

BÁO CÁO CUỐI KÌ MÔN KHDL

Dự đoán giá điện thoại cũ

Nhóm 6

THÀNH VIÊN:

- Võ Hoàng Bảo
- Trương Bảo Ngọc
- Nguyễn Hoàng Quân

Tổng Quan

01

Crawl Data

Mô tả dữ liệu

02

Trích xuất đặc trưng

03

Mô hình hóa dữ liệu

I.Thu tập dữ liệu

cÔNG CỤ

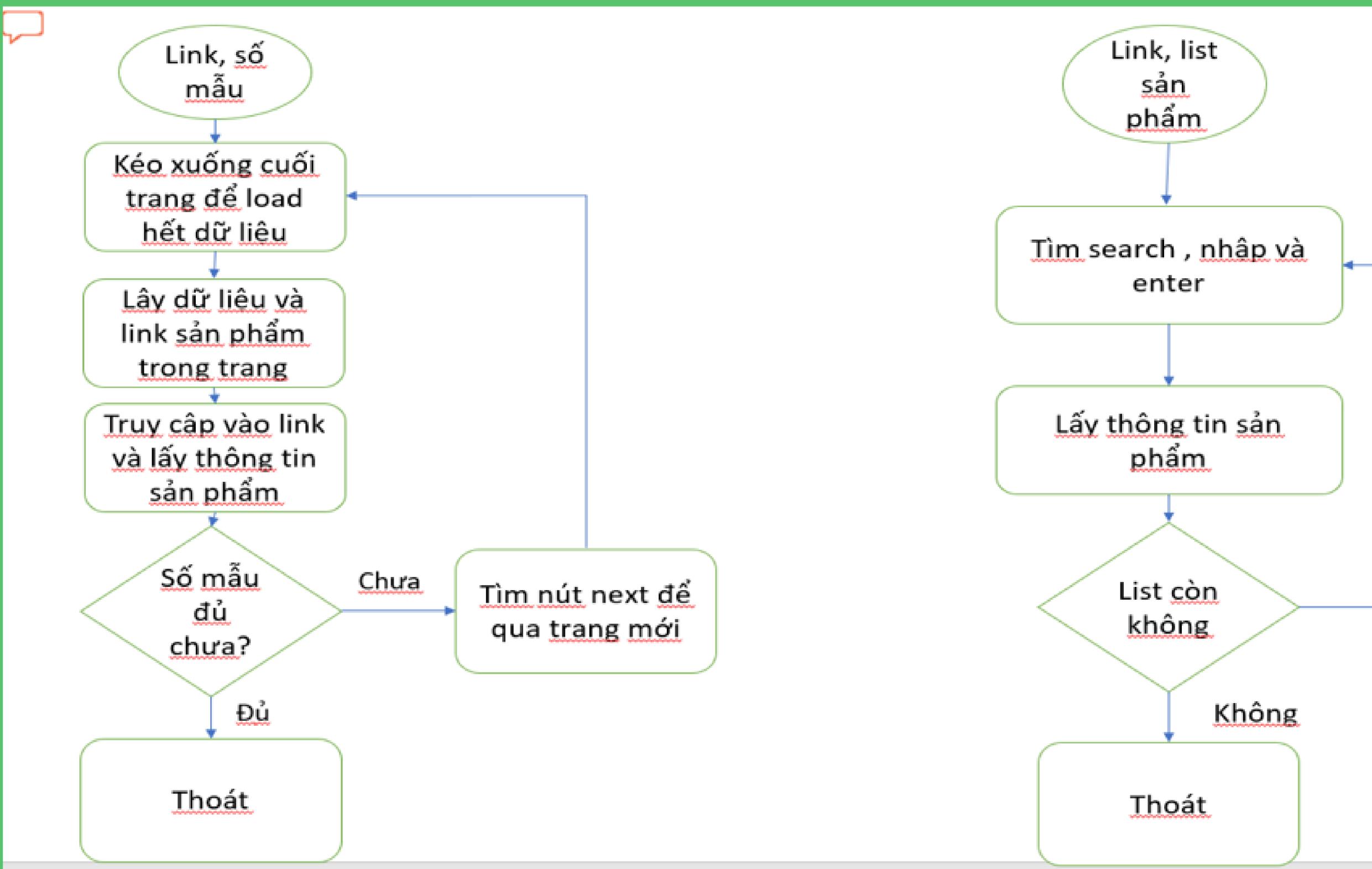
Selenium

- Điều khiển và tương tác với trình duyệt web
- Mở trình duyệt, điều hướng, điền biểu mẫu, nhấp chuột,..
- Thu nhập dữ liệu

BeautifulSoup

- Phân tích cú pháp HTML và XML
- Dùng các phương thức lấy thông tin từ các phân tử HTML

Sơ đồ mô tả quá trình thu nhập dữ liệu



Các bước thực hiện

Chợ Tốt | Nhà Tốt | Chợ Tốt Xe | Việc Làm Tốt

chợTỐT

Quản lý tin Đơn Hàng C

Điện thoại Samsung

Phú - 47 Phút Trước - Tp Hồ Chí Minh

xác sam sung A20 lỗi cảm ứng như hình
550.000 đ

Cửa Hàng Thành V... - 49 Phút Trước - Tp Hồ Chí Minh

Samsung s10 8G - 128 GB
4.000.000 đ

Samsung s10 8G - 128 GB
4.000.000 đ

Zero Channel - 51 Phút Trước - Nam Định

samsung A41 xách tay nhật
1.990.000 đ

Quảng cáo bị đóng bởi Google

Quảng cáo bị đóng bởi Google

1 2 3 4 5 6 7 8 9 >

Samsung s10 8G - 128 GB

4.000.000 đ

Samsung s10 8G - 128 GB
Máy mới 99 %

Nhấn để hiển số: 082785 ***

Hãng: Samsung **Dòng máy:** Galaxy S10

Tình trạng: Đã sử dụng (chưa sửa chữa) **Màu sắc:** Trắng

Dung lượng: 8 GB **Xuất xứ:** Đang cập nhật

Chính sách bảo hành: Đang cập nhật **Loại bảo hành:** Bảo hành Nhà cung cấp

Khu Vực

Xã Việt Hùng, Huyện Trực Ninh, Nam Định

LẤY NHÀ MUA HÀNG NHANH

Data columns (total 9 columns)		
#	Column	Non-Null Cou
---	---	---
0	names	10 non-null
1	price	10 non-null
2	link	10 non-null
3	brand	10 non-null
4	type	10 non-null
5	condition	10 non-null
6	preserve	10 non-null
7	storage	10 non-null
8	result	10 non-null

dtypes: int64(1), object(8)

Input: <https://www.chotot.com/mua-ban-dien-thoai-apple-cu-sdmd1ec3>

Input: Link sản phẩm ở data bước 1

Demo 10 mẫu

Các bước thực hiện

DEVICES		apple iPhone 11					
Apple iPhone 11			Released 2019, September 20 194g, 8.3mm thickness iOS 13, up to iOS 16.4.1 64GB/128GB/256GB storage, no card slot	 6.1" 828x1792 pixels 2160p			
Apple iPhone 11 Pro Max							
Apple iPhone 11 Pro							
MORE DEVICE RESULTS		 REVIEW	 OPINIONS	 COM			
ALL VERSIONS		A2221	A2111	A2223			
Versions: A2221 (International); A2111 (USA, Canada, Puerto Rico)							
NETWORK		Technology		GSM / CDMA / HSPA / EVDO			
 LAUNCH		Announced		2019, September 10			
		Status		Available. Released 2019, Se			

Input: gsmarena.com
List_type

Thông tin chi tiết

(total 15 columns):

	Non-Null Count	Dty
	-----	---
10	non-null	obj
10	non-null	int
10	non-null	obj
age	10 non-null	obj
(1),	object(14)	

Mô tả dữ liệu

RangeIndex: 10350 entries, 0 to 10349

Data columns (total 15 columns):

#	Column	Non-Null Count	Dtype												
0	names	10350	non-null	object	names	0									
1	price	10350	non-null	object	price	0									
2	link	10350	non-null	object	link	0									
3	brand	10350	non-null	object	brand	0									
4	type	10350	non-null	object	type	0									
5	condition	10350	non-null	object	condition	0									
6	preserve	10350	non-null	object	preserve	0									
7	storage	10078	non-null	object	storage	272									
8	result	10350	non-null	int64	result	0									
9	display	10323	non-null	object	display	27									
10	camera	10323	non-null	object	camera	27									
11	pin	10323	non-null	object	pin	27									
12	tag	10323	non-null	object	tag	27									
13	date	10323	non-null	object	date	27									
14	ram_storage	10323	non-null	object	ram_storage	27									
dtypes: int64(1), object(14)															

Big data

Data columns (total 15 columns):

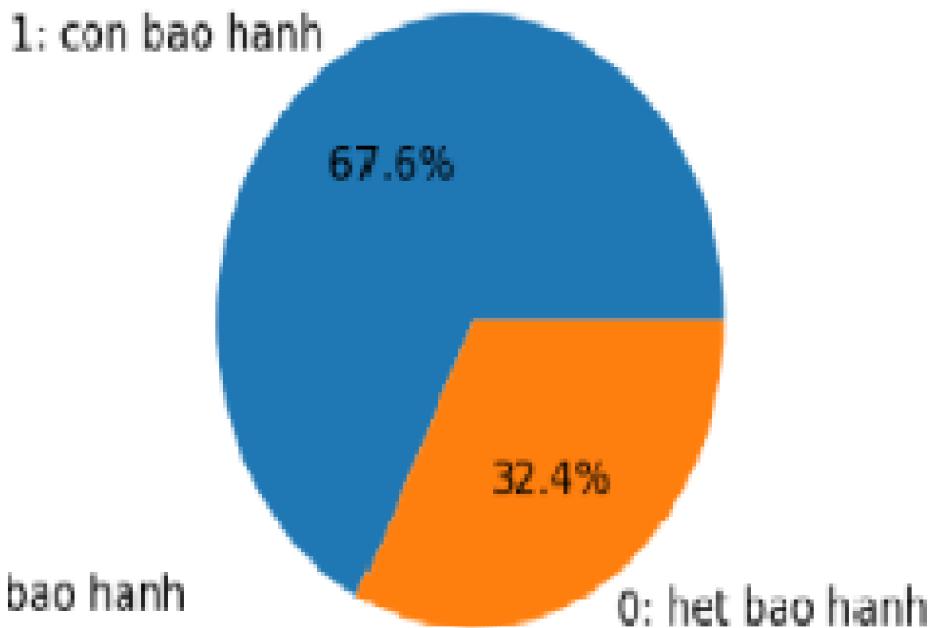
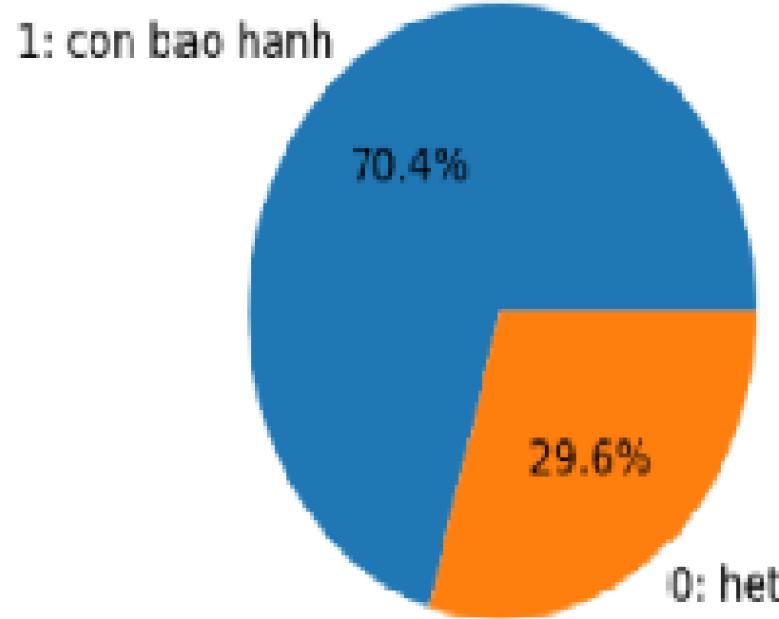
#	Column	Non-Null Count	Dtype												
0	names	1150	non-null	object	names	0									
1	price	1150	non-null	object	price	0									
2	link	1150	non-null	object	link	0									
3	brand	1150	non-null	object	brand	0									
4	type	1150	non-null	object	type	0									
5	condition	1150	non-null	object	condition	0									
6	preserve	1150	non-null	object	preserve	0									
7	storage	1127	non-null	object	storage	23									
8	result	1150	non-null	int64	result	0									
9	display	1145	non-null	object	display	5									
10	camera	1145	non-null	object	camera	5									
11	pin	1145	non-null	object	pin	5									
12	tag	1145	non-null	object	tag	5									
13	date	1145	non-null	object	date	5									
14	ram_storage	1145	non-null	object	ram_storage	5									
..															

Small data

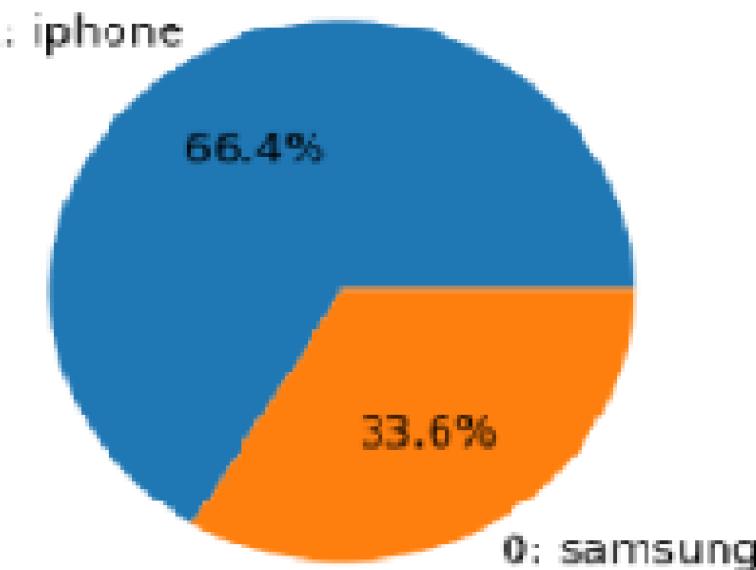
-Dữ liệu big data có 10350 mẫu và dữ liệu small data có 1150 mẫu bao gồm 13 đặc trưng

Mô tả dữ liệu

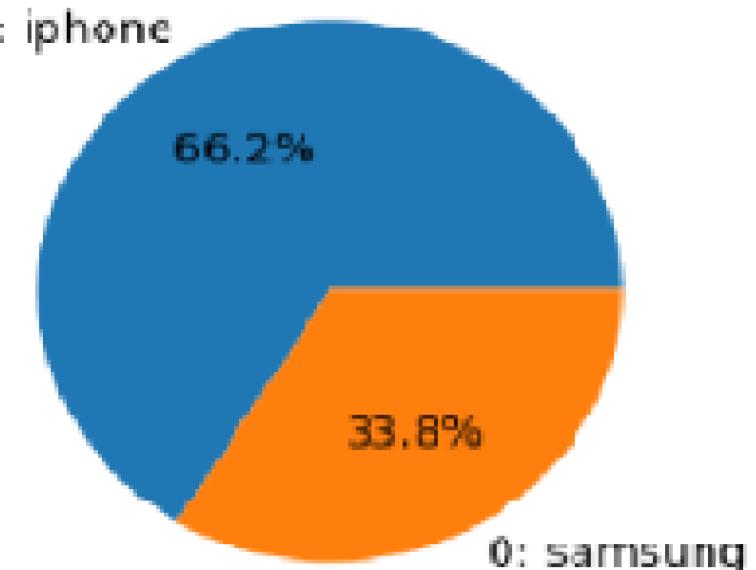
Distribution of preserve



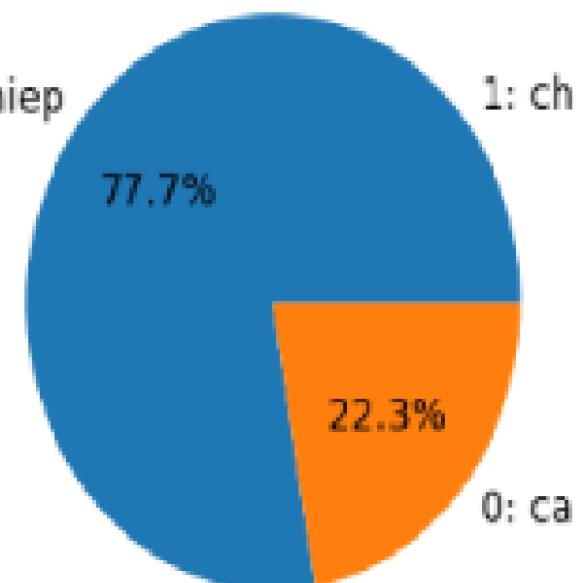
Distribution of brand



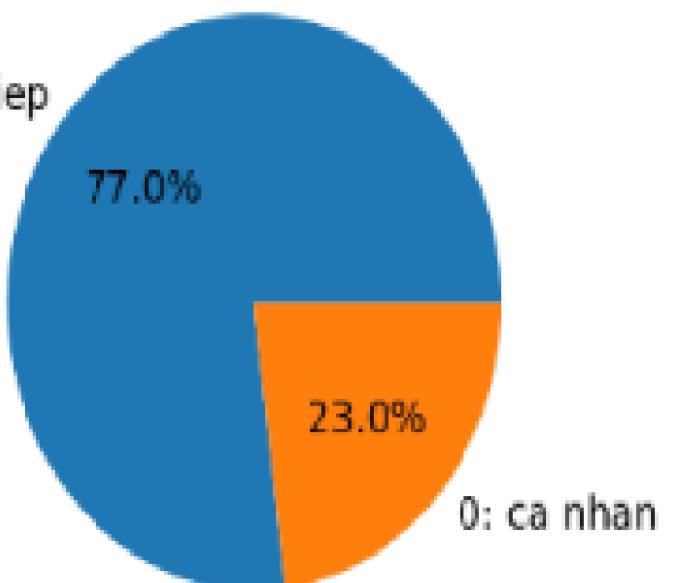
Distribution of brand



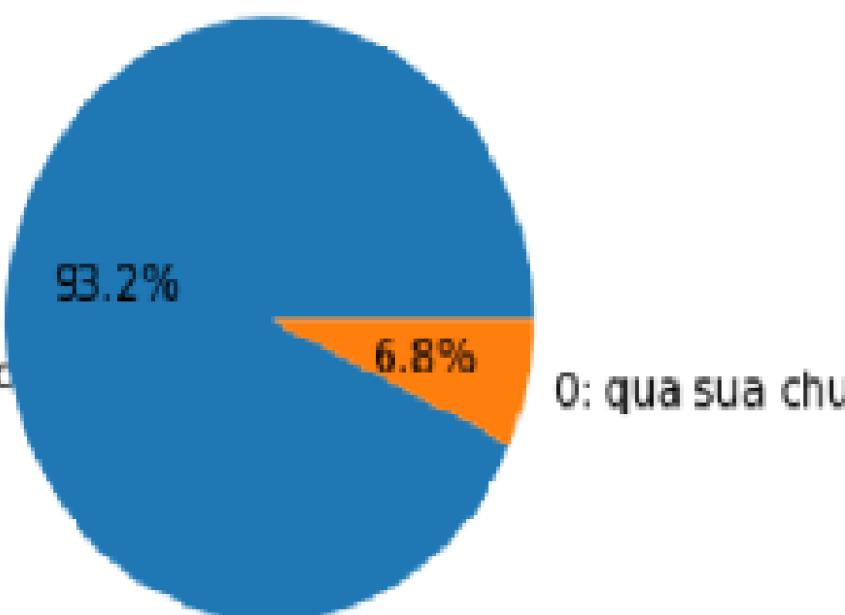
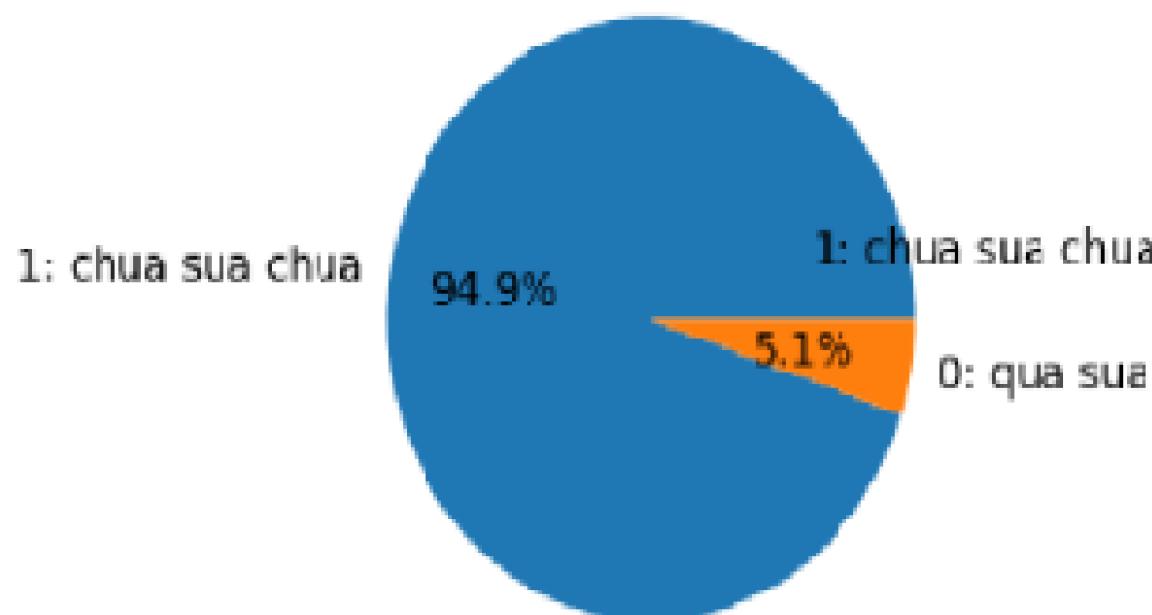
Distribution of result



Distribution of result

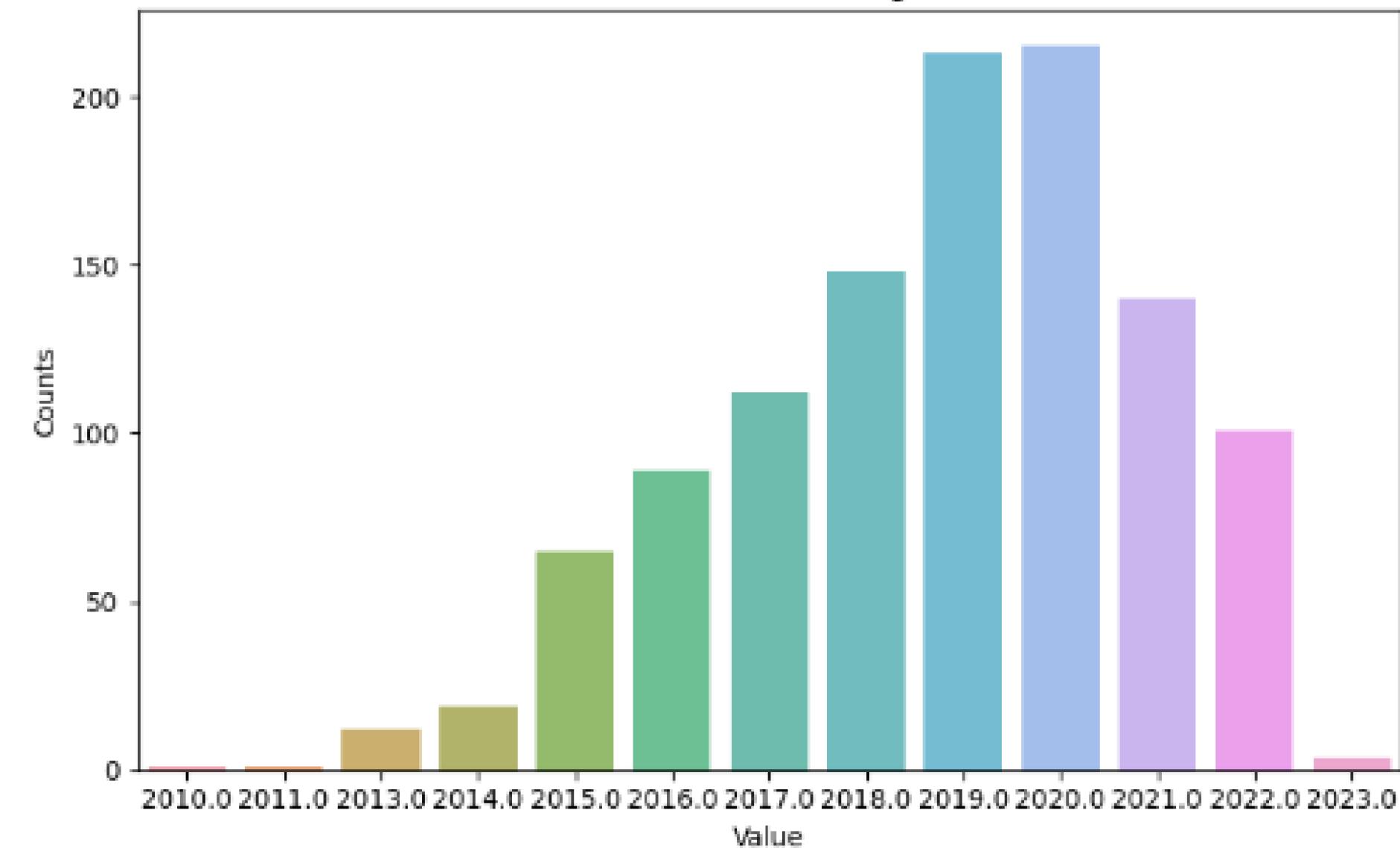


Distribution of condition

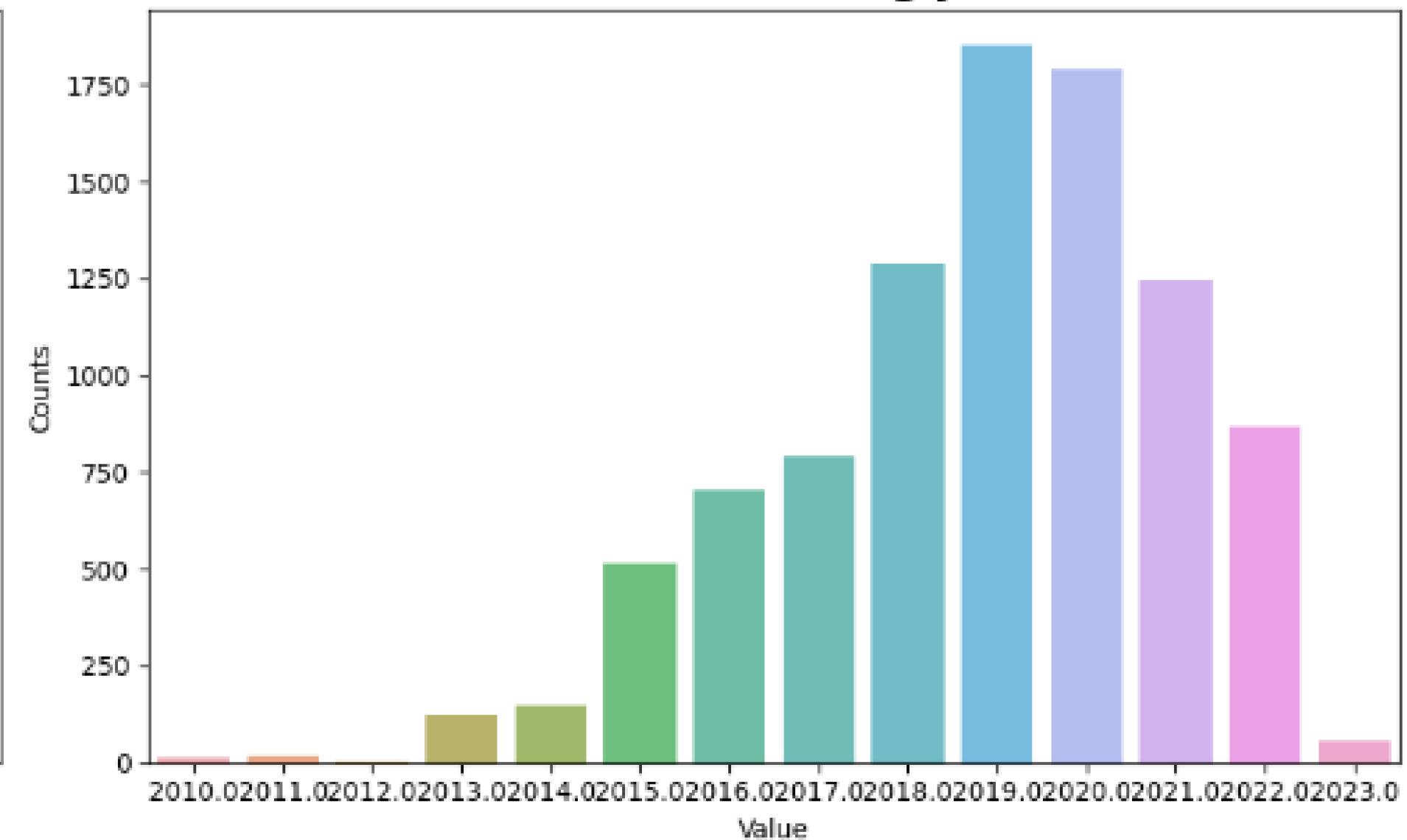


Mô tả dữ liệu

Distribution of year

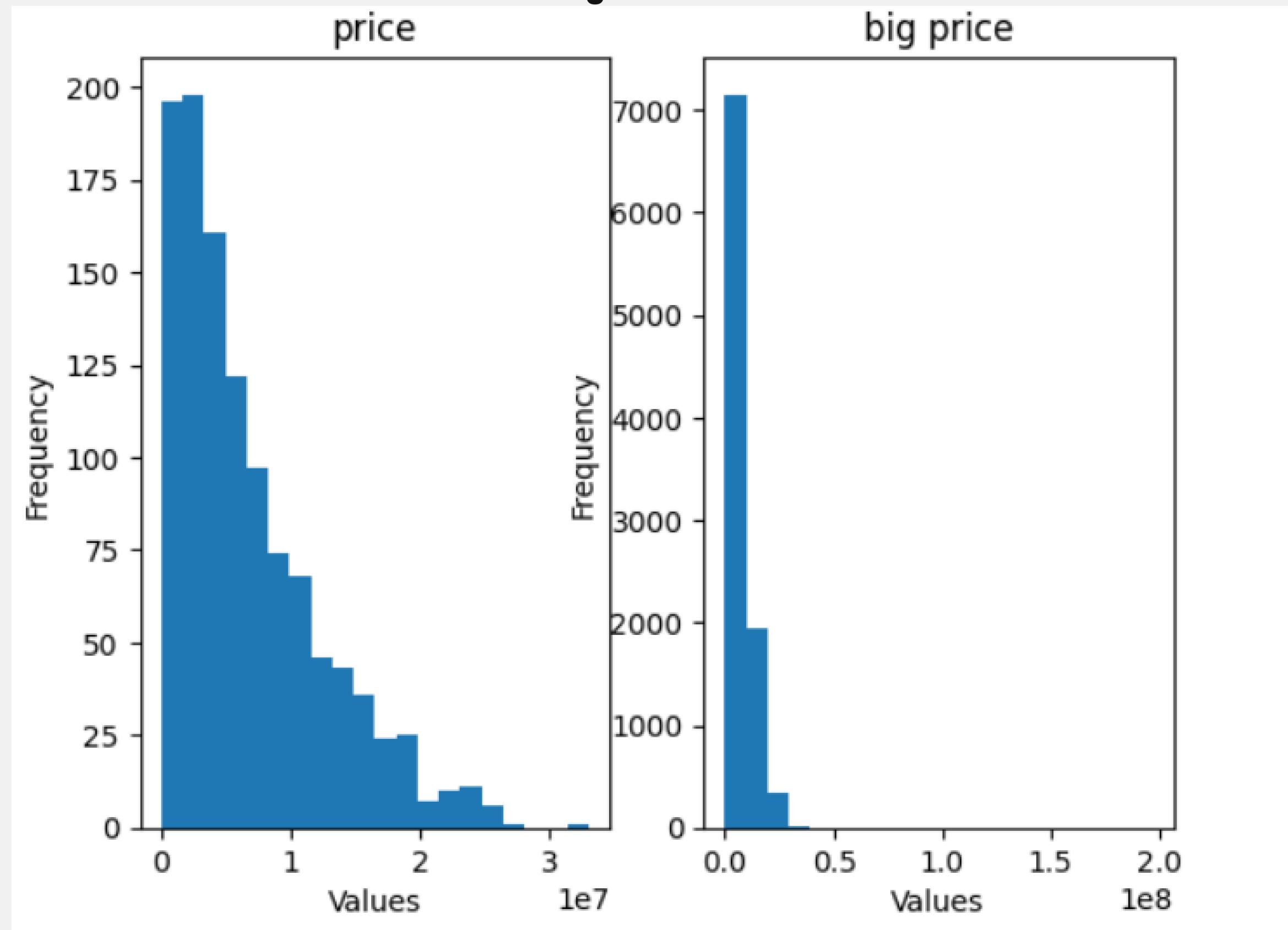


Distribution of big year



•Year: Dữ liệu thể hiện năm sản xuất đa số tập trung tại năm 2017 đến năm 2022, nhiều nhất là 2 năm 2019 và 2022.

Mô tả dữ liệu



• Price: Đa số dữ liệu đều nằm trong khoảng từ 1 đến 3 triệu. Ngoài ra bên dataset_10k có một số mẫu điện thoại có giá rất cao. Cao nhất trong tầm 20 triệu. Phân bố giảm dần theo chiều tăng của mức giá

II . Trích xuất đặc trưng



Làm sạch dữ liệu

	price	brand	type	condition	preserve	storage	result	display	camera	pin	date	ram_storage
0	9.900.000 đ	apple	iphone 12 pro	đã sử dụng (chưa sửa chữa)	đang cập nhật	128 gb	1	6.7"	12MP	3687mAh	Released 2020, November 13	128GB 6GB RAM, 256GB 6GB RAM, 512GB 6GB RAM
1	1500000	samsung	galaxy a30	đã sử dụng (chưa sửa chữa)	hết bảo hành	64.0 gb	0	6.4"	16MP	4000mAh	Released 2019, March	32GB 3GB RAM, 64GB 3GB RAM, 64GB 4GB RAM
2	8.500.000 đ	apple	iphone 11 pro max	đã sử dụng (chưa sửa chữa)	hết bảo hành	64.0 gb	0	6.5"	12MP	3969mAh	Released 2019, September 20	64GB 4GB RAM, 256GB 4GB RAM, 512GB 4GB RAM
3	11.000.000 đ	apple	iphone 12 pro	đã sử dụng (chưa sửa chữa)	đang cập nhật	128 gb	1	6.7"	12MP	3687mAh	Released 2020, November 13	128GB 6GB RAM, 256GB 6GB RAM, 512GB 6GB RAM
4	16.900.000 đ	samsung	galaxy z fold3	đã sử dụng (chưa sửa chữa)	còn bảo hành	256.0 gb	1	7.6"	12MP	4400mAh	Released 2021, August 27	256GB 12GB RAM, 512GB 12GB RAM
5	1500000	apple	iphone 6s	đã sử dụng (chưa sửa chữa)	còn bảo hành	64.0 gb	0	4.7"	12MP	1715mAh	Released 2015, September 25	16GB 2GB RAM, 32GB 2GB RAM, 64GB 2GB RAM, 128G...
6	5.500.000 đ	apple	iphone xr	đã sử dụng (chưa sửa chữa)	1 tháng	64 gb	0	6.1"	12MP	2942mAh	Released 2018, October 26	64GB 3GB RAM, 128GB 3GB RAM, 256GB 3GB RAM
7	4.500.000 đ	apple	iphone xs	đã sử dụng (chưa sửa chữa)	1 tháng	64 gb	0	6.5"	12MP	3174mAh	Released 2018, September 21	64GB 4GB RAM, 256GB 4GB RAM, 512GB 4GB RAM
8	260.000 đ	apple	iphone 5	đã sử dụng (chưa sửa chữa)	đang cập nhật	16.0 gb	0	4.0"	8MP	1560mAh	Released 2013, September 20	16GB 1GB RAM, 32GB 1GB RAM, 64GB 1GB RAM
9	9200000	apple	iphone 11 pro max	đã sử dụng (chưa sửa chữa)	hết bảo hành	64.0 gb	0	6.5"	12MP	3969mAh	Released 2019, September 20	64GB 4GB RAM, 256GB 4GB RAM, 512GB 4GB RAM

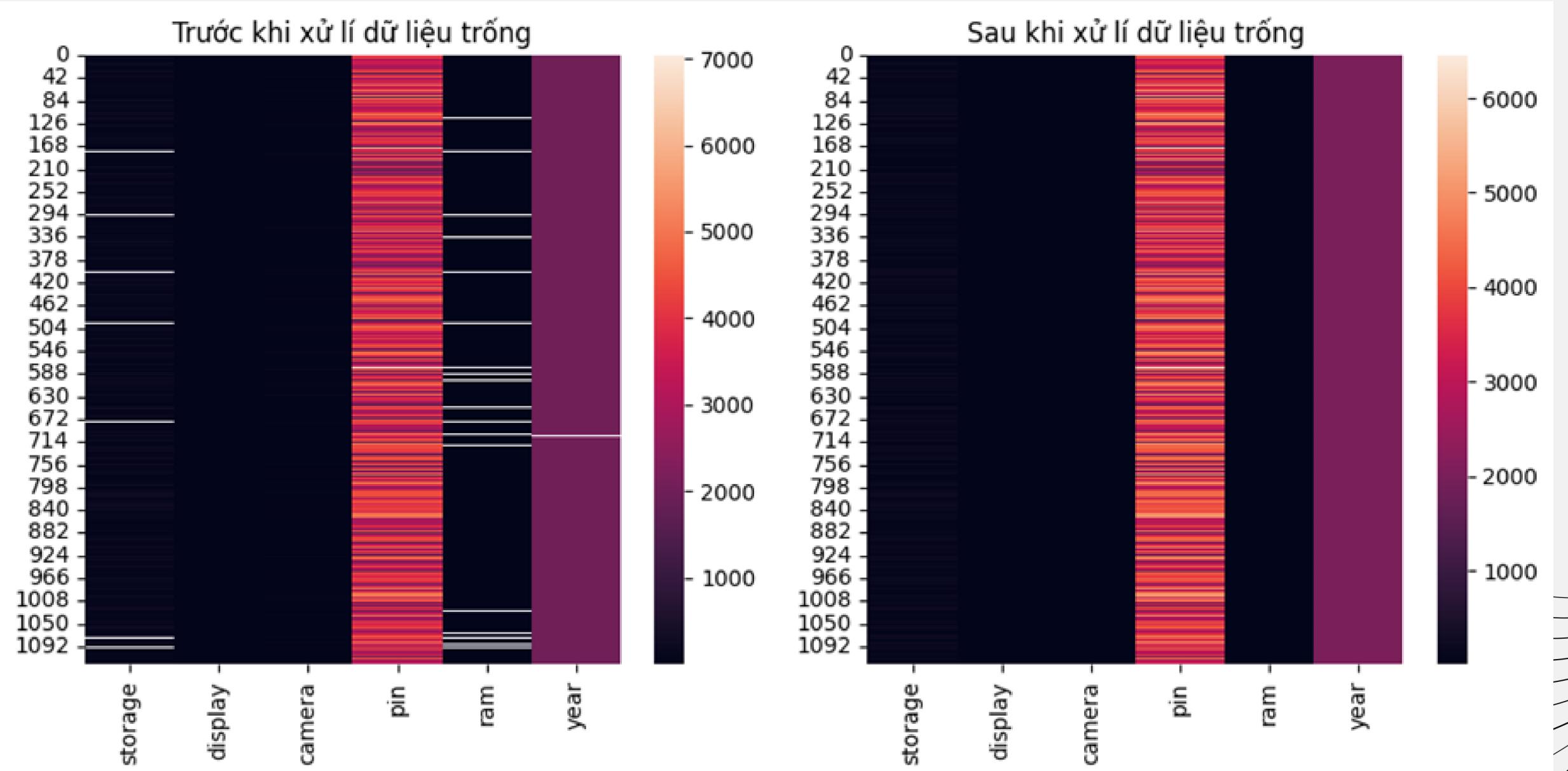
Dữ liệu trước khi được làm sạch

	price	brand	type	condition	preserve	storage	result	display	camera	pin	ram	year
0	9900000.0	1	iphone 12 pro	1	0	128.0	1	6.7	12.0	3687.0	6.0	2020.0
1	1500000.0	0	galaxy a30	1	0	64.0	0	6.4	16.0	4000.0	4.0	2019.0
2	8500000.0	1	iphone 11 pro max	1	0	64.0	0	6.5	12.0	3969.0	4.0	2019.0
3	11000000.0	1	iphone 12 pro	1	0	128.0	1	6.7	12.0	3687.0	6.0	2020.0
4	16900000.0	0	galaxy z fold3	1	1	256.0	1	7.6	12.0	4400.0	12.0	2021.0
5	1500000.0	1	iphone 6s	1	1	64.0	0	4.7	12.0	1715.0	2.0	2015.0
6	5500000.0	1	iphone xr	1	1	64.0	0	6.1	12.0	2942.0	3.0	2018.0
7	4500000.0	1	iphone xs	1	1	64.0	0	6.5	12.0	3174.0	4.0	2018.0
8	260000.0	1	iphone 5	1	0	16.0	0	4.0	8.0	1560.0	1.0	2013.0
9	9200000.0	1	iphone 11 pro max	1	0	64.0	0	6.5	12.0	3969.0	4.0	2019.0

Dữ liệu sau khi đã làm sạch và loại bỏ các chi tiết không cần thiết

Xử lý dữ liệu trống

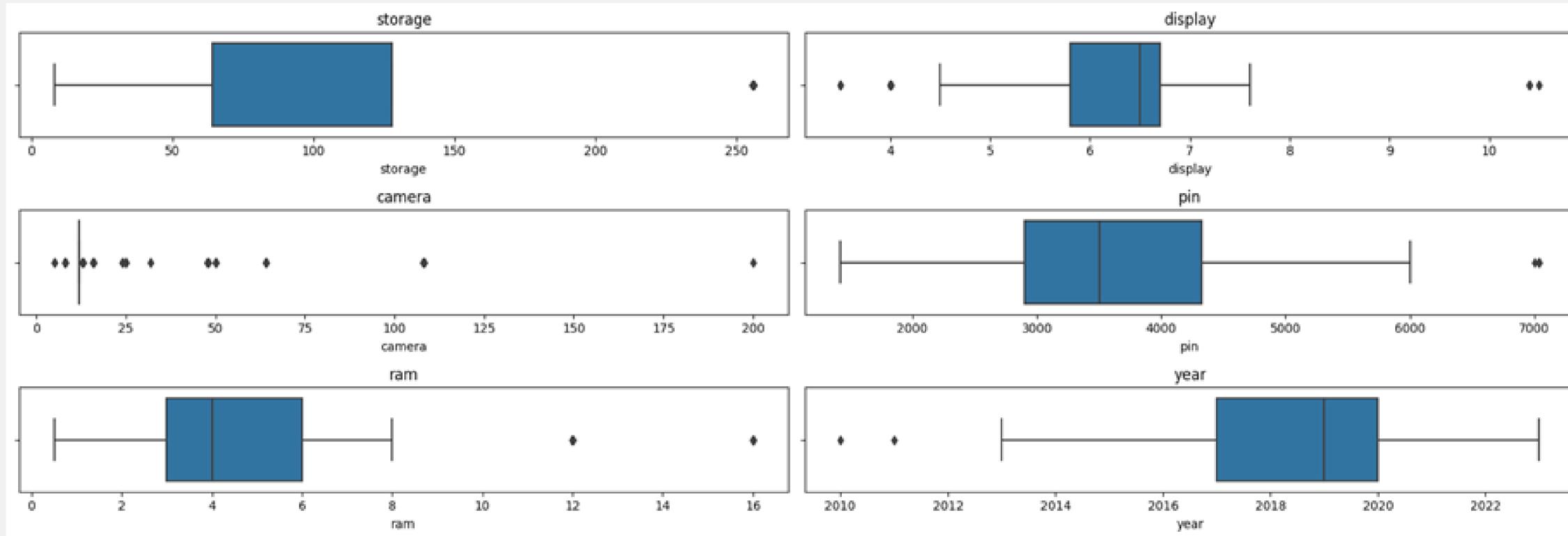
price	0
brand	0
type	0
condition	0
preserve	0
storage	23
result	0
display	5
camera	5
pin	5
ram	60
year	6
dtype:	int64



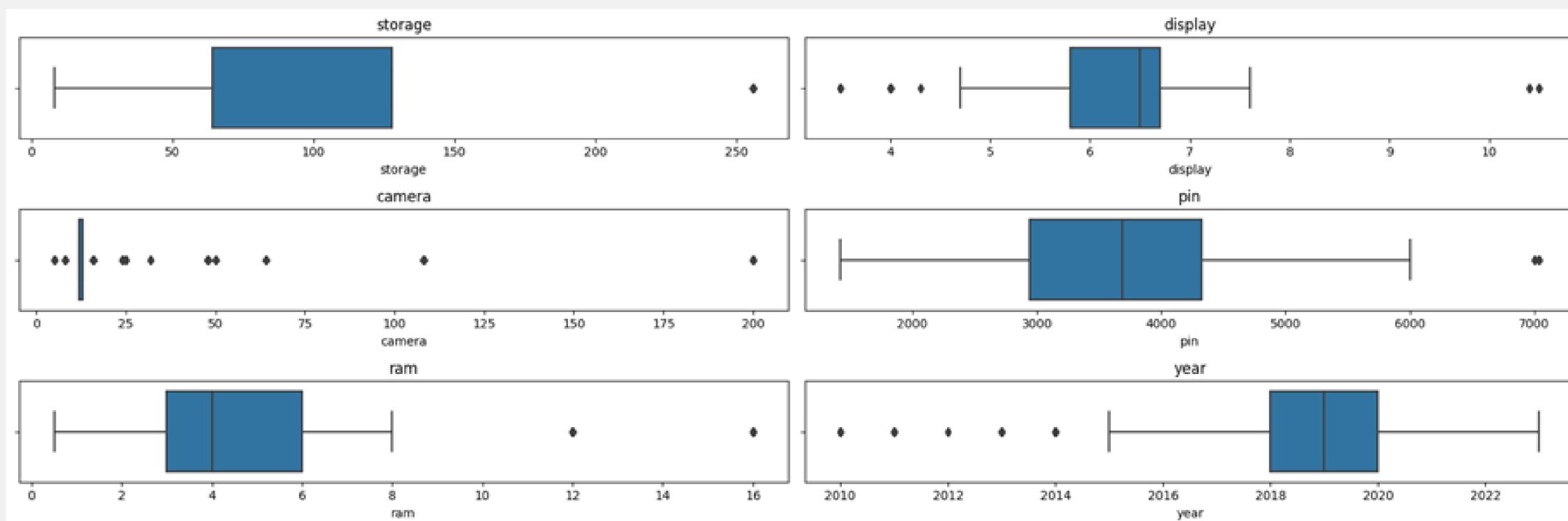
Sử dụng heatmap để thể hiện sự thay đổi trước và sau khi xử lý dữ liệu trống cho các đặc trưng

Xử lí ngoại lệ

Trước khi xử lí ngoại lệ :



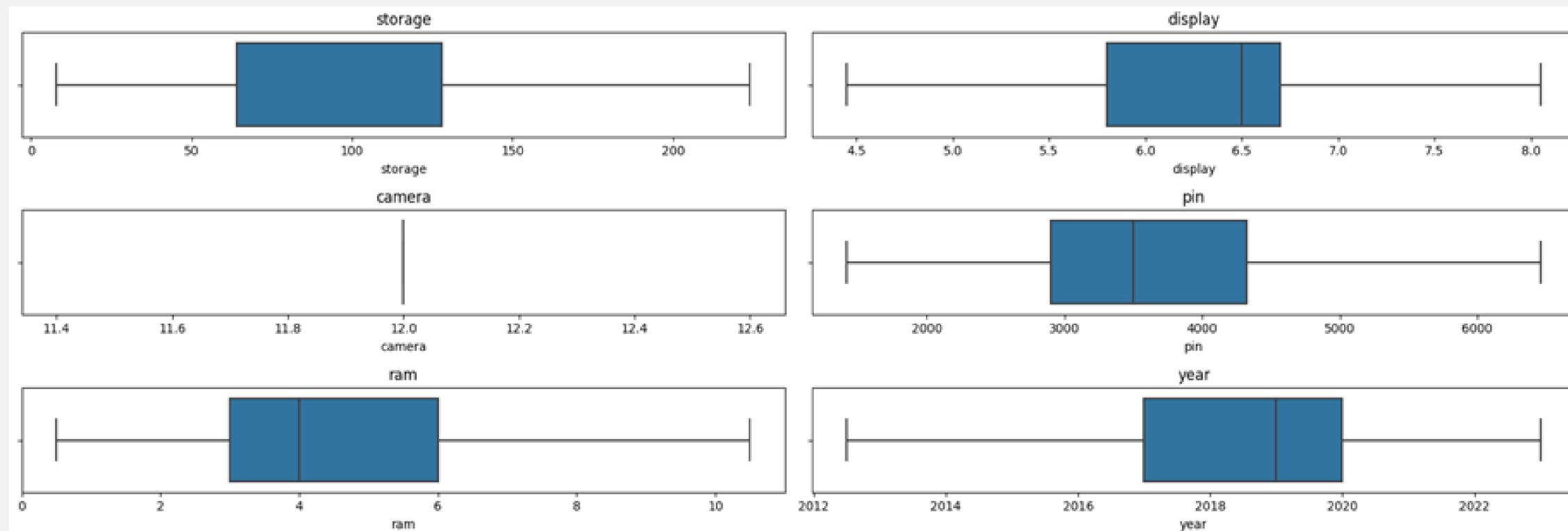
Phân phối dữ liệu của các đặc trưng ở small dataset



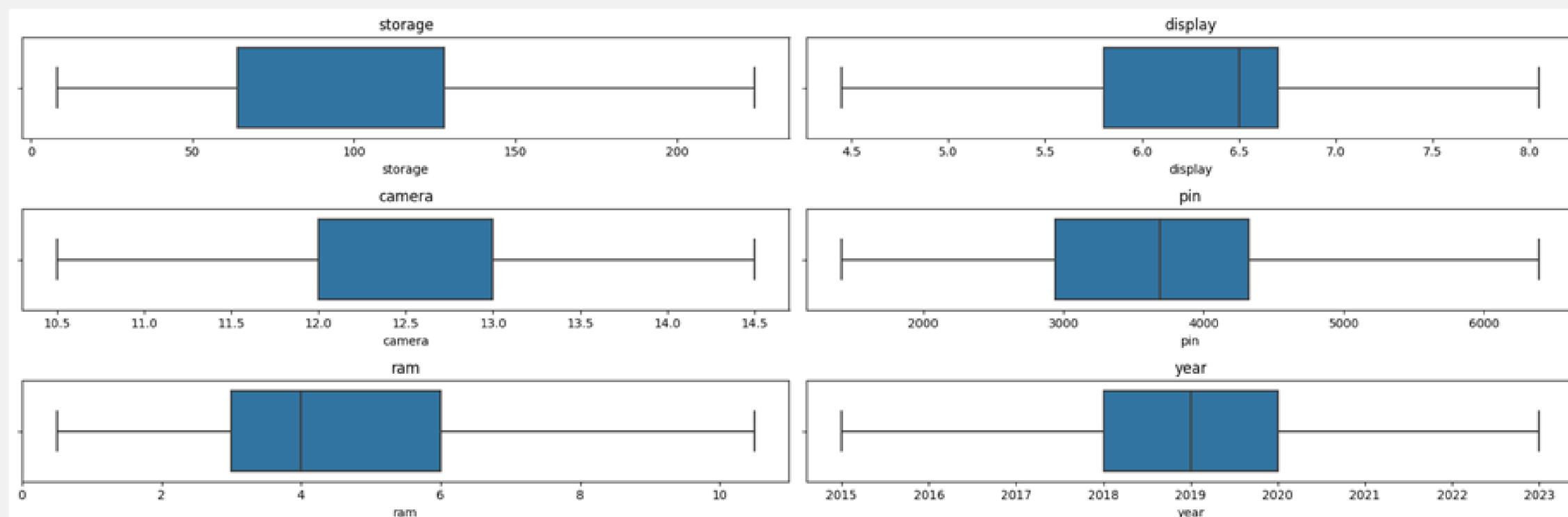
Phân phối dữ liệu của các đặc trưng ở big dataset

Xử lý ngoại lệ

Sau khi xử lí ngoại lệ :



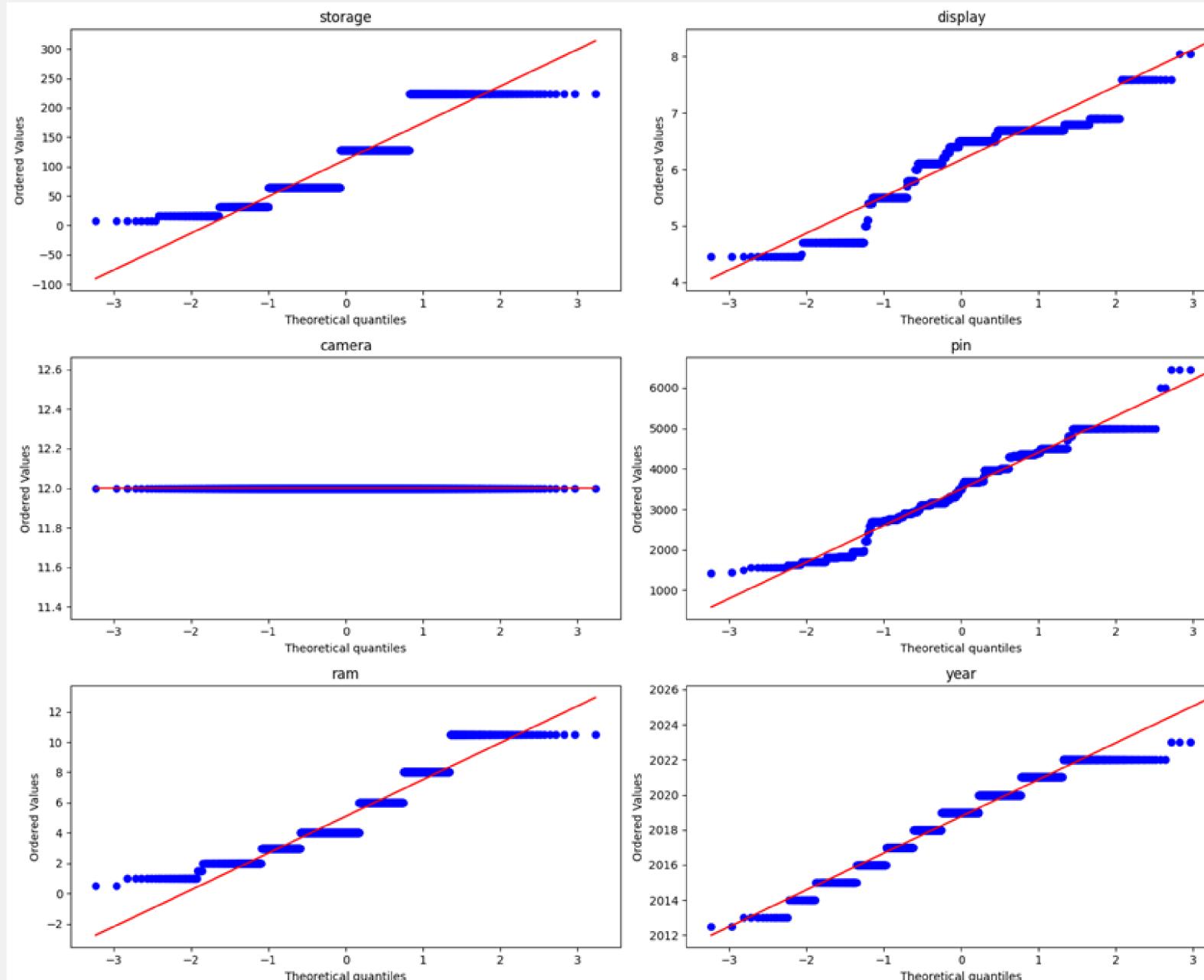
Phân phối dữ liệu của các đặc trưng ở small dataset



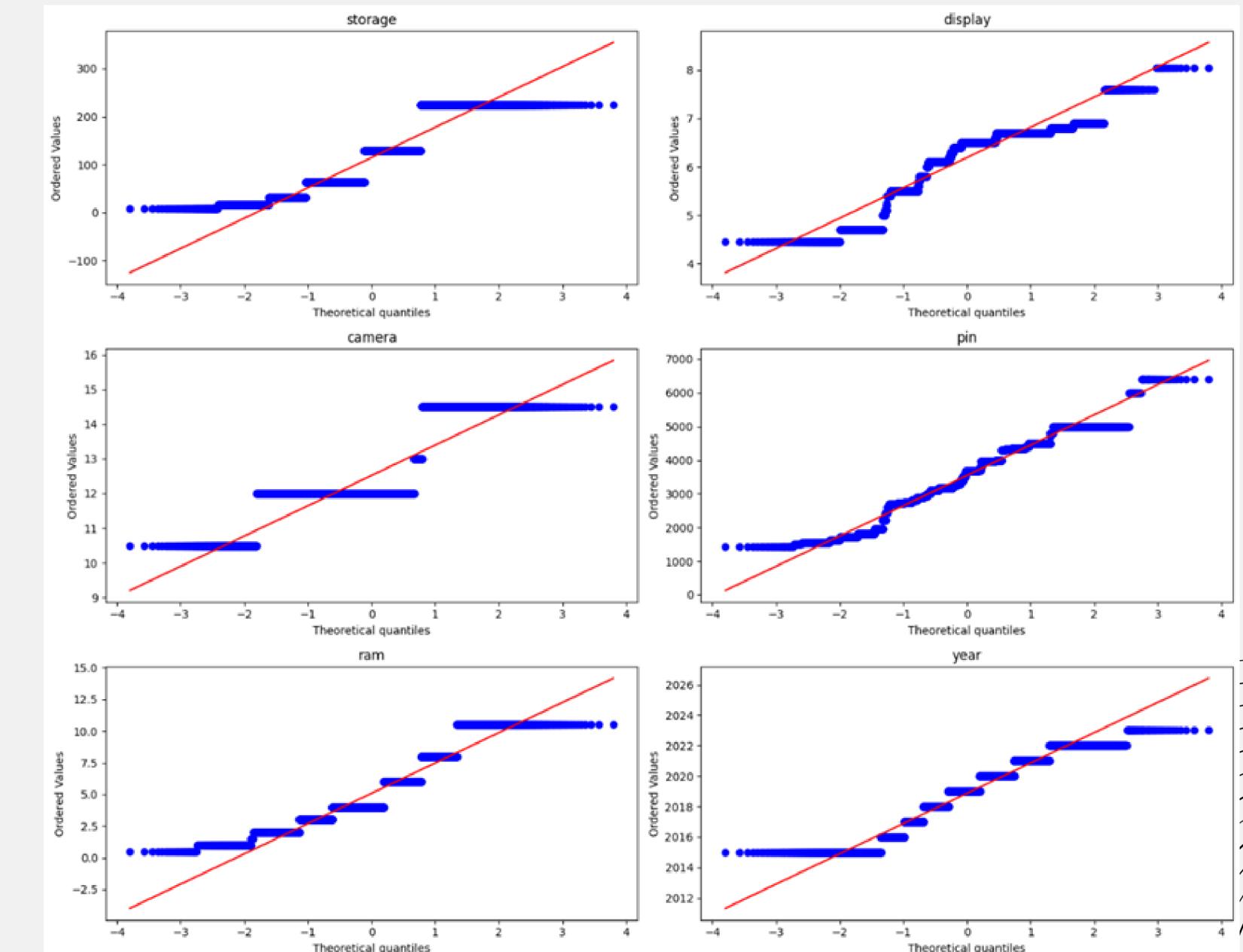
Phân phối dữ liệu của các đặc trưng ở big dataset

Chuẩn hóa đặc trưng

Trước khi chuẩn hóa :



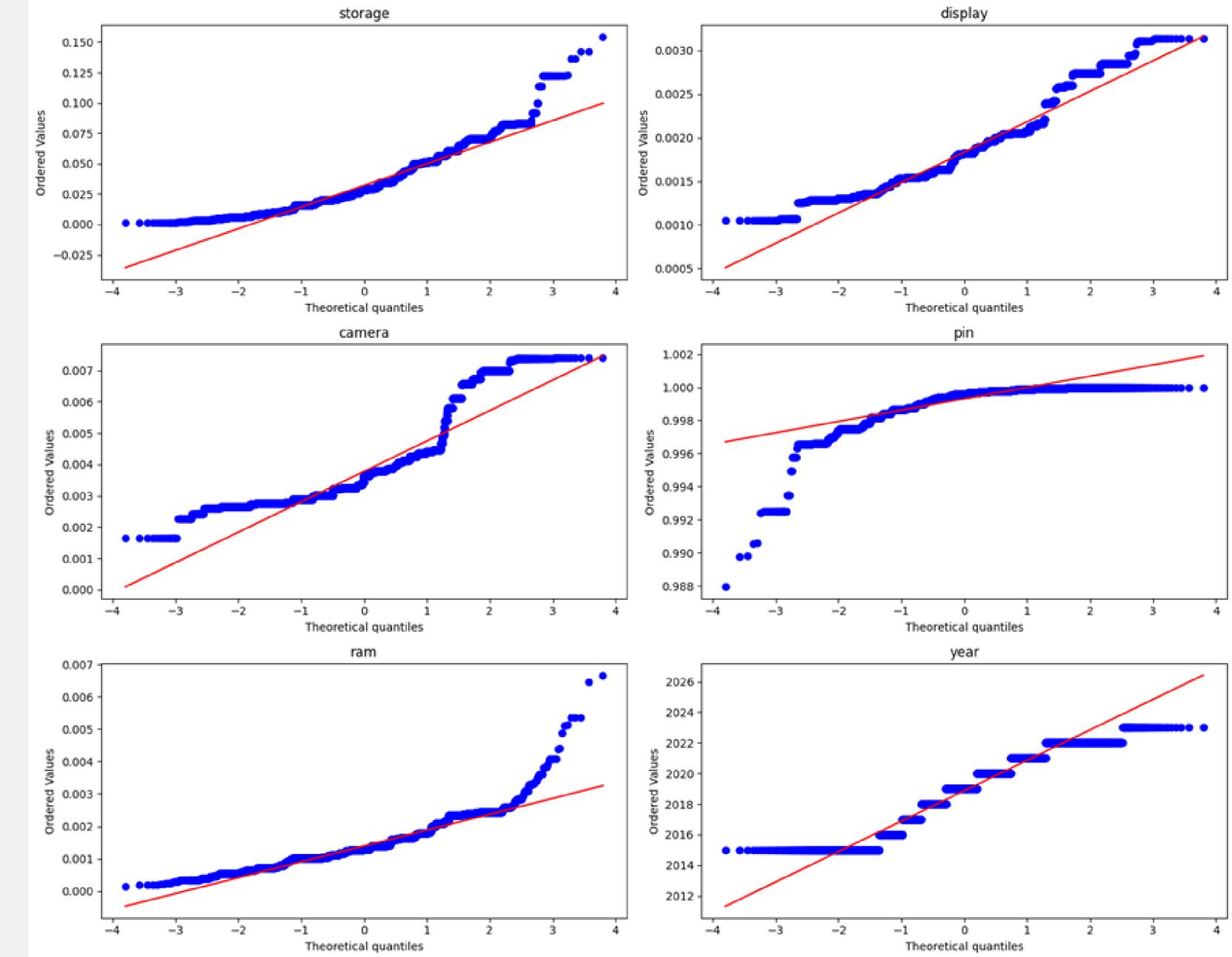
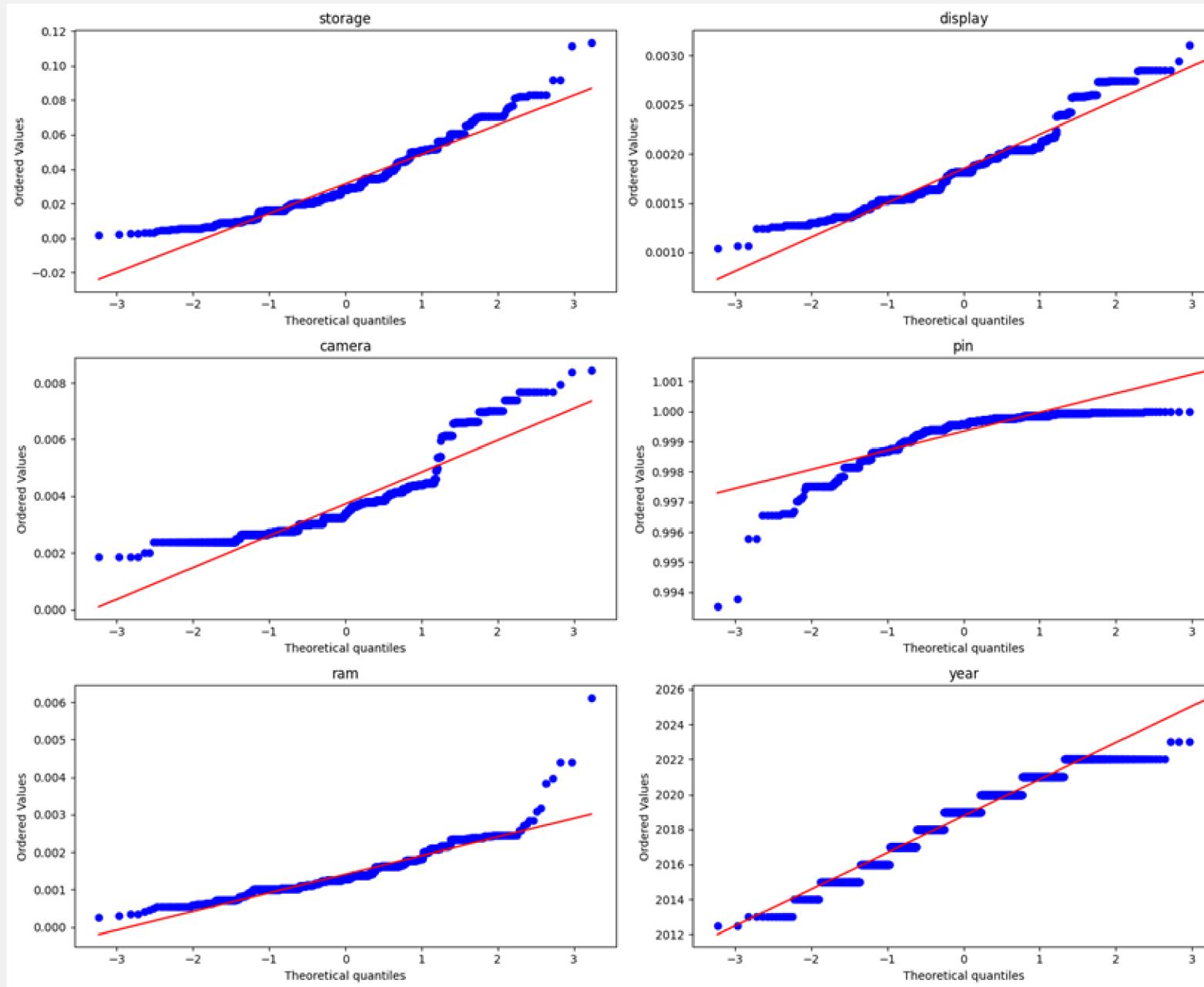
Prob plot của các đặc trưng ở
small dataset



Prob plot của các đặc trưng ở
big dataset

Chuẩn hóa đặc trưng

Sau khi chuẩn hóa dùng Normalizer :

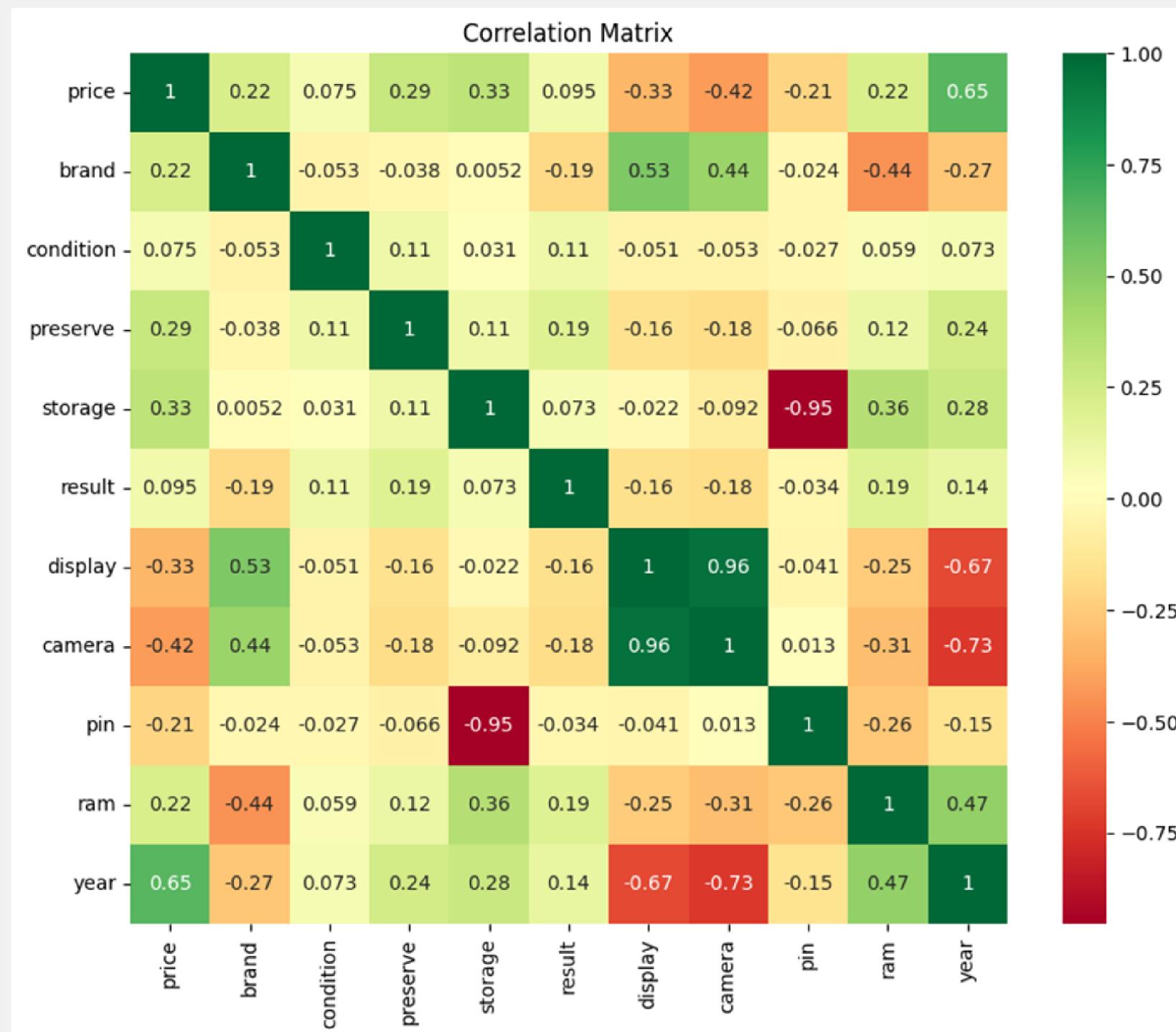


Prob plot của các đặc trưng ở
small dataset

Prob plot của các đặc trưng ở
big dataset

Lựa chọn đặc trưng

Sử dụng ma trận tương quan để thể hiện độ tương quan giữa các đặc trưng đối với price và dùng heatmap để thể hiện sự tương quan đó :



=> Ta nhận thấy đặc trưng có tương quan lớn với price là year và các đặc trưng có tương quan thấp là condition và result. Các đặc trưng còn lại phần lớn nằm trong khoảng 0.2 - 0.3.
Vì vậy ta chọn ngưỡng 0.2 nhằm lấy vừa đủ các đặc trưng quan trọng.

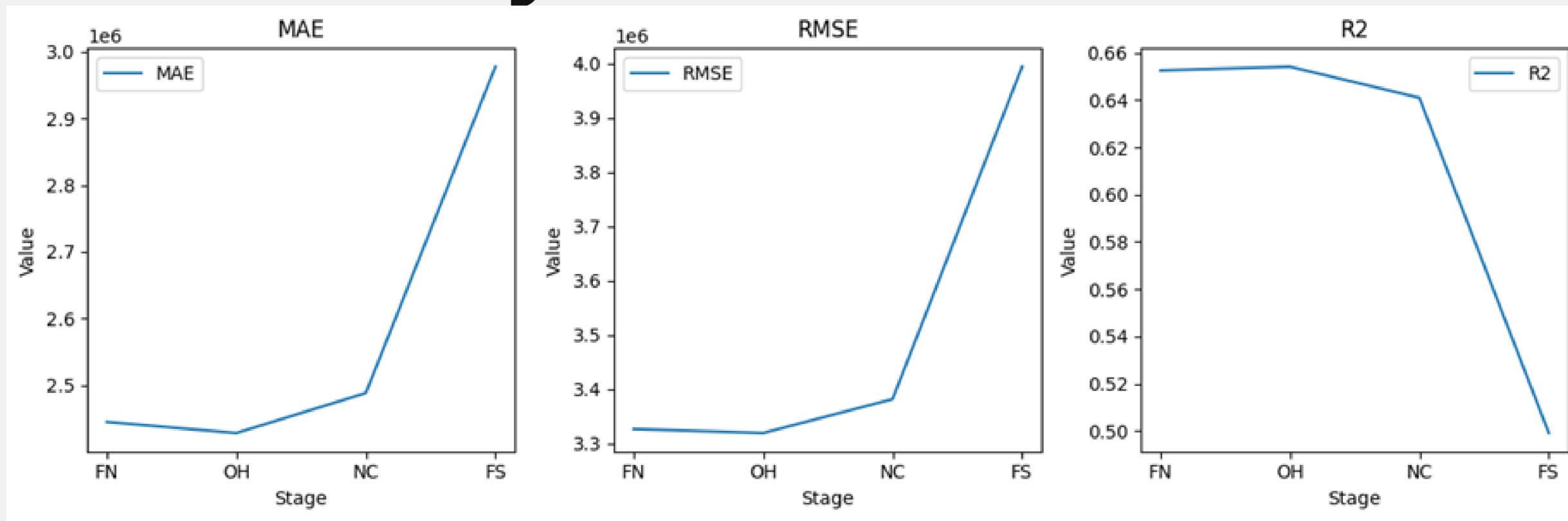
III. Mô hình hóa dữ liệu



Lựa chọn mô hình

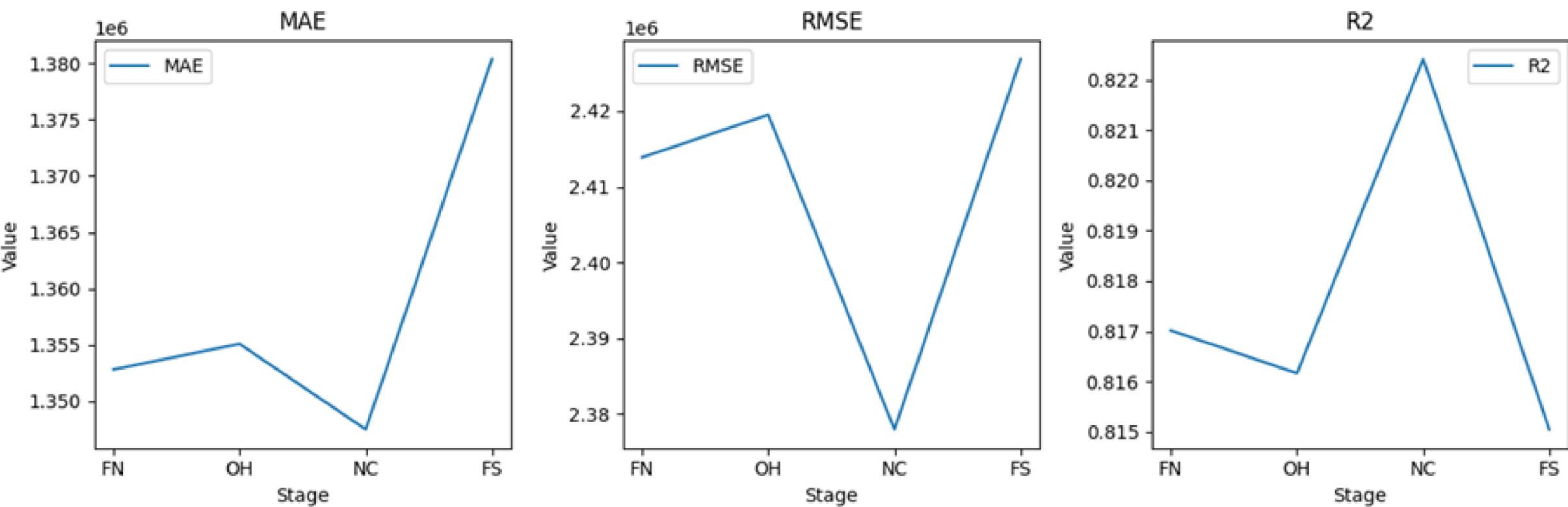
- Linear Regression: Linear Regression là một mô hình hồi quy đơn giản trong việc học máy. Nó là một phương pháp thống kê để xác định mối quan hệ tuyến tính giữa các biến đầu vào (độc lập) và biến đầu ra (phụ thuộc)
- Random Forest Regression: Random Forest Regression là một mô hình dựa trên nguyên tắc "rừng ngẫu nhiên". Nó kết hợp nhiều cây quyết định (decision tree) trong quá trình huấn luyện và tạo ra sự đa dạng trong dự đoán. Điều này giúp giảm thiểu hiện tượng overfitting và cải thiện khả năng dự đoán tổng quát của mô hình

Kiểm thử tiền xử lý



Linear Regression

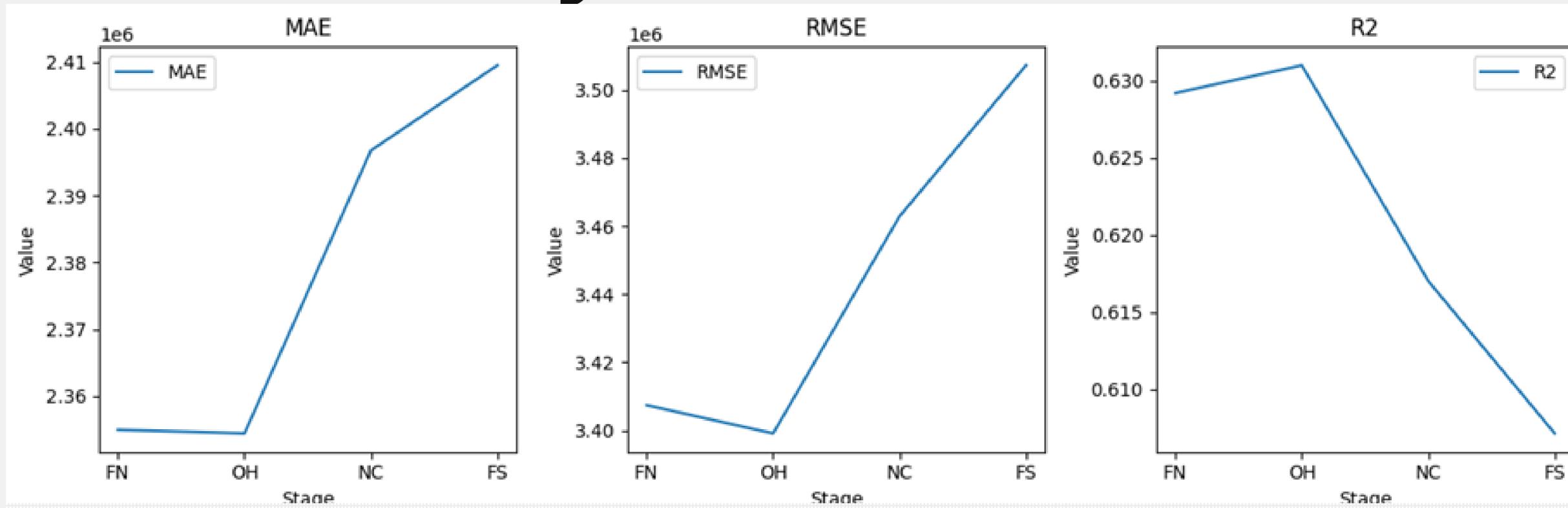
Best metrics value: MAE: 2428231.004813155 RMSE: 3318906.701552667 R2: 0.65408174167379



Random Forest Regression

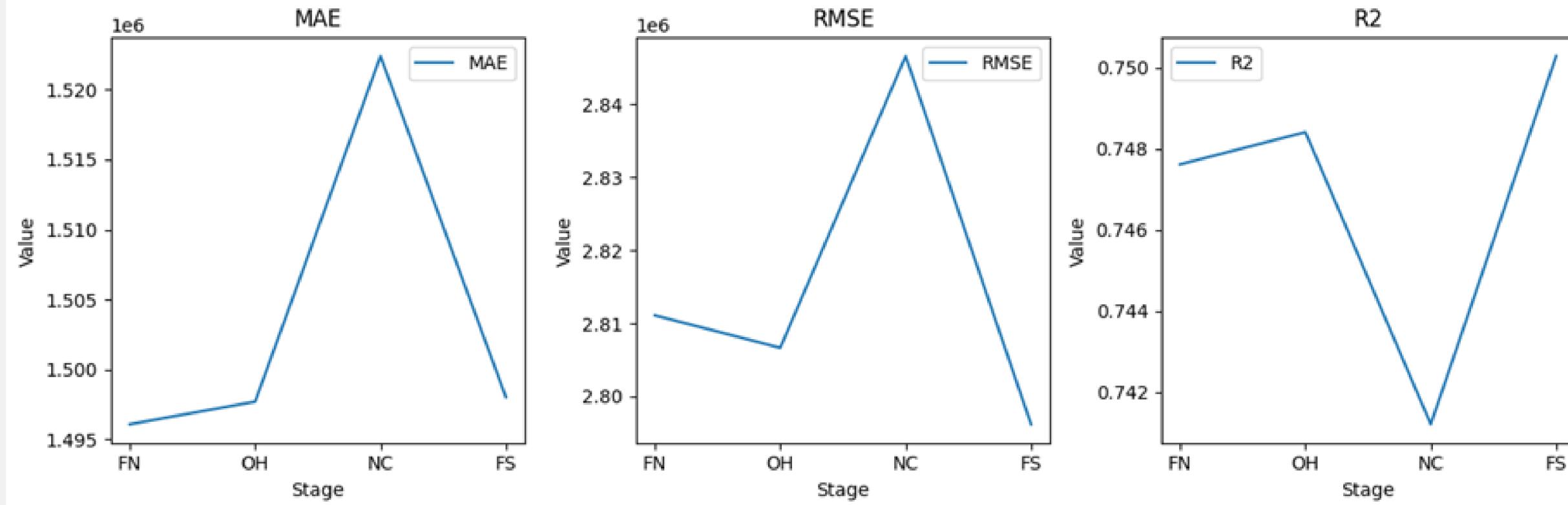
Best metrics value: MAE: 1347485.8191810192 RMSE: 2377989.52640211 R2: 0.8224162296105847

Kiểm thử tiền xử lý



Linear Regression

Best metrics value: MAE: 2354408.8512621955 RMSE: 3399076.0944626066 R2: 0.630994536619345



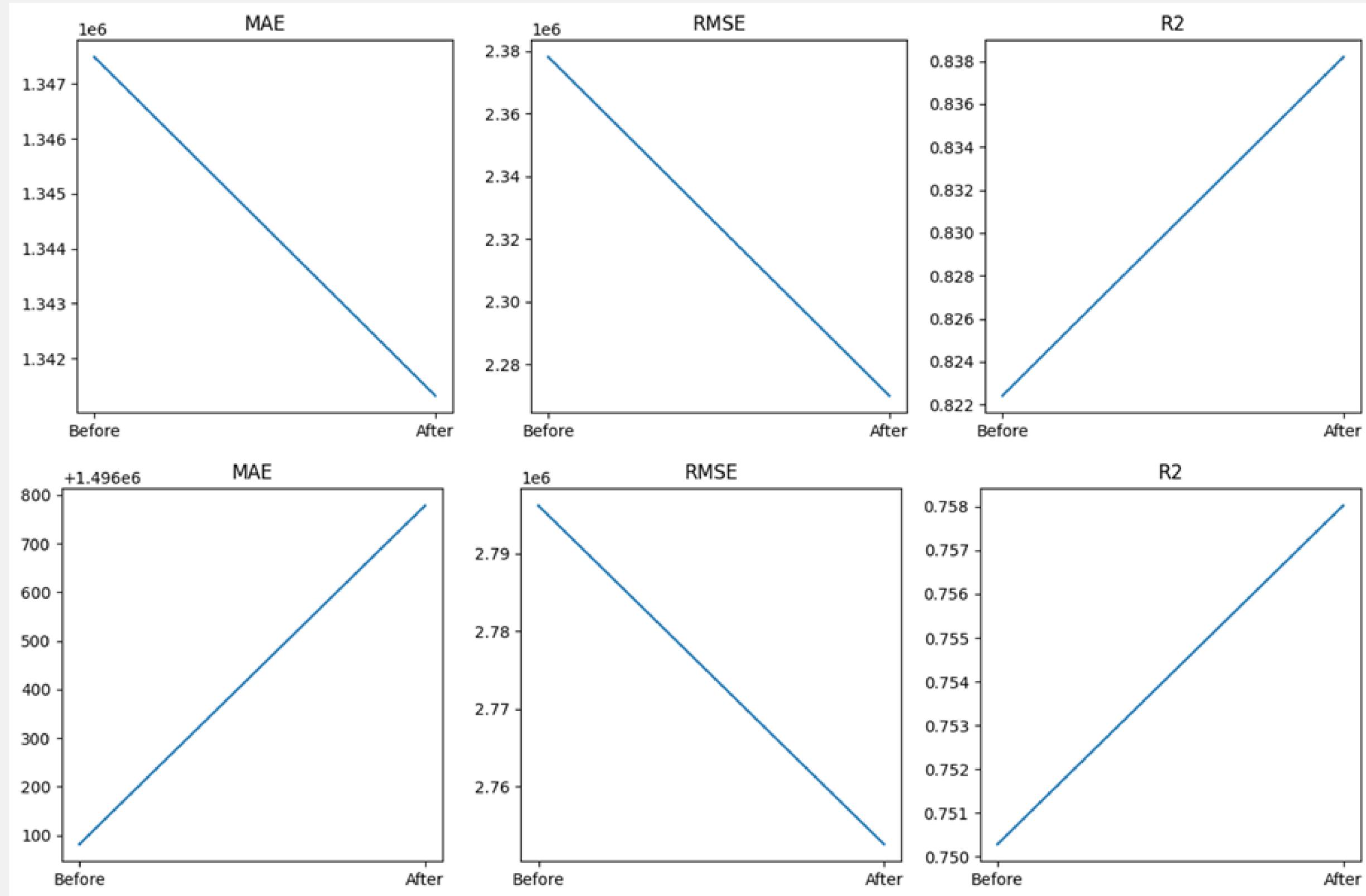
Random Forest Regression

Best metrics value: MAE: 1496081.5533366832 RMSE: 2796190.0746765975 R2: 0.7502851316728109

HyperParameter Tunning

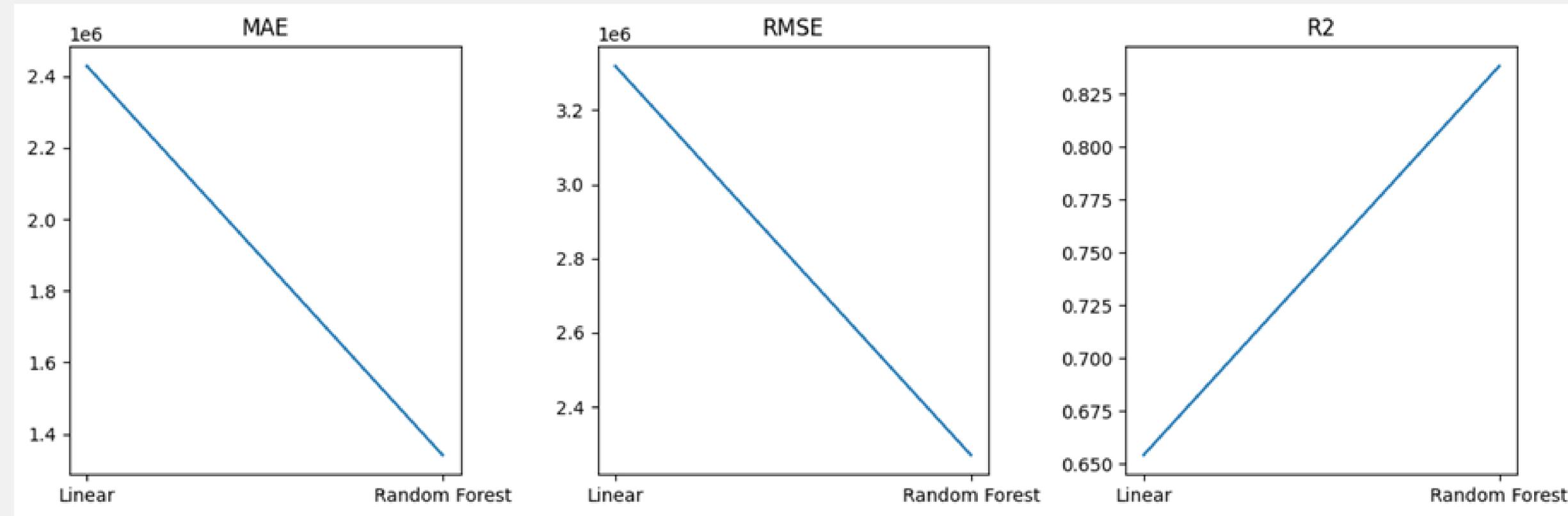
- 'bootstrap': [True, False], quyết định liệu có lấy mẫu ngẫu nhiên tái chọn để tạo ra mỗi cây trong Random Forest hay sử dụng toàn bộ mẫu để tạo ra mỗi cây trong Random Forest
- 'max_depth': [None, 5, 10], None: Các cây trong Random Forest không có giới hạn độ sâu. 5, 10: Giới hạn độ sâu của các cây trong Random Forest. Điều này giới hạn số lượng nút và giúp kiểm soát độ phức tạp của mô hình.
- 'max_features': ['auto', 'sqrt'], 'auto': Số lượng đặc trưng được sử dụng để xây dựng mỗi cây trong Random Forest được tự động điều chỉnh. 'sqrt': Số lượng đặc trưng được sử dụng để xây dựng mỗi cây là căn bậc hai của số lượng đặc trưng ban đầu.
- 'min_samples_leaf': [1, 2, 4], Số lượng mẫu tối thiểu cần có trong mỗi lá của cây quyết định. Giá trị nhỏ hơn sẽ tạo ra các cây phức tạp hơn và dễ gây overfitting.
- 'min_samples_split': [2, 5, 10], Số lượng mẫu tối thiểu yêu cầu để chia một nút trong cây. Giá trị nhỏ hơn sẽ tạo ra cây phức tạp hơn và dễ gây overfitting.
- 'n_estimators': [100, 200, 300], Số lượng cây trong Random Forest

HyperParameter Tunning



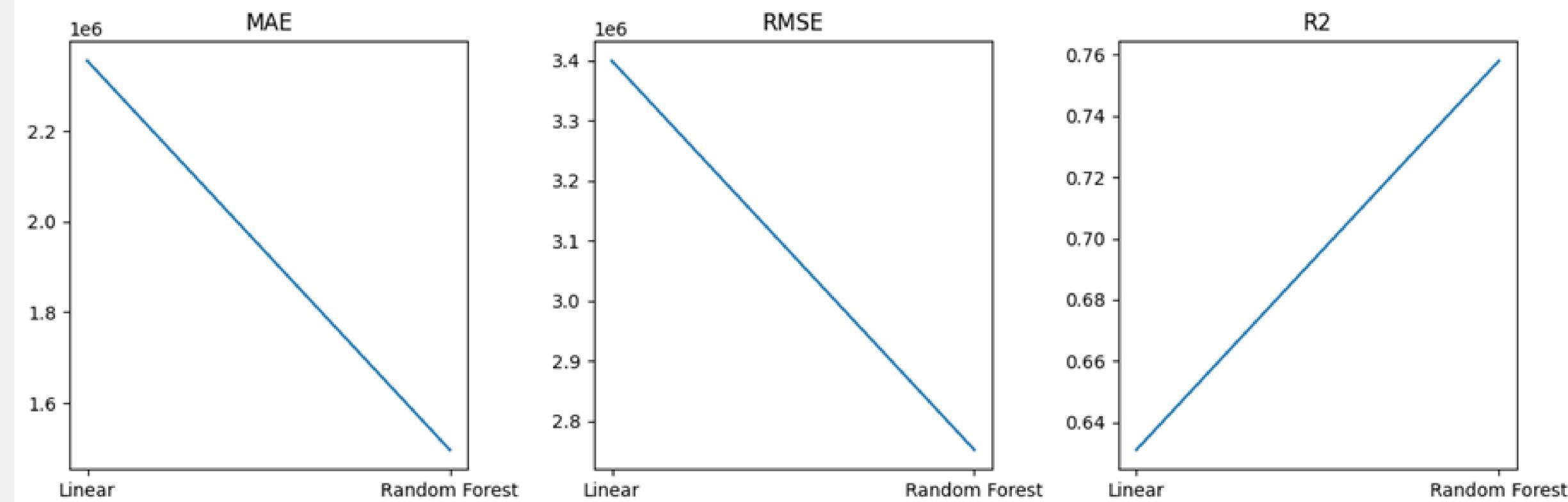
=> Đều cho thấy việc điều chỉnh tham số gia tăng độ hiệu quả của mô hình

So sánh 2 mô hình



Linear Regression: MAE: 2428231.004813155 RMSE: 3318906.701552667 R2: 0.65408174167379

Random Forest Regression: MAE: 1341324.1543589572 RMSE: 2269913.6731243203 R2: 0.8381912215193819



Linear Regression: MAE: 2354408.8512621955 RMSE: 3399076.0944626066 R2: 0.630994536619345

Random Forest Regression: MAE: 1496778.4213080446 RMSE: 2752547.709871014 R2: 0.7580192978493049

Kết quả

- Điểm chung:
 - Linear Regression ở cả 2 bộ dữ liệu đều có thể được cải thiện với kĩ thuật lọc ngoại lệ Outlier Handling
 - Random Forest Regression hiệu quả hơn so với Linear Regression ở cả 2 bộ dữ liệu
- Điểm khác:
 - Random Forest ở BigDS có thể cải thiện bởi kĩ thuật chuẩn hóa Normalizer Scaling, ngược lại khi ở SmallIDS có thể cải thiện bởi Outlier Handling và Feature Selection

Kết luận

1. KẾT QUẢ ĐẠT ĐƯỢC

- Đã thỏa mãn các yêu cầu tiểu luận đưa ra bao gồm: Thu thập dữ liệu, trích xuất đặc trưng, huấn luyện mô hình
- Mô hình đầu ra có độ hiệu quả cao, có thể áp dụng vào việc dự đoán giá điện thoại cũ

2. HƯỚNG PHÁT TRIỂN

- Thử nghiệm bộ dữ liệu với các mô hình dự đoán giá trị liên tục khác như Support Vector Regression hoặc Gradient Boosting Regression