TRỰC QUAN HOÁ DỮ LIỆU

ĐỒ ÁN CUỐI KỲ

BÁO CÁO

Tài liệu này báo cáo về các phân tích về Nông nghiệp, Lâm nghiệp và Thuỷ sản trong nước từ năm 2004 đến năm 2019 mà nhóm đã thực hiện trong môn học Trực quan hoá dữ liệu.

Thành viên nhóm 2:

1753008 – Phạm Huỳnh Nhật

1753065 – Nguyễn Nhật Khoa

1753085 – Nguyễn Công Phúc

Khoa Công nghệ Thông tin

Đại học Khoa học Tự nhiên TP HCM

Tháng 09/2020

Mục lục

1 Thông tin nhóm 3

Thông tin nhóm 2 3

2 Đề tài 3

2.1 Tổng quan đề tài 3

Tên đề tài 3

Mô tả đề tài 4

2.2 Dữ liệu 4

Nguồn dữ liệu 4

Mô tả dữ liệu 4

3 Trực quan dữ liệu 6

Sử dụng biểu đồ cột (Bar chart) 6

Sử dụng biểu đồ ngăn xếp (Stack combo chart) 9

Sử dụng biểu đồ Scatter 12

Sử dụng biểu đồ đường (Line chart) 13

Sử dụng biểu đồ gấp khúc (Bump chart) 14

Sử dụng biểu đồ tròn (Pie chart) 15

Sử dụng bản đồ (Mapping) 16

Sử dụng Heatmap 17

4 Phân tích 18

4.1 Rút chọn đặc trưng 18

Giảm chiều 18

4.2 Xử lý biến định tính 20

4.3 Cross - Validation 21

4.4 Xây dựng mô hình 22

Cải tiến mô hình 25

5 Nguồn tham khảo 27

1
Thông tin nhóm

Thông tin nhóm 2

MSSV	Họ tên	Email
1753008	Phạm Huỳnh Nhật	1753008@student.hcmus.edu.vn
1753065	Nguyễn Nhật Khoa	1753065@student.hcmus.edu.vn
1753085	Nguyễn Công Phúc	1753085@student.hcmus.edu.vn

2
Đề tài

2.1
Tổng quan đề tài

Tên đề tài

Phân tích tình hình mua bán xe ở Việt Nam gần đây

Mô tả đề tài

Trong bài báo cáo này, nhóm sẽ trình bày về tình hình mua bán xe ở Việt Nam gần đây. Nhóm sẽ trực quan dữ liệu dưới nhiều biểu đồ tương ứng với nhiều khía cạnh khác nhau (các thuộc tính khác nhau). Các biểu đồ được trình bày thể hiện cho thấy được cách quan sát dữ liệu từ đơn giản đến phức tạp, từ quan hệ độc lập đến quan hệ phụ thuộc. Sử dụng cùng lúc nhiều loại biểu đồ kết hợp với màu sắc đồng bộ, giúp cho người đọc, người xem có cách nhìn trực quan một cách dễ dàng những vấn đề mà nhóm trình bày.

2.2
Dữ liệu

Nguồn dữ liệu

Dữ liệu được sử dụng trong bài báo cáo này được lấy từ trang https://oto.com.vn/ .

Mô tả dữ liệu

Dữ liệu sẽ bao gồm một số thuộc tính như sau: car_model, km, imp_exp, km_1, imp_exp_1, car_type, out_color, in_color, door_num, seat_num, new_old, car_year, title, price, area, poster_name, poster_add, poster_tel.

Để trực quan trong bài báo cáo này thì nhóm đã bỏ đi một số thuộc tính không cần thiết và thêm thuộc tính brand vào như sau:

- car_model: Mẫu xe
- km: Số km đã đi
- imp_exp: Nhập khẩu/lắp ráp trong nước
- car_type: Loại xe
- out_color: Màu bên ngoài xe
- in_color: Màu bên trong xe
- door_num: Số cửa
- seat_num: Số chỗ ngồi

- new_old: Xe mới/Xe cũ
- car_year: Năm sản xuất của xe
- price: Giá bán
- area: Khu vực giao dịch
- brand: Hãng xe

Sẽ có nhiều câu hỏi được đặc ra dành cho dữ liệu trên, ví dụ như là: Hãng nào được ưa chuộng nhất, loại xe nào giá thành cao nhất, độ ưa chuộng các hãng với các dòng xe trải qua từng đời như thế nào, … Những câu hỏi này sẽ được trả lời ở phần Trực quan dữ liệu.

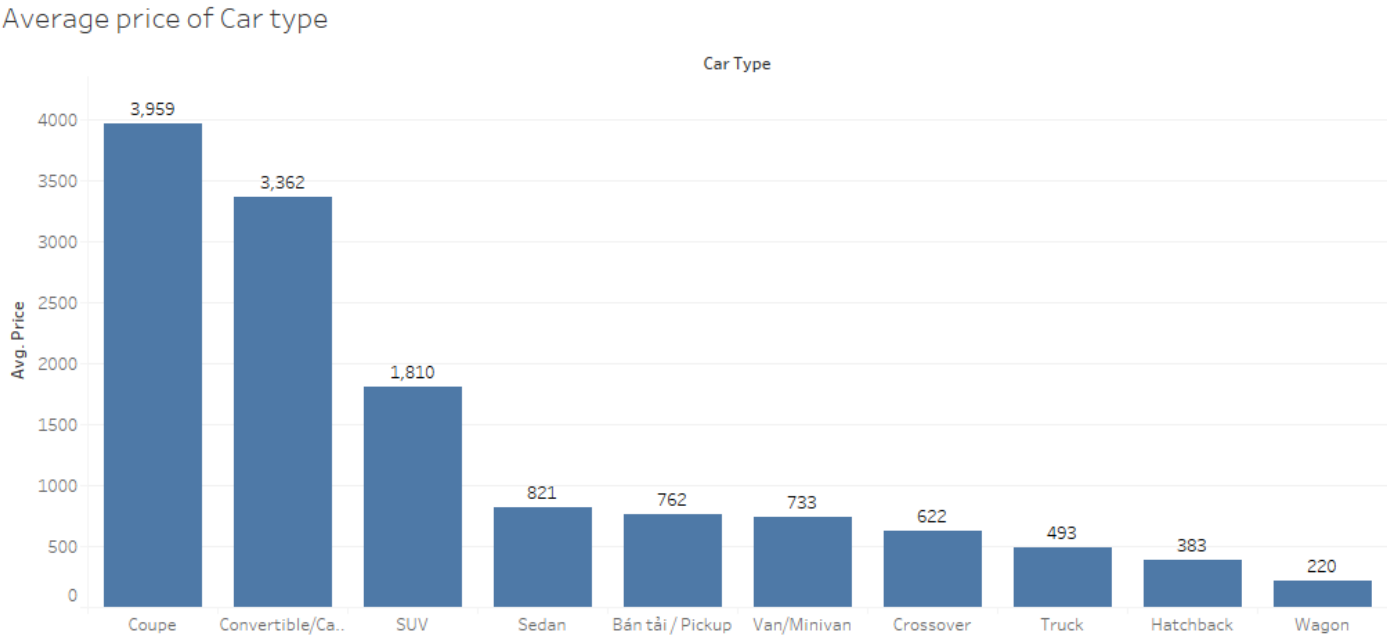
Dữ liệu có khoảng 39000 mẫu, sau khi được tiền xử lý thì còn khoảng 27000 mẫu.

3

Trực quan dữ liệu

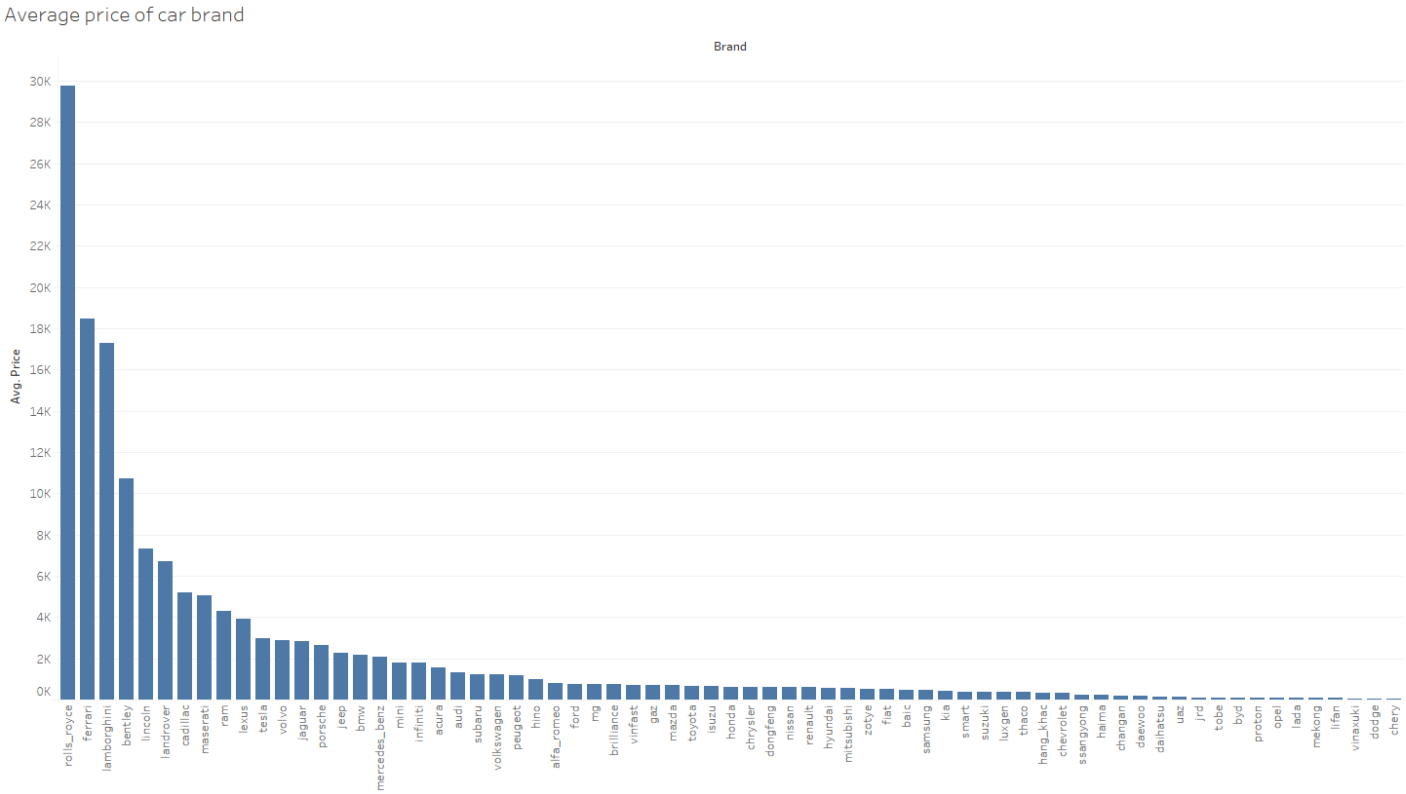
Sử dụng biểu đồ cột (Bar chart)

- Giá thành trung bình theo từng loại xe



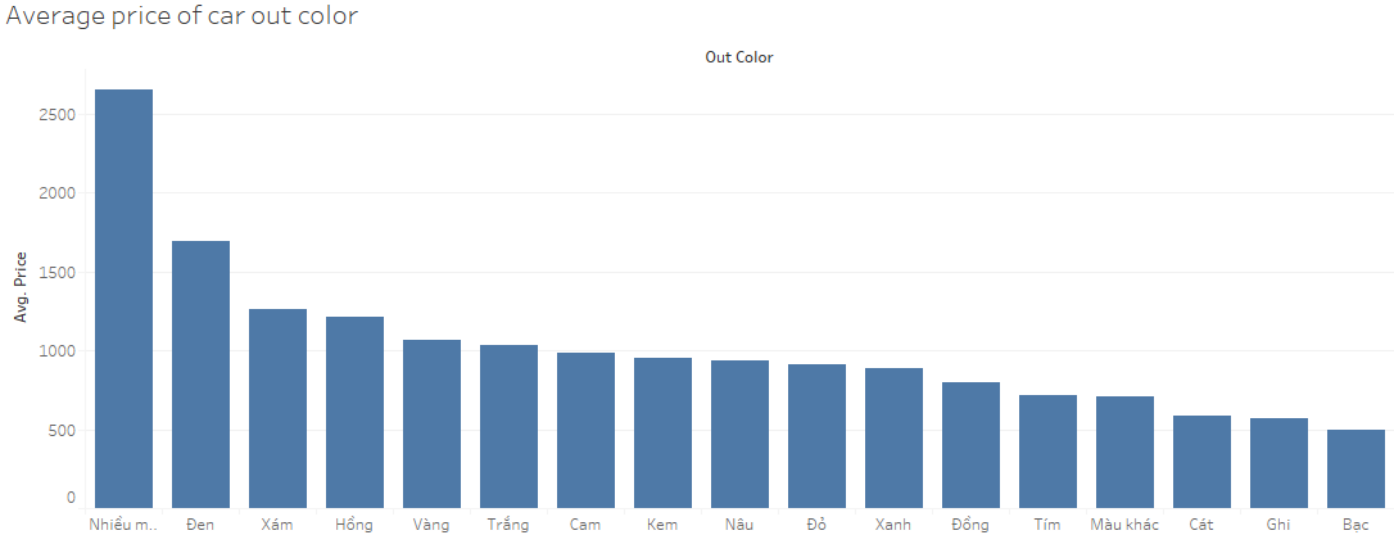
- Phân tích và đánh giá:
  Dựa vào biểu đồ ta có thể thấy được rằng, những xe thuộc loại "Coupe" có giá trung bình cao nhất và xe thuộc loại "Wagon" có giá trung bình thấp nhất. Điều này là rõ ràng bởi vì Coupe là nói chung cho các dòng siêu xe, trung bình giá khoảng 4 tỷ đồng. Những loại xe từ hạng 3 trở đi có giá tiền tầm trung. Từ đó suy ra biểu đồ này thể hiện được sự ảnh hưởng của loại xe đối với giá thành.

- Giá thành trung bình theo từng hãng xe



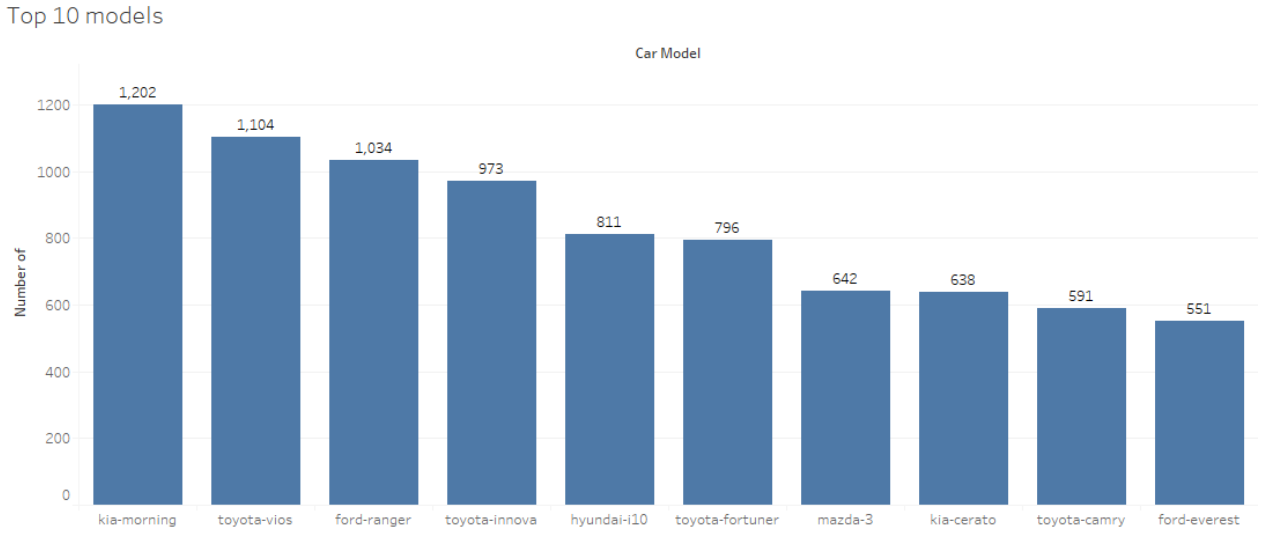
- Phân tích và đánh giá:
  Rõ ràng, các hãng xe thuộc dòng siêu xe đứng top đầu của biểu đồ. Kế đến là các hãng xe quen thuộc được sử dụng ở Việt Nam. Các xe thuộc những hãng này được sử dụng rộng rãi bởi các công ty taxi hay là xe công nghệ. Bởi giá thành tầm trung và chi phí bảo dưỡng, thuế thấp. Điều này rất phù hợp cho việc đáp ứng nhu cầu công việc.

- Giá thành trung bình theo từng màu bên ngoài xe



- Phân tích và đánh giá:
  Ngoài những yếu tố về hãng xe, hay loại xe thì màu sắc cũng ảnh hưởng tới giá cả của xe. Ta có thể thấy xe càng nhiều màu thì giá thành càng cao. Kế tiếp là màu đen thông dụng, được ưa chuộng bởi màu đen rất sang trọng và phù hợp đường xá, khí hậu Việt Nam (nóng ẩm, mưa nhiều). Chính vì vậy, màu đen cao hơn hẳn những màu cơ bản khác. Những màu còn lại thì giá không chênh lệch nhiều.

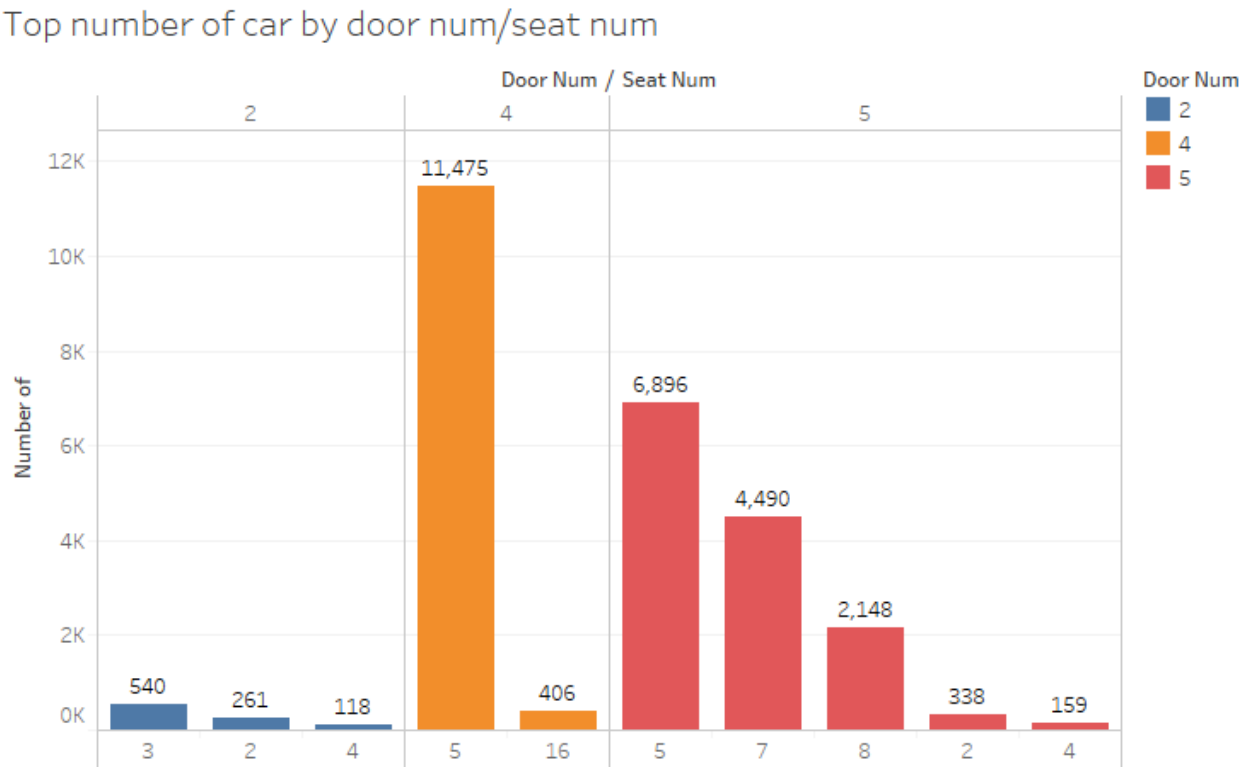
- Top 10 mẫu xe được bán thành công nhất



- Phân tích và đánh giá:
  Dễ dàng nhìn ra được, mẫu xe "Kia-morning" và "Toyota-vios" có số lượng giao dịch lớn nhất. Bởi vì giá thành khá phù hợp với đa số người Việt. Kiểu dáng nhỏ gọn, tiết kiệm diện tích và chi phí đỗ xe. Tiếp theo là mẫu xe bán tải "Ford-Ranger" rất được ưu chuộng tại Việt Nam. Sau đó là hãng xe "Toyota-innova" quen thuộc với các hãng taxi trong nước.

Sử dụng biểu đồ ngăn xếp (Stack combo chart)

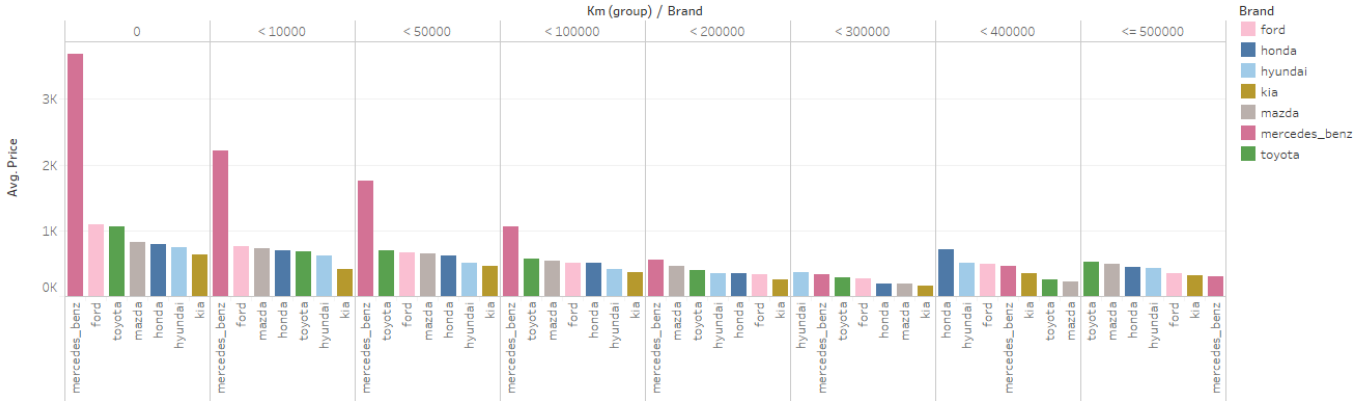
- Các xe có số lượng cửa và chỗ ngồi nào được bán nhiều nhất



- Phân tích và đánh giá:
  Dễ dàng nhìn ra ở Việt Nam, xe có 4 cửa 5 chỗ là chiếm đa số về số lượng, vì nó được sử dụng rộng rãi nhất nên nhu cầu mua bán của loại này rất cao. Nhất là xe gia đình hay trong lĩnh vực taxi và các dịch vụ thuê xe hợp đồng. Đối với loại xe có 2 cửa, thì đó chính là những loại siêu xe, xe tải và các loại xe khác. Loại này thì ít được giao dịch mua bán hơn.

- Giá trung bình dựa trên quảng đường đã đi của một số hãng xe

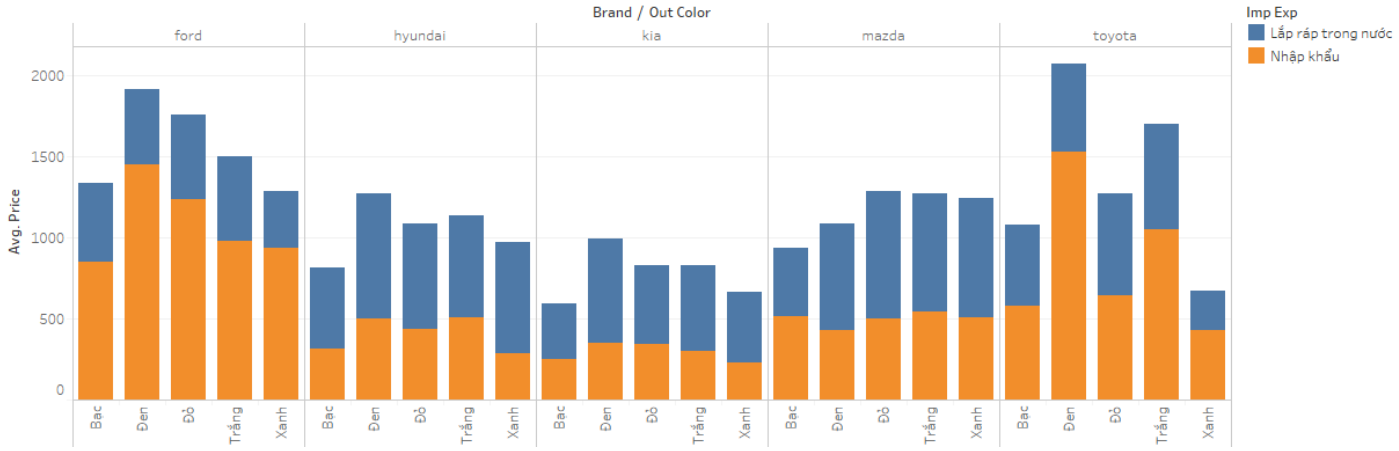
Top 7 by km and brand



- Phân tích và đánh giá:
  Ở biểu đồ này, dòng "Mercedes Benz" có giá cao vượt trội cho dù là xe cũ hay xe mới. Nó chỉ bị rớt giá khi số km đã đi hơn 400000. Điều này chứng tỏ "Mercedes Benz" khá bền, khá danh tiếng nên khó mất giá. Những hãng xe còn lại thì giá trị có xu hướng giảm khi đi càng nhiều km. Nhưng không giao động mạnh như "Mercedes Benz".

- Giá thành trung bình của các hãng xe được bán nhiều nhất với loại xe nhập và xe lắp ráp trong nước

Stack combo chart for price base on brand and out color

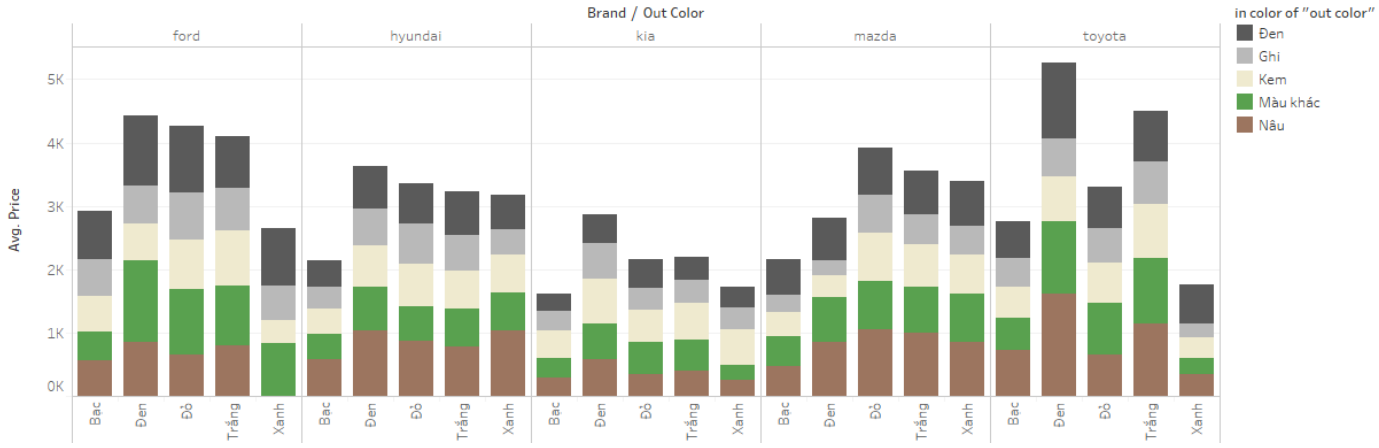


- Phân tích và đánh giá:
  Không chỉ màu sắc, hãng xe ảnh hưởng tới giá cả mà nguồn gốc là xe nhập hay xe lắp ráp trong nước cũng ảnh hưởng tới giá cả. Lấy hãng "Ford" là ví dụ, ta có thể thấy được giá thành khi xe được nhập từ nước ngoài về có giá cao hơn hẳn

xe được lắp trong nước. Điều này là do một phần bởi các loại thuế hay là các chi phí phát sinh khác đã đưa giá xe tăng lên nhiều. Ngược lại, về hãng "Kia", các xe hãng này được lắp ráp trong nước có giá cao hơn. Vì chi phí nhập linh kiện khá là cao, ảnh hưởng trực tiếp đến giá xe. Ngoài ra, các xe "Kia" nhập khẩu thường là xe cũ nên giá sẽ thấp hơn các loại xe mới được lắp trong nước.

- Giá thành trung bình của các hãng xe được bán nhiều nhất với các màu sắc trong và ngoài xe

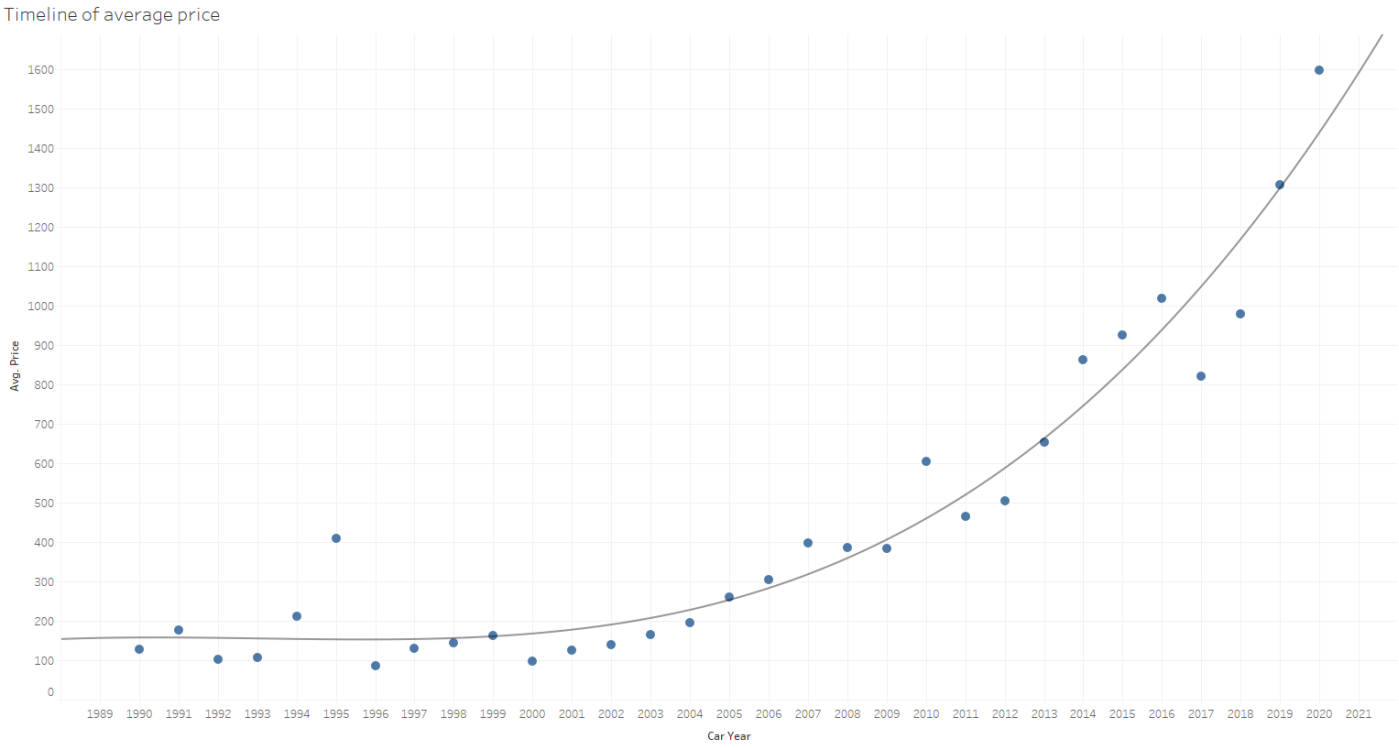
Stack combo chart for price base on brand and out color



- Phân tích và đánh giá:
  Nhìn vào biểu đồ này, ta thấy được xét một hãng xe, từ một màu bên ngoài xe bất kỳ, ta có cái nhìn tổng quan, có thể so sánh được với loại màu đó thì các màu khác nhau bên trong xe sẽ có sự khác biệt về giá cả. Ví dụ với hãng "Ford" với màu bên ngoài là Bạc thì với các màu bên trong khác nhau sẽ không có sự chênh lệch quá nhiều về giá cả. Còn với hãng "Toyota" có màu bên ngoài là màu Đen, thì màu Nâu bên trong sẽ có giá khoảng 1,6 tỷ. Giá này cao hơn hẳn với những màu bên trong nội thất còn lại.

Sử dụng biểu đồ Scatter

- Xu hướng giá của xe qua từng đời



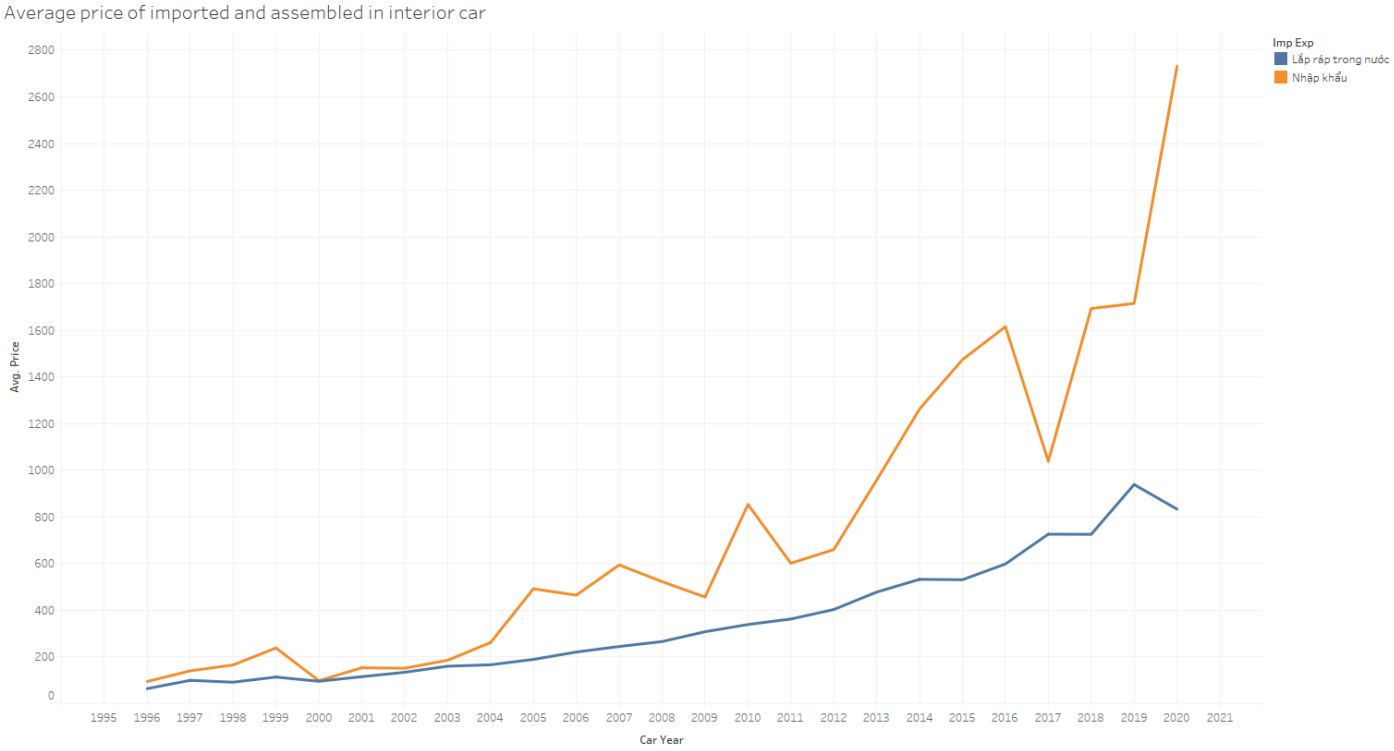
- Phân tích và đánh giá:

Avg. Price = 0.0678166*Car Year^3 + -405.491*Car Year^2 + 808172*Car Year + -5.36914e+08
R-Squared: 0.941776
P-value: < 0.0001

Nhìn chung, giá xe sẽ tỉ lệ thuận với năm sản xuất xe, nói cách khác giá xe có xu hướng tăng dần theo thời gian. Một cách rõ ràng hơn thì nó có mối quan hệ hồi quy (giống như ảnh trên). Từ đó có thể dự đoán với một năm nào đó bất kỳ trong tương lai, ta có thể đoán được giá trị trung bình của chiếc xe đó.
Rõ ràng, mô hình dự đoán có điểm R-Squared rất cao (khoảng 0.94), suy ra được mô hình được xây dựng khá tốt để dự đoán. Bởi vì chỉ số P-value giữa 2 biến rất nhỏ (< 0.0001), cho thấy mối tương quan chặt chẽ giữa 2 trường dữ liệu.

Sử dụng biểu đồ đường (Line chart)

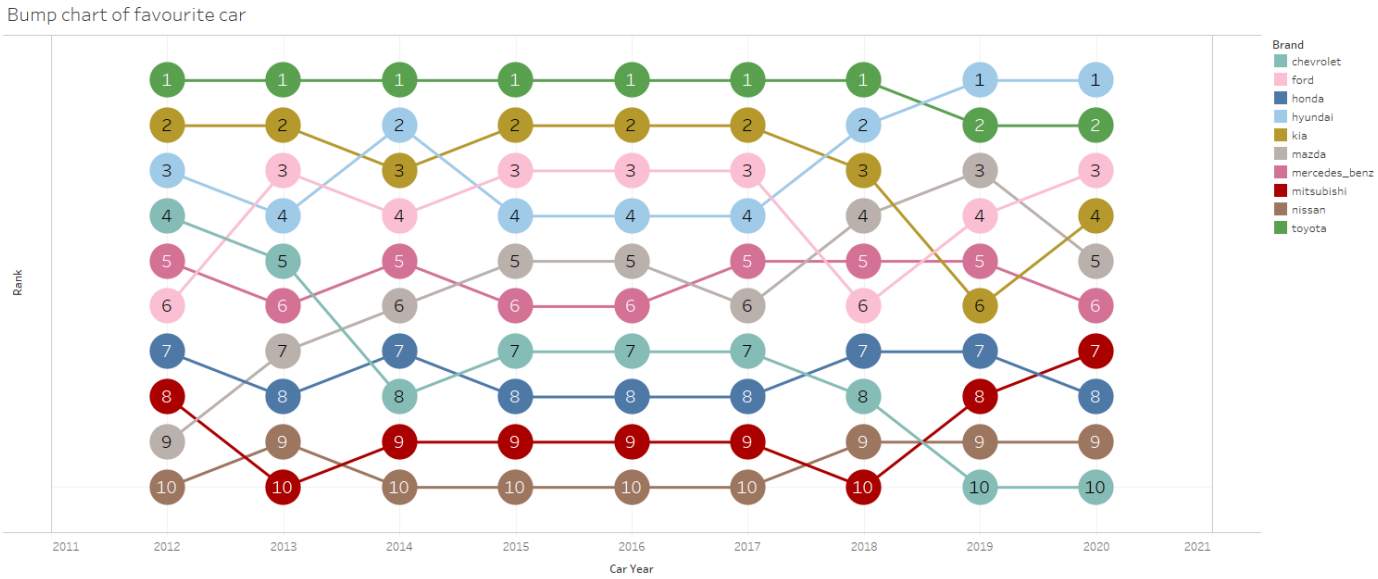
- Giá thành trung bình đối với loại xe nhập và xe lắp ráp trong nước



- Phân tích và đánh giá:
  Nhìn chung, cả hai đường biểu diễn cho giá xe thông qua các năm đều có xu hướng tăng. Nhưng các xe được lắp ra chỉ tăng nhẹ theo thời gian. Còn với xe nhập khẩu thì giao động mạnh. Rõ ràng nhất là tăng cực mạnh từ năm 2017 đến nay, bởi năm 2017 là cột mốc áp dụng thuế cực mạnh lên những xe nhập khẩu.

Sử dụng biểu đồ gấp khúc (Bump chart)

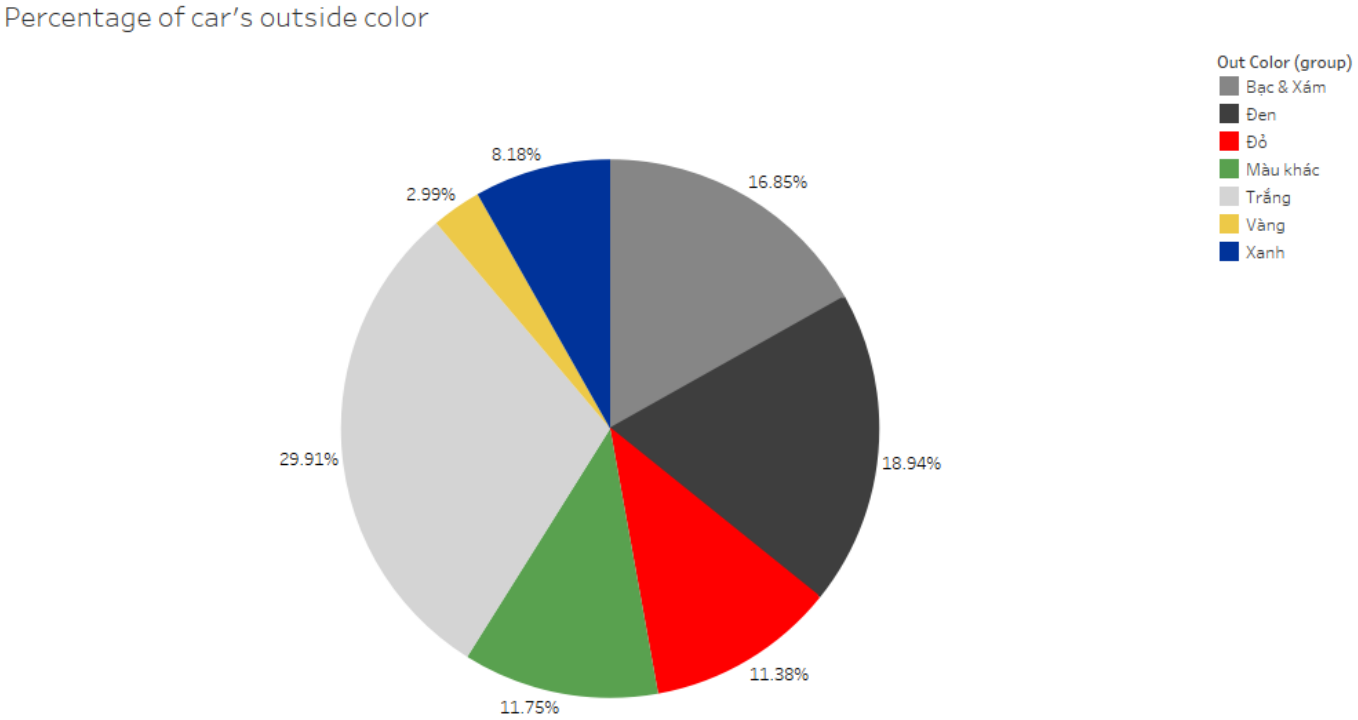
- Thứ hạng các hãng xe trải qua từng đời



- Phân tích và đánh giá:
  Màu xanh lá được đại diện cho "Toyota" luôn nằm ở vị trí thứ nhất với các dòng xe từ năm 2012 đến năm 2018. Còn từ năm 2019 đến 2020, nhường vị trí đầu cho "Hyundai", và xuống top 2. Một ví dụ cho việc dao động thất thường nhất là "Ford". Trong thời gian khảo sát thì việc mua bán xe của hãng "Chevrolet" hay còn gọi là thị hiếu của người Việt với hãng này giảm rõ rệt với các dòng xe sau này.

Sử dụng biểu đồ tròn (Pie chart)

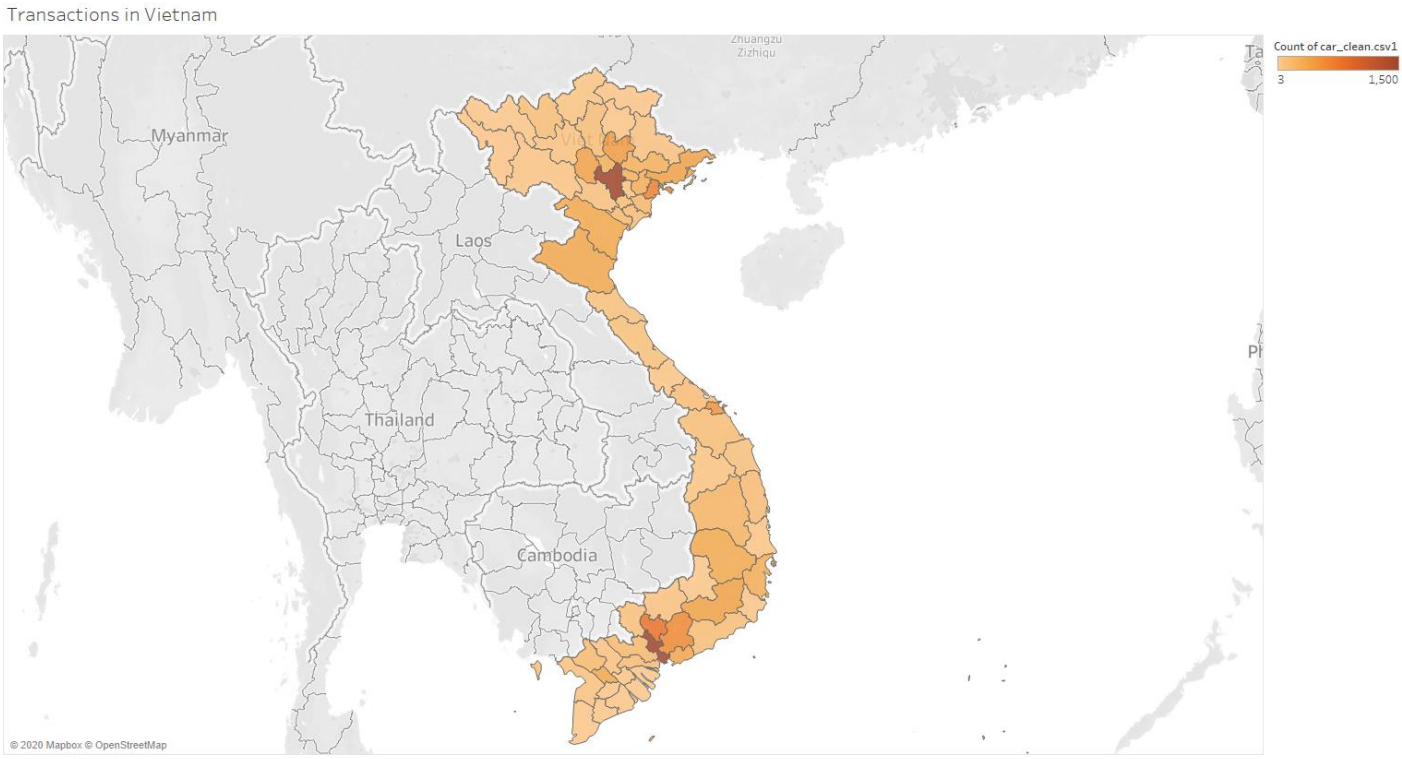
- Tỉ lệ của các xe có màu bên ngoài khác nhau



- Phân tích và đánh giá:
  Rõ ràng nhìn ra được màu trắng luôn được ưu chuộng với bất kỳ dòng xe nào với tỉ lệ gần 30%. Mặc dù màu đen có giá cao hơn nhưng số lượt giao dịch mua bán xe màu trắng lại cao hơn. Kế đến là những xe có tông màu "Bạc, Xám" và Đen với tỉ lệ từng loại gần 20%. Từ đây ta rút ra được ý nghĩa màu sắc sẽ ảnh hưởng đến giá cả.

Sử dụng bản đồ (Mapping)

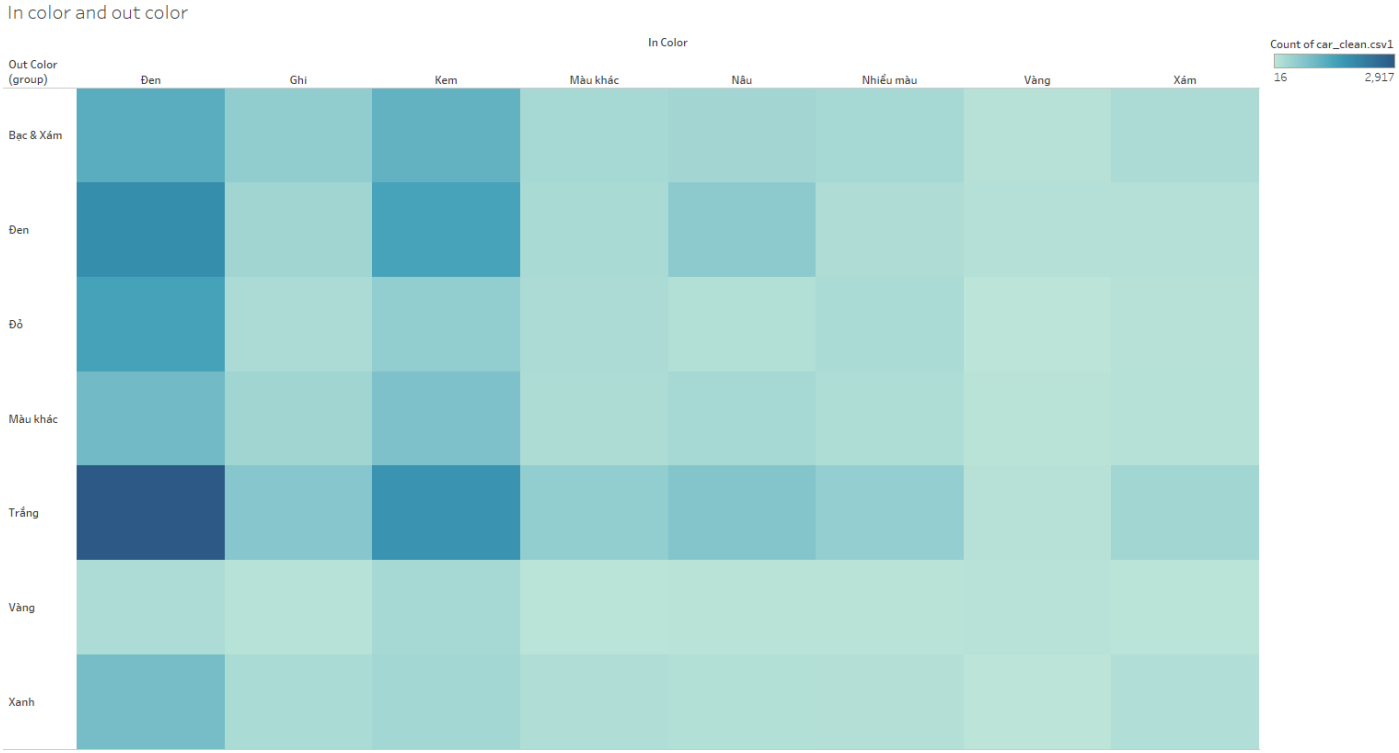
- Số lượng giao dịch mua bán xe trên từng tỉnh thành ở Việt Nam



- Phân tích và đánh giá:
  Tổng lượng giao dịch ở các tỉnh thành ở Việt Nam được biểu diễn dựa vào mức độ đậm nhạt của tông màu trên. Rõ ràng ở Thủ đô Hà Nội và Thành phố Hồ Chí Minh có lượng giao dịch vượt trội. Bởi đây là 2 thành phố phát triển nhất Việt Nam. Các tỉnh lân cận 2 thành phố trên cũng có số lượng giao dịch cao hơn hẳn những tỉnh xa hơn. Về miền Trung thì rõ ràng ở Đà Nẵng được nổi trội hơn hẳn những tỉnh khác.

Sử dụng Heatmap

- Mối tương quan giữa màu trong xe và màu ngoài xe



- Phân tích và đánh giá:
  Nhìn vào Heatmap này, ta có thể thấy được rằng xe có màu trắng bên ngoài và bên trong màu đen sẽ được giao dịch mua bán nhiều nhất. Đa số nội thất bên trong có màu Đen và màu Kem sẽ được quan tâm hơn những màu khác bởi vì bên trong xe rất khó vệ sinh và chi phí vệ sinh của 2 màu này không quá cao.

4

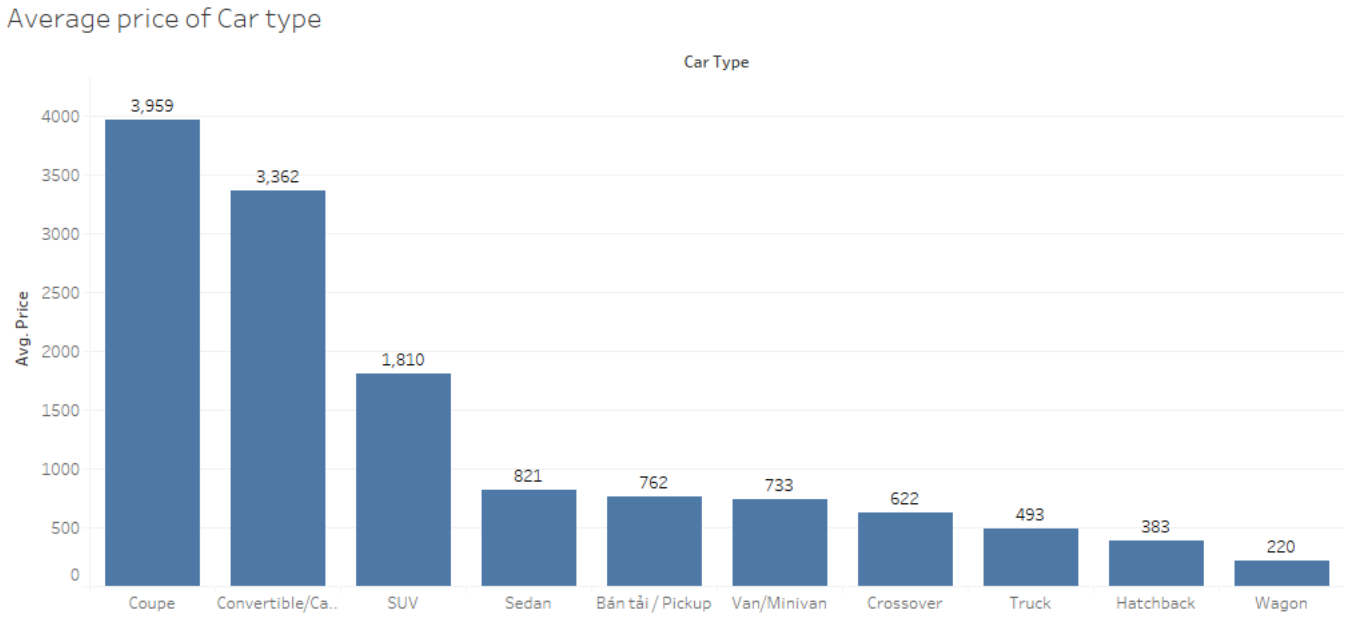
Phân tích

4.1

Rút chọn đặc trưng

Giảm chiều

Nguyên tắc: Chia giỏ dựa theo giá trung bình, các thuộc tính cần xử lý là imp_exp, car_type, out_color, in_color, door_num, seat_num, new_old, brand. Trong đó brand là một thuộc tính được tách từ car_model.

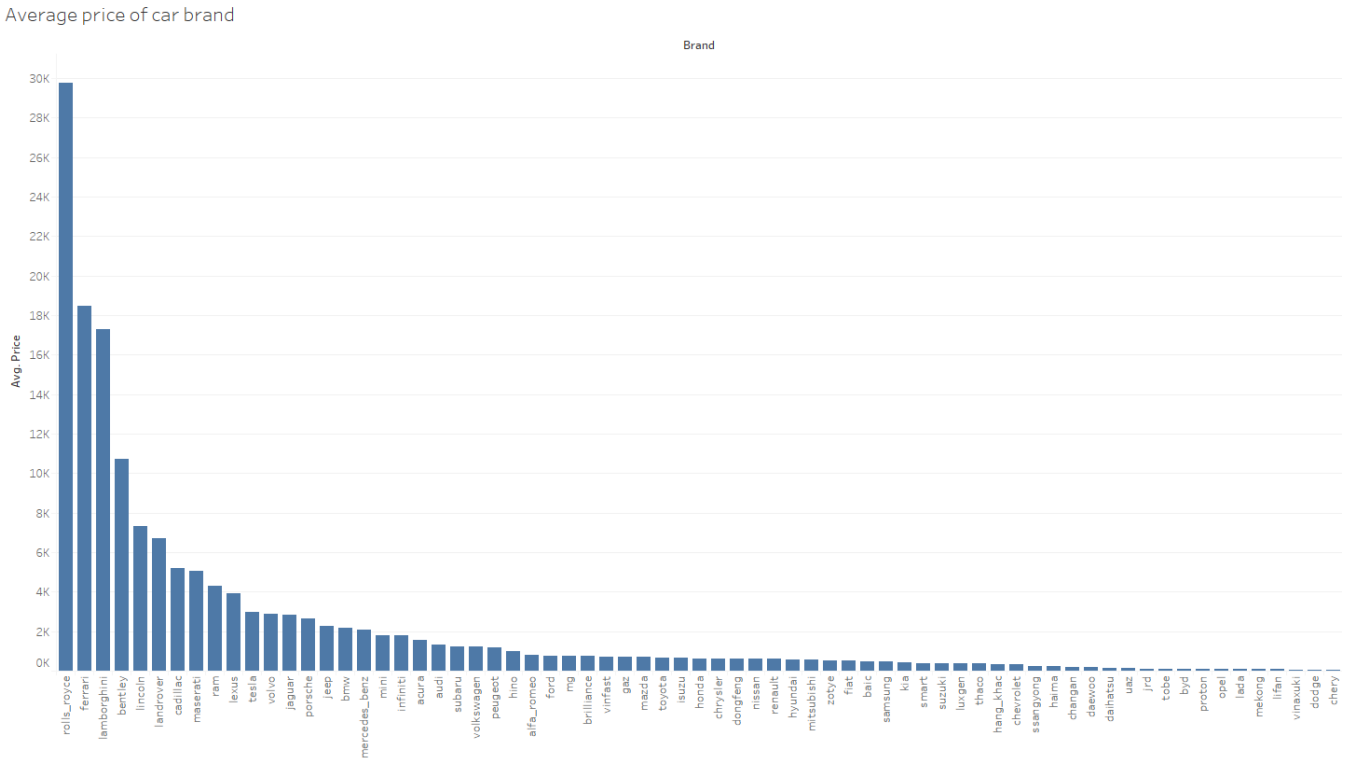


Ví dụ như ta có biểu đồ giá tiền trung bình theo từng kiểu xe như trên. Có thể được chia thành “7 bins” như sau:

- Convertible/Cabriolet, Coupe

- SUV
- Sedan
- Van/Minivan, Bán tải / Pickup
- Crossover
- Truck
- Wagon, Hatchback

Đối với dữ liệu có nhiều unique như brand (= 69)



Ta nhận thấy rằng, sự phân bố giá tiền không đều. Đặc biệt là ở mức trên 1 tỷ đồng. Nên ta cần chia giỏ dựa trên các phân vị. Ở đây là 10 phân vị:

	Giá tiền (triệu đồng)
10%	87.000000
20%	177.342857
30%	371.503734
40%	526.292216
50%	635.928734

60%	755.733333
70%	1277.319837
80%	2400.352803
90%	5056.289655
Lớn 90%	Lớn hơn 5056.289655

Sau khi xử lý hết tất cả các biến, ta được bảng mô tả biến định tính như sau:

	imp_exp	car_type	out_color	in_color	door_num	seat_num	new_old	brand
count	27629	27629	27629	27629	27629	27629	27629	27629
unique	2	7	7	9	4	9	2	10
top	Lắp ráp trong nước	car_type_bin5	out_color_bin4	in_color_bin5	door_num_bin3	seat_num_bin5	Xe cũ	brand_bin6
freq	16359	9662	9088	9591	15217	18463	17910	7909

4.2
Xử lý biến định tính

Đối với mô hình hồi quy tuyến tính, dữ liệu đầu vào phải là số nên các biến định tính cần phải được "Dummy". Ví dụ như imp_exp gồm có 2 giá trị là "lắp ráp trong nước" và "nhập khẩu". Nên cần tách thành 2 thuộc tính mới là "lắp ráp trong nước" và "nhập khẩu".

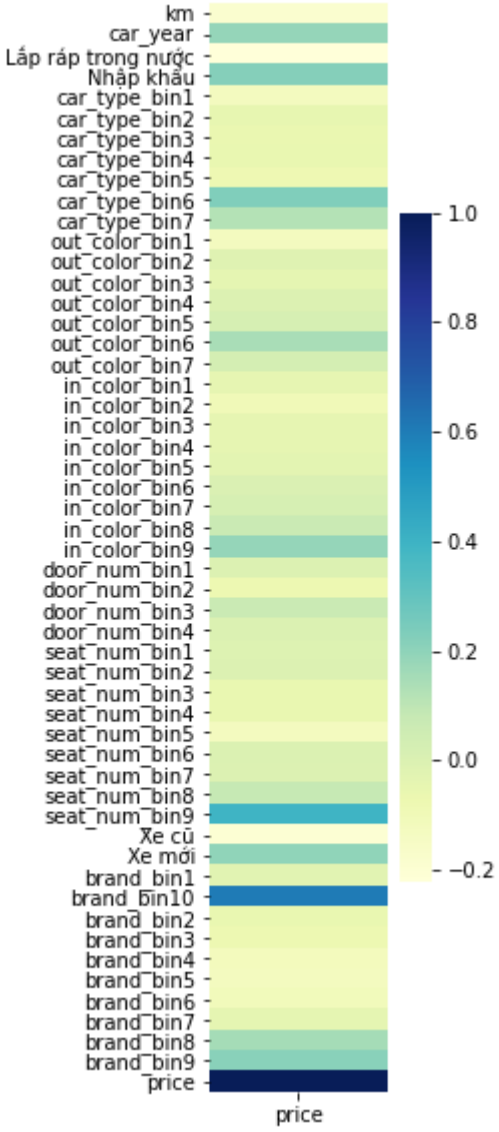
Sau khi Dummy các dữ liệu định tính xong, ta nhận thấy rằng chúng đều ở dưới dạng nhị phân. Trong khi đó các biến định lượng như "km" và "car_year" lại mang giá trị rất lớn. Nên ta cần chuẩn hoá chúng lại trên cùng một miền giá trị.

	km	car_year	Lắp ráp trong nước	Nhập khẩu	car_type_bin1	car_type_bin2	car_type_bin3	car_type_bin4	car_type_bin5
1777	0.340	0.433333	1	0	0	0	0	0	1
9426	0.000	0.933333	1	0	0	1	0	0	0
12199	0.138	0.733333	1	0	0	0	1	0	0
7743	0.000	1.000000	1	0	0	0	0	0	1
26674	0.193	0.033333	1	0	0	0	0	0	0

car_type_bin6	...	brand_bin10	brand_bin2	brand_bin3	brand_bin4	brand_bin5	brand_bin6	brand_bin7	brand_bin8	brand_bin9	price
0	...	0	1	0	0	0	0	0	0	0	58
0	...	0	0	0	0	1	0	0	0	0	665
0	...	0	0	0	1	0	0	0	0	0	326
0	...	0	0	0	0	1	0	0	0	0	540
1	...	0	0	0	0	0	0	0	0	0	50

4.3
Cross - Validation

Chia tập huấn luyện và tập kiểm thử để kiểm thử mô hình. Sau khi chia ta được bảng mô tả hệ số tương quan dành cho tập huấn luyện.



4.4
Xây dựng mô hình

Xây dựng mô hình dựa trên công thức Original Least Square (OLS).

Cách xây dựng: Loại bỏ những cột có P-value > 0.05 (threshold). Quá trình lặp lại nhiều lần cho tới khi tất cả các cột đều có P-value <= 0.05. Đồng thời cũng sử dụng kiểm định đa cộng tuyến với quy tắc như sau:

Giá trị VIF	tương quan
$[1, 2)$	không có mối tương quan giữa biến độc lập này và bất kỳ biến nào khác
$[2, 5)$	có một mối tương quan vừa phải, nhưng không nghiêm trọng
$[5, 10)$	có một mối tương quan và khá nghiêm trọng
$[10, \infty)$	chắc chắn có đa cộng tuyến

Mô hình được đánh giá dựa trên R2, kiểm định F, AIC, BIC. Trong đó R2 và F càng lớn càng tốt. AIC và BIC càng nhỏ càng tốt.

Sau quá trình xây dựng mô hình, ta được bảng kết quả như sau:

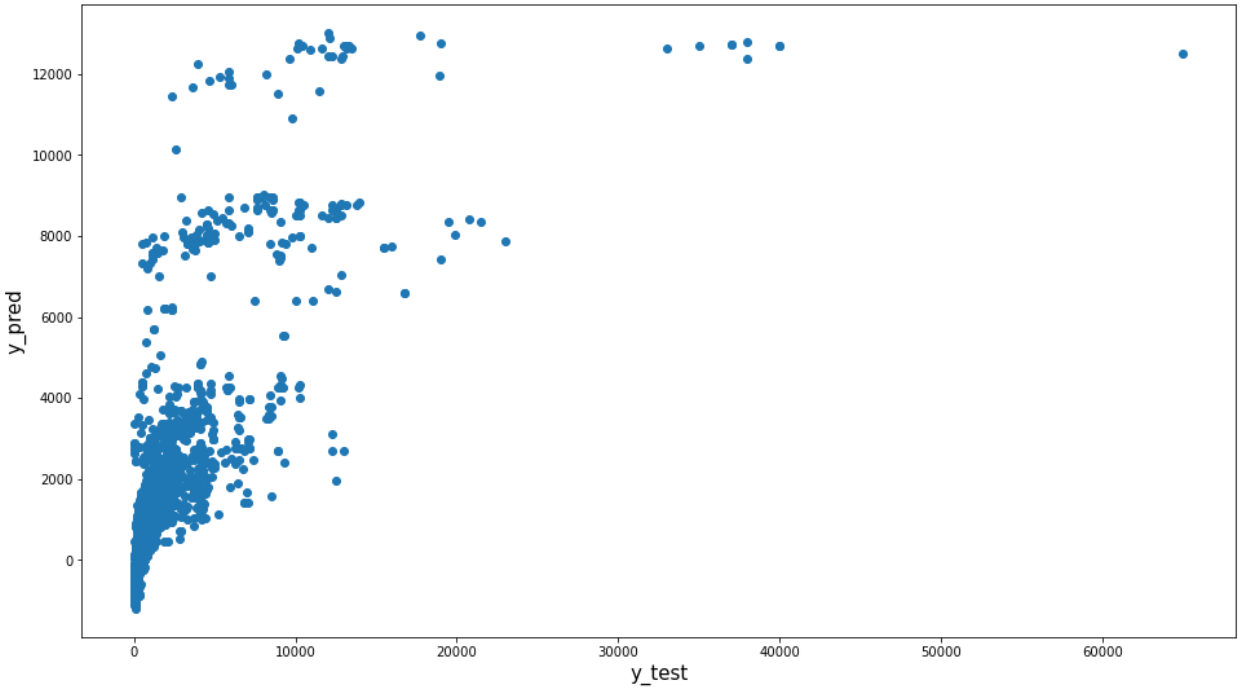
	coef	std err	t	P>|t|	[0.025	0.975]
const	-170.4256	73.888	-2.307	0.021	-315.252	-25.599
car_year	1938.2822	74.698	25.948	0.000	1791.868	2084.696
Nhập khẩu	194.7513	21.517	9.051	0.000	152.576	236.926
car_type_bin6	182.5887	31.512	5.794	0.000	120.823	244.354
car_type_bin7	-763.3692	120.092	-6.357	0.000	-998.759	-527.980
out_color_bin3	-66.5057	23.932	-2.779	0.005	-113.414	-19.597
out_color_bin5	126.2721	48.772	2.589	0.010	30.674	221.870
out_color_bin6	250.9833	26.532	9.460	0.000	198.978	302.989
in_color_bin4	-94.1217	41.143	-2.288	0.022	-174.765	-13.479
in_color_bin5	-79.7551	21.829	-3.654	0.000	-122.542	-36.968
in_color_bin8	313.3556	59.034	5.308	0.000	197.644	429.067
door_num_bin2	106.3201	25.318	4.199	0.000	56.695	155.945
seat_num_bin4	294.5529	42.244	6.973	0.000	211.752	377.354
seat_num_bin8	209.4706	34.187	6.127	0.000	142.462	276.479
seat_num_bin9	3945.7154	90.770	43.469	0.000	3767.799	4123.631
Xe mới	390.1080	25.693	15.183	0.000	339.747	440.469
brand_bin1	-1491.5308	211.925	-7.038	0.000	-1906.921	-1076.141
brand_bin10	5971.8033	73.883	80.828	0.000	5826.987	6116.620
brand_bin2	-976.1894	86.148	-11.332	0.000	-1145.046	-807.333
brand_bin3	-1272.7399	58.101	-21.906	0.000	-1386.623	-1158.857
brand_bin4	-1317.5340	42.100	-31.295	0.000	-1400.053	-1235.015
brand_bin5	-1362.2575	37.458	-36.368	0.000	-1435.678	-1288.837
brand_bin6	-1188.1898	37.063	-32.058	0.000	-1260.837	-1115.542
brand_bin7	-1217.1923	40.955	-29.720	0.000	-1297.468	-1136.917
brand_bin9	1253.8198	66.942	18.730	0.000	1122.608	1385.032

Với các tiêu chí đánh giá là:

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.557
Model:	OLS	**Adj. R-squared:**	0.557
Method:	Least Squares	**F-statistic:**	1086.
Date:	Sat, 19 Sep 2020	**Prob (F-statistic):**	0.00
Time:	23:28:41	**Log-Likelihood:**	-1.7910e+05
No. Observations:	20721	**AIC:**	3.583e+05
Df Residuals:	20696	**BIC:**	3.585e+05
Df Model:	24		
Covariance Type:	nonrobust		

Biểu đồ lan truyền (sai lệch) của y_test và y_pred:



Các điểm được thể hiện càng gần đường chéo chính thì chứng tỏ tỉ lệ đúng càng cao.

Cải tiến mô hình

Sau khi vấn đáp, Thầy có yêu cầu nhóm về tìm hiểu cách nâng cao độ tin cậy của mô hình. Nói cách khác, tăng điểm R-squared. Sau khi tìm hiểu, nhóm đã tìm ra cách để cải tiến như sau:

- Thử chọn mô hình hồ quy tuyến tính theo dạng đa thức bậc K
- Với K = 2, số lượng đặc trưng từ 51 tăng lên 1431 đặc trưng, dùng 1431 đặc trưng này xây dựng mô hình và ta được bảng mô tả kết quả:

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.814
Model:	OLS	**Adj. R-squared:**	0.808
Method:	Least Squares	**F-statistic:**	134.9
Date:	Sun, 20 Sep 2020	**Prob (F-statistic):**	0.00
Time:	19:28:04	**Log-Likelihood:**	-1.7013e+05
No. Observations:	20721	**AIC:**	3.416e+05
Df Residuals:	20069	**BIC:**	3.467e+05
Df Model:	651		
Covariance Type:	nonrobust		

Rõ ràng, mô hình có số điểm R-squared tăng vượt bậc từ 0.557 (với K = 1) lên 0.814 (với K = 2). Tuy nhiên con số 1431 đặc trưng là tương đối tốn kém khi nói về tài nguyên sử dụng. Vì vậy nên ta cần tạo ra một mô hình có số lượng đặc trưng tối ưu hơn bằng cách sử dụng quy trình loại bỏ những đặc trưng mang p value lớn một cách tuần tự. Sau quá trình tối ưu ta được mô hình có bảng mô tả như sau:

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.811
Model:	OLS	**Adj. R-squared:**	0.809
Method:	Least Squares	**F-statistic:**	383.1
Date:	Sun, 20 Sep 2020	**Prob (F-statistic):**	0.00
Time:	19:20:51	**Log-Likelihood:**	-1.7027e+05
No. Observations:	20721	**AIC:**	3.410e+05
Df Residuals:	20490	**BIC:**	3.428e+05
Df Model:	230		
Covariance Type:	nonrobust		

Tuy độ khớp mô hình so với tập train (và trên tập test cả 2 có điểm số là 0.75) không có quá nhiều thay đổi nhưng F-statistic đã được cải thiện hơn rất nhiều. Hơn nữa số lượng đặc trưng mà mô hình này dùng chỉ nằm ở mức 1/3 (tương đương với 343 đặc trưng) so với mô hình ban đầu. Điều này có thể tạo nên một sự khác biệt lớn về mặt tốc độ tính toán của hai mô hình. Cách poly trong bài báo cáo này gọi là Brute Force và thường thì không ai làm như thế nếu dữ liệu lớn.

5

Nguồn tham khảo

[1]:
Ebook PHÂN TÍCH THỐNG KÊ của Thầy Bùi Tiến Lên được cung cấp trên MOODLE.

https://courses.ctda.hcmus.edu.vn/

[2]:
https://www.kaggle.com/goyalshalini93/car-price-prediction-linear-regression-rfe

[3]:
https://playfairdata.com/tableau-201-make-dynamic-dual-axis-bump-charts/

[4]:
https://help.tableau.com/current/pro/desktop/en-us/buildexamples_maps.htm#:~:text=A%20map%20view%20with%20one,already%20have%20in%20the%20view.

[5]:

https://www.tutorialgateway.org/stacked-bar-chart-in-tableau/

[6]:
https://kb.tableau.com/articles/howto/stacked-bar-chart-multiple-measures

[7]:

https://oto.com.vn/