Untitled

Guilherme Oliveira Casagrande

2023-12-04

## Análise dos dados coletados

O processo de raspagem de dados foi implementado em Python, fazendo uso de bibliotecas como Selenium para interação com páginas web dinâmicas, FastAPI para a criação de uma API que recebe e armazena dados, e MySQL para o armazenamento em um banco de dados. A coleta de dados envolveu a navegação pelo site de imóveis Vivareal, extração de informações importantes sobre os imóveis de suas páginas, e posterior armazenamento em um banco de dados para análise posterior.

A avaliação da raspagem de dados revelou algumas considerações importantes para a qualidade e aplicabilidade dos dados coletados. Inicialmente, ao raspar dados dos sites Lopes Imóveis e Vivareal, ficou evidente que a porcentagem da amostra resultante utilizável para o intuito do projeto era significativamente pequena, especialmente devido à inadequação dos dados provenientes do site Lopes Imóveis para utilização efetiva em nosso projeto. Diante desse desafio, reconhecemos a necessidade de explorar mais fontes de dados para enriquecer a amostra.

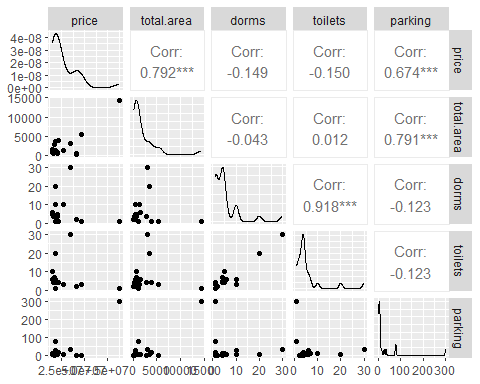
Além disso, observamos que, para um projeto mais abrangente e informativo, teria sido interessante explorar um tema que envolvesse mais variáveis manipuláveis. A ampliação do escopo permitiria uma análise mais completa e a identificação de padrões mais sutis no mercado imobiliário de luxo em São Paulo.

Uma constatação relevante foi a predominância de dados relacionados a galpões, pontos comerciais, terrenos e fazendas, constituindo a maioria da amostra coletada. Esse excesso de dados em categorias específicas, especialmente galpões, pontos comerciais e fazendas, levou à exclusão de metade da amostra, uma vez que esses tipos de propriedades não atendiam ao propósito central do projeto. Reconhecemos que, em futuras raspagens, seria fundamental implementar limites mais precisos para evitar a inclusão excessiva de categorias que não contribuem significativamente para os objetivos do projeto.

## Descrição dos dados e variáveis coletadas

1. ID Único: Um identificador único atribuído a cada entrada no conjunto de dados, garantindo a distinção entre diferentes registros.
2. Fonte (Site Vivareal e Lopes Imóveis): A indicação da fonte dos dados, que é o site de imóveis Vivareal, esclarecendo a origem das informações.
3. Número de Dormitórios, Banheiros e Vagas (separadamente): Informações sobre a quantidade de dormitórios, banheiros e vagas de garagem, detalhando a configuração interna da propriedade.
4. Preço: O valor monetário associado a cada propriedade, refletindo o preço de venda ou locação.
5. Área Total: A medida em metros quadrados representa a área total da propriedade, fornecendo uma indicação do tamanho físico do imóvel.
6. Endereço Completo: O endereço completo, incluindo rua, número, bairro, cidade e estado, permitindo uma localização precisa da propriedade.
7. Bairro Específico:O bairro específico destaca a região na qual a propriedade está situada.
8. Tipo do Imóvel: A classificação do imóvel, indicando se é uma casa, apartamento, terreno, prédio, ou outra categoria específica.
9. Latitude e Longitude do Imóvel (separadamente): As coordenadas geográficas que representam a localização exata da propriedade no mapa, possibilitando análises espaciais e visualizações em sistemas de mapeamento.

## Regressão linear



VIF com todas as variáveis

|  | x |
| --- | --- |
| total.area | 2.87 |
| dorms | 6.58 |
| parking | 2.86 |
| toilets | 6.75 |

VIF sem toilets

|  | x |
| --- | --- |
| total.area | 2.70 |
| dorms | 1.02 |
| parking | 2.73 |

## StepAIC

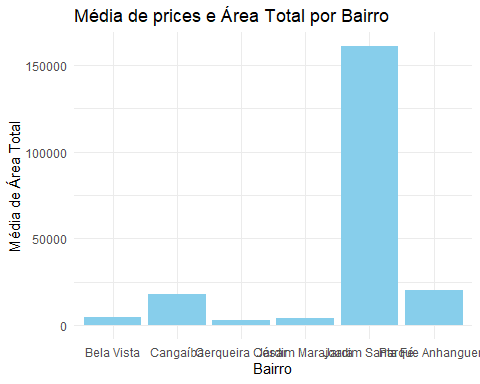
## Start: AIC=811.19  
## price ~ total.area + dorms + parking  
##   
## Df Sum of Sq RSS AIC  
## - parking 1 2.2773e+13 2.2652e+15 809.44  
## - dorms 1 7.0208e+13 2.3126e+15 809.96  
## <none> 2.2424e+15 811.19  
## - total.area 1 1.1781e+15 3.4206e+15 819.74  
##   
## Step: AIC=809.44  
## price ~ total.area + dorms  
##   
## Df Sum of Sq RSS AIC  
## - dorms 1 8.4180e+13 2.3494e+15 808.35  
## <none> 2.2652e+15 809.44  
## - total.area 1 3.9068e+15 6.1720e+15 832.50  
##   
## Step: AIC=808.35  
## price ~ total.area  
##   
## Df Sum of Sq RSS AIC  
## <none> 2.3494e+15 808.35  
## - total.area 1 3.9635e+15 6.3129e+15 831.06

## Modelo

Resultados finais

|  | Estimate | Std. Error | t value | Pr(>|t|) |
| --- | --- | --- | --- | --- |
| (Intercept) | 18894426.70 | 2460225.74 | 7.68 | 0 |
| total.area | 4297.17 | 689.85 | 6.23 | 0 |

## Gráfico de dispersão



## Clusterização

