1. Análise exploratória dos dados (EDA)

Iniciamos o processo tratando os valores faltantes ou zeros nas variáveis numéricas. Removemos os 11 anúncios sem preço e preenchemos os valores nulos nas colunas 'ultima\_review' e 'reviews\_por\_mes' com zeros. Com os dados numéricos tratados, geramos uma matriz de correlação para essas variáveis.

Conforme mostrado na Figura 1, apenas a longitude apresenta uma correlação negativa fraca com o preço. No entanto, é importante destacar que a ausência de correlação linear não descarta a possibilidade de existir alguma relação não linear entre as variáveis. Portanto, os modelos de machine learning podem ser capazes de identificar padrões mais complexos nos dados que não são evidentes pela simples análise de correlação linear.

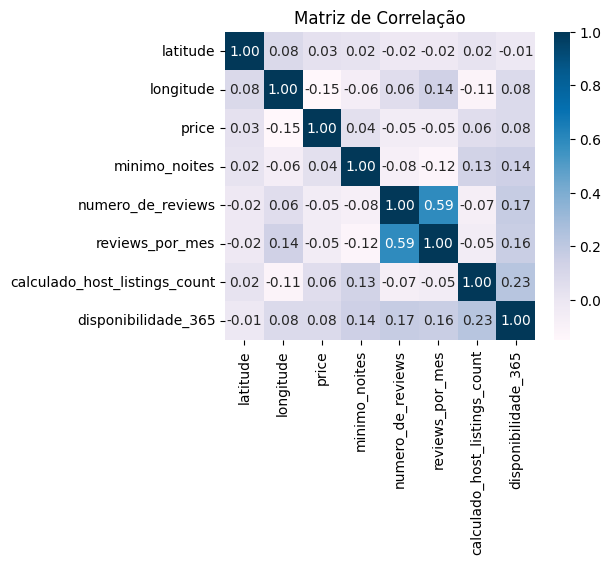


Figura 1 - Matriz de correlação para as variáveis numéricas.

Analisando as variáveis "room\_type", "bairro\_group" e "bairro", observamos que elas influenciam no preço. No entanto, para simplificar o modelo de predição, não será utilizada a variável "bairro", devido à sua grande quantidade de entradas únicas. Assim, utilizaremos apenas a variável "bairro\_group", que contém o mesmo tipo de informação geográfica de bairro, porém é mais abrangente. Essa simplificação não resulta em perda significativa de informação, uma vez que, como veremos, latitude e longitude fornecem informações sobre a localização dos imóveis e sua relação com o preço.

Conforme a Figura 2, os preços de quartos compartilhados (“shared rooms”) anunciados na plataforma são pouco menores do que os preços de “private room”. Além disso, há poucas opções de quartos compartilhados na plataforma, o que sugere uma possível estratégia de investimento. Em áreas estratégicas, imóveis e quartos privados podem ser disponibilizados como quartos compartilhados, aumentando a lucratividade por imóvel.

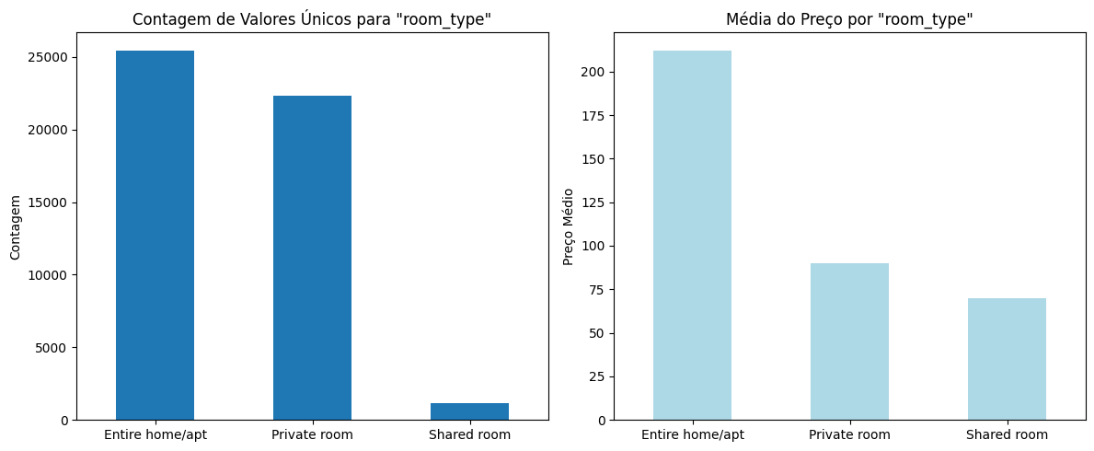


Figura 2 - Quantidade de imóveis e preço médio por tipo de acomodação.

Ao analisar a Figura 3, podemos concluir que Manhattan apresenta o maior preço médio de imóveis, mesmo com uma grande disponibilidade de imóveis na plataforma. Isso sugere que a demanda nessa localidade é alta.

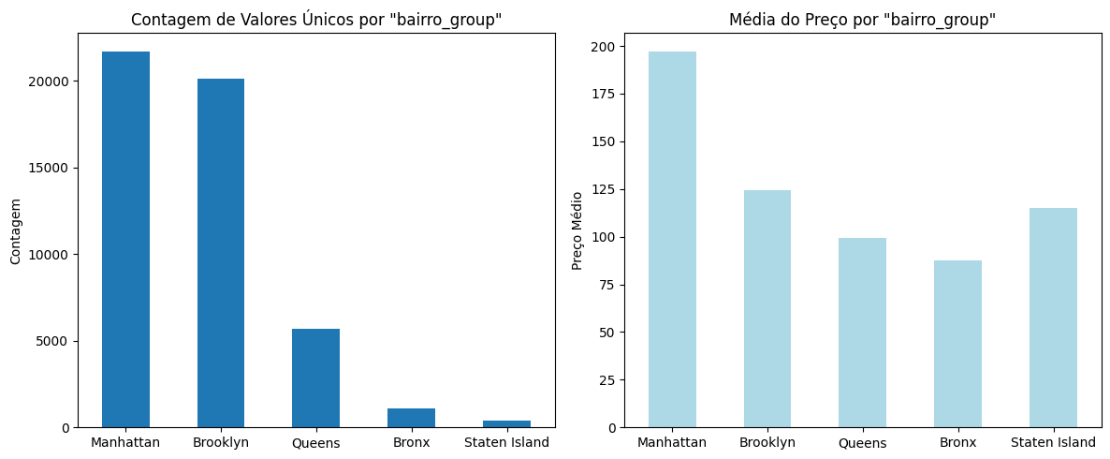


Figura 3 - Quantidade de imóveis e preço médio por grupo de bairros.

A partir da Figura 4, constatamos visualmente que a região sul da ilha de Manhattan concentra uma maior quantidade de anúncios com valores elevados (acima de 200). Isso sugere que a margem de lucro para a locação de quartos compartilhados na parte sul de Manhattan pode ser maior.

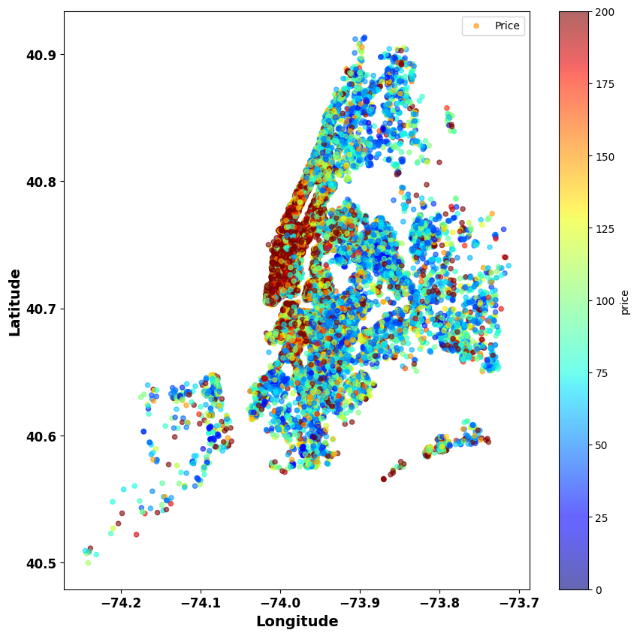
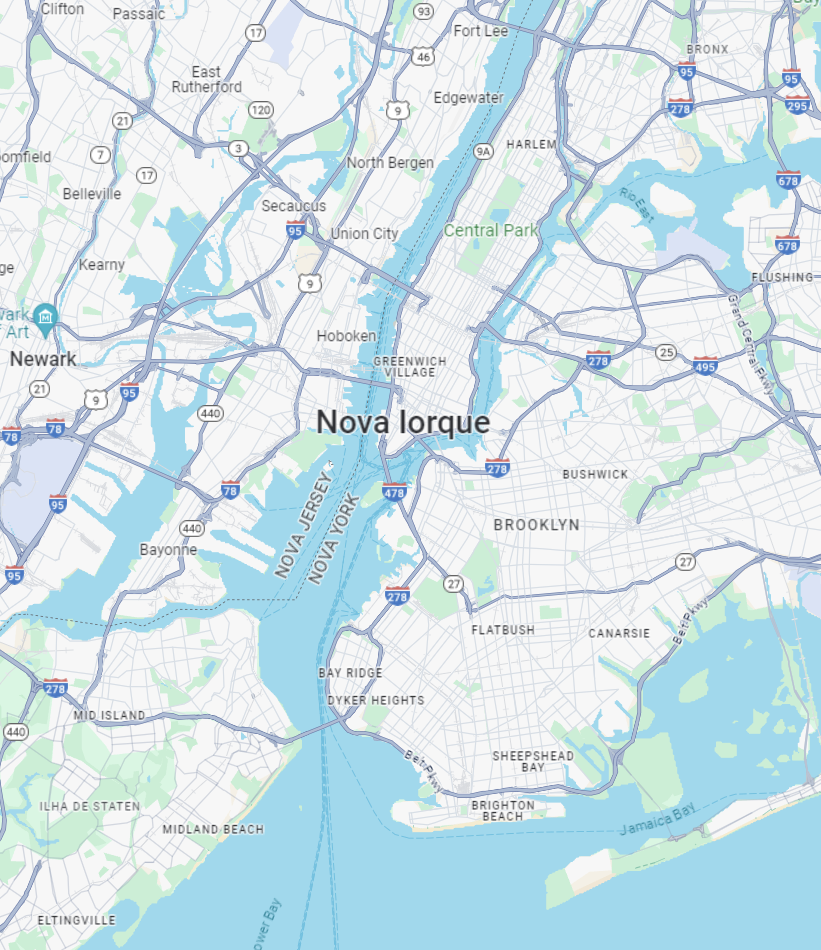
 

Figura 4 – Mapa de calor dos preços dos imóveis por latitude e longitude (esquerda) e mapa de Nova Iorque (direita). fonte: <https://www.google.com.br/maps>.

Os gráficos de dispersão das variáveis “mínimo\_noites” e “disponibilidade\_365” (Figura 5) sugerem que essas variáveis não têm uma influência direta sobre o preço dos imóveis. No entanto, decidimos incluí-las no modelo devido à possibilidade de conterem informações relevantes para capturar nuances ou padrões nos dados que podem não ser imediatamente aparentes pela visualização gráfica.

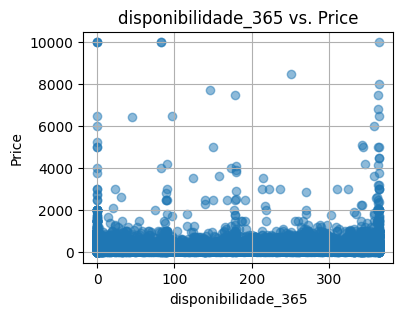
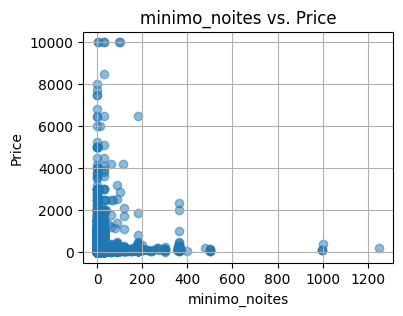


Figura 5 – Gráfico de dispersão do preço para mínimo de noites (esquerda) e disponibilidade ao longo do ano (direita).

1. Perguntas:
   1. A análise realizada sugere que uma estratégia interessante para investir em um apartamento para alugar na plataforma é focar na região ao sul de Manhattan (Lower Manhattan). Conforme destacado na Figura 4, essa área está bem valorizada, o que pode resultar em um bom retorno sobre o investimento. Além disso, considerando a quantidade reduzida de quartos compartilhados disponíveis na plataforma como um todo e proporcionalmente em Manhattan, investir em um apartamento nessa região pode ser altamente lucrativo. Isso ocorre porque a demanda por hospedagem compartilhada tende a ser alta em áreas turísticas como Manhattan. Portanto, ao escolher investir em um apartamento para alugar na plataforma, considerar a região sul de Manhattan e a modalidade de locação de quartos compartilhados maximizará o retorno do investimento.
   2. Conforme a Figura 1 e a Figura 5, não foi identificada uma correlação clara entre o número mínimo de noites e a disponibilidade ao longo do ano com o preço dos imóveis. No entanto, é importante ressaltar que essas variáveis ainda podem conter informações relevantes que não são facilmente capturadas graficamente ou pela análise de correlação linear. Portanto, mesmo que não haja uma correlação evidente, é prudente incluir essas variáveis no modelo de previsão, pois elas podem contribuir para capturar padrões e assim prever os preços dos imóveis de forma mais precisa.
   3. Sim. Palavras como Luxury e Furnished quando presentes no nome indicam, na média, lugares de mais alto valor. Essas informações podem ser fornecidas aos modelos de machine learning para melhorar os resultados das regressões.
2. As variáveis mais relevantes para a previsão do preço, como localização geográfica (latitude, longitude, bairro), características do anúncio (número mínimo de noites, número de reviews, disponibilidade ao longo do ano) tipo de acomodação (room\_type) e descrição do anúncio (nome) são utilizadas em um modelo de machine learning para a predição.

As variáveis categóricas foram convertidas em variáveis dummy, incluindo a criação de novas colunas a partir do nome do anúncio para indicar se as palavras 'Luxury' e 'Furnished' estavam presentes.

As variáveis numéricas latitude e longitude foram transformadas por normalização, para evitar que elas dominem o processo de treinamento do modelo devido à escala dos valores. Por outro lado, as demais variáveis numéricas foram assumidas como normalmente distribuídas e, portanto, padronizadas para manter a média próxima de 0 e o desvio padrão próximo de 1.

O tipo de problema que estamos resolvendo é de regressão, pois estamos prevendo um valor contínuo (o preço dos imóveis). Dentre os modelos que poderiam ser utilizados para esse tipo de problema, alguns dos mais comuns incluem regressão linear, árvores de decisão, florestas aleatórias, redes neurais e técnicas de grandient boosting, como o XGBoost. As redes neurais e o XGBoost estão frequentemente entre as técnicas vencedoras em competições que envolvem regressão.

Para problemas de regressão, medidas comuns de performance incluem o erro médio absoluto (MAE) e o erro quadrático médio (MSE). O MAE é uma escolha popular porque é fácil de interpretar e é menos sensível a outliers em comparação ao MSE. Desse modo, o MAE é utilizado como medida da performance dos modelos.

Inicialmente a função MLPRegressor da biblioteca Scikit-Learn foi utilizada na tarefa de regressão, o MAE para essa implementação foi de 72.64 nos dados de validação. Também foi implementada uma rede neural utilizando o TensorFlow que atingiu um MAE de 59.49 nos dados de validação. Entretanto, um modelo de XGBoost acabou sendo selecionado, pois, além de um MAE de 58.41 nos dados de validação, esse modelo é mais simples do que a implementação no TensorFlow. Nos dados de teste o MAE do XGBoost chegou a 53.76 e a inspeção visual de alguns resultados de teste demonstra uma boa capacidade do modelo.

1. A sugestão de preço utilizando o modelo proposto é de 233. Este valor é ligeiramente superior ao preço de 225 informado na plataforma. Portanto, ao compararmos a previsão do modelo com o valor real, este apartamento parece ser um ótimo investimento.

(Observação: O apartamento analisado faz parte do conjunto de dados. Para evitar a contaminação dos dados de treinamento, este registro foi removido logo após a importação dos dados.)

1. O modelo e os scalers utilizados para normalizar e padronizar os dados foram salvos no arquivo modelo\_e\_scalers.pkl.
2. Repositório de código público para acessar os arquivos:

Git\_hub\_arquivos\_123456