Desafio Cientista de Dados

Introdução

Este foi um trabalho / desafio proposto no programa Lighthouse da Indicium.

O programa procura testar conhecimentos dos conceitos estatísticos de modelos preditivos, criatividade na resolução de problemas e aplicação de modelos básicos de machine learning.

O desafio era:

> ”Você foi alocado(a) em um time da Indicium que está trabalhando atualmente junto a um cliente no processo de criação de uma plataforma de aluguéis temporários na cidade de Nova York. Para o desenvolvimento de sua estratégia de precificação, pediu para que a Indicium fizesse uma análise exploratória dos dados de seu maior concorrente, assim como um teste de validação de um modelo preditivo.
>
> Seu objetivo é desenvolver um modelo de previsão de preços a partir do *dataset* oferecido, e avaliar tal modelo

> utilizando as métricas de avaliação que mais fazem sentido para o problema.

⚠ **Você encontra o trabalho completo no repositório de entrega no GitHub → ACESSE AQUI**

Objetivos

- Faça uma análise exploratória dos dados (EDA), demonstrando as principais características entre as variáveis e apresentando algumas hipóteses de negócio relacionadas.
- Responda também às seguintes perguntas:
    1. Supondo que uma pessoa esteja pensando em investir em um apartamento para alugar na plataforma, onde seria mais indicada a compra?
    2. O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?
    3. Existe algum padrão no texto do nome do local para lugares de mais alto valor?
- Inclui mais 2 perguntas por conta
    1. Imóveis com mais reviews têm preços mais altos?
    2. Imóveis com alta disponibilidade têm preços mais baixos?
- 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?
- Supondo um apartamento com as seguintes características, qual seria a sugestão de 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',
 '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}
```

- Salve o modelo desenvolvido no formato .pkl.
- A entrega deve ser feita através de um repositório de código público que contenha:
    1. README explicando como instalar e executar o projeto
    2. Arquivo de requisitos com todos os pacotes utilizados e suas versões
    3. Relatórios das análises estatísticas e EDA em PDF, Jupyter Notebook ou semelhante conforme passo 1 e 2.
    4. Códigos de modelagem utilizados no passo 3 (pode ser entregue no mesmo Jupyter Notebook).
    5. Arquivo .pkl conforme passo 5 acima.

▼ Relatório da EDA

Introdução

A análise exploratória de dados, ou EDA, é uma etapa muito importante no processo de ciência de dados, principalmente em projetos de modelos de previsão.

Antes de elaborar qualquer análise ou modelo precisamos estudar o dataset, identificar sua estrutura, padrões e incosistências e, com base nisso, gerar objetivos e perguntas que serão relevantes para a identificação e tradução de insights. Essas perguntas são resolvidas ao longo do desenvolvimento do trabalho.

Como estamos pensando em previsão de preços, **este é um problema de regressão.**

Análise

Temos um dataset com 48.894 linhas e 16 colunas, composto da seguinte forma:

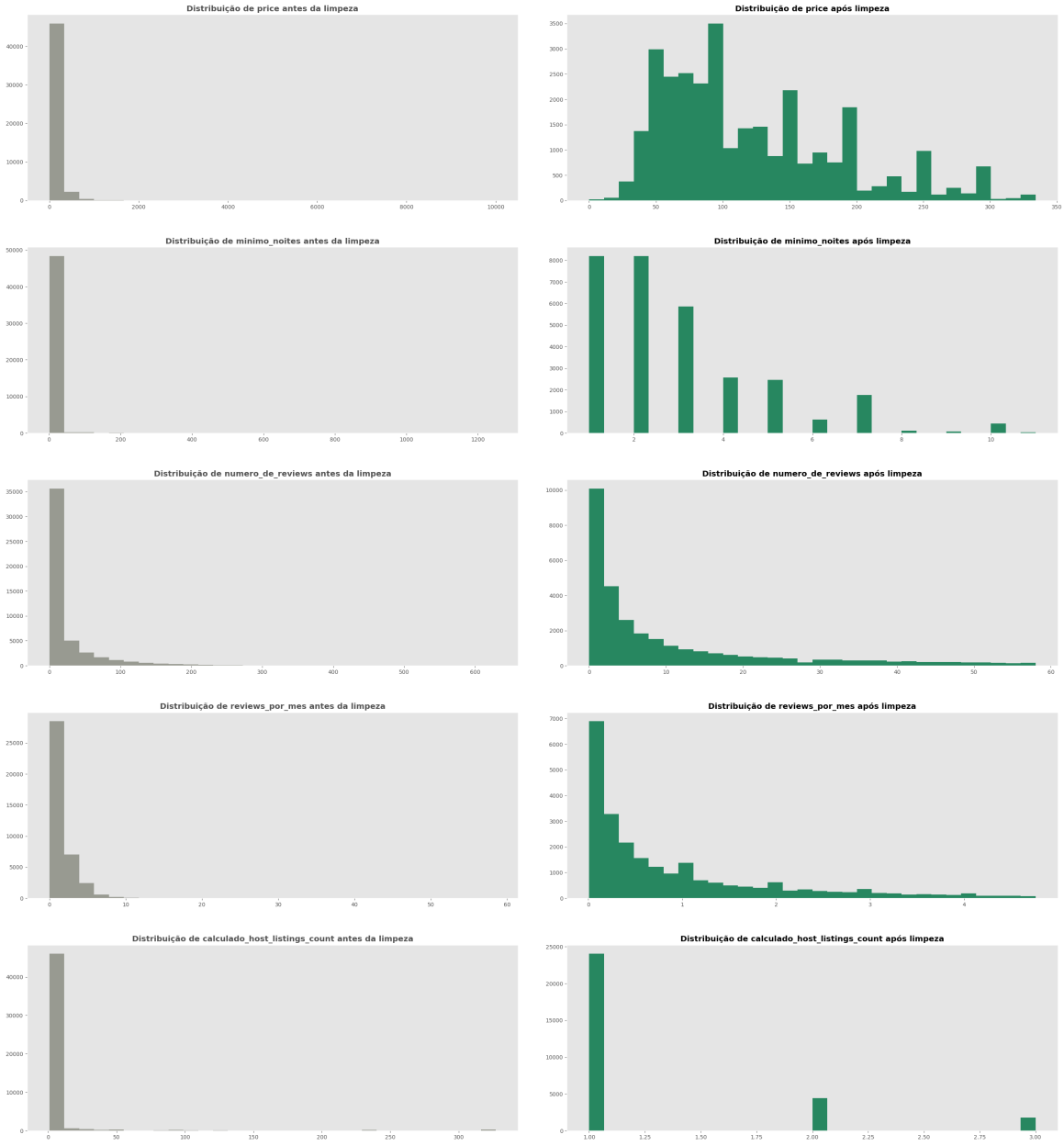
```
#   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              48894 non-null  int64
 1   nome                            48878 non-null  object
 2   host_id                         48894 non-null  int64
 3   host_name                       48873 non-null  object
 4   bairro_group                    48894 non-null  object
 5   bairro                          48894 non-null  object
 6   latitude                        48894 non-null  float64
 7   longitude                       48894 non-null  float64
 8   room_type                       48894 non-null  object
 9   price                           48894 non-null  int64
 10  minimo_noites                   48894 non-null  int64
 11  numero_de_reviews               48894 non-null  int64
 12  ultima_review                   38842 non-null  object
 13  reviews_por_mes                 38842 non-null  float64
 14  calculado_host_listings_count   48894 non-null  int64
 15  disponibilidade_365             48894 non-null  int64
```

Verificação de ausentes, duplicatas e outliers:

- As colunas reviews_por_mes e ultima_review apresentam 20,56% de valores ausentes.
- A coluna host_name apresenta 0,04% de valores ausentes.

- A coluna nome apresenta 0,03% dos valores ausentes.
- Não há duplicatas no conjunto de dados.
- 5 das 6 variáveis apresentam outliers
- A maioria das variáveis tem distribuição assimétrica à direita, tendo a maioria dos valores concentrados em números baixos.
- O alto desvio padrão em várias colunas indica grande variabilidade nos dados.

Distribuição das variáveis categóricas

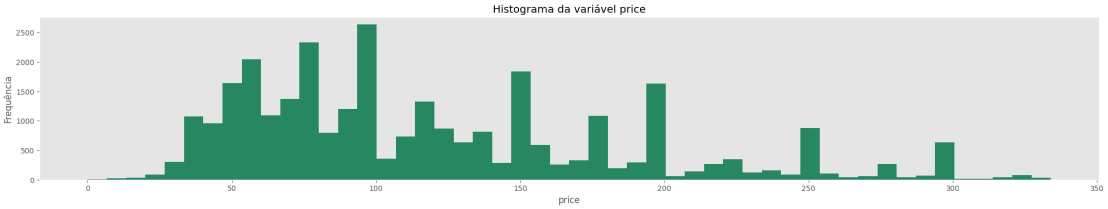


Distribuição das variáveis numéricas após limpeza. Fonte: Autor

- Verificamos que existiam outliers e que as distribuições não são normais.

▼ Análise Univariada

price

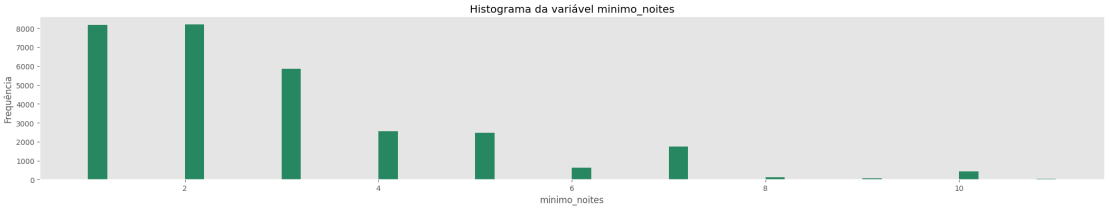


Distribuição da variável numérica price após limpeza. Fonte: Autor

- A maior parte dos dados se concentra entre 50 e 150, indicando que a maioria dos preços se encontra nessa faixa, o que sugere que a maioria dos produtos ou serviços são mais baratos.
- A cauda direita do histograma é mais longa que a esquerda, indicando uma assimetria positiva. Isso significa que há alguns valores de preço muito altos que "puxam" a média para cima, enquanto a maioria dos valores está concentrada em uma faixa mais baixa.

---

minimo_noites

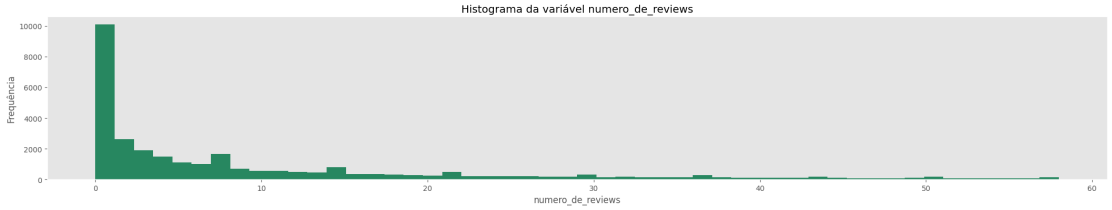


Distribuição da variável numérica minimo_noites após limpeza. Fonte: Autor

- A maioria das propriedades exige um número mínimo de noites relativamente baixo. Isso indica demanda de estadias curtas.
- Distribuição assimétrica à direita
- A média do número mínimo de noites é de aproximadamente 2,87 noites
- O desvio padrão é de 1,95 noites. Isso indica que há uma variação considerável no número mínimo de noites exigido pelas propriedades.
- Foi feito o teste estatísticos de chi-quadrado para esta variável. Há forte evidência de que há uma associação significativa entre o número mínimo de noites e o preço.

---

numero_de_reviews



Distribuição da variável numérica numero_de_reviews após limpeza. Fonte: Autor

- A maioria das propriedades possui um número relativamente baixo de avaliações, com um pico nas primeiras faixas.
- O histograma apresenta distribuição assimétrica à direita
- A média do número de reviews é de aproximadamente 9,97
- O desvio padrão é de 13,44. Isso indica que há uma variação considerável no número de reviews entre as propriedades.
- O teste estatístico sugere uma forte evidência de que há uma associação significativa entre o número de reviews e o preço.

---

reviews_por_mes

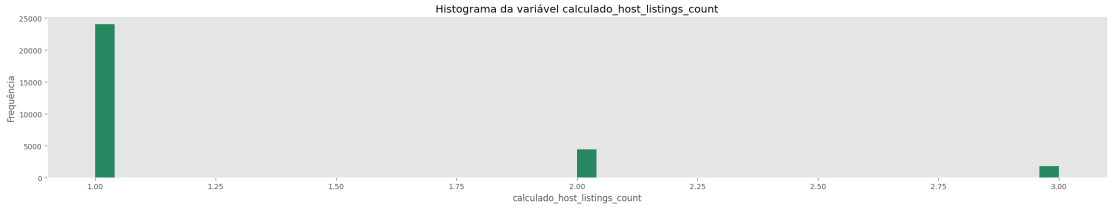
Distribuição da variável numérica reviews_por_mes após limpeza. Fonte: Autor

Nota-se que:

- A maioria das propriedades recebe um número relativamente baixo de reviews por mês.
- A distribuição é assimétrica à direita.
- A média de reviews por mês é de aproximadamente 0,89. Isso significa que, em média, as propriedades recebem menos de 1 review por mês.
- O teste estatístico sugere uma forte evidência de que há uma associação significativa entre o número de reviews por mês e o

preço.

calculado_host_listing_counts

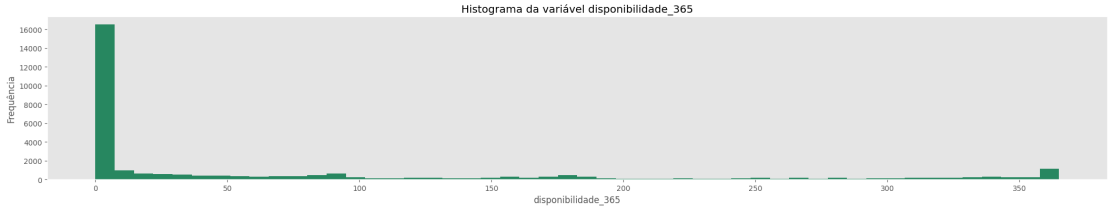


Distribuição da variável numérica calculado_host_listing_counts após limpeza. Fonte: Autor

Nota-se que:

- A grande maioria das observações está concentrada em 1.0.
- A média é de aproximadamente 1,27. Isso significa que, em média, os anfitriões possuem 1,27 listagens.
- O desvio padrão é relativamente baixo. Isso indica que a maioria dos anfitriões tem um número pequeno de listagens.
- O teste estatístico retornou que não há evidências suficientes para afirmar que anfitriões com múltiplas listagens cobram preços significativamente diferentes daqueles com apenas uma listagem

disponibilidade_365



Distribuição da variável numérica disponibilidade_365 após limpeza. Fonte: Autor

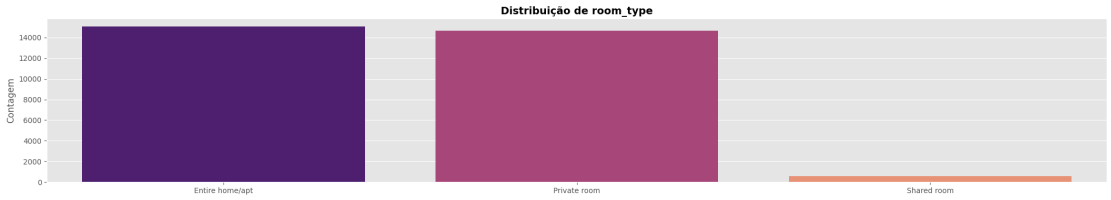
Nota-se que:

- A maioria dos imóveis tem baixa disponibilidade ao longo do ano.
- A distribuição é altamente assimétrica à direita, com uma longa cauda. Isso indica que há um grande número de imóveis com baixa disponibilidade e alguns com disponibilidade muito alta.

- A média de disponibilidade é de 72 dias no ano
- 50% dos imóveis tem um dia disponível ou menos
- Há evidências suficientes para afirmar que existe uma relação significativa entre a faixa de disponibilidade e a faixa de preço dos imóveis.

▼ Distribuição das variáveis categóricas

room_type

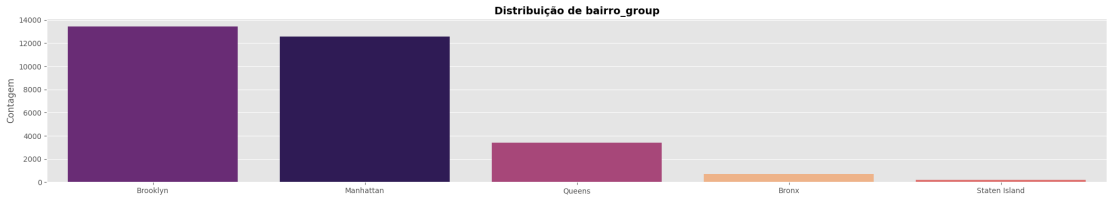


Distribuição da variável categórica room_type. Fonte: Autor

O gráfico acima apresenta a distribuição dos tipos de quartos do dataset.

- Apartamentos e casas no seu todo são o tipo mais comum, seguido de quartos privativos e quartos compartilhados.

bairro_group

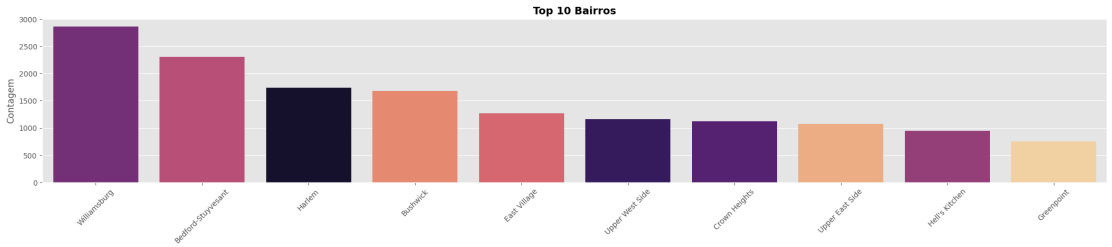


Distribuição da variável categórica bairro_group. Fonte: Autor

O gráfico acima apresenta a distribuição dos bairros onde os anúncios estão localizados.

- Manhattan é o bairro com maior número de anúncios, seguido do Brooklyn e do Queens.

bairro

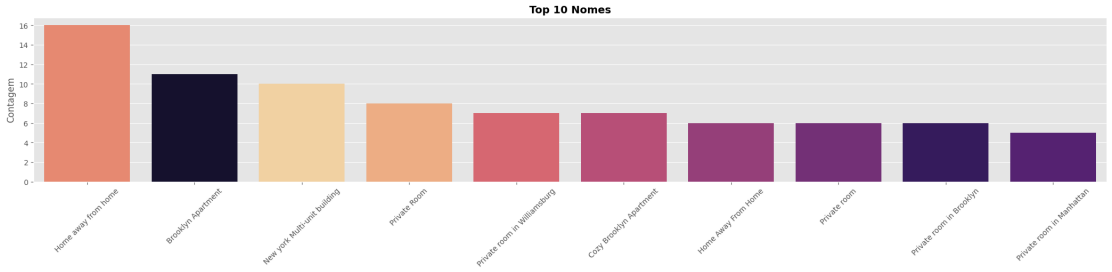


Distribuição da variável categórica bairro. Fonte: Autor

O gráfico acima apresenta a distribuição das áreas onde os anúncios estão localizados.

- Williamsburg é a área com maior número de anúncios, seguido de Bedford-Stuyvesant e do Harlem.

nome



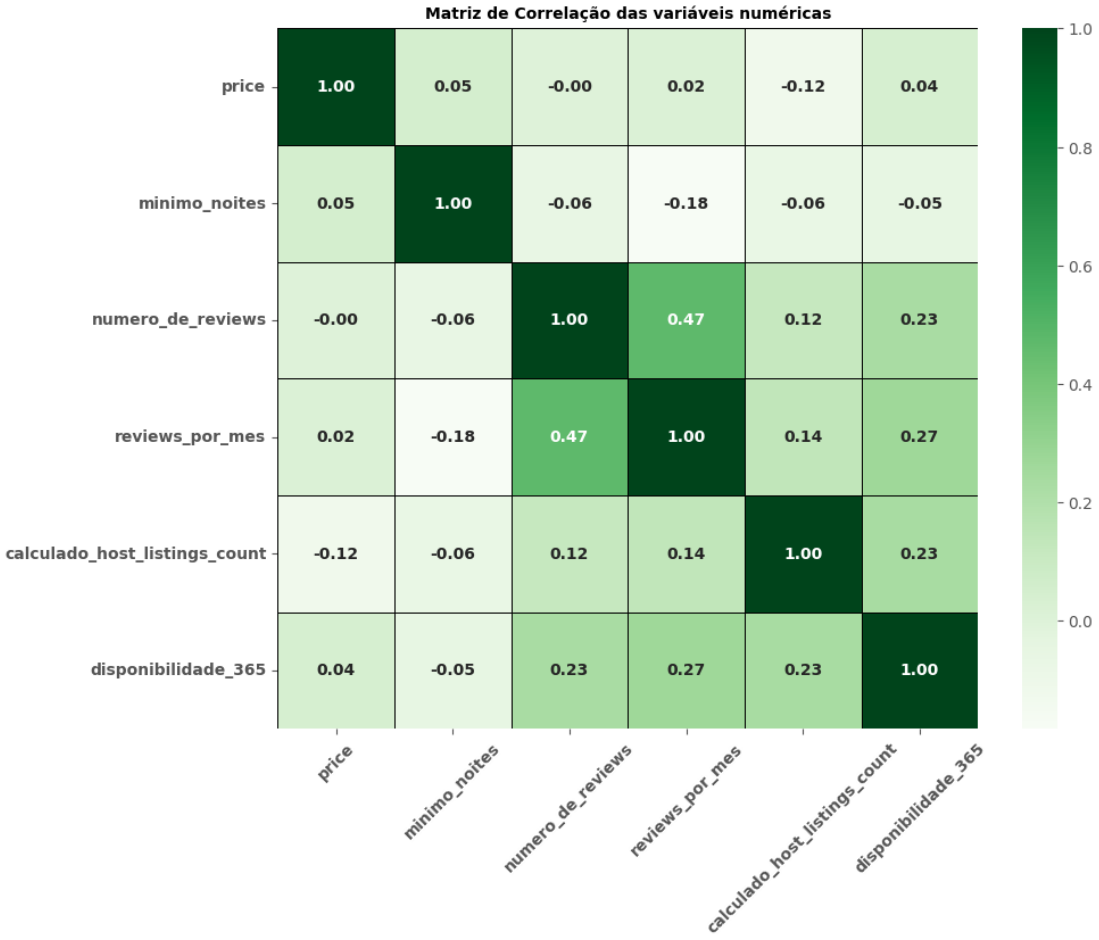
Distribuição da variável categórica nome. Fonte: Autor

O gráfico acima apresenta a distribuição das áreas onde os anúncios estão localizados.

- Home away from home tem o maior número de anúncios, seguido de New York Multi-unit building e Brooklyn Apartment.

▼ Correlação entre as variáveis

Correlação entre as variáveis numéricas



Correlação entre as variáveis numéricas. Fonte: Autor

Os resultados da correlação e do heatmap acima apresentam os níveis de relação entre as variáveis numéricas do DataFrame. Se nota:

- **price**
  - Tem uma correlação negativa muito fraca com todas as variáveis, indicando que anúncios de hosts com várias listagens podem ter preços ligeiramente menores.
  - Correlações com as outras variáveis são próximas de zero, indicando que o preço não é fortemente influenciado por estas variáveis, mas sim, por algum outro fator externo.
- **minimo_noites**
  - As noites mínimas têm pouca influência sobre as outras variáveis.
  - As correlações são muito fracas, indicando pouca relação com as demais variáveis.
- **numero_de_reviews**

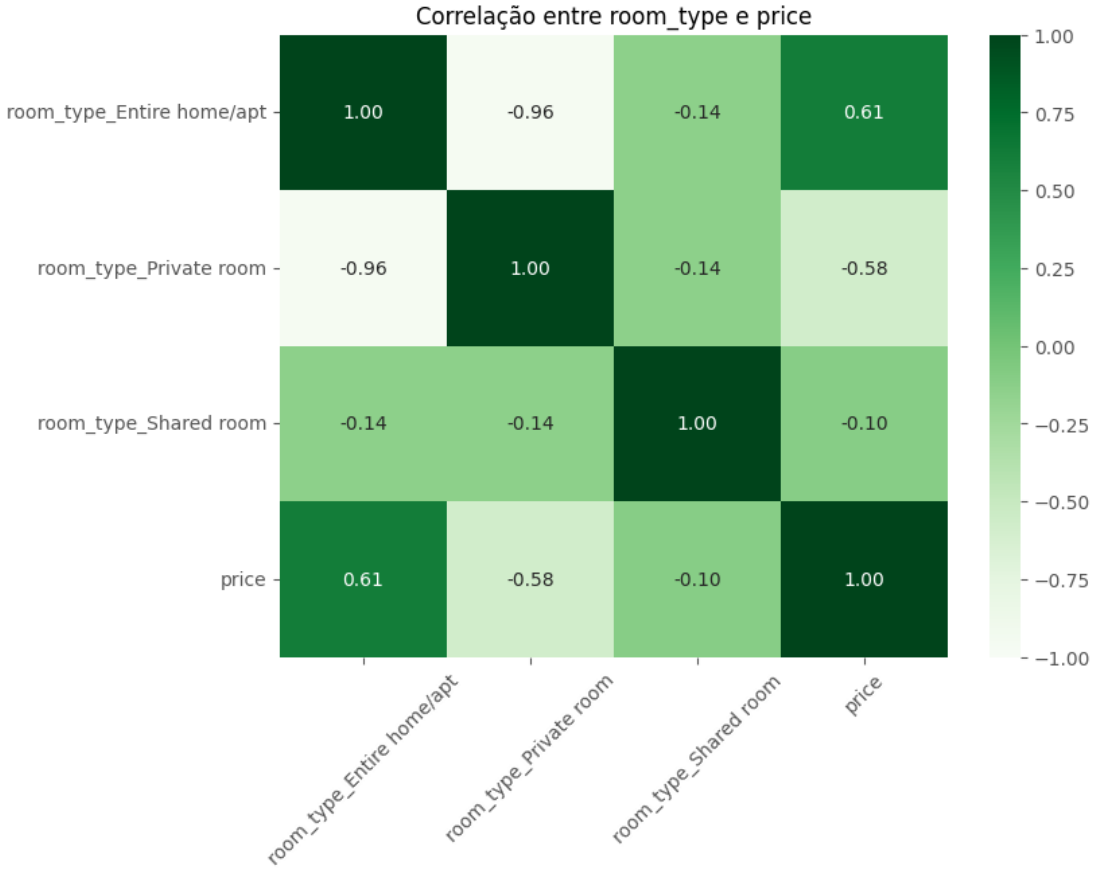
    - Tem uma correlação correlação positiva muito fraca com disponibilidade_365 (0,23), o que sugere que anúncios com mais avaliações tendem a ter mais disponibilidade por ano.
    - Anúncios mais ativos (com mais avaliações) tendem a estar pouco mais disponíveis.
- **reviews_por_mes**
    - Tem uma correlação positiva moderada com numero_de_reviews, sugerindo que anúncios com mais avaliações totais recebem mais avaliações por mês
- **calculado_host_listings_count**
    - Tem uma correlação positiva muito fraca com disponibilidade_365 (0,23), sugerindo que hosts com mais listagens tendem a ter anúncios com maior disponibilidade.

As demais correlações foram muito fracas.

Conclusões Principais

- O preço dos anúncios não é fortemente influenciado pelas outras variáveis analisadas, outro fator impacta esta variável.
- Anúncios com mais avaliações totais (numero_de_reviews) e mais avaliações por mês (reviews_por_mes) tendem a ter uma pequena maior disponibilidade.
- A disponibilidade tem correlação fraca com o número de avaliações, sugerindo que anúncios mais disponíveis são mais ativos e gerenciados por hosts experientes.
- Anúncios com um número maior de noites mínimas tendem a ter levemente menos avaliações por mês, possivelmente porque são menos atraentes para reservas de curta duração.

Correlação entre as variáveis categóricas



Correlação entre a variáveis room_type e price. Fonte: Autor

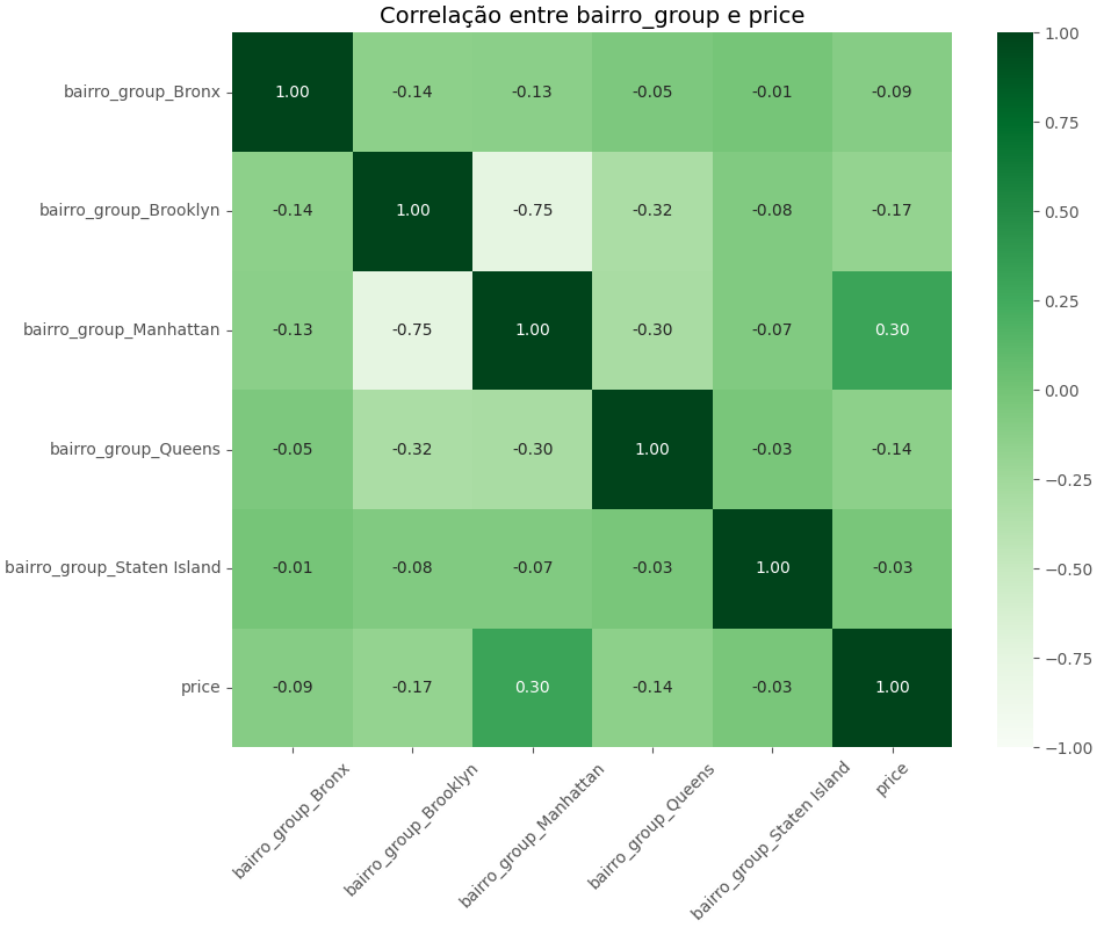
Os resultados da correlação e do heatmap acima apresentam os níveis de relação entre a variável categórica room_type e a númerica price do DataFrame. Se nota:

- Há uma correlação positiva forte (0,61) entre anúncios do tipo "Entire home/apt" e o preço. Isso significa que, em geral, anúncios desse tipo tendem a ter preços mais altos.
- Há uma correlação negativa forte (-0.58) entre anúncios do tipo "Private room" e o preço. Isso indica que anúncios desse tipo tendem a ter preços mais baixos.
- Há uma correlação negativa muito fraca (-0.10) entre anúncios do tipo "Shared room" e o preço. Isso sugere que o tipo "Shared room" praticamente não influencia o preço.
- Há uma correlação negativa muito forte entre room_type_Entire home/apt e room_type_Private room:
    - Isso é esperado, pois um anúncio não pode ser ao mesmo tempo "Entire home/apt" e "Private room". Quando uma categoria é 1, a

outra tende a ser 0.

Conclusões principais

- Anúncios do tipo "Entire home/apt" estão associados a preços mais altos (correlação positiva de 0.61).
- Anúncios do tipo "Private room" estão associados a preços mais baixos (correlação negativa de -0.58).
- Anúncios do tipo "Shared room" praticamente não influenciam o preço (correlação muito próxima de zero).



Correlação entre a variáveis bairro_group e price. Fonte: Autor

Os resultados da correlação e do heatmap acima apresentam os níveis de relação entre a variável categórica bairro_group e a númerica price do DataFrame. Se nota:

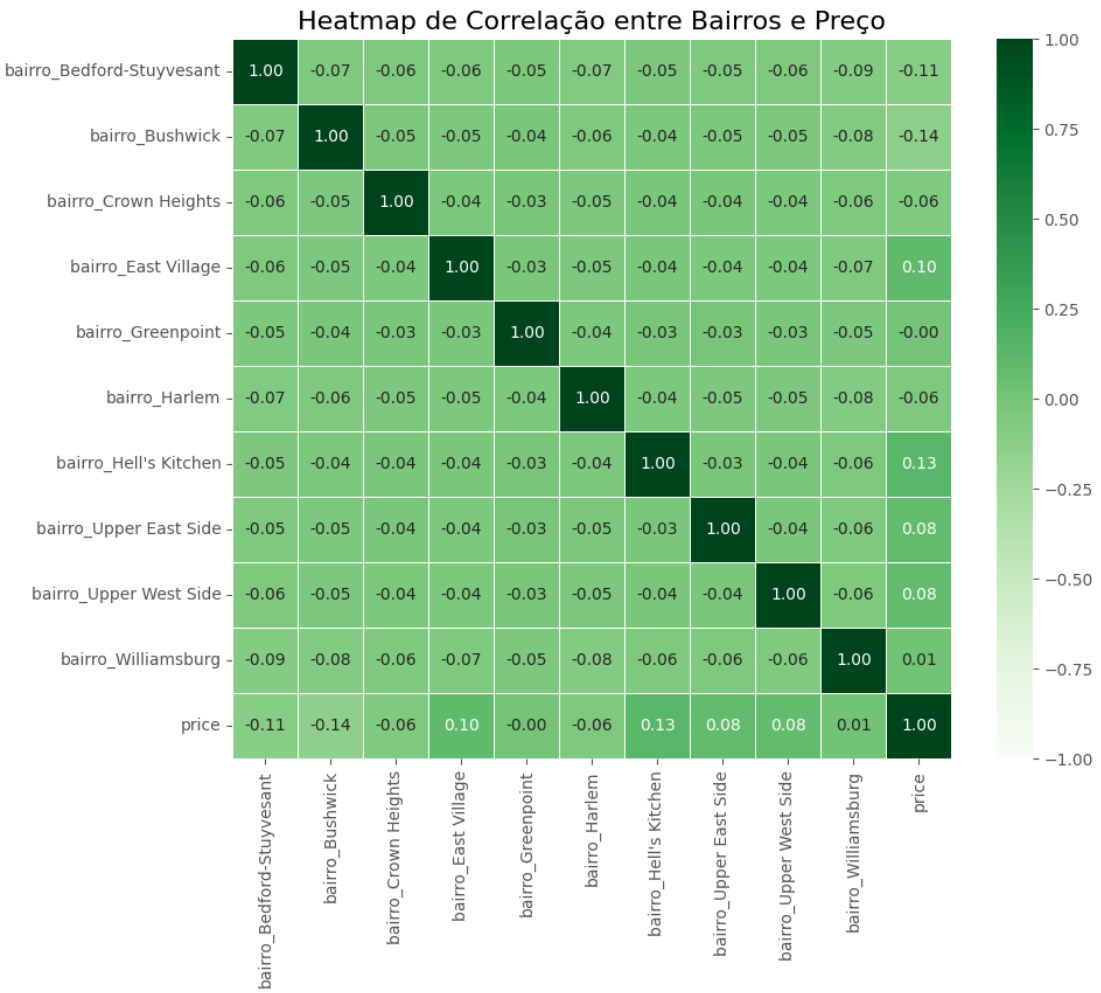
- Há uma correlação positiva moderada (0.30) entre anúncios em Manhattan e o preço. Isso significa que anúncios em Manhattan

tendem a ter preços mais altos.

- As demais correlações com o preço são muito fracas.

Conclusões principais

- Manhattan está associada a preços mais altos (correlação positiva de 0.3).
- Brooklyn e Queens estão associados a preços ligeiramente mais baixos.
- Bronx e Staten Island praticamente não influenciam o preço (correlações próximas de zero).
- Os bairros são mutuamente exclusivos (um anúncio pertence a apenas um bairro). Por isso, as correlações entre eles são negativas.



Correlação entre a variáveis bairro e price. Fonte: Autor

Os resultados da correlação e do heatmap acima apresentam os níveis de relação entre a variável categórica bairro e a númerica price do DataFrame. Se nota:

- Todas as correlações foram muito fracas.

Conclusões principais

- Os bairros com correlação negativa (Bedford-Stuyvesant, Bushwick, Crown Heights, Harlem) tendem a estar associados a preços mais baixos, com Bushwick sendo o mais forte nessa tendência.
- Os bairros com correlação positiva (East Village, Hell's Kitchen, Upper East Side, Upper West Side) tendem a estar associados a preços mais altos, com Hell's Kitchen sendo o mais forte nessa tendência.
- Greenpoint e Williamsburg têm correlações próximas de zero, indicando que a presença desses bairros não influencia significativamente o preço.

Correlação entre a variáveis nome e price. Fonte: Autor

Os resultados da correlação e do heatmap acima apresentam os níveis de relação entre a variável categórica nome e a númerica price do DataFrame. Se nota:

- A maioria dos nomes listados tem uma correlação muito próxima de zero, indicando que a presença desses nomes não influencia significativamente o preço dos imóveis.
- Nomes dos imóveis não parecem ter um impacto significativo no preço, já que as correlações são muito próximas de zero. Isso sugere que outros fatores (como localização, tamanho, comodidades, etc.) podem ser mais importantes na determinação do preço dos imóveis.

Respondendo as perguntas

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

Analisando os resultados na EDA os três melhores bairros para compra visando alugar na plataforma são:

- West Village → Potencial: Alta demanda
  - Tem o preço médio mais alto
  - Tem um número significativo de reviews
  - Tem baixa disponibilidade
- Chelsea → Potencial: Alta popularidade
  - Tem o segundo preço médio mais alto
  - Tem o maior numero de reviews
  - Tem disponibilidade moderada
- Nolita → Potencial: Público que paga mais com menos reviews
  - Tem preço médio alto
  - Tem poucos reviews
  - Tem disponibilidade média

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

- Foram feitos testes de Kruskal Wallis e Games Howell para verificar a influência na mediana dos preços.
- As variáveis foram transformadas em faixas:
  - minimo_noites
    - Estadia curta (0-3)
    - Estadia média (4-7)
    - Estadia longa (8-11)
  - disponibilidade_365

- Muito Baixa: 0 a 30 dias
- Baixa: 31 a 90 dias
- Média Baixa: 91 a 150 dias
- Média Alta: 151 a 210 dias
- Alta: 211 a 270 dias
- Muito Alta: 271 a 365 dias

---

minimo_noites

O preço dos aluguéis varia de forma significativa dependendo do número mínimo de noites exigido para a reserva.

Principais conclusões:

- A faixa "4-7" tem preços significativamente maiores do que "0-3" e "8-11".
- A faixa "0-3" tem preços significativamente mais baixos do que "4-7", mas mais altos do que "8-11".

---

disponibilidade_365

A diferença entre as faixas de disponibilidade_365 e preço é significativa

Principais conclusões:

- Imóveis com menor disponibilidade de dias no ano (como "Muito Baixa") tendem a ser mais baratos.
- Nem todas as comparações de faixas mostram diferenças significativas, o que indica que a disponibilidade tem um efeito mais relevante em alguns casos do que em outros.

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

Sem aprofundar em nada estatístico ou de modelagem, podemos observar que:

- Uso de palavras de alto impacto: Termos como "luxuoso", "exclusivo", "palácio", "mansão", "penthouse", "espetacular", "privilégio" e "exclusivo" são frequentemente utilizados para destacar a qualidade e o status do imóvel.

- Há menção de bairros prestigiados, isso pode atrair compradores interessados em morar em áreas de alto padrão.
- Destaque para características únicas: "vista deslumbrante", "terraço privativo", "jardim", "lareira", "piscina" e "academia", pode aumentar o interesse e justificar um preço mais alto.
- Empregar palavras como "melhor", "mais bonito", "mais luxuoso" e "único" enfatiza a exclusividade e o alto valor do imóvel.
- A descrição dos imóveis é predominantemente textual, o que dificulta a análise quantitativa direta.
- Seria necessário aplicar técnicas de processamento de linguagem natural (NLP) para extrair informações relevantes e quantificáveis.

Imóveis com mais reviews têm preços mais altos?

Sim, como visto na análise univariada, existe forte evidência de que há uma associação significativa entre o número de reviews e o preço.

No entanto, essa relação não é necessariamente linear e pode ser influenciada por outros fatores.

Imóveis com alta disponibilidade têm preços mais baixos?

Sim, a análise univariada mostra que a maioria dos imóveis com preços baixos estão na categoria de alta disponibilidade.

Isso sugere que os proprietários de imóveis com alta disponibilidade podem estar dispostos a cobrar preços mais baixos para atrair inquilinos.

Há evidências suficientes para afirmar que existe uma relação significativa entre a faixa de disponibilidade e a faixa de preço dos imóveis

Quais variáveis e/ou suas transformações você utilizou e por quê?

As variáveis utilizadas foram:

- bairro_group

- bairro
- room_type
- minimo_noites
- numero_de_reviews
- reviews_por_mes
- calculado_host_listings_count
- disponibilidade_365

Elas foram escolhidas com base no resultado do teste estatístico de associação entre variáveis. Por mostrarem associação, podem ser bons preditores para o modelo.

As variáveis que não foram escolhidas, como as id’s e latitude x longitude, são de identificadores únicos, ou tem correlação direta com alguma variável escolhida (como bairro, por exemplo).

Variáveis como bairro_group, bairro, e room_type são categóricas e precisam ser codificadas (por exemplo, usando One-Hot Encoding) para que possam ser utilizadas em modelos de machine learning.

Foi necessário normalizar ou padronizar as variáveis numéricas (como minimo_noites, numero_de_reviews, reviews_por_mes, calculado_host_listings_count, e disponibilidade_365) para garantir que todas tenham a mesma escala.

Qual tipo de problema estamos resolvendo (regressão, classificação)?

Como estamos pensando em previsão de preços, este é um problema de regressão.

Qual modelo melhor se aproxima dos dados e quais seus prós e contras?

O modelo que melhor performou foi o LightGBM, com um RMSE base de **55.49** e otimizado, com **55.17**.

Prós: Velocidade, eficiência, precisão e capacidade de lidar com dados desbalanceados.

Contras: Sofre de overfitting com mais facilidade, é um pouco mais complexo e sensível.

Qual medida de performance do modelo foi escolhida e por quê?

A medida escolhida foi a RMSE (Root Mean Squared Error).

O RMSE é expresso na mesma unidade da variável target, o que facilita a interpretação do erro. Ele indica o erro médio em dólares.

Ele penaliza mais os erros grandes, o que é útil para identificar modelos que cometem previsões muito distantes dos valores reais.

É amplamente utilizado para avaliar modelos de regressão.

Supondo um apartamento com as seguintes características, qual seria a sugestão de 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',
 '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}
```

```
# Faz a previsão com o modelo otimizado
prediction = best_lgb_model.predict(teste)
print(prediction)
```

```
[261.47852646]
```

Print do código de predição e seu retorno. Fonte: Autor

Executar o modelo preditivo com base nas características fornecidas retorna um preço de **$261.47**

Se chegou até aqui, agradeço pela leitura do trabalho.

Fique à vontade para fazer críticas, perguntas e sugestões.

Meus links:

LinkedIn: https://www.linkedin.com/in/ianrstoltz098/

E-mail: mailto:ian.rstoltz@gmail.com

Portfólio: **Clique aqui!**

Ian Stoltz - Analista de Dados
