1. EDA - Desafio Ciência de Dados (Indicium)

Gabriela Brito

Analisando estatísticas do dataset

	bairro_group	room_type	price	minimo_noites	numero_de_reviews	calculado_host_listings_count	disponibilidade_365
count	48894	48894	48894.000000	48894.000000	48894.000000	48894.000000	48894.000000
unique	5	3	NaN	NaN	NaN	NaN	NaN
top	Manhattan	Entire home/apt	NaN	NaN	NaN	NaN	NaN
freq	21661	25409	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	152.720763	7.030085	23.274758	7.144005	112.776169
std	NaN	NaN	240.156625	20.510741	44.550991	32.952855	131.618692
min	NaN	NaN	0.000000	1.000000	0.000000	1.000000	0.000000
25%	NaN	NaN	69.000000	1.000000	1.000000	1.000000	0.000000
50%	NaN	NaN	106.000000	3.000000	5.000000	1.000000	45.000000
75%	NaN	NaN	175.000000	5.000000	24.000000	2.000000	227.000000
max	NaN	NaN	10000.000000	1250.000000	629.000000	327.000000	365.000000

price

Média de preços: $152.72
Desvio padrão: $240.15
Menor valor: $0
Maior valor: $10.000

minimo_noites

Média de mínimo de noites: 7 noites
Desvio padrão: 20 noites
Menor número mínimo: 1 noite
Maior número: 1250

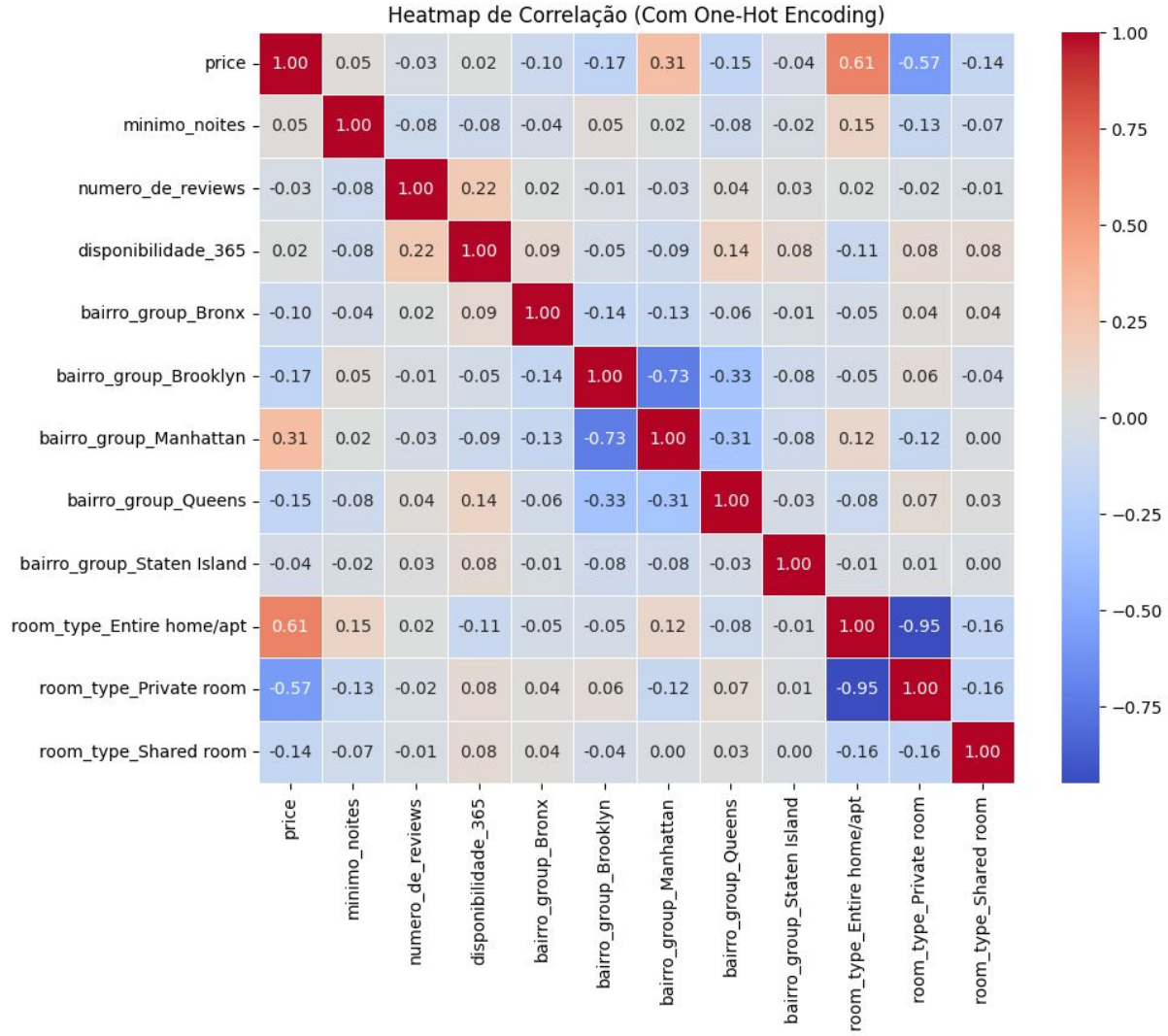
disponibilidade

Média: 112 dias
Desvio padrão: 131 dias
Menor valor: 0 dias
Maior disponibilidade: 365 dias

Análise estatística do slide anterior

Algumas colunas foram retiradas da análise anterior, como id, host_name…
Os dados apresentados mostram que o **min** e **máx** das colunas price, minimo_noites, disponibilidade_365, numero_de_reviews e calculado_host_listings_count indicam a presença de outliers que precisam ser removidos para melhor funcionamento do treinamento do modelo.
Esse fato também está enviesando a média e desvio padrão dos valores.

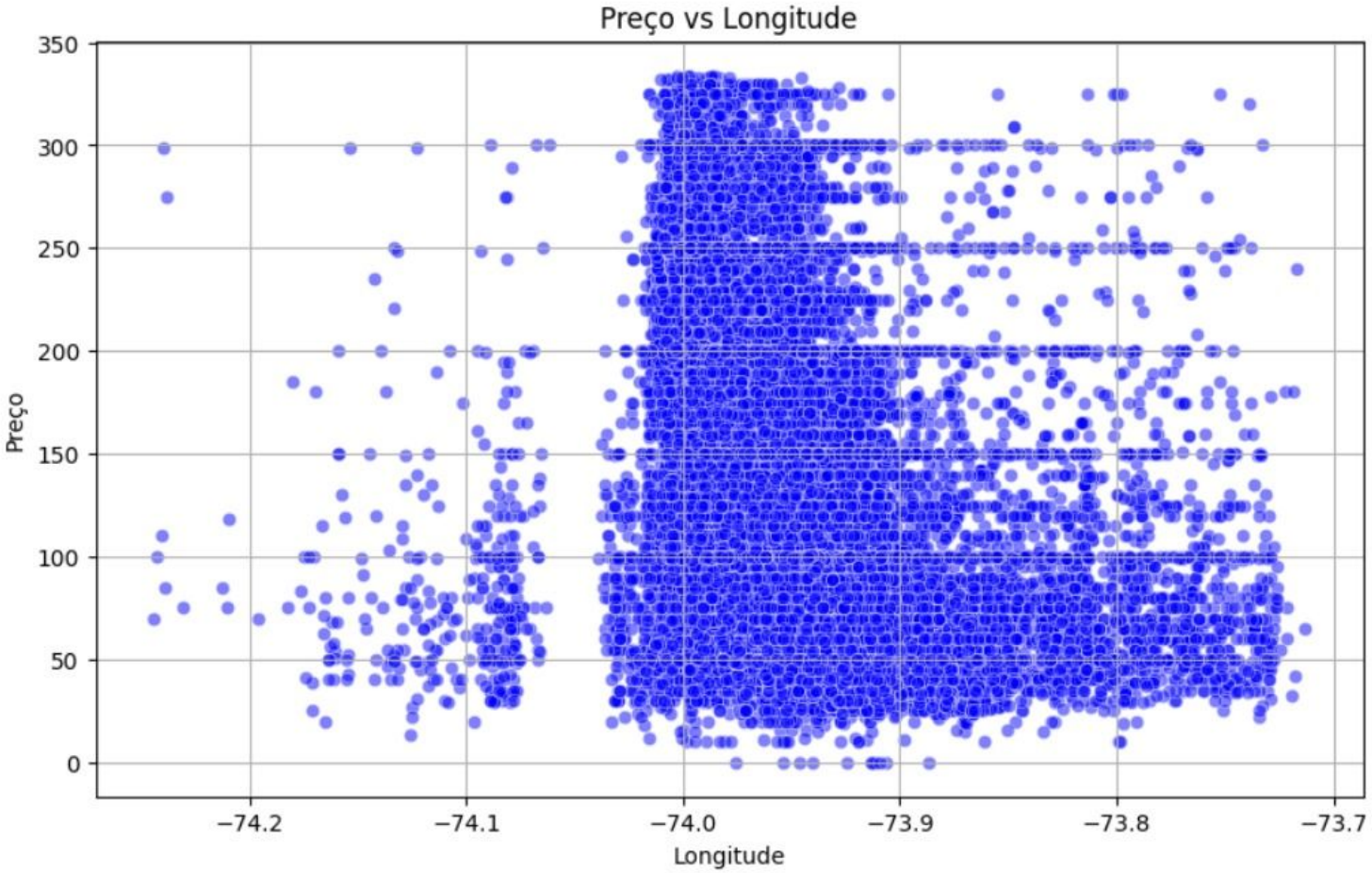
Já as colunas **bairro_group** e **room_type** que, a principio, por intuição, são as colunas que mais devem ter correlação com price, estão com valores nulos devido ao fato de que são colunas com valores categóricos. Necessitando de tratamento adequado (nesse caso, foi escolhido o One-Hot Encoder) para que sejam avaliados com mais detalhes. Porém, a princípio, já se sabe que existem três tipos de tipos de quarto: Entire house/apto (o mais frequente), Private Room e Shared Room. Além disso, os bairros são 5: Bronx, Queens, Brooklyn, Staten Island e Manhattan(o mais frequente).



Após o processo de One-Hot Encoding, percebe-se que as features que possuem mais correlação com **Price** são o tipo de quarto **Entire home/apt (0.61)** e a localização do bairro **Manhattan (0.31)**. Não por engano, são também os mais frequentes.

Outras correlações importantes são entre **numero_de_reviews e disponibilidade_365 (0.22)**; **minimo_noites e room_type_Entire home/apt (0.15)**.

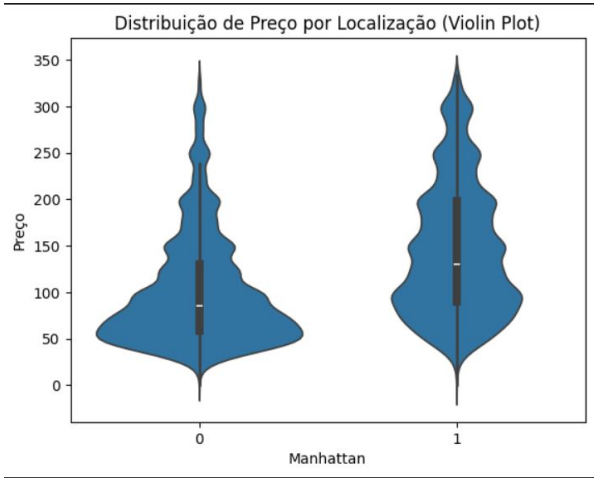
E, ainda, as correlações perfeitamente negativas como **Private Room x Entire home/apt** e **Manhattan x Brooklyn**, indicando que quando um apartamento é alugado na especificação de um, a do outro cai já que são concorrentes.

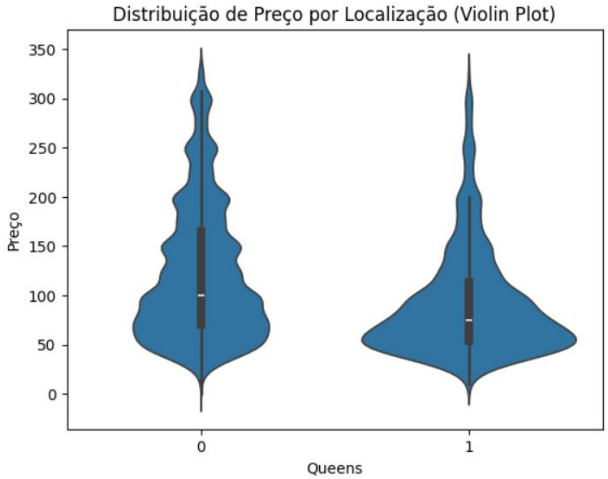


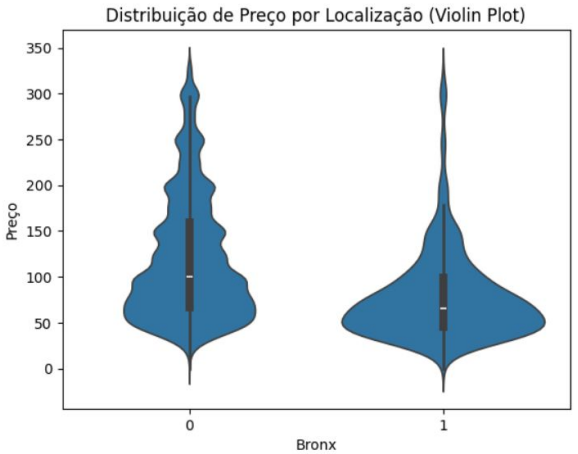
Nesse gráfico, fica mais evidente a relação preço x localização, além do número de apartamentos.

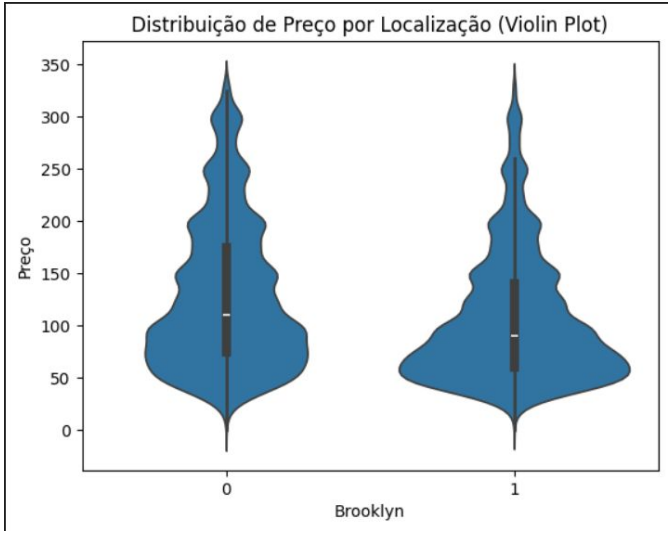
Nesse caso, percebe-se que a região de Staten Island possui pouquíssimos apartamentos disponíveis em sua região.

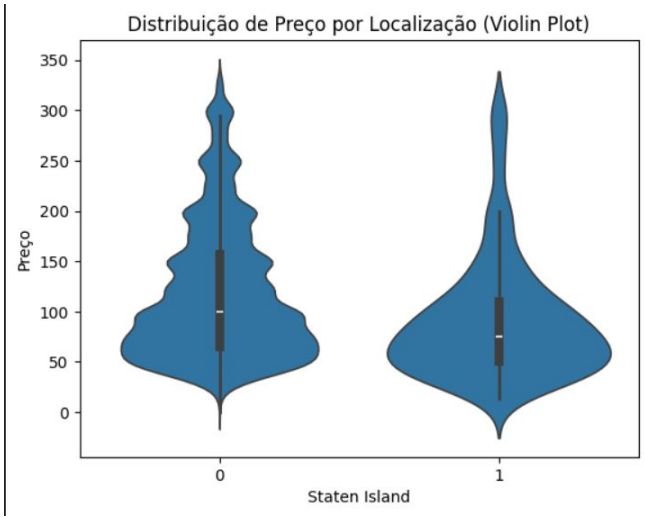
Além disso, os maiores preços concentram-se no meio do gráfico, que refere-se aos bairros centrais.



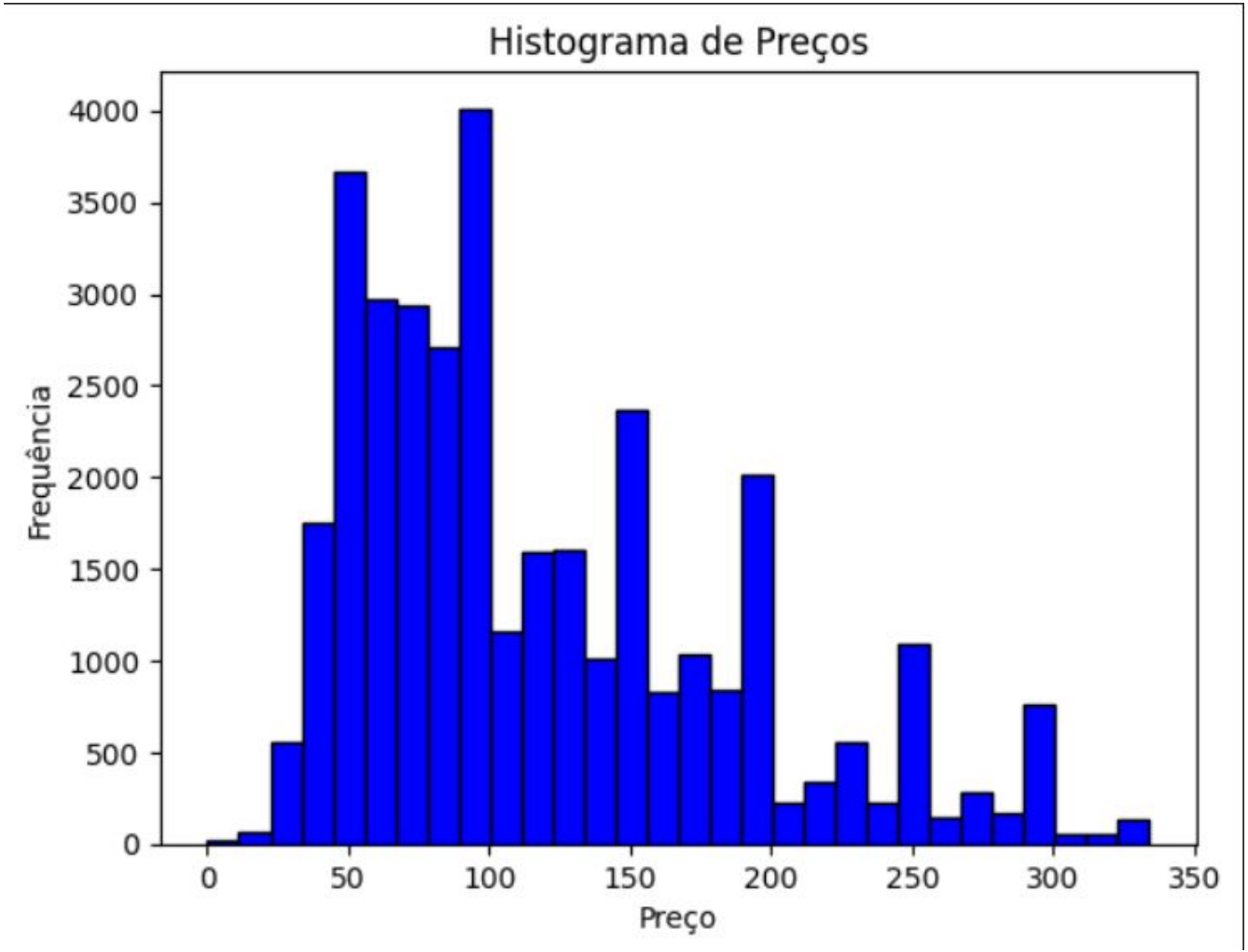








0 são os outros bairros, 1 é o bairro da label do eixo X. Nesse caso, os gráficos (violin plot) são interessantes para observar que os maiores preços concentram-se na região de Manhattan.



Aqui é um histograma de preços x frequência (número de apartamentos).
Ao analisá-lo, percebe-se que existe uma grande quantidade de apartamentos em certos intervalos de preço e logo decai drasticamente.
Como a discrepância da grande quantidade de apartamentos no intervalo $90 - $100 e pouquíssimos apartamentos no intervalo $100-$110.

2ª questão

Gabriela Brito

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

Neste caso, o bairro de Manhattan seria o mais indicado pra compra, já que o preço médio de aluguel lá é mais alto. Há ainda outros pontos que poderiam ser levados em consideração, como analisar a disponibilidade e o número de reviews (na matriz de correlação já foi apresentada uma relação entre eles), mas seria interessante analisar se impactaria na rentabilidade do negócio.

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

A correlação do número mínimo de noites resultou em 0.05 e da disponibilidade x preço foi 0.03. Ambas com valores baixos, indicando que não se relacionam como acontece com a relação dos bairros e tipo de espaço com o preço.

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

9150	Furnished room in Astoria apartment	Queens
29237	1-BR Lincoln Center	Manhattan
17691	Luxury 1 bedroom apt. -stunning Manhattan views	Brooklyn
12341	Quiet, Clean, Lit @ LES & Chinatown	Manhattan
40432	2br - The Heart of NYC: Manhattans Lower East ...	Manhattan
6529	Spanish Harlem Apt	Manhattan
30267	Beautiful/Spacious 1 bed luxury flat-TriBeCa/Soho	Manhattan
4376	Film Location	Brooklyn
29661	East 72nd Townhouse by (Hidden by Airbnb)	Manhattan
45665	Gem of east Flatbush	Brooklyn

Considerando esses os 10 apartamentos mais caros da lista, talvez o único padrão textual seja a conformidade gramatical do uso de maiúsculas e minúsculas, visto que se encontra, ao longo da lista, nomes como `MARTIAL LOFT 3: REDEMPTION (upstairs, 2nd room)` (sendo este o apto mais barato). Mas nada que seja discrepante e/ou de relevância.

3ª questão

Gabriela Brito

a. **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?**

Estamos nos deparando com um problema de REGRESSÃO dado que é um problema de previsão de preços, em que o output é um valor numérico.

PRÉ PROCESSAMENTO:
Para esse tipo de problema, iniciei com o pré-processamento dos dados removendo colunas que classifiquei como irrelevantes, como ‘id’, ‘nome’ ‘host_id’, ‘latitude’ e ‘longitude’. Latitude e longitude poderiam ter sido usadas com relevância, mas é uma redundância a coluna de bairros_group, que foi a que decidi utilizar. Além da intuição e remoção de redundância, também utilizei a função nunique() pra saber as colunas que possuíam quase que integralmente valores únicos e por isso não faziam sentido na previsão.

Também precisei remover outliers, visto que dentre a grande maioria das colunas escolhidas para o treinamento do modelo (preço, disponibilidade, minimo_noites…) apresentavam outliers (tanto min quanto max) e precisaram ser retirados pra melhor desempenho. Pra isso, foi utilizado o método IQR.

Além disso, as duas principais colunas ‘room_type’ e ‘bairro_group’ eram colunas de valores categóricos e, por esse motivo, precisei transformá-las com o One-Hot Encoder com get_dummies()

a. **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?**

DIVISÃO E TREINAMENTO:

Os dados foram divididos em treino (80%) e teste (20%), assumi valores padrões dado o tamanho do dataset apresentado.

Além disso, utilizei como hiperparâmetro a biblioteca GridSearchCV que, em ultima instancia, faz multiplas combinações pra encontrar os melhores valores pra os hiperparâmetros do modelo que, no caso, foi o Random Forest.

AVALIAÇÃO:
o modelo conseguiu prever para os dados de teste e as métricas utilizadas foram MSE, MAE e R^2. O score não foi o suficiente, visto que um modelo bem treinado tem um score ≥ 0.6 e, no meu caso, deu 0.47. Estava mais baixo, fiz alguns ajustes e consegui melhorar, mas ainda assim não tive tempo de torná-lo ainda melhor como desejaria.
Ainda, acredito que algumas outras features (faltantes no dataset) poderiam ser bem interessantes pra análise, como a área. Talvez, a criação de novas features, baseadas nas já existentes seja interessante.

4ª questão

Gabriela Brito

a. 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?

Testando no modelo que implementei, o preço previsto para o apartamento com essas descrições foi de $202.39. (código está no notebook do treinamento do modelo)