DESAFIO LIGHTHOUSE:

Relatório de Entrega

Data by Nick Youngson CC BY-SA 3.0 Alpha Stock

Visão geral do relatório

Este relatório foi criado para o desafio do Lighthouse, promovido pela Indicium, e apresenta as principais estatísticas dos ***datasets cars_train*** disponibilizados pelo demandante. Objetiva apoiar a compreensão do cenário e de potenciais análises, bem como disponibilizar EDA que solucione as questões propostas pelo desafio.

Conteúdo

EDA

Visão geral do banco de dados tratado

Informações do trabalho

division	metrics	value
dataset	dataset	d_treino
dataset	dataset type	data.frame
dataset	target	not defied
job	samples	29.584 / 29.584 (100%)
job	created	
job	created by	Elig Cassiane Arse da Silva

Tabela 1: Informações do trabalho

Para realização desta análise desconsiderou-se a variável ID, pois uma vez que ela trata exclusivamente da identificação de cada observação do banco de dados não fornece informações úteis para análise estatística, pois não contém informações sobre as características ou atributos dos dados. Portanto, não contribui para a análise exploratória e não influencia nos padrões ou relações encontradas entre as outras variáveis. Seguindo o mesmo pressuposto, as variáveis, num_portas, entrega_delivery, garantia_de_fábrica, revisoes_dentro_agenda, veiculo_licenciado, ipva_pago, elegivel_revisao, revisoes_concessionaria, veiculo_único_dono, dono_aceita_troca, -veiculo_alienado e veiculo_alienado foi excluída por ser composta exclusivamente de NAs ou então de dado categórico ou lógico que só continha um nível.

Estrutura dos dados

division	metrics	value
size	observations	29,584
size	variables	17
size	values	502,928
size	memory size (MB)	3
duplicated	duplicate observation	0
missing	complete observation	29,584
missing	missing observation	0
missing	missing variables	0
missing	missing values	0

division	metrics	value
data type	numerics	4
data type	integers	1
data type	factors/ordered	12
data type	characters	0
data type	Dates	0
data type	POSIXcts	0
data type	others	0

Tabela 2: Estrutura e tipos de dados

A tabela 2 apresenta um resumo sobre a estrutura da base de dados, incluindo informações sobre o tamanho do

conjunto de dados, a existência de dados duplicados e valores ausentes, bem como os tipos de dados presentes na base. Essas informações são úteis para entender a qualidade dos dados, a quantidade de informações disponíveis, a complexidade dos tipos de dados e outras características relevantes para análises e modelagem dos dados.

Análise Univariada

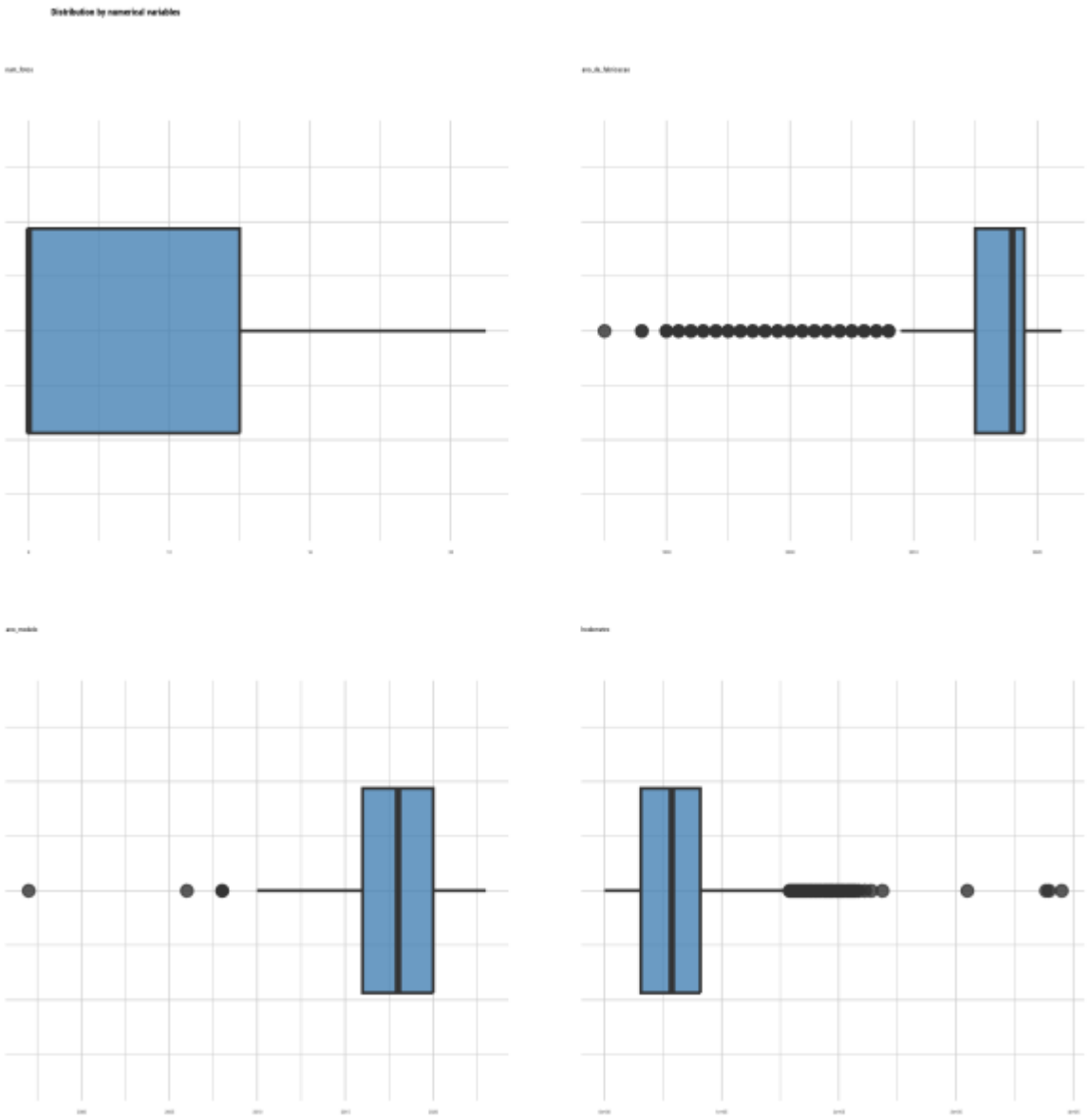
Estatísticas descritivas

Variáveis numéricas

variables	missing	mean	sd	min	Q1	median	Q3	max
num_fotos	0	10.31	3.48	8.00	8.00	8.0	14.0	21
ano_de_fabricacao	0	2,016.76	4.06	1,985.00	2,015.00	2,018.0	2,019.0	2,022
ano_modelo	0	2,017.81	2.67	1,997.00	2,016.00	2,018.0	2,020.0	2,023
hodometro	0	58,430.59	32,561.77	100.00	31,214.00	57,434.0	81,953.5	390,065
preco	0	133,023.88	81,662.87	9,869.95	76,571.77	114,355.8	163,679.6	1,359,813

Tabela 3: Estatística descritiva de variáveis numéricas

Conforme a tabela, os carros anunciados tem, em média, 10 fotos. Os números de fotos variam em torno da média em cerca de 3.48 unidades. Quartis (Q1, median e Q3) de 8, 8 e 14, respectivamente: A maioria dos carros tem entre 8 e 14 fotos em seus anúncios.



variables	data types	distinct	skewness	kurtosis	zero	negative	outlier
num_fotos	numeric	14	1.01	-0.63	0	0	0
ano_de_fabricacao	integer	35	-2.49	9.19	0	0	1,147
ano_modelo	numeric	17	-0.50	-0.52	0	0	6
hodometro	numeric	26,004	0.59	1.20	0	0	158

Tabela 4: Estatística descritiva de variáveis numérica

Em média, os carros são fabricados em torno do ano de 2016. O carro mais antigo fabricado é 1985 e o mais novo é 2022. Em média, os carros têm um ano de modelo de cerca de 2017. Os modelos variam em torno da média em cerca de 2.67 anos.

Em média, os carros têm cerca de 58,430.59 quilômetros no rodados.A quilometragem varia em torno da média em cerca de 32,561.77 km. Observando os quartis (Q1, median e Q3) de 31,214, 57,434 e 81,953.5, respectivamente: A maioria dos carros tem uma quilometragem entre 31,214 e 81,953.5 km.

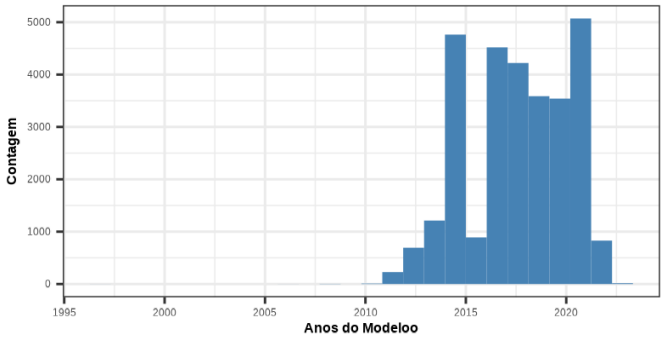
DESAFIO LIGHTHOUSE:

variables	data types	distinct	skewness	kurtosis	zero	negative	outlier
preco	numeric	29,565	2.15	10.97	0	0	1,448

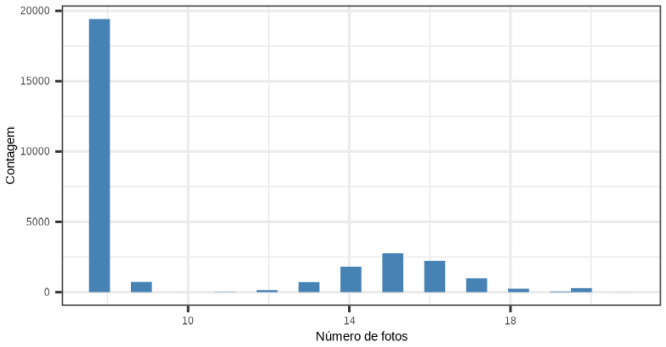
Em média, os carros têm um preço de aproximadamente 133,023.88.

Os preços variam em torno da média em cerca de 81,662.87. A maioria dos carros custam entre 76,571.77 e 163,679.6.

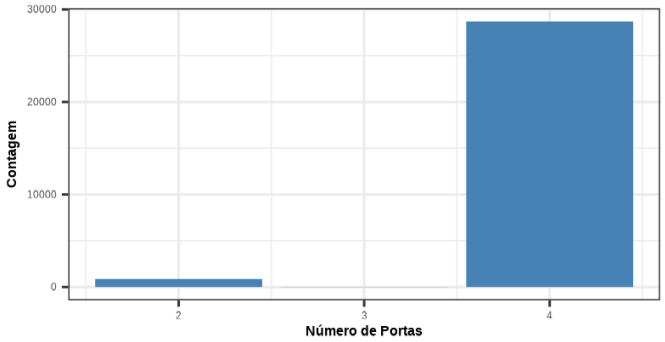
Distribuição dos Anos do Modelo dos Veículos



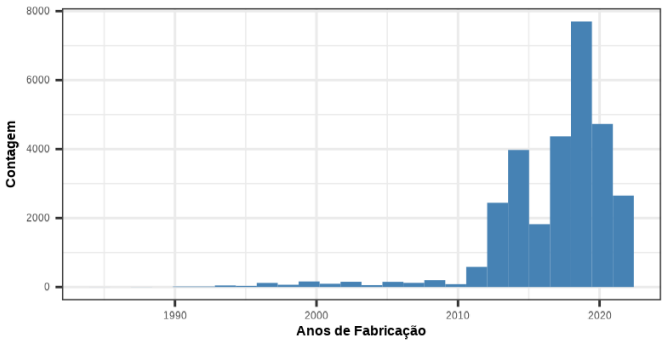
Distribuição do Número de Fotos por Anúncio



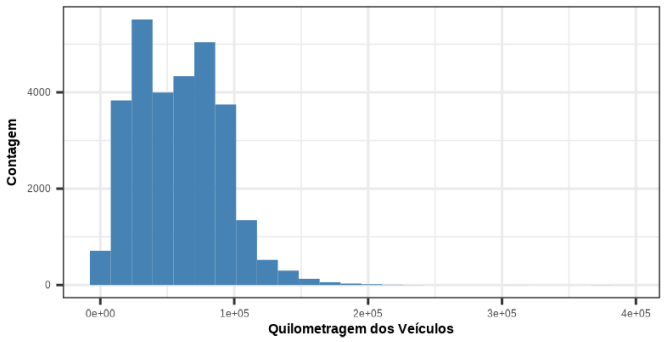
Distribuição do Número de Portas dos Veículos



Distribuição dos Anos de Fabricação dos Veículos

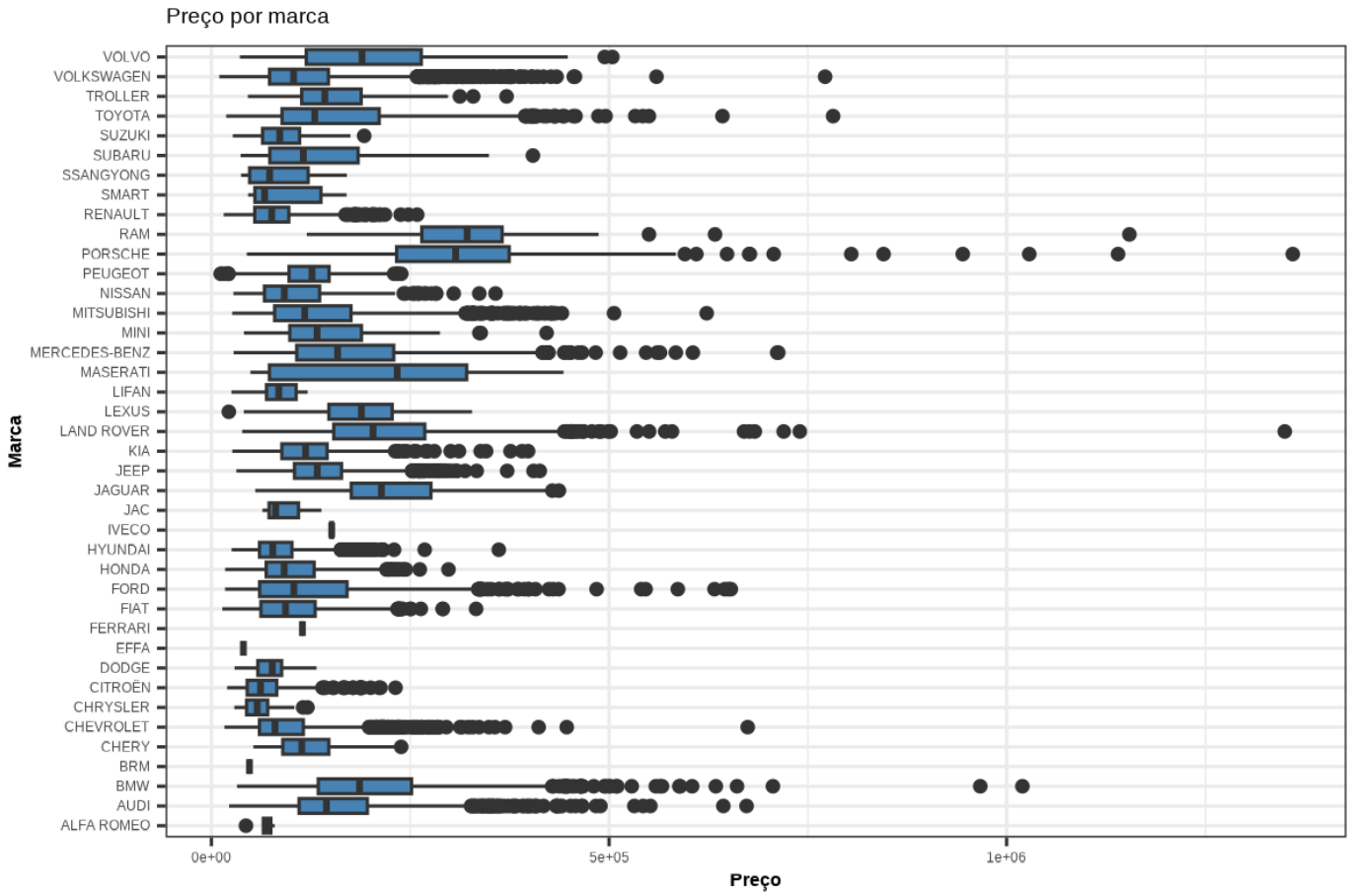


Distribuição da Quilometragem dos Veículos

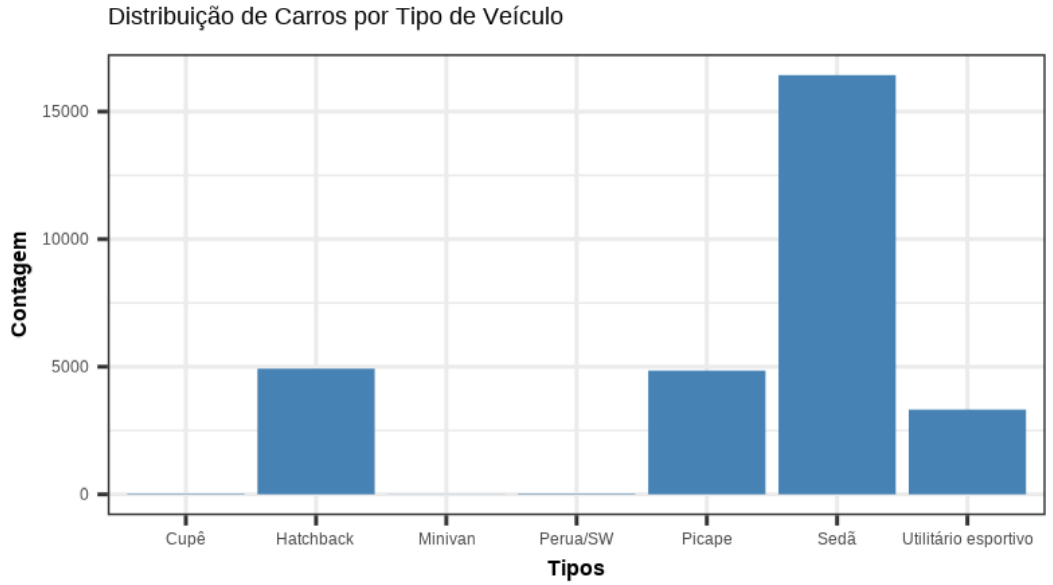


Variáveis categóticas

A seguir, veremos a distribuição das variáveis por preço, já que ele é a nossa variável preditiva.

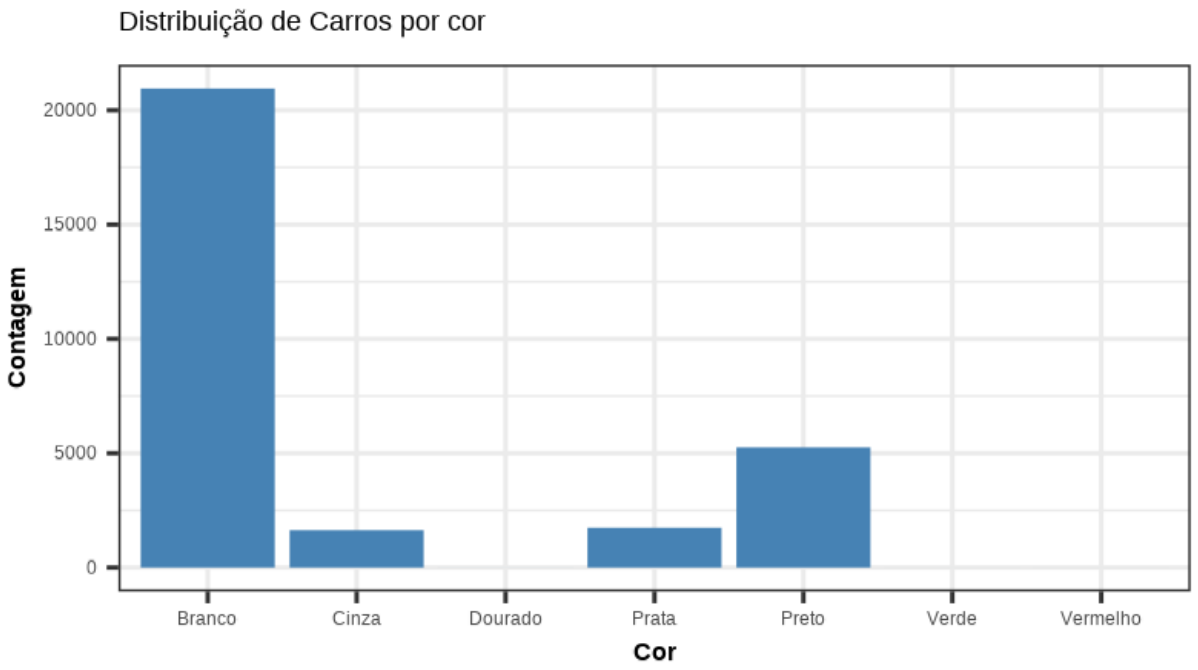


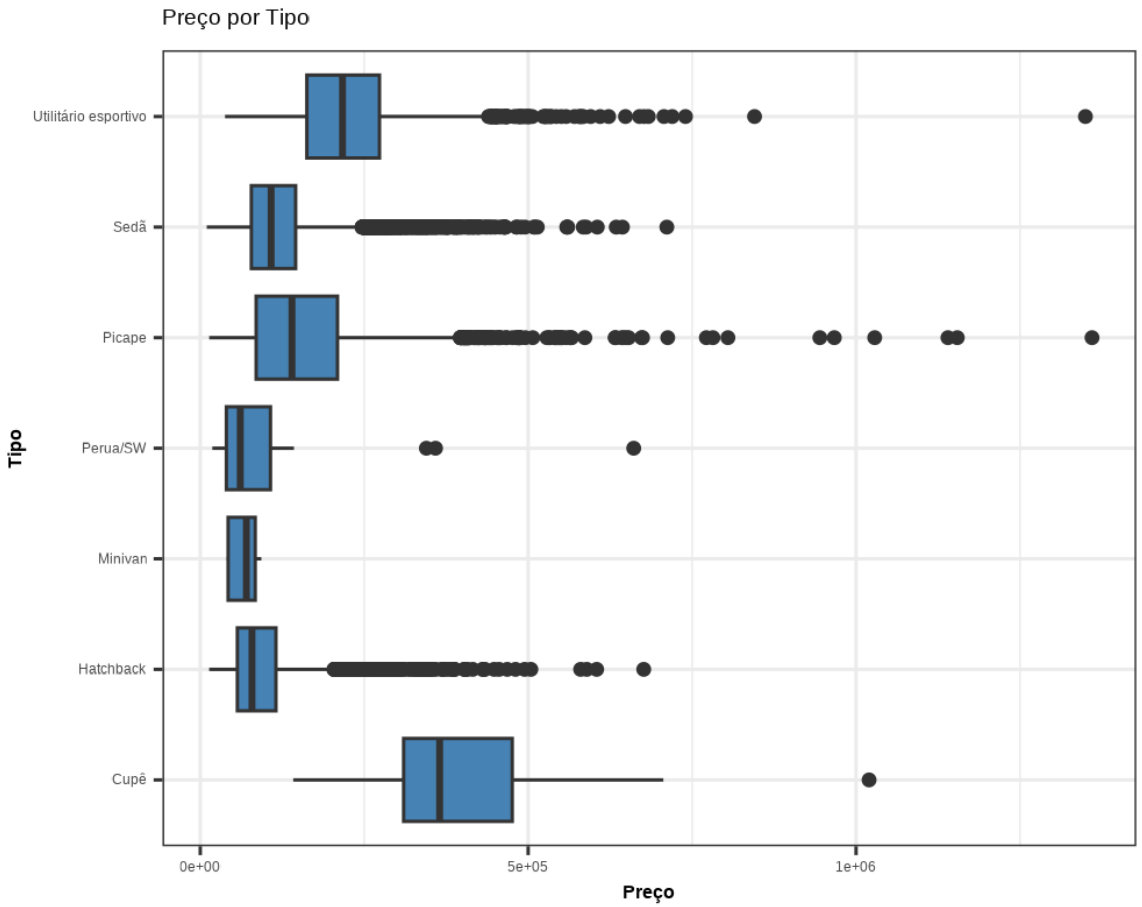
Existe uma grande variação no preço por marca, indicando que este pode ser um fator relevante no preço de um carro.

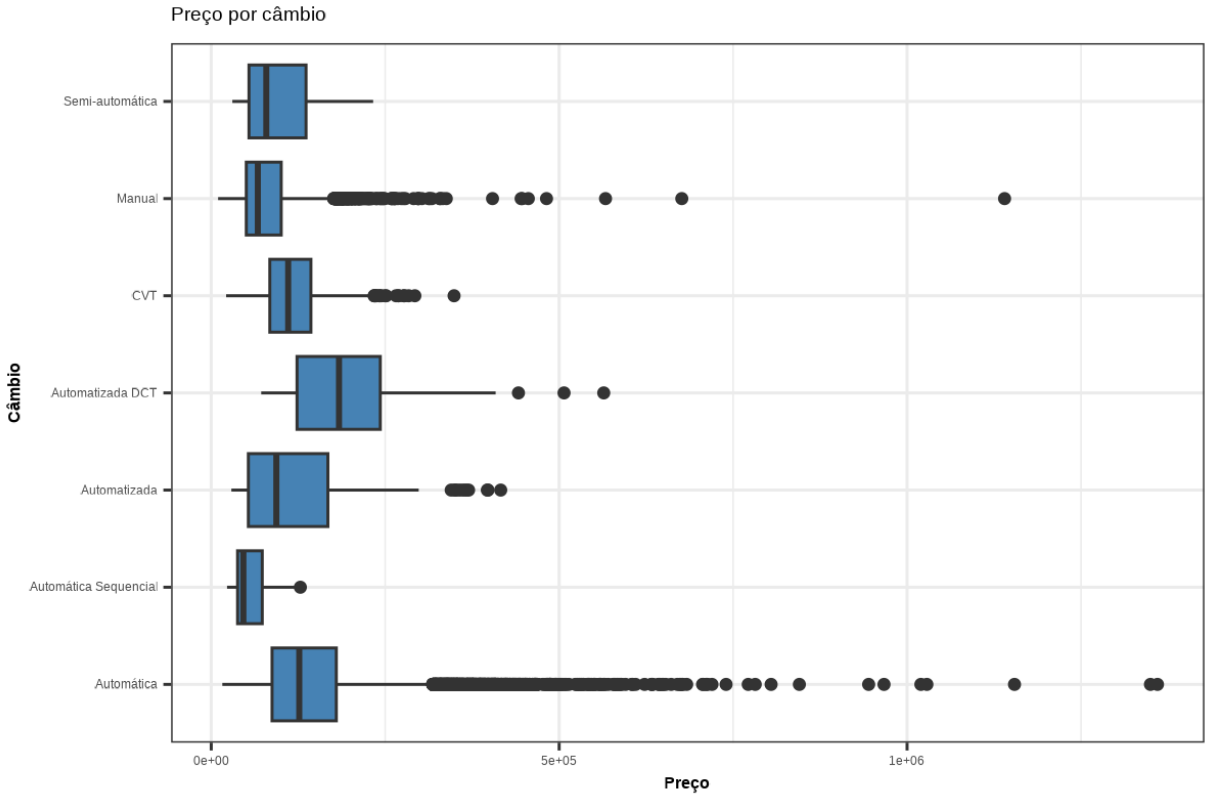


A maioria dos carros são Sedã

A frequência das cores dos carros usados disponíveis para venda pode ajudar a identificar as cores mais populares ou mais comuns entre os veículos.


Distribuição de Carros por cor
Contagem
20000
15000
10000
5000
0
Branco
Cinza
Dourado
Prata
Preto
Verde
Vermelho
Cor


Preço por Tipo
Tipo
Utilitário esportivo
Sedã
Picape
Perua/SW
Minivan
Hatchback
Cupê
0e+00
5e+05
1e+06
Preço



Análise Bivariada

Comparar Variáveis Numéricas

first variable	second variable	correlation coefficient
num_fotos	ano_de_fabricacao	0.02592
num_fotos	ano_modelo	0.02763
num_fotos	hodometro	0.03046
num_fotos	preco	-0.03151
ano_de_fabricacao	ano_modelo	0.86031
ano_de_fabricacao	hodometro	-0.72829
ano_de_fabricacao	preco	0.23917
ano_modelo	hodometro	-0.79006
ano_modelo	preco	0.21485
hodometro	preco	-0.35922

Tabela 6: Coeficiente de correlação

Conforme a tabela 6, não há nenhum indício de correlação linear positiva entre as variáveis.

Análise multivariada

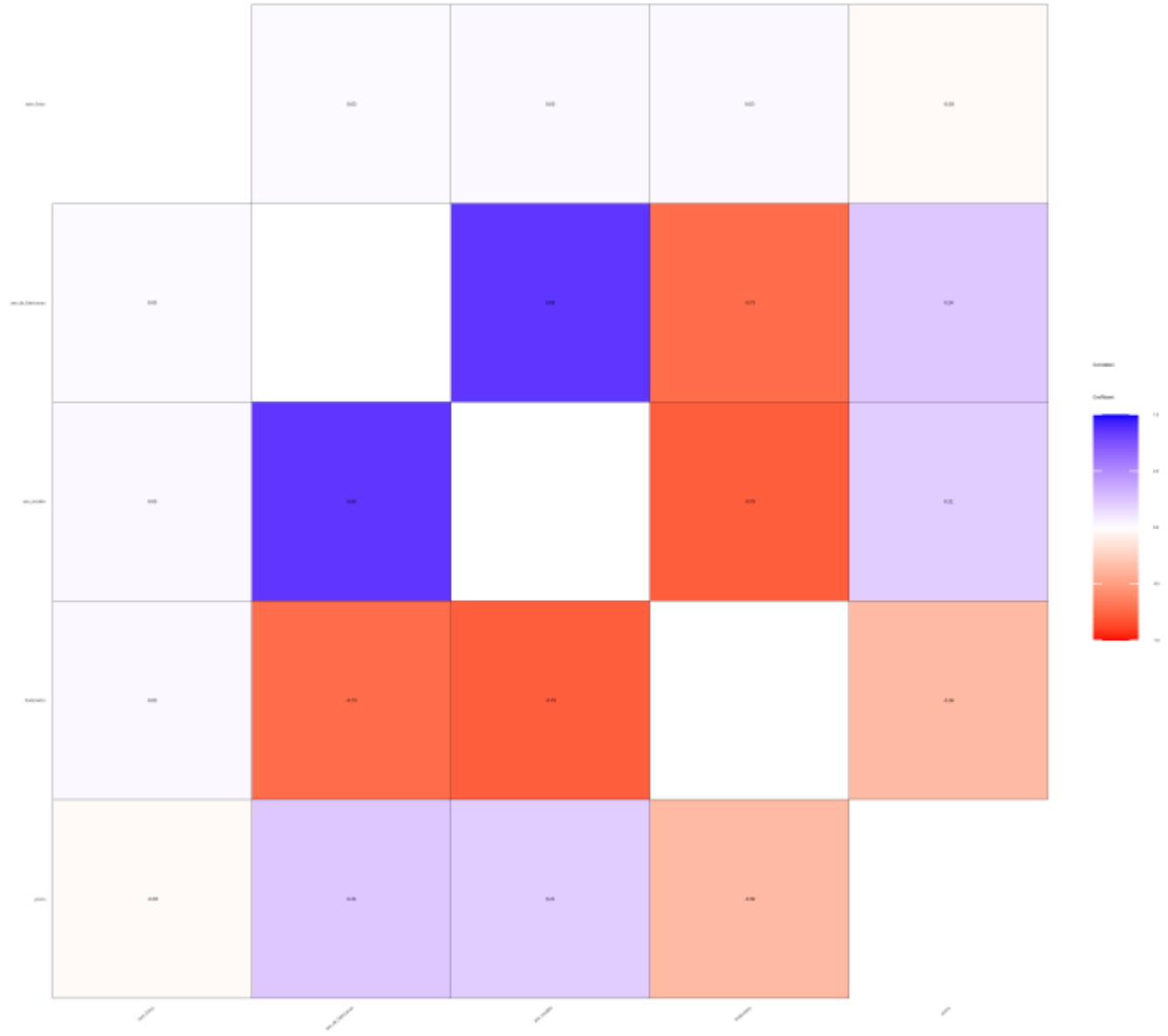
first variable	second variable	correlation coefficient
num_fotos	ano_de_fabricacao	0.02592
num_fotos	ano_modelo	0.02763
num_fotos	hodometro	0.03046
num_fotos	preco	-0.03151
ano_de_fabricacao	ano_modelo	0.86031
ano_de_fabricacao	hodometro	-0.72829
ano_de_fabricacao	preco	0.23917
ano_modelo	hodometro	-0.79006
ano_modelo	preco	0.21485
hodometro	preco	-0.35922

Análise de correlação

Matriz de Coeficientes de Correlação

	second variable				
first variable	num_fotos	ano_de_fabricacao	ano_modelo	hodometro	preco
num_fotos	NA	0.026	0.028	0.030	-0.032
ano_de_fabricacao	0.026	NA	0.860	-0.728	0.239
ano_modelo	0.028	0.860	NA	-0.790	0.215
hodometro	0.030	-0.728	-0.790	NA	-0.359
preco	-0.032	0.239	0.215	-0.359	NA

Gráfico de Correlação



Estas informações nos ajudaram a conhecer melhor o banco, entender quais variáveis podem ser eliminadas do modelo de regressão a fim de melhorar a execução

Hipótese 1

Qual o estado com menor média de preço?
Fiz uma média entre os preços de todas as regiões e o Piauí

Hipótese 2

Qual a cor de carro mais vendida?
Branco

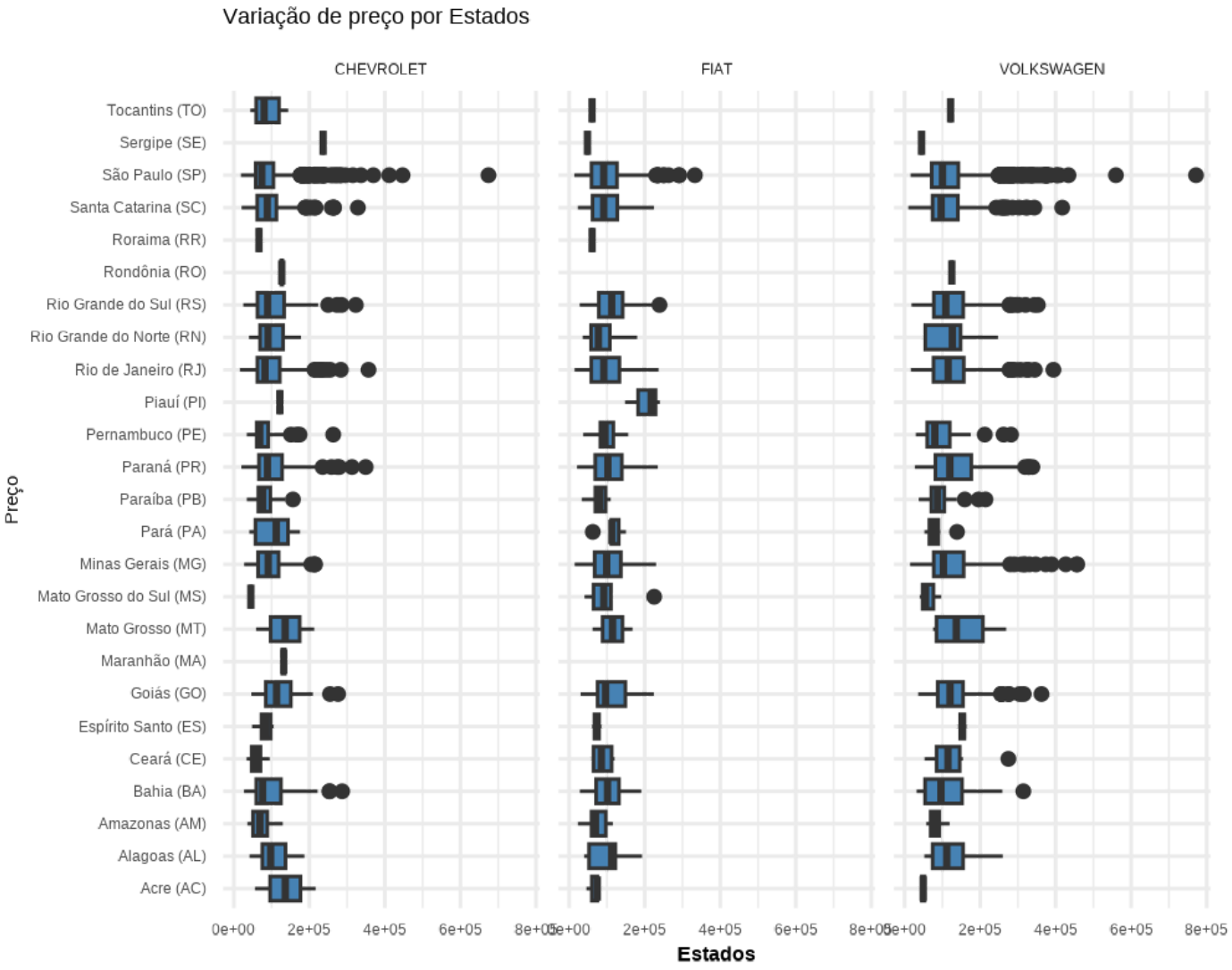
Hipótese 3

Qual estado tem maior variação da média de preço?
Mato Grosso

Respondendo às perguntas de negócio

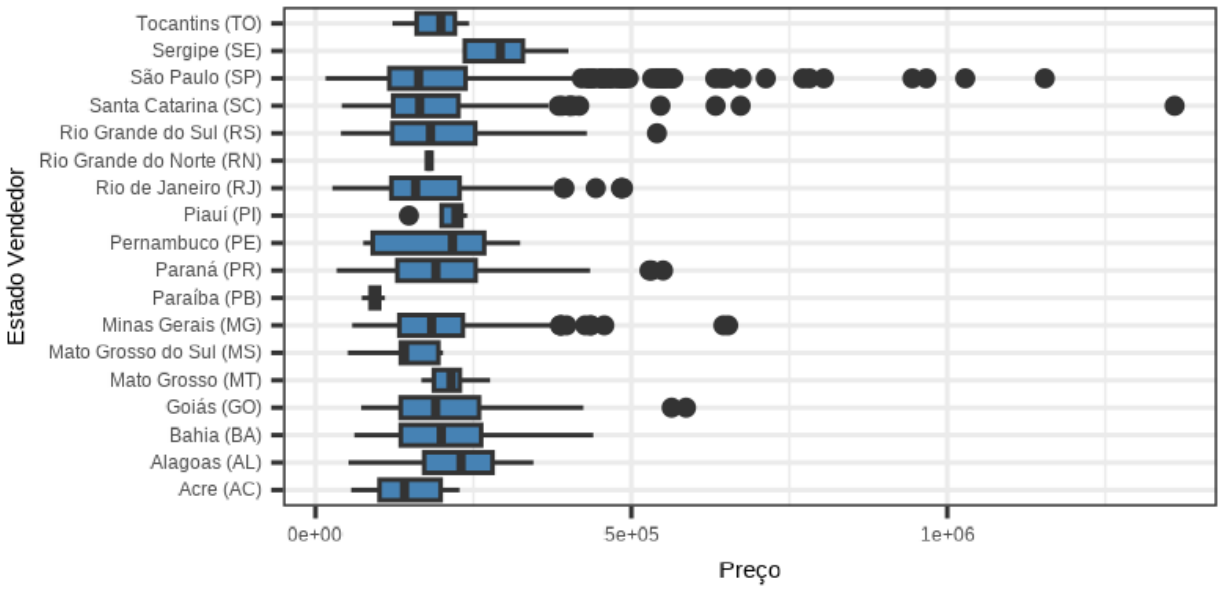
a) Qual o melhor estado cadastrado na base de dados para se vender um carro de marca popular e por quê?

O preço médio de venda de carros populares das marcas Chevrolet, Fiat e Volkswagen, segundo os dados da base, é de 106430.
Avaliando a variação de preços de carros populares, identifica-se que o estado do Mato Grosso do Sul é o melhor para vender carros neste mercado, isso porque, para as três marcas de carros populares selecionadas na análise, esse estado apresenta pouca variação de preço. O que é positivo para a sustentabilidade do negócio a longo prazo, tendo em vista que gera previsibilidade para os vendedores, reduz o risco de desvalorização exacerbada em períodos curtos de tempo e também reduz a necessidade de ajuste frequente do preço para acompanhar o mercado. Ver gráfico abaixo

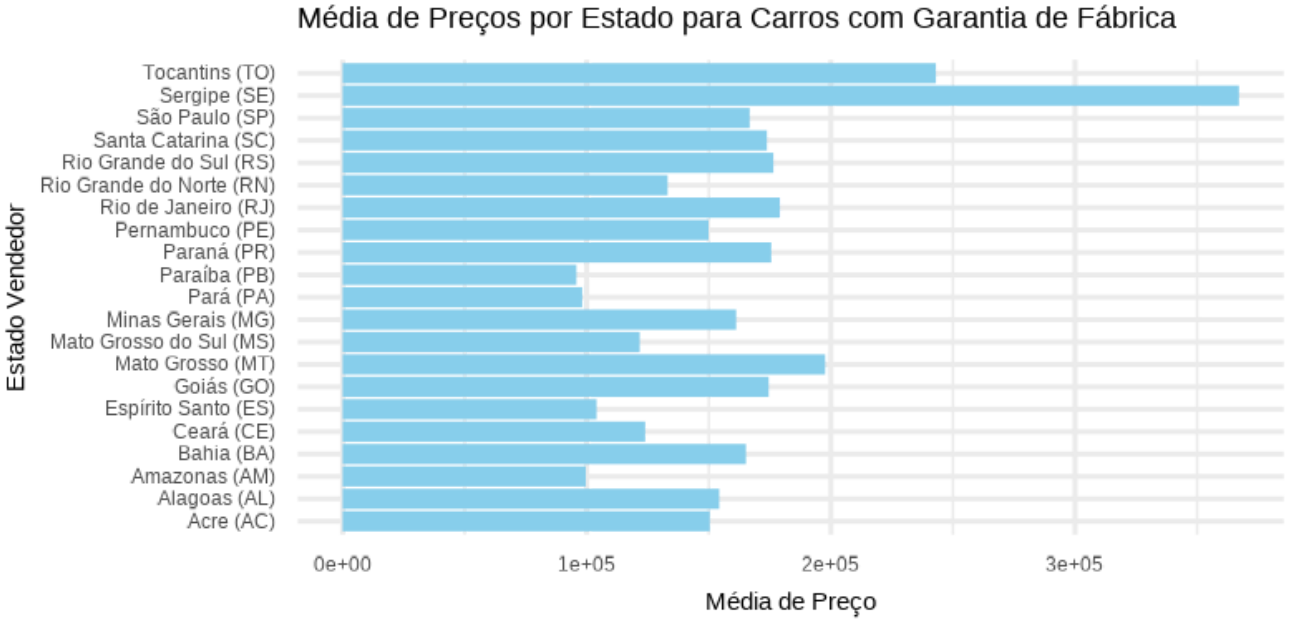


b) Qual o melhor estado para se comprar uma picape com transmissão automática e por quê?

A média de preço de picapes automáticas é de 190559.3, considerando a variação de preços, o melhor estado para comprar uma picape é Paraíba, pois conta com menor valor, e baixíssima variação, com os preços concentrados em torno da média, conforme gráfico abaixo.



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



O estado com menor média de preços para comprar um carro com garantia de fábrica é a Paraíba.

Previsão do preço

1. Explique como você faria a previsão do **preço** a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê? Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?

   Como a variável a ser predita é contínua, trata-se de um problema de regressão. A variável alvo é o preço, e as demais variáveis foram utilizadas como preditoras para fazer a previsão.

Foram excluídas as seguintes variáveis: id (por ser coluna única), num_fotos, num_portas, cidade_vendedor, versao, cor, entrega_delivery, elegivel_revisao, dono_aceita_troca e garantia_de_fábrica.

As variáveis que restaram foram tratadas para a lidar com valores faltantes (NA) substituindo-os pela mediana para as variáveis numéricas e pela moda para as variáveis categóricas. Em seguida, as variáveis categóricas foram transformadas em variáveis dummy (hot encoding).

O modelo utilizado é Random Forest, que é um algoritmo de aprendizado de máquina baseado em árvores de decisão. Escolhi esse modelo pela sua capacidade de lidar com um grande número de variáveis preditoras, sua resistência ao overfitting e boa performance geralmente obtida em problemas de regressão. Contudo, é um modelo mais difícil de interpretar em comparação com uma única árvore de decisão e pode ser mais lento em grandes conjuntos de dados, como é o caso da nossa base.

Para avaliar a performance do modelo, a métrica escolhida foi o RMSE (Root Mean Squared Error). O RMSE é bastante utilizado em problemas de regressão porque fornece uma medida da dispersão dos erros do modelo.