**FACULDADE DE TECNOLOGIA DA ZONA LESTE**

ALEKSANDER ALENCAR JUNIOR

NICHOLAS FELICIO BARBOSA

**EXTRAÇÃO E ANÁLISE DE DADOS APLICADO A UM SITE DE LEILÕES DE AUTOMÓVEIS.**

São Paulo

2022

ALEKSANDER ALENCAR JUNIOR

NICHOLAS FELICIO BARBOSA

**EXTRAÇÃO E ANÁLISE DE DADOS APLICADO A UM SITE DE LEILÕES DE AUTOMÓVEIS.**

Trabalho de Conclusão de Curso apresentado à Faculdade de Tecnologia da Zona Leste, como requisito parcial para a obtenção do diploma de Graduação no Curso de Análise e Desenvolvimento de Sistemas.

São Paulo

2022

Desenho com traços pretos em fundo branco

Descrição gerada automaticamente com confiança média

ALEKSANDER ALENCAR JUNIOR

NICHOLAS FELICIO BARBOSA

EXTRAÇÃO E ANÁLISE DE DADOS APLICADO A UM SITE DE LEILÕES DE AUTOMÓVEIS.

Monografia apresentada no curso de Análise e Desenvolvimento de Sistemas na Faculdade de Tecnologia da Zona Leste orientado pelo Professor Me. Wellington Pinto de Oliveira.

Aprovado em:

Banca Examinadora

Prof. Me. \_\_Wellington Pinto de Oliveira\_\_Instituição: \_FATEC ZL\_\_

Julgamento: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Assinatura: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Msc. \_\_Leandro Colevati dos Santos\_Instituição: \_FATECZL\_

Julgamento: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Assinatura: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Me \_Ricardo Satoshi Oyakawa\_Instituição: \_FATEC ZL\_

Julgamento: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Assinatura: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

São Paulo, 09 de Novembro de 2022

JUNIOR, Aleksander Alencar. BARBOSA, Nicholas Felicio; **EXTRAÇÃO E ANÁLISE DE DADOS APLICADO A UM SITE DE LEILÃO DE AUTOMÓVEIS**, 71 p. Trabalho de conclusão de curso de graduação, Faculdade de Tecnologia da Zona Leste.

**RESUMO**

Carro é um bem móvel muito popular e o mercado de usados passou por elevado crescimento no ano de 2021, porém diversos obstáculos podem surgir na hora de escolher e comprar um carro como preço, modelo, marca e entre outras características, então como uma ferramenta de extração e análise de dados pode identificar os lotes de carros desfavoráveis com base nas suas características; Para este trabalho foi essencial realizar estudos sobre técnicas e tecnologias utilizadas para o processo de extração, mineração e análise de dados, tarefas que compõe a descoberta de conhecimento em base de dados e os dados mostraram que lotes com estado do chassi com os valores danificado, amassado, ilegível ou enferrujado; tipo de origem judicial; quilometragem acima dos 300.000; ar-condicionado danificado; direção hidráulica/elétrica danificada, são altamente desfavoráveis em relação aos demais valores dessas mesmas características.

**Palavras Chave**: Extração de Dados; Análise de Dados; Ciência de dados; Descoberta de Conhecimento em Base de Dados

JUNIOR, Aleksander Alencar. BARBOSA, Nicholas Felicio; **EXTRAÇÃO E ANÁLISE DE DADOS APLICADO A UM SITE DE LEILÃO DE AUTOMÓVEIS**, 71 p. Trabalho de conclusão de curso de graduação, Faculdade de Tecnologia da Zona Leste.

**ABSTRACT**

Car is a very popular asset and the used market experienced high growth in the year 2021, but several obstacles can arise when choosing and buying a car such as price, model, brand and among other characteristics, so as a tool to receive data analysis can identify unfavorable car batches based on their characteristics; For this work, it was essential to carry out studies on techniques and technologies used for the warranty process, data mining and analysis, tasks that make up the discovery of knowledge in the database and the known data that lots with the state of the chassis with the damaged values, dented, illegible or rusted; type of judicial origin; mileage over 300,000; damaged air conditioning; damaged hydraulic/electric steering, are highly unfavorable in relation to the other values of these same characteristics.

**Keywords**: Data Extraction; Data Analysis; Data Science; Knowledge Discovery in Databases

**Lista de Figuras**

[**Figura 1 - Dados, processamento e informação** 13](#_Toc120641205)

[**Figura 2 - Interdisciplinaridade da Ciência de Dados** 14](#_Toc120641206)

[**Figura 3 - Ciclo de vida Ciência de Dados** 15](#_Toc120641207)

[**Figura 4 - Diferenças de dados estruturados** 18](#_Toc120641208)

[**Figura 5 - Processo de Extração de dados** 20](#_Toc120641209)

[**Figura 6 - Frequência de perguntas sobre JSON no StackOverflow** 21](#_Toc120641210)

[**Figura 7 - Processo de mineração de dados** 23](#_Toc120641211)

[**Figura 8 - Fases do KDD** 31](#_Toc120641212)

[**Figura 9 - Características do veículo** 41](#_Toc120641213)

[**Figura 10 - Fluxograma do processo** 42](#_Toc120641214)

[**Figura 11 - HTTP Get** 43](#_Toc120641215)

[**Figura 12 - Contagem total de páginas** 43](#_Toc120641216)

[**Figura 13 - O uso de Regex** 44](#_Toc120641217)

[**Figura 14 - Montando lista de URLs** 44](#_Toc120641218)

[**Figura 15 - Definindo key e value no dicionário de informações capturadas** 45](#_Toc120641219)

[**Figura 16 - Uso de XPath** 46](#_Toc120641220)

[**Figura 17 - Iterando sobre lista no HTML** 46](#_Toc120641221)

[**Figura 18 - Conversão em datetime** 47](#_Toc120641222)

[**Figura 19 - Sumário da base de dados** 48](#_Toc120641223)

[**Figura 20 - Verificação de dados nulos** 49](#_Toc120641224)

[**Figura 21 - Dataframe após a limpeza de dados** 50](#_Toc120641225)

[**Figura 22 - Fluxo de componentes para a visualização dos dados** 51](#_Toc120641226)

[**Figura 23 - Componente de Tabela de Dados** 52](#_Toc120641227)

[**Figura 24 - Lotes com as características desfavoráveis** 60](#_Toc120641228)

**Lista de Gráficos**

[**Gráfico 1 - Distribuição de Marcas** 54](#_Toc120641195)

[**Gráfico 2 - Distribuição de Cor** 54](#_Toc120641196)

[**Gráfico 3 - Distribuição de Tipo de Combustível** 55](#_Toc120641197)

[**Gráfico 4 - Distribuição de Tipo de Câmbio** 55](#_Toc120641198)

[**Gráfico 5 - Distribuição de Tipo de Origem** 56](#_Toc120641199)

[**Gráfico 6 - Distribuição por Estado do Chassi** 56](#_Toc120641200)

[**Gráfico 7 - Dispersão de Km X Ano** 57](#_Toc120641201)

[**Gráfico 8 - Dispersão de Km X Ano (origem)** 58](#_Toc120641202)

[**Gráfico 9 - Dispersão de Km X Lance Atual** 59](#_Toc120641203)

[**Gráfico 10 - Dispersão de Ano X Lance Atual** 59](#_Toc120641204)

**Lista de Abreviaturas e Siglas**

CSV Comma-separated Values

DOM Document Object Model

HTML Hypertext Markup Language

HTTP Hypertext Transfer Protocol

JSON JavaScript Object Notation

KDD Knowledge Discovery in Database

URL Uniform Resource Locator

WEKA Waikato Environment for Knowledge Analysis

FENABRAVE Federação Nacional da Distribuição de Veículos Automotores

**SUMÁRIO**

[1 INTRODUÇÃO 8](#_Toc120641229)

[2 FUNDAMENTAÇÃO TEÓRICA 11](#_Toc120641230)

[2.1 Dados e Informação 11](#_Toc120641231)

[2.2 Tipos de Fontes de Dados 16](#_Toc120641232)

[2.3 Extração e Manipulação de Dados 18](#_Toc120641233)

[2.3.1 Extração de Dados 19](#_Toc120641234)

[2.3.1.1 Extração de HTML 20](#_Toc120641235)

[2.3.1.2 Extração de JSON 20](#_Toc120641236)

[2.3.1.3 Extração de PDF 21](#_Toc120641237)

[2.3.2 Mineração de Dados 22](#_Toc120641238)

[2.3.2.1 Algumas Técnicas de Mineração de Dados 25](#_Toc120641239)

[2.3.2.2 Softwares para mineração de dados 26](#_Toc120641240)

[2.4 Descoberta de Conhecimento em Base de Dados (KDD) 30](#_Toc120641241)

[2.5 Trabalhos Relacionados 33](#_Toc120641242)

[3 METODOLOGIA 34](#_Toc120641243)

[3.1 Variáveis e Escopo do Projeto 34](#_Toc120641244)

[3.2 Materiais e Métodos 35](#_Toc120641245)

[3.2.1 Python 36](#_Toc120641246)

[3.2.2 Pandas 37](#_Toc120641247)

[3.2.3 Pandasql 37](#_Toc120641248)

[3.2.4 Google Colaboratory 38](#_Toc120641249)

[3.2.5 Request 38](#_Toc120641250)

[3.2.6 XPath 39](#_Toc120641251)

[3.2.7 Locators 39](#_Toc120641252)

[3.2.8 Regex 40](#_Toc120641253)

[3.2.9 Orange 40](#_Toc120641254)

[3.3 Extração dos Dados 41](#_Toc120641255)

[3.4 Análise dos Dados 48](#_Toc120641256)

[4 RESULTADOS 53](#_Toc120641257)

[4.1 Coleta dos dados 53](#_Toc120641258)

[4.2 Análise Exploratória dos Dados 53](#_Toc120641259)

[5 DISCUSSÃO DE RESULTADOS 61](#_Toc120641260)

[6 Considerações Finais 63](#_Toc120641261)

[6.1 Trabalhos Futuros 64](#_Toc120641262)

[7 REFERÊNCIAS BIBLIOGRÁFICAS 66](#_Toc120641263)

# INTRODUÇÃO

Carro é um bem móvel muito popular, seja por sua facilidade de locomoção ou por conforto de deslocamento para o local desejado. O mercado de carros usados pode ser uma alternativa para quem procura não pagar tão caro em um carro. Segundo Giordan (2022), o mercado de usados passou por elevado crescimento no ano de 2021, sendo mais de 10.000.000 de veículos usados comercializados, porém diversos obstáculos podem surgir na hora de escolher e comprar um carro como preço, modelo, marca e entre outras características. Além disso, fatores externos como falta de componentes no mercado para fabricação, crises econômicas e até mesmo pandemias, podem afetar o preço do carro. De acordo com Rodrigues (2022), a crise de semicondutores no período de 2020 a 2021, somada a motivos políticos, inflação e ao surto de COVID-19 pelo mundo, agravou no aumento do preço dos carros não só no Brasil, mas em outros países também, como nos Estados Unidos, fazendo com que alguns modelos estejam até 28% mais caros em relação a anos anteriores.

Uma das alternativas de procurar um carro usado com um preço mais atrativo é buscar veículos de leilões usados dado que o mercado de leilões de veículos usados no formato online cresce cada vez mais no Brasil, um dos motivos é o fato dos usuários dessa plataforma poderem participar dos leilões em diferentes localidades do país, arrematar o bem no lugar diferente do qual o usuário se encontra e o preço mais atrativo. De acordo com a matéria do Auto Esporte (2020), a grande vantagem de participar de leilões é o preço atrativo do veículo a ser arrematado, podendo ser até 20% abaixo do valor da tabela FIPE (Fundação Instituo de Pesquisas Econômicas). O objetivo de todos os participantes de leilões é arrematar o bem desejado por um preço que considere aceitável, normalmente menor que o valor de mercado.

A maioria das plataformas online de leilões de veículos usados dispõe de informações suscintas, obtendo apenas informações básicas como, fotos, condição do carro, quilometragem, pequenos defeitos, dados do edital do lote e entre outras características e, também, aqueles que estão participando presencialmente no leilão não podem realizar testes para melhor avaliação do estado do veículo, de forma que não há tantas informações para auxiliar numa tomada de decisão. Sendo assim, é natural que os usuários adotem estratégias, conhecimentos prévios e, até mesmo, questões afetivas, como apoio à tomada de decisão.

Dado o contexto apresentado a respeito dos obstáculos de se comprar um carro usado com base nas suas características e preço num site de leilão e, que grande parte de sites e plataformas online de leilões de carros usados e seminovos dispõem informações suscintas em relação as características dos carros, é relevante realizar uma extração e análise de dados da plataforma Sodré Santoro, a maior organização de leilões presenciais e online da América Latina, para gerar um relatório que contém informações com a identificação dos lotes desfavoráveis para arremate, informando e alertando o leitor quais lotes não são interessantes.

O presente trabalho responderá a seguinte questão: como uma ferramenta de extração e análise de dados pode identificar os lotes de carros desfavoráveis com base nas suas características?

O objetivo geral deste trabalho é desenvolver uma ferramenta de extração e análise de dados, transcrever em um relatório, que consiga indicar lotes dos carros com características desfavoráveis e alertar o leitor sobre esses índices.

Para alcançar o objetivo geral, foi proposto desenvolver um *bot* *web crawler* com o objetivo de percorrer as páginas de lotes do site de leilão de carros usados Sodré Santoro, a maior organização de leilões presenciais e online da América Latina, com as seguintes carrocerias, hatch, suv, sedan, crossover e picapes que não possuem sinistro para extrair os dados referente aos lotes.

Realizar a análise dos dados coletados que visa a investigar quais são os lotes desfavoráveis com base nas características dos carros e transcrever para um relatório.

Este trabalho está disposto da seguinte forma: No capítulo 2 é abordado a fundamentação teórica, onde os conceitos essenciais ao trabalho são discorridos. O capítulo 3 é a explicação de como foi realizado o trabalho, quais tecnologias e como elas foram utilizadas, ou seja, a metodologia. Os resultados são demonstrados no capítulo 4. No capítulo 5 é apresentado a discussão dos resultados. O capítulo 6 é dedicado as considerações finais.

Este trabalho está disposto da seguinte forma: No capítulo 2 é abordado a fundamentação teórica, onde os conceitos essenciais ao trabalho são discorridos. O capítulo 3 é a explicação de como foi realizado o trabalho, quais tecnologias e como elas foram utilizadas, ou seja, a metodologia. Os resultados são demonstrados no capítulo 4. No capítulo 5 é apresentado a discussão dos resultados. O capítulo 6 é dedicado as considerações finais.

# FUNDAMENTAÇÃO TEÓRICA

Esta seção é destinada a explicar os principais conceitos que serão tratados neste trabalho: Dados e Informação, Extração e Mineração de dados e Descoberta de Conhecimento em Base de Dados (KDD). Além disso, trabalhos relacionados serão abordados em relação ao presente trabalho.

## Dados e Informação

Com a chegada da era digital a produção de dados aumentou espantosamente, de forma que dispositivos com maiores capacidades de armazenamento são produzidos para suportar tal demanda. A demanda por dados é tão grande que nas principais plataformas de entretenimento em apenas 60 segundos são consumidos: 70.017 horas assistidas na Netflix, 452.000 tweets enviados no Twitter, mais de 80.000 acessos ao Facebook, 3.5 milhões de buscas feitas no Google, 46.200 postagens realizadas no Instagram e entre outras plataformas (POPOMARONIS, 2017). Os dados são muito importantes e valorosos, o processamento dos dados pode levar a resultados expressivos em diversas áreas, seja em empresas, estudos ou pesquisas acadêmicas.

Os dados por característica não representam algo significativo e que por si só não expressa nenhuma mensagem que promova a compreensão a respeito de determinado contexto. O seu conteúdo pode ser letras, números, palavras e entre outros símbolos. De acordo com Setzer (1999, p. 2) dados são elementos de uma série de símbolos que sejam quantificáveis como, palavras, algarismos, letras, mesmo que não há significado individual para quem está lendo. Além disso, a sua descrição pode ser demonstrada formalmente e estruturalmente.

Segundo SOARES et al (2017, p. 4), os dados digitais podem ser caracterizados em três tipos: Estruturados, Semiestruturados e Não-estruturados:

* Estruturados: São dados que facilmente são interpretados por um computador por possuírem um padrão fixo e constante. Um exemplo disso é um banco de dados, onde os dados podem ser organizados através de linhas e colunas.
* Semiestruturados: São uma forma de dados estruturados que não está de acordo com uma modelagem de dados definida, como bancos de dados relacionais ou outras formas de tabelas de dados. Um exemplo de dados semiestruturados são os e-mails, arquivos XML e páginas HTML.
* Não-estruturados: São o oposto dos dados estruturados, ou seja, não tem uma estrutura bem definida e não estão dispostos de forma organizada a fim de serem processados efetivamente por computadores. Alguns exemplos: textos diversos, relatórios, imagens, arquivos de áudio e arquivos de vídeo, chats, e-mails, entre outros.

Para melhor compreensão dos dados, um arquivo que dispõe das características de carros sobre leilão de carros usados como, cor, quilometragem, ano, marco, preço e entre outros atributos, esses dados por si só não contém algum significado, é necessário processá-los para que possa atribuir e assimilar alguma informação útil.

Em suma, dados são armazenados em pequenas, médias, grandes e gigantes quantidades e, com um armazenamento e tratamento adequado para esses dados, podem vir a se tornar dados úteis, dados que podem ser transformados, futuramente, em informação, e levar ao auxílio na tomada de decisão.

A informação surge a partir do tratamento que é feito sob os dados recebidos. Dados que são organizados ou processados de tal forma que se obtém um valor agregado. Quando os dados se tornam organizados, arranjados e relacionados de uma maneira significativa, eles se tornam informação, ou seja, para que o dado seja transformado em informação, é necessário um processamento desse dado. Segundo Becker (2015, p. 35) os dados são matérias essenciais para a formulação da informação e que quando organizados e contextualizados agregam valor e significado para quem utiliza essa informação. A informação quando interpretada e contextualizada por quem a consome pode apoiar à tomada de decisões, ou seja, os processos desde os dados, passando pela informação e ao chegar na tomada de decisão são subjetivos, pois implica-se na interpretação e contextualização por parte de cada indivíduo.

Assim como no exemplo anterior a respeito do arquivo de dados de leilão de carros, o processamento de dados trata, organiza e é realizada alguma análise com o objetivo de obter informações desses dados. Como situação hipotética, os dados foram organizados em uma tabela de acordo com seus atributos e identificou-se que os carros da marca X possui a maior frequência e um valor médio menor em relação as outras marcas. A partir dessa informação, quem a consumir pode ter um melhor entendimento dessa base de dados e influenciar no seu processo decisório em uma futura compra de carros usados num leilão.

Uma melhor compreensão de como o dado se torna informação é demonstrada na Figura 1, a qual ilustra o fluxo pelo qual o dado percorre até se tornar informação, passando pelo processamento e, por fim, se tornando informação.

**Figura 1 - Dados, processamento e informação**

Diagrama

Descrição gerada automaticamente

Fonte: Adaptado de Monteiro (2019)

Como foi exposto, os dados são o alicerce para a informação. Os dados sozinhos não são capazes de retratar um contexto ou situação por completo. De outro lado, a informação possui conteúdo compreensível, de forma que comunica uma situação ou cenário.

Para a transformação de dados em informação útil é necessário processá-los, a ciência de dados é uma das áreas que realizam o processamento dos dados com o objetivo de extrair informações úteis. Rauntenberg e Carmo (2019, p. 58) associam a ciência de dados como a camada dos métodos de transformação dos dados em informação. Por causa do Big Data, essa área tem uma demanda crescente devido a necessidade de extração de informações úteis de bases de dados complexas. Como ilustra a Figura 2, a área de Ciência de Dados envolve os principais conceitos: Ciência da Computação, Matemática e Estatística e Conhecimento do Domínio.

**Figura 2 - Interdisciplinaridade da Ciência de Dados**

Diagrama, Diagrama de Venn

Descrição gerada automaticamente

Fonte: Adaptado de Conaway (2010)

Os profissionais que atuam com a ciência de dados necessitam ter o domínio dessas áreas, pois o processamento de dados nesse contexto utiliza-se de métodos que envolvem a relação desses campos, ou seja, por mais que eles sejam especialistas em um determinado campo da ciência de dados, é necessário conhecer a base das outras áreas, como por exemplo a análise de dados e visualização dos dados, os quais exigem conhecimentos matemáticos e estatísticos. Além disso, as outras disciplinas se correlacionam com o tema de aprendizado de máquina que busca aplicar algoritmos para extrair padrões, antes não identificados, nas bases de dados. O especialista no domínio é essencial para auxiliar no processo decisório, utilizando as informações geradas através das outras etapas.

Segundo Bugnion, Manivannan e Nicolas (2017), dentro da Ciência de Dados há o Ciclo de Vida Ciência de Dados, que são as etapas pelas quais os dados passam com o objetivo de extrair informações úteis. Na figura 3 é ilustrado o caminho que os dados percorrem em cada etapa e que pode ser realizado de forma iterativa independente do modelo de algoritmo de aprendizado de máquina escolhido.

**Figura 3 - Ciclo de vida Ciência de Dados**

Diagrama

Descrição gerada automaticamente

Fonte: Rauntenberg e Carmo (2019)

* **Obtenção de Dados:** Essa etapa do processo estabelece a escolha e avaliação dos dados primários e seus metadados, sejam eles de monitoramento de um conjunto de sensores, dados obtidos da web e entre outros.
* **Ingestão de Dados:** Nesta etapa a transformação e os conjuntos de dados obtidos de diversas origens e de formatos variados em uma única base de dados.
* **Exploração de Dados:** É a etapa a qual a investigação dos dados acontece, de forma que é possível definir questões iniciais em relação aos dados. Além disso, a noção de quais tipos de dados estão sendo utilizados
* **Definição dos Parâmetros:** Está relacionado ao tipo de algoritmo de aprendizado de máquina que será utilizado, ou seja, para um determinado tipo de algoritmo de máquina requer parâmetros específicos.
* **Implementação do Modelo:** Através dos algoritmos de máquina que foram definidos e seus parâmetros escolhidos, de forma iterativa, os modelos envolvem técnicas de treinamento e testes dos algoritmos, ajustando os parâmetros de acordo com a avaliação. O modelo mais apropriado é aquele que melhor reflete estatisticamente os dados utilizados.
* **Utilização do Modelo:** Após o modelo estar consolidado é capaz de utilizá-lo para extrair informações úteis em relação aos dados interessados.
* **Tomada de Decisão:** A última parte do processo está centrada na junção das informações adquiridas através da análise de dados e do gestor que irá tomar decisões com o auxílio das informações. Nessa etapa, também, está relacionada a questão da visualização dos dados, expondo os resultados obtidos da análise em gráficos, tabelas ou relatórios.

Exposto os conceitos de dados, informação e como a ciência de dados é importante na transformação de dados em informações úteis, nota-se que esse é um processo interativo, ou seja, as questões inicialmente estabelecidas podem ser polidas durante o processo e ser ajustadas de acordo com a demanda desejada.

## Tipos de Fontes de Dados

Com a passar dos anos negócios que usam informação para a melhor tomada de decisão e operação mais eficiente tiveram crescimento enormes, essa informação e gerados pelas grandes quantidades de dados coletados e processados para identificar informações útil paras necessidades do negócio, de acordo com "The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things” a quantidade de dados desestruturados em 2020 é de 40 ZB(IDC, 2014), o provimento do Big Data como um serviço facilita muito a implementação de outros serviços como tecnologia de pesquisa, deep data analytics para identificar padrões ocultos. Com a coleta e processamento dos dados armazenados na big data é possível retirar um grande valor para as companhias com a redução do tempo humano gasto e implementação de outros serviços e “insights” que só são possíveis com uma quantidade exorbitante de dados.

O conceito de big data vem da premissa que existem datasets tão grandes e com tantas estruturas que que excede a capacidade de ferramentas de programação tradicionais como databases, softwares e etc, para coletar, armazenar e processar em um tempo razoável e principal supera a capacidade de percepção do humano em entender a informação, segundo Agrawal et al (2011), heterogenia, escala, propensão, complexidade e problemas com privacidade impedem que seja possível tirar valor dos dados em um ambiente de big data, essas estruturas e pluralidade de tipos de dados e como os computadores modernos conseguem armazenar e atualizar dados. O conceito de big data vem com o objetivo de superar a condição humana de processamento e armazenamento de dados e consequentemente a de assimilá-los.

Os dados podem ser classificados como dados primários e secundários e podem vir de fontes internas ou externas, segundo Ayswarrya G (2022), estrategista de conteúdo da Atlan, dados primários são os que você ou sua organização criam, exemplo, entrevistar pessoas para obter feedback sobre seu produto, os dados da entrevista são dados primários. Dados secundários, os dados que você coleta de outra pessoa e você ou sua organização controlam, exemplo, dados do Google Analytics para saber quantas pessoas visitam seu site, são dados secundários. Os dados internos são os dados privados que sua organização possui, controla ou coleta, os dados de vendas ou dados financeiros de sua organização são bons exemplos de dados internos e dados externos são coletados de fontes externas à sua organização, por exemplo fazer crawling de sites de terceiros e guardá-los em um banco de dados, este exemplo seria calcificado como dado primário e externo.

Dados estruturados são organizados para encaixar perfeitamente em planilhas e bancos de dados relacionais, por outro lado, esse tipo de dado precisa de sistema para dar contexto a esses dados. Dados desestruturados como texto, áudio, imagem, vídeos ou NoSQL são compreendidos diretamente por humanos, mas pode ser desafiadora de analisar por programas ou em grandes quantidades, segundo Tobias Geisler Mesevage estima se que 20% dos dados sejam estruturados e 80% não, o número de dados desestruturados tende a crescer muito mais rápido.

**Figura 4 - Diferenças de dados estruturados**

Interface gráfica do usuário, Diagrama

Descrição gerada automaticamente  
Fonte: Adaptado de Altexsoft (2020)

## Extração e Manipulação de Dados

A extração de dados web pode ser executada de diferentes formas desde as mais manuais com os famigerados comandos CTRL+C e CTRL+V, até softwares altamente sofisticados que conseguem capturar centenas de informações por minuto com diversos processos decisórios imbuídos no programa, as técnicas amplamente usadas são:

1. Copia e Cole.
   1. Processo altamente manual executado por humanos que é suscetível a erro e pode ser entediante quando há muitos dados;
2. Regular expressions
   1. Abordagem poderosa para extrair textos de páginas web essa abordagem é empoderada se usada com auxílio de alguma linguagem de programação;
3. Hypertext Transfer Protocol (HTTP) Programming
   1. Técnica para extração de dados fazendo post e gets para servidores web;
4. Hyper Text Markup Language (HTML) Parsing
   1. Linguagem de query semiestruturadas como o XQuery pode ser usado para transformar páginas HTML em conteúdo;
5. E. Document Object Model (DOM) Parsing
   1. Linguagens de consulta (Query Language) como XPATH e CSS Selector fornecem a capacidade de navegar e extrair conteúdo de um DOM (Document Object Model) usando uma variedade de critérios;
6. Software de Web Scraping
   1. Há softwares que automaticamente reconhecem a interface de uma estrutura de dados e armazena esse conteúdo localmente removendo a necessidade de escrever manualmente um código de Web Scraping.

### Extração de Dados

O Web Scraping é uma excelente técnica para extração de dados desestruturados de websites e transformar em dados estruturados que podem ser armazenados e analisados em um banco dados, essa técnica também e conhecida por outros nomes com web data extraction, web data scraping, web harvesting ou screen scraping.

O objetivo de um Web Scraping é pegar dados como preços de produtos em lojas, preços de ações, relatórios e etc. e transformar isso em uma estrutura compreensível como CSV (Comma-separated values), banco relacional ou não relacional (Sirisuriya, 2015). Na figura 4 é apresentado o processo de extração de dados.

**Figura 5 - Processo de Extração de dados**

Texto

Descrição gerada automaticamente com confiança média

Fonte: Adaptado de Sirisuriya (2015)

#### Extração de HTML

A maior parte de informação acessível ao público na internet é exibido como HTML para os usuários, essa informação vem de maneira semiestruturada dentro de tags para que a forma do texto tenha sentido para o leitor, ferramentas como Python e bibliotecas de extração de informação em htmls como o requests\_html e be Beautiful Soup são baseadas em padrão de textos que identificam o início e o fim de dados relevantes, não é usado inteligência artificial para entender os conteúdos e é eficiente para analisar grande volume de informação(J. Hammer), contudo, se há mudança no padrão do HTML é necessário que em contrapartida o hajam alteração na lógica de extração, algo que facilita a extração de informação e o uso de XPath para o requests\_html e Regex em soma ao Beautiful Soup, assim que extraído pode se fazer a sumarização da informação e passa de semiestruturada para estruturada, isso é vantajoso já que informação estruturada é melhor aproveitada por programas de computador e pelo negócio em geral.

#### Extração de JSON

JSON (JavaScript Object Notation) é o formato mais popular para requisições e respostas de API's, ainda falta um esquema padronizado ou definição de metadados que permite aos desenvolvedores especificar a estrutura de documentos JSON, de acordo com Stack Overflow, agora são feitas mais perguntas sobre JSON do que sobre outros formatos de intercâmbio de dados. Para fazer um request para uma API usando Python usamos a biblioteca Requests que prove o suporte necessário para o envio e recebimento de dados, no qual podemos enviar um dicionário com os parâmetros e autenticação quando necessário e receber uma resposta que pode ser facilmente convertida em um dicionário Python com o método "response.json()". No geral JSON podem ser muito facilmente manipulados por Python e os dados normalmente vem estruturados e bem sumarizados.

**Figura 6 - Frequência de perguntas sobre JSON no StackOverflow**

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Fonte: Seva Safris(2019)

#### Extração de PDF

O PDF que foi desenvolvido pela Adode em 1993, é um formato de arquivo para representar documentos de maneira independente do aplicativo, do hardware e do sistema operacional usados para criá-los. Os pdf podem ser armazenados bancos de dados por ByteStrings que quando passada por interpretadores de pdf como Adobe Acrobat Reader e Google chrome tem sua informação reestabelecida e visualizada, para extrair o conteúdo de um pdf usando bibliotecas do python como pdfminer e pdftotext que conseguem extrair o texto que posteriormente pode ser seleciona e sumarizado com o uso de Regex. A principal dificuldade de se extrair dados de pdf é que além de não ter uma estrutura de dados interna, o texto contido vem corrido e pode apresentar artefatos presentes de outros elementos textuais.

### Mineração de Dados

O conceito de mineração de dados é comumente confundido com diversos aspectos do processamento de dados. De acordo com Aggarwal (2015, p. 1), o conceito de mineração de dados é o estudo da coleta, limpeza, processamento, análise e obtenção de insights úteis que agregarão valor a partir dos dados.

A mineração de dados assim como a mineração de ouro consiste em retirar o ouro de pedras e areia, analogamente podemos nomear como "Knowledge Mining From Data", O termo mineração é o processo ativo de encontrar um pequena quantidades de pepitas preciosas de uma grande quantidade de matéria bruta, essa escolha de palavras se tornou popular por carregar o termo "mineração" e "dados", mas existem outros termos com significado similares, como por exemplo Knowledge Mining From Data, Knowledge Extraction, Data/Patterns Analysis, Data Archaeology e Data Dredging.

O termo é tratado por alguns como sinônimo de outro termo conhecido o KDD ou Knowledge Discovery in Databases, outros veem o processo de mineração como um passo essencial dentro do processo de KDD.

O processo de KDD consiste em 7 etapas:

1. Limpeza de dados (Remover barulho e dados inconsistentes)
2. Integração de dados (Combinar diferentes fontes de dados)
3. Seleção de dados (Recuperar dados relevante para a tarefa de análise)
4. Transformação de dados (Quando os dados são propriamente transformados em formulários para as operações de sumarização e agregação)
5. Mineração de dados (processo de extração de padrões dentro dos dados)
6. Avaliação de padrões (Identificar os padrões realmente interessantes baseado nas medidas de interesse)
7. Apresentação de conhecimento (Onde técnicas de visualização e representação de conteúdo são apresentados ao usuário)

(HAN e PEI, 2011)

Para Berry e Linoff (1997, p.7), a mineração de dados é a exploração e análise de grandes quantidades de dados com o objetivo de descobrir padrões e regras significantes.

Assim como o processo de KDD, a mineração de dados está sujeita a um processo e sequência de etapas a serem percorridas, as quais são sequencialmente lógicas. Podemos compreender que a mineração de dados faz parte do processo de KDD. Na figura 5 é demonstrado o processo de mineração de dados que consiste em 3 etapas.

Na etapa da Coleta de Dados, podem ser utilizados dispositivos, técnicas e meios automatizados para realizar a coleta dos dados, como: hardware, sensores, formulários, *bots* especializados em extração de dados (*crawlers*) e entre outras opções. Vale ressaltar que a escolha correta da fonte a ser coletada é interessante e pode afetar o desempenho da análise durante o processo.

**Figura 7 - Processo de mineração de dados**

Diagrama

Descrição gerada automaticamente

Fonte: Adaptado de Aggarwal (2015)

Os dados possuem tipos variados, podem ser categóricos, quantitativos, texto e entre outros tipos, por isso, a partir dos dados coletados é realizada a atividade de limpeza dos dados, a qual visa padronizar e moldar os dados a determinados algoritmos de mineração de dados, visto que os diferentes algoritmos requerem padronizações para a sua execução.

A parte final do processo, processamento analítico dos dados. Essa etapa final define quais algoritmos ou técnicas que serão usadas para o problema estabelecido. Determinados tipos de algoritmos e técnicas irão se adaptar ou resolver melhor o problema do que outros, por isso, é importante projetar ou escolher as técnicas e métodos analíticos mais adequados.

Há inúmeras ferramentas e softwares para mineração de dados que assessoram no processo de mineração de dados. Weka (*Waikato Environment for Knowledge Analysis*), uma das ferramentas open source mais populares, foi desenvolvida pela Universidade de Waikato na Nova Zelândia, suas funcionalidades contam com diversos tipos de algoritmos voltados a aprendizado de máquina e mineração de dados (FRANK et al., 2016). O programa abrange desde a etapa de pré-processamento até algumas opções de visualização dos dados, além disso, a ferramenta traz dados experimentais para testar os algoritmos disponíveis.

#### Algumas Técnicas de Mineração de Dados

Existem inúmeras técnicas de mineração de dados, porém outras são mais populares por resolver problemas mais recorrentes ou realizar a extração de conhecimento nas bases de dados.

Cada técnica de mineração se enquadra melhor em certos tipos de problemas e tarefas. Podemos defini-las em dois grupos: preditivos e descritivos. Na mineração de dados preditiva, os atributos são utilizados como alvo para prever novas instâncias de valores, ou seja, a sua finalidade é a criação de um modelo preditivo que prevê o valor alvo definido do atributo.

Por outro lado, no modelo de mineração de dados descritiva implica na descrição de padrões escondidos nos dados. O modelo é focado para identificar a relação entre os dados e não prever valores. Um caso de uso para o modelo descritivo é a tarefa de agrupar e reconhecer regras de associação que descreve a relação dos atributos.

De acordo com Camilo e Silva (2009, p.8), as tarefas que são realizadas durante o processo de mineração de dados são divididas e classificadas em grupos, além disso, é necessário julgar quais são as tarefas que melhor se adequam com base nos objetivos pré-determinados. As tarefas mais comuns são:

* **Descrição (*Description*)**: empregada para descrever os padrões e tendências encontradas pelos dados. Esta tarefa é frequentemente aplicada em técnicas de análise exploratória de dados para certificar a influência de certas variáveis no resultado obtido.
* **Classificação (*Classification*)**: a Classificação procura reconhecer a qual classe um certo registro faz parte que pode ser utilizado em dados não classificados com o objetivo de categorizá-los. O modelo selecionado explora o grupo de registros estabelecidos, com cada registro que possui a indicação de qual classe o pertence, o modelo visa “aprender” como classificar um novo registro, ou seja, a classificação determinará a qual grupo, que foi classificado anteriormente, o dado que mais coincide.
* **Estimação (*Estimation*) ou Regressão (*Regression*)**: semelhante a tarefa de Classificação, entretanto a Regressão é utilizada quando o registro é definido por um valor numérico viabilizando a estimar o valor de uma determinada variável a qual esse tipo se assemelha a mesma grandeza.
* **Predição (*Prediction*)**: essa tarefa visa identificar um valor futuro de um atributo preestabelecido, similar às tarefas de Classificação e Estimação. Podem ser utilizados alguns métodos de Classificação e Regressão para a predição, entretanto, com devidas ressalvas.
* **Agrupamento (*Clustering*)**: procura reconhecer um valor limitado de possíveis agrupamentos ou clusters a partir dos dados, de forma que os registros com valores parecidos são agrupados em um mesmo cluster.
* **Associação (*Association*)**: A tarefa de Associação tem como finalidade a identificação de relações entre os atributos.

#### Softwares para mineração de dados

Com a grande demanda por extração de informações úteis, seja em negócios ou pesquisas acadêmicas, atualmente, há diversas soluções e ferramentas disponíveis para uso, tanto comerciais quanto open-source, utilizando códigos, scripts e bibliotecas ou até mesmo através de fluxos de componentes, não necessitando codificar os algoritmos.

De acordo com Mikut e Reischl (2011), há grupos de métodos que se baseiam em estatística clássica e outros métodos que se originaram do campo da inteligência artificial. Esses métodos são desenvolvidos com base em artigos teóricos de protótipos de software desenvolvido por organizações ou instituições e distribuídos apenas os algoritmos de sucesso. Os algoritmos podem ser de acesso público ou comerciais e, normalmente, alguns dos algoritmos são similares. Muitas companhias utilizam soluções de mineração de dados em suas bases de dados, popularizando certos grupos de métodos, como por exemplo, a inteligência artificial. Além disso, algumas empresas ganharam popularidade através dos seus produtos que fornecem soluções de mineração de dados, empresas como IBM e Oracle se destacam nesse contexto.

As ferramentas open-source não deixam em nada a desejar, elas estão cada vez mais populares e ganhando espaço na área de mineração de dados, justamente por atender às demandas dos usuários. Atualmente, ferramentas como WEKA são relevantes na área, pois seus componentes são utilizados em outras ferramentas de mesmo objetivo. Programas matemáticos que utilizam scripts orientados são utilizados para a mineração de dados, mesmo que o seu propósito não seja para esse tipo de solução, entretanto, as funções presentes auxiliam na tarefa de mineração e visualização dos dados.

Há dezenas de ferramentas disponíveis para mineração de dados e algumas são mais completas que outras, possibilitando o processo de KDD por completo. Algumas ferramentas são mais utilizadas por certos grupos pelo fato de atenderem suas necessidades da melhor forma, ou seja, usuários que estão procurando ferramentas que proporcionam soluções de mineração de dados em negócios irão optar por determinadas ferramentas do que um pesquisador que também está utilizando alguma ferramenta de mineração de dados para sua pesquisa.

De acordo com Wolff (2020), algumas das ferramentas mais populares atualmente, sendo elas para aplicações comerciais ou pesquisas, de código aberto ou que necessitam pagá-las para utilizar:

* **RapidMiner:** A ferramenta RapidMiner é uma das plataformas open-source e gratuita mais populares atualmente a qual muitos cientistas de dados utilizam. A plataforma contém diversos algoritmos para mineração de dados como, aprendizado de máquina, aprendizado profundo e entre outros. É possível, também, realizar o pré-processamento dos dados para padronizá-los ao algoritmo escolhido. O seu uso é através de uma interface que permite a manipulação dos componentes apenas arrastando e soltando, permitindo que usuários que não possuem experiência em programação possam criar alguns modelos.
* **Weka:** A ferramenta Weka foi desenvolvida pela universidade Waikato na Nova Zelândia com a linguagem de programação Java. A ferramenta é de código aberto e possui licença GPL (Generic Public License) e seu uso é gratuito. Esse software é munido de diversos algoritmos para mineração de dados, permitindo diversas tarefas desse contexto, realizar o pré-processamento dos dados e uma visualização simplória dos dados. Há uma interface gráfica que facilita a utilização e visualização dos algoritmos para o usuário, entretanto, é necessário possuir algum conhecimento dos algoritmos para entender como usá-los e interpretar os resultados. Além disso, a ferramenta é bem famosa no meio acadêmico, devido a seu fácil uso para quem tem o mínimo de conhecimento em mineração de dados e está procurando alguma solução para sua pesquisa nesse sentido.
* **Orange:** A ferramenta Orange é gratuita e open-source, oferece os algoritmos de aprendizado de máquina e soluções para mineração de dados. Muitos cientistas e professores a utilizam pelo fato de possuir interface gráfica, facilitando a manipulação dos fluxos de componentes e a visualização e ilustração dos conceitos da ciência de dados. A ferramenta também oferece o pré-processamento dos dados e diversas opções de manipular os dados de acordo com o algoritmo desejado. Um dos pontos relevantes da ferramenta é a visualização dos dados, que oferece numerosos tipos de gráficos, distribuições estatísticas e entre outras projeções para os seus dados. A possibilidade de utilizar extensões conta como ponto positivo, pois agrega ao arsenal de técnicas de mineração de dados, como a extensão de mineração de regras de associação.
* **Oracle Data Mining:** A ferramenta Oracle Data Mining apresenta recursos muito poderosos, assim como as outras ferramentas, é possível utilizá-la para soluções em mineração de dados. A ferramenta é um componente do Oracle Advanced Analytics Option, um produto da Oracle com foco no Big Data. O software é capaz de realizar tarefas de predição, descrição, regressão, classificação, detecção de anomalia e demais tarefas da área de mineração de dados, consultas em SQL em grandes bases de dados é uma vantagem dentre as outras ferramentas e opera com fluxos de componentes de arrastar e soltar numa interface gráfica. É possível utilizar a versão experimental pelo período de 30 dias, após isso é necessário pagar.
* **IBM SPSS Modeler:** Essa ferramenta proporciona recursos de ciência e análise de dados para as empresas que desejam extrair valor das suas bases de dados, por isso possui grande representatividade no mercado. Equipada com algoritmos de aprendizado de máquina, é possível realizar diversas tarefas de mineração de dados, proporcionando a descoberta de padrões e comportamentos nas bases de dados. A ferramenta disponibiliza diversas formas de visualizar os dados e uma interface gráfica que permite arrastar e soltar os componentes desejados. É possível realizar as operações na nuvem ou localmente. Assim como a ferramenta Oracle Data Mining, pode-se utilizar da versão experimental por 30 dias, após esse prazo é necessário pagar pelo produto.

Vale mencionar a biblioteca Scikit-Learn que é feita em Python e dispõe de diversos algoritmos de aprendizado de máquina. O projeto começou em 2007 como projeto do Google Summer e após 3 anos outros programadores deram continuidade no programa. A biblioteca combinada com outras bibliotecas e frameworks, como por exemplo, NumPy, Matplotlib e Pandas, oferecem um conjunto de manipulação e análise de dados completo, desde o pré-processamento até o uso de diversos tipos de algoritmos de aprendizado de máquina (SCIKIT-LEARN, 2022).

As bibliotecas podem ser utilizadas em plataformas cloud, como o Google Colaboratory, a qual é possível executar os scripts junto das bibliotecas remotamente.

Como foi exposto, há uma vasta gama de ferramentas de mineração de dados e suas soluções, disponíveis de forma gratuita, open-source, código aberto ou pagas. Entretanto, um dos desafios em relação as ferramentas é que alguns requerem mais conhecimento técnico dos algoritmos de mineração de dados e conceitos de ciência de dados do que outras, que buscam um uso mais intuitivo, não precisando codificar, apenas utilizando fluxo de componentes através de uma interface gráfica.

## Descoberta de Conhecimento em Base de Dados (KDD)

Com grandes volumes de dados armazenados nos bancos de dados de diversas empresas e instituições, o dado se tornou um recurso essencial para análises e projeções estatísticas com o objetivo de usá-las em tomadas de decisão, apoiando estratégias que buscam vantagens competitivas ou benefícios comerciais e científicos. Devido as enormes quantidades de dados armazenados, não é mais viável analisar, sintetizar e extrair conhecimento como era feito tradicionalmente. O processo de extração de conhecimento em base de dados é usado para suprir a imensa demanda atual disposta pelo mercado.

A descoberta de conhecimento em base de dados permite a identificação de padrões e comportamentos através de técnicas que utilizam dados, gráficos e algoritmos, de forma que o produto seja a obtenção de conhecimento (DANTAS et al., 2008). Além disso, as técnicas envolvem a interdisciplinaridade de áreas de estatística, inteligência artificial, banco de dados e sistemas de informação.

A utilização do KDD, descoberta de conhecimento em base de dados, é consumida por diversos setores que colhem e armazenam grandes quantidades de dados, principalmente de usuários e consumidores de produtos e serviços. Na área comercial, por exemplo, o marketing emprega o KDD em suas bases de dados para extrair categorias de compradores de produtos específicos e, até mesmo, identificar padrões e prever determinados comportamentos de consumo. No campo acadêmico, o KDD é utilizado na astronomia a fim de identificar, catalogar e classificar objetos em imagens do céu (FAYYAD, 1996).

De acordo com FAYYAD et al. (1996, p.39), o KDD abarca um processo como todo a descoberta de conhecimento útil que tenha valor a partir dos dados. O processo é composto por fases que têm como propósito a manipulação, processamento e análise dos dados para gerar conhecimento. As fases do KDD serão abordadas na seção seguinte.

As Fases da Descoberta de Conhecimento em Base de Dados

O processo de KDD (Knowledge Discovery in Databases) é interativo e iterativo, passando por etapas importantes que tratarão e processarão os dados de acordo com as técnicas adotadas. A compreensão do domínio estabelecido e as hipóteses a serem testadas também são imprescindíveis. Na figura 5 ilustra de maneira compreensível como percorrem as fases do KDD.

**Figura 8 - Fases do KDD**

Diagrama

Descrição gerada automaticamente

Fonte: Adaptado de Faayad (1996)

Segundo FAAYAD et al (1996, p.40), após compreender o domínio apresentado e formular as hipóteses, a próxima etapa é a seleção dos dados, a qual é selecionada a massa ou base de dados que será submetida às próximas etapas e, consequentemente, a extração do conhecimento.

Em seguida, o Data Cleaning e Pré-Processamento concentra-se na eliminação de ruídos. As estratégias definidas de tratamento de valores nulos encontrados serão utilizadas para garantir a qualidade dos dados.

Após assegurar a integridade dos dados, a fase seguinte é a Transformação dos Dados, a qual foca em apenas utilizar os dados relevantes e que irão agregar valor ao usuário. Além disso, é nessa fase que os dados serão transformados e manipulados para se adaptarem ao padrão do algoritmo de mineração desejado.

A seguir, a fase de Mineração de Dados é uma das mais importantes de todo o processo de KDD, pois é nela que os padrões serão encontrados através dos algoritmos e técnicas que servirão de informação para a elaboração de uma análise mais profunda. Nessa fase, também, é escolhido o tipo de algoritmo e técnica que melhor se encaixa com o objetivo previamente estabelecido, de forma que o resultado contenha mais valor à análise.

A fase seguinte é a de Avaliação ou Interpretação dos Dados, nessa fase será avaliado e interpretado os padrões extraídos. Caso o resultado seja insatisfatório, há a possibilidade de iterar novamente as fases anteriores.

A última fase estabelece o Conhecimento adquirido. O conhecimento consolidado pelas fases predecessoras será introduzido ao sistema desejado ou relatado e documentado aos indivíduos de interesse. O produto do KDD, o conhecimento, irá agregar valor a quem utilizá-lo, seja no meio acadêmico ou comercial.

## Trabalhos Relacionados

DE LUCENA (2018) desenvolveu um *Web Crawler* para extrair os dados do fórum do jogo Warframe e realizar uma análise de dados, entre a atividade do fórum do jogo e a quantidade média de jogadores na plataforma PC, através do processo de KDD, o qual inclui a mineração de dados. Para realizar seu objetivo, foi desenvolvido um *crawler* para extrair dados, previamente definidos, do fórum. Após a extração, os dados foram submetidos ao processo de KDD e ferramentas de mineração de dados. Os resultados obtidos foram em partes, supostos pela autora, de que quando há aumento na média de jogadores na plataforma analisada, a atividade do fórum aumenta também, o oposto também é válido. Além disso, o padrão encontrado foi que a maior atividade do fórum é dos usuários da plataforma PC.

PEREIRA (2017) realizou mineração de dados de páginas de notícias sobre delitos e infrações de trânsito utilizando técnicas de extração de dados. Partindo da questão que o conhecimento obtido a partir da mineração de dados poderia ajudar a população e autoridades a identificar zonas de risco na cidade em que o autor reside, foram identificadas palavras-chave que relacionam a criminalidade. A forma como os dados foram estruturados e analisados, permitem que o conhecimento obtido seja usado para propostas de planos de segurança pública por parte das autoridades locais.

BORGES (2018) realizou o Web Scraping dos dados de carros do site da FIPE com objetivo de analisar a variação de preço dos automóveis de forma que a análise seja uma ferramenta para suporte à tomada de decisão na escolha de um veículo. A extração é feita através de técnicas de Web Scraping e os dados foram salvos em um banco de dados propriamente estruturado para os dados dos automóveis e, posteriormente, analisados de forma estatística.

Embora os trabalhos relacionados possuam conceitos e termos similares, algumas técnicas o diferenciam do presente trabalho.

# METODOLOGIA

Para este trabalho foi essencial realizar estudos sobre técnicas e tecnologias utilizadas para o processo de extração, mineração e análise de dados, tarefas que compõe a descoberta de conhecimento em base de dados. Através dos estudos foi possível identificar quais ferramentas mais viáveis para o desenvolvimento do software responsável pela extração dos dados e quais ferramentas para realizar a análise de dados.

## Variáveis e Escopo do Projeto

O projeto engloba algumas variáveis relacionadas aos dados dos carros extraídos da plataforma Sodré Santoro, as quais são relevantes para análise. De acordo com Miragaya (2021), em seu artigo no site Quatro Rodas, há diversas vantagens em adquirir veículos através de leilões, porém, é necessário se atentar a algumas variáveis pertinentes ao leilão. As variáveis utilizadas neste projeto foram:

* Placa: armazena o número final da placa do carro de acordo com o edital do leilão.
* Modelo: o modelo do carro.
* Marca: o nome do fabricante do veículo.
* Cor: a cor que o carro se encontra.
* Ano: o ano do modelo do veículo
* Quilometragem: quantidade quilômetros rodados.
* Lance Atual: valor em reais(R$) do veículo quando o dado foi extraído.
* Local: o estado federativo do pátio em que o veículo se encontra.
* Combustível: tipo de combustível, sendo as opções de flex, gasolina e diesel.
* Origem: é a origem do lote ou veículo, podendo ser dos tipos: seguradoras, frotas, financiamento, particular, judicial ou lojistas.
* Estado do Chassi: representa a condição e estado do chassi, podendo ser: amassado, danificado, enferrujado, ilegível ou íntegro.
* Câmbio: o tipo de câmbio, com as opções de: manual, automático e semi-automático.
* Ar-Condicionado: se o carro tem ar-condicionado ou não.
* Direção Hidráulica/Elétrica: se o carro possui de tecnologia de direção hidráulica ou elétrica.
* Kit Gás: se o carro dispõe ou suporta o kit gás.
* Blindagem: se a lataria do veículo dispõe de blindagem ou não.

Características com maior frequência apontam para um atributo favorável em relação ao carro já que

Após as variáveis serem selecionadas, através da biblioteca Pandas, elas foram manipuladas e submetidas ao processo de KDD com o objetivo de obter informações e padrões relacionados aos carros para serem utilizadas como suporte à tomada de decisão.

A ferramenta Orange foi utilizada para a visualização e mineração de dados, de forma que o programa permite a criação de um fluxo através dos seus componentes.

## Materiais e Métodos

Os recursos tecnológicos empregados para o desenvolvimento do software de extração de dados e para a análise de dados são: softwares *Open Source* ou versões gratuitas que são disponibilizadas na web. O termo *Open Source* relaciona-se ao acesso livre, distribuição livre e podendo realizar a modificação do código fonte. A expressão está atrelada ao conceito de que o software foi desenvolvido sob uma licença que possibilita a inspeção, uso, modificação e redistribuição do código (GOMES, 2014).

Linguagem de Programação, Tecnologias e Ferramentas Utilizadas:

* Python 3.10;
* Pandas 1.3.5;
* Google Colab (Google Colaboratory);
* Pandasql 0.7.3;
* Requests;
* Locators;
* Regex;
* XPath (XML Path Language);
* Orange.

### Python

O Python foi escolhido pela familiaridade com a ferramenta por alguns membros do grupo e a facilidade de trabalhar com dados no desenvolvimento do programa de extração de dados e *frameworks* para a análise dos dados.

A linguagem de programação foi desenvolvida por Guido van Rossum em 1995 como sucessor de uma linguagem chamada ABC. Apesar de Guido ser o criador do Python, ele conta com diversas contribuições da comunidade.

Python é uma linguagem de programação de alto nível e orientada a objetos, famigeradamente por ser de fácil uso, tanto pela quantidade imensa de conteúdo consolidada por membros da comunidade de programação e outras áreas, quanto pela grande quantidade de *frameworks* que possibilitam inúmeras atividades utilizando os dados, nesse contexto, a mineração de dados. Além disso, possui características que a tornam amigável para quem está começando na área de programação por ser bem documentada (PYTHON, 2022).

A linguagem de programação foi utilizada na maioria das fases do trabalho, ou seja, na extração, armazenamento e análise dos dados. Os *frameworks* utilizados possuem enormes guias, conteúdos e bibliotecas consolidadas pela comunidade, os quais foram essenciais no auxílio e desenvolvimento desse trabalho.

### Pandas

Para a realização do processo de análise de dados a biblioteca Pandas foi essencial no presente trabalho, pois ela permite realizar manipulações com os dados da base de dados para atingir os objetivos e hipóteses estabelecidas.

A biblioteca Pandas feita em Python teve início no ano de 2008 e no ano seguinte, em 2009, se tornou open source. O seu objetivo é preencher a lacuna que a linguagem de programação Python possui no campo de estatística, essa foi uma das justificativas para o seu desenvolvimento pois à época Python não possuía tantas bibliotecas que auxiliassem os usuários nessa área.

Essa ferramenta contém diversas funções e recursos que proporcionam a solução de diversos obstáculos no campo de estatística e análise de dados, principalmente como o *framework* viabiliza a construção de conjuntos de dados estruturados e o tratamento de dados faltantes (MCKINNEY, 2011).

Atualmente a biblioteca conta ativamente com o apoio da comunidade. Pandas é capaz de se equiparar com softwares e ferramentas de mesmo propósito, continua implementando mais recursos e possibilitando indivíduos da área científica e áreas relacionadas a análise de dados a optarem pelo seu uso.

### Pandasql

A biblioteca Pandasql permite a execução de queries em DataFrames do Pandas utilizando a sintaxe SQL. Essa biblioteca é útil para quem busca a utilização da sintaxe SQL no Pandas, facilitando a seleção dos dados desejados utilizando do conhecimento de SQL.

### Google Colaboratory

Para a realização da etapa de análise de dados foi utilizado a plataforma Google Colaboratory ou Colab que é um produto do Google Research. A plataforma permite a execução de código Python arbitrário pelo navegador o qual é voltado para aprendizado de máquina, análise dedados e educação.

O Colab é uma aplicação de notebooks hospedados do Jupyter, ou seja, os arquivos são armazenados em formato ipynb, não requisita de nenhum tipo de configuração para executar e o seu acesso não necessita arcar com custos financeiros (COLABORATORY, 2022).

Nesse contexto, o arquivo .csv que contém a base de dados foi carregado no Colab e, posteriormente, os dados foram tratados e analisados. Além disso, a biblioteca Pandas foi importada, complementando o processo de análise.

### Request

A biblioteca Request foi utilizado na parte de extração de dados pelo fato de que funciona enviando uma requisição para o servidor de destino passando parâmetros como o método HTTP, header, cookies, params, data e payloads e, também, possibilitando o uso de API’s (Application Programming Interface) caso necessário; O custo computacional é extremamente mais baixo e o algoritmo trabalha de forma bem mais performática (HERBERT, 2020).

### XPath

Para o tratamento dos arquivos XML foi utilizada a XPath, XML Path Language, uma linguagem de consulta usada para lidar com arquivos do tipo XML e HTML.

A XPath foi essencial no desenvolvimento pois nem sempre quando queremos fazer uma raspagem de dados em um website precisamos de todos os elementos. Às vezes, apenas um dado em si, serve, seja ele em forma de lista ou tabela, por exemplo.

O diferencial dela é conseguir encontrar pontos específicos em documentos HTML através de expressões que funcionam como caminhos que selecionam um ou mais “nós”. Funcionando como se fosse uma árvore de diretórios onde cada expressão é usada para extrair informações desses “nós”.

Assim, retornando um conjunto de elementos especificados através de um caminho em forma de uma expressão XPath, colaborou na consulta, extração de dados e, consequentemente, no desenvolvimento do trabalho.

### Locators

Uma ferramenta utilizada para localizar elementos de uma página por meio de comandos é o Locators, utilizando o Selenium.

Como sabemos, para interagir com um elemento de uma página primeiro precisamos encontrar esse tal elemento, e é aí que fazemos o uso do Locators, pois nele podemos localizar os elementos através da ajuda dos métodos “findElement()” que nos retorna um WebElement ou “findElements()” que nos retorna uma lista de WebElements.

Com o Locators, podemos procurar os elementos através do id, pelo nome, por texto do link (Link Text), por parte do texto do link (Partial Link Text), pelo XPath, pela tag, pela classe (class) ou até mesmo pelo CSS Path (CSS Selector).

Resumindo, são diversas opções e temos a liberdade de escolher a que melhor se encaixa dentro do nosso contexto. Assim, fazendo uma busca eficiente.

### Regex

Outra ferramenta usada no desenvolvimento do trabalho foi o Regex, ou, Regular Expression. Ela é uma expressão que a princípio, esteticamente, não agrada quem está utilizando, pois ela é escrita através de caracteres e metacaracteres, que quando formados determinam uma função.

Através dela foi possível fazer a detecção de padrões como, por exemplo, no caso: nome, ano, data, placa, cor, km, etc. Assim, a extração das informações que foram necessárias.

Essencial, pois, com ela, podemos realizar buscas por padrões que definem qual é o formato esperado do termo que estamos procurando, por exemplo, buscar todos os termos que tenham um formato de data (dia, mês e ano) mesmo sem saber a data que estamos procurando. É possível realizar validações, por exemplo, validar nas datas encontradas se sua sequência de caracteres atende ao padrão estabelecido. E é possível fazer substituições, já que no Regex há um recurso capaz de extrair apenas uma parte de um padrão para utilizá-la em um destino.

### Orange

A ferramenta Orange dispõe de diversas opções para a realização do KDD, desde a transformação, limpeza, pré-processamento e visualização dos dados, até a disponibilidade de diversos modelos e algoritmos de aprendizado de máquina. A interface é amigável e de fácil uso, de modo que arrastar, soltar e interligar os componentes desejados possibilita a criação de um fluxo do processo.

A escolha dessa ferramenta se deve pela facilidade do seu uso e da não necessidade de programação específica na área de mineração de dados e descoberta de conhecimento em base de dados.

## Extração dos Dados

O site Sodré Santoro é divido em 5 tipos de lotes de leilão (Veículos, Sucatas, Materiais, Imóveis e Judiciais). Estes possuem opções de filtragens para selecionar os lotes de maior interesse. Para este trabalho optou-se por selecionar apenas o lote de veículos (carros), devido as características combinarem melhor com um indivíduo que deseja comprar um carro de leilão, seja uso próprio ou revenda. Depois de selecionada a categoria de carros e que não possuem sinistro, selecionamos os dados que seriam salvos de cada lote, que foram: Placa, KM(Quilometragem), Origem, Câmbio, Direção Hidráulica/Elétrica, Blindagem, Cor, Combustível, Estado do Chassi, Ar-Condicionado, Kit Gás, URL, local, Data e Câmbio (Moto). Na figura 7 ilustra o esquema das características do carro que foram capturadas.

**Figura 9 - Características do veículo**

Uma imagem contendo Interface gráfica do usuário

Descrição gerada automaticamente

Fonte: Sodré Santoro (2022)

Para definir quais objetos serão desenvolvidos foi desenvolvido um fluxograma com a definição do processo a ser executado, o processo decisório a ser programado e os loops a serem repetidos.

**Figura 10 - Fluxograma do processo**

Diagrama

Descrição gerada automaticamente

Fonte: dos Autores (2022)

Para fazer uma mineração efetiva usando Requests, demanda o estudo das chamadas feitas pelo Front-End para o Back-End. Uma forma simples de identificar essas chamadas é pelo DevTools presente em navegadores como Chrome e Firefox, no desenvolvimento do projeto utilizamos o Chrome. No Chrome para utilizar o DevTools basta apertar F12, ir na aba Network e procurar pela chamada que retorna à informação desejada, essa chamada é geralmente do tipo “Document” ou “Fetch”. Neste momento estava na página inicial buscando a URL que nos levaria para página de leilões de Carros, Motos e peças.

**Figura 11 - HTTP Get**

Texto

Descrição gerada automaticamente

Fonte: dos autores (2022)

Pontos importantes nessa função é o Request que retorna o HTML (Hyper Text Markup Language) e o uso do LXML para gerar um DOM (Document Object Model) interativo no Python, exemplo: Para encontrar o elemento usamos.

Outra parte importante da captura de dados é a iteração sobre lista capturando inicialmente o total de itens e dividindo pelo número de itens em cada página, seguindo iterando pelo total de páginas.

**Figura 12 - Contagem total de páginas**

Texto

Descrição gerada automaticamente

Fonte: dos autores (2022)

O número total de itens foi extraindo usando o Regex ou Regular Expression que provê uma forma concisa e flexível de identificar cadeias de caracteres de interesse, como caracteres particulares, palavras ou padrões de caracteres. Expressões regulares são escritas numa linguagem formal que pode ser interpretada por um processador de expressão regular, um programa que serve um gerador de analisador sintático ou examina o texto e identifica as partes que se casam com a especificação dada (Goyvaerts, 2021).

**Figura 13 - O uso de Regex**

Texto

Descrição gerada automaticamente

Fonte: dos autores (2022)

O código abaixo (Figura 11) tem o principal objetivo de guardar em uma lista a URL (Uniform Resource Locator) dos links com o uso da propriedade html.links que retornaram todos os links contidos naquela página, também se necessário filtrar os links usando condicionais.

**Figura 14 - Montando lista de URLs**

Texto

Descrição gerada automaticamente

Fonte: dos autores (2022)

Também é não preferível evitar usar loops encadeados na captura de lista, e sim capturar as URL’s para os itens e em seguida iterar uma 2ª vez, o número de itens capturados não altera, porém existes uma melhora significativa na legibilidade do código.

Com as URL’s em uma lista podemos iterar sobre ela e capturar item por item, em caso de uma necessidade maior ainda por performance pode se utilizar facilmente de “multithreads”, mas aqui não se aplica.

Os dois métodos para captura eficiente de informação presentes na web são o uso de XPath e Regex.

O uso de XPath e especialmente simples quando a informação no DOM está atrelada a um id, e possível gerar um XPath copiado diretamente da DevTools, porém idealmente devemos escrever XPath mais eficientes como os ensinados em “XPath in Selenium with All Tactics and Examples [2021 Guide]”.

Para gravar essas informações obtidas durante a raspagem definimos cada propriedade a ser gravada em uma “key” do dicionário Python com chamadas respectivas para buscar a informação e tratá-la.

**Figura 15 - Definindo key e value no dicionário de informações capturadas**

Texto

Descrição gerada automaticamente

Fonte: dos autores (2022)

Essas chamadas podem ser bem simples com retorno direto do XPath ou com tratativas para retorno de lista ou identificação de padrão de datas.

**Figura 16 - Uso de XPath**

Texto

Descrição gerada automaticamente

Fonte: dos autores (2022)

Função que pode tratar as informações contidas em uma lista primeiramente contando o número de elementos e usando a iteração de index para localizar de forma sequencial os elementos, na figura abaixo as informações contidas em cada item da lista são guardadas como chave e valor em um dicionário Python.

**Figura 17 - Iterando sobre lista no HTML**

Texto

Descrição gerada automaticamente

Fonte: dos autores (2022)

Função que lida com data e a converte em classe datetime do Python, para usar o texto que foi retornada da função XPath, podemos o usar datetime.strptime passando o texto obtido e o formato da data, como lote há apenas a informação do dia, mês e horário a classe por padrão utiliza o ano 1900, basta alterar a propriedade “year” para o ano atual.

**Figura 18 - Conversão em datetime**

Texto

Descrição gerada automaticamente

Fonte: dos autores (2022)

## Análise dos Dados

Seguindo os passos para o KDD, com a base de dados consolidada na etapa de extração de dados, o próximo passo antes de realizar a análise é realizada a limpeza e pré-processamento nos dados coletados. Utilizando a biblioteca Pandas, através do Colab (Google Colaboratory), foi realizado uma análise da estrutura dos dados para verificar dados nulos. As funções da biblioteca Pandas permitem realizar a verificação da estrutura da base e se há dados nulos. A figura 7 detalha quais tipos de dados que cada coluna possui.

**Figura 19 - Sumário da base de dados**

Tabela

Descrição gerada automaticamente

Fonte: dos autores (2022)

Prosseguindo com a investigação por dados nulos, a função isnull() permite identificar quais colunas possuem os dados nulos, de forma que facilita o tratamento desses dados. A figura 8 mostra uma tabela contendo quais colunas possuem dados nulos sendo a coluna Câmbio (Moto) a única que possui dados nulos.

**Figura 20 - Verificação de dados nulos**

Tabela

Descrição gerada automaticamente

Fonte: dos autores (2022)

Ainda na etapa de limpeza de dados e pré-processamento, foi feito o procedimento de redução das colunas url, placa, date e Câmbio (Moto) pois não são relevantes para utilizar na análise. Os nomes das colunas restantes foram modificados para facilitar a legibilidade e manipulação delas, como por exemplo, a substituição dos espaços entre as palavras pelo caractere *underscore,* a eliminação das pontuações de acento agudo e circunflexo nas palavras que as contém e algumas palavras que estavam em inglês foram traduzidas para o português.

Com as devidas colunas e dados selecionados, a etapa seguinte do KDD consiste na transformação dos dados, ou seja, os dados serão formatados para se encaixar melhor nas análises e algoritmos de mineração de dados.

As principais colunas passaram por tratamento de dados, como exemplo, na coluna de lance\_atual foi retirado o cifrão ($), o R do real e a vírgula que delimita as casas decimais, porém o tipo de dado dessa coluna estava como string sendo feita a sua conversão para int.

Na coluna ano os dados são array, sendo que alguns registros apresentavam mais de um elemento, nesse caso apenas o primeiro elemento foi selecionado e convertido para o tipo int.

A coluna local foi tratada para que apenas a sigla federativa fosse definida como dado da coluna.

A coluna nome, que representa o modelo dos carros, foi tratada para que apenas a marca e o modelo do carro sejam representados e cada primeira letra do modelo seja maiúscula. Na figura 9 é demonstra como ficou os cinco primeiros registros do DataFrame após a limpeza dos dados.

**Figura 21 - Dataframe após a limpeza de dados**

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Fonte: dos autores (2022)

Após a limpeza e seleção dos dados, foi feita a estatística descritiva para investigar e obter informações relevantes a partir dos dados consolidados na etapa anterior, utilizando as variáveis de maior interesse em relação as características dos carros. Segundo Guedes (2005), a estatística descritiva tem como propósito juntar as medidas dos dados e obter uma visão macro da variação dos valores, sendo que sua representação pode ser realizada através de gráficos, tabelas e medidas descritivas.

Utilizando o software Orange, possibilitou a criação de um fluxo de componentes para a visualização e análise dos dados, de forma que através de gráficos e medidas descritivas foi possível obter informações úteis a respeito dos dados. A figura 21 ilustra o fluxo de componentes criado para a visualização dos dados.

**Figura 22 - Fluxo de componentes para a visualização dos dados**

Diagrama

Descrição gerada automaticamente

Fonte: dos autores (2022)

O fluxo começa com o componente de importação de arquivo CSV que permite a importação do arquivo único CSV que contém os dados extraídos da plataforma Sodré Santoro.

Após a exportação, o componente de editar domínio possibilita renomear, checar os tipos e corrigir alguns valores das variáveis.

Logo em seguida, o componente de selecionar colunas foi utilizado para descartar as variáveis de data e URL, as quais não são pertinentes ou interessantes para serem incluídas na visualização de dados.

Os três últimos componentes são para a visualização dos dados em forma de gráficos e medidas descritivas. O componente Scatter Plot é um gráfico de dispersão o qual ilustra duas ou mais variáveis e se existe alguma relação de causa e efeito nelas. O componente de Distribuição permite a visualização das frequências das variáveis através de gráficos de barras. Por último, o componente de recursos estatísticos dispõe de medidas descritivas das variáveis.

Após a análise da base de dados foi possível identificar quais são as características desfavoráveis e assim realizar uma query no próprio Pandas para selecionar os lotes apenas com essas características desfavoráveis. Para isso é necessário a importação da biblioteca Pandasql, a qual permite utilizar sintaxe SQL em DataFrames do Pandas. Para a realização da query as seguintes variáveis foram utilizadas: Estado do chassi, Ar-Condicionado, Direção Hidráulica/Elétrica, Quilometragem e Origem.

O resultado da query é armazenado em um DataFrame e exportado para um arquivo CSV para ser analisado e visualizado no software Orange.

No programa Orange, o componente de tabela de dados facilita a visualização dos dados em uma tabela. A figura 23 ilustra o fluxo criado para a visualização da base de dados com apenas os lotes de características desfavoráveis.

**Figura 23 - Componente de Tabela de Dados**

Diagrama

Descrição gerada automaticamente

Fonte: dos autores (2022)

# RESULTADOS

## Coleta dos dados

As técnicas de extração de dados abordadas nos capítulos anteriores foram utilizadas para a extração de dados no site sodresantoro.com.br, o qual é direcionado para leilões de automóveis usados e aplicado o filtro para apenas carros leves. A extração possibilitou a criação de uma base de dados em um arquivo de extensão .CSV que pode ser manipulado e realizar análise dos dados.

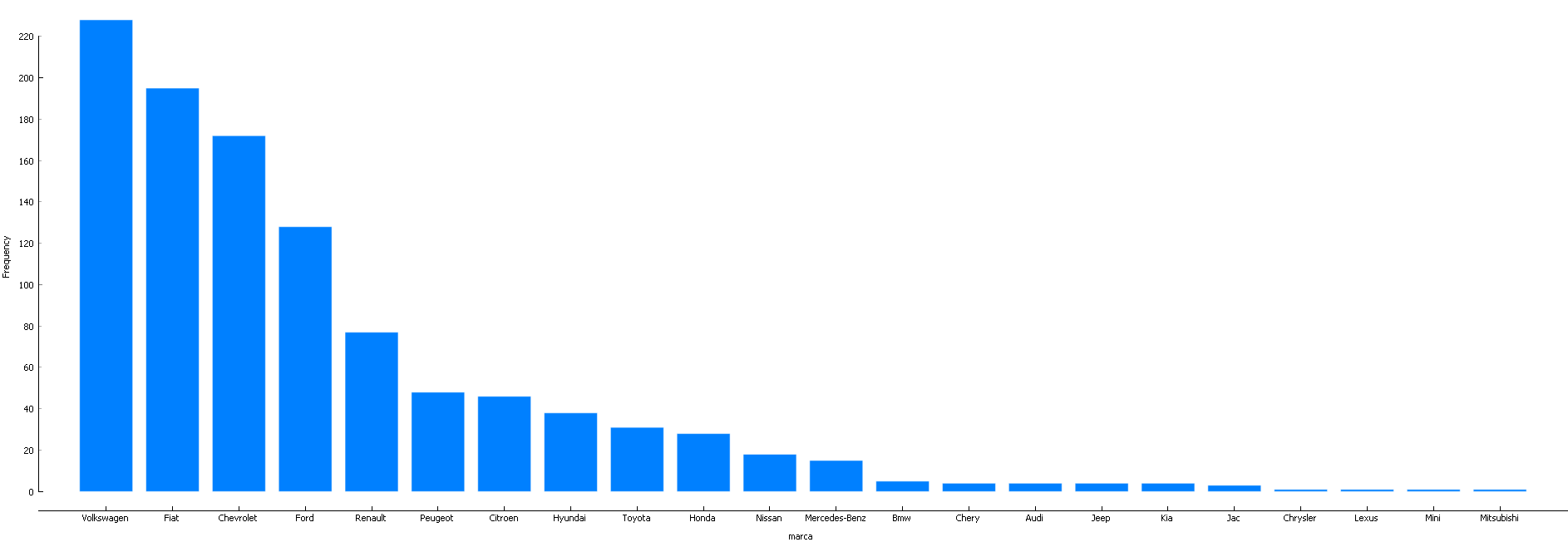
O Script desenvolvido acessa a URL de veículos usados e extrai os dados previamente selecionados para consolidar a base dedados e posteriormente a análise. A performance é consideravelmente satisfatória, visto que a execução leva em torno de 6.5 segundos, vale ressaltar que a performance está associada a quantidade de itens capturados, podendo ter o tempo de execução maior ou menor. Além disso, a captura dos dados foi realizada diariamente entre o período de 17 de abril a 25 de outubro de 2022.

Cada extração gera um arquivo que, posteriormente, seus dados são adicionados a um único arquivo geral, ou seja, esse único arquivo geral possui os dados das demais extrações realizadas. Este arquivo ao final do período acumulou 8340 registros de carros anunciados no leilão, eliminando as duplicatas restaram 1068 registros.

## Análise Exploratória dos Dados

Após a coleta dos dados, e passado as etapas abordadas no KDD, utilizando o software Orange para visualização dos dados, foram gerados gráficos pertinentes a análise.

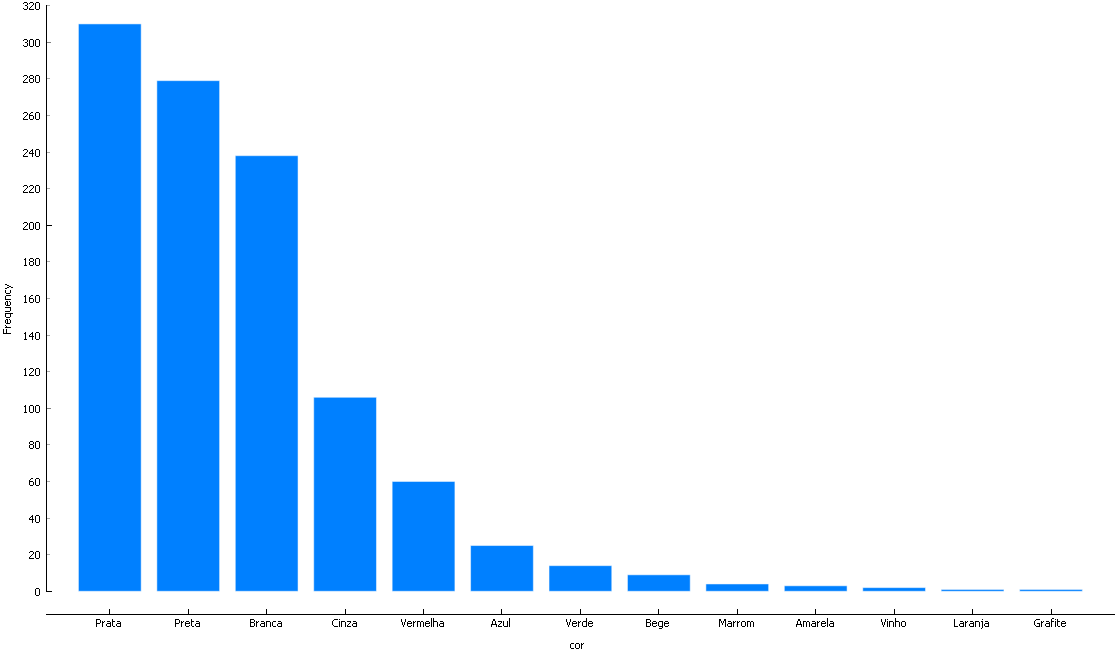
**Gráfico 1 - Distribuição de Marcas**



Fonte: dos autores (2022)

A distribuição de marcas representa a frequência de carros por marca. É notório a prevalência de quatro marcas em relação as demais, sendo elas: Volkswagen, Fiat, Chevrolet e Ford. Essas marcas foram as mais populares dentro da amostra de dados. De acordo com o relatório mensal da FENABRAVE (2022) de setembro de 2022 de índices e números de carros seminovos e usados, os modelos mais populares foram referentes as quatro marcas citadas acima.

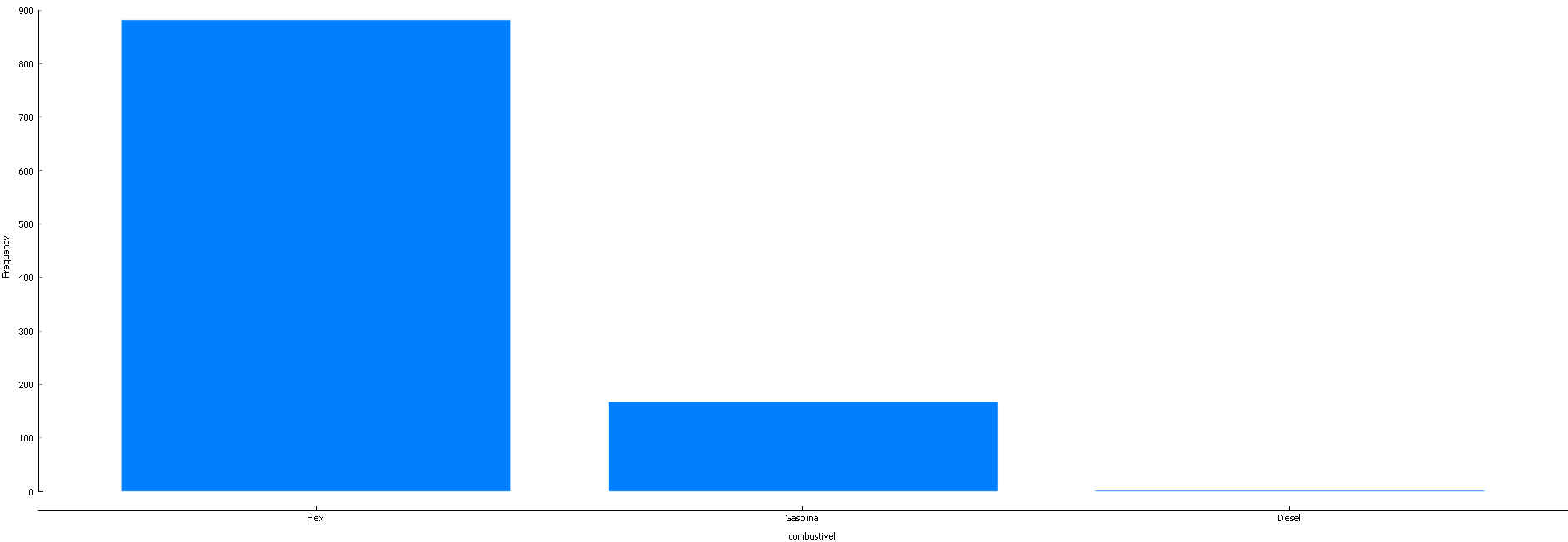
**Gráfico 2 - Distribuição de Cor**



Fonte: dos autores (2022)

O gráfico 2 apresenta a preferência pelas cores prata, preta, branca e cinza, ou seja, há uma certa tendência e maior volume de carros dessas cores. As cores menos comuns que se destacam são: bege, marrom, amarela, vinho, laranja e grafite.

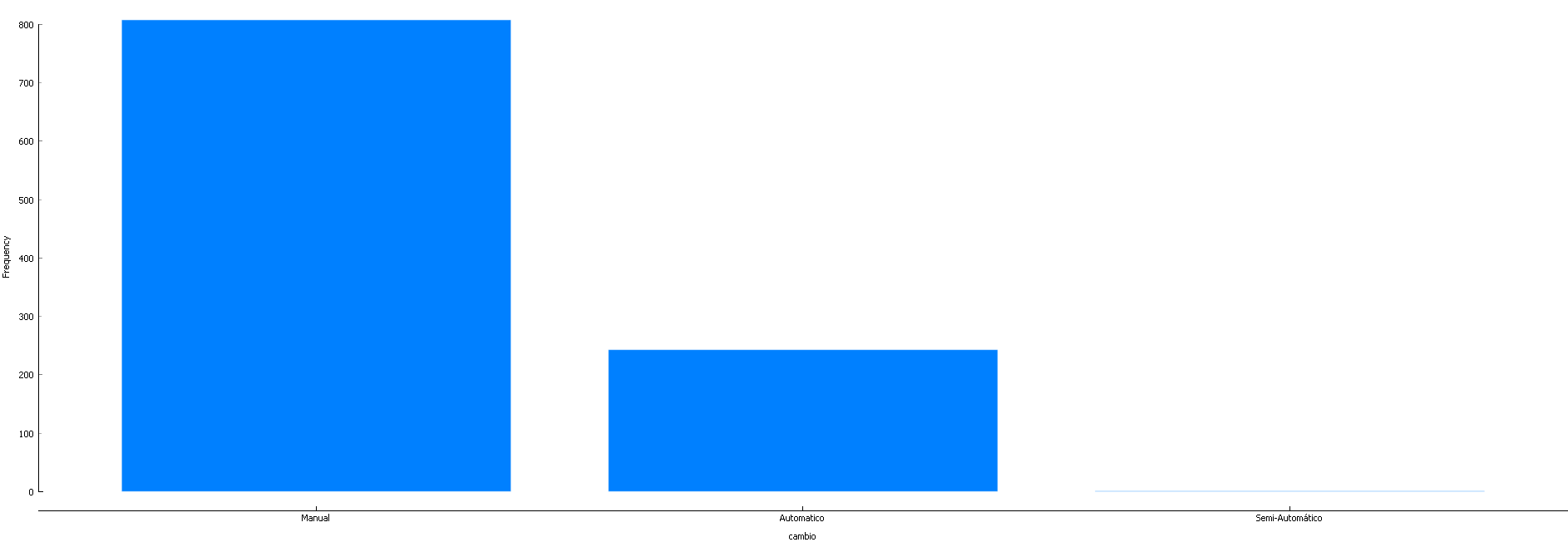
**Gráfico 3 - Distribuição de Tipo de Combustível**



Fonte: dos autores (2022)

A maioria dos veículos possuem como tipo de combustível Flex, tipo muito popular no país pela versatilidade de suportar tanto etanol quanto gasolina. O carro que utiliza apenas gasolina apresenta uma leve característica desfavorável pelo fato de não suportar o uso de etanol.

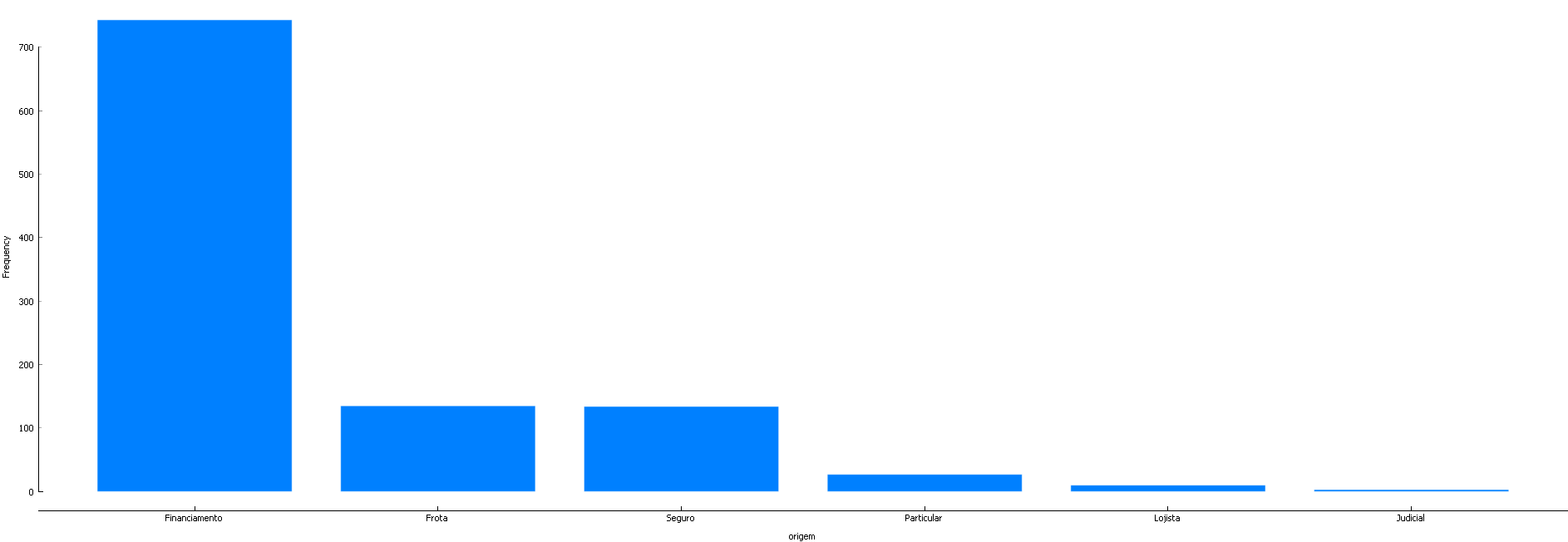
**Gráfico 4 - Distribuição de Tipo de Câmbio**



Fonte: dos autores (2022)

Na base de dados há certa tendência e preferência por carros que utilizam o câmbio manual, embora pesquisas apontem que o interesse por carros automáticos aumentou em mais de 50% no primeiro semestre de 2021 em relação ao de 2020 (CARVALHO, 2021).

**Gráfico 5 - Distribuição de Tipo de Origem**



Fonte: dos autores (2022)

A maior parte dos veículos são de origem de Financiamento. As origens de Frota e Seguro estão quase que equiparadas. Segundo Miragaya (2021), em seu artigo no site Quatro Rodas, os tipos de origem são relevantes, sendo o tipo de origem Judicial uma característica altamente desfavorável, pois necessita da contratação de advogado para resolver os trâmites judiciais pós-leilão.

**Gráfico 6 - Distribuição por Estado do Chassi**

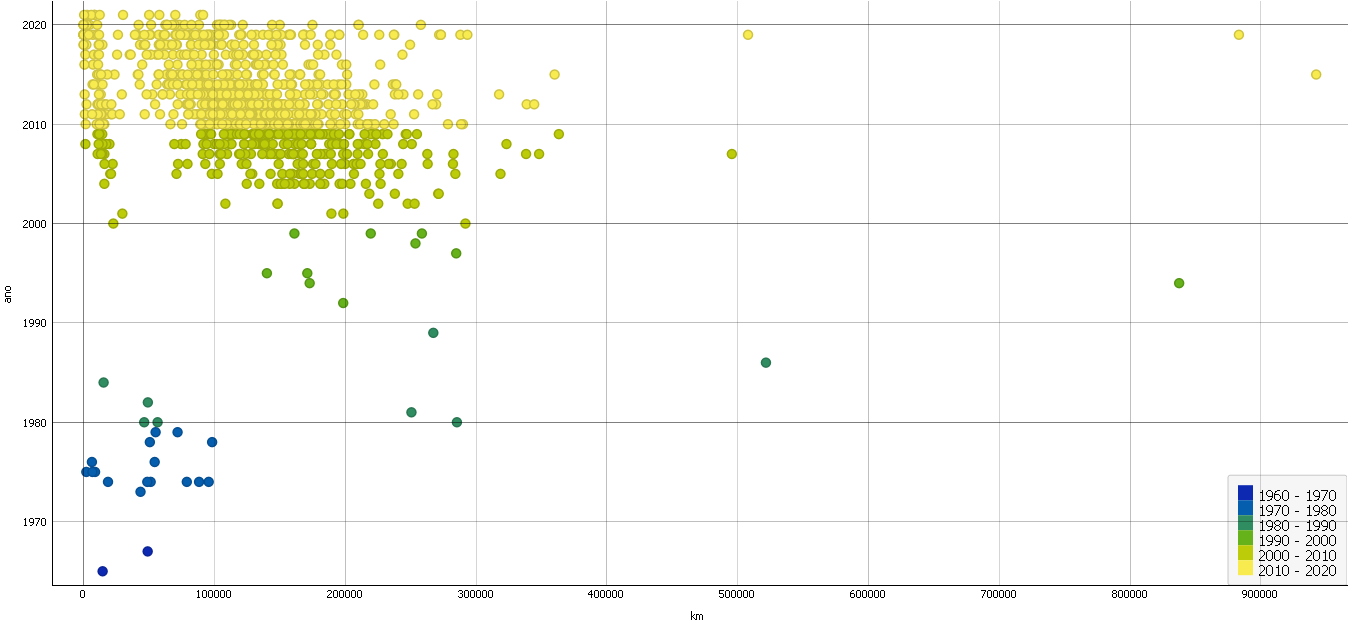
Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente com confiança média

Fonte: dos autores (2022)

A maior faixa de volume dos lotes apresenta o estado do chassi como integro, uma característica altamente favorável, em contrapartida, os outros valores como: remarcado, enferrujado, ilegível, sem acesso, danificado e amassado, se enquadram como características altamente desfavoráveis, pois a avaria no chassi do carro diminui em até 30% o preço de valor de mercado do carro, problemas para financiamento e dificuldades ao realizar o seguro do carro com seguradoras (DESPACHANTE, 2021).

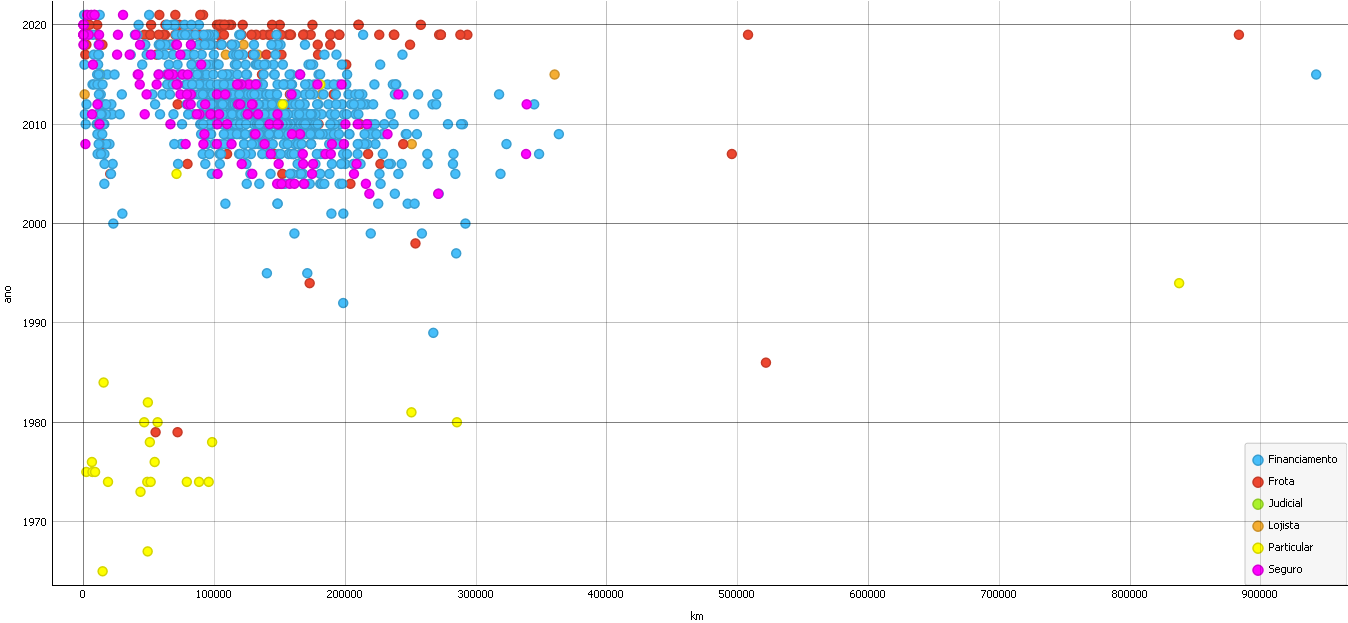
**Gráfico 7 - Dispersão de Km X Ano**



Fonte: dos autores (2022)

É notório uma alta concentração de carros que possuem o ano de fabricação entre 2010 e 2020, além disso, a maioria de veículos nesse intervalo de ano têm a quilometragem entre 100.000 e 200.000 quilômetros. É interessante notar que os veículos na faixa dos anos 1960 a 1970 estão abaixo de 100.000 quilômetros rodados.

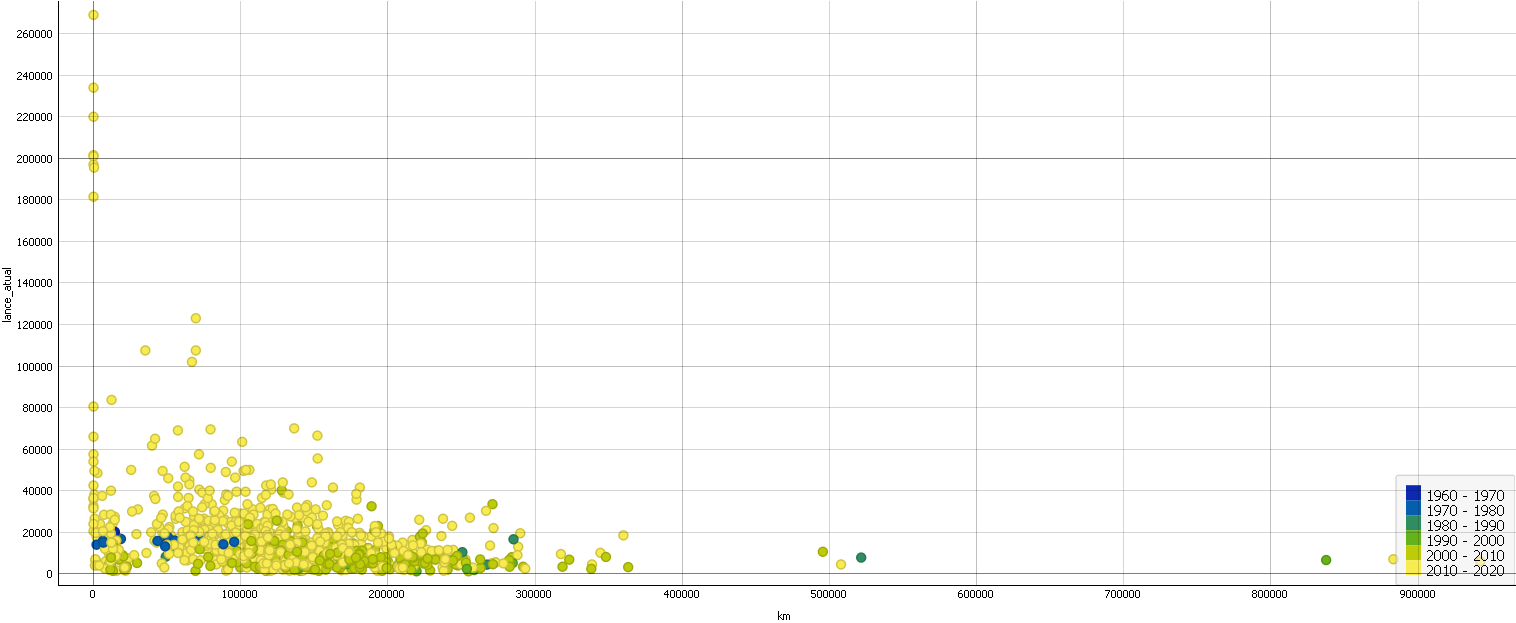
**Gráfico 8 - Dispersão de Km X Ano (origem)**



Fonte: dos autores (2022)

O gráfico 7 apresenta certas similaridades ao gráfico 6, entretanto, é possível perceber que os veículos abaixo dos anos de 1980 são de origem particular, isto é, são carros que possivelmente têm maiores cuidados e manutenção constante por serem antigos. Miragaya (2021), afirma que carros abaixo dos anos de 1980 são considerados clássicos e com certo potencial de valor histórico.

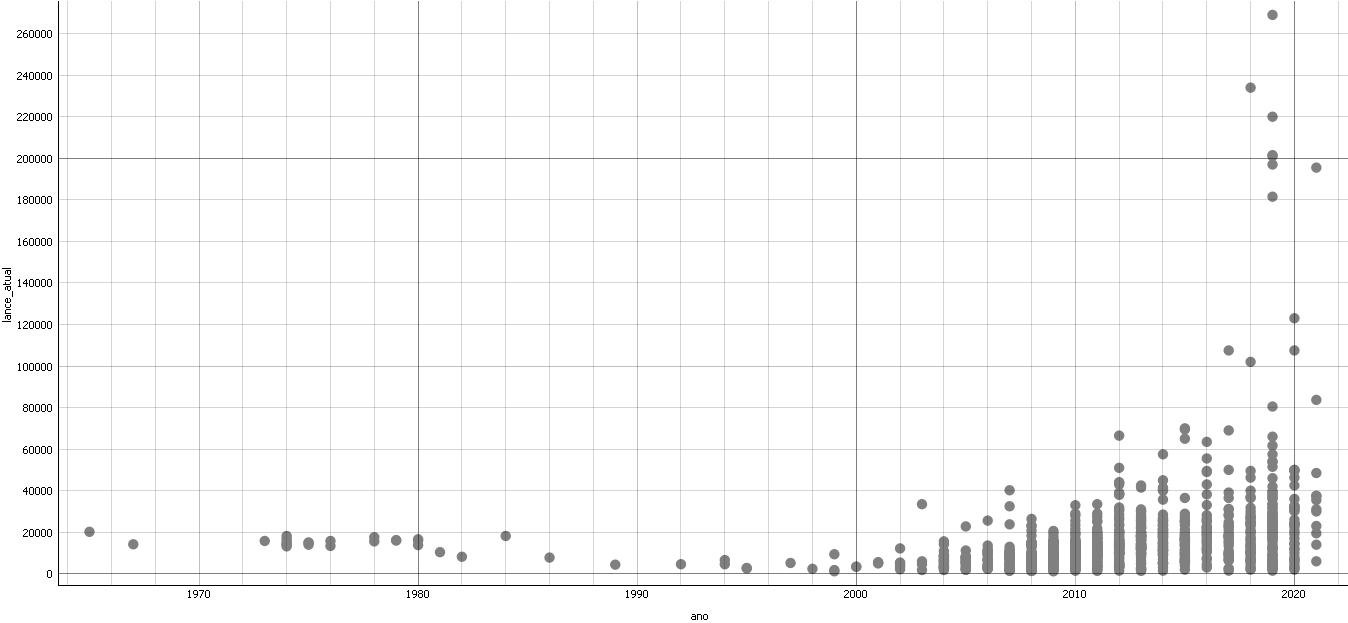
**Gráfico 9 - Dispersão de Km X Lance Atual**



Fonte: dos autores (2022)

Carros seminovos e usados possuem lances mais altos, ou seja, a alta quilometragem é uma das caraterísticas desfavorável. Carros acima dos 300.000 quilômetros rodados não possuem valores de Lance Atual maiores que R$20.000.

**Gráfico 10 - Dispersão de Ano X Lance Atual**



Fonte: dos autores (2022)

Existe maior concentração de carros na faixa dos anos 2010 a 2020, sendo nessa faixa com a maior quantidade de Lance Atual de valores mais altos. Carros mais recentes de 2020 são os que possuem maiores valores de lance atual.

**Figura 24 - Lotes com as características desfavoráveis**

Uma imagem contendo Tabela

Descrição gerada automaticamente

Fonte: dos autores (2022)

O total de carros identificados com as características desfavoráveis foi de 92 carros, sendo que a figura 24 ilustra uma pequena amostra do conjunto. É possível visualizar muitos lotes com valores danificado para o ar-condicionado e outros com o estado do chassi com valores diferentes de íntegro, valores esses que refletem negativamente no preço de mercado do carro.

# DISCUSSÃO DE RESULTADOS

Em 7 meses foram extraídos 8352 dados de carros do site de leilão Sodré Santoro. Foi necessário realizar extrações diárias para acompanhar os novos leilões e obter o valor dos lances mais recentes.

Os gráficos demonstram claramente características mais populares dos lotes analisados como marcas e cores e, também evidenciam algumas características que são consideradas alertas em um lote, as quais são desfavoráveis para optar pelo arremate do lote, como por exemplo, o tipo de origem ser judicial, lotes que apresentam o estado do chassi como danificado, ilegível, remarcado, amassado, enferrujado ou sem acesso, alta quilometragem, ultrapassando dos 300.000 quilômetros rodados e entre outras características.

Algumas características influenciam muito negativamente no valor de mercado do carro, esses são pontos altamente relevantes para serem levados em consideração, pois além de atuar negativamente no preço do carro, carros com essas características terão maiores dificuldades para contratarem algum tipo de seguro ou até mesmo financiamento.

É notório que existem parcelas maiores de marcas de carros como, Volkswagen, Fiat, Chevrolet e Ford. A preferência por essas mesmas marcas também apareceu em diversos relatórios mensais da Fenabrave (Federação Nacional da Distribuição de Veículos Automotores) na transação de veículos seminovos e usados no ano de 2021 e 2022. Além disso, há uma oferta de carros que se encontram na faixa de ano de 2010 a 2020.

O estado do chassi é uma característica muito relevante, pois caso o estado de conservação dessa parte do carro esteja em péssimas condições, avariada ou com a impossibilidade da leitura de informações pode diminuir em até cerca de 30% o valor de mercado do carro e problemas na contração de seguro ou financiamento. Além disso, carros que rodaram boa parte no litoral, frequentemente sofrem com a exposição da maresia, causando enferrujamento com maior facilidade nessa e outras partes do carro, necessitando de cuidados e manutenções preventivas mais constantes.

No tipo de origem vale ressaltar que para lotes que são de origem judicial são mais burocráticos, sendo altamente recomendável a contratação de um advogado para resolver trâmites judiciais pós-leilão. Os lotes de origem particulares são vistos com bons olhos, pois boa parte dos lotes são de carros abaixo do ano de 1990, ou seja, podem ser carros clássicos e que passam por manutenção constantes (MIRAGAYA, 2021). Na análise foi possível verificar que a maior parte dos lotes de origem particular são carros abaixo do ano de 1990.

A análise aponta que ainda há maior oferta por carros que tenham o tipo de câmbio manual, embora Carvalho (2021), em seu artigo no site Quatro Rodas, aponte estudos realizados pela OLX a qual constatou um aumento de mais de 50% nas vendas de carros automáticos em sua plataforma no primeiro semestre de 2021 em relação ao de 2020.

A quilometragem do carro compõe uma das características a ser observada, a alta quilometragem implica que o motor e outras peças sofreram foram bem desgastadas. De acordo com Sousa (2020), em seu artigo no site Auto Esporte, carros que ultrapassaram os 100.000 quilômetros rodados podem sofrer com problemas mecânicos. Vale destacar que o tipo de cenário que o carro mais rodou é relevante, carros que rodaram mais em estradas sofrem um desgaste menor no motor em relação aos que rodam na cidade, pelo fato de na estrada o motor conseguir atingir condições que sofre menos desgaste.

As características do carro são altamente relevantes, porém algumas outras informações podem fazer muita diferença, se o carro rodou mais em regiões litorâneas ou se rodou mais em estradas. Essas informações agregam à análise e, posteriormente, no processo decisório.

# Considerações Finais

A linguagem de programação Python foi fundamental, pois a sua utilização foi feita para a extração e limpeza dos dados e a importação da biblioteca Pandas. Além disso, a plataforma Google Colaboratory proporcionou o tratamento dos dados em máquina virtual.

O objetivo do presente trabalho foi desenvolver uma ferramenta de extração e análise de dados de carros usados de leilão da plataforma Sodré Santoro, transcrever em um relatório, que consiga indicar lotes dos carros com características desfavoráveis e alertar o leitor sobre esses índices.

A primeira etapa foi conceituar os principais assuntos que compõe a descoberta de conhecimento em base dados, abordando sobre dados, informação, extração e manipulação de dados, tipos de fontes de dados e mineração de dados. Algumas das ferramentas que podem ser utilizadas em algumas das etapas da descoberta de conhecimento foram descritas.

A extração usando as bibliotecas Request\_HTML da linguagem de programação Python para o desenvolvimento do *bot* foi desafiadora, já que, é necessário identificar e reproduzir as requisições feita pelo browser, porém assim que o desenvolvimento foi concluído, a extração de dados ocorre de maneira extremamente performática e consistente.

A última etapa do trabalho foi realizar a análise e visualização dos dados. A plataforma do Google Colaboratory foi utilizada para importar a biblioteca Pandas, para realizar a etapa de limpeza dos dados e exportar a base, já limpa, para ser analisada no software Orange. A limpeza de dados apresentou certos obstáculos, pois os dados estavam com tipos de dados diferentes do que foram utilizados na análise, algumas inconsistências e valores nulos. A biblioteca Pandas demonstrou ser versátil e bem documentada em trabalhar com arquivos de extensão CSV e explorar estruturalmente a base de dados.

O software Orange mostrou-se extremamente intuitivo, de fácil usabilidade e entendimento. A possibilidade de criar fluxo de componentes facilita o processo de manipulação e visualização dos dados. Os componentes apresentam opções para configurá-los e, de forma interativa, obter os resultados desejados, além disso, é possível exportar apenas os resultados desejados para uma página HTML. A detentora do software mantém uma documentação bem vasta e esclarecedora de cada componente.

Os resultados da análise e visualização de dados mostraram que lotes com estado do chassi com os valores danificado, amassado, ilegível ou enferrujado; tipo de origem judicial; quilometragem acima dos 300.000; ar-condicionado danificado; direção hidráulica/elétrica danificada, são altamente desfavoráveis em relação aos demais valores dessas mesmas características.

Vale ressaltar que algumas informações não disponibilizadas no site são altamente relevantes como, se o carro rodou mais em cidade ou região litorânea e se a maior parcela da quilometragem rodada foi na cidade ou em estrada.

Os resultados da análise alertam o leitor a respeito dos lotes com as características desfavoráveis mencionadas, de forma que é altamente recomendável evitar os lotes que possuam essas características negativas, pois esses carros podem apresentar complicações em questões de manutenção, revenda ou contratação de seguradoras.

## Trabalhos Futuros

Como expectativa para trabalhos futuros, pode ser realizada a etapa de mineração de dados utilizando algum algoritmo de máquina supervisionado para classificar de forma automatizada os lotes com base nas características desfavoráveis. É também viável utilizar algoritmos de máquina não supervisionado de agrupamento e analisar semelhanças ou padrões entre os grupos estabelecidos.

Uma outra possibilidade é programar o *crawler* para capturar dados de carros de outros sites de leilões e submeter esses dados ao processo de KDD, observando quais são as diferenças e semelhanças de resultados em relação a este presente trabalho.

# REFERÊNCIAS BIBLIOGRÁFICAS

AGGARWAL, C. C. **Data Mining**: The Textbook. Nova York: Springer, 2015.

AGRAWAL, Divyakant et al. Challenges and opportunities with Big Data 2011-1. 2011.

AYSWARRYA, G. Data Sources 101. KDnuggets. 2022. Disponível em: https://www.kdnuggets.com/2019/10/data-sources-101.html. Acesso em: 03 de nov. de 2022.

BERRY, Michael JA; LINOFF, Gordon S. **Data mining techniques: for marketing, sales, and customer relationship management**. John Wiley & Sons, 2004.

BECKER, João Luiz. **Estatística básica: transformando dados em informação**. Bookman editora, 2015.

BORGES, Thiago da Cunha; GANIMI, Zeus Olenchuk. Extração de dados com web scraping para análise da variação de preço de veículos automotores. 2018.

Bugnion, P; Manivannan, A; Nicolas, P. R. **Scala: Guide for Data Science Professionals**. Birmingham: Packt Publishing, 2017.

CAMILO, Cássio O., SILVA, João C. da – **Mineração de Dados: conceitos, tarefas, métodos e ferramentas**, Relatório Técnico, 2009.

CAMPOS, E. S. et al. Heurística do Julgamento: Uma Análise dos Lances em Leilões Sob a Perspectiva do Valor Justo e da Maldição do Vencedor, **Veredas Favip-Revista Eletrônica de Ciências**, volume 9, n. 2, p. 118 – 142, mar. 2016.

CARVALHO, Isadora. **Interesse por carros automáticos está quase 70% maior no Brasil**. Quatro Rodas. 12 out 2021. Disponível em: https://quatrorodas.abril.com.br/auto-servico/interesse-por-carros-automaticos-esta-quase-70-maior-no-brasil/. Acesso em: 21 de nov. de 2022.

CHASSI Remarcado: você sabe porque alguns veículos possuem?. **Despachante**, 2021. Disponível em: https://despachante.com/blog/carro/chassi-remarcado-voce-sabe-porque-alguns-veiculos-possuem/. Acesso em 09 de nov. de 2022.

Conamay, Drew (2010). **The data science venn diagram**. Drew Conway, (2010). Disponível em: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram. Acesso em: 21 de mar. De 2022.

DANTAS, Eric Rommel G. et al. O uso da descoberta de conhecimento em base de dados para apoiar a tomada de decisões. **V Simpósio de Excelência em Gestão e Tecnologia-SEGeT**, v. 1, p. 50-60, 2008.

DE LUCENA, T. A. M. **Um Crawler para Extração e a Análise de Dados do Fórum do Warframe**. Monografia (Bacharel em Ciência da Computação) - Faculdade de Ciências Exatas e Naturais, Universidade do Estado do Rio Grande do Norte. Rio Grande do Norte, p. 37. 2018.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37-37, 1996.

FENABRAVE. Informativo Usados Fenabrave Setembro 2022. São Paulo. 2022. 8 pg. Disponível em: https://online.fliphtml5.com/ordey/oxaj/#p=8. Acesso em: 03 nov. de 2022.

FLEISHER, R. M. M. C. Le coadic, yves françois. a ciência da informação. tradução de maria yêda f. s. de filgueiras gomes. brasília: briquet de lemos, 1996. **A ciência da Informação**, v. 1, n. 2, 1996.

FRANK, E.; HALL, M. A.; WITTEN, I. H. **The WEKA Workbench**. Nova Zelândia: Morgan Kaufmann, 2016. Disponível em: https://www.cs.waikato.ac.nz/ml/weka/Witten\_et\_al\_2016\_appendix.pdf. Acesso em: 05 abr. 2022.

GOMES, Tânia Margarida dos Santos. **Ferramentas open source de Data Mining**. 2014. Tese de Doutorado.

GOOGLE RESEARCH. **Google Colaboratory**. Site: Documentation, 2022. Página de documentação. Disponível em: https://research.google.com/colaboratory/intl/pt-BR/faq.html. Acesso em: 12 de abr. de 2022.

GOYVAERTS, Jan. **Regular Expressions Tutorial Learn How to Use and Get The Most out of Regular Expressions**; Disponível em: https://www.regular-expressions.info/tutorial.html. Acessado em: 10 de maio de 2022.

GUEDES, Terezinha Aparecida et al. Estatística descritiva. **Projeto de ensino aprender fazendo estatística**, p. 1-49, 2005.

HERBERT, Anthony. Como começar a usar a biblioteca Requests em Python. **Digital Ocean**, 2020. Disponível em: https://www.digitalocean.com/community/tutorials/how-to-get-started-with-therequests-library-in-python-pt. Acesso em: 14 de jan. de 2022.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data mining**: concepts and techniques. Elsevier, 2011.

MCKINNEY, Wes et al. pandas: a foundational Python library for data analysis and statistics. **Python for high performance and scientific computing**, v. 14, n. 9, p. 1-9, 2011.

MIKUT, Ralf; REISCHL, Markus. Data mining tools. **Wiley interdisciplinary reviews: data mining and knowledge discovery**, v. 1, n. 5, p. 431-443, 2011.

MIRAGAYA, Fernando. **Tudo que você precisa saber para antes de comprar carro em leilão**, Quatro Rodas, 8 jul 2021. Disponível em: <https://quatrorodas.abril.com.br/auto-servico/leiloes-de-carros-seguradoras-monta-banco-divida-classicos/>. Acesso em: 31 de out. de 2022.

MONTEIRO, Alexandre Almeida. **Maldição do vencedor” (winner´s curse)? Uma análise das ofertas nos leilões da agência nacional do petróleo, gás natural e biocombustíveis (ANP)**. Monografia (Bacharelado em Economia) – Instituto de Economia, Universidade Federal do Rio De Janeiro, Rio de Janeiro, p.55. 2015.

PEREIRA, G.H.V., **Mineração de Dados de Páginas de Notícias Sobre Delitos e Infrações de Trânsito Usando Python e Web Scraping**, Dourados: UEMS, 2017.

PEREIRA, Luiz Gustavo Simão. Por dentro dos pátios, de frente para a tela: Apropriações e indicativos da presença de smartphones em um leilão na região metropolitana de São Paulo. **Ponto Urbe. Revista do núcleo de antropologia urbana da USP**, n. 27, 2020.

POPOMARONIS, T. **Here's an Incredible Breakdown of What Happens on the Internet (Every 60 Seconds)**. Inc., 4 maio 2017. Innovate. Disponível em: https://www.inc.com/tom-popomaronis/every-60-seconds-an-incredible-breakdown-of-what-happens-on-the-internet.html. Acesso em: 18 de mar. de 2022.

PROVOST, Foster; FAWCETT, Tom. Data science and its relationship to big data and data-driven decision making. **Big data**, v. 1, n. 1, p. 51-59, 2013.

PYTHON SOFTWARE FOUNDATION. Python Language Site: Documentation, 2022. Página de documentação. Disponível em: https://docs.python.org/3/license.html. Acesso em: 11 de abr. de 2022.

Rautenberg, Sandro; Carmo, Paulo Ricardo Viviurka do. (2019). Big Data e Ciência de Dados: complementariedade conceitual no processo de tomada de decisão. **Brazilian Journal of Information Studies: Research Trends**. v. 13, n. 1, p. 56-67, 2019.

SAFRIS, Servis. A Deep Look at JSON vs. XML, Part 1: The history of Each Standar. 2019. **Toptal**. Disponível em: <https://www.toptal.com/web/json-vs-xml-part-1>. Acesso em: 01 de out. de 2022.

SETZER, Valdemar W. Dado, informação, conhecimento e competência. **DataGramaZero Revista de Ciência da Informação**, n. 0, v. 28, 1999.

SCIKIT-LEARN. Scikit-learn, 2022. About us. Disponível em: <https://scikit-learn.org/stable/about.html>. Acesso em: 16 de ago. de 2022.

SIRISURIYA, De S. et al. A comparative study on web scraping. 2015.

SOARES, David Jose et al. Empresas orientadas a dados e análises: a tecnologia está a serviço da tomada de decisão?. 2017.

SOUSA, Renan. Comprar carro com alta quilometragem significa mau negócio?. **Auto Esporte**, 2020. Disponível em: https://autoesporte.globo.com/servicos/noticia/2020/10/comprar-carro-com-alta-quilometragem-significa-mau-negocio.ghtml. Acesso em: 21 de nov. de 2022.

SOUZA, Renan. Por que carros de leilão são mais baratos?. **AutoEsporte**, 2020. Disponível em: https://autoesporte.globo.com/carros/noticia/2020/10/por-que-carros-de-leilao-sao-mais-baratos.ghtml. Acesso em: 24 de nov. de 2021.

Structured vs Unstructured Data: Compared and Explained. **Altexsoft**, 2020. Disponível em: <https://www.altexsoft.com/blog/structured-unstructured-data/>. Acesso em: 03 de nov. de 2022.

WOLFF, Rachel. 10 Best Data Mining Tools in 2022. **Monkey Learn**. 22 dez. 2020. Disponível em: <https://monkeylearn.com/blog/data-mining-tools/>. Acesso em: 04 de ago. de 2022.