



Escrever título (escolher no final)

Universidade Federal da Paraíba - CCEN

Gabriel de Jesus Pereira

11 de maio de 2024

Índice

Resumo	3
Capítulo 1	4
Introdução	4
Objetivos	4
Objetivo Geral	4
Objetivos Específicos	4
Organização do Trabalho	4
Capítulo 2	5
Recursos Computacionais	5
Linguagem de Programação R	5
Quarto	5
Linguagem de Programação Python	5
Web Scraping	5
Metodologia	6
Os dados e o procedimento adotado para sua obtenção	6
Capítulo 4	8
Resultados	8
Capítulo 5	9
Conclusão e Discussão	9

Resumo

Capítulo 1

– Fazer antes da conclusão –

Introdução

Objetivos

Objetivo Geral

Objetivos Específicos

Organização do Trabalho

Capítulo 2

– Fazer depois da metodologia –

Recursos Computacionais

Linguagem de Programação R

Quarto

Linguagem de Programação Python

Web Scraping

Metodologia

Os dados e o procedimento adotado para sua obtenção

O Web scraping, também conhecido como extração de dados da web, é uma técnica utilizada para o processo de coleta de dados estruturados da web de maneira automatizada. É um processo que vem sendo constantemente utilizado por instituições públicas e privadas para a construção de produtos que utilizam algoritmos de aprendizagem de máquinas, observa ofertas e descontos, faz análise de mercado ou monitoração de marcas.

Neste projeto, para fins de estudo e análise do mercado imobiliário, os dados foram coletados por meio de extração de dados do site do Zap Imóveis. O Zap Imóveis é um site do Grupo OLX que reúne ofertas do mercado imobiliário e que funciona como uma plataforma dinâmica para facilitar a conexão entre quem deseja alugar, comprar ou vender um imóvel; podendo servir também para corretores ou outros profissionais do setor de imóveis. Este projeto foi possível graças as informações que foram coletadas do site do Zap Imóveis em dois diferentes períodos do ano de 2023. O primeiro deles, as informações foram coletadas utilizando variados pacotes para raspagem de dados e proxies rotativas da linguagem de programação R, a fim de evitar ser bloqueado pelos mecanismos de segurança do site. Na segunda etapa, os dados foram coletados empregando a linguagem de programação Python com as bibliotecas Scrapy e Playwright, que serve para web crawling e web scraping, e o Playwright que serve para testes em aplicativos da web, mas que neste caso foi utilizado para manejar páginas dinâmicas.

Desta forma, com a ideia de modelar o valor do imóvel e analisar o mercado imobiliário, foram coletados aqueles variáveis que estavam disponíveis no site do Zap Imóveis e que poderiam de alguma forma ser significativas ao tentar explicar o valor do imóvel durante a sua modelagem. Assim, no total foram coletadas 23 variáveis, das quais 8 são quantitativas e 15 qualitativas nominais, sendo 13 de caráter dicotômico. No entanto, nem todas essas variáveis foram coletadas diretamente do Zap Imóveis, a latitude e longitude foram obtidas pela geocodificação do endereço utilizando o pacote tidygeocoder da linguagem de programação R. Portanto, temos as seguintes variáveis:

- Valor do imóvel: esta é a variável dependente, aquela que será modelada e será o principal objeto de estudo deste trabalho;
- Área: área do imóvel em m^2 ;
- Condomínio: valor pago pelo condomínio;

- IPTU: imposto cobrado de quem tem um imóvel urbano;
- Banheiro: quantidade de banheiros presentes na propriedade;
- Vaga de estacionamento: quantidade total de vagas de estacionamento;
- Quarto: quantidade de quartos no imóvel;
- Latitude: posição horizontal medida em frações decimais de graus;
- Longitude: posição vertical que, assim como a latitude, é medida em frações decimais de graus;
- Tipo do imóvel: foram obtidos 7 tipos de imóveis, apartamentos, casas, casas comerciais, casas de condomínio, casas de vila, coberturas, lotes comerciais e de condomínio;
- Variáveis dicotômicas que indicam se o imóvel tem ou não aquela característica (representado como 1 ou 0, respectivamente): área de serviço, academia, elevador, espaço gourmet, piscina, playground, portaria 24 horas, quadra de esporte, salão de festa, sauna, spa e varanda gourmet;

No entanto, devido a observações feitas durante o estudo, nem todas essas variáveis foram utilizadas para a modelagem do valor dos imóveis, seja por conter muitos valores ausentes ou por não ter se mostrado significativo para o que se desejava explicar. Ainda, como a coleta destes dados foram feitas em dois momentos distintos, temos dois bancos de dados, um com 29712 observações e o outro com 14956. Por fim, essas duas bases de dados foram unidas e, para não correr o risco de conter imóveis repetidos, aqueles que tinham o mesmo número de identificação foram removidos.

Capítulo 4

Resultados

Capítulo 5

Conclusão e Discussão