RELATÓRIO EDA (Análise Exploratória de Dados)

O objetivo desta análise exploratória é demonstrar as principais características entre as variáveis e levantar hipóteses de negócio relacionadas. E oferecer resposta às seguintes perguntas para a tomada de decisão:

1 - Supondo que uma pessoa esteja pensando em investir em um apartamento para alugar na plataforma, onde seria mais indicada a compra?
2 - O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?
3 - Existe algum padrão no texto do nome do local para lugares de mais alto valor?
4 - Explique como você faria a previsão do preço a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê?
5 - Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras?
6 - Qual medida de performance do modelo foi escolhida e por quê?
7. Supondo um apartamento com as seguintes características:
{'id': 2595,
'nome': 'Skylit Midtown Castle',
'host_id': 2845,
'host_name': 'Jennifer',
'bairro_group': 'Manhattan',
'bairro': 'Midtown',
'latitude': 40.75362,
'longitude': -73.98377,
'room_type': 'Entire home/apt',
'minimo_noites': 1,
'numero_de_reviews': 45,
'ultima_review': '2019-05-21',
'reviews_por_mes': 0.38,
'calculado_host_listings_count': 2,
'disponibilidade_365': 355}
Qual seria a sua sugestão de preço?

Iniciei a análise exploratória com o carregamento do arquivo "teste_indicium_precificacao.csv" onde verifiquei que a base de dados fornecida possui 48.894 linhas e 16 colunas o que aparentemente pode

ser uma quantidade razoável de dados. Além disso foi verificado à primeira vista que nas 10 primeiras e 5 últimas linhas as colunas "números de reviews" possuem quantidade igual a zero. E que nas colunas "ultima review" e "review por mes" estava com um NAN (Not A Number). E que os o total de valores nulos são de 10.052, o que demonstra que os dados precisam passar por um tratamento para não comprometer o modelo e os insights que nos possibilitará responder às perguntas anteriores.

Após análise dos valores nulos, foi constatado que as colunas "ultima_review (última avaliação)" e "reviews_por_mes (avaliações por mês)", possuem o total de 10.052 valores nulos cada.

Primeira hipótese
A partir da análise prévia foi possível também verificar, que o número de reviews, a quantidade de reviews e data da avaliação (postagem) são irrelevantes no valor dos aluguéis. Pois, a análise exploratória possibilitou identificar que há valores de aluguéis mais altos e com poucos o nenhum review (avaliação).

Segunda hipótese
A partir dos cálculos estatísticos de período de ocupação por ano por tipo de imóvel, é possível concluir que os imóveis completos (entire home/apt) possuem a melhor relação tempo de ocupação x preço em relação aos quartos privativos ou quartos compartilhados. Veja a comparação abaixo:
Entire home/apt
Ficam ocupados em média 111 dias e o preço médio é $211,79.
Private room
Ficam ocupados em média 111 dias e o preço médio e $89,78
Shared room
Ficam ocupados em média 162 dias e o preço médio é $70,13.

Portanto, isto confirma que há alta demanda por este tipo de imóvel, possibilitando um maior potencial de lucro em relação aos outros. Todavia, Supondo que uma pessoa esteja pensando em investir em um apartamento para alugar na plataforma, a análise da correlação bairro x preço confirmada pela análise gráfica, indica que o melhor local para a aquisição do imóvel é o bairro de Manhattan.

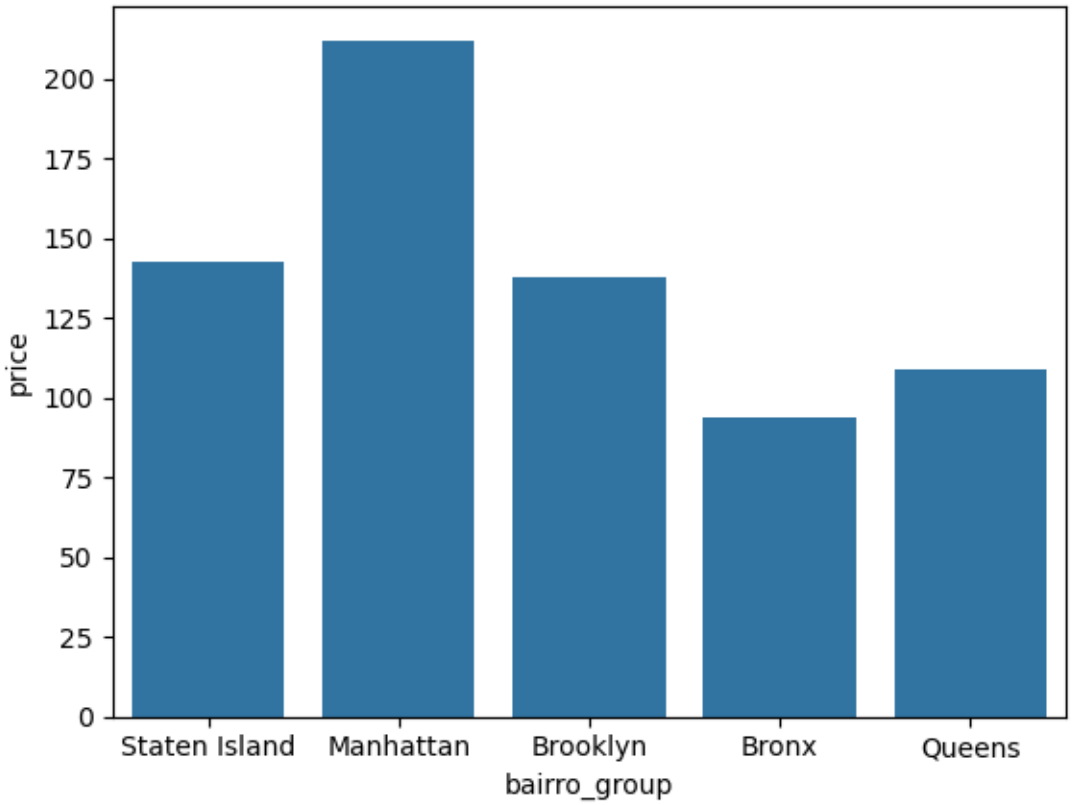
No que diz respeito a se o número mínimo de noites e a disponibilidade ao longo do ano, se interferem no preço. A análise de correlação indica que essas variáveis não possuem impactos significativos no preço.

Quanto à existência de algum padrão no texto do nome do local para lugares de mais alto valor, a análise do impacto das palavras nos anúncios em relação ao preço revelou que para a palavra "Penthouse", o preço manteve próximo dos $300; a palavra "luxury" manteve o preço acima dos $250 e as palavras "view", "loft" e "Duplex" mantiveram o preço do aluguel acima dos $200 mas abaixo dos $250 dólares. Inclusive, acrescentar palavras como estas podem ser mais descritivas e aumentar o valor percebido do cliente em relação ao imóvel.

Portanto, após a análise exploratória, temos subsídios suficientes para responder às perguntas propostas acima:

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

Resposta: Manhattan, conforme a análise de correlação preço x bairro.



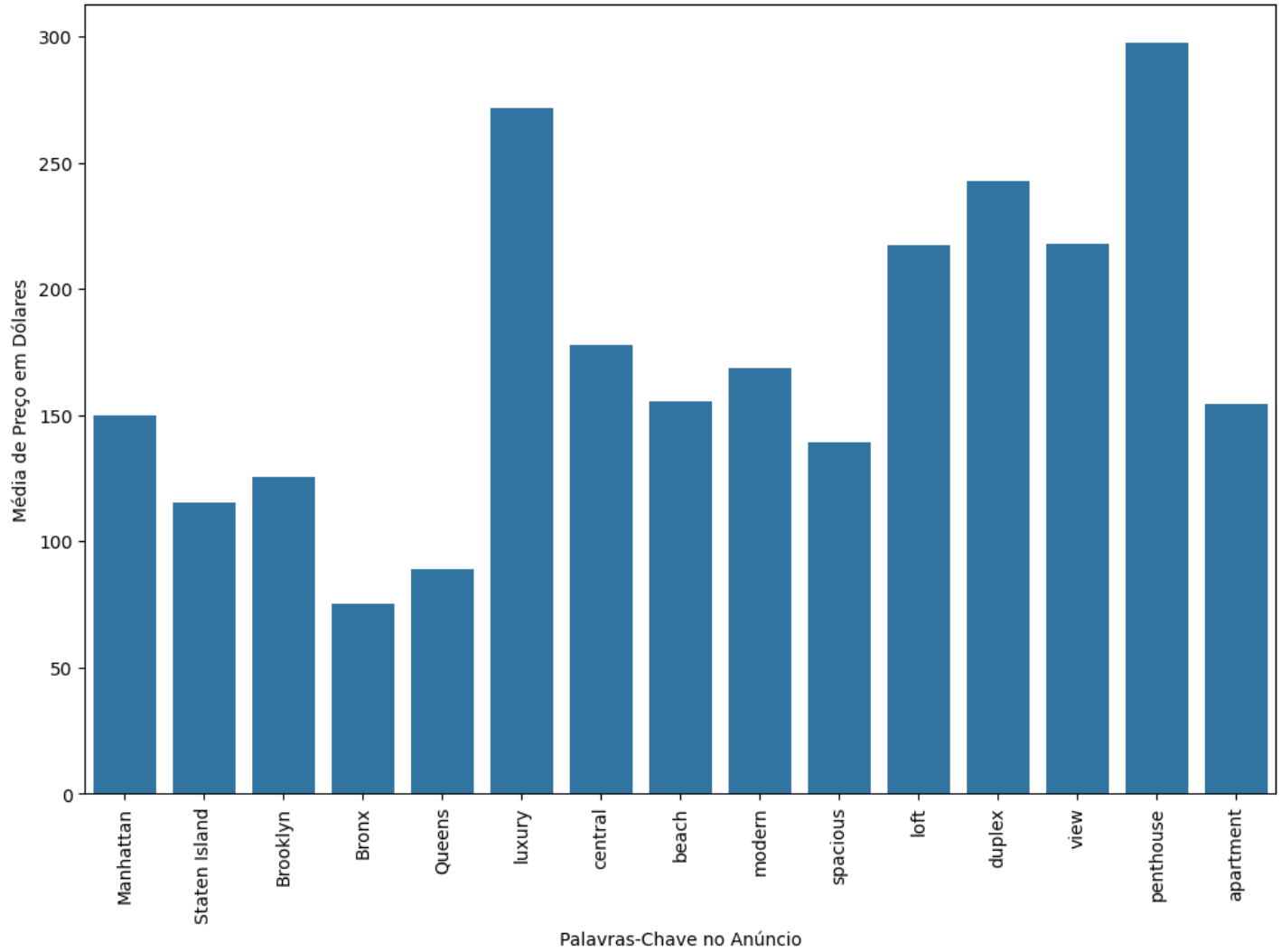
2 - O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?

Resposta: As variáveis “mínimo de noites” e a disponibilidade ao longo do ano *não interferem* no preço. Pois, possui correlação muito fraca conforme mostra a tabela abaixo.

	minimo_noites	disponibilidade_365	price
minimo_noites	1.000000	0.144320	0.042799
disponibilidade_365	0.144320	1.000000	0.081833
price	0.042799	0.081833	1.000000

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

Resposta: Sim. Há algum impacto para os títulos que continham palavra “Penthouse”, o preço manteve próximo dos $300; para palavra "luxury" manteve o preço acima dos $250 e no que diz respeito às palavras "view", "loft" e "Duplex" mantiveram o preço do aluguel acima dos $200 mas abaixo dos $250 dólares.



4 - Explique como você faria a previsão do preço a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê?

Respostas: Eu faria a previsão do texto a partir dos dados, baseado na mediana de preços praticados por proprietários de imóveis com as mesmas características. As variáveis que eu utilizei foram a

localização geográfica (latitude, longitude), palavras chaves que remetem às características do imóvel, o tipo de acomodação e o nome do distrito de Nova York (bairro). A escolha destas variáveis foi em decorrência do grau de relevância e impacto no preço da diária do aluguel.
No entanto, alguns destes dados requereu transformações para não comprometer o modelo. Como por exemplo, a transformações dos dados do tipo object para o tipo inteiro (0 ou 1) para que o modelo interprete melhor os dados de forma representativa.
Foi realizada também o escalonamento das variáveis longitude e latitude como técnica de normalização de dados com o intuito de melhorar o desempenho do modelo deixando a escala destas variáveis uniforme.

5 - Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras?

Respostas: É um problema de regressão. E o modelo que pelo menos na teoria, se aproxima dos dados seria o Linear Regression (Regressão Linear). Ele é de fácil implementação e interpretação. Porém, é muito sensível a dados incomuns, como outliers e não captura bem as relações não lineares. Quero dizer as "features" (variáveis) e o target (preço).

6 - Qual medida de performance do modelo foi escolhida e por quê?

Respostas: Eu escolhi a MAE (Erro Médio Absoluto) e R^2 (Coeficiente de Determinação). A primeira, porque o problema em questão trata-se de valor monetário, tornado assim, fácil a interpretação e a segunda para contrapor com a medida MAE e analisar melhor o quanto o modelo estava acertando nas predições após o treinamento.

7. Supondo um apartamento com as seguintes características:

{'id': 2595,
'nome': 'Skylit Midtown Castle',
'host_id': 2845,
'host_name': 'Jennifer',
'bairro_group': 'Manhattan',
'bairro': 'Midtown',
'latitude': 40.75362,
'longitude': -73.98377,
'room_type': 'Entire home/apt',
'minimo_noites': 1,
'numero_de_reviews': 45,

'ultima_review': '2019-05-21',

'reviews_por_mes': 0.38,

'calculado_host_listings_count': 2,

'disponibilidade_365': 355}

Qual seria a sua sugestão de preço?

Resposta: Como se trata de um apartamento inteiro, bem localizado, no centro de Manhattan e de fácil acesso a varias atrações turísticas. A minha sugestão de preço é de $660 dólares.

Este preço está compatível com o principal concorrente do nosso cliente.

Fonte: Site Airbnb