Introdução

Este relatório apresenta a solução para o desafio de ciência de dados proposto pela Indicium. O objetivo é realizar uma análise exploratória dos dados (EDA) de aluguéis temporários na cidade de Nova York e desenvolver um modelo preditivo para prever o preço dos imóveis. O desafio também inclui a resposta a perguntas específicas relacionadas ao negócio, como a escolha do local ideal para investimento e a influência de variáveis como o número mínimo de noites e a disponibilidade ao longo do ano no preço dos imóveis.

Análise exploratória dos dados(EDA)

Descrição do dataset

O dataset contém 48.894 linhas e 16 colunas, incluindo informações como ID, nome, host ID, host nome, bairro, área do local, latitude, longitude, tipo de quarto, Preço em dólares, Mínimo de noites, número de avaliação, data da última avaliação, avaliações por mês, quantidade total de imóveis que um host possui, número de dias em que o anúncio está disponível para reserva.

Tratamento de dados nulos

Antes de iniciar a análise, foi identificado que algumas colunas continham valores nulos:

- **Nome do anúncio, Nome do anfitrião, Última avaliação e Média de avaliações por mês.**

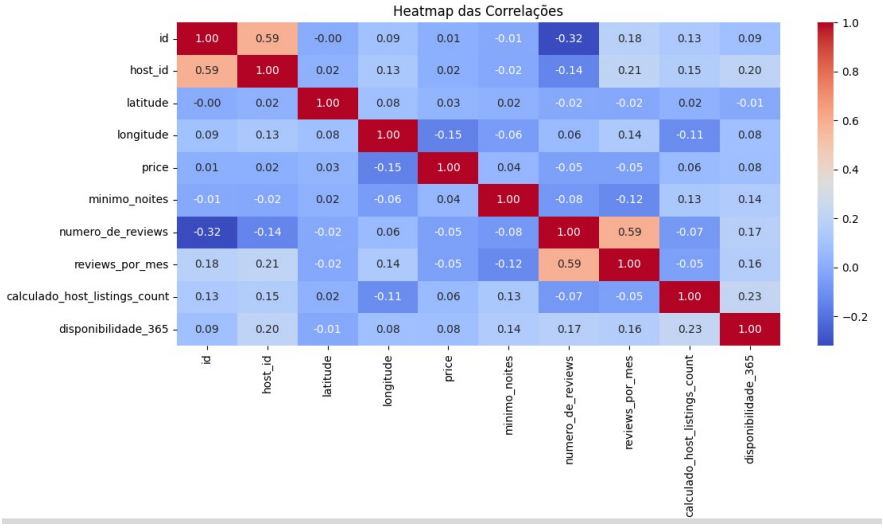
Para garantir uma análise mais consistente, esses valores foram preenchidos da seguinte forma:

- **Nome do anúncio** → "Sem nome"
- **Nome do anfitrião** → "Indefinido"
- **Última avaliação** e **Média de avaliações por mês** → 0

Esse tratamento permitiu manter todas as informações no dataset sem a necessidade de remover linhas, garantindo uma análise mais completa e precisa.

Correlação entre as colunas

A primeira etapa da análise foi calcular as correlações entre as variáveis. O resultado obtido foi o seguinte:



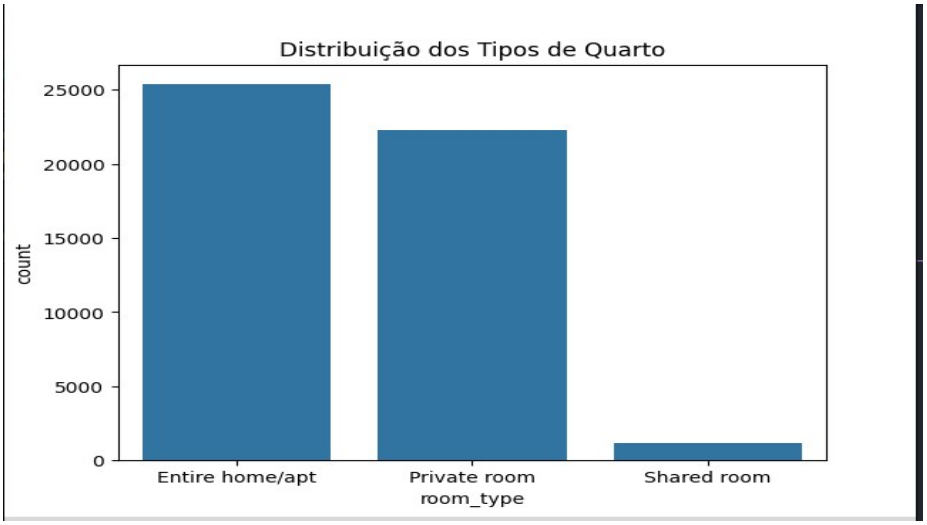
Fonte: Gráficos gerados em Python com *matplotlib* e *seaborn*, utilizando dados do arquivo CSV fornecido.

Observa-se que o número total de avaliações e as avaliações por mês possuem uma correlação positiva significativa (**0.59**), indicando que anúncios mais avaliados tendem a continuar recebendo novas avaliações com frequência.

Além disso, a quantidade de imóveis por host apresenta uma leve correlação com a disponibilidade anual (**0.23**), sugerindo que hosts com mais propriedades tendem a manter seus anúncios ativos por mais tempo. Já o preço não apresenta correlações fortes com nenhuma variável, indicando que outros fatores não incluídos no dataset podem influenciar seu valor.

Verificação dos tipos de quartos

Após a análise de correlação, examinei a distribuição dos tipos de quartos em todos os imóveis, bem como a quantidade de cada um deles, O gráfico abaixo ilustra esses padrões.



Fonte: Gráfico gerado em Python com *matplotlib* e *seaborn*, utilizando dados do arquivo CSV fornecido.

O gráfico apresenta a distribuição dos tipos de acomodação disponíveis na base de dados Observa-se que a maioria dos anúncios são de casas/apartamentos inteiros, seguidos por quartos privados. Já os quartos compartilhados representam uma parcela muito pequena das listagens.

Isso sugere que a preferência dos anfitriões e hóspedes está voltada para acomodações que oferecem maior privacidade, enquanto opções compartilhadas são menos comuns na plataforma.

Dados estatísticos

Além disso, realizei análises estatísticas da coluna de preços. Os resultados estão ilustrados na tabela abaixo.

```
Métricas individuais:
Média: 152.72
Mediana: 106.00
Moda: [100]
Desvio Padrão: 240.16
Variância: 57675.20
Valor Mínimo: 0
Valor Máximo: 10000
Percentil 25%: 69.00
Percentil 75%: 175.00
```

Fonte: Tabela gerada em Python com *pandas*, utilizando dados do arquivo CSV fornecido.

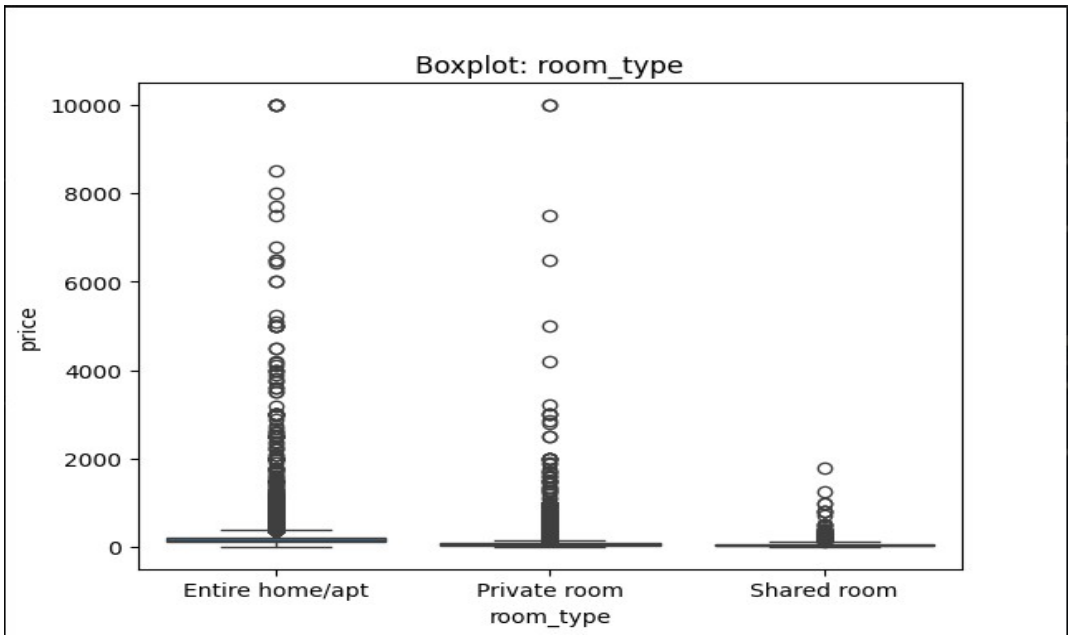
Os dados apresentam uma média de 152,72, mas a mediana de 106 indica que a distribuição pode estar assimétrica, possivelmente enviesada por valores altos. A moda de 100 sugere que esse é o valor mais frequente na amostra.

O desvio padrão elevado (240,16) e a variância alta (57.675,20) indicam que há uma grande dispersão nos valores. Além disso, a presença de um valor máximo de 10.000 e um mínimo de 0 confirma a existência de outliers.

Os percentis (25%: 69,00 e 75%: 175,00) mostram que a maior parte dos dados está concentrada dentro desse intervalo, mas com uma cauda longa à direita, reforçando a possível assimetria.

Supondo algumas hipóteses

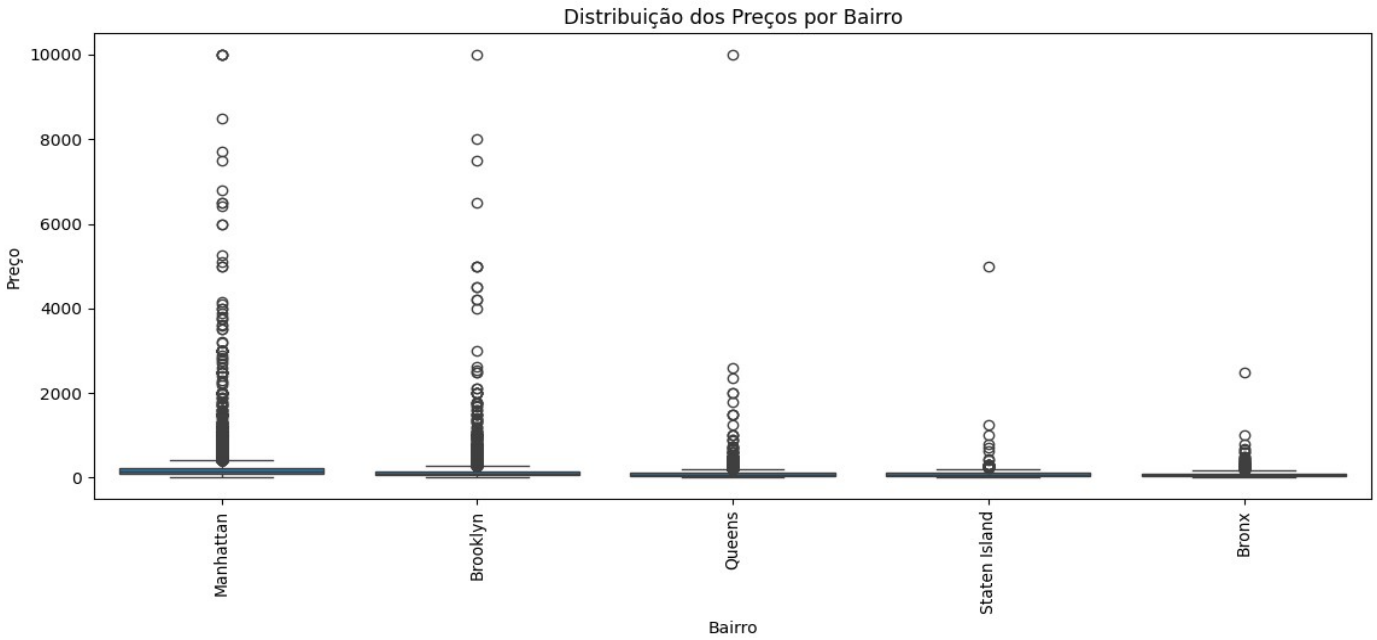
Após refletir sobre a pergunta "Casas, apartamentos e quartos privados tendem a ser mais caros do que quartos compartilhados?", cheguei a uma conclusão com base na análise do gráfico boxplot abaixo.



Fonte: Gráfico gerado em Python com *matplotlib* e *seaborn*, utilizando dados do arquivo CSV fornecido.

Com base na análise do gráfico boxplot, podemos concluir que, de maneira geral, as opções de acomodação privada, como casas, apartamentos e quartos privados, apresentam preços mais elevados em comparação com quartos compartilhados. A distribuição dos valores mostra uma concentração de preços mais altos para as acomodações privadas e uma variação de preços significativamente menor para os quartos compartilhados. Essa tendência reforça a ideia de que as opções privadas tendem a ser mais caras, conforme evidenciado pelos dados apresentados.

Também investiguei se a localização influencia no preço.



Fonte: Gráfico gerado em Python com *matplotlib* e *seaborn*, utilizando dados do arquivo CSV fornecido.

A análise do gráfico "Distribuição dos Preços por Bairro" evidencia uma variação significativa nos preços das acomodações conforme a localização. Manhattan e Brooklyn registram os valores mais altos, enquanto Queens, Staten Island e Bronx oferecem opções mais acessíveis. Isso reforça que a localização é um fator determinante na precificação dos imóveis

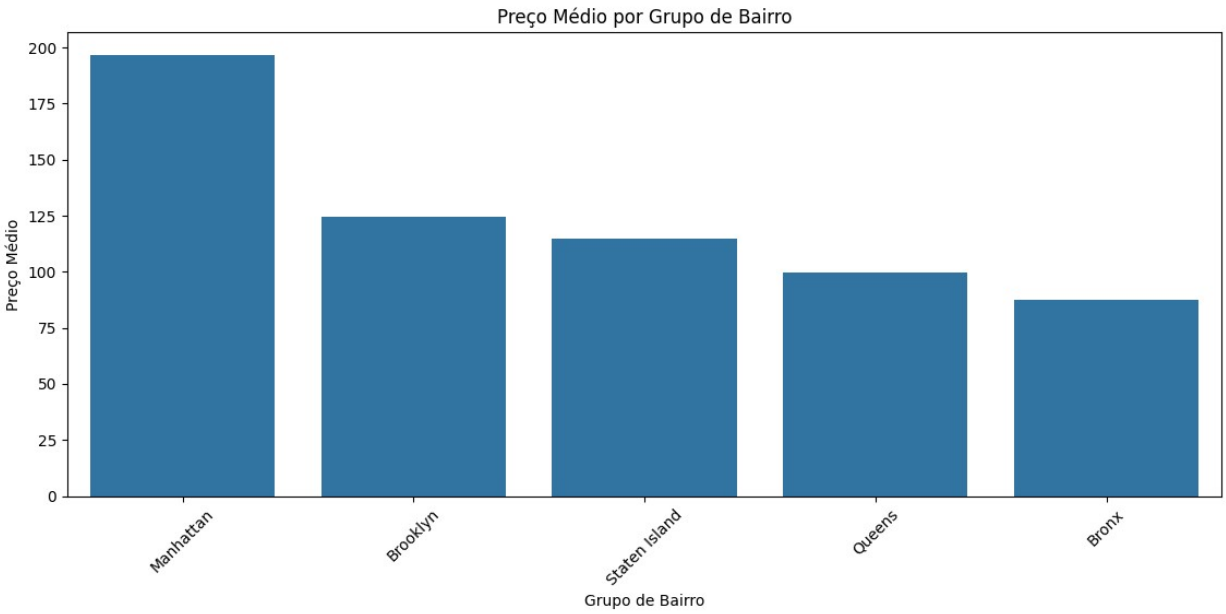
Análise de investimento e fatores que influenciam o preço

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

Para quem busca investir em regiões com custos mais baixos, o Bronx, com preço médio de $87,50, e o Queens, com média de $99,50, surgem como opções atrativas. Ambos oferecem valores mais acessíveis em comparação a outras áreas, representando boas oportunidades de custo-benefício. Essa tendência pode ser observada no gráfico e na tabela a seguir.

```
Preço médio em cada bairro.
bairro_group
Manhattan        196.875814
Brooklyn         124.381983
Staten Island    114.812332
Queens            99.517649
Bronx             87.496792
Name: price, dtype: float64
```

Fonte: Tabela gerada em Python com *pandas*, utilizando dados do arquivo CSV fornecido.

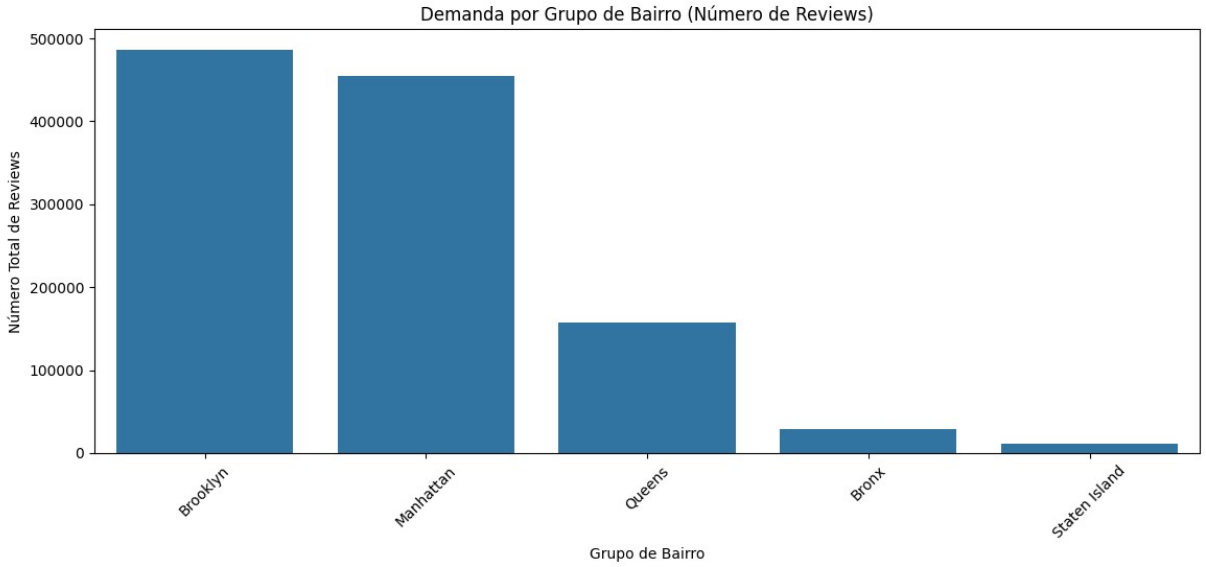


Fonte: Gráfico gerado em Python com *matplotlib* e *seaborn*, utilizando dados do arquivo CSV fornecido.

Mas, se considerarmos a demanda, ou seja, o número de avaliações, vale a pena investir no Brooklyn e Manhattan, que possuem um volume maior de avaliações. Isso indica que essas áreas têm maior procura, o que pode ser um indicativo de maior interesse e potencial de rentabilidade. Portanto, embora o custo seja mais alto, a maior demanda pode justificar o investimento nessas localidades. O gráfico e tabela a seguir ilustram essa tendência.

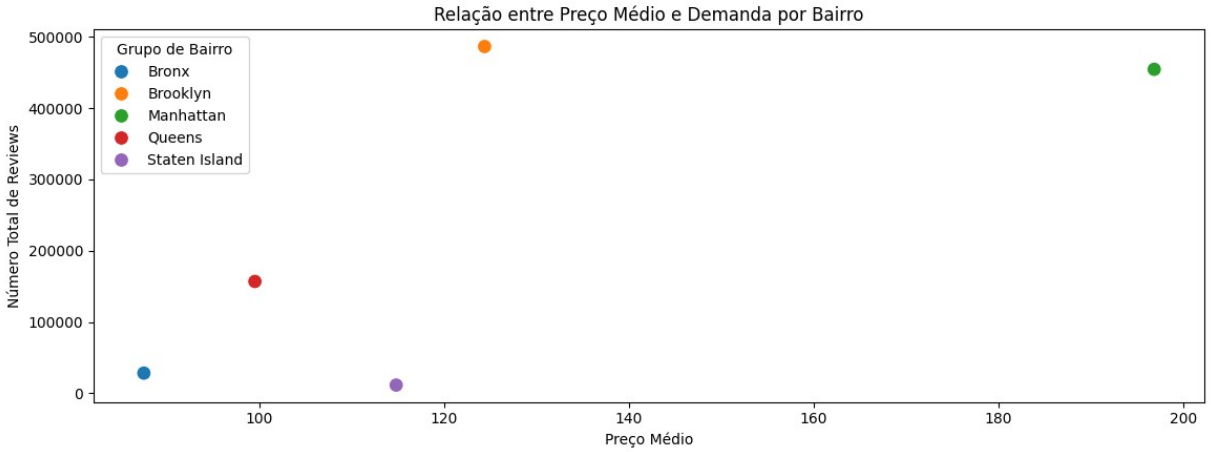
```
Número de avaliações em cada Bairro Group:
bairro_group
Brooklyn         486565
Manhattan        454569
Queens           156950
Bronx             28371
Staten Island     11541
Name: numero_de_reviews, dtype: int64
```

Fonte: Tabela gerada em Python com *pandas*, utilizando dados do arquivo CSV fornecido.



Fonte: Gráfico gerado em Python com *matplotlib* e *seaborn*, utilizando dados do arquivo CSV fornecido.

Pensando em demanda e número de avaliações, o mais indicado é investir no Brooklyn, pois, além de ter um custo relativamente acessível, apresenta um alto número de reviews, como mostrado no gráfico abaixo. Isso indica uma forte procura pela região, tornando-a uma opção atraente tanto em termos de preço quanto de popularidade.

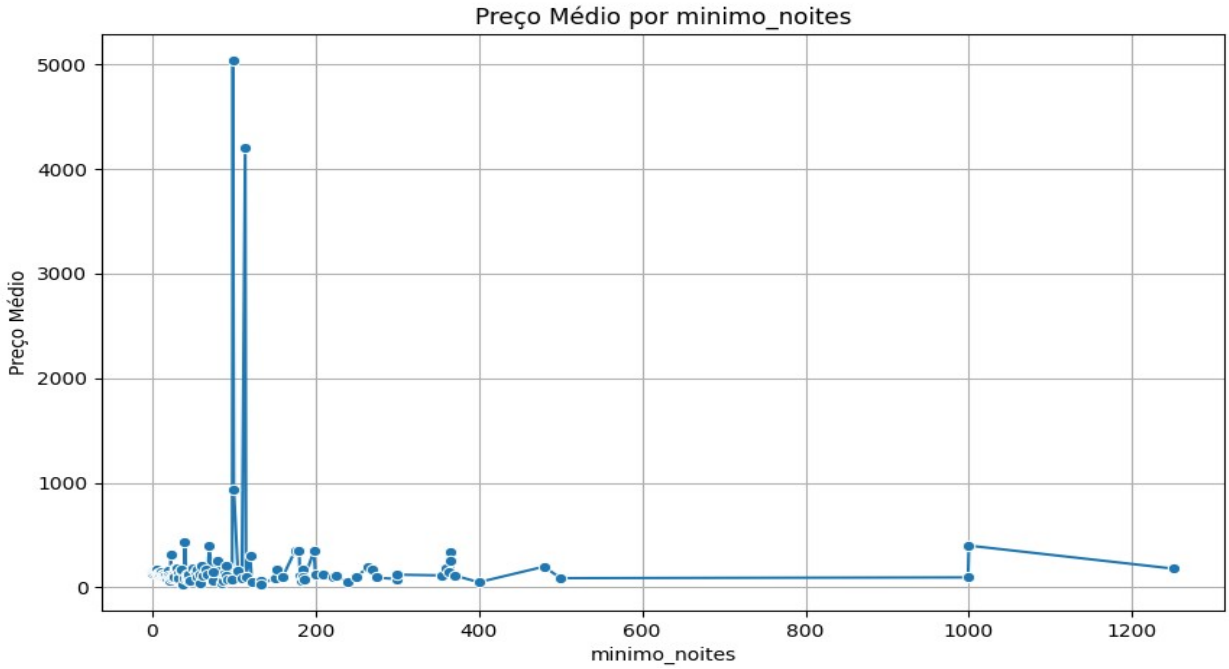


Fonte: Gráfico gerado em Python com *matplotlib* e *seaborn*, utilizando dados do arquivo CSV fornecido.

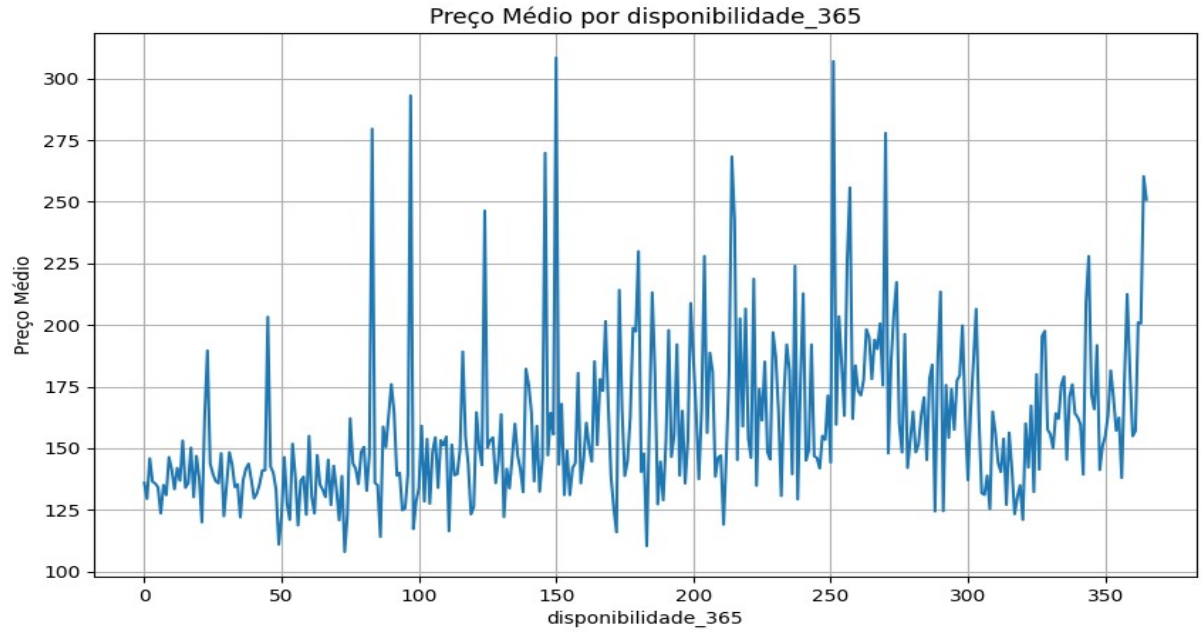
2 - O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?

O número mínimo de noites não parece ter uma influência significativa no preço. No gráfico abaixo, observa-se que, para um intervalo entre 0 e 200 noites, os preços variam amplamente, abrangendo tanto valores elevados quanto mais acessíveis, sem apresentar uma tendência clara de aumento ou redução conforme o número mínimo de noites cresce.

Da mesma forma, a disponibilidade ao longo do ano não demonstra um impacto consistente nos preços. O gráfico indica uma grande variação nos valores à medida que a disponibilidade aumenta, sem evidenciar um padrão definido. Os gráficos a seguir ilustram esse cenário.



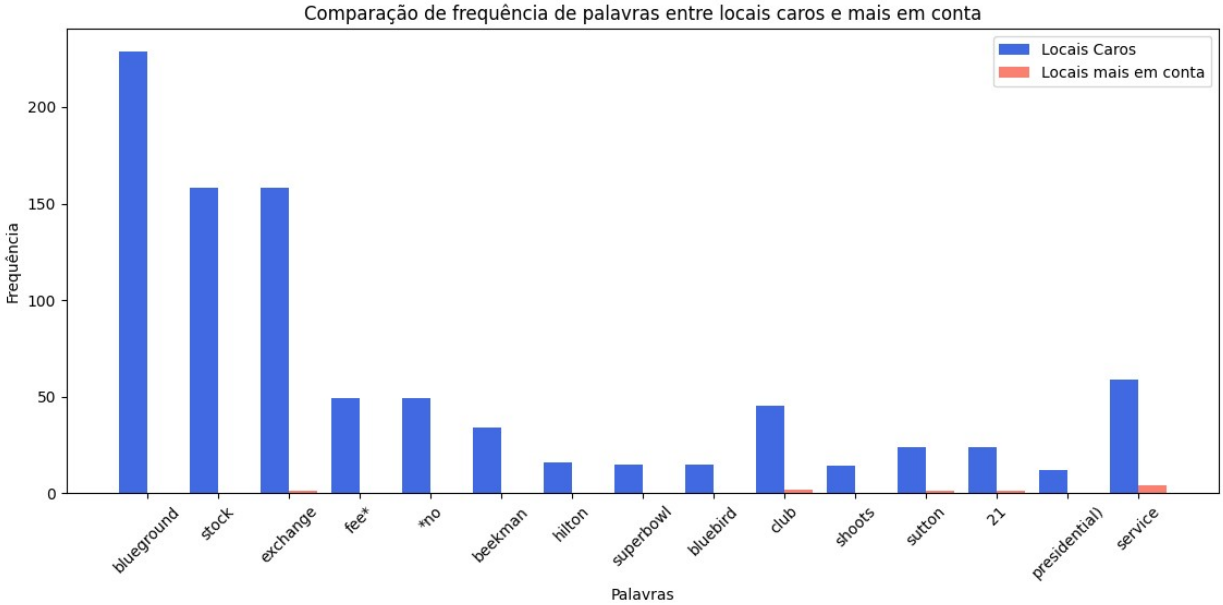
Fonte: Gráfico gerado em Python com *matplotlib* e *seaborn*, utilizando dados do arquivo CSV fornecido.



Fonte: Gráfico gerado em Python com *matplotlib* e *seaborn*, utilizando dados do arquivo CSV fornecido.

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

Sim, o gráfico abaixo demonstra que existem padrões no texto do nome do local para acomodações de maior valor, sugerindo que certos termos podem estar associados a preços mais altos.



Fonte: Gráfico gerado em Python com *matplotlib* e *seaborn*, utilizando dados do arquivo CSV fornecido.

Previsão de Preços de Imóveis: Estratégia e Modelagem

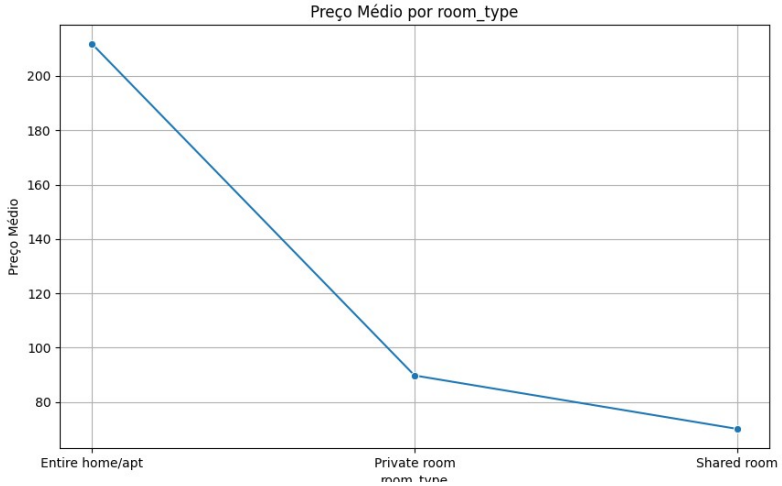
Definição do Problema

O objetivo do projeto é prever o preço de um imóvel com base em suas características. Esse tipo de problema se enquadra na categoria de regressão, pois a variável alvo (preço) é um valor contínuo, e não uma categoria discreta.

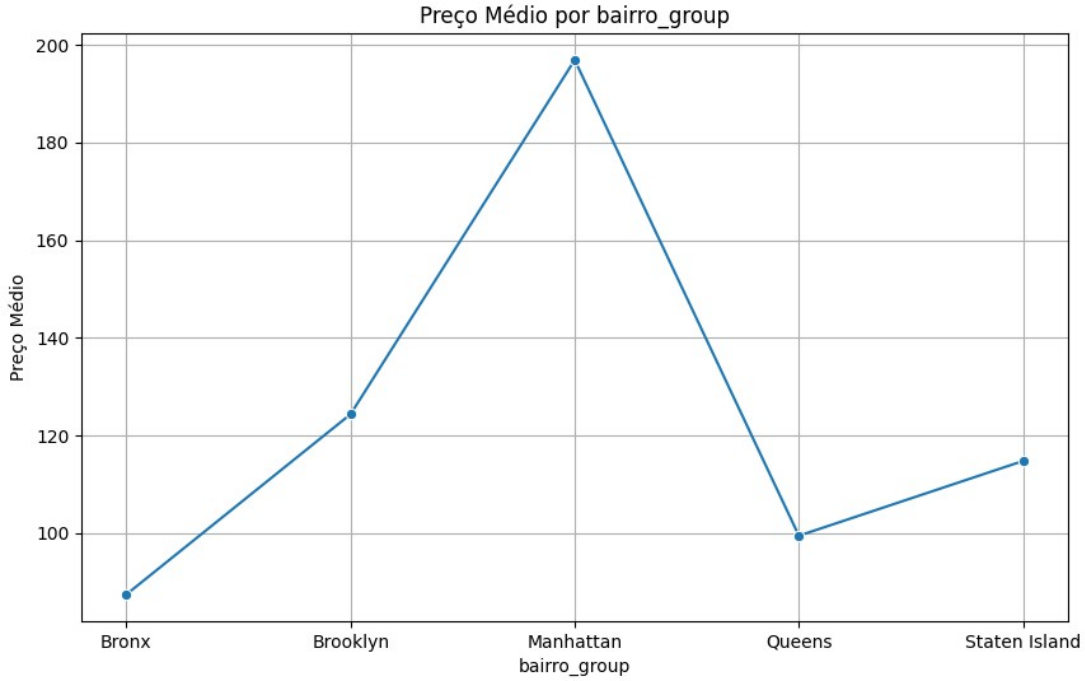
Seleção de Variáveis e Transformações

A escolha das variáveis foi baseada em uma análise exploratória de dados (EDA), onde identifiquei os fatores que mais influenciam o preço do imóvel. As principais variáveis selecionadas foram:

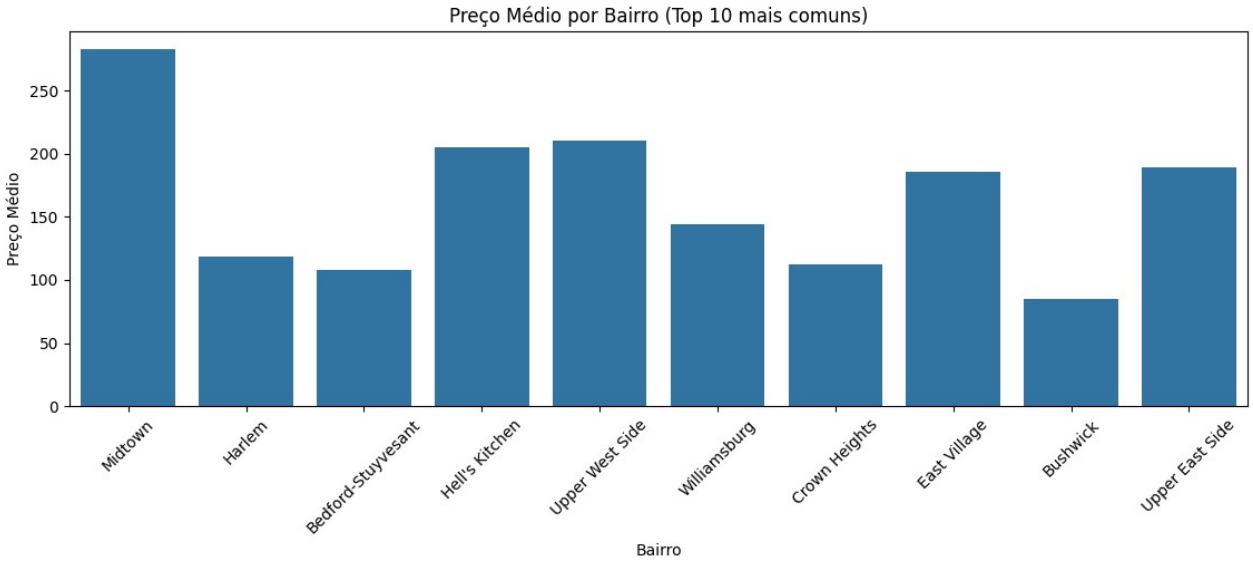
bairro_group – A localização do imóvel impacta diretamente seu valor no mercado.
room_type – O tipo de acomodação influencia significativamente o preço do aluguel.
bairro – Alguns bairros têm preços consistentemente mais altos ou mais baixos, tornando essa variável essencial.



Fonte: Gráfico gerado em Python com *matplotlib* e *seaborn*, utilizando dados do arquivo CSV fornecido.



Fonte: Gráfico gerado em Python com *matplotlib* e *seaborn*, utilizando dados do arquivo CSV fornecido.



Fonte: Gráfico gerado em Python com *matplotlib* e *seaborn*, utilizando dados do arquivo CSV fornecido.

As análises gráficas demonstraram que essas variáveis possuem forte correlação com o preço, justificando sua inclusão no modelo.

Modelo Escolhido: Regressão Linear

Para modelar a relação entre as variáveis e o preço do imóvel, utilizei a Regressão Linear, um modelo simples e interpretável.

Vantagens:

Fácil implementação e interpretação,

Treinamento rápido, mesmo com grandes volumes de dados

Os coeficientes fornecem insights sobre a influência de cada variável.

Desvantagens

Assume uma relação linear entre as variáveis independentes e a variável dependente, o que pode limitar a precisão caso existam padrões não lineares complexos.

Sensível a outliers, que podem distorcer os resultados.

Métrica de Performance: RMSE (Root Mean Squared Error)

Para avaliar o desempenho do modelo, utilizei o RMSE (Erro Quadrático Médio da Raiz), pois:
Penaliza erros maiores, tornando as previsões mais confiáveis.
Mantém a unidade original da variável de saída (preço), facilitando a interpretação.
É amplamente utilizado em problemas de regressão.

Estimativa de Preço para um Imóvel Específico

Com base no modelo desenvolvido, qual seria a previsão de preço para o seguinte apartamento?

- **ID:** 2595
- **Nome:** Skylit Midtown Castle
- **Host ID:** 2845
- **Nome do Anfitrião:** Jennifer
- **Bairro Grupo:** Manhattan
- **Bairro:** Midtown
- **Latitude:** 40.75362
- **Longitude:** -73.98377
- **Tipo de Acomodação:** Entire home/apt
- **Mínimo de Noites:** 1
- **Número de Avaliações:** 45
- **Última Avaliação:** 21/05/2019
- **Reviews por Mês:** 0.38
- **Total de Imóveis do Anfitrião:** 2
- **Disponibilidade no Ano:** 355 dias

Qual seria a previsão de preço para este imóvel segundo o modelo?

Com base no modelo desenvolvido, a previsão de preço para o seguinte apartamento é $281,19 por noite.

O valor estimado foi calculado com base nas variáveis analisadas e no modelo de regressão linear desenvolvido.

Conclusão

Com base na análise exploratória dos dados e nas respostas às perguntas específicas do desafio, podemos tirar as seguintes conclusões:

1. **Preço das Acomodações**: As opções de acomodação privada, como casas, apartamentos e quartos privados, tendem a ser mais caras do que os quartos compartilhados, conforme evidenciado pelo gráfico boxplot. Isso é consistente com as expectativas de que a privacidade tem um preço maior.
2. **Influência da Localização**: A localização desempenha um papel crucial na definição dos preços, com Manhattan e Brooklyn apresentando os preços mais elevados, enquanto bairros como Queens, Staten Island e Bronx oferecem preços mais acessíveis. Isso indica que o mercado de aluguéis temporários em Nova York é sensível à localização, com áreas mais procuradas refletindo preços mais altos.
3. **Análise de Investimento**: Ao considerar o investimento em uma propriedade para alugar, o Bronx e o Queens se destacam como boas opções devido aos preços mais baixos. No entanto, Brooklyn e Manhattan, com maior número de avaliações, sugerem uma maior demanda e, portanto, um potencial de rentabilidade mais elevado. A análise de demanda, representada pelo número de avaliações, é um fator importante a ser considerado ao avaliar o potencial de retorno do investimento.
4. **Fatores que Influenciam o Preço**: O número mínimo de noites e a disponibilidade ao longo do ano não apresentaram uma relação clara com o preço, com os gráficos mostrando grande variação nos preços sem uma tendência definida. Isso sugere que esses fatores não são determinantes para o preço das acomodações, e outros aspectos, como a localização e o tipo de acomodação, são mais influentes.