NHẬP MÔN KHOA HỌC DỮ LIỆU
ĐỒ ÁN CUỐI KỲ

DỰ ĐOÁN GIÁ XE

GVHD: Lê Nhựt Nam

20120505 - Nguyễn Duy Khang
20120510 - Nguyễn Hữu Anh Khoa
20120588 - Lê Quang Thọ
19120633 - Nguyễn Anh Quốc

NỘI DUNG

1. Giới thiệu đồ án
2. Thu thập dữ liệu
3. Tiền xử lí dữ liệu
4. Trực quan hóa dữ liệu
5. Mô hình hóa dữ liệu và đánh giá

GIỚI THIỆU ĐỒ ÁN

Chủ đề: Dự đoán giá xe theo dữ liệu trên web được bán tại Úc.

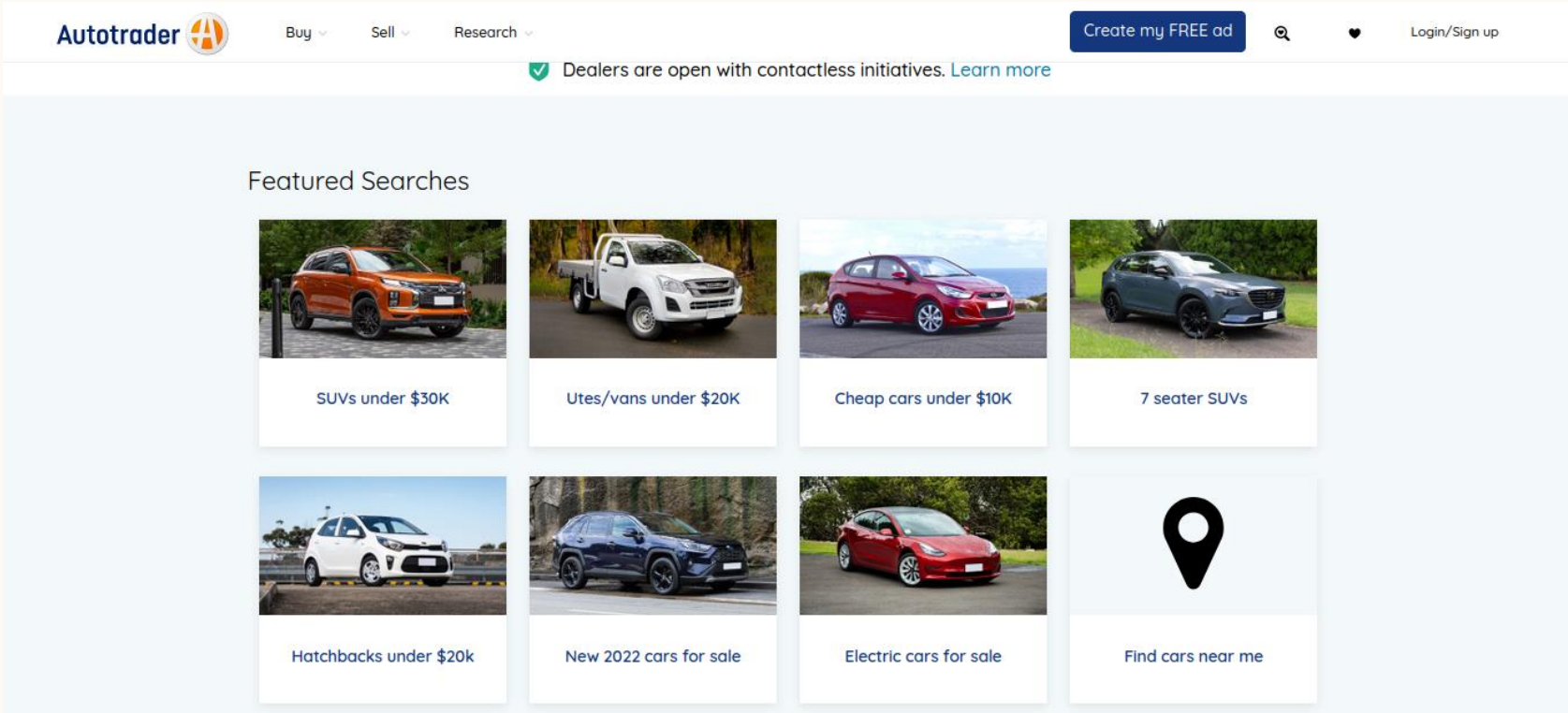
Câu hỏi: Liệu chúng ta có thể dự đoán giá xe thông qua các đặc trưng của xe ?

Lợi ích: đem những thông tin có ích cho những người mua xe, muốn tìm hiểu về xe.

Link github: https://github.com/huukhoa2112/NMKHDL-Project1

THU THẬP DỮ LIỆU

Nguồn dữ liệu: https://www.autotrader.com.au/for-sale



Màn hình chính của trang web

THU THẬP DỮ LIỆU

Công cụ để crawl dữ liệu: thư viện Scrapy của Python

Nhóm đã check trong file robots.txt của trang web thì web cho phép thu thập dữ liệu

```
import urllib.robotparser
```

```
rp = urllib.robotparser.RobotFileParser()
rp.set_url("https://www.autotrader.com.au//robots.txt")
rp.read()
```

```
rp.can_fetch("*", "https://www.autotrader.com.au//for-sale")
```

```
True
```

```
rp.can_fetch("*", "https://www.autotrader.com.au/car/12817873/renault/arkana/nsw/narellan/suv")
```

```
True
```

Sau khi crawl dữ liệu về máy, bộ dữ liệu có hơn 21000 dòng và 15 cột.

- ID: mã xe
- Name: tên xe
- Price: giá xe
- Brand: hãng xe
- Model: đời xe
- Variant: biến thể từ mô hình xe
- Series: dòng xe
- Year: năm sản xuất xe
- Gearbox: loại hộp số của xe
- Body Type: loại thiết kế
- Fuel Type: loại xăng dầu
- Status: tình trạng xe
- Kilometers: số km xe đi được
- CC: số phân khối
- Color: màu sắc xe
- Seating Capacity: số chỗ ngồi

TIỀN XỬ LÍ DỮ LIỆU

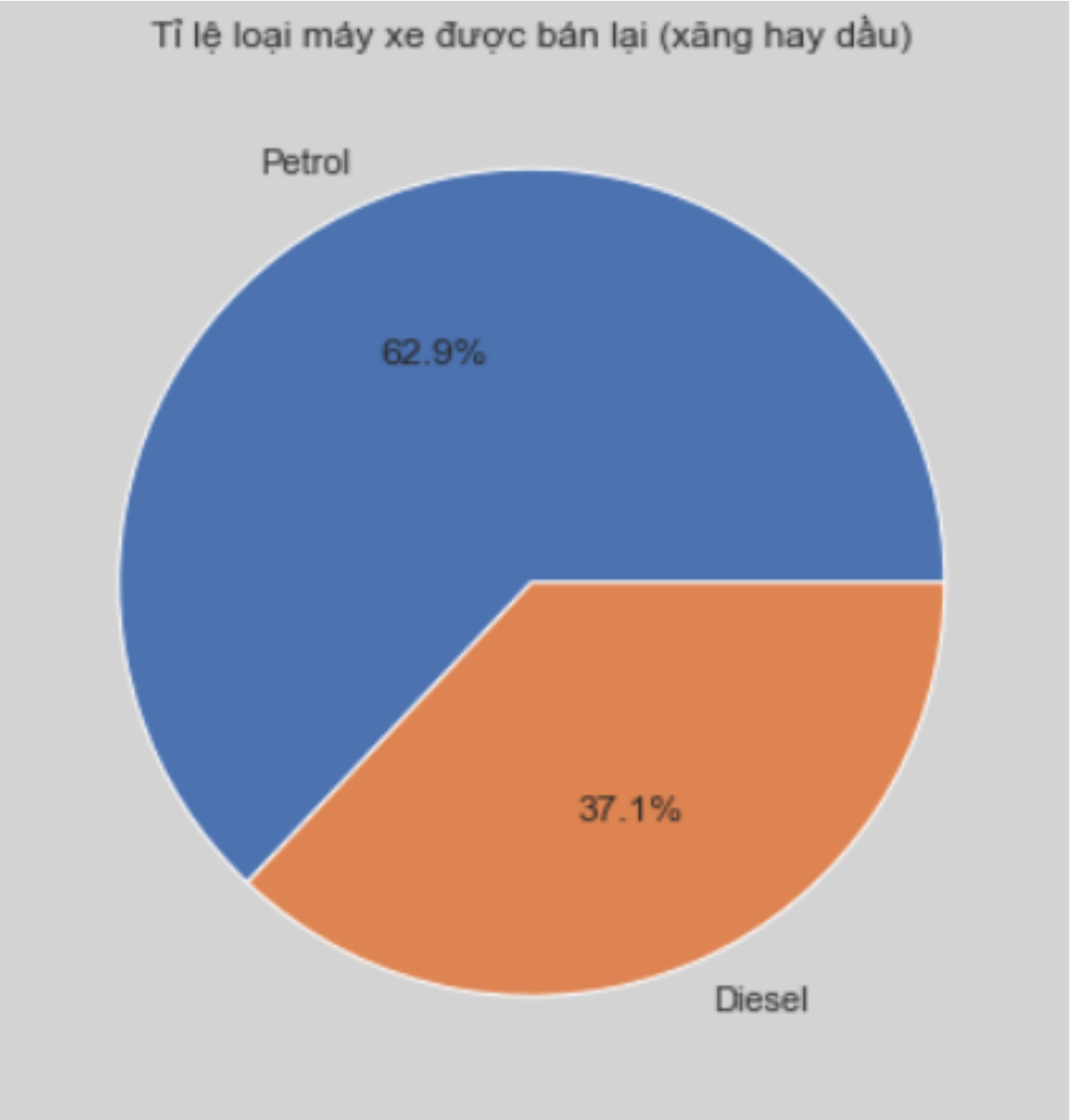
Sau khi thu thập dữ liệu, dữ liệu chứa các giá trị thiếu (NaN) và các cột vẫn chưa định dạng của kiểu dữ liệu vì thế ta phải tiến hành tiền xử lí dữ liệu.

- Đầu tiên, check xem có dòng nào bị lặp không (bỏ index=ID) sau đó ta xóa.
- Tiếp tục có 4 cột ta cần chuyển kiểu dữ liệu string sang int là {"Price", "Kilometers", "CC", "Seating Capacity"}
    - Đối với cột "Price", đầu tiên ta cần replace "$" sang string empty.
    - Đối với cột "Kilometers", đầu tiên ta cần replace " km" sang string empty.
- Sau khi chuyển sang kiểu dữ liệu int, ta xóa các dòng có dữ liệu trống ở trong 4 cột này.

TRỰC QUAN HÓA DỮ LIỆU

ĐẶT CÂU HỎI

Câu hỏi 1: Xe động cơ chạy bằng loại nhiên liệu nào được bán lại nhiều hơn?



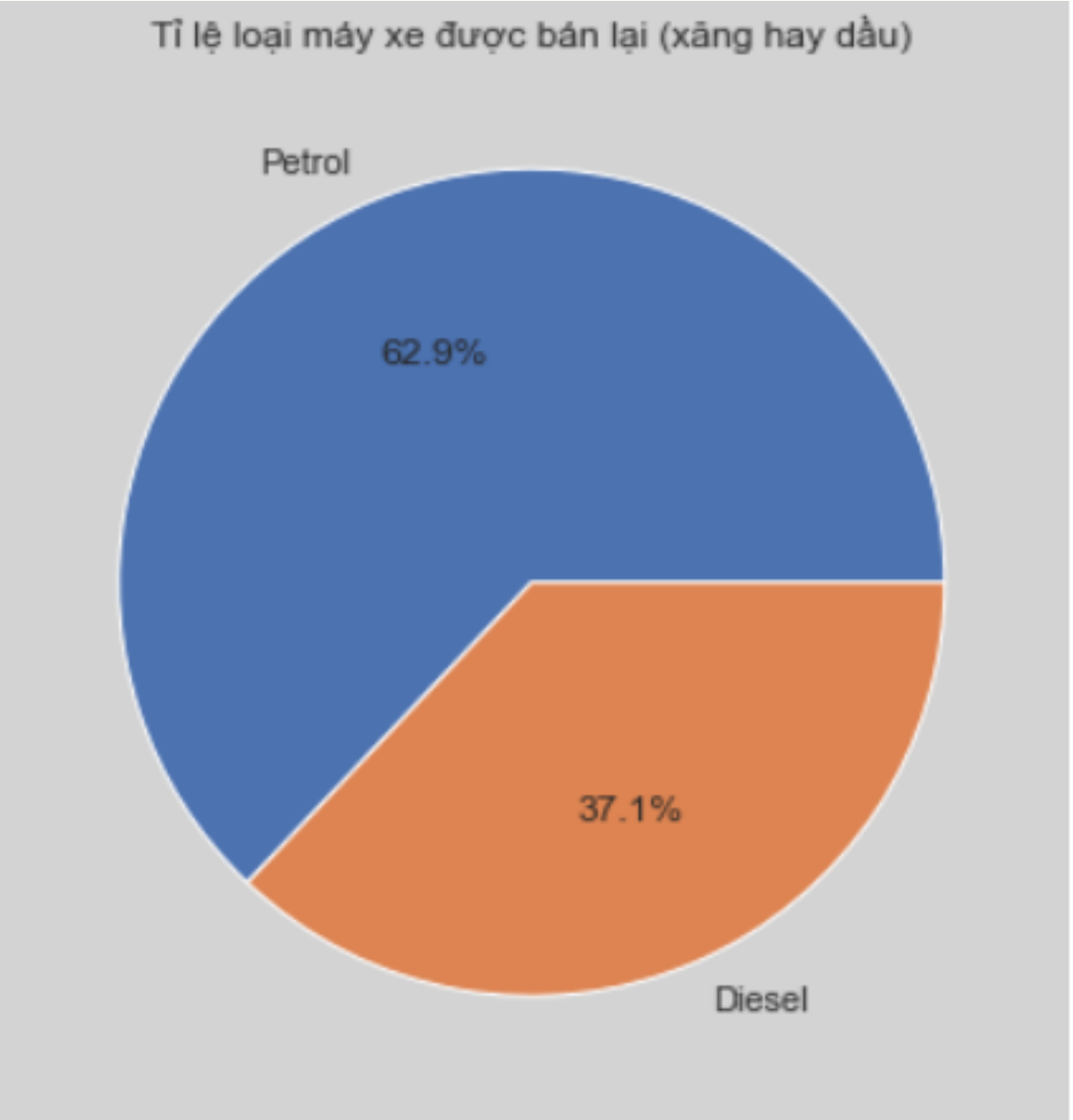
TRỰC QUAN HÓA DỮ LIỆU

Xe chạy bằng xăng bị bán lại nhiều hơn xe dầu (gần gấp đôi).

Vậy tại sao xe xăng lại bị bán lại nhiều như vậy, có thể do một số nhược điểm sau trong quá trình sử dụng:

- Tốn nhiên liệu hơn xe máy dầu.
- Dễ bốc cháy ở nhiệt độ cao gây nguy hiểm.

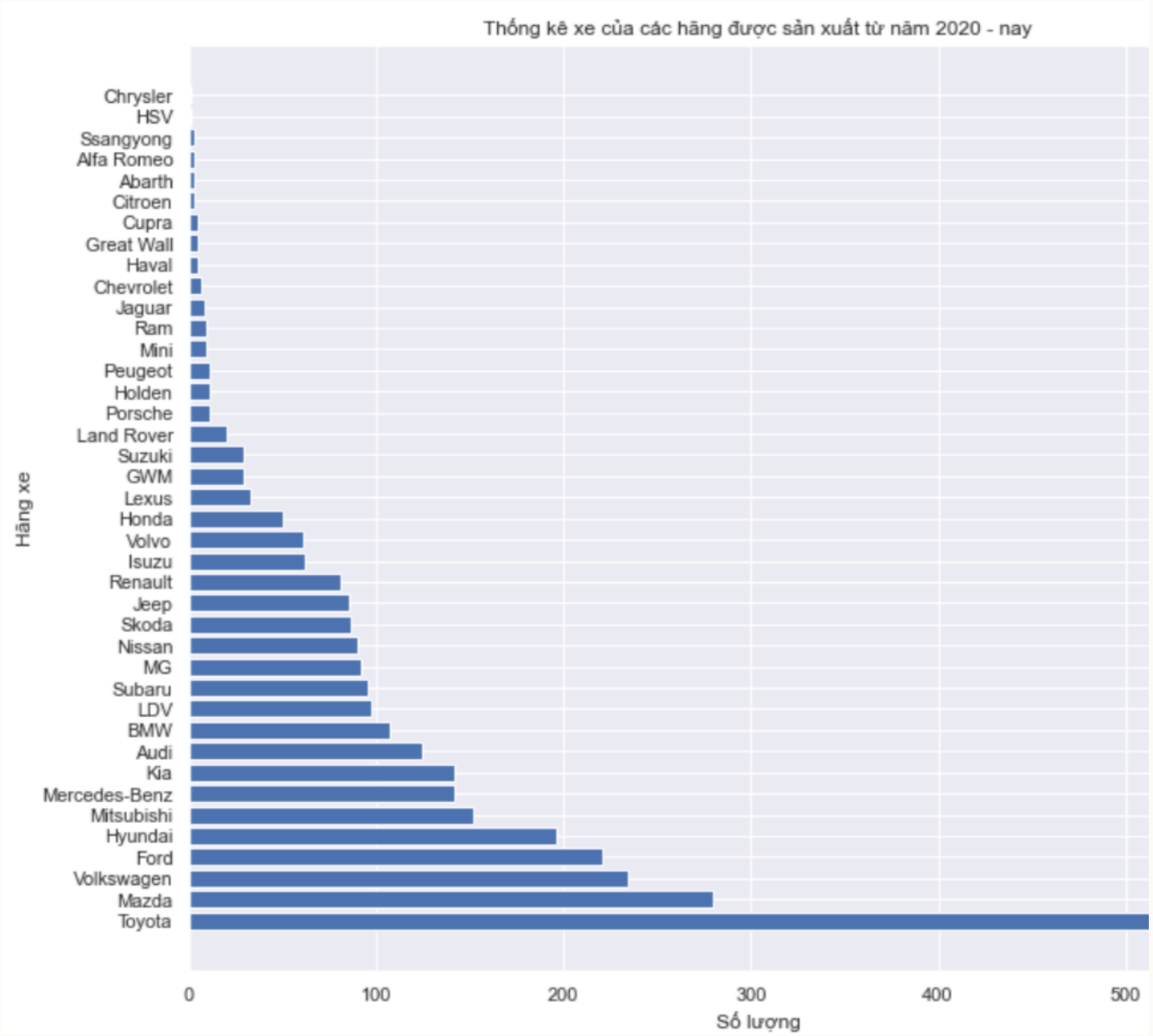
Ở bối cảnh mà tình hình năng lượng thế giới đang căng thẳng, việc những chiếc xe xăng tốn nhiều nguyên liệu hơn dễ bị bán lại hơn có vẻ cũng hợp lý.



TRỰC QUAN HÓA DỮ LIỆU

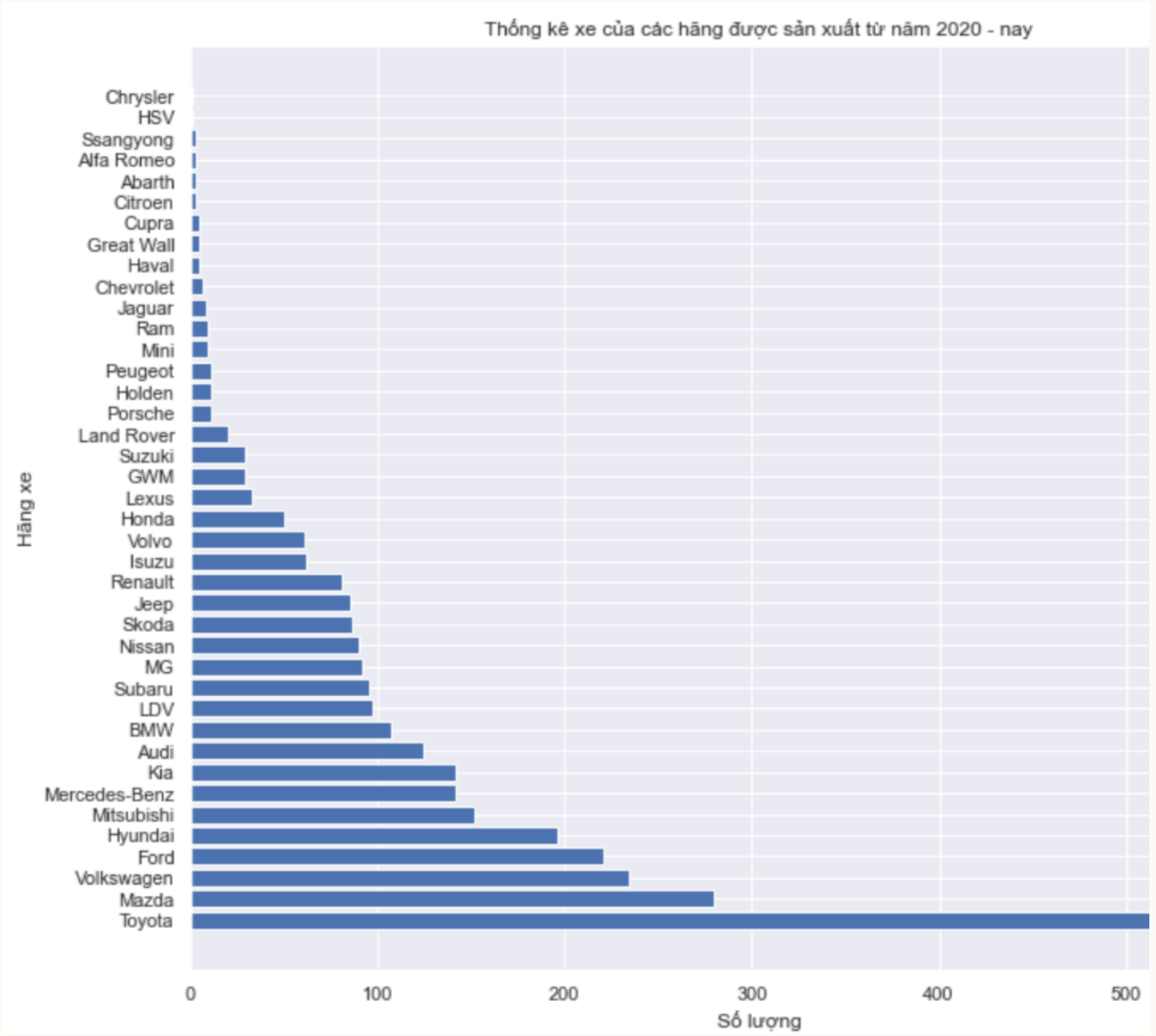
Câu hỏi 2:

Những hãng xe nào sản xuất nhiều xe nhất trong những năm gần đây (2020 - nay) và những hãng xe nào sản xuất ít xe nhất hoặc thậm chí ngừng sản xuất xe?



TRỰC QUAN HÓA DỮ LIỆU

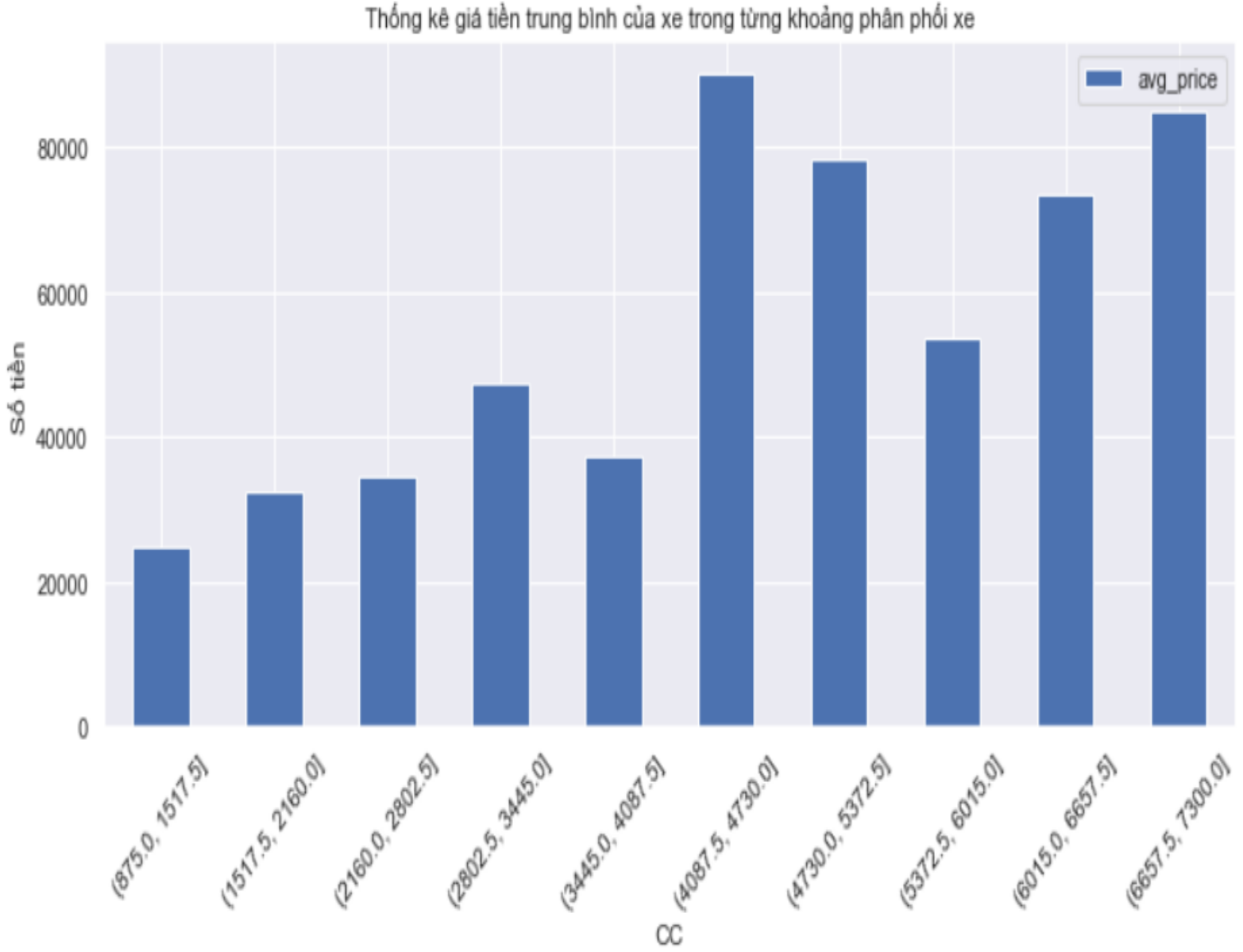
Một số hãng xe dường như không còn sản xuất xe mới nữa (trong phạm vi tập dữ liệu này), ví dụ như Chrysler và HSV, hoặc một số hãng khác có rất ít xe được sản xuất trong 2 năm gần đây. Mặt khác ta cũng có thể thấy những hãng xe có rất nhiều xe mới, dẫn đầu không có gì bất ngờ khi là Toyota. Kế tiếp là Mazda, Volkswagen, Ford, ...



TRỰC QUAN HÓA DỮ LIỆU

Câu hỏi 3: Giá xe trung bình của các xe trong các khoảng phân phối của xe ?

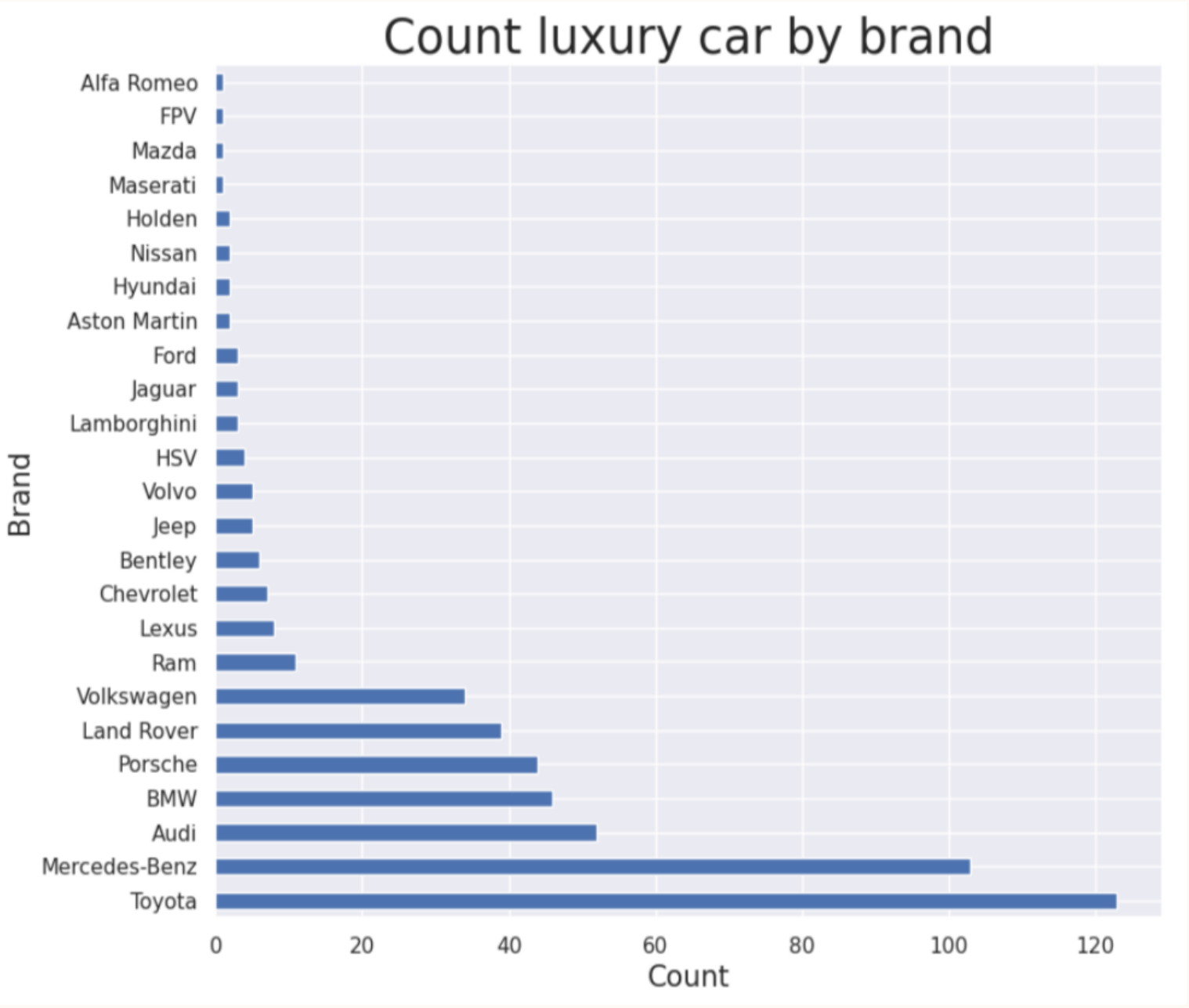
Ta thấy phân phối xe cao hay thấp chưa quyết định nhiều đến giá xe nó còn phụ thuộc vào nhiều vào yếu tố khác. Yếu tố phân khối của xe chỉ góp 1 phần trong quyết định giá xe.



TRỰC QUAN HÓA DỮ LIỆU

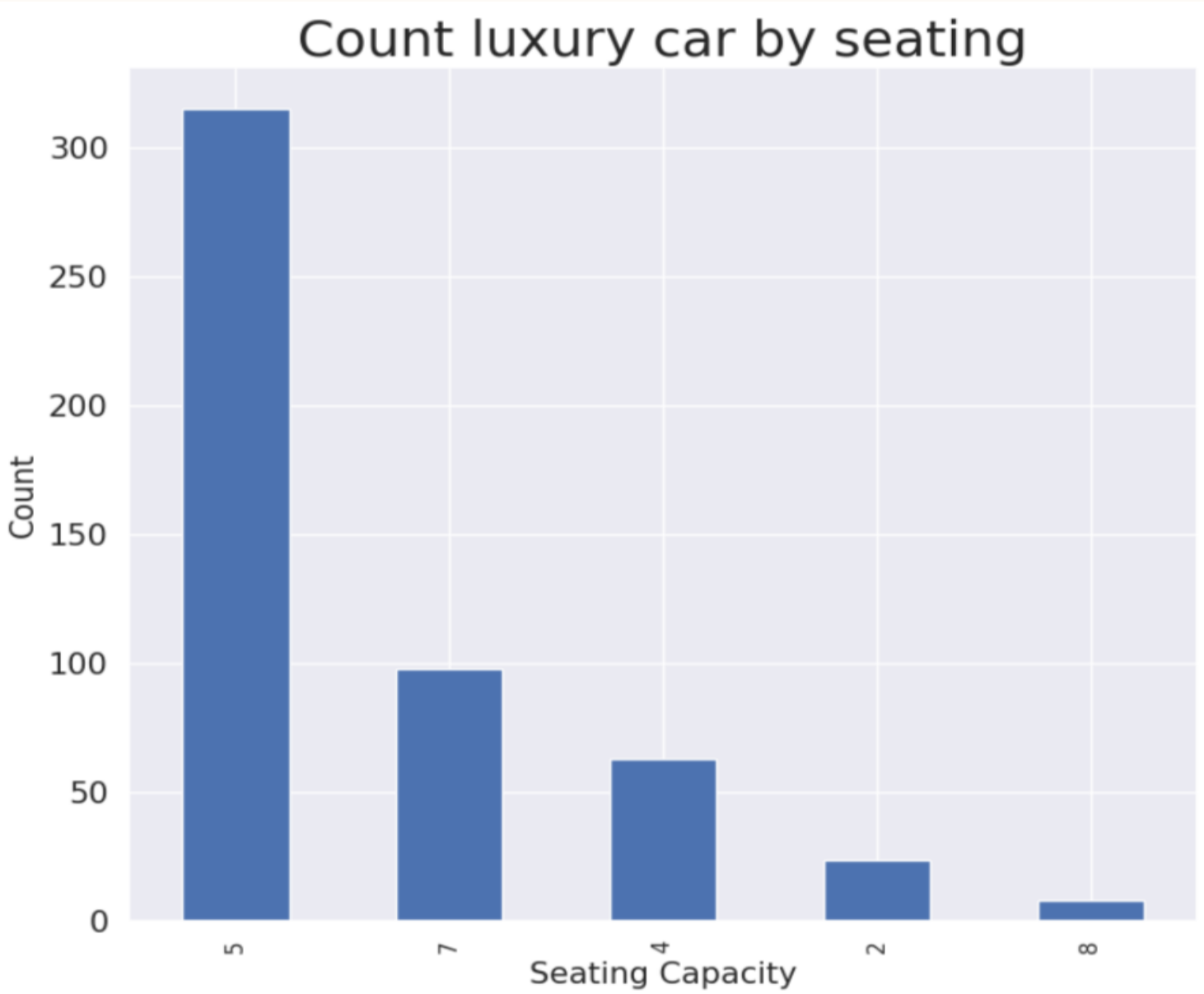
Câu hỏi 4: Với góc nhìn của người bán xe sang, (với xe sang là xe có định giá trên 100k $), tôi muốn biết những chiếc xe sang được sản xuất trong 3 năm gần đây thuộc những thương hiệu nào và phân bố số chỗ ngồi của nó như thế nào?

- Số lượng xe sang tập trung chủ yếu là Toyota và Mercedes - Benz.



TRỰC QUAN HÓA DỮ LIỆU

- Số loại chỗ ngồi nhiều nhất là 5 và 7 chỗ



TRỰC QUAN HÓA DỮ LIỆU

Câu hỏi 5:

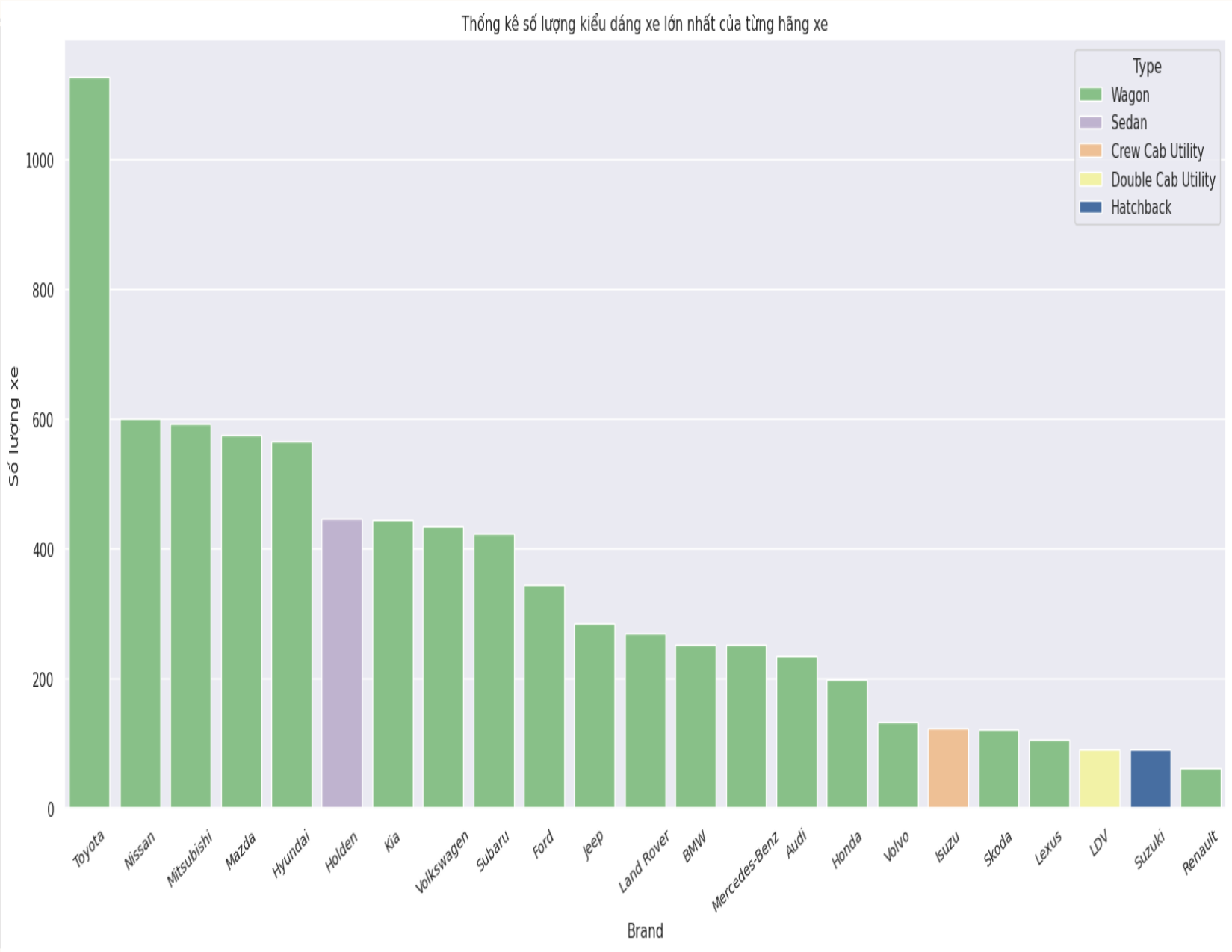
Kiểu dáng xe nào mà người dùng ưa thích nhất của từng hãng xe?

- Kiểu dáng Wagon khá được nhiều người ưa chuộng ở mọi hãng xe.

+Chưa phổ biến tại Việt Nam nhưng ở nước ngoài họ rất ưa chuộng.

+ Khả năng tiết kiệm nhiên liệu khá tốt.

+ Khoang xe rộng rãi.

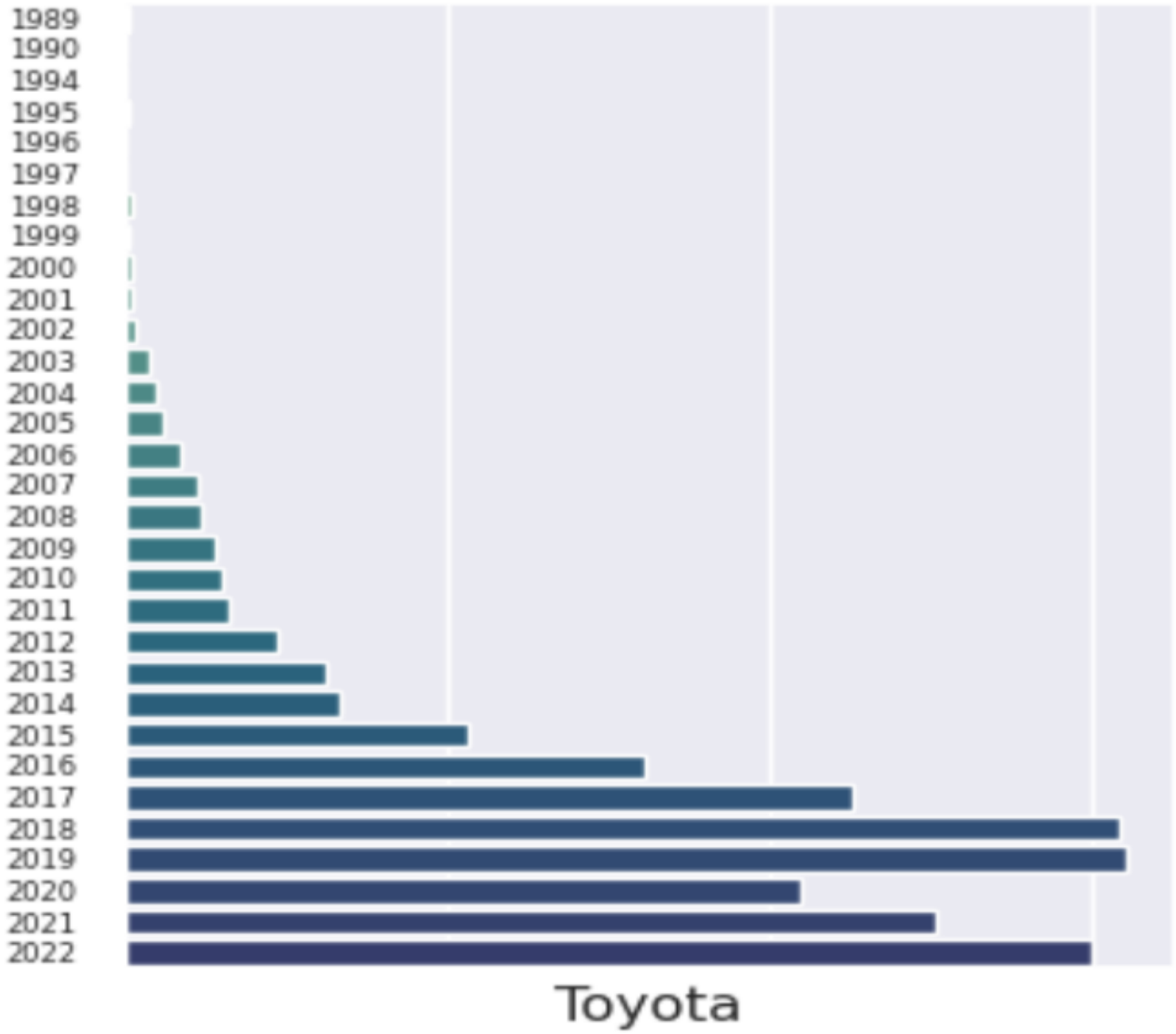


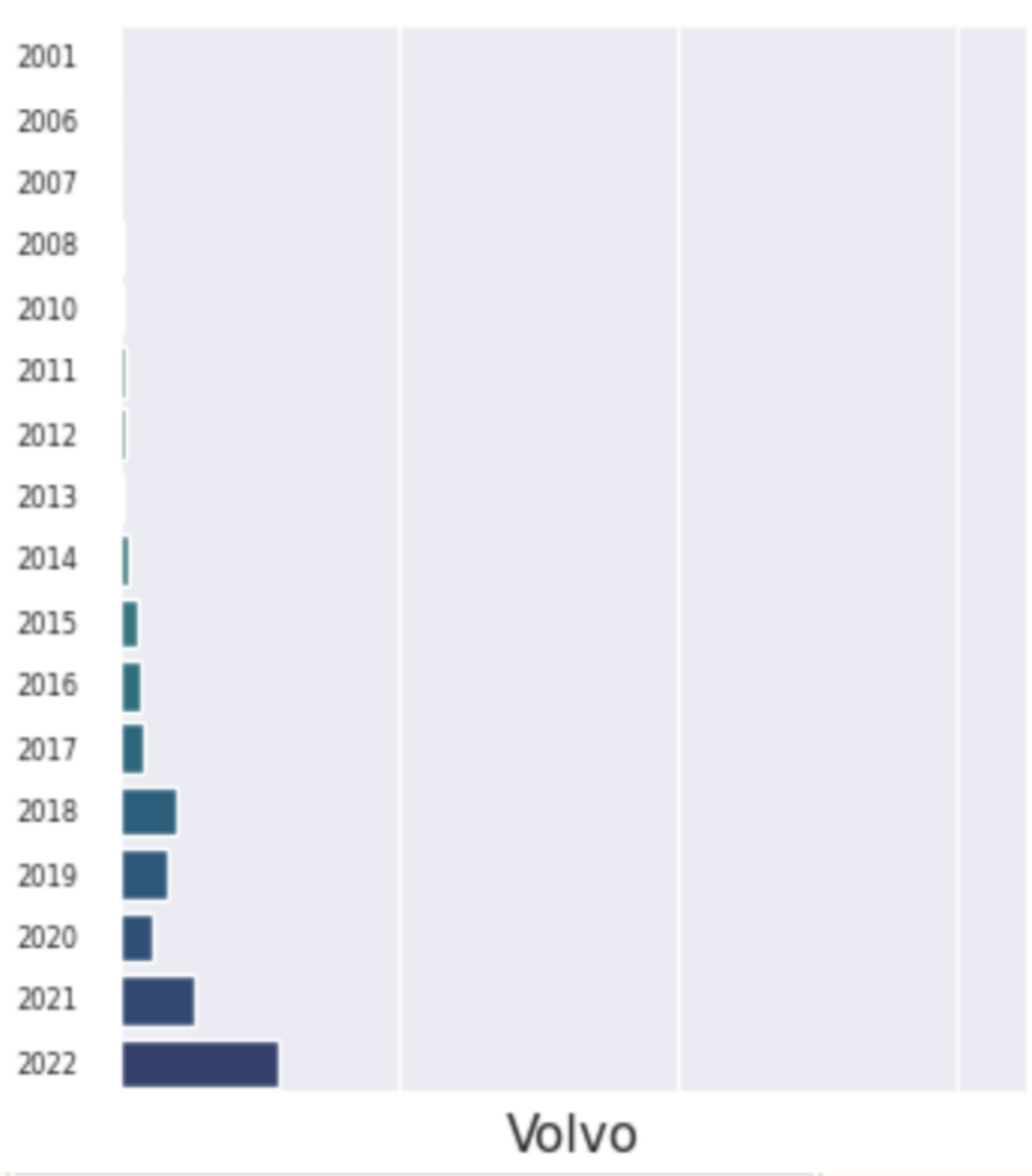
TRỰC QUAN HÓA DỮ LIỆU

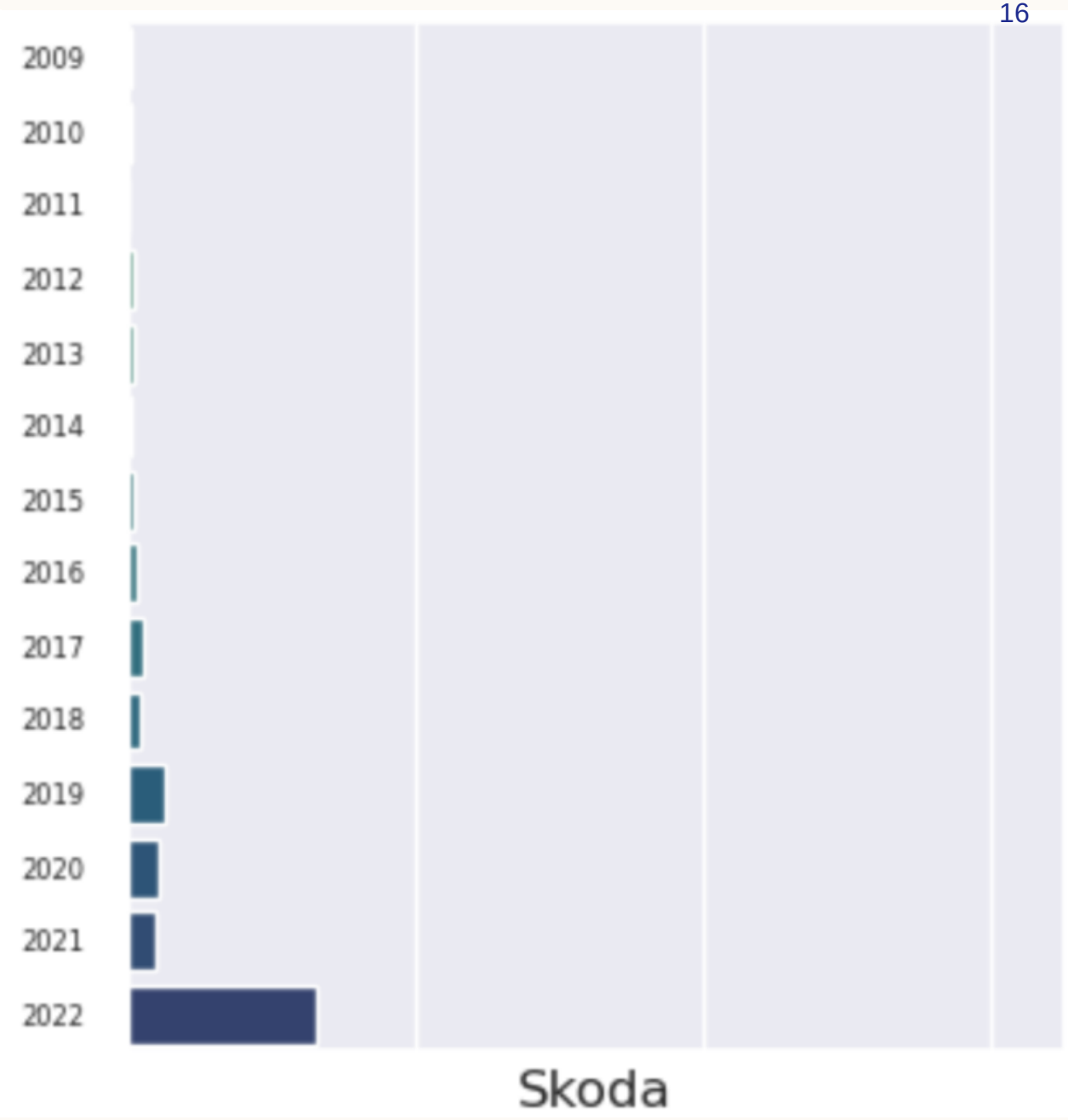
Câu hỏi 6:

Tổng doanh thu của từng hãng xe qua từng năm là bao nhiêu?

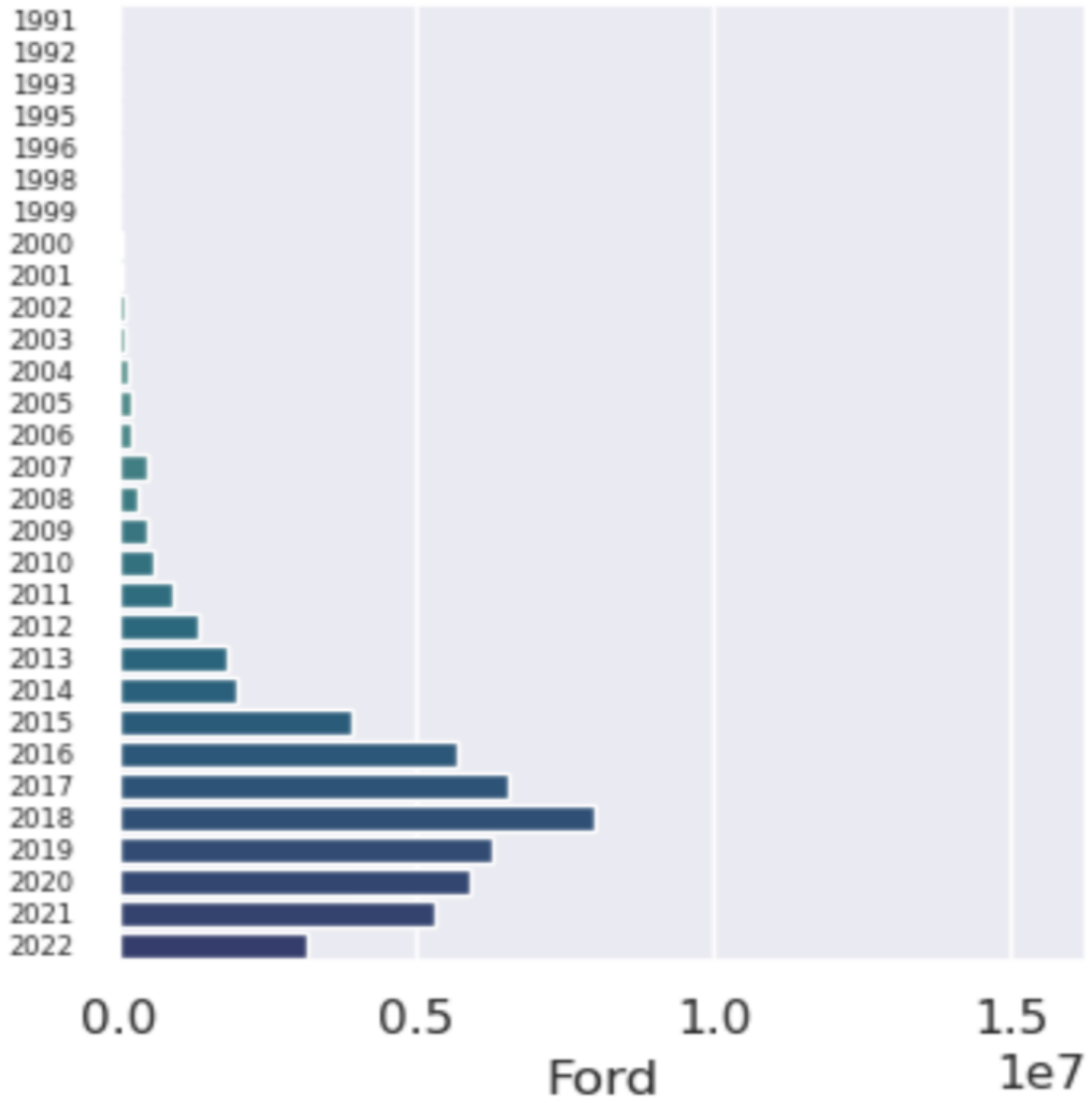
- Toyota có doanh thu cao nhất qua các năm.
- Các hãng như Skoda, Renault, Volvo có doanh thu tăng đáng kể vào năm 2022.
- Nhiều hãng xe có doanh thu cao vào năm (2016-2017) nhưng đến nay nó lại có xu hướng giảm.
- Lexus, Suzuki và Honda có khá ít người bán lại có lẽ ít người dùng hoặc xe vẫn còn tốt nên người dùng không muốn bán.

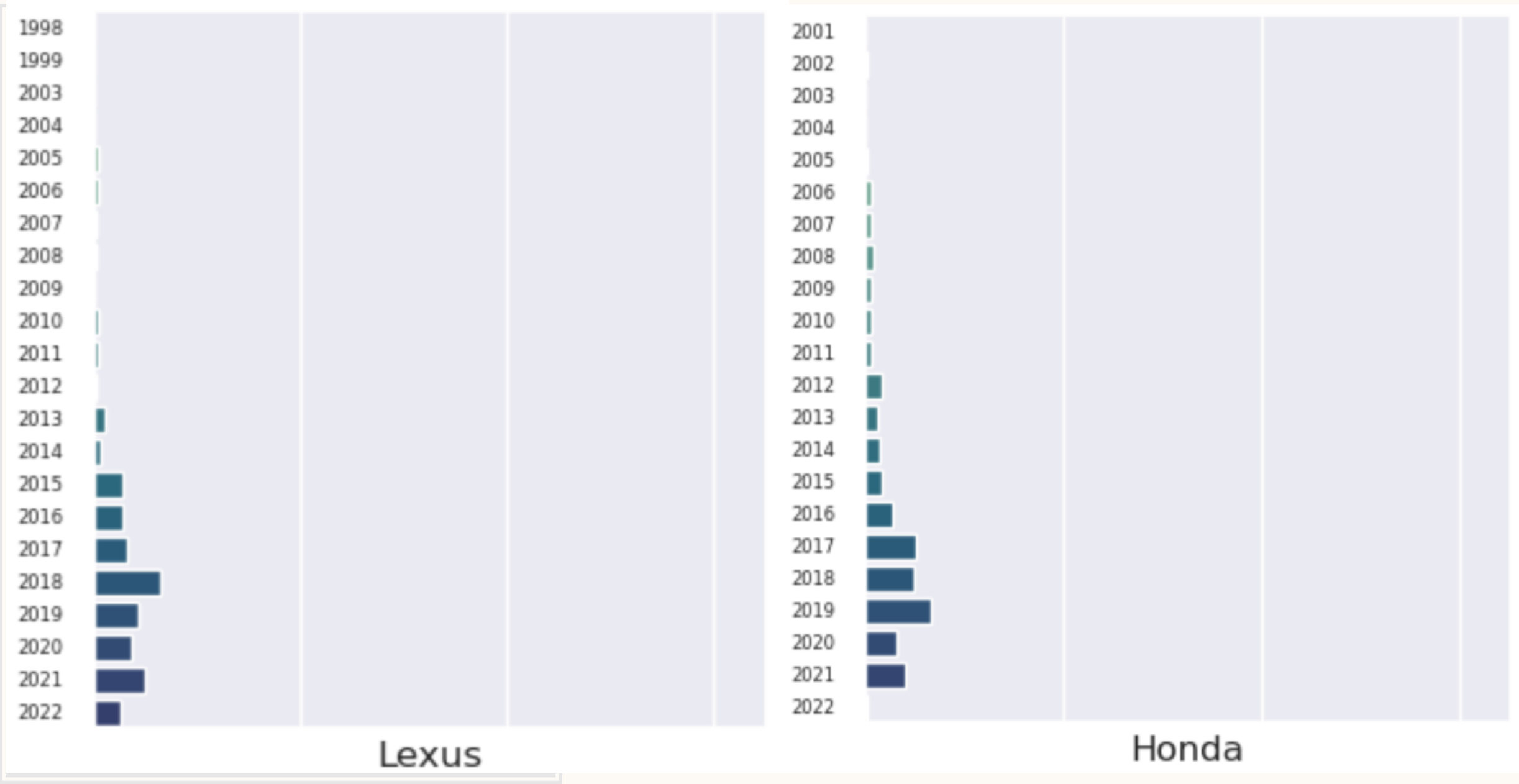



2001
2006
2007
2008
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Volvo


2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Skoda


2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Kia


1991
1992
1993
1995
1996
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
0.0
0.5
1.0
1.5
Ford
1e7


1998
1999
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Lexus
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Honda

MÔ HÌNH HÓA DỮ LIỆU VÀ ĐÁNH GIÁ

Bài toán: Dự đoán giá xe dựa vào các thuộc tính.

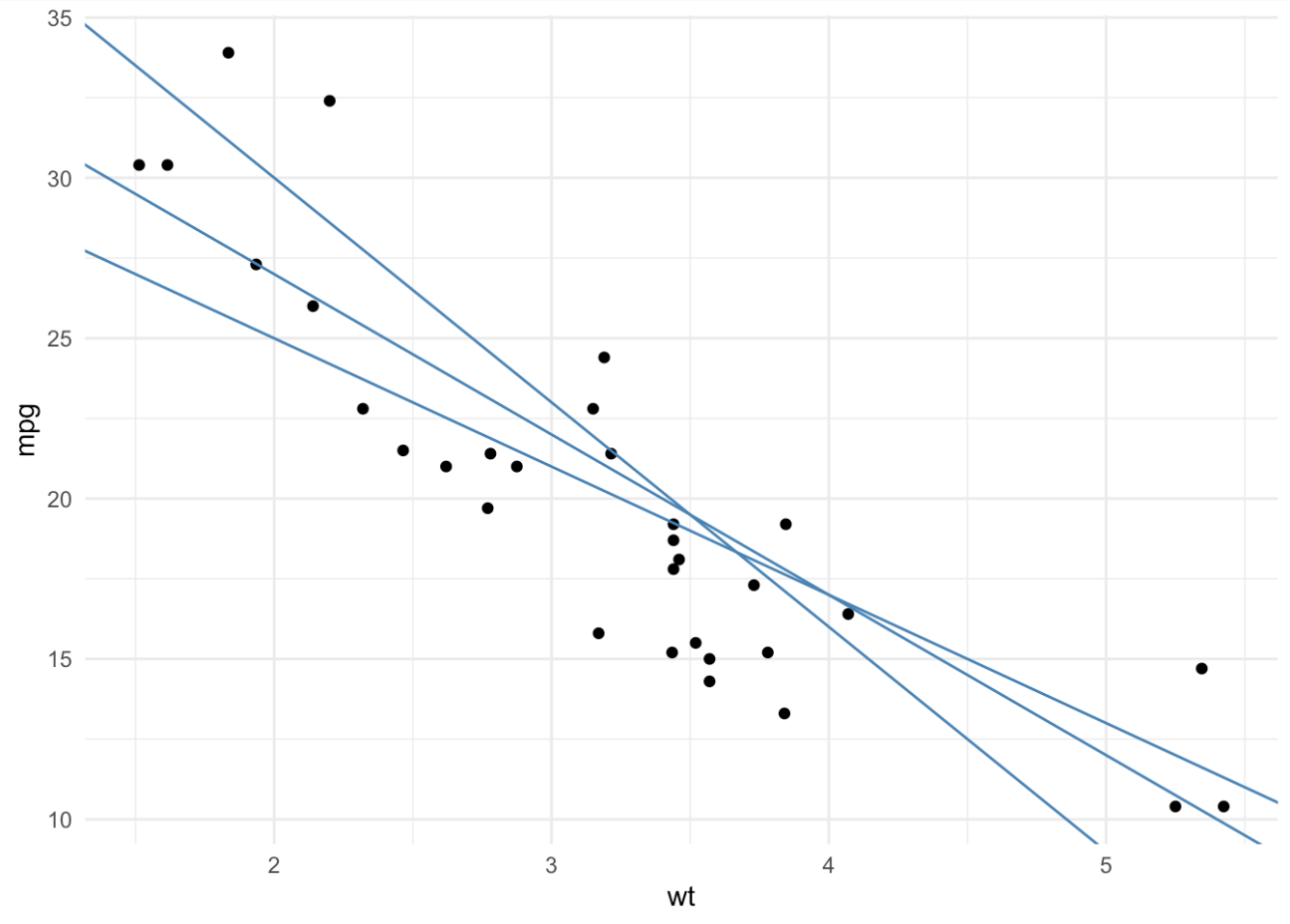
Các mô hình được chọn:

1. Multiple Linear Regression
2. Decision Tree Regressor
3. Random Forest
4. KNeighborsRegression

MÔ HÌNH HÓA DỮ LIỆU VÀ ĐÁNH GIÁ

1. Multiple Linear Regression

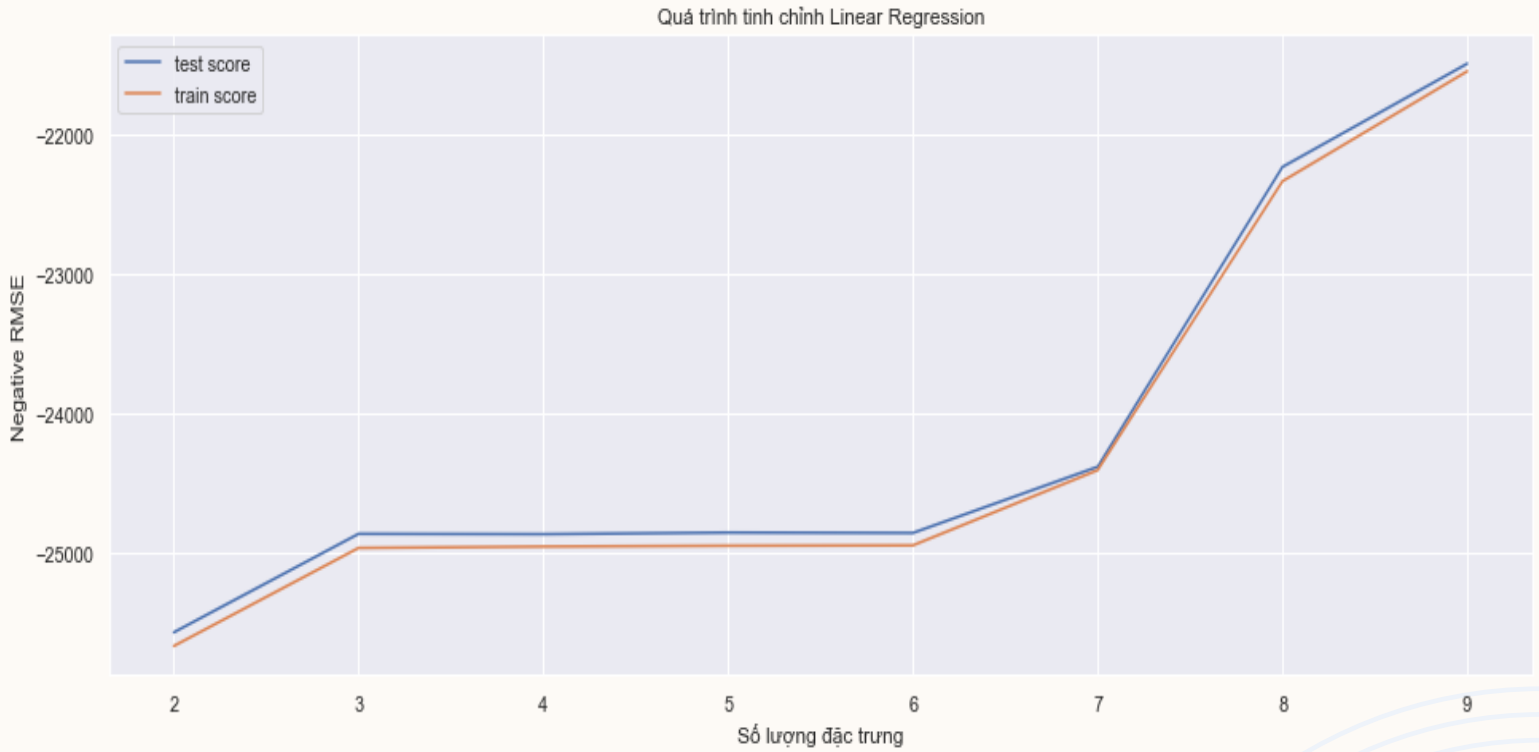
Là một kỹ thuật thống kê có sử dụng một số biến giải thích để dự đoán kết quả của một biến phản ứng. Mục tiêu của hồi quy tuyến tính nhiều (MLR) là để mô hình hóa các mối quan hệ tuyến tính giữa các biến giải thích (độc lập) và phản ứng (phụ thuộc) biến.



MÔ HÌNH HÓA DỮ LIỆU VÀ ĐÁNH GIÁ

- Tinh chỉnh số lượng đặc trưng của mô hình. Số lượng đặc trưng tốt nhất là 9.

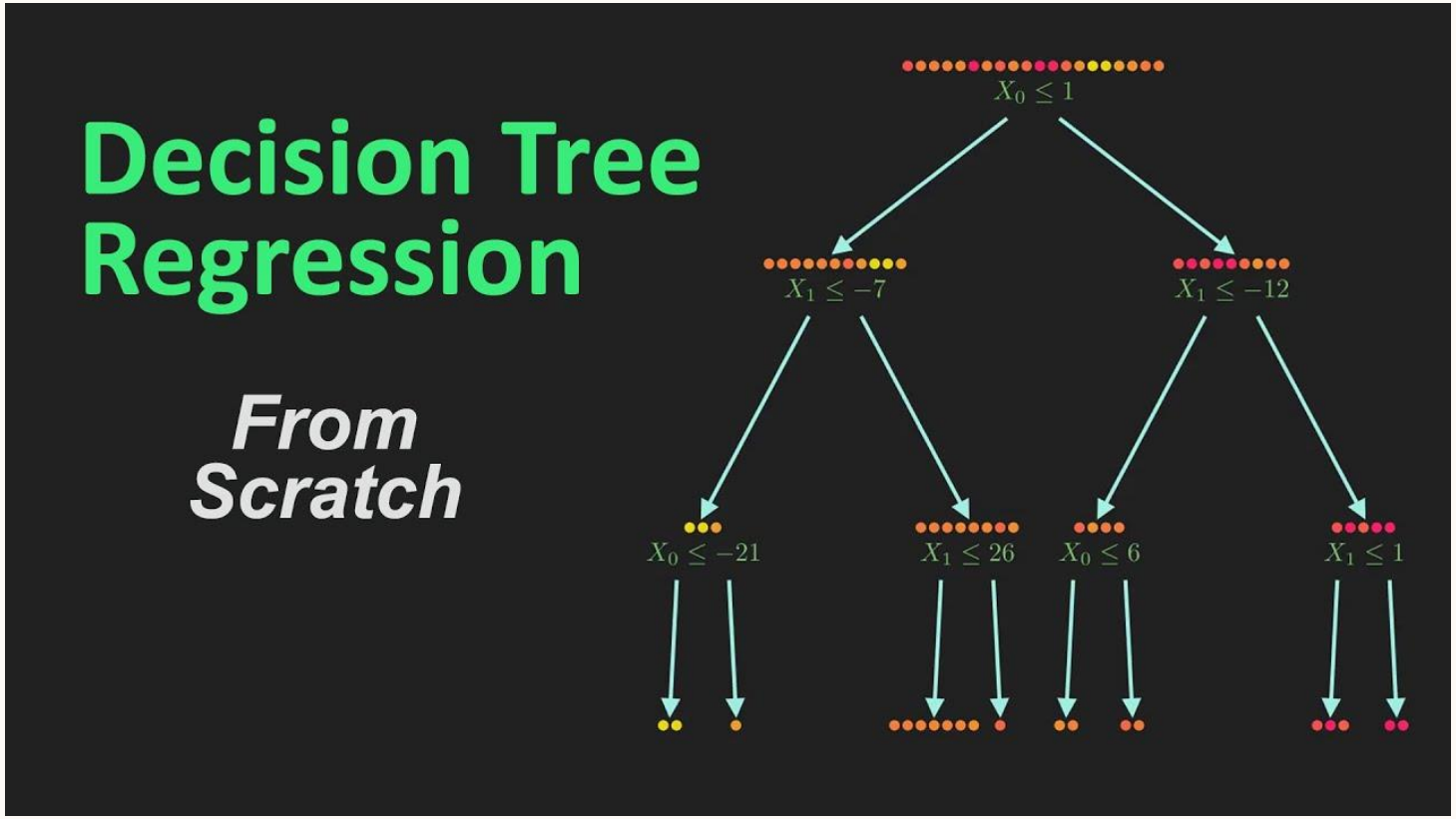
- Sau khi tìm ra được tham số tốt nhất, fit mô hình MLN với tham số đó và dự đoán dựa trên mô hình đã huấn luyện. Sau đó tính RMSE của dự đoán với tập giá trị đúng. Kết quả RMSE khoảng 24261.



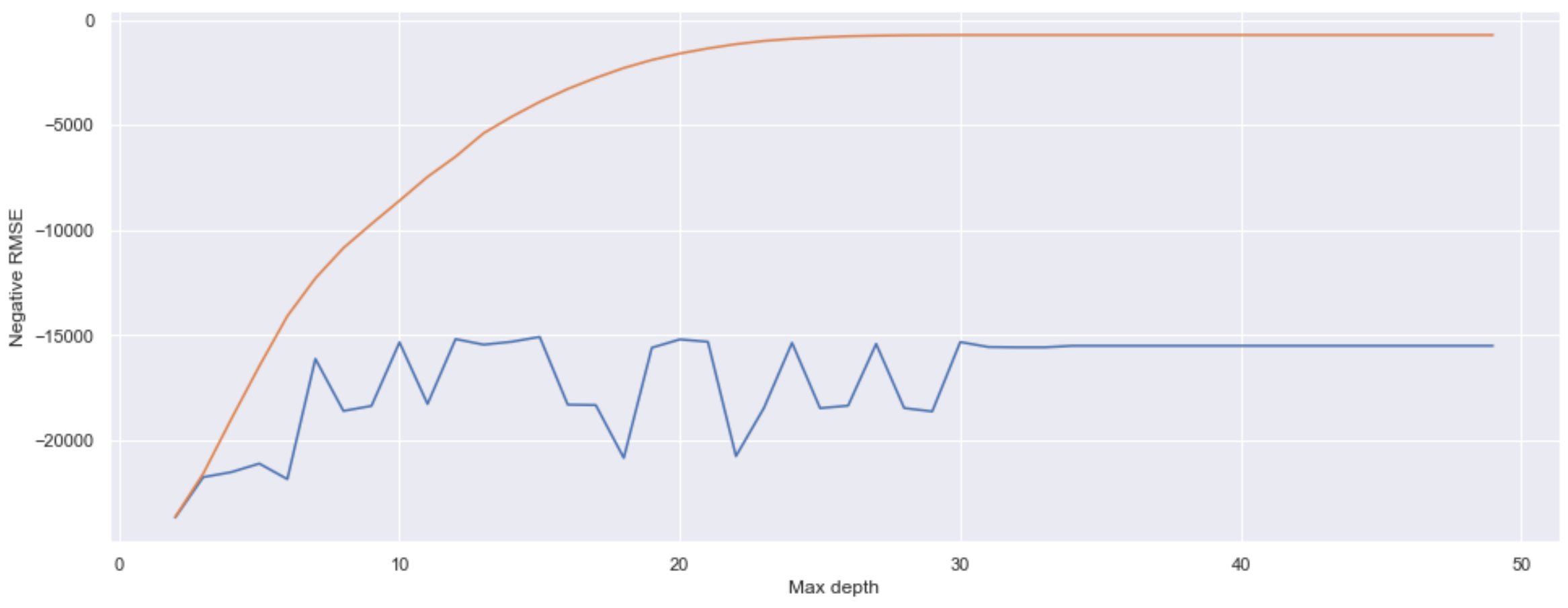
MÔ HÌNH HÓA DỮ LIỆU VÀ ĐÁNH GIÁ

2. DecisionTreeRegressor

Là một mô hình supervised learning, áp dụng vào bài toán regression. Mỗi nút trong (internal node) tương ứng với một biến; đường nối giữa nó với nút con của nó thể hiện một giá trị cụ thể cho biến đó. Mỗi nút lá đại diện cho giá trị dự đoán của biến mục tiêu, cho trước các giá trị của các biến được biểu diễn bởi đường đi từ nút gốc tới nút lá đó.



Đối với mô hình này siêu tham số quan trọng nhất cần tinh chỉnh là "max_depth". Ta dùng k – cross – validation, với k=5. Ta chọn các max_depth trong khoảng từ 2 đến 50. Ta thấy giá trị max_depth tốt nhất nằm ở khoảng 15 (khi Nagative RMSE lớn nhất).



MÔ HÌNH HÓA DỮ LIỆU VÀ ĐÁNH GIÁ

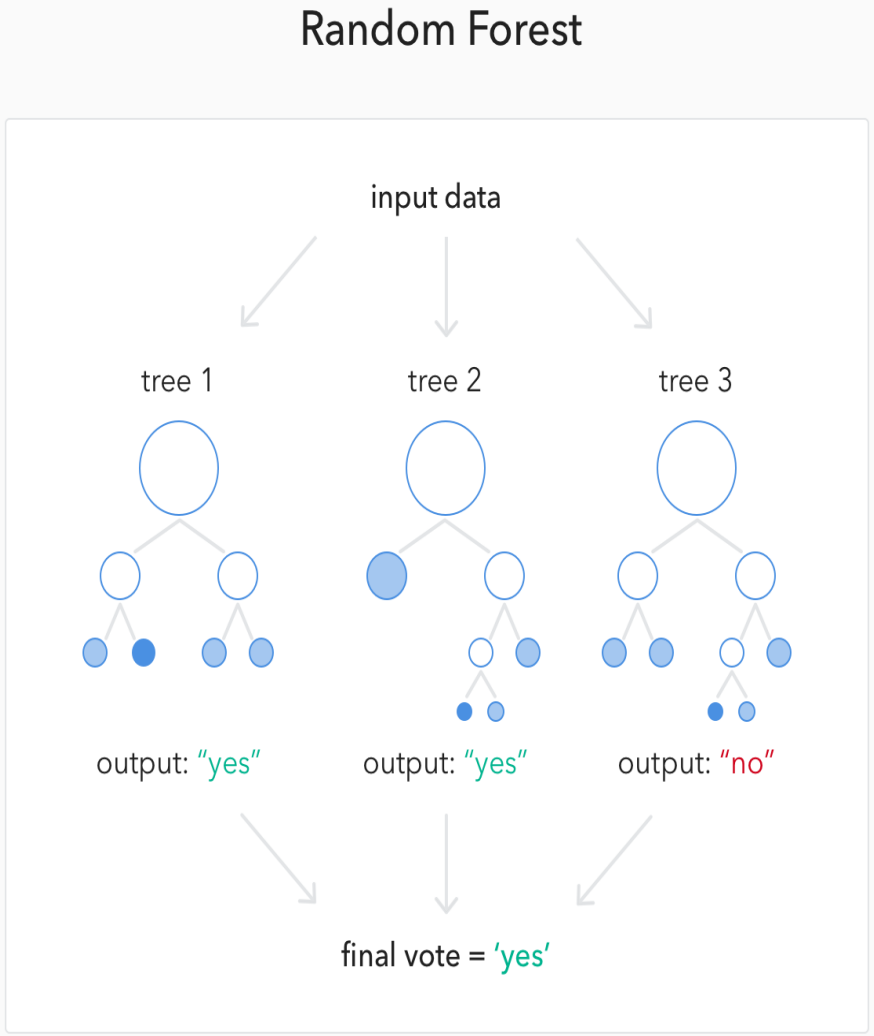
Sau khi chọn được max_depth = 15, truyền giá trị này vào mô hình và huấn luyện. Sau đó tiến hành dự đoán giá xe trong tập test. Ta tính được giá trị rmse vào khoảng 15246.3.

ID	y	y_predict
12774801	49790	62291.0
12688104	60114	53900.0
12834886	14990	12044.0
12688835	28990	34584.0
12747814	26850	38990.0
...	...	...
12788222	10980	10643.0
12793059	20990	19548.0
12849363	28990	28513.0
12803606	27990	26876.0
12829181	26999	19990.0

MÔ HÌNH HÓA DỮ LIỆU VÀ ĐÁNH GIÁ

3. RandomForest

Random Forest xây dựng nhiều cây quyết định bằng Decision Tree, mỗi cây quyết định sẽ khác nhau (có yếu tố random). Sau đó kết quả dự đoán được tổng hợp từ các cây quyết định.

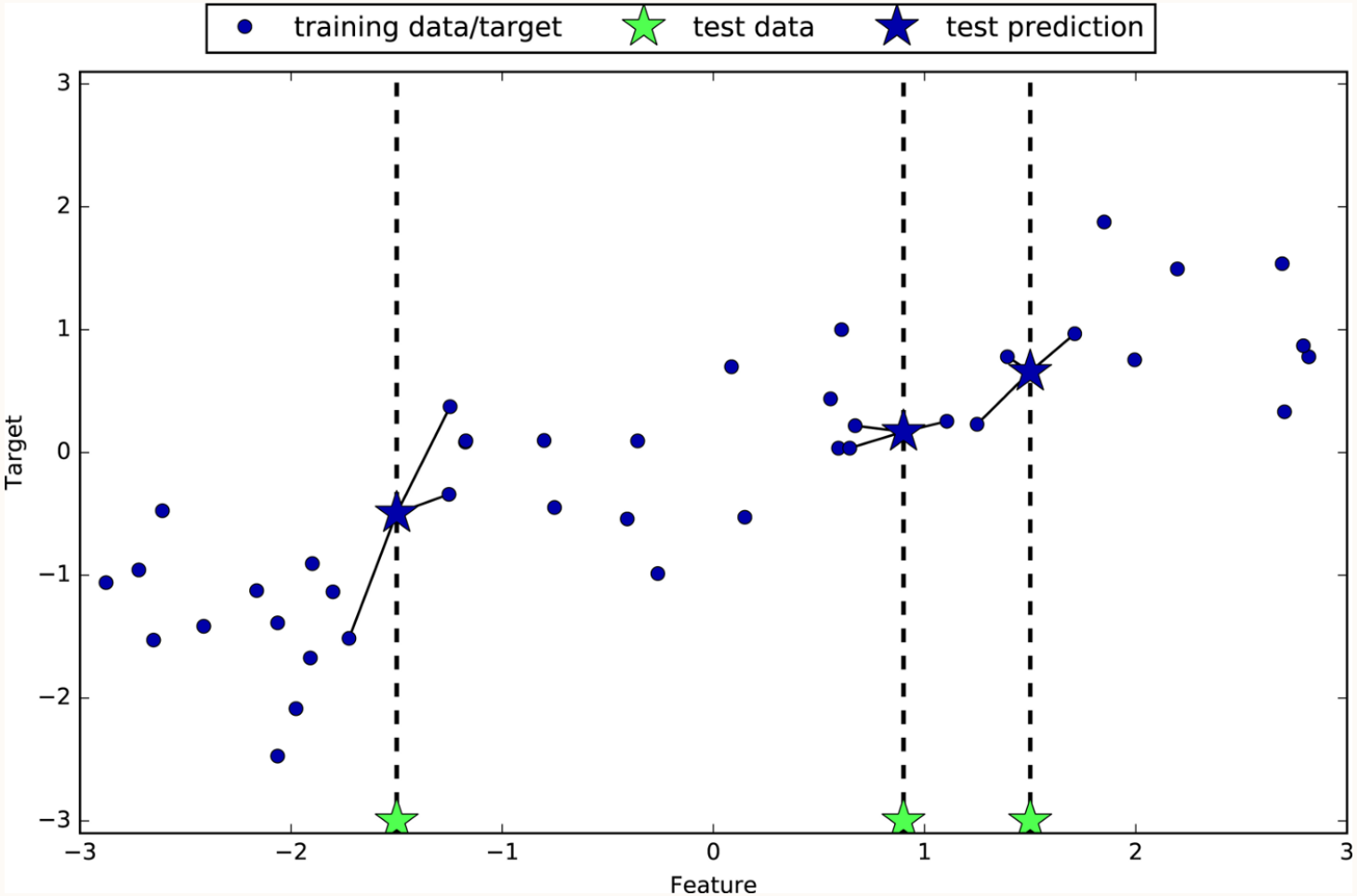


Theo biểu đồ ta có thể thấy chiều sâu lớn nhất của cây tốt nhất với mô hình RFR là 19.
Chúng ta có thể dùng best_params để tìm ra siêu tham số tốt nhất. {'max_depth': 19}
Và ta tính được RMSE: RandomForestRegressor(max_depth=19, random_state=0) :
13123.848014626898

MÔ HÌNH HÓA DỮ LIỆU VÀ ĐÁNH GIÁ

4. KNeighborsRegression

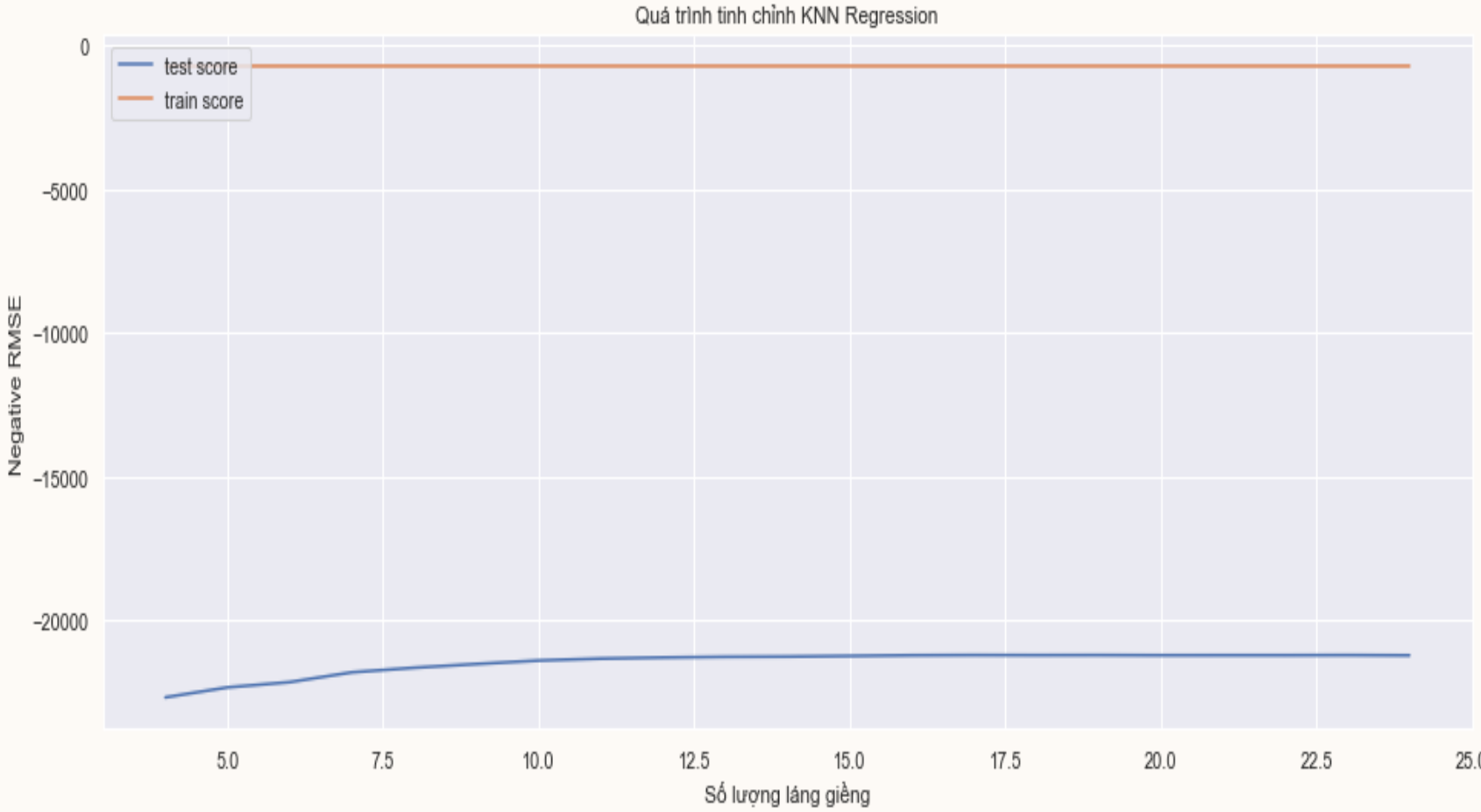
Là thuật toán học máy có giám sát, đơn giản và dễ triển khai. Thường được dùng trong các bài toán phân loại và hồi quy.



MÔ HÌNH HÓA DỮ LIỆU VÀ ĐÁNH GIÁ

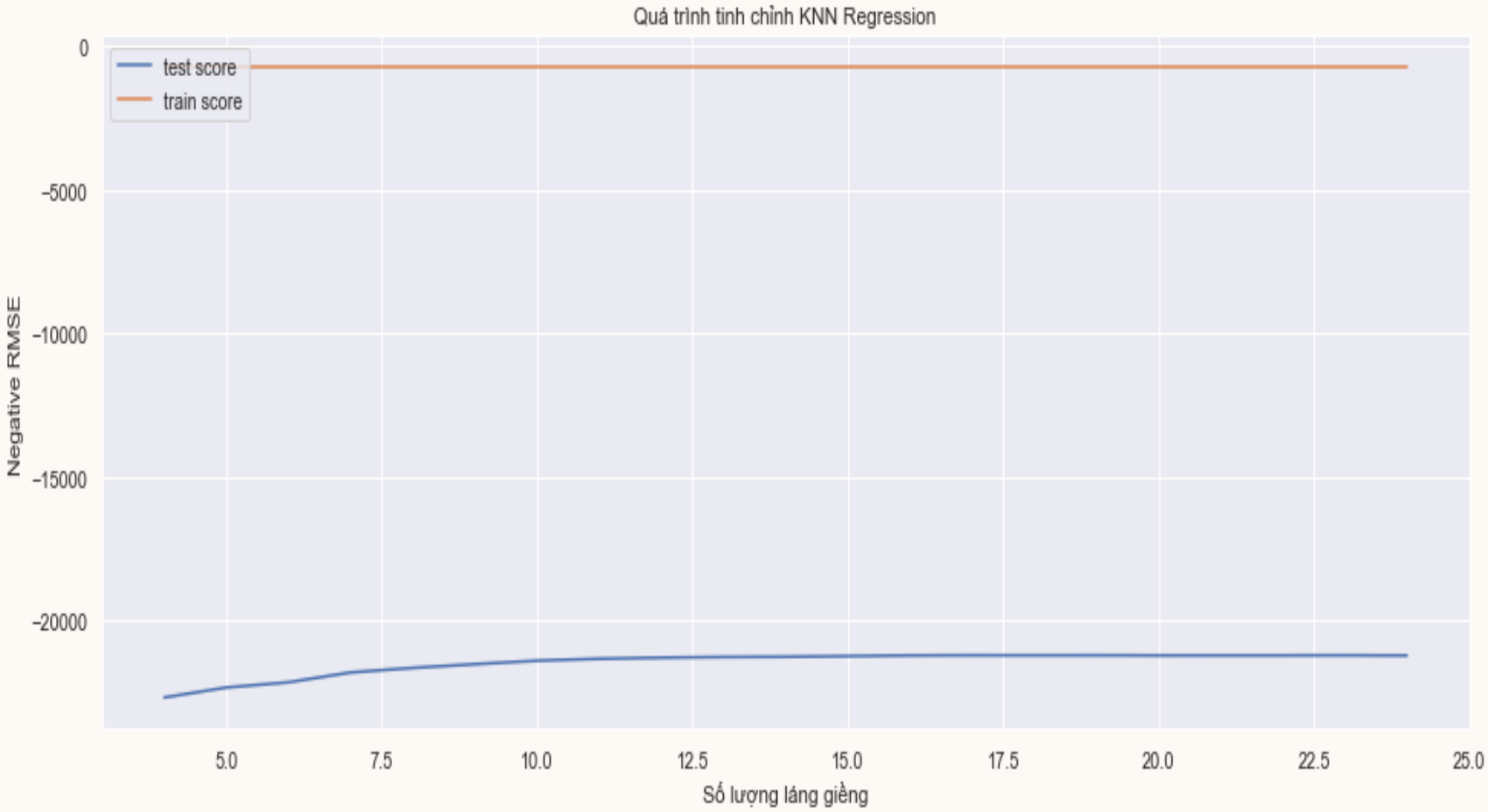
- Với mô hình KNN, siêu tham số quan trọng nhất là số lượng các neighbor.
- Dùng GridSearchCV để tinh chỉnh siêu tham số là số lượng các neighbor (chạy từ 4->24).
- Sử dụng 5 - fold cross validation để đánh giá trên từng fold của tập train. Sử dụng độ đo là Negative RMSE (càng lớn càng tốt).

Ta thấy số lượng neighbor tốt nhất 17. Có thể dùng best_params để tìm ra siêu tham số tốt nhất. Kết quả là 17.



MÔ HÌNH HÓA DỮ LIỆU VÀ ĐÁNH GIÁ

Fit mô hình KNN với tham số đó và dự đoán dưa trên mô hình đã huấn luyện. Sau đó tính RMSE của dự đoán với tập giá trị đúng. Tính toán RMSE của mô hình. Kết quả vào khoảng 23870.



MÔ HÌNH HÓA DỮ LIỆU VÀ ĐÁNH GIÁ

5. Đánh giá bốn mô hình đã làm ở trên

Lưu các kết quả dự đoán của 4 model và kết quả chính xác vào dataframe và file csv.

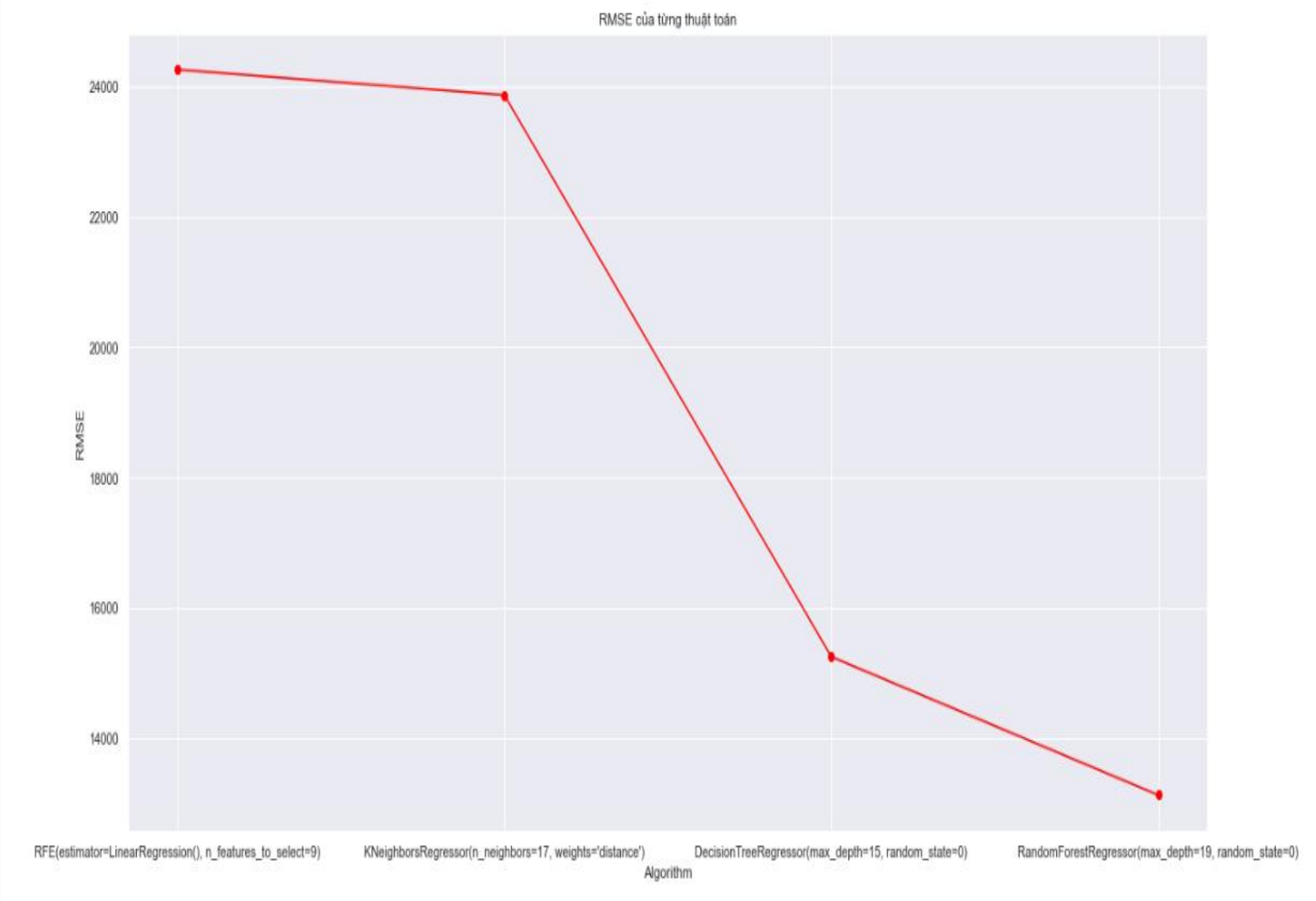
:

ID	y	y_predict_LR	y_predict_KNR	y_predict_DTR	y_predict_RFR
12774801	49790	62494.714215	62020.662438	62291.000000	55080.448333
12688104	60114	55182.158357	44482.459401	53900.000000	54691.150000
12834886	14990	25643.898096	23095.382384	12044.000000	14795.414424
12688835	28990	30216.983090	22741.119730	34584.166667	33281.162973
12747814	26850	28441.351434	19656.415871	38990.000000	29110.430000
...	...	...	...	...	...
12788222	10980	8189.058690	14783.341154	10643.000000	10529.052344
12793059	20990	21154.863459	21590.904043	19548.461538	20475.487487
12849363	28990	36480.016434	36448.858707	28513.250000	28772.658804
12803606	27990	32451.132507	40818.945533	26875.615385	28048.733020
12829181	26999	33813.063013	39904.812771	19990.000000	26943.017358

MÔ HÌNH HÓA DỮ LIỆU VÀ ĐÁNH GIÁ

5. Đánh giá 4 bốn mô hình đã làm ở trên

Vẽ biểu đồ so sánh độ RMSE của 4 thuật toán với tham số tốt nhất. Ta có thể sắp xếp độ "tốt" với tập dữ liệu này lần lượt là mô hình RandomForest, DecisionTreeRegressor, `KNeighborsRegression`, và MultipleLinearRegression.



MÔ HÌNH HÓA DỮ LIỆU VÀ ĐÁNH GIÁ

Ưu, nhược điểm của các mô hình

Multiple Linear Regression

* Ưu điểm:
- Nhanh chóng để mô hình hóa.
- Không cần quá nhiều dữ liệu.
- Đơn giản để hiểu, nó rất có giá trị cho các quyết định kinh. Doanh.

* Nhược điểm:
- Nhạy cảm với nhiễu.
- Không biểu diễn được những mô hình phức tạp.
- Các biến phải độc lập với nhau nhưng trong thực tế rất khó.

MÔ HÌNH HÓA DỮ LIỆU VÀ ĐÁNH GIÁ

Ưu, nhược điểm của các mô hình

Decision Tree Regressor

* Ưu điểm:
- Dễ hiểu, rõ ràng, không cần quá nhiều dữ liệu.
- Có thể xử lý cả dữ liệu có giá trị bằng số và dữ liệu có giá trị là tên thể loại.
- Có thể xử lý một lượng dữ liệu lớn trong thời gian ngắn.
- Dễ chuẩn bị dữ liệu.

* Nhược điểm:
- Không đảm bảo xây dựng được cây tối ưu.
- Dễ dẫn đến Overfitting.

MÔ HÌNH HÓA DỮ LIỆU VÀ ĐÁNH GIÁ

Ưu, nhược điểm của các mô hình

Random Forest

* Ưu điểm:
- Khả năng dự đoán tốt.
- Chuẩn bị và train dữ liệu dễ dàng.
- Phù hợp với dữ liệu lớn.
- Có thể xử lý với các trường dữ liệu bị khuyết.

* Nhược điểm:
- Có thể bị overfitting.
- Không thực sự tối ưu với bài toán hồi quy.
- Nếu dữ liệu có các biến phân loại với các mức thuộc tính khác nhau thì đây có thể là một vấn đề lớn vì thuật toán sẽ ưu tiên những biến có nhiều giá trị hơn, điều này có thể gây ra rủi ro dự đoán.

MÔ HÌNH HÓA DỮ LIỆU VÀ ĐÁNH GIÁ

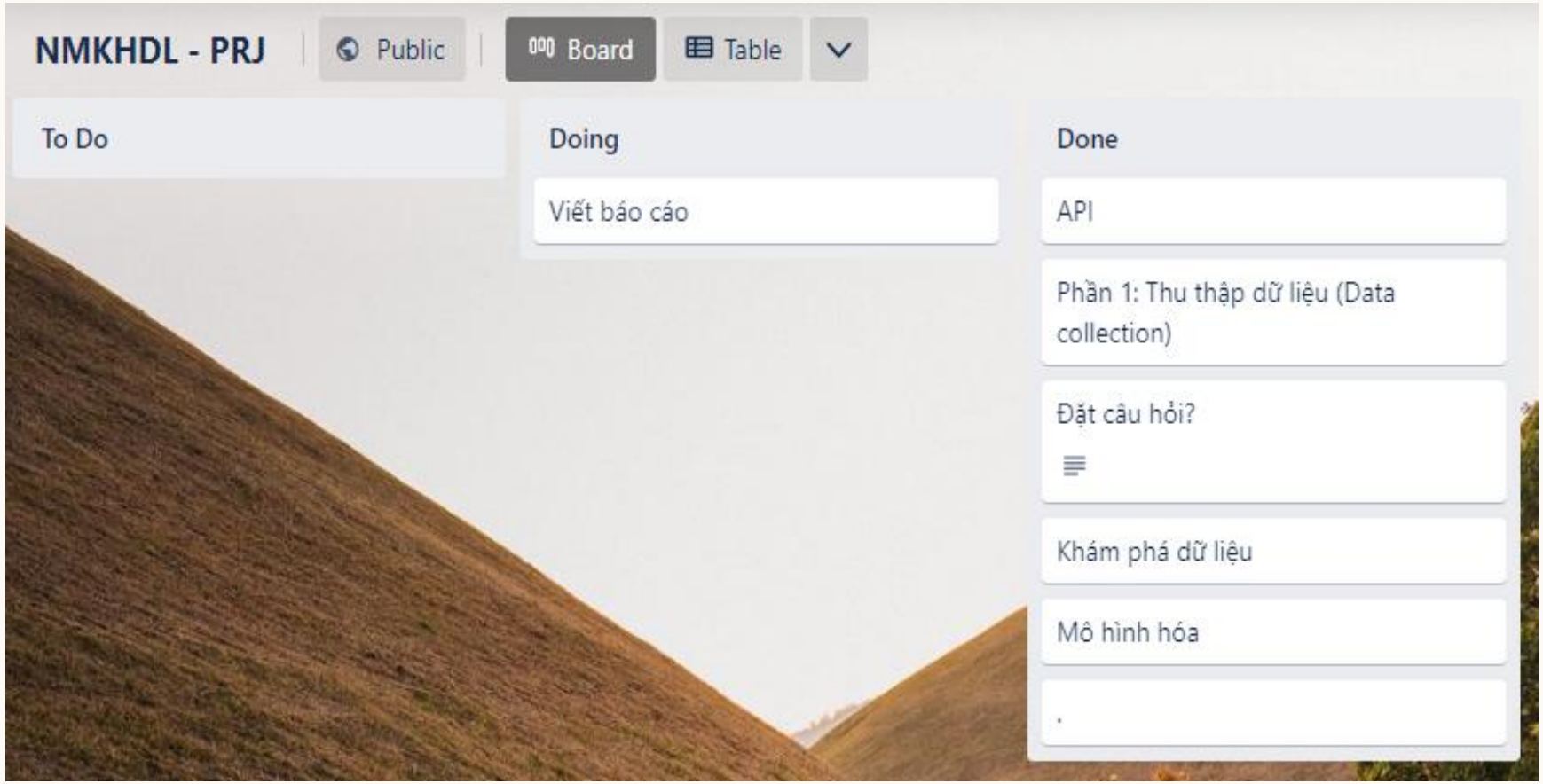
Ưu, nhược điểm của các mô hình

K - Neighbors Regression

* Ưu điểm:
- Trực quan, đơn giản.
- Đơn giản vì không cần tham số.
- Xử lý tốt với dữ liệu nhiễu.
- Có thể đưa ra dự đoán mà không cần thời gian đào tạo.

* Nhược điểm:
- Chậm.
- Mô hình hoạt động tốt với số lượng biến đầu vào nhỏ nhưng khi số lượng biến tăng lên, nó gặp khó khăn trong việc dự đoán đầu ra của điểm dữ liệu mới.
- Khó chọn giá trị K hợp lý.
- Nhạy cảm với nhiễu.
- Không xử lý được các trường dữ liệu bị thiếu.

THEO DÕI QUÁ TRÌNH BẰNG TRELLO



THANK YOU