Análise Exploratória de Dados

Precificação de Aluguéis Temporários - NYC

Amaury Nogueira Neto

Fevereiro de 2025

Sumário

1. Introdução

Este relatório apresenta a análise exploratória de dados (EDA) realizada sobre um conjunto de dados de aluguéis temporários em Nova York. A análise inclui estatísticas descritivas, distribuição dos dados por meio de gráficos, análise espacial e textual, e sugestões para investidores.

Além disso, são discutidas hipóteses de negócio que visam identificar áreas com potencial de investimento, a influência de variáveis operacionais (como mínimo de noites e disponibilidade) na precificação, e a relevância de determinados termos nos títulos dos anúncios para justificar preços diferenciados.

2. Visualização Inicial dos Dados

	id	nome	host_name	bairro_group	price	minimo_noites	numero_de_reviews	disponibilidade_365
0	2595	Skylit Midtown Castle	Jennifer	Manhattan	225	1	45	355
1	3647	THE VILLAGE OF HARLEM....NEW YORK !	Elisabeth	Manhattan	150	3	0	365
2	3831	Cozy Entire Floor of Brownstone	LisaRoxanne	Brooklyn	89	1	270	194
3	5022	Entire Apt: Spacious Studio/Loft by central park	Laura	Manhattan	80	10	9	0
4	5099	Large Cozy 1 BR Apartment In Midtown East	Chris	Manhattan	200	3	74	129

3. Estatísticas Descritivas

	count	mean	std	min	25%	50%	75%	max
id	48894	1.90175e+07	1.09829e+07	2595	9.47237e+06	1.96774e+07	2.91522e+07	3.64872e+07
host_id	48894	6.76214e+07	7.86112e+07	2438	7.82274e+06	3.07955e+07	1.07434e+08	2.74321e+08
latitude	48894	40.729	0.0545294	40.4998	40.6901	40.7231	40.7631	40.9131
longitude	48894	-73.9522	0.0461571	-74.2444	-73.9831	-73.9557	-73.9363	-73.713
price	48894	152.721	240.157	0	69	106	175	10000
minimo_noites	48894	7.03009	20.5107	1	1	3	5	1250
numero_de_reviews	48894	23.2748	44.551	0	1	5	24	629
reviews_por_mes	38842	1.37325	1.68045	0.01	0.19	0.72	2.02	58.5
calculado_host_listings_count	48894	7.14401	32.9529	1	1	1	2	327
disponibilidade_365	48894	112.776	131.619	0	0	45	227	365

4. Valores Ausentes

	Valores Ausentes
id	0
nome	16
host_id	0
host_name	21
bairro_group	0
bairro	0
latitude	0
longitude	0
room_type	0
price	0
minimo_noites	0
numero_de_reviews	0
ultima_review	10052
reviews_por_mes	10052
calculado_host_listings_count	0
disponibilidade_365	0

5.1 Distribuição de Preços

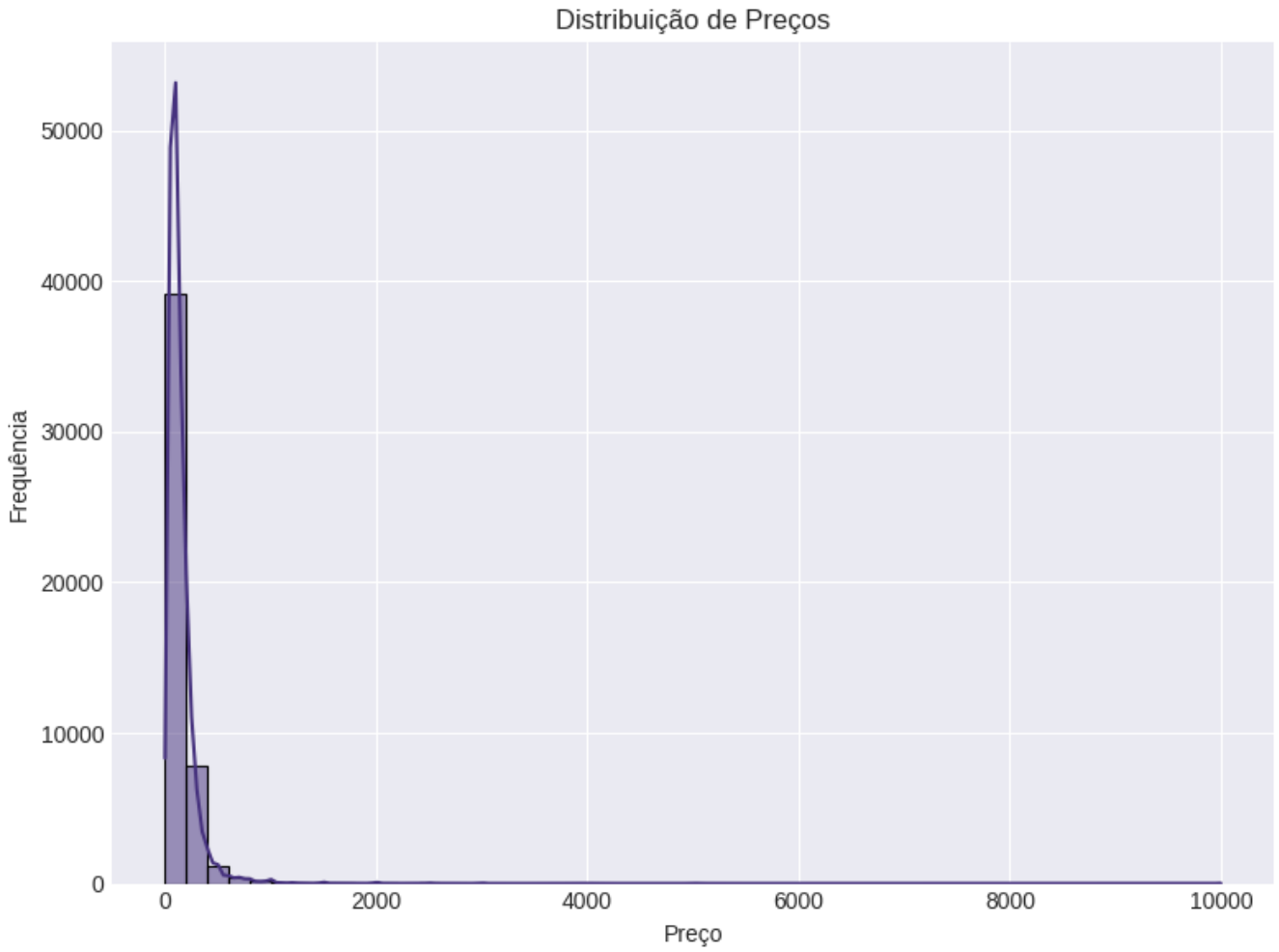
Gráfico: Distribuição de Preços

Descrição: Este gráfico apresenta a distribuição dos preços dos aluguéis.

Eixo X: Preço dos aluguéis.

Eixo Y: Frequência (contagem) dos preços.

Interpretação: A linha de densidade (KDE) indica a tendência geral da distribuição. Observa-se que a maioria dos preços se concentra em uma faixa específica, com possíveis outliers representando imóveis de luxo ou localizados em áreas de alto custo.



5.2 Distribuição de Mínimo de Noites

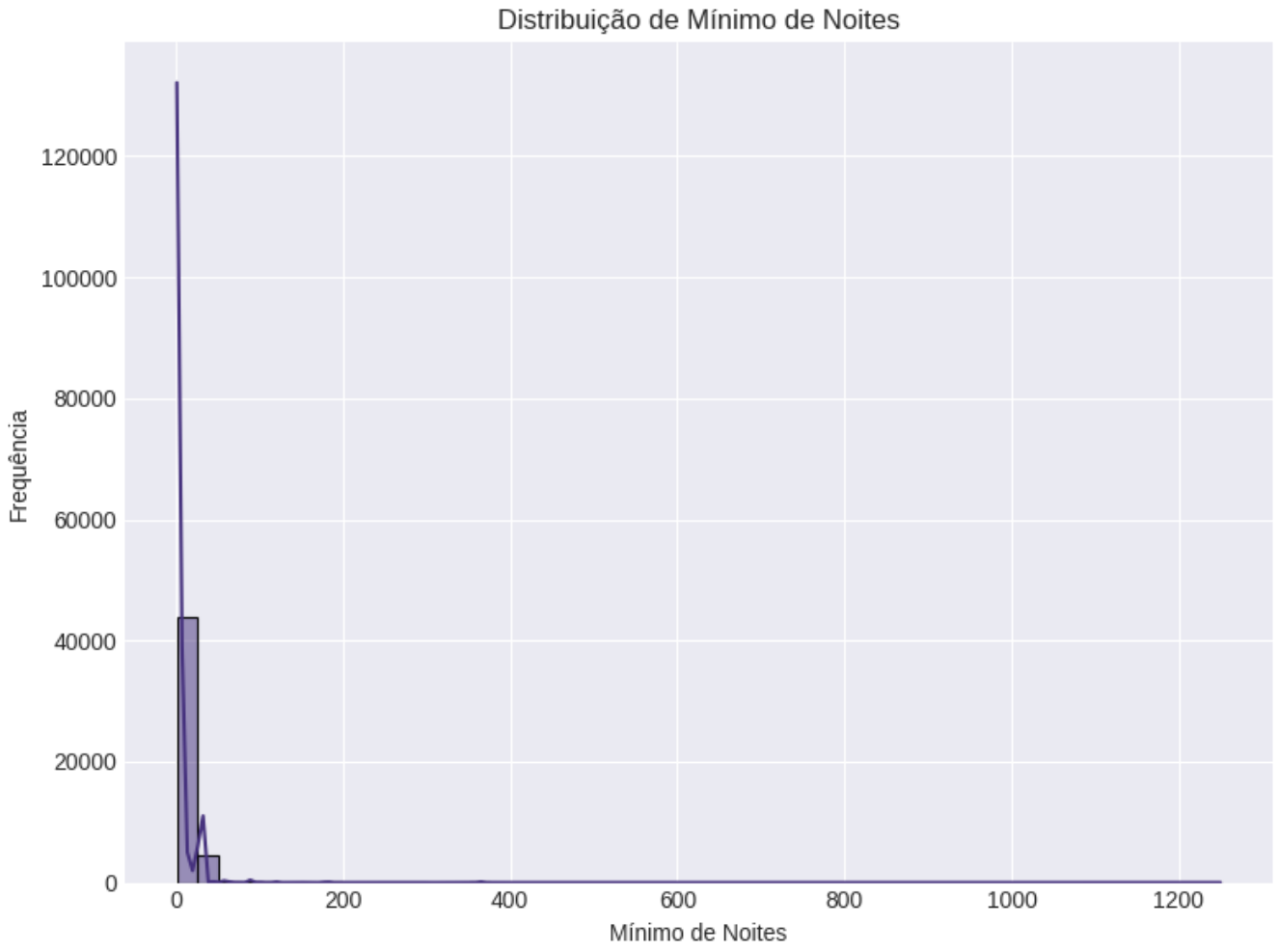
Gráfico: Distribuição de Mínimo de Noites

Descrição: Este gráfico mostra a distribuição do número mínimo de noites exigidas para reserva.

Eixo X: Número mínimo de noites.

Eixo Y: Frequência.

Interpretação: É possível identificar os padrões de exigência dos anfitriões. Picos em valores como 1 ou 3 noites podem indicar políticas comuns de reserva.



5.3 Distribuição do Número de Reviews

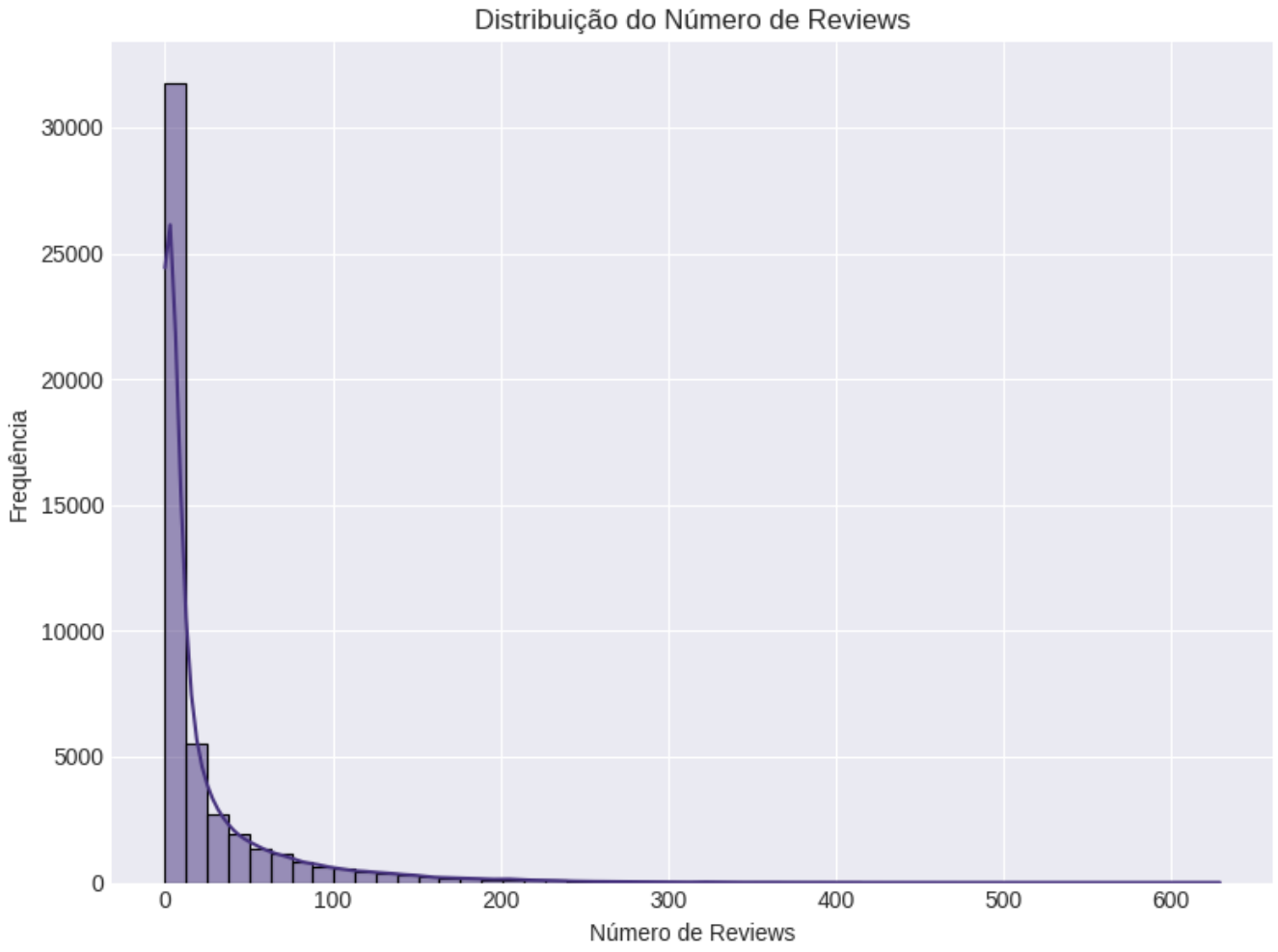
Gráfico: Distribuição do Número de Reviews

Descrição: Este gráfico ilustra a distribuição do número de avaliações (reviews) recebidas pelos imóveis.

Eixo X: Número de Reviews.

Eixo Y: Frequência.

Interpretação: A maioria dos imóveis apresenta poucos reviews, o que pode indicar que muitos são novos ou têm baixa taxa de ocupação. Outros imóveis com muitos reviews podem representar propriedades bem estabelecidas.



5.4 Distribuição da Disponibilidade Anual

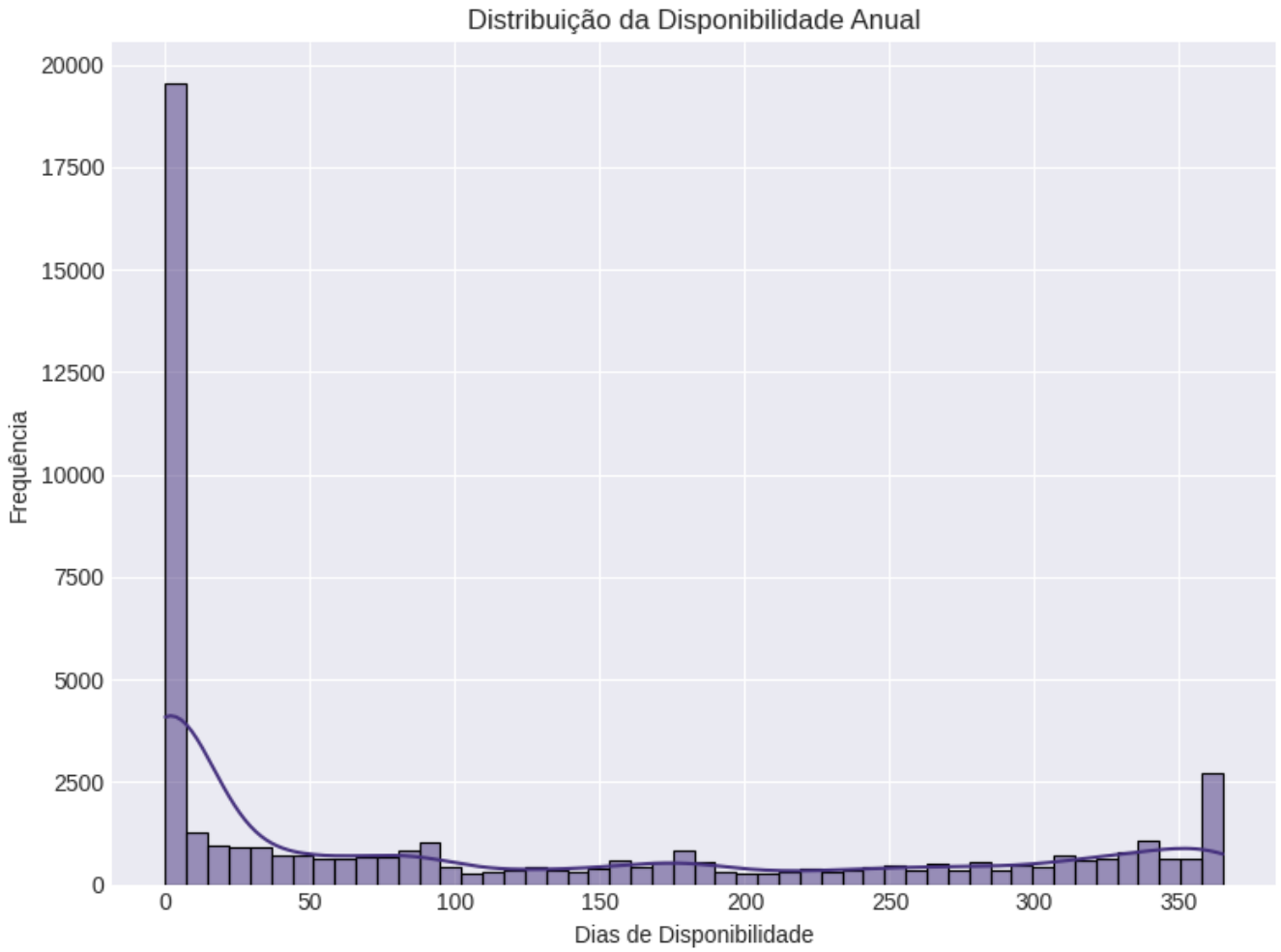
Gráfico: Distribuição da Disponibilidade Anual

Descrição: Este gráfico exibe a distribuição do número de dias que os imóveis estão disponíveis ao longo do ano.

Eixo X: Número de dias disponíveis (disponibilidade anual).

Eixo Y: Frequência.

Interpretação: Permite identificar a sazonalidade e as políticas de disponibilidade. Valores mais altos podem indicar imóveis com alta disponibilidade, enquanto valores baixos podem sugerir reservas frequentes.



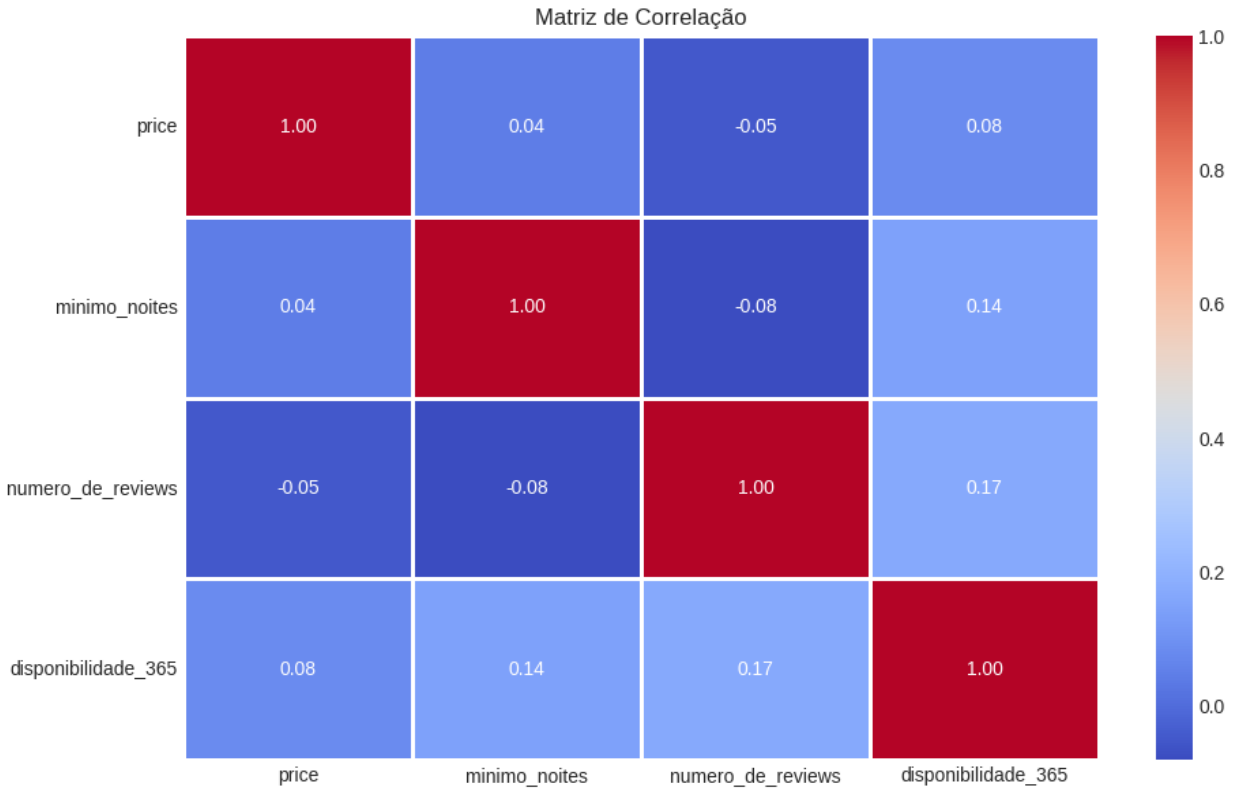
6. Matriz de Correlação

Gráfico: Matriz de Correlação

Descrição: Esta matriz apresenta a correlação entre as variáveis numéricas: Preço, Mínimo de Noites, Número de Reviews e Disponibilidade Anual.

A matriz utiliza cores para indicar a intensidade da correlação, com anotações numéricas.

Interpretação: Valores próximos de 1 ou -1 indicam uma forte correlação positiva ou negativa, respectivamente. Valores próximos de 0 sugerem pouca ou nenhuma correlação.

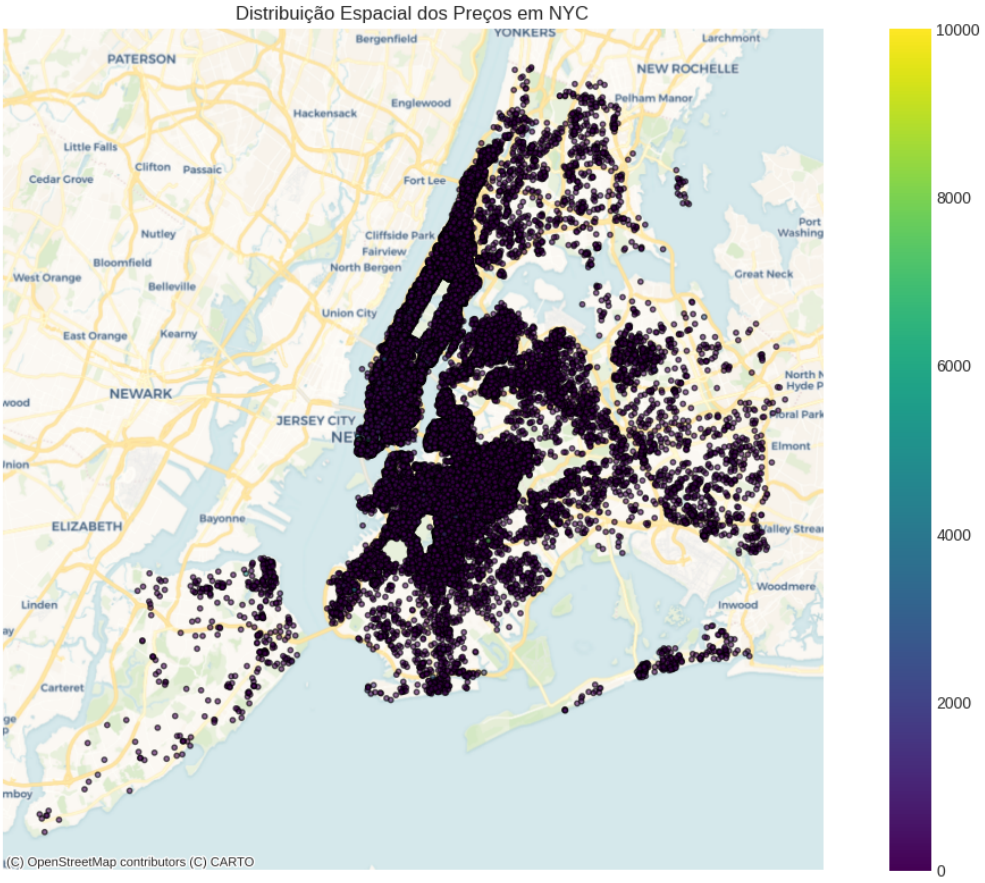


7. Distribuição Espacial dos Preços

Gráfico: Distribuição Espacial dos Preços

Descrição: Este mapa geoespacial apresenta a localização dos imóveis em NYC, coloridos de acordo com o preço. A legenda indica os intervalos de preços.

Interpretação: Áreas com cores mais intensas podem representar bairros com preços mais elevados. Os pontos individuais mostram a distribuição dos imóveis na cidade.



8. Palavras mais Comuns em Anúncios

Análise de Texto: Palavras mais Comuns em Anúncios

Descrição: Nesta seção, utilizamos o TF-IDF para identificar os termos mais comuns nos títulos dos anúncios, separados por faixas de preço (alto e baixo).

Interpretação: Os termos mais frequentes podem indicar características valorizadas em anúncios de alto preço, enquanto os termos em anúncios de baixo preço podem destacar atributos diferentes.

Alto preço: Termos mais comuns em anúncios de alto preço: 1BR, apartment, bedroom.

Baixo preço: Termos mais comuns em anúncios de baixo preço: brooklyn, beautiful, apt.

9. Insights e Hipóteses de Negócio

A partir da análise geoespacial, observa-se que bairros centrais, como Manhattan, apresentam preços elevados. Entretanto, áreas emergentes ou bairros próximos, como certas regiões de Brooklyn, podem oferecer um melhor equilíbrio entre custo e potencial de valorização.

A análise dos gráficos de mínimo de noites e disponibilidade sugere que:

- Imóveis com mínimo de noites baixo tendem a ter preços competitivos para atrair reservas frequentes;
- Imóveis com alta disponibilidade podem adotar estratégias diferenciadas de precificação, visando maximizar a ocupação em períodos específicos.

Adicionalmente, a análise de texto dos nomes dos anúncios indica que termos que enfatizam atributos de luxo ou localização privilegiada estão associados a preços mais altos, enquanto termos mais genéricos podem ser encontrados em anúncios de menor valor.

10. Justificativa das Escolhas no Pipeline e Validação do Modelo

No pipeline de modelagem, foram selecionadas variáveis numéricas (latitude, longitude, minimo_noites) que capturam informações sobre a localização e condições de reserva, sendo padronizadas com StandardScaler para evitar que diferenças de escala prejudiquem o desempenho.

Variáveis categóricas, como bairro_group e room_type, foram transformadas com TargetEncoder, pois essa técnica codifica as categorias com base na média do alvo, facilitando a captura de relações não lineares entre cada categoria e o preço.

A variável textual (nome) foi processada com TfidfVectorizer para extrair padrões relevantes dos anúncios, possibilitando identificar termos associados a atributos valorizados, como 'luxury' ou 'designer'.

Para a validação do modelo, utilizou-se StratifiedKFold com uma variável de estratificação (price_strata) para manter a distribuição dos preços nos conjuntos de treino e teste. Essa abordagem robusta, aliada ao uso de métricas como MAPE, R^2 e RMSE, assegura uma avaliação realista da performance e a generalização do modelo para toda a faixa de preços.

11. Respostas às Perguntas Específicas do Desafio

a. Investimento em Apartamento para Aluguel:

- Análise: A análise geoespacial e a matriz de correlação indicam que, embora bairros centrais como Manhattan apresentem preços elevados, áreas emergentes ou regiões adjacentes (por exemplo, certas áreas de Brooklyn) podem oferecer um melhor custo-benefício, combinando preços de aquisição mais moderados com alta demanda e potencial de valorização.

b. Influência do Número Mínimo de Noites e da Disponibilidade:

- Análise: Os gráficos demonstram que um número mínimo de noites baixo tende a favorecer preços competitivos, aumentando a rotatividade de reservas, enquanto a alta disponibilidade pode indicar estratégias de precificação diferenciadas para captar maior volume de reservas ou para aproveitar sazonalidades específicas.

c. Padrão no Texto do Nome dos Anúncios:

- Análise: A análise textual revela que anúncios contendo termos como 'luxury', 'designer' e 'central' estão frequentemente associados a imóveis de alto valor, enquanto termos mais genéricos são predominantes em anúncios com preços inferiores.

12. Explicação da Previsão do Preço

A previsão do preço foi realizada através de um pipeline de modelagem que integra diversas etapas de pré-processamento e a aplicação de um modelo preditivo.

Variáveis utilizadas e suas transformações:

- **Variáveis Numéricas:** Foram incluídas latitude, longitude e minimo_noites. Essas variáveis foram padronizadas utilizando StandardScaler para que diferenças de escala não afetem a performance do modelo.
- **Variáveis Categóricas:** As variáveis bairro_group e room_type foram transformadas com TargetEncoder, pois essa técnica codifica cada categoria com base na média do preço, permitindo capturar relações não lineares entre a categoria e o valor do imóvel.
- **Variável Textual:** O nome do anúncio foi processado com TfidfVectorizer para extrair padrões e identificar termos que podem estar associados a preços diferenciados, como 'luxury' ou 'designer'.

Tipo de problema: Estamos resolvendo um problema de **regressão**, uma vez que o objetivo é prever um valor contínuo (o preço do imóvel).

Modelo escolhido: O modelo que melhor se aproxima dos dados é o **LGBMRegressor**.

- **Prós:** É rápido, eficiente para grandes volumes de dados, e lida bem com variáveis de diferentes tipos (numéricas, categóricas e textuais).
- **Contras:** Pode ser sensível a hiperparâmetros e, sem uma adequada validação, pode apresentar overfitting.

Métrica de performance: Foi escolhida a **MAPE (Mean Absolute Percentage Error)**, pois permite interpretar o erro em termos percentuais, o que facilita a compreensão do desempenho do modelo em relação à variação dos preços.

13. Conclusão

A análise exploratória revelou padrões importantes nos preços dos aluguéis em NYC, desde a distribuição dos valores, as políticas de reserva e as correlações com a localização, até os atributos dos anúncios. As hipóteses de negócio, fundamentadas na análise geoespacial, nas variáveis operacionais e na extração de termos relevantes dos anúncios, sugerem oportunidades de investimento em áreas emergentes e reforçam a importância de estratégias de precificação ajustadas ao mercado.

Adicionalmente, o pipeline de modelagem foi desenvolvido para integrar variáveis numéricas, categóricas e textuais, resolvendo um problema de regressão por meio do LGBMRegressor. A escolha deste modelo, juntamente com a utilização de métricas como MAPE, R^2 e RMSE, garante uma avaliação realista da performance e a robustez do sistema preditivo.

Em suma, os insights obtidos e as justificativas das escolhas metodológicas fornecem uma base sólida para futuras análises e decisões estratégicas, tanto para a otimização da plataforma quanto para investimentos direcionados no mercado de aluguéis temporários.