Desafio técnico de Ciências de dados

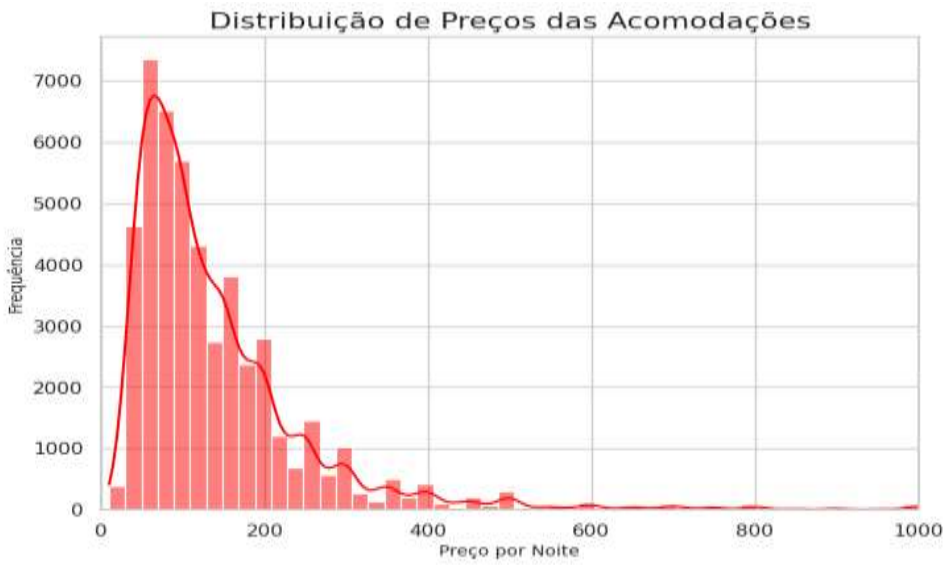
Matheus Mata

Análise exploratória dos dados (EDA)

Comecei utilizando o ‘df_testeIn.info()’ para visualizar melhor as colunas, análisar se há valores nulos e quais tipos de dados nós temos para uma possível transformação. Após essa análise utilizei o ‘df_testeIn.isnull().sum()’ para saber quantos valores nulos tem cada coluna e isso é importante para treinamento de modelos de machine learning pois dados faltantes podem prejudicar o treinamento.

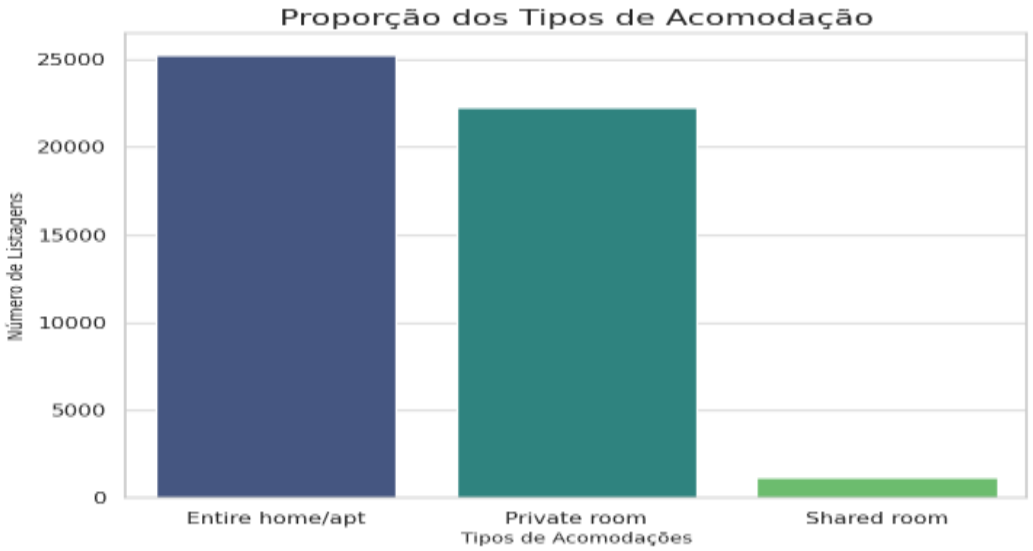
Utilizei o ‘df_testeIn.describe()’ para análise de dados estatísticos e identificar possíveis discrepâncias indicando outliers. Após isso fiz a limpeza de dados substituindo valores vazios e removendo linhas com possíveis outliers. Utilizei o ‘df_testeIn.head()’ para visualizar as primeiras linhas do dataframe já com os dados tratados para ver como ficou com o tratamento.

Construção de gráficos exploratórios

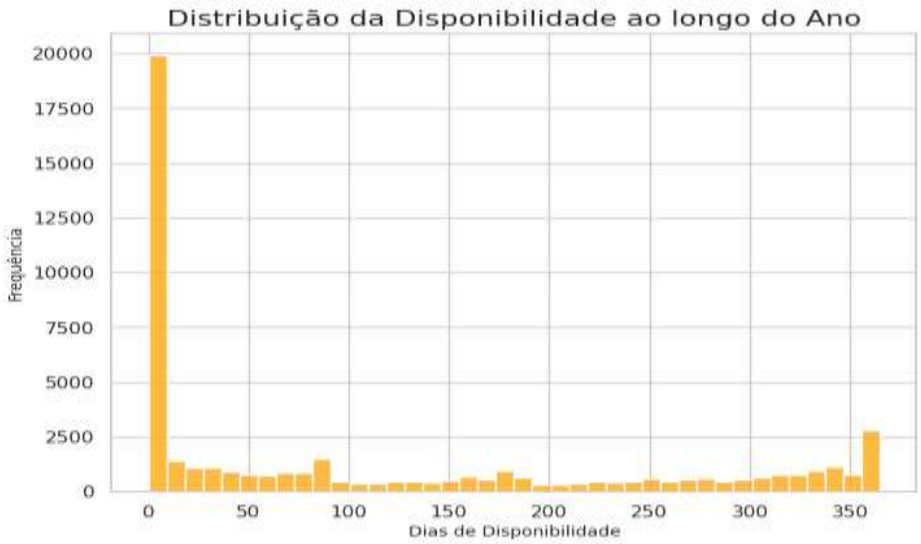


Nesse histograma podemos observar que a maioria dos preços está concentrada abaixo de $500 por noite, com uma alta densidade em valores baixos na faixa de $50 e $200.

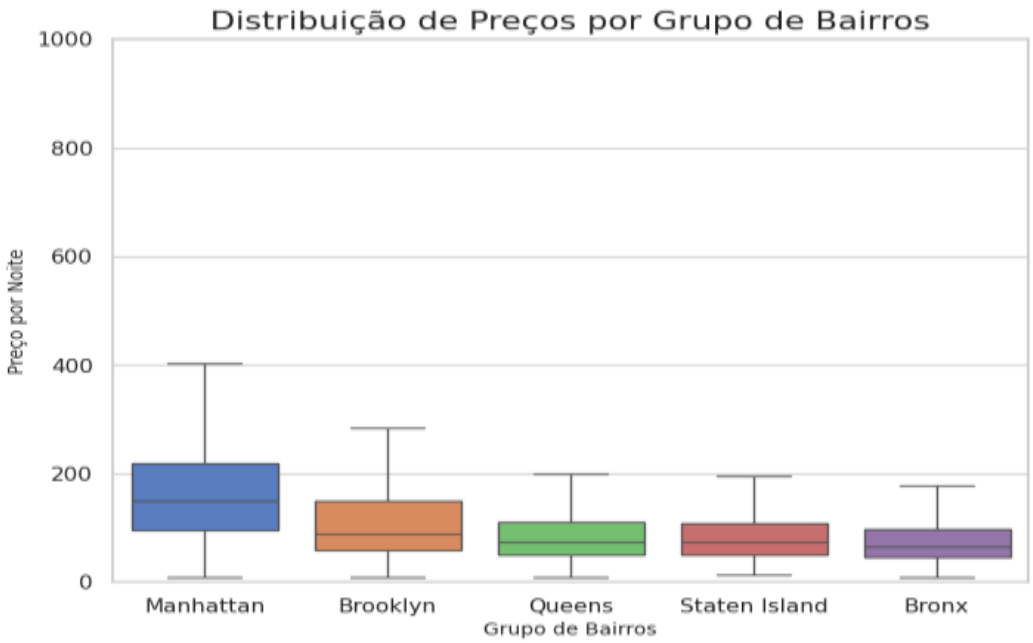
A curva de densidade mostra um decaimento acentuado para preços mais altos. Esses resultados sugerem uma preferência por acomodações com preços mais acessíveis.



Nesse gráfico de barras podemos que a acomodação “Entire home/apt” é o tipo mais comum, seguido por "Private room" . Tipos como "Shared room" e "Hotel room" têm uma frequência significativamente menor.



Nesse gráfico podemos observar que muitos imóveis estão disponíveis o ano inteiro , enquanto outros têm disponibilidade limitada. A concentração em 0 dias aparenta ser listagens que estão desativadas ou indisponíveis.



Nesse boxplot podemos observar que "Manhattan" apresenta os preços mais altos, enquanto "Bronx" tem valores mais baixos. Essa variação reflete diferenças na atratividade e infraestrutura de cada grupo de bairros.

Conclusões

A análise exploratória dos preços por grupos de bairros revelou que Manhattan é a região com os preços mais altos, isso pode revelar que é a área mais cara e

provavelmente a mais turística. O Bronx apresentou médias dos preços mais baixos, podendo ser uma opção mais acessível para um público com poder aquisitivo mais baixo e bairros como Brooklyn mostraram-se intermediários em termos de custo, combinando acessibilidade e atratividade.

A dispersão dos preços varia entre os grupos, por exemplo, Manhattan possui maior variação, sugerindo que possui acomodações econômicas e também propriedades de luxo, enquanto o Bronx tem uma faixa de preços mais consistente. A decisão de excluir possíveis outliers permitiu focar nos preços representativos, reforçando que a maioria das acomodações está em uma faixa acessível, abaixo de $1000 por noite.

Para consumidores, bairros mais econômicos como o Bronx podem ser atrativos para economia, enquanto Manhattan atende consumidores que buscam luxo e proximidade de possíveis atrações.

Esse tipo de análise nos ajuda a entender melhor o mercado de locação de imóveis e qual público de cada localização e com essa informação podemos orientar estratégias de marketing e otimização de preços para diferentes perfis de público.

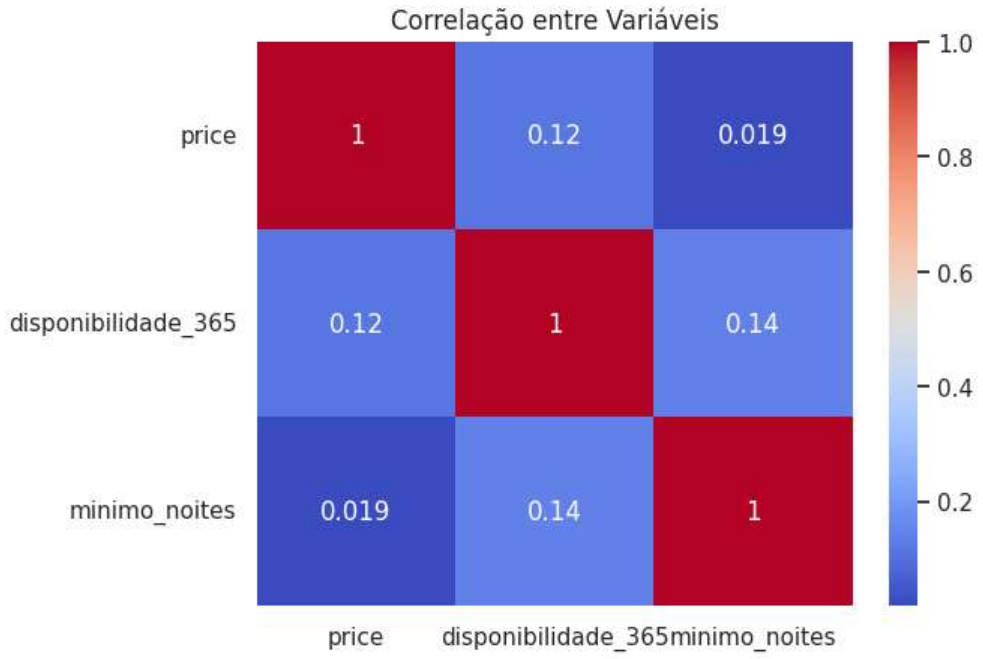
2.

a – A escolha em qual apartamento investir vai depender de quanto a pessoa tá disposta a investir por exemplo se ela está disposta a investir um pouco mais alto eu diria que Manhattan é o ideal porque o bairro possui os preços mais altos e maior atratividade para turistas dispostos a pagar por acomodações próximas a pontos turísticos, eventos e áreas de negócios podendo trazer um retorno maior com um tempo.

b – Analisei as correlações entre as variáveis para descobrir se há uma interferência das variáveis (mínimo_noites e disponibilidade_365) nos preços.

Após análise utilizando uma tabela e um heatmap podemos observar que a correlação entre elas são bem fracas ou inexistente indicando que há pouca ou nenhuma interferência das variáveis no preço

	price	disponibilidade_365	minimo_noites
price	1.000000	0.117918	0.019487
disponibilidade_365	0.117918	1.000000	0.143983
minimo_noites	0.019487	0.143983	1.000000



c - Sim, existem padrões observáveis nos nomes de locais associados a valores mais altos. As palavra como "East" e “Village” sugerem localização de alto padrão, a palavra “park” sugere pontos turísticos como o Central Park. Os termos “luxury” e “loft” são os termos com mais altos preços já que sugerem um estilo de vida mais sofisticado e de poder aquisitivo mais alto o número “2” pode indicar o número de listagem por host elevando mais o preço.

```
[('in', 1906),
 ('2', 1203),
 ('bedroom', 1147),
 ('apartment', 947),
 ('apt', 828),
 ('luxury', 655),
 ('village', 633),
 ('the', 630),
 ('w', 613),
 ('loft', 601),
 ('1', 598),
 ('park', 550),
 ('east', 531),
 ('of', 526),
 ('with', 522)]
```

3 - Como objetivo é prever o preço de uma acomodação com base em suas características isso indica ser um problema de regressão, pois estamos lidando com uma variável dependente contínua que é preço.

Primeiro eu removi colunas ultima_review, id, 'nome', 'host_id', 'host_name' que não influenciarão no treino do modelo, pois não estão ligadas diretamente ao preço codifiquei variáveis categóricas como bairro_group, bairro, room_type porque os modelos de ML para ter um funcionamento correto precisa que as variáveis estejam transformadas em números em seguida separei as variáveis independentes e dependente para melhor leitura dos modelos e separei os conjuntos de dados em dados de treino e teste que foi 80% para treino e 20% para teste o que é um número padrão.

Utilizei validação cruzada para treinamento do modelo que é uma técnica muito importante para avaliar o quão bem o modelo consegue fazer as previsões em dados novos e desconhecidos e evitar um overfitting e como o problema é uma questão de regressão usei como métricas para avaliação do modelo as métricas MAE, RSME, R^2 que são muito utilizadas para medir a acurácia do modelo em problemas que envolvem regressão.

Para treinamento utilizei os 3 modelos mais comuns para previsão de preços e problemas que envolvem regressão. Os modelos escolhidos foram Linear Regression, KNN e Random Forest. Fazendo uma breve descrição das métricas para entendermos qual modelo performou melhor:

• **MAE (Mean Absolute Error):** Representa o erro médio absoluto, ou seja, a média da diferença absoluta entre os valores reais e as previsões do modelo. Quanto menor o MAE, melhor.
• **RMSE (Root Mean Squared Error):** Representa a raiz quadrada do erro quadrático médio. É uma métrica que penaliza mais os erros grandes. Quanto menor o RMSE, melhor.
• **R^2 (Coeficiente de Determinação):** Indica a proporção da variância dos dados que é explicada pelo modelo. Varia de 0 a 1, onde 1 indica que o modelo explica perfeitamente a variância dos dados. Quanto maior o R^2, melhor.

Depois de avaliar os resultados do treinamento dos modelos podemos observar que o Modelo Random Forest se saiu melhor que os outros, então utilizarei esse modelo para previsão dos preços e salientando que os hiperparâmetros dos modelos escolhi o que mais é comum usar em problemas desse tipo. As vantagens do Random Forest são muitas as principais são:

- É um modelo mais flexível que a Regressão Linear e consegue capturar relações mais complexas entre as variáveis.
- É menos sensível a outliers do que outros modelos, como o KNN.
- A utilização de múltiplas árvores no Random Forest ajuda a reduzir o risco de overfitting, ou seja, o modelo se ajustar muito aos dados de treinamento e não generalizar bem para novos dados gerando uma performance ruim.

Mas todo modelo tem suas desvantagens e as principais são:

- O Random Forest é mais difícil de interpretar. É complicado entender como cada variável está influenciando a previsão final, já que o modelo se baseia em várias árvores de decisão complexas.
- A visualização do modelo e de suas decisões também é mais complexa, dificultando a análise do processo de tomada de decisão.
- O treinamento de um Random Forest pode ser mais demorado que o de modelos mais simples, especialmente quando o número de árvores e a profundidade das árvores são grandes. E o consumo de memória também é bem grande.

4 – Minha sugestão de preço do apartamento depois do treinamento do modelo e fazer a previsão são de $206,59.