

RESUMO

Ante um cenário mercadológico cada vez mais direcionado a automação de processos rotineiros, tem-se na Engenharia de Avaliações o desafio claro quanto a obtenção de dados de mercado, de forma qualificada, por meio de técnicas automatizadas. Essa evolução, nutrida pelo crescente aumento de competitividade, bem como da necessidade de diminuição de tempo de entrega dos trabalhos, dentre outros fatores, traz intrinsecamente, oportunidades e desafios, que devem ser estudados e testados por meio de técnicas aplicáveis, respeitando sempre as diretrizes normativas. Neste cenário, diante das várias possibilidades, destaca-se uma alternativa interessante: a obtenção de dados do mercado imobiliário através do processo de raspagem de dados, em especial daqueles provenientes de plataformas eletrônicas de anúncios. Destaca-se positivamente tratar-se de dados com ampla disponibilidade e de abrangência local, regional e nacional, bem como da possibilidade de apuração via ambiente remoto. Tem-se em contraponto a baixa qualidade da declaração dos dados pelos responsáveis e da falta de padronização entre as plataformas. Perante este desafio, desenvolveu-se um estudo de caso objetivando a verificação de um processo de raspagem de dados direcionados a um estudo de avaliação, visando analisar a captação, modelagem e o tratamento dos dados obtidos.

Palavras-chave:

Raspagem de Dados; Avaliações de Imóveis; Mercado imobiliário; Engenharia de Avaliações.

ABSTRACT

Faced with a market scenario increasingly directed towards the automation of routine processes, there is a clear challenge in Valuation Engineering in terms of obtaining market data, in a qualified manner, through automated techniques. This evolution, nourished by the increasing competitiveness, as well as the need to reduce the delivery time of the works, among other factors, intrinsically brings opportunities and challenges, which must be studied and tested through applicable techniques, always respecting the guidelines normative. In this scenario, given the various possibilities, an interesting alternative stands out: obtaining real estate market data through the data scraping process, especially those from electronic advertising platforms. It stands out positively that it is data with wide availability and of local, regional and national scope, as well as the possibility of verification via remote environment. In contrast, the low quality of data declaration by those responsible and the lack of standardization between platforms. Faced with this challenge, a case study was developed with the aim of verifying a data scraping process directed to an evaluation study, aiming to analyze the collection, modeling and treatment of the data obtained.

Keywords:

Web Scraping; Real State Appraisals; Real State Market; Evaluation Engineering.

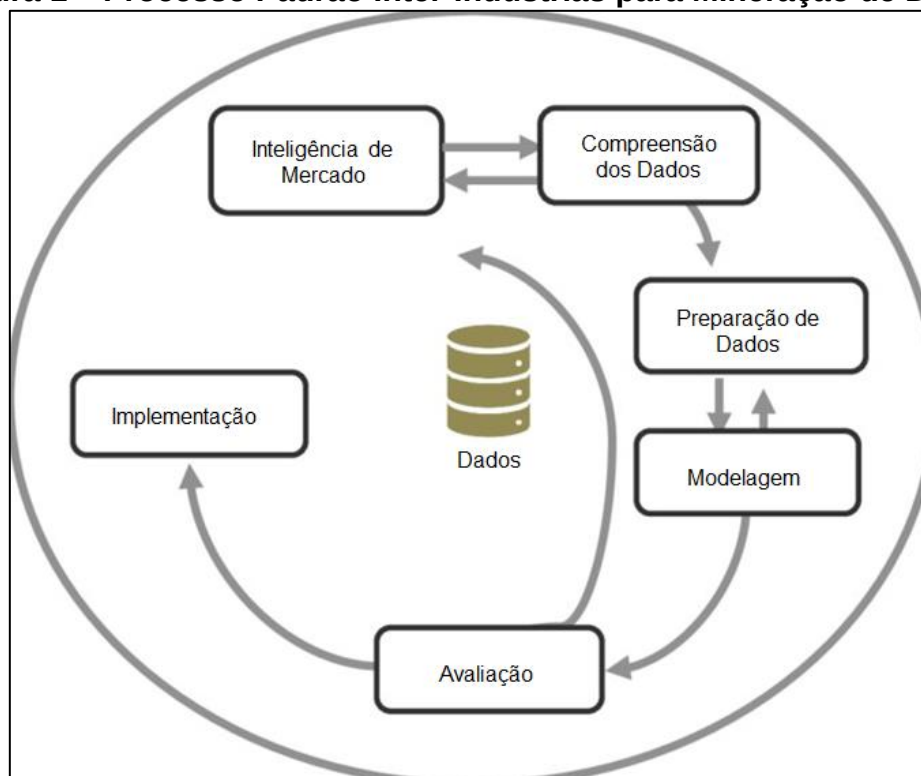
REALIZAÇÃO

1. INTRODUÇÃO

A raspagem de dados, também conhecida como Web Scraping, é o processo de extração automatizada de informações de websites. A técnica envolve o uso de um software ou script para percorrer páginas da web, identificar e extrair dados relevantes, como texto, imagens, links ou qualquer outro tipo de conteúdo estruturado. Ainda, permite coletar grandes quantidades de dados de maneira eficiente, eliminando a necessidade de extração manual.

Na área de avaliação imobiliária, a raspagem de dados pode ser utilizada para coletar informações sobre propriedades, como endereço, características físicas tais como área construída, área total, preço de venda ou aluguel, histórico de transações, entre outros. Ao extrair esses dados de várias fontes, como sites de imobiliárias, buscadores e agrupadores de dados imobiliários ou agências governamentais, os profissionais de avaliação imobiliária podem obter uma visão abrangente do mercado, bem como obter informações qualificadas a serem aplicadas na inteligência de mercado de uma empresa.

Figura 1 – Processo Padrão Inter-Indústrias para Mineração de Dados



Fonte: O Processo da Metodologia CRISP-DM (CHAPMAN, Pete et al, 2019 - Adaptado)

Importante abordar que ao fazer uso das técnicas de raspagem de dados na avaliação imobiliária, é importante estar ciente das questões legais e éticas envolvidas. É necessário respeitar os termos de serviço dos sites e as políticas de privacidade, bem como garantir que a coleta de dados esteja em conformidade com as regulamentações aplicáveis, como leis de proteção de dados e direitos autorais.

1.1. Objetivos

Este trabalho tem por fim fazer a análise dos dados obtidos experimentalmente através de aplicação da raspagem de dados, para avaliação de riscos e oportunidades do uso da técnica. Para tanto, faz-se o uso de algoritmo pré desenvolvido de plataforma de Web Scraping, em 02 portais imobiliários, de 02 tipologias macro diferentes, numa região pré determinada. Assim, serão analisadas características declaradas textualmente independente e em blocos. Ressalta-se que não serão analisadas e conferidas características por meio de análise visual das imagens disponibilizadas dos imóveis.

2. METODOLOGIA

A metodologia desenvolvida é descrita a seguir:

Figura 2 – Desenvolvimento da Raspagem e Análise de Dados



Fonte: Elaborado pelo Autor

2.1. Aspectos Normativos

A coleta de dados realizada neste trabalho procedeu de acordo com as definições dadas pela “NBR 14.653 – Avaliação de bens – Parte 1 – Procedimentos gerais”, em especial ao Item “6.4 – Coleta de Dados”, bem como de seus subitens 6.4.1 a 6.4.3. Ainda, foram observados os Procedimentos de Excelência normatizados no Item 5 da mesma Normativa, em especial quanto ao subitem 5.3 – Propriedade Intelectual e suas repercussões quanto a Lei Geral de Proteção de Dados – LGPD.

2.2. Plataforma de Raspagem de Dados

Para a realização do processo de Web Scraping (Raspagem de Dados) foi escolhida a plataforma “Octoparse”, que pode ser acessada no site oficial da plataforma (<https://www.octoparse.com/#>). Para tanto, foi adotada a versão “v8.6.2.050811”, na modalidade “Free Plan”.

Trata-se de uma plataforma com solução pré definida sem codificação para Web Scraping para transformar páginas em dados estruturados com cliques. Conta tanto com algoritmos pré definidos de extração de dados, bem como da possibilidade de personalização para cada site, opção adotada para este trabalho.

2.3. Plataforma de Anúncio de Imóveis e Ambientes de Coleta de Dados

Para a concepção deste trabalho, foram analisadas e selecionadas 02 (Duas) plataformas com boa disponibilidade de anúncios, bem como boa qualidade na estruturação e divulgação dos dados de cada imóvel.

Para tanto, foram escolhidas as plataformas Viva Real e Netimóveis, de forma a possibilitar a realização de comparativo de dados obtidos de ambas as plataformas, e que se diferenciam, além da independência dos anúncios, das declarações de campos textuais, bem como da descrição interna elaborada pelos Anunciantes.

Diante desta seleção, fez-se necessário a categorização e separação dos ambientes de coleta de dados, apresentados a seguir:

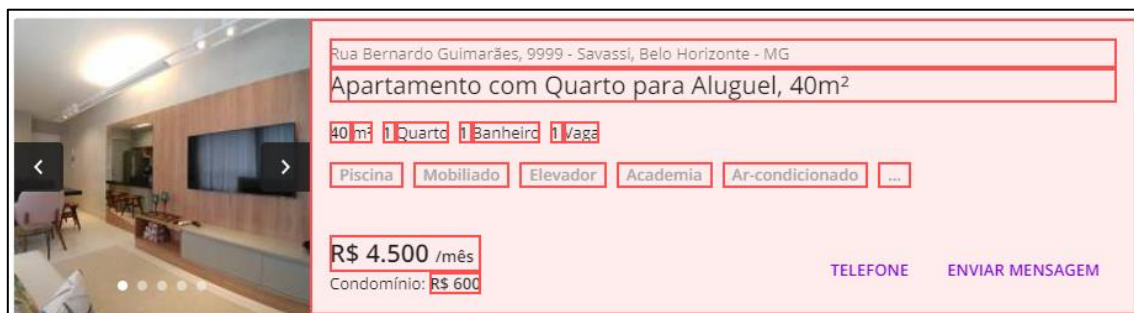
2.3.1. Catálogo Externo

Considerou-se como Catálogo Externo a página inicial de listagem dos anúncios filtrados por tipologia. Neste ambiente, são encontrados os resumos de cada anúncio imobiliário, com características principais, aqui listadas.

- I. Pré Visualização da Imagem do imóvel;
- II. Descrição parcial ou completa do endereço;
- III. Breve descrição da tipologia do imóvel;
- IV. Finalidade de transação do anúncio;
- V. Metragem dada em m²;
- VI. Número de Dormitórios;
- VII. Número de Banheiros;
- VIII. Número de Vagas de Garagem;
- IX. Valor de Venda/Aluguel Pretendido pelo Ofertante;
- X. Valor de Condomínio;
- XI. Grupo de facilidades/comodidades da Área Comum e/ou da Unidade Privativa do Imóvel, limitados até 05 (Cinco) itens;
- XII. Links para contato e informações.

Ressalta-se que desta listagem, não há obrigatoriedade quanto ao preenchimento de todos os campos. Desta forma, mantida a estrutura proposta, os anúncios divergem entre si, a partir das informações disponibilizadas pelo Anunciante.

Figura 3 – Modelo de Disposição de Informações via Catálogo Externo



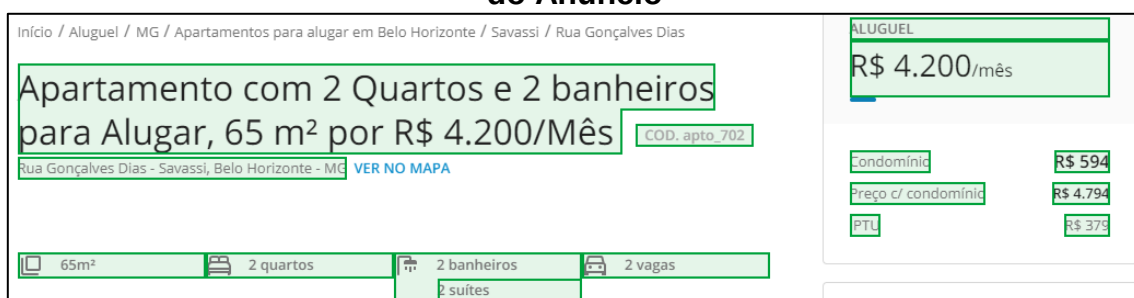
Fonte: Portal Viva Real – Acessado em 22 de Maio de 2023

2.3.2. Corpo Interno do Anúncio

Já de forma independente e separada, tem-se em seguida a apresentação de informações detalhadas através do Corpo Interno do Anúncio. Esta seção tem, além dos campos já descritos anteriormente no Catálogo Externo:

- (i) Um nível maior de detalhamento dos campos textuais;
- (ii) Caixa separada com características completas das facilidades/comodidades da área comum e da unidade privativa;
- (iii) A descrição textual independente elaborada pelo Autor do anúncio, conforme exemplo.

Figura 4 – Blocos Textuais com Maior Nível de Detalhamento via Corpo Interno do Anúncio



Fonte: Portal Viva Real – Acessado em 22 de Maio de 2023

Figura 5 – Caixa em Separado com Características Detalhadas de Facilidade/Comodidades da Área Comum e da Unidade Privativa via Corpo Interno do Anúncio

Características		
Academia	Espaço gourmet	Salão de festas
Área de lazer	Armário embutido	Armário na cozinha
Cozinha americana	Janela grande	Sala de jantar
Hidromassagem	Acesso para deficientes	Sauna
Box blindex	Escada	Porcelanato
Sala de almoço		

Fonte: Portal Viva Real – Acessado em 22 de Maio de 2023

Figura 6 – Descrição textual elaborada de forma independente pelo Autor

Apartamento para aluguel com 40 metros quadrados com 1 quarto em Savassi - Belo Horizonte - MG
O apartamento no bairro Savassi tem 40 metros quadrados com 1 quarto sendo 1 suite e 1 banheiro
Possui área de fitness, jardim, playground, salão para festas e eventos.
Vai lhe possibilitar curtir os dias mais quentes na piscina, todo o conforto do ar condicionado nos dias mais quentes. Elevador para mais praticidade no dia-a-dia.
já tem os móveis incluídos.

Fonte: Portal Viva Real – Acessado em 22 de Maio de 2023

Ressalta-se que essa estrutura não permanece estática, variando em extensão de acordo com cada anúncio, e tendo variações entre plataformas quanto aos campos e características declaradas.

Importante ainda destacar que, todo o armazenamento de dados foi realizado utilizando planilhas do Excel.

2.4. Sequência Lógica de Obtenção de Dados

Apresenta-se a seguir o algoritmo de sequência lógica desenvolvido para raspagem de dados, separado por etapas:

Figura 7 – Sequência Lógica de Obtenção de Dados



Fonte: Elaborado pelo Autor

3. ESTUDO DE CASO E APRESENTAÇÃO DOS DADOS

Apresenta-se a seguir, o Estudo de Caso desenvolvido a partir da Raspagem de Dados obtida através do uso da Plataforma *Octoparse*.

3.1. A Região da Savassi

Fez-se a escolha da Savassi como região para o estudo de caso por se tratar de um bairro localizado no município de Belo Horizonte, conhecido por sua localização privilegiada, bem como da grande disponibilidade de imóveis disponíveis tanto para venda, quanto para locação, bem como de diversas tipologias, favorecendo uma oferta maior de dados, bem como uma análise mais ampla dos mesmos.

No mercado imobiliário residencial, a Savassi é um dos bairros mais desejados da cidade. Possui uma variedade de opções de apartamentos e casas, que vão desde unidades mais antigas e charmosas até empreendimentos mais modernos e sofisticados. Os imóveis na região costumam ter preços mais elevados devido à sua localização central, facilidades de acesso e proximidade de serviços, comércio, restaurantes, bares e vida cultural intensa.

Quanto ao mercado imobiliário comercial, a Savassi é uma região bastante atrativa para negócios. É conhecida por abrigar uma grande quantidade de escritórios, empresas, lojas e estabelecimentos comerciais. Os imóveis comerciais na Savassi são valorizados devido ao fluxo intenso de pessoas e à presença de uma ampla variedade de serviços e comércio na região. É um destino popular para empresas que buscam uma localização estratégica, próximo a importantes vias de acesso e com boa visibilidade para seus negócios.

3.2. Tipologia de Imóveis

Ao fazer uma análise inicial dos anúncios de ambas as plataformas, verificou-se que as principais diferenças nas declarações de dados e características se dão pela classificação da tipologia dos imóveis.

Desta forma, optou-se, para cada uma das plataformas, executar a rotina de raspagem de dados de forma separada para os imóveis de tipologias (a) residências, divididos entre (a.1) apartamentos e (a.2) casas, bem como dos imóveis (b) comerciais, declarados como (b.1) salas, (b.2) lojas, (b.3) andares corridos, (b.4) casas comerciais, (b.5) prédios inteiros e (b.6) pontos comerciais, de acordo com as classificações permitidas em cada site e declarada pelos Anunciantes.

3.3. Classe de Dados

Ante a modelagem apresentada no Capítulo 2 deste trabalho, definiu-se, para fins de armazenamento, estruturação e análise dos dados obtidos, a divisão em classes apresentada a seguir.

3.3.1. Classe Textual

Considerou-se como dados de Classe Textual, quando disponíveis, divididos por campos, os itens a seguir:

- Link Independente de Acesso ao Anúncio do imóvel;
- Descrição da Tipologia do imóvel;
- Endereço do Imóvel, separando por níveis as informações referentes a Logradouro, Numeração, Bairro e Município;
- Código de Identificação do Imóvel, referente ao Anunciante;
- Identificação do Anunciante;
- Contato do Anunciante;
- Registro do Anunciante no Órgão de Classe;
- Descrição Interna elaborada pelo Anunciante.

3.3.2. Classe Numérica ou Quantitativa

Considerou-se como dados de Classe Numérica ou Quantitativa, divididos por campos, os itens a seguir:

- Área Privativa;
- Número de Dormitórios;
- Número de Banheiros;
- Número de Vagas;
- Valor do Aluguel do Imóvel;
- Valor de Condomínio;
- Valor de IPTU.

REALIZAÇÃO

3.3.3. Classe Qualitativa

Por fim, considerou-se como dados de Classe Qualitativa, todos os itens que compõe características declaradas referentes a facilidades/comodidades das áreas comuns e unidades privativas dos imóveis. São os exemplos mais relevantes, elevador, interfone, salão-de-festas, academia, piscina, garagem, cozinha, sauna, ar-condicionado, dentre outros menos de menor frequência.

3.4. Panorama Geral dos Dados

Apresenta-se a seguir, o Panorama Geral dos Dados obtidos, de forma não refinada, através de anúncios para fins de aluguel, no período entre 15 e 22 de Maio de 2023, divididos por Plataforma de Anúncio e Tipologia do Imóvel:

Tabela 1 – Panorama Geral de Imóveis

Panorama Geral de Imóveis			
	Viva Real	Netimóveis	Total
Residencial	256,00	56,00	312,00
Comercial	502,00	277,00	779,00

Fonte: Elaborado pelo Autor

Assim, totalizaram-se, na fase de pré-refinamento, um total de **1.091,00** imóveis verificados durante a execução da raspagem de dados.

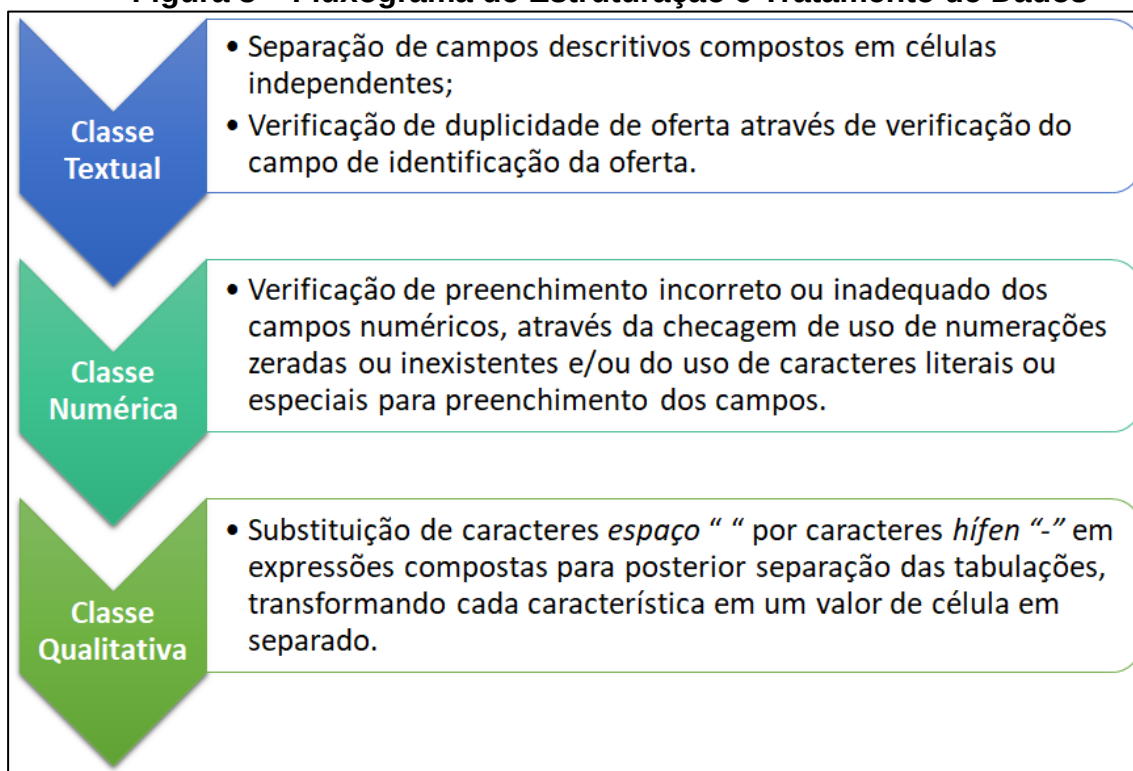
3.5. Fluxograma de Estruturação e Tratamento de Dados

Diante da obtenção dos dados de forma bruta, faz-se necessária a estruturação dos dados. São objetivos do refinamento dos dados:

01. Checagem e remoção de possíveis duplicatas;
02. Verificação de Preenchimento incorreto dos campos das Classes obtidas;
03. Padronização e estruturação das informações obtidas em campos de mesma formatação.

Desta forma, apresenta-se o fluxograma adotado durante o tratamento dos dados:

Figura 8 – Fluxograma de Estruturação e Tratamento de Dados



Fonte: Elaborado pelo Autor

4. ANÁLISE TÉCNICA

A análise técnica se baseará inicialmente nos dados obtidos após as etapas de estruturação e tratamento de dados, trazendo luz às inconsistências das declarações dos dados de Classe Textuais e Numéricas.

Em sequência, faz-se uma análise com foco na geração de inteligência de mercado, através da análise dos campos de Classe Qualitativa.

4.1. Inconsistências de Dados

Demonstra-se a seguir, as principais verificações quanto a inconsistências nas declarações de dados.

4.1.1. Duplicidade de Anúncios

O primeiro passo para a análise dos dados se dá através da verificação, através de campos de identificação dos anúncios, bem como de seus links do respectivo site, da duplicidade de anúncios. Realizada a verificação, tem-se o cenário a seguir:

Tabela 2 – Remoção de Duplicatas de Imóveis

Panorama de Remoção de Duplicatas de Imóveis						
	Viva Real			Netimóveis		
	Iniciais	Duplicatas	Restantes	Iniciais	Duplicatas	Restantes
Residencial	256,00	7,00	249,00	56,00	0,00	56,00
Comercial	502,00	10,00	492,00	277,00	72,00	205,00

Fonte: Elaborado pelo Autor

Assim, do levantamento inicial de dados, foram subtraídos 89 anúncios repetidos, restando 1.002 anúncios válidos.

4.1.2. Informações Cadastrais Incompletas de Endereço

Verifica-se também, para fins cadastrais, a qualidade da declaração dos endereços de cada um dos anúncios remanescentes, aqui apresentados quantitativamente quanto aos níveis mais detalhados de informação:

- ✓ Bairro;
- ✓ Logradouro;
- ✓ Numeração Completa e Válida.

Tabela 3 – Declaração incompleta de Endereço

Panorama de Declaração de Informação Cadastral Incompleta de Endereço						
	Viva Real			Netimóveis		
	Bairro	Logradouro	Nº Completo	Bairro	Logradouro	Nº Completo
Residencial	47,00	133,00	69,00	12,00	23,00	21,00
Comercial	59,00	220,00	213,00	60,00	91,00	54,00

Fonte: Elaborado pelo Autor

4.1.3. Indisponibilidade de Imagens para Conferência de Características

Verificou-se ainda, os anúncios que não possuem imagens para verificação das características declaradas pelo Anunciante, apresentadas a seguir:

Tabela 4 – Indisponibilidade de Imagens para Conferência

Panorama de Indisponibilidade de Imagens para Conferência				
	Viva Real		Netimóveis	
	Disponíveis	Indisponíveis	Disponíveis	Indisponíveis
Residencial	249,00	0,00	49,00	7,00
Comercial	469,00	23,00	161,00	44,00

Fonte: Elaborado pelo Autor

4.2. Observações e Inteligência de Mercado

Num olhar distinto da busca por inconsistências na declaração dos dados, pode-se verificar também informações valiosas quando analisados dados da Classe Qualitativa. Assim, desenvolvem-se algumas perspectivas interessantes a seguir:

4.2.1. Quantidade de Declarações de Características Qualitativas Distintas

Realizando-se o levantamento quantitativo das características declaradas na Classe Qualitativa, pode-se observar o seguinte universo de distribuição:

Tabela 5 – Características Qualitativas – Levantamento Quantitativo

Declaração de Características Qualitativas - Qtde de Termos Independentes		
	Viva Real	Netimóveis
	Caixa de Características Interna	Caixa de Características Interna
Residencial	128,00	60,00
Comercial	86,00	32,00

Fonte: Elaborado pelo Autor

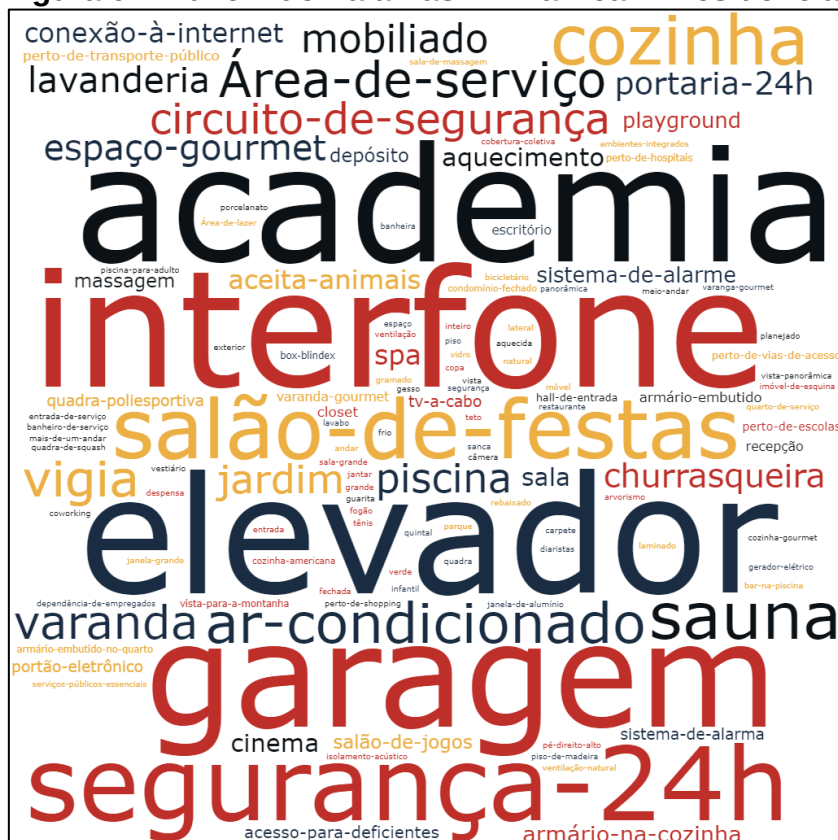
4.2.2. Histograma de Frequência de e Nuvem de Palavras das Características Qualitativas

Em seguida, realizou-se, quantitativamente, o levantamento do histograma de frequência dos termos, para cada um dos cenários considerados.

Porém, antes, faz-se uma listagem de eventos percebidos nessas análises:

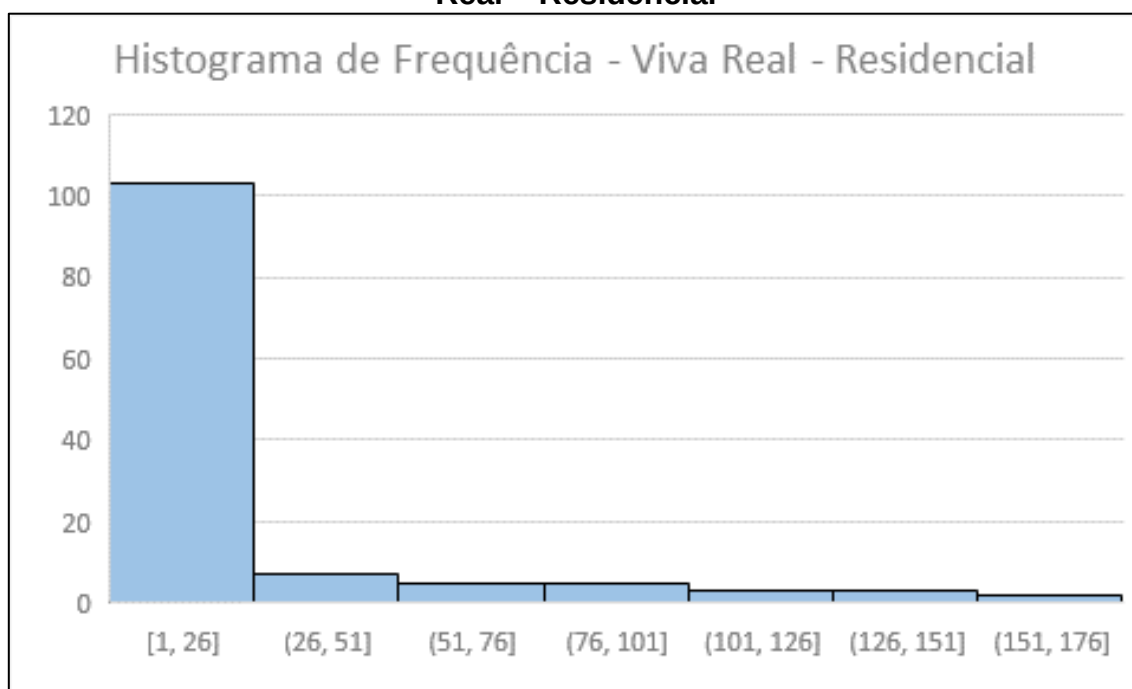
- Estão disponíveis 249 anúncios de imóveis residências do Viva Real, todos declarados com vagas de garagem. Entretanto, na Caixa de Características Internas, a ocorrência de declaração do mesmo termo “*garagem*” aparece somente 122 vezes, gerando uma possível inconsistência de informação;
- Elevadores são as características qualitativas mais declaradas, quando analisada a tipologia de imóveis residenciais, em ambos os portais;
- O portal Viva Real tem atribuições mais ligadas a facilidades referentes ao uso da área comum, enquanto o portal Netimóveis permite declarações de características quantitativas das unidades privativas como Classe Qualitativa;
- Ao menos 66,67% dos termos qualitativos declarados, independente de portal ou tipologia, aparecem com uma frequência inferior a 25,00% do termo de maior frequência em cada tipologia;
- O Portal Viva Real apresenta maior quantidade estruturada de campos qualitativos declarados em relação ao Netimóveis.

Figura 9 – Nuvem de Palavras – Viva Real – Residencial



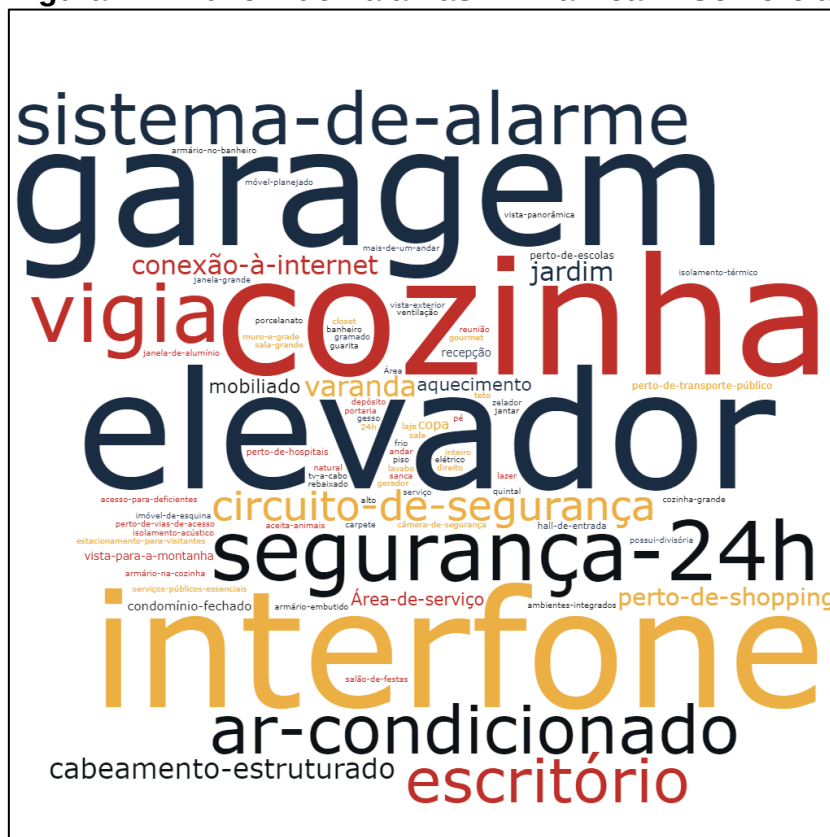
Fonte: Elaborado pelo Autor

Figura 10 - Histograma de Frequência das Características Qualitativas – Viva Real – Residencial



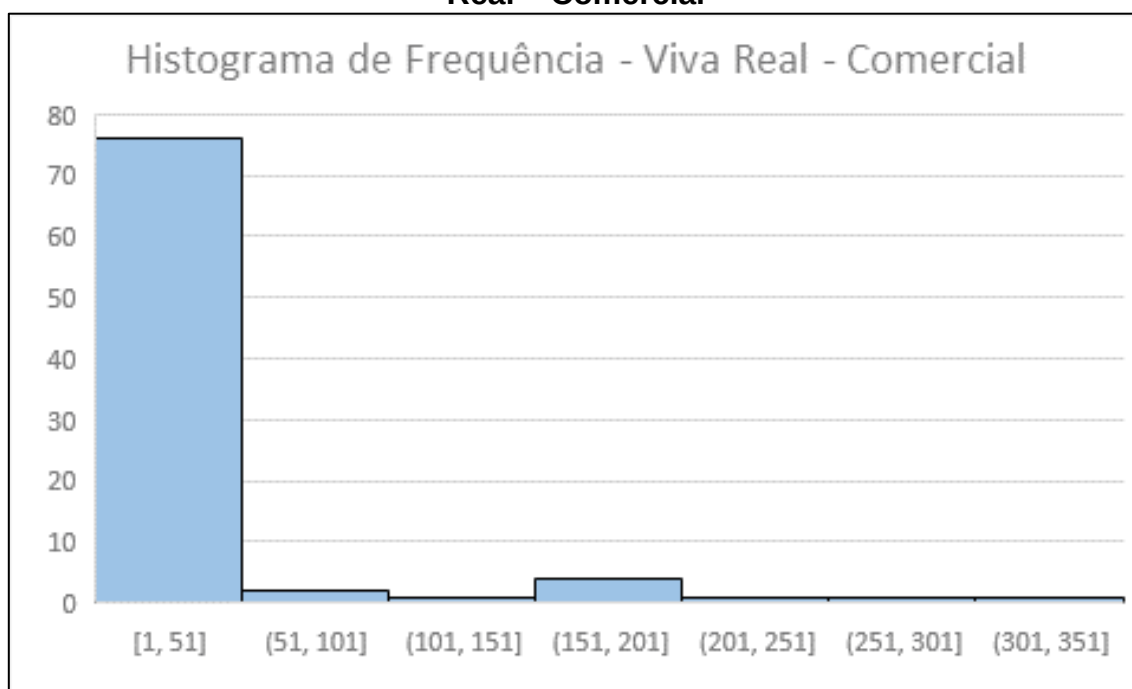
Fonte: Elaborado pelo Autor

Figura 11 - Nuvem de Palavras – Viva Real – Comercial



Fonte: Elaborado pelo Autor

Figura 12 - Histograma de Frequência das Características Qualitativas – Viva Real – Comercial



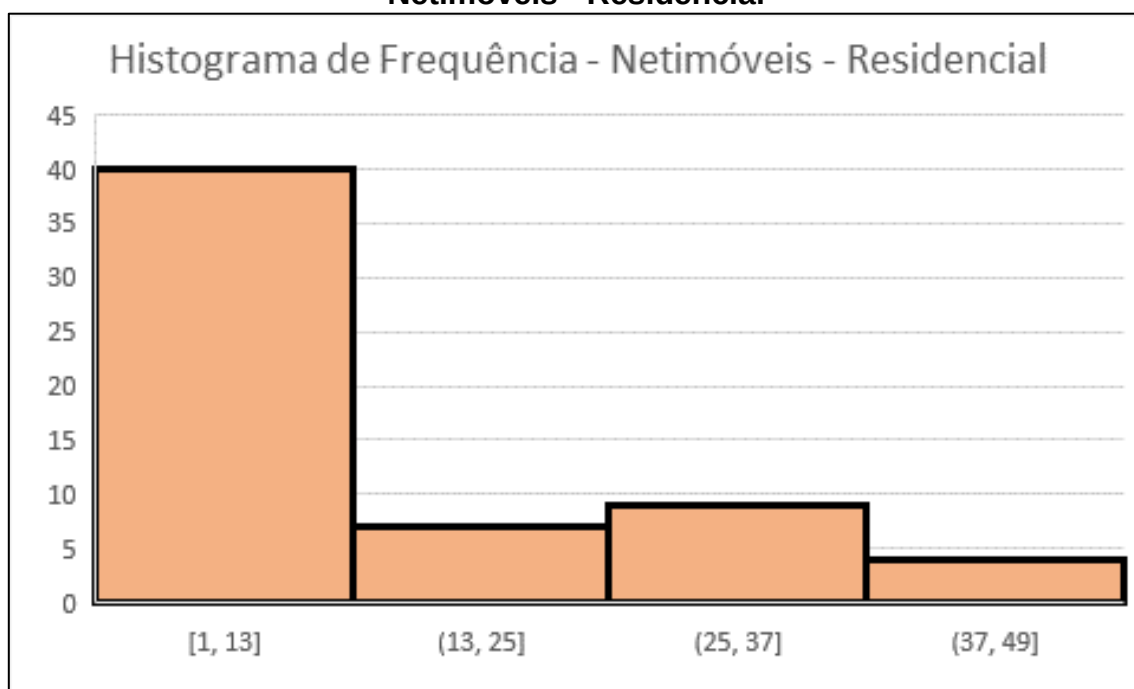
Fonte: Elaborado pelo Autor

Figura 13 – Nuvem de Palavras – Netimóveis – Residencial



Fonte: Elaborado pelo Autor

Figura 14 - Histograma de Frequência das Características Qualitativas – Netimóveis - Residencial



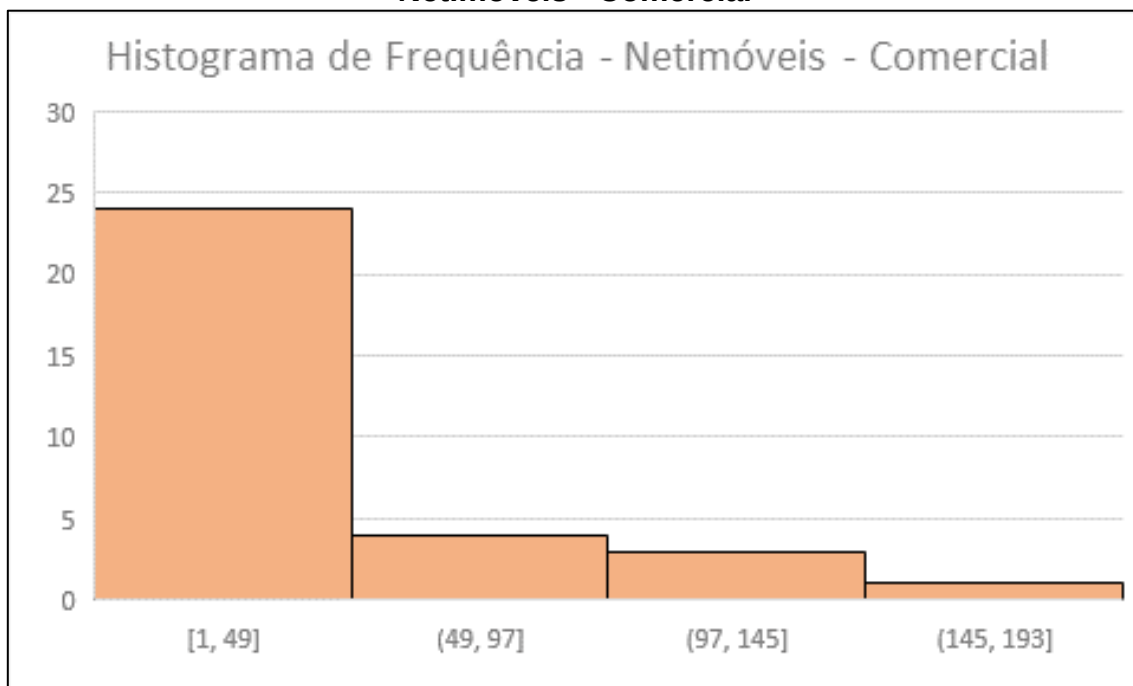
Fonte: Elaborado pelo Autor

Figura 15 – Nuvem de Palavras – Netimóveis - Comercial



Fonte: Elaborado pelo Autor

Figura 16 – Histograma de Frequência das Características Qualitativas – Netimóveis - Comercial



Fonte: Elaborado pelo Autor

5. CONCLUSÕES (CONCLUSÕES E RECOMENDAÇÕES)

Dado o Estudo de Caso em tela, pode-se chegar as seguintes constatações:

5.1. Riscos

- ❖ Diferentes estruturas de publicidade de dados entre as plataformas geram um nível alto de incompatibilidade para comparações diretas dos respectivos dados, sendo um fator de dificuldade, a princípio, para uso em larga escala;
- ❖ A não realização de conferência visual, técnica que demanda investimentos mais robustos em recursos computacionais e tempo de execução, diminui o nível de confiabilidade da análise dos dados;
- ❖ A qualidade da declaração das informações e o uso correto dos campos adequados para cada Classe de Dados, exercida de forma independente pelo Anunciante, é primordial para a qualidade da extração de dados e pode afetar diretamente a qualidade das análises;
- ❖ A estruturação do portal e das permissões de tipologia de declaração de dados interfere diretamente na qualidade e disponibilidade de informações.

5.2. Oportunidades

- ❖ A análise das Classes Textuais e Numéricas, depois de eliminadas inconsistências, podem se tornar fontes valiosas de dados para o uso nas avaliações imobiliárias, pois são de estruturação e tratamento de nível mais fácil e rápido quando comparados a Classe Qualitativa;
- ❖ A geração de um histograma de frequência, bem como uma nuvem de palavras, pode indicar qualitativamente quais são os termos mais declarados, e por consequência, com maior percepção de valor para aquele determinado mercado;
- ❖ A realização da Raspagem de Dados feita periodicamente pode reduzir as inconsistências e o descarte de dados, bem como fornece informações valiosas quanto a velocidade de comercialização de um determinado imóvel e das flutuações de preços declarados no mercado.

5.3. Considerações finais e observações para trabalhos futuros

O levantamento de informações via Raspagem de Dados, realizado de forma rápida, demonstra-se muito mais adequado à geração de conhecimento de mercado do que à aplicação direta nos modelos de avaliação imobiliária.

Tem-se como sugestões para trabalhos futuros, (i) a realização de Raspagem de Dados utilizando de maior abrangência de regiões diferentes, bem como da (ii) execução de rotina de Web Scraping periódicas, para verificação de flutuações e alterações durante um período pré determinado de tempo, e, por fim, da realização da (iii) inclusão da verificação visual das imagens disponibilizadas junto aos anúncios.

REFERÊNCIAS BIBLIOGRÁFICAS

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 14653-1 – Norma brasileira para avaliação de bens – Parte 1: procedimentos gerais**. São Paulo: ABNT, 2019.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 14653-2 – Avaliação de bens - Parte 2: Imóveis urbanos**. São Paulo: ABNT, 2011.

CHAPMAN, Pete et al. CRISP-DM 1.0: **Step-by-step data mining guide**. SPSS inc, 2000. Disponível em: <<https://www.the-modeling-agency.com/crisp-dm.pdf>>. Acesso em: 22 mai. 2023

MACHADO, Bruno Guia Neto. **Avaliação Imobiliária com Redes Neurais Artificiais**. Instituto Superior de Engenharia de Lisboa, Janeiro de 2022. Dissertação de Mestrado.

MOAIAD, Khder. **Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application**. International Journal of Advances in Soft Computing and its Applications, November 28, 2021.

SANTOS, João Manuel Azevedo. **Real Estate Market Data Scraping and Analysis for Financial Investments**. September 10, 2018. Mestrado Integrado em Engenharia Informática e Computação.