Relatório do desafio de ciência de dados do processo seletivo Indicium

Nome: João Vitor Vargas Soares

Sumário

Contexto

Relatório do desafio de ciência de dados feito pelo candidato para o processo seletivo da empresa Indicium. Este documento foi desenvolvido com base no arquivo fornecido pela empresa, cobrindo todos os tópicos solicitados.

Análise Exploratória de Dados (EDA)

Para a realização da análise exploratória de dados, foi feita a investigação de dados do dataset fornecido. O arquivo conta com 48894 linhas e 16 colunas. A descrição de cada coluna está presente abaixo.

- **id** - Atua como uma chave exclusiva para cada anúncio nos dados do aplicativo
- **nome** - Representa o nome do anúncio
- **host_id** - Representa o id do usuário que hospedou o anúncio
- **host_name** - Contém o nome do usuário que hospedou o anúncio
- **bairro_group** - Contém o nome do bairro onde o anúncio está localizado
- **bairro** - Contém o nome da área onde o anúncio está localizado

- **latitude** - Contém a latitude do local
- **longitude** - Contém a longitude do local
- **room_type** - Contém o tipo de espaço de cada anúncio
- **price** - Contém o preço por noite em dólares listado pelo anfitrião
- **minimo_noites** - Contém o número mínimo de noites que o usuário deve reservar
- **numero_de_reviews** - Contém o número de comentários dados a cada listagem
- **ultima_review** - Contém a data da última revisão dada à listagem
- **reviews_por_mes** - Contém o número de avaliações fornecidas por mês
- **calculado_host_listings_count** - Contém a quantidade de listagem por host
- **disponibilidade_365** - Contém o número de dias em que o anúncio está disponível para reserva

Pôde-se então verificar quais variáveis são categóricas e quais são quantitativas, o que é mostrado na tabela abaixo.

Variáveis categóricas	**Variáveis quantitativas**
id	latitude
name	longitude
host_id	price
host_name	mínimo_noites
bairro_group	numero_reviews
bairro	reviews_por_mes
room_type	calculado_host_listings_count
	disponibilidade_365
	ultima_review

Tabela 1: Classificação das variáveis

Em seguida, foram analisadas informações relacionadas ao dataset para que se entenda seu formato e a qualidade dos dados. As características analisadas foram a quantidade entradas únicas em cada coluna, a quantidade de valores zero em cada coluna e o número de linhas duplicadas. Todos esses fatores pareciam estar normais, não havendo necessidade de performar uma limpeza de dados até esse ponto.

Observou-se a distribuição das variáveis numéricas para se ter mais informações sobre os dados. A Tabela 2 foi gerada contendo dados estatísticos sobre cada coluna.

	price	disponibilidade_365	minimo_noites	numero_de_reviews	reviews_por_mes
total	48894	48894	48894	48894	38842
média	152,72	112,78	7,03	23,27	1,37
std	240,15	131,62	20,51	44,55	1,68
min	0	0	1	0	0,01
25%	69	0	1	1	0,19
50%	106	45	3	5	0,72
75%	175	227	5	24	2,02
max	10000	365	1250	629	58,5

Tabela 2: Dados estatísticos das colunas selecionadas

Como se pode ver as colunas *price* e *minimum_nights* chamam atenção. A coluna *price* possui média de 152,7, porém possui valor máximo de 10.000. Além disso, possui valores iguais a zero. Com relação à coluna *mínimo_noites*, sua média é de 7,03, porém o valor máximo é de 1250.Isso nos leva a concluir que há outliers que devem ser tratados na hora da limpeza dos dados.

Para visualizar a distribuição dos dados, gráficos do tipo histograma foram plotados, mostrando novamente que as colunas *price* e *minimum_nights* estão mal distribuídas. Os histogramas estão apresentados na Figura 1 a seguir. As outras colunas aparentam estar corretas segundo interpretação.

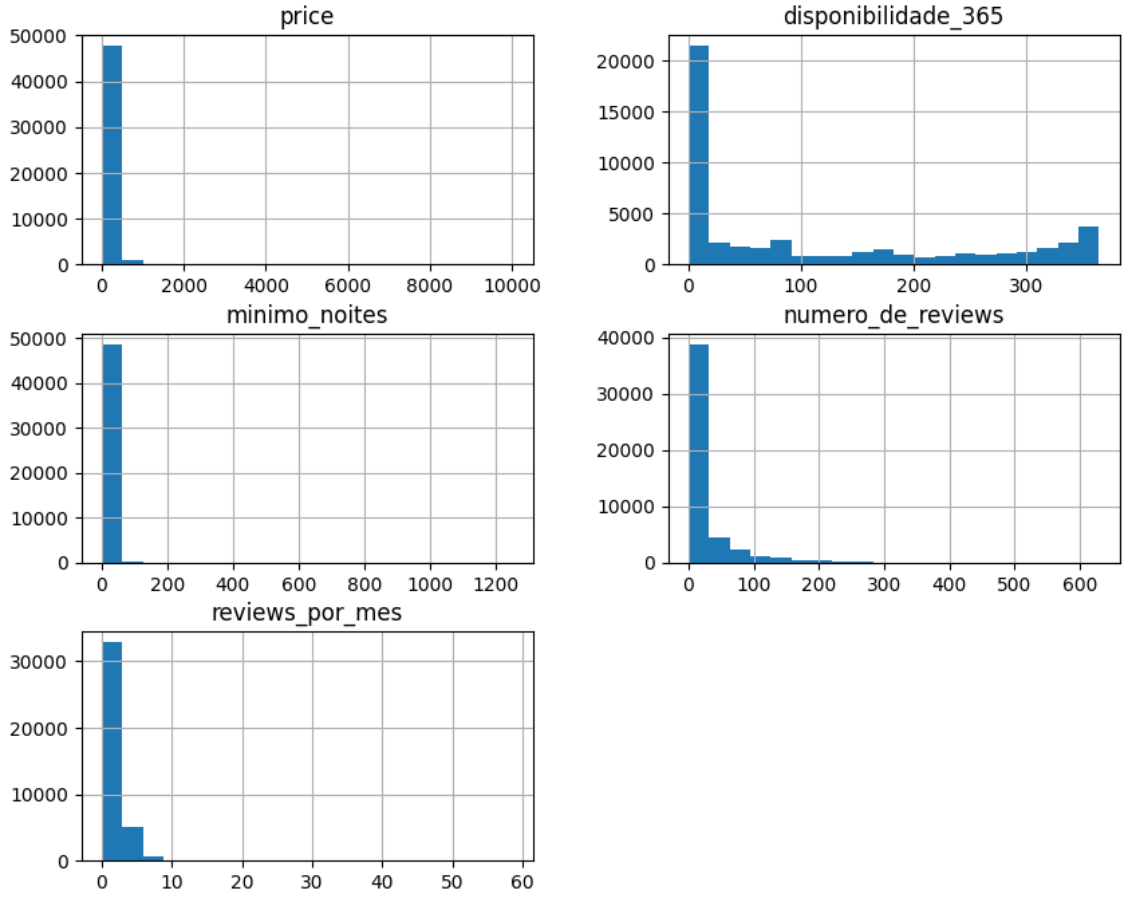

Figura 1: Histogramas das distribuições das variáveis quantitativas

Para se identificar os outliers das variáveis *price* e *minimum_nights* foram usados gráficos do tipo boxenplot, que são parecidos com boxplots porém, por possuírem mais quantis, é possível melhor analisar as distribuições na cauda, onde estão presentes os outliers. Primeiramente foi plotado o gráfico para o preço, onde foi identificado que a partir de US$600, os valores não contribuíam e poderiam distorcer a análise. Eles representam apenas 1,59% de todas as listagens, e os valores onde o preço é zero apenas 0,02%. É possível ver também que mais de 50% dos valores se encontram na faixa de US$70 e US$110. Foi utilizada escala logarítmica no eixo x para que a visualização ficasse mais clara. O gráfico boxenplot para o preço está presente na Figura 2.

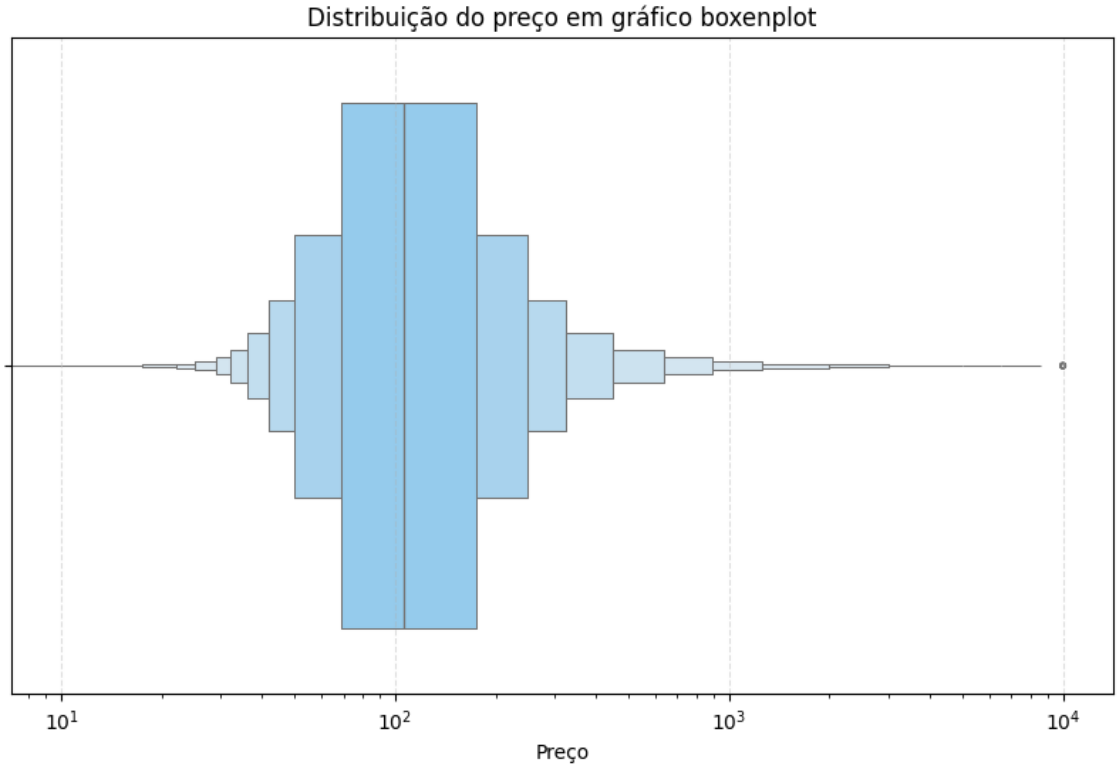

Figura 2: Gráfico boxenplot mostrando a distribuição do preço

O segundo gráfico boxenplot plotado foi o de mínimo de noites, mostrado na Figura 3. Nele é possível identificar que 50% dos valores se encontram entre 0 e 4 noites, e os valores acima de 30 noites, por representarem apenas 1,53 das listagens totais, podem influenciar na qualidade da análise.

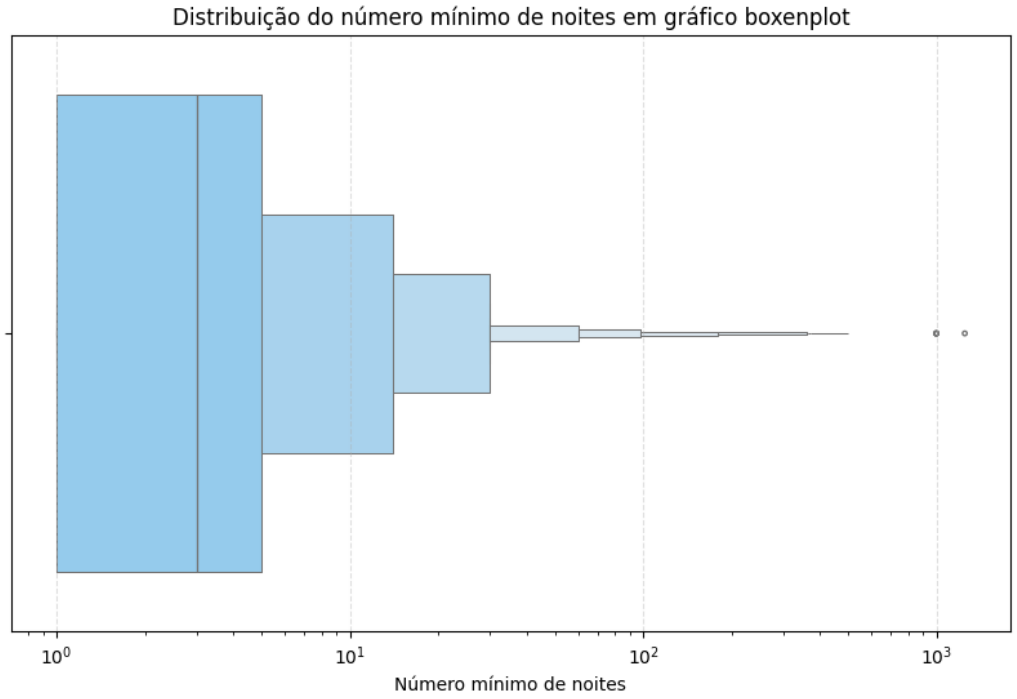

Figura 3: Gráfico boxenplot mostrando a distribuição do mínimo de noites

Após identificados os outliers, foi feita uma limpeza nos dados, gerando uma nova tabela que contém as listagens referentes a preços diferentes de zero e menores que US$600, além dos números mínimos de noites menores que 30 dias. Para verificar se a limpeza de dados foi efetiva, novos histogramas foram plotados para as variáveis em questão, como mostrado na Figura 4. Como se pode ver, as distribuições de preço e mínimo de noites se mostram muito mais claras.

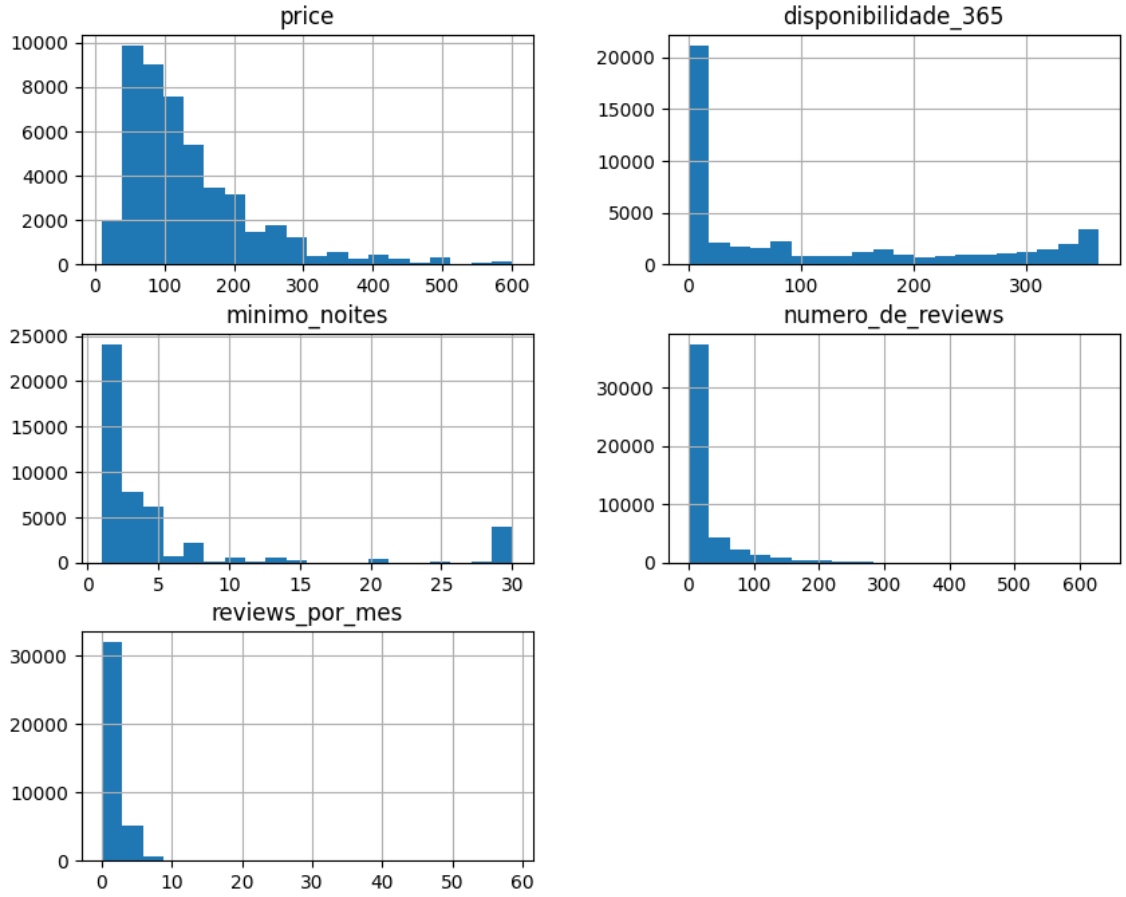

Figura 4: Histogramas das distribuições das variáveis quantitativas após limpeza de dados

Após a limpeza dos dados, a análise dos dados pode ser iniciada. O primeiro atributo a ser analisado é o tipo de quarto. Na Figura 5, que mostra qual o percentual de listagens referente a cada tipo de quarto, é possível ver que a categoria *Entire home/apt* é a que possui mais anúncios, representando mais de 50% dos mesmos. Em seguida, segue a categoria *Private room*, com aproximadamente 48%, e por fim a categoria *Shared room*, com aproximadamente 2%.

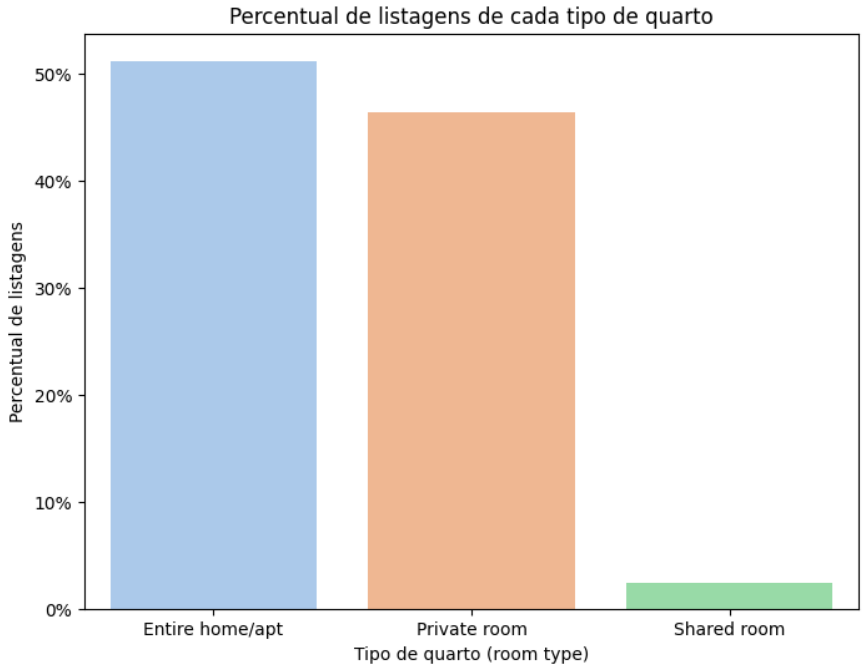

Figura 5: Gráfico em barras do percentual de listagens de cada tipo de quarto

O próximo gráfico, presente na Figura 5, mostra o preço médio de cada tipo de quarto. O maior preço médio é da categoria *Entire home/apt*, com valor médio de aproximadamente US$185, seguido pelo *Private room* com valor aproximado de US$80 e por fim o *Shared room* com valor aproximado de US$65. Com isso, vemos que há uma grande diferença entre o valor de um apartamento/casa inteira e das duas outras categorias.

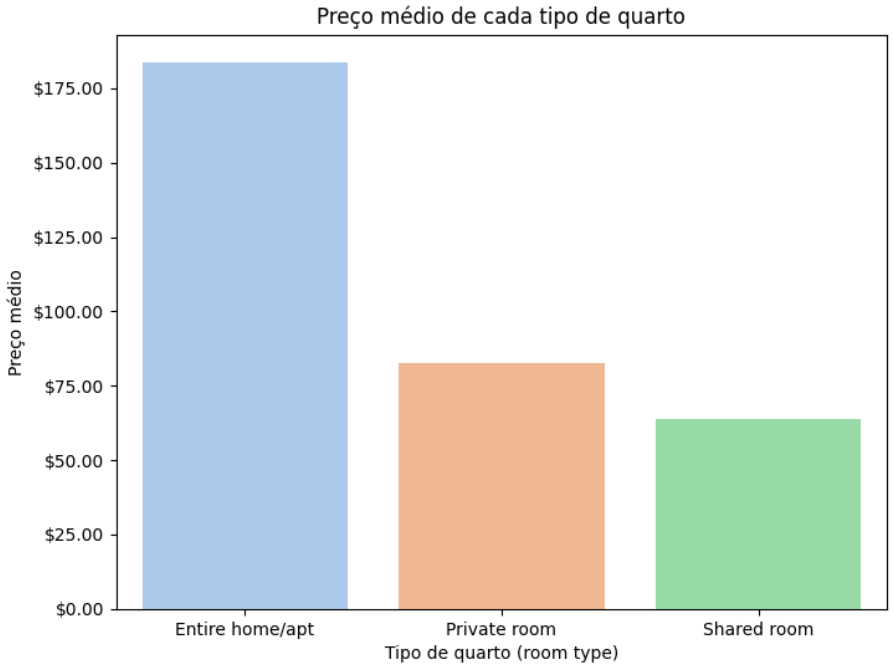

Figura 5: Gráfico em barras mostrando o preço médio de cada tipo de quarto

O próximo atributo a ser analisado foi o de bairros, onde há no total 221 possíveis opções. Por conta da grande quantidade de bairros, optou-se por utilizar apenas os bairros que representam 80% das listagens, ou seja, chegando ao valor de 36 bairros. A Figura 6 mostra a quantidade percentual de anúncios por bairro. Nela, é possível identificar que os três bairros com maior número de anúncios são Williamsburg, Bedford-Stuyvesant e Harlem, contendo aproximadamente 21% de todas as listagens.

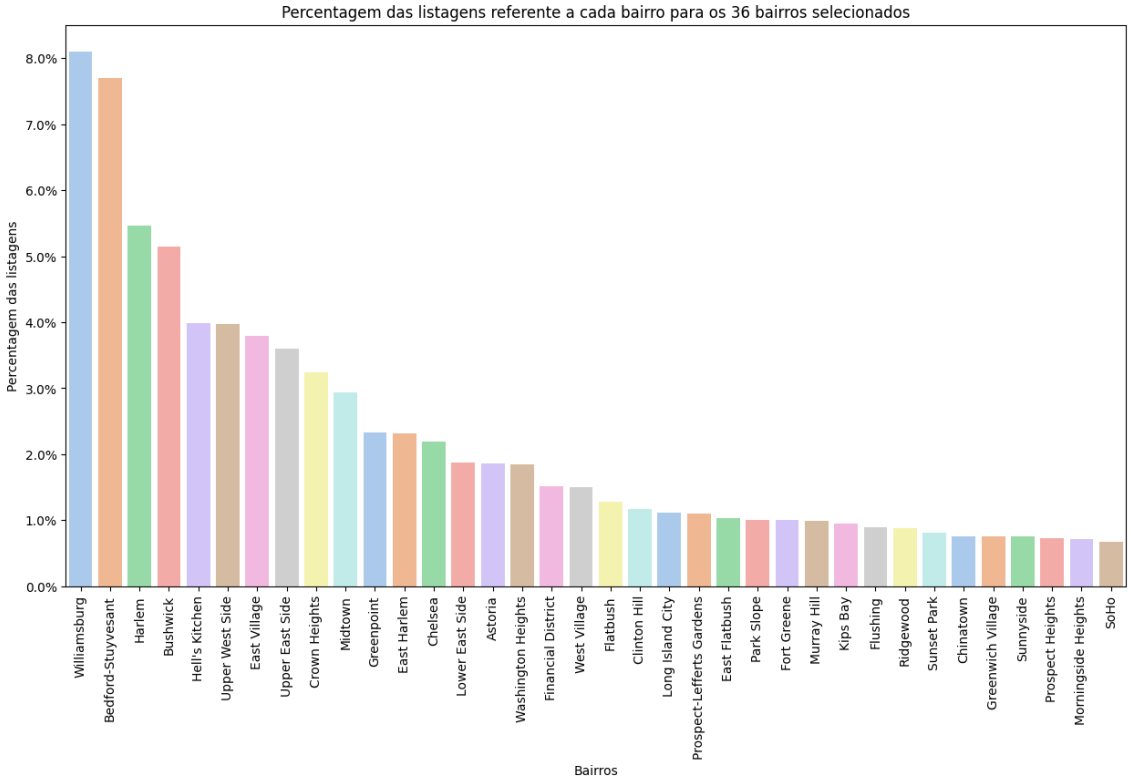

Figura 6: Gráfico de barras mostrando percentagem das listagens referentes a cada bairro selecionado

O preço médio de cada bairro também foi analisado para os 36 bairros selecionados, como mostrado na Figura 7. Pode-se verificar que os bairros Tribeca, Neponsit e NoHo possuem os maiores valores médios, chegando a mais de US$250.

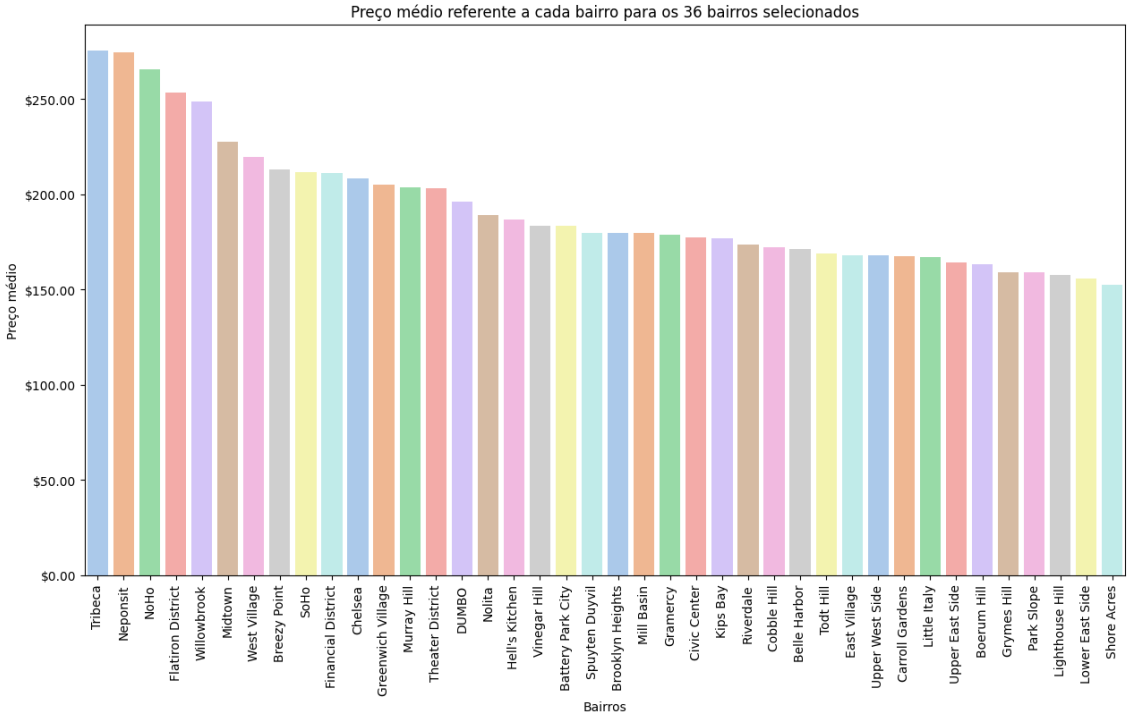

Figura 7: Gráfico de barras mostrando o preço médio referente a cada bairro selecionado

Os grupos de bairro foram analisandos também para se identificar o preço médio em cada grupo de bairro. Na Figura 8, é possível ver que Manhattan possui o maior valor médio de mais de US$160, sendo quase o dobro que o do bairro Bronx. Os bairros que seguem e seus respectivos valores aproximados são Brooklyn a US$110, Queens a US$95, Staten Island a US$90 e por fim Bronx a US$80.

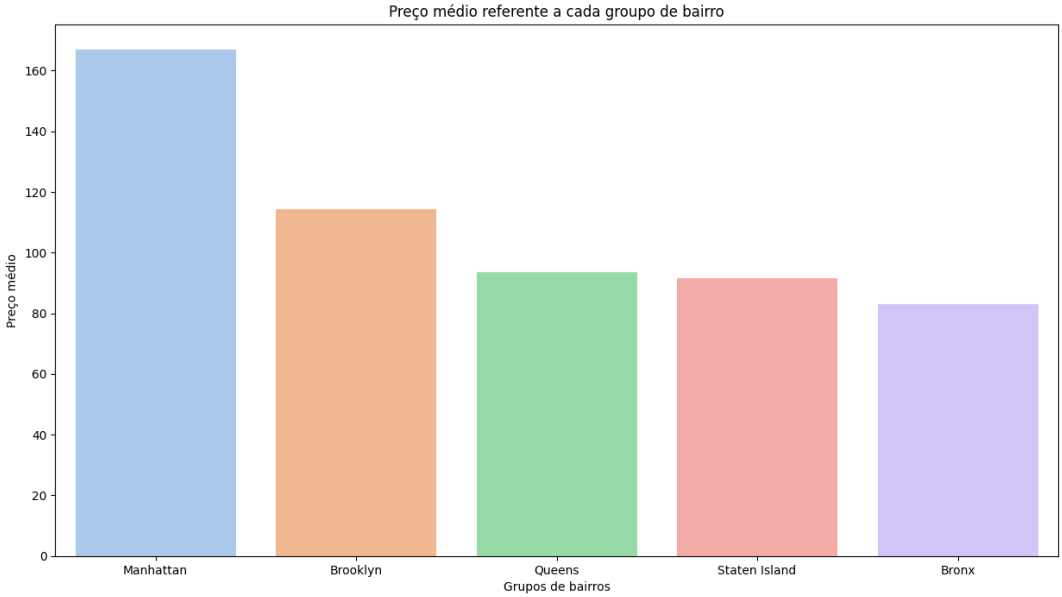

Figura 8: Gráfico de barras mostrando o preço médio referente a cada grupo de bairro

Para se identificar a relação entre algumas variáveis, o índice de correlação foi utilizado. Foram calculadas as correlações para as seguintes variáveis: *price*, *minimo_noites, disponibilidade_365, numero_de_reviews, reviews_por_mes, host_id.* Em seguida, um mapa de calor contendo os valores das correlações entre as variáveis foi plotado, como mostrado na Figura 8. Pode-se ver que todas elas possuem correlação baixa entre si. As únicas variáveis que possui um nível significativo é a de *numero_de_reviews* e *reviews_por_mes,* o que faz sentido, uma vez que uma depende da outra. Pode-se ver também que as correlações entre o preço e as outras variáveis é muito baixa, mostrando que elas não interferem na definição do preço.

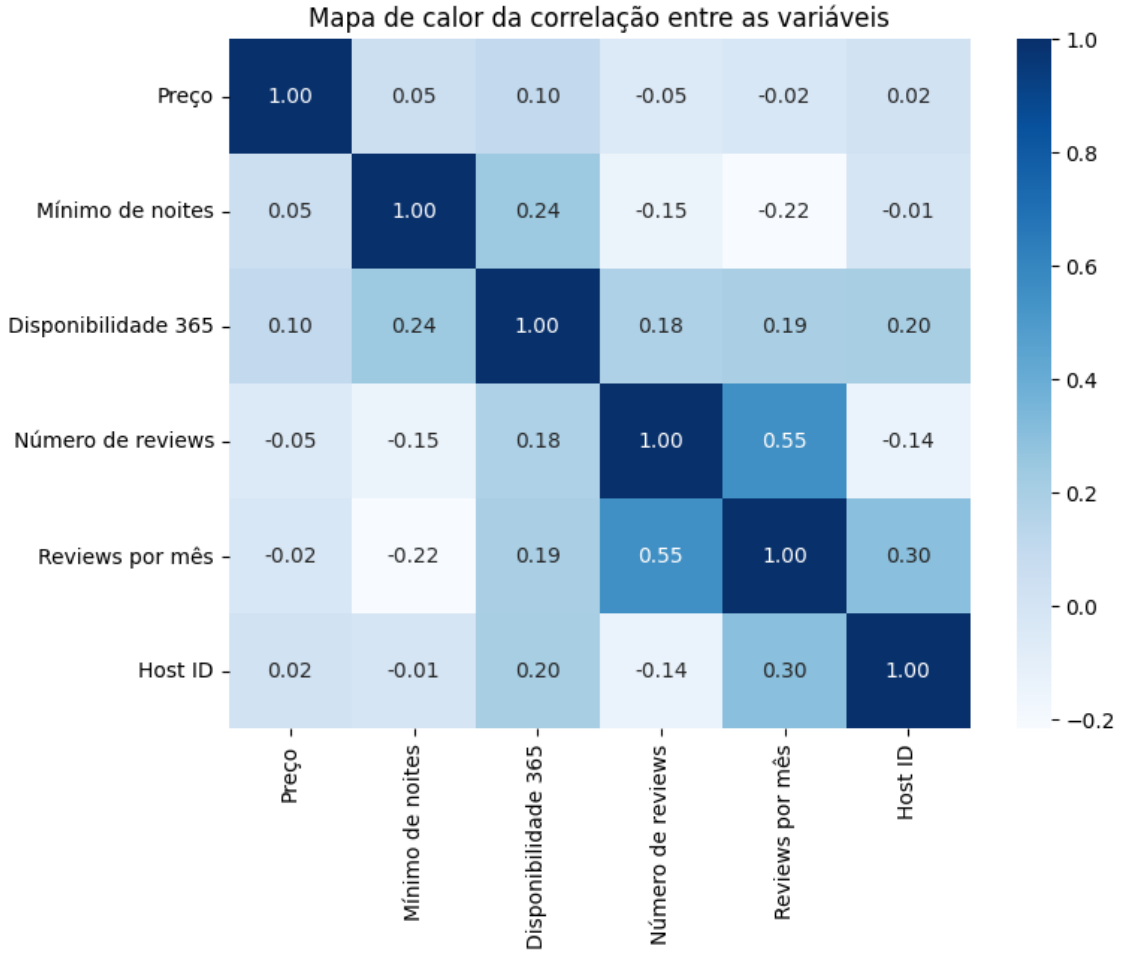

Figura 9: Mapa de calor das correlações entre as variáveis selecionadas

A próxima análise foi da disposição espacial dos anúncios por tipo de quarto e por grupo de bairro. A Figura 10 mostra a distribuição dos bairros de acordo com a latitude e longitude, onde é possível ver que a menor parte dos anúncios de se encontra no Bronx e em Staten Island, e a maior parte dos anúncios nos outros três bairros.

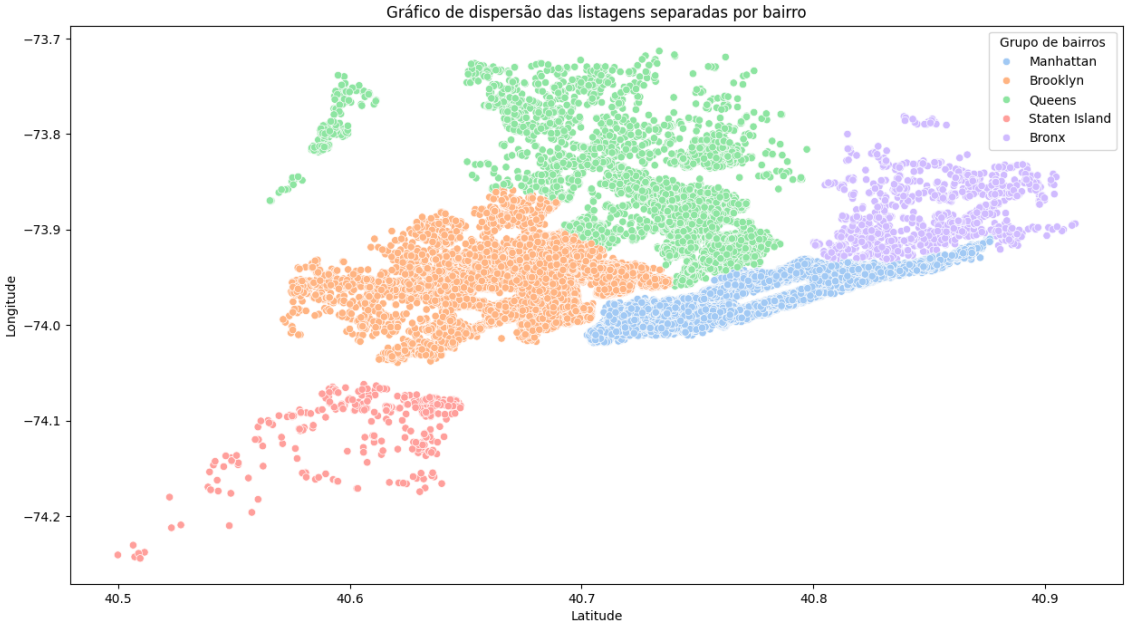

Figura 10: Gráfico de dispersão dos anúncios divididos por bairro

Na Figura 11, é possível ver que há uma grande quantidade de anúncios de quarto privado e de apartamento/casa inteira espalhados pela cidade. Além disso, mostra também que há poucos anúncios de quarto compartilhado no dataset.

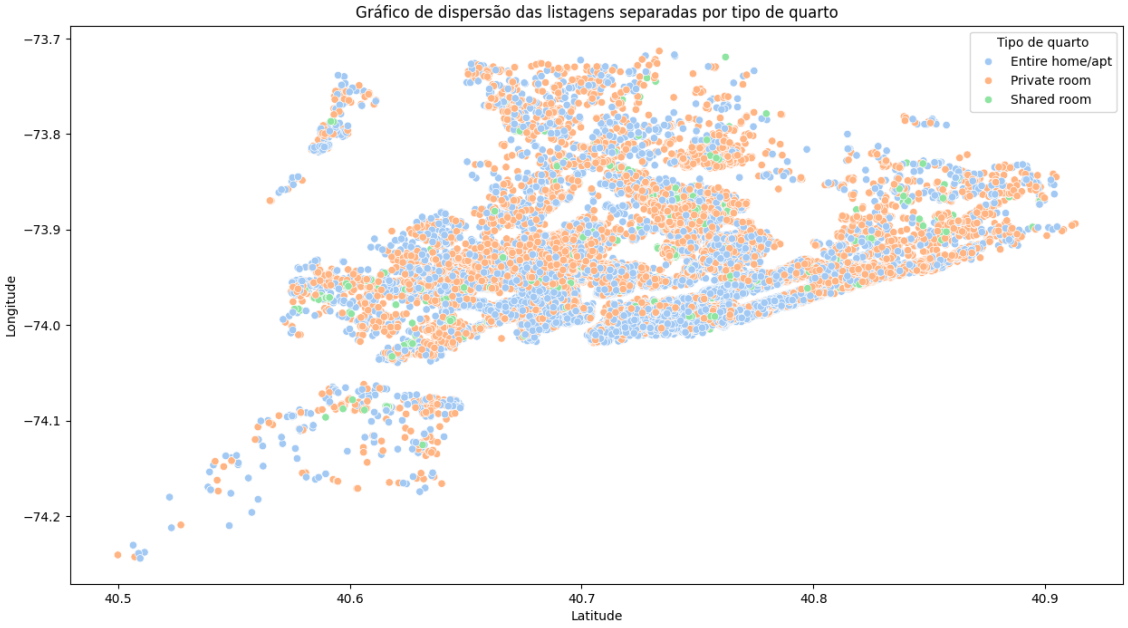

Figura 11: Gráfico de dispersão dos anúncios divididos por tipo de quarto

Finalmente, a última análise feita foi com relação aos nomes dos anúncios. Por se tratar de textos sem formatação definida, a melhor forma de se identificar padrões é utilizando word clouds. Essa ferramenta cria representações gráficas a partir de palavras e a frequência de ocorrência delas

em um texto. Quanto maior a palavra na imagem, mais vezes ela aparece no texto. Com essa ferramenta, é possível identificar palavras que mais aparecem nos anúncios. A análise foi feita para anúncios de valores altos, acima de US$600, e foi gerado o word cloud da Figura 12. Como é possível ver, algumas palavras não representam padrões para anúncios de maior valor, uma vez que são palavras comuns de anúncios de imóveis. Algumas palavras se destacam na identificação de anúncios de alto preço, como: luxury, large, spacious, penthouse. Essas palavras são relevantes e se mostram comuns em anúncios de maior valor.

Figura 12: Word cloud da frequência de ocorrência de palavras no nome dos anúncios

Respostas às perguntas

Após a análise, as três perguntas referentes a ela podem ser respondidas.

Pergunta 1: Supondo que uma pessoa esteja pensando em investir em um apartamento para alugar na plataforma, onde seria mais indicada a compra?

O local mais indicado para a compra de um apartamento é aquele onde há maior retorno, ou seja, nos bairros onde o preço médio da diária é mais alto. Além disso, deve-se levar em conta a quantidade de anúncios no bairro, uma vez que locais onde há muito anúncios podem ter uma menor demanda. O grupo de bairros selecionado foi Manhattan, que possui preço médio de aluguel de mais de 60 dólares. Além disso, os bairros com maior valor de aluguel são Tribeca,

Neponsit e NoHo, e eles não estão presentes entre os bairros com mais anúncios. Sendo assim, esses três bairros seriam a melhor indicação para a compra do apartamento.

Pergunta 2: O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?

Como visto no mapa de calor de correlações, as variáveis número mínimo de noites e disponibilidade ao longo do ano não são fatores relevantes em relação ao preço dos aluguéis. Elas possuem correlação de 0.05 e 0.10, respectivamente, o que mostra que há uma correlação muito fraca entre elas e o preço.

Pergunta 3: Existe algum padrão no texto do nome do local para lugares de mais alto valor?

A análise de padrão foi feita a partir da utilização de um word cloud, mostrado na Figura 12. As palavras identificadas como um padrão de locais de maior valor foram as seguintes: luxury, large, spacious, penthouse

Previsão dos preços

A previsão dos preços pode ser vista como um problema de regressão, uma vez que se busca estimar um valor numérico, e não uma categoria. Assim, foram escolhidos dois algoritmos de regressão: regressão linear múltipla e random forests.

Os dois modelos foram feitos a partir de três variáveis categóricas (bairro, bairro_group e room_type) e uma variável quantitativa (price). Elas foram escolhidas pois, como mostrado na análise de dados exploratória, possuem interferência direta no preço dos imóveis anunciados. Para poder utilizar as variáveis categóricas nos modelos a serem construídos, foi necessária a criação de variáveis *dummy*.

As medidas de performance escolhidas foram erro médio quadrado (RMSE) e coeficiente de determinação dos datasets de teste e de treino (R^2). O RMSE foi escolhido porque fornece uma estimação de quão bem o modelo é capaz de prever o valor em questão, sendo usado para avaliar a acurácia. Já o

coeficiente de determinação foi escolhido porque ele determina a proporção da variância na variável dependente que pode ser explicado pela variável dependente, ou seja, o quão bem os dados se encaixam no modelo de regressão.

Regressão linear múltipla

Para a regressão linear, os dados foram separados em teste e treino, com proporções de 20% e 80% respectivamente, atribuindo ao x as variáveis categóricas e ao y o preço. Com isso, o modelo foi treinado e teve score de 0.43 quando aplicado ao dataset de teste.

Para avaliar a performance do modelo, foram calculadas as medidas mencionadas acima. O erro médio quadrado foi de 68,16. O coeficiente R^2 de teste e treino foram de 0,43 e 0,41 respectivamente.

Random forests

Para o random forests, os dados também foram separados em teste e treino, com proporções de 20% e 80% respectivamente, atribuindo ao x as variáveis categóricas e ao y o preço. Com isso, o modelo foi treinado e teve score de 0.42 quando aplicado ao dataset de teste.

Para avaliar a performance do modelo, foram calculadas as medidas mencionadas acima. O erro médio quadrado foi de 68,46. O coeficiente R^2 de teste e treino foram de 0,42 e 0,41 respectivamente.

Escolha do melhor modelo

A partir dos resultados, pode-se escolher o modelo que melhor se aproxima dos dados. Como visto, os valores de erro médio quadrado e dos coeficientes R^2 são similares para os dois modelos, o que nos mostra que eles são equivalentes nesse tipo de classificação.

A escolha do modelo também pode ser feita de acordo com suas características. As vantagens de usar a regressão linear são a melhor interpretabilidade e eficiência computacional, sendo adequada para conjuntos de dados menores e menos complexos. Já as desvantagens são suas possíveis

suposições rígidas e sensibilidade a outliers, limitando sua aplicabilidade. As vantagens do algoritmo de random forest são sua alta precisão e robustez contra outliers, lidando bem com relações não lineares e complexas nos dados. Suas desvantagens são que ele é menos interpretável e pode exigir ajuste de hiperparâmetros. Assim, para o caso desse dataset, os dois modelos se mostram efetivos.

Previsão do valor do desafio

Para se fazer a sugestão de preço, foi feita a previsão dos valores com os modelos criados. As informações a serem colocadas para previsão estão presentes no trecho de código abaixo.

```
{'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}
```

As variáveis utilizadas foram o bairro, o grupo de bairro e o tipo de quarto. Aplicando esses valores aos algoritmos, obtiveram-se os valores US$252,94 e US$222,76 para os modelos de regressão linear múltipla e random forest respectivamente. Como pode-se ver, o valor que mais se aproxima do valor real é o do algoritmo de random forest.