Relatório de Análise Exploratória e Modelagem

Akyla de Aquino Pinto

Email: akylaaquino@gmail.com

---

1. Introdução

Este relatório apresenta os principais resultados obtidos durante a análise exploratória dos dados e o desenvolvimento dos modelos preditivos. O objetivo deste documento é fornecer uma visão geral dos dados, identificar padrões e apresentar os resultados da modelagem realizada.

---

2. Análise Exploratória

2.1. Descrição dos Dados

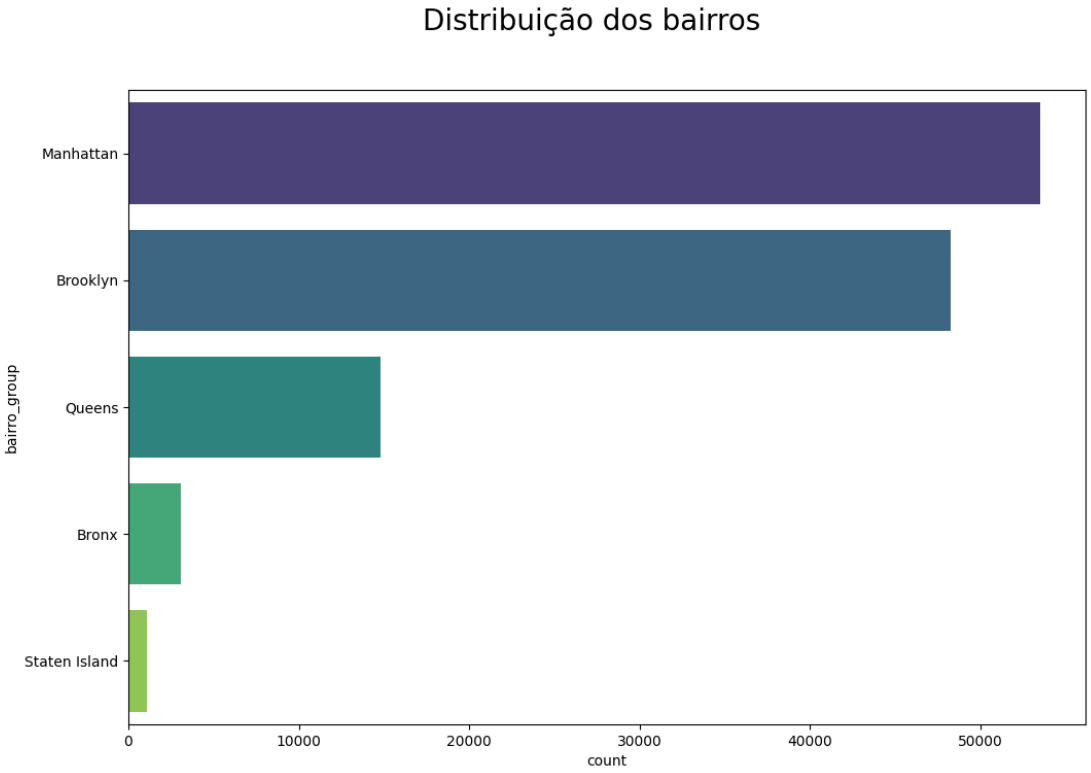
O nosso dataset de análise trata informações de Airbnb's de Nova York, nele teremos variáveis categóricas e numéricas, totalizando 16 variáveis, onde 6 delas são categóricas e 10 numéricas. Foi concatenado um dataset mais atual, baixado do site oficial que disponibiliza o dataset, Inside Airbnb.

Nesta etapa, foram verificadas as características dos dados, tais como a presença de valores ausentes, outliers e a distribuição das variáveis. Essa análise foi fundamental para identificar os passos necessários para a limpeza e o pré-processamento dos dados.

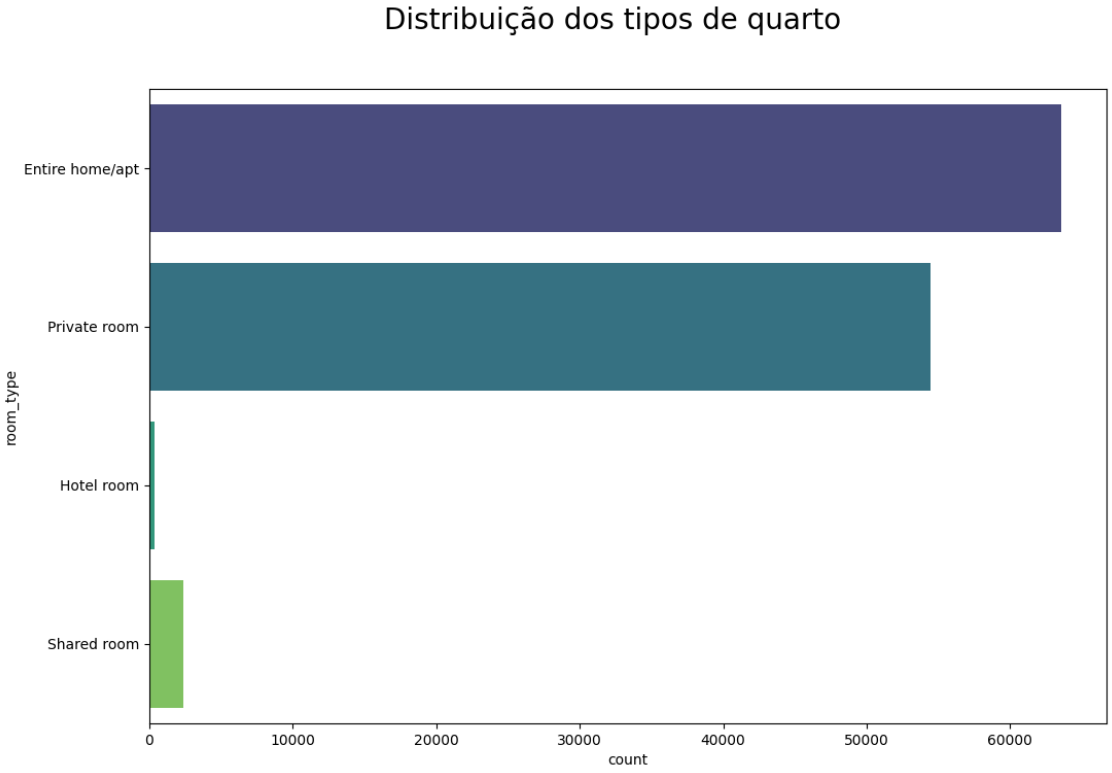
2.2. Visualizações e Estatísticas

Foram elaborados diversos gráficos para melhor compreender os dados:

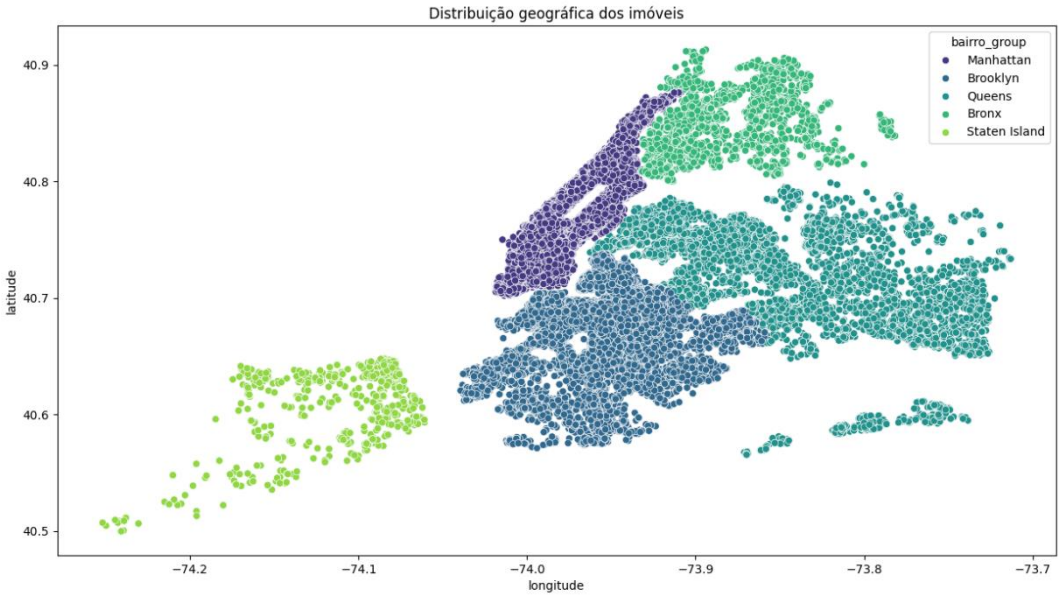
- Distribuição dos imóveis quanto ao bairro_group:



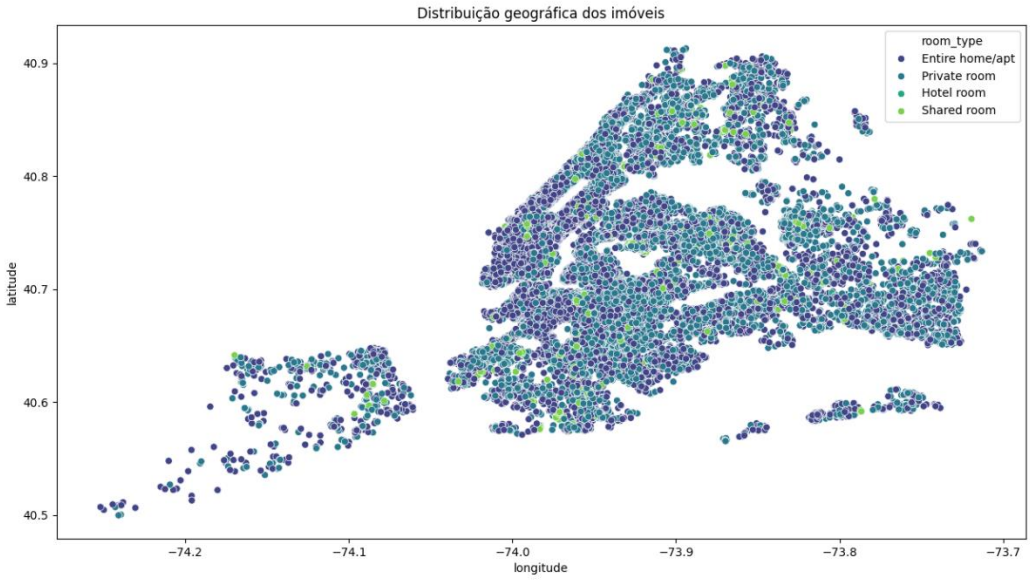
- Distribuição dos imóveis quanto a tipo de quarto:



- Distribuição geográfica quanto ao bairro_group:



- Distribuição geográfica quanto ao tipo de quarto:



As estatísticas descritivas também foram calculadas (média, mediana, desvio padrão, etc.), possibilitando uma visão inicial dos comportamentos dos dados. Há uma grande discrepância nos dados quanto aos preços, sendo alguns bem altos, isso influência em como os modelos de IA funcionam, para isso foram realizadas algumas abordagens:

- **Coluna média_bairro**: essa coluna foi criada para ajudar o modelo a se adaptar aos dados, como há um grande número de bairros, mais de 200, essa coluna serve para dar a média de preço de cada bairro.
- **Coluna price_log:** Essa coluna foi criada para normalizar os preços, devido a diferença de locais de luxo terem um grande preço, foi feita uma normalização utilizando logaritmo. Posteriori é feito uma ação inversa após a regressão realizada pelo modelo.
- **Colunas categóricas remanescentes:** Foi realizado One-Hot-Encoding nas demais coluans categóricas para o modelo não lidar com dados categóricos.

---

3. Modelagem

3.1. Modelos Testados

Diferentes algoritmos foram avaliados para prever a coluna price_log:

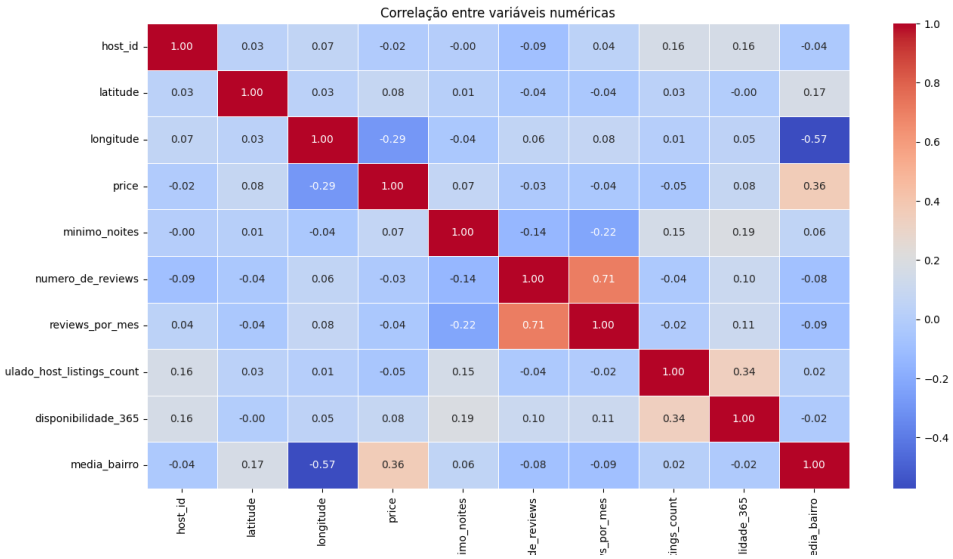
- **Regressão Lasso:** ótima para realizar regressão no qual queira evitar overfiting.
- **Random Forest:** Geralmente quando o foco é no desempenho e precisão do modelo, esse modelo teve o melhor resultado, mas devido ao seu grande peso (500 mb), não foi possível realizar commit ao github.
- **Light GBM e XGBoost:** Modelo baseado em árvores de decisão em conjunto. Parecido com random forest, mas com menor peso de memória. O XGBoost foi o segundo melhor modelo segundo as métricas realizadas, portanto, usamos ele como base para testar os dados de teste enviados.

- **Métodos de avaliação dos modelos:**
    - Erro Médio Absoluto (MAE): Média das diferenças absolutas dos valores previsto e os valores reais.
    - Erro Quadrático Médio (MSE): Médis dos quadrados das diferenças entre os valores reais e os previstos.
    - Raiz do Erro Quadrático Médio (RMSE): Raiz quadrática do MSE, ajuda a trazer o erro para escala dos valores originais.
    - R2 score: Mede a proporção da variância dos dados, sendo o melhor caso próximo de 1 e pior próximo de 0.

3.2. Pré-processamento e Seleção de Variáveis

Antes da modelagem, foram realizadas etapas de:

- Limpeza dos dados (tratamento de valores ausentes e outliers)
- Normalização/Padronização das variáveis
- Seleção das variáveis mais relevantes para a construção dos modelos utilizando a correlação de Pearson, a maioria das variáveis tem correlação, com a coluna price bem próxima de 0, ou seja, não afetam muito o valor em si. Nesse contexto foi decidido seguir com todas as variáveis (menos as retiradas anteriormente).



3.4. Resultados Obtidos

O modelo que apresentou o melhor desempenho foi o **[nome do modelo]**, com os seguintes resultados:

- **MAE:**
  - Regressão Lasso: 0.42
  - LGBM: 0.32
  - Random Forest: 0.21
  - XGBoost: 0.31
- **MSE:**
  - Regressão Lasso: 0.31
  - LGBM: 0.20
  - Random Forest: 0.11
  - XGBoost: 0.19
- **RMSE:**
  - Regressão Lasso: 0.56
  - LGBM: 0.45

    - Random Forest: 0.33
    - XGBoost: 0.43
- **R2:**
    - Regressão Lasso: 0.39
    - LGBM: 0.60
    - Random Forest: 0.78
    - XGBoost: 0.63

---

4. Conclusão

A análise exploratória permitiu identificar os principais padrões e comportamentos dos dados, enquanto a modelagem indicou que o modelo **Random Forest** é o mais adequado para a tarefa proposta. Recomenda-se:

- Realizar ajustes finos (tuning) no modelo selecionado.
- Implementar monitoramento contínuo do desempenho em cenários reais.

---

5. Próximos Passos

1. **Otimização do Modelo:** Ajuste dos hiperparâmetros e validação cruzada.
2. **Integração:** Implementação do modelo em ambiente de produção.
3. **Monitoramento:** Acompanhamento periódico do desempenho do modelo.
4. **Atualização dos Dados:** Inclusão de novos dados para refinar o treinamento do modelo.

Respondendo perguntas

1 - Supondo que uma pessoa esteja pensando em investir em um apartamento para alugar na plataforma, onde seria mais indicada a compra?

R: Uma resposta bastante abrangente, seria mais indicado obter mais informações quanto aos orçamentos que ela pretende gastar ou outras informações precisas. Mas nesse caso a melhor recomendação seria ver qual o bairro que tem em média mais disponibilidade de apartamentos por ano. Que no caso seriam os bairros Fort Wadsworth, Chelsea, Staten Island e Eastchester, onde Fort Wadsworth está disponível 365 dias do ano.

2 - O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?

R: Sim, pois o valor mínimo a ser pago depende do numéro mínimo de noites a ser pago e da disponibilidade do ano.

3 - Existe algum padrão no texto do nome do local para lugares de mais alto valor?

R: Realizando uma wordcloud podemos ver as palavras que mais aparecem nos nomes, que no caso são: Apartment, NYC, Manhattan, Beautiful e outros...

4 - Explique como você faria a previsão do **preço** a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê? Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?

R:

- Como queremos prever uma variável, estamos lidando com uma regressão.
- Utilizei transformações de One-Hot-Encoding em variáveis categóricas para os modelos de IA conseguirem lidar melhor com os dados, criei uma coluna com o logarítmo de price para normalizar devido os preços de apartamentos de luxo serem muito altos, isso ajuda o modelo a não ter discrepância na previsão.
- O melhor modelo foi o Random Forest, ótimos resultados no geral e boa pontuação nas métricas, porém, muito pesado para carregamento.
- Foram escolhidas MAE,MSE, RMSE e R2, como queremos analisar regressão então precisamos lidar com a diferença dos dados reais vs os previstos pelos modelos.

5 - Supondo um apartamento com as seguintes características,

{'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',

'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}

Qual seria a sua sugestão de preço?

R: Previsão do modelo deu 283