Link Google Colab: Link

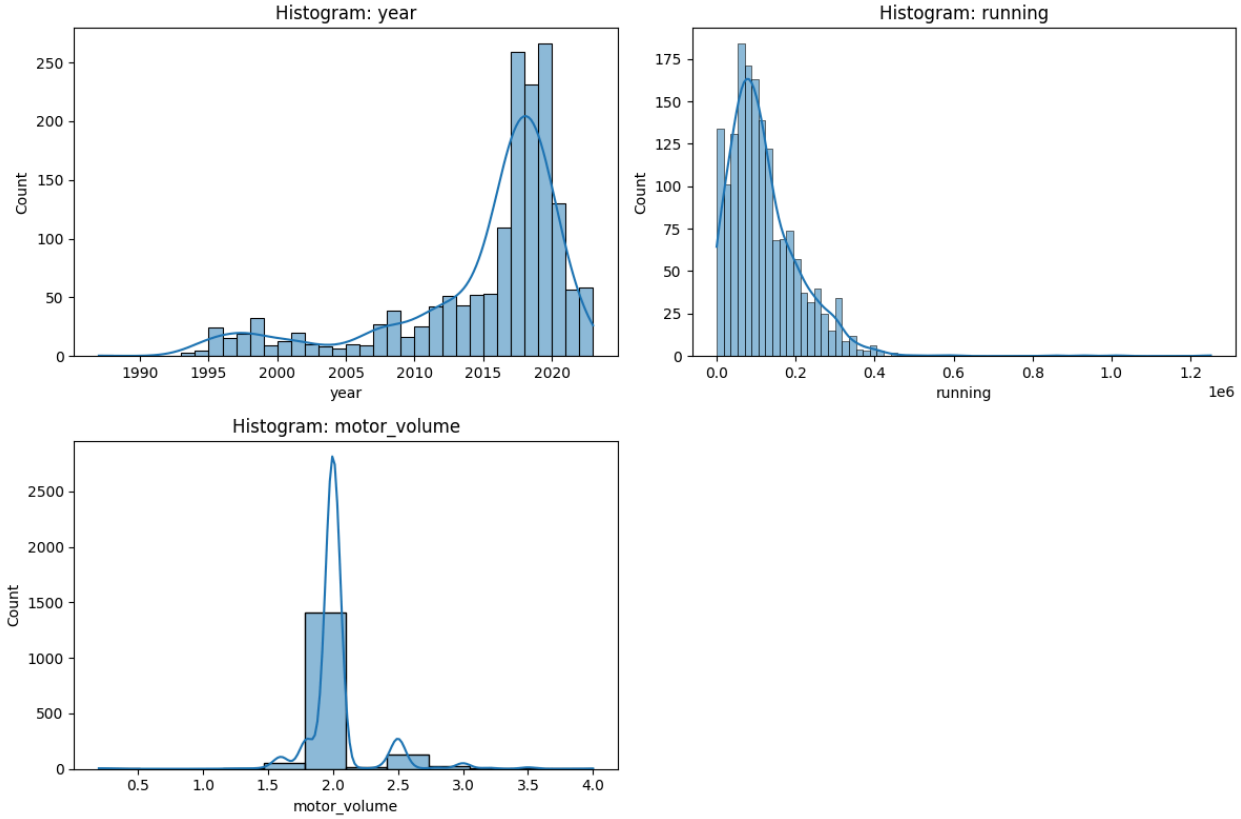
Họ tên	MSSV
Nguyễn Võ Ngọc Bảo	23520131
Vũ Việt Cương	23520213
Ngô Phương Nam	23520974

BÁO CÁO BƯỚC 2 - PHÂN TÍCH DỮ LIỆU VỚI EDA

Môn học: Lập trình Python cho Máy học

1. Phân tích sự phân phối dữ liệu (Data Distribution)

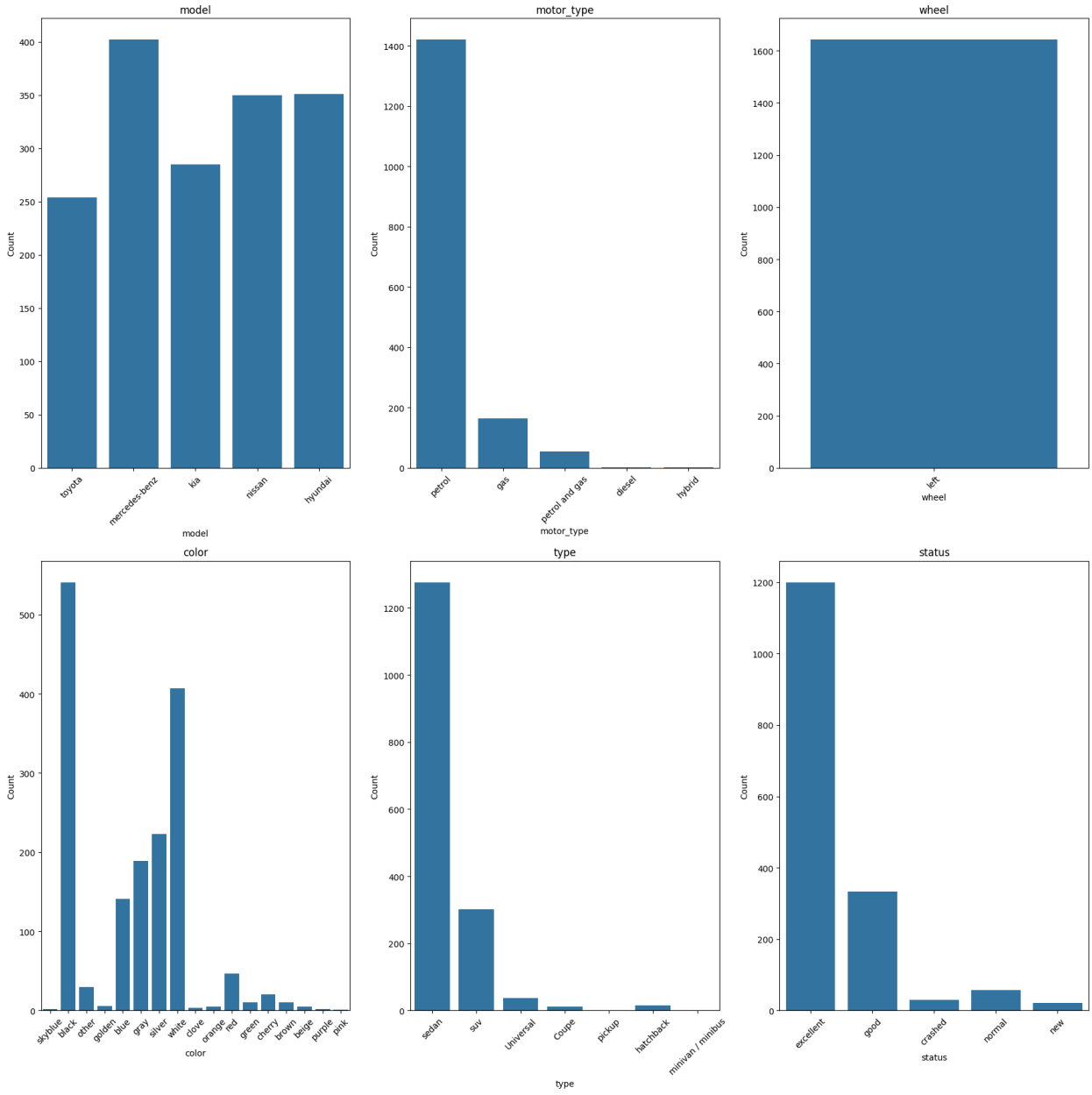
- Numerical Features: year, running, motor_volume



- Đối với biến 'year': phân phối lệch phải nhẹ. Số lượng mẫu giảm dần rõ rệt từ khoảng năm 2005. Giai đoạn trước năm 2000 có rất ít mẫu → Có thể phản ánh rằng xe được sản xuất từ những năm gần đây.
- 'Running': phân phối lệch trái mạnh. Phần lớn xe có số km đã chạy thấp, số xe chạy trên 400000 miles rất ít.
- 'Motor_value': tập trung ở một khoảng giá trị nhất định. Đa phần xe có dung tích xi lanh khoảng 2.0L. Có một số ít xe có dung tích xi lanh lớn hơn 3.0L hoặc nhỏ

hơn 1.5L. Cho thấy rằng thị trường ở đây là các dòng xe phổ thông, không thiên về xe có phân phối lớn.

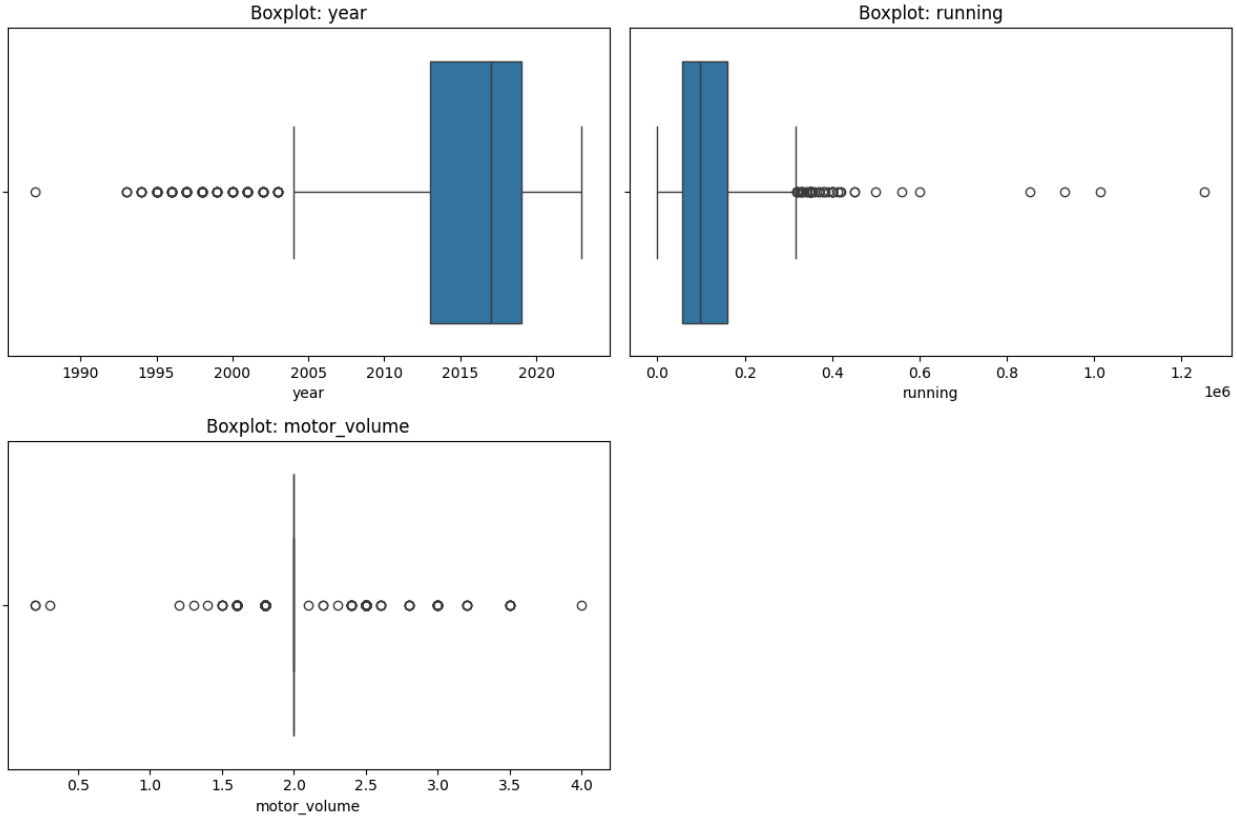
- Categorical features: model, motor_type (loại nhiên liệu), wheel (vị trí tay lái), color, type, status.



- ‘Model’: các hãng xe phổ biến là mercedes-benz với ~400 mẫu, theo sau là kia, nissan và hyundai ~350 mẫu và toyota ít mẫu hơn ~250 mẫu.
- ‘motor_type’: áp đảo là petrol (xăng) với hơn 1400 mẫu. Rất ít xe sử dụng gas, petrol/gas hoặc diesel. Gần như không có xe hybrid.→ Dataset chủ yếu là xe chạy xăng truyền thống.
- ‘wheel’ : tất cả các xe đều có vị trí tay lái ở bên trái.

- ‘Color’: màu xe phổ biến nhất là màu đen (>500) sau đó là white, sliver, gray và blue. Các màu pink, gold, orange, green rất hiếm.
- ‘Type’: chủ yếu là sedan và suv. Các loại xe như universal, coupe, pickup chiếm tỉ lệ rất nhỏ.
- ‘Status’: phần lớn là excellent, good (tình trạng tốt) và new hoặc normal. Một số ít là crashed.

2. Phân tích outlier



- Biến ‘year’: tập trung chủ yếu từ năm 2012 đến 2022. Với outliers là một vài giá trị nằm trước năm 2005, đặc biệt trước 1995.
- Biến ‘running’: phân bố chính dưới 300000. Có nhiều giá trị vượt biên 400000 thậm chí hơn 1000000.
- Biến ‘motor_volumn’: phân bố chính quanh 2.0L. Một số dòng xe có giá trị nhỏ hơn 1.0L hoặc lớn hơn 3.5L.

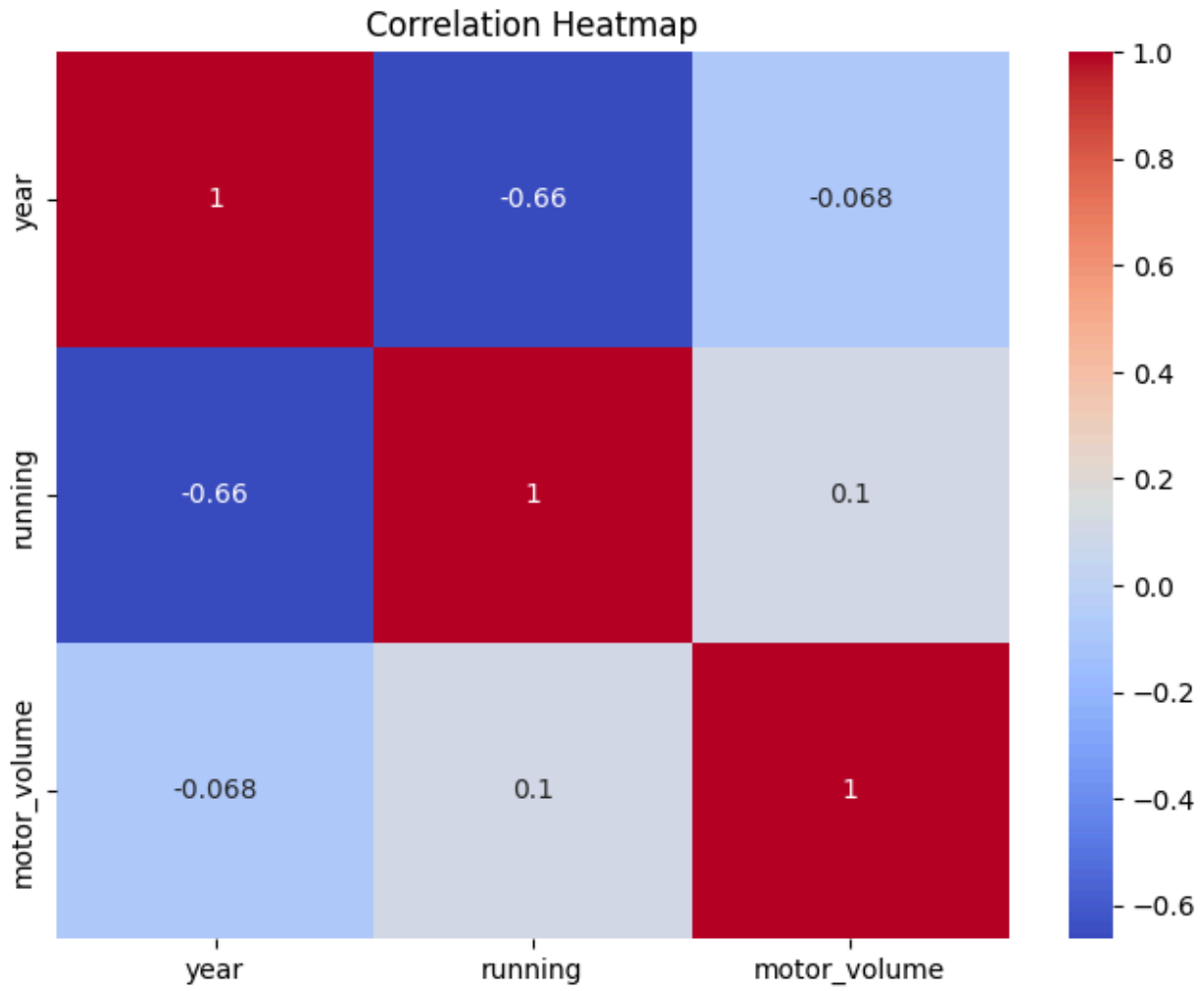
3. Phân tích đơn biến

	year	running	motor_volume	price
count	1642.000000	1.642000e+03	1642.000000	1642.000000
mean	2014.805725	1.192104e+05	2.035018	15982.633374
std	6.587573	9.676625e+04	0.253069	7176.084647
min	1987.000000	1.000000e+01	0.200000	462.000000
25%	2013.000000	5.632690e+04	2.000000	12000.000000
50%	2017.000000	9.878604e+04	2.000000	15750.000000
75%	2019.000000	1.609139e+05	2.000000	18500.000000
max	2023.000000	1.251708e+06	4.000000	87000.000000

Số lượng: 1642 mẫu cho mỗi biến.
Năm sản xuất: Trung bình 2014.8 (min 1987, max 2023).
Quãng đường: Trung bình 119,210 km (min 10, max 1,251,708). Phân tán lớn.
Dung tích động cơ: Trung bình 2.035 lít (min 0.2, max 4.0). Tập trung ở 2.0 lít.
Giá: Trung bình 15,982.63 (min 462, max 87,000). Có sự phân tán.

4. Phân tích đa biến

4.1. Feature - Feature



- Tương quan giữa year và running: Hệ số tương quan là -0.66. Điều này cho thấy có một mối tương quan nghịch mạnh giữa năm sản xuất và quãng đường đã đi. Nói cách khác, những xe có năm sản xuất gần đây (xe mới hơn) thường có quãng đường đã đi ít hơn, và ngược lại, những xe có năm sản xuất cũ hơn thường có quãng đường đã đi nhiều hơn

→ Có mối quan hệ tuyến tính nghịch khá mạnh.

- Tương quan giữa year và motor_volume: Hệ số tương quan là -0.068. Giá trị này rất gần với 0, cho thấy có một mối tương quan rất yếu hoặc hầu như không có mối tương quan tuyến tính đáng kể nào giữa năm sản xuất và dung tích động cơ. Năm sản xuất của xe dường như không liên quan đến việc xe đó có dung tích động cơ lớn hay nhỏ.

→ Hầu như không có mối quan hệ tuyến tính.

- Tương quan giữa running và motor_volume: Hệ số tương quan là 0.1. Điều này cho thấy có một mối tương quan thuận yếu giữa quãng đường đã đi và dung tích động cơ. Có một xu hướng nhỏ là những xe có dung tích động cơ lớn hơn có thể đi được quãng đường dài hơn một chút, nhưng mối quan hệ này không mạnh mẽ.

→ Có mối quan hệ tuyến tính thuận rất yếu.

4.2. Feature - Target

- Biến year so với price:

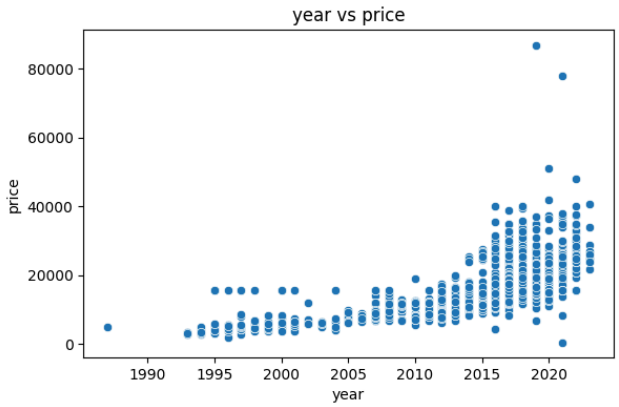
- Xu hướng chung: Có một xu hướng tăng lên rõ rệt của giá xe theo thời gian. Các xe đời cũ (trước năm 2000) thường có giá thấp hơn đáng kể so với các xe đời mới (sau năm 2010).
- Độ phân tán: Giá xe có xu hướng phân tán rộng hơn ở những năm gần đây. Điều này có thể là do sự đa dạng về mẫu mã, tính năng và phân khúc giá của các xe mới.
- Điểm nổi bật: Có một số xe đời rất cũ (trước năm 1995) vẫn có giá khá cao, có thể là các dòng xe cổ hoặc có giá trị sưu tầm. Từ khoảng năm 2015 trở đi, có vẻ như xuất hiện nhiều xe có giá cao đột biến

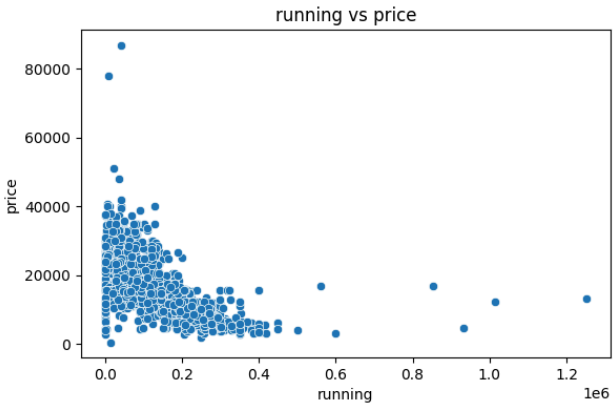
- Biến running (quãng đường đã đi) so với price:

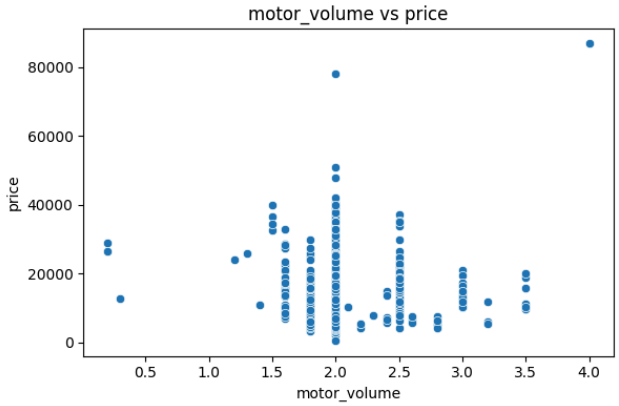
- Xu hướng chung: Nhìn chung, giá xe có xu hướng giảm khi quãng đường đã đi tăng lên. Các xe có quãng đường đi rất ít thường có giá cao hơn.
- Độ phân tán: Độ phân tán của giá khá lớn ở các mức quãng đường đã đi khác nhau, đặc biệt là ở quãng đường đi thấp. Điều này cho thấy quãng đường đã đi không phải là yếu tố duy nhất quyết định giá xe.
- Điểm nổi bật: Có một số xe đã đi một quãng đường rất dài (trên 1 triệu km) vẫn có giá đáng kể, có thể là do chất lượng bảo dưỡng tốt hoặc các yếu tố khác.

- Biến motor_volume (dung tích động cơ) so với price:

- Xu hướng chung: Mối quan hệ giữa dung tích động cơ và giá xe có vẻ phức tạp hơn và không có xu hướng tuyến tính rõ ràng như hai biến trên.
- Tập trung: Có vẻ như phần lớn các xe tập trung ở dải dung tích động cơ từ khoảng 1.0 đến 2.5 lít.
- Giá cao: Một số xe có dung tích động cơ nhỏ (dưới 1.0 lít) và một số xe có dung tích động cơ lớn (trên 3.0 lít) có thể có giá cao, nhưng số lượng không nhiều bằng các xe có dung tích trung bình. Điều này có thể phản ánh sự phổ biến của các dòng xe hạng trung.
- Độ phân tán: Giá xe phân tán khá rộng ở các mức dung tích động cơ khác nhau, cho thấy dung tích động cơ không phải là yếu tố duy nhất quyết định giá.


year vs price
price
year
0
20000
40000
60000
80000
1990
1995
2000
2005
2010
2015
2020


running vs price
price
running
1e6
0
20000
40000
60000
80000
0.0
0.2
0.4
0.6
0.8
1.0
1.2


motor_volume vs price
price
motor_volume
0
20000
40000
60000
80000
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0