Desafio Cientista de Dados – Lighthouse Indicium

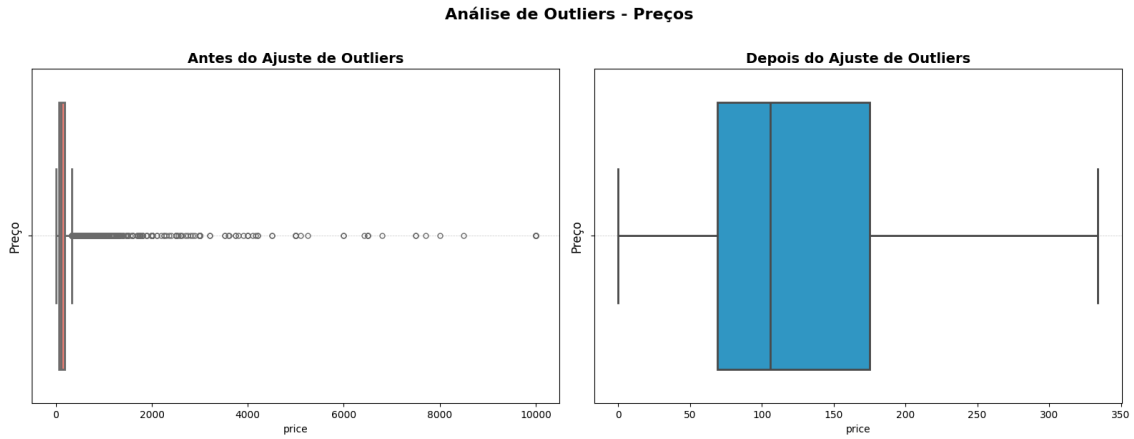
Beatriz Almeida Felício

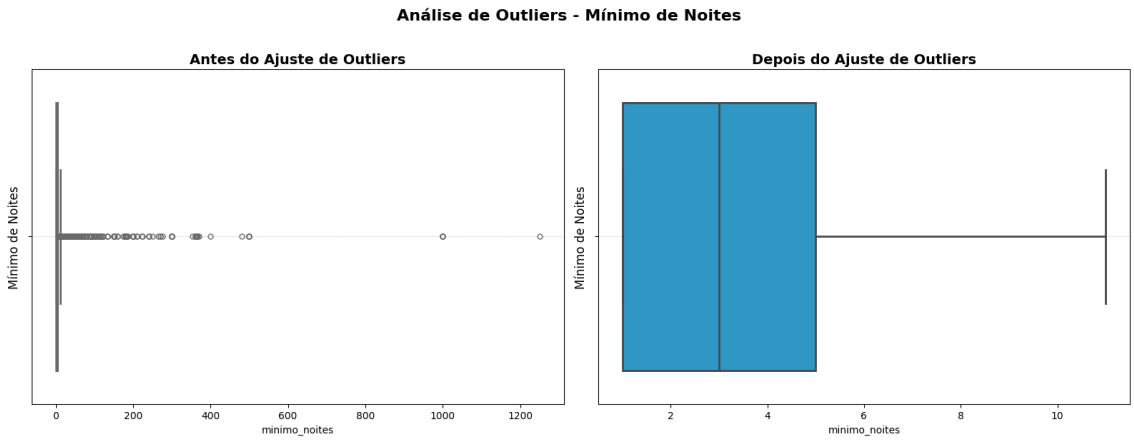
Descrição da atividade

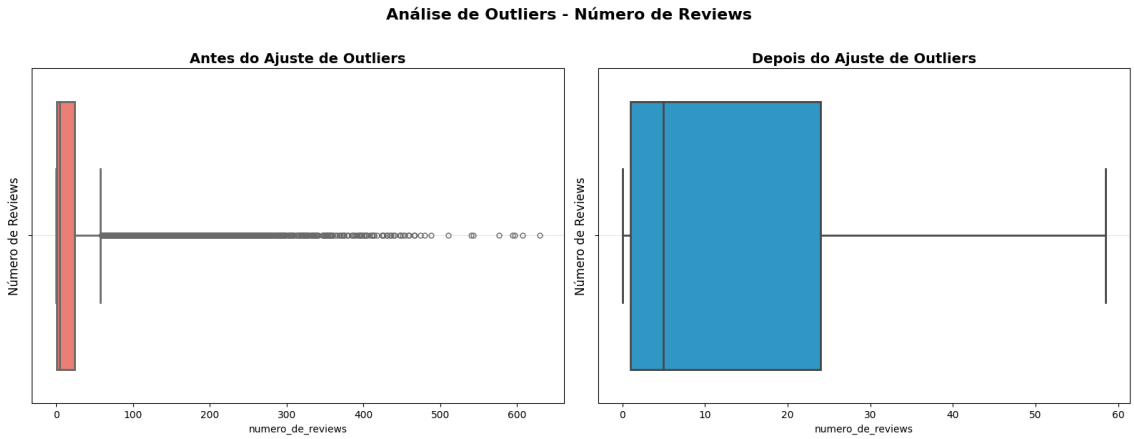
Análise Exploratória dos Dados (EDA)

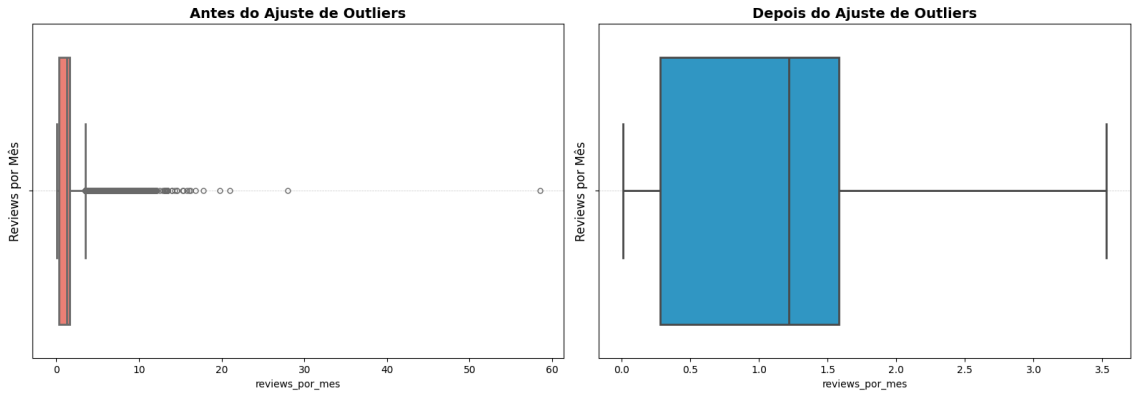
A análise exploratória foi iniciada com a aplicação de técnicas descritivas para compreender a estrutura e os detalhes do conjunto de dados. Inicialmente foram verificadas as informações disponíveis para identificar os tipos de dados presentes e planejar as melhores abordagens, considerando se as colunas eram qualitativas ou quantitativas, visto que essas características demandam métodos específicos. Durante o processo também foi avaliada a existência de valores ausentes e realizada uma análise estatística preliminar que evidenciou a presença de outliers.

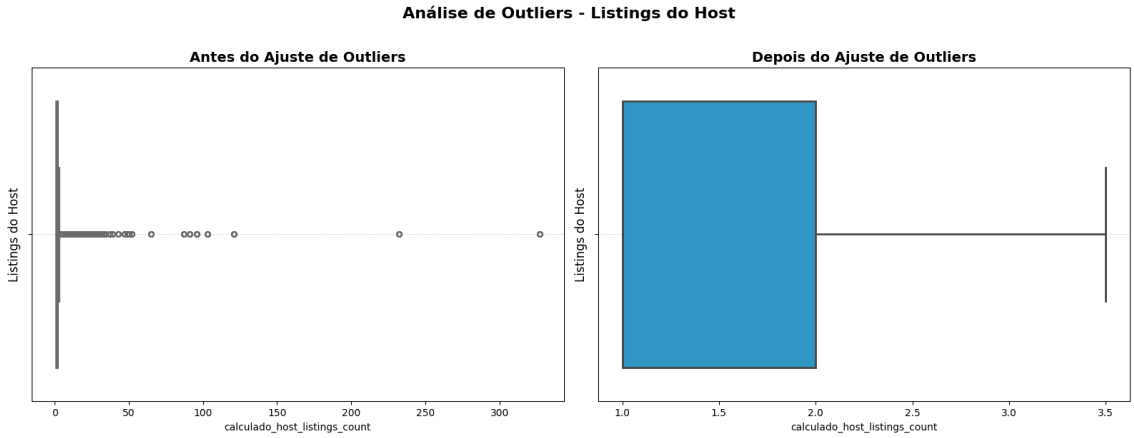
O dataset em análise contém um total de 48.894 linhas e 16 colunas, isso reforça a importância e necessidade de um tratamento minucioso dos dados para garantir a qualidade das etapas subsequentes. Com base nas informações levantadas, o primeiro passo foi tratar os valores ausentes. Para as variáveis qualitativas foram utilizados gráficos de visualização para identificar os valores mais frequentes e depois foi aplicado o cálculo da moda para substituí-los. Já para as variáveis quantitativas os valores ausentes foram substituídos pela média. Em seguida foi conduzida uma análise de outliers utilizando boxplots e para o tratamento foi aplicado o ajuste do IQR (Intervalo Interquartil), no qual os limites foram calculados com base nos quartis e os valores fora desses limites foram substituídos pelos próprios limites. Esse método foi bastante eficaz no tratamento de outliers e, a seguir, é apresentada a eficácia do tratamento de outliers aplicado na análise.





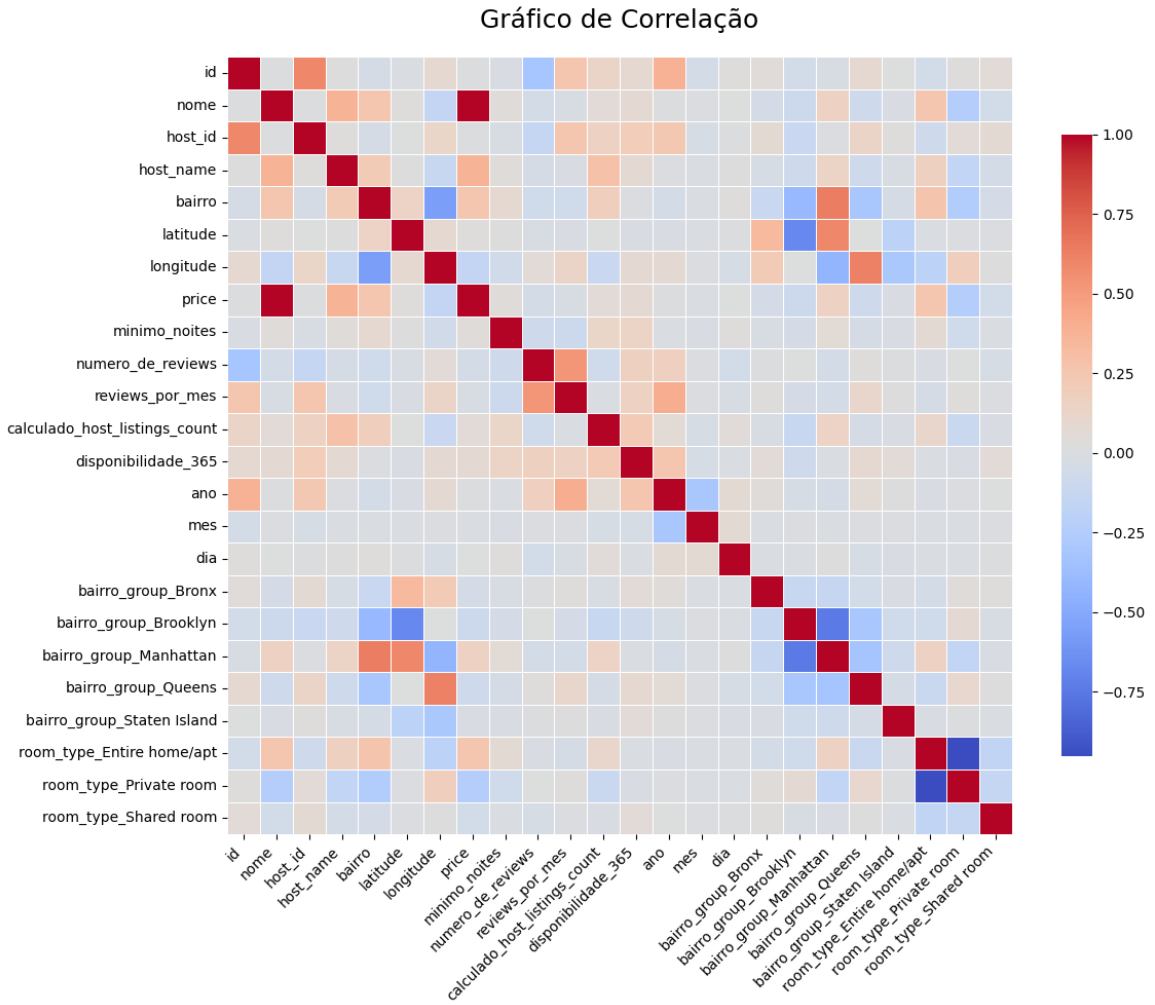






A coluna de datas inicialmente armazenada como string (object) foi convertida para o formato datetime para facilitar a análise de séries temporais. Posteriormente, essa coluna foi desmembrada em três novas colunas representando o dia, o mês e o ano porque os modelos de machine learning não conseguem trabalhar diretamente com o formato datetime. Além disso, foi realizado o encoding de variáveis categóricas para convertê-las em representações numéricas. Para colunas de alta cardinalidade foi utilizado o *target encoding*, que substitui os valores categóricos por estatísticas relacionadas à variável alvo "price", preservando a relação com o objetivo do modelo. Já para colunas de baixa cardinalidade foi aplicada a técnica de *one-hot encoding*. Essas transformações foram essenciais para garantir que os dados estivessem no formato adequado para treinar o modelo de machine learning, contribuindo diretamente para a eficácia e precisão das previsões.

Com o pré-processamento concluído, foi calculada a matriz de correlação das variáveis para avaliar a força das conexões entre elas e identificar as relações mais relevantes para as análises de negócio. Por fim, foram realizadas análises direcionadas às perguntas propostas no desafio, que serão detalhadas posteriormente neste relatório e na etapa final foi realizado o treinamento do modelo de machine learning para prever os preços, cujos resultados também serão apresentados nas próximas seções. Além disso, os gráficos detalhados que complementam a análise estão disponíveis no notebook principal do projeto.



Principais Características entre as Variáveis

- **Relação entre Preço e Nomes do Anúncio:**
  Os preços mais altos por noite demonstram forte correlação com o nome do anúncio e com o nome do host. Isso sugere que determinados hosts ou anúncios possuem maior prestígio ou valor percebido, resultando em tarifas mais elevadas.

- **Relação entre Preço e Tipo de Espaço:**
  Imóveis completos, como casas ou apartamentos, apresentam preços significativamente maiores em comparação com quartos privados, indicando que o tipo de acomodação é um fator que influencia no valor cobrado.

- **Relação entre Preço e Bairro:**
  A localização do imóvel, incluindo o bairro e a área em que ele se encontra, possui uma grande influência sobre os preços de aluguel, isso destaca a importância desse fator na determinação do valor. Esse impacto é evidente em áreas como Manhattan, onde a demanda e a valorização imobiliária são significativamente maiores, refletindo diretamente nos preços praticados.

- **Popularidade (Avaliações):**
  O número total de avaliações está diretamente relacionado à frequência de comentários mensais e à disponibilidade do imóvel ao longo do ano. Isso sugere que imóveis com maior disponibilidade tendem a ser mais populares, recebendo um volume maior de avaliações.

Hipóteses de Negócio

- **Investimento em bairros com preços médios competitivos e com alta concentração de anúncios**

  Os bairros que apresentam uma maior concentração de anúncios juntamente com preços médios competitivos podem ser os mais promissores para novos investimentos na plataforma de aluguéis temporários porque a maior disponibilidade aumenta a chance de ter mais reservas e, consequentemente, maior faturamento. Por isso, seria relevante identificar esses bairros estratégicos e direcionar campanhas de marketing ou parcerias para incentivar novos anfitriões a listar suas propriedades nessas áreas, aproveitando o potencial de demanda existente.

- **Programas de incentivo para hosts com maiores números de listagens**

  Hosts com um número maior de listagens demonstram um maior nível de experiência e têm maior potencial de gerar receita porque eles podem otimizar suas operações e ter uma base de clientes fiel. Com isso, seria importante desenvolver programas de incentivo ou parcerias exclusivas para hosts experientes, com foco em melhorar sua performance e ampliar sua rede de clientes, promovendo a fidelização do host e a otimização de receitas na plataforma.

- **Revisão da política de número mínimo de noites para reservas**

  Anúncios com uma exigência de mínimo de noites mais baixa atraem mais hóspedes porque isso proporciona mais flexibilidade, principalmente para viajantes de curta duração, negócios ou etc. Uma política de mínimo de noites mais baixa pode aumentar as chances de reserva, especialmente durante a baixa temporada, por isso seria importante incentivar anfitriões a revisar sua política de mínimo de noites e, caso fosse viável, reduzir essa exigência para aumentar a acessibilidade ao anúncio, visando melhorar a taxa de ocupação.

- **Estratégia de impulsionamento de anúncios com avaliações mais recentes**

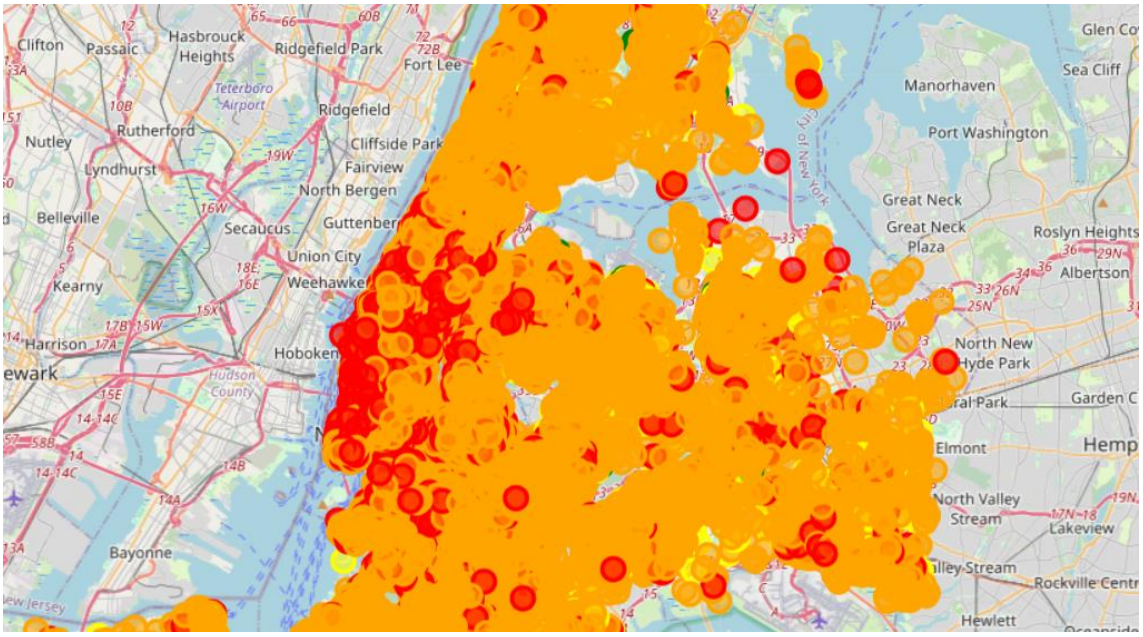
  A data das últimas avaliações é um indicativo da confiança e da reputação de um anúncio porque imóveis que possuem avaliações mais recentes geram maior atratividade e maior número de reservas. Então uma estratégia seria utilizar as análises de temporalidade das avaliações para promover anúncios que se destacam pela sua recente popularidade e para incentivar anfitriões a manterem um ciclo contínuo de feedback e melhorias, maximizando a atratividade do anúncio.

Perguntas de Negócio e Respostas

a. Onde seria mais indicada a compra de um apartamento para alugar na plataforma?

Foi realizada uma análise considerando as variáveis **preço por noite das reservas** e **popularidade das reservas**. Com base nos resultados a conclusão foi que áreas com maiores preços e alta popularidade representam um bom investimento para aquisição de apartamentos destinados à locação porque oferecem potencial para gerar um retorno financeiro significativo. O mapa gerado na análise é interativo e permite uma análise detalhada das regiões no código-fonte. Como exemplo, a imagem abaixo apresenta uma visão geral de alguns bairros. Os círculos no mapa são coloridos para indicar o potencial de investimento de cada área:

- **Verde**: Excelente investimento.
- **Amarelo**: Bom investimento.
- **Laranja**: Potencial intermediário.
- **Vermelho**: Menos recomendado.

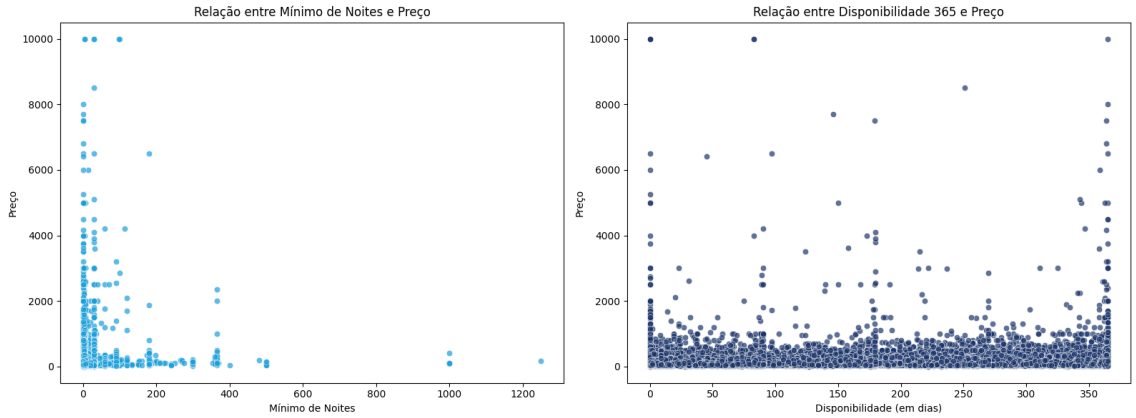


Com base na análise completa elaborada no código, é possível ver que os bairros mais recomendados são **Silver Lake**, **Richmondtown** e **Eltingville**, devido à boa relação entre preço e popularidade. O relatório detalhado, que apresenta a priorização dos bairros conforme a porcentagem de círculos verdes, amarelos, laranjas e vermelhos, está disponível em um arquivo de texto na pasta de dados.

```
Análise de Bairros para Aluguel (Ordenado por Avaliação de Potencial Investimento):

Bairro: Silver Lake
Total de imóveis: 2
Yellow (Bom para aluguel): 100.00%


Bairro: Richmondtown
Total de imóveis: 1
Yellow (Bom para aluguel): 100.00%


Bairro: Eltingville
Total de imóveis: 3
Yellow (Bom para aluguel): 66.67%
Red (Menos recomendado): 33.33%
```

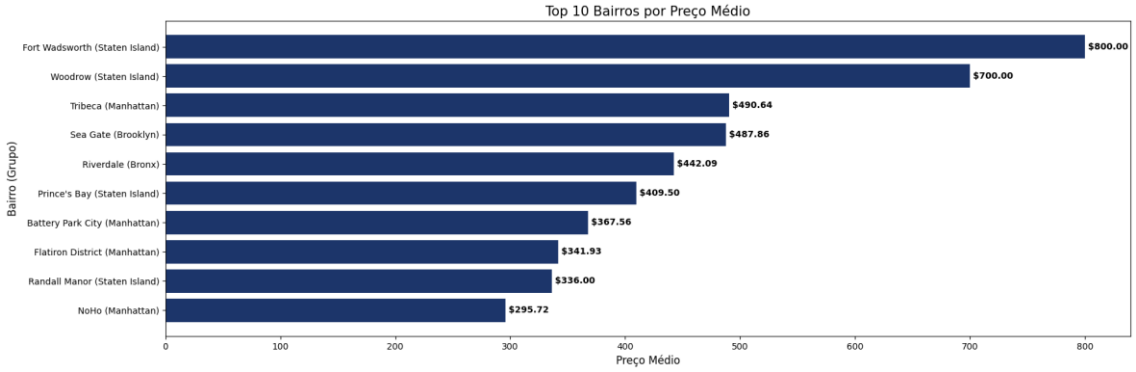
b. O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?

Foi realizada uma análise utilizando a matriz de correlação e gráficos de dispersão. A matriz indicou uma correlação linear próxima de zero, isso sugere que não há uma relação linear clara entre as variáveis analisadas e o preço. Nos gráficos de dispersão, embora também não seja possível observar uma relação linear, alguns padrões foram identificados. Na relação entre o número mínimo de noites exigido para reserva e o preço foi observado que a maioria das ofertas com preços baixos possui um número mínimo de noites reduzido, enquanto valores mais altos de noites mínimas apresentam uma menor concentração de preços elevados, ainda que existam casos isolados. Já na relação entre disponibilidade anual e preço foi visto que os preços mais baixos estão distribuídos de forma uniforme em diferentes níveis de disponibilidade, enquanto propriedades com alta disponibilidade exibem maior dispersão nos preços, incluindo alguns valores significativamente altos. Esses padrões indicam que, apesar da ausência de uma correlação linear, fatores como número mínimo de noites e disponibilidade anual podem influenciar o comportamento dos preços.



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

Para identificar possíveis padrões entre os nomes dos locais e os valores mais altos, foi realizado um agrupamento considerando o nome do bairro e da área onde os anúncios estão localizados. Essa análise permitiu identificar quais regiões apresentam os maiores valores e onde estão situadas. Foi observado que os anúncios com valores mais elevados estão concentrados frequentemente em Staten Island e Manhattan. Além disso, muitos nomes seguem um padrão composto por duas palavras, geralmente incluindo o nome do bairro ou da área seguido de um complemento geográfico e/ou histórico. Como por exemplo, "Fort Wadsworth" faz referência a um ponto histórico que continha construções de relevância cultural. Dessa forma, muitos nomes de bairros refletem características históricas, geográficas ou culturais, destacando a influência desses elementos nos altos valores observados.



Explicação referente ao item 3:

Para realizar a previsão dos preços de aluguel de apartamentos temporários na cidade de Nova York utilizei todas as variáveis disponíveis no conjunto de dados, aplicando transformações e tratamentos necessários para garantir uma melhor preparação dos dados para o modelo. Entre as principais etapas, destaquei a transformação da coluna de data, que foi dividida em três colunas separadas — dia, mês e ano. Além disso, foi realizada a transformação das variáveis categóricas para numéricas utilizando técnicas como o *One-Hot Encoding* para variáveis de menor cardinalidade, enquanto o *Target Encoding* foi aplicado em variáveis categóricas com maior cardinalidade. O tratamento de dados ausentes foi feito com imputação utilizando a média ou moda conforme o contexto da variável para garantir consistência no modelo. Outliers também foram cuidadosamente identificados e tratados por meio de métodos estatísticos, como o intervalo interquartil (IQR), para evitar distorções nos resultados.

O problema em questão é do tipo regressão porque o objetivo é prever um valor contínuo, nesse caso o preço de aluguel. Para modelar essa previsão optei pela Regressão Linear devido à sua simplicidade e eficiência computacional. Esse modelo é capaz de oferecer insights claros sobre como cada variável impacta o preço e é eficiente em cenários com relações lineares entre as variáveis. Entre as vantagens da regressão linear estão sua facilidade de implementação, baixo custo computacional e interpretabilidade direta. No entanto o modelo é sensível a outliers e isso exige um bom pré-processamento.

Para avaliar o desempenho do modelo utilizei as métricas *MAE* (Erro Médio Absoluto), *RMSE* (Raiz do erro quadrático médio) e R^2 (Coeficiente de Determinação). O MAE foi escolhido porque mede o erro médio absoluto entre os valores previstos e observados, sendo intuitivo e expresso na mesma unidade da variável dependente para facilitar a interpretação. Já o RMSE penaliza mais severamente erros maiores porque ele eleva os desvios ao quadrado antes de calcular a média e extrai a raiz, sendo útil para cenários em que desvios críticos devem ser destacados. Por fim, o R^2 avalia a proporção da variabilidade total dos dados explicada pelo modelo, fornecendo uma visão geral da qualidade do ajuste e da eficácia das previsões. Além disso, utilizei validação cruzada (K-Fold) para garantir a robustez e a generalização do modelo porque essa abordagem permite avaliar o desempenho em diferentes subdivisões dos dados, mitigando problemas como overfitting e proporcionando uma visão mais confiável sobre a capacidade do modelo em novos dados.

Os resultados obtidos mostraram que o modelo de regressão linear teve um bom desempenho nas métricas escolhidas, indicando que foi capaz de capturar bem as relações entre as variáveis e prever os preços de forma consistente. Essa combinação de análise, tratamento de dados e escolha de métricas permitiu que o modelo que elaborei se mostrasse eficiente e confiável no contexto proposto no desafio.

```
MAE (média): 1.6877
RMSE (média): 12.8725
R2 (média): 0.9971

Detalhes por fold:
MAE: [1.72637223 1.69892577 1.76475333 1.75318197 1.49528574]
RMSE: [12.66769692 14.53748225 13.67951056 12.84996251 10.62777185]
R2: [0.99678455 0.99673179 0.99753334 0.99710967 0.99722881]
```

Supondo um apartamento com as seguintes características. Qual seria a sua sugestão de preço?

```
{'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}
```

De acordo com o modelo elaborado, a previsão foi de R$224,46, um valor muito próximo do real presente no dataset que é de R$225. Essa pequena diferença demonstra a eficácia e a precisão do modelo que desenvolvi, evidenciando sua capacidade de capturar os padrões nos dados com alto grau de confiabilidade. Além disso, esse desempenho também mostra que o modelo está bem ajustado, sem sinais evidentes de overfitting ou underfitting. Isso tudo reforça a qualidade da abordagem adotada e a robustez dos parâmetros selecionados durante o treinamento e validação.

Conclusões

A análise exploratória e o pré-processamento realizados no conjunto de dados foram fundamentais para garantir a qualidade e confiabilidade das etapas subsequentes, permitindo uma modelagem mais eficiente e assertiva. A estruturação das variáveis, o tratamento de dados ausentes, outliers e a transformação das informações categóricas foram realizados com técnicas adequadas e resultou em um conjunto de dados otimizado para o treinamento do modelo de machine learning. Os resultados das análises propostas pelo desafio confirmaram a relevância de fatores como localização, tipo de espaço e popularidade no preço dos anúncios, reforçando hipóteses importantes para a tomada de decisões estratégicas no contexto de locações temporárias em Nova York.

A utilização de regressão linear como modelo preditivo se mostrou adequada ao problema proposto e alcançou resultados robustos e consistentes que foram evidenciados pelas métricas de desempenho (MAE, RMSE e R^2). A previsão realizada para o caso específico demonstrou precisão e alinhamento com os valores reais, destacando a eficácia do modelo em capturar padrões do conjunto de dados.

Esses resultados fornecem insights valiosos para otimizar a operação em plataformas de aluguel de curto prazo, como o direcionamento de investimentos para bairros estratégicos, desenvolvimento de programas para anfitriões experientes e ajustes em políticas de reservas para maximizar ocupação e receita. Dessa forma, a análise realizada representa uma base sólida para tomadas de decisão assertivas e planejamentos futuros, oferecendo soluções práticas e orientadas a dados para aprimorar a performance da plataforma.

Referências

Como Estilizar Gráficos no Seaborn. Disponível em: <https://hub.asimov.academy/tutorial/como-estilizar-graficos-no-seaborn/>. Acesso em: 24 jan. 2025.

DUARTE, R. **Métricas de Avaliação em Modelos de Regressão em Machine Learning**. **Sigmoidal**Carlos Melo, , 1 nov. 2023. Disponível em: <https://sigmoidal.ai/metricas-de-avaliacao-em-modelos-de-regressao-em-machine-learning/>. Acesso em: 27 jan. 2025

GABRIELPASTEGA, W. BY. **Outliers: Como definir, detectar e tratar — Parte 2**. Disponível em: <https://medium.com/@gabrielpbreis/outliers-como-definir-detectar-e-tratar-parte-2-5240149f8f98>. Acesso em: 24 jan. 2025.

GARG, S. **How to encode categorical data**. Disponível em: <https://garg-shelvi.medium.com/category-encoders-c2a9bb192f0a>. Acesso em: 25 jan. 2025.

Importing an ipynb file from another ipynb file? Disponível em: <https://stackoverflow.com/questions/20186344/importing-an-ipynb-file-from-another-ipynb-file>. Acesso em: 24 jan. 2025.

matplotlib.dates — Matplotlib 3.2.2 documentation. Disponível em: <https://matplotlib.org/3.2.2/api/dates_api.html>. Acesso em: 24 jan. 2025.

MOSCARDE, G. **Folium: Criando mapas interativos com dados reais de forma simples**. Disponível em: <https://moscarde.medium.com/folium-criando-mapas-interativos-com-dados-reais-de-forma-simples-c20ab89b5c79>. Acesso em: 25 jan. 2025.

PYTHON, R. **Python Folium: Create web maps from your data**. Disponível em: <https://realpython.com/python-folium-web-maps-from-data/>. Acesso em: 25 jan. 2025.

PYTHON, R. **Python mappings: A comprehensive guide**. Disponível em: <https://realpython.com/python-mappings/>. Acesso em: 26 jan. 2025.

RABELLO, E. B. **Cross Validation: Avaliando seu modelo de Machine Learning**. Disponível em: <https://medium.com/@edubrazrabello/cross-validation-avaliando-seu-modelo-de-machine-learning-1fb70df15b78>. Acesso em: 27 jan. 2025.

Regressão linear. Disponível em: <https://developers.google.com/machine-learning/crash-course/linear-regression?hl=pt-br>. Acesso em: 26 jan. 2025.

ROSA, H. N. **O que é Target Encoding e como aplicá-lo. - Heitor Nunes Rosa**. Disponível em: <https://medium.com/@heitornunesrosa/o-que-%C3%A9-target-encoding-e-como-aplic%C3%A1-lo-85996e1bf00d>. Acesso em: 25 jan. 2025.

SANTOS, G. R. **Criando Mapas Interativos e Choropleth Maps com Folium em Python**. Disponível em: <https://medium.com/data-hackers/criando-mapas-interativos-e-choropleth-maps-com-folium-em-python-abffae63bbd6>. Acesso em: 25 jan. 2025.

Seaborn: Statistical data visualization — seaborn 0.13.2 documentation. Disponível em: <https://seaborn.pydata.org/>. Acesso em: 24 jan. 2025.