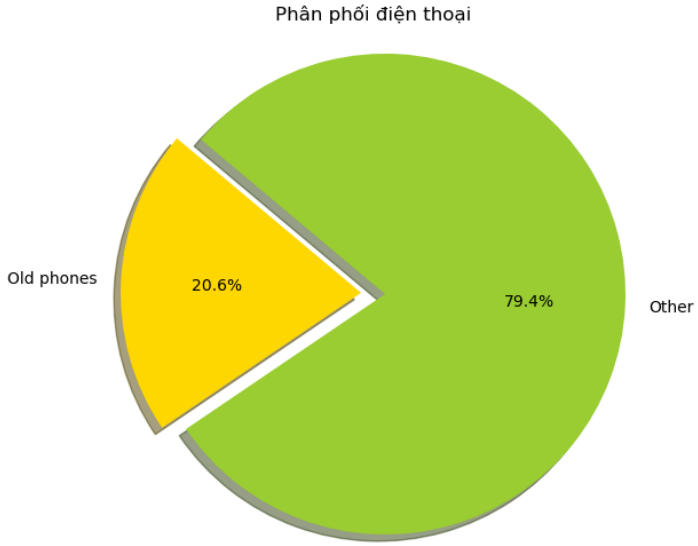


4.2.12 Có bao nhiêu điện thoại cũ trong bộ dữ liệu?

Điện thoại cũ là điện thoại đã qua sử dụng, không còn trong tình trạng mới 100%. Chúng có thể được bán lại sau khi người dùng sử dụng một thời gian hoặc đã qua tân trang, sửa chữa. Đặc điểm của điện thoại cũ:

- Có thể có dấu hiệu hao mòn như trầy xước, pin giảm hiệu suất.
- Giá bán thường rẻ hơn so với điện thoại mới.
- Chất lượng và bảo hành phụ thuộc vào tình trạng máy và nơi bán.

Điện thoại cũ là lựa chọn phổ biến cho những ai muốn tiết kiệm chi phí mà vẫn sở hữu một thiết bị có tính năng tốt.

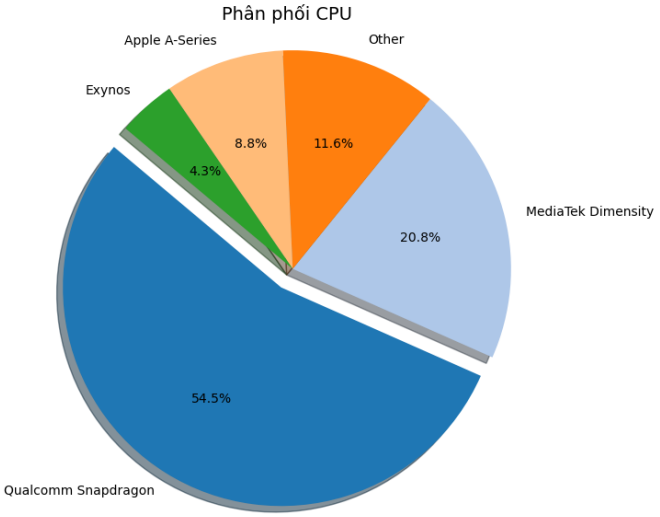


4.2.13 CPU?

CPU là bộ xử lý trung tâm, quyết định hiệu năng và tốc độ xử lý của thiết bị. Các loại CPU chính bao gồm:

- **Qualcomm Snapdragon**: Hiệu năng mạnh mẽ, tối ưu năng lượng, phổ biến trên nhiều smartphone cao cấp và tầm trung.
- **MediaTek Dimensity**: Hiệu năng tốt, giá cạnh tranh, hỗ trợ công nghệ 5G, thường xuất hiện trên các thiết bị tầm trung.
- **Apple A-Series**: CPU độc quyền của Apple, nổi bật với hiệu năng đơn nhân mạnh mẽ và tối ưu hóa phần mềm cho iPhone, iPad.
- **Exynos**: Được Samsung phát triển, cung cấp hiệu năng ổn định và khả năng xử lý đồ họa tốt.

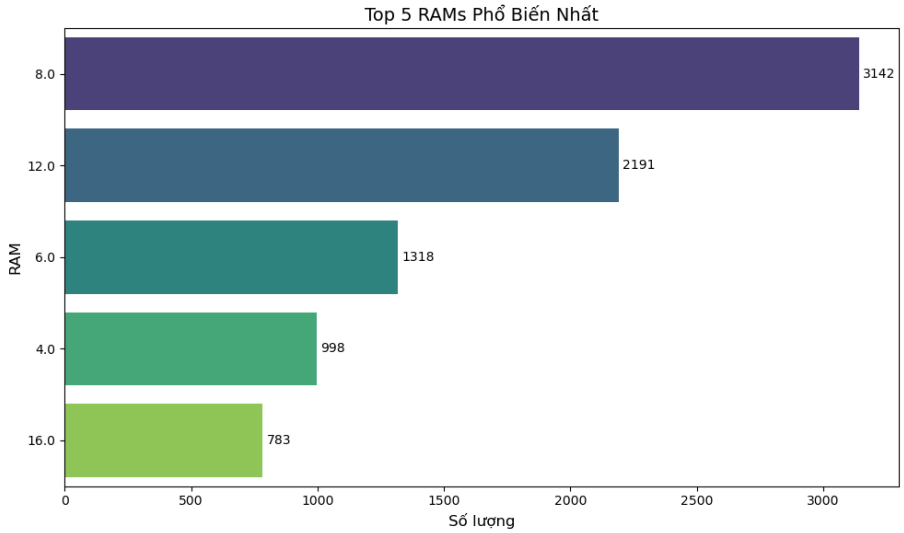
Các CPU này có nhiều phiên bản khác nhau, liên tục được cải tiến để đáp ứng nhu cầu về hiệu năng, tiết kiệm pin và hỗ trợ công nghệ mới như **AI**, **5G** và **đồ họa cao cấp**.



4.2.14 RAM?

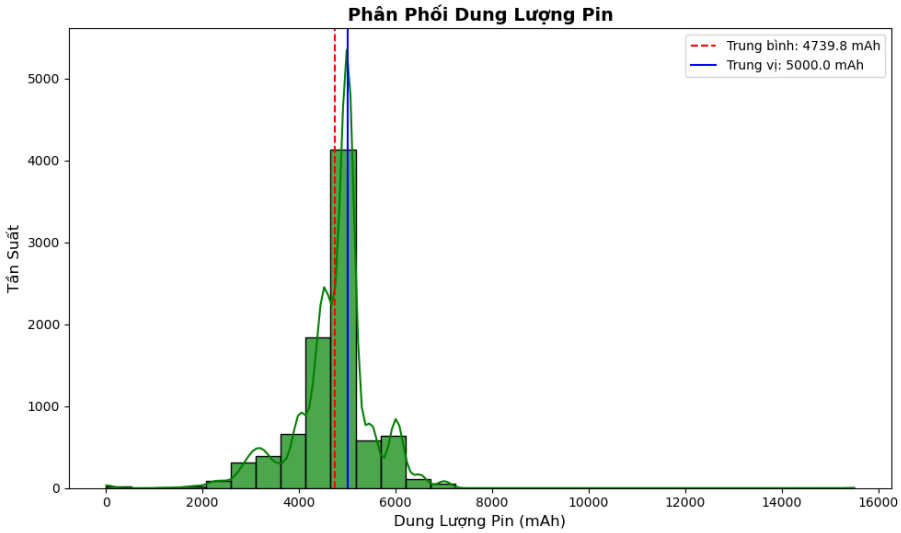
RAM trong điện thoại là bộ nhớ tạm thời, đóng vai trò quan trọng trong việc lưu trữ và truy cập dữ liệu của các ứng dụng đang hoạt động. RAM ảnh hưởng trực tiếp đến khả năng **đa nhiệm** và **hiệu suất hoạt động** của thiết bị. **Dung lượng RAM** cũng có nhiều mức khác nhau tùy theo phân khúc điện thoại:

- **Phổ thông**: 4GB - 6GB.
- **Tầm trung**: 8GB - 12GB.
- **Cao cấp**: 12GB - 16GB hoặc cao hơn.



4.2.15 Dung lượng Pin?

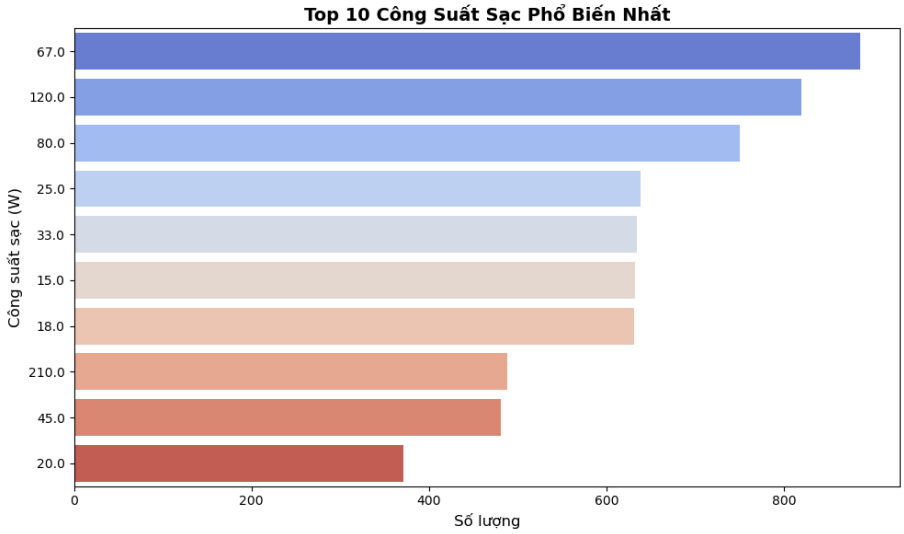
Dung lượng pin là chỉ số đo khả năng lưu trữ năng lượng của pin điện thoại, thường được tính bằng đơn vị **mAh** (milliampere-hour). Dung lượng pin càng cao, thời gian sử dụng điện thoại càng lâu trước khi cần sạc lại.



4.2.16 Công suất sạc?

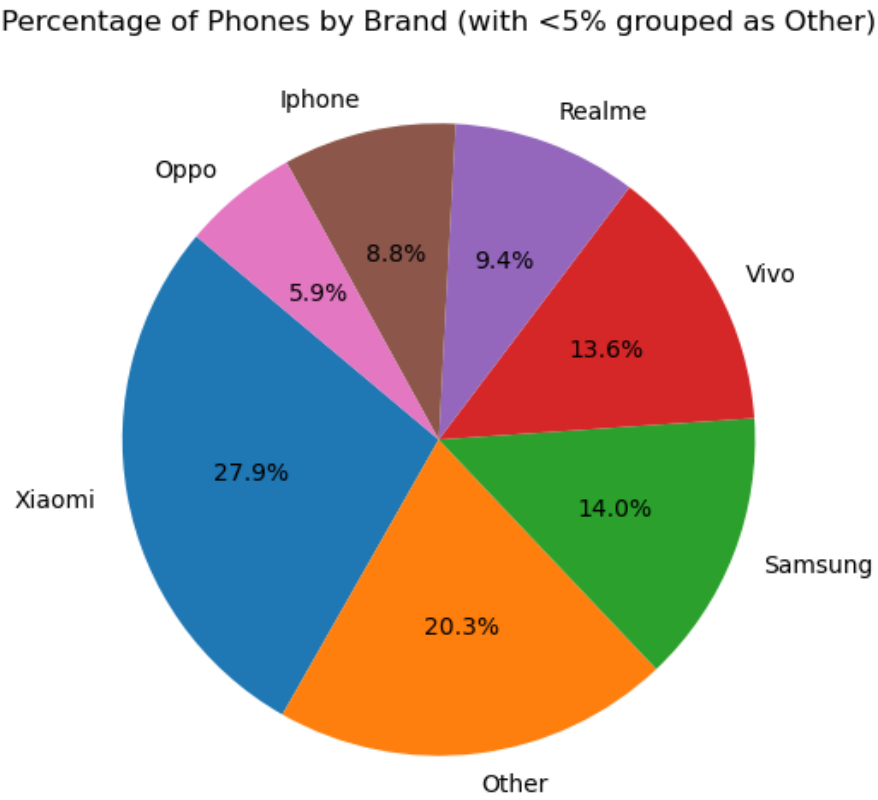
Công suất sạc là chỉ số đo tốc độ sạc pin của điện thoại, thường tính bằng **Watt (W)**. Công suất càng cao, thời gian sạc pin càng nhanh. **Xu hướng hiện nay**:

- Điện thoại ngày càng được trang bị **công suất sạc lớn** (từ **33W** đến **120W**), đặc biệt trên các dòng flagship và gaming phone.
- Xu hướng sạc siêu nhanh giúp rút ngắn thời gian sạc, đáp ứng nhu cầu sử dụng liên tục của người dùng.



4.2.17 Hãng điện thoại?

Hãng điện thoại là **thương hiệu** hoặc **nhà sản xuất** cung cấp các thiết bị di động trên thị trường. Mỗi hãng điện thoại có dòng sản phẩm riêng với thiết kế, tính năng và công nghệ khác nhau, phục vụ nhiều phân khúc khách hàng.



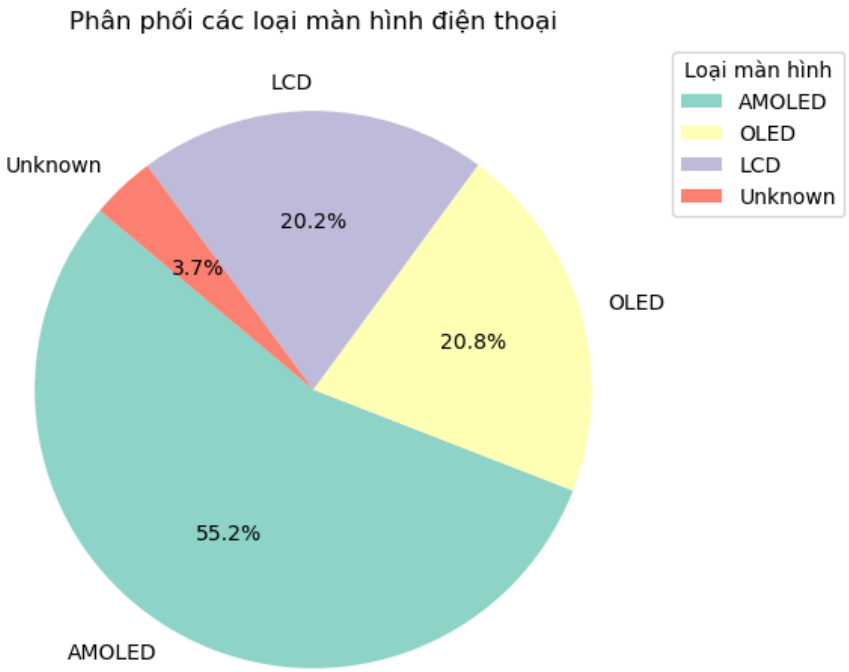
4.2.18 Loại màn hình?

Màn hình điện thoại là thành phần hiển thị thông tin và là mặt tương tác chính với người dùng, thường tính bằng **inch**. Màn hình càng lớn, trải nghiệm xem càng thú vị. **Các loại màn hình phổ biến**:

- **LCD (Liquid Crystal Display)**: Đây là loại màn hình sử dụng tinh thể lỏng, phổ biến với mức giá tầm trung. Màn hình LCD có độ sáng tốt nhưng màu sắc không bằng các loại màn hình khác.
- **OLED (Organic Light-Emitting Diode)**: Màn hình này sử dụng các hạt phát quang hữu cơ, cho màu đen sâu và màu sắc rực rỡ. Các điện thoại cao cấp thường trang bị màn hình OLED.
- **AMOLED (Active Matrix Organic Light-Emitting Diode)**: Là phiên bản nâng cấp của OLED, với khả năng tiết kiệm năng lượng tốt hơn và màu sắc hiển thị đậm nét hơn.

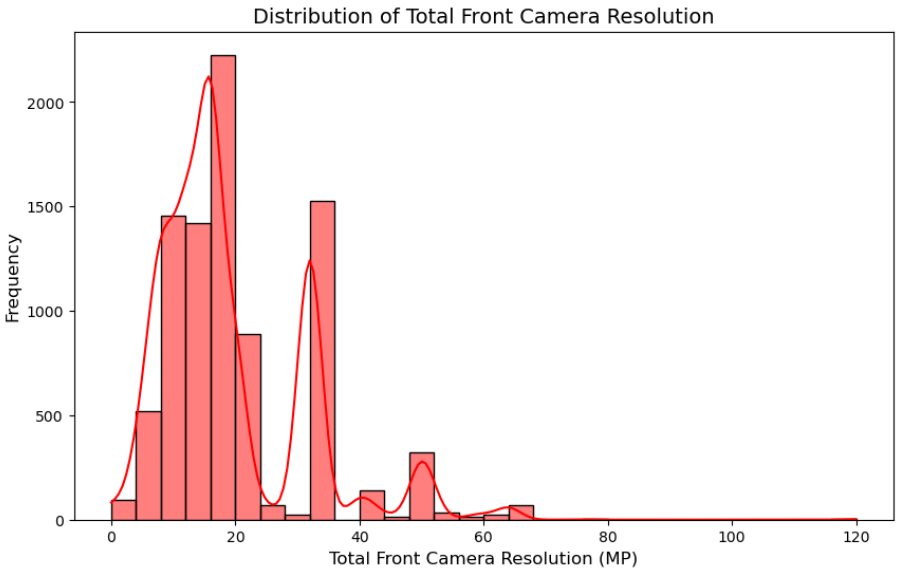
Xu hướng hiện nay:

- Nhiều hãng điện thoại chú trọng phát triển màn hình có độ phân giải cao và tần số quét lớn (từ **90Hz** đến **120Hz**) để đáp ứng nhu cầu trải nghiệm mượt mà trong các tác vụ và chơi game.



4.2.19 Độ phân giải cam trước?

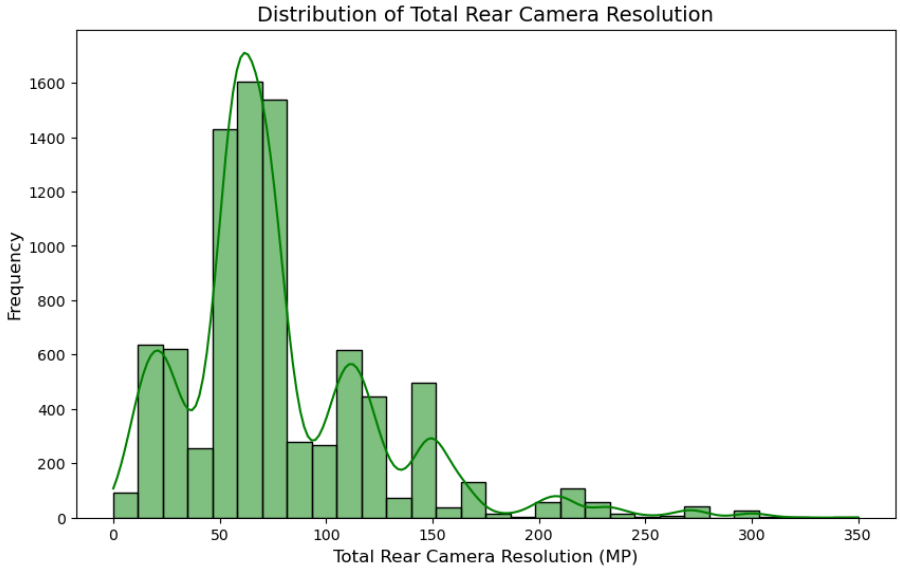
Độ phân giải camera trước là chỉ số đo lường số lượng điểm ảnh mà camera trước của điện thoại có thể chụp được, thường tính bằng **megapixel (MP)**. Độ phân giải càng cao, chất lượng hình ảnh càng sắc nét, chi tiết hơn, đặc biệt khi chụp selfie hoặc gọi video. Xu hướng hiện nay là điện thoại được trang bị **nhiều camera** để tăng cường khả năng chụp ảnh và quay video, đáp ứng nhu cầu đa dạng của người dùng. Camera trước cũng ngày càng được nâng cấp với độ phân giải cao, tích hợp AI và các công nghệ hiện đại để mang lại trải nghiệm chụp ảnh selfie tốt nhất.



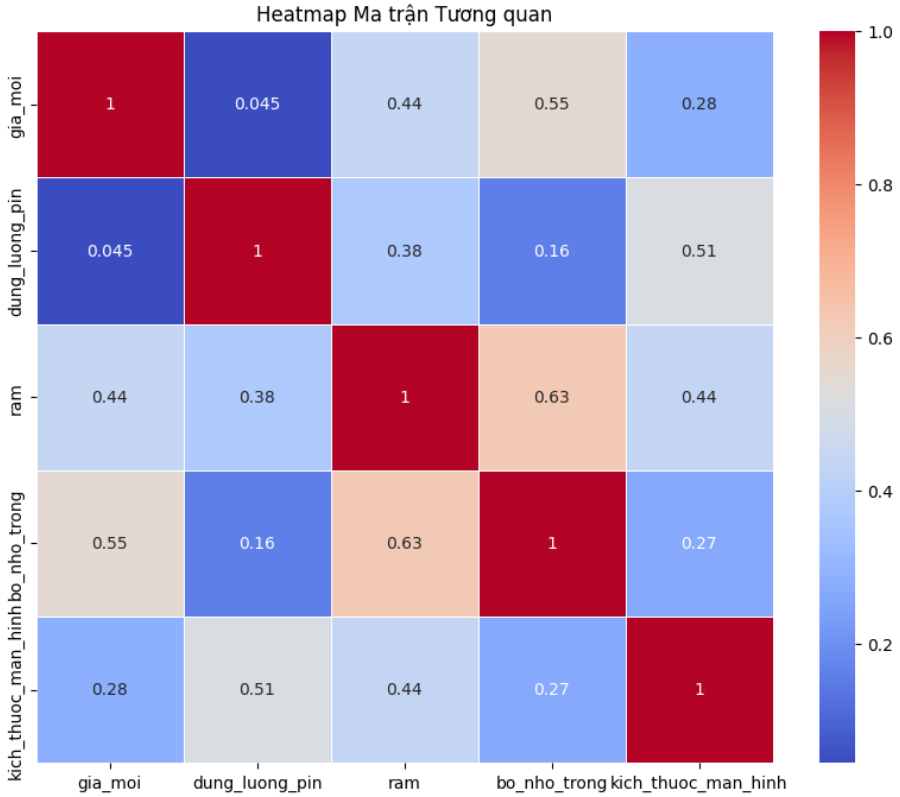
4.2.20 Độ phân giải cam sau?

Độ phân giải camera sau ngày càng cao, đặc biệt là camera chính với độ phân giải từ **48MP** trở lên, thậm chí đạt **108MP** hoặc **200MP** trên các thiết bị cao cấp. Điện thoại hiện đại thường có **nhiều**

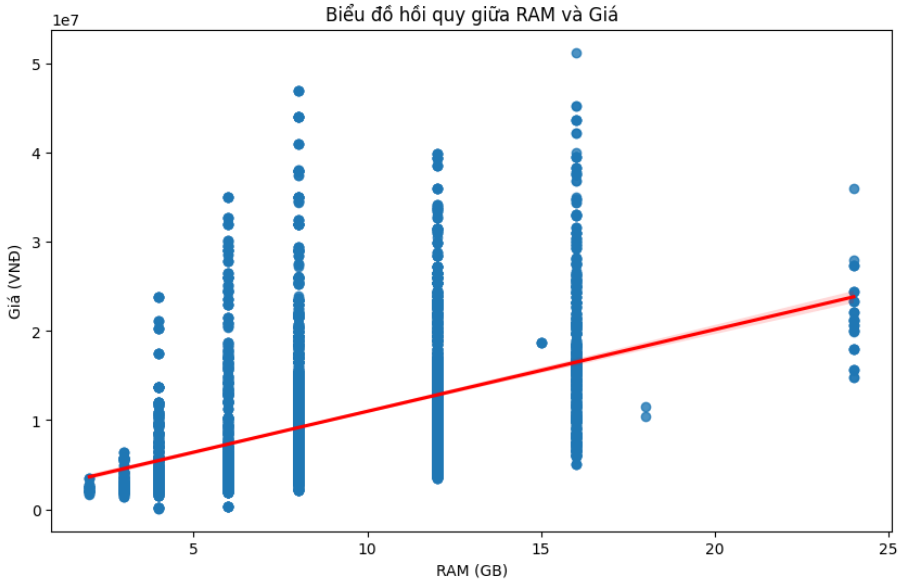
camera sau để phục vụ các nhu cầu khác nhau như chụp ảnh góc rộng, zoom xa, xóa phông và chụp cận cảnh, kết hợp với công nghệ **AI** và quay video chất lượng cao (**4K/8K**).

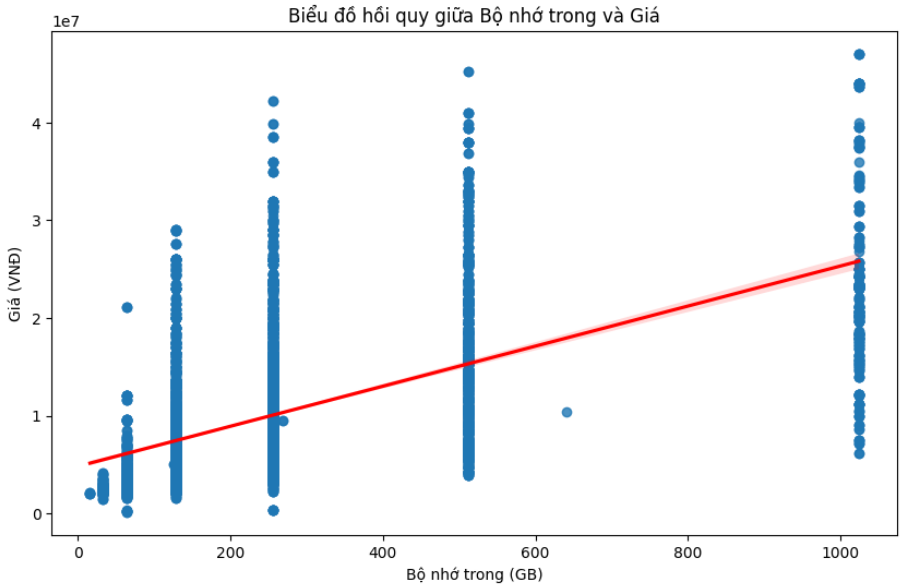


4.2.21 Phân tích tương quan giữa: Giá, Dung lượng pin, RAM, Bộ nhớ trong, Kích thước màn hình?



4.2.22 Mối quan hệ giữa Giá và các thuộc tính: RAM, Bộ nhớ trong





5 Defining the Problem (Xác định các vấn đề)

Sau khi tiền xử lý và khám phá dữ liệu ta sẽ đi đặt ra các vấn đề, câu hỏi cần giải đáp dựa trên dữ liệu đã được làm sạch.

5.1 Mức giá giảm có ảnh hưởng đến lượt đánh giá và mức độ hài lòng của khách hàng không?

5.1.1 Mục đích

- Đánh giá tác động của giá cả đến hành vi tiêu dùng của khách hành để tìm ra những chiến thuật kinh doanh tốt nhất.

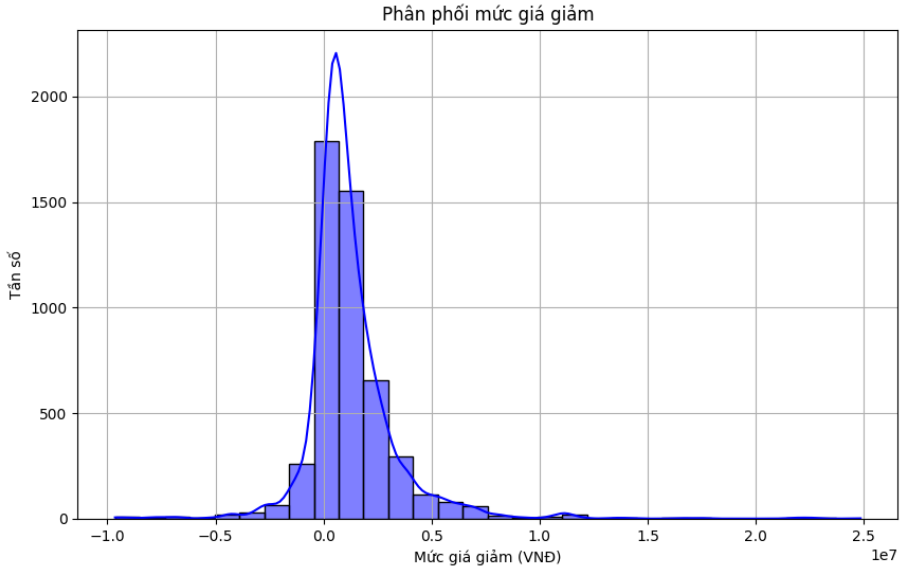
- Đưa ra các quyết định kinh doanh thông minh dựa trên dữ liệu đã được xử lý.

5.1.2 Tiền xử lý

- Ta sẽ tính toán dựa trên các cột gia_cu (giá cũ), gia_moi (giá mới), và so_luong_binh_luan (số lượng bình luận). Và loại bỏ các hàng có giá trị bị thiếu (dropna).
- Tạo cột mới có tên **Mức giá giảm**, là kết quả của phép trừ giữa gia_cu và gia_moi. Điều này xác định mức giảm giá cho mỗi sản phẩm.
- Tạo cột mới có tên **Có giảm giá**, trong đó mỗi giá trị là True nếu Mức giá giảm > 0 (tức là giá cũ cao hơn giá mới), và False nếu không.

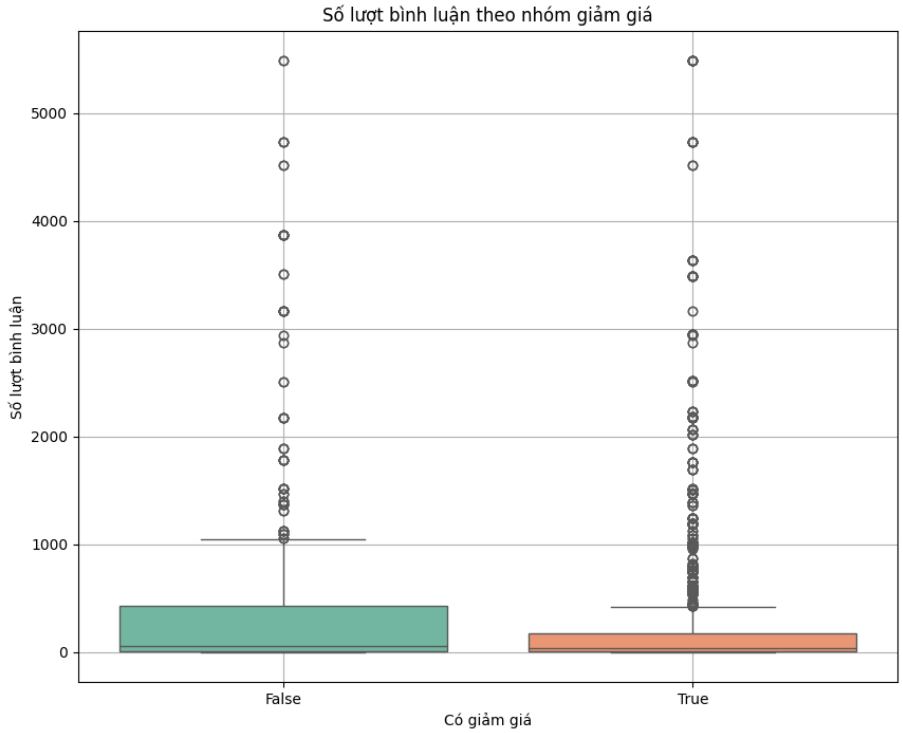
5.1.3 Trực quan hóa và nhận xét

Trực quan phân phối dữ liệu của mức giá giảm



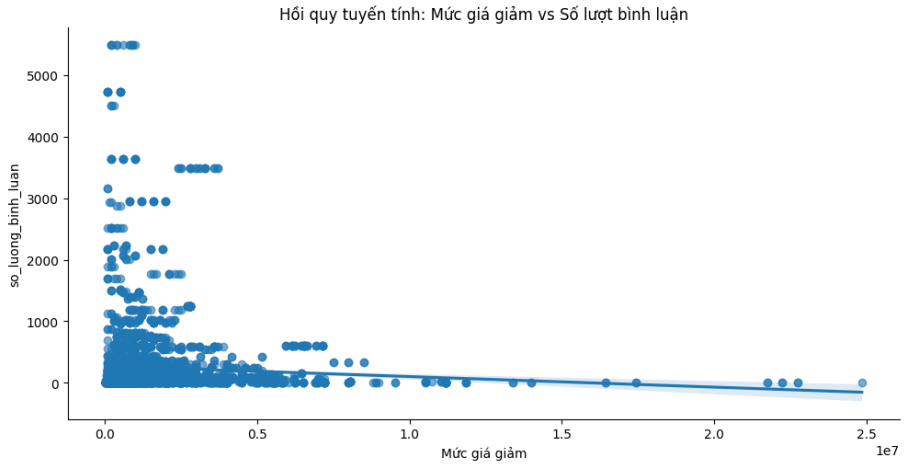
- Nhận xét
  - Biểu đồ histogram cho thấy ‘Mức giá giảm‘ chủ yếu tập trung gần giá trị 0, với 1 ít sản phẩm có mức giá giảm lớn hơn nhưng lại không đáng kể.
  - Điều này cho thấy rằng hầu hết các sản phẩm không có mức giảm giá quá lớn

Phân tích nhóm (Có giảm giá vss Không giảm giá)



- Nhận xét
  - Biểu đồ trên ta có thể thấy không có sự khác biệt quá lớn giữa số lượt bình luật của 2 nhóm (có giảm giá và không giảm giá)
  - Các giá trị ngoại lai (outliers) của nhóm không giảm giá có phạm vi phân bố rộng hơn
  - Trung bình của sản phẩm không giảm giá > sản phẩm giảm giá. Điều này cũng cho thấy rằng lượt bình luận của sản phẩm không giảm giá nhiều hơn.

Hồi quy tuyến tính



- Nhận xét
  - Hệ số tương quan giữa 'Mức giá giảm' và 'Số lượt đánh giá' là -0.0063 rất gần với 0.
  - Từ hệ số tương quan ta có thể kết luận là hầu như không có mối tương quan tuyến tính giữa 2 thuộc tính này. Có thể là do 'Mức giá giảm' không phải là yếu tố then chốt ảnh hưởng đến 'Số lượng đánh giá'.
  - Đường hồi quy gần như phẳng, thể hiện mối quan hệ yếu giữa 2 thuộc tính này hoặc không tồn tại.

- Số lượt đánh giá (5000+) rải rác không đồng đều và có thể thấy nó khong phụ thuộc rõ ràng vào mức giá giảm.

Kết luận: Mức giá giảm không phải là yếu tố quyết định khiến khách hàng để lại đánh giá. Thay vào đó, người tiêu dùng có xu hướng đánh giá sản phẩm dựa trên chất lượng, trải nghiệm sử dụng, hoặc sự hài lòng tổng thể. Điều này nhấn mạnh rằng doanh nghiệp cần tập trung vào việc cải thiện chất lượng sản phẩm và dịch vụ hơn là chỉ dựa vào các chiến lược giảm giá để thu hút phản hồi từ khách hàng.

5.2 Thời gian bảo hành có ảnh hưởng đến sự hài lòng của khách hàng không?

5.2.1 Mục đích

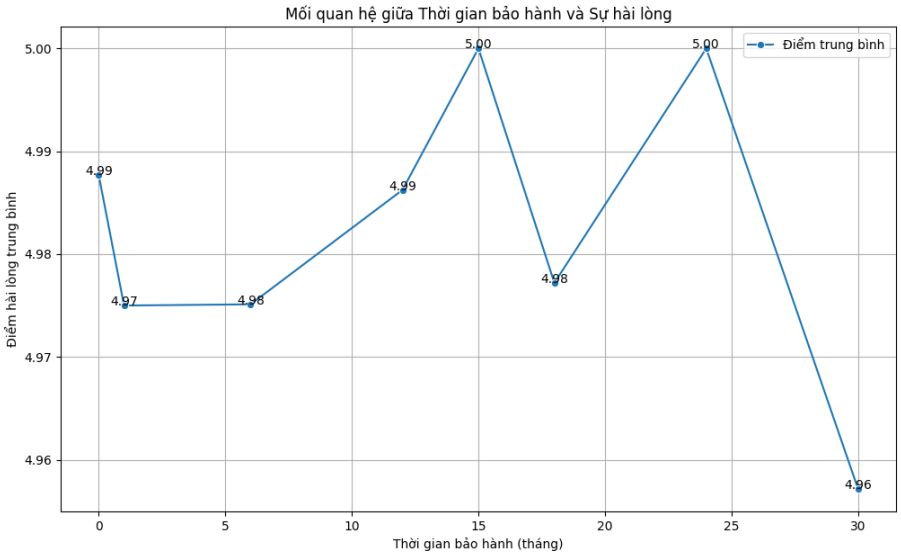
- Để làm rõ xem mối liên quan giữa thời gian bảo hành và điểm đánh giá.
- Tìm kiếm chiến lược kinh doanh thông minh giựa trên những đúc kết tính toán. Nếu thời gian bảo hành có tác động đến điểm đánh gái thì doanh nghiệp nên cân nhắc điều chỉnh chính sách bảo hành của mình để nâng cao sự hài lòng cũng như trung thành của khách hàng.

5.2.2 Tiền xử lý

- DataFrame được tạo từ hai cột: thoi_gian_bao_hanh (thời gian bảo hành) và danh_gia (điểm đánh giá).
- Loại bỏ các giá trị thiếu để đảm bảo dữ liệu đầy đủ và chính xác.
- Trung bình điểm đánh giá được tính theo từng khoảng thời gian bảo hành (tháng).
- Dữ liệu này giúp nhận biết liệu thời gian bảo hành dài hơn có làm tăng mức độ hài lòng hay không.

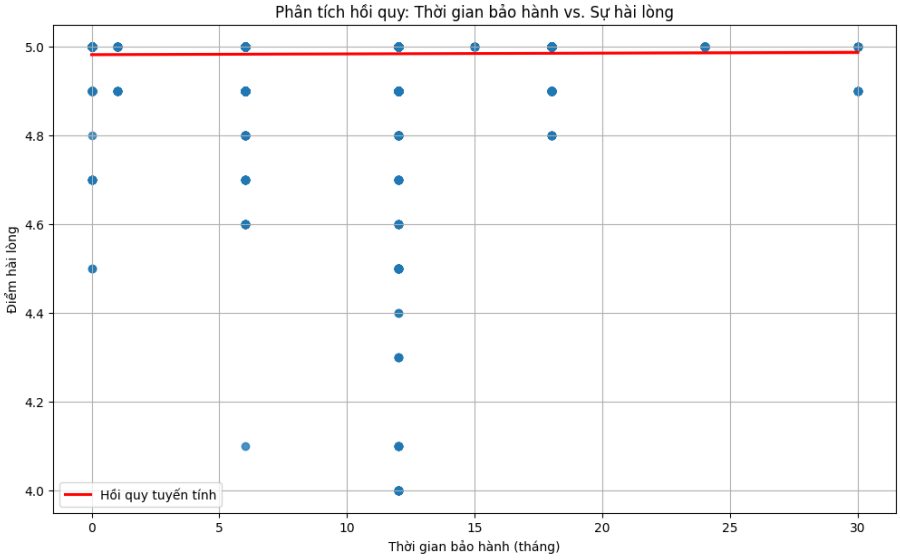
5.2.3 Trực quan hóa và nhận xét

Biểu đồ đường (Thời gian bảo hành vs. Sự hài lòng)



- Nhận xét
  - Biểu đồ trên ta thấy không có xu hướng rõ ràng giữa thời gian bảo hành và mức độ hài lòng
  - Điểm trung bình dao động từ 4.96 đến 5.00, cho thấy sự khác biệt rất nhỏ giữa các mức thời gian bảo hành.
  - Biểu đồ đường cho thấy các điểm hài lòng dao động nhẹ ở các mức thời gian bảo hành khác nhau không có xu hướng răng hay giảm rõ ràng. Có thể là do thời gian bảo hành không phải là yếu tổ then chốt đến điểm hài lòng của khác hàng.

Vẽ biểu đồ hồi quy



- Nhận xét
  - Đường hồi quy gần như nằm ngang cho thấy thời gian bảo hành không có liên hệ tuyến tính đáng kể so với sự hài lòng
  - Các dữ liệu rải rác xung quanh mức đánh giá cao (gần 5), cho thấy sự ổn định của đánh giá khách hàng bất kể là thời gian bảo hành nào
  - Hệ số tương quan 0.011 rất gần 0. cũng cho thầy không có mối liện hệ giữa thời gian bảo hành và điểm đánh giá.

Kết luận: Thời gian bảo hành không có ảnh hưởng đáng kể đến sự hài lòng của khách hàng. Điều này cho thấy khách hàng đánh giá sự hài lòng dựa trên những yếu tố khác, chẳng hạn như chất lượng sản phẩm, trải nghiệm sử dụng, hoặc dịch vụ hỗ trợ, thay vì chỉ phụ thuộc vào thời gian bảo hành. Do đó, doanh nghiệp nên tập trung cải thiện các yếu tố này để nâng cao trải nghiệm khách hàng.

5.3 Dòng điện thoại nào được quan tâm nhiều nhất từ trước đến nay dựa trên số lượt đánh giá và hỏi đáp? (Top 10)

5.3.1 Mục đích

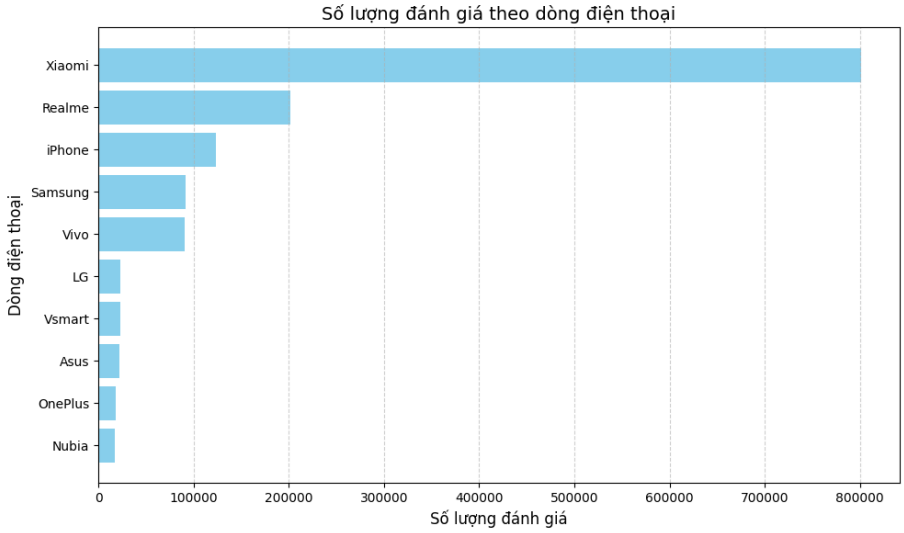
- Tìm hiểu xem trang web ‘Mobilecity‘ thu hút được đại đa số lượt tiếp cận sản phẩm của các dòng điện thoại nào.
- Giúp chúng ta nhận định được các dòng điện thoại thu hút sự quan tâm khách hàng, đưa ra được các chiến lược chăm sóc khách hàng hoặc các chiến dịch quảng cáo tập trung vào các dòng điện thoại nổi trội, từ đó nâng cao khả năng tăng doanh số.

5.3.2 Tiền xử lý

- Chọn cột hang_dien_thoai (hãng điện thoại) và so_luong_binh_luan (số lượng bình luận).
- Loại bỏ các giá trị thiếu (dropna).
- Nhóm dữ liệu theo hãng điện thoại, tính tổng số lượng bình luận cho từng hãng (sum).
- Sắp xếp theo số lượng bình luận giảm dần (sort_values) và lấy top 10 hãng được quan tâm nhiều nhất.

5.3.3 Trực quan hóa và nhận xét

Biểu đồ barh



- Nhận xét
  - 'Xiaomi' là dòng điện thoại có số lượt đánh giá cao nhất với khoảng 800.000 lượt, cao hơn gần 4 lần so với vị trí thứ 2 là 'Realme' (200.000 lượt).
  - Xếp theo sau là 'Iphone' với khoảng 120.000 lượt.
  - 'Samsung' và 'Vivo' khá ngang bằng nhau với khoảng 90.000 lượt.
  - Các dòng còn lại thu hút khá ít lượt đánh giá, khoảng 20.000 lượt cho các dòng 'LG', 'Vsmart', 'Asus', 'OnePlus', 'Nubia'.
  - 5 dòng điện thoại thu hút nhiều lượt đánh giá nhất là 'Xiaomi', 'Realme', 'Iphone', 'Samsung', 'Vivo'.

Kết luận: Dựa vào thông tin đánh giá của các dòng điện thoại này thì hệ thống cửa hàng sẽ biết được độ hài lòng của khách hàng với các dòng điện thoại hút khách này, từ đó có kế hoạch cải thiện và phát triển chất lượng tiếp thị khách hàng/bảo hành sản phẩm/quảng cáo/...Có thể tập trung nhiều hơn nữa những nguồn lực của cửa hàng để mở rộng khả năng tiếp cận của khách hàng đến với các dòng điện thoại này.

5.4 Dòng điện thoại nào được quan tâm nhiều nhất từ trước đến nay dựa trên số lượt đánh giá và hỏi đáp? (Top 10)

5.4.1 Mục đích

- Tìm hiểu xem trang web 'Mobilecity' thu hút được đại đa số lượt tiếp cận sản phẩm của các dòng điện thoại nào.
- Giúp chúng ta nhận định được các dòng điện thoại thu hút sự quan tâm khách hàng, đưa ra được các chiến lược chăm sóc khách hàng hoặc các chiến dịch quảng cáo tập trung vào các dòng điện thoại nổi trội, từ đó nâng cao khả năng tăng doanh số.

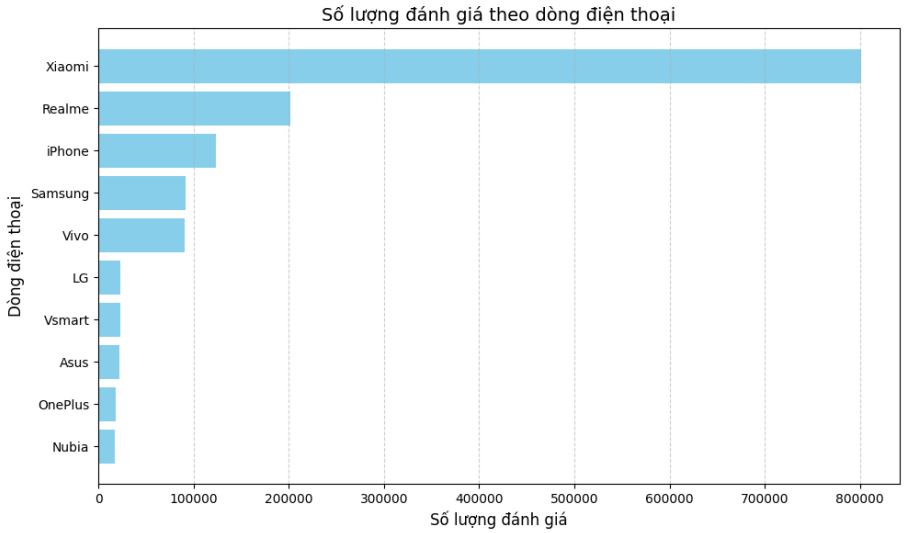
5.4.2 Tiền xử lý

- Chọn cột hang_dien_thoai (hãng điện thoại) và so_luong_binh_luan (số lượng bình luận).
- Loại bỏ các giá trị thiếu (dropna).
- Nhóm dữ liệu theo hãng điện thoại, tính tổng số lượng bình luận cho từng hãng (sum).

- Sắp xếp theo số lượng bình luận giảm dần (sort_values) và lấy top 10 hãng được quan tâm nhiều nhất.

5.4.3 Trực quan hóa và nhận xét

Biểu đồ barh



- Nhận xét
  - 'Xiaomi' là dòng điện thoại có số lượt đánh giá cao nhất với khoảng 800.000 lượt, cao hơn gần 4 lần so với vị trí thứ 2 là 'Realme' (200.000 lượt).
  - Xếp theo sau là 'Iphone' với khoảng 120.000 lượt.
  - 'Samsung' và 'Vivo' khá ngang bằng nhau với khoảng 90.000 lượt.
  - Các dòng còn lại thu hút khá ít lượt đánh giá, khoảng 20.000 lượt cho các dòng 'LG', 'Vsmart', 'Asus', 'OnePlus', 'Nubia'.
  - 5 dòng điện thoại thu hút nhiều lượt đánh giá nhất là 'Xiaomi', 'Realme', 'Iphone', 'Samsung', 'Vivo'.

Kết luận: Dựa vào thông tin đánh giá của các dòng điện thoại này thì hệ thống cửa hàng sẽ biết được độ hài lòng của khách hàng với các dòng điện thoại hút khách này, từ đó có kế hoạch cải thiện và phát triển chất lượng tiếp thị khách hàng/bảo hành sản phẩm/quảng cáo/...Có thể tập trung nhiều hơn nữa những nguồn lực của cửa hàng để mở rộng khả năng tiếp cận của khách hàng đến với các dòng điện thoại này.

5.5 Trong các thông số kỹ thuật của một chiếc điện thoại, những thông số nào thường có ảnh hưởng lớn nhất đến giá bán?

5.5.1 Mục đích

- **Thông số kỹ thuật** gồm những thông tin chính: màn hình (kích thước, tần số quét, loại màn hình), camera, chip xử lý (CPU), RAM và bộ nhớ, Pin và sạc, hệ điều hành, Thiết kế và chất liệu, ...
- Các mục đích chính của việc trả lời câu hỏi:
  - Đối với nhà sản xuất: Phân tích các yếu tố ảnh hưởng đến giá giúp họ định giá sản phẩm hợp lý, tối ưu hóa thiết kế để tăng sức cạnh tranh trên thị trường.
  - Đối với khách hàng: Khi hiểu rõ nhu cầu của mình, khách hàng có thể xác định được chiếc điện thoại mong muốn thuộc khoảng giá nào và đánh giá xem số tiền bỏ ra có tương xứng với các tính năng, thông số kỹ thuật của sản phẩm hay không.

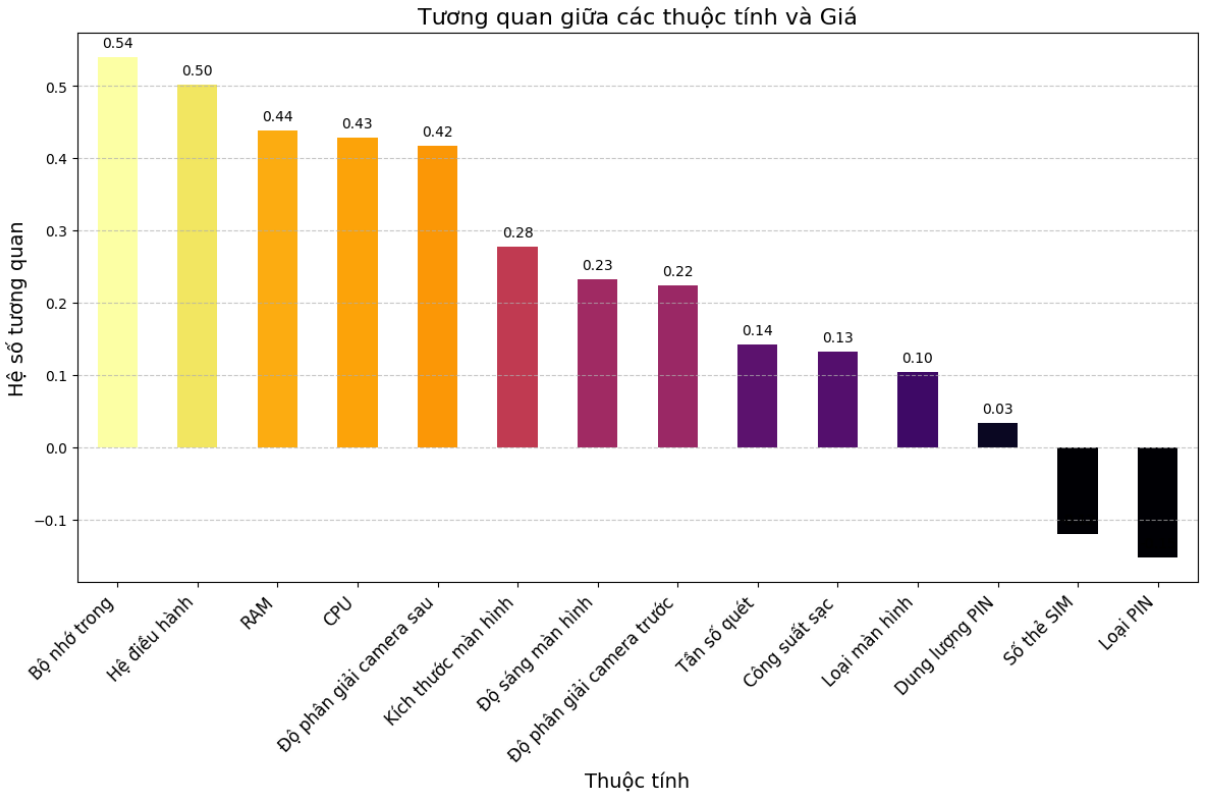
– Đối với tư vấn viên: Việc nắm rõ thông số kỹ thuật nào ảnh hưởng nhiều đến giá giúp tư vấn viên đưa ra lời khuyên chính xác, giúp khách hàng chọn được sản phẩm phù hợp nhất với yêu cầu của họ.

5.5.2 Tiền xử lý

- Tạo DataFrame mới: Tạo một bản sao từ dữ liệu gốc để xử lý, loại bỏ các giá trị thiếu (gia_moi) và chọn các thông số kỹ thuật quan trọng.
- Số hóa thuộc tính:
  - Hệ điều hành: Chuyển đổi phiên bản thành số thực và gán trọng số cho Android và iOS.
  - CPU: Sử dụng tiến trình sản xuất (nm) và tốc độ xử lý (GHz) để đánh giá hiệu suất CPU.
  - Loại màn hình: Gán giá trị dựa trên tần suất xuất hiện (AMOLED > OLED > LCD).
  - Loại PIN: Gán giá trị ưu tiên (Si/C > Li-Po > Li-Ion).
  - Camera: Tổng độ phân giải của camera trước và sau, cộng thêm giá trị bổ sung dựa trên số lượng camera.
  - Thiết kế: Loại bỏ vì không thể số hóa chính xác.
- Xử lý giá trị thiếu:
  - Điền giá trị thiếu bằng mean cho hầu hết các cột.
  - Với Số thẻ SIM, điền bằng giá trị nhỏ nhất để đảm bảo tính hợp lý.

5.5.3 Trực quan hóa và nhận xét

Trực quan hóa bằng biểu đồ cột



- **Nhận xét một số yếu tố trong thiết kế ảnh hưởng đến giá của một chiếc điện thoại:**

  - **Chất liệu khung và mặt lưng**:
    * **Khung kim loại** (nhôm, titan, thép) và **mặt lưng kính/gốm/da** cao cấp hơn nhựa, góp phần tăng giá trị sản phẩm.
    * Kính **Gorilla Victus** vượt trội hơn các loại kính thông thường, nâng cao giá trị và độ bền.
  - **Khả năng kháng nước/bụi**:
    * Các chuẩn **IP68/IP69** đảm bảo khả năng kháng nước/bụi vượt trội hơn so với chuẩn **IP53/IP54**, làm tăng giá sản phẩm.
  - **Màn hình**:
    * Công nghệ màn hình: **AMOLED** $>$ **OLED** $>$ **LCD**.
    * Thiết kế màn hình cong và sử dụng kính cường lực cao cấp làm tăng giá trị thẩm mỹ và sử dụng.
  - **Tính năng bổ trợ**:
    * Hỗ trợ bút cảm ứng, trigger gaming, đèn LED RGB, hoặc các thiết kế đặc biệt như **điện thoại gập** hoặc **gaming phone**.
  - **Thiết kế**:
    * Thiết kế nguyên khối, hợp tác với các thương hiệu cao cấp, hoặc đạt tiêu chuẩn quân đội nâng cao giá trị sản phẩm.

- **Nhận xét**
  - Các thông số kỹ thuật có tương quan mạnh nhất với giá bán:
    * **Bộ nhớ trong**: Đáp ứng nhu cầu lưu trữ dữ liệu ngày càng lớn.
    * **Hệ điều hành**: Ảnh hưởng bởi phiên bản và mức độ tối ưu hóa.
    * **RAM**: Tăng khả năng xử lý đa nhiệm.
    * **CPU**: Quyết định hiệu năng, đặc biệt ở các dòng flagship.
  - Các thông số kỹ thuật có tương quan trung bình với giá:
    * **Camera sau**: Quan trọng với người dùng yêu thích chụp ảnh.
    * **Kích thước màn hình**: Cải thiện trải nghiệm giải trí.
  - Các thông số kỹ thuật có ảnh hưởng nhỏ nhưng vẫn đáng kể:
    * **Công suất sạc**: Phổ biến với tốc độ sạc nhanh.
    * **Loại màn hình và Tần số quét**: Quan trọng với các dòng điện thoại gaming hoặc giải trí cao cấp.
  - Các thông số kỹ thuật có tương quan thấp với giá:
    * **Dung lượng PIN, Số thẻ SIM, và Loại PIN**: Không có sự khác biệt lớn giữa các phân khúc giá.

Kết luận: Những thông số ảnh hưởng mạnh nhất đến giá của một chiếc điện thoại là **Bộ nhớ trong, Hệ điều hành, RAM, và CPU**. Các yếu tố khác như **camera sau và màn hình** đóng vai trò bổ trợ, trong khi các yếu tố như **PIN, SIM** có ảnh hưởng nhỏ hơn.

5.6 Kiểu thiết kế điện thoại nào phổ biến nhất hiện nay, dựa trên các kiểu thiết kế của các mẫu điện thoại hiện có trong cửa hàng?

5.6.1 Mục đích

- Lý do: Các mẫu thiết kế xuất hiện phổ biến trên các điện thoại phần nào phản ánh được thị hiếu và nhu cầu của người mua. Theo quy luật cung - cầu, việc phân tích các mẫu thiết kế thịnh hành mang lại nhiều giá trị và ý nghĩa thực tiễn.

- Mục đích chính:

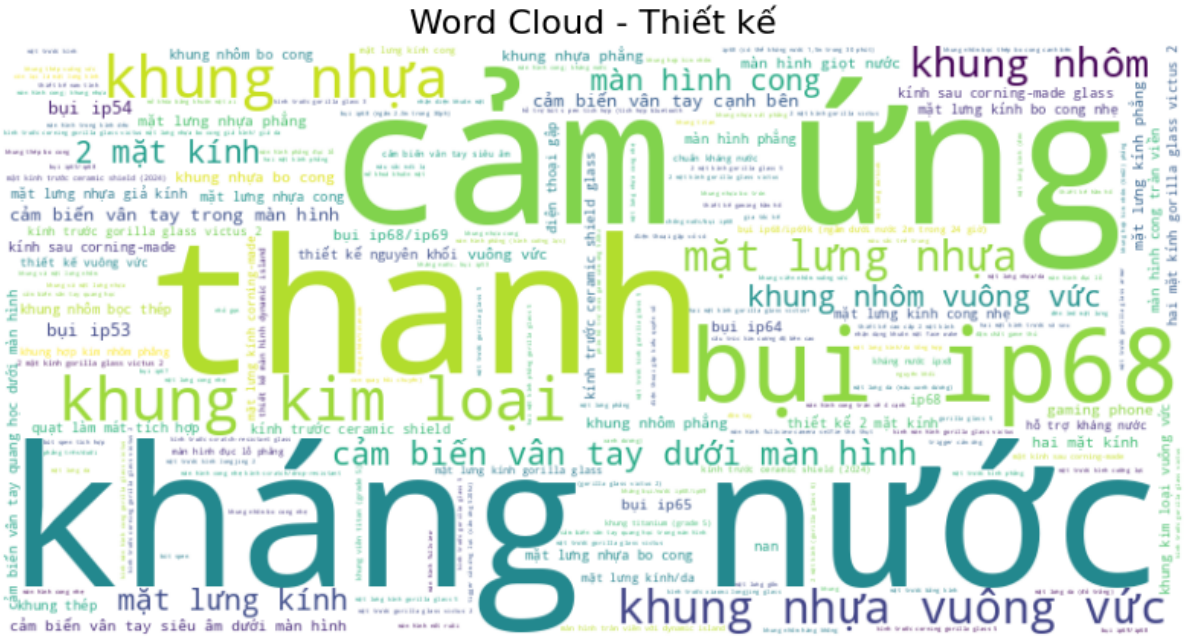
- Đối với nhà sản xuất: Hiểu được các kiểu thiết kế điện thoại nào đang được ưa chuộng, từ đó tối ưu hóa dây chuyền sản xuất và tập trung phát triển các sản phẩm phù hợp với thị hiếu thị trường.
- Đối với người tiêu dùng: Cung cấp thông tin về các kiểu thiết kế phổ biến, giúp họ lựa chọn sản phẩm đáp ứng nhu cầu cá nhân và bắt kịp xu hướng.
- Đối với các chủ cửa hàng: Nắm bắt xu hướng thị trường để điều chỉnh kế hoạch nhập hàng, quản lý tồn kho hiệu quả và xây dựng chiến lược kinh doanh phù hợp.
- Đối với các bên khác: Như nhà nghiên cứu thị trường, nhà đầu tư – thông tin này hỗ trợ việc phân tích xu hướng tiêu dùng, xây dựng chiến lược tiếp thị và đưa ra quyết định đầu tư chính xác.

5.6.2 Tiền xử lý

- Tạo DataFrame mới: Tạo design_df chứa cột Thiết kế từ dữ liệu gốc.
- Xử lý dữ liệu:
  - Loại bỏ giá trị thiếu và chuyển tất cả giá trị về dạng chuỗi.
  - Chuẩn hóa văn bản
- Sử dụng hàm explode để tách danh sách kiểu thiết kế thành các hàng riêng lẻ.
- Sử dụng value_counts() để đếm số lần xuất hiện của từng kiểu thiết kế.

5.6.3 Trực quan hóa và nhận xét

Biểu đồ wordcloud



- Nhận xét
  - Thiết kế phổ biến: Các yếu tố như kháng nước, chống bụi (IP68), và cảm ứng xuất hiện nổi bật, cho thấy đây là những đặc điểm được người dùng và các nhà sản xuất tập trung.
  - Vật liệu và khung: Khung nhựa, khung kim loại, và mặt lưng kính là các loại thiết kế phổ biến, phản ánh sự đa dạng trong chất liệu cấu thành điện thoại.

- Tính năng đặc biệt: Chuẩn IP68 xuất hiện thường xuyên hơn so với các chuẩn thấp hơn (như IP54), phản ánh xu hướng hướng tới khả năng bảo vệ cao hơn ở các dòng sản phẩm hiện đại.

Kết luận: Các yếu tố thiết kế được tập trung nhiều nhất hiện nay bao gồm tính năng kháng nước, chống bụi, chất liệu cao cấp như kim loại, kính, và các tiện ích hiện đại như cảm biến vân tay hoặc màn hình cong. Đây là các yếu tố quan trọng để tăng giá trị sản phẩm và thu hút khách hàng.

5.7 Hãng điện thoại có ảnh hưởng đến giá cả không?

5.7.1 Mục đích

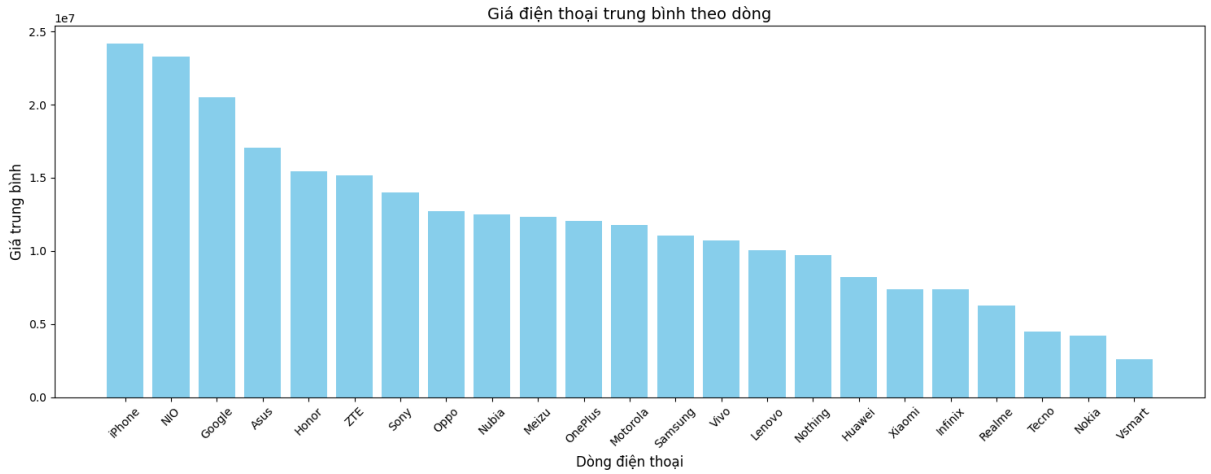
úp xác định vị trí của các thương hiệu trên thị trường.Phần nào xác định được chiến lược của thương hiệu trên thị trường (nhắm tới tới khách hàng có thu nhập thấp hay thu nhập cao trong xã hội).

5.7.2 Tiền xử lý

- Loại bỏ các mẫu điện thoại cũ bằng cách loại bỏ các dòng có từ khóa "cũ"trong cột ten.
- Lọc dữ liệu với các cột hang_dien_thoai và gia_moi, đồng thời loại bỏ các giá trị thiếu (dropna).
- Sử dụng groupby theo hang_dien_thoai để tính giá trị trung bình (mean) và số lượng mẫu (count) của từng hãng. Sắp xếp các hãng theo giá trị trung bình gia_moi giảm dần.
- Chỉ giữ lại các hãng có số lượng mẫu điện thoại lớn hơn 10 (count $>$ 10) để đảm bảo tính đại diện.
- DataFrame cuối cùng chứa tên hãng (hang_dien_thoai), giá trung bình (mean), và số lượng mẫu (count), sẵn sàng để phân tích mối quan hệ giữa hãng và giá cả.

5.7.3 Trực quan hóa và nhận xét

Biểu đồ bar



- Nhận xét
  - Điện thoại khác nhau có liên quan mật thiết tới giá cả, điều này cho thấy có sự phân khúc về giá điện thoại của các hãng trên thị trường tùy thuộc vào nhóm đối tượng khách hàng mà hãng đó hướng tới:
    * iPhone ưu tiên hướng tới những khách hàng có thu nhập khá cao. Điều này dễ hiểu khi họ được xem là hãng điện thoại đi trước thời đại, mang những công nghệ hiện đại nhất vào chiếc điện thoại của mình. Bên cạnh iPhone thì Google và NIO cũng có giá trung bình rất cao (ở mức hơn 20 triệu)

* Phần lớn hãng điện thoại sẽ hướng đến giá tầm trung, phù hợp hơn với người lao động thông thường, đây là nơi mà số lượng khách hàng tiềm năng tập trung chủ yếu. Một số hãng tiêu biểu: Samsung, Oppo, Vivo, Huawei.
* Số ít các hãng lựa chọn sẽ hướng đến khách hàng của mình là người có thu nhập thấp. Ví dụ Realme, Nokia, Vsmart.

Kết luận: Giá điện thoại phụ thuộc chặt chẽ vào hãng, phản ánh sự phân khúc khách hàng. iPhone, Google, và NIO phục vụ nhóm thu nhập cao với giá trên 20 triệu đồng. Samsung, Oppo, Vivo tập trung phân khúc tầm trung, trong khi Realme, Nokia, và Vsmart hướng đến khách hàng thu nhập thấp. Cửa hàng nên điều chỉnh danh mục sản phẩm và chiến lược tiếp thị phù hợp để đáp ứng nhu cầu từng nhóm, tối ưu hóa doanh số và duy trì cạnh tranh.

6 Data Modeling (Xây dựng mô hình)

Data Modeling (Xây dựng mô hình) là một bước quan trọng trong quy trình Data Science, nơi các mô hình toán học, thống kê hoặc học máy (machine learning) được sử dụng để trích xuất giá trị từ dữ liệu. Đây là quá trình kết nối dữ liệu với các thuật toán nhằm dự đoán, phân loại hoặc khám phá thông tin hữu ích.

6.1 Tiền xử lý dữ liệu cho các mô hình

Bước 1: Loại bỏ các cột không ảnh hưởng đến giá dựa vào kiến thức: loại bỏ cột **'ten', 'duong_dan', 'loai_dien_thoai', 'mau_sac', 'thiet_ke', 'cpu'**.

Bước 2: Loại bỏ các dòng trùng lặp: sau khi tính toán bằng cách **df.duplicated().sum()** ta thấy có 5548 dòng trùng lặp, ta sẽ thực hiện loại bỏ nó bằng cách **df.drop_duplicates()**.

Bước 3: Loại bỏ các cột dựa vào phân tích: Ta sẽ xem tỷ lệ dữ liệu thiếu ở của các thuộc tính hiện tại. Nhận thấy rằng cột **do_sang_man_hinh** và **gia_cu** có tỉ lệ các giá trị thiếu rất cao so với phần còn lại. Ta sẽ loại bỏ 2 cột này:

Bảng 10: *Tỉ lệ giá trị thiếu của các thuộc tính*

Features	**Missing ratio (%)**
gia_cu	43.3
do_sang_man_hinh	42.5
loai_pin	27.9
tan_so_quet	24.6
gia_moi	19.1
bo_nho_trong	7.4
cong_suat_sac	7.3
kich_thuoc_man_hinh	5.7
loai_man_hinh	5.4
dung_luong_pin	2.3
ram	0.5
thoi_gian_bao_hanh	0.0
do_phan_giai_cam_sau	0.0
so_the_sim	0.0
la_dien_thoai_cu	0.0
danh_gia	0.0
hang_dien_thoai	0.0
he_dieu_hanh	0.0
so_luong_binh_luan	0.0
do_phan_giai_cam_truoc	0.0

Bước 4: Loại bỏ các hàng tồn tại ít nhất một giá trị NaN: để giữ tính thực tế của dữ liệu, chúng em không điền giữ liệu thiếu mà sẽ xóa đi.

Bước 5: Xử lý các cột có dạng Non-numeric:

Bảng 11: *Thông tin Dtype của các cột*

Column	Dtype
he_dieu_hanh	object
hang_dien_thoai	object
loai_man_hinh	object
loai_pin	object
do_phan_giai_cam_sau	object
do_phan_giai_cam_truoc	object

Ta sẽ áp dung OneHotEncoder và LabelEncoder là hai phương pháp phổ biến để xử lý dữ liệu chuỗi (categorical data) trong học máy (machine learning).

- LabelEncoder
  - Biến đổi các giá trị chuỗi trong một cột thành các số nguyên (integer).
  - Được sử dụng khi có các giá trị với thứ bậc tự nhiên (ordinal data), ví dụ: ["low", "medium", "high"].
  - Dùng khi mô hình cần các biến số nguyên mà không yêu cầu xử lý đặc biệt (ví dụ: trong cây quyết định).
- OneHotEncoder
  - Biến đổi các giá trị chuỗi thành các cột nhị phân (binary columns) đại diện.
  - Dùng khi dữ liệu không có thứ bậc (nominal data).
  - Dùng khi sử dụng các mô hình phi tuyến hoặc yêu cầu phân tích dựa trên nhị phân

6.2 Tạo các mô hình

6.2.1 Model: XGBRegressor, DecisionTreeRegressor, RandomForestRegressor

(a) XGBRegressor

Lý do chọn:

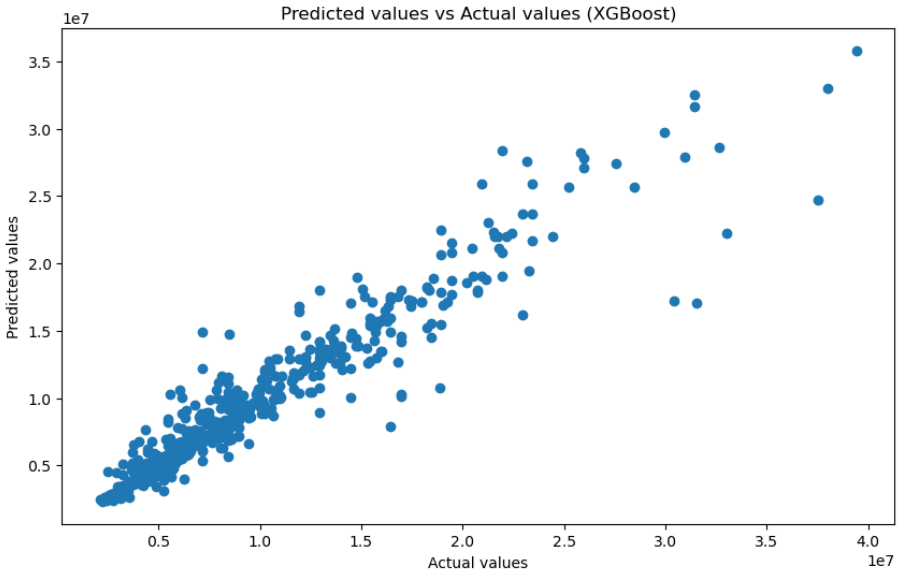
- Phù hợp với dữ liệu phức tạp: Bộ dữ liệu có **1,577 mẫu** và **50 đặc trưng** (11 số thực, 5 số nguyên, 34 nhị phân), đòi hỏi mô hình mạnh để xử lý tốt cả dữ liệu liên tục và nhị phân.
- Khả năng phi tuyến: Nắm bắt được mối quan hệ phức tạp giữa các đặc trưng như `ram`, `bo_nho_trong`, và `dung_luong_pin`.
- Chống overfitting: Với số lượng đặc trưng lớn, `XGBRegressor` sử dụng regularization và early stopping để kiểm soát độ phức tạp.
- Hiệu quả tính toán: Tối ưu với song song hóa và xử lý dữ liệu thưa, phù hợp với tập dữ liệu vừa phải.
- Dự đoán chính xác: Boosting giúp cải thiện lỗi, đặc biệt với bài toán dự đoán giá trị liên tục như giá điện thoại.

Đánh giá mô hình: Ta sẽ đánh gái mô hình dựa trên các chỉ số đánh giá mô hình.

Bảng 12: *Các chỉ số đánh giá mô hình*

Chỉ số	**Giá trị**
Mean Absolute Error (MAE)	1,165,279.62
Mean Squared Error (MSE)	4,029,285,920,587.60
R^2	0.90

Trực quan hóa: vẽ biểu đồ so sánh giá trị thực tế và giá trị dự đoán



(b) DecisionTreeRegressor

Lý do chọn:

- Dễ hiểu và trực quan:
  - Cấu trúc cây quyết định đơn giản, dễ giải thích và trực quan hóa, phù hợp khi cần giải thích các yếu tố ảnh hưởng đến giá điện thoại.
- Khả năng xử lý dữ liệu phức tạp:
  - Bộ dữ liệu có **1,577 mẫu** và **50 đặc trưng** (11 số thực, 5 số nguyên, 34 nhị phân). `DecisionTreeRegressor` có thể xử lý tốt cả đặc trưng số và nhị phân mà không cần chuẩn hóa hay chuyển đổi phức tạp.
- Mô hình phi tuyến:
  - Cây quyết định có khả năng nắm bắt các mối quan hệ phi tuyến giữa các đặc trưng, ví dụ: sự kết hợp giữa `ram`, `bo_nho_trong`, và `dung_luong_pin` đến giá.

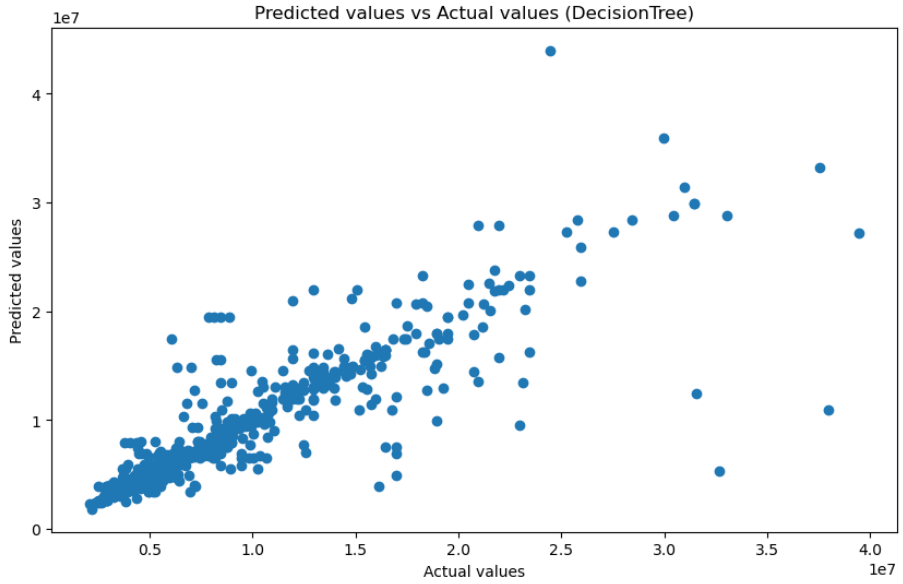
- Không cần giả định dữ liệu:
  - Không đòi hỏi giả định về phân phối dữ liệu hoặc mối quan hệ tuyến tính giữa các đặc trưng, phù hợp với tập dữ liệu thực tế có thể không tuyến tính.
- Chống ảnh hưởng bởi đặc trưng không liên quan:
  - Tự động chọn lọc các đặc trưng quan trọng (như `kich_thuoc_man_hinh`, `tan_so_quet`,...), giảm ảnh hưởng từ những đặc trưng ít liên quan đến giá.
- Hiệu quả tính toán với tập dữ liệu nhỏ:
  - Với 1,577 mẫu, `DecisionTreeRegressor` có thời gian huấn luyện nhanh, không đòi hỏi tài nguyên tính toán lớn.

Đánh giá mô hình:

Bảng 13: *Các chỉ số đánh giá mô hình*

Chỉ số	**Giá trị**
Mean Absolute Error (MAE)	1,611,120.41
Mean Squared Error (MSE)	11,261,320,106,041.40
R^2	0.73

Trực quan hóa: vẽ biểu đồ so sánh giá trị thực tế và giá trị dự đoán



(c) RandomForestRegressor

Lý do chọn:

- Hiệu quả trên dữ liệu phức tạp:
  - Bộ dữ liệu có **1,577 mẫu** và **50 đặc trưng** (11 số thực, 5 số nguyên, 34 nhị phân). `RandomForestRegressor` có thể xử lý tốt cả đặc trưng liên tục và nhị phân.
- Khả năng phi tuyến cao:
  - `RandomForestRegressor` xây dựng nhiều cây quyết định và tổng hợp kết quả, cho phép nắm bắt tốt các mối quan hệ phi tuyến giữa các đặc trưng (như `ram`, `bo_nho_trong`, và `kich_thuoc_man_hinh`) đến giá.

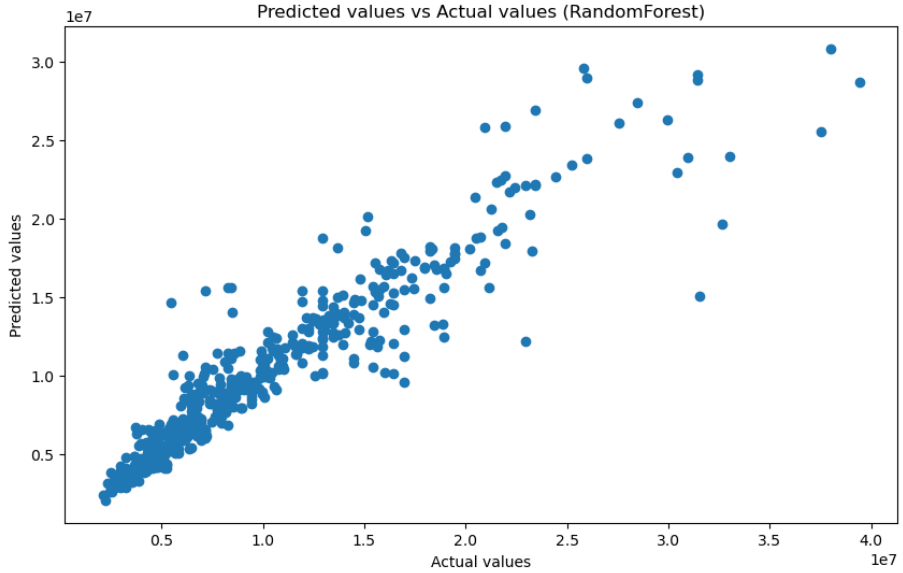
- Giảm overfitting:
  - Random Forest giảm overfitting so với `DecisionTreeRegressor` bằng cách sử dụng kỹ thuật bagging (xây dựng nhiều cây trên các tập dữ liệu con ngẫu nhiên) và chọn ngẫu nhiên một tập con các đặc trưng tại mỗi nút phân chia.
- Đánh giá tầm quan trọng của đặc trưng:
  - `RandomForestRegressor` cung cấp thông tin về mức độ quan trọng của từng đặc trưng (như `dung_luong_pin`, `tan_so_quet`,...), giúp hiểu rõ yếu tố nào ảnh hưởng nhiều nhất đến giá điện thoại.
- Hiệu suất tốt và khả năng mở rộng:
  - Hoạt động hiệu quả trên tập dữ liệu vừa phải, đồng thời dễ mở rộng lên tập dữ liệu lớn nhờ tính song song của thuật toán.
- Khả năng xử lý nhiễu và đặc trưng không liên quan:
  - Do sử dụng trung bình kết quả từ nhiều cây, mô hình giảm độ nhạy với nhiễu và đặc trưng không liên quan, đảm bảo dự đoán ổn định hơn.
- Không cần giả định dữ liệu:
  - Không yêu cầu giả định về phân phối dữ liệu hay mối quan hệ giữa các đặc trưng, phù hợp với tập dữ liệu thực tế.

Đánh giá mô hình:

Bảng 14: *Các chỉ số đánh giá mô hình*

Chỉ số	**Giá trị**
Mean Absolute Error (MAE)	1,316,223.53
Mean Squared Error (MSE)	5,176,328,243,890.02
R^2	0.87

Trực quan hóa: vẽ biểu đồ so sánh giá trị thực tế và giá trị dự đoán



6.2.2 Đánh giá và so sánh các mô hình: GBRegressor, DecisionTreeRegressor, RandomForestRegressor

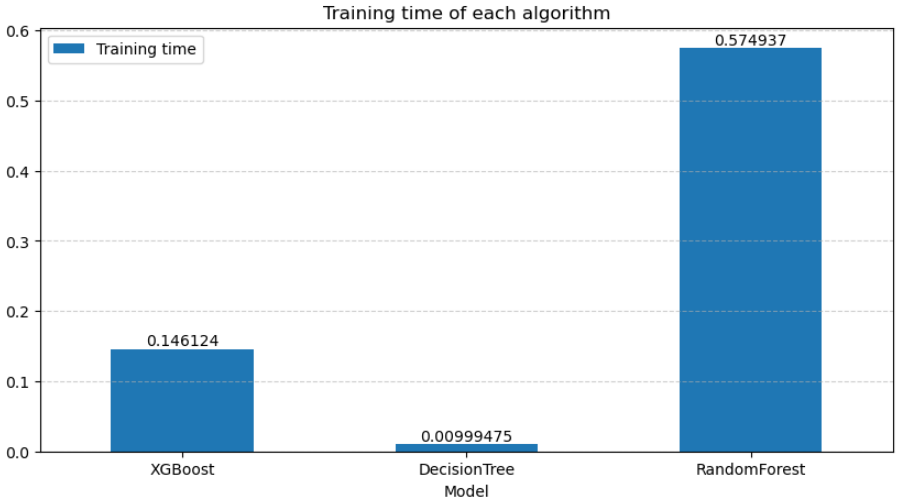
(a) Chỉ số đánh giá

Bảng 15: *So sánh các mô hình*

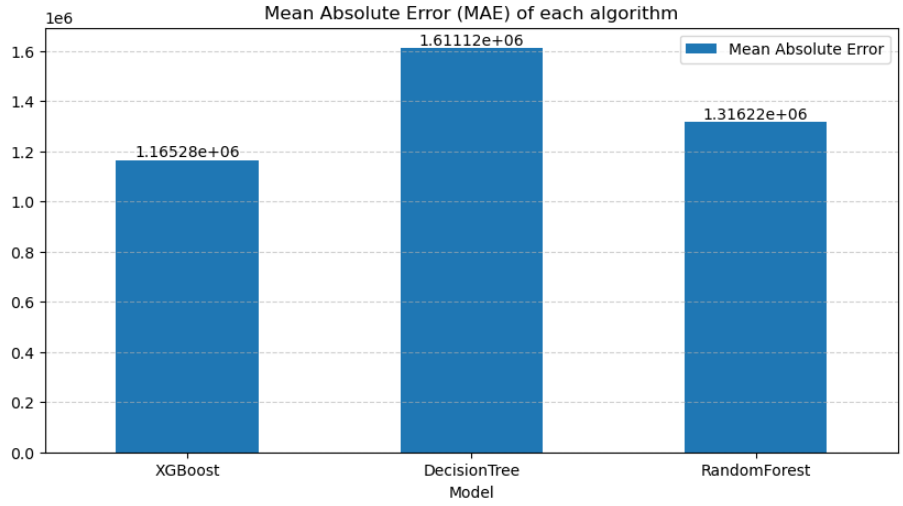
Model	Training Time (s)	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R^2
XGBoost	0.146124	1,165,280.00	4,029,286,000,000.00	0.902306
DecisionTree	0.009995	1,611,120.00	11,261,320,000,000.00	0.726957
RandomForest	0.574937	1,316,224.00	5,176,328,000,000.00	0.874494

(b) Trực quan hóa từng chỉ số đánh giá của các mô hình

Thời gian huấn luyện



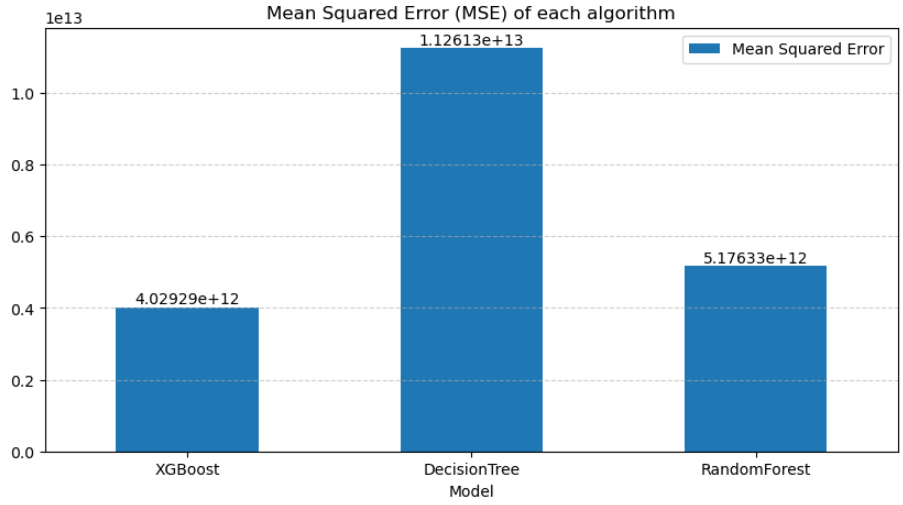
- Nhận xét
  - **DecisionTreeRegressor (0.01 giây)**:
    - Đây là mô hình có thời gian huấn luyện nhanh nhất, do chỉ xây dựng một cây quyết định. Điều này phù hợp với đặc điểm của Decision Tree, vốn là thuật toán đơn giản, hiệu quả và ít tốn tài nguyên tính toán.
    - Tuy nhiên, nhược điểm là dễ bị overfitting, đặc biệt với dữ liệu có nhiều đặc trưng và mối quan hệ phức tạp.
  - **XGBoost (0.146 giây)**:
    - XGBoost có thời gian huấn luyện dài hơn Decision Tree nhưng vẫn nhanh và tối ưu, nhờ khả năng song song hóa và tối ưu thuật toán Gradient Boosting.
    - Mặc dù thời gian huấn luyện cao hơn Decision Tree, XGBoost thường đạt hiệu suất tốt hơn, đặc biệt với bài toán hồi quy nhờ khả năng học từ lỗi của các mô hình trước đó.
  - **RandomForestRegressor (0.575 giây)**:
    - Random Forest có thời gian huấn luyện dài nhất do phải xây dựng và tổng hợp kết quả từ nhiều cây quyết định. Tuy nhiên, điều này mang lại tính ổn định và giảm overfitting so với Decision Tree.
    - Với bộ dữ liệu cỡ trung bình (1,577 mẫu), thời gian này vẫn chấp nhận được, nhưng đối với tập dữ liệu lớn hơn, có thể cần điều chỉnh số lượng cây hoặc tăng cường tài nguyên tính toán.



Mean Absolute Error (MAE)

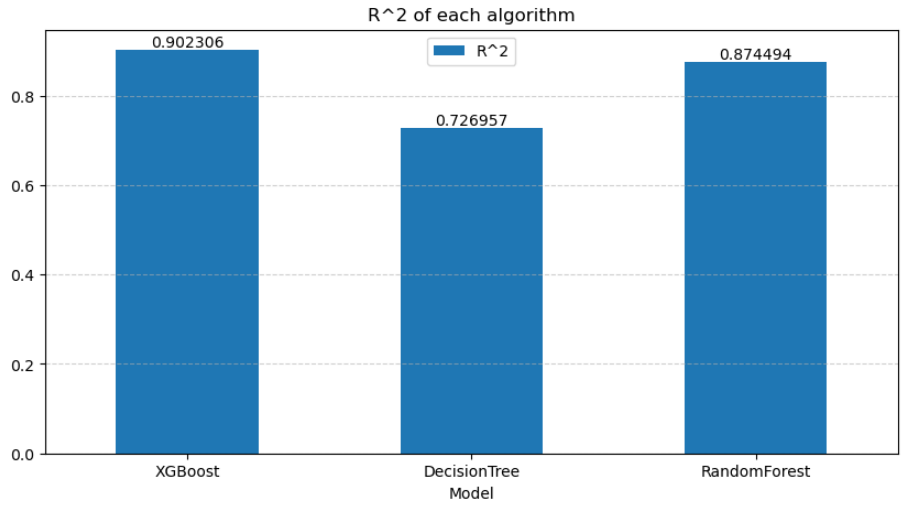
- Nhận xét:
  - **XGBoost (1.165 triệu)**:
    - Đây là mô hình có MAE thấp nhất, cho thấy XGBoost dự đoán gần đúng giá trị thực hơn so với các mô hình còn lại.
    - XGBoost thường đạt hiệu suất cao nhờ khả năng học từ lỗi của các mô hình trước đó trong quá trình boosting, đồng thời nắm bắt tốt các mối quan hệ phi tuyến giữa các đặc trưng.
  - **DecisionTreeRegressor (1.6 triệu)**:
    - Decision Tree có MAE cao nhất trong số các mô hình, cho thấy dự đoán ít chính xác hơn.
    - Nguyên nhân chính là Decision Tree dễ bị overfitting khi hoạt động độc lập, dẫn đến hiệu suất không ổn định, đặc biệt với dữ liệu kiểm tra.
  - **RandomForestRegressor (1.316 triệu)**:
    - Random Forest đạt MAE trung bình, tốt hơn Decision Tree nhưng kém XGBoost.
    - Việc tổng hợp nhiều cây giúp Random Forest ổn định hơn so với một Decision Tree đơn lẻ, nhưng hiệu quả vẫn thua XGBoost vì không tối ưu hóa lỗi giữa các cây như boosting.

Mean Squared Error (MSE)



- Nhận xét:
  - **XGBoost (4.03×10^{12})**:
    - **MSE thấp nhất**, cho thấy XGBoost có độ chính xác cao và dự đoán sát với giá trị thực hơn so với các mô hình khác.
    - XGBoost tối ưu hóa lỗi dựa trên boosting, nên thường vượt trội trong các bài toán hồi quy nhờ giảm thiểu lỗi dự đoán ở từng bước học.
  - **DecisionTreeRegressor (1.126×10^{13})**:
    - **MSE cao nhất**, gấp hơn 2 lần XGBoost, cho thấy Decision Tree có sai số lớn hơn đáng kể.
    - Nguyên nhân chính có thể là do Decision Tree dễ bị overfitting khi hoạt động độc lập, dẫn đến dự đoán kém chính xác trên tập kiểm tra.
  - **RandomForestRegressor (5.18×10^{12})**:
    - MSE của Random Forest cao hơn XGBoost nhưng thấp hơn Decision Tree, nhờ việc sử dụng nhiều cây quyết định và tổng hợp kết quả, giúp cải thiện độ chính xác và giảm sai số.
    - Tuy nhiên, Random Forest không tối ưu hóa lỗi tuần tự như XGBoost, nên hiệu suất vẫn kém hơn.

R2 Score



- Nhận xét:
  - **XGBoost (0.9023)**:
    - **R^2 cao nhất**, cho thấy XGBoost giải thích được 90.23% phương sai của biến mục tiêu `gia_moi` dựa trên các đặc trưng đầu vào.
    - Điều này thể hiện XGBoost là mô hình hiệu quả nhất trong việc dự đoán giá điện thoại, nhờ khả năng nắm bắt tốt các mối quan hệ phức tạp giữa các đặc trưng.
  - **DecisionTreeRegressor (0.727)**:
    - R^2 thấp nhất (72.7%), cho thấy mô hình chỉ giải thích được một phần phương sai của biến mục tiêu.
    - Điều này phản ánh hạn chế của Decision Tree khi sử dụng một cây quyết định duy nhất, dễ bị overfitting và kém chính xác trên tập kiểm tra.
  - **RandomForestRegressor (0.8745)**:

* R^2 cao thứ hai (87.45%), gần với XGBoost nhưng thấp hơn một chút.
* Random Forest cải thiện đáng kể so với Decision Tree bằng cách tổng hợp nhiều cây, giúp giảm overfitting và tăng khả năng giải thích phương sai của biến mục tiêu.

Kết luận

- **XGBoost**:
  - **Hiệu suất vượt trội** trên tất cả các chỉ số: MAE thấp nhất, MSE thấp nhất, và R^2 cao nhất.
  - **Ưu điểm**: Thời gian huấn luyện nhanh, khả năng dự đoán chính xác và phù hợp cho bài toán có nhiều đặc trưng và mối quan hệ phi tuyến phức tạp như dự đoán giá điện thoại.
  - **Nhược điểm**: Cần một chút tinh chỉnh để đạt hiệu suất tối ưu.
  - **Kết luận**: Là lựa chọn tốt nhất cho bài toán này, khi cân nhắc cả thời gian huấn luyện và độ chính xác.
- **Random Forest**:
  - **Hiệu suất trung bình**: MAE và MSE thấp hơn Decision Tree nhưng kém XGBoost, R^2 đạt gần với XGBoost.
  - **Ưu điểm**: Dự đoán ổn định nhờ tổng hợp nhiều cây quyết định, giảm thiểu overfitting so với Decision Tree.
  - **Nhược điểm**: Thời gian huấn luyện cao nhất, dễ bị chậm nếu dữ liệu lớn hoặc số cây quá nhiều.
  - **Kết luận**: Là lựa chọn tốt nếu cần mô hình dễ triển khai và ổn định, dù độ chính xác kém hơn XGBoost.
- **Decision Tree**:
  - **Hiệu suất thấp nhất**: MAE và MSE cao nhất, R^2 thấp nhất, cho thấy dự đoán kém chính xác và ít giải thích được phương sai của biến mục tiêu.
  - **Ưu điểm**: Thời gian huấn luyện cực nhanh, mô hình đơn giản và dễ hiểu, phù hợp với bài toán cần giải pháp nhanh chóng và dễ diễn giải.
  - **Nhược điểm**: Dễ bị overfitting, kém chính xác trên tập kiểm tra, đặc biệt với dữ liệu phức tạp.
  - **Kết luận**: Phù hợp nếu ưu tiên tốc độ huấn luyện hoặc cần mô hình đơn giản để giải thích, nhưng không phù hợp cho bài toán đòi hỏi độ chính xác cao.

Tổng kết

- **Ưu tiên chính xác và hiệu quả**: Chọn **XGBoost**.
- **Ưu tiên ổn định và dễ triển khai**: Chọn **Random Forest**.
- **Ưu tiên tốc độ và sự đơn giản**: Chọn **Decision Tree**.

Với yêu cầu dự đoán giá điện thoại, **XGBoost** là lựa chọn tối ưu nhất.

Phần III
Tổng kết

1. Khó khăn gặp phải

- Việc tìm nguồn dữ liệu hợp lý, mang tính thực tế cao, vẫn gặp phải nhiều khó khăn, vì dữ liệu thường bị giới hạn về mặt ý nghĩa hoặc không đầy đủ cho mục đích phân tích. Ngoài ra, kỹ năng cào dữ liệu còn cần được cải thiện.

- Xử lý dữ liệu phức tạp, đặc biệt là với các chuỗi dữ liệu, đòi hỏi phải chọn ra được các đặc trưng có ý nghĩa để phục vụ cho việc phân tích và xây dựng mô hình.
- Gặp khó khăn trong việc số hóa các kiểu dữ liệu phân loại (categorical data), đặc biệt là khi dữ liệu phân loại không có thứ tự rõ ràng hoặc có quá nhiều giá trị không phổ biến. Việc chuyển đổi các giá trị phân loại thành dạng số hợp lý là một thách thức lớn, đòi hỏi phải lựa chọn kỹ thuật thích hợp như one-hot encoding hoặc label encoding.

2. Bài học kinh nghiệm rút ra

- Cải thiện kỹ năng phân tích dữ liệu, từ đó có được cái nhìn sâu sắc hơn về các phương pháp phân tích khác nhau và cách khai thác ý nghĩa từ dữ liệu. Quá trình phân tích giúp rút ra những kết luận có giá trị dựa trên các mẫu dữ liệu thực tế.
- Bổ sung kỹ năng cào dữ liệu và chọn ra các thuộc tính có ý nghĩa cho quá trình phân tích, điều này giúp giảm độ phức tạp và làm tăng hiệu quả mô hình. Cào dữ liệu từ web hoặc sử dụng API giúp thu thập thông tin lớn từ các nguồn trực tuyến và cải thiện chất lượng dữ liệu.
- Sử dụng một cách đa dạng và kết hợp nhiều thư viện bổ trợ như ‘pandas‘ (xử lý dữ liệu), ‘matplotlib‘/‘seaborn‘ (vẽ đồ thị), ‘scikit-learn‘ (máy học), và ‘tensorflow‘/‘keras‘ (deep learning) để tối ưu hóa quá trình phân tích và mô hình hóa dữ liệu.
- Không nên đưa code vào slides khi thuyết trình.
- Chuẩn hóa dữ liệu cho mô hình: cần chú trọng về tính tối ưu hơn và các bước xử lý dữ liệu trước khi đưa vào mô hình.
- Slides nên đánh số trang rõ ràng.
- Nên có kết luận:
  - Câu hỏi xong thì kết luận gì?
  - Mô hình xong thì kết luận gì?

3. Nếu có thêm thời gian

- Có thể suy nghĩ thêm về các câu hỏi phân tích có thể trả lời được bằng dữ liệu, ví dụ như tìm ra các yếu tố ảnh hưởng đến một biến mục tiêu cụ thể hoặc các dự báo về tương lai dựa trên dữ liệu hiện tại.
- Nếu có thêm thời gian, có thể xây dựng mô hình gợi ý sản phẩm dựa trên nhu cầu của người mua, điều này sẽ giúp cải thiện trải nghiệm người dùng và tối ưu hóa các đề xuất trong các hệ thống thương mại điện tử.

4. Không gian làm việc

- Github, Canva, Google meet, Zalo, Google doc...vv

Tài liệu tham khảo

[1] Trang web *MobileCity*
URL: https://mobilecity.vn/dien-thoai

[2] Thư viện pandas.
URL: https://pandas.pydata.org/docs/.

[3] Thư viện Seaborn.
URL: https://seaborn.pydata.org/.

[4] Thư viện Matplotlib.
URL: https://matplotlib.org/.

[5] Tutorial Data Science.
URL: https://github.com/academic/awesome-datascience.

[6] Thư viện hỗ trợ học máy.
URL: https://scikit-learn.org/stable/.

[7] Top 50 đồ thị.
URL: https://www.machinelearningplus.com/plots/top-50-matplotlib-visualizations-the-master-pl

[8] Khóa học Python.
URL: https://www.udemy.com/course/python-for-data-science-and-machine-learning-bootcamp/?couponCode=KEEPLEARNING.