Desafio Cientista de dados

Análise de dados dataset

Juliana Marian Arrais

Contents

1 Introdução **2**

1.1 Objetivos 2

1.1.1 Objetivos específicos 3

2 Metodologia **4**

2.1 Recursos 4

2.2 Dados utilizados 5

2.2.1 Pré-processamento de dados 8

2.3 Modelos 10

3 Resultado e Discussão **10**

4 Modelo **29**

4.1 Linear Regression 32

4.2 Ridge 33

4.3 Lasso 35

4.4 Comparação dos três modelos 37

5 Conclusão **38**

1 Introdução

Em muitas ocasiões, torna-se essencial vender o carro usado, mas é desafiador determinar com precisão o preço justo do veículo. A depreciação de um carro é influenciada por diversos fatores, o que requer que o proprietário esteja ciente do valor real de seu veículo no mercado. Felizmente, com o rápido avanço do aprendizado de máquina, esse problema pode ser resolvido de forma eficiente e economizando tempo humano.

Neste contexto, apresentaremos uma solução abrangente para um problema semelhante, utilizando técnicas de aprendizado de máquina. Abordaremos todo o processo, desde a coleta de dados até a implementação do modelo, com o objetivo de prever o valor justo de um carro usado.

Essa abordagem permitirá ao proprietário do carro obter uma estimativa precisa e objetiva do preço do veículo, evitando desvalorizações injustas e maximizando o retorno financeiro na venda.

Através dessa solução de ponta a ponta, buscaremos simplificar o processo de determinar o valor de um carro usado, proporcionando maior confiabilidade e praticidade para os proprietários que desejam vender seus veículos com segurança e assertividade.

1.1 Objetivos

O objetivo deste relatório é realizar uma análise das principais estatísticas do conjunto de dados contendo variáveis utilizadas para prever o preço de carros fornecido pela empresa Incidium.

1.1.1 Objetivos específicos

- Descrever graficamente essas variáveis (features), apresentando as suas principais estatísticas descritivas. Comente o porquê da escolha destas estatísticas e o que elas nos informam.

- Faça uma EDA. Nesta EDA, crie e responda 3 hipóteses de negócio. Além disso, responda também às seguintes perguntas de negócio:

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

  - 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?

  - 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)?

  - E Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?

- Envie o resultado final do modelo em uma planilha com apenas duas colunas (id, preco).

- A entrega deve ser feita através de um repositório de código público que contenha:

  - README explicando como instalar e executar o projeto

– Arquivo de requisitos com todos os pacotes utilizados e suas versões

– Relatórios das análises estatísticas e EDA em PDF, Jupyter Notebook ou semelhante conforme passo 1 e 2.

– Códigos de modelagem utilizados no passo 3.

– Arquivo final com o nome predicted.csv conforme passo 4 acima.

– Todos os códigos produzidos devem seguir as boas práticas de codificação.

2 Metodologia

Nesta seção, descreveremos a metodologia desenvolvida para a realização deste relatório.

2.1 Recursos

Todo o desenvolvimento desse trabalho envolvendo a análise de dados, os treinamentos e a análise de resultados foi realizado na plataforma Google Colab[1]. Essa ferramenta possibilita o acesso a boas placas de vídeo, sendo fácil de instalar e de utilizar pacotes e bibliotecas de Python e Jupyter Notebook[2] como ferramente principal, além de ter uma boa integração com o Google Drive.

Para as tarefas de carregamento de dados e análise de resultados, as principais bibliotecas utilizadas foram Pandas[3], NumPy[4], Matplotblib[5], Seaborn[6]. Para as tarefas de Machine Learning, utilizou-se o Scikit-learn[7]

[1] https://research.google.com/colaboratory/faq.html
[2] https://jupyter.org/
[3] https://pandas.pydata.org/
[4] https://numpy.org/
[5] https://matplotlib.org/
[6] https://seaborn.pydata.org/
[7] https://scikit-learn.org/stable/

O relatório foi elaborado utilizando o editor LaTeX no Overleaf[8], uma plataforma online para edição de documentos com formatação técnica. Para agilizar o processo de pesquisa, também foram feitas consultas ao ChatGPT[9]. Todos os dados, informações e códigos utilizados estão disponíveis na Pasta Google Drive e no repositório do GitHub[10] (GitHub). O relatório completo pode ser acessado em Relatório Overleaf.

2.2 Dados utilizados

Para este projeto foram utilizados os dados fornecidos pela empresa Indicium[11]. Um dataset para treinamento chamado cars_training[12] composto por 29584 linhas, 28 colunas de informação (features) e a variável a ser prevista ("preco"). Um segundo dataset para teste chamado de cars_test[13] composto por 9862 linhas e 28 colunas, sendo que este dataset não possui a coluna "preco". O link com o desafio encontra-se no Google Drivre[14].

Os dados possuem variáveis informativas que compõem todas colunas e são descritas abaixo:

- id: Contém o identificador único dos veículos cadastrados na base de dados
- num_fotos: contém a quantidade de fotos que o anuncio do veículo contém
- marca: Contém a marca do veículo anunciado
- modelo: Contém o modelo do veículo anunciado

[8] https://www.overleaf.com
[9] https://chat.openai.com/
[10] https://github.com/
[11] https://indicium.tech/
[12] https://drive.google.com/file/d/1OUTEJwGrSZbePYNlmUQppfXmGvdFk8Ci/view?usp=drive$_l ink$
[13] https://drive.google.com/file/d/1BrA4XFjERtJDSZvk5$_z 3IhXo3etP2wNq/view?usp = drive_l ink$
[14] https://docs.google.com/document/d/1mHIskC1n1VINLvn10$_T IEL9fnT_O ayG$ – $X_g rdlxVnWY/edit?usp = drive_l ink$

- versao: Contém as descrições da versão do veículo anunciando. Sua cilindrada, quantidade de válvulas, se é flex ou não, etc.
- ano_de_fabricacao: Contém o ano de fabricação do veículo anunciado
- ano_modelo: Contém o modelo do ano de fabricação do veículo anunciado
- hodometro: Contém o valor registrado no hodômetro do veículo anunciado
- cambio: Contém o tipo de câmbio do veículo anunciado
- num_portas: Contém a quantidade de portas do veículo anunciado
- tipo: Contém o tipo do veículo anunciado. Se ele é sedã, hatch, esportivo, etc.
- blindado: Contém informação se o veículo anunciado é blindado ou não
- cor: Contém a cor do veículo anunciado
- tipo_vendedor: Contém informações sobre o tipo do vendedor do veículo anunciado. Se é pessoa física (PF) ou se é pessoa jurídica (PJ)
- cidade_vendedor: Contém a cidade em que vendedor do veículo anunciado reside
- estado_vendedor: Contém o estado em que vendedor do veículo anunciado reside
- anunciante: Contém o tipo de anunciante do vendedor do veículo anunciado. Se ele é pessoa física, loja, concessionário, etc
- entrega_delivery: Contém informações se o vendedor faz ou não delivery do veículo anunciado

- troca: Contém informações o veículo anunciado já foi trocado anteriormente
- elegivel_revisao: Contém informações se o veículo anunciado precisa ou não de revisão
- dono_aceita_troca: Contém informações se o vendedor aceita ou não realizar uma troca com o veículo anunciado
- veiculo_único_dono: Contém informações o veículo anunciado é de um único dono
- revisoes_concessionaria: Contém informações se o veículo anunciado teve suas revisões feitas em concessionárias
- ipva_pago: Contém informações se o veículo anunciado está com o IPVA pago ou não veiculo_licenciado: Contém informações se o veículo anunciado está com o licenciamento pago ou não
- garantia_de_fábrica: Contém informações o veículo anunciado possui garantia de fábrica ou não
- revisoes_agenda: Contém informações se as revisões feitas do veículo anunciado foram realizadas dentro da agenda prevista
- veiculo_alienado: Contém informações se o veículo anunciado está alienado ou não
- preco (target): Contém as informações do preço do veículo anunciado

O conjunto de dados de teste se encontra disponível no link: Dataset teste e o conjunto de dados para treino em: Dataset treino.

2.2.1 Pré-processamento de dados

O pré-processamento de dados desempenha um papel fundamental no desenvolvimento de modelos de Machine Learning e para a análise de dados, visando obter resultados precisos e confiáveis. Essa etapa tem como objetivo preparar os dados brutos de entrada para o modelo selecionado, garantindo que estejam em um formato adequado e prontos para o treinamento.

A primeira etapa consiste em limpar os dados, pois geralmente os conjuntos de dados contêm ruídos, valores ausentes e inconsistências, como dados *outliers*.

Para o modelo de Machine Learning foram excluidas as variáveis *id* e *num_fotos*. A exclusão de variáveis desnecessárias é uma prática comum em análise de dados e modelagem, pois pode ajudar a melhorar a eficiência e a precisão do modelo. A variável *id* é um identificador e não possui nenhuma relação com o preço do carro. A variável *num _fotos* é interessante para os compradores, porém, é provável que a inclusão dessa variável não agregue muita informação para a previsão do preço do veículo. Além disso, a quantidade de fotos é um número discreto e não contínuo, o que pode dificultar sua interpretação em alguns modelos de Machine Learning.

Foram utilizados os métodos .head(), .info() e .isnull() para visualizar as primeiras linhas do conjunto de dados e obter informações sobre cada variável. Durante essa análise, foram identificadas variáveis que continham dados ausentes (NaN, *Not a Number*, e elas estão listadas abaixo:

- 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

Na análise de dados, é frequente encontrar datasets que possuam valores ausentes ou faltantes. Essas lacunas podem surgir por diversas razões, como erros na coleta ou inserção de dados, falhas nos sensores de medição ou porque algumas informações não foram fornecidas. Para representar esses valores ausentes em um DataFrame ou matriz numérica, é utilizado o marcador "NaN" (*Not a Number*). Quando um valor está ausente ou não pode ser expresso numericamente, o Pandas atribui o valor NaN a essa posição no DataFrame.

Durante o pré-processamento dos dados, é crucial detectar os valores NaN e tomar decisões sobre como tratá-los adequadamente. Existem diversas estratégias para lidar com esses valores ausentes, como o preenchimento utilizando estatísticas como média ou mediana, a utilização do último valor válido fornecido ou até mesmo a opção de removê-los do conjunto de dados por completo. A escolha da melhor abordagem dependerá da natureza dos dados e do contexto específico da análise. É importante realizar uma análise cuidadosa para garantir que o tratamento dos valores NaN não comprometa a integridade e a qualidade dos resultados obtidos.

Os modelos de machine learning geralmente não aceitam valores NaN como entrada. Valores ausentes podem causar problemas durante o treinamento do modelo e levar a resultados imprecisos ou inesperados. Portanto, as variáveis de entrada para o modelo foram

selecionadas excluindo todas aquelas que continham valores NaN. Essa decisão foi tomada em função da limitação de tempo para analisar detalhadamente cada coluna do conjunto de dados.

2.3 Modelos

Existem diversos modelos de machine learning que podem ser utilizados para prever os preços dos carros com base em dados históricos ou informações relevantes sobre os veículos. Segundo o artigo de [1] alguns dos modelos mais comuns são: Regressão Linear, Regressão de Árvore de Decisão, *Random Forest* (Floresta Aleatória), *Gradient Boosting*, *Support Vector Machines* (SVM), *K-Nearest Neighbors* (KNN), *XGBoost*.

No artigo [1] os autores citam que esse problema de previsão de preços de carros usados é um problema de regressão.

3 Resultado e Discussão

A EDA *(Exploratory Data Analysis)* é a etapa inicial e fundamental no processo de análise de dados. Consiste em explorar e investigar os dados de forma visual e descritiva para obter insights, identificar padrões, relações e tendências entre as variáveis. Envolve o uso de estatísticas descritivas, visualizações de dados e análise de correlações. A EDA é essencial para a compreensão completa dos dados e orienta a seleção de técnicas analíticas adequadas.

Inicialmente, realizou-se uma análise dos preços para verificar a presença de outliers. A média da variável "preços" foi de $R\$133.023,87$. Essa média é relativamente alta em relação a carros usados, porém, ela ficou elevada devido à presença de carros de luxo no dataset, que possuem valores muito superiores aos de carros populares. O carro mais caro do dataset

custa $R\$: 1.359.812.8923$ e o mais barato custa $R\$: 9.869.95$.

A contagem das marcas presentes no conjunto de dados para identificar as que possuem maior quantidade de registros está registrada na figura 1. E na figura 2 contém o gráfico de pizza das marcas, que é uma forma comum de representar a distribuição proporcional de categorias em um conjunto de dados. A marca Volkswagen corresponde a 15,5% do total do dataset. Na figura 3 contém a contagem das 5 marcas mais comuns no dataset.

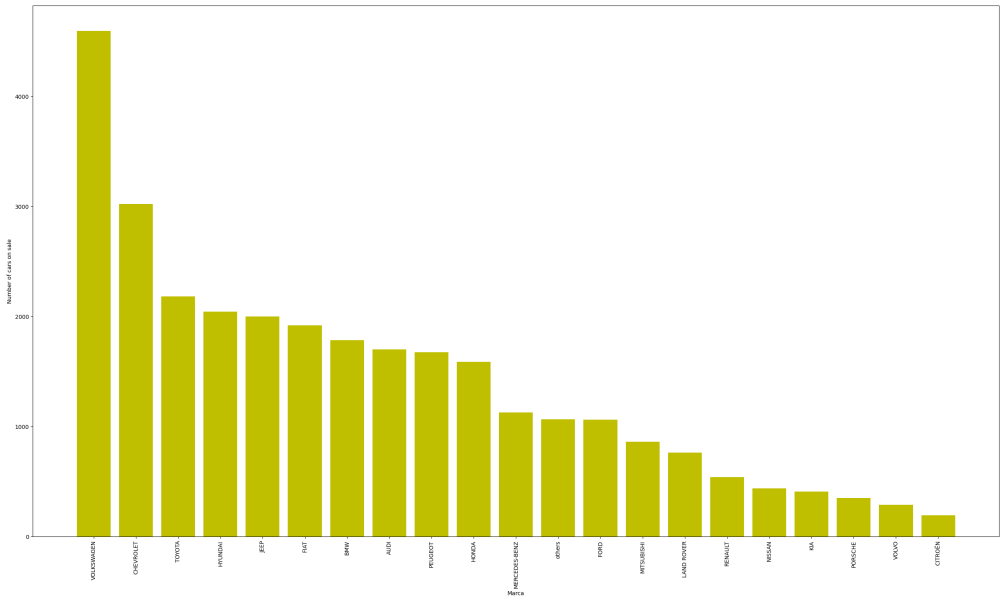

Figure 1: Contagem das marcas no dataset
Font: Elaborada pela autora (2023).

Outro fator importante para a venda de um carro é análise do câmbio. Na figura 4 observamos que 76,2% dos carros possuem câmbio automático e 16,9% correspondem ao câmbio manual. E no gráfico 5 é possível verificar que mais da metade dos vendedores são pessoas física.

Além disso, foram criados gráficos para visualizar a distribuição dos carros à venda neste dataset entre os estados brasileiros (figura 6 e na figura 7).

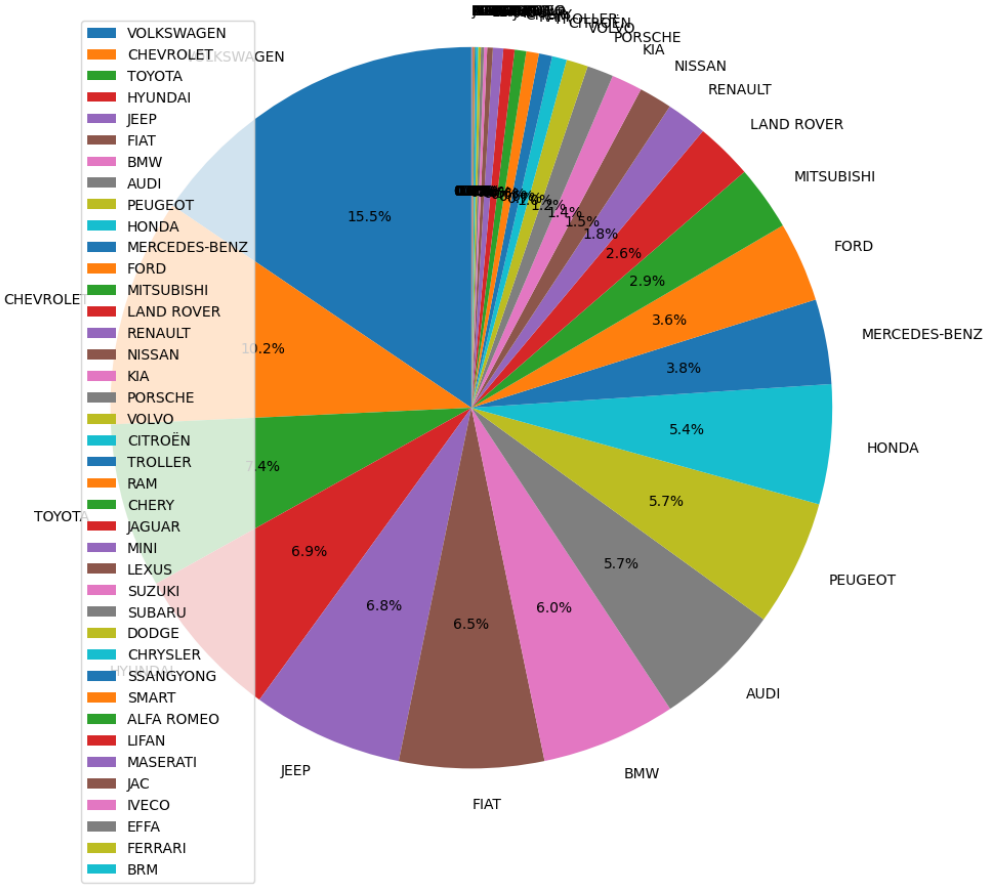

Figure 2: Gráfico pizza das marcas
Font: Elaborada pela autora (2023).

A fim de uma futura apresentação para o cliente, foi elaborado o *Wordclouds* para as variáveis *marca* (figura 8) e *cor* (figura 9). O *Wordclouds* são representações visuais de dados de texto, onde o tamanho de cada palavra na nuvem corresponde à sua frequência ou importância dentro do texto fornecido. Em uma nuvem de palavras, as palavras que aparecem com mais frequência no texto são exibidas com um tamanho de fonte maior, enquanto as palavras menos frequentes são mostradas com tamanhos de fonte menores.

Como mencionado anteriormente, na figura 8, as marcas com maior frequência no dataset

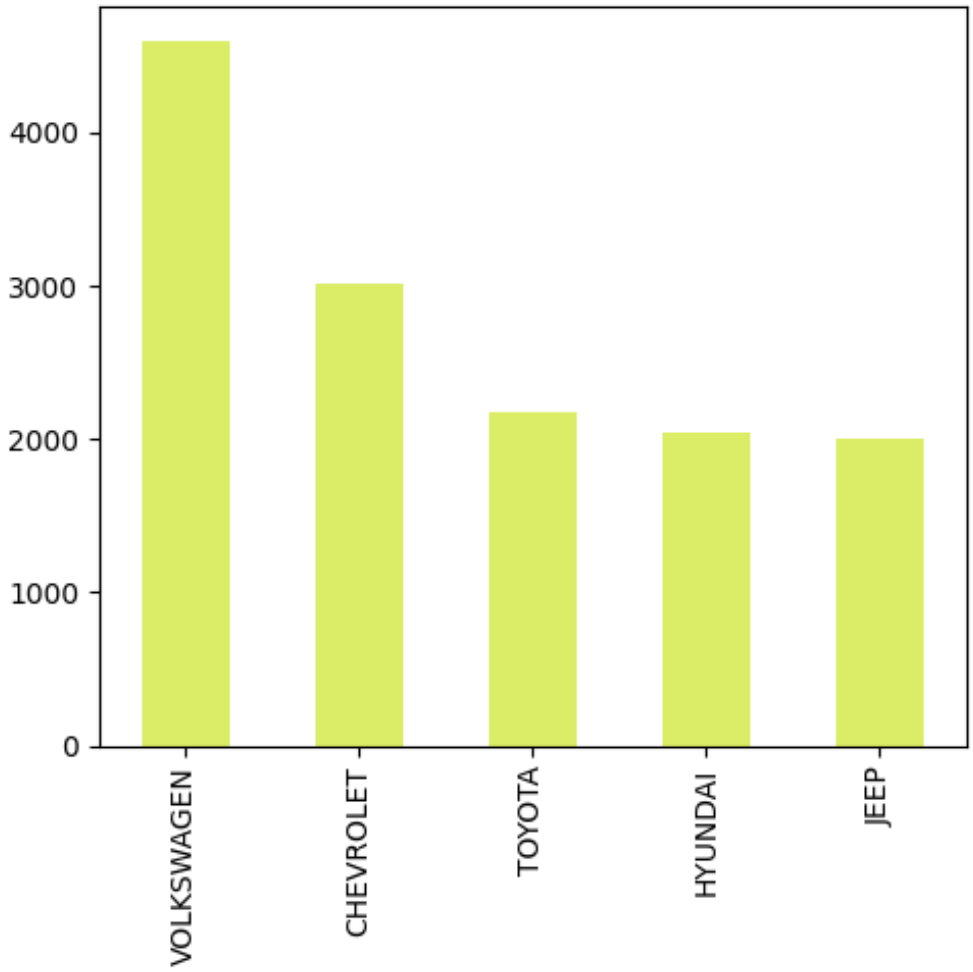

Figure 3: As 5 marcas mais comuns no dataset
Font: Elaborada pela autora (2023).

são Volkswagen e Chevrolet. Além disso, na figura 9, podemos observar que a cor que mais corresponde ao dataset é a cor Branco.

Para a variável Preço foi plotado o gráfico de frequência de distribuição na figura 10 com todos os dados do dataset, ou seja, com outliers. Esse gráfico é uma forma de representar visualmente como os valores do preço estão distribuídos em um conjunto de dados. Essa visualização é importante para entender a distribuição dos valores e identificar padrões,

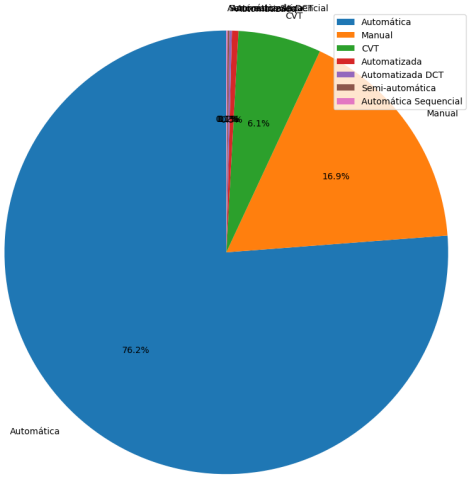

Figure 4: Gráfico pizza do câmbio
Font: Elaborada pela autora (2023).

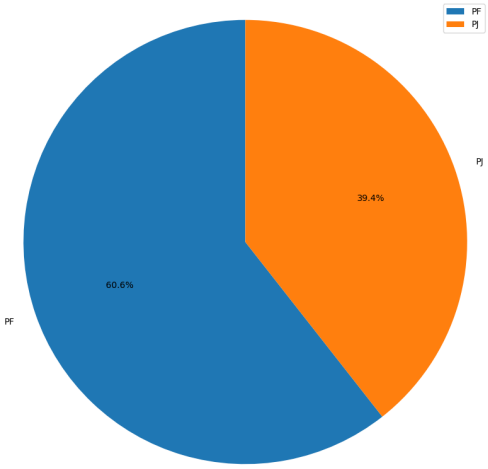

Figure 5: Gráfico pizza sobre o vendedor
Font: Elaborada pela autora (2023).

tendências ou características importantes dos dados.

Para identificar e tratar outliers nos dados de preço, foi utilizado a variável $price_upper_limit$ com percentil 75, onde um valor que separa os 75% maiores valores dos 25% menores em um conjunto de dados ordenados com o coeficiente de dispersão 1.5. Portanto, a formula

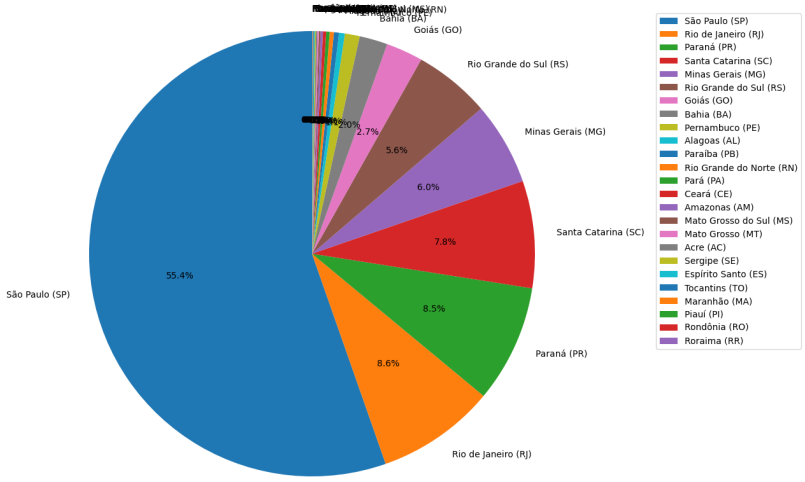

Figure 6: Gráfico pizza com a porcentagem dos estados brasileiros
Font: Elaborada pela autora (2023).

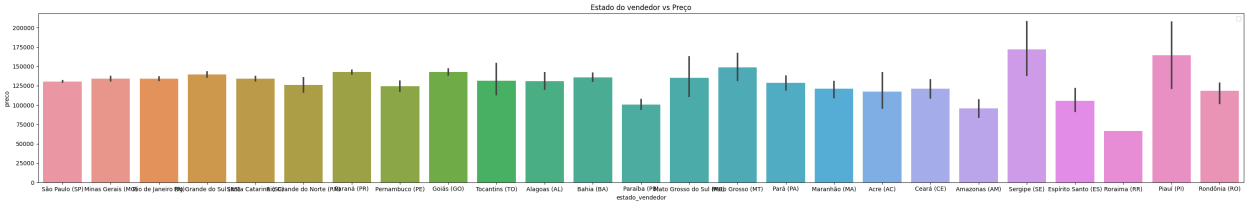

Figure 7: Gráfico relação preço dos carros disponíveis vs estados
Font: Elaborada pela autora (2023).

utilizada foi de: $price\_upper\_limit = price\_percentile75 + 1.5 * price_iqr$. Onde stamos estabelecendo um limite superior para os preços no conjunto de dados. Quaisquer preços acima desse limite são considerados valores extremos ou outliers e podem ser tratados de diferentes maneiras, como remoção, transformação ou substituição por valores mais adequados. A frequência de distribuição dos preços com esse pré-processamento encontra-se na figura 11.

Uma outra relação interessante é entender como a quilometragem (hodômetro) se ajusta em relação aos valores da variável "preço". Para isso, foi criado um gráfico de dispersão

Figure 8: Wordclouds para a variável marca
Font: Elaborada pela autora (2023).

Figure 9: Wordclouds para a variável cor
Fonte: Elaborada pela autora (2023).

(*scatter plot*) com uma linha de regressão linear ajustada aos dados. A figura 12 mostra claramente que à medida que a quilometragem aumenta, o preço tende a diminuir. No entanto, é importante destacar que pode haver outliers no dataset. Conforme observado no gráfico de regressão, os carros com quilometragem acima de 400.000 podem ser considerados discrepantes e seria recomendado removê-los do conjunto de dados para evitar distorções em

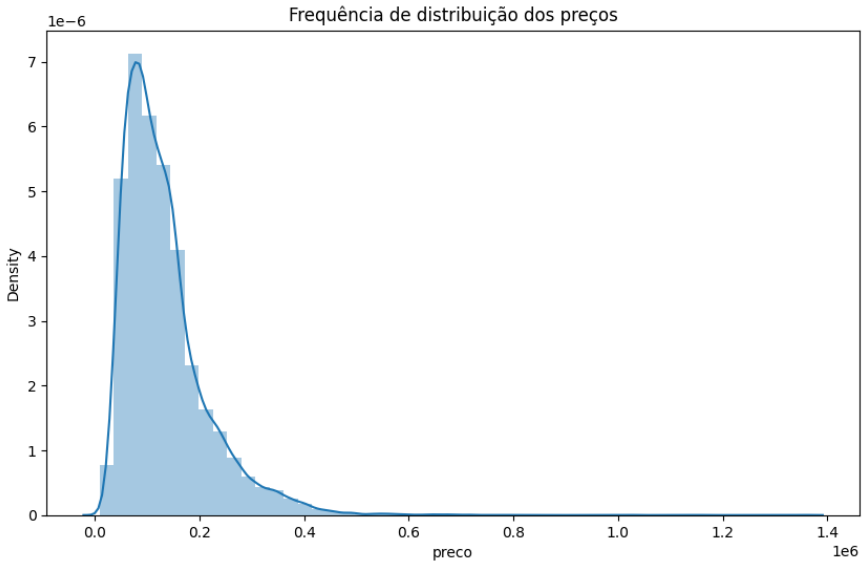

Figure 10: Frequência de distribuição dos preços com outliers
Fonte: Elaborada pela autora (2023).

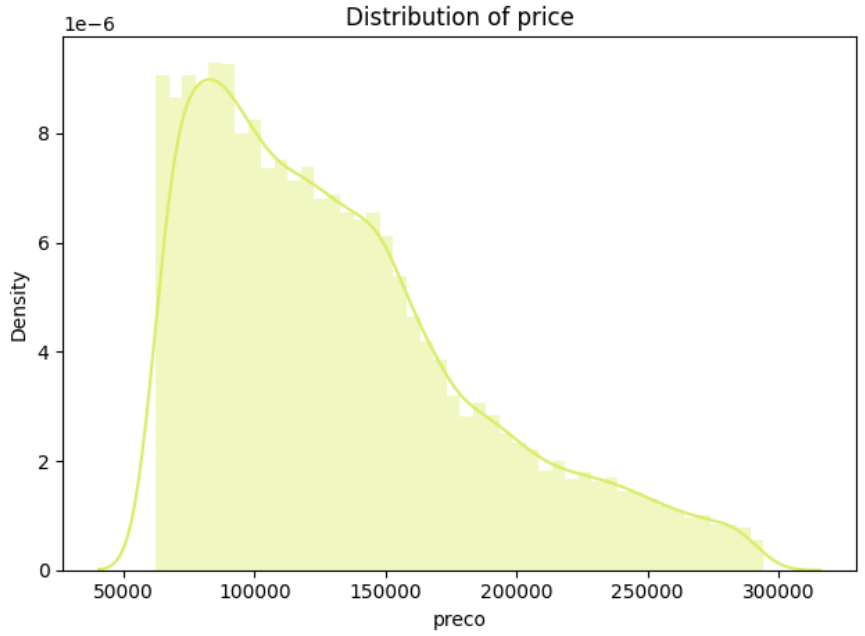

Figure 11: Frequência de distribuição dos preços
Fonte: Elaborada pela autora (2023).

análises posteriores.

Na figura 13 se encontra o gráfico com os cinco estados com o maior número médio em relação ao hodometro. O estado de Roraima é o estado brasileiro que tem a média de quilometragem mais alta de todo o dataset. E na figura 14 mostra qual tipo de carro contém a média do hodometro mais alta. A perua/SW são os tipos que contém mais quilometragem do dataset.

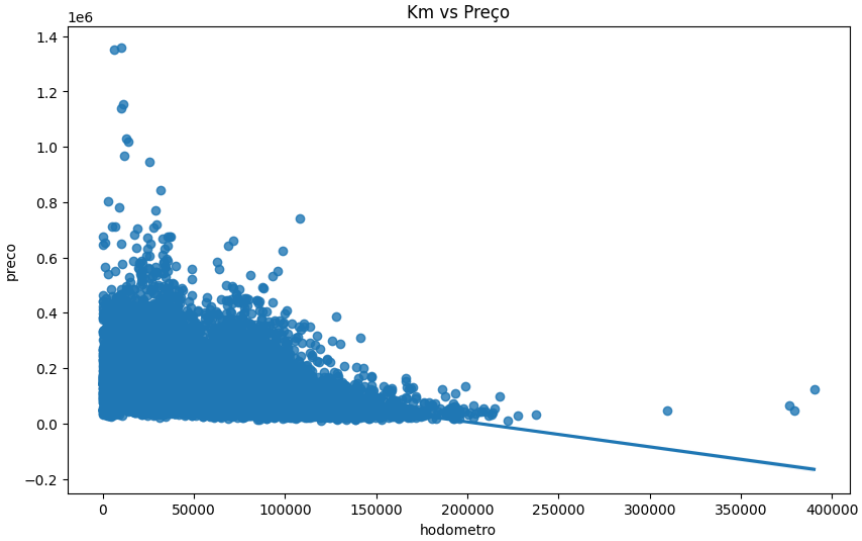

Figure 12: Hodometro vs Preço
Fonte: Elaborada pela autora (2023).

Na análise representada na Figura 15, foi também aplicada a técnica de regressão linear para investigar a relação entre o ano de fabricação e o preço dos carros. Conforme esperado, verificou-se que carros mais recentes apresentam valores mais elevados, enquanto carros mais antigos tendem a ter preços mais baixos. No entanto, é relevante mencionar que carros muito antigos podem ser identificados como possíveis anomalias. Essa observação indica que a tendência geral de aumento dos preços com o passar dos anos pode ser interrompida ou não se aplicar aos veículos mais antigos, possivelmente devido a fatores específicos relacionados

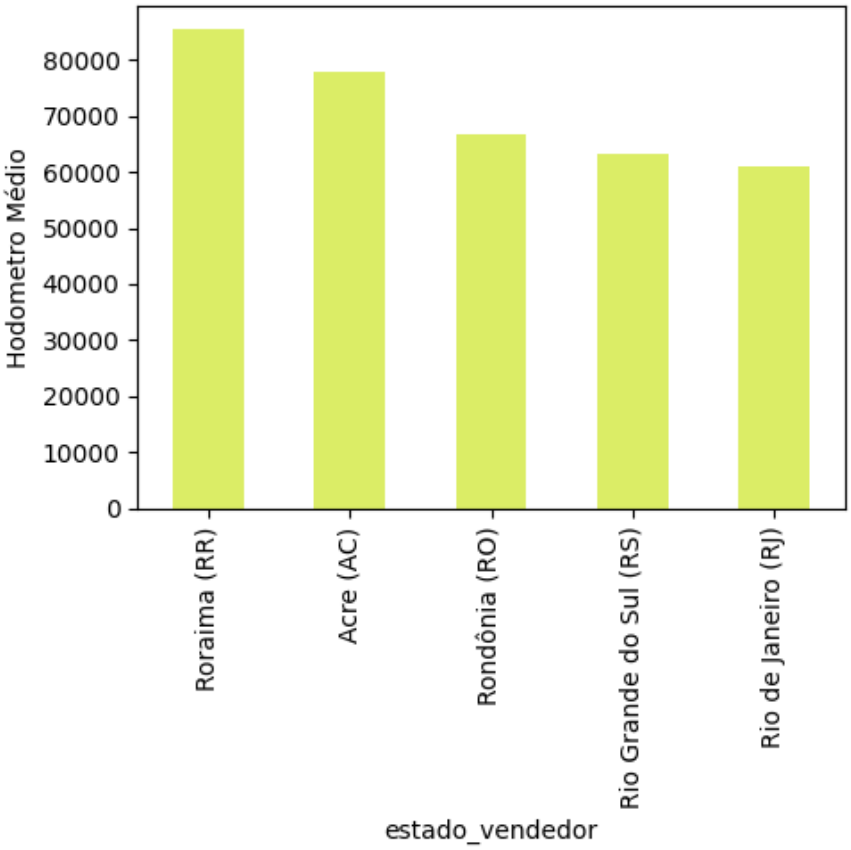

Figure 13: Os cinco estados com o maior número médio em relação ao hodometro
Fonte: Elaborada pela autora (2023).

a essa faixa temporal, como raridade, antiguidade e condições de conservação. Para melhor visualização, também podemos observar na imagem 16.

O *heatmap* (mapa de calor) é uma ferramenta gráfica utilizada para visualizar e representar matrizes de dados através de cores e é amplamente utilizado para identificar padrões, tendências e relacionamentos entre duas dimensões diferentes nos dados. Na figura 17 podemos observar a distribuição e intensidade dos valores presentes na matriz.

O câmbio do carro desempenha um papel fundamental na experiência de direção e pode influenciar diretamente a satisfação do proprietário com o veículo. Por isso, é importante

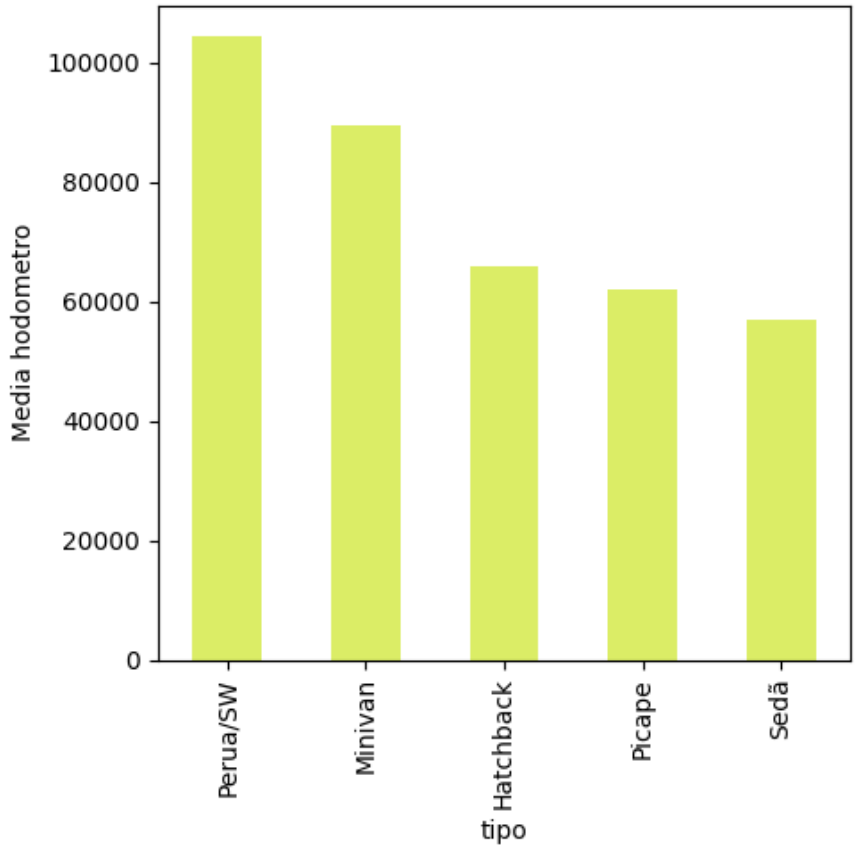

Figure 14: Os cinco estados com o maior número médio em relação ao hodometro
Fonte: Elaborada pela autora (2023).

analisar cuidadosamente as opções disponíveis e escolher aquela que melhor se adapta às necessidades individuais. No gráfico 18 foi plotado a relação preço vs tipo de câmbio. Como observado no gráfico, o câmbio "Automaticada CVT" (Transmissão Continuamente Variável) possui o maior valor do mercado. Esse tipo de câmbio não possui marchas fixas, mas sim uma ampla faixa de relação de marcha contínua. Oferece aceleração suave e eficiente, além de melhorar a economia de combustível. O câmbio mais em conta segundo o gráfico é o câmbio "Automática sequencial". Esse tipo combina elementos de um câmbio manual com a automação das trocas de marcha. Proporciona mudanças de marchas mais rápidas e suaves,

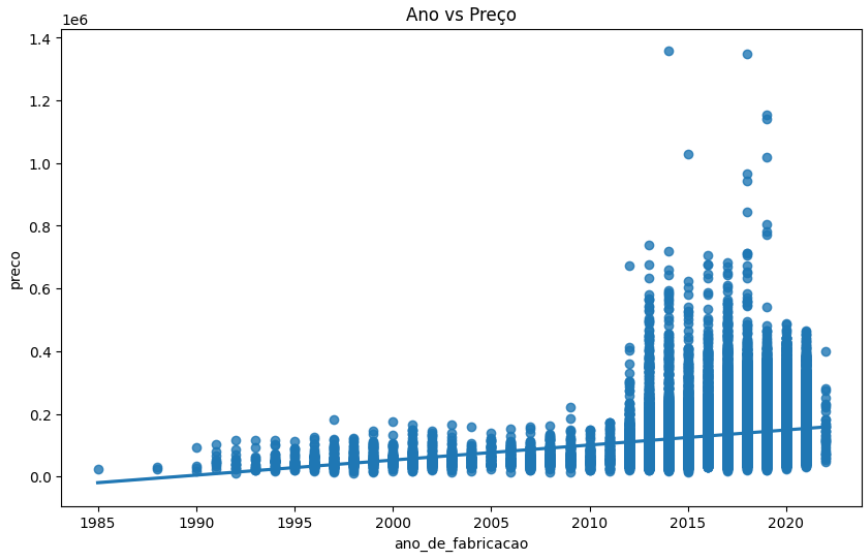

Figure 15: Ano vs Preço
Fonte: Elaborada pela autora (2023).

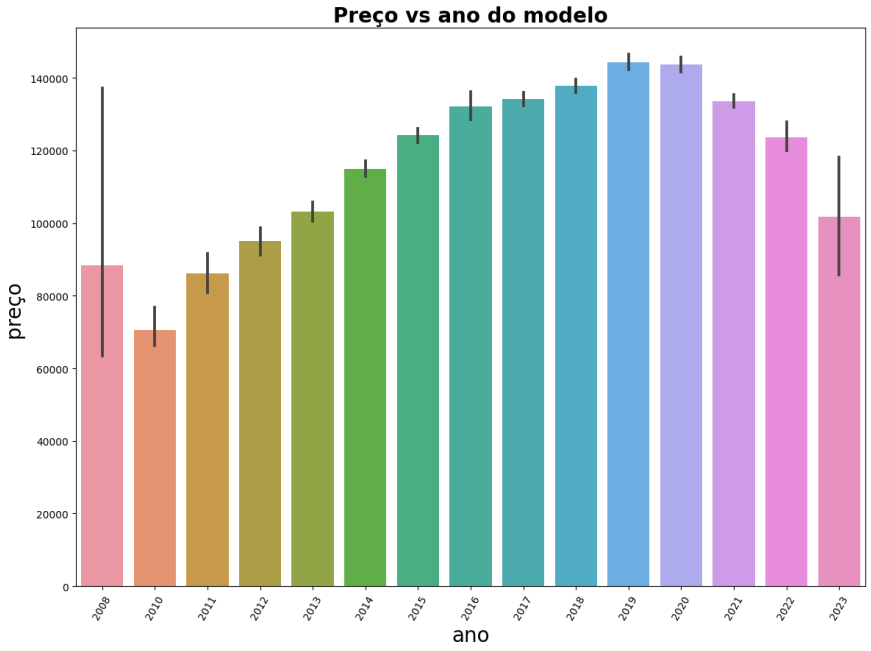

Figure 16: Ano vs Preço
Fonte: Elaborada pela autora (2023).

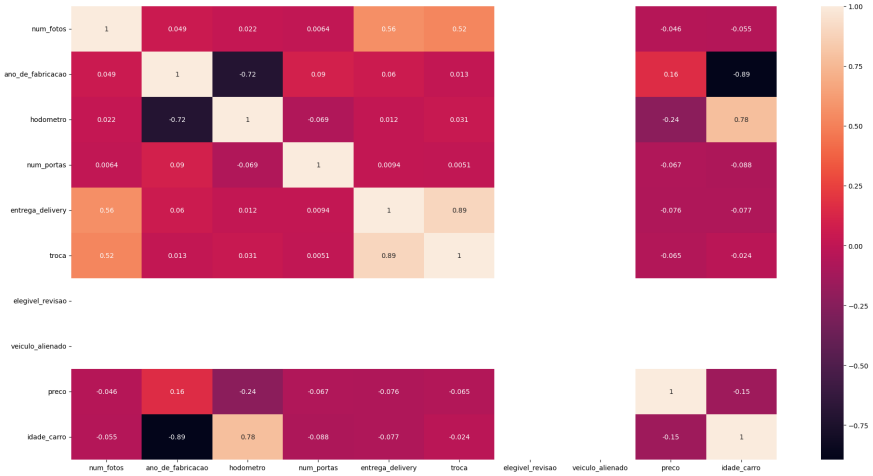

Figure 17: Heatmap
Fonte: Elaborada pela autora (2023).

tornando a condução mais esportiva e eficiente.

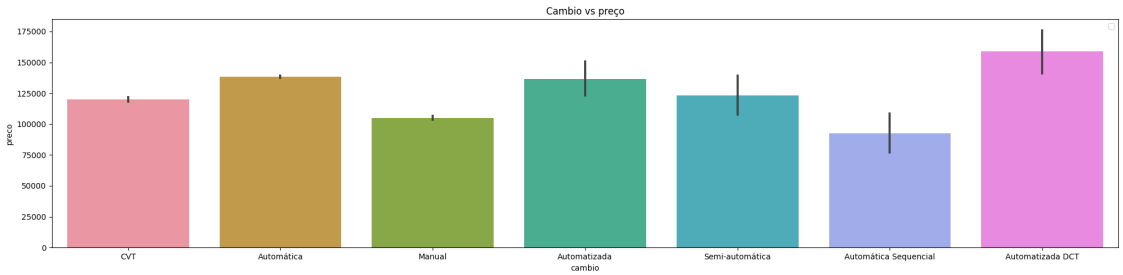

Figure 18: Heatmap
Fonte: Elaborada pela autora (2023).

Intuitivamente, os carros com câmbio manual tendem a ser mais baratos do que os carros com câmbio automático ou outros tipos de transmissão. Essa diferença de preço ocorre principalmente devido à menor complexidade da tecnologia e dos componentes envolvidos em um câmbio manual, o que resulta em um custo de fabricação menor.

Entretanto, pelo gráfico analisado os carros com câmbio manual não aparecem como os mais baratos, isso pode ser um indicativo da presença de outliers no dataset.

Portanto, a presença de outliers nos preços dos carros ou na variável câmbio podem estar influenciando na aparente discrepância entre os preços dos carros com câmbio manual e automático no gráfico. É importante investigar esses valores atípicos e considerar a possibilidade de tratá-los adequadamente, seja removendo-os, transformando-os ou aplicando outras técnicas estatísticas, a fim de obter uma análise mais precisa e representativa dos dados.

Outro fator importante na hora de comprar um carro é o ano de fabricação. O ano de fabricação de um veículo pode influenciar significativamente seu preço, condição e desempenho geral. Carros mais novos costumam ter tecnologias mais avançadas, maior eficiência energética e menor desgaste em comparação com carros mais antigos.

Além disso, o ano de fabricação também pode afetar a disponibilidade de peças de reposição e a facilidade de manutenção do veículo. Cada comprador terá suas próprias prioridades e preferências em relação ao ano de fabricação, mas é essencial realizar uma análise cuidadosa e pesquisar informações sobre o modelo específico do carro desejado para tomar uma decisão informada e satisfatória. Na figura 19 é possível observar a relação entre o preço e o ano de fabricação.

Em um dataset de carros usados, é de se esperar que o gráfico da figura 19 tenham preços mais altos com base no tipo de carros possa apresentar variações e não necessariamente um aumento linear com o passar do tempo, como ocorreria em um conjunto de dados de carros novos. Isso acontece porque os carros usados são afetados por uma série de fatores além da inflação e valorização ao longo dos anos. Algumas das principais razões para a variação nos preços de carros usados podem incluir: condição de carro, popularidade do modelo, oferta e demanda regional e ano de fabricação.

Além do câmbio e do ano de fabricação, outro fator importante e decisivo na hora de comprar um carro é o tipo de veículo. As diferentes categorias ou tipos de carro são listados

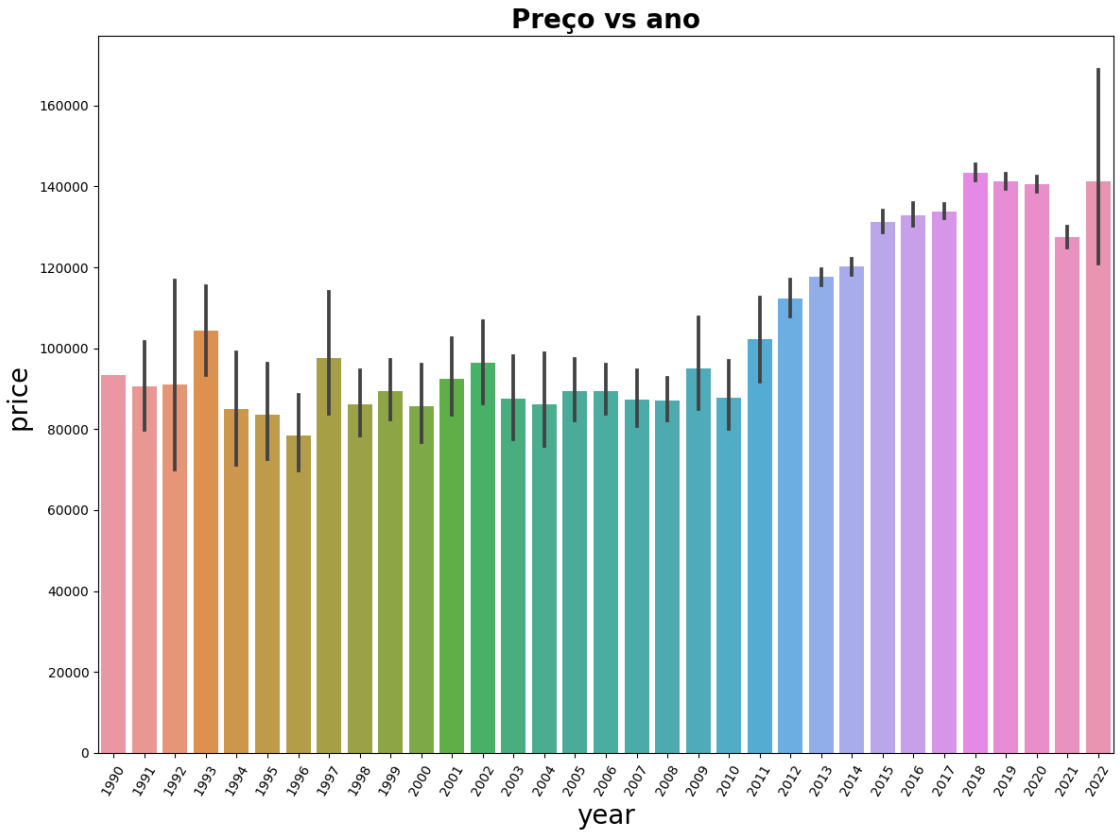

Figure 19: Preço vs ano de fabricação
Fonte: Elaborada pela autora (2023).

nas figuras 20 e 21. Essas categorias podem variar desde carros de passeio tradicionais, SUVs (Utilitários Esportivos), sedãs, hatchbacks, picapes, até veículos esportivos, elétricos ou utilitários.

A escolha do tipo de carro dependerá das necessidades individuais do comprador, estilo de vida, preferências pessoais e propósitos de uso do veículo. Porém, no dataset consta que a maior parte dos carrros são do tipo sedã.

Na figura 22 é uma visualização que mostra a relação entre o preço dos carros e duas variáveis: o tipo de câmbio (por exemplo, manual ou automático) e o tipo de carro (por

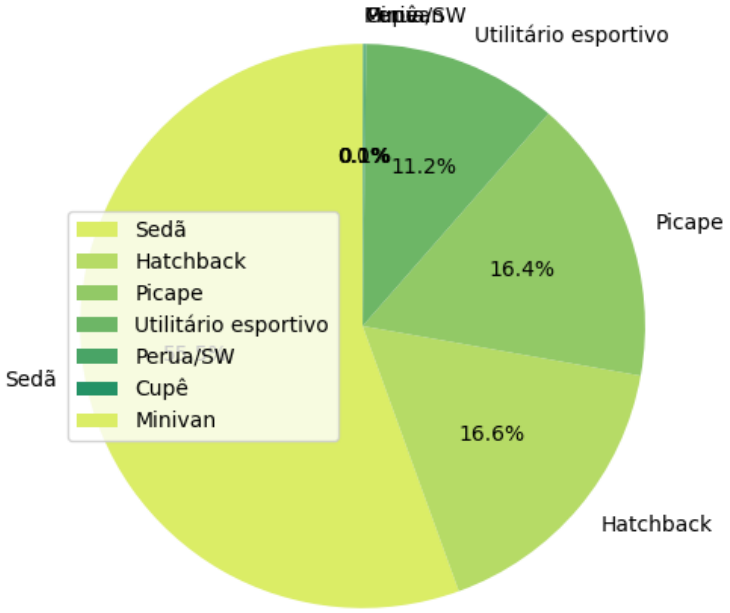

Figure 20: Tipo de veículos
Fonte: Elaborada pela autora (2023).

exemplo, sedã, SUV, hatchback, etc.). É importante frisar que a variável de câmbio "Automática sequencial" continua com os valores baixos. Isso pode ser um indicativo de carros com preços muito discrepantes em relação à maioria dos outros. O mesmo gráfico foi plotado em relação aos estados brasileiros e se encontram na figura 23.

Também foi criado um gráfico mostrando os 5 tipos de carros na figura 24. Nesse gráfico específico, os tipos de carros estão representados no eixo x, enquanto os valores são plotados no eixo y. Cada barra vertical ou ponto no gráfico corresponde a um tipo de carro específico e indica o valor do preço mais alto observado para esse tipo. Essa visualização é útil para identificar os tipos de carros que têm os preços mais altos em seu conjunto de dados. Ao analisar o gráfico, pode-se facilmente identificar quais tipos de carros têm valores

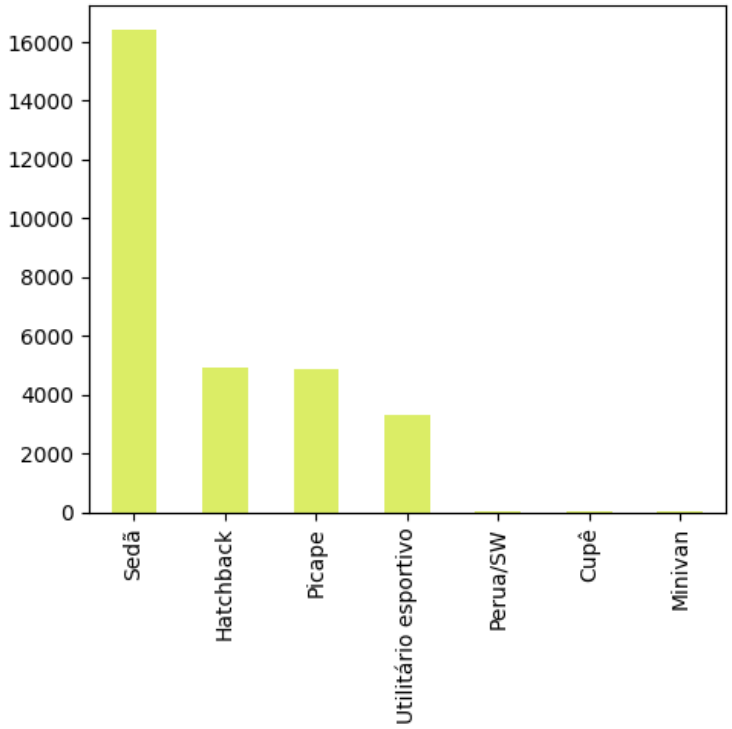

Figure 21: Tipo de veículos
Fonte: Elaborada pela autora (2023).

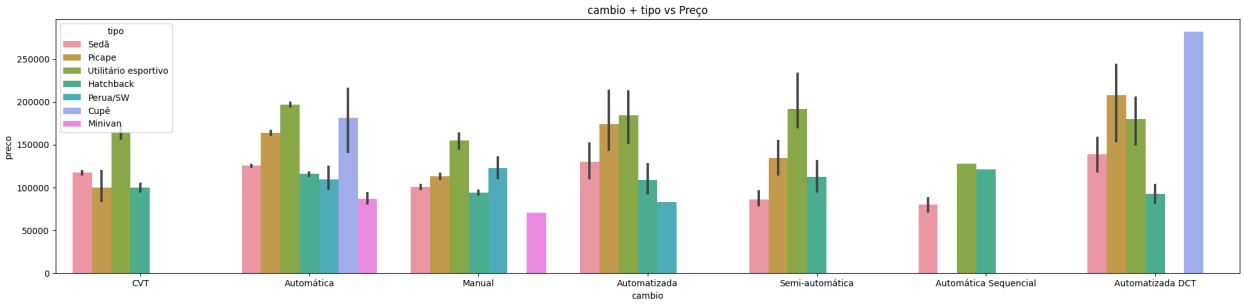

Figure 22: cambio + tipo vs Preço
Fonte: Elaborada pela autora (2023).

extremamente elevados em relação aos outros.

E no gráfico 25 foi plotado o tipo de carro mais o tipo de câmbio em relação ao preço.

Para este relatório foi levantando três questionamentos inicais que são:

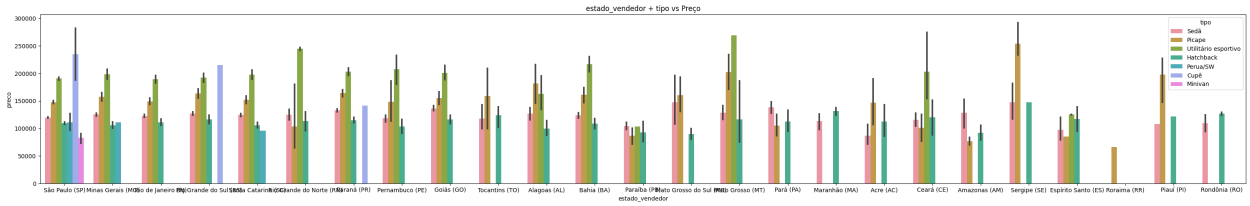

Figure 23: Estados + tipo vs Preço
Fonte: Elaborada pela autora (2023).

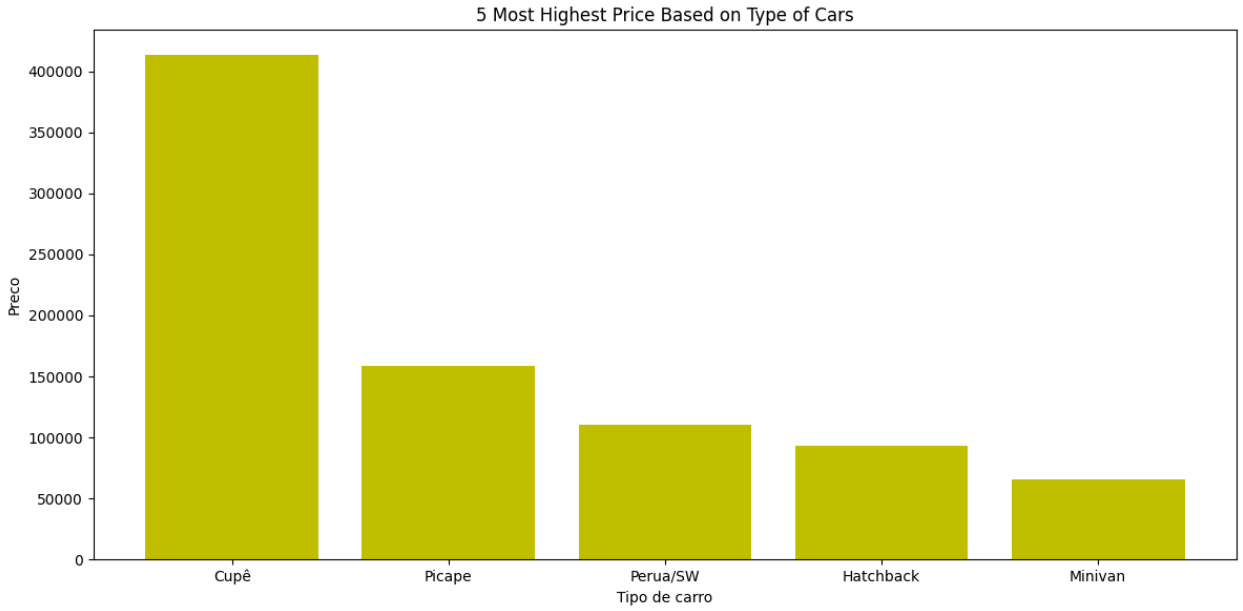

Figure 24: Tipos de carros vs preço
Fonte: Elaborada pela autora (2023).

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

- 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?

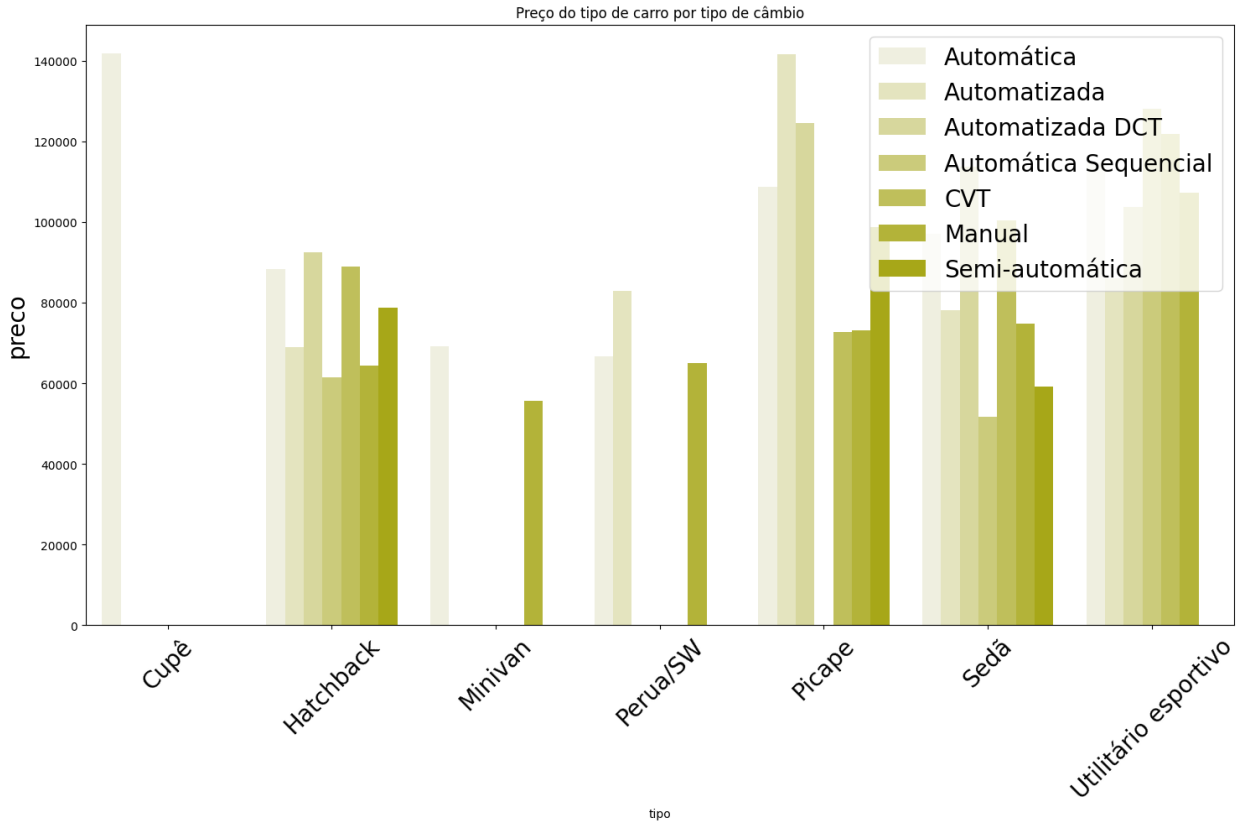

Figure 25: 5 preços mais altos com base no tipo de carroso
Fonte: Elaborada pela autora (2023).

Primeiramente, é importante ressaltar que a definição de "carros populares" pode variar dependendo do contexto e do país. No Brasil, a definição tradicional de carros populares refere-se a veículos com preços acessíveis, que geralmente são mais vendidos e populares no mercado nacional. Esses carros costumam ter características como menor preço, consumo de combustível mais econômico e menor custo de manutenção.

Anteriormente, foram consideradas as 5 marcas que mais aparecem no dataset, que são Volkswagen, Chevrolet, Toyota, Hyundai e Jeep. No entanto, é válido destacar que a marca Jeep, por exemplo, é conhecida por produzir veículos com preços mais elevados e não é

comumente associada a carros populares no Brasil. Para definir de forma mais precisa quais marcas podem ser consideradas como "carros populares", seria necessário realizar uma pesquisa mais abrangente e analisar critérios específicos, como faixa de preço, motorização, consumo de combustível e outras características relevantes para o contexto brasileiro.

O estado de São Paulo que contém o maior número de carros. Para esse mesmo estado, foi constado que a marca Chevrolet é o que tem maior número de carros disponíveis, seguido da Volkswagen, fiat, toyota e Hyundai. Portnao, o estado de São Paulo é o melhor estado para comprar um carro de marca popular. O melhor estado para se comprar uma picape com transmissão automática também é em São Paulo, pois foi o estado com o maior número de picapes com câmbio automática do dataset. O estado de São Paulo é o que tem a maior quantidade de carros com garantia de fábrica, onde constam um total de 2307, onde 1203 sao do tipo seda. O segundo estado com melhor garantia é o Parana com 212 carros tipo sedã. Os piores estado para comprar uma Picape com garantia são os estados do Acre, Espírito Santo e o Tocantins que só tem 1 carro disponivel em cada estado com garantia de fábrica.

4 Modelo

Para essa estapa foi utilizado o site Kaggle para a implementação do dataset através da biblioteca Scikit-learn. Neste contexto, a regressão linear é uma técnica estatística utilizada para modelar a relação entre uma variável dependente (a variável que queremos prever) e uma ou mais variáveis independentes (as variáveis que utilizamos para fazer a previsão). O link do código está disponível em: sklearn.

Para essa etapa foram utilizados 3 diferentes classes:

- LinearRegression: É uma classe que implementa o modelo de regressão linear padrão.

- Ridge: É uma classe que implementa a regressão linear com regularização L2 (também conhecida como regularização Ridge).
- Lasso: É uma classe que implementa a regressão linear com regularização L1 (regularização Lasso).

Nesta fase do processo, as colunas "id" e "num_fotos" foram excluídas e, para o pré-processamento, também foram removidas as colunas que continham valores ausentes (NaN). Essa escolha foi motivada pela falta de tempo disponível para realizar um pré-processamento e uma análise exploratória de dados completa.

Ao remover as colunas "id" e "num_fotos", a intenção foi para eliminar informações que não são relevantes para a análise ou modelagem dos dados. A coluna "id" geralmente é um identificador único para cada registro, e a coluna "num_fotos" pode representar o número de fotos associadas a cada registro, mas essas informações podem não ser relevantes para a análise pretendida.

A exclusão das colunas com valores ausentes (NaN) é uma prática comum para simplificar o pré-processamento dos dados. Os valores ausentes podem causar problemas durante a análise e modelagem, e lidar com eles exige tempo e cuidado para escolher a abordagem mais adequada, como preenchimento com valores médios, moda ou outras técnicas de imputação. Ao eliminar as colunas com valores ausentes, essa etapa foi agilizada, embora seja importante ter em mente que a eliminação de dados pode levar à perda de informações. No gráfico 26 é possível identificar as váriaveis que contém NaN no dataset fornecido pela empresa.

É importante lembrar que a decisão de remover colunas ou lidar com valores ausentes deve ser tomada com base no contexto específico do projeto e nos objetivos da análise. Em situações com restrições de tempo, é compreensível que algumas etapas mais complexas sejam

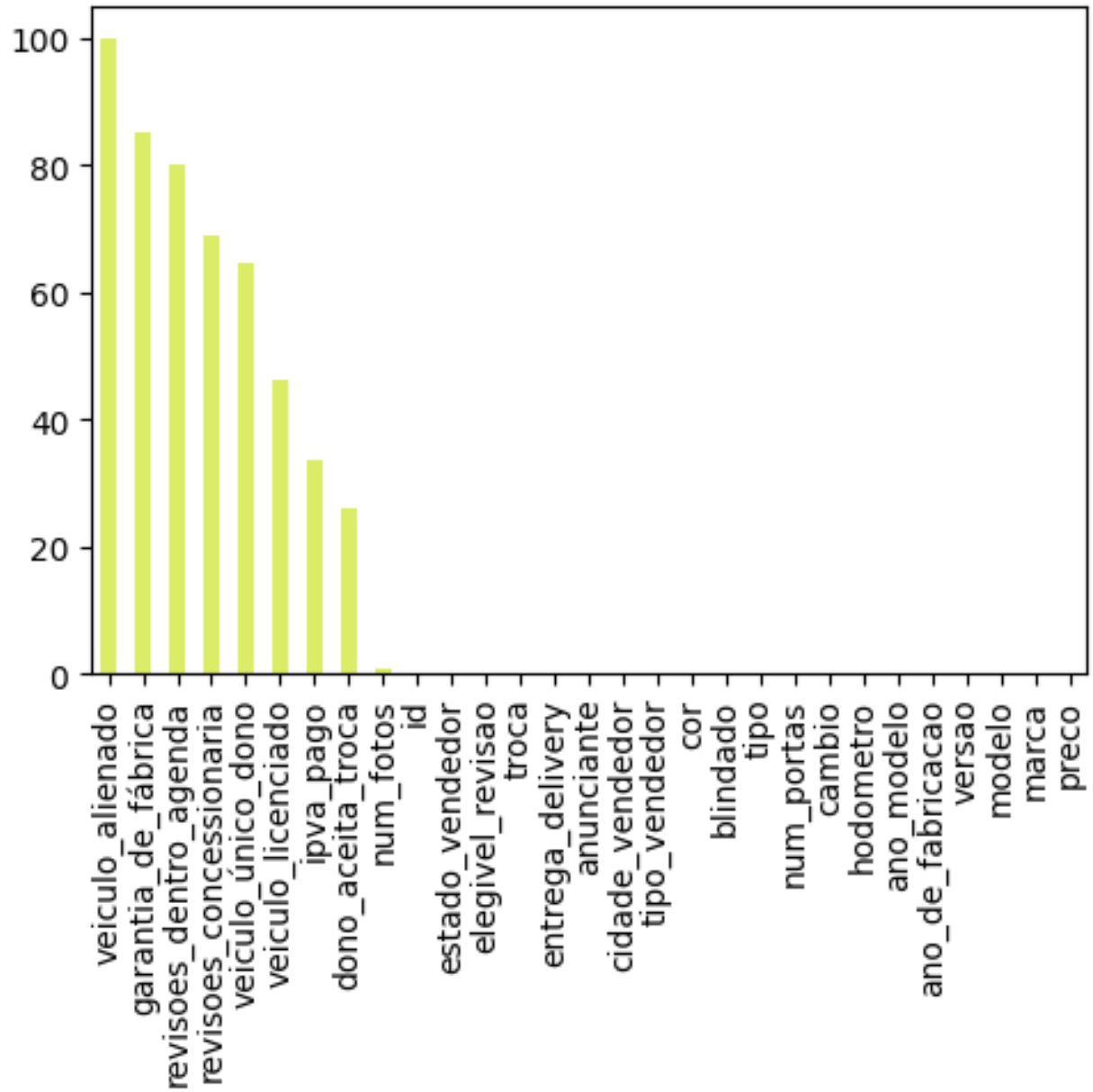

Figure 26: Variáveis que contém Nan
Fonte: Elaborada pela autora (2023).

simplificadas, mas é fundamental ter consciência das implicações dessa escolha na qualidade e robustez dos resultados finais.

As variáveis excluidas foram: dono_aceita_troca, veiculo_único_dono, revisoes_concessionaria, ipva_pago, veiculo_licenciado, garantia_de _fábrica, revisoes_dentro_agenda e veiculo_alienado. Restaram 19 variáveis.

4.1 Linear Regression

O dataset foi dividido em variáveis dependentes (y) e independentes (X). A variavel dependente é a nossa targert, e nesse caso queremos achar o preço. Também foi dividio os dados em recursos numéricos e categóricos separadamente.

Para a primeira classe *LinearRegression() Model* obteve o valor de:

- Training Score:0.73
- Testing Score:-11781049716707925054392893 44.00

Os "Scores" da classe LinearRegression() se referem à avaliação do desempenho do modelo de regressão linear em dados de treinamento e teste. Os Scores são valores que indicam o quão bem o modelo se ajustou aos dados e quão bem ele generaliza para dados desconhecidos.

O Training Score é um indicador de quão bem o modelo se ajustou aos dados de treinamento. Ele é calculado com base nos dados utilizados para treinar o modelo. Um valor de Training Score mais próximo de 1 indica que o modelo se ajustou bem aos dados de treinamento. O Training Score é de 0.73, o que indica que o modelo teve um desempenho razoavelmente bom ao se ajustar aos dados de treinamento. Para o Testing Score é o indicador de quão bem o modelo generaliza para dados desconhecidos (ou seja, dados que não foram utilizados durante o treinamento do modelo). Ele é calculado com base nos dados de teste, que são dados que o modelo nunca viu antes. Um valor de Testing Score negativo (como no seu exemplo) é incomum e geralmente indica que o modelo está tendo um desempenho

muito ruim na generalização para os dados de teste. Isso pode ocorrer quando o modelo está sofrendo de overfitting.

As métricas para essa clase são:

- Mean Squared Error:747469765059042485276745837665111244 8.00
- Root Mean Squared Error:27339893289093913 60.00
- Mean Absolute Error:172001793492287936.00
- r2_score:-11781049716707925054392893 44.00

O gráfico dos coeficientes do modelo de regressão linear é uma visualização na figura 27 e mostra os valores dos coeficientes atribuídos a cada variável independente do modelo. Os coeficientes representam as inclinações (ou pesos) que as variáveis independentes têm na previsão da variável dependente do modelo.

4.2 Ridge

A classe LinearRegression() não nos permite controlar sua complexidade, então é muito provável que superajuste os modelos quando o conjunto de dados for relativamente pequeno. Para a segunda classe foi utilizada a Ridge, que utiliza a regularização l2. As técnicas de regularização restringem explicitamente um modelo para evitar o overfitting.

Os resultados obtidos para essa classe foram:

- Training Score:0.78
- Testing Score:0.71

E para as métricas:

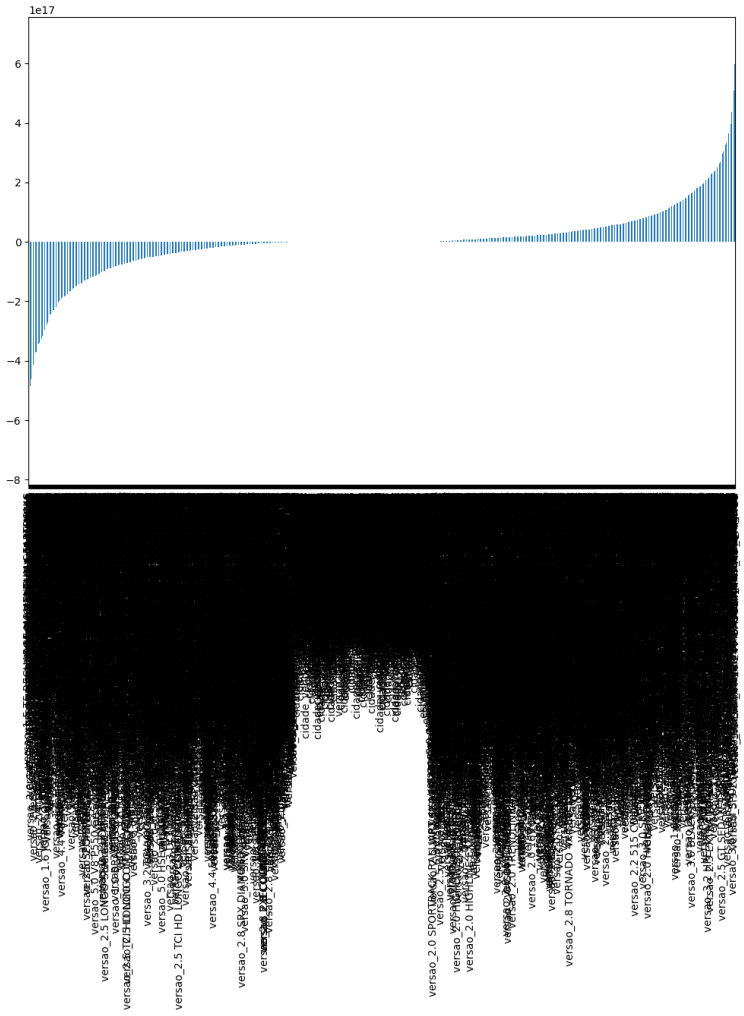

Figure 27: Gráfico modelo de regressão linear
Fonte: Elaborada pela autora (2023).

- Mean Squared Error:1845779827.64
- Root Mean Squared Error:42962.54
- Mean Absolute Error:28270.77
- r2 _score:0.71

A visualização do gráfico da classe Ridge está na figura 28.

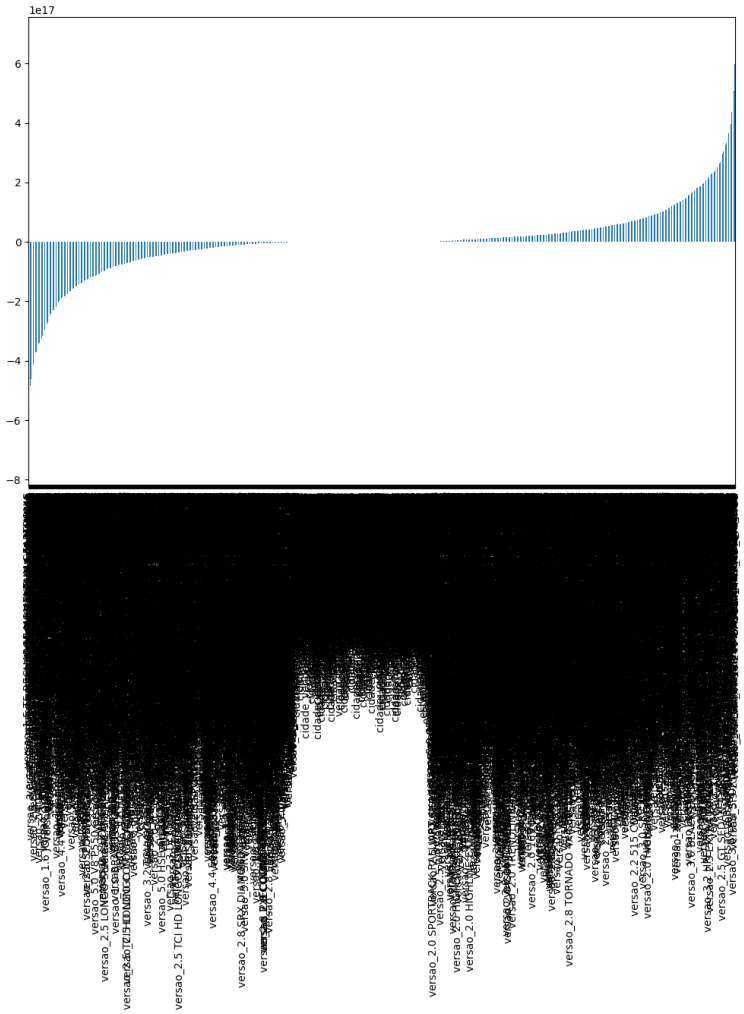

Figure 28: Gráfico modelo Ridge
Fonte: Elaborada pela autora (2023).

4.3 Lasso

A função Lasso É um modelo linear que utiliza a técnica de regularização L1. Essa regularização reduz a magnitude do coeficiente, no entanto, ao contrário de Ridge, reduz a magnitude de alguns dos recursos a zero. Por isso, negligencia alguns dos recursos comple-

tamente.

Os resultados obtidos para essa classe foram e o gráfico está na figura 29.

- Training Score:0.78
- Testing Score:0.71

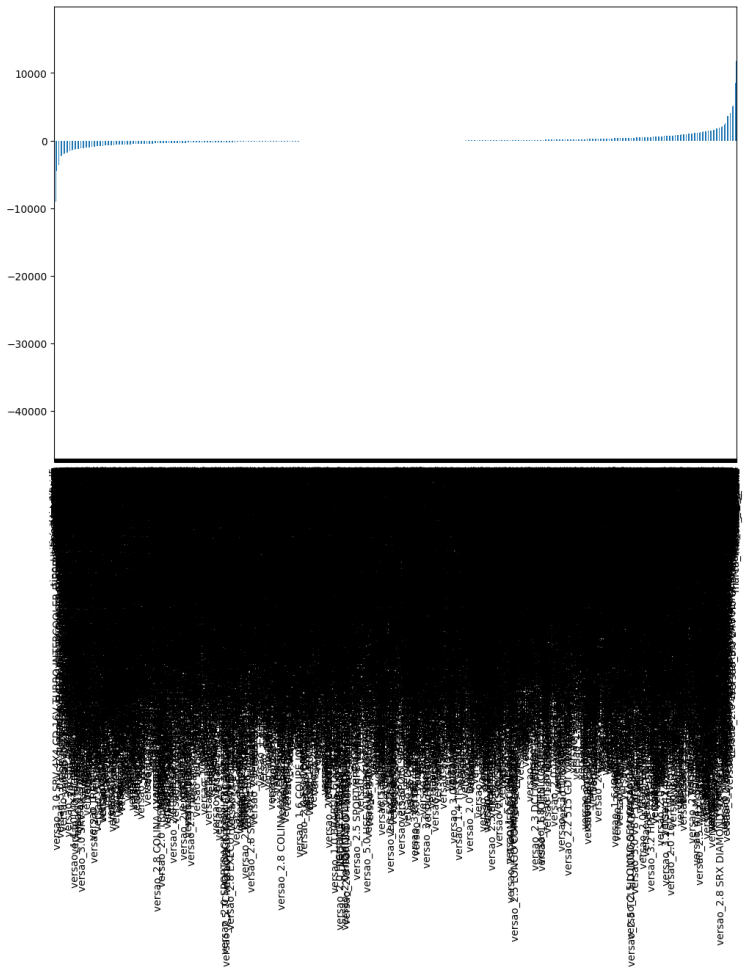

Figure 29: Gráfico modelo Lasso
Fonte: Elaborada pela autora (2023).

4.4 Comparação dos três modelos

A comparação entre os três Modelos Lineares estão plotados na figura 30. A partir do gráfico abaixo comparando o coeficiente de recursos independentes, fica claro que o modelo LinearRegression tem a maioria dos coeficientes diferentes de zero e são de grande magnitude e a maioria de seus valores estão fora de y-lim, que são representados por blocos quadrados azuis.

Já o modelo Ridge e Lasso a maioria de seus valores estão na linha horizontal e poucos estão muito próximos da linha horizontal, devido à sua menor magnitude representada pelo laranja e verde, respectivamente.

É importante ressaltar que, para uma comparação mais consistente, o ideal seria aplicar pelo menos três técnicas diferentes de machine learning e, em seguida, realizar uma comparação entre elas. Cada técnica pode levar a resultados diferentes e pode haver uma variação nos coeficientes atribuídos às variáveis independentes.

Após selecionar a técnica que obtém o melhor valor, é possível realizar o refinamento através da otimização dos hiperparâmetros do modelo. Os hiperparâmetros são configurações que afetam o desempenho e a capacidade de generalização do modelo. A busca pelos melhores hiperparâmetros visa melhorar ainda mais o desempenho do modelo e encontrar uma configuração que se ajuste melhor aos dados.

A comparação entre diferentes técnicas e a otimização dos hiperparâmetros são etapas cruciais na construção de um modelo eficiente e preciso. Essas práticas ajudam a garantir que o modelo selecionado seja o mais adequado para resolver o problema em questão e possa oferecer resultados confiáveis e robustos.

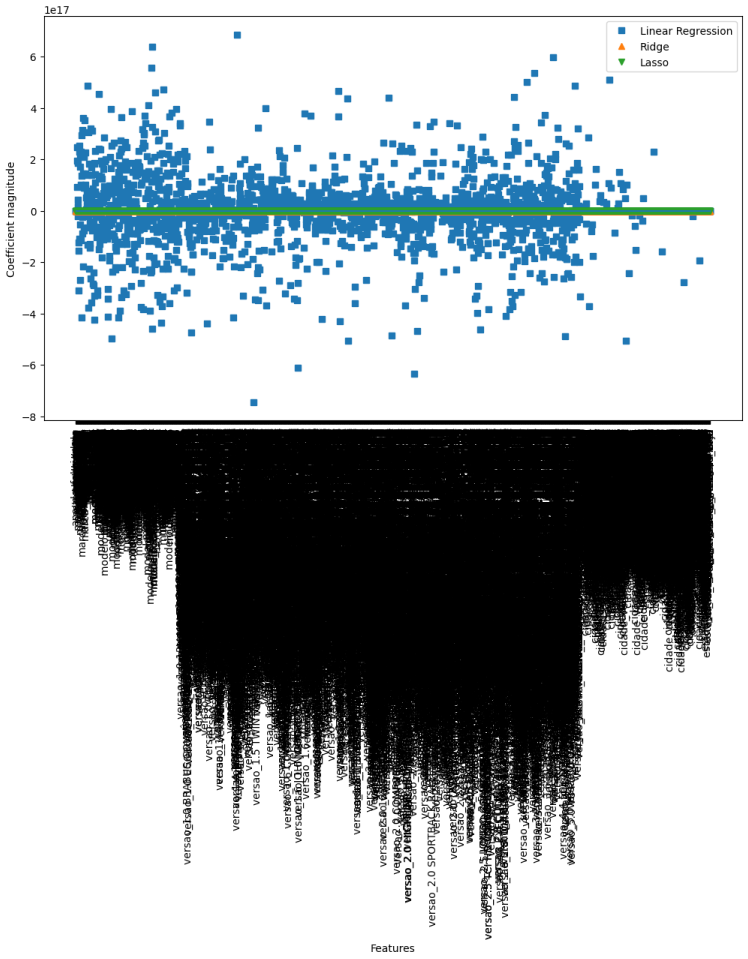

Figure 30: Comparação dos três modelos
Fonte: Elaborada pela autora (2023).

5 Conclusão

Neste relatório, foi realizada uma análise detalhada do dataset fornecido pela empresa Indicium, que contém variáveis relacionadas à venda de carros. O objetivo dessa análise foi extrair insights significativos e compreender melhor os padrões e tendências presentes nos dados.

Foram exploradas várias variáveis relevantes no dataset, incluindo informações sobre os

carros, como marca, modelo, ano de fabricação, tipo de câmbio e tipo de carro. Além disso, foram analisados os preços dos carros e outras características relevantes que poderiam influenciar nas vendas. Foram utilizadas diversas técnicas estatísticas e ferramentas de visualização de dados para realizar essa análise. Gráficos e outras representações visuais foram empregados para tornar os resultados mais compreensíveis e acessíveis.

Ao longo do relatório, foram apresentadas conclusões sobre os insights obtidos a partir dos dados, destacando aspectos importantes para o mercado de venda de carros e oferecendo sugestões e recomendações com base nos resultados encontrados. O relatório visa fornecer informações valiosas que possam auxiliar a empresa Indicium a tomar decisões mais informadas e estratégicas no setor de venda de carros, bem como para aprimorar suas estratégias de marketing e vendas.

Todo o processo de análise foi conduzido com rigor e precisão, garantindo a qualidade dos resultados obtidos. No entanto, devido à falta de conhecimento específico na área de programação por parte da redatora deste trabalho, é recomendável que um especialista com experiência e expertise na área valide os resultados do dataset. A validação dos resultados por um profissional especializado é fundamental para assegurar a precisão das conclusões e a correta interpretação dos dados. Esse especialista poderá identificar possíveis erros ou viéses na análise, bem como realizar verificações adicionais para confirmar a confiabilidade dos resultados.

Outros fatores importantes para garantir um bom resultado incluem uma análise cuidadosa das variáveis que foram excluídas devido a valores NaN. Embora excluir variáveis seja uma solução rápida, é essencial considerar que essa abordagem pode resultar na perda de informações relevantes para o modelo.

Em vez disso, é recomendável adotar um tratamento mais cuidadoso dos valores ausentes,

como preenchê-los com estatísticas adequadas ou utilizar outras técnicas de imputação de dados. Dessa forma, é possível preservar a integridade e a representatividade do conjunto de dados para o modelo de machine learning.

Implementar diferentes modelos de machine learning e fazer uma comparação entre eles é fundamental para interpretar os resultados. A análise dos coeficientes do modelo e a comparação entre diferentes técnicas de machine learning são etapas essenciais no processo de construção e refinamento de modelos. Essa abordagem sistemática ajuda a garantir que o modelo final esteja bem ajustado aos dados e seja capaz de fazer previsões precisas em novos dados.

É importante reconhecer que cada abordagem possui suas próprias vantagens e limitações, e a escolha do método mais adequado depende das características específicas dos dados. As considerações sobre a disponibilidade e qualidade dos dados, a complexidade do problema e os recursos computacionais disponíveis também são relevantes ao selecionar o modelo mais apropriado.

No entanto, é compreensível que essas técnicas mais avançadas podem requerer um estudo mais extenso e uma abordagem mais complexa, o que está fora do escopo deste trabalho. Portanto, é importante reconhecer as limitações e ajustar as análises conforme a disponibilidade de recursos e conhecimento na área de machine learning.

Um fato relevante desta conclusão é que, devido à limitação de tempo, não foi possível realizar pesquisas extensas para um levantamento teórico completo sobre o problema em questão. Além disso, uma análise mais profunda do mercado de carros seria necessária para obter uma compreensão mais abrangente do assunto.

Além disso, o processo de implementação de todos os algoritmos selecionados, seguindo uma abordagem científica, também requereria um tempo substancial. É importante recon-

hecer que a pesquisa e implementação de algoritmos de machine learning demandam rigor, experimentação e validação cuidadosa para obter resultados confiáveis e significativos.

Embora essas limitações possam afetar a amplitude e profundidade dos resultados, é essencial ressaltar que a análise realizada até o momento já proporcionou insights valiosos e informações úteis para a compreensão do dataset em questão.

References

[1] Dibya Ranjan Das Adhikary, Ronit Sahu, and Sthita Pragyna Panda. "Prediction of used car prices using machine learning". In: *Biologically Inspired Techniques in Many Criteria Decision Making: Proceedings of BITMDM 2021*. Springer, 2022, pp. 131–140.