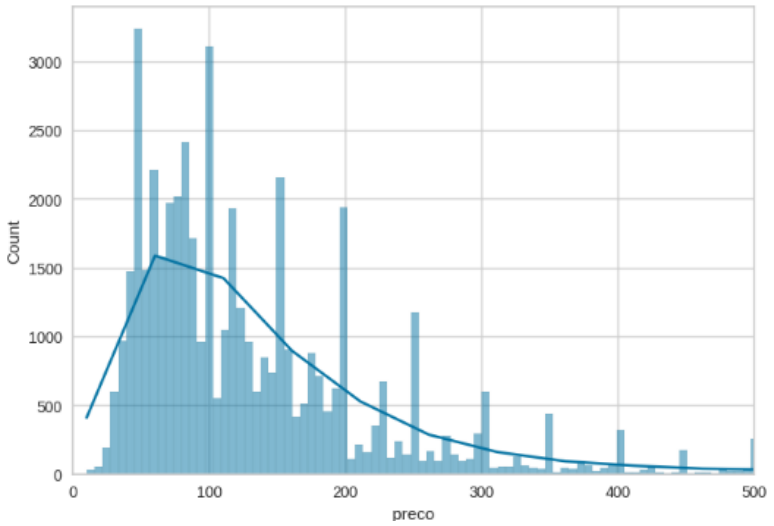
Relatório EDA - Desafio Indicium
Nome : Brena dos Santos Freitas

1. Faça uma análise exploratória dos dados (EDA), demonstrando as principais características entre as variáveis e apresentando algumas hipóteses de negócio relacionadas. Seja criativo!

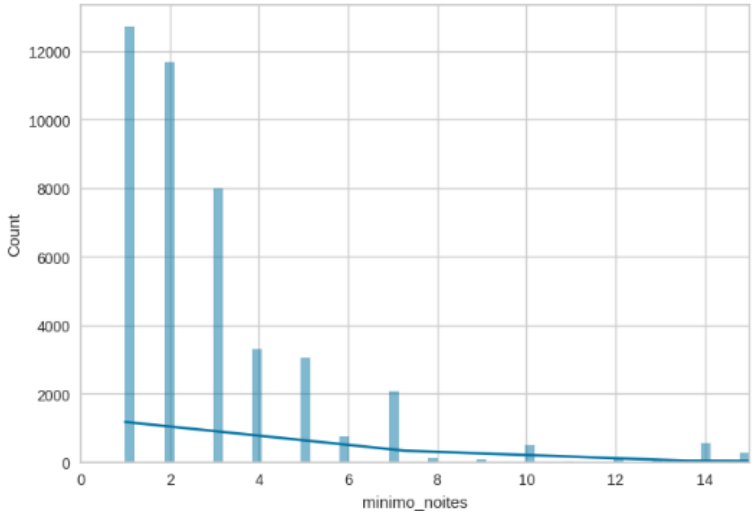
Imagem 1: Histograma de preço



Fonte : Autoria própria

Ao observar o histograma de preço , percebe-se que ele está mais distorcido para a esquerda, ou seja, há frequência de aluguéis com valores mais baixos do que aluguéis com valores mais altos, assim os valores mais frequentes estão entre 40 até 200 dólares por noite.

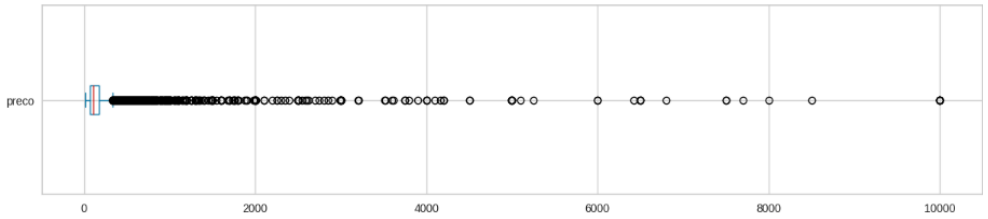
Imagem 2: Histograma de mínimo noites



Fonte : Autoria própria

Ao observar o histograma de mínimo noites, percebe-se que ele está distorcido à direita o que indica que a faixa de quantidade mínima de noites e 1 até 7 noites é a faixa mais frequente entre os aluguéis.

Imagem 3: Boxplot para a variável preço



Fonte : Autoria própria

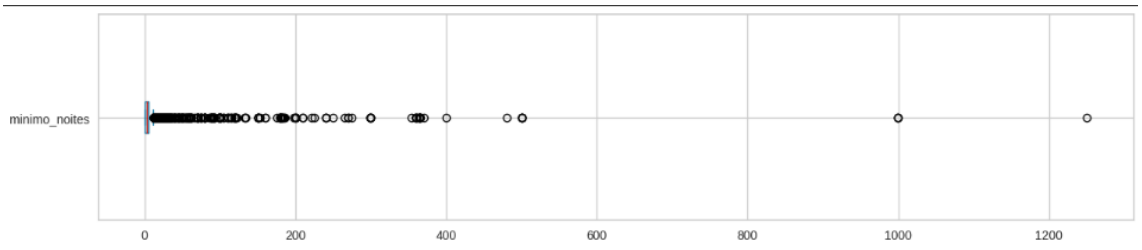
Imagem 4: Resultado do cálculo lugares com aluguéis acima e abaixo de $1000

```
Há 298 lugares com preços maiores que 1000.
Há 48655 lugares com preços menores ou igual a 1000.
```

Fonte : Autoria própria

Observando o boxplot para a variável preço , percebe-se alguns valores outliers, ou seja, são valores discrepantes em relação ao resto do conjunto de dados. Ao calcular os valores acima de 1000(mil) dólares, percebe-se que a uma quantidade de somente 298 lugares que cobram acima de 1000(mil) dólares, e há cerca de 48000(quarenta e oito mil) lugares que cobram abaixo de 1000(mil) dólares. Esses outliers podem trazer prejuízos na análise de dados, sendo necessário tratar esses valores de alguma forma, como retirá-los do conjunto de dados.

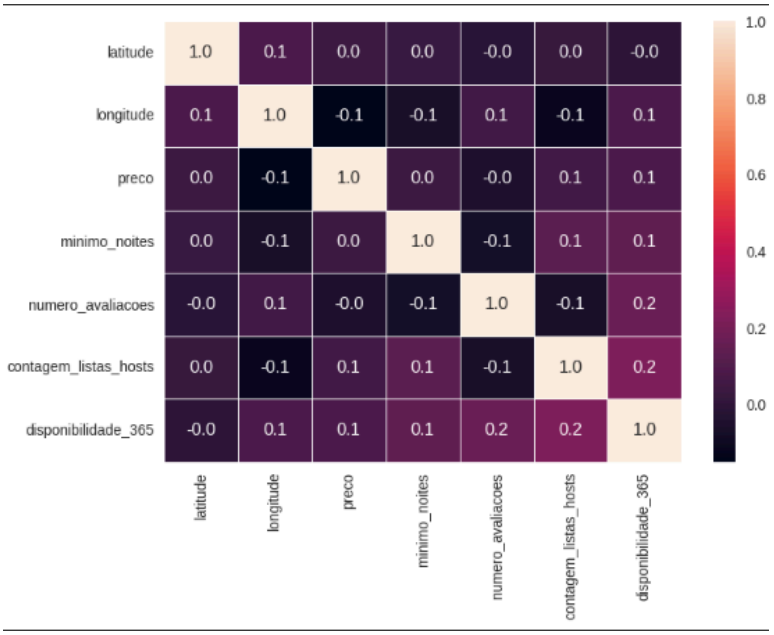
Imagem 5: Boxplot para a variável mínimo noites



Fonte : Autoria própria

Observando o boxplot para a variável minimo_noites, percebe-se alguns valores outliers, ou seja, são valores discrepantes em relação ao resto do conjunto de dados.

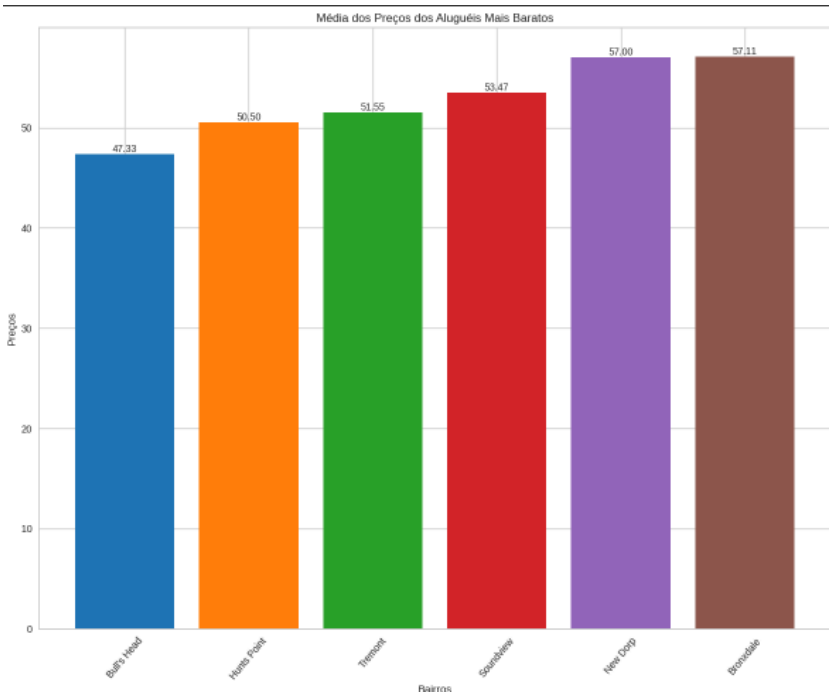
Imagem 6: Boxplot para a variável mínimo noites



Fonte : Autoria própria

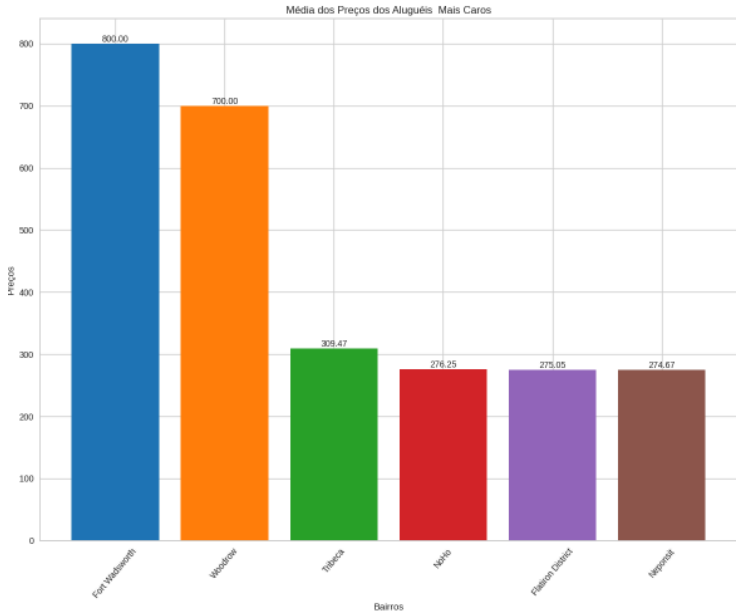
2. Responda também às seguintes perguntas:
   a. Supondo que uma pessoa esteja pensando em investir em um apartamento para alugar na plataforma, onde seria mais indicada a compra?

Imagem 7 : Média de preços dos Aluguéis mais baratos



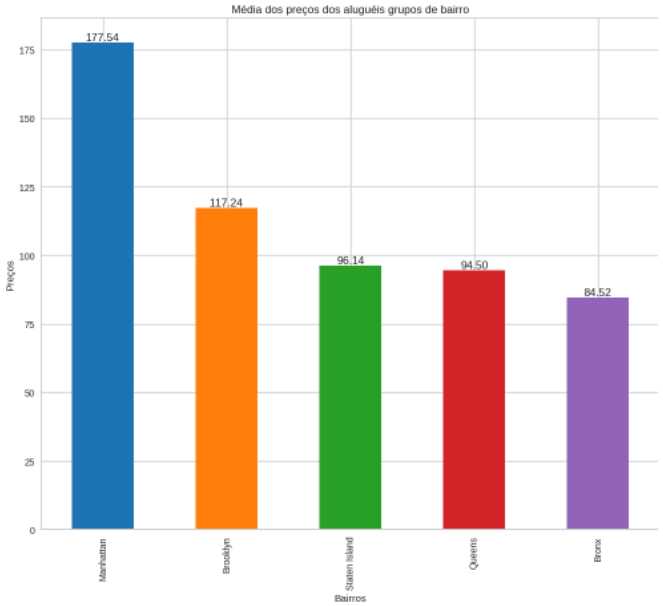
Fonte : Autoria própria

Imagem 8 : Média de preços dos Aluguéis mais caros



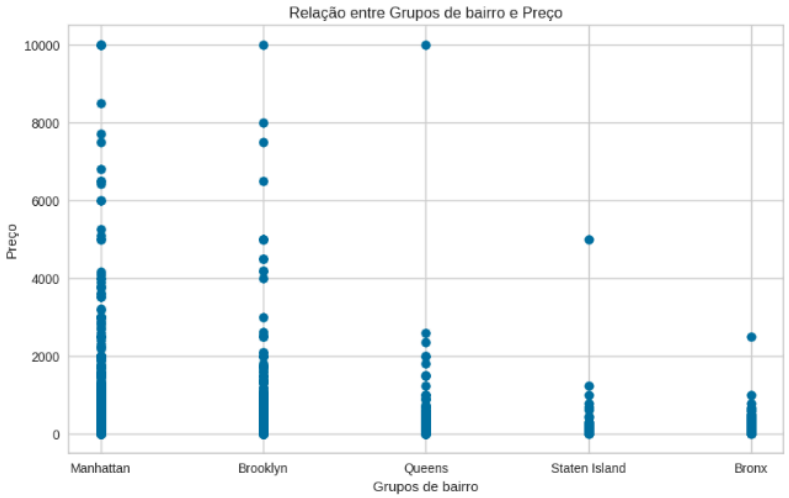
Fonte : Autoria própria

Imagem 9 : Média de preços dos Aluguéis nos grupos de bairro



Fonte : Autoria própria

Imagem 10 : Média de preços dos Aluguéis nos grupos de bairro



Fonte : Autoria própria

A partir da visualização da média de preços dos aluguéis, é válido dizer que o investimento em alugar um apartamento pode variar dependendo da condição de cada indivíduo (o quanto pode investir para alugar um local), público-alvo , localização do imóvel. Quanto à localização, há locais que são mais baratos, mas estão em regiões menos seguras de New York.

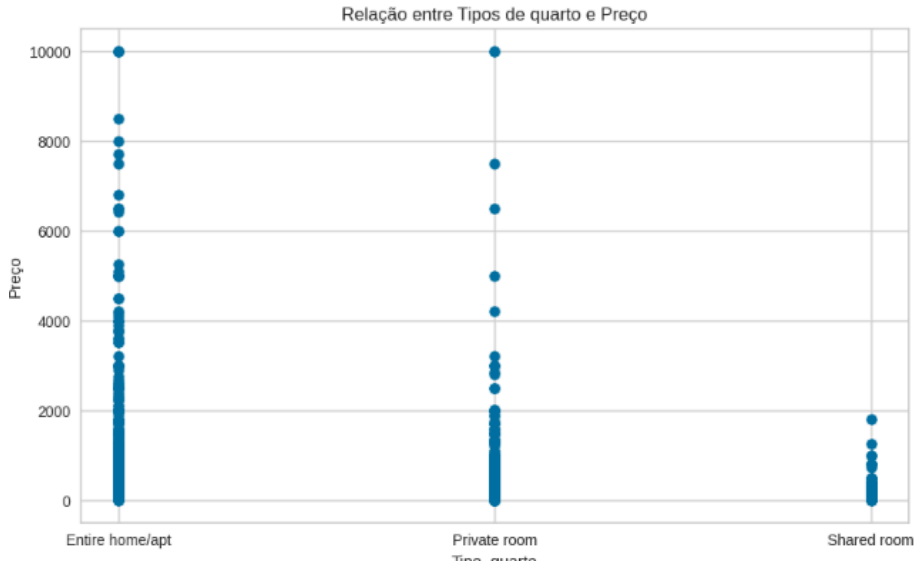
Em situação em que um indivíduo pretenda alugar em um lugar mais barato, pode-se indicar os bairros do gráfico “Média de preços dos Aluguéis Mais Baratos”. Em vista desses bairros em que há menor precificação de aluguéis, em suma estão localizados no Bronx, que é a região mais pobre de New York, é distante de pontos turísticos importantes da cidade e não é tão seguro.

Sobre o perfil(público-alvo), é necessário avaliar se o indivíduo é um turista e prefere uma localização perto dos principais pontos turísticos ou se é um estudante e vai passar longos períodos em determinado local e quer buscar um lugar barato.

No caso em que os indivíduos tiverem uma boa condição e optarem por investir em uma boa localização, podem optar por escolherem alugar em Manhattan , onde se concentram os aluguéis mais caros de acordo com o gráfico “Média dos preços de aluguéis mais caros” e também no gráfico “ Média dos preços de aluguéis grupos de

bairro”. Em Manhattan, há bastante segurança , é perto de importantes pontos turísticos da cidade de New York e é perto de outros serviços importantes.

Imagem 11: Relação entre tipos de quarto e preço



Fonte : Autoria própria

Imagem 12: Código de agrupamento entre ‘tipo_quarto’ e ‘bairro_preco’

```
[61] private_room = data[data['tipo_quarto'] == 'Private room'].groupby('bairro_grupo')['tipo_quarto'].count()

     # Encontrar o grupo de bairro com a maior contagem de quartos privados
     bairro = private_room.idxmax()
     maior_concentracao_p = private_room.max()

     print(f'O grupo de bairro com a maior concentração de quartos privados é "{bairro}", com {maior_concentracao_p} quartos privados.')

     O grupo de bairro com a maior concentração de quartos privados é "Brooklyn", com 10131 quartos privados.

[40] shared_room = data[data['tipo_quarto'] == 'Shared room'].groupby('bairro_grupo')['tipo_quarto'].count()

     # Encontrar o grupo de bairro com a maior contagem de quartos privados
     bairro_shared_room = shared_room.idxmax()
     maior_concentracao_s = shared_room.max()

     print(f'O grupo de bairro com a maior concentração de quartos compartilhados é "{bairro_shared_room}", com {maior_concentracao_s} quartos compartilhados.')

     O grupo de bairro com a maior concentração de quartos compartilhados é "Manhattan", com 480 quartos compartilhados.

[41] entire_home =data[data['tipo_quarto'] == 'Entire home/apt'].groupby('bairro_grupo')['tipo_quarto'].count()

     bairro_entire_home = entire_home.idxmax()
     maior_concentracao_e = entire_home.max()

     print(f'O grupo de bairro com a maior concentração de lugares do tipo entire_home/apt é "{bairro_entire_home}", com {maior_concentracao_e} entire home')

     O grupo de bairro com a maior concentração de lugares do tipo entire_home/apt é "Manhattan", com 13192 entire home
```

Fonte : Autoria própria

Além disso, o tipo de quarto para se alugar também consta como uma das características importantes ao escolher a localização, visto que os “shared room”(Quartos compartilhados) em sua maioria são mais baratos do que alugar em um “Entire home/apt” ou “Private/room”, que são casas/ apartamentos e quartos privados, respectivamente.

Observando o agrupamento entre as variáveis “tipo_quarto” e “preço”, conclui-se que o Brooklyn possui 10131 quartos privados, ou seja, concentra a maior quantidade de “Private Room” para alugar. Já Manhattan, possui 480 quartos compartilhados(“Shared room”) e 13192 lugares do tipo “Entire home/apt”, ou seja, concentra esses dois tipos de imóveis.

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

Imagem 13 : Correlação entre mínimo de noites, preço e disponibilidade

	preco	minimo_noites	disponibilidade_365
preco	1.000000	0.035409	0.117837
minimo_noites	0.035409	1.000000	0.219442
disponibilidade_365	0.117837	0.219442	1.000000

Fonte : Autoria própria

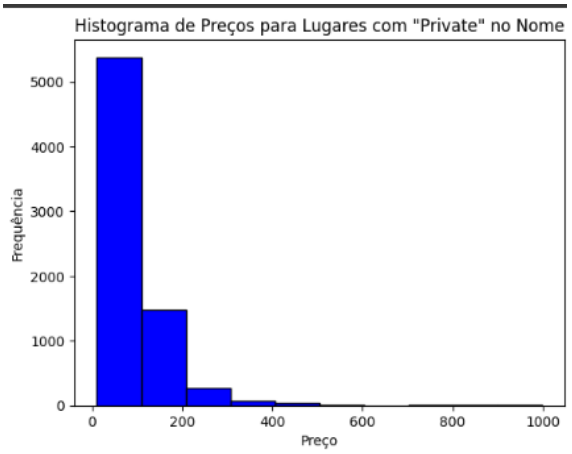
Imagem 14 : Distribuição de densidade para mínimo de noites

Fonte : Autoria própria

Por meio do cálculo de correlação linear é percebido que as variáveis preço, minimo_noites e disponibilidade 365 possuem uma correlação linear baixa.

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

Imagem 15 : Histograma de preços para lugares com “private” no nome



Fonte : Autoria própria

Ao observar o histograma, percebe-se que os lugares com “private” no nome, em sua maioria custam menos de 200 dólares.

3. 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?

Para melhorar a análise de dados, foi necessário o tratamento de dados como, a retirada de linhas com valores nulos/ausentes, retirada de linhas com valores duplicados. Houve a transformação da coluna “tipo_quarto” para outras 3 colunas na tabela do conjunto de dados. Estas colunas são “Entire home/apt”, “Private room”, “Shared room”, que foram transformadas por meio do One Hot encoder, um tipo de codificador para variáveis categóricas, em que para cada categoria indica 1(para verdadeiro) e 0 (para falso).

Além disso, utilizou-se o Label Encoder para decodificação das variáveis categóricas “bairro” e “ bairro_grupo”, em que cada campo passou a ter uma valor numérico. Após isso, houve o escalonamento dos atributos para que estes ficassem na mesma escala a fim de melhorar o treinamento do conjunto de dados.

O desafio envolve um problema de regressão, visto que estamos querendo prever um valor numérico, no caso, o preço dos aluguéis dos imóveis para alugar na cidade de New York. Não é considerado do tipo de classificação, pois para este caso

estaríamos querendo prever uma categoria específica, como no caso de prever se algo é 0 ou 1.

Para a criação do modelo, foi necessário a divisão dos previsores da classe ‘preço’, a qual queremos prever.

Ainda existem algumas inconsistências, o que dificulta com que os modelos aprendam padrões significativos, o que explique a acurácia baixa nos modelos, além da precisão , f1-score, recall e score baixos. Além disso, é importante reajustar os parâmetros de cada tipo de modelo para melhorar a precisão deles.