PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Josmar Baruffaldi Cristello

Análise e Previsão de Preços Listados na Plataforma Airbnb para Imóveis em Toronto, Ontário, Canada

Belo Horizonte
2021

Josmar Baruffaldi Cristello

Análise e Previsão de Preços Listados na Plataforma Airbnb para Imóveis em Toronto, Ontário, Canada

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte
2021

Agradecimentos

Agradeço primeiramente a meus pais, Josmar e Marina que desde cedo me incentivaram, proporcionando aprendizados extremamente valiosos para minha trajetória, além de sempre fornecer suporte em momentos de dificuldade e necessidade. Sem a ajuda deles, jamais teria tornado esse sonho realidade.

A minha mãe, Marina, que faleceu durante a elaboração desse trabalho. Sua imagem ficará sempre comigo, e nunca esquecerei do que você significa para mim. Obrigado por tudo.

Aos amigos do período escolar, Élvio Junior e Pedro Abreu, que apesar de terem seguido trilhas diferentes, até hoje participam de certa forma da caminhada da minha vida. Agradeço também a todos amigos que fiz na PUC/MG e que me auxiliaram em cada degrau da pós-graduação, fazendo as madrugadas de estudo serem menos solitárias, e simplificando matérias que muitas vezes pareciam incompreensíveis. Em especial, ao Rafael Mendes.

A todos meus colegas de trabalho da Teadit, que ofereceram ajuda sempre que necessário, e compreenderam minha necessidade de me ausentar em alguns dias, para poder me dedicar a esse projeto. A todos que auxiliaram direta ou indiretamente fizeram parte da minha formação: Muito obrigado, pessoal!

- Josmar Cristello

SUMÁRIO

TABELA DE FIGURAS

1. Introdução

1.1. Contextualização

O Airbnb é uma plataforma que foi criada em 2007 [1]. Seus criadores eram dois amigos que moravam juntos, precisavam de dinheiro, e se aproveitaram de uma oportunidade. Uma grande conferência de design estava acontecendo em São Francisco, onde moravam, e todos os hotéis locais estavam ocupados. Eles decidiram, então, criar a plataforma (que na época se chamava AirBedandBreakfast).

Desde então, a plataforma cresceu significativamente e hoje em dia tem mais de 7 milhões de propriedades listadas, e está disponível em 100.000 cidades em 220 países [2].

Essa plataforma cresceu tanto resolvendo um problema importante. O mesmo problema que levou os seus criadores a cria-la originalmente: Unir pessoas com espaço disponível querendo uma renda extra, e pessoas que precisam de um lugar para ficar, e não querem depender de hotéis. Seja por motivo de custo, disponibilidade ou praticidade.

Selecionar um imóvel com bom custo-benefício, entretanto, não é trivial. A plataforma oferece inúmeras opções de customização do anúncio, centenas de possíveis amenidades, descrição do anfitrião, diversas categorias de avaliação, o texto em si das avaliações, dentre outras.

Esse excesso de atributos faz com que seja difícil, tanto para um anfitrião, quanto para um hóspede de tomar decisão. O anfitrião vai ter como interesse maximizar o lucro do seu imóvel ou espaço disponível, enquanto o hóspede tipicamente terá como interesse principal maximizar o custo-benefício de sua estadia.

1.2. O Problema Proposto

Diante desse contexto, este estudo tem como objetivo determinar quais são os principais atributos que contribuem para o preço listado de uma propriedade no Airbnb, tão como criar um modelo de predição para esses preços. Para o estudo, a cidade de Toronto, em Ontario, foi selecionada. Toronto é uma cidade turística que recebe 27.5 milhões de visitantes anualmente, e é o principal destino turístico no Canada [3].

Para facilitar o entendimento do problema, este estudo inclui a técnica dos 5W's, que consiste em fazer as seguintes perguntas:

Why? Uma boa modelagem do preço permite a um potencial hóspede maximizar seu custo-benefício, e permite um potencial anfitrião maximizar seu lucro.

Who? Parte dos dados foram coletados pela InsideAirbnb, e são referentes às propriedades listadas na plataforma Airbnb. A outra parte dos dados (geográficos) foram fornecidos pela plataforma Mapquest.

What? São dois objetivos: Predizer o preço listado de um imóvel na plataforma Airbnb a partir destes atributos mais importantes e entender quais são os atributos mais importantes e o impacto destes

Where? Imóveis de toda a cidade de Toronto, localizada na província de Ontário, no Canada.

When? O banco de dados apresenta anúncios cadastrados de anfitriões que estão desde 2008 no Airbnb.

2. Coleta de Dados

2.1. Fonte dos Dados

Foram utilizadas três bases de dados, de duas fontes diferentes. A primeira fonte foi referente aos dados dos imóveis listados na plataforma do Airbnb em Toronto (listing.csv), tão como as avaliações destas propriedades (review.csv). Esses dados foram obtidos do site *InsideAirbnb* [4] no link http://insideairbnb.com/get-the-data.html [5].

O autor deste estudo considerou fazer obtenção dos dados direto no Airbnb (web-scraping), mas não o fez, porque em abril de 2021 isso é contra os termos e serviço do Airbnb [6]. O InsideAirbnb fornece dados extraídos do Airbnb, sob a licensa CCO 1.0 (“public domain dedication”) [7].

A segunda fonte de dados foi a API (Application Programming Interface) de geolocalização reversa do site MapQuest [8]. Essa API foi utilizada para, através de pares de latitude e longitude, obter o município de cada anúncio.

O MapQuest fornece um plano gratuito que permite até 15 000 buscas por mês [9]. Utilizações dos dados para uso acadêmico não são restringidas pelos termos e serviço do site [10]. No total, esse trabalho contemplou 15 833 anúncios do Airbnb, cada um com um par de latitude e longitude, então a API foi acessada em dois meses subsequentes para que todas as buscas pudessem ser executadas no plano gratuito.

2.2. Obtenção dos Dados

Para as primeiras duas bases de dados (Lista de anúncios e de avaliações), a obtenção foi feita através do script download_airbnb.ipynb. Foram utilizadas as bibliotecas gzip, shutil, wget, os e pathlib para fazer download dos dados em formato compactado .gzip e posteriormente descompactar estes arquivos.

A terceira base de dados, utilizada para processar pares de latitude e longitude e obter o município, através do script GEO_process_coordinates.ipynb. Para estes dados, foram utilizadas as bibliotecas pandas e geocoder. Foram feitas solicitações individuais para cada par de latitude e longitude (obtidas da base de dado dos anúncios), com a resposta da API (geocoder) foi armazenada em uma dataframe (pandas).

Como a resposta era demorada, a cada 1000 anúncios, os dados foram armazenados em um arquivo .csv individual, o que gerou um total de 16 arquivos (geolocation_data1000.csv, ... geolocation_data15832.csv). Isso evitava perdas de conexão, diminuía a memória utilizada durante a execução e reduzia e quaisquer outros problemas que podem acontecer com conexões mais lentas. Finalmente, todos os arquivos (biblioteca glob) foram unidos em uma única dataframe, e exportados para um arquivo final, geolocation_data.csv com todas as 15 832 respostas. Esse arquivo foi, posteriormente, processado no arquivo de processamento principal.

2.3. Descrição dos Dados

A primeira base de dados (lista de anúncios) pode ser encontrada no arquivo listings.csv, e possui 36.23 MB. Essa base de dados possui um total de 15.832 linhas, cada uma descrevendo um anúncio do site Airbnb, na região de Toronto, coletado no dia 2021-02-08. Possui também 73 colunas, que descrevem diferentes atributos das propriedades relacionadas a estes anúncios. Os tipos dos dados, e as colunas em si estão descritos na Tabela 1. A coluna que vamos tentar prever nesse estudo está destacada em negrito, price.

A segunda base de dados (lista de avaliações), reviews.csv, possui 108.61 MB. Essa base de dados tem 424.060 linhas, cada uma representando uma avaliação. Além disso, possui 6 colunas, descrevendo propriedades referentes a essas avaliações. Os dados estão descritos na Tabela 2.

Finalmente, a terceira base de dados foi a API do site Mapquest. Esta API retorna diversas colunas como o endereço exato da rua, código de CEP, dentre outros. Todavia, esse estudo estava exclusivamente interessado em obter o município correspondente a latitude e longitude, portanto esse foi o tipo de único dado extraído, no formato de string.

Tabela 1. Nome, Descrição e tipo dos dados no arquivo listings.csv.

Nome da coluna/campo	Descrição	Tipo
listing_url	URL do anúncio (site do Airbnb)	String
scrape_id	Código único da obtenção dos dados	Int
last_scraped	Data que os dados foram obtidos (insideAirbnb)	Date
name	Nome / Título do anúncio, feito pelo anfitrião	String
description	Descrição do anúncio, feito pelo anfitrião	String
neighborhood_overview	Descrição do bairro, feito pelo anfitrião	String
picture_url	URL da imagem utilizada no anúncio	String
host_id	Código que identifica unicamente o anfitrião	int
host_url	URL para a página do anfitrião	String
host_name	Nome do anfitrião	String
host_since	Data de cadastro na plataforma (Anfitrião)	Date
host_location	Localização auto-reportada pelo anfitrião	String
host_about	Descrição do anfitrião	String

host_response_time	Tempo que o anfitrião demora para responder	String
host_response_rate	Percentual de perguntas respondidas	String
host_acceptance_rate	Percentual de reservas que o anfitrião aceita	String
host_is_superhost	Se o anfitrião é considerado “Super Anfitrião” [11]	Bool
host_thumbnail_url	URL da thumbnail da foto do anfitrião	String
host_picture_url	URL da foto do anfitrião	String
host_neighbourhood	Bairro auto-reportado pelo anfitrião	String
host_listings_count	Quantidade de anúncios do anfitrião tem	Int
host_total_listings_count	Quantidade de anúncios do anfitrião tem	Int
host_verifications	Verificações no Airbnb que o anfitrião já fez	Array
host_has_profile_pic	Se o anfitrião tem ou não uma foto	Bool
host_identity_verified	Se a identidade do anfitrião foi verificada	Bool
neighbourhood	Bairro do anúncio	String
neighbourhood_cleansed	O bairro, pelo *Insideairbnb*	String
neighbourhood_group_cleansed	O grupo do bairro pelo *Insideairbnb*	String
latitude	Latitude do anúncio	Float
longitude	Longitude do anúncio	Float
property_type	Tipo de propriedade, descrito pelo anfitrião	String
room_type	Tipo de acomodação, descrito pelo anfitrião	String
accommodates	Número de pessoas acomodadas pelo imóvel	Int
bathrooms	Número de banheiros do imóvel	Float
bathrooms_text	Descrição dos banheiros do imóvel	String
bedrooms	Número de quartos do imóvel	Int
beds	Número de camas do imóvel	Int
amenities	Lista de amenidades oferecidas	Array
price	Preço da reserva de uma diária	String
minimum_nights	Mínimo número de noites para uma reserva	Int
maximum_nights	Máximo número de noites para uma reserva	Int
minimum_minimum_nights	Menor 'minimum_nights', 365 noites a frente	Int
maximum_minimum_nights	Maior 'minimum_nights', 365 noites a frente	Int
minimum_maximum_nights	Menor 'maximum_nights', 365 noites a frente	Int
maximum_maximum_nights	Maior 'maximum_nights', 365 noites a frente	Int
minimum_nights_avg_ntm	'minimum_nights' médio, 365 noites a frente	Int
maximum_nights_avg_ntm	'maximum_nights' médio, 365 noites a frente	Int
calendar_updated	Número atualizações do calendário	Int
has_availability	Se a propriedade está disponível neste momento	Bool
availability_30	Disponibilidade nos próximos 30 dias	Int

availability_60	Disponibilidade nos próximos 60 dias	Int
availability_90	Disponibilidade nos próximos 90 dias	Int
availability_365	Disponibilidade nos próximos 365 dias	Int
calendar_last_scraped	Data de obtenção dos dados do calendário	Date
number_of_reviews	Número de avaliações do anúncio	Int
number_of_reviews_ltm	Avaliações nos últimos 12 meses	Int
number_of_reviews_l30d	Avaliações nos últimos 30 dias	Int
first_review	Data da primeira avaliação	Date
last_review	Data da avaliação mais recente	Date
review_scores_rating	Somatório das avaliações (próximas colunas)	Int
review_scores_accuracy	Avaliação da precisão do anúncio	Int
review_scores_cleanliness	Avaliação da limpeza da propriedade	Int
review_scores_checkin	Avaliação da experiência de check-in	Int
review_scores_communication	Avaliação da comunicação com anfitrião	Int
review_scores_location	Avaliação da localização da propriedade	Int
review_scores_value	Avaliação do custo-benefício	Int
license	Licensa da propriedade (se existente)	String
instant_bookable	Se pode ser reserva sem aprovação	Bool
calculated_host_listings_count	Propriedades que o anfitrião tem na região	Int
calculated_host_listings_count_entire_homes	Propriedades do anfitrião, do tipo propriedade inteira	Int
calculated_host_listings_count_private_rooms	Propriedades do anfitrião, do tipo quarto particular	Int
calculated_host_listings_count_shared_rooms	Propriedades que o anfitrião, do tipo quarto compartilhado	Int
reviews_per_month	Número de avaliações recebidas por mês	Float

Tabela 2. Nome, Descrição e tipo dos dados no arquivo reviews.csv.

Nome da coluna/campo	Descrição	Tipo
listing_id	Código da listagem a qual a avaliação pertence	Int
id	Código único da avaliação	Int
date	Data que os dados foram obtidos	Date
reviewer_id	Código único do hóspede	Int
reviewer_name	Nome do hóspede	String
comments	Texto da avaliação	String

3. Processamento/Tratamento de Dados

Esta seção, tão como o restante do trabalho, utilizou a linguagem de programação Python 3.9.1 64-bytes. Foram utilizadas as bibliotecas pandas e numpy. O banco de dados inicial (listings.csv) tinha uma grande quantidade de colunas não tratadas, valores ausentes, inconsistentes ou que não contribuíam de forma relevante para a análise.

A estratégia nessa seção foi: Reduzir a redundância dos dados (Removendo atributos iguais), remover atributos com grande quantidade de dados inválidos (>70%), e adotar diferentes métodos para preencher os valores ausentes. Esses métodos podem ser: Remoção das linhas faltantes, preenchimento por mediana ou média, imputação manual, e finalmente o tratamento da coluna como dado categórico.

As etapas descritas nessa seção são a descrição da execução do script Notebook.ipynb.

3.1. Importação dos Dados

Inicialmente, os dados arquivo listing.csv foram carregados para um dataframe da biblioteca Pandas [12]. O arquivo possuía um total de possui 15 832 linhas e 73 colunas (ou Atributos). Cada linha nesse arquivo correspondia a um anúncio de um imóvel no site Airbnb para a região de Toronto.

3.2. Conversão dos Dados

O atributo que é a variável de interesse para o modelo (‘price’) é interpretada como string. Isso acontece porque os valores foram armazenados com símbolo indicador da moeda. Exemplos: “$469.00”, “1,396.00”.

Portanto, os valores foram substituídos com remoção do símbolo da moeda ($) e remoção do indicador de milhar (,). Finalmente, os valores foram convertidos para formato float.

Os atributos host_response_rate e host_acceptance_rate, que são valores percentuais também são interpretados como string. Sendo assim, ambas as colunas tiveram o símbolo percentual (%) removido, e foram convertidos para o formato float.

Os atributos com formato booleano, são importadas com valores “t”, ou “f”, ou seja, também interpretadas como strings. Assim, essas colunas tiveram seus valores “t” convertidos para o número 1, e os valores “f” convertidos para o número 0.

Finalmente, as colunas de data (host_since, last_scraped, first_review, last_review), que também são interpretadas como strings, foram convertidas para datas. Após a conversão das colunas de data, um novo atributo foi criado para cada uma delas descrevendo esse tempo em frequência diária.

A coluna host_since, uma das colunas de data, descreve a data em que um anfitrião se cadastrou na plataforma. Já a coluna last_scraped descreve a data em que os dados foram obtidos pela insideAirbnb. Assim, a nova coluna criada, host_since_days, é a subtração das duas mencionadas anteriormente, e passa a descrever o número de dias desde que o anfitrião se cadastrou na plataforma. O resultado esse processo é demonstrado na Tabela 3. O mesmo procedimento é repetido para criar as colunas first_review_days e last_review_days.

Tabela 3. Novo atributo de data, frequência diária.

id	last_scraped	host_since	host_since_days
1419	09-02-21	08-08-08	4568
8077	09-02-21	22-06-09	4250
23691	11-02-21	15-03-10	3986
27423	10-02-21	04-05-10	3935
30931	09-02-21	22-06-09	4250

3.3. Remoção de Colunas Não Utilizadas e Dados Inválidos

Todas as colunas de texto (name, description, host_about, ...) foram removidas. Por motivos de performance, essas colunas serão tratadas separadas em um script separado de NLP (Processamento de Linguagem Natural). Os módulos de NLP foram separados pois são relativamente pesados computacionalmente.

As colunas de metadados, ou seja, colunas que armazenam dados da aquisição dos dados em si por parte do insideAirbnb também foram removidas. Essas são scrape_id, last_scraped e calendar_last_scraped. Finalmente, as colunas que representam urls são removidas listing_url, picture_url, host_url, host_thumbnail_url, host_picture_url. No total, 17 atributos que não contribuiriam para o modelo ou para a análise dos dados foram removidos.

Finalmente, foram removidas um total de 1301 linhas que apresentam atributos de availability_365 = 0 e number_of_reviews = 0 simultaneamente. Estas representam propriedades sem nenhuma data disponível de reserva no próximo ano (365 dias), e que além disso, nunca tiveram nenhuma avaliação de hóspedes. Essas propriedades, muito provavelmente, são utilizadas apenas para especulação imobiliária ou estão pausadas no Airbnb pelo anfitrião.

3.4. Colunas Relacionadas ao Anfitrião

Inicialmente, os imóveis que tiveram valores ausentes em diversos atributos relacionados ao anfitrião foram removidos. Essas propriedades possuíam data de cadastro do anfitrião faltantes (host_since), tão como tempo resposta do anfitrião (host_response_time e host_response_rate). Como são diversos atributos faltantes, o preenchimento com mediana ou média poderia mascarando os dados. Finalmente, outro fator que contribui para a remoção dos dados foram o fato delas representarem apenas 16 propriedades (0.1% dos dados).

A colunas host_response_time tem um total de 5502 valores ausentes. No caso dessa coluna, cada valor faltante indica que o anfitrião ainda não recebeu nenhuma pergunta, e, portanto, não forneceu nenhuma resposta. Por isso, os valores ausentes foram preenchidos por “no response yet”. A distribuição dos valores é, então, disposta na Tabela 4.

Tabela 4. Distribuição da coluna host_response_time.

Valor	Quantidade	% Total
no response yet	6648	43,9%
within a few hours	5550	36,7%
within a day	1692	11,2%
a few days or more	1251	8,3%

As colunas calculated_host_listings_count, host_total_listing_count, e host_listing_count apresentam valores que descrevem as propriedades que o anfitrião possui. A primeira coluna apresenta a quantidade de propriedades da região em questão (Toronto). As outras duas são iguais entre si e descrevem todas as propriedades do anfitrião, tanto em Toronto quanto em outras localizações (e.g. em outros países). Como o objetivo desse estudo é se concentrar em Toronto, as últimas duas colunas foram removidas, e calculated_host_listings_count foi mantida.

Os colunas calculated_host_list_count... _entire_homes, _private_rooms e _shared_rooms foram todas removidas. Essas colunas não são necessárias, visto que elas são calculadas pelo InsideAirbnb e interpretam o tipo de propriedade a partir de processamento do texto, e isso será feito de forma direta por esse estudo mais adiante.

As colunas host_has_profile_pic e has_availability representam, respectivamente, se o anfitrião possui, ou não, uma foto de perfil e se se o imóvel está disponível. Conforme a Figura 1 ambas as colunas apresentam mais de 95% dos dados em uma só categoria (1, ou verdadeiro, para ambos). Portanto, essas não ajudariam nem na análise dos dados nem na construção do modelo, e foram removidas.

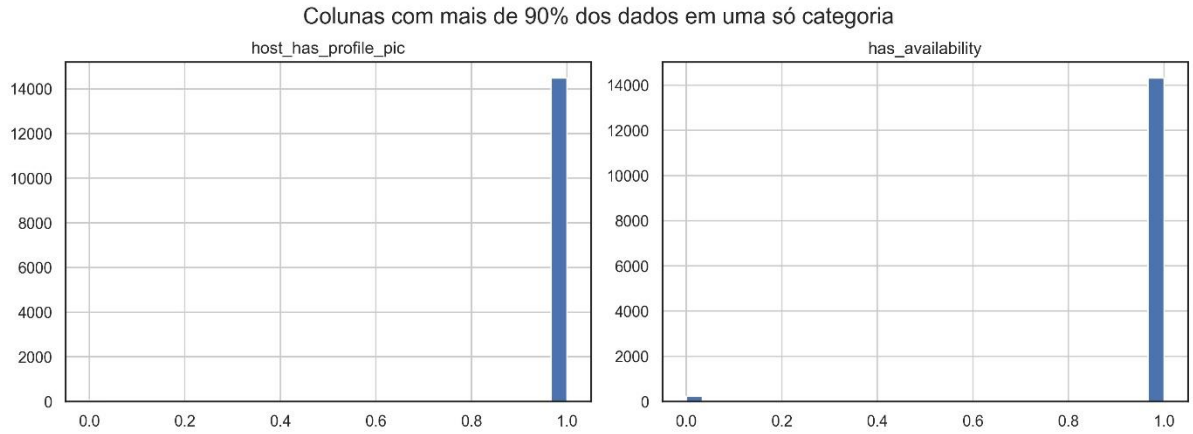


Figura 1. Colunas com 95% dos dados em uma só categoria.

Após essa etapa inicial de processamento, as colunas do anfitrião têm sua distribuição demonstrada nos histogramas da Figura 2.

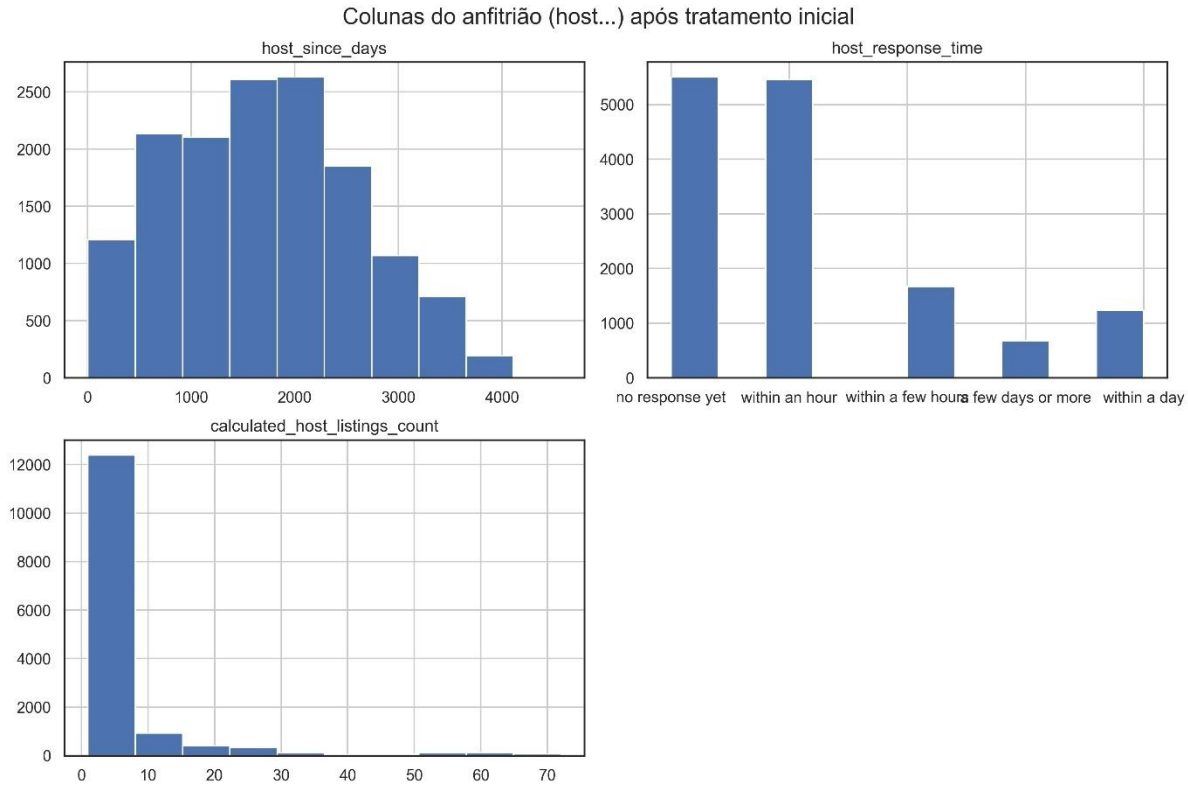


Figura 2. Distribuição das colunas do anfitrião após processamento.

As colunas host_response_rate e host_acceptance_rate apresentam um total de 5502 e 3993 valores ausentes, respectivamente. Em ambos os casos, uma ausência de resposta significa que o anfitrião ainda não respondeu, ou que o hóspede não avaliou o anfitrião. Ambas as colunas são, por padrão, numéricas.

Para preservar a informação e significação dos valores ausentes, elas foram agrupadas em bins e categorizadas, ou seja, não podendo mais ser tratadas como valores numéricos. Os valores ausentes foram convertidos em “no response yet”, e a distribuição das colunas podem ser vistas na Tabela 5.

Tabela 5. Valores das colunas host_response_rate e host_acceptance_rate.

host_response_rate		**host_acceptance_rate**	
Bin / Categoria	**Quantidade**	**Bin / Categoria**	**Quantidade**
no response yet	5502	no response yet	3993
0-60%	897	0-74%	2236
60-91%	1103	74-92%	2399
90-99%	760	92-99%	2641
100%	6269	100%	3262

As colunas com altíssimo número de valores ausentes (>70%) são removidas. Essas são: neighbourhood_group_cleansed, bathrooms e calendar_updated, todas com 100% dos valores ausentes, e provavelmente representam colunas importadas incorretamente pelo insideAirbnb. A coluna license, que representa as licensas do imóvel, também é removida, visto que apresenta 74% dos valores nulos.

3.5. Colunas Relacionadas aos Imóveis

Existem diversas colunas que descrevem o mínimo e máximo número de noites disponíveis para aluguel. Essas colunas já foram descritas na Tabela 1, e são redundantes e correlacionadas. As colunas minimum_nights e maximum_nights foram mantidas, enquanto as minimum_minimum_nights, maximum_minimum_nights, minimum_nights_avg_tm, minimum_maximum_nights, maximum_maximum_nights e maximum_nights_avg_tm foram removidas. Como são muitas colunas, a distribuição delas é demonstrada somente no fim desse estudo, no Anexo 1.

De forma similar, existem diversas colunas que descrevem a disponibilidade do imóvel pelos próximos 30, 60 90 e 365 dias. Como o último caso é o mais abrangente e inclui informação dos anteriores, as colunas availability_30, availability_60, availability_90 foram removidas, enquanto a coluna availability_365 foi mantida. A comparação gráfica entre essas colunas é demonstrada na Figura 3.

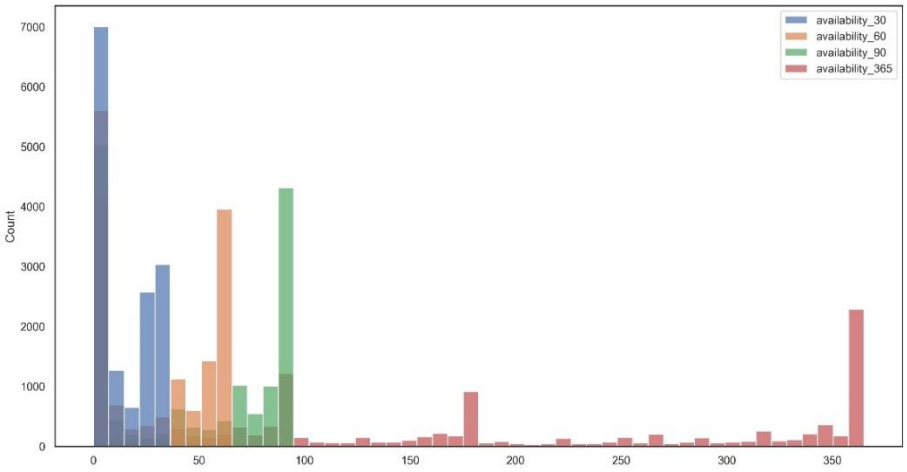


Figura 3. Comparação de todas as colunas que descrevem de disponibilidade.

A coluna property_type descreve o tipo de imóvel, e é preenchida pelo próprio anfitrião. Essa coluna tem 61 valores únicos que segmentam muito os dados. Por isso, ela foi manualmente agrupada, e cada um dos 61 valores foi convertido em quatro possíveis categorias: ‘House’, ‘Apartment’ e ‘Hotel’. Casos que não se enquandravam em nenhuma dessas categorias, eram definidos como ‘Other’.

Exemplos: “Private room in house”, “Private room in townhouse” e “Entire townhouse” foram categorizados como “House”. “Private room in apartment” e “Private room in guest suite” foram categorizados como “Apartment”. Os valores únicos dessa coluna pós-conversão são demonstrados na Tabela 6.

Tabela 6. Valores únicos da coluna property_type após processamento.

property_type	Quantidade	% Total
Apartment	8960	61,7%
House	5330	36,7%
Hotel	161	1,1%
Other	80	0,6%

A coluna room_type não precisou ser processada, mas é demonstrada aqui para exemplificar como nenhuma informação é perdida com a conversão anterior. O atributo room_type categoriza as acomodações em “Entire home/apt”, “Private room”, “Shared room” e “Hotel Room”, e, portanto, quando combinado com a property_type, descreve completamente um imóvel. Esses valores são demonstrados na Tabela 7.

Tabela 7. Valores únicos da coluna room_type.

room_type	Quantidade	% Total
Entire home/apt	9359	64,4%
Private room	4917	33,8%
Shared room	210	1,4%
Hotel room	45	0,3%

A coluna que descreve a quantidade de banheiros tem formato de string. Exemplos: “1 bath”, “2.5 baths”, “shared half-bath”. Aqui a conversão foi feita em duas etapas. Primeiro, todas as instâncias de “half bathroom” foram convertidas no número 0.5. Depois, todo o texto foi removido, e os números foram extraídos, transformando a coluna em um valor numérico que representa o número de banheiros da

propriedade. Nota que “half bathroom” ou 0.5 banheiros são, conforme o Airbnb, definidos como um banheiro que possui apenas privada e pia, sem chuveiro [13].

A coluna com a quantidade de camas (beds) e de quartos (bedrooms) já estão em formato numérico, mas apresentam um total de 168 e 1100 valores ausentes, respectivamente. Como esses valores representam menos de 10% dos dados, e a remoção deles ia representar uma grande perda de informações, eles foram preenchidos com a mediana dos dados. Finalmente, a distribuição dos dados dessas colunas após processamento é demonstrada na Figura 4.

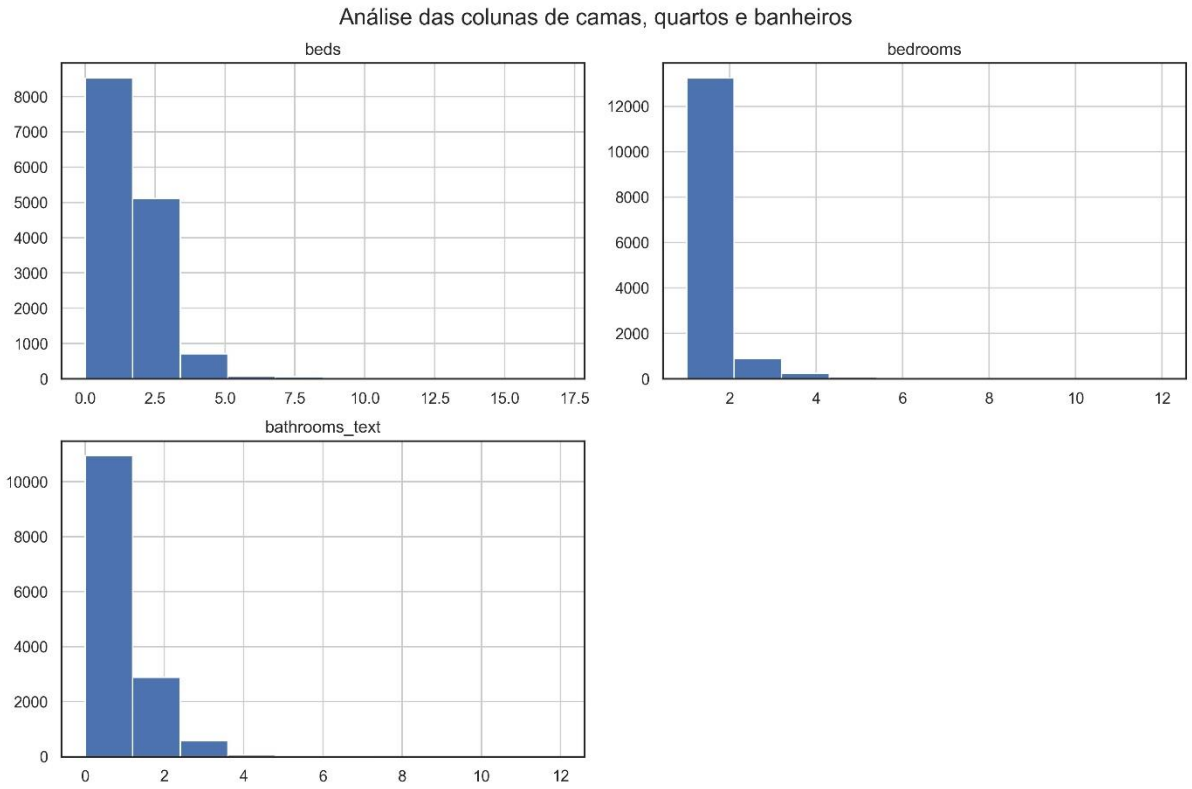


Figura 4. Histograma das colunas beds, bedrooms e bedrooms_text.

A coluna que descreve as amenidades de cada propriedade armazena uma lista com as diferentes amenidades de cada propriedade. Essas, muitas vezes são digitadas pelo anfitrião. No total, as 15.832 propriedades tem um total de 378.502 amenidades.

Para processar essa coluna, as listas foram, primeiro, separadas em cada uma das amenidades. Depois, o texto foi tratado, e os artefatos de obtenção dos dados foram removidos. Elas foram então contabilizadas por frequência e as 43 amenidades mais comuns (Aparecem 2200 vezes ou mais) foram selecionadas. Finalmente, uma coluna booleana é criada para cada uma das amenidades selecionadas.

Por exemplo, uma propriedade que tinha na coluna Amenity [‘Stove’, ‘Gym’] passa a ter duas colunas, am_Stove e am_Gym, ambas com valor 1. Como o imóvel exemplificado não tem nenhuma outra das 41 amenidades, todas essas outras colunas passam a ter valor 0.

3.6. Colunas Relacionadas às Avaliações

As colunas first_review_days e last_review_days são numéricas e representam o número de dias da primeira e da última avaliação de um anfitrião, respectivamente. Ambas as colunas apresentam 2361 valores ausentes. Neste caso, valores ausentes não significam nota 0, e sim que o imóvel/anúncio que ainda não foi avaliado por nenhum hóspede.

Portanto, para preservar o máximo de informações, ambas as colunas foram categorizadas em bins e os valores ausentes são substituídos pelo valor ‘no review yet’. Isso faz com que o modelo não trate mais as colunas de forma numérica, e sim categórica. Essa transformação é demonstrada nas Tabela 8 e Tabela 9.

Tabela 8. Descrição dos valores first_review_days, após processamento

Bin / Categoria	Quantidade	Fração do total
0-6 months	1084	11,6%
6-12 months	861	9,2%
1-2 years	3174	33,9%
2-4 years	4188	44,7%
4+ years	2818	30,1%
no review yet	2363	25,2%

Tabela 9. Descrição dos valores last_review_days, após processamento

Bin / Categoria	Quantidade	Fração do total
0-6 months	5147	42,4%
6-12 months	2055	16,9%
1-2 years	2673	22,0%
2-4 years	1599	13,2%
4+ years	651	5,4%
no review yet	2363	19,5%

As colunas review_score_rating, review_score_accuracy, review_score_cleanliness, review_score_checkin, review_score_communication, review_score_location e review_score_value descrevem diferentes notas da avaliação doo imóvel. A primeira coluna, review_score_rating é simplesmente o somatório das outras 6 colunas, então ela foi removida para evitar problemas de multicolinearidade na construção do modelo.

Todas essas colunas são numéricas, com um total de 2550 valores ausentes. Como são muitas colunas, a distribuição de duas colunas é demonstrada nas Figura 5, e todas as 6 colunas são demonstradas no Anexo 2Anexo 1. Um valor ausente nesse contexto significa um imóvel que ainda não foi avaliado, e não um imóvel com avaliação ruim. Portanto, eles foram categorizados em bins, e todos os valores ausentes vão ser descritos com “no review yet”. A Tabela 10 descreve os valores dessas colunas pós-processamento.

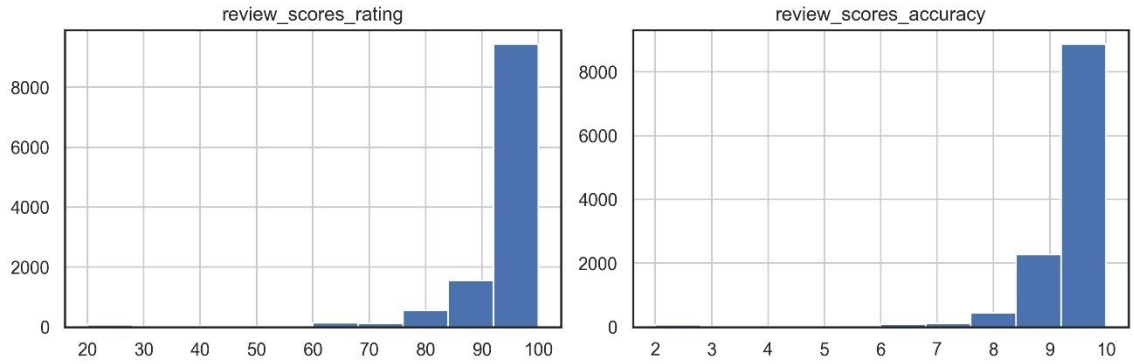


Figura 5. Histograma das colunas review_scores pré-tratamento.

Tabela 10. Frequência das colunas review_scores pós-tratamento.

	Frequência					
Bin / Categoria	**accuracy**	**cleanliness**	**checkin**	**communication**	**location**	**value**
no review yet	2582	2580	2582	2581	2582	2582
0-5	116	160	85	102	51	133
5-8	632	1206	435	448	457	915
8-9	2271	3258	1564	1493	1925	3742
10	8872	7269	9807	9849	9458	7101

As colunas number_of_reviews, number_of_reviews_l30d, number_of_reviews_ltm e reviews_per_month descrevem o número das avaliações do imóvel em diferentes períodos. Elas são fortemente correlacionadas, como demonstrado na Figura 6. Para evitar problemas de colinearidade, as colunas number_of_reviews_ltm e number_of_reviews_l30d foram removidas, e as outras duas foram mantidas.

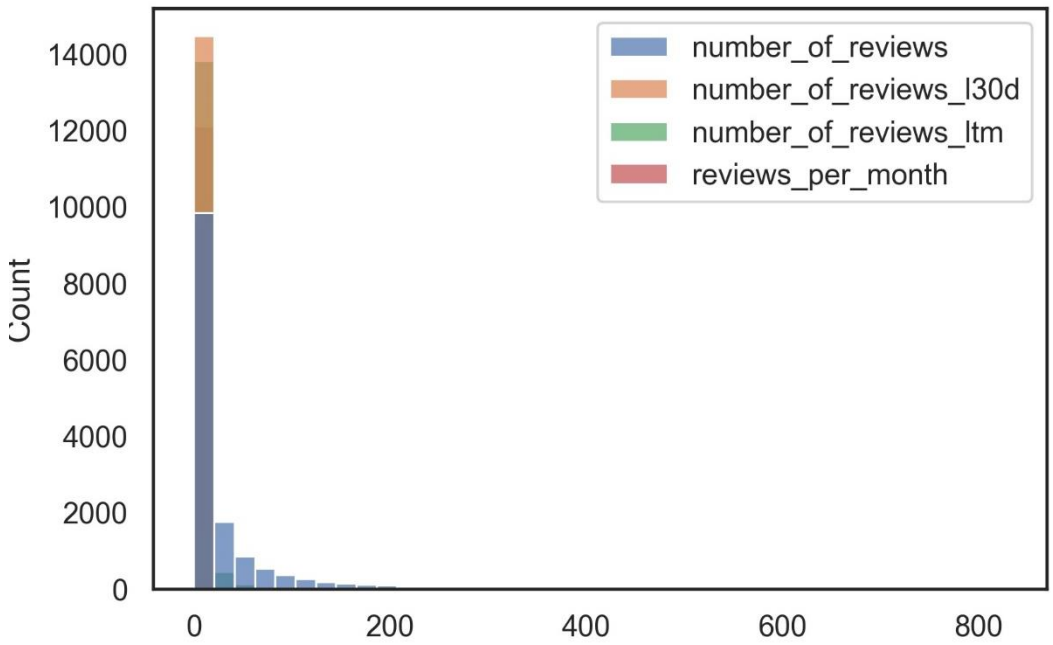


Figura 6. Distribuição das colunas que descrevem o número de avaliações.

Embora o atributo number_of_reviews não apresenta valores ausentes, a reviews_per_month apresenta um total de 2361 valores ausentes. Por isso, todos esses serão preenchidos com o número “0”, para indicar que esses anúncios ainda não foram avaliados por nenhum hóspede.

3.7. Coluna de Interesse: price

A coluna price, ou seja, o preço listado para a diária é extremamente assimétrico, conforme demonstrado na Figura 7. A coluna tem uma mediana de CAD$ 92, mas com um valor máximo de CAD$ 13000! Essa assimetria é causada por relativamente poucos outliers, apenas 279 anúncios (~2% do total) tem um valor de diária maior que CAD$501. Após investigação de alguns casos, muito provavelmente são anúncios que o anfitrião lista com um valor elevado temporariamente para não pausar o anúncio, mas também não alugado no período. Como esses outliers não contribuiriam com a análise ou interpretação, todos esses 279 anúncios foram removidos. A nova distribuição é mostrada na Figura 8.

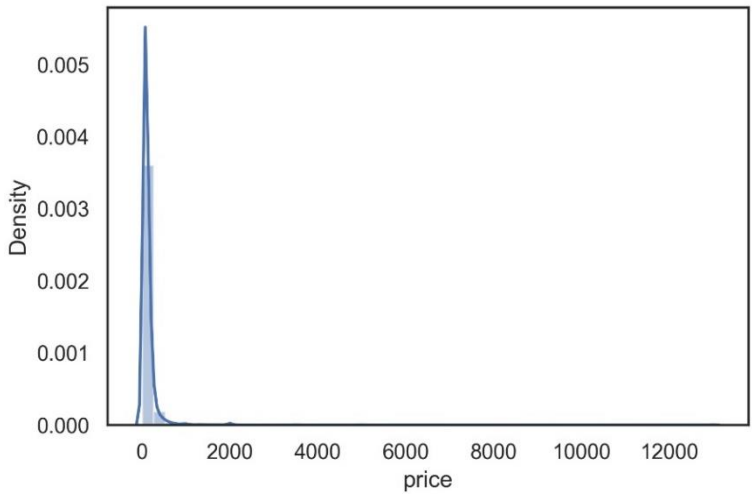


Figura 7. Distribuição de densidade da coluna price, antes do processamento.

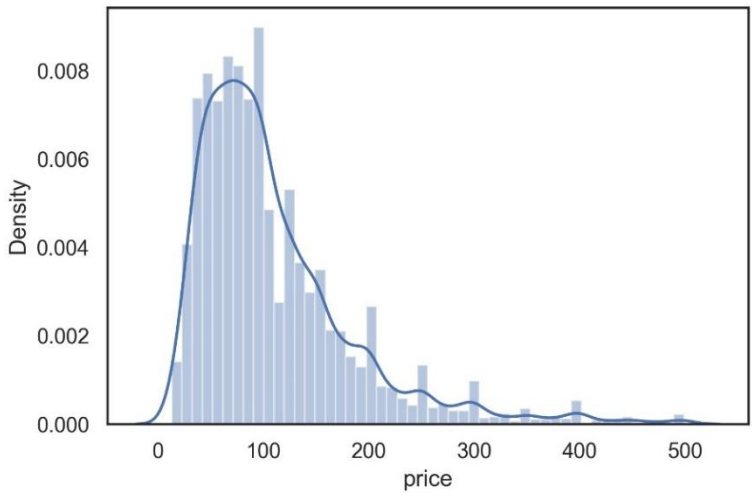


Figura 8. Distribuição de densidade da coluna price, após o processamento.

É particularmente interessante notar uma maior densidade em números inteiros (100, 200, 250, 300, 350, 400), demonstrando o efeito da psicologia na hora da definição do preço. Isso não é surpreendente, e já foi amplamente estudado em artigos de psicologia [14].

3.8. Processamento de Linguagem Natural (NLP): Avaliações e Descrições

Essa parte do processamento é feito pelos scripts NLP_process_reviews.ipynb e NLP_process_descriptions.ipynb, para as avaliações e descrições, respectivamente.

Conforme descrito anteriormente, as avaliações estão armazenadas no arquivo reviews.csv e são feitas pelos hóspedes. Uma mesma propriedade pode ter 0 ou mais avaliações. Já as descrições estão armazenadas no listings.csv e são feitas pelos anfitriões, e, portanto, cada anúncio pode ter no máximo uma descrição.

Essas colunas são apresentam uma grande quantidade de dados. São um total de 424.060 avaliações, e 15.267 descrições. Para conseguir utilizá-las tanto na análise, quanto nos modelos de machine learning, a estratégia selecionada foi análise de sentimento, seguido de cálculo de polaridade.

Em ambos os casos, a tecnologia utilizada foi a biblioteca ntlk com vader_lexicon. Essa tecnologia foi selecionada porque o vader_lexicon é particularmente adepto em lidar com linguagem informal, gírias e siglas (e.g. LOL, OMG, nah, meh, poggers, me_irl), visto que ele é treinado especificamente para redes sociais [15].

Primeiro, a coluna de interesse foi carregada em uma dataframe do pandas, o texto foi, então, processado: Pontuação, artefatos de importação e valores numéricos foram removidos. Depois, o texto foi processado com o vader_lexicon, e uma nova coluna com a polaridade dessa coluna foi criada.

Para as avaliações, foi criada a coluna review_polarity e para as descrições, description_polarity. Essa polaridade contém valores que vão de -1 a +1, onde -1 representaria um comentário completamente negativo, e um positivo representaria um

completamente positivo. A distribuição da polaridade das avaliações e descrições são demonstradas nas Figura 9 e Figura 10.

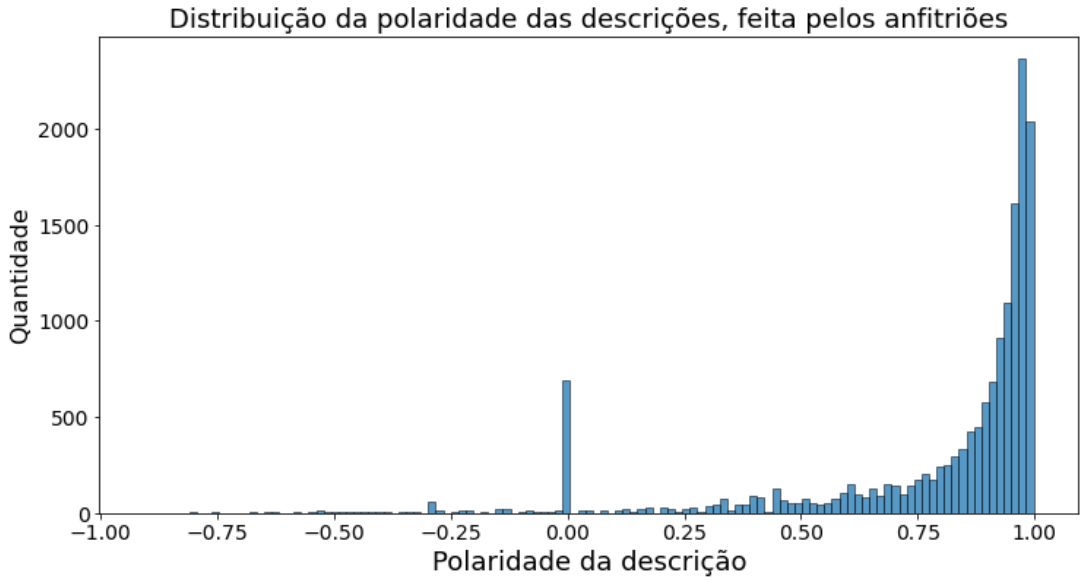


Figura 9. Distribuição da polaridade das descrições

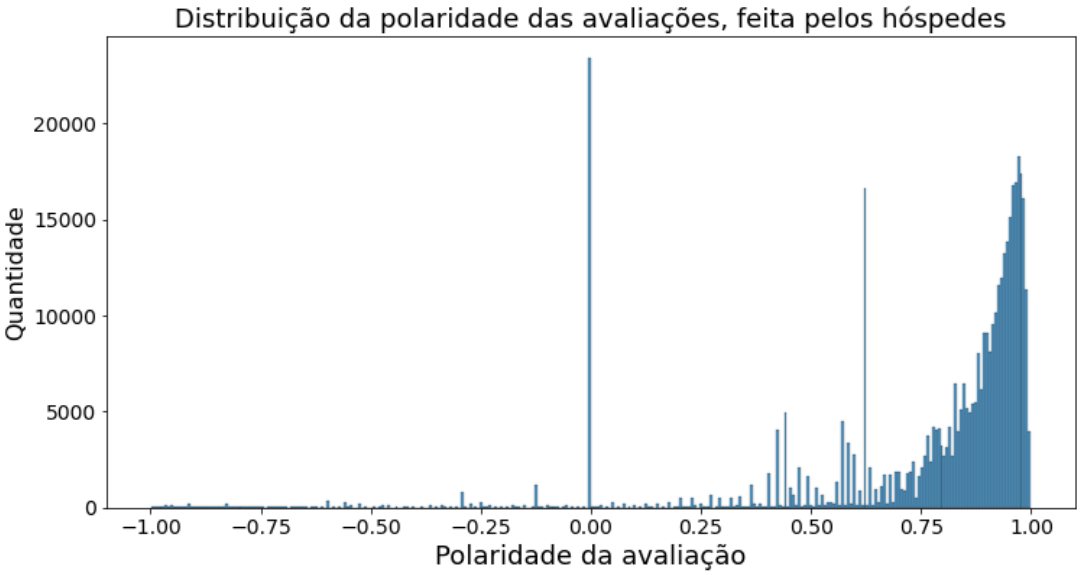


Figura 10. Distribuição da polaridade das avaliações.

Em ambos os casos, é possível ver um grande pico de distribuição no valor 0 (absolutamente neutro). Isso corresponde a avaliações com pouquíssimas palavras, e palavras curtas. Isso é mais comum nas avaliações do que nas descrições. É

também interessante notar, em ambos os casos, uma extrema assimetria: Em geral, tanto as descrições quanto as avaliações são positivas.

Como forma de verificar e exemplificar o que é considerado uma polaridade positiva ou negativa, exemplos de melhor e pior avaliação podem ser encontrados, por extenso e sem alteração, no Anexo 3 e Anexo 4. Similarmente, um exemplo de melhor e pior descrição no Anexo 5 e Anexo 6. Esses exemplos acompanham o código ID, para permitir que o leitor possa rastrear dados relacionados nos bancos de dados. Observação que nem todas as avaliações/descrições puderam ser anexadas, para evitar termos vulgares. Finalmente, O Anexo 7 apresenta uma nuvem de palavras com as palavras mais comuns das avaliações com polaridade positiva (>0.5).

3.9. Processamento de Geolocalização Reversa: Latitude e Longitude

O site insideAirbnb fornece a latitude e a longitude dos anúncios, mas não fornece o município. Ele fornece, também, o bairro dos anúncios (neighbourhood_cleansed), mas essa coluna tem 140 valores únicos, com o valor mais frequente correspondendo a apenas 20 anúncios. A distribuição dessa coluna é demonstrada no Anexo 8, que por ser muito segmentada, acaba não sendo útil para a construção do modelo. Portanto, para que fosse possível utilizar as informações de geolocalização, duas estratégias foram adotadas.

A primeira estratégia foi utilizar geolocalização reversa, ou seja, utilizar a latitude e longitude para buscar o endereço equivalente, e extrair o nome do município correspondente. Para isso, a biblioteca geocoder foi utilizada, e a API do Mapquest forneceu os dados de busca reversa. Para cada par de longitude e latitude, foi buscada um município. Esse, então, foi armazenadado na coluna geo_city e adicionado ao dataframe principal.

A distribuição de anúncios por municípios é demonstrada na Tabela 11. Como os municípios Unionville, Thornhill, Concord e Mississauga tinham, combinados, apenas 6 registros (<0.1% do total), esses registros foram removidos da análise.

Tabela 11. Distribuição do novo atributo, geo_city.

geo_city	Quantidade
Toronto	10129
North York	1758
Scarborough	945
Etobicoke	612
York	463
East York	243
Unionville	2
Thornhill	2
Concord	1
Mississauga	1

A segunda estratégia utilizada para geolocalização foi a criação de clusters através do algoritmo k-means. A criação de clusters é muito útil em mineração de dados, e particularmente em uma situação quando se deseja agrupar diversos pontos em uma matriz. O algoritmo selecionado, k-means, tem a desvantagem de precisar calcular a distância entre todos os objetos (par latitude e longitude) com o centroide do cluster, o que faz com que ele não seja muito eficiente computacionalmente [16]. Isso não foi um problema considerável para a quantidade de dados nesse estudo.

Para isso, a tecnologia utilizada foi a biblioteca sklearn. Foram criados no total 8 clusters, e armazenados na coluna geo_cluster. A distribuição pode ser vista na Tabela 12. É possível ver como a distribuição, embora não perfeitamente homogênea, é mais balanceado do que a coluna geo_city. Nota que a distribuição não ser completamente homogênea é um fato esperado, visto que a distribuição das propriedades em si também não é homogênea.

Tabela 12. Distribuição do novo atributo, geo_cluster

geo_cluster	Quantidade
0	6197
1	620
2	1427
3	721
4	1319
5	2513
6	618
7	735

3.10. Estado Final do Dataframe após Processamento/Tratamento

Depois das etapas anteriores, o dataframe estava limpo, sem valores ausentes, tratado e devidamente categorizado. Inicialmente o dataframe tinha 15.832 linhas e 73 colunas. Ao fim do tratamento, tinha 14.150 linhas 81 colunas. Das colunas originais, 42 foram removidas, e 51 novas foram criadas.

4. Análise e Exploração dos Dados

4.1. Análise e Exploração Inicial dos Dados

Em Toronto, são listados um total de 14.150 propriedades, distribuídos em 6 municípios. Dessas, 8.497 estão imediatamente disponíveis para hóspedes nos próximos 30 dias, e o restante já se encontra reservado.

O número de propriedades por anfitrião será explorado mais a fundo posteriormente, mas é extremamente assimétrico (Assimetria de 4.0). Cada anfitrião tem, na mediana, 2.0 propriedades listadas. Mas alguns chegam a listar 69 propriedades por si só. Por outro lado, 48% dos anfitriões têm apenas 1 propriedade.

Em média, um hóspede pagaria uma diária de CAD$110 para conseguir uma propriedade na região de Toronto na data de coleta desses dados, 02 de fevereiro de 2021. Isso ainda é mais barato que um hotel que, de acordo com budgetyourtrip.com [17], custa uma média CAD$ 224 por dia.

A Figura 11 mostra o cadastro de novos anfitriões na plataforma, e a primeira avaliação de propriedades por mês, no período de 2010 a 2020. É possível ver claramente que a plataforma está em crescimento.

Também é possível ver uma queda em 2020, tanto na tendência de novos anfitriões quanto de primeiras avaliações. Isso faz sentido, considerando que esse ano foi marcado pela pandemia global Covid-19, e os efeitos no setor de turismo e hospitalidade já foram amplamente estudados [18].

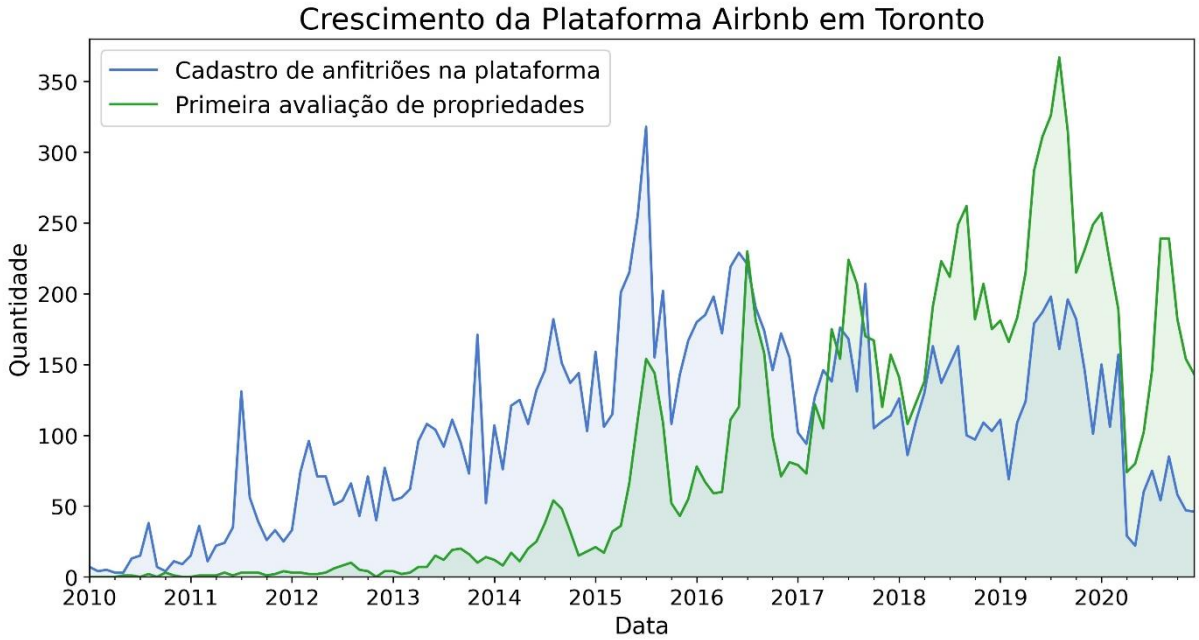


Figura 11. Novos anfitriões e primeiras avaliações, de 2010 a 2020.

O valor da diária, embora distribuído entre CAD$10 e CAD$500, está fortemente concentrada próximo dos valores de CAD$100. Apresenta uma mediana de CAD$90,0 e média de CAD$110,2, conforme Figura 12.

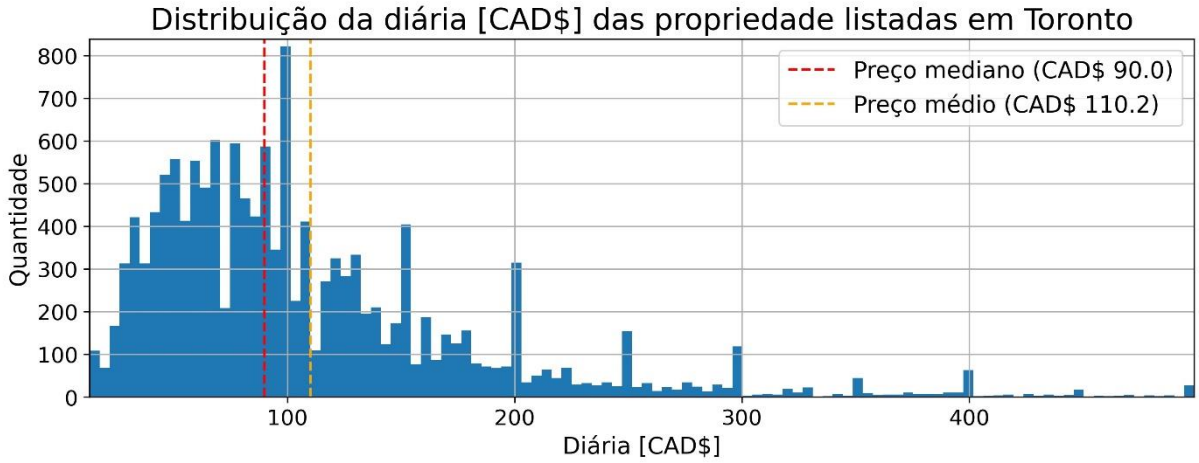


Figura 12. Distribuição de diária [CAD$] das propriedades listadas em Toronto.

4.2. Anfitriões, Quantidade de Propriedades, Atributos Associados

Embora a mediana de propriedades por anfitrião seja de 2 propriedades, alguns possuem muito mais que isso. A Figura 13 mostra os 10 anfitriões com mais propriedades, e quantas propriedades cada um deles tem organizado por ID. O

anfitrião com mais propriedades possui 69, enquanto o décimo maior possui 30 propriedades.

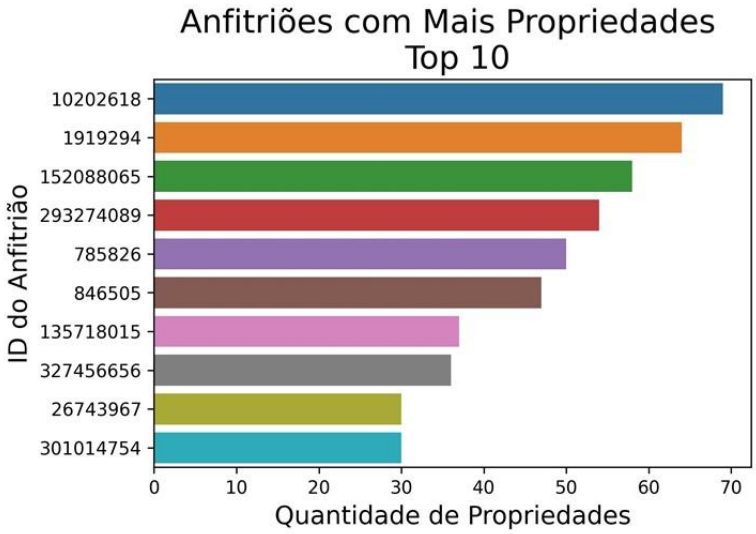


Figura 13. Os 10 Anfitriões com mais propriedades, por quantidade.

A Figura 14 apresenta os valores de diária agrupados pela quantidade de propriedades de cada anfitrião, com erro de 1 desvio padrão. A partir de 10 propriedades é possível ver uma tendência de maiores preços. Isso não quer dizer necessariamente que anfitriões com mais de 10 propriedades simplesmente aumentam o preço, mas pode indicar que eles ou tem propriedades melhores (e, portanto, preços maiores), ou sabem usar melhor o sistema para obter um preço mais vantajoso.

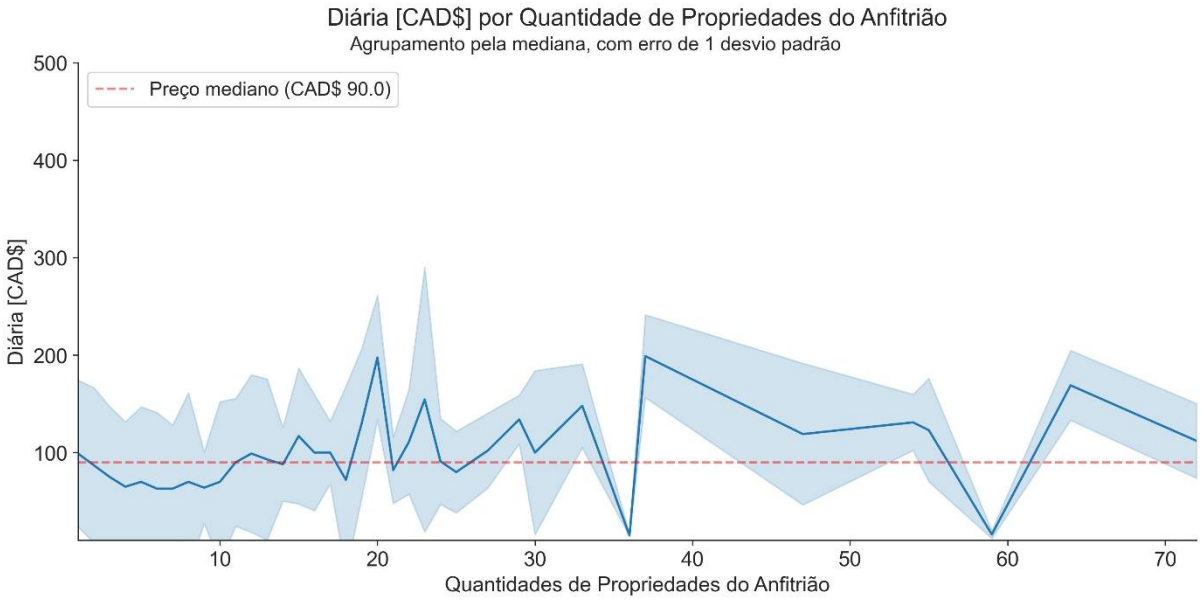


Figura 14. Correlação da diária por quantidade de propriedades (agrupadas).

Além da quantidade de propriedades, os anfitriões têm dois atributos booleanos: Se ele é ou não um “super anfitrião” (host_is_superhost) e se a identidade dele foi ou não verificada (host_identity_verified). Um super anfitrião, é definido pelo Airbnb, como um anfitrião que atingiu diversos requerimentos [11]:

- Pelo menos 100 noites reservadas em, no mínimo, 3 reservas diferentes.
- Uma resposta de pelo menos 90%.
- Um cancelamento de 1% ou menos.
- Uma nota de avaliação maior ou igual que 4.8 nos últimos 365 dias.

A Figura 15 analisa a quantidade de anfitriões que são considerados superhosts e o efeito desse atributo no preço mediano. Conforme esperado, poucos anfitriões têm essa categoria exclusiva (apenas 29%). Por outro lado, e de forma surpreendente, o impacto no preço da diária é muito pequeno. Isso sugere que, mesmo se um hóspede estiver tentando minimizar o preço pago pela diária, é valido procurar anúncios feitos por superhosts, para garantir um maior conforto e melhor experiência pelo mesmo preço (melhor custo benefício).

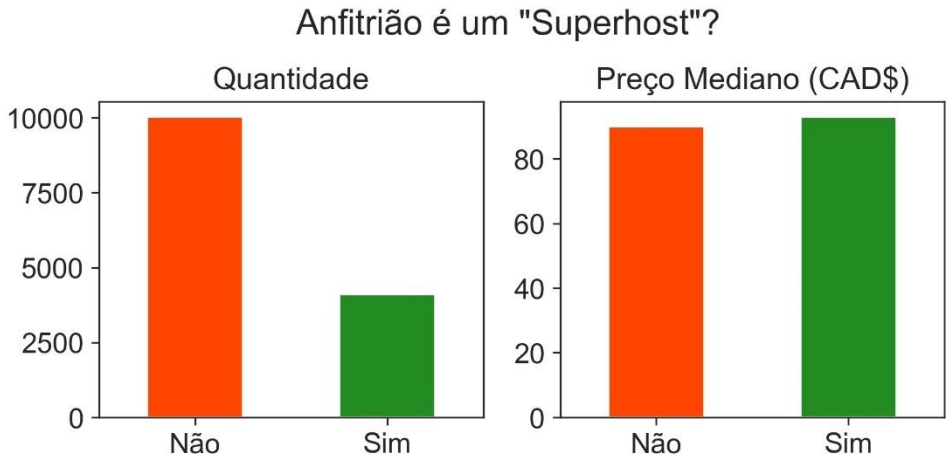


Figura 15. Quantidade de superhosts, e o efeito deles no preço da diária.

A Figura 16 analisa a quantidade de anfitriões que tem sua identidade verificada, e o efeito disso no preço da diária. É possível ver que a maioria dos anfitriões tem sua identidade verificada (83%), e que isso tem um efeito significativo, com o preço mediano para anfitriões não verificados sendo de CAD$80, e para verificados de CAD$93.

Isso não significa, necessariamente, que as propriedades de anfitriões não verificados são piores. Mas pode indicar que eles têm menos experiência, são novos e listam menos detalhes nas propriedades, e por isso acabam listando preços menores também.

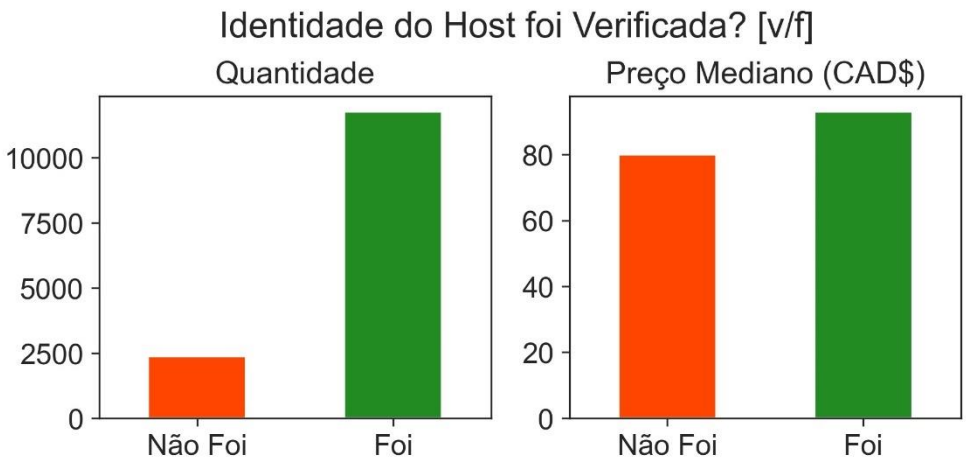


Figura 16. Quantidade de anfitriões verificados, e o efeito deles no preço da diária.

4.3. Geolocalização - Municípios

A cidade de Toronto é dividida em seis municípios principais, esses são demonstrados na Figura 17, conforme disponibilizado pela Wikipedia [19]. Esses

municipios são: Old Toronto, East York, Scarborough, North York, York e Etobicoke. Essa informação concorda com os dados obtidos por geolocalização reversa.

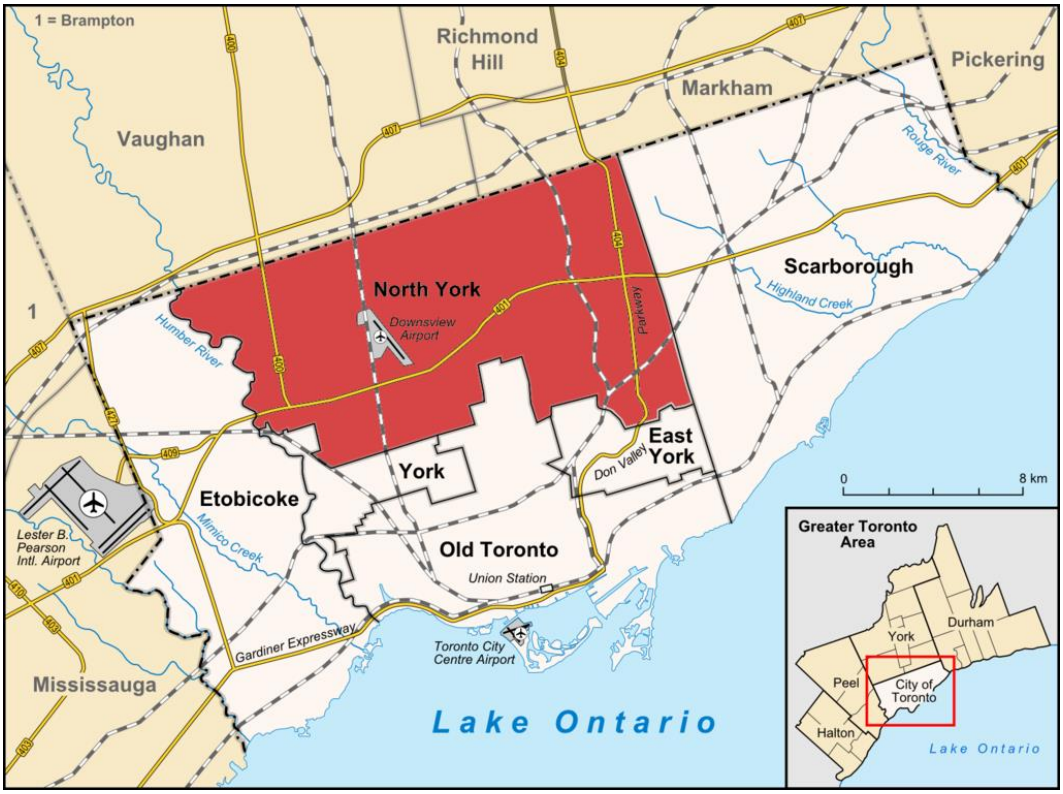


Figura 17. Municípios de Toronto [19].

Para enriquecer a visualização das propriedades sendo analisadas até o momento, a Figura 18 mostra o mapa de Toronto com as propriedades agrupadas por latitude e longitude. Para isso, foi utilizada a biblioteca Folium [20], que disponibiliza os dados fornecidos pela OpenStreetMap [21]. O Anexo 9 também disponibiliza uma visualização alternativa, mostrando um mapa de calor dessas propriedades sob o mapa de Toronto.

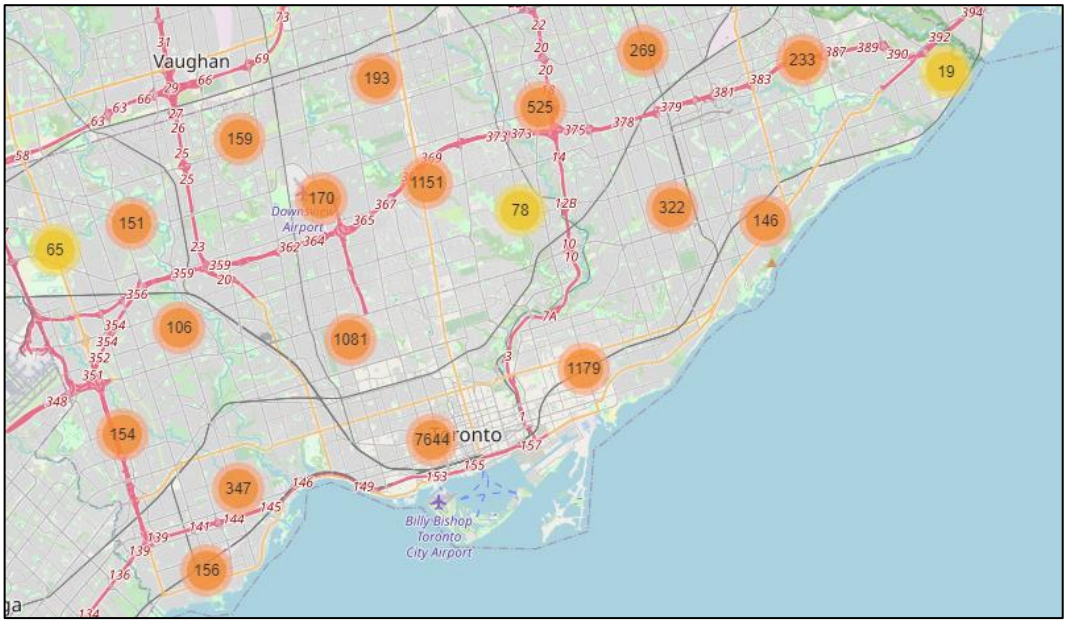


Figura 18. Propriedades listadas no dataframe, agrupadas geograficamente.

A Figura 19 mostra um gráfico de dispersão de latitude e longitude, com categorização dos pontos feito pelo preço da diária da propriedade. Esse gráfico demonstra claramente que a maioria das propriedades está concentrada no município de Old Toronto. Ele mostra, também, que o valor da diária é, em geral, maior na zona sul de Toronto, próximo ao Lake Ontario.

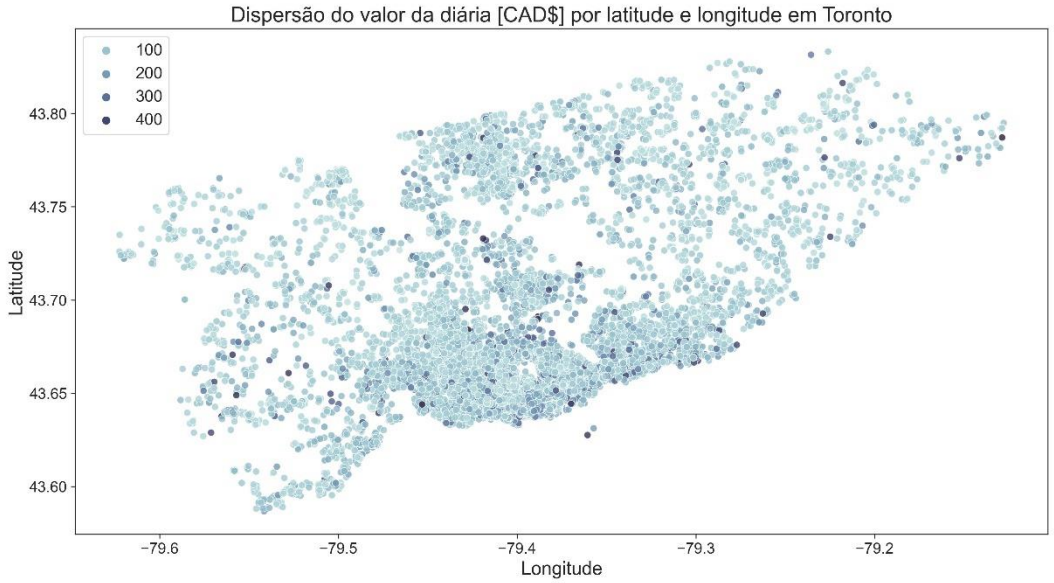


Figura 19. Dispersão de Latitude e Longitude, colorido por faixa de preço.

A Figura 20 mostra um boxplot dos preços de diária por município. Conforme esperado, Old Toronto tem preços significativamente maiores. Além disso, é relevante

notar que Old Toronto também tem 71.5% das propriedades, seguido de North York com 12.4% das propriedades, e os outros 4 municípios combinados somam apenas os 16% restantes.

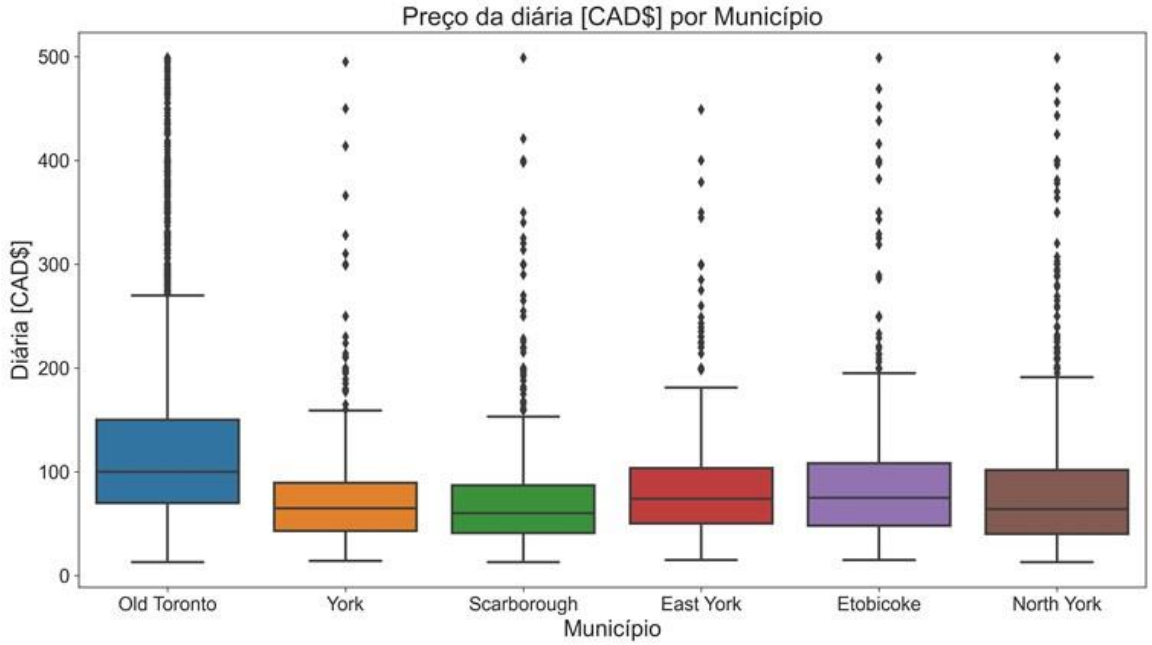


Figura 20. Distribuição de preços por município, em Toronto.

4.4. Geolocalização - Clusters

Anteriormente, foram definidos 8 clusters por Kmeans. Esses clusters podem ser vistos na Figura 21, em um gráfico de dispersão por latitude e longitude. Na Figura 21 um boxplot compara os preços por cluster. Isso contrasta com o boxplot por municípios, e mostra como a segmentação por cluster foi benéfica - visto que é possível notar uma maior diferença dentre os diferentes clusters.

Intuitivamente, o gráfico também faz sentido, pois o cluster 1 e 2 são geograficamente similares aos pares de latitude e longitude de Old Toronto (que a Figura 20 mostra ter um preço maior). Além disso, os clusters 1, 2 e 6 apresentam preços maiores, conforme esperado pela análise da Figura 19

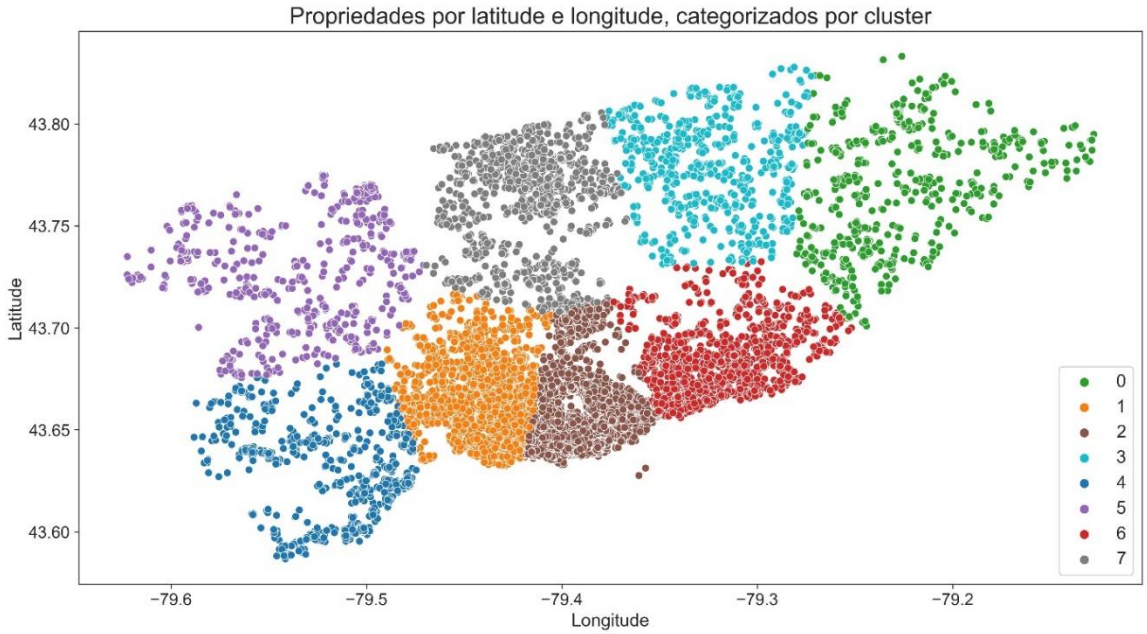


Figura 21. Dispersão de Latitude e Longitude, categorizado por ID do cluster

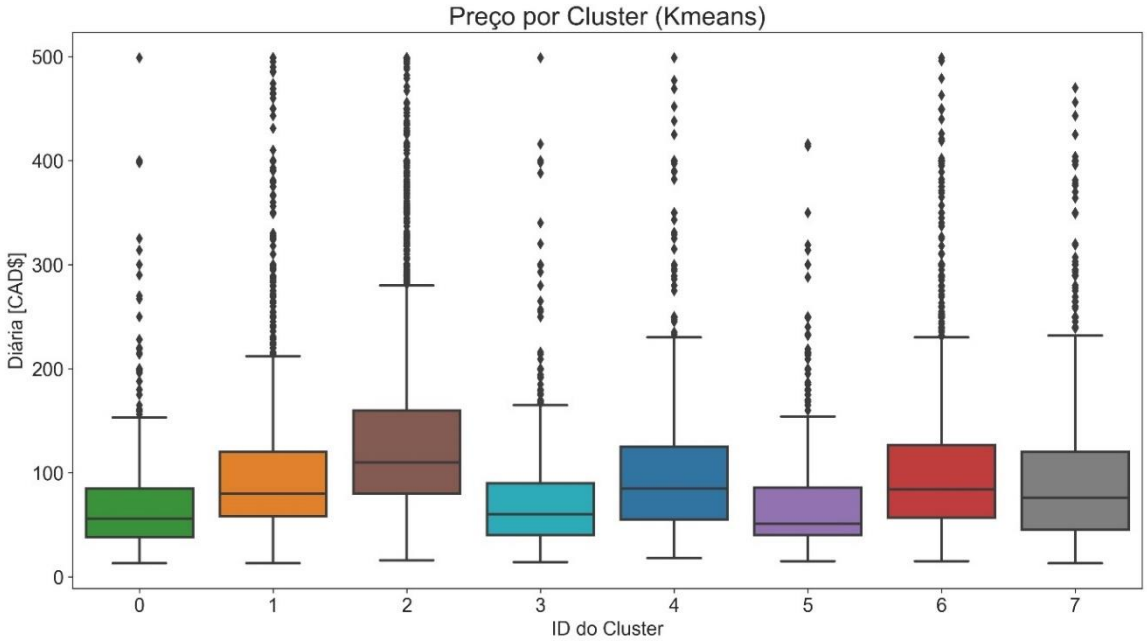


Figura 22. Distribuição de preços por cluster, em Toronto.

4.5. Tipos de Propriedade e Acomodação

A Figura 23 mostra os tipos mais comuns de propriedade e de acomodação. A grande maioria dos imóveis são apartamentos (“Apartment”, 62%) seguido de casas (“House”, 36%). Hoteis e outros representam apenas 2% dos dados.

O tipo mais comum de acomodação é o imóvel inteiro (“Entire home/apt”, 64%), seguido de quarto particular (“Private room”, 34%), mas restante do espaço compartilhado. Quartos compartilhado e quartos de hotéis representam 2% dos dados.

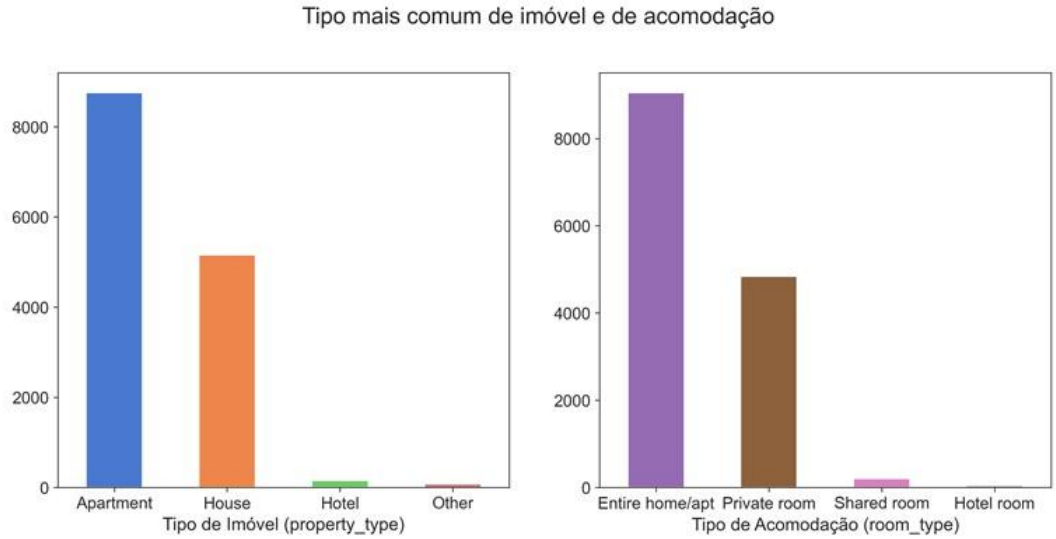


Figura 23. Tipo mais comum de imóvel e de acomodação.

O Anexo 10 analisa os tipos de imóvel quantificados por município. É possível ver que os padrões não se mantem. Enquanto Old Toronto tem mais apartamentos do que casas, conforme a Figura 23 mostra, essa relação se inverte para os outros municípios. York, Scarborough, East York, Etobicoke e North York tem mais casas do que apartamentos. Isso mostra como Old Toronto, por ter muito mais propriedades, distorce os dados.

O Anexo 11 analisa os tipos de acomodação por município. Aqui, novamente, Old Toronto apresenta a mesma relação que a Figura 23, mais imóveis inteiros do que quartos individuais. Já para Scarborough e North York essa relação se inverte, e quartos individuais passam a ser mais comuns do que imóveis inteiros. Para os outros municípios, as quantidades são praticamente iguais.

A Figura 24 mostra os valores de diária por tipo de imóvel. É possível ver que os apartamentos e hotéis apresentam valores diárias superiores, e que as casas são os tipos de imóveis mais baratos. Em todos os casos, a dispersão de dados é bem elevada.

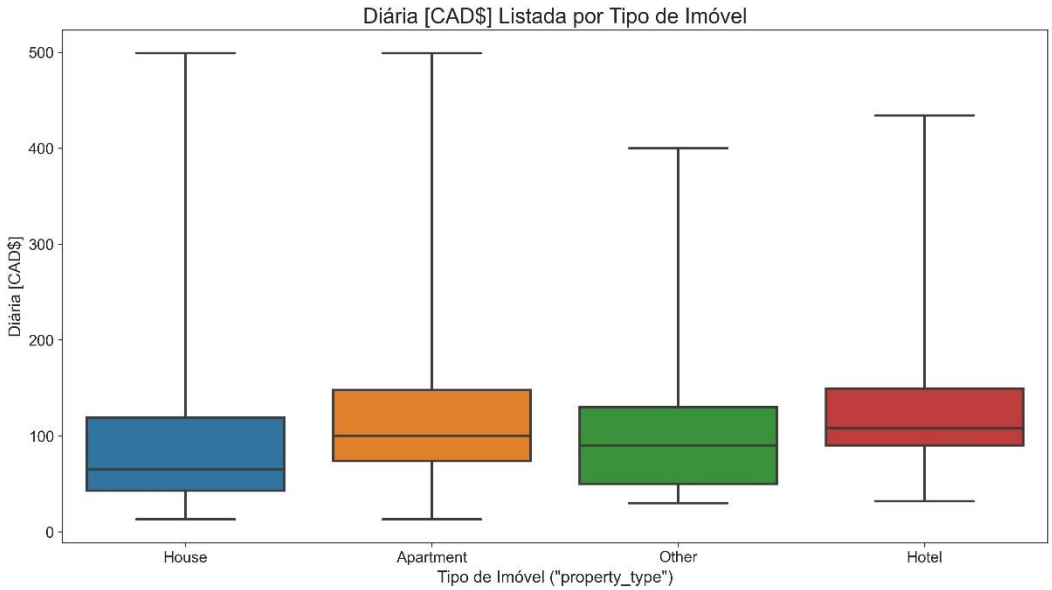


Figura 24. Boxplot do valor da diária [CAD$] por tipo de imóvel.

A Figura 25 faz essa comparação, por tipo de acomodação. Aqui, a diferença é muito mais significativa, e é possível perceber que os anúncios que oferecem o imóvel inteiro têm diárias mais caras do que os outros tipos. O segundo tipo de acomodação mais caro são quartos particulares e quartos de hotéis. Finalmente, os mais baratos são quartos compartilhados.

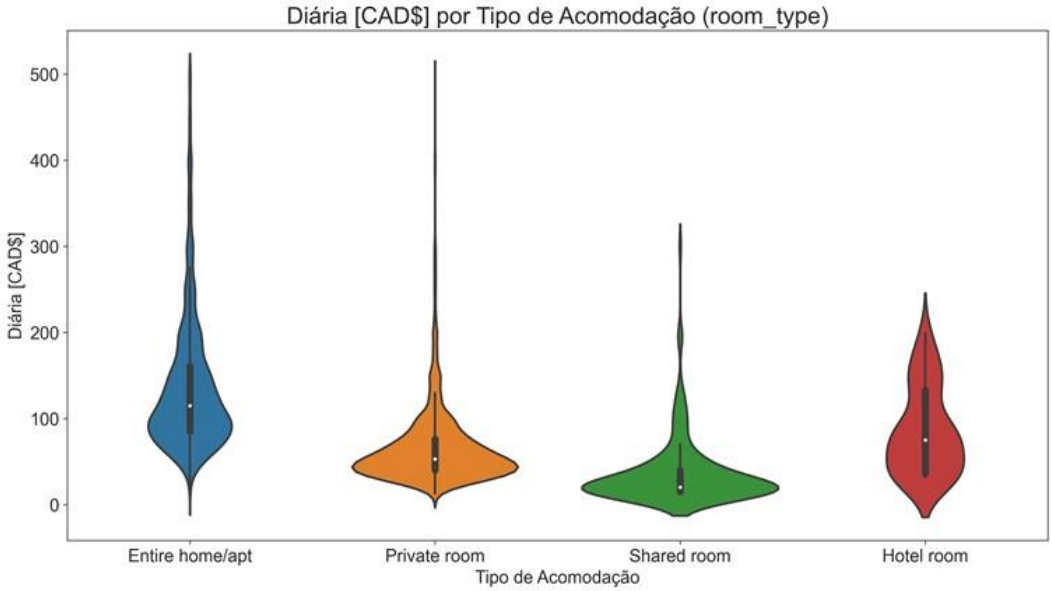


Figura 25. Violinplot do valor da diária [CAD$] por tipo de acomodação.

Finalmente, um imóvel pode, ou não, ser reservável instantaneamente. Ou seja, caso ele seja, o anfitrião não precisa fazer nenhuma confirmação – Um hóspede em potencial pode reservar sem precisar esperar por nada. Por outro lado, imóveis que não se enquadram nessa categoria, precisam da confirmação do anfitrião. A Figura 26 mostra que a grande maioria dos imóveis não pode ser reservado imediatamente, e que o efeito desse atributo no preço é muito pequeno.

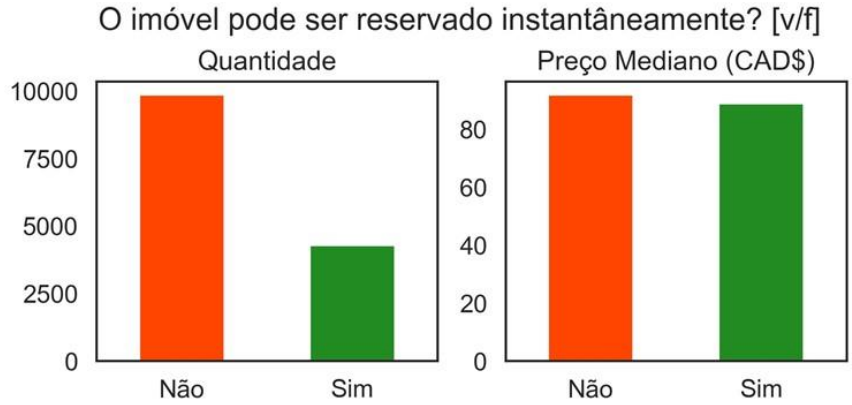


Figura 26. Número de imóveis imediatamente reserváveis, e o efeito deles no preço.

4.6. Quantidade de Pessoas Acomodadas

A Figura 27 mostra que o tipo mais comum de propriedade acomoda até duas pessoas. Além disso, a grande maioria das propriedades acomoda entre 1 e 4 pessoas (85%), e apenas 15% acomodam 5 ou mais pessoas.

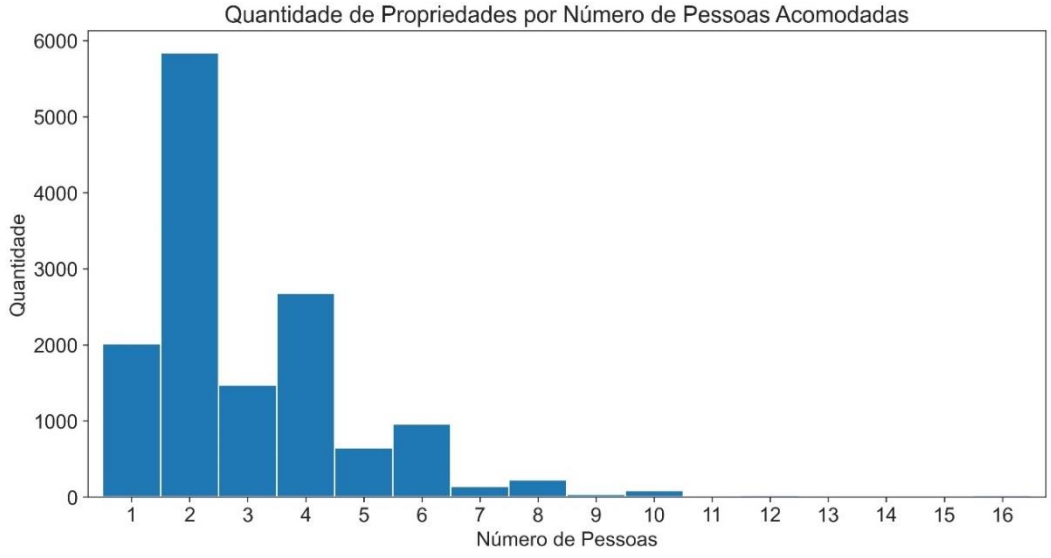


Figura 27. Quantidade de propriedades por número de pessoas acomodadas.

Na Figura 28 é possível ver uma forte correlação positiva entre número de pessoas acomodadas e preço do imóvel. Quanto mais pessoas acomodadas, na mediana, maior o preço da diária.

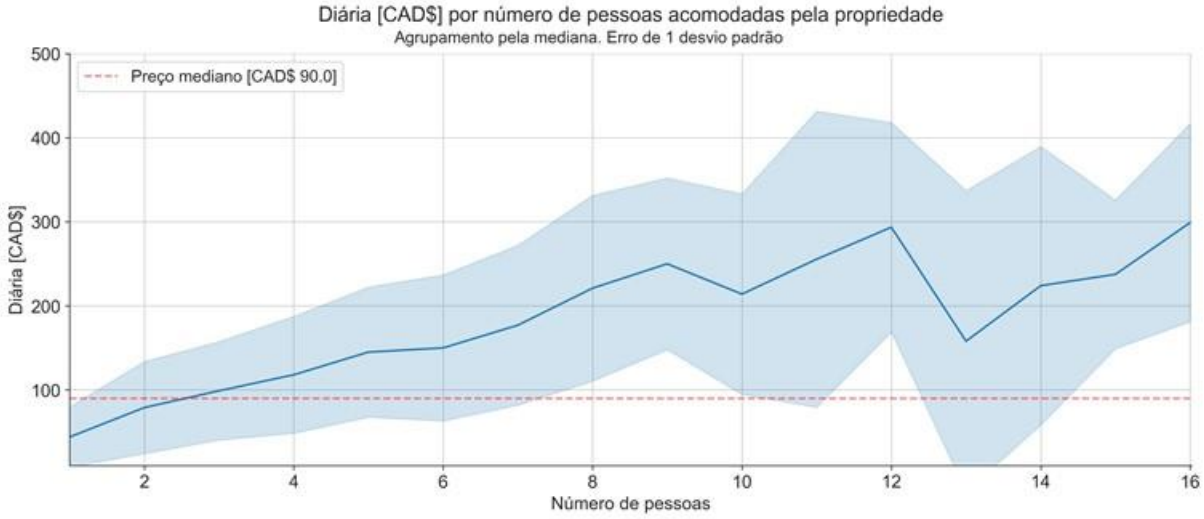


Figura 28. Preço por pessoas acomodadas, agrupado pela mediana.

Apesar do preço do imóvel ser maior para mais pessoas acomodadas, do ponto de vista do hóspede, ainda pode ser economicamente viável selecionar uma propriedade que acomode mais pessoas. Isso é demonstrado na Tabela 13, onde fica claro que por mais que o preço da diária aumente, o preço da diária por pessoa diminui. Em outras palavras, economicamente é melhor selecionar uma propriedade pro exato número de pessoas que estão alugando o imóvel.

Tabela 13. Diária por pessoa, por número de pessoas acomodadas.

Pessoas Acomodadas	Diária (Mediana)	Diária / Pessoa
1	44	44
2	79	40
3	99	33
4	118	30
5	145	29
6	150	25
7	177	25
8	221	28
9	250	28

10	214	21
11	256	23
12	294	24
13	158	12
14	224	16
15	238	16
16	299	19

4.7. Amenidades

O modelo considera muitas amenidades, é inviável analisar todas elas graficamente. Por isso, apenas as amenidades que têm um efeito de pelo menos 15% na mediana do preço da diária serão demonstrados. Os outros continuaram sendo incluídos nos modelos, apesar de não serem demonstrados nessa seção. Após filtrar por esses valores, encontramos que as principais amenidades são:

- Televisão (TV) – Efeito Positivo
- Secadora de Roupas (Dryer) – Efeito Positivo
- Lavadora de Louças (Dishwasher) – Efeito Positivo
- Elevador (Elevator) – Efeito Positivo
- Academia (Gym) – Efeito Positivo
- Quarto com Tranca (Lock on Bedroom Door) – Efeito Negativo
- Piscina (pool) – Efeito Positivo

De todos esses, talvez o mais surpreendente seja o quarto com tranca ter um efeito negativo no preço. Mas isso pode ser explicado pois os únicos imóveis que tem quarto com tranca são, também, imóveis compartilhados. Conforme análises da Figura 25, imóveis compartilhados tem preços de diárias menores. Finalmente, o gráfico de frequência e preço para cada uma dessas amenidades é disponibilizado nos Anexo 12 ao Anexo 18

5. Criação de Modelos de Machine Learning

5.1. Preparação – Remoção de Colunas

Algumas colunas foram utilizadas até agora exclusivamente para permitir diferentes visualizações e traçar conclusões sobre os dados. Essas colunas não serão utilizadas nos modelos de machine learning, e, portanto, serão removidas.

Inicialmente, host_since, last_review e first_review foram removidas. Essas três colunas são colunas de datas e já tiveram atributos que criados a partir delas: host_since_days, last_review_days e first_review_days. Os atributos criados são numéricos e serão incluídos no modelo.

Depois a latitude e longitude foram removidas. Isso porque o modelo vai utilizar, conforme concluído na seção de análise e exploração dos dados, os clusters produzidos a partir da latitude e longitude. Pelo mesmo motivo, a coluna geo_city, com o nome do município, foi removida. Finalmente, a coluna host_id será removida, pois apresenta um código de identificação (metadados) que não é relevante para o modelo.

Após a remoção dessas colunas, o dataframe passou a ter 14150 linhas e 74 colunas. Ou seja, o número de linhas permaneceu igual, e 7 colunas foram removidas.

5.2. Preparação – Criação de Dummies

A criação de “dummies” é o processo de converter uma variável categórica em N atributos novos, onde N é o mesmo valor de níveis de categoria [22]. Foi por esse motivo que diversos atributos foram convertidos em categoria ao longo desse estudo. Aqui, foram criados dummies para:

- Colunas do Anfitrião e do imóvel: host_response_time, host_response_rate, host_acceptance_rate, property_type, room_type.
- Colunas de Avaliação: first_review_days, last_review_days, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value
- Coluna de NLP: review_polarity.
- Coluna de Geolocalização: geo_cluster.

Após a criação dos dummies, o dataframe passa a ter 14150 linhas com 136 colunas. Ou seja, adição de 55 colunas.

5.3. Preparação – Correlação Entre as Variáveis

Por mais que, durante o estudo, cuidado tenha sido tomado para evitar correlação entre as variáveis (através de transformação, e remoção de redundâncias), ainda é importante verificar o grau de correlação. Para isso, foi utilizada matriz de correlação. Essa matriz pode ser vista graficamente através do mapa de calor, na Figura 29. Como

esse dataframe possui muitos atributos, a matriz de correlação foi dívida em duas partes, para facilitar a visualização do leitor. Antes dessa divisão, o autor confirmou que a divisão seria feita na parte correta para minimizar as correlações altas, e, portanto, não alterar a interpretação.

É possível ver que temos correlações altas ao longo de diversos atributos. Vamos remover apenas os que tem as maiores correlações. A correlação máxima encontrada foi de 1.0, e mínima de -0.96.

O critério de remoção foi uma correlação, em módulo, acima de 0.99. Não surpreendente, as colunas removidas são todas as 7 colunas de avaliações e host_response com valor “no response yet” ou “No review yet”, já que nesses casos, uma ausência de avaliação impacta todas as colunas por igual.

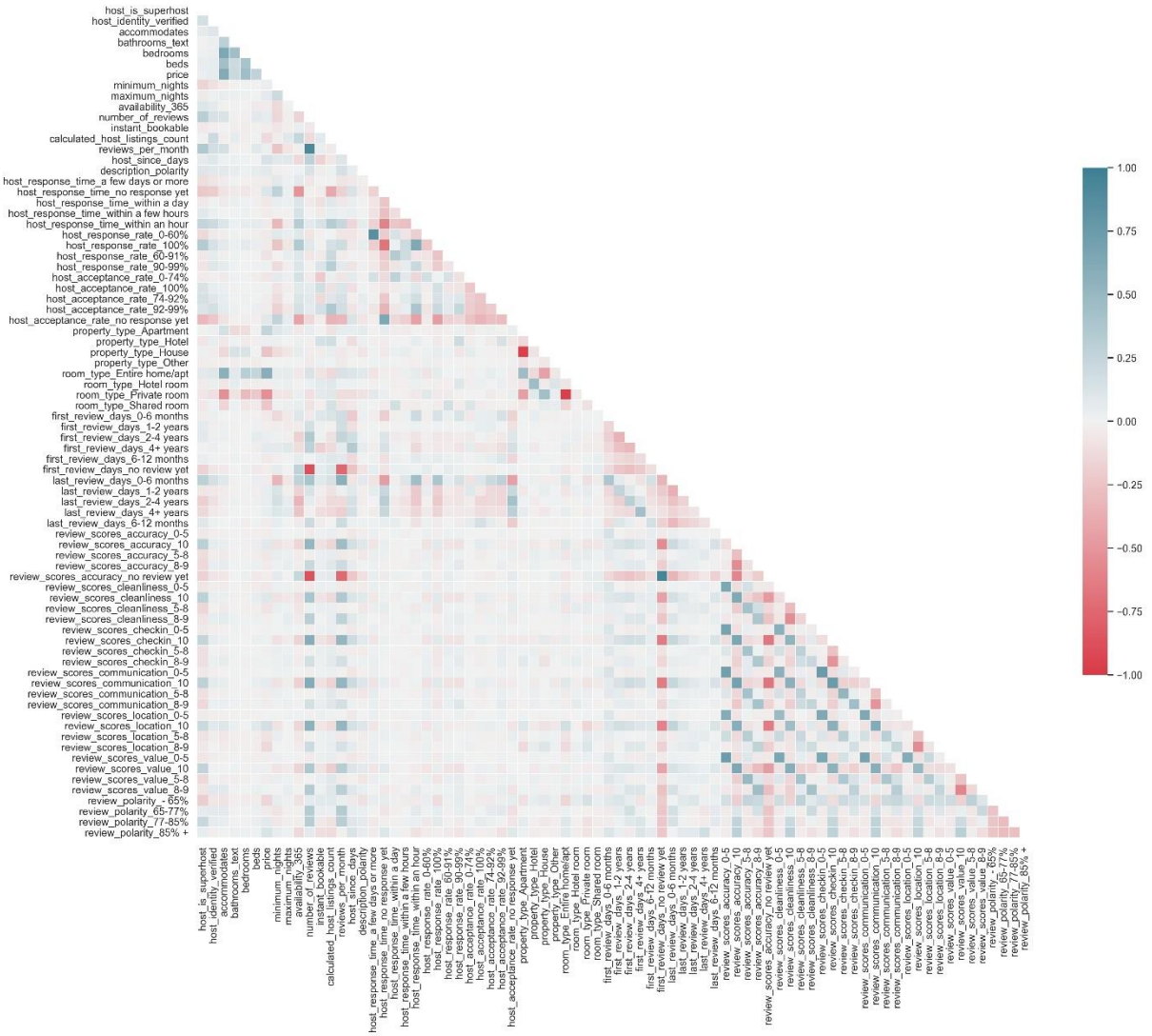


Figura 29. Mapa de Calor da Matriz de correlação – Parte 1.

A segunda parte da matriz de correlação pode ser vista na Figura 30. A correlação máxima encontrada foi de 0.93, e a mínima de -0.33. Nenhum valor vai ser removido como consequência disso, mas isso traz algumas interpretações aos dados.

É possível ver que essas correlações são coerentes com a expectativas. Por exemplo, imóveis com máquinas de secar roupas (Dryer), frequentemente também tem máquinas de lavar (Washer). Propriedades com pratos e talheres (Dishes and Silverware), frequentemente tem geladeiras (Refrigerator). Espaços que tem micro-ondas (microwave), também tem geladeiras. Finalmente, casas e apartamentos com Stove (Fogão), também tem Oven (Forno).

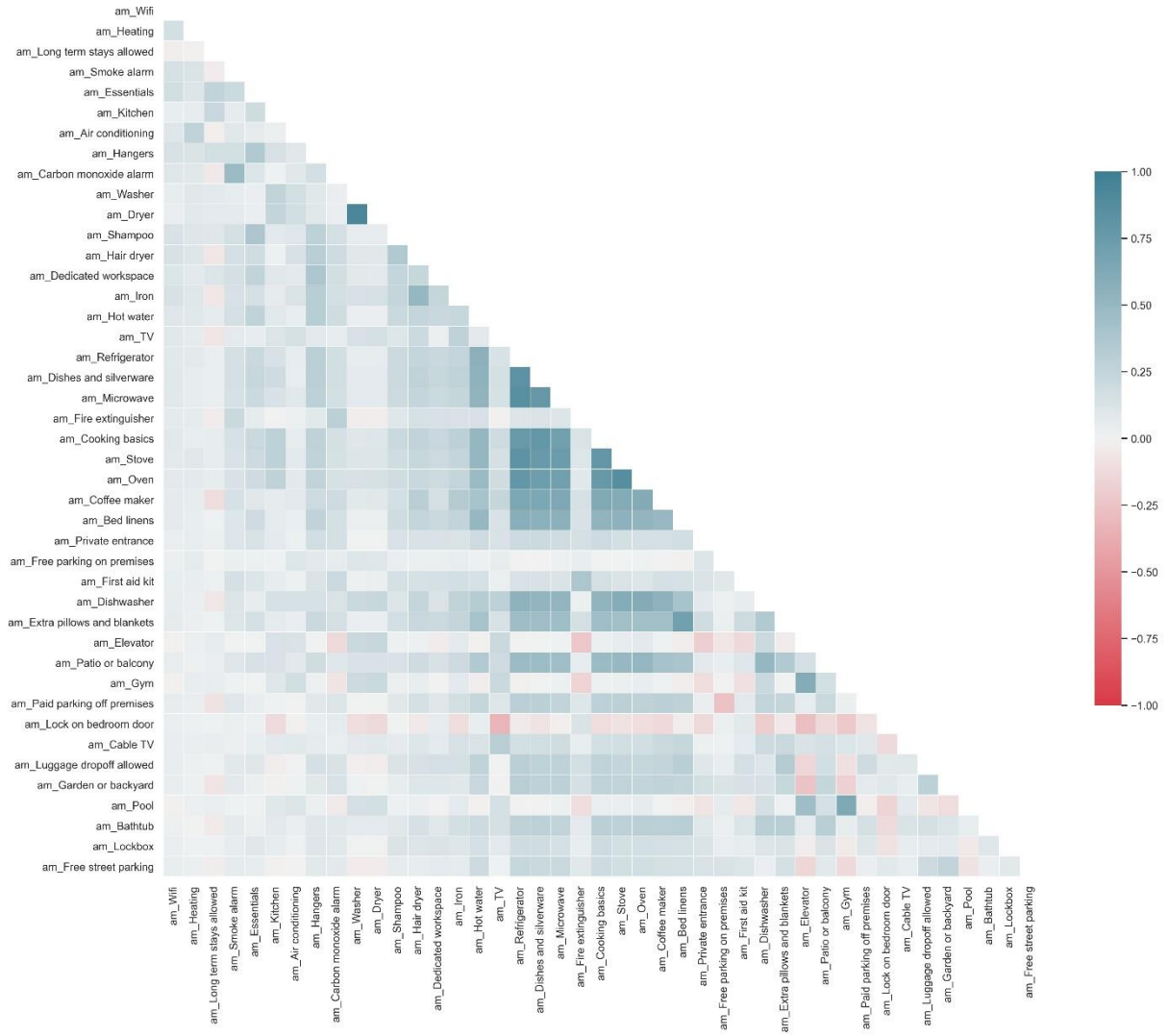


Figura 30. Mapa de Calor da Matriz de correlação – Parte 2.

5.4. Preparação – Transformação dos Valores Numéricos

Ao longo de todo o estudo foi possível ver que os valores são, na grande maioria das vezes, fortemente assimétricos. Transformação logarítmica é ideal para dados assimétricos [23]. Por isso, todas essas colunas serão transformadas. Antes da transformação, todos os valores 0.0 são substituídos por 0.01, visto que valores 0 não podem ser transformados logaritimicamente. Como são muitas colunas, a distribuição delas antes da transformação pode ser vista no Anexo 19, e após transformação no Anexo 20.

5.5. Preparação – Normalização e Divisão dos Dados

Finalmente, a coluna de interesse é separada (“price”), e os atributos de predição são normalizados através de um scaler da biblioteca sklearn (preprocessing.StandardScaler). Isso é feito para escalar os atributos pela sua própria variância, e é uma exigência para os algoritmos de machine learning que serão utilizados mais à frente.

Os dados são então particionados em 80% para o grupo de treino (training set) e 20% para o grupo de teste (testing set), utilizando um random_state de 1092867, para permitir reprodução dos resultados.

5.6. Definição de Modelo e Métricas

Como os dados têm um número relativamente elevado de features com alta correlação, dois algoritmos que são robustos com relação a isso são selecionados. Random Forest (Biblioteca sklearn [24], RandomForestRegressor) e Gradient Boosting (Biblioteca XGBoost [25], XGBRegressor).

Serão utilizadas duas métricas. A primeira é o R^2, ou Coeficiente de Determinação. Essa métrica é a soma dos quadrados dos erros do valor previsto, com relação ao valor médio do real. O valor desejável é o mais próximo de 1.0 (ou seja, não existiriam erros)

A segunda métrica selecionada foi o MSE, ou Erro Quadrático Médio (Mean Squared Error). Essa métrica mede as médias dos quadrados dos erros do valor previsto em relação ao valor real. O valor desejável é o mais próximo de 0 o possível.

5.7. Aplicação dos Modelos

Primeiro, o algoritmo de Random Forest é aplicado, com 250 estimadores (n_estimators). Com esse algoritmo, obtemos as métricas conforme Tabela 14.

Tabela 14. Métricas do algoritmo Random Forest.

Random Forest - 1			
Training MSE	0,018	**Test MSE**	0,126
Training R^2	0,955	**Test R^2**	0,688

.

Em seguida, o algoritmo de gradient boost é aplicado. Com esse, obtemos as métricas da Tabela 15.

Tabela 15. Métricas do algoritmo Gradient Boost.

Gradient Boost (XGBoost) - 1			
Training MSE	0,041	**Test MSE**	0,128
Training R^2	0,898	**Test R^2**	0,683

5.8. Filtro dos Atributos mais Importantes

Ainda utilizando a biblioteca de gradient boosting, é possível analisar quais atributos são mais importantes, e quais são poucos relevantes. Dos 128 atributos, apenas 22 tem mais de 0.5% de importância. Esses são demonstrados na Figura 31.

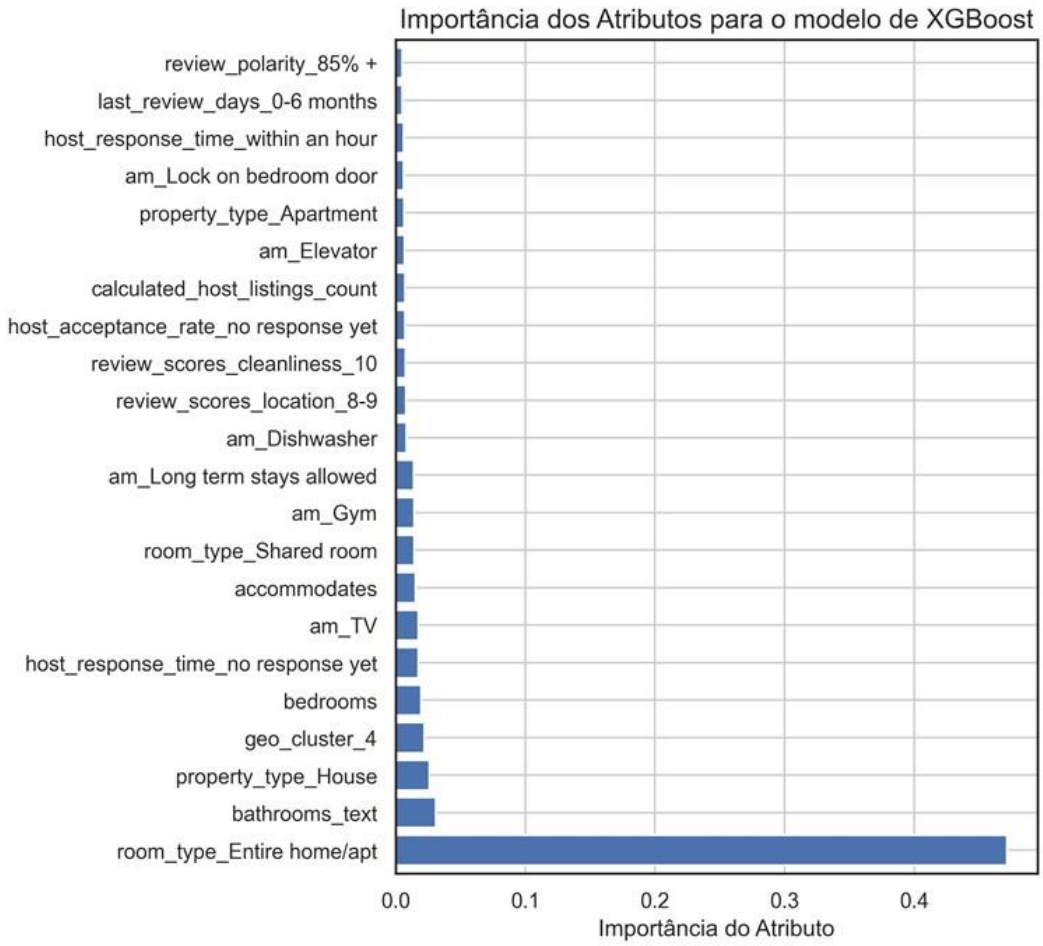


Figura 31. Atributos com mais de 0.5% de importância

Usando esses atributos mais importante como base, um total de 66 atributos são removidos. Essa remoção é feita apenas de atributos não conectados com os 22 mais importantes. Por exemplo, review_polarity_85% é considerado o 22° mais importante. Portanto, review_polarity_-65% review_polarity_77-85%, etc., são mantidos. Isso diminui o total de atributos de 128 para 62

5.9. Interpretação dos Atributos mais Importantes

Como esperado pelas análises feitas na seção de análise e exploração dos dados, o atributo mais importante para o preço é o tipo de acomodação. Particularmente, imóvel inteiro (entire home/apt) com a maior importância, seguido de quarto compartilhado (shared room). O tipo de imóvel também é relevante, sendo o tipo de casa (house) o mais impactante, seguido de apartamento (apartment).

A quantidade de pessoas acomodadas pelo imóvel, também visto anteriormente, também foi considerado um fator importante. Das amenidades, apenas televisões (TV), academia (Gym), Lava-louças (Dishwasher) e elevadores (Elevator) e tranca na porta do quarto (Lock on bedroom door) entraram na lista final. Finalmente, foi possível ver que a categorização das latitudes e longitudes em cluster fez efeito, pois geo_cluster_4 foi o quarto atributo mais importante – Ele era um dos clusters correspondente a latitude e longitude de Old Toronto.

5.10. Reaplicação dos Modelos com Menos Atributos

Os dois modelos são reaplicados após a redução dos modelos. Os novos valores obtidos podem ser vistos nas Tabela 16 e Tabela 17.

Tabela 16. Métricas do algoritmo Random Forest – Após Remoção.

Random Forest – 2			
Training MSE	0,018	**Test MSE**	0,126
Training R^2	0,955	**Test R^2**	0,689

Tabela 17. Métricas do algoritmo Gradient Boost – Após Remoção.

Gradient Boost (XGBoost) - 2			
Training MSE	0,043	**Test MSE**	0,125
Training R^2	0,894	**Test R^2**	0,691

6. Apresentação dos Resultados

6.1. Seleção do Melhor Modelo

Em todos os testes feitos, os modelos utilizando Random Forest apresentaram melhores métricas no grupo de treino, mas piores no grupo de teste. Com critério de selecionar o melhor algoritmo no grupo de teste, os modelos feitos com Gradient Boosting se sobressaíram.

O melhor modelo de predição, e, portanto selecionado como conclusão desse trabalho, foi o Gradient Boosting (XGBoost) após a remoção dos atributos com baixa importância. Embora as métricas (R^2 ou MSE) não tenham melhorado significativa com a remoção desses atributos, o que melhorou foi o tempo de execução e necessidade de recursos para o processamento – que também é um fator importante na decisão de um modelo.

Tabela 18. Melhor modelo de predição

Gradient Boosting (XGBoost) - 2			
Training MSE	0,043	**Test MSE**	0,125
Training R^2	0,894	**Test R^2**	0,691

Embora a precisão não tenha melhorado significativamente com remoção de atributos, o tempo de execução e a exigência de recursos foram significativamente reduzidas. Portanto, os melhores modelos são após a remoção das colunas com baixa importante.

6.2. Recomendações aos Anfitriões e Hóspedes

Esse estudo apresentou diversos gráficos analisando os atributos mais importantes e sua distribuição. Esses não serão repetidos aqui por redundância, mas podem ser encontrados na seção de Análise e Exploração dos Dados e Anexo. Esses dados podem ser utilizados por anfitriões para maximizar seu lucro, ou por hóspedes, para maximizar seu custo-benefício

Como conclusões, é interessante notar que um hóspede que deseje maximizar seu custo-benefício deve sempre procurar propriedades que acomodem o exato número de pessoas que está viajando com ele. Ou seja, um hóspede viajando em um grupo de 3 alugando uma propriedade para 3 pessoas vai pagar um menor valor por pessoa do que um hóspede viajando sozinho.

A localização também se mostrou um atributo crítico, e, portanto, é possível conseguir propriedades com menores preços em locais mais afastados da zona sul de Toronto. Outro fator relevante é o tipo da acomodação: Casas são mais baratas

que apartamento, e quartos compartilhados mais baratos que propriedades que oferecem o apartamento inteiro.

Finalmente, se viajar com mais pessoas, selecionar um local mais afastado ou alterar o tipo do imóvel não forem opções viáveis para um hóspede, este pode considerar remover de sua busca amenidades que ele não tem interesse. Imóveis sem televisões, academia, e lava-louças também apresentam uma alta correlação com o preço.

6.3. Limitações e Sugestões para Estudos Futuros

A maior limitação desse estudo foi o fato de as propriedades possuírem apenas o preço listado. Uma análise muito mais rica seria possível se, além do preço listado, os dados disponíveis de todos as reservas já feitas no passado também estivessem disponíveis.

Como sugestão adicional, seria interessante expandir esse estudo para outras cidades que são grandes pontos turísticos [26], como: Bangcock (Tailândia), Londres (Reino Unido), Paris (França), Dubai (Emirados Árabes Unidos) e Nova Iorque (Estados Unidos). Com essa comparação, seria interessante ver se os atributos importantes são os mesmos, e caso não sejam, se poderiam ser aplicados para obter bons resultados em outras cidades.

7. Links

Os scripts referenciados nesse relatório podem ser encontrados no repositório do github: https://github.com/notidentical/Airbnb-Toronto-Analysis. Como o Github limita o tamanho de arquivos em 100 MB [27], esse repositório não inclui a fonte de dados com as avaliações (reviews.csv). Isso não é um problema, pois o download desse arquivo, tão como dos outros arquivos, ainda é feito pela execução do script download_airbnb.ipynb. De qualquer forma, um link alternativo do projeto inteiro (incluindo todas as fontes de dados) é fornecido pelo Google Drive: https://drive.google.com/file/d/102uJpRA2bD4f-RFY0TNUZ_tvRbV4f7cM/view.

O download dos arquivos é feito, conforme já mencionado, pelo script download_airbnb.ipynb, as descrições e avaliações são processadas pelos scripts NLP_process_descriptions.ipynb e NLP_process_reviews.ipynb, e os municípios são

obtidos pelo script GEO_process_coordinates.ipynb. Finalmente, a análise em sí é feita pelo script analysis.ipynb.

Finalmente, a apresentação pode ser encontrada no youtube, através do link: https://www.youtube.com/watch?v=0capktXIFPs.

REFERÊNCIAS

1. ABOUT-US. **Airbnb**, 2021. Disponivel em: <https://news.airbnb.com/about-us/>. Acesso em: Abril 2021.

2. AIRBNB-STATISTICS. **Stratosjets**, 2021. Disponivel em: <https://www.stratosjets.com/blog/airbnb-statistics/>. Acesso em: Abril 2021.

3. TOURISM. **Toronto**, 2021. Disponivel em: <https://www.toronto.ca/business-economy/industry-sector-support/tourism/>. Acesso em: Abril 2021.

4. ABOUT. **Inside Airbnb**, 2021. Disponivel em: <http://insideairbnb.com/about.html>. Acesso em: Abril 2021.

5. GET-THE-DATA. **Inside Airbnb**, 2021. Disponivel em: <http://insideairbnb.com/get-the-data.html>. Acesso em: Abril 2021.

6. AIRBNB. **Termos de Serviço**, 2021. Disponivel em: <https://www.airbnb.com.br/help/article/2908/termos-de-servi%C3%A7o>. Acesso em: Abril 2021.

7. CC01.0. **Creative Commons**, 2021. Disponivel em: <https://creativecommons.org/publicdomain/zero/1.0/>. Acesso em: Abril 2021.

8. GEOCODING API. **Mapquest Developer**, 2018. Disponivel em: <https://developer.mapquest.com/documentation/geocoding-api/reverse/get/>. Acesso em: Abril 2021.

9. PLANS. **Mapquest Developer**, 2021. Disponivel em: <https://developer.mapquest.com/plans>. Acesso em: Abril 2021.

10. LEGAL. **Developer Mapquest**, 2021. Disponivel em: <https://developer.mapquest.com/legal>. Acesso em: Abril 2021.

11. SUPERHOST?, H. D. I. B. A. How. **Airbnb**, 2021. Disponivel em: <https://www.airbnb.com/help/article/829/how-do-i-become-a-superhost>. Acesso em: Abril 2021.

12. MCKINNEY, W. pandas: a Foundational Python Library for Data Analysis and Statistics, 2011.

13. MEANING of Indicating bathroom as 1 or 0.5. **Community Center**, 2021. Disponivel em: <https://community.withairbnb.com/t5/Help/Meaning-of-Indicating-bathroom-as-1-or-0-5/td-p/105155>. Acesso em: Abril 2021.

14. KUANGJIE, Z.; WADHWA, M. This Number Just Feels Right: The Impact of Roundedness of Price Numbers on Product Evaluations. **Journal of Consumer Research**, v. 41, n. 5, p. 1172-1185, Fevereiro 2015. ISSN 10.1086/678484.

15. GILBERT, E.; HUTTO, C. J. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. **Conference: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media**, Ann Arbor, MI, Janeiro 2015.

16. NA, S.; XUMIN, L.; YONG, G. Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm. **IEEE**, Abrl 2010. ISSN 11261758.

17. TRAVEL Budget for Toronto. **Budget Your Trip, LLC**, 2021. Disponivel em: <https://www.budgetyourtrip.com/canada/toronto>. Acesso em: Abril 2021.

18. BOROS, L.; DUDÁS, G.; KOVALCSIK, T. The effects of COVID-19 on Airbnb. **Hungarian Geographical Bulletin**, v. 4, n. 69, p. 363-381, Dezembro 2020. ISSN 10.15201/hungeobull.69.4.3.

19. FILE:NORTH York Locator.png. **Wikipedia**, 2021. Disponivel em: <https://en.wikipedia.org/wiki/File:North_York_Locator.png>. Acesso em: Abril 2021.

20. FOLIUM. **Python-visualization**, 2013. Disponivel em: <https://python-visualization.github.io/folium/>. Acesso em: Abril 2021.

21. ABOUT. **Open Street Map**, 2021. Disponivel em: <https://www.openstreetmap.org/about>. Acesso em: Abril 2021.

22. DUMMY variable (Statistics). **Semantic Scholar**, 2021. Disponivel em: <https://www.semanticscholar.org/paper/Dummy-variable-(-statistics-)/0607ecffe2f5fafba982c24e243ed3d388cd0298>. Acesso em: Abril 2021.

23. CURRAN-EVERETT, D. Explorations in statistics: the log transformation. **Advances in Physiology Education**, v. 40, n. 2, p. 343-347, Dezembro 2016.

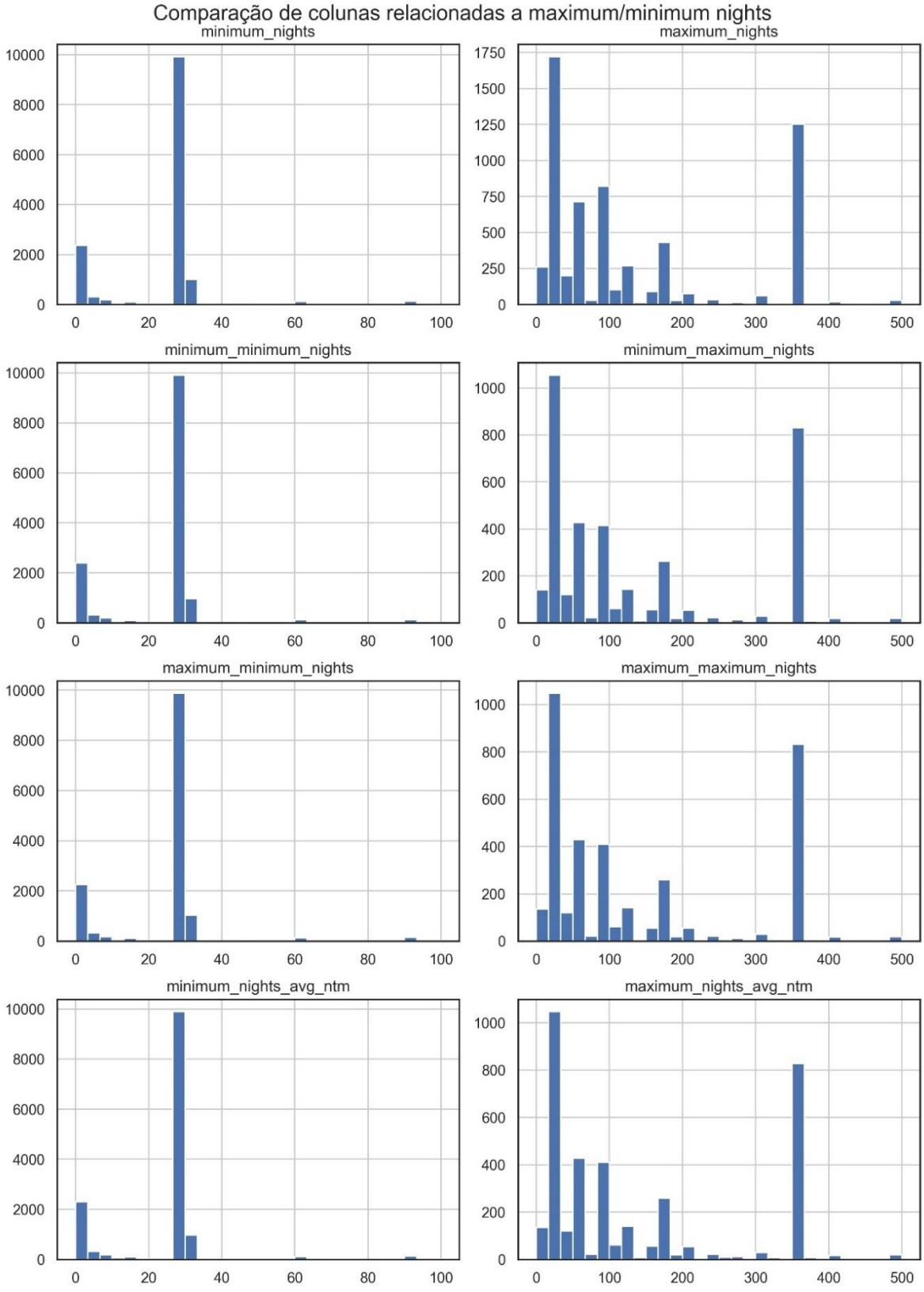
24. ABOUT us. **scikit-learn**, 2021. Disponivel em: <https://scikit-learn.org/stable/about.html>. Acesso em: Abril 2021.

25. CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. **ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, v. 22, p. 785-794, Agosto 2016. ISSN 10.1145/2939672.2939785.

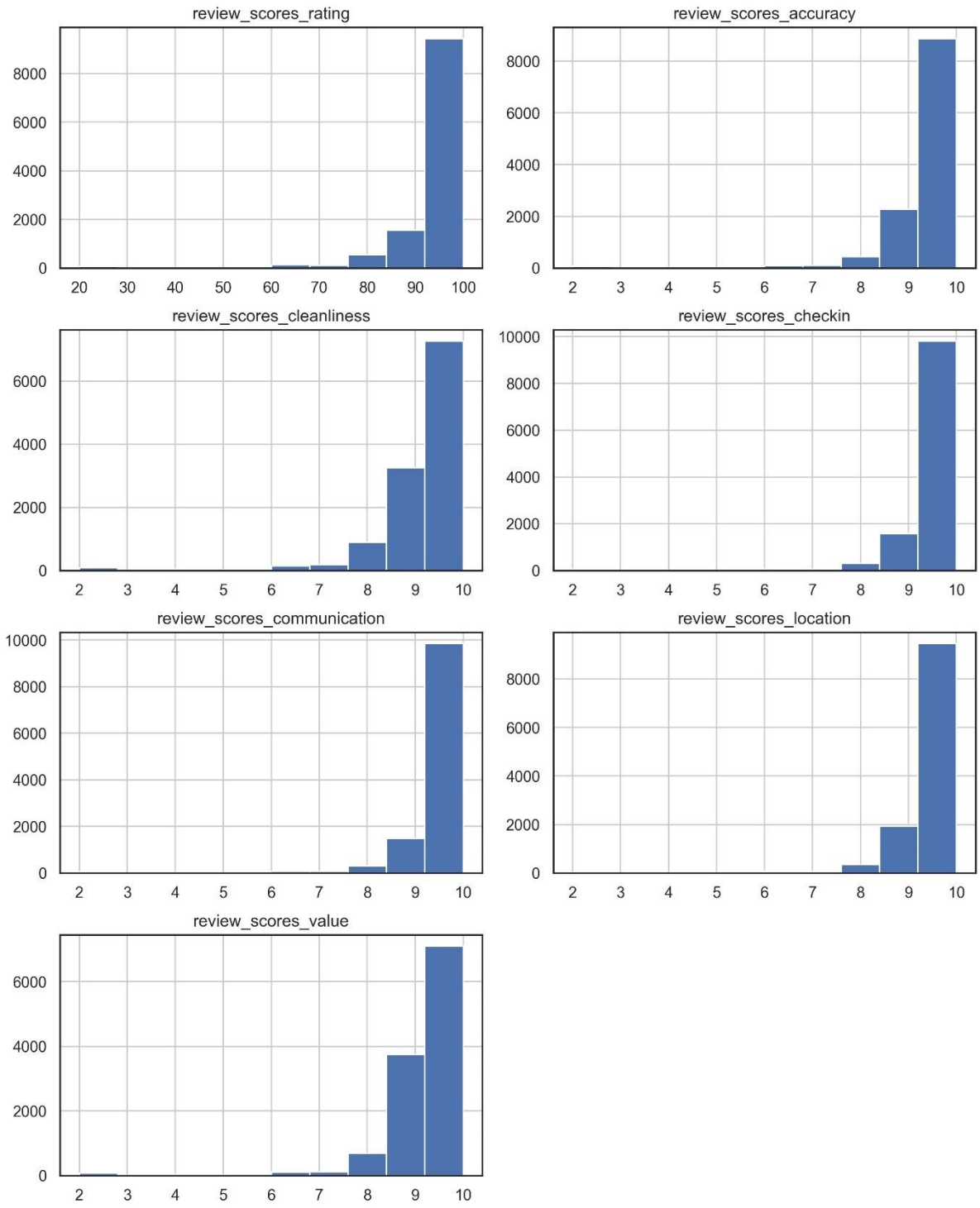
26. BUSINESS Insider. **The 19 most visited cities around the world in 2019**, 2019. Disponivel em: <https://www.businessinsider.com/most-visited-cities-around-the-world-ranked-2019-9#7-new-york-city-136-million-13>. Acesso em: Abril 2021.

27. CONDITIONS for large files. **github**, 2021. Disponivel em: <https://docs.github.com/en/github/managing-large-files/conditions-for-large-files#warning-for-files-larger-than-50-mb>. Acesso em: Abril 2021.

APÊNDICE



Anexo 1. Comparação das colunas que descrevem maximum e minimum nights.



Anexo 2. Distribuição das colunas review_scores.

Patrick's home was absolutely amazing! We were so grateful every day of our family vacation that we chose his home to stay in. It is roomy and beautiful and had everything we could possibly need. Comfortable beds with nice sheets. Lots of towels and a fully stocked kitchen. Hundreds of books to look through, and interesting art to admire made downtime fun. It was great to have lots of space for spreading out and relaxing between outings. The giant jacuzzi tub was amazing! It was chilly when we were there, but we still appreciated the amazing view of the city from the fabulous roof top deck. We all had to go out every night to see, and photograph the CN tower, as it is lit up differently every night. We loved the two fireplaces and gas wood stove. The location was wonderful, too. It was very easy to get anywhere we wanted to go. We used the street cars and subways and never drove our car once. Kensington market was such a fun neighborhood. We saw many of the classic tourist sights in Toronto, but our favorite times were spent strolling around Patrick's neighborhood sampling all the great eats and shopping in the fun, quirky stores. Chinatown is right around the corner, and we truly enjoyed this area as well. We never once felt unsafe. The square the house sits on is being redone, and I'm sure will be beautiful when finished. It looks like a nice play ground is being installed. Another plus to this location is how surprisingly quiet it is, for being right in the city! It is a four level single family home, which means no one above or below! Patrick was an amazing host and answered all of our questions promptly. He provided lots of information tailored to our interests, which, I must admit, involved food for us most of the time! He has lived in Toronto for a long time, and is a wealth of information. We were thrilled that we got to spend some time getting to know him, and will always treasure our time together. I am sure he is always an impeccable host, and would be happy to check in on his guests, or give them all the privacy they wanted. We felt comfortable and at home from the minute we stepped through the front door, to the minute we left. If we return again, I hope we can stay here, possibly in the summer to enjoy the rooftop deck and cute backyard. The stairs in the house would not work well with young children, but Patrick explains that very well in his description. We loved it, however, and felt truly lucky that Patrick shared his lovely, interesting, one-of-a-kind home with us!

ID: 247436497

Anexo 3. Melhor avaliação (polarity_score = +0.9996). Texto não alterado.

First of all: the location of the apartment is absolutely magnificent and the view to the CN Tower is really spectacular. We couldn't get enough of the view within the 5 days. The communication with Ramy was also very easy and uncomplicated. He always responded very quickly, which was really great. Under other circumstances this review would have been better, but unfortunately our stay at the beginning was a bit different than expected. We are aware that Ramy is absolutely not to blame for all this, but it has an impact on our evaluation. We have received all information regarding check-in from Ramy in advance. Registration with the concierge was really uncomplicated. Arrived at the top of the apartment, there was a problem with the lock, so we were not able to open the door as discussed. We then had to wait over an hour for Ramy to solve the door problem with his key. Unfortunately the door could not be opened even with his key, so we had to wait another 45 minutes for the locksmith. It took about 3 hours until we were finally in the apartment, but unfortunately without a working latch, because this was removed by the lock service due to the defect. We only had the remaining lock, which worked separately to the latch, to lock the door. Ramy was so kind and refunded us some money because of this "inconvenience", which we found really great, because from our point of view this was not self-evident. By the way, the repair could not be done during our 6-day stay, as a weekend and a rather unfriendly employee of the locksmith company stood in the way. In the same night at the day of arrival there was also a fire alarm, during which 3-5 fire engines with blue lights arrived and the building was inspected by the firemen. Fortunately, it was just a false alarm (probably not the first of its kind), but the instructions through the building's speakers were very hard to understand and not very informative. - For example, it was not said whether the building would be evacuated and whether the elevator could be used. From our point of view, this information matters if you are staying at the 57th floor (where the apartment is located). Generally speaking regarding the apartment: It is very clean and really modern. Unfortunately, the bathroom is a bit small and there is no possibility to leave your cosmetics or your toilet bag in the bathroom, as there is not enough storage space. Probably because of the height of the floor it was not possible to open the windows, unfortunately. Everything ran exclusively over the A/C, which took us some time getting used to it. The stone items on the wall are very decorative, but not very high quality. During our stay one of these items detached itself from the wall without our external influence. From our point of view, the kitchen could be equipped a little better. For example, pans and scissors are missing. Altogether a few more plates and cutlery would be desirable, because for 4 plates and the appropriate cutlery it is not really worthwhile to put the dishwasher into operation. A washing machine and a dryer are available in the apartment. But first we had to empty the dryer ourselves in order to use it, because the cleaner did not empty the towels of the previous guests. A small guest book with all information about the apartment and the building would be recommendable, e.g.: important phone numbers; where is the fitness room located; information about the house rules; information about the use of the technical equipment; information about the garbage disposal etc.

ID: 540026446

Anexo 4. Segunda pior avaliação (polarity_score = -0.9985). Texto não alterado.

Recently renovated- walls in bedrooms, all floors

3 storey townhouse. With Parking, quiet/100 walking score neighbourhood.

Private rooftop patio with BBQ. 2 bedrooms- masterbedroom with Queensize matress, second with queen bed and a large Queensize airmatress for 2ppl

Perfect location- shopping/restaurants/bars and proximity - Kensigton, Chinatown, QueenWest, downtown

Other things to note

Now its been 2 times my place got completely trashed and tons of things got stolen. I dont really feel like going though all the trobles with police and insurance so be aware of it. So from now on No parties, management is not happy about noise complains and I dont want to be kicked out from my own home

ID: 15465840

Anexo 5. Segunda pior descrição (polarity_score = -0.9063). Texto não alterado.

Our 2 bedroom home is in the heart of Riverdale, a block from the TTC, 2 beautiful parks, and the vibrant Danforth and Leslieville areas. We are minutes from downtown, but in an amazing residential area with mature trees and great people. Newly decorated and furnished, this is a beautiful and comfortable house in a prime neighbourhood. A perfect summer oasis and home away from home.

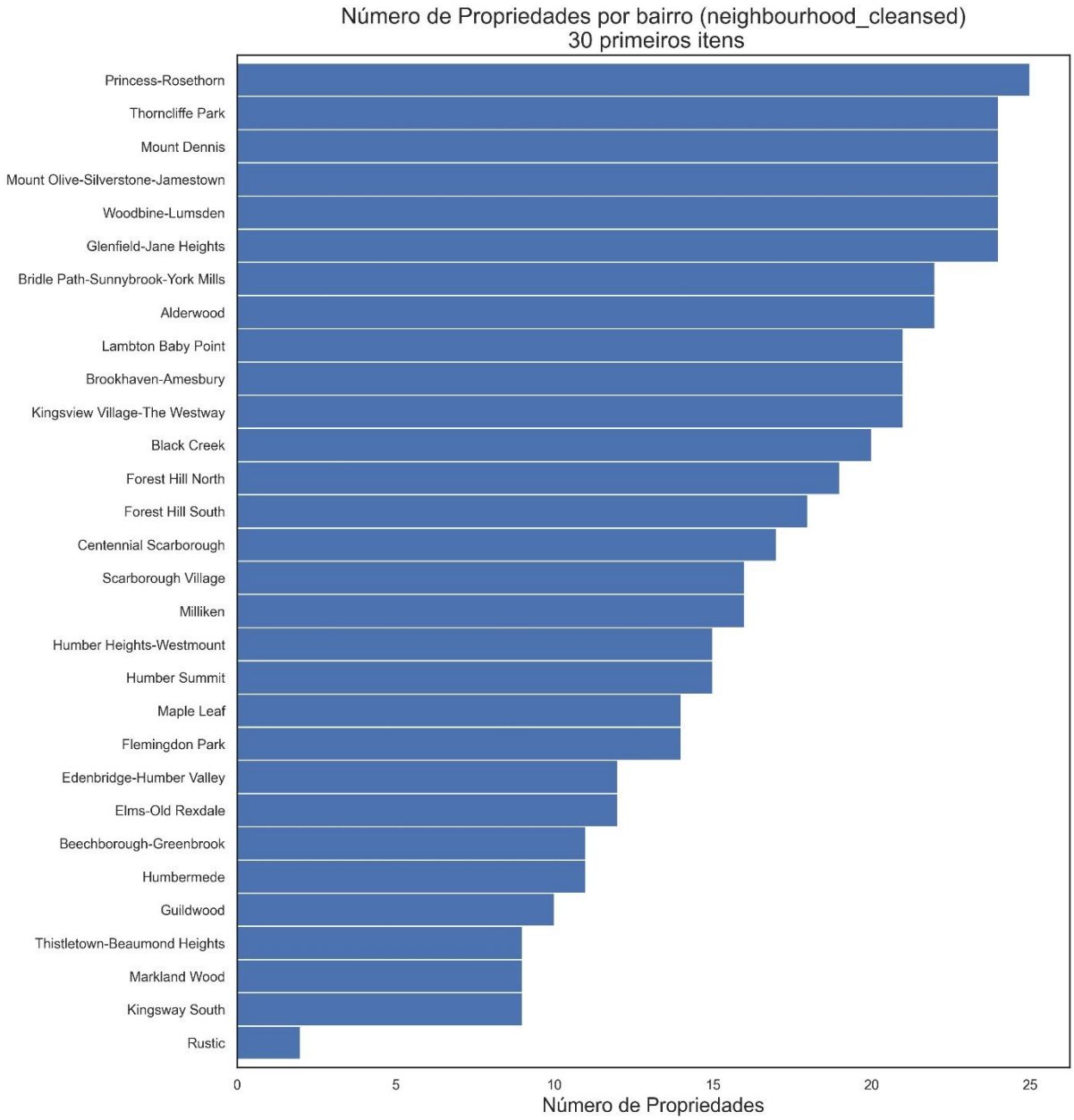
The space

This newly renovated, 100+ year old home has all the comfort and amenities you could want in one of Toronto's most popular and vibrant neighbourhoods. This is our home, and not a rental property, so all beds, furniture, kitchen ware etc has been chosen with care for style and comfort. The backyard features 2 decks with a small table for 4 perfect for breakfast in the sun, and a larger patio with a table that seats 6/8. The lovely covered front porch has great seating for end of the afternoon sunshine. The house is ideally suited for a family.

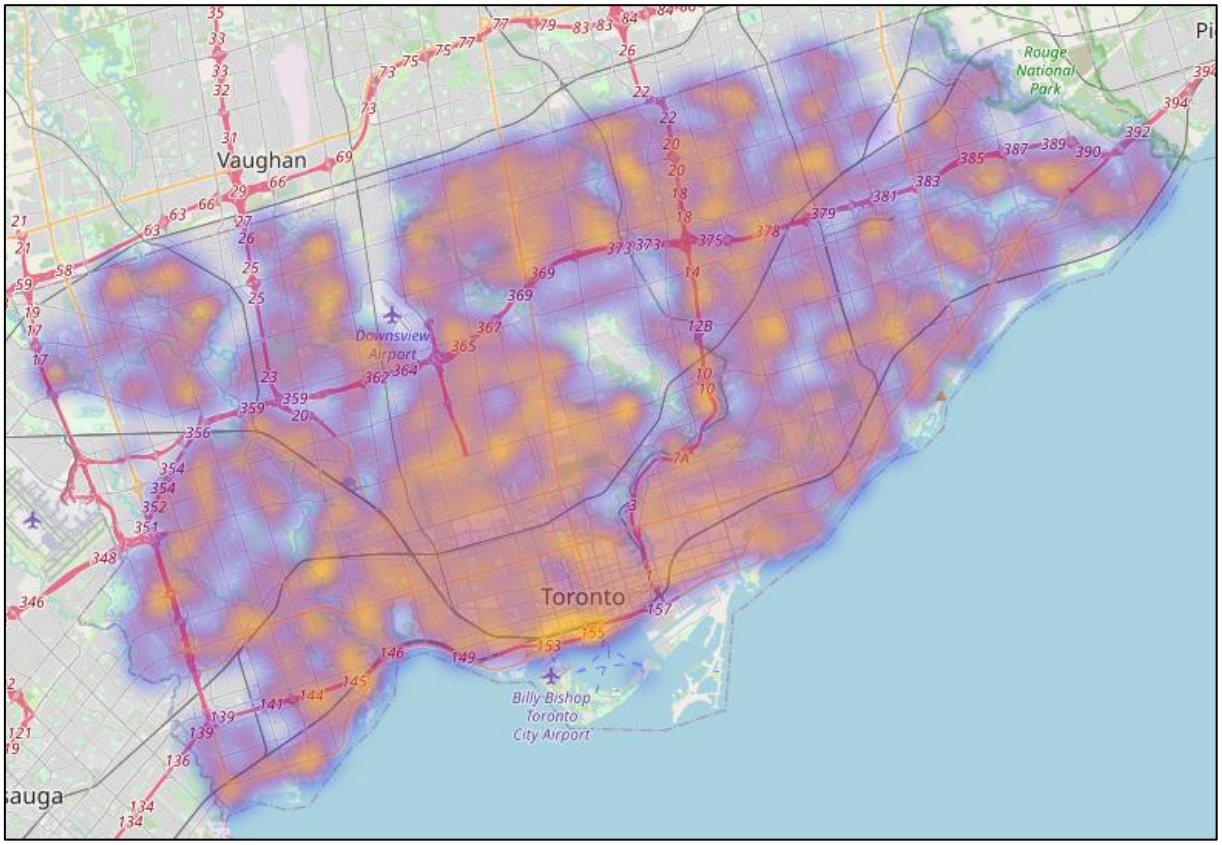
ID: 2979246

Anexo 6. Melhor descrição (polarity_score = +0.9979). Texto não alterado.

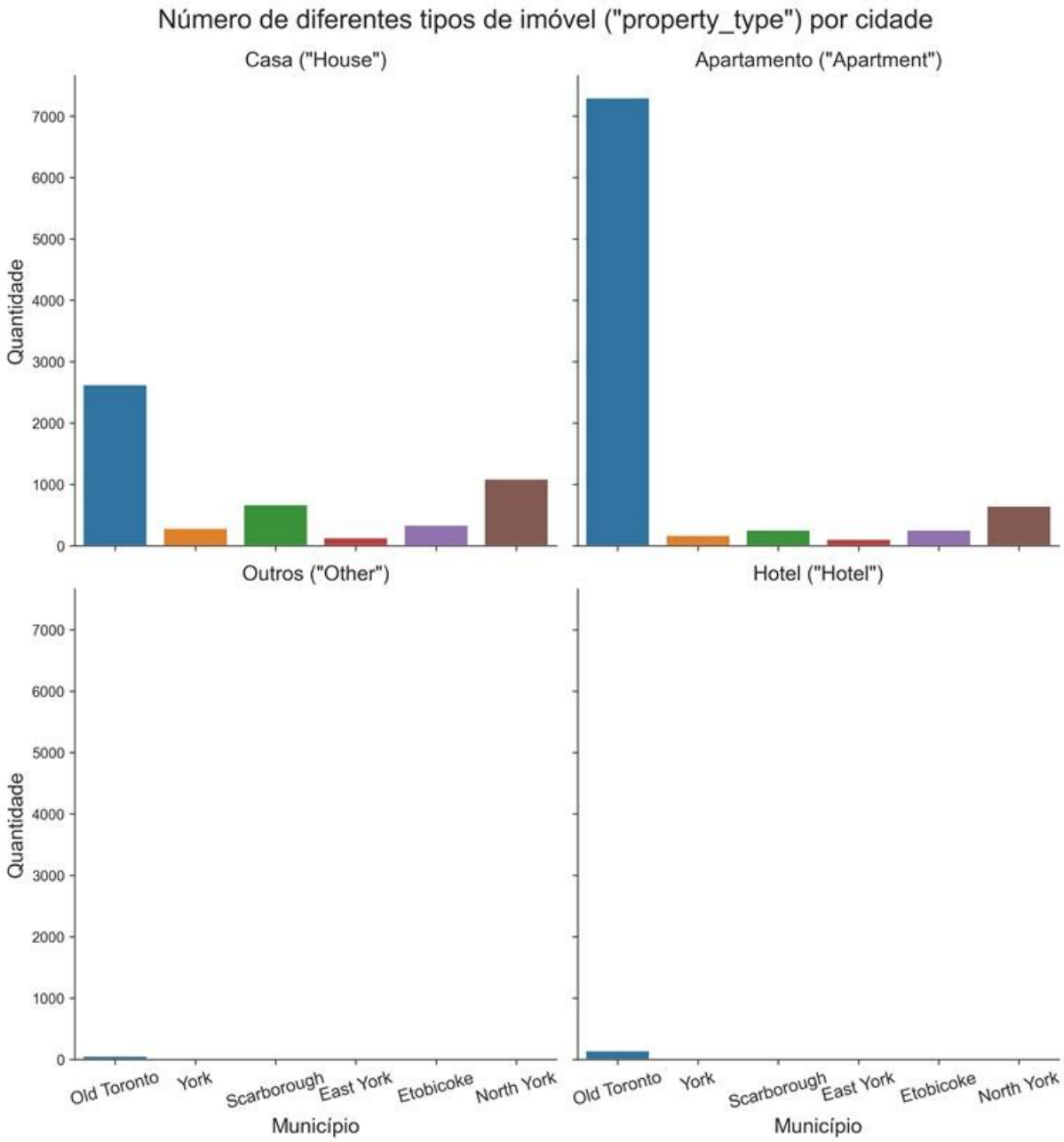
Anexo 7. Nuvem de palavras das palavras mais usadas nas avaliações positivas. Máscara (Formato) utilizado da CN Tower, atração de Toronto.



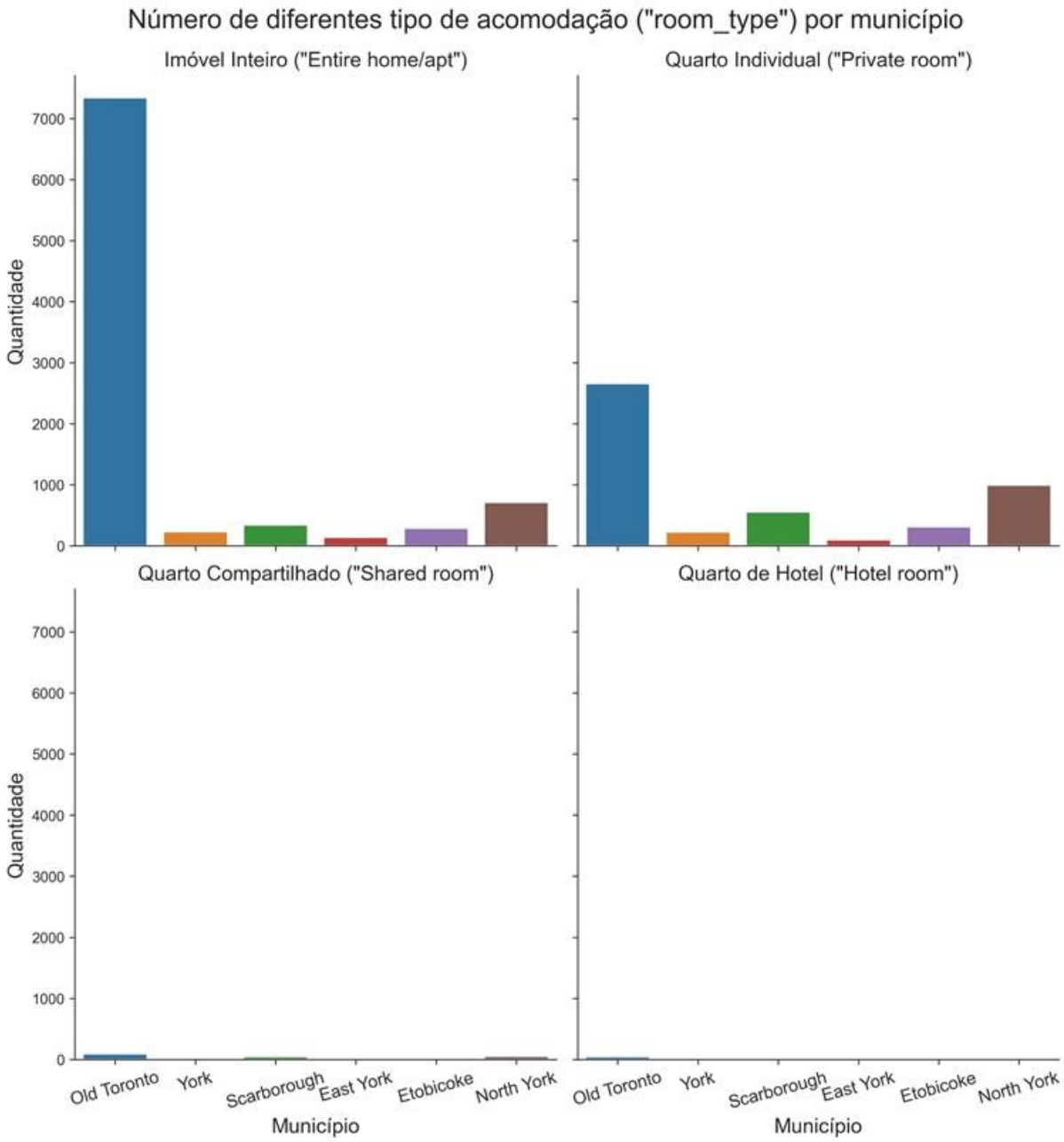
Anexo 8. Distribuição de propriedades por bairro (neighbourhood_cleansed).



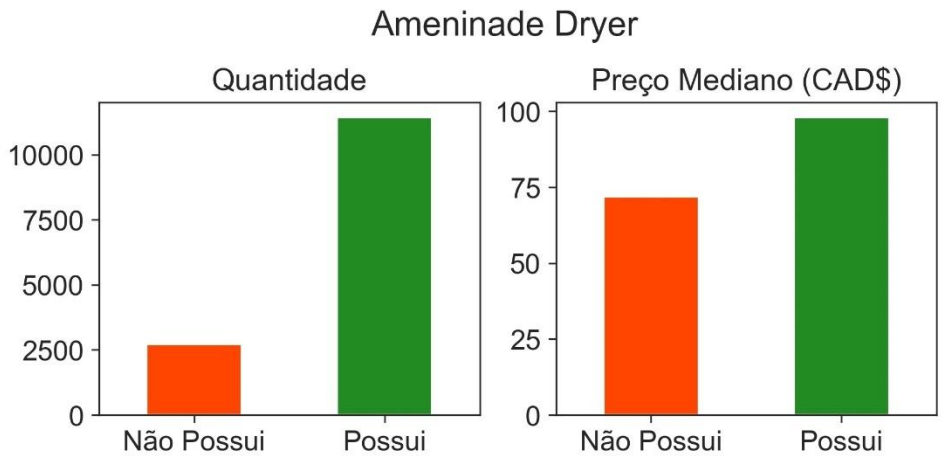
Anexo 9. Mapa de calor das propriedades listadas no dataframe.



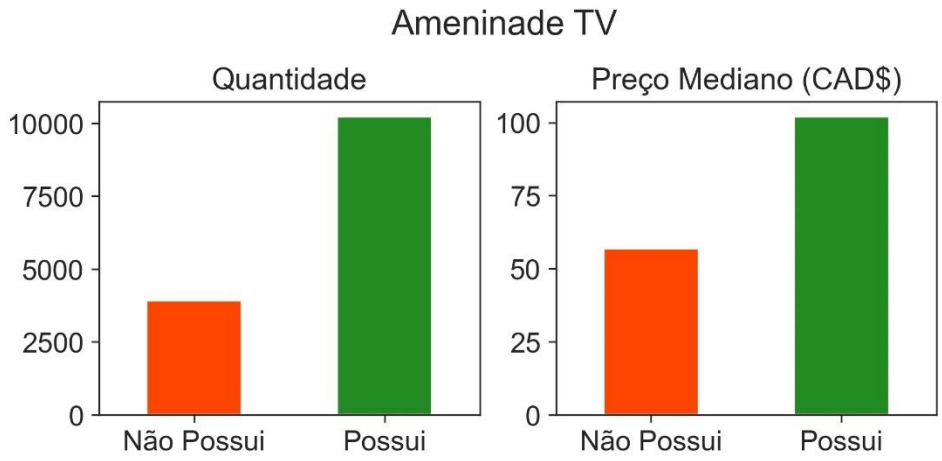
Anexo 10. Tipos de imóvel por município.



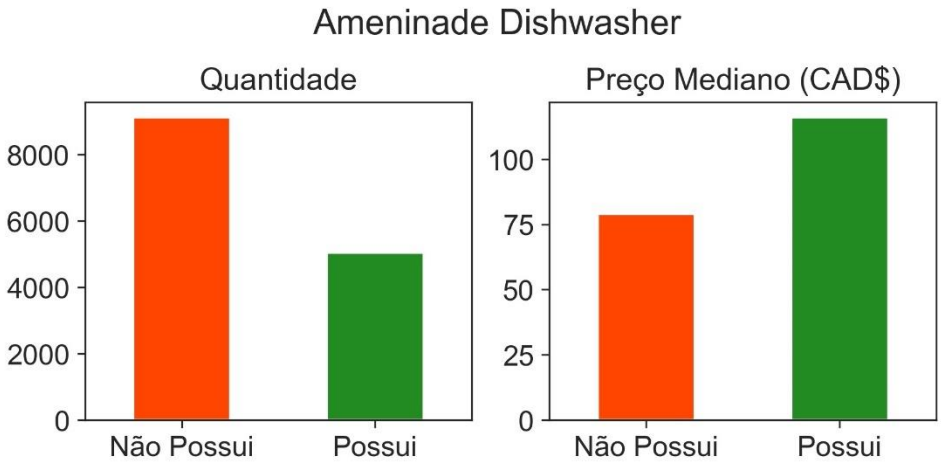
Anexo 11. Tipos de acomodação por município.



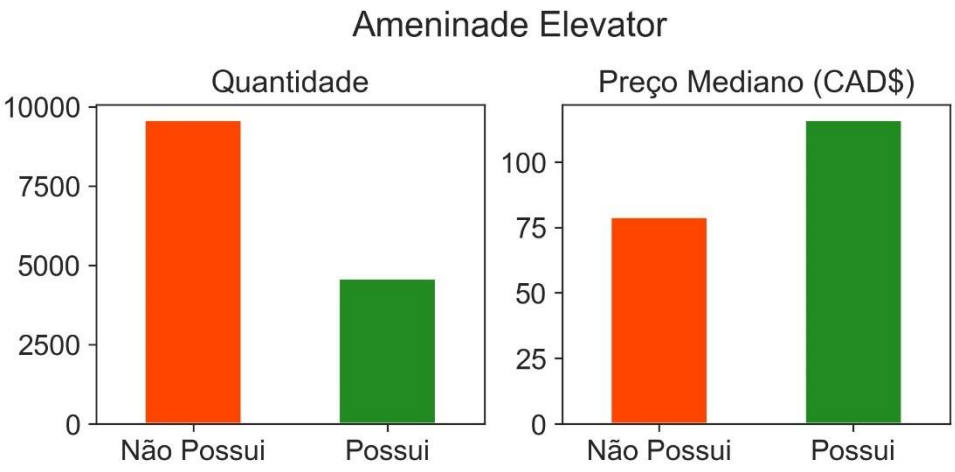
Anexo 12. Frequência e impacto no preço da secadora de roupas (Dryer).



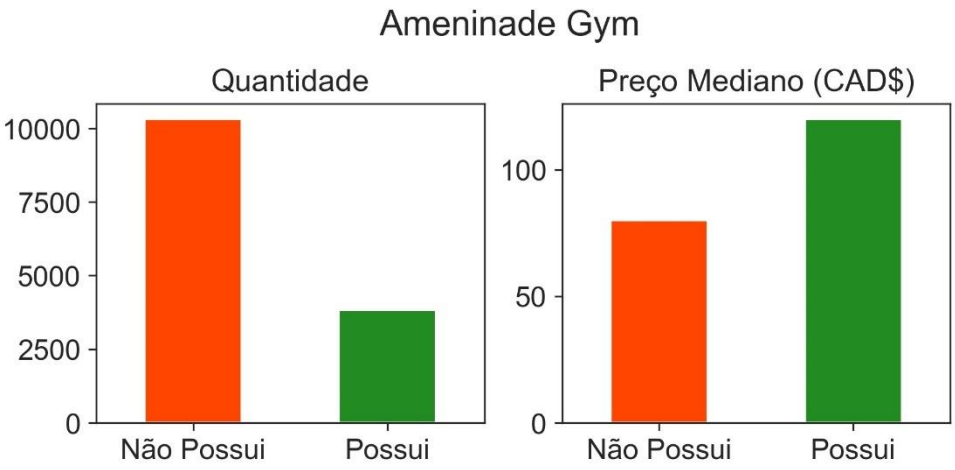
Anexo 13. Frequência e impacto no preço da televisão (TV).



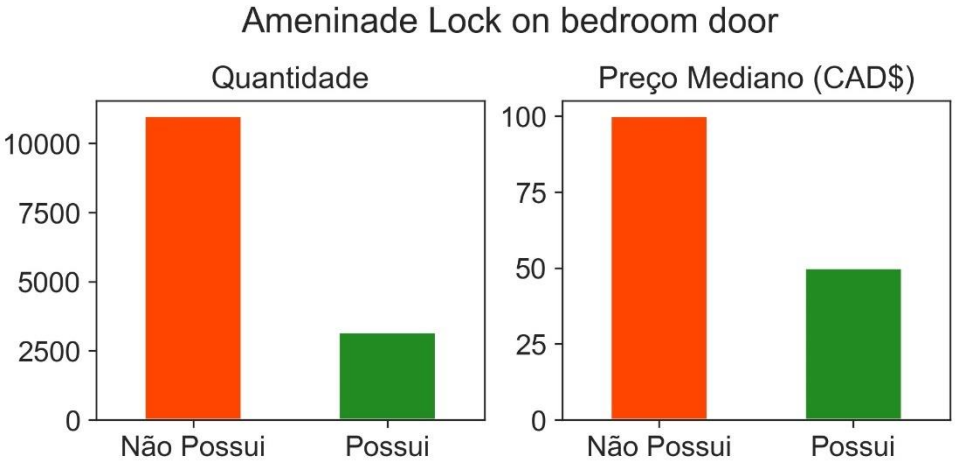
Anexo 14. Frequência e impacto no preço da lava-louças (Dishwasher).



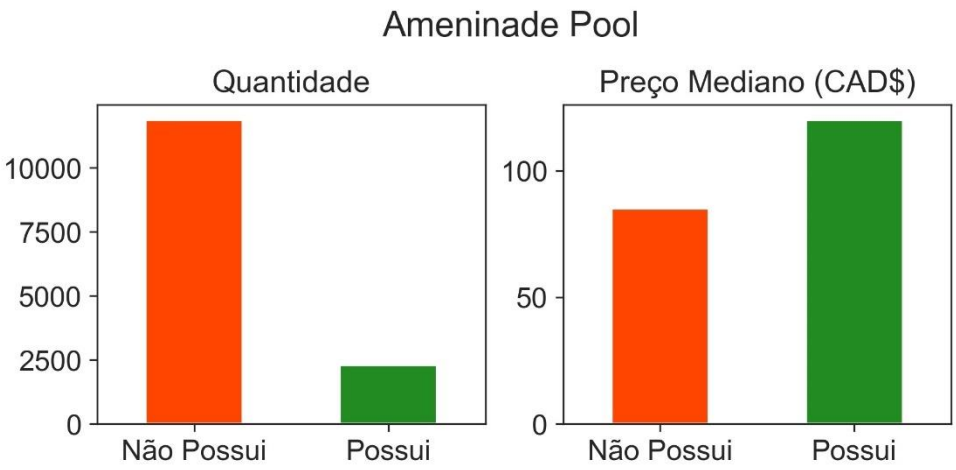
Anexo 15. Frequência e impacto no preço do Elevador (Elevator).



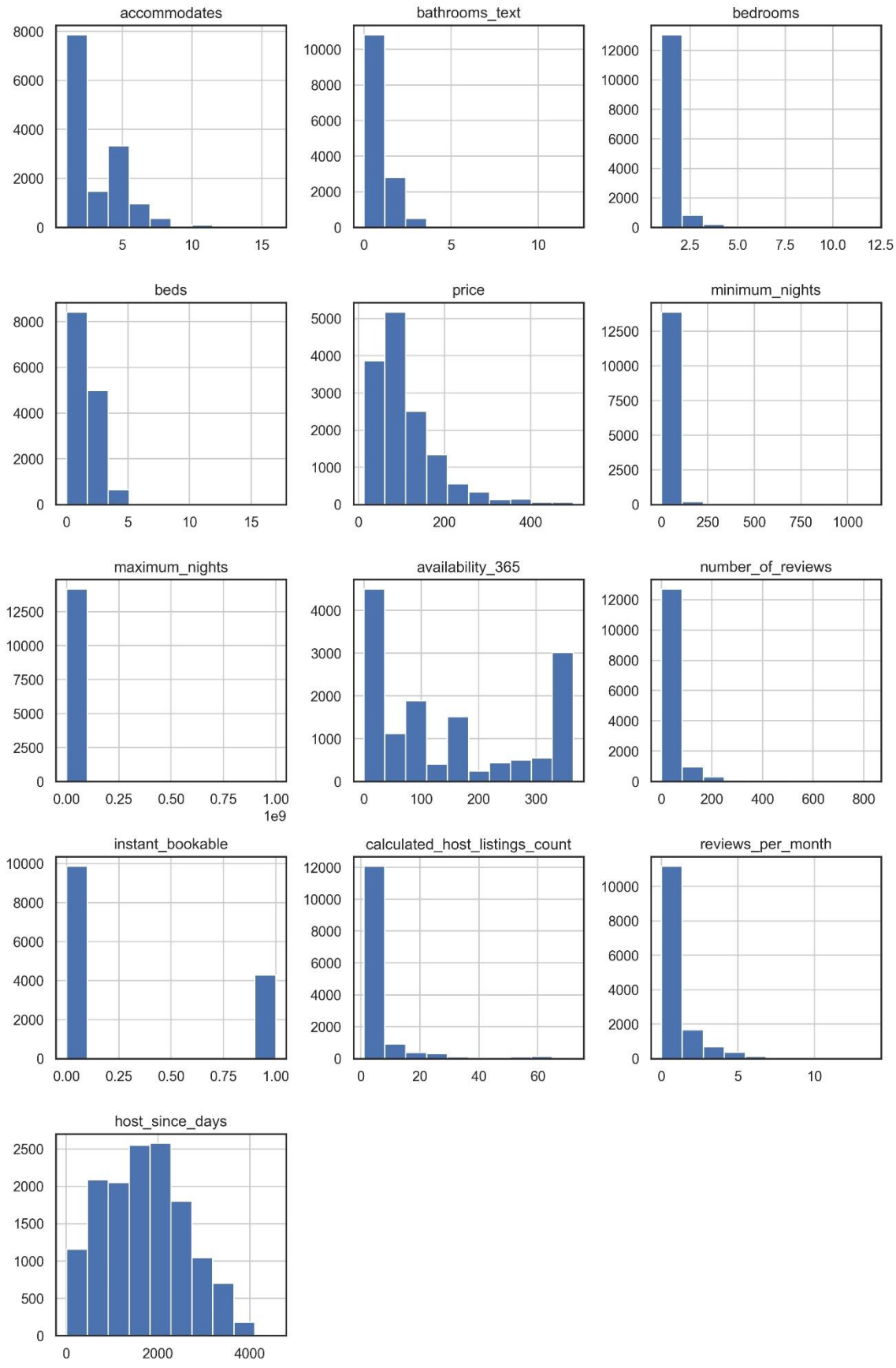
Anexo 16. Frequência e impacto no preço da Academia (Gym).



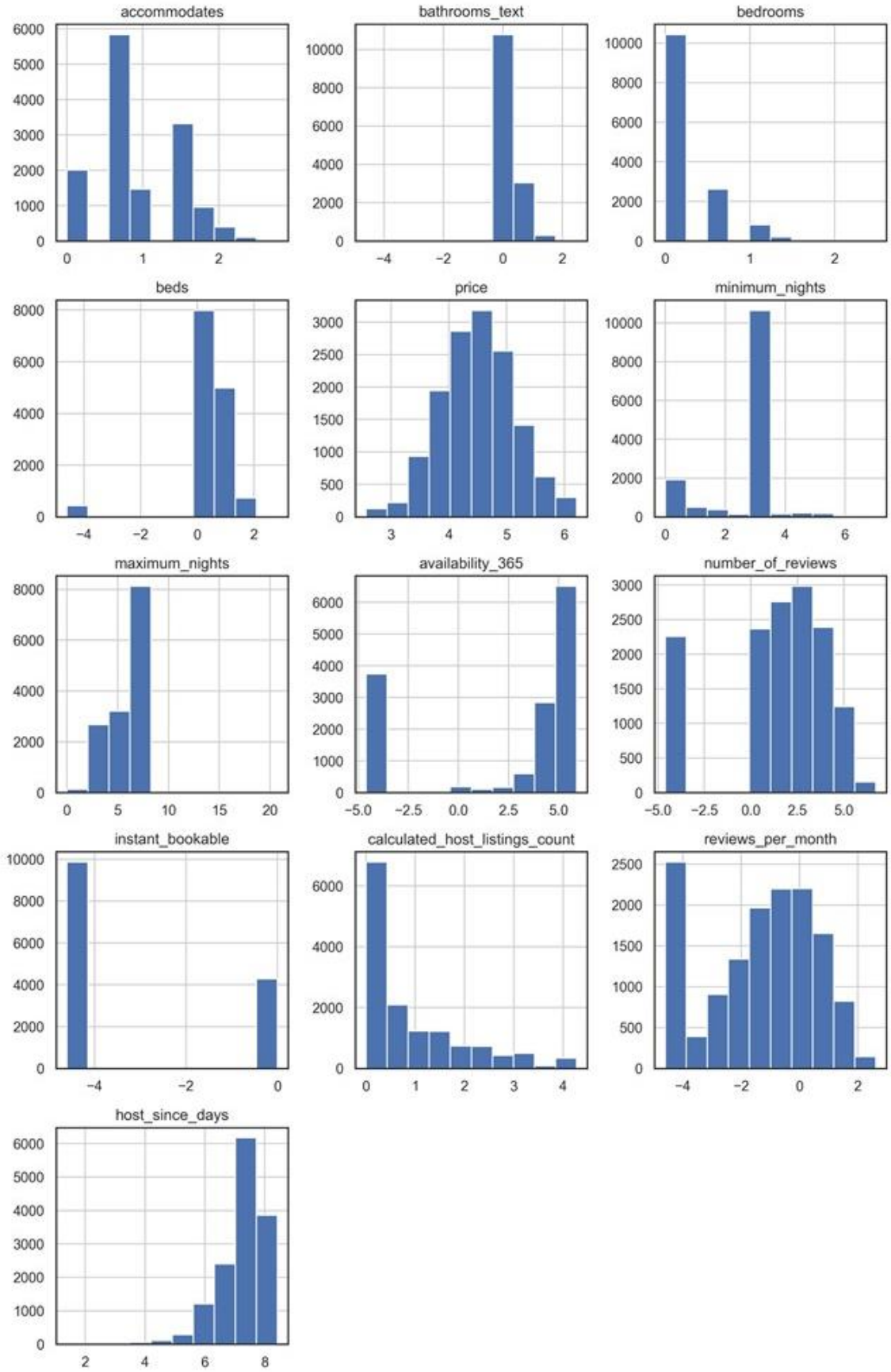
Anexo 17. Frequência e impacto ne tranca no quarto (lock on bedroom door).



Anexo 18. Frequência e impacto no preço da Piscina (Pool).



Anexo 19. Distribuição das colunas numéricas, antes da transformação.



Anexo 20. Distribuição das colunas numéricas, após da transformação.