PDF Relatório

1. Faça uma análise exploratória dos dados (EDA), demonstrando as principais características entre as variáveis e apresentando algumas hipóteses de negócio relacionadas. Seja criativo!

2. Responda também às seguintes perguntas:

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

Preço médio por grupo de bairro:

bairro_group

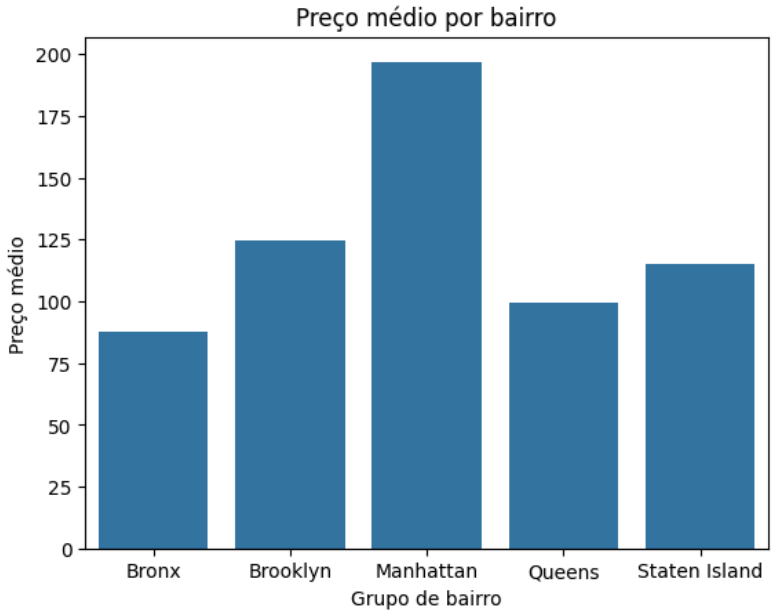
Bronx 87.496792

Brooklyn 124.381983

Manhattan 196.875814

Queens 99.517649

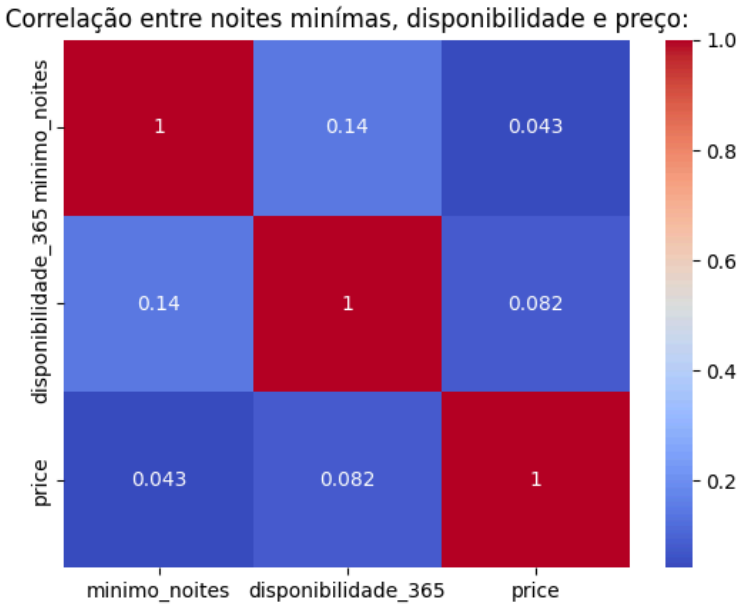
Staten Island 114.812332



Sendo assim, seria recomendado alugar no bairro Bronx, pois possui menores preços médios.

Porém no caso de compra, seria mais interessante bairros com maior preço médio de aluguel como Manhattan.

O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço? Sim, segue abaixo gráfico gerado no notebook demonstrando a afirmação:



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

Sim, existem palavras bem mais comuns em lugares de alto valor: [('in', 2181), ('2', 1332), ('bedroom', 1297), ('apartment', 1112), ('apt', 933), ('luxury', 737), ('village', 728), ('the', 727), ('1', 684), ('loft', 661)]. Um exemplo seria a palavra luxury sendo utlizada constantemente.

1. 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?

O problema que estamos resolvendo é de regressão. Para prever o preço, seria importante utilizar variáveis como latitude, longitude, bairro, disponibilidade, room type e a quantidade de reviews por mês, pois através delas é possível levantar dados de localização, quantas

pessoas utilizam o apartamento por mês e o tipo de quarto. Para medida da performance, foi utilizado o RMSE, pois possui melhor sensibilidade ao erro.

E o modelo foi Random Forest e seus prós são seu desempenho consistente e capacidade de capturar relações não lineares, e contras são consome mais memória e tempo e dificuldade de interpretação completa.

1. 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?

Preço sugerido para o apartamento de exemplo: 287.4433333333333