Relatório LightHouse Indicium

Análise Exploratória dos Dados (EDA)

No meu primeiro contato com a estrutura dos dados e o tipo de desafio, tive algumas ideias iniciais e defini algumas hipóteses sobre quais variáveis interferiam no preço de um imóvel para alugar:

1. A localidade do imóvel tem um impacto significativo no preço
2. O tipo de locação impacta o preço

Além disso tive ideias de métricas a buscar na análise, como o potencial de rentabilidade e visualização da distribuição de imóveis por bairro em conjunto com estações de metrô para discutirmos algumas coisas.

Metodologia da Análise

Primeiramente criei uma forma de limpar o dataset antes de começar com as análises, a função que desenvolvi busca nas colunas se existe algum valor que o **pandas** identificou como NaN e substitui por um valor adequado.

Na minha análise verifiquei que as colunas “nome”, “host_name”, “ultima_review” e “reviews_por_mes” possuiam valores nulos, e troquei eles por “Não informado”, “Não informado”, NaT (Not a Time) e 0 respectivamente.

Além disso identifiquei outliers em linhas com “price” ou “disponibilidade_365” igual a 0, o que deve inferir imóveis em situação inativa, onde talvez o proprietário definiu esses parâmetros para 0 ao invés de excluir o imóvel do sistema.

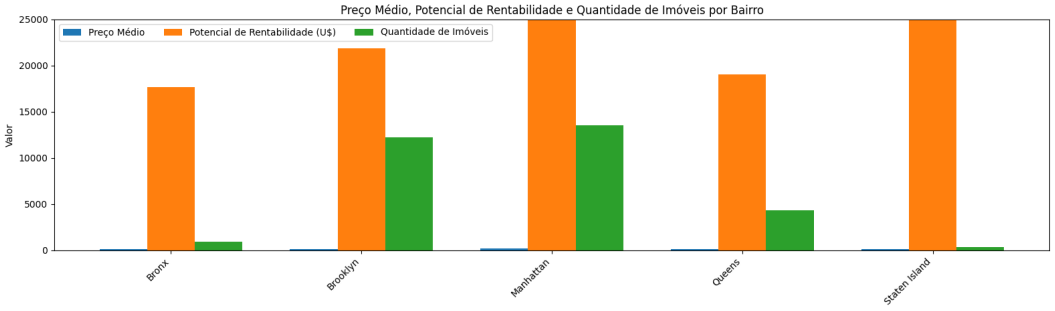
Análise do preço visando o bairro

Nesta etapa eu gerei métricas como o Preço Médio, Desvio Padrão, Quantidade de Imóveis e Potencial de Rentabilidade por bairro.

Para o potencial de rentabilidade eu calculei o preço médio e multipliquei pela disponibilidade média (média de disponibilidade_365) daquele bairro, isso nos gera uma ideia do quanto um imóvel médio pode faturar naquele bairro.

Descrição estatística

Bairro	Preço Médio	Desvio Padrão do Preço	Preço Mínimo	Preço Máximo	Quantidade de Imóveis	Disponibilidade Média	Potencial de Rentabilidade (U$)
Bronx	89.105147864184	112.67676126111807	10	2500	913	197.93647316538883	17637.158709117062
Brooklyn	132.92695070192622	189.0459603432529	10	8000	12252	164.34067907280445	21845.305345431756
Manhattan	214.20207979939522	325.02942577890997	10	10000	13559	178.89121616638394	38318.87056068263
Queens	100.02978129362494	108.58446711853789	10	2600	4298	190.42903676128432	19048.574899186933
Staten Island	114.2296072507553	291.52650770815325	13	5000	331	225.01510574018127	25703.38715418808



Com os gráficos das métricas observamos que Brooklyn e Manhattam possuem o maior preço médio de locação, porém Manhattam tem o **MAIOR** desvio padrão, o que indica uma enorme variação nos preços dos imóveis, Manhattam também tem a segunda disponibilidade média mais baixa, o que impacta o potencial de rentabilidade do bairro.

Análise do preço visando o intervalo de disponibilidade durante o ano e intervalo de noites mínimas.

Nesta análise gerei algumas métricas em intervalos de imóveis, nas categorias de disponibilidade durante o ano e intervalo de noites mínimas.

Tive os seguintes resultados:

Intervalo de Disponibilidades Durante o Ano e Preço

Intervalo de Disponibilidade	Preço Médio	Desvio Padrão do Preço	Preço Mínimo	Preço Máximo	Quantidade no Intervalo
(-0.001, 50.0]	138.63794968893697	141.02255105778877	10	6419	7394
(50.0, 100.0]	148.73166701094814	279.4170978175427	10	10000	4841
(100.0, 250.0]	163.2863994097393	236.80083476040895	10	7703	8132
(250.0, 300.0]	180.0517361111111	232.97786165044718	10	8500	2880
(300.0, 365.0]	183.886627189736	330.0722869269859	10	9999	8106

Intervalo de Noites Mínimas e Preço

Intervalo de Noites Mínimas	Preço Médio	Desvio Padrão do Preço	Preço Mínimo	Preço Máximo	Quantidade no Intervalo
(0, 50]	160.8316611284963	241.45705927286195	10	10000	30997
(50, 100]	272.6698113207547	840.1951953493401	24	9999	212
(100, 250]	300.2421052631579	821.777260754636	35	6500	95
(250, 300]	131.5	46.47042069962354	65	199	6
(300, 365]	266.7241379310345	439.0543325108542	50	2350	29
(365, 1250]	137.57142857142858	96.52079432051798	45	400	14

É visto que em relação a disponibilidade, o intervalo que possui o maior preço médio são os imóveis que tem disponibilidade entre 300 e 365, mas este intervalo também é o que tem o maior desvio padrão, o que nos mostra uma grande variação no dataset. Enquanto o primeiro intervalo (0 e 50) tem o menor desvio padrão e o preço médio relativamente baixo dentre o dataset.

Já em relação as noites mínimas identificamos imóveis que requerem entre 50 e 100 noites mínimas são a maioria entre o dataset, e também tem um valor relativamente médio entre todos os intervalos. É importante destacar também que os outros intervalos possuem um número extremamente pequeno de imóveis o que torna o desvio padrão muito elevado em alguns casos, e faz com que as métricas não sejam muito confiáveis.

> 💡 Irei abordar mais análises do EDA ao responder as perguntas a seguir, definidas no desafio.

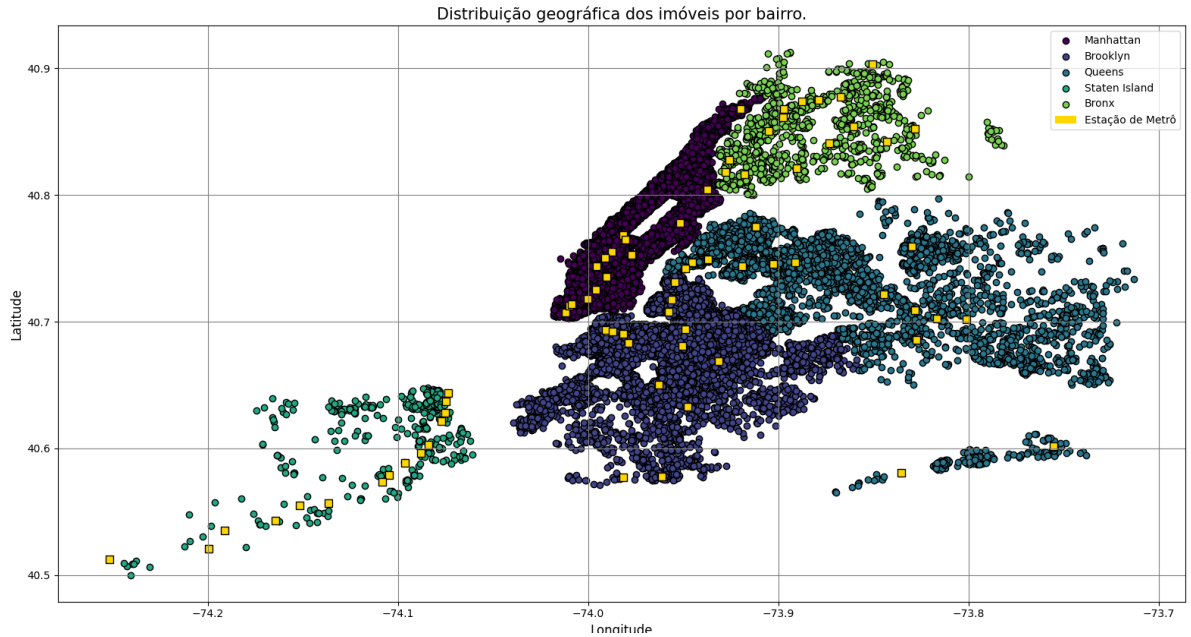
1. Supondo que uma pessoa esteja pensando em investir em um apartamento para alugar na plataforma, onde seria mais indicada a compra?

Com base na análise do preço, em relação aos bairros, nos podemos inferir que o Brooklyn é melhor bairro para se investir, visando o lucro com base nas informações que temos.

Por ser o segundo bairro com maior preço médio e terceiro menor em desvio padrão do preço ele oferece métricas confiáveis no que se diz ao potencial de rentabilidade.

Porém para essa decisão o comprador do imóvel também deve se atentar em outras informações além das métricas, como locomoção e facilidade de acesso a serviços básicos, na minha análise eu também trouxe a distribuição geográfica dos imóveis por bairro e introduzi alguns pontos de estação de metrô, para termos uma noção da mobilidade em cada bairro. Vemos que além de ter o maior preço médio, o bairro do Brooklyn também possui boa dispersão de estações de metrô.

Com essas análises combinadas podemos reiterar que o Brooklyn é um local indicado para a compra.



2. O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?

Apesar de as análises iniciais da correlação dessas duas variáveis durante o EDA não demonstrarem grandes interferências nos preços, realizei uma análise de correlação e tive o seguinte resultado:

minimo_noites	numero_de_reviews	reviews_por_mes	calculado_host_listings_count	disponibilidade_365
0.13273634245973515	-0.11209770735578682	-0.11714499191254839	-0.13074426726419075	0.05829030938949284

Esta tabela mostra a correlação das variáveis com o preço, utilizando o método spearman de cálculo de correlação. Ele mostra que o mínimo de noites de estadia é a variável com maior correlação com o preço entre as demais, porém ainda sim é uma identificação muito fraca.

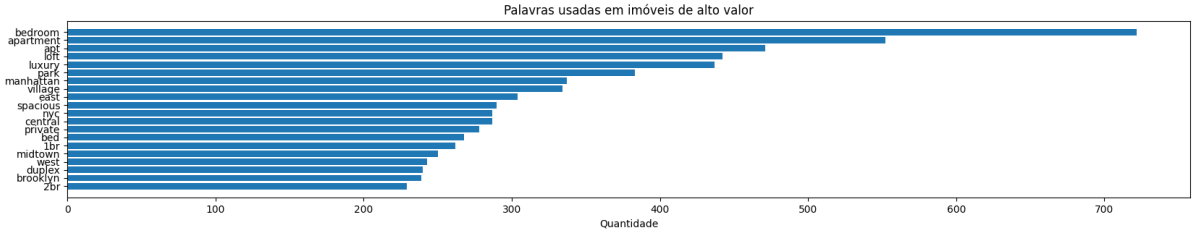
Isto muda quando utilizamos fomos para o passo de criação de um modelo de machine learning, onde além de tudo ele identificou que o minimo de noites e a disponibilidade ao longo do ano tem um forte impacto no preço.

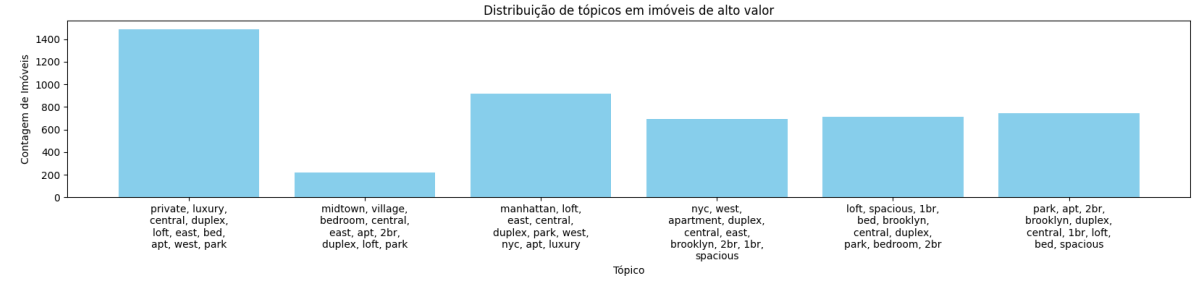
	Minimo de noites	Disponibilidade durante o ano
Modelo otimizado	5° Mais impactante	3° Mais impactante
Modelo não otimizado	4° Mais impactante	3° Mais impactante

O que nos confirma que existe sim uma interferência no preço com base no minimo de noites e na disponibilidade daquele imóvel.

3. Existe algum padrão no texto do nome do local para lugares de mais alto valor?

Para esta etapa utilizei um modelo **LDA (Latent Dirichlet Allocation)** não supervisionado e uma vetorização dos nomes para identificar nomes mais comuns e conjunto de nomes (tópicos) mais comuns em imóveis com o preço acima da média, usei como métrica preço 1.5 vezes mais caros que a média.





Com esse resultado identificamos que existe sim um padrão nos nomes de imóveis de alto valor, principalmente no que se diz a respeito ao 1° tópico, com palavras como “private”, “luxury“ e “duplex”

4. Explique como você faria a previsão do **preço** a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê? Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?

Esse é um problema de regressão, onde estamos buscando prever o valor do preço com base nas outras variáveis.

Além da etapa inicial de limpeza dos dados que ja tinha comentado, também removi os outliers de preço = 0 e disponibilidade_365 = 0 para remover os imóveis inativos, além disso também removi a coluna de “nome” do imóvel, pois percebi que os modelos estavam se ajustando demais ao nome, invés de

buscar relações entre as características físicas do imóvel. Isso poderia gerar um viés no modelo onde ele ignoraria a localidade, tipo de imóvel e etc e prediria o preço com base apenas no nome.

Para essa etapa eu testei o modelo de RandomForestRegressor algumas vezes e utilizei Randomzied Search para otimizar o modelo da melhor forma, tive dificuldades pois meu computador não é completamente ótimo para o tipo de problema de regressão, ainda mais com grandes quantidades de dados.

Para minha análise usei as métricas: RMSE (Raiz do Erro Quadrático Médio) pela grande variação que observei no dataset, e R^2 para avaliar o modelo de forma geral.

Não pude testar muitos modelos por conta do tempo limitado, outras demandas que tive que fazer em conjunto com o desafio e o estado do meu workstation para problemas complexos de machine learning. Mas testei o RandomForestRegressor com e sem otimização e observei que ao otimizar o modelo, mesmo que com poucas iterações obtive um resultado melhor.

5. Com base no modelo, a sugestão de preço do apartamento com as características seria de : U$ 278.18

Vídeo explicativo do projeto: https://drive.google.com/drive/folders/1DSwaC-Zed3Nspe9fdHHK7eO3m5AwZ02j?usp=sharing