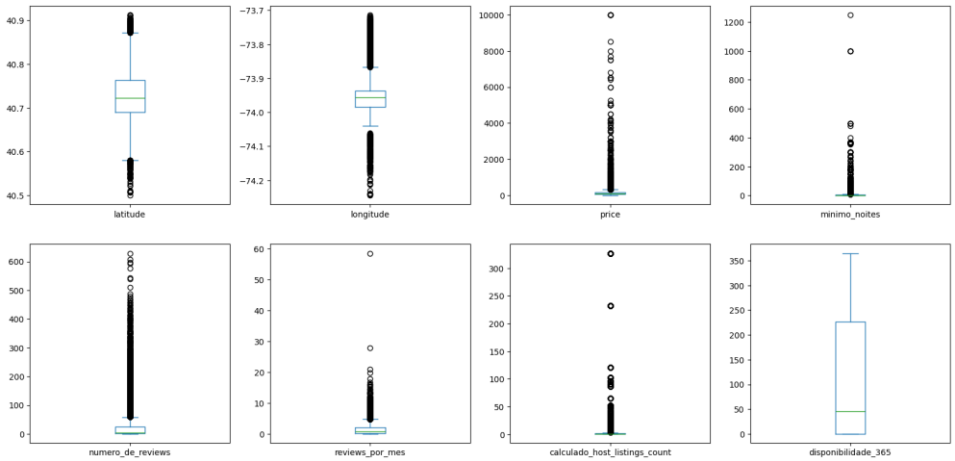
Relatório Indicium Desafio Cientista de Dados

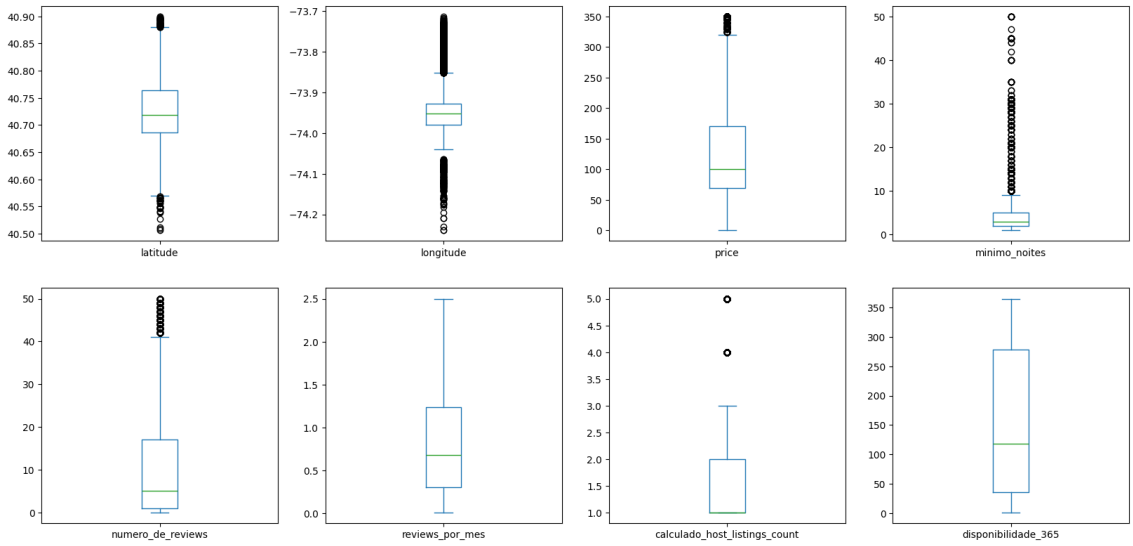
Esse é um relatório que faz parte do desafio proposto pelo programa LightHouse, onde tinha como objetivo de **“desenvolver um modelo de previsão de preços a partir do *dataset* oferecido, e avaliar tal modelo utilizando as métricas de avaliação que mais fazem sentido para o problema”. Para isso, foi feito um notebook que u**tiliza bibliotecas como **pandas, seaborn, sklearn e matplotlib** para análise de dados e modelagem. Com ele foi capaz observar alguns pontos no arquivo sobre os preços de casas para alugar em Nova Yorque. O relatório será dividido em etapas para facilitar a compreensão do que foi feito e suas análises a partir disso.

1. Análise de Dados

- O dataframe foi carregado a partir do arquivo “teste_indicium_precificacao .csv” e mostrado para ter uma primeira visualização sobre os dados que seriam trabalhados. Tal dataframe continha 48894 linhas e 16 colunas, sendo separado pelas colunas **'id', 'nome', 'host_id', 'host_name', 'bairro_group', 'bairro', 'latitude', 'longitude', 'room_type', 'price', 'minimo_noites', 'numero_de_reviews', 'ultima_review', 'reviews_por_mes', 'calculado_host_listings_count' e 'disponibilidade_365'.**
- Em primeiro momento algumas colunas foram removidas, como **id, host_id, nome** e ***host_name***, por serem irrelevantes para a previsão de preços nesse primeiro momento.
- Foram gerados boxplots para visualizar a distribuição dos dados e poder visualizar melhor a distribuição dos dados e seus outliers.
- A seguir, os gráficos gerados pelo Boxplot:



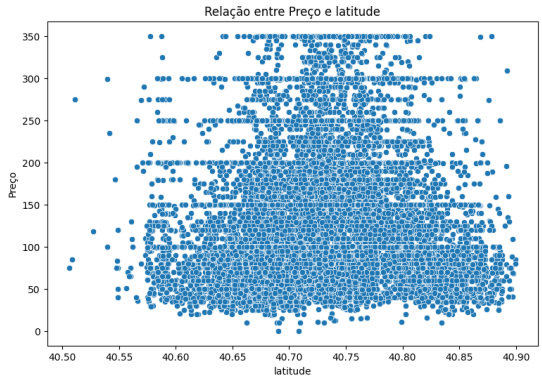
- Boa parte dos outliers foram removidos para variáveis como **preço**, **número de reviews**, **mínimo de noites**, **reviews por mês** e **calculado host listings count**. Não foram removidos TODOS os outliers para que o modelo e visualização de dados não ficasse tão restringida.
- Foi utilizado preço abaixo de 350, número de reviews acima de 50, mínimo de noite acima de 50, reviews por mês acima de 2.5, calculado listings Count acima de 5, e casas que não tinha disponibilidade de dias.Com isso perdemos um pouco de informação, mas garantimos uma análise mais detalhada e um modelo de previsão melhor.
- Dados de longitude foram mantidos entre 40.5 e 40.9 enquanto os de latitude foram mantidos entre -74.24 e -74.7 devido aos limites da cidade. Os outliers de longitude e latitude foram mantidos para não concentrar apenas em imóveis próximos ao centro.
- Com isso tivemos uma nova distribuição que já permite uma melhor visualização dos dados.

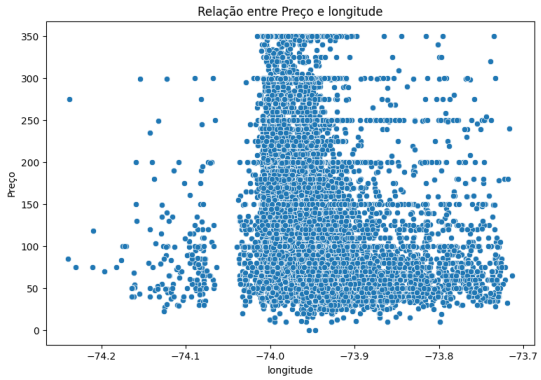


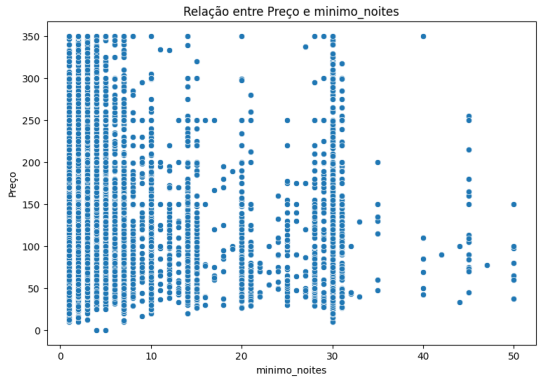
- Para apresentar algumas hipóteses que levam a precificação foram utilizados gráficos do tipo scatterplot e kdeplot.
- Nos gráficos de scatterplot apesar de podermos ter uma visualização geral da relação de preços ele ainda fica muito poluído devido ao alto número de pontos.
- Portanto pra uma melhor visualização o kdeplot é a melhor opção já que oferece um heatmap e um gráfico mais limpo.

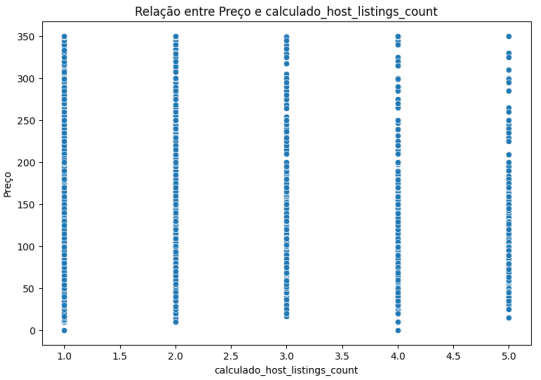
2. Gráficos para a análise exploratória

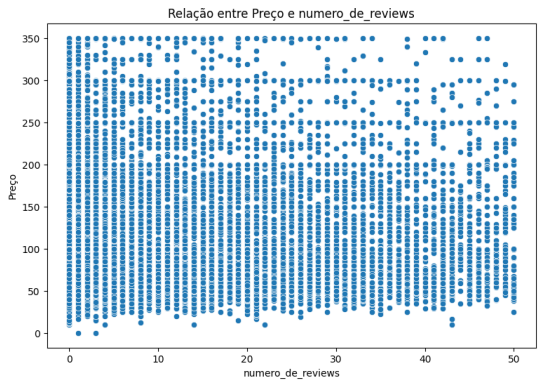
- **Scatterplot:**

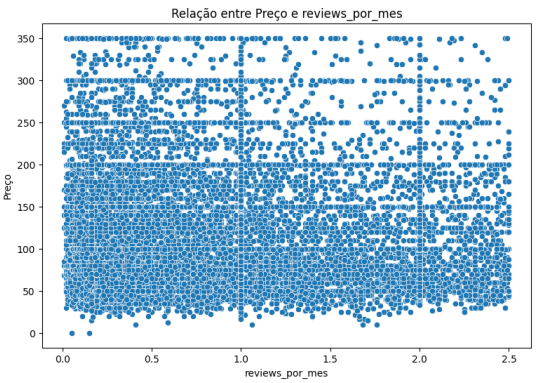


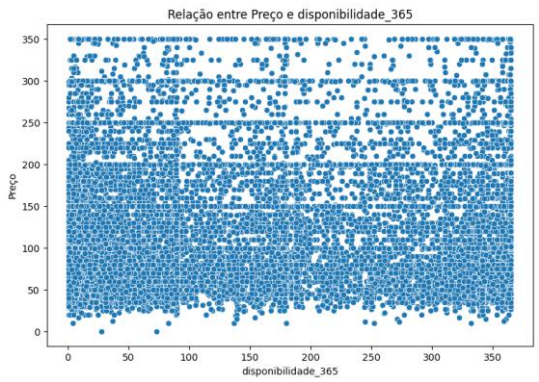




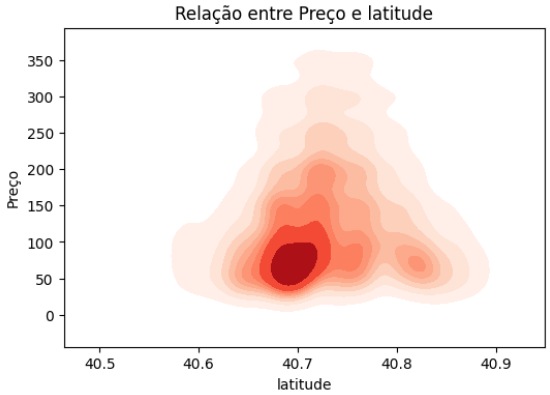


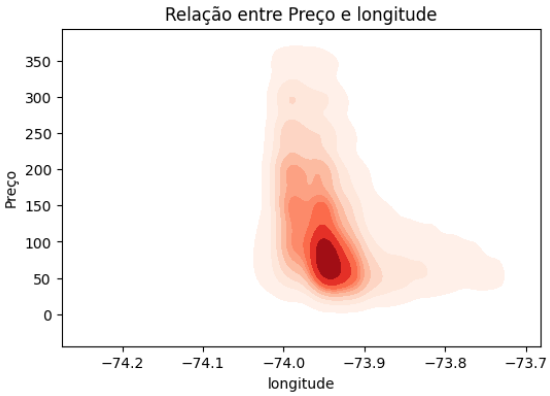


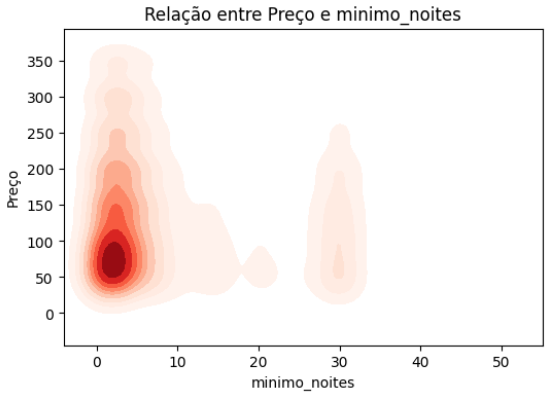


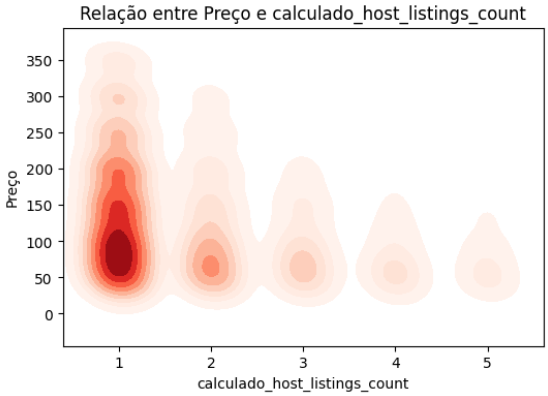


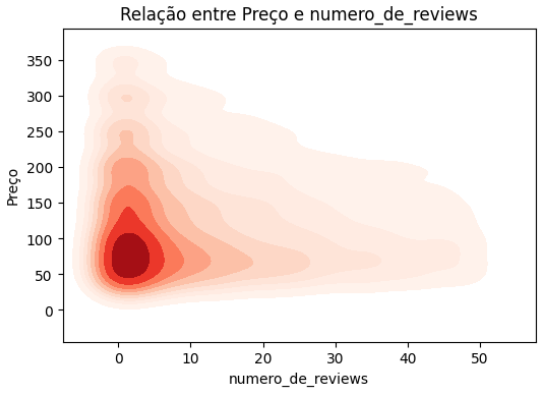
- **Kdeplot**

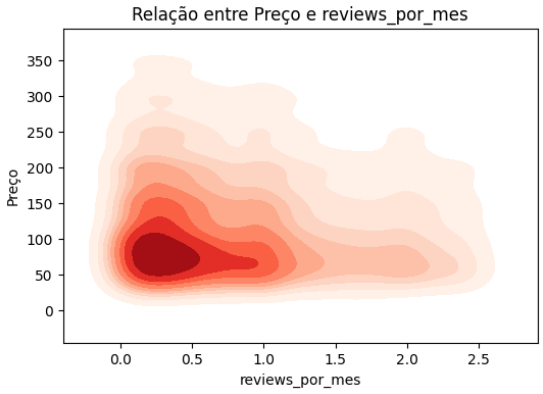

Relação entre Preço e latitude
Preço
latitude
40.5
40.6
40.7
40.8
40.9

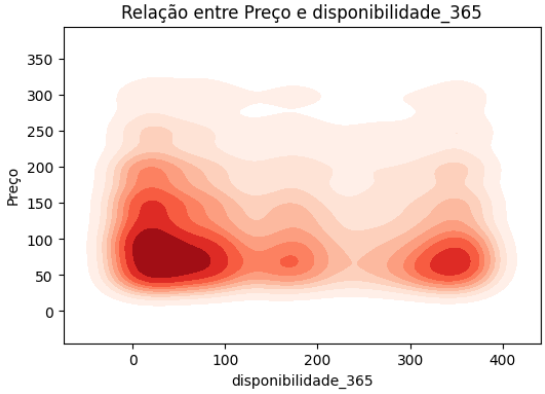

Relação entre Preço e longitude
Preço
longitude
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7


Relação entre Preço e minimo_noites
Preço
minimo_noites


Relação entre Preço e calculado_host_listings_count
Preço
calculado_host_listings_count

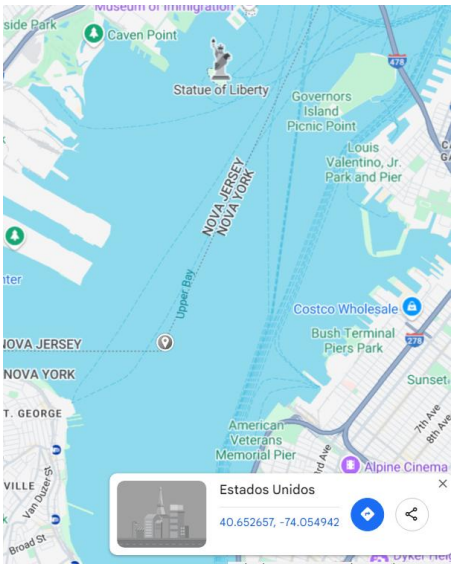

Relação entre Preço e numero_de_reviews
Preço
numero_de_reviews


Relação entre Preço e reviews_por_mes
Preço
reviews_por_mes


Relação entre Preço e disponibilidade_365
Preço
disponibilidade_365

3. Levantamento de hipóteses

- Vemos que as casas de maiores valores estão localizadas mais ao centro da cidade, assim como uma grande quantidade de casas que se localizam mais ao centro.
- O fato de não ter casas localizadas na longitude -74.05 se deve ao fato de ser localizado no Upper Bay, ou seja, no mar aberto.



- Há uma grande concentração de casas em que o mínimo de noites se concentra entre 0 e 10 dias ou de 30 dias, o que mostra que os donos das casas preferem por estadias menores ou de alugueis sejam de no mínimo 1 mês. Pode-se inferir que valores muito altos de **mínimo de noites** podem afastar hóspedes ocasionais, reduzindo o valor do imóvel.

- Vemos que em relação ao **Count_list_hosts** mostra que os donos de vários imóveis tendem a ter locais de preços mais inferiores, enquanto os de preços mais altos são poucos donos que possuem várias casas nessa faixa de preço.

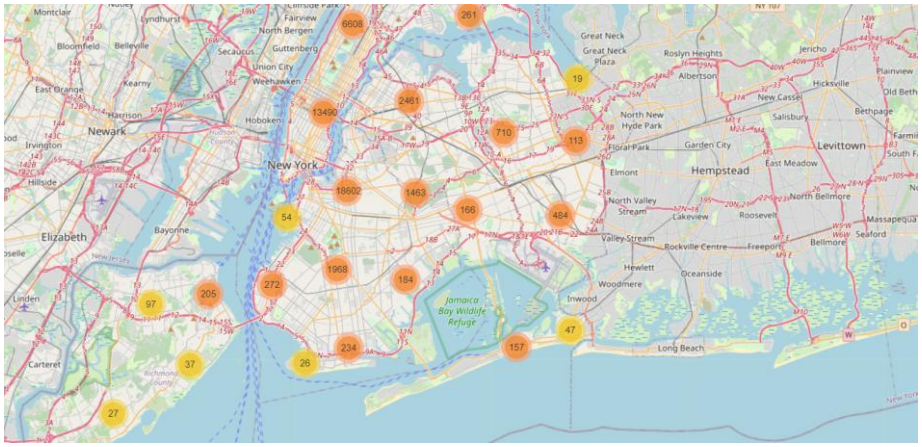
- Em relação ao número de reviews pode-se dizer que casas de altos valores tem menores números de reviews e esses dados junto com o numero de reviews por mês indica baixa rotatividade desses lugares, ou seja, o alto preço afasta novos hóspedes.

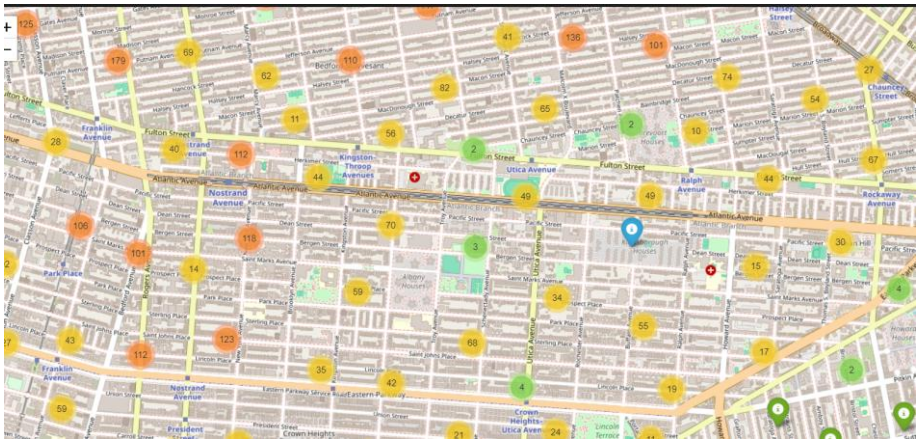
- A **disponibilidade ao longo do ano** pode indicar a frequência de ocupação, impactando a precificação.

4. Respostas para as perguntas

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

- Pode-se analisar a relação entre bairros e preços médios para identificar os mais rentáveis. Se o objetivo é retorno financeiro, é interessante escolher locais com alta demanda e preços mais elevados, mas sem um valor inicial muito alto.
- Para isso, no código foi criado um mapa utilizando o **Folium** que mostra os bairros a relação de preços e localizações. Onde temos um gradiente de pontos sendo quanto mais verde mais barato e mais vermelho seriam os mais caros.



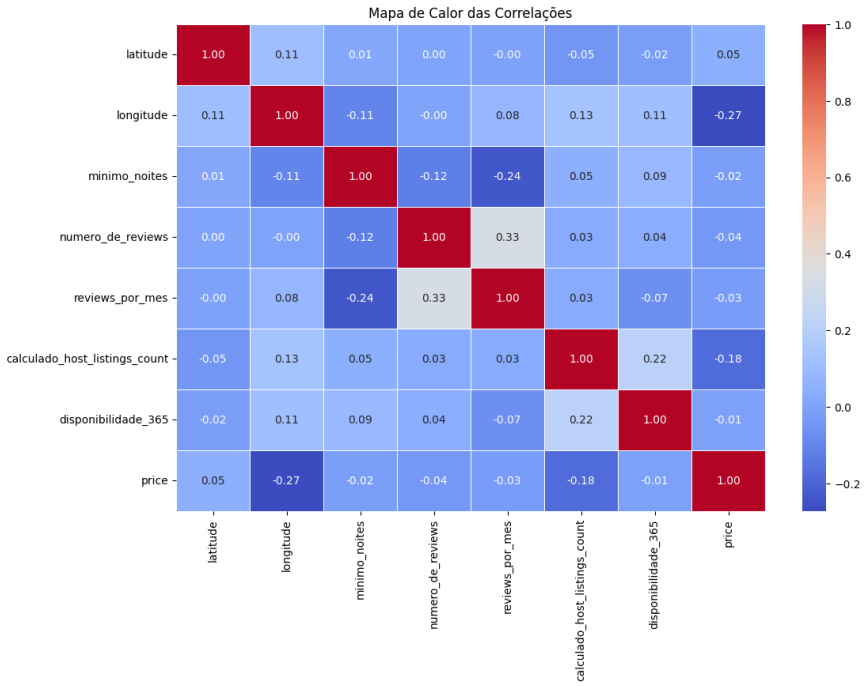


- Áreas de pontos **vermelhos** indicam regiões onde os preços são mais altos.
- Se o objetivo for alto retorno de investimento, buscar áreas com poucos imóveis disponíveis e preços elevados pode ser estratégico.
- Se o objetivo for alta ocupação, regiões com imóveis mais acessíveis (azul ou verde) podem ser mais vantajosas.

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

- Sim, como mostrado na análise inicial, foram removidos valores extremos dessas variáveis. Pode-se inferir que valores muito altos de mínimo de noites podem afastar hóspedes ocasionais, reduzindo o valor do imóvel. A disponibilidade ao longo do ano pode indicar a frequência de ocupação, impactando a precificação.

- Usando a função correlation, podemos ver essa correlação entre preço e as variáveis entre número mínimo de noites e a disponibilidade. Mostramos que quanto menor o número de noites maior o preço assim como a disponibilidade de dias, já que ambos tem um correlação negativa, ou seja, inversamente proporcionais.



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

- Filtrando o dataframe para pegar as 10% de casas mais caras para alugar podemos ver as seguintes palavras que são mais repetidas:

nome	
in	1094
2	675
Bedroom	389
Luxury	350
Village	319
with	317
Apartment	313
of	305
3	294
Loft	292
Park	281
w/	278
Manhattan	273
&	262
East	262
+	258
the	258
Apt	256
by	256

- Com isso vemos que as palavras bedroom, Luxury, village, apartment, Loft, Manhattan e East são palavras chaves muito utilizadas nesses tipos de casas. Isso nos informa que os imóveis mais caros tendem a enfatizar luxo, localização e o tipo de acomodação no nome. Quem deseja atrair um público disposto a pagar mais pode usar esse padrão para nomear seus anúncios na plataforma

5. Modelo Utilizado

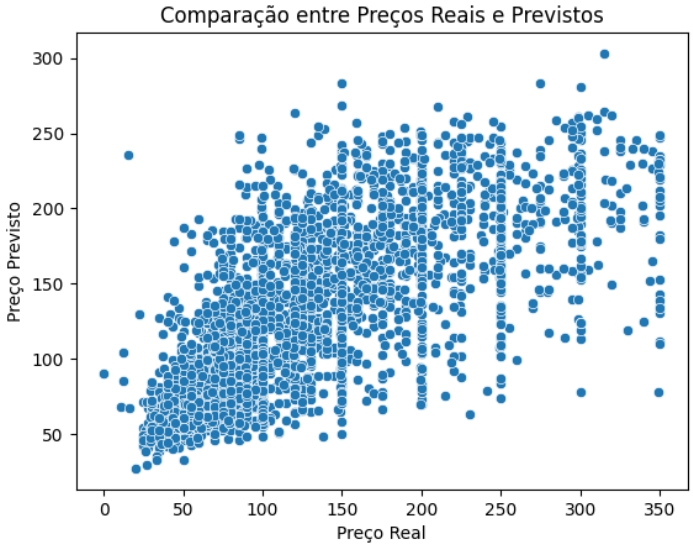
- Para fazer a previsão de preço utilizamos o método de Machine Learning com um modelo que melhor ser adapte aos dados.
- No modelo de **Machine Learning** utilizado, as variáveis foram selecionadas com base na sua relevância para a previsão de preços de aluguel na plataforma. Sendo esses:
    - **Latitude e longitude:** A localização é um dos fatores mais importantes para a precificação de imóveis. Regiões mais valorizadas geralmente possuem preços mais altos.
    - **room_type:** Tipo de acomodação (quarto inteiro, compartilhado ou imóvel inteiro) influencia diretamente no preço. Imóveis inteiros tendem a ser mais caros do que quartos privativos.
    - **minimum_nights:** Quanto maior o número mínimo de noites exigido, menor pode ser a flexibilidade para hóspedes, impactando a demanda e o preço.
    - **number_of_reviews**: Imóveis com muitas avaliações geralmente são mais populares e podem ter preços mais competitivos.
    - **reviews_per_month:** Indica a frequência com que o imóvel recebe hóspedes. Um imóvel com alta ocupação pode ter um preço mais ajustado à demanda.
    - **calculated_host_listings_count**: Mostra quantos imóveis o anfitrião possui. Anfitriões com muitos imóveis podem ter estratégias diferentes de precificação.
    - **availability_365**: Mede quantos dias o imóvel está disponível no ano. Imóveis muito disponíveis podem ter preços mais baixos para garantir ocupação.

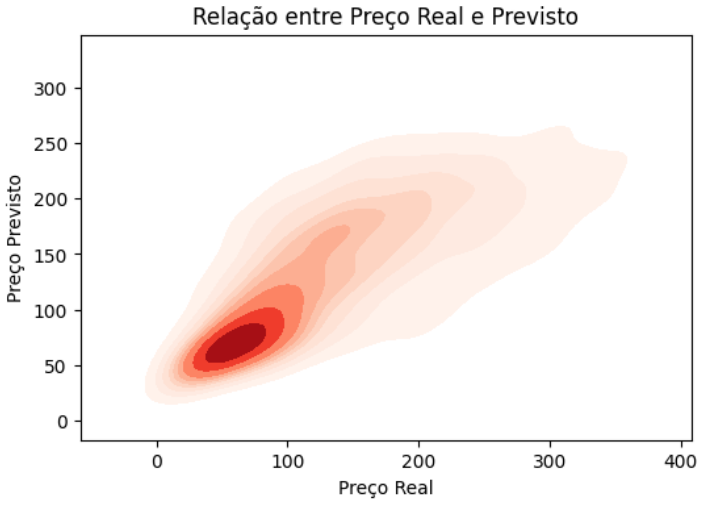
- Por que outras variáveis foram removidas?
    - **ID do imóvel e do anfitrião**: Não afetam diretamente o preço.
    - **Nome do imóvel**: Embora pudesse indicar luxo, foi removido pois não é uma variável numérica.
    - **Host_name**: Informação do anfitrião não impacta a precificação.
- O modelo escolhido foi o **RandomForestRegressor**, da biblioteca Scikit-learn.
- Foi utilizado um pipeline que inclui pré-processamento dos dados:
    - **Imputação de valores ausentes**
    - **OneHotEncoding** para variáveis categóricas
    - **StandardScaler** para variáveis numéricas
- O dataset foi dividido em treino e teste.
- A performance do modelo foi validada com **cross-validation**.

- **Justificativas para o Modelo:**
    - O **RandomForestRegressor** foi escolhido por ser um modelo robusto para problemas de regressão com muitos dados.
    - Ele lida bem com dados categóricos e numéricos, além de ser menos sensível a outliers do que modelos lineares.
    - Permite captar relações complexas entre variáveis e captura interações entre elas.
- Resultados do modelo:
    - O modelo apresentou um bom resultado de início para uma faixa de preço de até 350 doláres. Podemos ver isso a partir das métricas de MAE (mean Absolute error), MSE (mean_squared_error), RMSE (Root mean squared error) e R^2
    - O modelo mostrou um erro absoluto em relação à média foi de 36,21 dólares e o RMSE se matendo com o mesmo numero de casas que o MAE. Além de que um R^2 acima de 50% para tal modelo demonstra que o modelo resultou bem, podendo ser melhorado mais pra frente com uma melhor filtragem dos dados e escolhendo faixas de preços específicas.

```
MAE: 36.212637803480256
MSE: 2547.932153934389
RMSE: 50.477045812273815
R²: 0.537086714285117
```

- Também se fez-se gráficos que mostra a relação entre preços reais e preços previstos pelo modelo





- Com esses gráficos podemos visualizar a tendência entre o preço real e o preço previsto

6. Previsão de um novo apartamento

- 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?
  - De acordo com o modelo teríamos um apartamento no valor de Preço sugerido: $217.96

7.Conclusão

Com essa análise concluímos que o modelo funciona bem, ainda que com suas limitações. Outras variáveis que podem ser implementadas futuramente para a melhoria do modelo seriam as de taxas de criminalidade por localização, número de farmácias, escolas e hospitais ao redor, número de estações policiais por bairro e custo de vida. Pois, essas informações também são cruciais para definir o nível de conforto que um morador terá e como isso influencia nos preços das casas.