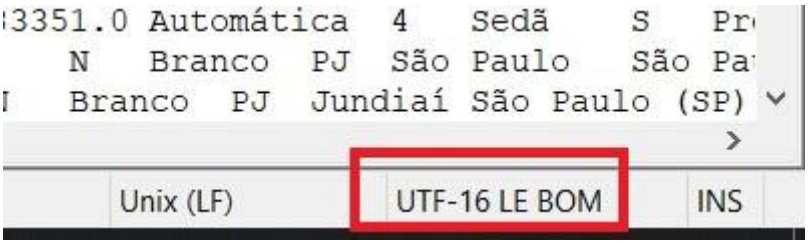
Relatório Indicium:

Vaga: Cientista de Dados
Candidato: Icaro Luis Mendes Ribeiro Moura
Email: icaroluis4@gmail.com
Repositório público: github.com/icaroluis4/desafio-indicium

1.Leitura

Primeiramente a base de dados 'train.csv' foi aberta em um editor de texto a fim de ter conhecimento do encoder do arquivo que é o utf-16-LE BOM.



Após uma breve leitura do arquivo em python das primeiras linhas, verificou-se o delimitador '\t' como separador dos campos no arquivo CSV.

```
['\ufeffid\tnum_fotos\tmarca\tmodelo\tversao\tano_de_fabricacao\tano_modelo\thodometro\
['3007162238985394196138630974698992223 92\t8.0\tNISSAN\tKICKS\t1.6 16V FLEXSTART SL 4P
['2796398421341295883064695661502886442 14\t8.0\tJEEP\tCOMPASS\t2.0 16V FLEX LIMITED AUT
['5641446081062104890029567823653817198 1\t16.0\tKIA\tSORENTO\t2.4 16V GASOLINA EX 7L AW
['5686250982684993342808637239015940554 5\t14.0\tVOLKSWAGEN\tAMAROK\t2.0 HIGHLINE 4X4 CD
['3389809757532003438945199098555980271 97\t8.0\tSSANGYONG\tKORANDO\t2.0 GLS 4X4 16V TUR
```

Verificou-se um número exorbitante de casas decimais após o ponto na coluna "**preco**" o que não faz sentido no mundo real, apenas duas casas após o ponto é o suficiente para definir os centavos dos preços.

Antes

Depois

A coluna “**veiculo_alienado**” foi retirada da base de dados uma vez que não possui um campo sequer preenchido.

```
26  revisoes_dentro_agenda   5910 non-null   object
27  veiculo_alienado         0 non-null      float64
```

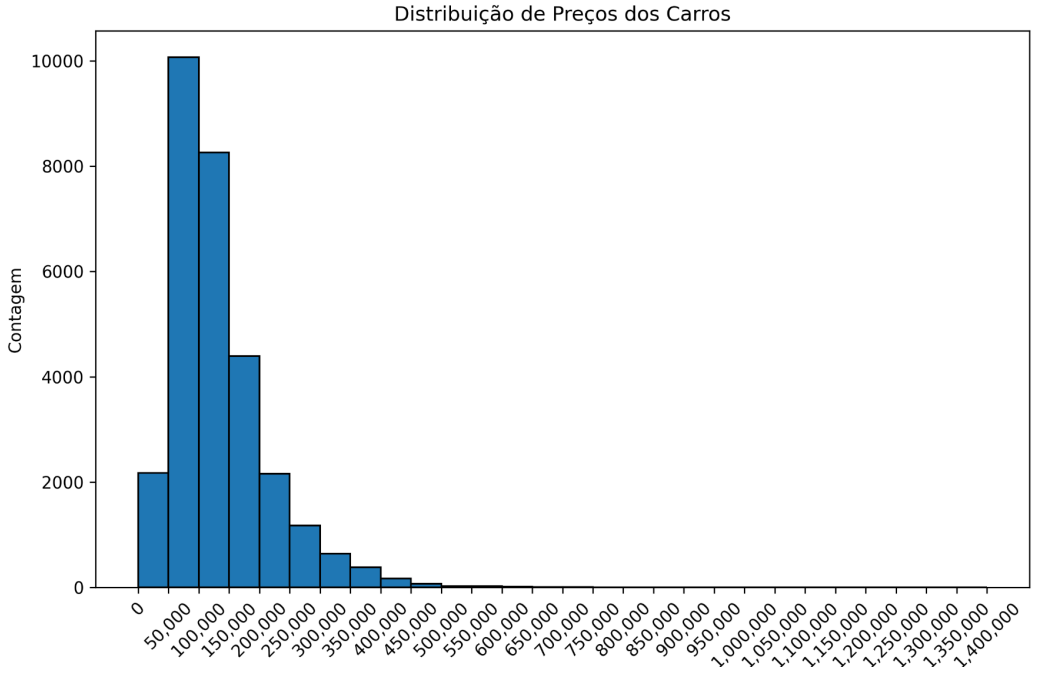
Features Numéricas
Serão apresentadas as principais estatísticas seguidas de alguns gráficos essenciais para exploração dos dados e por fim uma análise das principais features numéricas.

2.Preço

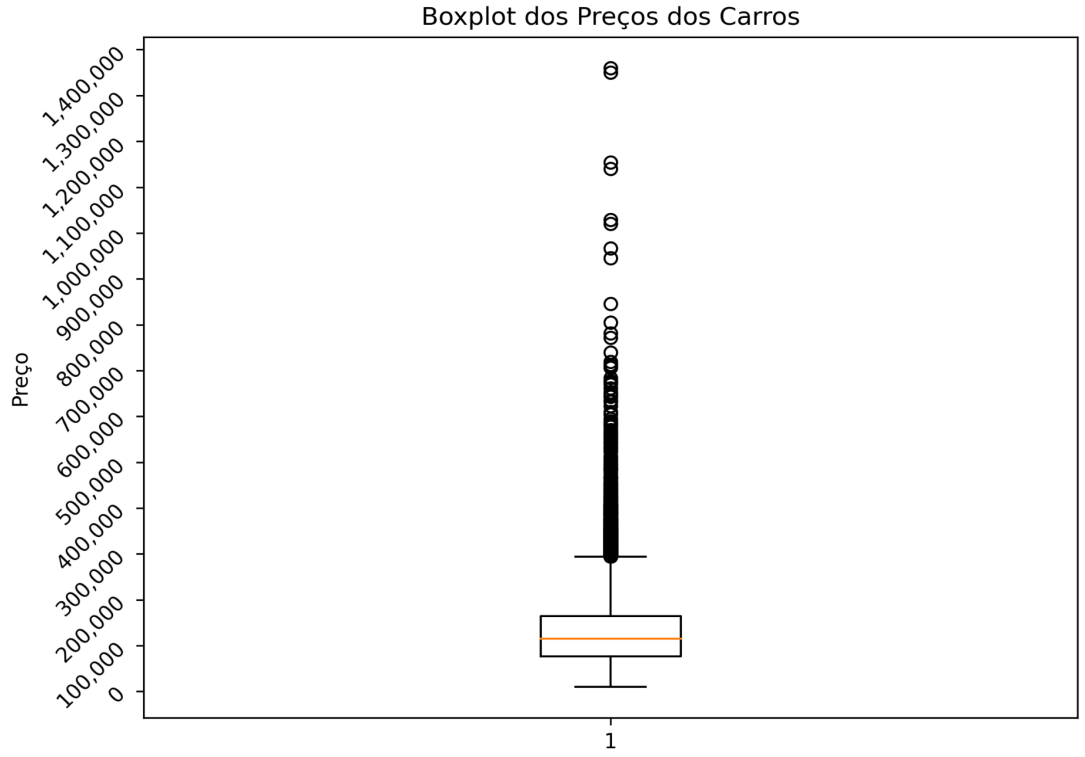
2.1 Principais Estatísticas

```
Estatísticas da coluna 'preco':
Média: 133023.87985769333
Mediana: 114355.795
Desvio Padrão: 81662.87224046292
Mínimo: 9869.95
Máximo: 1359812.89
```

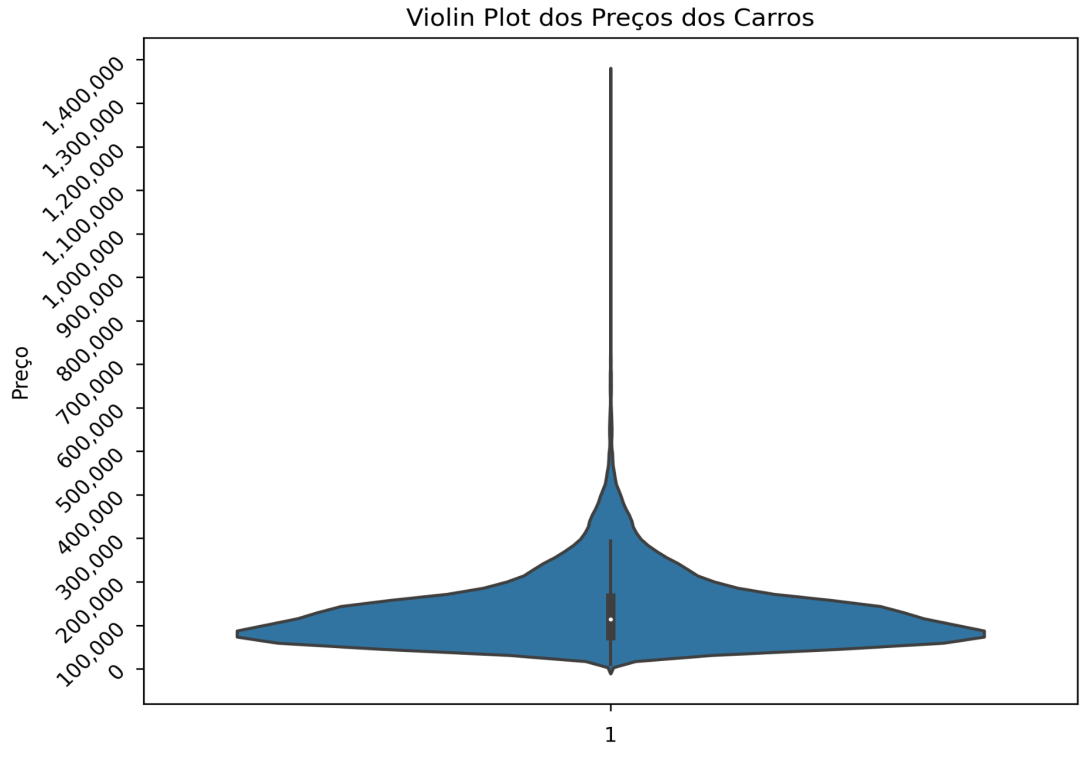
2.2 Histograma com intervalos a cada 50 mil



2.3 Gráfico Box Plot



2.4 Gráfico Violin Plot



2.5 Análise

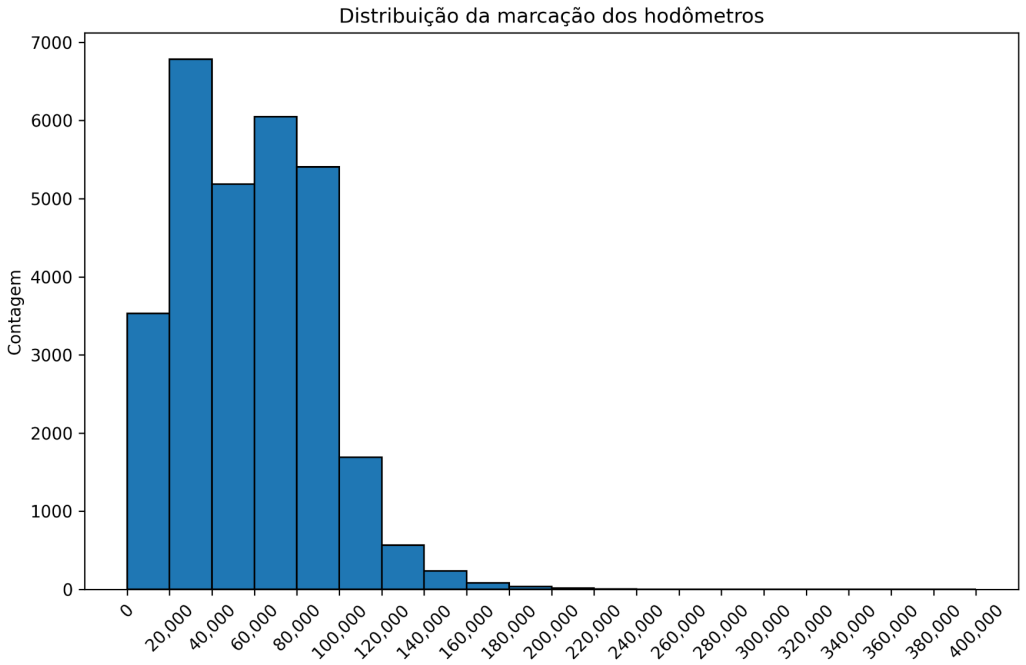
O preço é a feature central da base dados, baseado nas estatísticas encontradas pode-se observar uma grande quantidade de dispersão entre os dados com um desvio padrão massivo, baseado no histograma sabemos que a grande maioria dos preços está concentrada no intervalo entre 50 mil e 100 mil, baseado no gráfico box-plot podemos observar uma enorme quantidade de outliers algo que inclusive prejudica a visualização dos dados, desta forma se faz necessário um violin-plot para melhor visualização da densidade dos dados

3.Hodômetro

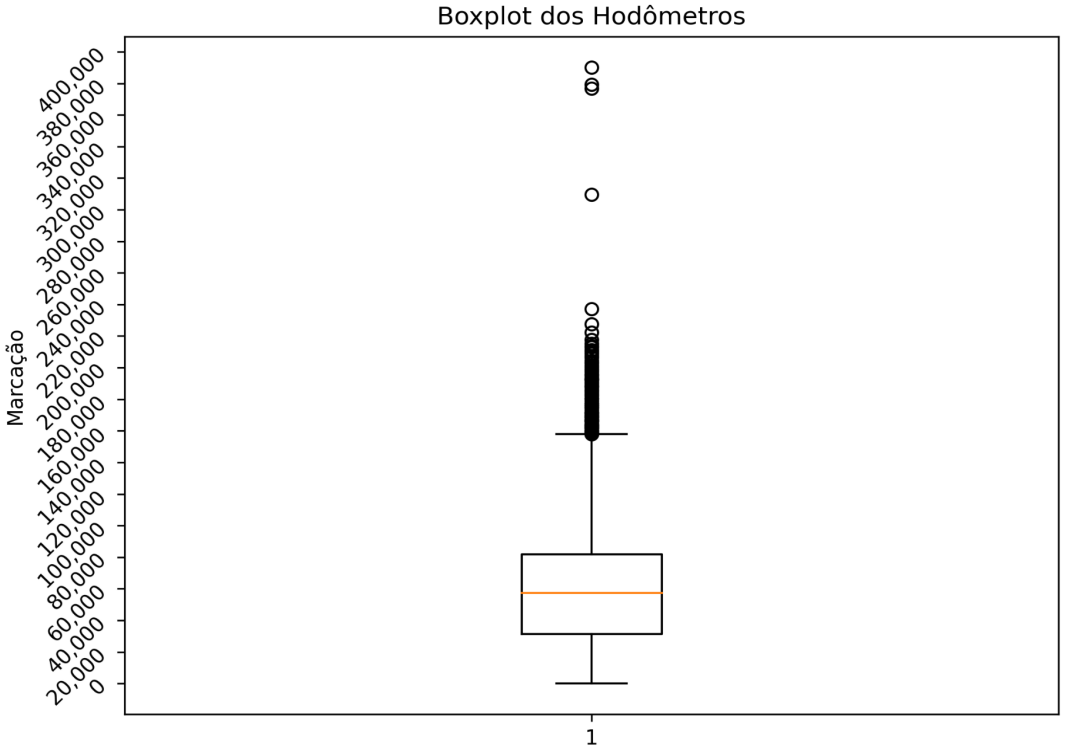
3.1 Principais Estatísticas

```
Estatísticas da coluna 'hodometro':
Média: 58430.59207679827
Mediana: 57434.0
Desvio Padrão: 32561.76930909199
Mínimo: 100.0
Máximo: 390065.0
```

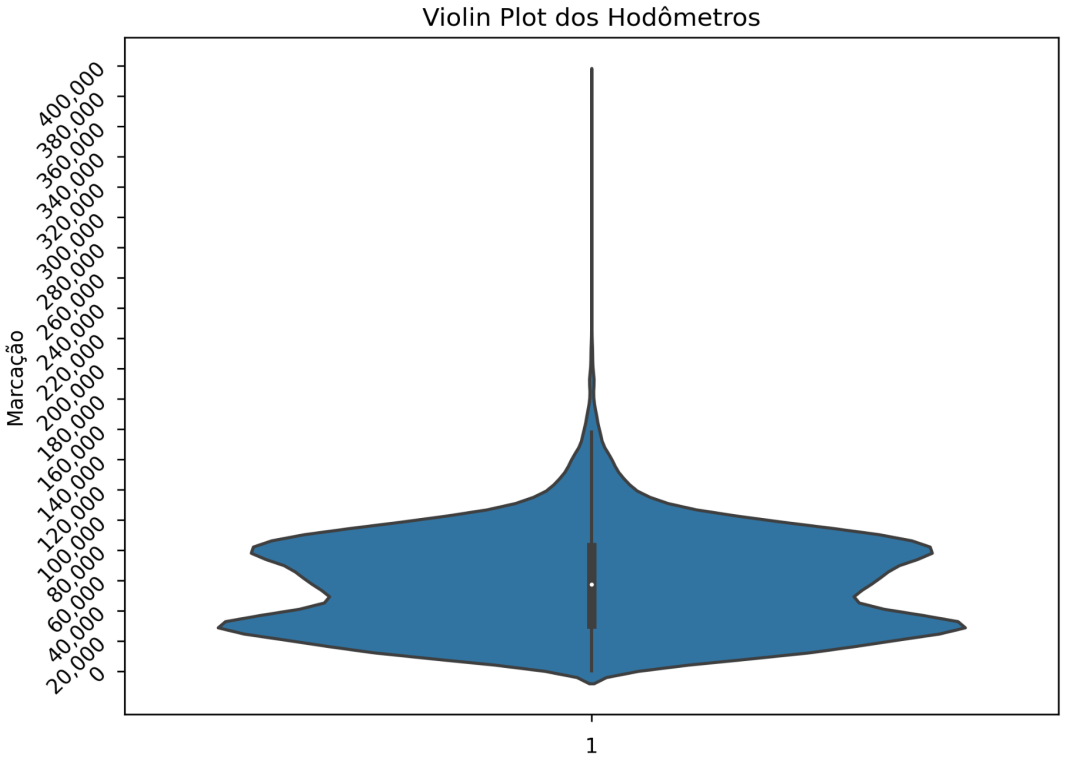
3.2 Histograma



3.3 Gráfico Box Plot



3.4 Gráfico Violin Plot



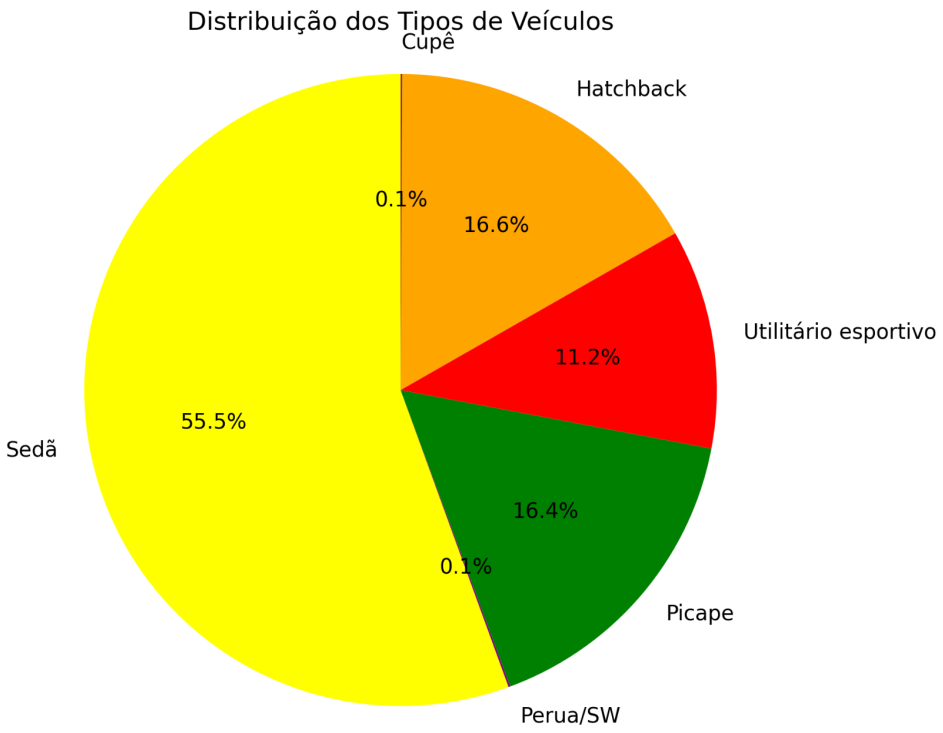
3.5 Análise

É possível notar que a média e a mediana são relativamente próximas o que nos leva a crer que existe uma simetria na distribuição dos dados. O histograma confirma a concentração dos dados em torno de um grande intervalo entre 0 a 100 mil. O gráfico box plot nos dá uma clara visualização da existência de muitos outliers mas isso não parece afetar a concentração dos dados, uma melhor visualização da densidade com o gráfico violin plot mostra que as duas grandes concentrações são nos valores aproximados de 30 mil e 80 mil.

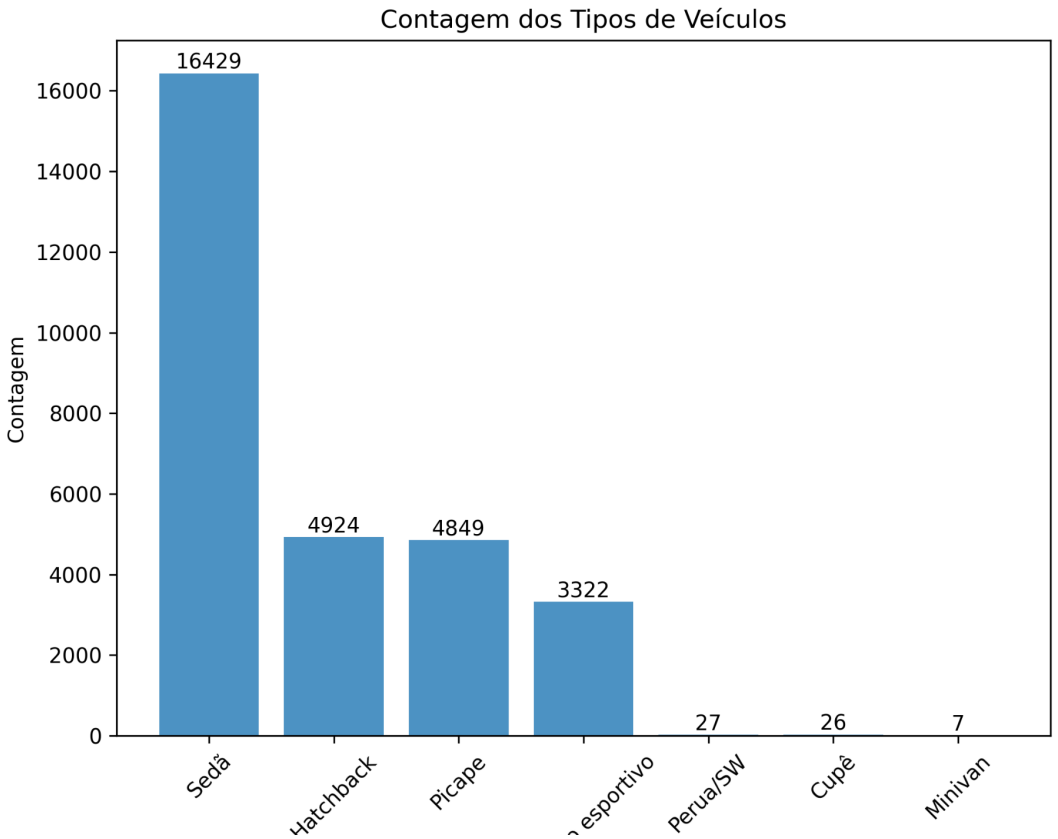
Features Categóricas
Serão apresentados alguns gráficos essenciais para exploração dos dados e por fim uma análise das principais features categóricas

4.Tipo

4.1 Tipo Gráfico de Pizza



4.2 Tipo Gráfico de Contagem

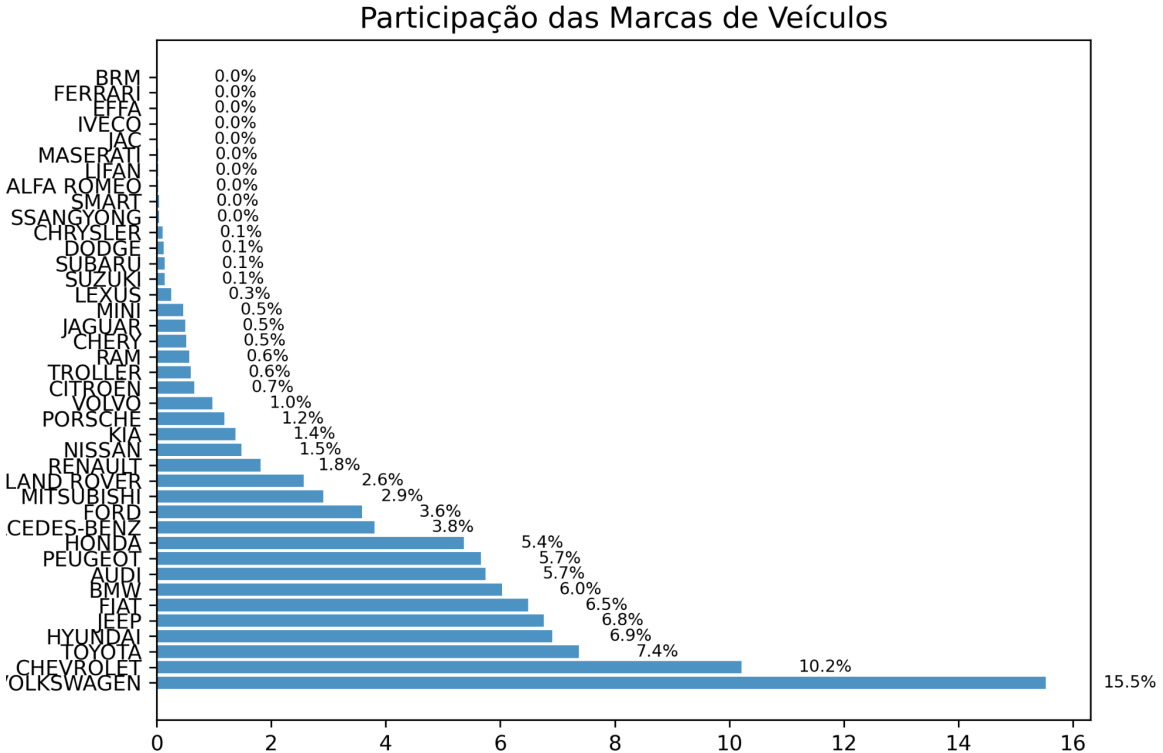


4.3 Análise

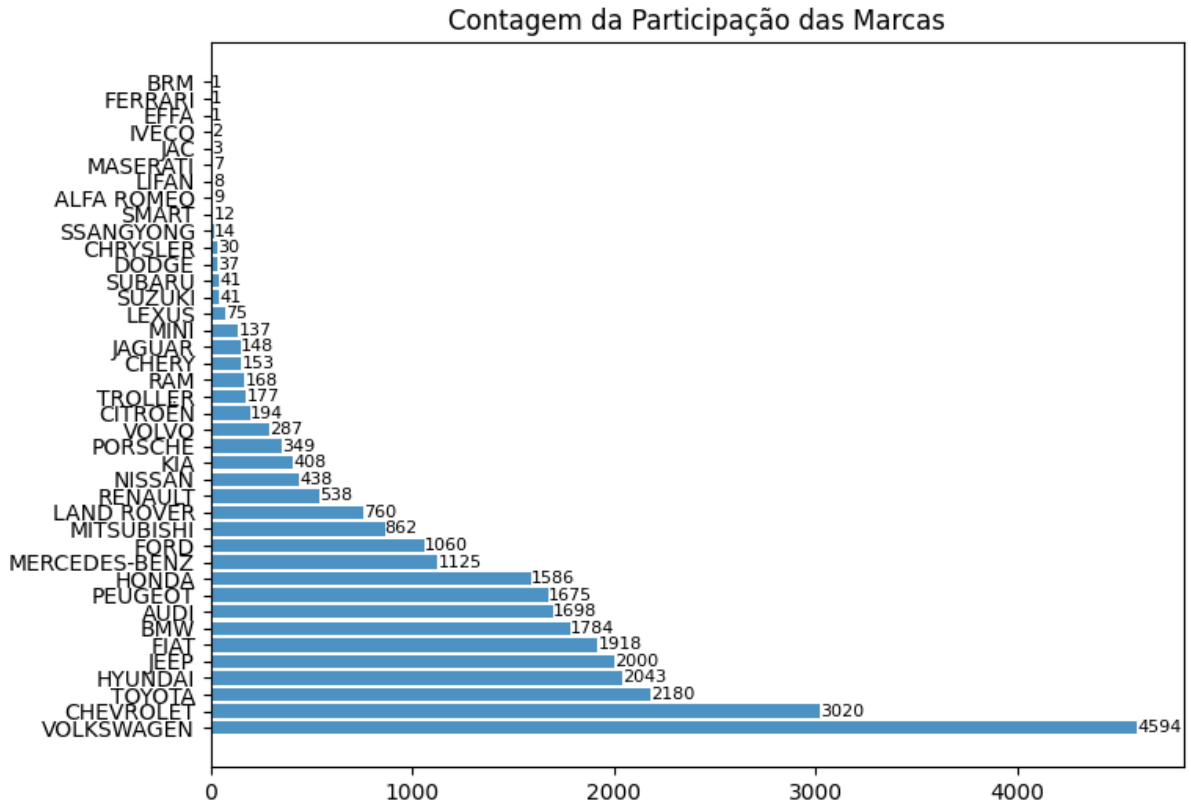
Através do gráfico de Pizza e de Contagem é possível observar que o grande fluxo de vendas da base de dados se concentra em carros do tipo sedan e a grande minoria se encontra em carros do tipo Perua/Sw , Cupê e Minivan.

5. Marca

5.1 Marca Porcentagem da Participação



5.2 Contagem da Participação

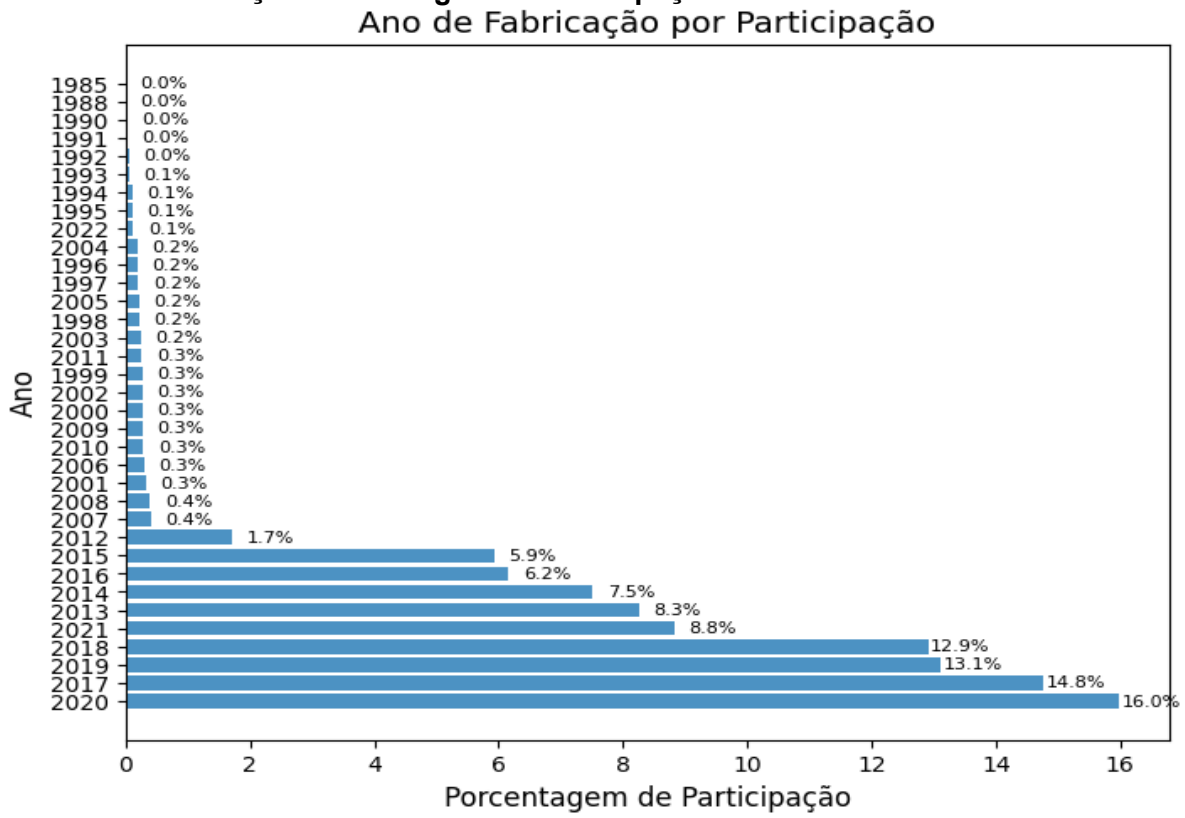


5.3 Análise

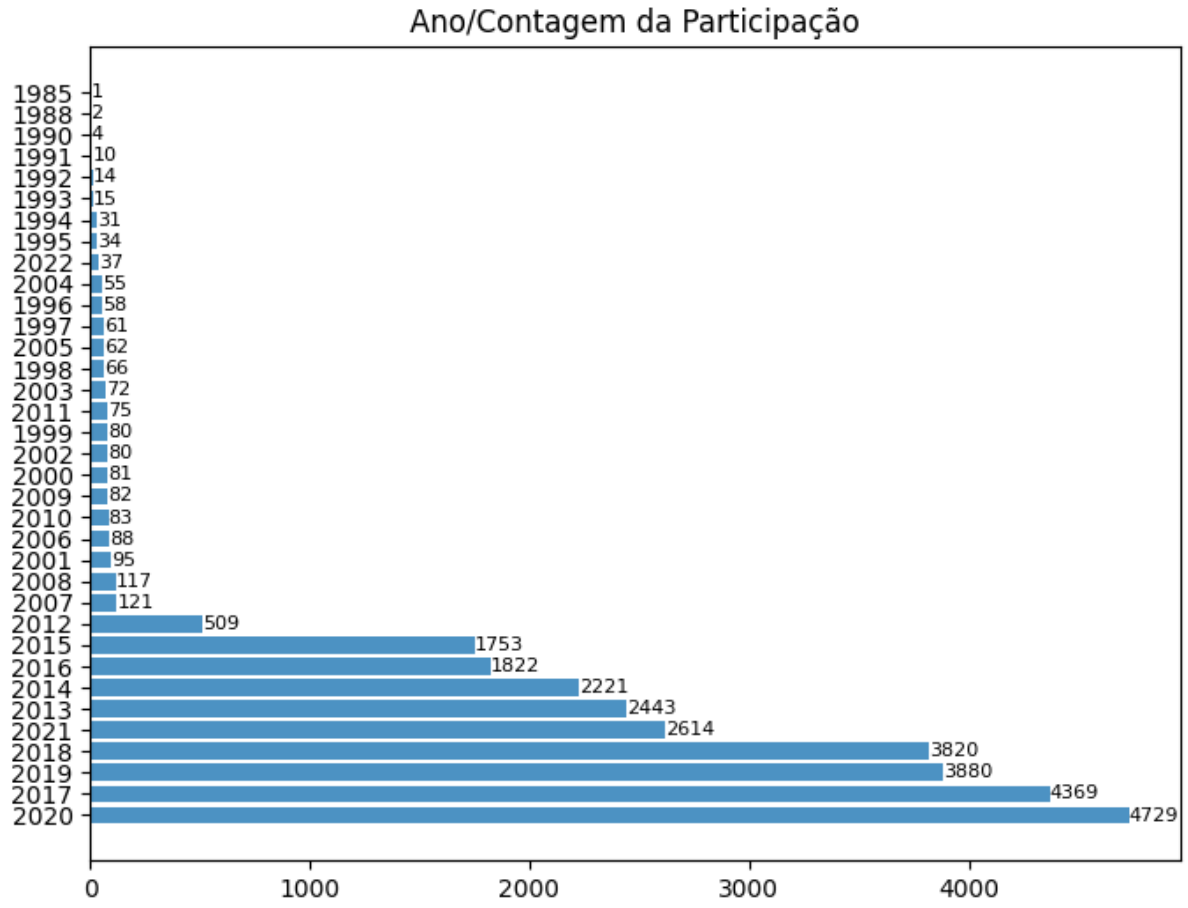
É possível perceber a partir dos gráficos que a maior participação nas vendas é da marca Volkswagen, seguida pela Chevrolet e Toyota. Os valores com 0.0% na verdade são valores com participação menor que 0.1% no data frame.

6. Ano

6.1 Ano de Fabricação Porcentagem de Participação



6.2 Ano de Fabricação contagem por ano

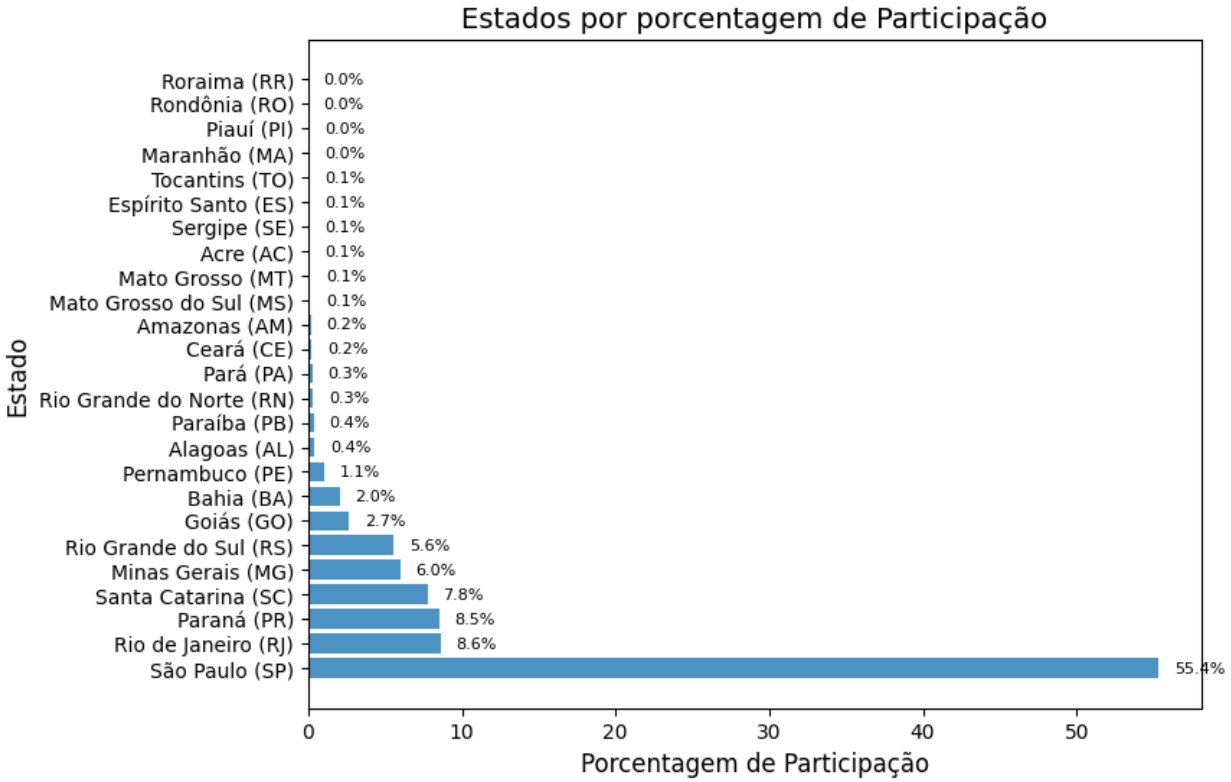


6.3 Análise

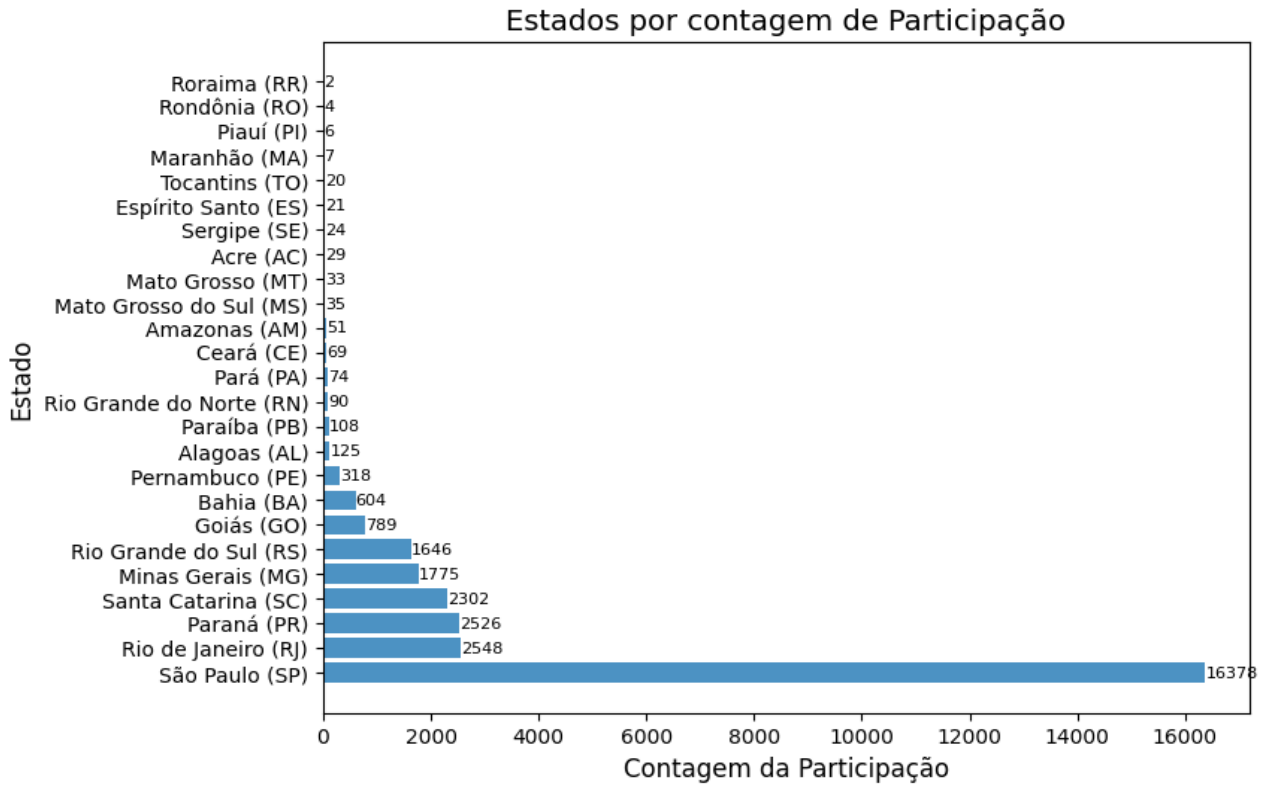
É possível notar que a grande parcela dos carros fabricados que foram vendidos foram fabricados nos últimos 10 anos

7. Estados

7.1 Estados por porcentagem de Participação nas vendas



7.2 Estados por contagem de Participação

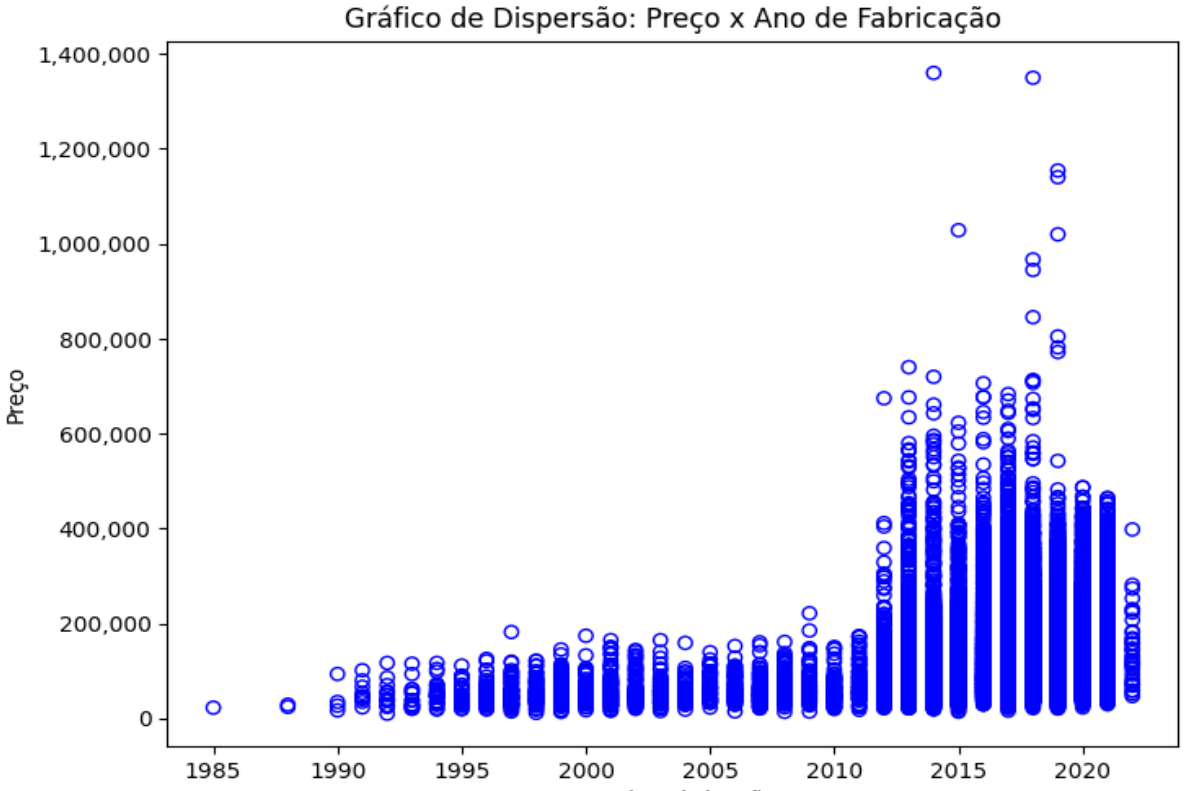


7.3 Análise

Após a plotagem dos gráficos fica nítido que o estado de São Paulo tem disparadamente a maior participação nas vendas presentes do Data Frame.

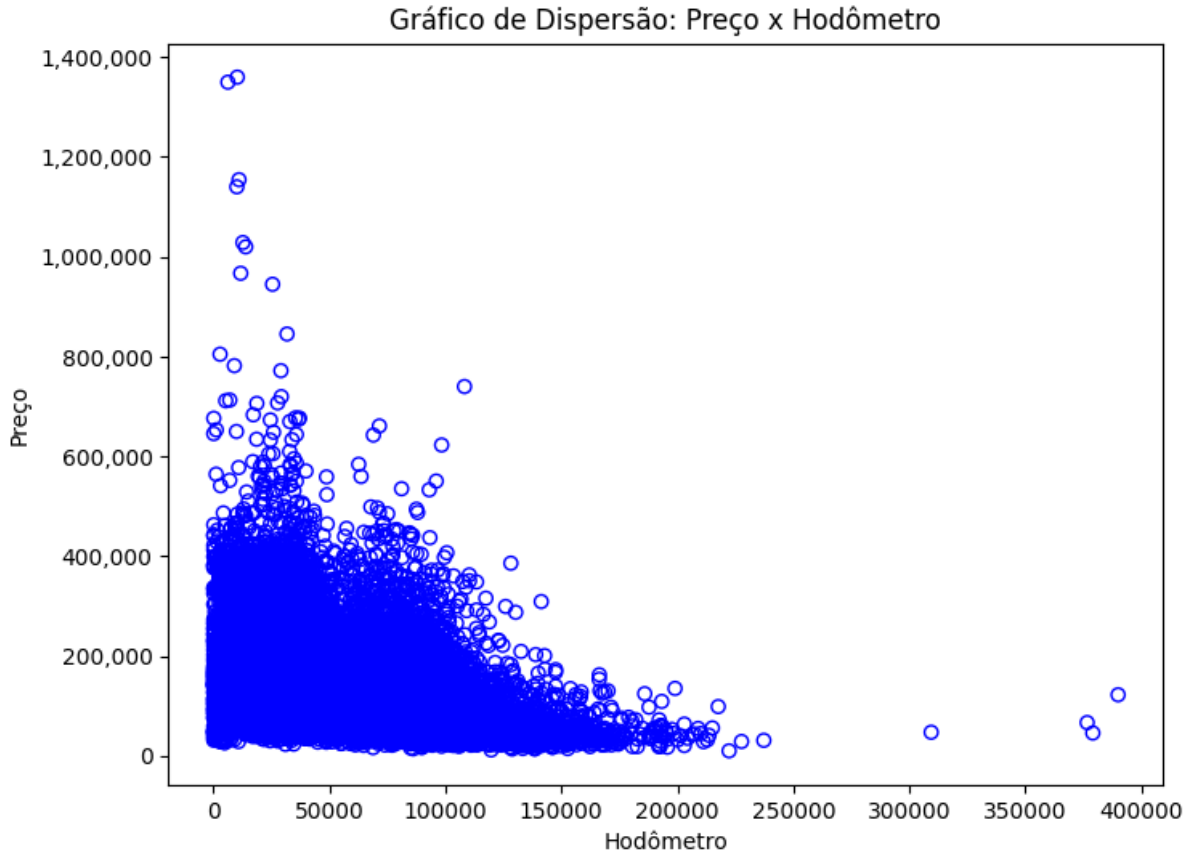
8.EDA

1.Hipótese: Existe uma relação inversa entre o preço médio dos carros e a idade dos veículos? Quanto mais velhos os carros, maior a tendência de redução no preço da venda ?
R: Para responder a esta hipótese se faz necessário utilizar um gráfico de dispersão Preço *versus* Ano de Fabricação.



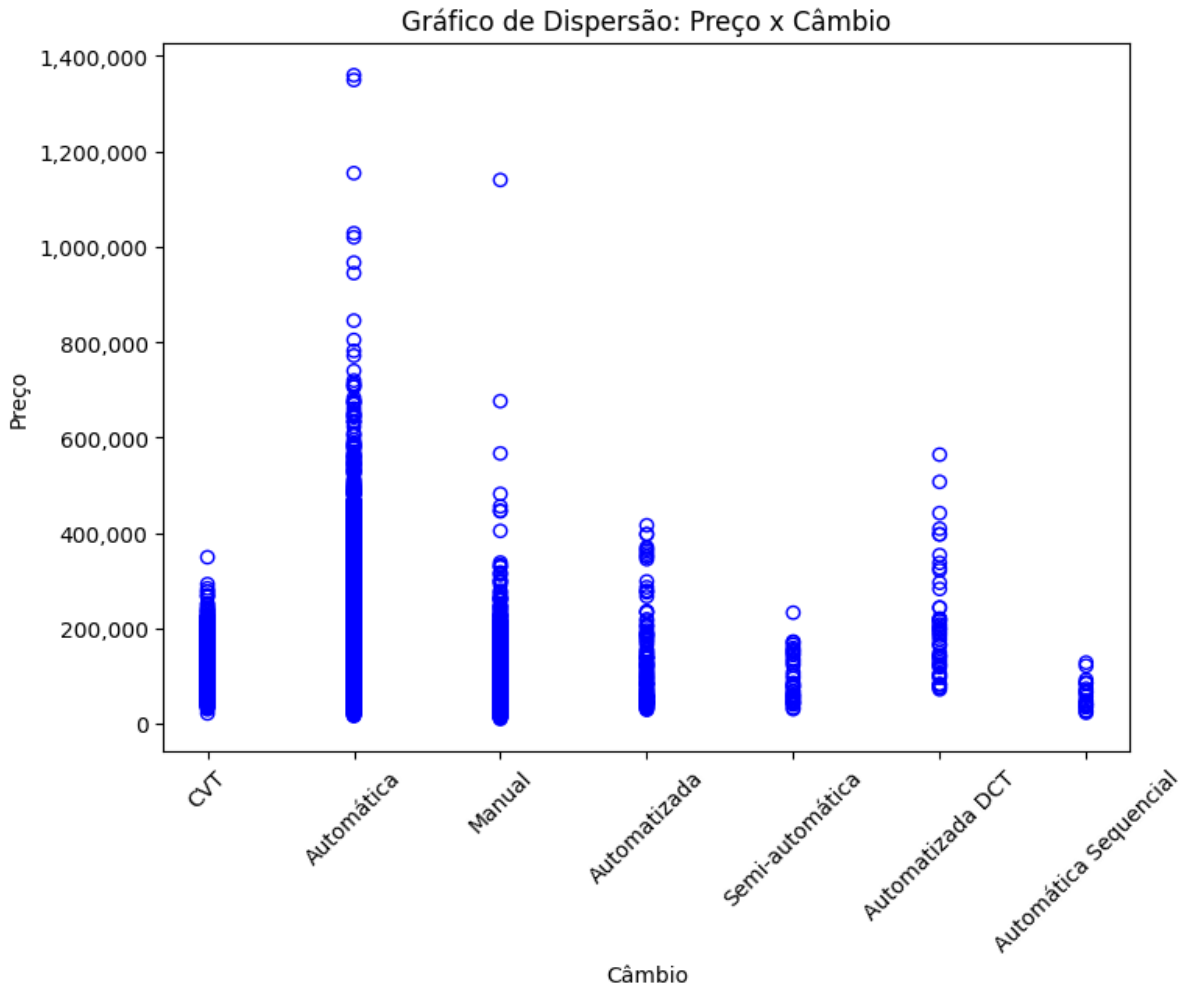
É possível observar uma clara tendência de carros com ano de fabricação mais antigos estarem ligados a preços de vendas mais baixos, muito raramente um carro fabricado antes 2010 atinge a faixa dos 200 mil.

2.Hipótese: Existe uma relação inversa entre o preço médio dos carros e as marcações dos hodômetros dos veículos? Quanto maior a quilometragem dos carros, maior a tendência de redução no preço da venda ?
R: Para responder a esta hipótese se faz necessário utilizar um gráfico de dispersão Preço *versus* Hodômetro.

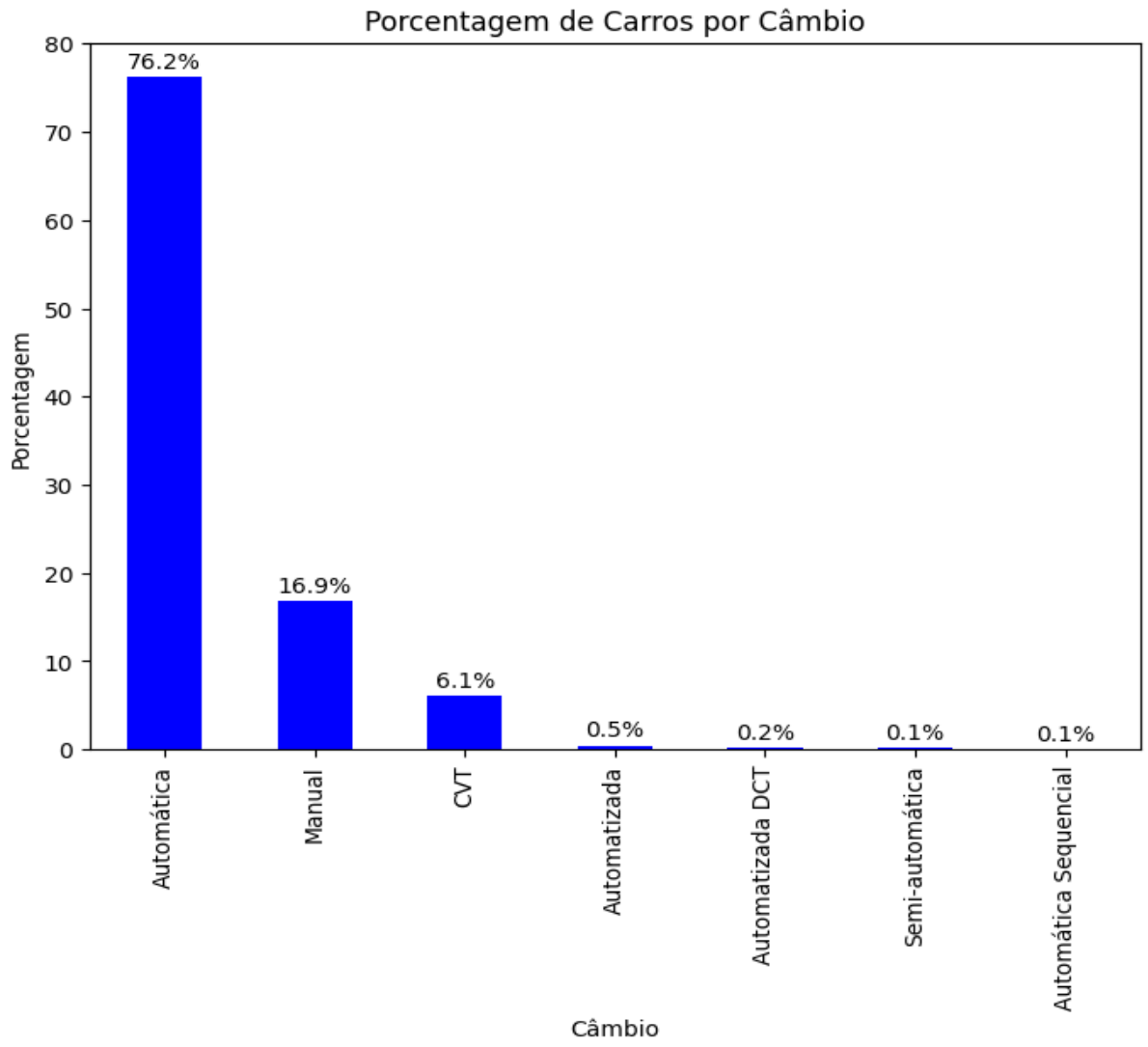


É possível constatar que carros acima de 150 mil Km marcados no hodômetro raramente são vendidos com um preço acima de 200 mil reais, logo existe uma tendência observável de que carros com alta quilometragem tendem a ter preços de venda mais baixos.

3.Hipótese: Existe alguma preferência de câmbio ? Carros com um tipo específico de câmbio tendem a ter o preço médio de venda mais alto ?
R:Para responder a esta hipótese se faz necessário utilizar um gráfico de dispersão Preço *versus* Câmbio.



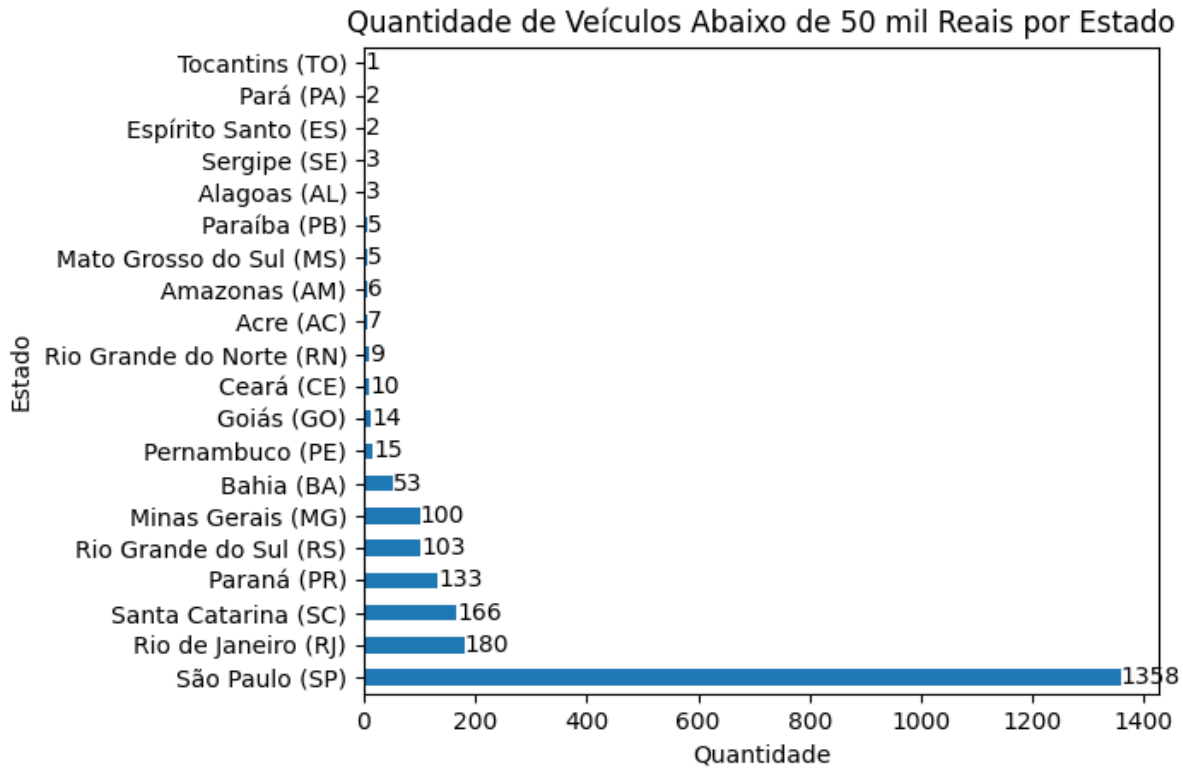
Não é possível afirmar com certeza que um carro com um câmbio específico tende a valorizar seu preço de venda, apesar do câmbio “Automático” possuir valores de vendas bem elevados para a média dos outros câmbios, o grosso de suas vendas pode estar na faixa de 0 a 100 mil reais, por isso se faz necessário saber a porcentagem dos carros presente no Data Frame por câmbio.



76.2% de todos os carros presentes no Data Frame possuem o campo “Automática” na coluna câmbio, reforçando a tese de que a grande maioria dos carros que possuem esse câmbio não necessariamente refletirá em preços mais altos de venda.

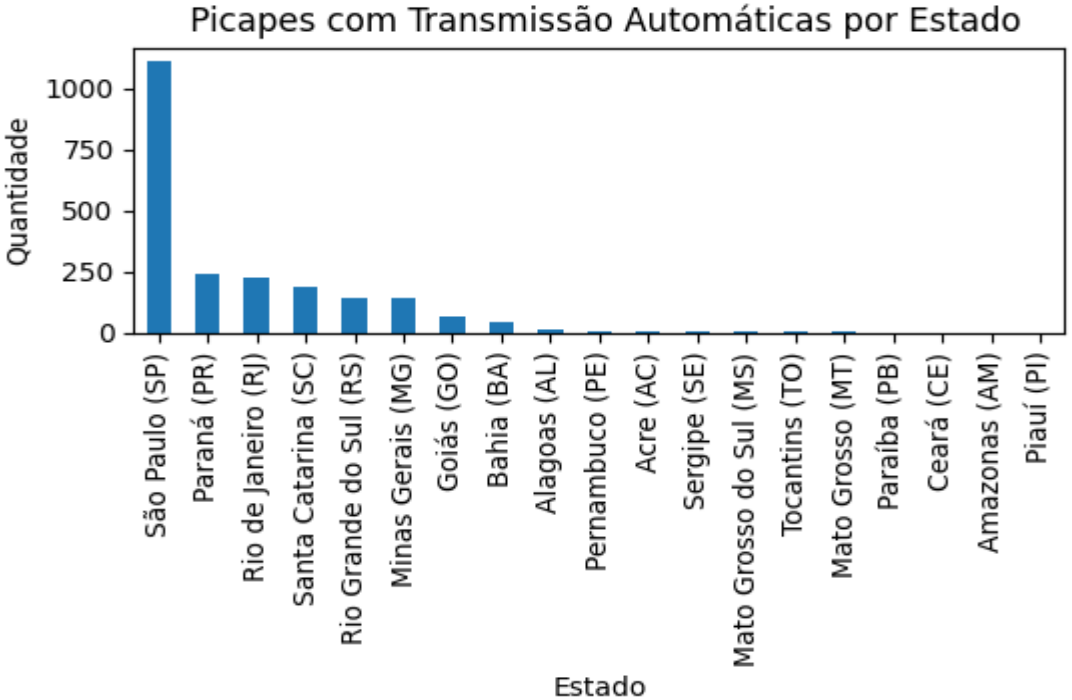
a.Pergunta: Qual o melhor estado cadastrado na base de dados para se vender um carro de marca popular e por quê?

R: Vamos considerar um carro popular como um veículo que seja vendido abaixo de 50 mil reais.Logo:



É possível constatar que o melhor estado para se vender um carro é o estado de São Paulo devido a alta demanda observável.

b.Pergunta: Qual o melhor estado para se comprar uma picape com transmissão automática e por quê?
R:Para responder essa pergunta primeiro precisamos saber a quantidade de Picapes com Transmissão Automática por Estado. Aqui se encontra uma pegadinha devido a muitas vezes as pessoas confundirem câmbio com transmissão, encontraremos a informação sobre a transmissão do veículo na coluna "**modelo**" e não na coluna "**câmbio**", Logo:

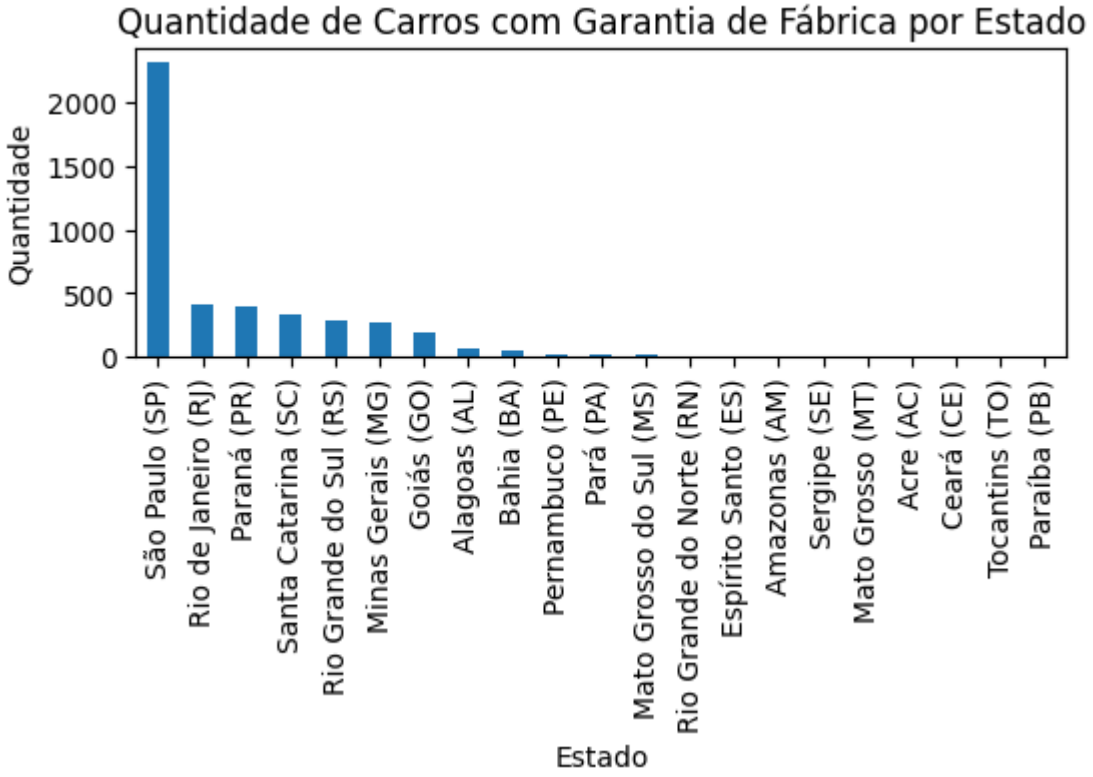


Porém isso não necessariamente significa que o estado de São Paulo é o melhor para se comprar uma picape com transmissão automática, para isso precisamos saber o preço médio da venda, por isso faremos um filtro nos estados com pelo menos 100 vendas feitas de picapes com transmissão automática para encontrar o melhor preço médio. Logo:

estado_vendedor	
Minas Gerais (MG)	202995.50
Paraná (PR)	205766.56
Rio Grande do Sul (RS)	208011.46
Rio de Janeiro (RJ)	189944.80
Santa Catarina (SC)	190214.25
São Paulo (SP)	196964.02

O melhor preço médio para se comprar uma picape com transmissão automática acabou por se confirmar ser no estado de São Paulo.

c.Pergunta: Qual o melhor estado para se comprar carros que ainda estejam dentro da garantia de fábrica e por quê?
R:Para responder a essa pergunta precisamos filtrar os dados para incluir apenas os carros que possuem garantia de fábrica, depois calcular a contagem de carros com garantia de fábrica por estado e por fim calcular o preço médio por estados que tenham pelo menos 100 vendas feitas com carros dentro da garantia de fábrica.



Cálculo do preço médio:

estado_vendedor	
Goiás (GO)	174521.41
Minas Gerais (MG)	161206.27
Paraná (PR)	175578.25
Rio Grande do Sul (RS)	176442.24
Rio de Janeiro (RJ)	179109.86
Santa Catarina (SC)	173798.44
São Paulo (SP)	166751.08

Apesar do estado de São Paulo ter o maior fluxo de vendas o melhor preço médio fica com o estado de Minas Gerais, portanto o melhor estado para comprar um carro dentro da garantia de fábrica é Minas Gerais

9. Modelo

Como o objetivo é fazer a previsão de preços que é um valor numérico a partir dos dados, sabemos que trata-se de um problema de **regressão linear**.

9.1 Variáveis

No processo de previsão, foram utilizadas as seguintes variáveis:

Hodômetro: Essa variável indica o número de quilômetros rodados pelo carro. É uma feature importante, pois geralmente carros com menor quilometragem tendem a ter um preço mais alto.
Ano de fabricação: O ano de fabricação do carro é uma variável relevante, pois carros mais recentes costumam ter um valor de mercado mais alto.
Estado do vendedor: A localização geográfica do vendedor pode influenciar os preços dos carros, pois há diferenças regionais no mercado automotivo.
Marca: A marca do carro é uma característica importante, pois diferentes marcas podem ter valores de mercado distintos.
Tipo: O tipo de carro, como sedan, hatchback, SUV, etc., pode influenciar o preço, já que diferentes tipos de veículos têm demandas e características diferentes.
Câmbio: O tipo de câmbio do carro, como manual ou automático, pode influenciar no preço.
Anunciante: O anunciante do carro também pode ter um impacto no preço, pois vendedores particulares e concessionárias podem ter diferentes políticas de preços.
Modelo e **versão**: Essas informações específicas do veículo podem ser úteis para prever o preço, uma vez que diferentes modelos e versões podem ter valores de mercado distintos.
Transformações aplicadas nas variáveis:

As colunas categóricas (estado do vendedor, marca, tipo, câmbio, anunciante, modelo e versão) foram convertidas em variáveis categóricas e codificadas usando a técnica de codificação **one-hot**, para representá-las numericamente no modelo. Essa transformação permite que o algoritmo aprenda as relações entre as categorias e os preços.

9.2 Modelo utilizado:

Foi utilizado o modelo XGBoostRegressor, um algoritmo baseado em árvores de decisão otimizado para regressão. O XGBoost é conhecido por sua eficácia em competições de ciência de dados ele é capaz de lidar com combinações de variáveis categóricas e numéricas e capturar relações complexas entre as variáveis é um modelo amplamente utilizado em problemas de regressão.
Porém, é importante ressaltar que o XGBoost requer um ajuste cuidadoso de hiperparâmetros para obter o melhor desempenho. O processo de ajuste de hiperparâmetros pode ser demorado, pois há vários hiperparâmetros a serem otimizados. Além disso, assim como outros algoritmos baseados em árvores de decisão, o XGBoost pode ser sensível a ruídos e outliers nos dados. Se esses outliers não forem tratados adequadamente, eles podem afetar negativamente o desempenho do modelo, levando ao *overfitting*.

A medida de performance escolhida foi o Mean Squared Error (MSE), que calcula a média dos erros quadrados entre os valores reais e previstos. O MSE é uma métrica comumente utilizada em problemas de regressão e penaliza erros maiores de forma quadrática. A raiz quadrada do MSE (RMSE) também foi calculada para facilitar a interpretação do erro.

10.Conclusão e Agradecimento

Agradeço a equipe da Indicium por permitir aplicar meus conhecimentos prévios e aprender vários novos nesse desafio, certamente este é um projeto que vai para o meu portfólio, a adição de labels novas no Data Frame de teste foi algo muito desafiador de resolver e tomou uma parte dos meus neurônios, no mais sigo trilhando meu caminho no campo da Ciência de Dados, espero um dia poder participar de um time tão ávido em resolver problemas como o da Indicium, obrigado.