Relatório da Análise Estatística e Exploratória de Dados (EDA) e Modelagem Preditiva

1 Introdução

Este relatório apresenta uma análise exploratória de dados (EDA) e modelagem preditiva com o objetivo de identificar padrões relevantes e prever preços no mercado de aluguel de apartamentos em Nova York. A análise abrange características dos preços, disponibilidade e variáveis textuais, buscando responder a questões de negócio que podem orientar futuros investimentos.

2 Processamento inicial dos dados

2.1 Carregamento e limpeza

- Os dados foram carregados a partir de um arquivo CSV contendo **48.894** registros.
- Após a remoção de registros com valores nulos na coluna `price` e preços menores ou iguais a zero, o dataset foi reduzido para **48.883** registros.

2.2 Divisão dos dados

- O conjunto de dados foi dividido em 80% para treino e 20% para teste.
- O conjunto de treino contém **39.106** registros e o conjunto de teste **9.777** registros.

3 Análise estatística dos preços

3.1 Estatísticas descritivas

Preço máximo: $10.000

Preço mínimo: $10

Preço médio: $152,76

3.2 Distribuição dos preços

- Remoção de outliers acima do 99º percentil para uma análise mais precisa.
- A distribuição apresenta uma assimetria à direita, indicando a presença de algumas propriedades com preços muito altos.

3.3 Média vs. Mediana

- A média dos preços é superior à mediana, sugerindo que os outliers afetam significativamente o preço médio.

4 Análise das variáveis categóricas

4.1 Tipo de quarto (`room type`)

- A distribuição dos tipos de quartos mostra predominância de **Private Room** e **Entire Home/Apt**, com menor presença de **Shared Room**.

4.2 Localização (`bairro group`)

- Análise de frequência dos bairros revela que **Manhattan** possui maior oferta e apresenta os preços mais altos, seguido por **Brooklyn**.

5 Análise de correlação

5.1 Disponibilidade e Preço

- A correlação entre o número de noites mínimas e o preço é fraca, indicando que a exigência de estadias mínimas não impacta significativamente o preço cobrado.
- Propriedades com maior disponibilidade anual tendem a ter preços mais baixos, sugerindo que anfitriões priorizam a ocupação constante em detrimento de tarifas mais altas.

5.2 Padrões nos títulos dos anúncios

- A análise textual confirmou que títulos contendo palavras como *luxury*, *spacious* ou *central* estão consistentemente associados a preços mais elevados, refletindo a percepção de maior valor agregado por parte dos locatários.

6 Modelagem preditiva

6.1 Preparação dos Dados para o Modelo

- As variáveis categóricas foram codificadas utilizando **One-Hot Encoding**.
- Variáveis numéricas foram mantidas em seus valores originais.

6.2 Seleção do modelo

- O modelo escolhido para prever os preços foi o **Random Forest Regressor**, devido à sua capacidade de lidar com dados não-lineares e múltiplas variáveis categóricas.
- O modelo foi treinado com hiperparâmetros padrão.

6.3 Treinamento e avaliação

- O modelo foi treinado com o conjunto de treino e avaliado no conjunto de teste.
- O desempenho foi medido utilizando o **Erro Quadrático Médio (RMSE)** e o $\mathbf{R^2}$.

6.4 Resultados do modelo

- **RMSE:** 73,08
- **MAE:** 44,65
- $\mathbf{R^2}$**:** 0,51
- O modelo explicou 51
- As variáveis mais importantes para o modelo foram: **localização**, **tipo de quarto** e **disponibilidade anual**.

7 Hipóteses de negócio

7.1 Localização ideal para investimento

- Manhattan aparece como a região com o preço médio mais alto:

  Preço médio em Manhattan: $\$196,88$

  Brooklyn: $\$124,44$

  Queens: $\$99,51$

  Bronx: $\$87,57$

  Staten Island: $\$114,81$

- **Conclusão:** Se o objetivo é maximizar o preço do aluguel, investir em Manhattan parece ser a melhor escolha. No entanto, considere também a ocupação e a competitividade do mercado na região. Brooklyn pode ser uma boa alternativa, com preços mais acessíveis para compra e ainda com boa rentabilidade.

7.2 Impacto da disponibilidade e número mínimo de noites

- Correlação entre preço e número mínimo de noites: 0,043
- Correlação entre preço e disponibilidade: 0,082
- Ambas as correlações são positivas, mas muito fracas. Isso sugere que:
  - Aumentar o número mínimo de noites ou a disponibilidade ao longo do ano tem pouco impacto direto no preço.
  - O modelo de regressão OLS também confirma:
    - Coeficiente para número mínimo de noites: 0,5012 (a cada noite mínima adicional, o preço sobe cerca de $0,50 em média).
- **Conclusão:** Embora haja uma ligação estatística, ela é fraca. Fatores como localização e tipo de quarto têm impacto muito maior.

7.3 Importância do título do anúncio

- Palavras mais comuns em locais caros:
  - 'luxury', 'loft', 'village', 'park', 'spacious', 'manhattan'
- Palavras mais comuns em locais baratos:
  - 'room', 'private', 'cozy', 'brooklyn', 'bushwick', 'sunny'
- **Conclusão:** Listagens caras tendem a usar palavras como *luxury*, *loft*, *village*, o que transmite uma imagem de sofisticação. Listagens baratas frequentemente usam termos como *room*, *cozy*, e referenciam bairros como Brooklyn e Bushwick. Para quem quer atrair mais inquilinos dispostos a pagar mais, usar descrições que destacam luxo, espaço e localização premium pode ser eficaz.

8 Conclusão

A análise identificou padrões importantes que podem guiar decisões de investimento. Localização, disponibilidade e estratégias de marketing (como a escolha de títulos) são fatores chave para maximizar o retorno sobre o investimento. O modelo de **Random Forest Regressor** demonstrou ser eficaz na previsão de preços, oferecendo uma ferramenta valiosa para precificação automatizada.