Análise Preliminar e Pré-processamento na Machine Learning

Ernesto Gurgel Valente Neto, Nome da Professora

[1]Pós-graduando em Ciência de Dados – Centro Universitário Farias Brito
Caixa Postal 607175-230 – Ceará – CE – Brasil.

[2]D.ra Ananda Freire Mestre e Doutora Engenharia de Teleinformática – Local da Orientadora – Ceará – CE – Brasil.

{[1]gurgelvalente@gmail.com, [2]anandalf@gmail.com}

Abstract. *This article describes the methodologies applied to data analysis, observation techniques and application of data transformations on a dataset of vehicle sales. The goal is to present knowledge acquired in the graduation of Data Science and related impacts on the observations made on the dataset and Machine Learning model, as well as express the results obtained from the observations.*
Keywords: *Data Science; Data Analysis; Data Transformations; Machine learning.*

Resumo. *O presente descreve as metodologias aplicadas à análise de dados, técnicas de observação e aplicação de transformações de dados sobre um conjunto de dados de venda de veículos. Seu objetivo é apresentar os conhecimentos estudados na graduação de Ciência de Dados e seus impactos relacionados a partir das observações feitas sobre o conjunto de dados e modelo de Machine Learning, bem como expressar os resultados obtidos pelas observações.*
Palavras-chave: *Ciência de Dados, Análise de Dados, Transformações de dados, Machine learning.*

INTRODUÇÃO

No desenvolvimento deste artigo, são empregados técnicas e conhecimentos para o pré-processamento, análise e desenvolvimento de um modelo capaz de predizer valores de veículos.

Na última década a quantidade de dados e informação produzida pela humanidade aumentou consideravelmente em decorrência, da humanidade e da conectividade, — "como resultado, 2,5 quintilhões de bytes de dados são gerados diariamente, oriundo de imagens digitais, vídeos, sensores inteligentes, transações eletrônicas, sinais de GPS, entre outros" (EATON; et al, 2012).

Avaliando a importância e impacto dos dados na atualidade, pode-se pressupor que a humanidade, nunca antes na história, formalizou uma quantidade tão diversa de dados, sendo ambas oriundas de fontes diversificadas, como trabalhos, catalogações, vendas, referências, fatos acadêmicos e científicos, comportamento humano, natural, mídia, áudio, literatura etc. Com a conectividade e o acesso à informação, esses números criaram verdadeiros nichos de dados e fatos a serem explorados e transformados em conhecimento; esse conjunto crescente de informações permitiu ao cientista de dados analisar comportamentos, padrões, bem como perfis de maneira cada vez mais precisa, criando mapas tais quais quando aplicamos a metodologias de Ciência Dados para desenvolver modelos de *Machine Learning* capazes de compreender ou expressar fatos da nossa sociedade e comportamento. Na literatura atual é notável que cientistas de dados aplicam abordagens voltadas à análise de dados, tratamentos e técnicas como ferramentas para aplicação de *Machine Learning* e *Deep Learning* dessa forma, este artigo visa a análise e aplicação dessas mesmas abordagens de maneira comparativa.

Este artigo visa utilizar um *dataset* para experimentação de técnicas de visualização de dados, análise de *outliers*, estatísticas e conhecimentos estudados na literatura de Ciência de Dados para criar um modelo de *Machine Learning* que seja capaz de predizer valores finais de preços dos bens de serviço observados, considerados pela literatura como um problema de regressão, definido quando precisamos prever determinados valores. Em cada etapa, serão documentadas as observações e os conhecimentos adquiridos durante o desenvolvimento deste artigo, assim como o impacto de *outliers* nos modelos de regressão aplicados.

1. INTRODUÇÃO METODOLÓGICA

As etapas descritas a seguir envolvem desenvolvimento em ciência de dados e abordam a necessidade da compreensão da medição e a avaliação da modelagem e internalização do conhecimento. Desse modo, o cientista de dados explora e internaliza informações a respeito do que é avaliado, identificando relações, classificando dados existentes e observando seus comportamentos, assim possibilitando o teste de hipóteses e o desenvolvimento de modelos de *Machine Learning* apropriados.

Em linhas gerais, a análise de dados é um conjunto de atividades que devem ser desempenhadas, desde a seleção dos dados até a produção do conhecimento, que é o principal produto da análise. A análise de dados envolve o processamento de coleções de objetos em busca de padrões consistentes, de forma a detectar relacionamentos sistemáticos entre variáveis componentes desses objetos e gerar conhecimento não facilmente detectado. Dá-se o nome de processo de análise de dados à especificação do encadeamento desse conjunto de atividades. As atividades que compõem o processo de análise de dados podem ser organizadas em quatro etapas: seleção, pré-processamento, métodos de análise e avaliação (Han et al., 2006).

A linguagem de programação selecionada foi *Python*, por contar com um conjunto de bibliotecas que facilita o desenvolvimento e análise do modelo de *Machine Learning* a ser desenvolvido. Em seguida, foi realizada uma pesquisa em conjuntos diversos de *datasets* que pudessem expressar uma problemática que poderia ser selecionada para ser resolvida, como a previsão do valor final de um produto. Nessa etapa, foram selecionados inicialmente *datasets* diversos da base de dados do *Kaggle*, enquanto se analisava o tamanho da base de dados e a diversidade de informações disponíveis que pudessem ser levadas em consideração para análise.

Definidos os dados trabalhados, passaram a ser importados conjuntos de bibliotecas em *Python* como facilitador da análise, permitindo que os dados sejam internalizados e o conhecimento sobre as características do *dataset* estudado. Durante a exploração, antes de ser realizada qualquer tentativa de tratamento na base de dados, foram catalogados e expressos numericamente e graficamente diversas perspectivas que pudessem aumentar a compreensão sobre o conjunto de dados, tais quais como a contagem de marcas, modelos e sua distribuição em relação ao todo, a quilometragem agrupada em faixas, tipos de combustíveis utilizados, cidades em que as vendas dos veículos eram realizadas, contagem dos anos de veículos vendidos e modelos de cada veículo.

O próximo fator foram as análises, expressas graficamente sobre quantidades dos itens correspondentes de cada variável da base de dados, visibilizando relações entre as variáveis e impacto do preço de cada veículo. Durante essa etapa, foi analisada a *Correlação de Pearson* para verificar a sensibilidade e importância de cada uma das variáveis observadas. Em posse de uma compreensão possibilitada pelas informações

disponíveis, e estabelecido o objetivo de previsão dos valores finais dos veículos, foi aplicada uma avaliação de *outliers* sobre o conjunto da base de dados. Também foram aplicadas contagens e avaliação dos impactos da remoção dos mesmos, através de técnicas de visualização e cruzamento das mesmas com valores, que são analisadas nesse momento, e revisadas durante todo o processo. A necessidade da aplicação de técnicas de tratativa de "anomalias", assim como o conhecimento adquirido durante a fase de observação, permitem mensurar a necessidade de substituição dos *outliers* por valores médios, agrupamentos, remoção, ou outras que aproximem mais os dados.

2. METODOLÓGICA

O conhecimento parte de uma metodologia amplamente utilizada e observada em meios acadêmicos, artigos científicos, que busca compreender melhor as coleções de objetos da base de dados, os principais campos de análise, aplicando técnicas de pré-processamento para desenvolver conhecimento que possa agregar valor à análise das informações e, assim, criar um modelo de *Machine Learning* que possa solucionar a problemática de predição de valores.

2.1 EXPLORAÇÃO DA INFORMAÇÃO

Primeiramente, parte do trabalho desse artigo consistia na análise dos dados e nos conjuntos que representavam o problema. Os campos da *database* utilizados na análise podem ser visualizados a seguir.

Coluna	Tamanho	Tipo de Dados
Unnamed	117927	non-null int64
generation name	87842	non-null object
year	117927	non-null int64
mark	117927	non-null object
model	117927	non-null object
vol_engine	117927	non-null int64
city	117927	non-null object
price	117927	non-null int64
province	117927	non-null object

fuel	117927	non-null object
mileage	117927	non-null int64

Tabela 1: Tabela df.info de dados

A próxima etapa corresponde à análise de campos com valores *null* ao qual podem não agregar nenhum valor no conjunto de dados, ou a exploração dos dados que indiquem necessidade de avaliação para tratamentos ou campos com nenhuma importância. Nessa etapa, os primeiros campos analisados e indicados são "*Unnamed*" e "*generation_name*", respectivamente. O primeiro representa um índice da base de dados; o segundo, a concatenação do texto da geração do veículo e o intervalo possível entre fabricação do modelo, que possui uma quantidade significativa de valores *null* [30.085 de 117.927], representando 25,5% da amostra. Assim, ambos foram escolhidos para remoção.

Conseguinte à relação da distribuição das quantidades únicas de dados de cada objeto da base de dados, onde os nomes foram respectivamente renomeados seguindo o padrão anterior em tradução livre, os mesmos podem ser observados a seguir.

Coluna	**Distribuição Única**
Distribuição da Marca	23
Distribuição do Modelo	328
Distribuição do Ano	54
Distribuição da Quilometragem	35394
Distribuição do Motor	508
Distribuição do Combustível	6
Distribuição da Cidade	4427
Distribuição do Estado	23
Distribuição dos Preços	9310

Tabela 2: Distribuição de Dados

Uma análise mais detalhada para exploração da representação das unidades únicas é realizada a seguir, exemplificando alguns dos campos explorados para maior análise e entendimento dos dados ao qual se tratava.

- Modelos de Marcas: ['opel' 'audi' 'bmw' 'volkswagen' 'ford' 'mercedes-benz' 'renault' 'toyota' 'skoda' 'alfa-romeo' 'chevrolet' 'citroen' 'fiat' 'honda' 'hyundai' 'kia' 'mazda' 'mini' 'mitsubishi' 'nissan' 'peugeot' 'seat' 'volvo'];
- Modelos de carros: ['combo' 'vectra' 'adam' 'agila' 'ampera' 'antara' 'astra' 'corsa' 'crossland-x' 'frontera' 'grandland-x' 'insignia' 'vivaro' 'zafira' 'a3' 'karl' 'meriva' 'mokka' 'omega' 'signum' 'tigra' '80' 'a1' 'a2' 'a4', …., etc.];
- Modelos do Motor: [1248 1499 1598 1400 1368 1600 1799 1796 1994 1998 2498 1995 2198 1910 2171 3175 2792 1597 2958 3000 1398 1364 999 1229 1200 1100 996 1242 1000 1199 973 998 0 1991 2231 3195 2405 2200 2000 2400 2384 1389, …., etc.];
- Tipo de Combustível: ['Diesel' 'CNG' 'Gasoline' 'LPG' 'Hybrid' 'Electric'];
- Cidade: ['Janki' 'Katowice' 'Brzeg' 'Kolonia Gościeńczyce' 'Augustówek' 'Bledzew', …., etc.];
- Estado: ['Mazowieckie' 'Śląskie' 'Opolskie' 'Dolnośląskie' 'Lubelskie' 'Wielkopolskie' 'Warmińsko-mazurskie' 'Małopolskie' 'Podkarpackie' 'Kujawsko-pomorskie' 'Pomorskie' 'Podlaskie' 'Łódzkie' 'Świętokrzyskie' 'Zachodniopomorskie' 'Lubuskie' 'Berlin' 'Wiedeń' 'Niedersachsen' 'Moravian-Silesian Region' '(' 'Trenczyn' 'Nordrhein-Westfalen'].

2.2 VISUALIZAÇÃO DOS CONJUNTOS DE DADOS

A seguir são utilizadas técnicas de visualização de dados para comparação da disposição dos anos no *dataset*. Nesse momento, torna-se necessária a observação das relações dos objetos estudados, porém somente alguns serão demonstrados a seguir:

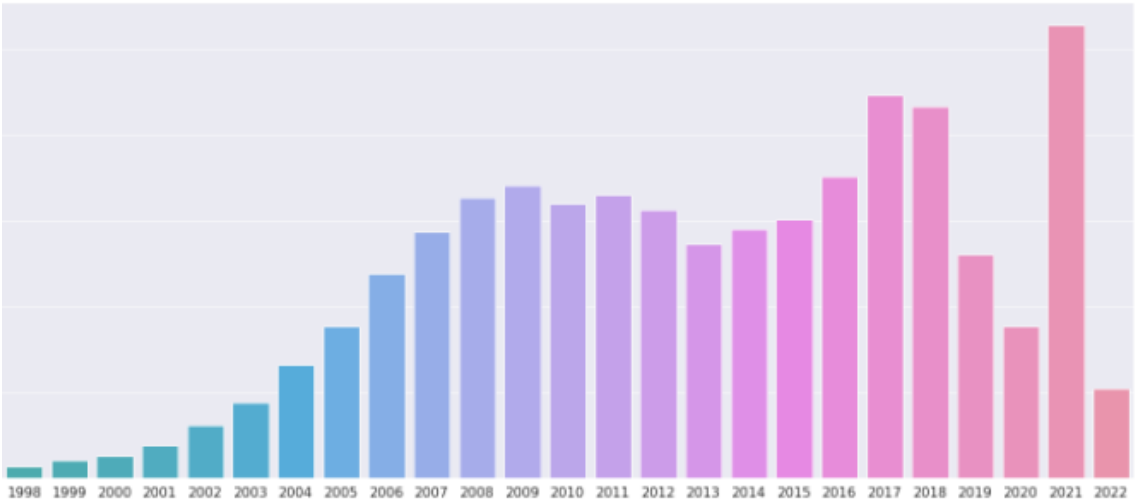

Figura 1: Contagem das quantidades de veículos de cada ano

A contagem de concentração em anos revela que do conjunto de dados iniciado em 1945 até o ano de 2022, apresenta os índices mais altos de concentração situados a partir do ano de 2002, enquanto os anos anteriores demonstram menor índice de

representatividade e permite a especulação da possibilidade de se criar um agrupamento para o conjunto de dados. Foram observados 2.472 valores abaixo do ano de 2002 de 117.927 representando 0,02% da amostra.

O histograma a seguir foi desenvolvido para demonstrar a representatividade dos tipos de quilometragem acumuladas em subconjuntos numéricos para expressar seus valores e quantificá-los.

```
df["mileage_hist"] = pd.cut(df["mileage"],
                            bins=[0., 25000, 50000, 75000, 100000, 2000000.],
                            labels=[1, 2, 3, 4, 5])
```

Tabela 3: Código do histograma Quilometragem

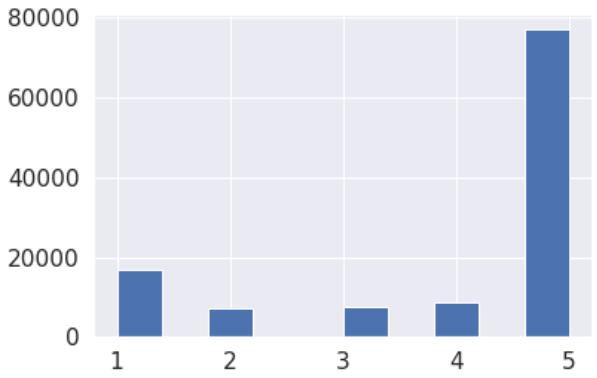

Figura 2: Quantificação de Quilometragem

Nesse conjunto, é possível observar que os grupos 3, 4 e 5 apresentam maior representatividade das quilometragens observadas na base de dados, porém, ainda assim, são necessárias maiores observações e experimentações antes de tirar quaisquer conclusões sobre quais tratamentos deveriam ser utilizados. Observa-se que a distribuição apresentada na figura 2 indica que a base de dados analisada possui veículos seminovos e usados, como pode ser observado nas faixas de agrupamento de quilometragem.

A próxima etapa de análise, apresenta a quantidade de cada tipo de combustível disponível na base de dados, na qual podem ser observadas certas predominâncias e preferências de carros consumidos. Os dados respectivamente representam *Diesel*, *CNG* (Gás Natural), Gasolina, *LPG* (Gás Liquefeito de Petróleo), *Hybrido* (Combinação Diesel/Gasolina) e Elétrico.

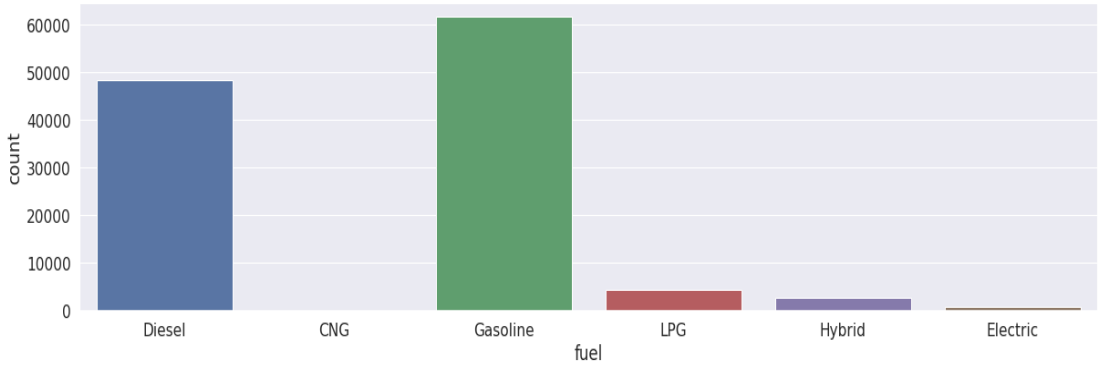

Figura 3: Quantificação de Tipos de Combustível

Com base na figura 3, pode-se afirmar que existem três principais categorias de combustíveis a serem indicados, *Gasoline*, *Diesel* e Outros (Subgrupo representado por *CNG*, *LPG*, *Hybrid* e *Eletric*). Nesse momento, é realizada a observação da possível necessidade de agrupamentos desses subconjuntos de dados, porém ainda existe a necessidade de avaliação dos primeiros modelos antes de validar qualquer hipótese, indicando até o momento a necessidade como inconclusiva. Os valores a seguir são indicativos aproximados dos números.

1. 48476 *Diesel* de 117927, representando 41,1% da amostra.
2. 47 *CNG* de 117927, representando 0,0003% da amostra.
3. 61597 *Gasoline* de 117927, representando 52,2% da amostra.
4. 4301 *LPG* de 117927, representando 0,04% da amostra.
5. 2621 *Hybrid* de 117927, representando 0,022% da amostra.
6. 885 *Electric* de 117927, representando 0,008% da amostra.

A próxima representação está relacionada à distribuição das cidades, buscando detectar as quantidades e proporções das amostras das cidades. Durante a análise, a distribuição aparenta possuir homogeneidade nas quantidades de cidades, não sendo identificada nenhuma necessidade de tratamento nesse momento. O gráfico pode ser visualizado na próxima figura a seguir:

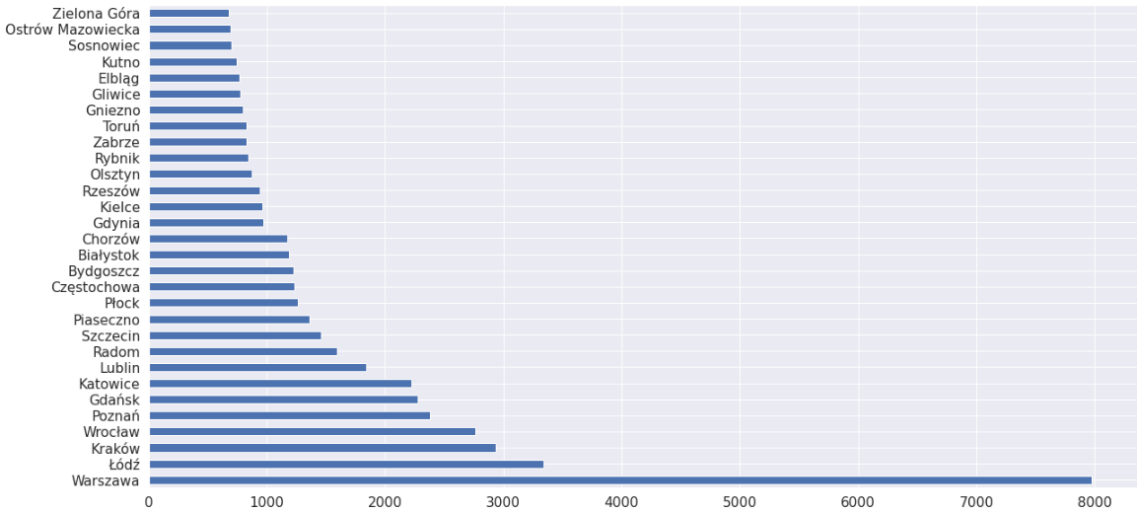

Figura 4: Quantificação de Cidades

A próxima representação é uma amostra da distribuição dos dados analisados das províncias, sendo assim, as demais cidades marginalizadas dessa representação apresentam pouca representatividade. Enquanto se cogita a questão de remoção desses conjuntos de dados ou agrupação, ainda se espera os primeiros testes do modelo de *Machine Learning*, visto que podem representar características importantes da distribuição de vendas, assim como uma preocupação de que agrupamentos consecutivos causem um desvio de viés que represente a realidade.

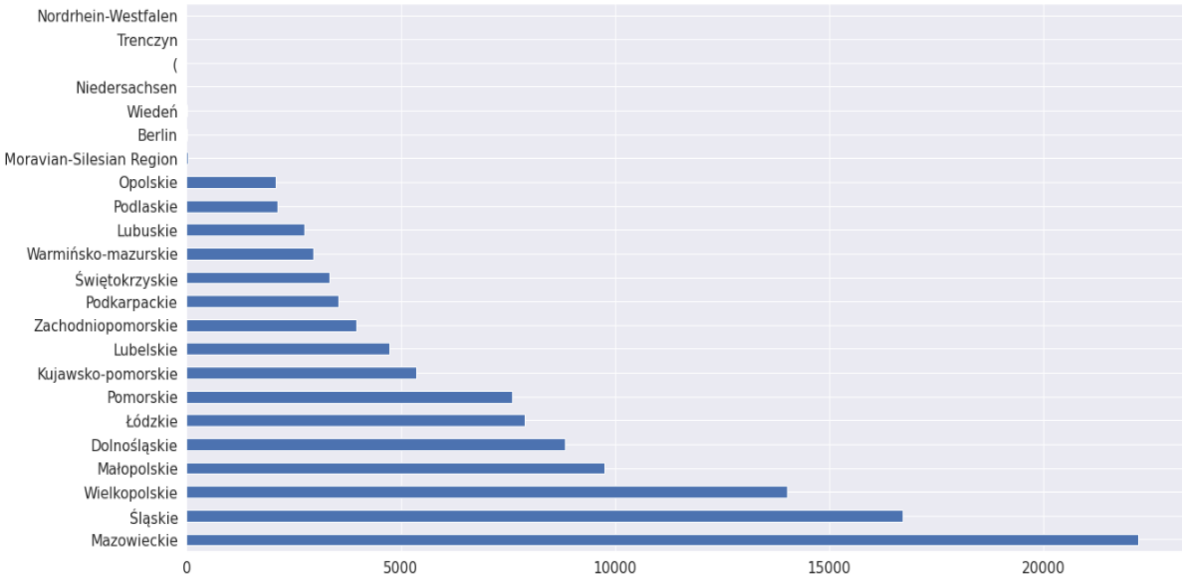

Figura 5: Quantificação de Províncias

Na figura anterior, é possível observar dados homogeneamente distribuídos nas províncias abaixo de *Opolskie* até *Mazowieckie.* A respeito de *Nordrhein-Westfalen* a *Moravian-Silesian Region*, respectivamente, os valores encontrados na distribuição foram 1, 1, 1, 2, 3, 35 indicando uma distorção na distribuição da amostra. A próxima etapa de análise do conjunto de dados foi realizada pelo agrupamento de valores contendo preços, tornando perceptível principais faixas, representado a seguir:

```
df["price_hist"] = pd.cut(df["price"],
                          bins=[0., 5000, 10000, 25000, 100000, 250000, 500000],
                          labels=[1, 2, 3, 4, 5, 6])
```

Tabela 4: Código do histograma de Valores

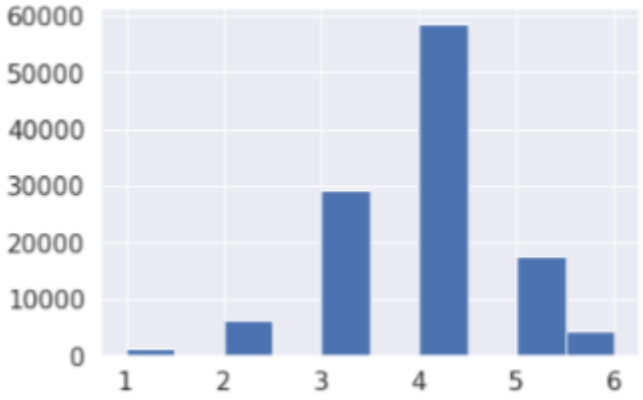

Figura 6: Histograma de Valores

Outras análises desta seção podem ser encontradas no Apêndice deste artigo, que contém todo o conjunto de informações exploradas e o código desenvolvido durante a análise.

2.3 COMPORTAMENTO E PREFERÊNCIA

O próximo ponto em questão de observação são os itens mais comuns e menos comuns enumerados na análise do *dataset*. Os objetos que compõem esses conjuntos podem ser explicativos se observados e avaliados quanto a questões de possível necessidade de agrupamento, ou eliminação se assim decidido, assim como podem revelar as características e comportamentos.

Indicativos de preferência comuns, são, respectivamente:

- Marca (maior preferência): [('audi', 12031), ('opel', 11914), ('bmw', 11070), ('volkswagen', 10848), ('ford', 9664)];
- Modelo (maior preferência): [('astra', 3331), ('seria-3', 2944), ('a4', 2912), ('golf', 2592), ('a6', 2496)];
- Ano (maior preferência): [(2021, 10559), (2017, 8909), (2018, 8647), (2016, 7021), (2009, 6828)];
- Motor (maior preferência): [(1598, 10206), (1968, 8121), (1995, 6545), (1997, 5340), (1998, 4498)];
- Combustível (maior preferência): [('Gasoline', 61597), ('Diesel', 48476), ('LPG', 4301), ('Hybrid', 2621), ('Electric', 885)];
- Cidade (maior preferência): [('Warszawa', 7972), ('Łódź', 3341), ('Kraków', 2936), ('Wrocław', 2764), ('Poznań', 2382)];
- Estado (maior preferência): [('Mazowieckie', 22219), ('Śląskie', 16706), ('Wielkopolskie', 14016), ('Małopolskie', 9756), ('Dolnośląskie', 8838)]
- Preço$ (maior preferência): [(19900, 1336), (39900, 1154), (29900, 1139), (18900, 1100), (14900, 1010)].

Indicativos de preferência menos comuns, são, respectivamente:

- Marca (menor preferência): [('chevrolet', 608), ('alfa-romeo', 704), ('mini', 1088), ('mitsubishi', 1120), ('honda', 2176)];
- Modelo (menor preferência): [('ampera', 10), ('frontera', 18), ('omega', 20), ('karl', 27), ('expert', 32)];
- Ano (menor preferência): [(1945, 1), (1974, 1), (1952, 1), (1983, 1), (1978, 2)];
- Motor (menor preferência): [(2935, 1), (1959, 1), (2319, 1), (2783, 1), (4415, 1)];

- Combustível (menor preferência): [('CNG', 47), ('Electric', 885), ('Hybrid', 2621), ('LPG', 4301), ('Diesel', 48476)];
- Cidade (menor preferência): [('Bledzew', 1), ('Augustówek', 1), ('Kolonia Gościeńczyce', 1), ('Kamionka Dolna', 1), ('Krężnica Okrągła', 1)];
- Estado (menor preferência): [('Nordrhein-Westfalen', 1), ('Trenczyn', 1), ('(', 1), ('Niedersachsen', 1), ('Wiedeń', 2)];
- Preço$ (menor preferência): [(417200, 1), (134558, 1), (328300, 1), (348930, 1), (290934, 1)];

2.4 VISUALIZAÇÃO DE OUTLIERS

Conjuntos de análises sobre *outliers* foram dirigidos durante o desenvolvimento deste artigo. Os dados referentes cruzados respectivamente em relação ao preço de cada item analisado, observando cada conjunto de lados correlacionado aos seus respectivos *outliers* nos preços. Os pontos observados no experimento, por grau de importância e mapeamento, são respectivamente os conjuntos de *outliers* do tipo de combustível, quilometragem agrupada e os subconjuntos de anos citados no capítulo anterior. Esses fatores foram decididos como importantes após pesquisa da perspectiva dos consumidores sobre os veículos.

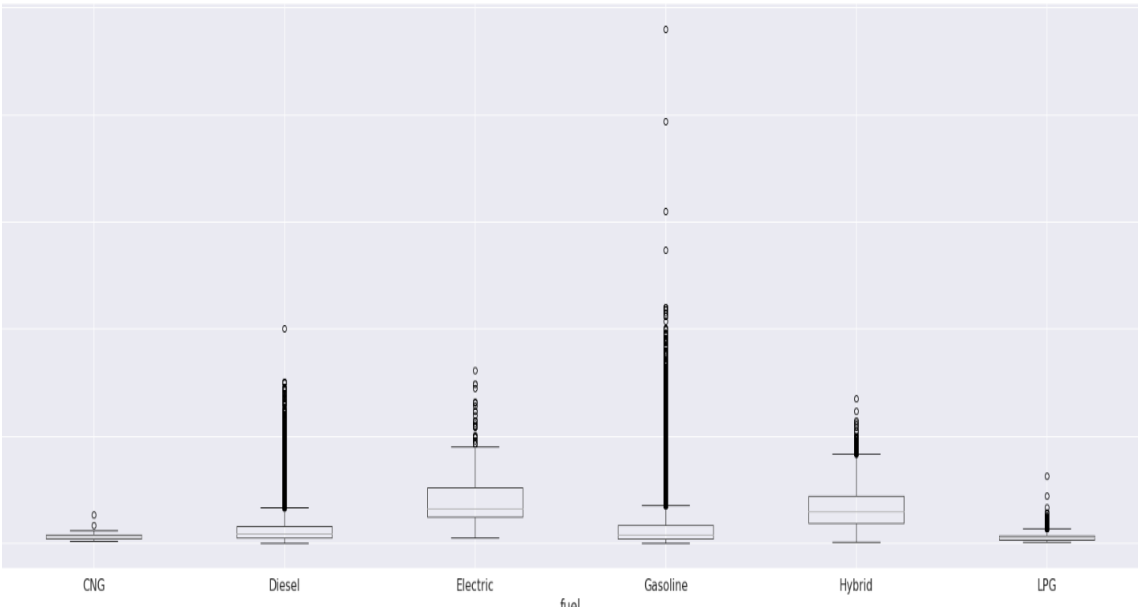

Figura 7: Outliers do preço sob conjunto Combustível

É possível notar conjuntos expressivos de concentração de variações, principalmente entre os conjuntos de veículos ao qual utilizam combustível a Diesel e Gasolina, assim como em veículos elétricos, híbridos, LPG e CNG em menor grau. Nesse ponto, como nas etapas anteriores, optou-se por observar o comportamento do modelo de *Machine Learning* antes de tomar quaisquer decisões enquanto, por outro

lado, também criar um "*data frame"* somente como esses dados dos *outliers* e outro sem eles, a fim de testar comparativamente o comportamento de modelo de *Machine Learning*.

O próximo ponto a seguir, é a análise dos *outliers* do conjunto de dados da quilometragem cumulativamente através de agrupamento, visto que, nos primeiros experimentos seria inviável visualizar individualmente seus *outliers* sem agrupá-los.

```
df["mileage_hist"] = pd.cut(df["mileage"],
                    bins=[0., 25000, 50000, 75000, 100000, 2000000.],
                    labels=[1, 2, 3, 4, 5])
df.boxplot(by = 'mileage_hist', column =['price'], figsize=(60,10), grid = True)
```

Tabela 5: Código do histograma geração de outliers

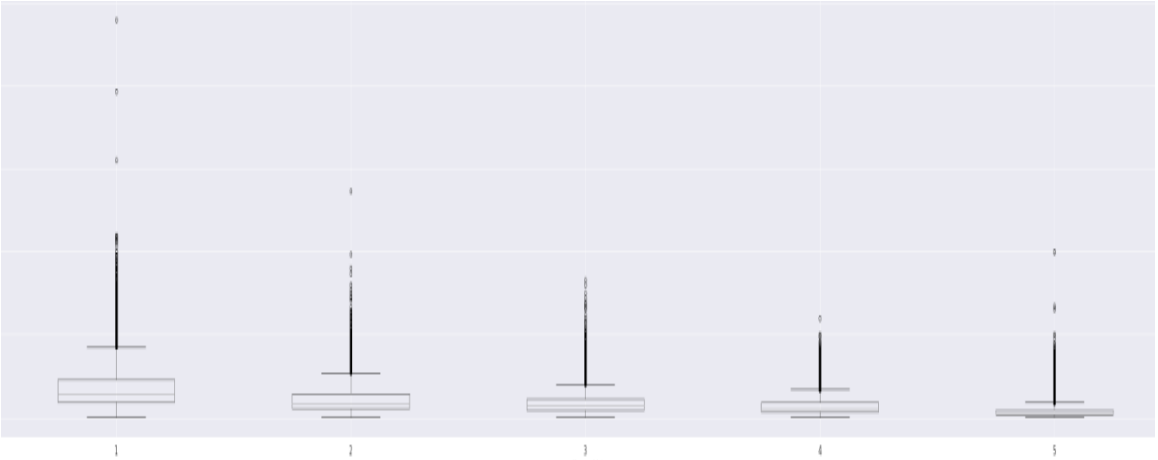

Figura 8: Outliers do preço sob conjunto de Quilometragem

Por meio do agrupamento de faixas de quilometragem, é possível a comparação de cinco conjuntos com preços estabelecidos de veículos, assim, demonstrada-se a variação dos *outliers* de cada uma dessas categorias. Nota-se que existem variações extremas em cada agrupamento.

Demonstrativamente, é segmentado um conjunto da amostra de dados para explicar a observação da relação dos *outliers* e ano. Consta no *dataset* extraído a informação que tal campo representa o ano de fabricação do veículo, sendo assim, é possível notar que, ao longo do tempo, existem crescentes conjuntos de *outliers* em relação aos veículos, esse fator pode ser explicado principalmente pelas novas formas de combustível, valores e marcas com grandes distorções de preços, porém seriam necessárias mais análises para conclusões.

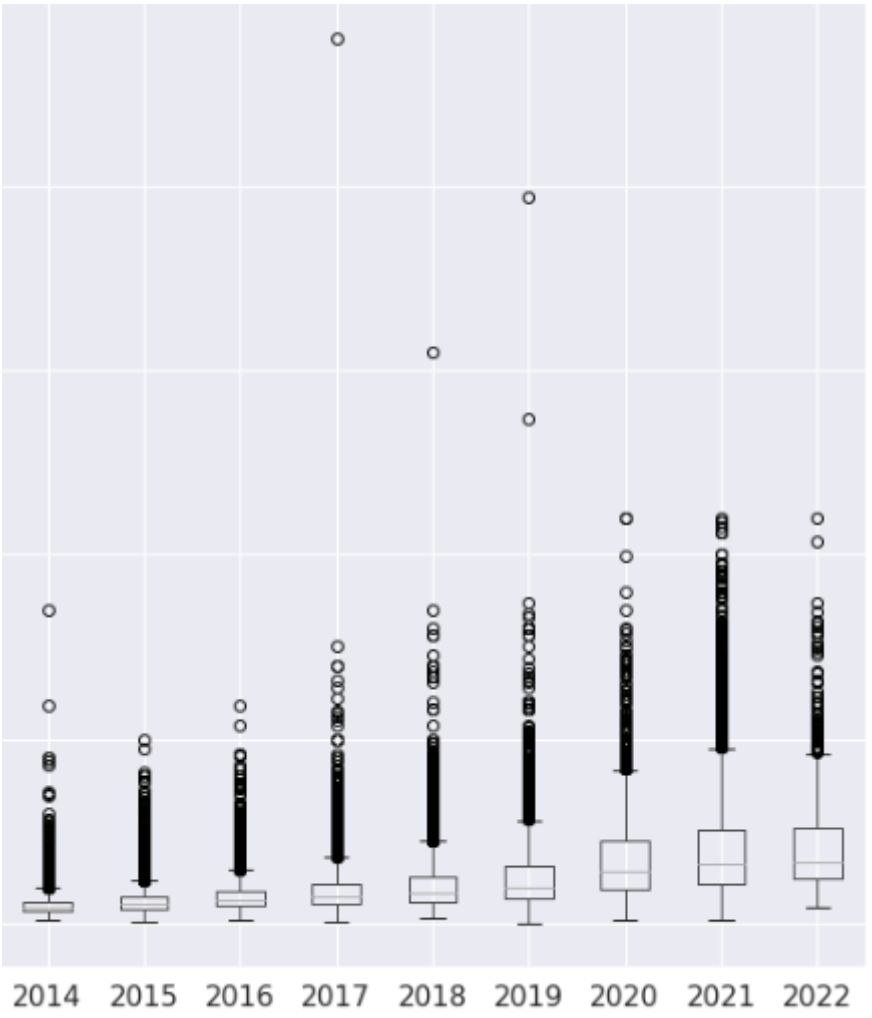

Figura 8: Outliers do Preço sob Ano de Fabricação

Na seção de apêndice, podem ser encontradas maiores análises sobre os conjuntos de *outliers* observados, assim como outras perspectivas de análise. Através das observações realizadas e comparações entre duas metodologias, *Z-Score (Escore Padrão),* também conhecida como *standard score* e *IQR* (*Amplitude interquartil*)*,* foram observados respectivamente 2.537 *outliers* de 117.927 registros, o que equivale a 0,02151% da amostra removidos pelo cálculo *Z-Score* e 10.064 *outliers* de 117.927, equivalente a 0,0853% da amostra removidos pelo cálculo *IQR*.

2.5 COMPORTAMENTO SOBRE A VARIÁVEL ALVO

Os fatores que impactam no valor final dos veículos, ou seja a variável *target*, apresentam sua composição a partir das variações de quilometragem, ano e tipo de combustível. A base de dados aparenta possuir uma distribuição homogênea quanto às cidades e estados onde os itens são vendidos, assim como uma quantidade crescente de *outliers*.

Durante as análises anteriores, foi observada a relevância dos conjuntos de dados expressos por ano, quilometragem e combustível. Foram colhidas informações sobre fatores de impacto no preço dos carros em jornais, informações do cotidiano, globo[1] e minuto seguros[2], tais informações são observadas a seguir:

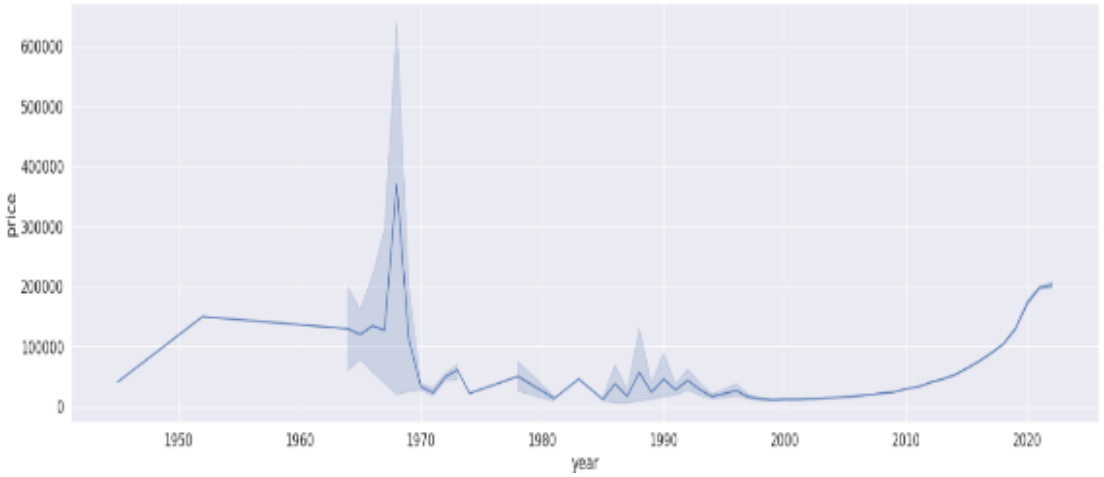

Figura 9: Preço sob Ano de Fabricação

Dessa forma, pode ser observado um valor crescente no preço dos veículos ao longo dos anos, com exceção de valores observados entre a década de 1950 e 1970. Segundo a revista online ISTOÉ[3], entre a década de 1950 e 1970, alguns veículos antigos possuem valor elevado na década atual em que esse artigo é produzido, explicando assim a distorção de valores. Entretanto, observa-se que é ideal manter essa distorção, visto que o ponto observado é o valor do veículo no qual é impactado pelas preferências de coleção.

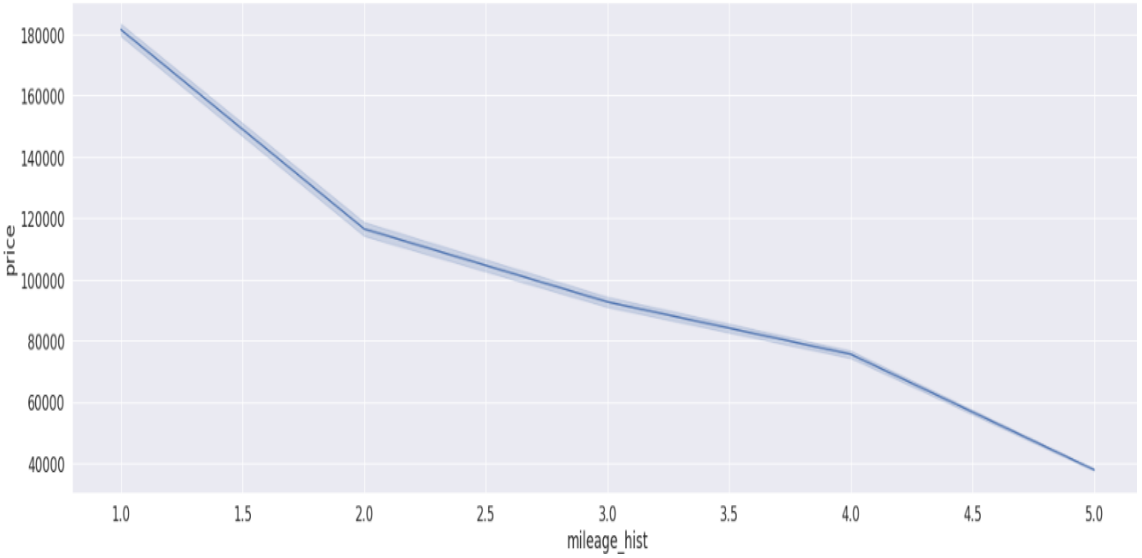

Figura 10: Preço sob Quilometragem

Realizando observações sobre os dados da quilometragem do veículo encontram-se indícios da afirmação em jornais da atualidade, apontada pelo declínio de

valores com base na quilometragem, representando uma importante característica a ser levada em consideração.

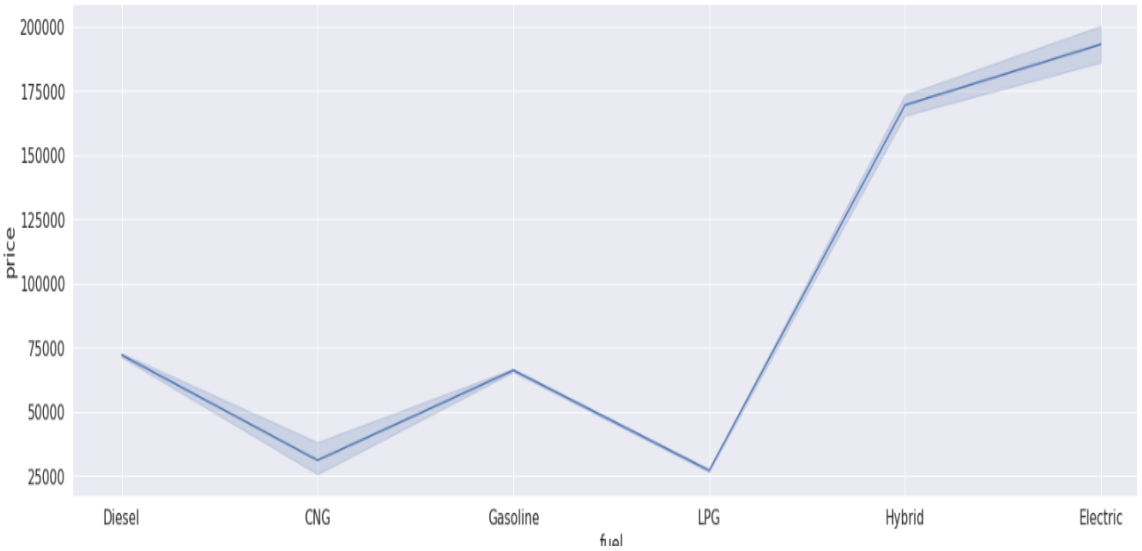

Figura 11: Preço sob tipo de combustível

O último ponto analisado sobre o veículo é representado na figura 11, e pelos valores correspondentes, ou seja, como a variável alvo do modelo pode ser impactada pela relação dos tipos de combustíveis. Nota-se a proximidade dos valores entre *Diesel* e *Gasoline*, porém, drástica variação entre os modelos híbridos e elétricos. Além disso, outras análises complementares podem ser encontradas no apêndice deste artigo.

2.6 ANÁLISE DA DISTRIBUIÇÃO NORMAL

Dentre os métodos analisados para o tratamento de *outliers* do conjunto de dados para este artigo foi selecionado o método Z-SCORE, no qual é necessária uma análise da distribuição normal do conjunto de dados.

Na figura 12, é analisada a distribuição normal, também conhecida como distribuição gaussiana, que mede o comportamento e o relacionamento dos eventos, assim como a probabilidade de ocorrerem de maneira correlacionada. A distribuição normal representa estatísticas discretas e contínuas, onde é possível traçar uma reta de avaliação. Observa-se nesta análise um gráfico de quantil a quantil, ou seja, *Q-Q*, que não forma uma distribuição perfeita de linha de 45º graus no *dataset* analisado, representando uma distribuição não normal.

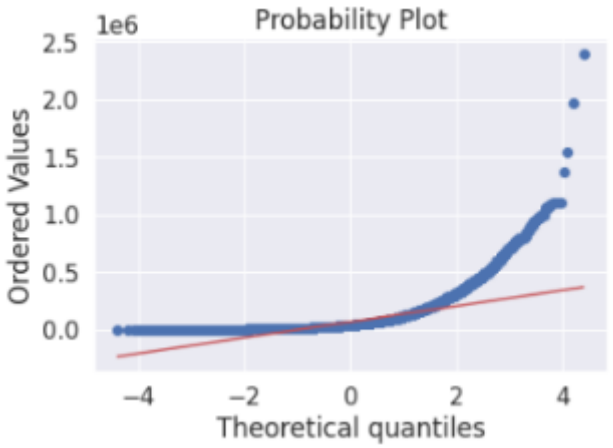

Figura 12: Quantile-Quantile Distribuição Normal

Observa-se que somente um conjunto de dados segue a reta prevista da *Theoretical Quantiles*, indicando não possuir uma distribuição *gaussiana*. A afirmativa anterior é confirmada ao aplicar-se o teste de *Shapiro* e *D'Agostino's K2,* que resulta em p = 0 para ambos os resultados, rejeitando H = 0 em ambos os casos e indicando uma distribuição não normal. Como conclusão dessa análise, o método *Z-SCORE* não pode ser aplicado sobre os conjuntos de *outliers* nos modelos de regressão. Uma abordagem para solucionar o problema da aplicação da metodologia *Z-SCORE* é a transformação na base *logarítmica* e aplicação do teste de hipóteses de *Shapiro-Wilk* e *D'Agostino's K²* sobre os conjuntos de dados, entretanto, não será abordada neste artigo.

2.7 ANÁLISE DA CORRELAÇÃO DE DADOS

A seguir, como abordagem da metodologia de exploração do conjunto em análise, são mensurados os comportamentos por meio matemático. As tabelas a seguir servem como base de como se encontram os relacionamentos dos dados, os valores representados no coeficiente de *Correlação de Pearson,* no qual é mensurada correlação de duas variáveis em escala métrica.

Nas tabelas a seguir é possível verificar o relacionamento entre as variáveis existentes que foram observadas neste trabalho. O coeficiente da *Correlação de Pearson* apresenta um intervalo de 1, o que indica uma relação positiva onde uma variável aumenta, assim, aumentará respectivamente a outra relacionada. Um intervalo 1 negativo representa uma relação perfeita entre as variáveis, porém indicando uma inclinação negativa, e indicando em 0 nenhuma associação entre as variáveis. A seguir será apresentado o relacionamento do *dataset* dos dados não-tratados e dados tratados, respectivamente, por metodologias utilizadas em ciência de dados.

Na tabela 6 a seguir, é possível identificar os relacionamentos medidos pelo cálculo da *Correlação de Pearson*.

-	ANO	KM	MOTOR	PREÇO
ANO	1.000000	-0.731958	-0.161557	0.596181
KM	-0.731958	1.000000	0.206169	-0.542808
MOTOR	-0.161557	0.206169	1.000000	0.299669
PREÇO	0.596181	-0.542808	0.299669	1.000000

Tabela 6: Correlação dos dados – não tratados.

Analisados os dados da tabela 6, é possível afirmar:

- ANO e PREÇO possuem um forte relacionamento positivo.
- O relacionamento entre a KM e o ANO é fortemente negativo.
- O ano possui um relacionamento fraco e negativo com o MOTOR.
- A KM do veículo afeta significante o preço.
- O MOTOR e a KM possuem uma relação positiva.
- O MOTOR e o ANO possuem uma relação negativa.

A metodologia aplicada, considerada no tratamento de dados nesse artigo foi a aplicação e remoção de valores extremos que se dá pelo *IQR* (*Amplitude interquartil*), que possibilita a definição de limites para aceitação e rejeição de valores estimados. Sendo, IQR = Q3 – Q1, Q1 representa o primeiro quartil e Q3 o terceiro quartil, implicando no limite inferior igual a Q1 - 1.5 * IQR e o limite superior igual a Q3 + 1.5 * IQR.

-	ANO	KM	MOTOR	PREÇO
ANO	1.000000	-0.690955	-0.286711	0.737301
KM	-0.690955	1.000000	0.351611	-0.591232
MOTOR	-0.286711	0.351611	1.000000	0.100293
PREÇO	0.737301	-0.591232	0.100293	1.000000

Tabela 7: Correlação dos dados tratados – método IQR.

Na tabela 7, é possível notar algumas mudanças nos relacionamentos das variáveis categorias.

- ANO e PREÇO apresenta aumento na representação positiva.
- O relacionamento entre a KM e o ANO apresenta redução no relacionamento negativo.
- O ANO em relação ao MOTOR apresenta aumento no relacionamento negativo.
- A KM apresenta aumento no relacionamento negativo do PREÇO.

- O MOTOR e a KM apresentam aumento na relação positiva.
- O MOTOR e o ANO apresentam aumento na relação negativa.

A metodologia *Z-Score,* utilizada para identificar *outliers* medindo a distância numérica dos pontos até a média da amostra em desvios padrões, não será utilizada, pelos seguintes motivos;

1. Sendo um método sensível a valores extremos pode não identificar possíveis outliers;
2. Durante a análise, foram observadas distorções em diversos aspectos, estas apresentadas nesse artigo e encontradas no apêndice;
3. Por o conjunto de dados não se tratar de uma distribuição normal, ou seja, distribuição *gaussiana* e este artigo não aborda metodologias de normalização de distribuição.

2.8 INDICATIVOS DO DATASET

A seguir, é apresentado um resumo das análises características da base de dados estudada no desenvolvimento deste artigo:

Características gerais da base de dados analisada:

- A base de dados apresentada representa os anos de 1945 a 2022;
- Dentro do *dataset,* existem 23 marcas diferentes, 328 modelos, 508 motores, 6 tipos de combustível, 4.427 cidades e 9.000 preços para veículos;
- Os dados em sua pluralidade apresentam boa distribuição homogênea;
- Dados apresentam conjuntos significativos de *outliers* após determinadas datas, indicando principalmente a virada do século 21;
- *Outliers* representam 0,0085% da base de dado;
- Tamanho do *dataset* Tratado: 107.863;
- Tamanho do *dataset* Não-Tratado: 117.927;
- *generation_name*, 30.085 *Null's*;
- *Unnamed*, *df.dropped* (removida coluna de numeração).

Itens de maior representação do *dataset*:

- **MARCA**: *Audi, Opel, BMW, Volkswagen, Ford*;
- **MODELO**: *Astra, Série-3, A4, Golf , A6*;
- **ANO**: 2021, 2017, 2018, 2016, 2009;
- **MOTOR**:1598, 1968, 1995, 1997, 1998;
- **COMBUSTÍVEL**: *Gasoline*, *Diesel*, *LPG*, *Hybrid*, *Electric;*

- **CIDADES**: Warszawa, Łódź, Kraków, Wrocław, Poznań;
- **ESTADO**: *Mazowieckie, Śląskie, Wielkopolskie, Małopolskie;*
- **PREÇO**: $$19.900, $39.900, $29.900, $18.900, $14.900
- **PREÇOS MAIS ALTOS**: *Mercedes, BMW, Audi, Volvo, Alfa-Romeo*;

Principais indicativos da análise da correlação:

- **ANO**: KM apresenta forte correlação negativa assim como o PREÇO;
- **KM**: PREÇO E ANO apresentam uma forte correlação negativa;
- **MORTOR**: ANO e possui uma correlação negativa e KM uma correlação positiva;
- **PREÇO**: ANO possui uma forte correlação positiva e KM uma forte correlação negativa.

2.9 IMPLEMENTAÇÃO DO MODELO

Após análises, as variáveis categóricas são escolhidas através das observações da *Correlação de Pearson*, fatores relacionados à atualidade, mercado, assim como pelas técnicas de visualização. Logo após a escolha das variáveis categorias, estas são convertidas numericamente utilizando uma biblioteca em *Python LabelEncoder*. A metodologia utilizada transforma os dados entre valores de 0 até N-1, convertendo tipos de dados não numéricos em dados numéricos, permitindo que assim modelos de regressão escolhidos sejam utilizados. O resultado da aplicação dessa metodologia pode ser visualizado na tabela a seguir:

year	mileage	vol_engine	price	Mark	Model	Fuel
2015	139568	1248	35900	15	89	1
2018	31991	1499	78501	15	89	1
2015	278437	1598	27000	15	89	1
2016	47600	1248	30800	15	89	1
2014	103000	1400	35900	15	89	0

Tabela 9: Conversão de dados para tipos numéricos

Com as conversões de dois *datas frames*, respectivamente, *DF_IQR,* tratado pelo cálculo para renovação de *outliers* interquartile, e *dfNormal,* sem nenhum tratamento dos dados, logo são instanciados os conjuntos de dados das variáveis preditoras. Os dados são separados em conjuntos de treino e teste, para os *data frames,* separando 70% dos dados para treino e 30% dos dados para teste, como amplamente estudado na literatura, representados a seguir:

```
X2 = dfNormal.drop(columns="price")
y2 = dfNormal["price"]
X_train2, X_test2, y_train2, y_test2 = train_test_split(X2, y2, test_size = 0.3, random_state=42)
X2 TREINO: (82548, 6), X2 TESTE: (35379, 6), Y2 TREINO: (82548,), Y2 TESTE: (35379,)
```

Tabela 10: Exemplo divisão dados Treino e Teste

Em busca de um melhor entendimento sobre o comportamento dos dados, assim como comparativamente em busca da metodologia adequada de tratamentos para as anomalias encontradas durante a análise exploratória dos dados e da análise de desempenho, dois modelos são escolhidos e treinados para cada *data frame*. Os modelos testados progressivamente são, *Linear Regression, Random Forest Regressor,* escolhidos assim de menor para maior grau de complexidade.

Assim, comparativamente, modelos de *Machine Learning* usarão dois *data frames* diferentes: um *dataframe* não-tratado, um *dataframe* tratado pela metodologia *IQR*, ambos codificados utilizando a biblioteca *LabelEncoder.* Comparativamente serão medidas as taxas dos dados treino e teste utilizando metodologia *score* e *r2 score* e análises de resíduos com o objetivo de fornecer um parâmetro como critério de avaliação e aceitação dos modelos. Os modelos serão discutidos na seção a seguir.

2.10 COMPARAÇÃO DOS MODELOS

A seguir são demonstrados dois modelos de *Machine Learning* e comparados os *dataframe* tratado e não tratado. O *DTCdf* representa os dados tratados, assim como *X_train, X_test, y_train, y_test* enquanto o *dfNormal* dados não tratados *X_train2, X_test2, y_train2, y_test2*. Os Resultados do treinamento podem ser visualizados na tabela a seguir:

	FUNCTION	SCORE	R2 SCORE
LINEAR REGRESSION IQR	regr.score(X_train,y_train)	0.6337	–
	regr.score(X_test,y_test)	0.6403	–
	r2_score(y_test,y_pred)	–	0.6403
LINEAR REGRESSION DFNORMAL	regr.score(X_train2,y_train2)	0.5660	–
	regr.score(X_test2,y_test2)	0.5833	–
	r2_score(y_test2,y_pred2)	–	0.5424

Tabela 11: Resultados dos Modelos Parte I – *Linear Regression.*

Na seção 3 serão descritos os resultados das pontuações de treino e teste, assim como a medida *score* e *r2 score* para modelos de dados tratados e não tratados. Na figura 13 seguir é apresentado gráfico de dispersão de resíduos e o *Theoretical Quantiles* do modelo de Regressão Linear;

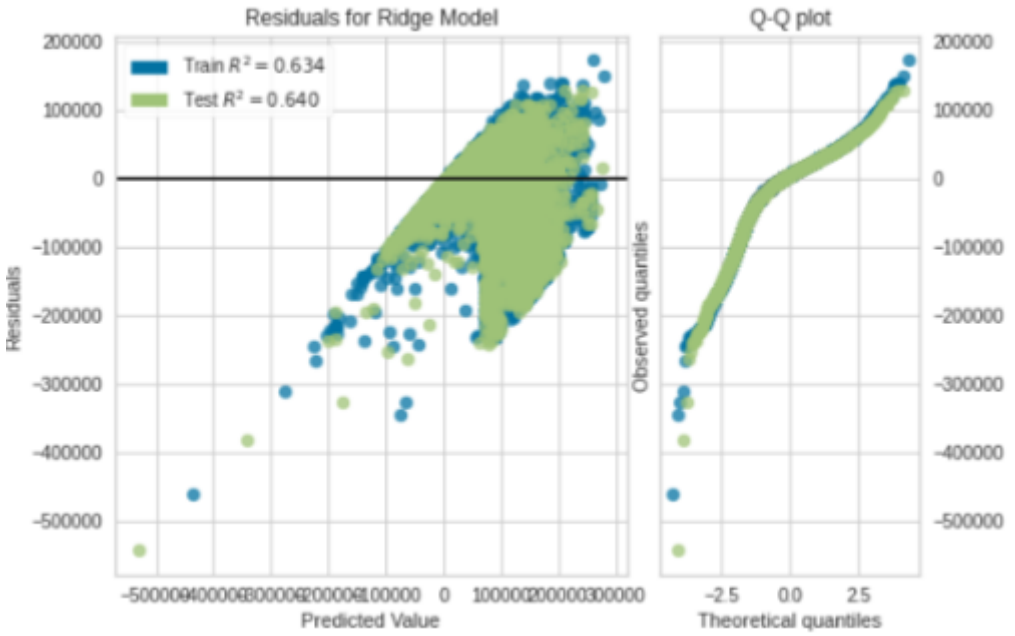

Figura 13: Dispersão e Resíduos (*DTCdf* dados tratados – IQR*).*

Na figura 13 é possível observar os resíduos, assim como *Q-Q plot*, representando a inclinação e ajuste do modelo. Essas e outras informações serão debatidas na seção de resultados.

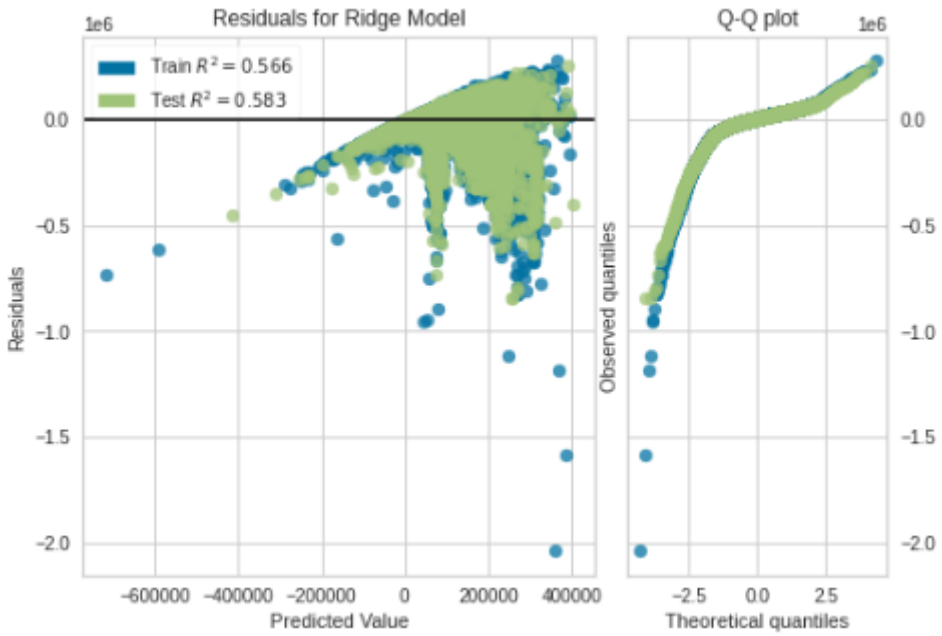

Figura 14: Dispersão e Resíduos (Df Normal – dados não-tratados).

Na figura 14, é observado o gráfico residual, assim como *Q-Q plot*. A análise será debatida na seção de resultados.

	FUNCTION	SCORE	R2 SCORE
RANDOM FOREST REGRESSOR IQR	RFR.score(X_train,y_train)	0.9873	–
	RFR.score(X_test,y_test)	0.9434	–
	r2_score(y_test,y_predRFR)	–	0.9434
RANDOM FOREST REGRESSOR DFNORMAL	RFR.score(X_train2,y_train2)	0.9856	–
	RFR.score(X_test2,y_test2)	0.9397	–
	r2_score(y_test2,y_predRFR2)	–	0.9397

Tabela 12: Resultados dos Modelos Parte III – *Random Forest Regressor*

Respectivamente na seção 3 serão descritos os resultados. O gráfico de dispersão de resíduos do modelo *Random Forest Regressor* é apresentado na figura 15, a seguir;

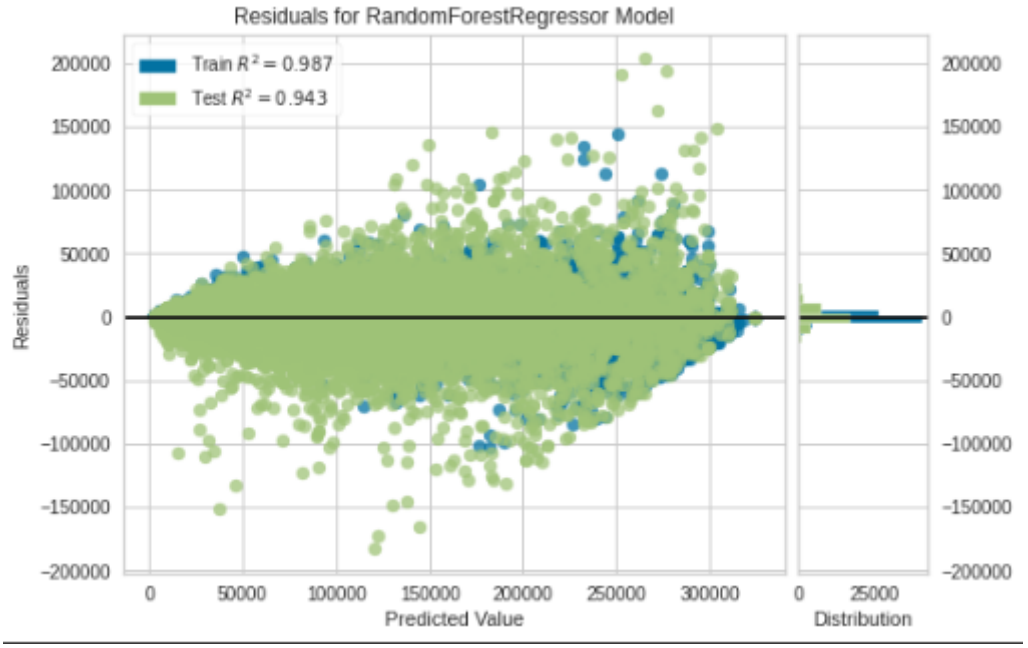

Figura 15: Dispersão e Resíduos (*Random Forest Regressor* tratados − IQR).

A figura 15 representa um gráfico de resíduos, assim como a *Distribution*, distribuição associada ao erro, representando o ajuste do modelo em torno do eixo zero. Essas e outras informações serão debatidas na seção de resultados.

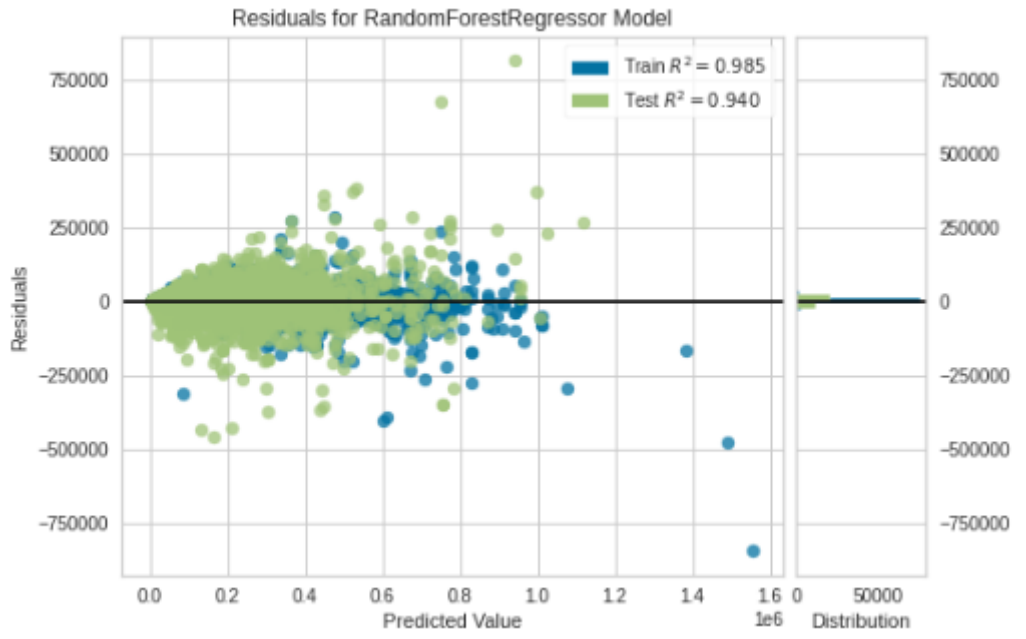

Figura 16: Dispersão e Resíduos (*Random Forest Regressor* − não tratados).

Na figura 16, é observado o gráfico residual, assim como sua distribuição. Essas e outras informações serão debatidas na seção de resultados.

3. DESCRIÇÃO DOS RESULTADOS

Nessa seção, são discutidos comparativamente os resultados obtidos, usando como parâmetro o mesmo *dataset,* assim instanciado com e sem tratamento do conjunto de dados, e progressivamente aplicando modelos mais complexos de *Machine Learning* como comparativos.

3.1 LINEAR REGRESSION

Linear Regression é um algoritmo de *Machine Learning* que traça uma reta a partir de uma relação através de diagramas de dispersão, baseando-se na correlação entre as variáveis categóricas, ou seja para cada X, implica em Y. Logo, a reta sintetiza os relacionamentos, e o modelo de *Machine Learning* pode ser aplicado para realizar previsões baseado em valores quantitativos que explicam a variável categoria X e determina como medida a variável alvo Y de forma crescente ou decrescente no relacionamento.

Utilizado para avaliar o desempenho do modelo o *score method of regressors* ou *score*, método de pontuação de precisão, que é calculado pelo coeficiente de determinação de R2, indica a pontuação do algoritmo de 0,6337 para dados de *X Train* e *Y Train* e 0,6403 para dados de *X Test* e *y Teste*, empregados em um *dataframe* tratado pelo cálculo do intervalo interquartil (*IQR*).

O desempenho do modelo de *Regressão Linear* pelo método *score* indica a pontuação do algoritmo de 0,5660 para dados de *X Train 2* e *Y Train 2* e 0,5833 para dados de *X Test 2* e *y Teste 2*, empregados em um *dataframe* não-tratado pelo cálculo do intervalo interquartil (*IQR*).

A próxima medida de pontuação utilizada é *R2 score,* que avalia o desempenho do modelo de *Machine Learning*, representando r^2, conhecido como coeficiente de determinação, como medida da variância nas previsões do conjunto de dados e, representando a diferença entre o valor das amostras no conjunto de dados e as previsões realizadas. O modelo de *Regressão Linear* apresentou avaliação de desempenho de 0,6403 para dados de *y test* e *Y pred* (previsto)*,* representando o *data frame* tratado pelo método *IQR* e 0,5424 para dados de *y teste 2* e *Y pred 2* (previsto), representando o *data frame* não tratado pelo método *IQR*.

3.1.1 LINEAR REGRESSION − RESIDUAL E THEORITICAL QUANTILES

Na literatura, os resíduos são calculados pela diferença do valor residual observado e o valor previsto, que indica pontos de melhoria para o modelo. Uma dispersão próxima ao eixo central indica um modelo mais preciso, apontando fortes correlações dos conjuntos de dados observados entre dados reais e resultados previstos. Como estudado na literatura, o uso mais comum do gráfico de resíduos é para demonstrar a variação em relação ao erro do regressor. A distribuição ou dispersão de

dados aleatoriamente em torno do eixo horizontal aponta um modelo apropriado ao conjunto de dados.

Comparativamente, o gráfico de resíduos, apresentado na figura 13, representa dados tratados pelo método *IQR,* enquanto a figura 14, representa dados não tratados. Nota-se uma maior distribuição uniforme dos resíduos apontados pelo método *IQR* em torno de zero, quando comparado com o modelo de dados não tratados.

O gráfico de resíduos nas figuras 13 e 14 apresenta também o *Theoretical Quantile*, ou *Q-Q Normal*, que representa um teste empírico para explorar a diferença da distribuição normal em torno do eixo X demonstrando os desvios da distribuição teórica comparada aos dados previstos da distribuição matemática. É possível observar os ruídos inerentes ao mundo real que impactam na deformação da cauda, representando pontos mais distantes da média. Quanto à observação feita sobre os *quantiles,* representados pelo *Theoritical Quanitles* percebe-se um maior ajuste aos pontos mais distantes do modelo *Linear Regression* tratado pelo método *IQR*, que representa os pontos de melhoria que podem ser indicados removendo variáveis ou tratando os subconjuntos de dados.

Em ambas as análises, observa-se maior ajuste aos ruídos representados do mundo real ao utilizar o método *IQR* no modelo de *Linear Regression*.

3.2 RANDOM FOREST REGRESSOR

Random Forest é um algoritmo que aplica metodologia de árvore de decisão, utilizando os hiperparâmetros de uma árvore para o processo de combinação das variáveis categorias X e Y previsto. Combinados a K número de dados a árvore escolhe N árvores para construir etapas, empregando o método *ensemble methods (bagging)* para combinar um algoritmo de ajuste a subconjuntos de modelos de treino e previsões de vários modelos e selecionando aleatoriamente subconjuntos usados em cada amostra para resolver problemas de classificação e regressão.

Durante a análise, o desempenho do modelo *Random Forest Regressor* apresentou o *score* de precisão de 0,9873 para dados de *X Train* e *Y Train* e 0,9434 para dados de *X Test* e *y Teste*, empregados em um *data frame* tratado pelo cálculo do intervalo interquartil (*IQR*).

O desempenho do modelo pelo método *score* indica a pontuação de 0,9856 para dados de *X Train 2* e *Y Train 2* e 0,9397 para dados de *X Test 2* e *y Test 2*, empregados em um *data frame* não-tratado pelo cálculo do intervalo interquartil (*IQR*).

A medida de pontuação *R2 score* para o modelo de *Random Forest Regressor,* apresenta avaliação de desempenho de 0,9434 e 0,9397, para dados tratados e não tratados respectivamente.

3.2.1 RANDOM FOREST REGRESSOR − RESIDUAL MODEL E DISTRIBUTION

A análise a seguir está relacionada ao *Random Forest Regressor* para dados tratados pela metodologia *IQR* e não tratados, representados pela análise do *Residual Model* e *Distribution*, respectivamente nas figuras 15 e 16 dos gráficos de resíduos de dados tratados e dados não tratados.

A apresentação de resíduos representa as mesmas características associadas ao modelo de observação e predição, citados na seção anterior de *Linear Regression*, quanto à relação de distribuição de resíduos de dados previstos e observados sobre o erro é substituída a avaliação do *Theoretical Quantile* pela distribuição aleatória de dados de resíduos que é representada ao redor de suas dimensões no gráfico de *Distribution*.

Ambos os modelos de *Machine Learning* apresentam boa performance. O primeiro modelo tratado com metodologia *IQR* apresenta maior quantidade de resíduos, assim como um histograma *Distribution* do erro mais distribuído em torno de zero. O segundo modelo de *Machine Learning* de dados não-tratados apresentam menor quantidade de resíduos em torno do eixo e menor erro concentrado em torno de zero.

5. CONCLUSÃO

O avanço da tecnologia moderna permitiu registrar uma grande quantidade de dados sobre diversas coleções de objetos que representam a realidade das atividades, interações, fatos e relações humanas com o mundo. Conforme esses dados são expressos por meio de conhecimentos científicos, torna-se possível organizar e reorganizar informações, através de análises, experimentações e tratativas, possibilitando atingir conclusões que agregam valor, na literatura, citadas como produção de conhecimento.

Durante o experimento, foi constatada a importância da internalização dos dados, seus relacionamentos, avaliação de métricas matemáticas, técnicas de visualização, experimentação e exploração do conhecimento explícito sobre o assunto. Dessa maneira é perceptível que toda essa informação e experimentação serve como um termômetro para o cientista de dados construir adequadamente um modelo de *Machine Learning* que possa resolver problemáticas estabelecidas.

Durante a observação o *linear regression,* modelo menos complexo, apresentou possuindo menor grau de desempenho sem nenhuma tratativa de *outliers*, porém a tratativa permitia melhor assimilação dos conjuntos, apresentando um melhor desempenho. O modelo *random forest regressor* se ajustou bem ao *dataset*, tanto nos dados tratados quanto não tratados, indicando indícios de overfitting para dados tratados.

Portanto, para modelos de *Machine Learning* menos complexos como *Linear Regression* é possível observar a necessidade do tratamento de dados das províncias, a remoção dos itens de menor representatividade, tratamento das variações de preços sobre a quilometragem, ano e o agrupamento ou remoção dos tipos de combustível de menor relevância.

Apesar das diferenças de desempenho em ambos os modelos, comparativamente *Linear Regression* e *Random Forest Regressor*, em modelos mais complexos como *Random Forest Regressor* é possível pressupor que os *outliers* representam um contexto da realidade da venda de veículos, onde nesses casos se apresente interesse em analisar casos excepcionais. Conclui-se assim, que, em determinadas análises, quando aplicados modelos complexos os *outliers* podem na verdade ser a representação de comportamentos existentes.

6. REFERÊNCIAS

Tratamento de dados: uma abordagem prática para aprendizagem de atenuação de ruídos e eliminação de outliers, Saulo Neves Matos. 14º Simpósio Brasileiro de Automação Inteligente

EATON, Chris; DEUTSCH, Tom; DEROOS, Dirk; et al. Understanding Big Data. Analytics for enterprise class hadoop and streaming data. Nova York, McGraw Hill, 2012.

Han, J., Kamber, M., e Pei, J. (2006). Data Mining: Concepts and Techniques. Morgan Kaufmann. Department of Computer Science University of Illinois.

Ciência de Dados, Fabio Porto e Artur Ziviani, Laboratório Nacional de Computação Científica (LNCC).

Por que os preços dos carros dispararam no mundo. G1 Globo, 16/11/2021. Disponível em:
https://g1.globo.com/economia/noticia/2021/11/16/por-que-os-precos-dos-carros-dispararam-no-mundo.ghtml

William Bonner aumenta coleção de carros clássicos com modelo que pode chegar a R$ 100 mil ISTOÉ, 09/01/2022. Disponível em:

https://www.istoedinheiro.com.br/william-bonner-aumenta-colecao-de-carros-classicos-com-modelo-que-pode-chegar-a-r-100-mil/

Moda faz crescer procura por antigos. Veículos com mais de 30 anos servem como chamariz para seminovos; com 20 anos, a valorização anual é de 10%. Folha de São Paulo, 2005. Disponível em:

https://www1.folha.uol.com.br/fsp/veiculos/cv2011200501.htm

7. APÊNDICE

A seguir link disponível do código no *Github* ao qual pode ser utilizado neste trabalho para maiores análises de outros gráficos e dados explorados, assim como para replicação.

Link, https://github.com/ernesto-arq/Artigo_Ciencia_de_Dados.