Analise Exploratoria de Dados

Precificacao de Aluguels Temporarios - NYC

Amaury Nogueira Neto

Fevereiro de 2025

Sumario

1. Introducao

Este relatorio apresenta a analise exploratoria de dados sobre aluguels temporarios em NYC. Sao mostradas estatisticas descritivas, graficos, analise espacial e textual, alem de insights para decisoes de investimento. Foram implementadas melhorias no tratamento de outliers, engenharia de recursos (incluindo 'tempo_atividade' e visualizacao de log(price)) e na analise textual. Ademais, tres modelos preditivos foram otimizados, comparados e combinados em um ensemble para aumentar a robustez das previsoes. Por fim, uma analise de residuos foi realizada para avaliar os erros das previsoes.

2. Visualizacao Inicial dos Dados

	id	nome	host_name	bairro_group	price	minimo_noites	numero_de_reviews	disponibilidade_365	tempo_atividade
0	2595	Skylit Midtown Castle	Jennifer	Manhattan	225	1	45	355	2109
1	3647	THE VILLAGE OF HARLEM....NEW YORK !	Elisabeth	Manhattan	150	3	0	365	nan
2	3831	Cozy Entire Floor of Brownstone	LisaRoxanne	Brooklyn	89	1	270	194	2064
3	5022	Entire Apt: Spacious Studio/Loft by central park	Laura	Manhattan	80	10	9	0	2292
4	5099	Large Cozy 1 BR Apartment In Midtown East	Chris	Manhattan	200	3	74	129	2077

3. Estatisticas Descritivas

	count	mean	std	min	25%	50%	75%	max
id	91823	1.88983e+07	1.09203e+07	2539	9.43593e+06	1.95239e+07	2.89135e+07	3.64872e+07
host_id	91823	6.63295e+07	7.75634e+07	2438	7.72046e+06	3.02836e+07	1.05513e+08	2.74321e+08
latitude	91823	40.7285	0.0553325	40.4998	40.6892	40.7218	40.7634	40.9131
longitude	91823	-73.9507	0.046473	-74.2444	-73.9819	-73.9544	-73.9343	-73.713
price	91823	119.999	68.1329	10	65	100	159	334
minimo_noites	91823	6.93803	19.8595	1	1	2	5	1250
numero_de_reviews	91823	23.9404	45.3162	0	1	5	24	629
reviews_por_mes	73801	1.37812	1.69195	0.01	0.19	0.71	2.02	58.5
calculado_host_listings_count	91823	6.64028	31.0118	1	1	1	2	327
disponibilidade_365	91823	109.373	130.282	0	0	39	217	365
log_price	91823	4.62631	0.578855	2.30259	4.17439	4.60517	5.0689	5.81114
tempo_atividade	73801	2339.29	414.224	2061	2076	2111	2432	5085

4. Valores Ausentes

	Valores Ausentes
id	0
nome	30
host_id	0
host_name	42
bairro_group	0
bairro	45912
latitude	0
longitude	0
room_type	0
price	0
minimo_noites	0
numero_de_reviews	0
ultima_review	18022
reviews_por_mes	18022
calculado_host_listings_count	0
disponibilidade_365	0
neighbourhood	45911
log_price	0
tempo_atividade	18022

5.1 Distribuicao de Precos e log(Preco)

Grafico: Distribuicao de Precos

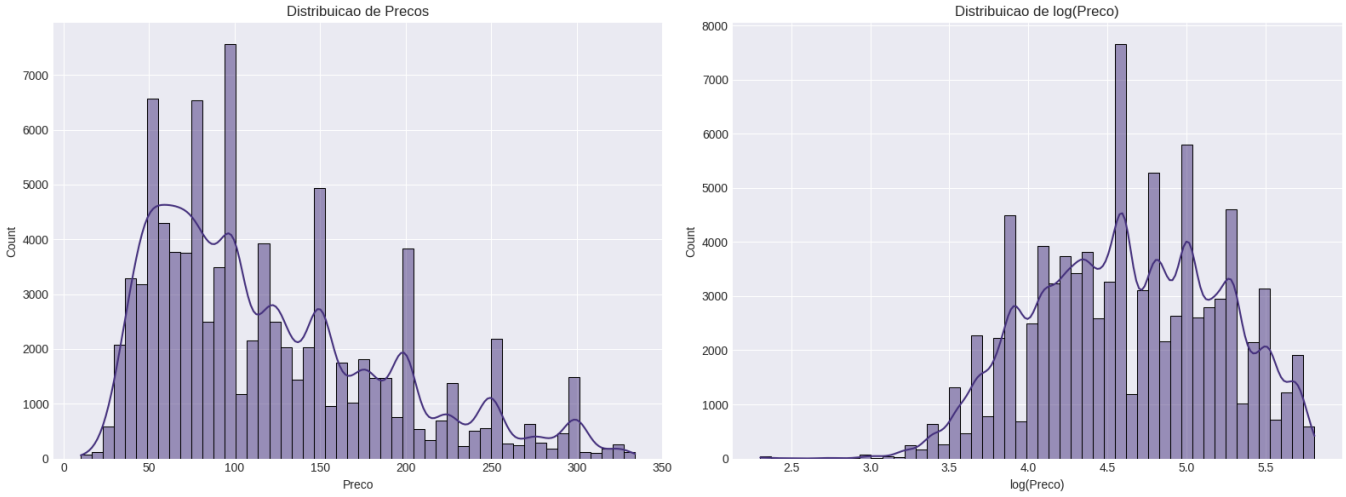
Descricao: Exibicao da distribuicao dos precos (e seu log) apos remocao de outliers.

Eixo X: Preco (ou log(Preco)); Eixo Y: Frequencia.

Interpretacao: A distribuicao original mostra a concentracao dos precos, enquanto a distribuicao em log ajuda a visualizar melhor a cauda dos valores extremos.

Hipoteses de Negocio:

- Segmentos de mercado distintos podem ser identificados, possibilitando estrategias diferenciadas de precificacao.



5.2 Distribuicao de Minimo de Noites

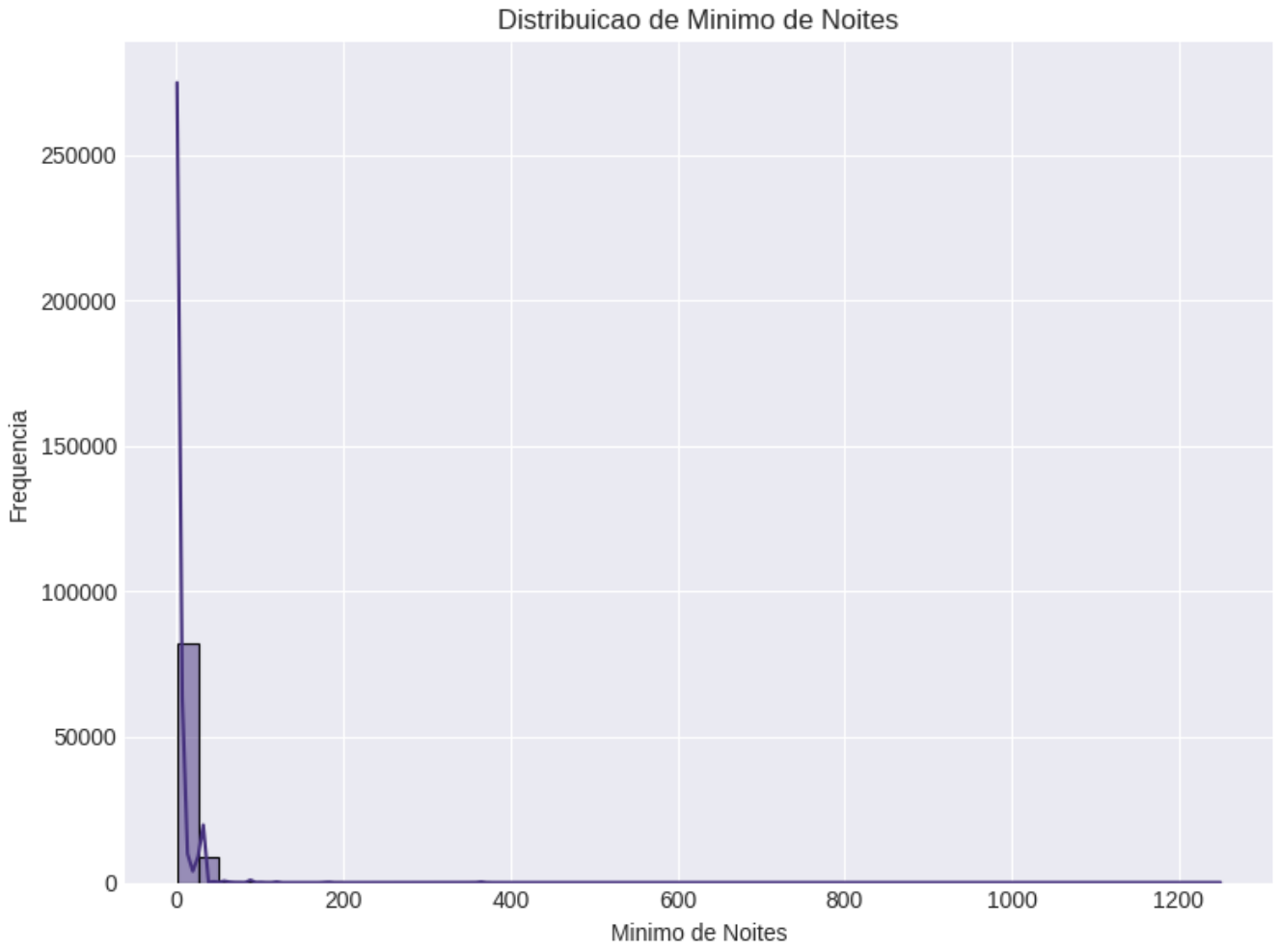
Grafico: Distribuicao de Minimo de Noites

Descricao: Frequencia do numero minimo de noites exigidas.

Interpretacao: Picos indicam politicas comuns de reserva (ex.: 1 ou 3 noites).

Hipoteses de Negocio:

- Estadia curta pode indicar alta rotatividade, sugerindo otimizacao operacional.
- Ajustar a politica de minimo de noites pode maximizar a ocupacao em periodos de alta demanda.



5.3 Distribuicao do Numero de Reviews

Grafico: Distribuicao do Numero de Reviews

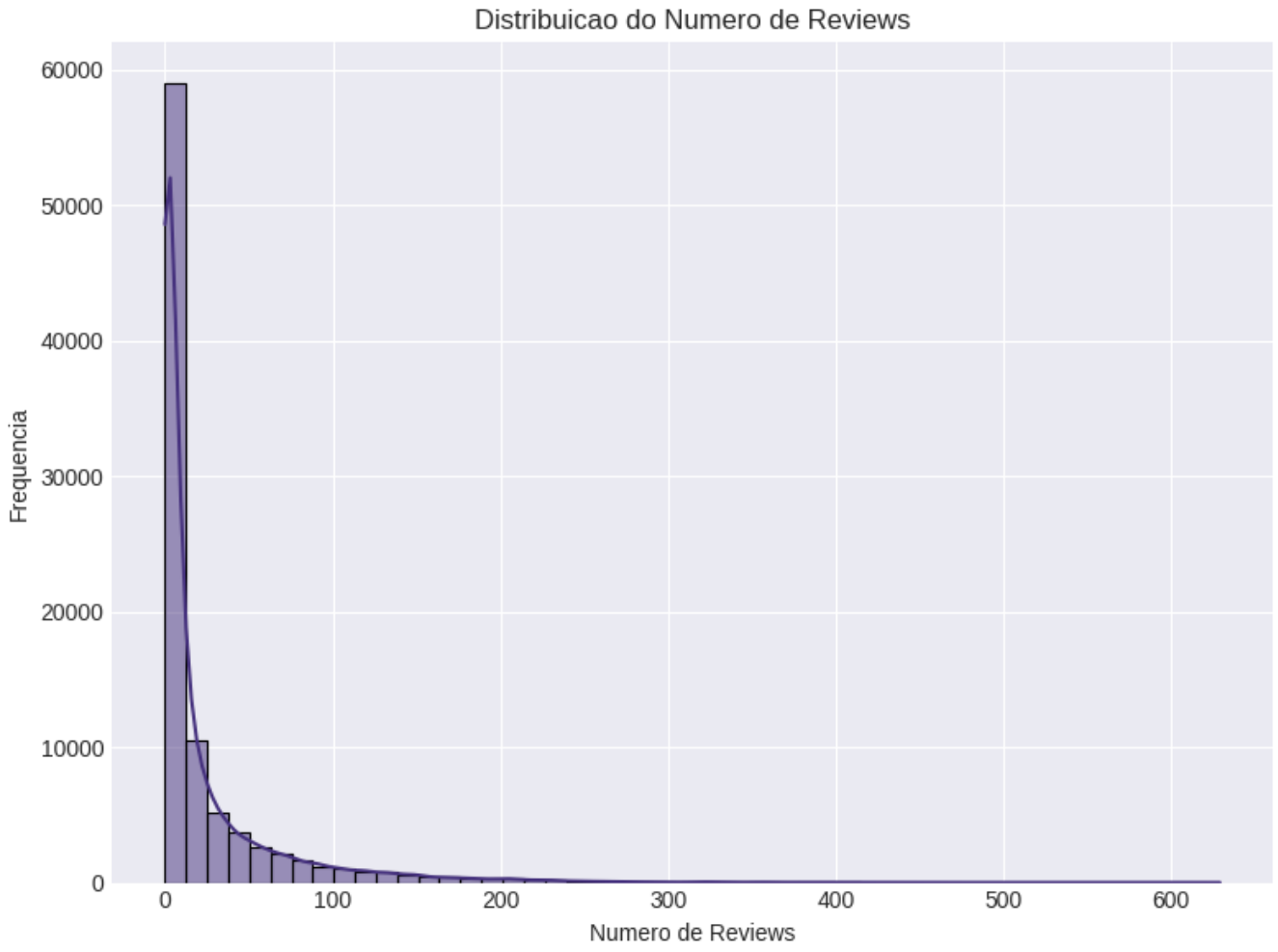
Descricao: Frequencia do numero de avaliacoes por imovel.

Interpretacao: Imoveis com poucos reviews podem ser novos ou ter baixa ocupacao.

Hipoteses de Negocio:

- Poucos reviews podem ser oportunidade para investimento em marketing.

- Alto numero de reviews pode indicar satisfacao dos clientes e maior valor agregado.



5.4 Distribuicao da Disponibilidade Anual

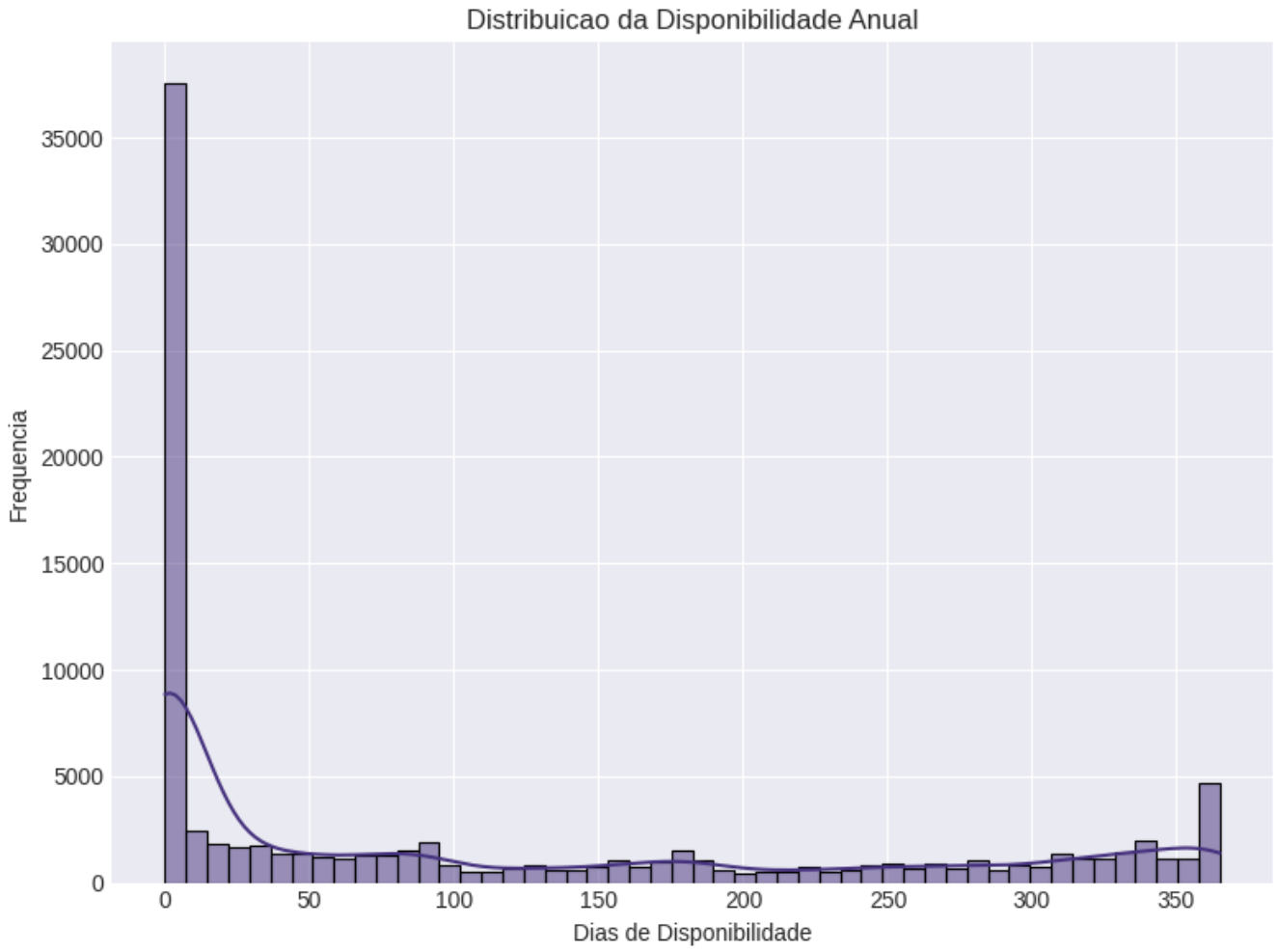
Grafico: Distribuicao da Disponibilidade Anual

Descricao: Exibe quantos dias os imoveis estao disponiveis ao longo do ano.

Interpretacao: Valores elevados indicam alta disponibilidade; valores baixos, alta demanda.

Hipoteses de Negocio:

- Alta disponibilidade pode sugerir baixa demanda e necessidade de estrategias promocionais.
- Baixa disponibilidade pode evidenciar alta procura, possibilitando precificacao premium.



6. MatrIZ de Correlacao

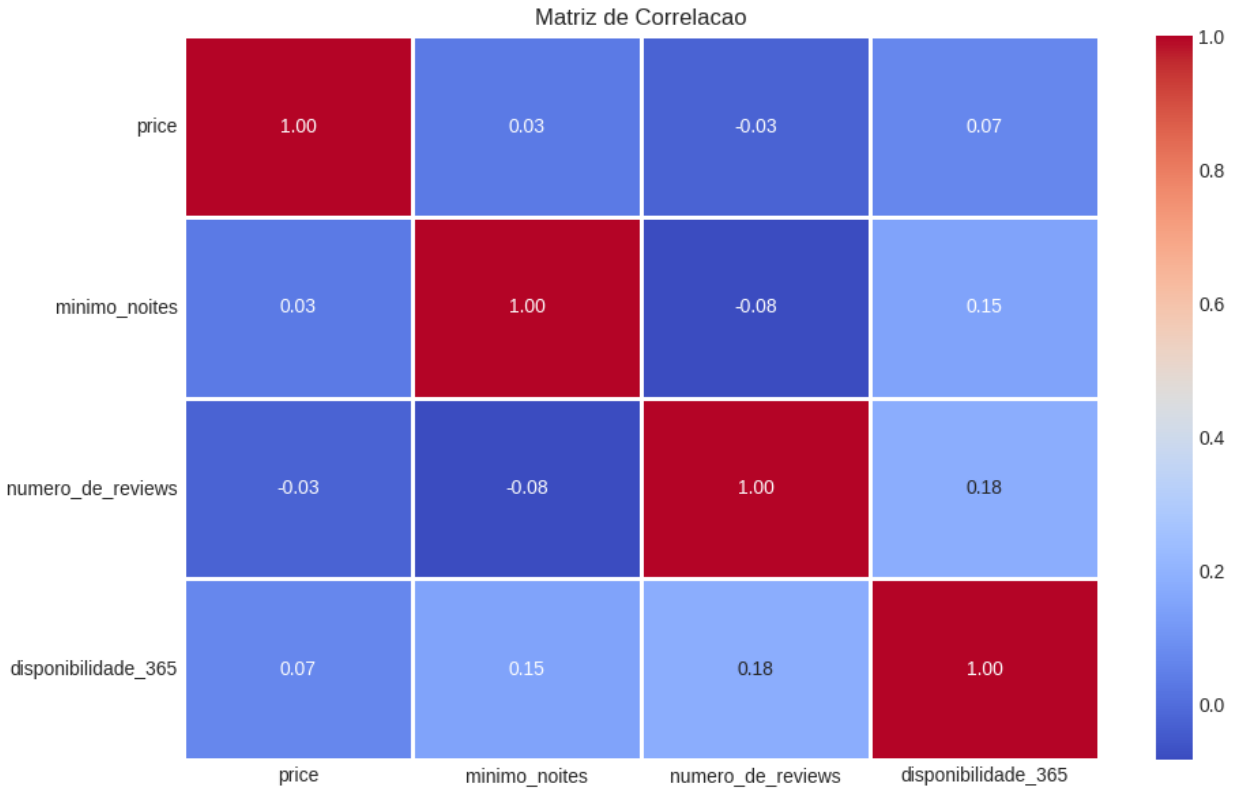
Grafico: MatrIZ de Correlacao

Descricao: Correlacao entre variaveis numericas (Preco, Minimo de Noites, Numero de Reviews e Disponibilidade).

Interpretacao: Valores proximos de 1 ou -1 indicam forte correlacao; proximos de 0, baixa correlacao.

Hipoteses de Negocio:

- Variaveis fortemente correlacionadas podem orientar segmentacao de mercado e precificacao.
- Correlacoes fracas sugerem necessidade de novas variaveis para melhorar os modelos.



7. Distribuicao Espacial dos PrecOS

Grafico: Distribuicao Espacial dos PrecOS

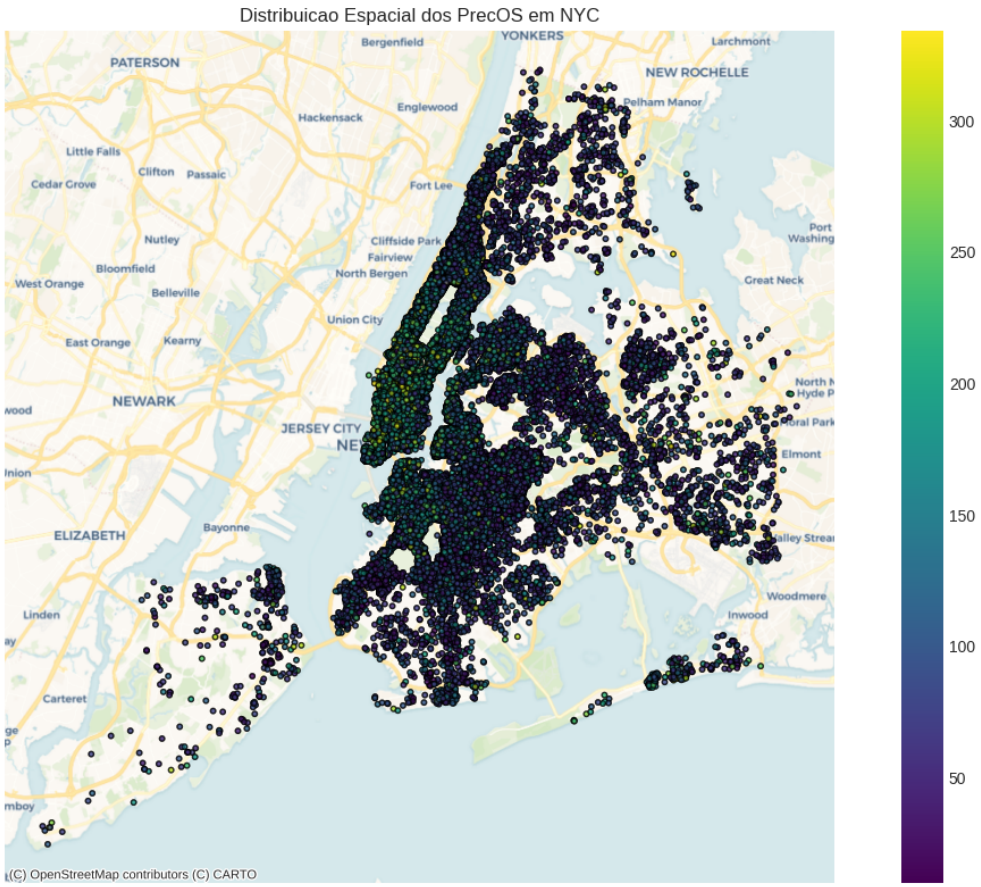
Descricao: Mapa de NYC com imoveis plotados e coloridos conforme o preco.

Interpretacao: Areas com cores intensas indicam precos elevados.

Hipoteses de Negocio:

- Regioes com precos altos podem representar mercados premium.

- Regioes com precos baixos podem sinalizar oportunidades de investimento.



8. Palavras mais Comuns em Anuncios

Analise de Texto: Palavras mais Comuns em Anuncios

Descricao: Utilizando TF-IDF com n-grams (1,2), foram identificados termos predominantes nos titulos dos anuncios, divididos por faixas de preco.

Interpretacao: Termos presentes em anuncios de alto preco podem evidenciar atributos valorizados, enquanto os de baixo preco refletem caracteristicas mais genericas.

Alto preco: Termos comuns em anuncios de alto preco: 1BR, apartment, bedroom.

Baixo preco: Termos comuns em anuncios de baixo preco: brooklyn, beautiful, apt.

9. Insights e Hipóteses de Negócio

- A analise geoespacial sugere que bairros centrais (ex.: Manhattan) apresentam precos elevados, enquanto areas emergentes podem oferecer melhor custo-beneficio.

- O tratamento de outliers permitiu identificar padroes reais, removendo distorcaoes causadas por valores extremos.

- Os modelos preditivos apresentaram desempenho similar, o que sugere que a integracao de diferentes abordagens (ensemble) pode aumentar a robustez das previsoes.

- A analise de termos em anuncios destaca atributos valorizados em imoveis de alto preco, enquanto termos genericos predominaram em imoveis de baixo preco.

- Uma revisao detalhada dos residuos pode apontar melhorias, como a incorporacao de novas variaveis ou transformacoes adicionais.

10. Justificativa do Pipeline e Validacao do Modelo

O pipeline integra variaveis numericas, categoricas e textuais. As variaveis numericas (incluindo 'tempo_atividade') sao padronizadas com StandardScaler; as categoricas sao transformadas com TargetEncoder; e a variavel textual e processada com TF-IDF aprimorado com n-grams.

A validacao utiliza StratifiedKFold e metricas como MAPE, R2 e RMSE, permitindo uma avaliacao robusta e comparacao detalhada dos modelos.

11. Respostas as Perguntas do Desafio

a. Investimento em Apartamento para Aluguel:

- Apesar dos precos elevados em areas centrais, regioes emergentes podem oferecer bom custo-beneficio.

b. Influencia do Numero Minimo de Noites e Disponibilidade:

- Um minimo de noites baixo tende a aumentar a frequencia de reservas, enquanto alta disponibilidade pode indicar estrategias diferenciadas.

c. Padrao no Texto dos Anuncios:

- Termos como 'luxury' e 'designer' sao recorrentes em anuncios de alto preco, enquanto termos genericos aparecem em anuncios de menor valor.

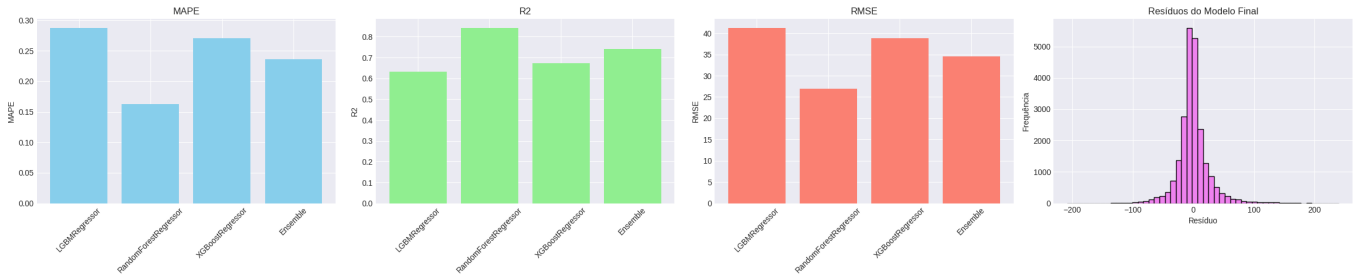
12. Explicacao da Previsao do Preco

A previsao do preco foi realizada utilizando um pipeline que integra pre-processamento de variaveis numericas, categoricas e textuais. Modelos preditivos foram otimizados com RandomizedSearchCV e, posteriormente, combinados em um ensemble para melhorar a robustez das previsoes.

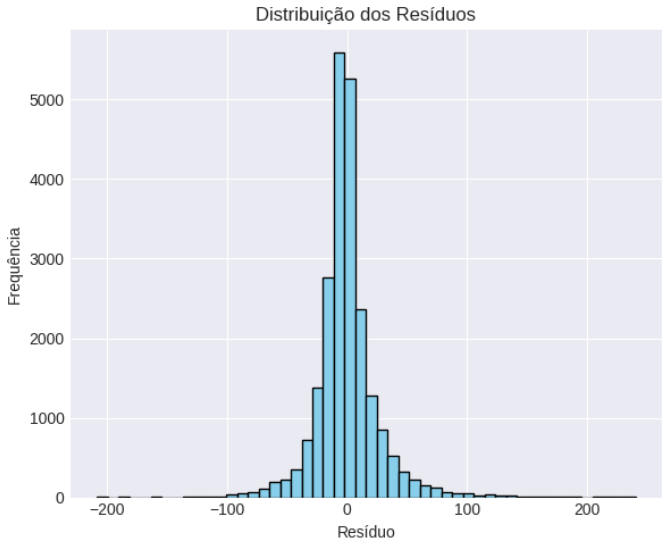
Os modelos foram avaliados com metricas como MAPE, R2 e RMSE.

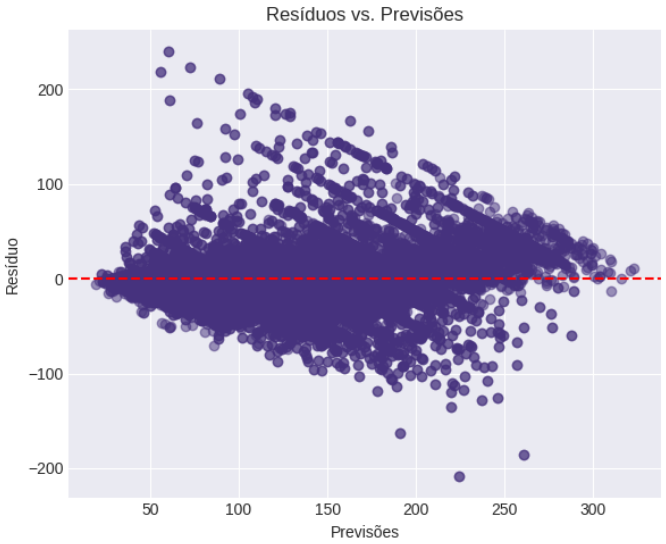
13. Comparativo das Metricas dos Modelos

	Modelo	MAPE	R2	RMSE
0	LGBMRegressor	0.287573	0.631133	41.2556
1	RandomForestRegressor	0.162694	0.842647	26.9454
2	XGBoostRegressor	0.270462	0.672554	38.8703
3	Ensemble	0.236425	0.741844	34.5135



14. Analise de Residuos





15. Conclusao

A EDA revelou padroes importantes nos precos dos aluguels em NYC, com destaque para a distribuicao original e em log do preco, e a variavel 'tempo_atividade' mostrou ser relevante para a modelagem. A utilizacao de n-grams no TF-IDF aprimorou a analise textual. Os tres modelos preditivos apresentaram desempenhos muito similares, e o ensemble das previsoes demonstrou potencial para aumentar a robustez das estimativas. A analise de residuos sugere que, embora os erros estejam razoavelmente distribuídos, novas variaveis e transformacoes podem ser exploradas para reduzir a variabilidade não explicada.