TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA TOÁN – CƠ – TIN HỌC

TÊN ĐỀ TÀI:
DỰ ĐOÁN GIÁ Ô TÔ ĐÃ QUA SỬ DỤNG

BÀI TIỂU LUẬN KẾT THÚC HỌC PHẦN

Học phần: Machine Learning and Data Science

Giảng viên hướng dẫn: TS.Nguyễn Thị Minh Huyền

Thực hiện : Phạm Quang Hiếu

MỤC LỤC

GIỚI THIỆU VÀ MỤC TIÊU BÀI TOÁN.....6
1. Xác định tập dữ liệu.....6
1.1. Xác định các biến trong bộ dữ liệu.....6
1.2. Khai phá và xử lý dữ liệu missing.....7
1.2.1. Kiểm tra các quan sát đầu và cuối trong bộ dữ liệu.....8
1.2.2. Kiểm tra và xử lý các giá trị null (rỗng) trong bộ dữ liệu.....9
2. Phân tích biến đơn và xử lý dữ liệu.....11
2.1. Biến địa điểm.....11
2.2. Năm sản xuất (year).....13
2.3. Số km đã chạy (Kilometers_Driven).....14
2.5. Hộp số (Transmission).....15
2.6.Chủ sở hữu (Owner_Type).....16
2.7. Mức tiêu thụ nhiên liệu (Mileage).....17
2.9. Công suất của động cơ (Power).....18
2.10. Số ghế (Seats).....19
2.11. Nhãn hiệu xe (name).....19
2.12. Biến giá (Price).....20
3. Phân tích sự tương quan giữa biến giải thích (x) và biến phụ thuộc (y).....21
3.1.Tương quan giữa biến giá và các biến định tính.....21
3.2. Tương quan giữa biến giá và các biến định lượng.....21
4. Phân cụm dữ liệu.....24
5. Đánh giá tương quan giữa các cụm dữ liệu.....26
6. Lựa chọn đặc trưng cho mô hình đầy đủ.....27
7. Lựa chọn mô hình học máy để áp dụng cho bộ dữ liệu.....29
7.1. Mô hình hồi quy hồi quy tuyến tính (linear regression).....29
7.2. Mô hình Random Forest.....30
8. Áp dụng mô hình vào bộ dữ liệu.....31

8.1. Mô hình hồi quy 31
8.2. Mô hình Random forest 33
9. Áp dụng mô hình vào bộ dữ liệu với biến Giá được đưa về phân phối chuẩn 33
9.1. Mô hình hồi quy 33
9.2. Mô hình Random forest 34
10. Kết luận 34
TÀI LIỆU THAM KHẢO 35

DANH MỤC HÌNH ẢNH

Hình 1 – Tổng quan về bộ dữ liệu 7
Hình 2 – Kiểm tra các quan sát đầu và cuối trong bộ dữ liệu 8
Hình 3 – Kiểm tra các giá trị null (rỗng) trong bộ dữ liệu 9
Hình 4 – Kiểm tra các giá trị null trong bộ dữ liệu sau khi đã xử lý 10
Hình 5 – Tách đơn vị đo và chuyển dữ liệu về dạng số 10
Hình 6 – Biểu đồ biểu diễn địa điểm phân bố xe cũ 11
Hình 7 – Các thành phố có lượng xe cũ từ cao xuống thấp 11
Hình 8 – Biểu đồ biểu diễn phân phối xe cũ theo năm 13
Hình 9 – Biểu đồ biểu diễn phân phối số km đã chạy 14
Hình 10 – Biểu đồ biểu diễn phân phối loại nhiên liệu 15
Hình 11 – Biểu đồ biểu diễn phân phối loại hộp số 16
Hình 12 – Biểu đồ biểu diễn phân phối về chủ sở hữu 16
Hình 13 – Biểu đồ biểu diễn phân phối mức tiêu thụ nhiên liệu 17
Hình 14 – Biểu đồ biểu diễn phân phối dung tích động cơ 18
Hình 15 – Biểu đồ biểu diễn phân phối công suất của động cơ 18
Hình 16 – Biểu đồ biểu diễn phân phối số ghế xe 19
Hình 17 – Biểu đồ biểu diễn phân phối về nhãn hiệu xe 19
Hình 18_1 – Biểu đồ của biến giá (lệch phải) 20
Hình 18_2 – Biểu đồ của biến giá (chuẩn hóa) 20
Hình 19 – Bảng ANOVA phân tích tương quan giữa biến giá và các biến định tính 21
Hình 20 – Biểu đồ phân tích tương quan giữa biến giá và các biến định lượng 21
Hình 21 – Biểu đồ phân tích tương quan giữa biến Price và Engine 22
Hình 22 – Biểu đồ phân tích tương quan giữa biến Price và Power 23
Hình 23 – Biểu đồ phân tích tương quan giữa biến Price và biến Mileage 23
Hình 24 – Biểu đồ elbow point plot 25
Hình 25 – Biểu đồ phân cụm phân cấp 26
Hình 26 – Đánh giá tương quan giữa các cụm dữ liệu 26

Hình 27 - Bảng ANOVA phân tích tương quan giữa các cụm dữ liệu 27

Hình 28 - Mô hình Linear Regression 30

Hình 29 - Mô hình Random forest 31

Hình 28 – Áp dụng mô hình hồi quy regression cho bộ dữ liệu 32

Hình 29 – Áp dụng mô hình Random forest cho bộ dữ liệu 33

Hình 30 – Áp dụng mô hình hồi quy regression cho bộ dữ liệu với biến Giá được đưa về phân phối chuẩn 33

Hình 31 – Áp dụng mô hình hồi quy Random forest cho bộ dữ liệu với biến Giá được đưa về phân phối chuẩn 34

GIỚI THIỆU VÀ MỤC TIÊU BÀI TOÁN

Xe cũ có một thị trường rất tiềm năng và việc có thể dự đoán giá trị thị trường ô tô đã qua sử dụng có thể giúp ích cho cả người mua và người bán. Có rất nhiều cá nhân quan tâm đến thị trường ô tô đã qua sử dụng tại một số thời điểm trong cuộc sống của họ vì họ muốn bán xe hoặc mua một chiếc xe đã qua sử dụng. Trong quá trình này, việc trả quá nhiều hoặc bán ít hơn sẽ là một góc khuất lớn đối với giá trị thị trường.

Bộ dữ liệu gồm thông tin về 6019 xe ô tô cũ với 13 biến dự báo, **mục tiêu là xây dựng mô hình hồi quy tuyến tính bội để dự đoán giá của 1 chiếc xe ô tô đã qua sử dụng** thông qua các biến đầu vào như: Tên/thương hiệu của xe, năm sản xuất, số km đã chạy,...

Dữ liệu được sử dụng trong dự án này được tải xuống từ Kaggle.

1. Xác định tập dữ liệu

Bộ dữ liệu car_data được cung cấp bởiKaggle:

https://www.kaggle.com/iabhishekmaurya/used-car-price-prediction, với các thông số được thu thập tại thị trường Ấn Độ.

1.1. Xác định các biến trong bộ dữ liệu

Bộ dữ liệu gồm 12 biến (7 biến định tính và 5 biến định lượng):

1. Name: Tên/thương hiệu của loại xe - Biến định tính (định danh)
2. Location: Khu vực - Biến định tính (định danh)
3. Year: Năm sản xuất - Biến định tính (thứ tự)
4. Kilometers_Driven: Số km đã chạy - Biến định lượng (liên tục)
5. Fuel_Type: Loại nhiên liệu - Biến định tính (định danh)

6. Transmission: Hộp số (bộ truyền chuyển động) - Biến định tính (định danh)
7. Owner_Type: Chủ sở hữu - Biến định tính (thứ tự)
8. Mileage: Mức tiêu thụ nhiên liệu: Biến định lượng (rời rạc)
9. Engine: Dung tích động cơ - Biến định lượng (rời rạc)
10. Power: Sức mạnh động cơ (mã lực) - Biến định lượng (rời rạc)
11. Seats: Số ghế ngồi - Biến định lượng (rời rạc)
12. Price: Giá - Biến định lượng (liên tục)

1.2. Khai phá và xử lý dữ liệu missing

Sử dụng tập dữ liệu với các biến giải thích đã cho để dự đoán giá của bất kỳ chiếc xe đã qua sử dụng nào.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6019 entries, 0 to 6018
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Unnamed: 0         6019 non-null   int64  
 1   Name               6019 non-null   object 
 2   Location           6019 non-null   object 
 3   Year               6019 non-null   int64  
 4   Kilometers_Driven  6019 non-null   int64  
 5   Fuel_Type          6019 non-null   object 
 6   Transmission       6019 non-null   object 
 7   Owner_Type         6019 non-null   object 
 8   Mileage            6017 non-null   object 
 9   Engine             5983 non-null   object 
 10  Power              5983 non-null   object 
 11  Seats              5977 non-null   float64
 12  New_Price          824 non-null    object 
 13  Price              6019 non-null   float64
dtypes: float64(2), int64(3), object(9)
memory usage: 658.5+ KB
```

Hình 1 – Tổng quan về bộ dữ liệu

Nhận xét

Bộ dữ liệu gồm 6019 quan sát với 13 biến (cột id không sử dụng). Tuy nhiên, nhiều trường dữ liệu đang bị định dạng chưa chính xác và sẽ cần phải xử lý trước khi phân tích sâu hơn. Cụ thể:

- Cột id: Không sử dụng

- Year: Dữ liệu dạng số nhưng đây là biến định tính (có thứ tự) vì năm sản xuất không sử dụng để tính toán.
- Mileage: Mức tiêu thụ nhiên liệu là dữ liệu định lượng (rời rạc), vì dữ liệu thể hiện kèm với đơn vị đo nên đang bị hiểu thành dữ liệu định tính. Dữ liệu này cần xử lý đồng bộ và tách đơn vị đo.
- Engine: Dung tích động cơ là dữ liệu định lượng (rời rạc), vì dữ liệu thể hiện kèm với đơn vị đo nên đang bị hiểu thành dữ liệu định tính. Dữ liệu này cần xử lý tách đơn vị đo.
- New_Price: Giá xe khi mua mới là dữ liệu định lượng (rời rạc), vì dữ liệu thể hiện kèm với đơn vị đo tiền tệ nên đang bị hiểu thành dữ liệu định tính. Dữ liệu này bị thiếu (missing) khá nhiều nên có thể cân nhắc bỏ trường thông tin này.

1.2.1. Kiểm tra các quan sát đầu và cuối trong bộ dữ liệu

```
#Kiểm tra 5 dòng đầu tiên của bộ dữ liệu
data_train.head()
```

	Unnamed: 0	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price
0	0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5.0	NaN	1.75
1	1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5.0	NaN	12.50
2	2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5.0	8.61 Lakh	4.50
3	3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7.0	NaN	6.00
4	4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5.0	NaN	17.74

```
#Kiểm tra 5 dòng cuối cùng của bộ dữ liệu
data_train.tail()
```

	Unnamed: 0	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price
6014	6014	Maruti Swift VDI	Delhi	2014	27365	Diesel	Manual	First	28.4 kmpl	1248 CC	74 bhp	5.0	7.88 Lakh	4.75
6015	6015	Hyundai Xcent 1.1 CRDi S	Jaipur	2015	100000	Diesel	Manual	First	24.4 kmpl	1120 CC	71 bhp	5.0	NaN	4.00
6016	6016	Mahindra Xylo D4 BSIV	Jaipur	2012	55000	Diesel	Manual	Second	14.0 kmpl	2498 CC	112 bhp	8.0	NaN	2.90
6017	6017	Maruti Wagon R VXI	Kolkata	2013	46000	Petrol	Manual	First	18.9 kmpl	998 CC	67.1 bhp	5.0	NaN	2.65
6018	6018	Chevrolet Beat Diesel	Hyderabad	2011	47000	Diesel	Manual	First	25.44 kmpl	936 CC	57.6 bhp	5.0	NaN	2.50

Hình 2 – Kiểm tra các quan sát đầu và cuối trong bộ dữ liệu

1.2.2. Kiểm tra và xử lý các giá trị null (rỗng) trong bộ dữ liệu

```
Unnamed: 0              0
Name                    0
Location                0
Year                    0
Kilometers_Driven       0
Fuel_Type               0
Transmission            0
Owner_Type              0
Mileage                 2
Engine                 36
Power                  36
Seats                  42
New_Price            5195
Price                   0
dtype: int64
```

Hình 3 – Kiểm tra các giá trị null (rỗng) trong bộ dữ liệu

*** Nhận xét:**

Các giá trị rỗng sẽ làm giảm độ chính xác khi đo đếm và vẽ sơ đồ cho dữ liệu. Qua kiểm tra ta thấy giá trị null nằm trong các trường Engine, Power, Seats, New_Price. Cụ thể:

- Cột New_Price chứa nhiều giá trị rỗng (5195/6019 điểm dữ liệu là g iá trị null). Tuy nhiên khi xét tương quan (correlation) biến new_price với biến Price ta thấy hệ số tương quan =0.15 cho ta thấy biến new_price vẫn có thể có ý nghĩa trong mô hình.

  Hướng xử lý: Với các ô dữ liệu chứa giá trị null ta thay thế bởi giá trị 0, các ô có giá trị new_price thay thế bởi giá trị 1.

- Các biến Mileage, Engine, Power, Seats có chưa các điểm dữ liệu Null tuy nhiên số lượng dữ liệu Null trên tổng số mẫu thu thập tương đối nhỏ. Do vậy có thể đánh giá các biến trên vẫn có ý nghĩa thống kê.

  Hướng xử lý: Các biến liên tục Mil eage, Engine, Power ta thay thế các giá trị null bằng giá trị trung bình (mean) của các dữ liệu sẵn có. Vớ i biến rời rạc Seats ta sử dụng giá trị mode để thay thế cho giá trị null.

```
data_train.isnull().sum()
```

```
Unnamed: 0          0
Name                0
Location            0
Year                0
Kilometers_Driven   0
Fuel_Type           0
Transmission        0
Owner_Type          0
Mileage             0
Engine              0
Power               0
Seats               0
New_Price           0
Price               0
dtype: int64
```

Hình 4 – Kiểm tra các giá trị null trong bộ dữ liệu sau khi đã xử lý

*** Nhận xét:**

- Tổng số quan sát vẫn giữ đủ 6019 quan sát.

- Các cột dữ liệu vốn là định lượng nhưng đang được hiểu là định tính do gắn với đơn vị đo cần được chuyển về dạng số.

	Company	Mileage(km/kg)	Engine(CC)	Power(bhp)
0	MARUTI	26.60	998.0	58.16
1	HYUNDAI	19.67	1582.0	126.20
2	HONDA	18.20	1199.0	88.70
3	MARUTI	20.77	1248.0	88.76
4	AUDI	15.20	1968.0	140.80

```
data_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6019 entries, 0 to 6018
Data columns (total 18 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Unnamed: 0         6019 non-null   int64  
 1   Name               6019 non-null   object 
 2   Location           6019 non-null   object 
 3   Year               6019 non-null   int64  
 4   Kilometers_Driven  6019 non-null   int64  
 5   Fuel_Type          6019 non-null   object 
 6   Transmission       6019 non-null   object 
 7   Owner_Type         6019 non-null   object 
 8   Mileage            6019 non-null   object 
 9   Engine             6019 non-null   object 
 10  Power              6019 non-null   object 
 11  Seats              6019 non-null   float64
 12  New_Price          6019 non-null   int64  
 13  Price              6019 non-null   float64
 14  Company            6019 non-null   object 
 15  Mileage(km/kg)     6019 non-null   float64
 16  Engine(CC)         6019 non-null   float64
 17  Power(bhp)         6019 non-null   float64
dtypes: float64(5), int64(4), object(9)
memory usage: 846.5+ KB
```

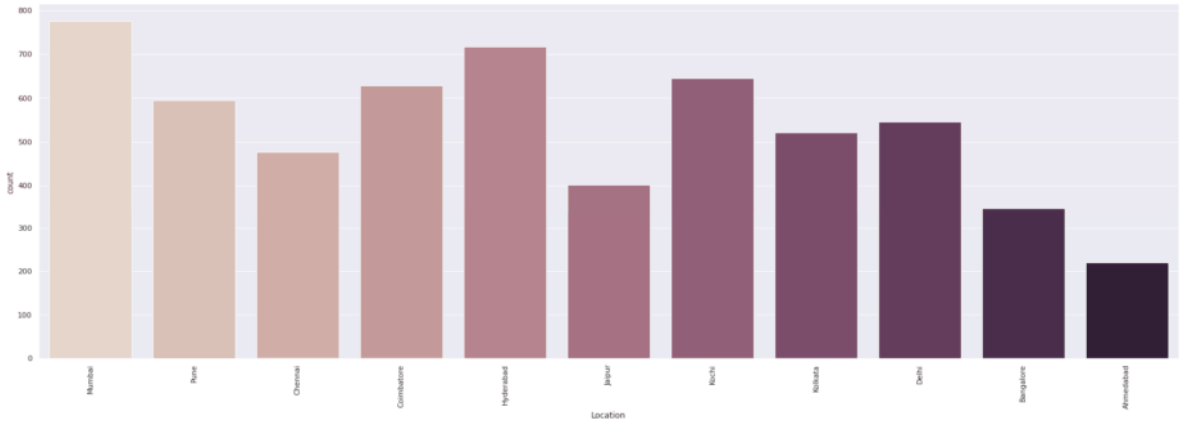
Hình 5 – Tách đơn vị đo và chuyển dữ liệu về dạng số

*** Nhận xét:**

Sau khi tách đơn vị đo và chuyển dữ liệu về dạng số thì ta tiến hành xóa các cột dữ liệu ban đầu để giúp dễ theo dõi, tránh trùng lặp và xử lý nhanh hơn.

2. Phân tích biến đơn và xử lý dữ liệu

2.1. Biến địa điểm



Hình 6 – Biểu đồ biểu diễn địa điểm phân bố xe cũ

```
Mumbai        775
Hyderabad     718
Kochi         645
Coimbatore    629
Pune          594
Delhi         545
Kolkata       521
Chennai       476
Jaipur        402
Bangalore     347
Ahmedabad     220
Name: Location, dtype: int64
```

Hình 7 – Các thành phố có lượng xe cũ từ cao xuống thấp

*** Nhận xét:**

Lượng xe cũ tập trung chủ yếu ở Mumbai và Hyderabad, điểm chung cả 2 thành phố đều có cấu trúc dân cư phức tạp, thuộc nhiều tầng lớp trong xã hội có mức thu nhập từ thấp tới cao. Phân tích kỹ hơn về 6 thành phố theo thứ tự ở hình 7 sẽ cho thấy cái nhìn tổng quan và sự ảnh hưởng của các yếu tố kinh tế, xã hội tới nhu cầu của ô tô đã qua sử dụng:

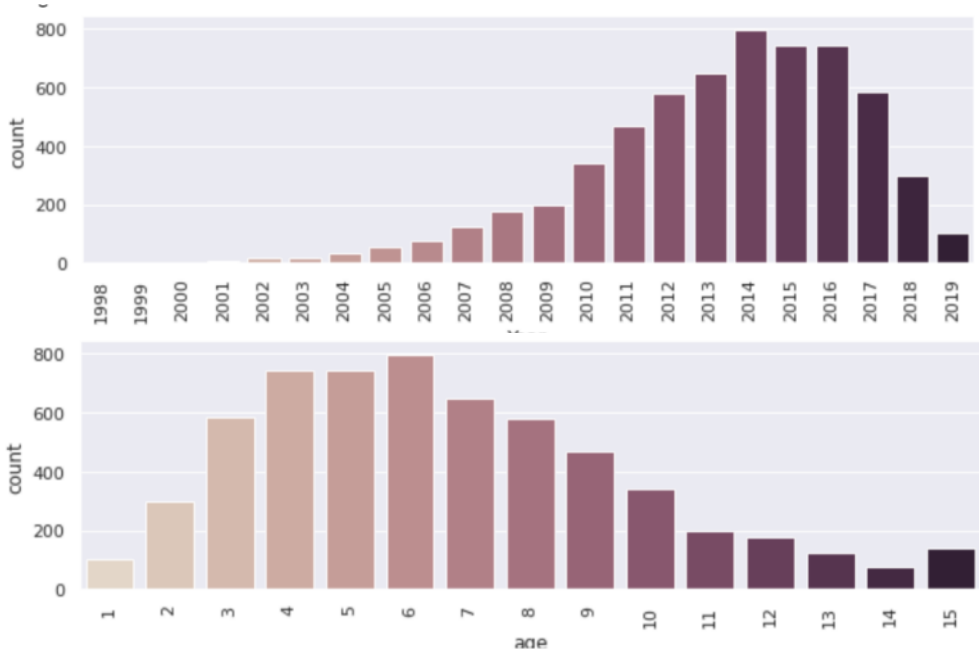
- Thành phố Mumbai là thủ phủ của bang Maharashtra, là thành phố đông dân nhất Ấn Độ, và theo một số cách tính toán là thành phố đông dân nhất thế giới với

một dân số ước tính khoảng 22 triệu người (thời điểm năm 2019). Cùng với các ngoại ô xung quanh, Mumbai tạo thành một vùng đô thị đông dân thứ 6 thế giới với dân số khoảng 20 triệu người.Mumbai là thủ đô thương mại và giải trí của Ấn Độ, thu hút người nhập cư từ khắp nơi trê n đất nước Ấn Độ do thành phố này có nhiều cơ hội kinh doanh và mức số ng. Điều này lý giải tại sao lượng xe cũ ở Mumbai là rất cao. Hiện nay vấn đề ô nhiễm khí thải từ xe cũ cũng đang là vấn đề đau đầu đối với giới chức của thành phố này.

- Thành phố Hyderabad là 1 thành phố tương đối nhỏ với dân số chỉ bằng 1/3 thành phố Mumbai (thành phố đông dân nhất Ấ n Độ) nhưng lượng xe cũ khá cao. Điều này được lý giải bởi thành phố Hyderabad là thành phố du lịch nổi tiếng được mệnh danh là "Thành phố ngọc trai" nên dịch vụ vận chuyển, đáp ứng nhu cầu của khách du lịch tạo nên nhu cầu cao về các loại xe cũ.
- Thành phố Kochi, còn gọi là Cochin, là một thành phố cảng lớn ở miền tây nam Ấn Độ, nằm bên biển Ả Rập. Kochi đứng đầu bang Kerala về tổng số lượt khách du lịch nội địa và quốc tế, và đứng thứ sáu về các điểm du lịch Ấn Độ và là một trong 28 thành phố Ấn Độ nằm trong danh sách 440 thành phố toàn cầu đang lên (một nghiên cứu năm 2011 thực hiện bởi Học viện toàn cầu McKinsey) mà được dự đoán sẽ chiếm giữ 50% GDP thế giới.
- Thành phố Coimbatore còn có tên là Kovai, là thành phố ở bang Tamil Nadu, phía nam Ấn Độ. Thành phố Coimbatore có các ngành công nghiệp: chế biến cà phê và dầu thực vật, thuộc da và sản xuất hàng dệt đặc biệt ngành công nghiệp may mặc cực kỳ phát triển tại đây.
- Thành phố Pune là thành phố lớn thứ nhì ở bang Maharashtra và là thành phố đông dân thứ bảy Ấn Độ (khoảng 5tr dân). Pune được biết đến với các ngành sản xuất ô tô, cũng như các viện nghiên cứu về công nghệ thông tin , đời sống của người dân tương đối cao, lượng xe cũ chỉ ở mức trung bình.
- Thủ đô Delhi có dân số và GDP cao thứ 2 Ấn Độ với kinh tế chủ yếu là thương nghiệp, tập trung nhiều lao động nước ngoài tay nghề cao với mức thu nhập cao. Đây cũng là thành phố du lịch nổi tiếng và phương tiện giao thông công cộng hiện đại nên chủ yếu người lao động có thu nhập trung bình sẽ lựa chọn phương ti ện giao thông công cộng thay vì đầu tư 1 chiếc xe cũ.

Tóm lại, qua phân tích ta có thể thấy mức sống, sự đô thị hóa, sự phát triển của các phương tiện giao thông công cộng và sự phát triển của các khu công nghiệp, ngành du lịch đóng vai trò quan trọng tới nhu cầu sở hữu xe ô tô của dân cư.

2.2. Năm sản xuất (year)




```
2014    797
2015    744
2016    741
2013    649
2017    587
2012    580
2011    466
2010    342
2018    298
2009    198
2008    174
2007    125
2019    102
2006     78
2005     57
2004     31
2003     17
2002     15
2001      8
1998      4
2000      4
1999      2
Name: Year, dtype: int64
```

Hình 8 – Biểu đồ biểu diễn phân phối xe cũ theo năm

*** Nhận xét:**

- Mô hình bị lệch trái do có những điểm ngoại lệ trong dữ liệu và các giá trị ngoại lai có thể gây thiệt hại nặng tới hiệu suất của một mô hình thống kê. Mô hình hồi quy sẽ cho kết quả rất tệ khi được huấn luyện qua dữ liệu sai lệch.

Hướng xử lý: Quy đổi year sang tuổi của xe, do mô hình lệch trái các xe có tuổi >15 năm quy về 15 năm.

- Loại xe được bán nhiều tập trung vào loại được sản xuất cách chừng 5 - 7 năm. Điều này có thể được lý giải khá dễ dàng:

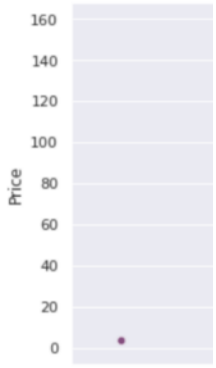
> Mức khấu hao trung bình của xe ô tô là từ 10 – 15 năm. Theo các chuyên gia, khi mua xe ô tô cũ tốt nhất nên chọn xe đã qua sử dụng dưới 6 nă m. Bởi khi này giá trị sử dụng của xe còn nhiều.Tất nhiên tình trạng xe ô tô phụ thuộc vào rất nhiều yếu tố như loại xe, dòng xe, thương hiệu và đặc biệt là cách chăm sóc, bảo dưỡng của chủ xe… Ví dụ như các dòng xe Nhật như xe Toyota, xe Honda, xe Mitsubishi… nổi tiếng là các dòng xe ô tô bền bỉ và ổn định hơn. Nếu được chăm sóc kỹ, lái xe đúng cách, bảo dưỡng đúng lịch bảo dưỡng của hãng… sẽ giữ phong độ tốt hơn.

Tuy nhiên nhìn chung, xe ô tô sử dụng từ 1 – 6 năm chưa hao mòn, xuống cấp nặng. Phụ tùng, linh kiện chưa phải thay thế nhiều. Hệ thống vận hành như động cơ, hộp số, hệ thống lái… đa phần vẫn còn hoạt động ổn định. Tình trạng ngoài thất như sơn xe hay tình trạng nội thất như hệ thống ghế xe, hệ thống giải trí màn hình – loa, hệ thống điều hoà… gần như vẫn còn tốt. Khi này, ngoại trừ việc tình trạng xe khá mới thậm chí rất mới thì xe còn nằm trong thời gian bảo hành chính hãng. Thế nên mọi thứ sẽ có sự đảm bảo cao hơn.

Trái lại, dù là xe phổ thông bình dân hay xe sang thì sau 10 năm chắc chắn đã hao mòn, xuống cấp nhiều do đó, rủi ro trục trặc, hư hỏng là rất cao mà xét về khía cạnh an toàn cũng không thực sự đảm bảo. Vì thế, dù giá xe ô tô cũ 10 năm rất rẻ, rất hấp dẫn nhưng rủi ro đi kèm cũng rất cao. Đây là lý do xe cũ đời sâu hay xe cũ 10 năm bị người sử dụng liệt vào nhóm những xe cũ không nên mua…

2.3. Số km đã chạy (Kilometers_Driven)

```
60000    80
45000    69
65000    67
50000    60
55000    58
         ..
70920     1
75014     1
32005     1
25858     1
83969     1
Name: Kilometers_Driven, Length: 3038, dtype: int64
```



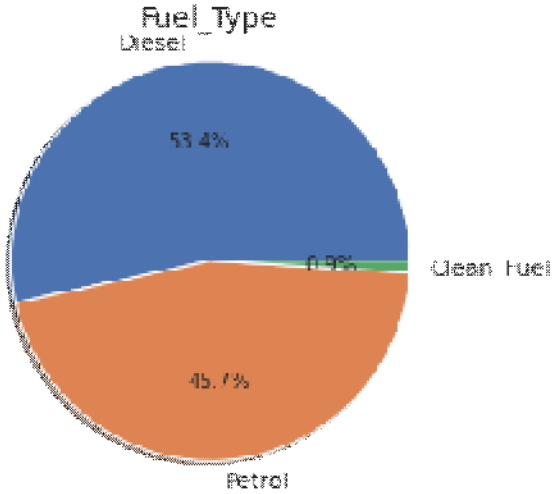
Hình 9 – Biểu đồ biểu diễn phân phối số km đã chạy

Nhận xét:

Do bộ dữ liệu khi nhập chứa cả loại dữ liệu đã được làm tròn và chưa được làm tròn nên lượng các điểm dữ liệu rất lớn mà chưa thật sự mô tả đúng bản chất(gồm 3093 điểm dữ liệu khác nhau). Do đó, ta có thể gộp dữ liệu thành các nhóm bằng cách sử dụng kmean để chia các dữ liệu thành 10 cụm dữ liệu khác nhau.

Phần lớn số xe cũ được ưa chuộng đã chạy được khoảng 40.000km – 60.000km. Đối với xe gia đình, trung bình dòng ôtô phổ thông có thể đi được khoảng 10.000 - 15.000 km/năm. Như vậy, số km đã đi tương đương với loại xe có số tuổi từ 4 – 5 năm, phù hợp với các tiêu chí đã phân tích.

2.4. Loại nhiên liệu (Fuel_Type)



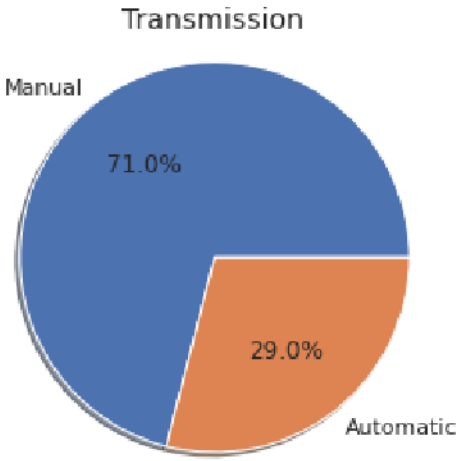

```
Diesel        3205
Petrol        2746
Clean_Fuel      68
Name: Fuel_Type, dtype: int64
```

Hình 10 – Biểu đồ biểu diễn phân phối loại nhiên liệu

Nhận xét: Có 4 loại nhiên liệu tương ứng với các xe được bán, gồm: Diesel, petrol, CNG, LPG. Trong đó, loại xe chạy nhiên liệu Diesel chiếm 53.3%

Cụ thể:

- Diesel: Nhiên liệu phổ thông sử dụng nhiều trong ngành công nghiệp ô tô (53.3%)
- Petrol: Được sử dụng phổ thông (45.6%)
- Clean_fluel(LPG, CNG, Electric): Khí hóa lỏng và năng lượng điện thân thiện với môi trường được sử dụng ít, chỉ khoảng 20triệu xe/thế giới sử dụng các loại này do trở ngại trong việc đầu tư thiết bị chuyển đổi nhiên liệu, bình gas (ước tính từ 1.200 – 1.500 USD/ôtô); thiếu hạ tầng dịch vụ cung ứng LPG/CNG (kho chứa nhiên liệu, cột nạp LPG/CNG cho phương tiện)…



2.5. Hộp số (Transmission)

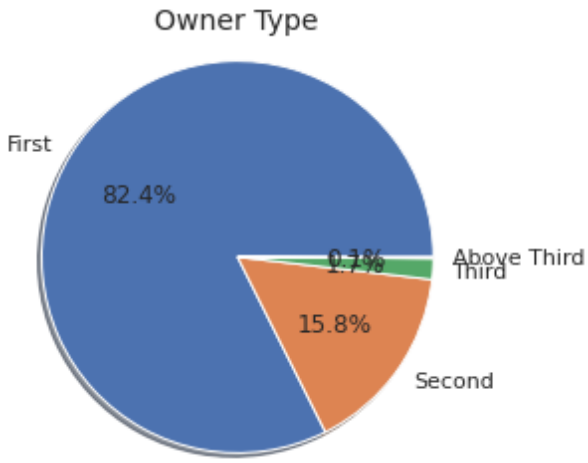
```
Manual       4170
Automatic    1702
Name: Transmission, dtype: int64

count       5872
unique         2
top       Manual
freq        4170
Name: Transmission, dtype: object
```

Hình 11 – Biểu đồ biểu diễn phân phối loại hộp số

*** Nhận xét:** Có 2 loại hộp số là số tay và tự động, trong đó, loại xe số tay chiếm đa số với 71%. Hầu hết các xe cũ thường là bản số sàn(Manual) do vậy số lượng xe số sàn chiếm đa số trong số lượng xe cũ, bên cạnh đó giá trung bì nh của 1 xe cũ b ản số tự động cũng cao gấp 4 lần xe số sàn do vậy thị phần không nhiều do không phù hợp với kinh tế của nh ững người có nhu cầu mua xe cũ.

2.6.Chủ sở hữu (Owner_Type)




```
First             4839
Second             925
Third              101
Fourth & Above       7
Name: Owner_Type, dtype: int64
```

Hình 12 – Biểu đồ biểu diễn phân phối về chủ sở hữu

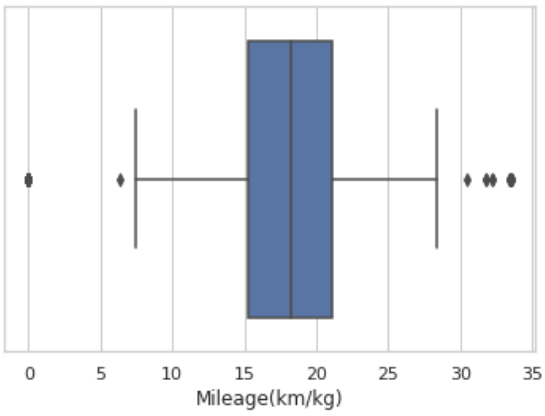
*** Nhận xét:** Phần lớn số xe cũ được bán mới qua 1 đời chủ, điều này cho thấy người mua xe chú trọng đến tính ổn định, an toàn của chiếc xe (xe trải qua nhiều đời chủ có thể là xe có độ an toàn thấp, sửa chữa nhiều).

Nếu một chiếc xe ô tô sử dụng suôn sẻ thì khi xe bắt đầu xuống cấp tức khoảng 4 năm trở đi người ta mới bán đi và tậu xe mới. Trừ trường hợp xe bị tai nạn, xe đem lại xui xẻo, hoặc gặp sự cố,…Một trường hợp nữa đó là chủ nhân muốn chạy theo xu hướng nên bán xe ô tô cũ và đổi xe mới nhưng rất ít gặp. Do đó nếu như qua 2 đời chủ tức là xe đó có đời sản xuất cách thực tại khoảng 6 năm trở đi. Đối với các loại xe ô tô cũ này người có nhu cầu sẽ không lựa chọn do:

- Thứ nhất, xe chạy lâu ngày động cơ đã hư hỏng nhiều.
- Thứ hai, xe có thể đã trải qua những lần đại tu.
- Thứ ba, có thể là chiếc xe không may mắn, hay gây tai nạn,…

Nếu mua phải những chiếc xe này về người sử dụng sẽ tốn không ít tiền sửa vặt hoặc phải nhanh chóng bán tháo đi và chắc chắn sẽ lỗ nặng vì thực tế thì những người am hiểu về xe sẽ chẳng ai muốn tậu một chiếc xe đã quá cũ thế này.

2.7. Mức tiêu thụ nhiên liệu (Mileage)



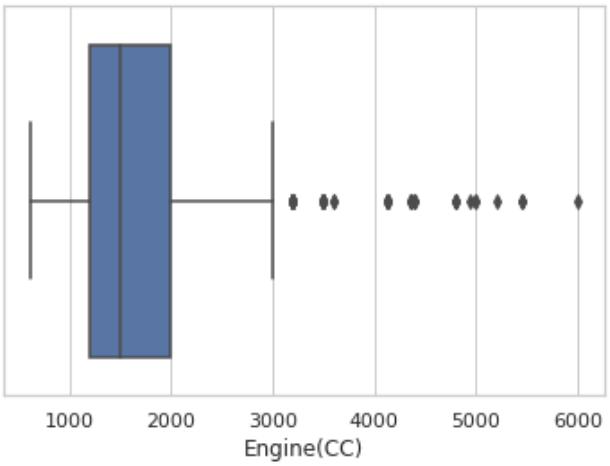

```
18.90    172
17.00    149
18.60    119
20.36     88
21.10     87
         ...
9.70       1
13.29      1
12.97      1
15.11      1
14.45      1
Name: Mileage(km/kg), Length: 429, dtype: int64
```

Hình 13 – Biểu đồ biểu diễn phân phối mức tiêu thụ nhiên liệu

Nhận xét:

Hầu hết các xe đã qua sử dụng có mức độ tiêu thụ nhiên liệu dao động từ 15 - 21 (km/kg/lit), đây là mức tiêu hao nhiên liệu khá lý tưởng , phù hợp với mong muốn của người có nhu cầu sử dụng xe cũ.

2.8. Dung tích động cơ (Engine)



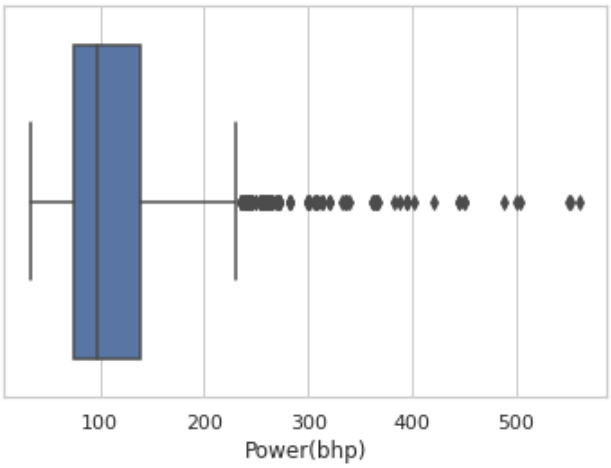

```
1197 CC    606
1248 CC    512
1498 CC    304
998 CC     259
2179 CC    240
           ...
1422 CC      1
1978 CC      1
2487 CC      1
2720 CC      1
2112 CC      1
Name: Engine, Length: 146, dtype: int64
```

Hình 14 – Biểu đồ biểu diễn phân phối dung tích động cơ

*** Nhận xét:**

Hầu hết các xe đã qua sử dụng phù hợp nhu cầu người sử dụng có dung tích động cơ dao động từ 1200 - 2000 CC. Điều này cho thấy loại xe có dung tích động cơ này thường là loại nhỏ hoặc vừa, hiệu suất vừa đủ, tiện di chuyển trong thành phố, tiết kiệm nhiên liệu.

2.9. Công suất của động cơ (Power)



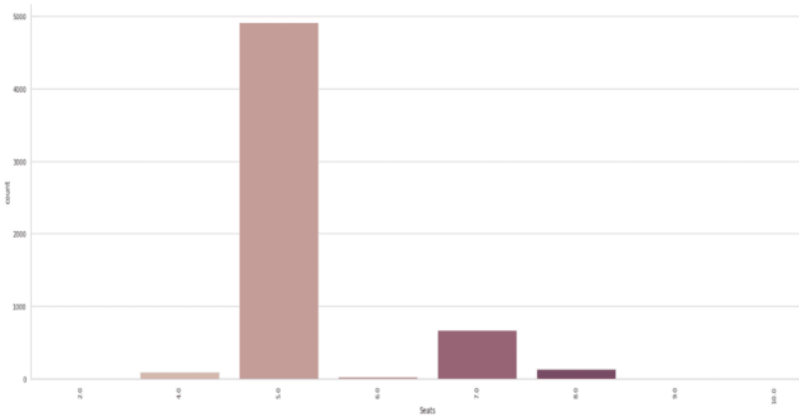

```
74 bhp        235
98.6 bhp      131
73.9 bhp      125
140 bhp       123
78.9 bhp      111
              ...
402 bhp         1
321 bhp         1
41 bhp          1
144 bhp         1
281.61 bhp      1
Name: Power, Length: 372, dtype: int64
```

Hình 15 – Biểu đồ biểu diễn phân phối công suất của động cơ

*** Nhận xét:**

Hầu hết các xe đã qua sử dụng có sức mạnh động cơ dao động quanh mức 100 bhp. Điều này cho thấy dòng xe cũ được ưa chuộng là loại nhỏ hoặc vừa, có công suất trung bình, phù hợp đi trong thành phố.

2.10. Số ghế (Seats)

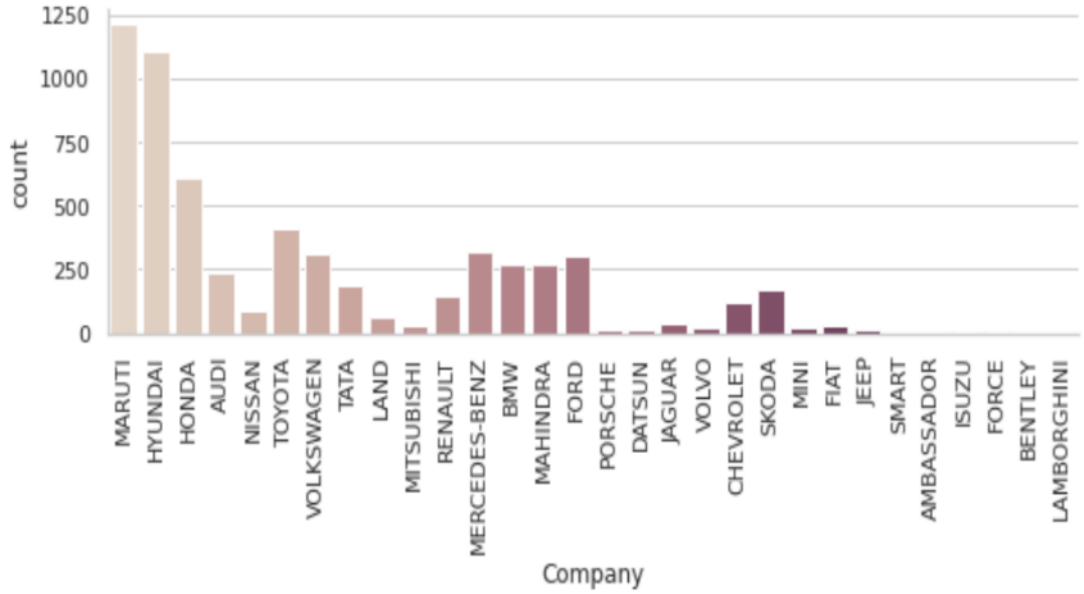


```
5.0     5014
7.0      674
8.0      134
4.0       99
6.0       31
2.0       16
10.0       5
9.0        3
0.0        1
Name: Seats, dtype: int64
```

Hình 16 – Biểu đồ biểu diễn phân phối số ghế xe

*** Nhận xét:** Dòng xe 5 ghế có số lượng chiếm đại đa số vì phù hợp với gia đình có từ 4 - 6 người, thuận tiện trong di chuyển, không chiếm diện tích gara và dễ dàng trong việc tìm kiếm nơi đỗ tại các địa điểm công cộng.

2.11. Nhãn hiệu xe (name)

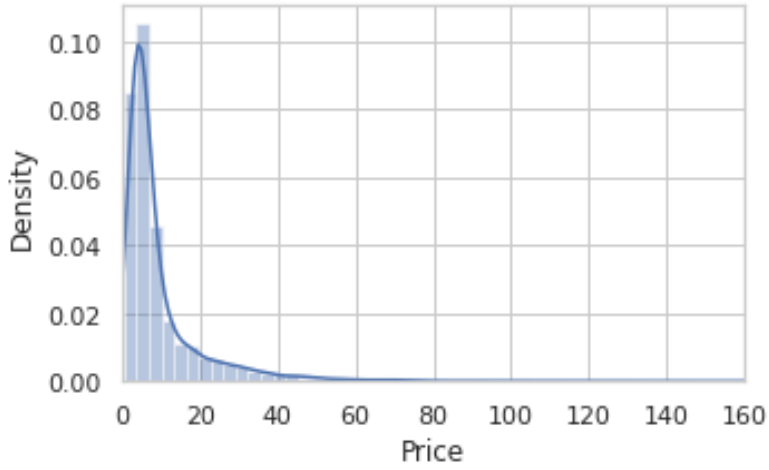


Hình 17 – Biểu đồ biểu diễn phân phối về nhãn hiệu xe

*** Nhận xét:** 3 thương hiệu xe được ưa chuộng nhất thuộc về Maruti, Huyndai và Honda. Thị trường Ấn Độ có nhiều mẫu xe ô tô phổ biến nhưng trong đó dòng xe Maruti rất được yêu thích vì đáp ứng các tiêu chuẩn cho gia đình, cá nhân và chỗ ngồi tiện dụng. Dòng xe

này có ưu điểm là mẫu mã, kiểu dáng thiết kế hiện đại, nội thất sang trọng và giá cả hợp lý. Điều này cho thấy thị hiếu của người sử dụng khá quan tâm tới kiểu dáng của xe.

2.12. Biến giá (Price)



Hình 18_1 – Biểu đồ của biến giá (lệch phải)

*** Nhận xét:** Từ biểu đồ cho ta thấy phân bố của biến price không tuân theo phân bố chuẩn - đang bị lệch phải.Đồ thị lệch phải chứng tỏ ít người có thể (hoặc chấp nhận) mua xe với mức giá 160.000 Rupee.

Lưu ý: Đơn vị tiền tệ của Ấn độ ký hiệu là "lakh" được hiểu như sau:

Một lakh là một đơn vị trong hệ thống đánh số Ấn Độ tương đương với một trăm ngàn. Ví dụ, ở Ấn Độ 150.000 rupee trở thành 1,5 lakh rupee

Hướng xử lý: Sử dụng log để chuẩn hóa lại phân bố biến price.

Hình 18_2 – Biểu đồ của biến giá (chuẩn hóa)

3. Phân tích sự tương quan giữa biến giải thích (x) và biến phụ thuộc (y)

3.1.Tương quan giữa biến giá và các biến định tính

	sum_sq	df	F	PR(>F)
Company	159713.068750	29.0	122.309827	0.000000e+00
Location	13124.641946	10.0	29.147839	2.764997e-55
Fuel_Type	6908.530188	3.0	51.142658	1.278965e-32
Transmission	8759.191358	1.0	194.528353	1.625838e-43
Owner_Type	7306.103531	3.0	54.085825	1.785741e-34
Residual	262287.162643	5825.0	NaN	NaN

Hình 19 – Bảng ANOVA phân tích tương quan giữa biến giá và các biến định tính

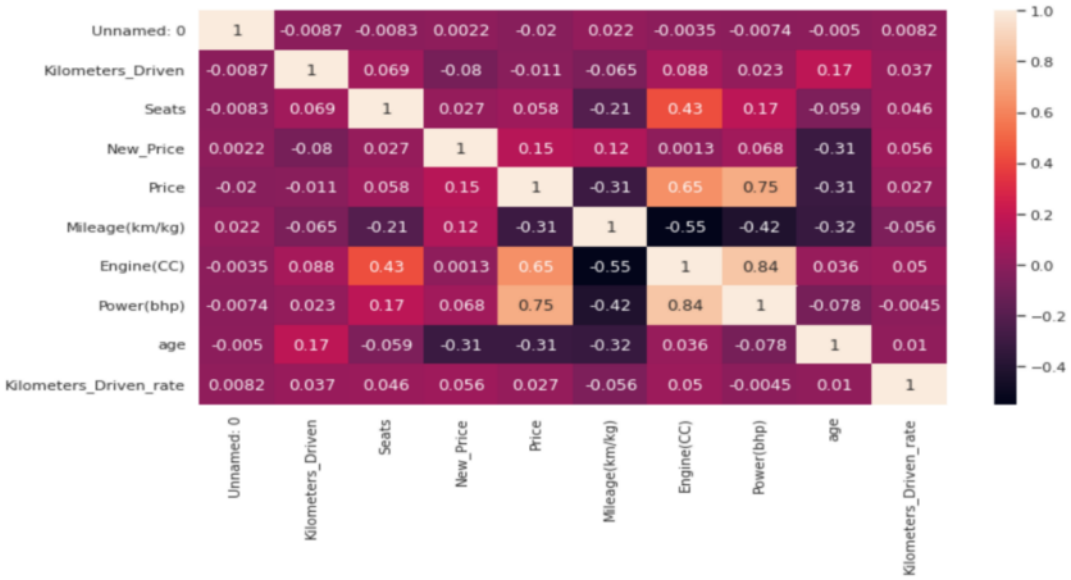
*** Nhận xét:**

- Qua phân tích ANOVA về tương quan giữa các biến định tính Company, Location, Fuel_Type, Transmission, Owner_Type ta thấy P value tính ra đều tương đối bé, qua đó có thể thấy sự thay đổi về Giá thành có sự tương quan tới các biến trên .

- Đối với mô hình dạng tuyến tính, để có thể tính toán đòi h ỏi phải dữ liệu dạng string phải biến đổi để tính toán được. Do vậy, ta sắp xếp lại thứ tự thành dạng: First = 1; Second = 2; Third:3; Fourth & Above:4;

- Mặt khác, về mặt dung lượng bộ nhớ và tốc độ xử lý thì dùng thứ tự dạng số sẽ tối ưu hơn thứ tự dạng string.

3.2. Tương quan giữa biến giá và các biến định lượng



Hình 20 – Biểu đồ phân tích tương quan giữa biến giá và các biến định lượng

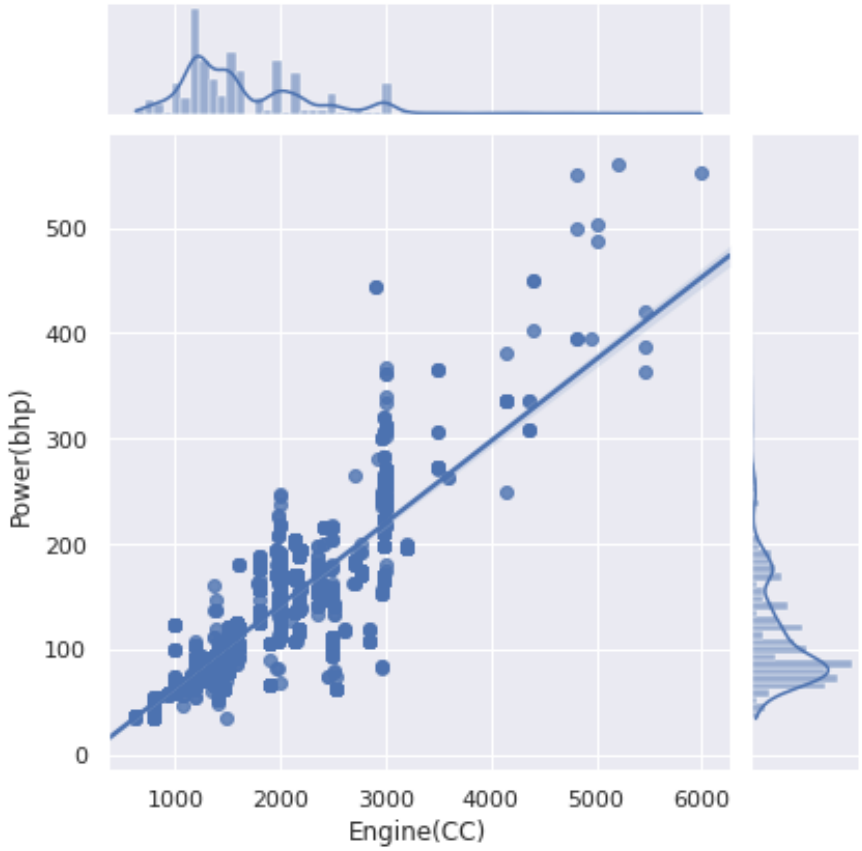
*** Nhậnxét:**

Ta thấy hệ số tương quan của biến Price với các biến Power và Engine tương đối lớn điều đó chứng tỏ các biến có sự tương quan thuận chặt chẽ với nhau. Mặt khác biến Mileage có tương quan nghịch với Price. Trong khi đó biến Kilometer_Driven và biến Seats có hệ số tương quan tương đối bé có vẻ như các thông số về Seats và Kilometer_Driven không có nhiều sự ảnh hưởng tới giá cả.

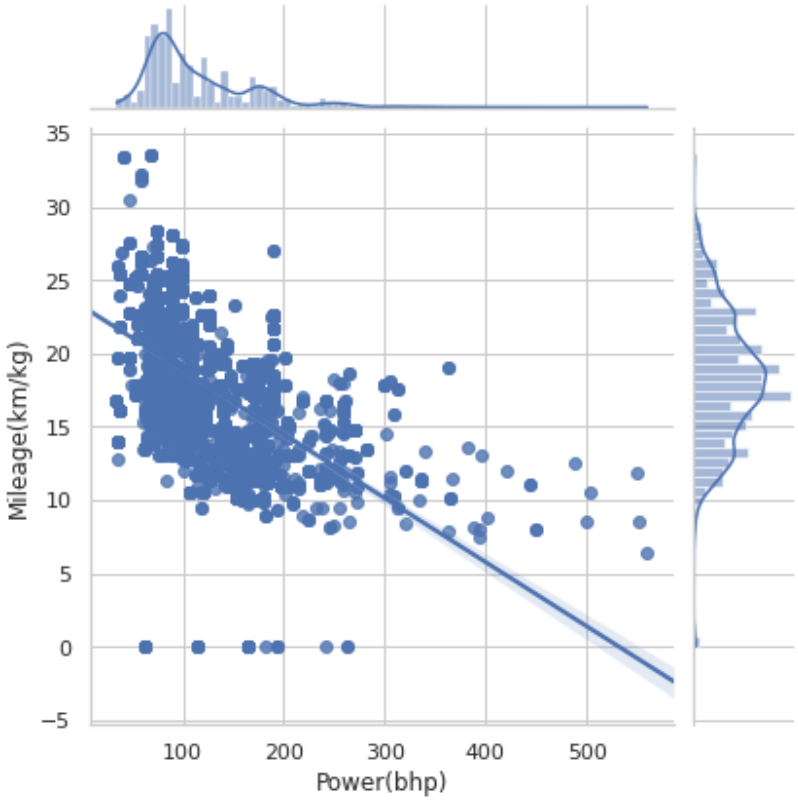
Mileage: Mức tiêu thụ nhiên liệu: Biến định lượng (rời rạc)

Engine: Dung tích động cơ - Biến định lượng (rời rạc)

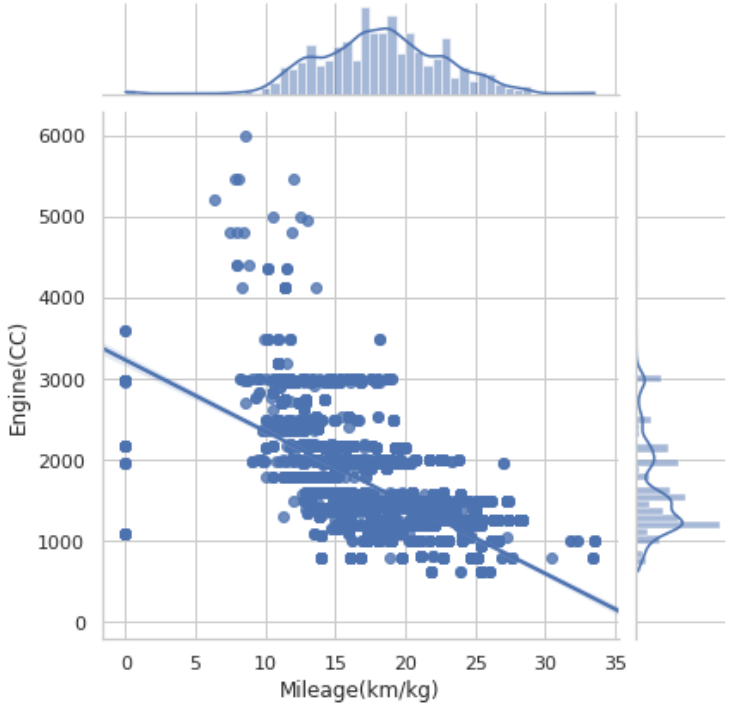
Power: Sức mạnh động cơ (mã lực) - Biến định lượng (rời rạc)



Hình 21 – Biểu đồ phân tích tương quan giữa biến Price và Engine



Hình 22 – Biểu đồ phân tích tương quan giữa biến Price và Power



Hình 23 – Biểu đồ phân tích tương quan giữa biến Price và biến Mileage

*** Nhận xét:**

Ta thấy các biến Engine, Power, Mileage có vẻ như có ảnh hưởng tương quan với nhau thể hiện qua hệ số đánh giá độ tương quan r tương đối lớn. Điều đó cũng phản ánh giống kết quả với đồ thị tương quan giữa các cặp biến với nhau. Ở đây ta thấy các điểm dữ liệu phân bố tập trung theo 1 đường thẳng cắt qua các điểm dữ liệu. Điều này cũng hoàn toàn hợp lý với logic thực tế khi các biến trên đều phụ thuộc lớn vào tùy từng loại động cơ ô tô. Do vậy ta có thể sử dụng phương pháp PCA để có thể giảm được số lượng biến trong mô hình.

```
machine_data = X_final[['Engine(CC)','Power(bhp)','Mileage(km/kg)']]
```

```
from sklearn.preprocessing import StandardScaler
scalar = StandardScaler()
# fitting
scalar.fit(machine_data)
machine_data = scalar.transform(machine_data)
```

```
from sklearn.decomposition import PCA
pca = PCA(n_components = 2)
pca.fit(machine_data)
x_pca = pca.transform(machine_data)
print(x_pca.shape)
print(sum(pca.explained_variance_ratio_))
```

```
(6019, 2)
0.9530948024661079
```

Kết luận:

Từ 3 biến đã chọn, sau khi áp dụng phương pháp PCA để giảm chiều dữ liệu ta thu được 2 biến mới với mức độ phản ánh dữ liệu là 95.3%.

4. Phân cụm dữ liệu

Sau khi xử lí các biến liên tục và encoding các biến rời rạc, chúng ta sẽ sử dụng 2 kĩ thuật phân cụm là K-Means và Phân cụm phân cấp.

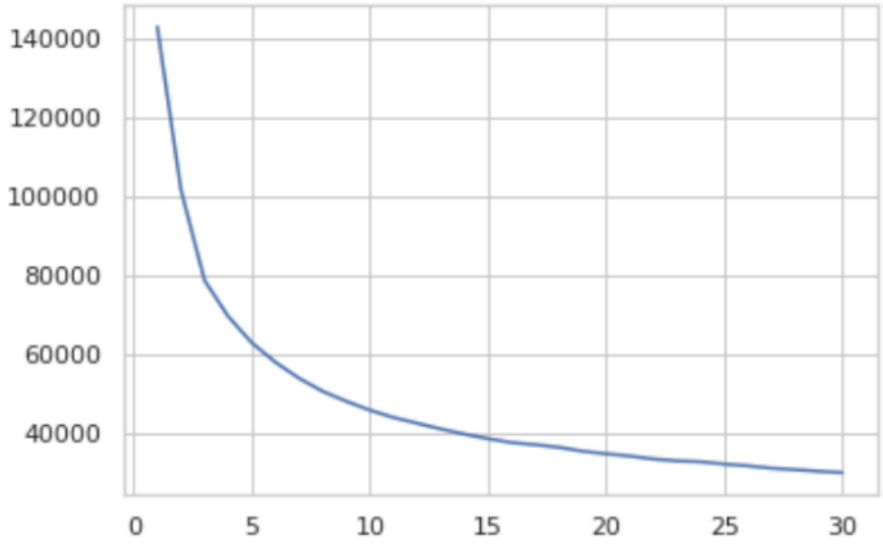
Đối với thuật toán k -mean, ta sử dụng Quy tắc Elbow để tìm ra giá trị số cụm tối ưu. Chúng ta thử k từ 1 đến 30 để tìm độ phân tán sau đó dựa vào biểu đồ elbow point plot để xác định điểm k tối ưu.

```
from sklearn.cluster import KMeans, AgglomerativeClustering

#k-means k = 1 to 30 to find an optimal k by Elbow
distortions = []
for k in range(1, 31):
    cluster = KMeans(n_clusters=k, random_state=0).fit(X_final)
    distortions.append(cluster.inertia_)

import seaborn as sns
sns.lineplot(x=range(1,31), y=distortions)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2cb898f910>
```

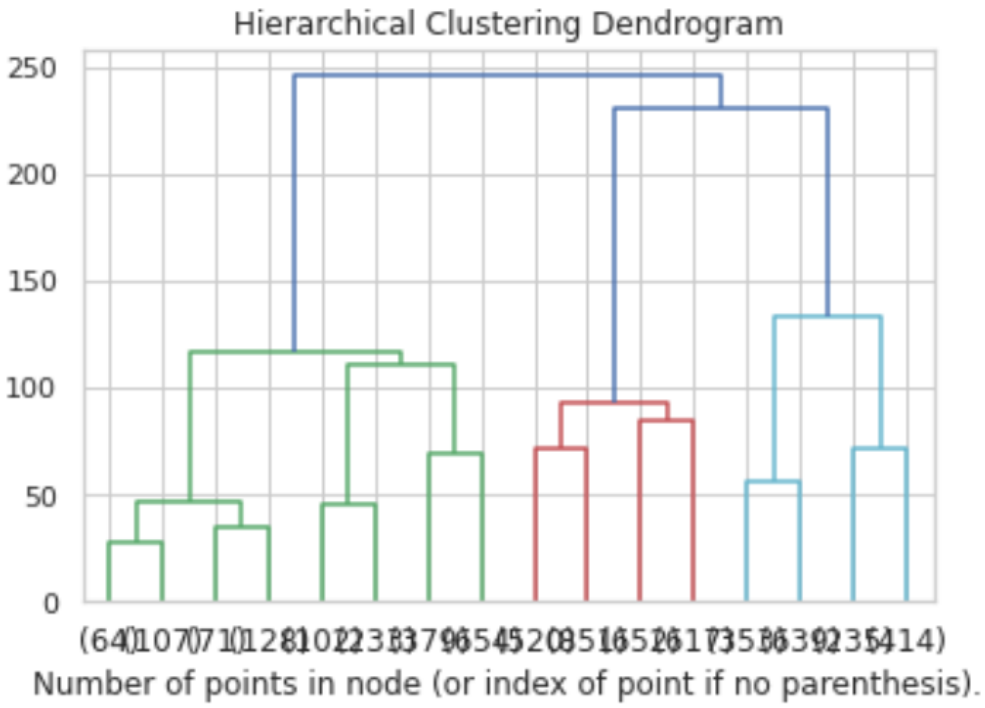


Hình 24 – Biểu đồ elbow point plot

*** Nhận xét:**

Từ đồ thị chúng ta thấy từ K gần 3 trở đi, mối quan hệ của K và độ phân tán trở nên tuyến tính nên ta chọn giá trị K tối ưu nằm trong vùng lân cận với 3.

Để chắc chắn, ta sử dụng thuật toán Phân cụm phân cấp như sau với distance_threshold = None để đưa ra mọi phân cụm có thể.

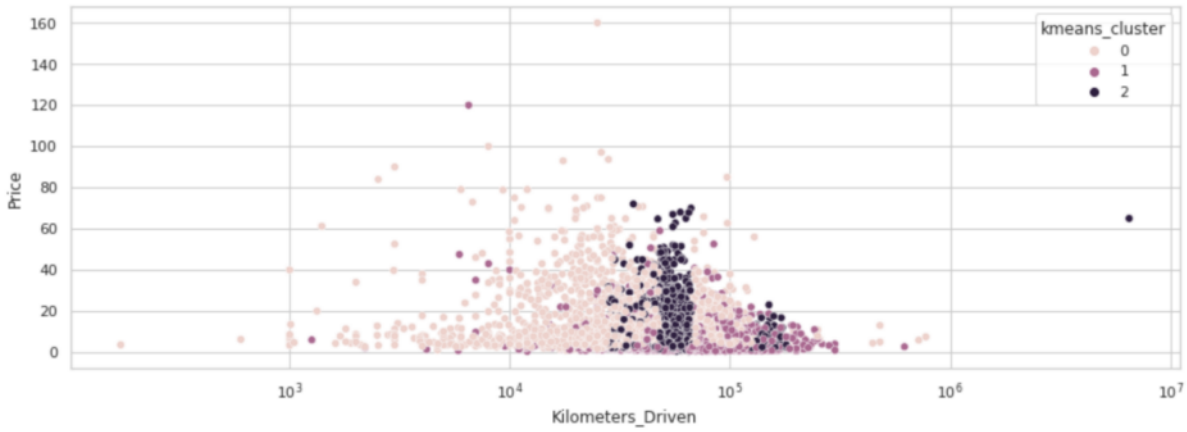


Hình 25 – Biểu đồ phân cụm phân cấp

Như trên biểu đồ dendrogram, ở cấp thứ 3 , bộ dữ liệu được gom từ 8 nhóm -> 3 nhóm và sau đó từ 3 nhóm ->2 nhóm. Do đó, ta có thể coi K=3 là điểm tối ưu, dựa vào việc kết hợp 2 thuật toán phân cụm.

Ta sẽ gán nhãn cụm cho các quan sát bằng thuật toán K-means với K=3.

5. Đánh giá tương quan giữa các cụm dữ liệu



Hình 26 – Đánh giá tương quan giữa các cụm dữ liệu

Để đánh giá phân phối của biến mục tiêu có khác nhau trên các cụm này hay không, ta sẽ sử dụng ANOVA:

	sum_sq	df	F	PR(>F)
kmeans_cluster	14847.461952	1.0	120.983821	7.090743e-28
Residual	738422.524747	6017.0	NaN	NaN

Hình 27 - Bảng ANOVA phân tích tương quan giữa các cụm dữ liệu

*** Nhận xét**

Ta thấy cụm không có sự tách biệt Pvalue rất nhỏ, do vậy có sự khác biệt về giá giữa các cụm.

Kết quả phân cụm được thể hiện như hình 26, tuy nhiên sự tách biệt của các cụm không thực sự rõ ràng và khi thêm biến số cụm không giúp cải thiện kết quả mô hình nên nhóm quyết định không đưa them biến số cụm vào trong mô hình.

6. Lựa chọn đặc trưng cho mô hình đầy đủ

Sau khi phân tích ta lựa chọn các biến cho mô hình đầy đủ như sau:

Stt	Tên biến	Mô tả
1	Name	Làm sạch, xử lý dữ liệu null, dữ liệu trùng lặp, véc tơ hóa dữ liệu.
2	Location	Làm sạch, xử lý dữ liệu null, dữ liệu trùng lặp, véc tơ hóa dữ liệu.
3	Year	Làm sạch, xử lý dữ liệu null, xử lý dữ liệu ngoại lệ (outlier), quy đổi số năm thành số tuổi của xe.
4	Kilometers_Driven	Làm sạch, xử lý dữ liệu null, gộp nhóm (bining) dữ liệu.
5	Fuel_Type	Làm sạch, xử lý dữ liệu null, gộp nhóm (bining) dữ liệu, véc tơ hóa dữ liệu.
6	Transmission	Làm sạch, xử lý dữ liệu null, dữ liệu trùng lặp, véc tơ hóa dữ liệu.
7	Owner_Type	Làm sạch, xử lý dữ liệu null, dữ liệu trùng lặp, véc tơ hóa dữ liệu.
8	PCA1, PCA2	Từ 3 biến Mileage, Engine, Power sau khi áp dụng phương pháp PCA để giảm chiều dữ liệu ta thu được 2 biến mới với mức độ phản ánh dữ liệu là 95.3%.
9	Seat	Làm sạch, xử lý dữ liệu null, dữ liệu trùng lặp.
10	New_price	Làm sạch, xử lý dữ liệu null bằng cách chuyển về dữ liệu factor (0,1).

7. Lựa chọn mô hình học máy để áp dụng cho bộ dữ liệu

7.1. Mô hình hồi quy hồi quy tuyến tính (linear regression)

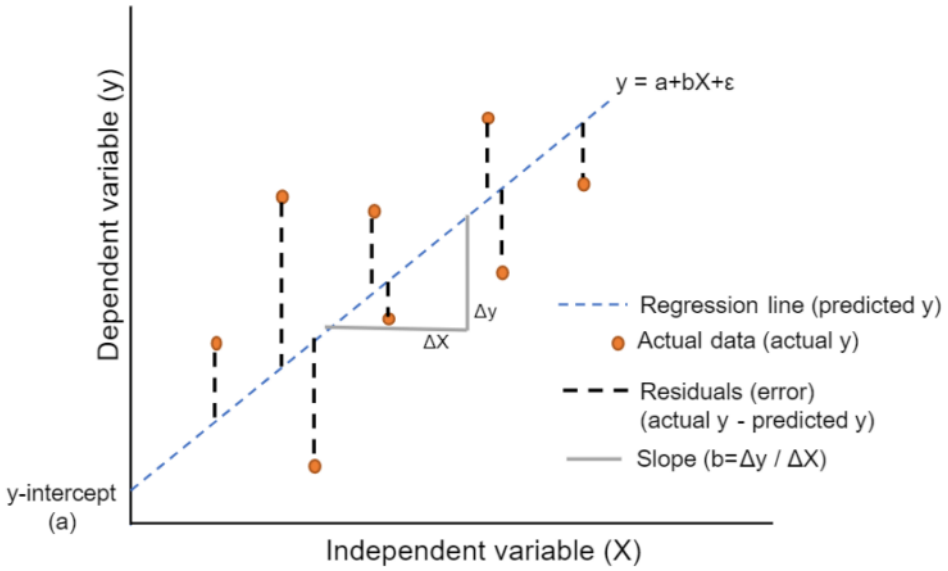
Mô hình có dạng:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

Trong đó:

- y là giá trị của biến phụ thuộc
- X là giá trị của biến độc lập
- β_0 là hệ số chặn
- β_1 là hệ số góc
- εlà sai số

Hàm mất mát được tính theo công thức:

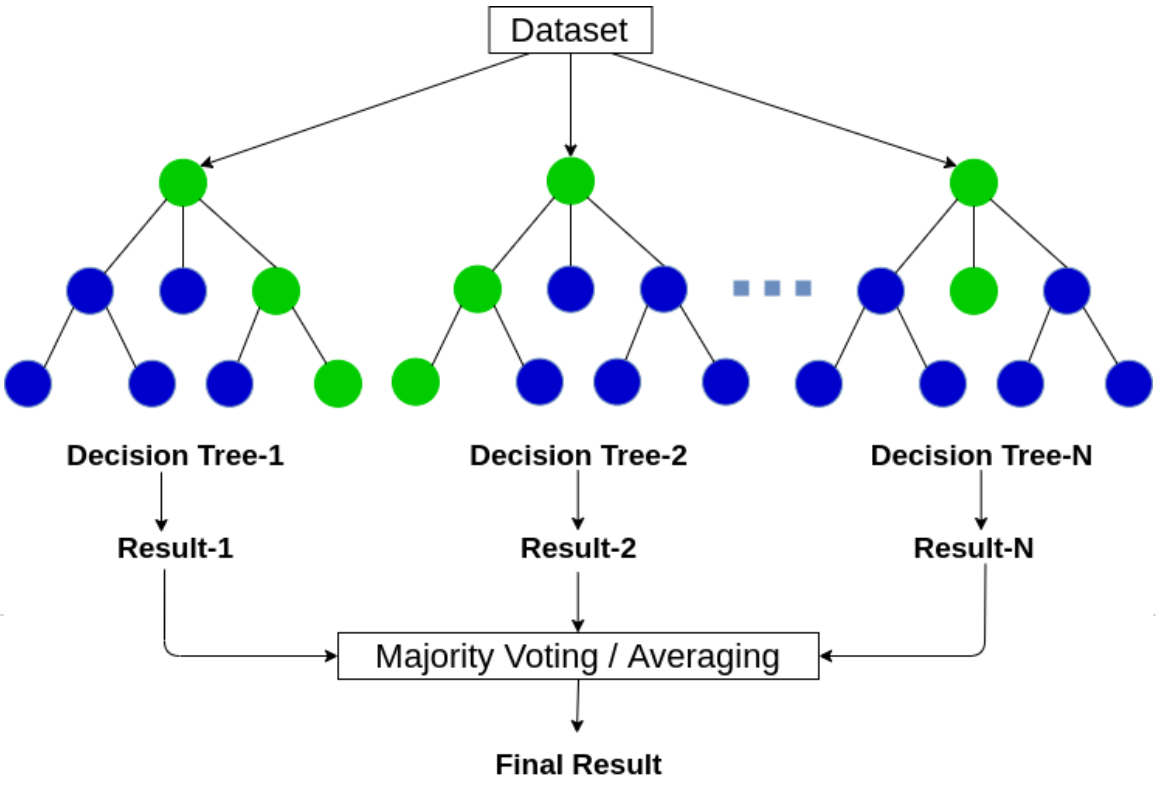


$$\sum_{i=0}^{n} (y_i - \beta_0 - \beta_1 X_i)^2$$

Hình 28 - Mô hình Linear Regression

Mô hình Linear Regression là một mô hình hồi quy đơn giản nhất, nhưng lại là một phương pháp được sử dụng nhiều nhất, với mục đích tìm ra một đường thẳng đi qua các điểm dữ liệu sao cho tổng bình phương phần dư là nhỏ nhất. Hồi quy tuyến tính được sử dụng rộng rãi trong thực tế do tính chất đơn giản hóa của hồi quy. Nó cũng rất dễ để ước lượng.

7.2. Mô hình Random Forest



Hình 29 - Mô hình Random forest

Phương pháp Random Forest sẽ xây dựng nhiều cây quyết định. Tuy nhiên mỗi cây quyết định sẽ khác nhau có y ếu tố ngẫu nhiên. Sau đó kết quả dự đoán được tổng hợp từ các cây quyết định. Thực nghiệm đã chỉ ra rằng phương pháp này hiệu quả hơn việc chỉ dùng một decision tree thông thường với khả năng tổng quát hóa tốt hơn, tránh hiện tượng overfit nhưng đánh đổi b ằng việc ta không thể hiểu cơ chế hoạt động của thuật toán này do cấu trúc quá phức tạp của mô hình này — do vậy thuật toán này là một trong những phương thức Black Box — tức ta sẽ bỏ tay vào bên trong và rút ra được kết quả chứ không thể giải thích được c ơ chế hoạt động của mô hình. Đó là sự đánh đổi giữa khả năng giải thích và khả năng dự báo.

Yếu tố ngẫu nhiên của Random Forest khi sinh ra mỗi cây trong mô hình thể hiện ở hai điểm

- Mỗi cây quyết định sẽ được học một tập con mẫu khác nhau sinh ngẫu nhiên t ừ dữ liệu tổng bằng phương pháp boostraping
- Việc lựa chọn feature tại mỗi nhánh của cây sẽ dựa trên feature cho kết quả tốt nhất trong số các feature được lấy ngẫu nhiên thay vì xét trên toàn bộ tất cả cá feature

8. Áp dụng mô hình vào bộ dữ liệu

8.1. Mô hình hồi quy

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
linear_reg = LinearRegression()
linear_reg.fit(X_train, y_train.T)
y_pred= linear_reg.predict(X_test)

print("R^2 on Traing set: %.2f " % linear_reg.score(X_train,y_train))
print("R^2 on Testing set: %.2f" % linear_reg.score(X_test,y_test))
print('Mean squared error: %.2f'% mean_squared_error(y_test,y_pred))
#print('Coefficient of determination: %.2f'% r2_score(y_test,y_pred))
```

```
R^2 on Traing set: 0.77
R^2 on Testing set: 0.77
Mean squared error: 0.20
```

Hình 28 – Áp dụng mô hình hồi quy regression cho bộ dữ liệu

*** Nhận xét:**

Ta thấy với mô hình hồi quy regression, $R^2 = 0.77$ là tương đối nhỏ; sai số trung bình bình phương (mean squared error) = 0.2

8.2. Mô hình Random forest

```
from sklearn.ensemble import RandomForestRegressor
rf_reg = RandomForestRegressor()
rf_reg.fit(X_train, y_train)
y_pred= rf_reg.predict(X_test)
print("R2 score on Traing set: %.2f"% rf_reg.score(X_train,y_train))
print("R2 score on Testing set: %.2f"% rf_reg.score(X_test,y_test))
print('Mean squared error: %.2f'% mean_squared_error(y_test,y_pred))
```

```
R2 score on Traing set: 0.99
R2 score on Testing set: 0.88
Mean squared error: 0.10
```

Hình 29 – Áp dụng mô hình Random forest cho bộ dữ liệu

*** Nhận xét:**

- Ta thấy với mô hình Random forest, R^2 trên tập train = 0.99, tuy nhiên R^2 trên tập test = 0.88 cho thấy mô hình đã bị overfitting ; sai số trung bình bình phương = 0.1

Với bộ dữ liệu trên, khi áp dụng 2 mô hình vào ta thấy độ chính xác còn thấp. Kết hợp với phân tích tại mục 2.12 về biến Price có xu hướng lệch phải, do vậy ta cần xem xét trường hợp áp dụng mô hình khi xử lý log với biến Price (đưa về phân phối chuẩn).

9. Áp dụng mô hình vào bộ dữ liệu với biến Giá được đưa về phân phối chuẩn

9.1. Mô hình hồi quy

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
linear_reg = LinearRegression()
linear_reg.fit(X2_train, y2_train.T)
y2_pred= linear_reg.predict(X2_test)

print("R2 score on Traing set: %.2f " % linear_reg.score(X2_train,y2_train))
print("R2 score on Testing set: %.2f" % linear_reg.score(X2_test,y2_test))
print('Mean squared error: %.2f'% mean_squared_error(y2_test,y2_pred))
```

```
R2 score on Traing set: 0.92
R2 score on Testing set: 0.90
Mean squared error: 0.09
```

Hình 30 – Áp dụng mô hình hồi quy regression cho bộ dữ liệu với biến Giá được đưa về phân phối chuẩn

*** Nhận xét:**

- Ta thấy với mô hình hồi quy regression, R^2 trên tập train = 0.92 và R^2 trên tập test = 0.90; sai số trung bình bình phương = 0.09.

- Kết quả này tốt hơn kết quả tại mục 8.1.

9.2. Mô hình Random forest

```
from sklearn.ensemble import RandomForestRegressor
rf_reg = RandomForestRegressor()
rf_reg.fit(X2_train, y2_train)
y2_pred= rf_reg.predict(X2_test)
print("R2 score on Traing set: %.2f"% rf_reg.score(X2_train,y2_train))
print("R2 score on Testing set: %.2f"% rf_reg.score(X2_test,y2_test))
print('Mean squared error: %.2f'% mean_squared_error(y2_test,y2_pred))
```

```
R2 score on Traing set: 0.99
R2 score on Testing set: 0.92
Mean squared error: 0.08
```

```
final_model = rf_reg.fit(X2, y2)
```

Hình 31 – Áp dụng mô hình hồi quy Random forest cho bộ dữ liệu với biến Giá được đưa về phân phối chuẩn

*** Nhận xét:**

- Ta thấy với mô hình Random forest, R^2 trên tập train = 0.99 và R^2 trên tập test = 0.92; sai số trung bình bình phương = 0.08.

- Kết quả này tốt hơn kết quả tại mục 9.1.

Sau khi áp dụng 4 mô hình, ta thấy mô hình Random forest cho bộ dữ liệu với biến Price được đưa về phân phố i chuẩn (mục 9.2) cho kết quả tốt nhất.Đối với bộ dữ liệu đang sử dụng thì mô hình nói trên là tốt nhất.

10. Kết luận

Việc thu thập và xử lý dữ liệu là một trong những khó khăn lớn nhất trong nghiên cứu giá xe ô tô đã qua sử dụng . Nghiên cứu này chú trọng vào việc phân tích, xử lý dữ liệu bị thiếu (missing), bị trùng lặp, dữ liệu ngoại lệ, đánh giá tương quan giữa các biến giải thích để tạo ra bộ dữ liệu chuẩn (clearning data) trước khi áp dụng vào mô hình dự đoán.

Trong quá trình thực hiện, nghiên cứu này tập trung vào việc xem xét vai trò, đánh giá tầm quan trọng của các biến giải thích và chọn ra mô hình được xem là tốt nhất. Chọn mô hình tối ưu hay chọn biến số liên quan có ý nghĩa quan trọng vì dữ liệu lớn, phức tạp, số

lượng biến giải thích lớn khiến cho quá trình phân tích thực sự là 1 thử thách. Khi phát hiện và chọn ra được các yếu tố liên quan sẽ giúp khám phá tốt hơn và cho ra kết quả tin cậy hơn. Kết quả này có nghĩa là các yếu tố ảnh hưởng được c họn có thể giải thích hầu hết các yếu tố quyết định giá xe ô tô đã qua sử dụng.

Nhìn chung, đóng góp của nghiên cứu này nằm ở phương pháp nghiên cứu sử dụng kết hợp nhiều phương pháp phân tích, xử lý dữ liệu và đánh giá các chỉ số trong mô hình, tạo điều kiện cho việc định giá xe ô tô đã qua sử dụng.

TÀI LIỆU THAM KHẢO

1. Bài giảng, tài liệu của giảng viên môn Machine learning.

2. Nguyễn Văn Tuấn (2020). "Mô hình hồi quy và khám phá khoa học".

3. Abhishek (2020). "Used Car Price Prediction", https://www.kaggle.com/iabhishekmaurya/used-car-price-prediction

4. Kajal Kumari (2021). "Car Price Prediction – Machine Learning vs Deep Learning" https://www.analyticsvidhya.com/blog/2021/07/car-price-prediction-machine-learning-vs-deep-learning/

5. Tarique Akhtar (2020). "Predicting Car Price using Machine Learning" https://towardsdatascience.com/predicting-car-price-using-machine-learning-8d2df3898f16