**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**

**TÍTULO DO PROJETO**

Belo Horizonte

2021

**Josmar Baruffaldi Cristello**

**TÍTULO DO PROJETO**

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 **SUMÁRIO**

[1. Introdução 7](#_Toc70818297)

[1.1. Contextualização 7](#_Toc70818298)

[1.2. O problema proposto 7](#_Toc70818299)

[2. Coleta de Dados 7](#_Toc70818300)

[2.1. Fonte dos Dados 7](#_Toc70818301)

[2.2. Obtenção dos Dados 8](#_Toc70818302)

[2.3. Descrição dos Dados 8](#_Toc70818303)

[3. Processamento/Tratamento de Dados 11](#_Toc70818304)

[3.1. Conversão dos Dados 12](#_Toc70818305)

[3.2. Remoção de Colunas Não Utilizadas e dados invalidos 13](#_Toc70818306)

[3.3. Tratamento de Colunas Relacionadas ao Anfitrião 14](#_Toc70818307)

[3.4. Tratamento de Colunas Relacionadas aos Imóveis 17](#_Toc70818308)

[3.5. Tratamento de Colunas Relacionadas a Avaliações 21](#_Toc70818309)

[3.6. Tratamento da Coluna de interesse: price 23](#_Toc70818310)

[3.7. Processamento de Linguagem Natural (NLP): Avaliações e Descrições 24](#_Toc70818311)

[3.8. Processamento de Geolocalização Reversa: Latitude e Longitude 27](#_Toc70818312)

[3.9. Estado final do dataframe após Processamento/Tratamento de Dados 28](#_Toc70818313)

[4. Análise e Exploração dos Dados 28](#_Toc70818314)

[4.1. Análise e Exploração inicial dos dados 28](#_Toc70818315)

[4.2. Anfitriões, Quantidade de Propriedades, Atributos Associados 30](#_Toc70818316)

[4.3. Geolocalização - Municípios 33](#_Toc70818317)

[4.4. Geolocalização - Clusters 36](#_Toc70818318)

[4.5. Tipos de Propriedade e Acomodação 38](#_Toc70818319)

[4.6. Quantidade de Pessoas Acomodadas 41](#_Toc70818320)

[4.7. Amenidades 42](#_Toc70818321)

[5. Criação de Modelos de Machine Learning 43](#_Toc70818322)

[5.1. Preparação – Remoção de Colunas 43](#_Toc70818323)

[5.2. Preparação – Criação de Dummies 44](#_Toc70818324)

[5.3. Preparação – Correlação Entre as Variáveis 44](#_Toc70818325)

[5.4. Preparação – Transformação dos valores numéricos 46](#_Toc70818326)

[5.5. Preparação – Scaling, e Divisão dos dados 47](#_Toc70818327)

[5.6. Definição de Modelo e Métricas 47](#_Toc70818328)

[5.7. Aplicação dos Modelos 47](#_Toc70818329)

[5.8. Filtro dos Atributos mais Importantes 48](#_Toc70818330)

[5.9. Interpretação dos Atributos mais Importantes 49](#_Toc70818331)

[5.10. Re-Aplicação dos Modelos com Menos Atributos 49](#_Toc70818332)

[5.11. Seleção do Melhor Modelo 50](#_Toc70818333)

[6. Apresentação dos Resultados 50](#_Toc70818334)

[7. Links 51](#_Toc70818335)

[REFERÊNCIAS 52](#_Toc70818336)

[APÊNDICE 54](#_Toc70818337)

TABELA DE FIGURAS

[Figura 1. Colunas com 95% dos dados em uma só categoria. 16](#_Toc70818338)

[Figura 2. Distribuição das colunas do anfitrião (host...), após processamento. 16](#_Toc70818339)

[Figura 3. Comparação de todas as colunas que descrevem de disponibilidade. 18](#_Toc70818340)

[Figura 4. Histograma das colunas beds, bedrooms e bedrooms\_text. 20](#_Toc70818341)

[Figura 5. Histograma das colunas review\_scores\_ pré tratamento. 22](#_Toc70818342)

[Figura 6. Distribuição das colunas que descrevem o número de avaliações. 23](#_Toc70818343)

[Figura 7. Distribuição de densidade da coluna price, antes do processamento 24](#_Toc70818344)

[Figura 8. Distribuição de densidade da coluna price, após o processamento 24](#_Toc70818345)

[Figura 9. Distribuição da polaridade das descrições 26](#_Toc70818346)

[Figura 10. Distribuição da polaridade das avaliações. 26](#_Toc70818347)

[Figura 11. Novos anfitriões e primeiras avaliações, de 2010 a 2020. 29](#_Toc70818348)

[Figura 12. Distribuição de diária [CAD$] das propriedades listadas em Toronto. 30](#_Toc70818349)

[Figura 13. Os 10 Anfitriões com mais propriedades, por quantidade de prop. 30](#_Toc70818350)

[Figura 14. Os 10 Anfitriões com mais propriedades, por quantidade de prop. 31](#_Toc70818351)

[Figura 15. Quantidade de superhosts, e o efeito deles no preço da diaria. 32](#_Toc70818352)

[Figura 16. Quantidade de anfitriões verificados, e o efeito deles no preço da diaria. 33](#_Toc70818353)

[Figura 17. Municipios de Toronto [11]. 34](#_Toc70818354)

[Figura 18. Propriedades listadas no dataframe, agrupadas geograficamente. 35](#_Toc70818355)

[Figura 19. Dispersão de Latitude e Longitude, colorido por faixa de preço. 35](#_Toc70818356)

[Figura 20. Distribuição de preços por município, em Toronto. 36](#_Toc70818357)

[Figura 21. Dispersão de Latitude e Longitude, categorizado por ID do cluster 37](#_Toc70818358)

[Figura 22. Distribuição de preços por cluster, em Toronto. 37](#_Toc70818359)

[Figura 23. Tipo mais comum de imóvel e de acomodação. 38](#_Toc70818360)

[Figura 24. Boxplot do valor da diária [CAD$] por tipo de imóvel. 39](#_Toc70818361)

[Figura 25. Violinplot do valor da diária [CAD$] por tipo de acomodação. 40](#_Toc70818362)

[Figura 26. Número de imóveis imediatamente reserváveis, e o efeito deles no preço. 40](#_Toc70818363)

[Figura 27. Quantidade de propriedades por número de pessoas acomodadas. 41](#_Toc70818364)

[Figura 28. Preço pelo número de pessoas acomodadas, agrup. pela mediana. 41](#_Toc70818365)

[Figura 29. Mapa de Calor da Matriz de correlação – Parte 1. 45](#_Toc70818366)

[Figura 30. Mapa de Calor da Matriz de correlação – Parte 2. 46](#_Toc70818367)

[Figura 31. Atributos com mais de 0.5% de importância 48](#_Toc70818368)

[Tabela 1. Nome, Descrição e tipo dos dados no arquivo listings.csv. 8](#_Toc70818527)

[Tabela 2. Nome, Descrição e tipo dos dados no arquivo reviews.csv. 10](#_Toc70818528)

[Tabela 3. Primeiras linhas das colunas de datas. 12](#_Toc70818529)

[Tabela 4. Value\_count da coluna *host\_response\_time*. 13](#_Toc70818530)

[Tabela 5. Valores das colunas host\_resposne\_rate e host\_acceptance\_rate 16](#_Toc70818531)

[Tabela 6. Valores únicos da coluna property\_type após processamento 17](#_Toc70818532)

[Tabela 7. Valores únicos da coluna room\_type. 18](#_Toc70818533)

[Tabela 8. Descrição dos valores first\_review\_days, após processamento 20](#_Toc70818534)

[Tabela 9. Descrição dos valores last\_review\_days, após processamento 20](#_Toc70818535)

[Tabela 10. Frequência das colunas review\_scores\_ pós tratamento. 21](#_Toc70818536)

[Tabela 11. Distribuição do novo atributo, geo\_city 26](#_Toc70818537)

[Tabela 12. Distribuição do novo atributo, geo\_cluster 27](#_Toc70818538)

[Tabela 13. Diária por pessoa, por número de pessoas acomodadas. 41](#_Toc70818539)

[Tabela 14. Métricas do algorítimo Random Forest. 46](#_Toc70818540)

[Tabela 15. Métricas do algorítimo Gradient Boost. 47](#_Toc70818541)

[Tabela 16. Métricas do algorítimo Random Forest – Após Remoção. 48](#_Toc70818542)

[Tabela 17. Métricas do algorítimo Gradient Boost – Após Remoção. 49](#_Toc70818543)

[Anexo 1. Comparação das colunas que descrevem maximum\_nights e minimum\_nights 54](#_Toc70818826)

[Anexo 2. Distribuição das colunas review\_score\_ 55](#_Toc70818827)

[Anexo 3. Melhor avaliação (polarity\_score = +0.9996). Texto não alterado. 56](#_Toc70818828)

[Anexo 4. Segunda pior avaliação (polarity\_score = -0.9985). Texto não alterado. 57](#_Toc70818829)

[Anexo 5. Segunda pior descrição (polarity\_score = -0.9063) . Texto não alterado. 58](#_Toc70818830)

[Anexo 6. Melhor descrição (polarity\_score = +0.9979) . Texto não alterado. 58](#_Toc70818831)

[Anexo 7. Núvem de palavras das palavras mais usadas nas avaliações positivas. Máscara (Formato) utilizado da CN Tower, atração de Toronto. 59](#_Toc70818832)

[Anexo 8. Distribuição de propriedades por bairro (neighbourhood\_cleansed). 60](#_Toc70818833)

[Anexo 9. Mapa de calor das propriedades listadas no dataframe. 61](#_Toc70818834)

[Anexo 10. Tipos de imóvel por município. 62](#_Toc70818835)

[Anexo 11. Tipos de acomodação por município. 63](#_Toc70818836)

[Anexo 12. Frequência e impacto no preço da secadora de roupas (Dryer). 64](#_Toc70818837)

[Anexo 13. Frequência e impacto no preço da televisão(TV). 64](#_Toc70818838)

[Anexo 14. Frequência e impacto no preço da lava-louças (Dishwasher). 64](#_Toc70818839)

[Anexo 15. Frequência e impacto no preço do Elevador (Elevator). 65](#_Toc70818840)

[Anexo 16. Frequência e impacto no preço da Academia (Gym). 65](#_Toc70818841)

[Anexo 17. Frequência e impacto ne tranca no quarto (lock on bedroom door). 65](#_Toc70818842)

[Anexo 18. Frequência e impacto no preço da Piscina (Pool). 66](#_Toc70818843)

[Anexo 19. Distribuição das colunas numéricas, antes da transformação. 67](#_Toc70818844)

[Anexo 20. Distribuição das colunas numéricas, após da transformação. 68](#_Toc70818845)

# Introdução

## Contextualização

Nessa seção você deverá introduzir e contextualizar o seu problema de interesse.

## O problema proposto

Nessa seção você deverá descrever o problema que você está propondo resolver através dos dados. Uma dica interessante é utilizar a técnica dos [5-Ws](https://its.unl.edu/bestpractices/remember-5-ws). Responder algumas dessas perguntas te dará uma melhor visão do seu problema e da solução.

(Why?) Por que esse problema é importante?

(Who?) De quem são os dados analisados? De um governo? Um ministério ou secretaria? Dados de clientes?

(What?): Quais os objetivos com essa análise? O que iremos analisar?

(Where?): Trata dos aspectos geográficos e logísticos de sua análise.

(When?): Qual o período está sendo analisado? A última semana? Os últimos 6 meses? O ano passado?

# Coleta de Dados

## Fonte dos Dados

Foram utilizadas três bases de dados, de duas fontes diferentes. A primeira fonte foi referente aos dados de propriedades do (listing.csv) e de avaliações (review.csv). Ambos de propriedades localizadas em Toronto, Ontário e listados no Airbnb. Esses dados foram obtidos do site *InsideAirbnb* [1] no link <http://insideairbnb.com/get-the-data.html> [2].

O autor deste estudo considerou fazer obtenção dos dados direto no Airbnb (web-scraping), mas isso é contra os termos e serviço do Airbnb [3]. O InsideAirbnb fornece dados extraídos do Airbnb, sob a licensa CCO 1.0 (“public domain dedication”) [4].

A segunda fonte de dados foi a API (Application Programming Interface) de geolocalização reversa do site MapQuest [5]. Essa API foi utilizada para, através de pares de latitude e longitude, obter a cidade de cada anúncio.

O MapQuest fornece um plano gratuito que permite até 15 000 buscas por mês [6]. Utilizações dos dados para uso acadêmico não são restringidas pelos termos e serviço do site [7]. 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.

## 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 a cidade foi obtida através do script geolocation\_process\_coordinates.ipynb. Para estes dados, foram utilizadas as bibliotecas pandas, geocoder e glob. 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) do 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 (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.

## 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 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. Essa API retorna diversas colunas como o endereço exato da rua, código de CEP, dentre outros. Todavia, esse estudo estava exclusivamente interessado em obter a cidade 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 | 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 quando cadastrou na plataforma | 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 | Bool |
| host\_thumbnail\_url | URL da thumbnail da foto do anfitrião | String |
| host\_picture\_url | URL da imagem utilizada na 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 que a prop. acomoda | Int |
| bathrooms | Número de banheiros da propriedade | Float |
| bathrooms\_text | Descrição dos banheiros da propriedade | String |
| bedrooms | Número de quartos da propriedade | Int |
| beds | Número de camas da propriedade | Int |
| amenities | Lista de amenidades oferecidos | 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 | 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 privado | Int |
| calculated\_host\_listings\_count\_shared\_rooms | Propriedades que o anfitrião, do tipo shared room | 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 |

# Processamento/Tratamento de Dados

Nessa seção, foram utilizadas as bibliotecas pandas, numpy do Python 3.9.1 64-byte. O banco de dados inicial (listings.csv) tinha uma grande quantidade de colunas, mas boa parte delas não eram tratadas, tinham dados em brancos, inválidos ou que não contribuíam para a análise.

A estratégia nessa seção foi: Reduzir a redundância dos dados (Removendo colunas completamente iguais), remover colunas com grande quantidade de dados inválidos (>70%), e encontrar maneiras de lidar com colunas com grande quantidade de dados faltantes. Essas maneiras podem ser: Remoção das linhas faltantes, preenchimento por mediana ou média, e finalmente o tratamento da coluna como dado categórico.

Essa seção começa com a importação dos dados do arquivo que contém os anúncios de Toronto: listings.csv. O arquivo, inicialmente, possui 15 832 linhas e 73 colunas que serão tratadas. Essa seção vai descrever, primeiro, a execução do script Notebook.ipynb.

## Conversão dos Dados

Por padrão, a coluna que é a variável de interesse para o modelo, e armazena os valores de preço de diária para cada anúncio em Toronto (‘price’) é interpretada pelo como *string*, por que os valores foram armazenados com símbolo indicador da moeda. Exemplos: “$469.00”, “1,396.00”. Portanto, os valores são substituídos com remoção do símbolo da moeda ($) e remoção do indicador de milhar (,). Finalmente, os valores são convertidos para formato float.

As colunas *host\_response\_rate* e *host\_acceptance\_rate*, que são valores percentuais que são interpretados como *string*. Sendo assim, ambas as colunas têm o símbolo percentual (%) removido, e são convertidos para o formato float.

As colunas que deveriam ter formato booleano, são importadas com valores “t”, ou “f”. Em outras palavras, interpretadas pelo Python como strings. Assim, essas colunas têm 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, são convertidas para datas, no formato (%d-%m-%y)

Após a conversão das colunas de data, essas têm uma nova coluna criada, através da engenharia de atributos. A coluna host\_since descreve a data em que um anfitrião se cadastrou na plataforma, enquanto a coluna last\_scraped descreve a data em que os dados foram obtidos pela insideAirbnb.

Assim, uma nova coluna é criada, com o nome host\_since\_days que, em vez de armazenar a data em que o anfitrião se cadastrou na plataforma, descreve o número de dias desde que ele se cadastrou – e, portanto, é a subtração das colunas host\_since e last\_scraped. O resultado esse processo é exemplificado na Tabela 3. O mesmo procedimento é repetido para criar as colunas first\_review\_days e last\_review\_days.

Tabela 3. Primeiras linhas das colunas de datas.

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

## Remoção de Colunas Não Utilizadas e Dados Inválidos

Todas as colunas de texto (name, description, host\_about, ...) são removidas. Por motivos de eficiência, essas colunas serão tratadas separadas no script de NLP (Processamento de Linguagem Natural), e, portanto, são removidos do script atual.

As colunas de metadados, ou seja, colunas que armazenam dados da aquisição dos dados em si também são removidas (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'). Nenhuma dessas 17 colunas contribuíram diretamente com o modelo, ou com a interpretação dos dados.

Finalmente, são removidas um total de 1301 linhas que apresentam atributos de availability\_365 = 0 e number\_of\_reviews=0 simultaneamente. Elas representam propriedades que não tem nenhuma data disponível para serem alugadas no próximo ano, e nunca tiveram nenhuma avaliação de hóspedes. Essas propriedades, muito provavelmente, são utilizadas apenas para especulação imobiliária OU estão pausadas no Airbnb

## Colunas Relacionadas ao Anfitrião

Inicialmente, as propriedades que têm dados em branco de diversas colunas relacionados ao anfitrião são removidas. Essas colunas possuem data de cadastro do host em branco (*host\_since*), tão como resposta do host (*host\_response\_time* e *host\_response\_rate*) em branco. Como essas apresentam várias colunas em branco, o preenchimento com mediana ou média poderia acabar 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 em branco. No caso dessa coluna, cada valor em branco significa que o anfitrião ainda não recebeu nenhuma pergunta, e, portanto, não forneceu nenhuma resposta. Nesse caso, os valores ausentes são substituídos por ‘no response yet’. A distribuição dos valores é, então, disposta na Tabela 4.

Tabela 4. Value\_count 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), enquanto as outras duas são exatamente iguais e descrevem todas as propriedades, inclusive as em outras localizações (e.g. em outros países). Por isso, as últimas duas colunas são removidas, e apenas calculated\_host\_listings\_count é mantida.

As colunas calculadas: calculated\_host\_list\_count... \_entire homes, \_private\_rooms e \_shared\_rooms são todas removidas. Essas colunas não são necessárias, visto que elas são produzidas pelo InsideAirbnb e interpretam o tipo de propriedade, e isso será feito pelo script 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 no modelo, e são removidas. Após etapa inicial de processamento, as colunas do anfitrião têm sua distribuição descrita nos histogramas da Figura 2.

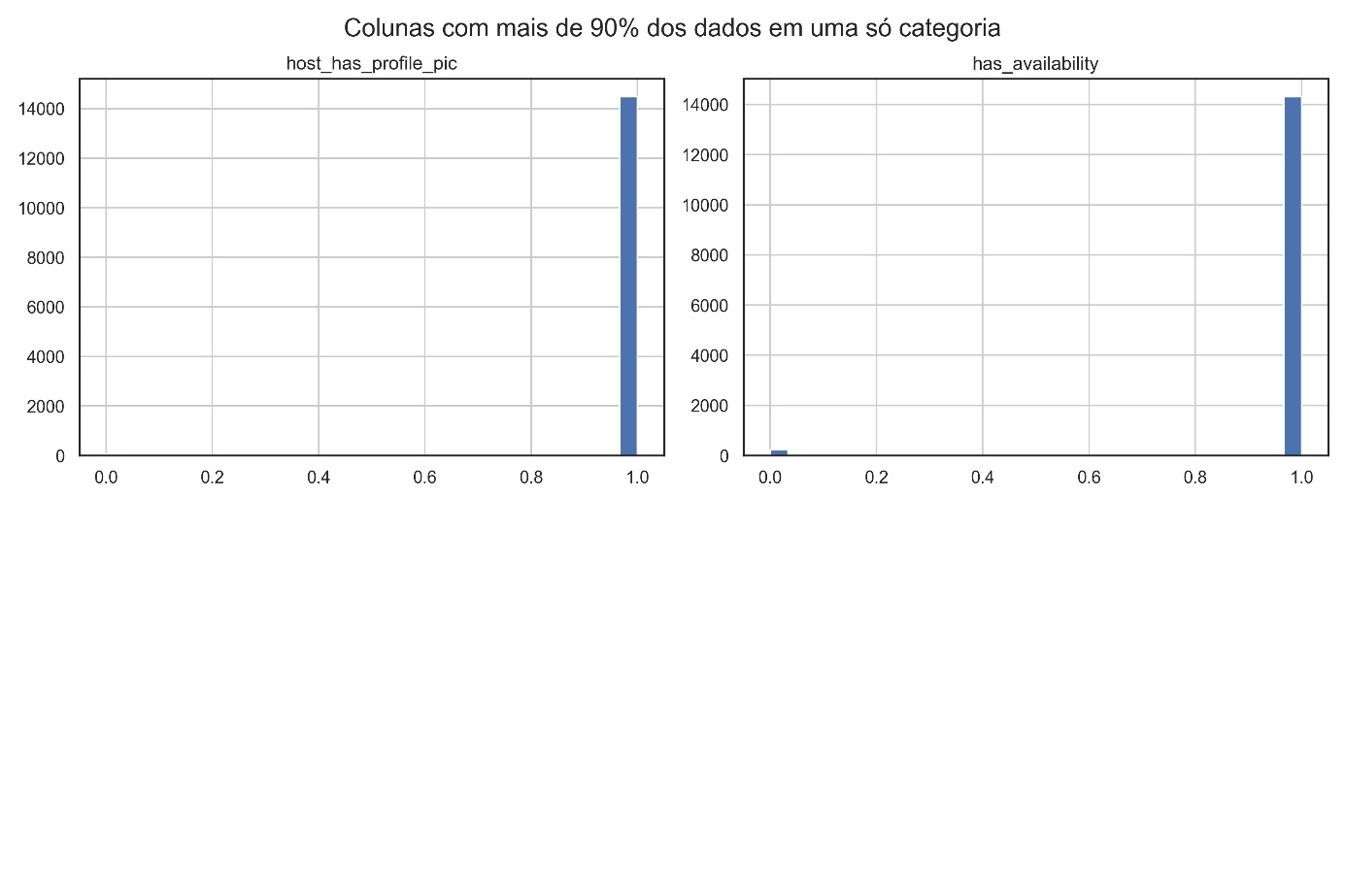


Figura 1. Colunas com 95% dos dados em uma só categoria.

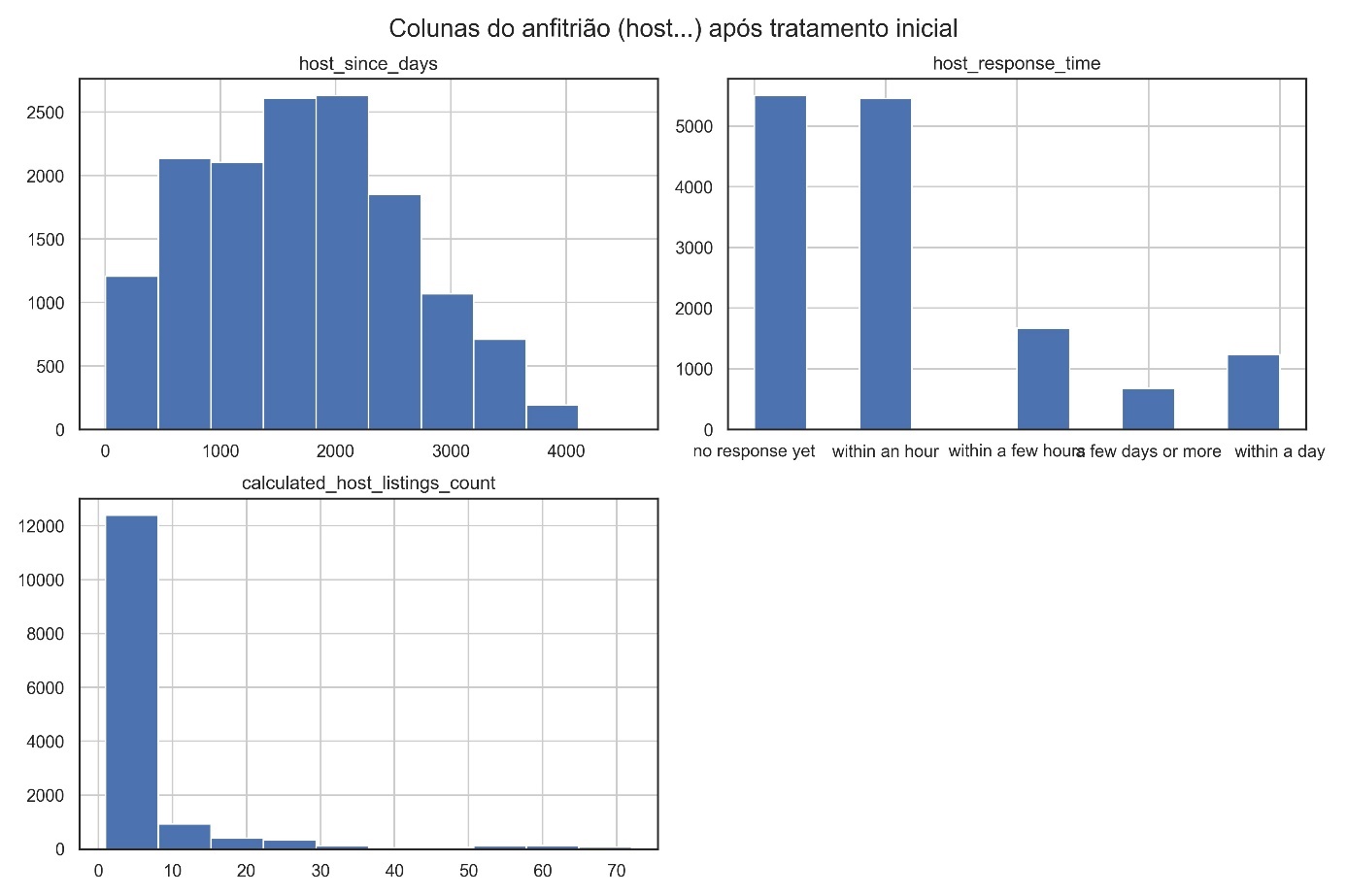


Figura 2. Distribuição das colunas do anfitrião (host...), após processamento.

As colunas host\_response\_rate e host\_acceptance\_rate apresentam um total de 5502 e 3993 linhas nulas, respectivamente. Entretanto, em ambos os casos, uma ausência de resposta significa que o anfitrião ainda não respondeu, ou que o hóspede não responde. Ambas as colunas são, por padrão, numéricas. Todavia, para que seus valores ausentes possam ser usados elas são agrupadas em bins e categorizadas. Os valores ausentes são, então, convertidos em “no response yet”, e as colunas pós conversão, são demonstradas na Tabela 5.

Tabela 5. Valores das colunas host\_response\_rate e host\_acceptance\_rate

|  |  |  |  |
| --- | --- | --- | --- |
| **host\_response\_rate** | | **host\_acceptance\_rate** | |
| **bin** | **quantidade** | **bin** | **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 colunas são: neighbourhood\_group\_cleansed, bathrooms e calendar\_updated, todas com 100% dos valores ausentes, e provavelmente representam colunas importadas incorretamente pelo insideAirbnb.com. A coluna license, que representa as licensas do imóvel, também é removida, visto que apresenta 74% dos dados ausentes.

## Colunas Relacionadas aos Imóveis

Existem diversas colunas que descrevem o mínimo e máximo número de noites disponíveis para aluguel. Essas colunas são descritas na Tabela 1, e são fortemente redundantes e correlacionadas. Devido a isso, as colunas minimum\_nights e maximum\_nights são mantidas, enquanto as minimum\_minimum\_nights, maximum\_minimum\_nights, minimum\_nights\_avg\_tm, minimum\_maximum\_nights, maximum\_maximum\_nights e maximum\_nights\_avg\_tm são removidas. Como são muitas colunas, a semelhança entre essas colunas é demonstrada no graficamente no fim desse estudo, no Anexo 1.

Similarmente, 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 são removidas, enquanto a coluna availability\_365 é mantida. A comparação gráfica entre essas colunas é demonstrada na Figura 3.

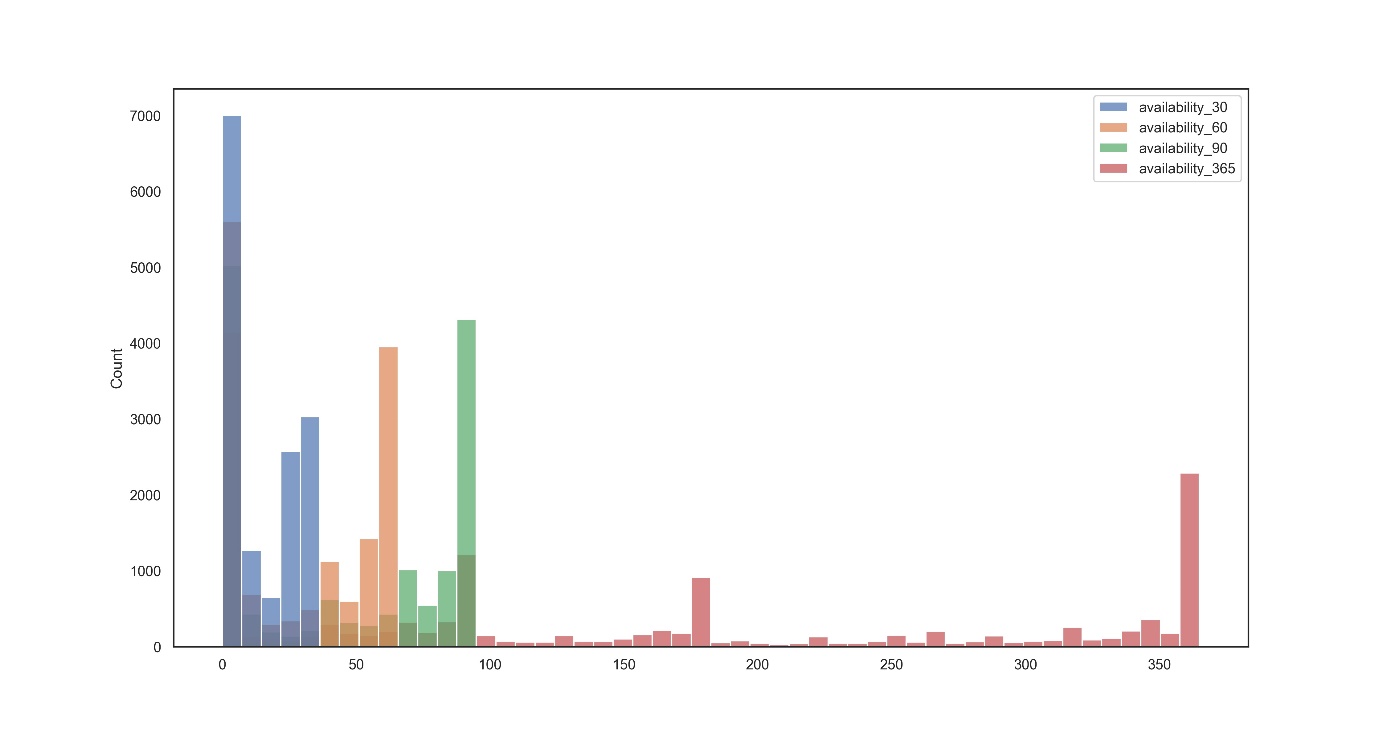


Figura 3. Comparação de todas as colunas que descrevem de disponibilidade.

A coluna property\_type descreve o tipo de imóvel, e é digitada pelo próprio anfitrião. Essa coluna tem 61 valores únicos que segmentam demais a quantidade dos dados. Por isso, ela é manualmente agrupada, e cada um dos 61 valores é convertido em quatro possíveis categorias: ‘House’, ‘Apartment’ e ‘Hotel’. Se o caso não se aplicar em nenhum desses três, é categorizado como “Other”.

Como exemplos: “Private room in house”, “Private room in townhouse” e “Entire townhouse” todos são categorizados como “House”. Similarmente, “Private room in apartment” e “Private room in guest suite” são 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** | **Fração do total** |
| Apartment | 8960 | 61.7 |
| House | 5330 | 36.7 |
| Hotel | 161 | 1.1 |
| Other | 80 | 0.6 |

A coluna room\_type não precisa ser processada, mas é demonstrada aqui para exemplificar como nenhuma informação é perdida com a conversão anterior. A coluna 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** | **Fração do 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 também vem em formato de texto (1 bath, 2.5 baths, shared half-bath). Aqui a conversão é feita em duas etapas. Primeiro, todas as instâncias de half bathroom são convertidas no número 0.5. Depois, todo o texto é removido, e os números são extraídos, transformando a coluna em numérica, e representando o número de banheiros na propriedade. Uma observação relevante é que “half bathroom” ou 0.5 banheiros são, conforme o Airbnb, definidos como um banheiro que possui apenas privada e pia, sem chuveiro [8].

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 são preenchidos com a mediana dos dados. Finalmente, a distribuição dos dados dessas colunas após processamento é demonstrada na Figura 4.

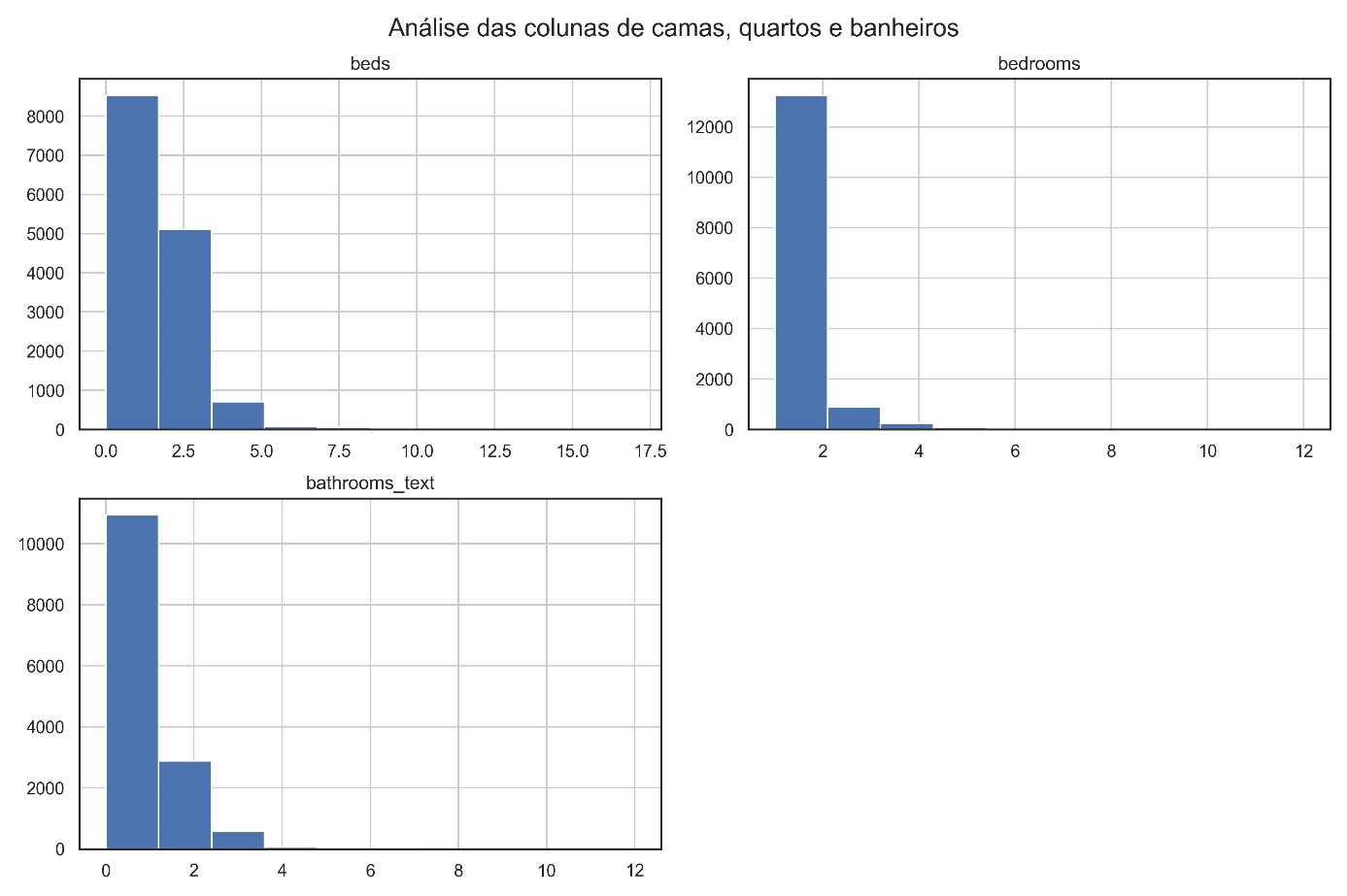


Figura 4. Histograma das colunas beds, bedrooms e bedrooms\_text.

A coluna que descreve as amenidades de cada propriedade apresenta um desafio. Essa coluna armazena uma lista com as diferentes amenidades por cada propriedade, que 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 são, primeiro, separadas em cada uma das amenidades. Essas têm seu texto tratado, e os artefatos de obtenção dos dados são removidos, e finalmente elas são contabilizadas por frequência. Depois, as primeiras 43 amenidades (Ou seja, amenidades que aparecem pelo menos 2200 vezes) são selecionadas. As amenidades selecionadas têm então, uma coluna booleana criada para cada uma delas.

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 essa propriedade do exemplo não tem nenhuma outra das 41 amenidades, todas essas outras colunas passam a ter valor 0.

## Colunas Relacionadas a Avaliações

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

Portanto, para evitar perder informações, ambas as colunas são 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

|  |  |  |
| --- | --- | --- |
| **first\_review\_days** | **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

|  |  |  |
| --- | --- | --- |
| **last\_review\_days** | **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 de uma propriedade em particular. A primeira coluna, review\_score\_rating é simplesmente o somatório das outras 6 colunas, então ela será removida para evitar problemas de multicolinearidade na construção do modelo. As próximas seis colunas descrevem as avaliações com respeito a diversos parâmetros, e serão mantidas e tratadas a seguir.

Todas essas colunas são numéricas, mas tem um total de 2550 valores ausentes. Como são muitas colunas, a distribuição de apenas duas colunas é demonstrada nas Figura 5, e todas as 6 colunas são demonstradas no Anexo 1. Novamente, um valor ausente nesse contexto significa um imóvel que ainda não foi avaliado, e não um imóvel com avaliação ruim. Portanto, vamos categorizar esses valores numericamente 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 (Categ.)** | **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 todas descrevem o número das avaliações do imóvel em diferentes períodos de tempo, e, portanto, são extremamente correlacionadas. Essa correlação é demonstrada na Figura 6. Para evitar problemas de colinearidade, as colunas number\_of\_reviews\_ltm e number\_of\_reviews\_l30d são removidas, e as outras duas são mantidas.



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

A coluna number\_of\_reviews não apresenta valores ausentes, mas a reviews\_per\_month apresenta 2361 valores ausentes. Portanto, todos esses serão preenchidos com o número “0”, para indicar que esses anúncios ainda não foram avaliados por nenhum hóspede.

## 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 investigar individualmente, muito provavelmente são anúncios que o anfitrião coloca um valor elevado temporariamente para não precisar o anúncio, mas também não ser alugado no período. Como isso não contribuirá na análise, todos esses 279 anúncios serão removidos. A nova distribuição é mostrada na Figura 8.

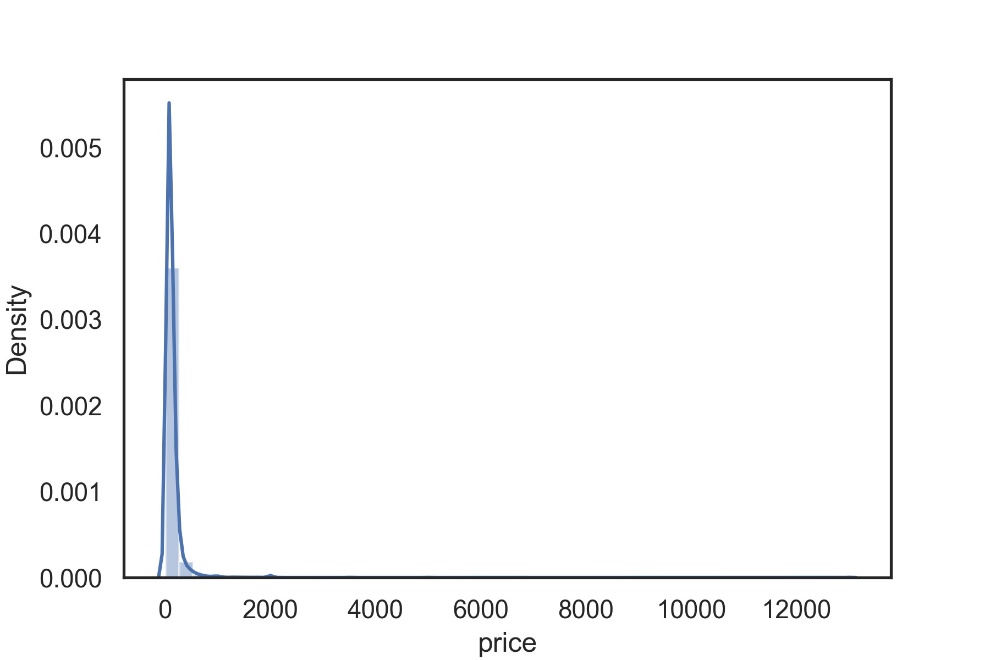


Figura 7. Distribuição de densidade da coluna price, antes do processamento

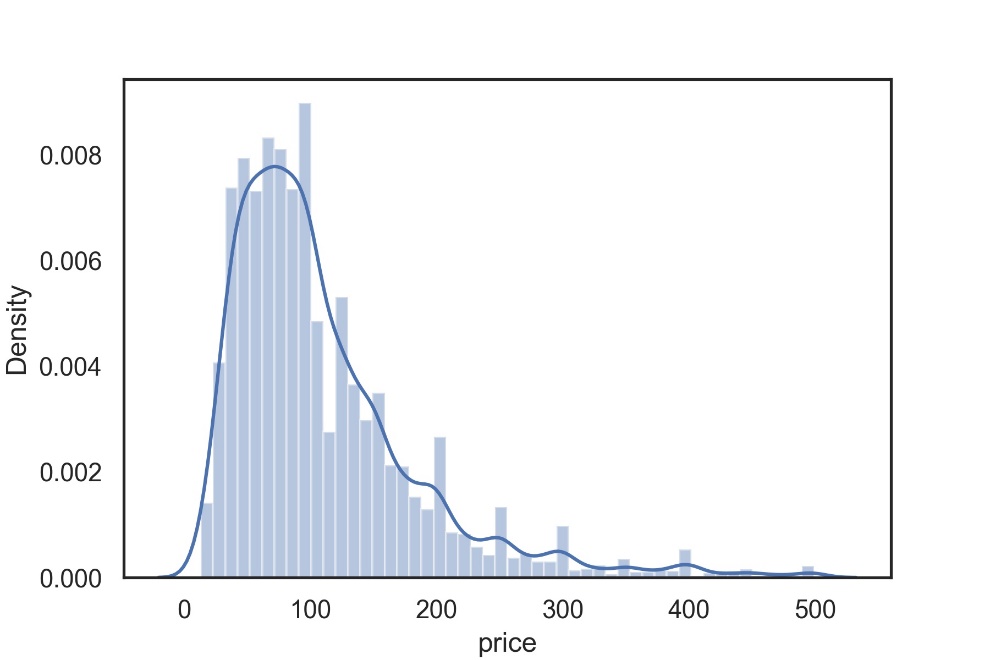


Figura 8. Distribuição de densidade da coluna price, após o processamento

É particularmente interessante notar uma densidade maior em números inteiros (100, 200, 250, 300, 350, 400), demonstrando o efeito da psicologia na hora da definição do preço

## 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 o processamento das avaliações e descrições, respectivamente.

As avaliações estão armazenadas no arquivo reviews.csv e são feitas pelos hóspedes, muitas vezes propriedades podem ter múltiplas 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 muito valiosas, mas também representam um grande número de dados. São um total de 424.060 avaliações, com 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), visto que ele é treinado especificamente para redes sociais [9].

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

Para as avaliações, foi criada a coluna review\_polarity e para as descrições, description\_polarity. Essa polaridade não contém informação sobre o texto ser negativo, neutro, ou positivo, e sim uma nota de -1 a +1 com a polaridade daquela frase. A distribuição da polaridade das avaliações e descrições são demonstradas nas Figura 9 e Figura 10.

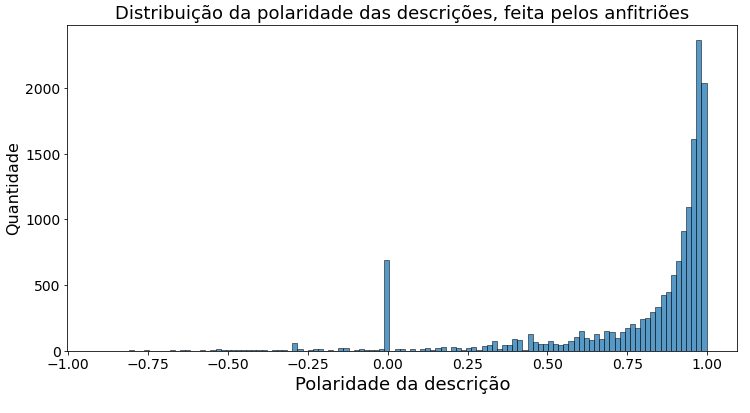


Figura 9. Distribuição da polaridade das descrições

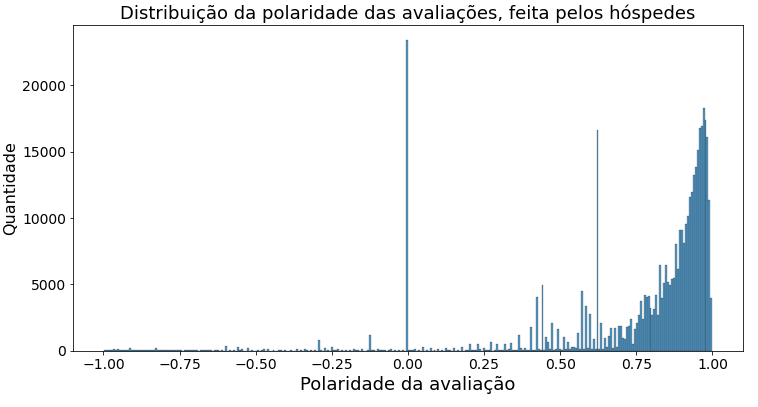


Figura 10. Distribuição da polaridade das avaliações.

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: Visto que, 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 são acompanham o código ID, para permitir que o leitor possa rastrear dados relacionados nos bancos de dados. Uma 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).

## 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. Essa coluna, por ser muito segmentada, acaba não sendo útil para a construção do modelo. A distribuição dessa coluna é demonstrada no Anexo 8.

Portanto, para que ainda fosse possível utilizar as informações de geolocalização de alguma forma, duas estratégias foram adotadas. A primeira delas foi utilizar geolocalização reversa, ou seja, utilizar a latitude e longitude para buscar o endereço equivalente, e depois extrair o nome do município desse endereço. 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, ou seja, cada anúncio, foi buscada um município. Esse, então, foi armazenada na coluna geo\_city e adicionada ao dataframe principal. A distribuição de anúncios por municípios pode ser vista 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. 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 ainda não perfeitamente homogêneo (visto que o número de propriedades também não é homogêneo), é mais balanceado do que a coluna geo\_city.

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 |

## Estado final do dataframe após Processamento/Tratamento de Dados

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.

# Análise e Exploração dos Dados

## Análise e Exploração inicial dos dados

Em Toronto, são listados um total de 14.150 propriedades, distribuídos em 6 cidades. 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$110.2 para conseguir uma propriedade na região de Toronto. Isso ainda é mais barato que um hotel que, de acordo com budgetyourtrip.com [10], custa uma média CAD$ 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 constante crescimento. Também é possível ver que 2020 foi um ano atípico, com uma grande queda tanto no número de cadastros novos de anfitriões e primeiras avaliações, mas que demonstrou sinais de melhora na segunda metade de 2020.

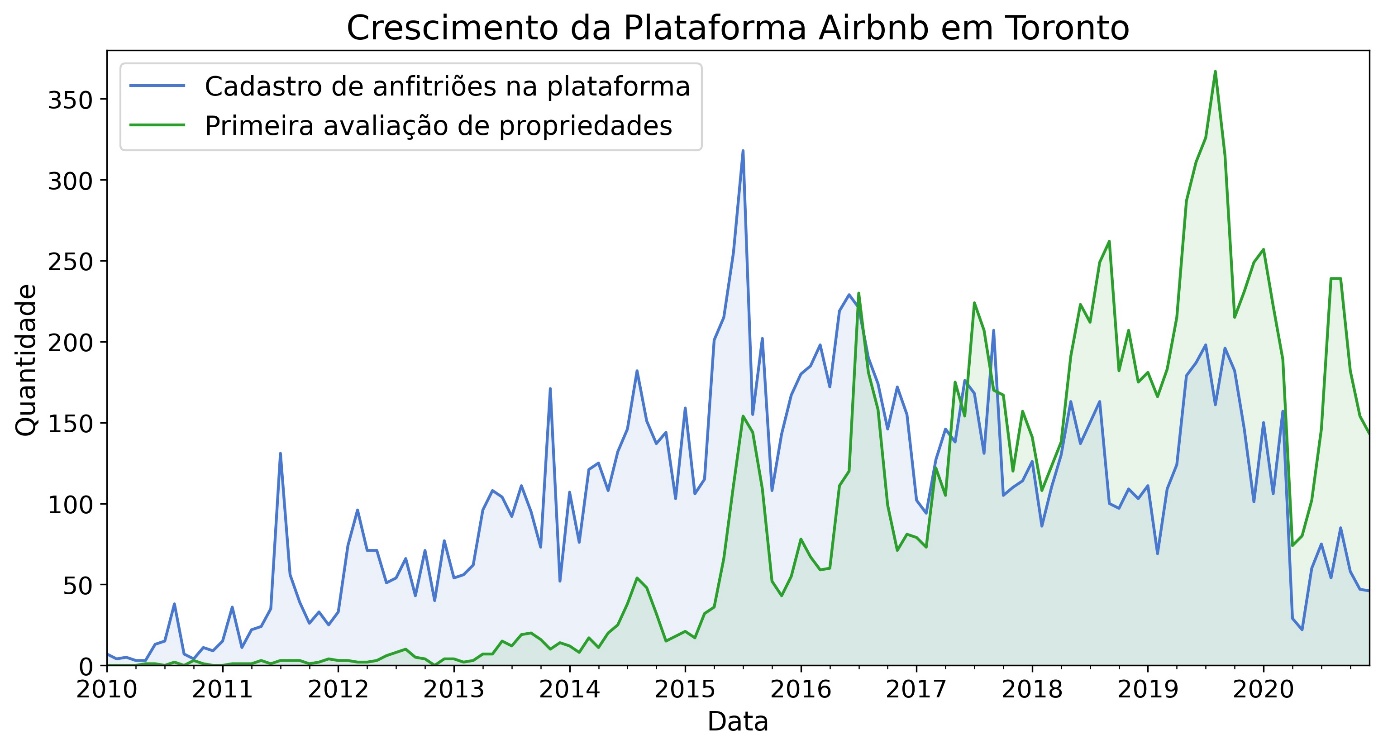


Figura 11. Novos anfitriões e primeiras avaliações, de 2010 a 2020.

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. Essa distribuição pode ser vista na Figura 12.

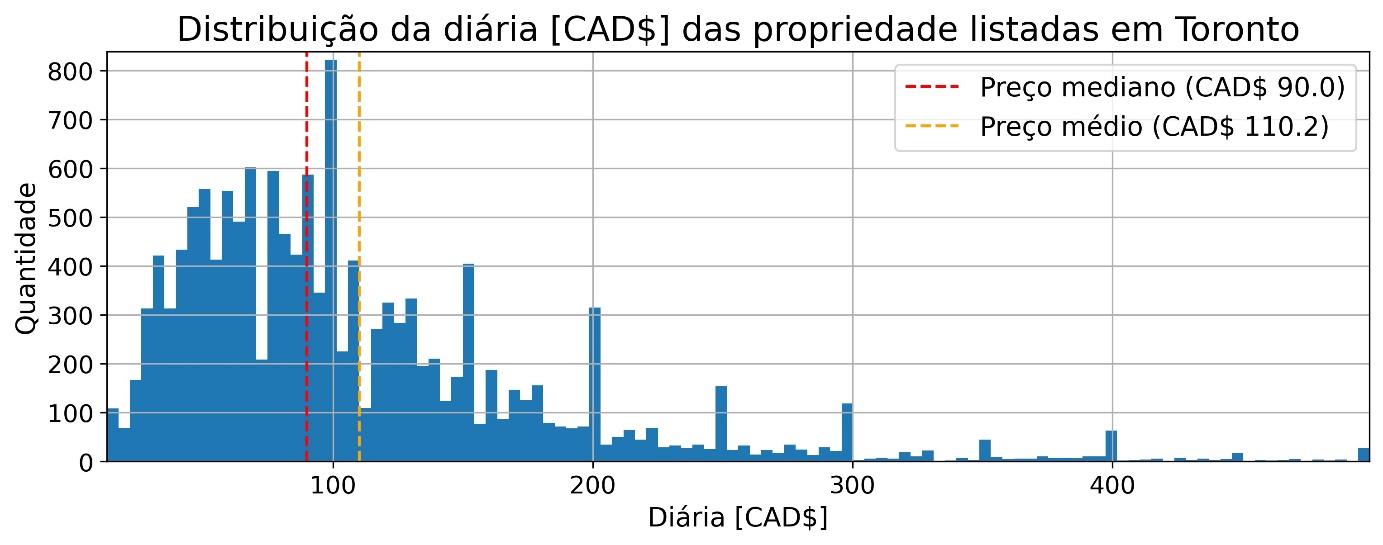


Figura 12. Distribuição de diária [CAD$] das propriedades listadas em Toronto.

## 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.

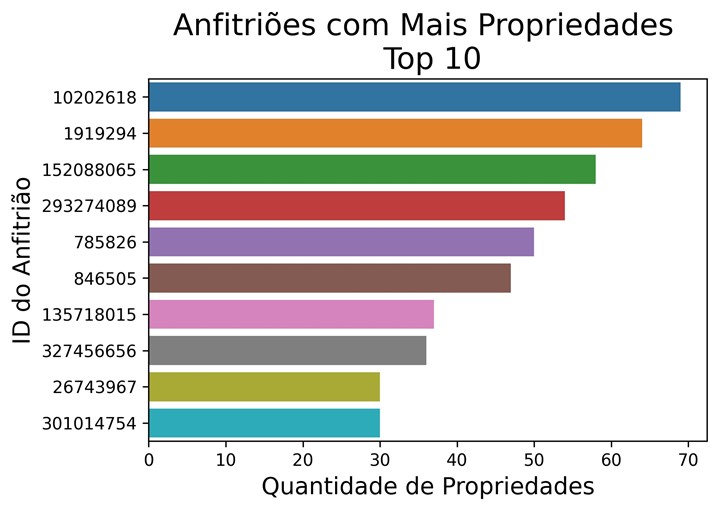


Figura 13. Os 10 Anfitriões com mais propriedades, por quantidade de prop.

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 indica que eles ou tem propriedades melhores (e, portanto, preços maiores), ou sabem usar melhor o sistema para obter um preço mais vantajoso.



Figura 14. Os 10 Anfitriões com mais propriedades, por quantidade de prop.

Além da quantidade de propriedades, os anfitriões têm duas propriedades booleanas: 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:

* 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 preço da diária é pouquíssimo impactado por esse atributo. Isso sugere que, mesmo se um hóspede estiver tentando minimizar o preço pago pela diária, vale a pena ele procurar anúncios feitos por superhosts, para garantir um maior conforto e melhor experiência.

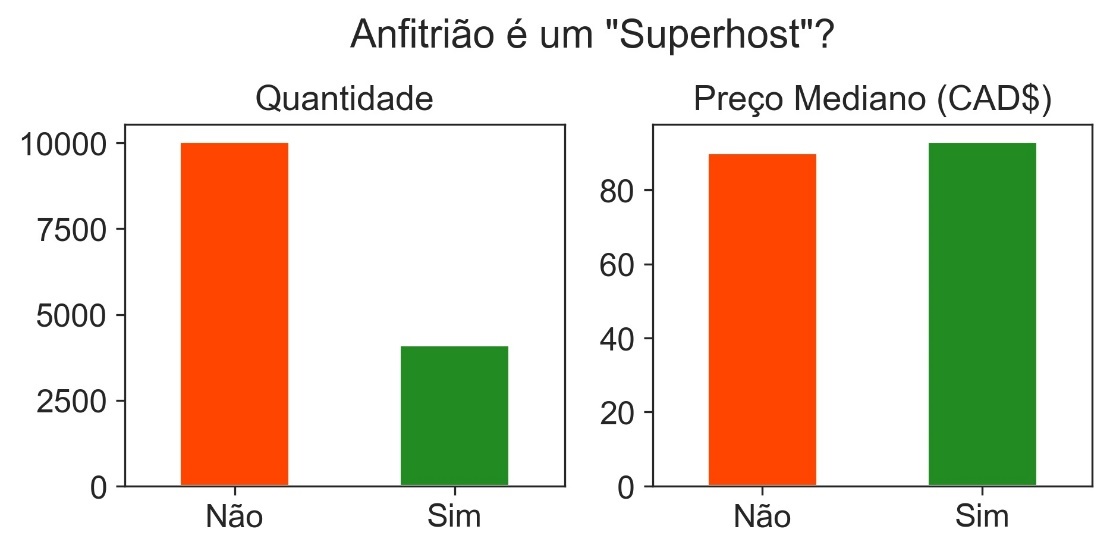


Figura 15. Quantidade de superhosts, e o efeito deles no preço da diária.

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 indica, 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.

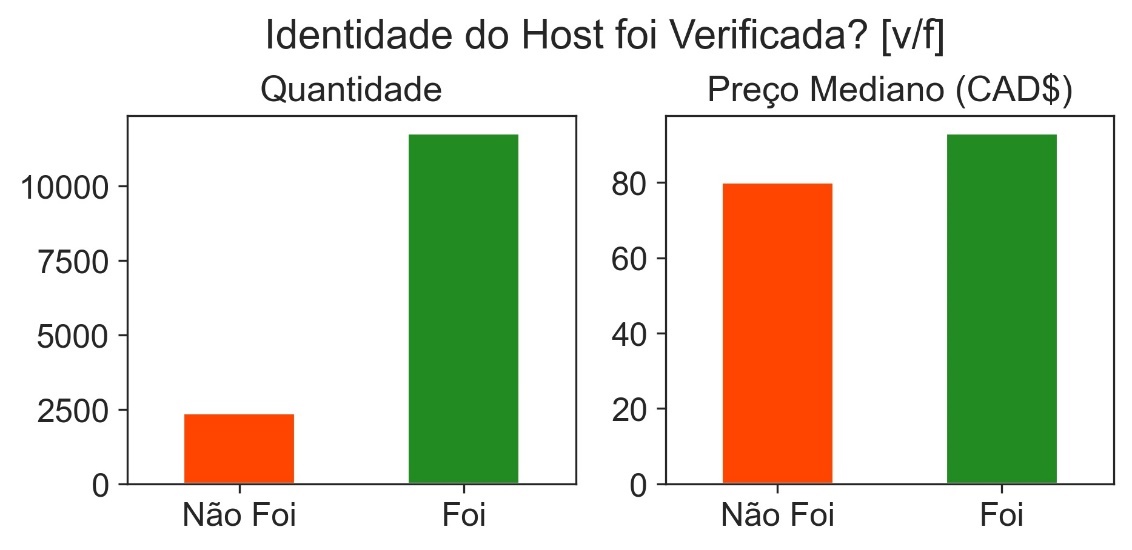


Figura 16. Quantidade de anfitriões verificados, e o efeito deles no preço da diária.

## Geolocalização - Municípios

A cidade de Toronto é dividida em seis municípios principais, esses são demonstrados na Figura 17, conforme disponibilizado pela Wikipedia [11]. 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 [11].

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, que disponibiliza os dados fornecidos pela OpenStreetMap [12]. O Anexo 9 também disponibiliza uma visualização alternativa, mostrando um mapa de calor dessas propriedades sob o mapa de Toronto.

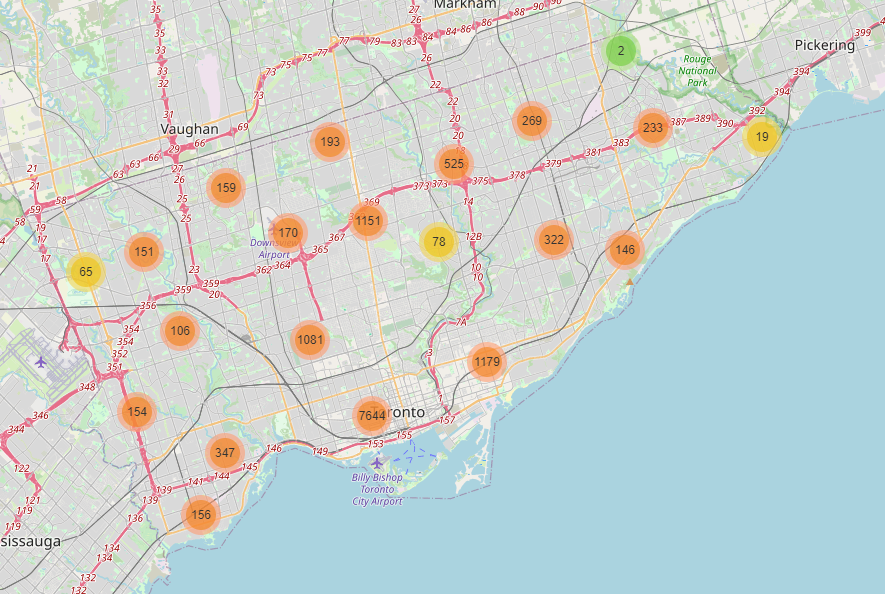


Figura 18. Propriedades listadas no dataframe, agrupadas geograficamente.

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 em Old Toronto. Ele mostra, também, que o valor da diária é maior em Toronto, e em geral, na zona sul de Toronto.

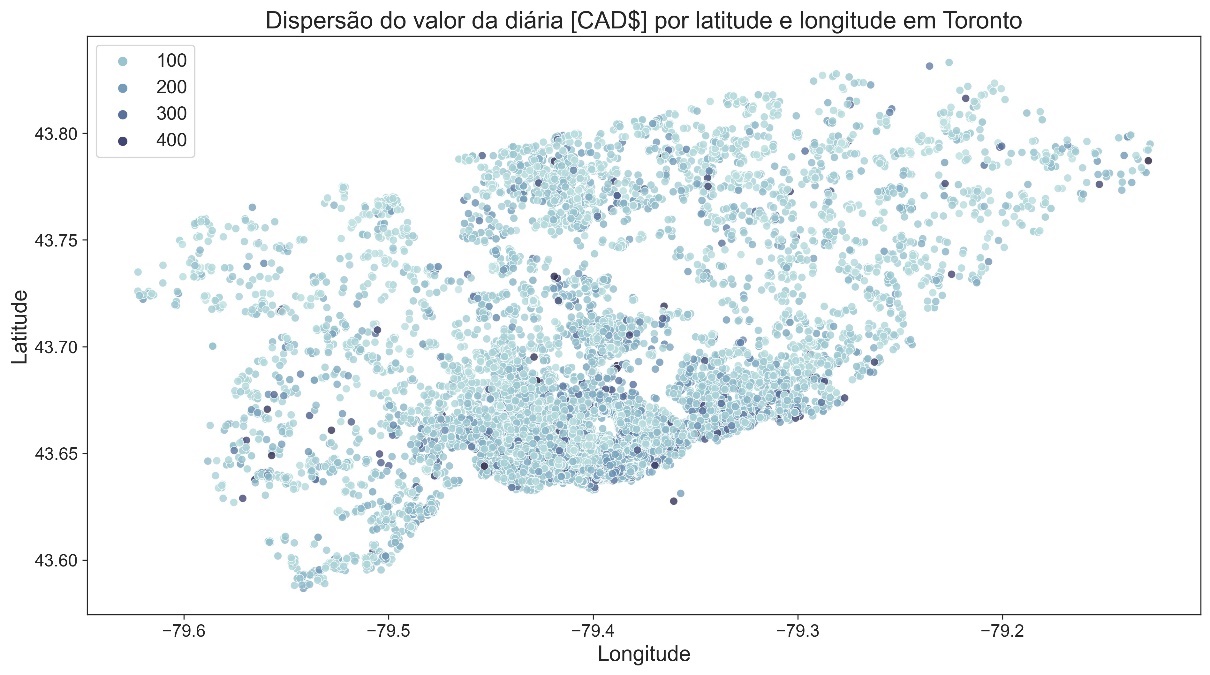


Figura 19. Dispersão de Latitude e Longitude, colorido por faixa de preço.

A Figura 20 mostra um boxplot dos preços de diária por município. Conforme esperado, Old Toronto tem preços significativamente maiores. Além disso, é relevante mencionar 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 os 16% restantes.

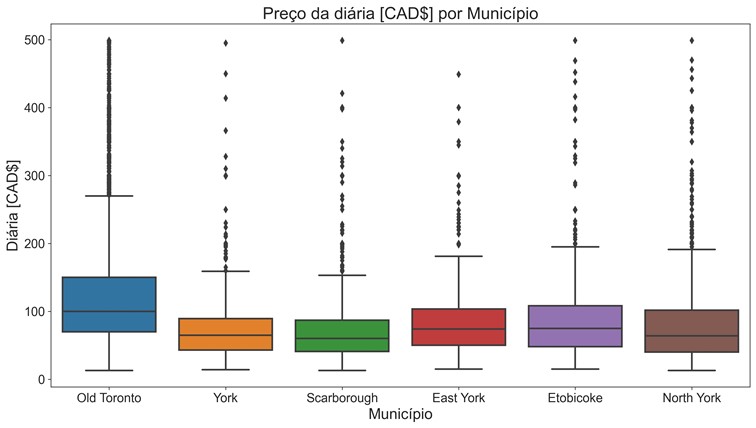


Figura 20. Distribuição de preços por município, em Toronto.

## 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 <> 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 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

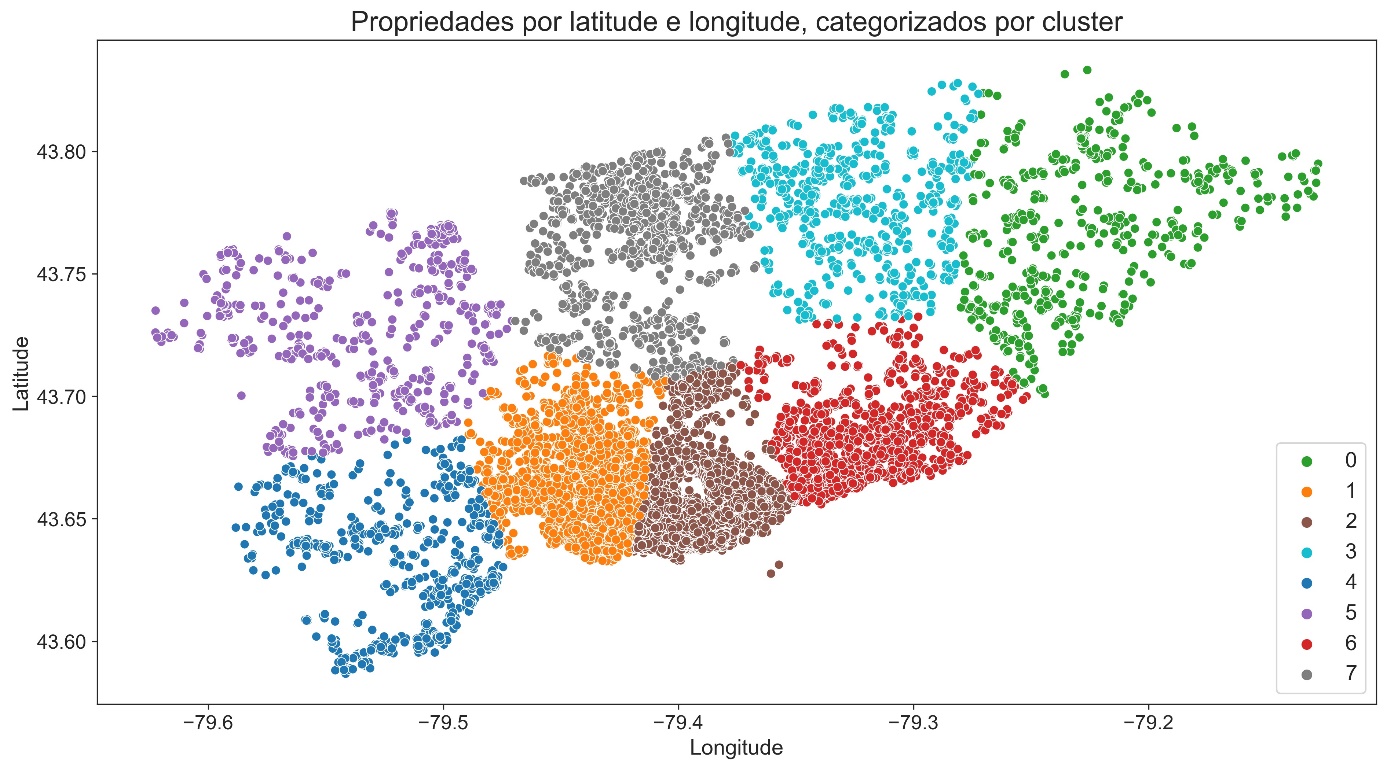


Figura 21. Dispersão de Latitude e Longitude, categorizado por ID do cluster

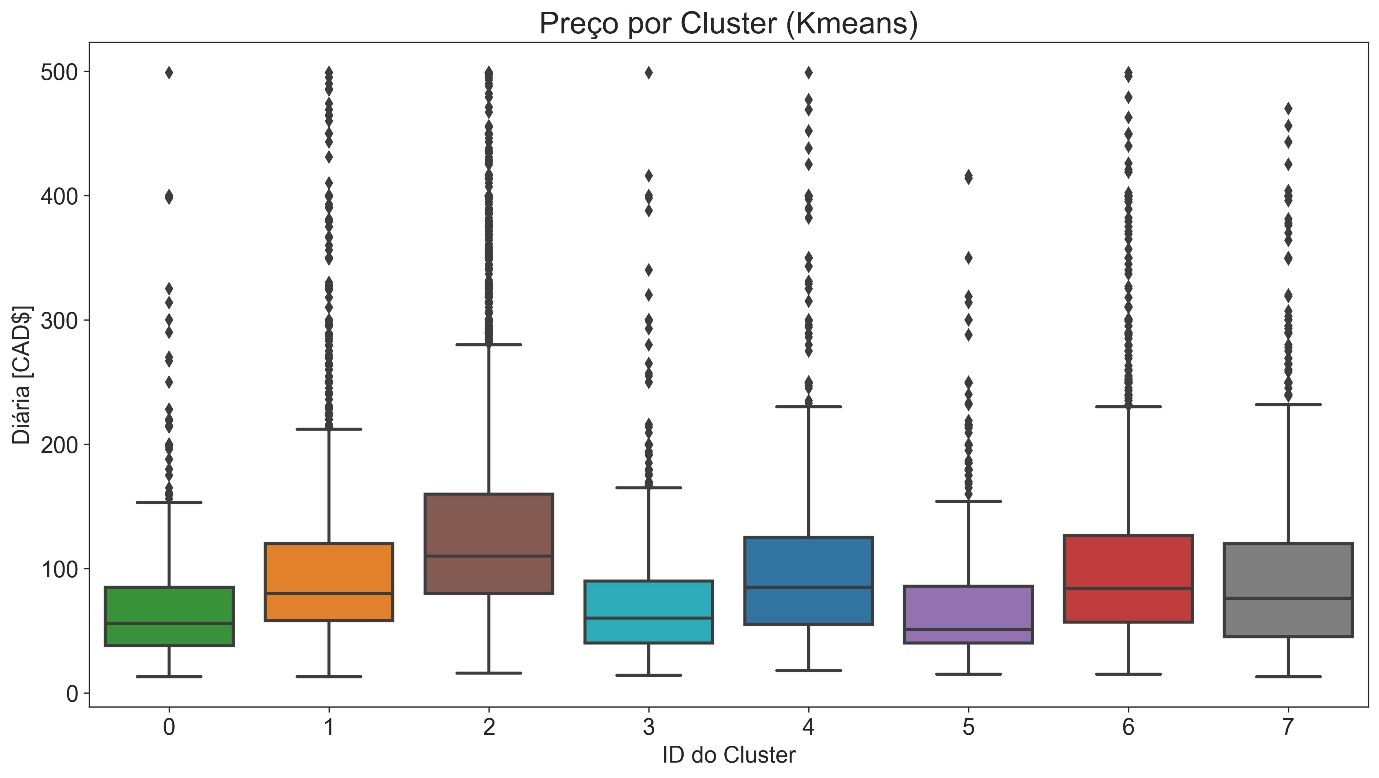


Figura 22. Distribuição de preços por cluster, em Toronto.

## Tipos de Propriedade e Acomodação

A Figura 23 mostra os tipos mais comuns de propriedade. 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 individual (‘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 1 analisa os tipos de imóvel quantificados por município. É possível ver que os padrões não se mantem. Enquanto 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.

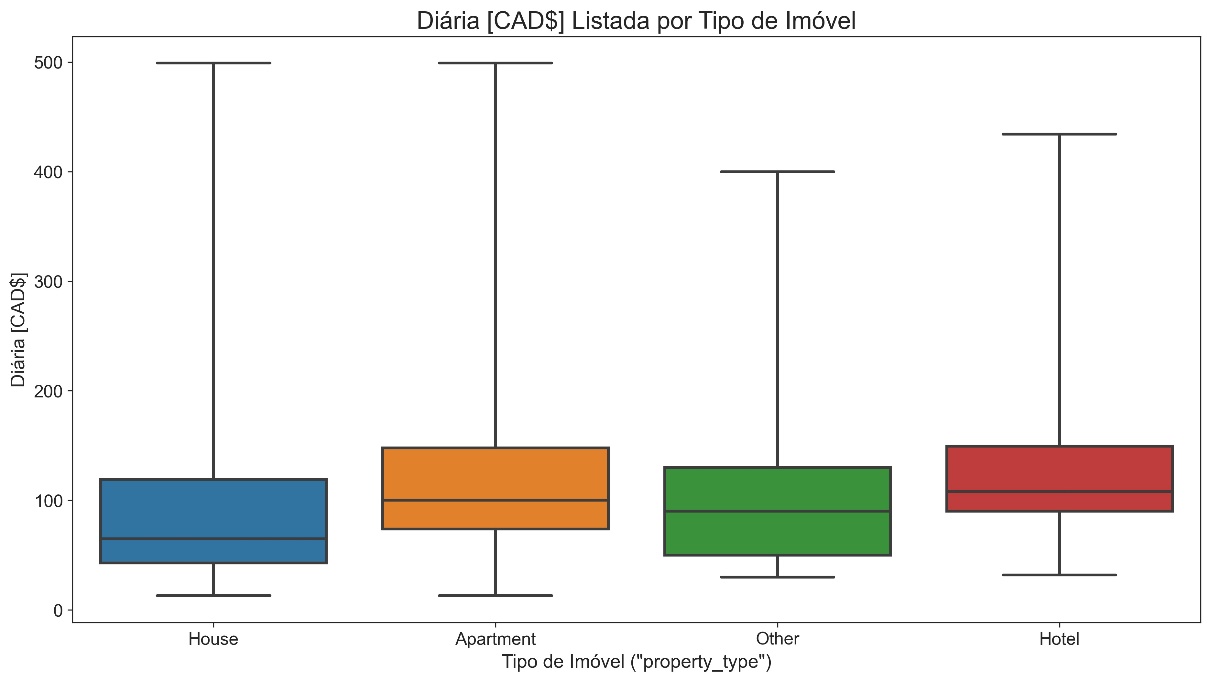


Figura 24. Boxplot do valor da diária [CAD$] por tipo de imóvel.

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 significativamente mais caras do que os outros tipos. O segundo tipo de acomodação mais caro são quartos privados e quartos de hotéis. Finalmente, os mais baratos são quartos compartilhados.

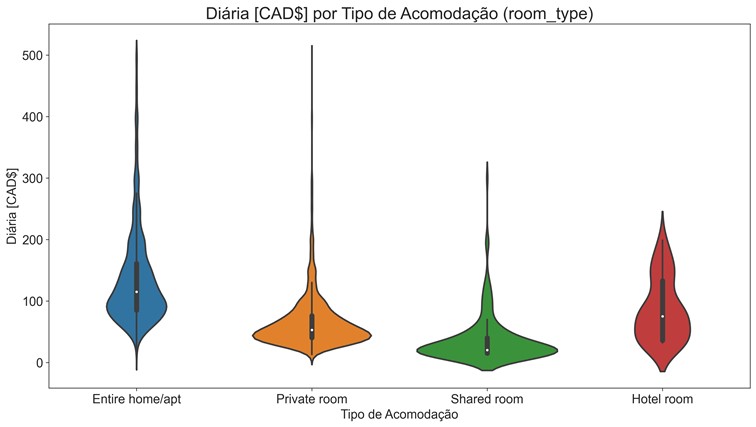


Figura 25. Violinplot do valor da diária [CAD$] por tipo de acomodação.

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.

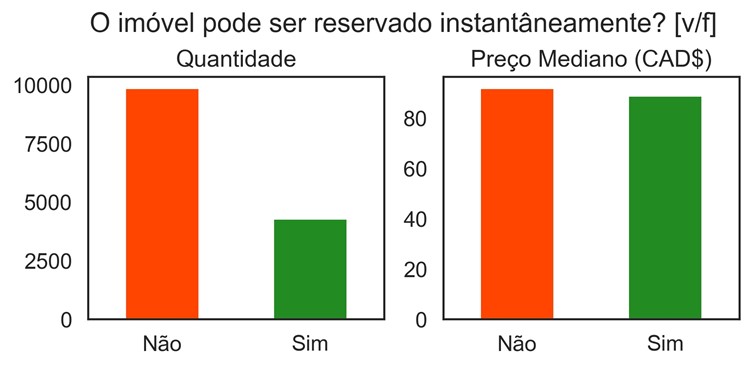


Figura 26. Número de imóveis imediatamente reserváveis, e o efeito deles no preço.

## 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.

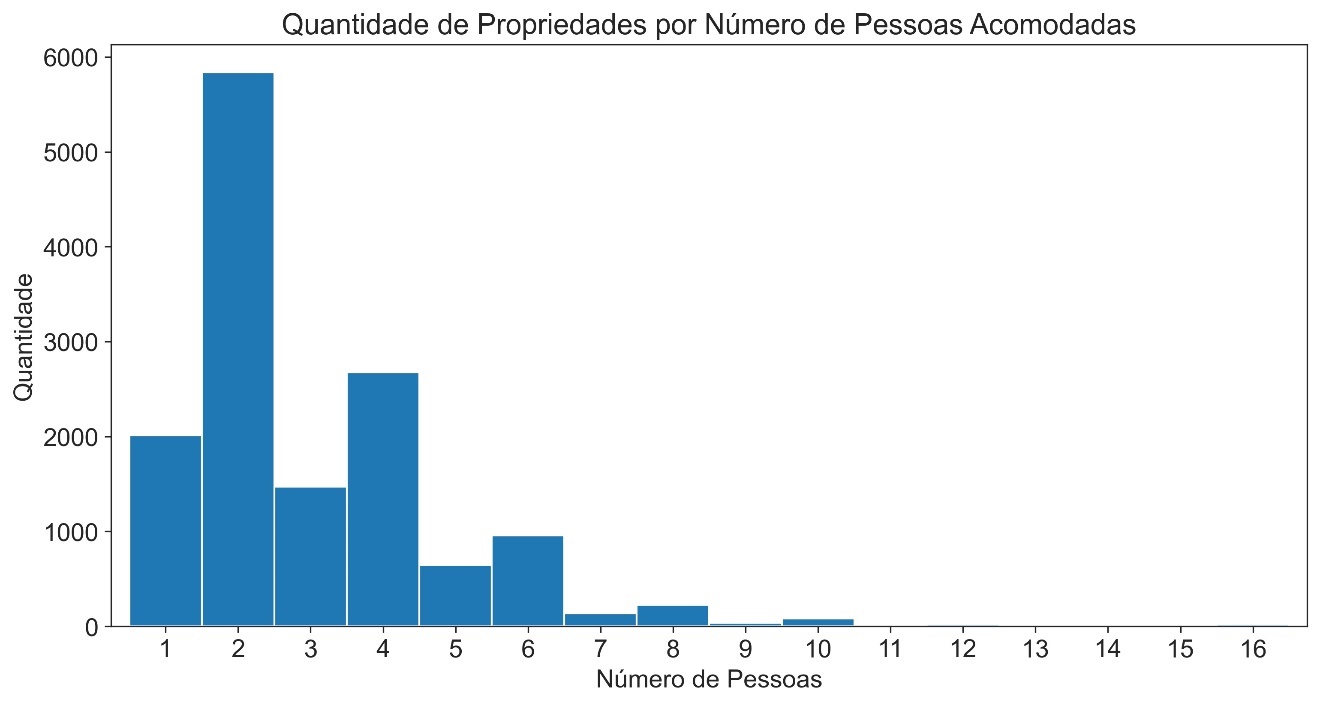


Figura 27. Quantidade de propriedades por número de pessoas acomodadas.

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.

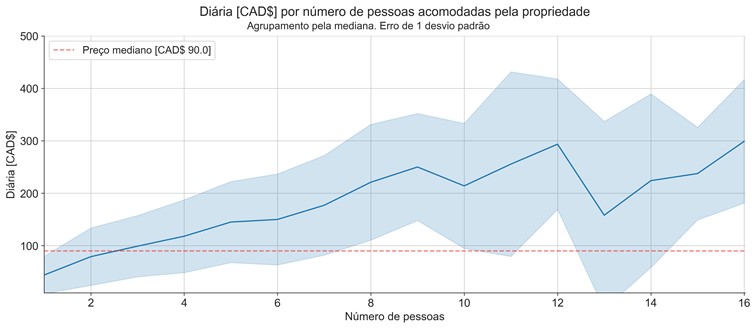


Figura 28. Preço pelo número de pessoas acomodadas, agrup. pela mediana.

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 <>, onde fica claro que por mais que o preço da diária aumente, o preço da diária por pessoa diminui. Em outras palavras, economicamente sempre 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 |

## Amenidades

O modelo considera <> amenidades, então fica 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 considerados. 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 acontece, muito provavelmente, pois os únicos imóveis que tem quarto com tranca são também imóveis compartilhados. E, 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

# Criação de Modelos de Machine Learning

## Preparação – Remoção de Colunas

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

As primeiras colunas removidas são host\_since, last\_review e first\_review. Essas três colunas são colunas de datas. Os atributos que foram criados a partir destas, host\_since\_days, last\_review\_days e first\_review\_days, que são colunas numéricas com os números de dias para cada um dos eventos serão incluídas no modelo.

Depois a latitude e longitude serão removidas. Isso porque o modelo vai utilizar, conforme concluído na seção de análise e exploração dos dados, os clusters que foram produzidos a partir da latitude e longitude. Pelo mesmo motivo, a coluna geo\_city, com o nome do município, será removida. Finalmente, a coluna host\_id será removida, pois representa apenas um código de identificação que não é útil para o modelo.

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

## 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. 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.

## Preparação – Correlação Entre as Variáveis

Por mais que, durante o estudo, cuidado tenha sido tomado para evitar um alto índice de correlação entre as variáveis (através de transformação, e remoção de redundâncias), ainda é importante verificar as redundâncias. Para isso, vamos usar uma 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.

Na primeira parte, 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 acima de 0.99 (em módulo). 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.

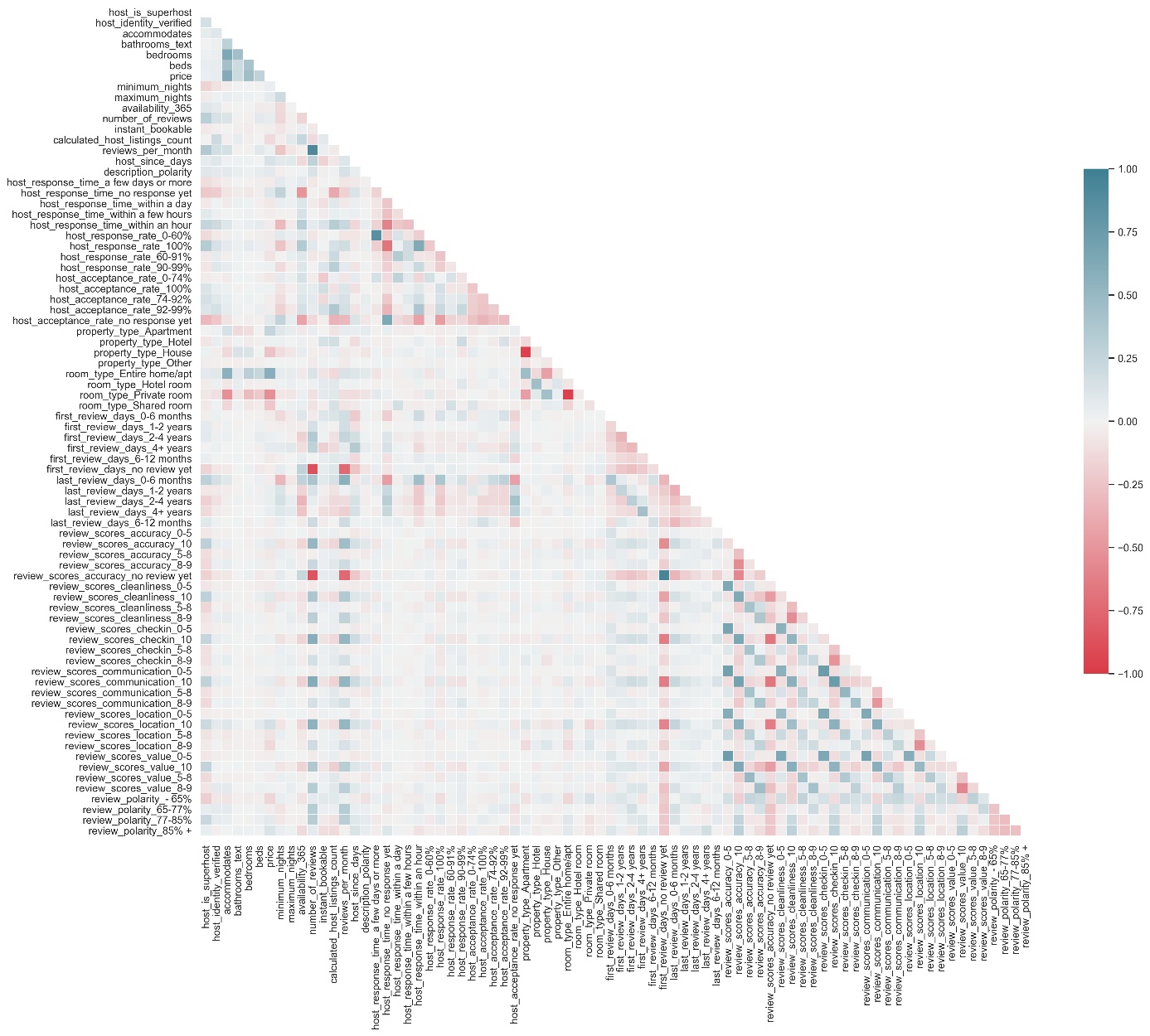


Figura 29. Mapa de Calor da Matriz de correlação – Parte 1.

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

É possível ver que os dados são coerentes nessas correlações. 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).

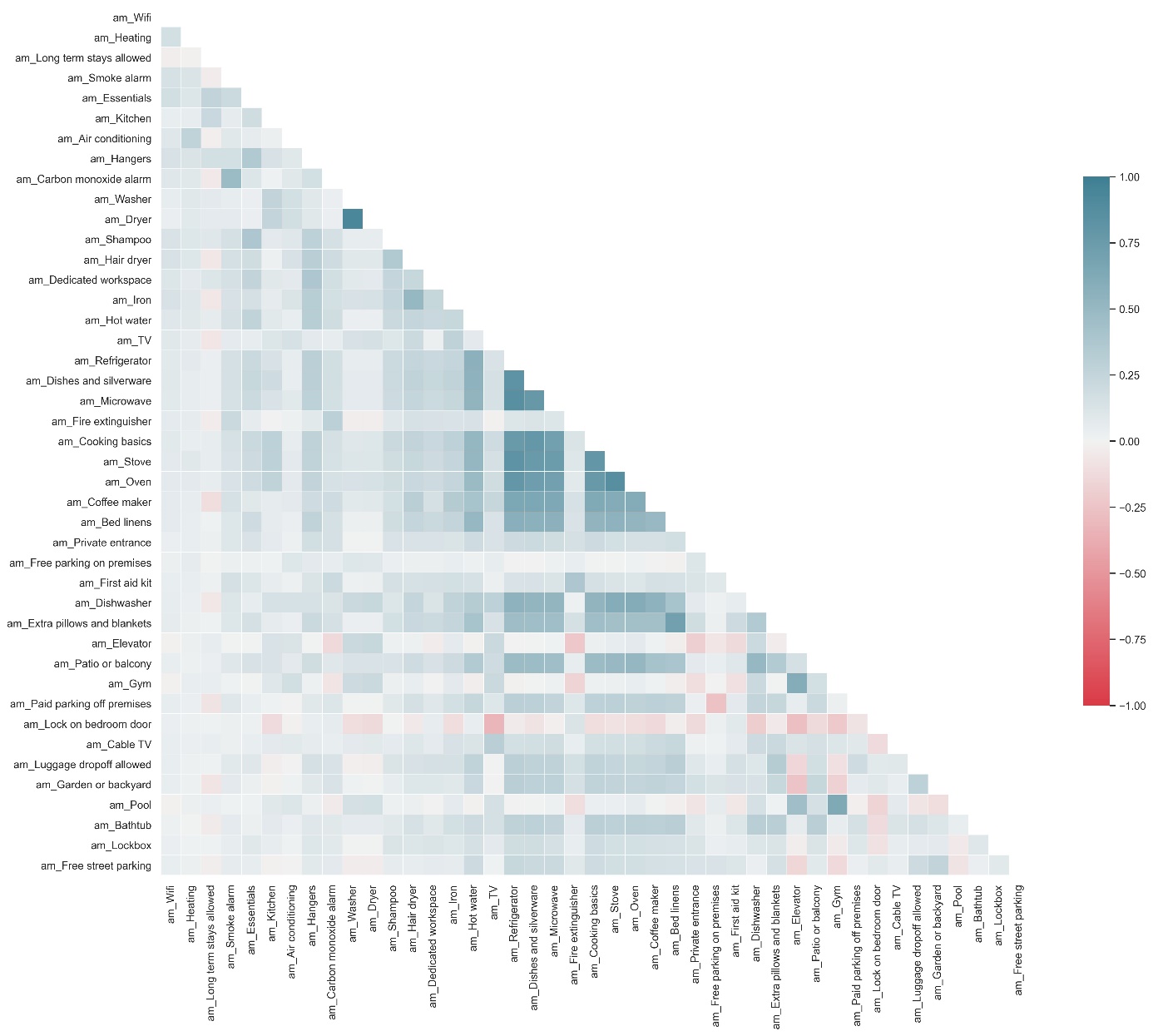


Figura 30. Mapa de Calor da Matriz de correlação – Parte 2.

## 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 [13]. 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.

## Preparação – Scaling, e Divisão dos dados

Finalmente, a coluna de interesse é separada (price), e os atributos de predição são 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.

## 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, RandomForestRegressor) e Gradient Boosting (Biblioteca XGBoost, XGBRegressor).

Serão utilizadas duas métricas. A primeira é o R², 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.

## 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 R²** | 0,955 | **Test R²** | 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 R²** | 0,898 | **Test R²** | 0,683 |

## 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% peso de importância. Esses são demonstrados na Figura 31.

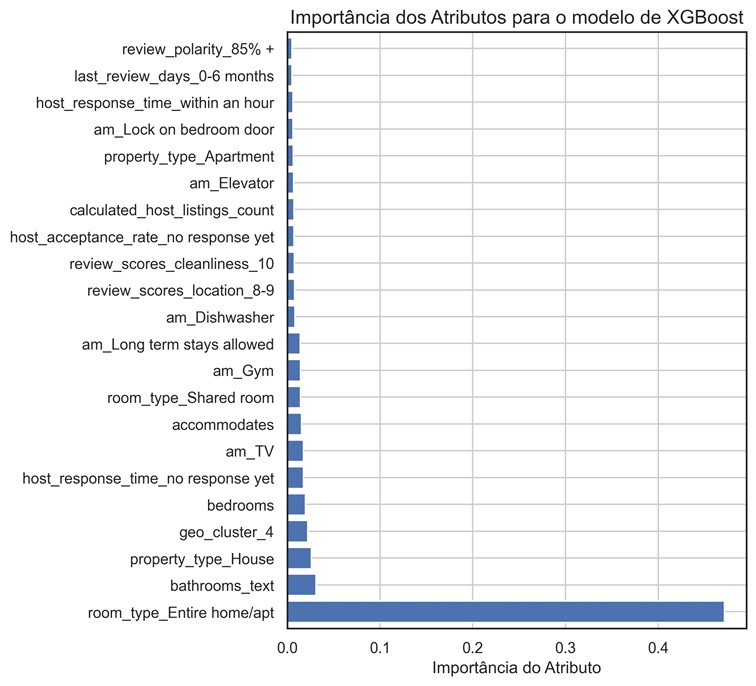


Figura 31. Atributos com mais de 0.5% de importância

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

## Interpretação dos Atributos mais Importantes

Como esperado pelas análises feitas na seção de análise e exploração dos dados, o atributo mais importante pro 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 importante, 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 correspondente a latitude e longitude de Old Toronto.

## 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 R²** | 0,955 | **Test R²** | 0,689 |

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Gradient Boost (XGBoost) - 2** | | | |
| **Training MSE** | 0,043 | **Test MSE** | 0,125 |
| **Training R²** | 0,894 | **Test R²** | 0,691 |

## Seleção do Melhor Modelo

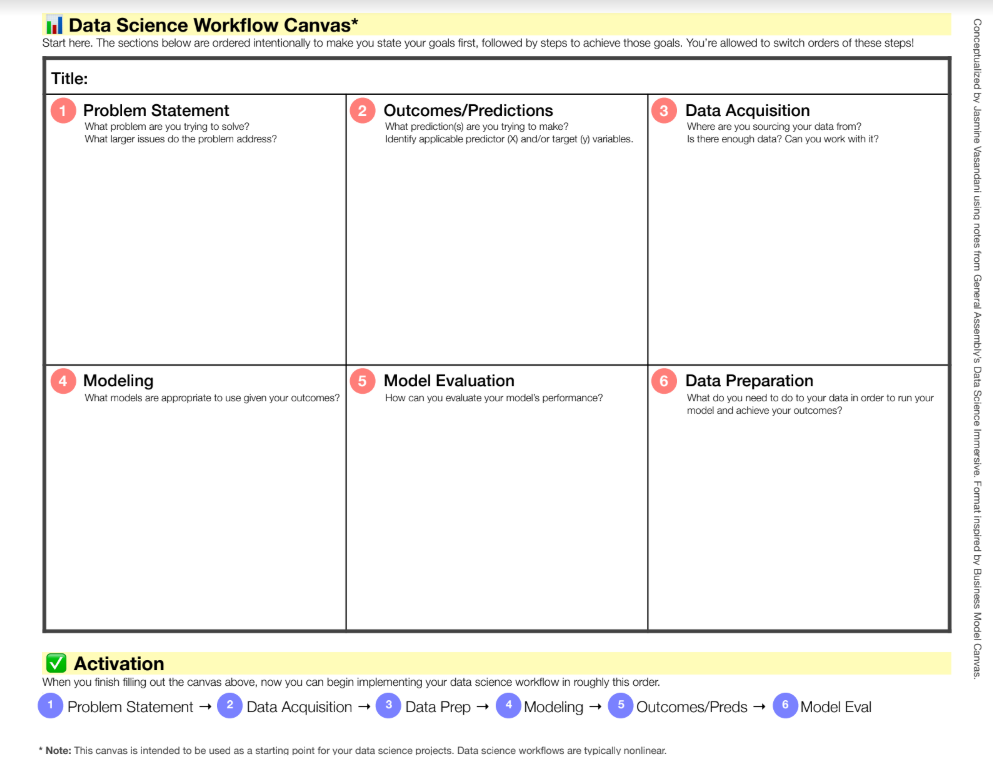
Embora a precisão não tenha melhorado significativamente com remoção de atributos, o tempo de execução e a exigência de recursos foram significativamente reduzidas. Portanto, os melhores modelos são após a remoção das colunas com baixa importante.

O algoritmo Random Forest apresentou uma melhor performance no grupo de treino, mas uma pior performance no grupo de teste. Com critério de selecionar o melhor algoritmo no grupo de teste, o algoritmo de Gradient Boost após a remoção dos atributos é selecionado.

Conforme o documento de instruções para o TCC, essa etapa é obrigatória. Nessa seção você irá descrever as ferramentas e algoritmos utilizados. Se você utilizou ferramentas visuais como Knime e Rapid Miner, coloque aqui um print do seu modelo e a descrição detalhada do workflow do seu modelo. Caso você tenha escrito scripts em Python, por exemplo, coloque aqui o seu script. Explique as *features* utilizadas, faça a comparação entre diferentes algoritmos/modelos, justifique a escolha por determinado modelo, os parâmetros utilizados, etc. Por fim, salienta-se que embora você possa utilizar ferramentas como KNIME e RapidMiner para testar protótipos do seu modelo de ML, encorajamos você a fazer seus modelos em Python ou R.

# Apresentação dos Resultados

Nessa seção você deve apresentar os resultados obtidos. Apresente gráficos, *dashboards*, conte a sua história de forma bastante criativa. Aqui você pode utilizar os modelos de Canvas propostos por Dourard (clique [aqui](https://www.louisdorard.com/machine-learning-canvas)) ou por Vasandani (clique [aqui](https://towardsdatascience.com/a-data-science-workflow-canvas-to-kickstart-your-projects-db62556be4d0)).



# Links

Aqui você deve disponibilizar os links para o vídeo com sua apresentação de 5 minutos e para o repositório contendo os dados utilizados no projeto, scripts criados, etc.

Link para o vídeo: youtube.com/...

Link para o repositório: github.com/...

x

x

x

REFERÊNCIAS

x

|  |  |
| --- | --- |
| 1. | ABOUT. **Inside Airbnb**, 2021. Disponivel em: <http://insideairbnb.com/about.html>. Acesso em: 4 Abril 2021. |
| 2. | GET-THE-DATA. **Inside Airbnb**, 2021. Disponivel em: <http://insideairbnb.com/get-the-data.html>. Acesso em: 04 Abril 2021. |
| 3. | AIRBNB. **Termos de Serviço**, 2021. Disponivel em: <https://www.airbnb.com.br/help/article/2908/termos-de-servi%C3%A7o>. Acesso em: 04 Abril 2021. |
| 4. | CC01.0. **Creative Commons**, 2021. Disponivel em: <https://creativecommons.org/publicdomain/zero/1.0/>. Acesso em: 04 Abril 2021. |
| 5. | GEOCODING API. **Mapquest Developer**, 2018. Disponivel em: <https://developer.mapquest.com/documentation/geocoding-api/reverse/get/>. Acesso em: 04 Abril 2021. |
| 6. | PLANS. **Mapquest Developer**, 2021. Disponivel em: <https://developer.mapquest.com/plans>. Acesso em: 04 Abril 2021. |
| 7. | LEGAL. **Developer Mapquest**, 2021. Disponivel em: <https://developer.mapquest.com/legal>. Acesso em: 04 Abril 2021. |
| 8. | 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: 04 Abril 2021. |
| 9. | 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. |
| 10. | TRAVEL Budget for Toronto. **Budget Your Trip, LLC**, 2021. Disponivel em: <https://www.budgetyourtrip.com/canada/toronto>. Acesso em: Abril 2021. |
| 11. | FILE:NORTH York Locator.png. **Wikipedia**, 2021. Disponivel em: <https://en.wikipedia.org/wiki/File:North\_York\_Locator.png>. Acesso em: Abril 2021. |
| 12. | ABOUT. **Open Street Map**, 2021. Disponivel em: <https://www.openstreetmap.org/about>. Acesso em: Abril 2021. |
| 13. | CURRAN-EVERETT, D. Explorations in statistics: the log transformation. **Advances in Physiology Education**, v. 40, n. 2, p. 343-347, Dezembro 2016. |
| 14. | STANDARD Motor Codes. **National Association of Rocketry**. Disponivel em: <http://www.nar.org/standards-and-testing-committee/standard-motor-codes/>. Acesso em: 19 mar. 2017. |

x

x

x

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

APÊNDICE

**Programação/Scripts**

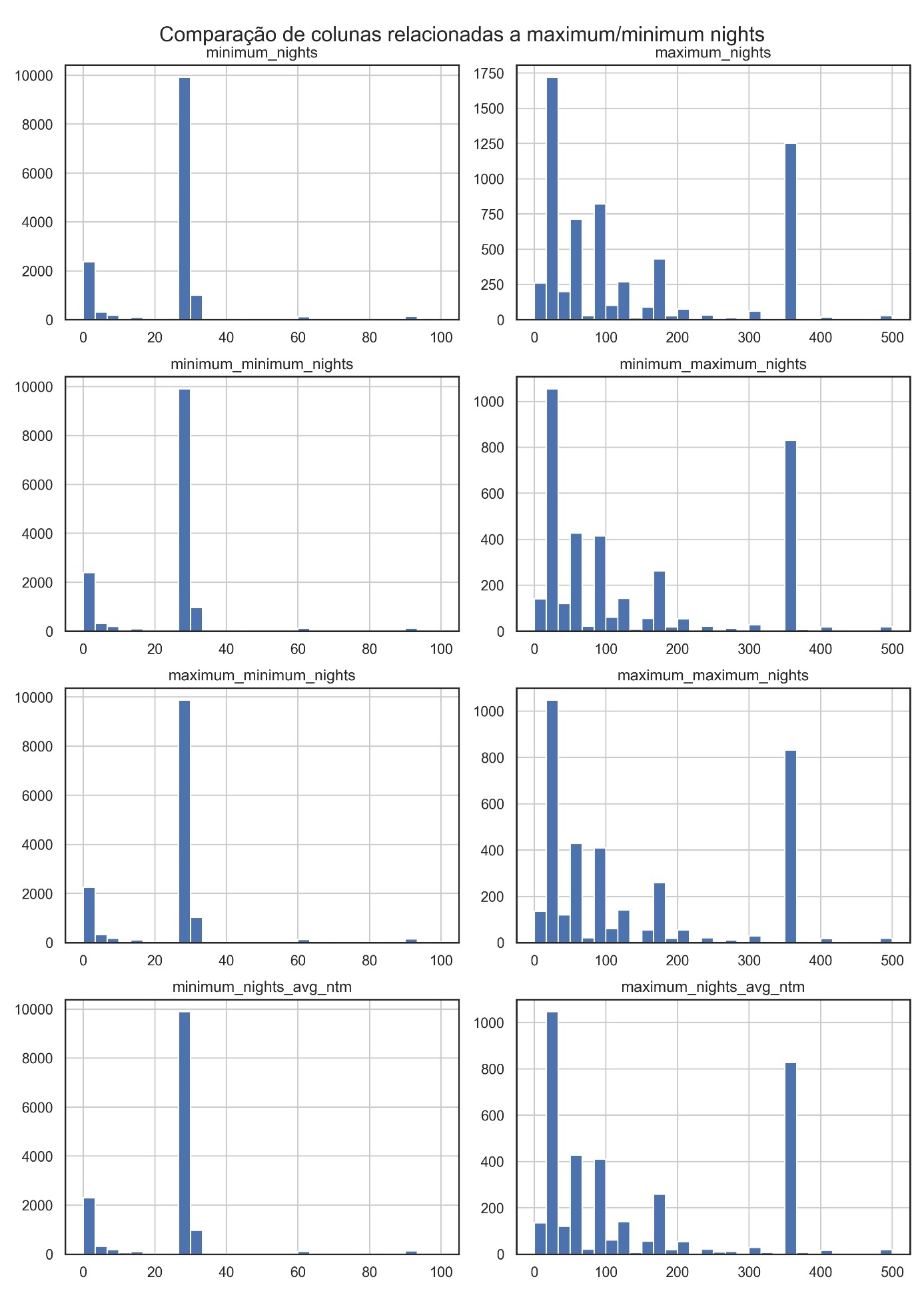
Cole aqui seus scripts em Python e/ou R.

**Gráficos**

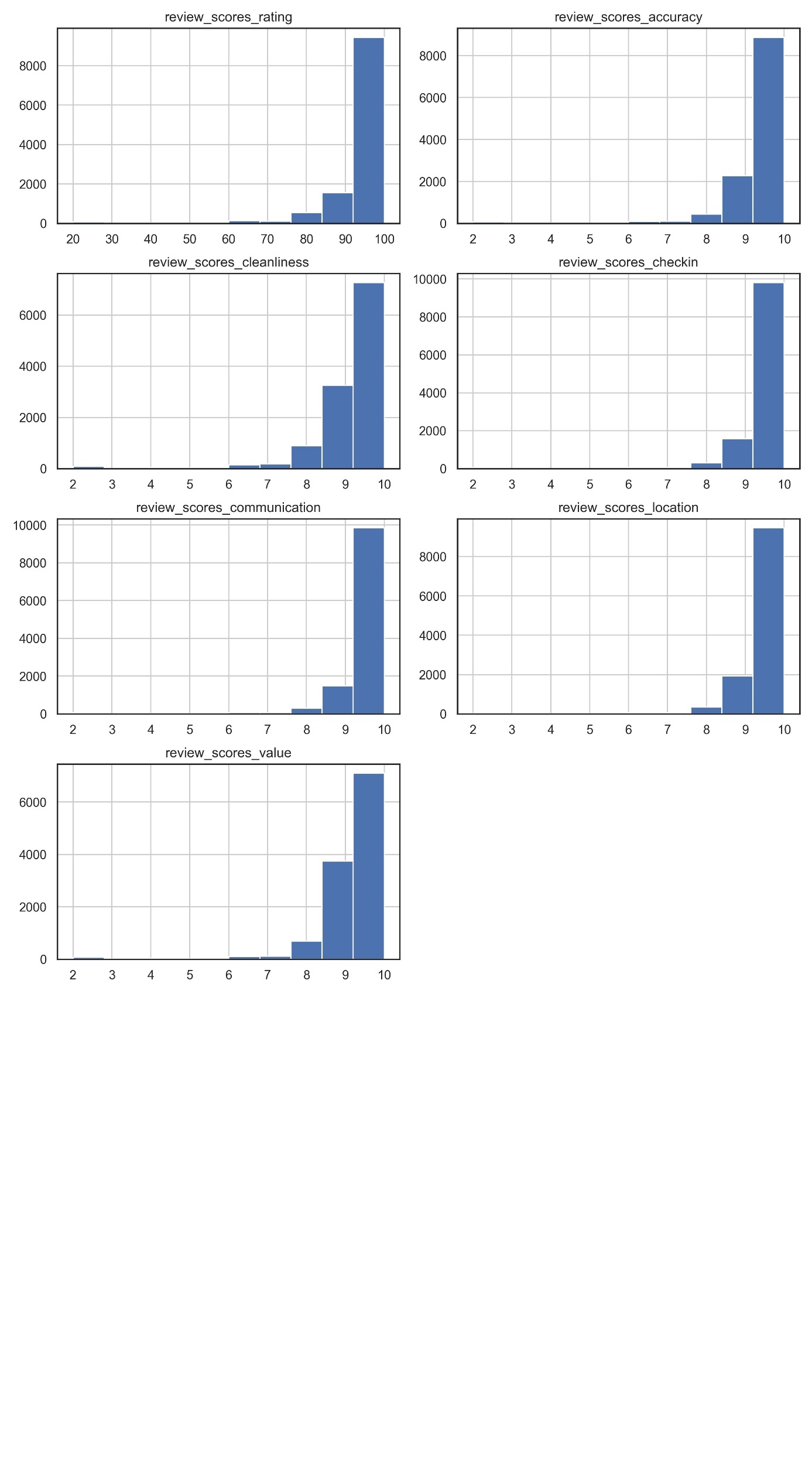
Cole aqui workflows (KNIME, RapidMiner, etc.), gráficos e figuras que você tenha gerado e não colocou no texto principal.

**Tabelas**

Cole aqui tabelas de dados que você tenha gerado e não colocou no texto principal.



Anexo 1. Comparação das colunas que descrevem maximum\_nights 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<br /><br />  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 |

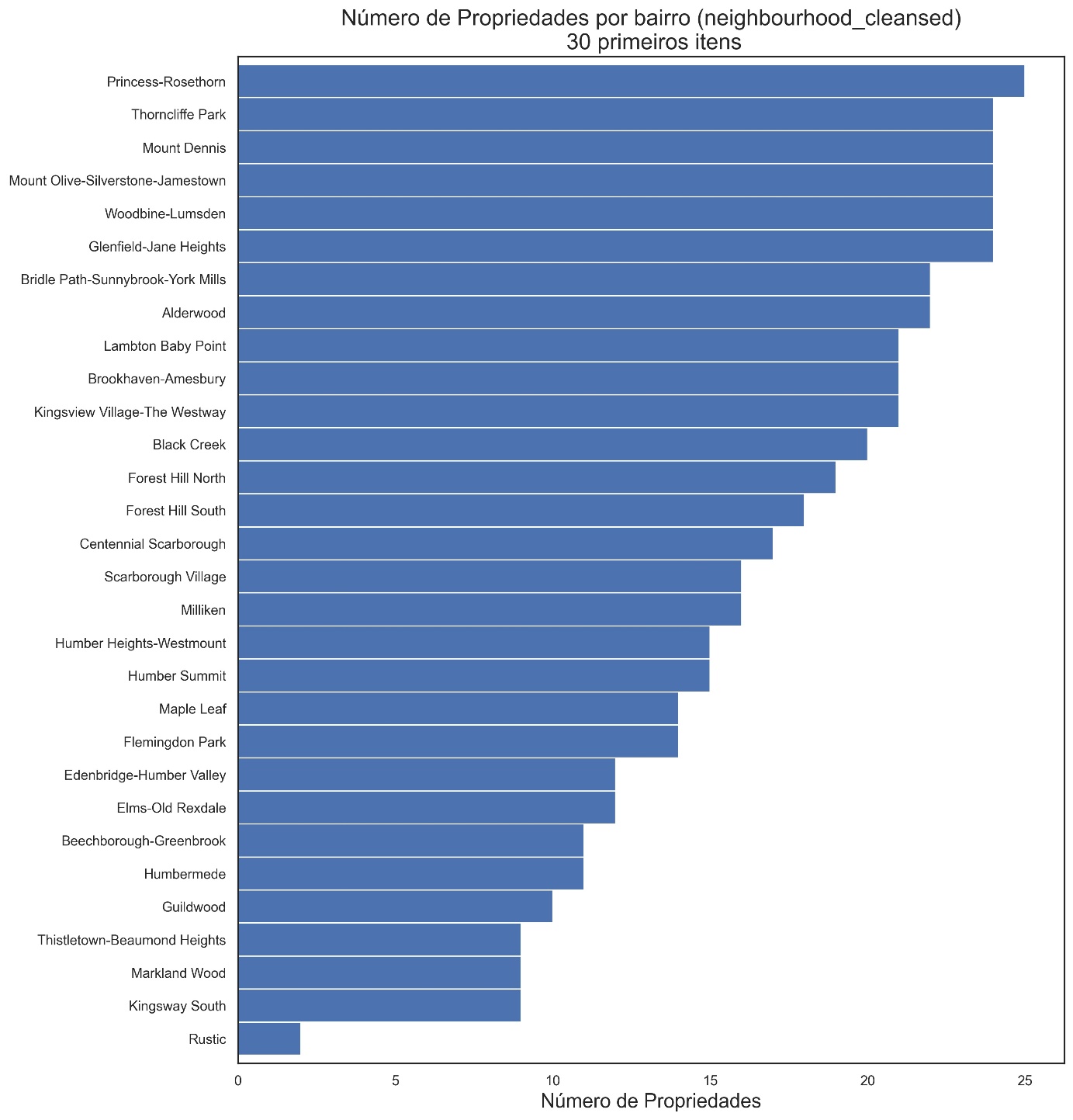
Anexo 5. Segunda pior descrição (polarity\_score = -0.9063). Texto não alterado.

|  |
| --- |
| 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 |

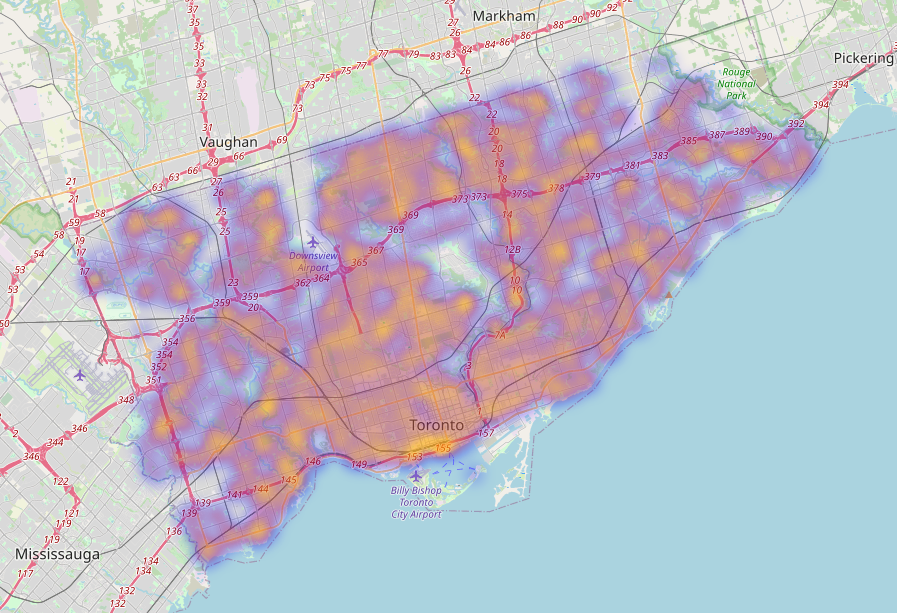
Anexo 6. Melhor descrição (polarity\_score = +0.9979). Texto não alterado.



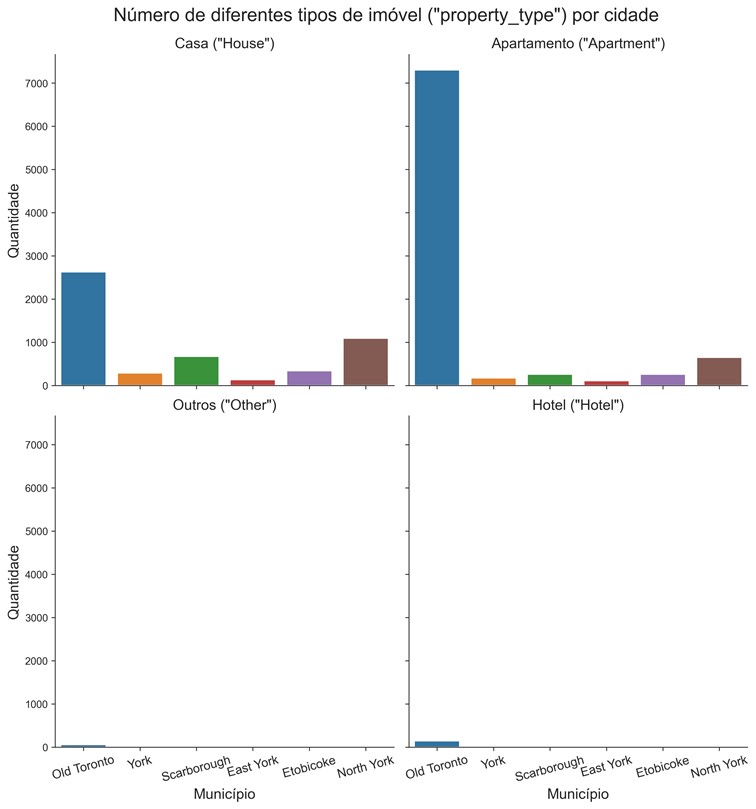
Anexo 7. Nuvem de palavras das palavras mais usadas nas avaliações positivas. Máscara (Formato) utilizado da CN Tower, atração de Toronto.



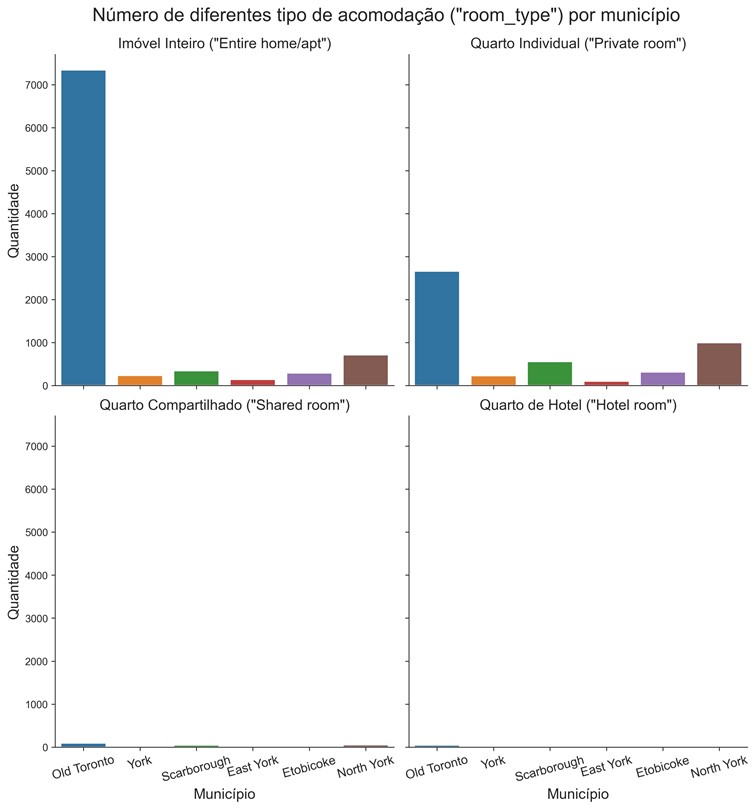
Anexo 8. Distribuição de propriedades por bairro (neighbourhood\_cleansed).



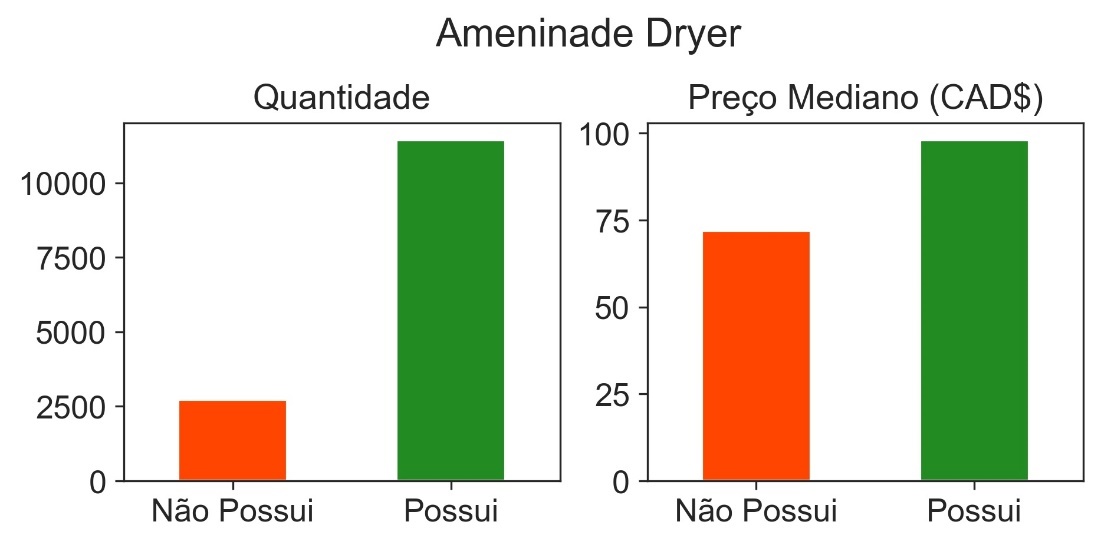
Anexo 9. Mapa de calor das propriedades listadas no dataframe.



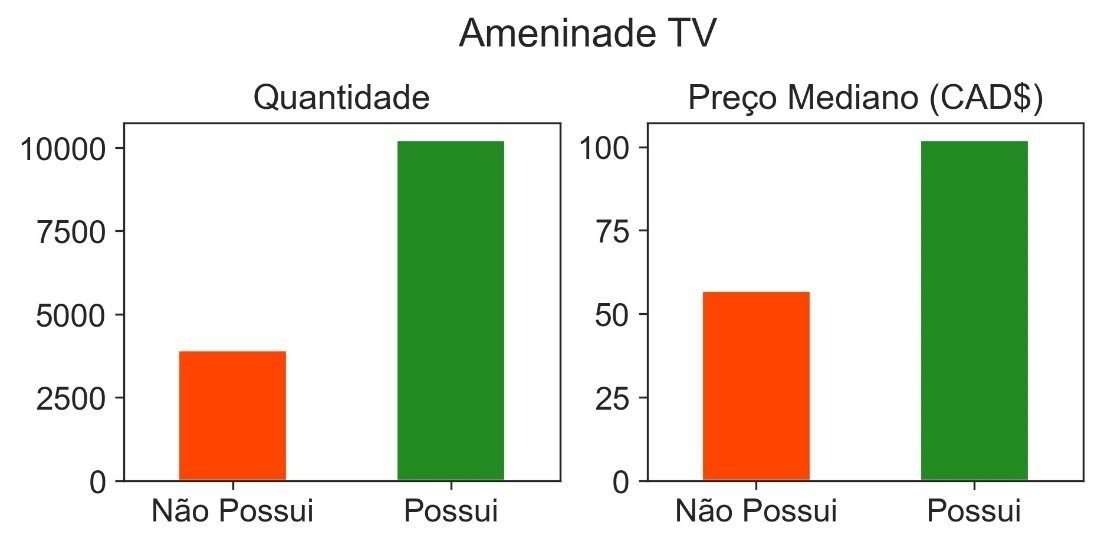
Anexo 10. Tipos de imóvel por município.



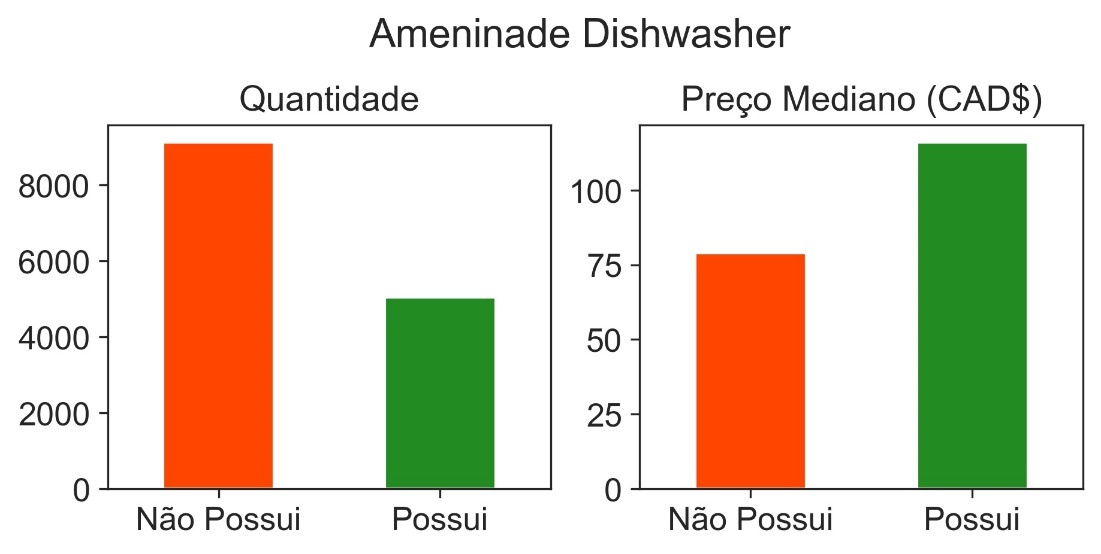
Anexo 11. Tipos de acomodação por município.



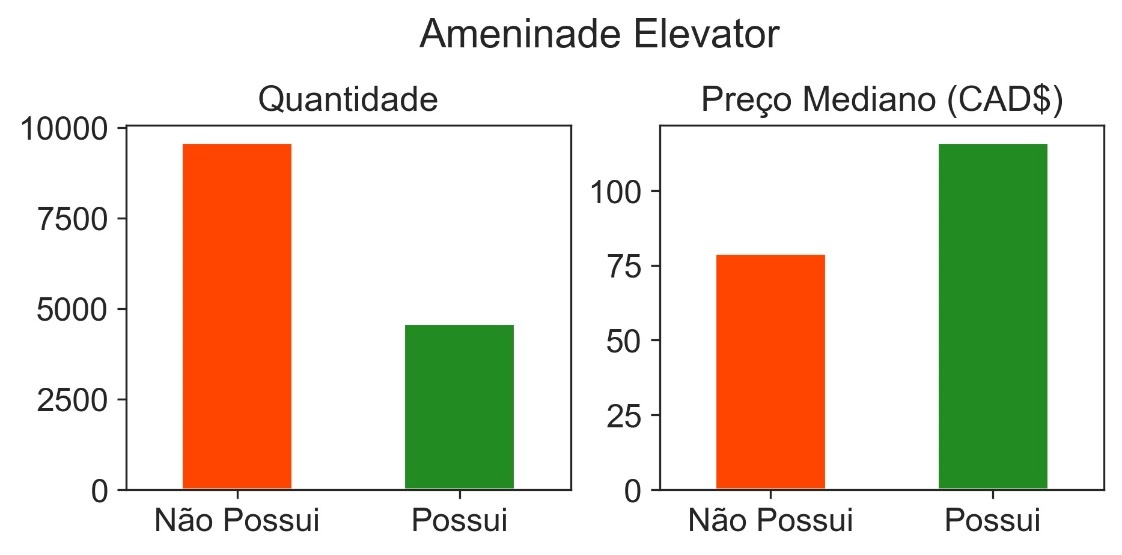
Anexo 12. Frequência e impacto no preço da secadora de roupas (Dryer).



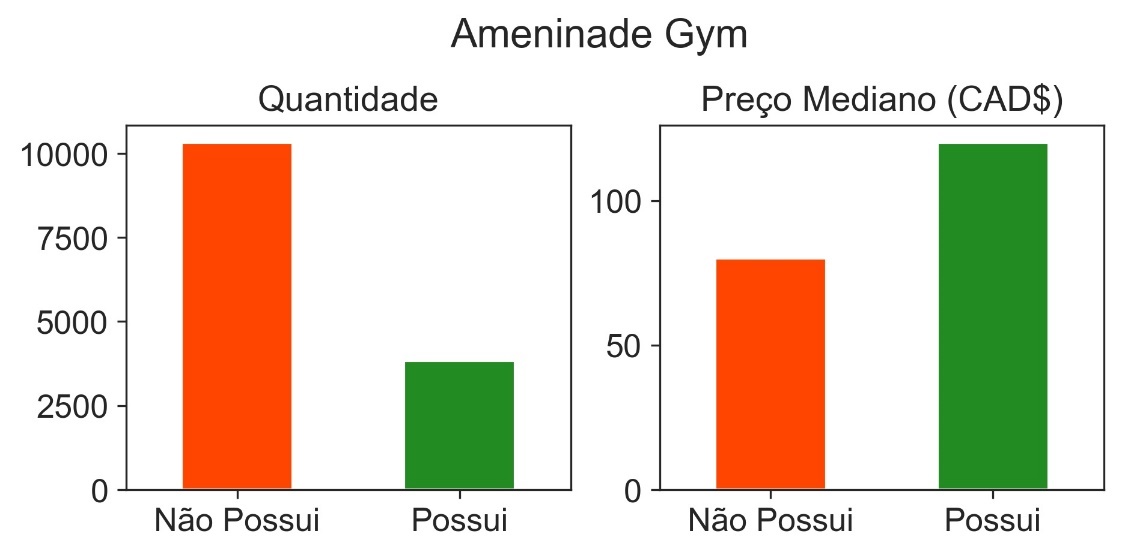
Anexo 13. Frequência e impacto no preço da televisão (TV).



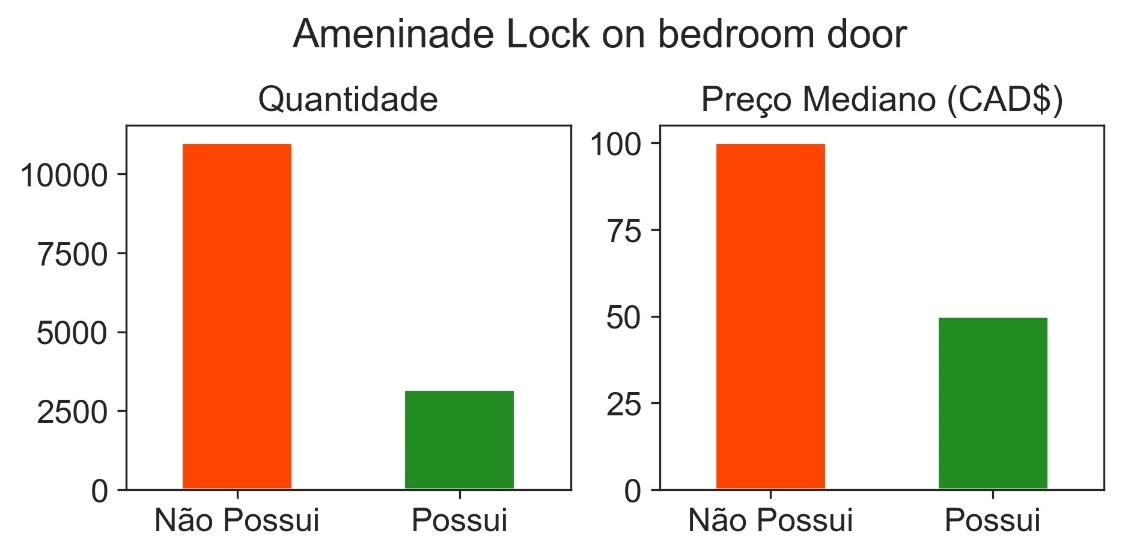
Anexo 14. Frequência e impacto no preço da lava-louças (Dishwasher).



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



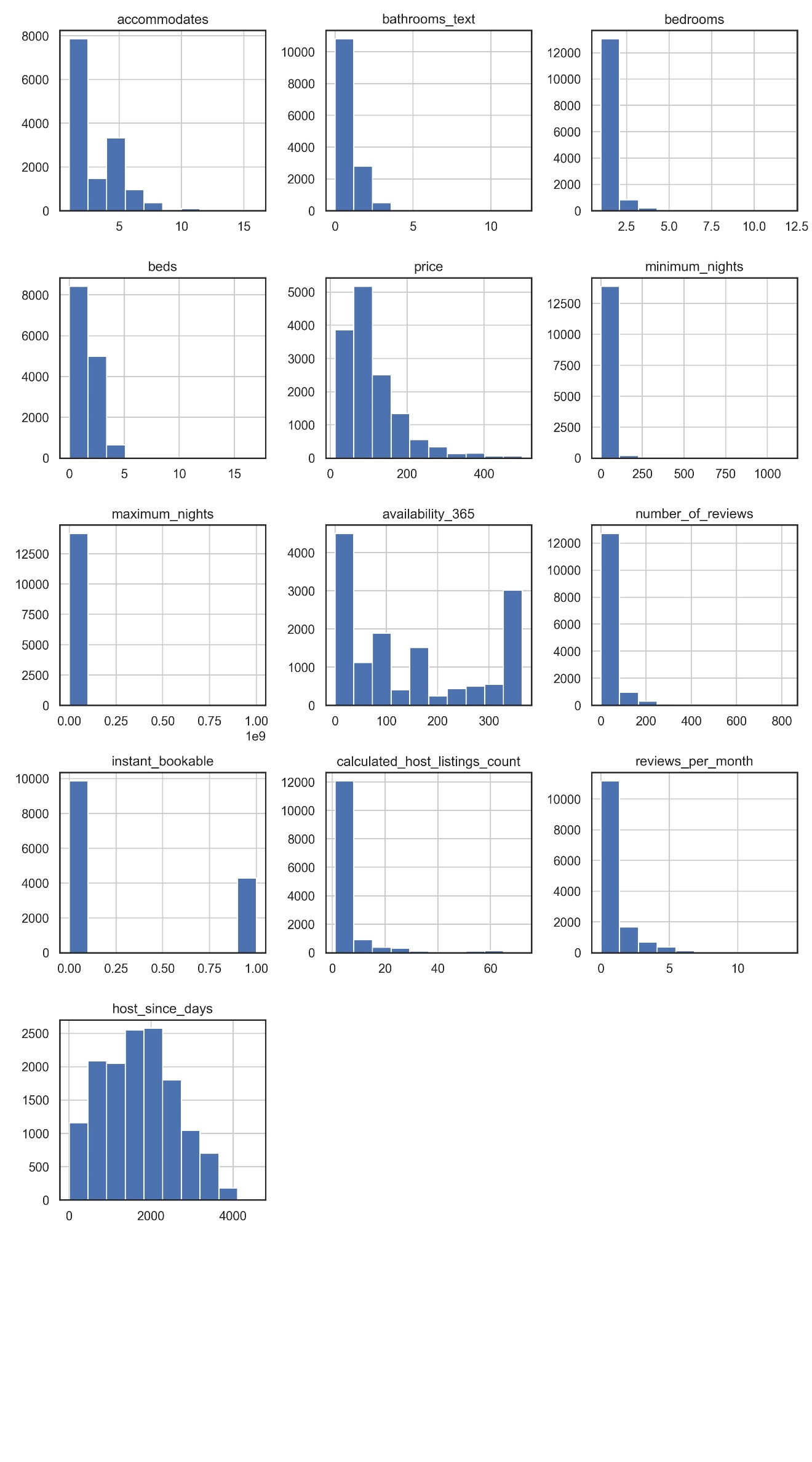
Anexo 16. Frequência e impacto no preço da Academia (Gym).



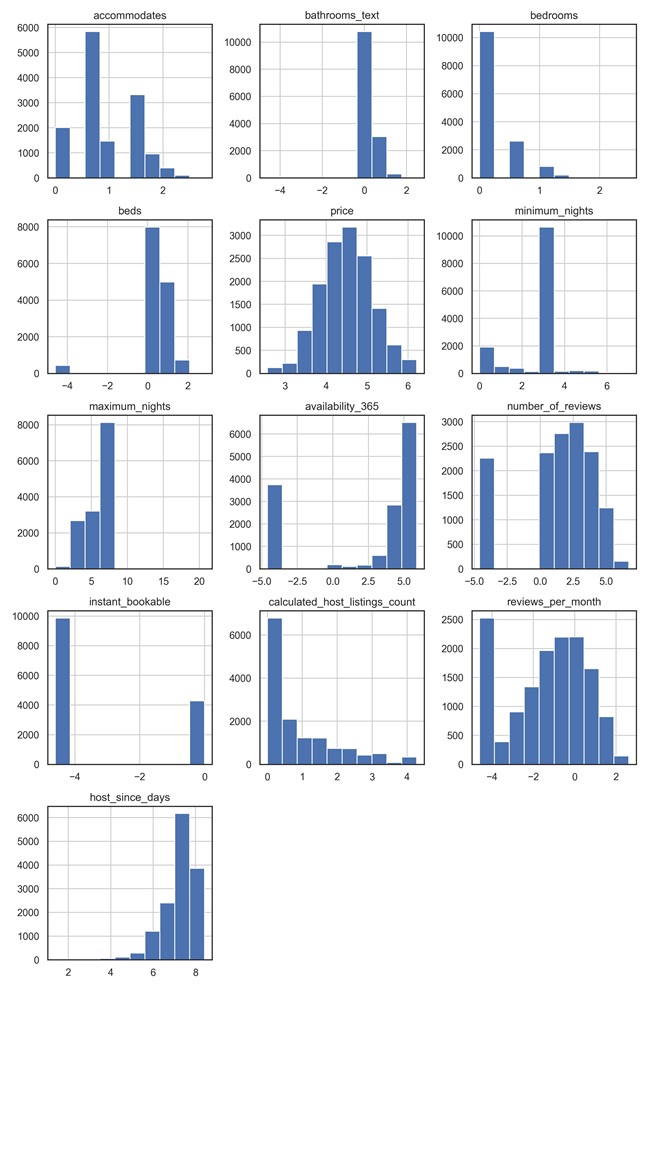
Anexo 17. Frequência e impacto ne tranca no quarto (lock on bedroom door).



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.