

# Relatório de Análise Exploratória e Modelagem Preditiva

 [github.com/faanogueira/MachineLearning](https://github.com/faanogueira/MachineLearning)

## 1. Análise Exploratória dos Dados (EDA)

A análise exploratória foi conduzida para entender melhor a distribuição dos preços por 'bairro\_group' e 'room\_type' com o uso de tabelas e gráficos (de pizza, barra, histograma e dispersão), as correlações entre variáveis são responsáveis por identificar padrões que impactam a precificação dos imóveis. Através de matriz de correlações, podemos identificar uma correlação positiva fraca entre a disponibilidade e a quantidade de imóveis. Os demais índices são nulos ou negativos.

### ◆ Estatísticas Gerais dos Dados

Foram analisadas as seguintes estatísticas:

- Média, mediana e desvio padrão dos preços (calculados com Pandas usando `.mean()`, `.median()` e `.std()`);
- Distribuição dos tipos de acomodação;
- Disponibilidade ao longo do ano e relação com o preço.

### ◆ Principais Insights

- **Bairros mais caros:** Manhattan apresenta os preços mais altos;
- **Tipo de acomodação:** Apartamentos inteiros têm valores significativamente mais altos do que quartos compartilhados;
- **Mínimo de noites:** Valores altos podem impactar a atratividade do anúncio.

## 2. Respostas às Perguntas

### a) Onde seria mais indicada a compra de um imóvel para alugar?

- Após a análise identificamos a região de Manhattan com o maior potencial de retorno financeiro, pois os preços médios são mais altos.

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

- Sim. Os imóveis com maior número mínimo de noites tendem a ter preços mais baixos, já os imóveis com menor mínimo de noites são inversamente proporcionais.
- A disponibilidade de 365 dias tende a estar associada a preços mais baixos, sugerindo maior concorrência.

### c) Existe algum padrão no nome dos locais de alto valor?

- Palavras como "Luxury", "Penthouse", "Exclusive" aparecem com maior frequência nos imóveis de maior preço.

## 3. Explicação da Modelagem

O *Random Forest Regressor* foi utilizado por trazer precisão e eficiência do modelo para prever a demanda e ajudar na tomada de decisões estratégicas. Após realizar a Análise Exploratória e a coleta de dados, removi os outliers com valores extremos (<25% e >75%), selecionei as variáveis e separei por características e alvo. Dividi o pipeline para treino e teste e finalizei com a previsão do preço.

### ◆ Variáveis utilizadas:

- Localização (bairro\_group, bairro, latitude, longitude)
- Tipo de acomodação (room\_type)
- Restrições de aluguel (minimo\_noites, disponibilidade\_365)
- Popularidade (numero\_de\_reviews)

### ◆ Modelo escolhido: Random Forest Regressor

- Prós: Lida bem com dados não-lineares e outliers, além de ser mais preciso e eficiente.
- Contras: Pode ser mais lento que modelos mais simples.

### ◆ Métrica de avaliação:

- **RMSE (Raiz do Erro Quadrático Médio)** foi escolhida para penalizar grandes desvios na previsão.

## 4. Previsão do Preço para o Imóvel Exemplo

Para o apartamento "**Skylit Midtown Castle**", o modelo previu um preço de **\$191** pois esse valor representa bem a mediana do mercado e evita possíveis distorções causadas por valores extremos.

Para sugerir um preço para o apartamento com essas características, podemos usar a média e a mediana dos preços de imóveis que pertencem ao mesmo grupo de bairro ('Manhattan') e ao mesmo tipo de acomodação ('Entire home/apt').