Maria Moura

Desafio Cientista de Dados

Relatório de Análise Estatística e Exploratória de Dados (EDA)

1. Introdução

Este relatório apresenta uma análise exploratória dos dados (EDA) relacionados a aluguéis de apartamentos em uma plataforma. O objetivo é identificar padrões, relações entre variáveis e formular insights de negócio. Todos os gráficos estão disponíveis no jupyter notebook

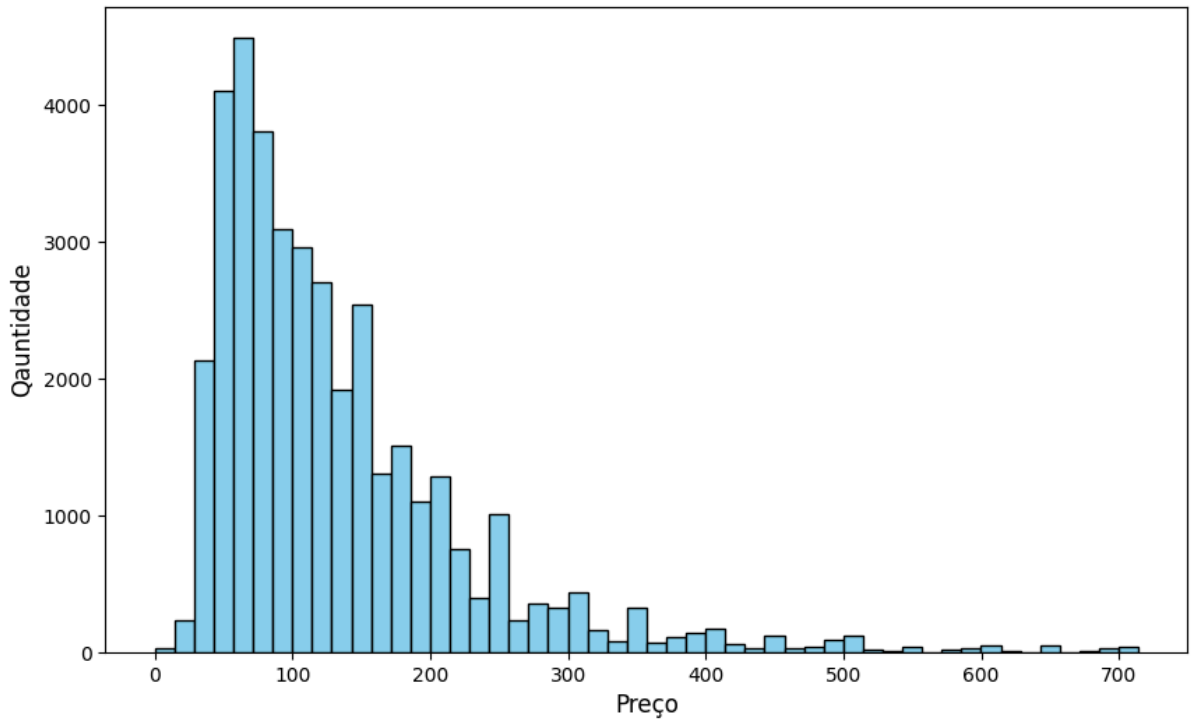
2. Análise Exploratória de Dados

2.1 Detecção e Remoção de Outliers

Utilizou-se o método do Z-score para detectar e remover outliers, garantindo que a análise fosse realizada com dados consistentes.

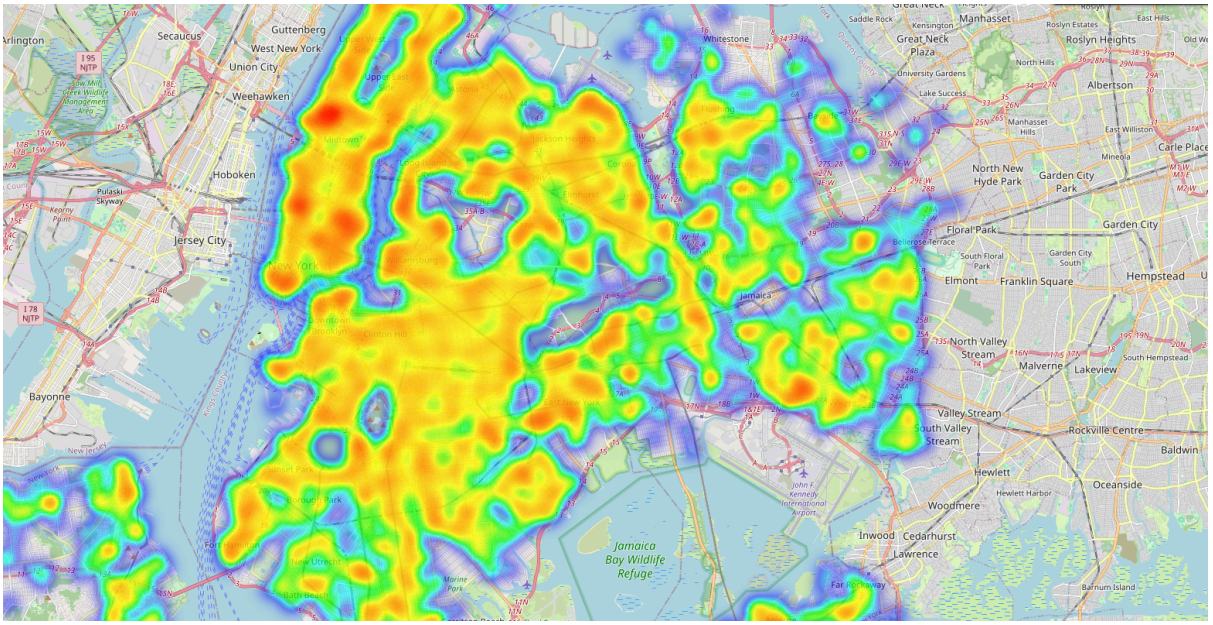
2.2 Análise de Preços

Foi identificado que os preços mais comuns variam entre **30 a 150 dolares**.



2.3 Mapa de Calor

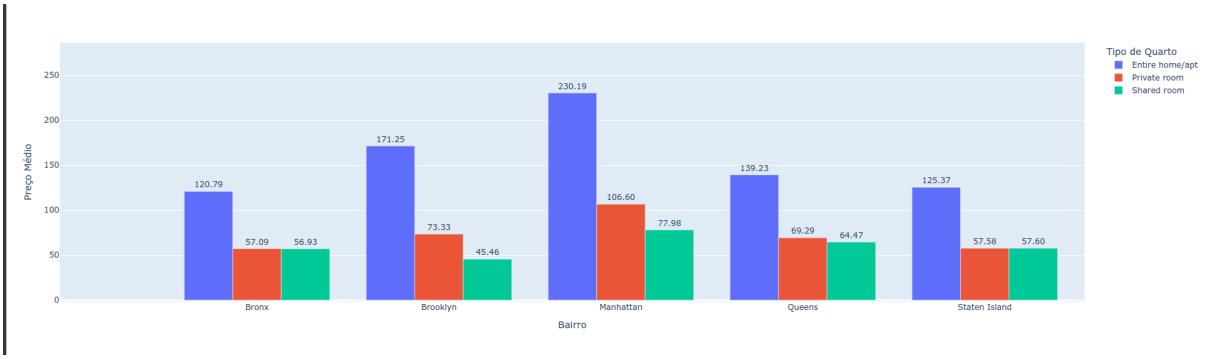
A análise por **latitude e longitude** mostrou que as regiões mais procuradas estão localizadas em **New York e Central Park**, confirmando uma alta demanda nessas áreas.



]

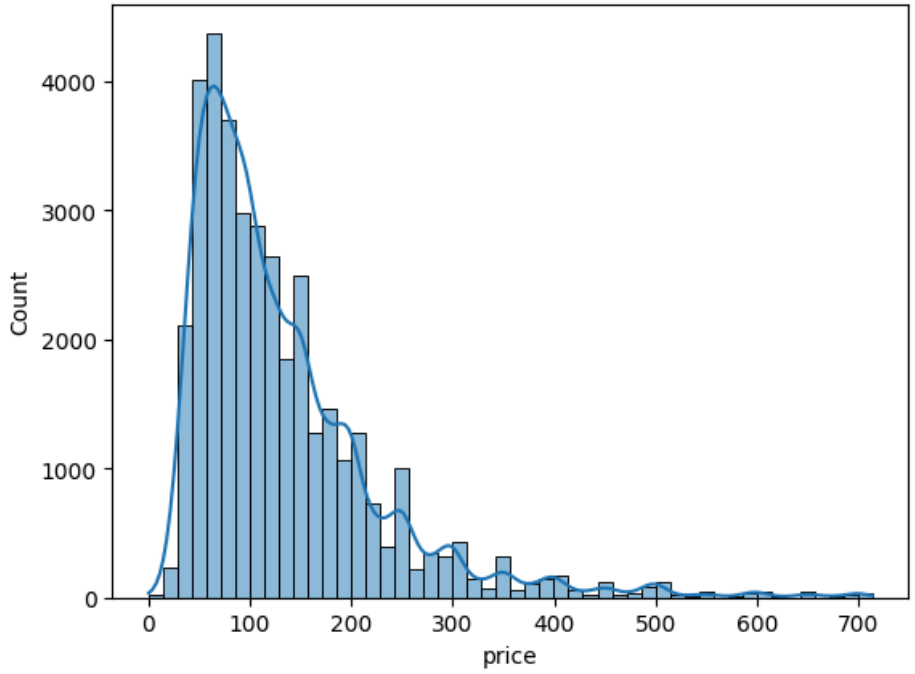
2.4 Correlação entre Tipo de Quarto, Preço e Bairro

Os dados mostram que **Manhattan** tem os maiores valores de aluguel, independentemente do tipo de acomodação (apartamento, casa, quarto privado ou compartilhado), seguido de **Brooklyn e Queens**.



2.5 Correlação entre Avaliações e Preço

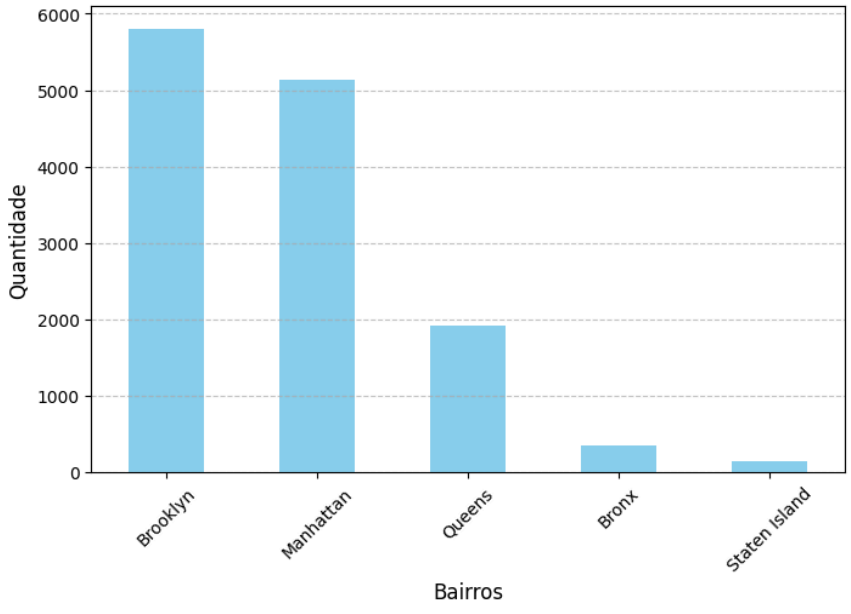
Os aluguéis com preço médio são os mais avaliados, sugerindo que os locais agradáveis e com valores acessíveis tendem a receber mais feedbacks.



3. Respostas para Perguntas Específicas

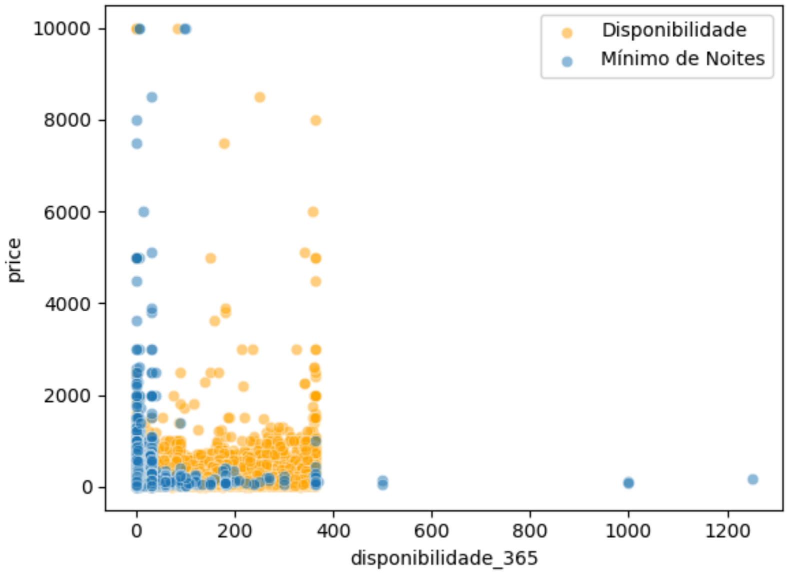
3.1 Onde investir em um apartamento para alugar?

Criamos uma função para identificar os aluguéis com avaliação acima da média e preço entre 50 a 200 dólares. **Brooklyn** lidera em número de anúncios, com quase **6.000 opções**.



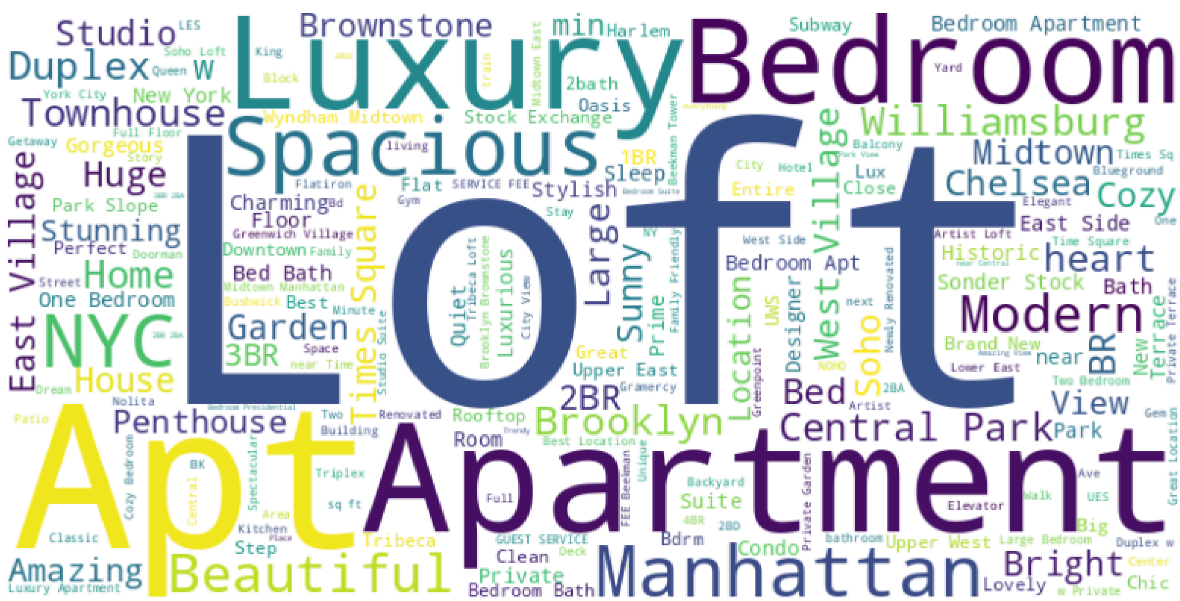
3.2 O número mínimo de noites e a disponibilidade afetam o preço?

Através de um **gráfico de dispersão**, observamos que **quanto menor o número mínimo de noites, maior é a disponibilidade do imóvel**.



3.3 Existe um padrão no nome do local para aluguéis mais caros?

Utilizando uma **nuvem de palavras** com os 10 aluguéis mais caros, notamos que termos como **"belo", "luxuoso", "espaçoso"** e nomes de bairros como **"New York" e "Manhattan"** aparecem com mais frequência.



4. Modelo Preditivo

4.1 Modelagem e Preparação dos Dados

Foram utilizadas as bibliotecas **Scikit-learn, XGBoost e Optuna** para otimização de hiperparâmetros. As variáveis independentes incluem **bairro, tipo de quarto, número de reviews, mínimo de noites, entre outras**. Variáveis categóricas foram convertidas para o tipo **"category"** para otimizar o treinamento.

4.2 Divisão dos Dados

Os dados foram divididos em **treino e teste** usando **train_test_split**, garantindo que o modelo seja avaliado com dados não vistos.

4.3 Otimização de Hiperparâmetros

O **Optuna** foi utilizado para encontrar os melhores valores para **número de estimadores, taxa de aprendizado e profundidade máxima da árvore**. O objetivo é minimizar o **erro absoluto médio (MAE)**.

4.4 Modelo Escolhido: XGBoost

O **XGBoost** foi escolhido por ser um dos algoritmos mais eficientes para problemas de regressão. Ele utiliza a técnica de **boosting**, criando várias árvores de decisão sequenciais que corrigem os erros anteriores.

4.5 Análise de Performance

A performance do modelo foi avaliada com o **erro absoluto médio (MAE)**. A análise de correlação mostrou que os dados têm pouca correlação entre si, o que pode dificultar a previsão precisa dos preços.

---

5. Conclusão

A análise exploratória indicou que **Manhattan, Brooklyn e Queens** são as regiões mais valorizadas. O modelo de predição utilizou **XGBoost** e **Optuna** para ajustar hiperparâmetros e minimizar o erro. Apesar da baixa correlação entre as variáveis, o modelo apresentou um desempenho aceitável para prever preços com base em características específicas dos aluguéis.
