Análise Exploratória dos Dados

- O Dataset apresentado possui aproximadamente 50 mil linhas de dados com 16 colunas de informações.

O Objetivo deste projeto é desenvolver uma análise que ajude a formular uma estratégia de precificação para imóveis de locação por temporada na cidade de Nova York e também testar um modelo preditivo que possa auxiliar os anunciantes a definir com mais precisão o preço dos aluguéis.

O problema de previsão de preços de aluguel é um clássico problema de Machine Learning no qual utilizamos algoritmos de regressão. Baseado em uma série de dados numéricos que sintetizam as categorias dos imóveis e dos preços, queremos prever qual será o preço de um imóvel novo na base de dados utilizando suas características para determinar a melhor faixa de valor.

Além do dataset fornecido, foram utilizadas algumas informações geográficas que ajudaram a entender melhor a relação entre localização e preço do aluguel. Mapas e pesquisas no site da data.ny.gov e na página data.cityofnewyork.us ajudaram a compreender melhor a dinâmica da cidade e acrescentaram alguns dados à análise.

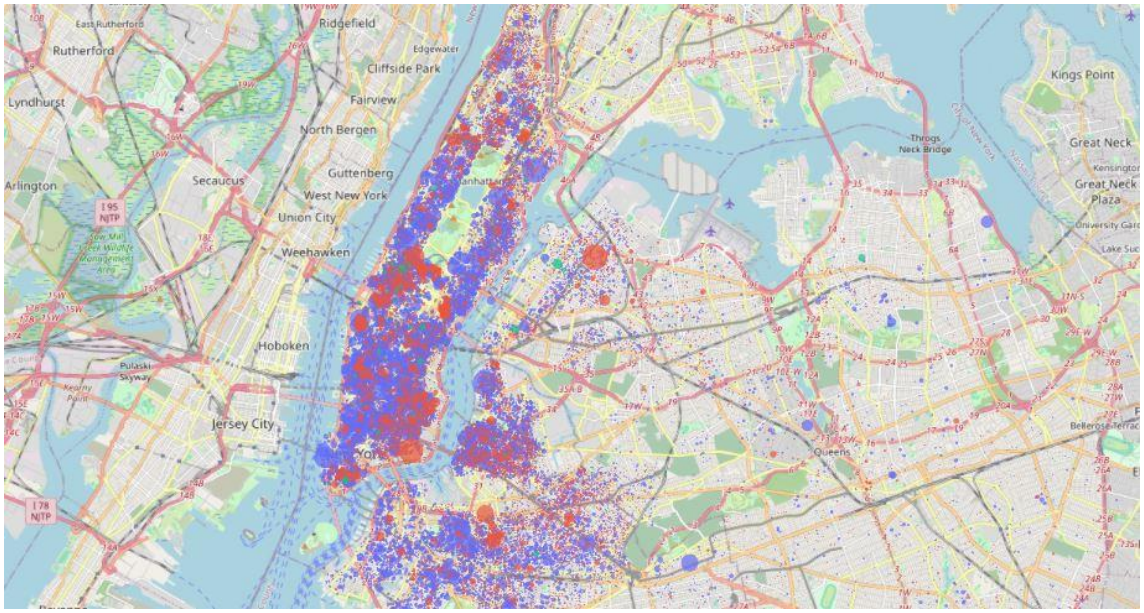
- **Tratamento dos dados e feature engineering**

Os dados apresentados estavam estruturados de uma forma coesa, com poucos missing values (dados vazios) e com tipagem correta. Foi necessário alterar alguns poucos tipos de dados e para a etapa de modelagem foi necessário escolher e tratar dados nulos, pois os algoritmos de ML normalmente não lidam bem com dados nulos.

Também foi necessário transformar dados categóricos em categorias numéricas, por exemplo: a coluna 'bairro_group' possuía os valores em formato de texto (string) que eram 'Manhattan', 'Queens', 'Brooklyn', etc. Estes valores foram alterados para números que correspondiam a cada região da cidade. O mesmo foi feito com a coluna 'room_type', que possuía três categorias e foi transformada em três códigos numéricos.

A coluna bairros possuía alta cardinalidade (muitos valores diferentes) por isso foi feita uma seleção dos 50 bairros com mais imóveis e estas categorias de bairros foram transformadas em colunas binárias utilizando a técnica de One Hot Encoding.

A distância de cada imóvel para um "centro" foi escolhida para enriquecer o dataset. Ao olharmos atentamente para o mapa abaixo, fica evidente um padrão de imóveis mais valorizados na ilha de Manhattan, nas adjacências e ao longo das linhas de transporte.



Mapa tipo Scatter Plot mostrando a distribuição dos imóveis, bem como o tipo e o valor.

Os círculos maiores indicam preços mais altos e as cores indicam o tipo de imóvel, se 'entire house / apartment', 'entire room' ou 'shared room'. Podemos notar como a concentração e o valor aumenta quanto mais próximos os imóveis estão do centro de Manhattan. Por isso defini o Central Park como centro e utilizei a distância de cada imóvel até este centro como uma possível feature importante para a predição.

Depois procuramos responder a uma das perguntas mais importantes do negócio: Quais seriam as regiões da cidade mais propensas a receber investimento caso um investidor queira adquirir um imóvel. Certamente as regiões mais valorizadas, como mostrado no mapa, seriam

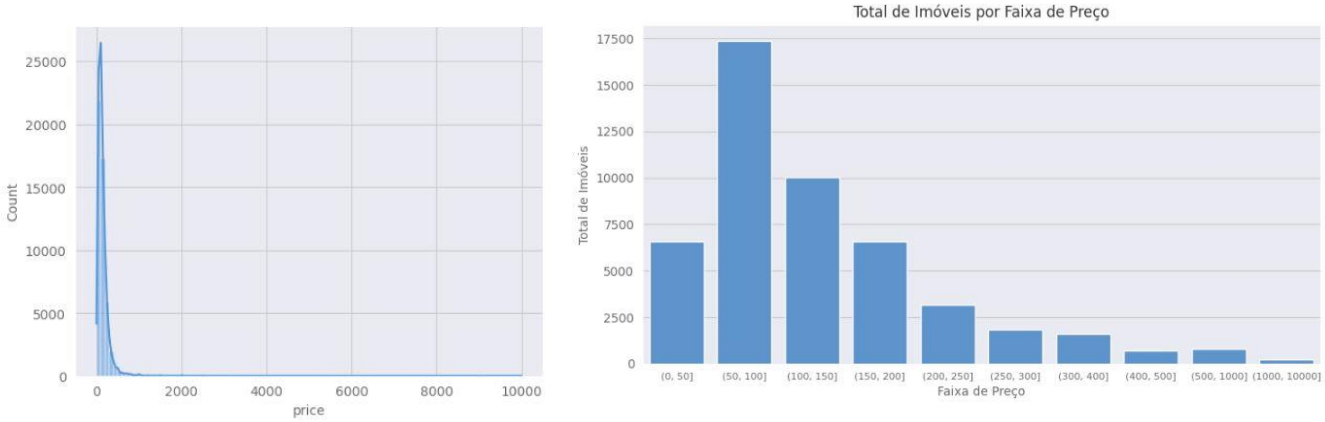
mais propensas a receber investimento, porém elas também possuem os maiores preços de aquisição, mas seriam mais rentáveis no médio e longo prazo, tanto por conta da valorização, quanto da liquidez e facilidade em alugar. Porém pondero que para fazer uma análise mais precisa, seria necessário incluir mais informações, como, por exemplo, a quantidade de vezes que cada imóvel foi alugado na plataforma ao longo de um período de tempo.

- **Análise Univariada**

Nesta etapa a análise se concentra em entender as distribuições de cada feature e como cada variável numérica ou categórica se comporta nesta amostra de dados. Iniciamos olhando para a categoria 'room_type', que possui mais de 51% dos imóveis na categoria 'entire house / apartment' e mais 45% deles pertencem à categoria 'entire room'. Aproximadamente 2,5% apenas são classificados na categoria 'shared room'. Isso demonstra claramente a predileção dos clientes em alugar espaços únicos onde possam ter privacidade. Mas como Nova York é uma cidade famosa por seus altos preços de estadia e hospedagem, existem muitos imóveis que são mais baratos do tipo 'entire room', onde o hóspede fica em um quarto em uma casa compartilhada, modelo que serve bastante bem para estudantes e pessoas viajando sozinhas.

Em seguida foi realizada uma avaliação da quantidade de imóveis por região da cidade, que confirmou a visualização do mapa acima, onde a maioria dos imóveis está localizada em Manhattan, seguida por Brooklyn e depois Queens. Foi feita a mesma análise em relação aos bairros e neste caso o bairro com mais imóveis para aluguel foi Williamsburg, seguido por Bedfort-Stuyvesant (no quadrante sudeste) e depois Harlem (ao norte da ilha de Manhattan).

A análise da distribuição dos preços revelou um fato interessante: a grande maioria dos preços está na faixa de 50 a 150 U$, mas há também imóveis que possuem preços variando até 10000 U$, o que poderíamos considera como outliers que enviesam a distribuição. Quando olhamos para esta distribuição em 'bins', compartimentos, podemos perceber uma distribuição mais clara, mas ainda assim bastante enviesada à esquerda, com predominância de valores entre 50 e 150.



Distribuição de todos os preços e distribuição em compartimentos. Mostra mais claramente a predominância de valores na faixa descrita.

A distribuição do número mínimo de noites mostra que a maioria dos imóveis possui mínimo de 1 a 3 dias de aluguel, o que condiz plenamente com o propósito da plataforma de aluguel de imóveis por temporada. Surpreendentemente, existem quase 10% dos imóveis que exigem 30 dias de aluguel. Quando a disponibilidade do imóvel é avaliada, também surge um fato interessante: muitos imóveis não possuem disponibilidade ou possuem disponibilidade muito pequena. Uma possibilidade é que os imóveis estejam já alugados e por isso não estão disponíveis, mas também pode significar que muitos proprietários estão ainda completando seus perfis ou haviam retirado a disponibilidade dos seus locais no momento da coleta dos dados.

O número de reviews também apresenta uma distribuição muito enviesada à esquerda, ou seja, aproximadamente 10% dos dados não possuem nenhuma review e outros 10% possuem até no máximo 5 reviews. Mas há casos de imóveis que possuem 300 ou até 600 reviews. O mesmo se aplica à quantidade de reviews por mês, sendo que a maioria dos imóveis que recebem reviews, recebe até 1 review por mês. Este fator depende muito da interação entre o proprietário e o hóspede e muitos hóspedes simplesmente não fazem reviews. Mais adiante iremos avaliar como estes vários fatores se relacionam com o preço, nosso variável alvo, aquela que desejamos prever com os modelos de Machine Learning.

- **Análise Multivariada**

Em primeiro lugar, a relação mais evidente que é do preço com a localização. Já foi descrito de diversas formas como locais mais centrais e principalmente na ilha de Manhattan possuem preços mais elevados do que o resto, porém através do gráfico abaixo isso fica ainda mais evidente. O que podemos observar é que os preços em Manhattan são claramente mais altos, sendo a média destes valores bem maior que o restante e com uma dispersão muito maior de valores altos (limitei a visualização até 1000 U$ para facilitar o entendimento). E m seguida vem a região do Brooklyn com preços mais baixos porém ainda mais altos que o restante e também possui uma ‘cauda’, uma dispersão mais longa de valores maiores. As outras três regiões possuem um comportamento bem similar, com o Bronx apresentando preços mínimos um pouco mais baixos.

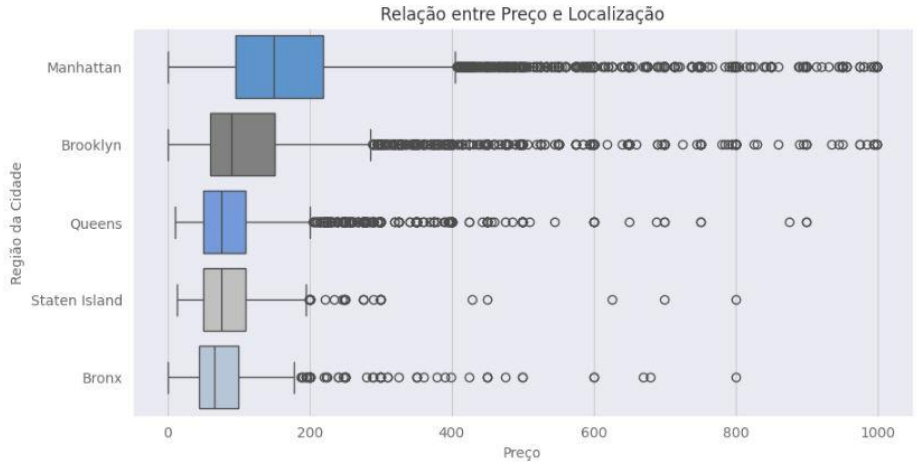


Gráfico Box Plot relacionando preço e localização (limite de preços U$ 1000)

A variável distância Central Park, quando relacionada através de um gráfico de dispersão com os preços, evidencia mais uma vez aquilo que já foi dito: a proximidade com o centro faz os preços ficarem mais altos e amplia o intervalo destes preços. O que podemos observar do gráfico abaixo é que há uma grande concentração de pontos na faixa até U$1000, que se distribui por praticamente todo o intervalo de distância, até aproximadamente 30km, mas valores maiores se concentram em localizações que estão a no máximo 10km do Central Park. Os imóveis mais baratos estão mais afastados, na região de Staten Island, entre 30 e 40 km de distância.

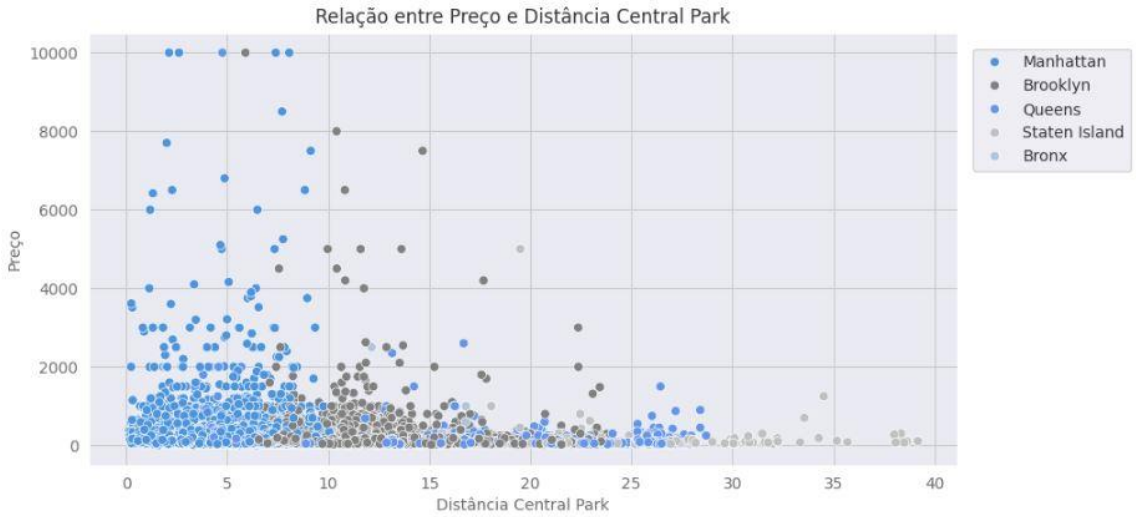


Gráfico tipo Scatter Plot (dispersão) mostrando relação entre Preços e Distância até o Central Park

Não é possível determinar uma relação direta entre o mínimo de noites e o preço. O que percebemos observando o gráfico de dispersão abaixo é que ele forma padrões de linhas verticais, ou seja, os imóveis de diferentes preços se concentram em números inteiros de dias, por exemplo, 7, 10, 30, 60 e múltiplos de 30 (múltiplos de meses).

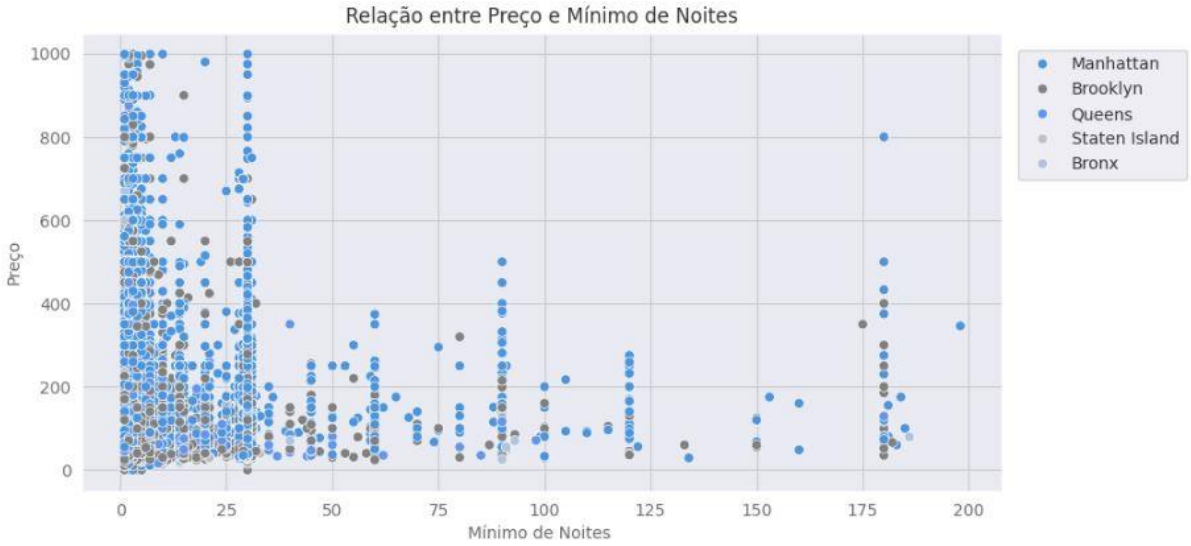


Gráfico tipo Scatter Plot (dispersão) mostrando relação entre Preços e Mínimo de Noites

A mesma impossibilidade de determinar uma relação clara entre preço e disponibilidade pode ser observada no gráfico que possui uma quase homogeneidade de dispersão tanto no eixo de preços quanto no eixo dos dias disponíveis.

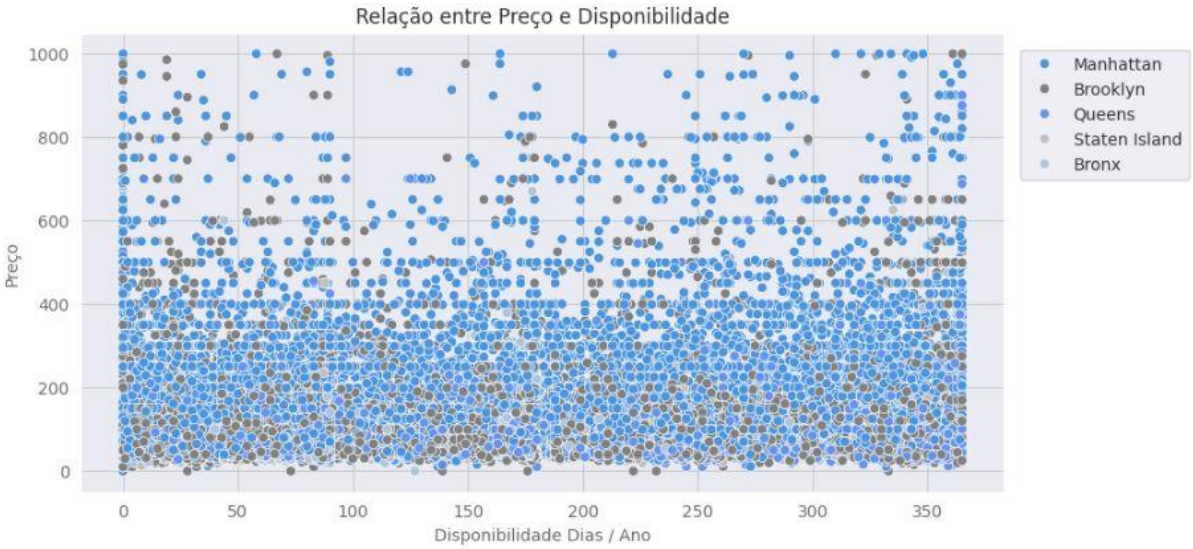


Gráfico tipo Scatter Plot (dispersão) mostrando relação entre Preços e Disponibilidade

Também realizei uma análise utilizando algoritmos de Natural Language como NLTK e Spacy para tentar compreender se havia algum tipo de padrão nos textos dos títulos dos anúncios que poderia ser 'lido' e compreendido por estas ferramentas de inteligência artificial que trabalham com linguagem. Não observei nada substancial que pudesse ser utilizado como preditor dos preços, além do que me pareceu óbvio: que o título de imóveis de alto padrão carregam adjetivos em inglês do tipo 'beautiful', 'luxury', ' penthouse', etc, para descrever o imóvel. Também me pareceu óbvio que anúncios que continham as palavras 'Manhattan', 'Tribeca', 'East Side', 'Park Avenue', ou seja, palavras que nomeiam a localização dos imóveis possuíam preços mais altos. De todo modo caberia um estudo mais detalhado no sentido de procurar possíveis adjetivos que chamem mais a atenção dos usuários, mas para isso precisaríamos também avaliar através de testes estatísticos se alterações nos anúncios refletiriam em maiores índices de locação.

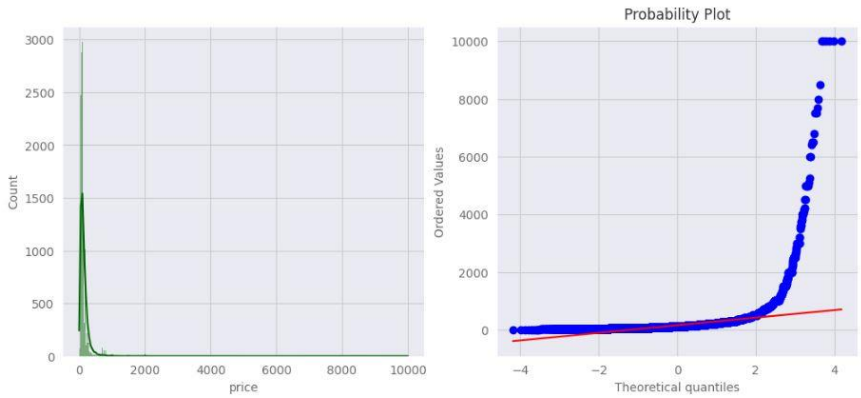
- **Modelagem**

Este é um problema de previsão de preço, portanto um problema de regressão. Temos uma série de variáveis ditas preditoras e uma variável alvo que desejamos descobrir o valor com o mínimo de erro possível. A qualidade das variáveis independentes influi diretamente na precisão e na confiabilidade dos modelos preditivos. Percebe-se através das análises exploratórias e do conhecimento do negócio (entendimento sobre a dinâmica urbana) que as variáveis preditoras mais importantes devem ser aquelas relacionadas à localização, distância do centro e tipo do imóvel. Essa intuição será confirmada no momento da modelagem e da avaliação dos resultados, como será descrito mais adiante.

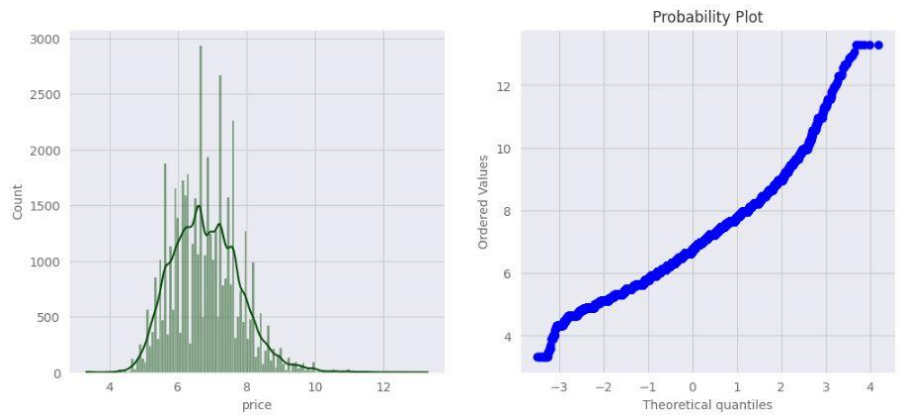
Algumas experiências foram feitas para determinar qual tipo de modelo / algoritmo iria obter um melhor resultado. Primeiro aplicamos um modelo de decision Tree, um modelo de árvore simples onde o objetivo foi servir como ponto de partida para outros modelos mais complexos. Depois foi aplicado um modelo de regressão do tipo Random Forest, um modelo que faz uso de técnicas de bagging e boosting para aprimorar o aprendizado de uma árvore subsequente em relação à anterior, um modelo muito eficiente, porém que pode ser computacionalmente custoso e lento. Por fim o modelo aplicado foi o XGBoost, modelo muito eficiente que tem por objetivo achar o ponto de menor erro através de um processo de gradiente descendente, onde o modelo vai se aproximando de minimizar o erro em cada iteração até encontrar o ponto ideal de equilíbrio da função de erro entre o real e o predito.

O modelo que obteve melhor performance final foi o XGBoost e as métricas de avaliação do modelo usadas foram o R2, que descreve o quão próximos os dados previstos estão da linha de regressão. Este valor varia de 0 a 1 e quanto mais próximo de 1 melhor a predição. A segunda métrica avaliada foi o MAE (mean absolute error) ou erro absoluto médio, quanto de erro em média o conjunto dos dados preditos obteve acima ou abaixo dos valores reais. Neste caso quanto menor o valor do MAE, melhor o modelo conseguiu acertar os valores preditos.

Foi necessário aplicar uma transformação logarítmica à variável preço, pois sua distribuição estava muito assimétrica e isso atrapalhou os primeiros experimentos . Foi utilizada um transformação log2 para deixar a curva mais simétrica, de acordo com os gráficos abaixo que demonstram antes e depois da transformação.



Distribuição da variável alvo (price) antes da transformação LOG2



Distribuição da variável alvo (price) após a transformação LOG2

Nos gráficos após a transformação podemos ver graficamente como a curva em forma de sino, típica das distribuições normais, apareceu com mais clareza e no gráfico tipo QQ plot, que relaciona os quartis reais com os quartis teóricos, vemos que a forma dos pontos se aproxima de uma reta a 45 graus que representa uma distribuição perfeitamente simétrica. Portanto ao final do processo, iremos precisar inverter a função LOG2 para chegar ao valor final.

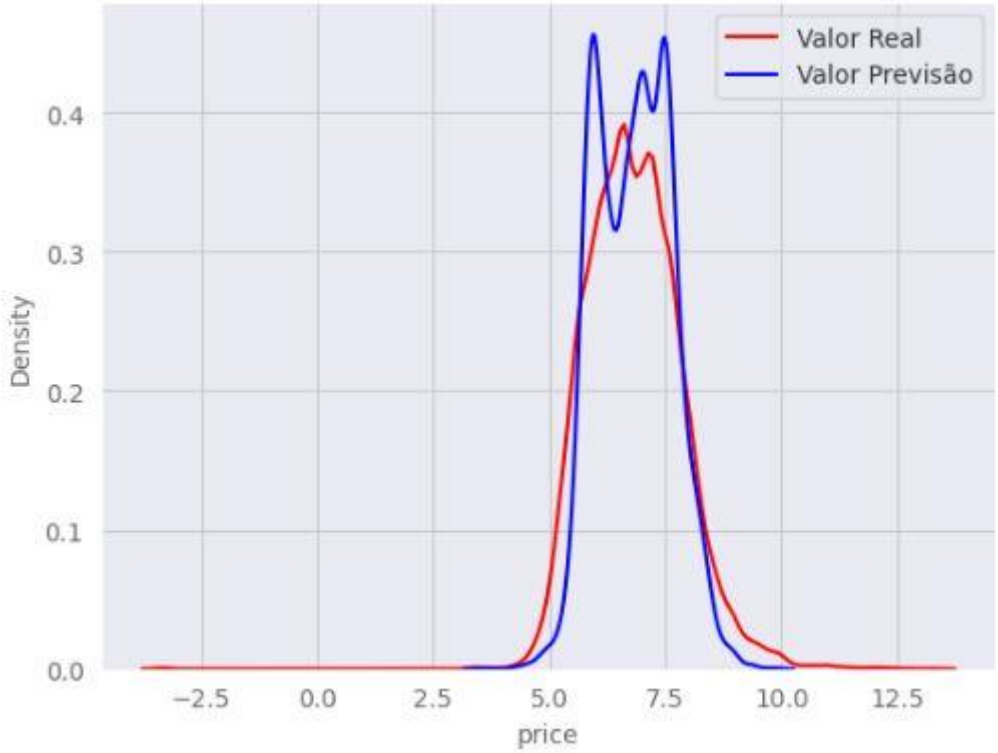


Gráfico de distribuição dos valores reais e preditos do modelo final XGBoost. Quanto mais próxima a curva previsão estiver da curva real, melhor a capacidade de previsão do modelo.

O modelo XGBoost final, após tunagem dos hiperparâmetros apresentou os seguintes resultados:

R2 = 0.558

MAE = 0.475

Isso significa que o modelo é capaz de explicar aproximadamente 55% da variabilidade dos valores reais, pode-se dizer de forma resumida que o modelo é capaz de “acertar” em 55% dos casos. O MAE abaixo de 0,5 é um bom indicador de quanto o modelo irá errar, quão perto dos valores reais ele será capaz de chegar.

O gráfico de resíduos abaixo ajuda a entender melhor quão “próximo do alvo” o modelo consegue acertar. A linha tracejada no zero seria um acerto perfeito e pra cima ou pra baixo são os resíduos do erro. Um bom gráfico de resíduos não possui um padrão linear e nem apresenta pontos muito dispersos, quanto mais próximos os pontos estiverem, melhor. Neste gráfico podemos observar alguns pontos mais afastados no canto superior esquerdo e no canto inferior direito, o que significa que o modelo está errando mais ao prever valores

extremos, tanto para valores baixos quanto para valores mais altos. Os valores médios, que são a maioria deles, estão bem ajustados.

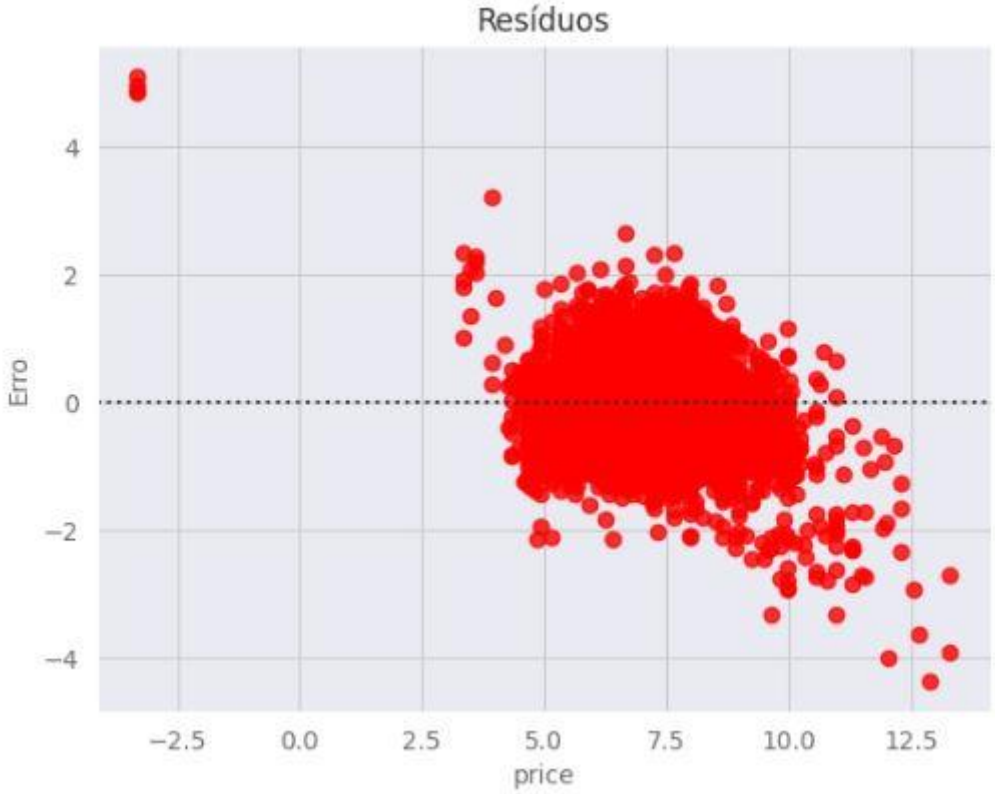


Gráfico de resíduos. O modelo acerta bem os casos médios, tendendo a ter um erro maior nos valores extremos.

- **Fazendo uma previsão com o modelo**

Supondo um imóvel com as características determinadas, foi possível testar o modelo para entender sua eficácia na previsão do 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',
 'price': 225, '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}
```

Sabemos que este imóvel localizado em Manhattan, no bairro Midtown, do tipo ‘entire room’, possui um preço de U$ 225. Após transformar e pré-processar os dados de teste da mesma

forma que pré-processamos os dados para treinar o modelo e retornarmos o valor da predição (log2) para a escala real (exp2), **a predição final foi U$ 277,82.**

Em Relação ao preço real, a previsão do modelo ficou 52,8 U$ acima do real, ou aproximadamente 19% de erro.

Evidentemente, este foi apenas um experimento. Muitos outros experimentos podem ser realizados, melhorando a qualidade da predição.

Uma forma de melhorar a qualidade do modelo é trabalhar com mais cuidado na etapa de feature enginnering, preparando features que sejam importantes para aprimorar a capacidade de aprendizado do modelo. Também é possível testar com mais ou menos variáveis e avaliar se a performance do modelo melhora.

Em Machine Learning, se há uma coisa interessante é esta possibilidade de estar sempre melhorando e testando novos métodos para melhorar a capacidade dos modelos, o que faz com que o interesse pelo projeto nunca acabe.