Desafio Cientista de Dados

Descrição do Problema

O problema se refere a um cliente que o *core business* é compra e venda de veículos usados e está com dificuldades na área de revenda dos automóveis usados em seu catálogo. Dessa forma, o 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 os mais próximos dos valores de mercado.

Descrição dos Dados

São fornecidos dois *datasets* com objetivo de prever a coluna "preco" a partir dos dados:

1. Um *dataset* para treinamento chamado *cars_train* composto por 29584 linhas, 28 colunas de informação (*features)* e a variável a ser prevista (“preco”) (Disponível em: *./dataset/ cars_train.csv)*.
2. Um segundo *dataset* para teste chamado de *cars_test* composto por 9862 linhas e 28 colunas, sendo que este *dataset* não possui a coluna “preco”. (Disponível em: *./dataset/cars_test.csv)*.

Arquivos

Lista de arquivos utilizados neste projeto:

- Dataset de treinamento: ./dataset/cars_train.csv
- Dataset de teste: ./dataset/cars_test.csv
- Notebook EDA: ./notebooks/EDA.ipynb
- Notebook modelagem: ./notebooks/modelagem_regressao.ipynb

- Modelo Treinado: ./models/model_regressao_linear.pkl
- Resultado do modelo: ./results/predicted.csv
- Códigos de modelagem: ./src/
- Link GitHub: *https://github.com/andressagomes26/marketplace_lh_cd.git*

Como instalar e executar o projeto

Para utilizar este projeto deve-se clonar o repositório do *github* e executar o seguinte comando dentro da pasta do projeto:

pip install -r requirements.txt

Para treinar o modelo de regressão linear pode-se executar o comando abaixo:

python src/train.py

Caso seja necessário alterar o caminho da pasta do dataset ou o caminho em que o modelo será salvo, pode-se enviar estes caminhos como argumento. Sendo eles:

-- path_dataset_train <caminho_dataset_treinamento>

-- path_dataset_teste <caminho_dataset_teste>

--path_model <caminho_modelo>

Para testar o modelo de regressão linear e salvar o arquivo 'predicted.csv' com o resultado do modelo pode-se executar o comando abaixo:

python src/test.py

Caso seja necessário alterar o caminho da pasta do dataset, o caminho em que os modelos serão salvos, ou o caminho para salvar o arquivo predicted.csv, pode-se enviar estes caminhos como argumento. Sendo eles:

-- path_dataset_train <caminho_dataset_treinamento>

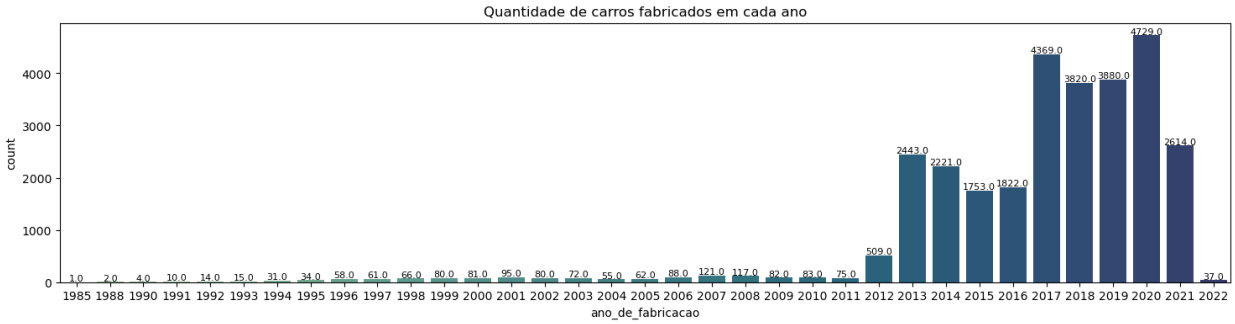
-- path_dataset_teste <caminho_dataset_teste>

--path_model <caminho_modelo>

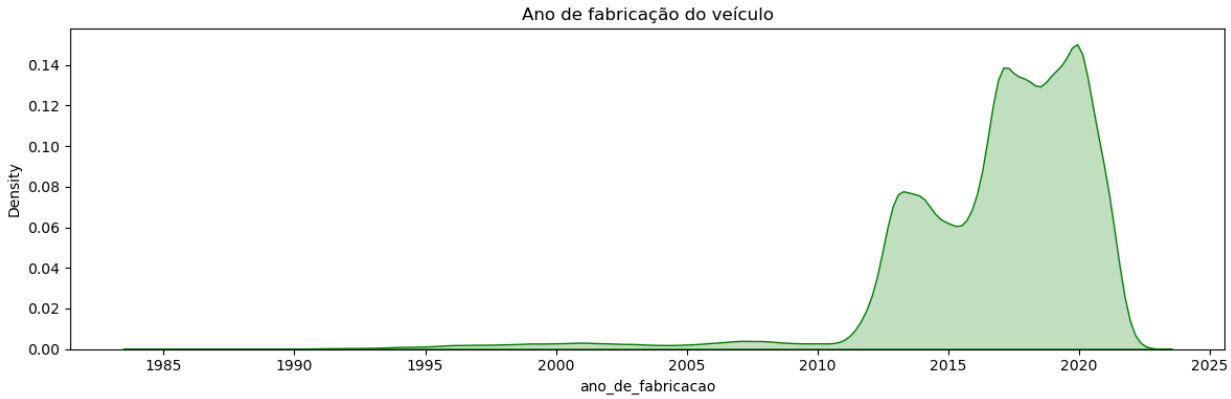
--path_predicted <caminho_predicted>

Análise Exploratória dos Dados (EDA)

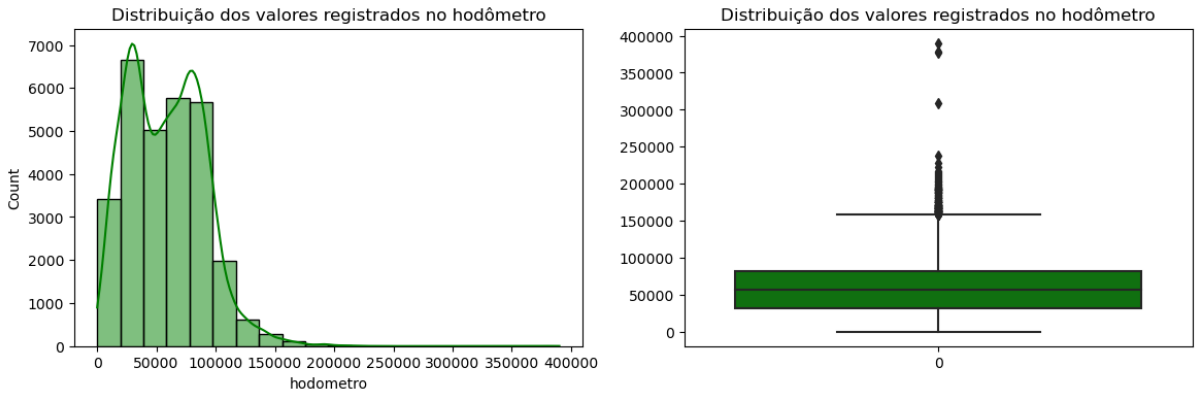
A priori, foi realizado uma análise das principais estatísticas da base de dados. O dataset de treinamento é composto por seis variáveis numéricas e 22 variáveis categóricas. Sobre a *feature* 'ano_de_fabricacao' é possível analisar a quantidade de veículos que foram fabricados em cada ano. O ano de 2020 foi o ano com a maior quantidade de carros fabricados, com um total de 4729 veículos.



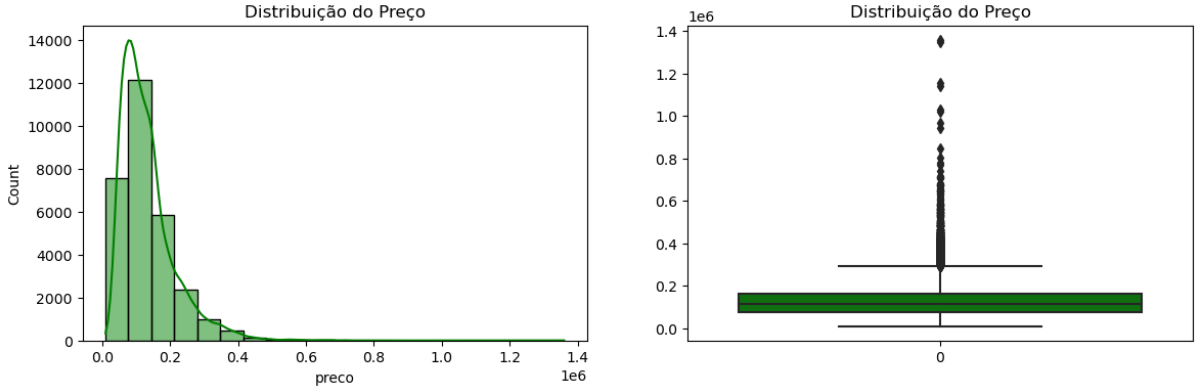
Além disso, é possível verificar que a concentração da quantidade de carros fabricados ocorre entre os anos de 2010 até 2021.



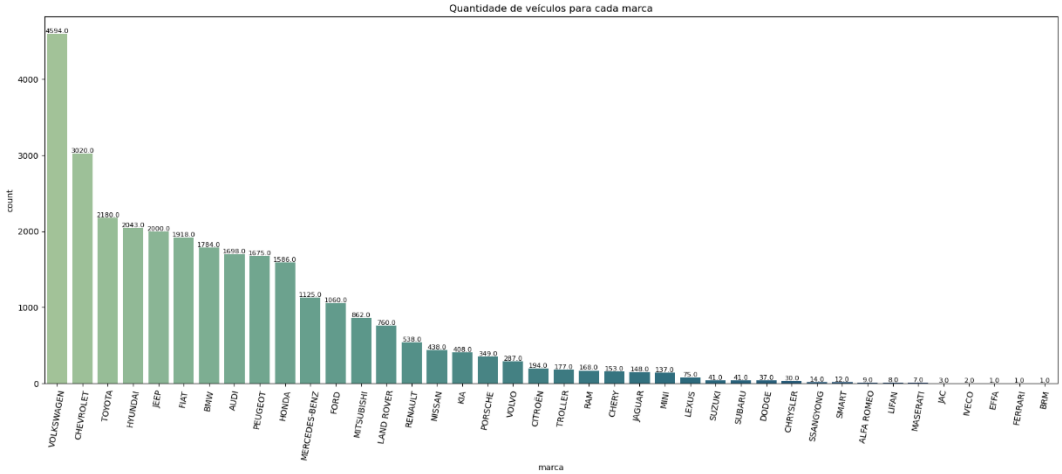
Sobre o atributo 'hodometro' é possível observar a distribuição de valores registrados no hodômetro dos veículos. A maioria dos veículos possuem valores no hodômetro na faixa de 0 até 150.000. Além disso, é possível notar que a mediana dos valores registrados no hodômetro é igual a 57.434. Vale ressaltar, que existem outliers, ou seja, alguns veículos possuem valores exorbitantes registrados no hodômetro, na faixa de 300.000 até aproximadamente 400.000.



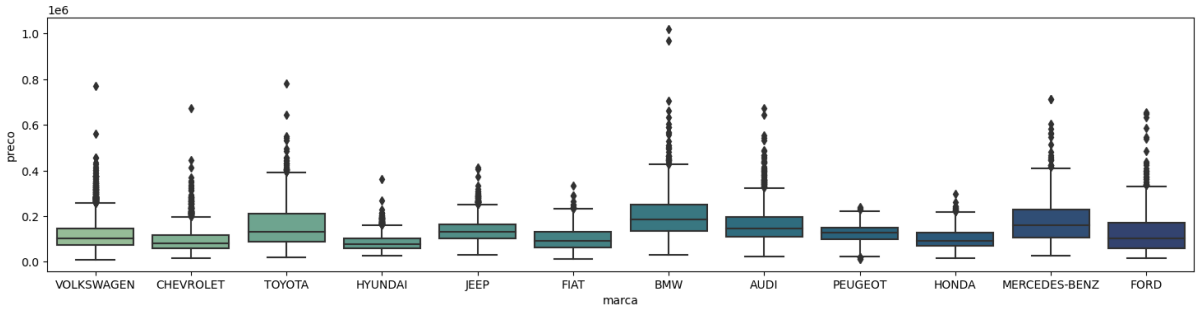
Em relação à *feature* 'preco' é possível verificar que a distribuição de preços dos veículos se concentra na faixa de 100.000 até 400.000. Existem veículos com preços mais elevados, atingindo aproximadamente valores até 1.4 le6.



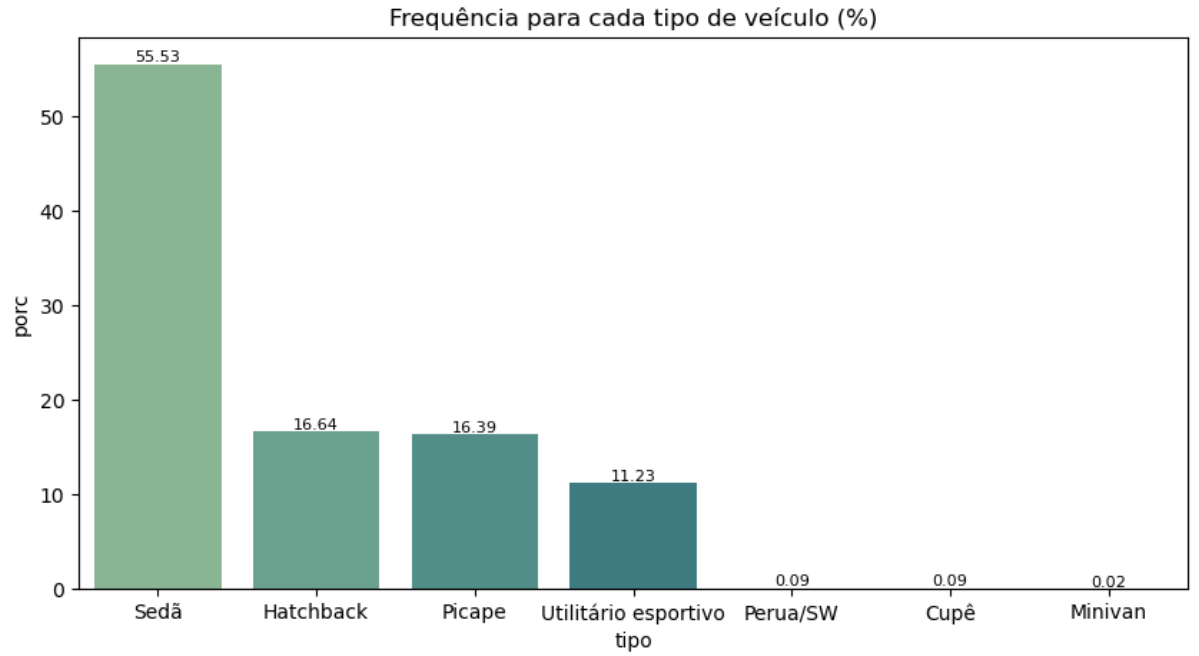
Ademais, é possível observar a quantidade de veículos por marcas. A maioria dos veículos são da Volkswagen (4594 veículos), seguida pelas marcas Chevrolet (3020 veículos) e Toyota (2180 veículos). É importante conhecer a quantidade de veículos por marcas, pois a marca pode ser um fator determinante para o preço do veículo.



Portanto, para analisar a distribuição de preços das marcas de veículos mais vendidas é válido analisar os *boxplots* para cada marca. Dessa forma, Volkswagen e Chevrolet, além de serem as marcas mais vendidas, são as marcas mais baratas. Já a marca BMW, sétima marca mais vendida, é a marca com a mediana de preços mais elevada.

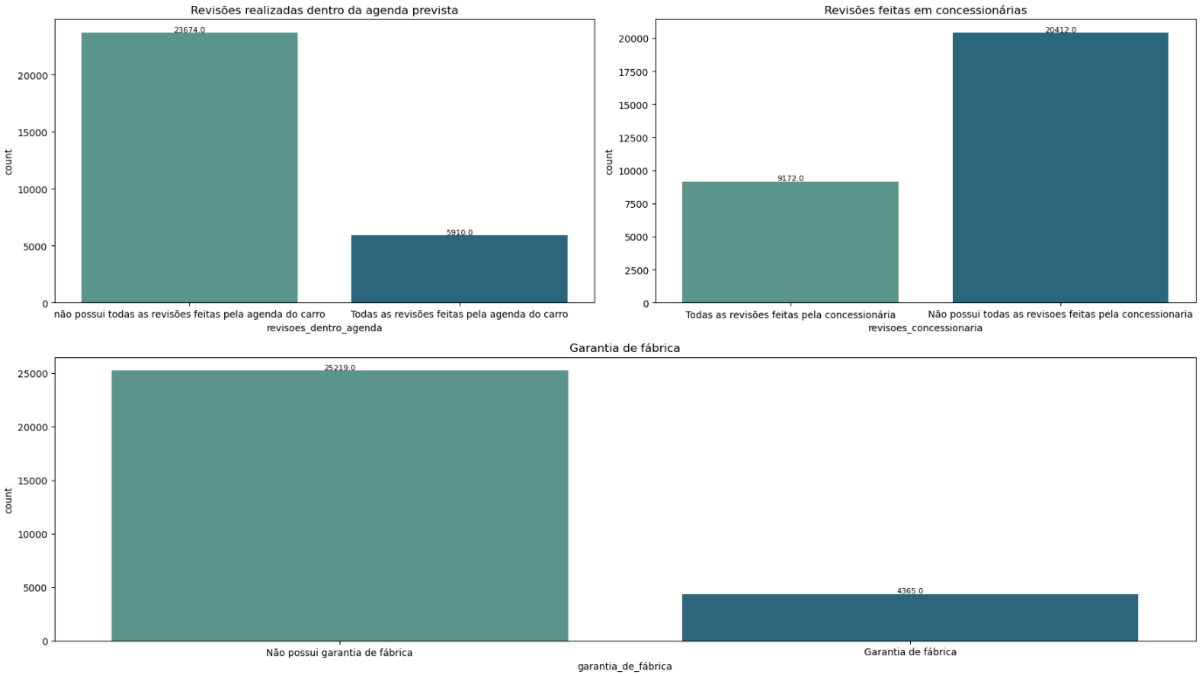


Além da marca, é possível observar como os veículos estão distribuídos de acordo com o tipo. Dessa forma, 55,53% dos veículos são do tipo Sedã, seguido pelo Hatcback com 16,64%. O tipo menos vendido é o Minivan com 0.02. Essas informações podem ser relevantes para o cliente para determinar qual tipo de veículo revender.

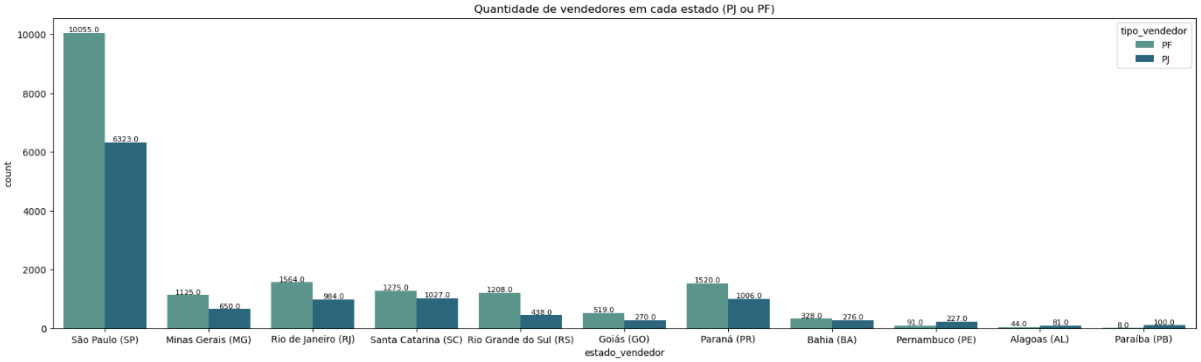


Além disso, outros fatores que podem ajudar na revenda dos veículos são as informações sobre as revisões e sobre a garantia. No gráfico a seguir, pode-se observar que a maioria dos veículos não possuem todas as revisões feitas dentro da agenda, não possui todas as revisões feitas pela concessionária, além de não possuir

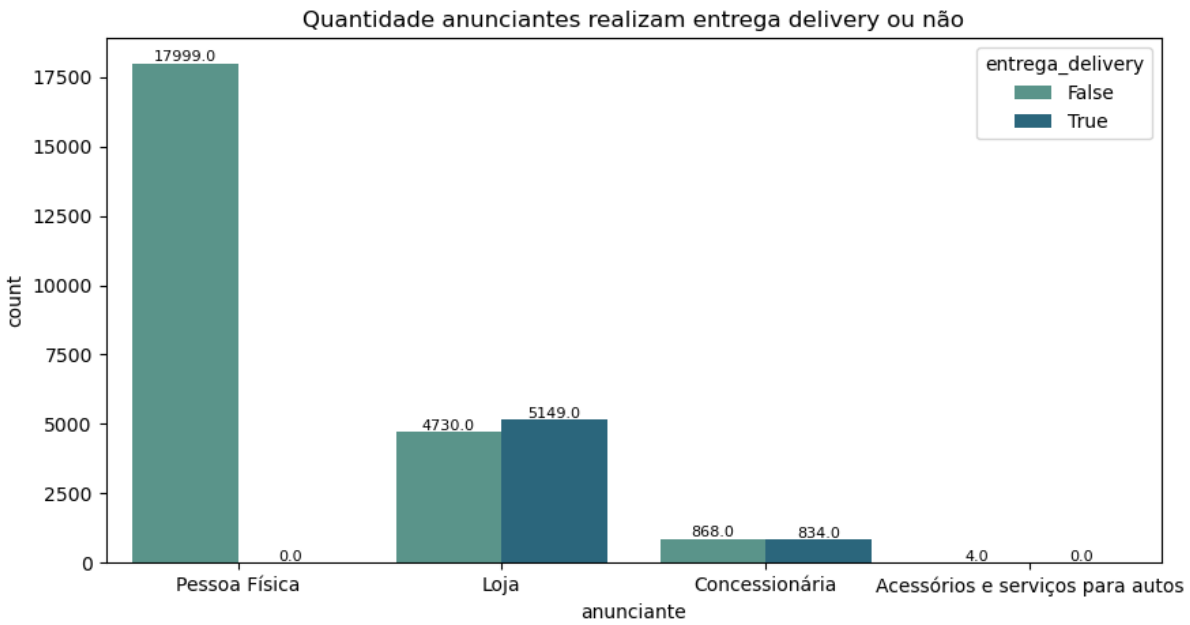
garantia de fábrica. Logo, tais fatores podem não ser essenciais no momento da revenda de veículos.



Além das características dos veículos, é possível analisar características do vendedor. Em quase todos os estados os vendedores de veículos são do tipo pessoa física (PF). Logo, pode-se concluir que não é necessário ser pessoa jurídica (PJ) para realizar a revenda de veículos.



Portanto, dos anunciantes que do tipo Pessoa física (PF) nenhum realiza entrega delivery. Esse tipo de entrega é realizado apenas por lojas ou por concessionárias.



Análise Exploratória dos Dados (EDA) - Hipóteses de negócio

Nessa etapa, foram criadas e respondidas três hipóteses de negócio:

1. **Hipótese de negócio 1:** Existe relação entre o valor registrado no hodômetro do veículo e o preço.
2. **Hipótese de negócio 2:** O tipo de veículo interfere na média de preço.
3. **Hipótese de negócio 3:** Existe associação entre as variáveis IPVA PAGO e Veículo licenciado.

Para responder a **Hipótese de Negócio 1** foi realizada a Análise Exploratória dos Dados e um teste de hipótese. Dessa forma, para entender se existe relação entre o valor registrado no hodômetro e o preço do veículo foi utilizado o Coeficiente de correlação de Pearson pois, deseja-se medir a relação entre as duas variáveis contínuas. Logo, foram definidas duas hipóteses:

- Hipótese Nula (H0): Não há correlação entre as variáveis
- Hipótese Alternativa (Ha): Há correlação entre as variáveis

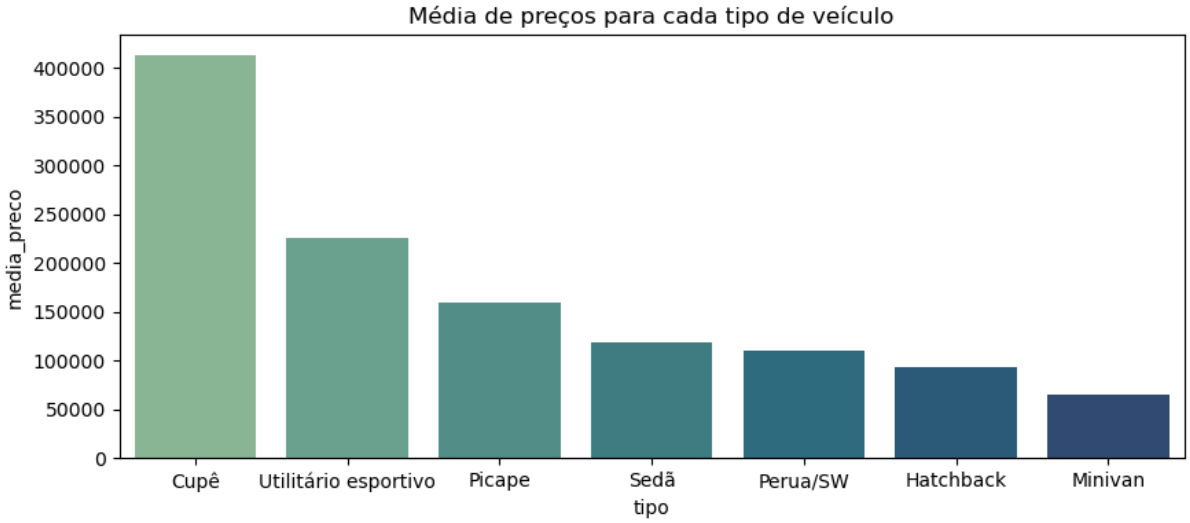
Analisando o Coeficiente de correlação de Pearson no teste de hipótese, decidiu-se rejeitar a H0 e aceitar Ha, ou seja, existe uma correlação entre valor registrado no hodômetro e o preço do veículo. Entretanto, como é possível observar no gráfico a seguir, existe uma correlação fraca e negativa, ou seja, à medida que o

valor registrado no hodômetro aumenta, o preço do veículo tende a diminuir. No entanto, a relação não é muito forte, o que significa que os dados não seguem um padrão consistente.

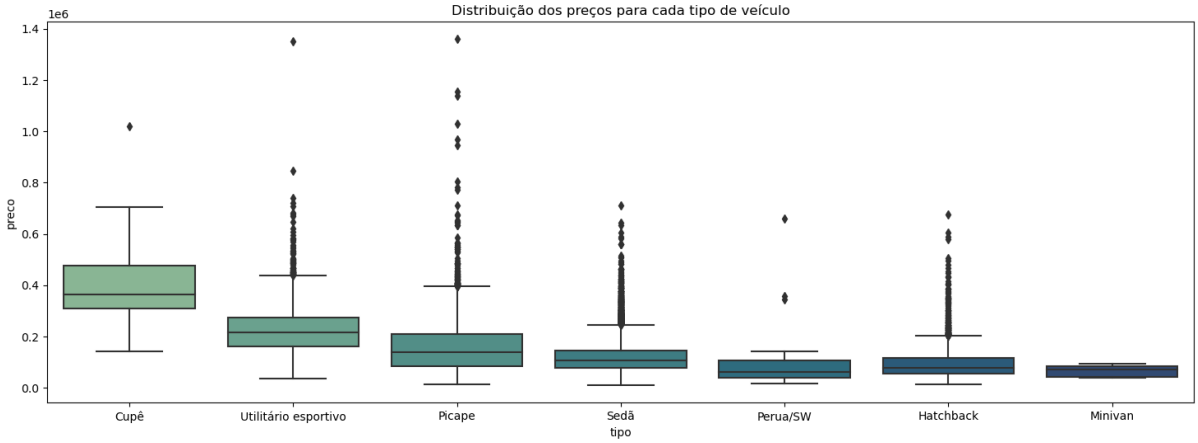
Ademais, para responder a **Hipótese de Negócio 2** foi utilizado a Análise Exploratória dos Dados e a Análise de Variância (ANOVA), pois, deseja-se comparar as médias da variável contínua ‘Preço’ entre os diversos grupos ‘Tipos de veículos’. Portanto, para definir se o tipo de veículo interfere na média de preço dos veículos foram criadas as hipóteses:

- Hipótese Nula (H0): Não há diferença significativa nos preços médios dos tipos de veículos.
- Hipótese Alternativa (Ha): Há diferença significativa nos preços médios entre os tipos de veículos.

Portanto, de acordo com o teste realizado decidiu-se rejeitar H0 e aceitar Ha, ou seja, existe uma diferença entre o preço médio dos diferentes tipos de veículos. Para consolidar a informação, é possível verificar o gráfico com o preço médio dos veículos de acordo com o tipo. Portanto, o tipo ‘Cupê’ possui o preço médio mais alto entre os tipos de veículo, enquanto, o tipo ‘Minivan’ possui o preço médio mais baixo.



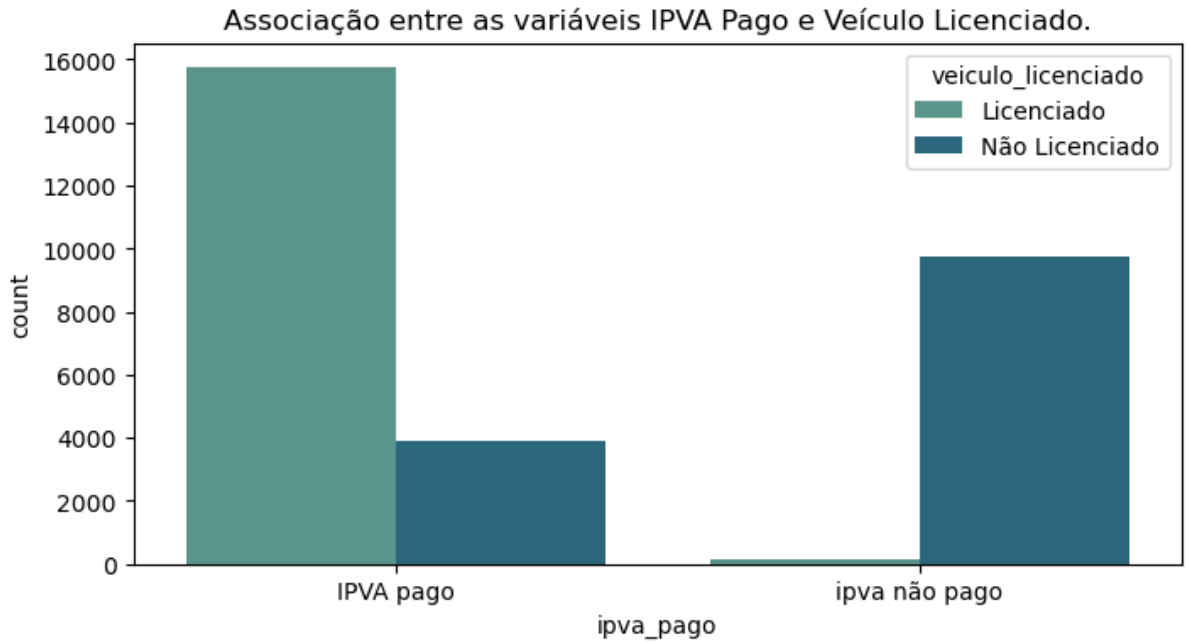
Além da média dos preços é possível observar a informação pela distribuição dos preços para cada tipo de veículo. Novamente, a mediana dos preços do ‘Cupê’ é a mais alta entre todos os tipos de veículo.



Por fim, para responder a **Hipótese de Negócio 3** foi utilizado a Análise Exploratória dos Dados e o teste do Qui-quadrado. Nessa hipótese, deseja-se entender se existe relação entre as variáveis categóricas IPVA pago e Licenciamento Pago. Isto posto, foram definidas as hipóteses:

- Hipótese Nula (H0): Não há associação entre "IPVA Pago" e "Licenciamento pago".
- Hipótese Alternativa (Ha): Há associação entre "IPVA Pago" e "Licenciamento pago".

Dessa forma, pelo teste a H0 foi rejeitada e Ha aceita, confirmando que há uma associação significativa entre as variáveis IPVA Pago e Licenciamento Pago. Logo, é possível observar, pelo gráfico a seguir, que geralmente quando o IPVA do veículo foi pago o Licenciamento também pode ser considerado pago. Da mesma forma, a maioria dos veículos que não possuem IPVA pago também não possuem Licenciamento pago.



Análise Exploratória dos Dados (EDA) - Perguntas de negócio

Nessa etapa, foram respondidas as seguintes perguntas de negócio:

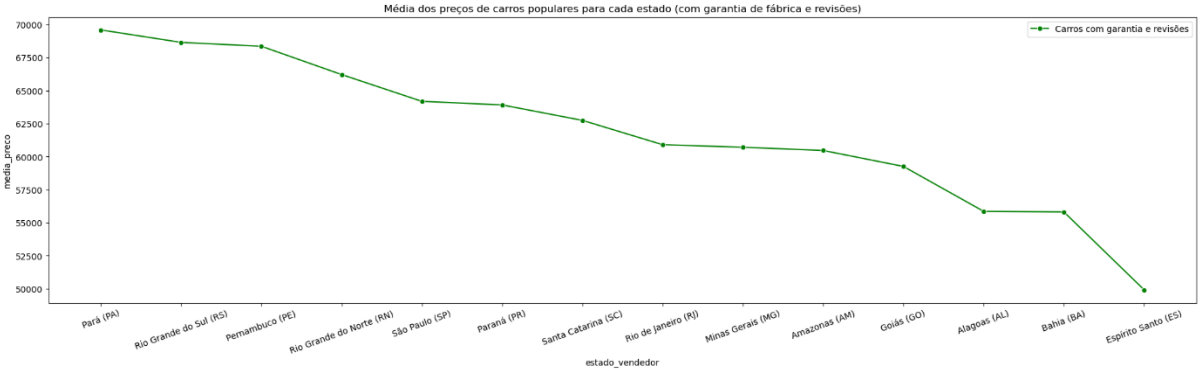
a. Qual o melhor estado cadastrado na base de dados para se vender um carro de marca popular e por quê?
b. Qual o melhor estado para se comprar uma picape com transmissão automática e por quê?
c. Qual o melhor estado para se comprar carros que ainda estejam dentro da garantia de fábrica e por quê?

Para definir qual o melhor estado para se vender um carro de marca popular foi considerados alguns pontos. A priori, definiu-se como carro popular os veículos com

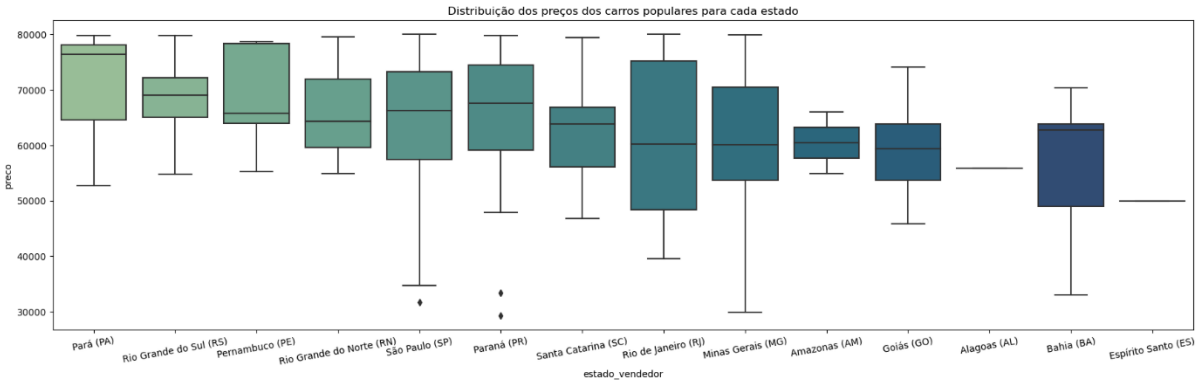
preço até 80.000.

Além disso, para definir qual o melhor estado, foram considerados aqueles que estados que oferecem veículos com garantia de fábrica e com revisões dentro da agenda. Esses fatores foram considerados, pois poderão dar ao cliente mais confiança na hora de efetuar a compra do carro popular. Dessa forma, foi analisada a média e a mediana dos preços dos carros populares em cada estado, para verificar aqueles em que a distribuição de preços é mais alta.

O gráfico a seguir realiza a comparação entre a média dos preços de veículos para cada estado considerando carros os carros que possuem com garantia de fábrica e revisões. Portanto, o estado que é possível vender carros de marca popular com média de preço mais alta é o estado Pará (PA), garantindo assim um maior lucro na vendo dos veículos.



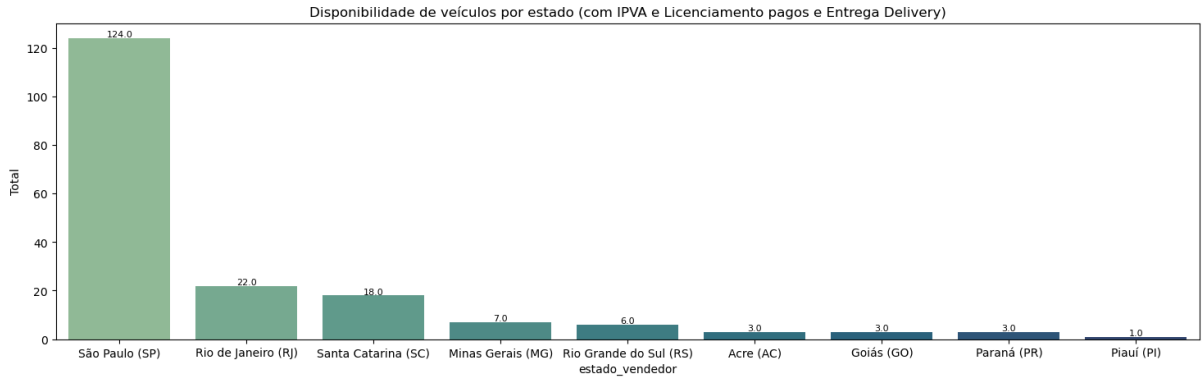
Além disso, pela distribuição dos preços dos veículos, novamente é possível observar que o estado Pará (PA) possui a mediana de valores mais alta para os carros de marca popular.



Para responder a segunda pergunta de negócio, foram considerados os

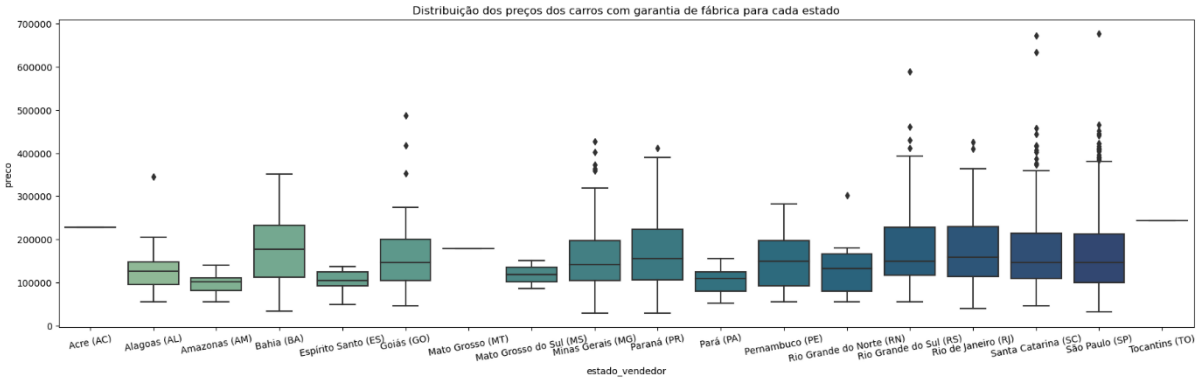
veículos do tipo 'Picape' com câmbio automático, excluindo os veículos com câmbio manual e semiautomático. Além disso, para determinar qual o melhor estado para se comprar o veículo, considerou-se a disponibilidade de veículos em cada um dos estados. Considerou-se também apenas os veículos que possuem IPVA Pago e Licenciamento Pago, por serem fatores de interesse na compra de um veículo. Por fim, o último fator de interesse seria se o vendedor entrega delivery, pois isso facilitaria a compra independentemente do estado.

Dessa forma, analisando os dados, o melhor estado para se comprar uma picape com transmissão automática seria São Paulo (SP), pois, é o estado com a maior disponibilidade de picapes com transmissão automática, com IPVA pago e Licenciamento pago e que a entrega é realizada via delivery. Logo, o estado de São Paulo (SP) é o que possui as características mais propícias para comprar o veículo.



Para responder qual o melhor estado para se comprar carros que ainda estejam dentro da garantia de fábrica, considerou-se apenas os veículos que possuem garantia de fábrica e os veículos que possuem revisões dentro da agenda, pois revisões realizadas regularmente podem ajudar a manter a garantia de fábrica válida e garantir o bom funcionamento do veículo. Ademais, o fator para determinar o melhor estado será a mediana dos preços. O estado que possuir a menor mediana de preços será o escolhido.

Portanto, pela análise exploratória dos dados, o melhor estado para se comprar carros que ainda estejam dentro da garantia de fábrica é o Amazonas (AM), pois, é o estado que possui a menor mediana dos preços do veículo que possuem garantia de fábrica e revisões dentro da agenda.



Pré-Processamento dos Dados

Na etapa de pré-processamento dos dados realizou-se o tratamento dos valores nulos. A priori, os valores das seguintes *features* foram substituídos por um valor que corresponde ao sentido oposto dos valores não nulos:

- 'dono_aceita_troca': 'Nao aceita troca';
- 'veiculo_único_dono': 'Não é único dono'
- 'revisoes_concessionaria': 'Não possui todas as revisoes feitas pela concessionaria',
- 'ipva_pago': 'ipva não pago',
- 'veiculo_licenciado': 'Não Licenciado';
- 'garantia_de_fábrica': 'Não possui garantia de fábrica';
- 'revisoes_dentro_agenda: não possui todas as revisões feitas pela agenda do carro'.

Ademais, os valores nulos de *'num_fotos'* foram substituídos pela mediana do número de fotos. Foi analisado se existiam valores duplicados, porém não se encontrou nenhuma ocorrência, logo, não foi necessário nenhum tratamento. Além disso, foram eliminados os atributos:

- Id: ID único por amostra;
- veiculo_alienado: pois todos os valores eram nulos.

Em seguida, foi realizado o tratamento das variáveis categóricas. Nessa etapa, foi utilizada a técnica *one-hot encoding*. Vale salientar, que como existiam uma quantidade de valores únicos para as *features* de treinamento superior a quantidade

valores únicos para as *features* de teste, foi necessário aplicar uma operação de união entre os valores únicos dos dois DataFrames antes de realizar *o one-hot encoding*, a fim de evitar que as colunas após o tratamento fossem diferentes para os datasets de treinamento e teste.

Ademais, foi realizada a normalização dos dados para as variáveis numéricas, a fim de, manter todos os valores em uma escala padrão:

- 'num_fotos';
- 'ano_de_fabricacao';
- 'ano_modelo';
- 'hodometro';
- 'num_portas'

Por fim, a última etapa consistiu em realizar a Análise de Componentes Principais (PCA) para reduzir a dimensionalidade dos dados, antes de serem enviados para o treinamento.

Modelagem

Para fazer a previsão do preço a partir dos dados utilizou-se algoritmos de aprendizado supervisionado, visto que, temos a disponibilidade de dados anotados. Dessa forma, como o objetivo é prever os rótulos de saída contínuos, referente ao preço dos veículos, foi utilizado o conceito de regressão.

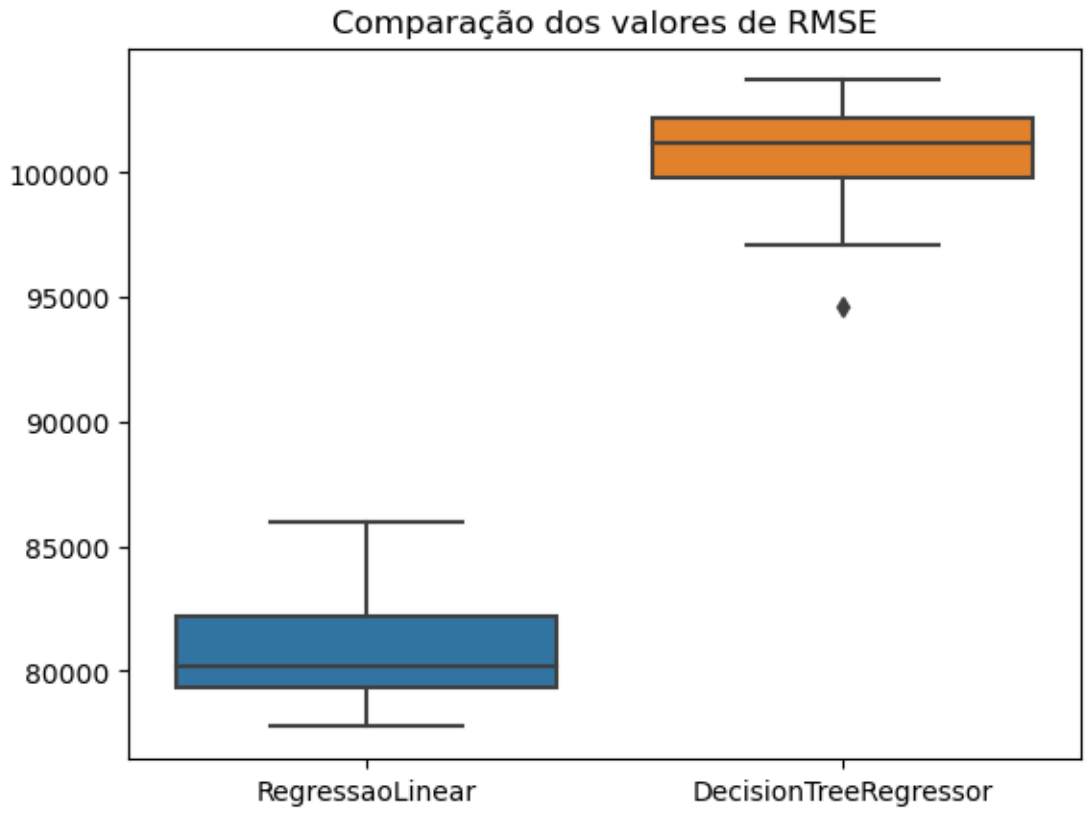
Logo, foi utilizada a validação cruzada para avaliar qual modelo melhor se adequa para o problema escolhido de acordo com as métricas de avaliação. Na etapa de modelagem utilizou-se os modelos de Regressão Linear e Decision Tree Regressor.

Avaliação

Para avaliar a performance e desempenho dos algoritmos de regressão foi utilizada a métrica *Root Mean Squared Erro (RMSE).* A métrica fornece a média dos erros em unidades originais da variável de resposta. Dessa forma, quanto menor o valor, melhor é o desempenho do modelo, indicando uma maior precisão nas previsões. Portanto, como é possível observar no gráfico a seguir, o modelo de

Regressão Linear apresentou valores menores para a métrica RMSE, sendo o modelo escolhido para esse problema.

- Regressão Linear: RSME = 80859 (média)
- Decision Tree Regressor: RSME = 100520 (média)



Portanto, com o modelo Regressão Linear pode-se prever os preços para os veículos do *dataset* de teste. O resultado do modelo foi salvo em uma planilha com o id do veículo e os preços preditos.

Código do projeto

O código do projeto está disponível no link a seguir:
https://github.com/andressagomes26/marketplace_lh_cd.git