Análise

Incialmente para a análise carregamos os dados com o arquivo analise.ipynb. As informações sobre variáveis nulas e tipos de dados foram passadas para uma tabela em excel para uma análise inicial.

```
In [6]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29584 entries, 0 to 29583
Data columns (total 29 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       29584 non-null  object
 1   num_fotos                29407 non-null  float64
 2   marca                    29584 non-null  object
 3   modelo                   29584 non-null  object
 4   versao                   29584 non-null  object
 5   ano_de_fabricacao        29584 non-null  int64
 6   ano_modelo               29584 non-null  float64
 7   hodometro                29584 non-null  float64
 8   cambio                   29584 non-null  object
 9   num_portas               29584 non-null  int64
 10  tipo                     29584 non-null  object
 11  blindado                 29584 non-null  object
 12  cor                      29584 non-null  object
 13  tipo_vendedor            29584 non-null  object
 14  cidade_vendedor          29584 non-null  object
 15  estado_vendedor          29584 non-null  object
 16  anunciante               29584 non-null  object
 17  entrega_delivery         29584 non-null  bool
 18  troca                    29584 non-null  bool
 19  elegivel_revisao         29584 non-null  bool
 20  dono_aceita_troca        21922 non-null  object
 21  veiculo_único_dono       10423 non-null  object
 22  revisoes_concessionaria  9172 non-null   object
 23  ipva_pago                19659 non-null  object
 24  veiculo_licenciado       15906 non-null  object
 25  garantia_de_fábrica      4365 non-null   object
 26  revisoes_dentro_agenda   5910 non-null   object
 27  veiculo_alienado         0 non-null      float64
 28  preco                    29584 non-null  float64
dtypes: bool(3), float64(5), int64(2), object(19)
memory usage: 6.0+ MB
```

```
In [4]: df.isnull().sum()

Out[4]: id                          0
        num_fotos                 177
        marca                       0
        modelo                      0
        versao                      0
        ano_de_fabricacao           0
        ano_modelo                  0
        hodometro                   0
        cambio                      0
        num_portas                  0
        tipo                        0
        blindado                    0
        cor                         0
        tipo_vendedor               0
        cidade_vendedor             0
        estado_vendedor             0
        anunciante                  0
        entrega_delivery            0
        troca                       0
        elegivel_revisao            0
        dono_aceita_troca        7662
        veiculo_único_dono      19161
        revisoes_concessionaria 20412
        ipva_pago                9925
        veiculo_licenciado      13678
        garantia_de_fábrica     25219
        revisoes_dentro_agenda  23674
        veiculo_alienado        29584
        preco                       0
        dtype: int64
```

Na tabela excel abaixo a coluna tipo contém o tipo apresentado no júpter notebook e entre parênteses a suspeita inicial de que tipo de dado representaria melhor a informação. Por exemplo a variável “blindado” em vez de se tratar como uma string “S” ou “N” podemos transformá-la para uma variável booleana. “ano_modelo” da mesma forma pode ser pensada como um inteiro em vez de um float. A coluna expectativa representa o quanto esperamos incialmente que a variável influencie no valor final do carro, nessa análise inicial podemos remover a variável “veiculo_alienado” por apresentar apenas valores nulos. Notamos que a variável “versão” contém a informação de tipo de combustível que pode ser útil se extraída para uma coluna própria. Também notamos que as “tipo_vendedor” e “anunciante” possivelmente contém a mesma informação duplicada de se tratar de pessoa física ou jurídica. Todas essas suspeitas iniciais serão avaliadas em seguida.

Variável	Tipo	Nulos	Expectativa	Observação
id	object(big number)	0	Zero	
num_fotos	float(int)	177	Baixa	
marca	object(string)	0	Media	
modelo	object(string)	0	Alta	
versao	object(string)	0	Baixa	Extrair tipo de combustível
ano_de_fabricacao	int	0	Alta	
ano_modelo	float(int)	0	Alta	
hodometro	float(int)	0	Alta	
cambio	object(string)	0	Alta	
num_portas	int	0	Media	
tipo	object(string)	0	Meda	
blindado	object(boolean)	0	Alta	
cor	object(string)	0	Baixa	
tipo_vendedor	object(string)	0	Baixa	Equivalente a anunciante
cidade_vendedor	object(string)	0	Media	
estado_vendedor	object(string)	0	Media	
anunciante	object(string)	0	Baixa	Equivalente a tipo_vendedor
entrega_delivery	boolean	0	Baixa	
troca	boolean	0	Media	
elegivel_revisao	boolean	0	Media	
dono_aceita_troca	object(boolean)	7662	Baixa	
veiculo_único_dono	object(boolean)	19161	Media	
revisoes_concessionaria	object(boolean)	20412	Media	
ipva_pago	object(boolean)	9925	Media	
veiculo_licenciado	object(boolean)	13678	Baixa	
garantia_de_fábrica	object(boolean)	25219	Media	
revisoes_dentro_agenda	object(boolean)	23674	Media	
veiculo_alienado	float(int)	29584	Zero	Zero devido ao fato de ter apenas valores nulos
preco	float(int)	0	Alvo	

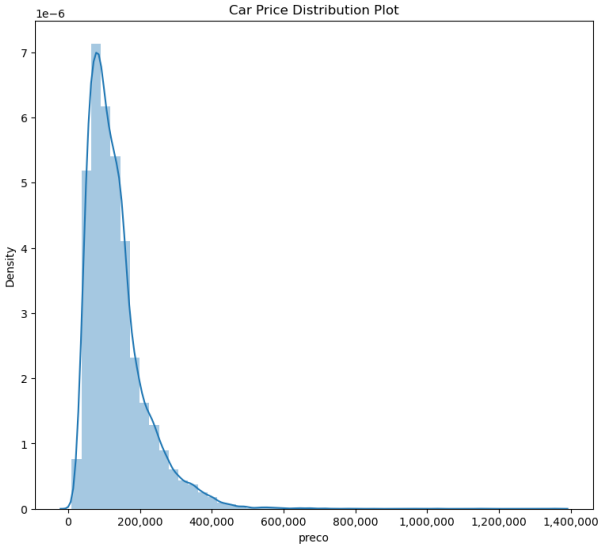
Em seguida plotarmos um gráfico de barras simples das váriáveis suspeitas de serem boleanas e pudemos confirmar de fato que as mesmas só apresentam dois valores possíveis. Em alguns casos o segundo valor se trata de null ou NaN que podemos considerar falso. Vale notar que a variável “elegivel_revisao” apresenta apenas um valor em todas as colunas, logo assim como “veiculo_alienado” ela não será útil na nossa análise.

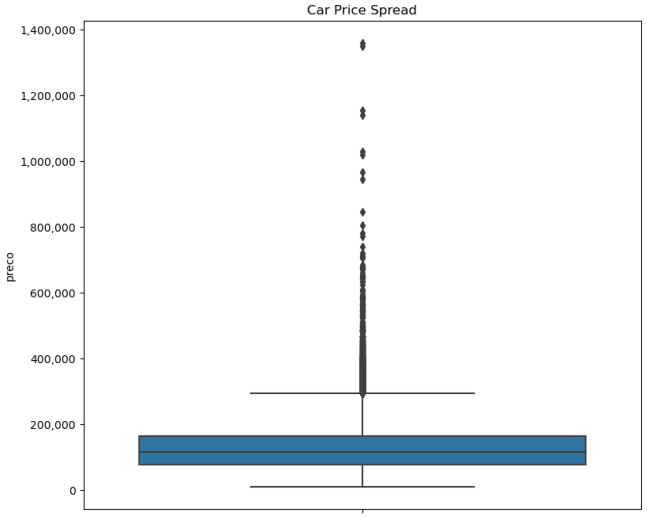
Ao compararmos as colunas “tipo_vendedor” e “anunciante” confirmamos nossa suspeita de serem equivalentes. Nelas ‘’PF” corresponde sempre a pessoa física e “PJ” a pessoa jurídica ou Concessionária em mais de 99% dos casos de onde podemos inferir que apenas a coluna “tipo_vendedor” pode ser usada daqui em diante.

No final após alterarmos as colunas necessárias temos um novo dataframe sem as colunas “versao”, “anunciante”, “elegivel_revisao”, “veiculo_alienado” bem como uma nova coluna “combustível” para prosseguir com o EDA.

EDA

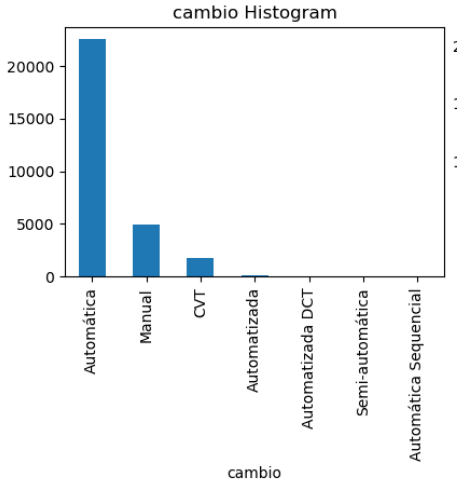
Plotando o gráfico de distribuição dos preços notamos que os mesmos estão concentrados entre 0 400 mil reais com outliers na casa dos milhões o que dependendo do que buscamos analisar pode ser um problema que exija um tratamento dos dados antes de uma resposta definitiva sobre determinadas perguntas.



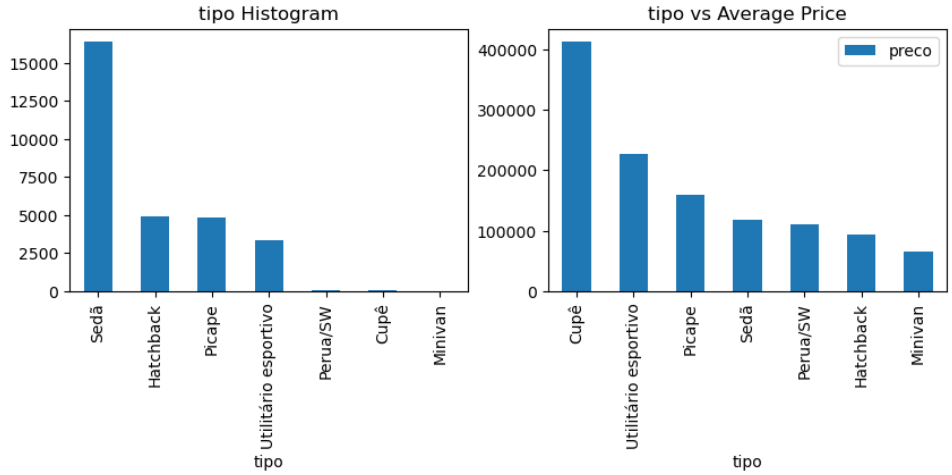


Plotamos em seguida o histograma e a média de preços em relação as variáveis categóricas e booleanas para podermos inferir algum tipo de informação. A seguir seguem algumas inferências. Volkswagem, Chevrolet e Toyota foram as marcas mais negociadas no período enquanto as marcas com a maior média de preço se trataram das marcas de luxo como já era o esperado.

Ao olharmos o histograma de tipo de câmbio notamos que o câmbio automático passou a dominar a venda de carros com o antigo câmbio manual perdendo bastante espaço.

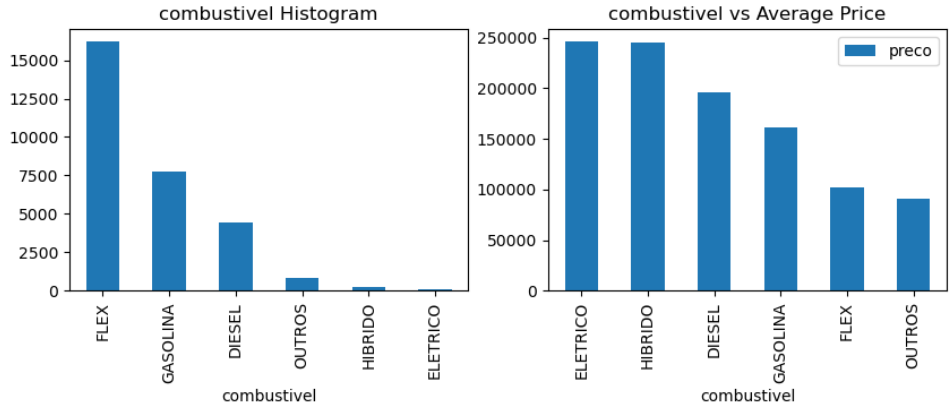


Sedans e hatchs são a maioria do mercado com o tipo cupê apresentando a maior média de preços

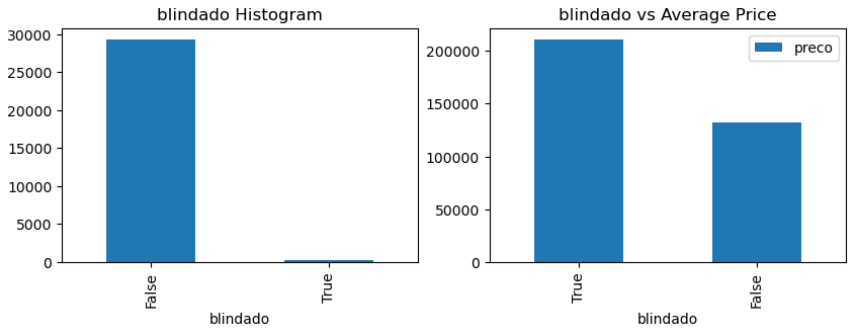


O estado de São Paulo como esperado apresentou o maior volume de vendas do país enquanto Sergipe apresentou a maior média nos preços dos carros. A maioria dos anúncios foram feitos por pessoas físicas sendo a diferença da média dos preços entre pessoa física e jurídica bem pequena.

Os carros flex e a gasolina dominaram o número de vendas enquanto os elétricos e híbridos tiveram a maior média de preços.



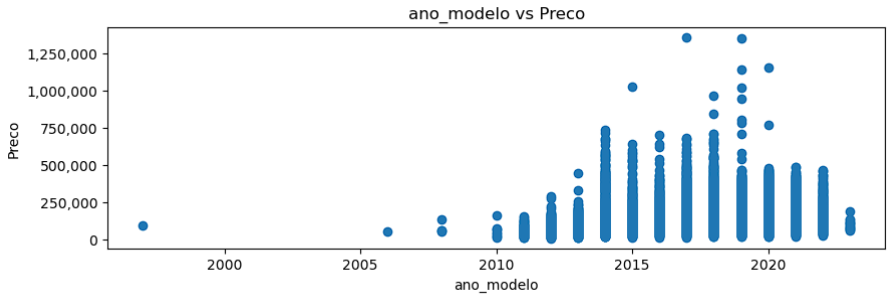
Carros blindados apesar de formarem uma quantidade pequena do mercado possuem um valor de venda muito alto se comparado aos carros sem blindagem.



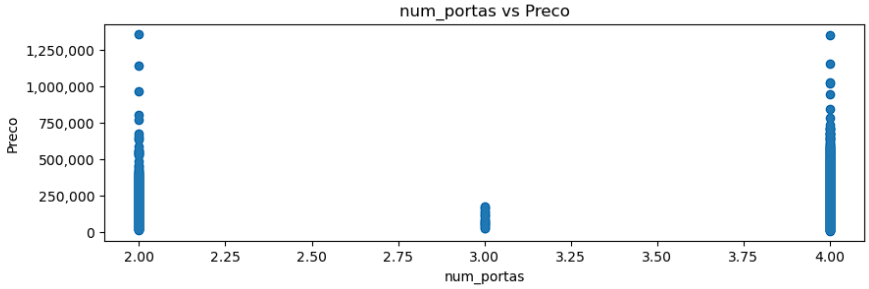
Entre as outras variáveis booleanas as que apresentaram algum impacto significativo na média foram a garantia de fábrica ainda estar ativa e o fato das revisões de fábrica serem feitas apenas na

concessionária. Podemos imaginar que os dois fatos estão relacionados já que normalmente carros na garantia fazem a revisão na concessionária sempre e estão mais novos. O que contribui pra uma menor desvalorização do preço.

Já nos gráficos das variáveis numéricas em relação ao preço notamos o esperado que carros com ano de fabricação mais recente apresentam valores maiores.



Um fato curioso é que carros com três portas apresentaram uma tendência de valores muito mais baixos que os de duas e quatro portas. Enquanto é esperado que os de quatro portas sejam mais caros podemos inferir que alguns carros de luxo de duas portas puxaram o valor dos mesmos para o alto.



Perguntas de negócio

Quanto as perguntas de negócio seguem as respostas.

Qual o melhor estado cadastrado na base de dados para se vender um carro de marca popular e por quê?

Consideramos carros populares aqueles com valores abaixo de R$ 120.000,00 reais e por melhor aquele que tivesse o maior volume de vendas. Como esperado o estado com mais vendas se tratou de São Paulo.

Qual o melhor estado para se comprar uma picape com transmissão automática e por quê?

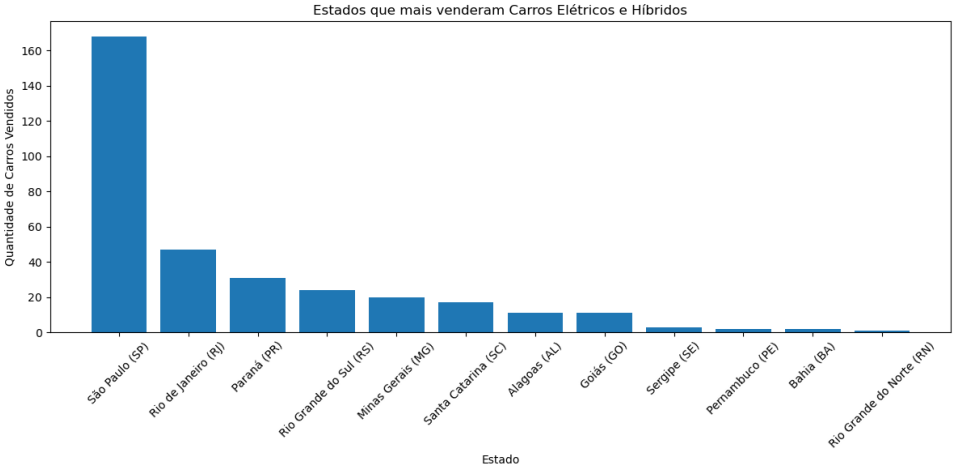
Qual o melhor estado para se comprar carros que ainda estejam dentro da garantia de fábrica e por quê?

Para as duas perguntas acima calculamos a média dos valores por estado e procuramos o estado com a menor média de preços. Como resultado o estado da Paraíba apresentou o menor valor em ambos os casos.

Notamos também em ambos os casos o estado de Sergipe como tendo os preços mais altos então refizemos a conta removendo os outliers que poderiam estar enviesando os resultados e obtivemos a mesma resposta. Mostrando que de fato a Paraíba é o estado com os menores valores para esse tipo de compra.

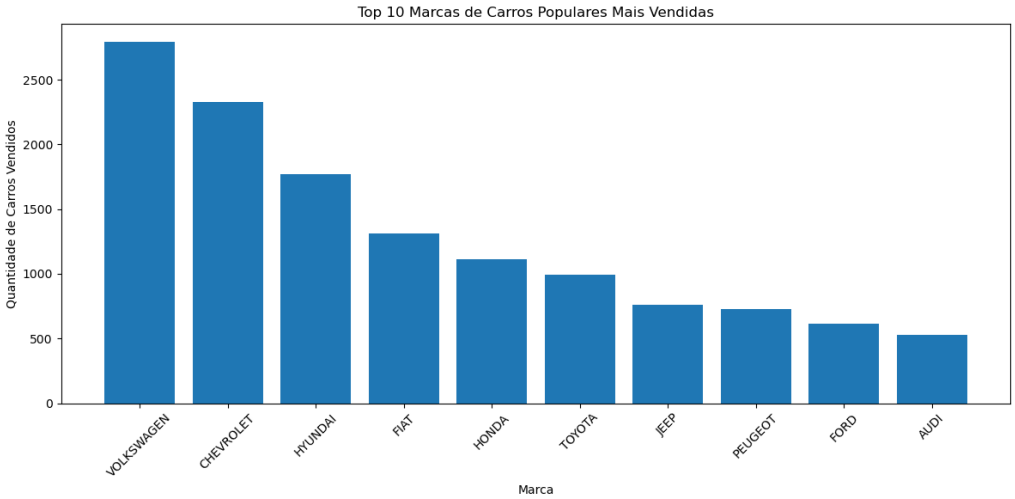
Qual os 3 estados mais venderam carros elétricos e híbridos?

São Paulo, Rio de Janeiro e Paraná



Quais as 5 marcas de carros populares mais vendidas?

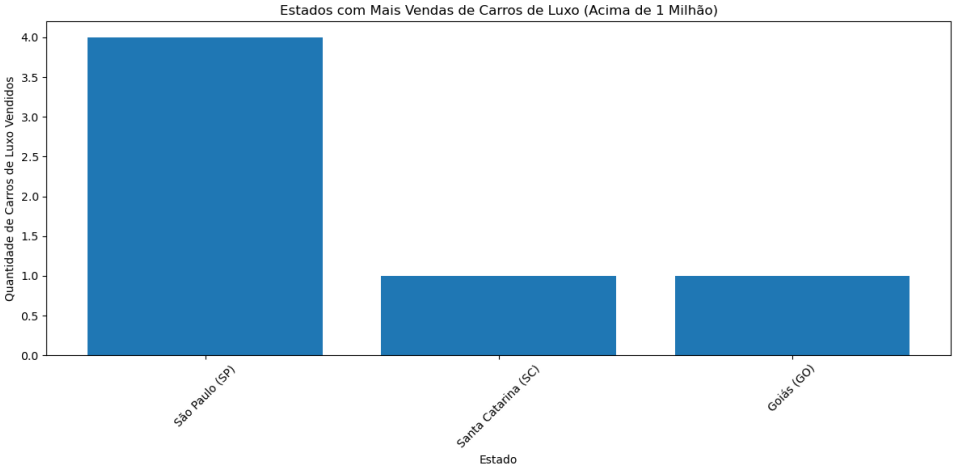
Volkswagen, Chevrolet, Hyundai, Fiat, Honda



Qual os 5 estados com mais vendas de carros de luxo(acima de 1 milhão)?

Notamos que apenas 3 estados tiveram registro na base de dados.

São Paulo, Santa Catarina, Goiás



Previsão

Se trata de um modelo de regressão pois estamos buscando um valor numérico ao contrário de um valor categórico. Foram testados e comparados vários modelos de regressão sendo que o algoritmo LightGBM se mostrou o mais eficiente de todos. A escolha das variáveis usadas tomou como base a observação da influência das mesmas no impacto do preço final e testes removendo algumas das variáveis escolhidas para verificar se não causavam overfitting. A medida de performance escolhida foi o r2_score.