RELATÓRIO: Desafio Cientista de Dados

AUTOR: João Diamantino

LINKEDIN: https://www.linkedin.com/in/joaodiamantino-datascience/

REPOSITÓRIO: https://github.com/JoaoDiamantino/CarsPricePredict_LightHouse

1.Análise estatística dos dados

Este relatório apresenta uma análise das principais estatísticas descritivas da base de dados '*train*' utilizando as variáveis disponíveis. O objetivo é fornecer uma visão geral das características e informações resumidas dos dados.

A análise estatística descritiva fornece uma visão geral das principais características das variáveis da base de dados. As estatísticas descritivas, como **média**, **mediana**, **desvio padrão** e **quartis**, nos ajudam a compreender **a distribuição e a variabilidade dos dados**. Os gráficos utilizados na análise visual fornecem uma representação visual das distribuições e possíveis valores atípicos.

Ao interpretar as estatísticas descritivas e os gráficos, é **possível identificar padrões**, **tendências e anomalias nos dados**, fornecendo insights valiosos para análises posteriores. É importante considerar o contexto do problema e a natureza das variáveis ao interpretar os resultados.

Cabe ressaltar que este relatório apresenta apenas uma análise preliminar das estatísticas descritivas das variáveis da base de dados. Análises mais aprofundadas e específicas podem ser realizadas de acordo com os objetivos e requisitos do projeto.

	num_fotos	ano_de_fabricacao	ano_modelo	hodometro	preco
count	29584.000000	29584.000000	29584.000000	29584.000000	2.958400e+04
mean	10.309931	2016.758552	2017.808985	58430.592077	1.002299e+16
std	3.481502	4.062422	2.673930	32561.769309	9.547504e+15
min	8.000000	1985.000000	1997.000000	100.000000	2.006292e+10
25%	8.000000	2015.000000	2016.000000	31214.000000	3.196583e+15
50%	8.000000	2018.000000	2018.000000	57434.000000	8.244891e+15
75%	14.000000	2019.000000	2020.000000	81953.500000	1.342533e+16
max	21.000000	2022.000000	2023.000000	390065.000000	6.549912e+16

Tabela 1: Descrição estatística das variáveis numéricas

Através da Tabela 1 é possível perceber comportamentos específicos das variáveis numéricas:

- A variável 'num_fotos' possui mais da metade de valores identificados com 8, pois o valor mínimo (min), 1° quartil (25%) e mediana (50%) são mostrados com este valor.

- A variável ‘ano_de_fabricação’ mostra que os valores representados são em maioria recentes, de 2015 a 2023. Anos mais antigos podem representar alguma anomalia ou outliers.
- A variável ‘hodômetro’ possui uma grande variabilidade de dados, pois além da grande diferença entre o mínimo e máximo (max), vemos o grande valor do desvio padrão (std) se formos analisar em conjunto coma média (mean)
- O mesmo valendo para a variável ‘preco’ que indica grande variabilidade dos dados, com valores bastante discrepantes.

Um boxplot, também conhecido como diagrama de caixa, é um gráfico estatístico que fornece informações sobre a distribuição de um conjunto de dados. Ele é composto por um retângulo (a "caixa") que representa os quartis Q1, Q2 (mediana) e Q3 dos dados. Uma linha vertical é traçada dentro da caixa na posição da mediana.

O **boxplot** é útil para identificar a assimetria dos dados, a presença de valores extremos e a comparação de distribuições entre diferentes grupos ou variáveis. Ele fornece uma representação visual compacta das principais estatísticas descritivas, permitindo uma rápida compreensão da distribuição dos dados.

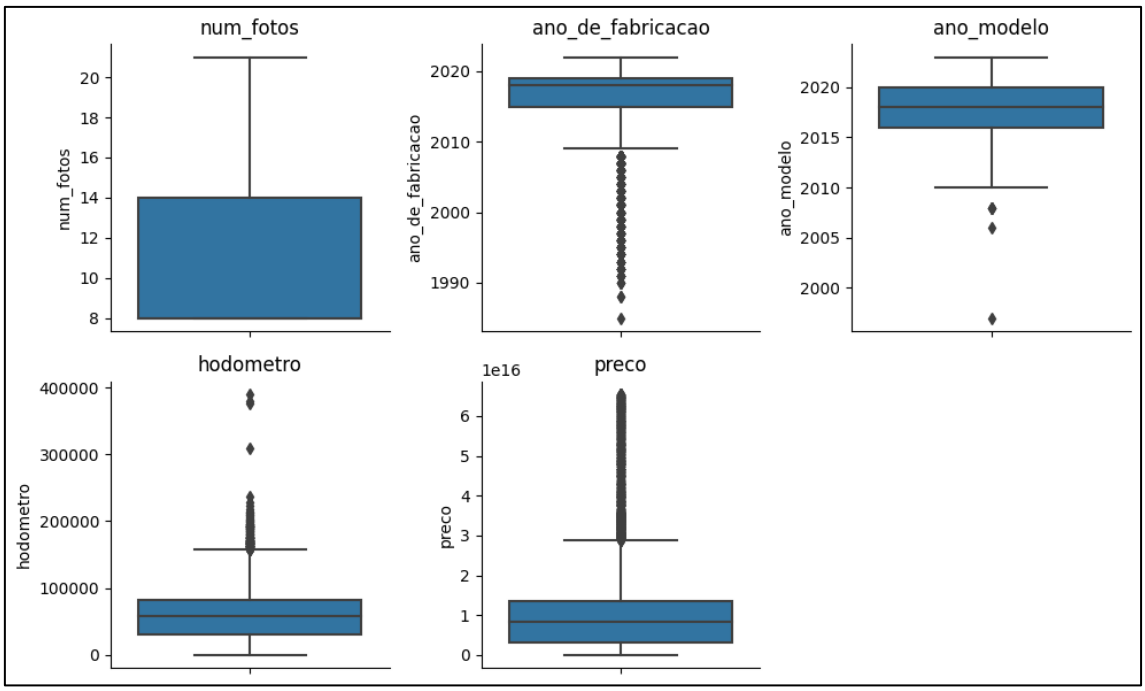


Figura 1: Boxplot das variáveis numéricas.

Um **histograma** é uma representação gráfica da distribuição de frequência de um conjunto de dados. Ele é construído através da divisão do intervalo dos dados em intervalos (chamados de bins) e contagem do número de observações que se enquadram em cada intervalo.

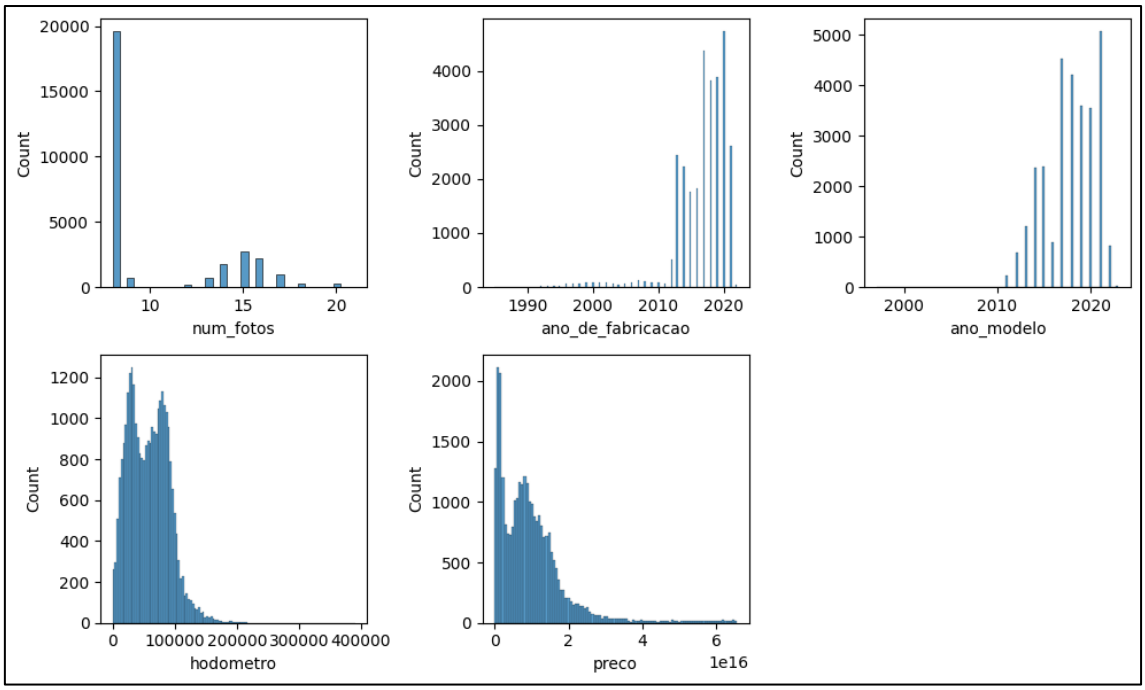


Figura 2: Histograma das variáveis numéricas

Um, também conhecido como **matriz de correlação**, é uma representação gráfica que mostra as relações de correlação entre as variáveis de um conjunto de dados. É uma ferramenta poderosa na Análise Exploratória de Dados para identificar associações entre variáveis e entender sua interdependência.

Nesse tipo de gráfico, cada célula da matriz representa a correlação entre duas variáveis. A intensidade da correlação é representada por cores ou tons, onde cores mais escuras indicam uma correlação mais forte (positiva ou negativa) e cores mais claras indicam uma correlação mais fraca ou inexistente.

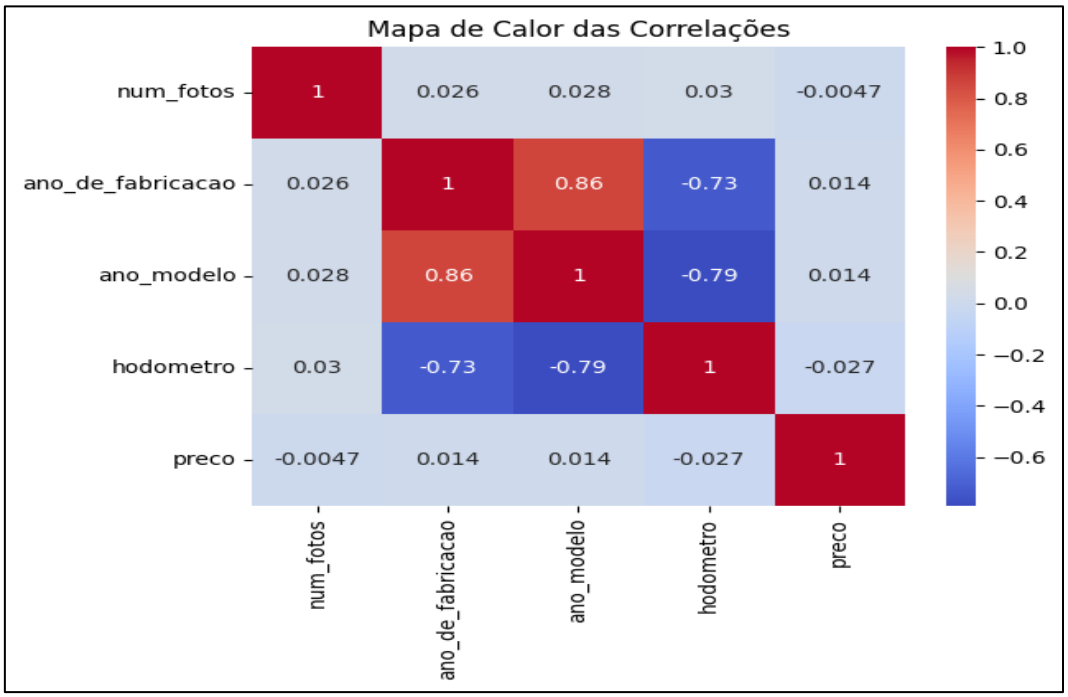


Figura 3Matriz de Correlação das variáveis numéricas

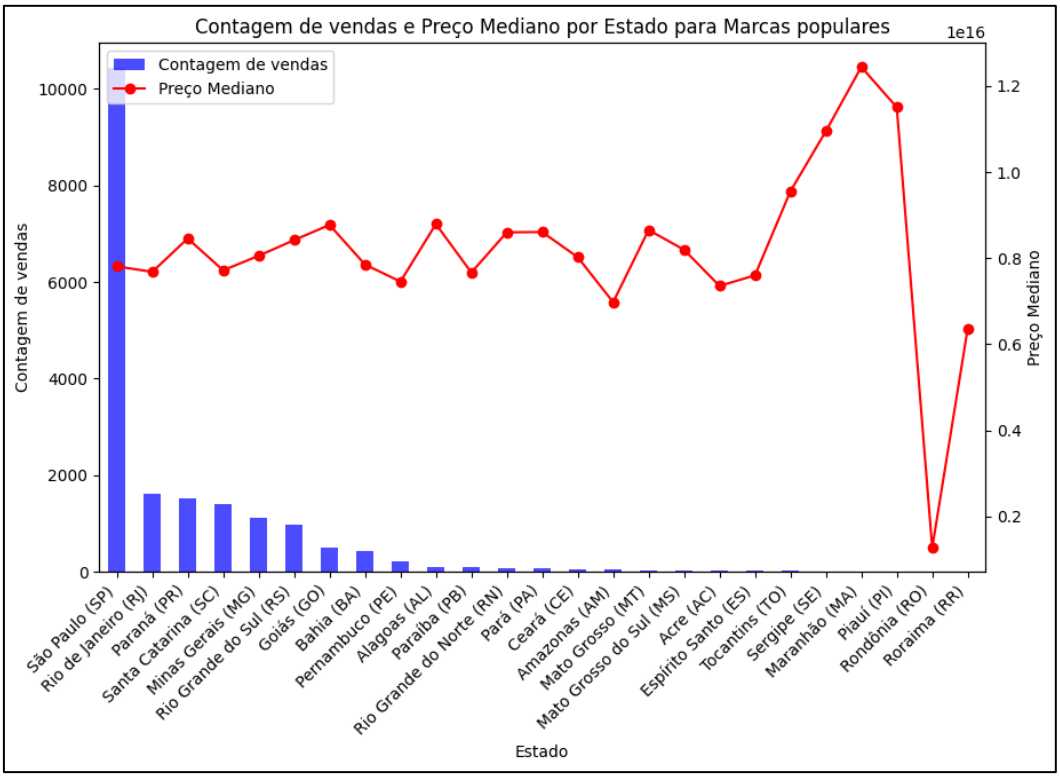
2. Análise Exploratória de Dados

A **Análise Exploratória de Dados** (EDA, do inglês Exploratory Data Analysis) é uma etapa fundamental na análise de dados, que tem como objetivo **explorar e entender os dados** disponíveis antes de aplicar técnicas estatísticas mais avançadas ou construir modelos preditivos.

Durante o processo de EDA, o analista examina e investiga os dados de forma sistemática, **buscando identificar padrões, tendências, relacionamentos e anomalias que possam estar presentes**. Essa abordagem exploratória permite uma compreensão mais profunda dos dados, bem como **a formulação de perguntas e hipóteses** para investigações futuras.

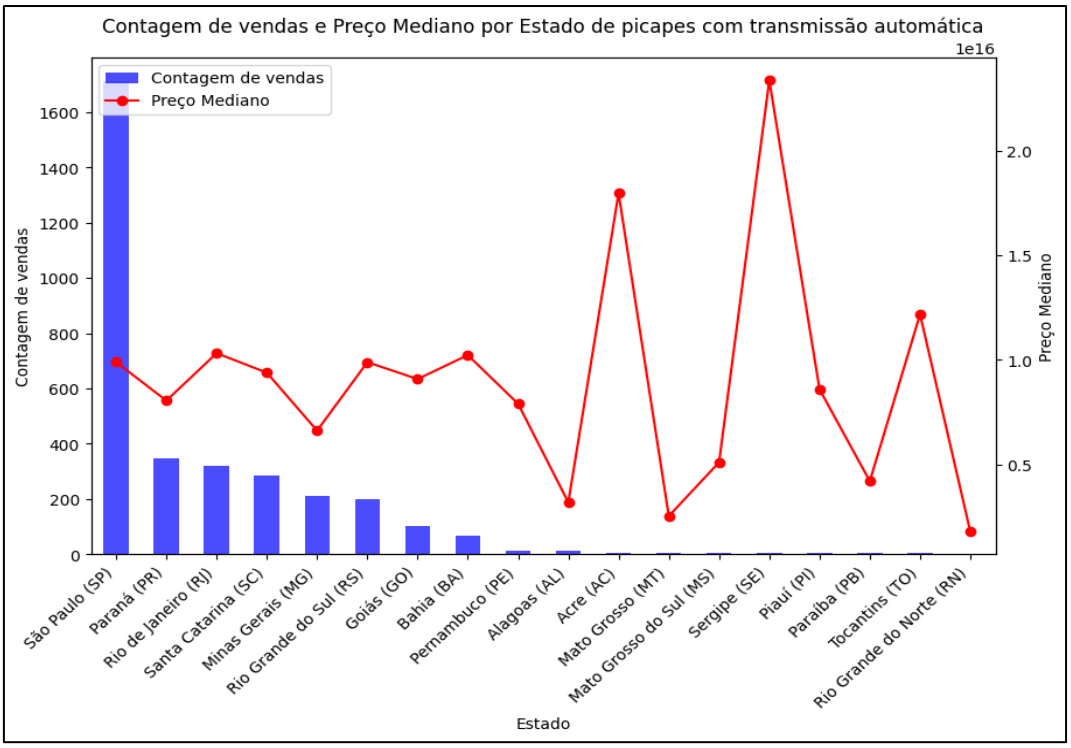
Vamos responder algumas hipóteses de negócio:

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



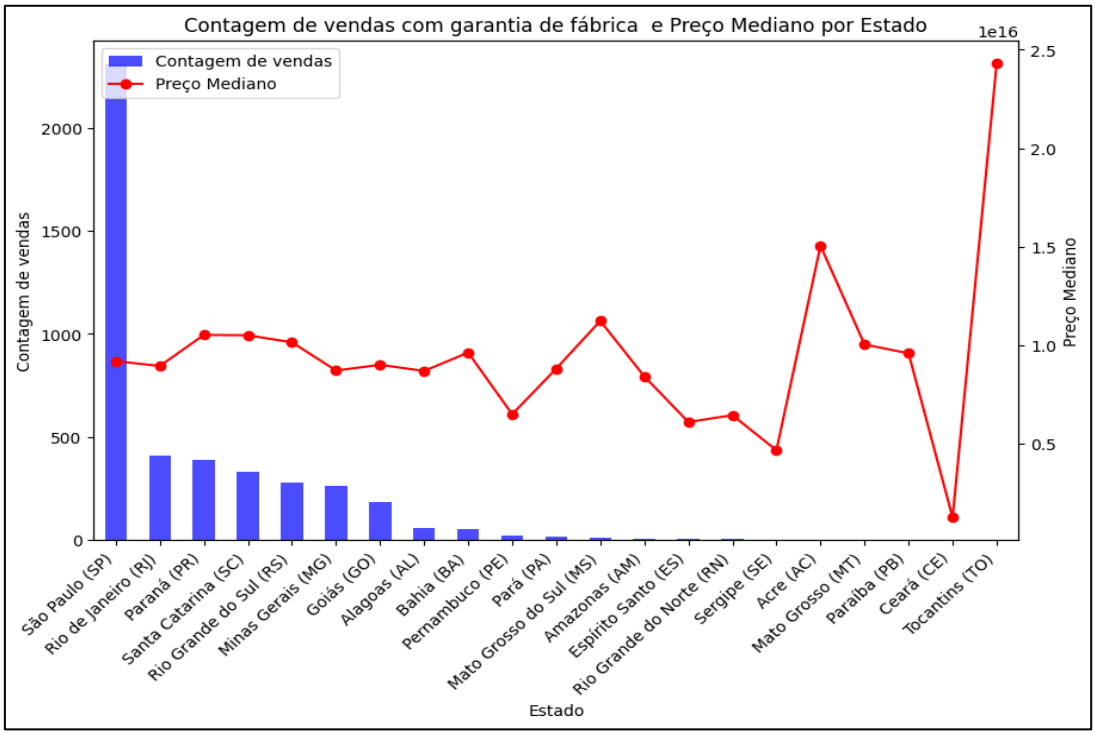
Como marca popular iremos considerar: Volkswagen, Fiat, Chevrolet,Citroen, Ford, Peugeot, Renault, Toyota, Honda, Hyunday. Como mostrado no gráfico abaixo, São Paulo, apesar de não possuir um preço mediano de venda elevado, é o estado com maior número de vendas, com larga frente em relação aos demais estados.

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



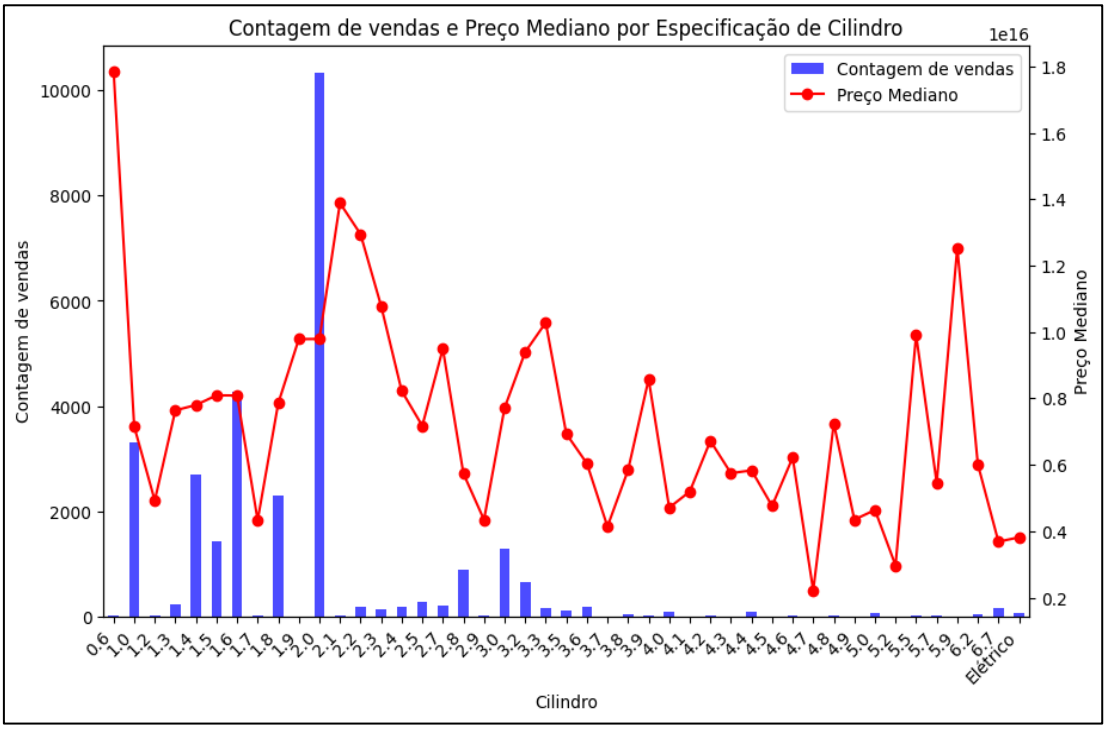
A contar pela disponibilidade de venda já registrada e pelo preço mediano de venda, Minas Gerais é o melhor estado para se comprar uma picape com transmissão automática. Podemos, ainda, mencionar os estados de Alagoas, Mato Grosso e Rio Grande do Norte, que apesar do pouco número de vendas registradas, possuem um preço mediano de venda muito abaixo dos demais estados.

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



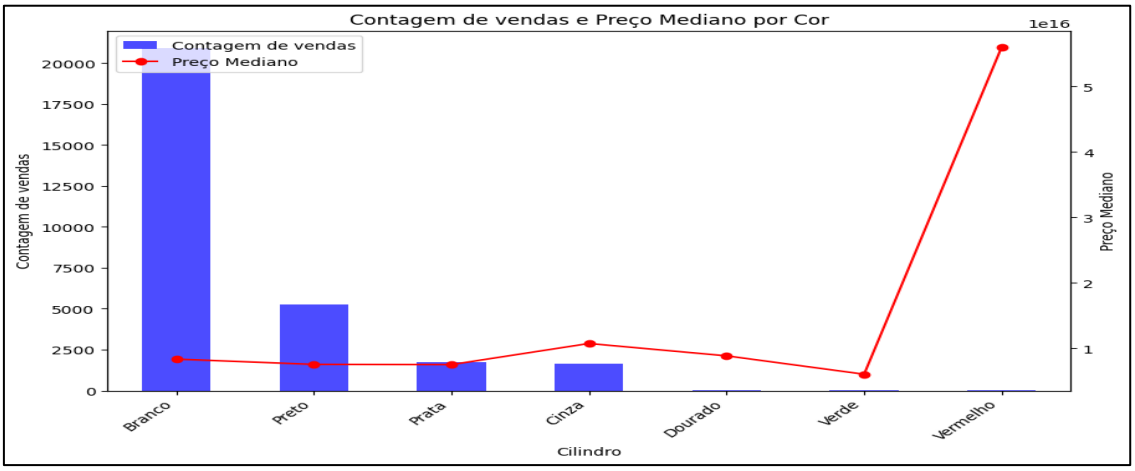
Apesar de uma quantidade inferior de vendas registradas em relação aos estados de maiores vendas, Pernambuco é o estado com o menor preço mediano de vendas com uma disponibilidade aceitável. Podemos mencionar o estado do Rio de Janeiro, como o estado com menor preço mediano dentre as localidades com mais vendas registradas.

4. Quais especificações de cilindrada são mais vendidas? Há variações no preço mediano entre eles?



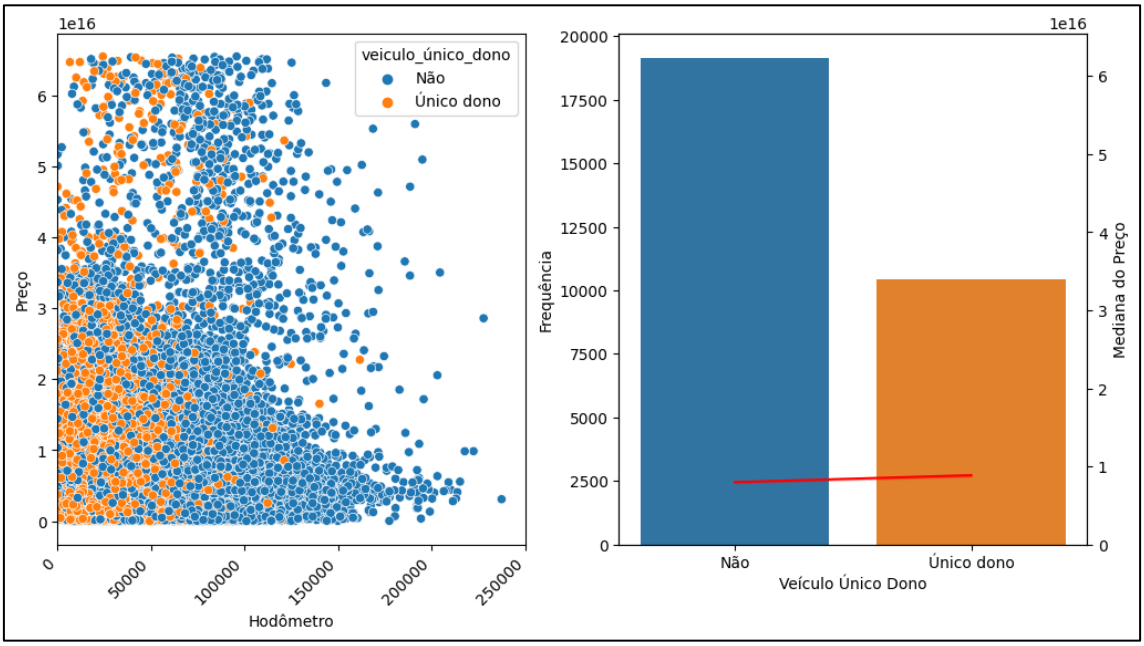
As especificações de cilindrada mais vendidas são 2.0, 1.6 , 1.0, 1.4 e 1.8, que juntos representam cerca de 77% das vendas. O preço mediano entre eles varia pouco, estando somente o 2.0 com um valor mais alto. (As especificações das cilindradas foram obtidas através da descrição das versões por meio de código na linguagem python)

5. *Quais cores são mais vendidas? Há variações no preço mediano entre elas?*



As cores mais vendidas são branco, preto, prata e cinza, representando cerca de 99% das vendas. Elas possuem pouca variação de preço mediano.

6. Carros de dono único tendem a ser muito mais valorizados? Isso sofre variações com os valores do hodômetro?



Segundo os gráficos, carros de único dono não são tão mais valorizados em relação aos demais. Possuem maior vantagem de vendas em faixas de alto preço com baixo hodômetro, o que garante essa leve superioridade em relação ao preço mediano, justamente por serem essencialmente carros com menor rodagem.

3. Modelagem Preditiva

Para predição do preço de vendas dos carros é necessário incluir **informações históricas** sobre preços passados, bem como **outros atributos relevantes**, como características do produto, dados econômicos ou informações do mercado. Com as informações disponíveis temos um caso de modelo **de regressão**.

Os modelos de regressão são técnicas estatísticas usadas para prever ou estimar um valor numérico contínuo com base em variáveis independentes. Essa técnica é um caso de modelo supervisionado de machine learning, do qual treinamos o comportamento do nosso modelo a partir de uma variável *target*, quem em nosso caso é o preço.

Algumas transformações nos dados foram feitas para possibilitar a análise e a performance do modelo. Dentre essas, podemos destacar:

- **Preenchimento de valores faltantes**. No caso, por exemplo, da variável 'num_fotos', os valores nulos foram substituídos pela moda da distribuição dos dados para essa variável, que no caso foi 8

- **Conversão de Variáveis**. Alterar o tipo das variáveis é importante para manipulação e análise dessas.
- **Feature Engineering**: Foram criadas algumas colunas derivadas de outras para poder melhor explicar a relação dos dados. Por exemplo, criamos o 'market_share" dos cilindros a partir da versão dos carros.
- **One-Hot-Encoder:** As variáveis categóricas foram transformadas em variáveis *dummies* para melhor ajuste nos algoritmos de predição.
- **Transformação Box-Cox:** A variável preço sofreu uma transformação para que tenha mais aderência a normalidade e assim permitir um melhor desempenho do modelo.
- **Remoção de Outliers:** As variáveis numéricas sofreram um tratamento para que valores acima (ou abaixo) de 3 desvios padrões fossem convertidos para esses valores mínimos.
- **Padronização dos dados:** Por terem escalas diferentes, as variáveis numéricas sofreram um transformação em que a média e o desvio padrão de suas distribuições fossem 0 e 1, respectivamente.

Os modelos baseados em árvores de decisão foram escolhidos devido a característica multinomial de algumas variáveis. No caso, o modelo de melhor performance foi o **Random Forest**.

O Random Forest é um algoritmo de aprendizado de máquina que combina múltiplas árvores de decisão para realizar previsões ou classificações. Possui como vantagens **a precisão**, pois produz resultados precisos e confiáveis, sendo capaz de lidar com dados de alta dimensionalidade; **resistência a *overfitting*,** pois as árvores individuais são treinadas em diferentes subconjuntos; **não sensível a outliers e é capaz de lidar com dados faltantes.**

Como desvantagens podemos citar a **dificuldade de interpretação** dos resultados, velocidade **de processamento**, espaço **de armazenamento** e sensibilidade a **dados desbalanceados.**

A métrica de performance foi, principalmente, o **RMSE** (Root Mean Square Error), ou Erro Quadrático Médio, é uma métrica comumente utilizada para avaliar o desempenho de modelos de regressão ou previsão. Ele mede a diferença entre os valores reais e os valores previstos pelo modelo, apresentando a média das diferenças ao quadrado. Ele é uma boa métrica por estar na mesma escala dos valores da variável terget.