Desafio Cientista de Dados

Marcos Paulo Cortina Trivelato

Resumo

Você foi alocado(a) em um time da Indicium que está trabalhando atualmente junto a um cliente que o *core business* é compra e venda de veículos usados. Essa empresa está com dificuldades na área de revenda dos automóveis usados em seu catálogo.

Para resolver esse problema, a empresa comprou uma base de dados de um *marketplace* de compra e venda para entender melhor o mercado nacional, de forma a conseguir precificar o seu catálogo de forma mais competitiva e assim recuperar o mau desempenho neste setor.

Seu objetivo é analisar os dados para responder às perguntas de negócios feitas pelo cliente e criar um modelo preditivo que precifique os carros do cliente de forma que eles fiquem o mais próximos dos valores de mercado.

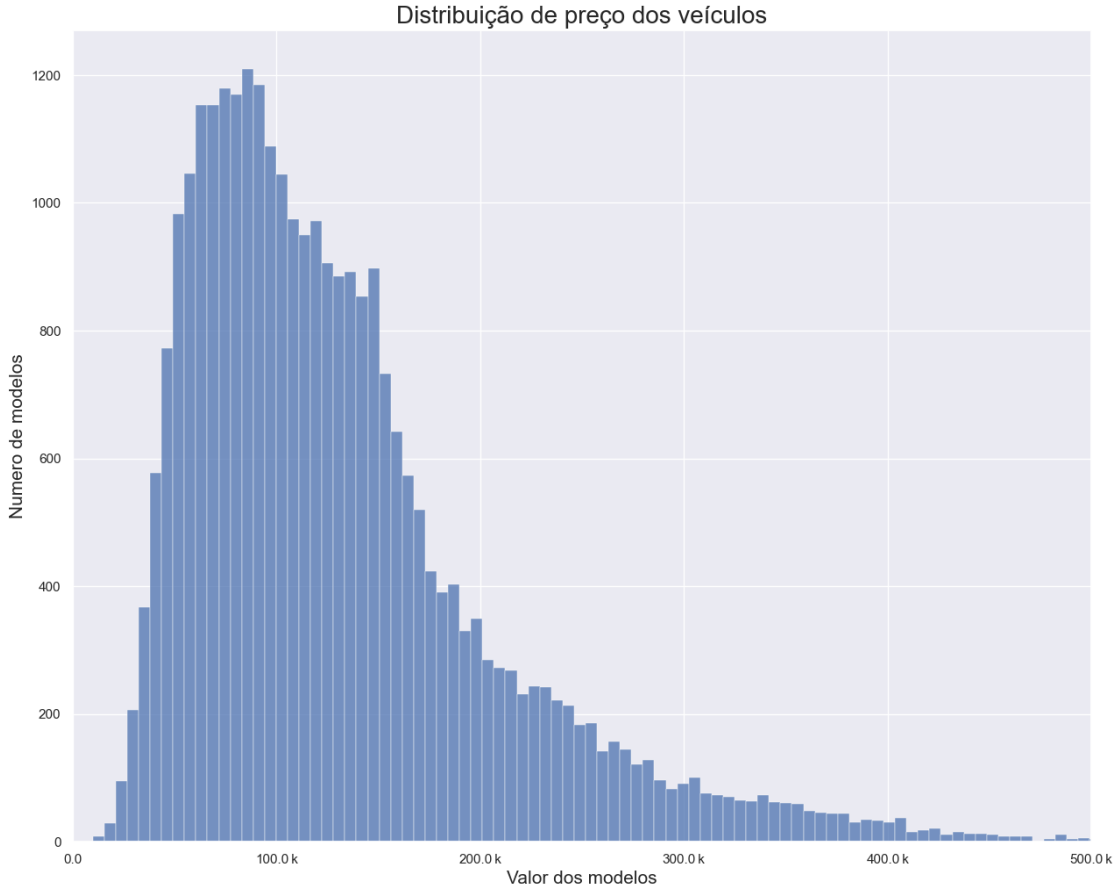
O link para o repositório completo é repositório.

Análise estatística dos Dados

A partir das variáveis dos dados fornecidos, vamos fazer análises gráficas e entender o que elas podem nos dizer. Para o notebook completo e melhor visualização dos gráficos acessar Análise Estatística.

Distribuição de preços

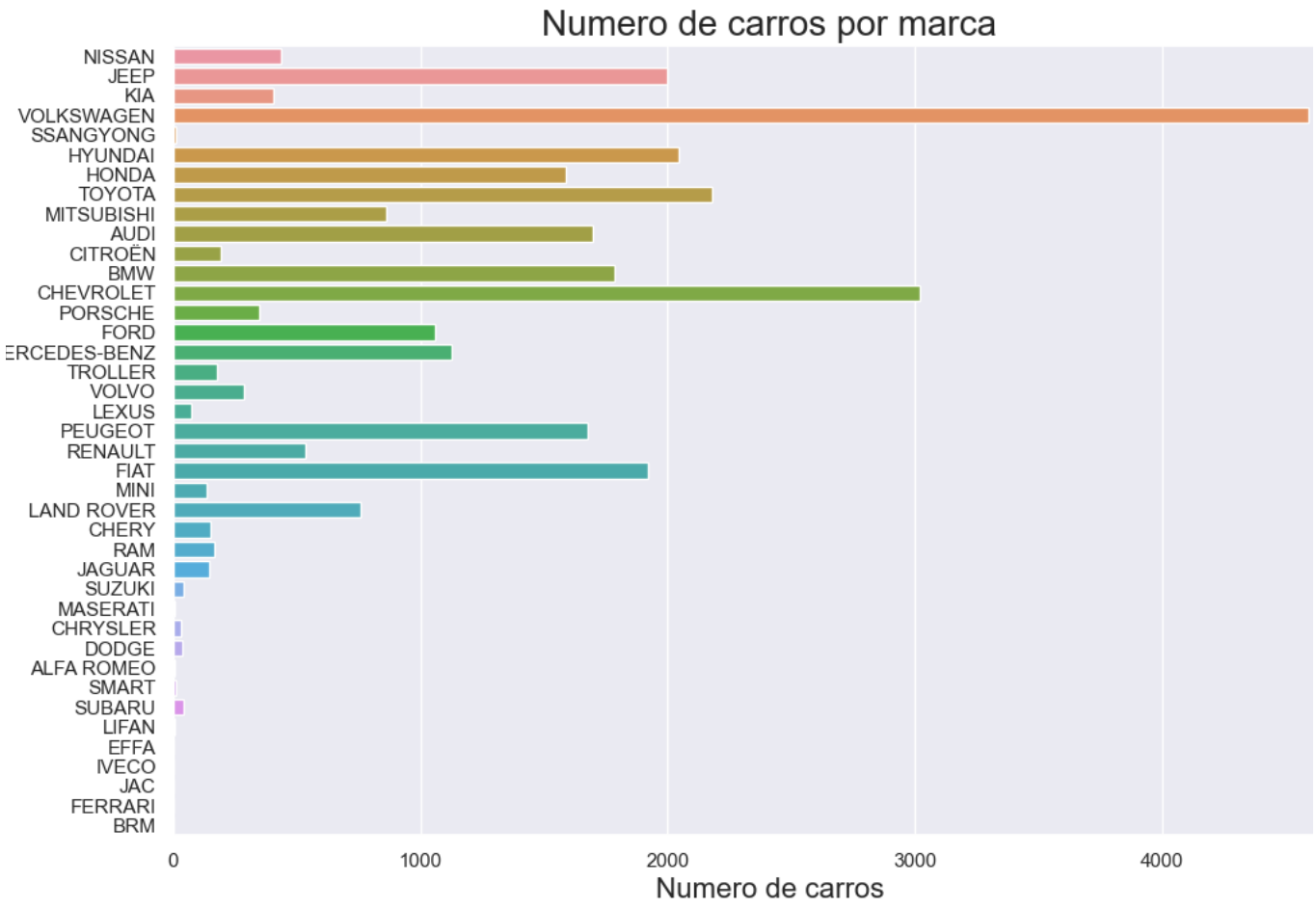
Primeiro vamos considerar a distribuição de preços dos veículos anunciados. Nele podemos perceber em qual valor de preço há mais modelos disponíveis. Percebemos que o pico de número de carros disponível está abaixo de 100.000R$, porém ainda com muitos carros nessa região e acima. Levando em consideração o momento dos preços de carros novos (Os 10 carros mais baratos do Brasil em 2023) e usados, quem está procurando opções mais baratas talvez procure um 0Km ao invés de um usado, enquanto estiver vendendo pode ter dificuldades

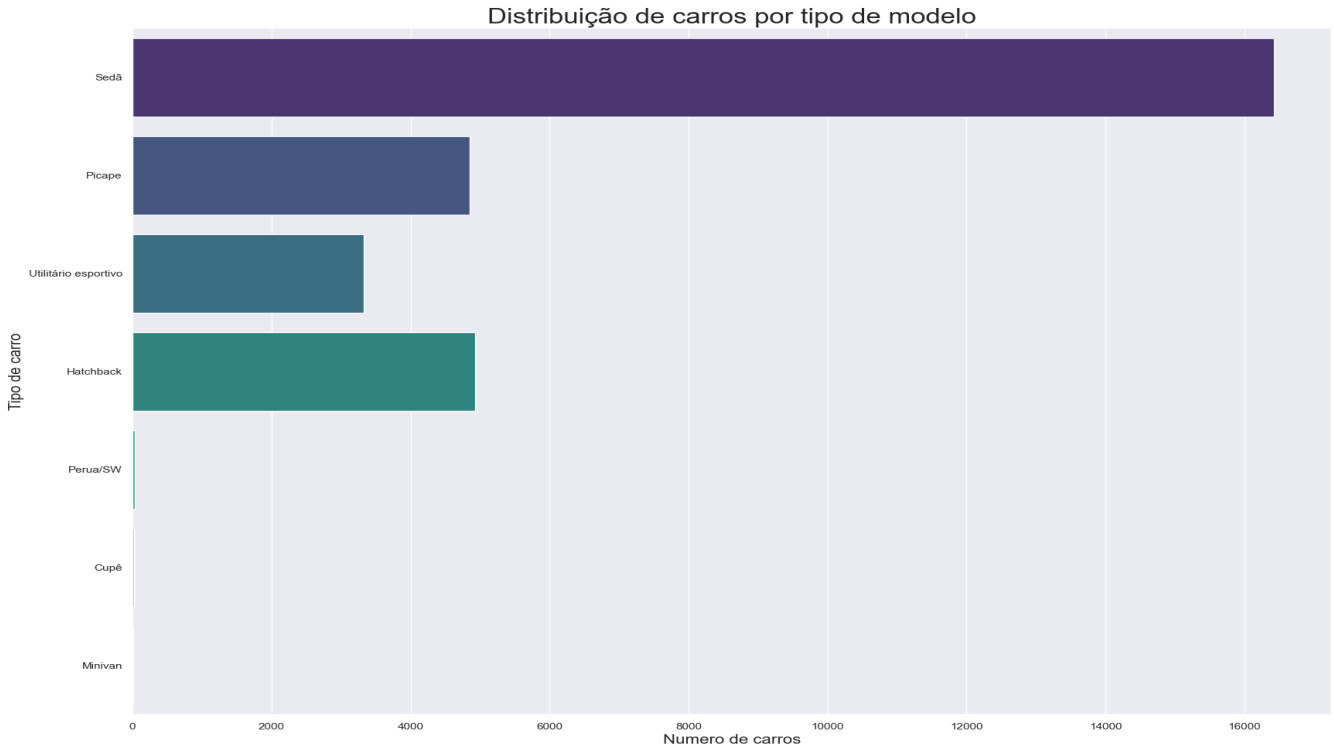


Marcas e Tipos

Podemos observar nos gráficos abaixo as marcas e modelos com mais abrangência no mercado. Observamos que marcas como Volkswagen e Chevrolet continuam entre as mais populares, enquanto marcas como Ferrari e Maserati, caracterizadas como marcas de luxo, tem um volume pequeno de veículos no mercado. Já quanto ao tipo dos modelos, observamos

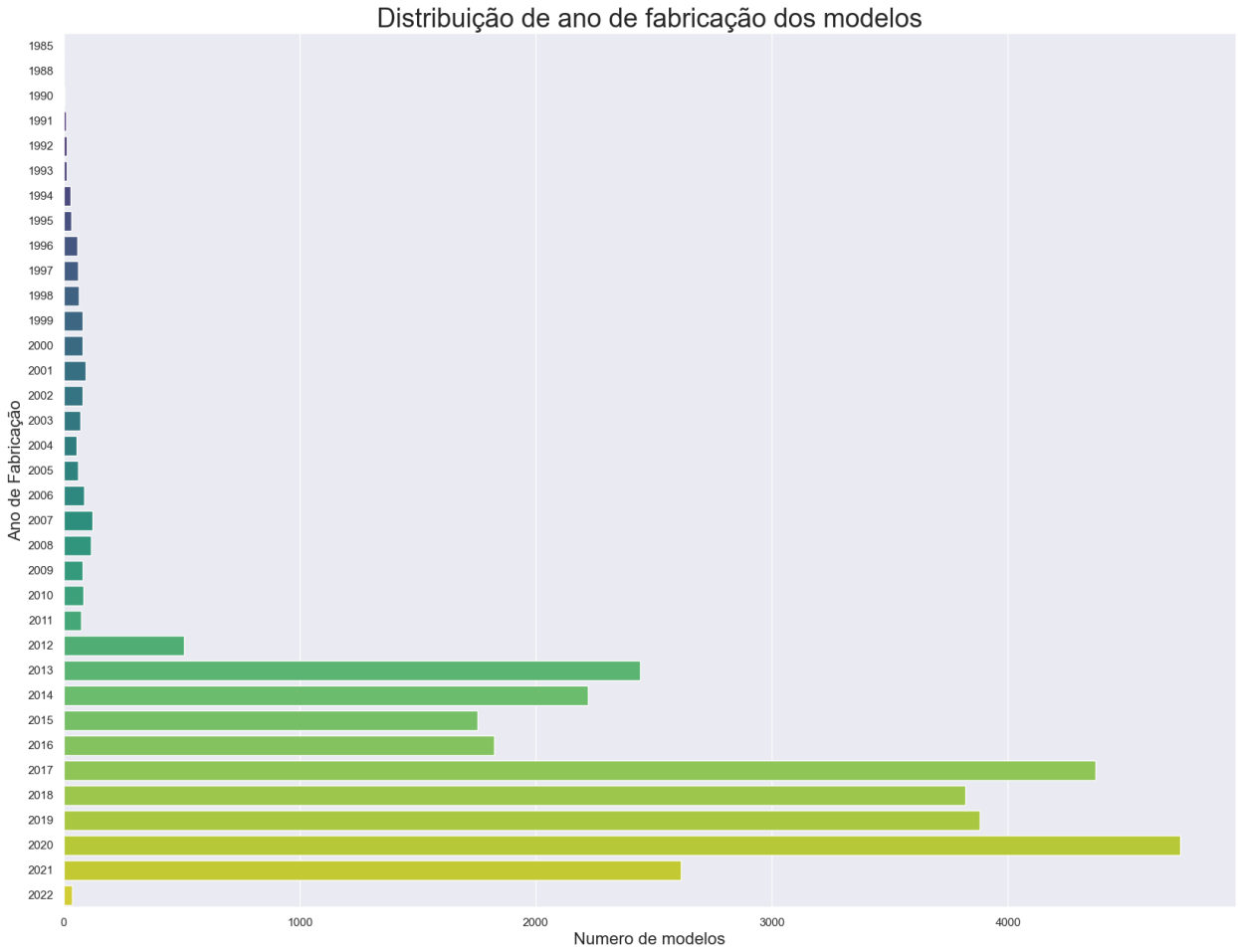
que os modelos Sedan são os mais mais populares, enquanto pickups e hatches estão próximos em volume no mercado.



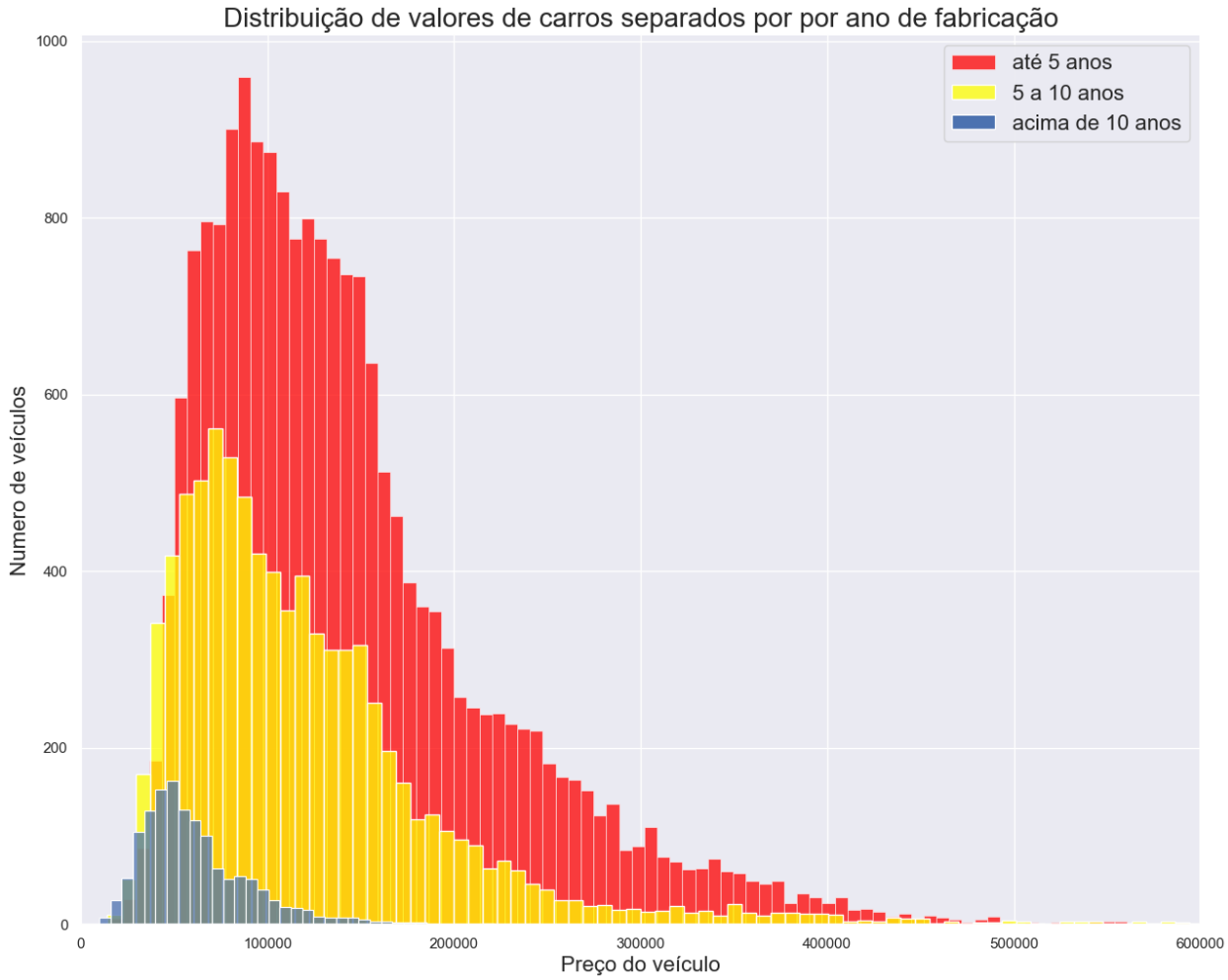


Ano de Fabricação

Observamos aqui que grande parte dos carros disponíveis para revenda ainda estão em uma intervalo de até dez anos de fabricação, ou seja ainda são modelos relativamente novos.

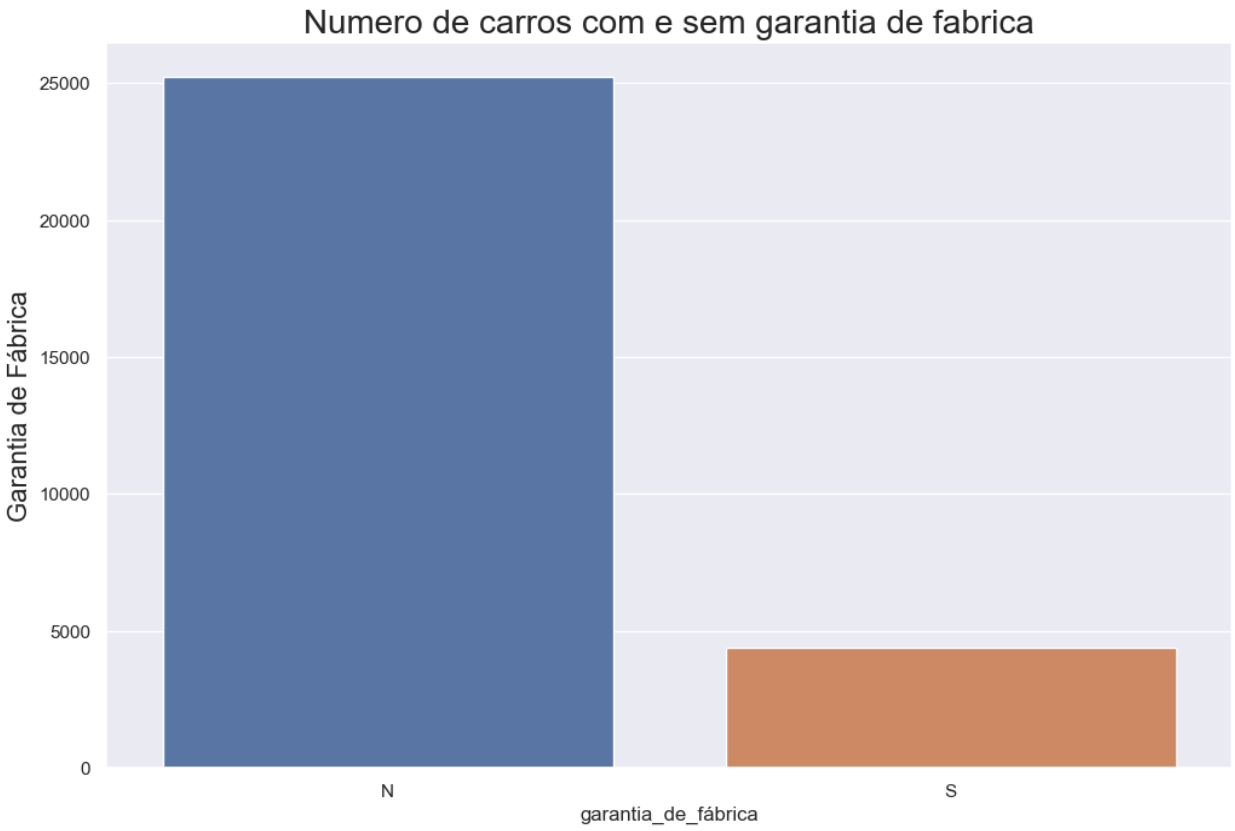


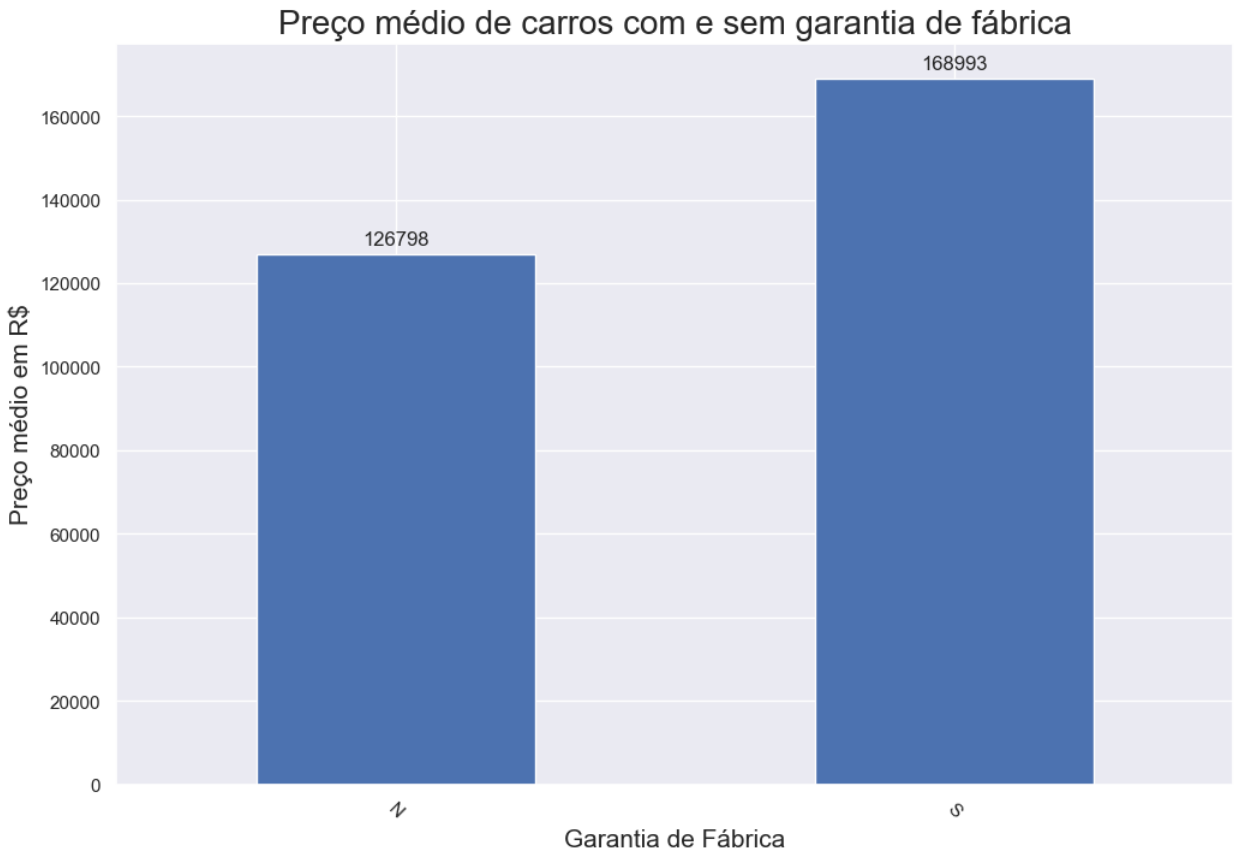
O gráfico de distribuição de valores por ano de fabricação, nos permite novamente observar que há uma predominância de carros mais novos, mas também uma tendência de queda de valor dos carros, observada pela posição dos picos. Eles estão separados aqui de 5 em 5 anos pois muitas marcas dão garantias estendidas de 5 e algumas até 10 anos nas peças dos carros.



Garantia de Fábrica

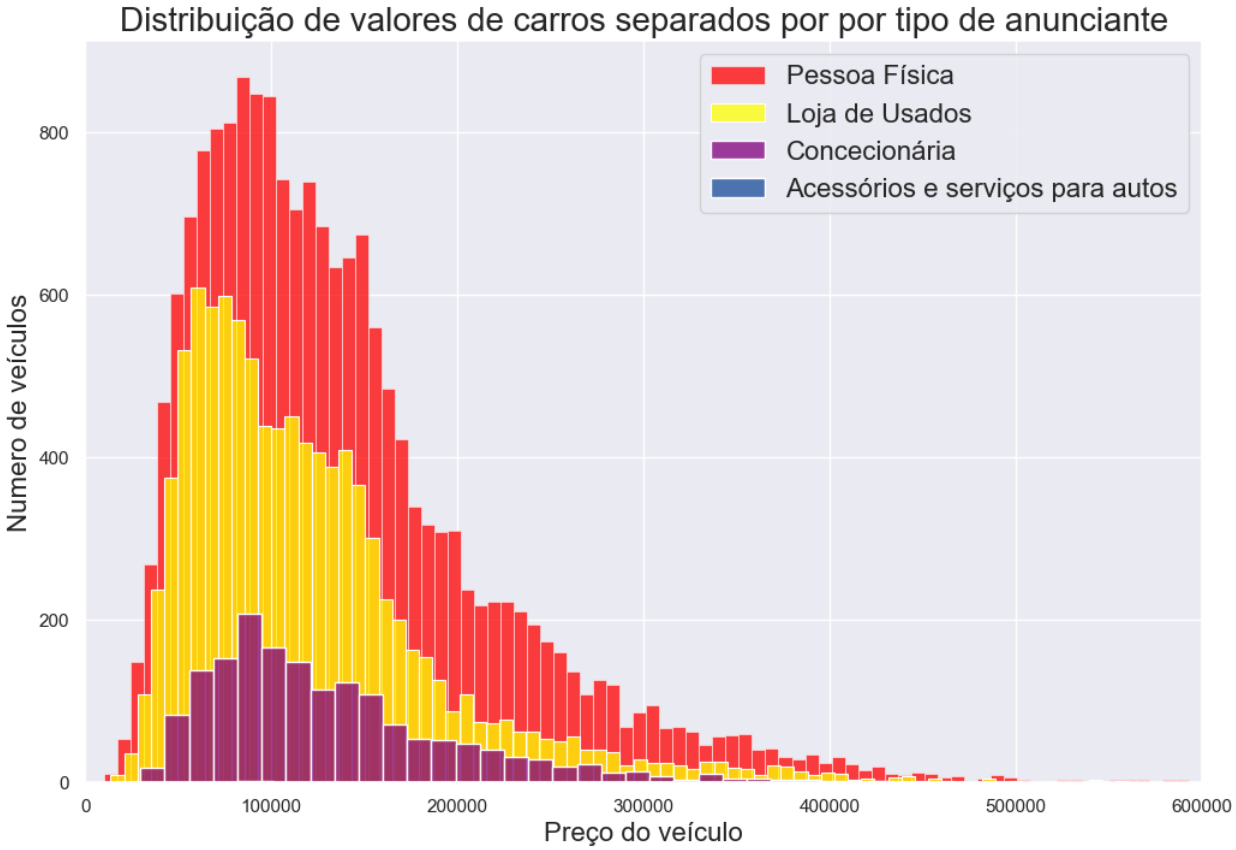
Apesar de a maior parte dos veículos ainda terem até 10 anos de fabricação, eles não têm em sua maioria garantia de fábrica. E isso possivelmente alavanca os preços daqueles que ainda possuem estas garantias.


Numero de carros com e sem garantia de fabrica
Garantia de Fábrica
25000
20000
15000
10000
5000
0
N
S
garantia_de_fábrica


Preço médio de carros com e sem garantia de fábrica
Preço médio em R$
160000
140000
120000
100000
80000
60000
40000
20000
0
126798
168993
N
S
Garantia de Fábrica

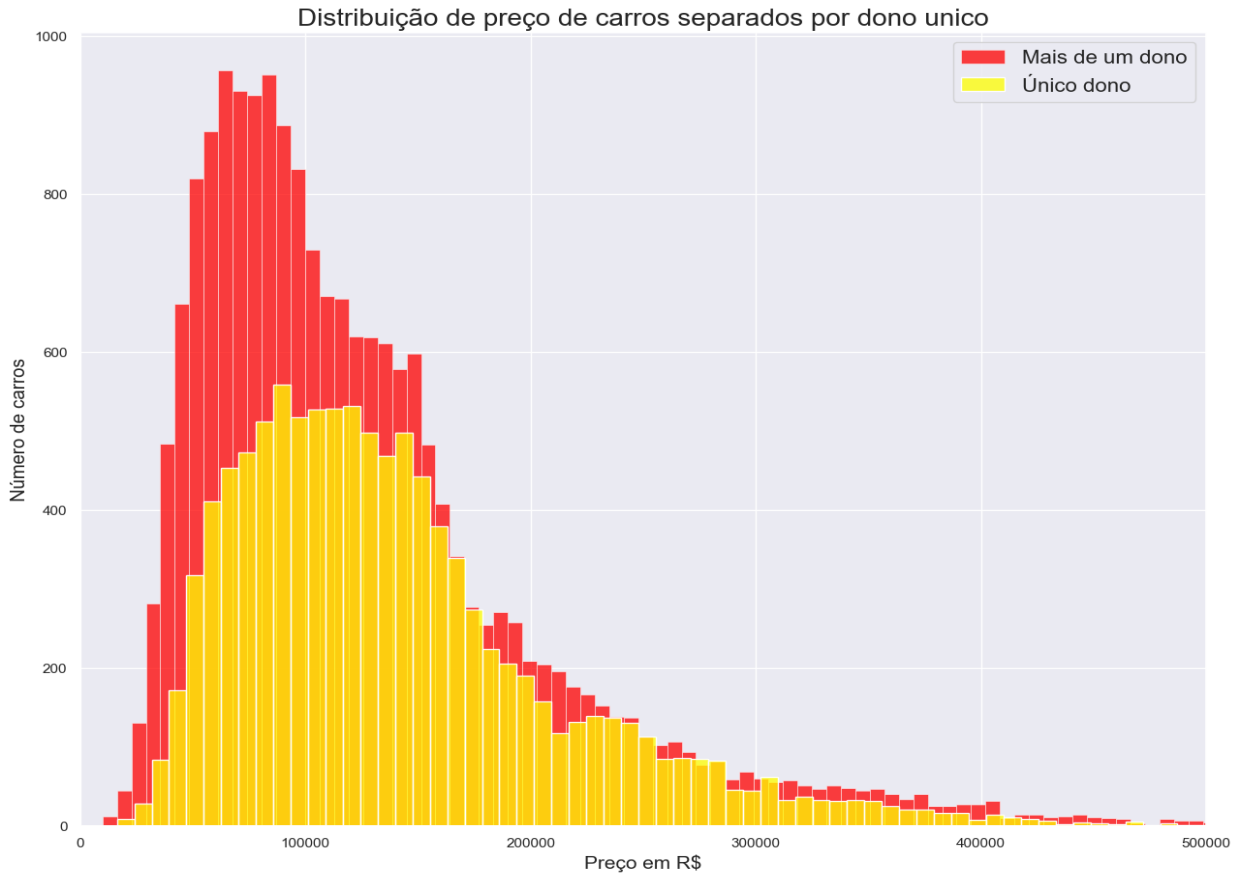
Anunciantes

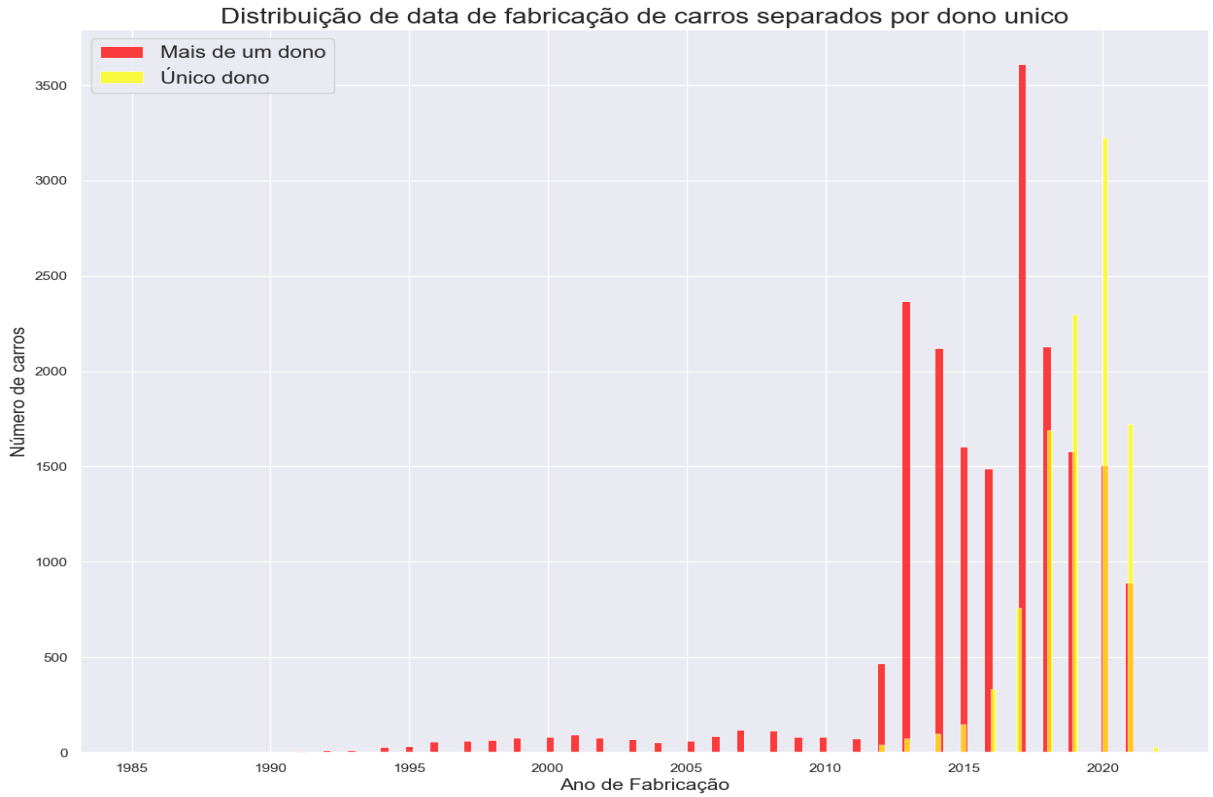
O Tipo do anunciante demonstra-se uma estatística que pouco altera o valor final do veículo vendido, apenas há uma pequena tendência dos carros das revendedoras estarem em uma região de preços abaixo dos 100 mil R$, isso indicaria que estas lojas têm tendência a obter carros que vendem mais rápido dando mais agilidades em seus fluxos de caixa.



Único Dono

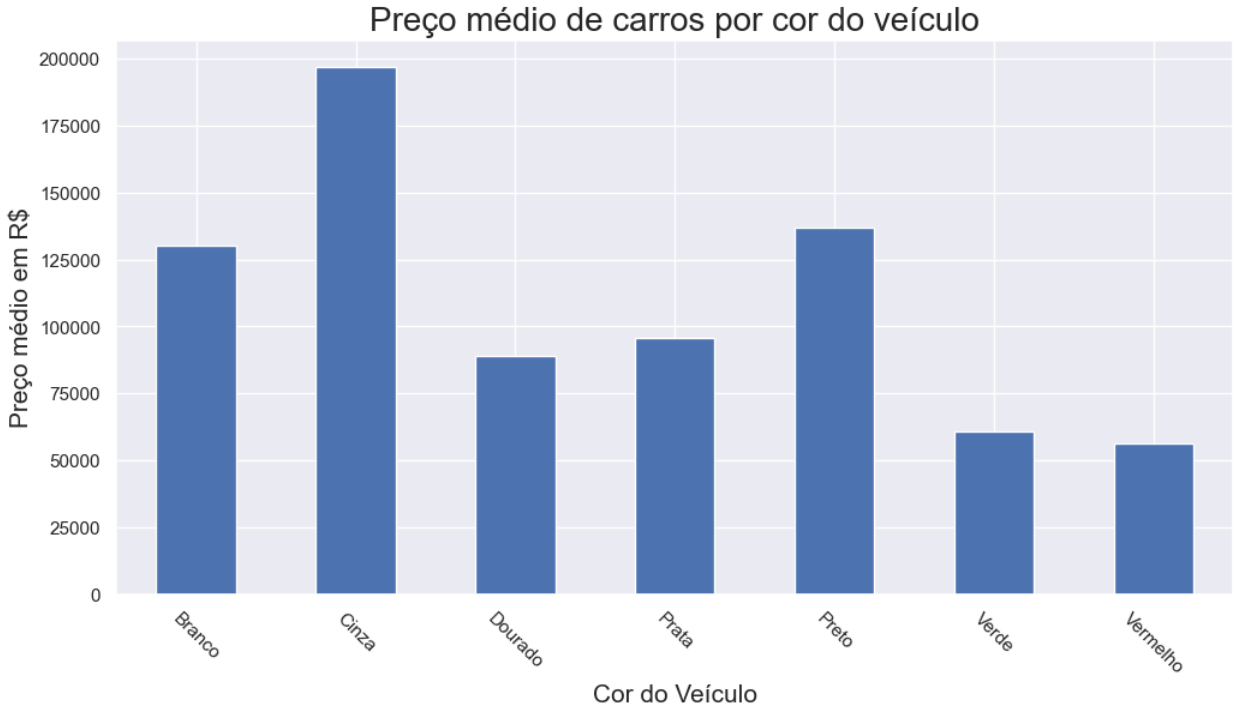
Observamos pelo gráfico abaixo que os carros com único dono tendem a ter um preço mais alto de mercado. Não só seu pico ocorre em valores mais altos, mas a curva tem uma tendência de queda mais suave do que a dos carros que já tiveram mais do que um dono. Isso está também ligado a idade do carro, pois carros mais velhos têm possibilidades maiores de terem tido mais de um dono.

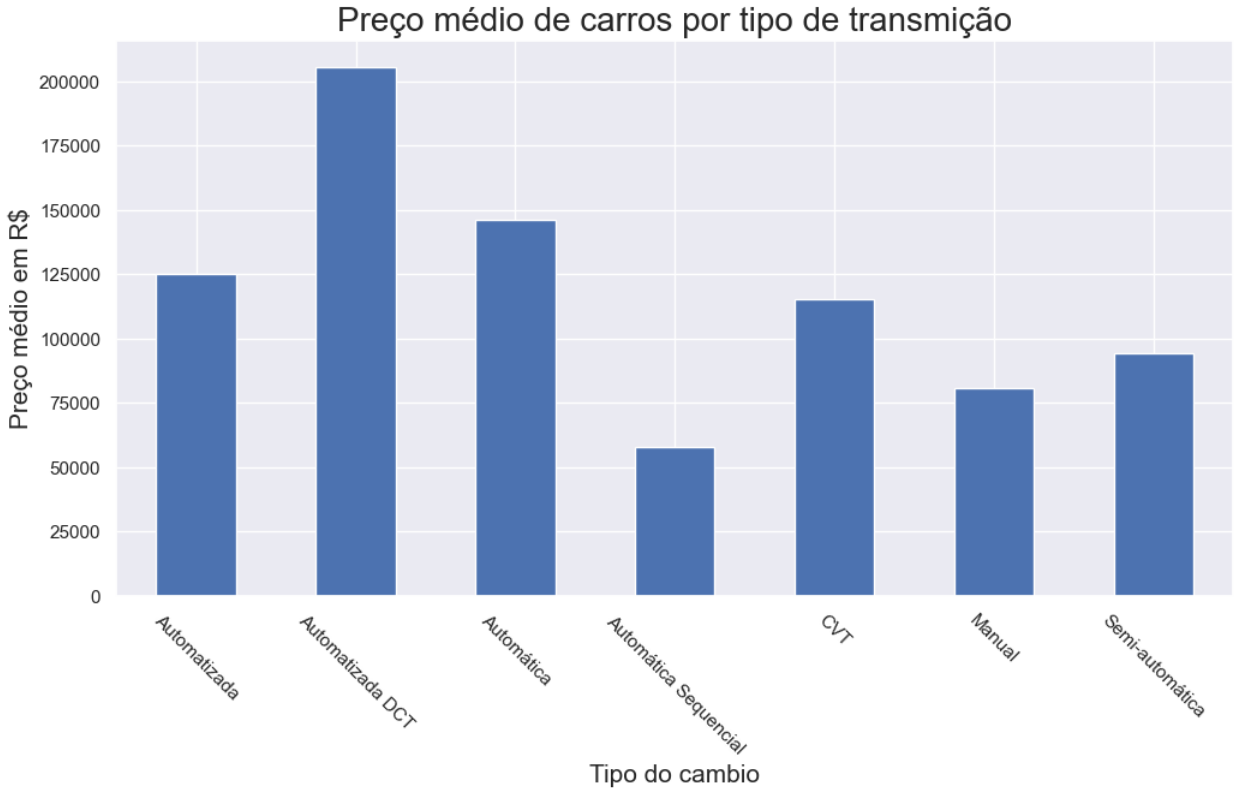

Distribuição de preço de carros separados por dono unico
Mais de um dono
Único dono
Número de carros
Preço em R$
0
200
400
600
800
1000
0
100000
200000
300000
400000
500000


Distribuição de data de fabricação de carros separados por dono unico
Mais de um dono
Único dono
Número de carros
Ano de Fabricação
0
500
1000
1500
2000
2500
3000
3500
1985
1990
1995
2000
2005
2010
2015
2020

Características

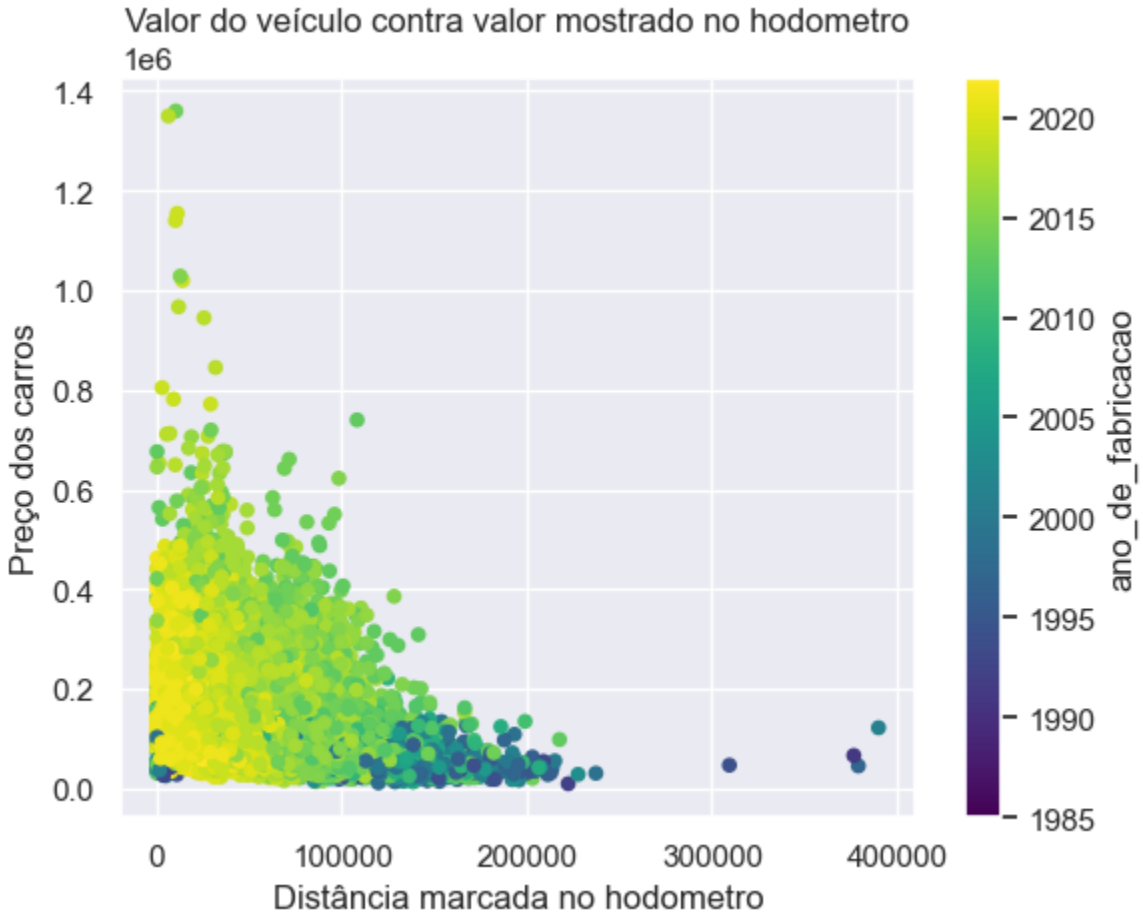
Algumas características dos veículos os tornam mais caros, em destaque as médias de preços de veículos pelo tipo de transmissão, e a cor do carro.





Distância Marcada no Hodômetro

Aqui vemos uma relação entre a distância marcada no hodômetro e os preços dos carros, carros que têm menor rodagem tendem a ter maior preço. Porém é necessário observar que a medida do hodômetro não significa carro novo, pois em algum momento este medidor volta a zero.



EDA

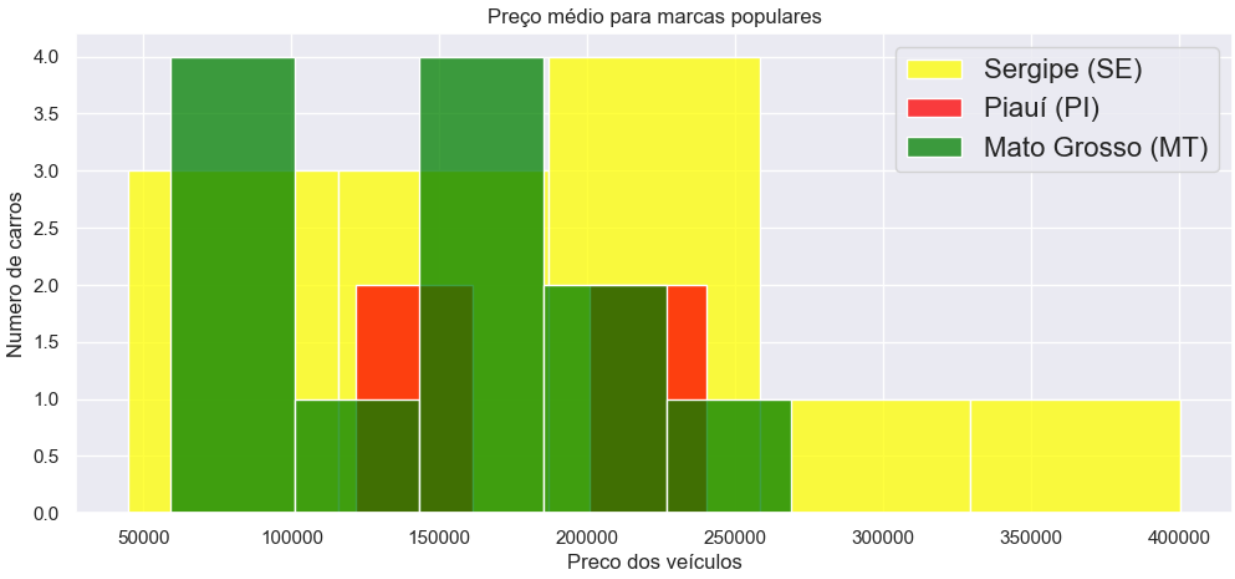
Nesta seção vamos responder a três perguntas através dos dados. O notebook para esta parte do desafio está em EDA

a. Qual o melhor estado cadastrado na base de dados para se vender um carro de marca popular e por quê?
b. Qual o melhor estado para se comprar uma picape com transmissão automática e por quê?
c. Qual o melhor estado para se comprar carros que ainda estejam dentro da garantia de fábrica e por quê?

Melhor estado para vender uma marca Popular

Segundo a Focus2Move, que mantém dados de vendas de carros ao redor do mundo, as 5 marcas mais vendidas em Junho de 2023 foram Fiat, Chevrolet, Volkswagen, Toyota e Hyundai (Marcas mais vendidas no Brasil). Portanto, para definir qual melhor estado para se vender uma marca popular filtramos os dados por essas 5 marcas.

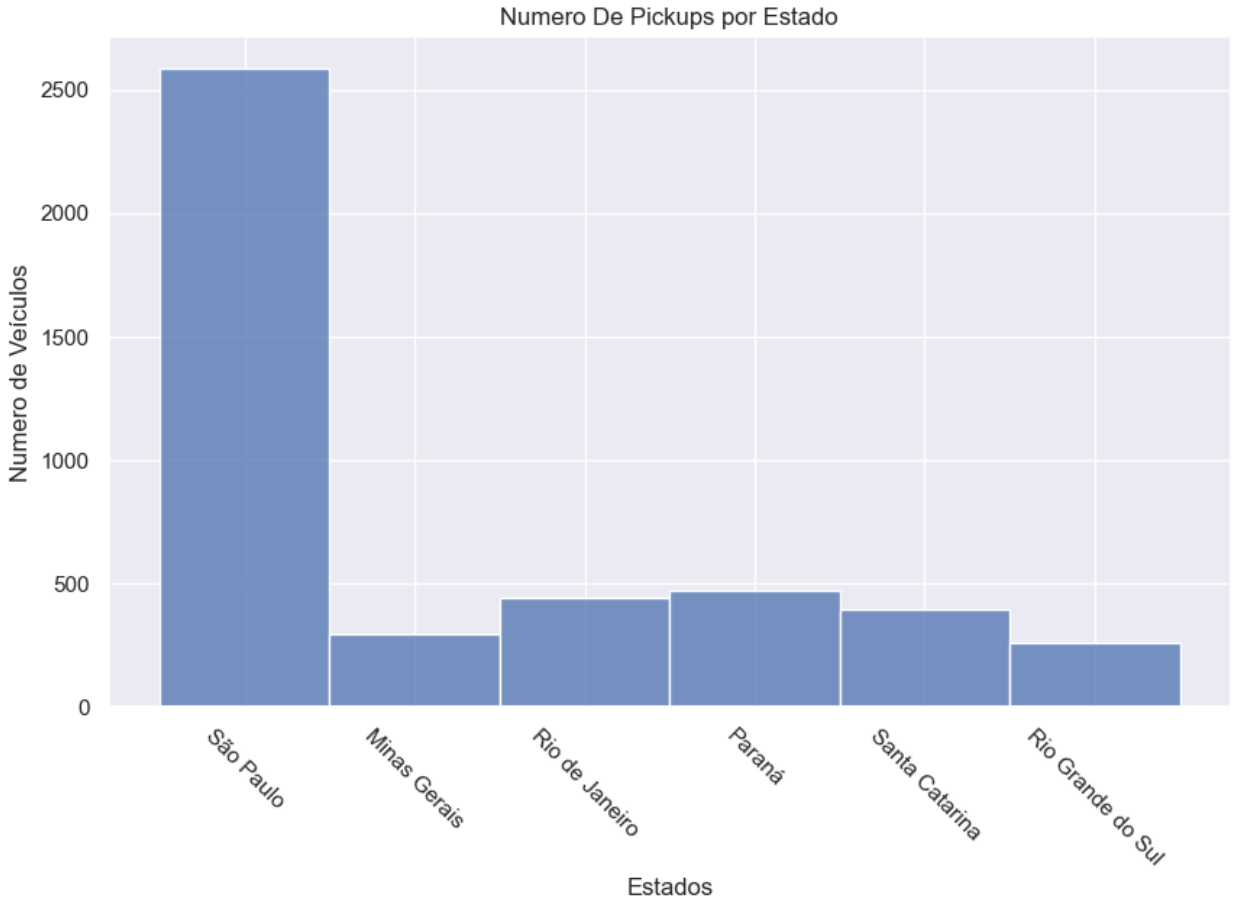
Para tomar esta decisão foram tirados as médias, desvios padrões, mínimos, máximos, dos preços dos carros das marcas populares. Abaixo está o gráfico com as distribuições dos estados com as maiores médias.



Como podemos observar o estado do Sergipe é o que tem a maior média com um valor de 188331R$, apesar do Piauí estar próximo com a média de 181713R$, temos poucos dados sobre estas marcas neste estado, as medidas para este estado podem estar enviesadas. Apesar de a média consideravelmente menor do estado do Mato Grosso 140977R$, este tem um desvio padrão menor (63789R$ no MT, contra 108336R$ no SE) e um valor mínimo maior (59041R$ no MT, contra 44862RS no SE), portanto este seria o estado de melhor possibilidade de vendas, já que tem uma média de preços alta, visando uma venda a um alto preço e uma variedade menor de preços no mercado o que ajuda a acelerar a venda.

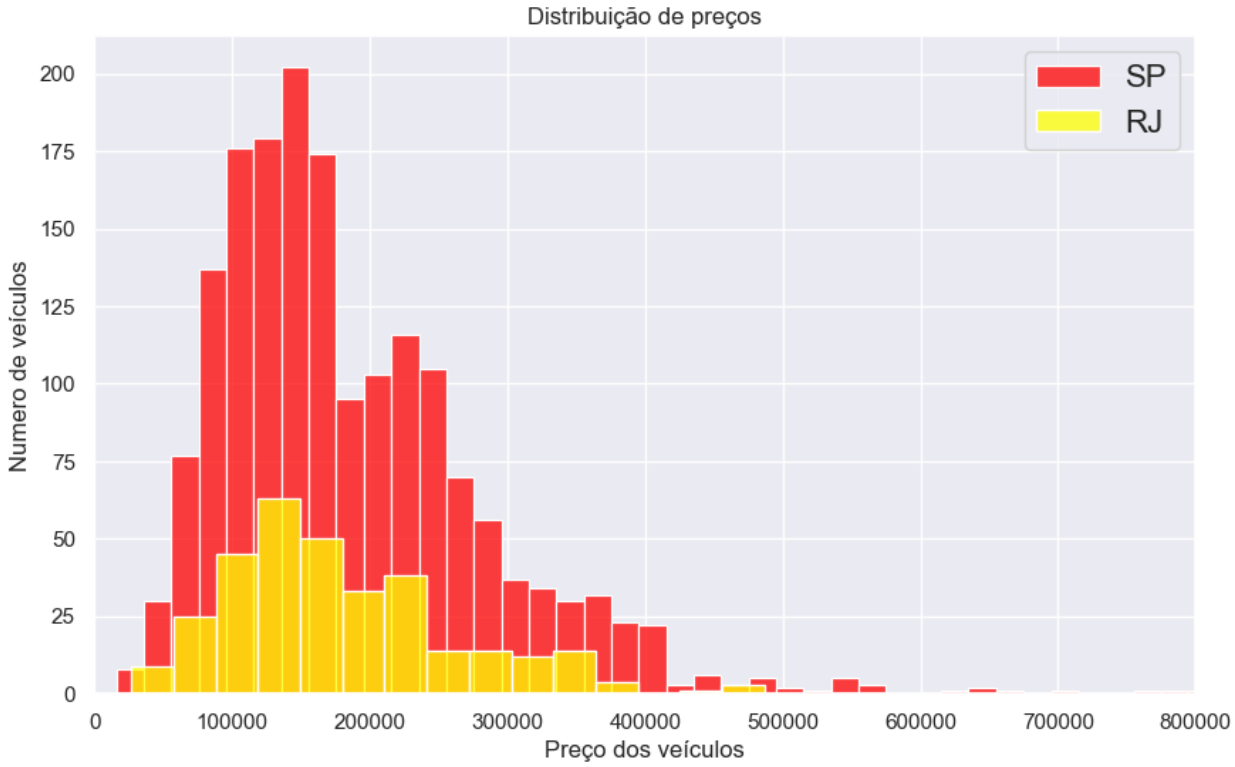
Melhor estado para comprar uma Pickup com Transmissão Automática

Para responder esta pergunta primeiro verificamos quais estados têm o maior número de pickups disponíveis para venda, pois isso garante mais variabilidade de preços para veículos com características parecidas.



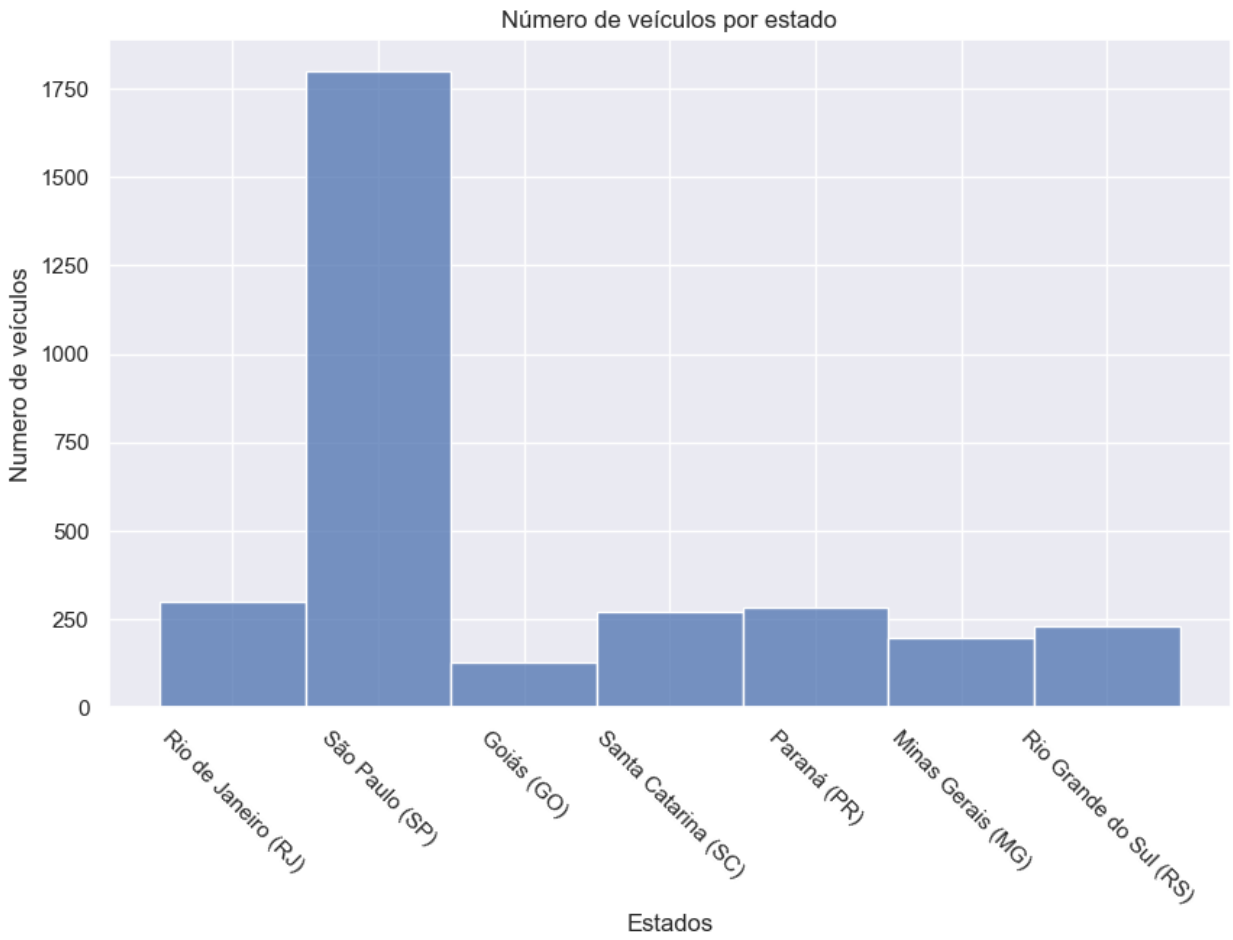
O estado do Rio de Janeiro tem a menor média de preço e variância, além do segundo menor preço mínimo, seria então o estado do RJ o mais propício a se comprar uma pickup a um baixo preço. Porém caso se esteja procurando mais variedade de modelos o estado de São Paulo terá disponível sem um aumento considerável na média dos preços.

	estado_vendedor	mean	std	min	max
3	Rio de Janeiro	154264	87253	16265	486870
5	São Paulo	156332	105206	13643	1154360
4	Santa Catarina	159073	111806	22727	1359813
0	Minas Gerais	163568	95840	17721	653173
1	Paraná	171580	94020	20759	550147
2	Rio Grande do Sul	171750	95629	27691	540420



Melhor estado para comprar veículos dentro da garantia

Assim como na análise anterior, vamos buscar o estado com mais variabilidade e combinar com menores preços. O estado de Minas é o local ideal para se buscar menor preços e o de São Paulo para buscar variedade sem comprometer preços.



	estado_vendedor	mean	std	min	max
1	Minas Gerais (MG)	158502	74254	29907	426790
6	São Paulo (SP)	168004	84211	31763	677129
0	Goiás (GO)	168883	87035	45812	486648
2	Paraná (PR)	171324	78341	29328	411923
5	Santa Catarina (SC)	174984	96625	44005	672933
3	Rio Grande do Sul (RS)	178862	90161	54743	589419
4	Rio de Janeiro (RJ)	180841	84383	39556	486870

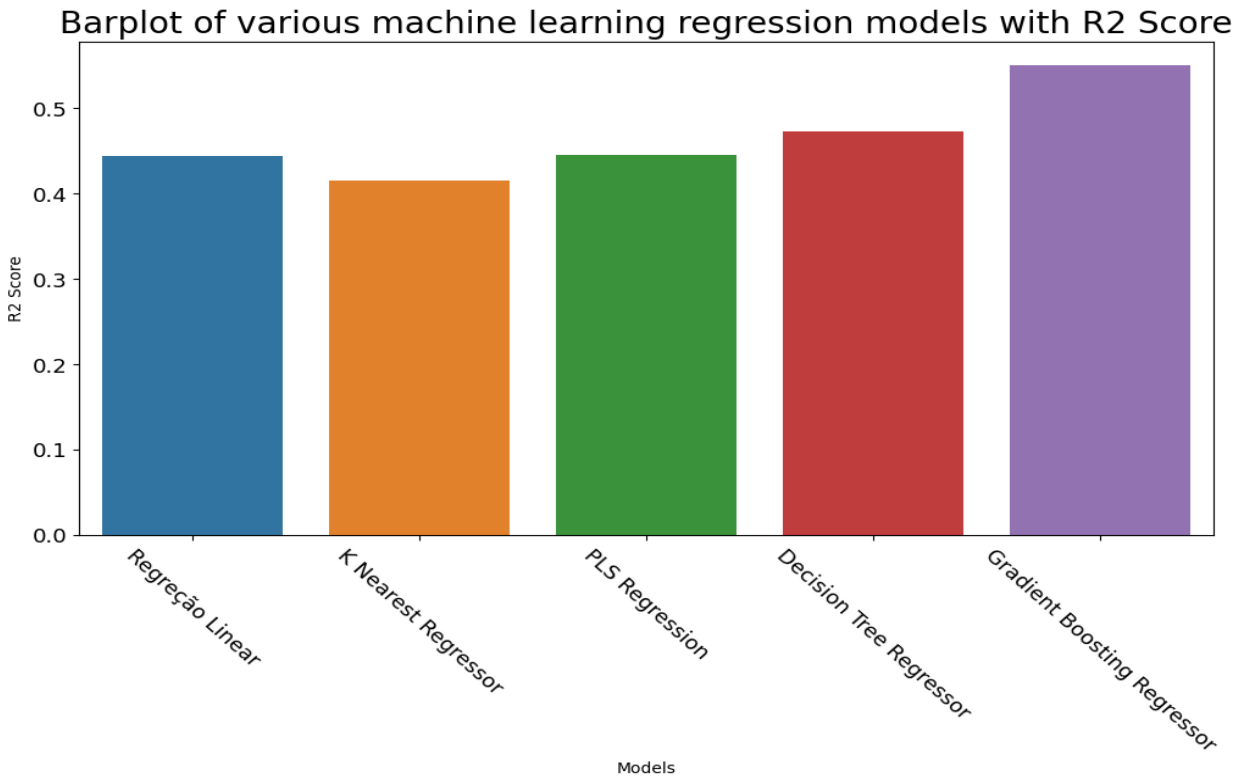
Modelagem e Machine Learning

A modelagem dos dados para previsão dos preços é feita através de uma regressão, apesar de muitos dos dados serem classificatórios usamos encoders, que levam em consideração a distribuição estatística das características, para transformá-los em dados numéricos e a partir deles fazer a regressão.

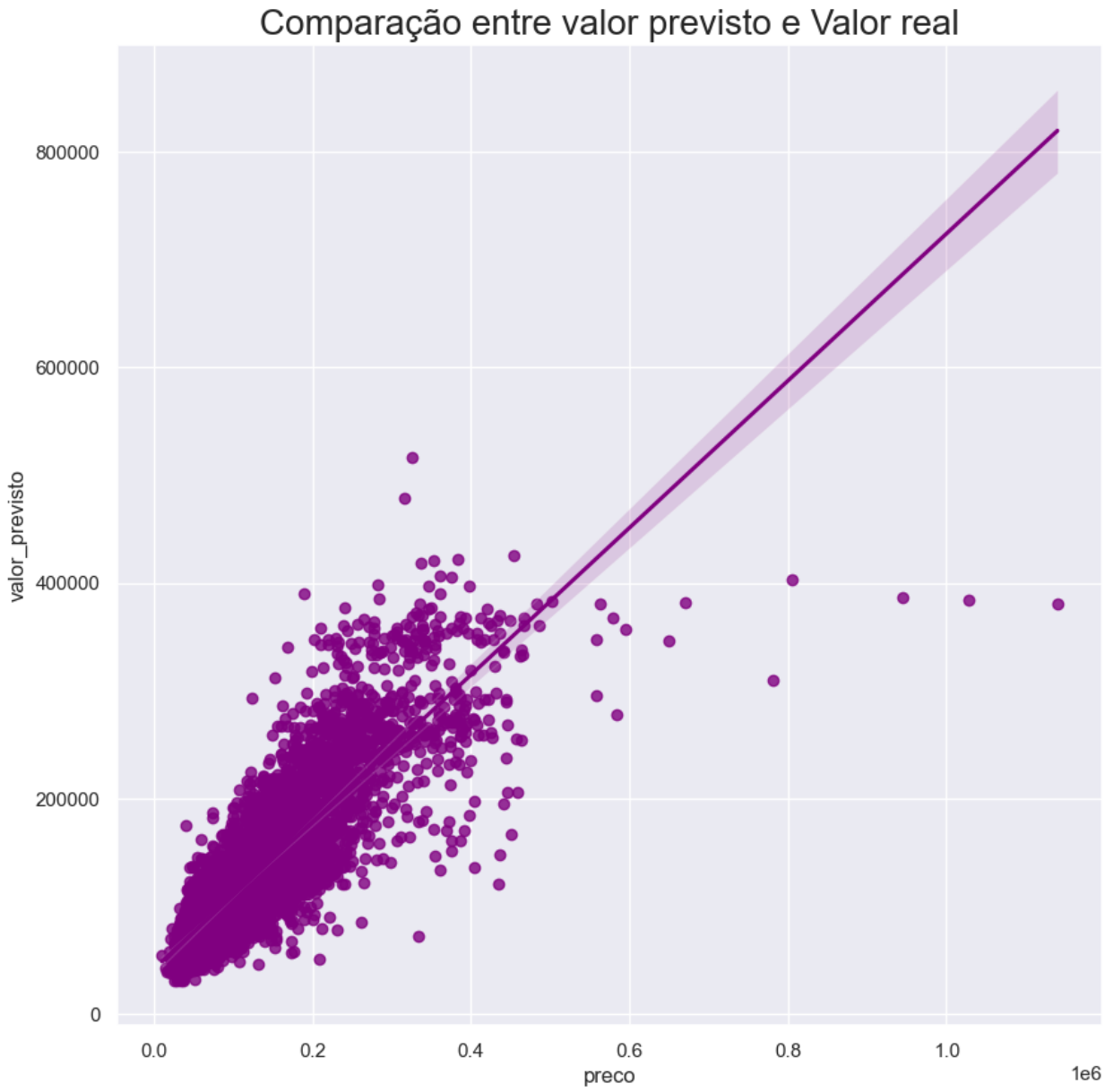
Para escolher o modelo foi levado em consideração o erro médio absoluto, o erro médio quadrado, e o R2 Score. Em todos o Gradient Boost Regressor teve a melhor performance.

Modelos de gradient boosting tendem a ser mais acurados pois ele corrige a si mesmo durante o processo, e é capaz de capturar padrões mais complexos. Porém eles podem gerar overfitting especialmente em bases de dados muito ruidosas.

Todos os procedimentos para obtenção do modelo de preços estão no notebook Modelo, e os resultados estão em predi



cted.csv.


Comparação entre valor previsto e Valor real
valor_previsto
800000
600000
400000
200000
0
0.0
0.2
0.4
0.6
0.8
1.0
preco
1e6