Desafio Cientista de Dados

LH_CD_LUCELIA

Repositório: https://github.com/LuceliaLima/LH_CD_LUCELIA

Relatório das Análises Estatísticas e EDA

Declaração do problema

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.

Objetivos

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.

Informações sobre a cidade de Nova York

A cidade de Nova York é mais populosa do que qualquer cidade brasileira. Com uma população de mais de 8 milhões de habitantes, Nova York é uma das cidades mais densamente povoadas dos Estados Unidos e é um centro cultural, econômico e político importante não só para o país, mas também para o mundo.

A cidade abrange cinco regiões chamadas **boroughs** são estes: Bronx, Brooklyn, Manhattan, Queens e Staten Island.

- **1 - Manhattan é o grande centro**, é o coração de Nova York, é conhecida por seus arranha-céus icônicos, como o Empire State Building e o One

World Trade Center. É o centro financeiro, comercial e cultural da cidade, abrigando instituições como Wall Street, a Broadway e a Times Square.

- **2 - Brookly** localizado ao sudoeste de Manhattan, é o **borough mais populoso de Nova York** e é conhecido por sua diversidade étnica e cultural e maiores instalações portuárias da cidade.
- **3 - Queens** é uma mistura de **áreas residenciais, comerciais e industriais bem diversificada**, com grandes aeroportos e estádios de tênis e beisebol.
- **4 - Bronx** localizado ao norte de Manhattan, é o **local mais pobre e violento da cidade e do país.**
- **5 - Staten Island** situado ao sul de Manhattan, é conhecido por seu **ambiente mais suburbano** em comparação com os outros boroughs. Possui grandes áreas verdes e praias.

Figura 1 - Boroughs de Nova York
Fonte: https://pt.wikipedia.org/wiki/Boroughs_de_Nova_Iorque

Dicionário de variáveis:

- **id** - Atua como uma chave exclusiva para cada anúncio nos dados do aplicativo
- **name** - O nome do anúncio (propriedade)
- **host_id** - O id do usuário que hospedou o anúncio
- **host_name** - Nome do usuário que hospedou o anúncio
- **bairro_group** - Nome do bairro onde o anúncio está localizado
- **bairro** - Nome da área onde o anúncio está localizado.
- **latitude** - Latitude do local
- **longitude** - Longitude do local
- **room_type** - Tipo de espaço de cada anúncio (tipo de quarto)
- **price** - Preço por noite em dólares listado pelo anfitrião
- **minimo_noites** - Número mínimo de noites que o usuário deve reservar
- **numero_de_reviews** - Número de comentários(avaliações) dados a cada listagem

- **ultima_review** - Data da última revisão dada à listagem
- **reviews_por_mes** - Número de avaliações fornecidas por mês
- **calculado_host_listings_count** - Quantidade de imóveis por anfitrião.
- **disponibilidade_365** - Número de dias em que o anúncio está disponível para reserva em 365 dias.

Conjuntos de Dados:

O conjunto de dados descreve a atividade de listagem e as métricas do em Nova York em 2019. Ele contém todas as informações necessárias para fazer previsões e fazer inferências sobre anfitriões. Nossos dados, que consistem em 48.895 linhas e 16 colunas.

Os dois principais conjuntos de dados:

• **Numérico**: Alguns atributos numéricos utilizados no desafio são id, host_id, latitude, longitude, mínima_noites, calculado_host_listings_count, disponibilidade_365, número_de_revisões, reviews_por_mes e variável price (classe).

• **Categórico:** Alguns atributos categóricos utilizados no desafio são nome, host_name, bairro_grupo, bairro e rom_type, ultima_review.

ETAPAS REALIZADAS:

Neste desafio foram realizados alguns conceitos estatísticos os dados de forma descritiva e estatística para determinar como as variáveis se correlacionam para gerar hipóteses úteis para futuras tomadas de decisão. É importante analisar os dados cuidadosamente para obter insights significativos que possam ajudar na tomada de melhores decisões de negócios e na compreensão do comportamento do cliente e do anfitrião.

Análise Exploratória de Dados (EDA):

Depois de carregarmos os dados, foi comparado a variável alvo, 'preço', com outras variáveis independentes. Como resultado deste processo, conseguir identificar vários aspectos e relações entre o alvo e as variáveis independentes. Isso nos deu uma idéia melhor de como os recursos se comportam quando comparados às variáveis de destino. Como por exemplos na etapa de pré-processamento dos dados:

- As colunas **host_name** e **nome** possuem menos de 1% de dados faltantes.
- As colunas **ultima_review** e **reviews_por_mes** possuem 20.55% dos dados faltantes. No total de 20.141 dados faltante em todo dataset.
- Foi realizado exclusão das colunas desnecessárias: id, host_name, ultima_review e reviews_por_mes.

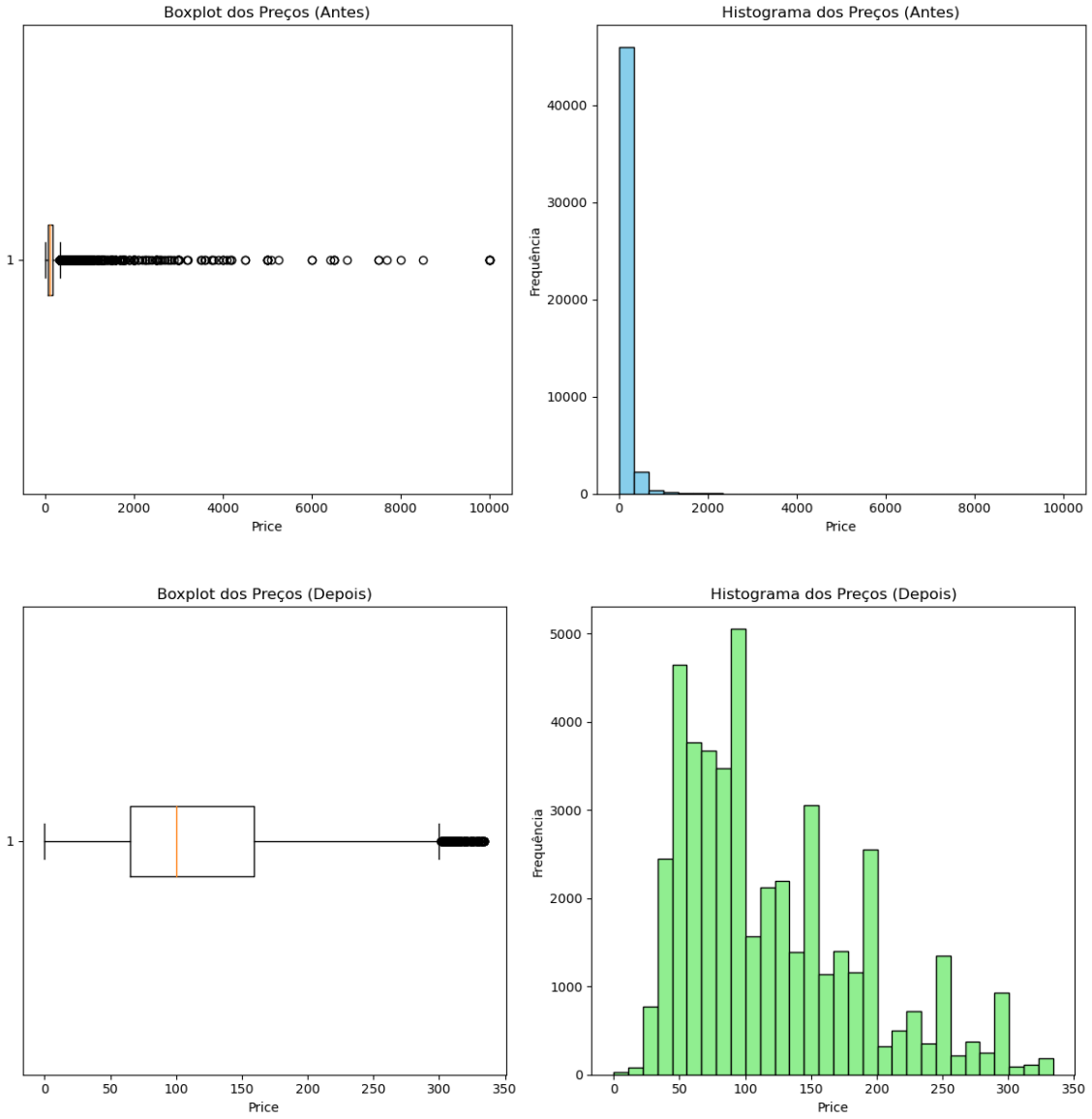
- Na coluna nome foi realizado a substituição dos dados faltante por no_nome e dataset não houve dados duplicados.

A partir do resumo estatístico dos dados pela função describe foi concluído que:

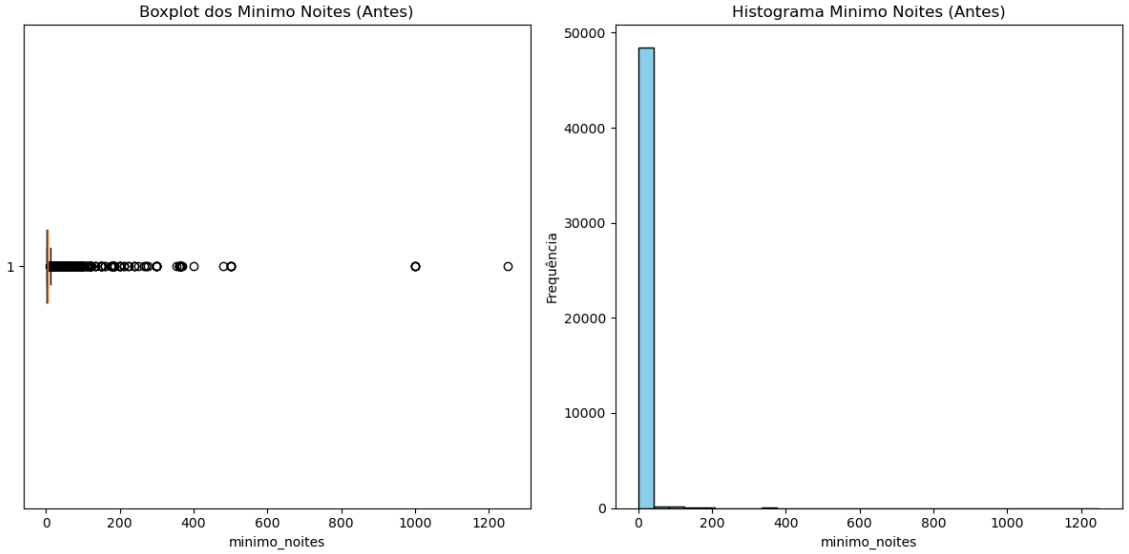
- 50% dos preços dos aluguéis estão abaixo 106 mil dólar e 50% está acima de 106.
- 75% dos preços dos aluguéis são menores ou iguais a 175 mil dólar.
- Na variável minimo_noites o máximo de noites é de 1250, ou seja, mais de um ano (365 dias).

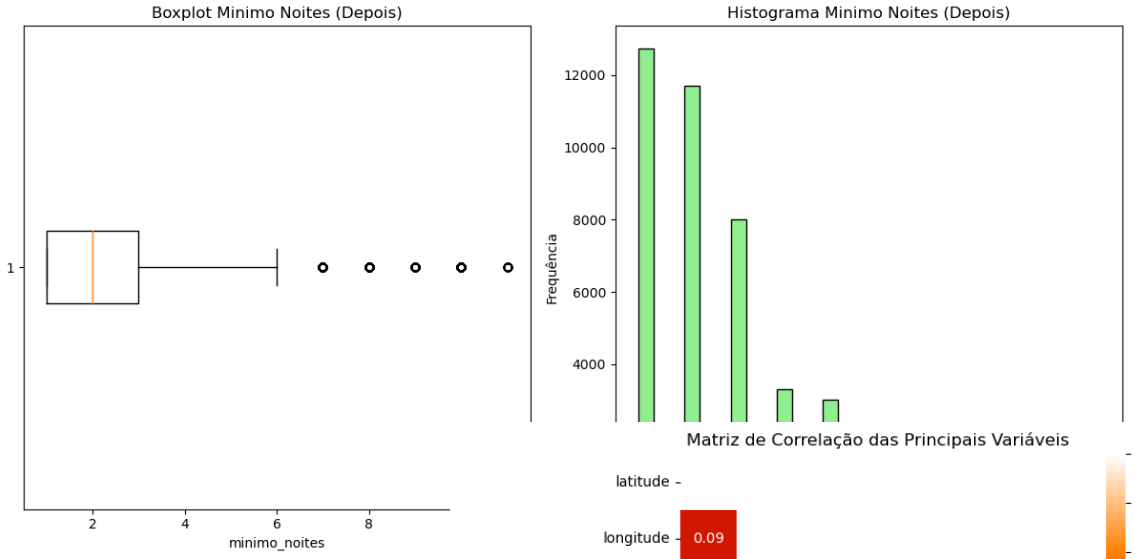
Na etapa de verificação da distribuição das variáveis e detecção de outliers foi utilizado histograma e boxplot. Procedimento foi realizado em duas variáveis: price e mínimo_noite. Onde na variável price foi realizado a remoção de outliers que corresponde 6.08% dos dados e a variável mínimo_noite corresponde em 13.51% dos dados removidos. Como podemos ver nas figuras abaixo do antes e depois da remoção, o quando melhorou muito a distribuição dos dados.

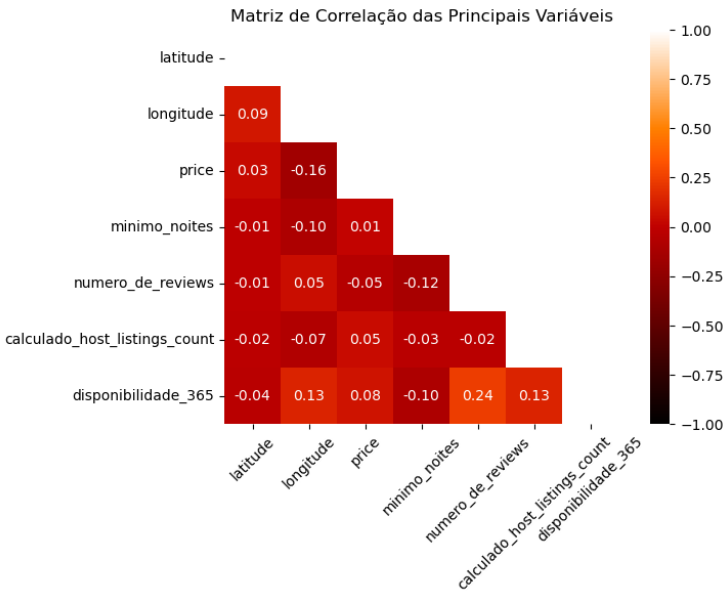
Boxplot e Histograma da variável Price (antes e depois da remoção dos outliers)



Boxplot e Histograma da variável minimo_noite (antes e depois da remoção dos outliers)







Correlação entre as diferentes variáveis:

Ao perceber a análise do coeficiente de correlação de Pearson, não foi encontrado nenhuma variável altamente correlacionada positiva e/ou negativa. Somente moderada

entre disponibilidade_365 e numero_de_reviews.

Principais características entre as variáveis e apresentando algumas hipóteses de negócio relacionadas:

O preço médio é significativamente mais alto (US$ 250.79) para as propriedades localizadas em Manhattan em comparação com outras áreas. Isso sugere que Manhattan é um local mais caro para alugar propriedades. O que para proprietários que desejam maximizar o retorno sobre o investimento, investir em propriedades localizadas em Manhattan pode ser uma estratégia lucrativa devido aos preços médios mais altos.

A maioria das propriedades listadas são casas/apartamentos (cerca de 176.19%), indicando que esse tipo de alojamento é mais comum na área. Os quartos privados têm uma participação significativa (cerca de 78.94%) no Brooklyn, sugerindo que essa área pode ser mais adequada para locação de quartos privados em comparação com outros tipos de acomodação.

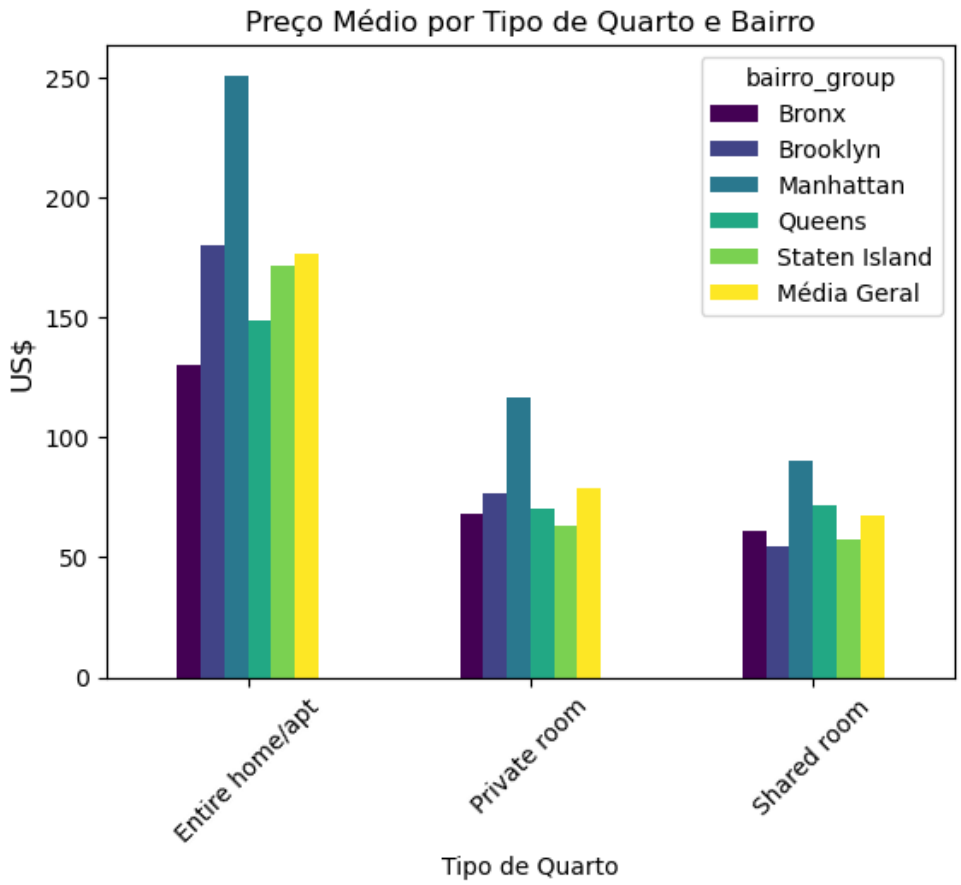
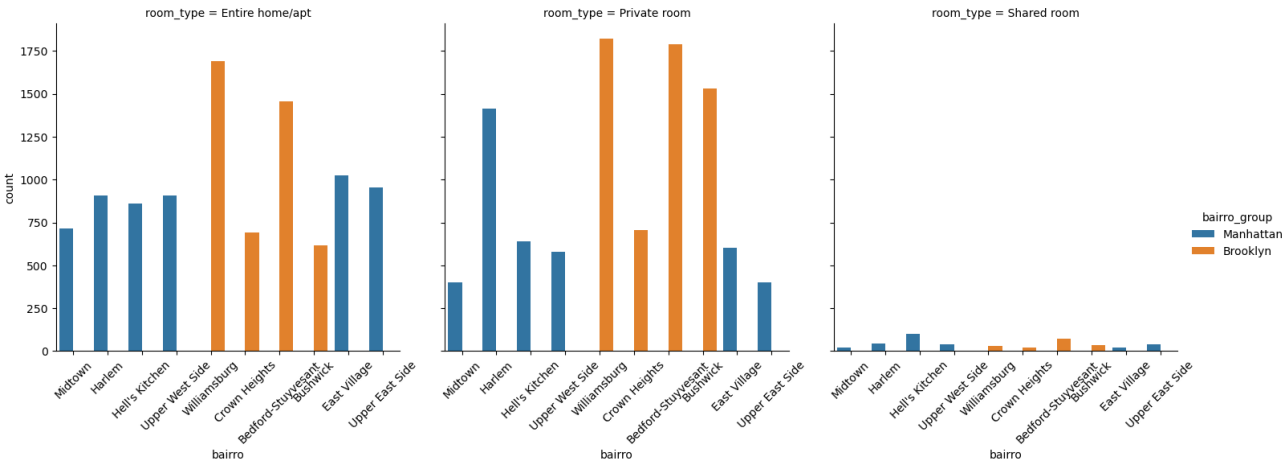


Gráfico de Barra contagem de cada tipo de quarto em Nova York entre os principais grupo de bairros (Manhattan e Brookly)



Uma observação notável é a escassez de anúncios do tipo 'quarto compartilhado' nos 10 bairros mais populosos. A presença desses anúncios está concentrada principalmente em Manhattan e Brooklyn, o que era esperado, considerando que são destinos turísticos populares. Entre esses bairros, Bedford-Stuyvesant e Williamsburg são os mais populares em Manhattan, enquanto Harlem é o mais popular em Brooklyn.

Houve outras análises entre as variáveis, que estão disponíveis no notebook como: preço vs número mínimo de noites, número de reviews vs preço e outros.

Respondendo às seguintes perguntas:

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

Os imóveis mais alugados em New York são os do tipo casa ou apartamento, sendo esses equivalente a 43,18% com preço em média de 210,24 dólar e quarto privado, sendo equivalente a 41,24% com preço em média de 89,74 dólar.

Investir em um apartamento em Manhattan pode ser uma escolha estratégica devido à alta demanda e aos preços médios mais elevados. Propriedades em Manhattan têm um preço médio de $210,24, indicando uma forte demanda nessa área. Investir em apartamentos pode ser mais vantajoso do que casas, considerando que apartamentos representam 43,18% dos aluguéis mais populares em Nova York.

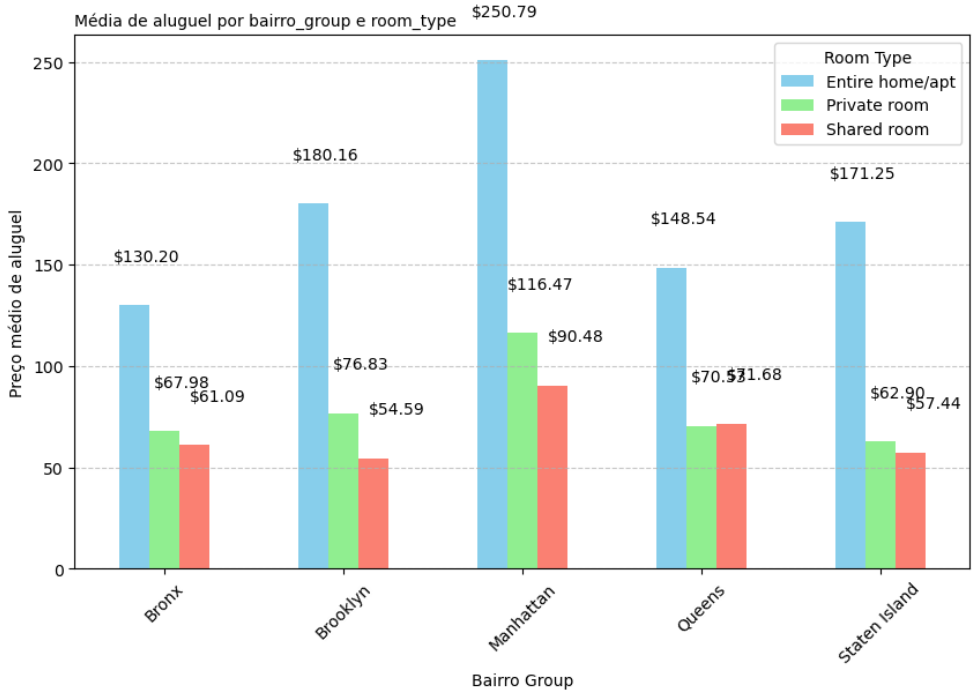
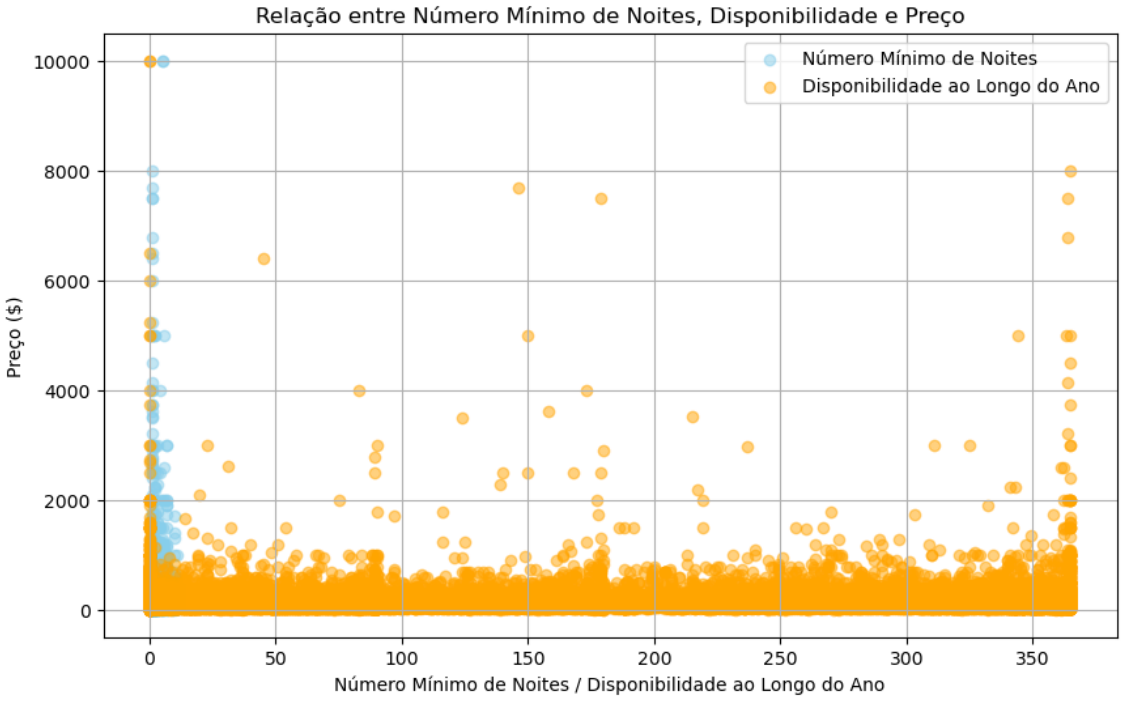


Figura 2 - Média de aluguel por bairro_group e room_type

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



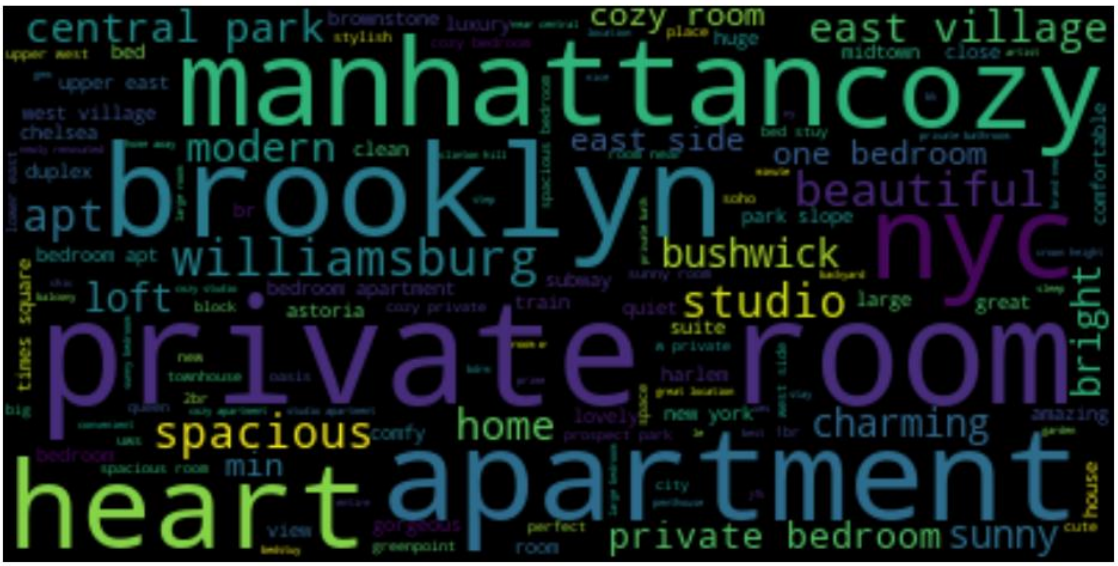
A disponibilidade ao longo do ano também pode afetar o preço das acomodações. Propriedades com alta disponibilidade, ou seja, mais dias disponíveis para reserva ao longo do ano, podem ter preços mais competitivos para atrair reservas e maximizar a ocupação. Por outro lado, propriedades com

disponibilidade limitada podem ter preços mais elevados, especialmente durante períodos de alta demanda, como feriados ou eventos especiais. Acomodações que exigem um número maior de noites mínimas podem ter preços mais elevados.

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

A presença de palavras como "Manhattan" e "Brooklyn" no nome do local sugere que a localização é um fator significativo no valor das acomodações. Manhattan, conhecido por ser um dos distritos mais desejáveis e com alto custo de vida em Nova York, pode ter acomodações com preços mais elevados em comparação com outras áreas. Por outro lado, áreas como Brooklyn podem oferecer opções mais acessíveis, mas ainda valorizadas.

Compreender o padrão no texto do nome do local pode informar as estratégias de marketing para proprietários de acomodações. Destacar a localização privilegiada, como "Manhattan", ou características exclusivas da propriedade, como "private room" ou "apartment", pode atrair viajantes dispostos a pagar mais pelo valor percebido. Os proprietários podem considerar ajustes de preços com base nas palavras-chave identificadas nos nomes dos locais.



Explique como você faria a previsão do **preço** a partir dos dados.

Para prever o preço a partir dos dados, podemos seguir os seguintes passos:

Realizar o pré-processamento dos dados, incluindo limpeza, tratamento de valores ausentes, codificação de variáveis categóricas e normalização. Identificar as features mais relevantes para a previsão do preço. Dividir o conjunto de dados em conjuntos de treinamento e teste. Selecionar um modelo de regressão adequado para prever o preço com base nas características dos dados e nos requisitos do problema. Treinar o modelo selecionado usando o conjunto de treinamento e ajustar os parâmetros. Avaliar o desempenho do modelo usando o conjunto de teste. Isso pode ser feito calculando métricas de desempenho, como RMSE, R^2, MAE, MSE, entre outras.

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

Variáveis Numéricas:

Selecionei todas as variáveis numéricas do conjunto de dados, excluindo as variáveis categóricas. Isso foi feito para garantir que apenas os recursos numéricos fossem considerados na modelagem, já que os modelos de machine learning geralmente trabalham melhor com dados numéricos.

Normalização dos Dados com MinMaxScaler:

Utilizei o MinMaxScaler para normalizar os dados numéricos. A normalização é uma prática comum em modelagem de machine learning para garantir que todas as variáveis tenham a mesma escala, o que pode ajudar os modelos a convergir mais rapidamente e a ter um desempenho mais consistente.

Divisão em Conjunto de Treino e Teste:

Dividi os dados normalizados em conjuntos de treino e teste, com 80% dos dados destinados ao treinamento do modelo e 20% para teste. Essa divisão é importante para avaliar a capacidade do modelo de generalizar para novos dados não vistos durante o treinamento.

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

Estou resolvendo um problema de regressão. Isso porque estamos tentando prever o preço de aluguel de apartamentos com base em diferentes

variáveis independentes, como número mínimo de noites, número de reviews, disponibilidade ao longo do ano, entre outros. Os modelos utilizados são todos regressores (Linear Regression (LR), Ridge Regression (Ridge), Lasso Regression (Lasso), ElasticNet Regression (ElasticNet), K-Nearest Neighbors (KNN), Decision Tree Regressor (CART), Random Forest Regressor (RF), Gradient Boosting Regressor (GBM), XGBoost Regressor (XGBoost), LightGBM Regressor (LightGBM) e CatBoost Regressor (CatBoost)). O que significa que são modelos projetados para realizar previsões numéricas.

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

Nessa análise foi realizada previsão com 11 modelos de regressão, do qual o modelo quem melhor se sobressaiu com bons resultados foi o **modelo LightGBM Regressor (LightGBM - Light Gradient Boosting Machine)** é o que melhor se aproxima dos dados, considerando as métricas de desempenho.

Prós:

- Excelente desempenho em termos de RMSE, R^2 e MAE, indicando uma boa capacidade de fazer previsões precisas.
- Boa eficiência computacional, sendo mais rápido do que alguns dos outros modelos de gradient boosting.
- Lida bem com conjuntos de dados grandes e com alta dimensionalidade.

Contra:

- Pode exigir mais ajustes de hiperparâmetros do que alguns modelos mais simples, como regressão linear.
- Pode ser mais difícil de interpretar do que modelos lineares.

Outros modelos com desempenho semelhante incluem XGBoost, CatBoost e GBM. Todos esses modelos são implementações eficientes de gradient boosting e geralmente oferecem bons resultados em uma variedade de problemas de regressão.

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

A medida de desempenho escolhida para avaliar os modelos foi o Root Mean Squared Error (RMSE), que é uma medida da diferença entre os valores previstos pelo modelo e os valores observados. O RMSE é amplamente utilizado em problemas de regressão porque penaliza mais fortemente os erros maiores, fornecendo uma avaliação mais robusta do desempenho do modelo em relação às previsões.

Além disso, foram fornecidas outras métricas de desempenho, como o coeficiente de determinação (R^2) e o erro absoluto médio (MAE), para fornecer uma visão mais abrangente do desempenho dos modelos. No entanto, o RMSE foi escolhido como a medida principal de desempenho devido à sua interpretabilidade e capacidade de capturar a magnitude dos erros de previsão.

4. 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',

'price': 225,

'minimo_noites': 1,

'numero_de_reviews': 45,

'ultima_review': '2019-05-21',

'reviews_por_mes': 0.38,

'calculado_host_listings_count': 2,

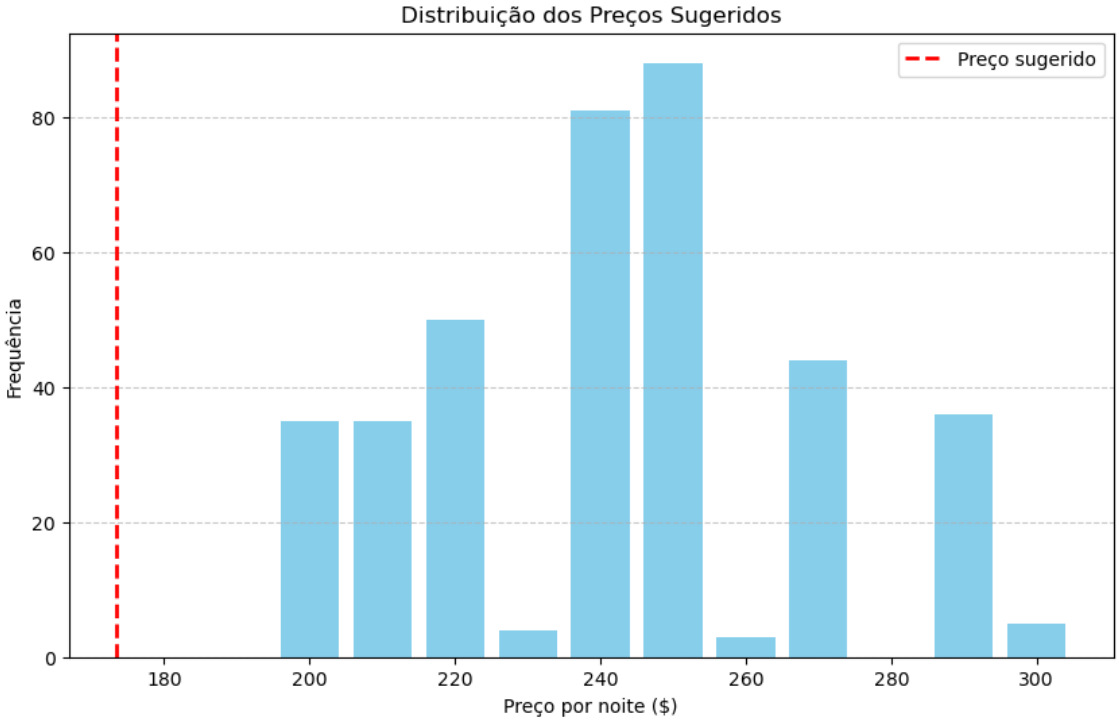
'disponibilidade_365': 355}

Qual seria a sua sugestão de preço?

Com base nos dados fornecidos para o apartamento 'Skylit Midtown Castle', podemos obter os seguintes insights de negócio: O preço atual do apartamento é de 225 por noite, a sugestão de preço calculada com base nos dados disponíveis é de **$ 173.39 por noite.** Isso indica que o preço atual está acima da sugestão calculada, sugerindo que pode haver uma oportunidade de ajuste de preço para melhorar a competitividade no mercado.

O apartamento recebeu um total de 45 avaliações. Um número relativamente alto de avaliações pode indicar uma demanda saudável pela propriedade e pode influenciar a percepção do valor pelos potenciais hóspedes.

A disponibilidade ao longo do ano é alta, com 355 dias disponíveis. Isso sugere que o apartamento está disponível para aluguel na maioria dos dias do ano, o que pode influenciar a estratégia de precificação.


Distribuição dos Preços Sugeridos
Preço sugerido
Frequência
0
20
40
60
80
180
200
220
240
260
280
300
Preço por noite ($)